EPA STRIVE Programme 2007-2013

Predictive Mapping

(2007-S-CD-1-S1)

ISIS Final Technical Report 5

Prepared for the Environmental Protection Agency
by
Teagasc and Cranfield University

Authors:

Mayr, T.R., Zawadzka, J.E., Corstanje, R., Lo Papa, Fealy, R.

ENVIRONMENTAL PROTECTION AGENCY
An Ghníomhaireacht um Chaomhnú Comhshaoil
PO Box 3000, Johnstown Castle, Co.Wexford, Ireland

Telephone: +353 53 916 0600 Fax: +353 53 916 0699
Email: info@epa.ie Website: www.epa.ie

ACKNOWLEDGEMENTS

This report is published as part of the Science, Technology, Research and Innovation for the Environment (STRIVE) Programme 2007-2013. The programme is financed by the Irish Government under the National Development Plan 2007-2013. It is administered on behalf of the Department of the Environment, Heritage and Local Government by the Environmental Protection Agency which has the statutory function of co-ordinating and promoting environmental research.

This project involved many staff and students based at Teagasc Johnstown Castle, Teagasc Athenry, Cranfield University and University College Dublin. Many thanks go to all the staff, PhD students, placement students and FAS placements who contributed to this project. We also wish to thank the Steering Committee, Dr Alice Wemaere (EPA), Ms Niamh Connolly (EPA), Dr Mark Gibson (Teagasc), Prof Peter Loveland, Dr Erika Micheli and Dr Arwyn Jones.

DISCLAIMER

The EPA STRIVE Programme addresses the need for research in Ireland to inform policymakers and other stakeholders on a range of questions in relation to environmental protection. These reports are intended as contributions to the necessary debate on the protection of the environment.

EPA STRIVE PROGRAMME 2007-2013

Published by the Environmental Protection Agency, Ireland

Details of Project Partners

Teagasc - Environment, Soils and Land-use Centre,
Johnstown Castle, Wexford, County Wexford, Ireland.
Tel.: +353 (0)53 9171200
E-mail: rachel.creamer@teagasc.ie

Teagasc – Spatial Analysis Group, Rural Economy and Development Programme.
Ashtown, Dublin 15, Ireland.
Tel.: +353 (0)1 8059953
E-mail: reamonn.fealy@teagasc.ie

Cranfield University - National Soil Resources Institute
Cranfield Campus, Cranfield, Bedfordshire, MK43 0AL, UK
Tel.:+44 (0)1234 750111
Email: t.mayr@cranfield.ac.uk

University College Dublin
School of Biosystems Engineering, Agriculture and Food Science,
Belfield, Dublin 4, Ireland.
Tel.: +353 (0)1 7162808
Email: nick.holden@ucd.ie

Executive Summary

The Irish Soil Information System (ISIS) project was established in 2008, following a comprehensive inventory of Irish soil data compiled by Daly and Fealy (2007) which highlighted that soil data coverage of Ireland was incomplete in both detail and extent. The ISIS project is funded under the Environmental Protection Agency STRIVE Research Programme 2007-2013 and co-funded by Teagasc. It was led by Teagasc with the participation of researchers from Cranfield University (UK) and University College Dublin. The overall objective of the ISIS project was to conduct a programme of structured research into the national distribution of soil types and construct a soil map, at 1:250,000 scale, which will identify and describe the soils according to a harmonised national legend. This map is now available in digital format and forms the basis of a new soil information system for Ireland (http://isis.teagasc.ie).

The ISIS project has utilised existing data and maps from the previous National Soil Survey (NSS) conducted by An Foras Talúntais (forerunner organisation to Teagasc). The NSS produced: mapping at 1:126,720 scale for 44% of the country; a General Soil Map of Ireland and a National Peatland map, both at 1:575,000 scale and other miscellaneous large scale mapping of experimental farms. In addition, more recent map products have been included such as the Indicative Soil and Subsoil mapping (Fealy and Green, 2009) with national coverage using GIS and remote sensing techniques.

Comparison of soil information at European scale has led to the requirement for the harmonisation and coordination of soil data across Europe, and, in light of the demands for soil protection on a regional basis within member states there is a growing need to support policy with a harmonised soil information system. The European Soil Bureau Network (ESBN) Technical Working Group dealing with Soil Monitoring and Harmonisation recommended a soil map of Europe at a scale of 1:250,000 as an economically feasible intermediate scale that can identify specific problems at regional scale (Montanarella and Jones, 1999).

The ISIS project adopted a combined methodology of utilising novel predicted mapping techniques in tandem with traditional soil survey applications. This unique combination at a national scale has resulted in the development of a new national soil map for Ireland. Building upon the detailed work carried out by the An Foras Talúntais (AFT) survey (known as *Terra Cognita*), the ISIS project generated soil-landscape models at a generalised scale of 1:250,000 for the counties of Carlow, Clare, Kildare, Laois, Leitrim, Limerick, Meath, Offaly, Tipperary South, Waterford, Westmeath, Wexford, West Cork, West Mayo and West Donegal. These soil-landscape models (also referred to as soilscapes) were used as the baseline data for statistical models (random forests, Bayesian belief networks and neural networks) to predict soil map units in counties where there was no map available (referred to as *Terra Incognita*). To validate the methodology, this work was supported by a 2.5 year field survey, in which 11,000 locations were evaluated for soil type, using an auger bore survey approach. These data were used to check the predicted soil mapping units (associations) for counties: Cavan, Dublin, East Cork, East Donegal, East Mayo, Galway, Kerry, Kilkenny, Louth, Monaghan, Roscommon, Sligo, Tipperary South and Wicklow, where a detailed soil

survey map was not available. Where new soil information was generated, due to previously unknown combinations of soil-landscape units, profile pits were selected at representative locations across the country. These 225 pits were described and sampled in detail and were used to generate a new soil classification system for the country. The final product is a unique combination of new and traditional methodologies and soils data from both the AFT and the ISIS project. The final, soil association map of Ireland consists of 58 associations (excluding areas of alluvium, peat, urban, rock or marsh) that are made up from 213 soil series. Associated representative profile information is available in the online soil information system.

A key component of the ISIS project has been the development of a soil and land information system and associated public web site. This system has been designed to hold the complete set of information deriving both from the ISIS field programme and modelling activity, as well as the previously existing legacy soils information available for Ireland. Drawing on this information system, the web site is designed to hold and disseminate this information online both in cartographic and tabular form to stakeholders. Prior to this development, there has been no harmonised computerised system in place to hold and manipulate national Irish soils data. The information system therefore addresses the pressing need and requirement for a publicly-accessible, integrated IT framework based upon contemporary informatics standards to serve the many and varied stakeholders having an interest in soils information in Ireland.

Technical Note on Soil Classification

Two Irish soil classification systems were developed during the ISIS project. An **Interim Soil Classification** was developed in the early stages of the project to enable the harmonisation and generalisation of the county soil maps published by An Foras Talúntais (AFT) and the rationalisation of the original AFT soil series. The **Interim Soil Classification** was used during the development of Work Packages (WP): WP1 and WP2, to produce the training data for the predictive mapping and for most of the field programme in WP3. In 2013/4, **the Interim Soil Classification** was modified following a World Reference Base style hierarchical approach that recognises Great Soil Groups and defines sub-groups by supplementary diagnostic horizons. The **Final Soil Classification** System was developed to provide a more user-friendly classification system that adopts the approach of a hierarchical key for recognition of Great Soil Groups and diagnostic horizons to define the sub-groups.

The **Final Soil Classification** System was subsequently implemented during the description of representative soil profiles, final map production and is included in the updated soil profile handbook, and national soil series list. This modified system is the **Final Soil Classification** system for Ireland that appears in the map and associated information system on the ISIS website.

This Final Technical Report was developed using the **Interim Soil Classification**, and describes a significant contribution to the production of the final New Soil Map of Ireland. Table B below details the differences between the **Interim** and the **Final Soil Classification** Systems.

The **Final Soil Classification** System for Ireland has 3 hierarchical levels:

1. Great Soil Groups:

The classification criteria for the Great Soil Groups (GSG) were based on recognisable features used by An Foras Talúntais (National Soil Survey of Ireland) to classify the soils of Ireland at Great Soil Group level. Table A provides an overview of the key criteria for recognizing the Great Soil Groups. The sequence follows World Reference Base (WRB) principles.

2. Soil Sub-groups:

The Irish Soil Classification of soil sub-groups (SSG) is based on the recognition of diagnostic horizons, properties and materials which, where possible, should be observed and measured in the field. The selection of diagnostic characteristics takes into account their relationship with soil forming processes. Diagnostic features are selected that are significant to soil management. Subgroups are named with a maximum of two diagnostic features that represent the most important processes occurring in the soil profile. Table B provides a look-up table between the interim and the modified classification systems, listing the Great Soil Groups and Sub-groups.

3. Soil Series

The classification of series is based on the same principles as the interim classification system. Within a sub-group a series is further defined by the nature of the soil texture and parent material.

4. Soil Associations

For mapping purposes, the soil series are combined to form soil associations that are identified by the most frequently occurring soil series and combinations of ancillary series. Each association is named after the key (lead) soil series, which is the most extensive soil in the association, e.g. Kilrush series is the dominant component in the Kilrush Association. To facilitate mapping, each soil association based on the Interim Classification is assigned an alphanumeric code that comprises the soil subgroup code (numeric) concatenated with a single alphabetic character, e.g. 711b for Kilrush Association. In the Final Soil Classification, the Kilrush Association is assigned the code 0700b in accordance with Tables A and B. With respect to classification terminology, the reports (3, 4, 5, 11 & 12) describing the predictive mapping programme refer only to soil association codes that relate to the Interim Soil Classification. However, the ISIS Soil Information System contains a translation table that links the interim soil association codes to the codes that relate to the Final Soil Classification. Thus the results of the predictive mapping can be linked to the final version of the New Soil Map of Ireland.

Table A: Sequencing of the Great Soil Groups (GSG) in the Final Irish Soil Classification

Criteria	GSG code	Great Soil Group (GSG)
Soils with thick organic layers	1	OMBROTROPHIC PEAT
	2	MINEROTROPHIC PEAT
Shallow or extremely gravelly soils	3	RENDZINAS
	4	LITHOSOLS
Soils influenced by water	5	ALLUVIAL SOILS
	6	GROUNDWATER GLEYS
	7	SURFACE-WATER GLEYS
Soils affected by Fe/Al chemistry increase	8	PODZOLS
	9	BROWN PODZOLICS
Soils with clay enriched subsoil	10	LUVISOLS
Relatively young or soils with limited profile development	11	BROWN EARTHS

For more details of the finalised Irish Soil Classification System please refer to the following documents:

ISIS Final Technical Report 10: Simo et al. (2014). The Irish Field Handbook for Soil Profile Descriptions. Available from http://erc.epa.ie.safer/reports

ISIS Final Technical Report 13: Simo et al. (2014). The Irish Soil Information System Map and Legend. Available from http://erc.epa.ie.safer/reports

ISIS Final Technical Report 9: Creamer et al. (2014). The Irish Soil Information System National Soil Series - Description and Classification of Representative Profiles. Available from http://erc.epa.ie.safer/reports

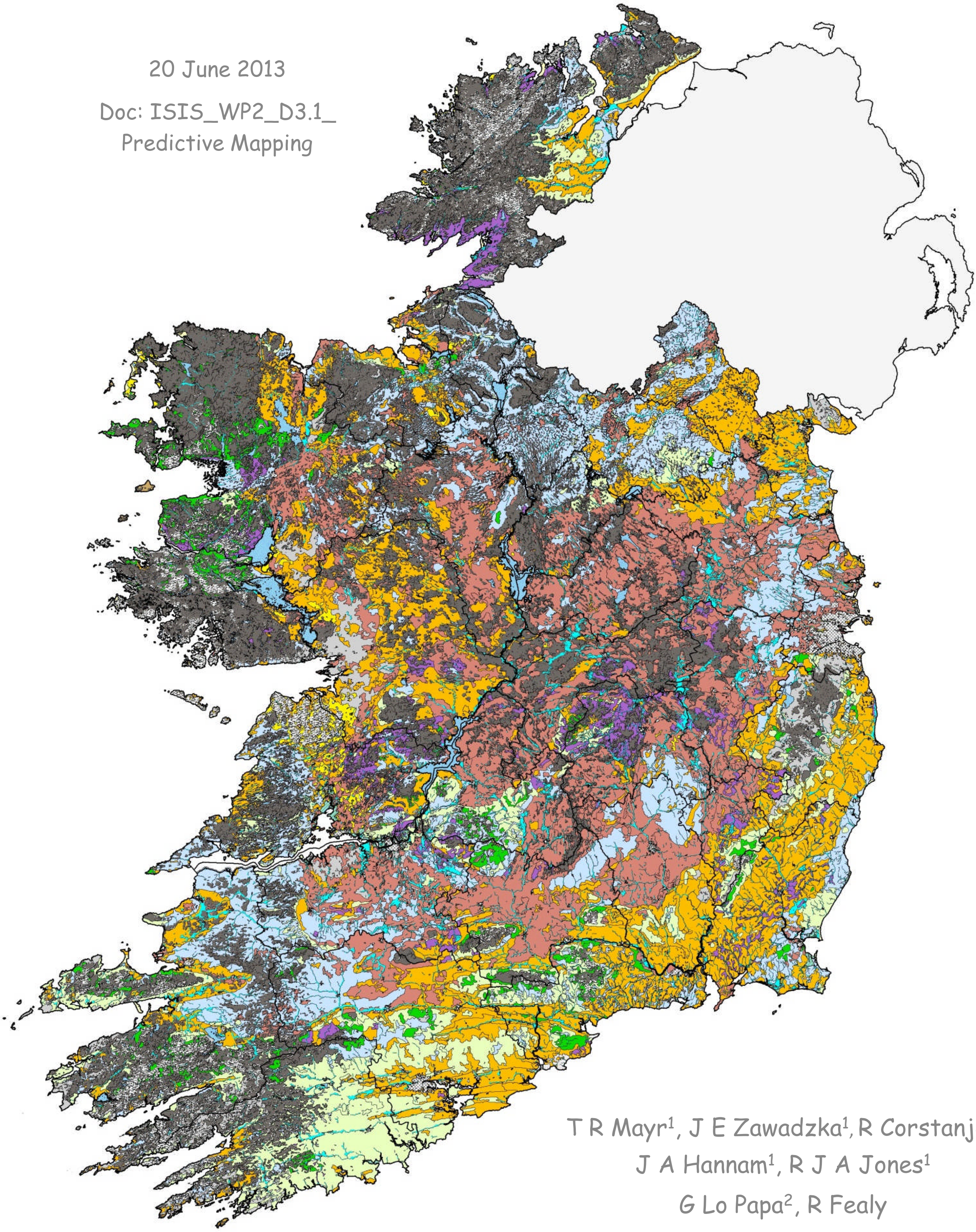
Table B Linkage between the Interim and Final Irish Soil Classifications for Soil Subgroups

Interim SSG_code	Interim Soil Subgroup (SSG)	SSG code	Soil Subgroup (SSG)
911	Raw Ombrotrophic Peat Soils	110	Natural Ombrotrophic Peat Soils
912	Earthy Ombrotrophic Peat Soils	170	Drained Ombrotrophic Peat Soils
913	Cut-over Ombrotrophic Peat Soils	180	Cut-over Ombrotrophic Peat Soils
914	Industrial Ombrotrophic Peat Soils	190	Industrial Ombrotrophic Peat Soils
921	Raw Minerotrophic Peat Soils	210	Natural Minerotrophic Peat Soils
922	Earthy Minerotrophic Peat Soils	270	Drained Minerotrophic Peat Soils
		280	Cut-over Minerotrophic Peat Soils
211	Typical Rendzinas	300	Typical Rendzinas
215	Histic Rendzinas	310	Histic Rendzinas
213	Humic Rendzinas	360	Humic Rendzinas
214	Stagnic Rendzinas		
212	Gleyic Rendzinas		
111	Typical Lithosols	400	Typical Lithosols
113	Histic Lithosols	410	Histic Lithosols
112	Humic Lithosols	460	Humic Lithosols
821	Typical Alluvial Gleys	500	Typical Alluvial Gley Soils
		510	Histic Alluvial Gley Soils
823	Typical Calcareous Alluvial Gleys	550	Typical Calcareous Alluvial Gley Soils
		551	Histic Calcareous Alluvial Gley Soils
824	Humic Calcareous Alluvial Gleys	556	Humic Calcareous Alluvial Gley Soils
822	Humic Alluvial Gleys	560	Humic Alluvial Gley Soils
811	Typical Brown Alluvial Soils	570	Typical Alluvial Soils
812	Gleyic Brown Alluvial Soils	572	Gleyic Alluvial Soils
813	Humic Brown Alluvial Soils	576	Humic Alluvial Soils
721	Typical Groundwater Gleys	600	Typical Groundwater Gleys
		610	Histic Groundwater Gleys
723	Calcareous Groundwater Gleys	650	Calcareous Groundwater Gleys
		651	Histic Calcareous Groundwater Gleys
724	Humic Calcareous Groundwater Gleys	656	Humic Calcareous Groundwater Gleys
722	Humic Groundwater Gleys	660	Humic Groundwater Gleys
		690	Anthropic Groundwater Gleys
711	Typical Surface-water Gleys	700	Typical Surface-water Gleys
712	Humic Surface-water Gleys	760	Humic Surface-water Gleys
		790	Anthropic Surface-water Gleys
611	Ferric Podzols	800	Typical Podzols
621	Typical Gley Podzols	820	Gleyic Podzols
622	Stagno-Gley Podzols	830	Stagnic Podzols
632	Iron-pan Stagno Podzols	843	Stagnic Iron-pan Podzols
612	HumoFerric Podzols	860	Humic Podzols
		890	Anthropic Podzols
631	Ferric Stagno Podzols		
511	Typical Brown Podzolics	900	Typical Brown Podzolics
512	Gleyic Brown Podzolics	920	Gleyic Brown Podzolics
514	Stagnic Brown Podzolics	930	Stagnic Brown Podzolics
		936	Humi-Stagnic Brown Podzolics
513	Humic Brown Podzolics	960	Humic Brown Podzolics
		990	Anthropic Brown Podzolics
411	Typical Luvisols	1000	Typical Luvisols
412	Gleyic Luvisols	1020	Gleyic Luvisols
		1026	Humi-Gleyic Luvisols
414	Stagnic Luvisols	1030	Stagnic Luvisols
		1036	Humi-Stagnic Luvisols
413	Humic Luvisols	1060	Humic Luvisols
1020	Technosols	1090	Anthropic Luvisols
311	Typical Brown Earths	1100	Typical Brown Earths
312	Gleyic Brown Earths	1120	Gleyic Brown Earths
		1126	Humi-Gleyic Brown Earths
314	Stagnic Brown Earths	1130	Stagnic Brown Earths
315	Humi-stagnic Brown Earths	1136	Humi-Stagnic Brown Earths
321	Typical Calcareous Brown Earths	1150	Typical Calcareous Brown Earths
322	Gleyic Calcareous Brown Earths	1152	Gleyic Calcareous Brown Earths
323	Stagnic Calcareous Brown Earths	1153	Stagnic Calcareous Brown Earths
		1156	Humic Calcareous Brown Earths
		1159	Anthropic Calcareous Brown Earths
313	Humic Brown Earths	1160	Humic Brown Earths
		1190	Anthropic Brown Earths
		1196	Humi-Anthropic Brown Earths

Predictive Mapping

20 June 2013

Doc: ISIS_WP2_D3.1_
Predictive Mapping



T R Mayr[1], J E Zawadzka[1], R Corstanje[1],
J A Hannam[1], R J A Jones[1]
G Lo Papa[2], R Fealy

[1]NSRI Cranfield University

[2]Teagasc

IRISH SOIL INFORMATION SYSTEM (ISIS)

Predictive Mapping

T.R. Mayr[1], J.E. Zawadzka[1], R. Corstanje[1], J.A Hannam[1], R.J.A. Jones[1]
G. Lo Papa[2], R. Fealy[2]

[1]National Soil Resources Institute
School of Applied Sciences
Cranfield University
Cranfield MK43 0AL, UK

[2]Teagasc
Agriculture and Food Research Agency
Kinsealy
Dublin, Ireland

Cranfield University
20 June 2013

Document: ISIS_WP2_D3.1_PredictiveMapping

Acknowledgement

This work is part of the Irish Soil Information System (ISIS) Project, managed by Teagasc (the Irish Agriculture and Food Development Authority) and co-funded by the Environmental Protection Agency (EPA) of Ireland through their Science, Technology, Research and Innovation for the Environment (STRIVE) Programme, as part of the National Development Plan 2007-2013.

Document: ISIS_WP2_D3.1_PredictiveMapping

This document should be cited as:

Mayr. T.R., Zawadzka, J.E., Corstanje, R., Hannam, J.A., Jones, R.J.A., Lo Papa, G. and Fealy, R. (2013). Predictive Mapping. ISIS Project Technical Report No.ISIS_WP2_D3.1, 78pp., Cranfield University and Teagasc

Tables

Figures

Appendices

1. Introduction

In the mid-1880s, Vasily Dokuchaev (1886) formulated the following hypothesis:

> *Any soil is always and everywhere a mere function of the following factors of soil formation: 1) nature (content and structure) of the parent rock, 2) the climate of the given terrain, 3) the mass and character of vegetation, 4) the age of the terrain, 5) the terrain topography. It immediately follows that (a) if the mentioned factors are the same in two different localities (however far apart they might be), the soils in the two localities should also be similar, and vice versa; consequently, b) if we have thoroughly studied these factors, we may predict in advance what the soil itself should be like. Next, (c) it is well know that momentum should not change if one force component increases or decreases by some value, while another force component changes by the opposite value; thus there should be a similar, to some extent, relationships between character of soil and character of its forming factors. Hence, it is clear that it is theoretically possible to state and solve, for example, the following problem: How would a given soil change if there is an increase of the terrain temperature by say 1-2degree, with a synchronous increase of meteoric water by 1-2 inches? How would soil change if there is an annual increase of the vegetative mass by 20 poods per desyatina with decrease of the temperature by 1-2degree?*

Later, Dokuchaev (1899) carried out the first step towards the formulisation of the problem. He proposed the first soil formation equation:

$\Pi = f\,(\text{К, О, Г})\ \text{В};$ $\qquad S = f\,(cl, o, p)\ a$

where П is soil К is climate, О is organism, Г is parent material, and В is age of the soil (topography was not included into the expression to a stenographer's mistake). Zakharov (1927) presented a general soil forming equation ideally describing the first law of soil science:

$\pi = f\,(\text{М.Г.П., Р.Ж.Орг., Кл., Возр.срр., Р- Ф})$ $\qquad S = f\,(p, o, cl, a, r)$

where π is soil, М.Г.П. is parent rock material, Р.Ж.Орг. is plant and animal organisms, Кл. is climate, Возр.срр. is the age of the terrain, and Р- Ф is topography.

Jenny (1941) adopted the Zakharov equation replacing Russian abbreviations with English ones:

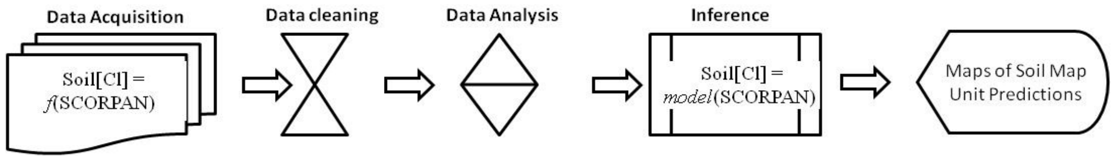
$$S = f(cl, o, r, p, t, \ldots)$$

where S is soil, cl, o, r, p, and t are soil-forming factors: climate, organisms, topography, parent material, and time, respectively. The ellipsis indicate that additional soil formers may be included.

McBratney (2003) proposed a Zakharov-like formulation for empirical quantitative descriptions of relationships between soil and other spatially distributed factors, using them as soil predictors:

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

where Sc is soil classes, Sa is a soil attributes, s is soil, other properties of the soil at a point; c is climate, local climatic properties; o is organisms, vegetation, fauna, and human activities; r is topography, morthometric variables; p is parent material, lithology; a is age, time; and n is space, spatial position. [Florinsky]



2. Inference engines

Three different inference engines were used for the predictive mapping, namely Belief Networks, Neural Networks and Random Forests. They were selected on the basis that they use three very different statistical concepts.

2.1 Bayesian Belief Networks

Belief networks are a vital tool in probabilistic modelling and Bayesian methods. They are one class of probabilistic graphical model. In other words they are a marriage between two important fields: probability theory and graph theory. It is this combination which makes them a powerful methodology within machine learning and statistics.

While statistical models are able to rank variable importance, they do not define the relationship between soil associations and the predictor variables, and are considered 'black-box' techniques. Bayesian belief networks represent a completely different approach to the statistical inference models used to date. BBNs are mathematical models which make predictions based on probabilities using a combination of measured data and expert opinion. They are of interest is that they can work well using limited datasets and offer a clear description of the relationships and interactions between predictor and target variables. Furthermore, they provide the means to formalise a soil surveyor's knowledge by into a set of rules and probabilities which can, in theory, be applied to regions beyond the study area.

The first stage of the modelling is to create a conceptual model diagram based on the previous modelling results and expert knowledge. The second stage is to allocate each node a set of values (these will be classes for both categorical data and the product of the discretization of continuous variables), Boolean where possible (Hough *et al.*, 2010). This relies on the discretization of continuous variables, one of the major challenges involved in BBN modelling, which is usually a decision made using expert knowledge (Uusitalo, 2007).

The next task is to define the relationships between variables using conditional probability, which is the probability of an event A given B. For example the probability of a certain soil type at a given location will be influenced by the presence of a certain parent material and climatic conditions, which needs to be accounted in the model structure. Here the prior conditional probability will be calculated using Bayes rule. For variables which are not influenced by others, such as rainfall, prior unconditional probabilities which is usually a combination of expert knowledge and observed data. BBN also allows a phase of 'structural learning' which shows, not only the variables influence soil associations most strongly, but also the effects variables have on predicting individual soil associations. The models were generated using NETICA and NETICA API software packages from Norsys Software Corporation. [Khaled Taalab]

2.3 Random Forests

Random Forest is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees. The algorithm for inducing a Random Forest was developed by Leo Breiman and Adele Cutler.

Breiman (2001) introduces RF modelling as a method of improving predictions made using a classification and regression tree approach. The key features of this method are that trees are constructed using a bootstrap of the entire dataset and the splits at each node are not made by the best predictor from the entire suite of input variables, but from the best of a randomly selected subset, which prevents overfitting (Liaw & Wiener, 2002). The performance of the model is assessed by predicting the mean square error (MSE) of the 'out of bag' portion of the data at each tree, then averaging over the entire forest.

$$MSE_{OOB} = n^{-1}\sum_{i=1}^{n}(z_i - \hat{z}_i^{OOB})^2$$

where z_i^{OOB} is the mean out of bag prediction for the ith observation. RF also provides a measure of fit comparable to the R² values of the other models. This 'pseudo R²' is labelled the 'percent variance explained' and is calculated using the following formula:

$$Var_{ex} = 1 - \frac{MSE_{OOB}}{\hat{\sigma}_y^2}$$

where $\hat{\sigma}_y^2$ is the total variance of the dependent variable calculated with n as divisor (rather than $n-1$) (Liaw & Wiener, 2002). The number of trees was set to 100 and the random test data proportion to 30 percent.

An interesting feature of RF is the ability to rank predictor variable in order of importance, which is done by measuring how much the error of the 'out of bag' estimates increases when data for a particular variable is 'removed' from the analysis and the other variables are left intact. This is done on a tree-by-tree basis for the entire forest. The models were generated using the STATISTICA9 (StatSoft Inc., 2011). [Khaled Taalab]

2.4 Neural Networks

Artificial Neural Networks (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the novel structure of the information processing system.

The principles of ANNs are well established (Bishop, 1995) with Maier & Dandy (2001) offering a practical guide for environmental modelling. The structure used for this modelling was a multilayer perceptron, a powerful predictive technique and the one most commonly applied in soil science (Agyare *et al*., 2007). Here, data is separated into a series of nodes, with similar nodes arranged into layers, typically, an input layer (containing the variables used for prediction), an output layer (where predictions are made) and in-between a single hidden layer which weights and transforms the data to extract meaningful relationships.

For each model the data was separated, with 70 percent used for training, 15 percent used for testing and 15 percent used for independent validation. Splitting the data allowed the number of hidden nodes to be tested, which is important as the optimum number of nodes will differ depending on the problem at hand and the number of input variables. It is recommended that the number of hidden nodes should be half the number of input variables plus the number of output variables (Statsoft, Inc., 2011). Generally, adding more nodes will increase the performance of the model, however, this may lead to overfitting the data. To avoid this, the AAN uses a back-propagation algorithm (Rumelhart *et al*., 1986) to test the performance of the network on both training and testing datasets. Training the network should reduce the 'error function' associated with predictions so when the error function of the testing dataset set plateaus or increases, it indicates the ANN has begun to overfit the data. The error function for regression is Sum of Squares error given below:

$$E_{SOS} = \sum_{i-1}^{N} (y_i - t_i)^2$$

where N is the number of training cases, y_i is the predicted value of the i^{th} case and t_i is the observed value. Ideally, when the differences in the error function are negligible, the network with the fewest nodes is chosen. As the test dataset plays a role in developing the ANN infrastructure, a validation data set is used to independently test the predictive power of the models, the best performing were selected using R^2 values and RMSE.

ANNs can also rank variables in order of importance, although they use a different method from RFs. Here, data for each variable is replaced in turn by its mean value from the training data and the effect on the error function is recorded. The variables are then ranked by the amount their omission increases the overall error function (Lou & Nakai, 2001). The learning rate for the ANNs was set to 0.1 and the analysis was carried out using STATISTICA9 (StatSoft Inc., 2011). [Khaled Taalab]

3. Methodology

An initial assessment of soil-landscape relationships were undertaken for Carlow, Laois, Leitrim, Meath, Tipperary South and Wexford to get a "feel" for available data. Predictive mapping for all of *terra incognita* was undertaken in two stages. In *Stage 1*, predictions were made on extrapolated soilscapes based on distance thresholds using feature space analysis. *Stage 2* was based on extrapolation of soilscapes based on induction rules (Bayesian Belief Network).

4. Inference models (Terra Cognita)

4.1 Training data

At the start of the modelling exercise the following problems were identified with regards to the 6 inch (to 1 mile) and County Soil Maps at 1:126,720 scale (Table 1):

- Different age of mapping (pre-, past- General Soil Map)
- Different complexity in mapping (Carlow, Kildare and Wexford) (Table 2)
- No border harmonisation between adjacent Counties
- Not pure soil series, many complexes
- Spatial accuracy (County maps)
- Un-correlated map units (particularly complexes)

Table 1: Number of map units with regards to 6 inch sheets, published county maps and rationalised soil associations

County	6 inch	County	1:250k
Carlow	49	43	18
Clare	62	64	39 (41)
Cork West	29	28	12 (14)
Donegal	70	63	21 (25)
Kildare	29	26	20
Laois	56	52	34
Leithrim	23	23	17 (18)
Limerick	94	90	31 (33)
Mayo	74	75	20 (23)
Meath	53	51	26 (27)
Offaly	82	84	24 (25)
Tipperary N	68	68	30 (31)
Waterford	124	97	23 (25)
Westmeath	44	40	21 (23)
Wexford	45	44	32 (33)

Number in brackets include non-soil classes

Consequently, the 1:250,000 soil association map produced by WP1 was used to provide the training data (Figure 1). For details of the rationalisation, harmonisation and generalisation procedure see Jones *et al.* (2011).

Table 2: Comparison of published County soil maps*

Country	No, of map units	No. of polygons	Average size (ha)
Carlow	43	521	172.20
Clare	64	2029	170.03
Kildare	26	412	411.30
Laois	52	1045	164.85
Leitrim	23	686	231.66
Limerick	90	1102	249.90
Mayo West	75	1234	257.14
Meath	51	754	310.94
Offaly	84	1735	115.13
Tipperary N	68	1479	138.59
Waterford	97	1135	162.73
Donegal West	59*	1616	64.88
Westmeath	40	976	188.56
Wexford	44	411	574.53

**Soilanno* field

4.2 Environmental covariates

Covariate analysis was undertaken using the Generalised Soil Associations for Carlow, Laois and Wexford (Appendix 1) based on the 260 environmental covariates collated for the project (Appendix 3):

Target unit

- Soil Associations

Sample selection

- Regular grid
- Remove urban, industrial complexes, surface water, peat and rock
- Areas adjacent to rivers (2 pixels each side) were excluded to avoid artefacts from ‘burning’ the river network into the DTM

Sample size (percent gird cells)

- 60 m spacing (11%)

Analysis

- Feature analysis (Chi-square) [Carlow, Laois, Leitrim, Meath, Tipperary & Wexford]
- Correlation analysis [Carlow, Laois and Wexford]

Final selection

- Final covariate selection detailed in Table 3

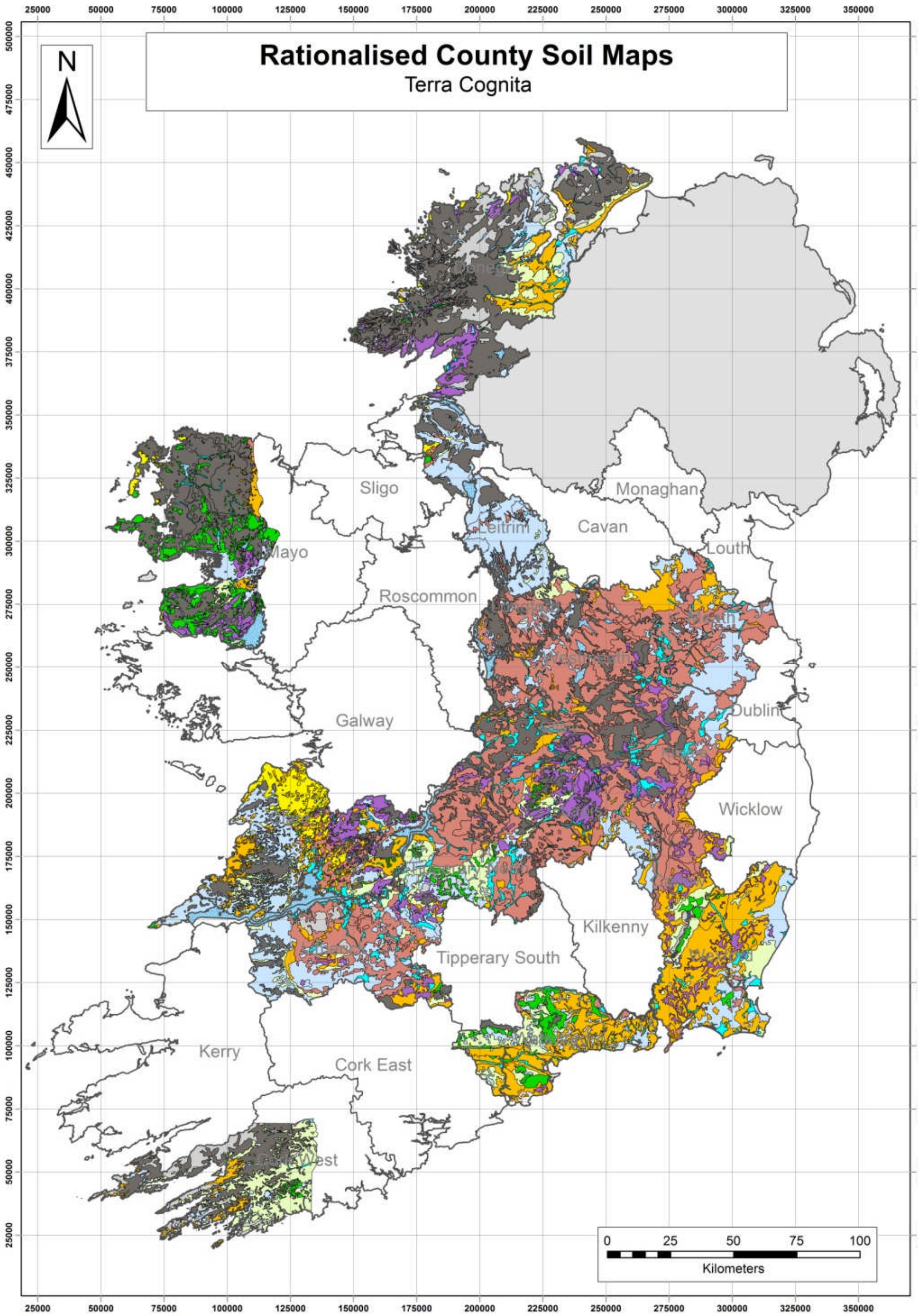


Figure 1: National Soil Associations (Terra Cognita)

Table 3: Environmental Covariates (*scorpan* factors)

s - Soil	• PCTZ2TOP
• General Soil Map	• Protection Index (Relief energie)
• DEM normalized by GSM soil units	• SAGA Wetness Index (t=5, floating point))
c - Climate	• Surface Curvature Index
• Annual mean temperature	• ZCR2ST
• Annual mean precipitation	• ZPIT2PEAK
• Annual mean solar radiation	• ZTOP2PIT
• Annual Mean evapotranspiration	**p – Parent material**
• Thornthwaite Global Humidity Index (THOR)	• Geology map (Version 2)
• Potential Soil Moisture Deficit (PSMD)	• Subsoil map (Version 4)
o - Organism	• Actual/potential drainage density ratio
• CORINE landcover map	**n – Landscape position**
• Habitat map	• Hammond landform subclass
r- Relief	• Iwahashi 16 classes
• Dissection Index (King, 1972)	• SOTER
• Downslope Flowpath Length	• Multi Resolution Valley Bottom Flatness (MRVBF) index
• DTM v0	• Catchment analysis (kmeans7)
• PMIN2MAX	• ConMap
• Elevation above pit	• Protection Index (Relief energie)
• LPIT2PEAK	

4.3 Stratification

Based on the new map of national soil associations, 37 soilscapes were manually delineated for *Terra Cognita* based on expert knowledge according to Hannam *et al*. (2012).

4.4 Sampling design

Sampling design and size is of major importance in the accuracy assessment process (Foody, 2002). One must set up the location and the intensity of sampling to capture the study area. However, practical constraints often limit the realisation of this statistical requirement (Foody, 2002). As recommended by Moran and Bui (2002), we applied a random weighted area sampling scheme. This procedure samples a number of pixels within each class, proportional to the extent of the soil landscape units. The underlying hypothesis below this sampling scheme is that more individuals are needed to characterise large units than smaller units. Besides, this method makes sure that small units are not under-represented, which may be the case with simple random or systematic sampling. [Grinand *et al*., 2007]

The number of sample points is to some extended dictated by practicalities of data preparation, i.e. handling the data in a number of software programmes. Table 4 lists the number of points under various sampling regimes as well as a perspective to traditional sampling schemes.

At the start of the project both Hawths Analysis Tool as well as the Biogeographic Sampling Design tool were evaluated. A description of both software products can be found in Appendix 2 Based on the analysis, the Biogeographic Sampling Design Tool was selected.

Table 4: number of samples in different sampling designs

	Area (km2)	25/km2	30/km2	35/km2	40/km2	50/km2
Carlow	897.18	22430	26915	31401	35887	44859
Kildare	1694.57	42364	50837	59310	67783	84729
Laois	1722.66	43067	51680	60293	68906	86133
Limerick	2753.85	68846	82616	96385	110154	137693
Meath	2344.49	58612	70335	82057	93780	117225
Offaly	1997.47	49937	59924	69911	79899	99874
Tippperary N	2049.74	51244	61492	71741	81990	102487
Waterford	1849.40	46235	55482	64729	73976	92470
Wexford	2362.63	59066	70879	82692	94505	118131

Intensity level	Inspection density	Typical publication scale	Minimum sized delineations
1 (Very high)	> 5 per ha	1:2,500	0.016 ha
2 (High)	1-5 per ha	1:10,000	0.25 ha
3 (Moderately high)	0.2-1 per ha (20-100 per km^2)	1:25,000	1.56 ha
4 (Medium)	5-20 per km^2	1:50,000	6.25 ha
5 (Low)	1-5 per km^2	1:100,000	25 ha
6 (Very low)	<1 per km^2	1:250,000	156 ha

4.5 Database collation

Training databases were developed for individual soilscapes linking the National Soil Associations (Hannam *et al*., 2011) with the environmental covariates.

Reference unit
- Soilscapes Maps (see ISIS_WP2_D1.2)

Target units
- Soil associations

Sample selection
- Stratified Random Sampling

Sample size (percent grid cells)
- 10 observations per km^2
- 298’466 observations

Masking

- Remove urban, industrial complexes, surface water, peat and rock
- Areas adjacent to rivers (2 pixels each side) were excluded to avoid artefacts from "burning" the river network into the DTM

Cleaning

- Removing cases with "empty" fields (i.e. geology)

Re-coding

- Re-coding categorical data layers as both subsoil maps and geology had comas in their fields.

4.6 Inference

Due to requirement for providing predictive maps for both planning and conducting the field survey, initial modelling (*Stage 1*) was developed at the same time as the field work progressed and was based on a concept of "Environmental Distance". Consequently, modelling during Stage 1 was undertaken in phases, with each phase using different training areas. Having completed Phase 1, the mapping exercise was repeated in *Stage 2* using all available reference areas using "Induction Rules". Table 5 lists the different soilscapes used in Stage 1 and Stage 2 predictive mapping.

4.6.1 Stage 1 – Environmental Distance

Both the soil association and soilscape delineations in Terra Cognita changed as the project progressed, and consequently, different Phases are based on different reference areas.

a) Phase 1

Kilkenny and Tipperary – January 2011 – Figure 2 and Table 6

Bayesian Belief Network

- **25** environmental covariates used (Appendix 4);
- Average training accuracy of **72%** (Table 7);
- Relative importance of environmental covariates are listed in Appendix 4;

Random Forest

- **31** environmental covariates (Appendix 11);
- Training accuracy of **77**% (Table 8);
- Relative importance of environmental covariates are listed in Appendix 11;
- Validation accuracy of **76**% (Table 8);

b) Phase 2

Cork and Kerry – February 2011 – Figure 3 and Table 6

Bayesian Belief Network

- **25** environmental covariates used (Appendix 5);
- Average training accuracy of **69%** (Table 9);
- Relative importance of environmental covariates are listed in Appendix 5;

Table 5: Soilscapes used in Stage 1 and Stage 2 predictive mapping

Phase 1 (KIL_TPS, DUB_WIC)	Phase 2 (COR_KER)	Phase 3 (MYE, GLW, RSC, SLG, CVN, LTH, MNG)	Stage 2
s1a			
	s1b	s1b	s1b
	s1c	s1cn	s1cn_updated_160812
	s1d	s1dn	s1dn_updated_160812
s2b	s2b	s2b	s2b
s2c	s2c	s2c	s2c_updated
s4	s4	s4	s4
s5	s5		s5
s6	s6	s6	s6
s8	s8	s8	s8
s9e	s9e	s9e	s9e
s9f		s9f	s9f
s10	s10	s10	s10
s12a	s12a	s12a	s12a
s15	s15	s15	s15
s16	s16	s16	s16
s17	s17	s17n	s17n
s20	s20n	s20n	s20n
s22	s22	s22	s22
s23	s23	s23	s23
s24	s24n	s24n	s24n
s25	s25		s25
	s26	s26	s26n*
	s27	s27	s27
	s28	s28	s28
	s29	s29	s29n_updated
		s30	s30_newDNGL
		s31	s31n_newDNG:
		s32	s32_updated
		s33	s33_updated
		s34	s34_updated
			s35_newDNGL
		s36	s36n_newDNGL
		s37	s37n_newDNGL
			s39

Table 6: Stage 1 soilscapes, number of sampling points and component associations

Soilscape	Sampling points	Soil Associations	Soilscape	Sampling points	Soil Associations
s1a	84,621	50	**s20**	2,345	12
s1b	3,827	28	**s20n**	2553	13
s1c	55,628	50	**s22**	2,082	10
s1cn	62,864	50	**s23**	4,423	11
s1d	21,618	36	**s24**	5,054	6
s1dn	28645	39	**s24n**	16,788	25
s2b	4,777	14	**s25**	4,705	11
s2c	12584	32	**s26**	6,313	28
s5	6,617	21	**s27**	5,237	19
s6	7,514	9	**s28**	6,059	12
s8	5,147	12	**s29**	10,801	8
s9e	15,503	22	**s30**	12,711	18
s9f	4,026	12	**s31**	3,904	15
s10	7,069	17	**s32**	13,218	16
s12a	6,065	9	**s33**	6,149	16
s15	1,611	6	**s34**	519	3
s16	3,574	12	**s35**	10,435	19
s17	14,378	19	**s36**	2,263	13
s17n	14378	22	**s37**	7,921	18

Table 7: Stage 1 Phase 1 BN inference performance (accuracy)

	with climate	without climate	optimised
s1a	50.49	50.49	50.49
s2b	78.21	78.21	78.21
s2c	62.25	62.25	62.25
s4	70.07	70.07	70.07
s5	68.29	68.29	68.29
s6	67.61	67.61	67.61
s8	77.31	77.31	77.31
s9e	54.47	54.47	54.47
s9f	63.49	63.49	63.49
s10	77.01	77.01	77.01
s12a	67.35	67.35	67.35
s15	88.52	88.52	88.52
s16	78.88	78.88	78.88
s17	66.42	66.42	66.42
s20	75.01	75.01	75.01
s22	77.09	77.09	77.09
s23	66.36	66.36	66.36
s24	79.40	79.40	79.40
s25	71.39	71.39	71.39
	70.51	71.99	73.18

Table 8: Stage 1 Phase 1 RF1 inference performance (accuracy)

	Risk estimate		Standard error		Accuracy	
Soilscape	**train**	**test**	**train**	**test**	**train**	**test**
s1a	0.4813	0.4831	0.0021	0.0031	51.87	51.69
s2b	0.2365	0.2631	0.0074	0.0115	76.35	73.69
s2c	0.4366	0.4462	0.0053	0.0081	56.34	55.38
s4	0.2002	0.2343	0.0092	0.0147	79.98	76.57
s5	0.2219	0.2461	0.0061	0.0096	77.81	75.39
s6	0.1547	0.1586	0.0050	0.0077	84.53	84.14
s8	0.1877	0.2062	0.0065	0.0103	81.23	79.38
s9e	0.2190	0.2127	0.0040	0.0060	78.10	78.73
s9f	0.2550	0.2584	0.0083	0.0124	74.50	74.16
s10	0.1873	0.1876	0.0056	0.0085	81.27	81.24
s12a	0.3109	0.3441	0.0071	0.0110	68.91	65.59
s15	0.0957	0.0932	0.0088	0.0132	90.43	90.68
s16	0.2549	0.2283	0.0088	0.0125	74.51	77.17
s17	0.2926	0.3072	0.0052	0.0081	70.74	69.28
s20	0.2304	0.2750	0.0104	0.0168	76.96	72.50
s22	0.1535	0.2110	0.0094	0.0164	84.65	78.90
s23	0.1754	0.1646	0.0069	0.0100	82.46	83.54
s24	0.1460	0.1602	0.0060	0.0093	85.40	83.98
s25	0.1542	0.1585	0.0063	0.0096	84.58	84.15
					76.87	75.59

* without ConMap

Table 9: Stage 1 Phase 2 BN1 inference performance (accuracy)

	with climate	**without climate**		**with climate**	**without climate**
s1b	61.08	62.34	**s15**	67.35	69.23
s1c	51.15	56.12	**s16**	88.52	87.71
s1d	58.56	60.49	**s17**	78.88	76.11
s2b	78.21	78.63	**s20n**	66.42	70.41
s2c	62.25	63.06	**s22**	72.54	73.21
s4	70.07	69.26	**s23**	77.09	79.11
s5	68.29	68.90	**s24n**	66.36	69.00
s6	67.61	69.43	**s25**	51.40	56.42
s8	77.31	77.91	**s26**	71.39	76.26
s9e	54.47	62.57	**s27**	65.69	67.57
s10	63.49	68.36	**s28**	59.57	62.97
s12a	77.01	77.41	**s29**	58.68	62.83
				66.88	69.10

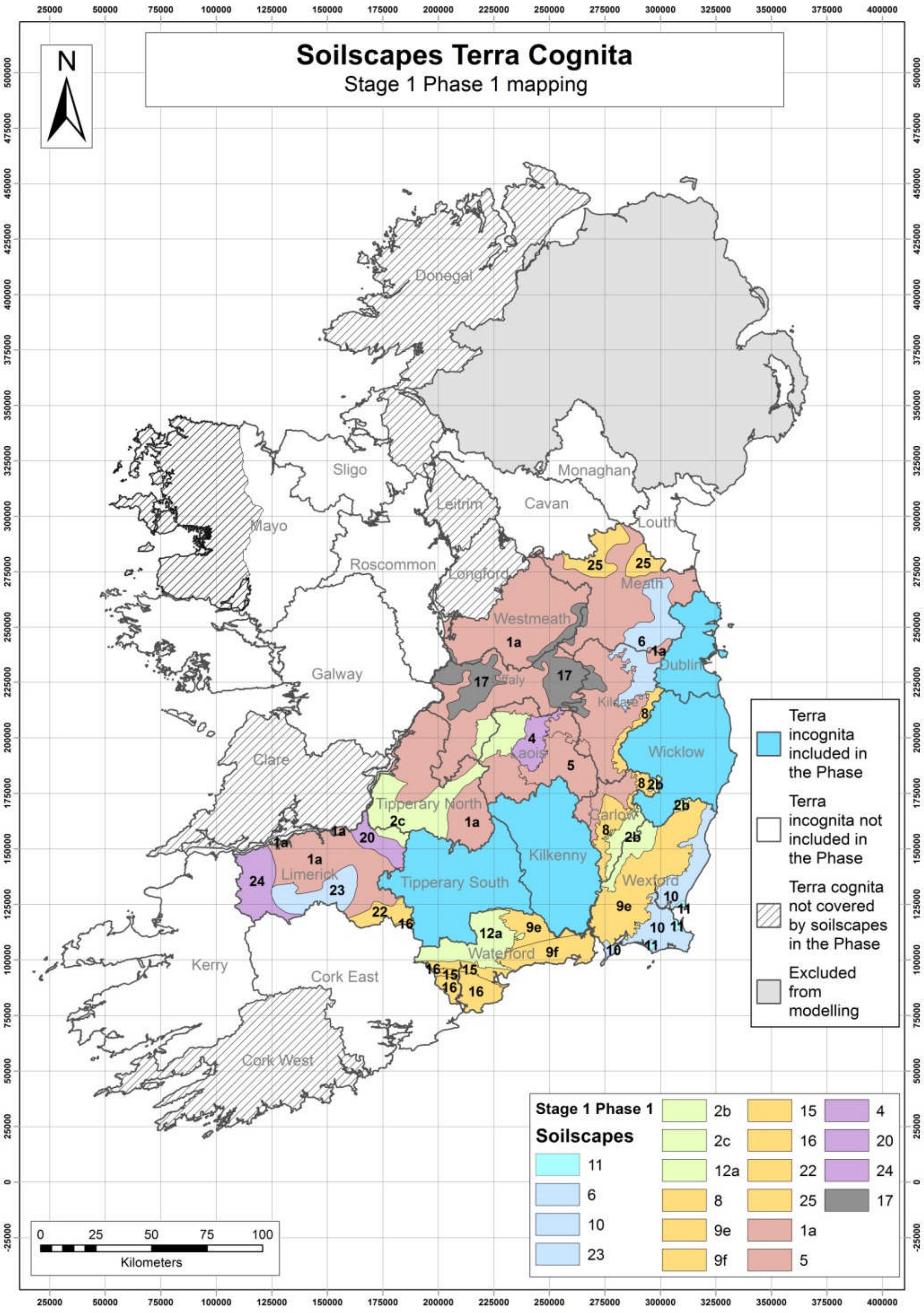


Figure 2: Stage 1 reference areas for Phase 1

Random Forest

- **31** environmental covariates (Appendix 11);
- Training accuracy of **74**% (Table 10);
- Relative importance of environmental covariates are listed in Appendix 11;
- Validation accuracy of **73**% (Table 10);

c) Phase 3

Galway, Mayo East, Roscommon and Sligo – April 2011 – Figure 4 and Table 6
Cavan, Louth and Monaghan – November 2012

Bayesian Belief Network

- **25** environmental covariates used (Appendix 6);
- Average training accuracy of **70%** (Table 11);
- Relative importance of environmental covariates are listed in Appendix 6;
-

Random Forest

- **31** environmental covariates (Appendix 11);
- Training accuracy of **75**% (Table 12);
- Relative importance of environmental covariates are listed in Appendix 11;
- Validation accuracy of **74**% (Table 12);

4.6.2 Stage 2 – Induction Rules

Models were generated for each of the 34 soilscapes identified in *Terra Cognita* (**Figure 5**). The number of cases available per soilscape is listed in Table 13.

a) Bayesian Belief Networks

Three 3 different Bayesian Belief Networks were constructed to assess different combinations of environmental covariates. In all networks the numerical data were binned into 30 classes were possible.

Network 1(with climate):

- **31** environmental covariates used (Appendix 7);
- Average training accuracy of **68%** (Table 14);
- Relative importance of environmental covariates are listed in Appendix 7;

Network 2 (without climate):

- **25** environmental covariates used (Appendix 8);
- Average training accuracy of **69%** (Table 14);
- Relative importance of environmental covariates are listed in Appendix 8;

Network 3 (forward selection):

- Stepwise selection (Appendix 9);
- Stopping condition of <0.1%
- Average training accuracy of **78%** (Table 14);
- Relative importance of environmental covariates are listed in Appendix 9**;**

Table 10: Stage 1 Phase 2 RF inference performance (accuracy)

	Risk estimate		Standard error		Accuracy	
Soilscape	**train**	**test**	**train**	**test**	**train**	**test**
s1b	0.3470	0.3833	0.0093	0.0141	65.30	61.67
s1c	0.4463	0.4451	0.0025	0.0038	55.37	55.49
s1d	0.3404	0.3564	0.0038	0.0060	65.96	64.36
s2b	0.2365	0.2631	0.0074	0.0115	76.35	73.69
s2c	0.4366	0.4462	0.0053	0.0081	56.34	55.38
s4	0.2002	0.2343	0.0092	0.0147	79.98	76.57
s5	0.2219	0.2461	0.0061	0.0096	77.81	75.39
s6	0.1547	0.1586	0.0050	0.0077	84.53	84.14
s8	0.1877	0.2062	0.0065	0.0103	81.23	79.38
s9e	0.2190	0.2127	0.0040	0.0060	78.10	78.73
s10	0.1873	0.1876	0.0056	0.0085	81.27	81.24
s12a	0.3109	0.3441	0.0071	0.0110	68.91	65.59
s15	0.0957	0.0932	0.0088	0.0132	90.43	90.68
s16	0.2549	0.2283	0.0088	0.0125	74.51	77.17
s17	0.2926	0.3072	0.0052	0.0081	70.74	69.28
s20n	0.2516	0.2923	0.0103	0.0163	74.84	70.77
s22	0.1535	0.2110	0.0094	0.0164	84.65	78.90
s23	0.1754	0.1646	0.0069	0.0100	82.46	83.54
s24n	0.3388	0.3460	0.0044	0.0068	66.12	65.40
s25	0.1542	0.1585	0.0063	0.0096	84.58	84.15
s26	0.3409	0.3641	0.0072	0.0109	65.91	63.59
s27	0.2554	0.2753	0.0072	0.0112	74.46	72.47
s28	0.2477	0.2550	0.0067	0.0101	75.23	74.50
s29	0.2892	0.3071	0.0052	0.0081	71.08	69.29
					74.42	72.97

* without ConMap

Table 11: Stage 1 Phase 3 BN inference performance (accuracy)

	with climate	without climate	optimised		with climate	without climate	optimised
s1b	61.08	62.34	62.02	**s22**	77.09	79.11	79.83
s1cn	50.05	55.34	58.73	**s23**	66.36	69.00	79.20
s1dn	61.78	62.89	62.17	**s24n**	51.40	56.42	61.56
s2b	78.21	78.63	80.15	**S25**	71.39	76.26	79.21
s2c	62.25	63.06	62.97	**s26**	65.69	67.57	71.15
s4	70.07	69.26	73.33	**s27**	59.57	62.97	65.97
S5	68.29	68.90	70.74	**s28**	58.68	62.83	68.95
s6	67.61	69.43	67.94	**s29**	58.53	62.29	66.06
s8	77.31	77.91	77.11	**s30**	65.54	69.61	70.62
s9e	54.47	62.57	65.23	**s31**	78.48	81.63	82.22
s9f	63.49	68.36	64.11	**s32**	63.19	64.77	68.67
s10	77.01	77.41	79.53	**s33**	63.33	67.05	70.95
s12a	67.35	69.23	64.86	**s34**	85.55	87.09	88.05
s15	88.52	87.71	88.95	**S35**	52.73	58.57	63.03
s16	78.88	76.11	77.31	**s36**	86.83	86.70	84.67
s17n	63.21	69.61	67.09	**s37**	56.81	61.56	62.02
s20n	72.54	73.21	74.34		67.37	69.86	71.55

Table 12: Stage 1 Phase 3 RF1 inference performance (accuracy)

	Risk estimate		Standard error		Accuracy	
Soilscape	**train**	**test**	**train**	**test**	**train**	**test**
s1b	0.3470	0.3833	0.0093	0.0141	65.30	61.67
s1cn	0.4480	0.4447	0.0024	0.0036	55.20	55.53
s1dn	0.3743	0.3740	0.0034	0.0052	62.57	62.60
s2b	0.2365	0.2631	0.0074	0.0115	76.35	73.69
s2c	0.4366	0.4462	0.0053	0.0081	56.34	55.38
s4	0.2002	0.2343	0.0092	0.0147	79.98	76.57
s6	0.1547	0.1586	0.0050	0.0077	84.53	84.14
s8	0.1877	0.2062	0.0065	0.0103	81.23	79.38
s9e	0.2190	0.2127	0.0040	0.0060	78.10	78.73
s9f	0.2550	0.2584	0.0083	0.0124	74.50	74.16
s10	0.1873	0.1876	0.0056	0.0085	81.27	81.24
s12a	0.3109	0.3441	0.0071	0.0110	68.91	65.59
s15	0.0957	0.0932	0.0088	0.0132	90.43	90.68
s16	0.2549	0.2283	0.0088	0.0125	74.51	77.17
s17n	0.3331	0.3264	0.0047	0.0072	66.69	67.36
s20n	0.2516	0.2923	0.0103	0.0163	74.84	70.77
s22	0.1535	0.2110	0.0094	0.0164	84.65	78.90
s23	0.1754	0.1646	0.0069	0.0100	82.46	83.54
s24n	0.3388	0.3460	0.0044	0.0068	66.12	65.40
s26	0.3409	0.3641	0.0072	0.0109	65.91	63.59
s27	0.2554	0.2753	0.0072	0.0112	74.46	72.47
s28	0.2477	0.2550	0.0067	0.0101	75.23	74.50
s29	0.2892	0.3071	0.0052	0.0081	71.08	69.29
s30	0.2518	0.2525	0.0046	0.0071	74.82	74.75
s31	0.1486	0.1401	0.0069	0.0100	85.14	85.99
s32	0.2505	0.2536	0.0045	0.0069	74.95	74.64
s33	0.2250	0.2651	0.0064	0.0102	77.50	73.49
s34	0.1072	0.0575	0.0167	0.0176	89.28	94.25
s36	0.1390	0.1799	0.0087	0.0146	86.10	82.01
s37	0.3207	0.3504	0.0063	0.0098	67.93	64.96
					74.88	73.75

* without ConMap

Table 13: Stage 2 soilscapes, number of sampling points and component associations

Soilscape	Sampling points	Soil Associations	Soilscape	Sampling points	Soil Associations
s1b	3,828	28	**s23**	4,423	11
s1cn	56,272	49	**s24n**	16,788	25
s1dn	26,142	38	**s25**	4,705	11
s2b	4,777	14	**s26**	6,314	28
s2c	11,108	32	**s27**	5,237	19
s5	6,617	21	**s28**	6,059	12
s6	7,514	9	**s29**	10,790	9
s8	5,147	12	**s30**	10,383	18
s9e	15,503	22	**s31**	17,836	27
s9f	4,026	12	**s32**	11,100	13
s10	7,069	17	**s33**	5,118	12
s12a	6,065	9	**s34**	435	4
s15	1,611	6	**s35**	10,435	18
s16	3,574	12	**s36**	13,256	24
s17n	14,378	22	**s37**	8,652	29
s20n	2,554	13	**s39**	455	7
s22	2,082	10			

Table 14: Stage 2 BN2 inference performance (accuracy)

	with climate	without climate	optimised		with climate	without climate	optimised
s1b	60.93	61.24	69.81	**s23**	67.99	69.32	83.70
s1cn	48.31	53.65	62.93	**s24n**	52.62	55.38	68.69
s1dn	58.65	59.66	68.31	**s25**	71.94	75.07	86.95
s2b	78.63	76.81	82.29	**s26**	66.69	67.15	72.14
s2c	62.10	61.70	67.10	**s27**	60.70	62.13	71.37
s5	69.65	68.02	76.21	**s28**	59.18	60.76	77.40
s6	66.50	68.63	83.64	**s29**	59.82	61.03	68.28
s8	77.04	78.10	83.25	**s30**	67.23	69.06	76.02
s9e	56.19	62.09	81.20	**s31**	60.70	62.61	81.66
s9f	65.00	66.87	74.76	**s32**	60.59	63.35	76.60
s10	77.13	76.14	85.77	**s33**	63.40	64.19	75.46
s12a	67.58	69.02	71.82	**s34**	88.74	90.58	97.70
s15	88.70	86.47	92.49	**s35**	57.23	58.62	68.00
s16	77.90	75.13	81.11	**s36**	76.31	76.35	79.02
s17n	61.68	68.38	73.13	**s37**	61.79	62.97	70.99
s20n	73.95	72.35	79.63	**s39**	91.21	88.79	92.53
s22	79.01	79.01	83.53		67.73	68.81	77.68

b) Neural Networks

Two Neural Networks were extracted based on 38 environmental covariates. In contrast to both Bayesian Networks and Random Forests, the relative importance of the environmental covariates is not available.

Network 1 (without ConMap)**:**
- **31** environmental covariates (Appendix 10);
- Training accuracy of **15%** (Table 15);
- No relative importance measure of environmental covariates are available;
- Validation accuracy of **22%** (Table 15);

Network 2 (without ConMap)**:**
- **31** environmental covariates (Appendix 10);
- Training accuracy of **15%** (Table 15);
- No relative importance measure of environmental covariates are available;
- Validation accuracy of **21%** (Table 15);

c) Random Forests

Network 1 (without ConMap):
- **31** environmental covariates (Appendix 11);
- Training accuracy of **75%** (Table 16);
- Relative importance of environmental covariates are listed in Appendix 11;
- Validation accuracy of **73%** (Table 16);

5. Predictive mapping of Soil Associations (Terra Incognita)

Counties for which no detailed soil survey maps exist encompass Cavan, Dublin, East Cork, East Mayo, Galway, Kerry, Kilkenny, Louth, Monaghan, Roscommon, Sligo, Tipperary South and Wicklow.

5.1 Deployment data

Deployment data were compiled for *terra incognita* as follows:
- 20 m grid sampling
- Urban, surface water, peat and rock and river buffer removed
- Areas adjacent to rivers (2 pixels each side) were excluded to avoid artefacts from “burning” the river network into the DTM
- Empty fields in the categorical data were marked with “no data”
- Data blocks of 500k points extracted from N to S
- Data blocks separated into individual soilscape files
- Soilscape files packaged into 250k blocks for processing

Using the DTM resolution as the primary processing unit, the task- in-hand can be established from Table 17, which lists the number of pixels to be processed for each County in *terra incognita*. A total of approximately 98’140’804 pixels, equivalent to approximately 3’925.632 km^2, had to be processed.

Table 15: Stage 2 NN inference performance (accuracy)

Soilscape	Training performance		Test performance		Validation performance	
	Network 1	Network 2	Network 1	Network 2	Network 1	Network 2
s1b	22.09	26.31	30.89	34.21	33.68	33.86
1cn	29.78	21.92	28.74	31.62	25.72	31.98
s1dn	23.68	21.75	24.19	28.36	26.91	28.77
s2b	9.75	12.47	16.76	18.58	17.32	15.22
2c	22.31	17.49	29.11	30.19	32.53	30.85
s5	13.21	11.66	18.35	17.44	18.75	19.25
s6	4.18	9.47	11.00	11.89	11.89	11.45
s8	14.38	5.86	18.39	17.36	18.65	18.13
s9e	8.63	8.82	15.48	16.77	16.60	16.69
s9f	15.82	16.35	24.38	22.39	27.53	24.71
s10	8.99	9.42	12.74	12.08	14.06	13.68
s12a	9.70	18.04	22.33	23.32	27.83	25.30
s15	2.48	4.87	9.54	9.13	10.79	9.96
s16	15.55	11.91	19.96	18.84	24.44	22.39
s17n	17.31	21.45	24.86	27.83	23.47	24.91
s20n	16.94	14.25	22.51	20.42	25.39	24.08
s22	6.24	10.97	12.50	15.71	12.18	15.71
s23	14.30	16.31	16.74	17.80	19.61	17.04
s24n	21.04	18.56	27.16	28.59	27.40	26.65
s25	9.32	9.07	10.35	11.21	11.49	11.63
s26	27.64	17.42	30.87	32.14	30.44	26.53
s27	12.93	17.48	24.33	28.79	24.84	23.44
s28	15.32	15.28	22.80	23.02	23.35	22.80
29n	17.99	20.60	25.46	26.21	26.21	26.02
s30	8.80	14.57	20.36	19.27	22.09	21.52
s31	10.87	10.84	16.00	15.59	14.36	14.73
s32	10.00	9.63	14.89	16.16	14.71	15.32
s33	14.26	14.54	17.73	18.12	20.86	20.73
s34	1.64	4.92	10.77	12.31	6.15	7.69
s35	23.48	18.66	27.48	25.88	29.39	27.99
s36	13.97	15.94	20.02	20.12	20.27	20.32
s37	21.94	18.39	26.37	27.14	29.38	29.22
s39	18.50	16.61	19.12	23.53	20.59	13.24
	14.64	14.60	20.37	21.27	21.48	20.96

* without ConMap

Table 16: Stage 2 RF2 inference performance (accuracy)

	Risk estimate		Standard error		Accuracy	
Soilscape	**train**	**test**	**train**	**test**	**train**	**test**
s1b	0.3470	0.3833	0.0093	0.0141	65.30	61.67
s1cn	0.4365	0.4435	0.0025	0.0038	56.35	55.65
s1dn	0.3660	0.3768	0.0036	0.0055	63.40	62.32
s2b	0.2365	0.2631	0.0074	0.0115	76.35	73.69
s2c	0.4421	0.4617	0.0056	0.0086	55.79	53.83
s5	0.2219	0.2461	0.0061	0.0096	77.81	75.39
s6	0.1547	0.1586	0.0050	0.0077	84.53	84.14
s8	0.1877	0.2062	0.0065	0.0103	81.23	79.38
s9e	0.2190	0.2127	0.0040	0.0060	78.10	78.74
s9f	0.2550	0.2584	0.0083	0.0124	74.50	74.16
s10	0.1873	0.1876	0.0056	0.0085	81.27	81.24
s12a	0.3109	0.3441	0.0071	0.0110	68.91	65.59
s15	0.0957	0.0932	0.0088	0.0132	90.43	90.68
s16	0.2549	0.2283	0.0088	0.0125	74.51	77.17
s17n	0.3331	0.3264	0.0047	0.0072	66.69	67.36
s20n	0.2516	0.2923	0.0103	0.0163	74.84	70.77
s22	0.1535	0.2110	0.0094	0.0164	84.65	78.90
s23	0.1754	0.1646	0.0069	0.0100	82.46	83.54
s24n	0.3388	0.3460	0.0044	0.0068	66.12	65.40
s25	0.1542	0.1585	0.0063	0.0096	84.58	84.15
s26	0.3409	0.3641	0.0072	0.0109	65.91	63.59
s27	0.2554	0.2753	0.0072	0.0112	74.46	72.47
s28	0.2477	0.2550	0.0067	0.0101	75.23	74.50
s29	0.2887	0.3157	0.0052	0.0082	71.13	68.43
s30	0.2546	0.2579	0.0051	0.0078	74.54	74.21
s31	0.2092	0.2080	0.0036	0.0056	79.08	79.20
s32	0.1917	0.1996	0.0045	0.0069	80.83	80.04
s33	0.2407	0.2548	0.0072	0.0111	75.93	74.52
s34	0.0805	0.1095	0.0158	0.0267	91.95	89.05
s35	0.3010	0.3086	0.0054	0.0083	69.90	69.14
s36	0.2650	0.2613	0.0046	0.0070	73.50	73.87
s37	0.3477	0.3744	0.0061	0.0095	65.23	62.56
s39	0.1704	0.2083	0.0213	0.0338	82.96	79.17
					74.80	73.47

* without ConMap

Table 17: Number of pixels processed (deployment)

	Area (m2)	Area (km2)	Grid cells
Cavan	1933789928.6	193,379.0	4,834,475
Cork East	7336781201.9	733,678.1	18,341,953
Dublin	923288609.0	92,328.9	2,308,222
Galway	6145357874.6	614,535.8	15,363,395
Kerry	4784389578.3	478,439.0	11,960,974
Kilkenny	2073286250.0	207,328.6	5,183,216
Louth	823871879.9	82,387.2	2,059,680
Mayo East	5278289211.6	527,828.9	13,195,723
Monaghan	1292391360.0	129,239.1	3,230,978
Roscommon	2545657181.8	254,565.7	6,364,143
Sligo	1833958526.0	183,395.9	4,584,896
Tipperary S	2255410000.0	225,541.0	5,638,525
Wicklow	2029850000.0	202,985.0	5,074,625
		3,925,632.2	98,140,804

5.2 Stratification

Stratification in *terra incognita* is based on 2 different approaches according to Mayr *et al*. (2013), namely environmental distance and rule induction. The *distance threshold* approach is based on feature space analysis, whereas the *rule induction* approach is based on a Bayesian Belief Network inference engine.

Stage 1 soilscape extrapolation is based on environmental distance approach (feature space analysis). During *Stage 1*, soilscape definitions changed from one Phase to the next due to operational requirements (Figures 2 to 4) with the final version shown in Figure 6. *Stage 2* soilscape extrapolation is based on induction rules using a Bayesian Belief Network (Figure 7).

5.3 Deployment

For the Bayesian Belief Networks and Random Forests, the respective rules were applied to the deployment files and processed using a dedicated FORTRAN program. The results mapped at 20m resolution in ESRI ArcGIS^{TM}.

Stage 1 deployment resulted in the following maps: Bayesian Belief Network (Figure 8), and Random Forest (Figure 9). Stage 2 deployment resulted in the following maps: Bayesian Belief Network 2 (Figure 10), Bayesian Belief Network 3 (Figure 11), and Random Forest (Figure 12).

5.4 Post-processing

Generalisation

A generalisation procedure was applied to up-scale the predicted soil association's *Terra Incognita* to a scale of 1:250,000 scale to achieve harmony with the associations delineated for *Terra Cognita*. A bespoke ArcGIS™ tool was written by G. LoPapa (Teagasc) to undertake this process.

Stage 1 post processing resulted in the following maps: Bayesian Belief Network (Figure 13), and Random Forest (Figure 14). Stage 2 post processing resulted in the following maps: Bayesian Belief Network 2 (Figure 15), Bayesian Belief Network 3 (Figure 16), and Random Forest (Figure 17).

Table 18: Mask composition

Water	Subsoil Map	Water	Water
Urban	Subsoil Map	Made	Made ground
Peat	Habitat Map	PH	Bog&Heath
	Habitat Map	PBCEBC	Bare Peat & Soil
	Habitat Map	F	Fen
	Habitat Map	FC	Cutover Fen
	Habitat Map	FR	Reclaimed Fen
	Habitat Map	RBF	Raised Bog / Fen
	Habitat Map	RBFC	Cutover Raised B
	Habitat Map	RBFR	Reclaimed Raised
	Habitat Map	UBB	Upland Blanket B
	Habitat Map	UBBC	Cutover Upland B
	Habitat Map	UBBCE	Cutover / Erodin
	Habitat Map	UBBR	Reclaimed UBB
	Habitat Map	LBB	Lowland Blanket
	Habitat Map	LBBCE	Cutover / Erodin
	Habitat Map	LBBR	Reclaimed LBB
Rock	Landuse Map	ER	Bare rock
	Landuse Map	CR	Rocky complex
Alluvial	Subsoil Map	Alluvium	Alluvium_undifferentiated
	Subsoil Map	Alluvium	Clayey
	Subsoil Map	Alluvium	Gravelly
	Subsoil Map	Alluvium	Silty
Sand/Dunes	Subsoil Map	Other-Aeolian	Aeolian_Sediments_undifferentiated
	Subsoil Map	Other-Aeolian	Blown_sand
	Subsoil Map	Other-Aeolian	Blown_sand_in_dunes
Tidal Marshes	Subsoil Map	Marine_deposits	Estuarine_sediments_(silts/clays)
	Subsoil Map	Other-Material	Marsh
	Subsoil Map	Other-Material	Tidal_marsh

Mask
The available data sets allowed us to re-introduce a number of key "soil" types which would have been difficult to map. These are alluvial, peat, rock, sand/dunes, tidal marshes, urban and water bodies (Table 18).

Cartography and GIS

As part of the post-processing procedure of the various map layers, common map layouts, colour schemes, labels and legends were used. Further to this, the data were provided in an ESRI File Geodatabase, as described in report 'ISIS_WP2_D44_NewSoilMapofIrelandMapv1DataDelivery'.

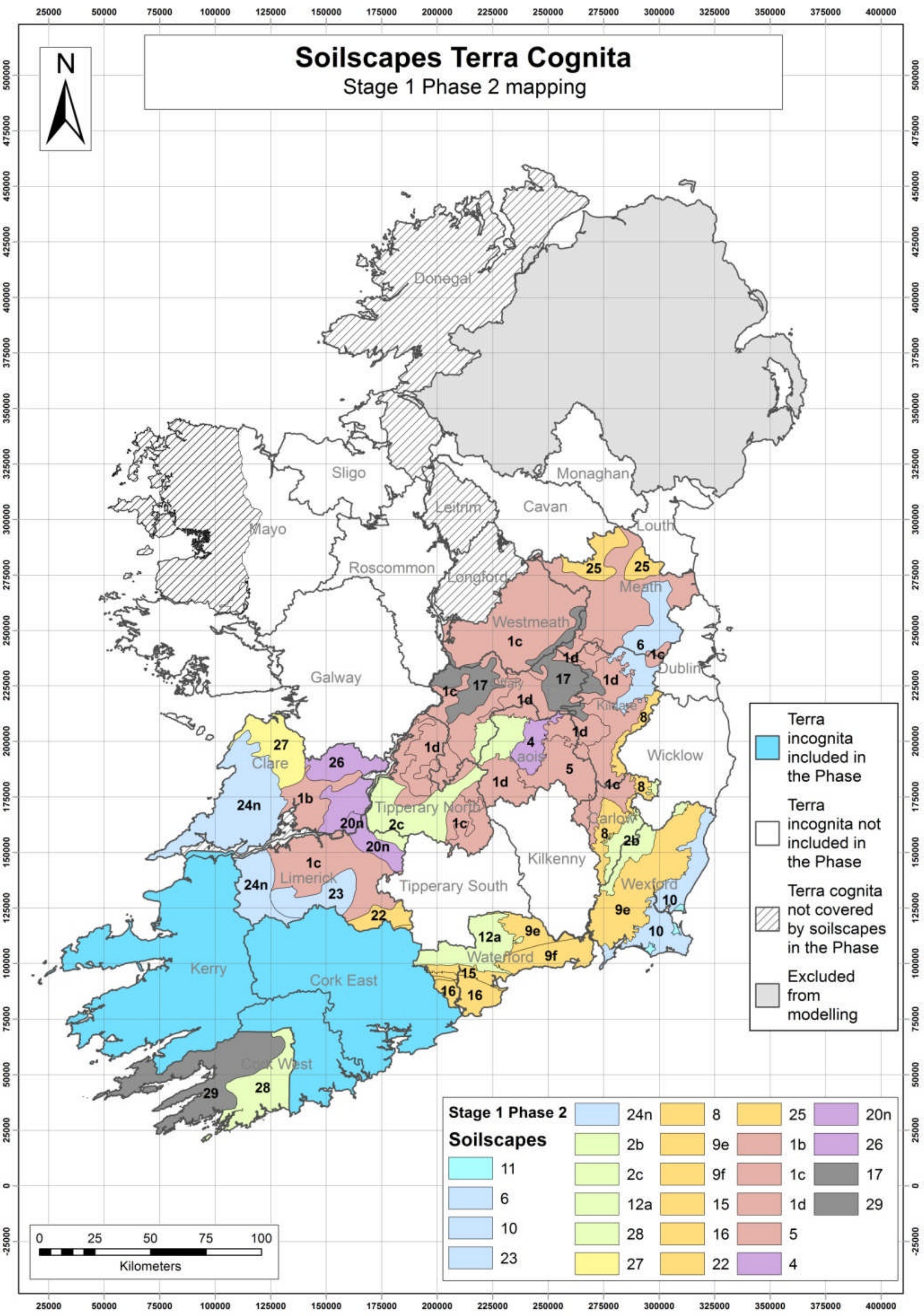


Figure 3: Stage 1 reference areas for Phase 2

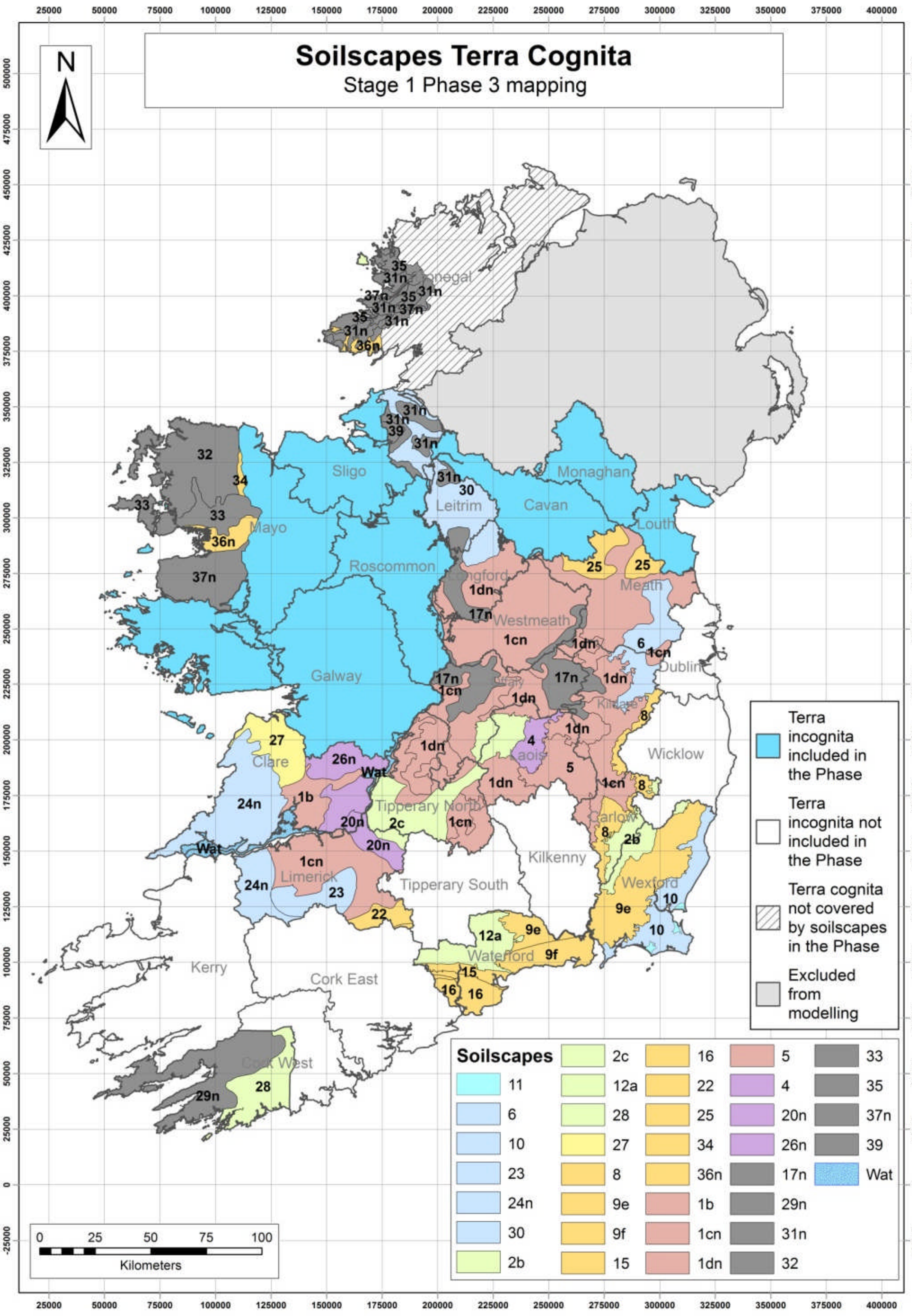


Figure 4: Stage 1 reference areas for Phase 3

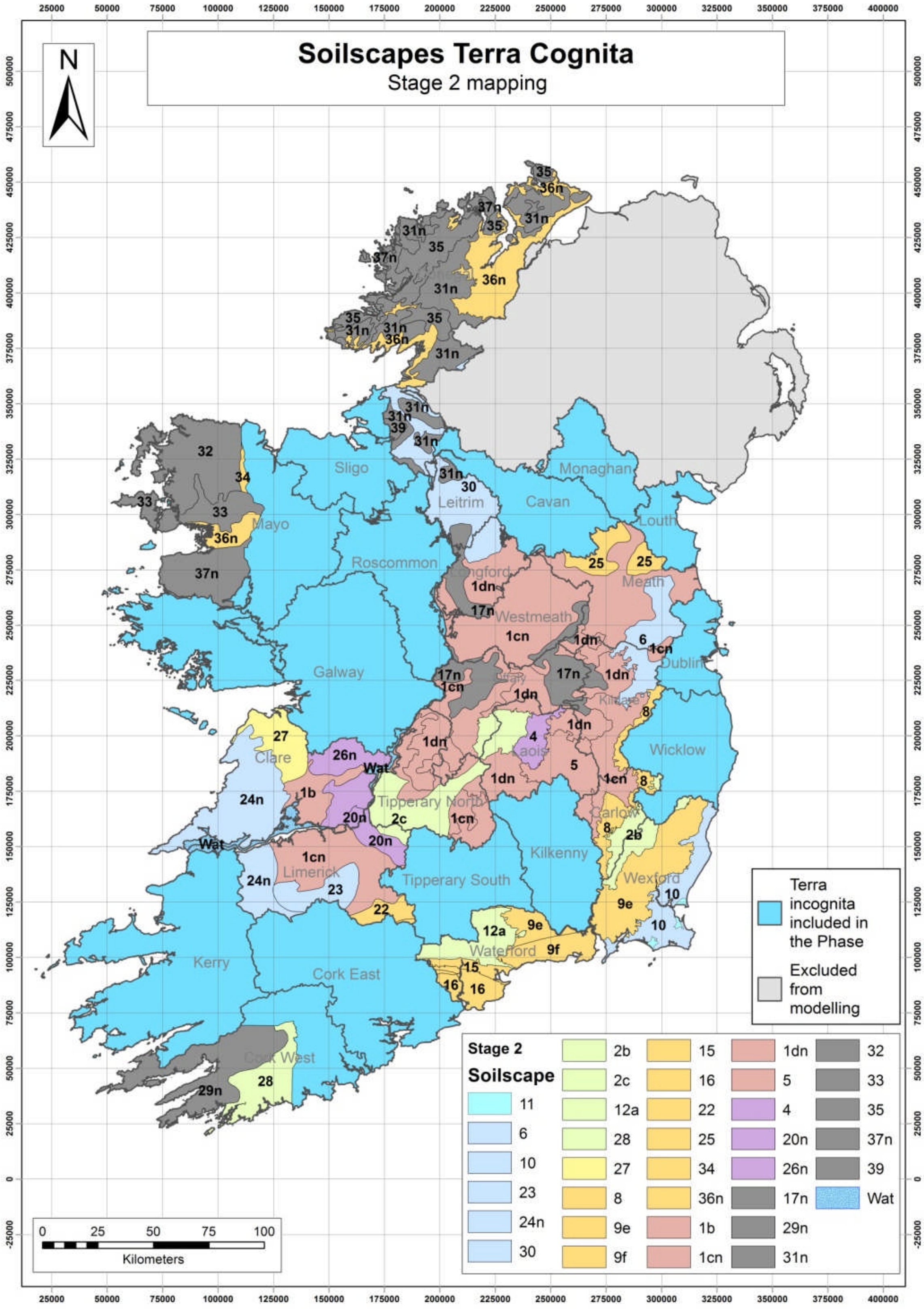


Figure 5: Stage 2 reference areas

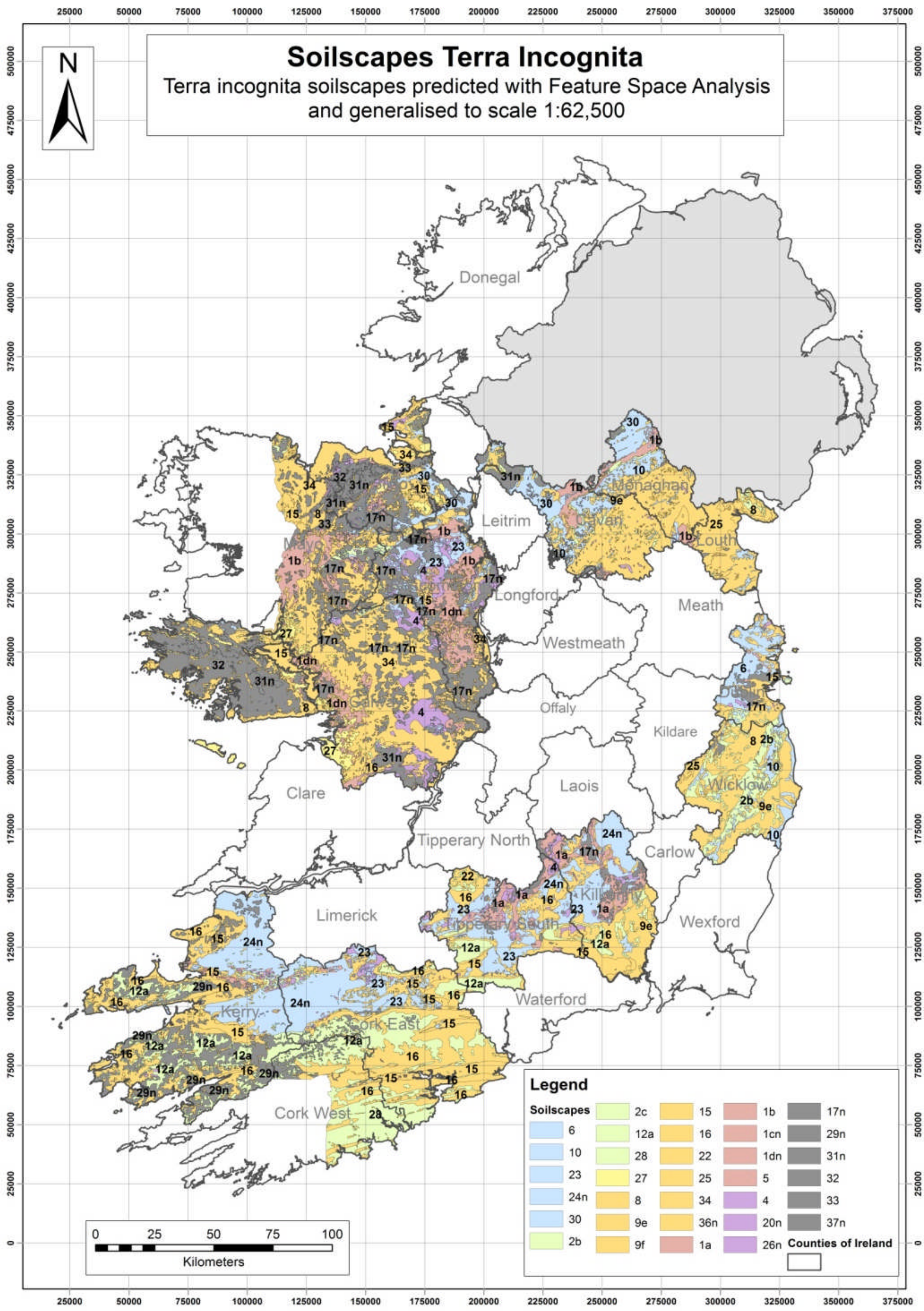


Figure 6: Stage 1 soilscapes based on Feature Space Analysis

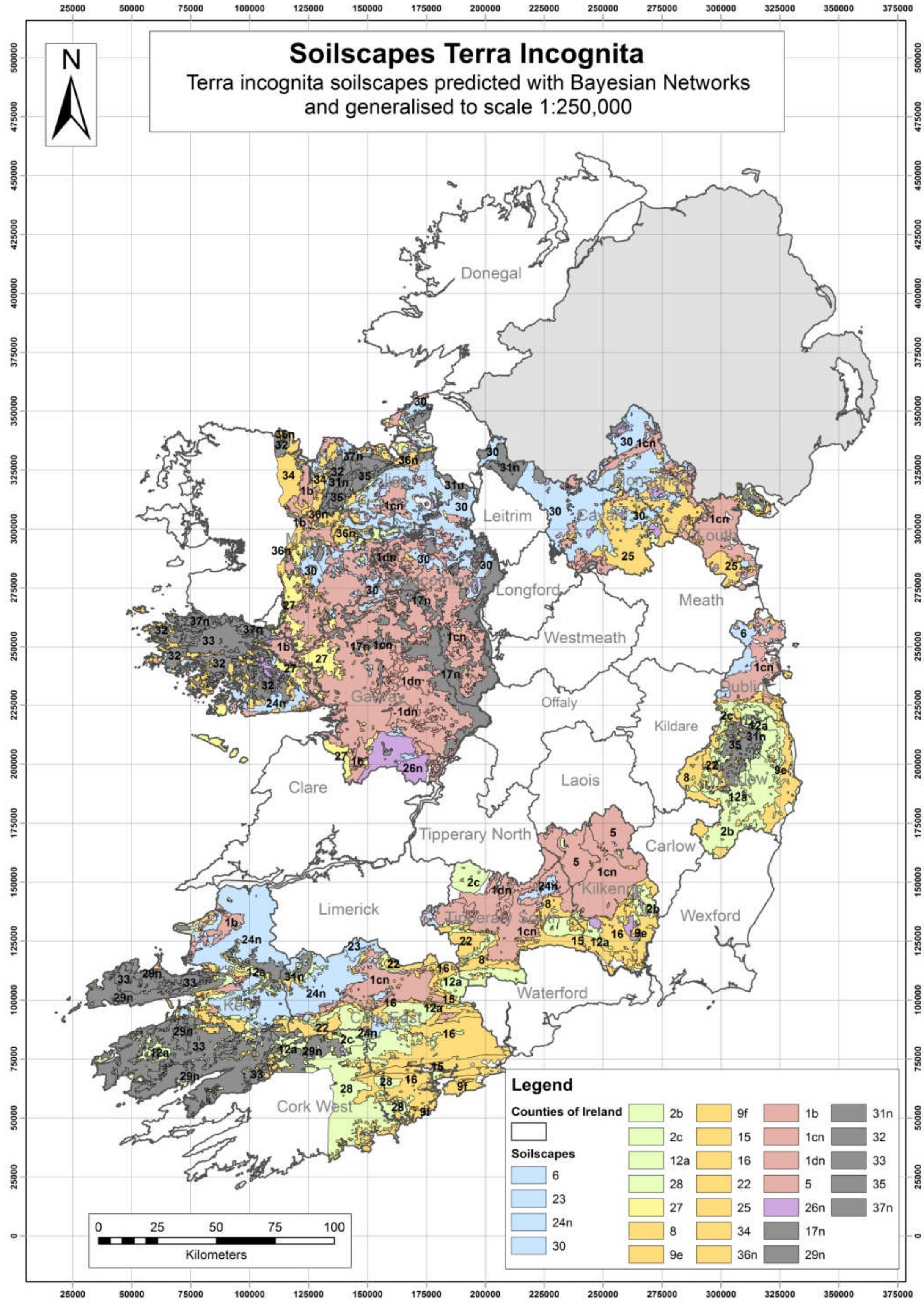


Figure 7: Stage 2 soilscapes based on Bayesian Belief Network (BN)

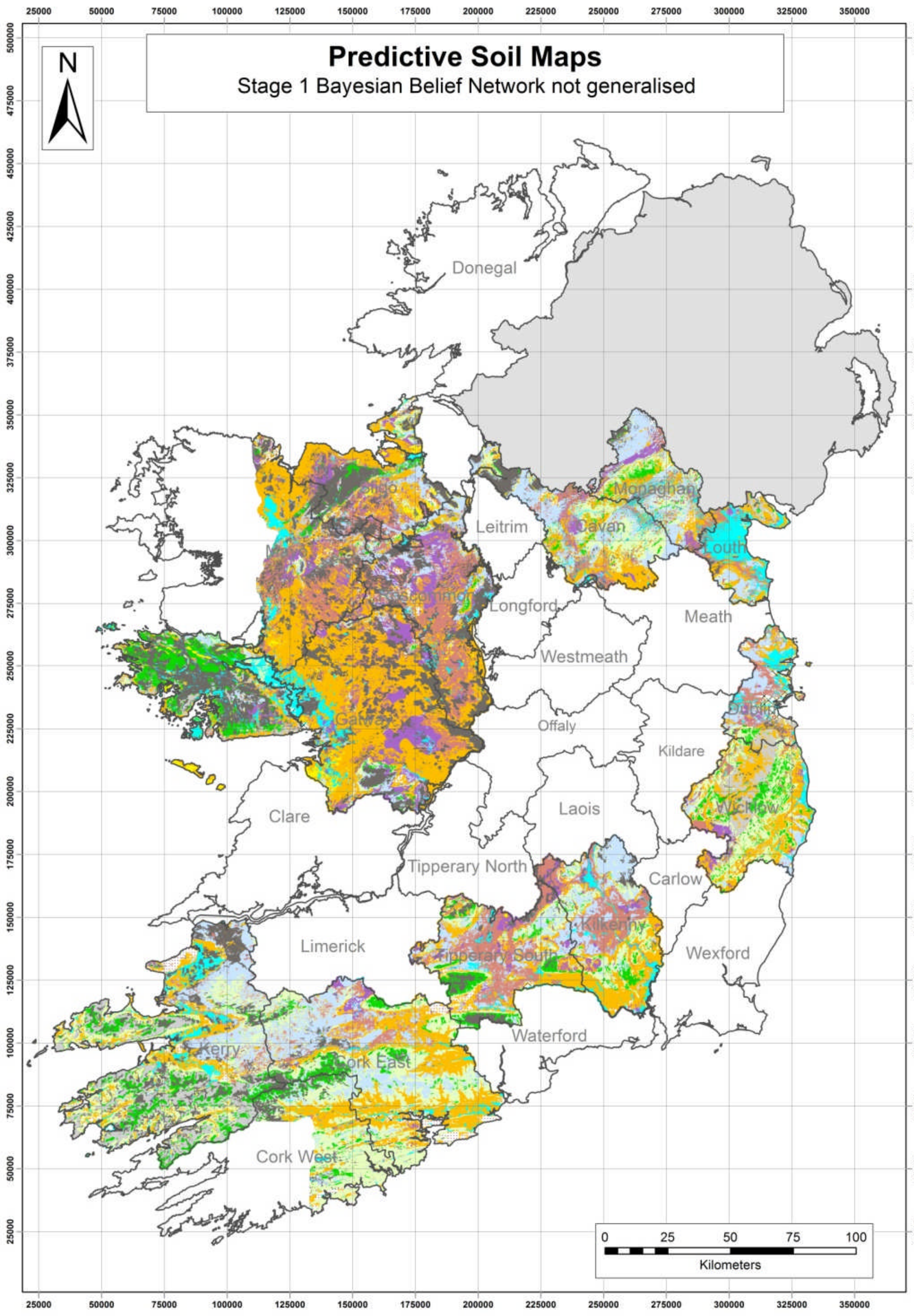


Figure 8: Stage 1 predictive soil map using Bayesian Belief Network – BN1 (not generalised)

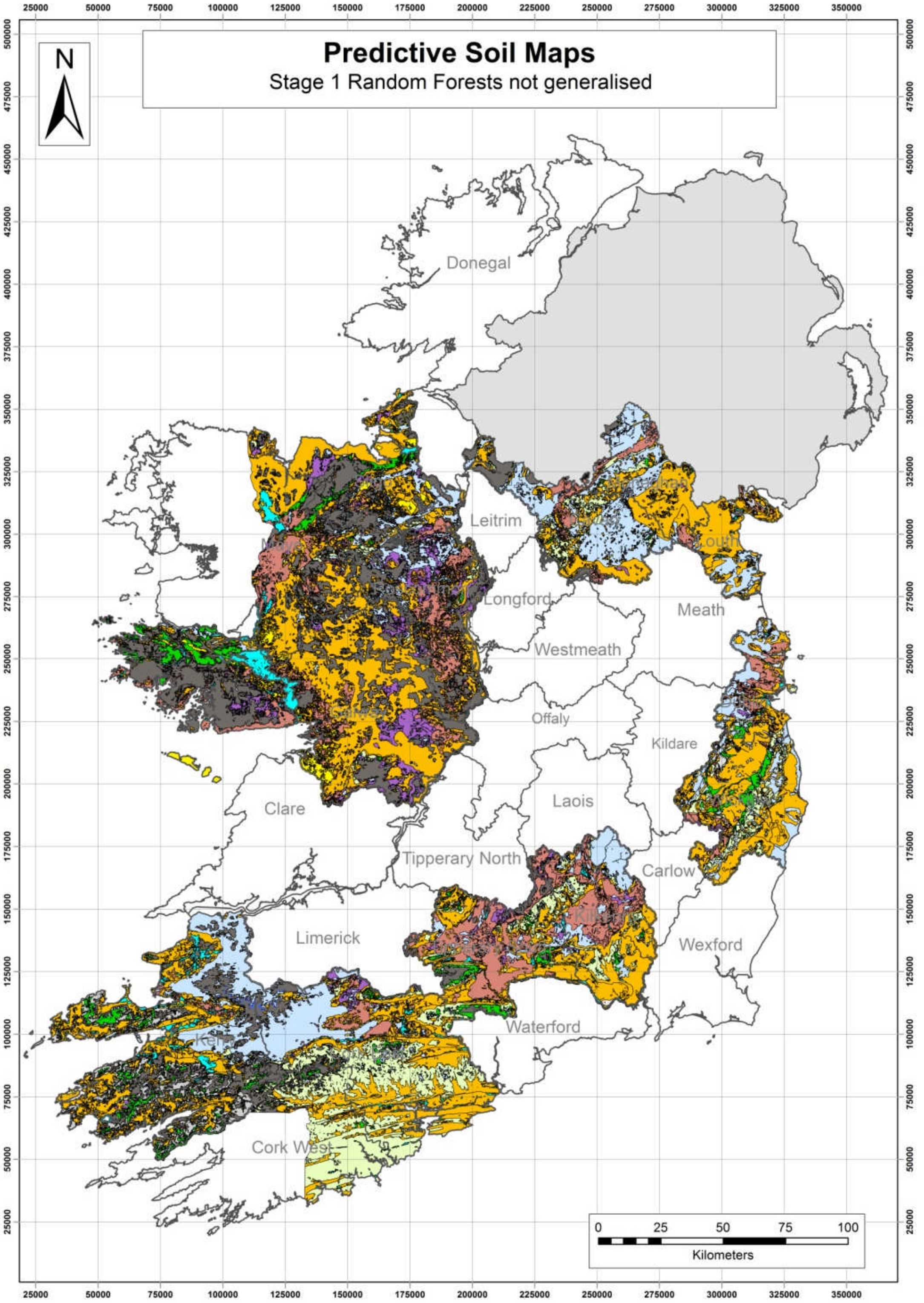


Figure 9: Stage 1 predictive soil map using Random Forests – RF1 (not generalised)

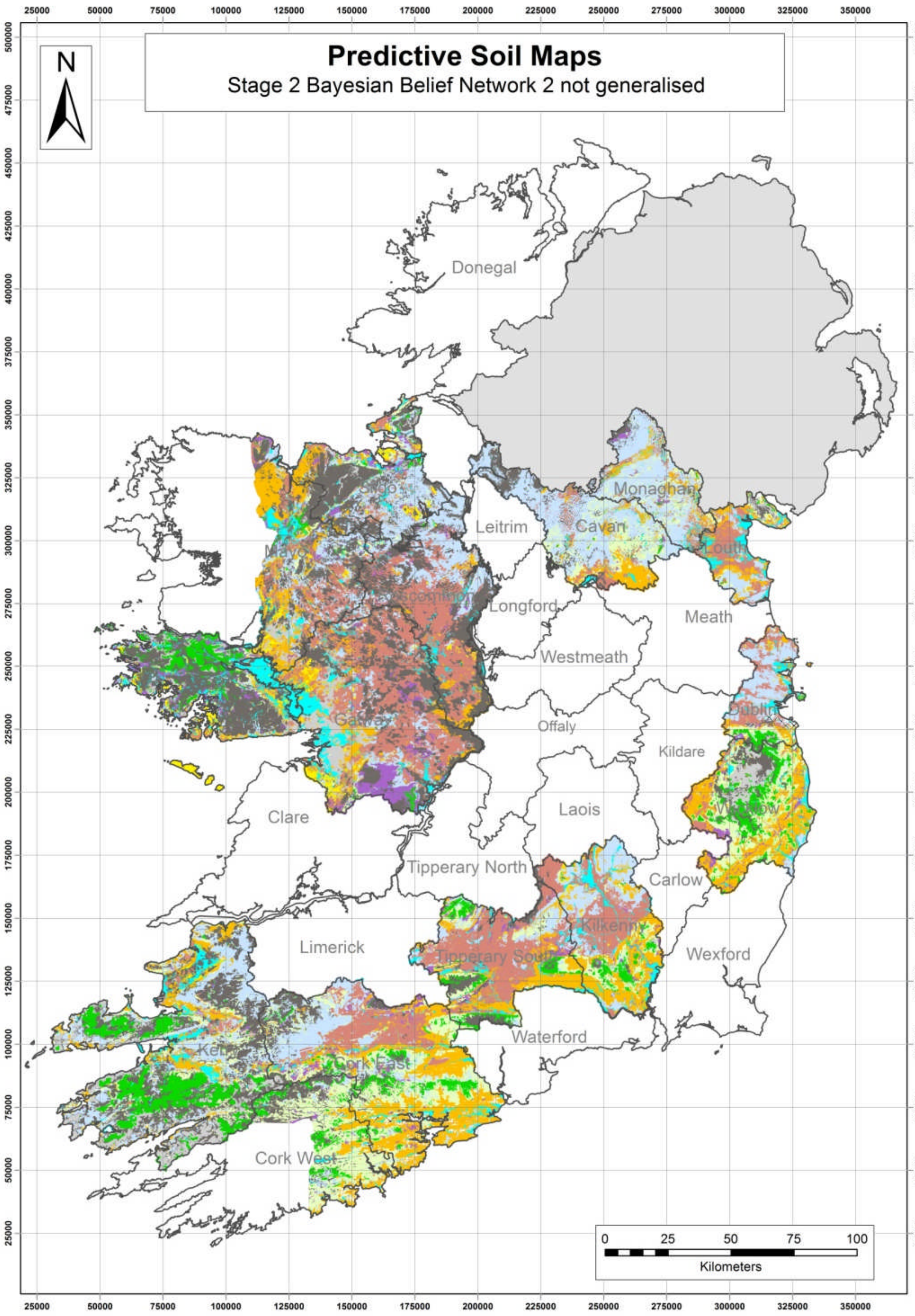


Figure 10: Stage 2 predictive soil map using Bayesian Belief Network -BN2 (not generalised)

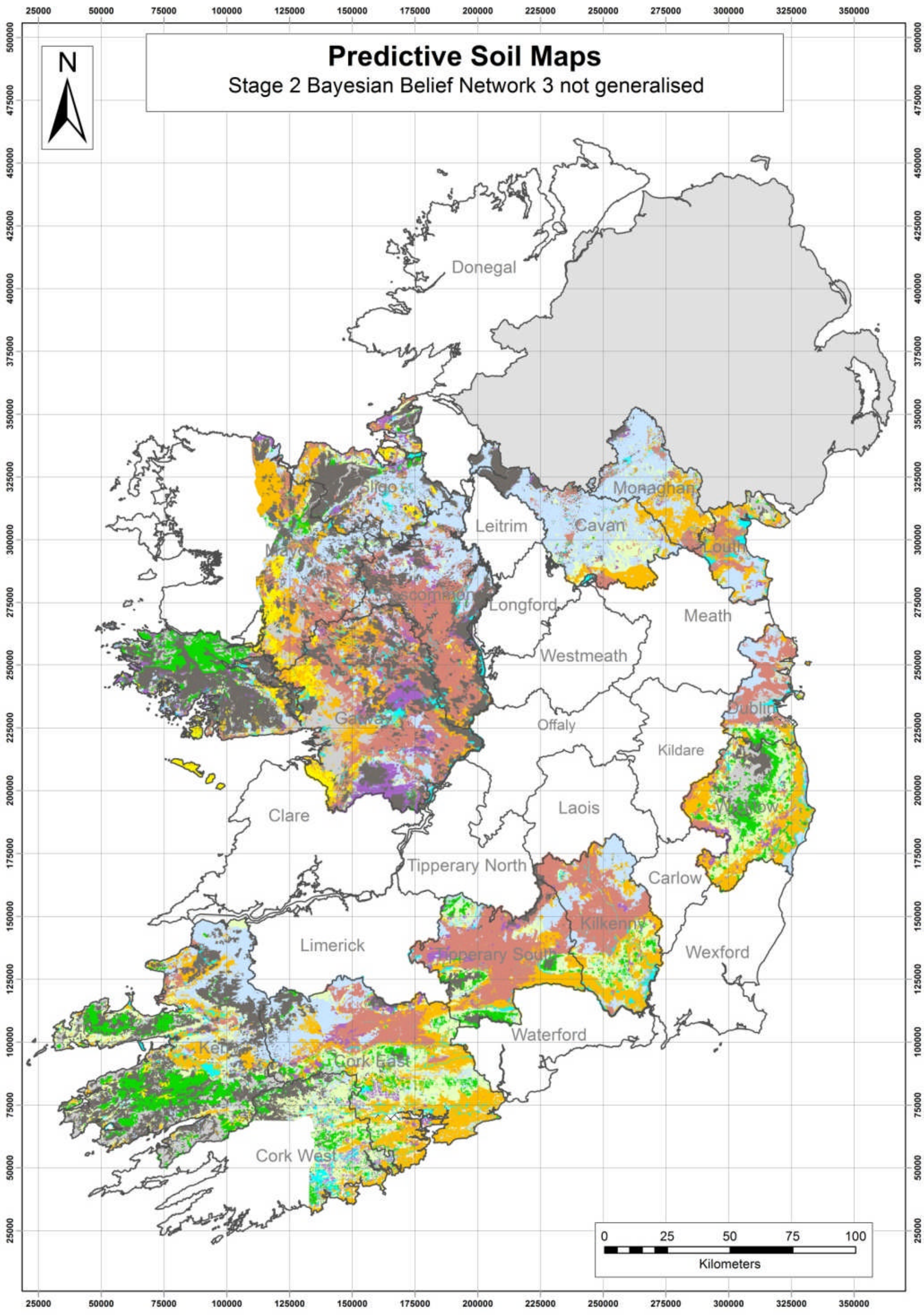


Figure 11: Stage 2 predictive soil map using Bayesian Belief Network - BN3 (not generalised)

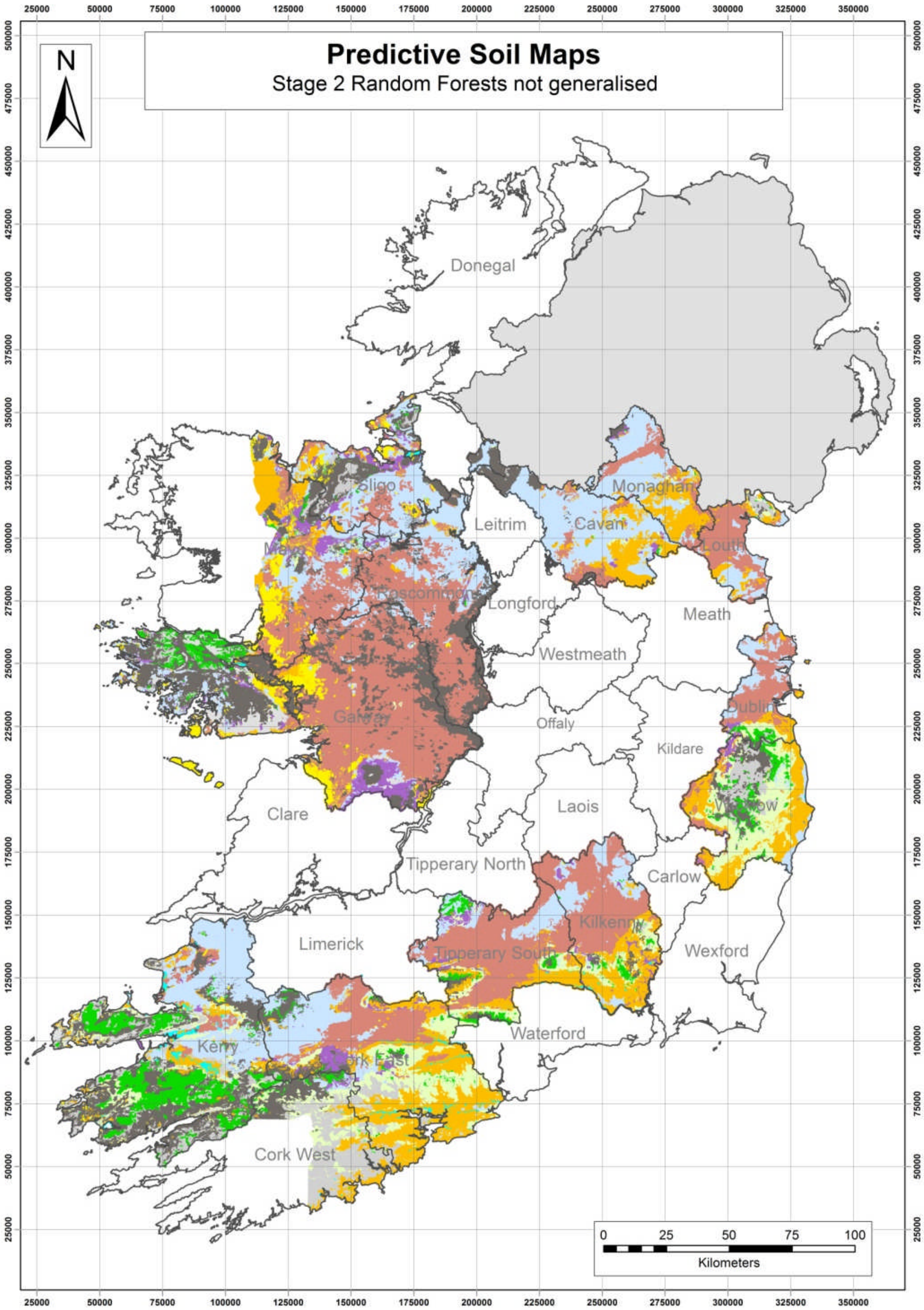


Figure 12: Stage 2 predictive soil map using Random Forests – RF2(not generalised)

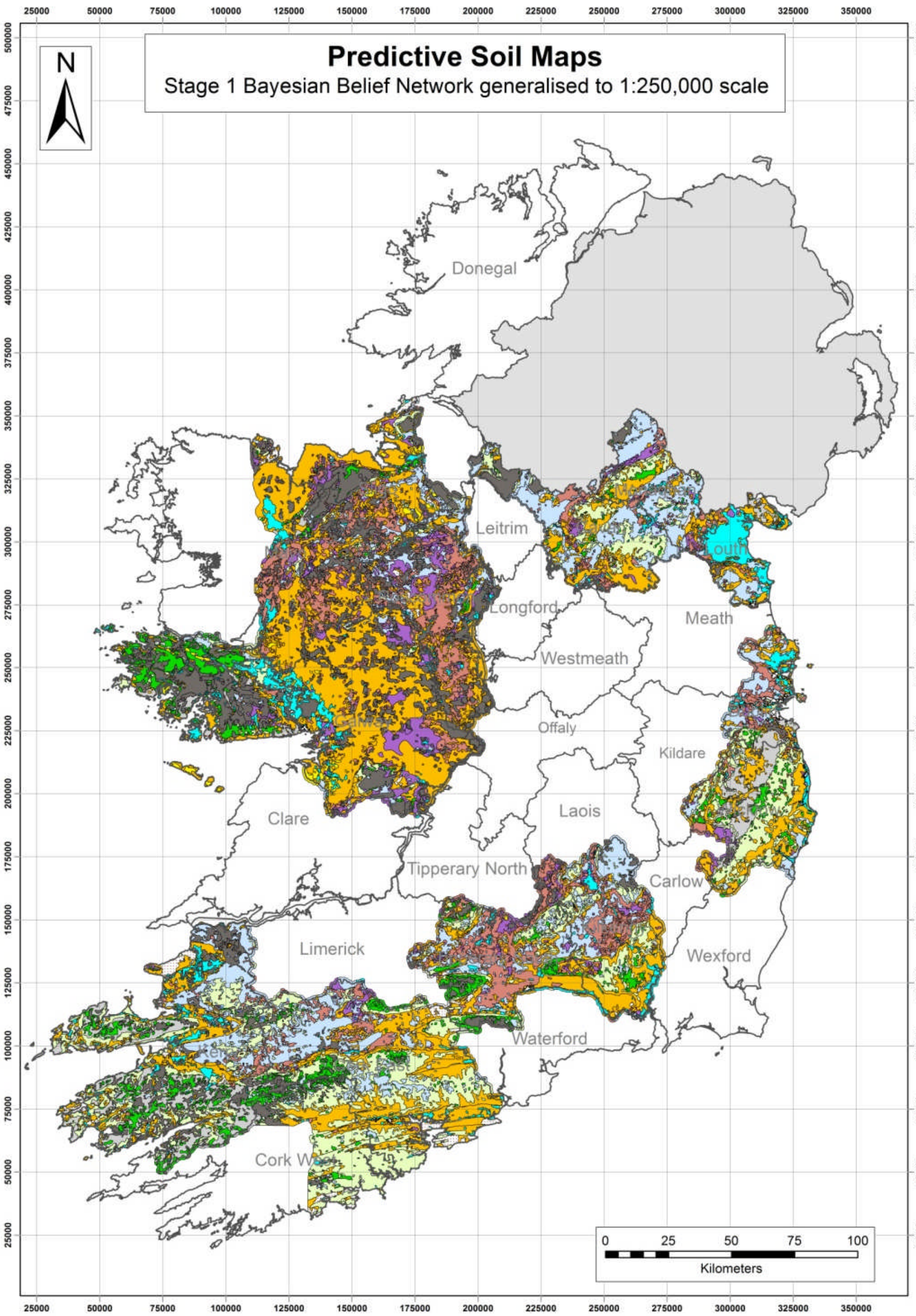


Figure 13: Stage 1 predictive soil map using Bayesian Belief Network – BN1 (generalised)

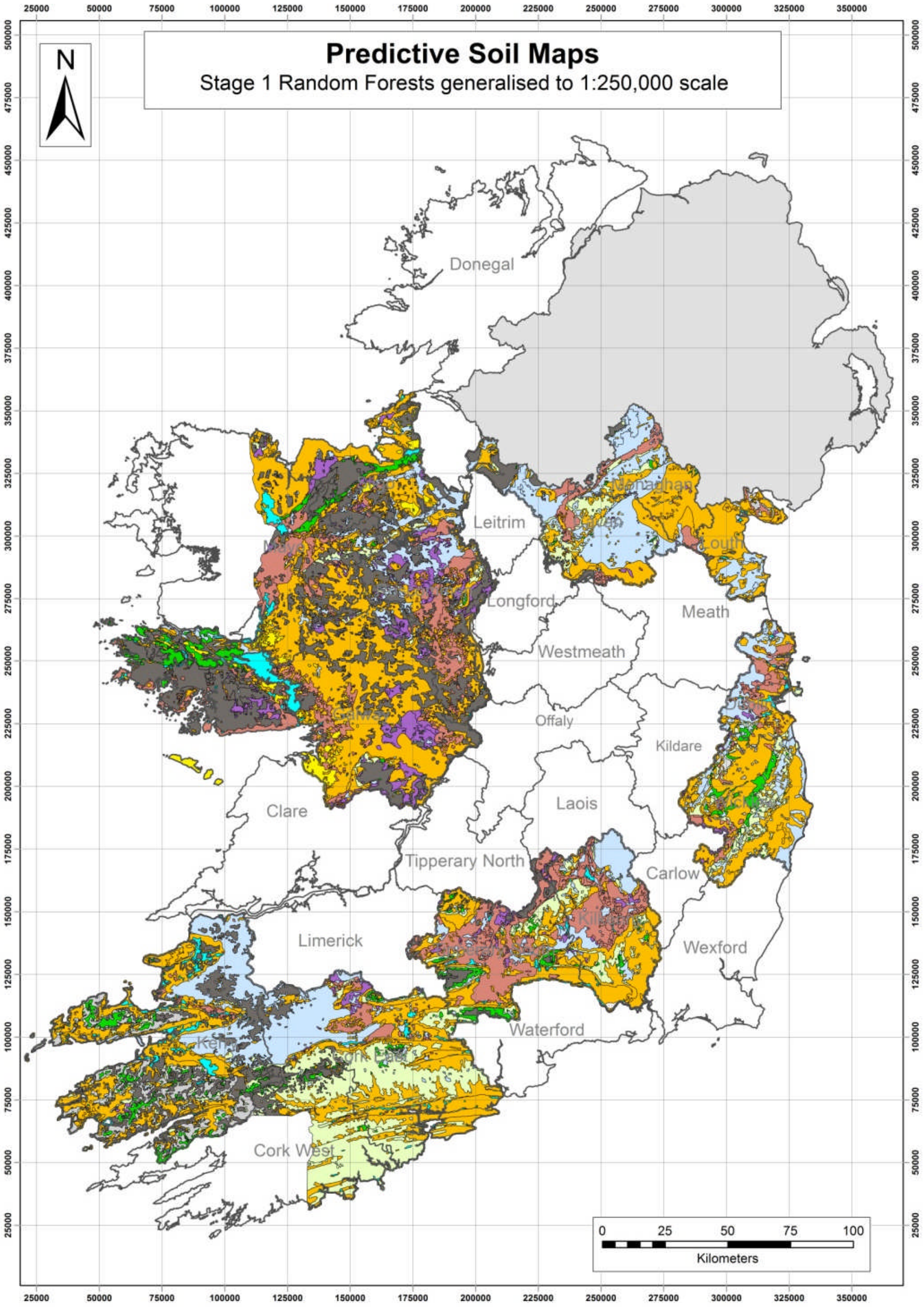


Figure 14: Stage 1 predictive soil map using Random Forests – RF1 (generalised)

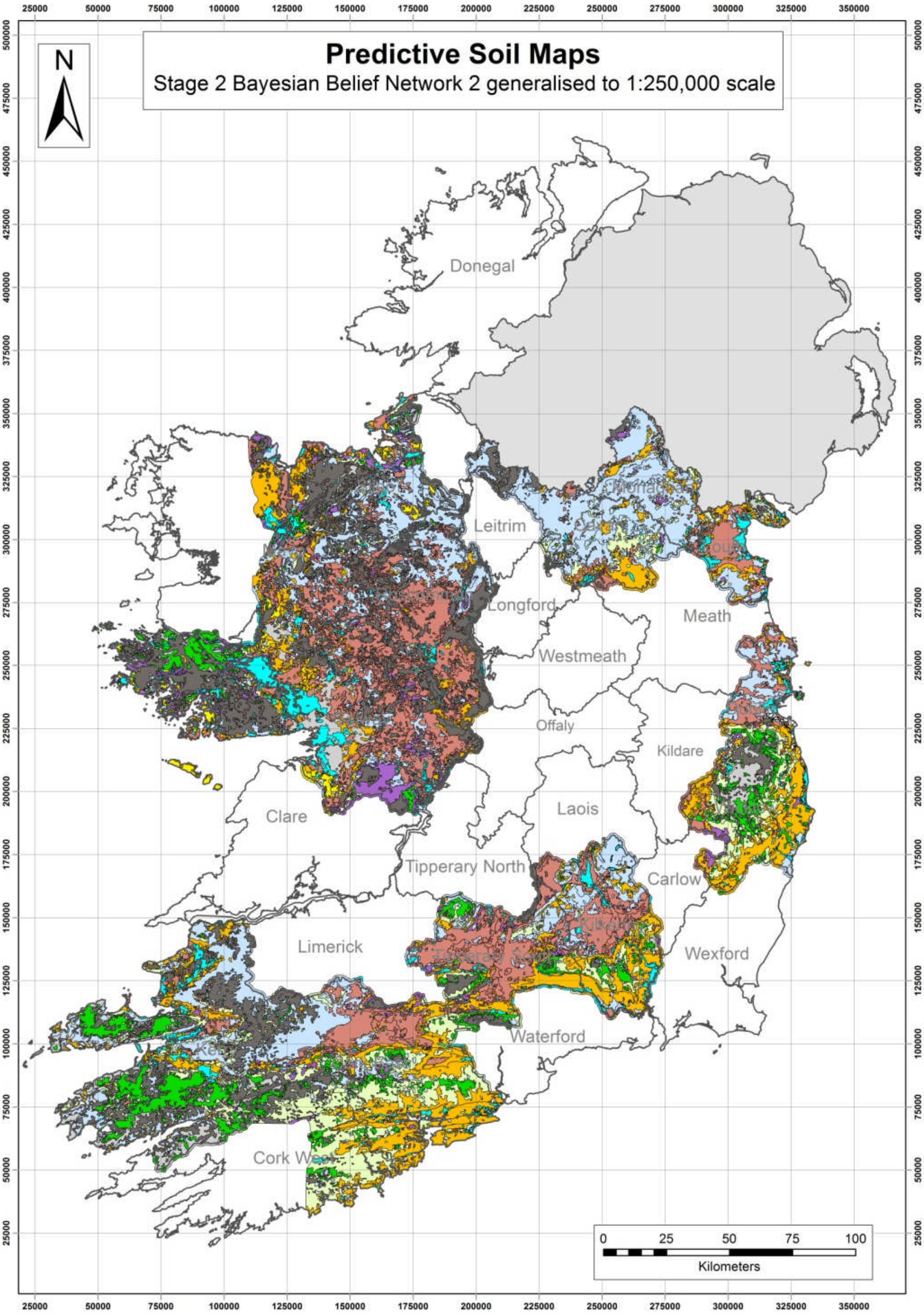


Figure 15: Stage 2 predictive soil map using Bayesian Belief Network – BN2 (generalised)

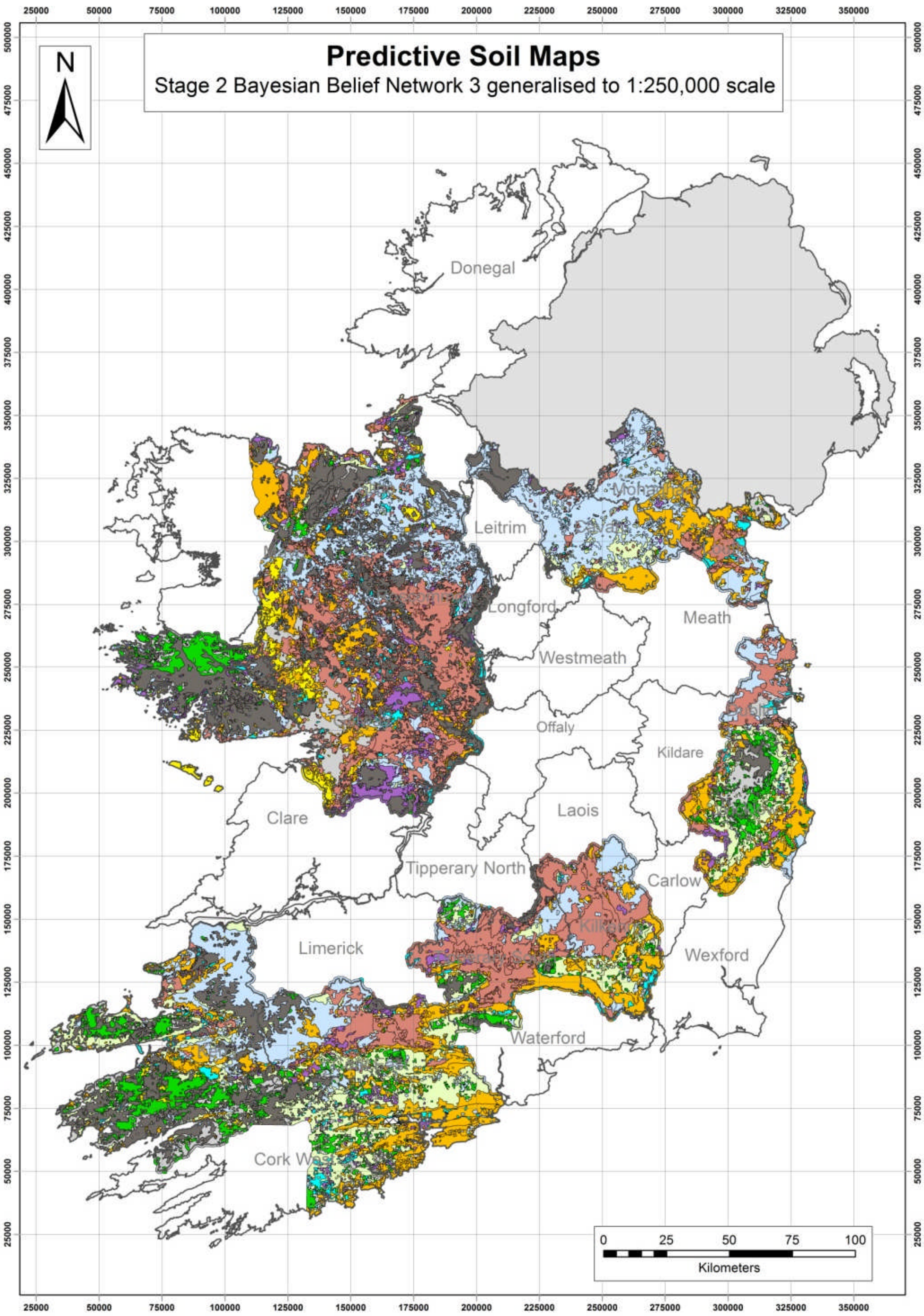


Figure 16: Stage 2 predictive soil map using Bayesian Belief Network – BN3 (generalised)

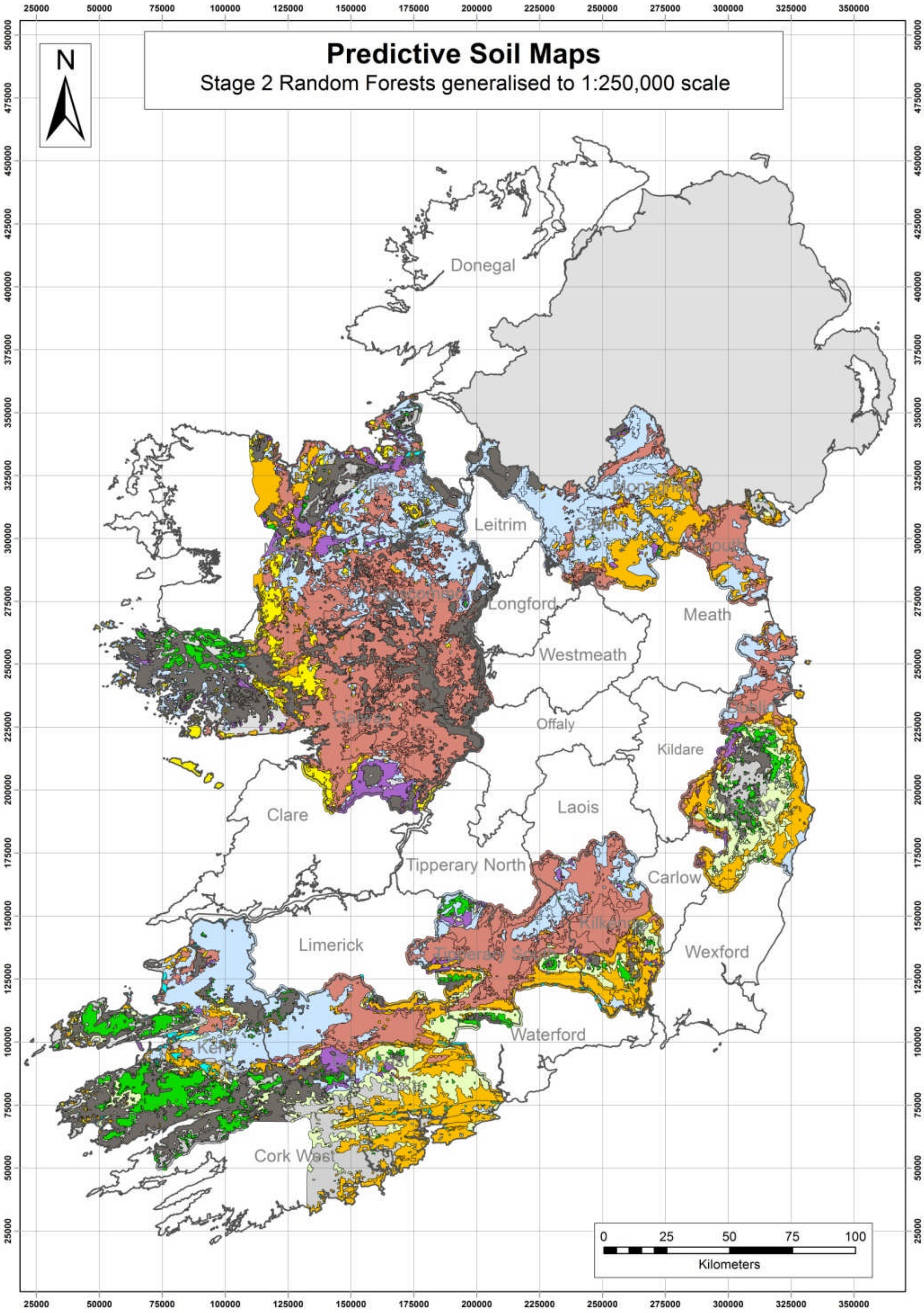


Figure 17: Stage 2 predictive soil map using Random Forests – RF2 (generalised)

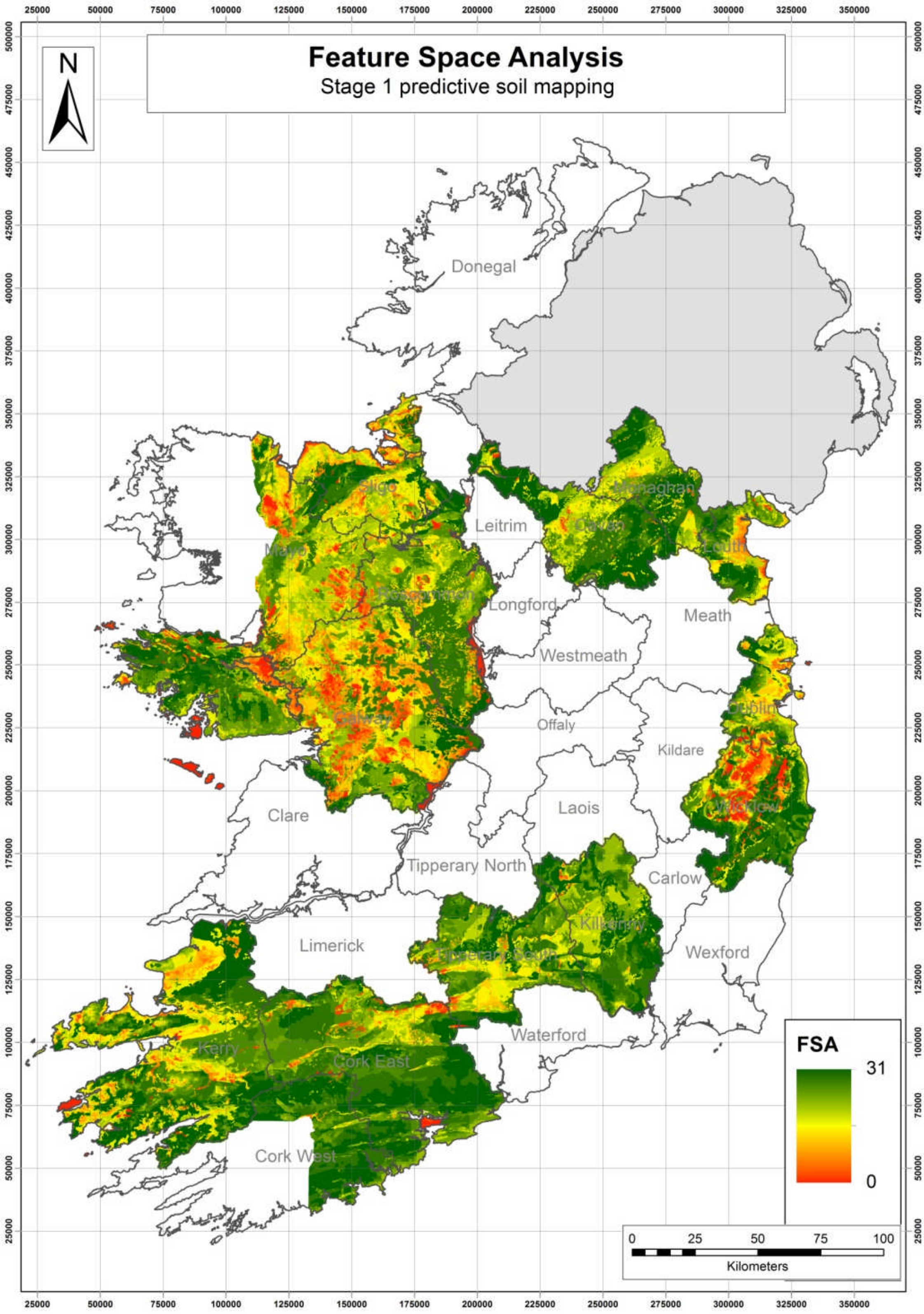


Figure 18: Stage 1 generic support assessment
(the legend shows the number of covariates available)

6. Support assessment

Methodology

Support assessment (generic)

The support assessment is expressed as the number of environmental covariates available in *terra incognita* compared with the relevant reference soilscape in *Terra Cognita*. The analysis is based on the number of classes in terms of categorical data and range in terms of numerical data.

Deployment assessment

Deployment assessment is expressed in terms of beliefs for each state of the response node in the case of Bayesian Networks and posterior (final) prediction probabilities for each response category in the case of Random Forests. In both cases the relevant values for the dominant association are used in the analysis.

Stage 1 – Environmental Distance

Support assessment (generic)

- Results of the feature space analysis for Stage 1 are shown in Figure 18
  Reference area: individual soilscapes
  Feature space analysis (0-31)

Deployment assessment

- Results for the Bayesian Belief Network model are shown in Figure 19.

Stage 2 – Induction Rule

Support assessment (generic)

- Results of the feature space analysis for Stage 2 are shown in Figure 20.
  Reference area: Individual soilscapes
  Feature space 0-31

Deployment assessment

- Results for BN2 are shown in Figure 21.
- Results for BN3 are shown in Figure 22.
- Results for RF are shown in Figure 23.

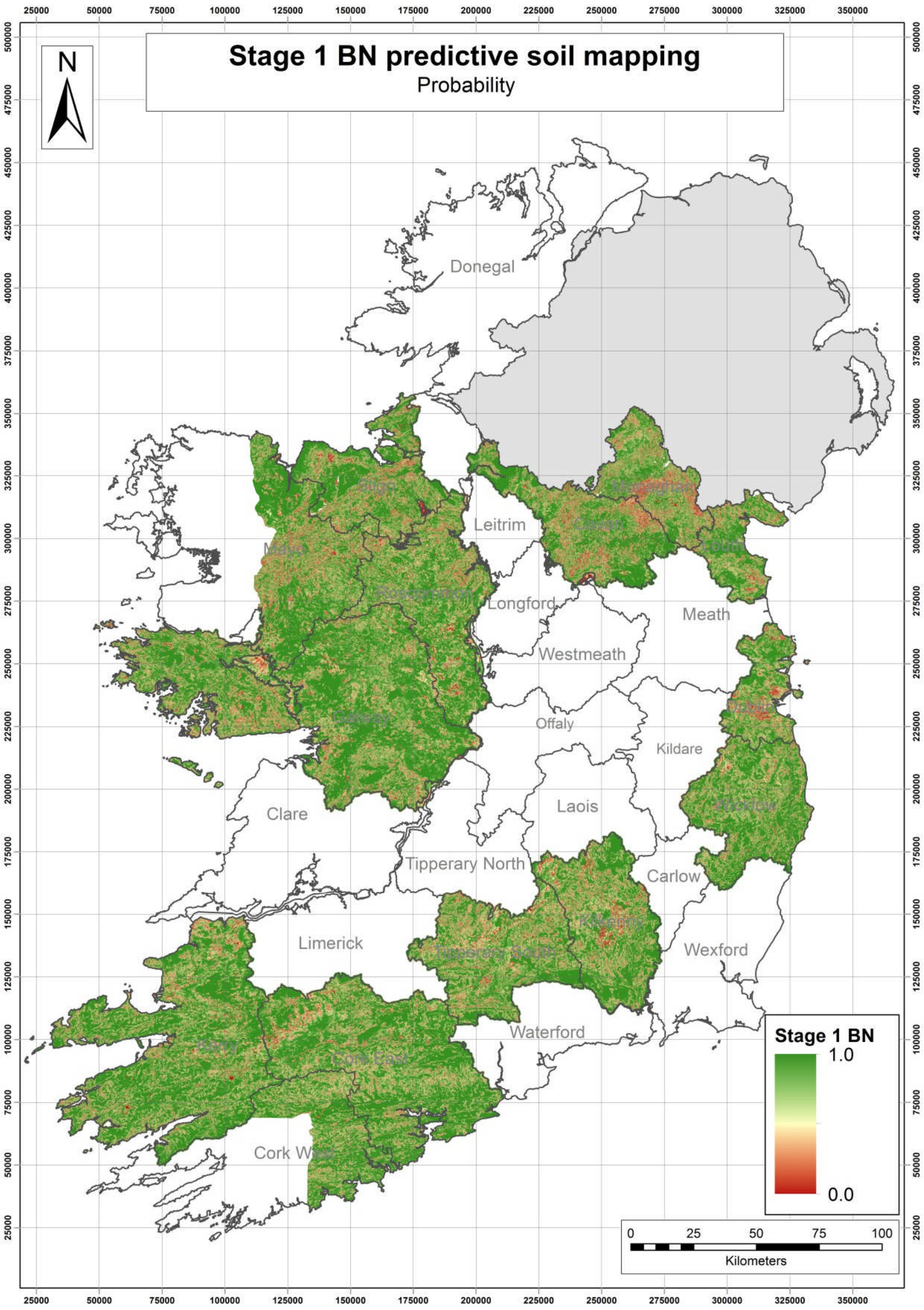


Figure 19: Stage 1 deployment assessment for BN – BN1

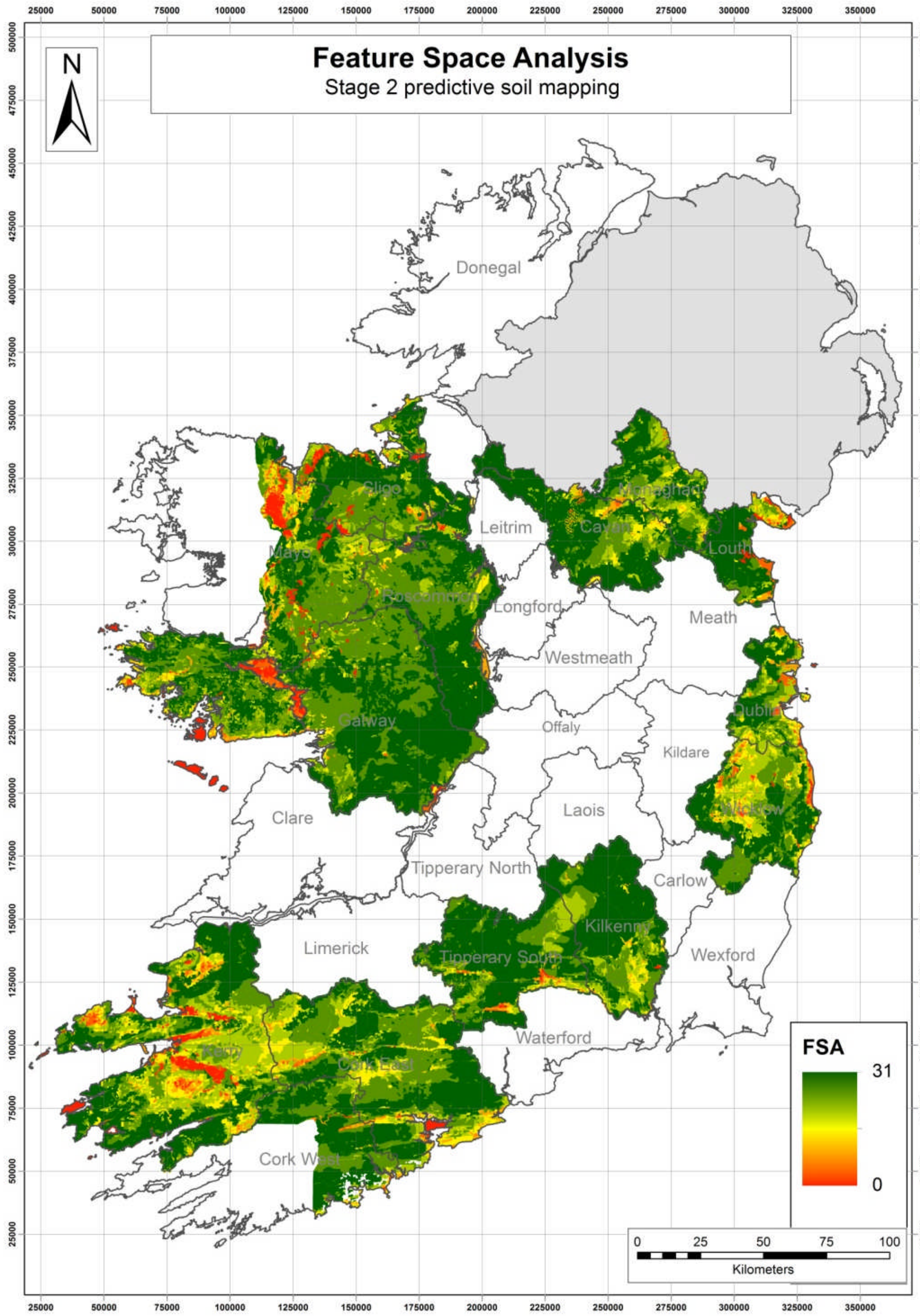


Figure 20: Stage 2 generic support assessment
(the legend shows the number of covariates available)

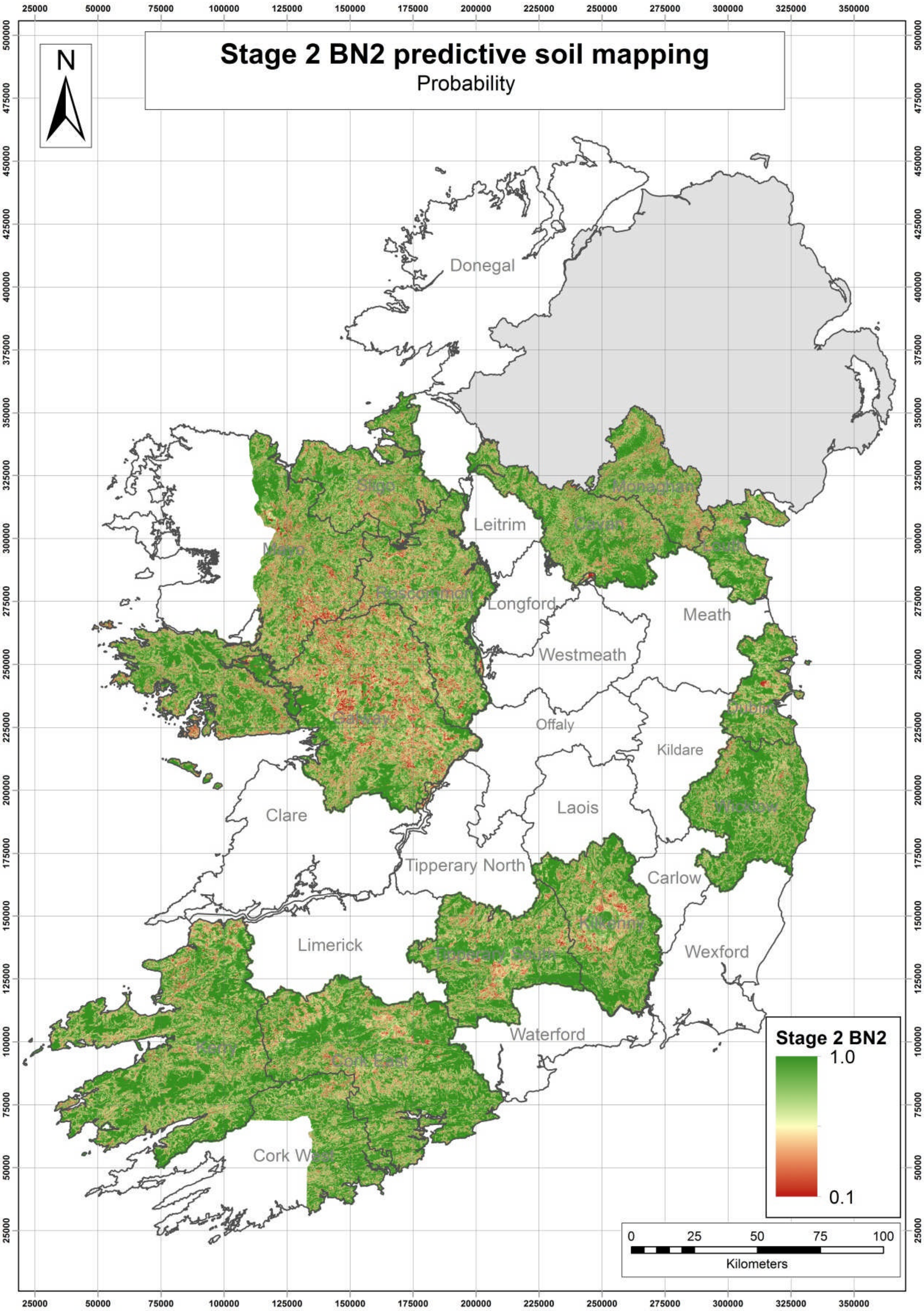


Figure 21: Stage 2 deployment assessment for Belief Network - BN2

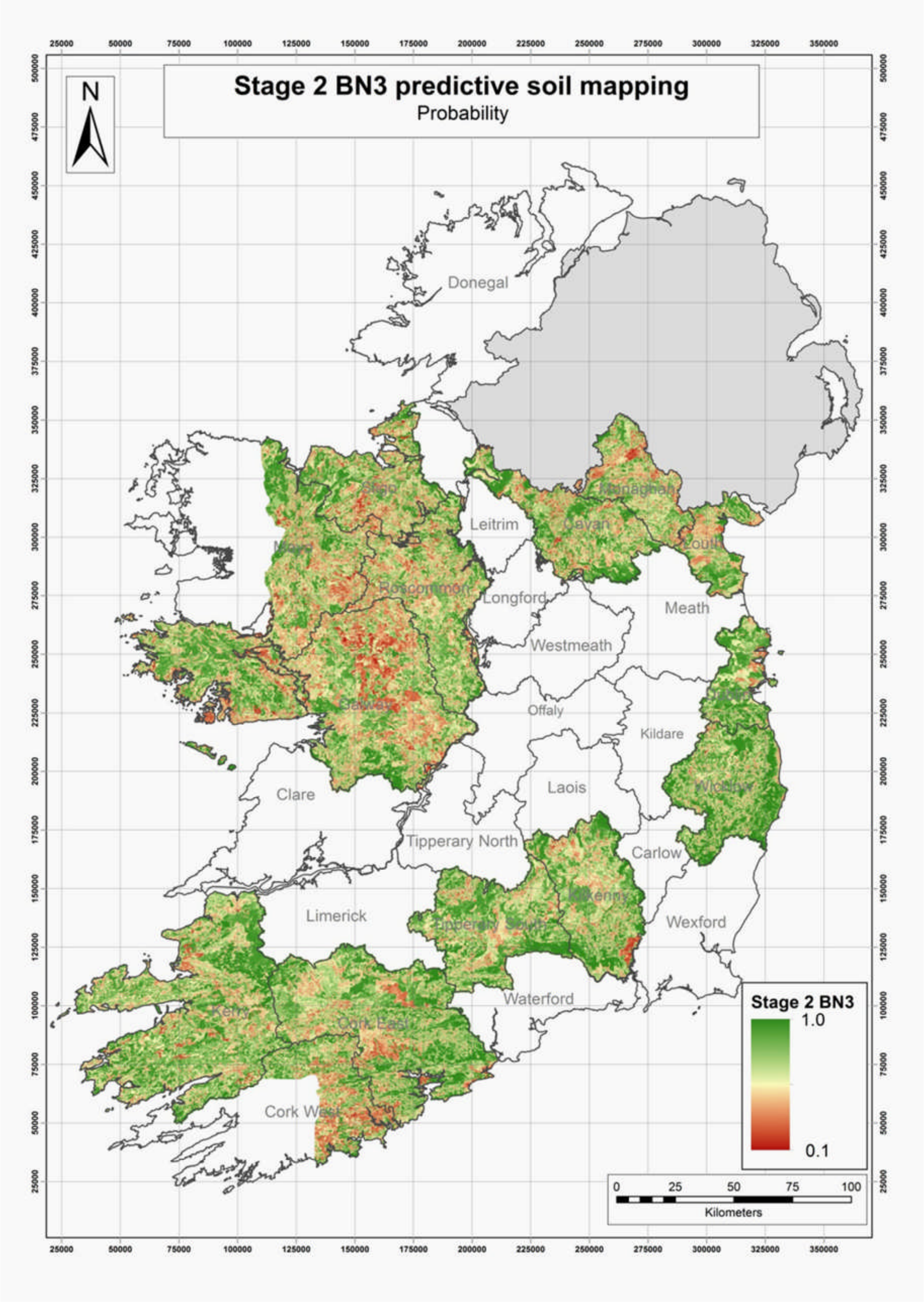


Figure 22: Stage 2 deployment assessment for Belief Network - BN3

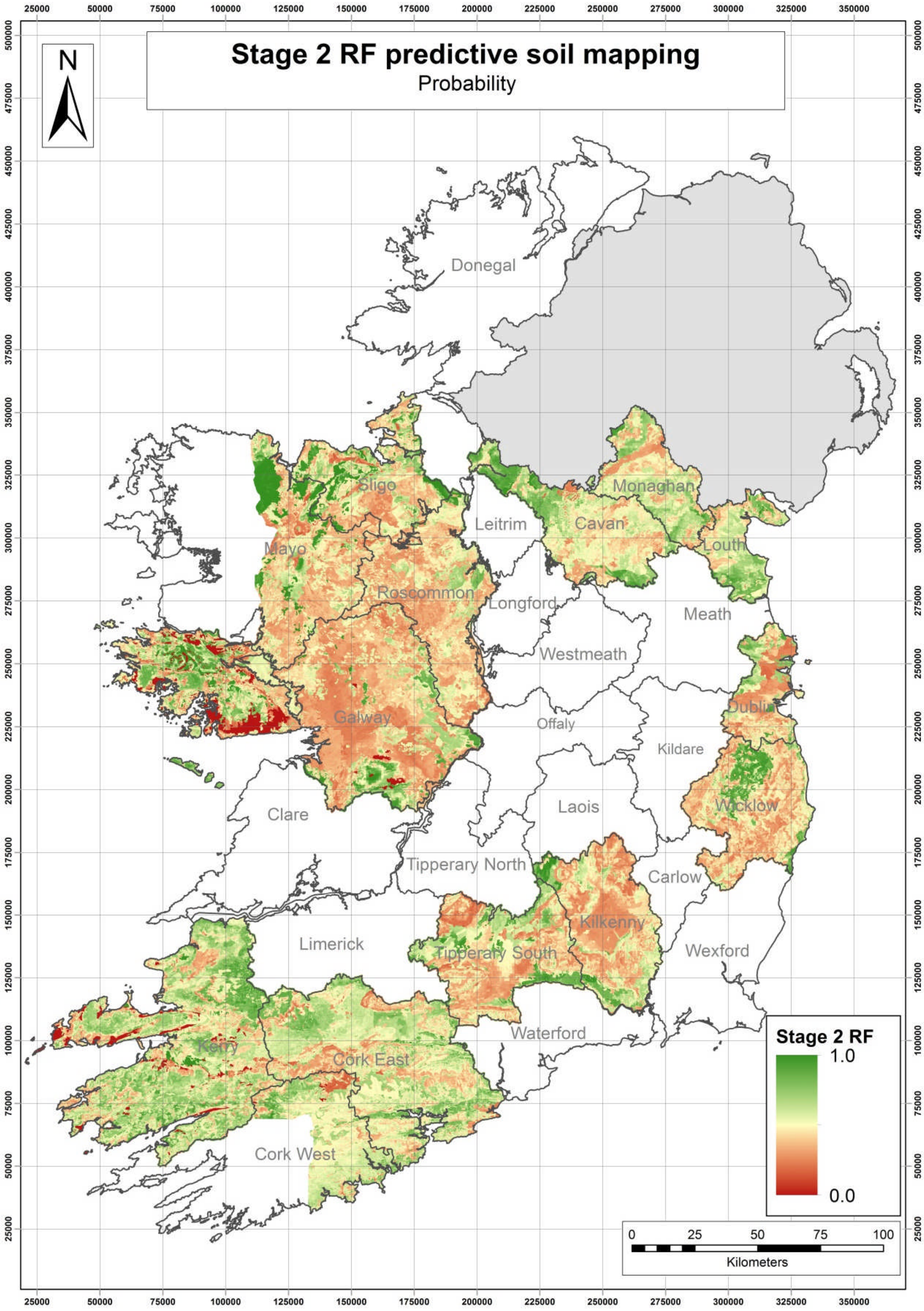


Figure 23: Stage 2 deployment assessment for Random Forest – RF2

7. Thematic assessment

Table 19 summaries the number of associations predicted by approach and Table 20 summaries the spatial extent.

Stage 1

Bayesian Belief Networks

- Extrapolation of soil associations predicted 121 of the 126 associations in *Terra Cognita* (Table 19) for raw predictions and 111 of 126 for the generalised map.
- The spatial extent of different soilscapes in *terra incognita* as predicted by Bayesian Belief Networks is listed in Table 20.

Random Forests

- Extrapolation of soil associations predicted 93 of the 126 associations in *Terra Cognita* (Table 19) for the raw predictions and 77 of the 126 for the generalised map.
- The spatial extent of different soilscapes in *terra incognita* as predicted by Random Forests is listed in Table 20.

Table 19: Thematic assessment – number of associations

Layer	Associations	Area (km2)
terra cognita	126	35696
ph1_bn_raw	121	35411
ph1_bn_250	111	35802
ph1_rf_raw	93	35402
ph1_RF_250	77	35848
ph2_BN2_raw	121	35370
ph2_BN2_250	117	35981
ph2_BN3_raw	121	35362
ph2_BN3_250	116	35987
ph2_RF_raw	93	35516
ph2_RF_250	82	36001

Stage 2

Bayesian Belief Networks

- Extrapolation of soil associations predicted 121 of the 126 associations in *Terra Cognita* (Table 19) for raw predictions and 117 of 126 for the generalised map.
- The spatial extent of different soilscapes in *terra incognita* as predicted by Bayesian Belief Networks is listed in Table 20.

Random Forests

- Extrapolation of soil associations predicted 93 of the 126 associations in *Terra Cognita* (Table 19) for the raw predictions and 82 of the 126 for the generalised map.
- The spatial extent of different soilscapes in *terra incognita* as predicted by Random Forests is listed in Table 20.

Table 20: Thematic assessment – spatial extent of associations

Terra	Terra	Stage 1	Stage 1	Stage 2	Stage 2	Stage 2
cognita	**cognita**	**BNk**	**RFk**	**BN2**	**BN3**	**RF**
Soil Assoc	**km2**	**km2**	**km2**	**km2**	**km2**	**km2**
112	39.1	487.0	12.0	63.9	85.6	1.3
113	155.5	19.9		484.8	586.8	
211	1235.7	1488.9	750.9	927.6	1022.7	1794.2
213	229.4	88.9	25.1	241.9	415.5	505.1
311	433.9	143.9	413.0	160.6	235.8	414.7
313	3508.8	8793.1	12333.8	4008.1	4096.6	4686.6
314	239.5	30.2	0.5	172.6	113.9	36.0
321	142.5	156.4	13.5	107.0	314.1	
411	324.1	201.3	38.4	508.4	698.6	71.0
414	6958.8	3813.8	3547.5	6453.0	6478.6	9225.0
511	694.3	79.1	472.4	93.2	109.3	481.0
512	1455.1	951.0	1313.9	1458.3	1382.4	629.3
513	86.5	39.7	198.9	74.6	84.7	470.2
514	924.6	4393.6	2259.6	1752.0	2148.3	1375.4
611	101.4	1.3	4.9	146.5	149.4	14.6
612	69.0	74.1	5.1	103.6	161.6	2.5
621	115.6	392.5	51.7	308.5	197.5	40.5
631	39.3	49.4	46.8	121.5	44.4	34.8
632	47.8	1.1		11.6	29.9	1.4
711	808.0	1099.8	877.5	1403.5	1522.3	1409.7
712	3790.8	2894.1	4721.7	5303.6	5529.5	6074.0
721	1624.1	2081.1	686.6	2121.6	1719.9	1319.2
722	199.4	31.0		17.9	6.5	29.6
723	1367.7	706.1	753.6	869.5	953.5	1004.8
724	349.3	690.2	359.6	57.2	112.0	
811	122.4	7.3		37.2	128.7	
812	78.2	7.8	0.5	11.7	13.9	
820	5.1	9.4				
821	178.1	235.1	38.0	138.0	159.4	0.8
822	266.3	632.2	1.0	238.5	114.2	
823	564.2	796.3	695.1	1047.0	408.1	190.1
824	19.9			11.0	9.6	
911	65.2	28.1	28.2	130.0	24.2	
913	6578.6	3747.3	4688.5	4968.5	5172.3	4387.4
914	382.5	182.7	83.7	416.8	145.8	3.3
921	932.3	803.8	1275.6	428.6	740.4	840.6
922	16.2			3.8	1.9	
1011	610.1	501.3	51.5	1388.9	811.4	419.7
Isl	39.5					
NoData	1.1		98.5			538.2
Rock	3.0	2.4				
SM	1.2	7.9		5.5	2.2	
Urban	60.3	78.9		68.3	52.4	
WATER	831.2	53.7		115.9	2.7	
	35695.6	35801.5	35847.6	35980.6	35986.6	36001.2

8. Conclusions

Satisfactory accuracy levels were only achieved with Bayesian Belief Networks and Random Forests. Random Forests produced better results for both Stage 1 and 2 modelling with the exception for the optimised BN model in Stage 2. However, in both Stage 1 and Stage 2 inference modelling, Bayesian Belief Networks predicted a higher number of individual soil associations compared to Random Forest models (Table 19).

However, detailed analysis of the results from all models highlight the fact that results for individual soilscapes can vary quite substantial both within specific models as well as between different models. Although these results are purely based on the training accuracies of individual inference engines, they indicate that there is scope to exploit those differences using a Multiple Classifier System approach.

A true assessment of the performance of individual models will only be possible using field observations.

9. Recommendations

The important message is that different combinations soilscape and soil association models appear to produce different results and these differences could be exploited. If the validation results will confirm these results, a combined map should be generated based on a Multiple Classifier System.

Acknowledgements

The authors would like to thank Khaled Taalab for his contributions (as indicated) on the descriptions of Bayesian Belief Networks, Neural Networks and Random Forests

References

Agyare, W. A., Park, S. J. and Vlek, P. L. G. (2007). Artificial neural network estimation of saturated hydraulic conductivity, Vadose Zone Journal, 6, 423-431

Bishop, C. M. (1995). Neural networks: a principled perspective, Neural Networks - Producing Dependable Systems (ERA 95-0973)

Breiman, L. (2001). Random Forests, Mach. Learning, 45, 5-32

Buol, S.W., Hole, F.D. and McCracken, R.J. (1973) Soil genesis and classification. Univ. Press, Ames, IA, USA

Dokuchayev, V.V. (1886). In Materials for land evaluation of the Nizhyny Novgorod Governorate. Natural and historical part: Report to the Nizhyny Novogorod Governorate zemstoo. Vol. 1: Key points in the history of land evaluation in the European Russia, with classification of Russian Soils (391p.). St. Petersburg: Evdokimov Press (in Russian)

Dokuchayev, V.V. (1899). Report to the transcaucasian statistical committee on soil taxonomy in general and especially for the Transcaucasia. Horizontal and vertical soil zones (19p). Tiflis, Russia. Office Press of the Civilian Affairs Commander-in-Chief in the Caucasus (in Russia)

Florinsky, I.V. (2012). Digital Terrain Analysis in Soil Science and Geology. Elsevier 379p.

Foody, G.M. (2002). Status of land cover classification accuracy assessment. Remote Sens. Environ. 80:185-201.

Grimm, R., Behrens, T., Maerker, M. and Elsenbeer, H. (2008). Soil organic carbon concentrations and stocks on Barro Colorado Island - Digital soil mapping using Random Forests analysis, Geoderma, 146, 102-113.

Grinand, C. Arrouays, D., Laroche, B. and Martin, M.P. (2008). Extrapolation regional soil landscapes from existing soil map: Sampling intensity, validation procedures and integration of spatial context. Geoderma143:180-190.

Hannam, J.A., Jones, R.J.A., Zawadzka, R C Palmer, J. Truckell, I.G. & Mayr, T.R. (2012). Soilscapes for Ireland: Terra Cognita. ISIS Project Technical Report No2.2, pp. 25, Cranfield University.

Hough, R. L., Towers, W. and Aalders, I. (2010). The Risk of Peat Erosion from Climate Change: Land Management CombinationsAn Assessment with Bayesian Belief Networks, Hum. Ecol. Risk Assess., 16, 962-976.

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

Jones, R.J.A., Hannam, J.A., Palmer, R.C., Truckell, I.G., Creamer, R.E. and McDonald, E. (2011). Classification, Harmonisation and Generalisation of County Soil Maps of Ireland. ISIS_WP1_D1.1, 136 pp, Cranfield & Teagasc.

Jones, R.J.A., Hannam, J.A., Palmer, R.C., Truckell, I.G., Creamer, R.E. and McDonald, E. (2012). Classification, Harmonisation and Generalisation of County Soil Maps of Ireland. ISIS_WP1_D1.1 ver2, 142 pp, Cranfield & Teagasc.

Liaw, A., and Wiener, M. (2002). Classification and Regression by RandomForest. R News: The Newsletter of the R Project, 2(3):18–22. (http://cran.r-project.org/doc/Rnews/).

Lou, W. G. and Nakai, S. (2001). Application of artificial neural networks for predicting the thermal inactivation of bacteria: a combined effect of temperature, pH and water activity, Food Res. Int., 34, 573-579.

Maier, H. R. and Dandy, G. C. (2001). Neural network based modelling of environmental variables: A systematic approach, Math. Comput. Model., 33, 669-682.

Mayr T.R., Zawadzka J. E., Corstanje R., Lo Papa G., Fealy R. (2013). Soilscapes for Ireland: Extrapolation to Terra Incognita. ISIS Project Technical Report No. ISIS_WP2_D2.3, 64 pp., Cranfield University & Teagasc.

McBratney, A.B., Mendonca Santos, M.L. and Minasny, B. (2003). On digital soil mapping. Geoderma, 117(1-2):3-52.

Moran, J.M. and Bui, E.N. (2002). Spatial data mining for enhanced soil map modelling. Int. J. Geogr. Inf. Sci. 16(6):533-549.

Prasad, A. M., Iverson, L. R. and Liaw, A. (2006). Newer classification and regression tree techniques: Bagging and random forests for ecological prediction, Ecosystems, 9, 181-199.

Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1986). Learning Representations by Back-Propagating Errors, Nature, 323, 533-536).

Uusitalo, L. (2007). Advantages and challenges of Bayesian networks in environmental modelling. Ecological Modelling 203 (3–4), 312–318.

Zakharov S.A. (1927). A course of soil science. Moscow: Gosizdat (in Russian) 440pp.

Appendices

Appendices 4 - 11 are only available as Microsoft Excel worksheets

Appendix 1: Variable selection

Report on the Statistical Analysis of the 20k sample from county Tipperary.

The objective of this analysis was two-fold:

i) To trial a set of multivariate techniques to determine which might be useful for prediction of soil classes from DTM derived variables
ii) To determine an effective method at deriving the best subset of variables for subsequent prediction modelling.

What follows is interim in that it reflects the current status of the analysis and it is my expectation that this will be continuously be updated as the different forms, techniques are analyzed and trailed.

Discriminant Analysis (DA)

The objective of DA is to obtain a subset of predictors that formulated in linear combination best discriminate between groups or classes. It is a simple method in which variable selection is uncomplicated much in the line with multiple regression (forward selection, backward selection, mixed, best subsets). Stepwise DA can be executed in Statistica (Statistics, Multivariate Exploratory Techniques, General DA or DA) or in R (Package klaR; function greedy.wilks). What follows are results from the Statistica package. A number of variables with extremely large numbers were standardized; no further transformations were considered. This may be an avenue of interest for further research.

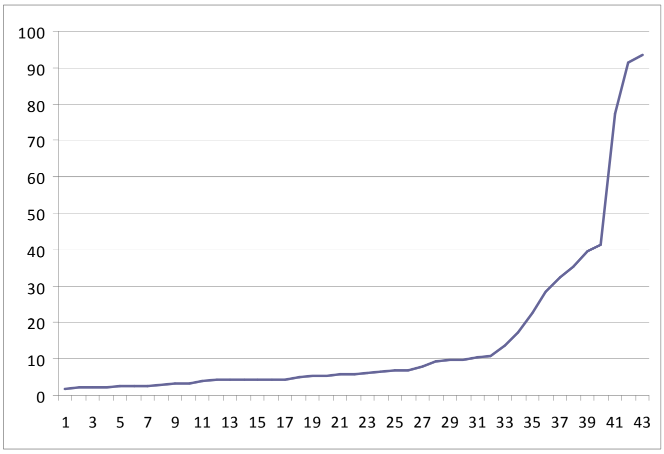
Overall performance. The overall ability of DA to discriminate between all series is with an accuracy of 39.89 %.

Subset performance.

Selection procedure	No selected	Percent correct (%)
Forward stepwise(F to enter = 1.0)	43	38.65
Best subsets	43	37.98
Backward stepwise(F to remove = 1.0)	43	39.75
Forward entry (p to enter = 0.005)	43	39.75
Backward removal (p to remove = 0.005)	43	39.75

As can be seen from the above table – little difference was observed between the different selection procedures. In all cases the same variables where selected (ASPECT, LNQAREA, NEWASPECT, PLAN, PROFILE, QAREA, QWETI, SLOPE, ANN_MAX, ANN_MEAN, "L2PIT", "L2STR", "LPIT2PK", "LSTR2DIV", "PPIT2PK", "PSTR2DIV", "N2CR", "N2PEAK", "N2ST", ANN_MIN, "PCTN2ST", "PCTZ2PIT", "PCTZ2ST", "PCTZ2TOP", "PMIN2MX", "Z2CR", "Z2PEAK", "Z2PIT", "Z2ST", "Z2TOP", "ZCR2PIT", "ZCR2ST", "ZPIT2PK", "ZTOP2PIT", TRI_TIPP, CROWT, DE_MART, LANG, P_PET, PRESCOTT, ANN_RADIAT, ANN_PREC, ANN_PE) . My expectation is that this was in part due to the difficulties discriminating a large number of series. Of interest here would be to repeat this exercise in which variable(s) are selected from an aggregate groups (climate or terrain) to predict a smaller number of more clearly defined soil series.

A simple approach to variable selection is based on the F statistic to enter. The adjacent figure shows to the F statistic on the y-axis and variable number on the x-axis. If an F to enter of 10 is chosen instead of F of 1; then the following variables would be used in the model: (LNQAREA, "L2STR", QWETI, "PPIT2PK", "PCTZ2PIT", DE_MART, "Z2ST", "PCTZ2TOP", LANG, "PMIN2MX", "ZCR2ST", ANN_PE, ANN_RADIAT; in order of increasing F-value). The associated percent correct in this case is 35.85 (error statistic of 64.19 %). In subsequent modelling in Statistica – this is the continuous subset used for Cart and Random Forests.



Other points of interest regarding DA

- DA allows for a formal test of differences between groups of classes, thereby identifying soil series that are hard to distinguish from one another (tests.xls).
- DA will supply the mean and std of each predictor for each class (i.e. soil series), so the parameter space can be identified and analyzed for each soil series ().

Random Forests and CART

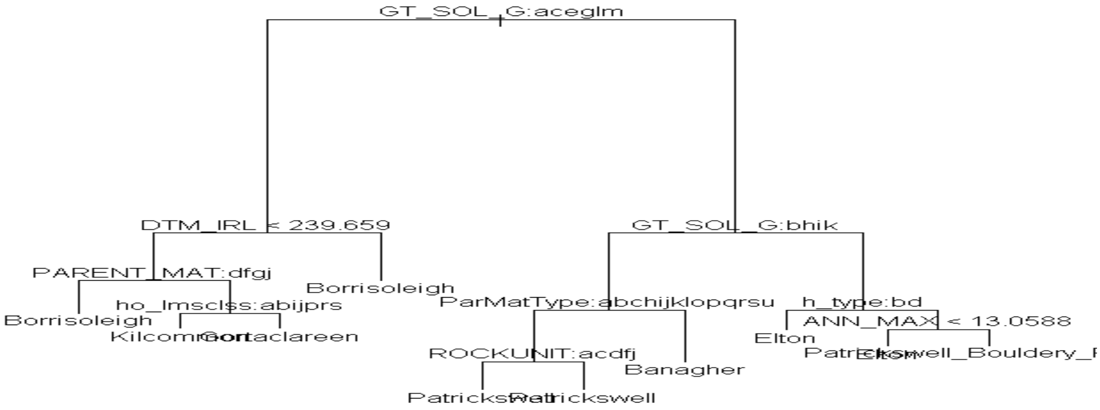
The objective of classification and regression trees (CART) and the complex algorithm of Random Forest (which I do not understand as yet completely) in which multiple classification regression trees are evaluated together. The advantage of using CART or the composite random forests is that both methods can easily accept categorical and continuous variables.

CART in R

CART in R is relatively simple and obtained under the package ‘tree’. The computational time is lengthy (overnight). Results from the classification of the soil series are:

- Variables actually used in tree construction: "GT_SOL_G" "DTM_IRL" "PARENT_MAT" "ho_lmsclss" "ParMatType" "ROCKUNIT" "h_type" "ANN_MAX"
- Number of terminal nodes: 10
- Residual mean deviance: 4.015 = 80270 / 19990
- Misclassification error rate: 0.573 = 11460 / 20001

The figure below presents the topology of the tree:



Note the low number of soil series that are actually depicted (vs 64 that have been requested for prediction. This in part because tree growth is limited to a depth of 31 by the use of integers to label nodes. The raw output from R further illustrates this (where * denotes a terminal node):

```
1) root 20001 123300 Patrickswell (2) GT_SOL_G:
Acid_Brown_Earth_75%,Blanket_Peat_100%_(High_level),Brown_Podzolic_80%,Gley_75%,Peaty_Podzol_75%,Podzol_70% 7203  39060
Borrisoleigh

    4) DTM_IRL < 239.659 4956  24580 Borrisoleigh

      8) PARENT_MAT:
Mostly_granite_sandstone,Ordovician_Silurian_Cambrian_shale_glacial_till,Ordovician_Silurian_Cambrian_shales_and_mica_schist,Sands
tone_granite_mica_schist_glacial_till 2899  10330 Borrisoleigh *

      9) PARENT_MAT: Mostly_sandstone,Peat,Sandstone_glacial_till 2057  11430 Kilcommon

       18) ho_lmsclss: A1a,A2a,B2a,B3a,C3a,C4a,C4b 1124   5416 Kilcommon *

       19) ho_lmsclss: A3a,A4a,B4a,B4b 933   4737 Gortaclareen*

    5) DTM_IRL > 239.659 2247  10390 Borrisoleigh *

  3) GT_SOL_G:
Basin_Peat_100%,Brown_Podzolic_60%,Gley_60%,Gley_90%,Grey_Brown_Podzolic_70%,Minimal_Grey_Brown_Podzolic_70%,Minimal_G
rey_Brown_Podzolic_80%,Water 12798  65880 Patrickswell

    6) GT_SOL_G: Basin_Peat_100%,Gley_90%,Grey_Brown_Podzolic_70%,Minimal_Grey_Brown_Podzolic_80% 8752  40180
PatrickswellCCCCCCC

     12) ParMatType:
Alluvium_undifferentiated,Basic_esker_sands_and_gravels,Bedrock_at_surface,Karstified_limestone_bedrock_at_surface,Lake_sediments
_undifferentiated,Limestone_sands_and_gravels_(Carboniferous),Limestone_till_(Carboniferous),Made_ground,Sandstone_and_shale_till
_(Lower_Palaeozoic),Sandstone_sands_and_gravels_(Devonian),Sandstone_sands_and_gravels_(Lower_Palaeozoic/Devonian),Sandstone
_till_(Devonian),Sandstone_till_(Lower_Palaeozoic/Devonian),Water 6645  25950 Patrickswell

       24) ROCKUNIT:
,Devonian_Old_Red_Sandstones,Dinantian_(early)_Sandstones_Shales_and_Limestones,Dinantian_Lower_Impure_Limestones,Silurian_M
etasediments_and_Volcanics 1413   7098 Patrickswell *

       25) ROCKUNIT:
Dinantian_Dolomitised_Limestones,Dinantian_Pure_Bedded_Limestones,Dinantian_Pure_Unbedded_Limestones,Dinantian_Upper_Impur
e_Limestones 5232  16490 Patrickswell *

     13) ParMatType: Cutover_Peat,Fen_Peat,Marl_(Shell) 2107   9783 Banagher*

    7) GT_SOL_G: Brown_Podzolic_60%,Gley_60%,Minimal_Grey_Brown_Podzolic_70%,Water 4046  19060 Elton

     14) h_type: OPM,PLA 1471   4849 Elton*

     15) h_type: PHM 2575  12570 Elton
```

```
30) ANN_MAX < 13.0588 2179   9772 Elton*

31) ANN_MAX > 13.0588 396   1408 Patrickswell_Bouldery_Phase*
```

CART is also possible in Statistica. Here we observe the problem that Statistica has limited the number of classes for prediction to < 50. The file 20k contains 65 series, so the 15 series least represented in the file where removed. These were; Carney (2), Louhgree_Landform (2), Slievereagh (2), Howardstown_Baggotstown (3), Rock_outcrop (3), Seefin (4), Kilcolgan (5), Puckane_Peaty_Phase (5), Kilgory (6), Knockaceol_Peaty_Phase (6), Kinvarra (8), Knockaceol (8), Derrygarreen (10), Milltownpass (13) and Coolalough (31); the total representing 0.5 % of all observed cases. In this case, there are no limits to the number of nodes. The tree figure is represented on the subsequent page.

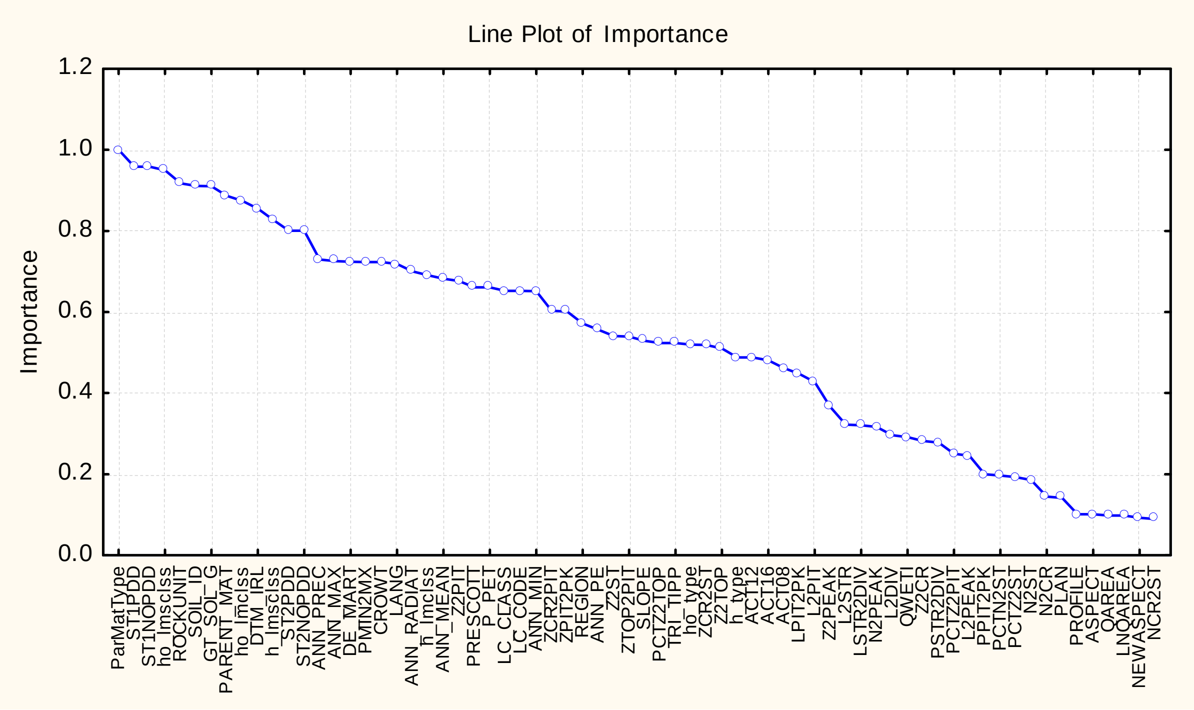
Performance measures:

Source	DF	-LogLike	RSquare
Model	833	21779.662	0.53
Error	19006	19279.14	.
C. Total	19839	41058.80	.
N	19856	.	.

Crossvalidation:

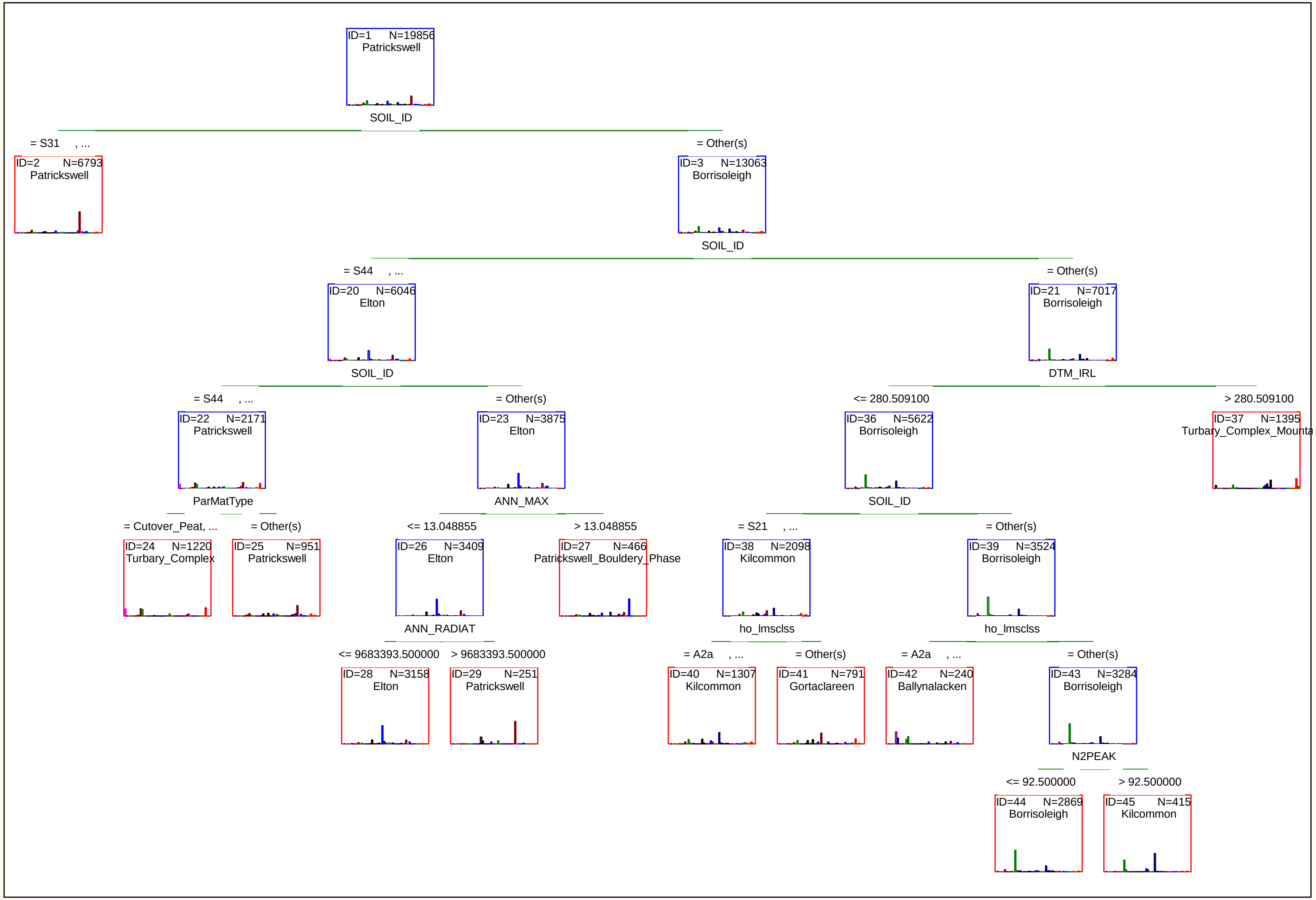
Risk estimates (CART) Dependent variable: Series Options: Categorical response

	Risk - estimate	**Standard - error**
Train	0.545377	0.003534
V-fold	0.570291	0.003514



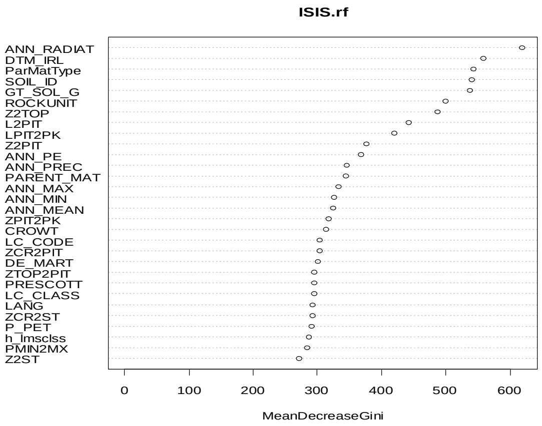
Tree 3 graph for Series

Num. of non-terminal nodes: 11, Num. of terminal nodes: 12



Random Forests in R

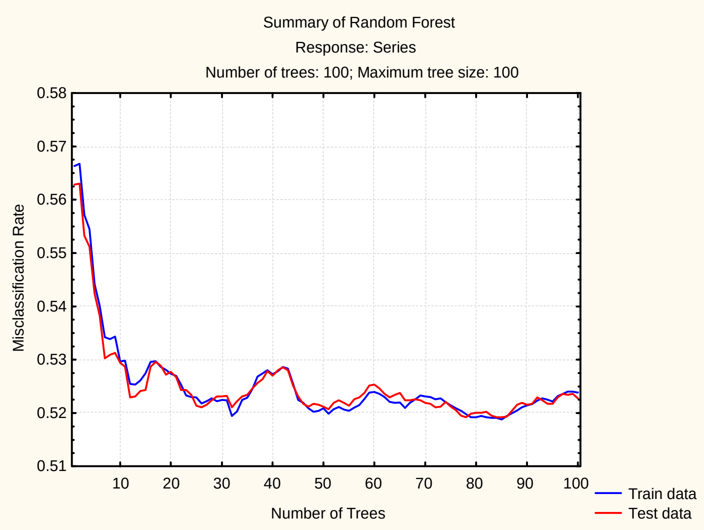
Overall Performance of the Random Forest is an OOB estimate of error rate: 22.68% (so 77.32 % accuracy). We are also able to obtain from this analysis the respective variable importance or sensitivity analysis, which is presented in the figure across.

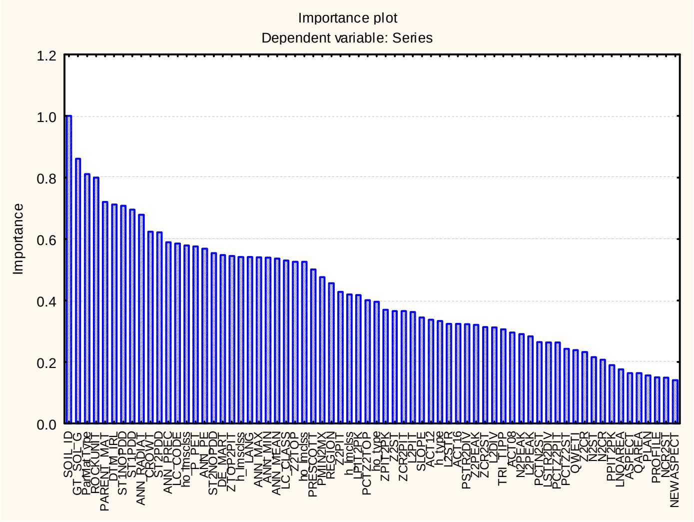


<u>Random Forests in Statistica.</u> Random Forest, as with CART has a limit of 50 soil series. It is otherwise simple to use.

Performance measures: **Crossvalidation:**

Source	DF	-LogLike	RSquare		Risk - Estimate	Standard - error
Model	637	14282.9361	0.56	**Train**	0.523952	0.004217
Error	13362	11106.3181	.	**Test**	0.523274	0.006522
C. Total	13999	25389.2542	.			
N	14012	.	.			

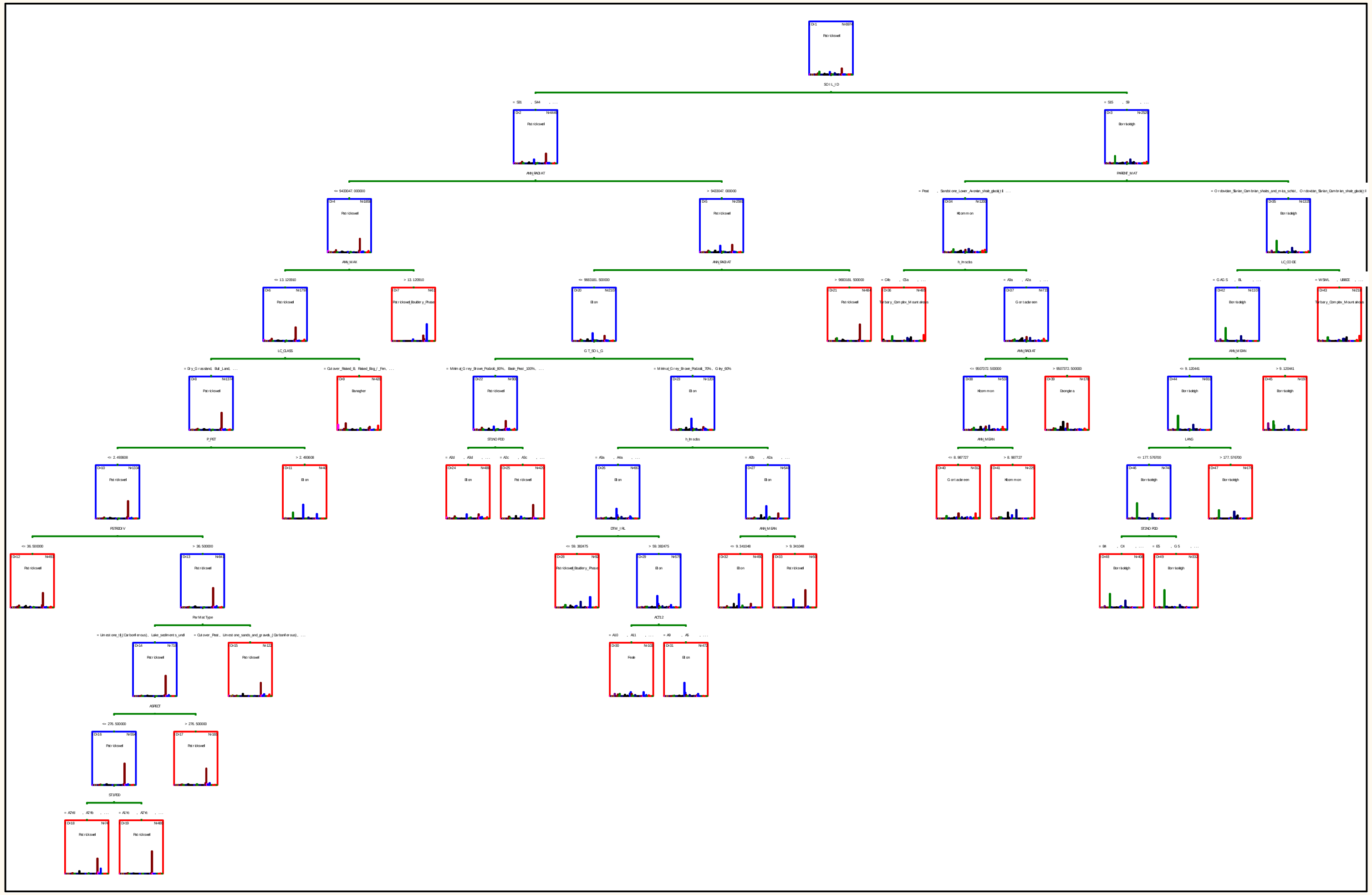




Tree graph for Series

Num. of non-terminal nodes: 24, Num. of terminal nodes: 25

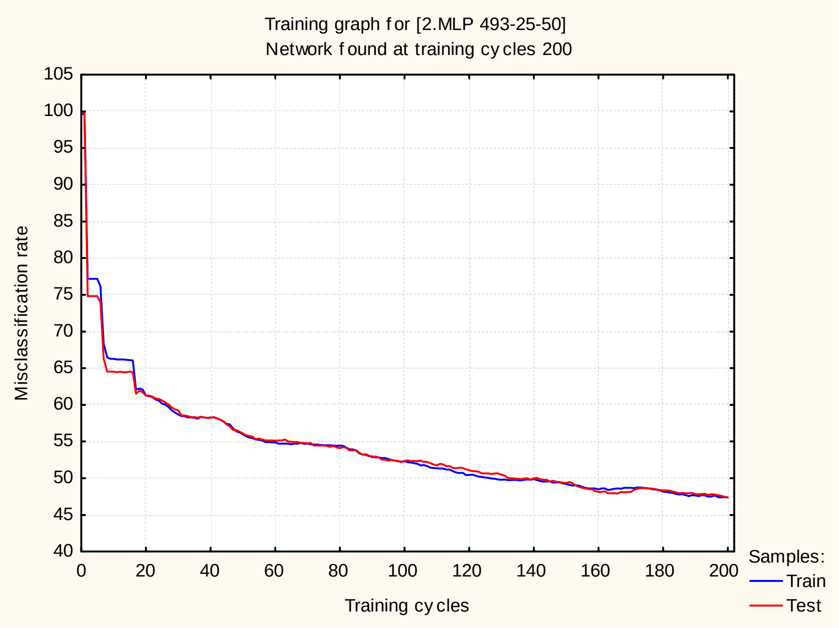
Tree number: 1



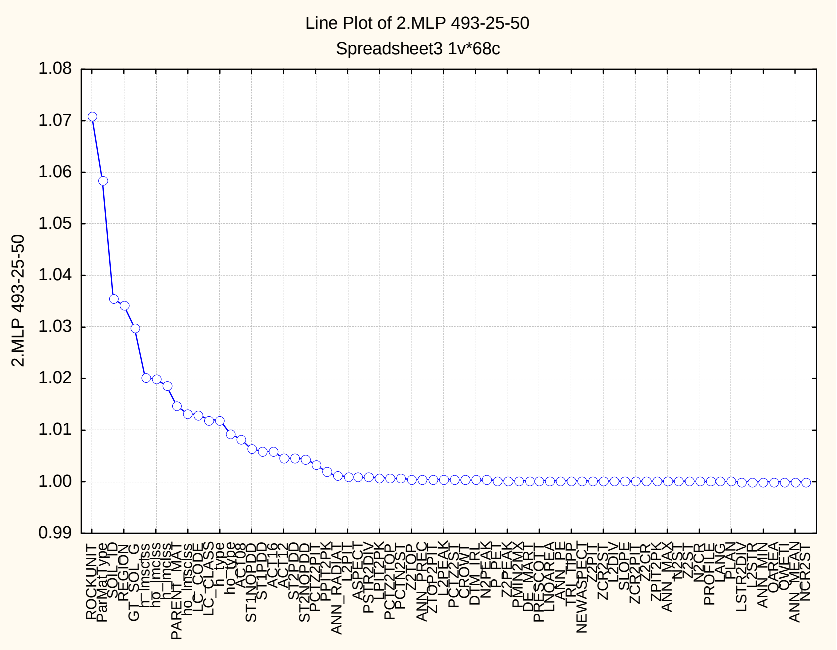
Neural Nets in Statistica

As part of the data miner package, Statistica contains a fairly extensive Neural Network package which I have not fully explored yet. There are three basic procedures for Neural Net fitting; i) fully automated, ii) custom neural net and one based on subsampling. What follows is based on a subsampling network, random subsampling. I found that the automated NN ran into memory problem very quickly. The software will produce a series of networks (I produced two but more can be entertained); overall performance was 47.50 % misclassification rate.

Performance measures



Variable selection. As with CART – the software will generate a sensitivity analysis of the network to the input variables:



Contingency Analysis and Correspondence analysis of the Netica Model. Tests table is analogous to the Analysis of Variance table for continuous data. The negative log likelihood for categorical data plays the same role as sums of squares in continuous data.

Overall performance. The overall performance of the Netica model at 20k can be represented in two ways – in the contingency analysis as anova or in terms of the kappa statistic.

	Source	DF	-LogLike	RSquare
20K				
	Model	4096	29199.81	0.473819
	Error	15840	32426.74	.
	C. Total	19936	61626.54	.
	N	20000	.	.

Some detail on the analysis (from the reference information of SAS software):

- The degrees of freedom for Model are used to compute the response rates for each factor level and are equal to (*s*–1)(*r*–1), where *r* is the number of response levels and *s* is the number of factor levels.
- LogLike is the negative log likelihood, which measures fit and uncertainty in the same manner that sums of squares does in continuous response situations:
- LogLike for C. Total is the negative log likelihood; (or uncertainty) when the probabilities are estimated by fixed rates for each response level. This corresponds to the total variation in an Analysis of Variance table.
- The Error –log likelihood (negative log likelihood) is the uncertainty calculated after fitting a model. This term is found by using separate observed response rates for each sample. The log likelihood fitting process minimizes this value. It is computed using
- Rsquare is the portion of the total uncertainty attributed to the model fit. An Rsquare of 1 means that the factors completely predict the categorical response. An Rsquare of 0 means that there is no gain from using the model instead of fixed background response rates. In categorical data analysis, high Rsquares are somewhat rare.

The kappa statistic:

The Kappa statistic (Agresti 1990) is shown in the Test table when the X and Y variables have the same set of values. Kappa measures the degree of agreement on a scale from zero to one. If two responses tend to agree, then most of the counts are on the diagonal. A Kappa of zero indicates that the responses are no more diagonal than expected from chance alone (from independent responses).

	Kappa Statistic	Std Error
20k	0.36	0.0034

Correspondence analysis:

Correspondence analysis is a multivariate technique to show which rows or columns of a frequency table have similar patterns of counts. In the correspondence analysis plot there is a point for each row and for each column. Define the row profile to be the counts in a row divided by the total count for that row. If two rows have very similar row profiles, their points in the correspondence analysis plot will be close together. Predictions are poor (poor correspondence) for those with relatively high numbers whereas correspondence is good for those with better predictions

Correspondence loadings for observed and Predicted soil series.

Observed	c1	c2	Predicted	c1	c2
Allen	-0.85	0.11	Allen	-0.84	0.11
Aughty	1.99	0.62	Aughty	2.03	0.64
Baggotstown	0.18	-0.14	Baggotstown	-0.03	-0.09
Baggotstown_Crush	-0.84	0.09	Baggotstown_Crush	-0.80	0.08
Ballincurra	-0.86	0.09	Ballincurra	-0.86	0.09
Ballynalacken	0.91	-0.23	Ballynalacken	0.94	-0.27
Ballynalacken_Knockshigowna	0.02	-0.13	Ballynalacken_Knockshigowna	0.03	-0.16
Ballyshear	-0.86	0.10	Ballyshear	-0.89	0.11
Ballyshear_Patrickswell	-0.79	0.07	Ballyshear_Patrickswell	-0.84	0.09
Banagher	-0.66	0.05	Banagher	-0.67	0.06
Boora_Complex	-0.92	0.13	Boora_Complex	-0.98	0.16
Borrisoleigh	1.04	-0.22	Borrisoleigh	1.16	-0.20
Borrisoleigh_Ballynalacken	1.14	-0.28	Borrisoleigh_Ballynalacken	1.13	-0.32
Borrisoleigh_Knockshigowna	1.38	-0.13	Borrisoleigh_Knockshigowna	1.55	0.01
Borrisoleigh_Steep_Phase	1.25	-0.21	Borrisoleigh_Steep_Phase	1.43	-0.12
Burren	-0.89	0.10	Burren	-0.93	0.11
Burren_Rocky_Phase	-0.82	0.07	Burren_Rocky_Phase	-0.86	0.09
Camoge	-0.61	0.00	Camoge	-0.67	0.02
Camoge_Milltownpass	-0.85	0.08	Camoge_Milltownpass	-0.90	0.09
Carney	-0.91	0.11	Carney	-0.93	0.12
Cooga	0.77	-0.25	Cooga	0.80	-0.30
Coolalough	-0.59	0.01	Coolalough	-0.66	0.02
Derrygarreen	1.04	-0.21	Derrygarreen	1.19	-0.18
Doonglara	0.66	-0.28	Doonglara	0.63	-0.31
Dovea	-0.45	-0.08	Dovea	-0.48	-0.09
Drombanny	-0.86	0.10	Drombanny	-0.93	0.12
Elton	-0.20	-0.11	Elton	-0.28	-0.09
Feale	0.36	-0.19	Feale	0.38	-0.21
Gortaclareen	0.59	-0.24	Gortaclareen	0.55	-0.24
Gortnamona	-0.80	0.09	Gortnamona	-0.73	0.07
Howardstown	-0.60	0.00	Howardstown	-0.65	0.01
Howardstown_Baggotstown	-0.69	0.01	Howardstown_Baggotstown	-0.77	0.04
Howardstown_Patrickswell	-0.72	0.04	Howardstown_Patrickswell	-0.77	0.06
Kilcolgan	-0.89	0.11	Kilcolgan	-0.96	0.13
Kilcommon	0.90	-0.17	Kilcommon	0.98	-0.18
Kilcommon_Peaty_Phase	1.31	0.03	Kilcommon_Peaty_Phase	1.52	0.18
Kilgory	0.17	-0.20	Kilgory	0.04	-0.20
Kinvarra	-0.89	0.10	Kinvarra	-0.93	0.12
Knockaceol	0.68	-0.33	Knockaceol	0.73	-0.40
Knockaceol_Peaty_Phase	1.45	0.25	Knockaceol_Peaty_Phase	1.56	0.31
Knockastanna	1.44	-0.01	Knockastanna	1.67	0.13
Knockastanna_Knockshigowna	1.83	0.30	Knockastanna_Knockshigowna	1.83	0.26
Knockastanna_Peaty_Phase	1.70	0.20	Knockastanna_Peaty_Phase	1.82	0.29
Knocknaskeha	0.48	-0.28	Knocknaskeha	0.53	-0.33
Knocknaskeha_Doonglara	0.18	-0.27	Knocknaskeha_Doonglara	0.22	-0.32
Louhgree_Landform	-0.56	0.04	Louhgree_Landform	-0.61	0.05
Milltownpass	-0.76	0.05	Milltownpass	-0.80	0.06
Mylerstown	-0.87	0.10	Mylerstown	-0.88	0.11
Patrickswell	-0.82	0.07	Patrickswell	-0.83	0.08
Patrickswell_Baggotstown	-0.80	0.07	Patrickswell_Baggotstown	-0.82	0.07
Patrickswell_Baggotstown_Elton	-0.17	-0.11	Patrickswell_Baggotstown_Elton	-0.19	-0.12
Patrickswell_Bouldery_Phase	-0.29	-0.08	Patrickswell_Bouldery_Phase	-0.39	-0.07
Patrickswell_Lithic_Phase	-0.87	0.10	Patrickswell_Lithic_Phase	-0.86	0.09

Observed	c1	c2	Predicted	c1	c2
Pollardstown	-0.66	0.03	Pollardstown	-0.66	0.03
Puckane	1.02	-0.24	Puckane	1.06	-0.21
Puckane_Gortaclareen	1.14	-0.10	Puckane_Gortaclareen	1.28	-0.04
Puckane_Peaty_Phase	1.43	0.13	Puckane_Peaty_Phase	1.54	0.16
Puckane_Slievereagh	0.52	-0.29	Puckane_Slievereagh	0.50	-0.28
Rock_outcrop	1.82	0.19	Rock_outcrop	1.96	0.23
Seefin	1.66	0.37	Seefin	1.78	0.46
Shannon_Banagher	-0.91	0.12	Shannon_Banagher	-0.97	0.15
Slievereagh	2.96	65.11	Slievereagh	3.18	79.57
Turbary_Complex	-0.82	0.10	Turbary_Complex	-0.77	0.10
Turbary_Complex_Mountainous	1.87	0.60	Turbary_Complex_Mountainous	1.91	0.72
Turbary_Knockastanna	1.85	0.27	Turbary_Knockastanna	1.93	0.33

Appendix 2: Sampling Tool

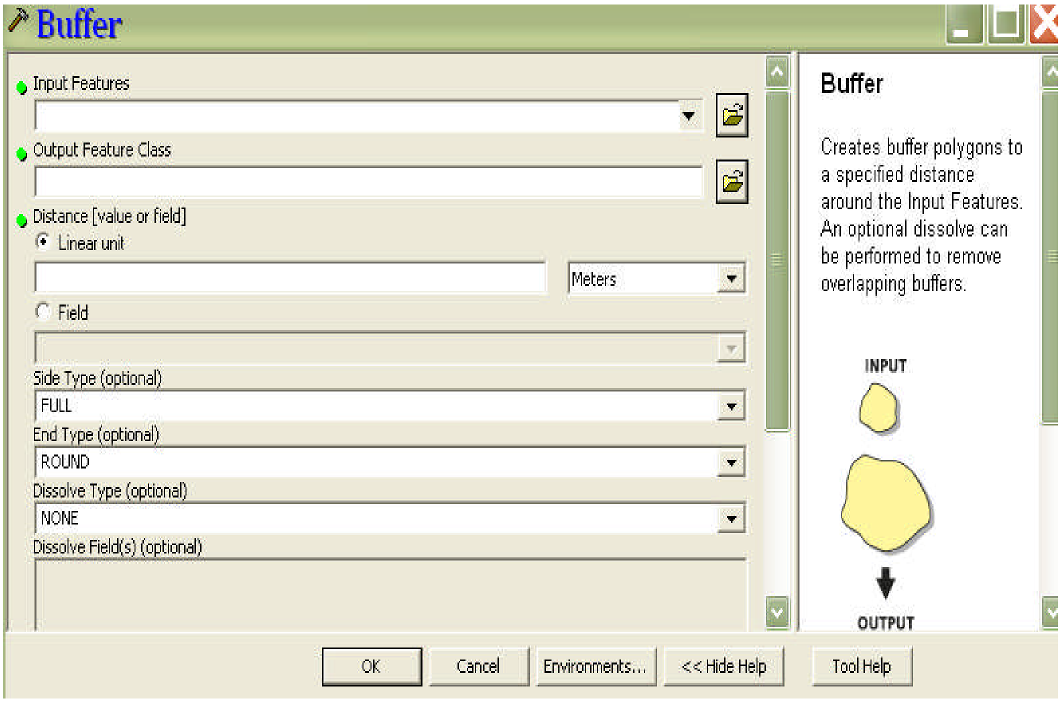
Report on sampling approaches for ISIS

There were two sampling approaches considered, simple random sampling and stratified random sampling. A number of tools exist that can do and three where considered:

- Hawths Analysis Tool
- NOAA sampling of biological populations
- Biogeographic Sampling design tool

The NOAA and biogeographic tools were obtained from the ESRI (http://arcscripts.esri.com/details.asp?dbid=15080 and http://arcscripts.esri.com/details.asp?dbid=15814) whereas a more up to date version of the Hawths tool can be obtained from (http://www.spatialecology.com/htools/rndpnts.php). Extraction of the zip file for the NOAA sampling tool was not successful so this tool was disregarded in the subsequent analyses and only the Hawths and Biogeographic analysis tools considered (which is also a NOAA product). In both sampling approaches, an initial filtering step is required to ensure that sampling points do not fall close or on class boundaries. This is particularly important in the stratified random sampling as stratification is driven by the classes.

Section 1: Buffering. A buffer strip can be obtained in ArcGIS™ using the tool Buffer (analysis). This can be found in the tool box, tab Index, scroll or type buffer in the keyword field. Select Buffer(analysis). This will generate the follow screen in Figure 1. For Input Features, select the feature layer for which a buffer has to be generated (e.g. soil map). Next identify the file name of the output file. For Distance select a distance for which the buffer needs to be generated. As in this case we wish to generate a buffer away from the boundary (i.e. around the boundary); this distance needs to be negative (e.g. -5 m). Next select outside only in side type and List for dissolve type. Click on the box of the layer for which you want to generate a buffer (e.g. Soil_ID) and hit OK. This should generate a buffer around the boundaries which is blank.



Buffer

Input Features

Output Feature Class

Distance [value or field]

● Linear unit

Meters

○ Field

Side Type (optional): FULL

End Type (optional): ROUND

Dissolve Type (optional): NONE

Dissolve Field(s) (optional)

Buffer

Creates buffer polygons to a specified distance around the Input Features. An optional dissolve can be performed to remove overlapping buffers.

Figure 1: Buffering tool screenshot

Section 2: Loading sampling tools. After unzipping these, an installation file needs to be run for Hawths Tools (htools_setup) and for the Biogeographical tool (SamplingTool_92SP6). Subsequently, in ArcGISTM, Hawths Tool can be found under tools → customize → toolbars; click on Hawths tools. Biogeographical tool can be found under tools → customize → commands (or just the command tab vs the toolbar tab earlier). Click and drag the Biogeographical icon (blue grid with red inserts) onto your toolbar.

Section 3: Simple random sampling. Both tools perform this adequately, Hawths under sampling tools and biogeographical tool does so directly.

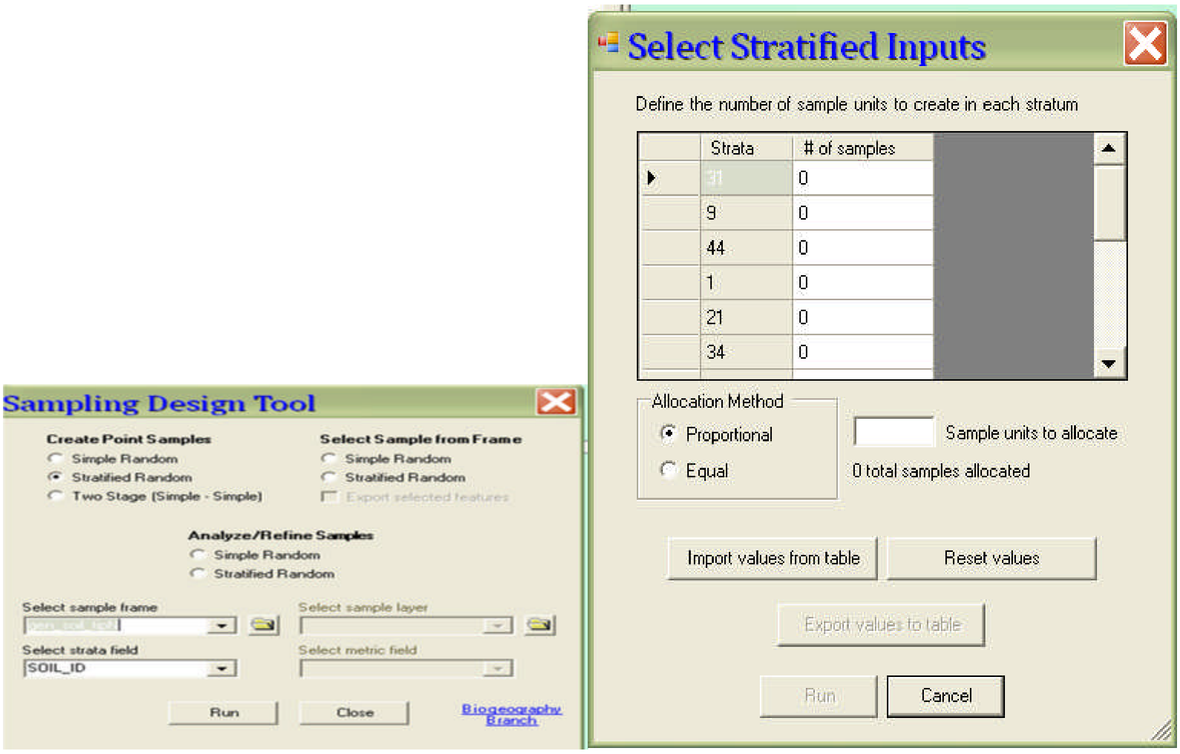
Section 4: Stratified random sampling. A comparison of the two tools was carried out in order to ensure that the stratification was based by area and not by counts of occurrence. This was tested on a data set which is summarized in the Table 1 in two forms, the count occurrences of each soil class and the respective areal extension. It is the latter that we wish the stratified sampling to represent. The last two columns are the observed probabilities obtained from the sampling occurrences, and reflect the number of samples per class.

Table 1. Table to show the observed and predicted occurrences of soil series.

			Observed Probabilitiesx	
Soil Class	**Count Prob**	**Areal Prob**	**Biogeographic**	**Hawths**
1	0.05797	0.016779	0.024975	0.054776
14	0.02174	0.026174	0.054945	0.029392
15	0.02174	0.002913	0.005994	0.018036
18	0.01449	0.002521	0.004995	0.017368
21	0.11594	0.104733	0.107892	0.126921
30	0.04348	0.064113	0.046953	0.04676
31	0.13043	0.199547	0.282717	0.115564
34	0.08696	0.199014	0.120879	0.072812
39	0.05072	0.073734	0.003996	0.032064
43	0.0942	0.034527	0.070929	0.110888
44	0.22464	0.05573	0.091908	0.228457
5	0.03623	0.010177	0.020979	0.045424
50	0.01449	0.034651	0.024975	0.014028
6	0.01449	0.003133	0.006993	0.018036
9	0.07246	0.072253	0.130869	0.069472

When comparing the sampling allocation by the two tools to the observed distributions of the soil series, then the biogeographic tool will respect the areal coverage of the soil class (corr = 0.8), which the Hawths Tool does not (corr = 0.3); inversely, the Hawths Tool respects the count probability (i.e. number of occurrences; corr = 0.98), which the Biogeographical tool does far less (corr = 0.6). The latter tool is also easier to use, far more transparent, better documented and significantly less computer intensive.

Section 5: Use of the biogeographic tool. The associated documentation to this tool is very good, so what follows is a short how to for stratified random sampling. Figure 2 shows a screenshot of the initial dialogue box for the tool. Select stratified random sampling, the feature with the classes (in this case the buffered feature) and field by which to stratify. After selecting 'run', a second screen will allow you to select the total number of samples needed. Select proportional for the areal coverage and equal for count coverage. In most of our cases, select proportional. Once a number of samples is selection, these are automatically allocated across the categories. Subsequently hit 'run' and follow-up box will allow you to specify the file in which the sampling points will be saved. A very similar approach is done in simple random sampling, in which only the total number of random points.



Create Point Samples: ● Stratified Random

Select strata field: SOIL_ID

Define the number of sample units to create in each stratum

Strata	# of samples
31	0
9	0
44	0
1	0
21	0
34	0

Allocation Method
● Proportional
○ Equal

Sample units to allocate

0 total samples allocated

Figure 2: Sampling tool screenshot

Appendix 3: Variable Derivatives

	Derivatives	**Source**	**Resolution**	**Number**
	c-climate			
1	Annual and Monthly mean Precipitation	EPA	1 km grid	
2	Annual and Monthly mean, min and max Temperature	EPA	1 km grid	
3	Annual and Monthly mean Solar Radiation	EPA	1 km grid	
4	Annual and Monthly mean Potential EvapoTranspiration	EPA	1 km grid	
5	Lang factor	Appendix 3b	1 km grid	
6	De Martonne Aridity Index	Appendix 3b	1 km grid	
7	Crowther Leaching factor	Appendix 3b	1 km grid	
8	FAO-UNEP Aridity Index	Appendix 3b	1 km grid	
9	Prescott Index	Appendix 3b	1 km grid	
10	Climatic Moisture Balance	Appendix 3b	1 km grid	
11	Thornthwaite Global Humidity Index	Appendix 3b	1 km grid	
12	Soil Climatic Index (Canadian)	Appendix 3b	1 km grid	
13	Emberger Pluviothermic Quotient	Appendix 3b	1 km grid	
14	Potential Soil Moisture Deficit	NSRI	1 km grid	
				14
	o-organisms			
15	Habitat Land cover map	EPA	1:50,000	
16	Landcover CORINE	JRC	1:100,000	
				2
	r-relief			
	LandmapR			
17	Slope gradient	LandMapR	20m	
18	Aspect	LandMapR	20m	
19	Profile Curvature	LandMapR	20m	
20	Planimetric Curvature	LandMapR	20m	
21	Diffuse upslope area	LandMapR	20m	
22	Wetness index	LandMapR	20m	
23	Ln of diffuse upslope area	LandMapR	20m	
24	New Aspect	LandMapR	20m	
25	Total distance to pit	LandMapR	20m	
26	Total distance to peak	LandMapR	20m	
27	Total distance from pit to peak	LandMapR	20m	
28	Relative distance from pit to peak	LandMapR	20m	
29	Total distance to channel	LandMapR	20m	
30	Total distance to divide	LandMapR	20m	
31	Total distance from divide to channel	LandMapR	20m	
32	Relative distance from cbannel to divide	LandMapR	20m	
33	Relief difference between cells and channel	LandMapR	20m	
34	Horizontal difference between cells and channel	LandMapR	20m	
35	Relief difference between cells and crests	LandMapR	20m	
36	Horizontal difference between cells and crests	LandMapR	20m	
37	Relief difference between cells and pits	LandMapR	20m	
38	Relief between cells and peaks	LandMapR	20m	
39	Horizontal distance between cells and peaks	LandMapR	20m	
40	Relief between cells and highest peak	LandMapR	20m	
41	Total elevation change from pit to peak	LandMapR	20m	
42	Elevation change from divide to channel	LandMapR	20m	
43	Elevation change from divide to pit	LandMapR	20m	
44	Total relief within catchment	LandMapR	20m	
45	Total horizontal distance change from divide to channel	LandMapR	20m	
46	Relative relief in respect to divide and channel	LandMapR	20m	
47	Relative relief in respect to pit and peak	LandMapR	20m	
48	Relative relief in respect to highest and lowest elevation	LandMapR	20m	
49	Relative horizontal distance in respect to divide and channel	LandMapR	20m	
50	Relative relief within dataset	LandMapR	20m	

	Derivatives	**Source**	**Resolution**	**Number**
	TAS			34
51	Average Flowpath Slope	TAS	20m	
52	Average upslope flowpath length	TAS	20m	
53	Maximum upslope flowpath length	TAS	20m	
54	Elevation above pit	TAS	20m	
55	Flow direction D8 (single-flow-direction)	TAS	20m	
56	Flow direction Rho8 (single-flow-direction)	TAS	20m	
57	Flow direction Dinf (multiple-flow-direction)	TAS	20m	
58	Flow direction FD8 (multiple-flow-direction)	TAS	20m	
59	Specific catchment area (From D8)	TAS	20m	
60	Specific catchment area (From D8 and log normalized)	TAS	20m	
61	Specific catchment area (From Dinf)	TAS	20m	
62	Specific catchment area (From Dinf and log normalized)	TAS	20m	
63	Specific catchment area (From DRho8)	TAS	20m	
64	Specific catchment area (From DRho8) and log normalized)	TAS	20m	
65	Specific catchment area (From FD8 with exponent parameter = 1)	TAS	20m	
66	Specific catchment area (From FD8 with exponent parameter = 1.1)	TAS	20m	
67	Specific catchment area (From FD8 with exponent parameter = 1.2)	TAS	20m	
68	Specific catchment area (From FD8 with exponent parameter = 1.5)	TAS	20m	
69	Number of upslope cells flow accumulation (From D8)	TAS	20m	
70	Number of upslope cells flow accumulation (From Dinf)	TAS	20m	
71	Number of upslope cells flow accumulation (From DRho8)	TAS	20m	
72	Total catchment area flow accumulation (from D8)	TAS	20m	
73	Total catchment area flow accumulation (from Dinf)	TAS	20m	
74	Total catchment area flow accumulation (from DRho8)	TAS	20m	
75	Number of Inflowing Neighbours (D8)	TAS	20m	
76	Number of Inflowing Neighbours (DRho8)	TAS	20m	
77	Number of Inflowing Neighbours (Dinf)	TAS	20m	
78	Compound Topographic index (Wetness Index) (from SCAD8)	TAS	20m	
79	Compound Topographic index (Wetness Index) (from SCADinf)	TAS	20m	
80	Compound Topographic index (Wetness Index) (from SCADRho8)	TAS	20m	
81	Compound Topographic index (Wetness Index) (from SCAFD8 with exponent parameter = 1)	TAS	20m	
82	Compound Topographic index (Wetness Index) (from SCAFD8 with exponent parameter = 1.1)	TAS	20m	
83	Compound Topographic index (Wetness Index) (from SCAFD8 with exponent parameter = 1.2)	TAS	20m	
84	Compound Topographic index (Wetness Index) (from SCAFD8 with exponent parameter = 1.5)	TAS	20m	
85	Maximum Downslope Elevation Change	TAS	20m	
86	Relative stream power index p=1	TAS	20m	
87	Relative stream power index p=2	TAS	20m	
88	Relative stream power index p=0.5	TAS	20m	
89	Relative stream power index p=1	TAS	20m	
90	Relative stream power index p=2	TAS	20m	
91	Relative stream power index p=0.5	TAS	20m	
92	Relative stream power index p=1.5	TAS	20m	
93	Relative stream power index p=3	TAS	20m	
94	Relative stream power index p=1	TAS	20m	
95	Relative stream power index p=2	TAS	20m	
96	Relative stream power index p=0.5	TAS	20m	
97	Relative stream power index p=1.5	TAS	20m	
98	Relative stream power index p=1	TAS	20m	
99	Relative stream power index p=2	TAS	20m	
100	Relative stream power index p=0.5	TAS	20m	
101	Relative stream power index p=1.5	TAS	20m	
102	Relative stream power index p=3	TAS	20m	
103	Downslope Flowpath Length	TAS	20m	
104	Maximum Downward Slope	TAS	20m	
105	Minimum Upslope Flowpath Length	TAS	20m	
106	Mean Upslope Elevation	TAS	20m	
107	Mean Upslope Slope	TAS	20m	
108	Network Wetness Index	TAS	20m	
109	Landform Classification	TAS	20m	
110	Surface Curvature Index	TAS	20m	

	Derivatives	Source	Resolution	Number
	SAGA			60
111	Analytical hillshading	SAGA	20m	
112	Slope	SAGA	20m	
113	Aspect	SAGA	20m	
114	Curvature	SAGA	20m	
115	Plan Curvature	SAGA	20m	
116	Profile Curvature	SAGA	20m	
117	Convergence index	SAGA	20m	
118	Curvature classification	SAGA	20m	
119	Watershed subbasins	SAGA	20m	
120	Catchment area	SAGA	20m	
121	Modified Catchment Area	SAGA	20m	
122	Specific Catchment area	SAGA	20m	
123	Catchment slope	SAGA	20m	
124	Wetness index	SAGA	20m	
125	Stream Power	SAGA	20m	
126	LS-Factor	SAGA	20m	
127	Altitude above channel network	SAGA	20m	
128	Channel network base level	SAGA	20m	
129	DEM normalized by Watershed subbasins	SAGA	20m	
130	Overland flow distance to channel network	SAGA	20m	
131	Vertical overland flow distance to channel network	SAGA	20m	
132	Horizontal overland flow distance to channel network	SAGA	20m	
133	Cell Balance	SAGA	20m	
134	Flow accumulation (Mass-flux method)	SAGA	20m	
135	Flow Path Length D8	SAGA	20m	
136	Flow Path Length MFD	SAGA	20m	
137	Flow width	SAGA	20m	
138	SAGA Wetness Index (t=5)	SAGA	20m	
139	SAGA Wetness Index (t=10)	SAGA	20m	
140	SAGA Wetness Index (t=15)	SAGA	20m	
141	SAGA Wetness Index (t=20)	SAGA	20m	
142	Difference between SAGA WIt5 and WIt10	SAGA	20m	
143	Difference between SAGA WIt5 and WIt15	SAGA	20m	
144	Difference between SAGA WIt5 and WIt20	SAGA	20m	
145	Mean difference (diff) between Saga Wetness Indices	SAGA	20m	
146	Mean diff between Saga Wetness Indices normalized by watershed subbasin	SAGA	20m	
147	Mean diff between Saga Wetness Indices normalized by catchments ('catchment new')	SAGA	20m	
148	Catchment Height (K)	SAGA	20m	
149	Catchment Slope (K)	SAGA	20m	
150	Catchment Height (Braunschweiger)	SAGA	20m	
151	Catchment Slope (Braunschweiger)	SAGA	20m	
152	Catchment Height (Dinf)	SAGA	20m	
153	Catchment Slope (Dinf)	SAGA	20m	
154	Catchment Aspect (Dinf)	SAGA	20m	
155	LS-Factor	SAGA	20m	
156	Visible Sky	SAGA	20m	
157	Diurnal Anisotropic Heating	SAGA	20m	
158	Downslope Distance Gradient	SAGA	20m	
159	Downslope Distance Gradient Difference	SAGA	20m	
160	Mass Balance Index	SAGA	20m	
161	Mid-Slope Positon	SAGA	20m	
162	Multiresolution Index of Valley Bottom Flatness (MRVBF)	SAGA	20m	
163	Multiresolution Index of the Ridge Top Flatness' (MRRTF)	SAGA	20m	
164	Normalized Height	SAGA	20m	
165	Protection Index	SAGA	20m	
166	Real Area Grid	SAGA	20m	
167	Surface Specific Point	SAGA	20m	
168	Sky View Factor	SAGA	20m	
169	Sky View Factor (Simplified)	SAGA	20m	
170	Slope Height	SAGA	20m	
171	Standardized Height	SAGA	20m	
172	Terrain View Factor	SAGA	20m	
173	Valley Depth	SAGA	20m	

	Derivatives	**Source**	**Resolution**	**Number**
	Spatial analyst			63
174	Mean curvature	Spatial analyst	20m	
175	Maximal curvature	Spatial analyst	20m	
176	Minimum curvature	Spatial analyst	20m	
177	Unspherical curvature	Spatial analyst	20m	
178	Total gaussian curvature	Spatial analyst	20m	
179	Difference curvature	Spatial analyst	20m	
180	Profile curvature	Spatial analyst	20m	
181	Tangential curvature	Spatial analyst	20m	
182	Flowpath curvature	Spatial analyst	20m	
183	Profile excess curvature	Spatial analyst	20m	
184	Tangential excess curvature	Spatial analyst	20m	
185	Total ring curvature	Spatial analyst	20m	
186	Total accumulation curvature	Spatial analyst	20m	
187	Land element classification	Spatial analyst	20m	
188	Backslope (Planar Slope)	Spatial analyst	20m	
189	Shoulder Slope	Spatial analyst	20m	
190	Foot slope	Spatial analyst	20m	
191	Hollow shoulder	Spatial analyst	20m	
192	Hollow foot	Spatial analyst	20m	
193	Hollow	Spatial analyst	20m	
194	Spur	Spatial analyst	20m	
195	Spure foot	Spatial analyst	20m	
196	Shoulder spur (Nose)	Spatial analyst	20m	
197	Peak	Spatial analyst	20m	
198	Pit	Spatial analyst	20m	
199	Plain	Spatial analyst	20m	
200	Ridge	Spatial analyst	20m	
201	Saddle	Spatial analyst	20m	
202	Channel	Spatial analyst	20m	
203	Maximum Membership	Spatial analyst	20m	
204	Entropy	Spatial analyst	20m	
205	Confusion Index	Spatial analyst	20m	
206	Topographic Position Index (kernel of 5 cells)	Spatial analyst	20m	
207	Topographic Position Index (kernel of 10 cells)	Spatial analyst	20m	
208	Topographic Position Index (kernel of 30 cells)	Spatial analyst	20m	
209	Slope Position Classification (kernel 5 cells)	Spatial analyst	20m	
210	Slope Position Classification (kernel 10 cells)	Spatial analyst	20m	
211	Slope Position Classification (kernel 30 cells)	Spatial analyst	20m	
212	Landforms from TPI (10-30)	Spatial analyst	20m	
213	Landforms from TPI (5-11)	Spatial analyst	20m	
214	Topographic Relative Moinsture Index	Spatial analyst	20m	
215	10 Class Landforms	Spatial analyst	20m	
216	Topographic wetness index	Spatial analyst	20m	
217	Flow Direction (Standard)	Spatial analyst	20m	
218	Flow Accumulation (Standard)	Spatial analyst	20m	
219	McNab's Landform Index	Spatial analyst	20m	
220	Terrain Shape Index	Spatial analyst	20m	
221	Pennock's 4 landforms	Spatial analyst	20m	
222	Pennock's 11 landforms	Spatial analyst	20m	
223	Pennock's 11 landforms filtered	Spatial analyst	20m	
224	Solar-radiation Aspect index	Spatial analyst	20m	
225	Hierarchically scaled assessment of Topographic position	Spatial analyst	20m	
226	Landform classification by topographic position	Spatial analyst	20m	
227	DEM normalized by GSM polygons	Spatial analyst	20m	
228	DEM normalized by GSM soil units	Spatial analyst	20m	
229	DEM normalized by Subsoil units	Spatial analyst	20m	
230	DEM normalized by Catchments	Spatial analyst	20m	
231	Openness	Spatial analyst	20m	
232	Topographic Ruggedness Index	Spatial analyst	20m	
233	Vector Ruggedness Measure (VRM)	Spatial analyst	20m	

	Derivatives	Source	Resolution	Number
	ArcGis			60
234	Actual drainage density based on LandMapR catchments		20m	
235	Actual drainage density based on fine catchments		20m	
236	Potential drainage density based on LandMapR catchments		20m	
237	Potential drainage density based on fine catchments		20m	
238	Actual/Potential drainage density ratio based on LandMapR catchments		20m	
239	Actual/Potential drainage density ratio based on fine catchments		20m	
240	Dissection Index based on LandMapR catchments		20m	
241	Dissection Index based on fine catchments		20m	
242	Relative Landscape Position		20m	
243	Relative Landscape Position (filtered)		20m	
	AML			10
244	Multiresolution Index of Valley Bottom Flatness		20m	
245	MRRTF		20m	
	Latest			2
246	Branch Length		20m	
247	Catchment morphometry		20m	
248	Percent Elevation between Channel and Divide		20m	
249	Potential Drainage Density		20m	
				4
	p-parent mateiral			
250	Subsoil map	EPA	1:50,000	
251	Bed rock geology	Geological Survey of Ireland		
252	General Soil Map	An Foras Talunta	1:575,000	
				3
	a-age			
253	Glaciation map			
				1
	n-landscape-context			
254	Hammond Landform Classification		20m	
255	SOTER Landform Classification		20m	
256	ACT (Automated Classification of Topography)		20m	
257	Dikau landform classification		20m	
258	Morap landform classification		20m	
259	Catchment analysis		20m	
260	ConMap		100m	
				7
Total				**260**

References

Bauer, J. / Rohdenburg, H. / Bork, H.-R. (1985): 'Ein Digitales Reliefmodell als Vorraussetzung fuer ein deterministisches Modell der Wasser- und Stoff-Fluesse', Landschaftsgenese und Landschaftsoekologie, H.10, Parameteraufbereitung fuer deterministische Gebiets-Wassermodelle, Grundlagenarbeiten zu Analyse von Agrar-Oekosystemen, (Eds.: Bork, H.-R. / Rohdenburg, H.), p.1-15.

Boehner, J., Koethe, R. Conrad, O., Gross, J., Ringeler, A., Selige, T. (2002): Soil Regionalisation by Means of Terrain Analysis and Process Parameterisation. In: Micheli, E., Nachtergaele, F., Montanarella, L. [Ed.]: Soil Classification 2001. European Soil Bureau, Research Report No. 7, EUR 20398 EN, Luxembourg. pp.213-222.

Boehner, J., Selige, T. (2006): Spatial Prediction of Soil Attributes Using Terrain Analysis and Climate Regionalisation' In: Boehner, J., McCloy, K.R., Strobl, J.: 'SAGA – Analysis and Modelling Applications', Goettinger Geographische Abhandlungen, Vol.115, p.13-27.

Bolstad, P.V., Swank, W., Vose J., 1998. Predicting Southern Applachian overstory vegetion with digital terrain data.. Landscape Ecology. 13: 271-283.

Dikau, R. 1989. The application of a digital relief model to landform analysis. In Three Dimensional Applications in Geographical Information Systems, ed. J. F. Raper, 51-77. London: Taylor and Francis.

Dikau, R., E. E. Brabb, and R. M Mark. 1991. Landform classification of New Mexico by Computer. U.S. Geological Survey Open File Report 91-634

Eyton, J. R. 1991. Rate-of-change maps. Cartography and Geographic Information Systems. 18: 87-103.

Freeman, G.T., 1991: 'Calculating catchment area with divergent flow based on a regular grid', Computers and Geosciences, 17:413-22.

Gallant, J.C., Dowling, T.I. (2003): 'A multiresolution index of valley bottom flatness for mapping depositional areas', Water Resources Research, 39/12:1347-1359'

Gruber, S., Peckham, S. (2008): Land-Surface Parameters and Objects in Hydrology. In: Hengl, T. and Reuter, H.I. [Eds.]: Geomorphometry: Concepts, Software, Applications. Developments in Soil Science, Elsevier, Bd.33, S.293-308.

Hantzschel, J., Goldberg, V., Bernhofer, C. (2005): 'GIS-based regionalisation of radiation, temperature and coupling measures in complex terrain for low mountain ranges'. Meteorological Applications, V.12:01, p.33–42, doi:10.1017/S1350482705001489.

Lea, N.L. (1992): 'An aspect driven kinematic routing algorithm', in: Parsons, A.J., Abrahams, A.D. (Eds.), 'Overland Flow: hydraulics and erosion mechanics', London, 147-175.

McNab, H.W. 1993. A topographic index to quantify the effect of mesoscale landform on site productivity. Canadian Journal of Forest Research. 23: 1100-1107.

McNab, H.W. 1989. Terrain shape index: quantifying effect of minor landforms on tree height. Forest Science. 35(1): 91-104.

Moeller, M., Volk, M., Friedrich, K., Lymburner, L. (2008): Placing soil-genesis and transport processes into a landscape context: A multiscale terrain-analysis approach. Journal of Plant Nutrition and Soil Science, 171, pp. 419-430, DOI: 10.1002/jpln.200625039.

Moore, I.D., Grayson, R.B., Ladson, A.R. (1991): 'Digital terrain modelling: a review of hydrogical, geomorphological, and biological applications' Hydrological Processes, Vol.5, No.1.

O'Callaghan, J.F., Mark, D.M., 1984: 'The extraction of drainage networks from digital elevation data', Computer Vision, Graphics and Image Processing, 28:323-344.

Olaya, V. Hidrologia computacional y modelos digitales del terreno. Alqua. 536 pp. 2004.

Parker, A. J. 1982. The Topographic Relative Moisture Index: An Approach to Soil-Moisture Assessment in Mountain Terrain. Physical Geography, 1982, 32, pp.160-168.

Pennock,D.J., Zebarth,B.J., and De Jong,E., 1987. Landform Classification and Soil Distribution in Hummocky Terrain, Saskatchewan, Canada. Geoderma, 40: 297-315.

Pennock,D.J. and Corre,M.D., 2001. Development and application of landform segmenation procedures. Soil & Tillage Research, 58: 151-162.

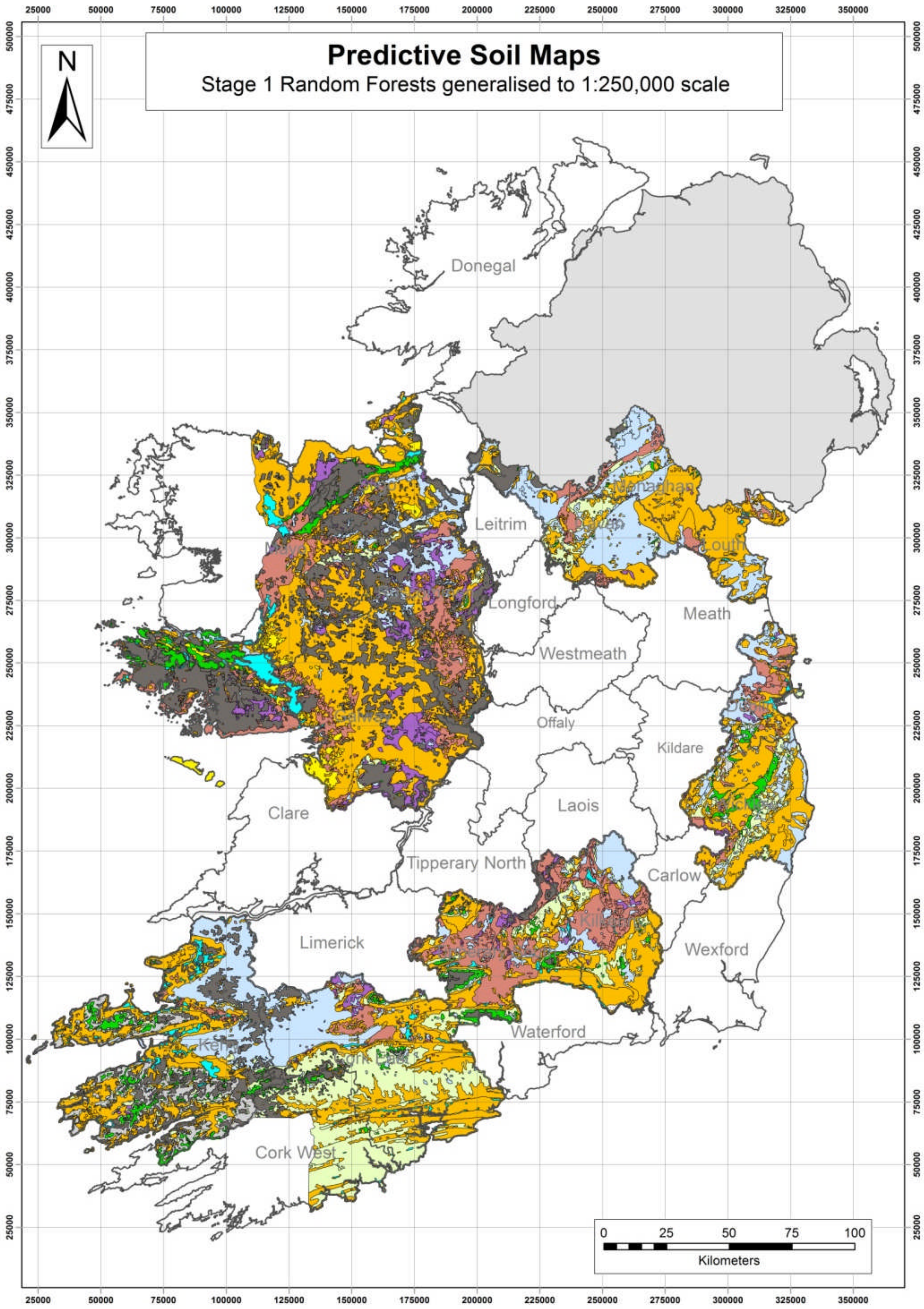
Peucker, T.K. and Douglas, D.H., 1975: 'Detection of surface-specific points by local parallel processing of discrete terrain elevation data', Computer Graphics and Image Processing, 4, 375-387.

Quinn, P., K. Beven, P. Chevallier and O. Planchon. 1991. The prediction of hillslope flow paths for distributed hydrological modelling using digital terrain models. Hydrological Processes. 5: 59-79.

Roberts. D. W., and Cooper, S. V., 1989. Concepts and techniques of vegetation mapping. In Land Classifications Based on Vegetation: Applications for Resource Management. USDA Forest Service GTR INT-257, Ogden, UT, pp 90-96.

Sappington, J.M., K.M. Longshore, and D.B. Thomson. 2007. Quantifiying Landscape Ruggedness for Animal Habitat Anaysis: A case Study Using Bighorn Sheep in the Mojave Desert. Journal of Wildlife Management. 71(5): 1419 -1426

Tarboton, D.G. (1997): 'A new method for the determination of flow directions and upslope areas in grid digital elevation models', Water Resources Research, Vol.33, No.2, p.309-319.



Predictive Soil Maps

Stage 1 Random Forests generalised to 1:250,000 scale