

## Potential of multi-objective models for risk-based mapping of the resilience characteristics of soils: demonstration at a national level

M. DEBELJAK<sup>1</sup>, D. KOCEV<sup>1</sup>, W. TOWERS<sup>2</sup>, M. JONES<sup>2</sup>, B. S. GRIFFITHS<sup>3,\*</sup> & P. D. HALLETT<sup>3</sup>

<sup>1</sup>Department of Knowledge Technologies, Jozef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia, <sup>2</sup>Macaulay Institute, Craigiebuckler, Aberdeen AB15 8QH, UK, and <sup>3</sup>Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA, UK

### Abstract

Policy makers rely on risk-based maps to make informed decisions on soil protection. Producing the maps, however, can often be confounded by a lack of data or appropriate methods to extrapolate using pedotransfer functions. In this paper, we applied multi-objective regression tree analysis to map the resistance and resilience characteristics of soils onto stress. The analysis used a machine learning technique of multiple regression tree induction that was applied to a data set on the resistance and resilience characteristics of a range of soils across Scotland. Data included both biological and physical perturbations. The response to biological stress was measured as changes in substrate mineralization over time following a transient (heat) or persistent (copper) stress. The response to physical stress was measured from the resistance and recovery of pore structure following either compaction or waterlogging. We first determined underlying relationships between soil properties and its resistance and resilience capacity. This showed that the explanatory power of such models with multiple dependent variables (multi-objective models) for the simultaneous prediction of interdependent resilience and resistance variables was much better than a piecewise approach using multiple regression analysis. We then used GIS techniques coupled with an existing, extensive soil data set to up-scale the results of the models with multiple dependent variables to a national level (Scotland). The resulting maps indicate areas with low, moderate and high resistance and resilience to a range of biological and physical perturbations applied to soil. More data would be required to validate the maps, but the modelling approach is shown to be extremely valuable for up-scaling soil processes for national-level mapping.

**Keywords:** Multi-objective regression trees, pedotransfer, soil resilience, digital mapping, soil risk-based maps

### Introduction

In the European Commission Thematic Strategy for Soil Protection (European Commission, 2006), it is recognized that soil is subject to a series of degradation processes or threats. The proposed Soils Framework Directive (European Commission, 2006) will require member states to identify risk areas on the basis of common elements to be taken into account, set risk reduction targets for those areas and estab-

lish programmes of measures to achieve them. Thus, the use of regional, risk-based maps indicating soil vulnerability to stresses is now commonplace and is likely to become more important because of growing concerns about soil sustainability worldwide, as evident in Europe following discussions of the EU Soils Framework Directive (European Commission, 2006). Policy makers rely on these maps to make larger scale decisions about soil protection and land-use strategies. Some recent examples of risk-based maps include subsoil compaction (Horn *et al.*, 2005), soil erosion risk (Davidson & Grieve, 2004), cobalt deprivation in grazing livestock (Suttle *et al.*, 2003), sensitivity of inland waters to acidification (Kernan *et al.*, 2004) and run-off risk from slurry application (Jordan *et al.*, 2007).

Correspondence: B. S. Griffiths. E-mail: bryan.griffiths@teagasc.ie

\*Present address: Teagasc, Environment Research Centre, Johnstown Castle, Wexford, Ireland.

Received November 2008; accepted after revision November 2008

There are various approaches to making regional, risk-based maps, including generalized linear models, classification and regression trees, neural networks, fuzzy systems and geostatistics (McBratney *et al.*, 2003). Generally, the quality of risk-based maps is limited by the spatial availability of data. Estimates of soil properties are therefore often made using pedotransfer functions, which can then be used to identify 'at-risk' soil categories based on specific soil characteristics. Often pedotransfer functions are simple multiple regression equations, but a major limitation to this approach is that it requires the 'function' being mapped to have a quantifiable relationship with specific soil properties. Model error can present considerable uncertainty in predicting soil processes from general properties (Chirico *et al.*, 2007). This can lead to the situation where models that can accurately predict 'function' are not suitable for extrapolation to the regional level as they either require extensive data input (Jordan *et al.*, 2007) or produce maps with considerable error (McBratney *et al.*, 2003).

The data collected also need to correspond to the data from which the underlying soil database was constructed, so that geographic information system (GIS) techniques can be used to generate the regional maps. Using simple multiple regression to generate pedotransfer functions can give considerable uncertainty in the resulting map if there is a poor correlation between the soil properties and the 'function'. By using regression, rather than process-based modelling approaches, there can also be heavy empiricism and non-causal links between soil properties and 'function'. Although a powerful aspect of using pedotransfer functions is the estimation of a particular soil process, ranging from hydraulic conductivity (Chirico *et al.*, 2007) to heavy metal sorption (Deurer & Bottcher, 2007), such estimates may not be necessary for risk-based mapping as the data are used to form broad classifications of the potential threat to soil resources. This opens the possibility of using other approaches to develop risk-based maps where multiple regression functions are not suitable.

One example where simple multiple regression equation approaches would not have been suitable for generating pedotransfer functions was a study on the resistance and resilience of some Scottish soils to experimentally applied biological and physical perturbations (Kuan *et al.*, 2007). Here, we adopted the definitions of resistance and resilience as described by Seybold *et al.* (1999), whereby resistance is the magnitude of the decline in capacity of the soil to function and resilience is the rate of recovery. Both are key measures of sustainability. The study by Kuan *et al.* (2007) found good correlations between soil properties and resilience to compression and copper but no such relationship with resilience to heat. Soil resilience is a key property in deciding on management or future land-use options; so, the provision of regional resilience maps could be a very useful aid for decision makers. This paper proposes the use of

data mining in order to obtain classification of soil factors affecting soil resilience as an alternative to deriving pedotransfer functions from simple multiple regression approaches. The machine learning technique of multiple regression tree induction was applied for deriving models on the effects of soil factors on the biological and physical resilience of Scottish soils based on many positive published results from applying machine learning methods (Struyf & Džeroski, 2006) and in particular in ecological modelling (Debeljak *et al.*, 2001, 2007, 2008; Džeroski, 2001; Jerina *et al.*, 2003).

This paper has two major objectives. The first is to determine whether any underlying relationships exist between soil properties and resilience/resistance capacity. Previous studies have tended to look only for partial relationships with no attempt to apply a systems approach to this question. Our hypothesis was that the explanatory power of such models with multiple dependent variables (multi-objective models) for the simultaneous prediction of interdependent resilience and resistance variables would be much higher than that in the piecewise approach taken by Kuan *et al.* (2007). The second is to demonstrate the feasibility of up-scaling results of models with multiple dependent variables using GIS techniques coupled with an existing, extensive soil data set. The aim was to demonstrate the potential of the methodology by producing initial soil resilience and resistance maps to identify the range and spatial distribution of biological and physical resilience of Scottish soils.

## Materials and methods

Two distinctive types of methods were applied in this study. To address the first objective, we used machine learning techniques to produce models with multiple dependent variables on data from representative soil types throughout Scotland. The second part of the study was based on data analysis and manipulation with various GIS tools and to apply the models with multiple dependent variables to a large data set of Scottish soil attributes to produce soil resilience and resistance maps.

### *The data sets*

The research was conducted on two data sets. The first was used to create models with multiple dependent variables using machine learning techniques and the second was to produce resilience maps using a GIS. It should be emphasized that the purpose of this study was to assess the utility of models with multiple dependent variables for the mapping of soil processes. The resulting maps would require validation and the models be further improved through additional data collection before they could provide reliable tools for decision makers.

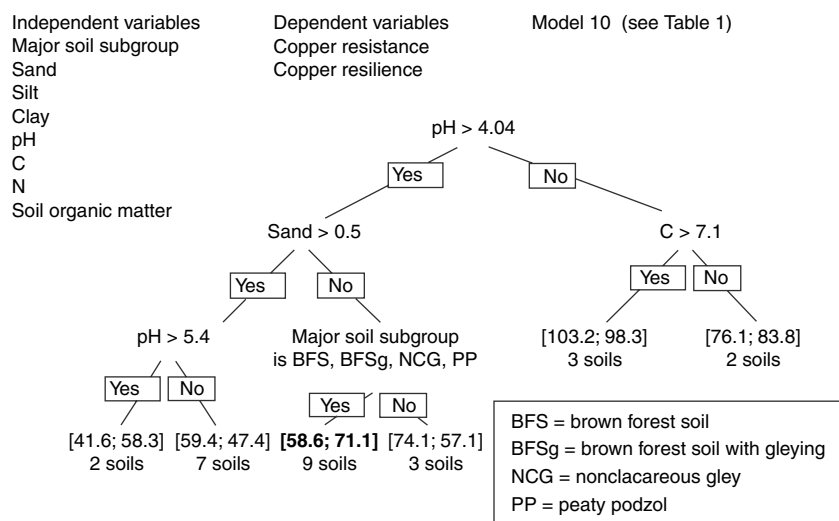
The data for the creation of models with multiple dependent variables were taken from the study by Kuan *et al.* (2007) in which soil was sampled from 26 sites around Scotland at a depth of 100–300 mm. Background measurements (or independent variables in the terminology of the models) included carbon (C) and nitrogen (N) content, organic matter content, soil electrical conductivity, pH, basal respiration, microbial biomass, dissolved organic carbon,  $\text{NO}_3^-$  and  $\text{NH}_4^-\text{-N}$ , microbial metabolic quotient and particle size distribution. Four measurements of resilience (dependent variables) were made at each of the 26 sampling sites. These were the effects of heat or copper on the short-term decomposition of added plant material, the physical compression and rebound characteristics, and the response and recovery to overburden stress at saturation. For each measure of stability, the data gave a measure of both resistance (instantaneous susceptibility to the perturbation) and resilience (recovery in function over time following perturbation). Thus, the final set of dependent variables consisted of heat resistance, heat resilience, copper resistance, copper resilience, compression resistance, compression resilience, overburden resistance and overburden resilience. The table which we used for the creation of data mining models comprised 26 rows (sites where soil samples were taken) and 18 columns (dependent and independent variables).

The second data set comprised of soil attributes across Scotland (Towers *et al.*, 2006). Scotland has been mapped at the scale of 1:250 000 and the map is supported by a database of approximately 13 000 soil profiles, including over

40 000 individual soil horizons, held in the Scottish soils database in the Macaulay Institute, Aberdeen, Scotland. A large number of variables are held for each profile and horizon, but those utilized in this study comprise major soil subgroup (according to the soil classification used in Scotland), soil texture (sand, silt and clay percentages), pH, organic C content, total N content and soil organic matter.

These data have recently been summarized in a variable data set called Scottish Soils Knowledge and Information Base (SSKIB; Lilly *et al.*, 2004). It includes summary statistics and expert knowledge on the chemical, physical and biological aspects of Scottish soils. The 1:250 000 scale national soil map of Scotland was used as the basis for the project. A variable data set (SSKIB) was developed for all the soils that occur on the national map. The common framework involved determining the proportions of soil types within each map unit of the 1:250 000 scale map and then developing typical horizon sequences for each soil type. Summary data such as carbon content or soil nutrients were determined from over 40 000 soil analyses and are a component part of SSKIB.

Scottish Soils Knowledge and Information Base was developed in an MS-Access relational database structure to allow linkages between the various component parts at different levels. At the highest level, there is information on the individual soil map units such as parent material and soil association (a grouping of soil series developed on similar parent materials). At the next level, there is information on the individual soil types (series) such as soil leaching potential, hydrology of soil types (HOST) class



**Figure 1** Multi-objective regression tree to predict the resistance and resilience of soils to copper perturbation. The independent variables are soil properties taken from soil survey data: major soil subgroup, sand, silt, clay, pH, C, N and soil organic matter. The dependent variables are the resistance and the resilience of the soil to copper stress (model 10, Table 1). The six stability classes are defined on both resistance and resilience, values shown in the leaves (lowest levels of the regression tree) are [resistance; resilience]. The number of soils from the original experiment in each class is given. Figures in bold refer to the example in the text.

(Boorman *et al.*, 1995) and proportion of component soil series. At the next level, the MS-Access table comprises information on typical horizon sequences for each soil series (cultivated and uncultivated where appropriate) and thicknesses of each horizon. The final level of information is summary statistics (mean, median, geometric mean, etc.) of the chemical and physical properties of each individual soil horizon for both cultivated and uncultivated soils. Information on the geochemical signature of upper horizons was also collated from previous work and extended to provide limited but extensive cover. This involved grouping soil associations based on their component rock types and stratigraphy. The main linking variables of SSKIB, essential for map production, are a unique soil series numeric code and horizon notation at lower levels and a series or map unit at higher levels.

#### Data mining

Regression trees are a representation of piecewise constant or piecewise linear functions, and models are given in a form of hierarchical structures of their elements (Figure 1). Like classical regression equations, they predict the numerical value of a dependent variable from the values of a set of independent variables (Breiman *et al.*, 1984). A regression tree, which has a form of inverse hierarchical structure, has a test in each inner node (junction from where two links go to the lower hierarchical levels) that tests the value of a certain independent variable, and in each leaf (the lowest level of a hierarchical tree) can be a linear equation or just a constant for predicting the value of the dependent variable.

Regression trees are able to predict the value of single numerical dependent variables, while multi-objective regression trees (MORTs) (Blockeel *et al.*, 1998) are capable of predicting several numerical dependent variables simultaneously (for all dependent variables only one tree is induced). The MORTs are an instantiation of the predictive clustering framework and is implemented in the CLUS system (Blockeel & Struyf, 2002). With this approach, the prediction is a vector of numeric values (one component of the vector for each dependent variable). MORTs are generalizations of the regression trees and have two main advantages over building a separate regression tree for each dependent variable: a single MORT is usually much smaller than the total size of the individual trees for all dependent variables, and a MORT specifies dependencies between the different dependent variables. In this research, we used the CLUS system for inducing (multi-objective) regression trees. In the CLUS system, the heuristic for selection of the tests in the internal nodes is the sum of the variations in the induced subsets and it is used for inducing both single-objective and multi-objective regression trees. More information about the CLUS system is available at <http://www.cs.kuleuven.be/~dtai/clus/> (last accessed 23 January 2009).







**Table 1** Quantitative evaluation of 16 models that best explain combinations of soil resistance and resilience to perturbations of heat, copper, compression and overburden when saturated

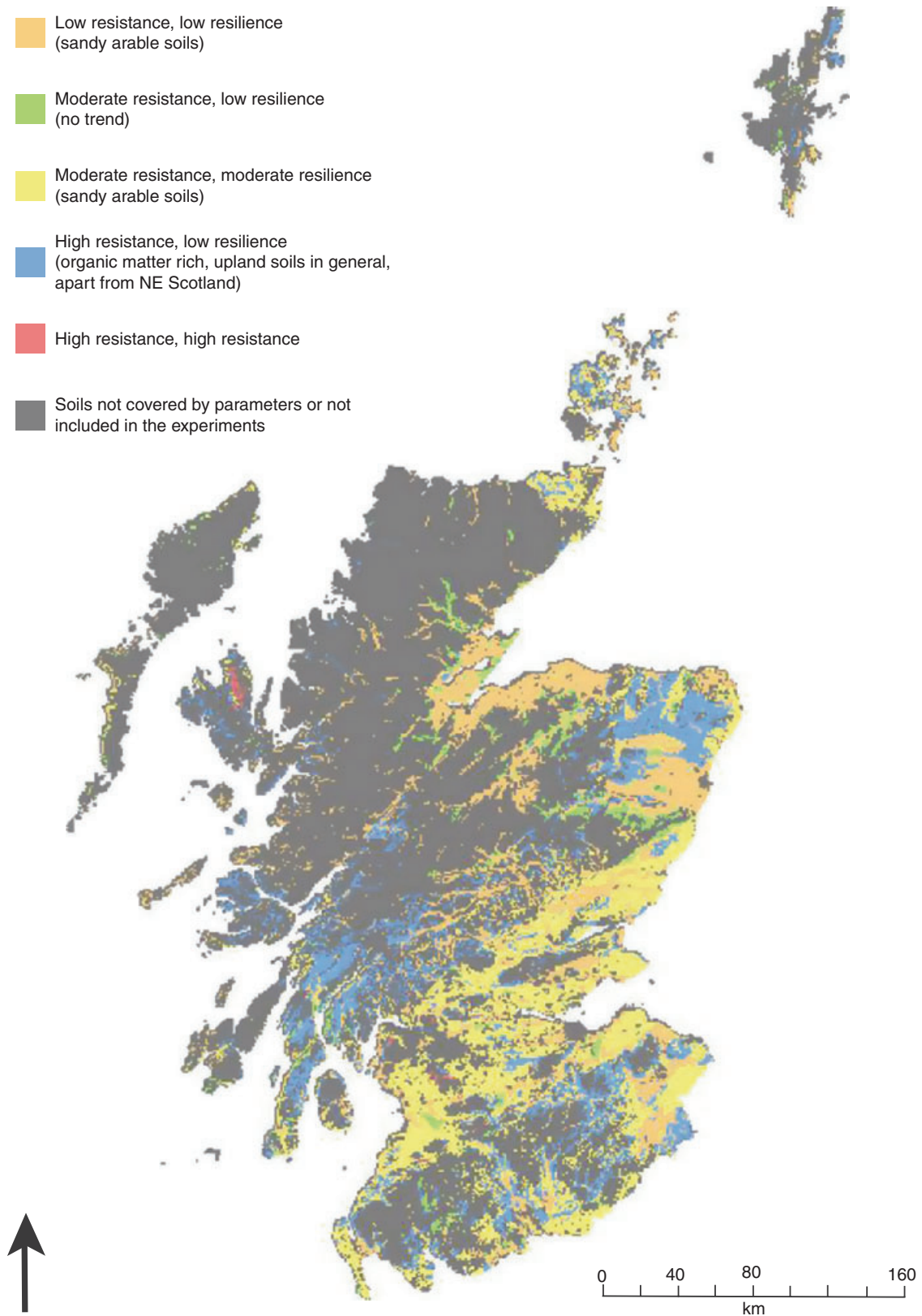
Model	Dependent variables	RMSE	CC
1	Heat resistance	15.07	0.16
2	Heat resilience	26.09	-0.17
3	Copper resistance	14.17	0.63
4	Copper resilience	25.15	0.24
5	Compression resistance	0.02	0.6
6	Compression resilience	0.28	0.55
7	Overburden resistance	0.07	0.28
8	Overburden resilience	0.23	0.33
9	Heat resistance	14.89	0.13
	Heat resilience	25.09	0.28
<b>10</b>	<b>Copper resistance</b>	<b>15.14</b>	<b>0.62</b>
	<b>Copper resilience</b>	<b>21.1</b>	<b>0.36</b>
<b>11</b>	<b>Compression resistance</b>	<b>0.24</b>	<b>0.71</b>
	<b>Compression resilience</b>	<b>0.02</b>	<b>0.48</b>
<b>12</b>	<b>Overburden resistance</b>	<b>0.1</b>	<b>-0.3</b>
	<b>Overburden resilience</b>	<b>0.26</b>	<b>0.11</b>
13	Heat resistance	15.72	-0.05
	Heat resilience	23.91	-0.21
	Copper resistance	12.45	0.72
	Copper resilience	23.04	0.19
14	Compression resistance	0.29	-0.28
	Compression resilience	0.02	0.01
	Overburden resistance	0.09	0.52
	Overburden resilience	0.22	0.47
<b>15</b>	<b>Copper resistance</b>	<b>0.63</b>	<b>0.81</b>
	<b>Copper resilience</b>	<b>0.99</b>	<b>0.44</b>
	<b>Compression resistance</b>	<b>1.06</b>	<b>0.36</b>
	<b>Compression resilience</b>	<b>0.55</b>	<b>0.85</b>
<b>16</b>	<b>Heat resistance</b>	<b>15.48</b>	<b>-0.31</b>
	<b>Heat resilience</b>	<b>20.89</b>	<b>-0.08</b>
	<b>Copper resistance</b>	<b>11.01</b>	<b>0.81</b>
	<b>Copper resilience</b>	<b>18.59</b>	<b>0.44</b>
	<b>Compression resistance</b>	<b>0.08</b>	<b>-0.14</b>
	<b>Compression resilience</b>	<b>0.21</b>	<b>0.28</b>
	<b>Overburden resistance</b>	<b>0.34</b>	<b>0.35</b>
	<b>Overburden resilience</b>	<b>0.01</b>	<b>0.85</b>

The evaluating criteria are the root mean-squared error (RMSE) and the correlation coefficient (CC), see text for details. Those models with the highest criteria, shown in bold, were selected for mapping.

Regression trees are widely used in modelling (Tan *et al.*, 2006) and are very easy to interpret and understand. The methods for the tree induction process are non-parametric (it does not require prior assumptions for the probability distribution of the dependent and the other variables) and they are not computationally expensive, even on large data sets. Moreover, the process of tree induction is not influenced by redundant variables and noise. In essence, regression trees are models that are interpretable, have reasonable predictive performance and can be obtained quite fast.

Soil biological stability to copper

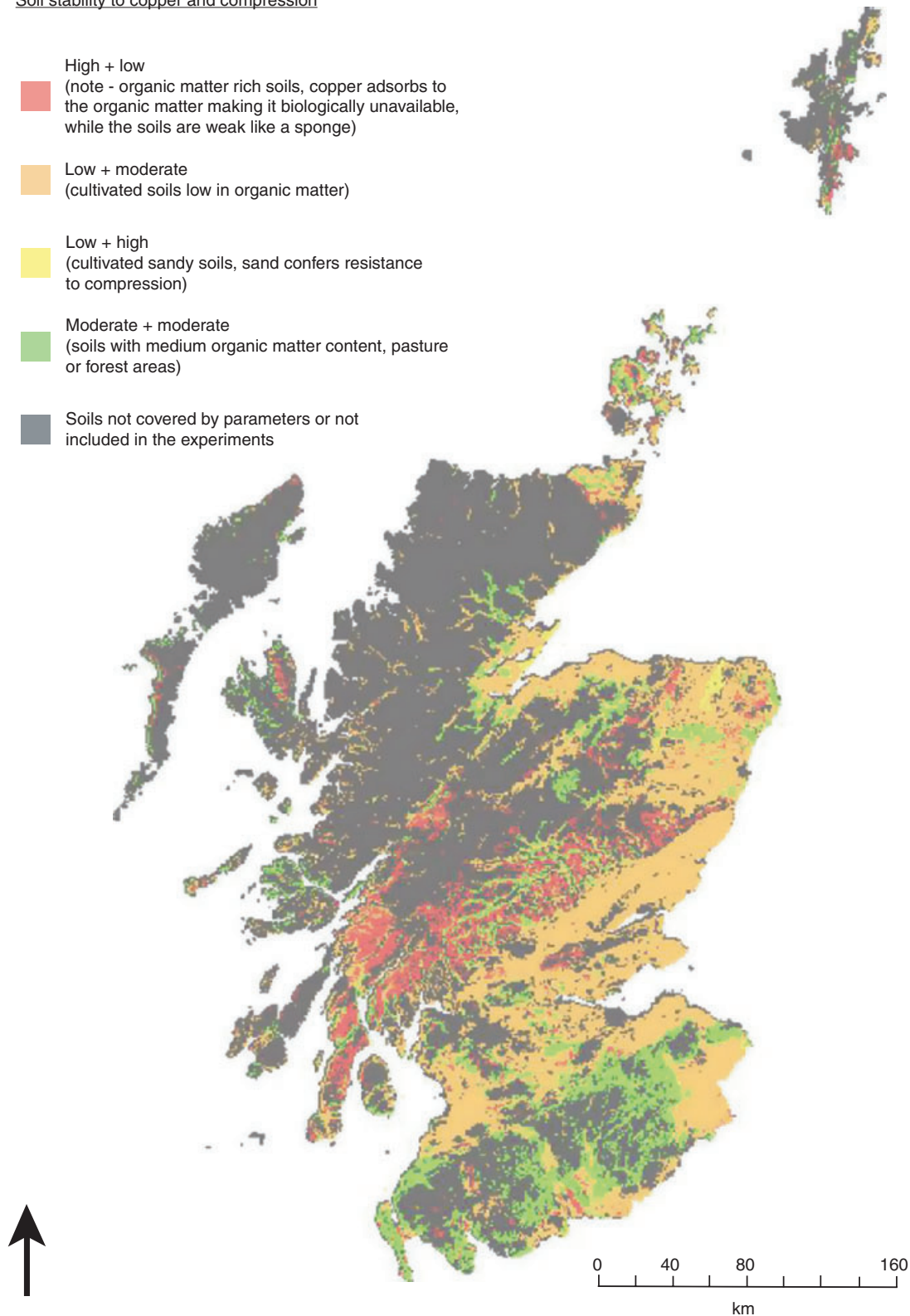
-  Low resistance, low resilience  
(sandy arable soils)
-  Moderate resistance, low resilience  
(no trend)
-  Moderate resistance, moderate resilience  
(sandy arable soils)
-  High resistance, low resilience  
(organic matter rich, upland soils in general,  
apart from NE Scotland)
-  High resistance, high resilience
-  Soils not covered by parameters or not  
included in the experiments



**Figure 2** Risk-based map of stability (resistance and resilience) to copper generated from the regression tree in Figure 1 (model 10; Table 1).

Soil stability to copper and compression

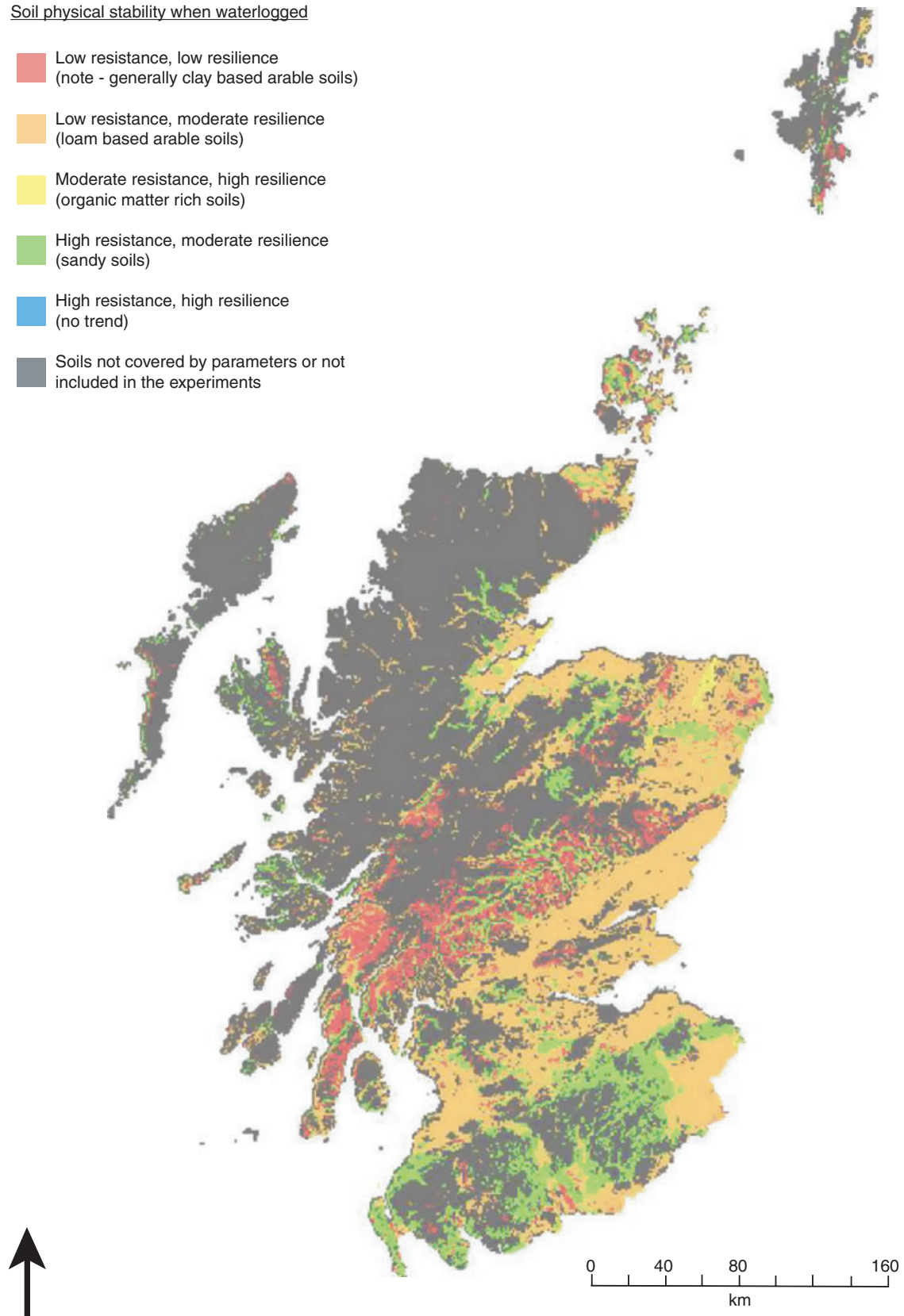
- High + low  
 (note - organic matter rich soils, copper adsorbs to the organic matter making it biologically unavailable, while the soils are weak like a sponge)
- Low + moderate  
 (cultivated soils low in organic matter)
- Low + high  
 (cultivated sandy soils, sand confers resistance to compression)
- Moderate + moderate  
 (soils with medium organic matter content, pasture or forest areas)
- Soils not covered by parameters or not included in the experiments



**Figure 3** Risk-based map of stability (resistance and resilience) to copper and compression (model 15; Table 1). The land area (km<sup>2</sup> and % of total area) covered by each of the categories is: high and low (5209, 7%), low and moderate (20 875, 26%), low and high (277, 0.4%), moderate and moderate (8421, 11%), not covered (43 218, 55%).





Soil physical stability when waterlogged

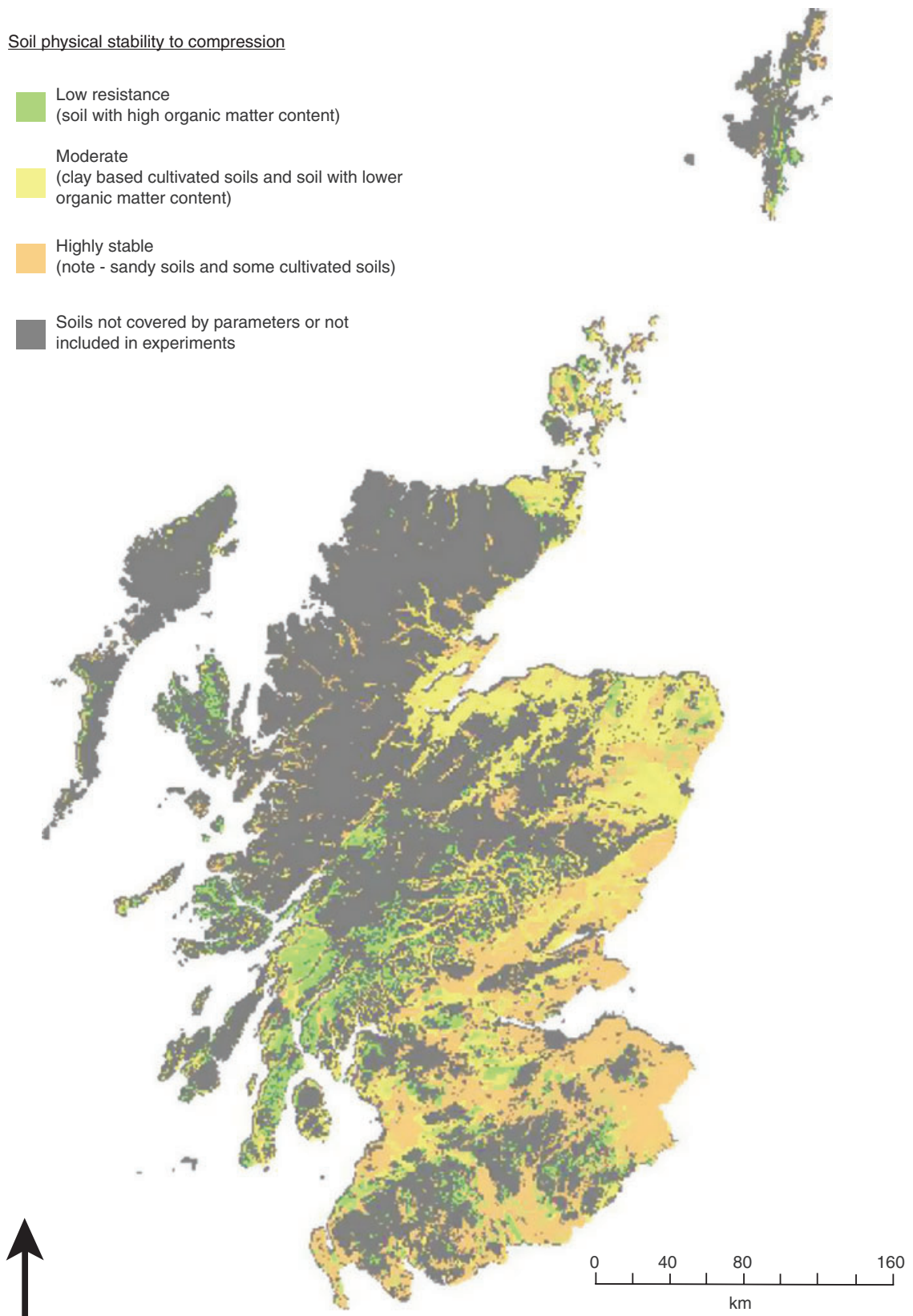
- Low resistance, low resilience  
(note - generally clay based arable soils)
- Low resistance, moderate resilience  
(loam based arable soils)
- Moderate resistance, high resilience  
(organic matter rich soils)
- High resistance, moderate resilience  
(sandy soils)
- High resistance, high resilience  
(no trend)
- Soils not covered by parameters or not  
included in the experiments



**Figure 4** Risk-based map of stability (resistance and resilience) to waterlogging (model 12; Table 1).

Soil physical stability to compression

-  Low resistance  
(soil with high organic matter content)
-  Moderate  
(clay based cultivated soils and soil with lower organic matter content)
-  Highly stable  
(note - sandy soils and some cultivated soils)
-  Soils not covered by parameters or not included in experiments



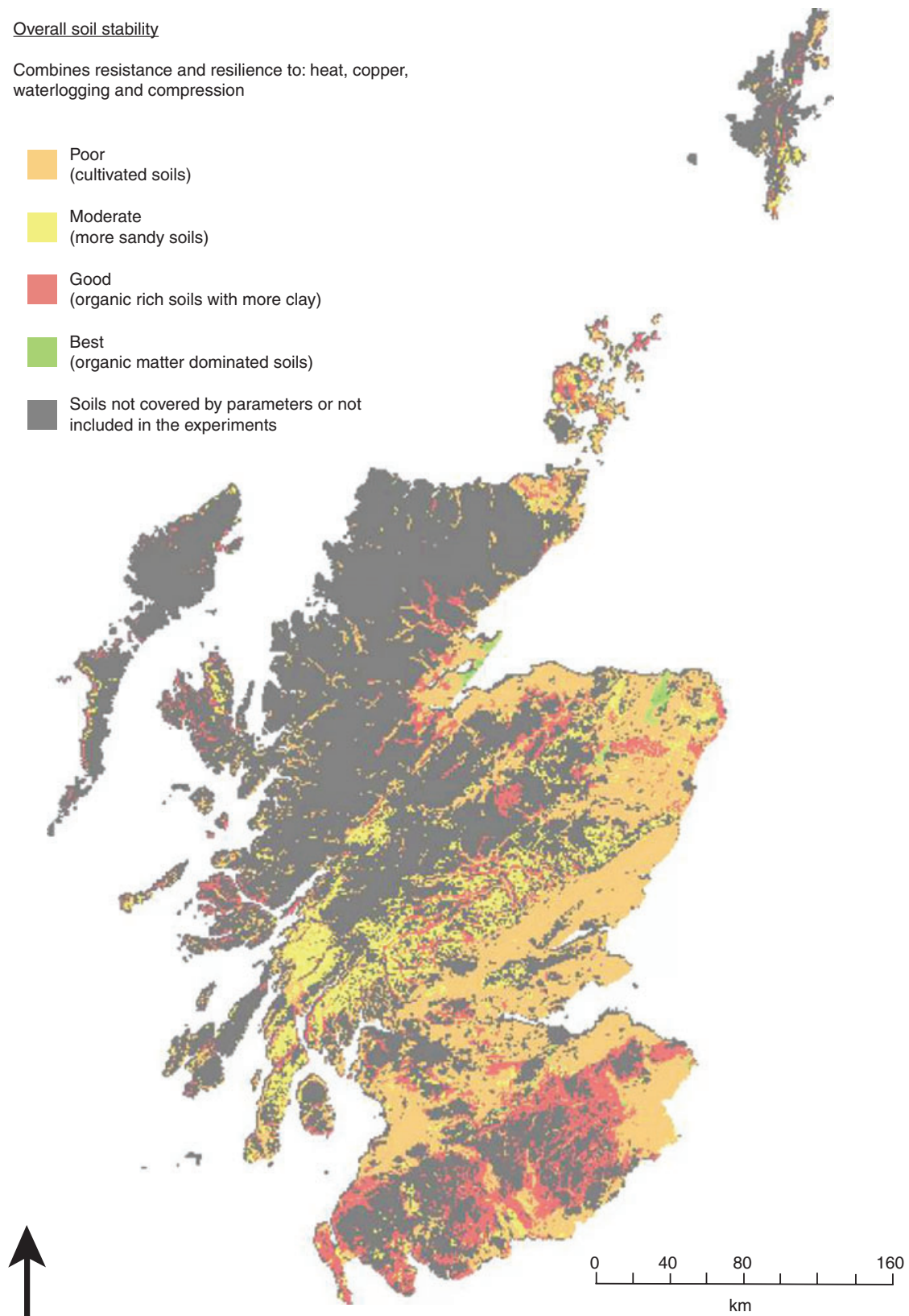
**Figure 5** Risk-based map of stability (resistance and resilience) to compression (model 11; Table 1).



Overall soil stability

Combines resistance and resilience to: heat, copper, waterlogging and compression

- Poor  
(cultivated soils)
- Moderate  
(more sandy soils)
- Good  
(organic rich soils with more clay)
- Best  
(organic matter dominated soils)
- Soils not covered by parameters or not included in the experiments



**Figure 6** Risk-based map of overall soil stability (resistance and resilience) determined from the response to four stresses (copper, heat, compression and waterlogging) (model 16; Table 1).

### Experimental design

To induce multi-objective regression trees, we selected from the first data set (e.g. 18 variables and 26 sites) those soil variables that were also present in the Scottish soils database. These were major soil subgroup, soil texture, pH, carbon, nitrogen and organic matter content. The CLUS system was applied and several models were generated. To evaluate the generated models, and to select the models which have been used to produce soil maps in the second part of the research, we used three types of criteria: quantitative, qualitative and contextual. The quantitative aspects of the models were evaluated by the Pearson correlation coefficient (CC) and the root mean-squared error (RMSE). For validation we used the leave one out (LOO) procedure because of the small number of samples. Use of other validation procedures such as 10-fold cross-validation and splitting the data set into training and testing part in our study were not feasible due to the sample size (Quinlan, 1986). The Pearson correlation coefficient was the most important quantitative measure of model performance. The qualitative aspect of the model was assessed from how well the model classified different soils into their expected ranges. We validated the internal structure of the models through assessing the accuracy of the expected relationships between the elements of the model from our existing knowledge about the soil properties. The models with the higher correlation coefficients with up to three hierarchical levels and the best explanatory power were selected for producing resilience maps.

### Generation of resilience maps

The model outputs in the form of regression trees were recreated as SQL (Structured Query Language) conditional queries in Microsoft Access using the SSKIB data. The median values for the relevant variables (e.g. organic carbon and total N) for the surface horizon of the soil were used as these would be equivalent to the samples taken in the original experiment (Kuan *et al.*, 2007). Regression trees were generated using results from Kuan *et al.* (2007) to relate soil properties with resilience. As the regression trees contain a number of thresholds for the independent variables, the SSKIB data set was 'sorted' into a defined number of classes based on these thresholds. This provided an output of the models by soil series which were subsequently linked to a 1-km grid data set of the 1:250 000 scale soil map. This contains information on the proportion of soil series within each 1-km cell to produce a spatial representation of this output. In this project, the dominant soil series in each 1-km grid cell was used for this extrapolation phase of the work. The SSKIB data offer a number of different options for the mapping phase but using median values and dominant soil was judged appropriate for the purpose of demonstrating the utility of the method-

ology. The different options could be useful with other mapping objectives.

In preparing the maps, the soil classes have been put into more descriptive categories rather than the specific estimates of stability generated by the regression trees. The class descriptions on all of the maps are based on the expert judgement of the authors and how the soils in that grouping responded in the experiments. These categories used the groupings 'low', 'moderate' and 'high' for each of resistance and resilience. The maps show either resistance or resilience to a particular stress, combined resistance and resilience to a stress, or the resistance and resilience to combined stresses. Maps were generated in identical fashion from the models highlighted in Table 1.

### Results and discussion

Several models were generated for each dependent variable or combination of variables. All induced models, regardless of their type, used the same set of independent variables: major soil subgroup, soil texture, pH; carbon, nitrogen and organic matter content. After comparison of the different model types, the ones with the best performance evaluation were selected for further interpretation and visualization through GIS (Table 1).

An example of a resulting regression tree for resilience and resistance to copper (model 10, Table 1) is shown in Figure 1. In this example, soils with a pH >4.04 have <50% sand and are either Brown Forest Soils, Brown Forest Soils with Gleying, Non-calcareous Gleys or Peaty Podzols and have a copper resistance of 58.6 and a copper resilience of 71.1 (shown in bold in Figure 1). There were six soil classes predicted by this model and they were used to make the soil map (Figure 2). In this case, one of the classes, soils with a pH <4.04 and a carbon content <7.6%, actually occupied such a small portion of the soil map that it is not represented (Figure 2).

By using those models with acceptable evaluation criteria (Table 1), we were able to generate risk-based soil maps of stability to copper (Figure 2), copper and compression (Figure 3), overburden (Figure 4), compression (Figure 5), heat, copper, compression and overburden (Figure 6). Stability incorporated both the resistance and resilience characteristics to the imposed perturbations. These outputs indicate resistance and resilience of soils at a very broad scale, based on national scale data. This output could be tested by selecting sites and soils from each of the classes and testing whether the soils have the inherent range of properties that allocated it to that class in the first place, thereby testing the veracity of the original data and assessing if the responses of each soil fitted into the expected pattern, thereby testing the model.

In general, the models were best able to fit stability to copper, compression and overburden to soil properties, while there was no good fit between the soil properties and

stability to heat. The lack of correlation between stability to heat and the measured soil properties was also noted in a straightforward correlation matrix of the original data set (Kuan *et al.*, 2007). Although subsequent studies on the same soils did reveal a correlation between resistance to heat and the concentration of soil organic carbon, experiments indicated that both the quantity and quality of substrate in the soil influenced stability to heat (Griffiths *et al.*, 2008). The lack of a good fit for the models concerning stability to heat precluded subsequent mapping, apart from when stability to heat was included in combination with other perturbations (Table 1).

The expert interpretations of the stability to individual perturbations were the easiest to make. Thus, sandy arable soils were less biologically stable to copper than were the organic matter-rich upland soils, which is related to the effects of pH and organic matter on the biological availability of copper (Kuan *et al.*, 2007). Similarly, the clay-based arable soils were less physically stable to an overburden stress than were the organic matter-rich and coarse-textured soils because of previously described resilient properties of organic-matter and resistance of sand particles (Kuan *et al.*, 2007). Although these latter soils could be identified as more stable than the clay-rich soils, they were soils that were highly resistant and resilient but which had no common characteristics. Organic matter-rich soils were less resistant to compressive stress than were the coarse-textured soils. The models were able to identify patterns of stability to combinations of perturbations, although they showed that soils were not equally stable to all perturbations. In a comparison of biological stability to copper and physical stability to compression, the organic matter-rich upland soils were highly stable to copper but highly vulnerable (not resistant) to compression, while pasture or forest areas with a medium content of organic matter were moderately stable to both perturbations. To get a more holistic picture of soil stability incorporating information from all four perturbations, the overall descriptions become necessarily less precise and show that the cultivated soils are more vulnerable and organic matter-rich soils the least.

Multi-objective regression trees have been applied successfully in ecological modelling (Debeljak *et al.*, 2001, 2007, 2008; Džeroski, 2001; Jerina *et al.*, 2003) and the current study has shown the usefulness of this approach for the mapping of soil resilience from a limited database. Where pedotransfer functions based on multiple regression analysis are not possible or where the relationships between soil properties defining the functions are poor, MORTs should provide a valuable tool for risk-based mapping of soil resources. For certain processes, for example cadmium sorption by soil (Deurer & Bottcher, 2007), fairly reliable parametric predictions are possible through modelling, thus allowing detailed mapping. Many soil processes, however, depend on numerous interacting variables that vary over space and time; so, modelling is fraught with

uncertainty (Le Bissonnais *et al.*, 2002; Chirico *et al.*, 2007). Moreover, classification and mapping of multiple processes, for example the biological and physical resilience as modelled in this paper, is not possible. MORTs extend regression tree approaches that have been previously applied to mapping soil processes (Bishop & McBratney, 2001). Where large national soil databases are available such as the SSKIB database in Scotland, the machine learning techniques employed in MORTs offers a valuable tool for generating risk-based regional maps of potential soil threats.

## Conclusions

This study demonstrates the flexibility and applicability of a data mining approach to the problem of regression with multiple dependent variables, even if the data set is not very large (we had 26 samples to induce multi-objective models). The models achieve relatively good performance as judged by statistical validation processes. Subsequent expert evaluation proved the existence of internal logic in the structure of most of the induced models. It is demonstrated that models induced with data mining can be easily linked with techniques for analysing spatially related data using a GIS. The combination of both approaches gives high added value to the final results (e.g. risk-based soil maps) compared with partial approaches.

In this example, we mapped soil resistance and resilience to biological and physical stresses. Many soil properties influence resistance and resilience and simple relationships using multivariate methods were found for only a few properties. For instance, increasing organic matter content resulted in soils being more resistant to copper stress but less resistant to compaction. While the data set which we applied in this research was clearly not sufficient to cover all the soil types across Scotland, it does demonstrate that a combination of data mining and GIS techniques is a useful means to generate risk-based maps. The identification of the cultivated areas as being of low overall stability indicates that the original database on resistance and resilience of soils needs expanding to cover more soils and to utilize more background information from the GIS database to improve the accuracy of the maps.

## References

- Bishop, T.F.A. & McBratney, A.B. 2001. A comparison of prediction methods for the creation of field-extent soil property maps. *Geoderma*, **103**, 149–160.
- Blokeel, H. & Struyf, J. 2002. Efficient algorithms for decision tree cross-validation. *Journal of Machine Learning Research*, **3**, 621–650.
- Blokeel, H., De Raedt, L. & Ramon, J. 1998. Top-down induction of clustering trees. In: *Proceedings of the 15th International Conference on Machine Learning* (ed. J. Shavlik), pp. 55–63. Morgan Kaufmann, San Francisco, CA.

- Boorman, D.B., Hollis, J.M. & Lilly, A. 1995. *Hydrology of soil types: a hydrologically-based classification of the soils of the United Kingdom*. Institute of Hydrology Report No.126. Institute of Hydrology, Wallingford.
- Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. 1984. *Classification and regression trees*. Wadsworth Publishing, Belmont, CA.
- Chirico, G.B., Medina, H. & Romano, N. 2007. Uncertainty in predicting soil hydraulic properties at the hillslope scale with indirect methods. *Journal of Hydrology*, **334**, 405–422.
- Davidson, D.A. & Grieve, I.C. 2004. *Trends in soil erosion*. Scottish Natural Heritage Commissioned Report F00AC106. Scottish Natural Heritage, Edinburgh, 12 pp.
- Debeljak, M., Džeroski, S., Jerina, K., Kobler, A. & Adamic, M. 2001. Habitat suitability modelling for red deer (*Cervus elaphus* L.) in south-central Slovenia with classification trees. *Ecological Modelling*, **138**, 321–330.
- Debeljak, M., Cortet, J., Demsar, D., Krogh, P.H. & Džeroski, S. 2007. Hierarchical classification of environmental factors and agricultural practices affecting soil fauna under cropping systems using Bt maize. *Pedobiologia*, **51**, 229–238.
- Debeljak, M., Squire, G.S., Demšar, D., Young, M.W. & Džeroski, S. 2008. Relations between the oilseed rape volunteer seedbank, and soil factors, weed functional groups and geographical location in the UK. *Ecological Modelling*, **212**, 138–146.
- Deurer, M. & Bottcher, J. 2007. Evaluation of models to upscale the small scale variability of Cd sorption in a case study. *Geoderma*, **137**, 269–278.
- Džeroski, S. 2001. Applications of symbolic machine learning to ecological modelling. *Ecological Modelling*, **146**, 263–273.
- European Commission. 2006. *Soil protection – the story behind the strategy*. Office for Official Publications of the European Communities, Luxembourg (ISBN 92-79-02066-8).
- Griffiths, B.S., Hallett, P.D., Kuan, H.L., Gregory, A.S., Watts, C.W. & Whitmore, A.P. 2008. Functional resilience of soil microbial communities depends on both soil structure and microbial community composition. *Biology and Fertility of Soils*, **44**, 745–754.
- Horn, R., Fleige, H., Richter, F.H., Czyz, E.A., Dexter, A., az-Pereira, E., Dumitru, E., Enarache, R., Mayol, F., Rajkai, K., de la Rosa, D. & Simota, C. 2005. SIDASS project. Part 5. Prediction of mechanical strength of arable soils and its effects on physical properties at various map scales. *Soil & Tillage Research*, **82**, 47–56.
- Jerina, K., Debeljak, M., Džeroski, S., Kobler, A. & Adamic, M. 2003. Modeling the brown bear population in Slovenia – a tool in the conservation management of a threatened species. *Ecological Modelling*, **170**, 453–469.
- Jordan, C., Higgins, A. & Wright, P. 2007. Slurry acceptance mapping of Northern Ireland for run-off risk assessment. *Soil Use and Management*, **23**, 245–253.
- Kernan, M., Hughes, M., Hornby, D., Bennion, H., Hilton, J., Phillips, G. & Thomas, R. 2004. The Use of a GIS-based inventory to provide a regional risk assessment of standing waters in Great Britain sensitive to acidification from atmospheric deposition. *Water, Air, & Soil Pollution: Focus*, **4**, 97–112.
- Kuan, H.L., Hallett, P.D., Griffiths, B.S., Gregory, A.S., Watts, C.W. & Whitmore, A.P. 2007. The biological and physical stability and resilience of a selection of Scottish soils to stresses. *European Journal of Soil Science*, **58**, 811–821.
- Le Bissonnais, Y., Montier, C., Jamagne, M., Daroussin, J. & King, D. 2002. Mapping erosion risk for cultivated soil in France. *Catena*, **46**, 207–220.
- Lilly, A., Towers, W., Malcom, A. & Paterson, E. 2004. *Report on a workshop on the development of a Scottish Soils Knowledge and Information Base (SSKIB)*. Macaulay Land Use Research Institute. Available at: [http://www.macaulay.ac.uk/workshop/SSKIB/SSKIBWorkshop\\_Report.pdf](http://www.macaulay.ac.uk/workshop/SSKIB/SSKIBWorkshop_Report.pdf) (last accessed 23 January 2009).
- McBratney, A.B., Santos, M.L.M. & Minasny, B. 2003. On digital soil mapping. *Geoderma*, **117**, 3–52.
- Quinlan, J.R. 1986. Induction of decision trees. *Machine Learning*, **1**, 81–106.
- Seybold, C.A., Herrick, J.E. & Brejda, J.J. 1999. Soil resilience: a fundamental component of soil quality. *Soil Science*, **164**, 224–234.
- Struyf, J. & Džeroski, S. 2006. Constraint based induction of multi-objective regression trees. Knowledge discovery in inductive databases. In: *Fourth International Workshop, KDID'05, Revised, Selected and Inductive Papers* (eds F. Bonchi & J.-F. Boulicaut), Vol. 3933, pp. 222–233. Lecture Notes in Computer Science, Springer.
- Suttle, N.F., Bell, J., Thornton, I. & Agyriaki, A. 2003. Predicting the risk of cobalt deprivation in grazing livestock from soil composition data. *Environmental Geochemistry and Health*, **25**, 33–39.
- Tan, P., Steinbach, M. & Kumar, V. 2006. *Introduction to data mining*. Addison-Wesley, Boston, MA.
- Towers, W., Grieve, I.C., Hudson, G., Campbell, C.D., Lilly, A., Davidson, D.A., Bacon, J.R., Langan, S.J. & Hopkins, D.W. 2006. *Scotland's soil resource – current state and threats*. Scottish Executive Environment and Rural Affairs Department. Available at: <http://www.scotland.gov.uk/Publications/2006/09/21115639/0> (last accessed 23 January 2009).