Farm-scale Digital Soil Mapping Techniques for Karaka and Patumahoe, South Auckland

April 2016

Technical Report 2016/013

David Palmer
Sharn Hainsworth

Landcare Research

Auckland Council
Technical Report 2016/013
ISSN 2230-4525 (Print)
ISSN 2230-4533 (Online)

ISBN 978-0-9941404-0-1 (Print)
ISBN 978-0-9941404-1-8 (PDF)

This report has been peer reviewed by the Peer Review Panel.

Review completed on 26 April 2016
Reviewed by two reviewers

Approved for Auckland Council publication by:

Name: Dr Lucy Baragwanath

Position: Manager, Research and Evaluation

Date: 26 April 2016

Recommended citation

Palmer, D and Hainsworth, S (2016). Farm-scale digital soil mapping techniques for Karaka and Patumahoe, south Auckland. Prepared by Landcare Research for Auckland Council. Auckland Council technical report, TR2016/013

Executive summary

Project and Client

Karaka and Patumahoe are rural areas bordering the Auckland city urban area and have a broad variety of land uses, including dairying, agriculture, lifestyle blocks, and amenities such as golf courses. Current soil maps of the Karaka and Patumahoe areas are at coarse resolution (1:50 000 map scale) and can be of limited use at farm management level. Auckland Council recently acquired Light Detection And Ranging (LiDAR) coverage of its region and engaged Landcare Research in 2014 to investigate methods for mapping soil classes at finer resolution farm-scale (1:10 000 – 1:5000 map scales), using digital data derived from LiDAR. Landcare Research has used these data to develop high resolution digital elevation models (DEM) and terrain attributes representing the Earth's surface. The outcome of this research is to determine the effectiveness of fine resolution (~5-m cell size) digital soil mapping and modelling (DSMM) to assess the extent to which the recurring pattern of soil continues across the landscape.

Objectives

- Using digital soil mapping techniques develop relationships between soil classes, terrain attributes, and radiometric covariate layers for farms located in Karaka Road and Gallagher Road.
- Undertake a validation exercise to determine the most appropriate digital soil model.
- Investigate and contrast maps derived from the digital soil models for the Karaka and Gallagher sites.
- Provide an indication of the area or distance over which the recurring pattern of soils is likely to continue, for future digital soil mapping extrapolation.

Methods

- Karaka dairy farm covers 62 ha of gently undulating landscape in close proximity to the Manukau Harbour. The farm rises from just above sea level along its northern stream boundary to around 35 m at its uppermost southern boundary. Soil survey was undertaken providing a total of 302 observations classified to the subgroup level of the New Zealand Soil Classification System. The Gallagher Road farm is at an average elevated of 53 m and covers an area of 118 ha. The farm occurs on the edge of the Pukekohe volcanic centre and is influenced by its proximity to these basaltic volcanoes.
- DEM derivatives – including elevation, slope, aspect, plan, profile, and total curvatures, up-slope contributing area, topographic wetness index, stream power index, sediment transport index, slope length, and landform elements – were developed using a 5-m cell size resolution DEM, derived from LiDAR data. Gamma radiometric data that included thorium, potassium, uranium, and total counts were also incorporated in the modelling.

- Geostatistical modelling was undertaken in the ‘R’ open source environmental using C5 decision trees, multinomial logistic regression (MNLR), and Random Forests modelling techniques. As a complement to these techniques, DSMART, a polygon disaggregation technique, was also investigated for potential map development.
- Environmental covariates in the form of DEM derivatives and gamma radiometrics were extracted for the 302 Karaka and 75 Gallagher surveyed sites and converted to a .csv format for modelling in R. Using the C5, MNLR, and Random Forests modelling techniques in R the relationships between terrain attributes, radiometrics, and Soil types were explored and developed. Cross-validation was undertaken with a 70% to 30% split for model and validation datasets to provide information around model accuracy. Validation from the DSMART model was undertaken using all soil observations and a confusion matrix.
- Final maps were developed from the digital models and explored using expert knowledge of the location (from pedologists) to provide understanding of the final maps. A reconnaissance of the Karaka area was undertaken to gain an indication of the distance over which the reoccurring pattern of soils is likely to continue beyond the surveyed location.

Results

- Statistics from the cross-validation of the Random Forests, MNLR and C5 decision tree modelling techniques were compared. Using the model dataset, the Random Forests model provides the best soil class predictions, followed by MNLR, and C5. More importantly, using the validation dataset, MNLR provided the best soil class predictions (31%), followed by Random Forests (30%), and C5 (25%). A combined Karaka and Gallagher model produced a model with a prediction accuracy of 39%. Interestingly, validation statistics for the DSMART model using the 302 sample observations indicate a prediction accuracy of 47%.
- Maps developed from the three models display a variety of results from a visual perspective. Using expert knowledge to ensure maps were pedologically plausible, the Random Forests model provides the best map. The C5 model has simplified the soil class predictions with only three soil classes represented. Conversely, the MNLR map provides good detail with all soil classes represented, but with some soil classes extending beyond their natural position in the landscape. The DSMART polygon disaggregation map provides a reasonable assessment of what was found in the field, but with some simplification taking place. Overall, smaller map units are not represented in the DSMART-derived map. The combined Karaka and Gallagher data produced a better map overall from a visual perspective, reducing the spatial extent of Organic soil classes predicted at lower elevations, but increasing the occurrence of Gley soil classes at the higher elevated Gallagher location.

Conclusions and future directions

- Fine-resolution detailed maps were successfully developed for the Karaka and Gallagher locations. The models of choice were Random Forests and the map polygon disaggregation technique DSMART. The future direction for this project is the filling of gaps in the Soil Regions, and the extrapolation of these models across the wider Karaka and Patumahoe areas and to test the efficacy of these maps with independent validation.

Table of contents

List of figures

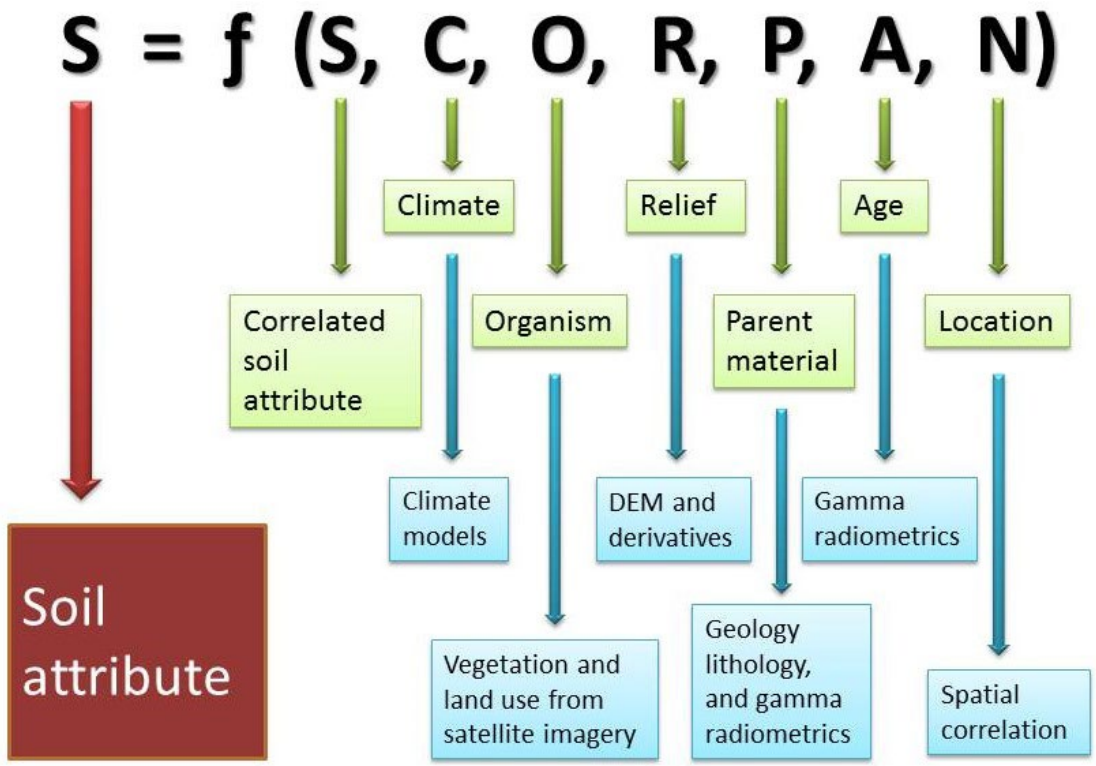
List of tables

1.0 Background

Historically in New Zealand, soil survey has been undertaken at map scales between 1:25 000 and 1:50 000, with S-map requiring a minimum map scale of 1:50 000. Recently there has been interest by land managers and regional authorities to develop maps at finer resolutions. Landcare Research uses data from the National Soil Database to develop pedotransfer functions (Minasny and McBratney 2002; Lilburne et al. 2014) that predict soil physical and chemical attributes. The S-map database (Lilburne et al. 2011) has been developed to deploy these pedotransfer functions relative to functional horizons (Webb 2003), soil families and their siblings. This approach logically assumes that soil functional horizons are highly correlated with soil physical and chemical properties.

The geostatistical component of digital soil mapping (DSM) provides methods for the development of models to produce maps at coarse or fine resolutions. Kempen et al. (2009) is an example of the updating of a 1:50 000 soil map using legacy soil data and a multinomial logistic regression approach. Grunwald (2009) puts forward the 'multi-criteria characterisation of recent digital soil mapping and modelling approaches', while Hastie et al. (2009) discusses some of the tools in data mining, inference, and machine learning for the prediction in statistical approaches in DSM. A recent development is the soil map polygon disaggregation technique DSMART (Holmes et al. 2014; Odgers et al. 2014). DSMART stands for the Disaggregation and Harmonisation of Soil Map Units Through Resampled Classification Trees. This is an iterative process where many realisations are generated using C4.5 (Quinlan 1993). Realisations are combined to provide a probability surface for each soil class using an existing soil map of known soil class composition (percentage of Soil types per polygon), and covariate layers representing SCORPAN factors (McBratney et al. 2003).

Jenny's (1941) equation of the soil-forming factors was intended as a mechanistic model for soil development where soil development is considered a function of climate, organisms (including humans), relief, parent material, and time. McBratney et al. (2003) extends this concept with SCORPAN where S is soil information from either an existing map, database, or from expert knowledge, C refers to climate, O to organisms (including human activity), R to relief (topography), P is parent material, A is age, and N refers to spatial position (Figure 1).



In a modern setting each factor can represent one or more continuous or categorical variables utilised in DSMM. An example of a geostatistical DSMM application is classification or decision trees (Lagacherie and Holmes 1997; Moran and Bui 2002; Bui and Moran 2003) where a tree structure is generated by partitioning the data recursively into a number of groups (Minasny and McBratney 2007). Overall, each division is chosen to maximise some error measure in the response variable of the resulting groups.

For many soil locations, as with this study, the soil parent material can strongly influence soil formation. Geological maps are a spatial source of information often used as a spatial covariate for DSM. The drawback of using geology is the coarse map scale at which these maps are developed and displayed for the Auckland Region. An alternative and supportive covariate for spatial modelling is radiometrics. Gamma-ray spectrometry, commonly termed radiometrics, is the measure of natural radiation in the Earth's surface (IAEA 2003). Radiometrics can provide insight into the top ~30–40 cm of the Earth's crust and can help to distinguish between certain soil and rock mineralogy.

Abundances of potassium (K), thorium (Th) and uranium (U) are measured by detecting the gamma rays produced during the natural radioactive decay of these elements, with the

number of gamma rays recorded in counts per second. Radiometric data are available for the Karaka area through the New Zealand Ministry of Economic Development.[1]

In this study, as in other studies, the ‘soil windows’ approach was adopted in this research to represent soil types of a region with the assumption that this pattern of soils and their relationships with soil-forming factors and SCORPAN continue across the landscape. If these relationships change or should a soil not be identified within the soil window, then our ability to predict soil types will be reduced. Indeed, if a soil class is not identified or missed, then it cannot be predicted within the modelling environment. It should also be remembered that the sampling of soil types is undertaken using a type of soil classification hierarchy. In our situation the New Zealand Soil Classification (NZSC) was used. Prediction of soil classes at the detailed end of the NZSC hierarchy (subgroup level) is often difficult because of the high variability associated with fine-scale mapping. It should also be remembered that fine-resolution covariates like terrain attributes have errors associated with them.

In this research we have used a variety of DSMM techniques – C5, MNLR, Random Forests, and DSMART – to explore the relationships between soil classes, terrain attributes, and radiometric covariate layers for the test sites in Karaka and Patumahoe.

2.0 Objectives

- Use digital soil mapping techniques develop relationships between soil classes, terrain attributes, and radiometric covariate layers for farms located in Karaka Road and Gallagher Road.
- Undertake a validation exercise to determine the most appropriate digital soil model.
- Investigate and contrast maps derived from the appropriate digital soil models for the Karaka and Gallagher sites.
- Provide an indication of the area or distance over which the recurring pattern of soils is likely to continue and use this to inform future DSM extrapolation.

3.0 Site Description

Karaka–Patumahoe is essentially a rural community that borders the Auckland city urban area to its north-east (Figure 2). To the south of Karaka stands the South Auckland Volcanic Field, which was active between c. 1.6 and 0.5 Ma (Lowe 2010). Pukekohe Hill, the youngest volcanic centre in this field underlain by basalt lavas, erupted about 0.56 Ma (Briggs et al. 1994; Edbrooke 2001). Distal tephras (Hamilton ash beds) also influence soil pedogenesis. In this area the Hamilton ash beds are between ~1.1 and ~3.5 m thick (Rae 1995), deposited incrementally millimetre-by-millimetre over the last c. 60,000 years (Lowe 2010), draped over older basalts of the region. A major characteristic of the Hamilton ash beds is their high clay contents, between 60% and 90%. The late Pliocene to early Pleistocene, non-marine sediments of the Puketoka Formation are typically 5–60 m thick in the Manukau area and interfinger with lava and tuff of the South Auckland Volcanic Field. Soil morphology throughout this region is also influenced by the underlying Waitemata Group. Parent materials are sandstones and mudstones that have been compacted, uplifted, folded and faulted, collectively called the Waitemata Group. Broadly, soils that have formed in parent materials of the Waitemata Group are from the Ultic Soil Order. The Soil types forming on these parent materials are more weathered because they exist within a landscape that has been stable for a long time. In contrast, where Hamilton ash is present, Granular and Allophanic soils dominate. Where the Hamilton ash formation is eroded exposing basalt, Brown soils tend to be predominant.

The location of the Karaka Road farm soil window is a few kilometres from the Manukau Harbour, rising to 35 m above sea level, with an average elevation of 13 m (Table 1). In contrast, the Gallagher Road farm soil window (at Patumahoe) is on average 53 m above sea level, ranging from 23 m at its lowest extent to 83 m elevation at its highest site. The two sites also have contrasting slopes, with the Gallagher farm nearly twice as steep, averaging 9.7 degrees compared with 4.3 degrees at the Karaka farm. Temperatures at both sites are similar, only increasing 0.2°C with increasing elevation. The Karaka site averages 1284 mm annual rainfall, compared with the higher Gallagher site with 1343 mm.

Although there are differences in climate variables, the temperature and rainfall differences remain relatively small and unlikely to be useful for modelling soil types. In contrast, the influences that topography and geology potentially have on soil-forming processes are likely to be strong. Therefore, we expect terrain attributes and radiometrics to play important roles in the development of DSMM for this region.

Table 1 **Karaka Road and Gallagher Road farms' general characteristics**

Property	Karaka			Gallagher		
	Min	Max	Mean	Min	Max	Mean
Elevation (m)[1]	2.8	34.6	12.8	23	82.7	53.4
Slope (°)[1]	0	25.7	4.3	0	33	9.7
Total annual rainfall (mm)[2]	-	-	1284	-	-	1343
Average annual temperature (°C)[2]	-	-	14.5	-	-	14.3

[1] Data derived from this research

[2] Data derived from long-term normalised climate data (Leathwick et al. 2002, 2003)

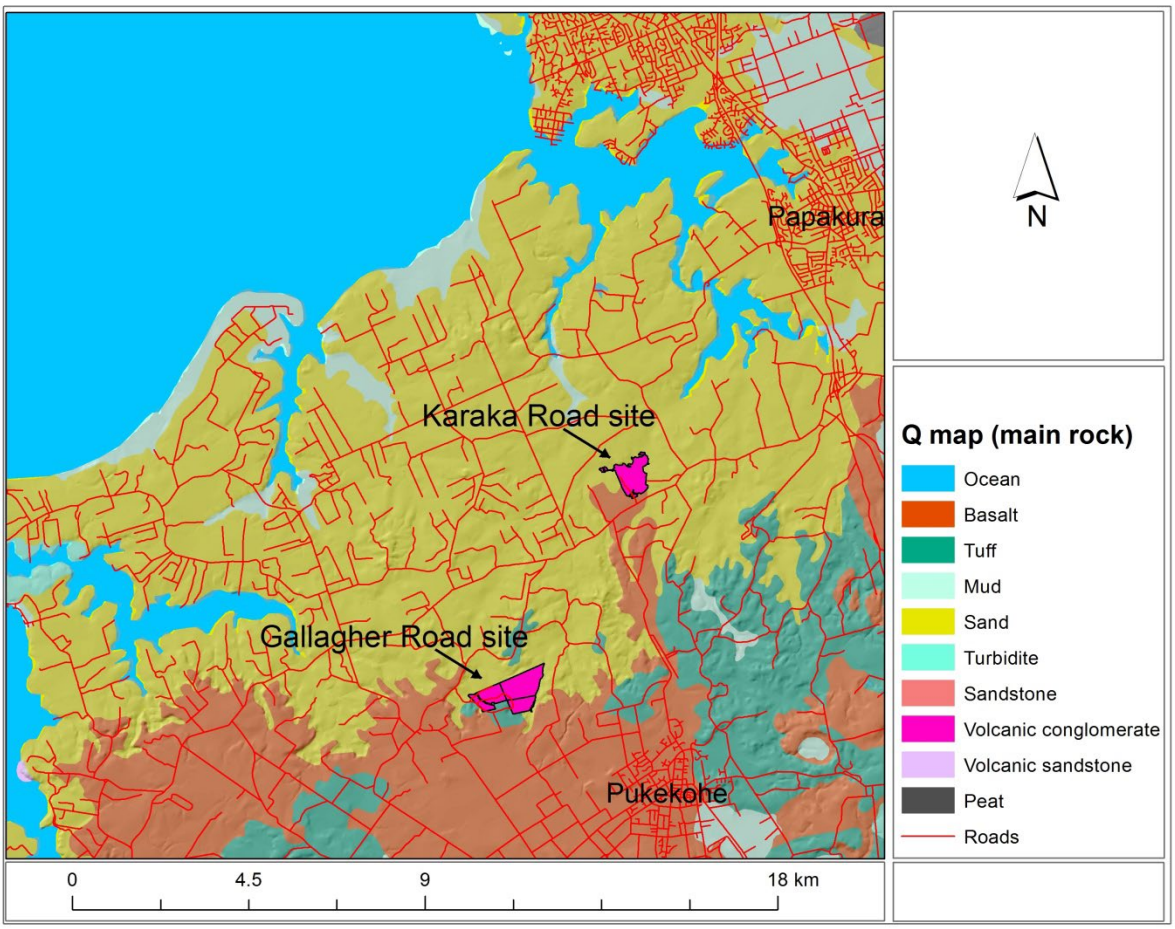


Figure 2 **Karaka and Gallagher farm locations over QMap geology with roads.**

4.0 Methods

4.1 Soil survey and soil observations

The 62-ha Karaka Road dairy farm (Figure 3) was sampled and surveyed by Landcare Research in May and June 2014. The 118-ha Gallagher Road farm (Figure 4) was surveyed by LandSystems in June and July 2014. At both locations, soil types were described using the New Zealand Soil Classification (NZSC) (Hewitt 2010), by digging soil pits and through soil auguring. Soil survey provided Karaka and Gallagher farms with 302 and 75 observations, respectively. A total of 19 soil subgroups were identified, falling within the Allophanic, Brown, Gley, Organic, Pumice, Raw, Recent, and Ultic soil types. All observation locations were geolocated using a Trimble S60 GPS set to the New Zealand Transverse Mercator (NZTM) projection.

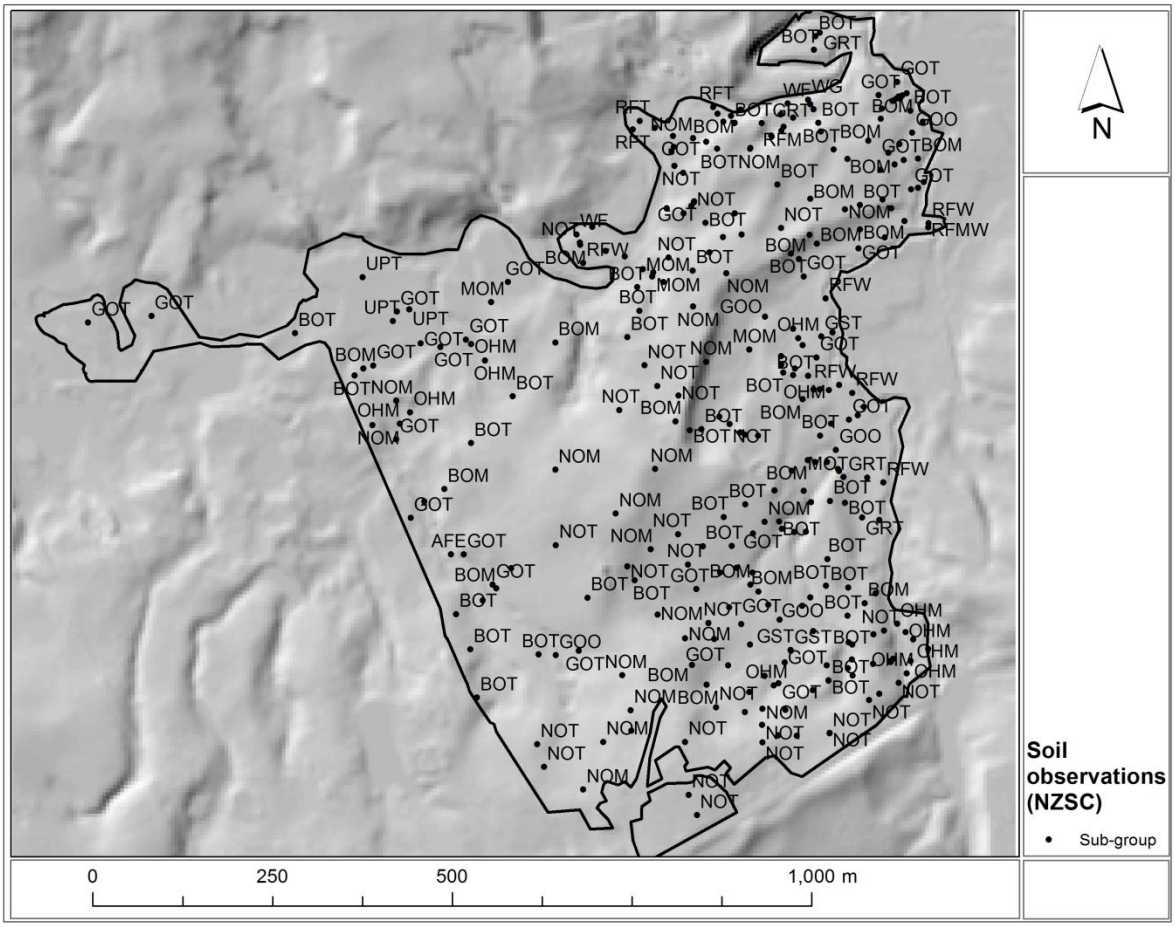


Figure 3 **Karaka Road dairy farm with 302 soil observations described to the subgroup level of the New Zealand Soil Classification. Black polygons are the farm boundary, with hillshade derived from the 5-m-cell-size-resolution digital elevation model to provide topographic relief.**

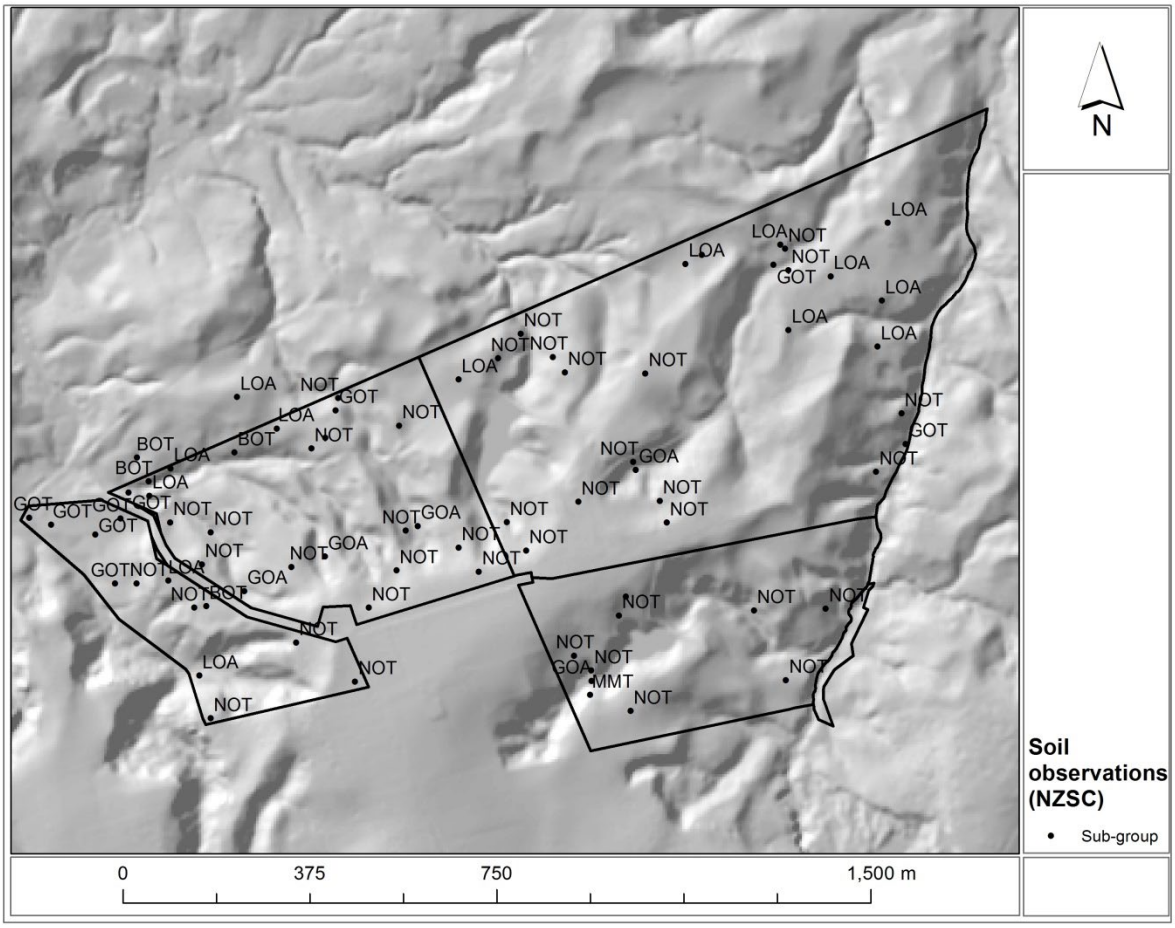


Figure 4 **Gallagher Road farm with 75 soil observations described to the subgroup level of the New Zealand Soil Classification. Black polygons are the farm boundary, with hillshade derived from the 5-m-cell-size-resolution digital elevation model to provide topographic relief.**

4.2 Soil covariate data collection and extraction

Terrain attributes including elevation, slope, aspect, plan, profile, and total curvatures, upslope contributing area, Topographic Wetness Index (TWI), Stream Power Index (SPI), Sediment Transport Index (STI), distance to stream, slope length, and landform elements were developed using a 5-m-cell-size-resolution DEM derived from LiDAR data provided by Auckland Council. Terrain attributes were developed using purpose-written Python scripts. For modelling details refer to Gallant and Wilson (1998, 2000), and Palmer et al. (2009). The ‘distance to stream’ terrain attribute layer was developed using TauDEM tools (Tarboton 2014), while landform elements (Schmidt and Hewitt 2004) were developed using a purpose-written Arc Macro Language (AML) script. Radiometric data were also used in the model development, which included thorium, potassium, uranium, and total counts. Radiometrics or gamma-ray spectrometry is described as a measure of natural radiation in the Earth’s surface (IAEA 2003) and provides information related to the top ~30–40 cm of the Earth’s crust and can help to distinguish between mineralogy. Potassium (K), thorium (Th) and uranium (U) values were provided through the New Zealand Ministry of Economic Development using data from Meyers. The ‘main rock’ class, from

QMap for the Auckland Region (Edbrooke 2001), was also used as a covariate layer in model investigation. All data were developed at a 5-m cell size resolution using the New Zealand Transverse Mercator (NZTM) projection. All covariates were extracted to the observation locations of soil samples (Section 5.1).

4.3 Statistical modelling and digital soil mapping (DSM)

DSM development was undertaken in the 'R' open-source environment. Models used included C5 decision trees, multinomial logistic regression (MNLR), and Random Forests™ (*randomForest* package: Liaw and Wiener 2002). DSMART, a map polygon disaggregation technique, was also used to investigate soil map development. C5 decision trees (*C50* package in R) fits classification tree models or rule-based models using the Quinlan (1993) C5.0 algorithm. C5 and C4.5 (depends on model version) are also implemented in DSMART developed by Odgers et al. (2014). MNLR ('multinom' function from the *nnet* package in R) not only enables the return of the most likely or probable prediction (class), but also the occurrence probabilities of the other soil classes considered. For discussions and applications to DSM refer to Kempen et al. (2009). Random Forests is an ensemble learning method for classification (and regression) that develops many decision trees during training that are later aggregated to give one single prediction for each observation in the dataset. For more information and discussion regarding Random Forests and DSM refer to Breiman (2001) and Grimm et al. (2008).

Initially, independent variables were assessed for correlation between covariates. Although, it is recognised that cross-validation methods used in the modelling techniques here, to some degree manage model overfitting due to strongly correlated covariates. A correlation coefficient matrix was used to identify covariate pairs with high correlations (>0.75). Highly correlated covariates were identified and avoided in the final models to avoid model overfitting. The C5, MNLR, and Random Forests modelling techniques in R were used to explore the relationships between terrain attributes, radiometrics, and soil types. Modelling techniques used a 70% to 30% split cross-validation between model and validation datasets, respectively. The 30% validation dataset was used to provide an indication of model prediction accuracy and precision from a confusion matrix.

For an overview of the final models, a confusion matrix was developed, representing the number of soils correctly classified, compared with the soils incorrectly classified, their numbers, and the class into which they were classified. The confusion matrix provides the model developer with an overview of not only correct classification, but also the number of observations involved. The confusion matrix can be undertaken for the model dataset, but more importantly, also for the validation dataset. Other validation statistics include overall accuracy, user's accuracy, producer's accuracy, and the kappa coefficient of agreement. For more details on validation of categorical prediction models refer to Congalton (1991).

Final maps were developed from the digital models and explored using the pedologists' (Sharn Hainsworth and Scott Fraser) expert knowledge of the location to provide understanding. A reconnaissance of the Karaka area was undertaken to assess the area or distance over which the recurring pattern of soils is likely to continue beyond the surveyed location.

4.4 DSMART modelling

The technique for disaggregating soil map polygons was also investigated for application across the Karaka–Patumahoe area. DSMART is the Disaggregation and Harmonisation of Soil Map Units Through Resampled Classification Trees (Holmes et al. 2014); modelling details are available in Odgers et al. (2014). DSMART uses the Python programming language to generate many realisations using classification trees implemented in C4.5 (Quinlan 1993) to predict soil classes as a function of input raster covariates. Within each polygon a fixed number of geographic coordinates are randomly selected and assigned soil classes according to the soil group proportions within the soil map polygon. These data were then modelled using C4.5 as a function of the co-related covariate raster layers. In this model we repeated the process 100 times providing 100 realisations of the potential soil group distribution. The soil class prediction probability was also calculated over the stack of realisations providing information on the first, second, and third highest predictions. In order to ensure that less common soils were represented in the final map, the number of samples per polygon was set to 15. Validation of outputs from the DSMART program was also undertaken using the Karaka and Gallagher farm soil observations, a confusion matrix, and validation statistics from the data.

5.0 Results

5.1 Model covariates

Initially modelling was undertaken for the Karaka and Gallagher farms separately. The final models and maps were developed using the Karaka and Gallagher datasets combined, providing not only improvements to validation statistics, but overall substantially improving maps based on expert knowledge. In the following sections we will focus on the final maps developed, and discuss the rationale for choosing these models and maps.

Through a process of elimination covariates for the combined Karaka and Gallagher observations were reduced to nine main covariates that relate to Soil types. The average importance plots (Figure 5) generated from the Random Forests model show that along with elevation, all the radiometrics (thorium (Th), Potassium (K), Uranium (U), total count), slope, distance to streams, Topographic Wetness Index (TWI), and landform elements were important in the modelling of soil types across these locations.

The most important model covariate used was elevation (Figure 6A, B). The overall elevation difference across the Karaka and Gallagher farms is less than 80 m. In general, elevations were much lower at the Karaka site compared with the higher elevations at the Gallager site. The second and third most important model covariates were the radiomatric layers thorium (Figure 6C, D) and potassium (Figure 6E, F). Overall thorium counts were lower at the Karaka farm, but higher at the Gallagher farm. Conversely, percentage of potassium counts were higher at Karaka, compared with Gallagher Road. Slope, distance to stream, and uranium were the forth, fifth, and sixth most important covariates (Figure 7), that retain a similar order of importance in the modelling context. Slope varies little across the Karaka farm (Figure 7A), compared with the Gallagher site (Figure 7B), where topographic relief was observed to be a more powerful driver of soil variability in the landscape. Slope across both sites delineates well the low-lying areas and upper terrace regions of these sites. Distance to streams (Figure 7C, D) identifies soils that are in close proximity to main water channels, and soils types that are further from fluvial processes and inputs. All of the radiometric data combined have the potential to delineate different parent materials, soil types, and mineralogy. Topographic Wetness Index (TWI), total radiometric counts, and landform elements are the final covariates used in the DSM process (Figure 8A–F). TWI can be influential because of its ability to delineate saturated versus non-saturated areas in the landscape. TWI is good for identifying low-lying areas in the landscape that are likely to be saturated for long periods (TWI > ~10), hence where reducing conditions will exist for much of the year and over time gleyed soil profile forms will develop. At some locations where wetting and drying processes are taking place, mottled soil types will occur, and at well-drained sites (low TWI values), gleyed and mottled soil profile forms are unlikely to occur. Total counts from the radiometrics add to the DSM (Figure 8C, D) by showing the spatial pattern of total radiometric counts. Landform elements (Figure 8E, F) provide insight into the position in the landscape in which soils occur. For, example the more stable sites on terraces, plateaus, and ridges tend to contain Granular or Allophanic Soils. Conversely, low-lying areas and channels tend to have soils developed under wetter conditions such as Gley Soils, and imperfectly drained soils. Although the modelling processes determined that landform elements are

less important in the models, this may be influenced by their larger number of classes, which may be leading to some redundancy.

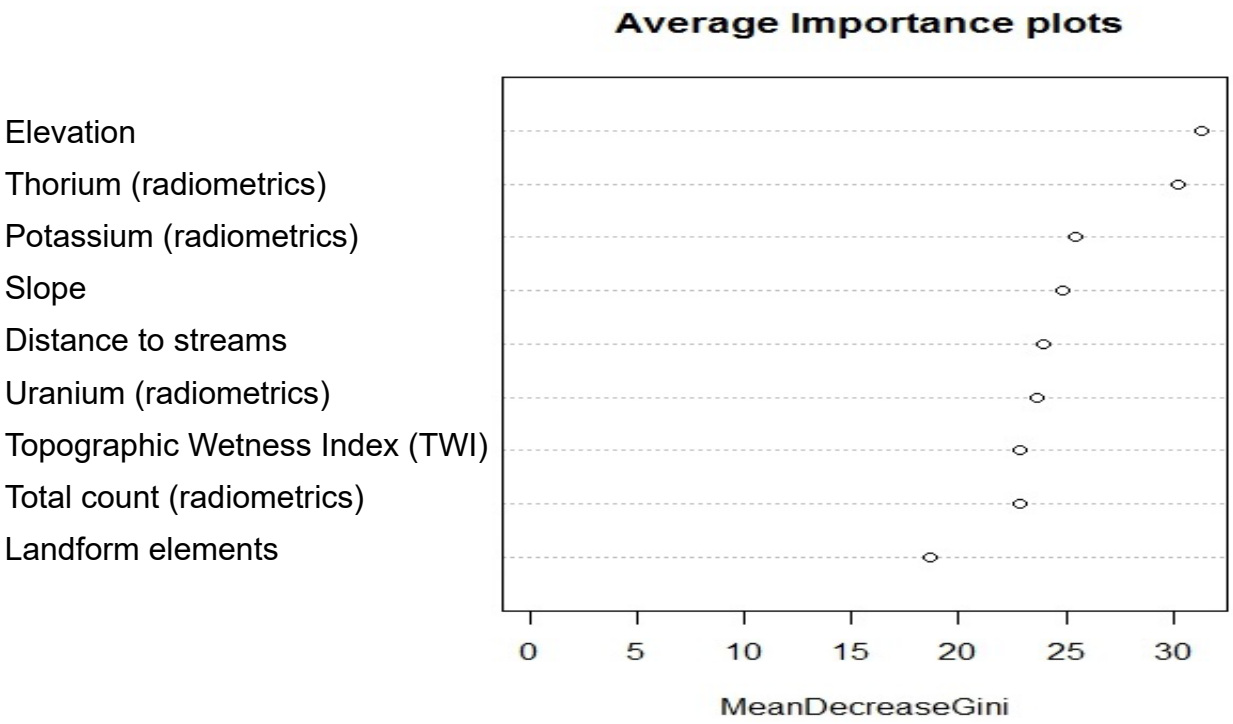


Figure 5 Average-importance plots showing the covariate importance in the final digital soil model.

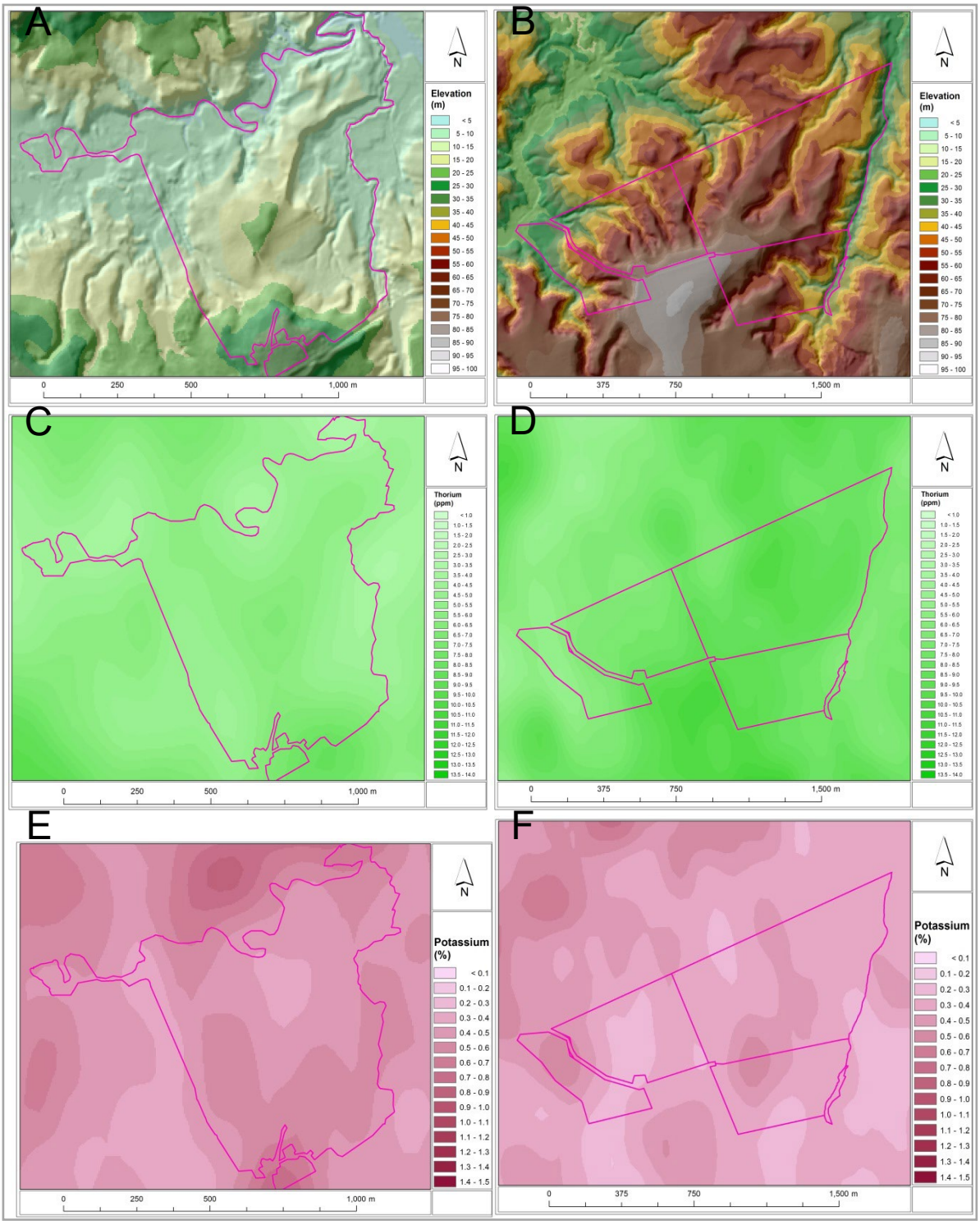


Figure 6 **Karaka (left), and Gallagher (right) farm locations illustrating (**A, B**) elevation, (**C, D**) thorium, (**E, F**) and potassium covariates used in the digital soil mapping and modelling process.**

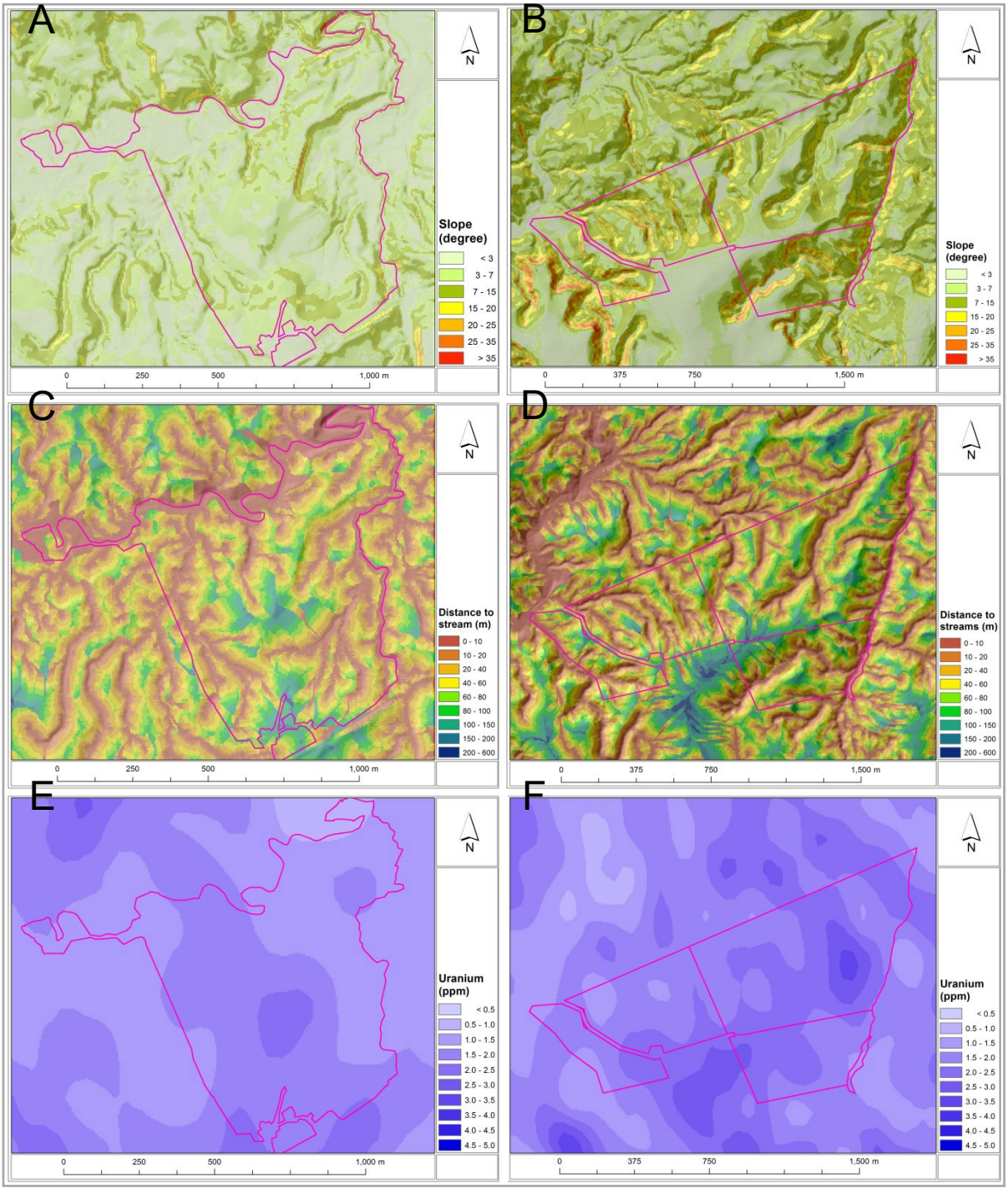


Figure 7 **Karaka (left), and Gallagher (right) farm locations illustrating (**A, B**) slope, (**C, D**) distance to stream, (**E, F**) and uranium covariates used in the digital soil mapping and modelling process.**

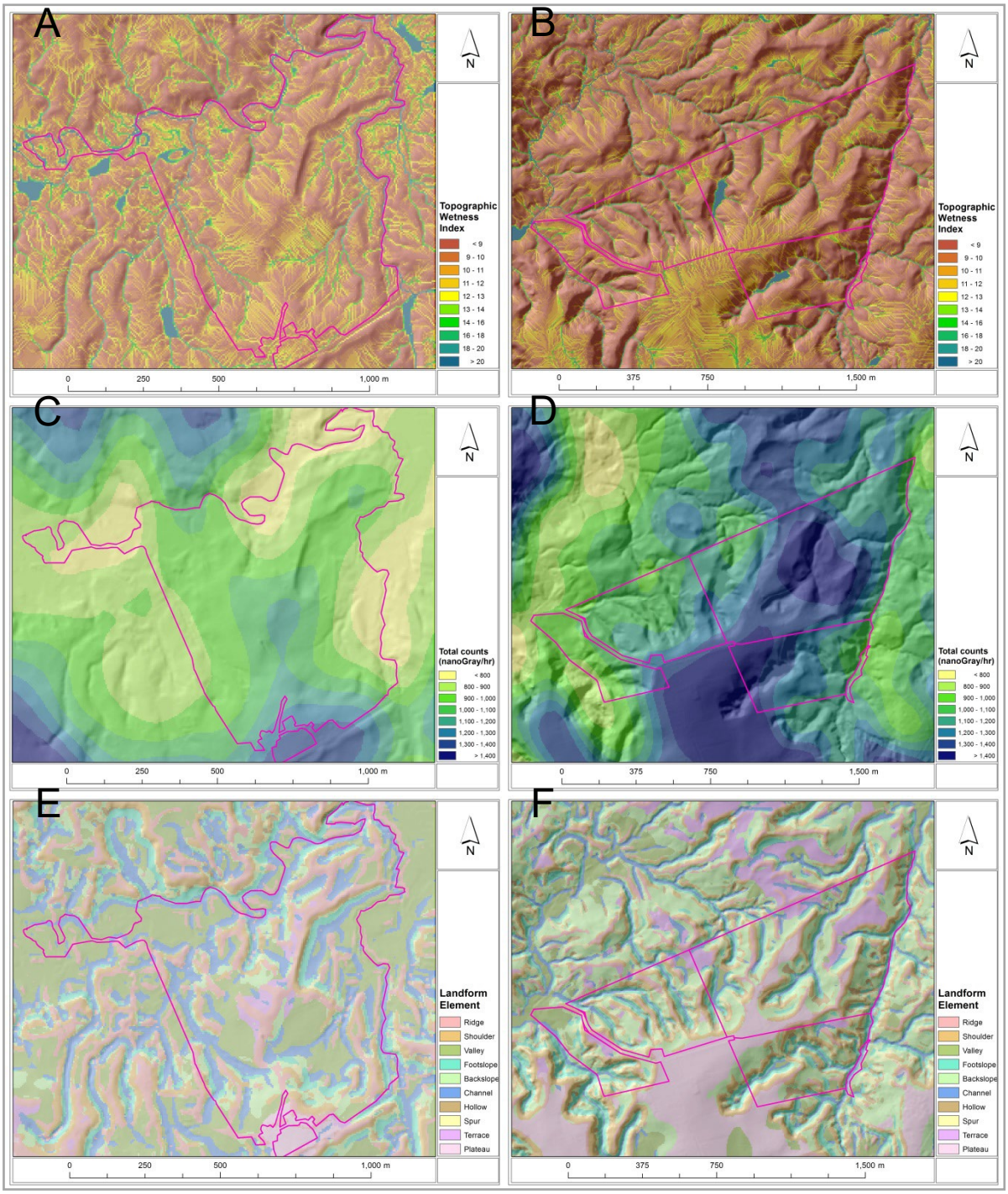


Figure 8 **Karaka (left), and Gallagher (right) farm locations illustrating (**A, B**) TWI, (**C, D**) gamma total counts, (**E, F**) and landform elements covariates used in the digital soil mapping and modelling process.**

Model and validation summary statistics are shown in Table 2. Model accuracy and kappa statistics for the Random Forests model show a perfect fit (typical for this modelling platform), whereas MNLR, and C5 decision trees have much lower model accuracy and kappa statistics. More importantly, validation summary statistics are similar ranging from

40% to 29% accuracy, and 0.29 to 0.14 kappa statistics for Random Forests, MNLR, and C5 models. Table 3 provides the interpretation for kappa statistics.

Table 2 **Summary statistics for the Karaka and Gallagher farms**

Statistic	Random Forests	Multinomial logistic regression	Decision trees C5	DSMART
Model accuracy (%)	100	56	42	-
Model kappa statistic	1.0	0.47	0.47	-
Validation accuracy (%)	40	39	29	47
Validation kappa statistic	0.29	0.28	0.14	0.35

Table 3 **Interpretation of the kappa statistic**

Kappa coefficient	Interpretation
<0.01	Less than chance agreement
0.01–0.20	Slight agreement
0.21–0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–0.99	Almost perfect agreement

Tables 4 and 5 show the model and validation confusion matrix from the Random Forests model for the Karaka and Gallagher farms (combined model). A perfect model or prediction would have all correct soil predictions in the diagonal green boxes. Misclassifications are found in other areas of the matrix. Random Forests model predictions are perfect; however, in the validation matrix, misclassifications do occur. In general Typic Orthic Brown Soils (BOT), Typic Orthic Gley Soils (GOT), and Typic Orthic Granular Soils (NOT) are the most common soil types represented in the validation matrix. There are misclassifications for most classes, but for the majority of the time, misclassified soils were taxonomically similar, or occur at similar positions in the landscape to the correctly-classified soil types.

Table 4 **Confusion matrix for the combined Karaka and Gallagher farms Random Forests model data**

	BOM	BOT	GOA	GOO	GOT	GRT	GST	LOA	MOM	NOM	NOT	OHM	RFM	RFMW	RFT	RFW	UPT	WF	WG
BOM	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BOT	0	45	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GOA	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GOO	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GOT	0	0	0	0	43	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GRT	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
GST	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0
LOA	0	0	0	0	0	0	0	13	0	0	0	0	0	0	0	0	0	0	0
MOM	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0
NOM	0	0	0	0	0	0	0	0	0	18	0	0	0	0	0	0	0	0	0
NOT	0	0	0	0	0	0	0	0	0	0	69	0	0	0	0	0	0	0	0
OHM	0	0	0	0	0	0	0	0	0	0	0	12	0	0	0	0	0	0	0
RFM	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
RFMW	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
RFT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0
RFW	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0
UPT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0
WF	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0
WG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Key to New Zealand Soil Classification soil types in Tables 4 and 5:

BOM = Mottled Orthic Brown; BOT = Typic Orthic Brown; GOA = Acidic Orthic Gley; GOO = Peaty Orthic Gley; GOT = Typic Orthic Gley; GRT = Typic Recent Gley; GST = Typic Sandy Gley; LOA = Acid Orthic Allophanic; MOM = Mottled Orthic Pumice; NOM = Mottled Orthic Granular; NOT = Typic Orthic Granular; OHM = Mellow Humic Organic; RFM = Mottled Fluvial Recent; RFMW = Mottled-weathered Fluvial Recent; RFT = Typic Fluvial Recent; RFW= Weathered Fluvial Recent; UPT = Typic Perch-gley Ultic; WF = Fluvial Raw; WG = Gley Raw.

Table 5 **Confusion matrix for the combined Karaka and Gallagher farms Random Forests validation data**

	BOM	BOT	GOA	GOO	GOT	GRT	GST	LOA	MOM	NOM	NOT	OHM	RFM	RFMW	RFT	RFW	UPT	WF	WG
BOM	0	3	0	1	2	0	0	0	0	0	0	0	1	1	0	0	0	0	0
BOT	3	11	0	0	1	0	0	0	0	1	3	0	0	0	0	0	0	1	0
GOA	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GOO	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GOT	4	4	0	0	8	0	0	1	0	2	2	3	0	1	2	3	1	0	0
GRT	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
GST	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LOA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MOM	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
NOM	0	1	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
NOT	1	5	1	0	2	0	0	1	0	7	19	0	0	0	0	0	0	0	0
OHM	0	0	0	2	1	0	0	0	0	0	0	2	0	0	0	0	0	0	0
RFM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RFMW	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
RFT	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0
RFW	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
UPT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
WF	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
WG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Key to New Zealand Soil Classification soil types in Tables 4 and 5:

BOM = Mottled Orthic Brown; BOT = Typic Orthic Brown; GOA = Acidic Orthic Gley; GOO = Peaty Orthic Gley; GOT = Typic Orthic Gley; GRT = Typic Recent Gley; GST = Typic Sandy Gley; LOA = Acid Orthic Allophanic; MOM = Mottled Orthic Pumice; NOM = Mottled Orthic Granular; NOT = Typic Orthic Granular; OHM = Mellow Humic Organic; RFM = Mottled Fluvial Recent; RFMW = Mottled-weathered Fluvial Recent; RFT = Typic Fluvial Recent; RFW= Weathered Fluvial Recent; UPT = Typic Perch-gley Ultic; WF = Fluvial Raw; WG = Gley Raw.

The final Karaka Random Forests map can be seen in Figure 9A. In general, Typic Orthic Granular Soils and Mottled Orthic Granular Soils (NOT and NOM, respectively) occur on the upper terraces and ridges of the landscape, whereas Gley Soils occur on the lower-lying flat areas of the landscape. The map also predicts Mellow Humic Organic Soils

(OHM) and Recent and Raw soils associated with the floodplains. Figure 9B provides a probability surface illustrating across a continuum the likeliness of each cell representing a Soil types being incorrectly (0) to correctly (1) assigned. The soil types with better prediction probabilities tend to occur in the upper landscape, compared with the poorer predictions occurring around the lower channel margins.

Figure 9C shows the Gallagher farm soil types (subgroup level of the NZSC). Overall, the soil types are predicted in the same position in the landscape as the Karaka farm; however, the main differences are that Acid Orthic Allophanic Soils (LOA) are occurring at the upper, probably more stable parts of the landscape (not found at the Karaka farm site). Acidic Orthic Gley Soils (GOA) occur along stream channels higher in the landscape (above the floodplains). Figure 9D provides a probability surface that shows the likeliness of each cell representing a Soil types being correctly assigned. Overall the Gallagher farm site has higher prediction probabilities compared with the Karaka farm site, with the majority of the Gallagher farm having probabilities above 0.5.

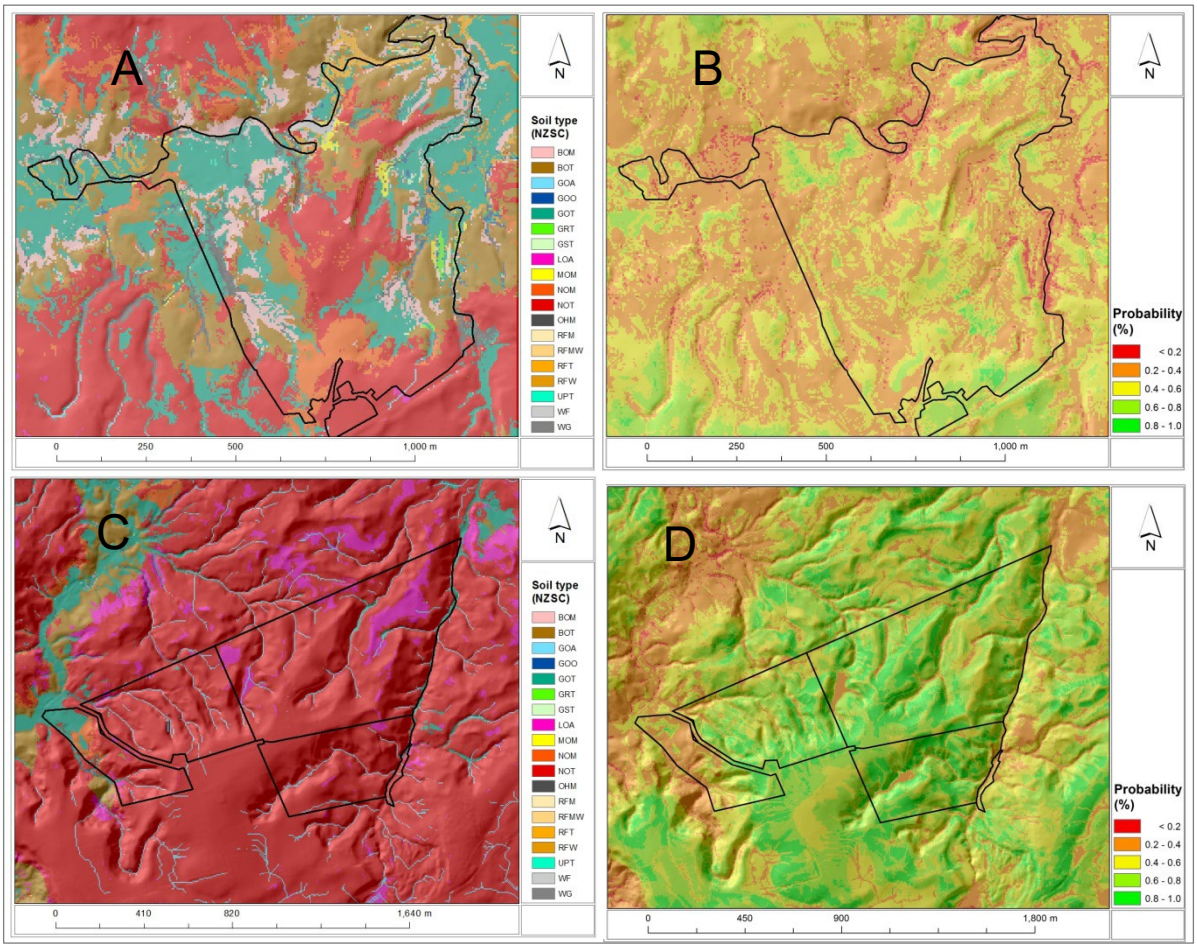


Figure 9 **Random Forests digital soil map of** A **Karaka farm (NZSC subgroup),** B **Karaka farm Soil types prediction probability,** C **Gallagher Road farm (NZSC subgroup), and** D **Gallagher Road farm Soil types prediction probability. NZSC = New Zealand Soil Classification.**

The Karaka DSMART map can be seen in Figure 10A. Overall, the map has a similar pattern to the Random Forests model with soils occurring in the correct position in the landscape. For example, OHM, RFT, WF, and WG soils are associated with the floodplains, compared with the NOT, NOM, BOT, and BOM soils that occur at the higher positions in the landscape. However, Brown Soils are generally over-represented at the

Karaka farm site when compared with the Random Forests model (Figure 9A). Figure 10C shows the Gallagher Road farm has a possible over-representation of LOA soils occurring at the upper parts of the landscape (not found at the Karaka farm site). GOA soils are also potentially over-represented along the stream and channel margins higher in the landscape. Figure 10B and D provide probability surfaces illustrating the likeliness of each cell representing a Soil types. Probability values range from <0.2 (poor prediction), and 0.8 to 1.0 (excellent prediction). Overall the Gallagher farm site has higher prediction probabilities compared with the Karaka farm site, with the majority of the Gallagher farm having probabilities above 0.5.

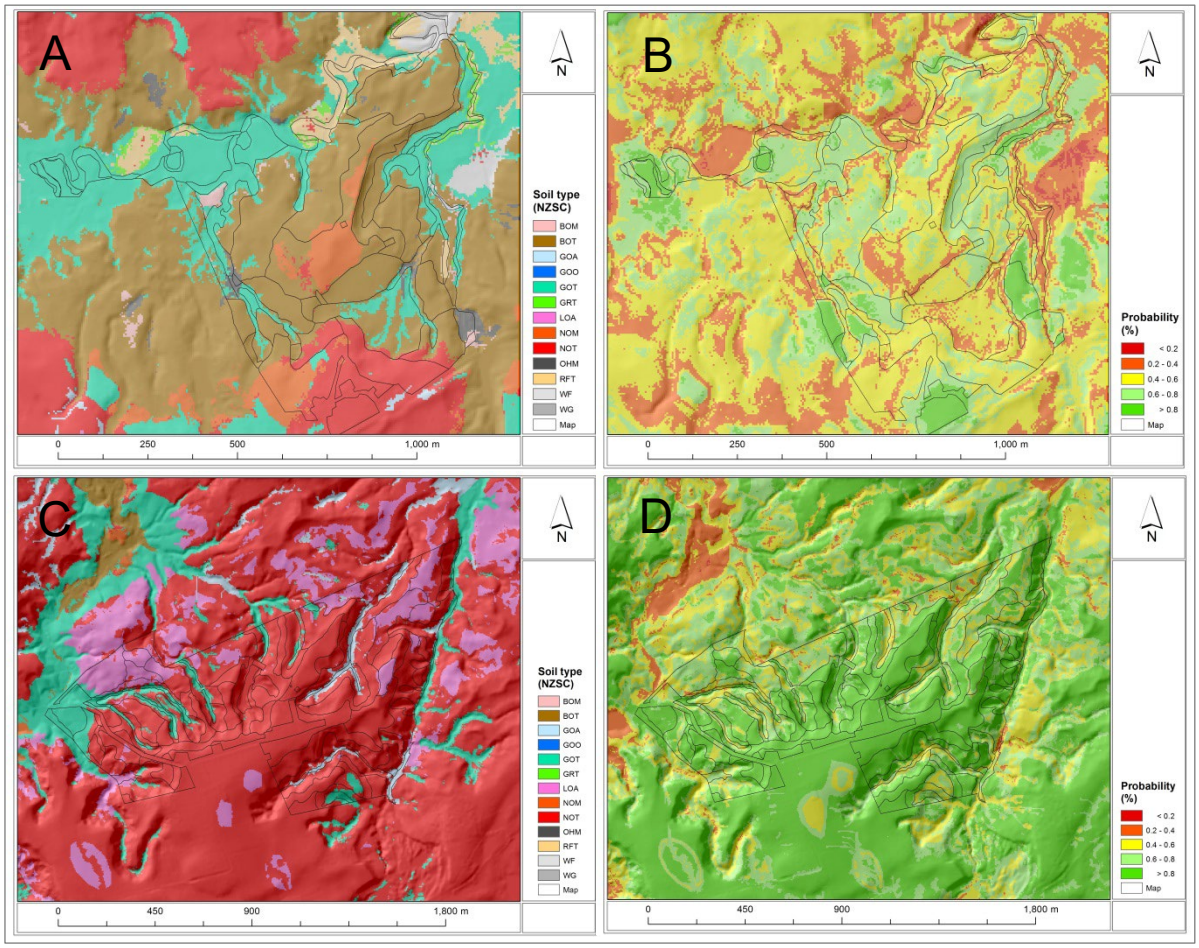


Figure 10 **DSMART digital soil map of** A **Karaka farm (NZSC subgroup),** B **Karaka farm soil-type probability,** C **Gallagher Road farm (NZSC subgroup), and** D **Gallagher Road farm soil-type prediction probability. NZSC = New Zealand Soil Classification.**

6.0 Discussion

6.1 Farm-scale model and map overview

Overall the Random Forests model (Figure 9) provided the best results of farm-scale maps for the Karaka and Gallagher farms. From an expert knowledge and pedological perspective soils occur in the correct position in the landscape and provide a good representation of reality. MNLR and C5 decision trees produced validation statistics that were slightly reduced when compared with the Random Forests model. However, from a qualitative review, the final maps from MNLR and C5 models have some concerns. Although the C5-decision-tree validation statistics were similar to the other models, only four soil types were represented and mapped, compared with 19 soil types observed in the field. All of the soil types in the MNLR map were predicted, but there were unusual patterns occurring, with some floodplain soils occurring in the upper parts of the landscape. For example, recent soil types only found on the lower-lying floodplains were extrapolated by the MNLR model to higher positions in the landscape (hills and terraces). If only considering soil types within the original observation window boundary, then soil predictions are considered reasonable from the investigating pedologist's perspective (expert knowledge).

The DSMART modelling technique results show a similar pattern of soil types on the Karaka and Gallagher farm as produced using Random Forests. Overall, soil types occur in the correct positions in the landscape, but the areas representing Brown and Allophanic soil types are substantially different to the Random Forests model results. From a validation perspective, the DSMART technique has the best prediction statistics (47%; Table 2). This validation statistic assumes independence, but in reality the surveyed observations were used to some degree by pedologists to develop map units, therefore incorporating bias to some extent. The fall-back position is that observations are used for model statistics. As a model statistic the DSMART predictions are ranked third out of the four models. From an expert-knowledge visual assessment the DSMART map provides a reasonable expectation of the types of soils occurring across the two farms, not foregoing the previously discussed concerns around soil-class area representation. Overall, the validation statistics and the visual assessment have guided us to selecting the map produced by the Random Forests model as the preferred map.

When considering the validation matrix for the Random Forests model (Table 5), the number of soils occurring in the diagonal column provides the number of soil classes correctly predicted. This statistic can only be true or false for a given class. Information that is every bit as important is the soil class given (predicted) when the prediction is incorrect. For example, the validation matrix shows that 19 of the NOT soil observations were correctly predicted, whereas three of the observations were predicted as BOT soils, and three observations as GOT soils, while one was classed as an RFT soil. Brown and Gley soils occur in the same position in the landscape as the NOT soils. Because of this overlapping environmental space, it is difficult for a model to separate the soil types using a limited covariate space. Conversely, the RFT soils do not occur in this position in the landscape, but are found on the floodplains, and therefore can be considered an incorrect prediction. This illustrates how validation statistics only provide binary information (correct versus incorrect), but in the real world taxonomically-similar soil types and soil types that occur in the same covariate space should be recognised. Another consideration is the

number of observations available for cross-validation statistics. Where observation numbers are low, outliers can substantially bias results. It should also be remembered that these are cross-validation statistics that are not truly independent of the modelling process. To be independent requires sampling for new observations that are never used in the modelling process that can be tested against model predictions.

While the authors are confident with the mapped results, it is pertinent to note and discuss the influence of model input covariates and the observation data available for the modelling process. In the initial stages of modelling, a variety of different covariates were tested, modelled, and mapped to choose the most appropriate covariates to predict soil types and to determine the best model and map. Other protocols to consider are the number of observations required to represent each soil types, and provide meaningful model and validation statistics. Above all else we need to ensure that all soil types have been observed and captured for the modelling process, because you cannot model a class that does not exist. Figures 5–8 illustrate the covariates used in the final model. The nine covariates used in the Random Forests modelling process were selected because the addition of further covariates provides no substantial improvement to model validation. If covariates are added or removed, the maps change the areas representing soil types, with the concomitant expansion or shrinkage of a class or classes.

The Soil Region also has an influence on model outcomes. A Soil Region is defined as a recurring pattern of soils under similar soil-forming factors, with these factors being represented by covariate layers. Modelling beyond the Soil Region will lead to a breakdown of the soil to environmental (soil-forming covariate) relationships. Although the final model combined observations from both the Karaka and Gallagher farm surveys, when modelled separately some of the mapped soil-type areas change. For example, the presence and extent of Allophanic Soil is noticeably different using (1) data only from the Gallagher farm, and (2) data combined from the Karaka and Gallagher farms (using the same input covariates). Also noted was the presence of Organic, Brown, and Recent soils in the valley bottoms and channels of the Gallagher farm when using the combined model. This does not mean these soil types do not occur at the Gallagher farm, only that they were not observed in the Gallagher soil survey.

From a stakeholder's perspective, soils in the lower parts of the landscape may be currently considered of less consequence compared with the Allophanic Soils in the higher, more stable parts of the landscape. Allophanic Soils have a greater ability to manage effluent and treat wastes. From a scientist's and environmentalist's perspective, however, the soils of the lower parts of the landscape may be every bit as important because of their potential vulnerability to leaching.

6.2 Future directions

Future directions must consider what Karaka and Pukekohe areas are of priority to be mapped at the farm scale. Discussions with the Franklin Local Board inform us that priority should be given to the intensively-farmed Pukekohe horticultural area. If we use the Soil Region concept to determine the extent to which the soil models apply, and to estimate how many representative areas (windows) will need to be sampled (like the Karaka and Gallagher Road farm sites) then we need to develop windows representative of the recurring pattern of soils across the region. Figure 2 is a GNS-developed map called QMap representing the main rock group for the Pukekohe and Karaka areas. QMap shows a clear delineation between the older basalts around Pukekohe (southern area), and the

sands and muds of the Karaka locality suggesting, from a geological perspective, that parent materials are likely to be different between these areas. Figure 11A and B show a more detailed picture of the likely recurring pattern of soil types across the Pukekohe and Karaka areas resulting from the use of radiometrics. The radiometric-based colour composite map (ternary) in Figure 11A delineates low-lying areas as red colourings (regions A), a western area with darker green and blue colours (region B), elevated coastal terraces as green, turquoise, and blue (regions C), a region closer to the Manukau Harbour with greens, yellows, and blues (region D), and an elevated area to the south shown as yellow, green, and blue colourings (region E). The darker areas on the map (region F) probably indicate areas with high water content or highly vegetated areas, and the urban areas are shown as stronger yellow colourings. The colours shown in this map (Figure 11A) are accentuated because the data has been clipped to this regional extent, which truncates the composite values in relation to the data mean when displayed. In Figure 11B the colours are not as accentuated and the pattern is more difficult to decipher; however, the general pattern remains. We should caution that low-lying areas can be prone to pockets of radon gas collecting in valleys, as well as to variations in soil moisture content, that can influence map colours. Therefore, map colours are used here to describe general patterns relative to geological units. These concerns highlight the need to undertake geochemical surveying in relation to gamma radiometrics in order to fully understand the implications for mapping soils digitally.

Overall, the radiometric data suggests a minimum of five soil windows representing Soil Regions of the Pukekohe and Karaka area. A Soil Region could potentially be a series of transects covering a greater area than an intensively-sampled small soil window. It would be prudent to state that radiometrics is a relatively new technology to New Zealand DSM. The airborne radiometric survey was flown at 200 m altitude with a mean terrain clearance of 60 m. As a result radiometric surfaces were developed at a coarse 50-m cell size resolution compared with our terrain attributes at 5 m. At this stage we think that of all the covariate layers we have investigated, radiometrics seems the most likely to assist us in delineating Soil Regions and providing an indication of parent material mineralogy.

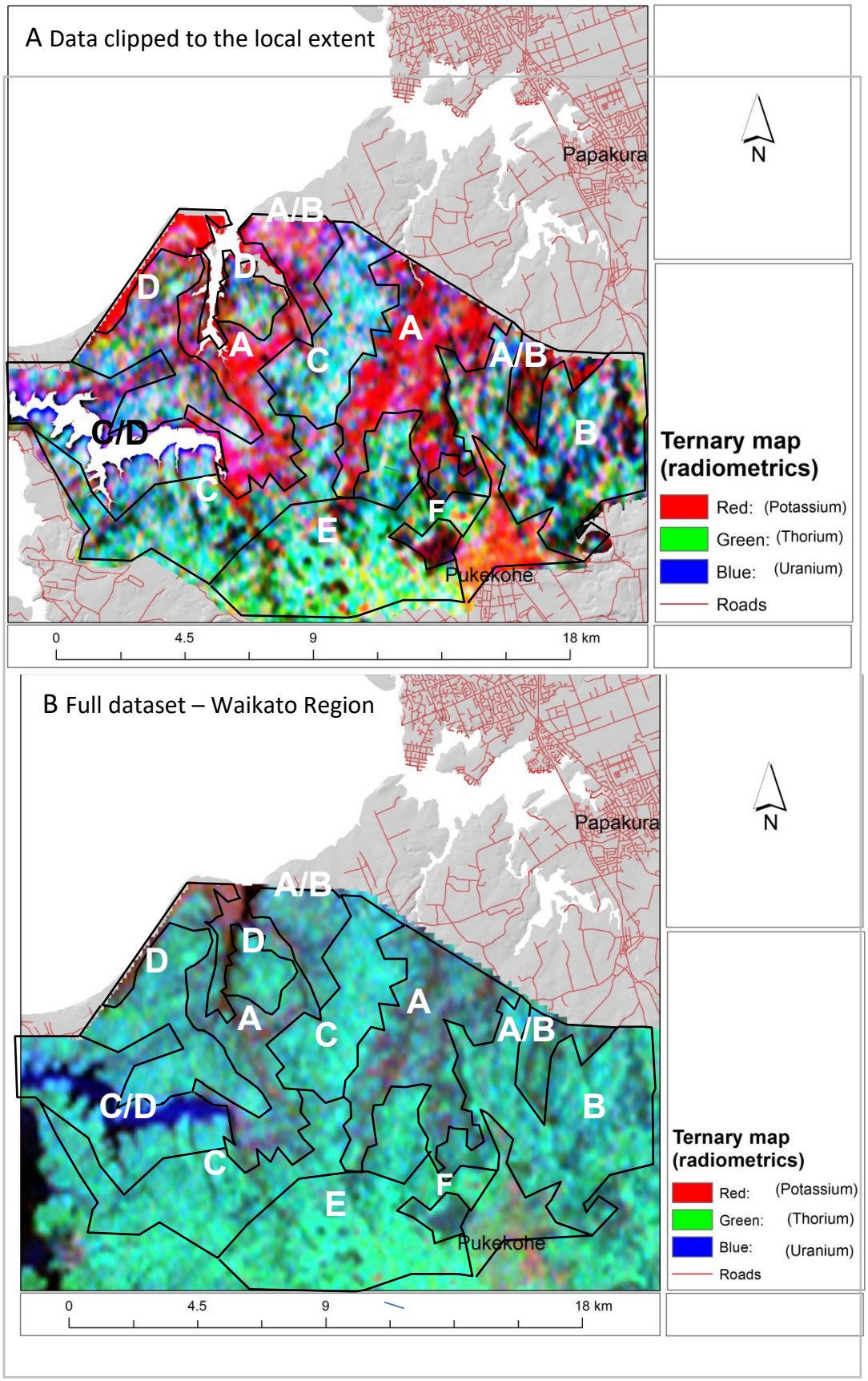


Figure 11 **Ternary map showing the relative radioelement abundance of potassium (red), thorium (green), and uranium (blue) for** A**, the local extent, and** B**, across the Waikato Region**

7.0 Conclusions and future work

7.1 Farm-scale model and map overview

Overall the Random Forests model (Figure 9) provided the best results of farm-scale maps for the Karaka and Gallagher farms. From an expert knowledge and pedological perspective soils occur in the correct position in the landscape and provide a good representation of reality. MNLR and C5 decision trees produced validation statistics that were slightly reduced when compared with the Random Forests model. However, from a qualitative review, the final maps from MNLR and C5 models have some concerns. Although the C5-decision-tree validation statistics were similar to the other models, only four soil types were represented and mapped, compared with 19 soil types observed in the field. All of the soil types in the MNLR map were predicted, but there were unusual patterns occurring, with some floodplain soils occurring in the upper parts of the landscape. For example, recent soil types only found on the lower-lying floodplains were extrapolated by the MNLR model to higher positions in the landscape (hills and terraces). If only considering soil types within the original observation window boundary, then soil predictions are considered reasonable from the investigating pedologist's perspective (expert knowledge).

The DSMART modelling technique results show a similar pattern of soil types on the Karaka and Gallagher farm as produced using Random Forests. Overall, soil types occur in the correct positions in the landscape, but the areas representing Brown and Allophanic soil types are substantially different to the Random Forests model results. From a validation perspective, the DSMART technique has the best prediction statistics (47%; Table 2). This validation statistic assumes independence, but in reality the surveyed observations were used to some degree by pedologists to develop map units, therefore incorporating bias to some extent. The fall-back position is that observations are used for model statistics. As a model statistic the DSMART predictions are ranked third out of the four models. From an expert-knowledge visual assessment the DSMART map provides a reasonable expectation of the types of soils occurring across the two farms, not foregoing the previously discussed concerns around soil-class area representation. Overall, the validation statistics and the visual assessment have guided us to selecting the map produced by the Random Forests model as the preferred map.

When considering the validation matrix for the Random Forests model (Table 5), the number of soils occurring in the diagonal column provides the number of soil classes correctly predicted. This statistic can only be true or false for a given class. Information that is every bit as important is the soil class given (predicted) when the prediction is incorrect. For example, the validation matrix shows that 19 of the NOT soil observations were correctly predicted, whereas three of the observations were predicted as BOT soils, and three observations as GOT soils, while one was classed as an RFT soil. Brown and Gley soils occur in the same position in the landscape as the NOT soils. Because of this overlapping environmental space, it is difficult for a model to separate the soil types using a limited covariate space. Conversely, the RFT soils do not occur in this position in the landscape, but are found on the floodplains, and therefore can be considered an incorrect prediction. This illustrates how validation statistics only provide binary information (correct versus incorrect), but in the real world taxonomically-similar soil types and soil types that

occur in the same covariate space should be recognised. Another consideration is the number of observations available for cross-validation statistics. Where observation numbers are low, outliers can substantially bias results. It should also be remembered that these are cross-validation statistics that are not truly independent of the modelling process. To be independent requires sampling for new observations that are never used in the modelling process that can be tested against model predictions.

While the authors are confident with the mapped results, it is pertinent to note and discuss the influence of model input covariates and the observation data available for the modelling process. In the initial stages of modelling, a variety of different covariates were tested, modelled, and mapped to choose the most appropriate covariates to predict soil types and to determine the best model and map. Other protocols to consider are the number of observations required to represent each soil types, and provide meaningful model and validation statistics. Above all else we need to ensure that all soil types have been observed and captured for the modelling process, because you cannot model a class that does not exist. Figures 5–8 illustrate the covariates used in the final model. The nine covariates used in the Random Forests modelling process were selected because the addition of further covariates provides no substantial improvement to model validation. If covariates are added or removed, the maps change the areas representing soil types, with the concomitant expansion or shrinkage of a class or classes. The Soil Region also has an influence on model outcomes. A Soil Region is defined as a recurring pattern of soils under similar soil-forming factors, with these factors being represented by covariate layers. Modelling beyond the Soil Region will lead to a breakdown of the soil to environmental (soil-forming covariate) relationships. Although the final model combined observations from both the Karaka and Gallagher farm surveys, when modelled separately some of the mapped soil-type areas change. For example, the presence and extent of Allophanic Soil is noticeably different using (1) data only from the Gallagher farm, and (2) data combined from the Karaka and Gallagher farms (using the same input covariates). Also noted was the presence of Organic, Brown, and Recent soils in the valley bottoms and channels of the Gallagher farm when using the combined model. This does not mean these soil types do not occur at the Gallagher farm, only that they were not observed in the Gallagher soil survey. From a stakeholder's perspective, soils in the lower parts of the landscape may be currently considered of less consequence compared with the Allophanic Soils in the higher, more stable parts of the landscape. Allophanic Soils have a greater ability to manage effluent and treat wastes. From a scientist's and environmentalist's perspective, however, the soils of the lower parts of the landscape may be every bit as important because of their potential vulnerability to leaching.

7.2 Future directions

Future directions must consider what Karaka and Pukekohe areas are of priority to be mapped at the farm scale. Discussions with the Franklin Local Board inform us that priority should be given to the intensively-farmed Pukekohe horticultural area. If we use the Soil Region concept to determine the extent to which the soil models apply, and to estimate how many representative areas (windows) will need to be sampled (like the Karaka and Gallagher Road farm sites) then we need to develop windows representative of the recurring pattern of soils across the region. Figure 2 is a GNS-developed map called QMap representing the main rock group for the Pukekohe and Karaka areas. QMap shows a clear delineation between the older basalts around Pukekohe (southern area), and the sands and muds of the Karaka locality suggesting, from a geological perspective, that

parent materials are likely to be different between these areas. Figure 12A and B show a more detailed picture of the likely recurring pattern of soil types across the Pukekohe and Karaka areas resulting from the use of radiometrics. The radiometric-based colour composite map (ternary) in Figure 12A delineates low-lying areas as red colourings (regions A), a western area with darker green and blue colours (region B), elevated coastal terraces as green, turquoise, and blue (regions C), a region closer to the Manukau Harbour with greens, yellows, and blues (region D), and an elevated area to the south shown as yellow, green, and blue colourings (region E). The darker areas on the map (region F) probably indicate areas with high water content or highly vegetated areas, and the urban areas are shown as stronger yellow colourings. The colours shown in this map (Figure 12A) are accentuated because the data has been clipped to this regional extent, which truncates the composite values in relation to the data mean when displayed. In Figure 12B the colours are not as accentuated and the pattern is more difficult to decipher; however, the general pattern remains. We should caution that low-lying areas can be prone to pockets of radon gas collecting in valleys, as well as to variations in soil moisture content, that can influence map colours. Therefore, map colours are used here to describe general patterns relative to geological units. These concerns highlight the need to undertake geochemical surveying in relation to gamma radiometrics in order to fully understand the implications for mapping soils digitally.

Overall, the radiometric data suggests a minimum of five soil windows representing Soil Regions of the Pukekohe and Karaka area. A Soil Region could potentially be a series of transects covering a greater area than an intensively-sampled small soil window. It would be prudent to state that radiometrics is a relatively new technology to New Zealand DSM. The airborne radiometric survey was flown at 200 m altitude with a mean terrain clearance of 60 m. As a result radiometric surfaces were developed at a coarse 50-m cell size resolution compared with our terrain attributes at 5 m. At this stage we think that of all the covariate layers we have investigated, radiometrics seems the most likely to assist us in delineating Soil Regions and providing an indication of parent material mineralogy.

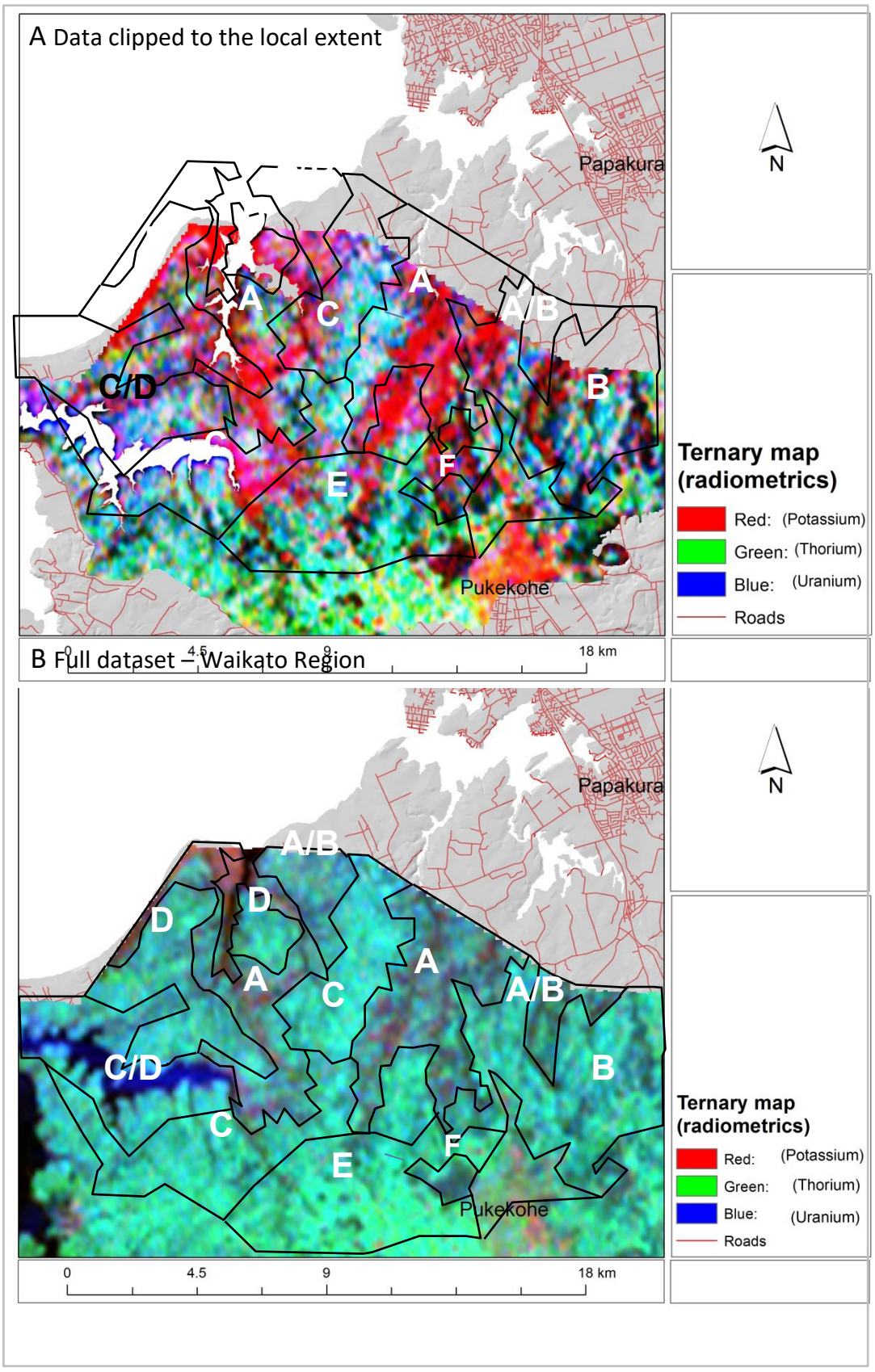


Figure 12 **Ternary map showing the relative radioelement abundance of potassium (red), thorium (green), and uranium (blue) for** A, **the local extent, and** B, **across the Waikato Region.**

8.0 Acknowledgements

This work was funded by Auckland Council under contract number ACPN 14971. Access to LiDAR used in this work was made available through Mike Martindale from Auckland Council. We thank and acknowledge Sinosteel Australia Pty Ltd for providing access to their gamma radiometric data. We thank Wesley Mansell for providing access for soil survey at his Karaka farm. We also thank Brian Gallagher, the owner of Gallagher Road farm and Reece Hill and Dan Borman of LandSystems for providing soil survey maps of that location. Thank you to Doug Hicks and Scott Fraser for their pedological knowledge in relation to this project and its outcomes. We also thank Jim Payne for his comments and review on this report.

9.0 References

Breiman L 2001. Random forests. Machine Learning 45: 5–32.

Briggs RM, Okada T, Itaya T, Shibuya H, Smith IEM 1994. K-Ar ages, paleomagnetism, and geochemistry of the South Auckland volcanic field, North Island, New Zealand. New Zealand Journal of Geology and Geophysics 39: 283–294.

Bui EN, Moran CJ 2003. A strategy to fill gaps in soil survey over large spatial extents: an example from the Murray–Darling basin of Australia. Geoderma 111: 21–44.

Congalton RG 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sensing of Environment 37: 35–47.

Edbrooke SW comp. 2001. Geology of the Auckland area. Institute of Geological and Nuclear Sciences 1:250,000 Geological Map 3. 1 sheet + 74 p. Lower Hutt, IGNS.

Gallant JC, Wilson JP 2000. Primary topographic attributes. In: Wilson JP, Gallant JC eds Terrain analysis: principles and applications. New York, John Wiley.

Grimm T, Behrens T, Marker M, Elsenbeer H 2008. Soil organic carbon concentrations and stocks on Barro Colorado Island: Digital soil mapping using Random Forests analysis. Geoderma 147: 102–113.

Grunwald S 2009. Multi-criteria characterization of recent digital soil mapping and modeling approaches. Geoderma 152: 195–207.

Hastie T, Tibshirani R, Friedman J 2009. The elements of statistical learning: data mining, inference, and prediction. 2nd edn. Springer. ISBN 0387848576.

Hewitt AE 2010. New Zealand soil classification. Landcare Research Science Series No. 1. Lincoln, Manaaki Whenua Press.

Holmes KW, Odgers NP, Griffin, EA, van Gool D 2014. Spatial disaggregation of conventional soil mapping across Western Australia using DSMART. In: Arrouays D, McKenzie N, Hempel J, Richer de Forges A, McBratney AB eds GlobalSoilMap: Basis of the global spatial soil information system. Boca Raton, FL, CRC Press. Pp. 273–278.

IAEA 2003. Guidelines for radioelement mapping using gamma ray spectrometry data. Vienna, Austria, International Atomic Energy Agency. ISBN 92-0-108303-3.

Jenny H 1941. Factors of soil formation, a system of quantitative pedology. New York, McGraw-Hill.

Kempen B, Brus DJ, Heuvelink GBM, Stoorvogel JJ 2009. Updating the 1:50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach. Geoderma 151: 311–326.

Lagacherie P, Holmes S 1997. Addressing geographical data errors in a classification tree soil unit prediction. International Journal of Geographical Information Science 11: 183–198.

Liaw A, Wiener M 2002. Classification and regression by randomForest. R News 2(3): 18–22.

Leathwick JR, Wilson G, Stephens RTT 2002. Climate surfaces for New Zealand. Landcare Research Contract Report LC9798/126. Hamilton, Landcare Research. 22 p.

Leathwick J, Wilson G, Rutledge D, Wardle P, Morgan F, Johnston K, McLeod M, Kirkpatrick R 2003. Land environments of New Zealand. Wellington, Ministry for the Environment, and Hamilton, Manaaki Whenua Landcare Research. 184 p.

Lilburne LR, Webb TH, Hewitt AE, Lynn IH, de Pauw B 2011. S-map database manual v1.2. Lincoln, Manaaki Whenua – Landcare Research.

Lilburne L, Webb T, Palmer D, McNeill S, Hewitt A, Fraser S 2014. Pedo-transfer functions from s-map for mapping water holding capacity, soil-water demand, nutrient leaching vulnerability and soil services. In: Currie LD, Chistensen CL eds Nutrient management for the farm, catchment and community. Occasional Report No. 27. Palmerston North, Fertilizer and Lime Research Centre, Massey University. http://flrc.massey.ac.nz/publications.html.

Lowe DJ 2010. Pukekohe silt loam, Pukekohe Hill. In: Lowe DJ, Neall VE, Hedley M, Clothier B, Mackay A. Guidebook for pre-conference North Island, New Zealand 'Volcanoes to Ocean' field tour (27–30 July, 2010). 19th World Soils Congress, International Union of Soil Sciences, Brisbane. Soil and Earth Sciences Occasional Publication No. 3, Palmerston North, Massey University. Pp. 1.12–1.23.

McBratney AB, Mendonça Santos ML, Minasny B 2003. On digital soil mapping. Geoderma 117: 3–52.

Minasny B, McBratney AB 2002. Uncertainty analysis for pedotransfer functions. European Journal of Soil Science 53: 417–429.

Minasny B, McBratney AB 2007. Incorporating taxonomic distance into spatial prediction and digital mapping of soil classes. Geoderma 142: 285–293.

Moran CJ, Bui E 2002. Spatial data mining for enhanced soil map modelling. International Journal of Geographical Information Science 16: 533–549.

Odgers NP, Sun W, McBratney AB, Minasny D, Clifford D 2014. Disaggregating and harmonising soil map units through resampled classification trees. Geoderma 214: 91–100.

Palmer DJ, Höck BK, Dunningham AG, Lowe DJ, Payn TW 2009. Developing national-scale terrain attributes for New Zealand (TANZ). FRI Bulletin 232. Rotorua, Forest Research Institute.

Quinlan JR 1993. C4.5: Programs for machine learning. San Mateo, CA, Morgan Kaufmann.

Rae AJ 1995. Application of geographical information systems to a soil-paleosol drainage sequence, Pukekohe Hill, South Auckland, New Zealand. Unpublished dissertation for Diploma in Applied Science, University of Waikato, Hamilton. 80p.

Schmidt J, Hewitt A 2004. Fuzzy land element classification from DTMs based on geometry and terrain position. Geoderma 121: 243–256.

Tarboton D 2014. Hydrology research group http://hydrology.usu.edu/taudem/taudem5/downloads.html. (accessed August 2014).

Webb TH 2003. Identification of functional horizons to predict physical properties for soils from alluvium in Canterbury, New Zealand. Australian Journal of Soil Research 41: 1005–1019.

Wilson JP, Gallant JC 1998. Terrain-based approaches to environmental resource evaluation. In: Lane SN, Richards KS, Chandler JH eds Landform monitoring, modelling, and analysis. New York, John Wiley. Pp. 219–240.

Wilson JP, Gallant JC 2000. Digital terrain analysis. In: Wilson JP, Gallant JC eds Terrain analysis: principles and applications. New York, John Wiley

Find out more: phone 09 301 0101, email rimu@aucklandcouncil.govt.nz or visit aucklandcouncil.govt.nz **and knowledgeauckland.org.nz**

Auckland Council
Te Kaunihera o Tāmaki Makaurau