COMPARISON OF DIFFERENT INTEGRATION TECHNIQUES OF ANCILLARY INFORMATION INTO AN OBJECT-BASED CLASSIFICATION PROCESS

M. Förster [a, *], B. Kleinschmit [a]

[a] Berlin University of Technology, Department of Landscape Architecture and Environmental Planning, Str. d. 17. Juni 145 (EB 5), 10623 Berlin, Germany; michael.foerster@tu-berlin.de, birgit.kleinschmit@tu-berlin.de

Commission I, WG I/5 and WG IV/3

KEY WORDS: ancillary information, QuickBird, geo-factors, significance analysis of microarrays, SAM, cluster analysis

ABSTRACT:

As a contribution to a better understanding of the integration of different data sets into knowledge-based classification processes the presented study compares different data-fusion techniques which combine spectral and textural information of a QuickBird scene with ancillary data layers and a knowledge base for the identification of forest structures and habitats. The approach compares a fuzzy logic classifier and a crisp rule-based integration technique. Both methods combine spectral and textural information with ancillary data-layers and a knowledge base. For the fuzzy logic approach, the possibility of assignment for each object is combined with a fuzzy knowledge base, which consists of information about the possibility of existence of tree species and geo-factors for a GIS-database consisting of slope, aspect, and height of a medium resolution DTM, soil type, and forestry site maps. Therefore, for all woodland species of this specific region, an index of location factors was developed. A fuzzy set for each class concerning each geo-factor is defined, containing membership functions.

In order to develop a better understanding of the integration patterns of single geo-factors a significance analysis was carried out. This was done by an ISODATA Clustering and a significance analysis of microarrays (SAM), which is abundantly applied in Bioinformatics. A sequence of classifications with various applied geo-factors for different classes was performed. Accuracy assessments based on test samples were calculated and the results arranged in a microarray. The results from the ISODATA Clustering and the SAM showed a high significance of the combined utilization of all additional geo-factors. Moreover, classes with smaller percentages of covered area and smaller ecological niches depend more on the application of ancillary information than other forest types. Additionally, two types of geo-factors were detected as insignificant, which is due to an inappropriate data quality of a soil-map and the lack of significance of the geo-factor aspect.

1. INTRODUCTION

Additional geo-data and knowledge is widely used in remote sensing classification processes (Maselli et al., 1995). Often the integration is performed via fuzzy-logic approaches (Stolz and Mauser, 1996). In the case of most environmental knowledge-based classification systems the terms of assignment of a class to a geo-factor are formulated in a consciously uncertain linguistic term. Therefore a way of describing this uncertainty by smooth transitions between classes seems a reasonable approach of enhancing the accuracy of the results.

However, the influence of a single geo-data rule set on the classification accuracy is not sufficiently examined. Moreover, the effect of fuzzy rules in comparison to the same rules applied as crisp rule sets has to be evaluated given the intensive effort to develop a fuzzy rule-base. Therefore the presented approach compares fuzzy to non-fuzzy geo-data integration approaches as well as the significance of single knowledge to the classification accuracy with the help of a significance analysis of microarrays (see Figure 1).

The analysis was carried out in the natural environment of forest structures and habitats. The example of forest classification is especially suitable for showing chances and challenges of data integration techniques, because a long-term information about silvicultural practices and ecological woodland development is available together with a good (geo)database provided by the local forestry administration.

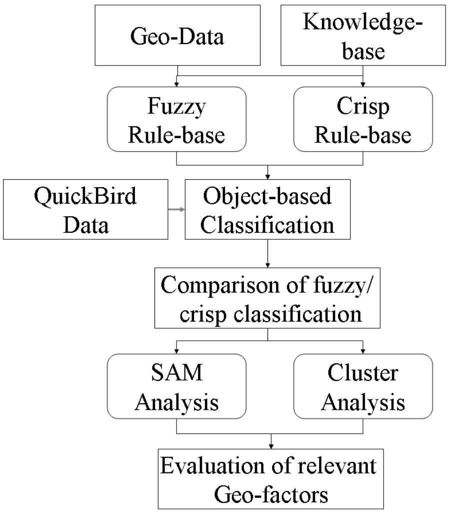

Geo-Data
Knowledge-base
Fuzzy Rule-base
Crisp Rule-base
QuickBird Data
Object-based Classification
Comparison of fuzzy/ crisp classification
SAM Analysis
Cluster Analysis
Evaluation of relevant Geo-factors

Figure 1. Evaluation scheme of the integration of ancillary data

* Corresponding author

For the purpose of data-integration additional information about the natural behaviour of forest structure within classification processes are used. Especially parameters concerning potential natural forest locations, such as elevation, aspect, wetness or soil acidity, were taken into account.

2. DATA AND METHODS

For the presented approach the satellite data were delineated in a multi-scale segmentation process (Burnett and Blaschke, 2003). This task was performed in an object-oriented approach using the software eCognition (Benz et al., 2004). The segmentation levels of different resolutions were delineated and assigned to hierarchical organized groups of objects, such as forest habitats, crown combinations and crown types of single tree species and combined with a fuzzy knowledge-base (for further details on the Chapter 2.1 and 2.2 see Förster and Kleinschmit 2006).

2.1 Data

For the presented investigation QuickBird data were used as VHSR satellite image. In summer 2005 data were acquired from the forested site called “Angelberger Forst”, which is situated in the pre-alpine area of Bavaria, Germany. Within this area, there are different semi-natural mixed forest types, including Beech forests, Alluvial forests types, such as Black Alder, Spruce, and a smaller amount of Larch and Afforestation of different woodland types. The scene was acquired on 11.08.2005 and had a cloud coverage of 10 % and an off-nadir angle of 11.3 degree.

As a geo-database a digital terrain model (DTM 5 and DTM 25), a conceptual soil map (CMS; 1 : 25.000), a forestry site map (FSM) as well as a silvicultural site map (SSM) were used. Especially the forestry maps consist of several attributes (e.g. soil type, water regime) which were considered separately. For the DTM, several standard parameters, such as slope, aspect and different forms of curvature (Dirnböck et al., 2003) were derived. The knowledge base to build up rule sets for potential forest types were available from a previous project in cooperation with the Bavarian State Institute of Forestry (Förster and Kleinschmit, 2006; Kleinschmit et al., 2006). These rules were complemented by silvicultural rules attained from local forest rangers and silvicultural literature (Walentowski et al., 2004). A register of location factors was developed, consisting of the DTM parameters, soil type from a conceptual soil map, and available water, soil substrate, and availability of nutrients from a forestry site map. Therefore, for all woodland species of this specific climatic region, an index of location factors was developed based on knowledge about Bavarian woodland types (Walentowski et al., 2005).

The QuickBird image was geometrically corrected and the pan-sharpening of the original data to a merged resolution of 0.6 m was performed (Zhang, 2002). The scenes were segmented at three landscape scales, but only the single tree / small tree group level (Scale Parameter (SP) 15, shape factor 0.1, compactness 0.5) was used for the methodological approach of investigating the influence of the geo-factors.

2.2 Classification with Fuzzy Rules

For the implementation of additional knowledge an input membership function and an output membership function were defined for each geo-factor. These membership functions were combined with the help of a rule base. For each object, the rule-base is defuzzicated by the weighted means method (Tilli, 1993). This possibility can be graphically visualized and implemented in the fuzzy-classification base of eCognition for all possible classes.

In combining the fuzzy sets and the hierarchical classification results the approach uses the minimum (AND-) rule, which specifies that the most unacceptable factor is the critical value for the forest type to occur. In a next step the minimum possibility of each possible class will be compared. The class with the highest membership will be assigned to the object.

2.3 Classification with Crisp Rules

In contrast to the fuzzy implementation a crisp rule-base was established. This classification is using an identical rule set for the additional data. However, the rules were integrated via thresholds (see figure 2 for comparison). As threshold, the weighted mean of the fuzzy rule was used.

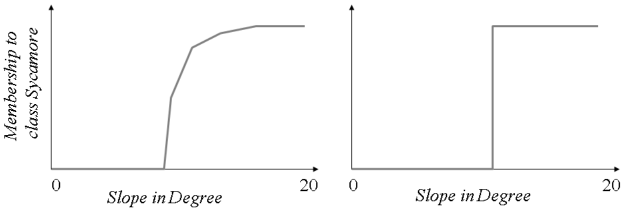


Figure 2. Exemplary fuzzy membership (left) and threshold integration (right) of the class Sycamore to Slope

2.4 Accuracy Assessment and Significance Analysis

The results of the classification processes including different sets of ancillary information were compared to test samples taken from silvicultural maps, field work and aerial photographs (see table 1). However, the comparison of fuzzy to non-fuzzy classifications was calculated only when including all geo-factors, due to the high significance of this combination.

Classification type	**Number of Variations**
All geo-factors included	1
All geo-factors excluded (classification with all or single spectral bands)	5
Single geo-factors included in all classes	6
Single geo-factors excluded from all classes	6
Single geo-factors included in single classes	32
Sum of performed classifications	**50**

Table 1. Classification setting for significance analysis of the additional geo-information

The different classification accuracies for single classes and the overall accuracy have to be evaluated in order to find the most significant geo-factor or the combination of geo-factors which were included. This can be done in ordering the results from highest to lowest accuracy. Although the ordering of the results give a good overview, the significance of the single accuracy of classes would be neglected. Moreover, the combination of the different geo-factors could interfere with each other, obscuring certain accuracy results or indicating only random fluctuations. Therefore the significance of the accuracies was evaluated by ISODATA clustering and by a Significance Analysis of Microarrays (SAM).

2.4.1 ISODATA Cluster Analysis: The cluster analysis is suitable for detecting patterns in multidimensional arrays. Therefore the 50 accuracies for each of the eight defined classes were classified in order to find clusters of significance. The calculation involved three to seven clusters, while the results of five derived clusters were best to interpret.

2.4.2 Significance Analysis with Microarrays (SAM): Microarrays are frequently used in Bioinformatics and in medical research mainly for pattern research of large gene experiments (Li et al., 2007). Hence, methods were developed to determine the significance of changes within the arrays. One of these techniques is the significance analysis of microarrays, which assigns a score to each data set on the basis of change in the results relative to the standard deviation of the measurements (Tusher et al., 2001).

Instead of genes, the different variations of the accuracy assessment for test data were used. To account fluctuations which are specific to a data set, a statistic is defined based on the ratio of change in data expression to standard deviation in the data for a certain geo-factor. The ‘‘relative difference’’ d(*i*) is:

$$d(i) = \frac{\overline{x}_I(i) - \overline{x}_U(i)}{s(i) + s_0} \qquad (1)$$

where $x_I(i)$ and $x_U(i)$ are defined as the average levels of expression for data set (*i*) in states I and U, respectively. s(*i*) is the standard deviation of repeated accuracy measurements under different inclusions of geo-factors. Repeated permutations of the data are used to determine if the expression of the data sets are significantly related to the response (in this case classification accuracy). From the permutations a expected relative difference $d_E(i)$ is calculated. To identify potentially significant changes in expression, a scatter plot of the observed relative difference d(*i*) vs. the expected relative difference $d_E(i)$ is used (see figure 4). The classification accuracy data which are represented by points displaced from the $d(i) \cong d_E(i)$ line by a distance greater than a self defined threshold are called positively (above the line) and negatively (below the line) significant.

3. RESULTS AND DISCUSSION

Pure spectral classification and classifications with ancillary data integrated via thresholds or fuzzy logic were compared class by class and for overall accuracy.

In a second step the fuzzy-logic based classifications with the various geo-factors included were evaluated. The results were assessed with the software R (Venables and Smith, 2006) and SAM (Chu et al., 2007).

3.1 Classification with and without Fuzzy Logic

The outputs of the compared classifications indicate a strong influence of the integrated ancillary information. The classification accuracy in all classes is highest when incorporating the rules via fuzzy logic (see Table 2). The classification with additional data of thresholds obtained better results than the spectral classification in some classes (e.g. Black Alder). However, the major improvement of including additional data and knowledge is only seen, when the cognitive uncertainties and the imprecise formulation of the information are acknowledged via fuzzy logic. Especially the detection of species with small ecological niches is improved when using fuzzy logic. With pure classification a tree type such as Black Alder is not distinguishable spectrally from other deciduous forest while showing the highest classification accuracy with fuzzy-logic knowledge integration.

Forest Type	**Ancillary Data Fuzzy Logic**	**Ancillary Data Crisp rules**	**Pure Classification**
Beech	0.81	0.74	0.75
Beech – young	0.32	0.18	0.15
Spruce	0.74	0.74	0.74
Spruce – old	0.42	0.32	0.32
Black Alder	0.98	0.34	0.17
Afforestation	0.97	0.95	0.95
Larch	0.96	0.59	0.59
Sycamore	0.88	0.72	0.68
Overall Accuracy	**0.77**	**0.70**	**0.64**

Table 2. Accuracy assessment for tree-type species for a pure classification in comparison to classifications with ancillary data of crisp rules and fuzzy-logic rules.

Nevertheless, the overall accuracy for most tree-types is certainly not sufficient for a reliable classification with VHSR-data. A further improvement is possible with careful integration of other additional data (such as LIDAR data) and a more efficient usage of silvicultural site conditions and the knowledge about it.

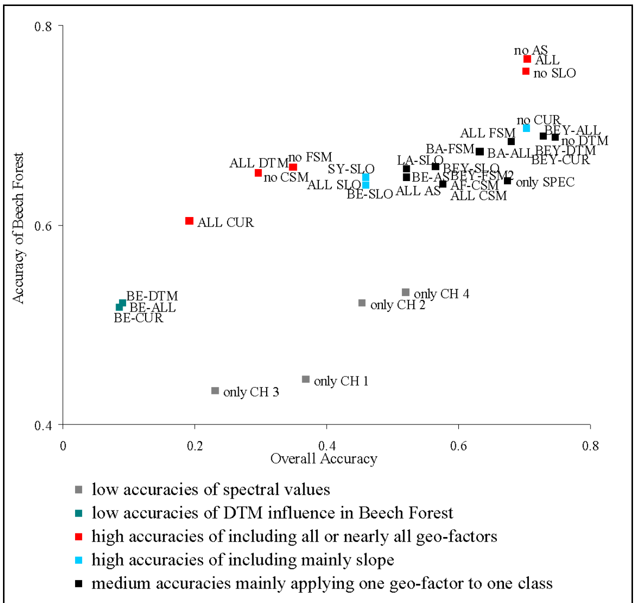
3.2 Significance of the Geo-Factors

The ISODATA cluster analysis of the accuracy results was used to find coherent patterns. An exemplary scatterplot of Beech Forest Accuracies and Overall Accuracies is shown in Figure 3 for five clusters. The first detected pattern is a cluster of low accuracies, when applying only a spectral classification and not using all spectral bands. Even the classification calculated with all spectral bands (named “only SPEC” in figure 3) is only assigned to the medium accuracy cluster.

One cluster (marked in red) identified classes with high accuracies, when all or nearly all geo-factors were included (names like “no AS” or “no SLO” mean that just Slope or Aspect were not integrated in the classification process). A very similar cluster (blue) shows high accuracies, when mainly slope was included in the classification process. Hence, it can be stated that including the whole set of rules and geo-factors improves the classification. Moreover, the factor Slope seems to be very distinct in classification accuracy response.

Another discrete cluster demonstrates very low accuracies when the DTM is included in the classification of Beech Forest. This feature still has to be evaluated, but it could indicate that either the rules are not well evaluated or that the terrain is not influencing the occurrence of Beech at all. Therefore, these rules might be excludable.

Most of the classification accuracies, however, are within a cluster called medium accuracies which mainly apply one geo-factor to one class (e.g. geo-factor slope to class Black Alder).



Abbreviations:
ALL...all geo-factors or all rules of a geo-factor
no...a certain geo-factor is not included
only CH1, CH2, CH3, CH4...one spectral band
only SPEC... all spectral channels are included
CUR...Curvature geo-factors are included
DTM...DTM geo-factors are included
SLO...Slope geo-factor is included

AS...Aspect geo-factor is included
CSM...Conceptual Soil Map geo-factor
FSM...Forestry Site Map geo-factors
BE...Beech Forest
SY...Sycamore
BEY...Beech (young)
BA...Black Alder
LA...Larch

Figure 3. Exemplary scatterplot of the ISODATA clustering results for accuracy of Beech Forest with the Overall Accuracy. The clustering was performed with five classes.

It can be concluded that the cluster analysis can detect interpretable patterns of the influence of additional data. Unfortunately, the ISODATA clustering could not give a detailed differentiation of the influence of single classes or geo-factors (one medium class). Moreover, this technique provides little information about the statistical significance of single classification approaches.

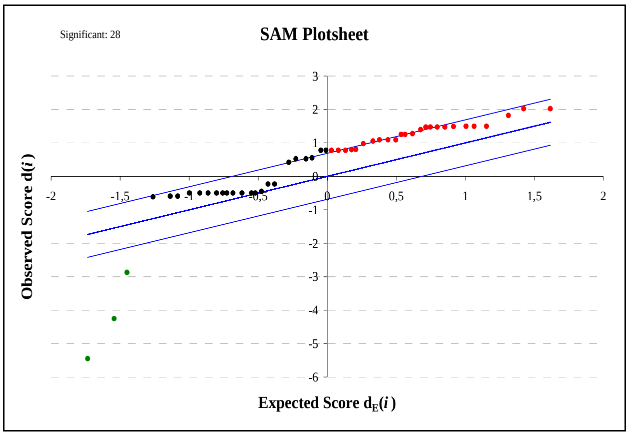


Figure 4. Plotsheet of the SAM method. Red values are positively significant green values are negatively significant.

In contrast, the SAM technique detects explicitly only significant classifications. The threshold of significance (delta value) was applied with 0.71. Figure 4 visualizes the observed score d(*i*) and the expected score from the permutations of the initial data $d_E(i)$ in a plot. 28 of the classifications were evaluated as significant. While only three accuracies had a negative significance (green), 25 of the input data sets were valued as positively significant (red).

The results of this calculation are listed in a table of significant values, which is summarized in table 3 beginning with the top entries of the highest significance.

Positively significant geo-factors	**Negatively significant geo-factors**
All geo-factors included	Only spectral band 3 included
All geo-factors except aspect included	Only spectral band 2 included
All geo-factors except slope included	Only spectral band 1 included
Slope in Beech-young included	
FSM in Beech-young included	
All geo-factors in Black Alder included	
FSM in Black Alder included	
All geo-factors except curvature included	
All geo-factors except CSM included	
DTM for all classes included	
All geo-factors except FSM included	
All geo-factors in Beech-young included	
DTM in Beech-young included	
Curvature in Beech-young included	
Slope in Sycamore included	

Table 3. List of positively and negatively significant geo-factors

As the ISODATA clustering was already indicating, SAM shows a high significance of the combined utilization of all additional geo-factors. Moreover, classes with smaller percentages of covered area and smaller ecological niches depend more on the application of ancillary information than other forest types. This is visible in the detected high significances of the classifications of “Beech-young” and “Black Alder” or “Sycamore” in combination with different geo-factors, such as the forestry site map (FSM) or terrain parameters.

Moreover, the ISODATA clustering and the SAM agree in giving a negative significance to classification, where the spectral value is only partly utilized. Additionally, two types of geo-factors were detected as less significant. In the SAM-order the accuracies using the Conceptual Soil Map (CSM) and the parameter aspect of the DTM were not detected as negatively significant. However, all accuracies which included only these geo-factors were close to a negative significance. This might be due to an inappropriate data quality of the soil-map and the lack of significance of the geo-factor aspect. In further studies in this pre-alpine region it can be considered to omit these additional information as not significant.

4. CONCLUSION AND OUTLOOK

This article presents a series of experiments of classifying QuickBird data with different kinds of information and integration techniques and testing their significance.

In conclusion it can be stated that the integration of ancillary data and knowledge is valuable to the classification success for forest types. The best classification results were achieved when utilizing a fuzzy-logic integration including all possible information. However, more integration techniques, such as neuronal networks or genetic algorithms should be applied in comparison.

Moreover, two methods of evaluating the significance of multiple classification accuracies were presented. While the ISODATA clustering showed valuable information about patterns in the accuracy data, the SAM technique can identify significant influences of different types of additional data to different class accuracies. This contributes to a better evaluation of classification results, especially when using multiple spatial data-types.

Further studies could involve SAM for testing the significance of hyperspectral bands for certain classes, the evaluation of the quality of training data, or the evaluating of the significance of scales when segmenting objects.

References

Benz, U., Hofmann, P., Willhauck, G., Lingenfelder, I. and Heynen, M., 2004. Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS-ready information. ISPRS Journal of Photogrammetry & Remote Sensing, 58: 239-258.

Burnett, C. and Blaschke, T., 2003. A multi-scale segmentation / object relationship modelling methodology for landscape analysis. Ecological Modelling, 168: 233-249.

Chu, G., Narasimhan, B., Tibshirani, R. and Tusher, V.G., 2007. SAM-Significance Analysis of Microarrays-Users guide and technical document, Stanford, pp. 1-41.

Dirnböck, T., Dullinger, S., Gottfried, M., Ginzler, C. and Grabherr, G., 2003. Mapping alpine vegetation based on image analysis, topographic variables and Canonical Correspondence Analysis. Applied Vegetation Science, 6: 85-96.

Förster, M. and Kleinschmit, B., 2006. Integration of Ancillary Information into Object-based Classification for Detection of Forest Structures and Habitats. In: S. Lang, T. Blaschke and E. Schöpfer (Editors), 1st International Conference on Object-based Image Analysis. ISPRS, Salzburg, Austria, pp. 1-6.

Kleinschmit, B. et al., 2006. Erfassung von Wald-Lebensraumtypen in FFH-Gebieten - Fernerkundung am Taubenberg und im Angelberger Forst. LWF Wissen, 51: 1-39.

Li, J., Tang, X., Zhao, W. and Huang, J., 2007. A new framework for identifying differentially expressed genes. Pattern Recognition, 40.

Maselli, F., Conese, C., Filippis, T. and Romani, M., 1995. Integration of ancilliary data into a maximum likelihood classifier with nonparametric priors. Journal of Photogrammetry and Remote Sensing, 50(2): 2-11.

Stolz, R. and Mauser, W., 1996. A fuzzy approach for improving landcover classification by integrating remote sensing and GIS data. In: E. Parlow (Editor), Progress in Environmental Remote Sensing Research and Applications. Balkema, Rotterdam, pp. 33-41.

Tilli, T., 1993. Mustererkennung mit Fuzzy-Logik. Franzis, München, 336 pp.

Tusher, V.G., Tibshirani, R. and Chu, G., 2001. Significance analysis of microarrays applied to the ionizing radiation response. PNAS, 98: 5116-5121.

Venables, W.N. and Smith, D.M., 2006. An Introduction to R.

Walentowski, H., Ewald, J., Fischer, A., Kölling, C. and Türk, W., 2004. Handbuch der natürlichen Waldgesellschaften Bayerns. Geobotanica, Freising, 441 pp.

Walentowski, H., Fischer, M. and Seitz, R., 2005. Fir-dominated forests in Bavaria. waldökologie online(2): 18-39.

Zhang, Y., 2002. Problems in the fusion of commercial high-resolution satellite images as well as Landsat 7 images and initial solutions. Int. Arch. Photogram. Remote Sens., 34(4).