CRANFIELD UNIVERSITY

MAREK SMID

MODELING AND DATA FUSION FOR SOIL MAPPING

SCHOOL OF APPLIED SCIENCES
Geographical Information Management

MSc
Academic Year: 2011 - 2012

Supervisor: Dr. Ronald Corstanje
September 2012

CRANFIELD UNIVERSITY

SCHOOL OF APPLIED SCIENCES
Geographical Information Management

MSc Thesis

Academic Year 2011 - 2012

MAREK SMID

Predictive Modeling and Data Fusion for Soil Mapping

Supervisor: Dr. Ronald Corstanje
September 2012

This thesis is submitted in partial fulfilment of the requirements for the degree of MSc in Geographical Information Management

© Cranfield University 2012. All rights reserved. No part of this publication may be reproduced without the written permission of the copyright owner.

ABSTRACT

Stochastic predictive models and data fusion principles are used in wide range of industries (economics, management, systematic, biomedicine, robotics, meteorology and climatology). The project implements these principles to soil science. Predictive modeling is used to research prediction of phenomena, their changes or comparison with their current status. Predictions and classification models are increasingly used in Geographic Information Management (GIS) to determine different properties and help to delivery better spatial outputs. Currently in most of the projects different prediction methods are applied. Their outputs are compared with ground truth data and the best fitting model is chosen. This project seeks to explore an approach in which, instead of selecting particular method, multiple methods are applied and their outputs are combined such that the optimal thematic classification is based on data fusion of predictions.

The methodology is illustrated by a case study in South Tipperary – Republic of Ireland. Predicted covariate is Soil class. The training and deployment data sets were generated by extracting values for twelve environmental variables from GIS layers. Three different stochastic predictive models were executed using this data. Each of those models was preceded in different software environment: R-project for Random Forest, STATISTICA10 for Artificial Neural Network and Netica for Bayesian Belief Network. Three different data fusion approaches were applied for combine models together. VB.net and Netica SWs were used for data fusion.

The results were analysed by Data mining tools and cross tabulation was used for comparison between models. To evaluate outputs of models against ground truth data, the Cramer’s V statistic was used. This study has shown that the most important of environmental variables for forming soil classes are GSM, Actual/Potential drainage ratio and the parent material. Predictions of different models are best matching within Luvisol and Brown Podzolic soil classes. In the same time models tends to misclassify between this two soil classes. Additionally RF model disagree with ANN and classifies certain areas of Luvisol into Stagnogley soil class. Combining multiple thematic classifications using data fusion provides better and more justifiable thematic map than any single classifier.

Keywords:

Random Forest

Artificial Neural Network

Bayesian Belief Network

Predictive thematic classification

Probabilistic forecasting

ACKNOWLEDGEMENTS

My profound thanks to Dr Ron Corstanje, my Supervisor, for the guidance and constructive feedback provided during the research.

A big thank you also goes to Dr Thomas Mayr for his support and assistance.

I would like to express my sincere gratitude to the Joanna Zawadzka, Fabio Veronesi and Stefano Cavazzi for their help and constructive discussion during the project.

Thanks to my family for their support and understanding, while I was away for this programme. I say a big thank you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
AOI	Area of Interest
BBN	Bayesian Belief Network
BC	Boundary Conditions
CPT	Conditional Probability Table
DF	Data Fusion
GAM	General Additive Model
GIS	Geographical Information System
GLM	General Linear Model
IC	Initial Conditions
LOO	Leave-One-Out cross validation
LUIM	Land Use Impact Model
MC	Model Classes
MLP	Multilayer Perceptron
MSE	Mean Square Error
MP	Model Parameters
SW	Software
RBF	Radial Basic Function
RDA	Redundant Analysis
RF	Random Forest
RSME	Root Mean Square Error

1 Introduction

Predictive modeling is used to research prediction of phenomena, their changes or comparison with their current status. Prediction and classification models are increasingly used in Geographic Information Management (GIS) to determine different properties and help to delivery better spatial outputs. Currently in most of the projects different prediction methods are applied. Their outputs are compared with ground truth data and the best fitting model is chosen to obtain final result. This project seeks to explore an approach in which, instead of selecting particular method, multiple methods are applied and their outputs are combined such that the optimal thematic classification is based on data fusion of predictions. The aim of the project is a development of sophisticated classification modeling approach, to mapping land use, land form or land cover, utilizing multiple thematic classification methods to produce better and more justifiable thematic classification then any single classifier.

The soil mapping case study was used to develop robust classification model to supersede the single classifier approach. Area of Interest (AOI) is South Tipperary. South Tipperary is a county in Republic of Ireland. It is part of the South-East Region and it is located in the province of Munster. The county is mostly a flat area with hills chain on western edge. AOI was chosen for extremely high availability of ground truth data.

The project is divided to two main parts:

- Development of predictive classification models
- Models comparison and Fusion of obtained results

In first part 3 of the most modern and sophisticated predictive classification models were executed (Random Forest, Artificial Neural Networks and Bayesian Belief Networks). Models were developed in different software environment (STATISTICA 10, R-project and Netica). This was given by different needs of particular models and different limitation of particular programs. These factors above will be described in more detail later in the thesis.

In second part the performance of particular models was evaluated. The Data Fusion of all models was executed mainly in ArcGIS Bayesian Classification Tool developed by University of Queensland and University of Melbourne. The tool is using a risk assessment framework for land degradation known as LUIM (Land Use Impact Model). As an input data for LUIM weighted Bayesian Belief networks were used. The weights were determined based on comparison of particular models against ground validation.

The last chapter assesses the combined output of models in order to define spatial distribution of values from each model. The behavior of final data fusion model was investigated.

1.1 Development of predictive classification models

The description of the data is not enough for understanding more sophisticated processes. Therefore models of real system are created. It is possible to understand the model as simplified expression of researched phenomena. Models can be categorised based on different conditions. Bellow is fundamental classification of mathematic models by (Haefner, 2005):

It is model describing explicitly mechanistic processes?

YES Processed oriented or mechanistic models (e.g. hydrological models or models of population dynamic)

NO Descriptive models but without description of physical or chemical processes (e.g. biogeographic models)

It is model describing future state of the system or system`s conditions?

YES Dynamic models (time dependent)

NO Static models (time independent)

It is model describing spatial structure?

YES Spatially heterogeneous models

- Discrete – represented by grid with specific resolution where each cell can contain spatial heterogeneity.
- Continuous – each point in space is defined by coordinates and has specific properties.

NO Spatially homogenous models

Does model contain any random component?

YES Stochastic models – contains elements of the random variability.

NO Deterministic models, often in a form of differential equations.

According to this classification system this thesis is focused to STOCHASTIC SPATIALLY HETEROGENOUS MODELS.

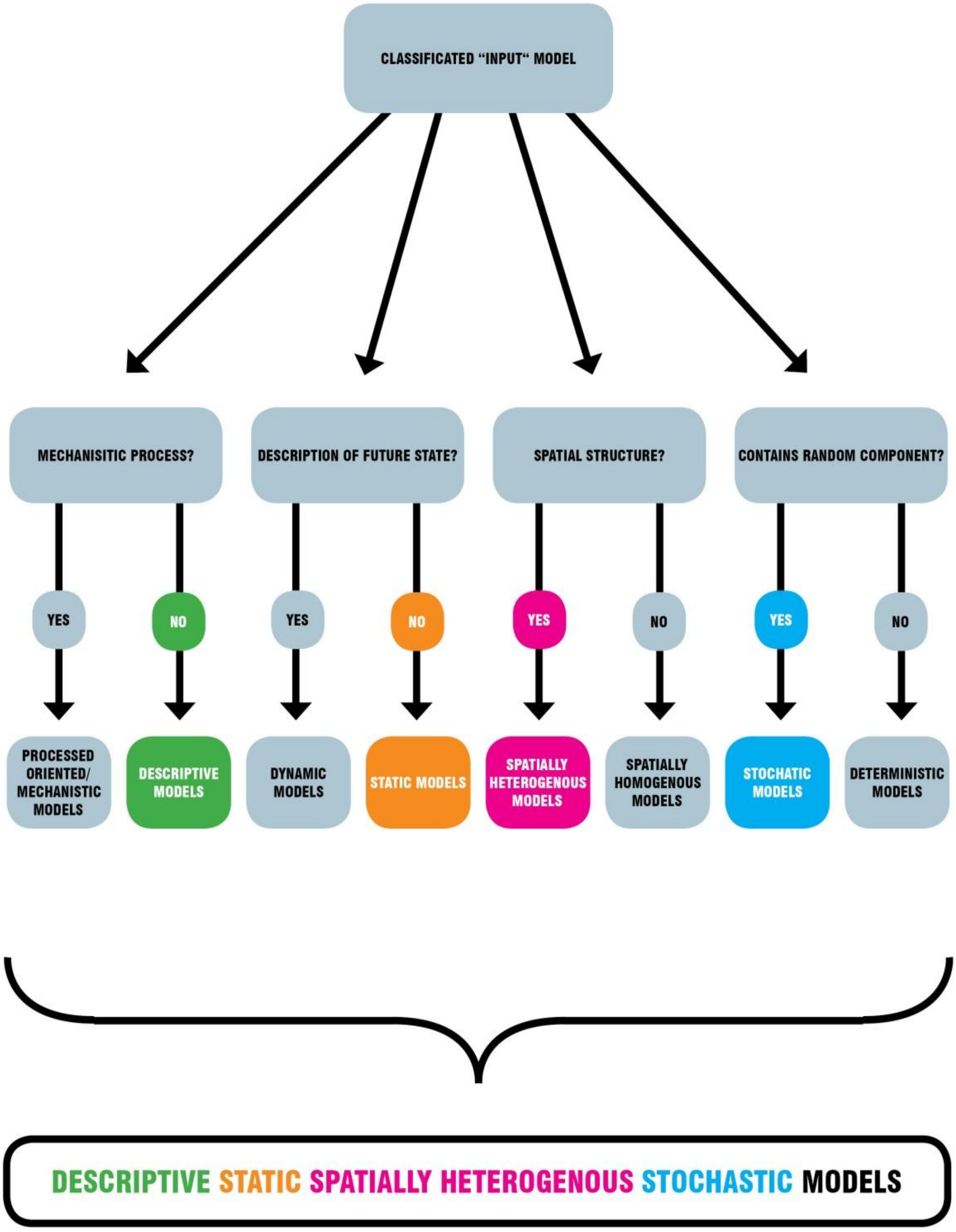

CLASSIFICATED "INPUT" MODEL
MECHANISITIC PROCESS?
DESCRIPTION OF FUTURE STATE?
SPATIAL STRUCTURE?
CONTAINS RANDOM COMPONENT?
YES
NO
YES
NO
YES
NO
YES
NO
PROCESSED ORIENTED/ MECHANISTIC MODELS
DESCRIPTIVE MODELS
DYNAMIC MODELS
STATIC MODELS
SPATIALLY HETEROGENOUS MODELS
SPATIALLY HOMOGENOUS MODELS
STOCHATIC MODELS
DETERMINISTIC MODELS
DESCRIPTIVE STATIC SPATIALLY HETEROGENOUS STOCHASTIC MODELS

Figure 1.1: The categorization of used models

Stochastic models are characterised by 3 of fundamental properties: the precision, the possibility of generalisation and the actual state description. It has been suggested that only 2 of this 3 properties can be increased to detriment of remaining property (Levins, 1966) Due to this criterion above it is possible to divide models into 3 groups (Guisan and Zimmermann, 2000):

- Analytic models, focused on precision and possibility of generalisation.
- Mechanistic models. Models well representing relations between variables. The models are focused on the possibility of generalisation and description of actual state.
- Empiric models, where explaining and explained variables are not necessary in direct casual relation. These models are focused on catching of actual state description.

Empirical models can provide the same or better precision then mechanistic models. At the same time empirical models have lower requirements to expert knowledge about researched covariates. The modelling process is not only computation itself but it is complex process with chain of steps. These steps are most commonly described as bellow (Guisan and Zimmermann, 2000):

- Sampling design
- Ground truth data collection
- Data analysis and model development
- Calibration and validation of model
- Model interpretation, comparison with reality

1.1.1 Types of Covariates

Due to classical mathematical classification of variables 3 types of covariates are recognised:

Qualitative variables

Semiquantitative variables

Quantitative variables

The relation between variables is main classification criterion (Legendre and Legendre, 1998).

With quantitative (categorical) variables is only possible to determine, whether 2 values are identical or not. Examples of quantitative variable are soil classes, geological units or land use. Special case is binary variable. The Binary variable can be processed like quantitative, semiquantitative or qualitative variable. The binary variable is most commonly used to express occurrence/non-occurence of particular specie.

Another type of variable is semiquantitative (ordinal) variable. The order of values can be determined with ordinal covariates. Example of semiquantitative variable is abundance score.

The most used covariates are quantitative covariates. It is possible to execute all kinds of mathematical operations with quantitative variables. The quantitative covariates can be categorised to discrete variables. The examples of quantitative covariates are different measurements, elevation or concentration.

From statistical point of view the covariates are divided to dependent variables and explaining variables (predictors). Depends on the type of dependent variable type of model is chosen. Classification models are chosen for categorical variable and regression models for continuous variable.

1.1.2 Data Fusion

Nowadays, models such as bioclimatic ‘envelope’ model are used in predicting of animal and plant distribution under future climate scenarios (Thuiller et al., 2005). However, the variance is relatively high in comparison with other models. Therefore, a solution is to utilize several models, ensembles (Araujo and New, 2007), and use appropriate techniques to investigate resulting range of projections. An ensemble was introduced to statistical science in 1878 by J. Willard Gibbs. A prediction ensemble is defined as a multiple simulations (copies) across several initial factors so called as a initial conditions (IC), model classes (MC),model parameters (MP) and boundary conditions (BC). Detailed explanations of separate segments are stated in (Araujo and New, 2007).

The simplest and most widely used method for modelling involves combination of single IC, MC, MP and BC to produce final Projection (P). However, it is possible to utilize more than one of mentioned features. Artificial neural network (Figure 1.2 d) and Random Forests (Figure 1.2 b and d) are two types of modelling techniques.

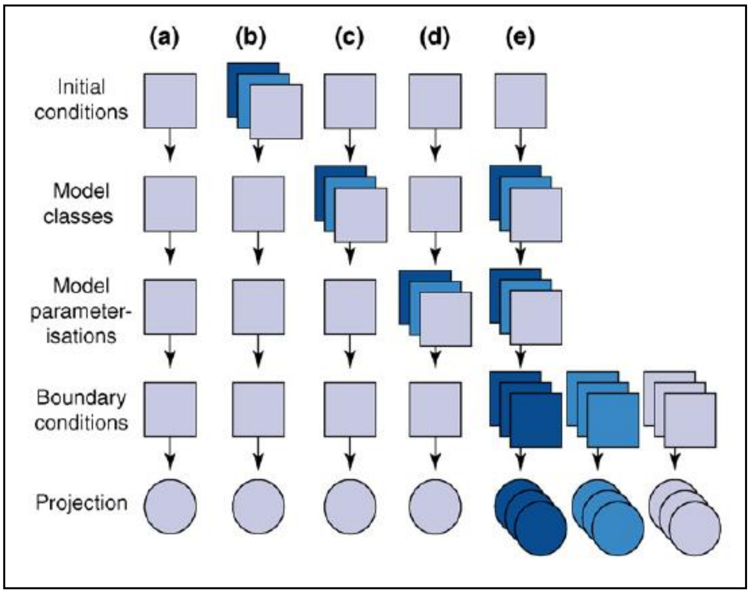


Figure 1.2: Fitting models of distribution and the production ensembles of forecasts. The squares represent different steps of model; circles are predictions of the model.

Combination of forecasts provides lower mean error in comparison with any single forecasts. This attitude was used in variety of fields, such as economics, management, systematic, biomedicine or meteorology (Benestad, 2004; Bernsel et al., 2009; Cramer et al., 2001). Usually, is the ‘best’ model chosen as a reference and this model is judged to be one in which outputs corresponded with observed data (Sanders, 1963; Winkler, 1989). Nevertheless, MP does not always describe real similarity with observed data. Instead of judging the ‘best’ model from an ensemble, exploring the range of projection is more meaningful. Further, ‘bounding box’ is defined or probability distribution functions (PDFs) are generated for medium to large ensemble sizes. Ensemble members are just subset of all possible IC-MC-MP-BC and represent only limited of reality. It is suggested to not estimate any ensemble average or confidence limits for the average. Measure is calculated for the ensemble of forecasts and it is possible when combining forecasts for consensus produce weighted and unweighted averages. For example, artificial neural networks are run several times and then the mean prediction is used. Furthermore, stacking is one of the analogous procedures for estimating weights. Unweighted methods can be used only in case when model predictions are equally robust (Araujo and New, 2007). Probabilistic forecasting can be considered as a most robust of ensemble forecasting. Nevertheless, we need to accept that models are different from reality and that there are some models which represent key aspects of the real system.

1.1.2.1 Ensembles in practices

The idea of combination of forecasts is necessary where a single model does not suffice and it is not close to the truth in all circumstances. However, it should not have been seen as an alternative to traditional approach to build better models with improved data. Better individual forecasts will get better combined forecast. It is necessary in practical decision-making to consider trade-off between costs of models and accuracy of potential modelling. However, it is always reduced the probability that forecast is far from truth with probabilistic and consensus approaches (Araujo and New, 2007; Winkler, 1989).

1.1.2.2 Conclusion

Ensemble forecasting has clear advantages over single-model forecasts. Different approaches were proposed with various applications in practices (Araujo and New, 2007) and examples of problematic where Ensemble forecasting is used are meteorology, prediction of species distribution, business, and hydrology.

1.1.3 Overview of used stochastic models

The knowledge of alternative methods and good orientation of data are needed to choose appropriate predictive method. Stochastic methods can be classified as parametric or non-parametric methods. The data often does not meet all of the classical preconditions such as distribution of dependent variable, independency of parameters or linearity of modeled relation. If assumptions above are met, then usually parametric methods are more successful. More robust non-parametric models are chosen usually when assumptions are not met (Breiman, 2001).

The predictive modeling can be also categorized based on nature of studied phenomena. The common case of predictive modeling is the classification. In classification the unknown object is classified into one of given categories. Examples of these methods are Classification Trees and Forests, Logistic Regression, Discriminate Analysis or Neural Networks.

The second group of models is dealing with continues variable. These models most commonly are trying to solve the regression problem. Due to expected relations of predictors and dependent variable is possible to use Multiple Regression, General Linear Model (GLM), General Additive Model (GAM) or Redundant Analysis (RDA).

During the course of this project different methodologies of stochastic modeling were investigated (Random Forest, Artificial Neural Network and Bayesian Belief Network). In following paragraphs these methods will be described in more detail.

1.1.3.1 Random Forest (RF)

The decision forest is superstructure of decision trees. Forest can be used for classification or regression. Some problems of decision trees (e.g. instability) are removed when forest approach is applied (Friedman et al., 2001). The structure of Random Forest is on Figure 1.3.

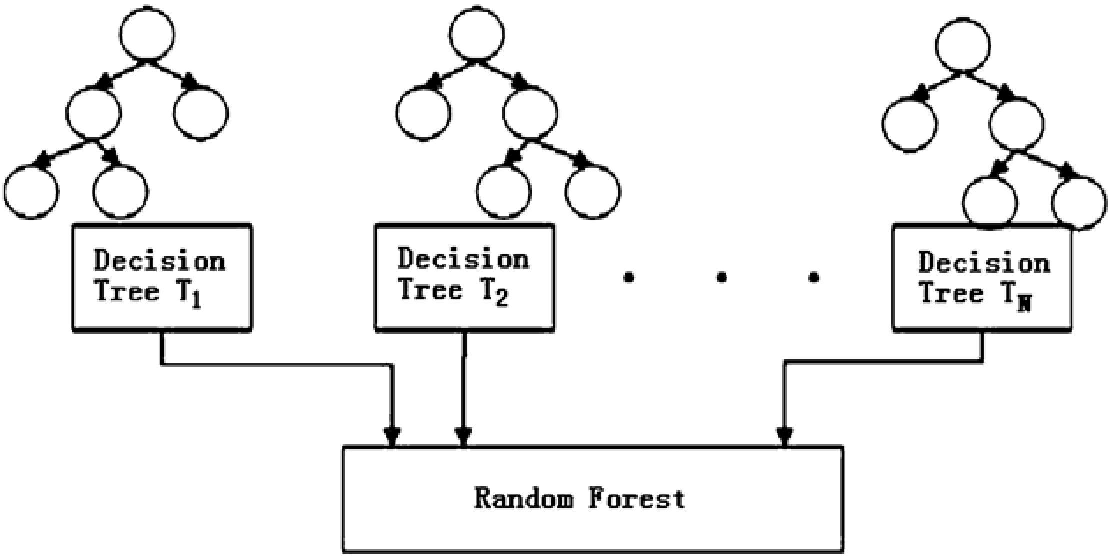


Figure 1.3: Diagram of general structure of RF

In Random Forest, the main advantage of decision trees-readability is lost. In principle, trees are based on a bootstrap of the entire dataset. The splits at the nodes are made from the best of randomly selected subsets. These are not made by the best predictor from the entire ensemble of input variables (Liaw and Wiener, 2002).

1.1.3.2 Artificial Neural Network (ANN)

The essence of ANN was well established by (Bishop, 1995). Maier and Dandy (Maier and Dandy, 2001) described implementation of ANN’s to environmental modeling. In principle Neural Networks are static non-linear models. A lot of types of ANN are recognized such as Radial Basic Function (RBF) or Multilayer perceptron (MLP) (Ripley, 1996).

In this project MLP was applied. This structure is most commonly used in soil science (Agyare et al., 2007). Layers are composed from the nodes (neurons). In this architecture 3 types of layers are recognized: 1) an input layer (containing the variables used for prediction, 2) ‘’hidden’ layer’’ situated in-between input and output layer and 3) an output layer.

The hidden layer contains weights and transforms the data to extract meaningful relationships. The number of hidden layers is given by:

a) The type of neural network.
b) The complexity of researched problem.

Inputs are certain observations of predictors. The values entering the other hidden layers are called ‘’inputs’’ as well (Friedman et al., 2001).

Following paragraphs describes function of neuron in more detail.

Inputs are multiplied by weights. The results of these operations are summarized. The sum is compared with threshold value. If weighted sum of inputs is lower than threshold value (constant) the transformation (activation) function is applied. Weights dedicate the importance of certain input for final result (e.g. classification). Several types of activation functions are recognized (Linear function, Hyperbolic-tangential function or Gaussian function) (Ripley, 1996).

However Neural Networks have ability to rank importance of variables, the different method than in RF is used. The data for each variable is replaced by mean value from the training data and the effect on error function is measured. The variables are then ranked by the amount of their omission increases the overall error function (Lou and Nakai, 2001).

In application of ANN for creation of predictive maps, network does not provide predictions in areas where data differ from those of the training dataset. This leads to blank areas without predictions. However the accuracy of the final map is not compromised. An Example of use of ANN is bioclimatic model SPECIES (Pearson et al., 2002) for modeling of plants habitats based on future climatic scenarios in UK.

The structure of ANN is shown on figure (Figure 1.4).

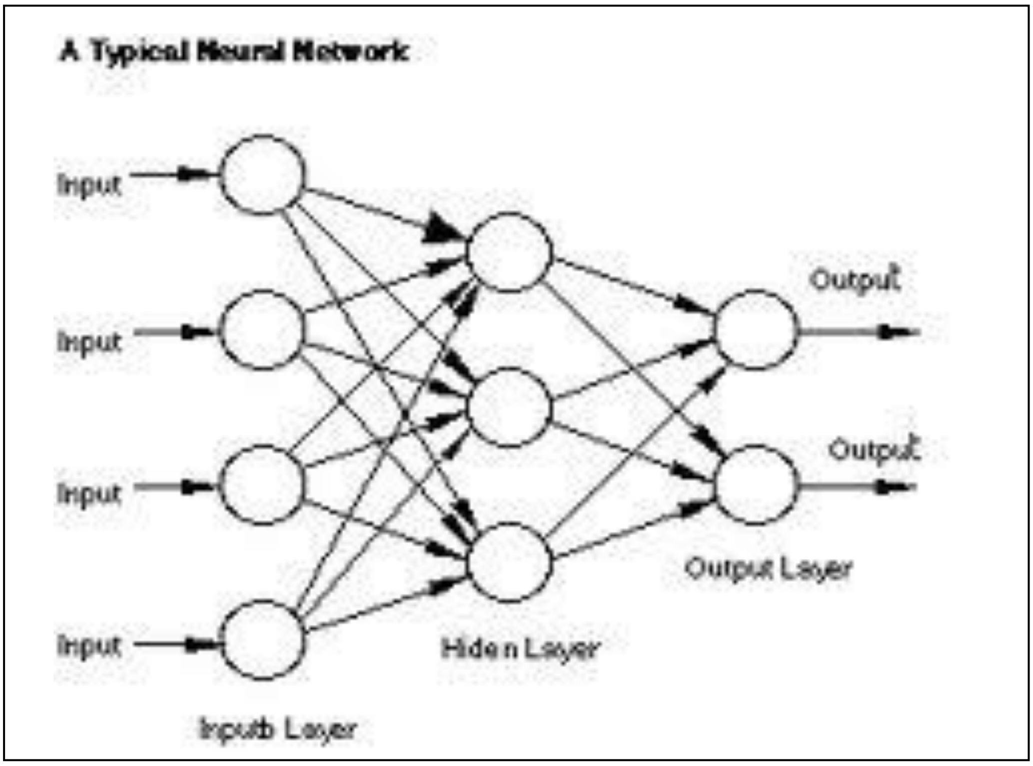


Figure 1.4: Simplified structure of Artificial Neural Network

1.1.3.3 Bayesian Belief Network (BBN)

To combine ensemble of predictors to single model, the non-linear modeling approach is needed. This means techniques such as Random Forest (Liaw and Wiener, 2002); (Wiesmeier et al., 2011) or Artificial Neural Networks (Zhao et. al, 2010). The weak point of these approaches is that they are"black-box" techniques. It is not easy to look inside the black box to describe relations between response and predictor variables (Suuster et al., 2012). Completely different approach is Bayesian Belief Network. BBN is graphical model applying the probabilities. These probabilities are based on expert knowledge or measurements from real world. They represent cause and effect relationships based on a conceptual diagram, linking a range of covariates (Hough et al., 2010). BBN is technique where the knowledge is well structured. The advantage is explicitly accounting of uncertainty and variability (Dlamini, 2011). However to include empirical data is easy, it is not necessary and it is possible to construct BBN structure only with information from expert knowledge (Kuhnert and Haynes, 2009).

In case of application of BBN to a digital soil mapping, the expert knowledge is used instead of empirical data. It is cheaper and it can be easily extended when additional data are available (Jensen, 1996).

BBNs can fuse qualitative and quantitative information (Hough et. al, 2012). The qualitative section of the model is composed of nodes and linking arrows. The nodes express variable of interest. The arrows express cause and effect relationships. The causality is represented by direction of arrows (Hough et al., 2010).

When nodes are linked into others, they are known as”parent nodes”, these which are being linked into are called”child nodes” (Jensen, 1996).

The Conditional Probability Table (CPT) represents the quantitative part of model and contains series of probability distributions and determines impact of change in parent node to child node and conversely (Das, 2004). Before CPTs are populated is necessary to discretised continues covariates into categorical form. This is needed for learning algorithm used by Netica^{TM}. Discretisation will decrease the complexity of the CPTs within the network (Kuhnert and Kayes, 2009).

The structure of BBN is shown on Fig. (Figure 1.5)

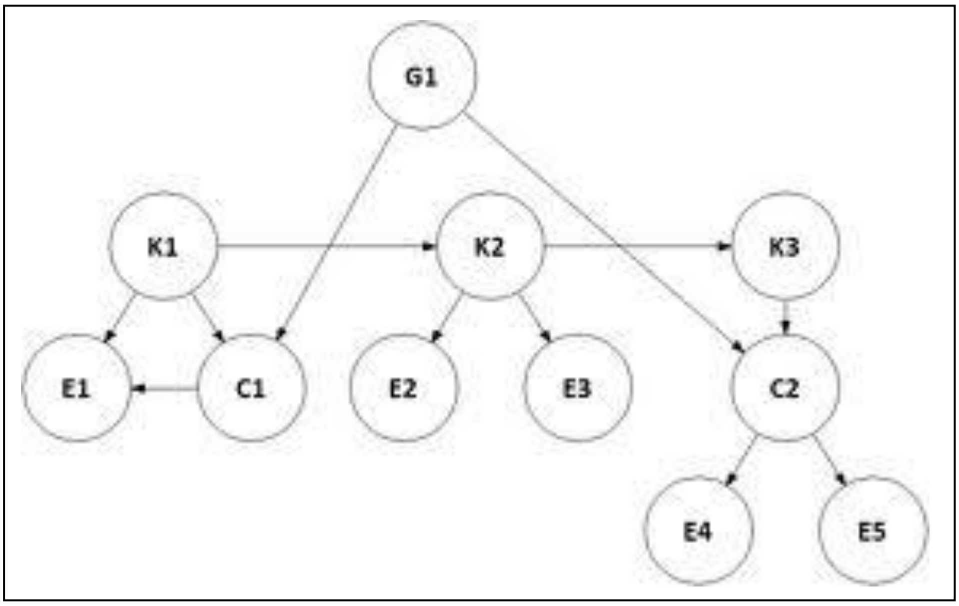


Figure 1.5: Simplified structure of Bayesian Belief Network

1.1.4 Validation techniques

In case that large number of observations is available, ideal is dividing the sample to 3 subsamples: the training data set, the validation data set and the testing data set. The training data set is used to create the model. The validation subsample is utilized to select the best model. The testing subsample is used to estimate error of model's performance. There is no general rule how to spread the data into these subsamples. The most common is following division of the data: 50% the training subsample, 25% the validation subsample and 25% the testing subsample (Friedman et al., 2001).

In reality usually is not ideal situation. The number of data does not allow this division of the entire dataset. The validation techniques are used for these cases. Validation techniques is possible to categorize to analytical (e.g. information criterions) and techniques which estimates model's objectivity by repetition of observations.

Estimation of overall model's error by validation techniques is mostly used to select appropriate model. This estimation is utilized also to determine stability of model, to selection of input variables and to determine model's complexity. The complexity of the model means final number of terminal nodes (during tree pruning) or number of hidden layers of Artificial Neural Network.

The general rule is to select the simplest model in terms of complexity and number of explaining variables (principle of parsimony). In the same time this model has to explain the largest amount of information. The result is compromise between these 2 requirements above.

The Validation techniques described below are based on repetition of use of observations.

1.1.4.1 k-fold-crossvalidation

In cross validation dataset is divided to k independent subsamples (usually k = 10).The one of the subsamples is utilized for testing (observations are not used for development of the model). All of other subsamples is (k – 1) are used to model's development. In total is developed k models tested on k ensembles.

From the results of testing ensembles is possible to define model’s stability (e.g. standard deviation of test data).The advantage of cross validation is that for testing procedure, the independent ensemble is used (Bishop, 1995).

If the number of cross validation subsets is even to number of observations, the method is called ’’leave-one-out’’ (LOO) cross validation. The LOO cross validation was designed to really small data sets. It is appropriate for generalization of error estimation of continuous functions (e.g. Root Mean Square Error). It is not optimal to discrete error estimation (Friedman et al., 2001).

The cross validation is most commonly used to define optimal number of terminal nodes in development of decision trees (De'ath and Fabricius, 2000).

The cross validation is appropriate for smaller datasets. For larger datasets, the bootstrap method is recommended (Friedman et al., 2001). For select subsample of variables in linear regression, 10-fold and 5-fold cross validation providing better results than LOO (Breiman and Spector, 1992).

1.1.4.2 Simple splitting

The other validation technique is simple splitting of the dataset into testing and training ensembles. Only 1 subsample (testing ensemble) is used to general error estimation (Van Houwelingen and Le Cessie, 1990). The cross validation provides better results than simple splitting for smaller data sets (Goutte, 1997).

The others ”re-sampling” validation techniques are not commonly used in distribution prediction modeling. These are mostly ” bootstrap” and ”jacknife” methods. Both of these methods are commonly used in machine learning (Ripley, 1996; Breiman, 1996).

1.1.4.3 Bootstrap

The bootstrap method is based on repetitive random selections from original entire data set (Efron, 1979). Ensemble is randomly split to the testing and the training part (given as percentage by user). Testing ensembles are always independent. Samples can be repeated in each subset. The advantage is possibility of application of this technique to small data sets. The bootstrap often provides better results than cross validation (Efron, 1983). On other hand, for certain methods (e.g. decision trees) bootstrap provides worst results (Friedman et al., 2001).

Many of others bootstrap methods were developed. The Random Forest and the bagging are used to forest development and randomization of estimate of variable's importance (Breiman, 2001). In Artificial Neural Networks is bootstrap used to define number of confidence intervals of their results (Tibshirani and Efron, 1993). In biological research is bootstrap used mostly in ecotoxicology (Halfon, 1985).

2 Methodology

2.1 Area of Interest and sampling strategy

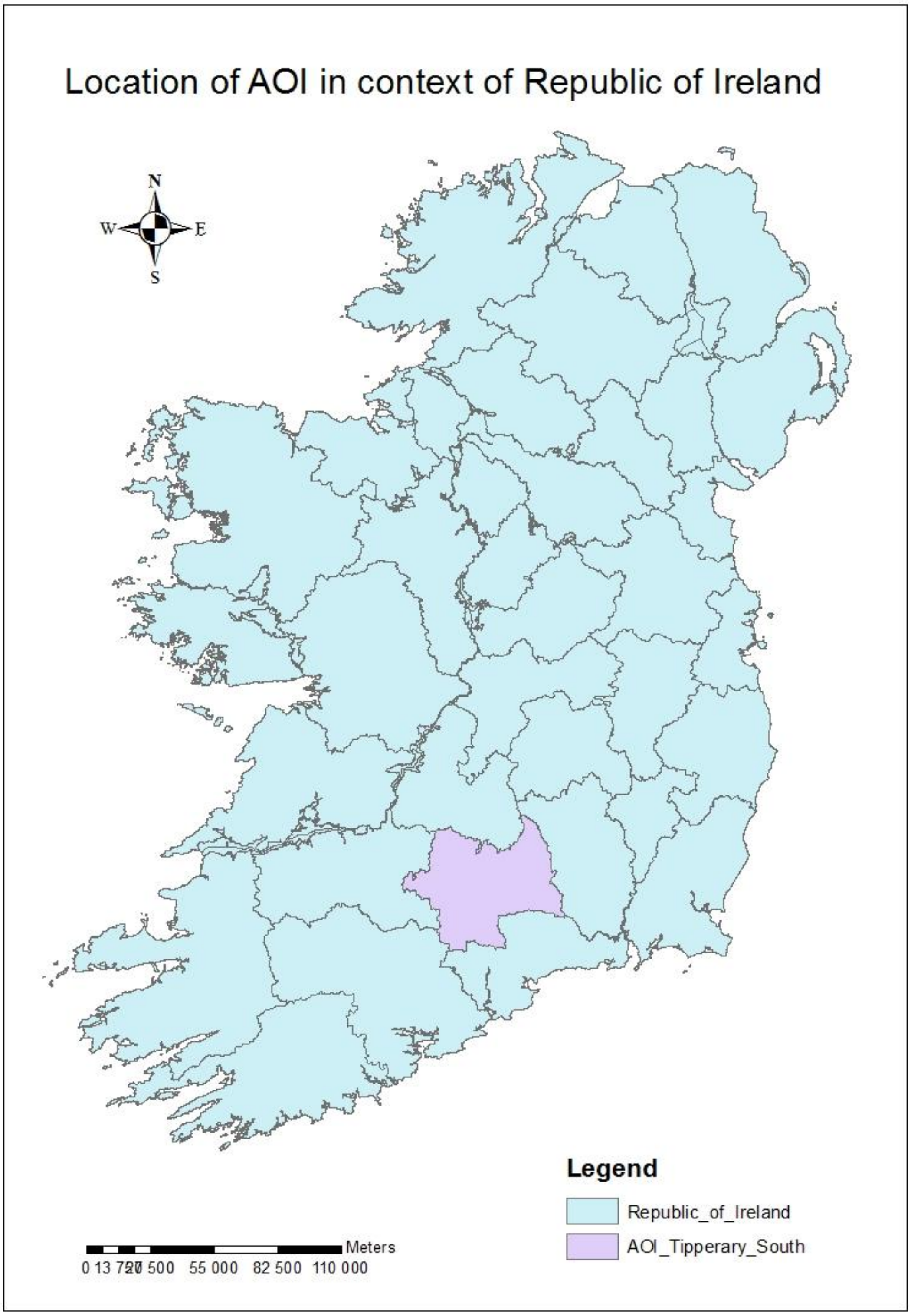


Figure 2.1: Location of AOI in context of Republic of Ireland

The wider Area of Interest (AOI) contains 4 different counties located in south of the Republic of Ireland. The spatial distribution of different soil classes was predicted for South Tipperary. Area of surrounding counties (North Tipperary, Limerick and Waterford) was used to training of models.

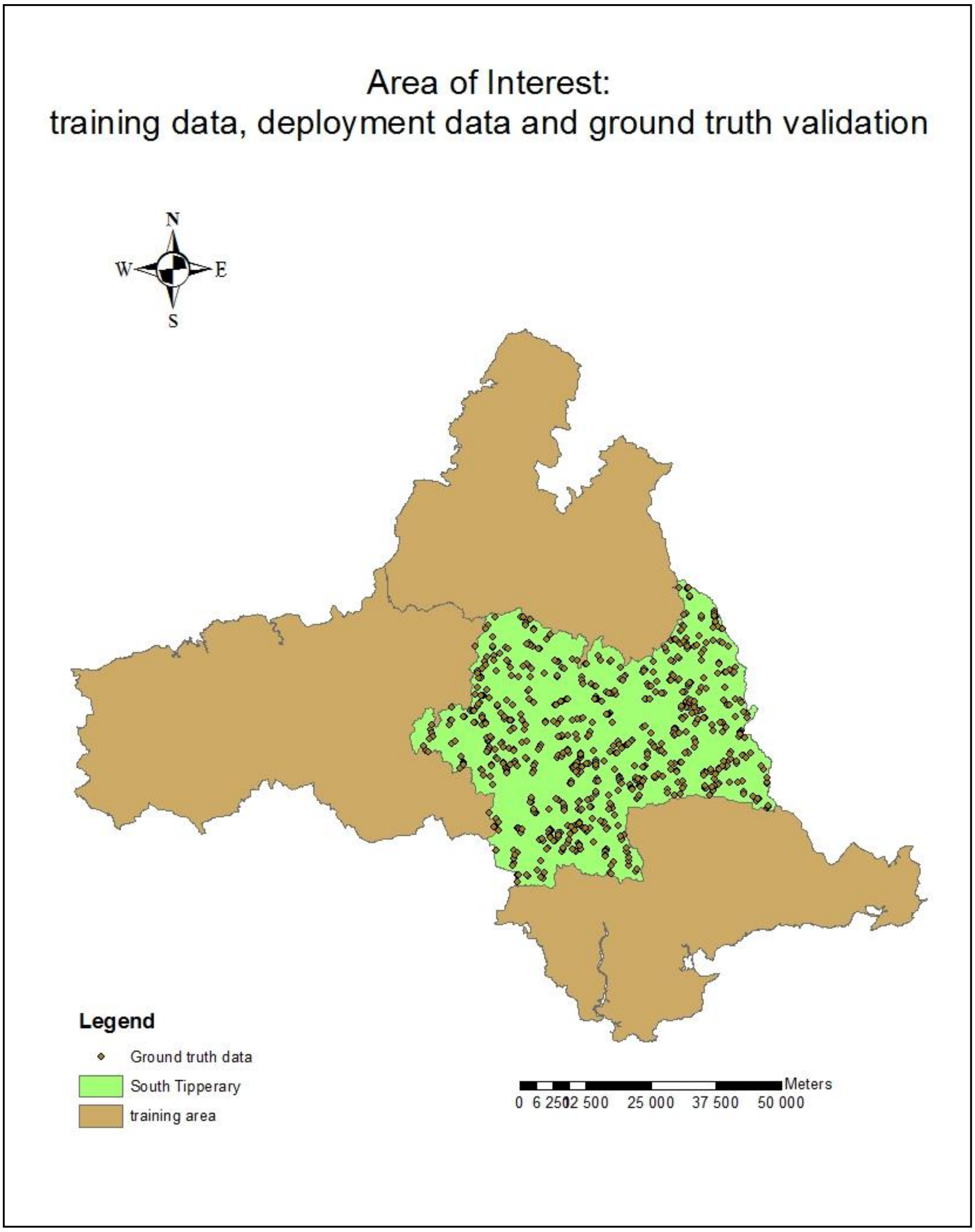


Figure 2.2: training area, deployment area and locations of ground truth validation data

The training area was sampled by Hawth's Analysis tool for ArcGIS. Sampling strategy in training area was designed like Unaligned Stratified Random Sample. The sampling density was 10 points per squared kilometre. The total number of points in training area was nearly 60 000 points. The deployment area was sampled by regular grid of points with 100m cell size. This grid was chosen because is appropriate for creation of the map afterwards (the raster with pixel 100x100m) and because this way nearly 250 000 points were created (250 000 rows is a limit for attribute table in ArcGIS).

2.2 Used datasets - covariates

The selection of variables was based on the framework scorpan-SSPFe (soil spatial prediction function with spatially auto correlated errors). According to McBratney, A.B. (2003), the scorpan was derived from (Jenny's, 1941) famous equation. The scorpan framework contains following factors:

s – soil (observed or measured attributes of the soil)|

c – climate (climatic properties of environment)

o – organism (including land cover and natural vegetation)

r – topography (terrain attributes and classes)

p – parent material (including Lithology)

a – age (time factor)

n - space

The values of each covariate were extracted for each point from vector and raster datasets in software environment of ArcGIS 10.0.

Environmental variables used in this project are in Table 2.1 .

Table 2.1: Environmental variables used in the project.

GSM	Generalised Soil Map	Categorical
SBS	Parent material	Categorical
COR	Land use CORINE	Categorical
aso2	SOTER terrain unit	Categorical
clo_2_i	Landscape position	Categorical
aje2	Actual/Potential drainage density ratio	Continuous
ajg2	Dissection Index	Continuous
rann	Solar radiation	Continuous
eann	Evaporation	Continuous
epm2	Relative relief within dataset	Continuous
ezz1	Total relief within catchment	Continuous
gna5	Downslope Flowpath Lenght	Continuous
gnh5	Surface Curvature Index	Continuous

The predicted variable was soil class. To evaluated model's performance was used dataset containing 1155 ground validation points with soil classes.

All of the environmental variables used in this project are soil associations based on the Irish National Soil Taxonomy. In the following paragraphs the individual covariates are described in more detail.

GSM – Generalised Soil Map

The GSM stands for Generalised Soil Map. GSM is the dataset containing soil classes. The soil class was predicted covariate in models. Soil class is categoric value describing the systematic categorization of soils based on distinguishing characteristics and soil properties. Values of the GSM covariate were extracted within sampling points from vector dataset.

SBS – Parent material

The parent material is the initial state of the solid matter making up a soil. It can consist of consolidated rocks and can also contain unconsolidated deposits such as river alluvioum, lake or marine sediments, glacial tills, loess (silt-sized, wind deposit particles), volcanic ash and organic matter. Parent materials influence soil formation through their mineralogical composition, their texture and their stratification (Encyclopaedia Britannica, 2011). The values of the SBS covariate were extracted within sampling points from vector dataset.

COR – Land use

This is categorical covariate extracted from database of CORINE (Coordination of Information in the Environment) program. The CORINE program was initiated by European Union in 1985. The CORINE dataset is an inventory of land cover in 44 classes, and presented as cartographic product of scale of 1:100 000 (Commission of the European Comities, 1995).

aso2 – SOTER terrain unit

The SOTER (World Soil and TERrain Digital Database) project was initiated by the international Society of Soil Science (ISSS) in 1986 (ISSS, 1986). The SOTER is intended to have global coverage at 1:1 000 000 scale (Bjates, 1990; ISRIC, 1993), which goal was later degraded to 1:5 000 000 scale. The Food and Agricultural Organisation (FAO) of United Nations and the International Soil Reference and Informaton Centre (ISRIC) joined this project and supported the idea of having a global scale soil and terrain database useful for series of applications. A small international committee was appointed to develop a “universal map legend system” and to define a minimum necessary set of soil and terrain attributes suitable for compilation of a small-scale soil resource map.

The SOTER unit delineation is based on 2 primary soil formation phenomena: terrain and lithology. Each SOTER unit represents a unique combination of terrain and soil characteristics. The two major differentiating criteria are applied in step-by-step manner, leading to a more detailed identification of the land area under consideration. Physiography is the first differentiating criterion to be used to characterize a SOTER unit. The term physiogrphy is used in this context as description of the landforms on the Earth´s surface. It can be best described as identifying and quantifying the major landforms, on the basis of dominant gradient of their relief intensity. The use of these variables, in combination with hypsometric classification and a factor characterizing the degree of dissection, can make a wide subdivision of an area and delineate it on the map. Further classification of the SOTER unit according to the parent material needs to be done to complete the delineation procedure (Dobos et al., 2005).

clo_2_i – Landscape position

The landscape position is an categorical covariate describing the soil and soil formind conditions according to the shape of surface, its slope and position on the landscape. The landscape position plays a role in the amount and quality of the soil profile. The soils that developed on higher elevations and sloping areas are generally excessively derived or well drained. Soils that occur at lower elevations such as in swales, adjacent to drainage-ways and water bodies, and within depressions generally receive surface runoff from higher elevations and often have a seasonal high water table at a shallow depth. (Guzman and Al-Kaisi, 2011)

aje2 – Actual/Potential drainage density ratio

The drainage density is defined as total length of channels per unit of area. It is a fundamental property of natural terrain that reflects local climate, relief, geology and other factors (Tucker et al., 2001). The aje2 is continuous variable and represents the ratio between actual and potential drainage density.

ajg2 – Dissection index

The dissection index is the alymetric (relief based) characteristic. It is describing the ratio between relative and absolute relief per unit area or grid. The dissection index as a function of elevation allows us to quantitatively summarize the morphology of landscape. (Jha and Kappat, 2009)

rann – Solar radiation

In general the solar radiation is the total frequency spectrum of the electromagnetic radiation produced by the Sun. In terms of this project, according to the fact that Earth´s atmosphere deflects or filters the majority of the Sun´s radiation, the solar radiation is a continuous covariate. This continuous covariate expresses the amount of the energy radiated by the Sun per area unit of the Earth´s surface. The values within sampling points were extracted from the vector dataset.

eann – Evaporation

The eann is a continuous covariate. The eann expresses the amount/intensity of the Evaporation. The evaporation is the process by which water is converted from its liquid form to its vapour form and thus transferred from land and water masses to the atmosphere.

epm 2 – Relative relief within dataset

The relative relief is the vertical difference in elevation between the highest and the lowest points of a land surface within specified horizontal distance or in limited area. It is also known as local relief. The emp2 is the continuous variable.

ezz1 – Total relief within catchment

The total relief within catchment is relief-based measure. It expresses the maximum elevation above catchment outlet. In a large catchment, most of the relief will arise from the elevation change along the main stream between the outlet and head water. Total relief within catchment is a continuous covariate.

gna 5 – Downslope flowpath length

A flowpath length is defined as the distance from divide to the first adjacent downslope channel. As such only the distance from drainage divides cells to the downstream channel are considered in the length calculation. It should be noted that the drainage divide is not only located at the upstream boundary of the subcatchment, but also within the subcatchment as defined by local ridges in the topography. Thus, the flowpath length is generally shorter than the average distance to the drainage divide that forms the upstream boundary of the subcatchment. The gna5 is continuous covariate extracted within sampling points from the raster dataset.

gnh5 – Surface Curvature Index

The Surface Curvature Index is a measure of total curvature and magnitude of slope gradient. The gnh5 indicates predominantly convex slope (positive values) or predominantly concave slope (negative values). Surface Curvature Index is a continuous variable extracted from the raster dataset.

2.2.1 Derivation of covariates

The focus and final output of this study are soil properties derived by spatial interference system. There are 3 main areas of concern affecting the accuracy of those predictions: soil information, environmental covariates and interference system.

The soil information is assumed correct and it creates the foundations for the parameterization of models. The mapping was based on field observation by the soil surveyor but is subject to the interpretation of individual (Cavazzi, 2013). The 2 soil information covariates used in this project were the Parent material and information from Generalised Soil Map.

The environmental covariates used in this project are possible to categorize to 2 main groups*:*

“The Ecological variables” - derived from combination of census and remote sensing data (i.e. Land use, Solar radiation and Evaporation).

“The Geographical variables” – derived from DEM. These terrain attributes used in the project are the Landscape position, the Actual/Potential drainage density ratio, the Dissection Index, the Relative relief, the Total relief within catchment, the Downslope flowpath length and the Surface Curvature Index. These covariates are indicative of soil-landscape relationships controlling the spatial distribution of physical, chemical and biological soil properties and the water balances (Florinsky. 2011; Cavazzi, 2013).

The special covariate according to this categorization is the SOTER unit. This variable is a combination between the soil information and DEM derived covariate.

2.3 Software limitation for creation of the dataset

Some data was excluded from training and deployment datasets. This need was given by nature of used stochastic models and limitation of software tools.

Due to the nature of the predictive techniques, it was possible to maintain only the combinations of predictors with dependent variable, which occurred in both training and deployment datasets. In other words, because the models were trained in different area than deployed, the training dataset does not contain all of the soil classes occurring in deployment area. However this soil classes were in AOI very marginal.

The software limitations were this 2:

The R-project is not able to handle categorical variable with more than 32 categories.

STATISTICA10 has a limit for dependent variable 50 categories.

The soil class, the parent material and SOTER terrain unit were variables which were not meet the requirements from limitation above. From soil classes and SBSs vectors were created frequency tables in Microsoft Excel. The least occurring categories were identified and excluded from entire dataset.

2.3.1 Modification of SOTER Terrain Units

The SOTER Terrain Units delivered by An SRTM-based procedure is special kind of categorical variable classifying localities in terms of landscape position. Units take the form of 4 digit code. Each number within a code stands for different characteristic. First number expresses the slope in percentage. The second number represents Relief intensity in metres. The third digit is Potential drainage density index and last digit of the code express Elevation.

Each unique combination of numbers is 1 category. In area of interest 94 different SOTER categories occurred. Two main approaches were soliciting to deal with this issue. The first option was to exclude whole variable. It would be relevant because other covariate representing landscape position (clo_2_i from CONMAP) was included in dataset. The second option was transformation of the vector in some meaningful way with maintaining most of the information.

In first instance the digit of Potential Drainage Density was excluded because was identical for all points all over the area of interest.

Afterwards the categories were merged together until total number of categorize was 32. The main criterion for merging was elevation with priority to merge categories with low value of elevation. This criterion was given by AOI because in context of Southern Ireland, the landscape is relatively flat and variation in soil classes is not based on elevation.

2.4 Used stochastic models

The principles of the stochastic predictive methods used in this project were described in the Introduction section. Some information will be added in the following section to demonstrate what was carried out in this study. The following information will aid readers to guide them in their future work.

2.4.1 Random Forest

Random Forest was executed in R-statistical computing language. The Package"randomForest" was used. The full title is"Breiman and Cutler's random forests for classification and regression" and is downloadable at URL http://stat-www.berkely.edu/users/breiman/RandomForests.

The number of trees in the forest (n_{tree}) and the number of variables tested at each node (m_{try}) are 2 of user-defined parameters. The model's performance is evaluated by Mean Square Error (MSE) of the " out of bag" portion of the data at each tree, averaging over the entire forest as shown equation 1-1.

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

Where z_i^{OOB} is mean out of bag predictions for the *i*th observation. The RF differs from the other 2 models used in this project by ability to evaluate importance of each variable. This is done by measuring how much the ''out of bag'' estimate error raises when data for a certain covariate is temporary removed from the model. This is done all over the decision forest.

The simplest version of used programming code (without commands to plot charts etc.), is possible to found in Figure A-2

2.4.2 Artificial Neural Network

ANN was preceded in R-project software environment. Two of the packages to construct ANNs are available:

1) the neural net package
2) the nnet package

The nnet package was chosen for purpose of this project. Package is possible to download at: URL http://www.stats.ox.ac.uk/pub/MASS4/.

The recommended number of hidden nodes is half of input variables plus the number of output variables. Generally, the higher number of nodes, the better is performance of the model. However, this may cause over fitting the data. To avoid this, the Artificial Neural Networks uses a back-propagation algorithm

(Rumelhart et al., 1986). The back-propagation is used to evaluate the performance of the model on both training and testing data. The error function for regression is the Sum of Squares error shown at equation 1-2.

$$\mathbf{E_{SOS}} = \sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{y_i} - \mathbf{t_i})^{\mathbf{2}} \qquad \textbf{(2-2)}$$

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. If the differences in the error function are insignificant, the structure with the fewest neurons is chosen. The best performing models are selected using of R^2 and RMSE.

The simplest version of used programming code (without commands to plot charts etc.), is possible to found in table A-1.

2.4.3 Bayesian Belief Network

According to Chatterjee and Bandopadhyay, the basic procedure in NN learning is to start with untrained network, present a training pattern to the input layer, pass the signals through the net, and determine the output, which is function of weights. The outputs obtained from the model are compared with the target output values of the same training pattern and any observed difference corresponds to an error. The error function is same scalar function of the weights and is minimised so that network outputs match the target output. Thus, the weights are adjusted to reduce this measure of error. The training error on a pattern to be summed over the output units is the squared difference between the target output and network output obtained from the model

$$\boldsymbol{E_d(W)} \equiv \sum_{\boldsymbol{k=1}}^{\boldsymbol{N}} \frac{\mathbf{1}}{\mathbf{2}} \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} (\boldsymbol{t_k^j} - \boldsymbol{z_k^j})^2 = \sum_{\boldsymbol{j=1}}^{\boldsymbol{N}} \frac{\mathbf{1}}{\mathbf{2}} |\boldsymbol{t^j} - \boldsymbol{z^j}|^2 \qquad \textbf{(2-3)}$$

where K represents the number of output units and, without loss of generality, only one output unit is present (K=1), t and z represent the target and the network output vector; D = (t^1, t^2,..., t^N) represents the target output data from the training set, N represents number of training patterns, W≡ (W_1,W_2,..., W_B)

represents weight vector in the network and B is the total number of weights and biases the network.

Traditional NN models do not provide a confidence interval of estimated values of output variables. Probabilistic interpretation in the maximum likelihood method can be used for NN learning process. Unlike the initialization of single weights during learning in the traditional NN model, the BBN model initializes the distribution of weights (Radford, 1996). Initialized weight distributions, known as prior distributions, are updated by Bayesian rule using training data. Suppose that patterns in training set are independently drawn from a distribution p(x,t), where $x \equiv (x_1, x_2, ..., x_d)$, and target output t is a deterministic non-linear function z(x), plus zero-mean Gaussian noise.

Before Bayesian learning of a MPL neural network, the prior probability distribution of the network weights W needs to be defined. After choosing the Gaussian model of prior probability distribution of network weights and an expression for the likelihood function, the Bayes´ theorem can be used to find the posterior probability distribution of network weights (Chatterjee and Bandopadhyay, 2009).

The following diagram displays the structure of applied Bayesian Belief Network

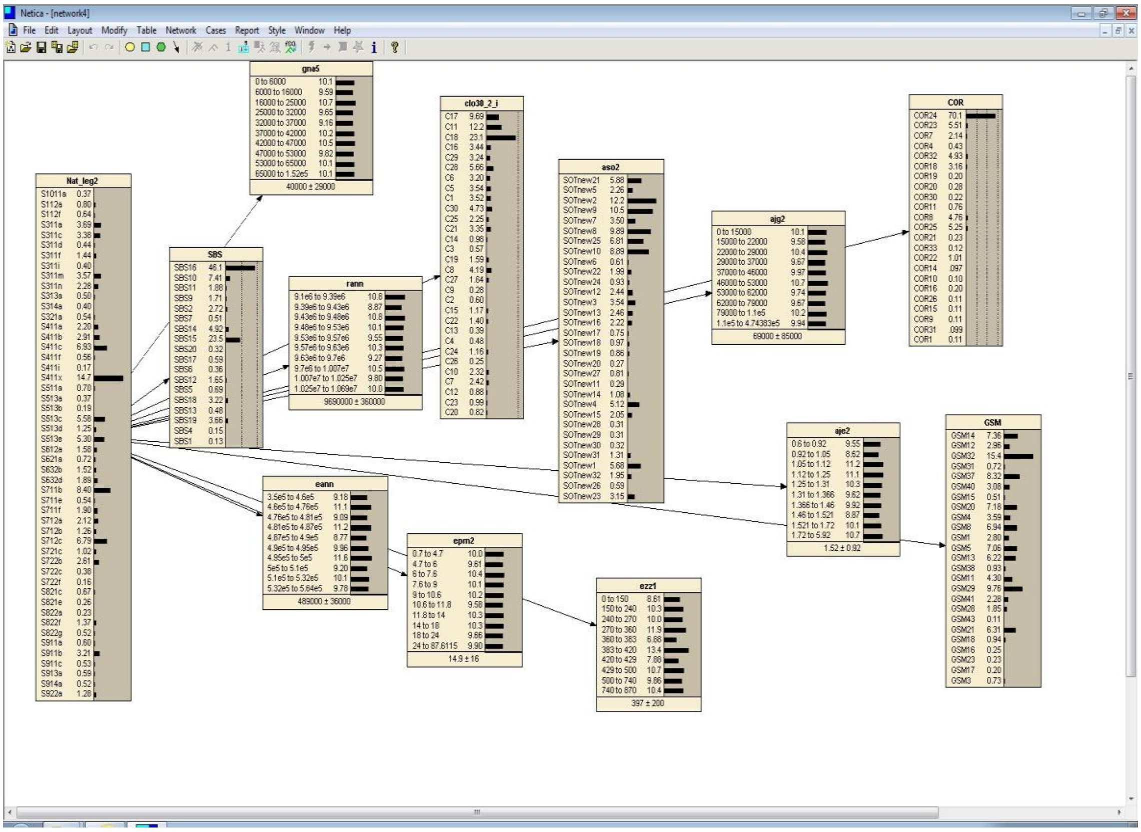


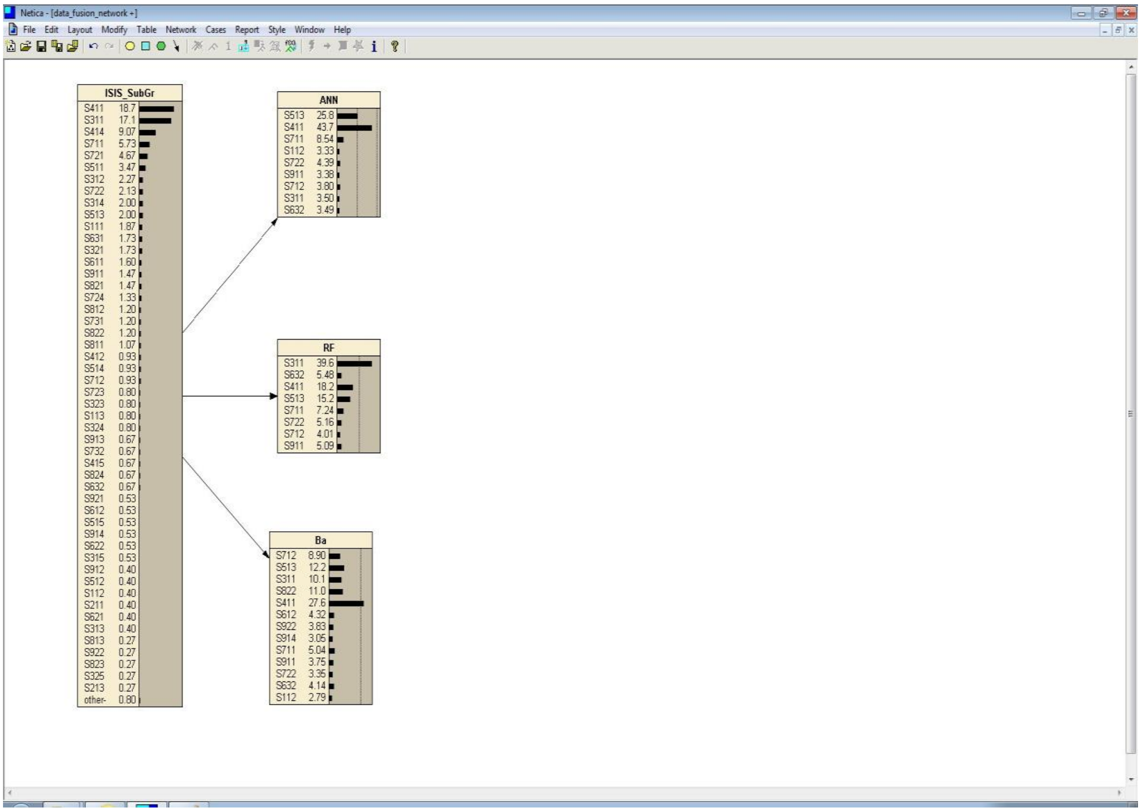
Figure 2.3: Structure of used BBN.

2.5 Data Fusion

The fundamental principles of Data fusion of different models were described in Introduction section. Here will be all of 5 approaches, tested in this study, specifically described.

2.5.1 Data Fusion -The first approach

The first approach is using another simple Bayesian Belief Network to combine the outputs of all models together.



ISIS_SubGr	
S411	18.7
S311	17.1
S414	9.07
S711	5.73
S721	4.67
S511	3.47
S312	2.27
S722	2.13
S314	2.00
S513	2.00
S111	1.87
S631	1.73
S321	1.73
S611	1.60
S911	1.47
S821	1.47
S724	1.33
S812	1.20
S731	1.20
S822	1.20
S811	1.07
S412	0.93
S514	0.93
S712	0.93
S723	0.80
S323	0.80
S113	0.80
S324	0.80
S913	0.67
S732	0.67
S415	0.67
S824	0.67
S632	0.67
S921	0.53
S612	0.53
S515	0.53
S914	0.53
S622	0.53
S315	0.53
S912	0.40
S512	0.40
S112	0.40
S211	0.40
S621	0.40
S313	0.40
S813	0.27
S922	0.27
S823	0.27
S325	0.27
S213	0.27
other-	0.80

ANN	
S513	25.8
S411	43.7
S711	8.54
S112	3.33
S722	4.39
S911	3.38
S712	3.80
S311	3.50
S632	3.49

RF	
S311	39.6
S632	5.48
S411	18.2
S513	15.2
S711	7.24
S722	5.16
S712	4.01
S911	5.09

Ba	
S712	8.90
S513	12.2
S311	10.1
S822	11.0
S411	27.6
S612	4.32
S922	3.83
S914	3.05
S711	5.04
S911	3.75
S722	3.35
S632	4.14
S112	2.79

Figure 2.4: The structure of BBN used in DF – The first approach

This approach assumes that all input models have equal importance and that ground truth validation data are really reliable.

The other 3 approaches are incorporating the weights.

2.5.2 Determination of the Weights

The weights were obtained from 2 main components:

-The convergence (degree of similarity) between different soil classes.

- The number of ground truth points supporting different soil classes.

The frequency table of ground validation points within different soil classes was done. The proportion of these frequencies was created to standardise the data.

The soil classes subgroups were signed to the table in order to most similar soil type were situated next to each other. The similarity scores were extracted by vb.net script counting distance (number of columns) between soil classes. The similarity score value range from 0 to 9. The zero means the most similar soil classes and 9 means the most different soil classes.

Because of the final weight is based on multiplication of convergence and number of points supporting each soil class, the reciprocal values of similarity score were counted. **The full vb.net script is shown at appendix 3.**

2.5.3 Data fusion – The second approach

The table containing ground truth validation data and predictions of Random Forest, Artificial Neural Network and Bayesian Belief Network was constructed. The weights were signed to extra columns. By simple function in Microsoft Excel the best prediction (with highest weight) for each point was extracted. The final map was created based on the extracted values. The excel function is shown below because is not standard operation and future readers may find it useful:

$$\text{If(E2>G2;If(E2>I2;D2;H2);If(G2>I2;F2;H2))} \quad \textbf{(2-4)}$$

2.5.4 Data fusion – The third approach

The simple Bayesian Belief Network was created (identical with the first approach). This time the weights were signed like probabilities to Conditional Probabilities tables in Netica software environment.

2.5.5 Data fusion – The fourth approach

The weights were incorporated in form of covariates. Both of the datasets (training and deployment) contains only 3 covariates. Each of these covariates comprises calculated weights for each model in each sampling point. The Random Forest predictive model in STATISTICA 10 was executed with this data.

2.5.6 Data fusion – The fifth approach

The weights were incorporated in form of covariates. The weights were incorporated in form of covariates. Both of the datasets (training and deployment) contains only 3 covariates. Each of these covariates comprises calculated weights for each model in each sampling point. The Bayesian Belief Network model in Netica software environment was executed with this data.

2.6 Statistical Analysis

The different statistical indicators were gained, depends on particular model. Examples of these statistical indicators are sensitivity to findings, error rates (for confusion matrixes), importance of variables, risk estimate or standard error. The confusion matrixes were used to compare the models against each other.

2.6.1 The Kappa Coefficient

The evaluation of the statistical significance of difference in accuracy between two thematic maps derived from remotely sensed data has often been based on the Kappa coefficient calculated for each map. The Kappa coefficient of agreement for a thematic map is based on the comparison of the predicted and actual class labels for each case in the testing set and may be calculated from:

$$\boldsymbol{\widehat{K} = \frac{P_o - P_c}{1 - P_c}} \tag{2-5}$$

Where P_o is the proportion of cases in agreement (i.e. correctly allocated) and P_c is proportion of agreement that is expected by chance. The derived coefficient provides an estimate of the accuracy of the map which together with that derived from another map is the basis of most map comparisons. Specifically, the map comparison seeks to determine if the difference in the derived estimation can be inferred to indicate a difference in associated population parameters of accuracy. The significance of the difference in accuracy between two maps with Kappa coefficients, $\widehat{K}_1$ and $\widehat{K}_2$, may be evaluated with the normal curve deviate

$$\boldsymbol{z} = \frac{\widehat{\boldsymbol{K}}_{\boldsymbol{1}} - \widehat{\boldsymbol{K}}_{\boldsymbol{2}}}{\sqrt{\widehat{\boldsymbol{\sigma}}^{\boldsymbol{2}}_{\boldsymbol{k_1}} - \widehat{\boldsymbol{\sigma}}^{\boldsymbol{2}}_{\boldsymbol{k_2}}}} \qquad \textbf{(2-6)}$$

where $\hat{\sigma}^2_{k_1}$ and $\hat{\sigma}^2_{k_2}$ represent the estimated variances of the derived coefficients. The significance of the difference between the two Kappa coefficients is then assessed by comparing the value of z calculated from equation above against tabulated values. For the simple situation of determining if there is a difference between two Kappa coefficients the null hypothesis (H_0), of no significant difference, would be rejected at the widely used 5 percent level of significance if |z| > 1.96 (Congalton, 1983). The approach for accuracy comparison based on equation above has been widely used for the comparison of classification accuracy statements in the remote sensing (Foody, 2004).

2.6.2 The Cramer's V

The Cramer's V was main statistic applied to results. Because of different models predicted different number of soil classes, the confusion matrixes were not applicable to evaluate performance of the model against ground truth data. Therefore was necessary find appropriate statistical technique. The Cramer's V was chosen to main evaluation of performance of different models and performance of different Data Fusion approaches.

Cramer's V is named after Swedish mathematician and statistician Harald Cramér. Cramer's V is a way of calculating in tables which have more than 2x2 rows and columns. It is used as post-test to determine strengths of association after chi-square has determined significance. V is calculated by first calculating chi-square, then using equation 1-3.

$$\mathbf{V} = \mathbf{SQRT}(\mathrm{X}^2/(\textbf{n-k-1}))^{\mathbf{2}} \qquad \textbf{(2-7)}$$

Where X^2 is chi-square and k is number of rows or columns in the table.

Chi-square says that there is significant relationship between variables but it does not say just how significant and important this is. Cramer’s V varies between 0 and 1. The lower value indicates little association between variables, the higher value indicates strong association (Cramer, V., 1999).

3 Results and discussion

3.1 Result of different classifiers

3.1.1 Random Forest

The number of trees and the maximum size of the tree are 2 user-defined parameters for Random Forest. The progress in automatic learning of algorithm is shown in chart Figure 3.1.

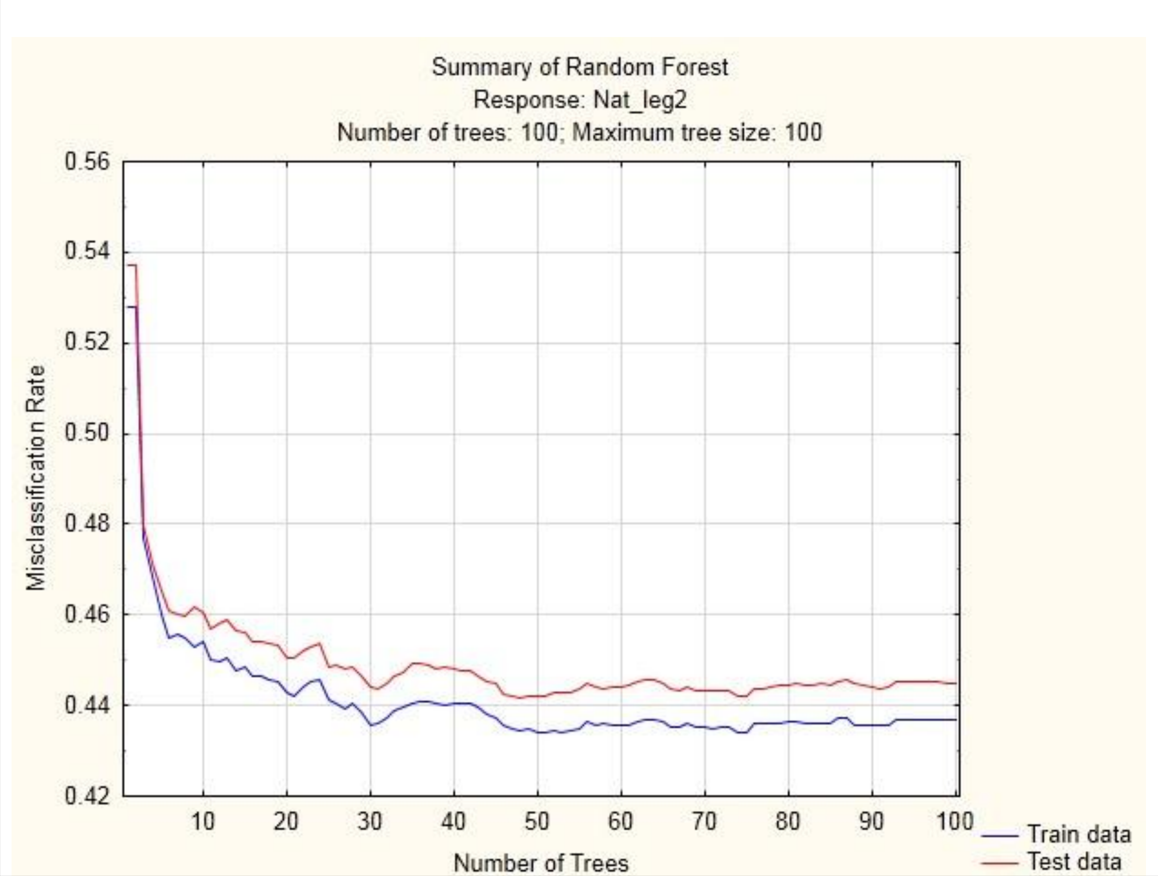


Figure 3.1: Process of Random Forest development.

The Random Forest modelling approach has interesting ability to rank covariates in terms of their importance for final result. The ranked variables are displayed in chart Figure 3.2.

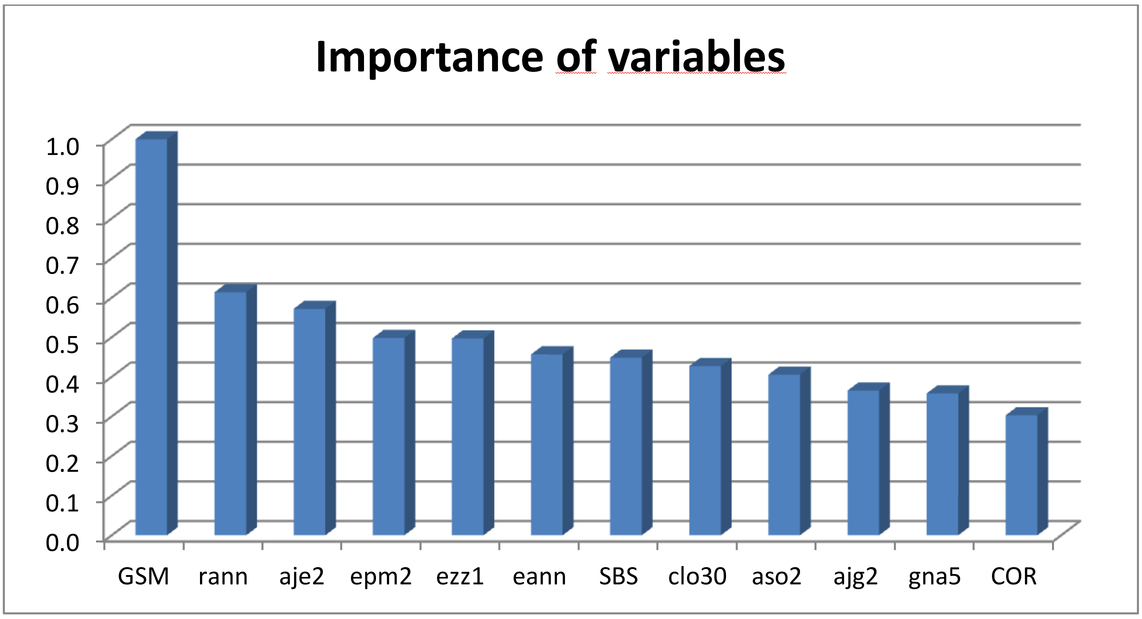


Figure 3.2: Ranking of importance of variables.

The parent material (SBS) is not ranked as one of the most important variables. It is not expected due to processes in real nature (Jenny, 1980). The low rank of the parent material covariate (7th position), is caused by fact that SBS is identical almost all over the area of interest. Therefore the parent material does not have a big impact from statistical point of view.

The risk estimates were 0.437 for training dataset and 0.445 for the testing dataset. The standard error was 0.0027 in training and 0.0041 in testing ensemble. The final result of the Random Forest model is shown in following map.

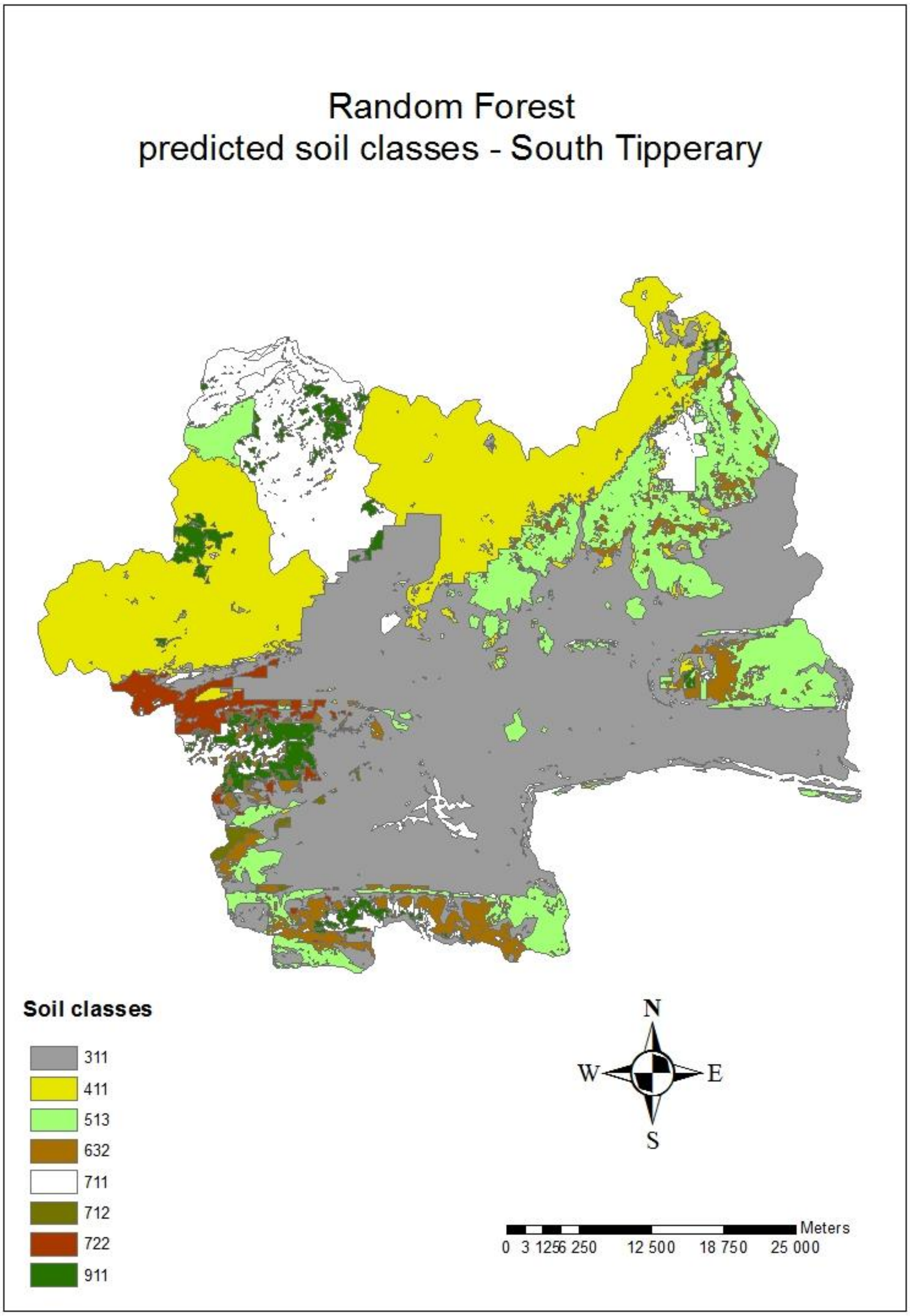

Random Forest
predicted soil classes - South Tipperary
Soil classes
311
411
513
632
711
712
722
911
N
W
E
S
Meters
0 3 125 6 250 12 500 18 750 25 000

Figure 3.3

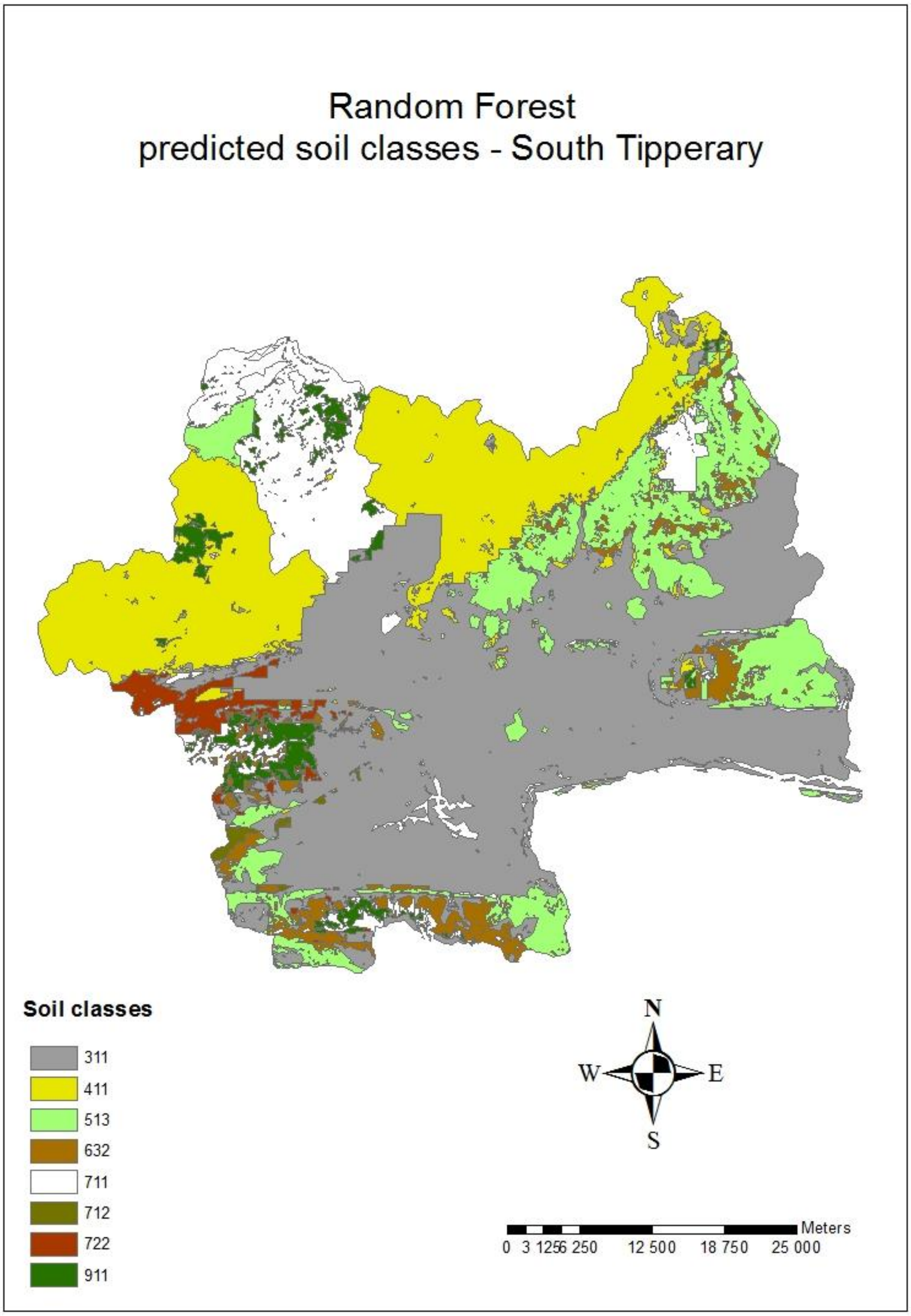


Figure 3.3: RF – predicted soil classes.

The result of Cramer’s V statistic was 0.376. The Cramer’s V asses the Random Forest model as the worst one from the tested classifiers.

3.1.2 Artificial Neural Network

ANN was executed in R-project and in STATISTICA10 software environment. The value of chi-squared of comparison AAN model against ground validation was 897.027. The result of Cramer’s V was 0.401. The map on

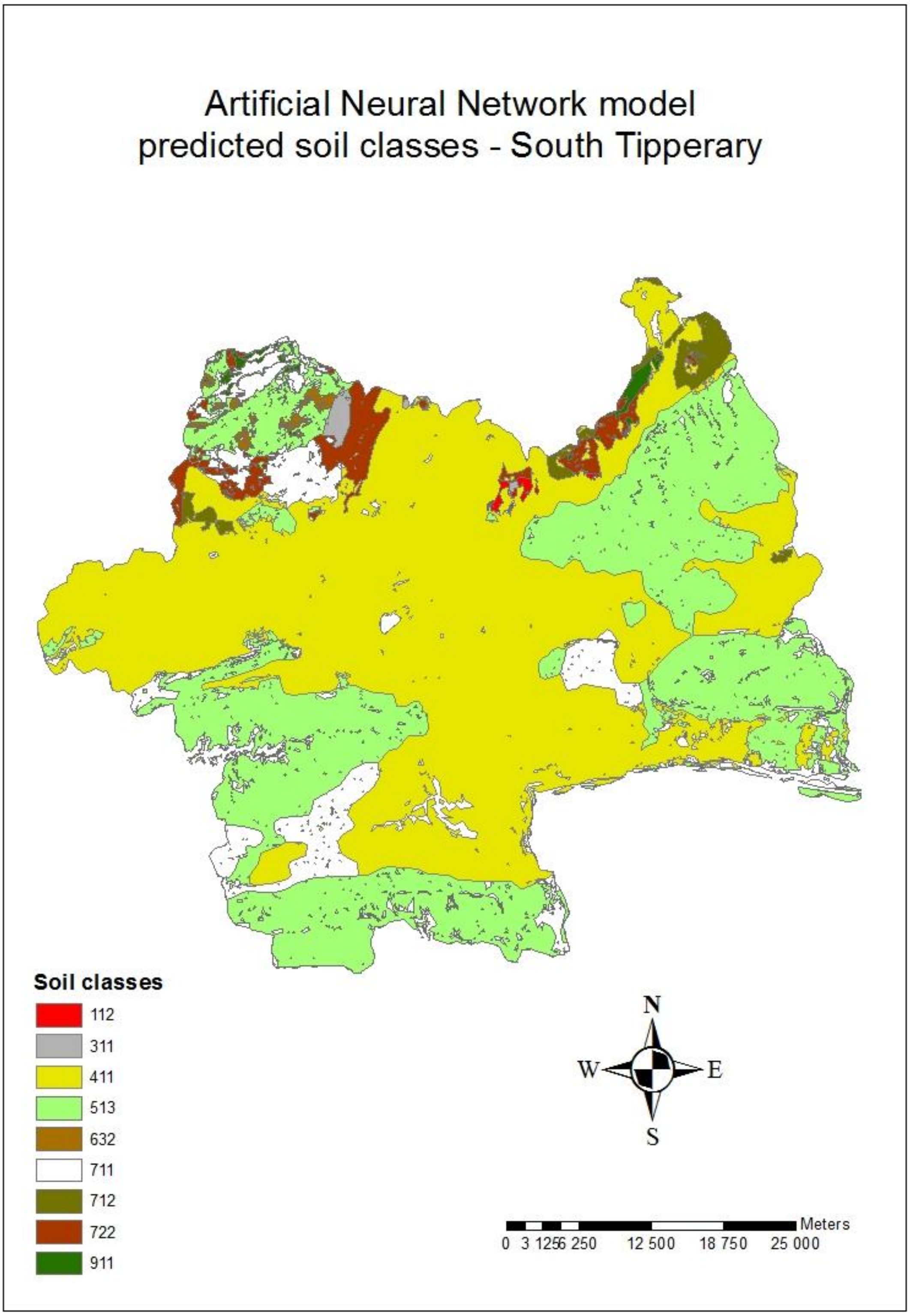


Figure 3.4: ANN – predicted soil classes.

3.1.3 Bayesian Belief Network

The error rate is derived from confusion matrix establishing relationships in Bayesian Belief Network. The error rate was 38.18%. In other words, model has 61.82% chance to make the correct decision.

Netica provides measurement of sensitivity of predicted variable to findings. These are based on entropy reductions (Ni, 2011). The sensitivity of predicted variable to each covariate is shown in table Table 3.1.

Table 3.1: Sensitivity to findings.

Node	**Sensitivity**
GSM	2.317
rann	1.433
eann	1.118
aje2	1.049
clo_2_i	1.012
SBS	1.008
ezz1	0.981
aso2	0.973
epm2	0.796
ajg2	0.627
gna5	0.578
COR	0.421

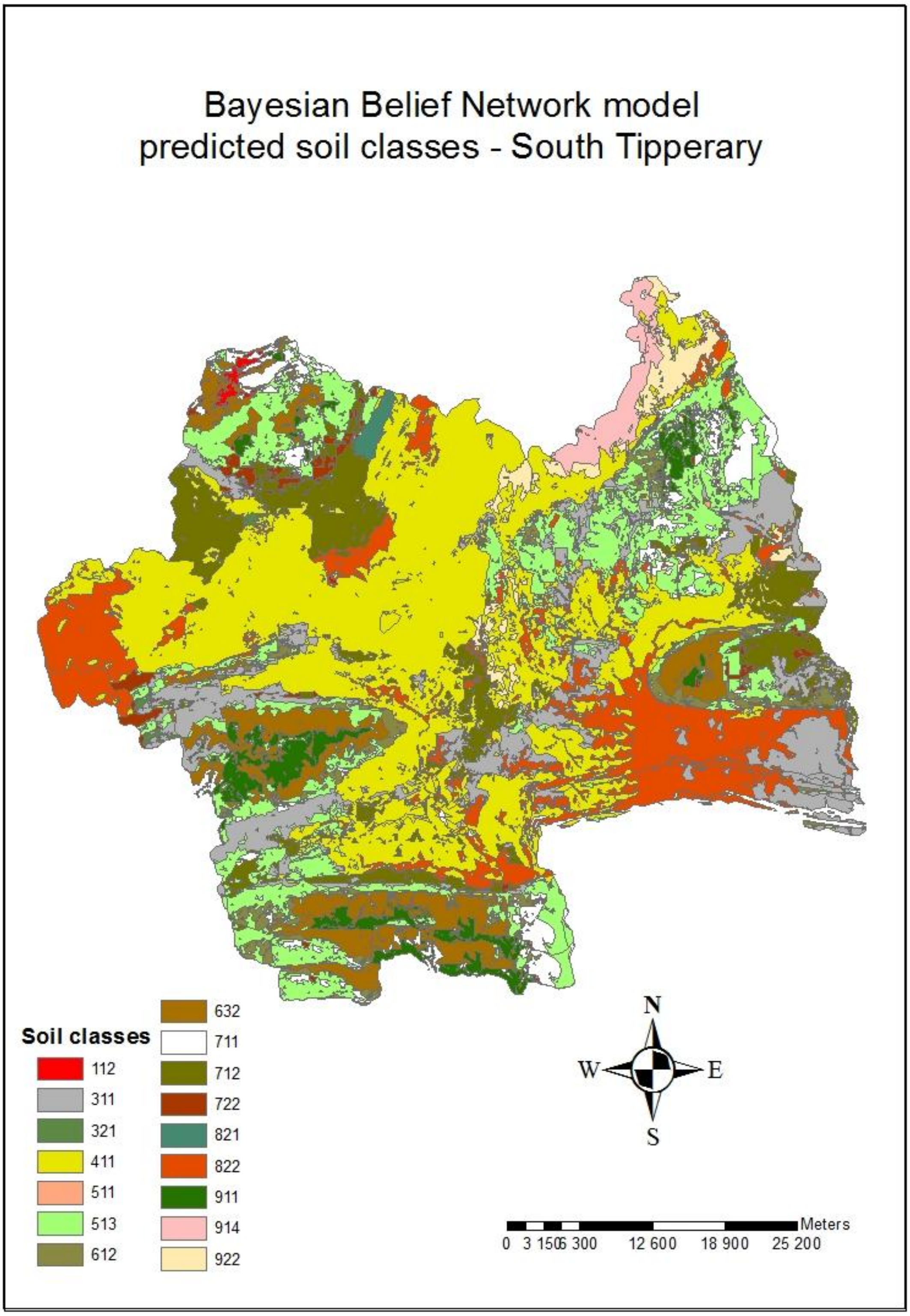


Figure 3.5: BBN –predicted soil classes.

3.1.4 The disproportion in number of predicted soil classes

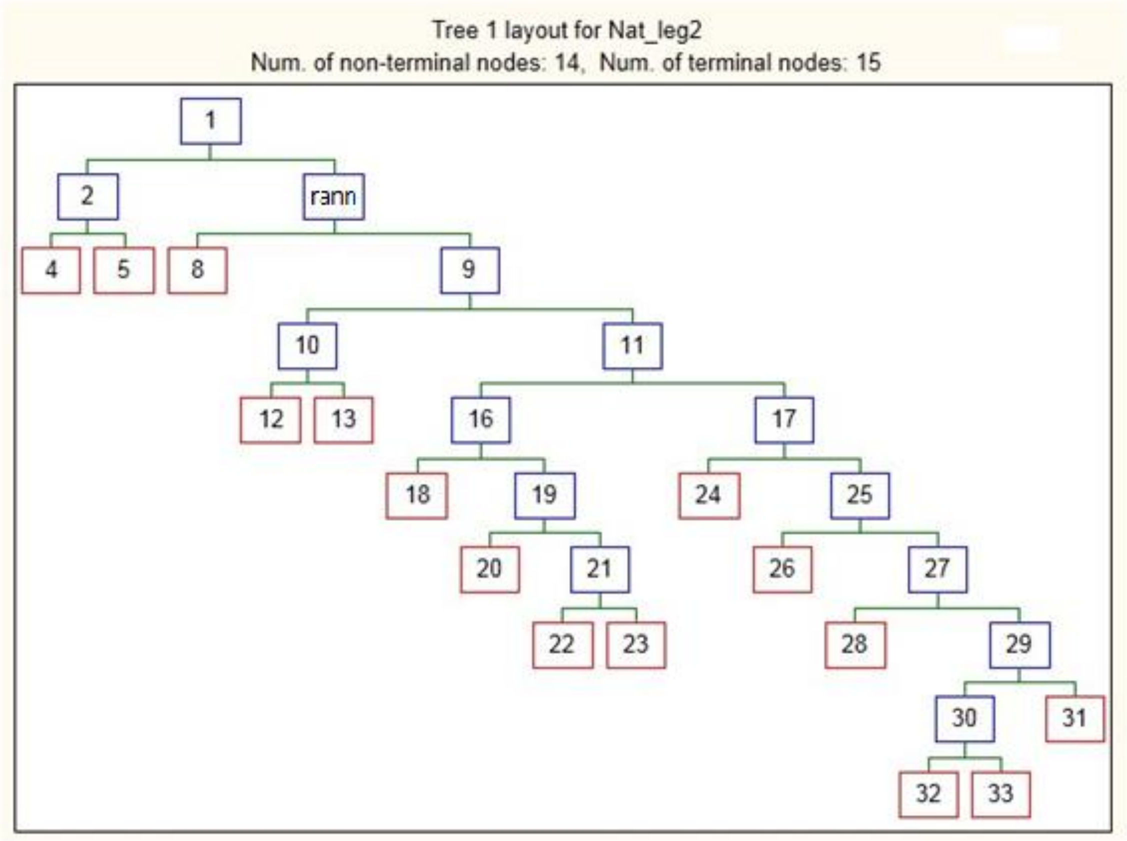
One issue arose during analysing results of different models so significantly that a separate section is devoted to closer analysis. BBN predicted 16 soil classes while RF 9 and ANN 8 soil classes.

To identify which of used variables caused this disproportion between models, the visual comparison between results of single classifiers and original layers of covariates was made. This was done with focus to the most important variables (ranking ability of RF). The Solar radiation covariate was identified as the variable with the most similar spatial pattern to RF prediction.

The Solar radiation variable divides the land into wide bands and is given in form of raster with low resolution (pixel size 1000x1000).

RF and ANN are black box models and it is difficult to investigate processes inside. However STATISTICA allows the user to display structure of the trees in RF.

The default number (100) of the trees in RF was used in the project. After detail investigation, with the focus on solar radiation variable, the 2 main types of the trees (Figure 3.6 and Figure 3.7) were found.



Tree 1 layout for Nat_leg2
Num. of non-terminal nodes: 14, Num. of terminal nodes: 15
1
2
rann
4
5
8
9
10
11
12
13
16
17
18
19
24
25
20
21
26
27
22
23
28
29
30
31
32
33

Figure 3.6: Structure of the tree in Random Forest.

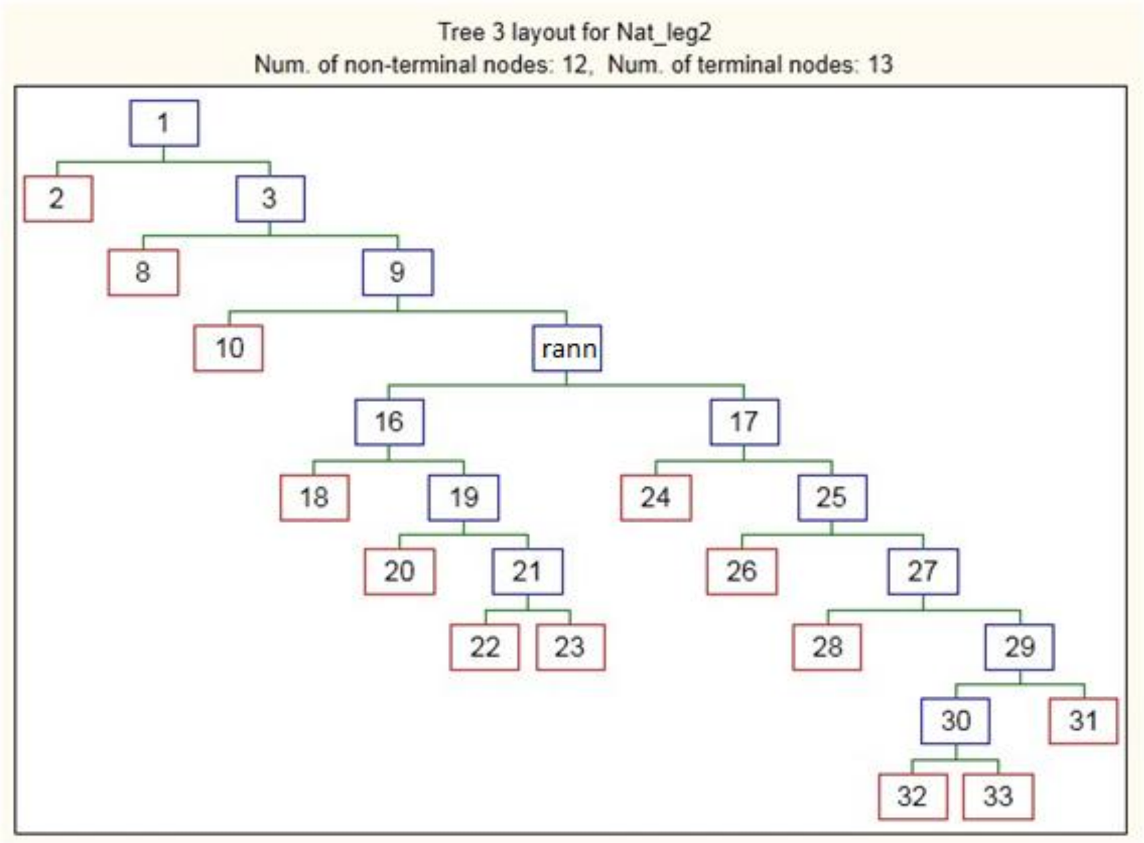

Tree 3 layout for Nat_leg2
Num. of non-terminal nodes: 12, Num. of terminal nodes: 13
1
2
3
8
9
10
rann
16
17
18
19
24
25
20
21
26
27
22
23
28
29
30
31
32
33

Figure 3.7: Structure of the tree in Random Forest.

The red squares on the picture above represent terminal nodes. In other words, they are different predicted soil classes. Note that decision based on solar radiation variable is always in high level in decision tree.

For instance on Figure 3.7, the solar radiation variable is in 4th level of the tree. When decision in this node is made, the soil classes in terminal nodes 24, 26, 28, 31, 32 and 33 are excluded. In rest of the tree are categories which finally appear in result of the model.

The situation on Figure 3.6 is even worst; the Soil radiation variable is in the second level, which leads to even higher exclusion.

It is more difficult to look inside black box of ANN. However due to spatial distribution of rainfall covariate and distribution of soil classes predicted by ANN, this solar radiation variable most likely excludes soil classes from result of ANN model.

The recommendation for further work is to try to re-run all of the models without the rainfall variable.

3.2 Confusion matrixes

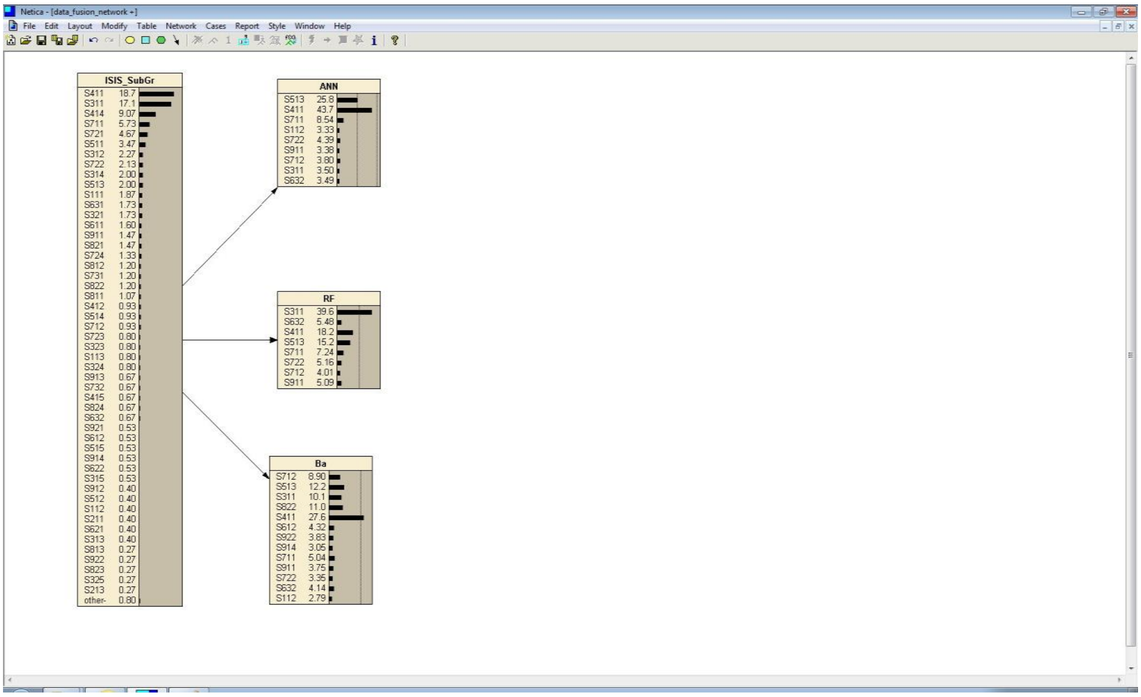
The cross-tabulation was applied to compare behaviour of different models against each other. Complete confusion matrixes can be found in appendix (Table A 1-3).

The cross-tabulation has shown that predictions of different models are best matching within Luvisol and Brown Podzolic soil classes. In the same time models tends to misclassify between these 2 soil classes. Additionally RF model disagree with ANN and classifies certain areas of Luvisol into Stagnogley soil class.

3.3 Results of different DF methods

3.3.1 DF – The first approach

This approach is based on simple Bayesian Belief Network. The structure of this BBN is shown on figure (Figure 3.8).



Netica - [data_fusion_network +]
File Edit Layout Modify Table Network Cases Report Style Window Help

ISIS_SubGr	
S411	18.7
S311	17.1
S414	9.07
S711	5.73
S721	4.67
S511	3.47
S312	2.27
S722	2.13
S314	2.00
S513	2.00
S111	1.87
S631	1.73
S321	1.73
S611	1.60
S911	1.47
S821	1.47
S724	1.33
S812	1.20
S731	1.20
S822	1.20
S811	1.07
S412	0.93
S514	0.93
S712	0.93
S723	0.80
S323	0.80
S113	0.80
S324	0.80
S913	0.67
S732	0.67
S415	0.67
S824	0.67
S632	0.67
S921	0.53
S612	0.53
S515	0.53
S914	0.53
S622	0.53
S315	0.53
S912	0.40
S512	0.40
S112	0.40
S211	0.40
S621	0.40
S313	0.40
S813	0.27
S922	0.27
S823	0.27
S325	0.27
S213	0.27
other-	0.80

ANN	
S513	25.8
S411	43.7
S711	8.54
S112	3.33
S722	4.39
S911	3.38
S712	3.80
S311	3.50
S632	3.49

RF	
S311	39.6
S632	5.48
S411	18.2
S513	15.2
S711	7.24
S722	5.16
S712	4.01
S911	5.09

Ba	
S712	8.90
S513	12.2
S311	10.1
S822	11.0
S411	27.6
S612	4.32
S922	3.83
S914	3.05
S711	5.04
S911	3.75
S722	3.35
S632	4.14
S112	2.79

Figure 3.8: Structure of BBN used in DF – the first approach

The results of the single thematic classifiers (RF, ANN and BBN) were used as covariates in this model.

The error rate derived from confusion matrix establishing relationships in data fusion BBN was 71.31%. However, in comparison of results, this approach supersedes any single classifier (Table 3.2:).

Table 3.2: Sensitivity to findings II

Node	**Sensitivity**
ANN	0.267
BBN	0.259
RF	0.227

The higher is value, the higher is impact of input covariate on the final result. Sensitivity to findings corresponds with result of Cramer’s V and ranks performance of different models in the same order. The result of Cramer’s V is 0.422154 and confirms the relevance of this approach.

The final output of this data fusion method is shown on Figure 3.9.

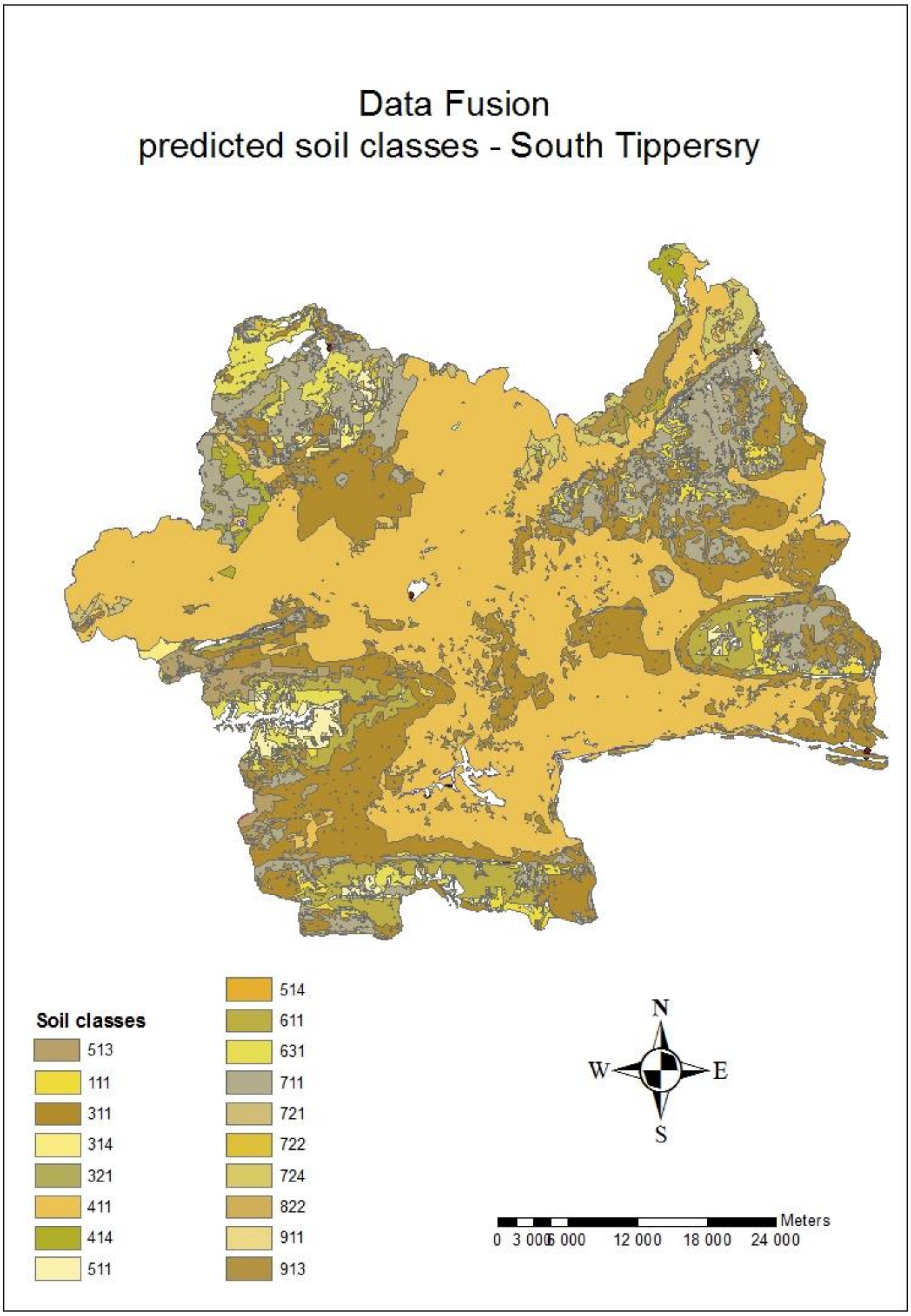


Figure 3.9: The first approach of DF - predicted soil classes.

3.3.2 DF – The second approach

The 2nd approach is using Microsoft Excel to extract predicted soil class with the highest weight from table containing predictions of all of the classifiers. The principle of this approach was explained in Methodology section in more detail. The final result is shown on Figure 3.10.

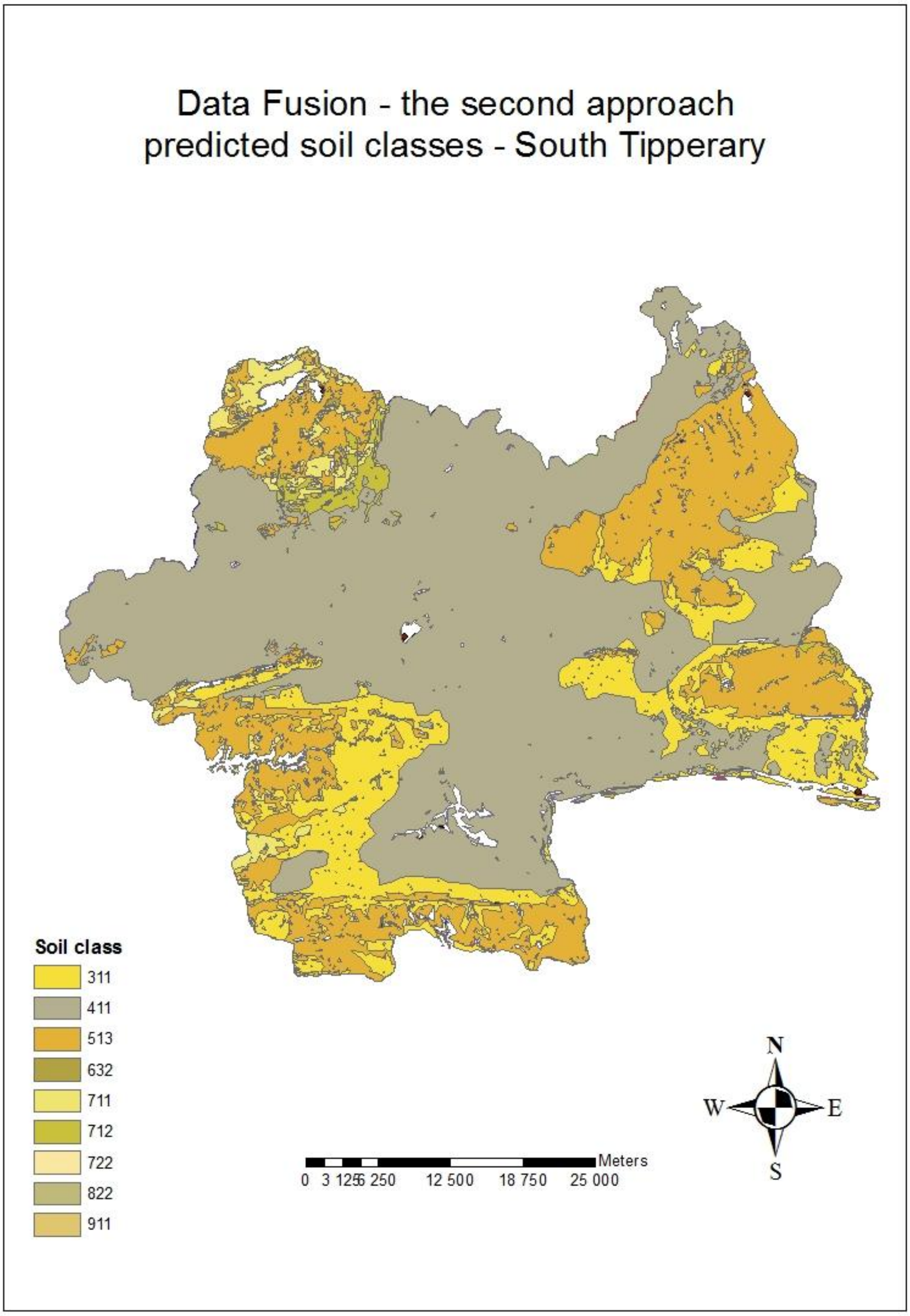


Figure 3.10: The second approach of DF – predicted soil classes.

The result of Cramer's V statistic is 0.422426 which is slightly better than 1st approach.

3.3.3 DF – The third approach

This approach was simple BBN with weights signed like probabilities to CPTs.

This method predicted only 1 soil class over all South Tipperary. It is caused by fact that the weight of this particular soil class was significantly higher than the other weights.

Because of result of this method was obviously meaningless, the Cramer's V statistic was not calculated and final map was not produced. However, it is not necessarily a dead end. It could be interesting to standardize the values of weights in different way and re-run the experiment.

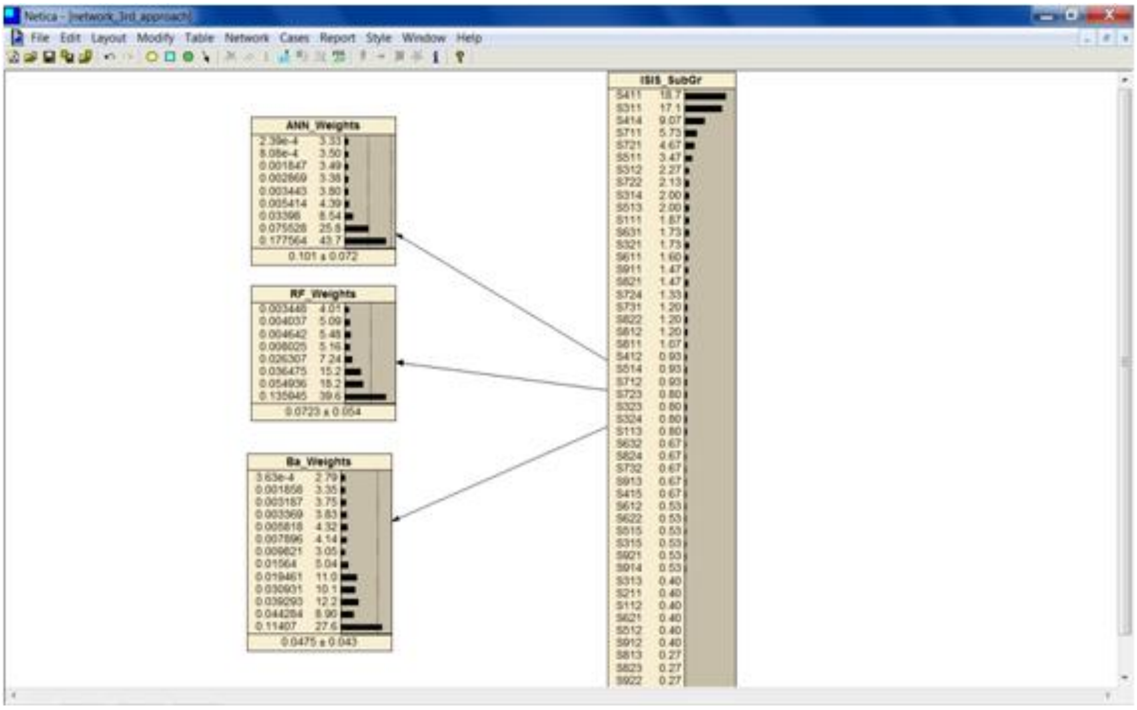
3.3.4 DF – The fourth approach

The RF model was applied. Both the training and the deployment data contains only 3 covariates – the weights delivered from each model.

The result was very poor. Only 2 soil classes were predicted. Because of result of this method was obviously meaningless, the Cramer's V statistic was not calculated and final map was not produced.

3.3.5 DF – The fifth approach

The BBN model was applied. Both the training and the deployment dataset contain only 3 covariates – the weight derived from each model. The structure of the BBN is shown on Figure 3.11



ANN_Weights	
2.39e-4	3.33
8.08e-4	3.50
0.001847	3.49
0.002869	3.38
0.003443	3.80
0.005414	4.39
0.03398	8.54
0.075528	25.8
0.177564	43.7
0.101 ± 0.072	

RF_Weights	
0.003445	4.01
0.004037	5.09
0.004642	5.48
0.008025	5.16
0.026307	7.24
0.036475	15.2
0.054936	18.2
0.135945	39.6
0.0723 ± 0.054	

Ba_Weights	
3.63e-4	2.79
0.001858	3.35
0.003187	3.75
0.003369	3.83
0.005818	4.32
0.007896	4.14
0.009821	3.05
0.01564	5.04
0.019461	11.0
0.030931	10.1
0.039293	12.2
0.044284	8.90
0.11407	27.6
0.0475 ± 0.043	

ISIS_SubGr	
S411	18.7
S311	17.1
S414	9.07
S711	5.73
S721	4.67
S511	3.47
S312	2.27
S722	2.13
S314	2.00
S513	2.00
S111	1.87
S631	1.73
S321	1.73
S611	1.60
S911	1.47
S821	1.47
S724	1.33
S731	1.20
S822	1.20
S812	1.20
S811	1.07
S412	0.93
S514	0.93
S712	0.93
S723	0.80
S323	0.80
S324	0.80
S113	0.80
S632	0.67
S824	0.67
S732	0.67
S913	0.67
S415	0.67
S612	0.53
S622	0.53
S515	0.53
S315	0.53
S921	0.53
S914	0.53
S313	0.40
S211	0.40
S112	0.40
S621	0.40
S512	0.40
S912	0.40
S813	0.27
S823	0.27
S922	0.27

Figure 3.11: Structure of BBN used in the fifth approach of DF.

The result of this data fusion method is on Figure 3.12.

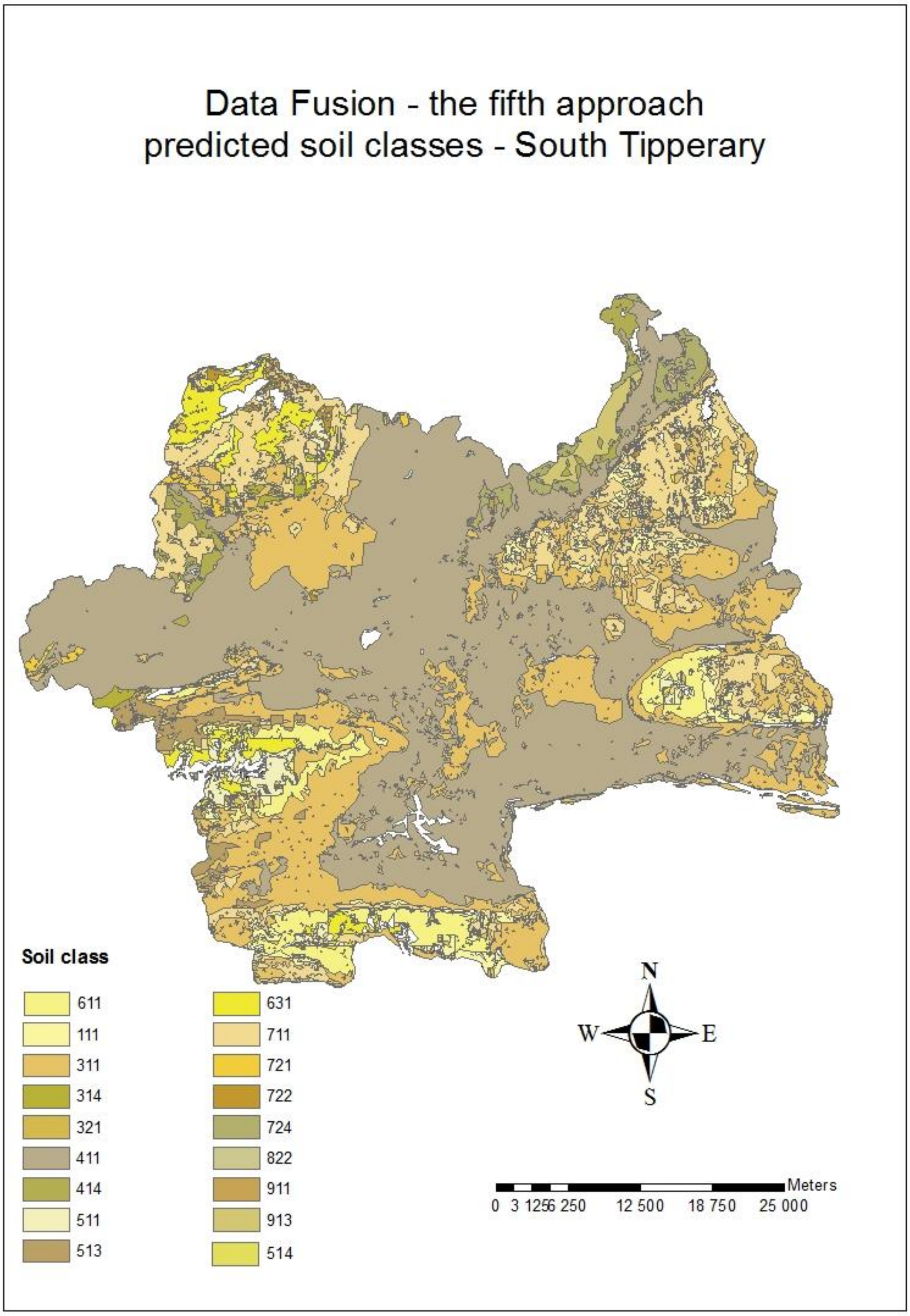


Figure 3.12: The fifth approach of DF – predicted soil classes.

The Cramer’s V statistic assesses this data fusion method as the best of the methods tested in this research. The result of Cramer’s V is 0. 49123.

3.4 Cramer’s V

The Cramer’s V statistic was used to compare performance of the different methods against Ground Truth validation data (GT).

Cramer’s V statistic was described in more detail in the Methodology section. Table 3.3. contains summary of Cramer’s V statistic results.

Table 3.3: The results of Cramer’s V statistic

Cramer's V (Assesment of single classifiers)	
GT/RF	0.376273
GT/ANN	0.40109
GT/BBN	0.410203
Cramer's V (Assessment of Data Fusion methods)	
GT/DF-1.approach	0.422154
GT/DF-2.approach	0.422426
GT/DF-5.approach	0.491295

The lower value indicates little association between variables, the higher value indicates strong association. Complete results of Cramer’s V statistic can be found in appendix (Tab. A 4-9).

4 Conclusion

The most important result is the methodology itself. The Data Fusion principles were applied in field of the soil science. This study has shown that combining multiple thematic classifications using Data Fusion provides better and a more justifiable thematic map rather than any single classifier.

Predictions of single classifiers (RF, ANN and BNN) are best matched within Luvisol and Brown Podzolic soil classes. At the same time models tend to misclassify between these 2 soil classes. Additionally RF model disagree with ANN and classifies certain areas of Luvisol into Stagnogley soil class.

The most important environmental variables from a statistical point of view were General Soil Map, Solar radiation and Actual/Potential drainage density ratio.

Cramer's V statistic evaluated Bayesian Belief Model as the best performing single classifier. The 5th data fusion approach (BBN with weights for each soil class incorporated as covariates) was considered as the best performing of all data fusion methods tested in this research.

5 Recommendations for further work

It should be highlighted that the aim of this project was not to gain the best prediction but the approach development. However, due to the random aspect in Random Forest model, to obtain a better prediction, the following suggestion is recommended. The "Monte Carlo" approach should be applied, where a higher number of forests (e.g. 1000) will be constructed. This can be done for example in R-project by a simple loop. The result of RF will be based on the variance of all of these forests.

The R-project and STATISTICA10 were used for experiments in this project. However in R users can define more parameters, STATISTICA behaved more user-friendly. The general recommendation is to use STATISTICA or decrease the number of elements in the data frame (decrease AOI or decrease sample density).

The solar radiation variable was identified as a cause of excluding soil classes inside black boxes (RF and ANN). Therefore a re-run of the experiment without solar radiation covariate in original datasets and comparison of these results is recommended.

REFERENCES

Agyare, W., Park, S. and Vlek, P. (2007), "Artificial neural network estimation of saturated hydraulic conductivity", *Vadose Zone Journal,* vol. 6, no. 2, pp. 423-431.

Bishop, C. M. (1995), "Neural networks for pattern recognition", .

Breiman, L. (1996), "Bagging predictors", *Machine Learning,* vol. 24, no. 2, pp. 123-140.

Breiman, L. (2001), "Random forests", *Machine Learning,* vol. 45, no. 1, pp. 5-32.

Breiman, L. and Spector, P. (1992), "Submodel selection and evaluation in regression. the x-random case", *International Statistical Review/Revue Internationale de Statistique,* , pp. 291-319.

Cavazzi, S.; Corstanje, R.; Mayr, T.; Hannam, J. and Fealy, R. (2013) “Are fine resolution digital elevation models always the best choice in digital soil mapping?”, *Geoderma*

Chatterjee, S. and Bandopadhyay, S. (2009), “Orebody modelling with uncertainty: a Bayesian Neural Network Approach”

Cramér, H. (1999). *Mathematical Methods of Statistics*, Princeton University Press

Congalton, R.G., Oderwald, R.G. and Mead, R.A. (1983), “Assessing Landsat classification accuracy using discrete multivariate statistical techniques”, *Photogrammetric Engineering and Remote Sensing*

De'ath, G. and Fabricius, K. E. (2000), "Classification and regression trees: a powerful yet simple technique for ecological data analysis", *Ecology,* vol. 81, no. 11, pp. 3178-3192.

Dlamini, W. M. (2011), "Application of Bayesian networks for fire risk mapping using GIS and remote sensing data", *GeoJournal,* vol. 76, no. 3, pp. 283-296.

Dobos, E., Darrousssian, J. and Montarella, L. (2005), “An SRTM-based procedure to delineate SOTER terrain units on 1:1 and 1:5 million scales”

Efron, B. (1979), "Bootstrap methods: another look at the jackknife", *The annals of Statistics,* vol. 7, no. 1, pp. 1-26.

Efron, B. (1983), "Estimating the error rate of a prediction rule: improvement on cross-validation", *Journal of the American Statistical Association,* , pp. 316-331.

Foody, G.M. (2003), “Thematic comparison: Evaluating the statistical significance of differences in classification accuracy”

Friedman, J., Hastie, T. and Tibshirani, R. (2001), *The elements of statistical learning,* Springer Series in Statistics.

Goutte, C. (1997), "Note on free lunches and cross-validation", *Neural computation,* vol. 9, no. 6, pp. 1245-1249.

Guisan, A. and Zimmermann, N. (2000), "Predictive habitat distribution models in ecology", *Ecological Modelling,* vol. 135, no. 2-3, pp. 147-186.

Guzman, J.G. and Al-Kaisi, M.M., (2011), “Landscape position effect on selected soil physical properties of reconstructed prairies in south-central Iowa”, *Journal of Soil and Water Conservation*

Haefner, J. W. (ed.) (2005), *Modeling Biological Systems: Principles and applications,* 2nd ed, Springer.

Halfon, E. (1985), "The bootstrap and the jackknife in ecotoxicology or nonparametric estimates of standard error", *Chemosphere,* vol. 14, no. 9, pp. 1433-1440.

Hough, R. L., Towers, W. and Aalders, I. (2010), "The Risk of Peat Erosion from Climate Change: Land Management Combinations—An Assessment with Bayesian Belief Networks", *Human and Ecological Risk Assessment,* vol. 16, no. 5, pp. 962-976.

Jensen, F. V. (1996), *An introduction to Bayesian networks,* UCL press London.

Jha, V.C. and Kapat, S. (2009), “Rill and Gully erosion risk of lateritic terrain in south-western Birshum district, west Bengalia, India”, Sociedade & Natureza

Legendre, P. and Legendre, L. (1998), *Numerical ecology,* Elsevier Science & Technology.

Levins, R. (1966), "The strategy of model building in population biology", *American Scientist,* vol. 54, no. 4, pp. 421-431.

Liaw, A. and Wiener, M. (2002), "Classification and Regression by randomForest", *R news,* vol. 2, no. 3, pp. 18-22.

Lou, W. 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 Research International,* vol. 34, no. 7, pp. 573-579.

Maier, H. R. and Dandy, G. C. (2001), "Neural network based modelling of environmental variables: A systematic approach", *Mathematical and Computer Modelling,* vol. 33, no. 6, pp. 669-682.

McBratney, A.B. (2003)"*On digital soil mapping*", Elsevier Science & Technology.

Pearson, R. G., Dawson, T., Berry, P. and Harrison, P. (2002), "SPECIES: a spatial evaluation of climate impact on the envelope of species", *Ecological Modelling,* vol. 154, no. 3, pp. 289-300.

Ripley, B. D. (1996), "Pattern recognition via neural networks", *a volume of Oxford Graduate Lectures on Neural Networks, title to be decided.Oxford University Press.[See http://www.stats.ox.ac.uk/ripley/papers.html.],* .

Rumelhart, D. E., Hintont, G. E. and Williams, R. J. (1986), "Learning representations by back-propagating errors", *Nature,* vol. 323, no. 6088, pp. 533-536.

Sommer, M. (2003), ''*Hierarchical data fusion for mapping soil units at field scale''*, Elsevier Science & Technology.

Suuster, E., Ritz, C., Roostalu, H., Kõlli, R. and Astover, A. (2012), "Modelling soil organic carbon concentration of mineral soils in arable land using legacy soil data", *European Journal of Soil Science,* .

Tibshirani, R. J. and Efron, B. (1993), "An introduction to the bootstrap", *Monographs on Statistics and Applied Probability,* vol. 57, pp. 1-436.

Tucker, G.E.; Catani, F.; Rinaldo, A. And Bras, R.L. (2001). "Statistical analysis of drainage density from digital terrain data"

Van Houwelingen, J. and Le Cessie, S. (1990), "Predictive value of statistical models", *Statistics in medicine,* vol. 9, no. 11, pp. 1303-1325.

Wiesmeier, M., Barthold, F., Blank, B. and Kögel-Knabner, I. (2011), "Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem", *Plant and Soil,* vol. 340, no. 1, pp. 7-24.

Araujo, M. B. and New, M. (2007), "Ensemble forecasting of species distributions", Trends in Ecology & Evolution, vol. 22, no. 1, pp. 42-47.

Benestad, R. E. (2004), "Tentative probabilistic temperature scenarios for northern Europe", Tellus Series A-Dynamic Meteorology and Oceanography, vol. 56, no. 2.

Bernsel, A., Viklund, H., Hennerdal, A. and Elofsson, A. (2009), "TOPCONS: consensus prediction of membrane protein topology", Nucleic acids research, vol. 37, pp. W465-W468.

Cramer, W., Bondeau, A., Woodward, F., Prentice, I., Betts, R., Brovkin, V., Cox, P., Fisher, V., Foley, J., Friend, A., Kucharik, C., Lomas, M., Ramankutty, N., Sitch, S., Smith, B., White, A. and Young-Molling, C. (2001), "Global response of terrestrial ecosystem structure and function to CO2 and climate change: results from six dynamic global vegetation models", Global Change Biology, vol. 7, no. 4, pp. 357-373.

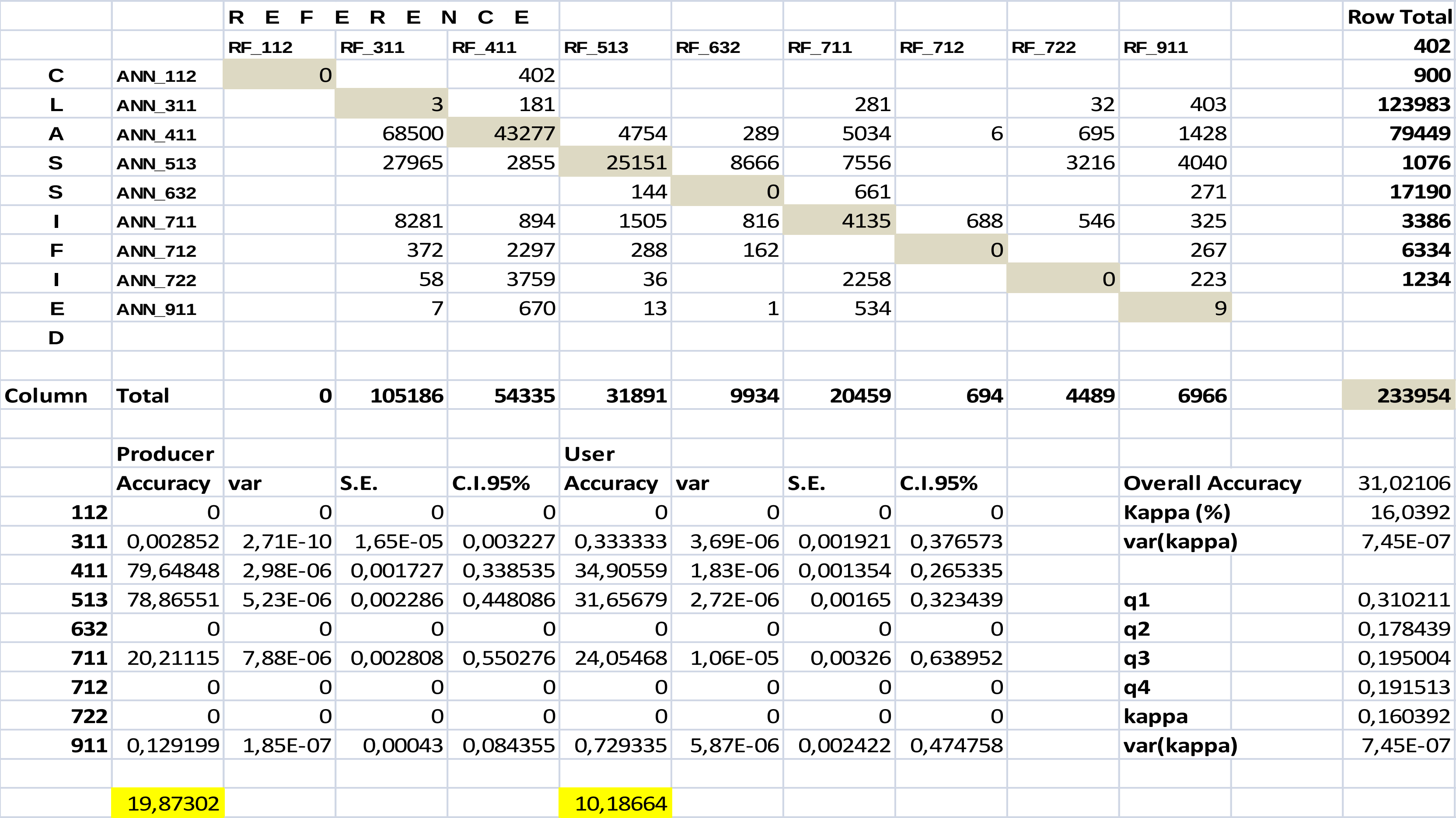
Sanders, F. (1963), "On subjective probability forecasting", Journal of Applied Meteorology, vol. 2, no. 2, pp. 191-201.

Thuiller, W., Lavorel, S., Araujo, M., Sykes, M. and Prentice, I. (2005), "Climate change threats to plant diversity in Europe", Proceedings of the National Academy of Sciences of the United States of America, vol. 102, no. 23, pp. 8245-8250.

Winkler, R. L. (1989), "Combining Forecasts - a Philosophical Basis and some Current Issues", International Journal of Forecasting, vol. 5, no. 4.

A Appendix

Table A-1: Confusion matrix (ANN vs RF)



		REFERENCE										Row Total
		RF_112	RF_311	RF_411	RF_513	RF_632	RF_711	RF_712	RF_722	RF_911		**402**
C	**ANN_112**	0		402								**900**
L	**ANN_311**		3	181			281		32	403		**123983**
A	**ANN_411**		68500	43277	4754	289	5034	6	695	1428		**79449**
S	**ANN_513**		27965	2855	25151	8666	7556		3216	4040		**1076**
S	**ANN_632**				144	0	661			271		**17190**
I	**ANN_711**		8281	894	1505	816	4135	688	546	325		**3386**
F	**ANN_712**		372	2297	288	162		0		267		**6334**
I	**ANN_722**		58	3759	36		2258		0	223		**1234**
E	**ANN_911**		7	670	13	1	534			9		
D												
Column	**Total**	**0**	**105186**	**54335**	**31891**	**9934**	**20459**	**694**	**4489**	**6966**		**233954**
	Producer				**User**							
	Accuracy	**var**	**S.E.**	**C.I.95%**	**Accuracy**	**var**	**S.E.**	**C.I.95%**		**Overall Accuracy**		31,02106
112	0	0	0	0	0	0	0	0		**Kappa (%)**		16,0392
311	0,002852	2,71E-10	1,65E-05	0,003227	0,333333	3,69E-06	0,001921	0,376573		**var(kappa)**		7,45E-07
411	79,64848	2,98E-06	0,001727	0,338535	34,90559	1,83E-06	0,001354	0,265335				
513	78,86551	5,23E-06	0,002286	0,448086	31,65679	2,72E-06	0,00165	0,323439		**q1**		0,310211
632	0	0	0	0	0	0	0	0		**q2**		0,178439
711	20,21115	7,88E-06	0,002808	0,550276	24,05468	1,06E-05	0,00326	0,638952		**q3**		0,195004
712	0	0	0	0	0	0	0	0		**q4**		0,191513
722	0	0	0	0	0	0	0	0		**kappa**		0,160392
911	0,129199	1,85E-07	0,00043	0,084355	0,729335	5,87E-06	0,002422	0,474758		**var(kappa)**		7,45E-07
	19,87302				10,18664							

Table A-2: Confusion matrix (BBN vs ANN)

		REFERENCE																	
		ANN_112	ANN_311	ANN_321	ANN_411	ANN_511	ANN_513	ANN_612	ANN_632	ANN_711	ANN_712	ANN_722	ANN_821	ANN_822	ANN_911	ANN_914	ANN_922		Row Total
C	Ba_112	0			85		68		48	236		9			244				690
L	Ba_311		2		8197		9389			3987		496			2				22073
A	Ba_321			0	73		207			114									394
S	Ba_411	14	69		75145		4945			2675	397	4			5				83254
S	Ba_511				30	0													30
I	Ba_513		569		3846		24286		119	3745	2	595			7				33169
F	Ba_612				175		5425	0		90									5690
I	Ba_632				1		13502		663	398		113			92				14769
E	Ba_711				30		5574			1007		140			118				6869
D	Ba_712		138		12547		4522			1756	663	1809			6				21441
	Ba_722		3		443		886			912		787			3				3034
	Ba_821		7		360		4			4		587	0		5				967
	Ba_822		18		18643		4271			1557	480	81		0	4				25054
	Ba_911		32		1		6337		246	641					88				7345
	Ba_914		5		1317					11	432	1267			634	0			3666
	Ba_922	388	57		3090		33			57	1412	446			26		0		5509
Column	Total	402	900	0	123983	0	79449	0	1076	17190	3386	6334	0	0	1234	0	0		233954

	Producer				User							
	Accuracy	var	S.E.	C.I.95%	Accuracy	var	S.E.	C.I.95%		Overall Accuracy		43,8723
1	0	0	0	0	0	0	0	0		Kappa (%)		26,01251
2	0,222222	2,46E-06	0,00157	0,307642	0,009061	4,1E-09	6,41E-05	0,012557		var(kappa)		8,15E-07
3	0	0	0	0	0	0	0	0				
4	60,60912	1,93E-06	0,001388	0,271983	90,25993	1,06E-06	0,001028	0,20141		q1		0,438723
5	0	0	0	0	0	0	0	0		q2		0,241389
6	30,56804	2,67E-06	0,001634	0,32035	73,21897	5,91E-06	0,002431	0,476558		q3		0,33557
7	0	0	0	0	0	0	0	0		q4		0,318215
8	61,6171	0,00022	0,014826	2,905825	4,489133	2,9E-06	0,001704	0,333955		kappa		0,260125
9	5,858057	3,21E-06	0,001791	0,351064	14,66007	1,82E-05	0,004268	0,836476		var(kappa)		8,15E-07
10	19,58063	4,65E-05	0,006819	1,336615	3,092207	1,4E-06	0,001182	0,231711				
11	12,42501	1,72E-05	0,004145	0,812373	25,93935	6,33E-05	0,007957	1,55963				
12	0	0	0	0	0	0	0	0				
13	0	0	0	0	0	0	0	0				
14	7,13128	5,37E-05	0,007326	1,435876	1,198094	1,61E-06	0,001269	0,248822				
15	0	0	0	0	0	0	0	0				
16	0	0	0	0	0	0	0	0				
	22,00127				13,30418							

Table A-3: Confusion matrix (Rf vs BBN)

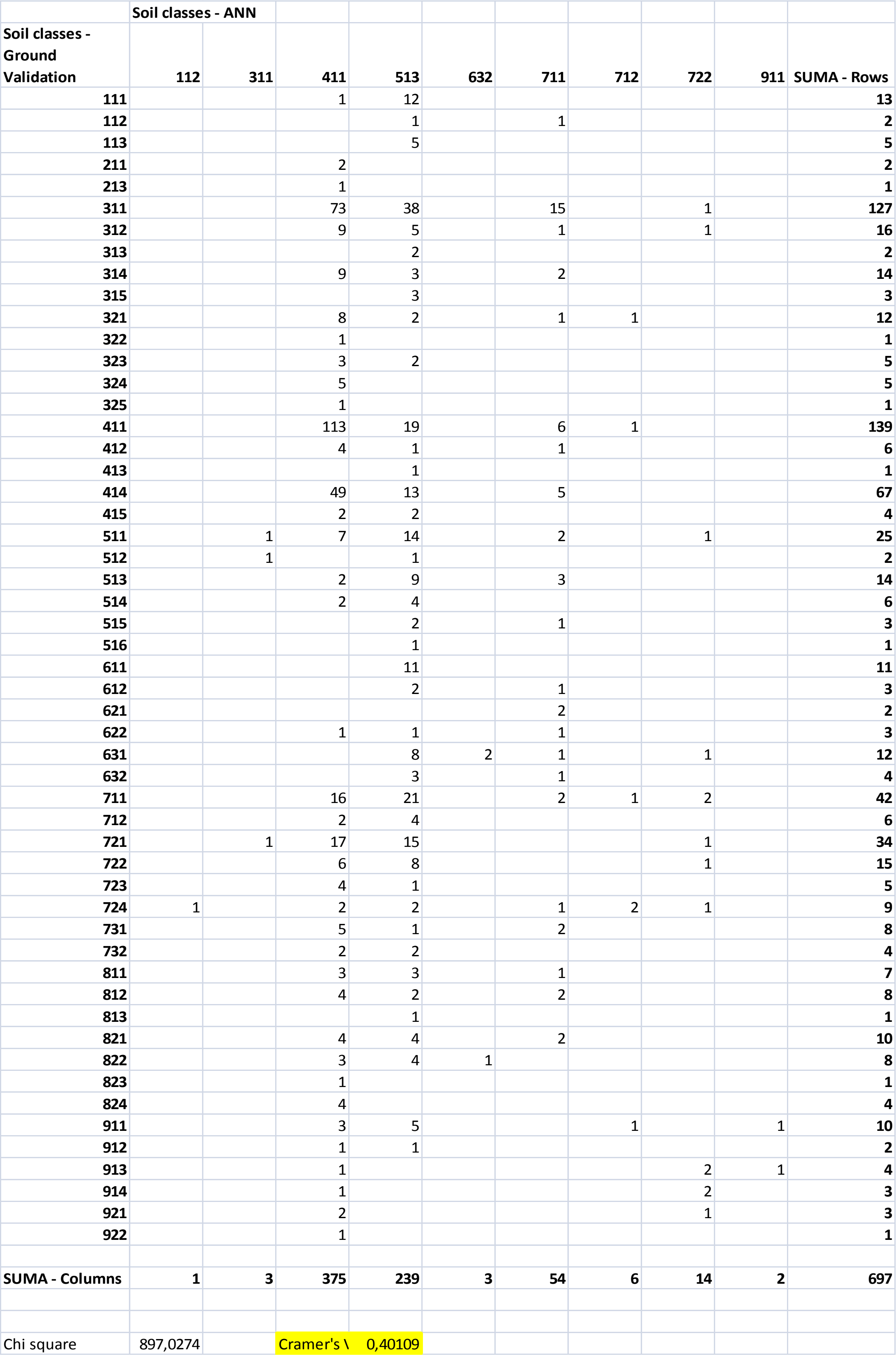
		REFERENCE																	
		RF_112	RF_311	RF_321	RF_411	RF_511	RF_513	RF_612	RF_632	RF_711	RF_712	RF_722	RF_821	RF_822	RF_911	RF_914	RF_922		Row Total
C	Ba_112	0			85		68		48	236		9			244				690
L	Ba_311		2		8197		9389			3987		496			2				22073
A	Ba_321			0	73		207			114									394
S	Ba_411	14	69		75145		4945			2675	397	4			5				83254
S	Ba_511				30	0													30
I	Ba_513		569		3846		24286		119	3745	2	595			7				33169
F	Ba_612				175		5425	0		90									5690
I	Ba_632				1		13502		663	398		113			92				14769
E	Ba_711				30		5574			1007		140			118				6869
D	Ba_712		138		12547		4522			1756	663	1809			6				21441
	Ba_722		3		443		886			912		787			3				3034
	Ba_821		7		360		4			4		587	0		5				967
	Ba_822		18		18643		4271			1557	480	81		0	4				25054
	Ba_911		32		1		6337		246	641					88				7345
	Ba_914		5		1317					11	432	1267			634	0			3666
	Ba_922	388	57		3090		33			57	1412	446			26		0		5509
Column	Total	402	900	0	123983	0	79449	0	1076	17190	3386	6334	0	0	1234	0	0		233954

	Producer				User						
	Accuracy	var	S.E.	C.I.95%	Accuracy	var	S.E.	C.I.95%		Overall Accuracy	43,8723
112	0	0	0	0	0	0	0	0		Kappa (%)	26,01251
311	0,222222	2,46E-06	0,00157	0,307642	0,009061	4,1E-09	6,41E-05	0,012557		var(kappa)	8,15E-07
321	0	0	0	0	0	0	0	0			
411	60,60912	1,93E-06	0,001388	0,271983	90,25993	1,06E-06	0,001028	0,20141		q1	0,438723
511	0	0	0	0	0	0	0	0		q2	0,241389
513	30,56804	2,67E-06	0,001634	0,32035	73,21897	5,91E-06	0,002431	0,476558		q3	0,33557
612	0	0	0	0	0	0	0	0		q4	0,318215
632	61,6171	0,00022	0,014826	2,905825	4,489133	2,9E-06	0,001704	0,333955		kappa	0,260125
711	5,858057	3,21E-06	0,001791	0,351064	14,66007	1,82E-05	0,004268	0,836476		var(kappa)	8,15E-07
712	19,58063	4,65E-05	0,006819	1,336615	3,092207	1,4E-06	0,001182	0,231711			
722	12,42501	1,72E-05	0,004145	0,812373	25,93935	6,33E-05	0,007957	1,55963			
821	0	0	0	0	0	0	0	0			
822	0	0	0	0	0	0	0	0			
911	7,13128	5,37E-05	0,007326	1,435876	1,198094	1,61E-06	0,001269	0,248822			
914	0	0	0	0	0	0	0	0			
922	0	0	0	0	0	0	0	0			
	12,37572				13,30418						

Table A-4: Cramer’s V (GT vs. RF)

Počet z COUNT	Soil classes - RF								
Soil classes - Ground Validaton	311	411	513	632	711	712	722	911	SUMA - Rows
111	2	1	7	3					13
112	1		1						2
113	2		1	1	1				5
211	2								2
213		1							1
311	69	21	23	4	6	1	1	2	127
312	9	4	1		1		1		16
313			2						2
314	7	2	1	1	2			1	14
315	2		1						3
321	9	2		1					12
322	1								1
323	3	2							5
324	3	2							5
325		1							1
411	95	27	15	1	1				139
412	3	3							6
413			1						1
414	35	18	10	1	2			1	67
415	2		1	1					4
511	8	4	6	1	1	1	1	3	25
512							1	1	2
513	5		3	1	1	2	2		14
514	2		4						6
515			3						3
516	1								1
611	4		4	1			2		11
612	2				1				3
621							2		2
622	1				1		1		3
631	1	1	1		3		3	3	12
632	1			1	1			1	4
711	7	11	13		6		1	4	42
712	2		1		1		2		6
721	15	11	3	2	2		1		34
722	6	2	2		5				15
723	4		1						5
724	1	4	2	1	1				9
731	2	4	2						8
732	3						1		4
811	6					1			7
812	6	1				1			8
813							1		1
821	5	1	2		1		1		10
822	2	2	3		1				8
823		1							1
824	1	3							4
911	3	4	1	1			1		10
912		1			1				2
913		4							4
914		3							3
921		2						1	3
922	1								1
SUMA - Columns	334	143	115	21	39	6	22	17	697
Chi square	690,7756		Cramer's V	0,376273					

Table A-5: Cramer’s V (GT vs. ANN)

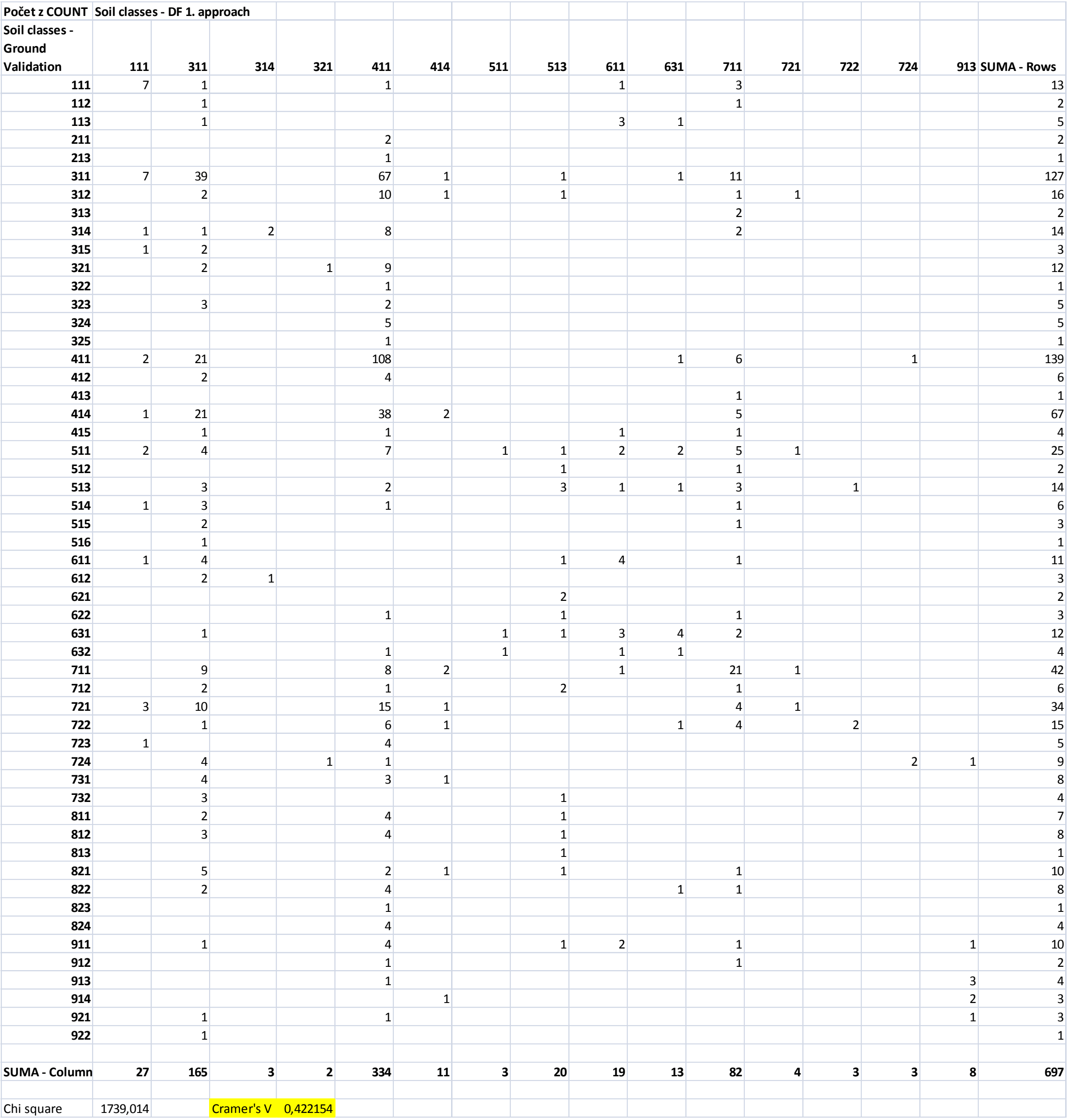


	Soil classes - ANN									
Soil classes - Ground Validation	**112**	**311**	**411**	**513**	**632**	**711**	**712**	**722**	**911**	**SUMA - Rows**
111			1	12						**13**
112				1		1				**2**
113				5						**5**
211			2							**2**
213			1							**1**
311			73	38		15		1		**127**
312			9	5		1		1		**16**
313				2						**2**
314			9	3		2				**14**
315				3						**3**
321			8	2		1	1			**12**
322			1							**1**
323			3	2						**5**
324			5							**5**
325			1							**1**
411			113	19		6	1			**139**
412			4	1		1				**6**
413				1						**1**
414			49	13		5				**67**
415			2	2						**4**
511		1	7	14		2		1		**25**
512		1		1						**2**
513			2	9		3				**14**
514			2	4						**6**
515				2		1				**3**
516				1						**1**
611				11						**11**
612				2		1				**3**
621						2				**2**
622			1	1		1				**3**
631				8	2	1		1		**12**
632				3		1				**4**
711			16	21		2	1	2		**42**
712			2	4						**6**
721		1	17	15				1		**34**
722			6	8				1		**15**
723			4	1						**5**
724	1		2	2		1	2	1		**9**
731			5	1		2				**8**
732			2	2						**4**
811			3	3		1				**7**
812			4	2		2				**8**
813				1						**1**
821			4	4		2				**10**
822			3	4	1					**8**
823			1							**1**
824			4							**4**
911			3	5			1		1	**10**
912			1	1						**2**
913			1					2	1	**4**
914			1					2		**3**
921			2					1		**3**
922			1							**1**
SUMA - Columns	**1**	**3**	**375**	**239**	**3**	**54**	**6**	**14**	**2**	**697**
Chi square	897,0274		Cramer's \	0,40109						

Table A-6: Cramer’s V (GT vs. BBN)

Počet z COUNT	Soil classes - BBN													
Soil classes - Ground Validation	112	311	411	513	612	632	711	712	722	822	911	914	922	SUMA - Rows
111		1		6	4	1	1							13
112		1		1										2
113						4	1							5
211			2											2
213			1											1
311		18	52	12	4		6	14	2	14	2		3	127
312		4	5	1				2		4				16
313				2										2
314		2	4	1			1	2	2	2				14
315				2	1									3
321			8				1			3				12
322			1											1
323		1	1	1				1					1	5
324			4							1				5
325			1											1
411		14	74	11	1		2	6	1	26			4	139
412		1	5											6
413				1										1
414		9	31	10			3	10		4				67
415				1		1		1		1				4
511		3	3	7	2	3	1			4	2			25
512		1						1						2
513		3		8		1	1				1			14
514		1		2	1					1	1			6
515		1		2										3
516					1									1
611		1		3	1	4	2							11
612				1	1				1					3
621		2												2
622		1	1	1										3
631	1			5	1	3			1		1			12
632	1					1				1	1			4
711			12	14	1	1	2	9		1	2			42
712		1	1	1				1	1		1			6
721		6	8	4	1			6		8	1			34
722		1	2	5		1	2	1		3				15
723		1	2		1					1				5
724			4					1		2		1	1	9
731			2				1	2		2			1	8
732				1	1			2						4
811		1	2					1		3				7
812		1	5	1				1						8
813				1										1
821		1	5				1	3						10
822		2	2	1						2	1			8
823				1										1
824			3							1				4
911			2	3		2						1	2	10
912			1	1										2
913			1									3		4
914												3		3
921			1							1		1		3
922								1						1
SUMA - Columns	2	78	246	111	21	22	25	65	8	85	13	9	12	697
Chi square	1407,383		Cramer's V	0,410203										

Table A-7: Cramer's V (GT vs. DF 1. approach)



Počet z COUNT	Soil classes - DF 1. approach														
Soil classes - Ground Validation	111	311	314	321	411	414	511	513	611	631	711	721	722	724	913	SUMA - Rows
111	7	1			1				1		3					13
112		1									1					2
113		1							3	1						5
211					2											2
213					1											1
311	7	39			67	1		1		1	11					127
312		2			10	1		1			1	1				16
313											2					2
314	1	1	2		8						2					14
315	1	2														3
321		2		1	9											12
322					1											1
323		3			2											5
324					5											5
325					1											1
411	2	21			108					1	6			1		139
412		2			4											6
413											1					1
414	1	21			38	2					5					67
415		1			1				1		1					4
511	2	4			7		1	1	2	2	5	1				25
512								1			1					2
513		3			2			3	1	1	3		1			14
514	1	3			1						1					6
515		2									1					3
516		1														1
611	1	4						1	4		1					11
612		2	1													3
621								2								2
622					1			1			1					3
631		1					1	1	3	4	2					12
632					1		1		1	1						4
711		9			8	2			1		21	1				42
712		2			1			2			1					6
721	3	10			15	1					4	1				34
722		1			6	1				1	4		2			15
723	1				4											5
724		4		1	1									2	1	9
731		4			3	1										8
732		3						1								4
811		2			4			1								7
812		3			4			1								8
813								1								1
821		5			2	1		1			1					10
822		2			4					1	1					8
823					1											1
824					4											4
911		1			4			1	2		1				1	10
912					1						1					2
913					1										3	4
914						1									2	3
921		1			1										1	3
922		1														1
SUMA - Column	**27**	**165**	**3**	**2**	**334**	**11**	**3**	**20**	**19**	**13**	**82**	**4**	**3**	**3**	**8**	**697**
Chi square	1739,014		Cramer's V	0,422154												

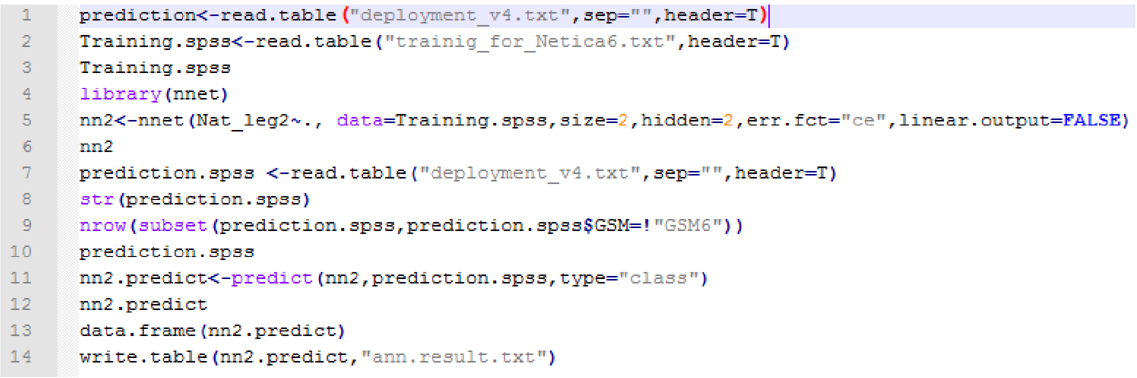
Table A-8: Cramer's V (GT vs. DF 2. approach)

Soil classes - Ground Validation	311	oach	513	711	712	SUMA - Rows
111	2	1	10			13
112	1		1			2
113	2		3			5
211		2				2
213		1				1
311	20	74	31	1	1	127
312	3	10	2		1	16
313			2			2
314	1	9	2	2		14
315	2		1			3
321	3	9				12
322		1				1
323	2	3				5
324		5				5
325		1				1
411	11	114	14			139
412	2	4				6
413			1			1
414	9	49	9			67
415		2	2			4
511	3	9	12	1		25
512			1		1	2
513	3	2	7	2		14
514	2	2	2			6
515			3			3
516	1					1
611	4		7			11
612	2			1		3
621				2		2
622		1	2			3
631	1	1	10			12
632	1		2	1		4
711	4	18	19		1	42
712	1	2	3			6
721	9	17	6		2	34
722	2	6	6	1		15
723		4	1			5
724	1	6	2			9
731	1	6	1			8
732	1	2	1			4
811	3	3		1		7
812	3	4	1			8
813			1			1
821	3	4	3			10
822	1	3	3	1		8
823		1				1
824		4				4
911	3	4	3			10
912		1	1			2
913		4				4
914		3				3
921		3				3
922		1				1
SUMA - Columns	**107**	**396**	**175**	**13**	**6**	**697**
Chi square	497,5001		Cramer's V	0,422426		

Table A-9: Cramer’s V (GT vs. DF 5. approach)

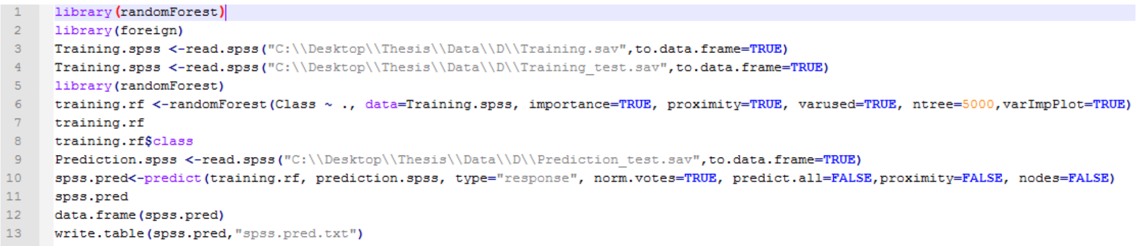
	Soil classes - BBN															
Soil classes - Gound Validation	**111**	**311**	**314**	**321**	**411**	**414**	**511**	**513**	**611**	**631**	**711**	**721**	**722**	**724**	**913**	**SUMA - Rows**
111	7	1			1				1		3					**13**
112		1									1					**2**
113		1							3	1						**5**
211					2											**2**
213					1											**1**
311	7	39			67	1		1		1	11					**127**
312		2			10	1		1			1	1				**16**
313											2					**2**
314	1	1	2		8						2					**14**
315	1	2														**3**
321		2		1	9											**12**
322					1											**1**
323		3			2											**5**
324					5											**5**
325					1											**1**
411	2	21			108					1	6			1		**139**
412		2			4											**6**
413											1					**1**
414	1	21			38	2					5					**67**
415		1			1				1		1					**4**
511	2	4			7		1	1	2	2	5	1				**25**
512								1			1					**2**
513		3			2			3	1	1	3		1			**14**
514	1	3			1						1					**6**
515		2									1					**3**
516		1														**1**
611	1	4						1	4		1					**11**
612		2	1													**3**
621								2								**2**
622					1			1			1					**3**
631		1					1	1	3	4	2					**12**
632					1		1		1	1						**4**
711		9			8	2			1		21	1				**42**
712		2			1			2			1					**6**
721	3	10			15	1					4	1				**34**
722		1			6	1				1	4		2			**15**
723	1				4											**5**
724		4		1	1									2	1	**9**
731		4			3	1										**8**
732		3						1								**4**
811		2			4			1								**7**
812		3			4			1								**8**
813								1								**1**
821		5			2	1		1			1					**10**
822		2			4					1	1					**8**
823					1											**1**
824					4											**4**
911		1			4			1	2		1				1	**10**
912					1						1					**2**
913					1										3	**4**
914						1									2	**3**
921		1			1										1	**3**
922		1														**1**
SUMA - Columns	**27**	**165**	**3**	**2**	**334**	**11**	**3**	**20**	**19**	**13**	**82**	**4**	**3**	**3**	**8**	**697**
Chi square	2355,296		Cramer's V	0,491295												

Figure A.1: ANN program



```
prediction<-read.table ("deployment_v4.txt",sep="",header=T)
Training.spss<-read.table("trainig_for_Netica6.txt",header=T)
Training.spss
library(nnet)
nn2<-nnet(Nat_leg2~., data=Training.spss,size=2,hidden=2,err.fct="ce",linear.output=FALSE)
nn2
prediction.spss <-read.table("deployment_v4.txt",sep="",header=T)
str(prediction.spss)
nrow(subset(prediction.spss,prediction.spss$GSM=!"GSM6"))
prediction.spss
nn2.predict<-predict(nn2,prediction.spss,type="class")
nn2.predict
data.frame(nn2.predict)
write.table(nn2.predict,"ann.result.txt")
```

Figure A.2: RF program



```
library(randomForest)
library(foreign)
Training.spss <-read.spss("C:\\Desktop\\Thesis\\Data\\D\\Training.sav",to.data.frame=TRUE)
Training.spss <-read.spss("C:\\Desktop\\Thesis\\Data\\D\\Training_test.sav",to.data.frame=TRUE)
library(randomForest)
training.rf <-randomForest(Class ~ ., data=Training.spss, importance=TRUE, proximity=TRUE, varused=TRUE, ntree=5000,varImpPlot=TRUE)
training.rf
training.rf$class
Prediction.spss <-read.spss("C:\\Desktop\\Thesis\\Data\\D\\Prediction_test.sav",to.data.frame=TRUE)
spss.pred<-predict(training.rf, prediction.spss, type="response", norm.votes=TRUE, predict.all=FALSE,proximity=FALSE, nodes=FALSE)
spss.pred
data.frame(spss.pred)
write.table(spss.pred,"spss.pred.txt")
```

Figure A.3: VB.net script

```
Option Explicit On
Imports Microsoft.Office.Interop
Imports Microsoft.VisualBasic.FileIO

Public Class FrmMain

    Private Sub Btn_Txtfile_Click(ByVal sender As System.Object, ByVal e As System.EventArgs) Handles Btn_Txtfile.Click
        Dim browsefile As String = ""
        browsefile = browse(True)
        If browsefile <> "" Then Txt_Txtfile.Text = browsefile
    End Sub
    Private Sub btn_subgrp_Click(ByVal sender As System.Object, ByVal e As System.EventArgs) Handles btn_subgrp.Click
        Dim browsefile As String = ""
        browsefile = browse(False)
        If browsefile <> "" Then txt_subgrps.Text = browsefile
    End Sub
    Function browse(ByVal TextFile As Boolean) As String
        If TextFile Then
            OpenFileDialog1.Filter = "Txt (*.txt)|*.txt"
        Else
            OpenFileDialog1.Filter = "excel (*.xlsx;*.xls)|*.xlsx;*.xls"
        End If
        OpenFileDialog1.FilterIndex = 1
        OpenFileDialog1.RestoreDirectory = False
        OpenFileDialog1.FileName = ""

        If OpenFileDialog1.ShowDialog() = System.Windows.Forms.DialogResult.OK
Then
            Return OpenFileDialog1.FileName
        Else
            Return ""
        End If

    End Function


    Private Sub Btn_Exit_Click(ByVal sender As System.Object, ByVal e As
System.EventArgs) Handles Btn_Exit.Click
        Me.Close()
    End Sub

    Private Sub Btn_OK_Click(ByVal sender As System.Object, ByVal e As
System.EventArgs) Handles Btn_OK.Click
        '***************************************
        ' validate user entry
        Dim Valid As Boolean = True

        ' check file names are valid
        If My.Computer.FileSystem.FileExists(Txt_Txtfile.Text) = False Then
            MsgBox("Please select a valid text file")
            Valid = False
        End If
        If My.Computer.FileSystem.FileExists(txt_subgrps.Text) = False Then
            MsgBox("Please select a valid excel spreadsheet")
            Valid = False
        End If


        ' **************************************
        ' if all OK process files
        If Valid = True Then
            Lookupsubgrps(Txt_Txtfile.Text, txt_subgrps.Text)
            Me.Close()
        End If
    End Sub

    Private Sub Lookupsubgrps(ByVal PredictedFile As String, ByVal SubgrpList As
String)
        Dim xlapp As Excel.Application = New Excel.Application
        Dim XlWb_subgrp As Excel.Workbook = xlapp.Workbooks.Open(SubgrpList)
        Dim xlws As Excel.Worksheet = XlWb_subgrp.Worksheets(1)

        ' Define subgroup dictionares
        Dim lookupgrp As Dictionary(Of String, String) = New Dictionary(Of
String, String)

        ' calculate number of columns
        Dim ColCount As Integer = 0
        For Each c As Excel.Range In xlws.Rows(1).cells
            If c.Value = "" Then
                Exit For
            Else
                ColCount += 1
            End If
        Next

        ' Load similar subgroups
        Dim jcolrev As Integer = (xlws.UsedRange.Columns.Count) - 1
        For jcol As Integer = 1 To xlws.UsedRange.Columns.Count
            Dim Strkey As String = jcol & "!" & jcolrev
            Dim StrValue As String = ""
            Dim irow As Integer = 2   ' ignore heading

            Do While Not xlws.Cells(irow, jcol).Value Is Nothing
                If StrValue = "" Then
                    StrValue = xlws.Cells(irow, jcol).Value.ToString
                Else
                    StrValue = StrValue & "!" & xlws.Cells(irow,
jcol).Value.ToString
                End If
                irow += 1
            Loop
            lookupgrp.Add(Strkey, StrValue)

            jcolrev -= 1
        Next


        '***********************************
        ' create output file
        Dim outfile = My.Computer.FileSystem.GetParentPath(SubgrpList) & "\" &
"Subgroup_scores.txt"
        My.Computer.FileSystem.WriteAllText(outfile, "X" & vbTab & "Y" & vbTab &
"Predicted" & vbTab & "Observed" & vbTab & "Score" & vbCrLf, _
                                            False, System.Text.Encoding.ASCII)

        ' read predicted/observed group
        Dim filename As String = PredictedFile
        Dim fields As String()
        Dim delimiter As String = vbTab
        Using parser As New TextFieldParser(filename)
            parser.SetDelimiters(delimiter)
            fields = parser.ReadFields() ' headings
            While Not parser.EndOfData
                Dim SCORE As Integer = 999

                ' Read in the fields for the current line
                fields = parser.ReadFields()
                ' Add code here to use data in fields variable.
                Dim predicted As String = fields(2)
                Dim Observed As String = fields(3)

                ' loop through dictionary finding predicted group
                ' Loop through the items based on key
                For Each predictedSCORE As String In lookupgrp.Keys
                    Dim sublist() As String =
Split(lookupgrp.Item(predictedSCORE), "!")

                    ' loop through all subgroups in list
                    For Each grp In sublist
                        Dim newSCORE As Integer = 0
                        ' if subgroup is predicted then find observed
                        If grp = predicted Then

                            ' for each predicted group loop through dictionary
for observed group
                            newSCORE = getscore(lookupgrp, predictedSCORE,
Observed, False)
                            If newSCORE < SCORE Then SCORE = newSCORE

                        End If
                    Next
                Next

                My.Computer.FileSystem.WriteAllText(outfile, fields(0) & vbTab &
fields(1) & vbTab & predicted & vbTab & Observed & vbTab _
                                                    & SCORE & vbCrLf, True,
System.Text.Encoding.ASCII)
            End While
        End Using

        GoTo tidyup

subend:
        MsgBox(Err.Description & " " & Err.Source & " " & Err.Erl)
tidyup:
        ' tidy up
        If Not XlWb_subgrp Is Nothing Then XlWb_subgrp.Close()
        xlapp.Quit()
        xlws = Nothing
        XlWb_subgrp = Nothing
        xlapp = Nothing
    End Sub

    Private Function getscore(ByVal lookup As Dictionary(Of String, String),
ByVal predictedSCORE As String, ByVal Observed As String, _
                                ByVal reverse As Boolean) As Integer
        Dim score As Integer = 999
        ' set predicted scores
        Dim PScores() As String = Split(predictedSCORE, "!")
        Dim predictedfwd As Integer = PScores(0)
        Dim predictedrev As Integer = PScores(1)

        For Each observedscore As String In lookup.Keys
            Dim OScores() As String = Split(observedscore, "!")
            Dim Observedfwd As Integer = OScores(0)
            Dim Observedrev As Integer = OScores(1)

            Dim newscore As Integer = 0
            Dim sublist() As String = Split(lookup.Item(observedscore), "!")
            For Each grp In sublist
                If grp = Observed Then

                    ' simple fwd/backward calculation
                    newscore = Math.Abs(predictedfwd - Observedfwd)
                    If newscore < score Then score = newscore

                    ' if predicted > Observed then need to loop forward
                    If predictedfwd > Observedfwd Then
                        newscore = Math.Abs(predictedrev + Observedfwd)
                        ' if predicted < Observed then need to loop backward
                    Else
                        newscore = Math.Abs(predictedfwd + Observedrev)
                    End If
                    If newscore < score Then score = newscore
                End If
            Next
        Next

        Return score
    End Function

End Class
```

Table A-10: National Soil legend

National_assoc_code	Association_Name	Texture_Substrate type	Ancillary1_code	Ancillary1_name	Ancillary2_code	Ancillary2_name	Ancillary3_code	Ancillary3_name	Ancillary4_code	Ancillary4_name	Ancillary5_code	Ancillary5_name	Comment	notes
112a	Rineanna	Loamy over lithoskeletal limestone	321	Ballincurra	411	Elton							includes Rineanna-Ballincurra complex	
112b	Crumpaun	Loamy over lithoskeletal limestone	314	Loughmuirran	711	unnamed (Clayey/shale)							wetter	
112c	Knockeyon_1	Loamy over lithoskeletal sandstone	632	Forth Commons									includes Slievereagh	
112d	Knockeyon_2	Loamy over lithoskeletal sandstone		Rock	911	Aughty							W Cork- hill and mountain complex rock is probably the dominant part of the association	

112d	Knockshigowna	Loamy over lithoskeletal slate and shale	311	Clonroche	513	Borrisoleigh							Borrisoleigh - Knockshigowna complex and	
113a	Carrigvahanagh	peat over lithoskeletal acid igneous rock		Rock	632	Blackstairs	722	Ballywilliam	911	Aughty			rock covers 50-80% of the area	
113b	Bantry	Peat over lithoskeletal sandstone and shale		Rock	911	Aughty							Hill and mountain complex D in West Cork	Bantry is new sereis name for the Schull and Ross Carbery rocky phases (W Cork)
211a	Seafield	Sandy stoneless drift	311	Dooyork	722	Ballyknockan							Dooyork series occurs along the edge of sand dunes. Some imperfectly and poorly drained soils (Ballyknockan) on landward side of dune sands (West Donegal)	
211b	Kilcolgan	Fine loamy drift with limestones	211	Kilcolgan bouldery phase	321	Kinvarra							drift over limestone rock; includes Kilcolgan bouldery phase	

213a	Burren	Fine loamy over limestone bedrock		Rock	111	unnamed (loamy/lithosklst)	321	Ballincurra	213	Kilcolgan			Include rocky phase with up to 25% bare rock; Co Clare has rocky phases; no Ballincurra at Leitrim	change to 213 as humose
213b	Burren rocky phase	limestone pavement	211	Burren									Also includes all Burren rocky phases	
311a	Clonroche	fine loamy drift with sillceous stones	311	Baunreagh Steep Phase	513	Borrisoleigh	712	Gortaclareen			711	Kilrush	Clonroche over drift; Baunreagh Steep Phase; 311f Ballynalacken -> 311a; 311j Cloverfield -> 311a; Baldswintown = Clonroche. Knockshigowna as minor component	easy rolling topography, Baunreagh Steep phase (Laois p.210 profile desc); Borrisoleigh, extensive in N Tipperary. Incudes Ballynalacken in Co Clare
311b	Kill_1	fine loamy drift with igneous & metamorphic stones	311	Kill lithic phase	313	Carrigogunnel							Derk shaley phase; Co Clare- Derk in felsitic drift	

311c	Clashmore_1	Coarse loamy drift with siliceous stones	312	Broadway	722	Puckane							rationalised with other Wick associations Old Ross, Knocknaskeha 1 and 2 and Kilfergus includes gleyic BE (Broadway) and 722 Puckane. Knockeyon as minor component	gently undulating
311d	Ballyvorheen	Sandy drift with siliceous stones	311	Clashmore	612	Portlaw							Related to Cooga series but keep separate for now; 311m->311d	
311e	Kells	Coarse loamy drift over hard shale	513	Rathkenny	711	Kilrush							Gley soils in valleys? maybe Street a minor component (small areas lumped during rationalisation). Also Ew on drumlins	also includes Keeloge in Carlow and Parknackle in Carlow (Kilrush)

311f	Ballylanders	Fine loamy over slate or shale bedrock	513	Cupidstown hill	313	Ridge	311	Clonroche					Largest areas (Clare 11.93kha) FLy over rock; Limerick CoLy over rock maybe?; Choose either FLy or CoLy and non-humose?	includes Ballindaggan in Carlow and Wexford and ridge in Laois and Hughstown in Kildare; Includes Ballynalacken in Co Clare
311g	Knocksquire	Coarse loamy over acid igneous bedrock		Rock										
311h	Borris	Coarse loamy drift with igneous and metamorphic stones	722	Ballywilliam										
311i	Broomhill	fine loamy drift over sandstone bedrock	311	Clashmore	1011	Monatray							some elements of man-made - reclaimed with beach sand particularly the Bannow	includes Wexford and Waterford Broomhill units
311j	Randallsmill	Coarse loamy stoneless drift												

311k	Broughillstown	coarse loamy over calcareous gravels	321	Baggotstown									different to Baggotstown as lead is typical BE rather than calc BE	
311l	Kennycourt	loamy drift with limestones	311	Clonroche	311	Ballylanders							described as a GBP but not enough clay increase. Most are freely drained soils on lst drift with some shallower shale bedrock soils probaly also some local shale drift	
311m	Kill_2	coarse loamy drift with igneous & metamorphic stones	721	Tramore	611	Ballyscanlon							includes Tramore (fragic) in Wexford	

311n	Clashmore_2	Coarse loamy drift with siliceous stones	711	Kilrush	411	Dungarvan							Dungarvan name has been used for the 411 subsis. Dungarvan unit in waterford is reclassified as Clashmore as it is a BE. Included Killadandaan now rationalised to Kilrush	
311o	Dovea	Fine silty drift with limestones	321	Ballincurra	712	Howardstown							Map code 249	
311p	**Dooyork**	**Sandy stoneless drift**											**Remove this association from national legend. Dooyork is an ancillary in 211a**	

311q	UN15 (cLy_Rk-LIM)	Coarse loamy over limestone bedrock	213	Burren	311	cLy over DR-LIM	213	Crush	513	Crossmolina			Predominantly Kinvarra series found in West Mayo (not to be confused with Kinvarra elsewhere in national legend) - coarse loamy RK-LIM	
313a	Ashgrove	Fine loamy drift with siliceous stones	811	Clohamon	711	Kilrush							Keep intact, though likely to be a different kind of map unit to rest of Ireland	
313b	Wonderhill	Fine loamy over lithoskeletal basic igneous rock	111	UN01 (Ly/basic igneous)	112	Carrigvahanagh							find names for shallow components	
313c	Schull	Coarse loamy drift with siliceous stones	1011	Ardmore (Schull plaggen)	712	Driminidy	511	Ross Carbery					Found in coastal areas on Western seaboard. Schull rocky phase now Bantry series	West Cork - includes Bantry complex A

313d	Bantry Bay	Coarse loamy dnse blue-grey drift with siliceous stones												
314a	Moord	fine loamy drift with sillceous stones	712	Gortaclareen	711	Kilrush							Moord in Waterford and Co Clare. Ambrosetown in Wexford now rationalised to Moord	Tullig rationalised to Moord in Co Clare
321a	Baggotstown	Coarse loamy over calcareous gravels	321	unnamed fine loamy/calc gravels	213	Crush	411	Patrickswell/ Elton					Badsey is FLy var of Baggotstown- modal profile = NTipp p95 - needs a name. This unit covers Baggotstown and Baggotstown-Crush units	Found on esker and moraine features. Crush found on the kame crests, Baggotstown elsewhere also includes some Patrickswell
321b	Ballincurra	Fine loamy over limestone bedrock	112	Rineanna	411	Elton	321	Kilfenora					refers to units described as Ballincurra	Kilfenora 2nd ancillary soil in Co Clare
321c	Kinvarra	Fine loamy over clayey drift with limestones	321	Kilfenora	313	Kilfergus							Co Clare	

411a	Patrickswell_1	loamy drift with limestones	321	Baggotstown	311	Ladestown	411	Elton					Incorporate bouldery phase (Tipperary)? Also included v minor Howardstown and Mylerstown minor components (Tipp)	well drained Patrickswell unit; redefined texture as loamy; horseheath anc1 now in lead series defn
411b	Patrickswell_2	loamy drift with limestones	411	Patrickswell lithic phase	321	Ballincurra							also include BE e.g. Baggotstown or ladestown. Also includes Mylerstown as minor component (tipp)	well drained Patrickswell with lithic phase
411c	Patrickswell_3	loamy drift with limestones	724	Mylerstown	411	Patrickswell lithic phase	922	Banagher	724	Ballyshear			wetter Patrickswell unit Mylerstown changed from 711 to 723. Howardstown also needed as a subsid and locally Coolalough (822)	Undulating 60-120m, 2deg slopes

411d	Kellistown	Coarse loamy drift with igneous and metamorphic stones	722	Puckane	722	Newtown							correlate Kellistown with Patrickswell ? Newtown becomes Kellistown as contain same units.	Clowater has a peaty top,Newtown is humose; both are SL-LS Kellistown undulating to rolling topography at 60-120m, Newtown found on concave sloes, flattish topography and local depressions, Clowater found in depressions.
411e	Mortarstown_1	fine loamy over clayey drift with limestones	414	Rathowen	411	Patrickswell							Text suggest stagnogleyic ; keep Offaly separate	
411f	Dunboyne	fine loamy drift with siliceous stones	711	Kilrush	414	Rathowen	311	Ladestown					Map says limestone till; text says Irish sea till (calc) intermixed with local limestone tills; Quat map says Shale enriched compact till of Irish Sea provenance.	411i-m -> 411f; 411n Clooncarine -> 411f; 411p Ballynamona -> 411f; 411e Graceswood->411f
411g	Mortarstown_2	Fine loamy over clayey drift with limestones	411	Rathowen	311	Kinvarra							Mortarstown -Kinvara complex	

411h	Athy	Coarse loamy over calcareous gravels	321	Baggotstown	723	Athy poorly drained component							poorly drained component needs new name- no other 723s that are relevant	
411x	Elton_1	Fine loamy drift with limestones	321	Baggotstown	711	Kilrush	822	Camoge	411	Patrickswell	213	Burren-Ballincurra complex	Howardstown as minor component	a few small areas of Burren-Ballincurra where limestone protrudes (Co Clare)
414a	Crosstown	Coarse loamy drift with siliceous stones	412	Crossabeg	412	Johnstown	411	Elton					includes 414A Fethard; Rathowen = Crosstown also includes RCP 414f Ballydoole in Limerick and Ballinbranagh in Carlow	414A -> 414a; 414e-> 414a
414b	Rathowen_1	Fine loamy drift with limestones	311	Ladestown	711	Kilrush							includes RCP 414a, c and d	
414c	Rathowen_2	Fine loamy drift with limestones	311	Ladestown	922	Banagher							wetter 414b with drained fen peat in hollows	Ladestown-Rathowen-Banagher
414d	Fethard	fine loamy drift with sillceous stones	721	Ballinruan									Ballinruan now rationalised to Kilpierce	

511a	Cooga_1	Coarse loamy drift with siliceous stones	511	UN06 (cLy_DR/GRN)	311	Clashmore	513	Knockboy	722	Puckane			includes RCP 511a and 511e and Clonnin complex (Offaly); includes Doonglara in Co Clare	
511b	Kiltealy	Sandy drift with igneous and metamorphic stones	512	Tomard (Cullion)	722	Ballywilliam	313	Carrigogunnel					now includes RCP 511b Kilnageer. Will also have brown earths associated with the Bpodz (313 Carrigogunnel - West Donegal)	
511c	Screen	Sandy stoneless drift	722	Ballyknockan	511	Carne								
511d	Cupidstown hill	Loamy over shale bedrock	112	Knockshigowna									equivalent to Knockastanna series mapped in Limerick and NTipp and includes Knockastanna-Knockshigowna complex	

511e	Ross Carbery	Coarse loamy drift with siliceous stones	511	Cooga	722	Puckane (Glassheenahielan)	811	Ilen					differnt from Cooga as PM is compact fine sand till in W Cork	Possible inclusion of 712 Drimidy
511f	Cooga_2	Coarse loamy drift with siliceous stones	632	Killinga	911	Aughty								
512a	Clonegall	coarse loamy drift with siliceous stones	722	Puckane									line work to be rationalised to include raheenleigh unit with clonegall. Includes cardtown Laois	
512b	Tomard	Loamy over slate or shale bedrock											perhaps add other ancillary soils	
513a	Knockboy	Coarse loamy drift with siliceous stones	513	Knockaceol	712	Puckane	311	Ballyvorheen					Merge with Cooga? Knockeyon is a minor component	
513b	Rathkenny	fine loamy drift with siliceous stones	311	Kells	511	UN31 (Ly/San)	712	Gortaclareen					Merge 513a and 513b	includes RCP 513b, 513f (Ballybrood)

513c	Borrisoleigh	Fine loamy over mudstone, shale or slate bedrock	311	Clonroche	112	Knockshigowna	612	Carrickbyrne	711	Kilrush			Includes steep phase and complexes that have these component soils. Includes Slievecoilta from Wexford and ridge from carlow and mountcollins limerick	includes Carrickbyrne association in Wexford
513d	Knockaceol	Coarse loamy over sandstone bedrock	112	Knockeyon	513	Knockboy	632	Forth Commons					Include peaty phase in Tipperary	
513e	Knockboy	Coarse loamy drift with siliceous stones	611	Ballycondon	711	Newport	722	Slieve Bloom					Waterford; Dodard correlated to Slieve Bloom	
514a	Corriga	fine loamy drift with siliceous stones	611	Meline	711	Kilrush							Leitrim unit; drift derived from greywackes, siltstone, mudstones and shales of Ballyhaise	

611a	Black Rock Mountain	Loamy over gneiss and schist bedrock	611	Stonepark	112	UN02 (Ly/GN&SC)		Rock					Stonepark is over mica schist drift (Leitrim); unnamed ranker/schist bedrock; includes	Black rock mountain in Wexford - in carlow BRM should be renamed as it is a 632/shale (see 632b)
611b	Slievebeag	Loamy over shale bedrock												
612a	Portlaw	Loamy drift with siliceous stones	513	Borrisoleigh	513	Knockboy	611	Ballycondon					Waterford	
621a	Ahuan	Loamy drift with siliceous stones	712	Gortaclareen	611	Drumslig							Waterford	
631a	Kiladoon	Loamy drift with siliceous stones	513	Knockboy	511	Cooga	911	Aughty						
632a	Blackstairs	Sandy over granite bedrock	112	Carrigvahanagh		Rock							Carlow and Wexford on blackstairs mountains	
632b	Knockastanna	Loamy over shale bedrock	511	Cupidstown hill	911	Aughty	112	Knockeyon	712	Gortaclareen			this is also the black rock mountain in carlow. Puckane (722) also a component	
632c	Forth Commons	Loamy over sandstone bedrock	911	Aughty		Rock							includes seefin	

632d	Monavullagh	Sandy over sandstone bedrock (conglomerate)	632	Glenary									Waterford. Knockalisheen is Sdy/drift w siliceous stones	
632e	Killinga	Loamy drift with siliceous stones (fine grained sstone till)	511	Ross Carbery	513	Knockboy	911	Aughty	511	Cooga			includes Rossmore from Laois p218-221. Also includes Reanascreena (W cork) reclaimed version of Killinga now classfied as Knockboy	
632f	Drumsleed	Sandy drift with siliceous stones	722	Puckane	911	Aughty	632	Killinga	113	UN				
711a	Macamore	Fine loamy over clayey drift with limestones	721	Kilpierce	711	UN09 (fLy over Cey/DR_SIL)							seen at vets house in Wexford- should this be undiff gley	
711b	Kilrush	Fine loamy drift with siliceous stones	712	Gortaclareen	922	Banagher	311	Ladestown	513	Borrisoleigh	911	Auchty	see notes for soils now classified as Kilrush.	
711c	Drumkeeran	Clayey drift with siliceous stones	712	Cluggin									small unit but review later	

711d	Straffan	Fine loamy drift with limestones	723	Mylerstown	411	Elton							no profile descriptions. Straffan 65% (+10%), 10% Mylerstown, Elton 10%	
711e	Clohernagh	Fine loamy drift with siliceous stones (fragic)	311	Clonroche									found only in Waterford (so far....) could be rationalised later with Kilrush	
711f	Newport	Coarse loamy drift with siliceous stones	722	Slieve Bloom	513	Knockboy							Waterford - wet unit	
711g	Tramore	Coarse loamy drift with igneous and metamorphic stones	311	Clashmore	311	Kill							Dense till of Tramore series is fragic (?), Bulk density of 1.77gcm^{-3} at 45cm ie SWG	
712a	Gortaclareen_1	Fine loamy drift with siliceous stones	722	Newtown?	822	Camoge	911	Allen	513	Knockboy				
712b	Cluggin	Clayey drift with siliceous stones	414	Crosstown										

712c	Howardstown_1	Clayey drift with limestones	724	Ballyshear	411	Patrickswell	922	Banagher					includes all Howardstown (inc 712 beta) and Sawyerstown (Kildare)	
712d	Gortaclareen_2	Fine loamy drift with siliceous stones	514	Corriga	711	Ballyhaise lithic phase							Ballyhaise-Corriga also includes drumlin complex (Bantry) in W Cork - Ballyhaise has been rationalised to Gortaclareen	
712e	Ballinamoor	Fine loamy drift with limestones	711	Straffan	911	Allen							includes 712alpha (Raheenduff) and RCP Ballinamore _1 and _2	
712f	Driminidy	Coarse loamy stoneless drift	911	Aughty	913	Turbury	511	Ross Carbery	913	Turbary	313	Schull	till is fine sand and very compact-seperate unit as different from drift with silicesous stones in other areas	includes Bantry complex B from West Cork

712g	Gortaclareen_3	Fine loamy drift with siliceous stones	414	Crosstown	312	Broadway								
721a	Kilpierce	Fine loamy drift with siliceous stones	721	UN10 (Ly over lithoskeletal SH & SL)	711	UN09 (fLy/Cey DR_SIL)	711	Kilrush	A	Undifferentiated alluvium			Wexford	
721b	Knockroe	coarse loamy drift with siliceous stones												
722a	Newtown	coarse loamy drift with igneous and metamorphic stones	722	Puckane	411	Kellistown	731	Greenane					** could be rationalised with Kellistown as the component soils are the same. associated with other coarse textured g/w gleys (Clowater peaty top) and better drained kellistown series	flattish topography and depressions; substrate type same as for Ballywilliam but BS of drift = 100%, but Ballywilliam drift is very acid (pH=5.0)

722b	Puckane_1	Coarse loamy drift with siliceous stones	721	Kilpierce	712	Gortaclareen	911	Allen					includes all Puckane units except association with Slievereagh and Raheenleigh and Oulartleigh	
722c	Puckane_2	Coarse loamy drift with siliceous stones	112	Knockeyon	311	Clashmore	513	Knockboy					Puckane-Slievereagh. Also Gortaclareen (712) needed	
722d	Slieve Bloom	Coarse loamy (upland) drift with siliceous stones	632	Knockastanna	511	Cooga	911	Aughty	722	Slieve Bloom undulating phase			includes Slieve Bloom steep phase; substrate type same as for Puckane but reserved pro tem for Uplands	includes Bawnrush
722e	Ballywilliam	Coarse loamy drift with igneous and metamorphic stones	712	Tramore	913	Aughty cut-over							occurs in valley bottoms on slopes of Blackstairs mountains	also includes Belmont, Ballinrush,Toberbride,Templeshanbo

722f	Puckane3	Coarse loamy drift with siliceous stones	712	Gortaclareen	311	Clashmore							occurs in waterford Puckane lead but with surface water gleys (permeable over impermeable substrate at >80)	unsure whether this should be an undifferentiated Gley??
723a	Mylerstown	Fine loamy drift with limestones	911	Allen	411	Patrickswell	724	Ballyshear					Ntipp WMeath are ZL/ZCL adjacent to cutover peat	see notes for rationalised soils
724a	Ballyshear	Fine loamy drift with limestones	724	Ballintemple	723	Mylerstown	411							
811a	Clohamon	Coarse loamy river alluvium	821	Lyre	812	Rearymore	821	Vicarstown					also profile in Laois p190; Vicarstown= Laois profile p188	also includes Liffey in Kildare
812a	Rearymore	Fine loamy river alluvium	812	UN12 (Zy RIV ALL)	811	UN12 (Zy RIV ALL)	821	Kilmannock var	811	Clohamon			From Alluviums in Offaly, Laois and Kildare to be reviewed! Kilmannock var is silty/river alluvium	use feale profile from Laois p182
813a	Milltownpass	Sandy stoneless drift	811	Clohamon										

820a	Finnery	alluvium											various soils typical & humic gleys + Pollagh soils (Laois)	
821x	River Burren	variable texture river alluvium											rationalise with Boyne?	
821a	Kilmannock	silty estuarine alluvium											also includes 'slob' in Waterford	
821b	Vicarstown	Clayey river alluvium	821	Feale	821	Kilmannock var								
821c	Feale	Fine loamy river alluvium	821	Boyne	821	Vicarstown								
821d	Kilgory	Sandy river alluvium	811	Aherlow										
821e	Boyne	Silty river alluvium	822	UN49_cLy_RIV-ALL	811	Clohamon	311	UN03 fLy_GRN					Boyne alluvium complex need to find soil names for 822 and 311 components	
821f	Lyre	Coarse loamy river alluvium	811	Clohamon	822	UN49								
822a	Coolalough	clayey lacustrine alluvium	824	Coolanick									Camoge-Miltownpass ; Clayey in Limerick; Fine loamy in WestMeath	Coolalough Fly NTipp profile = coolanick

822b	Millquarter	Fine silty lacustrine alluvium												
822c	Griston	Sandy lacustrine alluvium											very small extent only in Limerick, peaty top	
822d	Wexford slob	Silty marine alluvium												
822e	Shannon	Fine silty estuarine alluvium	821	Vicarstown	922	Banagher	824	Drombanny					Shannon-Banagher	
822f	Camoge	Clayey river alluvium	822	Coolfin	822	UN50 (cLy over calc GR)	813	Milltownpass					851 ancillary soils need completing	also Polloagh soils in Laois p192 wmeath profile p79 Miltownpass
822g	Coolfin	Fine silty river alluvium	811	Suir	921	Kilbarry	811	Finisk					waterford - complex association in alluvial positions. Finisk is probably too small to show and could be removed from the assoc.	
823a	Kilmore Slob	Sandy marine alluvium	823	Kilmore Slob varient									Variant is heavier texture	
824a	Drombanny	Carbonatic-loamy lake marl	823	Dunsany	212	Carney								
911a	Allen	peat	913	Turbary	922	Banagher	912	Garrynamona					Raw peat-raised bog in lowland (oligotrophic)	

911b	Aughty_1	peat	913	Aughty cutover	913	Turbary							Raw peat-blanket bog uplands (oligotrophic)	
911c	Knockmealdown	peat over rock	632	Glenary	113	Carrigvahanagh							Waterford peat over rock with podzol subsid	
911d	Aughty_2	peat	113	Bantry	632	Killinga	632	Drumsleed peaty phase		Rock			Mountain complex A, B and C and also includes units with Raheen (allen) + Hill & mountain complex C in West Cork	
921a	Pollardstown	peat	922	Banagher	913	Gortnamona								
913a	Turbary	peat	911	Allen	913	Gortnamona	912	Garrynamona	921	Pollardstown	913	Auchty Cut-over	Raw peat cut over (upland)	
913b	Aughty_Cutover	peat (cutover	113	Carrigvahanagh										
922a	Banagher_1	peat	913	Turbary	911	Allen							Earthy eutrophic lowland Fen peat	
922b	Banagher_2	peat & peaty alluvium	922	Ardrum									Ardrum association in Leitrim and Longford, includes peaty silty alluvium	

914a	Clonsast	peat	913	Turbary	913	Gortnamona	911	Allen					Industrial peat - milled and machined	
1011a	Monatray	loamy drift with siliceous stones	1012	Ardmore	1011	Curragh	1011	Schull plaggen					occurs in S Waterford 'plaggen' soils	