Towards Seamless Validation of Land Cover Data

Ekaterina Chuprikova[a], Lukas Liebel[b], and Liqiu Meng[c]

[a] *Chair of Cartography, Technical University of Munich, Germany; e.chuprikova@tum.de*
[b] *Chair of Remote Sensing Technology, Technical University of Munich; lukas.liebel@tum.de*
[c] *Chair of Cartography, Technical University of Munich, Germany; liqiu.meng@bv.tum.de*

Abstract: This article demonstrates the ability of the Bayesian Network analysis for the recognition of uncertainty patterns associated with the fusion of various land cover data sets including GlobeLand30, CORINE (CLC2006, Germany) and land cover data derived from Volunteered Geographic Information (VGI) such as Open Street Map (OSM). The results of recognition are expressed as probability and uncertainty maps which can be regarded as a by-product of the GlobeLand30 data. The uncertainty information may guide the quality improvement of GlobeLand30 by involving the ground truth data, information with superior quality, the know-how of experts and the crowd intelligence. Such an endeavor aims to pave a way towards a seamless validation of global land cover data on the one hand and a targeted knowledge discovery in areas with higher uncertainty values on the other hand.

Keywords: GlobeLand30, Land cover, Spatial-temporal uncertainty, Uncertainty visualization, Citizen Science, Probabilistic modeling

1. Introduction

Global land cover (GLC) is an important data source for environmental applications that helps us to understand the dynamic of our changing planet (Ban et al. 2015). Despite the development of remote sensing technology, the task of mapping and monitoring of our environment is still very challenging. Numerous efforts have led to GLC datasets of different resolutions such as GLC2000 (1km), GlobCover (300m), GLCF water mask (250m) and the latest publicly available GlobeLand30 (30m) produced by National Geomatics Center of China. Recent studies within the land cover community are increasingly committed to the validation of GlobeLand30 data (Brovelli et al. 2015, Sun et al. 2015, Arsanjani et al. 2016). The validation results revealed an overall accuracy between 46% and 83% for different study areas. Obviously some areas have higher degree of uncertainty than other areas, which necessitates further quality improvements of the GlobeLand30 data.

Due to multiple error sources and inaccuracies involved in data acquisition, processing and classification, the assigned land cover classes may differ from the ground truth and therefore they are uncertain. It is crucial to account for uncertainty that is useful for the evaluation of the data quality, decision-making practices and the creation of a general awareness. Apart from numerous definitions, we treat the uncertainty in this article as an indicator of distrust between the classified data in referenced datasets. Consequently, the higher the uncertainty, the more likely the actual land cover class of a particular location needs to be validated. Our work aims to model the GlobalLand30 data uncertainty based on probabilistic methods and to create a visualization support for the data refinement. We use the Bayesian Network framework for the uncertainty analysis of GlobeLand30 data. The Bayesian Networks are probabilistic graphical models that utilize probabilities obtained from the auxiliary data or expert notions to determine quantitative values of the uncertainty. In this work we focus on areas or GlobeLand30 classes that have higher degree of uncertainty. Volunteered geographic information (VGI), in particular, the Open Street Map (OSM) is adopted as an inexpensive auxiliary data source. In order to adapt the extracted dataset for our use case, we conducted a comprehensive analysis of the suitable OSM tags and assigned the closest GlobeLand30 class code to the OSM tags in our study area. The resulting maps will then visually guide the experts or non-expert users to explore the areas where a high degree of mismatch between the GlobeLand30 results and OSM information occurs.

2. Study Area and Data

This research investigates uncertain areas of the Global Land Cover classification, adopting the datasets of GlobeLand30, CORINE and Volunteered Geographic data based on OpenStreetMap for the area of Upper Bavaria (German: Oberbayern), Germany. Even though, none of the reference data sets is ground truth, we perform the data fusion in order to reveal variations in the classification. The test area of Upper Bavaria was selected due to availability of high quality datasets. This could reduce the complexity of the analysis and provide a basis for the study case. Moreover, the test area reveals a sufficient diversity in distribution of land classes, thus provides a solid background for exploring probability of each land class. Although the three datasets are different from each other, they are unified to the resolution of 30m and land cover classification based on GlobeLand30 schema.

2.1 GlobeLand30

In 2010 China launched a project with the aim to identify global land cover classes in resolution of 30 meters. The

Proceedings of the International Cartographic Association, 1, 2017. This contribution underwent single-blind peer review based on submitted abstracts | https://doi.org/10.5194/ica-proc-1-23-2017 |

GlobeLand30 data set is freely available and comprise 10 major classes of land cover, including cultivated areas, forests, grassland, shrub land, wetland, water bodies, tundra, artificial surfaces, bare land and permanent snow and ice. The classification is available for two base-line years, 2000 and 2010. The GlobeLand30 was produced based on more than 20,000 Landsat and Chinese HJ-1 satellite images (see www.globallandcover.com). Previous studies (Chen et al. 2015) indicate that the validation based on comparison with other data sets achieved an overall classification accuracy of over 80%. In the same time the GlobeLand30 data may show significant variation in different parts of the world and in some remote areas data presents the overall accuracy of 46% (Sun et al. 2015). Such a difference is mainly due to the difficulty to distinguish some classes such as forest, shrub land and grassland as well as the low availability of reference data. For this reason, our approach will involve auxiliary data, including VGI data, in order to support analysis of uncertain areas.

2.2 Open Street Map Land Cover

Since the term Volunteered Geographic Information (VGI) was introduced by Goodchild (2007), knowledgeable amateurs have contributed large amounts of spatially referenced data to different web portals. The OpenStreetMap (OSM) (openstreetmap.org) is, without any doubt, the most wide spread and well recognized project. The database comprises vector data, which is attributed with a great variety of labels and might serve as source data for various cartographic products. Since every contributor can freely edit the database without supervision, the OSM data is heterogeneous in terms of quantity and quality. Assessing the accuracy of the OSM is, therefore, an essential task to facility the scientific usage of this data source. Several studies have reported some encouraging results in terms of the overall accuracy and completeness (Helbich et al. 2012, Neis et al. 2012).

The validation of land cover classification requires manually labeled high-quality ground truth data for accuracy assessment. VGI-based approaches have been proposed and online communities, such as GEO-Wiki (Fritz et al. 2009, geo-wiki.org) are contributing data, specifically labeled for this task. Thus far, however, the coverage of the contributed data does not allow for exhaustive validation of large region datasets. In contrast to that, the OSM contributors are very actively collecting a broad range of thematic data with close to complete spatial coverage in certain areas (Ribeiro and Fonte 2015). Thus, utilizing the OSM as a source for land cover ground truth data is a promising approach. Since the OSM data is not specifically tailored to the needs of land cover map validation, various methods for transforming the original data to a more suitable representation have been developed (Fonte et al. 2015). While only a portion of the OSM attributes is valuable for a derived land cover map, the coverage is still high enough to be usable, especially in urban areas (Ribeiro and Fonte 2015).

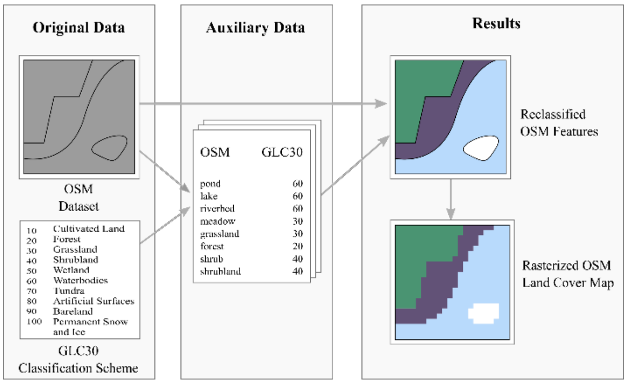

Fig. 1. Workflow used for creating land cover map from OSM data.

In our research, we implemented a method for deriving a land cover reference map from the OSM database as shown in Fig. 1. In order to preserve the entire content of the database, we use a complete XML-encoded extract of the OSM database, representing our study area, instead of pre-processed Shapefiles, as suggested by Fonte et al. (2016). For an efficient processing of the large data amount, we use a PostGIS database for our experiments. For the derivation of the land cover map, a subset of the OSM tags, namely “amenity”, “building”, “historic”, “land use”, “leisure”, “natural”, “shop”, “tourism”, and “waterway” is considered. We define a mapping from the OSM attributes to the classes used in the GLC30 classification scheme. The mapping is only conducted for polygon features, since point and line features do not provide immediate information about the coverage of an area. Exploiting additional information implicitly contained in point and line features, might be possible in general, using assumptions about specific feature classes, such as empirically determined road widths. In order to keep the used data as noise-free as possible, this was omitted in our experiments. In a final step, the vector data is rasterized to a 30 m grid with an appropriate Minimum Mapping Unit (MMU), merging small features with their neighboring features. The final OSM land cover map of our study area is shown in Fig. 2. The total coverage of our reference map is about 71% of the total study area.

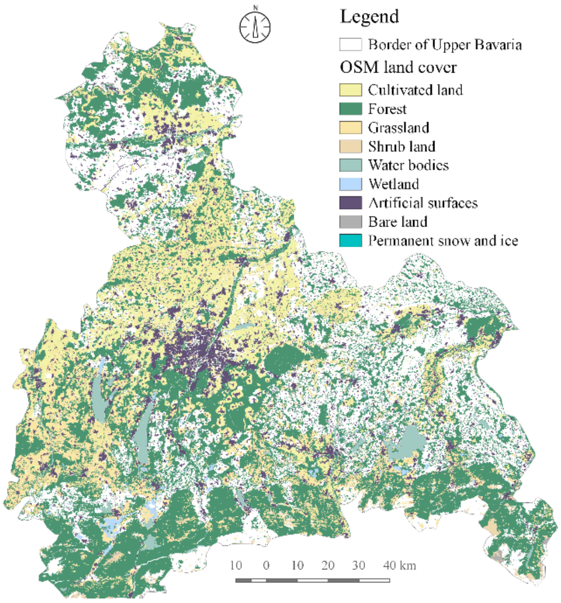

Fig. 2. Land cover map derived from XML-encoded extract of the OSM database.

2.3 CORINE Land Cover - CLC2006

The land cover mapping for the countries of European Union is realised within the programme CORINE (“Coordination of Information on the Environment”). Based on CORINE Land Cover 2000 for Germany, the data was updated and land cover product CLC2006 was produced as a vector database. According to Keil et al. (2011) the main data sources of the land cover and land use mapping were satellite images of Landsat 7 (for 2000) and IRS-P6 LISS III as well as SPOT-4 and SPOT-5. This product is following common European wide CLC nomenclature and consists of 44 classes, where 37 classes are relevant for Germany. Therefore, CLC2006 is characterized as result of the GIS derivation considering the land cover changes. The CLC2006 has 25m of minimum mapping unit (MMU) for the polygons. The data is released in the projections of Gauss-Kruger Zone 3, Gauss-Kruger Zone 4 or UTM Zone 32.

Several authors (Gallego 2001, Brovelli et al. 2015, Arsanjani et al. 2016) have proposed to assess the land cover quality using CORINE datasets for different study areas. Arsanjani et al. (2016) studied the validation of GlobeLand30 against CORINE for the area of Germany and showed that overall accuracy is 92%. Another solution was described in Brovelli et al. (2015) who indicated that the overall accuracy values of third level CORINE Land Cover are generally higher than 80%. In both cases the validation was made using confusion matrix that represents comparisons of a land cover map against the referenced dataset. In our study CLC2006, further called CORINE, was used as a variable for constructing Bayesian Network. Therefore, the polygon map CLC2006 was gridded for use in the model with the grid cell larger than MMU. The resolution of the gridded map is 30m. Moreover, we scaled down the class complexity in order to provide consistent classification for all the data sources. That is to say, 44 CORINE land cover classes were assigned to 10 classes in line with the GlobeLand30 classification (see the Table 1).

GlobeLand30 land cover classification	CORINE Pixel values	GlobeLand30 Pixel values
Cultivated	32 - 41	10
Forest	42 - 45	20
Grassland	46 - 47	30
Scrubland	48 – 49	40
Wetland	55 - 59	50
Water bodies	60 - 64	60
Tundra	-	70
Artificial surface	21 – 31	80
Bare land	50 – 53	90
Permanent ice and snow	54	100

Table. 1. CORINE (CLC2006) reclassification based on the GlobeLand30 land cover definition.

3. Methodology

Land cover classes derived from the satellite imageries are commonly used for various environmental studies. However, due to errors involved in the data acquisition and processing, uncertainties are introduced. The uncertainties of classified remote sensing data can be measured using different approaches such as comparison against ground truth, analysis of classification statistics (e.g., measures of separation in spectral space), comparisons between different land cover products (Quaife and Cripps 2016), or using an alternative method that involves Internet users for the data validation (Fritz et al., 2009). The scientific community proposes various methods to validate the classes and to model the quality and uncertainty. This article provides an approach of Bayesian Networks, which is widely used in diverse scientific domains such as medicine, weather forecasting and social science. Bayesian Networks have the property to address the spatial and temporal complexity of the dataset and produce reasoning based on evidences. Recently, Bayesian Networks are gaining importance for the quantitative and qualitative analysis of the remote sensing data (Quaife and Cripps 2016) because they are able to handle measurable information as well as qualitative criteria such as expert opinion or preferences based on the contribution of wide public of non-experts.

The underlying mathematical model of a Bayesian network is based on components such as Directed Acyclic Graphs (DAG) and conditional probability tables (CPTs) (Darwiche 2008). The nodes in DAG represent random variables, such as land cover classifications, and arrows among them describe dependencies among these variables. The process is organised in one-way direction, so that the child doesn’t transfer any feedback to the parent. On this note, the Bayesian approach is able to model the proportions of true values in selected pixels at each location across the whole study area. The data analysis might include integration of multiple land cover data, auxiliary data sets and expert knowledge representing the data of same nature and study area using consistent manner. The output of such evaluation consists of the probability maps and uncertainty information. Furthermore, this data might be visualized via geoportal, where the land cover data is disseminated or by means of stand-alone visualizations.

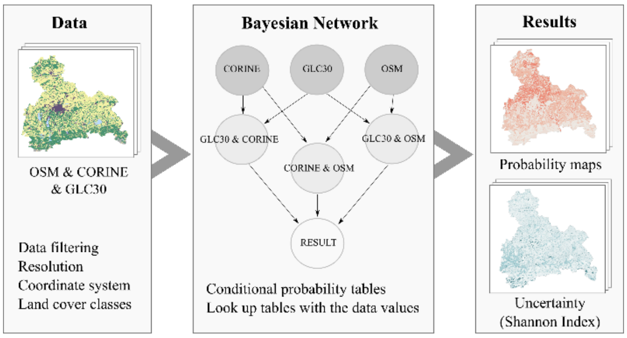

Fig. 3. Pipeline for using Bayesian Network.

The procedure for using Bayesian Network is depicted in Fig. 3. The graphical model defines the probability relations among the variables. In this research we treat the land cover classification GlobeLand30, CORINE and land cover derived from VGI data, namely OSM, as variable for the constructing Bayesian Network. We assume that all these three datasets are potentially misclassified, therefore each classification can be described in relation with other, assigning prior probabilities of each class. If we consider one of the variable (land cover classification) that doesn't have a parent node, the prior probability for this variable is assigned based on the possible values: $P\,(X_i = x_i)$, where $x_1,\ldots,x_n$ are all possible values (i.e., instantiations) of variable X_i. The nodes at the next level have parents, therefore the conditional probability is assigned. Hence, each variable is described by the probability under the state of its parent. If $x_i$indicates the values of variable X_i and $\rho\alpha_i$ indicated the set of values for X_i's parents, then $P(x_i|\rho\alpha_i)$ indicates the conditional probability, for example $P\,(X_3 = x_3\,|X_1 = x_1,\, X_2 = x_2)$.

Based on the probability theory the conditional probability is

$P(a|b) = P(a\,|b)/P(b)$ *or* $P(a,b) = P(a\,|b) \cdot P(b)$, (1)

Where $P\,(a,b)$ is a joint probability of event $a \wedge b$. Therefore,

$$P(a\,|b) \cdot P(b) = P(b|a)P(a) \qquad (2)$$

$$P(a\,|b) = P(b|a) \cdot P(a)/P(b) \qquad (3)$$

Equation 3 is the main concept supporting the Bayesian Networks modelling. Using this equation the probabilities of event P(a) can be updated based on the new evidence related to event b. Hence, based on the Bayesian theory it is possible to update the knowledge of a land cover class considering new/additional evidence (conditional probability). The reasoning is based on the degree of belief (posteriori probability).

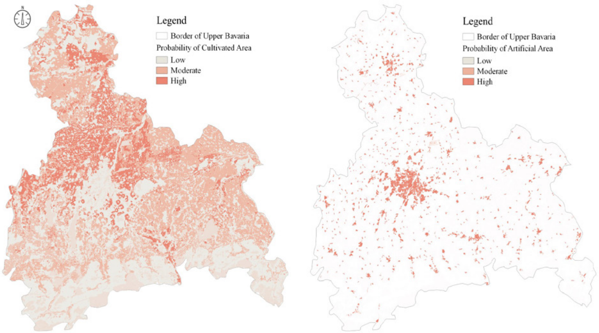

Fig. 4. Probability maps of cultivated and artificial land cover classes.

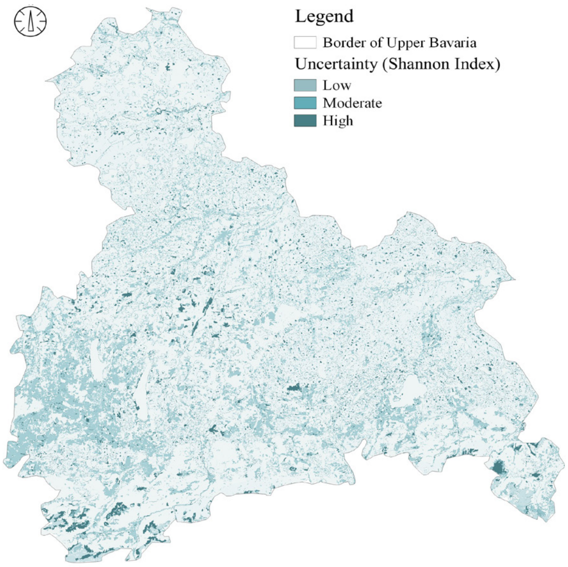

Fig. 5. Uncertainty (Shannon Index) map based on the data fusion of GlobeLand30, CORINE and land cover derived from OSM data.

Moreover, the effectiveness of the Bayesian Networks lies also in the ability to compute the conditional probability of the descendant nodes as well as parent nodes. One of the challenges when applying BN is to define the probability functions. This can be made via discretization process where the values of a variable are represented by discrete quantities. Therefore the range of the values is split into intervals defined by the land cover classification. The simulation was realized using R statistics and package for the spatial implementation of Bayesian Networks and mapping "bnspatial". The outputs of the simulation processes are the expected state of target node (i.e. the state with the highest relative probability) and the uncertainty, expressed as entropy by the Shannon index. The example of the output maps is illustrated in the Fig. 4 and 5.

4. Results

The Bayesian Networks are an effective technique for the analysis of remote sensing data, especially for the reasoning based on diverse sources with varying degrees of reliability. The outputs of this research work are posterior probability maps, and the map of uncertainty measured as Shannon index (entropy) (see the figure 3). The maps of posterior probability depict updated prior probability of land cover class occurrence considering the information from different input datasets. The map of uncertainty was elaborated to depict the diversity in the data and it illustrates Shannon entropy applied for the land cover classification. Therefore, the latter map shows the amount of various information that each location might be assigned to. The aim of the output maps is to highlight the areas with the highest degree of the ambiguity and provide visual guidance for the further land cover validation as well as for a deeper understanding of the dynamics related to the high uncertainty. As it can be seen from the Fig. 6, the uncertainty of a polygon is high, and when we compare to the original data sets, it is evident that the highlighted area was defined in inconsistent manner. Hence, the attention for the validation should be placed at such areas. Moreover, such data may be utilized to guide the implementation of decision support tools.

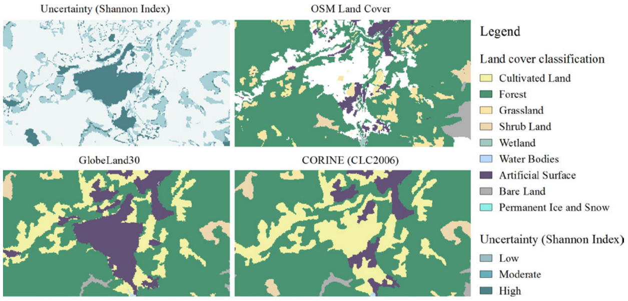

Fig. 6. Interpretation of uncertain areas based on land cover classification from GlobeLand30, CORINE (CLC2006) and Open Street. Map.

5. Conclusion

The proposed approach offers a technique for the identification of uncertain areas associated with the output from each additional source that can be further visualized within a geoportal of GlobeLand30. The available uncertainty information may guide the quality improvement of GlobeLand30 by involving the ground truth data, information with superior quality, the know-how of experts and the crowd intelligence. This may finally pave a way towards a seamless validation of global land cover data, which is beyond the current validation approach by image processing experts using a limited amount of sample data for selected regions. Moreover, it will trigger a targeted knowledge discovery in the areas with higher uncertainty values.

Based on the results we see a possible solution in using statistical methods of data fusion to discover uncertain information. The originality of our approach lies in using the Bayesian Networks technique for analysis of global land cover data along with the data derived from VGI, namely OSM. This study shows that Bayesian Networks with multiple datasets have potential to reveal the hidden patterns that cannot be found by linear processing. However, the findings have number of possible limitations, namely the quality of existent reference data, quality of volunteered contributions to the OSM and prior expert knowledge about the used data. Therefore, due to the mentioned shortcomings, we cannot claim that the results highlight misclassified areas. However, the discovered patterns show higher degree of entropy and enable visual guidance for data exploration. The data obtained indicate that the more precise the prior expert knowledge and the higher data quality of the referenced datasets, the better results could be achieved. Therefore, under ideal conditions, the GLC classification could be significantly improved.

Further research is needed to develop an application for the visual and analytical reasoning under uncertainty of land cover classification. Much research addressed the issue of data uncertainty and its signification. Therefore different visualization techniques were implemented and tested for a variety of data types. However, it was addressed by MacEachren (2015) that just little attention has been given to reasoning/decision-making under uncertainty. The Bayesian approach can integrate different GIS layers and the expert knowledge to analyse probability of each class occurrence and its related uncertainty. Therefore beyond the data fusion and data visualisation, the further work will contribute with the decision-making environment for the remotely-sensed data analysis.

6. Acknowledgements

This work is supported by the TUM International Graduate School of Science and Engineering (IGSSE) and Chair of Cartography, TU Munich.

7. References

Arsanjani, J. J., See, L., and Tayyebi, A. (2016). Assessing the suitability of GlobeLand30 for mapping land cover in Germany. International Journal of Digital Earth, 9(9):873–891.

Ban, Y., Gong, P., and Giri, C. (2015). Global land cover mapping using Earth observation satellite data: Recent progresses and challenges. ISPRS Journal of Photogrammetry and Remote Sensing, 103(February):1–6.

Belward, A. S. and Skøien, J. O. (2014). Who launched what, when and why; trends in global land-cover observation capacity from civilian earth observation satellites. ISPRS Journal of Photogrammetry and Remote Sensing, 103:115–128.

Brovelli, M., Molinari, M., Hussein, E., Chen, J., and Li, R. (2015). The First Comprehensive Accuracy Assessment of GlobeLand30 at a National Level: Methodology and Results. Remote Sensing, 7(4):4191–4212.

Chen, J., Chen, J., Liao, A., Cao, X., Chen, L., Chen, X., He, C., Han, G., Peng, S., Lu, M., Zhang, W., Tong, X., and Mills, J. (2015). Global land cover mapping at 30m resolution: A POK-based operational approach. ISPRS Journal of Photogrammetry and Remote Sensing, 103:7–27.

Cope, M. (2015). Commentary: Geographies of digital lives: Trajectories in the production of knowledge with user-generated content. Landscape and Urban Planning, 142:212–214.

Darwiche, A. (2008). Chapter 11 Bayesian Networks - Handbook of Knowledge Representation. Foundations of Artificial Intelli-gence, 3(07):467–509.

Fonte, C. C., Bastin, L., See, L., Foody, G., and Lupia, F. (2015) Usability of VGI for Validation of Land Cover Maps. International Journal of Geographical Information Science, 29(7), 1269–1291.

Fonte, C., Minghini, M., Antoniou, V., See, L., Brovelli, M. A., Milčinski, G. (2016). An Automated Methodology for Converting OSM Data into a Land Use/Cover Map. Proceedings of the 6th International Conference on Cartography & GIS, 462–473.

Fritz, S., McCallum, I., Schill, C., Perger, C., Grillmayer, R., Achard, F., Kraxner, F., and Obersteiner, M. (2009). Geo-wiki.org: The use of crowdsourcing to improve global land cover. Remote Sensing, 1(3):345–354.

Gallego, J. (2001). Comparing CORINE Land Cover with a more detailed database in Arezzo (Italy). Jrc, pages 1–8.

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. GeoJournal, 69(4):211– 221.

Helbich, M., Amelunxen, C., Neis, P. (2012). Comparative Spatial Analysis of Positional Accuracy of OpenStreetMap and Proprietary Geodata. Proceedings of GI_Forum, 24–33.

Keil, M., Bock, M., Esch, T., Metz, A., Nieland, S., and Pfitzner, A. (2011). CORINE Land Cover Aktualisierung 2006 für Deutsch-land. Workshop CORINE Land Cover 2000 in Germany and Europe and its use for Environmental Applications, TH-62-04-1:67.

Kinkeldey, C. (2014). A Concept for Uncertainty-Aware Analysis of Land Cover Change Using Geovisual Analytics.” ISPRS International Journal of Geo-Information 3(3): 1122-1138.

MacEachren, A. M. (1992). Visualizing uncertain information. Cartographic Perspective, vol. 13, no. 13, pp. 10–19.

MacEachren, A. M. (2015). Visual Analytics and Uncertainty: It’s not about the Data. EuroVis Workshop on Visual Analytics (Eu-roVA).

Neis, P., Zielstra, D., and Zipf, A. (2012). The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007–2011. Future Internet, 4(1), 1–21.

Quaife, T. and Cripps, E. (2016). Bayesian Analysis of Uncertainty in the GlobCover 2009 Land Cover Product at Climate Model Grid Scale. Remote Sensing, 8(4):314.

Ribeiro, A., and Fonte, C. C. (2015). A Methodology for Assessing OpenStreetMap Degree of Coverage for Purposes of Land Cover Mapping. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, II-3/W5, 297–303.

Sun, B., Chen, X., Zhou, Q., Tong, K., Kong, H., and Asia, C. (2015). Uncertainty Assessment Of Globeland30 Land Cover Data Set Over. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLI-B8.

Townshend, J. and Masek, J. (2012). Global characterization and monitoring of forest cover using Landsat data: opportunities and challenges. Journal of Digital Earth, 5(5):373–397.