RESPECTING HIERARCHICALLY STRUCTURED TAXONOMIES IN SUPERVISED IMAGE CLASSIFICATION: A GEOLOGICAL CASE STUDY FROM THE WESTERN CANADIAN SHIELD

E. M. Schetselaar [a,*] and C. F. Chung[b]

[a] Department of Earth Systems Analysis, International Institute for Geoinformation Science and Earth Observation, Hengelosestraat 99 7514 AE Enschede The Netherlands, schetselaar@itc.nl

[b] Geological Survey of Canada Geological Survey of Canada, Ottawa, 601 Booth St. Ottawa, Canada K1A 0E8, chung@NRCan.gc.ca

Commission IV, WG IV/1

KEY WORDS: Hierarchical, Knowledge Base, Classification, Integration, Data Mining, Geology, Geophysics

ABSTRACT:

Supervised image classification is based on assembling statistics between site-specific ground observations and remotely sensed measurements. If supervised image classification is applied within the context of a particular theme (e.g. vegetation, soil, lithology, land use), one is often confronted with extracting the statistical correlations from a hierarchically arranged network of taxonomic classes spatially abstracted and hierarchically generalized over a range of mapping scales. In practice, however, supervised image classification often appears to be based on a pragmatic approach, a priori categorizing the samples into classes from various levels or from a subset of the hierarchic network. Such approaches are suspect, since the sampling often appears to be biased towards maximizing the discrimination potential of the multivariate data set at cost of representing the categories identified by direct ground observation. The classification performance is, as a result, often assessed within the context of arbitrarily defined class schemas that only partly correspond to the schemas obtained by field surveys. Clearly, to gain more insight in how supervised classifiers are behaving with respect to ground observations, sampling procedures are required that respect the hierarchy of the taxonomy obtained in ground surveys. Herein we report the results of classification experiment applied to gamma-ray spectrometry and aeromagnetic data where samples are extracted at ground stations for each level of a four-level hierarchically arranged class network of bedrock lithology. The number of classes in this network ranges from n = 2 for the highest level and to n = 14 for its lowest. A number of classification experiments suggest that classification performance can be improved if the estimation of prior probabilities at a more detailed level in the taxonomy is conditioned by spatial patterns at more general levels in the taxonomy. This improvement in performance may even apply when such patterns are obtained by classification of the same data and sample set.

1. INTRODUCTION

Image classification methods have since the launch of the first earth orbiting remote sensing platforms been routinely applied for mapping land cover and natural resource themes. Regardless the application, one of the most complicating factors inherent to the image classification problem, is the incongruence between what and how physical properties are measured and how geospatial objects are identified and classified through the eyes of the geoscientific expert. Even when the sensing device is an optical system sensitive to the wavelength range of human vision, human cognition is rarely based on spectral properties alone. For example, it is well known to earth scientists that colour (i.e. perceived relative spectral properties between 400 and 700 nm) is often one of the least diagnostic and most misleading criteria for classifying a particular mineral or rock type (as it is frequently a function of the abundance of trace elements in the crystal lattice respectively state of weathering).

Clearly, image classification is no exception to any other application of multivariate statistics, namely to explore the universe between what an expert sees and categorizes with what is measured. Unique to the image classification task, however, is the uncertainty to what extent observed features match patterns in fields of measurements. In most multivariate statistical applications there is an unequivocal relationship between the category and the set of measurements on the category. In remote sensing applications, however, the object of interest "blends-in" with its immediate neighbourhood and the criteria used to classify may vary from general to detailed over a range of scales within a taxonomy of information classes. A geological example of class inheritance of geospatial objects over such a hierarchy would be: terrane, subterrane, lithotectonic domain, lithologic assemblage, plutonic suit, granitoid pluton, granite, monzogranite. A large variety of criteria are employed within such a hierarchy based on a multitude of geoscience themes, such as: regional tectonic synthesis, geochronology, lithogeochemistry, mineralogy, texture and structure.

The problem is that it is unclear, how the above criteria are related to the patterns in the remote sensing measurements that provides the discrimination potential. Any attempt to assign ground based information classes to measurements by multivariate statistical methods, therefore, should not only be based on samples of information classes at a particular level within a taxonomy, but also exploit the a priori knowledge on how the information classes are ordered within it. This paper presents image classification methods and experiments that respect the taxonomic knowledge of earth scientists by which geospatial objects have been classified.

2. A STRATIFIED BAYSIAN IMAGE CLASSIFICATION ALGORITM

In general form, the posterior probability that an m-dimensional vector of measurement X at pixel p belongs to class C_j can be expressed in Bayesian formula:

$$P\{C_j \mid X_l\} = \frac{P\{X_l \mid C_j\}P\{C_j\}}{P\{X_l\}} \qquad (1)$$

The term $P\{ X_l \mid C_j \}$ in this expression is the probability that the measurement vector will take on the value X_l given that the pixel is a member of class C_j. This probability can be determined by sampling a population of measurement vectors for pixels known to be a member of class C_j. In practice, because of the limited availability of measurement vectors with known class membership, this probability is estimated by assuming a particular distribution, such as the multivariate normal or the multivariate student t-distribution (McLachlan, 1992). $P\{X_l\}$ in expression 1 is the probability of the occurrence of measurement vector X_l. The term $P\{C_j\}$ in expression 1 is the prior probability that a pixel will be a member of class C_j. This probability is estimated by computing the mixing proportions of the total number of samples N_j of class C_j to the total number of samples over all the classes.

Dependent on the assumption for the distributions to estimate $P\{X_l \mid C_j\}$, a classification decision rule is defined whereby the pixel p is allocated to the class with the highest posterior probability, provided that it is above a threshold, below which p is assigned the label "unclassified". Assuming, for example, multivariate normal distributions for the classes and including estimation of priors, expression 1 can be written as:

$$P\{C_j \mid X_l\} = \frac{f_j(X_l)P\{C_j\}}{P\{X_l\}} \qquad (2)$$

Where $f_j(X_l)$ is the probability density function of a multivariate normal distribution for class C_j. Using a Maximum likelihood classifier, a decision rule can be formulated substituting class variance-covariance matrices V_j and class mean vectors m_j computed from the sample set to parameterize multivariate normal class distributions:

DR1: Choose j that minimizes:

$$\ln|V_j| + (X_l - m_j)'V_j^{-1}(X_l - m_j) - 2\ln P(C_j) \qquad (3)$$

The incorporation of prior probabilities in the decision rule led some workers to the development of methods to improve on the estimation of prior probabilities. Strahler (1980) showed how thematic classes from one or more additional variables can be used to refine the estimation of prior probabilities. Gorte (1998) extended such concepts to non-parametric classification methods (based on the k-nearest neighbour method) and used iterative estimation of class mixing proportions to obtain local priors. In this paper, the estimation of prior probabilities is based on a stratified classification over a taxonomy based on geoscientific knowledge of the study area. Thematic maps or classified patterns at general levels within this hierarchy are used as collateral variables to estimate prior probabilities for the classification at a more detailed level within the hierarchy. The classification is called stratified, because it proceeds stepwise from general to detail over the taxonomy. Distinctive properties of the classification problem at more general levels in the hierarchy leads to the formulation of assumptions on the sample set and spatial distribution of classes that can be exploited to improve classification performance at lower levels in the hierarchy.

We start by redefining C_j, the class membership of the j^{th} class, as C_j^k : the class membership of the j^{th} class at the k^{th} level in a taxonomy ranging from [k-q, …,k-1 , k], where q {1, .. , k-1} is an index variable between referring to the level above the k^{th} level in the taxonomy. We rewrite expression 1 and define the posterior probability that a pixel will be a member of class C_j^k given X and the fact that p is a member of the i^{th} class at the $(k-1)^{th}$ level:

$$P\{C_j^k \mid X_l, C_i^{k-1}\} = \frac{P\{X_l \mid C_i^{k-1}, C_j^k\}P\{C_i^{k-1}, C_j^k\}}{P\{X_l, C_i^{k-1}\}} \qquad (4)$$

If the j^{th} class descends directly from one and only one class at the $(k-1)^{th}$ level, then it can be shown that:

$$P\{X_l \mid C_i^{k-1}, C_j^k\} = P\{X_l \mid C_j^k\} \qquad (5)$$

If, instead, there are multiple inheritances in the taxonomy (e.g. a class may descend from more than one superclass) expression 5 is the assumption stating that the class probability densities do not vary with the class membership at the $(k-1)^{th}$ level. This assumption states that $P\{X_l \mid C_j^k\}$ e.g. the distribution of the measurement vectors for a particular class C_j is invariant with the classes C_i^{k-1} at higher levels in the taxonomy. This assumption is violated, for example, when the measurement vectors of classes with limited spatial extent are affected by different 'background' distributions of measurement vectors at more general levels in the taxonomy. Another example where this assumption is violated would be chemical alterations affecting X_l (by metamorphic or metasomatic processes) conditioned by the superclass in which a particular class is enclosed.

The joint probability $P\{C_i^{k-1}, C_j^k\}$ in expression 4 can be rewritten in the form of a conditional probability:

$$P\{C_i^{k-1}, C_j^k\} = P\{C_j^k \mid C_i^{k-1}\}P\{C_i^{k-1}\} \qquad (6)$$

Substituting expressions 5 and 6 into 4 we obtain:

$$P\{C_j^k \mid X_l, C_i^{k-1}\} = \frac{P\{X_l \mid C_j^k\} P\{C_j^k \mid C_i^{k-1}\} P\{C_i^{k-1}\}}{P\{X_l, C_i^{k-1}\}} \qquad (7)$$

Since $$\sum_{j=1}^{J} \frac{P\{X_l \mid C_j^k\} P\{C_j^k \mid C_i^{k-1}\} P\{C_i^{k-1}\}}{P\{X_l, C_i^{k-1}\}} = 1,$$

$$P(X_l, C_i^{k-1}\} = \sum_{j=1}^{J} P\{X_l \mid C_j^k\} P\{C_j^k \mid C_i^{k-1}\} P\{C_i^{k-1}\}$$

So that $P\{C_i^{k-1}\}$ cancels from de numerator and denominator after substitution in expression 7, we obtain:

$$P\{C_j^k \mid X_l, C_i^{k-1}\} = \frac{P\{X_l \mid C_j^k\} P\{C_j^k \mid C_i^{k-1}\}}{\sum_{j=1}^{J} P\{X_l \mid C_j^k\} P\{C_j^k \mid C_i^{k-1}\}}$$

$$= \frac{f_j(X_l) P\{C_j^k \mid C_i^{k-1}\}}{\sum_{j=1}^{J} f_j(X_l) P\{C_j^k \mid C_i^{k-1}\}} \qquad (8)$$

The term $P\{C_j^k \mid C_i^{k-1}\}$ in expression 8 can be directly computed from random samples that are hierarchically stratified into k levels.

Accordingly, the classification decision rule can be states as:

DR2: Choose j that minimizes:

$$\ln |V_j| + (X_l - m_j)' V_j^{-1} (X_l - m_j) - 2 \ln P(C_j \mid C_i^{k-1}) \qquad (9)$$

Note that because of the assumption stated in expression 5, this expression could have simply been obtained by substituting the conditional prior into expression 3. If this assumption is not valid, however, the joint conditional probabilities of the form $P\{X_l \mid C_i^{k-1}, C_j^k\}$ must be computed from all the occurring combinations between j classes at the k^{th} level and i classes at the $(k-1)^{th}$ level.

3. A CASE STUDY PREDICTION OF UNITS FROM A BEDROCK TAXONOMY OF THE WESTERN CANADIAN SHIELD

The above-derived method was tested in a number of classification experiments to predict a 4-level taxonomy of bedrock units from gamma-ray spectrometry and aeromagnetic data gridded on 100 x 100 meter pixels acquired over the western margin of the exposed Canadian Shield. The five grid (image) variables are: K: potassium channel (fig. 1a), eTh,: thorium channel (fig. 1b), eU uranium channel (fig. 1c), total magnetic field and residual magnetic field (fig 1dl). These five grids were augmented by their 9 x 9 average filtered derivatives, in analogy to the approach of Switzer (1980). This method exploits the spatial autocorrelation between pixels within each of the variables under the assumption that the alternation of bedrock units occurs on a large scale than that of a single pixel (Switzer, 1980). The statistical relationships between the image variables (Figure 1) and bedrock units was estimated at the field stations of several geological mapping projects, yielding 3528 samples, each having values of the K, eTh, eU, total magnetic field and residual magnetic field. The bedrock taxonomy is shown in Figure 2.

In a previous study samples of bedrock units were amalgamated from the second and third level of this taxonomy (Schetselaar et al., 2000). This classification coincided for 70% to geological maps of the study area compiled on 1 : 50.000 scale (McDonough et al., 2000 and references therein). As can be seen from Figure 2 the relationships between the first, second and third levels within this taxonomy are defined by single inheritance. The relationships between the fourth and third level (between mylonitic units and protoliths) however is defined by multiple inheritance. These were forced to single inheritance by masking the shear zones with the mylonitic units from the three levels above it. This was necessary because the relationships between original rock units (protoliths) and mylonitic units could not be recovered from the digital map database of the study area. The number of classes in this network ranged from n = 2 for the highest level and to n = 14 for its lowest. The hierarchic network was structured downwards according to lithotectonic domains (n = 2), basement-cover-plutonic assemblages (n = 4), bedrock units (n = 12 & n = 14). The bedrock classification was based on mineralogical, textural and structural field diagnostics.

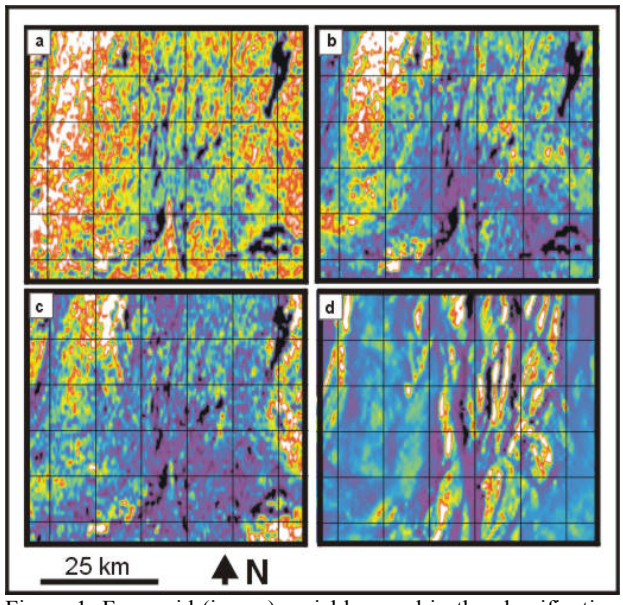

Figure 1. Four grid (image) variables used in the classification experiments, a) Potassium Channel (K); b) Thorium Channel (eTh); c) Uranium Channel (eU); residual magnetic field grid.

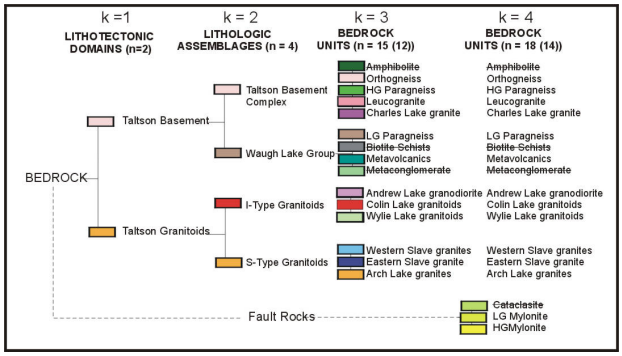

Figure 2. Bedrock taxonomy of the Taltson magmatic zone (Canadian Shield of NE Alberta) after McDonough et al. 2000 (and references therein). Classes that are striped out do not have enough samples (n < 10) for estimating the class distributions.

At the fourth level two mylonitic units were added to the third level. This additional differentiation was made because the fabric and mineralogy within these units are altered by high shear strains to an extent that they their field diagnostics are not representative to their undeformed equivalents at the third level.

First a standard Bayesian classification was applied to all levels separately, computing overall priors from the samples. The classified map patterns and their coincidence with compiled geological maps resulting from these classification experiments are shown in Figure 2a-h respectively Figure 3. Note that the fall-off rate of the coincidence percentages for the classes is higher than the overall percentages. This is obviously due to the large bias towards the classes of great spatial extent at the first and second levels of the taxonomy. The average class coincidence percentage are rapidly decreasing with increasing number of classes, ranging from 80 at the first level to 40 percent at the fourth level.

Next, two types of stratified classification experiments were conducted were class information at more general levels conditioned the classifications at more detailed levels of the taxonomy:

1. The computation of priors at k=2 and k=3 conditioned by units of map patterns at k=1 and k=2 of the taxonomy. This experiment is comparable to situations where detailed bedrock lithology classification is conditioned by regional geological maps showing lithotectonic domains and lithologic assemblages. This scenario is considered realistic in reconnaissance mapping projects where regional geological maps (typically between 1:250,000 and 1,000000 scales) are available.

2. The computation of priors at k=2 and k=3 conditioned by units of a map pattern obtained from a non-parametric classification method at k=1. Note that in this case no additional map layers or interpretations are used in the classification. The stratification is based on the same sample set used for the non-stratified classification experiments and proceeds stepwise from general to detailed levels in the taxonomy.

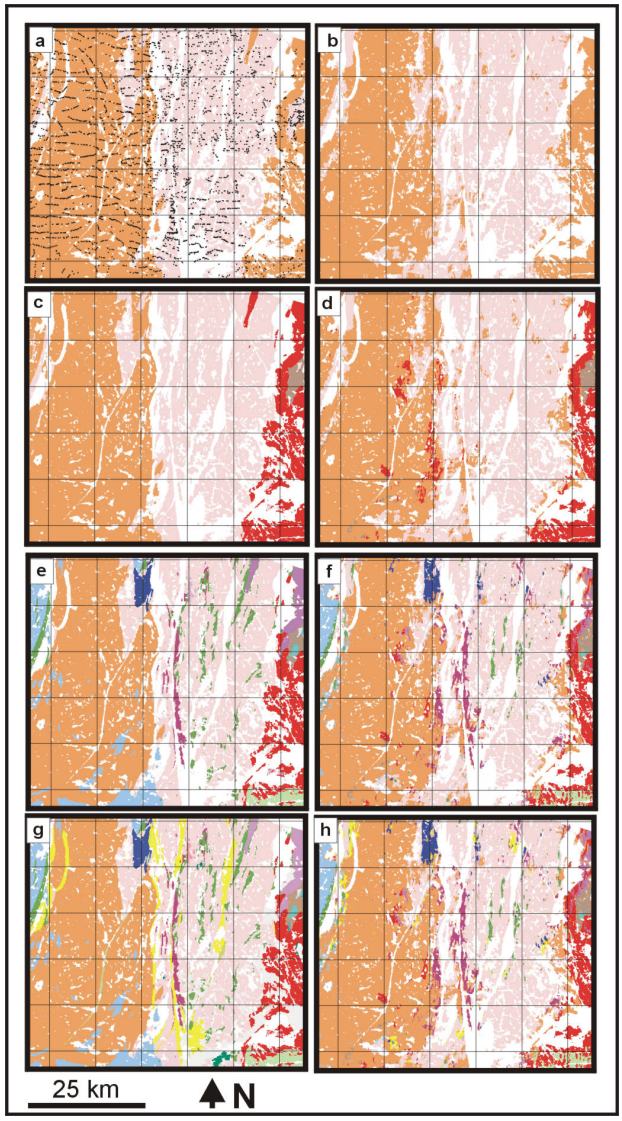

Figure 3. Classified patterns and image variables. (a) map compilation lithotectonic domains with overlay of sample locations; (b) classification lithotectonic domains; (c) map compilation lithologic assemblages; (d) classification lithologic assemblages; (e) map compilation bedrock units; (f) classification bedrock units; (g) map compilation bedrock units (including fault rocks); (h) classification bedrock units (including fault rocks).

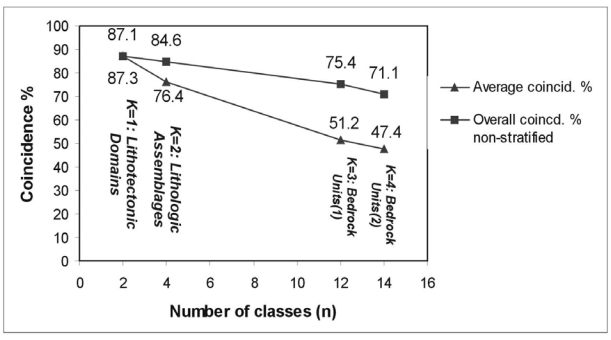

Figure 4. Overall and average coincidence percentages between classifications and map compilations for the four levels of the bedrock taxonomy.

The classified map patterns and coincidence percentages resulting from these experiments are shown in Figure 5 and Figure 6 respectively. The first method obviously appears to yields a considerable increase of the coincidence percentages (ca. 10%). This initial result suggests that it may be useful to integrate regional scale maps in a stratified classification approach. Alternatively, spatial patterns representing lithotectonic domains and lithologic assemblages can be outlined on remotely sensed data. Such regional units, for example, can often be easily outlined using anomaly patterns, texture and shape on colour-enhanced grid representations of aeromagnetic and gravity data, whereas it is often very difficult to use such grid representations to assign individual anomalies to particular bedrock units.

The second method was based on non-parametric estimation of the class probability distributions for lithotectonic domains at k=1. It appeared that the Maximum Likelihood classification at k=1 did not result in an increase of coincidence percentage at k=2 and k=3. Apparently the number of misclassified pixels of large units at k=1 was not compensated by reduction of overlap of class probability distributions or refinement in the estimation of priors at lower levels. Only in situations where the number of misclassified pixels is low with respect to the misclassified pixels due to overlap between class probability distributions at higher k, an improvement in classification performance is to be expected.

An attempt was made to improve the classification of lithotectonic domains at k=1, exploiting the following two characteristics of the classification problem at general levels (e.g. k=1 or k=2) of the bedrock taxonomy:

1. The large number of samples available ($n_{\text{Taltson Basement}}$ = 1708 and $n_{\text{Taltson Granitoids}}$ = 1512) at k=1 well spread over the entire study area. This permits direct estimation of $P\{C_j|X_l\}$ from the samples instead of estimation under the assumption of multivariate normal distribution.

2. The fact the lithotectonic domains form units of spatial dimensions at least two orders of magnitude larger with respect to the pixel size of 100 x 100 metres. This permits extensive post-classification smoothing without the risk of eliminating classes or to exploit the spatial distribution of the samples themselves in the classification.

In theory direct estimation could proceed by cross labeling all combinations between the five image variables. In practice, however, this results in computationally inhibitive in allocating memory for all unique combinations between the five image variables. (for this dataset stored in 8 bits, it would require 255^5 =1.0782 x 10^{12} combinations). Although the combinations could be evaluated over larger bin intervals, we selected an alternative method by evaluating the probabilities over each variable separately before their multiplication, assuming that they are conditionally independent. Using this approach a classification decision rule is stated as follows:

DR3: Choose j that maximizes:

$$P\{C_j\}\prod_{m=1}^{M} P\{X_l^m \mid C_j\} \qquad (10)$$

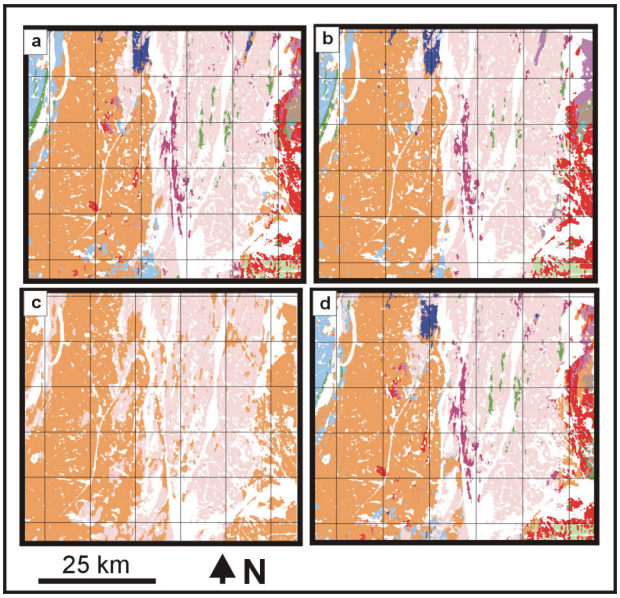

Figure 5 Results of stratified image classification experiments. a) classification bedrock units stratified on map compilation of lithotectonic domains; (b) classification bedrock units stratified on map compilation of lithologic assemblages; (c) non-parametric classification lithotectonic domains; (d) classification bedrock units stratified on classification lithotectonic domains (figure 5c).

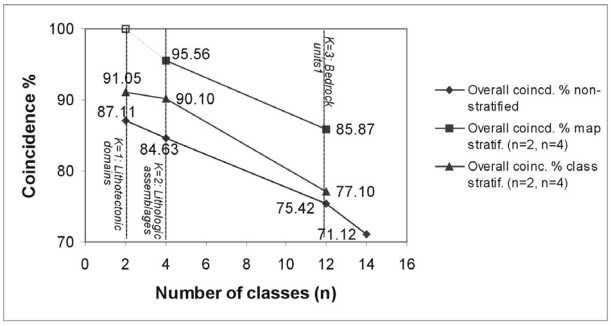

Figure 6. Coincidence percentages with geological map compilations for non-stratified Bayesian classification, for Bayesian classification stratified over map patterns at k=1 and k=2., for Bayesian classification at k=2 and k=3 stratified over non-parametric classification at k=1.

Because the lithotectonic domains are large contiguous units and the samples are evenly distributed, distance to samples was used as an additional variable to improve the discrimination potential of the multivariate dataset. By considering M variable, and computing coincidence percentages for all 20 combinations, it was found that the distance, potassium and magnetic grids provided the highest coincidence percentage (91%) improving the prediction of bedrock units with ca. 4 percent (figure 5c). The propagation of this classification result to levels k=2, and k=3 for computing priors, yielded an increase in coincidence with the geological map compilations of respectively 5.5 and 1.7 percent (figure 5d). Alternative non-parametric methods, such as the k-nearest neighbour classifier or algorithms based on artificial neural nets will be tested in the future to investigate if similar or higher classification performances can be obtained for the classifications at k=1 and k=2.

4. CONCLUSIONS

Initial results of a number of classification experiments suggest that classification performance can be improved by the estimation of prior probabilities at a particular level of a taxonomy that is conditioned by a more general level of the same taxonomy. In comparison to conventional approaches, the Bayesian stratified classification method provides mechanisms to introduce spatial data at more general levels in the classification, to which users have often better access or which can be easier derived through visual or automated image interpretation. An increase in classification performance may be obtained even when no additional data or visual interpretations are introduced, and spatial patterns associated to general levels in the taxonomy are also obtained by image classification. In addition the user can adapt algorithms and combination of image variables to each level within the hierarchy. This enhances the potential to adapt the classification methods to available map data and better exploit the intrinsic hierarchical structure of field knowledge.

REFERENCES

Gorte, B, 1998, *Probabilistic Segmentation of Remotely Sensed Images*, PhD thesis, ITC, 143 pp.

McDonough, M.R.M. McDonough, M.R.M., McNicoll, V.J. and Schetselaar, E.M., 2000, Age and kinematics of crustal shortening and escape in a two-sided oblique-slip collisional and magmatic orogen, Palaeoproterozoic Taltson Magmatic Zone, North-eastern Alberta, *Canadian Journal of Earth Sciences, special issue LITHOPROBE Alberta Basement Transect* 37(11), pp. 1549-1573.

McLachlan, G.LJ., 1992, *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, New York, 300 pp.

Schetselaar, E.M., Chung, C.F., Kim K., 2000. Classification of bedrock units in vegetated granite-gneiss terrain by the integration of airborne geophysical images and primary field data. *Remote Sensing of Environment* 71, pp. 89-105.

Strahler, A.H., 1980, The use of prior probabilities in maximum likelihood classification of remotely sensed data, *Remote Sensing of Environment* 10, pp. 135-163.

Switzer, P, 1980, Extension of linear discriminant analysis for statistical classification of remotely sensed satellite imagery, *Mathematical Geology* 12, pp. 367-376.