INT. J. REMOTE SENSING, 20 JULY, 2004,
VOL. 25, NO. 14, 2713–2723

Taylor & Francis
Taylor & Francis Group

Use of metatruth image concept to assess forest change detection accuracy at pixel level

K. NACKAERTS*, K. VAESEN, I. LIZARRAGA, B. MUYS and P. COPPIN

Laboratory for Forest, Nature and Landscape Research,
Vital Decosterstraat 102, 3000 Leuven, Belgium

(*Received 18 October 2001; in final form 21 August 2003*)

Abstract. Traditionally, the validation of a classified multispectral image only quantifies its correspondence to ground reference data containing thematic information generalized at the stand level, with stands represented as vector polygons. Little is known of the accuracy of such classifications at a scale below the stand. This study presents a methodology to assess classification accuracy at pixel level, i.e. sub-polygon, where the classification procedure is embedded in a change detection environment. A new type of reference data (Metatruth Image) was generated based on the integration of the outputs of various independent change detection procedures. The integration consisted of calculating for each pixel a probability distribution or pixel purity index for each change class by independent change detection procedures, defined by the number of times the pixel has been classified as a certain change class. First, the relationship between purity and accuracy was successfully validated. Next, the Metatruth Image was created based on 'high purity pixels'. Performing traditional accuracy assessment on the outputs of individual change detection procedures using the Metatruth Image as reference dataset, demonstrated that former outputs identified change events accurately at pixel level. As a consequence, traditional accuracy assessment at polygon level underestimates the true accuracy at pixel level of the change detection procedure in a systematic way with differences in kappa coefficients of agreement around 20%.

1. Introduction

Thematic maps are used in a wide variety of fields and applications, ranging from geology over climatology to forest management. Over the past several years, interest in the accuracy of maps has grown and with it research on the accuracy assessment of thematic maps. In a first approach, the global measures of error are based on the confusion matrix as the overall accuracy (Lunetta *et al.* 1991), the kappa coefficient of agreement (Rosenfield and Fitzpatrick-Lin 1986), the user's accuracy and producer's accuracy (Story and Congalton 1986). One advantage of these methods is that they yield a single overall thematic map accuracy index, usually presented as the fraction or percentage of correctly classified pixels. The kappa statistic was originally developed by Cohen (1960) to measure the observed

*Corresponding author; e-mail: kris.nackaerts@agr.kuleuven.ac.be

International Journal of Remote Sensing
ISSN 0143-1161 print/ISSN 1366-5901 online
http://www.tandf.co.uk/journals
DOI: 10.1080/0143116032000160453

agreement of categorical data (Landis and Kock 1977). Since then, the kappa statistic has received considerable attention in remote sensing applications (Congalton and Mead 1983, Rosenfield and Fitzpatrick-Lin 1986, Hudson and Ram 1987, Stehman 1992, Czaplewski 1994). A detailed description of the kappa coefficient of agreement can be found in Bishop *et al.* (1975) and Kalkhan (1994).

A more complicated model to calculate accuracy is by assigning to each pixel a different certainty estimate. This kind of model introduces the concept of spatial variation of uncertainty. Several fuzzy classification algorithms offer an estimation of the uncertainty at pixel level, such as the class membership probabilities (i.e. the probability a pixel belongs to a given class; Lowell 1994) and entropy (Leung *et al.* 1992). Allan (1999) introduced the concept of cell purity, where the cell purity is determined from the cell attributes, which result from the synthesis of n map realizations by n independent interpreters. For a cell, which does not have identical attributes over all interpretations, the purity may be indicative for a classification uncertainty of mixed pixels of heterogeneous classes or fuzzy classification (Goodchild 1992). The overall accuracy measure does not take this kind of uncertainty information into account and assumes all pixels are 'pure' pixels. These pure pixels are 'pure' in two ways: first of all, spatially, it is assumed that only one type of land cover occurs within the boundaries of these pixels. And secondly, thematically, there is no doubt about the class to which these pixels belong. Therefore it provides a poorer expression with respect to the quality of the less pure remainder, possibly the vast majority, of the dataset.

In forest management, digital change detection algorithms and vegetation indices or change indicators are important tools for monitoring the evolution of forest cover over time. But the type of change detection procedure can significantly affect the detection, identification and quantification of the change event (Colwell and Weber 1981). Furthermore, the accuracy of a change map is largely dependent on the change detection algorithm, change indicator(s), and classification system used to produce the map, because these factors may give different results for the same regions (Coppin and Bauer 1994). To determine the accuracy, traditionally, overall accuracy assessment measures (e.g. kappa coefficient of agreement) are calculated (Rosenfield and Fitzpatrick-Lin 1986). They are indicative for the correspondence between the produced change map and a ground reference dataset, the latter containing information on the actual change events, as observed on the ground. However, ground reference data contain some degree of both spatial and thematic generalization, as is inherently the case with all types of thematic map. This is mainly due to the high operational costs involved in creating such datasets. To develop a ground reference map of changes occurring in forests, the change events are generally described at the level of the forest management unit. The polygon of a certain forest management unit is said to have been affected e.g. by a storm, without mentioning that some areas inside the polygon may have been untouched.

Successful detection and identification of change events at sub-polygon level (e.g. pixel level) will aid the interpretation of digital change maps by experienced local forest managers. This will result in improved allocation of the available resources.

Therefore, this study will determine whether this intra-polygon information can be revealed in a cost-effective way i.e. by means of digital change detection procedures. The methodology to define this intra-polygon information is based mainly on the approaches to deal with pixel accuracy as mentioned above.

2. Materials

2.1. *Ground reference*

A high quality ground reference dataset was available for each time interval. In a first step, the forest management units were digitized. The description of the forest management units was available at the Beltrami County Land Department. They were defined irrespective of their dominant forest cover type, i.e. all sorts of cover types were represented e.g. aspen, birch, jack pine, red pine, black spruce. Second, management relevant information such as forest development stage (i.e. regeneration, immature, merchantable and over-mature) and forest density (understocked, well stocked and to be thinned) were determined for each forest management unit using colour infrared (CIR) aerial photography and information from local forest managers. In a third step, change events were identified by comparing changes in the forest development stage and forest density to critical parameters (as well as made available by local forest managers). This resulted in 11 change types: (1) clearcut, (2) clearcut with natural regeneration, (3) clearcut with plantation, (4) early regeneration development, (5) early plantation development, (6) plantation establishment, (7) flooding, (8) highgrading/selective cut, (9) complete vegetation removal, (10) storm damage and (11) no change. As this study was not interested in the detection and identification of the causal agent of a change event, the 11 change types were reclassed into three change classes: net canopy loss (1–3;7–10), net canopy gain (4–6) and no change (11).

2.2. *Satellite imagery*

For the initial change detection procedures two Landsat TM images (1984 and 1990) were selected during July–August. The images covered an area of 415 km^2 localized in Beltrami County, Minnesota, USA. The image pre-processing involved four processes: data calibration, scene rectification and registration, atmospheric normalization and correction, and interpretability enhancement through the generation of change detection indicators.

A detailed description of all change detection procedures used in this study can be found in Nackaerts et al. (2004) where the performance of various change detection routines consisting of specific change indicator/change algorithm combinations were compared. The outputs of all procedures tested then serve as input for the development of the Megatruth Image (described in §3), and therefore need some explanation (see also figure 1).

The change indicators used were divided into three different groups: (1) focusing on photosynthetic activity of green biomass: NDVI (Normalized Difference Vegetation Index), Kauth–Thomas' Greenness, VNIR (very near infrared) (TM4) and SAVI (Soil Adjusted Vegetation Index); (2) referring to soil characteristics: Kauth–Thomas' brightness and red (TM3); and a last group (3) focusing on ecosystem moisture condition: Kauth–Thomas' wetness, MIR (mid infrared) (TM5) and MG (mid infrared over green ratio) (TM5/TM2).

On the other hand, four change detection algorithms were selected for this study based on their methodological independency:

(1) Single band standardized image differencing is the most widely applied change detection algorithm for a variety of geographical environments (Coppin and Bauer 1996). It substracts each band from one image of one date from each corresponding band from a second of another date and

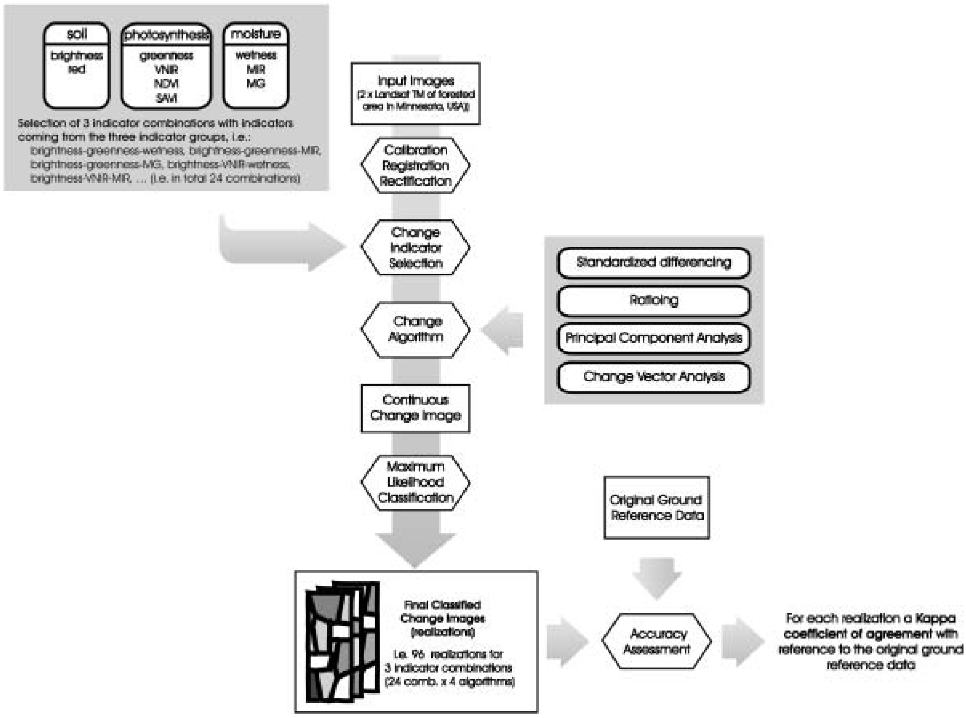

Figure 1. Overview of an individual change detection procedure.

divides the results by the sum of both. Pixel values different from 0 indicate changes.

(2) Image ratioing is one of the simplest and fastest change detection methods (Coppin and Bauer 1995). Two images from different dates are ratioed band-by-band. A value of 1 after ratioing corresponds to a pixel that has not changed and a value different from 1 to a pixel that has changed.

(3) Principal component analysis (PCA) is one of the most frequently applied linear transformation techniques. Principal component analysis transforms the multi-dimensional feature space into a new Euclidean space by means of a linear transformation, in such a way that maximal variability is explained by the first new axis (PC1). The second axis (PC2) is created orthogonal to PC1. If the change area is small proportional to the total study area, the first component will contain information about the areas with stable vegetation within image variability, the second one about the changed areas (Chavez and Kwarteng 1989).

(4) Modified change vector analysis (mCVA) is a conceptual extension of image differencing (Nackaerts *et al.* 2004). Two or more spectral variables represented by a vector in feature space are plotted at dates 1 and 2 for a given pixel. The magnitude and direction of the vector connecting these two vectors in a Euclidean space describe the magnitude and direction of spectral change between the two dates and are assumed to correspond to change on the ground.

The mCVA inherently requires at least two change indicators as input. Because of the grouping of biophysically related indicators into three groups, it was decided to use all three-input-band combinations with all three inputs coming out of a different group (i.e. 24 combinations). All combinations were submitted to the four change detection algorithms mentioned above, resulting in 96 realizations. The latter were continuous change images, on which the maximum likelihood classification was applied, resulting in categorized change images containing the change classes: net canopy loss, net canopy gain and no change.

3. Methodology

3.1. *The Metatruth Image*

The concept of the Metatruth Image is based on the assumption that through an integration of multiple change detection realizations (performed on a relatively large area) spatially more accurate information can be obtained than is obtained out of a map merely based on traditional (forest management relevant) ground observations. That this assumption is valid can be easily understood by the fact that in the case of traditional ground observations, the map information is based on a higher degree of generalization (i.e. at polygon level) than in the case of an image processing based analysis (i.e. at pixel level). This assumption is statistically tested in the second step of the procedure to generate the Metatruth Image as described below. A post validation was performed by comparing significant differences

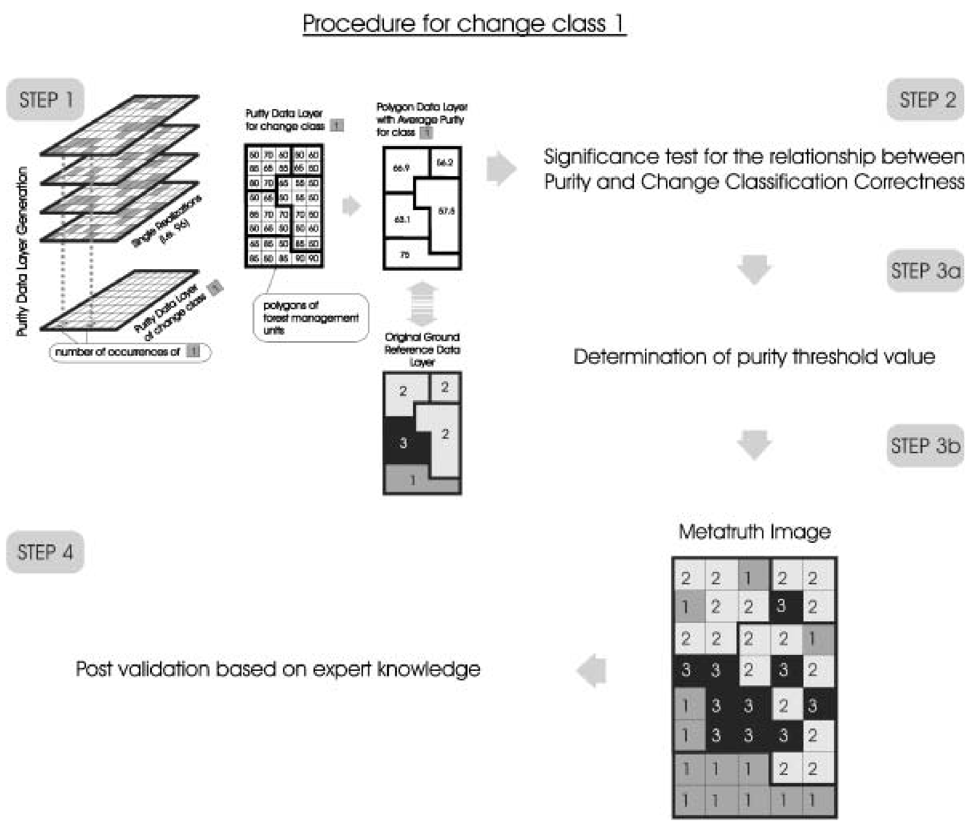

Figure 2. Overview of the Metatruth Image development.

between the Metatruth Image and the original ground reference data using the expert knowledge of the local foresters.

The procedure to obtain the Metatruth Image consists of three steps. An overview is given in figure 2.

First, for each change class a purity data layer is created in which each pixel value represents the number of times it has been classified as that specific change class, standardized by the number of realizations (i.e. 96 = 24 different input bands combined with four different change detection methods). This results in three purity data layers corresponding to canopy loss and canopy increase change events and no change.

In a second step, it is verified if a relationship exists between the calculated purity value and the correctness of a classification output, i.e. do high purity values for a specified change event correspond to a high certainty of correctness for that change event? As the ground reference data only contain thematic information at polygon level, the purity images were generalized to the same level. Therefore, the average purity value of each polygon is calculated.

The significance of the relationship between the purity values at polygon level and the change identification 'correctness' with respect to the available ground reference data is tested by means of a logistic regression approach. Polygon averaged purity values for a specific change class are hereby translated to a binary indicator: 0, identified as not belonging to change class *i*; 1, identified by the ground reference data as change class *i*. A Wald Chi-square statistic (Sharma 1996) and a corresponding significance level are calculated to test whether the relation found is significant, i.e. is there a relation between purity and change classification correctness?

The assumption that through multiple—methodologically independent—change detection realizations, spatially more accurate information can be obtained is now translated to the assumption that the significance of the correlation between purity values and classification correctness is independent of the level of generalization.

To validate this assumption, the polygon averaged purity values were stratified according to polygon size. The original dataset was sorted by polygon size and then arbitrarily subdivided into five groups containing an equal number of polygons. The same logistic regression approach as described above was performed on each of these five groups independently to test whether the significance of the purity–correctness relationship was size (i.e. scale or level of generalization) independent. Scale independency of the significance of this relationship justifies the interpretation of the Metatruth Image at pixel level.

Furthermore, purity threshold values have to be determined (step 3a). These serve to select pixels with a high degree of certainty, which will be used to develop the Metatruth Image (step 3b). The latter consists, for each pixel, of the change class corresponding to the highest pixel purity value that is larger or equal to the selected purity threshold value for the corresponding change class. The threshold value itself is defined based on the frequency data of the polygon averaged purity images. Since reference data are available for this level of generalization, the error of commission (Congalton 1991) can be calculated for each threshold value. This error corresponds to the ratio of the number of erroneously classified polygons to the total number of polygons classified to that specific change class at a given threshold level. When the assumption of a scale independent relationship between purity and change classification correctness is valid, a predefined level of accuracy can then be converted to the correct threshold value needed to create a Metatruth Image at pixel level.

3.2. *Accuracy assessment*

In order to test the performance of the traditional individual change detection procedures at sub-polygon level, kappa coefficients of agreement are calculated using both the original reference data ($\sim$management units, i.e. polygon level), and the Metatruth Image ($\sim$pixel level). Using paired-sample *t*-tests, the differences between the two approaches are evaluated. As such, it can be determined to what extent change detection realizations are capable of detecting and identifying changes at intra-polygon level.

4. Results and discussion

4.1. *Validation of the purity–change classification correctness relationship*

To test the significance of the relationship between purity and change classification correctness, a logistic regression was applied for all three change classes between average polygon purity and a binary change class indicator (see above). The Wald statistic for the model parameters (intercept and slope) showed highest degrees of significance (<0.0000). The same level of significance was found for the five different subsets corresponding to different polygon sizes. These results prove not only that there is a relationship between purity and change classification correctness but also that the significance of this relationship is scale independent. This justifies the use of accuracy information derived from the Metatruth Image generalized to polygon level, at another lower level of generalization such as the determination of the threshold value and application of this value at pixel level.

4.2. *The Metatruth Image*

In figure 3, the error of commission is plotted against its corresponding purity threshold value. This figure is based on the frequency data of the polygon averaged purity image for each change class and the ground reference data. The error of commission can easily be calculated as the ratio of the incorrectly classified polygons for a specific threshold value to the total sum of polygons. Since it has been demonstrated that the relationship between correctness and purity is scale independent, this threshold value can now be used at pixel level to create the desired Metatruth Image. In order to define a purity threshold value (see figure 3) (i.e. up to which purity value a sufficient degree of correctness is warranted), an error of commission of 10% was arbitrarily chosen. The resulting purity threshold values per change class were: 65.6 for class 1 (net canopy loss), 55.6 for class 2 (net canopy gain) and 32.5% for class 3 (no change) (figure 3).

Figure 3 also clearly supports the logistic regression approach to study the relationship between change classification correctness and in this study calculated purity values for each change class. On the one hand, high purity values are correlated to correctly detected change events. On the other hand, low purity values clearly correlate to inaccurate detection of change events. In general, the relationship highly resembles the sigmoid shape, characteristic for the logistic regression.

Finally, in order to develop the Metatruth Image, each purity image was thresholded by the above defined threshold values. For instance, a pixel having a pixel purity value for change class 1 of 60 (below the threshold) would be rejected

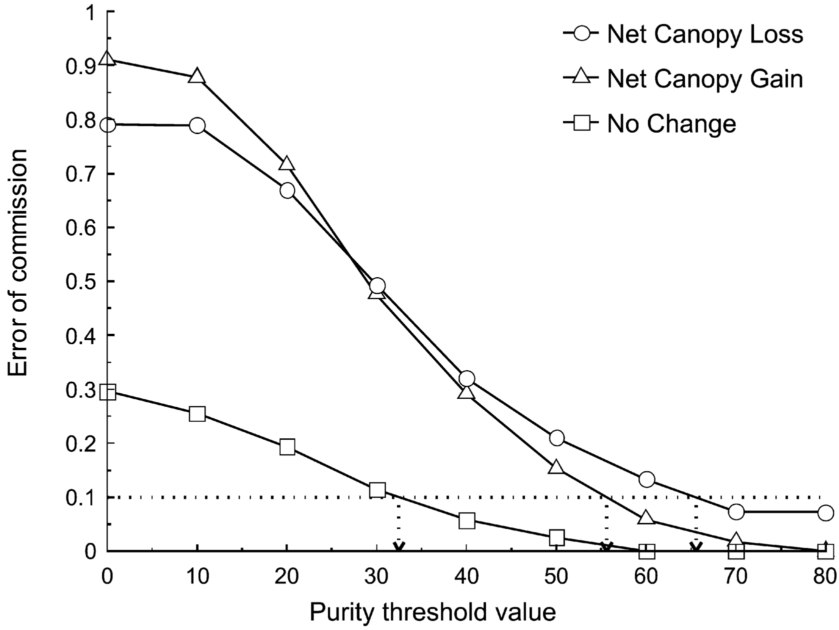

Figure 3. The effect of the purity threshold on the error of commission of the Metatruth Image.

while a pixel with a purity value for the same change class of 80 (above the threshold) would be selected. A pixel with a purity value for change class 2 of 60 would be identified in the Metatruth Image as change class 2. Approximately 95% of the original total area was selected to form the Metatruth Image. The other 5% non-classified pixels were too uncertain.

4.3. *Post validation of the Metatruth Image*

In order to verify if the Metatruth Image disposes of spatially more accurate change information, use was made of the expert knowledge of the local forest managers. Therefore, the latter image was compared to the original ground reference map, and the most remarkable differences were identified and proposed to the local foresters. It appeared that most confusion/distinction arose in polygons with predominant tree type 'aspen' and 'jack pine'.

For aspen polygons, a vast amount of pixels being assigned to change class 1 (net canopy loss), were classified as change class 3 (no change) in the Metatruth Image. According to local forest experts, aspen stands often show a high variability within the management unit. They are characterized by a thin canopy layer, which is extremely vulnerable to local stress (wind, nutrient deficiencies, etc.). Moreover, aspen stands have a vigorous growth and an enormous ability to reiterate from root suckers. As such, aspen polygons originally submitted to a canopy decrease and classified as such in the ground reference (i.e. as class 1, net canopy loss), in reality can (i.e. inside a polygon, at pixel level) have regenerated locally in the time interval under consideration, and can therefore locally be classified as class 3 (i.e. no change).

For jack pine stands, most confusion existed for polygons which were classified originally (i.e. in the original ground reference map) as class 3 (no change). Within these polygons in the Metatruth Image a substantial amount of pixels was classified as class 1 or 2. The confusion occurring here can be understood from the fact that these jack pine stands are characterized by a very heterogeneous horizontal structure: within the stand, some areas show very low tree density (due for example to wind damage), while other areas have high densities as a consequence of intensive natural regeneration. As such, within these polygons canopy decrease, canopy increase and no canopy change can occur easily next to each other.

Other disparities between the Metatruth Image and the ground reference map did occur, but were less clear and considered less substantial.

4.4. *Accuracy assessment*

Figure 4 shows an overview of the kappa coefficients of agreement for all change indicator/change algorithm combinations, once using the original ground reference dataset and once using the Metatruth Image. It can be clearly seen that all coefficients are much higher in the latter case.

Tables 1 and 2 support the same finding. Paired-sample *t*-tests, moreover, demonstrated that the difference in kappa between the two approaches was significant in all cases ($a = 0.05$) with an average increase of 22% and a maximum increase of 29%.

It appears that single change detection operations can detect and identify changes accurately at the pixel level, certainly more accurately than could be expected when reference is made to a polygon-based ground reference dataset. Moreover, at the pixel level the relative performance of the various change

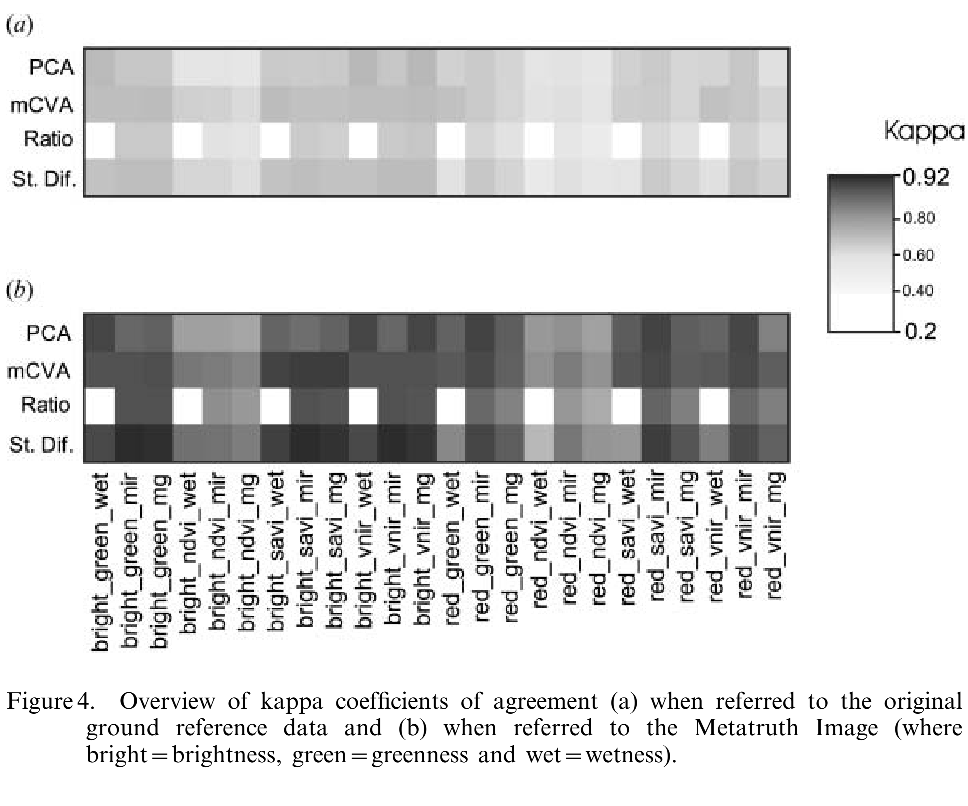

Figure 4. Overview of kappa coefficients of agreement (a) when referred to the original ground reference data and (b) when referred to the Metatruth Image (where bright = brightness, green = greenness and wet = wetness).

Table 1. Average kappa coefficients of agreement for all change detection algorithms (*a*) when referred to the Metatruth Image; (*b*) when referred to the original ground reference data.

Change detection algorithm	(*a*) Metatruth	(*b*) Original reference
PCA	0.8475	0.6159
mCVA	0.8738	0.6205
Ratio	0.6454	0.4780
Significant difference	0.8754	0.6349

indicators/change algorithms is similar to the performance mentioned in the study by Nackaerts *et al*. (2004): mCVA was demonstrated the best algorithm, while for the change indicators brightness performed best for the soil related indicators, MIR for the ecosystem moisture related ones, and greenness, VNIR and SAVI were superior to NDVI.

5. Conclusions

This study successfully validated the use of the pixel purity concept in a change detection context to extract pixel level reference data from traditional polygon level reference data: i.e. the Metatruth Image. Based on this new Metatruth Image, it was demonstrated that individual change detection routines are able to detect and identify change events accurately, not only at the forest management unit level (i.e. polygon level) as proven in an earlier studies (Coppin and Bauer 1994, Nackaerts *et al*. 2004) but also at pixel level. It is suggested that the highest accuracy is obtained by the use of the mCVA change detection algorithm, and the change indicators Kauth–Thomas (KT) brightness, greenness, TM4, SAVI and TM5.

Moreover, the results of this study indicate that the traditional accuracy measure, the kappa coefficient of agreement, underestimates the real change classification accuracy by on average 20% due to the generalization of the ground reference data. This means that the generalization of digital change detection documents to polygon level reduces its information content significantly. It is therefore advised that the expert knowledge of local forest managers is used for the interpretation and eventually generalization of digital change detection documents, taking sub-polygon patterns of change events into account.

Table 2. Average kappa coefficients of agreement for all change indicators (*a*) when referred to the Metatruth Image; (*b*) when referred to the original ground reference data.

Change indicator	(*a*) Metatruth	(*b*) Original reference
SAVI	0.8396	0.5978
NDVI	0.7318	0.5395
VNIR	0.8343	0.6071
Greenness	0.8363	0.6048
Brightness	0.8269	0.6059
Red	0.8066	0.5813
Wetness	0.7035	0.5259
MIR	0.8767	0.6266
MG	0.8514	0.6095

SAVI, Soil Adjusted Vegetation Index; NDVI, Normalized Difference Vegetation Index; VNIR, very near infrared; MIR, mid infrared; MG, mid infrared over green ratio.

References

ALLAN, R. C., 1999, Characterising local spatial uncertainty in the optimisation of thematic class areas. *Proceedings of the 3rd Symposium on Spatial Uncertainty, Quebec* (Chelsea, MI: Ann Arbor Press), pp.105–111.

BISHOP, Y. M. M., FEINBERG, S. E., and HOOLAND, P. W., 1975, *Discrete Multivariate Analysis—Theory and practice* (Cambridge, MA: MIT Press).

CHAVEZ, P. S. J., and KWARTENG, A. Y., 1989, Extracting spectral contrast in Landsat Thematic Mapper image data using selective Principal Component Analysis. *Photogrammetric Engineering & Remote Sensing*, **55**, 339–348.

COHEN, J., 1960, A coefficient of agreement of nominal scales. *Educational and Psychological Measurement*, **20**, 37–46.

COLWELL, J. E., and WEBER, F. P., 1981, Forest change detection. *Proceedings of the 15th International Symposium on Remote Sensing of Environment, 11–15 May 1981, Ann Arbor, MI*, pp. 839–852.

CONGALTON, R., 1991, A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment*, **37**, 35–46.

CONGALTON, R., and MEAD, R., 1983, A quantitative method to test for consistency and correctness in photointerpretation. *Photogrammetric Engineering & Remote Sensing*, **49**, 67–74.

COPPIN, P. R., and BAUER, M. E., 1994, Processing of multitemporal Landsat TM imagery to optimise extraction of forest cover change features. *IEEE Transactions on Geoscience and Remote Sensing*, **60**, 287–298.

COPPIN, P. R., and BAUER, M. E., 1995, The potential contribution of pixel-based canopy change information to stand-based forest management in the Northern US. *Journal of Environmental Management*, **44**, 69–82.

COPPIN, P. R., and BAUER, M. E., 1996, Digital change detection in forest ecosystems with remote sensing imagery. *Remote Sensing Reviews*, **13**, 207–234.

CZAPLEWSKI, R. L., 1994, Variance approximation for assessment of classification accuracy. Research paper, Rocky Mountain Forest & Range Experimentation Station, Forest Service, USDA, Fort Collins, CO, 94 pp.

GOODCHILD, M. F., 1992, Development and test of an error model for categorical data. *International Journal of Geographical Information Science*, **6**, 87–104.

HUDSON, W. D., and RAM, C. W., 1987, Correct formulation of the kappa coefficient of agreement. *Photogrammetric Engineering & Remote Sensing*, **53**, 421–422.

KALKHAN, M. A., 1994, Statistical properties of six accuracy indices using simple random and stratified random sampling: an application in remote sensing. PhD dissertion, Colorado State University, 134 p.

LANDIS, J. R., and KOCK, G. G., 1977, The measurement of observer agreement for categorical data. *Biometrics*, **33**, 154–174.

LEUNG, Y., GOODCHILD, M. F., and LIN, C. C., 1992, Visualization of fuzzy scenes and probability fields. *Proceedings of the Fifth International Symposium on Spatial Data Handling, Charleston, SC* (Columbia: International Geographical Union), pp. 480–490.

LOWELL, K. E., 1994, Probabilistic temporal GIS modelling involving more than two map classes. *International Journal of Geographical Information Systems*, **8**, 73–93.

LUNETTA, R., CONGALTON, R., FENSTERMAKER, L., JENSEN, J., MCGWIRE, K., and TINNEY, L., 1991, Remote sensing and geographic information system data integration: error sources and research issues. *Photogrammetric Engineering & Remote Sensing*, **57**, 677–687.

NACKAERTS, K., VAESEN, K., MUYS, B., and COPPIN, P., 2004, Performance of a modified change vector analysis in forest change detection. *International Jounal of Remote Sensing*, in press.

ROSENFIELD, G. H., and FITZPATRICK-LIN, K., 1986, A coefficient of agreement as a measure of thematic classification accuracy. *Photogrammetric Engineering & Remote Sensing*, **52**, 223–227.

SHARMA, S., 1996, *Applied Multivariate Techniques* (New York: John Wiley & Sons).

STORY, M., and CONGALTON, R., 1986, Accuracy assessment: a user's perspective. *Photogrammetric Engineering & Remote Sensing*, **52**, 377–399.

STEHMAN, S. V., 1992, Comparison of systematic and random sampling for estimating the accuracy of maps generated from remotely sensed data. *Photogrammetric Engineering &Remote Sensing*, **58**, 1343–1350.