AUTOMATIC QUALITY ASSESSMENT OF GIS DATA BASED ON OBJECT COHERENCE

Christian Becker and Joern Ostermann and Martin Pahl

Leibniz Universität Hannover
Institut für Informationsverarbeitung
Appelstr. 9a, 30167 Hannover, Germany
(becker, ostermann, pahl)@tnt.uni-hannover.de
http://www.tnt.uni-hannover.de

KEY WORDS: Land Cover, Updating, Change Detection, Semi-automation

ABSTRACT:

Spatially referenced data, stored as Geo Information System (GIS) data is needed for numerous applications like map services or administration. Thus, quality assessment of GIS data is of highest importance. Manual quality checks and updates are very time consuming, therefore support through automatic systems is required. Typically, semi-automatic systems are used to narrow down the list of objects that are getting reviewed in a manual process. Current systems introduce an additional layer of complexity, though. To select objects in a binary decision, the user is asked to define rules and to specify parameters. We present a system that provides a much more flexible selection of objects, without the need of any rules or parameters: Relations between up-to-date remote sensing data and GIS data are automatically identified and described by attributes. To determine a Coherence Model for each GIS feature the attribute space is evaluated using a multivariate Gaussian. Abnormality of GIS objects is determined by calculating the Mahalanobis distance of a GIS object's attributes to the mean. Evaluation shows that results outperform existing systems improving precision by a factor four.

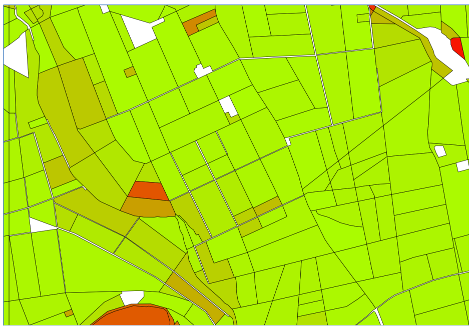


Figure 1: Choropleth map showing continuous evaluation result for GIS objects: (from red (conspicuous) to green (inconspicuous))

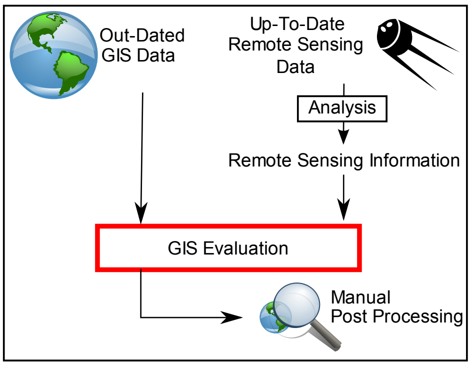


Figure 2: Structure of Quality Assessment Systems

1 INTRODUCTION AND RELATED WORK

Spatially referenced data in GIS (*Geo Information System*) databases are of great use for national and international organizations and companies. They provide a basis for administration, monitoring and various map services. Information about quality of GIS data and the area of changes are of highest importance for updates.

Increasing availability and quality of remote sensing data makes it interesting for quality assessment of GIS data: up-to-date remote sensing information provide a reference view for out-dated GIS data. On the other hand, increasing numbers of sensors and channels increase the complexity of the verification process. Thus, automatic solutions are required to keep up with the development.

Numerous research has been done in the area of quality assessment of GIS data. A first class of approaches focus on analysing remote sensing data. Results are simply projected on GIS data to show areas of change (Lacroix et al., 2006), (Leignel et al., 2010), (Olsen, 2004). Results are very detailed but often suffer from being cluttered and neglecting generalisation effects (Haunert, 2009). Furthermore, this approach requires algorithms that provide results for every GIS feature in the scene. However, currently only for special GIS features a reliable image analysis algorithm exists. In (Walter, 2004) the process of GIS quality assessment is subdivided into analysis and GIS evaluation as depicted in Figure 2. This structure can be seen in systems of a second class. An evaluation step is used to introduce a model for matching analysis results and GIS data. As GIS specifications are formulated in terms of GIS objects, this second class of approaches analyse GIS errors object-based: (Busch et al., 2004) present the system WIPKA-QS. First ortho-images are analysed by an analysis combining various approaches. GIS object based rules are introduced to specify for each GIS feature which analysis results are expected in a correct or incorrect GIS object. Optimisation of results and changing availability and quality of remote sensing data require regular development of rules by experts. In (Walter, 2004) the scene in also analysed object-based but is not requiring any rules. Relations between GIS objects and appearance in image data are trained to perform a maximum likelihood (re-)classification of GIS objects. Objects that end with another GIS feature than they began with are considered errors in the GIS. This approach has two obvious disadvantages: First, objects would have to be nearly

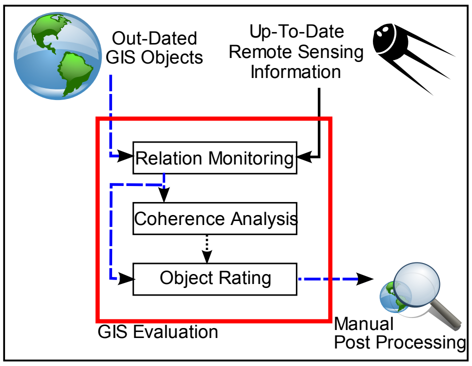


Figure 3: Internal structure of *Unsupervised* GIS *Evaluation*. Green dashed lines: GIS data, black solid lines: remote sensing information, dotted black line: *Coherence Model* used for GIS object rating.

completely wrong to be detected as errors. Partial errors cannot be found. Secondly, in cases where GIS features are hard to differentiate in the image, eg. greenland and cropland, erroneous classifications can be expected. In this paper we present a new system that is not requiring any rules while also not having the disadvantages of the system presented in (Walter, 2004). Furthermore, objects are given a continuous rating instead of a binary correct/false decision. In contrast to (Walter, 2004) we evaluate each GIS feature independently to avoid confusion caused by hard to distinguish GIS features. Hence, the multi-class classification becomes several one-class problems. Thus, the problem changes from finding a "correct" GIS feature for a GIS object to the problem to determine how "normal" a GIS object.

Details of our approach are presented in Section 2. Afterwards we demonstrate and discuss the performance in Section 3 before concluding in Section 4.

2 UNSUPERVISED GIS EVALUATION

As different features of GIS objects are evaluated independently for rest of section 2 we are describing the approach for one GIS feature only.

For our system we assume that up-to-date remote sensing data has already been analysed by appropriate algorithms (see also Figure 2) to provide *Remote Sensing Information*. All algorithms that provide the result as a label image of the scene may be used. Examples can be found in (Lacroix et al., 2006), (Leignel et al., 2010), (Helmholz et al., 2010) and in Section 3.

Fig. 3 visualizes the workflow of the GIS evaluation. In a first step relations between GIS data (blue dashed lines) and remote sensing information (black solid lines) are *monitored* for each GIS object. Monitoring results are stored in the objects in term of numerical *relation attributes*. GIS objects now can be considered to be points in an *attribute space* that is spaned by the relation attributes of all objects. Given that relations between GIS objects and their appearance in remote sensing data exist and have been captured successfully in the image analysis and monitoring steps, GIS objects will form clusters. Position, size and shape of the clusters reflect the coherence of appearance of GIS objects in remote sensing information. The coherence is modeled by fitting a probability distribution in attribute space (step *Coherence Analysis*). Finally, a GIS object's relative position in attribute space towards the cluster of all objects is used to determine an *abnormality rating* to it.

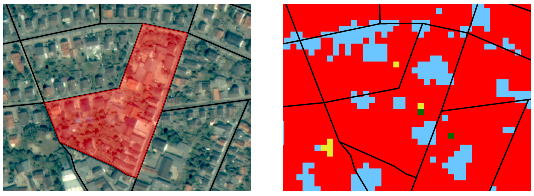


Figure 4: GIS object of feature *settlement* (left, highlighted in red) composed of multiple patches (right, *red: settlement, blue: industry, green: wood, yellow: cropland)*

2.1 Relation Monitoring

Relations between remote sensing information and GIS data are getting described by attributes for each GIS object.

In the image analysis results, a label image, each pixel is assigned to one of several *classes*. Groups of interconnected pixels with the same class are called *patches*. Now we define *Analysis Histogram* attributes: For each GIS object g and all analysis classes $i = 1, 2, \dots, n$ attributes $a_1, a_2, \dots, a_n$ are calculated:

$$a_i(g) := \frac{\sum_{p \in \mathrm{P}} \delta_i(p)|p \cap g|}{|g|} \tag{1}$$

with P all patches in the scene and interpreting patches and GIS objects as sets of pixels and

$$\delta_i(p) := \begin{cases} 1 & \text{if } p \text{ has analysis class } i \\ 0 & \text{else} \end{cases} \tag{2}$$

Analysis histogram attributes are used in the research field of landscape analysis (Mcgarigal and Marks, 1995). For quality assessment of GIS objects they have been used in (Walter, 2004). In contrast to approaches in (Lacroix et al., 2006), (Leignel et al., 2010), (Olsen, 2004) analysis classes are not uniquely mapped to each GIS feature. Instead, a GIS object is described by the percentage of all GIS classes in its area.

2.2 Coherence Analysis and Object Rating

Each GIS object can be placed in an attribute space using its analysis histogram attributes. Hence, objects with similar percentage of GIS classes are placed next to each other. Objects with other analysis classes are placed further away. To deduce a model of the distribution of objects the attribute space has to be evaluated in the *Coherence Analysis* step.

We opted to use a d-dimensional multivariate normal distribution. Its density is defined as

$$f(x = x_1, x_2, \dots, x_d) = \frac{1}{((2\pi|S|)^{d/2}} e^{(x-\mu)^T S^{-1}(x-\mu)} \tag{3}$$

The dimension d is the number of analysis histogram attributes, mean μ, covariance matrix S and $|S|$ the determinant. The distribution only depends on mean and covariance. Thus, it easily can be estimated. Fortunately, to measure distances we don't have to solve this equation. Instead we can determine the *Mahalanobis distance* (Mahalanobis, 1936):

Reference / System	Incorrect GIS Objects	Correct GIS Object
High Abnormality	tp *true positives*	fp *false positives*
Low Abnormality	fn *false negatives*	tn *true negatives*

Table 1: Evaluation matrix. The entries focus on the task to find incorrect GIS objects. Values are only meaningful for a specific abnormality rating as threshold between high and low abnormality.

For a GIS object $g = (a_1, a_2, \ldots, a_n)$ in attribute space, the Mahalanobis distance is defined as:

$$d(g, \mu) := \sqrt{(g - \mu)^{\mathrm{T}} S^{-1} (g - \mu)} \tag{4}$$

The Mahalanobis distance is a standard method for this application (Tan et al., 2005). The covariance matrix S in the formula is used to compensate pair-wise correlations between attributes. To interpret the Mahalanobis distance as normality measure, the distance of an object towards the Gaussian's mean μ must be determined. Objects near the centre are more normal than more distant objects. Thus, the *abnormality value* of an GIS object g is calculated by

$$\text{abnormality}(g) := d_{\text{mah}}(g, \mu) \tag{5}$$

$0 \leq$ abnormality $< \infty$, using the mean and covariance matrix determined for g's GIS feature. When we use in this paper the term high *normality* this corresponds to a low abnormality value and vice versa. The absolute values of abnormality have no spacial meaning.

3 RESULTS AND DISCUSSIONS

In this section we present experimental results for our system. The test area is located in central Germany. GIS data is taken from the German GIS data set ATKIS (Arb, 2011). We chose objects from GIS features 2111 (*settlement*) and 4107 (*forest*) for evaluation because they provide results for two very different GIS features: On the one hand, settlement objects tend to be very inhomogeneous in images, on the other hand, forest objects are much more regular. IKONOS imagery with 1 m resolution on four channels (*red*, *green*, *blue* and *near infrared*) is available to gain remote sensing information. The analysis step to process remote sensing data is not part of this contribution. It is only considered a necessity to create remote sensing information used as input data for our GIS evaluation. Thus, we are using a *Support Vector Machine* (SVM (Vapnik, 2000)) with RBF kernel. Features are *mean*, *covariance* and *Haralick features* (Haralick et al., 1973). A 25×25 pixel window is used. For performance reasons, every eighth pixel is classified, only. Original resolution is gained though nearest-neighbor up-scaling. We have trained the SVM with samples with *houses*, *halls*, *fields*, *trees*.

GIS objects have been labeled as correct and incorrect by an independent person to be used as reference. To evaluate our results, we compare the reference to the abnormality rating for the objects. This done by testing for all values of abnormality which objects have higher or lower abnormality and calculating values in Table 1. When this is done for all possible thresholds, the values can be depicted as graphs.

Examples for GIS objects, image and analysis results are shown in Figure 5. In Figures 5a and 5c two forest objects are shown, in Figures 5b and 5d two settlement objects. It can be seen that the image analysis is not very detailed. Up-scaling and large neighborhoods caused blocks of common texture. Texture borders tend to be confused with house textures or industry halls. We want to stress that the image analysis is only used as input. Systematic errors like this can easily be dealt with in our system. When all objects are infected, it seems to be normal, thus has no impact on the abnormality of objects.

Results can be seen in Figures 6 and 7 where True and False Positives/Negatives are shown. 4 of 427 forest and 30 of 751 settlement objects have been found from the reference to be incorrect GIS objects. All four incorrect forest objects can be found within the top five objects when ordered by abnormality. The fifth object is the upper object in Figure 5a (correct according to reference). Settlement objects are not composed of some dominant texture. Depending on population density and other social aspects, houses are mixed with green areas and trees. Nevertheless, results show that our algorithm can reliable identify incorrect GIS objects.

From the true/false positives/negatives we determine the following values:

$$\text{Precision} = \frac{\text{tp}}{\text{tp} + \text{fp}}, \quad 0 \leq \text{Precision} \leq 1 \tag{6}$$

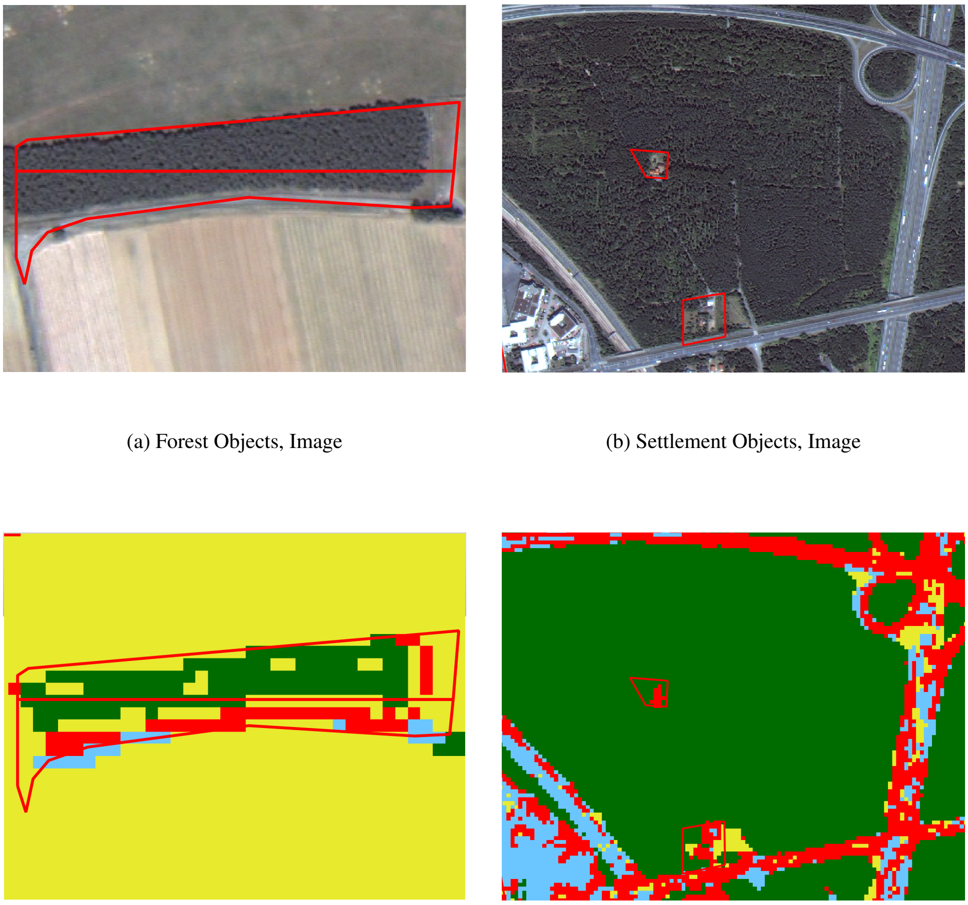
$$\text{Recall} = \frac{\text{tp}}{\text{tp} + \text{fn}}, \quad 0 \leq \text{Recall} \leq 1 \tag{7}$$

Precision and recall describe the quality of all objects that have been considered having a high/low abnormality at threshold p. On the one hand, a precision of 1 means that all objects with high abnormality are incorrect GIS objects. On the other hand, a recall of 1 indicates that all incorrect objects have high abnormality. Precision and recall have to be considered as tuple to get an impression of the quality. Results are shown in Figure 8. We want to evaluate our approach through comparison with alternative systems. Unfortunately, many systems provide no results that allow any comparison (eg. results in (Walter, 2004) include only three of four values from Table 1). For the Wipka system results for analysis of IKONOS images are presented (Helmholz et al., 2010), though. Wipka results do not provide a continuous rating but a binary decision. Therefore only single precision and recall values exist: precision=0.2, recall=0.73 (see also Fig. 8). Our result provides an increase of precision by factor four!

4 CONCLUSIONS

Existing quality assessment systems follow rather uniform approaches to tackle the task of decreasing the number of objects that is given to a human operator for GIS update. Based on automatic interpretation of remote sensing data, they allow to implement a rule set to select GIS objects for post processing (section 1). A disadvantage is that defining rules is not an option but an requirement. This task is assigned to end users that have to first appraise the analysis results before being able to formulate the rules.

We changed the GIS evaluation (Fig. 2) into an automatic evaluation, based on coherence of apperance (Fig. 3). No additional training data is required. We developed Analysis Histogram attributes to express relations between GIS data and analysis results. We fitted a multivariate Gaussian into the attribute space to describe coherence. Finally, we assigned each GIS object a abnormality rating.



(a) Forest Objects, Image

(b) Settlement Objects, Image

(c) Forest Objects, Image Analysis

(d) Settlement Objects, Image Analysis

Figure 5: Examples for objects with high unnormality rating. The lower forest object and both settlement objects are GIS errors according to reference.

Results have been presented checking ATKIS GIS data and using IKONOS imagery, interpreted by a SVM. We show that we can enhance precision by a factor 4 compared to existing systems.

We will proceed with tackling other problems using coherence analysis. An algorithm for finding change errors due to changed object borders at sub-object level has already been submitted for publication at ISPRS 2012.

REFERENCES

Arb, 2011. Amtliches Topographisch-Kartographisches Informationssystem.

Busch, A., Gerke, M., Grünreich, D., Heipke, C., Liedtke, C.-E. and Müller, S., 2004. Automated verification of a topographic reference dataset: System design and practical results. International Archives of Photogrammetry and Remote Sensing XXXV, Part B2, pp. 735–740.

Haralick, R. M., Shanmugam, K. and Dinstein, I., 1973. Textural features for image classification. Systems, Man and Cybernetics, IEEE Transactions on 3(6), pp. 610 –621.

Haunert, J.-H., 2009. Aggregation in Map Generalization by Combinatorial Optimization. Dissertation, Leibniz Universität Hannover, Germany.

Helmholz, P., Becker, C., Breitkopf, U., Büschenfeld, T., Busch, A., Grünreich, D., Heipke, C., Müller, S., Ostermann, J., Pahl, M., Vogt, K. and Ziems, M., 2010. Semiautomatic quality control of topographic reference datasets. In: ISPRS Commission 4 Symposium.

Lacroix, V., Idrissa, M., Hincq, A., Bruynseels, H. and Swartenbroekx, O., 2006. Detecting urbanization changes using spot5. Pattern Recognition Letters 27(4), pp. 226 – 233. Pattern Recognition in Remote Sensing (PRRS 2004).

Leignel, C., Caelen, O., Deibeir, O., Hanson, E., Leloup, T., Simler, C., Beumier, C., Bontempi, G., Warzée, N. and Wolff, E., 2010. Detecting man-made structure changes to assist geographic data producers in planning their update strategy. ISPRS Archive Vol. XXXVIII, Part 4-8-2-W9.

Mahalanobis, P. C., 1936. On the generalised distance in statistics. In: Proceedings National Institute of Science, India, Vol. 2number 1, pp. 49–55.

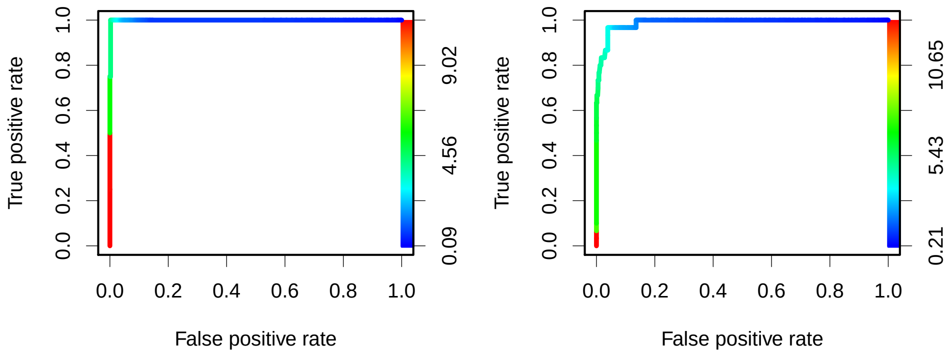
Mcgarigal, K. and Marks, B. J., 1995. FRAGSTATS: Spatial Pattern Analysis Program for Quantifying Landscape Structure. Department of Agriculture, Forest Service, Pacific Northwest Research Station.

Olsen, B. P., 2004. Automatic change detection for validation of digital map databases. In: In: International Archives of Photogrammetry and Remote Sensing, Vol. XXX IV, Part B2.

Tan, P.-N., Steinbach, M. and Kumar, V., 2005. Introduction to Data Mining, (First Edition). Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

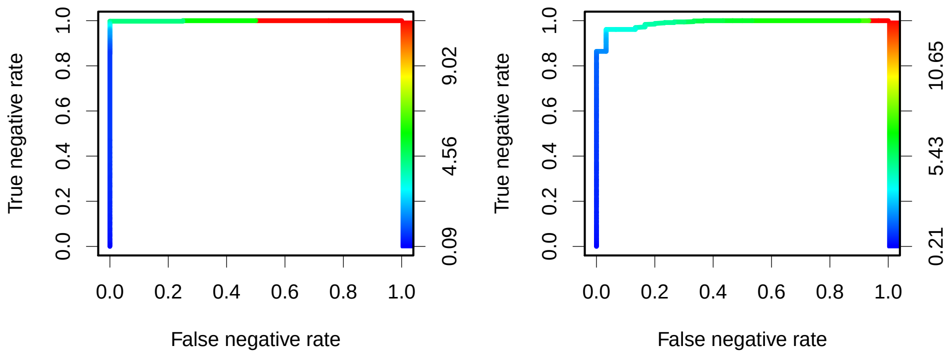
Vapnik, V., 2000. The Nature of Statistical Learning Theory. Springer, Berlin.

Walter, V., 2004. Object-based classification of remote sensing data for change detection. ISPRS Journal of Photogrammetry and Remote Sensing 58(3-4), pp. 225–238.



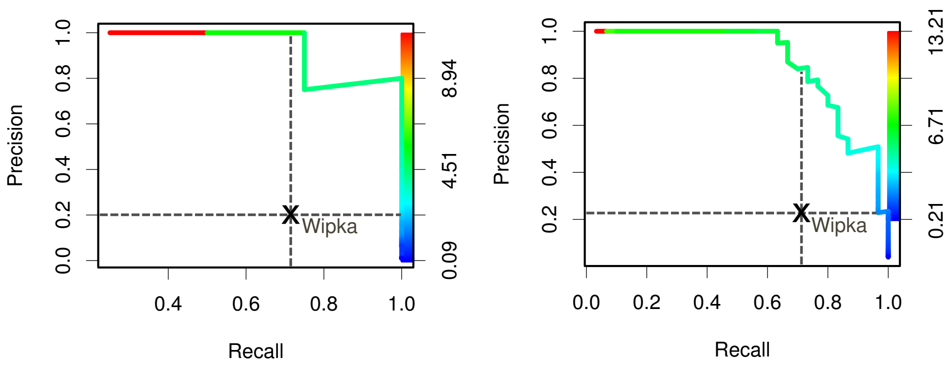
(a) Forest (b) Settlement

Figure 6: True Positives over False Positives for forest and settlement. Coherence rating used as threshold is visualized by color.



(a) Forest (b) Settlement

Figure 7: True Negatives over False Negatives for forest and settlement. Coherence rating used as threshold is visualized by color.



(a) Forest (b) Settlement

Figure 8: Precision over Recall for forest and settlement. Coherence rating used as threshold is visualized by color.