

A STRATEGY FOR QUALITY ASSURANCE OF LAND-COVER/LAND-USE INTERPRETATION RESULTS WITH FAULTY OR OBSOLETE REFERENCE DATA

P.Hofmann, P. Lohmann

Leibniz University of Hannover, Institute for Photogrammetry and GeoInformation
(hofmann/lohmann)@ipi.uni-hannover.de

KEY WORDS: Quality Assurance, Quality Control, Accuracy Assessment, Land Cover Classification

ABSTRACT:

Land cover mapping in most cases results in thematic maps which are in general not a perfect representation of the reality. They always contain errors due to the method used for its production or simply the difficulty that in many cases the land use is not equivalent to the land cover and semantic knowledge of an interpreter has to be used to derive the finally wanted object type. In order to assess the reliability of a particular land use map procedures for quality control and checking the geometry and thematic contents of the mapped objects have to be applied.

The result of an accuracy evaluation typically provides the user with an overall accuracy of the map and the accuracy for each class (object type) in the map. Conducting a proper accuracy assessment of a product created using remotely sensed data can be time and resource consuming and respectively expensive, especially if no appropriate or up-to-date reference material is available and ground truth data has to be collected.

In this paper a method is presented, which makes use of reference data (thematic maps), which have been produced in another context using class descriptions which slightly differ from the class descriptions used in the actual land use mapping and which have been generated some years before the actual mapping. It can be shown, that it is possible to use this “obsolete” data for accuracy checks, because means are implemented to find those areas in the reference data, which are out-of-date or which have been falsely assigned to the class which is inspected. The procedure allows to exclude these areas from the computation of accuracy measures. The method has been implemented in ArcGIS 9 using SQL-based algorithms.

1. INTRODUCTION

1.1 The DeCOVER project

Within the EC GMES initiative a consortium of 9 companies and the Institute of Photogrammetry and GeoInformation of the University of Hannover is collaborating in national project aiming in the development of a procedure for the revision of land cover data for public services named DeCOVER. Within DeCOVER a basic catalogue of land use objects is generated based on the classification of up-to-date satellite imagery with special emphasis to the use of the future RapidEye and TerraSAR-X satellite systems. The catalogue comprises 5 major classes which altogether ensemble 39 subclasses at a spatial resolution of 5m-Pixel. By implementing ontology based techniques for interoperability 3 major GIS databases are served, the CORINE CLC, ATKIS and the biotope and landuse catalogue BNTK. In order to avoid errors in the production of the DeCOVER basic data base procedures are implemented at each processing level to check for deficiencies and errors.

1.2 Quality assurance of GIS data

The process for evaluating data quality is a sequence of steps to produce and report a data quality result. A quality evaluation process consists of the application of quality evaluation procedures to specific dataset related operations performed by the dataset producer and the dataset user (ISO/TC 211, 2002).

A data quality evaluation procedure is accomplished through the application of one or more data quality evaluation methods.

Data quality evaluation methods are divided into two main classes, direct and indirect (see Fig 1).

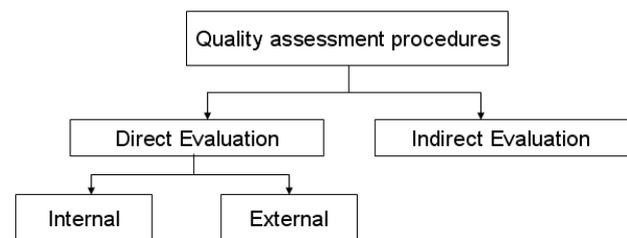


Figure 1: Types of Data Quality Assessment

Direct methods determine data quality through the comparison of the data with internal and/or external reference information. Indirect methods infer or estimate data quality using information on the data such as meta-information as being supplied with satellite data if it is used.

The direct evaluation methods are further subclassified by the source of the information needed to perform the assessment. Means of accomplishing direct assessments include for both automated or non-automated a full inspection or sampling (ISO/TC 211, 2003).

Data quality elements and data quality subelements which can be easily checked by automated means include format consistency, topological consistency, domain consistency, (i.e. boundary violations, specified domain value violations), completeness (omission, commission) and sometimes also temporal consistency.

Full quality control assures testing every item in the population specified by the data quality scope while sampling requires testing sufficient items in the population in order to achieve a data quality result. However, the reliability of the data quality result should be analysed when using sampling; especially, when using small sample sizes and methods other than simple random sampling. In the paper present we will focus on an effective method to control the thematic accuracy of mapping results by a pointwise comparison of the results with reference data that is not 100% reliable.

1.3 Data used within the project

The workflow within the DeCOVER project comprises a processing chain which is separated by the 5 major object categories (Urban, Water Bodies, Forest, Agricultural Land, Natural Areas) each of which should be checked by an independent data set. The 5 categories are classified in high resolution satellite images (SPOT, QUICKBIRD etc.) resampled to a spatial resolution of 5m by automatic object oriented classification techniques together with sophisticated interpretations based on auxillary data and semantic properties as being fixed within a predefined mapping guide. Since neither prevailing adequate reference data nor ground truth data exists, it has been investigated if the use of other existing landuse data could be successfully applied. Candidates for this type of data could be any of the above mentioned GIS databases, but also an independently produced landcover dataset, which has been produced by the company Infoterra GmbH, Germany. This dataset called LaND25 is based on a classification of satellite imagery (ETM) at a geometric accuracy of +25m and a thematic accuracy (top level classes) of 95%. In order to check the top level classes of DeCOVER a systematic sampling procedure has been selected using the LaND25 dataset processed to a grid of points with a horizontal and vertical distance of 180m and using the thematic content of LaND25 at these locations as being converted to that of the DeCOVER specification. For ease of use this represents the simplest and least expensive approach to sampling, yielding in site locations which in case of a too course grid not always results in samples that reflect the population about which inferences are to be drawn. However the minimum number of reference samples necessary to perform a significant statistical test (Goodchild, M. F., 1994) is by far exceeded with the given point distances.

2. QUALITY CONTROL IN A SEQUENTIAL PROCESS FLOW

2.1 Status of present quality measures

The quality of mapping results generated by interpreting remote sensing data is usually assessed by comparing the mapping results with a reference on a per pixel basis. This way, the results of the comparison are presented in a so called error matrix (Congalton, R. G. and Green, K., 1999; Lillesand, T. M. and Kiefer, R. W., 2000). Further common accuracy measurers are the Kappa Index of Agreement (KIA) and in some cases the measurers of Short and Helden are additionally used as parameters (Baatz, M. et al., 2004). However, all these measurers only indicate the degree of agreement between the mapping results derived from analysing the remote sensing data and those of the reference mapping. Additionally, the explanatory power of all these measurers depends strongly on the number of classes to be mapped and compared. Especially

when comparing the mapping results of only one class the comparison is reduced to counting and comparing the number of agreed pixels, false positives, i.e. pixel that have been classified in the reference but could not be confirmed by the classification and false negatives, i.e. pixels that have been classified in the mapping but could not be confirmed by the reference. Consequently, when comparing both mapping results only by one class and its complement, the error matrix is reduced to a schema as described in Tab. 1.

Table 1: Scheme of an error matrix for the comparison of two classes.

		Mapping Results			producers accuracy
		Non-Class	Class	Sum	
Reference Mapping Results	Non-Class	agreed non-class	false negatives		
	Class	false positives	agreed class		
	Sum				
	users accuracy				
	Total accuracy				

Further, the explanatory power of the KIA is changed to the evaluation of how well both mappings do fit to each other, i.e. how much the reference mapping agrees with the mapping result and vice versa. In accordance to Lillesand, T. M. and Kiefer, R. W., 2000 the KIA is calculated by:

$$KIA = \frac{N \cdot \sum_{i=1}^r x_{ii} - \sum_{i=1}^r (x_{i+} \cdot x_{+i})}{N^2 - \sum_{i=1}^r (x_{i+} \cdot x_{+i})} \quad (2.1)$$

whereas the meanings of the operands in (2.1) are explanatory shown in Tab. 2

Table 2: Explanatory distribution of the KIA-operands in the error matrix.

		Mapping Results		
		Non-Class	Class	Sum
Reference Mapping Results	Non-Class	x_{ii}	x_{fn}	x_{i+}
	Class	x_{fp}	x_{ii}	x_{+i}
	Sum	x_{+i}	x_{+i}	N

Since the KIA ranges from -1.0 to +1.0, the KIA has to be interpreted as follows: The more the KIA is close to -1.0 the less both mappings do agree. The closer the KIA is to 0.0, the more the mappings do agree by 50%. The closer the KIA is to +1.0 the more both mappings do agree. Note: the KIA does not explain, how far the mapping results meet the results from the reference mapping and vice versa. Additionally, when the KIA is close to 0.0 no conclusion about the quality of the mapping can be drawn. Therefore, the Producer's and User's Accuracy have to be interpreted.

Referring to Tab. 2 and Lillesand, T. M. and Kiefer, R. W., 2000 the User's Accuracy for each class is calculated by:

$$PA = \frac{x_{ii}}{x_{i+}} \quad (2.2)$$

and the Producer's Accuracy by:

$$UA = \frac{x_{ii}}{x_{+i}} \quad (2.3)$$

whereas the Producers Accuracy is closer to 1.0 the less the number of false positives is and vice versa and the Users Accuracy is closer to 1.0 the less the number of false negatives is and vice versa. Note: unlike in comparisons of multi-class mappings and references, UA and PA give no indication about possible confusions¹ of classes.

2.2 Quality control embedded into a sequential process chain

Within a sequential process chain, as like intended in the DeCOVER project, the quality of each successor products strongly depends on the output quality of its predecessor products. Consequently, controlling the intermediate results of each predecessor is vital in order to reduce error propagation and to guarantee the specified level of quality for the finally resulting products. As such it appears reasonable to control the products' quality in two stages: an intermediate stage, which is set up in the process chain and allows to control the quality of each predecessor result in a fast and reliable way but with reduced level of detail (plausibility check) and a final stage, which can be seen as a separate part of the process chain and complies the quality control of the consolidated final product and in the full level of detail (see Figure 1).

In order to reduce the necessary effort for the intermediate quality assurance (QA), methods of quality control (QC) have to fulfil two basic requirements:

1. be as reliable as possible and necessary in terms of the given product specifications
2. be as cost efficient as possible in terms of time consumption.

Consequently, the performance of QC procedures in the process chain is determined by a maximum of reliability and the effort necessary to achieve this reliability, i.e.: well balanced QC procedures are a good compromise between reliability and effort. Besides accepted methods of QC, the reliability of QA of course depends on the sampling strategy and sample size, while effort depends on the degree of automation and the precision of QC.

2.3 Reliability and ability of reference data

It is obviously, that evaluating the thematic quality of mapping results by using a reference mapping as a pointwise measurement of agreement/disagreement, the results of the comparison, i.e. the calculated parameters of accuracy assessment depend on the ability of the reference of being representative and comparable to the mapping of concern.

While the representativeness of the reference is given by the density of the sampling grid and can be adapted as necessary, the comparability of the reference data depends on three basic qualities:

1. semantic interoperability of the reference classes to the corresponding classes of the assessed mapping.
2. comparable mapping rules for corresponding classes, e.g. generalization rules, minimum mapping units etc.

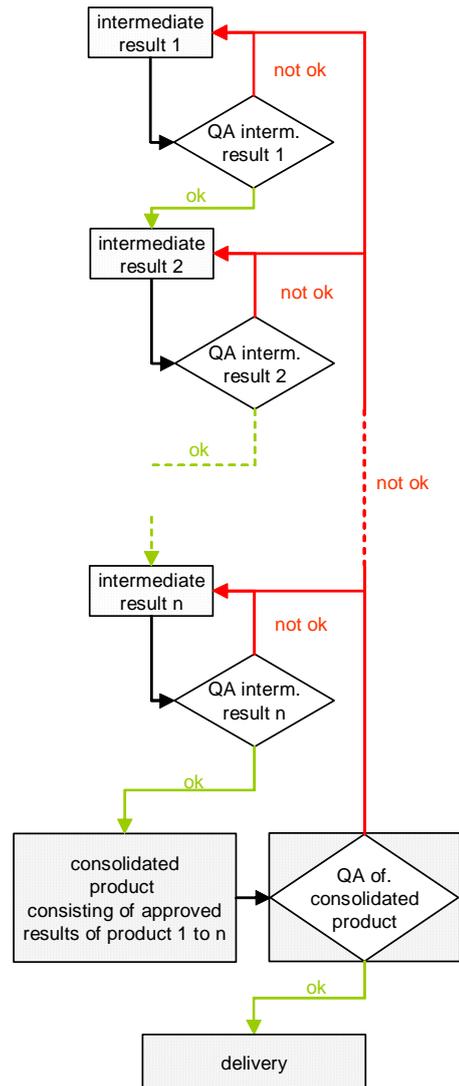


Figure 1: Stages of QA within the sequential process chain.

3. date of mapping the reference and the date of the mapping to be assessed.

Since all of these three aspects cannot be assumed as 100% conform in both mappings, the reference itself has to be assumed as partly erroneous². In consequence mapping mismatches between the reference mapping and the intermediate result do not necessarily indicate an error in the intermediate result. With respect to potential thresholds for accepting or rejecting intermediate results indicated by the QA parameters as presented in chapter 2.1, an intermediate result might be rejected although the mapping is correct. Thus, in order to avoid such miss-rejections the reference itself has to be checked against reality, i.e. against the satellite data used for mapping the product.

3. PLAUSIBILITY CHECK OF INTERMEDIATE RESULTS WITHIN THE PROCESS FLOW

¹ In some cases the error matrix is also called confusion matrix.

² In this context erroneous means the reference does not map reality the same way as the assessed mapping does.

3.1 General Procedure

In order to avoid unnecessary interruptions and delays in the progress of production, rejections of an intermediate product because of faulty mismatches between intermediate result and reference data have to be avoided. As demonstrated in chapter 2.3 such wrong rejections can be caused by an outdated reference or by differing class descriptions between reference and mapping. Consequently, the QA parameters are forged and indicate a less thematic accuracy as is in fact. In these cases it is necessary to verify whether the reference or the mapping is correct or incorrect. For verification only the satellite data used for the mapping in conjunction with the mapping guide must be used. If an error in the reference data is verified, the reference data should either be corrected accordingly or the respective parts of the reference data should be omitted during the calculation of the QA parameters. After adapting the reference data accordingly, it should be tested again whether a clearance of the intermediate result can be given or not (see Figure 2).

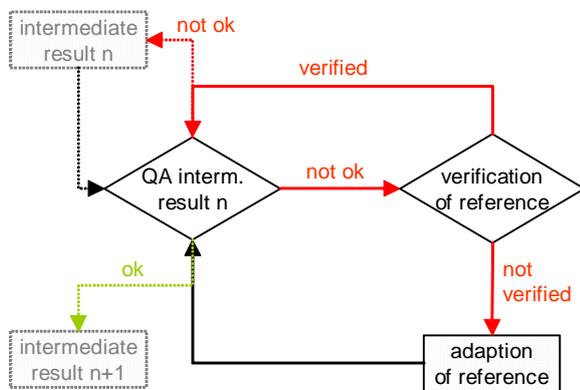


Figure 2: Verification of reference in case of disagreement between mapping and reference.

3.2 Handling of erroneous reference data

Since the verification process described in the chapter above can be very time consuming if every measurement point of disagreed comparison has to be examined, we developed a strategy which allows to accelerate this process and to keep the whole procedure of quality control credible. Therefore, we assume that relevant differences between mapping and reference occur spatially clustered – independent whether the reference is outdated, the class descriptions are differing, or the mapping is in fact false. Thus, a verification of errors should firstly focus on these clusters since in many cases the clusters indicate spacious mismatches, which are easier to explain and if necessary to correct as the following example illustrates:

In DeCOVER *airports* and *urban green* are mapped as a subclass of the category *urban*. According to the mapping guide, runways, taxiways, airport buildings and the green in-between are mapped as *airport* and thus assigned to the category *urban*. *Urban green* represents features such as parks, graveyards or the green of traffic islands. In the reference only airport buildings are classified as *urban* and features understood in DeCOVER as *urban green* are in the reference assigned to *grassland*. In consequence measurement points covering runways and taxiways as well as the green in-between the airport infrastructure or in parks, graveyards etc. appear as (spatially clustered) false-negatives (see Figure 3).

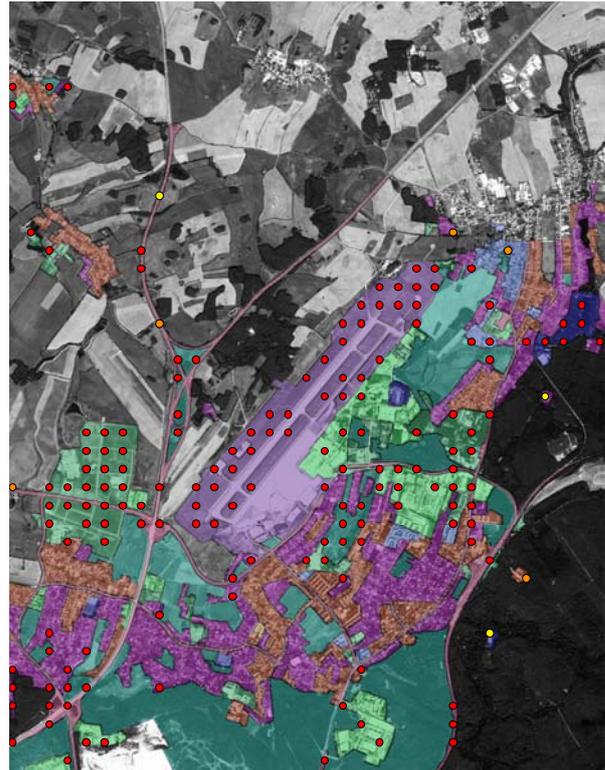


Figure 3: Example of clustered mismatches between reference and mapping.

The clustering of mismatching points can be determined by regarding the distance of a mismatching point to the next mismatching point. Classifying mismatching points according to their calculated distance to the next mismatching points finally identifies those points that should be taken firstly into account for a visual inspection and verification.

For an efficient ‘correction’ of the reference data, those points can be selected by their (non-interoperable) class assignments in both classifications and their distance to the next mismatching point by an SQL-statement of the following form:

```

select * from samples where [class-id reference] <>
[class-id mapping] and [class-id mapping] = [value
mismatching (sub-) class] and [distance] <= [threshold for
clustering]
  
```

This way, a whole bunch of points falsely indicating a mapping error can be adapted accordingly without being spatially selected. For the adaptation of the reference data set the operator has the choice of either deleting these mismatching points or to correct them in accordance with the satellite data used for the assessed mapping result and the respective mapping guide. Thereby, in cases where the reasons of mismatching are quite clear – such as in the example demonstrated – preferably the points should be corrected and not deleted, since by deleting points the sample size is reduced and might become too small in order to be representative. A deletion of points is only recommended for disperse mismatching control points, i.e. the mismatching points are not spatially clustered. Note: also disperse mismatching points have to be inspected and validated. They may be deleted only if the thematic mismatch cannot be verified by the satellite data.

3.3 Results

Within the DeCOVER project we have accomplished the sketched process of QA and QC exemplary for the category *urban* on a sub-area of a test-site that is covering parts of southern Saxony. The sub-area itself - representing parts of the city of Dresden - is of size of approx. 225km². As reference data we have used the LaND25 classification as described in chapter 1.3. In order to obtain a point mesh of sample points we generated a point layer with a regular distance of 180m which is the double density of points in order to have a representative sample size. The doubling was performed since we did not know a priori how many points might be deleted. These points obtained the classification values of the reference data and the DeCOVER classification by two respective spatial join operations. We then identified the false-positives and false-negatives and filled the values of the error matrix with the observed counts and calculated the QA values as described in chapter 2.1. In this first iteration of assessing the quality in terms of a plausibility check, the values for the QA measurer UA for *urban* of the DeCOVER classification was below the given threshold of 90% (see Table 3).

Table 3: Error matrix for category *urban* in test site sub-area before adapting mismatching points.

		DeCOVER				
		Not Urban	Urban	Sum	in %	producers accuracy [%]
LaND25	Not Urban	3590	558	4148	72,09	86,55
	Urban	69	1537	1606	27,91	95,70
	Sum	3659	2095	5754		
	in %	63,59	36,41			
	users accuracy [%]	98,11	73,37			
Total accuracy [%]		89,10				
KIA		0,75				

In order to visually verify or correct these results we determined clusters of false-positives and false-negatives as described in chapter 3.2 by calculating their distances to each other and classifying them according to their density. Because of the diagonal and orthogonal neighbourhoods of the points, the upper bound of the class with the highest density was determined at $180m \cdot \sqrt{2} \approx 255m$. The other class boundaries were determined according to their frequencies, so that finally five density classes were generated (see Figure 4). Then we visually inspected all mismatching points that were assigned to the first class, i.e. the clusters with highest density and compared the DeCOVER class-assignments with the satellite data (see Figure 3). In the case the mismatch between reference and mapping was obviously caused by any of the reasons as described in chapter 2.3 and 3.2, the class assignments of these points to the reference class were corrected. After having inspected all points indicated in red in Figure 4, and if necessary respectively corrected or deleted³, the error matrix was updated with the newly observed counts.

Now the values for the QA measurers were above the given thresholds (see Table 4) which would have led to an acceptance

³ Disperse clusters containing only two points.

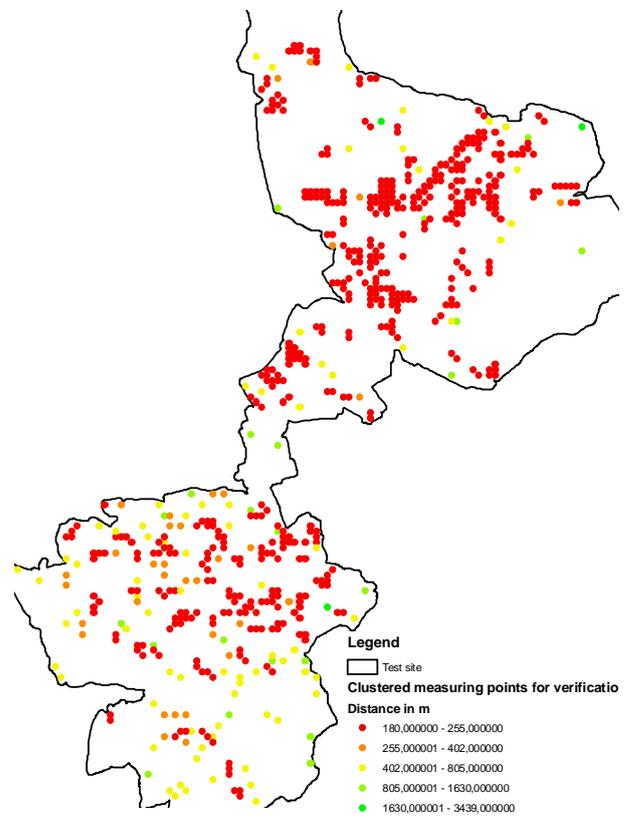


Figure 4: Spatial distribution of mismatching control points in test-site for category *urban*.

of the intermediate result and thus avoided a faulty rejection respectively.

Table 4: Error matrix for category *urban* in test site sub-area after adapting mismatching points.

		DeCOVER				
		Not Urban	Urban	Sum	in %	producers accuracy [%]
LaND25	Not Urban	3590	202	3792	65,99	94,67
	Urban	61	1893	1954	34,01	96,88
	Sum	3651	2095	5746		
	in %	63,54	36,46			
	users accuracy [%]	98,33	90,36			
Total accuracy [%]		95,42				
KIA		0,90				

3.4 Discussion

Although some mapping elements might not be controlled because they are not captured by the point-grid, we found that the methods for performing the QC presented in this article are a well balanced compromise between effort and achievable accuracy of controlling. Especially focusing only on those points, which give reason to assume that the results of the QA are forged - whereas the reasons for that forgery are a priori not quite clear - reduces the necessary effort remarkably but simultaneously keeps the QA credible. Additionally by adjusting the cluster defining distance between the points gives a controllable and comprehensible flexibility of adjusting the accuracy of control. Last but not least unavoidable interruptions of the production chain could be reduced to a minimum, since only in the case of confirmed deficiencies an intermediate result is not accepted. Nevertheless, there is still a remaining unknown

portion of potential errors that might not be detected, which is given by the case that both - reference and mapping - are not correct.

4. LITERATURE

Baatz, M. et al., 2004: eCognition Professional 4 User Guide, Definiens Imaging GmbH, Munich, Germany, pp. 155-160.

Congalton, R. G. and Green, K., 1999: Assessing the Accuracy of Remotely Sensed Data: Principles and Practices. Lewis Publishers. Boca Raton.

Goodchild, M. F., 1994: Sample Collection for Accuracy Assessment, in: Baily, M. et al.: "USGS – NPS Vegetation Mapping Program". <http://biology.usgs.gov/npsveg/aa/sect4.html> (accesses 11 May 2007).

Lillesand, T. M. and Kiefer, R. W., 2000: Remote Sensing and Image Interpretation. 4th Ed. John Wiley & Sons. New York, pp. 568-575.

ISO/TC 211, 2002: ISO 19113 Geographic information - Quality principles.

ISO/TC 211, 2003: ISO 19114 Geographic information - Quality evaluation procedures (draft).