CRANFIELD UNIVERSITY

Grant Campbell

THE APPLICATION OF DIGITAL SOIL MAPPING TO IMPROVE THE RESOLUTION OF NATIONAL SOIL PROPERTIES ACROSS GREAT BRITAIN

School of Water, Energy and Environment (SWEE)

Ph.D.
Academic Year: 2018-2019

Supervisors: R. Corstanje and J.A. Hannam (Cranfield University)
H.I.J. Black and A. Lilly (The James Hutton Institute)

October 2018

ABSTRACT

Many countries have created soil maps to illustrate the variety of soil properties and support how soils can be used. Traditional soil mapping by field survey and interpretation has been the most recognised form of soil mapping for many years and an effective way to capture a variable soil landscape. Such maps have enabled scientists and stakeholders to improve their understanding of relationships between soils and other landscape factors such as geology and land cover. However, with the amount of soil information growing and technology improving, Digital Soil Mapping (DSM) has been developed as an alternative approach to generate soil property predictions and to produce finer resolution soils data. Currently, DSM produces maps based on training of models with observed soils data and environmental covariates and then releases these to stakeholders to evaluate their utility. This PhD has taken a different approach by addressing stakeholder needs at the beginning of the process.

The overall aim of this PhD was to improve the spatial resolution of soil properties across Great Britain (GB), as informed by stakeholders. Three main aims were identified. The first assessed what current soils data and information stakeholders currently use, and what improvements they want to see from future soil-related products. The second aim, using information from the questionnaire survey and a comparison of laboratory and analytical methods, is to develop DSM which could be applied across the whole of GB. This was done by comparing two modelling approaches: Boosted Regression Trees (BRTs) and Multivariate Adaptive Regression Splines (MARS) for mapping soil properties (loss-on-ignition, texture and pH) across two pilot areas. The characteristics of MARS and BRT models at both training and deployment stages are examined. The third outcome investigated how well the soil properties mapped across GB, building on the development of DSM in the pilot areas and whether they reflect expert pedological knowledge. This section also focusses on how suitable an independent validation dataset is at evaluating soil property predictions.

This PhD has shown that stakeholders are aware of what soils data and information they are using and could clearly express what is needed to improve current maps. Wider use of soil information by non-soil experts would be improved by increasing data accessibility and user-friendly supporting materials. Fundamentally, most stakeholders require finer resolution than what is currently available which identifies an opportunity for DSM to fill some of this gap.

To address these gaps and develop DSM across GB, this PhD focussed on mapping soil properties that were directly comparable across Scotland and England & Wales and also key to many stakeholder information needs. After investigation of laboratory and analytical methods from the two national soil surveys of Scotland and England & Wales, soil loss on ignition, soil texture and soil pH were chosen for developing DSM for GB.

From the development of DSM, results showed that MARS models produced better statistical performances than BRTs for predicting soil properties within a training environment. However, when MARS models are deployed to larger areas, they extrapolate beyond their means and BRTs performed better. This is because MARS models perform more consistently when many variables are required. Furthermore, MARS models struggle with overfitting and missing data which subsequently leads to incorrect and unfeasible pedological relationships between soil properties. BRT models, despite not performing as well statistically, produce more consistent relationships between pedology and mapped soil properties. This is because BRT models introduce randomness in the boosting which reduces overfitting and improves the predictive performance. BRTs have shown to be more consistent in the mapping outputs than MARS because regressing to the mean is more favourable when most data matches up with one another. However, this does not necessarily mean that the full range of soils in these areas were being captured by the BRT model. This led to scaling up from the pilot areas to modelling soil properties across GB using a single regional BRT model and evaluating its performance.

BRT modelling results for GB at 2D and 3D predict well for pH and LOI but less so for texture. Going forward, more data are required to produce more consistent modelling outputs especially for areas across GB where soil properties are not currently being predicted well.

The GB modelling results also highlighted a poor performance of the model against an independent validation dataset. This is because the original data for both GB training and validation datasets were analysed and collected for different purposes. These datasets were taken at different time periods under a different sampling design. Furthermore, the data for both training and validation GB datasets were collected at different scales.

At present, these first versions of soil property DSM maps for GB have produced variable results. However, this exercise has shown that the development of reliable DSM maps would benefit from interaction between pedologists, modellers and stakeholders to ensure that mapped outputs are of sufficient quality at adequate finer resolution and can be usable. Such DSM maps, alongside management recommendations, will help to address many global challenges associated to soils. However, DSM is not the panacea for all mapping needs. Until such time that DSM fully develops into DSA and adequately incorporates the breadth of information available in traditional soil maps, mapping from field survey and observation will continue to be necessary for stakeholders.

Keywords: Digital Soil Mapping, stakeholders, soil properties, finer resolution, pedology, Boosted Regression Trees (BRTs), Multivariate Adaptive Regressive Splines (MARS), Great Britain, validation

ACKNOWLEDGEMENTS

This research work was conducted as part of a joint studentship by The James Hutton Institute and Cranfield University and co-funded by the Macaulay Development Trust and the Engineering and Physical Sciences Resource Council (EPSRC).

I must give my utmost appreciation to my fantastic supervision team of Professor. Ronald Corstanje and Dr. Jacqueline Hannam from Cranfield University and Dr. Helaina Black and Dr. Allan Lilly at the James Hutton Institute in Aberdeen, for their valued opinions throughout my PhD journey. I must also thank the now-retired Dr. Thomas Mayr for his input at various stages throughout my PhD. I must acknowledge the following people for their useful help, advice and friendly support along the way. At Cranfield: Dr. Stephen Hallett, Dr. Toby Waine, Prof. Jane Rickson, Prof. Jim Harris, Miss Joanna Zawadzka, Dr. Ezekiel Okonkwo, Dr. Oliver Pritchard, Dr. Fiona Fraser, Mr Tom Storr and Dr. Rebecca Whetton. At the James Hutton Institute: Dr. Matt Aitkenhead, Malcolm Coull, Dave Miller, Dr. David Miller, Willie Towers, Dr. Benjamin Butler, Dr. Nikki Baggaley, Dr. Zisis Gaskas, Dr. Laura Poggio, Dr. Katrin Prager, Dr. Mads Troldborg, Dr. Thomas Freitag. I must also thank the many people who have been there for me when times have been tricky, particularly Ian Alexander, my PhD pals Joseph Palmer and Tom Inglis and my ‘Hutton Mum’ Andrea Issott. Thanks for the friendly discussions when the chips have been down!

I’d like to dedicate my PhD to my loving family who have had to endure my highs and lows throughout notably: Shirley, Alistair, Calum, Hazel and Henry who have provided so much support and much needed hugs at stressful times! There are so many other good people who have offered me support and motivation along the way – I’ve mentioned too many so far, so I’ll just thank you all as a group here and hopefully you won’t all kill me!

“Leave nothing on the table”

TABLE OF CONTENTS

This Doctoral thesis has been structured as a collection of papers rather than as a single monograph. The publications from this research are in Appendix 7.

LIST OF FIGURES

LIST OF TABLES

GLOSSARY OF TERMS

Boosted Regression Trees A set of statistical models which utilises algorithms from the regression trees from the Classification and Regression Tree (CART) group of models and the boosting and combining a collection of models.

Covariates A set of environmental attributes measured at any location across the area of interest used to explain soil variation improving digital soil mapping prediction.

Digital Terrain Model An illustration of continuous elevation values over a topographic surface area.

Digital Soil Mapping The creation and population of spatial soil information systems by numerical models…inferring the spatial and temporal variations of soil types and soil properties…from soil observation and knowledge and related environmental variables (McBratney et al, 2003; Lagacherie et al, 2006).

Geographical Information System A collection of computer software and data used to model spatial processes, analyse spatial relationships and view and manage information about geographical places.

Major soil sub group The division of major soil group which separates different types of similar soils (e.g. alluvial soils – mineral or peaty).

Modelling the use of mathematical equations to simulate and predict real events and processes (Grunwald, 2009).

Multivariate Adaptive Regression Splines A range of techniques used for solving regression and classification issues, with the intention to predict the value of a set of dependent variables from a set of independent variables (Friedman, 1991).

Pedometrics Branch of soil science which deals with uncertainty associated in soil models due to deterministic or stochastic variation and lack of knowledge of soil properties and processes (McBratney et al, 2000).

Raster A spatial dataset that is made up by an array of equally sized cells arranged in rows and columns and made up of single or multiple bands.

Resampling The process of interpolating new values to a cell when transforming rasters to a new cell size.

Scale the measuring tool through which a landscape may be viewed or perceived (Levin, 1992).

Soil A 3D natural body found on Earth's surface which supports all organisms with properties resulting from factors such as climate and geology acting on earthy parent material.

Soil Series A group of soils, formed from a parent material, having horizons that, except for the texture of the A or surface horizon, are similar in all profile characteristics and in arrangement in the soil profile. Among these characteristics are colour, texture, structure, reaction, consistence, and mineralogical and chemical composition (USDA, 1990).

Terrain Attributes Data which is characteristic of the land surface usually derived from a DTM.

ABBREVIATIONS

2D	**Two - dimensional**
3D	**Three - dimensional**
AH	**Analytical Hillshading**
AMT	**Annual Mean Temperature**
AP	**Annual Precipitation**
BRT	**Boosted Regression Trees**
BSTC	**British Soil Texture Classification**
CART	**Classification and Regression Trees**
CI	**Convergence Index**
CNBL	**Channel Network Base Level**
DSA	**Digital Soil Assessment**
DSM	**Digital Soil Mapping**
DTM	**Digital Terrain Model**
FAO	**Food and Agricultural Organisation**
GB	**Great Britain**
GIS	**Geographical Information System**
GPS	**Geographical Positioning System**
GSM	**Global Soil Map**
GSP	**Global Soil Partnership**
HOST	**Hydrology of Soil Types**
ISO	**Isothermality**
LandIs	**Land Information System**
LCA	**Land Capability for Agriculture**
LCM	**Land Cover Map**
LOI	**Loss on Ignition**

LSFact	**Land Slope Factor**
MARS	**Multivariate Adaptive Regression Splines**
MDR	**Mean Diurnal Range**
MODIS	**Moderate Resolution Imaging Spectroradiometer**
NSI	**National Soils Inventory**
NSRI	**National Soils Research Institute**
NSIS	**National Soils Inventory of Scotland**
NVZ	**Nitrate Vulnerable Zones**
PDD	**Potential Drainage Density**
PLSR	**Partial Least Square Regression**
PSD	**Particle Size Distribution**
RMSE	**Root Mean Square Error**
RF	**Random Forest**
SOTER	**Soil Terrain Database**
SP	**Seasonal Precipitation**
ST	**Season Temperature**
TWI	**Topographic Wetness Index**
USDA	**United States Department of Agriculture**
VDCN	**Valley Depth to Channel Network**
WRB	**World Reference Base**
XSCurve	**Cross Sectional Curvature**

1 INTRODUCTION AND LITERATURE REVIEW

1.1. Introduction

Many countries have created soil maps to determine the variety of soil types (Bouma, 1989). In the mid-1880s, Russian scientist Vasily Dokuchaev proposed that the variation in soil type could be explained by factors such as parent material, climate, topography and time for soil formation. Several scientists subsequently implemented Dokuchaev's approach, most notably Hans Jenny (1941), who developed a mathematical equation to describe soil formation and define the spatial characteristics of soils (Textbox 1). It is the influential work of these pedological scientists that is now being adapted to address a growing demand for spatial soils information (Campbell et al, 2017; Grealish et al, 2015). This approach is relevant for this PhD as it explores the application of soil formation principles to modelling and mapping soils across Great Britain (GB).

$$s= f\,(cl, o, r, p, t)$$

Where s = soil

f= a function of

cl = climate,

o =organisms,

r = relief,

p= parent material

t = time

Textbox 1: Hans Jenny's equation for soil formation.

GB has two national soil surveys which have collected and mapped soils: The Soil Survey of England and Wales and the Soil Survey of Scotland. These extensive datasets are an ideal opportunity to explore and develop spatial modelling and mapping of soils. However, there is a need from stakeholders to provide unified GB soil maps (Mayr et al, 2006). There are also scientific and technical challenges to address in unifying soils data across the border, which these surveys provide an opportunity to explore.

1.1.1. Why do we map soils?

Soil maps are graphical representations of information about the spatial distribution of soil features (Yaalon, 1989). The earliest of these soil maps were produced in the 19th century to reflect agricultural-related activities. These were designed to map out areas where soil attributes had associated relationships with land use (Brevik et al, 2016). It is important that soils are mapped across different landscapes because there are many things humans and plants require which are reliant on soil e.g. agriculture, food and water (Yaalon, 1989; Blum, 2005). Soils are variable in 2D and 3D and are dynamic within the natural environment. There are known factors which interact alongside soils (e.g. land cover, geology, topography) (Brevik et al, 2016). Therefore, mapping soils are useful to provide a synthesis of a complex system which can be hard to describe and predict.

At present, most current maps are not appropriate enough for their required use (Brevik et al, 2016). Primarily, the scale of certain maps may not be appropriate depending on what information people require. The accuracy of current maps is variable and does not provide a quantitative output of how accurate the map is. Overall current maps are reasonable at 2D scale but fail to consider the dynamic nature of soils at 3D scales. This is a critical omission as soils and their associated properties change over 2D and 3D scales. Some soil maps may also have large gaps in coverage across areas. This is hugely dependent on where data has been collected, as some areas across a landscape have more data collected than others.

1.1.2. How do we map soils?

Traditional soil mapping is the most recognised way in which a soil landscape is captured (Scull et al, 2003). This involves making observations of soils and associated rock type often with the aid of aerial photographs. Observations based upon the surveyor’s expert knowledge are recorded to determine where to make inspections (Hudson, 1992). Soil profile pits and auger borings provide information of changes in soils across landscapes at 2D and 3D scales. These are used to produce and map descriptions of the chemical, biological and physical analyses of soil types.

Soils can be mapped at detailed scales (e.g. 1:10,000) to illustrate soil patterns in individual fields. However, soils can also be mapped across smaller scales (e.g. 1:500,000) to provide a general understanding of the soils across a country, continent or at global scales (FAO, 2018). In GB, soil maps are available at 1:250,000 scale, which are widely used for policy and decision-making at national level. Only a quarter of England and Wales is covered by finer scale maps (1:50,000 and 1:25,000) and very few at the scale of 1:10,000 (Mayr et al, 2006).

Brevik et al, (2016) provide a historical and present-day review of traditional soil mapping and Rossiter (2005) highlights the importance of mapping soils by creating a resource inventory. Most soil maps were originally created from soil surveys (Soil Survey Staff, 1993; Hartemink et al, 2013). In many cases, these soil maps are often the most useful way to access soil information (Valentine et al, 1981) but they vary in consistency and quality, with a great variety in soil classifications used and little validation of mapping accuracy. Such maps cannot provide detail on specific sites (Rossiter, 2005; Brevik et al, 2016) and lack exact georeferencing (Dobos et al, 2006; Jones et al, 2005). In addition, these map units were generalised to reflect information that soil mappers could interpret from base maps and field observations (Simonson, 1952). A significant amount of uncertainty is based around traditional maps from soil surveys. Identifying uncertainty from traditional maps is challenging as often these reflect mental pedological models developed by individual surveyors which are mainly subjective (Hudson, 1992). This means that many factors cannot be measured, and interpretations made are done using the surveyor’s expert knowledge (Jones et al, 2005; Hudson, 1992). Traditional soils mapping is expensive, time-consuming to carry out and relies on dwindling expert knowledge. Therefore, many researchers are investigating alternatives to estimate soil attributes (Scull et al, 2003; McBratney et al, 2003).

1.2. Moving from traditional mapping to predicted techniques

As geospatial technologies and scientific understanding improved, tools such as Geographical Information Systems (GIS), remote sensing, Global Positional Systems (GPS) and spatial statistics have greatly altered how soils have been mapped (Scull et al, 2003; McBratney et al, 2003). Increases in spatial information and tools have provided an opportunity to produce digital maps at finer resolutions and use larger amounts of spatial data from soil surveys (Brevik et al, 2016). The introduction of computers has led to increased opportunities to digitise and digitally manipulate soil data and maps and many traditional maps have now been digitised (Tomlinson, 1978). However, these have no statistical basis and as a result are not a digital soil map *per se* (Minasny and McBratney, 2016).

1.2.1. Pedometrics

Developments in computing technology have led to new statistical approaches being generated, making predictions of soil possible at unsampled locations (Mora Vallejo, 2008). This method is defined as pedometrics, which ‘deals with uncertainty in soil models that are due to deterministic or stochastic variation, vagueness and lack of knowledge of soil properties and processes’ (McBratney et al, 2000). A way of predicting soil properties is utilising geostatistics by interpolating soil properties from large numbers of observations collected over small areas (Goovaerts, 1999). Geostatistical models such as regression kriging (Voltz and Webster, 1990; Keskin and Grunwald, 2018) focus on extending to large areas where there are many spatial variations. Theoretically, geostatistical approaches are an advantage over data mining methods because their soil predictions integrate both correlations with the predictor variables and the spatial correlation between soil observations (Dobos et al, 2006).

Geostatistics were developed on densely sampled areas from field data (Hengl, 2006; McBratney et al, 2003; Minasny and McBratney, 2010). This information has been useful in developing geostatistics as soil properties vary over space. This spatial dependence can be quantified using variograms (Minasny and McBratney, 2010; McBratney and Gruijter, 1992). Variograms are models fitted to observed soil properties and these models can be used to express kriging weights and make a spatial prediction using kriging methods (McBratney and Gruijter, 1992). Various forms of kriging have been developed to map soil properties. As technology has increased over time, kriging has been updated using detailed environmental covariates derived from satellite imagery and Digital Terrain Models (DTMs). Pedometrics has used further techniques such as fuzzy set theory (e.g. McBratney and Odeh, 1997; Odeh et al, 1994; de Menezes et al, 2013) based on expert knowledge or terrain attribute clustering.

1.2.2. Digital Soil Mapping

Maps produced by kriging or other statistical-based approaches are an advancement on traditional mapping, as they provide quantitative understanding and new data can be updated on a regular basis. However, these approaches rarely take pedological understanding into account. Therefore, Digital Soil Mapping (DSM) has been developed to generate predictions of soil properties at unsampled locations by manipulating GIS data and considering Jenny's soil forming factors. This is based on the SCORPAN methodology which has developed Jenny's soil formation equation (Textbox 2; McBratney et al, 2003; Lorenzetti et al, 2015).

DSM can be defined as 'the creation and population of spatial soil information systems by numerical models…inferring the spatial and temporal variations of soil types and soil properties…from soil observation and knowledge and related environmental variables' (McBratney et al, 2003; Lagacherie, 2007).

$$S_c = f(s, c, o, r, p, a, n) + e$$

Where

S_c = the soil class or attribute to be modelled,

f= function

s= refers to existing soil information,

c = climatic condition at the site,

o = organisms,

r =local relief,

p = parent materials,

a =soil age,

n= spatial topology or spatial relationship

and e = associated error

Textbox 2: SCORPAN equation for DSM (McBratney et al, 2003)

DSM utilises measured soils data or surveyed information and relates these to available covariate data such as climate, land cover, relief and parent material. This is used to develop and apply models to predict soil properties in 2D and 3D (Behrens and Scholten, 2006; Dobos et al, 2006; Balkovic et al, 2013). DSM can also be referred to as predictive soil mapping and soil-landscape modelling in the literature (Scull et al, 2003), and has been summarised by Carré et al, (2007), (see Figure 1.1). DSM is mainly focussed on a range of mathematical techniques based on discovering, from a training dataset, relationships between the predictor variables and the predicted variable (Dobos et al, 2006; Crivelenti, et al, 2009). The most frequently used models by the DSM community centre around multiple regression models (e.g. Moore et al, 1993; Odeh et al, 1994), classification trees (Hastie et al, 2001) and neural networks (Behrens, 2005).

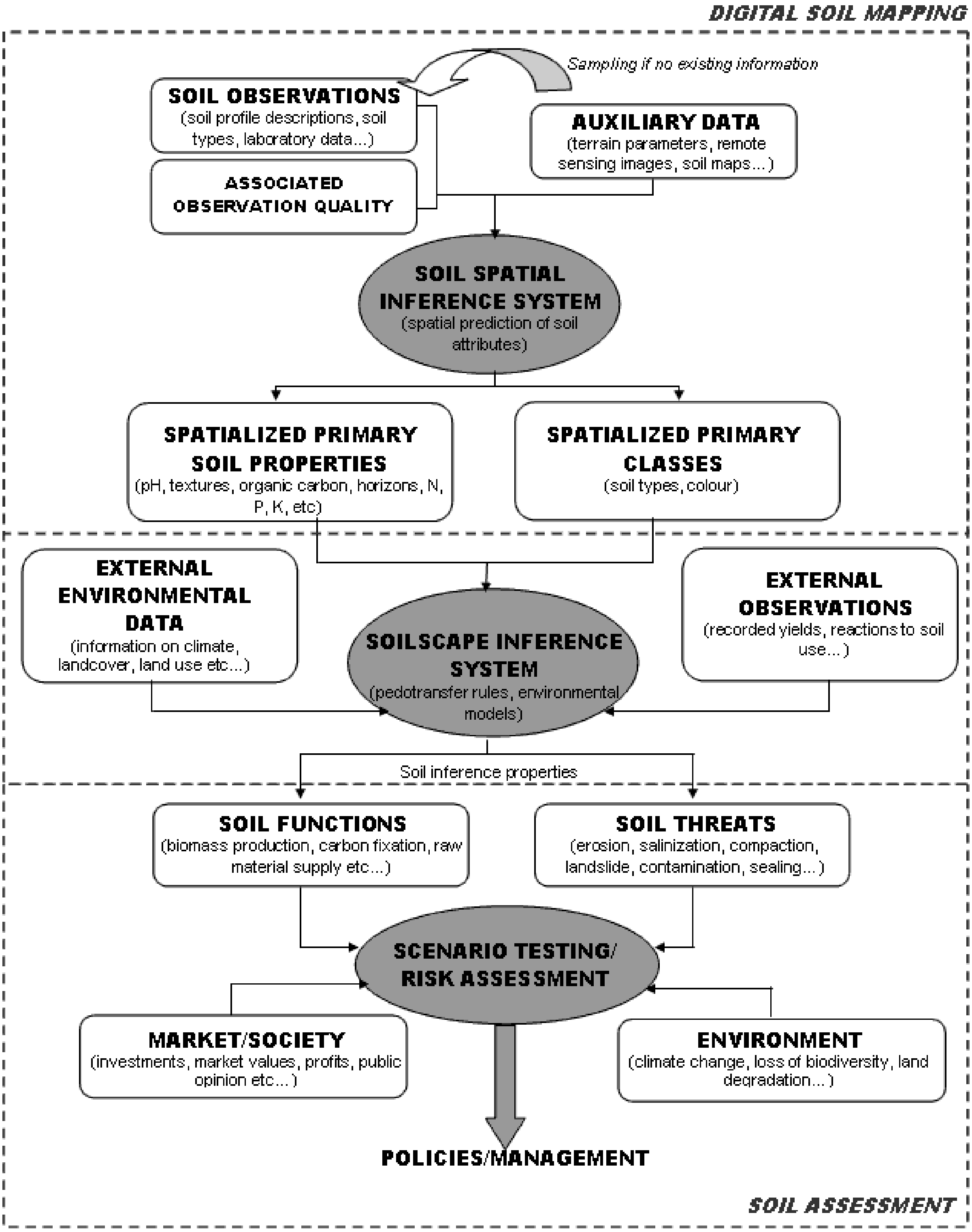

DIGITAL SOIL MAPPING
Sampling if no existing information
SOIL OBSERVATIONS (soil profile descriptions, soil types, laboratory data...)
AUXILIARY DATA (terrain parameters, remote sensing images, soil maps...)
ASSOCIATED OBSERVATION QUALITY
SOIL SPATIAL INFERENCE SYSTEM (spatial prediction of soil attributes)
SPATIALIZED PRIMARY SOIL PROPERTIES (pH, textures, organic carbon, horizons, N, P, K, etc)
SPATIALIZED PRIMARY CLASSES (soil types, colour)
EXTERNAL ENVIRONMENTAL DATA (information on climate, landcover, land use etc...)
SOILSCAPE INFERENCE SYSTEM (pedotransfer rules, environmental models)
EXTERNAL OBSERVATIONS (recorded yields, reactions to soil use...)
Soil inference properties
SOIL FUNCTIONS (biomass production, carbon fixation, raw material supply etc...)
SOIL THREATS (erosion, salinization, compaction, landslide, contamination, sealing...)
SCENARIO TESTING/ RISK ASSESSMENT
MARKET/SOCIETY (investments, market values, profits, public opinion etc...)
ENVIRONMENT (climate change, loss of biodiversity, land degradation...)
POLICIES/MANAGEMENT
SOIL ASSESSMENT

Figure 1.1: Digital Soil Assessment (including digital soil mapping) flow diagram (Carré et al, 2007)

1.2.3. DSM Critique

Scale is an important factor to consider when critiquing DSM. Scale is defined by Levin (1992) as ‘the measuring tool through which a landscape may be viewed or perceived’. In DSM context, scale is regarded as ‘the physical dimension of a phenomenon or process in space expressed in spatial units (pixel resolution and window size)’ (Cavazzi et al, 2013). Detailed reviews of scale can be found in work by Lam and Quattrochi (1992) and Goodchild and Quattrochi (1997); highlighting scale in four ways: cartographical, geographical, operational and spatial resolution. The issue of resolution and scale are important aspects in mapping. As Lam and Quattrochi (1992) argue, more than a single scale of observation may exist, thereby requiring measurement at several levels of resolution. This is particularly necessary due to a shift towards mapping soil functions, which will incorporate different resolutions and scales of data.

There has been a rapid evolution from traditional soils mapping to DSM over the past decade or so due to an ever-increasing need for accurate, reliable and quantified soil information (McBratney et al, 2003; Bui and Moran, 2003; Grunwald, 2009; Shi et al, 2009). McBratney et al, (2003) conducted a comprehensive literature review of DSM and these methods are now widely used in the soil science community.

Many authors have noted that DSM is an efficient way to update soils data based on improved resolution and understanding for mapping soil properties such as carbon, pH and texture (Scull et al, 2003; Hudson, 1992). DSM is useful in producing finer resolution outputs and can be linked to other global environmental datasets (Hengl et al, 2017). DSM is also noted as being less expensive than traditional soils surveying (Carré et al, 2007). However, this is assuming that field data are readily available for predictions. A major advantage of DSM is that it has huge capabilities of deriving uncertainties for predicted outcomes, allowing for errors to be tracked throughout the whole process: thereby producing increased quantitative understanding of soil variability as a level of precision (Carré et al, 2007; Scull et al, 2003).

DSM can be utilised alongside GIS support to provide improved accuracy assessments (Dobos et al, 2006; Sanchez et al, 2009).

However, there are known disadvantages to using DSM. A major weakness is that DSM is only as good as the input data used to develop the models for predictions. Currently, DSM is heavily reliant on legacy data from traditional soil survey with measurements, in some regions, which may be decades old (McBratney et al, 2003). Most DSM statistical methods are limited by the number of soils that can be mapped at once, while some processes can be hard to represent by available environmental variables (Brus et al, 2011). Successful DSM depends on the number of soil data points and environmental data layers. In areas where there is a lack of data, producing accurate maps with DSM has been challenging (Stoorvogel et al, 2009; Kempen et al, 2012). Ultimately, traditional field surveys and data collection will still be required to generate the basic data for DSM. However, different ways are needed to reduce the expense of field surveys. Alternative approaches have been suggested to characterise and measure soils in the field (e.g. remote sensing, digital data capture, in-field measurement of properties and processes and sensors) (Scull et al, 2003).

Uncertainty in DSM presents unique challenges (Brus et al, 2011). All maps, including DSM maps, are not 100% error free; and multiple error sources can be found (Heuvelink, 1998). These problems could potentially be magnified, so the final predictions and maps could contain large associated errors (Nelson et al, 2011). DSM is dependent on the statistical relationships between soil observations and environmental covariates at a range of locations. Most errors exist in the following:

- measurements in soil profiles.
- digitising of soil outlines on base maps.
- entering incorrect data into databases.
- classification of soil types and properties.
- generalisation of the overall sampled area.

- Interpreting or deriving connections between variables and soil properties (Heuvelink, 2014).

Understanding uncertainty in DSM is crucial as errors in decision-making can lead to serious consequences for all (Brus et al, 2011; Heuvelink, 2014). DSM produces quantifiable errors which can be explored to investigate where the main uncertainties lie. However, both digital soil maps and conventional maps produced previously contain errors, and are generally not validated (Grunwald, 2009; Brus et al, 2011).

Finally, DSM, in terms of its performance compared with traditional soil mapping is generally assumed to be an improvement. However, this is not always the case (Kempen et al, 2012; Bishop et al, 2001). Quite often, DSM maps provide poorer outputs than traditional maps. This is dependent on available data found in areas being investigated and the associated resolution. This has led to moving away from mapping specific soil properties to mapping associated functions.

1.2.4. Moving to Digital Soil Assessment (DSA)

Carré et al (2007) proposed moving beyond DSM to Digital Soil Assessment (DSA) (Figure 1.1), which involves integrating measured soils data with other datasets to produce 'functional' information more likely to be used by stakeholders e.g. agricultural land classification. DSA aims to provide a more dynamic understanding and mapping at both 2D and 3D scales. Many of these are 'functional' soil maps as they use information and data which influence functions such as agriculture and food production, environmental regulation and climate change (Blum, 2005). Such example maps include Land Capability for Agriculture (LCA) and the Hydrology of Soil Types (HOST) (Mayr et al, 2006; Boorman et al, 1995). However, functional maps require a lot of data, are difficult to deploy across larger areas and are problematic at understanding how the soil data or properties was converted into soil functional related information (Mayr et al, 2006).

Behrens and Scholten (2006) and Bouma et al (2012) reviewed the state of DSM in Germany and Holland. Both studies highlight an increasing demand for high-resolution maps to gain an improved understanding of environmental protection and land management issues. Mayr et al (2006) investigated methods of predicting soil functions within landscapes using available data and models in Great Britain. The main conclusions focus on acquiring more data than is currently available. Furthermore, there is an over-reliance on soil process models which provides difficulties when applied across different landscapes. DSM has the potential to overcome some of these constraints by considering the relationships between observed soil properties and other environmental covariates to improve the knowledge and understanding of soils areas.

In the future, DSA can be developed to improve DSM further as failing to do so could mean many digital soil maps become unusable (Finke, 2012). DSM can be useful in helping stakeholders make more informed decisions regarding policy and decision making, but it is vital that communicating this information is translated into usable materials which address specific stakeholder issues. Therefore, it is important that the requirements of end users of DSM maps are considered when developing new DSM products and when soil scientists are required to evaluate outputs (Finke, 2012).

Some global initiatives have been set up to enable new digital soil maps of the world to be produced using state-of-the-art approaches, responding to stakeholders needs for improved information on soils e.g. how much carbon is stored in soils or the pH of soil and suitability for agricultural purposes (Sanchez et al, 2009; Arrouays et al, 2014). An example of this is GlobalSoilMap.net (GSM) (GlobalSoilMap, 2011a, GlobalSoilMap, 2011b) which aims to investigate and understand the coverage of soil property data both at a 2D and 3D perspective at a much finer resolution than is currently available at present. There have already been many GlobalSoilMap products produced for regions and countries including Australia (e.g. Malone et al, 2014; Liddicoat et al, 2014; Searle, 2014), USA (e.g. Odgers et al, 2014), and France (Ciampalini et al, 2014; Vaysse et al, 2014).

1.3. Rationale of this work

Currently, DSM produces maps based on training of models with observed soils data and environmental covariates which is then released to stakeholders for them to consult and evaluate their utility. This PhD has approached this differently by addressing stakeholder needs at the beginning of the process. This is important to produce improved maps at finer resolution which match stakeholder expectations. It will be critical for stakeholders and the DSM community to consider outputs based on the modelling alongside relationships between soil properties and the underlying pedology. At present, limited work has been done to investigate approaches and consequences of carrying out DSM across different regions particularly when using different soils data sources with associated differences in methodologies. Furthermore, DSM uses a range of different models, but there has been little evaluation of which appropriate model(s) is best to be used. Within this PhD, there is an opportunity to explore these issues, particularly for GB where little DSM has been done across a national scale. Within GB, there were two distinct national soil surveys (Soil Survey of England and Wales and Soil Survey of Scotland), each with its own sampling and analytical approaches. Using data from both surveys, it would be useful to investigate whether there are appropriate ways to merge cross border soil properties and covariates for national scale DSM.

1.3.1. Principal aim

The main aim of this thesis is:

- to improve the spatial mapping and resolution of selected soil properties across GB informed by stakeholders.

1.3.2. Objectives

- The first objective is to explore what soils related data and information stakeholders currently use, and what desired improvements they want to see in soil information.
- The second objective is to develop DSM for GB, investigating how soil properties are mapped and modelled. This will involve using two recursive partition modelling

approaches across two pilot areas, with a view to comparing how both models work during model training and when deployed to wider areas.

- The third objective is to examine the DSM outputs across GB in relation to traditional mapping and a pedological understanding of the nature and distribution of soil and soil properties. It will also evaluate the suitability of an independent validation dataset on evaluating predictions of soil properties.

1.3.3. PhD chapter structure

After reviewing the current state of DSM in the academic literature and explained the knowledge gaps which this PhD plans to investigate, an outline is presented below, describing the content and structure of subsequent chapters.

- Chapter 2 reports on outcomes obtained from a questionnaire survey of non-soil science stakeholder needs across the UK and mainland Europe. This was to understand what soils data stakeholders currently use in their work or research, what issues they have whilst working with current data, and what improvements they would like to see to help support future work. This addresses objective one.
- Chapter 3 focusses on the collation, preparation and harmonisation of spatial soil property data required to support DSM across GB. Sampling and laboratory methods for key soil properties were investigated and compared, exploring the differences between methods and determination of harmonisation, if required. The requirements for data transformation are also discussed. This addresses part of objective two.
- Chapter 4 develops the DSM methodology for GB across two pilot areas: the north east Midland Valley in Scotland, and an area of west England and eastern Wales. DSM modelling and mapping of soil properties is explored using Boosted Regression Trees (BRTs) and Multivariate Adaptive and Regression Splines (MARS) models to compare the model performance during training and assess how the predicted maps of soil properties generated from the deployment of the models compares with other published soil maps. This also addresses part of objective two.

- Chapter 5 investigates DSM when applied to GB based on findings from Chapter 4. This section focuses on each stage of the DSM methodology; how well soil properties are mapped across GB and evaluating whether they reflect the pedological understanding of the nature and distribution of soils found in these areas. Residuals from an independent validation dataset were examined to evaluate how well each soil property is predicted across GB. This addresses objective three of the PhD thesis.
- Chapter 6 assesses the performance of DSM has done for mapping and modelling soil properties across GB, exploring the lessons learnt. The outcomes of the stakeholder survey in Chapter 2 are revisited and there is a discussion on whether results from DSM in GB could meet the stakeholder needs for finer resolution soil property maps in the future. The chapter also considers how we might move from DSM to Digital Soil Assessment.

References

Arrouays, D., McKenzie, N., Hempel, J., Richer de Forges, A., McBratney, A., 2014: GlobalSoilMap: Basis of the global spatial soil information system. Taylor and Francis Group: CRC Press.

Balkovic, J., Rampasekova, Z., Hutar, V., Sobocka, J. and Skalsky, R., 2013: Digital Soil Mapping from Conventional Field Soil Observations. *Soil and Water Research,* 8, 1, pp.13-25.

Behrens, T., Scholten, T., 2006: Digital soil mapping in Germany—a review. *Journal of Plant Nutrition and Soil Science*, 169, 3, pp. 434-443.

Behrens, T., Förster, H., Scholten, T., Steinrücken, U., Spies, E.-D., Goldschmidt, M., 2005: Digital Soil Mapping using artificial neural networks. *Journal of Plant Nutrition and Soil Science*, 168, pp.1-13.

Bishop, T.F.A., McBratney, A.B., Whelan, B.M., 2001: Measuring the quality of digital soil maps using information criteria. *Geoderma,* 103, 1, pp. 95-111.

Blum, W.E.H, 2005: Functions of soil for society and the environment. *Reviews in Environmental Science and Bio/Technology*, 4, pp. 75-79.

Boorman, D.B., Hollis, J.M., Lilly, A., 1995: Report No. 126 Hydrology of Soil Types; a hydrologically-based classification of the soils of the United Kingdom. Natural Environment Research Council.

Bouma, J, 1989: Using soil survey data for quantitative land evaluation. *Advances in Soil Science*, 9, pp. 177–213.

Bouma, J., Broll, G., Crane, T. A., Dewitte, O., Gardi, C., Schulte, R., and Towers, W., 2012: Soil information in support of policy making and awareness raising. *Current Opinion in Environmental Sustainability*, 4, 5, pp. 552-558.

Brevik, E.C., Calzolari, C., Miller, B. A., Pereira, P., Kabala, C., Baumgarten, A., Jordan, A., 2016: Soil mapping, classification, and pedologic modelling: History and future directions. *Geoderma*, 264, pp. 256-274.

Brus, D.J., Kempen, B., Heuvelink, G.B.M., 2011: Sampling for validation of digital soil maps. *European Journal of Soil Science*, 62, 3, pp. 394-407.

Bui, E.N., Moran, C.J., 2003: A strategy to fill gaps in soil survey over large spatial extents: an example from the Murray Darling Basin of Australia, *Geoderma*, 111, pp. 21-144.

Campbell, G.A., Lilly, A., Corstanje, R., Mayr, T.R., Black, H.I.J., 2017: Are existing soils data meeting the needs of stakeholders in Europe? An analysis of practice use from policy to field. *Land Use Policy*, 68, pp.211-223.

Carré, F., McBratney, A.B., Mayr, T., Montanarella, L., 2007: Digital soil assessments: Beyond DSM. *Geoderma*, 142,1–2, pp. 69-79.

Cavazzi, S., Corstanje, R., Mayr, T., Hannam, J., Fealy, R., 2013: Are fine resolution digital elevation models always the best choice in digital soil mapping? *Geoderma*, 195–196, pp. 111-121.

Ciampalini, R., Martin, M.P., Saby, N.P.A., Richer de Forges, A., Arrouays, D., 2014: Soil texture GlobalSoilMap products for the French region ‘Centre’, in: GlobalSoilMap: basis of the global spatial soil information system. Arrouays, D. et al., CRC Press, pp. 121–126.

Crivelenti, R.C., Coelho, R.M., Adami, S.F., de Medeiros Oliveira, S.R., 2009: Data mining to infer soil-landscape relationships in digital soil mapping. *Pesquisa Agropecuaria Brasileira*, 44, 12, pp. 1707-1715.

de Menezes, M.D., Godinho Silva, S.H., Owens, P.R., Curi, N., 2013: Digital Soil Mapping Approach based on fuzzy logic and field expert knowledge. *Ciencia E Agrotecnologia* 37, 4, pp. 287-298.

Dobos, E., Carré, F., Hengl, T., Reuter, H.I., Tóth, G., 2006: Digital Soil Mapping as a support to production of functional maps. EUR 22123 EN, Office for Official Publications of the European Community, Luxemburg.

Finke, P. A., 2012: On digital soil assessment with models and the Pedometrics agenda. *Geoderma*, 171: pp. 3-15.

GlobalSoilMap., 2011a: GlobalSoilMap.net: New Digital Soil Map of the world. [Accessed from https://www.google.co.uk/search?q=globalsoilmap.net&ie=utf-8&oe=utf-8&aq=t&rls=org.mozilla:en-GB:official&client=firefox-a&channel=sb&gfe_rd=cr&ei=aR3jVLWpOoSV8wOK4YHwBQ] [Last accessed 17th February 2015].

GlobalSoilMap., 2011b: Specifications Version 1: GlobalSoilMap.net products: Release 2.1. Technical report.

Goodchild, M.F. and Quattrochi, D.A., 1997: Scale in Remote Sensing and GIS. Lewis Publishers, pp. 1-11.

Goovaerts, P., 1999: Geostatistics in soil science: state-of-the-art and perspectives. *Geoderma*, 89, pp. 1 – 45.

Grealish, G.J., Fitzpatrick, R.W., Hutson, J.L., 2015: Soil survey data rescued by means of user-friendly soil identification keys and toposequence models to deliver soil information for improved land management. *GeoRes J*, 6, pp. 81-91.

Grunwald, S., 2009: Multi-criteria characterization of recent digital soil mapping and modelling approaches. *Geoderma*, 152, 3-4, pp. 195-207.

Hartemink, A.E., Krasilnikov, P., Bockheim, J., 2013: Soil maps of the world. *Geoderma*, pp. 207-208.

Hastie, T., Tibshirani, R., Friedman, J., 2001: The elements of statistical learning: data mining, inference and prediction. Springer Series in Statistics. Springer-Verlag, New York.

Hengl, T., Mendes de Jesus, J., Heuvelink, G.B.M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotic, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., Guevara, M.A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J.G.B., Ribeiro, E., Wheeler, I., Mantel, S. and Kempen, B., 2017: SoilGrids250m: Global gridded soil information based on machine learning. *PLoS ONE,* 12, 2: e0169748. https://doi.org/10.1371/journal.pone.0169748

Hengl, T., 2006: Finding the right pixel size. *Computers & Geosciences,* 32, 9, pp.1283-1298.

Heuvelink, G.B.M., 1998: Error propogation in Environmental Modelling with GIS. London: Taylor and Francis.

Heuvelink, G.B.M., 2014: Uncertainty quantification of GlobalSoilMap products in: GlobalSoilMap: basis of the global spatial soil information system. Arrouays, D. et al., CRC Press, pp. 433–439.

Hudson, B. D., 1992: Division S-5- Soil Genesis, Morphology & Classification: The soil survey as paradigm-based science. *Soil Science Society of America Journal*, 56, 3, pp. 836-841.

Jenny, H., 1941: Factors of Soil Formation: A system of quantitative pedology. Dover Publications, INC. New York.

Jones, R.J.A., Houskova, B., Bullock, P., Montanarella, L., 2005: Soil Resources of Europe, second edition. European Soil Bureau Research Report No.9, EUR 20559 EN, (2005), 420pp. Office for Official Publications of the European Communities, Luxembourg.

Kempen, B., Brus, D.J., Stoorvogel, J.J., Heuvelink, G.B.M., de Vries, F., 2012: Efficiency Comparison of Conventional and Digital Soil Mapping for Updating Soil Maps. *Soil Science Society of America Journal* 76, 6, pp. 2097-2115.

Keskin, H., Grunwald, S., 2018: Regression kriging as a workhorse in the digital soil mapper's toolbox, *Geoderma*, 326, pp. 22-41.

Lagacherie, P., 2007: Digital Soil Mapping: A State of the Art. In: Hartemink A.E., McBratney A., Mendonça-Santos M. (eds) Digital Soil Mapping with Limited Data. Springer, Dordrecht.

Lam, N.S., Quattrochi, D.A., 1992: On the issues of scale, resolution, and fractal analysis in the mapping sciences. *Professional Geographer,* 44, pp. 88-98.

Levin, S.A., 1992: The problem of pattern and scale in ecology. *Ecology*, 73, pp. 1943-1967.

Liddicoat, C., Kidd, D., Searle, R., 2014: Modelling soil carbon stocks using legacy site data, in Mid North region of South Australia. in: GlobalSoilMap: basis of the global spatial soil information system. Arrouays, D. et al., CRC Press.

Lorenzetti, R., Barbetti, R., Fantappié, M., L'Abate, G., Costantini, E.A.C., 2015: Comparing data mining and deterministic pedology to assess the frequency of WRB reference soil groups in the legend of small-scale maps. *Geoderma*, 154, pp. 138-152.

Malone, B.P., Minasny, B., Odgers, N.P., McBratney, A.B., 2014: Using model averaging to combine soil property rasters from legacy soil maps and from point data. *Geoderma*, 232-234, pp. 34-44.

Mayr, T., Black, H., Towers, W., Palmer, R., Cooke, H., Freeman, M., Hornung, M., Wood, C., Wright, S., Lilly, A., DeGroote, J., Jones, M., 2006: Novel methods for spatial prediction of soil functions within landscapes (SP0531). DEFRA, 26pp.

McBratney, A.B., Gruijter, J.d.,1992: A continuum approach to soil classification by modified fuzzy k-means with extragrades. *Journal of Soil Science*, 43, pp. 159-175.

McBratney, A.B., Odeh, I.O.A., 1997: Application of fuzzy sets in soil science: fuzzy logic, fuzzy measurements and fuzzy decisions. *Geoderma*, 97, pp. 293-327.

McBratney, A.B., Odeh, I.O.A., Bishop, T.F.A., Dunbar, M.S., Shatar, T.M., 2000: An overview of pedometric techniques for use in soil survey. *Geoderma*, 97,3–4, pp. 293-327.

McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003: On digital soil mapping. *Geoderma*, 117,1, pp. 3-52.

Minasny, B. and McBratney, A.B., 2010: Methodologies for global soil mapping. Digital soil mapping, Springer, pp. 429-436.

Minasny, B. and McBratney, A.B., 2016: Digital Soil Mapping: A brief history and some lessons. *Geoderma*, 264, pp. 301-311.

Moore, I.D., Gessler, P.E., Nielsen, G.A., Peterson, G.A., 1993: Soil attribute prediction using terrain analysis. *Soil Science Society of America Journal*, 57, pp. 443-452.

Mora-Vallejo, A., Claessens, L., Stoorvogel, J.J., Heuvelink, G.B.M., 2008: Small scale digital soil mapping in South-eastern Kenya. *Catena,* 76,1, pp. 44-53.

Nelson, M. A., Bishop, T.F.A., Triantafilis, J., Odeh, I.O.A., 2011: An error budget for different sources of error in digital soil mapping. *European Journal of Soil Science,* 62, 3, pp.417-430.

Odeh, I. O. A., McBratney, A.B., Chittleborough, D.J., 1994: Spatial prediction of soil properties from landform attributes derived from a Digital Elevation Model. *Geoderma*, 63, pp.197-214.

Odgers, N. P., Sun, W., McBratney, A.B., Minansy, B., Clifford, D., 2014: DSMART: An algorithm to spatially disaggregate soil map units. in: GlobalSoilMap: basis of the global spatial soil information system. Arrouays, D. et al., CRC Press, pp. 433–439.

Omuto, C., Nachtergaele, F., Rojas, R.V., 2013: State of the art report on Global and Regional Soil Information: Where are we? Where to go. Global Soil Partnership Technical Report.

Rossiter, D. G., 2005: Digital soil mapping: Towards a multiple-use Soil Information System. Geomatica [Accessed from http://www.css.cornell.edu/faculty/dgr2/Docs/CoGeo2005/PaperRossiterGeomatica2005.pdf] [Last Accessed 22nd August 2018].

Sanchez, P. A., Ahamed, S., Carré, F., Hartemink, A.E., Hempel, J., Huising, J., Lagacherie, P., McBratney, A.B., McKenzie, N.G., de Ourdes Mendonça-Santos, M., Minasny, B., Montanarella, L., Okoth, P., Pal, C.A., Sachs, J.D., Shephard, K.D., Vagen, T., Vanlauwe, B., Walsh, M.G., Winowiecki, L.A., Zhang, G.L., 2009: Digital soil map of the world. *Science,* 325, 5941, pp. 680-681.

Scull, P., Franklin, J., Chadwick, O.A., McArthur, D., 2003: Predictive soil mapping: a review. *Progress in Physical Geography*, 27, 2, pp. 171-197.

Searle, R. 2014: The Australian site data collation to support the GlobalSoilMap. in: GlobalSoilMap: basis of the global spatial soil information system. Arrouays, D. et al., CRC Press.

Shi, X., Long, R., Dekett, R., Philippe, J., 2009: Integrating different types of knowledge for digital soil mapping. *Soil Science Society of America Journal,* 73, 5, pp. 1682-1692.

Simonson, R.W. 1952: Lessons from the first half century of Soil Survey: II. Mapping of soils. *Soil Science*, 74, 4, pp. 323-330.

Soil Survey Staff, 1993: Soil Survey Manual. U.S. Department of Agricultural Handbook, No.18, Washington DC.

Stoorvogel, J.J., Kempen, B., Heuvelink, G.B.M., de Bruin, S., 2009: Implementation and evaluation of existing knowledge for digital soil mapping in Senegal. *Geoderma*, 149, 1–2, pp. 161-170.

Terribile, F., Coppola, A., Langella, G., Martina, M., Basile, A., 2011: Potential and limitations of using soil mapping information to understand landscape hydrology, *Hydrology and Earth System Sciences*, 15, pp. 3895-3933.

Tomlinson, R., 1978: Design considerations for digital soil map systems. 11th Congress of Soil Science, ISSS, Edmonton, Canada.

Valentine, K.W.G., Naughton, W.C., Navai, M., 1981: A questionnaire to users of soil maps in British Columbia: Results and implications for design and content. *Canadian Journal of Soil Science*, 61, pp. 123-135.

Vaysse, K, Arrouays, D., McKenzie, N.J., Coste, S., Lagacherie, P., 2014: Estimation of GlobalSoilMap.net grids cells from legacy soil data at the regional scale in Southern France. in: GlobalSoilMap: Basis of the Global Spatial Soil Information System. Arrouays, D. et al., CRC Press, pp. 133–138.

Voltz, M. and Webster, R., 1990: A comparison of kriging, cubic splines and classification for predicting soil properties from sample information, *Journal of Soil Science*, 41, pp.473-490.

Yaalon, D.H., 1989: The earliest Soil Maps and their logic. *Bulletin of the International Society of Soil Science*. International Society of Soil Science, Wageningen, 1989, p.24.

2 ARE EXISTING SOILS DATA MEETING THE NEEDS OF STAKEHOLDERS IN EUROPE? AN ANALYSIS OF PRACTICAL USE FROM POLICY TO FIELD

Abstract

Soils form a major component of the natural system and their functions underpin many key ecosystem goods and services. The fundamental importance of soils in the environment means that many different organisations and stakeholders make extensive use of soils data and information in their everyday working practices. For many reasons, stakeholders are not always aware that they are reliant upon soil data and information to support their activities. Various reviews of stakeholder needs and how soil information could be improved have been carried out in recent years. However, to date, there has been little consideration of user needs from a non-expert perspective. The aim of this study was to explore the use of explicit and hidden soil information in different organisations across Europe and gain a better understanding of improvements needed in soil data and information to assist in practical use by non-expert stakeholders. An online questionnaire was used to investigate different uses of soils data and information with 310 responses obtained from 77 organisations across Europe. Results illustrate the widespread use of soil data and information across diverse organisations within Europe, particularly spatial products and soil functional assessments and tools. A wide range of improvements were expressed with a prevalence for finer scale resolution, trends over time, future scenarios, improved accuracy, non-technical supporting information and better capacity to use GIS. An underlying message is that existing legacy soils data need to be supplemented by new up-to-date data to meet stakeholder needs and information gaps.

2.1. Introduction

Soils form a major component of our natural environment, performing an array of essential functions that underpin key ecosystem goods and services which plants and animals rely on (Costanza et al, 1997; Smith et al, 2015). The significance of soils within the environment has meant that stakeholders must use a wide variety of soils data and information in their decision making.

The concept of soil functions was first conceived during the early 1950s and has since been widely adopted in national and regional policy (Blum, 2005). From the mid-1900s onwards, soils functional aspects have been incorporated into assessment tools such as maps and models that assist decision makers across a wide range of soil-related issues from land use, cropping practises, protection of water bodies, and restoration of habitats to climate regulation. For instance, many assessments around agricultural productivity, such as the Land Capability for Agriculture in Scotland (Bibby et al, 1988) and laterally, the CAPRI model (Britz and Witzke, 2014), are based on soil maps. However, functional assessments have since extended across many other issues such as groundwater vulnerability (Environment Agency, 2013; Harter and Walker, 2001).

When exploring what needs to be improved in terms of soils data and information, it is important to understand the contemporary needs of stakeholders particularly where soils data and information may be implicit or part of an underlying model or assessment tool. There are various reviews of stakeholder needs and how these levels of information could be improved which have been carried out in recent years (Black et al, 2012, Prager and McKee, 2014, Valentine et al, 1981, Grealish et al, 2015, Omuto et al, 2013, GS Soil, 2010, Panagos et al, 2012). However, these reviews have generally assumed that stakeholders have some knowledge of soils or are fully aware that they are using soils data and information. The aim of this study is to understand soils data and information stakeholders' needs across Europe from a non-expert perspective.

Jones et al (2005) reviewed soils resources and information use across Europe and determined that these are traditionally associated through the function of food and fibre production, with increasing applications to other issues such as climate change and water resource management (Blum, 2005; Grealish et al, 2015, Haines-Young, 2012). Soil maps, data and information are used in many sectors besides soil science, such as farming, hydrology, land degradation, policy and environmental modelling (Valentine et al, 1981, Mather, 1988, GS Soil, 2010, Hallett et al, 2011, Omuto et al, 2013, Prager and McKee, 2014). Most soil information users indicated that key soil attributes are readily available (Wood and Auricht, 2011). However, improvements in a range of soil properties such as soil moisture, toxicity, biology and carbon are required (Auricht, 2004, Grealish et al, 2015).

Furthermore, engineering properties such as subsidence and corrosion are also of interest (Pritchard et al, 2015). These types of information are available but awareness of data accessibility and where to find them remains challenging. Information needs are also specific to stakeholder requirements and the spatial resolution of the undertaking. Black et al (2012) consulted a wide range of stakeholders in developing the Soil Monitoring Action Plan for Scotland with further consultation taking place with farmers and local authorities by Prager and McKee, (2014). Key improvements mentioned were finer spatial resolution, soil trends, soil biological and physical indicators and the extent of sealing.

The FAO (2012) identify three major challenges in addressing soil information availability. The first of these focusses on the importance of soil protection, particularly to the global modelling community as it will help mitigate and adapt to issues such as climate change and food security. A second consideration is soil monitoring, focusing on improving global soil data at finer scale resolution. The third examines advancing Digital Soil Mapping (DSM) and Digital Soil Assessment (DSA) techniques. DSM and DSA offers potential to map soil properties at detailed and broad scales (McBratney et al, 2003; Behrens and Scholten, 2006; Carré et al, 2007). However, it is not clear how any of these challenges reflect the needs of stakeholders, and difficulties remain around integrating the capability of models and the envisioned users of this data.

Stakeholder interaction and participation should be considered from the outset, and this is very rarely done (Reed, 2008). Studies by Bouma (2012) and Black et al (2012) highlighted that end-users were often not aware that they were using soils data and information so could not easily communicate further needs. It is therefore not straightforward to assume what the needs of envisioned users of ‘new’ soil information are, where this information is embedded in derived tools. A survey of non-expert users was designed to investigate their current needs and perceived gaps in their ability to deliver in their work activities. This information is vital in addressing how new soil tools and products, such as DSM and DSA, might (or might not) meet the stakeholder requirements and the likelihood of such products being of practical use. The aim is therefore to investigate what soils assessments and tools stakeholders currently use and what improvements, if any are required for future soil products/information sets.

2.2. Methodology

A detailed questionnaire was carried out to consider the range of soils data and information currently being used across Europe with a focus on explicit and hidden soils information being used by non-expert stakeholders: non-experts are people who use soils information or data in their everyday work but who are not expected to be academically trained soil scientists (Appendix 1).

The questionnaire was compiled using the web-based survey programme Qualtrics (Qualtrics, 2018: http://www.qualtrics.com/). In addressing the different uses of soils data and information, it was considered important to address functions of soils and contact stakeholders with close connections in and around these functions. Therefore, stakeholders were identified in order to be representative of the primary functions of soils (FAO, 2015: http://www.fao.org/resources/infographics/infographics-details/en/c/284478/) including biomass production, cultural heritage, regulating, biodiversity/habitats and infrastructure. A list of organisations across Europe, with named soil contacts, was draw up by accessing published materials, on-line searches and personal knowledge. The remit and primary activities of these organisations corresponded well with at least one of the soil functions and provided coverage across the soil functions. Stakeholders were based around commercial organisations, learned societies, non-governmental organisations (NGOs), local authorities and government organisations. A total of 98 organisations were contacted across 22 countries in Europe. Of these, 34 organisations can be considered trans-European in their activities i.e. no specific alignment with any one region or country. A pilot study of the questionnaire was conducted with staff at The James Hutton Institute (Aberdeen) and the Scottish Government's ethics committee; the questionnaire incorporated amendments following relevant feedback. The survey was carried out from July to August 2015 and was made accessible to stakeholders through an anonymous online link.

2.3. Questionnaire Results

2.3.1. What sectors use soils information?

There were 310 individual responses to the questionnaire from 77 out of the 98 organisations contacted and, from this, 93% of stakeholders said that they handled information about soil in their work.

Stakeholders were asked to identify what best describes the activities of their organisation. Stakeholders could tick more than one option for this question to obtain a broader understanding of activities associated with individual organisations. The top three activities were agriculture, research organisations (universities, institutes etc.) and conservation (Figure 2.1). Stakeholders who ticked 'other' ranged from people who worked in landscape photography, archaeology and oil and gas services. This shows that there is a wide array of stakeholders who have an interest in soils data and information and who may use certain tools and assessments related to activities within their organisation.

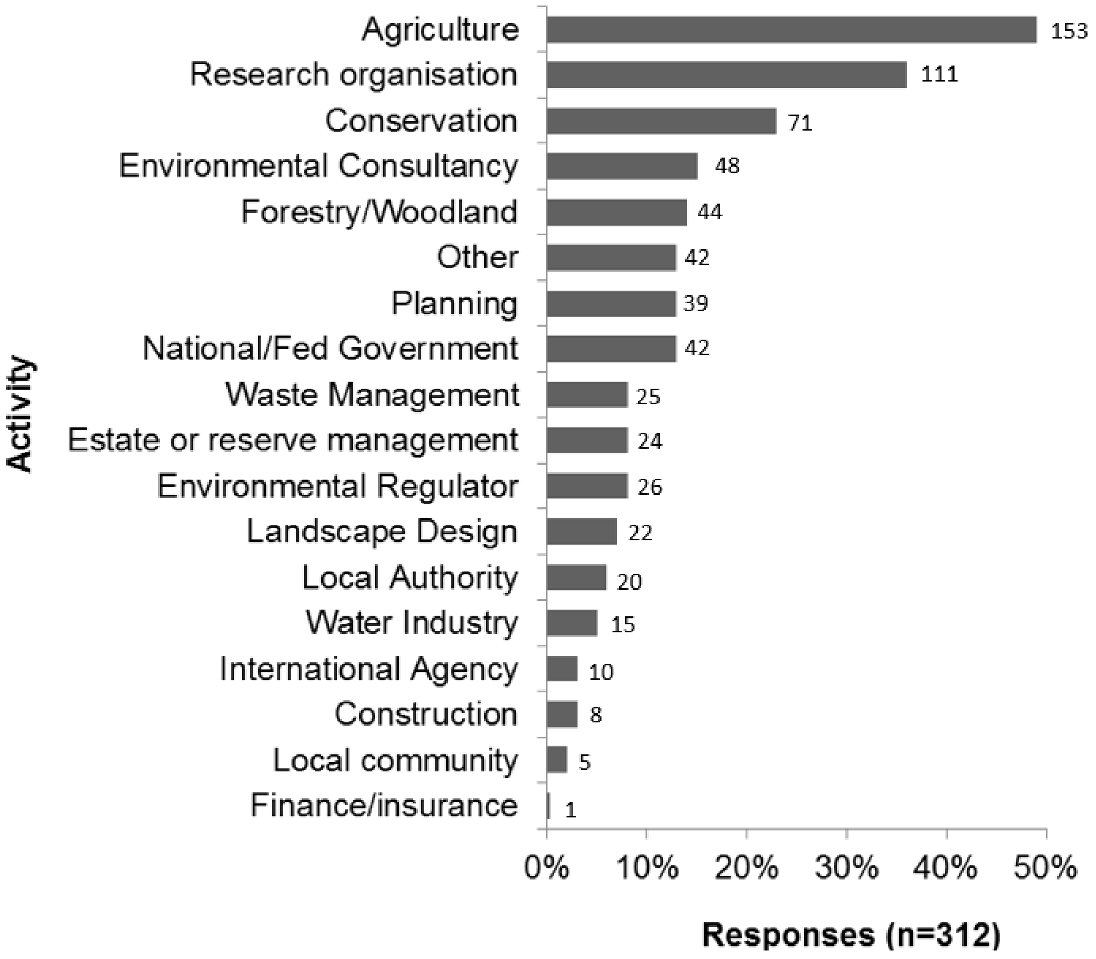


Figure 2.1: Range of organisations and the percentage of responses to the questionnaire (numbers for each are noted for each activity. This was to get an understanding as to the variety of organisations people worked for. N.B. Stakeholders could tick more than one option for this question.

2.3.2. Tools and assessments and awareness of embedded soils information

Stakeholders were encouraged to tick as many boxes as possible in terms of what tools and assessments they use in their work. These assessments are grouped by related soil functions. Most responses came from people who related to agricultural production and conservation of habitats and biodiversity. Respondents were asked about how aware they were that many of the assessments had soils information embedded within them, with 87% saying that they were '*aware*'.

In relation to ‘*Biomass Production’*, it was found that the two main tools predominantly used were agricultural land evaluation and fertiliser/pesticide usage assessments. In terms of assessments grouped under ‘*Infrastructure*’, it is the extraction of raw materials such as clay, sand and silt, followed by assessment of the impacts of soils on assets such as pipes and electric cables. Nitrate Vulnerable Zones (NVZs) were found to be the main assessment tool used by stakeholders closely associated with ‘*Environmental Regulation’* with soil erosion and diffuse pollution to water following closely behind.

Habitat suitability maps and land restoration assessments were the most commonly used assessments by stakeholders related to ‘*Habitats and Biodiversity’*.

The number of stakeholders requesting information on fundamental soil properties from the questionnaire was relatively high. Soil chemistry (primary contaminants) and other properties including soil acidity, alkalinity and carbon had the highest demand and application (Figure 2.2). Several other assessments which were not listed in the survey were also used by stakeholders including soil climate zones to identify nutrient demands of crops and grasslands.

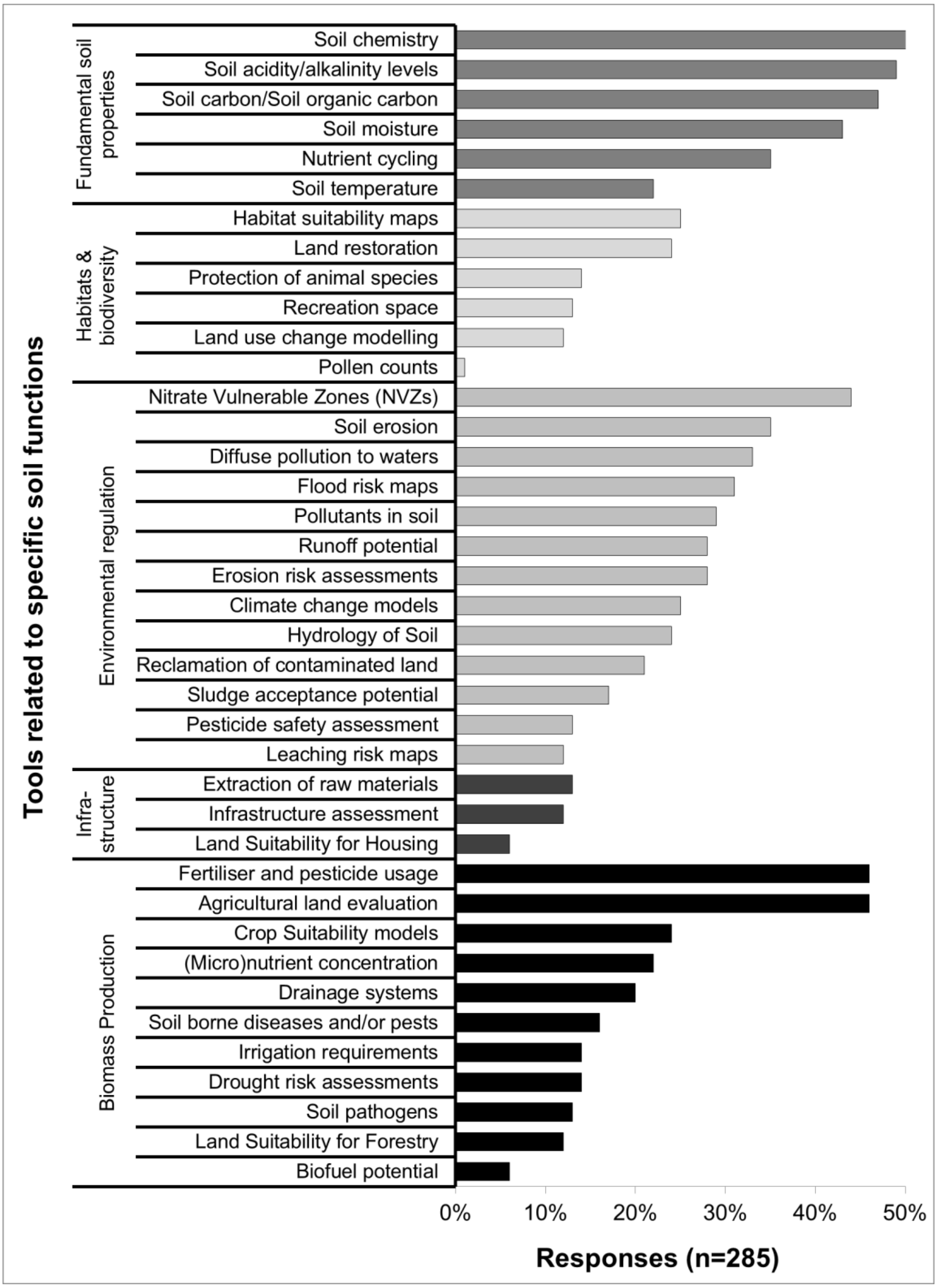


Figure 2.2: Tools and assessments and percentages used by respondents. These are broken up into their closest related function.

2.3.3. Sources of information, licencing and spatial importance

Respondents were asked to identify what sources they used to acquire soil information required for their work. The use of maps in either paper or digital format is the most prolific with 78% of respondents using them while 65% of respondents use Geographical Information Systems (GIS). Other sources consisted of social media websites and discussions with knowledge transfer exchange with stakeholders (11% of respondents).

Overall, most stakeholders found most sources that they used either '*very useful'* or '*useful*'. 95% found the use of maps, expert knowledge and field and laboratory analysis to be either *'useful'* or '*very useful'*. However, 11% reported that GIS systems were '*not very useful'* or '*not useful'* (Figure 2.3).

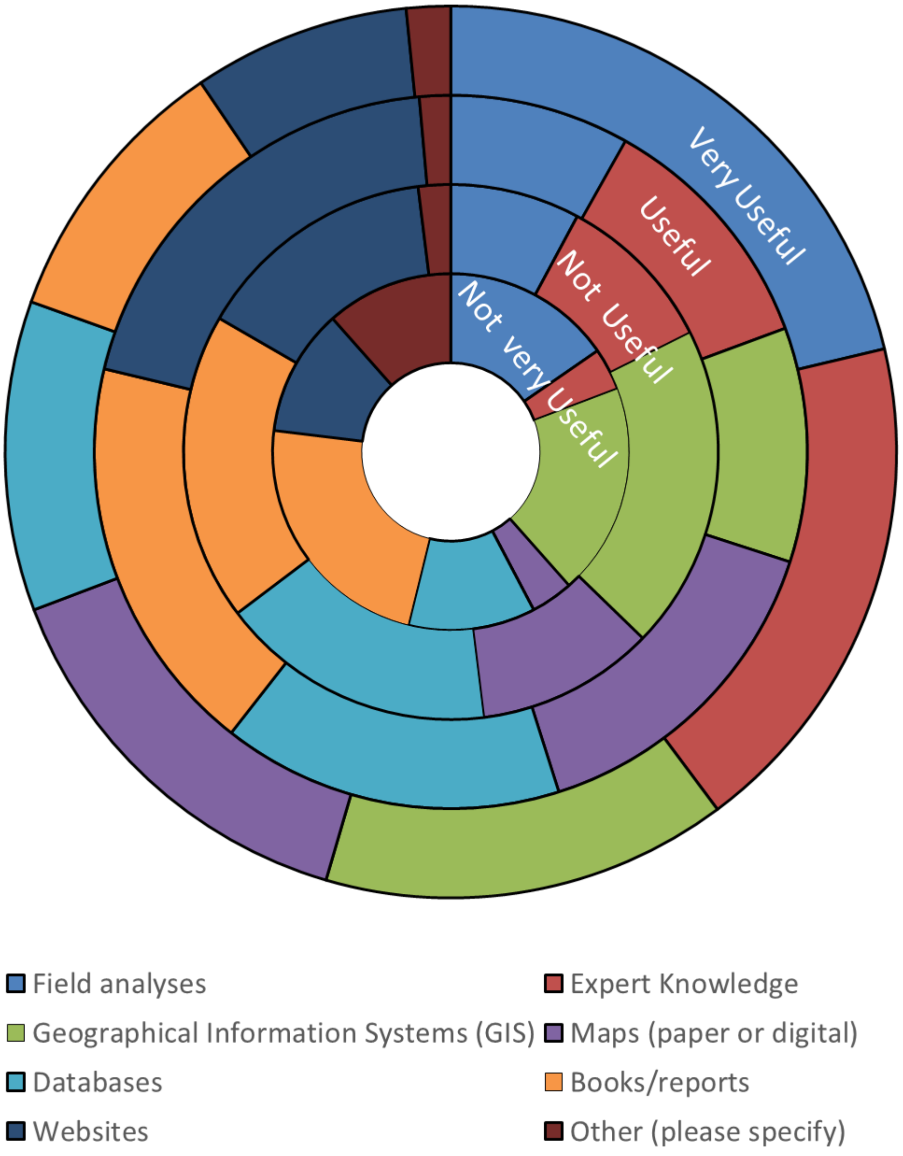


Figure 2.3: How stakeholders rated usefulness of sources. Outer circle represents the percentage of stakeholders who rated *'very useful'*. Inner circle represents those who rated *'not very useful'*.

When asked whether their organisation paid for licenced use of soils information, 49% said that their organisation did, 30% said '*no*' and 21% said that they '*didn't know*'.

Respondents were asked to assess the importance of spatial soils information for wider applications and end-user groups and as a result, an overwhelming 98% of the respondents said that this was '*very important'* or '*important*'. Previously, it was noted that 93% handled information about soil as part of their work. This extra 5% illustrates that those respondents who do not use or acknowledge soil as part of their work still see the importance of spatial soil information for wider applications and end-user groups.

2.3.4. Requested improvements to soil information and data

Improvements to soil data and information were a key issue addressed in this questionnaire. Respondents were asked what they would like to see improved in relation to the information they already use, and this has been summarised in Figure 2.4. Improvements were grouped post-survey to ease interpretation under four main themes: '*Uncertainty', 'Scale and Coverage', 'Metadata'* and *'Fundamental Data'.* 59% of stakeholders wanted soil information at a much finer resolution to what they currently use (Figure 2.4). Although not explicitly specified, many current national scale soil maps (particularly in GB) are at a resolution which is too coarse for in-field management of soils and to allow integration with other spatial datasets. With regards to '*Uncertainty*', respondents wanted improved accuracy and credibility of data sources. With regards to '*Scale and Coverage'*, as well as wanting information at finer scale resolution, respondents wanted to see improvements in co-ordinates of geographical locations (i.e. data in a format which they can georeference). With respect to '*Metadata*' issues, respondents requested improvements in the availability of associated documentation related to the data. Finally, under the category of '*Fundamental Data',* respondents wished to see improvements with trends over time and contemporary data. Respondents were then asked specifically if they would be interested in using any new information that might arise from improvements in spatial resolution/scale and uncertainty. From Table 2.1, it can be noted that there is a positive response to improvements regarding both issues. Other notable requirements ranged from improving map and data interpretations, and the ability to use multiple datasets or assessments.

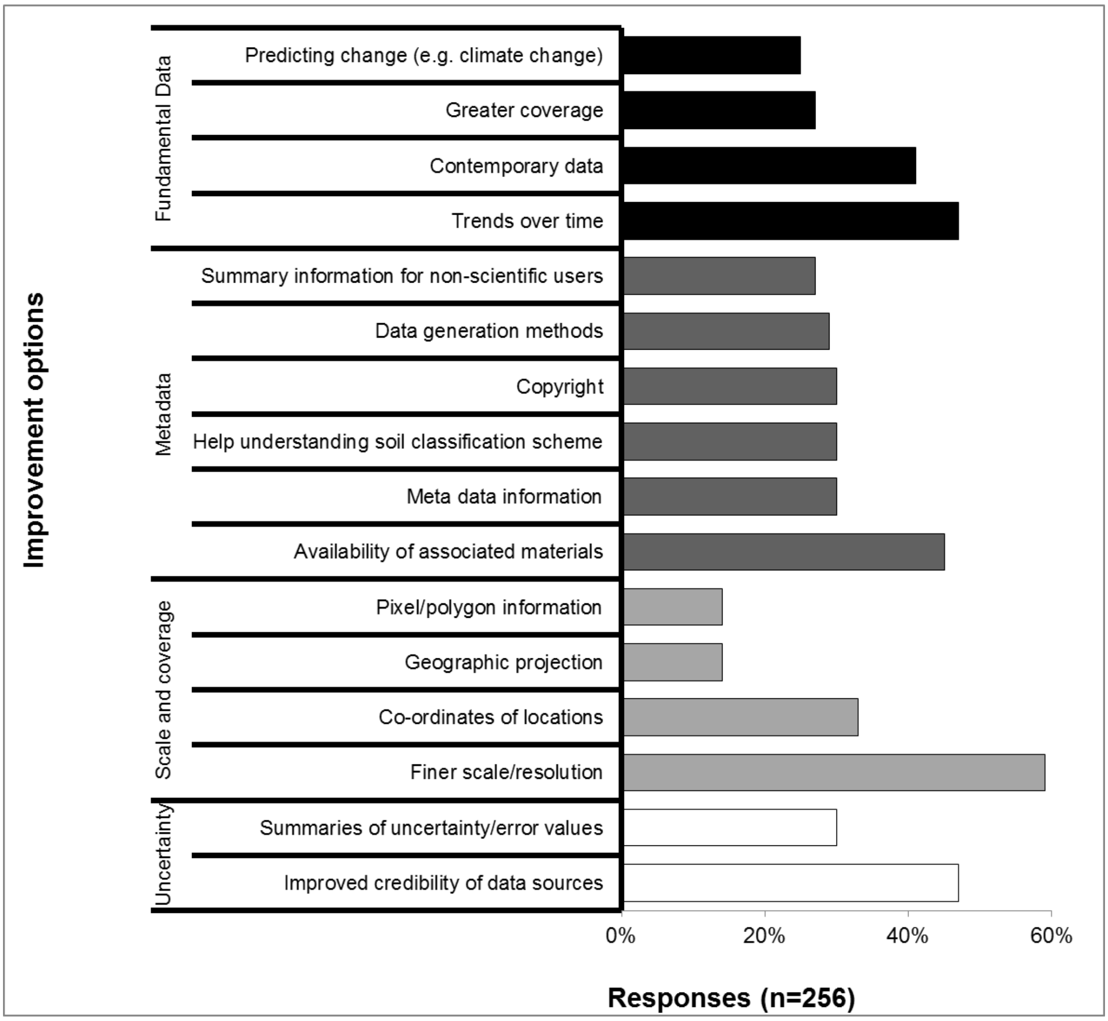


Figure 2.4: Improvement recommendations by the stakeholders

Issue	Yes	No	Total responses
Spatial resolution/scale	209	36	245
Summary of uncertainty/error values	159	57	216
Other (please specify)	10	7	17

Table 2.1: Would you be interested in any new information arising from an improvement in spatial resolution/scale or summary of uncertainty/error values.

There was a space at the end of the questionnaire for respondents to add any extra information that might be useful. The main themes that came out from the additional responses were opportunities to increase knowledge transfer between research and policy makers and the importance of education and training, which are vital in terms of increasing soil understanding.

2.3.5. Relationships between organisations and desired improvements

One of the main objectives of this study was to establish from the questionnaire what desired improvements were linked to the activities of stakeholders. To achieve this, responses were cross tabulated between activities of the organisations and the desired improvements the stakeholders had requested. This was undertaken using the Qualtrics software. The cross tabulations were then used to create heat maps using R Statistics software version 1.1 (R Core Team, 2013: https://www.r-bloggers.com/citing-r-or-sas/) (Figure 2.5). The legend indicates how the shading relates to the number of people who answered responses to both questions i.e. the darker the colour then the greater the correspondence between activities within that specific organisation and the requested improvements. From this it can be noted that, improvements in finer/scale resolution are being requested most by stakeholders whose activities revolve around agriculture or research but consistently needed across all organisational activity groups. Trends over time are also particularly related to those working in agriculture and research but also sought by stakeholders in conservation and national/federal or governmental agencies.

Using the same data, the crosstabs were converted into percentages to explore needs within activity groups. For most organisations (Figure 2.6), finer scale resolution and, associated, improved data accuracy predominated individual organisational user needs. Some organisations identified quite specific needs. In the finance/insurance category, these include improvements in contemporary data, finer scale resolution, improved coverage and methodology in how the data was generated. In the water sector, understanding soil classification and non-(expert) user summaries were identified as relatively high needs.

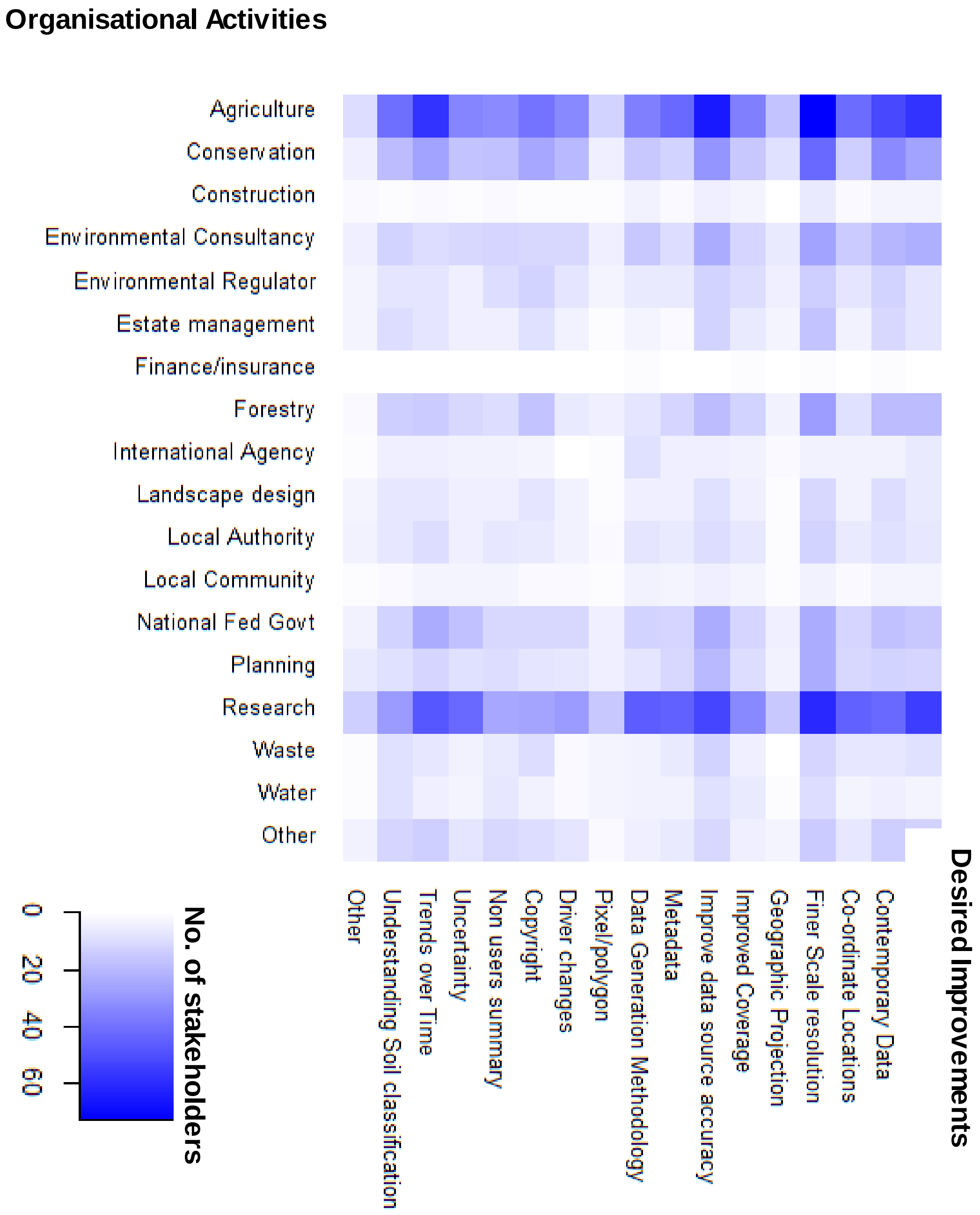


Figure 2.5: Heat map showing the cross tabulation of responses (n) between the activities of an organisation and suggested improvements in soil related information. The darker the colour indicates a greater number of responses.

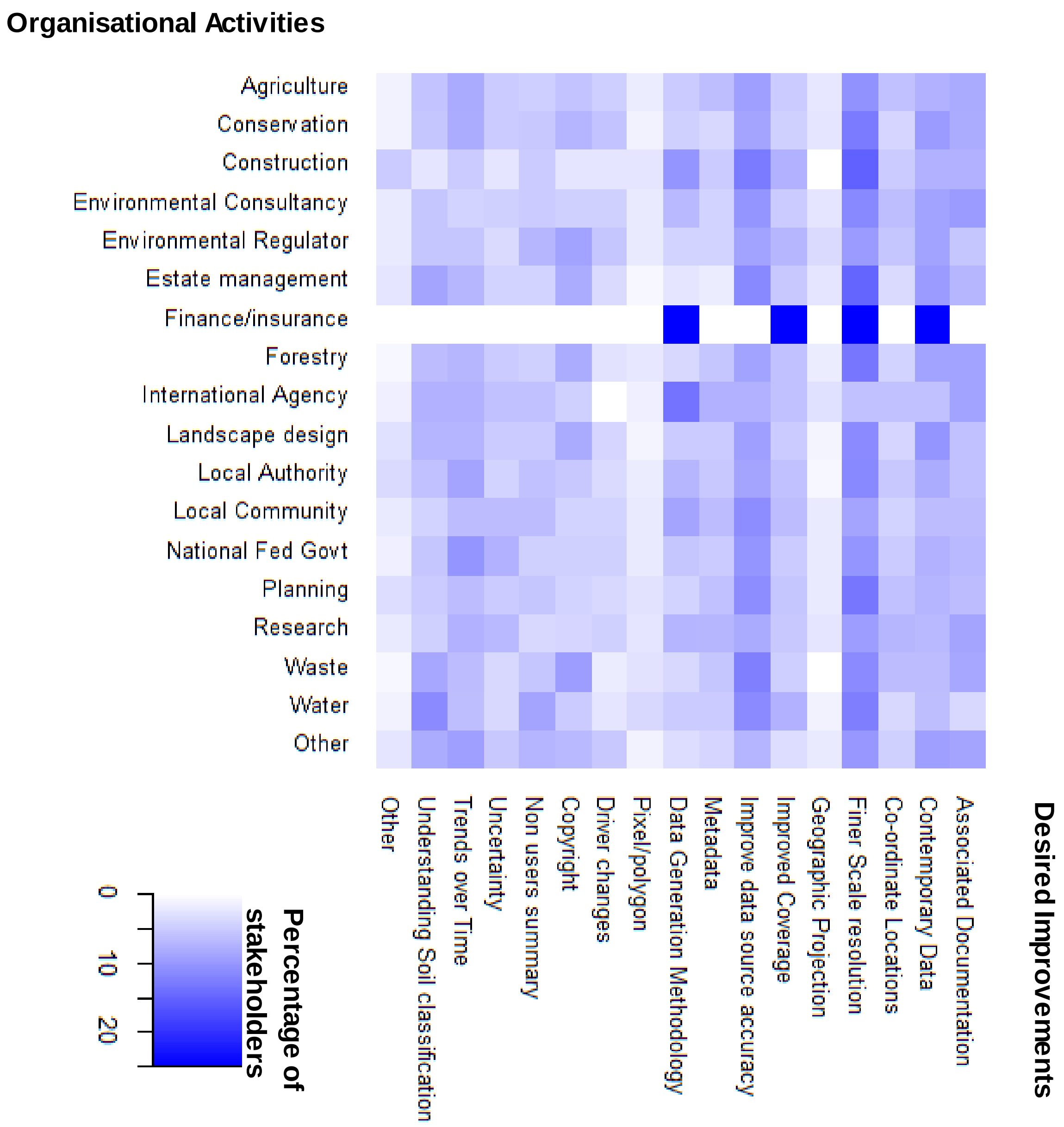


Figure 2.6: Heat map showing the cross tabulation of responses (%) between the activities of an organisation and suggested improvements in soil-related information. The darker the colour indicates a greater percentage of responses by each organisation.

2.4. Discussion

It is encouraging that many responses from this questionnaire were obtained from non-expert stakeholders across substantially different organisations. Many diverse sectors are using, wish to use or access soils information on a regular basis to support day-to-day work practices. Moreover, this survey demonstrates that soils data and information are widely used in a range of tools and assessments and are often integrated with other data sources such as historical data on climate and vegetation (e.g. where soil climate zones were used to establish nutrient demand for crops and grassland for regional animal manure management).

The survey responses also identified that there are barriers to accessing and using appropriate soil data. Overall, stakeholders find difficulties obtaining and collecting information for projects which are under licence or where they must pay for the use of it. Payment for use of data is particularly dependent on organisations procurement procedures and that different organisations are willing to pay varying amounts to obtain certain data for their work or projects (Montanarella and Vargas, 2012; Diafas et al, 2013). It is unclear how much these constitute a significant barrier to the use of soil information, as payment was not identified as one of the required key improvements from the questionnaire. However, improving accessibility would clearly benefit non-experts. Alongside this, there is a clear need to address technical understanding with needs identified for knowledge transfer between research and policy, education and training, improving associated supporting information, understanding soil classifications and non-expert user information. A need for more technical knowledge may well reflect a lack of soils in school and university level education. The level of responses suggests that there is demand (and opportunity) for soils training opportunities focussed on non-experts and practical applications. In parallel, there is also a clear need for increased skill capacity in GIS within organisations using spatial soil data and information. Without this, it is difficult to see how new spatial soil products, which are predominately GIS in nature, can be widely adopted for practical use.

Stakeholders used a variety of information sources and of these, it was notable that a high proportion of people found GIS to be the least useful source of information even though a high proportion of stakeholders use or want to use spatial information and that GIS is a widely used spatial information platform. This may be due to constraints around technical ability, accessibility to GIS software (although open-source GIS software is available e.g. QGIS) or could allude to a more fundamental problem with the GIS medium being inadequate for the assessments undertaken by the respondents.

Other sources of information that were mentioned ranged from the use of social media sites like Twitter, academic journal articles and discussions with other stakeholders. Although not used widely at present, social media does now present real and widespread opportunities to communicate with and inform non-experts. Interestingly, most people found field and laboratory analyses to be '*very useful'* or '*useful*' alongside maps, whether in paper or digital format, and expert knowledge. Reasons could be that stakeholders are utilising 'tacit knowledge' from field experts who acquired this information in the first place, thus using it as a validation tool (Hudson, 1992) and they are familiar enough with handling field and lab results. This may also reflect issues discussed about constraints with technical understanding and GIS skills limiting use of other soil data and information sources.

The questionnaire also indicated widespread requirements for information on future scenarios and trends over time. There is a significant amount of legacy soils data available but much of this is over 30 years old. There is however an underlying requirement for new information on soils to be able to determine current trends in soil properties and functions and to support modelling of future scenarios based on current conditions. Legacy data, on its own, cannot meet current user needs.

This survey indicates that several soil properties including texture (sand, clay and silt), contaminants, bulk density, pH and carbon have widespread use. These should be a priority in making more accessible and useable by addressing the needs for non-expert supporting materials, finer spatial resolution, trends over time etc. However, there are also other soil properties to be considered. Many of the answers in the questionnaire reflect instances where soil properties underpin soil functional assessments and tools. In such instances, the relevance of individual soil properties is '*hidden*' to the user and therefore the need for information on individual soil properties may not be fully expressed. This is a potential pit-fall to be recognised in any future assessments of stakeholder needs. Table 2.2 illustrates the soil properties used to derive these assessments using information gathered from previous documentation and literature (e.g. GlobalSoilMap, 2011a, GlobalSoilMap, 2011b, Mayr et al, 2006). This can be used in post-hoc identification of '*hidden*' soil properties in questionnaires, when exploring needs for soil functional assessments and in ensuring that all necessary soil properties are being considered in the improvement of existing mapping or development of new modelling and mapping, such as DSM and DSA (c.f. Mayr et al, (2006). Expressing the links between soil properties and soil functions can also be used as a tool in raising stakeholders' awareness of the wider range of soil properties which underpin the soil functional assessments and tools that they use regularly.

Related Soil Function	**Assessments**	**Organic carbon**	**pH**	**Clay**	**Silt**	**Sand**	**Coarse Fragments**	**ECEC**	**Bulk density of whole soil**	**AWC**	**Bulk Density of fine earth**
Biomass Production	Agricultural land evaluation			✓	✓	✓	✓		✓	✓	✓
	Biofuel potential		✓							✓	
	Crop Suitability models								✓	✓	✓
	Drainage systems		✓							✓	
	Fertiliser and pesticide usage	✓	✓	✓	✓	✓	✓	✓		✓	
	Irrigation requirements		✓							✓	
	Land Suitability for Forestry		✓	✓	✓	✓	✓				
	(Micro) nutrient concentration		✓								
	Soil borne diseases and/or pests										
	Soil pathogens										
	Drought risk assessments		✓							✓	
Environmental Regulation	Climate change models	✓	✓						✓	✓	✓
	Erosion risk assessments	✓		✓	✓	✓	✓			✓	
	Flood risk maps		✓	✓	✓	✓	✓			✓	
	Hydrology of Soil			✓	✓	✓	✓			✓	
	Leaching risk maps		✓					✓			
	Nutrient Vulnerable Zones		✓							✓	
	Pesticide safety assessment		✓								
	Pollutants in soil		✓					✓		✓	
	Reclamation of contaminated land		✓								
	Runoff potential		✓					✓		✓	
	Sludge acceptance potential		✓					✓			
	Soil erosion		✓	✓	✓	✓	✓				
	Diffuse pollution to waters									✓	
Fundamental Soil Properties	Nutrient cycling		✓					✓			
	Soil acidity/alkalinity levels		✓					✓			
	Soil carbon/organic carbon	✓							✓		✓
	Soil chemistry		✓					✓			
	Soil moisture	✓						✓	✓	✓	✓
	Soil temperature							✓			
Habitats and Biodiversity	Habitat suitability maps									✓	
	Land reclamation			✓	✓	✓	✓				
	Land use change modelling										
	Pollen counts										
	Protection of animal species										
	Recreational space										
Infrastructure	Extraction of raw materials	✓		✓	✓	✓	✓			✓	
	Infrastructure assessment			✓	✓	✓	✓				
	Land Suitability for Housing			✓	✓	✓	✓				

Table 2.2: Soil assessments against probable soil properties mapped as future work. Table adapted from: GlobalSoilMap (2011a, 2011b) and Mayr et al, (2006).

Most stakeholders stated, from the questionnaire, that they require information at finer spatial scale/ resolution than is currently being offered. An obvious focus for future work is to deliver finer spatial scale in the key soil properties identified by the stakeholders (i.e. bulk density, soil contaminants, pH, texture and carbon). However, one assumption is that finer spatial scale will lead to improved data and subsequent assessments. This may not be the case since scale is a complex parameter which is dependent on context and application (Goodchild, 1997, Wu and Li, 2009). Supported and promoted by FAO (FAO, 2018: http://www.fao.org/global-soil-partnership/pillars-action/4-information-and-data/en/), DSM is a major opportunity to generate soil property information at finer spatial scale than existing products, with the benefit of characterising properties of accuracy and precision (Goodchild and Quattrochi, 2013). Such predicted soil property products can then be used to make significant advances in modelling and mapping the soil functional assessments which are widely used by diverse stakeholders and organisations. However, it is imperative that such approaches are matched with field assessments to critically evaluate and validate the accuracy of predicting soil properties at finer spatial resolution using existing (generally legacy) data.

2.5. Conclusions

The questionnaire was designed to understand how soils data and information are being used by non-expert stakeholders for a range of purposes. The responses indicate that stakeholders are generally aware of the utility of soil data and soil functional assessments for their work however they may not be aware of the full range of soil properties underlying soil functional assessments. Stakeholders identified that better and wider use of existing (and future) soil information by non-experts could be enabled by improvements in data access and user-friendly supporting materials. Many stakeholders require finer spatial resolution than is currently offered, contemporary information on soils and trends over time for soil functions as well as properties. Established soil modelling such as the global initiatives in DSM and DSA can address some of these needs. However, a clear message from stakeholders is that existing legacy soils data needs to be supplemented by new up-to-date soil data which is fit for current and future uses. Requirements for contemporary data demand investments in new and novel monitoring and sampling at enough spatial resolution and frequency to enable assessments of the range of soil functions. These will, in turn, be used to deliver and shape a wide range of multi-organisational activities and policies. A question remains as to how long we can rely on legacy soil data to make decisions today and into the future?

References

Auricht, C., 2004: Natural Resources Atlas and Data Library - User Review. National Land and Water Resources Audit: Accessed from http://lwa.gov.au/products/er040794 [Last Accessed 20th July 2017].

Behrens, T., Scholten, T., 2006: Digital soil mapping in Germany—a review. *Journal of Plant Nutrition in Soil Science*, 169**,** pp. 434–443.

Bibby, J.S., Douglas, H.A., Thomasson, A.J., Robertson, J.S., 1982: Land Capability Classification for Agriculture. The Macaulay Land Use Research Institute, Aberdeen. ISBN 0 7084 0508 8.

Black, H., Bruneau, P., Dobbie, K., 2012: Soil Monitoring Action Plan. Accessed from (http://www.environment.scotland.gov.uk//media/59999/Soil_Monitoring_Action_Plan.PDF) [Last accessed 13th December 2016].

Blum, W.E.H., 2005: Functions of soil for society and the environment. *Reviews in Environmental Sci and Bio/Technology*, 4, pp. 75-79.

Bouma, J., Broll, G., Crane, T. A., Dewitte, O., Gardi, C., Schulte, R., and Towers, W., 2012: Soil information in support of policy making and awareness raising. *Current Opinion in Environmental Sustainability*, 4, 5, pp. 552-558.

Britz, W., Witzke, W., 2014: CAPRI model documentation. [Accessed from http://www.capri-model.org/docs/capri_documentation.pdf [Last accessed 21st July 2017].

Carré, F., McBratney, A.B., Mayr, T., Montanarella, L., 2007: Digital soil assessments: Beyond DSM. *Geoderma*, 142,1–2, pp. 69-79.

Costanza, R., d'Arge, R., de Groot, R.S., Farber, S., Grasso, M., Hannon, B., Limburg, K., Naeem, S., O'Neill, R.V., Paruelo, J., Raskin, R.G., Sutton, P., van den Belt, M., 1997: The value of world's ecosystem services and natural capital. *Nature*, 387 pp. 253–260.

Diafas, I., Panagos, P., Montanarella, L., 2013: Willingness to Pay for Soil Information Derived by Digital Maps: A Choice Experiment Approach. *Vadose Zone Journal*, 12, 4, DOI: 10.2136/vzj2012.0198.

Environment Agency., 2013: Groundwater protection: Principles and Practice (GP3) (August 2013 Version 1.1.) [Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/297347/LIT_7660_9a3742.pdf] [Last accessed 26th September 2015].

FAO., 2018: Pillar 4: Enhance the quantity and quality of soil data and information: data collection (generation), analysis, validation, reporting, monitoring and integration with other disciplines. [Available at: http://www.fao.org/global-soil-partnership/pillars-action/4-information-and-data/en/] [Last Accessed 18th September 2018].

FAO., 2015: Soil Functions. [Available at: http://www.fao.org/resources/infographics/infographics-details/en/c/284478/] [Last accessed 18th September 2018].

FAO., 2012: Towards Global Soil Information: Activities within the Geo Task Global Soil Data: Workshop report. Accessed from http://www.fao.org/fileadmin/templates/GSP/downloads/GSP_SoilInformation_WorkshopReport.pdf [Last accessed 10th September 2015].

GlobalSoilMap., 2011a: GlobalSoilMap.net: New Digital Soil Map of the world. [Accessed from https://www.google.co.uk/search?q=globalsoilmap.net&ie=utf-8&oe=utf-8&aq=t&rls=org.mozilla:en-GB:official&client=firefox-

a&channel=sb&gfe_rd=cr&ei=aR3jVLWpOoSV8wOK4YHwBQ] [Last accessed 17th February 2015].

GlobalSoilMap., 2011b: Specifications Version 1: GlobalSoilMap.net products: Release 2.1. Technical report.

Goodchild, M.F., 1997: Towards a geography of geographic information in a digital world. Computers, *Environment and Urban Systems,* 21, pp. 377-391.

Goodchild, M.F. and Quattrochi, D.A., 1997: Scale in Remote Sensing and GIS. Lewis Publishers, pp. 1-11.

Grealish, G.J., Fitzpatrick, R.W., Hutson, J.L., 2015: Soil survey data rescued by means of user-friendly soil identification keys and toposequence models to deliver soil information for improved land management. *GeoRes J*, 6, pp. 81-91.

Haines-Young, R., Potschin, M., Kienast, F., 2012: Indicators of ecosystem service potential at European scales: Mapping marginal changes and trade-offs, *Ecological Indicators*, 21, pp. 39-53.

Harter, T., Walker, L.G., 2001: Assessing the vulnerability of groundwater. [Accessed from www.dhs.ca.gov/ps/ddwem/dwsap/DWSAPindex.htm] [Last Accessed 20th July 2017].

GS Soil., 2010: Assessment and strategic development of INSPIRE compliant Geodata-Services for European Soil Data: D-2.3: Final report on consolidating soil related inventory and Theme Catalogue for soil data providers. *e*Content*plus*.

Hudson, H.D., 1992: Division S-5 – Soil Genesis, Morphology and Classification: The Soil Survey as Paradigm-based Science, *Soil Science Society of America Journal*, 56, pp. 836-841.

Jones, R.J.A., Houskova, B., Bullock, P., Montanarella, L., 2005: Soil Resources of Europe, second edition. European Soil Bureau Research Report No.9, EUR 20559 EN, (2005), 420pp. Office for Official Publications of the European Communities, Luxembourg.

Mather, A.S., 1988: New private forests in Scotland: characteristics and contrasts. *Area*, 20, 2, pp.135-143.

Mayr, T., Black, H., Towers, W., Palmer, R., Cooke, H., Freeman, M., Hornung, M., Wood, C., Wright, S., Lilly, A., DeGroote, J., Jones, M., 2006: Novel methods for spatial prediction of soil functions within landscapes (SP0531). DEFRA, 26pp.

McBratney, A. B., Mendonça Santos, M.L., Minasny, B., 2003: On digital soil mapping. *Geoderma*, 117,1, pp. 3-52.

Montanarella, L., Vargas, R., 2012: Global governance of soil resources as a necessary condition for sustainable development. *Current Opinion in Environmental Sustainability*, 4, pp. 1-6.

Omuto, C., Nachtergaele, F., Rojas, R.V., 2013: State of the art report on Global and Regional Soil Information: Where are we? Where to go. Global Soil Partnership Technical Report.

Panagos, P., Van Liedekerke, M., Jones, A., Montanarella, L., 2012: European Soil Data Centre: response to European policy support and public data requirements. *Land Use Policy*, 29, 2, pp. 329-338.

Prager, K. and McKee, A., 2014: Use and awareness of soil data and information among local authorities, farmers and estate managers. The James Hutton Institute Internal Report.

Pritchard O.G., Hallett, S.H., Farewell, T.S., 2015: Probabilistic soil moisture projections to assess Great Britain's future clay-related subsidence hazard. *Climatic Change*, 133, 4, pp. 635-650.

Qualtrics., 2018: Qualtrics. [Available at https://www.qualtrics.com/uk/] [Last accessed 18th September 2018].

R Core Team., 2013: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at [http://www.R-project.org] [Last accessed 18th September 2018].

Reed, M., 2008: Stakeholder participation for environmental management: A Literature Review. *Biological Conservation*, 141, 10, pp. 2417-2431.

Smith, P., Cotrufo, M.F., Rumpel, C., Paustian, K., Kuikman, P.J., Elliott, J.A., McDowell, R., Griffiths, R.I., Asakawa, S., Bustamante, M., House, J.I., Sobocká, J., Harper, R., Pan, G., West, P.C., Gerber, J.S., Clark, J.M., Adhya, T., Scholes, R.J., Scholes, M.C., 2015: Biogeochemical cycles and biodiversity as key drivers of ecosystem services provided by soils. *SOIL D*, 2, pp. 537-586.

Valentine, K.W.G., Naughton, W.C., Navai, M., 1981: A questionnaire to users of soil maps in British Columbia: Results and implications for design and content. *Canadian Journal of Soil Science*, 61, pp. 123-135.

Wood, B. and Auricht, C., 2011: ASRIS / ACLEP User Needs Analysis. [Accessed from http://www.clw.csiro.au/aclep/documents/ASRIS_User_Analysis.pdf] [Last Accessed 20th July 2017].

Wu, H., Li, Z.L., 2009: Scale Issues in Remote Sensing: A Review on Analysis, Processing and Modelling. *Sensors*, 9, 3, pp.1768-1793.

3 DSM PREPARATION AND DATA COLLECTION

3.1. A need for harmonised soil property datasets for GB

There were two national soil survey organisations collecting and mapping soil property data in Great Britain (GB); the Soil Survey of Scotland and the Soil Survey of England and Wales. With a move towards greater access to soils data (e.g. Global Soil Partnership, 2018; Omuto et al, 2013; Campbell et al, 2017; Hengl et al, 2017; Lawley et al, 2014) and increasing requests for cross-border datasets (Lilly, pers com, 2018), there is a clear need to provide unified soil property datasets. These unified datasets will be important for stakeholders and users especially in terms of dealing with key environmental challenges such as water resource management and the contribution of soil organic carbon data to the Global Soil Partnership (Campbell et al, 2017).

There are several challenges faced when attempting to harmonise the two GB soil datasets. Although the national soil maps have been produced at a common scale (1:250,000), there are some differences in how the map units were constructed and named. There are also differences in soil taxonomy between Scotland and England and Wales. Although both have a dataset containing ‘representative profiles’ collected with the intention to characterise the soil mapping and taxonomic units and datasets which have objective, grid samples collected to provide a statistical assessment of soil characteristics (National Soil Inventories), the sampling frame and grid size vary slightly between the two datasets along with the overall sampling depth. However, there is sufficient commonality within both to form a GB dataset. Finally, the laboratory and analytical methods to measure the various soil properties differ between the two datasets reflecting the characteristics of the soils and the specific soil property data that was required. The following sections describe the key datasets used in this study, considering the steps achieved to produce a harmonised dataset and the resulting GB soil property dataset.

3.2. Soil datasets

3.2.1. England and Wales

The soil property data for England and Wales were derived from the National Soils Inventory (NSI) and the Land Information System (LandIS) datasets (Hallett et al, 2017) which contains the representative profiles. For the NSI, 5662 sites were sampled across a regular 5 km grid square. Topsoil samples were collected, and soil profiles were described over the whole of England and Wales. The 5 km grid was based on the Ordnance Survey grid intersections but was offset by 1 km north and 1 km east to avoid sampling points falling on edges of published 1:25,000 scale maps. Urban areas and water bodies were not sampled, but all other locations with soil were sampled. At each site, 25 soil auger samples were taken at 4m intervals over a 20 x 20 m square centred on the grid intersection. The cores were then bulked to give a total of approximately 1 kg of moist soil for each site. The first round of sampling took place between 1978 and 1983. The LandIS representative soil profiles ('soil pits') were sampled to characterise the soil series mapped in England and Wales and sampling took place from 1935 and finished around 2000. There are over 11,000 soil pit site locations found in LandIS, comprising of over 47,000 horizons. Only associated data connected to the soil pit site locations was used. Georeferenced soil profile pits were excavated to 1.2 m depth where possible, described and samples taken from each soil horizon identified.

3.2.2. Scotland

The soil property data for Scotland also includes a National Soil Inventory and representative profiles as well as samples taken to characterise field scale experiments. The first National Soil Inventory of Scotland (NSIS1) took place between 1978 and 1988 (Lilly et al, 2010) and was based on a 5 km grid over the whole of Scotland (excluding the Orkney Islands). Soil profile descriptions and site information were collected at each Ordnance Survey 5 km grid intersection and samples were collected from each soil horizon (to a depth of around 80 cm where possible, and, at times, to 100 cm) at all 10 km intersects with soil and at some 5 km intersections. The soil at 721 10 km grid intersects were sampled, with 66 locations having no soil (the points were found to be in lochs or rivers, built up areas or on solid rock).

Between 2007 and 2009, there was a partial resampling of original NSIS (NSIS 2007-9) (Lilly et al, 2011). This was to investigate possible changes in soil properties, to compare sampling methods and to measure and assess new soil attributes as soil quality indicators. A total of 183 soil profiles were relocated on the 20 km grid, sampled and described.

The representative soil profiles of Scotland, collected to characterise soil map and taxonomic units, have data collected from the 1930s to present. These georeferenced soil profiles were excavated to around 1m (though some are deeper) and each major soil horizon was sampled. There are nearly 15,000 soil profiles in the Scottish Soils Database, over 7,000 of which are not part of the Inventory or from closely spaced grid and transect surveys which have accompanying analytical data. These 7,000 soil profiles are the main training dataset for Scotland.

3.3. Soil properties

The GlobalSoilMap.net (GSM) project was established with the aim of generating a new digital soil map of the world using state-of-the-art technologies for soil mapping and predicting soil properties (Sanchez et al, 2009). A standard approach to predict soil properties at depth for GSM was established. Currently, as part of the GSM criteria, twelve soil properties are predicted at each location (GlobalSoilMap, 2011a; GlobalSoilMap, 2011b). Of the twelve properties, five were selected to test the application of DSM techniques across GB. These are: loss on ignition (LOI), pH in water and soil texture (sand, clay and silt). These all hold agronomic significance and were also amongst the key properties identified by the stakeholders in the questionnaire survey in Chapter 2 (Campbell et al, 2017).

3.3.1. Organic Carbon/ Loss on Ignition (LOI)

Carbon was one of the main properties that the questionnaire survey showed was important for stakeholders (Campbell et al, 2017). There are different methods used to determine carbon documented in the literature such as Walkley Black (Defra, 2011), CHN analysers (Chapman et al, 2013) or by using Loss on Ignition.

Findings from Chapman et al, (2013) and Lilly et al, (2015) showed a difference in carbon measurements when the original NSIS samples were using a CHN elemental analyser. They found that the original values were 11.5% greater than the reanalysed archive sample. No such change was noted when the samples were reanalysed for LOI suggesting that the change was due to the analysers returning different values. LOI is also sensitive to different analytical methods, most importantly, the ignition temperature.

A range of ignition temperatures and durations are commonly used to determine LOI (Hoogsteen et al, 2015; Ball, 1964; Abella and Zimmer, 2007; Konen et al, 2002). For soils in England and Wales, temperatures used were at 850°C (Avery and Bascomb, 1982) and for Scotland, 900°C (Macaulay Institute for Soil Research, 1971). It has been noted by Ball (1964) that ignition temperatures at 375°C give different LOI values than those from ignition at 850°C, where there is the likelihood of structural water loss. However, there is also a significant risk of incomplete combustion at 375°C (Hoogsteen et al, 2015). Though, the lower temperatures are appropriate for soils with carbonates present.

The LOI values were found to be comparable between Scottish and England and Wales datasets. Thus, for the purposes of this study, LOI data from the representative profiles for England and Wales which were determined by combustion at 850°C (Hallett et al, 2017; Avery and Bascomb, 1982) and from the representative profiles for Scotland which were determined by combustion at 900°C (Macaulay Institute for Soil Research, 1971) were used.

3.3.2. pH (H_2O)

The questionnaire survey showed that soil pH was the second most important soil property for stakeholders (Campbell et al, 2017). Soil pH is a common measurement performed in soil chemical analyses (Davies et al, 1971), however, there are many different methods used throughout the world (Minasny and McBratney, 2011; Miller and Kissel, 2010; Kissel et al, 2009). These techniques and values will vary depending on the type of solution used and the ratio of soil to solution used (Minasny and McBratney, 2011). Two of the most common measures of soil pH used by the two GB Survey organisations are in a solution of soil and water and in a solution of soil, water and calcium chloride ($CaCl_2$).

In Scotland, the pH of water is measured on an air-dried soil to water ratio of 1:3 (Macaulay Institute for Soil Research, 1971) and for measuring the pH of $CaCl_2$, 0.01M $CaCl_2$ is added to the suspension. In England and Wales, measurements are made on a soil to water ratio of 1:2.5 and for measuring the pH of $CaCl_2$, 0.01M $CaCl_2$ is added to the suspension (Avery and Bascomb, 1979). More detailed information on the methods used is found in Appendix 2).

From Figure 3.1, pH data from Scotland and England and Wales can be seen to overlap with one another thereby suggesting that there is no systematic bias for soil pH despite the slightly different methods used across both datasets. As pH in water is part of the GlobalSoilMap.net criteria, this was another reason for including it in the GB modelling.

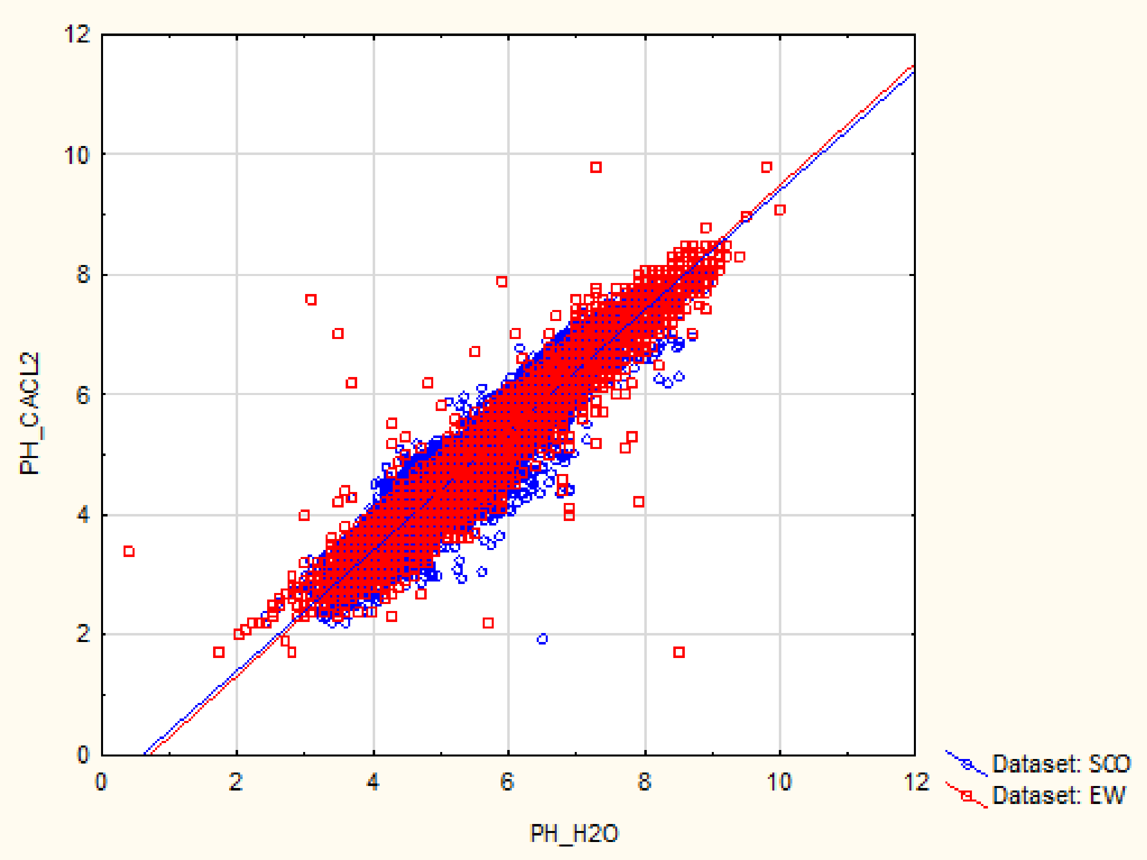


Figure 3.1: Overlaid relationships between pH in water and calcium chloride. The red markers indicate values from England and Wales (EW) dataset, and the blue markers represent values from the Scottish data (SCO).

3.3.3. Particle Size Distribution (PSD)

For England and Wales, Particle Size Distribution (PSD) was derived using the pipette method which involves the disaggregation of soil particles with hydrogen peroxide (Avery and Bascomb, 1979). This determination, however, has not been made for samples with organic carbon content greater than 15%, as these were classified as organic soils. In contrast, PSD in Scotland was measured using the hydrometer method. However, different particle size classes have been used in Scotland in comparison to England and Wales. In England and Wales, the particle size classes were predominantly based on the British Soil Texture Classification (BSTC) (Avery, 1973) although some data were in the USDA particle size classes. In Scotland, particle size classes mainly followed the United States Department of Agriculture (USDA) size classes (USDA, 1978). The primary difference between the BSTC and the USDA classes is the cut off used between the silt and sand fractions. The USDA particle size classes are ((<2 (clay), 2-50 (silt), 50-2000 (sand) in µm)) and for BSTC they are ((<2 (clay), 2-60 (silt), 60-2000 (sand) µm).

Initially, two different approaches were investigated for harmonisation of the particle size data. The first used a curve fitting approach (Nemes et al, (1999) and the second method involved developing regression equations (Minasny and McBratney, 2001). Further information on the comparison of harmonisation methods can be found in Appendix 2 but they are briefly outlined below.

The curve fitting approach (Nemes et al,1999) used data from the National Soil Inventory of Scotland 2007-9 (NSIS 2007-9) dataset as the proportion of particles in both USDA and BSTC particle size classes had been measured. A test data set from 300 mineral soil horizons was extracted from the main dataset, cumulative proportions were calculated, and the particle size classes converted to a log scale. The data were then plotted as particle size class against cumulative percentage. The hope was that a single curve could be used to estimate the proportion of particles in the 2-60 and 60-2000 µm size range, however, the test dataset showed that the curves have a wide range of shapes (Figure 3.2) meaning that each individual particle size dataset would have to be individually modelled making this impractical and time consuming.

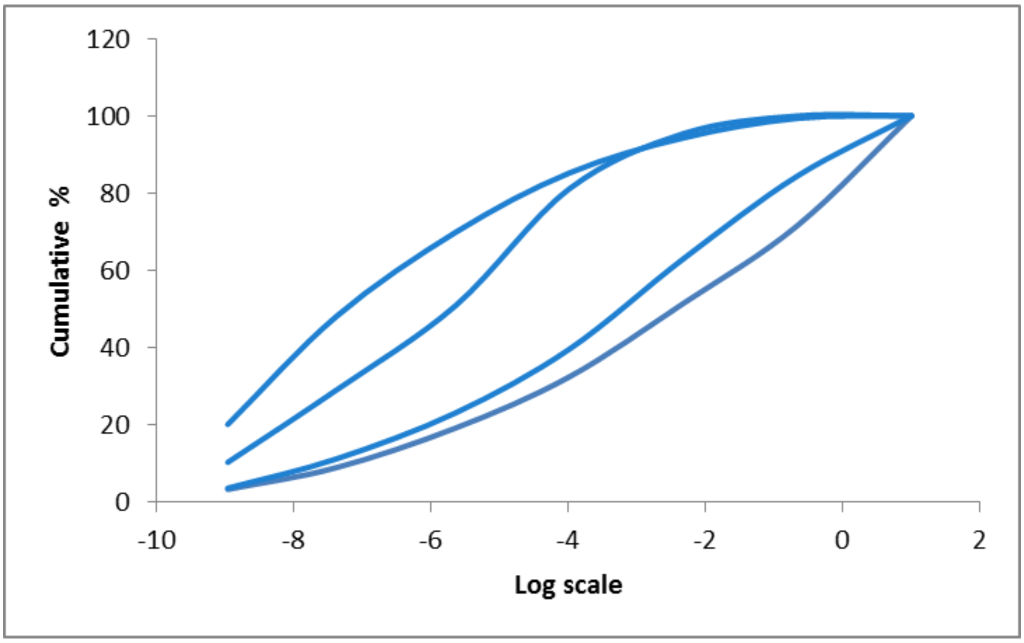


Figure 3.2: Example curve shape fits for example particle size classes

A second approach based on regression equations to predict the proportion of particles in the 2-60 and 60-2000 µm size range. The Scottish NSIS dataset (2007-9) contains particle size measurements in the following categories: <2, 2-6, 6-20,20-200, 200-2000, 20-2000, 2-20, 2-50, 2-60, 20-60,60-200, 200-600, 600-2000, 60-2000, 50-100, 100-250, 250-500, 500-1000, 125-200, 50-2000,1000-2000 µm, that is, both USDA and BSTC. These data were used to explore regression methods to predict the proportion of particles in the 2-60 and 60-2000 µm size range for Scotland and England and Wales.

The most successful regression equation used the proportion of particles in the 20-50 and 50-2000 µm (Table 3.1, R^2= 0.99). When the residuals were plotted, the majority were between +1 and -1 meaning that there is a very strong linear relationship (Figure 3.3)

n= 680	b	Standard error
Intercept	1.82	0.17
CLASS 20 – 50	-0.2	0.01
CLASS 50 – 2000	0.974	0.00
R= 0.99	R^2= 0.99	Adjusted R^2= 0.99

Table 3.1: Best regression analysis found from PSD classes for Scotland

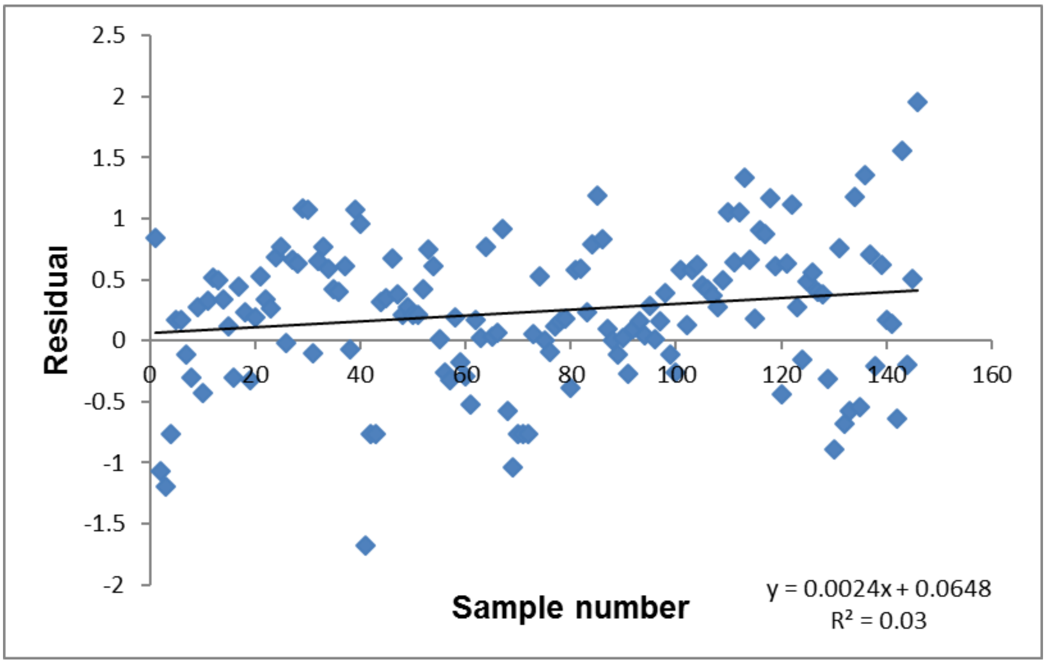


Figure 3.3: Graphed residuals found from PSD classes for Scotland.

When this regression was applied to England and Wales USDA data however, there was a large under prediction of the sand content up to 10%. This was attributed to the narrower range of silt and clay contents found in Scottish soils compared to those in England and Wales. Therefore, it was deemed to be inappropriate to apply the correction factor to the England and Wales USDA dataset.

A comparison between 161 measurements of both USDA and BSTC sand contents from the NSIS Scottish data showed an average of 3.6% difference in sand content which is within quoted experimental error (Dane and Topp, 1976, p283). As a result, BSTC and USDA classifications were deemed to be sufficiently similar across both Scotland and England and Wales with respect to potential errors in DSM modelling. Thus, no corrections were applied to the England and Wales or Scottish USDA particle size class data.

3.4. Covariates used for DSM

3.4.1. Soils (S)

Although national soil maps have been produced at a common scale (1:250,000) covering both Scotland and England and Wales (Soil Survey of Scotland Staff, 1981, Hallett et al, 2017) there are some differences in how the map units were constructed and named as well as differences in soil taxonomy.

The National Soil Map (NATMAP) of England and Wales is a digitised version of published 1:250,000 scale maps (Hallett et al, 2017). This is based on published soil maps produced at various scales and on reconnaissance mapping of unsurveyed areas. The legend is based on the soil associations identified by the most frequently occurring soil series or major soil sub group and by the arrangement of additional environmental information (Hallett et al, 2017). The map units are further distinguished by number codes and by dominant soil sub groups.

For this work, the dominant major soil group was used for DSM development and the soil polygons were preserved as a covariate.

The 1:250,000 National Soil Map of Scotland is a digitised version of a series of 7 paper maps published in 1981 (Soil Survey of Scotland Staff, 1981) derived from a combination of new soil survey work undertaken between 1978 and 1981 and an interpretation of existing detailed mapping completed over a 30-year period prior to 1978. The soil map units can be described as soil 'complexes' and are based on repeating landforms found across Scotland which comprise one or more specific soil types. The landforms are further subdivided based on the geological parent material. For the DSM, information on the spatial distribution of the Major Soil Subgroup and soil association was used as covariates. The Major Soil Group in England and Wales is the equivalent of the Major Soil Subgroup in Scotland. Both datasets were converted to raster format and then resampled to 100 m resolution.

3.4.2. Climate (C)

Global climate data was collected from World Climate, (World Climate, 2016; http://www.worldclim.org/), a global climate data website which provides free climate data for GIS and modelling purposes (Hijmans et al, 2005). Six of the climate datasets are commonly used in DSM and these were selected for use in this study (BioClim, 2016, http://www.worldclim.org/bioclim). The scales for these in their original form are at around 1km^2 grid scales and found in raster form.

The six parameters were:

- BIO1 – Annual Mean Temperature
- BIO2 – Mean Diurnal Range
- BIO3 – Isothermality
- BIO4 – Temperature Seasonality
- BIO12 – Annual Precipitation
- BIO15 – Precipitation Seasonality

3.4.3. Organisms (O)

Within GB, vegetation communities are a dominant soil forming factor. Therefore, because of this, land cover maps were used to represent the organisms soil forming factor. Land cover map (LCM2000) dataset is held by the Centre of Ecology and Hydrology CEH) and is derived from a classification of scenes from satellite image data covering GB. LCM2000 is categorised using classification from the Joint Nature Conservation Committee (JNCC) Broad Habitats, encompassing the full range of habitats across GB and Northern Ireland. LCM2000 was produced in both raster and vector formats at varying levels of detail and spatial resolution. Other datasets that were considered for use were the Land Cover Map (LCM1990) and Land Cover Map (LCM2007) also from the Centre of Ecology and Hydrology (CEH). The Land Cover for Scotland 1988 (LCS88) dataset based at the James Hutton Institute was also considered along with the European CORINE dataset found from the Joint Research Council. However, after investigations, the subclasses from LCM2000 were the solitary dataset used for all DSM work in this PhD because it covers the whole of GB. This dataset was converted to a raster format and then resampled to 100 m resolution.

3.4.4. Relief (R)

Several Digital Terrain Models (DTMs) were tested to ascertain the most effective for use in DSM of GB. To investigate which DTM is best to use for the whole of GB, the 25 m DTM from the Joint Research Centre (JRC) as well as 5 m and 50 m DTMs from the Ordnance Survey (OS) were investigated.

The 5 m DTM was obtained by a triangulated irregular network method. This was achieved by editing the mass points and breaklines or by automated techniques within a photogrammetric environment (Ordnance Survey, 2016a). The accuracy level of the 5 m DTM was found to be at 2 m root mean square error (RMSE). The 25 m DTM was obtained by an Advanced Spaceborne Thermal Emission Reflection Radiometer Global Digital Elevation Model (ASTER GDEM) and Shuttle Radar Topography Mission (SRTM) (Dufourmont, et al, 2014). The accuracy level of the 25 m DTM was found to be at 7 m root mean square error (RMSE). The accuracy level of the 50 m DTM was obtained by a triangulated irregular network method. This was achieved by editing the mass points and breaklines or by automated techniques within a photogrammetric environment (Ordnance Survey, 2016b). The accuracy level of the 50 m DTM was found to be at 4m root mean square error (RMSE).

Some other data was used for relief investigations. The CEH 1:50,000 digital river network dataset of Great Britain was produced as part of a long term project within the Institute of Hydrology between the mid-1970s and the late 1990s (Moore et al, 2000; https://catalogue.ceh.ac.uk/documents/7d5e42b6-7729-46c8-99e9-f9e4efddde1d). This dataset is constructed from source maps along with line work from rivers, lakes and estuaries to construct the river networks.

The high-water line for GB can be accessed from the Boundary-Line file on the Ordnance Survey Open Data source page (Ordnance Survey, 2016c). The external bounding line of the Boundary-Line dataset is digitised by aligning low water springs to the extent of the sea.

Catchment boundaries form part of the Integrated Hydrological Units (IHU) dataset for the UK (Centre for Ecology and Hydrology, 2016) were also utilised. The Hydrometric Areas are the coarsest units of the IHU in terms of spatial resolution. Figure 3.4 illustrates a flow chart for the process for DTM investigations.

A variety of approaches and criteria were undertaken to understand what is contained within the DTMs. These involved calculating the number of sinks or depressions found in the DTMs and conducting difference maps by taking a hydrologically correct DTM (i.e. filling in all the depressions) and subtracting it from the original DTM to investigate how hydrologically correct the DTMs are and to how consistent they are at illustrating characteristics of the relief.

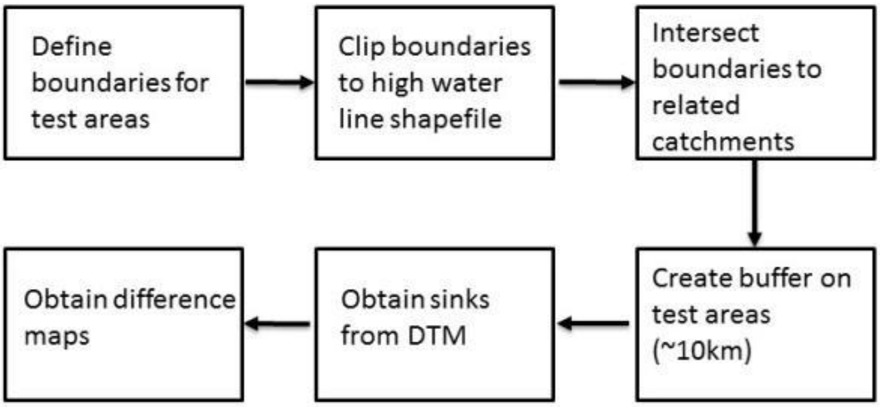

Define boundaries for test areas
Clip boundaries to high water line shapefile
Intersect boundaries to related catchments
Create buffer on test areas (~10km)
Obtain sinks from DTM
Obtain difference maps

Figure 3.4: DTM preparation process

Two sink fill analyses methods: Planchon and Darboux (2001) and Wang and Liu were carried out for GB to investigate which method is more reliable at determining relief parameters and to examine which is better operationally. Planchon and Darboux (2001) use an algorithm which involves filling in depressions with a layer of water and then removing excess water. This method is perceived to be versatile as the depressions are replaced with a horizontal or sloping surface. Compared with other methods, the Planchon and Darboux method represents increased consistency (Planchon and Darboux, 2001) as the algorithm is designed for analyzing the storage capacity of the soil and therefore, no attempt is made to determine the flow directions in the depressions. The other approach utilised is by Wang and Liu (2006) in which they utilise an algorithm to classify and fill surface depressions in DTMs. The original method has been improved to allow the creation of hydrologically rigorous elevation models. This means that it preserves a slope along the flow path of the DTM as well as filling in the depressions across the landscape. This is achieved by creating a minimum elevation difference between grid cells. Wang and Liu (2006) method is based on a novel concept of spill elevation and the least-cost search technique for each grid cell. This method is useful in identifying surface depressions, assigning flow directions and delineates watershed boundaries with one batch of processing.

The findings from the sink analyses along with the number of sinks found for each DTM for GB are found in Table 3.3. What can be highlighted from this is that the 50 m DTM produced the fewest number of sinks and the 5 m DTM obtained the most. It can also be noted that the DTM which produced the least difference output map between the filled DTM and the original DTM was at 50 m with the largest being at 5 m. This means that the 50 m DTM can be assumed as the most consistent DTM to use. Although, the DTMs at 5 m scale and 20 m scale obtained more sinks, these may not be accounted for in terms of just water bodies (e.g. lakes and reservoirs). The sink analysis might have assumed that some of the sinks were being categorised as quarries or small depressions, thereby giving higher residual differences than might have been anticipated. Superimposing a river network on top of the original DTM is inappropriate for this work as when trialled it introduced large anomalies due to the low grid resolution of the DTM at 50 m and the high drainage density in some parts of GB.

DTM technique used	**GB**
	Number of Sinks
5 m	244570
25 m	16092
50 m	79923
	Difference maps
5 m	0 – 245.62
25 m	0 – 298.34
50 m	0 – 176.23
	Sink filling method (50 m DTM)
Planchon and Darboux (2001)	0 – 176.08
Wang Liu (2006)	0 – 324.44

Table 3.2: DTM analyses for GB

The terrain attributes from the 50 m DTM were derived using System for Automated Geoscientific Analyses (SAGA) GIS. This is a free GIS package which can be downloaded. The version used for these analyses was SAGA 2.3.0 (Conrad et al, 2015). The DTM was then used to derive 14 terrain attributes (Table 3.3) which were then resampled to 100 m grid scale in connection with Global Soil Map specifications.

Terrain parameters	
Analytical Hillshading (AH)	Catchment Area
Slope	Topographic Wetness Index (TWI)
Aspect	LS Factor (LSFact)
Plan Curvature	Channel Network Base Level (CNBL)
Profile curvature	Channel Network Distance
Convergence Index (CI)	Valley depth (VD)
Closed Depressions	Relative slope position (RSP)

Table 3 3: List of terrain parameters obtained from sink analyses methods in SAGA GIS.

These terrain attributes are significant with respect to soil-landscape relationships as they have an influence on the control and distribution of physical, chemical and biological soil properties. The basic terrain analysis also considers Channel Network, Drainage Basins and Channel Density. Previous studies have illustrated that altering pixel sizes in computing terrain attributes used in DSM analysis (Smith et al, 2006), the importance of neighbourhood size (Zhu et al, 2001) and a combination of pixel and neighbourhood alterations (Roecker et al, 2008) have an influence on how useful a DTM can be.

Finally, it can be noted that the Planchon and Darboux (2001) method derived the least difference from the original DTM from the filled sinks compared to Wang Liu (2006) method and therefore this approach is used in this PhD study for obtaining subsequent relief parameters in SAGA GIS based upon the 50 m DTM.

3.4.5. Parent Material (P)

Geological information used for the DSM assessments have been obtained from the British Geological Survey (BGS). The DiGMapGB-250 dataset comprises geological map data at 1:250,000 scale but only provides bedrock information; there is no superficial, mass movement or artificial theme available onshore at this scale. Thus, it does not consider glacial deposits or recent sedimentary deposits such as alluvium or peat. This dataset was converted to raster format and then resampled to 100 m resolution.

3.4.6. Landscape Position (N)

Two landscape approaches were considered for DSM modelling: Hammond Landscape classification and the Soil and Terrain (SOTER) database.

The Hammond classification was created by the American Edwin H. Hammond, who originally classified landforms across the United States. His method of determining landform units is achieved by examining landforms within a square window of 6 x 6 miles on a 1:250,000 scale topographic map. From this, three elements are identified: slope, local relief and profile type.

These factors are categorised, and landform units are defined through combinations of these elements (Hrvatin and Perko, 2009).

The first part of the classification is slope unit which is defined by calculating the percentage of an area for each window which had a slope less than 8% (Hrvatin and Perko, 2009). This is categorised into 4 levels:

- Above 80% gently sloping terrain
- 50-80% gently sloping terrain
- 20-50% gently sloping terrain
- Below 20% gently sloping terrain

The second part of the classification is determined by local relief which is the difference between the maximum and minimum elevation (Hrvatin and Perko, 2009). This is categorised into 6 levels.

- 1: 0 – 30 m
- 2: 30 – 90 m
- 3: 90 – 150 m
- 4: 150 – 300 m
- 5: 300 – 900 m
- 6: 900 – 1500 m

The final element of Hammond's classification focusses on profile type. This is categorised into 4 levels by calculating the percentage of gently sloping terrain lying below or above the window's average elevation (Hrvatin and Perko, 2009).

- > 75% of gently sloping terrain lying in lowland areas.
- 50-75% of gently sloping terrain lying in lowland areas.
- 50-75% of gently sloping terrain lying in upland areas.
- > 75% of gently sloping terrain lying in upland areas.

By combining these three features, landform units are defined and set out on a grid. Hammond presented the classification results through boundaries between the landform units defined subjectively by edges of plains, low mountains, and large relief forms. As a result, the map is coarse in nature and highly subjective (Hrvatin and Perko, 2009).

The Soil and Terrain SOTER methodology is based on terrain and associated lithology. The classifications for the four terrain layers are described below (Figure 3.5).

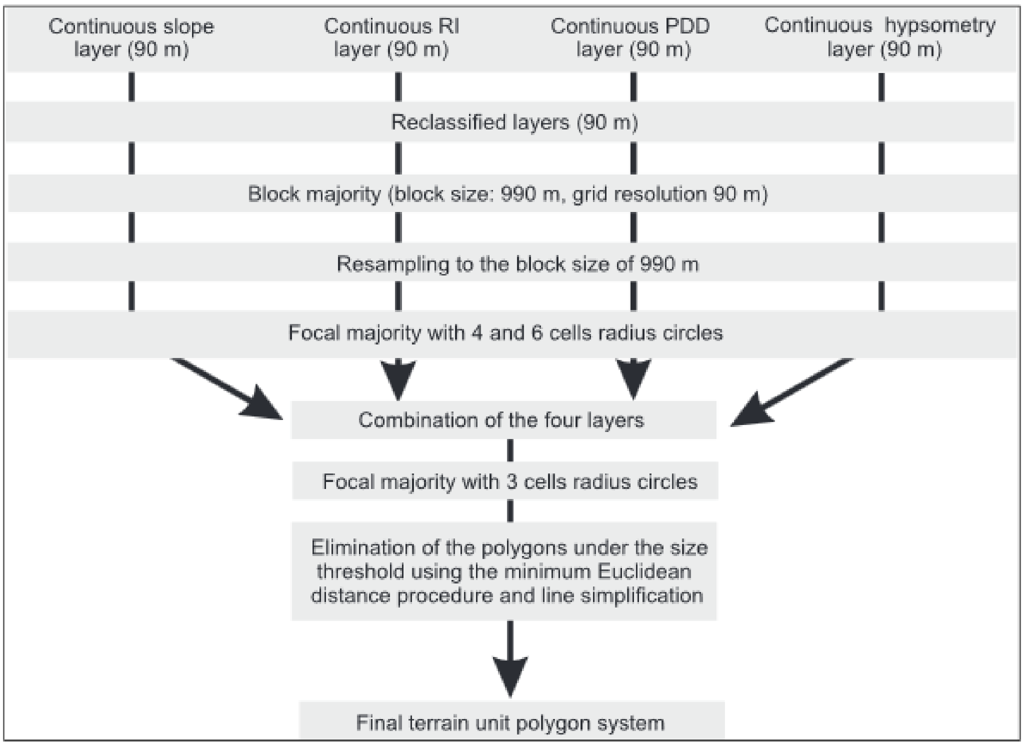

Continuous slope layer (90 m)
Continuous RI layer (90 m)
Continuous PDD layer (90 m)
Continuous hypsometry layer (90 m)
Reclassified layers (90 m)
Block majority (block size: 990 m, grid resolution 90 m)
Resampling to the block size of 990 m
Focal majority with 4 and 6 cells radius circles
Combination of the four layers
Focal majority with 3 cells radius circles
Elimination of the polygons under the size threshold using the minimum Euclidean distance procedure and line simplification
Final terrain unit polygon system

Figure 3.5: SOTER process from Dobos et al, (2010).

The SOTER modified classification scheme for acquiring a continuous slope layer is calculated from the slope function and uses the average maximum technique (Burrough, 1986). This is then classified into 7 SOTER slope class categories. The Relief Intensity (RI) classification is the difference in altitude between the highest and lowest points within a specified distance. The focal range of the relief is calculated and then reclassified into four RI SOTER classes. The use of a DTM makes it possible to derive an artificial drainage network which characterises the landscape. Dobos et al, (2005) refer to this as Potential Drainage Density (PDD). The flow direction and accumulation are obtained, and drainage is then derived from flow accumulation. As a result, the PDD is derived from focal sum modelling of the drainage based on the relief in the area. PDD values obtained from this are reclassified into 2 SOTER classes. Hypsometry (elevation) is land elevation based in relation to sea level and can be computed into 10 classes. These 4 components (slope, RI, PDD and hypsometry) are then combined to get final SOTER Landform Type (Dobos et al, 2010).

Both Hammond and the Soil Terrain Database (SOTER) are landscape map units which are derived from the DTM and used in this PhD project. These datasets were both converted to raster format and then resampled to 100 m resolution.

Thus, after reviewing and testing the suitability of various covariates as input datasets to develop a DSM model for GB, the following were selected (Table 3.4).

Covariate used	**Original Scale**	**Resampled to 100 m?**
Dominant Major Soil sub group (Scotland) (S)	1:250,000	Y
Dominant Major Soil Group (England and Wales) (S)	1:250,000	Y
Annual Mean Temperature (AMT)	1 km^2	Y
Annual Precipitation (AP)	1 km^2	Y
Isothermality (ISO)	1 km^2	Y
Mean Diurnal Range (MDR)	1 km^2	Y
Seasonal Precipitation (SP)	1 km^2	Y
Seasonal Temperature (ST)	1 km^2	Y
Land Cover Map (LCM2000) (O)	1 km	Y
Aspect (R)	100 m	Y*
Slope (R)	100 m	Y*
Analytical Hillshading (R)	100 m	Y*
Convergence Index (R)	100 m	Y*
Longitudinal Curvature (R)	100 m	Y*
Cross Sectional Curvature (R)	100 m	Y*
Land-Slope Factor (R)	100 m	Y*
Topographic Wetness Index (R)	100 m	Y*
Relative Slope Position (R)	100 m	Y*
Valley Depth to Channel Network (R)	100 m	Y*
Channel Network to Base Level (R)	100 m	Y*
ROCK (P)	1:250,000	Y*
SOTER (N)	100 m	Y*
HAMMOND (N)	100 m	Y*

Table 3.4: Main covariates used for DSM development work carried out in Chapter 4.

*created from 50 m DTM and resampled to 100 m.

References

Abella, S.R., Zimmer, B.W., 2007: Estimating organic carbon from loss-on-ignition in northern Arizona forests soils. *Soil Science Society of America Journal*, 71, pp.545-550.

Avery, B.W., Bascomb, C.L., 1982: Soil Survey Technical Monograph No.6. Soil Survey Laboratory Methods. Harpenden.

Avery, B.W., 1973: Soil classification in the soil survey of England and Wales, *European Journal of Soil Science*, 24, pp. 324-338.

Ball, D.F., 1964: Loss-on-ignition as an estimate off organic matter and organic carbon in non-calcareous soils. *Journal of Soil Science,* 15, pp.84-92.

BioClim., 2016: Bioclimatic variables. Accessed from http://www.worldclim.org/bioclim [Last Accessed 24th August 2018].

Campbell, G.A., Lilly, A., Corstanje, R., Mayr, T.R., Black, H.I.J., 2017: Are existing soils data meeting the needs of stakeholders in Europe? An analysis of practice use from policy to field. *Land Use Policy*, 68, pp.211-223.

Centre for Ecology and Hydrology., 2016: Integrated Hydrological units of the United Kingdom: Hydrometric Areas without a coastline. Accessed from: https://data.gov.uk/dataset/dc105e0b-89af-4cdb-86df-7105bfc04a1b/integrated-hydrological-units-of-the-united-kingdom-hydrometric-areas-without-coastline [Last Accessed 24th August 2018].

Chapman, S.J., Bell, J.S., Campbell, C.D., Hudson, G., Lilly, A., Nolan, A.J., Robertson, A.H.J., Potts, J.M., Towers, W., 2013: Comparison of soil carbon stocks in Scottish soils between 1978 and 2009. *European Journal of Soil Science*, 64, pp. 455-465.

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., Böhner., 2015: System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. Geoscientific Model Development, 8, pp.1991-2007, doi: 10.5194/gmd-8-1991-2015.

https://www.geosci-model-dev.net/8/1991/2015/gmd-8-1991-2015.html [Last Accessed 24th August 2018].

Dane, J.H., Topp, G.C., 1976: Methods of Soil Analysis. Part 4, Physical Methods. Soil Science Society of America, Book Series, No.5.

Davies, B.E., 1971: A statistical comparison of pH values of some English soils after measurement in both water and 0.01 M calcium chloride. *Soil Science Society of America Proceedings*, 35, 551–552.

Defra., 2011: Comparison of topsoil carbon changes across England and Wales estimated in the Countryside Survey and the National Soil Inventory. Final report of project: SP1101.

Dobos, E., Daroussin, J., Montanarella, L., 2005: An SRTM - based procedure to delineate SOTER Terrain Units on 1:1 and 1:5 million scales. European Commission Joint Research Centre. Institute for Environmental Sustainability EUR 21571 EN.

Dobos, E., Daroussin, J., Montanarella, L., 2010: A quantitative procedure for building physiographic units supporting a global SOTER database. Hungarian Geographical Bulletin, 59 92), pp.181-205.

Dufourmont, H., Gallego, J., Reuter, H., Strobl, P., 2014: EU-DEM Statistical Validation Report, August 2014.

GlobalSoilMap, 2011a: GlobalSoilMap.net: New Digital Soil Map of the world. [Accessed from https://www.globalsoilmap.net/] [Last accessed 17th February 2018].

GlobalSoilMap., 2011b: Specifications Version 1: GlobalSoilMap.net products: Release 2.1. Technical report.

Global Soil Partnership, 2018: Global Soil Partnership website. Accessed from: [http://www.fao.org/global-soil-partnership/en] [Last Accessed 23rd August 2018].

Hallett, S.H., Sakrabani, R., Keay, C.A., Hannam, J.A., 2017: Developments in land information systems: examples demonstrating land resource management capabilities and options. *Soil Use and Management.* 33, pp. 514-529.

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B. M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotic, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S. and Kempen, B., 2017: SoilGrids250m: Global gridded soil information based on machine learning. *PLoS ONE,* 12, 2: e0169748. https://doi.org/10.1371/journal.pone.0169748

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. *International Journal of Climatology*, 25: pp. 1965-1978.

Hoogsteen, M.J.J., Lantinga, E.A., Bakker, E.J., Groot, J.C.J., Tittonell, P.A., 2015: Estimating soil organic carbon through loss on ignition: effects of ignition conditions and structural water loss. *European Journal of Soil Science*, 66, pp.320–328.

Hrvatin, M., Perko, D., 2009: Suitability of Hammond's method for determining landform units in Slovenia. Acta geographica Slovenica**,** Accessed from: https://ojs.zrc-sazu.si/ags/article/view/1281 [Last Accessed 24th August 2018] doi: https://doi.org/10.3986/AGS49204.

Kissel, D.E., Sonon, L., Vendrell, P.F., Isaac, R.A., 2009: Salt concentration and measurement of soil pH. Communications in Soil Science & Plant Analysis, 40**,** 179–187.

Konen, M.E., Jacobs, P.M., Burras, C.L., Talaga, B.J., Mason, J.A., 2002: Equations for predicting soil organic carbon using loss-on-ignition for North Central U.S. soils. *Soil Science Society of America Journal.* 66, pp.1878- 1881.

Lawley, R., Emmett, B.A., Robinson, D.A., 2014: Soil observatory lets researchers dig deep. *Nature.* 509, pp.427

Lilly, A. and Chapman, S.J. 2015: Assessing changes in carbon stocks of Scottish soils: lessons learnt, IOP Conf. Series. *Earth and Environmental Sciences*. 25, 012016

Lilly, A., Miller, D., Towers, W., Donnelly, D., Poggio, L., Carnegie, P., 2015: Mapping Scotland's Soil Resources. Society of Cartographers Bulletin Vol 48.

Lilly, A., Bell, J.S., Hudson, G., Nolan, A.J., Towers, W., 2010: National Soil Inventory of Scotland (NSIS_1): site location, sampling and profile description protocols (1978-1988): Technical Bulletin, Macaulay Institute, Aberdeen.

Lilly, A., Bell, J.S., Hudson, G., Nolan, A.J., Towers, W., 2011: National Soil Inventory of Scotland 2007-2009: Profile description and soil sampling protocols. (NSIS2). Technical Bulletin, James Hutton Institute.

Macaulay Institute for Soil Research.,1971: Laboratory notes on Methods of Soil Analysis.

Miller, R.O., Kissel, D.E., 2010: Comparison of soil pH methods on soils of North America. *Soil Science Society of America Journal*, 74**,** pp. 310–316.

Minasny, B., McBratney, A.B., 2001: The Australian soil texture boomerang: A comparison of the Australian and USDA/FAO soil particle-size classification systems. *Australian Journal of Soil Research*, 39, 6, pp.1443-1451.

Minasny, B., McBratney, A.B., 2011: Models relating soil pH measurements in water and calcium chloride that incorporate electrolyte concentration. *European Journal of Soil Science*, pp.1-5.

Moore, R.V., Morris, D.G., Flavin, R.W., 2000: CEH digital river network of Great Britain (1:50000). Accessed from: https://catalogue.ceh.ac.uk/documents/7d5e42b6-7729-46c8-99e9-f9e4efddde1d [Last accessed 24th August 2018].

Nemes, A., Wösten, J.H.M., Lilly, A., Oude Voshaar, J.H., 1999: Evaluation of different procedures to interpolate particle-size distributions to achieve compatibility within soil databases. *Geoderma*, 90, pp.187-202.

Omuto, C., Nachtergaele, F., Rojas, R.V., 2013: State of the art report on Global and Regional Soil Information: Where are we? Where to go. Global Soil Partnership Technical Report.

Ordnance Survey, 2016a: OS Terrain 5 User Guide. Accessed from: http://digimap.edina.ac.uk/webhelp/os/data_files/os_manuals/os-terrain-5-user-guide_v1.1.pdf [Last accessed 24th August 2018].

Ordnance Survey., 2016b: OS Terrain 50 User Guide. Accessed from: https://www.ordnancesurvey.co.uk/docs/user-guides/os-terrain-50-user-guide.pdf [Last accessed 24th August 2018].

Ordnance Survey., 2016c: Boundary Line User Guide. Accessed from: https://www.ordnancesurvey.co.uk/docs/user-guides/boundary-line-user-guide.pdf [Last accessed 24th August 2018].

Planchon, O., Darboux, F., 2001: A fast, simple and versatile algorithm to fill the depressions of digital elevation models. *Catena*, 46, pp.159–176.

Roecker, S.M., Thompson, S.M., 2008: Scale effects on Terrain Attribute Calculation and Their Use as Environmental Covariates for Digital Soil Mapping. Proceedings of the 3rd Global Workshop on Digital Soil Mapping, Logan.

Sanchez, P. A., Ahamed, S., Carré, F., Hartemink, A.E., Hempel, J., Huising, J., Lagacherie, P., McBratney, A.B., McKenzie, N.G., de Ourdes Mendonça-Santos, M., Minasny, B., Montanarella, L., Okoth, P., Pal, C.A., Sachs, J.D., Shephard, K.D., Vagen, T., Vanlauwe, B., Walsh, M.G., Winowiecki, L.A., Zhang, G.L., 2009: Digital soil map of the world. *Science,* 325, 5941, pp. 680-681.

Soil Survey of Scotland Staff., 1981: Soil maps of Scotland at a scale of 1:250 000. Macaulay Institute for Soil Research, Aberdeen.

Smith, M.P., Zhu, A.X., Burt, J.E., Stiles, C., 2006: The effects of DEM resolution and neighbourhood size on digital soil survey. *Geoderma*, 137, pp. 58-69.

USDA.,1978: Soil Taxonomy. Agriculture Handbook no. 436. Washington D.C.: USDA, Soil Conservation Service.

Wang, L., Liu, H., 2006: An efficient method for identifying and filling surface depressions in digital elevation models for hydrologic analysis and modelling. *International Journal of Geographical Information Science,* 20, 2, pp. 193-213.

World Climate., 2016: WorldClim – Global Climate data: free climate data for ecological modelling and GIS. [Available from http://www.worldclim.org/] [Last Accessed 29th May 2018].

Zhu, A.X., Hudson, B., Burt, J., Lubich, K., Simonson, D., 2001: Soil mapping using GIS, expert knowledge and fuzzy logic. *Soil Science Society of America Journal*, 65, pp. 1463-1472.

4 EVALUATION OF TWO MODEL TYPOLOGIES AND THEIR BEHAVIOUR IN GENERATING SOIL PROPERTY PREDICTIONS: STUDIES FROM PILOT AREAS ACROSS GB

Abstract

Digital Soil Mapping (DSM) applies observed field data to a range of statistical models and covariates to produce mapping outputs. To quantify soil patterns associated within areas, modelling methods are applied. When considering which model to use in DSM, it is fundamental to investigate the model performance during training and development phases. The input data and how models behave are fundamentally important to modelling soil properties.

This chapter explores two critical components of DSM model development: model performance and model extrapolation to generate predictions at unsampled locations. This study compares Boosted Regression Trees (BRTs) and Multivariate Adaptive Regression Splines (MARS) models for mapping loss-on-ignition (LOI), soil pH and texture at 2D and 3D across two comparable GB pilot areas. These areas were selected based on having similar soil types, relief, land cover and land use. Modelling of these soil properties was focussed at six standard depth intervals, using GlobalSoilMap.net criteria. DSM maps were generated across the pilot areas at 100 m resolution. The main factors influencing the spatial distribution of these soil properties for all depths for both MARS and BRT models were soil maps, bedrock geology, land cover map, topographic wetness index and land classification classes. Our results suggest that MARS models produce better model performances than BRTs for predicting soil properties within the training data across the pilot areas at training level. However, when you deploy MARS models to areas beyond the training environment, they

extrapolate outside the suitable environment and as a result produce soil property values which do not make sense. BRT models, despite not being as strong statistically in a training capacity, are more consistent in illustrating relationships between soil properties and associated pedology. Future research will focus on improving the predictions of these soil properties to fully encapsulate the range of soils across GB.

4.1. Introduction

Digital Soil Mapping (DSM) has been discussed as a popular approach for producing improved soil property maps at finer resolution. This methodology uses observed field data and associated environmental covariates and a range of mathematical models to predict soil properties (Dobos et al, 2006; Behrens and Scholten, 2006). DSM has become useful because of increased costs of conducting soil surveys and a dwindling number of surveyor's who have the necessary expert knowledge (McBratney et al, 2003; Hudson, 1992). DSM has many advantages especially its efficiency at mapping and modelling soil properties and its ability to quantify the associated uncertainties (McBratney et al, 2003; Carré et al, 2007; Hudson, 1992, Minasny and McBratney, 2016). To quantify these soil properties, modelling methods—defined as 'the use of mathematical equations to simulate and predict real events and processes' (Grunwald, 2009) – are applied. These are created by developing statistical relationships between observed data and covariates (often indicated as model *'training'*) and applying these to generate predictions at unsampled locations (often represented as model *'deployment'*).

Grunwald (2009) conducted an exhaustive meta-analysis of models used in DSM, focusing on the range of mathematical models employed across a variety of geographical regions at various scales. The selection of models varies but generally two types of modelling typology are used. The first category of models is based on recursively partioning the predictor space (e.g. Classification Regression Trees (CART) (Breiman, 1984), Random Forests (RFs) (Breiman, 2001) and cubist regression model methods (Kuhn et al, 2012)). These models can capture complex interaction structures in a dataset and combine trees by ensemble methods to reduce the associated variance (Hastie et al, 2009; Nussbaum et al, 2018). A second group of models are regressive in nature, where the established relationship is linear (e.g. multiple regression (Hastie et al, 2009), partial least squares regression (PLSR) (Viscarra Rossel and Behrens, 2010) and neural networks (Behrens et al, 2005)). Certain mathematical models have characteristics which affect their performance. For instance, recursive partitioning models tend to overfit and regression-based models tend to be less effective where there are complex relationships present.

When assessing which appropriate model should be used in DSM, the model with the best performance parameters obtained during training is usually used. However, several considerations which affect model performance need to be evaluated. The first of these refers to the observed data for modelling soil properties. Observed data is usually represented by a soil core or representative profile which contains associated soil characteristics. As the model being used will be at a specific scale, how it applies to covariates with different grid scales will contribute to how well it performs. Furthermore, this will be dependent on location (Cavazzi et al, 2013) and is discussed extensively elsewhere (Leempoel et al, 2015; Pain, 2005; Thompson et al, 2001).

Model performance is assessed through cross-validation, either by retaining a random section of the data for validation, or by using a truly independent validation dataset. In all cases, the validation is affected by the scale of both the covariates and the validation dataset. These are known to be associated to the data model which will contribute to experiencing uncertainty and have been explored by others previously (Cavazzi et al, 2013; Corstanje et al, 2008).

A second aspect to consider when assessing model performance is when the models are applied to a grid of covariates i.e. moving from model training to model deployment. During this stage, the model typology is important as its behaviour under deployment is entirely dependent on the characteristics.

In this chapter, two crucial components of DSM model development are explored: i) model performance under training circumstances and ii) model deployment where the model is used to generate predictions at unsampled locations. This is done using two model types: one using a recursive partitioning approach (Boosted Regression Trees (BRTs)) and another which characterizes a regression-based methodology (Multivariate Adaptive Regression Splines (MARS)). The model performance and behaviour are illustrated during training phase and during model deployment. A hypothesis to this research is that good model performance at training level cannot be taken as an assurance of good model performance when deployed to unsampled locations.

This study will focus on comparing MARS and BRT models across two pilot areas in Scotland and England and Wales. Soils data found in these pilot areas are held by the successor Institutions of the Soil Survey of Scotland and Soil Survey of England and Wales (Lilly et al, 2010; Hallett et al, 2017). Both national Soil Surveys collected 'representative profiles' to characterize soil mapping and taxonomic units. Furthermore, profile datasets from both Surveys have objective, grid samples collected to provide a statistical assessment of soil characteristics (National Soil Inventories). These data sources accompanied with associated sampling and methodologies provide a unique opportunity to test BRTs and MARS modelling performance with the same covariates in comparable soil-land use environments. This helped to test which DSM method is best at predicting soil properties. Outcomes from this will reflect some of the issues of producing consistent DSM across larger or even global scale situations.

4.2. Models used

4.2.1. Boosted Regression Trees (BRTs)

One of the most commonly used modelling approaches for predicting soil properties are Classification and Regression based systems. Classification and Regression Trees (CART) are a rule-based method that generates a binary tree through 'recursive partitioning', a splitting process which creates a series of decision trees based on the predictor variables (Breiman et al, 1984; Prasad et al, 2006). Breiman et al, (1984) illustrate that since the 1980s, statisticians have developed CART providing new features such as Random Forests (RFs) and Boosted Regression Trees (BRTs) (Sutton, 2005).

BRTs are statistical models which create multiple boot-strapped regression trees without pruning and averaging the outputs (Elith et al, 2008). The boosting associated with BRT models refers to improving the accuracy of models by taking an average of many rules associated with variables. This will increase the performance of a classifier and produce low error rates (Bauer and Kohavi, 1999; Hastie et al, 2009). BRTs uses two algorithms: one which utilises regression trees from the classification and regression tree (decision tree) and the other involves building the boosting and combining the models (Elith et al, 2008).

In the first algorithm, the tree-based models partition the predictor space using a series of rules to identify areas where similar relationships between predictors lie. A constant is assigned to each area and regression trees are fitted using the mean response for observations in the area being investigated. This is assuming that the associated errors are normally distributed (Elith et al, 2008). The predictor variables in a dataset are split by a binary approach which are split further for the best model fit to be achieved. This is repeated many times until the prediction error is minimized (Hastie et al, 2009).

Compared to Random Forests, BRTs are more likely to be robust against over-fitting due to using randomly selected subsamples to fit the data at each segment (Friedman, 2002; Hastie et al, 2009). Another advantage of using BRTs is that a higher predictive accuracy and better interpretation of the relationships between variables is achieved (Friedman, 1991). There is however an overemphasis on categorical variables in comparison with continuous variables (Prasad et al, 2006).

4.2.2. Multivariate Adaptive Regression Splines (MARS)

The second most commonly used approaches in DSM modelling are those that are regression-based i.e. partial least square regression (PLSR) or multiple regression (Grunwald, 2009). Multivariate Adaptive Regression Splines (MARS) models are an example of this. MARS have become a popular choice in data mining because these models do not assume any type of relationship (e.g. linear, logarithmic) between the dependent and independent variables (Friedman, 1991). MARS models construct a relationship from a set of coefficients and basis functions which are determined from a regression of the data. Basis functions are known by two-sided truncated functions for linear or nonlinear expansion which show relationships between response and predictor variables (Friedman, 1991). These basis functions are known as (t-x) + and (x-t) + where t refers to parts of the linear regression which are determined from the data (Hastie et al, 2009).

MARS models choose a weighted sum from a set of basis functions that span all values from each predictor in a dataset. The algorithm then examines the input space and predictor values alongside interactions between variables. During this process, an accumulated number of basis functions are added to help maximize an overall goodness-of-fit. As a result, MARS determines the most important independent variables and significant interactions (Hastie et al, 2009).

However, one difficulty arising from MARS models is how sensitive they are at extrapolating relationships from the basis functions (Prasad et al, 2006). However, Nawar et al, (2015) argue that MARS models are more robust at predicting soil properties than other models such as Partial Least Square Regression (PLSR). MARS models are flexible in fitting complex, non-linear relationships and investigating the interactions and effects of certain variables, thus increasing model performance, something which PLSR does not do (Ghasemi et al, 2013). MARS is known to be better at estimating all soil properties than PLSR as it overcomes deviations that occur between predicted values and measured soil values at higher ranges (Prasad et al, 2006).

4.3. Study Areas

The two pilot study areas used are situated in the north eastern part of the Midland Valley (SCO), and west England and eastern Wales (EW) (Figure 4.1). These two areas were selected as they contain similar characteristics in terms of soil types, landscapes, land cover and land use (Veronesi et al, 2014).

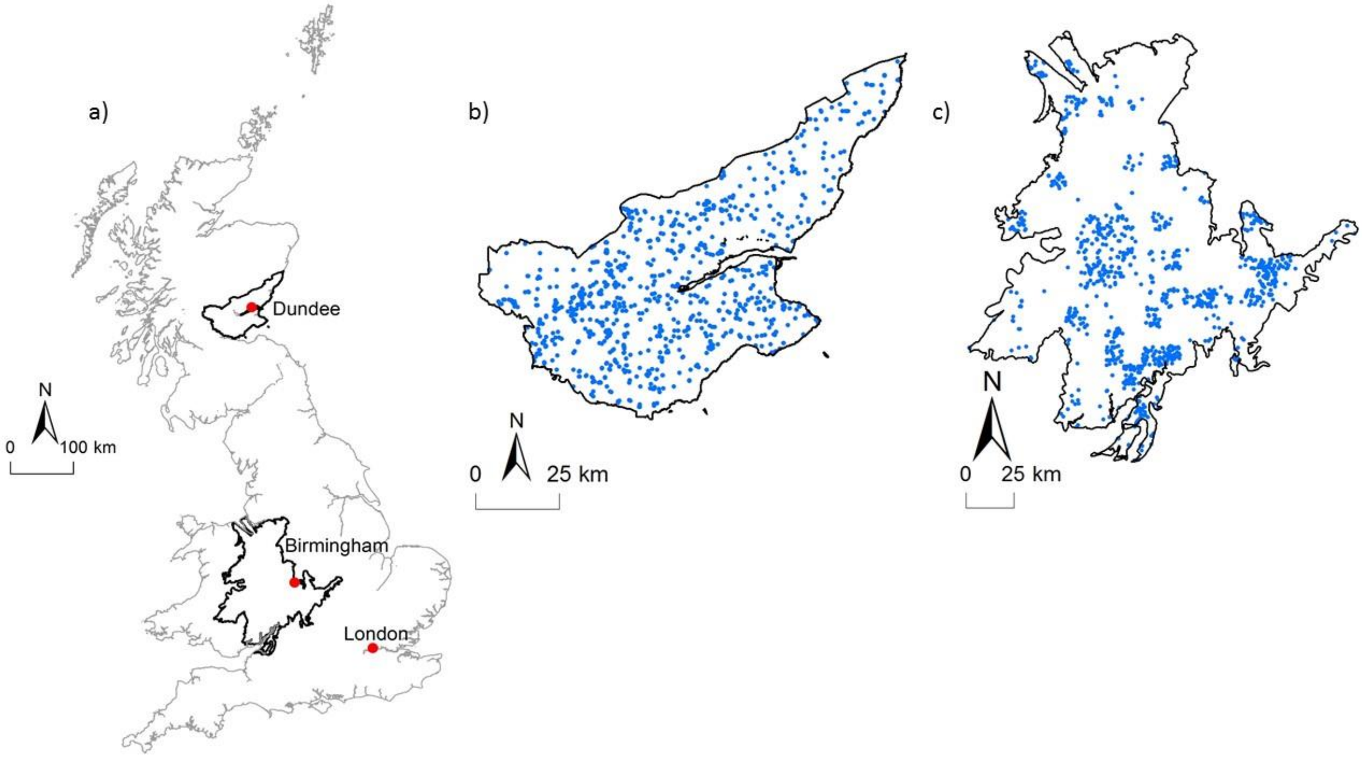


Figure 4.1: Location of the SCO and EW test areas a) with soil profile locations used for example pH training data in blue (b) and (c).

The SCO pilot area has a size of 5,384km^2 and covers the land between the Highland Boundary Fault in the north and the Forth Estuary in the south and from the North Sea in the east to River Earn in the west. Major population centres found in this area are Dundee, Perth and Kirkcaldy. The SCO area is comparable to the landscape units identified for stratification of Scotland's landscape by Scottish Natural Heritage (SNH, 2002). Within the SCO area, there is a variety of contrasting soil types ranging from brown earths, humus-iron podzols, non-calcareous gleys and peaty soils (blanket and basin peats, peaty gleys and peaty podzols).

The most extensive land cover types found in SCO constitute coniferous woodland, arable and improved grassland (Fuller et al, 2000).

The EW pilot area encompasses an area of 13,948km^2 between the major population centres of Liverpool, Birmingham and Shrewsbury. There is also a variety of soils ranging from podzols on sandy parent materials, slowly permeable or seasonally waterlogged clay rich soils to freely draining loamy soils. The most extensive land cover types in EW are arable, coniferous woodland and mountain, heath and bogs (Fuller et al, 2000).

4.4. Datasets

For the EW pilot area, the representative soil profiles were used for training the models. These representative profiles were sampled to characterise the soil series mapped in England and Wales which were collected from 1935 to around 2000 (Hallett et al, 2017). For the SCO pilot area, the representative profiles in the Scottish Soils Database of the Soil Survey of Scotland were used. The main horizons of representative profiles were sampled, and the samples analysed to determine a range of soil properties. The soil properties selected for modelling in both study areas were loss-on-ignition (LOI) (as an indicator for soil organic carbon), soil pH and texture. These properties were chosen based on a survey of stakeholder needs (Campbell et al, 2017) and as they were determined by similar laboratory methods (Macaulay Institute for Soil Research, 1971; Avery and Bascomb, 1982). The numbers of samples available for each soil property, for each pilot area, at depth, are shown below (Table 4.1).

Soil property	GSM depth range (cm)	Number of samples (SCO)	Mean	SD	Observed Range	Number of Samples (EW)	Mean	SD	Observed range
LOI	0-5	949	21.87	25.09	0.43 – 100.00	936	9.91	9.17	2.74– 88.43
	5-15	948	15.98	17.84	0.45 – 96.81	917	8.22	5.56	2.61– 66.80
	15-30	941	9.28	10.99	0.25 – 95.81	839	6.08	4.64	0.00– 79.15
	30-60	929	5.56	8.46	0.31 – 99.79	585	4.21	5.96	0.00– 99.15
	60-100	829	4.21	7.54	0.21 – 99.44	228	4.12	6.00	0.00– 66.04
	100-200	361	3.33	2.45	0.35 – 21.63	74	4.83	8.68	0.94– 66.01
pH	0-5	949	5.44	1.02	3.30 – 8.58	1096	5.96	1.00	2.34 – 8.36
	5-15	948	5.47	0.99	3.30 – 8.58	1095	6.01	0.98	3.12 – 8.75
	15-30	932	5.56	0.90	3.42 – 8.26	1092	6.17	0.99	3.32 – 8.56
	30-60	920	5.70	0.85	3.42 – 8.36	1063	6.40	1.05	2.74 – 8.75
	60-100	829	5.85	0.85	3.23 – 8.74	1063	6.79	1.17	2.26 – 9.02
	100-200	361	6.03	0.92	2.84 – 8.63	371	6.77	1.10	4.08 – 8.75
Sand	0-5	892	53.43	16.53	0.00 – 100.00	1020	36.70	23.70	0.00 – 93.14
	5-15	892	53.68	16.62	0.00 – 100.00	1019	36.86	23.84	0.00 – 94.65
	15-30	890	55.29	17.11	0.00 – 100.00	1013	37.11	24.55	0.00 – 96.96
	30-60	879	58.90	18.93	0.00 – 100.00	987	36.45	26.52	0.00 – 96.78
	60-100	781	60.19	20.39	0.00 – 100.00	902	36.04	29.09	0.00 – 99.41
	100-200	325	62.25	22.83	0.00 – 99.22	347	39.00	31.35	0.00– 100.00
Silt	0-5	892	28.82	12.22	0.00 – 84.97	1020	38.49	17.24	3.00 – 80.74
	5-15	892	28.69	12.15	0.00 – 78.00	1019	38.30	17.27	2.76 – 80.16
	15-30	890	27.65	12.02	0.00 – 78.07	1013	37.77	17.49	0.03 – 83.09
	30-60	879	25.28	12.57	0.00 – 77.61	987	36.56	18.07	1.63 – 81.50
	60-100	781	24.50	13.41	0.00 – 80.83	902	35.94	19.48	0.17 – 87.53
	100-200	325	23.86	15.14	0.00 – 71.31	347	35.47	21.94	0.00 – 98.71
Clay	0-5	892	13.61	8.05	0.00 – 41.97	1020	24.88	14.33	0.00 – 89.00
	5-15	892	13.83	8.03	0.00 – 44.96	1019	24.92	14.32	0.13 – 89.00
	15-30	890	14.54	8.50	0.00 – 48.86	1013	25.14	15.16	0.59 – 89.00
	30-60	879	14.69	9.53	0.00 – 57.77	987	16.69	26.98	1.04 – 89.00
	60-100	781	14.71	9.89	0.00 – 47.89	902	28.11	17.36	0.06 – 89.00
	100-200	325	13.31	10.11	0.00 – 48.99	347	25.83	17.31	0.00 – 97.43

Table 4.1: Sample size and summary statistics for each modelled soil property at depth specified by GlobalSoilMap for the SCO and EW pilot areas.

4.4.1. Comparison of Laboratory and Analytical Methods

As this study has utilised soil property data obtained by a range of laboratory methods obtained from two different soil surveys, it was necessary to establish what was required to produce a unified soil dataset before modelling and mapping in the SCO and EW pilot areas.

LOI was measured using ignition at 850°C for 30mins for the representative profiles for England and Wales (Hallett et al, 2017; Avery and Bascomb, 1982) and at 900°C on the representative profiles for Scotland (Macaulay Institute for Soil Research, 1971). Findings from Chapman et al, (2013) illustrated that LOI was a more consistent measurement of soil carbon for Scottish soils than the determination of carbon concentration. As a result, the England and Wales data was examined to assess compatibility. These data were originally presented as carbon concentration data. However, by applying a formula (LOI = 0.5 / OC) a unified dataset of LOI was developed for both datasets.

In Scotland, the pH of water is measured on a soil to water ratio of 1:3 (Macaulay Institute for Soil Research, 1971) and for measuring the pH of $CaCl_2$, 0.01M $CaCl_2$ is added to the suspension. In England and Wales, measurements are made on a soil to water ratio of 1:2.5 and for measuring the pH of $CaCl_2$, 0.01M $CaCl_2$ is added to the suspension (Avery and Bascomb, 1979). The pH data from Scotland and England and Wales were noted to overlap suggesting no systematic bias for soil pH despite the different methods used across Scotland and England and Wales.

Particle size distribution (PSD) in the England and Wales dataset was derived using the pipette method. This involves breaking up soil particles with hydrogen peroxide (Avery and Bascomb, 1979). PSD samples with organic carbon content greater than 15% were not determined as these are classified as organic soils. PSD in Scotland was measured using the hydrometer technique (Macaulay Institute for Soil Research, 1971). Different particle size classes are used in Scotland than in the England and Wales. In England and Wales, particle size measurements are based on the British Soil Texture Classification (BSTC) whereas in Scotland, texture classes are measured using the United States Department of Agriculture (USDA) classification. The primary difference is the cut off used between the silt and sand fractions where the USDA particle size distribution is (<2 (clay), 2-50 (silt), 50-2000 (sand in microns) and BSTC is <2 (clay), 2-60 (silt), 60-2000 (sand µm). From comparison work done, BSTC and USDA classifications were found to be sufficiently similar across both Scotland and England and Wales. The differences between USDA and BSTC sand and silt showed an average of 3.6% difference which is within quoted experimental error (Dane and Topp, 1976, p283). Thus, no corrections were applied to either England and Wales or Scottish USDA particle size class data. Therefore, there is an assumption that BSTC and USDA PSD are comparable in both EW and SCO pilot areas.

4.4.2. Environmental Covariates

To estimate soil properties at unsampled locations across the SCO and EW pilot areas, 24 grids of environmental covariates were derived based on the SCORPAN methodology (McBratney et al, (2003)). These environmental covariates included:

- Scotland: soil maps containing dominant major soil sub group (MSSG) (Hutton) and for England and Wales: major soil group (MSG) (LandIS). Both datasets are produced on a spatial scale of 1:250,000. MSSG and MSG are on the same level of classification in the soil hierarchies for Scotland and England and Wales and so are directly comparable.
- Climate variables for both pilot areas including Annual Precipitation (AP), Annual Mean Temperature (AMT), Isothermality (ISO), Mean Diurnal Range (MDR), Seasonal Precipitation (SP) and Seasonal Temperature (ST). These are available from www.worldclimate.org (World Climate, 2016)
- Land Cover Map 2000: (Fuller et al, 2000). This is available from: https://www.ceh.ac.uk/services/land-cover-map-2000
- A 50m Digital Terrain Model (DTM) created by Ordnance Survey and constructed from the 1:50,000 topographic maps. From this, 12 derivatives were created and obtained using SAGA GIS (Conrad et al, 2015). These were Aspect, Slope, Analytical Hillshading, Convergence Index, Longitudinal Curvature, Cross Sectional Curvature, Land-Slope Factor, Topographical Wetness Index, Relative Slope Position, Valley Depth, Valley Depth to Channel Network, Channel Network Base Level. 50m DTM is available from: https://www.ordnancesurvey.co.uk/business-and-government/products/terrain-50.html (OS Terrain 50, 2018)

- Digital Geological Map of Bedrock (DigMap250; ROCKS): from the British Geological Survey at a scale of 1:250,000. This is available from: (BGS, 2018) http://www.bgs.ac.uk/products/digitalmaps/digmapgb_250.html
- SOTER (World Soil and Terrain Digital Database) based on two primary soil formation phenomena: terrain and lithology. The classifications for the four terrain layers (slope, relief intensity, potential drainage density and hypsometry) are described in Dobos et al, (2010). These four components are then combined to get final SOTER Landform Type.
- Hammond Land Classification: information based on slope, local relief and profile type (Hrvatin and Perko, 2009).

Each covariate raster was resampled to a common resolution and extent of 100 m.

4.5. Modelling and mapping of soil properties

The variation of soil properties was modelled and mapped and modelled in 2D and 3D across the SCO and EW pilot areas. Firstly, depth functions were fitted for the representative soil profiles located in these pilot areas using an equal area mass preserving spline for each soil property (Bishop et al, 2001; Malone et al, 2009) at six standard soil depths (0-5, 5-15, 15-30, 30-60, 60-100, 100-200 cm), as defined by GlobalSoilMap.net specifications (GSM) (GlobalSoilMap, 2011) (Table 4.1). To produce these, R packages 'GSIF' (Hengl et al, 2017), 'aqp' (Beaudette et al, 2018), 'plyr' (Wickham, 2016) and 'sp' (Pebesma and Bivand, 2005) (R Core Team, 2013) were used. These datasets were inserted into ArcMap 10.2.1 and point shapefiles were created for each of the soil properties at specific GlobalSoilMap.net depth intervals across the whole of Scotland and England and Wales. These point shapefiles were then clipped to the SCO and EW pilot areas and sampled for each depth alongside a list of covariates. These datasets were saved as text files and implemented into R to be tested using BRT and MARS models.

After removal of null values and samples within 100 m of each other, the representative profile data was used for training the MARS and BRT models for the SCO and EW pilot areas. For MARS models, the ‘earth’ (Millborrow, 2017) and ‘epiR’ (Stevenson, 2015) packages were used. For the BRT models, ‘dismo’ (Hijmans et al, 2011) and ‘gbm’ (Ridgeway, 2017) packages were used. Statistical indices (e.g. R^2, RMSE) were recorded alongside predicted data range and observed data range.

After evaluating the performance of the training datasets, this data was applied to a deployment dataset created using ‘rgdal’ (Bivand et al, 2015), ‘raster’ (Hijmans, 2016) and ‘sp’ packages (Pebesma et al, 2018). This produced a stack of rasters in a table with each column representing a given covariate at 100 m resolution alongside predicted soil property values. These were used to generate raster maps within the SCO and EW pilot areas. The construction of these maps was generated at 100 m resolution, using ‘raster’ (Hijmans and Etten, 2013) package, saved as a TIF file and then transported into ArcMap 10.2.1. The R code used for this can be found in Appendix 7.

4.6. Results and Discussion

4.6.1. Main influencing covariates

The main covariates which influence predicting soil LOI, pH and texture properties across both pilot areas and at all depth were the soil types (MSSG and MSG), bedrock geology (ROCKS), land cover map (LCM2000), topographic wetness index (TWI), SOTER and Hammond. For both models, this was done by firstly determining which covariates are continuous or categorical. Then the models were run to assess the optimal performance of each. As a result, both BRT and MARS models determined how important each covariate is in influencing soil properties across the pilot areas in 2D and 3D.

4.6.2. BRT and MARS model comparisons

For LOI in both SCO and EW pilot areas, overall, the MARS model was better at predicting LOI values than the BRT model within the training dataset (see Tables 4.2 and 4.3). The R^2 is better for MARS than BRT at many depths and the RMSE is lower. In the SCO pilot area at 0-5 cm depth the BRT model produces a better R^2 than the MARS model (0.70 for BRT compared to 0.64 for MARS). However, despite this, MARS have a better overall model performance with R^2 values which range from 0.53 to 0.76 in comparison to values for BRT R^2 values 0.38 to 0.70. Similar model performances are evident for EW where MARS has a better model performance with R^2 values ranging from 0.54 to 0.94 in comparison to values for BRT R^2 values 0.13 to 0.47. RMSE values are also lower for the MARS model (range from 2.20 to 5.18) in comparison to BRT model (range from 4.34 to 7.39). The BRT model did not work well for depths below 30-60 cm for the EW pilot area due to a lack of variability in the dataset reflecting sparse and low value data at depth. In the subsoil there are much lower values of LOI (unless the soil is a peat). In the deployment table for EW, for example, there are near to or zero values in the observed LOI ranges which, despite a change in the model parameters, presented modelling issues so N/A values were recorded.

For pH in both the SCO and EW pilot areas, the models produced different results compared with LOI. There is agreement that the MARS model provides better predictions than the BRT model, but it is not as obvious as with the LOI results. For instance, in SCO pilot area, BRT models produce better R^2 values in the upper depths (0-5, 5-15, 15-30 cm). However, in general, the R^2 values between the models are very similar for both pilot areas. The R^2 for both models are around 0.50 at most depths and RMSE values range from 0.69 to 1.00.

For sand, silt and clay in both pilot areas, the results for the BRT and MARS models are similar. In the SCO pilot area, BRT models produce a better R^2 for sand at most depths but produce much higher RMSE in comparison to MARS. For Silt, the R^2 is similar for both MARS and BRT but comparable to sand at all depths, with the RMSE higher for BRT than it is for MARS. Clay likewise showed similar results with EW pilot area results for SCO. Overall, the

RMSE for the texture components is higher for BRT than it is for MARS. However, unlike the SCO pilot area, the clay component does produce higher RMSE at depths below 30 cm for MARS than it does for BRT.

	LOI				**pH**				**Sand**				**Silt**				**Clay**			
	MARS		BRT		MARS		BRT		MARS		BRT		MARS		BRT		MARS		BRT	
Depth (cm)	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
0-5	0.64	15.06	0.70	14.54	0.60	0.65	0.64	0.63	0.61	10.3	0.65	10.24	0.59	7.81	0.59	8.23	0.51	5.66	0.47	6.21
5-15	0.70	9.81	0.66	11.54	0.60	0.62	0.66	0.59	0.64	9.99	0.67	9.38	0.60	7.68	0.61	8.07	0.49	5.72	0.50	5.96
15-30	0.73	5.76	0.52	8.45	0.61	0.56	0.63	0.57	0.61	10.64	0.62	11.06	0.60	7.58	0.60	7.99	0.52	5.92	0.48	6.49
30-60	0.73	4.38	0.38	7.33	0.60	0.54	0.59	0.57	0.60	12.04	0.59	12.61	0.55	8.40	0.56	8.78	0.52	6.59	0.49	7.15
60-100	0.76	3.68	0.46	6.42	0.57	0.56	0.54	0.61	0.51	14.23	0.56	14.27	0.52	9.33	0.54	9.68	0.44	7.39	0.46	7.78
100-200	0.54	1.66	0.51	1.98	0.40	0.72	0.52	0.75	0.47	16.68	0.57	17.62	0.40	11.75	0.51	12.55	0.48	7.32	0.51	8.22

Table 4.2: Results from BRT and MARS modelling for predicting soil properties in SCO pilot area at GSM specified depths.

	LOI				pH				Sand				Silt				Clay			
	MARS		BRT		MARS		BRT		MARS		BRT		MARS		BRT		MARS		BRT	
Depth (cm)	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
0-5	0.68	5.18	0.47	7.39	0.52	0.70	0.51	0.73	0.72	12.62	0.73	12.73	0.73	8.91	0.73	9.17	0.63	8.68	0.62	9.49
5-15	0.52	3.83	0.36	4.63	0.53	0.67	0.53	0.71	0.74	12.21	0.73	12.75	0.74	8.86	0.74	9.10	0.64	8.63	0.62	9.24
15-30	0.54	3.16	0.29	4.34	0.54	0.68	0.52	0.72	0.69	13.77	0.72	13.43	0.74	8.91	0.73	9.34	0.59	9.70	0.60	10.05
30-60	0.72	3.17	0.14	5.82	0.58	0.68	0.57	0.71	0.67	15.32	0.68	15.30	0.71	9.79	0.70	10.14	0.54	11.32	0.59	11.09
60-100	0.88	3.11	N/A	N/A	0.26	1.00	0.28	1.08	0.62	18.03	0.65	17.95	0.64	11.73	0.64	12.06	0.55	11.60	0.57	12.12
100-200	0.94	2.21	N/A	N/A	0.63	0.67	0.60	0.81	0.59	20.03	0.67	19.74	0.68	12.35	0.69	13.59	0.59	11.14	0.52	13.65

Table 4.3: Results from BRT and MARS modelling for predicting soil properties in EW pilot area at GSM specified depths.

4.6.3. Mapping outputs for MARS and BRT models

From the R^2 and RMSE values and in terms of our understanding between observed predicted values, MARS is the better model to use and subsequently deploy to training data for LOI to a larger area. However, for pH and texture, both models are comparable for both MARS and BRT for both the SCO and EW pilot areas. A concordance correlation coefficient was undertaken on the training data for both pilot areas to see how powerful the BRT and MARS models are. From Table 4.4, it can be noted that MARS models consistently have better correlation coefficients for all soil properties than BRT outputs across both SCO and EW pilot areas and at all depths. However, when deploying the data, the MARS models produce unrealistic values. This is because they are extrapolated beyond the range of values in the training dataset, which is characteristic of the MARS modelling approach (Friedman, 1991). For example, in some of the maps (e.g. LOI at 5-15 cm; Figure 4.2a and Figure 4.2b) there are extremely high positive values (above 100%, depicted in orange) and extremely high negative values (less than 0%, depicted in black) which are clearly not possible. This is due to the MARS model predicting many overlapping splines that creates excess noise. This results in the model failing to decipher appropriate values in an area that it should have knowledge of dealing with better from the training data. There is also a boundary effect which is more profound in the MARS maps than the BRT maps. This is due to the influence of more organic soils being found in these areas and reflecting the change in land use to semi-natural.

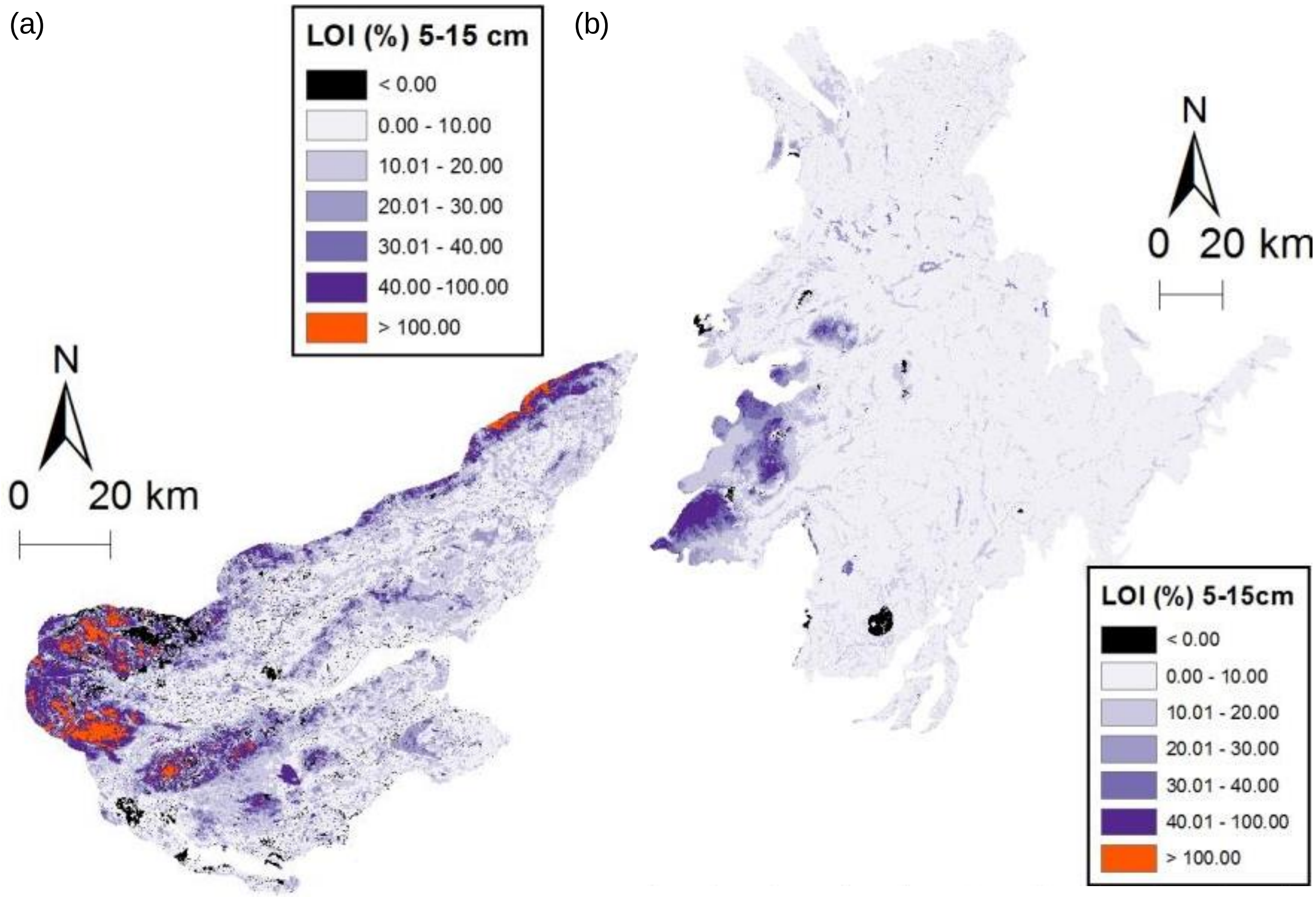


Figure 4.2: Mapped outputs from MARS models for LOI at 5-15 cm in pilot areas: a) SCO and b) EW

a)

Depth	LOI		pH		Sand		Silt		Clay	
	MARS	BRT	MARS	BRT	MARS	BRT	MARS	BRT	MARS	BRT
0-5	0.75	0.67	0.69	0.64	0.77	0.71	0.75	0.64	0.64	0.51
5-15	0.75	0,63	0.71	0.69	0.77	0.73	0.73	0.68	0.64	0.56
15-30	0.78	0.53	0.71	0.64	0.73	0.73	0.75	0.62	0.66	0.51
30-60	0.71	0.42	0.69	0.60	0.76	0.63	0.74	0.61	0.66	0.53
60-100	0.64	0.22	0.64	0.55	0.63	0.61	0.74	0.62	0.62	0.46
100-200	0.61	0.41	0.65	0.44	0.65	0.55	0.63	0.38	0.58	0.43

b)

Depth	LOI		pH		Sand		Silt		Clay	
	MARS	BRT	MARS	BRT	MARS	BRT	MARS	BRT	MARS	BRT
0-5	0.81	0.43	0.68	0.58	0.83	0.82	0.85	0.83	0.77	0.69
5-15	0.69	0.31	0.69	0.58	0.85	0.81	0.85	0.82	0.78	0.68
15-30	0.70	0.18	0.69	0.59	0.81	0.80	0.85	0.83	0.74	0.65
30-60	0.83	0.04	0.73	0.63	0.80	0.77	0.83	0.79	0.70	0.65
60-100	0.86	n/a	0.42	0.18	0.76	0.75	0.78	0.78	0.71	0.62
100-200	0.97	n/a	0.77	0.50	0.74	0.68	0.81	0.69	0.74	0.41

Table 4.4: Concordance correlation coefficient for a) SCO and b) EW pilot areas based on BRT and MARS modelling across all depths.

In contrast, the trained BRT models produce a lower R^2 value on average for each soil property at depth. However, when these models are deployed to a larger area, the values predicted in these areas and fit within an anticipated range (Figure 4.3a and 4.3b). To compare with what has been illustrated above, LOI at 5-15 cm, mapped from BRTs for the SCO pilot area show higher LOI values in the west, on the north boundary of the pilot area and lower LOI content further to the east and the coastline. For the EW pilot area, the highest LOI values can be found in the west of the pilot area and lower LOI values are situated towards the north and east.

The characteristic of the MARS model allows it to create models which initially over fit. However, these are later pruned using either a forwards or backwards stepwise cross validation to remove basis functions which are not required for the final model. The splitting rules are based on continuous smooth functions (Freidman et al, 1991). Furthermore, the MARS model could be creating relationships between variables which do not take place (e.g. calcareous soils on non-calcareous bedrock) or explain the changes or occurrences of specific soil properties in certain areas (e.g. Silt at 30-60 cm in Figure 4.4a and 4.4b and associated soil maps in Figure 4.7a and 4.7b).

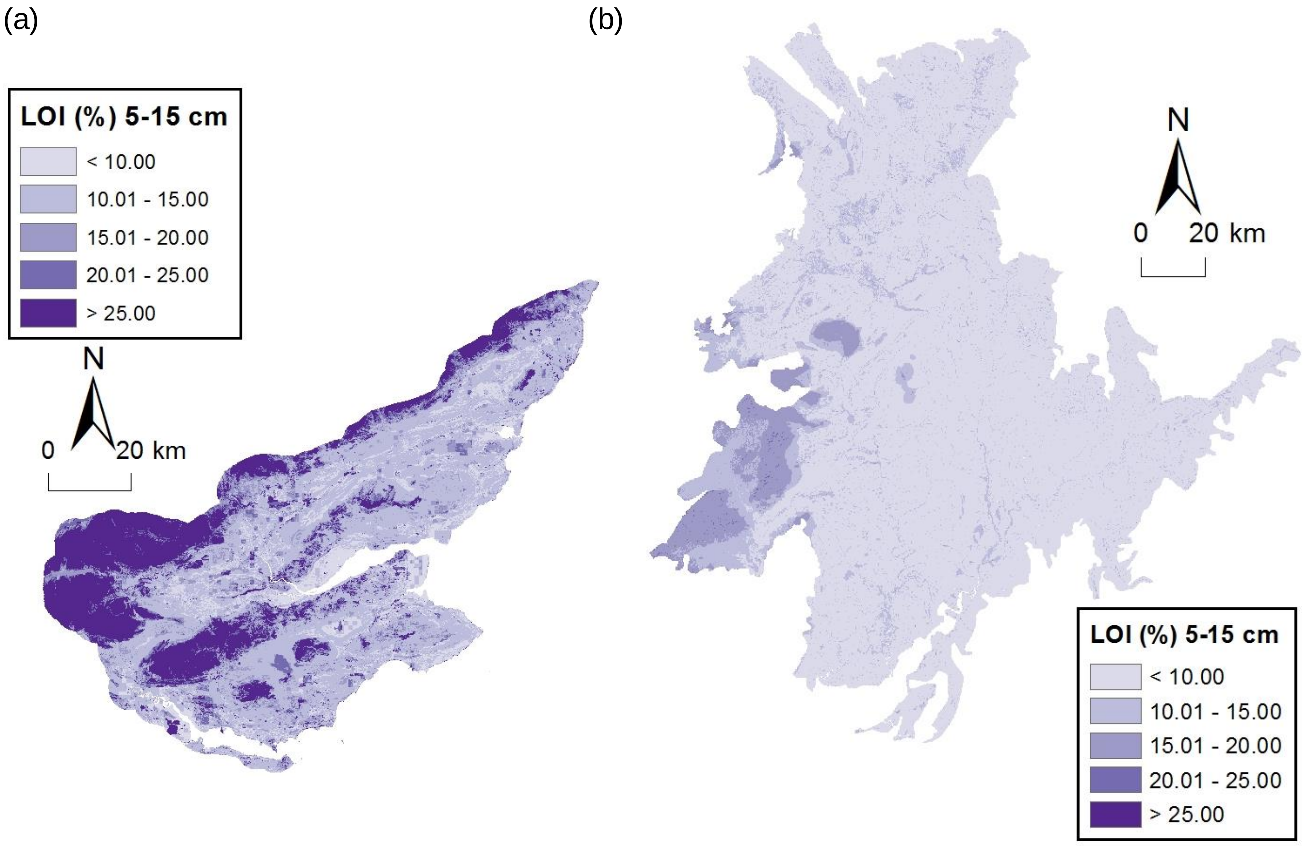


Figure 4.3: Mapped outputs from BRT models for LOI at 5-15 cm for pilot areas: a) SCO and b) EW.

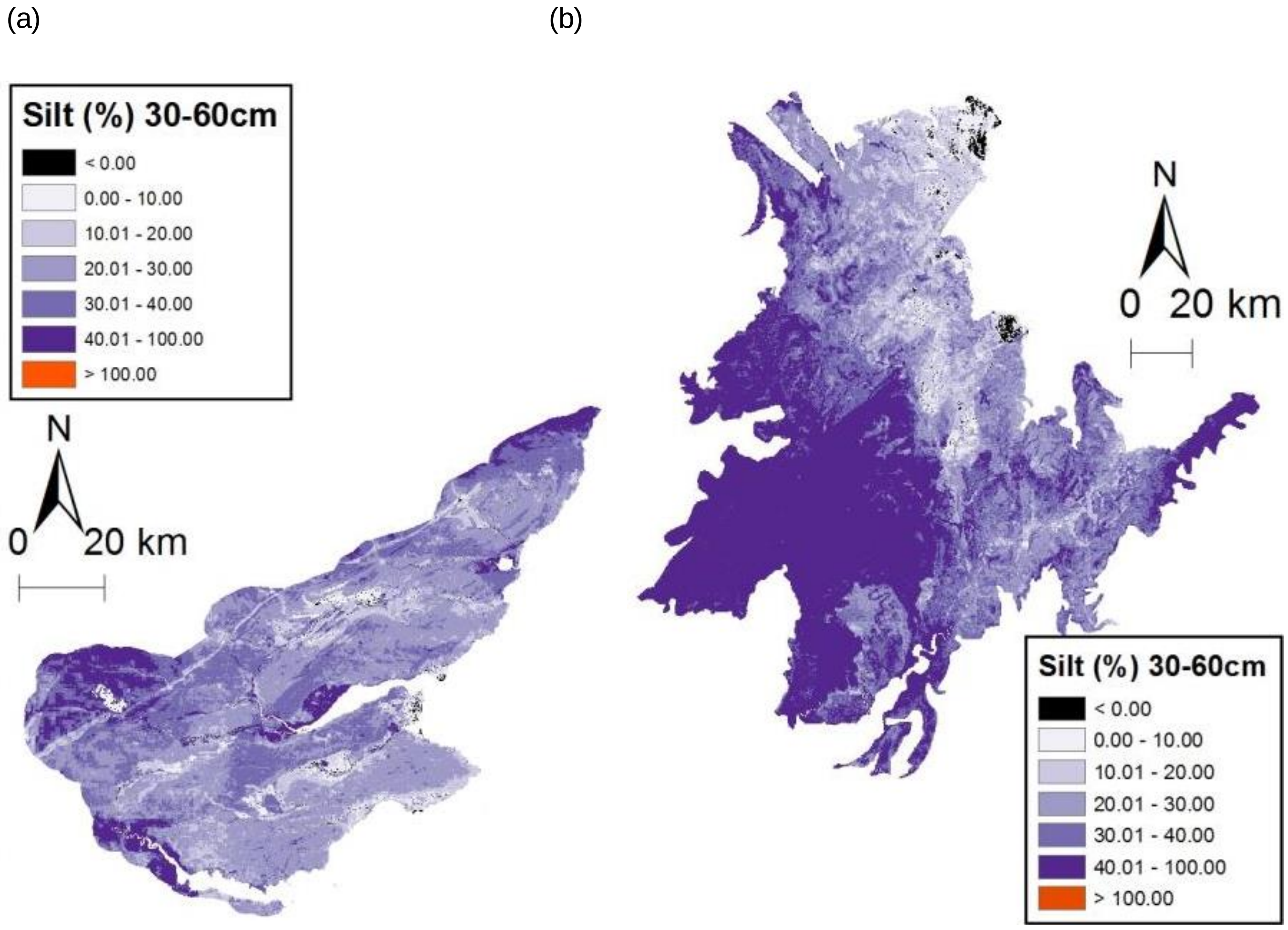


Figure 4.4: Mapped outputs from MARS models for silt at 30-60 cm for pilot areas a) SCO and b) EW.

MARS models are particularly suited to conditions which require many variables, particularly where there is non-linearity, multi-collinearity and a high degree of interaction amongst predictors (Hastie et al, 2009). However, MARS models have a susceptibility to overfitting (even with the pruning) and has issues in dealing with no/missing data. This has been illustrated with the LOI data particularly in the EW pilot area (Figure 4.3b). This generally is why MARS models work well in a training context because there is a tendency to over-fit. As a result, the statistical indices are improved as MARS models are usually based on statistical relationships between the covariates and the soil property being measured. This over-fitting may develop relationships which are not sensible from a pedological basis. For example, in some cases, there are strong dependencies on covariates leading to unusual spatial structures in deployment maps using MARS models, notably for LOI and silt (Figures 4.3a, 4.3b, 4.4a and 4.4b). These are strongly based on local relationships which have been trained effectively but, when deployed, can become unstable. This is undoubtedly a major reason for the unusual structures in some of the maps illustrated in this experiment.

The extrapolation procedures for MARS models produce extreme values (very high or low values) in many of the properties, particularly for LOI (Figure 4.3a and 4.3b) and particle size (e.g. silt and clay) (Figure 4.4a, 4.4b and Figure 4.5a and 4.5b). Therefore, because of the outcomes of this research, MARS models are not suitable for modelling and mapping the soil properties in this instance. There are also striped artefacts in some of the mapping outputs which has been influenced as a result of MARS modelling e.g. Figure 4.5b. This is largely as a result of coarser climatic data (precipitation and temperature) which has been resampled at much finer scale.

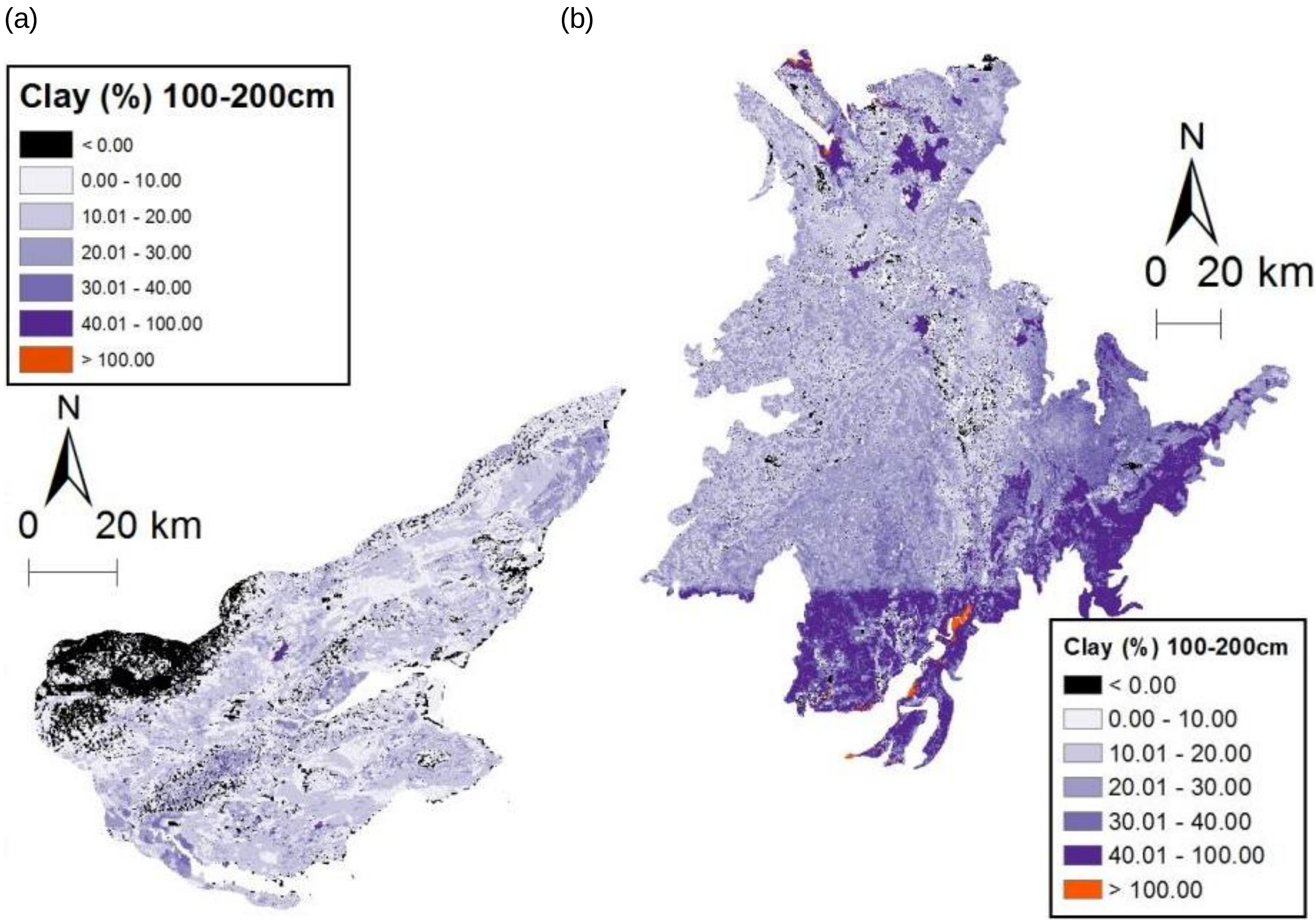


Figure 4.5: Mapped outputs from MARS models for clay at 100-200 cm for pilot areas a) SCO and b) EW.

In contrast, the BRT model represents the effect of each predictor after accounting for effects by other predictors. The related input data is weighted in subsequent trees. These are applied where data which has been poorly modelled by previous trees has a higher probability of being selected in the new tree. This sequential method is the boosted part of the regression tree meaning that new data has a chance of being included in a new tree (Breiman et al, 1984; Bauer and Kohavi, 1999). This differs from other models such as random forests as the data has an equal probability of being selected for the next tree. Splitting rules are based on binary splits on successive predictor variables.

BRTs introduce stochasticity in the boosting which improves the predictive performance and reduces overfitting the data. This effect has translated to a better performance in the mapping of the deployment data because regressing to the mean favours the predictive space where much of the data are consistent with one another (i.e. there are non-extreme values being picked up for all soil properties investigated). However, this does not necessarily capture the extreme values as well as one might anticipate. For instance, when investigating the BRT predicted range for pH (e.g. at 15-30 cm, Figure 4.6a and 4.6b) it can be seen that for both the SCO and EW pilot area there is a narrow range varying from acidic peat soils (around pH value of 4.00) to more alkaline soils (pH values above 6.50 and 7.00) (see Appendix 3a and 3b). This means that the full range of values for soil properties may not be being picked up in these areas based on information from the soil maps (Figure 4.7a and 4.7b). Going forward, partioning of these distinct soil environments or modelling them separately may improve the predictive performance.

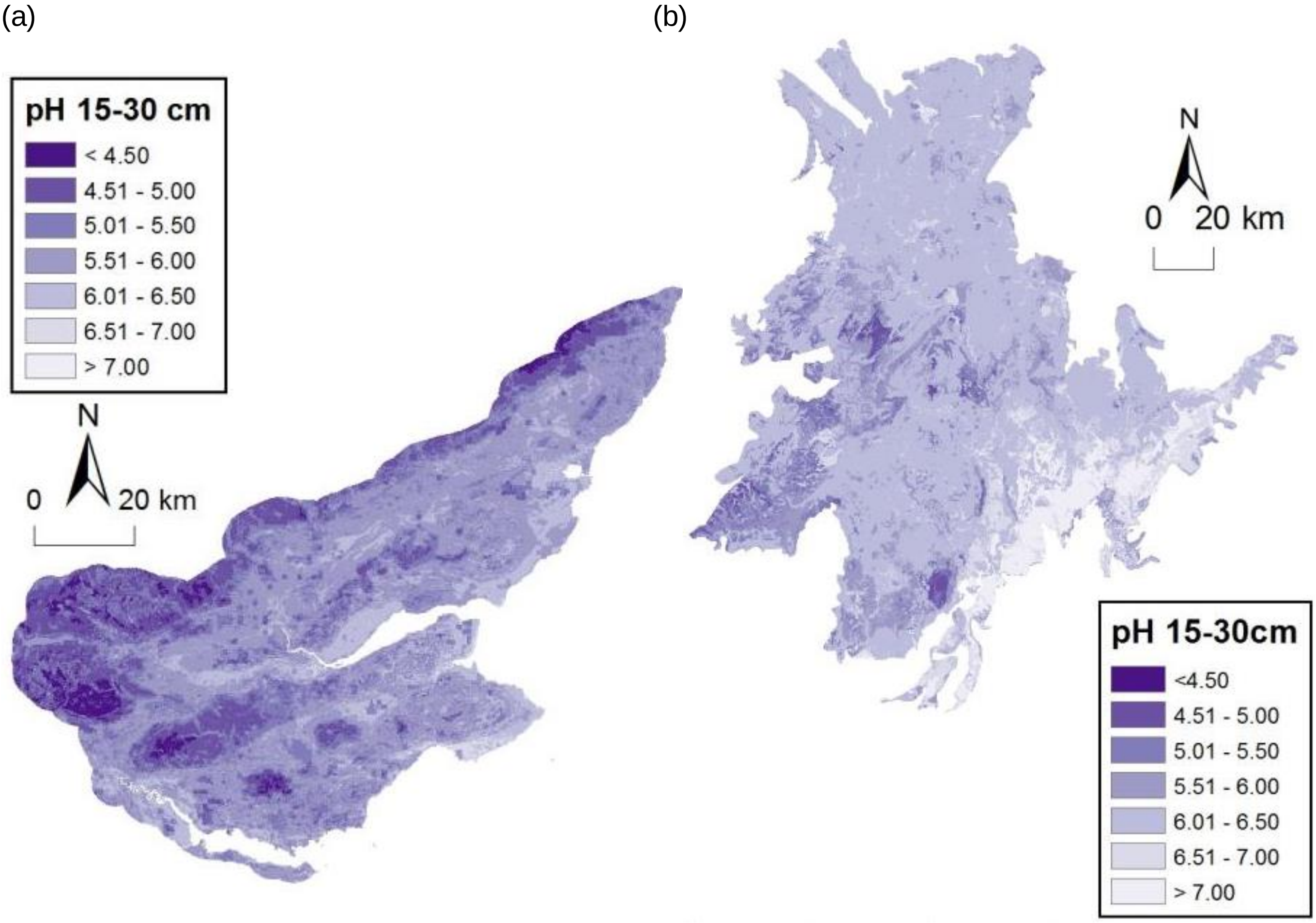


Figure 4.6: Mapped outputs from BRT models for pH at 15-30 cm in pilot areas: a) SCO and b) EW

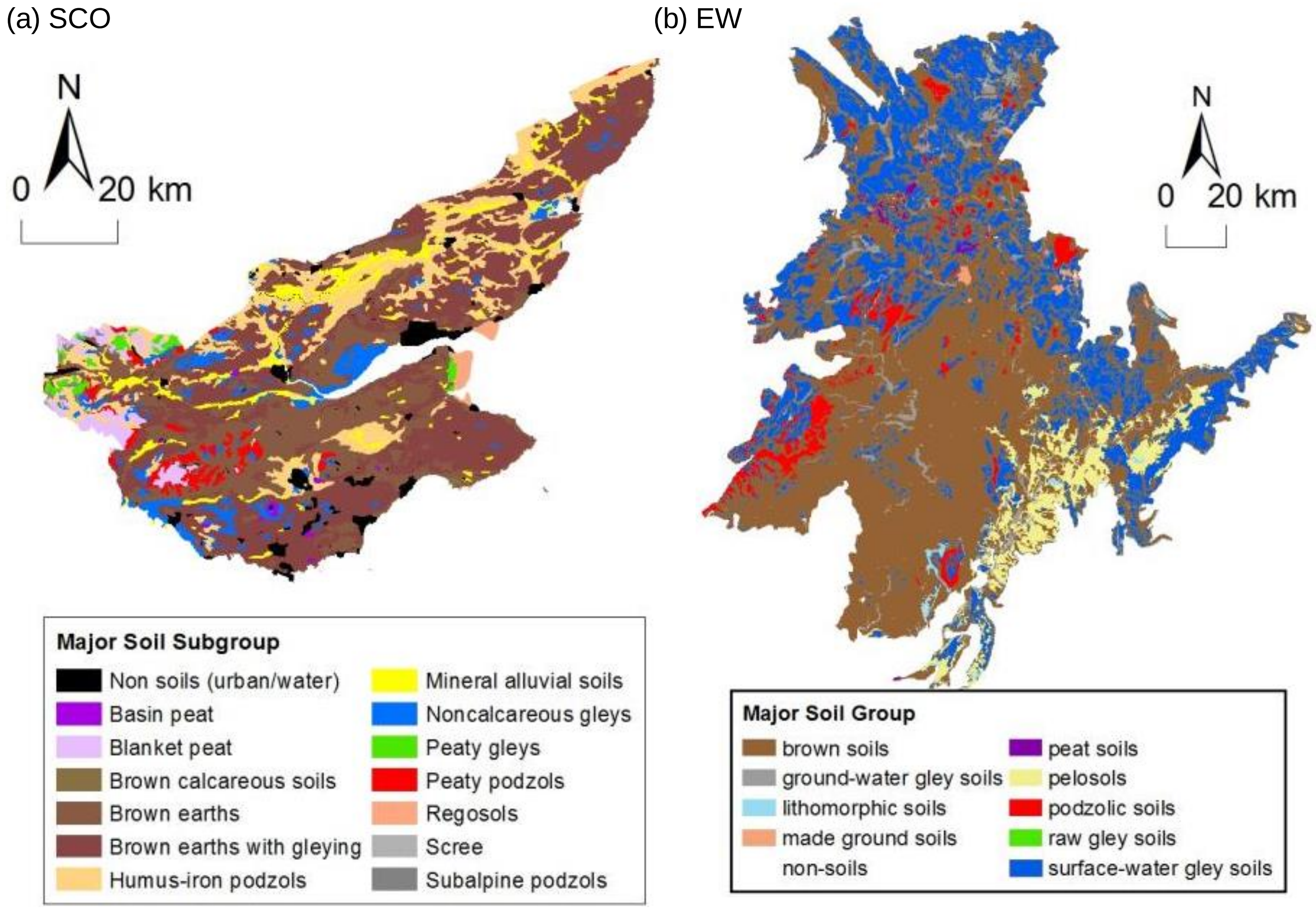


Figure 4.7: Major soil sub group map for the SCO pilot area a) and Major soil group map for the EW pilot area (b).

There is an assumption that the soil-landscape characteristics which have been found in these areas from both BRT and MARS models will be similar elsewhere. Another issue is indeed whether the correct model is used for these areas. For this study BRTs and MARS models have been assessed, however, there are many more models in the literature that could have considered. The main reason BRTs and MARS models were chosen is that it was thought that models would be good not only at training but also being applied more widely. Going forward, it will be crucial to make sure that whichever model(s) are chosen for predicting soil properties is predicting across all the feature space.

Future work will be needed in terms of masking out particular soils and features (water bodies, urban/unsurveyed areas) that are problematic for modelling or which simply do not exist beyond certain depths (e.g. ranker soils). To gain an increased understanding of how BRT models' model and map soil properties, it would be useful to scale up to the whole of GB to see if similar outcomes from this study are consistent at all depths on a much wider scale spatially and at depth.

4.7. Conclusions

This paper has explored the comparison between two recursive partitioning modelling methods; BRTs and MARS, on two pilot areas as part of the DSM development for future wider scale GB mapping. The results suggest that MARS models produce better model performances than BRTs for training and predicting soil properties. However, when predictions from MARS models are deployed, they extrapolate beyond the appropriate range of values of the property which is being predicted. This is because MARS models are suited to situations where many variables are required. Furthermore, MARS models struggle to deal with overfitting and missing data. As has been illustrated in the results and discussion, this over-fitting has led to relationships which do not make sense pedologically.

Conversely, BRT models are seen to be more consistent pedologically in terms of mapping soil properties. This is because these models represent the effect of each predictor after accounting for effects by other predictors. BRTs also introduce randomness in the boosting which improves the predictive performance and thereby reduces overfitting. As a result, BRTs have given a more consistent performance in the mapping deployment outputs because regressing to the mean favours the predictive space where the most data match up with one another. However, this does not necessarily mean the full range of soils in these areas has been captured. Future research will focus on refining the predictions of these soil properties to fully encapsulate the full range of soils across GB.

References

Avery, B.W., Bascomb, C.L., 1982: Soil Survey Technical Monograph No.6. Soil Survey Laboratory Methods. Harpenden.

Bauer, E., Kohavi, R., 1999: An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants, *Machine Learning*, 36, 1-2, pp. 105-139.

Beaudette, D.E., Roudier, P., O'Geen, A.T., 2013: Algorithms for Quantitative Pedology: A Toolkit for Soil Scientists. *Computers & Geosciences*. 52, pp. 258 - 268.

Behrens, T., Förster, H., Scholten, T., Steinrücken, U., Spies, E.D., Goldschmidt, M., 2005: Digital Soil Mapping using artificial neural networks. *Journal of Plant Nutrition and Soil Science*, 168, pp.1-13.

Behrens, T., Scholten, T., 2006: Digital soil mapping in Germany—a review. *Journal of Plant Nutrition and Soil Science*, 169, 3, pp. 434-443.

BGS., 2018: Digital Geological Map of Bedrock (DigMap250): [Available from http://www.bgs.ac.uk/products/digitalmaps/digmapgb_250.html] [Last Accessed 29th May 2018].

Bishop, T.F.A., McBratney, A.B., Whelan, B.M., 2001: Measuring the quality of digital soil maps using information criteria. *Geoderma* 103, 1, pp. 95-111.

Bivand, R., Keitt, T., Rowlingson, B., Pebesma, E., Sumner, M., Hijmans, R., Rouault, E., Warmerdam, F., Ooms, J., Rundel, C., 2015: rgdal: Bindings for the Geospatial data. http://CRAN.R-project.org/package=rgdal .

Breiman, L., 1984: Classification and Regression Trees Chapman &Hall/CRC, New York: Routledge.

Breiman, L., 2001: Machine Learning, 45,5. https://doi.org/10.1023/A:1010933404324

Campbell, G.A., Lilly, A., Corstanje, R., Mayr, T.R., Black, H.I.J., 2017: Are existing soils data meeting the needs of stakeholders in Europe? An analysis of practice use from policy to field. *Land Use Policy*, 68, pp.211-223.

Carré, F., McBratney, A.B., Mayr, T., Montanarella, L., 2007: Digital soil assessments: Beyond DSM. *Geoderma*, 142,1–2, pp. 69-79.

Cavazzi, S., Corstanje, R., Mayr, T., Hannam, J., Fealy, R., 2013: Are fine resolution digital elevation models always the best choice in digital soil mapping? *Geoderma*, 195–196, pp. 111-121.

Chapman, S.J., Bell, J.S., Campbell, C.D., Hudson, G., Lilly, A., Nolan, A.J., Robertson, A.H.J., Potts, J.M., Towers, W., 2013: Comparison of soil carbon stocks in Scottish soils between 1978 and 2009. *European Journal of Soil Science*, 64, pp. 455-465.

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., Böhner., 2015: System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. Geoscientific Model Development, 8, pp.1991-2007, doi: 10.5194/gmd-8-1991-2015. https://www.geosci-model-dev.net/8/1991/2015/gmd-8-1991-2015.html [Last Accessed 24th August 2018].

Corstanje, R., Kirk, G.J.D., Pawlett, M., Read, R., Lark, R.M., 2008: Spatial variation of ammonia volatilization from soil and its scale-dependent correlation with soil properties, *European Journal of Soil Science*, 59, pp. 1260-1270.

Dobos, E., Carré, F., Hengl, T., Reuter, H.I., Tóth, G., 2006: Digital Soil Mapping as a support to production of functional maps. EUR 22123 EN, Office for Official Publications of the European Community, Luxemburg.

Elith, J., Leathwick, J.R., Hastie, T., 2008: A working guide to boosted regression trees. *Journal of Animal Ecology*, 77: pp. 802–813.

Friedman, J.H., 1991: Multivariate Adaptive Regression Splines, the Annals of Statistics, Vol. 19, No. 1 (Mar. 1991), pp. 1-67.

Friedman, J.H., 2002: Stochastic gradient boosting, *Computational Statistics & Data Analysis - Nonlinear methods and data mining*, 38, 4, pp. 367-378.

Fuller, I., Smith, G.M., Sanderson, J.M., Hill, R.A., Thomson, A.G., Cox, R., Brown, N.J., Clarke, R.T., Rothery, P., Gerard, F.F., 2000: Land Cover Map 2000: a guide to the classification system, Centre for Ecology and Hydrology.

Ghasemi, J.B., Zolfonoun, E., 2013: Application of principal component analysis – multivariate adaptive regression splines for the simultaneous spectrofluorometric determination of dialkyltins in micellar media. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 115, pp. 357-363.

GlobalSoilMap., 2011. GlobalSoilMap.net: New Digital Soil Map of the world. [Accessed from https://www.globalsoilmap.net/] [Last accessed 17th May 2018].

Grunwald, S., 2009: Multi-criteria characterization of recent digital soil mapping and modelling approaches. *Geoderma*, 152, pp. 195-207.

Hallett, S.H., Sakrabani, R., Keay, C.A., Hannam, J.A., 2017: Developments in land information systems: examples demonstrating land resource management capabilities and options. *Soil Use and Management.* 33, pp. 514-529.

Hastie, T., Tibshirani, R., Friedman, J., 2009: The elements of statistical learning: Data Mining, Inference and Prediction, 2nd Edition, Springer Series in Statistics.

Hijmans, R.J., Phillips, S., Leathwick. J., Elith, J., 2011: Package ‘dismo’. Available online at: http://cran.r-project.org/web/packages/dismo/index.html.

Hijmans, R.J. and Etten, J.v., 2013: raster: geographic data analysis and modelling. R Package version 2.1-25. -< CRAN.R-project.org/package=raster>.

Hrvatin, M., Perko, D., 2009: Suitability of Hammond's method for determining landform units in Slovenia. *Acta geographica Slovenica*, Accessed from: http://ojs.zrc sazu.si/ags/article/view/1281. Date accessed: 24 Aug. 2016. doi: http://dx.doi.org/10.3986/AGS49204.

Hudson, B.D., 1992: Division S-5- Soil Genesis, Morphology & Classification: The soil survey as paradigm-based science. *Soil Science Society of America Journal* 56, 3, pp. 836-841.

Kuhn, M., Weston, S., Keefer, C., Coulter, N., 2012: Cubist models for regression. The comprehensive R Archive Network. http://cran.r-project.org/web/packages/Cubist/vignettes/cubist.pdf [Accessed 19th September].

Leempoel, K., Parisod, C., Geiser, C., Daprà, L., Vittoz, P., Joost, S., 2015: Very high-resolution digital elevation models: are multi-scale derived variables ecologically relevant? *Methods in Ecology and Evolution*, 6, pp. 1373-1383.

Lilly, A., Bell, J.S., Hudson, G., Nolan, A.J., Towers, W., 2010: National Soil Inventory of Scotland (NSIS_1): site location, sampling and profile description protocols (1978-1988): Technical Bulletin, Macaulay Institute, Aberdeen.

Macaulay Institute for Soil Research., 1979: Laboratory notes on Methods of Soil Analysis.

Malone, B.P., McBratney, A.B., Minasny, B., Laslett, G.M., 2009: Mapping continuous depth functions of soil carbon storage and available water capacity. *Geoderma*, 189-190, pp. 153-163.

McBratney, A. B., Mendonça Santos, M.L., Minasny, B., 2003: On digital soil mapping. *Geoderma*, 117,1, pp. 3-52.

Millborrow, S., 2017: earth: Multivariate Adaptive Regression Spline Models; R package, http://CRAN.R-project.org/package=earth.

Minasny, B., McBratney, A.B., 2016: Digital soil mapping: A brief history and some lessons, *Geoderma*, 264, pp. 301-311.

Nawar, S., Buddenbaum, H., Hill, J., 2015: Digital Mapping of Soil Properties using Multivariate Statistical Analysis and ASTER data in an arid region. *Remote Sensing,* 7, 2, pp. 1181-1205.

Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., Schaepman, M.E., Papritz, A., 2018: Evaluation of digital soil mapping approaches with larger sets of environmental covariates. *SOIL*, 4, pp. 1-22.

OS Terrain 50., 2018: OS Terrain 50. [Available from https://www.ordnancesurvey.co.uk/business-and-government/products/terrain-50.html] [Last Accessed 29th May 2018].

Pain, C.F., 2005: Size does matter: relationships between image pixel size and landscape process scales, MODSIM 2005 International Congress on Modelling and Simulation (2005), pp. 1430-1436.

Pebesma, E.J., Bivand, R.S., 2005: Classes and methods for spatial data in R. R News 5 (2), https://cran.r-project.org/doc/Rnews/.

Prasad, N., Iverson, L.R., Liaw, A., 2006: Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. *Ecosystems,* 9, pp.181-199.

R Core Team., 2013: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Ridgeway, G., 2009. Package ‘gbm’. R-News. 08:05:15. http://www.saedsayad.com/docs/gbm2.pdf.

SNH., 2002: Natural Heritage Zones: A national assessment of Scotland’s landscapes.

Stevenson, M., 2015: Tools for the analysis of epidemiological data (Version 0.9-62) [Software]. Retrieved from http://cran.r-project.org/web/packages/epiR/epiR.pdf.

Sutton, C.D., 2005: Classification and Regression Trees, Bagging and Boosting. In “Handbook of Statistics, Vol. 24”, pp. 303-329, ISSN:0169-7161.

Thompson, A.J., 2001: Digital elevation model resolution: effects on terrain attribute calculation and quantitative soil-landscape modelling, *Geoderma*, 100, pp. 67-89.

Veronesi, F., Corstanje, R., Mayr, T., 2014: Landscape scale estimation of soil carbon stock using 3D modelling. *Science of the Total Environment,* 487, pp. 578–588.

Viscarra Rossel, R.A., Behrens, T., 2010: Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma*,158, 1-2, pp.46-54.

Wickham, H., 2016: The Split-Apply-Combine Strategy for Data Analysis. J*ournal of Statistical Software*, 40, 1, 1-29. http://www.jstatsoft.org/v40/i01/.

World Climate., 2016: WorldClim – Global Climate data: free climate data for ecological modelling and GIS. [Available from http://www.worldclim.org/] [Last Accessed 29th May 2018].

5 MODELLING SOIL PROPERTIES FOR GREAT BRITAIN USING BOOSTED REGRESSION TREES

Abstract

Improving the resolution of soil property maps is a major requirement requested by stakeholders. However, communicating this information is a challenge due to outdated methods and inconsistencies with previous techniques. Digital Soil Mapping (DSM) has been promoted as a useful approach to address these issues. DSM is usually achieved by evaluating model statistics. However, it is also important to critically evaluate the mapping outputs to investigate whether these soil properties are being modelled effectively and whether they reflect pedological understanding.

This chapter predicts and maps loss-on-ignition (LOI), soil pH and texture across Great Britain (GB) using Boosted Regression Tree (BRT) models at 100 m resolution for specified depth intervals. Results reflect that BRT models work well across GB for predicting soil pH and LOI but perform poorly for texture properties.

This chapter also examines whether an independent validation dataset is useful in evaluating soil property predictions in comparison to modelled outputs. Results show inconsistencies across both training and independent validation datasets. Thus, future work should focus on intensively mapping or collecting more data from areas where there is sparse information to help produce improved maps and reduce the associated uncertainty.

5.1. Introduction

There is a growing demand to produce improved soil property maps for Great Britain (GB). At present, there are only a few unified datasets that cover GB (Reed, 2008; Campbell et al, 2017). Stakeholders have stated that they require improved soil information at finer resolution with associated metadata (Campbell et al, 2017; Sanchez et al, 2009). Despite the many efforts to generate new global or regional soil property maps, these are currently too coarse for many stakeholders, such as land managers, to make informed decisions at field or catchment scales (Grunwald et al, 2011; Sanchez et al, 2009; Reed, 2008, Hengl et al, 2017). However, soil property information is sometimes not available due to a lack of resources or foci of historical mapping programmes. As a result, alternative approaches, such as Digital Soil Mapping (DSM) have been developed to help improve the resolution and provide the new data (McBratney et al, 2003; Scull et al, 2003).

The focus of DSM is to utilise available soils data by developing statistical models alongside associated environmental covariates to produce new spatial predictions of soil properties (Dobos et al, 2006; Behrens and Scholten, 2006). DSM helps deal with the increased cost of traditional soil mapping and declining number of soil surveyors (McBratney et al, 2003; Hudson, 1992).

DSM has many useful advantages particularly in improving the quantitative understanding of soil variability (Scull et al, 2003; McBratney et al, 2003; Carré et al, 2007) by taking account of associated uncertainties in predicted soil properties that cannot be achieved through traditional soil mapping (McBratney et al, 2003; Carré et al, 2007; Hudson, 1992, Minasny and McBratney, 2016). There have been many examples of new soil property maps created by DSM at a range of scales (e.g. Hengl et al, 2017; Poggio and Gimona 2017; Minasny and McBratney, 2010; Minasny et al, 2010; Malone et al, 2009; Adhikari et al, 2014; Ballabio et al, 2014). However, many of these studies fail to address a critical component of DSM, which is to evaluate the outputs from the DSM models against what is expected based on expert pedological understanding.

In DSM there are generally three evaluation methods: i) model performance evaluated by statistical indices using cross-validation and independent validation (Piikki and Söderström, in press; Brus et al, 2013) ii) comparison of the statistics with other models and resulting maps (Nussbaum et al, 2018) and iii) how well the mapped outputs reflect expert knowledge of the spatial distribution of soil properties by pedologists and soil surveyors (Hudson, 1992; Reuter et al, 2008). The latter of these evaluation approaches has only been explored sporadically in DSM studies with few having used an observed, independent validation dataset to evaluate model outputs.

In this chapter, DSM predictions of soil properties at a national scale will be evaluated statistically against an independent validation dataset and in relation to expert pedological understanding. These soil properties – loss-on-ignition (as a proxy for soil organic carbon), pH and texture – were chosen based on a survey of stakeholder needs (Campbell et al, 2017) and to comply with Global Soil Map specifications (GlobalSoilMap, 2011).

5.2. Materials and Methods

5.2.1. GB soils

GB comprises of Scotland, England and Wales. Scotland has a large variety of soils which are mainly organic (Histosols), waterlogged (Stagnosols) or leached (Podzols) in nature (Figure 5.1). In Northern England, the main soil types are Stagnosols and Cambisols, with Histosols being found in upland areas. Central, eastern and south eastern England are characterised by Luvisols, Cambisols, Stagnosols and Leptosols. South west England and south west Wales are dominated by Cambisols. Podzols and Histosols are common in upland parts of Wales (Figure 5.1).

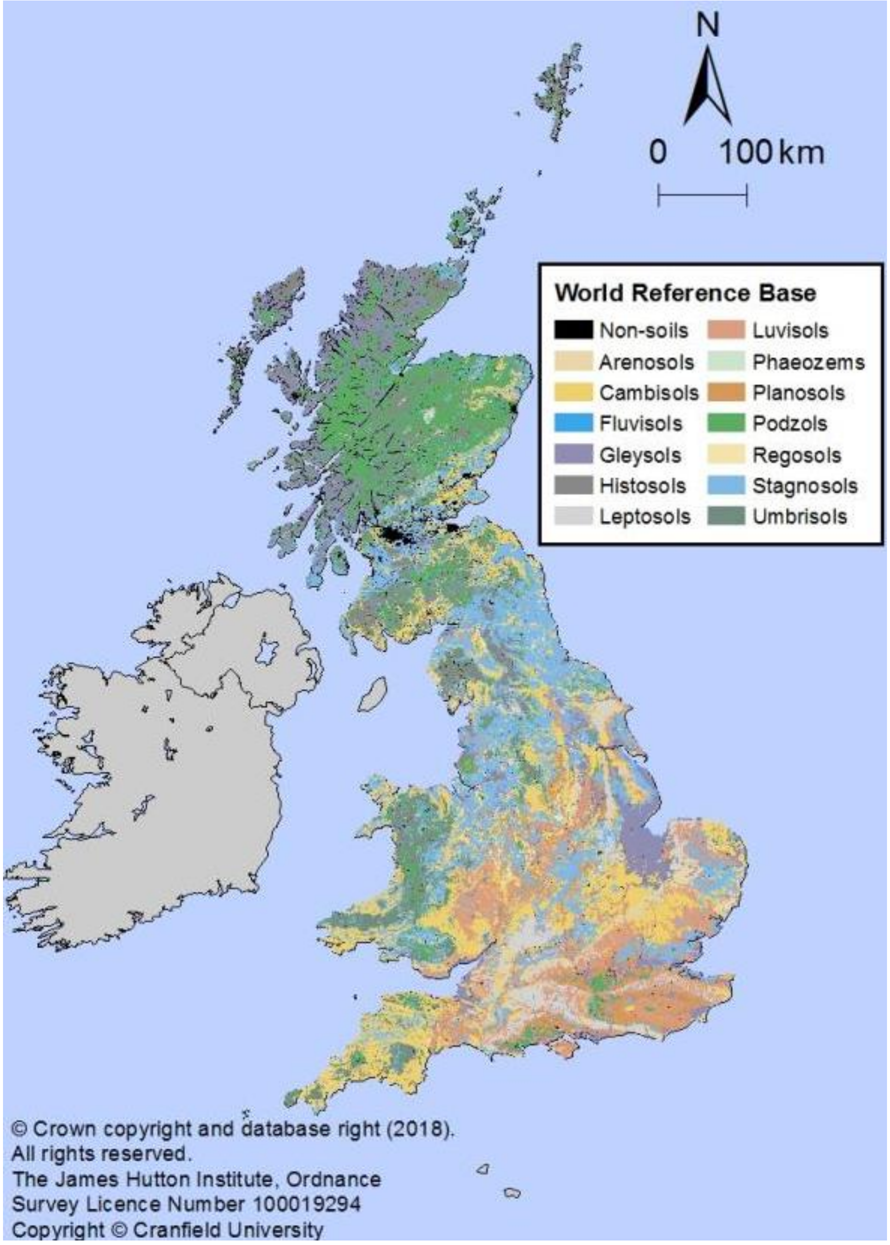


Figure 5.1: Map of the WRB (2006) Soil Reference Groups in GB.

5.3. Description of the datasets

For this study, the '*training*' dataset used to develop the models for the predicted soil properties were the representative soil profiles collected by the respective National Soil Survey organisations and now held by Cranfield University and The James Hutton Institute. The independent '*validation*' dataset comprises a combination of National Soil Inventory (NSI) and National Soil Inventory for Scotland (NSIS) samples.

5.3.1. Training Data

The soil property data used in this study were derived from analyses of soil horizon samples from representative profiles held in the Scottish Soils Database and in the LandIS database for England and Wales (Hallett et al, 2017). These were collected during national soil survey mapping programmes and the profiles were selected by the soil surveyors as they typified soil series. In England and Wales sampling took place primarily between the 1970s and 1980s while in Scotland sampling mainly took place between the 1960s and late 1980s. In addition, the National Soil Inventory of Scotland (NSIS 1978-88) dataset (Lilly et al, 2010) was also used, where there were few representative profiles (for example in the north and west). Those Inventory profiles that were subsequently resampled in the 2007-09 resampling programme in Scotland were excluded. In total, around 11,000 soil profiles have been sampled in Scotland giving over 54,000 individual samples. In England and Wales, samples were collected from over 11,000 soil profiles. For this study, between 12,361 and 14,156 profiles were selected from across GB. This is fewer than the total number of profiles available as not all profiles in the respective databases had a full complement of the soil properties being modelled over all depths.

5.3.2. Validation Data

An Independent validation dataset was obtained from regular grid sampling programmes. For England and Wales, the National Soil Inventory (NSI) topsoil (0-15 cm) dataset was used as a validation dataset. The NSI comprised 5662 sites sampled at 5 km Ordnance Survey grid intersections. For Scotland, data was obtained from the NSIS 2007-9 dataset, which was a partial resampling of original NSIS 1978-88 and took place between 2007 and 2009 (Lilly et al, 2011). This dataset comprised 183 sites on a 20 km grid and included samples taken from 0-15 cm depths using a soil auger. These were collected based on the same protocols as the NSI (Lilly et al, 2011). To create a unified GB soil validation dataset of topsoil properties, data from the NSIS 2007-9 auger samples (0-15 cm) were combined with the NSI data from the 0-

15 cm samples, giving a total of between 4985 and 5769 validation points across GB depending on the soil property.

5.3.3. Environmental Covariates

Based on previous DSM development work (Campbell et al, in prep), 8 covariates were used to represent the SCORPAN factors (Table 5.1). The 'S' soil component was taken as the dominant major soil subgroup in Scotland and the major soil group in England and Wales, as shown on the national 1:250,000 scale soil maps. Each dataset has different original scales which are shown in Table 5.1. and all were resampled onto a 100 m raster grid.

Dominant Major Soil Subgroup (Scotland) (1:250,000)
Major soil group (England and Wales) (1:250,000)
Land Cover Map 2000 (1km grid)
Topographic Wetness Index (based on 50m DTM)
Cross Sectional Curvature (based on 50m DTM),
Valley Depth to Channel Network (based on 50m DTM)
Convergence Index (based on 50m DTM)
Bedrock Geology (1:250,000)
Soil and Terrain dataset (SOTER) (based on 50m DTM)

Table 5.1: Covariates used for modelling of soil properties for GB representing different SCORPAN factors.

5.4. Mapping and modelling Methodology

Depth functions were fitted to the representative soil profile training datasets for each soil property (LOI, pH, sand, silt and clay) using an equal area mass preserving spline at six standard soil depth intervals (0-5, 5-15, 15-30, 30-60, 60-100, 100-200 cm) produced from GlobalSoilMap.net (GSM) specifications (Bishop et al, (2001); Malone et al, (2009); GlobalSoilMap, (2011); Hartemink et al, 2010)). To produce these, R statistics package and principally 'GSIF' (Hengl et al, 2017), 'aqp' (Beaudette and Roudier, 2018), 'plyr' (Wickham,

2016) and ‘sp’ (Pebesma et al, 2018) packages (R Core Team, 2013) were used. These datasets were imported into ArcMap 10.2.1 and point shapefiles were created for each soil property at the six standard GSM soil depth intervals across GB. Covariates are converted to rasters at 100 m resolution and sampled to provide a set of values for each location of a sampled soil profile in ArcMap 10.2.1. These datasets were saved as text files and imported into R.

After data cleaning (e.g. removing any null values and soils that were sampled within 100 m of each other), the representative profile training data was used to develop boosted regression tree (BRT) models to identify relationships between the soil properties and covariates. For the BRT models, ‘dismo’ (Hijmans, 2017) and ‘gbm’ (Ridgeway, 2017) packages were used (R Core Team, 2013). A ten-fold cross validation was used to evaluate the model performance and statistics (R^2, RMSE, mean, standard deviation, observed and predicted range) for each soil property were calculated. After developing the model using the training datasets, the model was applied to the covariate deployment dataset to predict soil properties at the specified depths on a 100 m grid across GB. This was created using R packages ‘rgdal’ (Bivand, 2018), ‘raster’ (Hijmans, 2016) and ‘sp’ (Pebesma et al, 2018) and produced predicted soil property values at the specified depths across the whole of GB. The resultant soil property maps were generated at 100 m resolution, using ‘raster’ (Hijmans, 2016) package, saved as a TIFF file and then imported into ArcMap 10.2.1.

The ‘raw’ predicted map was post-processed by masking urban and unsurveyed areas, rock dominated areas, major coastal features, man-made or disturbed soils, raw gley soils, lakes, rivers and other water bodies. Rankers (or Lithomorphic soils) were predicted to a maximum depth of 40 cm as this is the typical maximum depth observed for these soils.

For model evaluation using an independent dataset, R^2 and RMSE values were produced by correlating the predicted outputs produced from the BRT model and observed values from the independent validation dataset. This independent validation dataset was only available for one depth (topsoil at 0-15 cm) for GB. Residuals were calculated from the cross- validation and the independent datasets and mapped outputs were used to compare where soil properties are being over- or under-predicted across GB. The R code used for this can be adapted from similar work achieved in Chapter 4 (Appendix 7).

5.5. Results

5.5.1. BRT model performance

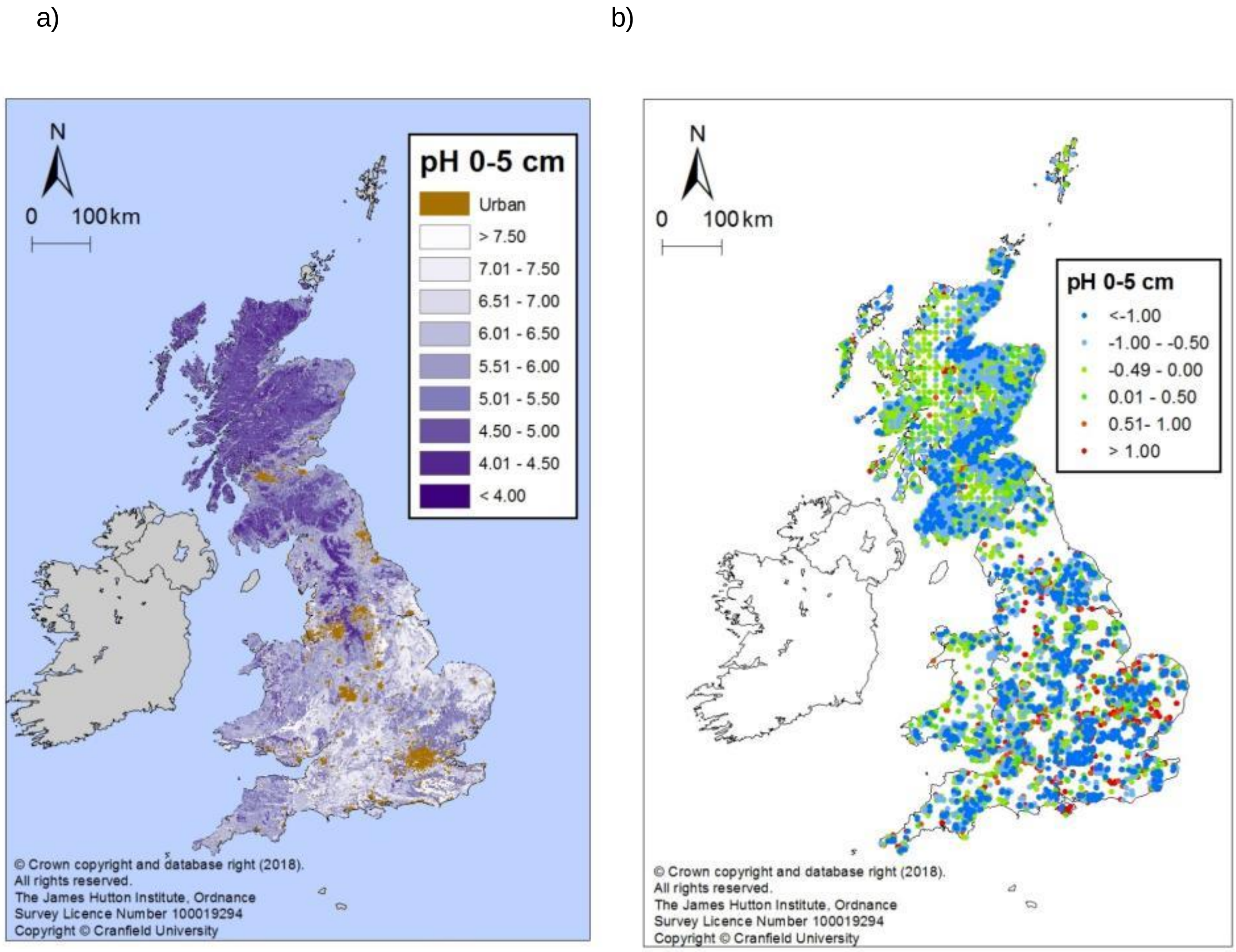
Table 5.2 presents BRT model performance statistics for each soil property at depth. The model performance for pH shows R^2 values from 0.56 at 0-5 cm to 0.31 at 60-100 cm. The RMSE values for pH range from 0.81 at 0-5 cm to 0.99 at 60-100 cm. The predicted BRT model ranges compare well overall to the observed data; however, the BRT model does not predict the pH of very acidic soils well (notably those with a pH < 3.72) or the pH of calcareous soils with pH > 8.33 (at 100-200 cm depth). Table 5.2 indicates observed soil pH values ranging from 2.00 to 9.59. There are very few examples of these samples at extreme low or high pH meaning that these values could represent specific location factors such as waste sites or methodological errors. For LOI, the model produces lower R^2 and higher RMSE values than for pH (ranging from 0.51 at 0-5 cm depth to 0.31 at 60-100 cm with RMSE values range from 19.42 at 0-5 cm to 11.30 at 100-200 cm depth). The BRT model predicts LOI well for the first four depths. However, beyond 60 cm depth the model fails to predict over the full observed range. The model performance of predicted sand content ranges from R^2 values of 0.53 at 0-5 cm depth to 0.35 at 100-200 cm. RMSE values range from 15.61 at 0-5 cm to 22.68 at 100-200 cm. Predictions of silt content vary with R^2 values ranging from 0.45 at 0-5 cm to 0.25 at 100-200 cm depth and RMSE values between 10.97 at 15-30 cm depth and 15.05 at 100-200 cm. The BRT model does not predict silt proportions greater than 70%. When predicting clay content, R^2 values range from 0.50 with a value of 9.11 RMSE at 0-5 cm depth to 0.38 at 100-200 cm with RMSE at 12.98. The BRT model fails to predict clay content greater than 66%.

Soil Property	Depths (cm)	R^2	RMSE	Mean	SD	Observed Range	Predicted Range	Samples (n)
pH	**0-5**	0.56	0.81	5.64	1.22	2.00 – 9.58	3.72 – 7.88	13929
	5-15	0.56	0.79	5.67	1.21	2.19 – 9.56	3.78 – 7.85	13918
	15-30	0.55	0.75	5.80	1.19	3.01 – 9.57	3.82 – 8.04	13780
	30-60	0.55	0.77	5.96	1.18	2.66 – 9.59	4.09 – 8.26	13269
	60-100	0.31	0.99	6.12	1.20	2.23 – 9.58	4.51 – 7.72	11788
	100-200	0.48	0.86	6.18	1.20	2.17 – 9.49	4.54 – 8.33	5332
LOI	**0-5**	0.51	19.42	24.58	27.73	0.28 – 100.00	6.06 – 94.20	12146
	5-15	0.50	16.81	20.10	23.84	0.27 – 100.00	1.98 – 98.39	12093
	15-30	0.43	14.58	13.66	19.23	0.16 – 100.00	5.33 – 90.93	11689
	30-60	0.35	13.61	8.88	16.64	0.09 – 100.00	5.35 – 75.57	10057
	60-100	0.31	13.16	6.77	15.12	0.06 – 100.00	5.34 – 54.01	7650
	100-200	0.32	11.30	4.71	11.33	0.14 – 100.00	3.72- 47.26	3551
Sand	**0-5**	0.53	15.61	48.16	23.20	0.00 – 100.00	6.06 – 100.00	11754
	5-15	0.50	15.96	48.30	23.19	0.00 – 100.00	6.65 – 100.00	11743
	15-30	0.50	16.00	48.97	23.70	0.00 – 100.00	0.30 – 100.00	11625
	30-60	0.48	17.54	50.43	25.36	0.00 – 100.00	7.65 – 100.00	11207
	60-100	0.38	19.28	51.52	27.31	0.00 – 100.00	9.34 – 100.00	9931
	100-200	0.35	22.68	55.52	28.08	0.00 – 100.00	3.38 – 93.38	4619
Silt	**0-5**	0.45	11.43	32.44	15.43	0.00 – 91.94	0.00 – 68.80	11754
	5-15	0.41	11.71	32.31	15.29	0.00 – 90.94	0.00 – 64.74	11743
	15-30	0.39	10.97	31.54	15.27	0.00 – 90.02	0.00 – 70.06	11625
	30-60	0.38	12.17	29.84	15.58	0.00 – 84.62	0.00 – 67.18	11207
	60-100	0.37	13.14	28.63	16.50	0.00 – 100.00	0.00 – 67.18	9931
	100-200	0.25	15.05	26.45	17.07	0.00 – 100.00	9.64 – 55.43	4619
Clay	**0-5**	0.50	9.11	17.17	13.61	0.00 – 94.00	0.00 – 56.76	11754
	5-15	0.50	9.46	17.50	13.92	0.00 – 94.00	0.00 – 58.23	11743
	15-30	0.46	9.57	18.14	14.36	0.00 – 94.00	0.00 – 66.38	11625
	30-60	0.48	10.85	19.07	15.67	0.00 – 94.00	0.28 – 52.43	11207
	60-100	0.48	11.75	19.61	16.54	0.00 – 94.35	1.39 – 54.09	9931
	100-200	0.38	12.98	17.85	16.23	0.00 – 98.95	2.76 – 44.25	4619

Table 5.2: Statistics for all soil properties predicted by BRT modelling for GB.

5.5.2. Predicted maps of soil properties

From the map of pH at 0-5 cm produced from BRT models, it can be noted that predicted acidic pH values of 4.5 or less are found mainly in the uplands of northern and southern Scotland and in north and central England (Figure 5.2a). This distribution is also seen for predicted pH at lower depths down to 30-60 cm. When compared to a map of residuals for pH at the same depth (Figure 5.2b), it is shown to be underpredicted across much of GB, particularly in the eastern Scotland and much of England and Wales.

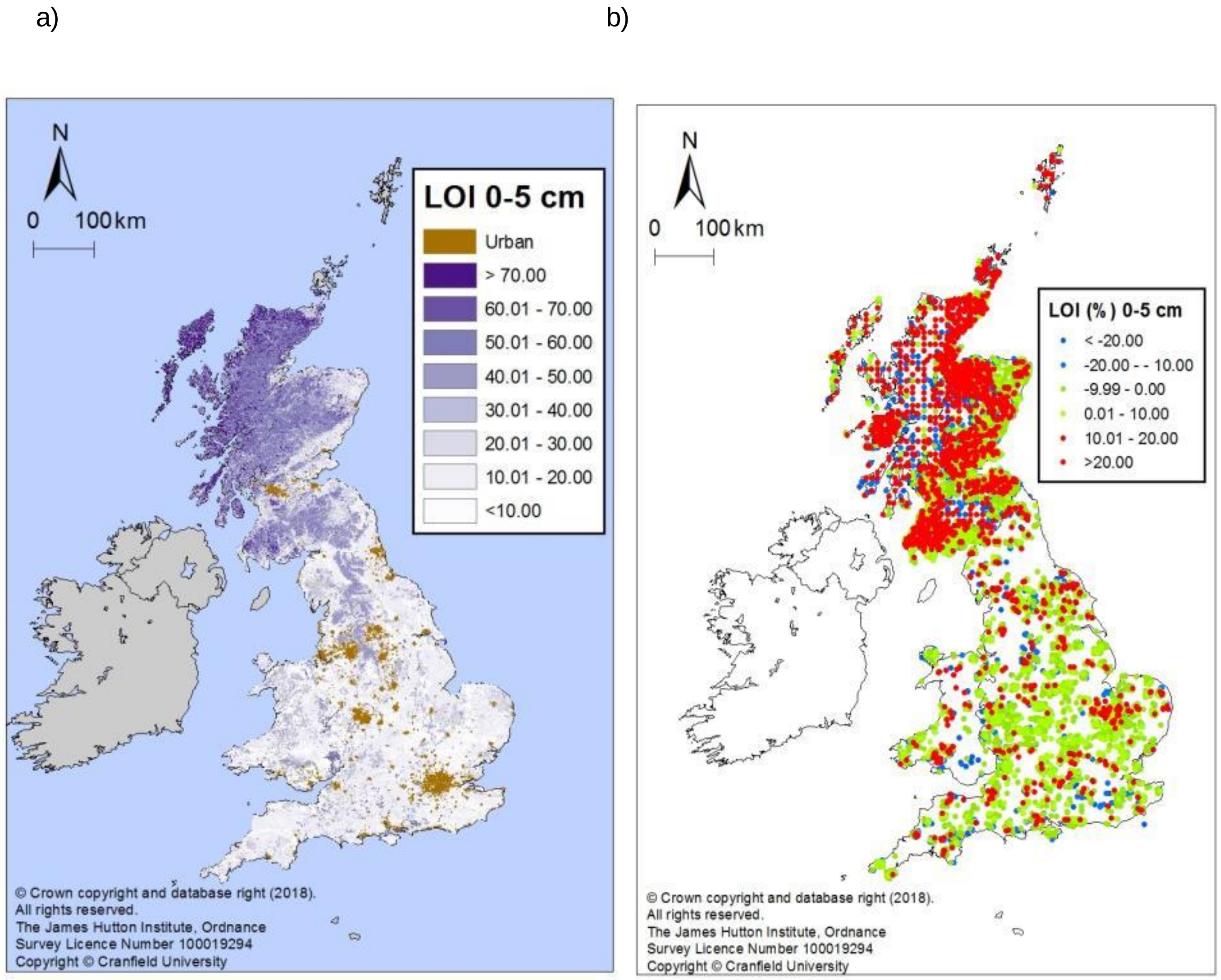


© Crown copyright and database right (2018). All rights reserved. The James Hutton Institute, Ordnance Survey Licence Number 100019294 Copyright © Cranfield University

© Crown copyright and database right (2018). All rights reserved. The James Hutton Institute, Ordnance Survey Licence Number 100019294 Copyright © Cranfield University

Figure 5.2: GB map of a) pH at 0-5 cm using BRT model and b) corresponding residual map.

The LOI maps show that greater values at 0-5 cm depths are found in northern and southern Scotland and central England (around 50-60%) with lesser values located in the south and east of England (less than 10%) (Figure 5.3a). This distribution is consistent for all depths down to 30-60 cm and below 60 cm where the BRT model has predicted LOI values around 20% or less for much of GB (see Appendix 4). The residual maps for all depths (Figure 5.3b), show LOI is shown to be overpredicting across much of Scotland and sporadically throughout England and Wales, However, this becomes less at lower depths (see Appendix 4).



© Crown copyright and database right (2018).
All rights reserved.
The James Hutton Institute, Ordnance
Survey Licence Number 100019294
Copyright © Cranfield University

© Crown copyright and database right (2018).
All rights reserved.
The James Hutton Institute, Ordnance
Survey Licence Number 100019294
Copyright © Cranfield University

Figure 5.3: GB map of a) LOI at 0-5 cm using BRT model and b) corresponding residual map.

Maps of sand content (at 0-5 cm, Figure 5.4a) illustrate larger values being predicted in the north and west Scotland (around 65-80%) and smaller values being predicted in the southern and eastern England (25-50%) and is consistent at all depths. When comparing mapped sand residuals at 0-5 cm (Figure 5.4b), it can be noted that sand is overpredicted across much of GB, although there is a slight underprediction in the far south of England and small areas of Wales.

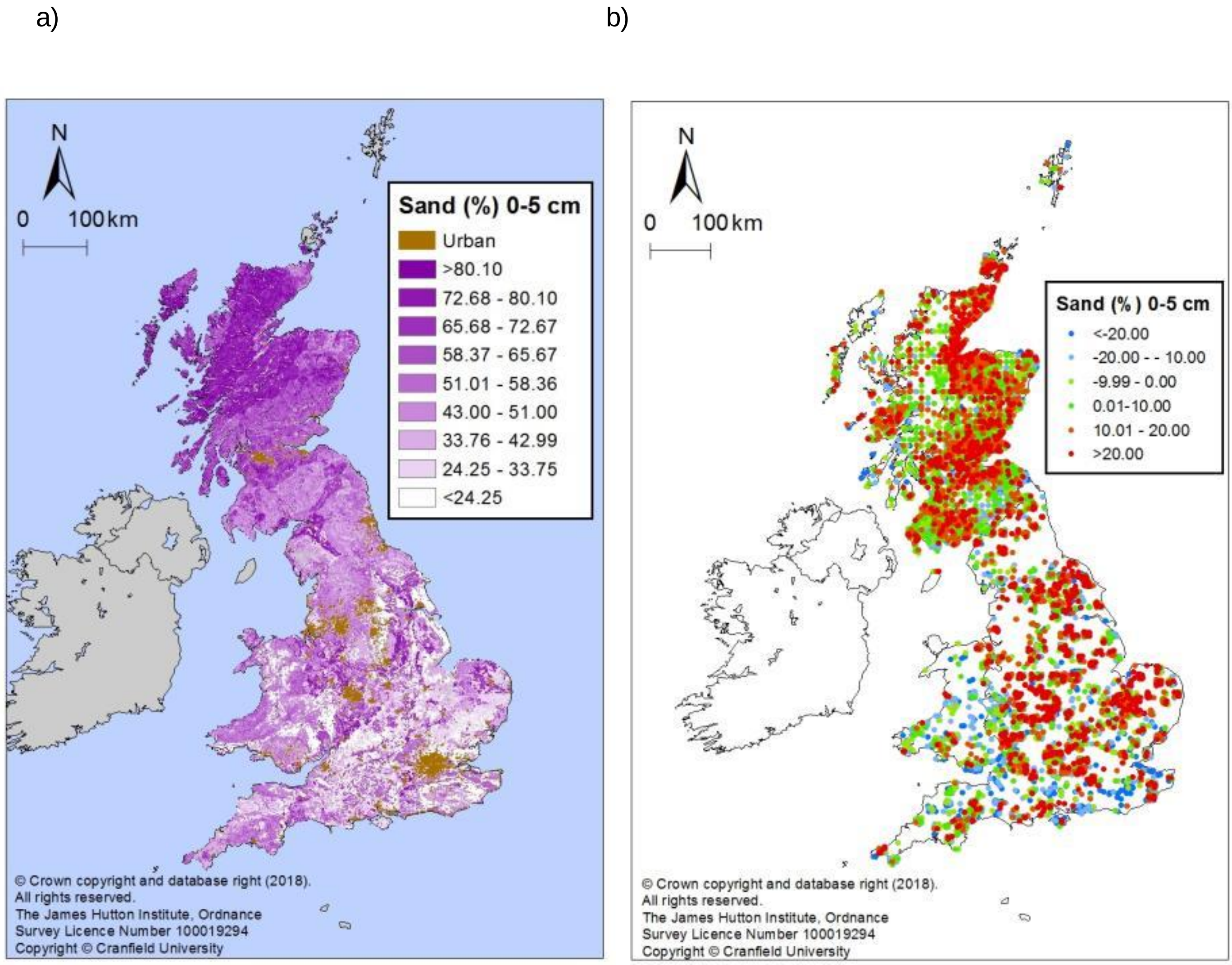

a)
b)
N
0 100km
Sand (%) 0-5 cm
Urban
>80.10
72.68 - 80.10
65.68 - 72.67
58.37 - 65.67
51.01 - 58.36
43.00 - 51.00
33.76 - 42.99
24.25 - 33.75
<24.25
© Crown copyright and database right (2018).
All rights reserved.
The James Hutton Institute, Ordnance
Survey Licence Number 100019294
Copyright © Cranfield University
N
0 100km
Sand (%) 0-5 cm
<-20.00
-20.00 - - 10.00
-9.99 - 0.00
0.01-10.00
10.01 - 20.00
>20.00
© Crown copyright and database right (2018).
All rights reserved.
The James Hutton Institute, Ordnance
Survey Licence Number 100019294
Copyright © Cranfield University

Figure 5.4: GB map of a) sand at 0-5 cm using BRT model and b) corresponding residual map.

The maps of predicted silt content for GB show contents in south eastern England and parts of Wales at around 50-60% with the least silt values (less than 30%) being found across Scotland notably in the far north and west (Figure 5.5a). This trend remains constant for all depths down the soil profile. The residual map of silt content for GB at 0-5 cm shows silt to be overpredicted across much of the country, particularly in eastern Scotland but sporadically across much of south England and southern Wales (Figure 5.5b).

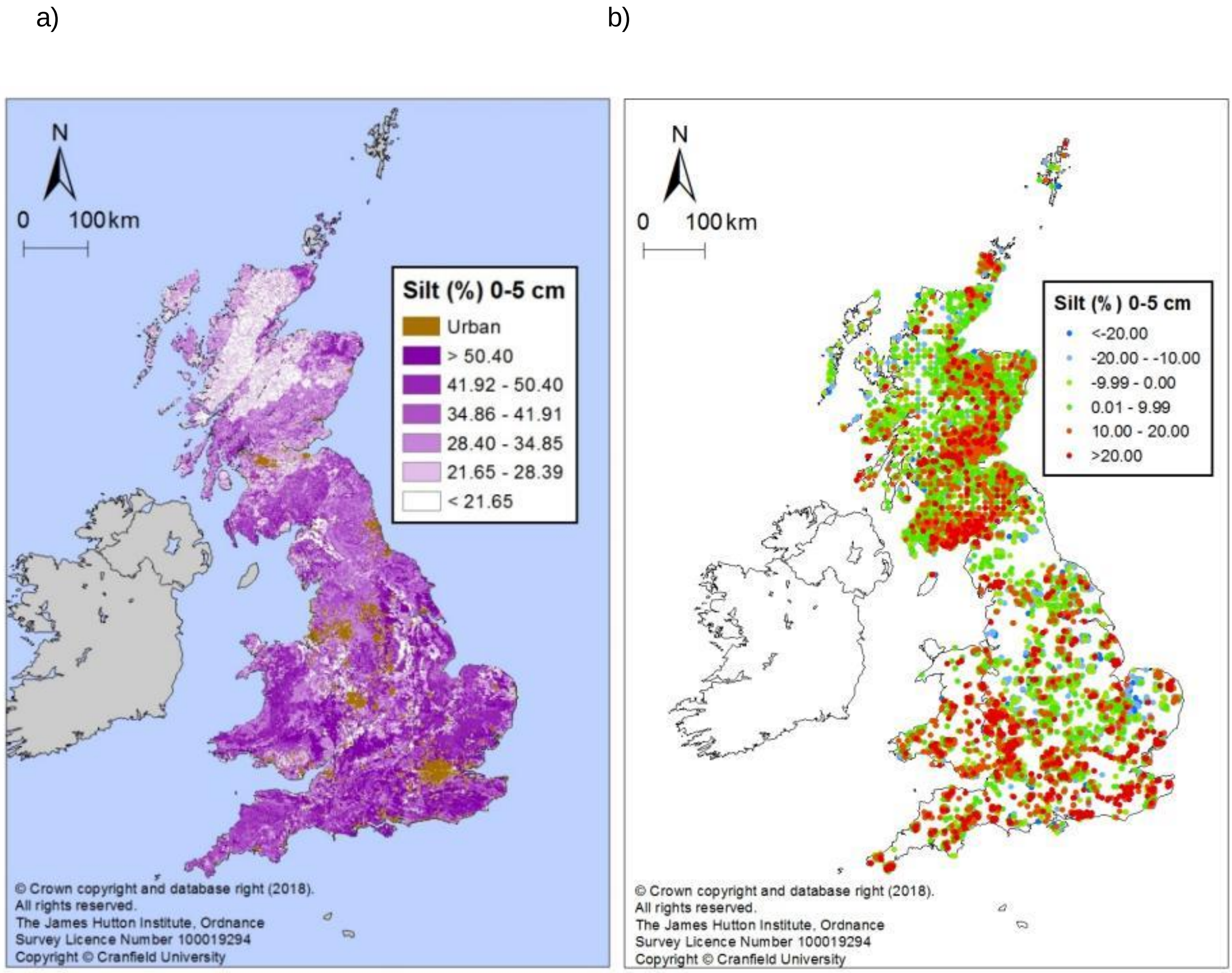
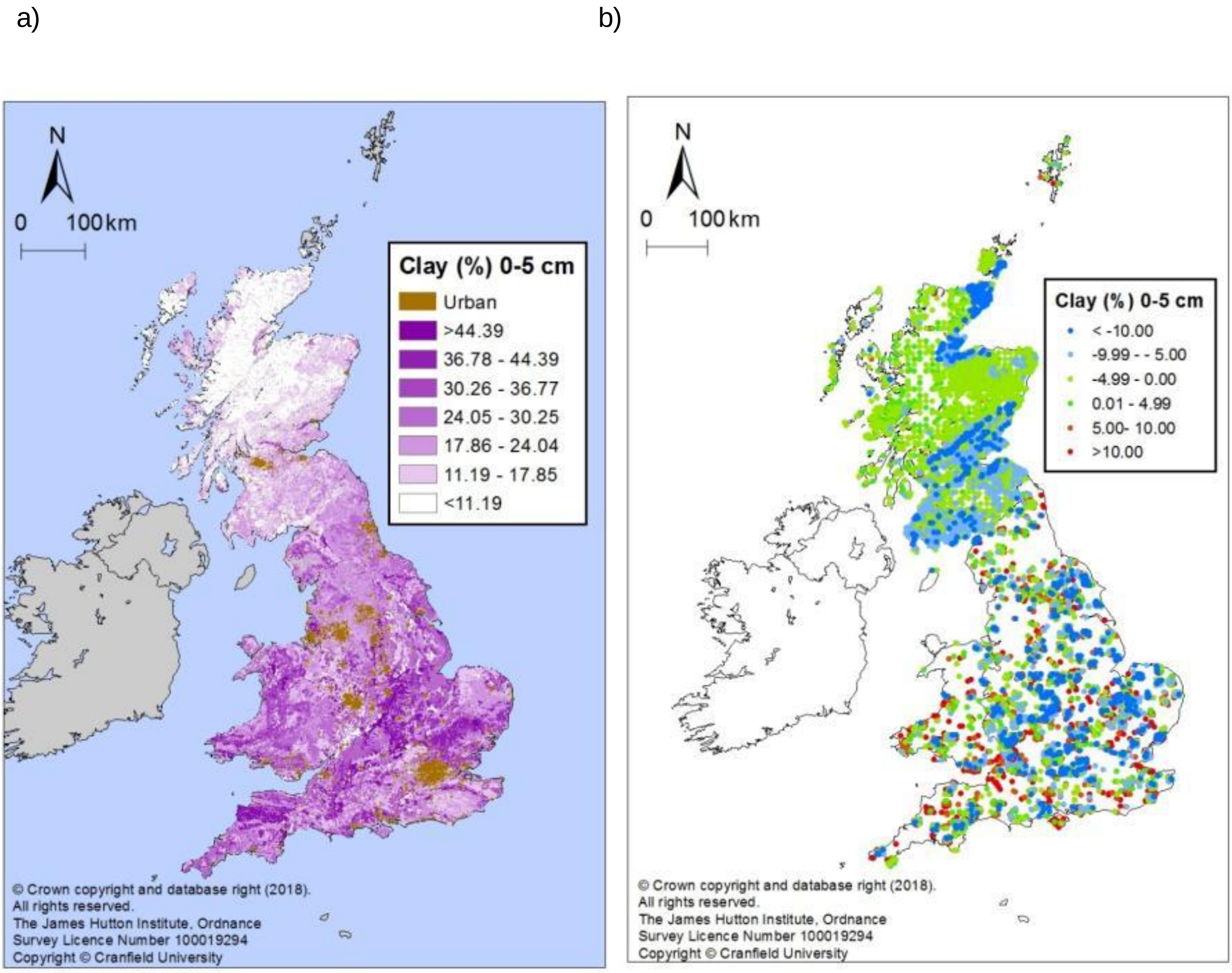

a)
b)
N
0 100km
Silt (%) 0-5 cm
Urban
> 50.40
41.92 - 50.40
34.86 - 41.91
28.40 - 34.85
21.65 - 28.39
< 21.65
© Crown copyright and database right (2018).
All rights reserved.
The James Hutton Institute, Ordnance
Survey Licence Number 100019294
Copyright © Cranfield University
N
0 100km
Silt (%) 0-5 cm
<-20.00
-20.00 - -10.00
-9.99 - 0.00
0.01 - 9.99
10.00 - 20.00
>20.00
© Crown copyright and database right (2018).
All rights reserved.
The James Hutton Institute, Ordnance
Survey Licence Number 100019294
Copyright © Cranfield University

Figure 5.5: GB map of a) silt at 0-5 cm using BRT model and b) corresponding residual map.

The maps of clay content show greater predicted values in the south of England, particularly in the east and south west (around 35% and above) with the smallest predicted clay values across much of Scotland (<25%) (Figure 5.6a). This trend is consistent for all depths, however, in Scotland there is a slight increase in predicted clay content at lower depths. When compared to the map of residuals at 0-5 cm (Figure 5.6b), clay is underpredicted across much of eastern Scotland and much of England and Wales.



© Crown copyright and database right (2018).
All rights reserved.
The James Hutton Institute, Ordnance
Survey Licence Number 100019294
Copyright © Cranfield University

© Crown copyright and database right (2018).
All rights reserved.
The James Hutton Institute, Ordnance
Survey Licence Number 100019294
Copyright © Cranfield University

Figure 5.6: GB map of a) clay at 0-5 cm using BRT model and b) corresponding residual map.

5.5.3. Independent validation and residual maps

Table 5.3 shows the independent validation of predicted soil properties taken at 0-15 cm for all soil properties across the whole of GB. These results indicate poorer R^2 values than those obtained using the cross-validation during the training of the BRT models, particularly sand, silt and clay (0.18 or lower). The RMSE is high for most of the soil property values particularly texture properties.

Soil property validation results							
Soil property	**R^2**	**RMSE**	**Mean**	**SD**	**Observed Range**	**Predicted Range**	**Samples (n)**
pH	0.16	1.25	5.38	1.11	3.10 – 9.20	3.99 – 8.04	5347
LOI	0.17	16.65	32.60	16.32	1.54 – 98.90	8.25 – 95.90	5355
Sand	0.04	24.72	45.27	20.21	4.85 – 94.41	10.73 – 85.79	4630
Silt	0.03	17.81	41.68	10.10	0.10 – 86.96	10.89 – 68.14	4630
Clay	0.02	14.69	27.03	13.57	1.86 – 63.13	2.53 – 57.60	4630

Table 5.3: Model predictions validated with an independent validation dataset for soil properties at 0-15 cm.

The residual output map for pH from the independent validation dataset (Figure 5.7) shows some spatial structure across GB. pH is being overpredicted across much of eastern England and underpredicting across much of Wales. However, in Scotland, there seems to be little structure as pH is being over and underpredicted. LOI is shown to be overpredicted across parts of western Wales and England. However, on whole, LOI seems to be predicted reasonably well across England and Wales as seen from the independent validation dataset (Figure 5.8). This same output shows little spatial structure in LOI predictions across Scotland. Sand content is shown to be overpredicted across much of eastern England and underpredicted across much of Wales. In Scotland, there again seems to be little structure although sand is underpredicted in eastern areas (Figure 5.9). Silt content is shown to be overpredicted across much of western England, Wales and across much of Scotland particularly in the east and south. Silt is underpredicted in parts of eastern England (Figure 5.10) and clay is overpredicted across much of Scotland when compared to the independent validation dataset. In England and Wales, the pattern is similar with clay shown to be underpredicted in the far south east. However, on the whole clay is still being overpredicted across much of England and Wales (Figure 5.11).

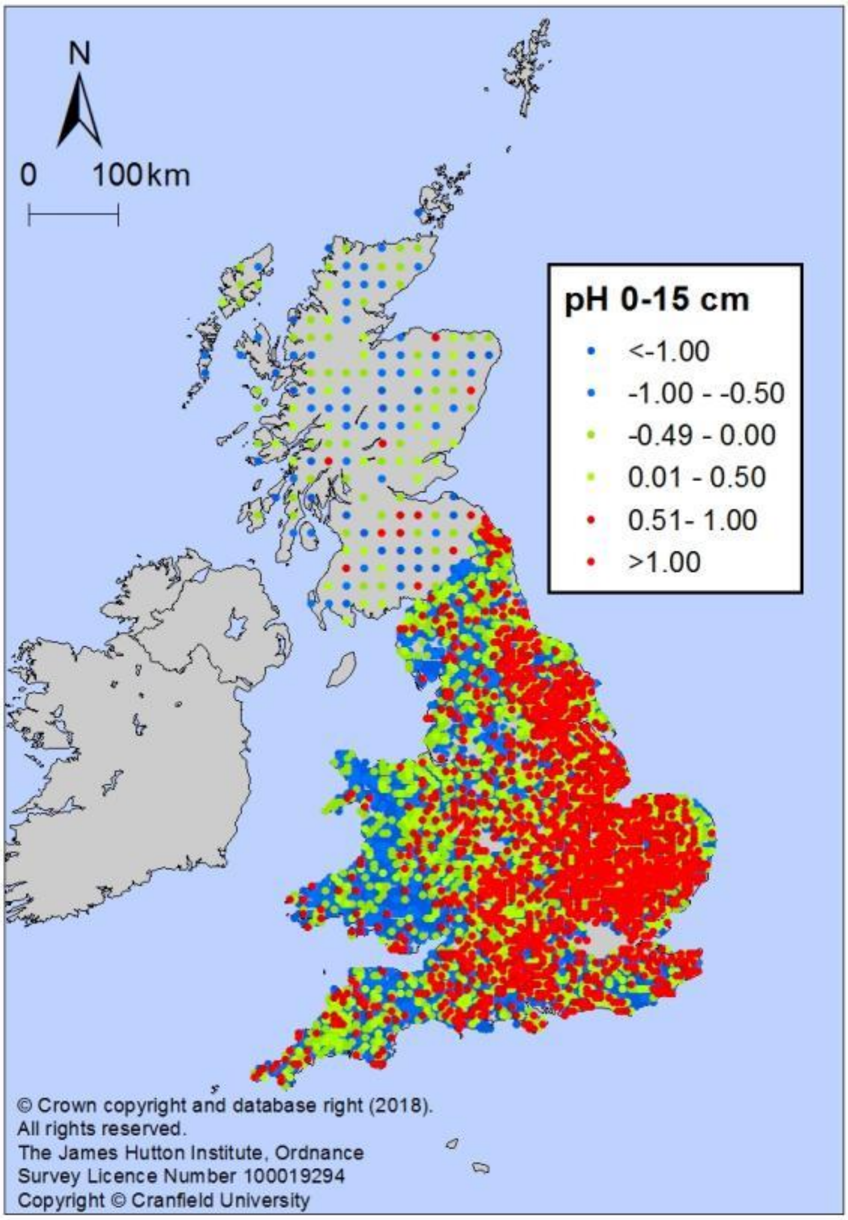


Figure 5.7: Residual map from GB independent validation dataset at 0-15 cm for pH

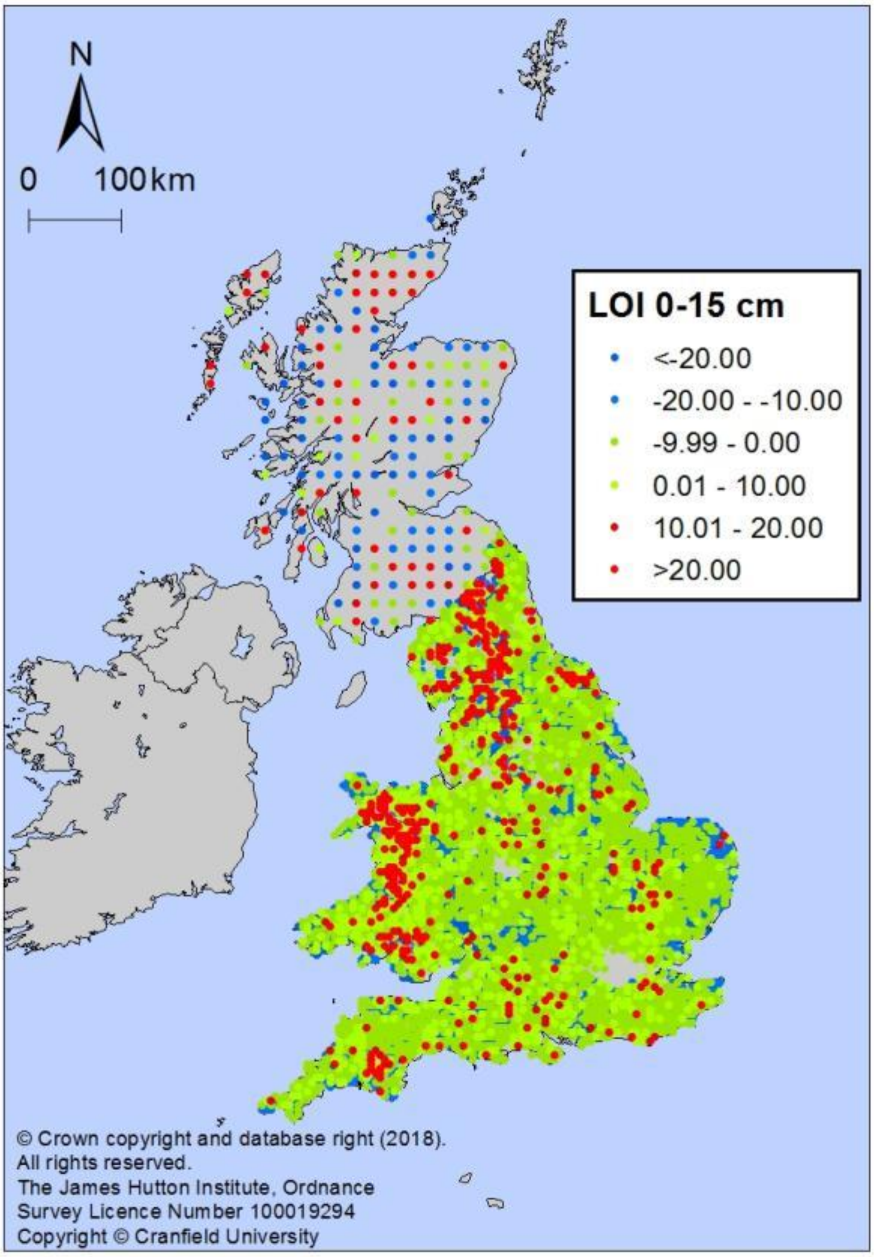

N
0 100km
LOI 0-15 cm
● <-20.00
● -20.00 - -10.00
● -9.99 - 0.00
● 0.01 - 10.00
● 10.01 - 20.00
● >20.00
© Crown copyright and database right (2018).
All rights reserved.
The James Hutton Institute, Ordnance
Survey Licence Number 100019294
Copyright © Cranfield University

Figure 5.8: Residual map from GB independent validation dataset at 0-15 cm for LOI

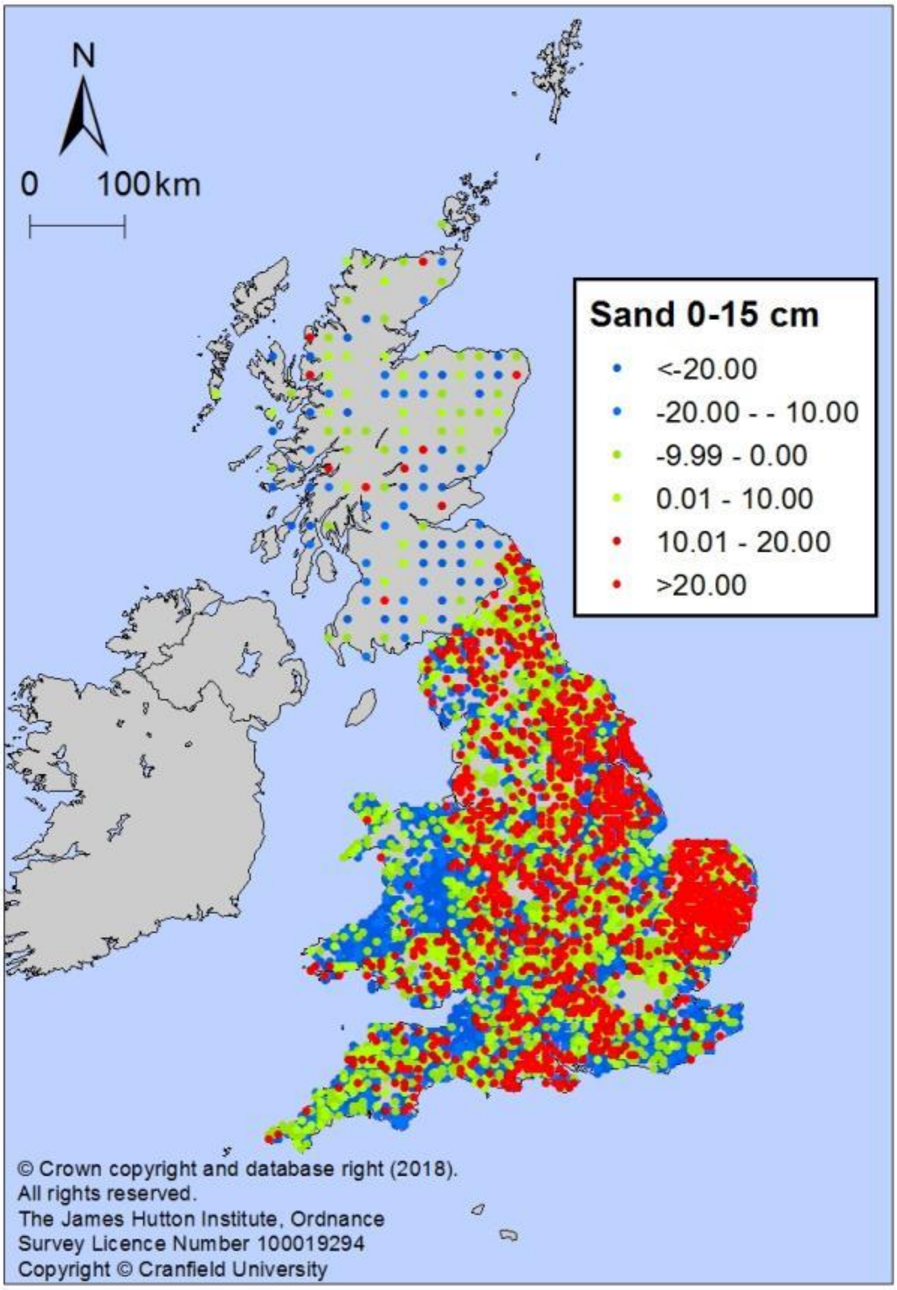

N
0 100km
Sand 0-15 cm
● <-20.00
● -20.00 - - 10.00
● -9.99 - 0.00
● 0.01 - 10.00
● 10.01 - 20.00
● >20.00
© Crown copyright and database right (2018).
All rights reserved.
The James Hutton Institute, Ordnance
Survey Licence Number 100019294
Copyright © Cranfield University

Figure 5.9: Residual map from GB independent validation dataset at 0-15 cm for Sand

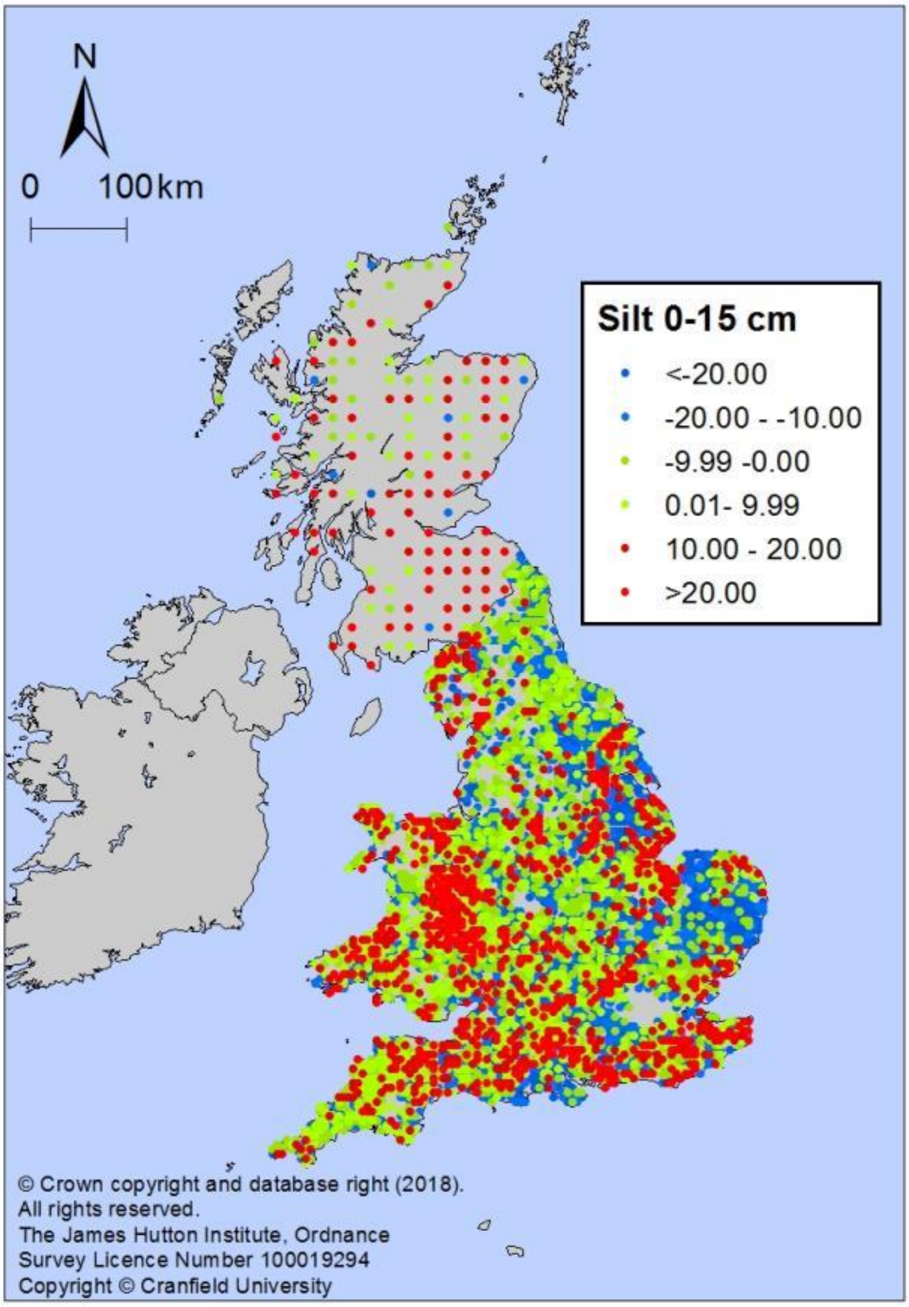

N
0 100km
Silt 0-15 cm
<-20.00
-20.00 - -10.00
-9.99 -0.00
0.01- 9.99
10.00 - 20.00
>20.00
© Crown copyright and database right (2018).
All rights reserved.
The James Hutton Institute, Ordnance
Survey Licence Number 100019294
Copyright © Cranfield University

Figure 5.10: Residual map from GB independent validation dataset at 0-15 cm for Silt

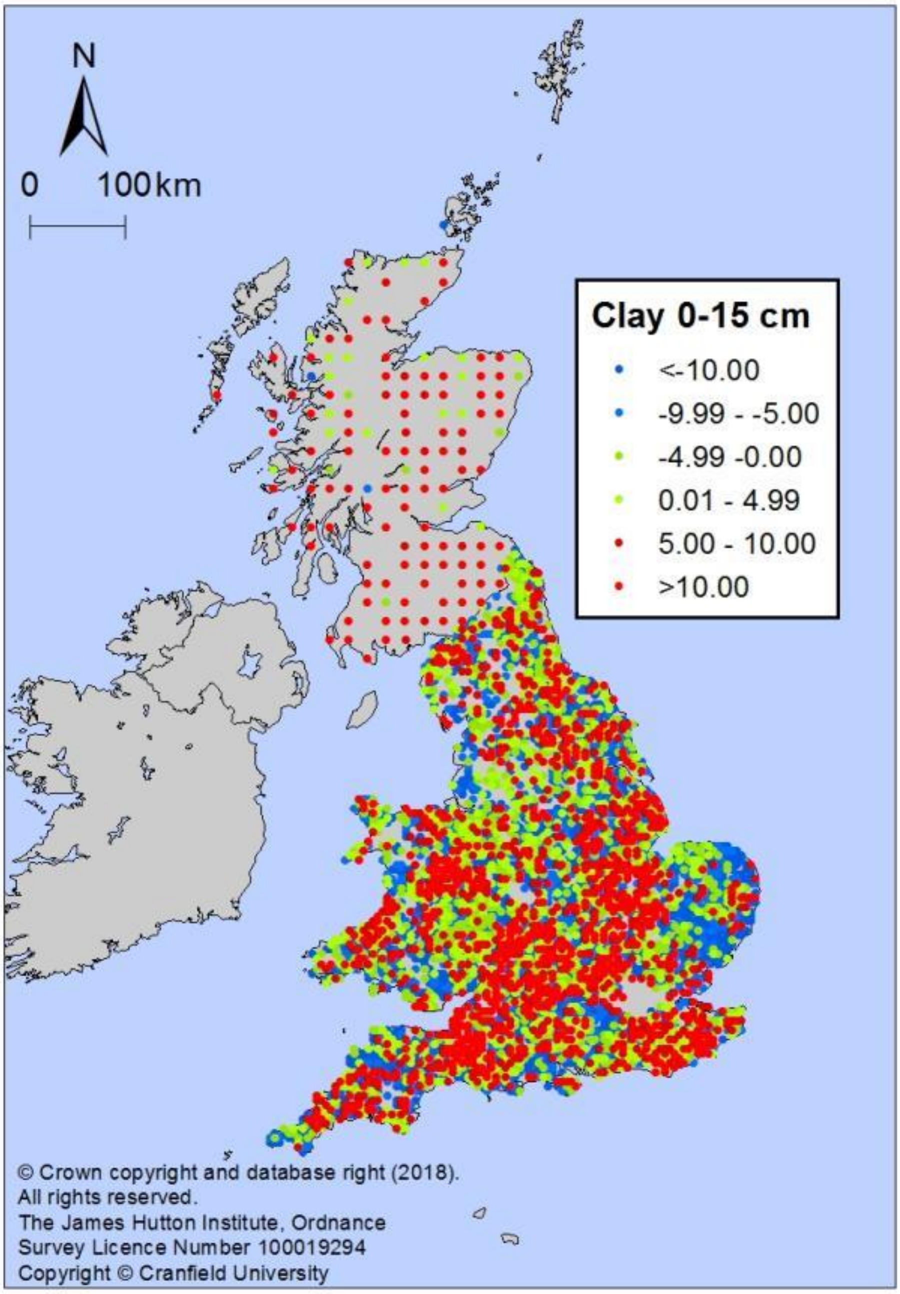

N
0 100km
Clay 0-15 cm
<-10.00
-9.99 - -5.00
-4.99 -0.00
0.01 - 4.99
5.00 - 10.00
>10.00
© Crown copyright and database right (2018).
All rights reserved.
The James Hutton Institute, Ordnance
Survey Licence Number 100019294
Copyright © Cranfield University

Figure 5.11: Residual map from GB independent validation dataset at 0-15 cm for clay

5.6. Discussion

5.6.1. Independent validation

The independent validation dataset showed significantly poorer R^2 values for all soil properties than when trained on the BRT model. LOI and pH performed badly (R^2 values 0.17 and 0.16 respectively) and texture performed even worse (R^2 between 0.04 and 0.06). However, previous work (Nussbaum et al, 2018) has shown that models only account for a moderate part of the variance in the validation data for texture and pH (around 0.25 mean square error; for GB, pH showed a RMSE of 1.25). Many other studies have also independently validated DSM against pH and other soil properties at multiple depths and reported poor R^2 values ranging from 0.1 to. 0.48 (Mulder et al, 2016) and negative to 0.75 (Vaysse and Lagacherie, 2015). These results are similar with this modelling work for GB (between 0.02 and 0.04 for sand, silt and clay to 0.17 for LOI).

Piikki and Söderström (2017) presented validation results on the Digital Soil Map of Sweden at many scales using a large validation dataset of soil analyses at farm scale. Predicted clay content error was reported to be less than 8% in 75% of the validation samples, while for sand it was 13%. The authors illustrate that larger scale predictions are difficult to validate on small scale farm measurements because there may be different processes which account for soil variability at these scales (e.g. geology and climate at national level). Other issues such as local farm management will affect changes in soil properties at farm scale. In comparison to the GB work, raster predictions were obtained at 100 m based on coarse scale national covariate data and against an independent dataset which focusses on a sample taken at a specific point in the landscape.

In order to improve mapping and modelling across GB, there is a need to collect more data from areas where only sparse information is available. This would enable more consistent modelling at regional scale. It is important to recognise that the training and validation datasets used in this GB study were collected and analysed at different times and used as novel independent comparisons. However, changes in some soil properties have been recorded across GB during this time (e.g. Chapman et al, 2015, Lilly and Baggaley, 2013, Bell et al, 2011, Bellamy et al, 2005, Reynolds et al, 2013) reflecting pollution recovery, climate change and land management changes (Kirk et al, 2010; Bell et al, 2011). These changes would be reflected in these independent datasets and as a result would influence the evaluation of the outcome of the training and validation. Therefore, uncertainties in the maps could reflect changes in soil properties rather than poor model performance. An alternative would be to use a sub-set of the same time period data sets for training and validation, but this has its own issues as this subset is no longer independent.

5.6.2. Pedological evaluation

The model performance statistics of the BRT models show the strength of the statistical relationship between the soil properties and covariates but when these are mapped a further evaluation of the model output can be achieved by comparing spatial patterns of soil properties with our understanding of how soils vary in different parts of GB. Therefore, it is important to evaluate these GB soil property maps and see if they reflect pedological knowledge spatially and at depth in addition to the model performance statistics.

In the north of Scotland, we would expect topsoil pH values between 3 and 5 associated with Histosols and histic Podzols and Gleysols. The predicted pH values obtained from the BRT model in these areas are between 4 and 5 which is within the expected range given the dominant soil types in these areas. Some Cambisols and Leptosols in south eastern England are calcareous with corresponding alkaline pH because they are formed on calcareous parent material (limestone and chalk). The BRT model predicted soils in these areas with a pH value of greater than 7.5 which matches with pedological understanding. There are some known calcareous soils found in coastal areas across Scotland, however, that are not predicted by the BRT model. These soils are not extensive in Scotland and thus are not well represented in the training dataset. Hengl et al (2017) produced global gridded soil properties (including pH) predicted from random forest and gradient boosting modelling approaches. Cross validation showed the ensemble models explained between 83% variation of pH with an overall average of 61%. These are an improvement on the results of the modelling for GB soils. However, the pH was still being overestimated in specific areas (e.g. in the rainforests of Tasmania, Australia (Hengl et al, 2017) and have limited ground observations to validate the predictions in these areas. Other studies such as Reuter et al, (2008) used regression kriging for mapping soil pH on a European scale. This study showed satisfactory R^2 values (between 0.4 and 0.5) which are comparable to this study for GB. The authors also provide useful commentary on pedological understanding across Europe. They observed more acidic pH values in granitic-based areas in Portugal and north Spain and in shallow soils across Scandinavia. This is also noted for northern Scotland too, similar to what has been noted in this study. More alkaline based pH values were concentrated in Mediterranean countries (e.g. south of France and Italy) due to calcareous parent material and in south eastern England, again, similar to what has been found from this study.

As expected, LOI values are greatest in the north of Scotland where the main soil types are Histosols and histic Podzols and Gleysols and in Northern England reflecting where Histosols and histic Stagnosols are known to be present. Although some areas of Histosols in North Wales do not have high predicted LOI. Lesser LOI values at depth match expected values for the mineral horizons of Podzols and Gleysols. Histosols are classified as having an organic horizon equal to or greater than 40 cm. As a result, we would expect to observe large LOI values at depth (e.g. at 30-60 cm) in Histosols. In the predicted maps, this is observed for only some of the areas containing Histosols in north and West Scotland (Appendix 4). In upland areas in Wales, LOI values diminish after 30 cm, where we would expect to see deeper peats (Histosols). There are some observed values of 70% for LOI at 60-100 cm, this is not being predicted in the mapping outputs at 60-100 cm (see Appendix 4 and 5). Podzols and Histic Gleysols have peaty surface soils and large LOI values in the topsoil (0-30 cm) and values will decrease in the underlying mineral horizons. Therefore, a reduction in LOI values at depths beyond 30 cm would be expected in these areas. This is shown in the predicted LOI at depths greater than 30 cm for the areas of Podzols in Scotland (Appendix 4). Comparable work was undertaken by Poggio and Gimona (2014) modelling soil organic carbon in 3D comparing regression kriging and depth function methods. Depth function modelling showed better performances of predicting soil organic carbon than regression kriging (R^2 = 0.60). This is more consistent than the results found in this study. However, the authors (Poggio and

Gimona, 2014) incorporated additional covariates such as MODIS (Moderate Resolution Imaging Spectroradiometer) data into the model.

For textural properties across GB, the BRT model was used to predict sand, silt and clay independently of each other and of the organic matter concentration as expressed through LOI. As a result, this has given an over prediction of soil particle size classes in areas dominated by organic surface horizons. One way to improve the modelling is to analyse soil particle size (sand, silt and clay) of the mineral fraction only. Poggio and Gimona (2017) masked pixels with greater probability of having organic soils and therefore only predicted soil particle size classes where there was a likelihood of mineral soils occurring. Results were in good agreement with values at validation locations. More importantly, their results produce better texture outputs than in this study albeit at coarser resolution (250 m). Another way to improve the modelling would be retrospectively rescale the sand, silt and clay using predicted LOI.

5.6.3. Effect of time on soil properties

As different datasets are being used across different time periods, it is important to understand how this may affect soil properties and the implications for using data for model training and validation. The NSIS and NSI have both been resampled once. For soil pH, both surveys indicated an increase in soil pH (Kirk et al, 2010, Black, pers comm, 2019). For LOI, both surveys recorded changes in soil organic matter (SOM) and soil organic carbon (SOC) content relating to different soil types, with significant losses from both surveys in arable soils. For Scottish soils, the carbon stock showed no change (Chapman et al, 2015) though carbon concentrations declined in both arable and improved grassland sites. Texture is unlikely to have changed over time but at present no one has assessed these changes in great detail.

Other surveys across GB such as Reynolds et al (2013), found that soils from 1978 to 2007 showed significant increases in pH across all land uses, which is consistent with reduction in acid rain (Kirk et al, 2010). Reynolds et al (2013) also noted changes in topsoil carbon concentration, with consistent declines in arable soils. The changes in SOC may reflect fluctuations in nitrogen deposition (Tipping et al, 2017) while management and climate change will also be factors in soil carbon variations (Bell et al, 2011).

Overall, these results indicate that the use of legacy data in DSM adds to the uncertainty when mapping for current uses.

5.7. Conclusions

This chapter has critically evaluated soil property predictions using cross validation and an independent validation dataset. The models have shown that cross GB modelling can be accomplished using the two national surveys. The BRT model performance indicates that soil pH and LOI are being predicted reasonably across GB and (mostly) at depth with texture predicted less well. A likely reason for this is the large range of data which has been used for both LOI and pH in comparison with texture. There is also a good array of covariates which can be used alongside the observed soil property data. Poor performance of the model against an independent validation dataset may be a consequence of the original data for both GB datasets being collected and analysed at different time periods and the data for both training and validation GB datasets being collected at different scales. Future work should therefore focus on collecting more data from areas where there is sparse information to produce more consistent modelling outputs. Furthermore, validation of the models will require increased expert knowledge to ensure that these and subsequent mapping outputs are both reliable and ultimately useful in the future.

Predictions of soil property outputs (soil LOI (as a proxy for soil organic carbon), pH and texture) across GB have been evaluated in relation to expert pedological understanding. The BRT models have produced reasonable statistical indices for predicting pH and illustrate strong pedological relationships between soils across GB and at depth. LOI is also shown to be predicting well across GB overall. However, at depths beyond 60 cm, the BRT model fails to predict large LOI values where we would expect to find deep organic horizons associated with peat soils (Histosols). Therefore, new data will be required to address this. The BRT model has failed to predict texture properties well since the model was used to predict each particle size class over the whole of GB including those areas dominated by organic surface horizons. This is an issue of ignoring other properties in soils, namely organic matter. To improve the modelling in future, an analysis of soil particle size (sand, silt and clay) using the mineral fraction only or by retrospectively rescaling sand, silt and clay using predicted LOI could be considered.

At present, these first versions of soil property DSM maps for GB are variable in terms of whether they can be used by stakeholders. However, these outputs have shown that developing reliable DSM maps in future would benefit from increased interaction between pedologists, modellers and stakeholders to produce usable mapping outputs of sufficient quality at finer resolution.

References

Adhikari, K., Kheir, R.B., Greve, M.B., Greve, M.H., Malone, B.P., Minasny, B., McBratney, A.B., 2014: Mapping soil pH and bulk density at multiple soil depths in Denmark. Chapter In: McKenzie et al., 2014: GlobalSoilMap: Basis of the global spatial soil information system. Taylor and Francis Group.

Avery, C.L., Bascomb, B.W., 1982: Soil Survey Technical Monograph No.6. Soil Survey Laboratory Methods. Harpenden.

Ballabio, C., Panagos, P., Montanarella, L., 2014: Mapping topsoil physical properties at European scale using the LUCAS database. *Geoderma*, 261, pp. 110-123.

Behrens, T., Scholten, T., 2006: Digital soil mapping in Germany—a review. *Journal of Plant Nutrition and Soil Science,* 169, 3, pp. 434-443.

Bell, M.J., Worrall, F., Smith, P., Bhogal, A., Black, H.I.J., Lilly, A., Barraclough, D., Merrington, G., 2011: UK land-use change and its impact on SOC: 1925-2007. *Global Biogeochemical Cycles, 25, GB4015, doi:10.1029/2010GB003881.*

Bellamy, P.H., Loveland, P.J., Bradley, R.I, Lark, R.M., Kirk, G.J.D., 2005: Carbon losses from all soils across England and Wales 1978 -2003, *Nature,* pp.245-248

Bishop, T.F.A., McBratney, A.B., Whelan, B.M., 2001: Measuring the quality of digital soil maps using information criteria. *Geoderma* 103, 1, pp. 95-111.

Campbell, G.A., Lilly, A., Corstanje, R., Mayr, T.R., Black, H.I.J., 2017: Are existing soils data meeting the needs of stakeholders in Europe? An analysis of practice use from policy to field. *Land Use Policy*, 68, pp.211-223.

Campbell, G.A., Lilly, A., Corstanje, R., Hannam, J., Black, H.I.J., in prep: Evaluation of two model typologies and their behaviour in generating soil property predictions: studies from pilot areas in England and Scotland.

Carré, F., McBratney, A.B., Mayr, T., Montanarella, L., 2007: Digital soil assessments: Beyond DSM. *Geoderma*, 142,1–2, pp. 69-79.

Chapman, S.J., Bell, J.S., Campbell, C.D., Hudson, G., Lilly, A., Nolan, A.J., Robertson, A.H.J., Potts, J.M., Towers, W., 2015: Comparison of soil carbon stocks in Scottish soils between 1978 and 2009. *European Journal of Soil Science,* 64, pp.455-465

Creamer, R, E., 2016: Irish Soil Information System: Soil Property Maps (2007-S-CD-1-S1), Report No. 204, EPA Research, Department of Communications, Climate Change and Environment. (available at www.epa.ie).

Dobos, E., Carré, F., Hengl, T., Reuter, H.I., Tóth, G., 2006: Digital Soil Mapping as a support to production of functional maps. EUR 22123 EN, Office for Official Publications of the European Community, Luxemburg.

Elith, J., Leathwick, J.R., Hastie, T., 2008: A working guide to boosted regression trees. *Journal of Animal Ecology*, 77: pp. 802–813.

FAO-Unesco-ISRIC.,1988: 1990: Revised Legend of the Soil Map of the World. World Soil Resources Report no. 60. FAO, Rome.

GlobalSoilMap., 2011: GlobalSoilMap.net: New Digital Soil Map of the world. [Accessed from https://www.google.co.uk/search?q=globalsoilmap.net&ie=utf-8&oe=utf-8&aq=t&rls=org.mozilla:en-GB:official&client=firefoxa&channel=sb&gfe_rd=cr&ei=aR3jVLWpOoSV8wOK4YHwBQ] [Last accessed 17th May 2018].

Grunwald, S., Thompson, J.A., Boettinger, J.L., 2011: Digital Soil Mapping and Modelling at Continental Scales: Finding Solutions for Global Issues. *Soil Science Society of America Journal*, 75, pp.1201-1213. doi:10.2136/sssaj2011.0025.

Hallett, S.H., Sakrabani, R., Keay, C.A., Hannam, J.A., 2017: Developments in land information systems: examples demonstrating land resource management capabilities and options. *Soil Use and Management.* 33, pp. 514-529

Hartemink, A.E., Hempel, J., Lagacherie, P., McBratney, A.B., McKenzie, N., MacMillan, R.A., Minasny, B., Montanarella, L., Mendonça Santos, M.L., Sanchez, P., Walsh, M., Zhang, G.L., 2010: GlobalSoilMap.net – A New Digital Soil Map of the World. In: Boettinger J.L., Howell D.W., Moore A.C., Hartemink A.E., Kienast-Brown S. (eds) Digital Soil Mapping. Progress in Soil Science, vol 2. Springer, Dordrecht.

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B. M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotic, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S. and Kempen, B., 2017: SoilGrids250m: Global gridded soil information based on machine learning. *PLoS ONE,* 12, 2: e0169748. https://doi.org/10.1371/journal.pone.0169748

Hudson, H.D., 1992: Division S-5- Soil Genesis, Morphology & Classification: The soil survey as paradigm-based science. Soil Science Society of America Journal 56, 3, pp. 836-841.

Kirk, G.J.D., Bellamy, P.H., Lark. R.M., 2010: Changes in soil pH across England and Wales in response to decreased acid deposition. Global Change Biology, 16, pp.3111-3119.

LeBron, I., Keith, A.M., Hughes, S., Reynolds, B., Robinson, D.A., Cooper, D.M., Emmett, B.A., 2011: Evaluation of the occurrence of soils with pH higher than 8.3 observed within the Countryside Survey, Defra project SP1304. Centre for Ecology and Hydrology. Natural Environment Research Council, Gwynedd, UK.

Lilly, A., Bell, J.S., Hudson, G., Nolan, A.J., Towers, W., 2010: National Soil Inventory of Scotland (NSIS_1): site location, sampling and profile description protocols (1978-1988): Technical Bulletin, Macaulay Institute, Aberdeen.

Lilly, A., Bell, J.S., Hudson, G., Nolan, A.J., Towers, W., 2011: National Soil Inventory of Scotland 2007-2009: Profile description and soil sampling protocols. (NSIS2). Technical Bulletin, James Hutton Institute.

Lilly, A., Baggaley, N.J., 2013: The potential for Scottish cultivated topsoils to lose or gain soil organic carbon. *Soil Use and Management,* 29, pp.39-47.

Malone, B.P., McBratney, A.B., Minasny, B., Laslett, G.M., 2009: Mapping continuous depth functions of soil carbon storage and available water capacity. *Geoderma*, 189-190, pp. 153-163.

McBratney, A. B., Mendonça Santos, M.L., Minasny, B., 2003: On digital soil mapping. *Geoderma*, 117,1, pp. 3-52.

Minasny, B., McBratney, A.B., 2016: Digital soil mapping: A brief history and some lessons, *Geoderma*, 264, pp.301-311.

Minasny, B., McBratney, A.B., Malone, B.P., Sulaeman, Y., 2010: Digital Soil Mapping of soil carbon © 2010 19thWorld Congress of Soil Science, Soil Solutions for a Changing World 1 – 6 August 2010, Brisbane, Australia. Published on DVD.

Minasny, B., A. McBratney., 2010: Methodologies for global soil mapping. Digital soil mapping, Springer: pp. 429-436.

Mulder, V., Lacoste, M., Richer-de-Forges, A.C., Arrouays, D., 2016: GlobalSoilMap France: High-resolution spatial modelling the soils of France up to two-meter depth. *Science of the Total Environment*, 573, pp.1352 - 1369.

Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., Schaepman, M.E., Papritz, A., 2018: Evaluation of digital soil mapping approaches with larger sets of environmental covariates. *SOIL*, 4, pp. 1-22.

Piikki, K., Söderström, M., 2017: Digital soil mapping of arable land in Sweden – Validation of performance at multiple scales, *Geoderma*, pp. 1-9 http://dx.doi.org/.

Poggio, L., Gimona A., 2014: National scale 3D modelling of soil organic carbon stocks with uncertainty propagation. An example for Scotland. *Geoderma*, 232, pp. 284–99

Poggio, L., Gimona, A., 2017: 3D mapping of soil texture in Scotland, *Geoderma Regional*, 9, pp. 5-16.

R Core Team., 2013: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Reed, M., 2008: Stakeholder participation for environmental management: a literature review. *Biological Conservation* 141, 10, pp. 2417-2431.

Reuter, H.I., 2008: Continental-scale Digital Soil Mapping using European Soil Profile Data: Soil pH. *Hamburger Beiträge zur Physischen Geographie und Landschaftsökologie* 2008,19, pp. 91-102.

Reynolds, B., Chamberlain, P.M., Poskitt, J., Woods, C., Scott, W.A., Rowe, E.C., Robinson, D.A., Frogbrook, Z.L., Keith, A.M., Henrys, P.A., Black, H.I.J., Emmett, B.A., 2013: Countryside Survey; National “Soil Change” 1978-2007 for Topsoils in Great Britain – Acidity, Carbon, and total nitrogen status. Vadose Zone Journal, 12, Ridgeway, G., 2009: Package ‘gbm’. R-News.08:05:15. http://www.saedsayad.com/docs/gbm2.pdf.

Sanchez, P. A., Ahamed, S., Carré, F., Hartemink, A.E., Hempel, J., Huising, J., Lagacherie, P., McBratney, A.B., McKenzie, N.G., de Ourdes Mendonca-Santos, M., Minasny, B., Montanarella, L., Okoth, P., Pal, C.A., Sachs, J.D., Shephard, K.D., Vagen, T., Vanlauwe, B., Walsh, M.G., Winowiecki, L.A., Zhang, G.L., 2009: Digital soil map of the world. *Science,* 325, 5941, pp. 680-681.

Scull, P., Franklin, J., Chadwick, O.A., McArthur, D., 2003: Predictive soil mapping: a review. *Progress in Physical Geography*, 27, 2, pp. 171-197.

SNH, 2002: Natural Heritage Zones: A national assessment of Scotland’s landscapes. [Accessed from https://www.nature.scot/sites/default/files/2017-06/B464892%20-%20National%20Assessment%20of%20Scotland%27s%20landscapes%20%28from%20NHF%29.pdf] [Last accessed 31st July 2018].

Terribile, F., Coppola, A., Langella, G., Martina, M., Basile, A., 2011: Potential and limitations of using soil mapping information to understand landscape hydrology, *Hydrology and Earth System Sciences*, 15, pp. 3895-3933.

Tipping, E., Davies, J.A.C., Henrys, P.A., Kirk, G.J.D., Lilly, A., Dragosits, U., Carnell, E.J., Dore, A.J., Sutton, M.A., Tomlinson, S.J., 2017: Long-term increases in soil carbon due to ecosystem fertilization by atmospheric nitrogen deposition demonstrated by regional-scale modelling and observations. *Scientific Reports,* 7, Article Number: 1890

Vaysse, K., Lagacherie, P. 2015: Evaluating Digital Soil Mapping approaches for Mapping GlobalSoilMap soil properties from legacy data in Languedoc-Roussillon (France), *Geoderma Regional*, 4, pp.20-30.

6 DISCUSSION

6.1. Revisit of aims and objectives

The main aim of this work was to improve the spatial resolution of soil properties across GB as informed by stakeholders to help aid future decision making, planning and policy development.

The first objective explored the soils-related information and data stakeholders currently use in their everyday working activities, and the desired improvements they would like to see from future work. From the survey results, it was noted that wider use of existing (and likely future) soil information by non-experts could be enhanced by improving data accessibility and increasing user-friendly supporting materials. Stakeholders also appreciated the need for fundamental soil properties such as soil chemistry, texture and carbon. Most stakeholders required finer spatial resolution of soils information than what is currently available to them. Finally, stakeholders wanted to gain more information on contemporary soils information and trends over time as well as improved subsequent functional maps and models derived from the soil properties.

The second stage of this thesis focussed on developing DSM for GB, investigating how soil properties are mapped and modelled. Soil properties of loss-on-ignition, pH and soil texture were chosen from the questionnaire survey and an assessment of laboratory and analytical techniques across Scotland and England and Wales. It was noted that during model training of the soil properties in two pilot areas, MARS produced better performances than BRT in both pilot areas. However, when MARS is deployed, it failed to predict soil properties beyond the values of the training dataset. Therefore, MARS models cannot be used for modelling and mapping soil properties across GB. On the other hand, the BRT models, which had poorer correlations within the training dataset, produce more consistent outputs for mapping the soil properties and better represent existing pedological knowledge. This led to BRT models being used for predicting soil properties across GB.

The third objective of this thesis examined DSM by mapping the soil properties across GB, investigating how well they mapped based on pedological understanding. This work also focussed on how suitable an independent validation dataset is for evaluating soil property predictions. The results illustrated that BRT models work reasonably across GB for pH and LOI but less so for texture properties. Furthermore, modelling and mapping soil properties across GB has provided inconsistent outcomes about whether an independent validation dataset is best to use to confirm if soil property outputs have been mapped effectively or not. Therefore, the results from the GB mapping only partially answer what the third objective set out to achieve.

6.2. Overview of the DSM process

What this thesis does differently from a range of academic papers is analyse the available methods and modelling approaches rather than be motivated by personal preferences of applying specific model types. A decision framework needs to be designed so that the stakeholder can be guided through the process and make an assured choice based on various stages of the DSM process. An example of a methodology which could be implemented is the recently published Soil Organic Carbon Mapping Cookbook (Yigini et al, 2018). Model results are associated with uncertainty, but these will become of interest to a range of stakeholders.

Throughout this PhD, it has been important to thoroughly investigate and evaluate each stage of the DSM process (Figure 6.1). There must be an understanding of where the input datasets and associated covariates have been collected and sourced from, and what work is required to make these useable. This is a potential challenge for soils data across GB, as some of soil properties have:

- Differing laboratory and analytical techniques.
- Been collected by different sampling strategies.
- Been collected over different time scales.

From this, it is important to use the training dataset (a snapshot of the whole area based on observed soil data points) as a basis for development and application of the specific model of choice considering the results of cross validation. The models based on a training dataset can then be deployed to a much larger area (e.g. 100 x 100 m grid) to provide the full landscape characteristics. It is fundamental to investigate the predictions from the training dataset and compare this with an independent validation dataset. Within the DSM community, there are generally three types of evaluation (Finke 2011):

- Model performance, evaluated by a range of statistical indices (R^2, RMSE etc.).
- Comparing the statistical indices with other models and resultant maps from other studies.
- How well the mapped outputs reflect expert knowledge and mental models produced by pedologists and soil surveyors (Hudson, 1992).

The latter of these evaluation approaches has not been explored enough in DSM academic literature, yet findings have been discussed in work by Vaysse and Lagacherie (2015), Stoorvogel et al, (2009) and Brus et al, (2011). These evaluation methods are important in telling us whether deploying a global model across a large area is effective for mapping soil properties or if different models are needed for unique soils. There are few studies where an independent validation dataset (completely independent from the training data) has been used to evaluate the model outputs (Piikki and Soderstrom, 2017).

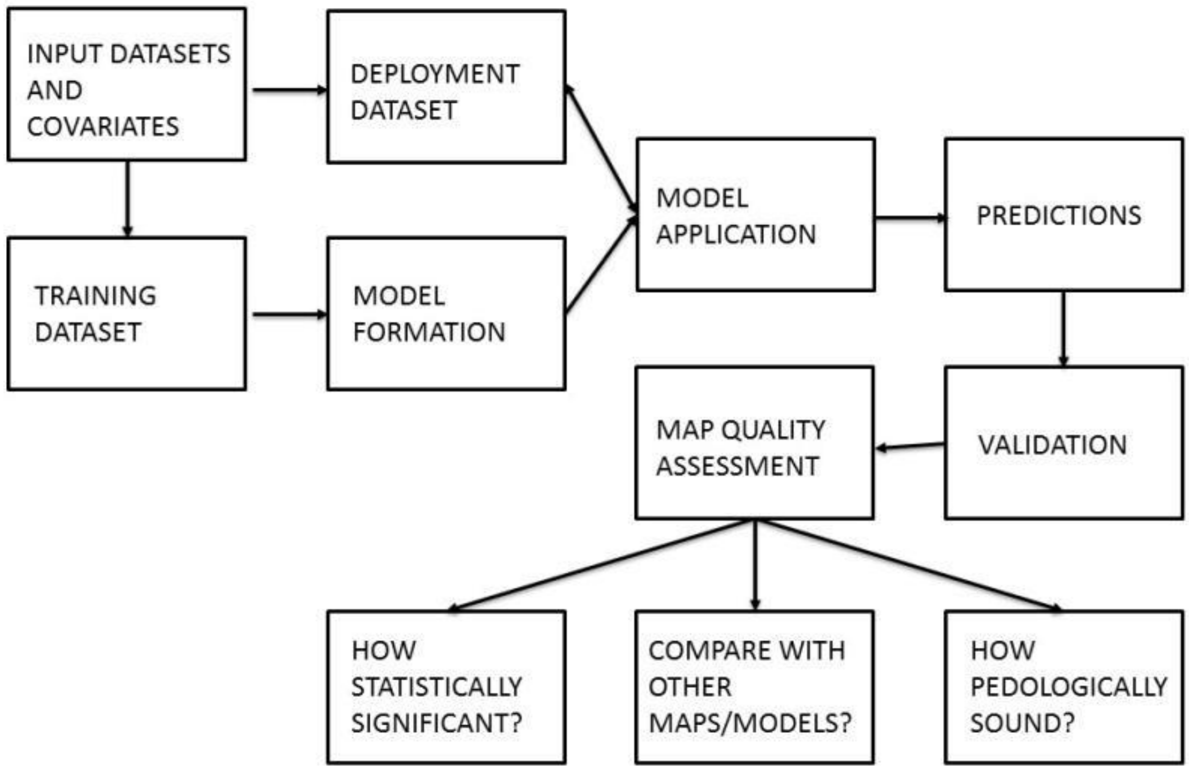


Figure 6.1: DSM process flow chart

As a result, from this DSM process, stakeholders will benefit from an evolution in the selection of the most appropriate DSM or modelling approach (Finke, 2011). This evolution may consist of:

- A decision framework development for choosing appropriate DSM methods in large extent projects which require some standardisation in quality measures of some DSM products.
- The increased application of ensemble DSM methods to make better predictions with narrower uncertainty bands.
- Decision support systems such as Bayesian Belief Networks (BBNs) can be filled with results of multiple concurrent modelling studies (e.g. Taalab et al, 2015).

This way it is important to engage stakeholders at every stage of the DSM process, so they can get an informed understanding of how the soil properties have been obtained and whether they match up with pedological expectations.

6.3. Stakeholder interests and challenges of GB mapping

In terms of maximising the usage of soils data and information, it was important to understand the range of stakeholder needs. Therefore, it was encouraging that many non-expert stakeholders across a range of different organisations answered the questionnaire survey in Chapter 2. The main outcomes from the survey illustrated that there is a requirement for up-to-date, finer scale resolution soils data and information (Campbell et al, 2017). Therefore, new soil property maps at finer scale resolutions need to be improved on what is currently available.

At present, there are different datasets being utilized from various soil surveys and soil taxonomic maps across GB from Scotland and England and Wales and many are hosted by the UK Soil Observatory (UKSO). It is important, therefore, to develop methods in which the soil property datasets across different surveys can be made comparable for GB on a national scale; as Chapter 3 has investigated using two national soil survey datasets. A major reason for this is that a range of different stakeholders could make more effective use of these maps for GB-wide applications, e.g. insurance, construction and governmental organisations. These were some of the lesser-known stakeholders and organisations (i.e. without a direct association with soil science) which answered the questionnaire in Chapter 2.

Digital Soil Mapping (DSM) is proposed as an appropriate toolkit for producing improved resolution soil property maps. A fundamental feature that was found from the questionnaire survey in Chapter 2 highlighted that many stakeholders are still requesting fundamental soil property data. Furthermore, from a critical evaluation of laboratory and analytical techniques in Chapter 3, the DSM modelling was developed across GB using a comparison between two recursive partitioning modelling approaches (BRTs and MARS models) for pH, texture and

LOI to improve the resolution of national soils data. In this process, it is important to evaluate the modelling performance on the outputs based from deployment of the models as well as at the training phase.

From the analyses of soil properties in the pilot study areas in Chapter 4, it was found that MARS models produced a higher R^2 and lower RSME in a training dataset than BRT models for most of the soil properties, at many of the recorded depths. However, although MARS models worked better in a training capacity, when deployed to predict properties at unknown sites, they tended to predict beyond the observed range of values, as illustrated by the mapped outputs in Chapter 4. On the other hand, despite showing weaker statistical indices in the training, the BRT models predicted outputs that were more consistent pedologically when mapped in a wider context. Yet this does not necessarily mean that the full range of soils in these areas was being captured by the model. This led to scaling up from the pilot areas to modelling soil properties across GB using a single regional BRT models and evaluating its performance.

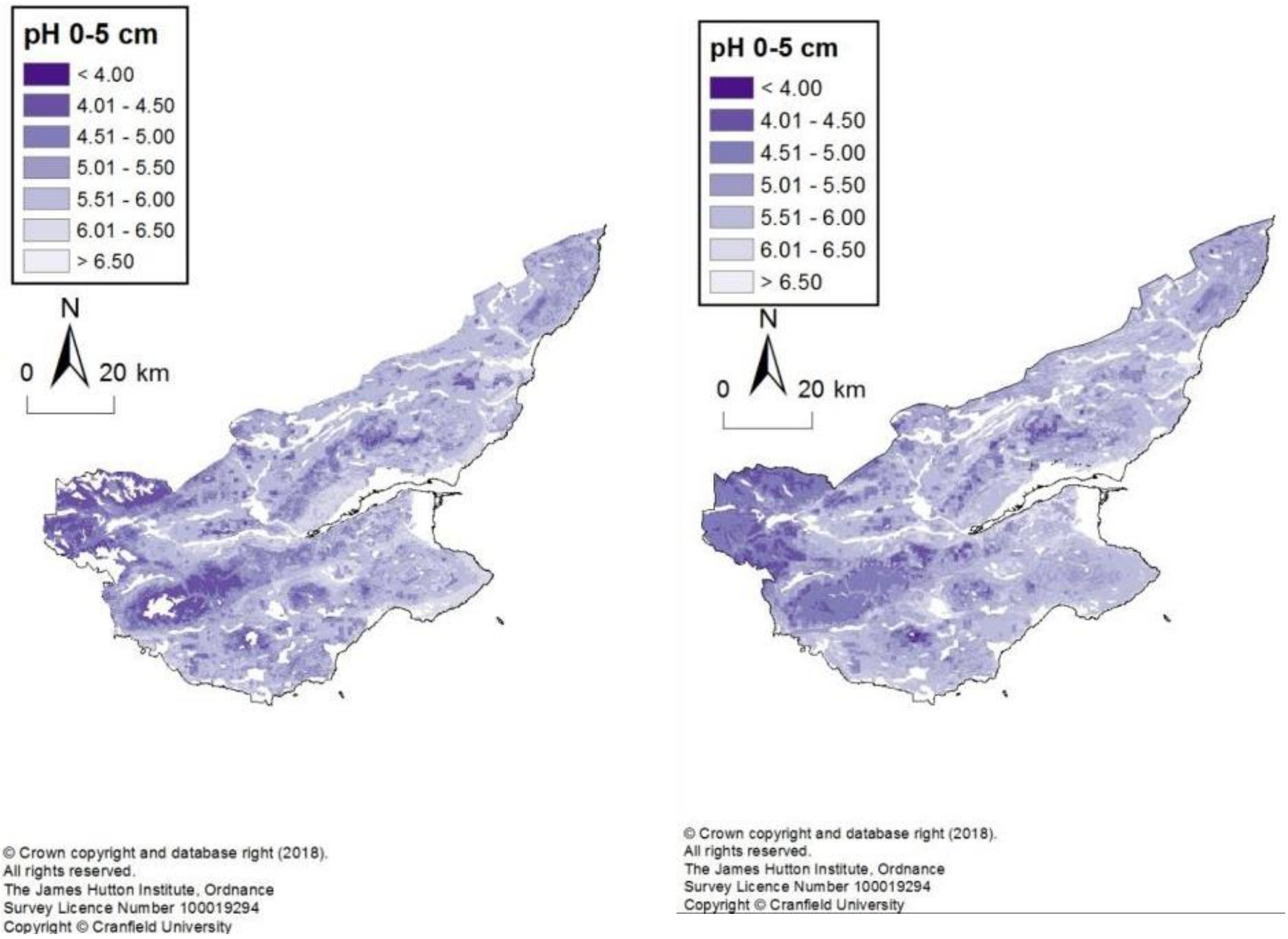
The GB modelling results, as illustrated in Chapter 5, predict pH and LOI reasonably well, and texture not well at all. One major issue is reflected in whether an independent validation dataset provides an appropriate evaluation of model performance. This is because the residual maps from training and validation soil property datasets have produced inconsistent results (see Chapter 5). Therefore, whilst it is acknowledged that stakeholders want to see improved finer scale resolution soil property maps across GB, from the development work done in this PhD it is difficult to say at present whether all requirements for stakeholders have been achieved; and further work still needs to be done to improve on these first versions of predicted soil property maps for GB. Maps of the predicted topsoil soil properties match reasonably well in terms of what we would expect from associated pedology, particularly for pH. However, for values further down the soil profile, the relationships between predicted soil property values and associated pedology becomes dissimilar. This is predominantly due to the number of soil profile points that are found in specific areas, as there are fewer samples at depth within the soil profile databases. Going forward, it would be much more effective to concentrate on areas where there is little data coverage and focusing on more extensive mapping in these areas and collecting more data in the process. In areas where we do have enough information, perhaps utilising other covariates such as MODIS may be an option. Furthermore, conducting feature space analyses would evaluate if the training dataset is representative of the overall covariate space that is used for mapping the whole of GB.

6.4. BRT model comparison in the test areas

In Chapter 4, it was shown that the BRT models predicted outputs were more consistent in pedological terms than MARS models. Thus, a single BRT modelling approach was used to model soil properties across GB with the model performance evaluated in Chapter 5. Table 6.1 reports the model performance for pH at 0-5cm for both test areas using the BRT approach solely within the test areas compared to a single BRT model for GB. The BRT model produced better results for the SCO test area compared with GB predictions (R^2 = 0.64 compared with 0.56; RMSE = 0.63 compared with 0.81). However, the opposite is shown for the EW test area where the R^2 was lower in the pilot area compared to GB modelling though RMSE is lower (R^2 = 0.51 compared to 0.56). The predicted ranges were similar for the SCO test area in comparison to GB but the predicted range for GB modelling was greater than that of the modelling of EW test area. The mapping outputs for pH at 0-5 cm across the SCO test areas shows more acidic pH soils being found in the west and more alkaline soils being found in the east and north. However, although both the pH mapping outputs are different to another, the residual maps produced for the SCO test area are reasonably similar (Figures 6.2a, 6.2b, 6.3a and 6.3b). The same can be said for the EW test area (Fig 6.4a and Fig 6.4b) where the majority of acidic pH soils is in the north, but the maps have produced different outputs when compared with one another. A reason for this might be the different covariates that are being used in each model and whether the most appropriate covariates are being considered in the BRT model. Furthermore, the GB model is likely to be skewed by the Scottish data which is a logical explanation to why the BRT model does not work as well in areas of England and Wales.

pH at 0-5cm	R^2	RMSE	Predicted Range
SCO	0.64	0.63	3.83 – 7.18
EW	0.51	0.73	4.25 – 7.32
GB	0.56	0.81	3.72 – 7.88

Table 6.1: Comparison of two independent pilot areas (SCO & EW) with GB modelling results for pH



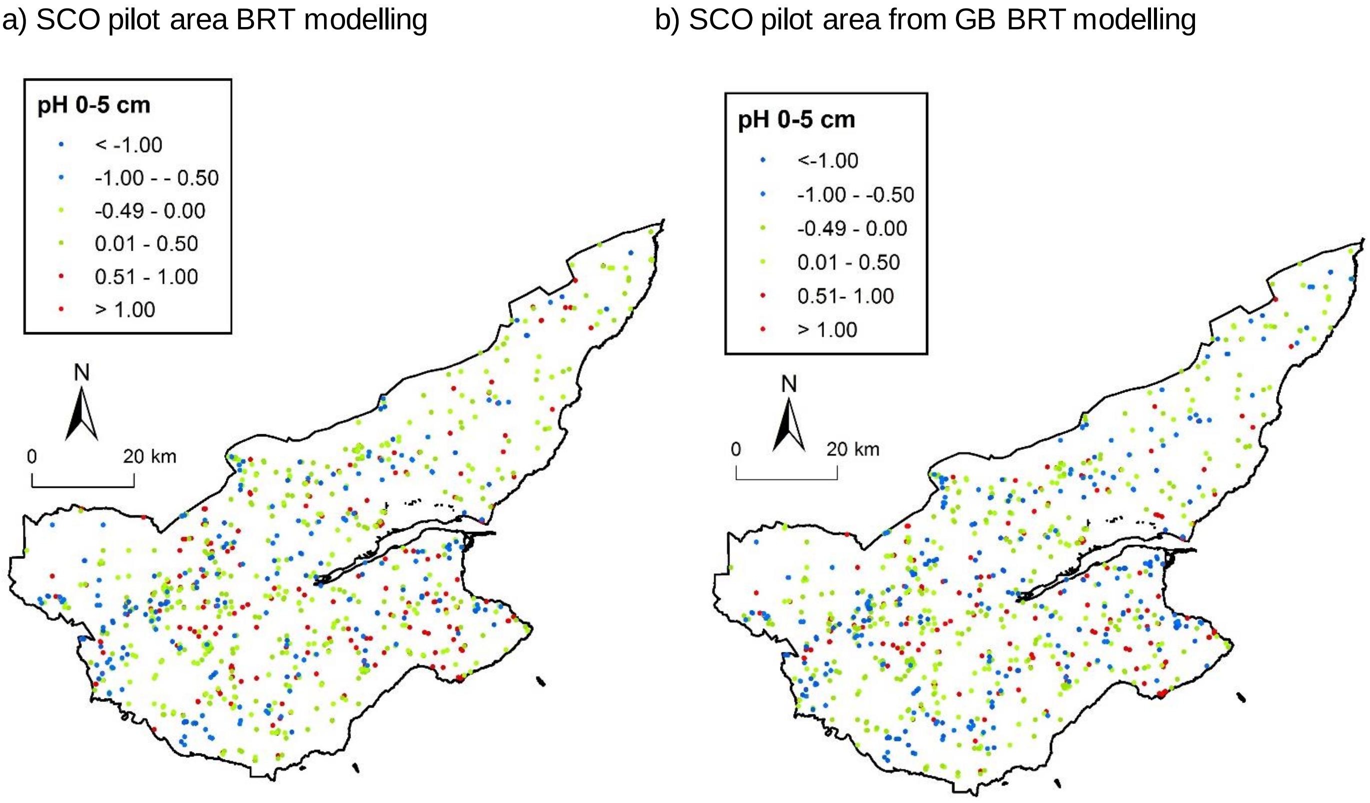
a) SCO pilot area BRT modelling

b) SCO pilot area from GB BRT modelling

© Crown copyright and database right (2018).
All rights reserved.
The James Hutton Institute, Ordnance
Survey Licence Number 100019294
Copyright © Cranfield University

© Crown copyright and database right (2018).
All rights reserved.
The James Hutton Institute, Ordnance
Survey Licence Number 100019294
Copyright © Cranfield University

Figure 6.2: Comparison of pH for SCO area at 0-5cm for a) pilot area modelling and b) GB BRT modelling



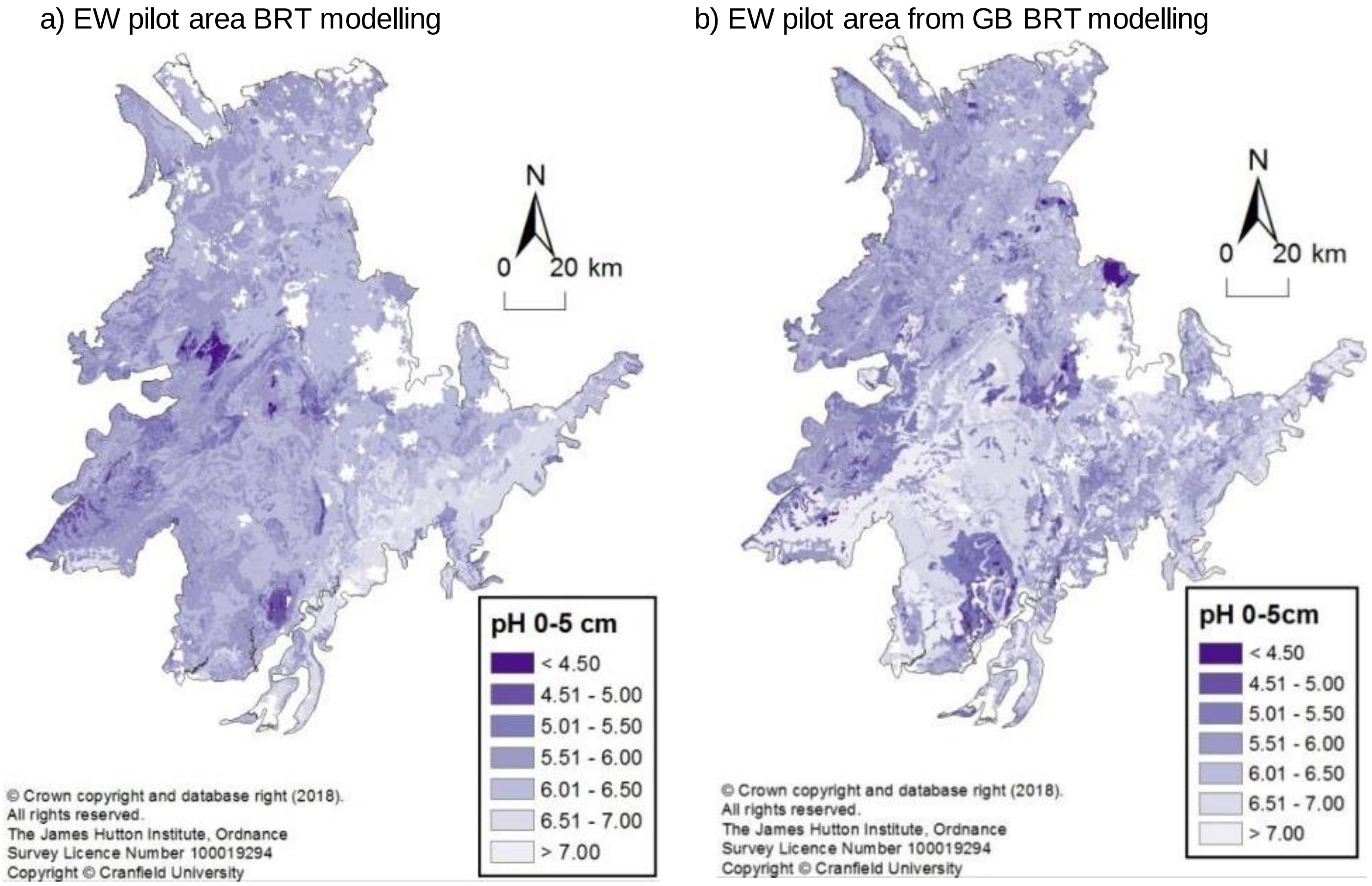
a) SCO pilot area BRT modelling

b) SCO pilot area from GB BRT modelling

© Crown copyright and database right (2018).
All rights reserved.
The James Hutton Institute, Ordnance
Survey Licence Number 100019294
Copyright © Cranfield University

© Crown copyright and database right (2018).
All rights reserved.
The James Hutton Institute, Ordnance
Survey Licence Number 100019294
Copyright © Cranfield University

Figure 6.3: Comparison of pH residuals for SCO area at 0-5cm for a) pilot area modelling and b) GB BRT modelling



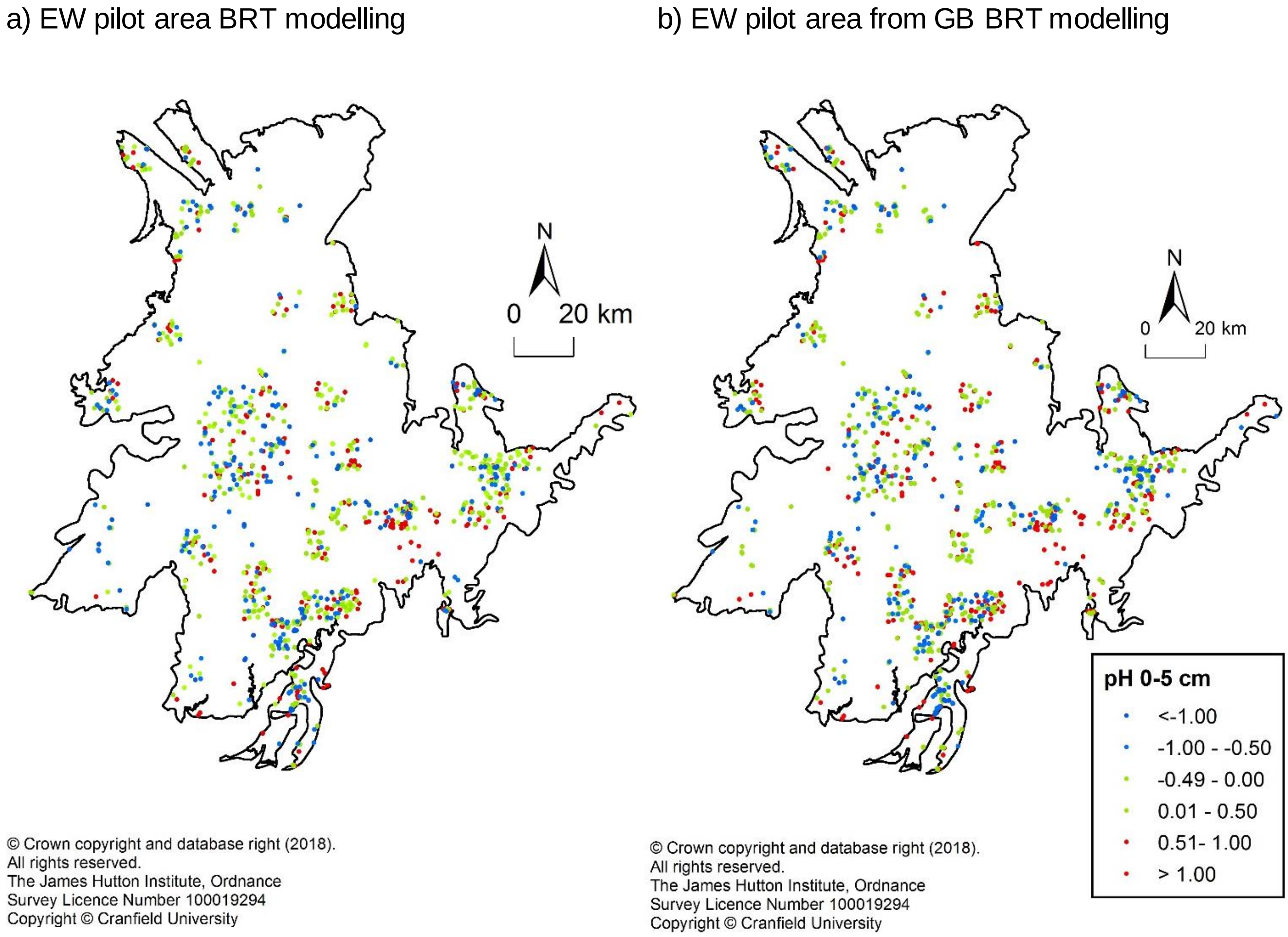
a) EW pilot area BRT modelling

b) EW pilot area from GB BRT modelling

© Crown copyright and database right (2018). All rights reserved. The James Hutton Institute, Ordnance Survey Licence Number 100019294 Copyright © Cranfield University

© Crown copyright and database right (2018). All rights reserved. The James Hutton Institute, Ordnance Survey Licence Number 100019294 Copyright © Cranfield University

Figure 6.4: Comparison of pH for EW area at 0-5cm for a) pilot area modelling and b) GB BRT modelling


a) EW pilot area BRT modelling

b) EW pilot area from GB BRT modelling

© Crown copyright and database right (2018).
All rights reserved.
The James Hutton Institute, Ordnance
Survey Licence Number 100019294
Copyright © Cranfield University

Figure 6.5: Comparison of pH residuals for SCO area at 0-5cm for a) pilot area modelling and b) GB BRT modelling

6.5. Relevance of results to develop functional mapping

The questionnaire survey indicated that several soil properties including texture (sand, clay and silt), pH and carbon/LOI have widespread use for stakeholders (Campbell et al, 2017). It was also noted that these soil properties should be made more accessible for addressing the needs of stakeholders, notably by providing supporting materials to enable data interpretation by non-experts, finer spatial resolution, and indication of trends over time. Many of the tools and assessments utilised by stakeholders in the questionnaire reflect instances where soil properties are used as part of known soil functional assessments and tools, or as individual soil properties.

It is important to understand what soil properties are used to derive these assessments, using information gathered from previous work (GlobalSoilMap, 2011a, GlobalSoilMap, 2011b, Mayr et al, 2006). This is particularly important when understanding the need~~s~~ for future soil functional assessments, and in ensuring that all essential soil properties are considered for improving existing - or developing new - modelling and mapping approaches (Mayr et al, 2006). In the survey, most stakeholders currently used soil property - only assessments utilising information such as soil chemistry and soil carbon. However, others used a range of current functional based assessments which are derived in part from or based upon soil property information. Such examples of models used are Nitrate Vulnerable Zones (NVZs) and agricultural land evaluation-based assessments. These assessments, along with many others, have used inputs such as soil texture and pH as well as factors such as land use, topography, soil depth and climate as well as known soil map units (Mayr et al, 2006). Other models such as the Hydrology of Soil Types (HOST) utilise soils in a different way by distinguishing them based on the parent material and the hydrological characteristics of soils (Gagkas and Lilly, in press). The HOST model is a classification scheme based on national UK soils maps which have been developed using expert knowledge to link hydrology factors with mental models of pathways of flow through the soil profile.

It is crucial to understand what soil properties are used to derive these assessments and what are considered the most important for stakeholders, as these models can be utilised in raising awareness of the range of soil properties which can help to underpin the soil functional assessments. This would help notably in making more informed land management decisions and shaping new related policies. At present, there is a consensus for new sets of models to overcome the limitations of existing models which are either too static or too complex and to deal with future scenarios (Mayr et al, 2006). Therefore, it is useful for soil functional models to be simplified for them to run on currently available soil property datasets and be operational to run spatially. This may be addressed through either adaptation of current models already being used, or the development of new soil functions models based on first principles (Mayr et al, 2006; Lawley et al, 2014). Models or predictive tools are needed for a range of soil functions with notable improvements required for particularly biological habitat (e.g. functional aspects of soil biodiversity) (Blum, 2005).

Improving accessibility of data would certainly benefit non-soil science experts. The UKSO is trying to do this by providing a platform of UK soil datasets from a variety of institutions such as BGS, CEH, The James Hutton Institute and Cranfield University. Furthermore, there is a clear need to address the technical understanding of soils-related data, particularly for knowledge transfer between research and policy, education and training, improving associated supporting information, understanding soil classifications and non-expert user information. A need for more technical knowledge is important because expert judgement will always contain a measure of bias reflecting both knowledge gaps and inherent natural variation, which are difficult to separate (Taalab et al, 2015)

The level of responses within the questionnaire survey suggests that there is demand (and potential) for training opportunities focussed on non-experts and practical applications within soils. Education and training for the modern soil scientist can be improved by developing and delivering short courses which focus on specific knowledge and understanding certain skillsets (Grunwald et al, 2011). This will help to provide opportunities for a range of organisations to be actively involved in advancing DSM and modelling. There are several DSM training workshops which have been set up over the last few years such as ones hosted by International Soil Reference and Information Centre (ISRIC) which shows that support for this is increasing. It should also be important for stakeholders and people working with soils data to gain increased knowledge and understanding working with Geographical Information Systems (GIS) using spatial soil data and information. Without this, it is difficult to see how new spatial soil products, which are predominately GIS in nature, can be widely adopted for practical use. Rossiter (2018) states that in the future no IT restrictions will influence what can be computed. The development of DSM has been concurrent with a rise in computer power (Minasny and McBratney, 2016), and this has led to a large response in the production of products (e.g. GlobalSoilMap) from a range of countries. Tools such as Google Earth Engine has useful in performing all tasks in DSM in one single operation (Padarian et al, 2015).

6.6. Future soil-functional mapping for GB?

Over the last 30 years, DSM has been noted as a credible approach in helping to improve information originally gathered from traditional soil surveying methods. However, DSM should not be the endpoint to completely meet stakeholder needs. DSM can be used as a basis for assessing functions of the soil, critical for all stakeholders. This is known as Digital Soil Assessment (DSA) and is made up of two main processes: a soil attribute spatial system, and an evaluation of soil functions and threats to soils (Carré et al, 2007a). Global BRT models perform well for predicting pH across much of GB but not as well for the other soil properties, therefore, stakeholder needs have only been partially met. Going forward, a stronger pedological understanding is needed within DSM to improve the quality of soil maps for a range of stakeholders to gain maximum benefit from them. However, the difficulty with this is incorporating expert knowledge into the evaluation of GB mapping outputs. This has been done to a certain degree by Taalab et al, (2015) and Angelini et al, (2017) and an approach such as this should be utilised far more in similar instances. As the functions of soils represent various ecological and socio-economical roles of soils, these factors also need to be considered in an assessment processes need to be considered (Haines-Young et al, 2011).

For subsequent DSA outputs to be accepted and utilised, certain issues need to be addressed. The first is to consider whether an assessment or risk map can be used depending on the output uncertainties (Greiner et al, 2018). Secondly, it is crucial to acknowledge the uncertainty for stakeholders who eventually want to use these maps. As some authors argue, to make decisions, (even if a map presents high uncertainties across the landscape) it may provide a good starting point to encourage discussions amongst stakeholders (Carré et al, 2007a; Reed, 2008). A work flow for DSM has been produced in this research (Figure 6.1) and it will be important for the DSM community to be aware of the end-user requirements and tailor their products accordingly (Reed, 2008; Campbell et al, 2017; Carré et al, 2007a). End-users would need to be educated in the use of these information products and its associated uncertainty (Grunwald et al, 2011; Greiner et al, 2018). The information generated by DSM needs to be useable to bring appreciable benefits to many stakeholders across a range of different fields. The modern soil scientist should be able to access and manipulate a range of data to represent soil and other environmental covariates (Grunwald et al, 2011).

Both Carré et al, (2007a) and Finke (2011) have framed their approach to DSA in the context of the European Commission's Soil Thematic Strategy, which is the most developed multi-national approach to soil security and wider environmental sustainability to date. Finke (2011) extends the thinking of DSA by highlighting some important factors to consider:

- There is a need to improve modelling approaches in DSM areas.
- The considerable data demands currently are not being met in global-scale models.
- The economic costs associated with various threats to soil and current challenges to effective mapping.
- The profitable use of uncertainty in risk-associated DSA involving stakeholders, researchers and policy makers.
- The applicable quality criteria for DSM and DSA products.

The future of DSA is reliant on an agreement on:

- The minimum number of datasets required for master soil properties.
- Map constructions at a range of scales focussing on the finest resolution possible and suitable.
- The greatest achievable mapping extent(s) (Finke, 2011).

In the context of GB, there is a consensus towards focussing on improving ecosystem services within DEFRA and other government agencies as well as other ranges of stakeholders (Mayr et al, 2006). The principal user groups are typically associated with agricultural and policy-makers whose main task is to minimise the effects of soil threats such as soil degradation, to preserve and maintain soil health, and to improve food security and household livelihoods (Haines-Young et al, 2012). Other users include research and modelling communities, farming associations, environmental management services and non-governmental organisations; and feedback should be encouraged from stakeholders, with appropriate quality standards developed for the incorporation of such information (Sanchez et al, 2009).

It is important to address the idea of what is good enough for the stakeholders but also coupled with what is deemed an appropriate mapped output by the pedologists and the DSM community. DSM is evolving from a science-driven procedure towards a user-driven process (Carré et al, 2007b), and this is reflected by an increase in DSM projects from small research areas towards regional, national and continental extents (Hartemink and McBratney, 2008; Panagos, 2017). Recently, Sanchez et al, (2009) revealed the GlobalSoilMap.net project, which is looking to produce a fine resolution 3D grid of soil properties similar to the outputs of this thesis.

Quantification of, and dealing with, uncertainty becomes increasingly relevant both in DSM and DSA (Table 6.2). The rich variety in DSM approaches being used (as well as measures to assess the soil prediction quality) is huge and needs standardization (Grunwald, 2009; Heuvelink et al, 2007).

Quality related to	In DSM-context expressed as
Grain size/scale	Positional quality: • Effective map scale • Location and width of boundaries
Input Errors	Analytical quality (measurement errors) • Measurement approach • Object uncertainty (matrix extraction), (sub) sampled volume, interferents, non-homogeneity, storage life
Completeness/need for inference	Data saturation
Semantic correctness	Accuracy: • Thematic maps • Single-value maps
Currency	Quality of legacy data
Logical consistency	Errors due to: • Generalisation • Harmonisation
Lineage	Errors in integration of heterogeneous data due to variation in: • Positional quality • Measurement error • Currency

Table 6.2: Quality aspects of digital soil maps and model studies (Finke, 2011).

6.7. Further work

Going forward, it will be crucial to evaluate the costs of gathering more data in certain areas against the benefit of mathematical models improving the mapping outputs. This trade-off is critical for the stakeholders and due to the large scale of projects like GSM, interactions with end-users should be kept at the forefront of discussions (Reed, 2008). This leads to the community asking questions on how useful DSM will be in the first instance which will dictate how important and relevant DSM is rather than honing the emphasis on trying to produce the best-looking DSM for a property that might not be required. However, this can be difficult as current soil maps may not be useful in their present state, or different people will have differing opinions on certain aspects. As this thesis has shown, it is important to focus on the mathematical outputs of certain models and assess how easily soil properties or functions of the soil can be mapped. In essence, there is still a lot of trial and error within

DSM to create quality-controlled mapping outputs which stakeholders would be comfortable using. At present, there is a great deal of global activity focussed on creating quantitative predictive models and associated digital maps of soil properties and soil functions (Banwart 2011). Ultimately, how useful these models are will depend on the confidence that stakeholders have in the information which is presented in these maps and models. One aspect that needs to be addressed is how best to communicate data uncertainty effectively (Richter et al, 2011). Although improving both availability and spatial resolution of soil maps will be useful and is a requirement of stakeholders, to the benefit of soil science and decision making in policy and land management. the issues of conveying uncertainty still remain a challenge.

Many users see this uncertainty as 'errors' in the data and can be become suspicious of using new information. Since traditional soil mapping and land evaluation are largely deterministic with a single outcome for a specific set of input data, it is likely stakeholders are less aware of existing uncertainties in either the data or boundaries. A DSM approach allows uncertainties to be propagated and visualised throughout. Therefore, in land evaluation tools using soils data, a user can see uncertainties associated with decision making for different options for a specified area of land or to further explore whether predicted soil properties are within expected or potential management ranges. This uncertainty information allows additional insight into the likely suitability of the land which hopefully will lead to more informed stakeholder decisions. The relevance of the uncertainties in decision making will be dependent on the decisions people face and there is further work required to understand different stakeholders' perceptions and uses of uncertainty in decision making (Fischhoff and Davis, 2014).

Currently, these GB soil property maps produced as first version outputs from this work are variable in terms of whether they can be used by stakeholders. Additional discussions with stakeholders, pedometricians and modellers is required to ascertain the suitability of these maps for developing soil functional work in the future.

References

Angelini, M.E., Heuvelink, G.B.M, Kempen, B., 2017: Multivariate mapping of soil with structural equation modelling. *European Journal of Soil Science*, 68, pp. 575-591.

Banwart, S. 2011: Save our soils. Nature, 152, Vol 474, pp. 151-152.

Blum, W.E.H., 2005: Functions of soil for society and the environment. Reviews in *Environmental Science and Bio/Technology*, 4 pp. 75-79.

Brus, D.J., Kempen, B., Heuvelink., 2011: Sampling for validation of digital soil maps. *European Journal of Soil Science*, 62, 3, pp. 394-407.

Campbell, G.A., Lilly, A., Corstanje, R., Mayr, T.R., Black, H.I.J., 2017: Are existing soils data meeting the needs of stakeholders in Europe? An analysis of practice use from policy to field. *Land Use Policy*, 68, pp.211-223.

Carré, F., McBratney, A.B., Mayr, T., Montanarella, L., 2007a: Digital soil assessments: Beyond DSM. *Geoderma*, 142,1–2, pp. 69-79.

Carré, F., McBratney, A.B., Minasny, B., 2007b: Estimation and potential improvement of the quality of legacy soil samples for digital soil mapping, *Geoderma*, 141, 1-2, pp.1-14.

Elith, J., Leathwick, J.R., Hastie, T., 2008: A working guide to boosted regression trees. *Journal of Animal Ecology*, 77: pp. 802–813.

Finke, P.A., 2011: On digital soil assessment with models and the Pedometrics agenda. *Geoderma*, 171–172, pp. 3–15.

Fischhoff, B., Davis, A.L., 2014: Communicating scientific uncertainty. Proceedings of the National Academy of Sciences of the United States of America. 111, 4, pp.13664-13671.

Friedman, J.H., 1991: Multivariate Adaptive Regression Splines, the Annals of Statistics, Vol. 19, No. 1 (Mar. 1991), pp. 1-67.

Gagkas, Z., Lilly, A., (in press): Downscaling soil hydrological mapping used to predict catchment hydrological response with Random Forests.

GlobalSoilMap., 2011a: GlobalSoilMap.net: New Digital Soil Map of the world. [Accessed from https://www.google.co.uk/search?q=globalsoilmap.net&ie=utf-8&oe=utf-8&aq=t&rls=org.mozilla:en-GB:official&client=firefox-a&channel=sb&gfe_rd=cr&ei=aR3jVLWpOoSV8wOK4YHwBQ] [Last accessed 17th February 2015].

GlobalSoilMap., 2011b: Specifications Version 1: GlobalSoilMap.net products: Release 2.1. Technical report.

Greiner, L., Nussbaum, M., Papritz, A., Zimmermann, S., Gubler, A., Grêt-Regamey, A., Keller, A., 2018: Uncertainty indication in soil function maps – transparent and easy-to-use information to support sustainable use of soil resources. SOIL, 4, pp.123–139, 2018; https://doi.org/10.5194/soil-4-123-2018.

Grunwald, S., 2009: Multi-criteria characterization of recent digital soil mapping and modelling approaches; *Geoderma*, 152, 3, pp. 195-207.

Grunwald, S., Thompson, J.A., Boettinger, J.L., 2011: Digital Soil Mapping and Modelling at Continental Scales: Finding Solutions for Global Issues. *Soil Science Society of America Journal*, 75, pp.1201-1213. doi:10.2136/sssaj2011.0025.

Haines-Young, R., Potschin, M., Kienast, F., 2012: Indicators of ecosystem service potential at European scales: Mapping marginal changes and trade-offs, *Ecological Indicators*, 21, pp. 39-53.

Heuvelink, G.B.M., Brown, J.D., van Loon, E.E., 2007: A probabilistic framework for representing and simulating uncertain environmental variables, International Journal of Geographical Information Science, 21, 5, pp.497-513 DOI: 10.1080/13658810601063951 .

Lawley, R., Emmett, B.A., Robinson, D.A., 2014: Soil observatory lets researchers dig deep. *Nature.* 509, pp.427

Mayr, T., Black, H., Towers, W., Palmer, R., Cooke, H., Freeman, M., Hornung, M., Wood, C., Wright, S., Lilly, A., DeGroote, J., Jones, M., 2006: Novel methods for spatial prediction of soil functions within landscapes (SP0531). DEFRA, 26pp.

Minasny, B., McBratney, A.B., 2016: Digital Soil mapping; a brief history and some lessons, *Geoderma*, 264, Part B, pp.301-311.

Padarian, J., Minasny, B., McBratney, A.B., 2015: Using Google's cloud –based platform for digital soil mapping, *Computers & Geosciences*, 83, pp. 80-88.

Piikki, K., Soderstrom, M., 2017: Digital soil mapping of arable land in Sweden – Validation of performance at multiple scales, *Geoderma*, pp. 1-9 http://dx.doi.org/.

Reed, M., 2008: Stakeholder participation for environmental management: a literature review. *Biological Conservation* 141, 10: pp. 2417-2431.

Rossiter, D.G., 2018: Past, present & future of information technology in pedometrics. *Geoderma*, 324, pp.131-137. doi: 10.1016/j.geoderma.2018.03.00.

Sanchez, P. A., Ahamed, S., Carré, F., Hartemink, A.E., Hempel, J., Huising, J., Lagacherie, P., McBratney, A.B., McKenzie, N.G., de Ourdes Mendonça-Santos, M., Minasny, B., Montanarella, L., Okoth, P., Pal, C.A., Sachs, J.D., Shephard, K.D., Vagen, T., Vanlauwe, B., Walsh, M.G., Winowiecki, L.A., Zhang, G.L., 2009: Digital soil map of the world. *Science,* 325, 5941, pp. 680-681.

Stoorvogel, J. J., Kempen, B., Heuvelink, G.B.M., de Bruin, S., 2009: Implementation and evaluation of existing knowledge for digital soil mapping in Senegal. *Geoderma*, 149, 1–2, pp. 161-170.

Taalab, K., Corstanje, R., Zawadzka, J., Mayr, T., Whelan, M.J., Hannam, J.A., Creamer, R., 2015: On the application of Bayesian Networks in Digital Soil Mapping, *Geoderma*, 259-260, pp.134-148.

Yigini, Y., Olmedo, G.F., Reiter, S., Baritz, R., Viatkin, K., Vargas, R., (eds)., 2018: Soil Organic Carbon Mapping Cookbook, 2nd edition. Rome, FAO. 220pp.

LIST OF APPENDICES

Appendix 1 – Hidden Soils Information Questionnaire

This appendix contains the questionnaire that was sent out to stakeholders to consider the range of soils data and information currently being used across Europe with a focus on explicit and hidden soils information being used by non-expert stakeholders. Outcomes from this can be found in Chapter 2 of the PhD thesis.

Dear Respondent,

My name is Grant Campbell and I am doing my PhD with Cranfield University and the James Hutton Institute. My project is investigating the importance of soil information (maps, data etc.) for decision-making, planning and policy development. I would be grateful if you could complete this short questionnaire in as much detail as possible as the results from the questionnaire will help formulate what information about the soil is useful and needed by the community. It will contribute to identifying what characteristics about the soil I intend to map in subsequent PhD work. The information collected will be confidential, completely anonymous, and only the aggregate (average or total) results will be reported for my PhD and in any subsequent scientific publications. It will be retained for the duration of the PhD and stored according to UK data protection regulations. If you would like any more information about the study, please do not hesitate to contact me on g.a.campbell@cranfield.ac.uk or at Grant.Campbell@hutton.ac.uk . The questionnaire should take no longer than 5-10 minutes to complete. You are free to miss out any question or to exit the questionnaire at any time. In most cases, you can answer more than one option (i.e. tick all that apply) as I hope to cover as much detail as possible about the use and effectiveness of information on soil. Thank you.

Please tick this box to acknowledge that you consent to the information you have given to be used for the purpose of the study.

□ I consent to the information being used

Q1 Do you use any information on soil as part of your work?

○ Yes

○ No

Q2 Which of the following best describes the activities of your organisation? Tick all that apply.

❑ Agriculture

❑ Conservation

❑ Construction

❑ Environmental Consultancy

❑ Environmental Advocacy (e.g. NGO's)

❑ Estate or reserve management

❑ Finance/ insurance

❑ Forestry/ woodland

❑ International Agency

❑ Landscape design

❑ Local Authority/ Councils

❑ Local Community (e.g. allotment associations)

❑ National/Federal Government Department or Agency

❑ Planning

❑ Research organisation (university, institutes etc)

❑ Waste management

❑ Water industry

❑ Other (please specify ____________________

Q3 Do you use any of the following information for your project(s)?

❑ Agricultural land evaluation

❑ Biofuel potential

❑ Climate change models

❑ Crop Suitability maps/models

❑ Drainage requirements

❑ Drainage systems (e.g. SUDS)

❑ Drought risk assessments

❑ Erosion risk assessments

❑ Extraction of raw materials (peat, sands, gravels, clays etc.)

❑ Fertiliser and pesticide usage

❑ Flood risk maps

❑ Habitat suitability

❑ Hydrology of Soil (e.g. HOST)

❑ Infrastructure assessment (pipes/electric cables etc.

❑ Irrigation requirements

❑ Land reclamation/restoration

❑ Land Suitability for Forestry

❑ Land Suitability for Housing

❑ Land use change modelling

❑ Leaching risk maps

❑ Micronutrient levels

❑ Nutrient Vulnerable Zones (e.g. Nitrate Vulnerable Zones (NVZs)

❑ Nutrient cycling

❑ Pesticide safety assessment

❑ Pollen counts

❑ Pollutant in soil

❑ Protection of animal species

❑ Reclamation of contaminated land

❑ Recreational space (e.g. green space, allotments)

❑ Recycling waste to land

❑ Runoff potential

❑ Sludge acceptance

❑ Soil acidity/alkalinity levels

❑ Soil borne diseases and/or pests

❑ Soil carbon/organic carbon

❑ Soil chemistry

❑ Soil erosion

❑ Soil moisture

❑ Soil pathogens

❑ Soil temperature

❑ Water pollution

❑ Other (please specify) ____________________

Q4 Of the following sectors, which are the most relevant to your work?

□ Agricultural production

□ Biofuel production

□ Building/ infrastructure

□ Climate change mitigation

□ Conservation of habitats and biodiversity

□ Contaminated land

□ Cultural heritage or archaeology

□ Environmental Impact Assessments

□ Extraction of raw materials (e.g. peat, sands, gravels, clays)

□ Flood regulation

□ Forestry production

□ Land use planning

□ Pests and diseases

□ Recreation (e.g. amenity woodland, tourism)

□ Recycling organic waste to land

□ Water supply and/or quality

□ Other (please specify) ____________________

Q5 Are you aware that the information you may use in your work has soils information embedded within it?

○ Yes, I was aware

○ No, I was not aware

○ I was not sure

Q6 What source(s) do you use to acquire the information you need?

□ Books/reports

□ Databases

□ Expert knowledge

□ Field analyses

□ Geographical Information Systems (GIS)

□ Maps (paper or digital)

□ Websites

□ Other (please specify ____________________

Q7 How accurate or useful do you find the available information for your purposes?

	Not very useful	Not useful	Useful	Very useful
Books/reports	❍	❍	❍	❍
Databases	❍	❍	❍	❍
Expert Knowledge	❍	❍	❍	❍
Field analyses	❍	❍	❍	❍
Geographical Information Systems (GIS)	❍	❍	❍	❍
Maps (paper or digital)	❍	❍	❍	❍
Websites	❍	❍	❍	❍
Other (please specify)	❍	❍	❍	❍

Q8 Does your organisation pay for the licence use of any of the information you have identified?

○ Yes

○ No

○ Don't Know

Q9 In your own opinion, what improvements could be made to make the information you use already more effective?

□ Associated documentation made available

□ Contemporary data

□ Co-ordinates of the geographical locations

□ Finer scale/resolution

□ Geographic projection

□ Greater coverage of the map

□ Improved accuracy/credibility of data sources

□ Meta data information

□ Methodology for data generation

□ Pixel/polygon-based information

□ Predicting change to drivers (e.g. climate change)

□ Relaxation of copyright

□ Summary interpretation for non-scientific users

□ Summaries of uncertainty/error values

□ Trends over time

□ Understanding/ additional information to help with soil classification scheme

□ Other (please specify) ____________________

Q10 Would you be interested in using any new information that might arise from an improvement in...

	Yes	No
Spatial resolution/scale	❍	❍
Summary of uncertainty/error values	❍	❍
Other (please specify)	❍	❍

Q11 How would you rate the importance of spatial soil information for wider applications and end users?

❍ Very important

❍ Important

❍ Not important

❍ Not very important

Q12 Is there any other information you wish to add that has not been discussed in the survey?

Appendix 2 – Harmonisation methods

This appendix contains additional information on the harmonisation methods and techniques attempted for LOI, pH and texture which were used in this PhD.

LOI/OC

Various techniques have been used for calculating soil organic carbon (OC). The Walkley Black method (WB) has been the most documented as it was one of the earliest ways of obtaining OC data (DEFRA, 2011). The soil organic carbon is oxidised by reacting potassium dichromate ($K_2Cr_2O_7$) in sulphuric acid (H_2SO_4). In England and Wales, soil organic carbon was measured by a modified WB method for the original samples which was carried out in 1978-1983 (Kalembasa & Jenkinson 1973; DEFRA, 2011). Sampling was restricted to the uppermost part of the soil.

In the laboratory, the contents were placed in a furnace at 850°C for 30mins. After ignition, the basin and contents are cooled in desiccator and reweighed. If the soil was calcareous, it was heated at 950°C for 2 hours to complete decompose the carbonates. LOI values (gkg^{-1}) have been converted to OC by the following equation (DEFRA, 2011)

$$OC = 0.5 \times LOI$$

For Scotland, soil organic carbon and LOI were measured (Macaulay Institute for Soil Research, 1971). Between 5 and 10g of 2mm air-dry soil was added to a crucible and weighed. These samples were then placed in an oven at 105°C for around 3 hours. The crucible and its contents were then removed, cooled and weighed. These samples were then placed in a muffle furnace and heated to between 800-900°C for around 2 hours. After cooling, these samples are then transferred to a desiccator, cooled and weighed again.

pH (H2O)

In Scotland, the pH of water is measured on a soil to water ratio of 1:3 (Macaulay Institute for Soil Research, 1971). After the soil and water have been mixed together, the solution is left to stand for 4 hours and the pH is recorded using an electrode probe. For measuring the pH of $CaCl_2$, 5 cm^3 of 0.01M $CaCl_2$ is added to a suspension where the mixture is shaken and left for a couple of hours before reading the pH. In England and Wales, measurements are made with 1:2.5 w/v suspensions in water and in 0.01M $CaCl_2$ (Avery and Bascomb, 1979). For mineral soils of <2mm mineral air-dry soil, 10g air-dry soil is added to a 50ml beaker. 25ml distilled water is then added to the beaker from which the contents are stirred and left for 10 minutes before the pH is recorded. To measure pH in calcium chloride values, 2ml 0.125M $CaCl_2$ was added to solution by pipette to reach effective concentration by 0.01M $CaCl_2$. This was then stirred, and the pH was measured and recorded using an electrode probe.

Particle Size Distribution

Fitted curves to Cumulative Particle Size Distributions

Different particle size classes have been used in Scotland in comparison to England and Wales. In England and Wales, the particle size classes were predominantly based on the British Soil Texture Classification (BSTC) (Avery and Bascomb, 1979) although some data were in the USDA particle size classes. In Scotland, particle size classes mainly followed the United States Department of Agriculture (USDA) size classes (USDA, 1978). The primary difference between the BSTC and the USDA classes is the cut off used between the silt and sand fractions. The USDA particle size classes are ((<2 (clay), 2-50 (silt), 50-2000 (sand) in µm)) and for BSTC they are ((<2 (clay), 2-60 (silt), 60-2000 (sand) µm).

The first potential harmonisation method that was investigated was curve fitting and interpolation (Nemes et al, 1999). The data used in the investigation were derived from the National Soil Inventory of Scotland as a wide range of particle size classes had been determined using laser diffraction including both USDA and BSTC particle size classes so provide an opportunity to investigate the possibility of converting PSD from USDA to BSTC (as majority of E&W data as BSTC). The data were split into the following fractions: <=2, 6.3, 20, 63, 200, 630 and 200 microns, summed together to make up to 100% and then divided by 1000 to change these measurements from micrometres to millimetres. The log measurement of the fractions to the base 2 was then calculated. A selection of these cumulative curves is illustrated in the graph below (Figure 2.1)

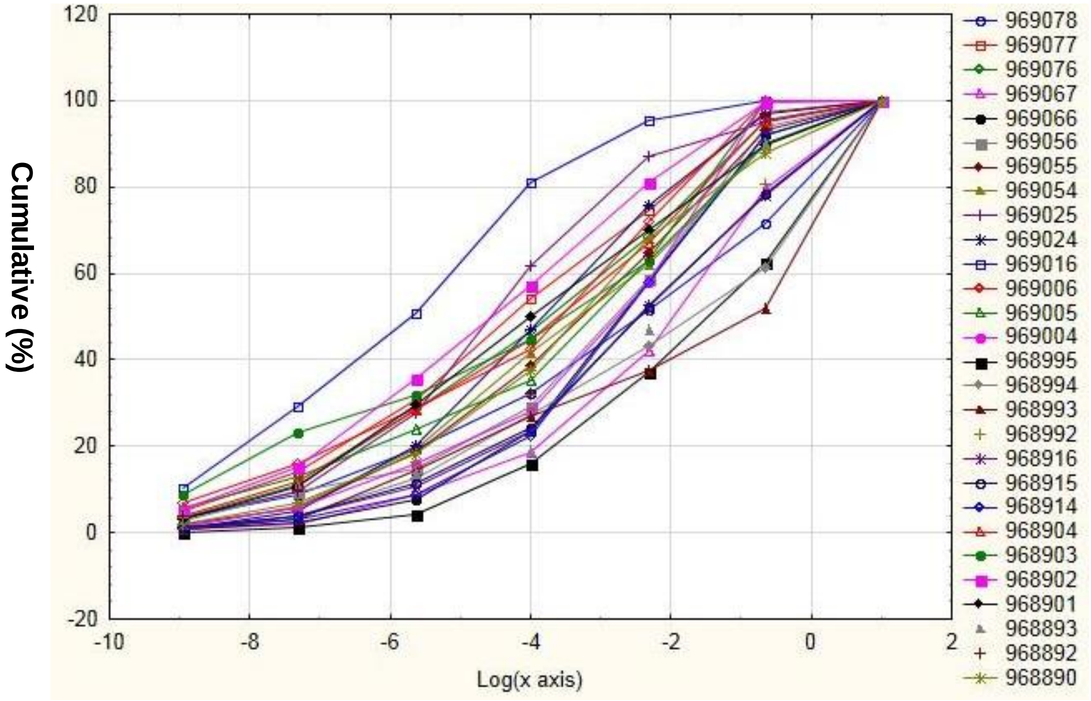


Figure 2.1: PSD cumulative curves for 30 of the 300 NSIS particle size data investigated.

What is evident from the graph above is that the shapes of the curves are very different and are determined by the distribution of the size ranges. Some of the samples measured have considerably larger sand contents and some will conversely see large average clay and silt contents. This mirrors similar results found by Nemes et al, (1999). Furthermore, for the Scotland data, due to the complexity of PSDs, there was no single equation that fitted the data from which the particles in the size range 2-60 µm could be predicted to align the Scottish data with that of England and Wales. Thus, if this approach was to be used, specific equations would need to be derived for a specific sample and that would be very time consuming to achieve.

Appendix 3 – BRT modelling results for soil properties in SCOT pilot area

This appendix contains additional statistical information on the soil properties used for BRT modelling at the SCOT test area which can be found in Chapter 4 of the PhD.

a)

LOI	Depths	R^2	RMSE	Mean	SD	Observed LOI range	Model LOI range	Samples (n)
	0-5	0.61	16.48	21.87	25.09	0.43 – 100.00	8.92 – 76.80	949
	5-15	0.58	12.60	15.98	17.84	0.45 – 96.81	10.02 – 55.81	948
	15-30	0.54	8.25	9.28	10.99	0.25 – 95.81	5.85 – 36.11	941
	30-60	0.48	6.71	5.56	8.46	0.31 – 99.79	2.22 – 29.33	929
	60-100	0.29	6.95	4.21	7.54	0.22 – 99.44	3.02 – 13.40	829
	100-200	0.49	2.01	3.33	2.45	0.35 – 21.63	2.59 – 5.80	361

b)

pH	Depths	R^2	RMSE	Mean	SD	Observed PH range	Model PH range	Samples (n)
	0-5	0.57	0.69	5.44	1.02	3.30 – 8.58	3.92 – 7.10	949
	5-15	0.58	0.66	5.47	0.99	3.30 – 8.58	4.03 – 7.30	948
	15-30	0.56	0.62	5.56	0.90	3.42 – 8.26	4.34 – 6.90	932
	30-60	0.54	0.60	5.70	0.85	3.42 – 8.36	4.64 – 7.20	920
	60-100	0.50	0.63	5.85	0.85	3.23 – 8.74	4.82 – 7.10	829
	100-200	0.53	0.71	6.03	0.92	2.84 – 8.63	5.23– 6.90	361

c)

Sand	Depths	R^2	RMSE	Mean	SD	Observed Sand range	Model Sand range	Samples (n)
	0-5	0.64	10.40	53.43	16.53	0.00 – 100.00	16.08 – 94.43	892
	5-15	0.66	10.00	53.68	16.62	0.00 – 100.00	13.32 – 97.77	892
	15-30	0.63	10.79	55.29	17.11	0.00 – 100.00	13.56 – 95.76	890
	30-60	0.56	13.16	58.90	18.93	0.00 – 100.00	17.88 – 90.18	879
	60-100	0.55	14.36	60.19	20.39	0.00 – 100.00	17.23 – 89.62	781
	100-200	0.63	15.56	62.25	22.83	0.00 – 99.22	23.77 – 91.75	325

d)

Silt	Depths	R^2	RMSE	Mean	SD	Observed Silt range	Model Silt range	Samples (n)
	0-5	0.58	8.34	28.82	12.22	0.00 – 84.97	6.99 – 55.80	892
	5-15	0.59	8.19	28.69	12.15	0.00 – 78.00	6.09 – 55.95	892
	15-30	0.58	8.09	27.65	12.02	0.00 – 78.07	5.18 – 55.29	890
	30-60	0.55	8.33	25.28	12.57	0.00 – 77.61	6.08 – 54.30	879
	60-100	0.53	9.77	24.50	13.41	0.00 – 80.83	8.35 – 51.90	781
	100-200	0.50	12.62	23.86	15.14	0.00 – 71.31	15.15 – 36.84	325

e)

Clay	Depths	R^2	RMSE	Mean	SD	Observed Clay range	Model Clay range	Samples (n)
	0-5	0.44	6.31	13.61	8.05	0.00 – 41.97	5.30 – 26.23	892
	5-15	0.48	6.03	13.83	8.03	0.00 – 44.96	4.51 – 29.06	892
	15-30	0.49	6.37	14.54	8.50	0.00 – 48.86	3.71 – 31.08	890
	30-60	0.48	7.22	14.69	9.53	0.00 – 57.77	4.06 – 31.77	879
	60-100	0.46	7.81	14.71	9.89	0.00 – 47.89	5.54 – 27.68	781
	100-200	0.51	8.32	13.31	10.11	0.00 – 48.99	8.34 – 20.67	325

Appendix 4 – BRT modelling results for soil properties in EW pilot area

This appendix contains additional statistical information on the soil properties used for BRT modelling at the EW test area which can be found in Chapter 4 of the PhD.

a)

LOI	Depths	R^2	RMSE	Mean	SD	Observed LOI Range	Model LOI range	Samples (n)
	0-5	0.47	7.39	9.91	9.17	2.74 – 88.43	8.29 – 28.58	936
	5-15	0.36	4.63	8.22	5.56	2.61 – 66.80	6.64 – 23.14	917
	15-30	0.29	4.34	6.08	4.64	0.00 – 79.15	5.64 – 10.06	839
	30-60	0.14	5.82	4.21	5.96	0.00 – 99.15	3.93 – 6.96	585
	60-100	N/A	N/A	4.12	6.00	0.00 – 66.04	N/A	228
	100-200	N/A	N/A	4.83	8.68	0.94 – 66.01	N/A	74

b)

pH	Depths	R^2	RMSE	Mean	SD	Observed pH Range	Model pH range	Samples (n)
	0-5	0.51	0.73	5.96	1.00	2.34 – 8.36	4.25 – 7.36	1096
	5-15	0.53	0.71	6.01	0.98	3.12 – 8.75	4.25 – 7.36	1095
	15-30	0.52	0.72	6.17	0.99	3.32 – 8.56	4.56 – 7.43	1092
	30-60	0.57	0.71	6.40	1.05	2.74 – 8.75	4.64 – 7.98	1063
	60-100	0.28	1.08	6.79	1.17	2.26 – 9.02	5.92 – 7.27	1063
	100-200	0.60	0.81	6.77	1.10	4.08 – 8.75	5.56 – 7.62	371

c)

Sand	Depths	R^2	RMSE	Mean	SD	Observed Sand Range	Model Sand range	Samples (n)
	0-5	0.73	12.73	36.70	23.70	0.00 – 93.14	9.51 – 76.82	1020
	5-15	0.73	12.75	36.86	23.84	0.00 – 94.65	9.39 – 76.10	1019
	15-30	0.72	13.44	37.11	24.55	0.00 – 96.96	8.77 – 77.64	1013
	30-60	0.68	15.30	36.45	26.52	0.00 – 96.78	9.30 – 80.59	987
	60-100	0.65	17.95	36.04	29.09	0.00 – 99.41	10.70 – 79.98	902
	100-200	0.67	19.74	39.00	31.35	0.00 – 100.00	17.48 – 69.83	347

d)

Silt	Depths	R^2	RMSE	Mean	SD	Observed Silt Range	Model Silt range	Samples (n)
	0-5	0.73	9.17	38.49	17.24	3.00 – 80.74	14.29 – 66.70	1020
	5-15	0.74	9.10	38.30	17.27	2.76 – 80.16	12.78 – 67.06	1019
	15-30	0.73	9.34	37.77	17.49	0.03 – 83.09	12.60 – 66.31	1013
	30-60	0.70	10.14	36.56	18.07	1.63 – 81.50	12.30 – 66.09	987
	60-100	0.64	12.06	35.94	19.48	0.17 – 87.53	11.95 – 62.71	902
	100-200	0.69	13.59	35.47	21.94	0.00 – 98.71	13.63 – 61.89	347

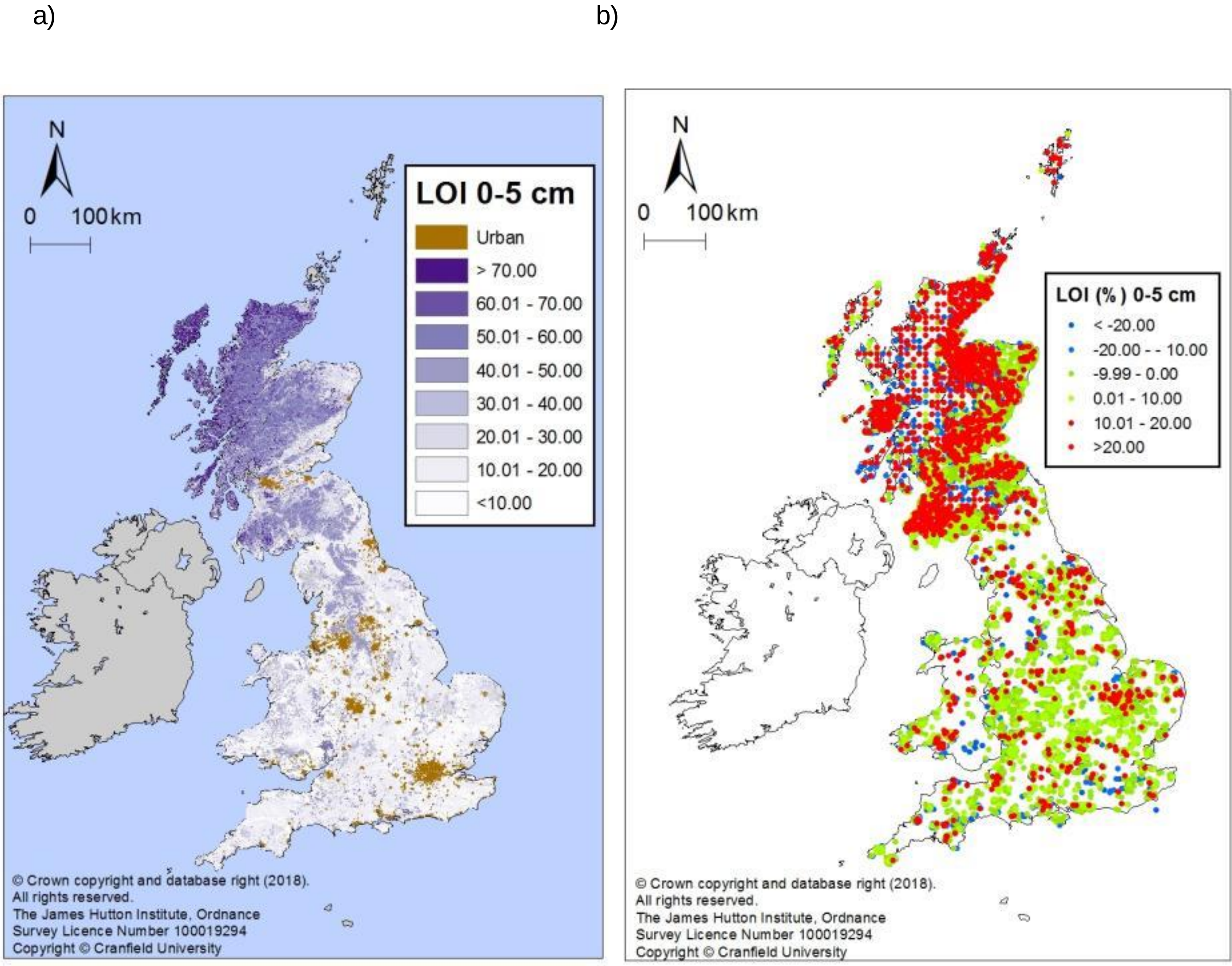
e)

Clay	Depths	R^2	RMSE	Mean	SD	Observed Clay Range	Model Clay range	Samples (n)
	0-5	0.62	9.49	24.88	14.33	0.00 – 89.00	14.76 – 52.11	1020
	5-15	0.62	9.24	24.92	14.32	0.13 – 89.00	14.25 – 52.44	1019
	15-30	0.60	10.05	25.14	15.16	0.59 – 89.00	13.60 – 52.37	1013
	30-60	0.59	11.09	16.69	26.98	1.04 – 89.00	12.01 – 53.87	987
	60-100	0.57	12.12	28.11	17.36	0.06 – 89.00	12.38 – 52.33	902
	100-200	0.52	13.65	25.83	17.31	0.00 – 97.43	14.83 – 43.15	347

Appendix 5 – LOI maps created by a) Boosted Regression Trees and b) associated residual maps at depth

This appendix contains maps of LOI at depth for GB along with associated residual maps. Information on the interpretation of this can be found in Chapter 5.

<u>0-5 cm</u>

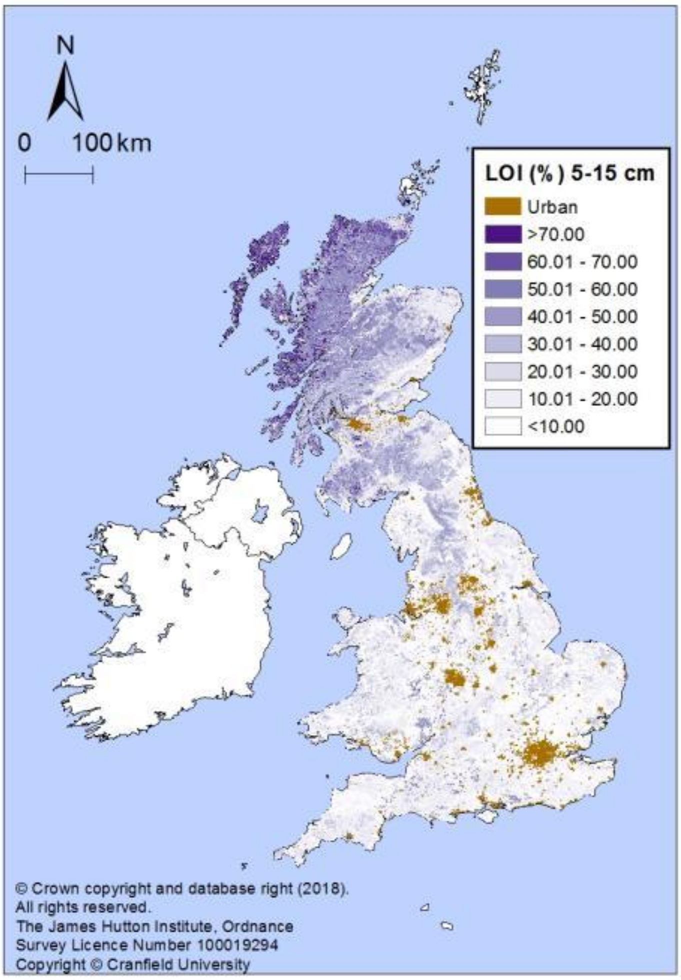


© Crown copyright and database right (2018). All rights reserved. The James Hutton Institute, Ordnance Survey Licence Number 100019294 Copyright © Cranfield University

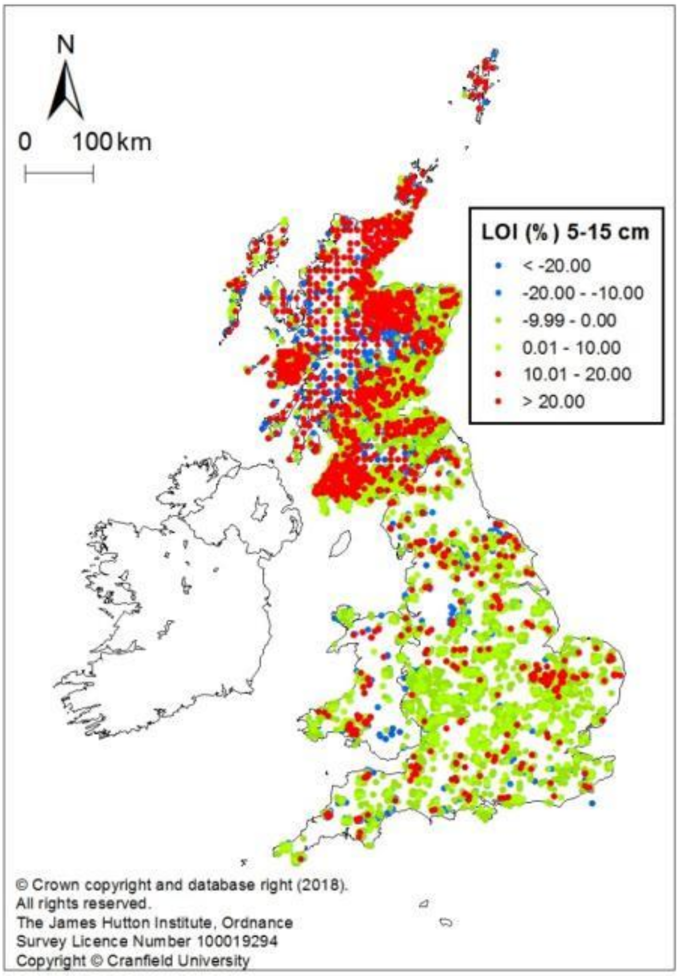
© Crown copyright and database right (2018). All rights reserved. The James Hutton Institute, Ordnance Survey Licence Number 100019294 Copyright © Cranfield University

5-15 cm

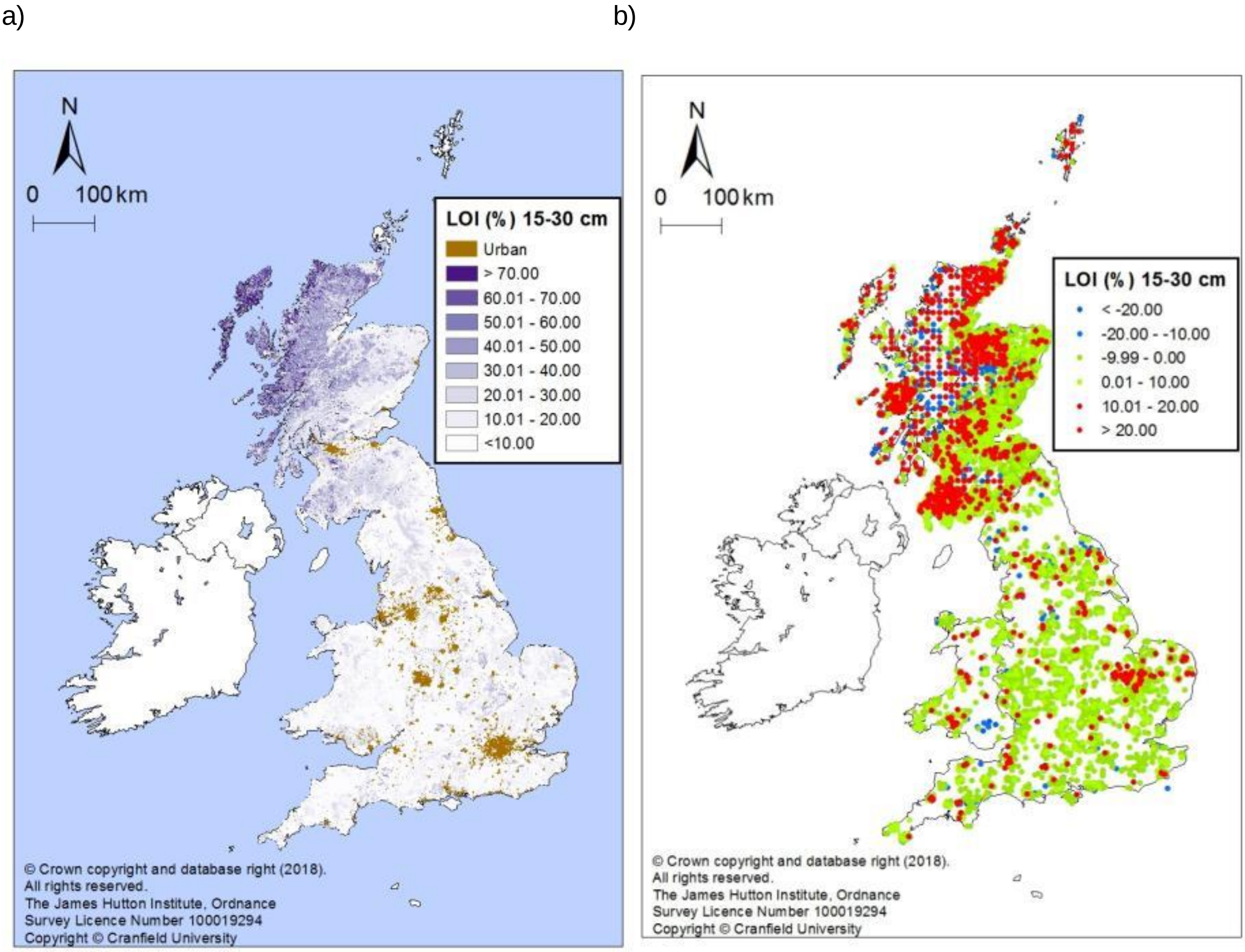
a)


N
0 100 km
LOI (%) 5-15 cm
Urban
>70.00
60.01 - 70.00
50.01 - 60.00
40.01 - 50.00
30.01 - 40.00
20.01 - 30.00
10.01 - 20.00
<10.00
© Crown copyright and database right (2018).
All rights reserved.
The James Hutton Institute, Ordnance
Survey Licence Number 100019294
Copyright © Cranfield University

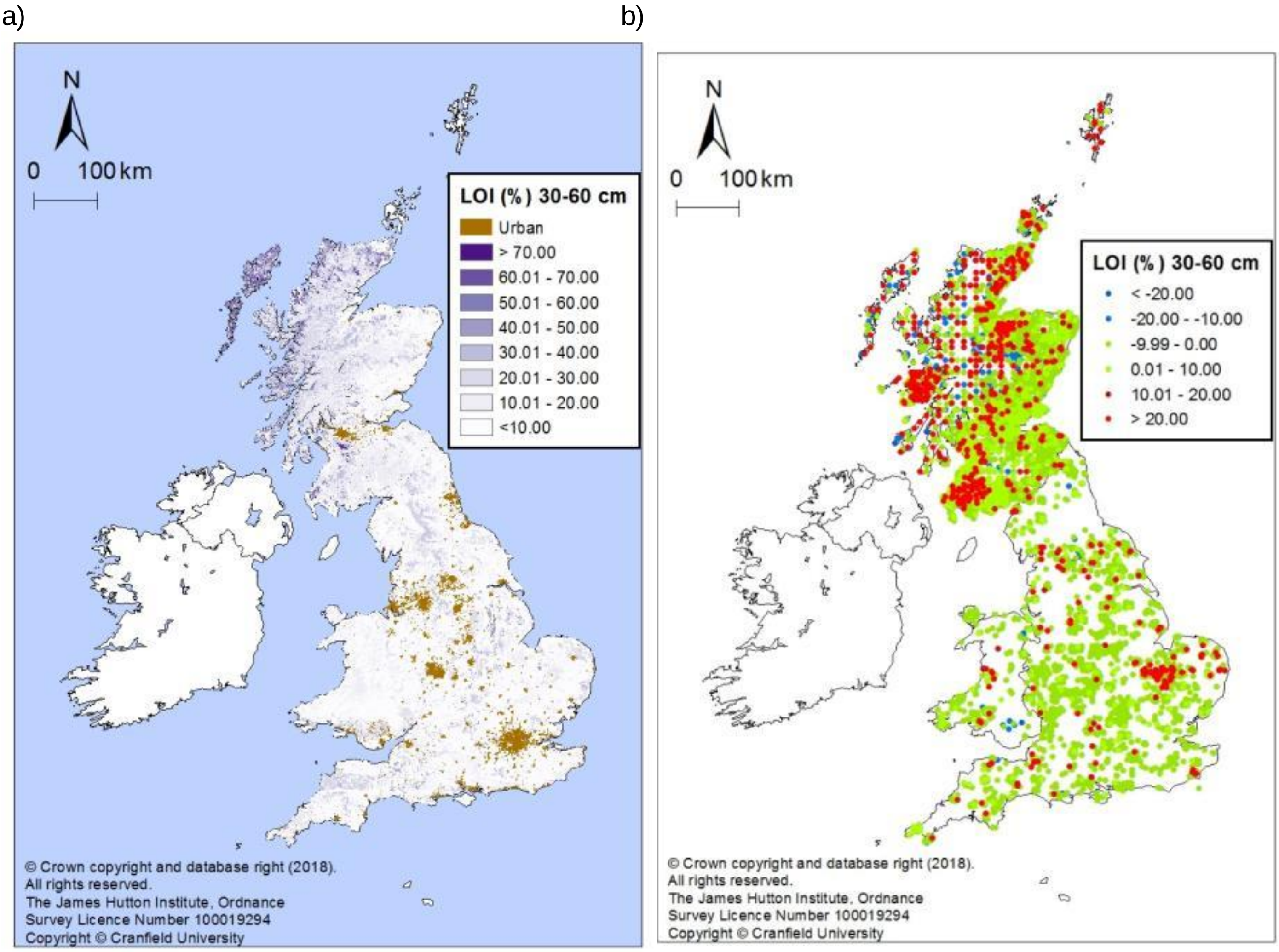
b)


N
0 100 km
LOI (%) 5-15 cm
< -20.00
-20.00 - -10.00
-9.99 - 0.00
0.01 - 10.00
10.01 - 20.00
> 20.00
© Crown copyright and database right (2018).
All rights reserved.
The James Hutton Institute, Ordnance
Survey Licence Number 100019294
Copyright © Cranfield University

15-30 cm

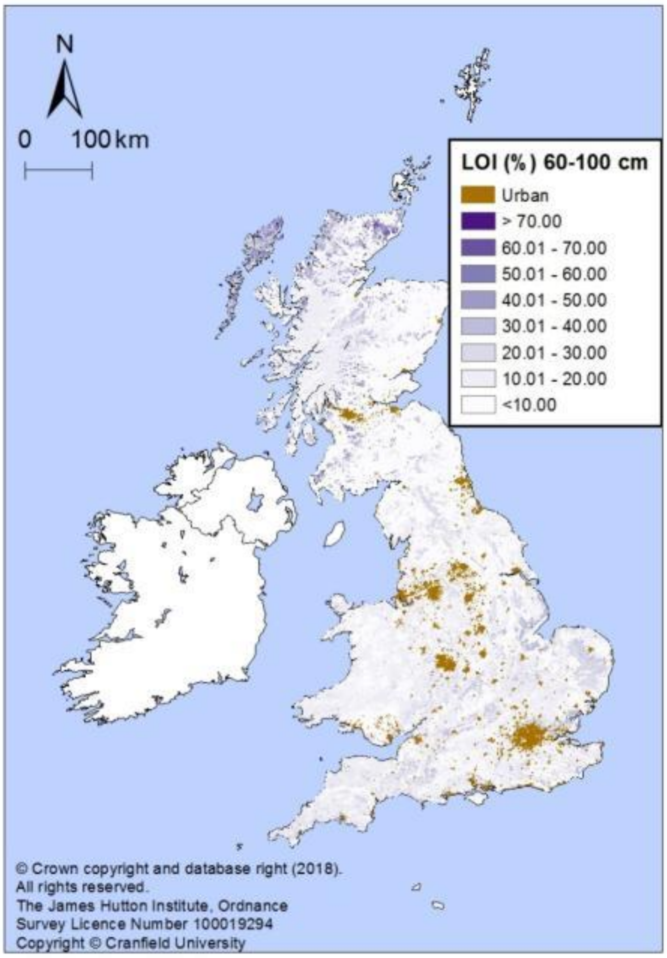

a)
N
0 100 km
LOI (%) 15-30 cm
Urban
> 70.00
60.01 - 70.00
50.01 - 60.00
40.01 - 50.00
30.01 - 40.00
20.01 - 30.00
10.01 - 20.00
<10.00
© Crown copyright and database right (2018).
All rights reserved.
The James Hutton Institute, Ordnance
Survey Licence Number 100019294
Copyright © Cranfield University
b)
N
0 100 km
LOI (%) 15-30 cm
< -20.00
-20.00 - -10.00
-9.99 - 0.00
0.01 - 10.00
10.01 - 20.00
> 20.00
© Crown copyright and database right (2018).
All rights reserved.
The James Hutton Institute, Ordnance
Survey Licence Number 100019294
Copyright © Cranfield University

30-60 cm

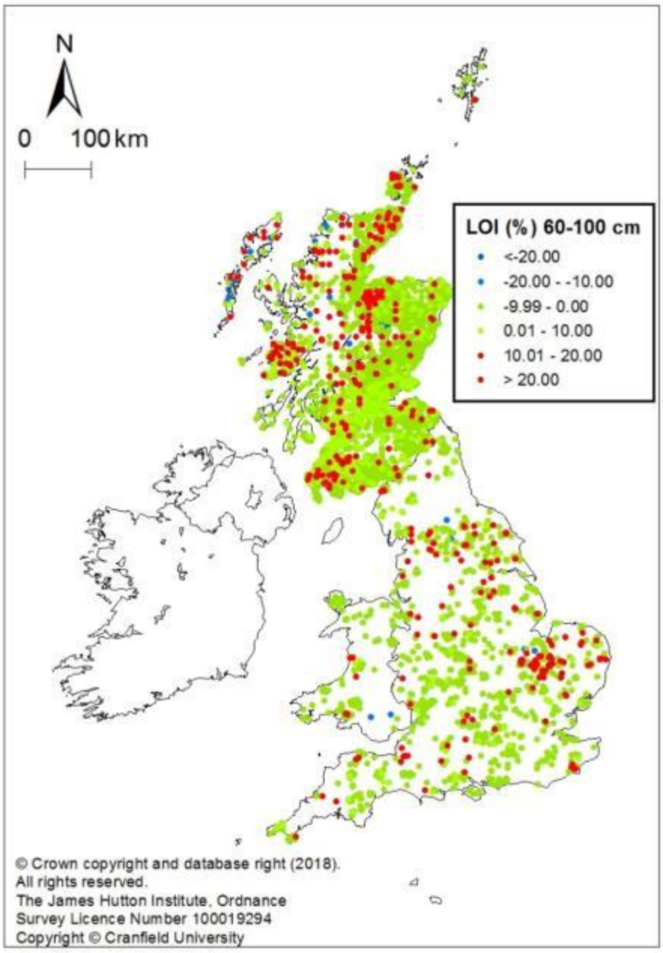

a)
N
0 100 km
LOI (%) 30-60 cm
Urban
> 70.00
60.01 - 70.00
50.01 - 60.00
40.01 - 50.00
30.01 - 40.00
20.01 - 30.00
10.01 - 20.00
<10.00
© Crown copyright and database right (2018).
All rights reserved.
The James Hutton Institute, Ordnance
Survey Licence Number 100019294
Copyright © Cranfield University
b)
N
0 100 km
LOI (%) 30-60 cm
< -20.00
-20.00 - -10.00
-9.99 - 0.00
0.01 - 10.00
10.01 - 20.00
> 20.00
© Crown copyright and database right (2018).
All rights reserved.
The James Hutton Institute, Ordnance
Survey Licence Number 100019294
Copyright © Cranfield University

60-100 cm

a)

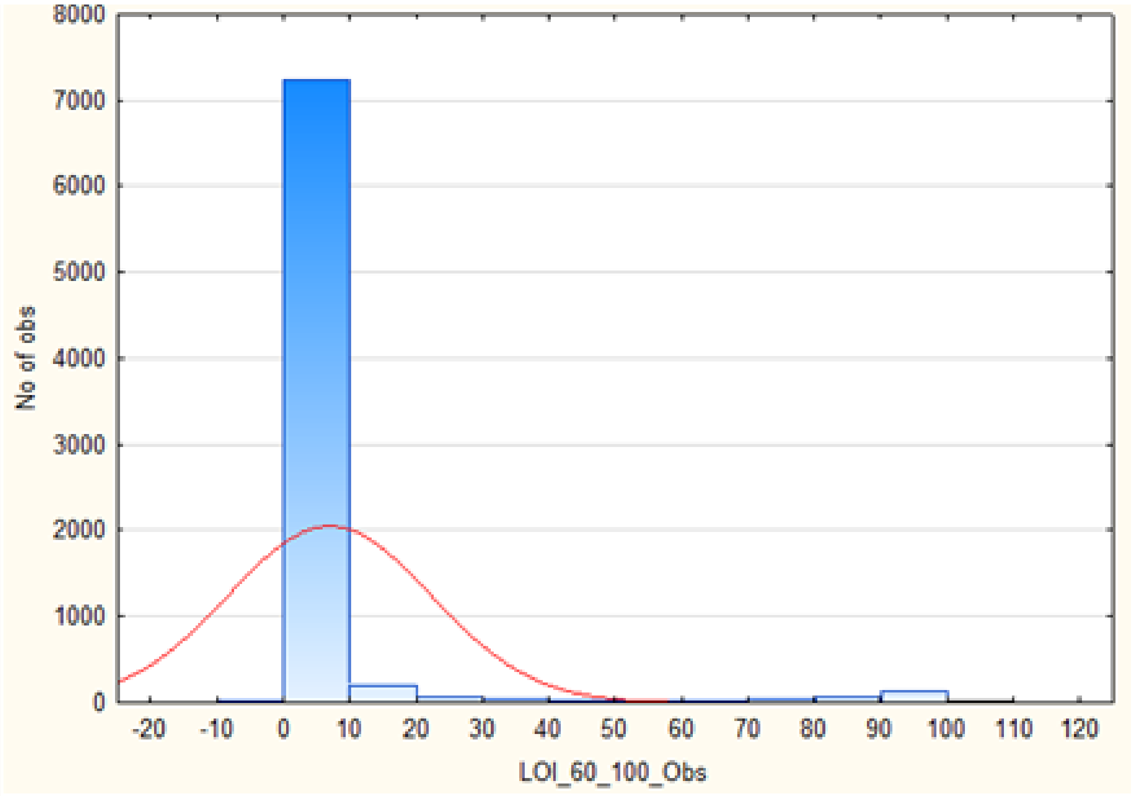

N
0 100 km
LOI (%) 60-100 cm
Urban
> 70.00
60.01 - 70.00
50.01 - 60.00
40.01 - 50.00
30.01 - 40.00
20.01 - 30.00
10.01 - 20.00
<10.00
© Crown copyright and database right (2018).
All rights reserved.
The James Hutton Institute, Ordnance
Survey Licence Number 100019294
Copyright © Cranfield University

b)


N
0 100 km
LOI (%) 60-100 cm
<-20.00
-20.00 - -10.00
-9.99 - 0.00
0.01 - 10.00
10.01 - 20.00
> 20.00
© Crown copyright and database right (2018).
All rights reserved.
The James Hutton Institute, Ordnance
Survey Licence Number 100019294
Copyright © Cranfield University

Appendix 6 – Number of LOI observations at 60-100 cm depth

This appendix contains a graph of LOI observations at 60-100cm depth. Information on this can be found in Chapter 5.



Appendix 7 – R Code

This appendix contains the R code used in this PhD research.

<u>Heat map Chapter 2</u>

```
dat1<-read.csv ("D:/Trial Grants.csv", sep=",")

dat1

str(dat1)

row. names(dat1) <-dat1$Activity

row. names(dat1)

dat1<-dat1[,2:12]

dat1_matrix<-data. matrix(dat1)

dat1_matrix

a<-colorRampPalette (c ("white", "blue")) (100)

library(gplots)

CLS<-c ('Associated Documentation’, ‘Contemporary Data', 'Co-ordinate locations', 'Finer
Scale resolution', 'Geographic projection', 'Improved coverage’,'Improve data source
accuracy', 'Metadata’, 'Data Generation Methodology', 'Pixel/polygon’, 'Driver changes',
‘Copyright’, ‘Non-users summary’, ‘uncertainty’, ‘Trends over time’, ‘Understanding soil
classification’)

RWS<-c (‘Agriculture', 'Conservation’, 'Construction', 'Environmental Consultancy',
'Environmental Regulator', 'Estate management’, 'Finance/insurance', 'Forestry',
'International Agency', 'Landscape design’, 'Local Authority', 'Local community', ‘National Fed
Govt’,’ Planning’, ’Research’, ‘Waste’, ‘Water’, ‘Other’)
```

```
heatmap.2(dat1_matrix, dendrogram = "none", Rowv = FALSE, Colv = FALSE, trace="none”, density.info="none”, col=a, key=TRUE, margins=c (14,14), keysize=1.0, key. xlab="“, key. title="“, labRow=RWS, labCol=CLS)
```

Creating splines

```
library (GSIF)

library(aqp)

library(plyr)

library(sp)

str(data)

depths(data)<-PRFL_NGR~ FIELD_SMPL_TOP + FIELD_SMPL_BTTM

class(data)

fittedsandadj2<-mpspline (data, "sandadj2", lam = 0.1, d = t (c (0,5,15,30,60,100,200)), vlow = 0, vhigh = 100, show. progress=TRUE)

fitstndsandadj2<- data. frame (fittedsandadj2$var.std, fittedsandadj2$idcol)
```

MARS Script Chapter 4 e.g. SCO test area

```
#The packages needed for this script are:

library(earth)

library(epiR)

#This loads your data

grant_data <- read.csv (file = "M:/Test_Areas_for_Modelling/Test_Areas/Fife/NEMV_newest_modelling/Silt_NEMV_100_200_GC1.csv", sep = ",", header = TRUE)

str(grant_data)
```

```
#Here we set soil property measurement to a vector...

Silt_100_200<- grant_data$Silt_100_200

#and then remove it from the rest of the data

grant_data["Silt_100_200"] <- NULL

#The following 9 lines of code are for ensuring that the variables are categorical

grant_data$soilsassoc <- as. factor(grant_data$soilsassoc)

grant_data$soilsmssg <- as. factor(grant_data$soilsmssg)

grant_data$LCM2000 <- as. factor(grant_data$LCM2000)

grant_data$ROCK <- as. factor(grant_data$ROCK)

grant_data$HAM <- as. factor(grant_data$HAM)

grant_data$SOTER <- as. factor(grant_data$SOTER)

grant_MARS <- earth (x = grant_data, y = Silt_100_200)

#This plots some kind of model summary

plot(grant_MARS)

#Here the fitted values are extracted from the model

MARS_fitted <- as. vector(grant_MARS$fitted.values)

#This plots the observed values against fitted values

plot (Silt_100_200, MARS_fitted, abline (0, 1))

grant_data_df<-data. frame (grant_data, MARS_fitted)

grant_data_df$MARS_fitted <- ifelse (grant_data_df$MARS_fitted < 0, 0,
grant_data_df$MARS_fitted)

### SOME SUMMARY STATISTICS ###

#R2
```

```
cor (MARS_fitted, Silt_100_200) ^2

#RMSE

sqrt (mean ((MARS_fitted - Silt_100_200) ^2))

#Bias

mean (MARS_fitted - Silt_100_200)

#compute the CCC

ccc <- epi.ccc (Silt_100_200, MARS_fitted, ci = "z-transform", conf. level = 0.95)

#Extract the estimate of CCC

ccc <- ccc$rho.c$est
```

BRT Script Chapter 4 e.g. SCO test area

```
library(dismo)

library(gbm)

grant_data <- read.csv (file = "M:/Test_Areas_for_Modelling/Test_Areas/Fife/NEMV_newest_modelling/Clay_NEMV_100_200_GC1.csv", sep = ",", header = TRUE)

str(grant_data)

grant_data$soilsassoc <- as. factor(grant_data$soilsassoc)

grant_data$LCM2000 <- as. factor(grant_data$LCM2000)

grant_data$ROCK <- as. factor(grant_data$ROCK)

grant_data$HAM <- as. factor(grant_data$HAM)

grant_data$SOTER <- as. factor(grant_data$SOTER)

Clay_100_200<- gbm. step (data=grant_data, gbm.x = 2:26, family = 'gaussian', gbm. y = 1, tree. complexity = 5, learning. rate = 0.001, bag. fraction = 0.5)
```

```
names (Clay_100_200)

summary (Clay_100_200)

grant_predict<- predict.gbm (Clay_100_200, n. trees = 1550)

hist(grant_predict)

range(grant_predict)

training<-data. frame (grant_predict, grant_data)

### SOME SUMMARY STATISTICS ###

plot (training$Clay_100_200, training$grant_predict)

abline (0,1)

#R2

cor (training$Clay_100_200, training$grant_predict) ^2

#RMSE

sqrt (mean ((training$grant_predict - training$Clay_100_200) ^2))

library(epiR)

#compute the CCC

ccc <- epi.ccc (training$Clay_100_200, training$grant_predict, ci = "z-transform", conf. level = 0.95)

#Extract the estimate of CCC

ccc <- ccc$rho.c$est
```

Stacking rasters (creating deployment)

```
library(rgdal)

library(raster)

library(sp)
```

```
a<-raster("soilsassoc.tif")

b<-raster("soilsmssg.tif")

c<-raster("AMT.tif")

d<-raster("AP.tif")

e<-raster("ISO.tif")

f<-raster("MDR.tif")

g<-raster("SP.tif")

h<-raster("ST.tif")

i<-raster("LCM2000.tif")

j<-raster("LCS88.tif")

k<-raster("Aspect.tif")

l<-raster("Slope.tif")

m<-raster("AH.tif")

n<-raster("CI.tif")

o<-raster("LongC.tif")

p<-raster("XSCurve.tif")

q<-raster("LSFact.tif")

r<-raster("TWI.tif")

s<-raster("RSP.tif")

t<-raster("VD.tif")

u<-raster("VDCN.tif")

v<-raster("CNBL.tif")
```

```
w<-raster("ROCK.tif")
x<-raster("HAM.tif")
y<-raster("SOTER.tif")
ca<-crop (a, a)
cb<-crop (b, b)
cc<-crop (c, c)
cd<-crop (d, d)
ce<-crop (e, e)
cf<-crop (f, f)
cg<-crop (g, g)
ch<-crop (h, h)
ci<-crop (i, i)
cj<-crop (j, j)
ck<-crop (k, k)
cl<-crop (l, l)
cm<-crop (m, m)
cn<-crop (n, n)
co<-crop (o, o)
cp<-crop (p, p)
cq<-crop (q, q)
cr<-crop (r, r)
cs<-crop (s, s)
```

```
ct<-crop (t, t)

cu<-crop (u, u)

cv<-crop (v, v)

cw<-crop (w, w)

cx<-crop (x, x)

cy<-crop (y, y)

com<-
ca+cb+cc+cd+ce+cf+cg+ch+ci+cj+ck+cl+cm+cn+co+cp+cq+cr+cs+ct+cu+cv+cw+cx+cy

cca<-crop (ca, com)

ccb<-crop (cb, com)

ccc<-crop (cc, com)

ccd<-crop (cd, com)

cce<-crop (ce, com)

ccf<-crop (cf, com)

ccg<-crop (cg, com)

cch<-crop (ch, com)

cci<-crop (ci, com)

ccj<-crop (cj, com)

cck<-crop (ck, com)

ccl<-crop (cl, com)

ccm<-crop (cm, com)

ccn<-crop (cn, com)

cco<-crop (co, com)
```

```
ccp<-crop (cp, com)

ccq<-crop (cq, com)

ccr<-crop (cr, com)

ccs<-crop (cs, com)

cct<-crop (ct, com)

ccu<-crop (cu, com)

ccv<-crop (cv, com)

ccw<-crop (cw, com)

ccx<-crop (cx, com)

ccy<-crop (cy, com)

stack_rstrs = stack (cca, ccb, ccc, ccd, cce, ccf, ccg, cch, cci, ccj, cck, ccl, ccm, ccn, cco,
ccp, ccq, ccr, ccs, cct, ccu, ccv, ccw ,ccx ,ccy,  bands = NULL, native = FALSE, RAT =
FALSE, quick = TRUE)

values<-rasterToPoints (stack_rstrs, fun= NULL, spatial =TRUE)

values.df<-as.data. frame(values)

values.df. naomit<-na. omit(values.df)

deployment<-data. frame (values.df. naomit)

deployment$soilsassoc <- as. factor(deployment$soilsassoc)

deployment$soilsmssg <- as. factor(deployment$soilsmssg)

deployment$AMT  <- as. integer(deployment$AMT)

deployment$AP  <- as. integer(deployment$AP)

deployment$ISO <- as. integer(deployment$ISO)

deployment$MDR  <- as. integer(deployment$MDR)
```

```
deployment$SP <- as. integer(deployment$SP)
deployment$ST <- as. integer(deployment$ST)
deployment$LCM2000 <- as. factor(deployment$LCM2000)
deployment$LCS88 <- as. factor(deployment$LCS88)
deployment$ROCK <- as. factor(deployment$ROCK)
deployment$HAM <- as. factor(deployment$HAM)
deployment$SOTER <- as. factor(deployment$SOTER)
deployment$x <- as. integer(deployment$x)
deployment$y <- as. integer(deployment$y)
str(deployment)
str(grant_data_df)
View(grant_data_df)
write. table (values.df. naomit, "Covariates_NEMV_22052017.txt", sep="\t", col. names=NA)
```

Using training data on deployment area (extrapolation)

```
grant_prediction <- predict (grant_MARS, newdata = deployment, type="response")
##create a dataframe from the predictions
pred_gcdf<-as.data. frame(grant_prediction)
str(pred_gcdf)
View(pred_gcdf)
#add xy coordinates to preddf
##Create xy coordinates
xy<-cbind (deployment$x, deployment$y)
```

```
formapping<-cbind (xy, pred_gcdf)


colnames(formapping)<-c("x","y","grant_MARS")

str(formapping)
```

Create raster map

```
library(raster)

coordinates(formapping)<- ~ x+y

gridded(formapping)<-TRUE

grant_MARS<-raster(formapping)

writeRaster (grant_MARS, filename="Clay_100_200_NEMV.tif", format="GTiff",
datatype="FLT4S”)
```

Appendix 8 – Publications

This appendix contains a list of the journal articles that have been prepared and/or published as part of this PhD thesis.

Journal Articles:

- Campbell, G.A., Lilly, A., Corstanje, R., Mayr, T.R., Black, H.I.J., 2017: Are existing soils data meeting the needs of stakeholders in Europe? An analysis of practice use from policy to field. *Land Use Policy*, 68, pp.211-223. [https://www.sciencedirect.com/science/article/pii/S0264837716312467]
- Campbell, G.A., Lilly, A., Corstanje, R., Hannam, J., Black, H.I.J., in revision: Evaluation of two model typologies and their behaviour in generating soil property predictions: studies from pilot areas in England and Scotland. *Geoderma*

Land Use Policy 69 (2017) 211–223

Contents lists available at ScienceDirect

Land Use Policy

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

Are existing soils data meeting the needs of stakeholders in Europe? An analysis of practical use from policy to field

G.A Campbell[a,b,*], A Lilly[a], R. Corstanje[b], T.R. Mayr[b], H.I.J Black[a]

[a] The James Hutton Institute, Craigiebuckler, Aberdeen, AB15 8QH, United Kingdom
[b] Cranfield University,College Road, Cranfield, Bedford, MK43 0AL, United Kingdom

ARTICLE INFO

Keywords:
Soils
Soil functions
Questionnaire
Stakeholders
Non-expert
Data
Improvements

ABSTRACT

Soils form a major component of the natural system and their functions underpin many key ecosystem goods and services. The fundamental importance of soils in the environment means that many different organisations and stakeholders make extensive use of soils data and information in their everyday working practices. For many reasons, stakeholders are not always aware that they are reliant upon soil data and information to support their activities. Various reviews of stakeholder needs and how soil information could be improved have been carried out in recent years. However, to date, there has been little consideration of user needs from a non-expert perspective. The aim of this study was to explore the use of explicit and hidden soil information in different organisations across Europe and gain a better understanding of improvements needed in soil data and information to assist in practical use by non-expert stakeholders. An on-line questionnaire was used to investigate different uses of soils data and information with 310 responses obtained from 77 organisations across Europe. Results illustrate the widespread use of soil data and information across diverse organisations within Europe, particularly spatial products and soil functional assessments and tools. A wide range of improvements were expressed with a prevalence for finer scale resolution, trends over time, future scenarios, improved accuracy, non-technical supporting information and better capacity to use GIS. An underlying message is that existing legacy soils data need to be supplemented by new up-to-date data to meet stakeholder needs and information gaps.

1. Introduction

Soils form a major component of our natural environment on Earth, performing an array of essential functions that underpin key ecosystem goods and services which we rely on (Costanza et al., 1997; Smith et al., 2015). The significance of soils within the environment has meant that stakeholders have to use a wide variety of soils data and information in their decision making.

The concept of soil functions was first conceived during the early 1950s and has since been widely adopted in national and regional policy (Blum, 2005). From the mid-1900s onwards, soils functional aspects have been incorporated into assessment tools such as maps and models that assist decision makers across a wide range of soil-related issues from land use, cropping practises, protection of water bodies, and restoration of habitats to climate regulation. For instance, many early assessments around agricultural productivity, such as the Land Capability for Scotland (Bibby et al., 1982) and laterally, the CAPRI model (Britz and Witzke, 2014), are based on soil maps. However, functional assessments have since extended across many other issues such as groundwater vulnerability (Environment Agency, 2013; Harter and Walker, 2001).

When exploring what needs to be improved in terms of soils data and information, we need to understand the contemporary needs of stakeholders particularly where soils data and information may be implicit or part of an underlying model or assessment tool. There are various reviews of stakeholder needs and how these levels of information could be improved which have been carried out in recent years (Black et al., 2012; Prager et al., 2014; McKee, 2014; Valentine et al., 1981; Grealish et al., 2015; Omuto et al., 2013; Houšková et al., 2010; Panagos et al., 2012). However, these reviews have generally assumed that stakeholders have some knowledge of soils or are fully aware that they are using soils data and information. The aim of this study is to understand soils data and information stakeholders' needs across Europe from a non-expert perspective.

Jones et al. (2005) reviewed soils resources and information use across Europe and determined that these are traditionally associated through the function of food and fibre production, with increasing applications to other issues such as climate change and water resource

[*] Corresponding author at: The James Hutton Institute, Craigiebuckler, Aberdeen, AB15 8QH, United Kingdom.
E-mail address: Grant.Campbell@hutton.ac.uk (G.A. Campbell).

http://dx.doi.org/10.1016/j.landusepol.2017.09.016
Received 14 December 2016; Received in revised form 10 August 2017; Accepted 11 September 2017
Available online 21 September 2017
0264-8377/

management (Blum, 2005; Grealish et al., 2015; Haines-Young, 2011). Soil maps, data and information are used in many sectors besides soil science, such as farming, hydrology, land degradation, policy and environmental modelling (Valentine et al., 1981; Mather, 1988; Houšková et al., 2010; Hallett et al., 2011; Omuto et al., 2013; Prager et al., 2014). The majority of soil information users indicated that key soil attributes are readily available (Wood and Auricht, 2011). However, improvements in a range of soil properties such as soil moisture, toxicity, biology and carbon are required (Auricht, 2004; Grealish et al., 2015).

Furthermore, engineering properties such as subsidence and corrosion are also of interest (Pritchard et al., 2015). These types of information are available but awareness of data accessibility and where to find them remains challenging. Information needs are also specific to stakeholder requirements and the spatial resolution of the undertaking. Black et al. (2012) consulted a wide range of stakeholders in developing the Soil Monitoring Action Plan for Scotland with further consultation taking place with farmers and local authorities by Prager and McKee, (2014). Key improvements mentioned were finer spatial resolution, soil trends, soil biological and physical indicators and sealing.

The FAO (2012) identify three major challenges in addressing soil information availability. The first of these focusses on the importance of soil protection, particularly to the global modelling community as it will help mitigate and adapt to issues such as climate change and food security. A second consideration is soil monitoring, focusing on improving global soil data at finer scale resolution. The third looks at advancing Digital Soil Mapping (DSM) and Digital Soil Assessment (DSA) techniques. DSM and DSA offers potential to map soil properties at detailed and broad scales (McBratney et al., 2003; Behrens and Scholten, 2006; Carré et al., 2007; Hartemink et al., 2008). However, it is not clear how any of these challenges reflect the needs of stakeholders, and difficulties remain around integrating the capability of models and the envisioned users of this data.

Stakeholder interaction and participation should be considered from the outset, and this is very rarely done (Reed, 2008). Studies by Bouma et al. (2012) and Black et al. (2012) highlighted that end-users were often not aware that they were using soils data and information so could not easily communicate further needs. It is therefore not straightforward to assume what the needs of envisioned users of 'new' soil information are, in particular where this information is embedded in derived tools. Here we planned a survey of non-expert users to investigate their current needs and perceived gaps in their ability to deliver in their work activities. This information is vital in addressing how new soil tools and products, such as DSM and DSA, might (or might not) meet the stakeholder requirements and the likelihood of such products being of practical use. Our aim is therefore to investigate what soils assessments and tools stakeholders currently use and what improvements, if any, required for future soil products/information sets.

2. Methodology

A detailed questionnaire was carried out to consider the range of soils data and information currently being used across Europe with a focus on explicit and hidden soils information being used by non-expert stakeholders: non-experts being people who use soils information or data in their everyday work but who are not expected to be academically trained soil scientists.

The questionnaire was compiled using the web-based survey programme Qualtrics (http://www.qualtrics.com/). In addressing the different uses of soils data and information, we considered it important to address functions of soils and contact stakeholders with close connections in and around these functions. Therefore, stakeholders were identified in order to be representative of the primary functions of soils (http://www.fao.org/resources/infographics/infographics-details/en/c/284478/) including biomass production, cultural heritage, regulating, biodiversity/habitats and infrastructure. A list of organisations across Europe, with named soil contacts, was drawn up by accessing published materials, on-line searches and personal knowledge. The remit and primary activities of these organisations corresponded well with at least one of the soil functions and provided coverage across the soil functions. Stakeholders were based around commercial organisations, learned societies, non-governmental organisations (NGOs), local authorities and government organisations. A total of 98 organisations were contacted across 22 countries in Europe. Of these, 34 organisations can be considered trans-European in their activities i.e. no specific alignment with any one region or country. A pilot study of the questionnaire was conducted with staff at The James Hutton Institute (Aberdeen) and the Scottish Government's ethics committee; the questionnaire incorporated amendments following relevant feedback. The survey was carried out from July to August 2015 and was made accessible to stakeholders through an anonymous online link.

3. Questionnaire results

3.1. What sectors use soils information?

There were 310 individual responses to the questionnaire from 77 out of the 98 organisations we contacted and, from this, 93% of stakeholders said that they handled information about soil in their work.

Stakeholders were asked to identify what best describes the activities of their organisation. Stakeholders could tick more than one option for this question in order to obtain a broader understanding of activities associated with individual organisations. The top three activities were agriculture, research organisations (universities, institutes etc.) and conservation (Fig. 1). Stakeholders who ticked 'other' ranged from people who worked in landscape photography, archaeology and oil and gas services. This shows that there is a wide array of stakeholders who have an interest in soils data and information and who may use certain tools and assessments related to activities within their organisation.

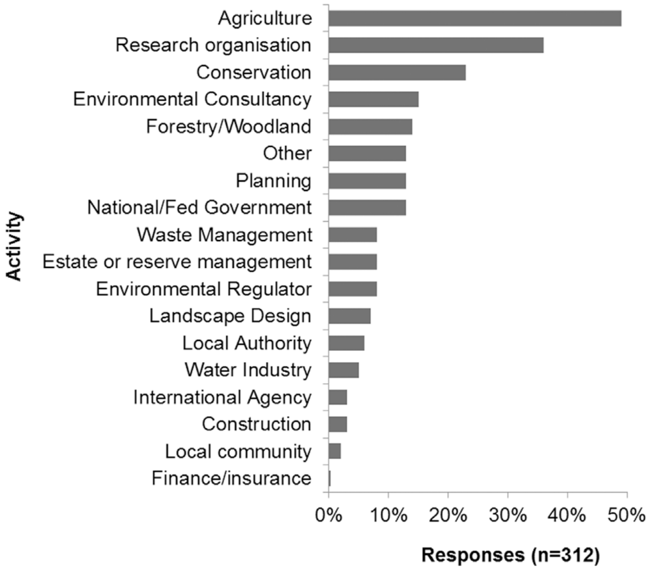


Fig. 1. Range and type of organisations and the percentage of responses to the questionnaire.
This was to get an understanding as to the variety of organisations people worked for. n.b. Stakeholders could tick more than one option for this question.

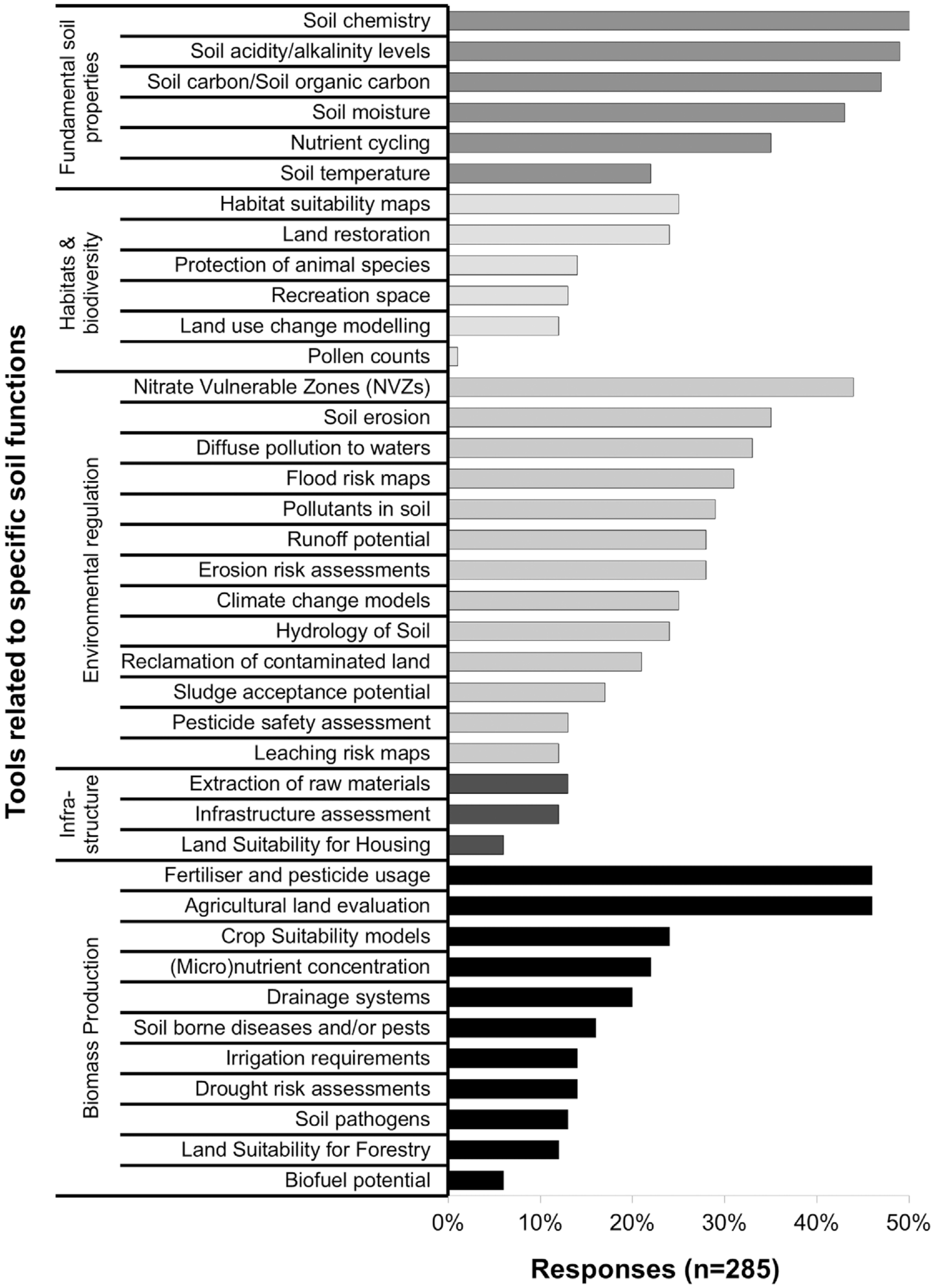


Fig. 2. Tools and assessments used and percentages used by respondents. These are broken up into their closest related soil function.

3.2. Tools and assessments and awareness of embedded soils information

Stakeholders were encouraged to tick as many boxes as possible in terms of what tools and assessments they use in their work. These assessments are grouped by related soil functions. Most responses came from people who were connected with agricultural production and conservation of habitats and biodiversity. Respondents were asked about how aware they were that many of the assessments had soils information embedded within them, with 87% saying that they were 'aware'.

In relation to '*Biomass Production'*, it was found that the two main tools predominantly used were agricultural land evaluation and fertiliser/pesticide usage assessments. In terms of assessments grouped under '*Infrastructure*', it is the extraction of raw materials such as clay, sand and silt, followed by assessment of the impacts of soils on assets such as pipes and electric cables. Nitrate Vulnerable Zones (NVZs) were found to be the main assessment tool used by stakeholders closely associated with '*Environmental Regulation'* with soil erosion and diffuse pollution to water following closely behind.

Habitat suitability maps and land restoration assessments were the

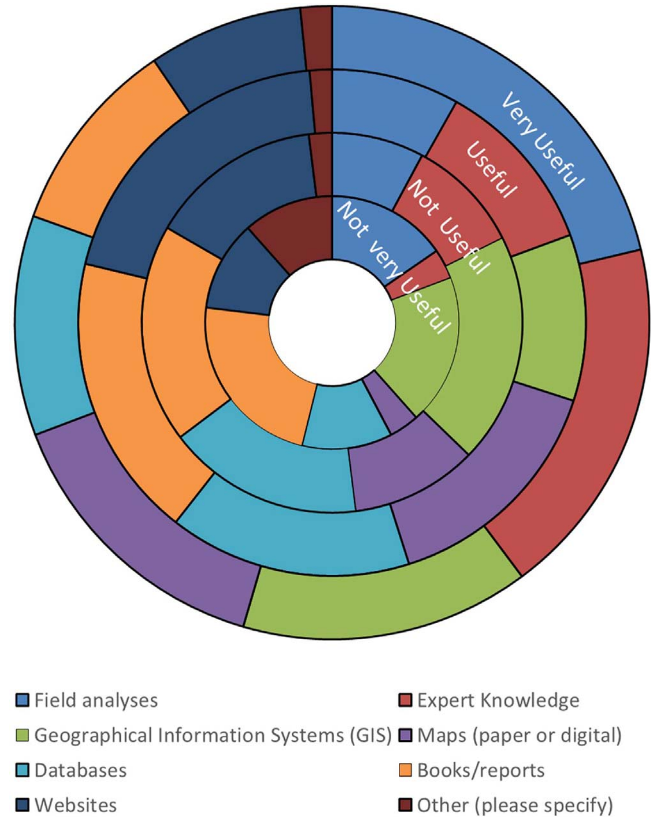


Fig. 3. How stakeholders rated usefulness of sources. Outer circle represents the percentage of stakeholders who rated '*very useful*'. Inner circle represents those who rated '*not very useful*'.

most commonly used assessments by stakeholders related to '*Habitats and Biodiversity'*.

The number of stakeholders requesting information on fundamental soil properties from the questionnaire was relatively high. Soil chemistry (primary contaminants) and other properties including soil acidity, alkalinity and carbon had the highest demand and application (Fig. 2). A number of other assessments which were not listed in the survey were also used by stakeholders including soil climate zones to identify nutrient demands of crops and grasslands.

3.3. Sources of information, licencing and spatial importance

Respondents were asked to identify what sources they used to acquire soil information required for their work. The use of maps in either paper or digital format is the most prolific with 78% of respondents using them while 65% of respondents use Geographical Information Systems (GIS). Other sources consisted of social media websites and discussions with knowledge transfer exchange with stakeholders (11% of respondents).

On the whole, most stakeholders found most sources that they used either '*very useful'* or '*useful*'. 95% found the use of maps, expert knowledge and field and laboratory analysis to be either *'useful'* or '*very useful'*. However, 11% reported that GIS systems were '*not very useful'* or '*not useful'* (Fig. 3).

When asked whether or not their organisation paid for licenced use of soils information, 49% said that their organisation did, 30% said '*no*' and 21% said that they '*didn't know*'.

Respondents were asked to assess the importance of spatial soils information for wider applications and end-user groups and as a result of this, an overwhelming 98% of the respondents said that this was '*very important'* or '*important*'. Previously, we saw that 93% handled information about soil as part of their work. This extra 5% illustrates that those respondents who do not use, or acknowledge soil as part of their work still see the importance of spatial soil information for wider applications and end-user groups.

3.4. Requested improvements to soil information and data

Improvements to soil data and information were a key issue addressed in this questionnaire. Respondents were asked what they would like to see improved in relation to the information they already use and this has been summarised in Fig. 4. We grouped improvements post-survey to ease interpretation under four main themes: '*Uncertainty', 'Scale and Coverage', 'Metadata'* and *'Fundamental Data'*. Most stakeholders wanted soil information at a much finer resolution or scale to what they currently use. With regards to '*Uncertainty*', respondents wanted improved accuracy and credibility of data sources. With regards to '*Scale and Coverage'*, as well as wanting information at finer scale resolution, respondents wanted to see improvements in co-ordinates of geographical locations (i.e. data in a format which they can georeference). With respect to '*Metadata*' issues, respondents requested improvements in the availability of associated documentation related to the data. Finally, under the category of '*Fundamental Data', w*e see that respondents wish to see improvements with trends over time and contemporary data. Respondents were then asked specifically if they would be interested in using any new information that might arise from improvements in spatial resolution/scale and uncertainty. From Table 1, we can see that there is a positive response to improvements regarding both of these issues. Other notable requirements ranged from improving map and data interpretations, and the ability to use multiple datasets or assessments.

There was a space at the end of the questionnaire for respondents to add any extra information that might be useful. The main themes that came out from the additional responses were opportunities to increase knowledge transfer between research and policy makers and also the importance of education and training, which are vital in terms of increasing soil understanding.

3.5. Relationships between organisations and desired improvements

One of the main objectives of this study was to establish from the questionnaire what desired improvements were linked to the activities of particular activities. To achieve this, responses were cross tabulations between activities of the organisations and the desired improvements the stakeholders had requested. This was undertaken using the Qualtrics software. The cross tabulations were then used to create heat maps using R Statistics software (https://www.r-bloggers.com/citing-r-or-sas/) (Fig. 5). The legend indicates how the shading relates to the number of people who answered responses to both of these questions i.e. the darker the colour then the greater the correspondence between activities within that specific organisation and the requested improvements. From this we can see, improvements in finer/scale resolution are being requested most by stakeholders whose activities revolve around agriculture or research but consistently needed across all organisational activity groups. Trends over time are also particularly related to those working in agriculture and research but also sought by stakeholders in conservation and national/federal or governmental agencies.

Using the same data, we converted the crosstabs into percentages to explore needs within activity groups. For the majority of organisations

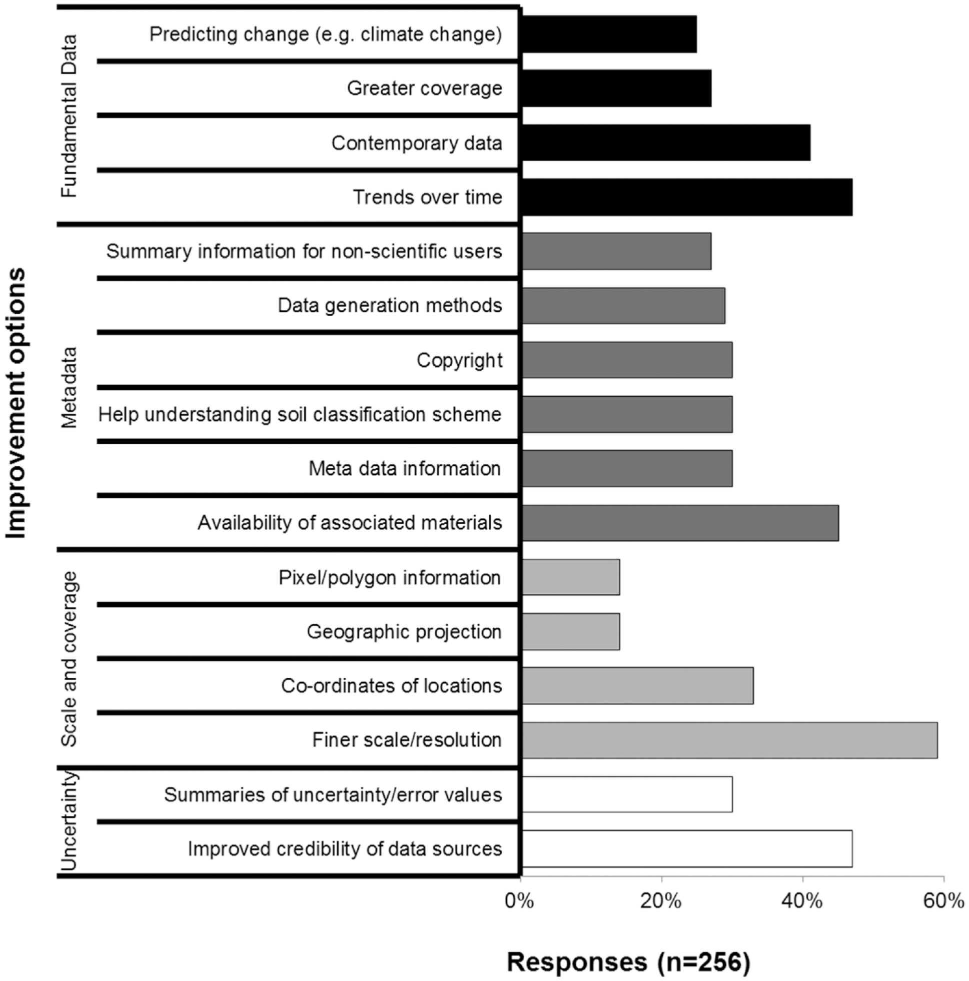


Fig. 4. Improvement recommendations by the stakeholders.

Table 1
Would you be interesting in any new information arising from an improvement in spatial resolution/scale or summary of uncertainty/error values.

Issue	Yes	No	Total responses
Spatial resolution/scale	209	36	245
Summary of uncertainty/error values	159	57	216
Other (please specify)	10	7	17

(Fig. 6), finer scale resolution and, associated, improved data accuracy predominated individual organisational user needs. Some organisations identified quite specific needs. In the finance/insurance category, these include improvements in contemporary data, finer scale resolution, improved coverage and methodology in how the data was generated. In the water sector, understanding soil classification and non-(expert) user summaries were identified as relatively high needs.

4. Discussion

It is encouraging that we were able to obtain a large number of responses from non-expert stakeholders across substantially different organisations. It is clear that many diverse sectors are using, wish to use or access soils information on a regular basis to support day-to-day work practices. Moreover, our survey demonstrates that soils data and information are widely used in a range tools and assessments and are often integrated with other data sources such as historical data on climate and vegetation (e.g. where soil climate zones were used to establish nutrient demand for crops and grassland for regional animal manure management).

The survey responses also identified that there are barriers to accessing and using appropriate soil data. Overall, it would seem that stakeholders find difficulties obtaining and collecting information for projects which are under licence or where they have to pay for the use of it. Payment for use of data is particularly dependent on organisations procurement procedures and that different organisations are willing to pay varying amounts in order to obtain certain data for their work or projects (Montanarella and Vargas, 2012; Diafas et al., 2013). It is unclear how much this constituents a significant barrier to the use of soil information, as payment was not identified as one of the key improvements from the questionnaire. However, improving accessibility would clearly benefit non-experts. Alongside this, there is a clear need to address technical understanding with needs identified for knowledge transfer between research and policy, education and training, improving associated supporting information, understanding soil classifications and non-expert user information. A need for more technical knowledge may well reflect a lack of soils in school and university level education. The level of responses suggests that there is demand (and opportunity) for soils training opportunities focussed on non-experts and practical applications. In parallel, there is also a clear need for

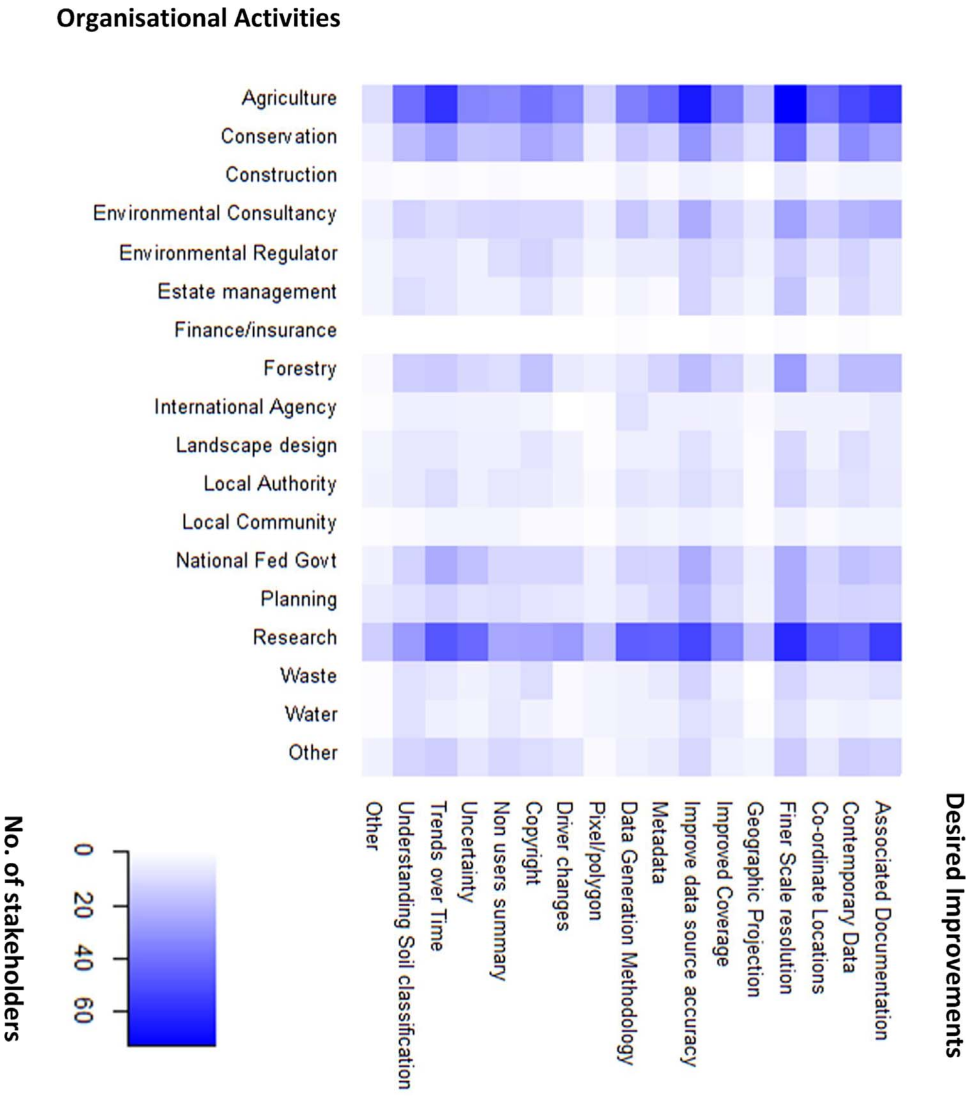


Fig. 5. Heat map showing the cross tabulation of responses (N) between the activities of an organisation and suggested improvements in soil related information. The darker the colour indicates a greater number of responses.

increased skill capacity in GIS within organisations using spatial soil data and information. Without this, it is difficult to see how new spatial soil products, which are predominately GIS in nature, can be widely adopted for practical use.

Stakeholders used a variety of information sources and of these, it was notable that a high proportion of people found GIS to be the least useful source of information even though a high proportion of stakeholders use or want to use spatial information and that GIS is a widely used spatial information platform. This may be due to constraints around technical ability, accessibility to GIS software (although open-source GIS software is available e.g. QGIS), or could allude to a more fundamental problem with the GIS medium being inadequate for the assessments undertaken by the respondents.

Other sources of information that were mentioned ranged from the use of social media sites like Twitter, academic journal articles and discussions with other stakeholders. Although not used widely at present, social media does now present real and widespread opportunities to communicate with and inform non-experts. Interestingly, most people found field and laboratory analyses to be '*very useful*' or '*useful*' alongside maps, whether in paper or digital format, and expert knowledge. Reasons could be that stakeholders are utilising '*tacit knowledge*' from field experts who acquired this information in the first place, thus using it as a validation tool (Hudson, 1992) and they are sufficient familiar with handling field and lab results. This may also reflect issues discussed about constraints with technical understanding and GIS skills limiting use of other soil data and information sources.

The questionnaire also indicated widespread requirements for information on future scenarios and trends over time. There is a significant amount of legacy soils data available but much of this is at over 30 years old which could be used more to explore future scenarios and trends over time. There is however an underlying requirement for new information on soils to be able to determine current trends in soil properties and functions and to support modelling of future scenarios based on current conditions. Legacy data, on its own, cannot meet current user needs.

Our survey indicates that a number of soil properties including texture (sand, clay and silt), contaminants, bulk density, pH and carbon have widespread use. These should be a priority in making more

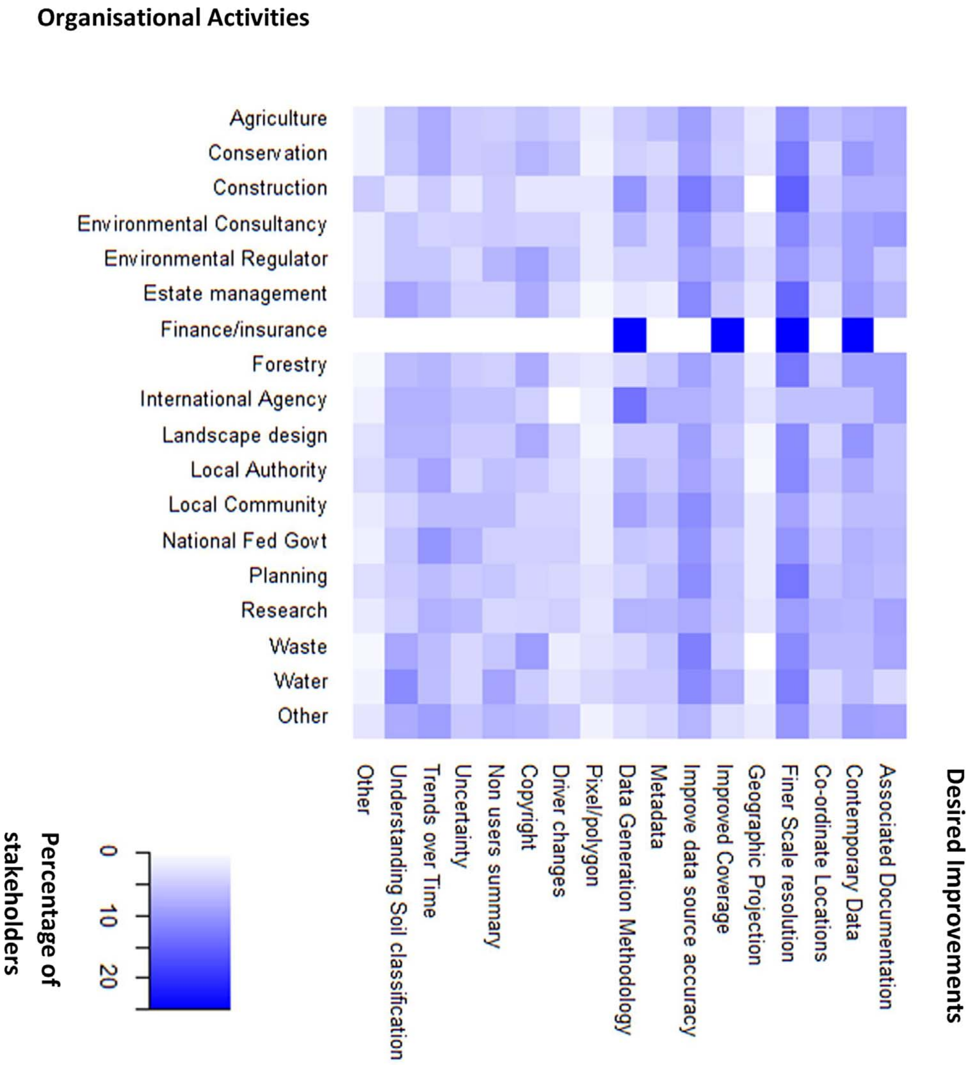


Fig. 6. Heat map showing the cross tabulation of responses (%) between the activities of an organisation and suggested improvements in soil related information. The darker the colour indicates a greater percentage of responses by each organisation.

accessible and useable by addressing the needs for non-expert supporting materials, finer spatial resolution, trends over time etc. However, there are also other soil properties to be considered. Many of the answers in the questionnaire reflect instances where soil properties underpin soil functional assessments and tools. In such instances, the relevance of individual soil properties is "*hidden*" to the user and therefore the need for information on individual soil properties may not be fully expressed. This is a potential pit-fall to be recognised in any future assessments of stakeholder needs. Table 2 illustrates the soil properties used to derive these assessments using information gathered from previous documentation and literature (e.g. GlobalSoilMap, 2011a,b; Mayr et al., 2006). This can be used in post-hoc identification of "*hidden*" soil properties in questionnaires, in particular when exploring needs for soil functional assessments and in ensuring that all necessary soil properties are being considered in the improvement of existing mapping or development of new modelling and mapping, such as DSM and DSA (c.f. Mayr et al. (2006). Expressing the links between soil properties and soil functions can also be used as a tool in raising stakeholders' awareness of the wider range of soil properties which underpin the soil functional assessments and tools that they use regularly.

Most stakeholders stated, from the questionnaire, that they require information at finer spatial scale/resolution than what is currently being offered. An obvious focus for future work is to deliver finer spatial scale in the key soil properties identified by the stakeholders (i.e. bulk density, soil contaminants, pH, texture and carbon). However, one assumption is that finer spatial scale will lead to improved data and subsequent assessments. This may not be the case since scale is a complex parameter which is dependent on context and application (Goodchild, 1997; Wu and Li, 2009). Supported and promoted by FAO (http://www.fao.org/global-soil-partnership/pillars-action/4-information-and-data/en/), DSM is a major opportunity to gain soil property information at finer spatial scale than existing products, with the benefit of characterising accuracy and precision properties (Cavazzi et al., 2013). Such predicted soil property products can then be used to make significant advances in modelling and mapping the soil functional assessments which are widely used by diverse stakeholders and organisations. However, it is imperative that such approaches are matched with field assessments to critically evaluate and validate the accuracy of predicting soil properties at finer spatial resolution using existing (generally legacy) data.

Table 2
Soil assessments mentioned in the questionnaire measured against probable soil properties that will be mapped as future work. Table adapted from: GlobalSoilMap (2011a, 2011b) and Mayr et al. (2006).

Related Soil Function	Assessments	Organic carbon	pH	Clay	Silt	Sand	Coarse Fragments	ECEC	Bulk density (whole soil)	Available Water capacity	Bulk Density (fine earth)
Biomass Production	Agricultural land evaluation			✓	✓	✓	✓		✓	✓	✓
	Biofuel potential		✓							✓	
	Crop Suitability models								✓	✓	✓
	Drainage systems		✓							✓	
	Fertiliser and pesticide usage	✓	✓	✓	✓	✓	✓	✓		✓	
	Irrigation requirements		✓							✓	
	Land Suitability for Forestry		✓	✓	✓	✓	✓				
	(Micro) nutrient concentration		✓								
	Soil borne diseases and/or pests										
	Soil pathogens										
	Drought risk assessments		✓							✓	
Environmental Regulation	Climate change models	✓	✓						✓	✓	✓
	Erosion risk assessments	✓		✓	✓	✓	✓			✓	
	Flood risk maps		✓	✓	✓	✓	✓			✓	
	Hydrology of Soil			✓	✓	✓	✓			✓	
	Leaching risk maps		✓					✓			
	Nutrient Vulnerable Zones		✓							✓	
	Pesticide safety assessment		✓								
	Pollutants in soil		✓					✓		✓	
	Reclamation of contaminated land		✓								
	Runoff potential		✓					✓		✓	
	Sludge acceptance potential		✓					✓			
	Soil erosion		✓	✓	✓	✓	✓				
	Diffuse pollution to waters									✓	
Fundamental Soil Properties	Nutrient cycling		✓					✓			
	Soil acidity/alkalinity levels		✓					✓			
	Soil carbon/organic carbon	✓							✓		✓
	Soil chemistry		✓					✓			
	Soil moisture	✓						✓	✓	✓	✓
	Soil temperature							✓			
Habitats and Biodiversity	Habitat suitability maps									✓	
	Land reclamation/restoration			✓	✓	✓	✓				
	Land use change modelling										
	Pollen counts										
	Protection of animal species										
	Recreational space										
Infrastructure	Extraction of raw materials	✓		✓	✓	✓	✓			✓	
	Infrastructure assessment			✓	✓	✓	✓				
	Land Suitability for Housing			✓	✓	✓	✓				

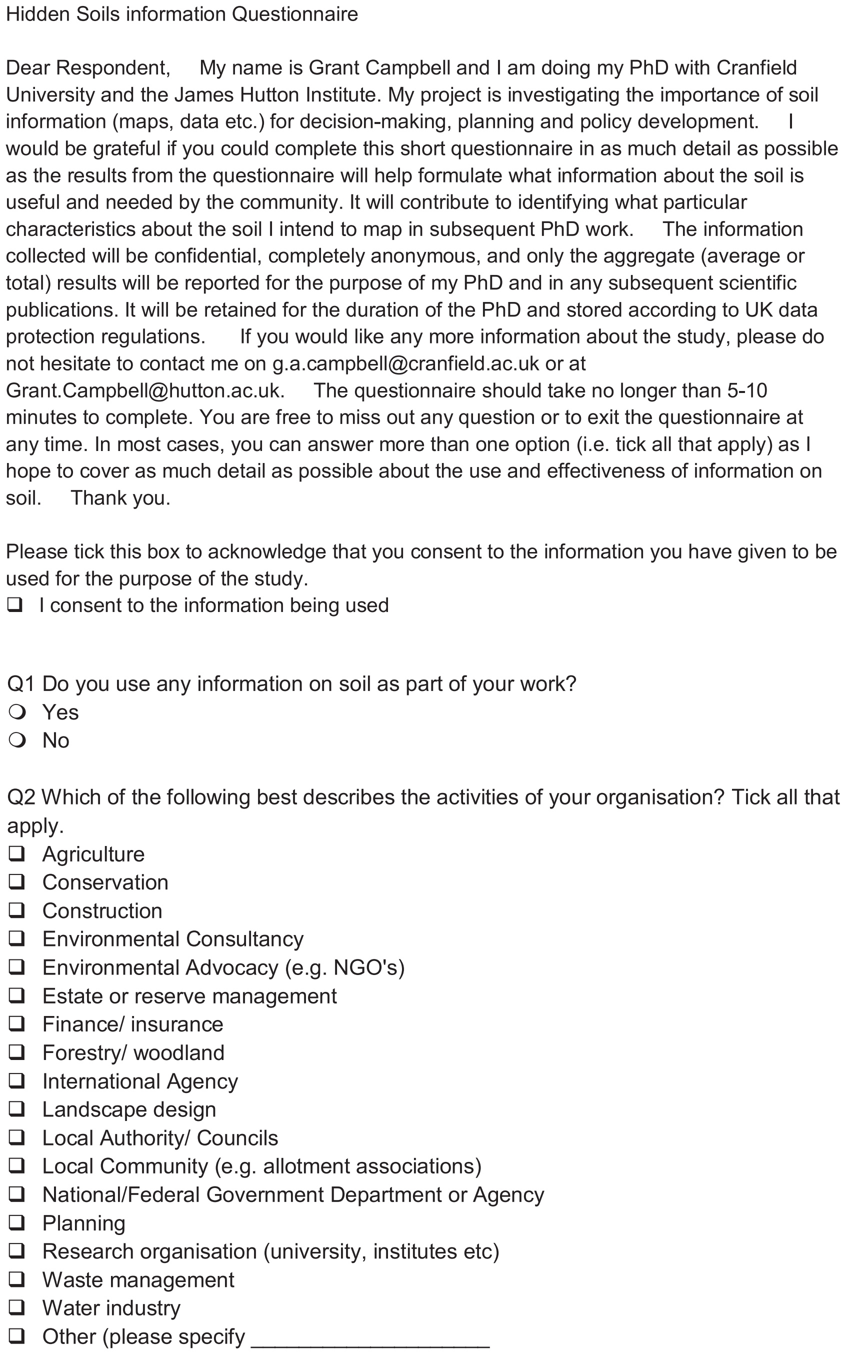
5. Conclusions

The questionnaire was designed to understand how soils data and information are being used by non-expert stakeholders for a range of purposes. The responses indicate that stakeholders are generally aware of the utility of soil data and soil functional assessments for their work however they may not be aware of the full range of soil properties underlying soil functional assessments. Stakeholders identified that better and wider use of existing (and future) soil information by non-experts could be enabled by improvements in data access and user-friendly supporting materials. The majority of stakeholders require finer spatial resolution than is currently offered, contemporary information on soils and trends over time for soil functions as well as properties. Established soil modelling such as the global initiatives in DSM and DSA can address some of these needs. However, a clear message from stakeholders is that existing legacy soils data needs to be supplemented by new up-to-date soil data which is fit for current and future uses. Requirements for contemporary data demand investments in new and novel monitoring and sampling at sufficient spatial resolution and frequency to enable assessments of the range of soil functions. These will, in turn, be used deliver and shape a wide range of multi-organisational activities and policies. A question still remains on how long we can rely on legacy soil data to make decisions today and into the future?

Acknowledgements

Thanks to all questionnaire participants who made this paper possible. Furthermore, the authors acknowledge Katrin Prager and Altea Lorenzo-Arribas at the James Hutton Institute for providing thoughts on questionnaire construction and how best to illustrate the results; Scottish Government who gave ethic approval to the questionnaire's use in Scotland; Oliver Pritchard for reviewing and suggesting improvements to the manuscript; Fiona Fraser for providing guidance on how to illustrate the cross-tabulation analysis using heat maps; the Rural and Environment ScienceAnalytical Services Division of the Scottish Government (RESAS) for funding support to Allan Lilly and Helaina Black and the sponsors EPSRC and Macaulay Development Trust for funding this PhD research; the reviewers for comments on a previous version of the paper.

Appendix A. Hidden Soils information Questionnaire



Hidden Soils information Questionnaire

Dear Respondent, My name is Grant Campbell and I am doing my PhD with Cranfield University and the James Hutton Institute. My project is investigating the importance of soil information (maps, data etc.) for decision-making, planning and policy development. I would be grateful if you could complete this short questionnaire in as much detail as possible as the results from the questionnaire will help formulate what information about the soil is useful and needed by the community. It will contribute to identifying what particular characteristics about the soil I intend to map in subsequent PhD work. The information collected will be confidential, completely anonymous, and only the aggregate (average or total) results will be reported for the purpose of my PhD and in any subsequent scientific publications. It will be retained for the duration of the PhD and stored according to UK data protection regulations. If you would like any more information about the study, please do not hesitate to contact me on g.a.campbell@cranfield.ac.uk or at Grant.Campbell@hutton.ac.uk. The questionnaire should take no longer than 5-10 minutes to complete. You are free to miss out any question or to exit the questionnaire at any time. In most cases, you can answer more than one option (i.e. tick all that apply) as I hope to cover as much detail as possible about the use and effectiveness of information on soil. Thank you.

Please tick this box to acknowledge that you consent to the information you have given to be used for the purpose of the study.

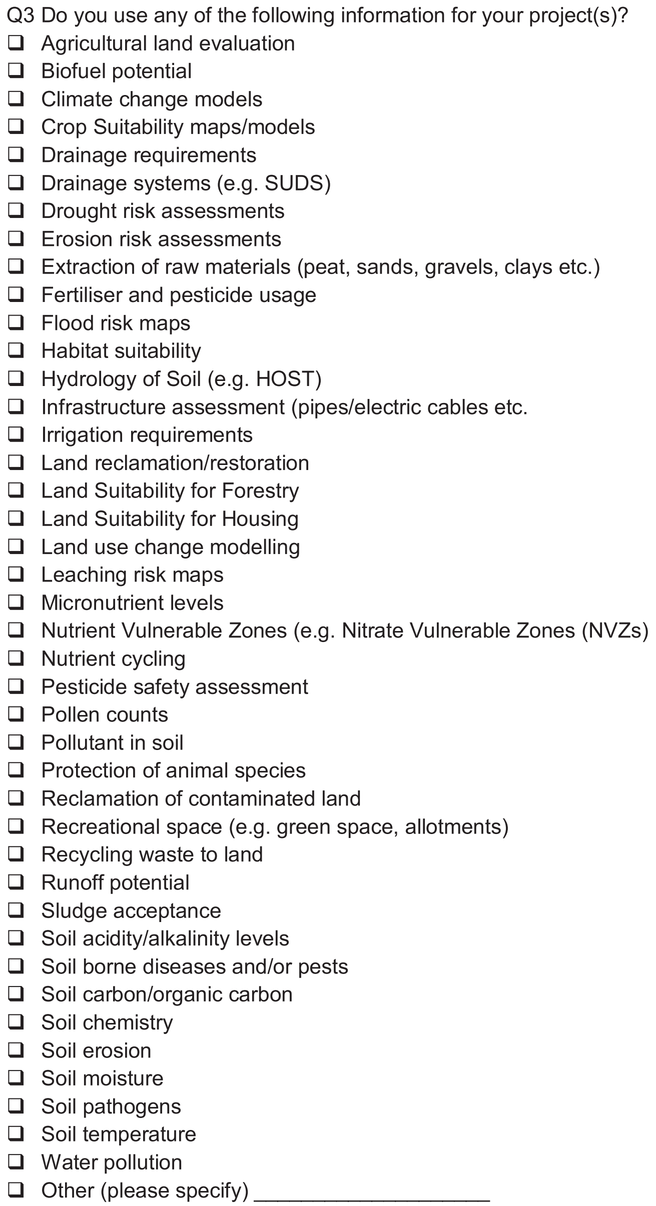
□ I consent to the information being used

Q1 Do you use any information on soil as part of your work?

○ Yes
○ No

Q2 Which of the following best describes the activities of your organisation? Tick all that apply.

□ Agriculture
□ Conservation
□ Construction
□ Environmental Consultancy
□ Environmental Advocacy (e.g. NGO's)
□ Estate or reserve management
□ Finance/ insurance
□ Forestry/ woodland
□ International Agency
□ Landscape design
□ Local Authority/ Councils
□ Local Community (e.g. allotment associations)
□ National/Federal Government Department or Agency
□ Planning
□ Research organisation (university, institutes etc)
□ Waste management
□ Water industry
□ Other (please specify ____________________



Q3 Do you use any of the following information for your project(s)?

□ Agricultural land evaluation
□ Biofuel potential
□ Climate change models
□ Crop Suitability maps/models
□ Drainage requirements
□ Drainage systems (e.g. SUDS)
□ Drought risk assessments
□ Erosion risk assessments
□ Extraction of raw materials (peat, sands, gravels, clays etc.)
□ Fertiliser and pesticide usage
□ Flood risk maps
□ Habitat suitability
□ Hydrology of Soil (e.g. HOST)
□ Infrastructure assessment (pipes/electric cables etc.
□ Irrigation requirements
□ Land reclamation/restoration
□ Land Suitability for Forestry
□ Land Suitability for Housing
□ Land use change modelling
□ Leaching risk maps
□ Micronutrient levels
□ Nutrient Vulnerable Zones (e.g. Nitrate Vulnerable Zones (NVZs)
□ Nutrient cycling
□ Pesticide safety assessment
□ Pollen counts
□ Pollutant in soil
□ Protection of animal species
□ Reclamation of contaminated land
□ Recreational space (e.g. green space, allotments)
□ Recycling waste to land
□ Runoff potential
□ Sludge acceptance
□ Soil acidity/alkalinity levels
□ Soil borne diseases and/or pests
□ Soil carbon/organic carbon
□ Soil chemistry
□ Soil erosion
□ Soil moisture
□ Soil pathogens
□ Soil temperature
□ Water pollution
□ Other (please specify) ____________________

Q4 Of the following sectors, which are the most relevant to your work?

□ Agricultural production
□ Biofuel production
□ Building/ infrastructure
□ Climate change mitigation
□ Conservation of habitats and biodiversity
□ Contaminated land
□ Cultural heritage or archaeology
□ Environmental Impact Assessments
□ Extraction of raw materials (e.g. peat, sands, gravels, clays)
□ Flood regulation
□ Forestry production
□ Land use planning
□ Pests and diseases
□ Recreation (e.g. amenity woodland, tourism)
□ Recycling organic waste to land
□ Water supply and/or quality
□ Other (please specify) ____________________

Q5 Are you aware that the information you may use in your work has soils information embedded within it?

○ Yes, I was aware
○ No, I was not aware
○ I was not sure

Q6 What source(s) do you use to acquire the information you need?

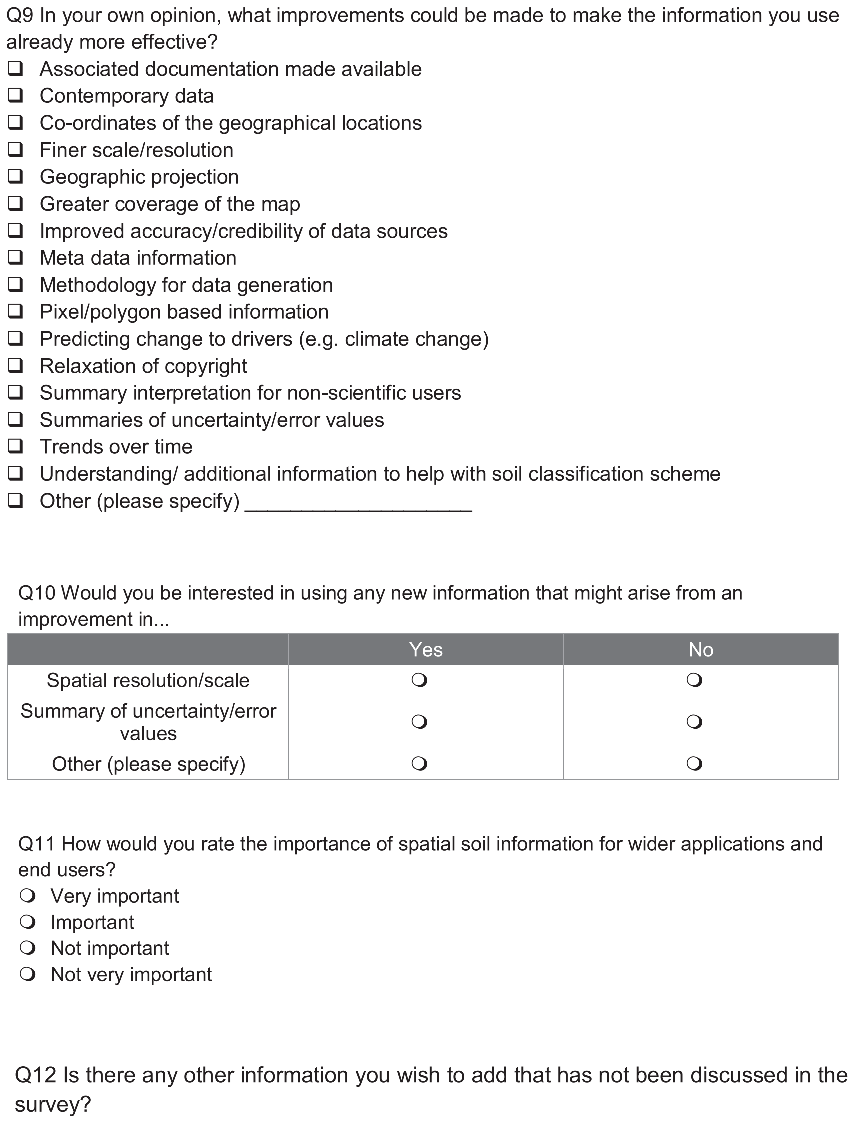
□ Books/reports
□ Databases
□ Expert knowledge
□ Field analyses
□ Geographical Information Systems (GIS)
□ Maps (paper or digital)
□ Websites
□ Other (please specify ____________________

Q7 How accurate or useful do you find the available information for your purposes?

	Not very useful	Not useful	Useful	Very useful
Books/reports	○	○	○	○
Databases	○	○	○	○
Expert Knowledge	○	○	○	○
Field analyses	○	○	○	○
Geographical Information Systems (GIS)	○	○	○	○
Maps (paper or digital)	○	○	○	○
Websites	○	○	○	○
Other (please specify)	○	○	○	○

Q8 Does your organisation pay for the licence use of any of the information you have identified?

○ Yes
○ No
○ Don't Know



Q9 In your own opinion, what improvements could be made to make the information you use already more effective?

- □ Associated documentation made available
- □ Contemporary data
- □ Co-ordinates of the geographical locations
- □ Finer scale/resolution
- □ Geographic projection
- □ Greater coverage of the map
- □ Improved accuracy/credibility of data sources
- □ Meta data information
- □ Methodology for data generation
- □ Pixel/polygon based information
- □ Predicting change to drivers (e.g. climate change)
- □ Relaxation of copyright
- □ Summary interpretation for non-scientific users
- □ Summaries of uncertainty/error values
- □ Trends over time
- □ Understanding/ additional information to help with soil classification scheme
- □ Other (please specify) ____________________

Q10 Would you be interested in using any new information that might arise from an improvement in...

	Yes	No
Spatial resolution/scale	○	○
Summary of uncertainty/error values	○	○
Other (please specify)	○	○

Q11 How would you rate the importance of spatial soil information for wider applications and end users?

- ○ Very important
- ○ Important
- ○ Not important
- ○ Not very important

Q12 Is there any other information you wish to add that has not been discussed in the survey?

References

Auricht, C., 2004. Natural Resources Atlas and Data Library ? User Review. National Land and Water Resources Audit. (Accessed from [Last Accessed 20th July 2017]). http://lwa.gov.au/products/er040794.

Behrens, T., Scholten, T., 2006. Digital soil mapping in Germany—a review. J. Plant Nutr. Soil sci. 169, 434–443.

Bibby, J., et al., 1982. Land Capability Classification for Agriculture. The Macaulay Land Use Research Institute, Aberdeen (ISBN 0 7084 0508 8).

Black, H., et al., 2012. Soil Monitoring Action Plan. ([Last accessed 13th December 2016]Accessed from). http://www.environment.scotland.gov.uk//media/59999/Soil_Monitoring_Action_Plan.PDF.

Blum, W.E.H., 2005. Functions of soil for society and the environment. Rev. Environ. Sci. BioTechnol. 4, 75–79.

Bouma, W., et al., 2012. Soil information in support of policy making and awareness raising. Curr. Opin. Environ. Sustainability 4 (5), 552–558.

Britz, W., Witzke, W., 2014. CAPRI Model Documentation. ([Last accessed 21st July 2017][Accessed from). http://www.capri-model.org/docs/capri_documentation.pdf.

Carré, F., et al., 2007. Digital soil assessments: beyond DSM. Geoderma 142, 69–79.

Costanza, R., d'Arge, R., de Groot, R., Farber, S., Grasson, M., Hannon, B., Limburg, K., Naeem, S., O'Neill, R.V., Paruelo, J., Raskin, R.G., Sutton, P., van den Belt, M., 1997. The value of world's service and natural capital. Nature 387, 253–260.

Diafas, I., et al., 2013. Willingness to pay for soil information derived by digital maps: a choice experiment approach. Vadose Zone J. 12 (4). http://dx.doi.org/10.2136/vzj2012.0198.

Environment Agency, 2009. Groundwater Protection: Principles and Practice (GP3) (August 2013 Version 1.1.). ([Last accessed 26th September 2015][Available at). https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/297347/LIT_7660_9a3742. pdf.

FAO, 2012. Towards Global Soil Information: Activities Within the Geo Task Global Soil Data: Workshop Report. ([Last accessed 10th September 2015] Accessed from). http://www.fao.org/fileadmin/templates/GSP/downloads/GSP_SoilInformation_WorkshopReport.pdf.

GlobalSoilMap, 2011a. GlobalSoilMap.net. New Digital Soil Map of the World. ([Last accessed 17th February 2015][Accessed from). https://www.google.co.uk/search?q=globalsoilmap.net&ie=utf-8&oe=utf-8&aq=t&rls=org.mozilla:en-GB:official&client=firefox-a&channel=sb&gfe_rd=cr&ei=aR3jVLWpOoSV8wOK4YHwBQ.

GlobalSoilMap, 2011b. Specifications Version 1: GlobalSoilMap.net Products: Release 2.1. Technical Report.

Goodchild, M.F., 1997. Towards a geography of geographic information in a digital world. Computers. Environ. Urban Syst. 21, 377–391.

Grealish, G.J., et al., 2015. Soil survey data rescued by means of user friendly soil identification keys and toposequence models to deliver soil information for improved land management. GeoRes J. 6, 81–91.

Harter, T., Walker, L.G., 2001. Assessing the Vulnerability of Groundwater. ([Last Accessed 20th July 2017][Accessed from). www.dhs.ca.gov/ps/ddwem/dwsap/DWSAPindex.htm.

Houšková, B., et al., 2010. Assessment and Strategic Development of INSPIRE Compliant Geodata-Services for European Soil Data.

Hudson, H.D., 1992. Division S-5?Soil genesis, morphology and classification: the soil survey as paradigm-based science. Soil Sci. Soc. Am J. 56, 836–841.

Jones, R., et al., 2005. Soil Resources of Europe, second edition. European Soil Bureau Institute for Environment & Sustainability JRC Ispra.

Mather, A.S., 1988. New private forests in Scotland: characteristics and contrasts. Area 20 (2), 135–143.

Mayr, T., et al., 2006. Novel Methods for Spatial Prediction of Soil Functions Within Landscapes (SP0531). DEFRA (26pp).

McBratney, A.B., et al., 2003. On digital soil mapping. Geoderma 117 (1), 3–52.

Montanarella, L., Vargas, R., 2012. Global governance of soil resources as a necessary condition for sustainable development. Curr. Opin. Environ. Sustain. 4, 1–6.

Omuto, C., et al., 2013. State of the art report on global and regional soil information: where are we? where to go? FAO Report.

Panagos, P., et al., 2012. European Soil Data Centre: response to European policy support and public data requirements. Land Use Policy 29 (2), 329–338.

Prager, K., McKee, A., 2014. Use and awareness of soil data and information among local authorities, farmers and estate managers. The James Hutton Institute Internal Report.

Pritchard, O.G., Hallett, S.H., Farewell, T.S., 2015. Probabilistic soil moisture projections to assess Great Britain's future clay-related subsidence hazard. Clim. Change 133 (4), 635–650.

Reed, M., 2008. Stakeholder participation for environmental management: a literature review. Biological Cons. 141 (10), 2417–2431.

Smith, P., et al., 2015. Biogeochemical cycles and biodiversity as key drivers of ecosystem services provided by soils. Soil D 2, 537–586.

Valentine, K.W.G., et al., 1981. A questionnaire to users of soil maps in British Columbia: results and implications for design and content. Can. J. of Soil Sci. 61, 123–135.

Wood, B., Auricht, C., 2011. ASRIS/ACLEP User Needs Analysis. ([Last Accessed 20th July 2017] [Accessed from). http://www.clw.csiro.au/aclep/documents/ASRIS_User_Analysis.pdf.

Wu, H., Li, Z.L., 2009. Scale issues in remote sensing: a review on analysis, processing and modelling. Sensors 9, 1768–1793.