CRANFIELD UNIVERSITY

Stefano Cavazzi

SPATIAL SCALE ANALYSIS OF LANDSCAPE PROCESSES FOR DIGITAL SOIL MAPPING IN IRELAND

School of Applied Sciences

Ph.D.
Academic Year: 2013 - 2014

Supervisors: T. Mayr and R. Corstanje

October 2013

CRANFIELD UNIVERSITY

School of Applied Sciences

Ph.D.

Academic Year 2013 - 2014

Stefano Cavazzi

SPATIAL SCALE ANALYSIS OF LANDSCAPE PROCESSES FOR DIGITAL SOIL MAPPING IN IRELAND

Supervisors: T. Mayr and R. Corstanje

October 2013

ABSTRACT

Soil is one of the most precious resources on Earth because of its role in storing and recycling water and nutrients essential for life, providing a variety of ecosystem services. This vulnerable resource is at risk from degradation by erosion, salinity, contamination and other effects of mismanagement. Information from soil is therefore crucial for its sustainable management. While the demand for soil information is growing, the quantity of data collected in the field is reducing due to financial constraints. Digital Soil Mapping (DSM) supports the creation of geographically referenced soil databases generated by using field observations or legacy data coupled, through quantitative relationships, with environmental covariates. This enables the creation of soil maps at unexplored locations at reduced costs. The selection of an optimal scale for environmental covariates is still an unsolved issue affecting the accuracy of DSM.

The overall aim of this research was to explore the effect of spatial scale alterations of environmental covariates in DSM. Three main targets were identified: assessing the impact of spatial scale alterations on classifying soil taxonomic units; investigating existing approaches from related scientific fields for the detection of scale patterns and finally enabling practitioners to find a suitable scale for environmental covariates by developing a new methodology for spatial scale analysis in DSM.

Three study areas, covered by detailed reconnaissance soil survey, were identified in the Republic of Ireland. Their different pedological and geomorphological characteristics allowed to test scale behaviours across the spectrum of conditions present in the Irish landscape. The investigation started by examining the effects of scale alteration of the finest resolution environmental covariate, the Digital Elevation Model (DEM), on the classification of soil taxonomic units. Empirical approaches from related scientific fields were subsequently selected from the literature, applied to the study areas and compared with the experimental methodology. Wavelet analysis was also employed to decompose the DEMs into a series of independent components at

varying scales and then used in DSM analysis of soil taxonomic units. Finally, a new multiscale methodology was developed and evaluated against the previously presented experimental results.

The results obtained by the experimental methodology have proved the significant role of scale alterations in the classification accuracy of soil taxonomic units, challenging the common practice of using the finest available resolution of DEM in DSM analysis. The set of eight empirical approaches selected in the literature have been proved to have a detrimental effect on the selection of an optimal DEM scale for DSM applications. Wavelet analysis was shown effective in removing DEM sources of variation, increasing DSM model performance by spatially decomposing the DEM. Finally, my main contribution to knowledge has been developing a new multiscale methodology for DSM applications by combining a DEM segmentation technique performed by k-means clustering of local variograms parameters calculated in a moving window with an experimental methodology altering DEM scales. The newly developed multiscale methodology offers a way to significantly improve classification accuracy of soil taxonomic units in DSM.

In conclusion, this research has shown that spatial scale analysis of environmental covariates significantly enhances the practice of DSM, improving overall classification accuracy of soil taxonomic units. The newly developed multiscale methodology can be successfully integrated in current DSM analysis of soil taxonomic units performed with data mining techniques, so advancing the practice of soil mapping. The future of DSM, as it successfully progresses from the early pioneering years into an established discipline, will have to include scale and in particular multiscale investigations in its methodology. DSM will have to move from a methodology of spatial data with scale to a spatial scale methodology. It is now time to consider scale as a key soil and modelling attribute in DSM.

Keywords: Digital Soil Mapping, Digital Elevation Models, terrain analysis, spatial scale, pixel resolution, window size, spatial data mining, geostatistics, wavelet, multiscale.

ACKNOWLEDGEMENTS

The present research work was conducted as part of the ISIS project, managed by Teagasc and co-funded by the EPA of Ireland through their Science, Technology, Research and Innovation for the Environment (STRIVE) Programme, as part of the National Development Plan 2007-2013.

I must acknowledge my supervisors Dr. Thomas Mayr and Dr. Ron Corstanje, my co-supervisors at Teagasc Mr. Reamonn Fealy and Dr. Rachel Creamer and all the other colleagues that helped and advised me along the way: Dr. Jacqueline Hannam, Dr. Robert Jones, Mr. Tim Brewer, Prof. Jane Rickson, Miss Joanna Zawadzka, Dr. Jim Halliday and a special mention to a fellow PhD student and now friend Dr. Fabio Veronesi.

I dedicate this thesis to my family: Maria, Fulvio and Gosia for their constant support and unconditional love.

”That’s another thing we’ve learned from your Nation,” said Mein Herr, “map-making. But we’ve carried it much further than you. What do you consider the largest map that would be really useful?”

“About six inches to the mile.”

“Only six inches!” exclaimed Mein Herr. “We very soon got to six yards to the mile. Then we tried a hundred yards to the mile. And then came the grandest idea of all! We actually made a map of the country, on the scale of a mile to the mile!”

“Have you used it much?” I enquired.

“It has never been spread out, yet,” said Mein Herr: “the farmers objected: they said it would cover the whole country, and shut out the sunlight! So we now use the country itself, as its own map, and I assure you it does nearly as well.”

Sylvie and Bruno Concluded, Lewis Carroll, 1893.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

ABBREVIATIONS

1D	One-dimensional
2D	Two-dimensional
AFT	An Foras Taluntais
ANOVA	Analysis of Variance
AOD	Above Ordnance Datum
b	Bits
B	Bytes
CART	Classification And Regression Tree
CORINE	Co-ordination of Information on the Environment
DB	Daubechies
DEM	Digital Elevation Model
DSM	Digital Soil Mapping
DWT	Discrete Wavelet Transform
EPA	Environmental Protection Agency of Ireland
EU	European Union
GIS	Geographical Information System
GSI	Geological Survey of Ireland
ICARUS	Irish Climate Analysis and Research Units
ISIS	Irish Soil Information System
KST	Keys of Soil Taxonomy
LIDAR	Light Detection And Ranging
MLA	Maximum Location Accuracy
MLD	Minimum Legible Delineation
MLP	Multilayer Perceptrons
NN	Neural Network
NSRI	National Soils Research Institute
NSS	National Soil Survey
OSI	Ordinance Survey of Ireland
PCA	Principal Components Analysis
RBF	Radial Basis Function

REML	Residual Maximum Likelihood
RF	Random Forest
SN	Scale Number
SOTER	Soil terrain database
STRIVE	Science, Technology, Research and Innovation for the Environment
SURE	Stein's Unbiased Risk Estimate
WRB	World Reference Base

GLOSSARY OF TERMS

Complex	A mapping unit of two or more kinds of soil occurring in such an intricate pattern that they cannot be shown separately on a soil map at the selected scale of mapping and publication. Generally, the name of a soil complex consists of the names of the dominant soils, joined by a hyphen (USDA, 1990).
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. Typical examples are terrain attributes derived from DEMs, remote sensing data (land use, etc.), climatic variables and geological maps.
DEM	The representation of continuous elevation values over a topographic surface by a regular array of z-values, referenced to a common datum (ESRI, 2010). These digital elevation models are typically used to represent Earth's terrain surface.
DSM	The creation and the population of a geographically referenced soil databases generated at a given resolution by using field and laboratory observation methods coupled with environmental data through quantitative relationships (Lagacherie *et al*., 2006).
GIS	An integrated collection of computer software and data used to view and manage information about geographic places, analyse spatial relationships, and model spatial processes. A GIS provides a framework for gathering and organizing spatial data and related information so that it can be displayed and analysed (Shekhar and Xiong, 2008).

Great Soil Group	Great Soil Groups are soils having the same kind, arrangement and degree of expression of horizons in the soil profile. They also have close similarity in soil moisture and temperature regimes and in base status (Gardiner and Radford, 1980a).
Phase	Soils of one series can differ in texture of the surface layer and in slope, stoniness, or some other characteristics that affect use of the soils by man. On the basis of such differences, a soil series is divided into phases (USDA, 1990).
Scale	The physical dimension of a phenomenon or process in space expressed in spatial units (pixel and roving window sizes).
Soil	A natural, three-dimensional body at the earth's surface that is capable of supporting plants and has properties resulting from the integrated effect of climate and living matter acting on earthy parent material, as conditioned by relief over periods of time (USDA, 1990).
Soil Series	A group of soils, formed from a particular type of 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).
Raster	A spatial data model that defines space as an array of equally sized cells arranged in rows and columns, and composed of single or multiple bands. Each cell contains an attribute value and location coordinates. Unlike a vector structure, which stores coordinates explicitly, raster coordinates are contained in the ordering of the matrix. Groups of cells that share the same value

	represent the same type of geographic feature (ESRI, 2010).
Resampling	The process of interpolating new cell values when transforming rasters to a new coordinate space or cell size. In the case of bilinear interpolation for example, a weighted average of the four nearest cells is used to determine a new cell value (ESRI, 2010).
Roving Window	The roving-window approach can be considered the standard filter technique in raster GIS operations and in image processing. It determines the new value for a given cell in a raster map using a mathematical function (mean, mode, standard deviation, etc.) of the cells values inside a n x n neighbourhood (with odd n) centred in the cell of interest. The window is moved one cell at a time across the raster map, until the whole area is processed (Grohmann and Riccomini, 2009).
Terrain Attributes	Data characterising the land surface geometry derived from elevation data. The local terrain attributes are calculated using a fixed size window around each cell (slope, aspect, etc.), while the regional one consider relation between cells and study and not-fixed surrounding area for each cell (Hengl and Reuter, 2009).

1 INTRODUCTION

In recent history, progress in information technology has resulted in increasing computational capacity, powerful Geographic Information System's tools (GIS), remote and proximal sensors and a vast amount of data such as digital elevation models (DEM). Soil scientists in the 21th century are now able to study and describe soils as dynamic entities in an interconnected landscape context (McBratney *et al.*, 2003; Lagacherie and McBratney, 2006; Grunwald, 2009; Kempen *et al.*, 2012).

Soil is an intricate system of interrelated physical, chemical and biological factors, much of it under human management. In order to gain a better understanding of this complexity, scientists have utilized both mathematical and statistical models for the quantification of its properties (Figure 1.1). This has created a new discipline in soil science termed "pedometrics", the science of developing quantitative techniques to predict soil properties from landscape attributes (Minasny *et al.*, 2008).

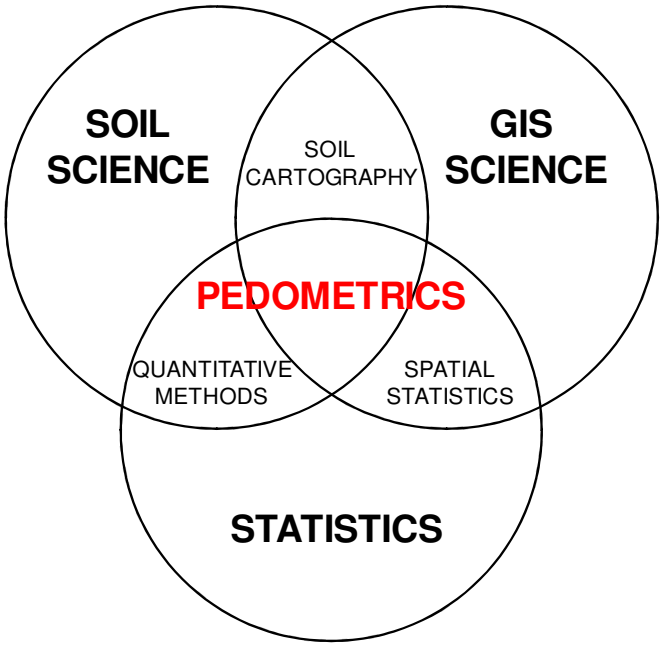


Figure 1.1 - Concept of Pedometrics.

This way of blending field observations (e.g. point observations or existing soil maps) with statistical spatial prediction techniques has created a new branch of research in soil science called Digital Soil Mapping (DSM). DSM enables us to predict soil taxonomic units or specific soil proprieties (organic carbon, texture, bulk density, etc.) in areas where information is required by spatially extending point observations of individual soil properties using mathematical and statistical techniques as well as estimating the uncertainty of such predictions. DSM, by formalising the relationship between soil forming factors and the landscape, aims to capture and model the intrinsic spatial variability naturally observed in soils.

After more than twenty years of intensive research and applications, DSM has emerged as a credible alternative to traditional soil mapping (Carre *et al.*, 2007) due to its low costs and fast deployment in comparison to conventional surveying methods. Despite its short history, the development of DSM has consisted of a rapid series of advancements (improvements in data mining and knowledge discovery, better selection of terrain attributes and ancillary data, etc.) as soil scientists expanded the scope and prediction power of their modelling. However, one of the fundamental concerns since the foundation of this technique, that still remains unsolved, is the issue of scale (Addiscott, 1998, Lagacherie, 2008). Thomson *et al.* (2001) have also suggested that this will become increasingly important with the fast development and implementation of regional soil-landscape models.

The choice of scale will be even more significant in the evolution of DSM as spatial data infrastructure will offer increasingly detailed environmental covariates, thanks to the improvements of sensor techniques. GIS science is already developing complex spatial models embedding hierarchical reasoning which supports multiscale representations of data and uncertainty over space and time (Eaglesona *et al.*, 2002). Scale frames the analysis and shapes the end result of DSM models suggesting that better attention and quantitative knowledge of the effects that soil forming processes have at different scales will improve our ability to map soils, thus enabling the up-scale or down-scale of soil information

at regional, national or global scale (Sanchez *et al.,* 2009). It is the impact of scale on DSM processes that is explored in this thesis.

1.1 Research context

The draft EU Soil Framework Directive (COM(2006)232) was introduced by the European Commission seeking to harmonize and raise the level of soil protection across Europe. There is therefore a strong need to bring together soil data across Europe (Jones *et al*., 2005) to reach a common understanding of the available soil resources and the threats challenging their sustainable management. In order to do that, it is essential to ensure data comparability. The European Soil Bureau Network recommended and endorsed the preparation of a soil map and associated information system at 1:250,000 scale. This scale was chosen as a reasonable intermediate level between the existing Soil Geographical Database of Eurasia at 1:1,000,000 scale and detailed national studies.

In Ireland, soil data exist at variable scales with a complete national coverage only available at 1:575,000 scale but with detailed information available at 1:127,560 scale covering 44% of the country (An Foras Taluntais soil survey from the 1960s, 70s and 80s). To rationalise and harmonise this information at the European target scale of 1:250,000, the Irish Soil Information System (ISIS) project was established, employing advanced DSM techniques in combination with traditional field survey methodologies for validation. The project is funded by the Irish Environmental Protection Agency (EPA), run by Teagasc with the technical support of the National Soil Resources Institute (NSRI) at Cranfield University.

By understanding the quantitative relationships between soil and environmental factors in the counties historically surveyed at detailed level, the project aims to create models to be used in the prediction of soil types in the remaining half of the country. In ISIS, the detailed soil maps were used as training for developing

soil-landscape relationships. The models were subsequently applied to similar areas for which no adequate soil information was available. At the same time a two and half years' survey campaign (2010-2012) has provided soil data on approximately 10,000 auger bores and 300 new profiles which were used as a ground truth to calibrate and validate the predictive models. The project providing the required soil information at EU 1:250,000 scale will form the basis for future soil research, management and policy in Ireland.

The present research work was conducted as part of the ISIS project, managed by Teagasc and co-funded by the EPA of Ireland through their Science, Technology, Research and Innovation for the Environment (STRIVE) Programme, as part of the National Development Plan 2007-2013. Therefore this thesis is concerned about the impact of spatial scale on mapping soil taxonomic units.

1.2 Digital Soil Mapping

As defined by the international working group on digital soil mapping of the International Union of Soil Science (IUSS), DSM is "*the creation and the population of a geographically referenced soil databases generated at a given resolution by using field and laboratory observation methods coupled with environmental data through quantitative relationships*" (Lagacherie *et al.*, 2006). This definition, focused on the quantitative relationships between environmental covariates and soil information, is deeply rooted in the pioneering work of Hans Jenny that conceptualized soil formation as a function of independent factors of pedogenesis in the famous equation:

$$S = f\ (cl, o, r, p, t, \ldots) \qquad (1.1)$$

where *S* (soil property) is a function of *cl* (regional climate), *o* (potential biota), *r* (topography), *p* (parent material), *t* (time) and the dots express factors not yet

known or specific to particular situations (Jenny, 1941). This mathematical relationship connecting observed soil properties with independent factors responsible for soil formation has allowed, in conjunction with technological and computational advancements (computer science, statistics and geostatistics in particular, remote sensing, GPS, GIS and numerical environmental data), a fertile environment for DSM to establish itself as a new discipline in soil science.

McBratney *et al.* in 2003 reviewed the development of this new branch of research and formalised DSM with the SCORPAN approach:

$$S_c = f\ (s, c, o, r, p, a, n) \quad ; \quad S_a = f\ (s, c, o, r, p, a, n) \tag{1.2}$$

where S_c (soil classes) or S_a (soil attributes) are a function of *s* (other properties of the soil at a point), *c* (climatic properties of the environment at that point), *o* (organisms, vegetation, fauna or human activity), *r* (topography, landscape attributes), *p* (parent material, lithology), *a* (age, the time factor) and *n* (spatial position). This new formulation of Jenny's soil formation function (expanded from the Vasily Dokuchaev equation) implicitly recognises the important missing aspect, that soils influence each other through spatial location. The new equation also better reflects the quantitative relationships between soil and soil forming factors, in light of the most recent techniques used in soil spatial prediction. It is centred around the idea that soil spatial variation can be estimated by statistical relationships, linking soil taxonomic units or soil proprieties with a set of environmental attributes at that particular location. If enough soil field observations and environmental covariates with a high data density are available, it is possible to use statistical techniques to exploit the existing relationships between the soil and its environment to subsequently extrapolate to unexplored locations.

The relationship between soil properties and landscape attributes has been confirmed as a central concept in soil science (Hudson, 1992). Terrain attributes are the most widely used covariates in DSM because of their primary role in soil formation and the broad availability of DEM (Behrens *et al.*, 2010b). DEMs are

representations of the endlessly varying topographic surface of the Earth, and they are a widespread data source for terrain analysis and other spatial applications. Terrain analysis provides a large number of high-resolution environmental information, quantitatively derived from DEM, including slope, aspect, plan curvature, etc. These topographic features are at the core of a wide range of landscape-scale environmental models (Gallant and Hutchinson, 1997). Terrain features describe the Earth's surface shape, position and connectivity, mediating the influence of all the other environmental factors in soil formation. These features control the local climatic characteristics like precipitation, solar radiation, thermal balance and wind speed; the activity of soil organisms like earthworms, bacteria, fungi and others, regulating the movement of water, gasses and soil particles. The amount of water runoff and patterns of drainage are two important aspects controlled by terrain, as clearly confirmed by the sequence of soils in the catena concept (Milne, 1934). Each soil in the catena has different characteristics, despite the same overall parent material and climate, due to the slope type (Schaetzl and Anderson, 2005). On steep slopes the rate of erosion by runoff is high, forming thin and dry soils while at the foot soils are deeper as the material is accumulated. Soil depths, moisture content and acidity vary along the slope, as seen in Ireland on impermeable acid parent material, where wet bogs with deep peats on the flatter aspects become dry shallow peaty podzols on the steeper facets. This fact explains the prevalence of poorly drained soils in flat areas but, on the other hand, differences in parent material can inhibit water movement and cause the opposite effect. Topography also controls the soil use and management practice for farming, animal rearing or forestry directly, so influencing man made impacts on soil formation.

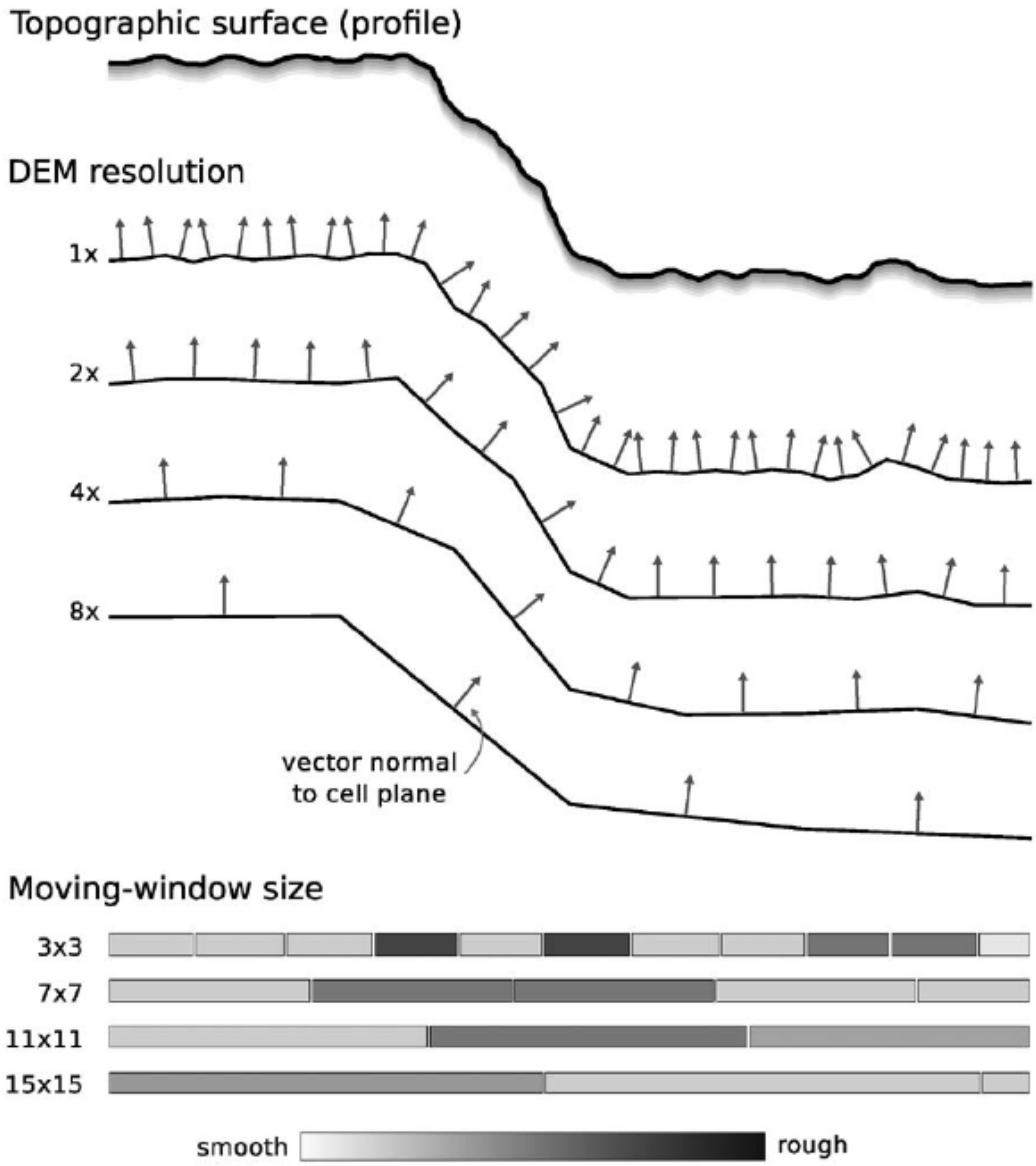


Figure 1.2 - Schematic representation of surface roughness changing with pixel and window sizes (from Grohmann *et al.*, 2010).

As presented by Grohmann *et al.* (2010) terrain features (roughness in this case) vary greatly with changes in pixel resolution and moving window size (Figure 1.2). These changes are going to have an effect on DSM analysis as the character of the terrain parameter changes with the alteration of scale, thus modifying the relationship with soil information.

1.3 The fundamental role of scale

The scale of raster based environmental covariates can be represented in two ways, grid size and window size, of which grid size is the most commonly considered to represent scale. The determination of an optimal grid size for environmental factors to use in soil prediction is still an unsolved issue with only few empirical guidelines available.

In DSM, the prediction power of soil attributes is highly dependent on finding the most suitable DEM resolution from which terrain attributes are derived. This presents major challenges because the DSM modelling scale has not only to encapsulate the scale at which the soil map units are represented on the map but also the scale at which the soil taxonomy characterises soil forming processes active in the landscape at a particular scale.

This problem is connected with the determination of the spatial scale of environmental phenomena or processes involved in soil formation. This is a critical problem because different pedogenetic laws and landscape processes operate at distinctive spatial scales (Florinsky and Kuryakova, 2000) and thus analysis based on data of one scale may not apply to another scale.

In statistical terms, observations made at a fine scale contain more variance than observations at a very coarse one. The greater assortment of observed processes and relationships at a detailed scale gives obviously more information, but also more noise. On the contrary, coarser scales can be too simplified and carry insufficient data.

The level of detail represented by a raster dataset depends on the pixel size so that the cell is small enough to capture the necessary information without undermining the analysis. Fine resolution data have a smaller pixel size and a higher feature spatial accuracy that affect the processing time of models and require large storage capacity. In particular, the increasing availability of high resolution DEM (1 m LIDAR for example) has the consequence of increasing data

storage and processing time exponentially, but might not increase the quality of the DSM models.

In DSM, a common approach is to use the finest DEM resolution available believing that this will improve the accuracy and precision of the prediction while on the contrary it is increasing the “noise” regarded as the unexplained variation inherently unpredictable. For example, it is likely that a lowland area could be represented effectively with a DEM of 100m resolution or more, whereas an upland or hill areas would probably need a much finer grid size (Pain, 2005). Furthermore, Thompson *et al.* (2001) have shown that higher-resolution DEM may not be necessary for generating useful soil-landscape models. Another concern is that most applications use algorithms running in small windows (usually 3 X 3 roving window) to perform terrain analysis, thus fixing the scale of resulting layers to the spatial resolution of the available DEM. This is expected to provoke mismatches between scale domains of terrain information and the environmental variable of interest (Smith *et al.*, 2006; Dragut *et al.*, 2009).

1.4 Research question

This research intends to explore the effect that environmental covariates have at different scales on our ability to map soils and to provide a guideline for the selection of the best DSM model inputs’ resolution enabling the up-scale or down-scale of soil information at regional, national or global scale. This could potentially improve the ability of DSM to provide reliable soil data currently demanded by the scientific community, practitioners and policymakers to address the global environmental issues that are threatening the planet.

1.4.1 Hypothesis

The resolution of environmental covariates affects the accuracy of soil predictions in DSM; therefore an analysis of spatial scale will result in an improvement in the accuracy of DSM model predictions.

1.4.2 Aims

The overall aim of this project is to explore a set of methodologies with which to identify the appropriate scale resulting in improved DSM performance in terms of predictive accuracy. The specific aims that have been identified are:

- To evaluate the role of scale and its impact on generating soil taxonomic predictions.
- To investigate and assess several methodologies suitable for the detection of scale patterns.
- To enable practitioners to find a suitable scale to be used in DSM analysis.

1.4.3 Objectives

1 To investigate the effects of scale on DSM analysis.

2 To assess the interaction between pixel and window sizes, with data mining classifiers, for the purpose of modelling soil taxonomic units.

3 To identify, from published literature, methodologies that can be used in quantitative scale detection.

4 To test these methods in the determination of the most suitable DEM pixel size for application in landscape-scale DSM.

5 To develop a multiscale approach for DSM.

6 To develop recommendations on scaling environmental covariates used for DSM.

1.5 Outline

After explaining the motivations and goals of the current thesis, an outline is presented to briefly describe the structure and content of the following chapters:

- Chapter 2 introduces the concepts of scale, its theoretical background and a review of the meanings of scale in the wider context of the environmental sciences. It then specifically reviews the current literature related to scale in DSM and pedometrics.

- Chapter 3 presents the study areas chosen for this research, thoroughly describing their geomorphological and pedological characteristics. It also describes the soil maps and DEM used in the study, the reasons why they were selected and their overall quality. A methodological section of this chapter presents the methods, techniques and approaches applied in the study.

- Chapter 4 is the foundation of the study, as an experimental methodology is applied to evaluate the effects of scale in DSM. The analysis of these results sets the benchmark for all the other approaches and techniques used in the following chapters.

- Chapter 5 critically evaluates the most common empirical approaches developed and used in the environmental sciences to manage scale. In light of Chapter 4 results, an indication of scientific robustness can be concluded for the application of these approaches in DSM.

- Chapter 6 presents wavelet decomposition, a technique capable to manage systems complexity. It is here applied to filter the DEM into a series of independent components at varying scales, to be then used in DSM analysis.

- Chapter 7 tries to tackle the challenge of developing a new multiscale approach using geostatistics. A moving window segmentation technique is developed and tested.

- Chapter 8 concludes the study reviewing all the aims and objectives set at the beginning of the thesis. A perspective on the new findings made in this research, their limitations and future work needed to further develop scale analysis in DSM is also examined.

2 LITERATURE REVIEW

This chapter provides the reader with a comprehensive literature review on the concept of scale. It introduces and describes the impact and importance of scale in DSM and the issues associated to its selection.

2.1 Scale

The concept of scale is perhaps best described by Levin (1992): "*scale represents the window of perception, the filter or the measuring tool through which a landscape may be viewed or perceived*". The environment cannot be studied, modelled or visualized in its full complexity and details. Scale is then important because of its role in features selection and information generalization, it is essentially a form of simplification.

Scale is a complex concept and has many different and often divergent meanings. It is also highly dependent on the context of study and its applications (Goodchild, 1997; Goodchild and Proctor, 1997, Wu and Li, 2009). Two main theoretical views conceptualising scale have emerged:

- Conceptual
- Functional

The *conceptual view* (Meentemeyer, 1989) divides scale depending on the notion of absolute or relative space in which it operates. Absolute space can exist independently from what is in it, while relative space exists only in relation to things and processes. The former view is associated primarily with maps and inventories while the latter with forms, functions and patterns.

The *functional view* (Cao and Lam, 1997; Marceau and Hay, 1999) focuses more on the uses and effects of scale and tries to use the interaction between absolute and relative scales in a more practical way. This has lead into the following classification of the meanings of scale according to spatial, temporal or spatio-temporal aspects (Figure 2.1).

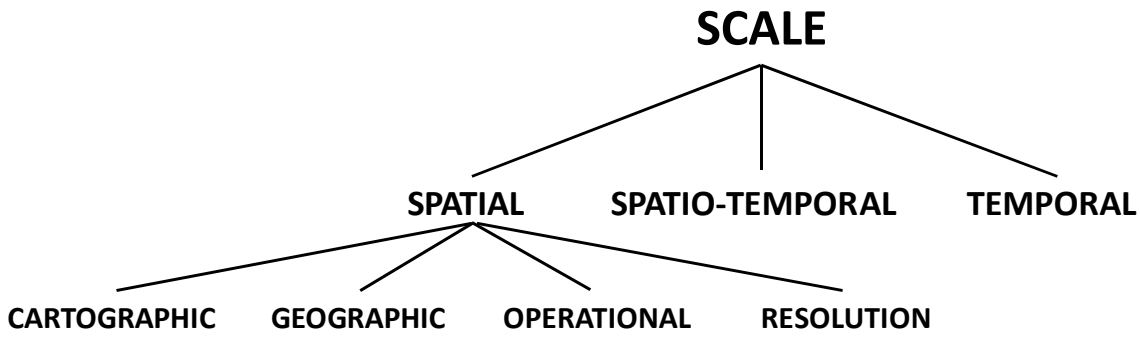


Figure 2.1 - Schematic representation of the concept of scale.

In DSM, scale primarily involves space and this will be the focus of this research. A comprehensive review of the significance of spatial scale in geography and environmental sciences can be found in Lam and Quattrochi (1992) and Goodchild and Quattrochi (1997), defining four spatial meanings of scale:

1. **Representative fraction or cartographic scale**

   The ratio used to scale a feature. It shows the relationship between one unit on the map and one unit on the ground. It is usually shown with a colon such as 1:250,000 meaning that 1 cm on the map represents 250,000 cm on the ground.

2. **Spatial extent or geographic scale**

   The extent or scope of a study or project. Spatial extent defines the total amount of information relevant to a project (Goodchild and Quattrochi, 1997). It is generally defined in terms of area or length, for example for a project area with a square shape, it is represented either by the area measured as the second power of the side or as a length by the size of one of the equal sides. For irregular shapes, it can be either the total area or as a length, it is calculated as the square root of the area.

3. **Process scale or operational scale**

   Process scale refers to the extent at which a phenomenon operates in the landscape. For example in the case of a complex process like erosion by water a range of scales can be identified. Rain splash redistribution and the

initiation of microrills occur at a scale of millimetres. Rill erosion on agricultural hillslopes operates at a scale of meters, while gully erosion can occur on a scale of hundreds of meters, or even kilometres.

4. **Spatial detail or spatial resolution**

   Defined as the shortest distance over which change is noted and thus having a unit of length. For example, the representation of spatial variation described by spatial resolution in a raster dataset is the length of a cell side, as variation within cells is not supported.

In GIS models the level of geographic detail cannot be fully represented in cartographic terms, for example the representative fraction loses significance because there is no set distance in the model to compare to the real world (Goodchild, 2001). In this context spatial resolution better encapsulates scale representing the level of spatial detail or the size of the smallest element in the dataset (in raster datasets pixel size).

The typical effects of spatial scale are:

- Accuracy and precision in both data and modelling. The real-world size of features may not be correctly represented within a GIS. For example, a soil boundary line with a width of 0.5 mm and map scale of 1:20,000 on the ground is actually an area with a width of 10 m.

- The way to collapse and aggregate data in order to make them workable and relevant to the problem investigated. For example, small-scale mapping makes displaying small and fragmented soil series impossible, requiring the creation of a new mapping unit (soil complex) for selected mapping or publication scales.

- The process with which to extract measures of variation and correlation, to make sense of the phenomena in question and to establish theories. Since spatial data are obtained through sampling with particular scales of measurement, the scales of variation observable in spatial data are

inextricably linked to the scales of measurement through which they were obtained (Atkinson and Tate, 2000).

- The approach used to communicate science through graphical representations and visualization. Information is selected and represented in a way that adapts to the scale of the map's display (screen, paper, etc.), not necessarily preserving real world details. Small-scale maps have more simplified features than larger-scale maps because they show a larger area in a smaller display.

Spatial scale of geographic information is still an unsolved issue and a major obstacle both conceptually and methodologically in all the environmental sciences due to the lack of formal laws and rules. As indicated by Jarvis (1995) this is a real scientific challenge as spatial data are scale dependent because of the heterogeneity of processes operating at multiple scales and their non-linear behaviour.

Models are created with a specific process scale in mind and need input data at a certain scale (Bierkens *et al.*, 2000) therefore requiring the input data to change scale. Changing scale or scaling (Figure 2.2), involves transferring information from one scale to another essentially in two types of way:

- Upscaling involves the generalization of information, coarsening the pixel size by interpolation reducing the resolution of the support.

- Downscaling on the contrary involves the decomposition or disaggregation of information, increasing the resolution of the support through model based, regression or stochastic simulation.

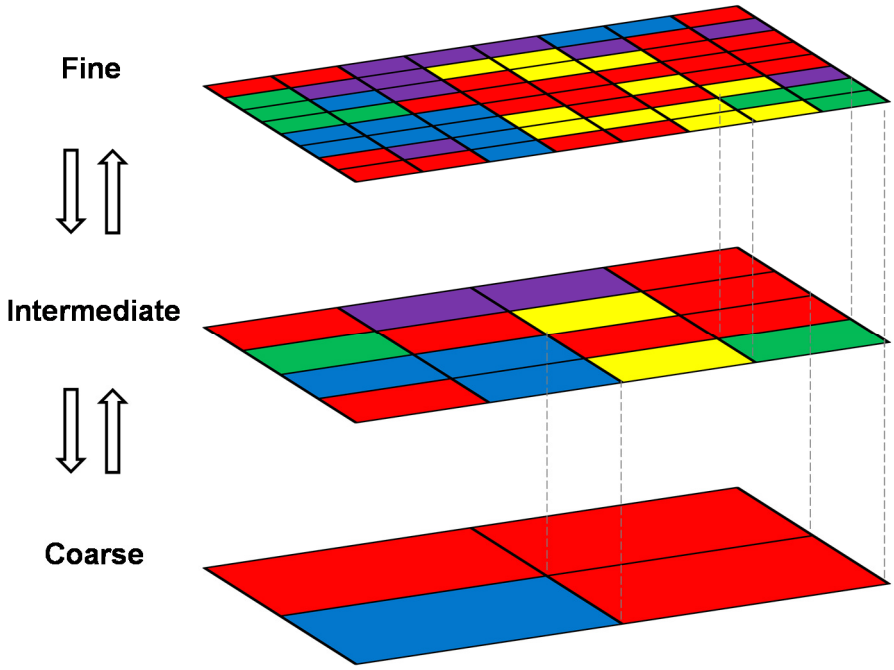


Figure 2.2 - Resampling.

Another operation commonly used in the manipulation of DEMs is passing a filter over the raster to reduce or remove small variations, revealing small-scale patterns or trends in the data. The roving window approach can be considered the standard filter technique in raster GIS operations and in image processing. It determines the new value for a given cell in a raster map using a mathematical function (mean, mode, standard deviation, etc.) of the cell values inside a *n* x *n* neighbourhood (with odd *n*) centred around the cell of interest. The window is moved one cell at a time across the raster map, until the whole area is processed (Grohmann and Riccomini, 2009).

In this research, scale will be regarded as the physical dimension of a phenomenon or process in space expressed in spatial units (pixel and roving window sizes).

2.2 Scale in DSM

Soil scientists are still debating if soil can be better symbolized by discrete physical bodies large enough to be classified in taxonomic systems (intrinsic scale / taxonomic scale) or if soil should be considered as a continuum and represented by raster layers. The consensus is that soils occur as a continuum, but classes are still used to describe soil map units. Conceptually soil taxonomy is a useful way of describing spatial variability in terms of scale, implicitly embedding a scale into the data, where coarse resolution variations are represented in the orders, intermediate variations in the suborders and fine resolution variations at the series level.

Traditional soil survey, based on categorical units, has elements of both science and art (Avery, 1987). Expert knowledge is intensively used to create classification maps generally produced during soil surveys in the field where surveyors would delineate soil classes on their survey sheets. The mix of expert knowledge, empirical data and supplementary data like aerial photographs or topographical maps, makes it challenging to produce replicable information. Moreover, as the definition of taxonomic units can be somehow interpretable, different soil surveyors could potentially create two different soil maps for the same area (Goodchild, 2009). Surveyors in the field need to balance the necessity to classify soils into taxonomic units that must be reasonably homogeneous with respect to the soil profile and at the same time are representative of a region appearing in reasonably large parcels of land. A compromise on the delineation of soil classes is generally made between spatial considerations to prevent fragmentation making the map of no practical value and representativity of the soil profiles to maintain data significance and quality (McBratney and Webster, 1981). Field soil investigations can involve free or grid surveying. The most widely used type of survey is “free”, where observations are made by the surveyor at irregular intervals and varying intensities reflecting the estimated complexity and predictability of the soil patterns. The consequence is

that the surveyor implicitly embeds a scale into their observations. How this expert knowledge on scale can be represented, should be of interest to scientists.

Classifying and producing maps has always been the main focus of soil surveying, primarily to identify soil types or other soil properties at a particular point in the landscape. The purpose of a soil map is then to supply users with information regarding soil types or properties of a described region. Its value lies in the quality of the information represented and the soil units identified, as the variability of soil conditions within the soil map units must be less than the overall variability in the landscape. Soil surveyors have traditionally relied on the soil map unit concept which is a section of the landscape with similar soil properties, geomorphology, hydrology, ecology, land use and other landscape features (Brus and Lark, 2013). This approach is strongly scale dependent as soil map units are equivalent to a class in soil taxonomy only at large-mapping scales. In complex landscapes or small-scale mapping (which shows less detail), a map unit may only represent the union of spatially related classes, imposing limitations on the information available.

2.2.1 Scale of soil spatial variation

Spatial heterogeneity of soil parameters has proved to be a major problem in the representation of soil properties (Sinowsky and Auerswald, 1999) especially when spatial variability differs significantly between the scale of observations and the one in which processes are active (Geng *et al*., 2010).

Table 2.1 - The hierarchical levels of scale in soil science (from Addiscott, 1998).

Scale	Unit
i + 4	region
i + 3	interacting catchments
i + 2	catena or catchment
i + 1	field (polypedon)
i	pedon
i - 1	profile horizon
i - 2	peds, aggregates
i - 3	mixtures
i - 4	molecular

In the context of soil science, the scale diagram (Table 2.1) shows the hierarchical levels of scale in soil systems, with the pedon as the base unit and defines other levels as a reference to it. Each element of level *i* is at the same time part of level *i* + 1 and includes element of level *i* - 1 as each pedon is part of a polypedon and includes profile horizons. Hierarchy theory allows to easily understand that scale affects not only how units are described by their characteristic list of descriptors but most importantly which units are described. Changing scale thus implies a change in organisational level as each component is nested within other levels. Nested hierarchies include levels which consist of, and contain, lower levels as presented in Table 2.1.

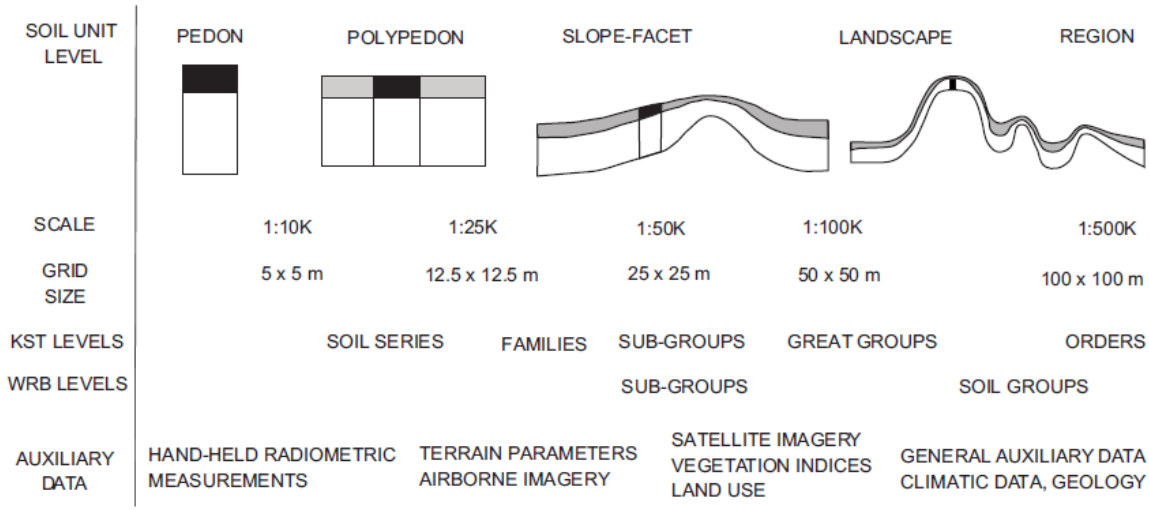


Figure 2.3 - Relationships between the level of soil unit, scale, grid resolution, taxonomy and auxiliary data (from Hengl, 2003).

Hengl (2003) effectively encapsulates this concept with a hierarchical level diagram of DSM (Figure 2.3). In this diagram it is possible to link soil unit levels with approximate cartographic scales, potential grid sizes, soil classification levels (KST and WRB) and typical environmental covariates used in DSM analysis as auxiliary data. It is not clear, however, if the grid size proposed is indicative of the soil unit represented or of the resolution of the DSM model to be used in the creation of spatial soil information.

Covariates, the environmental factors recognized as governing soil formation, vary at different scales and this spatial variation at some scales may be more strongly correlated with soil than at others (Lark, 2006). Soil forming factors have different domains with distinctive scales, for example geology operates at a coarser scale than land use.

Within this multitude of different domains, processes relate to each other in five broad ways:

- Joint processes that are equivalent across scales due to the hard links between them;

- Parallel processes that are consistent across scales due to the type of soft links between them;

- Iterative processes that are coherent across scales due to the type of soft links between them;

- Consecutive processes that are comparable across scales due to the type of soft links between them;

- Independent processes that are complementary across scales because the absence of any link between them.

Depending on the type of relation across scales of the processes, the strength of the scale link will differ: ranging from ‘Hard’ links for fully equivalent processes developed jointly across scales, through ‘Soft’ links with different degrees of linkage across one or more process scales, to ‘No’ links for processes developed independently that do not share any scale links between them.

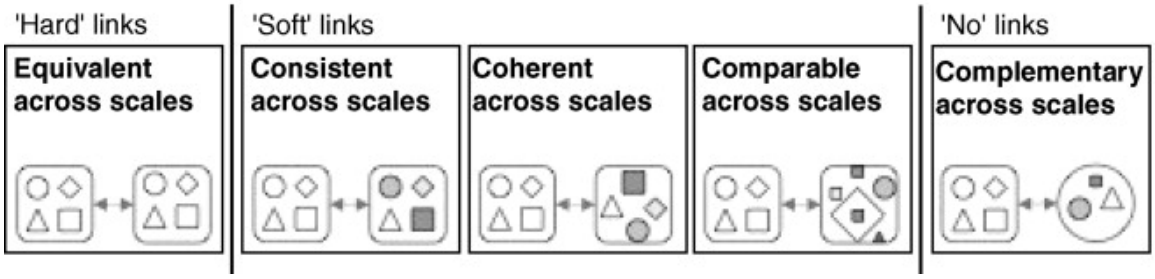


Figure 2.4 - Linkage between soil processes across scales (from Zurek and Henrichs, 2007).

Soil forming factors fall in the ‘No’ links and the ‘Soft’ links category (Figure 2.4), with different degrees of linkage between them and the soil type resulting from their interactions. Despite the need for more research to accurately evaluate their linkage with soil and their overall interaction at different scales, some relationships start to become apparent especially between relief and soil types (Pain, 2005). Therefore some scales will be more powerful in prediction than others and this statement should be taken into account when applying DSM techniques.

McBratney *et al.* (2003) reviewed pedometric methods for soil prediction in literature and suggested three main resolutions of interest (Table 2.2), which are:

- < 20 m (local extent);
- 20 m – 2km (catchment to landscape extent);
- > 2 km (national to global extent).

Table 2.2 - Suggested resolution and extent of digital soil maps (modified from McBratney *et al.*, 2003).

Pixel size and spacing	**Cartographic scale**	**Resolution 'loi du quart'** [a]	**Nominal spatial resolution**	**Extent** [b]
< 5 x 5 m	> 1:5,000	< 25 x 25 m	< 10 x 10 m	< 0.5 x 0.5 km
5 x 5 to 20 x 20 m	1:5,000 – 1:20,000	25 x 25 to 100 x 100 m	10 x 10 to 40 x 40 m	0.5 x 0.5 to 2 x 2 km
20 x 20 to 200 x 200 m	1:20,000 – 1:200,00	100 x 100 to 1,000 x 1,000 m	40 x 40 to 400 x 400 m	2 x 2 to 20 x 20 km
200 x 200 to 2,000 x 2,000 m	1:200,000 – 1:2,000,000	1,000 x 1,000 to 10,000 x 10,000 m	400 x 400 to 4,000 x 4,000 m	20 x 20 to 200 x 200 km
> 2,000 x 2,000 m	< 1:2,000,000	> 10,000 x 10,000 m	> 4,000 x 4,000 m	> 200 x 200 km

[a] According to the French soil scientist Jean Boulaine (1980), the smallest area discernible on a map is 0.5x0.5 cm or one quarter of a square centimetre, hence, the term 'loi du quart'.
[b] Calculated as minimum resolution times 100 (pixels) up to maximum resolution times 10,000 pixels (McBratney *et al.*, 2003).

Despite the abundance of information on pixel size, spacing, resolution and scale, the table has only an indicative purpose being an expanded version of a previous classification of soil information extents presented by McBratney *et al.*, (2000). The intent of the table was, and still is, to link soil demands with pedometric techniques: the local extent (< 2 km extent) answering precision agriculture demands, the catchment/landscape extent (2 - 200 km extent) for environmental and water management demands and finally the national/continental/global extent (> 200 km extent) answering climate change and food security demands.

The presented pixel size should be considered as a good approximation of what is expected by different audiences (policy makers, landscape managers and the wider scientific community) from digital soil maps as inputs for their policy, plans or analysis.

2.2.2 The issue of scale in DSM

Despite the recent growth and operational status of DSM, one existing and foreseeably growing issue for users of digital soil information is the disparity of spatial scales between what is required and what is actually available to adequately address soil-related questions posed to the soil science community (Grunwald *et al.*, 2011). While the demand for soil information is growing, the quantity of data collected in the field is reducing, mainly due to economic reasons (Sanchez *et al.*, 2009). In the absence of conducting new soil surveys and not being able to acquire the original legacy soil information (soil point data explored on the ground) as a means of creating user-specified soil information products, spatial scaling procedures will provide a useful solution (Malone *et al.*, 2013).

If the resolution of the output can be effortlessly identified and classified, the pixel size for environmental covariates used as inputs in DSM soil predictions is still an unsolved issue. DEMs are the finest resolution information currently available for DSM analysis; they are scale benchmarks directly influencing the scale of the model and the output. A common approach used by the DSM community is to apply the finest resolution available of DEM accepting that this will always improve the accuracy and precision of the prediction. As shown in Table 2.3 large scale DSM studies predicting soil taxonomic units from national to continental scale (from 900 km^2 to 1,600,000 km^2) use very fine resolution DEMs ranging from 25 to 100 m.

Table 2.3 - Summary of large scale DSM papers predicting soil taxonomic units (2007-2008) published in Geoderma and Soil Science Society of America Journal (modified from Grunwald, 2009).

Reference (Year)	Soil attribute	Region	Covariates	Spatial Extent	Spatial Resolution
Grinand *et al.* (2008)	unit	France	S, O, R, P	900 km^2	50 m
Minasny and McBratney (2007)	class	Australia	S, O, R [RS]	~ 3,500 km^2	25 m
Bockheim and McLeod (2008)	type	Antarctica	S, R	6,692 km^2	25 m
MacMillan *et al. (2007)*	eco type	Canada	S, C,O, R, P	82,000 km^2	25 m
Hengl *et al.* (2007)	group	Iran	S, O, R	1.6 mill km^2	100-1000 m

Scale analysis for DSM has not yet received the attention needed in the literature as stated by Behrens *et al.* (2010b), as only in the last ten years some scientists have started to focus on this critical aspect of DSM modelling (Malone *et al.*, 2013). As Papritz *et al.* (2005) suggested, soil information may be available at one spatial scale, but this might not be suitable for the purpose of the investigation and it may be required either at a finer or coarser scale.

The influence of DEM pixel size on DSM models has been discussed by Thomson *et al.* (2001), who investigated quantitative soil-landscape modelling by using empirical models for the prediction of the spatial distribution of soil attributes (A-horizon thickness, depth to secondary carbonates and a soil colour index related to soil organic carbon content). Smith *et al.* (2006) investigated which DEM resolution produced the most accurate digital soil surveys for a particular landscape, focusing on a GIS-based soil-mapping application. Their conclusion supported the idea that high resolution DEMs do not always produce the highest accuracy. On the contrary, Zhu *et al.* (2008) argued that for soil knowledge-based

soil mapping, DEM resolution was not as important as neighbourhood size in computing the required terrain attributes. Behrens *et al*. (2010a) adopted a data mining approach using random forest classifiers to assess scale properties of terrain attributes used in the prediction of soil types. Behrens *et al.* (2010a) concluded that each soil class was best predicted by different combinations of terrain attributes filtered at different scales. These findings suggest that the contradictory results presented in the literature still need re-examination with different data mining approaches and validation in larger areas with different soil landscapes. More research and time is still needed to answer this unresolved issue and to fully understand the complexity of scale, but intermediate goals such as improving prediction accuracy or reducing computational time and model complexity can be achieved. This can be done by applying existing techniques developed in other scientific domains to DSM or by creating a brand new approach capable of incorporating scale analysis in DSM, so ultimately advancing DSM and offering the possibility to improve knowledge of spatial variation on soil distribution.

3 MATERIALS AND METHODS

3.1 Study areas

Three study areas were selected in Ireland (Figure 3.1) in the counties of Leitrim, Meath and Tipperary North where detailed reconnaissance soil surveys had been carried out in the past.

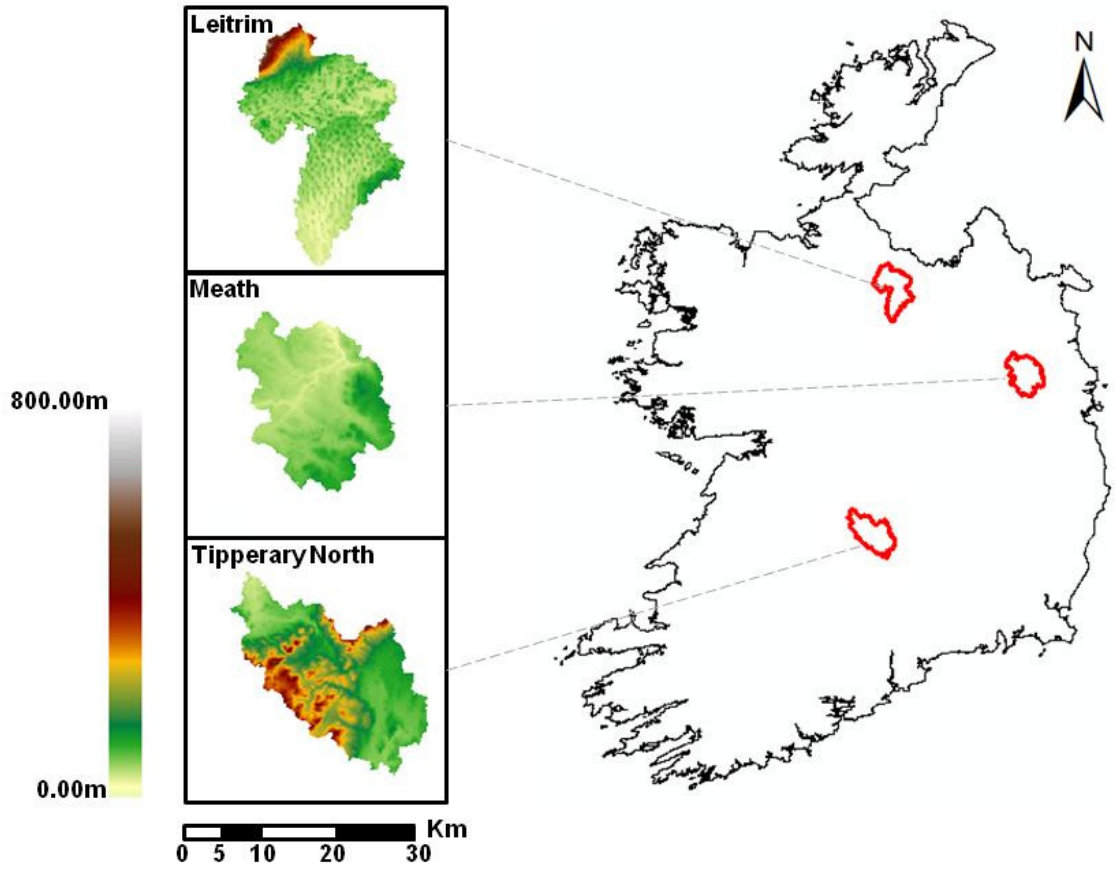


Figure 3.1 - Location of the three study areas in Ireland.

Area 1 of 370 km^2, located in the county of Leitrim, is situated in the Drumlin Belt and consists of hundreds of hillocks with an elongated shape formed by glacial movements on unconsolidated till.

Area 2 of 337 km^2, located in the county of Meath, is situated in the Central Plain and is a large low-lying region underlaid by limestone rocks and covered in glacial drift.

Area 3 of 374 km^2, located in the county of Tipperary North, is situated in the Southern Hills where shale plateaux had been eroded leaving steep slopes.

The selection of these three study areas was dictated by the fact that only few counties in Ireland were covered by the detailed reconnaissance soil survey (as presented in the Soil data section 3.2.1), so the chosen areas had to be within those surveyed counties. Moreover in order to accurately assess the effects of scale in DSM, sites with different soils and geomorphologies were selected, including fine resolution features like the drumlins in Leitrim, steep slopes in Tipperary North and coarse scale lowland in Meath, making these three areas ideal for the study.

3.1.1 Soils and landscapes of Ireland

Soils in Ireland are derived mainly from glacial deposits making the relationship between parent material and soil formation naturally complex (Gallagher and Walsh, 1943; Cruickshank, 1997). Transported drift is the main factor determining the landscape and soil types in Ireland in combination with the predominant temperate humid climate. As a result three pedogenetic processes have been described (Gardiner and Radford, 1980b) to have the strongest effects on soil formation:

- Leaching, whereby soluble nutrients and colloids are eluviated down the profile by percolating water;

- Gleisation, whereby ferric iron compounds in the soil are chemically reduced to ferrous compounds and segregated into mottles and concretions, under water-logged (anaerobic) conditions;

- Calcification, whereby calcium carbonate from weathering of limestone rich parent materials precipitate and accumulate resulting in calcareous soils, which at the surface are mixed with organic matter.

The characteristic soil landscape in Ireland comprises:

- large areas of fertile grey-brown podzolics soils (now termed luvisoils by Jones *et al*., 2011) in the well and moderately well drained areas of the plains;
- less fertile acid brown earths where the drift material is derived from acid parent materials which is poor in lime;
- gleyed soils where the internal soil drainage is poor because of impermeability or high water table;
- on hills thin acid peaty soils are juxtaposed with blanket peat.

In order to organize and systematically categorize soils based on common characteristics, a soil classification system was developed in Ireland based on the United States Department of Agriculture in 1938 (Gardiner and Radford, 1980b). Establishing hierarchies of soils types, based on their distinctive properties, soil taxonomy allows soil scientists to rationalise and understand the relationships between soils. The original soil classification system used in Ireland was based on two levels: Great Soil Groups and Soil Series. The Great Soil Groups divide the soils according to their soil profile characteristics, type and arrangement of horizons, soil moisture levels, temperature regimes and base status. There are ten Great Soil Groups (Podzols, Brown Podzolics, Brown Earths, Grey-Brown Podzolics, Blanket Peats, Gleys, Basin Peats, Rendzinas, Regosols and Lithosols). These groups are further subdivided at series level based on texture, parent material and drainage status. Soil Series are defined by Gardiner and Radford (1980a) as "*a collection of soil individuals essentially uniform in differentiating characteristics and in arrangement of horizons*". A series is usually named after the area in which it is most widely distributed or in which it is best expressed. The series can be separated into phases on the basis of certain features such as texture of the surface layer, slope, stoniness, or some

other characteristics that affect use of the soils by man. Soil series and phases of soils occur in an intricate pattern, making it difficult to be shown separately on a soil map. Therefore, they are joined into a soil complex. The name of a soil complex consists of the names of the dominant soils, joined by a hyphen.

Under the ISIS project, a new classification framework has been developed with two extra levels between Great Soil Groups and Soil Series, namely Soil Groups and Soil Sub-Groups. The General Soil Map (Gardiner and Radford, 1980a) only delineates Great Soil Groups because of scale limitations.

Table 3.1 - Soil classification of the three study areas by Great Soil Groups.

	Leitrim		**Meath**		**Tipperary North**	
	[km^2]	**[%]**	**[km^2]**	**[%]**	**[km^2]**	**[%]**
Podzols	-		-		17.73	5%
Brown Podzolics	-		-		124.16	33%
Brown Earths	-		-		27.51	7%
Grey-Brown Podzolics	-		234.12	70%	130.22	35%
Blanket Peats	30.25	8%	-		11.79	3%
Gleys	279.13	75%	91.97	27%	54.47	15%
Basin Peats	53.99	15%	11.02	3%	8.24	2%
Rendzinas	3.61	1%	-		-	
Regosols	-		-		-	
Lithosols	-		-		-	
Water	3.89	1%	-		-	

As shown in Table 3.1 the three study areas have a distinct soil landscape, according to the original soil county maps (Appendix C):

- Leitrim with 75% of its area covered in gleys is characterised by grey/blue waterlogged soils with poor drainage and high water table with poor drained peats on the drumlins;

- Meath with 70% of its area covered in grey-brown podzolics is characterised by soils with a calcareous parent material which offsets the effect of leaching with loss of nutrients and restricts the resulting podzolisation process;

- Tipperary North with 35% of grey-brown podzolics and 33% of brown podzolics is characterised by two distinctive areas the flat lowland with the grey-brown podzolics and the high relief area with the brown podzolics formed under the influence of podzolisation but less depleted and without iron pan.

3.1.2 Leitrim

Glacial movements had a great control over the undulating landscape of Leitrim, depositing huge amounts of drift and affecting the underlying geology. This material was transported only locally through lateral movements and maintains the same chemical composition of the bedrock (Aalen *et al.*, 1997). Soils are relatively young as soil formation started only after the last glacial episode (Midlandian glaciation about 15000 years ago) when the earlier land cover was removed from the surface.

Grasslands dominate the landscape with a dense network of hedgerows on the drumlins (Figure 3.2) and limited broad leaf woodlands along the main rivers. In between the drumlins in the narrowest strips, raised bogs are present and limited wetlands along the course of the Shannon.

Figure 3.2 - Leitrim characteristic drumlins (source: European Forum on Nature Conservation and Pastoralism).

Drumlins are small elongated hills of boulder clay deposited by glaciers, the elliptical shape (up to 800 m in length, 300 m in width and 120 m in height), created by the movement of unconsolidated till beneath the ice, is parallel to the direction of ice flow (Stokes *et al.*, 2011). The soils on the drumlins are poorly drained and impermeable due to the dense and compact clay of upper carboniferous shale composition (An Foras Taluntais, 1973).

3.1.3 Meath

Meath, with its relatively rich soils for agriculture and pasture, supports a diverse farming sector from potato production and grass growing to cattle rearing for beef or dairy. The agricultural landscape dominates with hedgerows, ditches and open drains delimiting fields (Figure 3.3).

Figure 3.3 - Agricultural landscape of Meath (source: National University of Ireland Galway, School of Geography and Archaeology).

The influence of carboniferous shales due to the glaciation movements of drifts from which limestone was dissolved out characterise the soil formation of Meath (Finch *et al.*, 1983). The soils of this lowland region are mainly grey-brown podzolics with gley in the low-lying positions.

3.1.4 Tipperary North

Soils in Tipperary North are older than the other two study areas as the weathering process had more time for the development of soil horizons in the profile as the south of the country was not affected by the last Midlandian glaciation but only by the previous Munsterian 100,000 years ago.

The underlying geology of Tipperary North is formed from two main rock formations associated with their own distinctive landscape. The limestone lowlands with their carboniferous series and covered by glacial sediments divided by the Silvermine Mountains (highest point at 468.8 m AOD) with mix composition (red sandstone, mudstone and slate) and different erosion rates (Finch and Gardiner, 1993). This has resulted in an extremely diverse soil landscape with seven different soil great groups present, dominated by brown podzolics and

grey-brown podzolics with gleys, podzols, brown earths, blanket peats and basin peats.

Figure 3.4 - Tipperary North (source: ENFO Environmental Information Service).

Grey-brown podzolics originated from the glacial till rich in limestone and have a heavy texture but good drainage so mostly suited to pastoral uses (Figure 3.4) while the brown podzolics located at the foothills of the Silvermine Mountains are well drained but poor in nutrients and very suitable for woodlands and forest plantations.

3.2 Data sets

3.2.1 Soil maps

In 1959 a National Soil Survey (NSS) of Ireland was launched by the then agricultural institute *An Foras Taluntais* (now Teagasc) with the purpose of surveying, classifying and mapping soil resources of Ireland (Lee *et al.* 2005). The NSS had been operated at three levels of resolution:

- Very detailed studies at field scale (1:2,500).

- Semi-detailed studies at county scale (1:126,720) with soil series as unit of mapping.

- A combined detailed and general reconnaissance with the purpose of deriving a map at national scale (1:575,000) with soil association as a unit of mapping.

These different levels of organization have resulted in a fragmented soil information system with partial coverage depending of the specific scale of study.

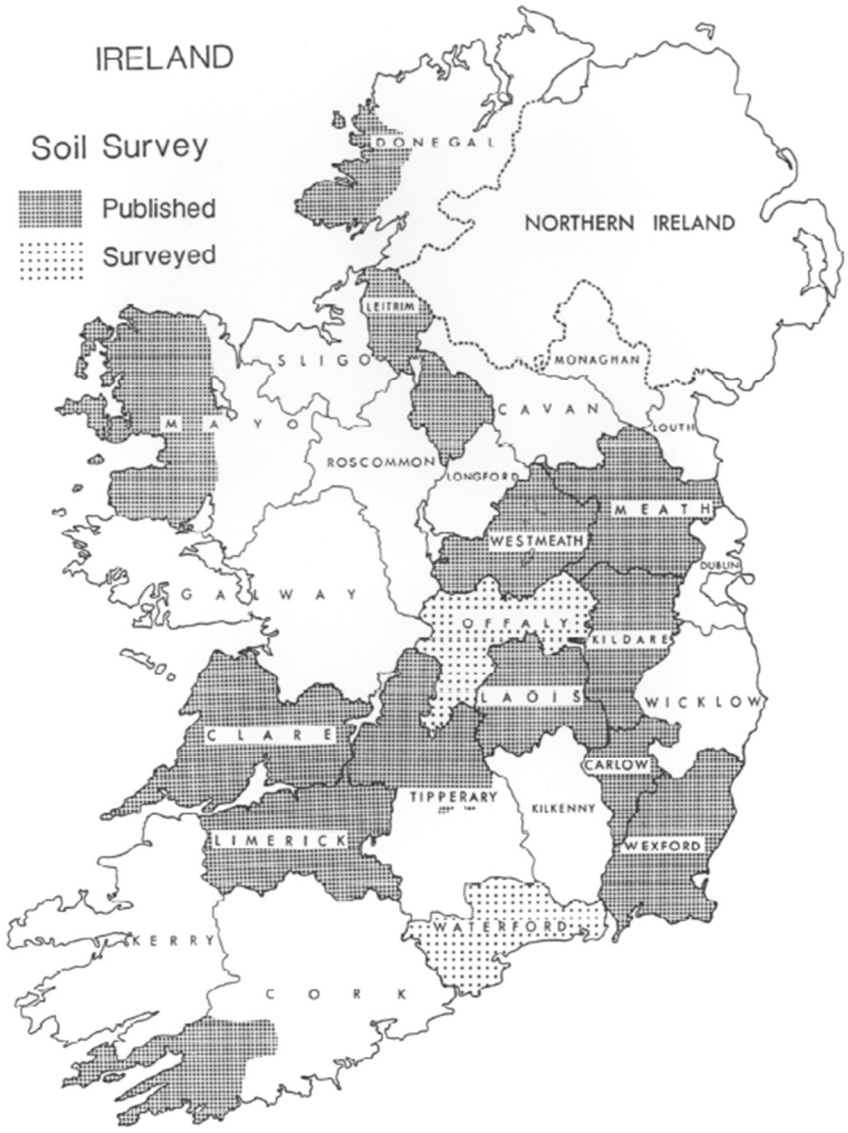


Figure 3.5 – Counties covered by detailed reconnaissance soil surveys in Ireland (from Lee *et al.*, 2005).

The Irish General Soil Map was first published in 1969, when only four counties and two regions had been surveyed at detailed reconnaissance scale, while surveys at reconnaissance level had been in operation in another ten counties. It lacked detail and precision for many areas (Gardiner and Radford, 1980a) and it was updated in a second edition released in 1980 (Appendix B) when further field work was completed. It covers the entire country at a scale of 1:575,000 with soils mapped in 44 associations (Gardiner and Radford, 1980a), while the detailed reconnaissance programme focused on soil series on a county basis that was discontinued in 1988 resulting in only 44% of the country surveyed (Figure 3.5) and the detailed studies were concentrating merely on individual agricultural experimental stations.

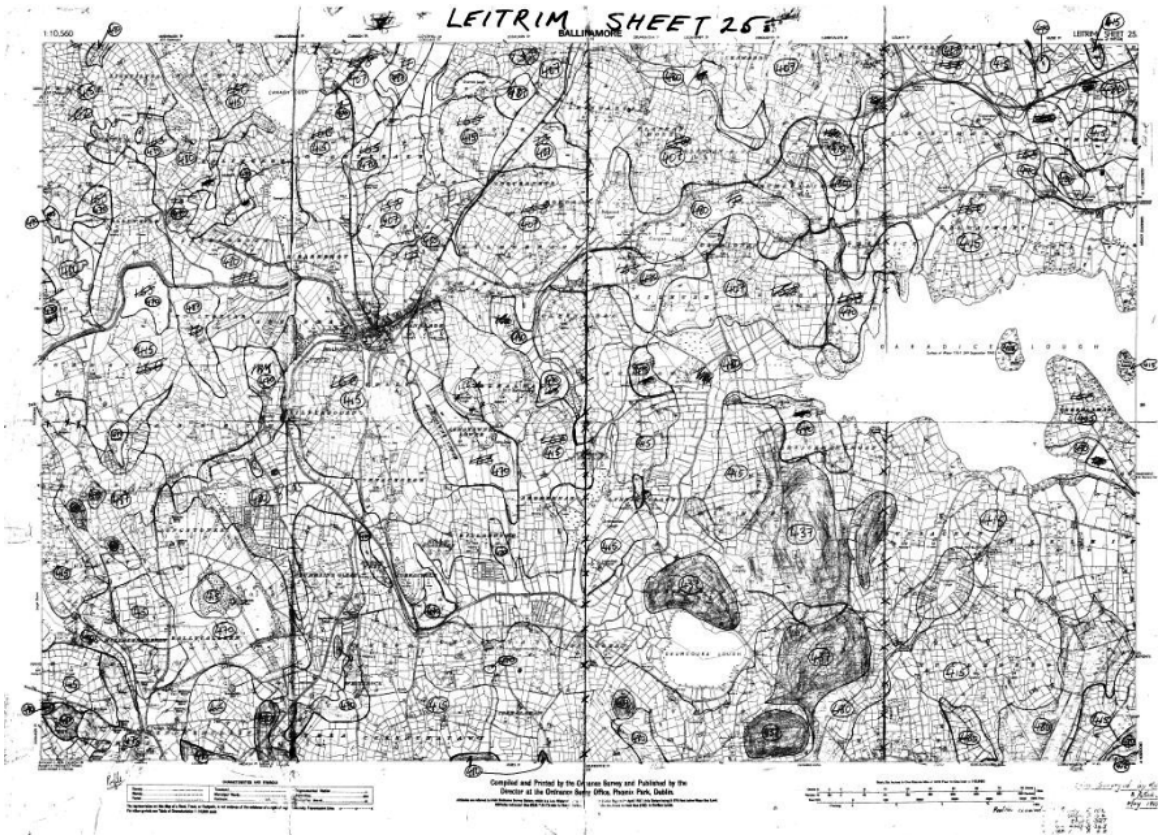


Figure 3.6 - Original survey sheets drawn by the soil surveyors in the field delimiting the different soil series detected.

Following the cessation of the field programme a new direction of research was soon taken involving the digitisation and data capturing of existing soil maps

(Figure 3.6) creating a soil information system and database. This plan was carried out on the detailed reconnaissance field maps drawn by the surveyors at a scale of 1:10,560 (6 inches to the mile) which were generalized to 1: 126,720 for publication (Lee *et al.*, 2005).

In this research all the DSM analysis performed are based on the original six inches maps digitised and provided by Teagasc (Figure 3.7).

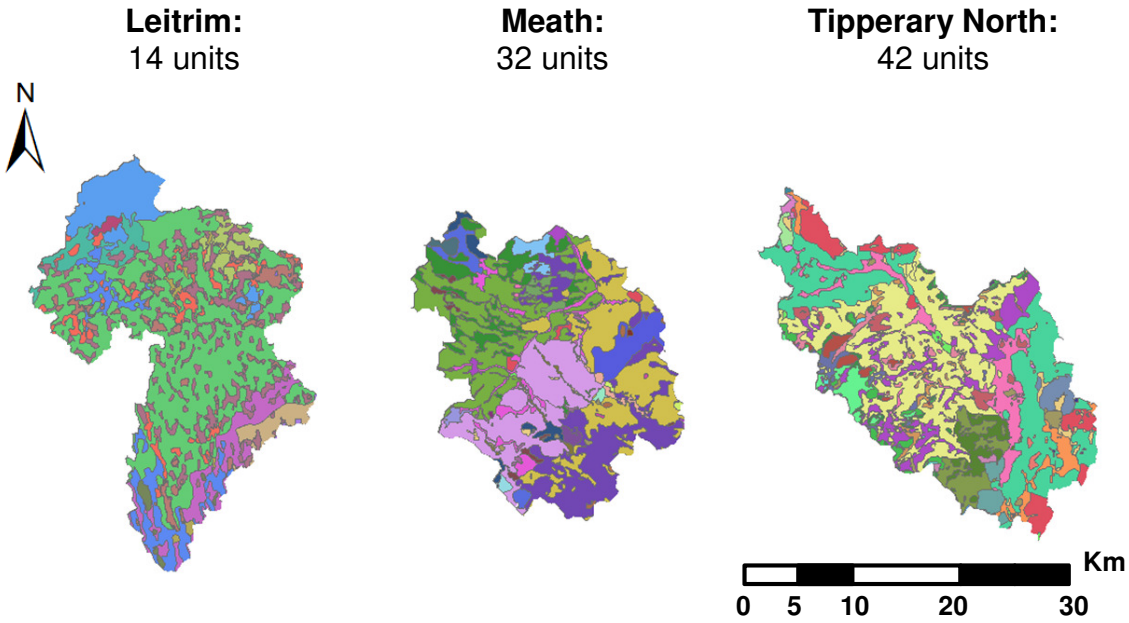


Figure 3.7 - Six inches to the mile maps for the three study areas classified by Great Soil Groups.

They contain a range of soil series, complexes and phases for each study area:

- Leitrim - Allen, Ardrum, Aughty, Ballinamore, Ballyhaise/Corriga Complex, Clooncarreen, Corriga, Drumkeeran, Drumkeeran Peaty Phase, Garvagh, Howardstown, Mortarstown/Kinvarra Complex, Rinnagowna and Unclassified;
- Meath - Allen, Ashbourne, Ashbourne Shaley Phase, Baggotstown, Baggotstown/Crush Complex, Ballincurra, Banagher, Boyne Alluvium, Burren, Camoge, Drombanny, Dunboyne, Dunboyne Gravelly Phase, Dunboyne Shaley Phase, Dunboyne/Ashbourne Complex, Dunsany, Feale, Glane, Gortnamona, Howardstown, Kells, Ladestown, Ladestown/Rathowen/Banagher Complex, Man Made, Patrickswell,

Patrickswell/Baggotstown/Elton Complex, Patrickswell/Howardstown Complex, Rathowen, Rathowen/Street Complex, Street, Turbary, Urban;

- Tipperary North - Allen, Aughty, Baggotstown, Ballincurra, Ballynalacken, Banagher, Borrisoleigh, Borrisoleigh Steep Phase, Borrisoleigh/Knockshigowna Complex, Burren, Burren Rocky Phase, Camoge, Camoge/Milltownpass Complex, Doonglara, Dovea, Drombanny, Elton, Feale, Gortaclareen, Howardstown, Kilcommon, Kilcommon Peaty Phase, Knockaceol, Knockaceol Peaty Phase, Knockastanna, Knockastanna Peaty Phase, Knockastanna/Knockshigowna Complex, Knocknaskeha, Knocknaskeha/Doonglara Complex, Large Rock Outcrop, Man Made, Patrickswell, Patrickswell Bouldery Phase, Patrickswell Lithic Phase, Patrickswell/Baggotstown/Elton Complex, Pollardstown, Puckane, Slievereagh, Turbary Complex, Turbary/Knockastanna Complex, Turbary/Mountainous Complex, Urban.

As a simple rule of thumb, when a paper map is digitized in GIS to create a dataset, the spatial resolution of the data is approximately 0.5 mm at the scale of the map (the so called pencil line) so in the case study the map at 1:10,560 has a spatial resolution of about 5.3 m.

3.2.2 Digital Elevation Model

A DEM is a grid based digital dataset of the topography of an area and, as every raster model, is covering continuously an area providing an elevation value for each raster cell. All pixels of the dataset are covering a defined area (pixel unit) and depending on the pixel size their number can change (Burrough and McDonnell, 1998). DEMs play a major role in DSM by providing information on topography and local landforms which have a clear impact on soils by controlling soil forming processes such as water and sediment movement (Florinsky, 1998). The most detailed DEMs available at national level are traditionally derived from existing contour lines and survey data from historical paper maps or can be generated at lower resolutions from remote sensing data mainly satellite imagery.

As technology moves forward and radar technology becomes less costly and more available, high resolution LIDAR DEMs will become the norm in the near future. This remote sensing technology uses laser scanning to collect elevation data by emitting thousands of pulses every second collecting a cloud of heights. It offers great vertical accuracy between 0.15 m and 0.25 m (Leica ALS50) and a point density up to 10 heights per square metre (O'Neill, 2009). LIDAR is the finest resolution information available in Ireland but covers only 39.7% of the country mainly cities and mayor urban areas (Fealy, 2006). The three study areas selected for the project lie in rural areas and have been only partially covered by LIDAR flights: Leitrim (7.2%), Meath (60.2%) and Tipperary North (8.58%). Given the sparse coverage of LIDAR data in Ireland, this type of elevation information was not deemed ready yet for national mapping applications, such as the ISIS project.

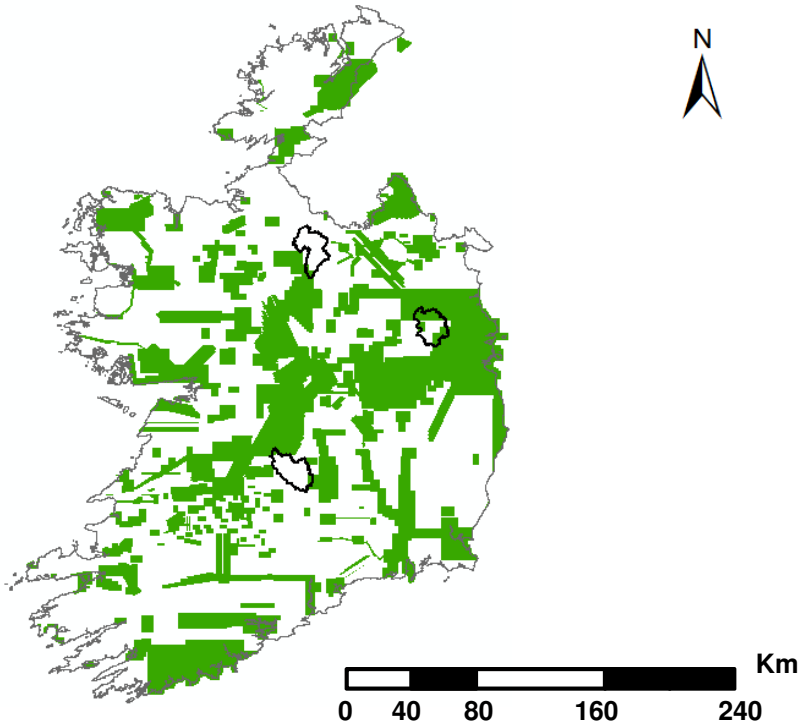


Figure 3.8 - LIDAR coverage for the Republic of Ireland (2005-2012 flights).

The finest resolution elevation data available in Ireland covering the entire country and selected for the ISIS project is the Environmental Protection Agency (EPA) DEM at 20 m resolution. This DEM is a GIS dataset in raster format of elevation

data (ARC/INFO numerical floating point decimal) at national level created in 2005 by the EPA using Ordnance Survey of Ireland (OSI) data and provided for this research by Teagasc. It was generated using the ANUDEM software (Hutchinson, 2007) by spline interpolation using OSI vector spot heights, drainage lines and contour data at scale 1:50,000 as inputs. The contour lines used had a 10 m vertical spacing (Preston and Mills, 2002).

In terms of accuracy, the DEM has been corrected both hydrologically and morphologically. All apparent height anomalies have been removed creating a hydrologically exact drainage network and improving the description of terrain shapes. In conclusion, this DEM with spatial resolution of 20 m is a good compromise for DSM analysis and is deemed suitable for this research.

The three study areas present different topographies recognizable from their unique descriptive statistics presenting measures of central tendency, statistical dispersion and shape of the distribution (Table 3.2). For example, the standard deviation of elevation is a measure of local relief, with the highest standard deviation found in areas with long and steep slopes such as on the hills of Tipperary North and on the drumlins of Leitrim.

Table 3.2 - Descriptive statistics of the three study areas DEMs.

	Leitrim	**Meath**	**Tipperary North**
Average Height [m]	94.83	77.29	161.97
Min Height [m]	7.73	29.00	47.00
Max Height [m]	584.46	155.61	468.83
Median Height [m]	74.77	71.20	138.67
Standard Dev. [m]	70.25	20.22	78.71
Skewness [ ]	3.26	0.84	0.84
Kurtosis [ ]	14.81	3.06	2.85

To visually appreciate the distributions of the three populations, histograms are presented in Figure 3.9. Tipperary North clearly shows two peaks characteristic

of a bimodal distribution suggesting that there are two separate populations: one with low values of elevation (flat area) and one with higher values and long tails possibly due to high variability (high relief). Meath has a very compact distribution with a small standard deviation and high frequency concentrated at the mean value typical of somewhat homogenous populations.

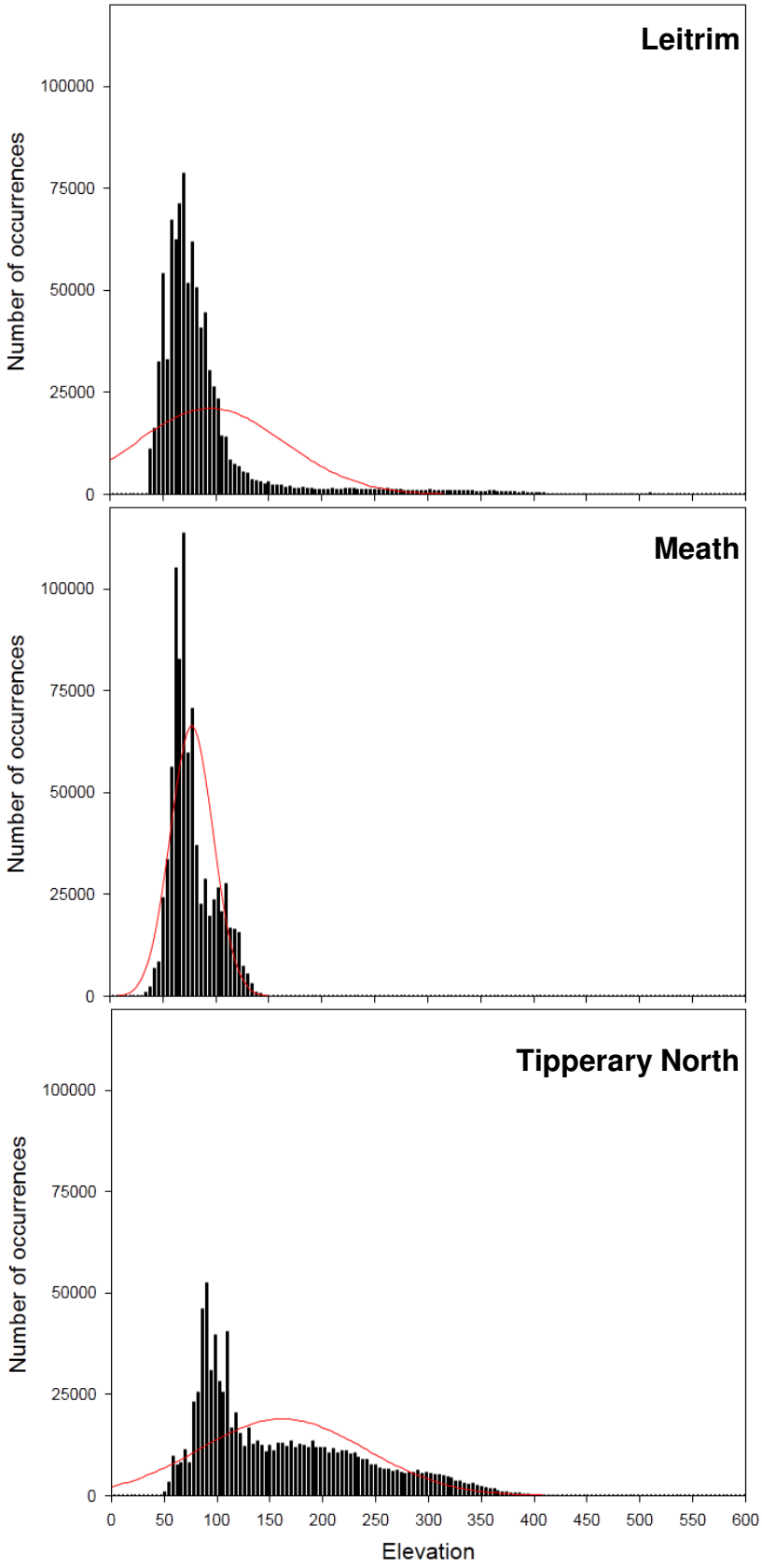


Figure 3.9 - Elevation histograms and distribution for the three study areas.

Leitrim has a long tail distribution highly skewed towards high values of elevation, this was expected as a small area of high relief is included in the north-west sub-catchment.

3.2.3 Terrain attributes

As already discussed in Chapter 2, the most suitable DEM resolution, from which terrain attributes are derived, depends on the scale of the processes controlling pedogenesis and this is landscape dependent. As a consequence there is a real need for a more landscape-scale pedology that could offer the connection between soil processes involved in soil formation and soil surveys, creating the foundations for up scaling soil information to regional, national and global scale (Pennock and Veldkamp, 2006). However, the predictive power of any soil-landscape model that employs data derived from DEM is clearly dependent on their quality and scale. If these generalizations are within the scale threshold of the landscape processes that are operating in the environment under study there are no problems but if they are greater or finer than the spatial resolution of these processes, any result derived must be treated with caution.

The DEM was used to derive eleven terrain attributes indicative of the soil-landscape relationships controlling the spatial distribution of physical, chemical and biological soil properties: slope gradient, aspect, curvature, plan curvature, profile curvature, slope height, valley depth, normalized height, standardized height, mid-slope position and convergence index (Behrens *et al.*, 2010b; Florinsky, 2011). The second-order finite difference algorithm of Zevenbergen and Thorne (1987) was used to calculate the local morphometric terrain attributes using SAGA GIS Terrain Analysis Morphometry library (Bock *et al.*, 2008).

3.3 Modelling techniques

The different modelling techniques developed for this research have been implemented into a multitude of software environments such as R, Statistica,

SAGA, ArcGIS and Matlab; enabling the computation of several mathematical algorithms.

3.3.1 Digital Soil Mapping

The environmental covariates available for DSM in Ireland are presented in Table 3.3, divided according to their SCORPAN characteristics and with a description of their spatial resolution or scale, depending on the type of information.

Table 3.3 - SCORPAN covariates of relevance to DSM in Ireland.

Covariates	Data	Scale or Resolution
Soil	General soil map of Ireland	1:575,000
	AFT county soil maps *	1:126,720
	AFT survey field maps *	1:10,560
Climate	ICARUS baseline climatology	1,000 m
	Met Eireann stations network	-
Organisms	CORINE 1990	100 m (25 ha)
	CORINE 2000	100 m (25 ha)
	CORINE 2006	100 m (25 ha)
Relief	EPA/Teagasc DEM	20 m
	OSI LIDAR *	1 m
	ASTER DEM	30 m
	GTOPO30 DEM	1,000 m
Parent material	GSI bedrock geology	1:100,000
	Quaternary map *	1:25,000
Age	Midlandian glaciation limits	~ 1:1,00,000
N - spatial position	Landform mapping (SOTER)	1:250,000

* partial coverage, not available for the whole country

In this research, two datasets were extensively used: the AFT survey field maps and the EPA/Teagasc DEM. The AFT survey field maps which show the delineation of soil series, made in the field by the surveyors, was the most

detailed information available at landscape scale on the variability of Irish soils. The EPA/Teagasc DEM was the most detailed elevation dataset at national coverage in Ireland and was the elevation information chosen for the overall ISIS project. This DEM was used to generate terrain attributes for the DSM scale analysis presented in this study.

3.3.2 Data Mining

With the increase of large and complex datasets new methods and techniques have been developed to make sense of data and extract meaningful patterns. Statistics, mathematics and information theory have contributed to the development of a multitude of different approaches for mining data in the detection of hidden structure and useful information to be applied in advancing research. The ones that have receive the most attention in DSM are: Random Forest (Grimm *et al.*, 2008; Behrens *et al.*, 2010b); Artificial Neural Networks (Behrens *et al.*, 2005; Lamorski *et al.*, 2008; Scull *et al.*, 2003; Zhu, 2000), Bayesian belief network (Marchant and Lark, 2007b; Reinds *et al.*, 2008) and Classification And Regression Trees (Moran and Bui, 2002; Scull *et al.*, 2005; Grinand *et al.*, 2008). In this research the focus would be on two well established data mining classifiers artificial neural network (NN) and random forest (RF). As the two are based on different assumptions and treat the data in distinctive ways, this should allow to quantify the degree of uncertainty associated with the DSM model allowing the experiments to focus on the comparison between different resolutions.

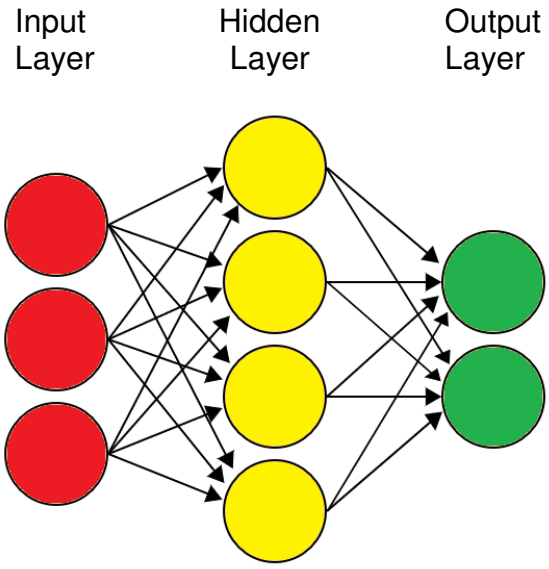


Figure 3.10 - An example of an artificial neural network.

Neural Networks are organised in a series of layers as presented in Figure 3.10, each input data is processed by the first layer, projected onto an intermediate hidden layer where each node is the weighted sum of the values in the input layer and finally re-projected to the output layer. The network is trained with the data entering in the input layer affecting both hidden and output layer so that the output layer starts to match the required output class. After the network has been trained each new input data will go through the network and be classified in one of the wanted categorises or if the network is uncertain to an intermediate value in between the most similar groups (Gershenfeld, 1999). This data mining technique has been successfully used in DSM to map soil properties (Ramadam *et al.*, 2005), to predict soil organic carbon across different land uses (Somaratne *et al.*, 2005), to model the spatial variation of soil loss from natural runoff (Licznar and Nearing, 2003) and by using terrain attributes derived from a DEM to map soil texture distribution (Zhao *et al.*, 2009).

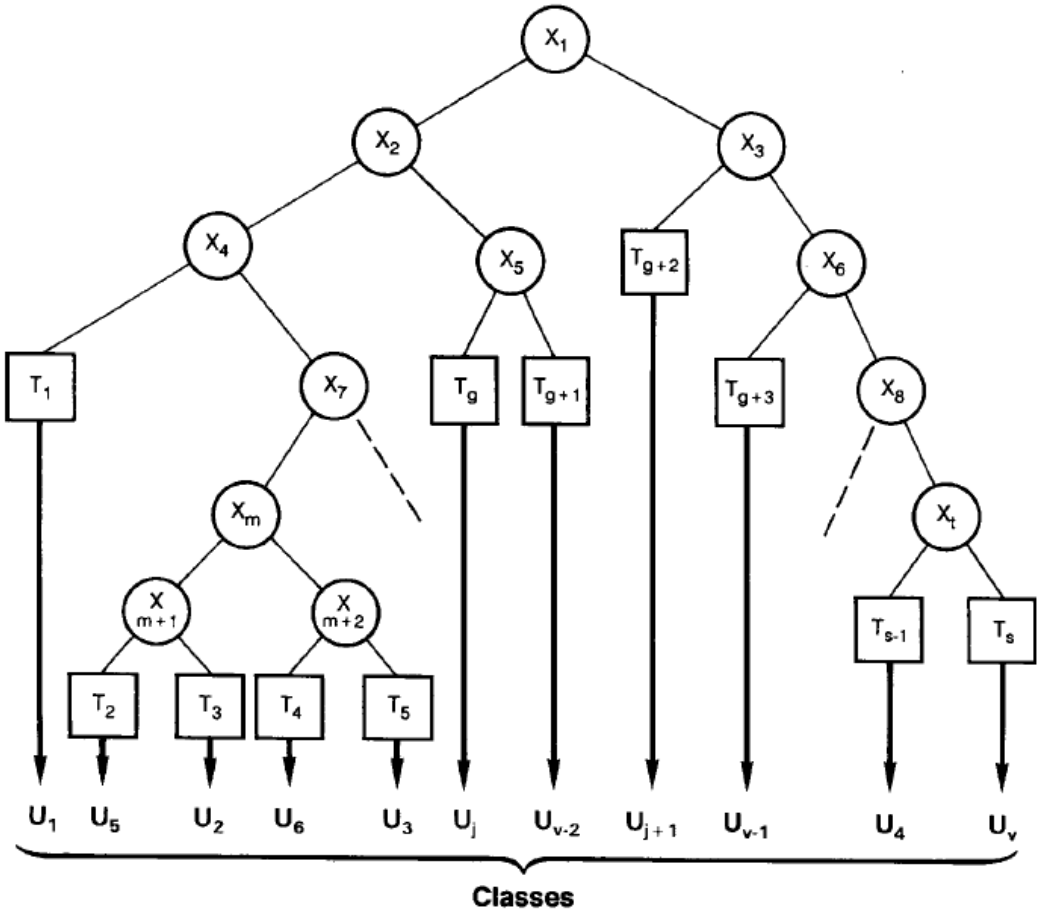


Figure 3.11 - An example of a classification tree used in DSM to predict soil map units (from Lagacherie and Holmes, 1997).

The Random Forest approach, based on the ramifications of a decision tree (Figure 3.11), is commonly used in DSM to produce rules for predicting soil map units at unvisited locations from a set of covariates available for the entire study area (Lagacherie and Holmes, 1997). The input data enters the tree and is then randomly divided into a smaller set belonging to a lower category based on the predictor variables which provide the best split, internally assessed with a function. It will then be divided up again into an even smaller set and so on till it reaches the final level previously set by the maximum number of trees. Once the training of the forest of trees is completed each data input will move down the each level being separated into smaller and smaller sets, finally reaching the wanted categorization (Breiman, 2001).

3.3.3 Empirical approaches

Empirical approaches to identify scale according to rough and quick practical guidelines have been developed and used in the wider environmental sciences from cartography, geomorphometry, hydrology and remote sensing to GIS and computer science. These empirical rules used in the selection of an appropriate pixel size are based on either the intrinsic properties of the data or the characteristics of the final resulting map. They are related to cartographic concepts such as size of delineation (Rossiter, 2003) or objects representations (Hengl, 2006), soil surveying conventions as sampling support density (Avery, 1987; McBratney, 1998), geomorphometric generalizations of terrain complexity (Hengl, 2006), hydrological characteristics of river networks and catchments (Sharma *et al*., 2011) and general statistics and information theory such as Shannon information content and Kolmogorov complexity (Hengl *et al*., 2013).

Although some of these empirical approaches are associated or inspired by scientific concepts, they are based on experience and practical knowledge rather than derived from scientific theory. For example, to calculate the optimum pixel size, a rule of thumb currently applied in all raster operations by the most popular GIS software and industry standard ESRI ArcGIS is to divide by 250 the width or height (whichever is shorter) of the extent of the feature dataset. A selection from the literature of the most promising empirical approaches, some of which have been previously suggested to the DSM community (Hengl, 2006) were tested and compared with the experimental results of the experimental methodology.

3.3.4 Wavelet

Soil changes with space and its variability depends on the interaction between soil forming processes that operate on different spatial and temporal scales. According to Si (2007) these variations can be divided into two broad categories based on their frequency of change:

- High frequency that vary repeatedly in space or time (cartographically equivalent of a small scale process)

- Low frequency that vary slowly in space or time (cartographically equivalent of a large scale process)

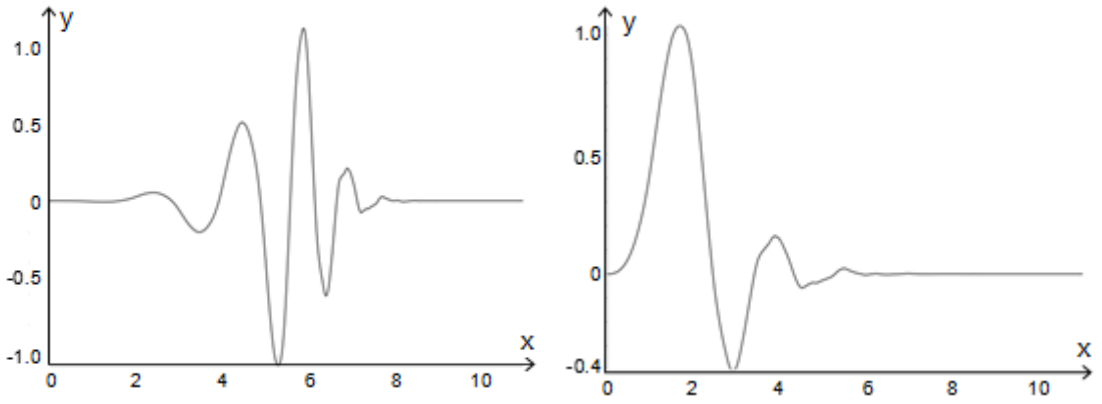
In landscapes where landforms are repetitive such as hummocky, rolling or undulating terrains, the continuous variation of soil results in data series with repetitive cycles of highs and lows (Pennock *et al.,* 2008). The examination of such data requires using techniques in which the total variance is partitioned by frequency. These techniques are referred to as spectral analysis (McBratney *et al.*, 2002). Spectral analysis has always interested soil scientist, geographers and geomorphologists (Pike and Rozema, 1975) and has recently benefitted from the work of geophysicist Jean Morlet, physicist Alex Grossman and mathematician Stephane Mallat in the late '80s that developed the wavelet theory (Grossman and Morlet, 1984; Mallat, 1989). This technique, designed to separate data at different scales from noise that does not have any correlation, is becoming the analysis of choice for scale analysis trying to explain soil information variation (Lark and Webster, 1999). This mathematical model is capable of analysing processes in terms of trends and localised features by partitioning the variation into scales at precise locations (Lark and Webster, 2001). Soil science has made use of wavelet in the analysis of 1D spatio-temporal transects (Lark *et al.*, 2004; Si and Farrell, 2004; Biswas *et al.*, 2008) and 2D for terrain attributes (Lark and Webster, 2004) and DEM decomposition (Mendonca-Santos *et al.*, 2007, Biswas *et al.*, 2013). The wavelet transform method in essence consists of the decomposition of a signal into a hierarchical set of approximations and details for each scale. It allows quantifying signal changes from one scale to another through dilations and translations of a single function called mother wavelet. The wavelet transform is based on the convolution of the following function (Biswas and Si, 2011):

$$W(s,\tau) = \int_{-\infty}^{\infty} y(x)\psi_{s,\tau}\,(\mathrm{x})\,\mathrm{dx} \tag{3.1}$$

where y is the measured parameter, x is the spatial or temporal spacing along the transect, s is the dilation factor, τ is the spatial or temporal translation of the function and ϕ is the mother wavelet:

$$\psi_{s,\tau}\,(x) = \frac{1}{\sqrt{s}}\,\psi_{s,\tau}\left(\frac{x-\tau}{s}\right) \tag{3.2}$$

There are many different mother wavelets with unique shapes and characteristics to better fit the signals analysed: crude wavelets (gaussian, morlet, Mexican hat); infinitely regular wavelets (Meyer); orthogonal and compactly supported wavelets (Daubechies, symlets, coiflets); biorthogonal and compactly supported wavelet (B-splines biorthogonal) and complex wavelets (complex Gaussian, complex morlet, complex Shannon, complex frequency B-spline) (Misiti *et al.*, 2012).



Wavelet function
(detail part)

Scaling function
(approximation part)

Figure 3.12 - Wavelet and scaling function of the db6 Daubechies wavelet (y = measured parameter and x = spatial or temporal spacing along the transect).

In this research the wavelet analysis presented in Chapter 6 has been performed using the Daubechies wavelet with six vanishing moments (Daubechies, 1990). The Daubechies (Figure 3.12) is a family of asymmetric, biorthogonal

wavelets characterized by a high number of vanishing moments for given support width and widely used in solving a broad range of problems in the geosciences (Labat, 2005).

3.3.5 Geostatistics

Geostatistics, initially empirically developed to predict probability distributions of recoverable ore reserves by mining engineers (Matheron, 1965), has since developed into an established branch of statistics with applications in all environmental disciplines requiring the analysis of spatial data. The main principle underlying geostatistics is the theory of regionalised variables (Goovaerts, 1997) which acknowledges that spatial attributes observed on a specific point location are a single realisation of a regional process expressed as:

$$Z(x) = \sum_{k=0}^{K} a_k \ f_k(x) + \varepsilon(x) \tag{3.3}$$

where $Z(x)$ is the variable under consideration as a function of spatial location, f_k is a known function of x related to its spatial location, a_k is a coefficient related to the specific situation and $\varepsilon(x)$ is the random residual from the trend.

Matheron understood that the variance of the random component depended only on the relative distance between observations (h) and not on where the observations were made. This was conceptualised as the intrinsic stationarity hypothesis (Webster and Oliver, 2007) from which the semivariance is the direct expression:

$$\gamma(h) = \frac{1}{2} \frac{1}{n(h)} \sum_{i=1}^{n(h)} (z(x_i) - z(x_i + h))^2 \tag{3.4}$$

where the semivariance γ (h) is calculated as half the variance of the increments with n(h) the number of paired data at a distance h (lag distance). From this

equation a variogram cloud can be created showing the spatial correlation of all the possible distances between paired points or averaging these values at each distance h.

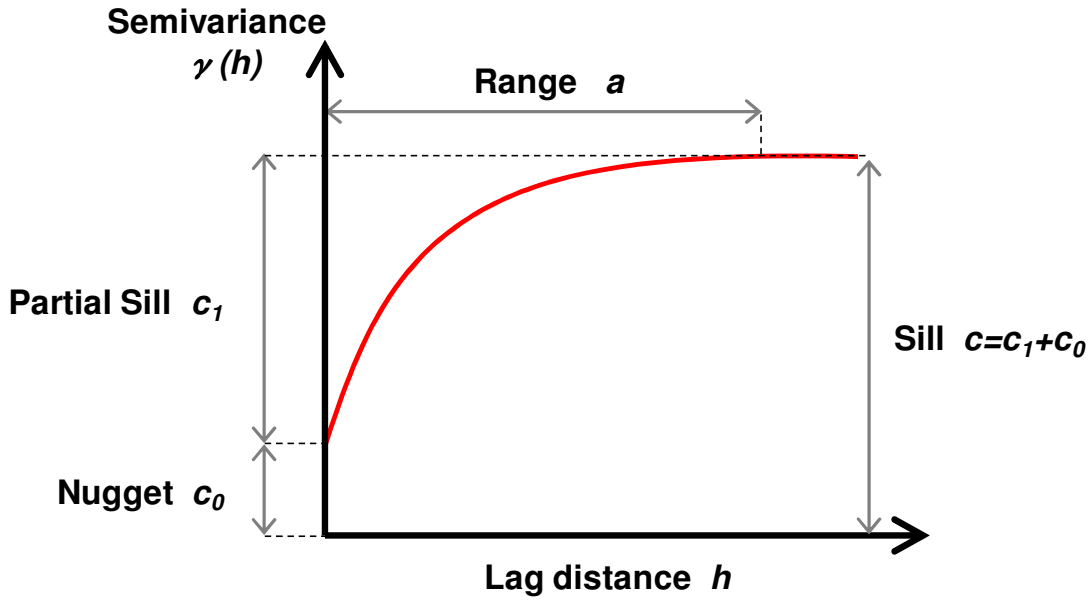


Figure 3.13 - Characteristic variogram model.

The resulting plot (Figure 3.13) of the fitted model at increasing lag distances is called semi-variogram or just variogram. The essential components of a variogram are:

- Nugget (c_0) the intercept on the variance axis of lag distance values equal to 0. It represents variability at distances smaller than the sample spacing and measurement errors;
- Sill ($c=c_0+c_1$) the maximum semivariance in the data at which the fitted model levels of, calculated adding the nugget to the partial sill (c_1);
- Range (a) the lag distance at which the variogram reaches the maximum semivariance, after this value points are not spatially autocorrelated anymore.

Mathematical models can be fitted to the experimental variogram (Exponential, Gaussian, Spherical, Power) for visualization of the spatial variation. This is a key

step in geostatistics as some authors still prefer to manually fit a model based on visual assessment and expert knowledge. The balance between accurately described correlation, especially a small lag distances, and the risk of overfitting needs to be found as the model has to correctly represent spatial variability (Webster, 2000). Variograms in this study were computed using the Residual Maximum Likelihood (REML) method (Patterson and Thomson, 1971; Marchant and Lark, 2007a) as it gives a better representation of the underlying variation in comparison to the classical method of moments especially when a regular grid sampling scheme is used (Lark and Cullis, 2004).

In this research variograms were used to describe spatial patterns and structures of elevation for the three investigated areas. In Chapter 7 a moving window variogram approach was used to classify spatial variation and develop a multiscale methodology for DSM analysis. All the variograms were created with the R software (R Development Core Team, 2011) using the gstat package (Pebesma and Wesseling, 1998).

4 EXPERIMENTAL METHODOLOGY

4.1 Introduction

The most suitable resolution of DEM to apply in soil mapping depends on the scale of the processes controlling pedogenesis and this is landscape dependent. As a consequence there is a real need for a more landscape-scale pedology (Pennock and Veldkamp, 2006). The choice of scale frames the analysis and shapes the end result of DSM analysis indicating that a better understanding and quantitative knowledge of scale will help to improve soil predictions.

The relationship between soil taxonomic units and landscape attributes has been confirmed as a central concept in soil science (Hudson, 1992). Terrain attributes are the most widely used predictors in DSM because of their primary role in soil formation and the broad availability of DEMs (Behrens *et al*., 2010b). DEMs are representations of the topographic surface of the Earth, and they are a widespread data source for terrain analysis and other spatial applications. Terrain analysis provides a number of high-resolution environmental information quantitatively derived from DEM including slope gradient, aspect, curvature, etc. and these topographic features are the core for a wide range of landscape-scale environmental models (Gallant and Hutchinson, 1997).

Thompson *et al*. (2001) have shown that higher-resolution DEM may not be necessary for generating useful soil-landscape models. Another concern is that most applications use algorithms running in small windows (usually 3 × 3 roving window) to perform terrain analysis, thus fixing the scale of resulting layers to the spatial resolution of the available DEM. This is expected to provoke mismatches between scale domains of terrain information and the environmental variable of interest (Dragut *et al*., 2009 and Smith *et al*., 2006).

The purpose of this chapter is to test and evaluate the role of spatial scale and its impact on generating soil class predictions by experimentally testing the interaction between pixel resolution and window size with two commonly used

data mining classifiers in DSM: Artificial Neural Network and Random Forest. The large number of scale combinations tested will allow an optimal scale to be established, providing the benchmark for comparing the results of the following chapters.

4.2 Materials and Methods

In order to address the problem previously described three different areas in terms of geomorphology and soil type were selected in Ireland. A detailed soil map obtained from legacy surveying was used as a dependent variable to train and test two separate DSM models. The soil information was classified at soil series level. The DEMs of the three areas were processed to obtain input terrain derivatives. A set of window and pixel size combinations was then run to obtain values for the model inputs, develop the DSM models on these values (and the soil map) and compare predictions to the mapped soil information. A measure of performance for each window by pixel size combination was obtained which was subsequently analysed using analysis of variance. As such the model behaviour can be experimentally described as a function of window and pixel size. Finally, the model performance by soil type was considered for the best and worst scale combinations tested, achieving the highest and lowest classification accuracy respectively.

4.2.1 DEM

The DEM was used to derive eleven terrain attributes: slope gradient, aspect, curvature, plan curvature, profile curvature, slope height, valley depth, normalized height, standardized height, mid-slope position and convergence index (Table 4.1) using SAGA GIS Terrain Analysis Morphometry library (Bock *et al.*, 2008). They are indicative of soil-landscape relationships controlling the spatial distribution of physical, chemical and biological soil properties and the overall energy and water balances. A large number of terrain attributes is

analysed which have been chosen for their ability to describe soil classes as presented by Behrens *et al.* (2010b) and Florinsky (2011), including:

- local morphometry attributes (slope gradient, aspect, curvature, plan curvature, profile curvature and convergence index);
- relative height and slope position attributes (slope height, valley depth, normalized height, standardized height and mid-slope position).

The selected terrain attributes have been recognised to affect pedogenesis (Florinsky, 2011) by governing the microclimate, the thermal balance, the water cycle, erosion processes, intra soil transport of nutrients and distribution of vegetation. For example, slope steepness drives erosion rate; slope position alters moist content; and aspect affects sun exposure and thermal regime. Topography is the result of both internal (geology) and external processes all operating at different scales, many of which are related to soil.

To calculate local morphometric terrain attributes fundamental for all the other variables, the second-order finite difference algorithm of Zevenbergen and Thorne (1987) was used. The procedure for optimizing DSM prediction power of soil data with environmental covariates will focus on the selection of an optimal scale, represented as the interaction between window and pixel sizes, correlated to the pedogenetic processes active at the landscape scale. The statistical relationship between soil taxonomic units and terrain derivatives will be used to select the scale at which terrain parameters correlate better with soil data and predict the most accurate soil information.

Table 4.1 - Investigated terrain attributes.

Terrain Attribute	Unit	Description
Local Morphometry		
Slope Gradient	[rad]	The angle of inclination of the topographic surface between the tangent and the horizontal planes
Aspect	[rad]	The clockwise angle from north of the projection to the horizontal plane of the topographic surface
Curvature	[m^{-1}]	The average of two orthogonal normal sections
Plan curvature	[m^{-1}]	The rate of change of the horizontal curvature
Profile curvature	[m^{-1}]	The rate of change of the vertical curvature
Convergence index	[%]	The index of convergence/divergence regarding to overland flow
Relative height and slope positions		
Slope height	[m]	The relative height difference to the immediate adjacent crest lines
Valley depth	[m]	The relative height difference to the immediate adjacent channel lines
Normalized height	[ ]	The height values are normalized to a range from 0 to 1
Standardized height	[ ]	The height values are standardized to have a mean of 0 and standard deviation of 1
Mid-slope position	[ ]	A classification of the slope position in both valley and crest directions

In order to fully analyse and incorporate the effects of scale on terrain attributes used in DSM, two techniques were applied to the three study areas' DEMs: smoothing (window size alteration) and resampling (pixel size alterations). Previous studies have shown the influence of pixel size alterations in computing

terrain attributes used in DSM analysis (Smith *et al.*, 2006), the importance of neighbourhood size (Zhu *et al.*, 2008) and a combination of both pixel and neighbourhood alterations (Roecker *et al.*, 2008; Behrens *et al.*, 2010a). In this study a standard 20 m resolution DEM was used to investigate the scale dependency of terrain attributes when converted to coarser resolutions. Thus the original DEM represents no smoothing (1 x 1 window) and a pixel size of 20 m. A series of DEMs were created from the original DEM. Firstly, the DEM was smoothed by applying different window sizes at 3 x 3, 5 x 5, 7 x 7, 9 x 9, 11 x 11, 13 x 13, 15 x 15, 17 x 17, 19 x 19 and 21 x 21. The resulting smoothed DEMs were re-sampled at 30, 40, 50, 60, 80, 100, 120, 140, 170, 200, 230, 260 m pixel sizes using bilinear interpolation. This resulted in 143 distinct datasets: the original DEM; 10 smoothed but not re-sampled DEMs; 12 re-sampled and not smoothed DEMs; and 120 smoothed and re-sampled DEMs. Figure 4.1 shows the resulting effects of different window sizes and re-sampled pixel sizes on DEM resolution.

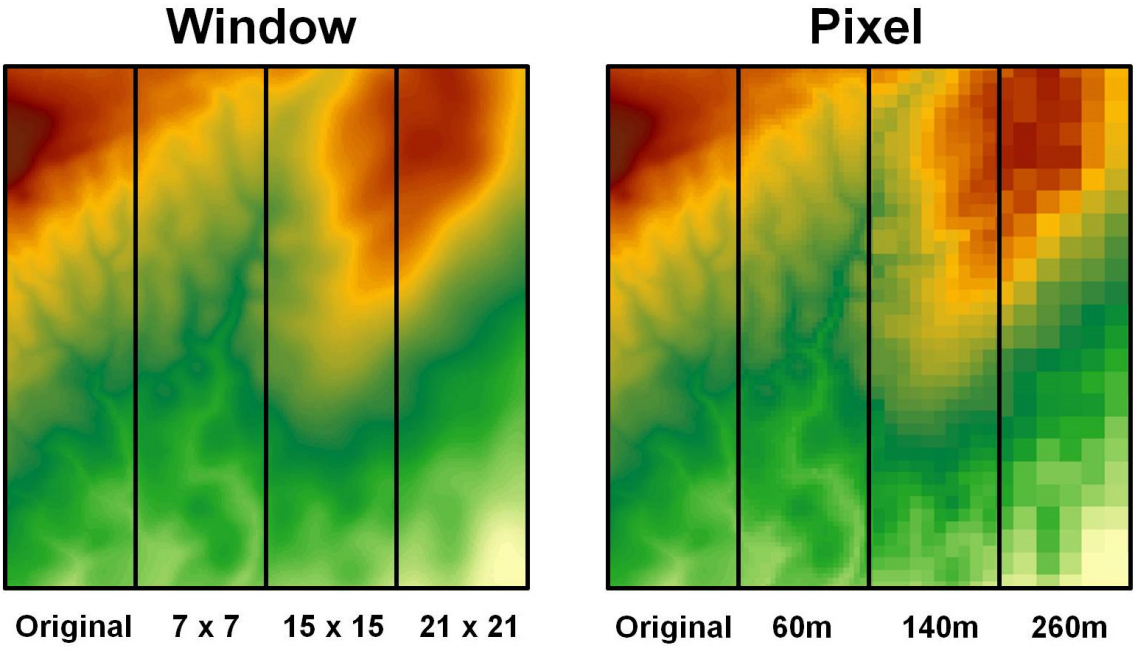


Figure 4.1 - Original DEM of Ireland at 20 m re-sampled at 50, 100 and 500 m pixel sizes (right) and with a 3 x 3, 5 x 5 and 21 x 21 window sizes (left).

Terrain attributes were derived for each of the 143 datasets and 4 points per km^2 were randomly extracted and fed into data mining inference systems to find an optimal scale; starting from the original DEM and incrementally increasing window and pixel sizes. The density of four points per square kilometre was chosen to be representative of the likely observational density during the field programme to produce the county soil maps.

A quantitative comparison has been made by calculating the validation performance of data miner classifiers. This will enable to better understand the role of scale and experimentally select an optimal scale to use as a benchmark in the second stage of the research.

4.2.2 DSM model development

The focus and final output of DSM are soil properties derived by a spatial inference system (Carre *et al.*, 2007). Dealing with fine resolution data and large extents creates uncertainties. There are three main areas of concern affecting the accuracy of these predictions: soil reference information, environmental covariates and inference systems. The soil information (6 inch maps) is assumed to be correct and it provides the training data for the models. The mapping was based on field observations by the soil surveyor but is subject to the interpretation of the individual. The environmental covariates, in this case the terrain attributes derived by the DEM, are under scrutiny in this first stage of the research. The inference systems, including data mining classifiers, consist of analytical processes designed to explore data in search of consistent patterns and systematic relationships between variables. The resulting relationships are used for prediction of an optimal scale through the comparison of individual validation performances. In order to minimise the uncertainties associated with the inference systems two distinct techniques that operate on different statistical assumptions were used in this research: Random Forest and Artificial Neural Network. These are regularly applied in DSM (McBratney *et al.*, 2003) and were chosen for their ability to handle datasets with many predictors, to deal with soil

predictors non-linear relationships and also to be robust to noise, outliers and overfitting (Viscarra Rossel and Behrens, 2010b).

Random Forest (RF) is a collection of CART-like trees following specific rules for classification or regression (Breiman, 2001). The trees are created using a different bootstrap sample of the data and each node is subsequently split using the best among a subset of predictors randomly chosen at that node to prevent overfitting. The number of predictors to be selected in a subset is calculated as the logarithm in base two of the total number of predictors plus one. This strategy allows the reduction of the number of factors required to just two: 1) the number of variables tested at each node and 2) the overall number of trees in the forest that was set to 100.

Artificial Neural Network or Neural Network (NN) is an interconnected group of artificial neurons processing information using a connectionist approach to computation. A dataset is used to train the neural network which discovers an approximate relationship, between a series of covariates and the response variable, by iteratively adjusting its parameters (Gershenfeld, 1999). In essence, a series of subsets with similar node arrangement creates an input layer, an output layer and in between a hidden layer, which weights the data to extract the significant information on its relationships. As presented by Nisbet *et al.* (2009), an automated network search with two strategies was used to develop the NN models: the most widely used Multilayer Perceptrons (MLP) with a maximum number of 15 hidden layers, and Radial Basis Function (RBF), a simpler network with faster learning algorithms set with a maximum number of 30 hidden networks. To prevent overfitting and increase performance, a weight decay of 0.001 was adopted and the data were separated into training (70%), testing (15%) and validation (15%) subsets allowing testing of the hidden layers.

Principal component analysis (PCA) was employed to detect structure in the relationships between the eleven terrain attributes and to assess their correlations and degree of redundancy.

Analysis of variance or ANOVA allows simultaneous comparison among means of several groups by partitioning the observed variance in a particular variable into components of different sources of variation. This was implemented using Statistica software (StatSoft, 2010) for the assessment of window, pixel and the interaction of the two showing their specific significance in optimal scale selection.

4.3 Results

The results are presented in Figures 4.2 - 4.6 and Table 4.2. Figure 4.2 illustrates how the most important six terrain attributes used in classification change according to scale. Figure 4.3 shows the PCA analysis of the terrain attributes used in the DSM models. Figures 4.4-4.5 summarise the validation performance results of both RF and NN across all the scale combinations tested. Table 4.2 contains the ANOVA analysis performed to discriminate between the effect of window, pixel or the interaction of the two and Figure 4.6 shows the classification accuracy by soil series for the best and worst scale combinations.

4.3.1 Terrain attributes

The effects of altering the spatial resolution of the DEM and window size are shown for the six most important classification attributes (slope gradient, aspect, curvature, slope height, mid-slope position and convergence index) by their change in dispersion calculated by the standard deviation (Figure 4.2). The random forest algorithm estimates the importance of a variable, by looking at how much prediction error increases when data for that variable are permuted while all others are left unchanged (Breiman, 2001).

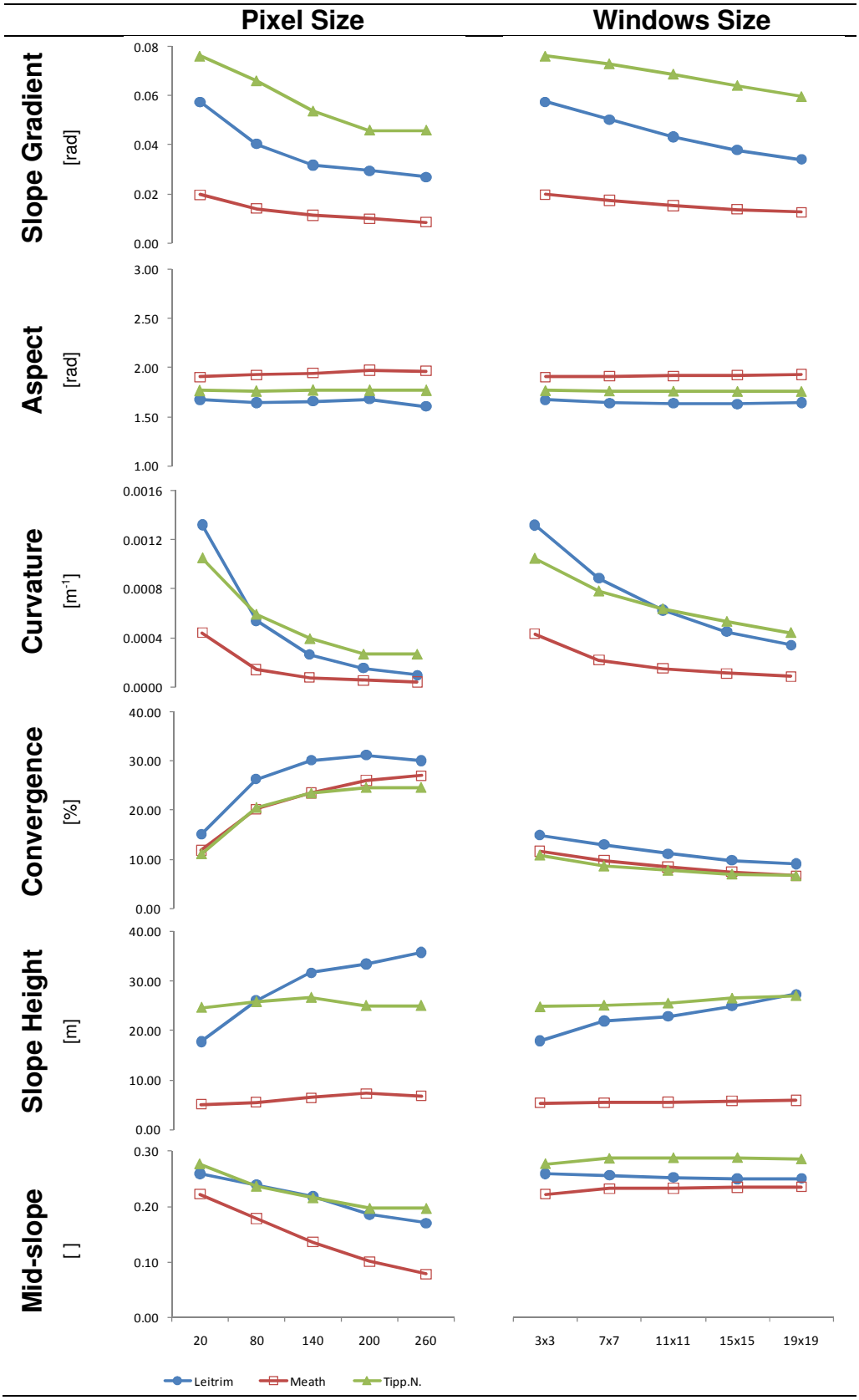


Figure 4.2 - Standard deviation of the six most important terrain attributes.

Statistical variability of slope gradient is reduced with the increase of both window and pixel size. In Leitrim and Tipperary North values decrease respectively from 0.058 to 0.026 and 0.076 to 0.045 with coarsening of spatial resolution and from 0.058 to 0.034 and 0.076 to 0.059 with the increasing of window size. Meath followed the same pattern of reduction but with a lower intensity decrease from 0.014 to 0.008 with pixel variation and from 0.020 to 0.014 with window variation. A similar declining pattern observed for slope gradient is followed by curvature. As expected the standard deviation of aspect does not change with scale since the full range of orientations can be anticipated at any scale. Convergence Index in contrast shows an unexpected divergent pattern: a sharp increase of standard deviation for all the three areas with the coarsening of pixel size and a low intensity decrease with the enlargement of window size. Slope height presents a clear difference between the behaviour of three tested areas with Leitrim characterized by the presence of the drumlins showing an increase from 17.78 to 35.85 with varying pixel size and from 17.78 to 27.22 with changing window size, meanwhile Meath and Tipperary North appear scale invariant with no significant change of standard deviation despite having very different values, respectively 5 and 25. Finally, mid-slope position follows the same sharp decrease observed with slope gradient and curvature at the increase of pixel size but unexpectedly not with window size in which results appear scale invariant.

As described by Wilson and Gallant (2000) standard deviation of topographic features is a measure of variability associated with landscape roughness. The terrain parameters tested demonstrate that statistical dispersion changes with resolution and window size alterations. This might have an effect on their predictive power in DSM modelling, for example by removing redundant information (low values of standard deviation), so improving classification accuracy.

4.3.2 Principal Component Analysis

Principal component analysis, by converting the set of eleven terrain attributes into a set of linearly uncorrelated values, has allowed evaluation of their level of

redundancy. In this case, redundancy means that some of the variables are correlated with one another, possibly because they measure a similar characteristic or their change follows a similar trend. Figure 4.3 shows the projection of the original eleven terrain attributes onto the two components, with the principal component on the horizontal axis and the second component on the vertical axis.

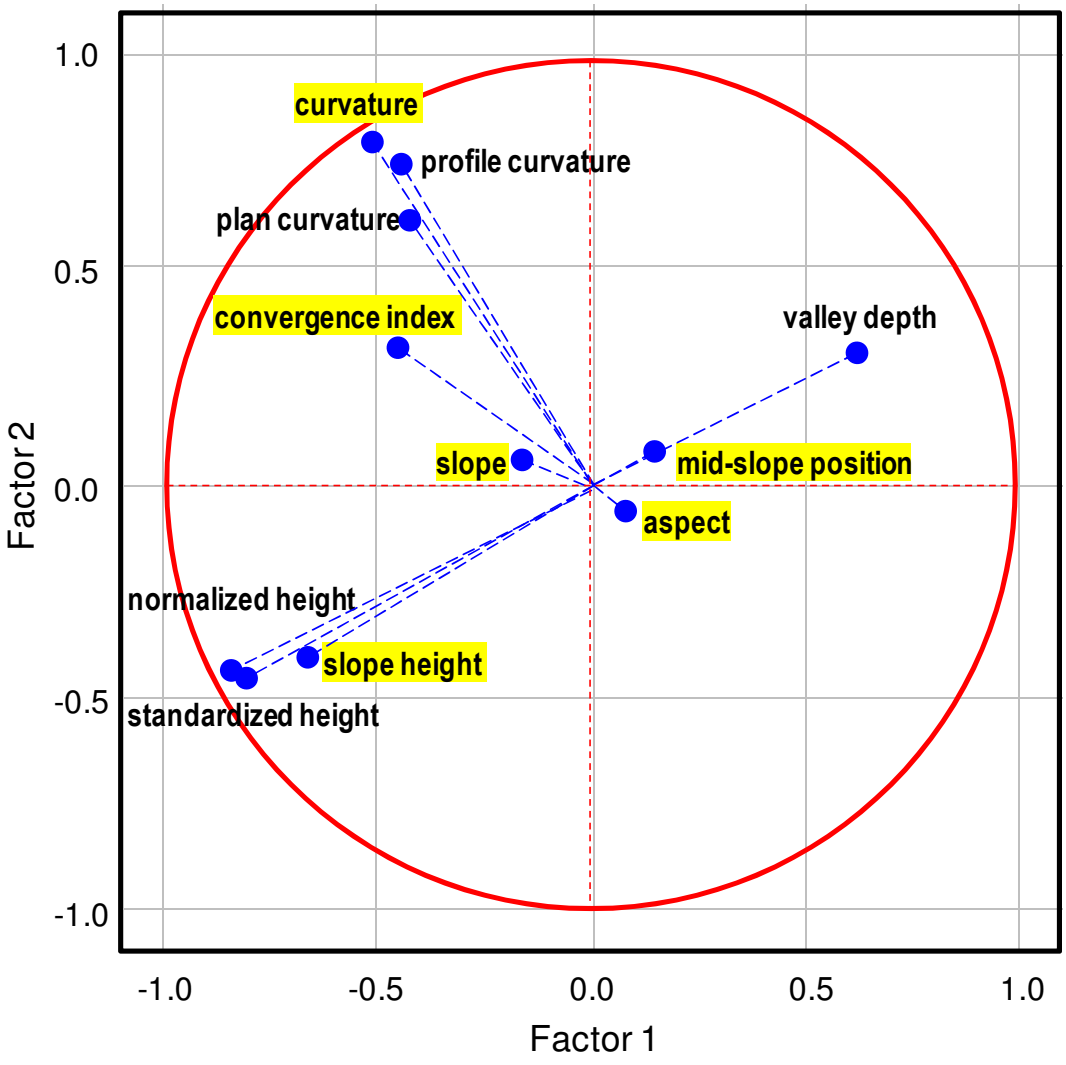


Figure 4.3 - PCA of the 11 terrain attributes used in the DSM models (highlighted are the six most important terrain attributes).

Some terrain attributes appear to cluster together as curvature, profile curvature and plan curvature indicating a higher correlation and high degree of redundancy. In this case this is due to the fact that curvature includes both maximum slope

and perpendicular directions. The same happens with slope height, normalized height and standardized height with similar negative correlations on both components. All the other terrain attributes have different principal components with mid-slope position and valley depth in the first quadrant (positive principle component and positive second component), aspect in the second quadrant (positive principle component and negative second component), slope gradient and convergence index in the fourth quadrant (negative principle component and positive second component). The six most important terrain attributes (slope gradient, aspect, curvature, slope height, mid-slope position and convergence index) according to the two DSM models account for the most of the variance in the observed terrain attributes.

4.3.3 DSM models

Results of the scale analysis are shown in Figure 4.4, where the validation performance of data miner classifiers is shown with window and pixel size effects. This has resulted in 143 possible combinations ranging from the original 20 m pixel size in a 1 x 1 window to the 260 m pixel size averaged over a 21 x 21 window. It should be noted that for the RF the validation performance has been calculated as (1 - misclassification rate) * 100 in order to provide a direct comparison with the value obtained for NN, where the validation performance of the best 5 networks was averaged to obtain a value for each scale combination of pixel and window sizes.

The three morphologically different study areas behave in unique ways. Leitrim characterized by fine resolution drumlins achieves optimum performance for the unsmoothed but re-sampled DEMs at 30 m (58.5% for NN and 58.0% for RF) gradually decreasing towards 260 m (51.6% for NN and 50.6% for RF). For the smoothed but not re-sampled DEMs the optimum performance is reached in a 3 x 3 window (58.0% for NN and 57.5% for RF) gradually decreasing towards 21 x 21 (48.2% for NN and 50.9% for RF). Of particular interest for this area is the interaction between window and pixel sizes because at the increase of the latter it is possible to note an alteration of validation performance in coarser resolutions.

What seems peculiar is the quick shift recognizable at 5 x 5 in which the validation performance trend line becomes completely flat (averaging at about 54% for both NN and RF) and preserves this trend till 11 x 11. Then the trend completely reverses from a decrease of validation performance towards coarse resolutions to an increase. To summarise this Figure 4.4a and 4.4b show the validation performance against the window and pixel sizes clearly evidencing two areas achieving 60% of successful classification: fine resolutions with small window sizes and coarse resolutions with large window sizes.

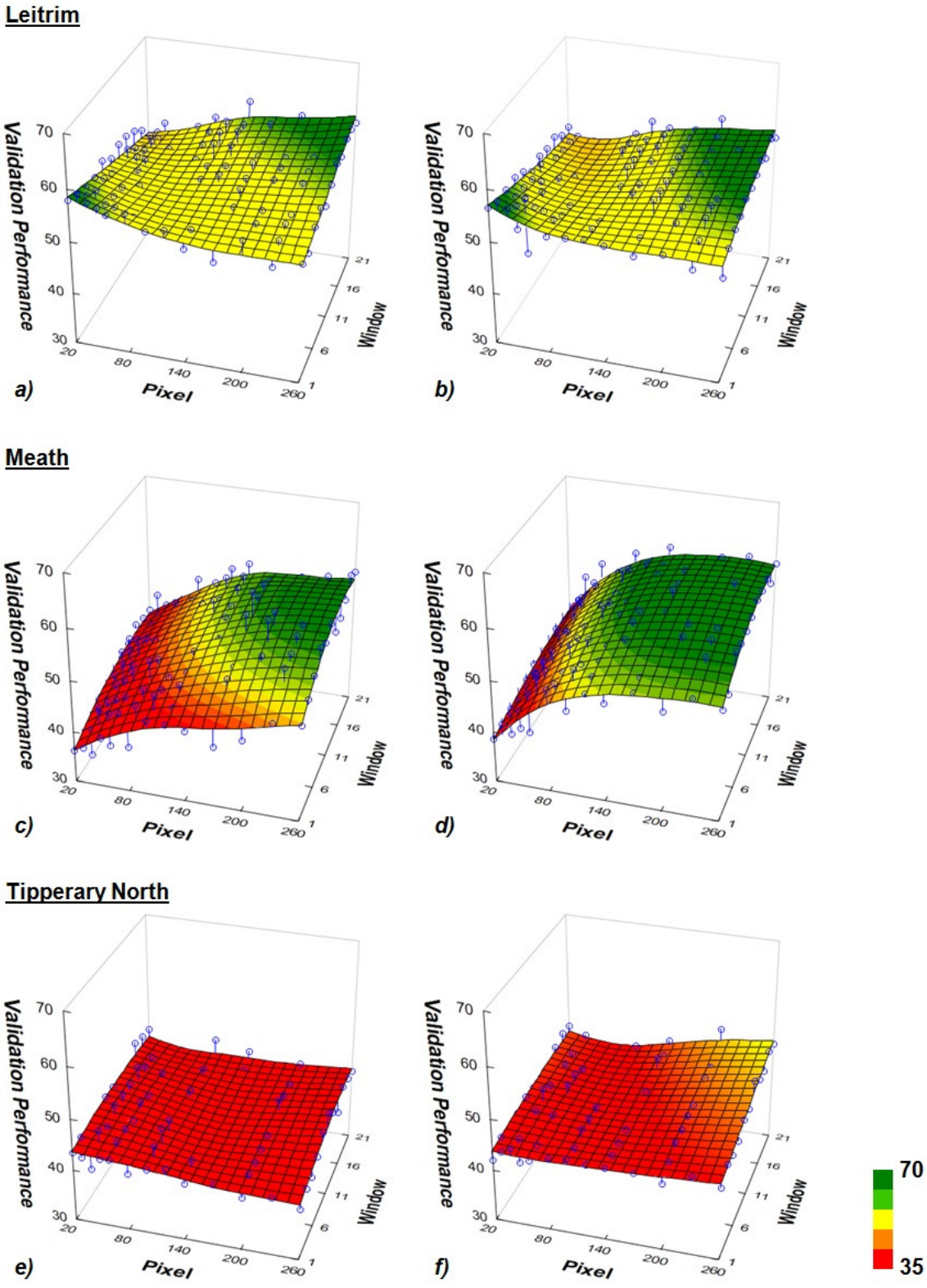


Figure 4.4 - 3D surface plots of validation performance against window and pixel sizes (NN on the left and RF on the right) of the three study areas.

On the contrary, Meath characterized by a flat surface shows a constant increase of performance towards coarser resolutions (above 140 m) across all window sizes. It is clearly shown in Figure 4.4c and 4.4d that performance gradually increases with pixel size for both classifiers. Window dimension is almost unresponsive in the RF model and shows weak dependence in the NN classifiers. Tipperary North is equally divided between steep slopes and flat plains but seems almost indifferent to changes in pixel size and only marginally affected by alteration of window size (Figure 4.4e and 4.4d) with a magnitude of change in order of 2% from its distinctive 42.5% for NN and 45.0% for RF validation performances. The trend line is generally flat with the line oscillating between a slight increase for the Neural Network and a slight decrease for the Random Forest. Notably, the two models performed very poorly at all scales and window sizes tested, suggesting that none of the terrain parameters investigated were able to discriminate between the low and high relief components. To investigate this apparent scale independence, Tipperary North was split into two homogeneous subareas; one almost completely flat and the other with steep slopes. The analysis was repeated and the results improved significantly (Figure 4.5) confirming the two distinct behaviour already seen for Leitrim and Meath; the low relief improving validation at coarser resolutions and being limitedly affected by the change in window size and the high relief area preferring fine resolutions with the above seen alteration at large window sizes.

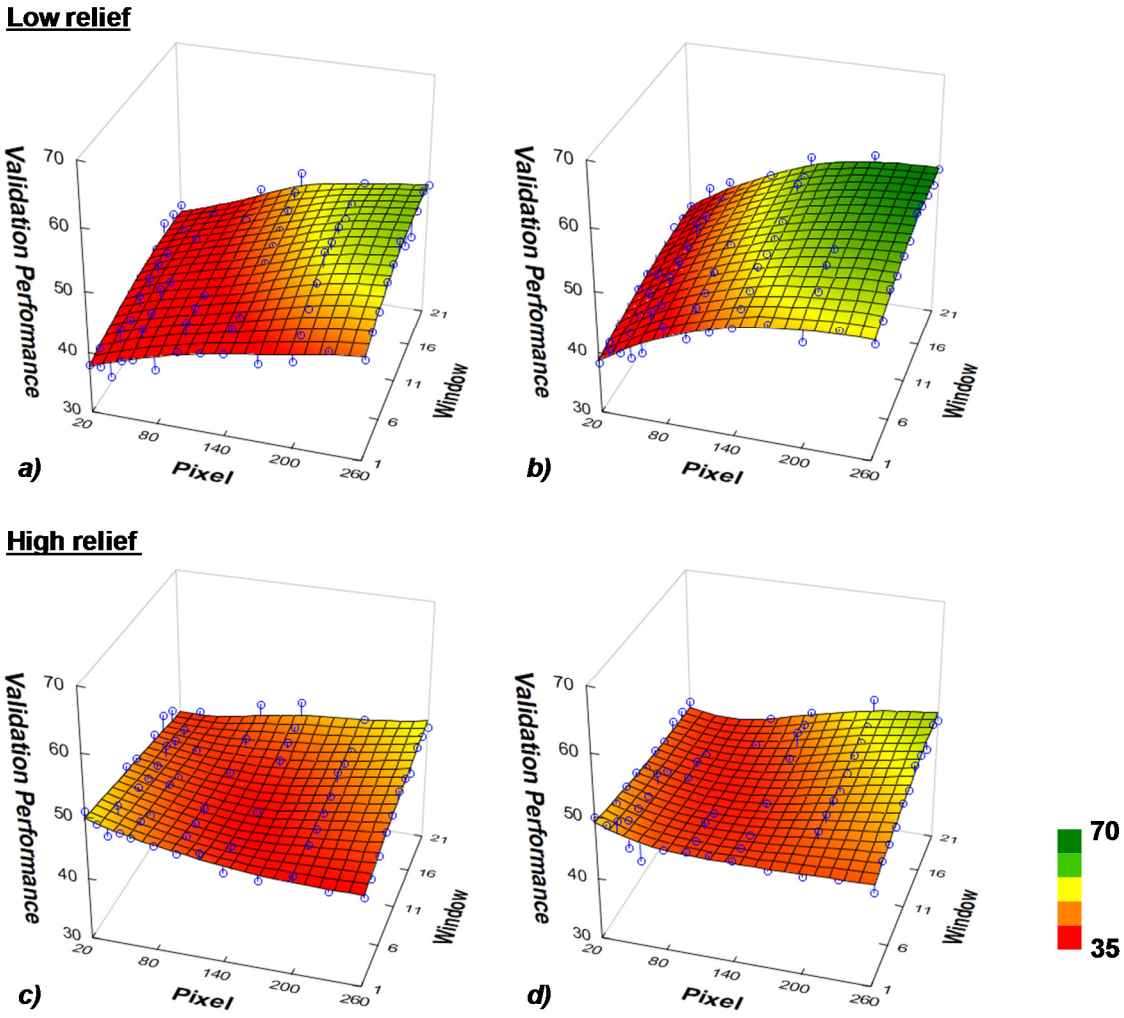


Figure 4.5 - 3D surface plots of validation performance against window and pixel sizes (NN on the left and RF on the right) of Tipperary North divided into low and high relief areas.

4.3.4 ANOVA

The effects of changing scale through the alteration of window and pixel sizes have been analysed with ANOVA and are presented in Table 4.2 showing respectively for Leitrim, Meath, Tipperary North the results of the analysis of variance. The null hypothesis is that there are no significant differences between the effects of pixel size, window size and the interaction of the two on the prediction of soil series using the two tested methodologies.

Table 4.2 - ANOVA results for the three study areas.

Leitrim

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
NN					
Pixel	1	1	71.08	26.06	<.0001*
Window	1	1	4.57	1.68	0.20
Pixel x Window	1	1	245.47	90.01	<.0001*
RF					
Pixel	1	1	162.24	53.60	<.0001*
Window	1	1	13.50	4.46	0.036*
Pixel x Window	1	1	159.10	52.56	<.0001*

Meath

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
NN					
Pixel	1	1	2905.86	487.99	<.0001*
Window	1	1	671.06	112.69	<.0001*
Pixel x Window	1	1	6.55	1.10	0.30
RF					
Pixel	1	1	3242.53	396.44	<.0001*
Window	1	1	401.87	49.13	<.0001*
Pixel x Window	1	1	4.42	0.54	0.46

Tipperary North

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
NN					
Pixel	1	1	28.76	8.87	0.0039*
Window	1	1	22.46	6.96	0.010*
Pixel x Window	1	1	7.01	2.16	0.15
RF					
Pixel	1	1	67.84	32.37	<.0001*
Window	1	1	63.046	30.087	<.0001*
Pixel x Window	1	1	12.43	5.93	0.017*

The simultaneous comparison among means of several groups, by partitioning the observed variance in a variable into components of different sources of

variation, has allowed the comparison of window size, pixel size and the interaction of the two showing their specific significance in optimal scale selection. These results show the significance of pixel size as a valuable factor in influencing performance validation for Leitrim and Meath. Notably, Leitrim shows also that the interaction between window and pixel sizes is highly significant confirming the results previously discussed. Meath shows window size as significant but not the interaction of the two factors. Tipperary North behaves erratically with the two data mining classifiers obtaining significantly different results. As previously indicated, the terrain attributes seem to operate independently of scale in Tipperary North, giving similar results across all spatial resolutions. Once the high relief and low relief components were separated and the ANOVA analysis was repeated, the results (not shown) confirmed the two distinct behaviour already seen for Leitrim and Meath with pixel and window size highly significant for the low relief area and pixel and the interaction with window size highly significant for the high relief component.

It is also worth mentioning that the two data mining techniques applied in this research generate almost identical patterns for Leitrim, Meath and Tipperary North divided in low and high relief areas, confirming reliability of the observed patterns.

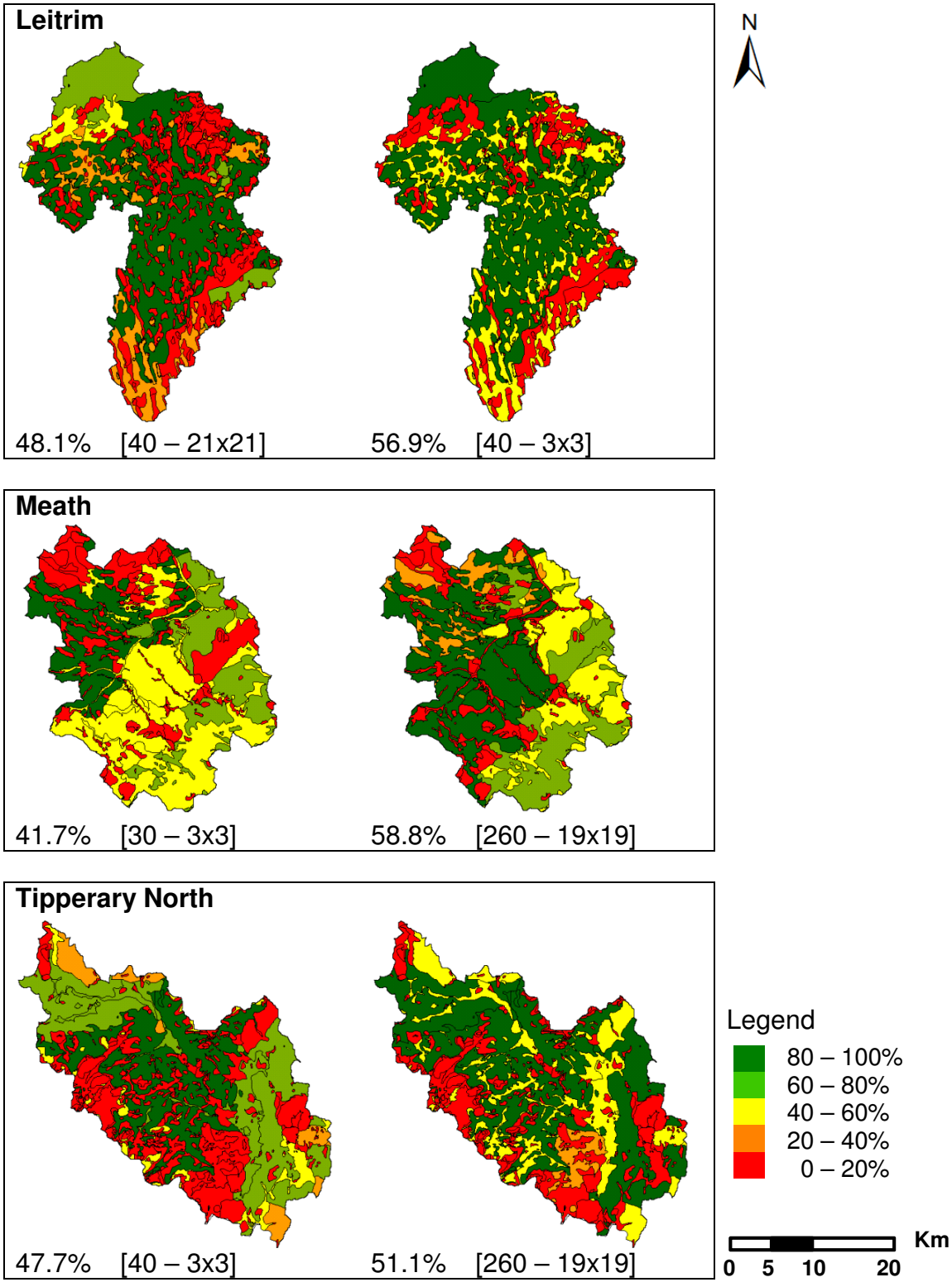


Figure 4.6 - Classification accuracy of samples aggregated by soil series. Best case (right) and worst case (left).

The proportion of soil classification accuracy of the best and worst scale combinations is shown in Figure 4.6. It varies from 48.1% to 56.9% for Leitrim (8.8% variation), 41.7% to 58.8% for Meath (17.1% variation) and 47.7% to 51.1% for Tipperary North (3.4% variation).

4.4 Discussion

The results presented suggest that the scale of the soil-topography relationship varies for both different types of soils and different types of geomorphological areas. Terrain attributes have been shown (Figure 4.2) to be sensitive to the scale of the source DEM and behave in different ways to this alteration, ultimately affecting DSM analysis as suggested by Behrens *et al*. (2010b). The effects of spatial scale based on pixel and window size alterations to map soil classes were proved to be significant, and it was found that the best pixel and window sizes varied with geomorphology and soil complexity.

In general validation performances ranging between 35% and 60% are comparable and consistent with previous studies that used machine learning models (Behrens *et al.* 2010b and Grinald *et al.* 2008). The results obtained by the two tested inference engines (NN and RF) are comparable and present the same scale behaviour for the three study areas. As a guideline two main patterns of behaviour were observed:

i) Flat homogeneous areas seem to prefer coarser resolutions (above 140 m in this study) across all the tested window sizes and

ii) Morphologically varied areas, with characteristic features such as the drumlins or abrupt changes in topography reflected in steep slopes, seem to prefer fine resolutions (30 m in this study) with small window sizes but also show good performance at coarser resolutions and large window sizes.

This generally confirms some of the hypothesized scale behaviour by Pain (2005) and Thompson *et al.* (2001); with finer scales required only in morphologically

more complex areas. The ANOVA results reflect this behaviour, where the morphological flat areas generally show improvement for both larger window and pixel sizes, and there is no interaction between these effects. In morphologically more complex areas, varying window size has no impact on performance, whereas pixel size does impact; varying both results in the complex interaction represented in Figure 4.4 and 4.5.

The presented experimental methodology, was not only instrumental in examining the effects of scale on spatial non-stationarity for soil-topography relationships, but also provided important insights on how scale affects a model's explanatory power as some areas were never correctly classified, failing to be related to the local topography. The scale dependence of terrain attributes can account for the appropriate classification of soil taxonomic units in many cases in the study areas, but remain cases where the soil class is incorrectly classified, even at the most optimal scale combination (Figure 4.6). These areas are in locations where other soil forming factors such as parent material have a greater influence on soil class categorisation than terrain factors, as in the case of the southern extent of Leitrim. Here, misclassified areas include similar soils with subtle differences in soil parent material. In the north of Meath an area always misclassified by the model includes a peat complex where vegetation cover and land use might be the main factors controlling soil class categorization. Equally, for some areas in Tipperary North podzols might be better categorised by parent material composition or other environmental covariates that provide detailed information on soil moisture, as this is a critical aspect in soil series classification. In addition, subsequent rationalisation of the soil taxonomic units used in the 6 inch mapping has indicated that some of the detailed criteria for the original differentiation in soil taxonomic units were not justified for the soil unit concept at this mapping scale. Thus misclassification can occur in the model if differentiation is expected from the 6 inch mapping but in practice the soils have very similar characteristics. General inaccuracy of the legacy soil data can also affect the overall classifier's accuracy and might be responsible in particular for some of the areas being consistently misclassified by the two models. The 6 inch soil maps

currently represent the most accurate detailed soil information available at the large scale in Ireland and were selected as part of the ISIS project.

4.5 Conclusions

In this chapter the effect of fine scale DEMs in DSM proved to not always be the best choice, questioning the common approach used in the DSM community of using the finest available DEM in DSM analysis. This current practice implicitly relies on the assumption that the overall effect of scale should balance out but the results show this depends on the morphology of the examined landscape. An exploratory test at different scales, like the presented one, could improve the final prediction of soil taxonomic units and certainly provide useful knowledge of the intrinsic characteristics of the area under scrutiny.

These experimental results set the scale benchmark for the next chapters in which different approaches will be tested and a new multiscale methodology developed. Two main patterns of scale behaviour have been described: flat areas obtaining the best classification accuracies at coarser pixel sizes and morphologically varied areas being influenced by the interaction of pixel and window alterations, obtaining the best accuracies at fine resolutions with small window sizes but also at coarser resolutions and large window sizes. This scale behaviour clearly indicates that the tested areas are not coherent in their scale response, as the scale of the soil-topography relationship varies for both different types of soils and different types of geomorphological area. The three areas will require further subdivisions needing an additional step in the DSM methodology able to segment the landscape. Moreover, this approach was computationally and labour intensive, so in the subsequent chapters alternatives will also be explored.

5 EMPIRICAL APPROACHES

5.1 Introduction

Despite the uncertainties related to choosing the most suitable pixel size to be used in DSM for analysis, visualization or modelling; DSM practitioners have to make a decision and select a pixel size. In the majority of cases, as seen in the previous chapter, the finest available DEM is chosen without proper consideration of its suitability for the data or process examined. This is probably due to the fact that common sense would suggest that using the most detailed dataset available will guarantee that the information needed to explain the process investigated must have been captured in the large amount of information stored and our models or statistical techniques will be able to exploit and make sense of that. Another reason might be the lack of a complete and definitive methodology to address the issue of scale with only empirical guidelines available. In order to guide the selection of an optimal pixel size, a series of empirical approaches have emerged in different applied fields from cartography, GIS, hydrology, remote sensing to computer science. The concept behind them is that they are easy to learn, straightforward to use and generally provide a unique or at least defined answer. Their use is general and extensive but has not been exhaustively evaluated yet with respect to DSM applications.

In the previous chapter the optimal DEM resolution for the three test areas was established experimentally. In this chapter the hypothesis that empirical approaches can be transferred to DSM to identify an optimal DEM pixel size will be tested by comparing these approaches with experimental results. A set of eight empirical approaches has been selected from the literature (Avery, 1987; McBratney, 1998; Rossiter, 2003; Hengl, 2006; Sharma *et al*., 2011; Hengl *et al*. 2013) which are representative of common empirical approaches based on: ESRI ArcGIS; sampling support; cartographic; object orientated; inflection points; information content; sink analysis and fractal dimension of stream networks.

These approaches are assumed to be quick ways to determine optimal scales and are based on very different assumptions.

The selected approaches will be tested in the context of DEM pixel size selection for the test areas, investigating their potential utility in DSM applications by comparing them with the results of the experimental methodology.

5.2 Theory

The raster model in DSM is seen as a useful data structure in which most of the technical characteristics of soil information are controlled by a single measure: pixel size (McBratney, 1998). Pixel size has a spatially explicit location and contains a single value for the soil target attribute in addition to a given set of environmental covariates. Information content of raster DEMs, used in DSM as covariates to derive terrain attributes, progressively decreases with the coarsening of pixel size. As presented in Chapter 4 (Experimental Methodology), this has proved beneficial for DSM analysis as particular pixel sizes are better at classifying soil classes. As previously discussed, in the literature issues related to scale and the choice of an optimal pixel size have been investigated in different applied fields with emerging empirical approaches used as potential solutions. A detailed review of the literature has been carried out selecting eight approaches focusing on all aspects of DSM modelling: the GIS software (ESRI ArcGIS), the soil survey (sampling support), the soil map (cartographic), the soil polygons (object orientated), the DEM topographic characteristics (inflection points), the DEM data content (information and complexity) and the hydrological characteristics of the study area (sink analysis and fractal dimension of stream network).

5.2.1 ESRI ArcGIS

Soil information, DEMs and terrain attributes are commonly managed, analysed and visualised using GIS software. The algorithm used in raster pixel size

selection by de facto standard in GIS (ESRI ArcGIS, the software used for this research) was tested.

5.2.2 Sampling support

The main objective of a soil survey is to provide information on the soil resources in a particular area, but this information will only be as useful as the precision and accuracy of the soil data gathered by the surveyors in the field (Avery, 1987). Soil surveyors have historically based their efforts on the delivery of two main pieces of work: a soil report and a soil map. A map was created based on the field sheets drawn on site, supported by the soil samples collected in the field. Incorporating precision, accuracy and scale, soil surveyors generally refer to survey intensity or inspection density to assess the quality and reliability of a map. This concept can be easily transferred to DSM, estimating an optimal pixel size from the area of the investigated region and the likely inspection density of the soil survey.

5.2.3 Cartographic

Cartography has historically involved the study of maps and map making processes, and more generally is involved with the way in which spatial information is communicated (Visvalingam, 1990). Maps as representations of reality need processes of selection, classification, displacement, symbolization and exaggeration to effectively accomplish their purpose of communication. The fundamental aspects of research in cartography are related to the design and editing of maps, projection systems and generalization techniques. Eliminating or simplifying characteristics of features that are not significant to the map's purpose has been the traditional focus of generalization. Another particularly interesting aspect, especially in the analysis of scale, is the area of generalization which is concerned with the reduction of features complexity at a particular scale. In cartography, fine pixel sizes are connected with large map scales and small areas, and coarse pixel sizes to small map scales and large areas. Both these aspects of scale have been analysed by Vink (1975) defining two metrics to quantify them: minimum legible delineation (MLD) and maximum location

accuracy (MLA). The work of Avery (1987) and Rossiter (2003) has used both these concepts to connect them with inspection density of soil surveys. More recently, Hengl (2006) has expanded this work to include optimal DEM pixel size selection for DSM applications using MLD and MLA.

5.2.4 Object orientated

Even though vector data formats are better suited to represent spatial objects, raster formats can be equally effective in the case of phenomena with abrupt changes or by using a thresholding operation. Pixel size influences the characteristics of objects, as these are scale dependent, controlling their number, size and shape which all vary with the coarsening of pixel size (Lillesand *et al.*, 2008). This concept can be extended to DSM for the selection of an optimal pixel size to represent soil polygons using DEMs.

5.2.5 Inflection points

DEMs are raster based data which describe the spatial distribution of elevation. Their data model is centred on GIS field-based ontology (Smith and Mark, 2003). The resulting raster data structure uses a regular grid of pixels to record the continuously changing elevation over an area, according to the variation of the phenomenon represented. In the literature there are approaches paying attention to the topographic characteristics of the DEM, like the inflection points (Kienzle, 2004). Contour lines join points of constant elevation and are an effective way of illustrating the shape of a surface, highlighting the relevant geomorphological changes of a landscape (Mackaness and Steven, 2006). Pixel size controls the detail of information in a DEM allowing the detection of relevant features such as peaks or valleys. It also controls smoothing or even removing them with the coarsening of pixel size. The variability of the landscape dictates the needed pixel size to accurately preserve these relevant features. The optimal pixel size should allow keeping this variability and maintaining the majority of the relevant geomorphological features (Borkowski and Meier, 1994).

5.2.6 Information and complexity

In computer science, information content or complexity has a central role in the relationship between information and computation, resulting from the combination of information theory with the mathematical foundation of computer science in algorithmic information theory. Shannon information theory (1965) and Kolmogorov algorithmic information theory of complexity (1965) were introduced with separate motivations but a similar aim to define and measure information. This work established a common unit of measurement, the bit (b). This unit, described as the amount of information in an object that could be explained by the length of the description needed to describe the object. These concepts underpinning modern computer science can be used to analyse the variability in elevation of DEMs. By considering complexity as a spatial phenomenon, Hengl *et al.* (2013) show how information content changes for soil polygon maps rasterized to different resolutions. This suggests that optimal pixel size could be determined by the size of compression algorithms (Allegrini *et al.*, 2003) and information content entropy (Wise 2012).

5.2.7 Sink analysis

Hydrology studies the movement and distribution of water on earth. This discipline is deeply interested in the analysis of DEMs, as terrain determines how and where water flows (Bloschl and Silvapalan, 1995). To model the flow of water and perform quantitative analysis, any DEM must be pre-processed to remove sinks, as surfaces with uninterrupted flow are needed. This is of extreme importance in hydrological modelling as sinks are areas that do not drain anywhere, causing the drainage network to be disconnected and have sections missing, essentially leading the flow algorithm into an endless loop of research. Tools have been created to deal with this problem, firstly analysing the DEM to locate any existing sinks and subsequently to fill the elevation of this incorrect depression (Maidment, 2002). Sinks can be generated by the interpolators used in DEM creation due to unsuitability of data density and spatial distribution (Sharma *et al.*, 2011). Sinks are particularly important in DSM as these

geomorphological features influence the spatial distribution of wet / dry conditions in floodplain areas, water moisture due to water stagnation and the complex relationships between water and soil affecting characteristics such as soil depth and overall soil formation. This concept can be exploited for the analysis of scale, as the number and total area covered by sinks should give an indication of the optimal resolution that maximises the water flow, so indirectly assessing the suitability of the DEM.

5.2.8 Fractal dimension of stream network

The pioneering work of Mandelbrot (1983) describes natural forms and processes as mathematical sets that exceed their topological fixed dimensions in regular Euclidean geometry. Based on this work, the concept of fractal and fractal dimension has been intensely investigated and applied in hydrology and other environmental sciences (Lanza and Gallant, 2006). Fractals are characteristically self-similar in the sense that a feature is precisely or closely similar to a part of itself (Falconer, 1990). Fractals can be exactly the same at every scale or have characteristic scales in which the pattern repeats itself. This principle is exploited for the description of complex forms or processes where they allow reduction of information to just one descriptor. This is extremely effective in scale analysis as fractals imply that variability exists at a range of scales, allowing this relationship to be quantified and compared across a series of different scales (Bloschl and Silvapalan, 1995). The main use of fractals in hydrology is in the analysis of stream networks where fractals can be applied in the detection of critical scales and this concept might be extended to DSM for the selection of an optimal DEM pixel size.

5.3 Materials and Methods

The eight selected empirical approaches were applied to the DEM of the three test areas and compared with the experimental results presented in Chapter 4.

The pedological and geomorphological complexity of the three areas should allow a full exploration of the different empirical approaches and assess their suitability in the selection of an appropriate DEM pixel size to be used for DSM applications.

5.3.1 ESRI ArcGIS

The selection of a suitable pixel size for a raster dataset in ArcGIS is handled by the software in the background providing the user with only a specific resolution as a default (with the possibility to manually change the value).

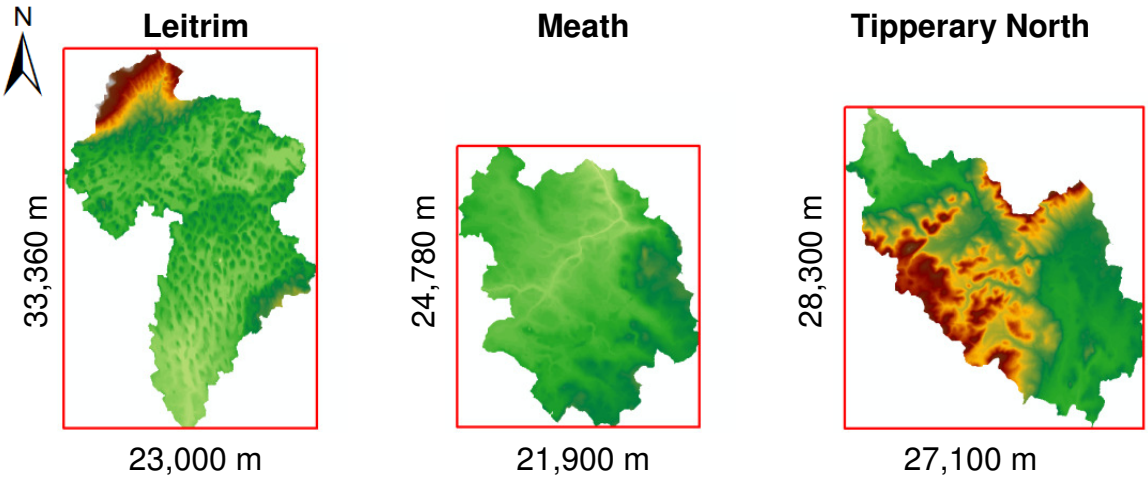


Figure 5.1 - Extent of the three study areas DEMs used by ArcGIS for pixel selection.

The provided pixel value is obtained with a simple rule, in which the system is taking the width or height whichever is shorter of the extent of the feature dataset (Figure 5.1) and divides it by 250.

5.3.2 Sampling support

A common rule of thumb in soil mapping (Avery, 1987) is to use an inspection density of four observations per square centimetre of produced map. This value is the average density that should be kept constant across the entire survey area to be able to uniformly inspect the soil resources ensuring consistency.

This concept can be exploited in the investigation of an optimal pixel size calculated as:

$$p = \sqrt{4 \times \frac{A}{N}} \times 10^2 \tag{5.1}$$

where p is derived by the square root of 4 times the total area (A) divided by the number of samples (N) and all multiplied by 100.

This can be further extended for the two most common approaches of surveying: random sampling (free survey, a) or systematic sampling (regular grid, b) obtaining:

$$a) \quad p = 0.25 \times \sqrt{\frac{A}{N}} \qquad\qquad b) \quad p = 0.5 \times \sqrt{\frac{A}{N}} \tag{5.2}$$

The difference between the two types of surveys depends on the fact that on a regular grid the distance between the point is predetermined and fixed but in a random sampling scheme there is a high probability to have clustered samples requiring roughly half the spacing between closest samples pairs (Hengl, 2006).

5.3.3 Cartographic

Cartographic techniques developed for a world of paper maps can be related to digital data represented in GIS systems (Goodchild, 2001) as national coverage DEMs are still obtained from interpolation of contour lines derived from digitised topographic maps. As previously introduced, scale as spatial resolution is strongly connected to traditional cartographic concepts of MLD and MLA.

MLD, which is the smallest area that can be represented at a particular map scale, is calculated as a function of the map representative fraction or scale number (SN):

$$MLD = SN^2 \cdot 0.000025 \quad (5.3)$$

According to Hengl (2006) MLD can be applied to calculate a suitable pixel size based on two assumptions: MLD can be considered equivalent to 4 pixel cells (Rossiter, 2003) and that MLD on the map is equal to 0.000025 m^2 (Vink, 1975) The resulting pixel values are then calculated as:

$$p \leq \sqrt{\frac{MLD}{4}} = SN \cdot 0.0025 \quad (5.4)$$

MLA, which is the smallest legible resolution, can range from a minimum 0.00025 m to a maximum 0.0001 m on the map (Vink, 1975). It is possible to use this range to estimate pixel size according to:

$$p \geq SN \cdot MLA = SN \cdot 0.00025\ (0.0001) \quad (5.5)$$

5.3.4 Object orientated

The correct representation of spatial objects, like soil polygons, with the smallest area and narrowest shape on the map is a function of pixel size. It is normally accepted that at least four pixels are needed to represent the object with the smallest area (a_{MLD}) and two for the object with the narrowest shape (w_{MLD}) (Hengl, 2006). This can be exploited in DSM to quantify the minimum pixel size necessary to correctly represent a soil map according to the formula:

$$p \leq \begin{cases} \frac{\sqrt{a_{MLD}}}{4} & S < 3 \\ \frac{w_{MLD}}{2} & S > 3 \end{cases} \quad (5.6)$$

where S is the shape complexity index calculated as the perimeter of the object divided by the boundary ratio of a circle of equal area.

5.3.5 Inflection points

A characteristic transect of 20km has been selected and extracted for the three study areas (Figure 5.2).

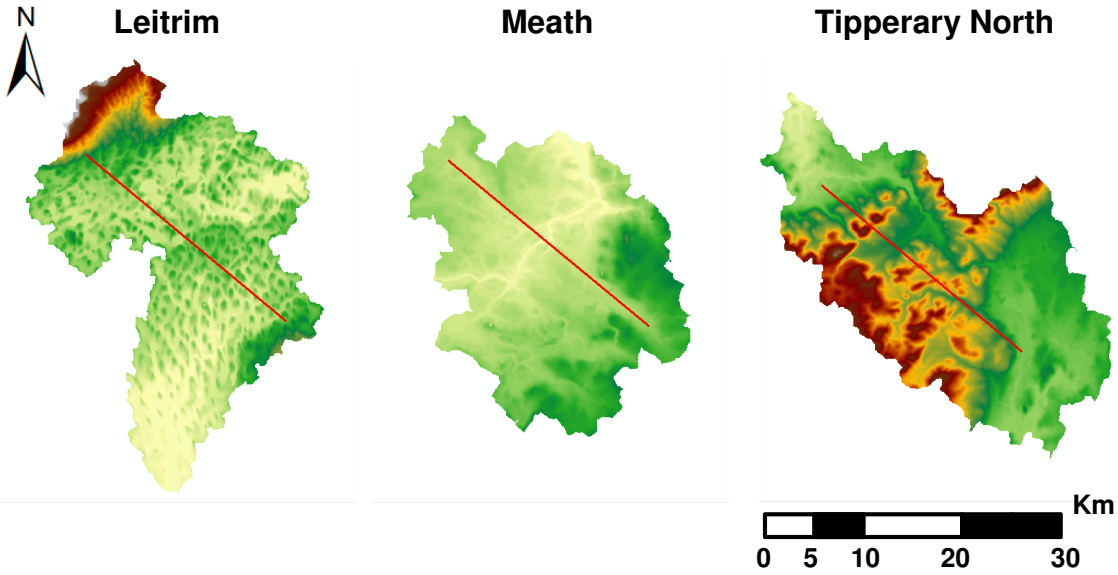


Figure 5.2 - Characteristic transects for the three study areas overlying the DEMs.

In order to calculate an optimal pixel size based on the terrain complexity, the elevation can be considered as the signal and the density of inflection points on the one dimensional transects the frequency of the signal, resulting in:

$$p \leq \frac{l}{2 \times n(\delta z)} \tag{5.7}$$

where the pixel size (p) is calculated as the total length of the transect (l) divided by two times (half the average spacing) the number of inflection points ($n(\delta z)$). The inflection points used in the formula are points on a curve at which the curvature or concavity changes from being concave upwards to concave downwards or vice versa. In simple words, these are the peaks or valleys where topography rapidly changes from climbing to descending in the case of a peak or from descending to climbing in the case of a valley. These have been visually assessed and manually determined for the three areas.

The same concept applies to a two dimensional situation where pixel size can be determined from the total length of contour lines (l) according to:

$$p = \frac{A}{2 \times \sum l} \tag{5.8}$$

where the total area (A) is divided by two times the summation of the contour lines ($\sum$ l).

5.3.6 Information and complexity

Kolmogorov complexity can be defined as the size of the smallest program that produces as output the raster investigated. This is closely related to compression algorithms (Kidner and Smith, 2003) where a sequence of numbers expressed in bits ($l_z(\omega)$) is compressed with an algorithm (z) to sequence (ω) of length (n) and the resulting complexity (K) can be calculated as:

$$K(\omega) = \lim_{n \to \infty} \sup \frac{l_{z\,(\omega^n)}}{n} \tag{5.9}$$

For Shannon entropy was the measure of unpredictability in the random variable "elevation" corresponding to its information content expressed in bits. This was firstly applied for the analysis of DEMs quality investigating the scale effects on derived terrain attributes used in hydrological and soil erosion modelling (Vieux, 1993; Vieux and Farajalla, 1994; Mendicino and Sole, 1997; Wang *et al.* 2001). These researches highlighted the potential of entropy as a measure of DEM quality proving the relationship with its information content, the effects of resampling at coarser pixel sizes and finally the effects of aggregation and smoothing on the reduction of entropy and loss of quality.

The entropy (H) was calculated as:

$$\mathrm{H} = -\sum_{1=1}^{m}(P_i \times \log_2 P_i) \quad (5.10)$$

Where $\mathrm{P_i}$ is the probability of a pixel being classified as class type i and m is the number of classes. The theory suggests that entropy should decrease with the decline of information content. In the context of DEM analysis this should equate to minimum values of entropy for areas with low or null variability such as flat regions and maximum entropy for highly variable regions. According to Sharma *et al.* (2011) if a DEM is oversampled with fine resolution pixel sizes or undersampled with coarse one, the low spatial variation resulting from redundancy of information (fine resolution) or loss of micro relief (coarse resolution) should reduce the level of entropy thus the relevant information content and so relate to the variability observed in the DSM covariates.

5.3.7 Sink analysis

The number of sinks applied in this analysis will be determined using the sink algorithm developed for the Arc Hydro tool (ESRI, 2010), which does not require any parameterization. For each DEM pixel size tested, the number of sinks will be integrated with the total area of sinks revealing the hydrological alteration caused by that particular scale.

5.3.8 Fractal dimension of stream network

Fractal dimension, a scaling index of fractal nature, was calculated with the Box-counting method (Sarkar and Chaudhuri, 1992; Taud and Parrot, 2005; Sun *et al.*, 2006) in which datasets are collapsed into smaller pieces of information according to the box size and shape. Analysis based on this way of gathering data allows complex patterns to emerge and examine how observations of detail change with scale. The advantage of this technique is that rather than changing the magnification at which data are observed, box counting consent to change the size of the box used to inspect the data or process analysed (Abedini and Shaghaghian, 2009). This is very important for a scale analysis based on fractal

dimension as otherwise the scaling properties of the method would be imposed on the data and in the end would damage the analysis.

As stated by Sun *et al.* (2006) the method to calculate the fractal dimension (F_D) can be easily described using three main steps:

- Counting the number of boxes (N) needed to cover all the stream network features and gradually increasing the size of the boxes using an iterative process in which each step (s) is determined following a power of 2;
- Creating a scatter plot of the transformed (log) number of boxes versus the transformed (log) number of steps and fit a regression line;
- Using the slope of the regression line to calculate the fractal dimension.

The final two steps can be incorporated and resolved through a formula to calculate the fractal dimension:

$$F_D = \frac{\log N}{\log s} \tag{5.11}$$

The stream network of the three study areas was generated for the five tested pixel sizes (20 m; 80 m; 140 m; 200 m and 260 m) using the DEMs previously processed during the sink analysis approach. For each resolution a flow direction and a flow accumulation datasets were generated using the accumulation threshold method via the ESRI tool Arc Hydro from which the stream networks lines were then converted. The threshold, which is the number of pixels used to identify a stream, was selected according to the methodology presented by Sharma *et al.* (2011). Threshold values for each DEM were calculated as a proportion (1%) of the number of pixel sizes, *e.g.* the threshold for the 20 m DEM (Leitrim = 9272; Meath = 8428 and Tipperary North = 9352) is sixteen times as that of the 80 m DEM (Leitrim = 579; Meath = 527 and Tipperary North = 584); forty-nine times as that of the 140 m DEM (Leitrim = 189; Meath = 172 and Tipperary North = 191); and so on. The proportionate selection of accumulation

thresholds should avoid the chance of over-densification of the stream network for finer resolution DEMs.

5.4 Results

The results are presented in Figures 5.3 - 5.12 and Tables 5.1 - 5.8. Table 5.1 illustrates pixel sizes estimated from sampling support. Table 5.2 illustrates pixel sizes estimated from MLD and their relationship with the relative scale numbers is presented in Figure 5.3; Table 5.3 illustrates pixel sizes estimated from MLA and their relationship with the relative scale numbers is presented in Figure 5.4; finally Figure 5.5 combines both MLD and MLA relationships summarising the results related to the cartographic concepts. Figure 5.6 presents the soil polygon patterns for the three study areas and Table 5.4 shows the pixel sizes estimated from aMLD and wMLD. Results from the inflection point approach are presented in Figure 5.7 (1D transects) and Figure 5.8 (2D contour lines). Complexity and information theory is presented in Table 5.5 (compression), Table 5.6 (entropy) and Figure 5.9 (normalized entropy behaviour). Based on the hydrological characteristics of the study areas, the last two approaches are summarised: Figure 5.10 illustrates the distribution of sinks for the three study areas; Table 5.7 presents the sink analysis parameters; and Figure 5.11 shows the sink analysis overall trend at increasing pixel sizes for the number of sinks and total sink area. Figure 5.12 displays the derived stream networks for the three study areas, while Table 5.8 contains the fractal dimensions, networks length and density. The results for each approach are now analysed in detail.

5.4.1 ESRI ArcGIS

Leitrim with an extent of 33,360 m x 23,000 m according to the ArcGIS rule used for pixel selection should have a pixel size of 92.0m as the width is smaller than the height (p = 23,000 / 250). Based on the same principle, Meath should have a pixel size of 87.6 m, as the width in this case is smaller than the height (p = 21,900

/ 250), and Tipperary North a value of 108.4 m, as even in this case the width is smaller than the height (p= 27,100 / 250).

5.4.2 Sampling support

Pixel sizes for each study area were calculated according to the inspection density rule of thumb for both a free survey and a regular grid sampling (Table 5.1).

Table 5.1 - Pixel sizes estimated from sampling support.

	Area [m^2]	Samples	p (random) [m]	p (systematic) [m]
Leitrim	370,877,182	1,483	125	250
Meath	337,101,600	1,348	125	250
Tipperary North	374,118,307	1,496	125	250

As a result of the same sampling density of 4 points per square kilometre the three study areas obtain equal pixel size results: 125 m for the random sampling and 250 m for the systematic sampling.

5.4.3 Cartographic

An estimated pixel size according to MLD is presented in Table 5.2 for the 1:10,560 detailed reconnaissance survey map, the 1:126,720 rationalised reconnaissance map and the 1:250,000 European target scale.

Table 5.2 - Pixel sizes estimated from MLD.

Scale	SN	MLD [m^2]	p [m]
1 : 10,560	10,560	2,788	26.4
1 : 126,720	126,720	401,449	316.8
1 : 250,000	250,000	1,562,500	625.0

The results for the 1:10,560 scale are rather coarse with a MLD of 2,788 m^2 and a pixel size of 26.4m, increasing to 401,449 m^2 and 316.8 m at the rationalised reconnaissance map scale level and suggesting a suitable pixel size of 625 m and a MLD of 1,562,500 m^2 for the European target scale (1:250,000). Using these results and expanding the equation, it is possible to generalize a relationship between pixel size and scale (Figure 5.3) highlighting a linear trend of the maximum pixel size according to MLD. The arrows in the figure show that pixel size should be equal or less than the pixel size, as represented by the trend line.

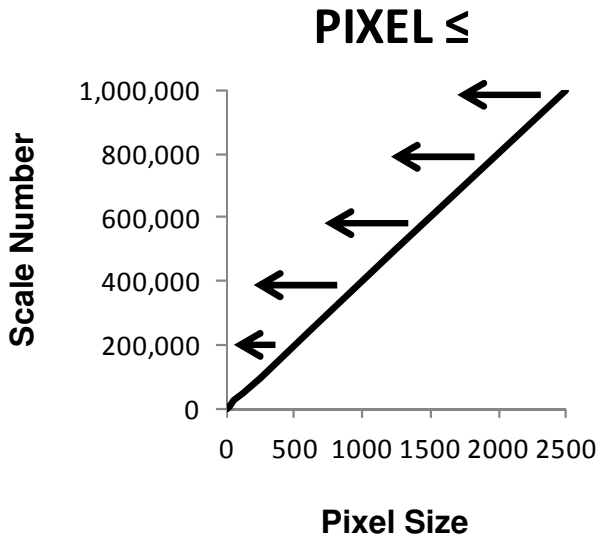


Figure 5.3 - Pixel size and scale number relationship using MLD.

An estimated pixel size according to MLA is shown in Table 5.3 for the two extreme values of the resolution range (from a minimum of 0.00025 m to a maximum of 0.0001 m). The results suggest that for the detailed reconnaissance survey map a pixel greater than 1.1 m (0.0001) or 2.6 m (0.00025) will allow the correct visualization on the map of soil features. For the 1:126,720 scale pixel values will range between 12.7 m and 31.7 m and at the target European scale the pixel limit should be set no smaller than 25.0 m, or for best legibility no smaller than 62.5 m.

Table 5.3 - Pixel sizes estimated from MLA for both 0.00025 and 0.0001 values.

Scale	SN	MLA [m]	p [m]	MLA [m]	p [m]
1 : 10,560	10,560	0.00025	2.6	0.0001	1.1
1 : 126,720	126,720	0.00025	31.7	0.0001	12.7
1 : 250,000	250,000	0.00025	62.5	0.0001	25.0

As previously done for MLD a graph presenting the relationship between pixel size and scale has been created for MLA (Figure 5.4) illustrating the limits (0.0001 m maximum limit of accuracy achievable on paper map and the 0.00025 commonly considered value for the smallest legible resolution) of the minimum pixel size according to MLA. The arrows in the figure show that pixel size should be equal to or greater than the pixel size represented by the trend line.

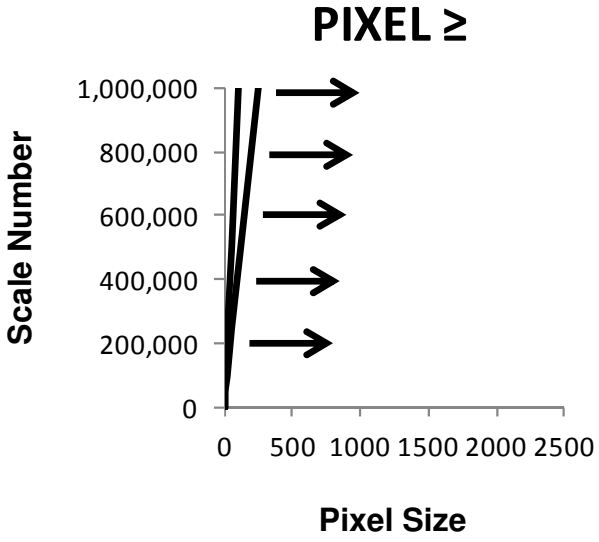


Figure 5.4 - Pixel size and scale number relationship using MLA

From the combination of the previously described relationships between pixel size and scale calculated for MLD and MLA a graph showing the optimum range according to cartographic concepts related to map design, visualization and generalization has been created (Figure 5.5). The arrows, in this case, show that pixel size should be contained between the pixel sizes represented by the two trend lines.

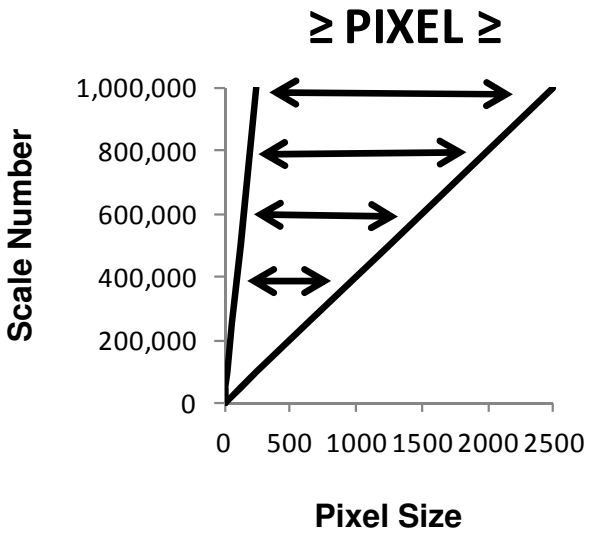


Figure 5.5 - Pixel size and scale number relationship combining MLD and MLA.

According to these two cartographic concepts the 1:10,560 detailed reconnaissance survey map should have a pixel size value between 2.6 m and 26.4 m; the 1:126,720 rationalised reconnaissance map between 31.7 m and 316.8 m and the European target scale of 1:250,000 a pixel size between 62.5 m and 625.0 m.

5.4.4 Object orientated

The soil polygons with the smallest area (a_{MLD}) and narrowest shape (w_{MLD}) were selected from the 6 inches soil map for the three study areas (Figure 5.6), which have respectively 332 polygons (average size of 1.11 km^2) for Leitrim, 227 polygons (average size of 1.48 km^2) for Meath and 455 polygons (average size of 0.82 km^2) for Tipperary North.

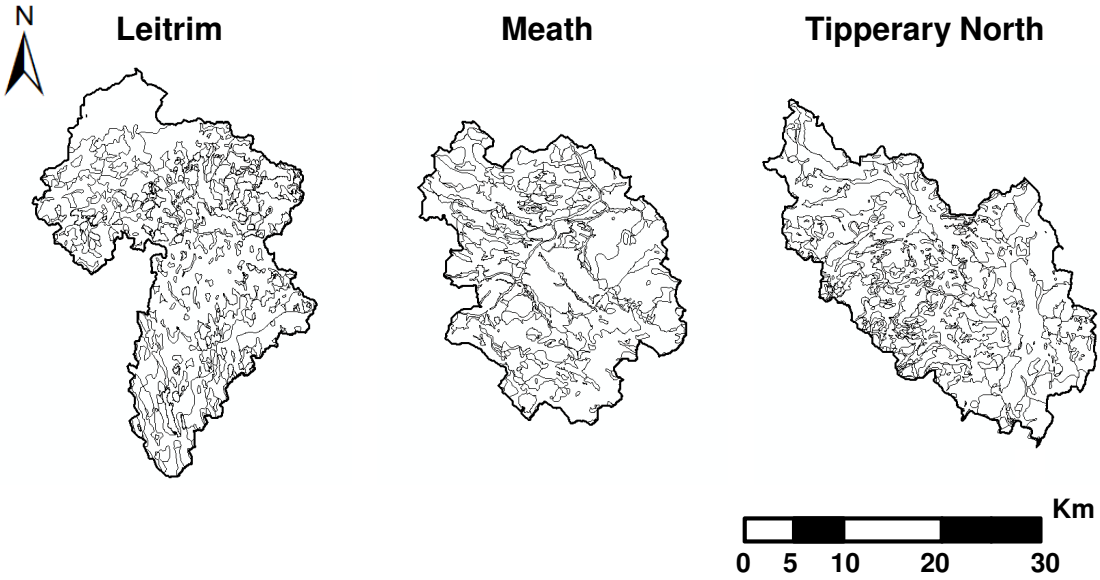


Figure 5.6 - Soil polygons patterns for the three study areas.

As presented in Table 5.4, a_{MLD} in the three study areas range from 3,382 m^2 for Tipperary North, 4,960 m^2 for Meath to 7,508 m^2 for Leitrim. w_{MLD} in the three study areas range from 38 m for Meath, 48 m for Tipperary North to 66 m for Leitrim.

Table 5.4 - Pixel sizes estimated from a_{MLD} and w_{MLD}.

	a_{MLD} [m^2]	p [m]	w_{MLD} [m]	p [m]
Leitrim	7,508	22	66	33
Meath	4,960	18	38	19
Tipperary North	3,382	15	48	24

Leitrim has the soil polygons with the largest a_{MLD} area and widest w_{MLD} shape, resulting in a pixel size of 22 m and 33 m respectively. Meath has the soil polygon with the narrowest w_{MLD} and an intermediate value for a_{MLD} resulting in very similar values of pixel size of 18 m and 19 m. According to the formula, Tipperary North, characterized by the soil polygon with the smallest a_{MLD}, requires a pixel

size of 15 m and to correctly represent the soil polygon with the narrowest shape, a pixel size of 15 m.

5.4.5 Inflection points

One dimensional

Leitrim with the highest number of inflection points, a total of 61 over the 20 km transect obtains an optimal a pixel size of 163.9 m. Tipperary North presents 43 inflection points with a corresponding pixel size of 232.6 m. Meath as expected has a very small number of inflection points, only 13 resulting in a pixel size of 769.2 m, this large pixel value is a consequence of the flat landscape that does not have much variation, except for the river network creating alterations of the otherwise smooth terrain (Figure 5.7).

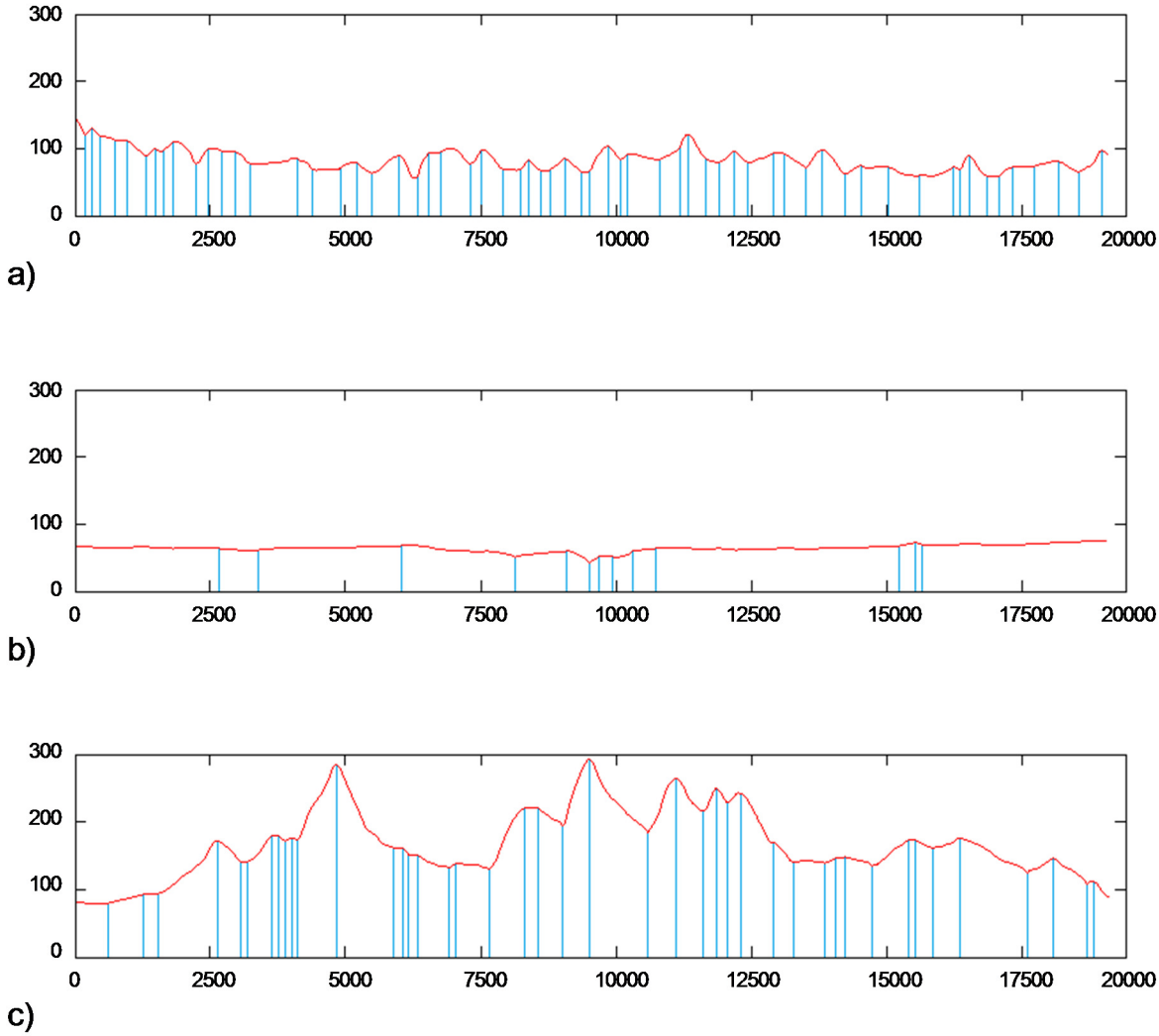


Figure 5.7 - Transects for the three study areas divided by the inflection points: a) Leitrim, b) Meath and c) Tipperary North.

The distribution of the inflection points varies greatly through the three study areas: Leitrim presents a dense and homogeneous distribution as the drumlins characterising its landscape produce a recurring pattern; Meath has very few inflection points with an average distance of 1,500 m between them; Tipperary North has a dense distribution but less homogeneous than Leitrim as the inflection points appears more concentrated.

Two dimensional

The same concept applies to a two dimensional situation where pixel size can be determined from the total length of contour lines (Figure 5.8)

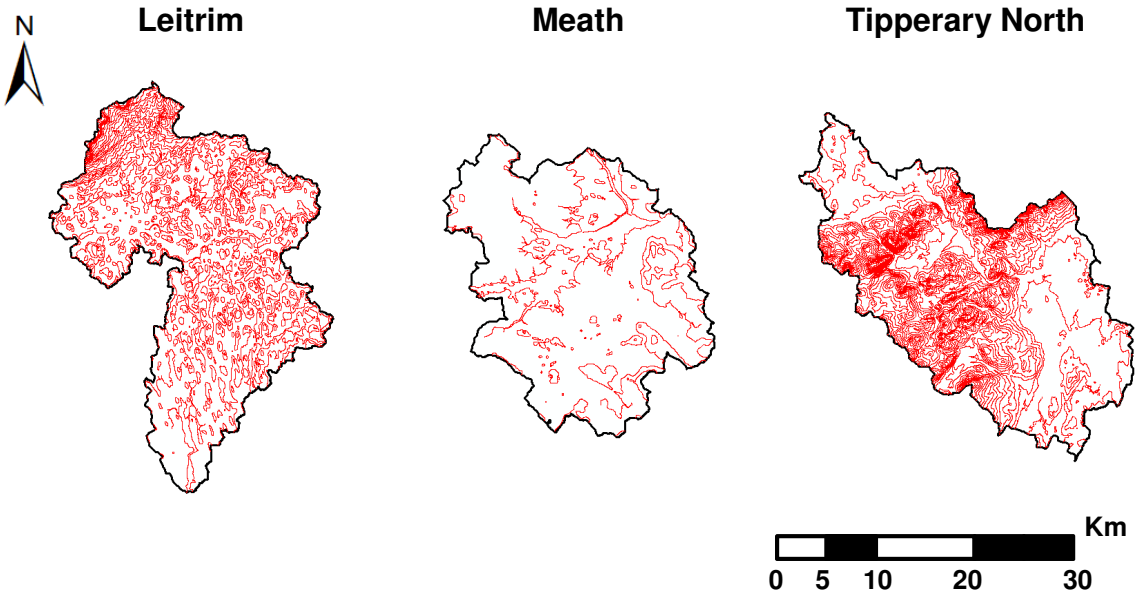


Figure 5.8 – Contour lines for the three study areas (5 m intervals).

Figure 5.8 perfectly illustrates the extremely different landscapes investigated, as Leitrim which is characterised by the drumlin features presents a dense but homogenous distribution of contour lines with a total of 1,431 km resulting in a pixel size of 129.6 m. Meath with its flat homogeneous landscape has hardly any contour lines totalling only 490 km mainly across the river channel in the north side of the study area with a consequent pixel size of 344.0 m. Tipperary North presents a very dense area of contour lines in the middle of the study area characterised by the high relief contrary to the low relief areas at the two opposite sides in the north and in the south with only few sparse contour lines, resulting in an average pixel size of 116.9 m.

5.4.6 Information and complexity

The results of the compression, performed in R using the gzip compression algorithm according to the procedure presented by Hengl *et al.* (2013), are

presented in Table 5.5. For comparison a simplified dataset with the same number of pixels of the tested resolution but containing only a single value (1) was created to evaluate the amount of information stored.

Table 5.5 - Results of the gzip compression algorithm for the tested resolutions: compression size for real data and simplified dataset, in brackets information density [B/km^2].

	Number of Pixels	Compression size [B]		Compression size of plain dataset [B]	
Leitrim					
20 m	927,190	2,962,717	(7,988)	12,318	(33)
80 m	57,944	196,957	(531)	2,790	(7)
140 m	18,932	65,125	(175)	1,691	(4)
200 m	9,269	32,895	(88)	1,184	(3)
260 m	5,478	19,748	(53)	980	(3)
Meath					
20 m	842,754	2,711,131	(8,042)	9,488	(28)
80 m	52,669	181,608	(538)	1,711	(5)
140 m	17,214	60,469	(179)	1,508	(4)
200 m	8,416	29,989	(89)	1,013	(3)
260 m	4,988	18,234	(54)	925	(3)
Tipperary North					
20 m	935,172	3,077,944	(8,227)	11,449	(30)
80 m	58,469	205,896	(550)	1,876	(5)
140 m	19,093	68,952	(184)	1,683	(4)
200 m	9,349	34,867	(93)	1,208	(3)
260 m	5,542	20,774	(55)	960	(3)

The three study areas have different amounts of information according to their sizes, measured in bytes (B). Tipperary North, the largest of the three areas has more information than Meath, the smallest. This also depends on the information density (values in brackets) where Leitrim unexpectedly has the lowest information density with 7,988 B needed to describe each square kilometre, Meath a slightly superior value of 8,042 B/km^2 and Tipperary North the highest

one with 8,227 B/km^2. The plain datasets, the ones that have the same geographical shape and number of pixels of the original data but simplified information content with just one value for elevation, show a different picture. Here it is Leitrim that requires the largest amount of information despite the lowest area covered in comparison with Tipperary North. This probably affected by the more fragmented nature of this area, as it has the longest perimeter of the three and a more elongated shape. Also worth mentioning is the profound effect in terms of information content that resampling at a larger pixel size has on the DEM.

Table 5.6 - Entropy and normalized entropy values for the different pixel sizes.

	Entropy	Normalized Entropy	Change
Leitrim			
20 m	4.989	0.418	-
80 m	4.986	0.523	10.5%
140 m	4.985	0.583	5.9%
200 m	4.985	0.577	-0.5%
260 m	4.986	0.566	-1.2%
Meath			
20 m	4.160	0.351	-
80 m	4.158	0.440	8.9%
140 m	4.162	0.491	5.1%
200 m	4.156	0.529	3.8%
260 m	4.157	0.562	3.3%
Tipperary North			
20 m	5.977	0.601	-
80 m	5.977	0.647	4.6%
140 m	5.976	0.698	5.1%
200 m	5.975	0.752	5.4%
260 m	5.967	0.797	4.5%

Vieux and Farajalla (1994) showed that resampling at coarser scales causes a loss of entropy in DEMs. The global entropy presented in Table 5.6 is influenced not only by the information content but also pixels number making it difficult to

assess the effect on information content from the reduction in the number of pixels. Stoy *et al.*, (2009) suggest normalizing entropy by dividing it with two times the natural logarithm of the total number of pixels. The normalized entropy values differ between the three study areas with the completely flat Meath achieving the lowest value of 0.351 at the original 20 m resolution, Leitrim 0.418 and Tipperary North 0.501 and change according to pixel size variation.

A coarser resolution will lead to a loss of information and hence a loss of normalized entropy. The resampling to a coarser pixel size should affect the elevation histogram distribution with fewer bins present, changing the relative proportions of remaining elevations, and ultimately leading to a loss of entropy.

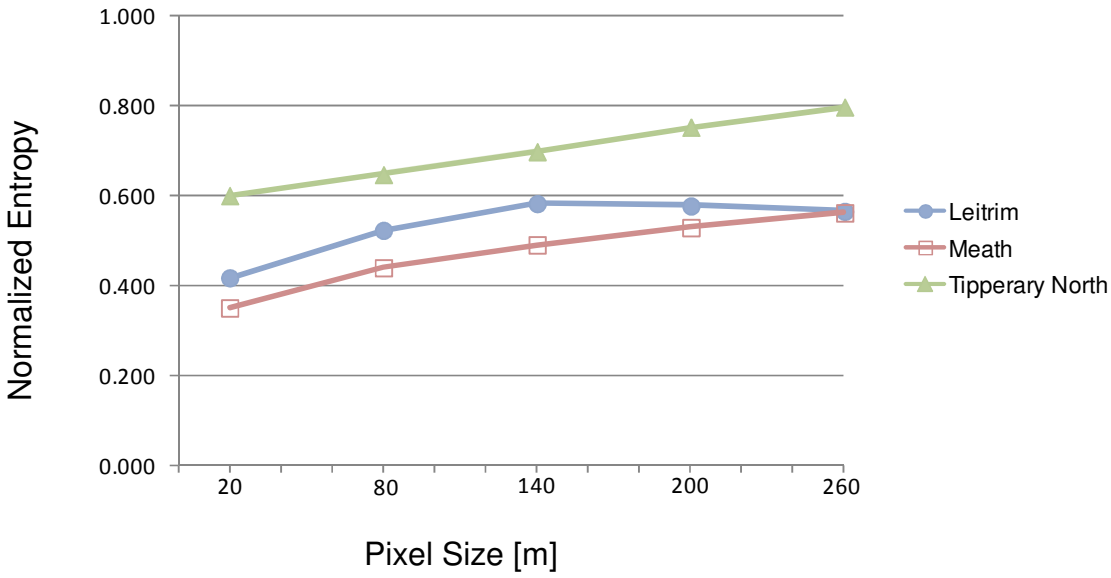


Figure 5.9 - Normalized entropy behaviour at different pixel sizes.

The areas behave very differently with the coarsening of window size as shown in Figure 5.9, Leitrim reaches the maximum level of normalized entropy of 0.583 at 140 m of resolution after which the values start to decrease, while both Meath and Tipperary North present a constant increase with no sign of flattening in the investigated interval.

5.4.7 Sink analysis

Leitrim presents a relevant number of sinks localized around the drumlins, this is probably due to the complex glacial formation of these distinctive geomorphological features that drives the runoff not to flow into a drainage network but to soak into the ground as infiltration. Meath with a characteristic flat landform and behaviour typical of alluvial plains shows a limited amount of sinks in the vicinity of the river and stream network probably due to the insufficient data density around these hydrological features created by the contours lines vertical spacing. With its composite landscape equally divided into lowlands and relief Tipperary North displays a limited amount of sinks in the alluvial plains and a good hydrological connectivity with a very restricted number of sinks in the more mountainous section (Figure 5.10).

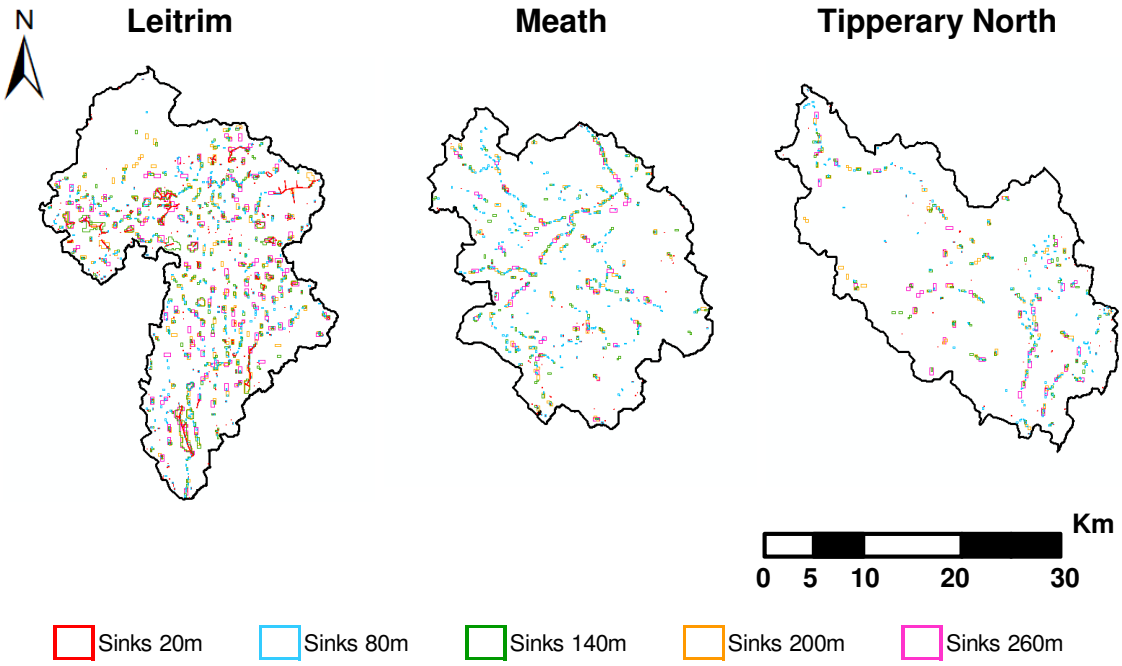


Figure 5.10 - Sink distribution at different pixel sizes for the three study areas.

Results of the sink analysis are shown in Table 5.7, these confirm the visual assessment previously described. Leitrim with the highest number of sinks (486 at 80 m) is definitely the area with the most challenging drainage network due to its complex glacial origin, the number of sinks at 20 m is 403 for a total area of

1,464,800 m^2 covering 0.39% of the study area, it then spikes at 80 m to 486 for a total area of 6,585,600 m^2 covering 1.78% of the study area. It later starts to display an opposite trend, a decrease of the number of sinks is associated with an increase of the total area. This pattern is caused by the growing pixel area coverage despite the fall in the number of pixels taken up by each sink, which is visible in the last column highlighting the number of pixels per sink.

Table 5.7 - Sinks analysis: number, total sinks area, percentage of the overall study area, total number of pixels and number of pixels per sink.

	Sinks	Sinks Area [m^2]	Sinks on Total Area	Total Pixels	Pixels per Sink
Leitrim					
20 m	403	1,464,800	0.39%	3,662	9.09
80 m	486	6,585,600	1.78%	1,029	2.12
140 m	253	17,404,800	4.69%	888	3.51
200 m	179	18,640,000	5.03%	466	2.60
260 m	115	19,536,400	5.27%	289	2.51
Meath					
20 m	118	94,000	0.03%	235	1.99
80 m	291	3,635,200	1.08%	568	1.95
140 m	122	4,468,800	1.33%	228	1.87
200 m	72	5,560,000	1.65%	139	1.93
260 m	44	5,475,600	1.62%	81	1.84
Tipperary North					
20 m	104	91,600	0.02%	229	2.20
80 m	182	2,265,600	0.61%	354	1.95
140 m	84	3,175,200	0.85%	162	1.93
200 m	53	4,040,000	1.08%	101	1.91
260 m	42	5,475,600	1.46%	81	1.93

To some extent, a similar pattern can be seen for Meath despite important differences, the first is the very low number of sinks existing in this study area (118 at 20 m, 291 at 80 m declining to a modest 44 at 260 m) with three quarters less sinks compared to Leitrim, the second observation is the extremely pronounced rise from the 20 m to the 80 m pixel size in which the number of

sinks, the sinks area and total number of pixels almost treble despite the constant number of pixels per sink. Tipperary North has the lowest number of sinks across the tested pixel sizes for the three study areas, the smallest sink areas detected and therefore the lowest percentage of sinks in the total area.

In order to clarify this behaviour, two graphs (Figure 5.11) were created, one for the number of sinks (a) and the other for the total sinks area (b). It is worth mentioning that all three areas have the highest number of pixels at 80 m for no obvious reason. This particular resolution seems to hold an amount of information that disrupt a uniform flow picked by the hydrological algorithm for sink selection. This effect could be not be based on interpolation artefacts created at this particular pixel size, it could have some physical or hydrological reason or simply be based on particular assumptions used for the selection of the number of cells used for analysis on which Arc Hydro is based on. The sinks area graph (b) shows a very similar trend for Meath and Tipperary North with the total area levelling off between 80 m and 140 m and between 140 m and 200 m for Leitrim.

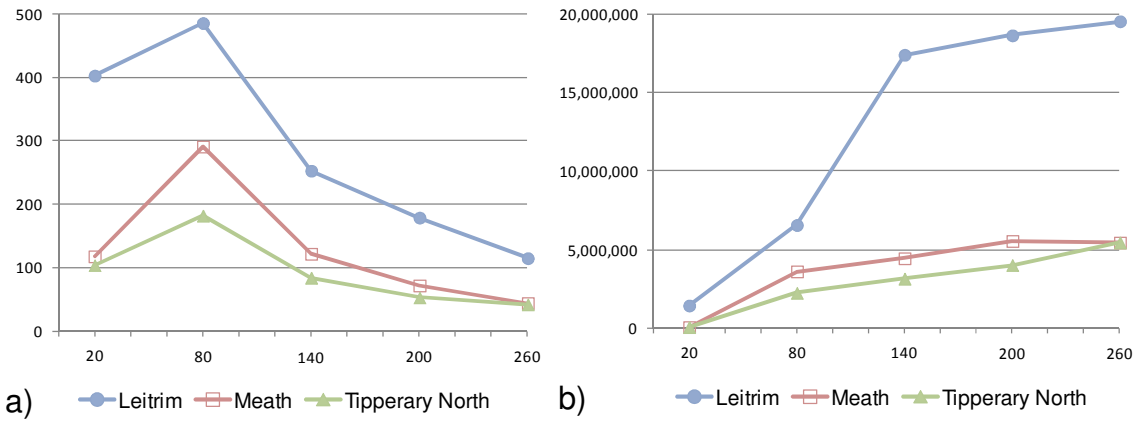


Figure 5.11 - Sink analysis overall trend at increasing pixel sizes for number of sinks (a) and total sinks area (b).

Combining these observations does not suggest a conclusive answer on which pixel size satisfy hydrologically the three study areas as the coarsest resolutions

lower the number of sinks and pixels involved in the formation of these artefacts but at the same time increase the total area of sinks areas.

5.4.8 Fractal dimension of stream network

The results, presented in Figure 5.12 and summarised in Table 5.8, show that Leitrim has the longest and most dense stream network in comparison with the other two areas with a total of 411,963 km (EPA reference water features) and 1.11 km/km^2 expected from its geomorphology characterised by drumlins. Meath on the other hand has the shortest and sparse network with only 316,680 km of streams and a drainage density of 0.94 km/km^2; Tipperary North with intermediate values of 384,055 km and 1.03 km/km^2 reflects the basin characteristics of the area landscape equally divided between high and low reliefs.

As shown on the right hand side of Figure 5.12, the magnified section of the drainage lines visibly illustrates the differences between the stream networks generated with different resolution DEMs. The shape of the drainage lines and the presence of artefacts are clearly visible expressions of the alterations that the topography experiences, such as more straight lines and less smooth bifurcations due to a flattening of local morphology.

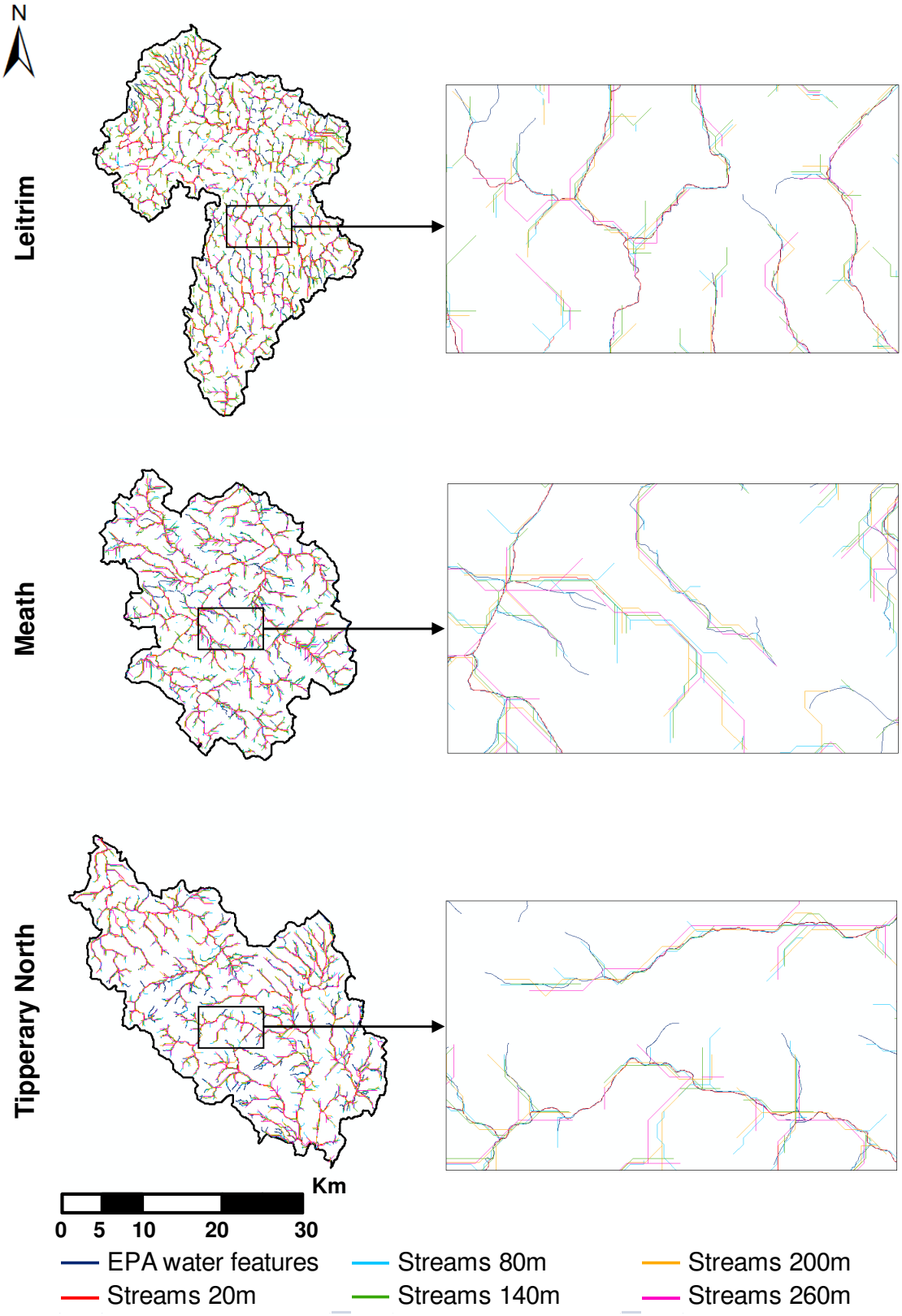


Figure 5.12 - Stream network derived at different pixel sizes for the three study areas.

To better appreciate these changes Table 5.8 presents, for each study area and pixel size analysed, the values of fractal dimension, total length of the stream network the overall network density and the resulting number of features.

Table 5.8 - Fractal analysis: fractal dimension, network length, network density and number of features.

	Fractal Dimension	Network Length [km]	Network Density [km/km^2]	Number of Features
Leitrim				
EPA reference network	1.32	411,963	1.11	42,182
20 m	1.37	323,470	0.87	6,572
80 m	1.45	338,313	0.91	2,142
140 m	1.52	486,937	1.31	1,921
200 m	1.51	370,050	1.00	1,069
260 m	1.50	303,060	0.82	672
Meath				
EPA reference network	1.38	316,680	0.94	11,948
20 m	1.33	147,551	0.44	3,091
80 m	1.53	405,434	1.20	2,418
140 m	1.57	394,377	1.17	1,466
200 m	1.55	286,094	0.85	787
260 m	1.57	345,035	1.02	752
Tipperary North				
EPA reference network	1.31	384,055	1.03	41,059
20 m	1.30	209,118	0.56	4,362
80 m	1.46	429,697	1.15	2,515
140 m	1.47	362,471	0.97	1,287
200 m	1.50	380,229	1.02	1,035
260 m	1.51	365,094	0.98	789

Leitrim, with the highest number of stream features (42,182), has the longest (411,963 km) and most dense (1.1 km/km^2) network in comparison with the other study areas, due to its complex drumlin landscape. Meath on the other hand has the lowest number of features (11,948), least dense (0.94 km/km^2) and shortest network (316,680 km), as expected for its flat landscape. Tipperary North with its

varied geomorphology in which flat areas are interrupted by the steep slopes of the Silvermine Mountains has intermediate values with 41,059 stream features, and a stream density of 1.03 km/km^2 and river length of 384,055 km respectively.

The results of the fractal analysis, presented in Table 5.8, show at least three interesting effects observed in all three study areas regardless of their local morphology:

- Fractal dimension of the stream network is increasing with pixel resolution and the trend appears to flatten above the 80 - 140 m pixel size;
- The length of the network behaves erratically, sharply decreasing at 20 m (-21% for Leitrim, -53% for Meath and -46% for Tipperary North) despite the fractal dimension of the network closely matching the EPA reference network;
- The number of features (straight segments of the network) fall from the very precise digitised EPA network to the computed versions decreasing with the increase of pixel size.

At a detailed visual inspection, the 20 m network closely matches the reference EPA network for the principal streams, but almost disappears completely for the secondary and minor streams, creating hardly any artefacts. This is surprising as the shape of the network better matches the EPA reference (Figure 5.11), supported by a very similar fractal dimension (Table 5.6), but just seems to prune the secondary and minor streams in the most flat and open areas. According to the fractal dimension, the three study areas should be best represented by the finest resolution tested at 20 m with the most similar fractal dimension value compared to the EPA reference network.

5.5 Discussion

The most promising empirical approaches, selected from the literature, were tested to identify an optimal DEM pixel size for DSM applications. They have

shown a diverse range of optimal pixel sizes accordingly to their inherent characteristics. Table 5.9 summarises the results obtained with the use of the eight empirical approaches allowing a detailed comparison with the optimal pixel sizes obtained in Chapter 4 using the presented experimental methodology.

Table 5.9 - Pixel size results according to the eight tested empirical approaches.

	Leitrim [m]	**Meath [m]**	**Tipperary North [m]**
Experimental methodology (no changes in window size)	**20**	**260**	**N/A**
ESRI ArcGIS	92.0	87.6	108.4
Sampling support:			
- random	125.0	125.0	125.0
- systematic	250.0	250.0	250.0
Cartographic	2.6 - 26.4	2.6 - 26.4	2.6 - 26.4
Object orientated:	≥22.0	≥18.0	≥15.0
Inflection points:			
- 1D	163.9	769.2	232.6
- 2D	129.6	344.0	116.9
Information and complexity:			
- complexity	20.0	20.0	20.0
- entropy	140.0	N/A	N/A
Sink analysis	-	-	-
Fractal dimension of stream network	20.0	20.0	20.0

The first approach, which tested the GIS software ESRI ArcGIS rule, provides different resolutions for the three study areas appearing very detailed and precise at first glance. This can cause a false sense of security for the software user not aware of the simple rule used in the selection. The user could also be persuaded

to believe that the value provided is indeed the most suitable resolution for the data. The lack of scientific robustness of the ESRI ArcGIS formula (the shortest value between envelope width or length divided by 250) appears clear if confronted with the previous chapter results. Meath, which is characterised by a flat lowland landscape, preferred coarse pixel sizes while in this case obtains the finest resolution between the three study areas.

The sampling support approach relies on the concept of inspection density. This directly links to the average number of inspections during a soil survey performed by the surveyors in which the “ideal” inspection density according to Avery (1987) is four observations per square centimetre of produced map. The resulting pixel sizes are the same for the three study areas as the value is calculated not taking into account the size or specific data content but estimated only from the map scale, making it an unreliable way to assess DEM pixel size for DSM applications.

The cartographic approach based on MLD and MLA (Vink, 1975) allows selecting the smallest size area that can be represented on the map and the smallest legible delineation as presented by Rossiter (2003). The interaction between these two cartographic characteristics has allowed selection of the most suitable scale range for a particular paper map scale. This technique provides the same answer for the three study areas not taking into account their sizes and their specific data content. Although these are major limitations for DSM applications, the value of this approach lies in the selection of the finest theoretical pixel size below which a paper map loses its capacity of storing meaningful information and the increase of data content is simply redundant. This could be of some use in DSM, not for the selection of the optimal DEM pixel size to be used in modelling, but for the management of storage usage and processing power for the production of paper soil maps.

The correct representation of the soil polygons with the smallest area and narrowest shape on the map is a function of pixel size. This concept presented by Lillesand *et al*. (2008) extends the finest legible resolution just discussed for the cartographic approach, redefining it with inherent data characteristics. The

pixel sizes calculated for the three study areas set the minimum pixel size needed in the representation of objects in maps already created, not in the selection of an optimal DEM pixel size for DSM modelling.

The inflection points approach, based on the progressive smoothing of geomorphological features with the coarsening of cell size, has proved very successful in discriminating morphologically homogeneous or varied landscapes and assigning different pixel values accordingly as presented by Kienzle (2004). The 1D analysis conducted on characteristic landscape transects has resulted in a value of 163.9 m for Leitrim, 769.2 m for Tipperary North and 232.6 m for Meath. The 2D version based on 5m contour lines for all study areas obtained pixel size values of 129.6 m for Leitrim, 116.9 m for Tipperary North and 344.0 m for Meath. These results follow the behaviours already seen for the experimental methodology presented in the previous chapter, accurately distinguishing between different morphologies in line with results from Borkowski and Meier (1994) and Mackaness and Steven, (2006). Although the general geomorphological differences were detected, this approach was not precise in detecting the optimal pixel sizes observed in Chapter 4. The limiting factor might have been the approximation of relevant geomorphological features in the transect analysis and the choice of a large contour line spacing for the 2D version.

The information and complexity approach (Hengl *et al.,* 2013) focused on the DEM data content. According to this approach, the three study areas have an optimal pixel size at the 20 m resolution. The problem with this result can be due to the fact that Kolmogorov complexity is not related to the quality of information as it measures the quantity of information contained in a dataset (Allegrini *et al*., 2003) and only indirectly accounts for its meaning. Also the compressed value is independent of the spatial distribution of the information and this has major limitations in DSM. On the other hand, the relationship between decreasing entropy and the decline of information content seems to capture the flattening of the curve at 140 m in the case of Leitrim. However, in the case of Meath and Tipperary North, the trend does not flatten in the investigated interval.

A disappointing lack of results characterises the sink analysis approach. It is not possible to find a pixel size value that hydrologically satisfies the three study areas as coarsening of pixel size lower the number of sinks and pixels involved in the formation of these artefacts but at the same time increases the total size of sinks areas. This contradiction does not allow obtaining any conclusive answer. This is in contrast with results presented by Sharma *et al.,* (2011) for a large mountainous area in north-east India. The reason behind this could have some physical or hydrological explanation or simply be based on particular assumptions used by the Arc Hydro tool used in the sink analysis which is not possible to parameterise (Maidment, 2002).

Finally, the application of the fractal dimension seems insufficient to accurately estimate an optimal DEM pixel size in this case. The three study areas obtain the same value of 20 m regardless of their size, data content or stream network characteristics. The 20 m network closely matches the EPA reference DEM for the principal streams with no artefacts created but at the same time almost completely misses the secondary and minor streams. The accumulation thresholds methodology chosen for the analysis (Sharma *et al.,* 2011) may have a strong impact on the lack of results, as it fixes an arbitrary scale for the identification of streams. Sun *et al*. (2006) suggested that fractal dimension should be explored in conjunction with other classification approaches such as texture or spectral analysis. The idea first presented by Wood and Snell (1957), that relief measured over a range of sampling scales can be used to predict relief characteristics at other scales, is still valid (Lanza and Gallant, 2006). The use of fractal dimension to analyse relief still needs to be thoroughly investigated by the research community (Wood, 1996).

5.6 Conclusions

In this chapter an extensive review of the most established and frequently used empirical approaches to identify an optimal scale, in a wide range of disciplines,

have been tested and compared. Some have been proved not to have any useful role in the selection of DEM pixel size in DSM due to the lack of formal scientific principles behind them, as in the case of the ESRI ArcGIS, or failing to take into account the intrinsic characteristics of the data under scrutiny (cartographic, object orientated, sampling support and information & complexity). Others cannot be taken from their specific discipline and applied in DSM due to their limitations of applicability, as in the case of the sink analysis and the other hydrological approach. The inflection points approach based on the smoothing of terrain variability with the coarsening of pixel size has been demonstrated to provide evidence applicable in the context of DSM. This general application seems the most promising one despite the fact that at the moment it can only be used in the comparison of scale proprieties between different areas rather than in the selection of the optimal pixel size.

In summary, comprehensive scale analysis of DEMs for DSM applications has appeared to be very demanding and beyond the scope of the tested empirical approaches. These approaches, which emerged from other scientific fields to address specific scale issues, cannot be transferred to DSM. More rigorous techniques are required in DSM to explore scale processes which are complex, localised and multiscale in nature.

6 WAVELET DECOMPOSITION

6.1 Introduction

Spatial variation in soil properties and processes is the result of complex, interrelated and scale dependent factors (Lark and Webster, 1999). In DSM, the drivers of soil variation are related to these factors through a set of inference models (McBratney *et al*., 2003). These models are applied to an area assuming that the spatial scale at which the model and the inputs operate is consistent across the entire geographic space. It is assumed that no scale dependency exist in that area and also stationarity in the relationships between the soil properties and covariates expressed in the model. It is evident from the results of Chapter 4 that both these assumptions are unlikely to be met. As a highly scale dependent and non-stationary process, soil variation can be difficult to quantify and to model as the previous chapters showed. Fine resolution features can change with greater frequency or amplitude in some localised areas than in others or coarse resolution features can be recurring at much larger scales than the one analysed making it difficult to be fully captured and characterised. As soil predictions are scale dependent, in order to improve DSM models it is critical to find suitable scale relationships between environmental covariates and soil processes using a technique able to analyse variation in the frequency domain (Mendonca-Santos *et al*., 2007).

In this chapter, one-dimensional wavelet analysis will be used to examine representative landscape transects and their relationship with the results of the DSM models presented in Chapter 4. In addition, two-dimensional wavelet analysis will also be used to spatially decompose the DEMs of Leitrim, Meath and Tipperary North to then derive terrain attributes and perform DSM modelling at each level of decomposition, with RF, as already presented in the experimental methodology (Chapter 4). As previously discussed, in nature soils variation is scale dependent, and this scale dependency is not random or equally distributed in a continuum but occurs as a function of interactive soil forming processes.

Wavelet decomposition is a method by which scale dependency at specific locations can be explored (Lark and Webster, 1999). This method should elucidate scale behaviour in a more straightforward and robust manner than the empirical approaches presented in Chapter 5. Soil as a complex system is the result of interconnected parts that as a whole exhibit properties not evident from the properties of the individual parts (Ibanez and Saldana, 2008). As DSM depends on finding the right relationships between soil and environmental covariates, the tested wavelet technique should be able to decompose DEMs into different scales, offering a valuable insight into the scale of variation of these covariates.

Wavelet decomposition was developed for signal processing in geophysical explorations particularly suited to analyse non-stationary data with high fluctuations and physical processes operating at a broad range of scales. Wavelet is the tool of choice in signal processing for compression and de-noising operations, it is particularly suited in the analysis of signals characterised by a large number of scale dependent processes (Labat, 2005). The use of wavelet analysis is intended to separate the signal or “real” data information from the noise which does not exhibit any correlation (De Bartolo *et al*., 2011). The basic aim of the wavelet analysis is to determine the frequency content of a signal while measuring its spatial variation. By considering the noise levels separately at each wavelet frequency (scale), this type of analysis allows adjusting a de-noising algorithm accordingly, capturing signal variation locally at a scale that matches the local detail.

In the field of pedometrics, since the late nineties, wavelet analyse has been applied for the investigation of scale properties of soil data (McBratney, 1998; Lark and Webster, 1999). The wavelet decomposition can address the problem of spatial dependency by partitioning separate spatial components (McBratney, 1998) that can then be mapped independently to be used in DSM analysis as presented by Mendonca-Santos *et al*. (2007). Milne and Lark (2009) used wavelet analysis to determine the scale dependency in soil process models, identifying

scales at which particular models captured particular processes and thereby improving model predictive performance. They also demonstrated that particular models perform better at different points in the landscape, in effect illustrating non-stationarity in model configuration. Wavelets decompose spatial variability both in according to a particular scale and at a particular geographic location. This method is therefore ideally suited to determine whether optimising the geographic space and spatial scale at which the DSM model is formulated has consequences for DSM performance.

6.2 Materials and Methods

The Discrete Wavelet Transform (DWT) is a type of numerical analysis created for time/space frequency transformations in which a wavelet function is discretely sampled over a signal capturing at the same time frequency (scale) and location of the information. It is especially suited for the analysis of regularly sampled data as the DEM (Milne and Lark, 2009). The wavelet comprises a set of localised functions which are non-zero for only a narrow window having a compact support. Transforming a signal with a chosen basis function results in a wavelet coefficient, describing the local variation of the signal within a scale interval.

Wavelet decomposition was applied to the three test areas presented in Chapter 3. The pedological and geomorphological complexity of the three investigated areas was deemed appropriate for the purpose of the wavelet analysis.

6.2.1 One-Dimensional DWT

Representative transects of 20,000 m were extracted from the EPA 20 m DEM with a Northwest-Southwest direction (Figure 6.1).

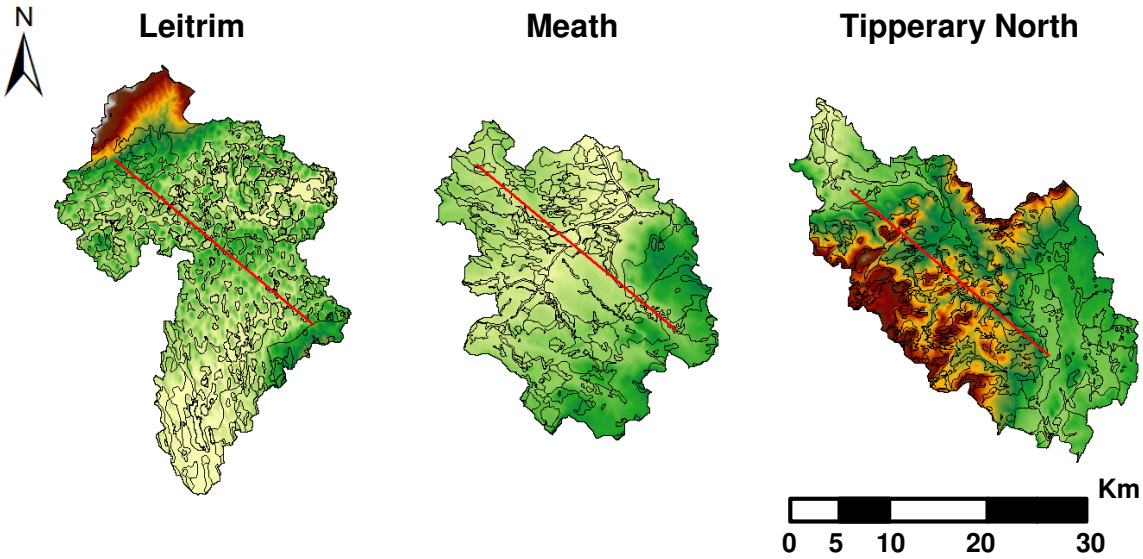


Figure 6.1 - Representative transects for the three study areas overlying the DEMs and 6 inch soil polygon boundaries.

The transects intersect the 6 inches soil maps of the study areas respectively:

- Leitrim, 11 map units (Allen, Ardrum, Ballyhaise/Corriga Complex, Drumkeeran, Garvagh, Howardstown, Mortarstown/Kinvarra Complex, Rinnagowna and Unclassified);
- Meath, 12 map units (Allen, Ashbourne, Boyne Alluvium, Drombanny, Dunboyne, Dunboyne Shaley Phase, Dunsany, Feale, Gortnamona, Patrickswell, Rathowen, Street);
- Tipperary North, 10 map units (Ballynalacken, Borrisoleigh, Borrisoleigh Steep Phase, Borrisoleigh/Knockshigowna Complex, Doonglara, Elton, Feale, Gortaclareen, Kilcommon, Knocknaskeha/Doonglara Complex).

From a geomorphological point of view the three profiles presented in Figure 6.2 exemplify the underlying landscape characteristics of the study areas captured by the DEM.

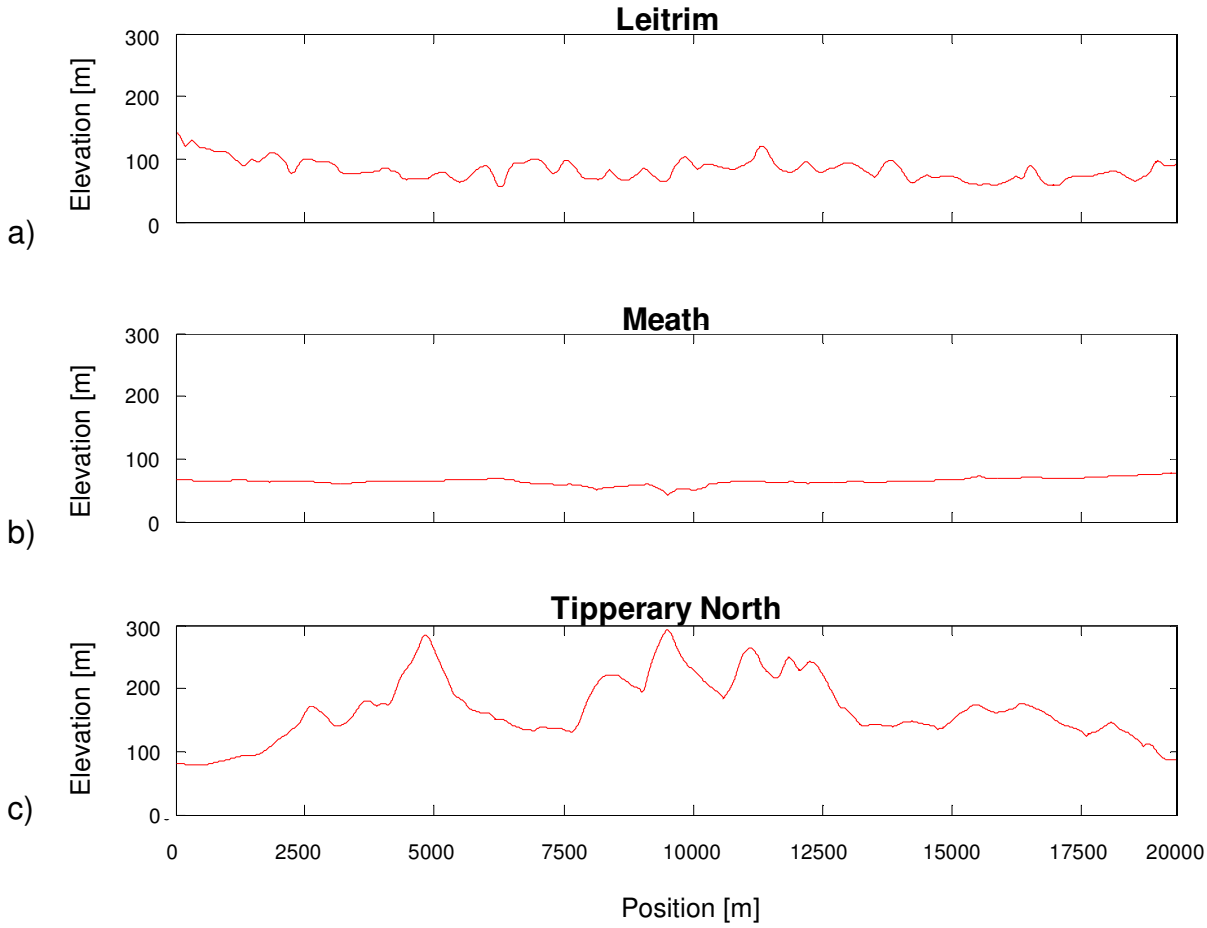


Figure 6.2 - Profiles of the one dimensional transects investigated.

Leitrim (Figure 6.2a) characterised by a drumlins landscape shows a constant number of elevation fluctuations along the transect, similar in their periods and amplitude, as the drumlins have an almost constant 100 m height and 800 m distance from crest to crest. Meath (Figure 6.2b) situated in the Meath plain is characterised by a flat landscape interrupted only by a localised fluvial erosion formation creating an almost constant profile. Tipperary North (Figure 6.2c) located in the Southern Hills, where shale plateaux had been eroded creating steep slopes, is characterised by a varied profile with peaks reaching 300 m in height and valleys falling below 100 m.

The representative transects for the three study areas were analysed with one-dimensional DWT performed employing Daubechies wavelet with six vanishing

moments (db6) using the Wavelet toolbox of Matlab software (Mathworks, 2011). Wavelet decomposition was applied to the transects at four scales corresponding to pixel resolutions of 40, 80, 160 and 320 m, creating four levels of approximation (a_1, a_2, a_3 and a_4) with associated detail (d_1, d_2, d_3 and d_4). The limit of four scales was chosen as it is the closest match to the size of the largest pixel resolution tested in Chapter 4.

A de-noising algorithm based on the thresholding method (Mallat, 1989) independently set with intervals using rigorous Stein's Unbiased Risk Estimate (SURE) thresholds (Rosas-Orea *et al*., 2005) was applied to the decomposed signal. The SURE method allows direct approximation of the mean-squared error of an estimate from the data, without requiring knowledge of the true parameter values. Therefore, instead of postulating a statistical model for the wavelet coefficients, it is possible to directly parameterise the de-noising algorithm as a sum of elementary nonlinear processes with unknown weights (Rosas-Orea *et al*., 2005). The resulting de-noised profile with associated residuals, will then be compared with the original profile and visually evaluated against the transect classification accuracy of the DSM model developed in Chapter 4 (using the EPA 20 m DEM).

6.2.2 Two-Dimensional DWT

Two-dimensional wavelet is a valuable technique for summarising and classifying sequences, functions and images. It can be used in DSM to explore environmental covariates used as inputs in predictive modelling (Mendonca-Santos *et al*., 2007). Two-dimensional wavelets are outer products of three one-dimensional wavelets. In essence, they are the result of extrapolating a matrix from the wavelet function (ψ) vector in the horizontal (6.1), diagonal (6.2) and vertical (6.3) direction and the scaling function ($\mathfrak{f}$) vector (6.4) of three one-dimensional wavelets (Daubechies, 1992), according to:

$$\psi^H(x,y) = \psi(\mathrm{x})\, \mathtt{f}(\mathrm{y}) \quad (6.1)$$

$$\psi^D(x,y) = \psi(\mathrm{x})\, \psi(\mathrm{y}) \quad (6.2)$$

$$\psi^V(x,y) = \mathtt{f}(\mathrm{x})\, \psi(\mathrm{y}) \quad (6.3)$$

$$\mathtt{f}(x,y) = \mathtt{f}(\mathrm{x})\, \mathtt{f}(\mathrm{y}) \quad (6.4)$$

The EPA 20 m DEM was analysed with two-dimensional discrete wavelet transform using the Wavelet toolbox of Matlab software (Mathworks, 2011). The procedure creates at each level an approximation (a) with associated horizontal, diagonal and vertical detail (d_H, d_D and d_V). The original image can be reconstructed by combining the approximation and the detail (Original = $a_1 + d_1$). Each following level is the result of adding the subsequent approximation and detail ($L_1 = a_2 + d_2$, $L_2 = a_3 + d_3$, $L_3 = a_4 + d_4$ and so on). The four resulting decomposition levels (L_1, L_2, L_3 and L_4) were used to derive eleven terrain attributes (slope gradient, aspect, curvature, plan curvature, profile curvature, slope height, valley depth, normalized height, standardized height, mid-slope position and convergence index) and develop DSM models at different scales (altering pixel and window sizes) using the methodology presented in Chapter 4. These results will be then compared with the classification accuracy of the DSM model applied to the original EPA 20 m DEM.

6.3 Results

The results are presented in nine figures (Figure 6.3 - 6.11) and one table (Table 4.1). Figure 6.3 illustrates the wavelet 1D decomposition at four levels of approximation with associated detail and detail coefficients for the representative profile in Leitrim, while Figure 6.4 and Figure 6.5 show the wavelet 1D

decomposition results for Meath and Tipperary North respectively. Figure 6.6 summarises the results of the de-noising operation (original profile, de-noised signal and residuals) matching them with the DSM model results for the same transect in Leitrim, while Figure 6.7 and Figure 6.8 summarises the same de-noising operation respectively for Meath and Tipperary North. Figure 6.9 displays the 2D wavelet decomposition of the DEM of Leitrim at four levels of approximation with associated detail component, while Figure 6.10 and Figure 6.11 display the 2D wavelet decomposition for Meath and Tipperary North respectively. Finally, Table 6.1 presents the classification accuracy of the DSM model for the three investigated areas using the spatially decomposed DEMs.

6.3.1 One-Dimensional DWT

Decomposition

The effect of decomposing the elevation profile of Leitrim with the one-dimensional DWT technique is presented in Figure 6.3 where it is possible to note that the first decomposition (a_1) at 40 m did not alter the profile significantly, while the detail (d_1) shows only one area at approximately 6,000 m, corresponding to two drumlins very close to each other creating a steep valley, having a moderate level of noise. The second decomposition (a_2) shows lower level of noise but more frequently and associated with the fluctuations in elevation related to the position of the drumlins. The third decomposition (a_3) shows a similar alternating pattern of the noise but at greater magnitude, with four areas achieving high values (± 2) at approximately 2,500, 6,000, 11,000 and 16,000 m. Finally, the fourth decomposition (a_4) shows a similar alternating pattern of the noise as a_3 but at greater magnitude, with four areas achieving high values (± 5) at approximately 0, 2,500, 6,000 and 10,000 m.

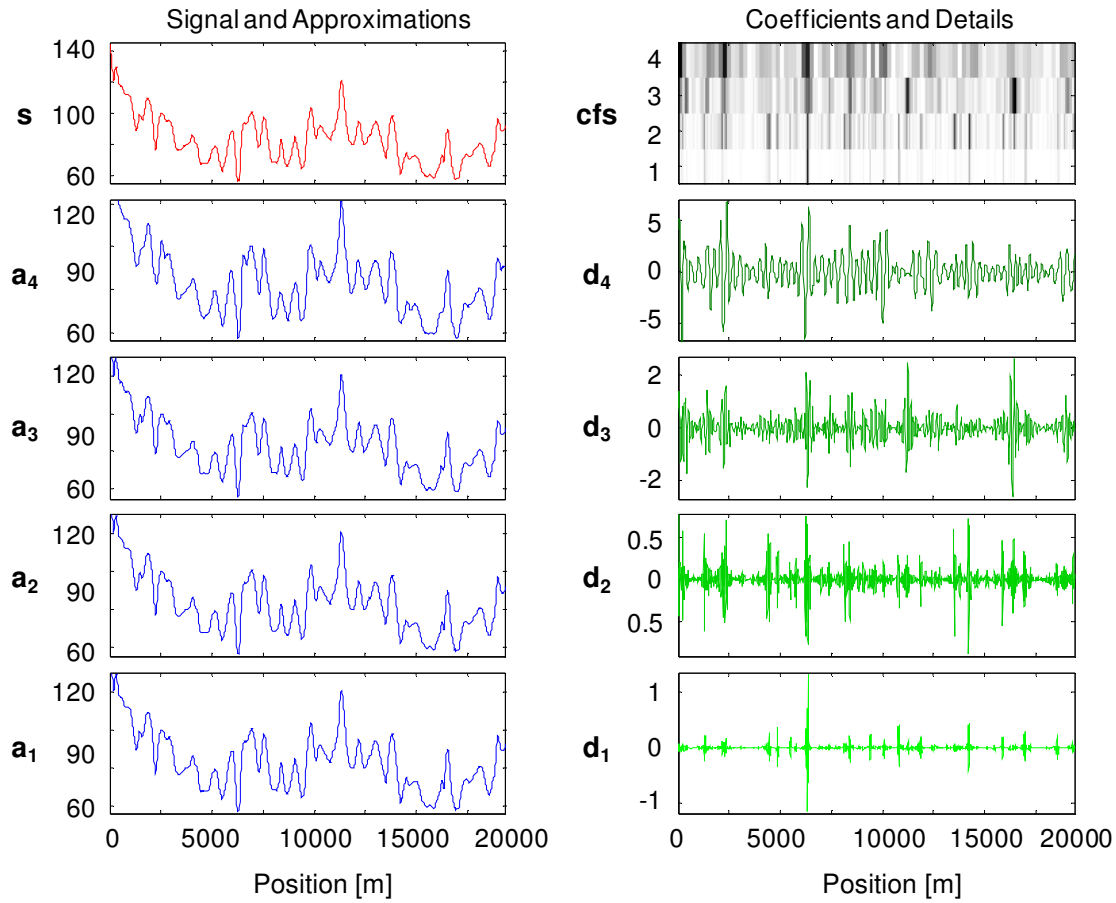


Figure 6.3 - 1D wavelet decomposition of Leitrim representative transect (s) at four levels of approximation (a_1, a_2, a_3 and a_4) with associated detail (d_1, d_2, d_3 and d_4) and details coefficients (cfs).

The final decomposition at level 4 (equivalent to 320 m pixel size) alters considerably the profile over-approximating the height of the drumlins at 0, 2,500, 6,000, 11,000 and 16,000 m by more than 5 m. To summarise the detail significance, in the top right corner of the figure, the detail coefficients are presented divided by scale (1, 2, 3 and 4) with darker values indicative of higher values of the coefficients. In general, the most relevant noise events appear on the transect at scale 4 at 0, 2,500 and 6,000 m and at scale 3 at 16,000 and 11,000 m.

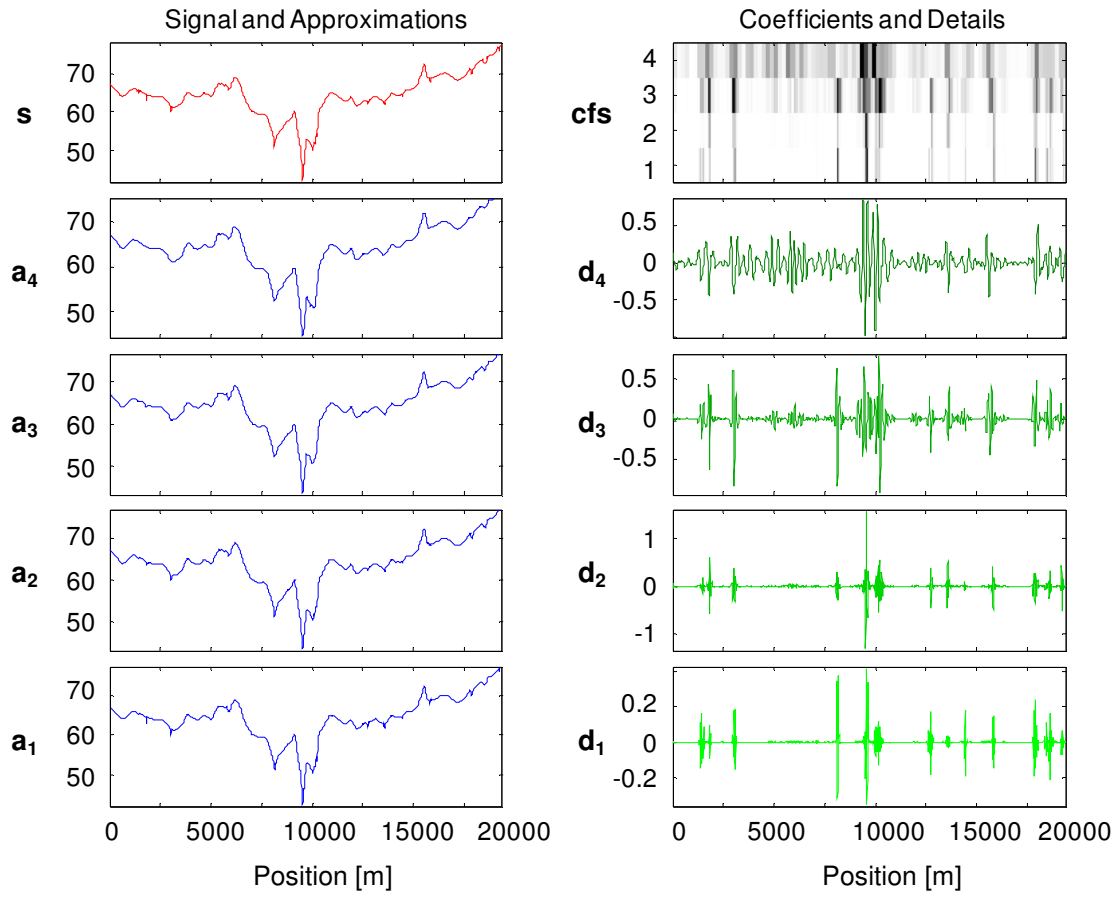


Figure 6.4 - 1D wavelet decomposition of Meath representative transect (s) at four levels of approximation (a_1, a_2, a_3 and a_4) with associated detail (d_1, d_2, d_3 and d_4) and details coefficients (cfs).

The one-dimensional wavelet decomposition for Meath (Figure 6.4) shows, as expected, very little noise at the four levels tested. The profile appears less homogeneous and more fragmented that what actually is in reality. This is due to the height exaggeration that makes the river channel, cutting the profile in the middle of the image, seem a prominent feature in the landscape. The difference in elevation between the river Boyne (42 m), its embankment (45 m) and the highest point in the floodplain (50 m) is lost over the transect length. In terms of noise, from the detail coefficients summary box it is possible to see two areas with high detail coefficients in the middle of the transect at approximately 10,000 m, where the profile intersects the river Boyne. At both, scale 3 and scale 4 this section of the profile shows the highest detail coefficients. It is worth mentioning

that the colour palette is adapted to each transect and set with the highest (black) and lowest (white) values obtained during the transect decomposition analysis.

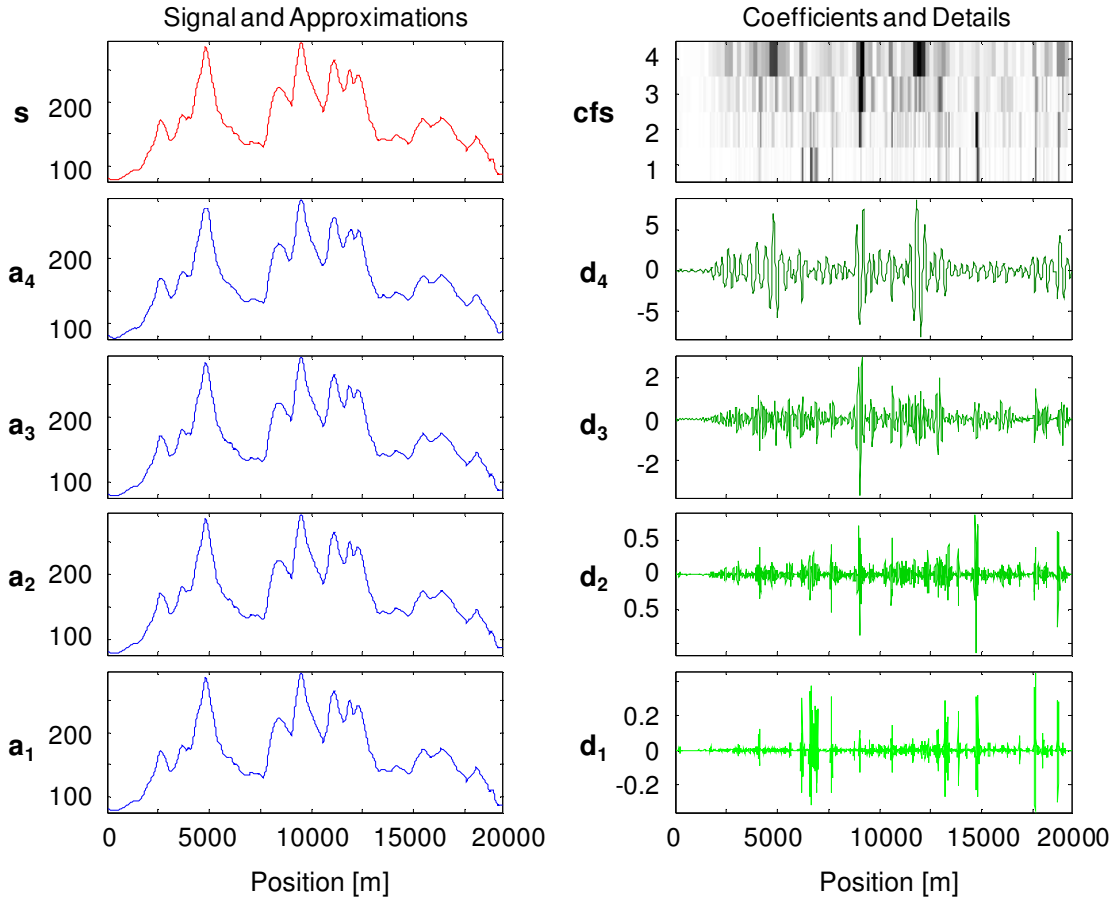


Figure 6.5 - 1D wavelet decomposition of Tipperary North representative transect (s) at four levels of approximation (a_1, a_2, a_3 and a_4) with associated detail (d_1, d_2, d_3 and d_4) and details coefficients (cfs).

The result of decomposing the transect of Tipperary North with the one-dimensional DWT technique is presented in Figure 6.5, where it shows very little noise at the first scale tested, obtaining low values (±0. 2) of detail. The second decomposition shows an increase in noise at about 9,000, 15,000 and 19,000 m. The third decomposition presents a marked increase (±2) at about 9,000 m and also a general increase of one area of the transect between 11,000 and 13,000 m. The final level of decomposition, scale 4, shows four marked areas of noise at 5,000, 9,000, 12,000 and 19,000 m. In conclusion, from the detail coefficients

summary box, it is possible to note 5 areas obtaining high values of detail: 15,000 m (scale 2), 9,000 m (scale 3) and 5,000, 9,000 and 12,000 m (scale 4).

De-noising

The comparison shown in Figure 6.6 was employed to visually compare the de-noised signal of Leitrim with the DSM model results obtained by RF for the EPA 20 m DEM (Figure 6.6a), as presented in Chapter 4. The classification accuracy of the DSM model ranges from 0 to 100% and has been classified into five groups (0-20%, 20-40%, 40-60%, 60-80% and 80-100%) to facilitate its interpretation.

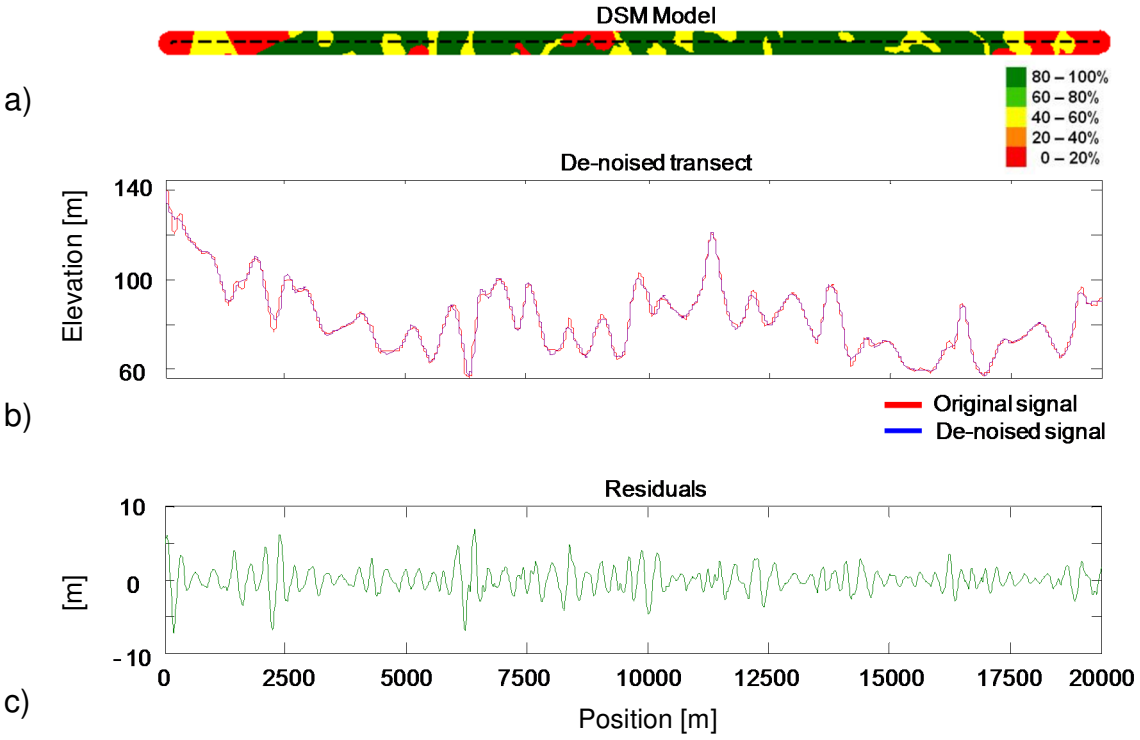


Figure 6.6 - Results of the de-noising operation on Leitrim representative transect: a) classification accuracy of the DSM model (Chapter 4) for the transect area; b) de-noised signal compared with the original profile and c) residuals of the noise removal process.

Leitrim displays three areas of poor classification: 0-2,500 m, 8,000-9,000 m and 17,500-20,000 m. The first area 0-2,500 m corresponds to an area detected by the 1D DWT with high noise values and partially reduced by the de-nosing

algorithm, as it is visible on Figure 6.6b, by removing a small peak at the beginning of the transect and raising the height of a valley at 2,500 m. For the remaining two areas of poor classification accuracy, two adjustments were made by the algorithm at 8,500 and 9,000 m reducing the height of the corresponding peaks and at approximately 19,000 m another peak was reduced in height.

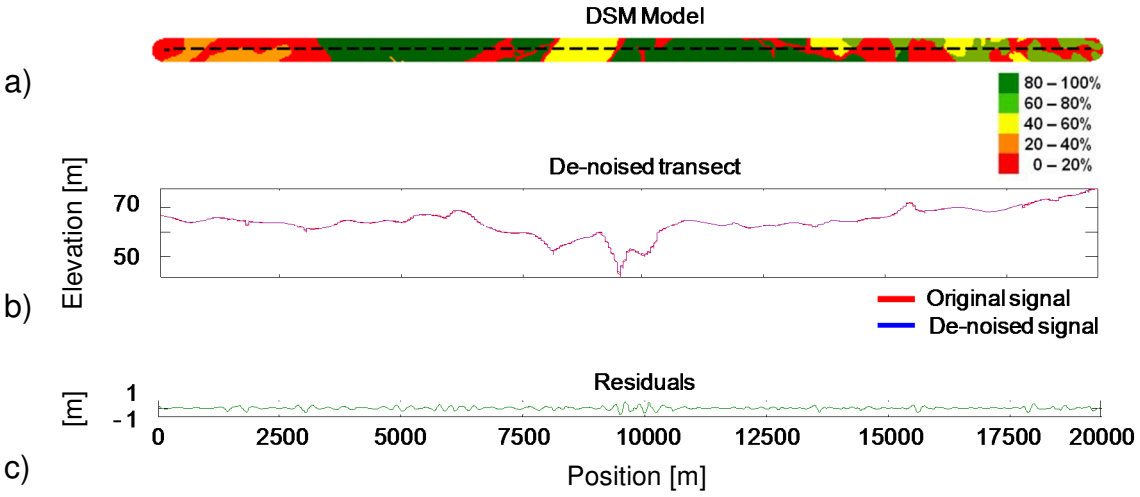


Figure 6.7 - Results of the de-noising operation on Meath representative transect: a) classification accuracy of the DSM model (Chapter 4) for the transect area; b) de-noised signal compared with the original profile and c) residuals of the noise removal process.

For Meath, Figure 6.7 shows very little change between the original profile and the de-noised one as previously shown in the 1D DWT decomposition where little noise detected on this transect in comparison with the other two investigated areas. Visually, it is very difficult to assess if the de-noising algorithm has changed the profile in a significant way as the two lines appear very close to each other. This is due to the fact that in order to facilitate the evaluation of findings, all the figures (Figure 6.6, Figure 6.7 and Figure 6.8) have been created with the same height scale to make the comparison with the DSM model results more realistic.

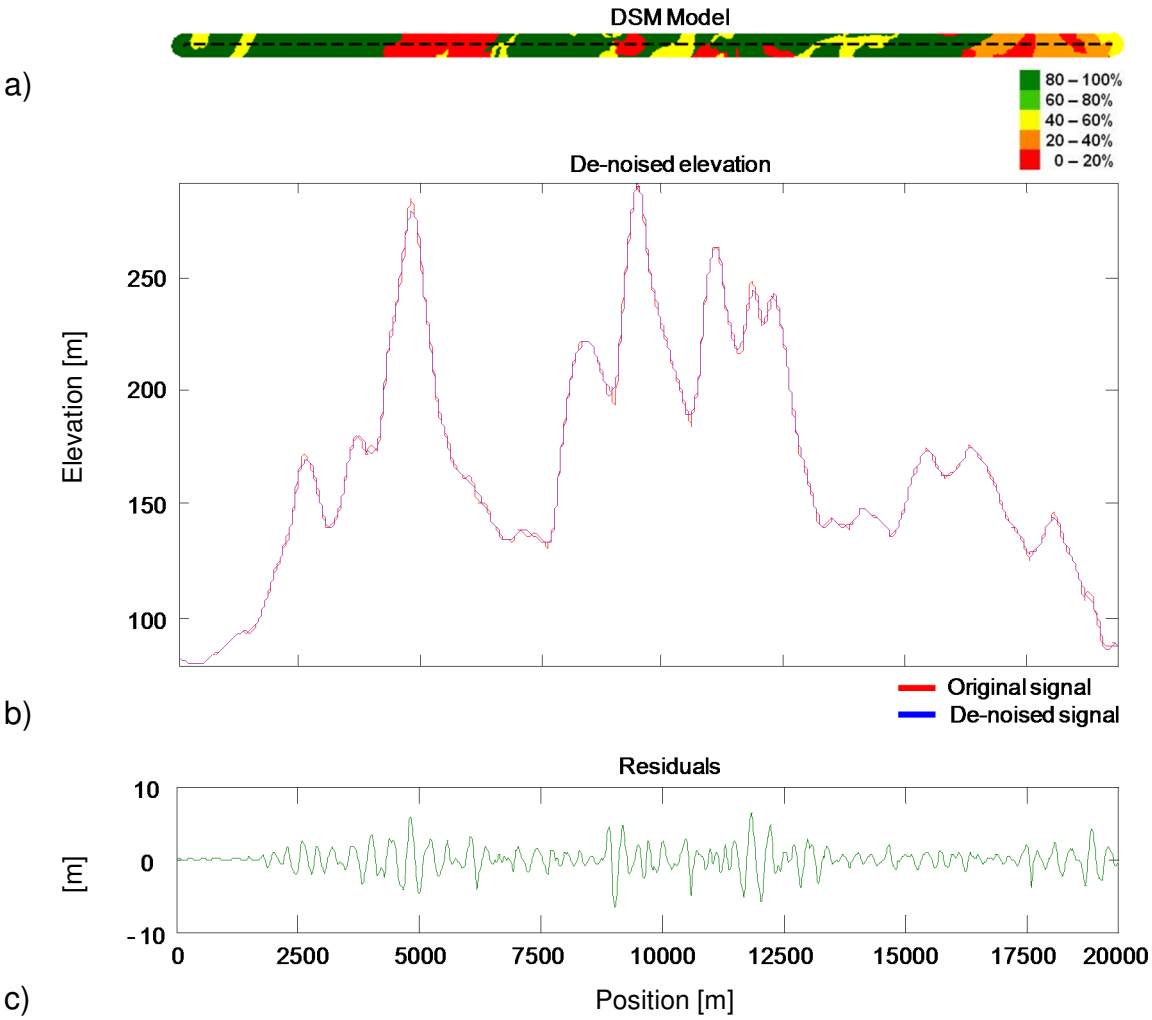


Figure 6.8 - Results of the de-noising operation on Tipperary North representative transect: a) classification accuracy of the DSM model (Chapter 4) for the transect area; b) de-noised signal compared with the original profile and c) residuals of the noise removal process.

The comparison between the results of the DSM model with the de-noising operation for Tipperary North is presented in Figure 6.8. The DSM model shows four areas of poor classification: 4,000-6,000 m, 9,000 m, 12,000 m and 16,000-20,000 m. The first area 4,000-6,000 m corresponds to a region on the transect previously detected by the 1D DWT with high noise values in which the thresholding algorithm has reduced the height of two peaks by more than 5 m reconstructing the profile from the four approximations and the processed details. Also, the peaks on the profile at 9,000, 12,000 and 12,500 m were modified by

the de-noising operation as well as raising the height of the valleys at 9,000, 11,000 and 12,000 m.

6.3.2 Two-Dimensional DWT

As previously presented for the 1D DWT analysis, a visual assessment is effective in the examination of large profile changes, such as peaks, drumlins or valley bottoms but has limitations in the detection of minor alterations from the de-noising algorithm on the profile in the case of low relief transects. A more comprehensive way of analysing scale relationships in DSM is to extend the analysis to a full spatial decomposition performed with the 2D DWT. Using the spatially decomposed DEMs as input for the DSM models, developed in Chapter 4, will allow a detailed comparison of classification accuracies.

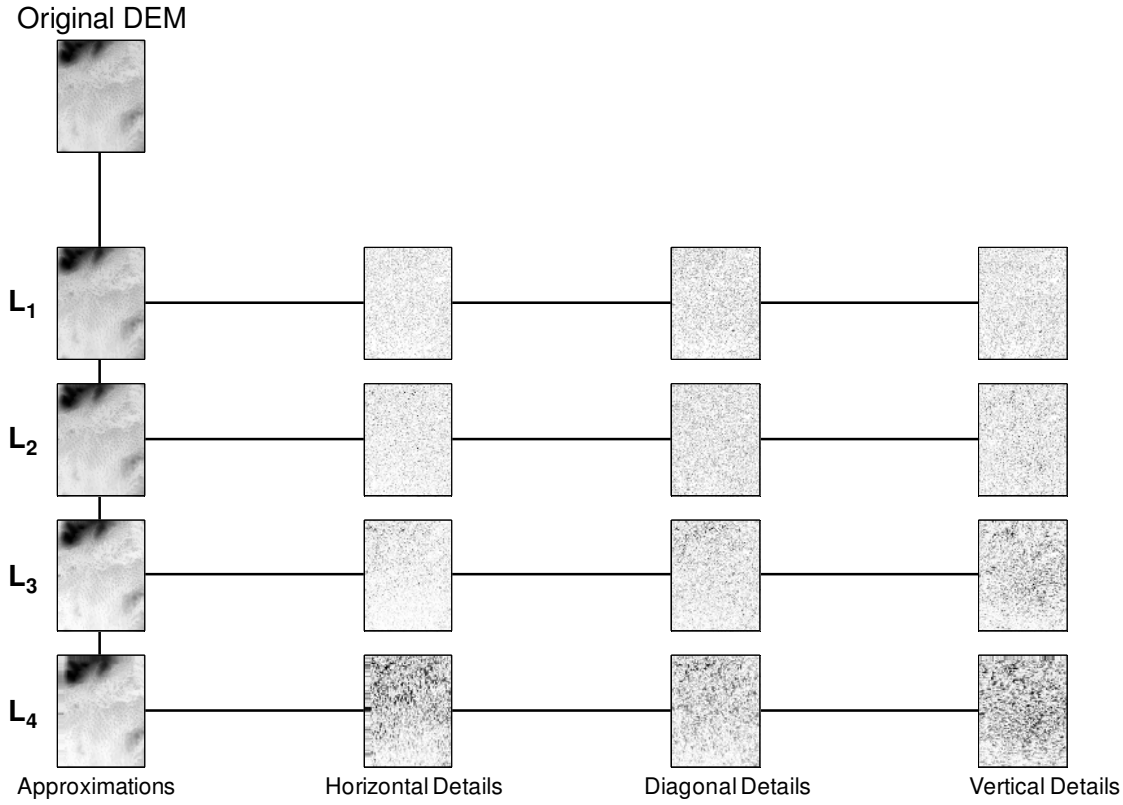


Figure 6.9 - Wavelet decomposition of the DEM of Leitrim at four levels of approximation with associated horizontal, diagonal and vertical components.

Figure 6.9 shows the two dimensional decomposition performed for Leitrim at four levels L_1, L_2, L_3 and L_4 generating four approximations with associated horizontal, diagonal and vertical details. The first and second level of decomposition did not alter the DEM significantly as already observed with the 1D DWT decomposition of the transect in Leitrim. The third level of decomposition presents a slight increase in the vertical detail as does the fourth one, mainly in the northwest and southeast corners of the DEM. These two areas denote the shift between the drumlin belt and area of higher relief (250 m in the northwest and 120 m in the southeast).

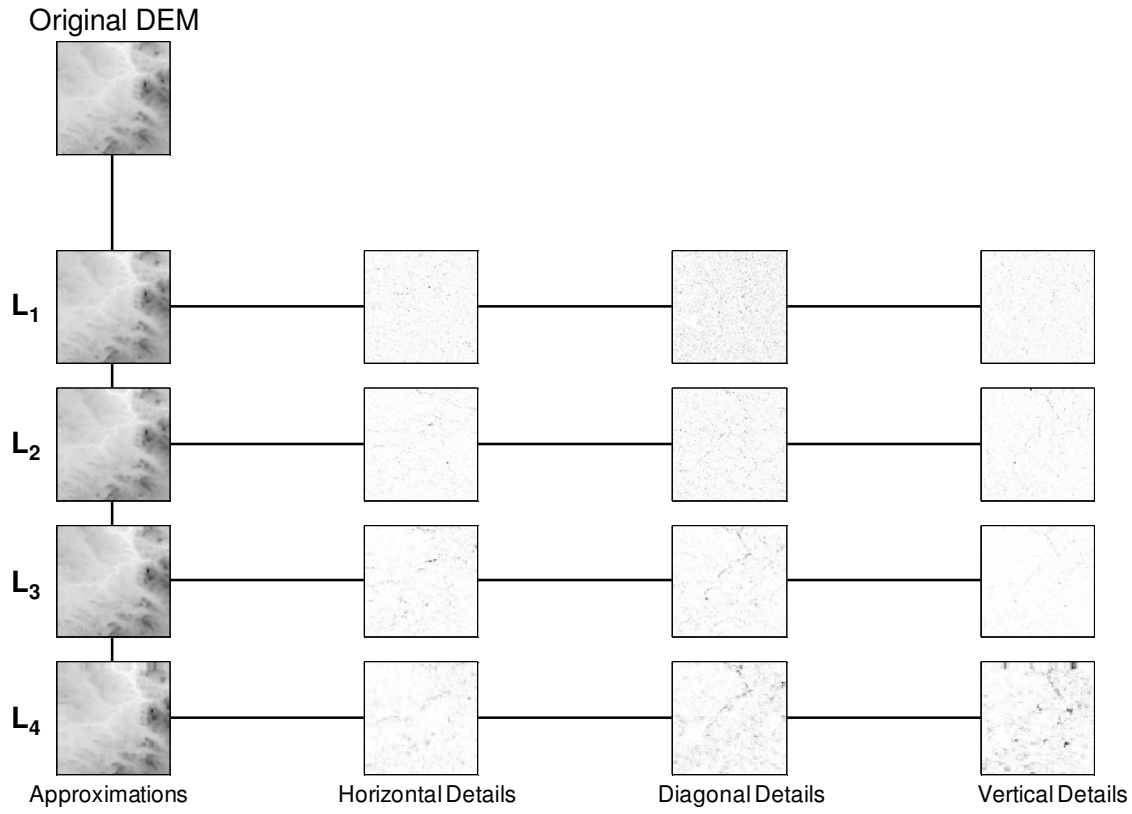


Figure 6.10 - Wavelet decomposition of the DEM of Meath at four levels of approximation with associated horizontal, diagonal and vertical components.

As previously discussed, the homogeneous DEM of Meath does not consent a visual assessment of the decomposition operation as the spatial variation of elevation is limited in a flat landscape with little abrupt differences in height.

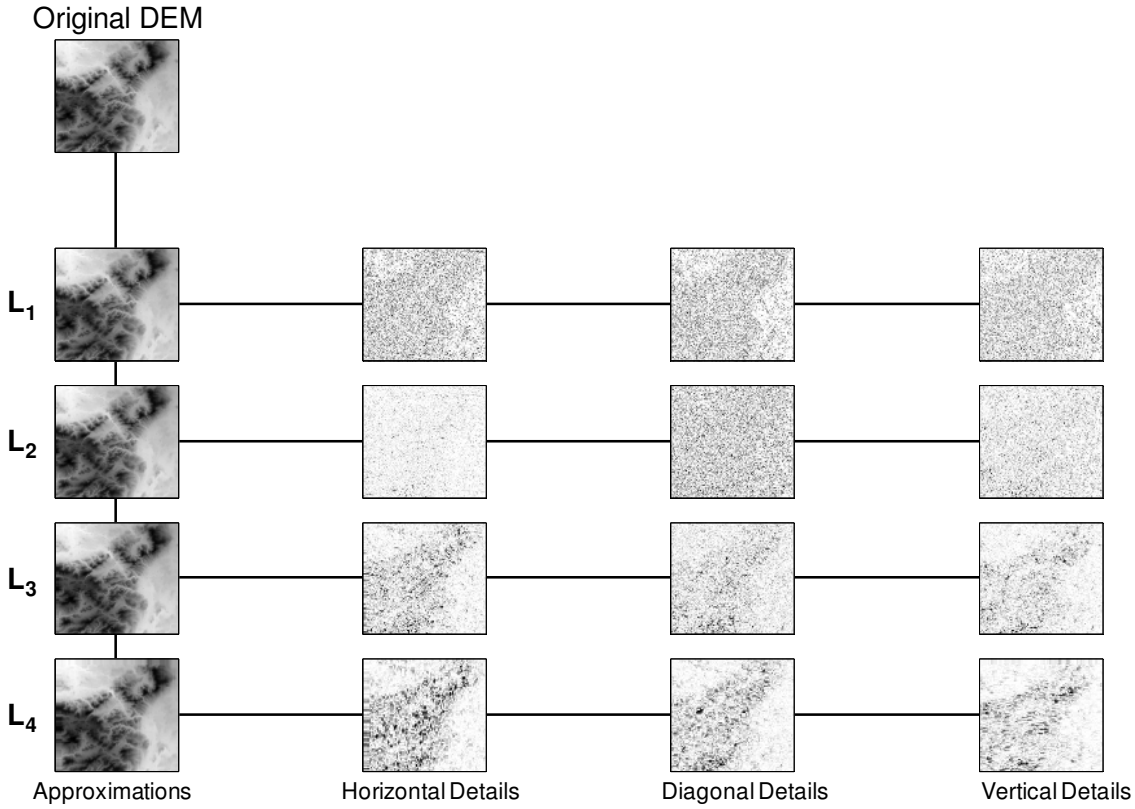


Figure 6.11 - Wavelet decomposition of the DEM of Tipperary North at four levels of approximation with associated horizontal, diagonal and vertical components.

The 2D DWT decomposition for Tipperary North (Figure 6.11) shows a more interesting pattern than the other two areas for L_1, as the associated horizontal diagonal and vertical detail components appear to capture a fair amount of fine resolution noise. The horizontal component of L_2 seems to have a low level of noise in comparison to the diagonal and vertical details. It is possible to observe a more structured level of detail in L_3 as the three directional details have higher values for the Southern Hills slopes. This area of high relief crosses the investigated area from southwest to northeast leaving two parallel areas of low relief in the opposite corners. This trend continues in L_4 where a large area of high relief seems to have even higher values of detail in comparison with the plain, also some of the prominent peaks obtain the highest values of detail.

In Table 6.1 the results of the DSM model performed with a classification tree (RF) are presented for the original EPA 20 m DEM and the four levels of decomposition (L_1, L_2, L_3 and L_4).

Table 6.1 - Classification accuracy of the DSM model for the three study areas using the spatially decomposed DEMs.

	Leitrim [%]	**Meath [%]**	**Tipperary North [%]**
Original DEM	56.9	38.8	44.6
L_1	58.1	39.8	44.7
L_2	56.5	41.5	44.9
L_3	56.1	43.9	46.0
L_4	57.9	46.3	45.8

The classification accuracy in Leitrim increases from 56.9% of the original DEM to 58.1% at L_1 to drop back to the previous value for L_2 at 56.5% just below the original value, remaining at a comparable low value for L_3 (56.1%) and increasing again for L_4 (57.9%). Meath with the lowest initial value of classification accuracy for the original DEM at 38.8% increases slightly at L_2 (39.8%) and again at L_3 (41.5%) to then remarkably rise to 43.9% at L_3 and at 46.3% at L_4, representing an improvement in accuracy by almost a fifth. Tipperary North shows no significant variation at L_1 (44.7%) and L_2 (44.9%) with only a slight rise at L_3 (46.0%) and L_4 (45.8%).

6.4 Discussion

Wavelet decomposition performed with the 2D DWT has been proved effective in dealing with elevation information, in the form of DEMs. The DWT appears a powerful technique in soil science to quantify signal changes from one scale to

another through the dilations and translations of a wavelet function as suggested by McBratney *et al.* (2003) and Lark (2005).

The one-dimensional technique applied to the three representative profiles was tested in very different pedological and geomorphological landscapes, such as the drumlin belt in Leitrim, the great plain of Meath in Meath and the Southern Hills in Tipperary North. It has shown to be particularly suited to areas with variable landscape like Leitrim with the periodic fluctuations typical of the drumlins or Tipperary North where the alternating peaks and valleys create a more fragmented signal with rapid changes in elevation. The profile of Meath was difficult to assess as the lack of abrupt changes in height made it difficult to visually appreciate the effects of the 1D DWT decomposition.

The aim of the second phase of the 1D DWT experiment was to assess if the different level of approximation with associated detail were in some way related to areas of poor performance of the DSM model presented in Chapter 4. The transect of Leitrim was particularly interesting as the first 2,500 m performed poorly in terms of classification accuracy of the DSM model as well as having high detail coefficients at L_4, suggesting a correspondence between the noise in the signal and the lack of predictive power of the model. The same happened for Tipperary North where the height of five peaks was reduced by the de-noising algorithm and three valleys were raised. The Meath profile has proved very difficult to visually asses as the lack of major morphological changes makes any variation undetectable.

The spatial decomposition with the 2D DWT seems to select only a small fraction of the total variance present in the DEM while still maintaining the general structure of the spatial variation, in line with results of Lark and Webster (2004). During the decomposition at the first level, the size of the wavelet used in the analysis is relatively small, offering accurate location resolution of fine scale phenomena, while at the increase of the wavelet size, larger and larger processes can be captured leading to coarse scale phenomena. The comparison of the DSM model trained using terrain attributes created with the decomposed

approximations of the DEM, shows a large improvement in the value of classification accuracy for Meath, a rise of 7% at the increase of the wavelet size to L_4. Also a minor increase of 1% was observed for Leitrim at L_1 and L_4 and for Tipperary North at L_3 and L_4 compared with the original EPA 20 m DEM. This seems to suggest that even a small reduction in variance, as in the case of Meath, can have a considerable impact on the classification accuracy of DSM models supporting the results that Mendonca-Santos *et al*. (2007) obtained for a flat area in NSW Australia. Removing redundant or artifactual information from the DEM 2D DWT seems to improve the way in which DSM models link topography with soil variation. The source of this variation is uncertain, it could be due to artefacts introduced during the DEM creation with the ANUDEM software by spline interpolation of contour lines (Hutchinson, 2007). Alternatively it could be caused by uncertainties in the original height information used for the creation of the contour lines. Oksanen and Sarjakoski (2006) suggested that DEM errors appear to be caused by spatial variation in different frequency classes: low-frequency errors (systematic errors in contour data) and high-frequency errors (noise between the DEM and the real terrain). In addition, the uncertainty could simply be resulting from redundant information not useful in identifying soil-terrain relationships exploited by the DSM model. It is worth mentioning that the results of the experimental methodology (Chapter 4) showed better classification accuracies at the optimum resolution for the three areas.

2D DWT decomposition is a robust tool to study the effect of scale on DEMs for DSM applications. This technique seems to avoid the problem of decrease in information content and of introducing artefacts due to changes in grid resolution, offering the possibility to create a DEM more suited in the creation of terrain attributes so important in DSM modelling. This use of the 2D DWT could offer an innovative way to gain new information from DEMs contributing to better predictions in DSM modelling. As Biswas *et al*. (2013) suggested, once the dominant scale has been identified using 2D DWT, the information could be used for scale-specific prediction of soil properties.

One limiting factor could be the restrictions on the scale size selection as this is implicitly imposed by the initial value as each subsequent level is its double. In light of the fast development of remote and proximal sensing technologies offering vast quantities of fine resolution data, spatial decomposition with wavelet could be extended to other environmental covariates used regularly in DSM such as climatic properties, land cover, land use or other soil properties.

6.5 Conclusions

In conclusion, 2D wavelet analysis has shown that by spatially decomposing a DEM it is possible to remove specific sources of variation, which might be unnecessary for DSM analysis, improving classification accuracy. The results obtained for the low relief homogeneous area seem to suggest that for this specific type of landscape wavelet decomposition could enhance the classification accuracy of DSM models used for soil taxonomic units. Although it improved classification accuracy in comparison with the original EPA 20 m DEM, the experimental methodology (Chapter 4) showed better classification accuracies at the optimal scale for the three study areas. The real contribution of wavelet decomposition was its ability to extract the relevant spectral scales for each area. However, this still leaves unresolved the issue of incoherent scale response observed for the three study areas. This issue might be better solved with a methodology capable of coping with multiple scales, confirming the assumption that an independent stratification approach is needed to appropriately take soil spatial variation into account.

7 MULTISCALE METHODOLOGY

7.1 Introduction

The soil forming processes that influence pedogenesis are composed of different nested features interacting with each other at various locations, with distinctive intensities and across multiple scales. These statistical relationships between soil taxonomic units or properties and environmental covariates are at the centre of DSM modelling (Lark, 2006). Corstanje *et al*. (2008a) used a nested analysis to determine whether a particular model form best represents soil processes at particular scales. It is clear from this work that scale at which a soil landscape model is formulated has consequences for the model performance and that the assumption that a single DSM model configuration across a unitary geographic space is one that needs exploring in more detail.

As previously discussed a multiscale methodology seems particularly suited for the intricate organisation of soil formation, especially in the context of DSM analysis, where this approach could contribute to enhance the modelling of soil spatial variability. Lagacherie (2008) reviewing the proceedings of the first international workshop on DSM, suggested that more work was needed to develop functions able to deal with the multiscale variations observed in soils. Behrens *et al*. (2010b) suggested that attention to techniques able to address scale issues in DSM is still limited. The experimental results presented in Chapter 4 showed that two main patterns of scale behaviour existed for the tested areas, the incoherent scale response across the areas suggested that further subdivisions were needed.

In this chapter a multiscale methodology based on geostatistics and spatial clustering will be used to examine three study areas with distinctive geomorphologies and soil types. By spatially characterising local statistics through moving window variograms, a segmentation of the DEMs will be implemented with k-means clustering. Each area will then undergo DSM modelling with RF as already presented in Chapter 4. A final comparison between

these results and the ones obtained without segmenting the DEM will be made and discussed.

7.2 Materials and Methods

As previously discussed in Chapter 3, the variogram is a central concept in geostatistics as it is used to analyse the structure of spatial variation in data. The global variogram represents the overall variation in elevation of the DEM, while local variograms characterise deviations from this. In the global approach a variogram is produced including in the calculation all the cells of the DEM, as previously discussed. The produced values at a specific location are dependent on all the values in the DEM. If the assumption of stationarity is not satisfied it should be possible to observe differences in the properties of the variogram caused by local sources of variation. If this is the case a more local approach in the calculation of the variogram should be able to detect these differences. A technique offering the possibility to calculate these local changes is the moving window (Haas, 1990).

A moving window variogram approach was used to classify spatial variation and develop a multiscale methodology for DSM analysis. Local variograms were used to describe spatial patterns and structures of the DEMs of the three investigated areas. The analysis was carried out using the EPA 20 m DEM for the three test areas presented in Chapter 3. All the variograms were created with R software (R Development Core Team, 2011) using the gstat package (Pebesma and Wesseling, 1998). K-means spatial clustering was performed in R with the package stats. The DSM models, created for each spatial cluster, were developed using RF as previously presented in Chapter 4.

The multiscale methodology presented in this chapter is composed of three sections:

- moving window variograms to compute local statistics;

- clustering of the computed parameters to segment the DEM;
- multiscale DSM modelling.

7.2.1 Geostatistics

A variogram is commonly presented as a graph showing a mathematical model describing the variability of data in relation to distance. The semivariance γ (h) is calculated as half the variance of the increments with lag $n(h)$ of the number of paired data from location x_i ($z(x_i)$ and $z(x_i + h)$), as stated:

$$\gamma(h) = \frac{1}{2}\frac{1}{n(h)}\sum_{i=1}^{n(h)}(z(x_i) - z(x_i + h))^2 \quad (7.1)$$

From the formula, it is possible to deduce that pairs of measurements with smaller values of h, in other words that are close to each other, have smaller variance as compared to measurements which are far apart. The variance gradually increases till the distance of separation reaches a value the range (a) beyond which the variance levels out and becomes independent of the distance. The maximum variance reached at that point is called sill ($c=c_0+c_1$), it is obtained by adding the variance realised with a hypothetical distance of 0 called the nugget (c_0) to the partial sill (c_1) that is the variance of the spatially structured component.

Different mathematical functions (models) can be used to fit to the experimental semi-variance values. The variogram model used in this research is the spherical one, as according to Nanos and Rodriguez (2012) it has been proved in modelling practice as the most convenient for multiscale variation. The spherical model is an adapted quadratic function for which at some distance *a* (range), pairs of points will no longer be autocorrelated and the variogram reaches an asymptote, according to:

$$\gamma(h) = \begin{cases} c_0 + c & for\ h\ > a \\ c_0 + c\left[1.5\left(\frac{h}{a}\right) - 0.5\left(\frac{h}{a}\right)^3\right] & for\ 0 < h \leq a \\ 0 & otherwise \end{cases} \quad (7.2)$$

Variograms in this study were computed using the REML method (Marchant and Lark, 2007a) as it gives a better representation of the underlying variation in comparison to the classical method of moments when used with data on a regular grid like the DEM in this case (Lark and Cullis, 2004).

In geostatistics soil properties are treated as the realizations of a regionalised random function, implicitly assuming a certain degree of stationarity (Webster, 2000). By assuming that the underlying stochastic process is stationary, the joint probability distribution of the random function is assumed independent from its geographic location (Corstanje *et al*., 2008b). In other words, it does not change over space but is the same for all the soil samples over an entire survey area. By computing a variogram selecting only a small number of cells of the DEM, it should be possible to obtain a local estimate of the variogram parameters for that specific neighbourhood.

7.2.2 Moving window variograms

A moving window technique to compute a variogram is, in essence, a predetermined mask centred on a specific cell of the DEM and considering in its calculations only the cells included in that particular neighbourhood. The window will then move to the adjacent cell, computing another variogram and so on (Fotheringham *et.al*,1996). In this research a square of 500 m was regarded as a reasonable size, large enough to adequately capture spatial variability and small enough to be representative of the local area. The moving window contained 625 cells in the local neighbourhood that is a sufficient number to accurately compute the variogram.

Considering a cell with coordinates *i* , *j* and a square 3 x 3 mask, a moving window technique examines in its calculation all the cells included between the cell in the preceding column and row (*i*-1 , *j*-1) and the cell in the following column and row (*i*+1, *j*+1), as shown in Figure 7.1.

i-1, *j*-1	*i*-1, *j*	*i*-1, *j*+1	*i*-1, *j*+2
i, *j*-1	*i*, *j*	*i*, *j*+1	*i*, *j*+2
i+1, *j*-1	*i*+1, *j*	*i*+1, *j*+1	*i*+1, *j*+2
i+2, *j*-1	*i*+2, *j*	*i*+2, *j*+1	*i*+2, *j*+2

Figure 7.1 - Moving window neighbourhood for a square 3 x 3 window centred on a cell with *i,j* coordinates (other shapes are also possible, such as circles or diamonds).

For every movement of the window a set of new variogram parameters is created, as these estimates have coordinates, it is possible to map them displaying how they change over space. By using a moving window method to locally estimate variograms, it is possible to define local statistics in terms of homogeneity of the data variation, including:

- a local distance parameter (*a*), which represents the maximum lag over which the random function is autocorrelated;
- local variance (*v*), the sum c_0+c_1;
- the spatial dependence ratio (*s*) calculated as the proportion $c_1/(c_0+c_1)$.

7.2.3 Multiscale segmentation

Clustering is a type of analysis used to group data into objects in such a way that data in one group are more similar to each other than to data in another group. The clustering method used in this research is k-means clustering (Hartigan, 1975). K-means is an unsupervised method intended for minimising the mean squared distance between the objects and their nearest centroid, considered as the multivariate means of the clusters (Brus *et al.*, 2006). The algorithm assumes that the data form a vector space and tries to find clustering around centroids $m_i \forall_i = 1 \ldots k$ which are obtained by minimizing the object, as follows:

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_i - \mu_i)^2 \tag{7.3}$$

where there are k clusters $S_i, i = 1,2,\ldots k$ and m_i is the centroid of all the points $x_j \vee S_i$. To assess accuracy v-fold cross validation was performed.

In unsupervised cluster analysis, one of the major challenges is to estimate the number of clusters. A technique used for assessing this is v-fold cross validation (Nisbet *et al.*, 2009). It involves partitioning a sample of data into complementary subsets, performing the analysis on one subset and validating the analysis on the other subset. In order to reduce variability, multiple rounds are performed using different partitions, and the validation results are averaged over the rounds.

By grouping the results obtained with the moving window variogram with k-means clustering, it should be possible to segment the DEM into areas characterised by similar local statistics in which the stationarity assumption is valid. Each area will then undergo DSM modelling with RF as presented in Chapter 4 and finally a comparison of results with the ones previously obtained without segmenting the DEM.

7.3 Results

The results are presented in five figures (Figure 7.2 - 7.6) and two tables (Table 7.1 and Table 7.2). Figure 7.2 illustrates the spherical variograms for the three study areas. The distribution of local statistics parameters, for the investigated areas, calculated with the moving window is presented in Figure 7.3, Figure 7.4 and Figure 7.5 showing the local distance parameter, variance and spatial dependence ratio respectively. Table 7.1 presents local statistics of the clustered areas for the investigated areas. Figure 7.6 displays the DEM segmented using k-means clustering of the local variogram parameters calculated with the moving window technique. Table 7.2 shows the classification accuracy of the DSM models created for the stratified EPA 20 m DEM of the three study areas. Table 7.3 presents the classification accuracy of the multiscale DSM models (pixel and window sizes alteration and stratification) of the three study areas. Finally, Table 7.4 summarises the classification accuracy of the study areas for the finest available DEM, pixel and window size alteration, stratification, and the new multiscale methodology.

7.3.1 Variograms

Spatial variability assessed using omni-directional variograms of elevation values (EPA 20 m DEM) was calculated for the three investigated areas and fitted using a spherical model (Figure 7.2). The figure contains considerably different variograms representing the underlying unique spatial structure of the three topographic areas.

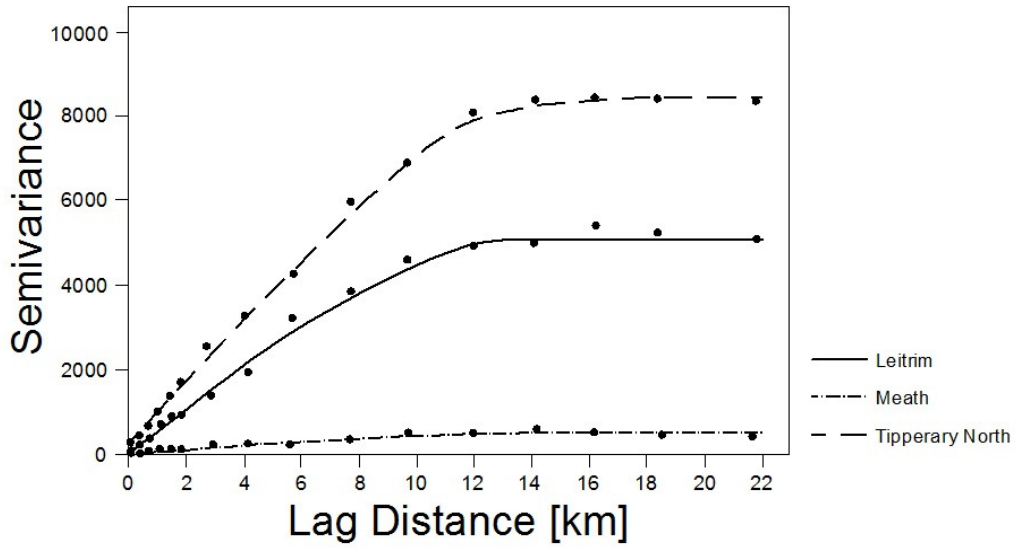


Figure 7.2 - The spherical global variograms of elevation value (EPA 20 m DEM) for Leitrim (a = 12,270; c = 4,800; c_0 = 10), Meath (a = 12,070; c = 650; c_0 = 10) and Tipperary North (a = 18,980; c = 9,550; c_0 = 260).

Tipperary North has the largest value of range and sill respectively 18,980 and 9,550 respectively, showing a strong long-distance correlation with a high degree of variance. It also has the largest value of nugget at 260, attributable to either measurement errors or spatial sources of variation at distances smaller than the elevation sampling interval. Leitrim is characterized by intermediate values of range and sill respectively 12,270 and 4,800 and also a very small nugget of 10. Meath has the same low value of nugget (10) and also has the lowest value of sill of 650 but an intermediate range (12,070) indicating low variance in the data caused by the homogenous flat topography. Despite the different spatial structures, with considerably different values of sill and range, all the three areas proved to have very similar values of spatial dependence calculated as the ratio between partial sill and the total sill ($c_1/(c_0+c_1)$) at 99.79% for Leitrim, 98.46% for Meath and 97.28% for Tipperary North. Since this ratio describes the proportion of the local variance which is spatially correlated these very high values show a strong spatial dependency. The examination of a soil-related environmental feature, such as elevation, through the analysis of its variogram is a valuable

guide for obtaining an order of magnitude of the scale at which covariates operate in the landscape, so influencing soil processes.

7.3.2 Moving Window Variograms

In order to spatially characterise local statistics in the three investigated areas, a moving window variogram technique was employed. For each cell of the original EPA 20 m DEM a local variogram was fitted and its parameters calculated. A total of 927,190 local variograms were computed for Leitrim, 842,754 for Meath and 935,172 for Tipperary North. The average window contained 625 cells and the smallest number of cells in a single window was 104 due to the edge effect. The local parameters calculated with the moving window approach are presented in the following three figures.

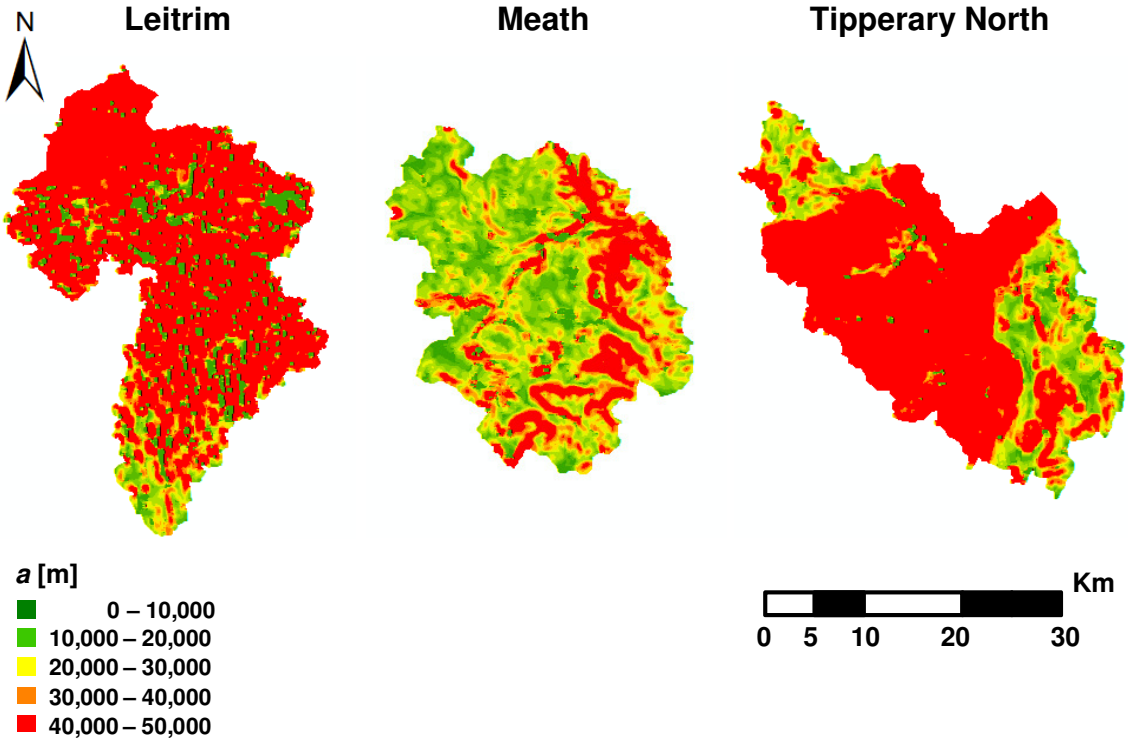


Figure 7.3 - Local distance parameter (*a*) for the investigated areas DEMs.

The local distance parameter (*a*) (Figure 7.3) shows a strong difference between areas on the drumlins and the south of Leitrim with values in the inter-drumlins plain and in the area of high relief in the north of the investigated site. Meath

displays two areas with high values of the range, one corresponding to the river network to the east and the other related to the rise in the overall height of the plain, by approximately 10 m, in the south. A very significant variation is shown in Tipperary North where the site is literally divided into two areas. In the centre the high relief region obtains high values of the range while in the north and south the plain has much lower values.

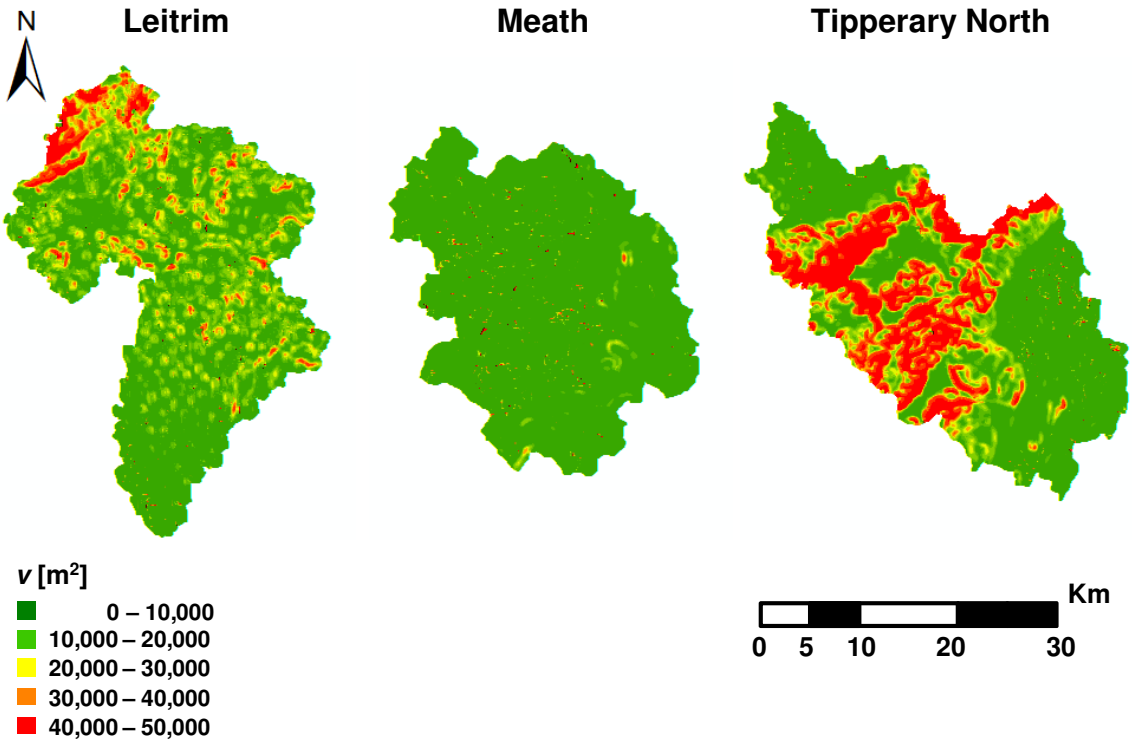


Figure 7.4 – Local variance (v) for the investigated areas DEMs.

In Figure 7.4 the local variance (v) is presented showing an interesting pattern in Leitrim, where the drumlins and the high relief area in the north obtain higher values than the inter-drumlin area and the south of the site. As the previously discussed global variogram indicated, the low value of sill (650) remains constant across the area. Tipperary North, on the other hand, shows a marked difference between the high relief area with high sill values and the low relief one with low sill values.

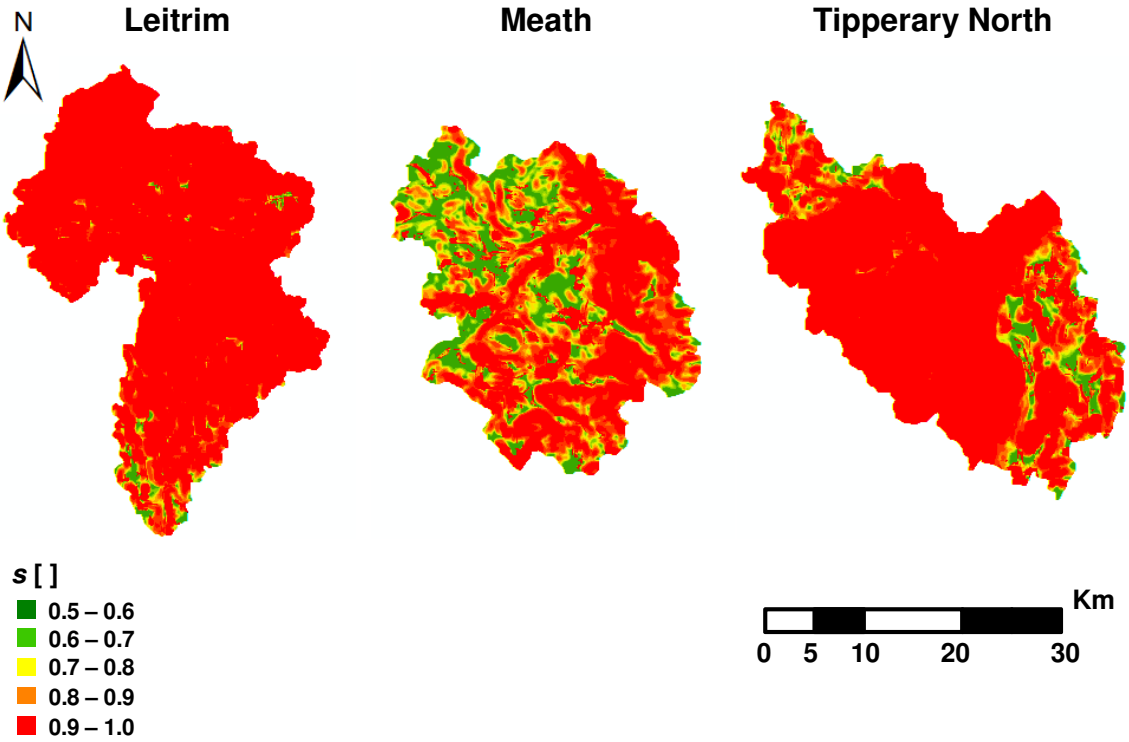


Figure 7.5 - Spatial dependence ratio (*s*) for the investigated areas DEMs.

Finally, in Figure 7.5 the spatial dependence ratio is presented. Leitrim shows a high degree of spatial dependence ratio (*s*), apart from an area in the south of the investigated site which shows a weaker spatial dependency. Interestingly, Meath has a marked distinction between the river network and the high plain area with strong spatial dependency and the rest of the site with low values. The low relief areas of Tipperary North display a low value of *s*, while the local variance in the centre of the site appears spatially correlated.

The presented local variograms parameters (*a*, *v* and *s*) demonstrate non-stationarity in the underlying stochastic process as they change over space. It is clear from Figure 7.3, Figure 7.4 and Figure 7.5 that spatial variation appears to have a structure over small areas where some regional patterns in these properties are evident. This has a direct consequence in terms of the soil processes controlled by topography that operate in the landscapes of the three study areas.

7.3.3 Multiscale DSM model

K-means clustering with v-fold cross validation was used to group the results of the moving window variograms as presented in Table 7.1. The clustering was performed using 500 iterations for the k-means and validated with v-fold cross validation (10 iterations with a set minimum number of 2 clusters and a maximum number of 25 clusters). The training errors for the three areas were: Leitrim (0.016), Meath (0.094) and Tipperary North (0.020). These errors measure the performance of the clustering which corresponds to the probability of misclassifying the data in the determination of an optimum cluster number.

Table 7.1 - Local statistics of the clustered areas DEMs for Leitrim, Meath and Tipperary North. Includes: local distance parameter (*a*), local variance (*v*) and spatial dependence ratio (*s*).

Clusters	***a*** [m]	***v*** [m^2]	***s*** [ ]	**Number of Cells**	**Coverage** [%]
Leitrim					
1	42,537	6,253	0.99	806,578	86.99
2	54,540	44,477	1.00	83,857	9.04
3	10,686	230	0.57	36,755	3.97
Meath					
1	18,046	858	0.82	703,684	83.50
2	56,943	9,130	1.00	139,070	16.50
Tipperary North					
1	52,623	42,001	0.99	544,832	58.26
2	18,910	670	0.84	275,408	29.45
3	21,920	240	0.59	70,979	7.59
4	10,713,096	1,996,060	1.00	26,278	2.81
5	7,327	130	0.27	17,674	1.89

Leitrim was divided into three clusters, the largest one covering 86.99% of the site with its centroid having high values of range (42,537) and similar values of variance (6,523) as indicated in the global variogram. The second area covering 9.04% of the site has high values of both range and sill and the third smaller group had low values of range and sill.

Meath was grouped into two clusters, a larger one (83.50% coverage) with similar values of range and sill as presented in the global variogram, 18,046 (*a*) and 858 (*v*) respectively; and a smaller one (16.50% coverage) with much higher local distance parameter (56,943) and local variance (9,130).

Tipperary North was divided into five clusters, with cluster 1 covering 58.26% of the area having high range and sill, 52,623 and 42,001 respectively and cluster 2, on the contrary, having low values of range (18,910) and sill (670) and covering 29.45% of the site. Cluster 3 has very similar values to the global variogram and as cluster 2, with a range of 21,920 and a sill of 240. The unreasonably high values of cluster 4 are most likely due to the fact that the moving window variogram did not achieve a value of sill in the neighbourhood and the algorithm returned a somewhat arbitrary number while fitting the model.

The spatial distribution of the clusters is presented in Figure 7.6. Leitrim shows an interesting pattern with a considerable fraction of the high relief area in the north of the site categorised at cluster 2, including the top of the drumlins scattered across the landscape. The area outside the drumlins belt in the south, which proved a real hindrance for the RF, as discussed in Chapter 4, is mainly being grouped as cluster 3.

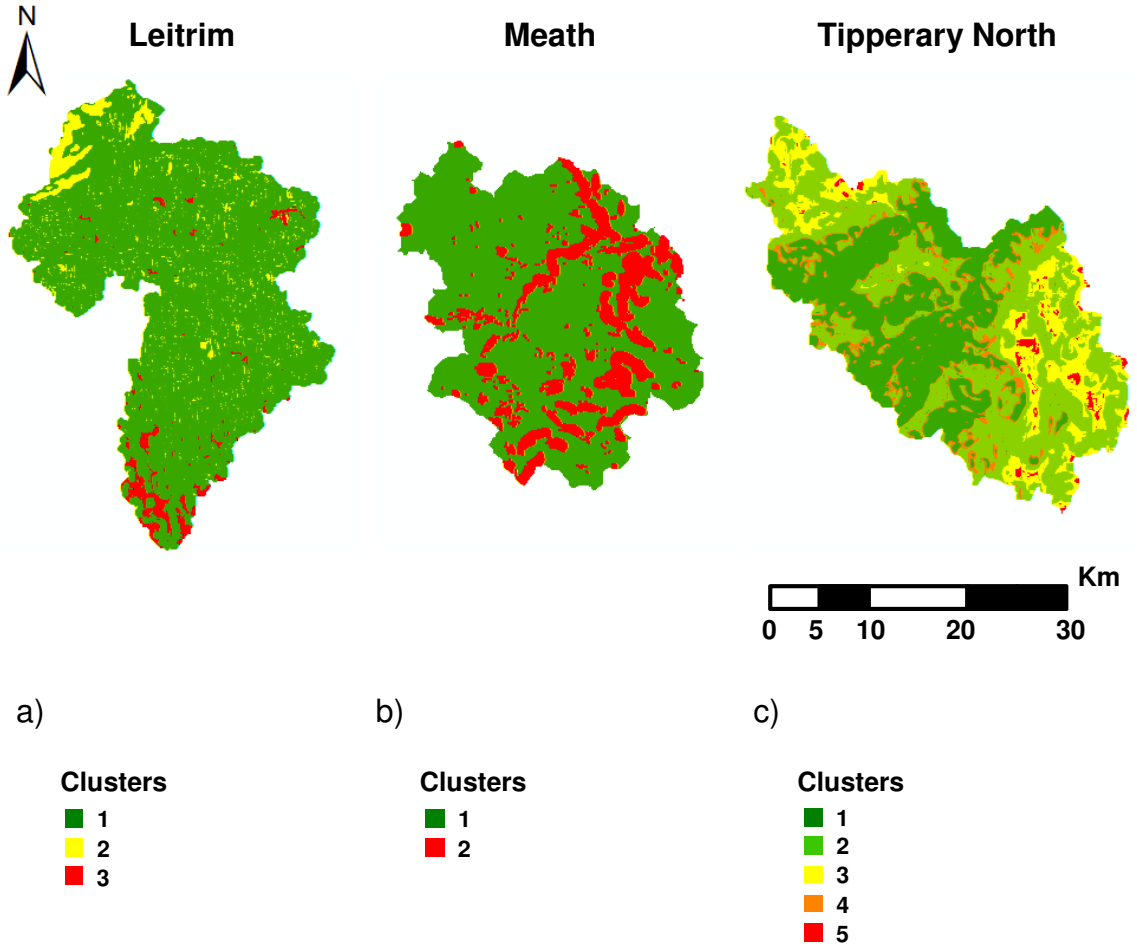


Figure 7.6 - DEM segmentation using k-means clustering of the local variogram parameters calculated with the moving window technique.

As expected, Meath looks divided accordingly to where the river network and the high plains are (red on Figure 7.6b). Tipperary North appears almost equally divided between cluster 1 for the high relief area and cluster 2 and cluster 3 for the plain. The group with the variograms that failed to achieve a sill, cluster 4, seems to follow cluster 1 as a contour line and separate it from cluster 2.

The EPA 20 m DEM was then used to create terrain attributes for each cluster and to train a RF model. The results of the modelling are presented in terms of classification accuracy which varies between 0% (no samples correctly classified) and 100% (all samples correctly classified). The classification accuracy of individual clusters using the EPA 20 m DEM without any pixel or window size alterations is shown in Table 7.2, allowing the effects of stratification to be tested.

Table 7.2 - Classification accuracy of the DSM models created for the stratified EPA 20 m DEM of the three study areas.

	Classification accuracy of stratified EPA 20 m DEM [%]
Leitrim	
Cluster 1	78.1
Cluster 2	10.9
Cluster 3	0.2
Total (Clusters 1, 2 and 3)	68.9
Meath	
Cluster 1	41.3
Cluster 2	38.3
Total (Clusters 1 and 2)	40.8
Tipperary North	
Cluster 1	70.4
Cluster 2	46.7
Cluster 3	42.3
Cluster 4	0.5
Cluster 5	5.5
Total (Clusters 1, 2, 3, 4 and 5)	58.1

Leitrim shows a large difference in classification accuracy between the three clusters. Cluster 1, dominating the large drumlin area characterised by gley soils, performs extremely well (78.1%), while Cluster 2, in the north of the study area where the blanket peats are present, and Cluster 3, in the south basin peats, are extremely inaccurate with only 10.9% and 0.2% respectively. The overall accuracy of the multiscale methodology presented is 68.9%.

Cluster 1 in Meath obtains 41.3% while Cluster 2, characterising the river network with a complex of grey-brown podzolics and gleys, achieves 38.3% of model classification accuracy, delivering a total of 40.8% for the entire study area.

The two largest clusters in Tipperary North, Cluster 1 and 2, perform very differently with the former achieving an extremely positive 70.4%, while the latter obtains a value of classification accuracy of 46.7%. Cluster 1 seems to represent the area of high relief where the brown podzolic soils are situated, while Cluster 2 characterises the area of limestone lowland dominated by the grey-brown podzolic soils. Cluster 3 with the waterlogged gley areas maintains a reasonable 42.3% while Cluster 4 and 5 achieve only 0.5% and 5.5%.

Table 7.3 - Classification accuracy of the multiscale DSM models (pixel and window sizes alteration and stratification) of the three study areas.

	Classification accuracy [%]	**Pixel and window sizes combination of optimal scale**
Leitrim		
Cluster 1	79.3	40 - 3x3
Cluster 2	10.9	20 - 3x3
Cluster 3	0.2	80 - 21x21
Total (Clusters 1, 2 and 3)	70.0	
Meath		
Cluster 1	68.6	260 - 19x19
Cluster 2	38.6	40 - 3x3
Total (Clusters 1 and 2)	63.6	
Tipperary North		
Cluster 1	71.2	40 - 3x3
Cluster 2	74.1	260 - 13x13
Cluster 3	48.6	260 - 11x11
Cluster 4	0.5	140 - 5x5
Cluster 5	5.6	40 - 3x3
Total (Clusters 1, 2, 3, 4 and 5)	67.1	

Table 7.3 presents the classification accuracy of the multiscale DSM models created by integrating the experimental methodology introduced in Chapter 4 (alteration of pixel and window size) with the previously presented stratification technique. In the case of Leitrim, the classification accuracy of Cluster 1 improves

only marginally by 1.2% (from 78.1% to 79.3%), while Clusters 2 and 3 do not change from the values obtained by the DSM models (10.9% and 0.2% respectively). Meath on the other hand, presents a sharp increase in classification accuracy with an improvement of 27.3% (from 41.3% to 68.6%) for Cluster 1, even though Cluster 2 does not present a considerable improvement, changing by only 0.3% (from 38.3% to 38.6%). Finally, Tipperary North matches the pattern observed for Meath, where Clusters 2 and 3 sharply increase to 74.1% and 48.6% respectively, while Clusters 1, 4 and 5 do not present any considerable change.

Table 7.4 - Classification accuracy of the three study areas for the finest available DEM; pixel and window size alteration; stratification; and the new multiscale methodology.

	Classification accuracy [%]
Leitrim	
Finest resolution DEM (EPA 20 m)	56.9
Pixel and Window size alteration (Chapter 4)	56.9
Stratification	68.9
Multiscale methodology (pixel and window sizes alteration and stratification)	70.0
Meath	
Finest resolution DEM (EPA 20 m)	38.8
Pixel and Window size alteration (Chapter 4)	58.8
Stratification	40.8
Multiscale methodology (pixel and window sizes alteration and stratification)	63.6
Tipperary North	
Finest resolution DEM (EPA 20 m)	44.6
Pixel and Window size alteration (Chapter 4)	51.1
Stratification	58.1
Multiscale methodology (pixel and window sizes alteration and stratification)	67.1

A summary table has been prepared to compare the classification accuracy of the three study areas using the different developed techniques (Table 7.4). The results obtained from the DSM models developed using the finest available DEM (EPA 20 m) are compared with the results of the experimental methodology presented in Chapter 4, the stratification analysis (Table 7.2) and the newly developed multiscale methodology (pixel and window size alteration in addition to stratification). In the case of Leitrim, results suggest that stratification on its own obtains similar classification accuracy (68.9%) in comparison to stratification and scale alterations (70.0%). For Meath, despite improving results by 2.0% (from 38.8% for the EPA 20 m DEM without stratification to 40.8% for the EPA 20 m DEM with stratification) and 4.8% (from 58.8% for the experimental methodology without stratification to 63.6% for the experimental methodology with stratification), stratification was outperformed by pixel and window size alterations which offered the greatest values of classification accuracy at 58.8% (38.8% without stratification) and 63.6% (40.8% without stratification) respectively. Tipperary North shows a similar high level of classification accuracy at 67.1% with the multiscale methodology while with the original EPA 20 m DEM it could only achieve 44.6%. Stratification has a relevant role to play in the improvement of the results of the DSM models in Tipperary North. This is clearly shown by the increase of 13.5% of classification accuracy with the EPA 20 m DEM performed on each individual cluster when stratified (from 44.6% without stratification to 58.1% with stratification) and the increase of 16.0% using the pixel and window size alterations (from 51.1% for the experimental methodology without stratification to 67.1% for the experimental methodology with stratification).

7.4 Discussion

Hutchinson and Gallant (2000) have recognised the need to identify appropriate DEMs scales for deriving various terrain processes and the need for effective methods to integrate terrain attributes and DSM modelling across different scales. The multiscale DSM methodology performed by clustering the local

variogram parameters calculated with a moving window technique and then used as an input variable for terrain attributes calculation, employed to train a RF model, has proved to increase the classification accuracy of soil taxonomic units. The proposed new methodology is a powerful technique for DSM, as it appears to support RF in the creation of more accurate relationships between terrain attributes and soil units, in comparison with the common practice of using a standard fine resolution DEM.

As a more intuitive manner of appreciating the spatial structure of DEMs, the clustering has allowed to make a more intuitive connection between the topography and the soil taxonomic units being classified. In some areas the relationship is extremely strong, as results of classification accuracy close to 70% prove, especially considering that only terrain features were used as predictors without accounting for other environmental covariates. The presented results clearly show that for other units, such as Clusters 2 and 3 in Leitrim and Clusters 4 and 5 in Tipperary North, the DSM models failed to relate terrain features to soil classes. This is possibly due to the fragmented nature of the areas, as the clusters appear too scattered to be modelled separately. Another explanation could be that the relationship between these clusters and topography, used as a predictor, is very weak and does not capture the soil formation processes active at particular locations. Limited areas, not accounting for more than 10% of the investigated sites and generally spatially concentrated, as in the south of Leitrim or some areas in the plain of Tipperary North, have posed a problem not only to the RF but also to the surveyors in the field. As the revised ISIS classification has demonstrated by rationalizing many of these areas. The different classification was simply not justified for the soil unit concept at this mapping scale in some areas, hence the different results between clusters.

It is also clear from these results that stratification on its own had an important role in the improvement of classification accuracy obtained by the newly developed methodology, as shown in Leitrim and Tipperary North. By segmenting the landscape, stratification seems to offer a direct way to connect DSM models

to specific soil forms, something that a single DSM model configuration across a unitary geographic space does not provide. This is in line with previous work by Corstanje *et al*. (2008a) where nested analysis was used to assess model scale relationships with soil processes. The addition of systematic scale alterations by varying pixel and window sizes, further improves the results, as shown in Meath and Tipperary North. Combining stratification with scale alterations has produced a comprehensive multiscale methodology. As suggested by Behrens *et al*. (2010b), multiscale techniques for DSM seem particularly suited to the intricate organisation of soil formation.

The experimental results presented in Chapter 4 showed that two main patterns of scale behaviour exist for the tested areas, suggesting that subdivision would address this scale incoherency. As previously discussed, the new multiscale methodology tackled that issue and also offered a way to identify the optimal pixel and window size combination for each cluster. This two-step approach is a way to address complexity and information generalization, while making the model computationally feasible. A remaining unsolved issue is how to make the model compute interacting pedogenetic factors at different scales within clusters. This might be solved by employing a mix of other SCORPAN factors, each operating on a different scale, so encapsulating the complexity of the soil forming processes operating.

In terms of the clustering, the k-means unsupervised clustering algorithm was selected for the presented research, as the focus was to compare the overall accuracy of multiscale methodologies with the common practice, but supervised clustering or other segmentation techniques that can classify the DEM according to its spatial structure should also be considered as suggested by Behrens *et al*., 2010b. Apart from the extensive computational time required by the moving window technique the main disadvantage of this approach, to calculate local statistics, is the selection of the window size that at the moment is rather arbitrary. The connection between the window size, dictating the neighbourhood of cells selected in the calculation of the local variogram, and its range, beyond which the

variance levels out and becomes independent, is not clear yet. It is worth also noting that for some cells, the variograms did not achieve a sill value, meaning that the spatial extent of the window was not large enough to capture the spatial variability at that location. For Leitrim and Meath this was not much of a concern as the total number of cells was insignificant, but for Tipperary North a total of 2.84% of the cells failed to deliver acceptable variogram parameters. Further investigation is needed to determine an optimal window size for this type of operations or to explore the feasibility of adaptive windows like the ones used in remote sensing for change detection (Gong and Corpetti, 2013). This issue is particularly important with regards to the clustering technique used in this research, as the k-means algorithm is sensitive to outliers and extremely large values of sill may substantially distort the extent of the clusters.

7.5 Conclusions

In this chapter a new multiscale methodology based on the analysis of the local variogram parameters, calculated using a moving window technique and k-means clustering with v-fold cross validation for segmentation of the DEM, was developed and tested. The results suggest that this methodology can achieve a higher level of classification accuracy in comparison with using the original EPA 20 m DEM, or the experimental methodology altering pixel and window sizes without segmenting the landscape. The overall improvement is substantial and consistent across various pedological and morphological conditions as shown for the three tested areas. This new modelling approach, by segmenting the landscape into areas in which different processes are active at different scales, incorporates the scale issue into the model form, as the scale effects become an inherent part of the RF inference.

8 CONCLUSIONS

The general conclusion of this research is that spatial scale analysis of environmental covariates enhances the practice of DSM, improving overall classification accuracy. The newly developed multiscale methodology can be successfully integrated in the current DSM analysis of soil taxonomic units performed with data mining techniques advancing the practice of soil mapping. By offering an innovative way to learn more about pedogenesis and soil variation in the landscape and by increasing the overall accuracy of DSM modelling, spatial scale analysis deserve and need more attention from the DSM community.

8.1 Review of the objectives

The hypothesis of this research was that the resolution of environmental covariates affected the accuracy of soil prediction in DSM and therefore spatial scale analysis would improve the accuracy of mapping soil taxonomic units. From the presented results it is possible to conclude that this is certainly true as demonstrated in Chapter 4 (experimental methodology), Chapter 6 (wavelet decomposition) and Chapter 7 (multiscale methodology).

General conclusions are listed according to the objectives set out for this research in Chapter 1 as follows:

Objective 1 - To investigate the effects of scale on DSM analysis.

Scale has been proved (Chapter 4) having a significant role in DSM analysis of soil taxonomic units, directly influencing classification accuracy of soil taxonomic units. The common practice in DSM to use the finest resolution of DEM available has been shown to have a detrimental effect on the prediction accuracy of soil taxonomic units.

Objective 2 - To test the interaction between pixel and window sizes, with data mining classifiers, for the purpose of modelling soil taxonomic units.

In order to characterise scale variation, both pixel and window size alterations were tested including their interaction (Chapter 4). Two patterns of behaviour emerged: flat homogeneous areas preferring coarser resolution DEMs independently of window size and morphologically varied areas preferring fine resolution DEMs with small window sizes but also coarser resolution DEMs associated with large window sizes.

Objective 3 - To identify, from published literature, methodologies that can be used in quantitative scale detection.

A set of eight empirical approaches, from related scientific fields, was selected in the literature and put to the test (Chapter 5) by comparing them with the results of the experimental methodology. Seven out of the eight failed to deliver meaningful results, some due to the lack of scientific principles behind them, others failing to take into account the intrinsic characteristics of the data or being too entrenched in their field of application and being inflexible. Although, the inflection points approach seemed the most promising one, it still needed some further refinement.

Objective 4 - To test the identified methods in the determination of the most suitable DEM pixel size for application in landscape-scale DSM.

2D wavelet analysis has shown in Chapter 6 that by spatially decomposing a DEM it is possible to remove specific sources of variation, which might be unnecessary for DSM analysis, improving classification accuracy. The 2D DWT could be introduced in DSM as a standard method to spatially decompose the DEM used in generating the required terrain attributes increasing model performance.

Objective 5 - To develop a multiscale approach for DSM.

A new multiscale methodology based on the analysis of the local variogram parameters calculated using a moving window technique, and k-means clustering with v-fold cross validation for segmentation of the DEM was developed and tested (Chapter 7). The results showed the highest level of classification accuracy in comparison with all the other tested techniques, making it a relevant tool for data mining based DSM.

Objective 6 - To develop recommendations on scaling environmental covariates used for DSM.

A list of practical recommendations is presented further in this chapter, including three main areas of advice: scale affects the results of DSM models; incorporate scale into DSM models; and include scale information into metadata.

8.2 Contribution to knowledge

The original contribution to knowledge of this research has been combining a DEM segmentation technique (performed by k-means clustering of local variograms parameters calculated in a moving window) with an experimental methodology altering DEM scales. This has created a new multiscale approach in DSM. A number of new findings have been made during this research work that enhanced the knowledge of DSM in relation to spatial scale. These include:

Improved understanding

Terrain attributes are sensitive to the scale of the source DEM and behave in different ways to this alteration, affecting DSM prediction accuracy. Hence, fine resolution DEMs are not always the best choice in DSM for the modelling of soil taxonomic units.

Two main patterns of scale behaviour have been described: flat areas obtaining the best classification accuracies at coarser pixel sizes and morphologically varied areas being influenced by the interaction of pixel and window alterations,

obtaining the best accuracies at fine resolutions with small window sizes but also at coarser resolutions and large window sizes.

Assessment of existing approaches

Generally accepted rules of thumb, commonly used to identify an optimum pixel size, have a detrimental effect on the final output of DSM modelling and could mislead practitioners.

DEMs decomposed with 2D DWT improve accuracy of DSM models by reducing the source of variation (redundant information that makes it difficult to link topographical change with soil variation).

Quantitative methodologies in scale detection

An exploratory DSM exercise at different scales altering pixel and window sizes will provide relevant knowledge of the area investigated, improving the final prediction of soil taxonomic units for mapping exercises at the landscape scale.

A new multiscale approach made by combining a DEM segmentation technique, including k-means clustering of local variogram parameters calculated with a moving window technique, with an experimental methodology altering DEM scales offers a way to significantly improve classification accuracy of soil taxonomic units.

8.3 Limitations

In terms of limitations associated with the results presented in this research, three main sources of constraints are identified as: the 6 inches soil map, the DEM and restrictions associated with the methodologies investigated.

If possible, soil forming processes should be observed and measured at the scale at which they take place. This is not always possible or feasible, and very often the legacy information available (point samples, soil polygons, etc.) is a reflection

of small scale mapping projects. It is recognised that good quality soil data are a necessary requirement to create meaningful DSM functions and accurately evaluate the quality of their outputs. A common problem with soil taxonomic units, including the 6 inch maps used in this research, is the lack of intra class variability measurements and the crisp boundaries between classes that are interpreted by data mining techniques as two different entities, while in the field there is some sort of gradual separation rather than a clear division. This fact could be responsible for the poor performance of the techniques used (RF and NN) in some areas of the investigated sites that did not respond to the scale alterations. This type of qualitative and categorical information represents the great majority of legacy data in soil science. It is still currently used by a wide range of users, so methodologies need to be developed using this type of information.

All the analyses were based on the EPA 20 m DEM created with the ANUDEM software (Hutchinson, 2007) by spline interpolation using OSI vector spot heights, drainage lines and contour data as inputs at the 1:50,000 scale. The DEM was corrected both morphologically, by removing all apparent height anomalies, and hydrologically, by enforcing a correct drainage network. Limitations in terms of accuracy might have affected low relief areas, where the 10 m contour intervals could have missed finer scale terrain features, limiting representation of topographic structures. The impact of this in DSM analysis using the 6 inch soil maps is considered negligible, as the soil surveyors in low relief areas would not have taken into account micro-topographic features in delineating soil series. This fact was stated by a soil surveyor involved in the ISIS project. The DEM used was a good compromise in terms of accuracy and spatial resolution, and was deemed suitable for the analysis of spatial scales in DSM, especially considering that it was the most accurate elevation information available at the national scale in Ireland.

The results of the experimental methodology presented in Chapter 4 set the scale benchmark for all the other techniques tested. Although the developed DSM models could largely account for the appropriate classification of soil classes in

Leitrim and Meath, they failed to explain soil spatial variation in Tipperary North. The apparent scale independence suggested that the method needed a further stage to separate areas with distinctive landscape characteristics. This was confirmed after Tipperary North was sub-divided into high and low relief areas as the accuracy of the models improved matching the scale behaviours previously observed for Leitrim and Meath. The application of the experimental methodology was also limited due to its computational and labour intensive nature. For this reason empirical approaches were selected from the literature and tested in Chapter 5. On close scrutiny, all their limitations in terms of data handling, lack of adaptability and insufficient scientific credentials made their use in DSM inadequate, as they failed to improve the classification accuracy of models.

A more established and increasingly popular technique in soil science, wavelet analysis, was tested with both 1D and 2D versions. The visual inspection of the 1D DWT on characteristic transects proved difficult for interpretation and practical implementation in DSM modelling. Limitations on solving the issue of incoherent scale response, observed for the three study areas, were also apparent in the 2D DWT during implementation. This was solved with the newly developed multiscale methodology which segments the landscape into clusters with homogeneous spatial structures and identifies optimal scale combinations of pixel and window sizes for each cluster. The model cannot compute interacting pedogenetic factors at different scales within clusters and this will limit its overall accuracy. The clustering technique is also sensitive to extreme values that interfere with the extent and spatial distribution of clusters. Some of these outliers derive from the fact that the moving window technique was not able to correctly capture the variogram parameters. Consequently, the spatial extent of the window was not large enough to capture the spatial variability at that location. This might have affected the performance of small scattered clusters across Leitrim and Tipperary North.

8.4 Recommendations

Supported by the results of this research, a series of recommendations are made to accurately incorporate spatial scale analysis into DSM operations.

Scale affects the results of DSM models

- It should not be assumed that fine resolution DEMs are always the best choice for DSM prediction of soil taxonomic units, as they are not.
- Particular attention should be put in the interpretation of terrain attributes, extensively used in DSM, as they are scale dependent but behave in different ways to scale alterations.

Incorporate scale into DSM models

- An exploratory scale test analysis should be included into a preliminary knowledge discovery stage before creating any final DSM model.
- In order to perform a multiscale methodology improving classification accuracy of soil taxonomic units, a strategy is to segment the DEM into homogeneous areas using k-means clustering of local variograms parameters calculated with a moving window and then performing DSM analysis on each cluster.
- Rules of thumb, supposed to help in the selection of an optimal scale, could mislead offering a false sense of security and their results should be critically evaluated if not avoided.

Include scale information into metadata

- Metadata of spatial information should always include scale, resolution and accuracy of the content.
- The current resolution of a dataset not always corresponds to the resolution of the information used to create it. It could have been

aggregated for visualization, smoothed to remove noise, interpolated to create a continuous raster or produced by a model with an embedded scale. This information should always be incorporated in the metadata to be then properly handled during analysis.

8.5 Future work

During this research project a number of questions for further investigation have emerged and are briefly discussed.

8.5.1 Exploring additional study areas

Further studies should be undertaken transferring the newly developed multiscale methodology to other parts of the world, validating the presented results on different pedological and geomorphological conditions. Moreover, areas with different types of human influence could be explored with the new methodology, testing if scale analysis can be used as a detection tool evaluating at what scales human activity is impacting the environment and which ecosystems and cycles are in more danger.

8.5.2 Soil data

Additional work is needed to understand how different soil classes behave at different scales. This could offer a new insight into the relationship between soil taxa and soil mapping units. The methodology developed for categorical soil data will have to be tested for quantitative soil properties exploring the role of scale analysis in the prediction of numerical variables with associated uncertainties.

8.5.3 Environmental covariates

In this research, terrain attributes derived from a DEM were used as environmental covariates. Further work needs to be carried out, when a suitable national coverage is achieved with Lidar. This will test whether the presented

results obtained with the interpolated EPA 20 m DEM can be transferred to DEMs derived from elevation points measured by remote sensing.

In order to remove the embedded scale set by the pixel based moving window technique used in the analysis of terrain attributes, an object based approach should be applied to the newly developed multiscale methodology.

It is also important to expand the scope of the multiscale spatial scale methodology to other environmental covariates valuable for DSM modelling such as climatic properties (temperature, rainfall, solar radiation, atmospheric pressure, etc.), biotic properties (land use, land cover, spectral indices, etc.), human activities (contamination, greenhouse gas emission, etc.) or other data that are rapidly being created with remote and proximal sensing technologies.

8.5.4 Spatial soil scaling theory

The complexity of scale will require more than the development of empirical methodologies in different scientific domains. This will need the establishment of a new multidisciplinary branch of research to create a theory of scale. This will certainly require advancements in theory and technology combined with a new statistical framework and experimental schemes to capture the complexity of soils as a system (Young *et al.*, 2008). Understanding the scaling behaviour of soil will allow estimation of the behaviour of soil processes at all scales. This paradigm of complexity involving processes at small scales that determine properties at larger scales will overall help to better understand soil.

8.6 Final remark

The main challenge in DSM research is to connect the spatial scales at which processes happen with the larger scales at which soil functioning is observed and units mapped. This research has tested several existing techniques and developed a new multiscale methodology to include spatial scale into DSM

operations. Incorporating spatial scale analysis of environmental covariates in DSM modelling has been proved beneficial to better capture soil spatial variation. The results presented in this research suggest that the multiscale methodology is the most effective way to take scale into account in DSM. The classification accuracies obtained for all the test areas with this new methodology were the highest in comparison with all the other techniques. Multiscale offers a more intuitive manner of appreciating scale behaviour and the connection between the topography and the soil taxonomic units being classified, helping DSM practitioners with their mapping activities.

The future of DSM, as it successfully progresses from the early pioneering years into an established discipline, will have to include scale and in particular multiscale in its methodology. DSM will have to move from a methodology of spatial data with scale to a spatial scale methodology. As stated by Burrough *et al.* (1994) *"gradually the general nature of soil variation, and its unpredictability, have led us to see variability as a key soil attribute rather than a nuisance, though this enlightened view is certainly not shared by everyone",* it is now time to consider scale as a key soil and modelling attribute rather than a nuisance.

REFERENCES

Aalen, F.H.A, Whelan, K., Stout, M., 1997. Atlas of the Irish rural landscape. Cork University Press, 352.

Abedini, M.J, Shaghaghian, M.R., 2009. Exploring scaling laws in surface topography. Chaos, Solitons & Fractals 42, 2373-2383.

Addiscott, T.M., 1998. Modeling concepts and their relation to the scale of the problem. Nutrient cycling in Agroecosystems 50, 239-245.

Allegrini, P., Benci, V., Grigolini, P., Hamilton, P., Ignaccolo, M., Menconi, G., Palatella, L., Raffaelli, G., Scafetta, N., Virgilio, M., Jang, J., 2003. Compression and diffusion: a joint approach to detect complexity. Chaos, Solitons & Fractals 15, 517-535.

An Foras Taluntais, 1973. County Leitrim Resource Survey: Land use potential (soils, grazing capacity, and forestry). Soil Survey Bulletin 29. An Foras Taluntais, Dublin, 110.

Atkinson, P.M., Tate, N.J., 2000. Spatial scale problems and geostatistical solutions: a review. Professional Geographer 52, 607-623.

Avery, B.W., 1987. Soil survey methods: a review. Technical Monograph Series 18. Soil Survey and Land Resource Centre, Silsoe, 86.

Behrens, T., Forster, H., Scholten, T., Steinrucken, U., Spies, E.D., Goldschmitt, M., 2005. Digital Soil Mapping using Artificial Neural Networks. Journal of Plant Nutrition and Soil Science 168, 1-13.

Behrens, T., Schmidt, K., Zhu, A. X., Scholten, T., 2010a, The ConMap approach for terrain-based digital soil mapping. European Journal of Soil Science, 61, 133-143.

Behrens, T., Zhu, A., Schmidt, K., Scholten, T., 2010b. Multi-scale digital terrain analysis and features selection for digital soil mapping. Geoderma 155, 175-185.

Bierkens, M.F.P., Finke, P.A., De Willigen, P., 2000. Upscaling and Downscaling Methods for Environmental Research. Kluwer Academic Publishers, Dordrecht, The Netherlands, 204.

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

Biswas A., Si, B.C., Walley, F.L., 2008. Spatial relationship between δ15N and elevation in agricultural landscapes. Nonlinear Process in Geophysics 15, 397-407.

Biswas, A., Si, B.C., 2011. Application of Continuous Wavelet Transform in Examining Soil Spatial Variation: A Review. Mathematical Geosciences 43, 379-396.

Biswas, A., Cresswell, H.P., Viscarra Rossel, R.A., Si, B.C., 2013. Curvelet transform to study scale-dependent anisotropic soil spatial variation, Geoderma, in press.

Bloschl, G., 1995. Hydrologic synthesis: Across processes, places, and scales. Water resources research 42, 1-3.

Bloschl, G., Silvaplan, M., 1995. Scale issues in hydrological modelling: a review. Hydrological processes 9, 251-290.

Bock, M., Bohner, J., Conrad, O., Kothe, R., and Ringeler A., 2008. SAGA: System for the Automated Geoscientific Analysis. Dept. of Physical Geography, Hamburg, Germany. http://www.saga-gis.org/en/index.html (last verified September 2013).

Bockheim, J.G., McLeod, M., 2008. Soil distribution in the McMurdo dry valleys, Antarctica. Geoderma 144, 43-49.

Borkowski, A., Meier, S., 1994. A procedure for estimating the grid cell size of digital terrain models derived from topographic maps. Geo-Informations-System 7, 2-5.

Boulaine, J., 1980. Pedologie Appliquee. Collection Sciences Agronomiques. Masson, Paris, 220.

Breiman, L., 2001. Random Forests. Machine Learning 45, 5-32.

Brus, D.J., de Gruijter, J.J., van Groenigen, J.W., 2006. Designing Spatial Coverage Samples Using the k-means Clustering Algorithm. In: Lagacherie, P., McBratney, A. B., Voltz, M., Digital Soil Mapping - An Introductory Perspective. Elsevier, Amsterdam, 3-22.

Brus, D.J., Lark, R.M., 2013. Soil Surveys. In: El-Shaarawi, A.H., Piegorsch, W.W., Encyclopedia of Environmetrics, 2027-2029.

Burrough, P.A., Bouma, J., Yates, S.R, 1994. The state of the art in pedometrics. Geoderma 62, 311-326.

Burrough, P.A., McDonnell, R.A., 1998. Principles of Geographical Information Systems. Oxford University Press, New York, 356.

Carre, F., McBratney, A.B., Mayr, T., Montanarella, L., 2007. Digital soil assessments: Beyond DSM. Geoderma 142, 69-79.

Cao, C., Lam, N.S., 1997. Understanding the scale and resolution effects in remote sensing and GIS. In: Quattrochi, D.A, Goodchild, M.F. Scale in Remote Sensing and GIS. CRC Press. 57-62.

Corstanje, R., Grunwald, S., Lark, R.M, 2008a. Inferences from fluctuations in the local variogram about the assumption of stationarity in the variance. Geoderma 143, 123-132.

Corstanje, R., Kirk, G.J.D., Lark, R.M., 2008b. The behaviour of soil process models of ammonia volatilization at contrasting spatial scales. European Journal of Soil Science 59:1271-1283.

Cruickshank, J.G., 1997. Soil and Environment: Northern Ireland. Agricultural and Environmental Science Department, Queen's University Belfast, 214.

Daubechies, I., 1990. The wavelet transform, time-frequency localization and signal analysis. IEEE Transactions on Information Theory 36, 961-1005.

Daubechies, I., 1992. Ten lectures on wavelets. CBMS-NSF Series in Applied Mathematics, 61. SIAM, Philadelphia, 377.

De Bartolo, S., Otten, W., Cheng, Q., Tarquis, A.M., 2011. Modelling soil system: complexity under your feet. Preface. Biogeosciences 8, 3139-3142.

Dobos, E., Carre, F., Hengl, T., Reuter, H.I., Toth, G., 2006. Digital Soil Mapping as a support to production of functional maps. Office for Official Publications of the European Communities, Luxemburg, 68.

Dragut, L., Schauppenlehner, T., Muhar, A., Strobl, J., Blaschke, T., 2009. Optimization of scale and parameterization for terrain segmentation: An application to soil-landscape modeling. Computers & Geosciences 35, 1875-1883.

Eaglesona, S., Escobarb, F., Williamsona, I., 2002. Hierarchical spatial reasoning theory and GIS technology applied to the automated delineation of administrative boundaries. Computers, Environment and Urban Systems 26, 185-200.

ESRI 2010. ArcGIS Desktop: Release 10.0. Redlands, California. Environmental Systems Research Institute.

Falconer, K., 1990. Fractal Geometry Mathematical Foundations and Applications. John Wiley, Chichester, 288.

Fealy, R., 2006. Landslide susceptibility mapping in Ireland. In: Creighton, R. Landslides in Ireland. Geological Survey of Ireland, Dublin, 32-46.

Finch, T.F., Gardiner, M.J., Comey, A., Radford, T., 1983. Soils of Co. Meath. Soil Survey Bulletin 37. An Foras Taluntais, Dublin, 162.

Finch, T.F., Gardiner, M.J, 1993. Soils of Tipperary North Riding. Soil Survey Bulletin 42. An Foras Taluntais, Dublin, 142.

Florinsky, I.V., 1998. Accuracy of local topographic variables derived from digital elevation models. International Journal of Geographical Information Science 12, 47-61.

Florinsky, I.V., 2011. Digital terrain analysis in soil science and geology. Elsevier Academic Press, Amsterdam, 379.

Florinsky, I.V., Kuryakova, G.A., 2000. Determination of grid size for digital terrain modeling in landscape investigations-exemplified by soil moisture distribution at a micro-scale. International Journal of Geographical Information Science 14, 815-832.

Fotheringham, A.S, Charlton, M., Brunsdon, C., 1996. The geography of parameter space: an investigation of spatial non-stationarity. International Journal of Geographical Information Systems 10, 605-627.

Gallagher, P.H, Walsh, T., 1943. Characteristics of Irish Soil Types – Part I. Proceedings of the Royal Irish Academy 42, 205-250.

Gallant, J.C., Hutchinson, M.F., 1996. Towards an Understanding of Landscape Scale and Structure. Third International Conference/Workshop on Integrating GIS and Environmental Modeling. Santa Fe, CA, 21-25.

Gallant, J.C., Hutchinson, M.F., 1997. Scale dependence in terrain analysis. Mathematics and Computers in Simulation 43, 313-321.

Gardiner, M.J., Radford, T., 1980a. Ireland: General Soil Map. An Foras Taluntais, Dublin.

Gardiner, M.J., Radford, T., 1980b. Soil Associations of Ireland and their Land Use potential. Explanatory bulletin to the Soil Map of Ireland 1980. Soil Survey Bulletin 36. An Foras Taluntais, Dublin, 142.

Geng, X., Fraser, W., VandenBygaart, B., Smith, S., Waddell, A., Jiao, Y., Patterson, G., 2010. Toward Digital Soil Mapping in Canada: Existing Soil Survey Data and Related Expert Knowledge. Digital Soil Mapping – Bridging Research, Environmental Application, and Operation, 473. Springer. 325-335.

Gershenfeld, N., 1999. The Nature of Mathematical Modelling. Cambridge University Press, Cambridge, 356.

Gong, X., Corpetti, T., 2013. Adaptive window size estimation in unsupervised change detection. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 6, 991-1003.

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

Goodchild, M.F., 2001. Metrics of scale in remote sensing and GIS. International Journal of Applied Earth Observation and Geoinformation 2, 114-120.

Goodchild, M.F., 2009. NeoGeography and the nature of geographic expertise. Journal of Location Based Services 3, 82-96.

Goodchild, M.F., Proctor, J., 1997. Scale in a digital geographic world. Geographical and Environmental Modeling 1, 5-23.

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

Goovaerts, P., 1997. Geostatistics for Natural Resources Evaluation. Oxford University Press, New York, 483.

Grimm, R., Behrens, T., Marker, M., Elsenbeer, H., 2008. Soil organic carbon concentrations and stocks on Barro Colorado Island - digital soil mapping using Random Forest analysis. Geoderma 146, 102-113.

Grinand, C., Arrouays, D., Laroche, B., Martin, M.P., 2008. Extrapolating regional soil landscapes from an existing soil map: sampling intensity, validation procedures, and integration of spatial context. Geoderma 143, 180-190.

Grohmann, C.H., Riccomini, C., 2009. Comparison of roving-window and search-window techniques for characterising landscape morphometry. Computers & Geosciences 35, 2164-2169.

Grohmann, C.H., Smith, M.J., Riccomini, C., 2010, Multiscale Analysis of Topographic Surface Roughness in the Midland Valley, Scotland. IEEE Transactions on Geoscience and Remote Sensing 99, 1-14.

Grunwald, S., 2009. Multi-criteria characterization of recent digital soil mapping and modelling approaches. Geoderma 152, 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, 1201-1213.

Haas, T.C., 1990. Lognormal and moving window methods of estimating acid deposition. Journal of the American Statistical Association 85, 950-963.

Hartigan, J.A., 1975. Clustering Algorithms (Probability & Mathematical Statistics). John Wiley & Sons Inc, 366.

Hengl, T., 2003. Pedometric mapping: bridging the gaps between conventional and pedometric approaches. PhD thesis, University of Wageningen, 214.

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

Hengl, T., Husnjak, S., 2006. Evaluating adequacy and usability of soil maps in Croatia. Soil Science Society of America Journal 70, 920-929.

Hengl, T., Toormanian, N., Reuter, H.I., Malakouti, M.J., 2007. Methods to interpolate soil categorical variables from profile observations: lessons from Iran. Geoderma 140, 417-427.

Hengl, T., Reuter, H.I., 2009. Geomorphometry: concepts, software, applications, Amsterdam. Elsevier, Oxford, 765.

Hengl, T., Nikolic, M., MacMillan, R.A., 2013. Mapping Efficiency and information content. International Journal of Applied Earth Observation and Geoinformation 22, 127-138.

Hudson, B.D. (1992). The soil survey as paradigm-based science. Soil Science Society of America Journal. 56, 836-841.

Hutchinson, M.F., 2007. ANUDEM Version 5.2. Fenner School of Environment and Society, Australian National University. http://fennerschool.anu.edu.au/research/products/anudem (last verified September 2013).

Hutchinson, M.F, Gallant, J.C., 2000. Digital elevation models and representations of terrain shape. In: Wilson, J.P, Gallant, J.C, Terrain Analysis, Principles and Applications. John Wiley & Sons, Chichester, 29-50.

Ibanez, J.J, Saldana, A., 2008. The continuum dilemma in pedometrics and pedology. In: Krasilnikov P., Carre, F., Montanarella, L., Soil geography and geostatistics: Concepts and applications. Joint Research Centre, Scientific and Technical Reports, 130-147.

Jarvis, P.G., 1995. Scaling processes and problems. Plant, Cell & Environment 18, 1079-1089.

Jenny H., 1941. Factors of Soil Formation. A System of Quantitative Pedology. McGraw Hill, New York, 281.

Jones, R.J.A, Houskova, B., Bullock, P., Montanarella, L., 2005. Soil Resources of Europe, Second edition. European Technical Report: EUR 20559 EN, Office for Publications of the European Communities, Luxemburg, 420.

Jones, R.J.A., Hannam, J.A., Creamer, R.E., MacDonald, E., Sills, P., Mayr, T.R., Shulte, R.P.O., 2011. Classification and rationalisation of soil series in Ireland. ISIS Technical Monograph No.ISIS_WP1_D2.1, Cranfield University & Teagasc, 99.

Kempen, B., Brus, D.J., Stoorvogel, J.J., Heuvelink, G.B.M., Vries, F., 2012. Efficiency comparison of conventional and digital soil mapping for updating soil maps. Soil Science Society of America Journal 76, 2097-2115.

Kienzle, S., 2004. The effect of DEM raster resolution on first order, second order and compound terrain derivatives. Transactions in GIS 8, 83-112.

Kidner, D.B., Smith, D.H., 2003. Advances in the Data Compression of Digital Elevation Models. Computers & Geosciences 29, 985-1002.

Kolmogorov, A.N., 1965. Three approaches to the quantitative definition of information. Problems of Information Transmission 1, 1-7.

Labat, D., 2005. Recent advances in wavelet analyses: Part 1. A review of concepts. Journal of Hydrology 314, 275-288.

Lagacherie, P., 2008. Digital soil mapping: a state of the art. In: Hartemink, A., McBratney, A.B., Mendonca-Santos, M.L., Digital Soil Mapping with Limited Data. Springer, Dordrecht, 3-14.

Lagacherie, P., McBratney, A. B., Voltz, M., 2006. Digital Soil Mapping - An Introductory Perspective. Elsevier, Amsterdam, 616.

Lagacherie, P., McBratney, A.B., 2006. Spatial Soil Information Systems and Spatial Soil Inference Systems: Perspectives for Digital Soil Mapping. In: Lagacherie, P., McBratney, A. B., Voltz, M., Digital Soil Mapping - An Introductory Perspective. Elsevier, Amsterdam, 3-22.

Lagacherie, P., Holmes, S., 1997. Addressing Geographical Data Errors in a Classification Tree for Soil Unit Prediction. International Journal of Geographical Information Science 11, 183-198.

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

Lamorski, K., Pachepsky, Y., Slawinski, C., Walczak, R.T., 2008. Using support vector machines to develop pedotransfer functions for water retention of soils in Poland. Soil Science Society of America Journal 72, 1243-1247.

Lanza, L.G., Gallant, J.C., 2006. Fractals and Similarity Approaches in Hydrology. Encyclopedia of Hydrological Sciences.

Lark, R.M., 2005. Spatial analysis of categorical soil variables with the wavelet transform. European Journal of Soil Science 56, 779-792.

Lark, R.M., 2006. Decomposing Digital Soil Information by Spatial Scale. In: Lagacherie, P., McBratney, A. B., Voltz, M., Digital Soil Mapping - An Introductory Perspective. Elsevier, Amsterdam, 301-326.

Lark, R.M., Webster, R., 1999. Analysis and elucidation of soil variation using wavelets. European Journal of Soil Science, 50,185-206.

Lark, R.M., Webster, R., 2001. Changes in variance and correlation of soil properties with scale and location: analysis using an adapted maximal overlap discrete wavelet transform. European Journal of Soil Science, 52, 547-562.

Lark, R.M., Cullis, B.R., 2004. Model-based analysis using REML for inference from systematically sampled data on soil. European Journal of Soil Science 55, 799-813.

Lark, R.M., Milne, A.E., Addiscott, T.M., Goulding, K.W.T., Webster, R., O'Flaherty, S., 2004. Scale- and location dependent correlation of nitrous oxide emissions with soil properties: An analysis using wavelets. European Journal of Soil Science 55, 611-627.

Lark, R.M., Webster, R., 2004. Analysing soil variation in two dimensions with the discrete wavelet transform. European Journal of Soil Science 55, 777-797.

Lee, J., Coulter, B., 2005. Application of Soils Data to Land Use and Environmental Problems in Ireland. European Soil Bureau – Research Reports 9, 187-191.

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

Licznar, P., Nearing, M.A., 2003. Artificial neural networks of soil erosion and runoff prediction at the plot scale. Catena 51, 89-114.

Lillesand, T.M., Kiefer, R.W, Chipman, J.W., 2008. Remote sensing and image interpretation. John Wiley & Sons, New York, 763.

Mackaness, W., Steven, M ., 2006. An Algorithm for Localised Contour Removal over Steep Terrain. The Cartographic Journal 43, 144-156.

MacMillan, R.A., Moon, D.E., Coupé. R.A, 2007. Automated predictive ecological mapping in a forest region of B.C., Canada, 2001–2005. Geoderma 140, 353-373

Maidment, D.R., 2002. Arc Hydro: GIS for Water Resources, ESRI Press, Redlands, California, 203.

Mallat, S.G., 1989. A theory for multiresolution signal decomposition: the wavelet representation. IEEE Transactions on Pattern Analysis and Machine Intelligence 11, 674–693.

Malone, B.P, McBratney, A.B, Minasny, B., 2013. Spatial scaling for Digital Soil Mapping. Soil Science Society of America Journal 77, 890-902.

Mandelbrot B.B., 1983. The Fractal Geometry of Nature. Freeman, San Francisco, 460.

Marceau, D.J., Hay, G.J., 1999. Remote sensing contributions to the scale issue. Canadian Journal of Remote Sensing 25, 357-366.

Marchant, B.P., Lark, R.M., 2007a. Robust estimation of the variogram by residual maximum likelihood. Geoderma 140, 62-72.

Marchant, B.P., Lark, R.M, 2007b. The Matern variogram model: implications for uncertainty propagation and sampling in geostatistical surveys. Geoderma 140, 337-345.

Matheron, G., 1965. Les variables regionalisees et leur estimation: une application de la theorie des fonctions aleatoires aux sciences de la nature. Masson, Paris, 306.

Mathworks, 2011. Matlab 2011 version a. The Mathworks Inc., Natick, MA.

McBratney, A.B., 1998. Some considerations on methods for spatially aggregating and disaggregating soil information. Nutrient Cycling in Agroecosystems 50, 51-62.

McBratney, A.B, Webster, R. 1981. Spatial dependence and classification of the soil along a transect in Northeast Scotland. Geoderma 26, 63-82.

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, 293-327.

McBratney, A.B., Minasny, B., Cattle, S.R., Vervoort, R.W., 2002. From pedotransfer functions to soil inference systems. Geoderma 109, 41-73.

McBratney, A.B., Mendonca Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3-52.

Meentemeyer, V., 1989. Geographical perspectives of space, time, and scale. Landscape Ecology 3, 163-173.

Mendicino, G., Sole, A., 1997. The information content theory for the estimation of the topographic index distribution used in topmodel. Hydrological Processes 11, 1099-1114.

Mendonca-Santos, M.L., McBratney, A.B., Minasny, B., 2007. Soil Prediction with Spatially Decomposed Environmental Factors. In: Lagacherie, P., McBratney, A.B., Voltz, M., Digital Soil Mapping: An Introductory Perspective. Elsevier, Amsterdam, 269-278.

Milne A.E., Lark R.M., 2009. Wavelet transforms applied to irregularly sampled soil data. Mathematical Geoscience 41, 661-678.

Milne, G. 1934. Some suggested units of classification and mapping particularly for east African soils. Soil Research 4, 183-198.

Minasny, B., McBratney, A.B., 2007. Spatial prediction of soil properties using EBLUP with the Matérn covariance function. Geoderma 140, 324-336.

Minasny, B., McBratney, A.B., Salvador-Blanes, S., 2008. Quantitative models for pedogenesis - A review. Geoderma 144, 140-157.

Misiti, M., Misiti, Y., Oppenheim, G., Poggi, J.M., 2012. Wavelet Toolbox User's Guide, The MathWorks Inc.

Moran, J.M., Bui, E.N., 2002. Spatial data mining for enhanced soil map modelling. International Journal of Geographical Information Science 16, 533-549.

Nanos, N., Rodriguez, J.A., 2012. Using a Multi-Scale Geostatistical Method for the Source Identification of Heavy Metals in Soils. In: Panagiotaras, D., Geochemistry - Earth's System Processes, InTech, 323-346

Nisbet, R., Elder, J., Miner, J., 2009. Handbook of Statistical Analysis and Data Mining Applications, Elsevier, Burlington, 864.

Oksanen, J., Sarjakoski, T., 2006. Uncovering the statistical and spatial characteristics of fine toposcale DEM error. International Journal of Geographical Information Science 20, 345-369.

O'Neill, V., 2009. OSI LiDAR datasets. In: 3rd Annual Irish Earth Observation Symposium. Dublin.

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

Papritz, A., Herzig, C., Borer, F., Bono, R., 2005. Modelling the spatial distribution of copper in the soils around a smelter in north-eastern Switzerland. In: Renard, P., Demougeot-Renard, H., Froidevaux, R., Geostatistics for environmental applications. Springer-Verlag, Heidelberg, 343-354.

Patterson, H.D, Thomson, R, 1971. Recovery of inter-block information when block sizes are unequal. Biometrika 58, 545-554.

Pebesma, E.J, Wesseling, C.G, 1998. Gstat: a program for geostatistical modelling, prediction and simulation. Computers & Geosciences 24, 17-31.

Pennock, D.J., Veldkamp, A., 2006. Advances in landscape-scale soil research. Geoderma 133, 1-5.

Pennock, D., Yates, T., Braidek, J., 2008. Soil sampling designs. In: Carter, M.R, Gregorich, E.G., Soil sampling and methods of analysis. Boca Raton: Canadian Soil Science Society, Taylor and Francis, 1-14.

Pike, R.J., Rozema, W.J., 1975. Spectral analysis of landforms. Annals of the Association of American Geographers 82, 1079-1084.

Preston, R., Mills, P., 2002. Generation of a Hydrologically corrected Digital Elevation Model for the Republic of Ireland. EPA Report.

Quattrocchi, D., Goodchild, M., 1997. Scale in remote sensing and GIS. Lewis Publishers, Boca Raton, 423.

R Development Core Team, 2011. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org (last verified September 2013).

Ramadam, Z., Hopke, P.K., Johnson, M.J., Scow, K.M., 2005. Application of PLS and back-propagation neural networks for the estimation of soil properties. Chemometrics and Intelligent Laboratory Systems 75, 23-30.

Reinds, G.J., Van Oijen, M., Heuvelink, G.B.M, Kros, H., 2008. Bayesian calibration of the VSD soil acidification model using European forest monitoring data. Geoderma 146, 475-488.

Roecker, S.M., Thompson, J.A., 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.

Rosas-Orea, M.C.E., Hernandez-Diaz, M., Guerrero-Ojeda, L., 2005. A comparative simulation study of wavelet based denoising algorithms. International Conference on Electronics, Communications, and Computers 2005, 125-130.

Rossiter, D.G., 2003. Methodology for Soil Resource Inventories, ITC, Enshede 110.

Sanchez, P.A., Ahamed, S., Carre, F., Hartemink, A.E., Hempel, J., Huising, J., Lagacherie, P., McBratney, A.B., McKenzie, N.J., Mendonca-Santos, M.L.,

Minasny, B., Montanarella, L., Okoth, P., Palm, C.A., Sachs, J.D., Shepherd, K.D., Vagen, T.G., Vanlauwe, B., Walsh, M.G., Winowiecki, L.A., Zhang, G.L., 2009. Digital Soil Map of the World. Science 325, 680-681.

Sarkar, N., Chaudhuri, B.B., 1992. An efficient approach to estimate fractal dimension of textural images. Pattern Recognition 25, 1035-1041.

Schaetzl, R.J, Anderson, S., 2005. Soils: Genesis and Geomorphology. Cambridge University Press. New York, 817.

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

Scull, P., Franklin, J., Chadwick, O.A., 2005. The application of classification tree analysis to soil type prediction in a desert landscape. Ecological Modelling 181, 1-15.

Shannon, C.E, 1948. The mathematical theory of communication. Bell System Technology Journal 27, 379-423.

Sharma, A., Tiwari, K.N., Bhadoria, P.B.S., 2011. Determining the optimum cell size of digital elevation model for hydrologic application. Journal of Earth System Science 120, 573-582.

Shekhar, S., Xiong, H., 2008. Encyclopedia of GIS. Springer, New York, 1370.

Si, B.C., Farrell, R.E., 2004. Scale dependent relationships between wheat yield and topographic indices: A wavelet approach. Soil Science Society of America Journal 68, 577-588.

Si, B.C., 2007. Spatial Scaling Analyses of Soil Physical Properties: A Review of Spectral and Wavelet Methods. Vadose Zone Journal 7, 547-562.

Sinowski, W., Auerswald, K., 1999. Using relief parameters in a discriminant analysis to stratify geological areas with different spatial variability of soil properties. Geoderma 89, 113-128.

Smith, B., Mark, D.M., 2003. Do mountains exist? Towards an ontology of landforms. Environment and Planning B: Planning and Design 30, 411-427.

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

Somaratne, S., Seneviratne, G., Coomaraswamy, U., 2005. Prediction of soil organic carbon across different land-use patterns: a neural network approach. Soil Science Society of America Journal 69, 1580-1589.

StatSoft, 2010. Statistica version 10. Statsoft Inc., Tulsa, OK.

Stokes, C.R., Spagnolo, M., Clark, C.D., 2011. The composition and internal structure of drumlins: Complexity, commonality, and implications for a unifying theory of their formation. Earth-Science Reviews 107, 398-422.

Stoy, P., Williams, M., Spadavecchia, L., Bell, R., Prieto-Blanco, A., Evans, J., Van Wijk, M., 2009. Using information theory to determine optimum pixel size and shape for ecological studies: aggregating land surface characteristics in Arctic ecosystems. Ecosystems 12, 574-589.

Sun, W., Xu, G., Gong, P., Liang, S., 2006. Fractal analysis of remotely sensed images: a review of methods and applications. International Journal of Remote Sensing 27, 4963-4990.

Taud, H., Parrot, J.F, 2005. Measurement of DEM roughness using the local fractal dimension. Geomorphologie: Relief, Processus, Environnement 10, 327-338.

Thompson, A.J., Bell, J.C., Butler, C.A., 2001. Digital elevation model resolution: effects on terrain attribute calculation and quantitative soil-landscape modeling. Geoderma 100, 67-89.

U.S. Department of Agriculture. Soil Conservation Service. Soil Survey Staff. 1990. Keys to Soil Taxonomy. Virginia Polytechnic Institute and State University, Blacksburg, 422.

Vieux, B.E., 1993. DEM aggregation and smoothing effects on surface runoff modeling. Journal of Computing in Civil Engineering 7, 310-338.

Vieux, B.E., Farajalla, N.S., 1994. Capturing the essential spatial variability in distributed hydrological modelling: hydraulic roughness. Hydrological Processes 8, 221-236.

Vink, A., 1975. Land Use in Advancing Agriculture, X. Springer, New York, 394.

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

Visvalingam, M., 1990. Trends and concerns in digital cartography, Computer-Aided Design 22, 115-130.

Wang, G., Gertner, G., Parysow, P., Anderson, A., 2001. Spatial prediction and uncertainty assessment of topographic factor for revised universal soil loss equation using digital elevation models. ISPRS Journal of Photogrammetry and Remote Sensing 56, 65-80.

Webster, R., 2000. Is soil variation random? Geoderma 97, 149-163.

Webster, R., Oliver, M.A., 2007. Geostatistics for Environmental Scientists. John Wiley & Sons, Chichester, 330.

Wilson, J.P., Gallant, J.C., 2000. Terrain Analysis, Principles and Applications. John Wiley & Sons, Chichester, 479.

Wise, S., 2012. Information entropy as a measure of DEM quality. Computers & Geosciences 48, 102-110.

Wood, J.D., 1996. The geomorphological characterization of Digital Elevation Models. PhD thesis. University of Leicester, United Kingdom.

Wood, W.F., Snell, J.B., 1957. The dispersion of geomorphic data around measures of central tendency and its application. US Army Quartermaster Research and Development Center, Research Study Report EA-8.

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

Young, I.M., Crawford, J.W., Nunan, N., Otten, W., and Spiers, A., 2008. Microbial distribution in soils: physics and scaling, Advances in Agronomy 100, 81-121.

Zevenbergen, L.W., Thorne, C.R., 1987. Quantitative analysis of land surface topography. Earth surface Processes and Landforms 12, 47-56.

Zhao, Z., Chow, T.L., Rees, H.W., Yang, Q., Xing, Z., Meng, F.R., 2009. Predict soil texture distributions using an artificial neural network model. Computers and Electronics in Agriculture 65, 36-48.

Zhu, A.X., 2000. Mapping soil landscape as spatial continua: The neural network approach. Water Resource Research 36, 663-677.

Zhu, A.X., Burt, J., Smith, M., Wang, R., Gao, J., 2008. The impact of neighbourhood size on terrain derivatives and digital soil mapping. In Zhou, Q., Lees, B., Tang, G.A. Advances in digital terrain analysis. Springer, Berlin, 333-348.

Zurek, M.B., Henrichs, T., 2007. Linking scenarios across geographical scales in international environmental assessments. Technological Forecasting and Social Change 74, 1282-1295.

Appendix A – Dissemination of the results

The initial findings presented in this thesis have been published/presented at the following forums:

JOURNAL PAPER

S. Cavazzi, R. Corstanje, T. Mayr, J. Hannam, R. Fealy. (2013). “Are fine resolution Digital Elevation Models always the best choice in digital soil mapping?” Geoderma 196, 111-121.

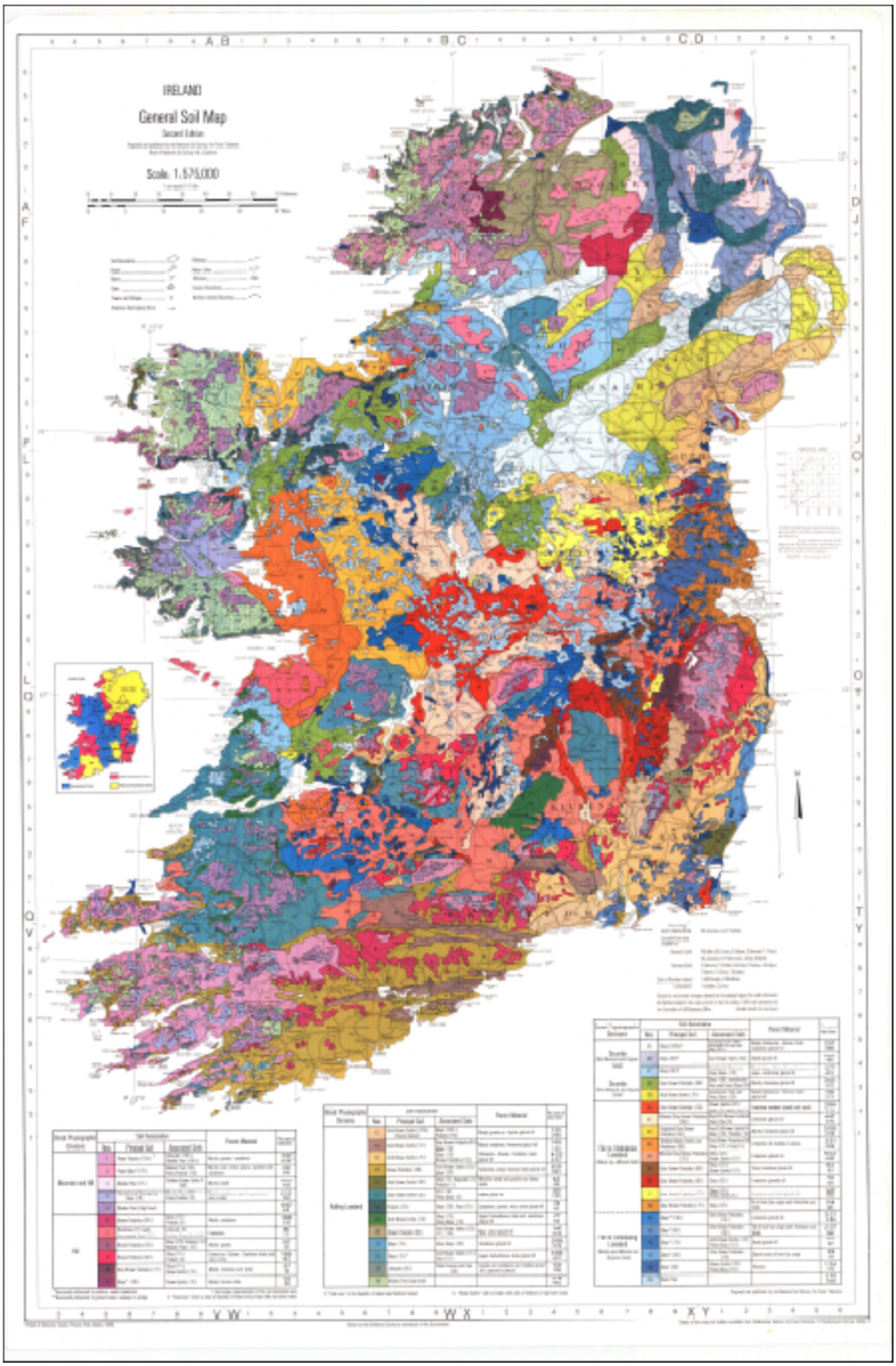
CONFERENCE PAPER

S. Cavazzi. (2012) “Spatial Scale in Digital Soil Mapping” Proceedings 1st AGILE PhD School.

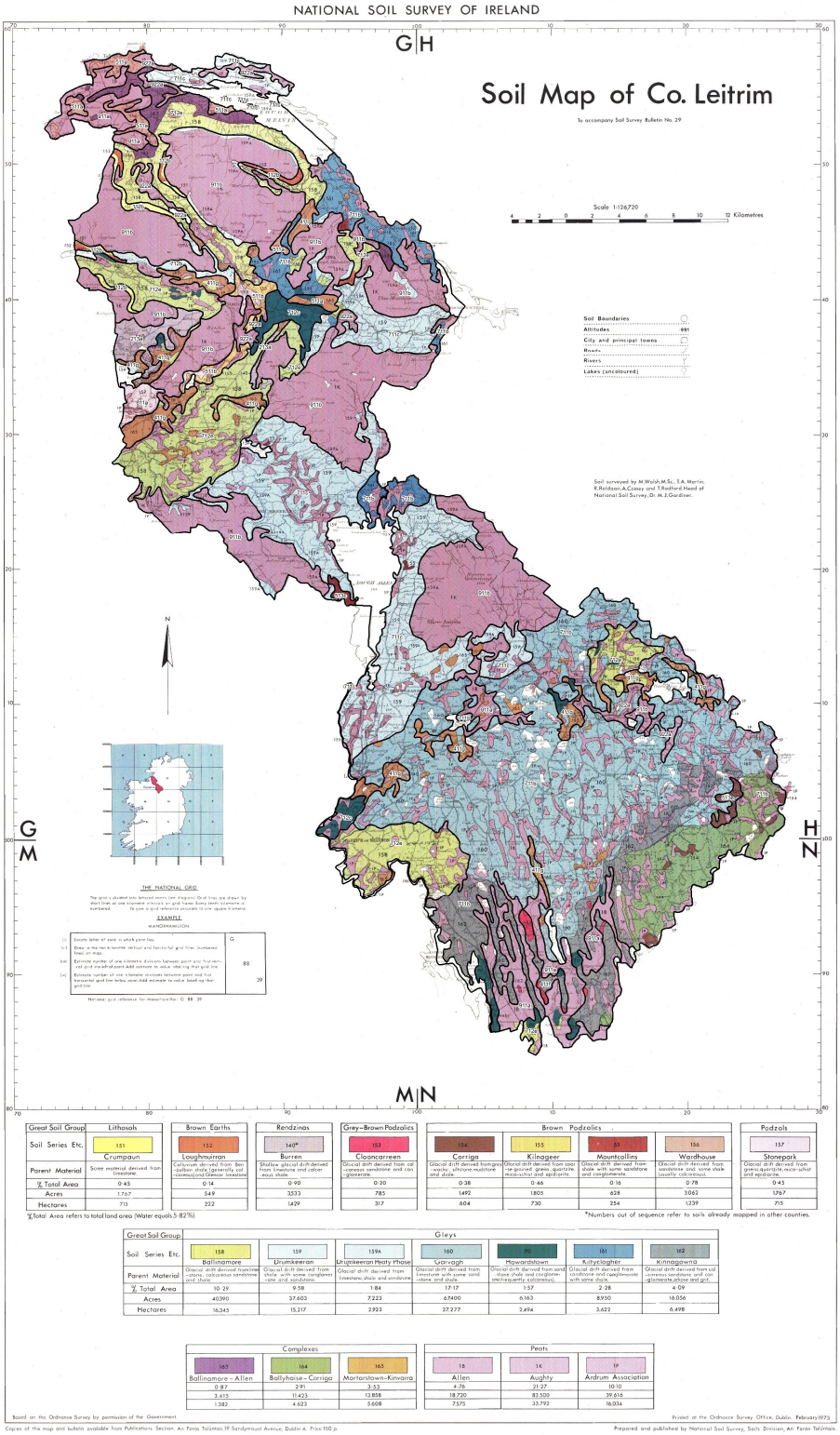
CONFERENCE POSTER

S. Cavazzi, R. Corstanje, T. Mayr. (2011) “Spatial scale analysis in DSM using machine learning techniques” Pedometrics.

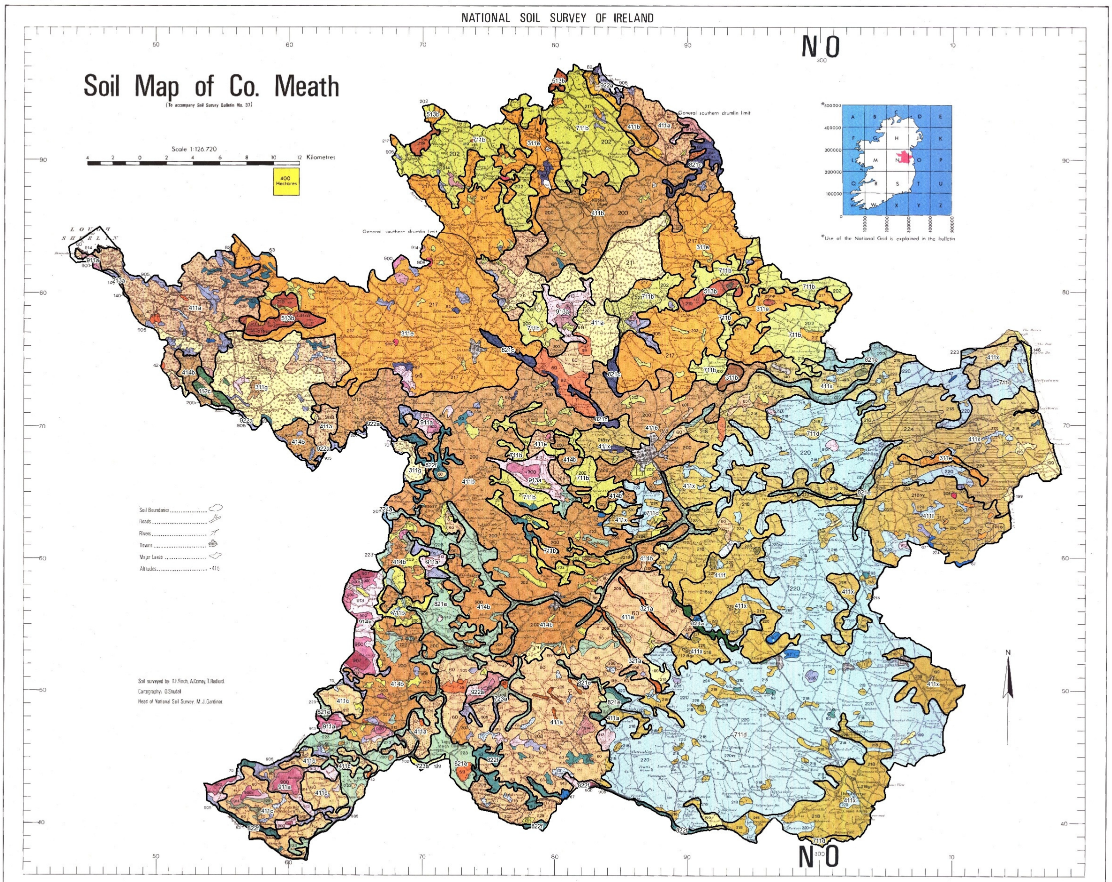
Appendix B – General soil map



Appendix C – County soil maps



NATIONAL SOIL SURVEY OF IRELAND



Great Soil Group	Regosol	Lithosol	Brown Earth					Rendzina	Grey Brown Podzolic									Brown Podzolic		Podzol
Soil Series	146 Seafield	204 Knockeyon	42 Baggotstown	43 Ballincurra	44sy Derk Shaley Phase	217 Kells	199 Ladestown	140 Burren	218 Dunboyne	218gv Dunboyne Gravelly Phase	218sy Dunboyne Shaley Phase	59 Elton	90 Mortarstown	60 Patrickswell	60L Patrickswell Lithic Phase	200 Rathowen	200c Rathowen Cherty Phase	219 Rathkenny	219sp Rathkenny Moderately Steep Phase	222 Slievebeag
Parent Material	Stabilised Sand	Chert Bedrock	Fluvioglacial Limestone	Limestone Till over Limestone Bedrock	Glacial drift, dominantly volcanic; with some Limestone, Shale and Sandstone	Drift from Lower Palaeozoic Shale	Fluvioglacial Shale, Chert and Limestone	Limestone Bedrock	Limestone and shale drift and Irish sea drift	Limestone and shale drift and Irish sea drift	Limestone and shale	Limestone Till	Limestone Till	Limestone Till	Limestone Till	Shale and Limestone Till	Shale Limestone and Chert Till	Glacial drift of Lower Palaeozoic Shale	Glacial drift of Lower Palaeozoic Shale	Lower Palaeozoic Shale
% Total Land Area	·046	·129	·073	·007	·099	11·814	·492	·085	10·285	·111	3·347	·617	·151	7·124	·084	13·904	·037	·354	·272	·017
Hectares	107	300	169	17	231	27,544	1,147	198	23,959	259	7,803	1,438	352	16,594	195	32,405	86	825	635	40

Urban= ·317%

Great Soil Group	Gley									
Soil Series	220 Ashbourne	220sy Ashbourne Shaley Phase	201 Ballyshear	63 Camoge	67 Drombanny	221 Dunsany	82 Feale	70 Howardstown	129 Mylerstown	202 Street
Parent Material	Limestone and shale drift and Irish sea drift	Limestone and shale drift and Irish sea drift	Gravelly Limestone Till	River Alluvium	Peaty Lake Alluvium	Alluvium of Namurian shale and limestone origin	River Alluvium	Limestone Till	Gravelly Limestone Till	Shale and Limestone Till
% Total Land Area	17·340	·122	·039	1·152	·140	·098	·641	·232	·179	8·720
Hectares	40,402	285	93	2,686	327	229	1,494	543	418	20,312

Peats

	Raised Bog	Raised Bog (cutover-reclaimed)	Fen Peat	Fen Peat Improved
Soil Series	900 Allen	913 Gortnamona	908 Pollardstown	905 Banagher
% Total Land Area	·806	1·635	·025	1·488
Hectares	1,877	3,812	59	3,471

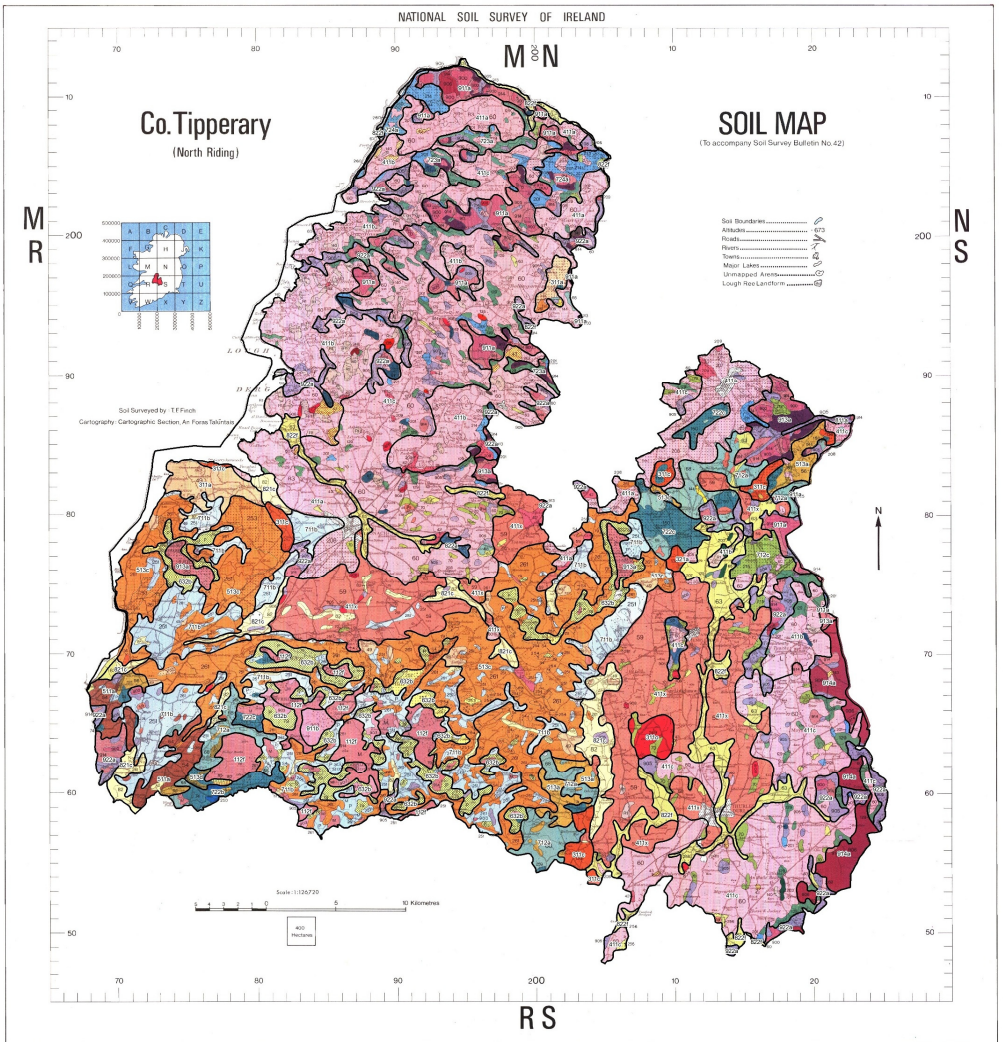
Peat Complexes

Industrial Sod Peat	Cutover Raised Bog
907 Clonsast	914 Turbary
·102	·612
238	1,427

Complexes

205 Patrickswell-Ladestown	208 Patrickswell-Baggotstown-Elton	209 Baggotstown-Crush	211 Ladestown-Rathowen-Banagher	212 Rathowen-Ladestown	223 Alluvial (Boyne Alluvium)	224 Dunboyne-Ashbourne	224sy Dunboyne-Ashbourne Shaley Phase	225 Dunboyne-Ladestown	226 Rathowen-Street	227 Street-Ladestown-Banagher	228 Glane	229 Patrickswell-Howardstown
·449	4·262	·141	3·756	1·507	4·438	1·564	·088	·078	·252	·295	·101	·339
1,043	9,929	328	8,757	3,518	10,346	3,644	205	183	588	687	235	790

Prepared and published by the National Soil Survey, An Foras Talúntais.
Based on the Ordnance Survey by permission of the Government. Printed at Ordnance Survey, Phoenix Park, Dublin. 1982
Copies of this map and bulletin available from Publications Section, An Foras Talúntais, 19 Sandymount Avenue, Dublin 4.



Great Soil Group	Regosol		Lithosol	Brown Earth						Rendzina			Grey Brown Podzolic				Brown Podzolic			
Soil Series	248 Carney	249 Milltown–pass	141 Slievereagh	42 Baggots–town	43 Ballincurra	49 Ballyna–lacken	248 Dovea	135 Kinvarra	136 Knockna–skeha	140 Burren	R1 Burren Rocky Phase	141 Kilcolgan	59 Elton	60 Patrickswell	L Patrickswell Lithic Phase	R3 Patrickswell Bouldery Phase	261 Borrisoleigh	34 Borrisoleigh Steep Phase	54 Cooga	56 Doonglara
Parent Material	Limestone Lake Alluvium	Limestone River Alluvium	Sandstone Mainly Bedrock	Limestone Fluvioglacial Drift	Shallow Limestone Till	Shale Till or Colluvium	Limestone/ Shale Till	Limestone Till	Sandstone/ Limestone/ Shale Till	Mainly Limestone Bedrock		Limestone Till	Limestone Till, some Sandstone and Shale				Shale Till and Colluvium		Sandstone Fluvioglacial Drift,some Shale	Sandstone Till or Colluvium
% Total* Land Area	·01	·06	·01	·31	·24	1·26	·34	·04	·64	·17	·92	·03	9·85	23·08	1·34	1·22	10·91	·40	·94	1·16
Hectares	22	129	20	621	486	2,527	682	79	1,284	344	1,834	56	19,650	46,055	2,745	2,444	21,772	807	1,881	2,314

*Urban areas etc.=1062ha 0·53 %

Great Soil Group	Podzol					Gley													
Soil Series	77 Knockaceol	P Knockaceol Peaty Phase	79 Knock–astanna	P Knock–astanna Peaty Phase	116 Seefin	201 Ballyshear	63 Camoge	66 Coolalough	250 Derry–garreen	67 Drombanny	82 Feale	68 Gortaclareen	70 Howards–town	251 Kilcommon	P Kilcommon Peaty Phase	144 Kilgory	129 Mylerstown	74 Puckane	P Puckane Peaty Phase
Parent Material	Sandstone Bedrock some Till and Colluvium		Shale Till and Colluvium		Sandstone Colluvium and Till	Limestone Till	Limestone/ Sandstone/ Shale–River Alluvium	Limestone Lake Alluvium	Sandstone Fluvioglacial Drift	Limestone Lake Alluvium	Shale River Alluvium	Sandstone/ Shale Till	Limestone Till	Shale Till and Colluvium		Sandstone River Alluvium	Limestone Till with some Shale	Sandstone/ Shale Till	
% Total Land Area	·04	·03	1·66	1·84	·02	·47	3·72	·15	·05	·26	2·40	2·38	·95	6·15	1·16	·03	2·15	·27	·03
Hectares	78	54	3,315	3,685	40	942	7,429	298	103	513	4,790	4,749	1,902	12,279	2,309	54	4,297	533	63

	Peat						
Soil Series	900 Allen	903 Aughty	905 Banagher	906 Boora (Complex)	908 Pollardstown	909 Gortnamona	914* Turbary (Complex)
% Total Land Area	1·63	·53	4·85	1·47	·18	·79	4·74
Hectares	3,248	1,056	9,686	2,935	367	1,572	9,469

R1 2–10% Rockiness
R3 25–50% Rockiness
S4 13°–19° slopes
L Lithic
M Mountainous
P Peaty
Large rock outcrop

	Complex																	
Soil Series	252 Turbary Knock–astanna	209 Baggotstown Crush	253 Borrisoleigh Ballynalacken	254 Borrisoleigh Knockshigowna	255 Ballynalacken Knockshigowna	214 Ballyshear Patrickswell	256 Camoge Milltownpass	213 Howardstown Baggotstown	215 Howardstown Patrickswell	257 Knockastanna Knockshigowna	258 Knocknaskeha Doonglara	207 Patrickswell Baggotstown	208 Patrickswell Baggotstown Elton	259 Puckane Gortaclareen	150 Puckane Slievereagh	260 Shannon Banagher
% Total Land Area	·45	·24	·90	·36	·23	·56	·21	·07	·41	·29	·26	1·20	1·99	·20	1·11	·18
Hectares	901	488	1,799	719	456	1,115	415	133	835	589	524	2,400	3,970	389	2,217	362

Prepared and published by National Soil Survey, An Foras Talúntais
Printed at Ordnance Survey, Phoenix Park, Dublin 1987
Based on the Ordnance Survey by permission of the Government