Ambiguity and Plausibility: Managing Classification Quality in Volunteered Geographic Information

Ahmed Loai Ali[1,3], **Falko Schmid**[1,2], **Rami Al-Salman**[1], and **Tomi Kauppinen**[1,4]

[1]Cognitive Systems Group, University of Bremen, Germany

[2]SFB/TR 8 Spatial Cognition, University of Bremen, Germany

[3]Faculty of Computers and Information, Assiut University, Egypt

[4]Department of Media Technology, School of Science, Aalto University, Finland

{*loai,schmid,rami,tomi.kauppinen*}*@informatik.uni-bremen.de*

ABSTRACT

With the ubiquity of technology and tools, current Volunteered Geographic Information (VGI) projects allow the public to contribute, maintain, and use geo-spatial data. One of the most prominent and successful VGI project is OpenStreetMap (OSM), where more than one million volunteers collected and contributed data that is obtainable for everybody. However, this kind of contribution mechanism is usually associated with data quality issues, e.g., geographic entities such as gardens or parks can be assigned with inappropriate classification by volunteers. Based on the observation that geographic features usually inherit certain properties and characteristics, we propose a novel classification-based approach allowing the identification of entities with inappropriate classification. We use the rich data set of OSM to analyze the properties of geographic entities with respect to their implicit characteristics in order to develop classifiers based on them. Our developed classifiers show high detection accuracies. However, due to the absence of proper training data we additionally performed a user study to verify our findings by means of intra-user-agreement. The results of our study support the detections of our classifiers and show that our classification-based approaches can be a valuable tool for managing and improving VGI data.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data mining and Spatial databases and GIS; I.2.1 [**Applications and Expert Systems**]: Cartography

General Terms

Management, Standardization.

SIGSPATIAL'14, November 04 - 07 2014, Dallas/Fort Worth, TX, USA
Copyright 2014 ACM 978-1-4503-3131-9/14/11$15.00
http://dx.doi.org/10.1145/2666310.2666392.

Keywords

Volunteered Geographic Information, Spatial Data Quality, Machine Learning, Geographic Information Systems

1. INTRODUCTION

During the last decade, the ubiquity of location-aware devices (e.g., smartphones) enables the public to collect, contribute, edit, and use geographic information — activities formerly exclusively conducted by national mapping agencies and professional organizations. The phenomenon is known as Volunteered Geographic Information (VGI) [12]. Due to its large success and openness, data generated by VGI projects became part of a common, globally available Spatial Data Infrastructure (SDI) and plays a significant role in Geographic Information Systems (GIS) [22].

The advancement of Web technologies and the availability of open source software lead to the increasing numbers of VGI projects, such as OpenStreetMap[1] (OSM). OSM is one of the most common VGI projects, with the aim to provide a free editable world map. A large number of contributors are producing and improving large scale geographic data sets covering many parts of the world [16]. OSM has no restriction about the spatial data to be contributed, and its rich data set enables numerous different applications — including but not limited to map provision, routing, planning, geo-visualization, and point of interests (POI) search. Applications require reliable and consistent data, which is not guaranteed with VGI data [10] in contrast to "official" data collected by authorities. Nevertheless, VGI is a potential alternative for authoritative data: it is typically open and free, dynamically and frequently updated, and employs crowdsourcing forces to ensure the quality [13].

The increasing number of OSM contributors, the vast amounts of daily contributions, and the loose classification system trigger questions about the resulting data quality. The large number of heterogeneous contributors fosters data of mixed quality: they have different perspectives, contribute for different purposes, and use different contribution technologies and tools. Data quality in VGI has been studied from different perspectives and identified a number of crucial constituents for quality issues and mechanisms.

[1]`http://www.openstreetmap.org`

In this work, we address VGI data quality from the perspective of classification plausibility. In OSM, there is no explicit classification system, just recommendations. If an "water" area is classified as "lake" or "pond" — the decision is up to the contributors and based on their conceptualization of space, and their knowledge and considerations of the provided recommendations. Due to a certain degree of conceptual ambiguity, in many cases multiple classes are applicable for an entity; if a piece of land is "grass" or "meadow", "garden" or "park" depends on the context and purpose of data collection. Additionally, missing hard constraints make it hard to clearly decide. As a result, a significant amount of data is inappropriately classified and can cause errors whenever addressed by algorithms, such as rendering, analysis, or routing algorithms.

However, in many cases one classification is more applicable than others, as comparable pieces of land might have certain comparable intrinsic properties: parks are usually more than just an area covered with grass, parks in many cases contain ways, trees, water bodies, etc.

In this paper, we attempt to tackle the problem of classification ambiguity and the resulting quality issues. In our approach we analyze the properties of potentially ambiguous classes with respect to their inherent structure. We use these properties and build classifiers with the aim to identify entities with a potentially inappropriate classification. To validate the promising results of our approach, we conducted a user study with a subset of the identified entities. Based on the findings of the intra-user-agreements of our participants, we have a strong support for the approach and the general applicability of automatic quality checking approaches. Our results also raise questions about remote (non-local) classification of entities of unclear characteristics.

2. RELATED WORK

In VGI, contributors produce geographic information without necessarily being educated surveyors or cartographers. The motivation for contribution can be highly diverse, and the quality of contributions also depends on the used equipments and methods. Thus, the combination of diverse educational backgrounds, different views on required data and its quality, as well as technical constraints lead to data of mixed quality. Due to the increasing significance of VGI questions concerning data quality, credibility, and reliability are increasingly studied [8, 10].

Quality of VGI data has various perspectives and notions: completeness, positional accuracy, attribute consistency, logical consistency, and lineage [6, 13]. As most VGI projects, OSM does not have data quality specifications or standard procedures as implemented by mapping agencies. The quality of VGI data can be assessed by two different methods: comparison with respect to reference data and intrinsic data analysis (which can be implemented by crowdsourcing approaches, social measures, or geographic consistency analysis [12, 13]). In [11, 15, 21] the authors compare OSM data to reference data, in [15, 21] the authors are able to show a high overall positional accuracy of OSM data in comparison with authoritative data. In terms of completeness, some studies conclude that some areas are well mapped and complete, however with a tight relation of completeness and urbanization [15, 25]. On the other hand, the following intrinsic methods and mechanisms are applied and proposed to ensure VGI data quality:

- *Crowdsourcing revision*: data quality can be ensured by means of crowdsourcing, thus by checking and editing of entities by multiple contributors.
- *Social measures*: this approach focuses on the assessment of contributors themselves as a proxy measure for the quality of their contributions [18].
- *Geographic consistency*: this approach analyzes the consistency of contributed entities with their geographic context, i.e., contextually implausible entities will be detected (e.g., a building in a lake).

Examples for intrinsic analysis methods are in e.g., [2] presenting 25 methods to assess VGI quality without the need for authoritative data. The methods are focused around "fitness for purpose" approach. In [19, 26] the authors analyze intrinsic information, such as tracking edits history, and contributor's reputation analysis. In [5, 18] the authors use trustworthiness as a proxy to assess the quality. [23] assesses data quality by analyzing the frequently edited entities by correlating the number of tags and the number of contributors associated with an entity.

Different aspects influence the quality of VGI data, e.g., the combination of loose contribution mechanisms, and the lack of strict mechanisms for checking the integrity of new and existing data are major sources of heterogeneous quality of VGI data [23]. Amongst others, semantic inconsistency is one of the essential problems of VGI data quality [8]: for instance, different classes represent the same geographic phenomena (*synonymy*), or one class describes different geographic phenomena (*polysemy*). In [24] and [30] the authors present methods for improving the semantic consistency of VGI. The analysis of semantic similarity is applied to enhance the quality of VGI through suggesting tags and detecting outliers in existing data [24, 30]. Another approach for tackling quality issues is the development of appropriate interfaces for the data generation and submission. In [28, 27] the authors demonstrate that task-specific interfaces support the generation of high quality data even under difficult conditions.

3. AMBIGUITY AND PLAUSIBILITY

In this work, we focus on the classification of entities as a facet of data quality. Classification ambiguity of spatial entities can be a fundamental source of data quality problems [6, 14]. Particularly in VGI, contributors are often nonexperts with no formal surveying or cartographic education. The diversity of cultural and educational backgrounds, conceptualization of spatial entities and understanding of recommendations lead to heterogeneous classifications. On the one hand local concepts should be preserved. While on the other hand as homogeneous data as possible is required to allow the development of global, uniform applications (e.g., map rendering or routing).

In OSM, the majority of contributors contribute data by annotating satellite imagery [10]. If mappers are not familiar with the area they map, this method makes it hard to identify the correct class for an entity: crucial details might not be visible on the (currently) low resolution imagery, or features can be wrongly interpreted. For instance a green area with scrub and trees might be classified as "scrub", "grassland", or "meadow". However this area could also be a "park" or a "garden". Without having local knowledge, some entities are hard to classify.

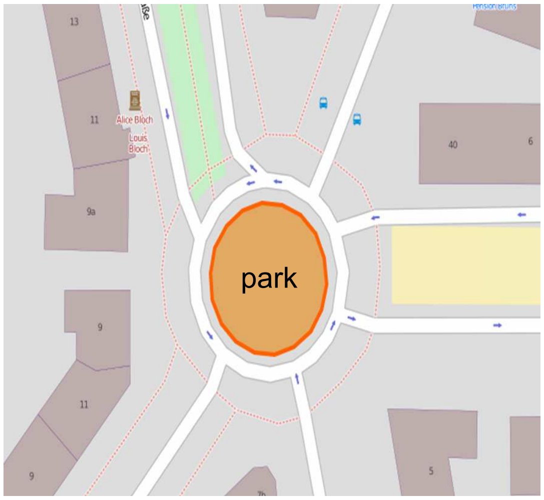


Figure 1: Inappropriate Classification: a "park" placed in a roundabout.

From other perspective, when mappers have local knowledge they contribute based on their personal perspectives [26], thus the diverse backgrounds and sometimes missing knowledge about the recommendations for contribution result in classification problems. In other cases, the recommendations themselves might be vague and an entity might belong to multiple classes. For example, an area covered by grass could be classified as a "grass", "meadow", or "grassland". Thus, an individual entity can have multiple valid classifications.

Whenever an entity can potentially belong to several classes, we call this *Classification Ambiguity*. Whenever we want to express the likelihood of an entity belonging to a specific class, we call it *Classification Plausibility*. In some cases the properties of the contributed entity indicate that the plausibility of an assigned class might be very low and indicate the class was most probably not chosen correctly. In this case we call it *Inappropriate Classification*. Figure 1 shows an example of a inappropriate classification: the green area in the center of a roundabout is tagged to be a "park" — typically parks are larger, have a certain degree of contained infrastructure, and are not placed in rather small roundabouts. According to OSM classification recommendations, this area should be "grass".

3.1 Classification by Tagging

In OSM, data is classified by means of tags of the form key = "$value$". Different tags are used to describe different properties, e.g., the tag $leisure$ = "$value$" is commonly used to describe entities with a recreational purpose, while $landuse$ = "$value$" reflects the primary use of the land by humans. In OSM tagging is not restricted and the same entity can be assigned with numerous combinations of tags. Nevertheless, some combinations are applicable, while others are misleading or contradictive. Our approach aims to check the classification integrity of an entity by inspecting its properties.

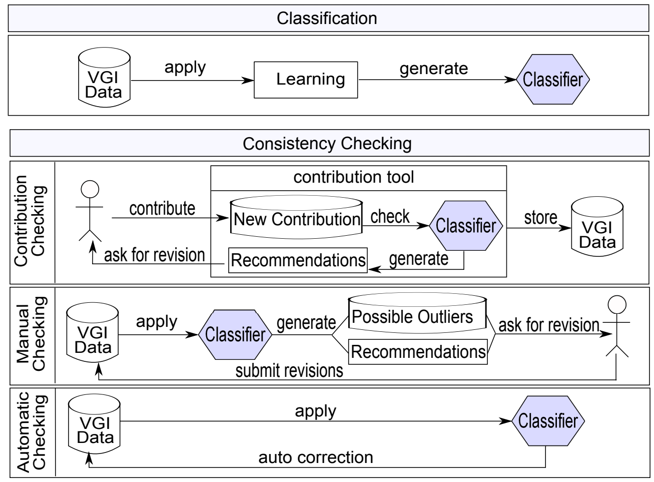


Figure 2: Learning-based approach to tackle classification plausibility.

4. LEARNING AND CROWDSOURCING

The increasing amount of VGI data - in particular OSM data - allows the application of machine learning algorithms as one of the possible methodologies to analyze and improve its data quality. We can select parts of certain entities in the database, learn their properties in form of a classifier, and apply the developed classifier on the entities of the database. The results tell us how well entities match to the learned properties. Figure 2 illustrates the approach of using learning for quality assurance as introduced in [1]. The approach consists of two phases: *Classification*, and *Consistency Checking*.

The *Classification* phase aims to develop a robust classifier based on data of sufficient quality. According to previous studies OSM data is of good quality in some areas [15, 21]; we can processes OSM data to extract an appropriate data set for learning the classifier. In the *Consistency Checking* phase, three scenarios for applying the developed classifier are possible: 1) *Contribution Checking* uses the classifier during the data contribution phase in an editor tool. The tool informs the contributor about the potential problematic data based on the classifier. The contributor can consider the hints generated by the tool and take action for correction if required. 2) *Manual Checking* refers to the manual validation of detected entities by volunteers, potentially inappropriately classified entities are presented to volunteers and validated by them. Regarding OSM data, there exists a number of applications, such as MapRoulette[2], MapDust[3], KeepRight[4], and OpenStreetBugs[5] improve the data quality. They typically check the integrity of entities against a set of rather static rules such as entities without name, roads without information about speed or driving direction, or entities marked by users for further inspection. If such systems or OSM editors are fed by entities detected by a learning approach as we propose, potential candidates with inappropriate classification can be identified and fixed by volunteers. 3) *Automatic Checking*, tries to automatically

[2] http://maproulette.org/
[3] http://www.mapdust.com/
[4] http://keepright.ipax.at/
[5] http://openstreetbugs.schokokeks.org/

detect and correct inappropriate classifications without human assistance.

However, as there is no clear reference data set to train the classifier, the results need to be interpreted with care. We deal with all kind of spatial real world entities, i.e., entities can belong to a certain class, although they might have rather unlikely characteristics (e.g., very small parks or huge private gardens).

4.1 Tackling Classification Plausibility

In this paper we are interested to check the classification plausibility of VGI data. One key idea is to preserve the locality of the data. During the classifier development, we maintain the locality within a given region for learning and applying the developed classifier. For example, learning from data of China and applying the extracted knowledge on data of the UK might return misleading results: they have different cultures (finding their expression also in the characteristics of spatial entities) and might have different conceptualizations of space. Thus, we follow the locality assumption of Tobler's law [29]. For this work we interpret Tobler's law as follows: cities in the same country have a closer concept for the same class of entity than cities of different countries, i.e., when we analyze data in Germany, we do not use this results to validate data in the UK.

5. CLASSIFICATION OF AMBIGUOUS AREAS

In our work, we focus on a set of classes with a certain degree of intrinsic ambiguity: areas that are typically rendered as green areas on a map. In OSM, amongst others these are entities tagged as "garden"," grass"," meadow", or " park". We chose these four classes as they represent a good example for classifications ambiguity. Conceptually, those entities have a certain degree of mutual ambiguity: parks and gardens share many characteristics, if a grass-covered area is just "grass", " meadow", or " garden" or " park" depends on the usage, conceptualization, or a legal definition.

The OSM recommendations[6] for the four classes are:

- Garden: *"a distinguishable planned space, usually outdoors, set aside for the display, cultivation, and enjoyment of plants and other forms of nature. The most common form is known as a residential garden, it is a form of garden and is generally found in proximity to a residence, such as the front or back garden."*
- Grass: *"a smaller areas of mown and managed grass for example in the middle of a roundabout, verges beside a road or in the middle of a dual-carriageway."*
- Meadow: *"a land primarily vegetated by grass plus other non-woody plants."*
- Park: *"an open, green area for recreation, usually municipal. These are outdoor areas, typically grassy or green areas, set aside of leisure and recreation. Typically open to the public, but may be fenced off, and may be closed; e.g., at night time."*

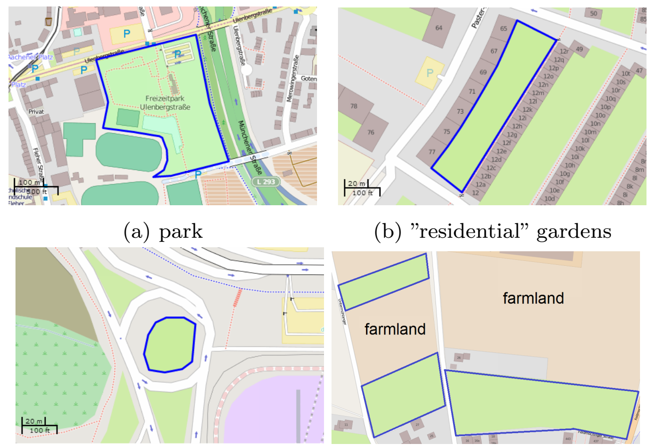
In OSM, these entities are contributed under various tags. They are commonly contributed with tags like *leisure* = "*value*", and *landuse* = "*value*".

[6] http://wiki.openstreetmap.org/wiki/

5.1 Selection of Classification Properties

To be able to distinguish between similar classes it is necessary to look into the characteristics and properties of each class. To develop a robust classifier we need to understand the properties of the entities to be classified. We apply not only the analytical methods, reflecting typical observable characteristics, but also statistical methods to explore the characteristics that are not immediately observable. In our approach we combine both methods.

Figure 3 shows typical entities of interest. Figure 3a depicts a "park" containing a playground, sport center, and paths. Figure 3b illustrates a *residential* "garden" surrounded by *residential* houses. Figure 3c shows a typical "grass" entity not containing other infrastructural entities and usually surrounded by or meet roads. Figure 3d shows "meadow" entities next to farmland and not containing other infrastructural entities.



(a) park (b) "residential" gardens

(c) grass (d) meadow

Figure 3: Samples of typical entities of interest.

These examples illustrate that geographic entities have basically two different types of properties: geometric (e.g., size and shape) and geographic properties (e.g. topological properties). In our previous work [1], we developed classifiers based on geometric properties to distinguish between entities of the classes "park" and "garden". This property is also observable in Figure 3: parks are usually larger than gardens. However, building classifiers for multiple classes requires the analysis of more properties, as size of entities can be similar, but their characteristics might be fundamentally different.

5.1.1 *Geometric Properties: Size*

Some entities are classifiable by considering their size. Figure 4 shows the average area of our entities of interest within the ten densest cities in Germany and the UK. "Meadows" and "parks" are usually larger than "grass" and "gardens". However, "meadows" and "parks" are as close as "grass" and "gardens". Thus, an entity's size will not be enough to distinguish between the four classes. In this study, we use the size of entities only as one of classification properties.

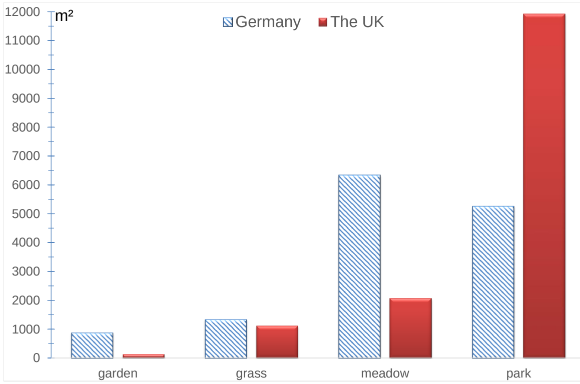


Figure 4: Average areas for the classes "garden", "grass", "meadow", "park" in Germany and the UK.

5.1.2 Analytical Context Properties

In addition to the OSM recommendations, the four entities of "garden", "grass", "meadow", and "park" are characterized by their internal and external context (see Figure 3 for examples). I.e., the kind of entities surrounded or contained in them influence and define their functionality and consequently their classification. For instance, "parks" typically contain other entities such as paths, playgrounds, and water bodies, whereas "grass" and "meadows" are rather unlikely to contain much infrastructure like this. Many of these relations are observable in the real world, and we tried to formulate a reasonable set of rules based on intensive visual analysis and data consultation.

We analyze the topological relations between pairs of entities by means of the 9-Intersection Model (9IM) [7]. As depicted in Figure 5, the 9IM distinguishes eight topological relations holding between two regions: *equal*, *disjoint*, *meet*, *overlap*, *contains*, *covers*, *inside*, and *coveredBy*.

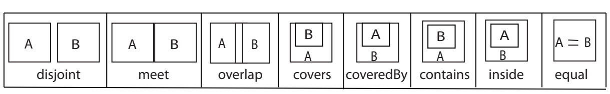


Figure 5: The eight distinct topological relations of the 9-intersection model.

In this study we consider three topological relations *meet*, *overlap*, and *contains*. These relations add distinct information to the classifier. We neglect the other relations due to three reasons: (a) *equal* and *covers* rarely hold among the entities of interest (e.g., a park is usually does not cover another entity), (b) *coveredBy* and *inside* are the inverse of *covers* and *contains* respectively, and (c) *disjoint* does not add additional information for the classification process. To find out about the characteristics of our example entities, we analyzed the features that are often contained by, overlap, or meet with "gardens", "grass", "meadow", "parks".

Following relations are part of the classifier, as they can be often observed in the data set:

- Meet with (areal) entities ($meet_A$): residential "gardens" often meet with (residential) houses. Additionally, as our analysis showed, "grass" often meet with houses as well, whereas "parks" and "meadows" are rather unlikely to meet with houses at all.
- Meet with (linear) entities ($meet_L$): in many cases, roads lead into and surround "parks" and public "gardens". They are often surrounded by fences as well.
- Overlap with (areal) entities ($overlap_A$): within a city, "parks" and "gardens" are often overlapped by residential areas, while "meadows" are usually overlapped with farmland entities.
- Overlap with (linear) entities ($overlap_L$): "grass" areas are often overlapped by roads, since they are often located next to highways and roundabouts.
- Contains (areal) entities ($contains_A$): one key property of the classifier is the containment property. The more entities are located inside the green area, the more likely the entity belongs to leisure-related entities, thus a "park" or public "garden".
- Contains (linear) entities ($contains_L$): "parks" and public "gardens" usually contain ways for bicycles, pedestrians, and sometimes cars, whereas "grass" or a "meadows" are unlikely to contain any of those entities.

5.1.3 Statistical Context Properties

In order to understand the characteristics of the geographic context of the interested entities, we investigate the keys of entities that are involved in the topological relations described above. Analytical context properties (as described in previous section) are observable in the environment and can be found in many instances. However, from the viewpoint of data, we can derive more properties based on the classification. To identify them, we utilized a straightforward statistical analysis to derive the set of keys that are both frequently hold in the relations to add distinct information to the classifier. We used all keys with an absolute occurrence of $\geq 2\%$ (below 2% there is a huge set of keys with rather low information gain, such as administrative boundaries). The selected keys for areal entities the keys are: "amenity" (5%), "building" (44%), "landuse" (23%), "leisure" (10%), "natural" (6%), and "sport" (2%). As well, for linear entities we selected the keys of: "barrier" (6%), "bicycle" (15%), "foot" (12%), "highway" (63%) and "waterway" (3%).

In general, the analysis of geometric properties (Section 5.1.1) and spatial context properties (Sections 5.1.2 and 5.1.3) can be adapted to the characteristics of any kind of areal geographic entities. Definitely, the kind of entities involved in the investigated topological relations will depend on the type of classes of interest.

5.2 Classifier Development

The development of a classifier involves two phases: training and validation. The aim of the *training* phase is to train the classifier to distinguish between classes based on the classification properties. In the *validation* phase we test the validity of the generated classifier [3].

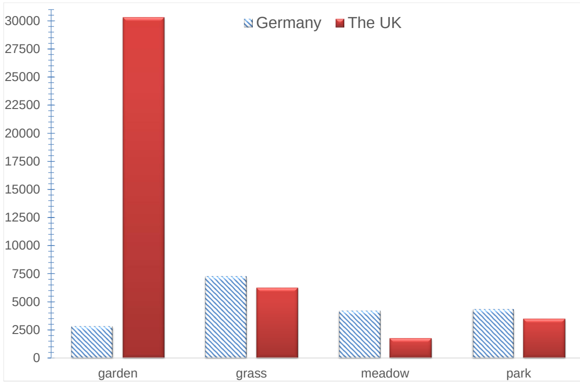


Figure 6: Number of "garden", "grass" ,"meadow" ,"park" entities in Germany and the UK.

5.2.1 Classifier Training

In this study, the training set consists of "park", "garden", "grass" and "meadow" entities extracted from OSM data set, D_{train} =$\{E_1, E_2, ..., E_n\}$. Each Entity E is represented by a set of properties and assigned to a class C, $E = \{$*size, meet*$_A$*, meet*$_L$*, overlap*$_A$*, overlap*$_L$*, contains*$_A$*, contains*$_L$*, amenity, building, landuse, leisure, natural, sport, barrier, bicycle, foot, highway, waterway*, $C\}$, where $C \in \{$garden, park, grass, meadow$\}$. The training process tries to identify a function, $f(E) = C$, to predict the class C of an entity E.

Building a classifier can be done by using *Eager Learning* (EL) or *Lazy Learning* (LL). In EL a training set is used to build a complete classifier before receiving any test entities. Bayesian classification, support vector machines (SVM), neural network (NN), and decision trees are examples for EL algorithms. On the contrary in LL, generalization beyond the training data is delayed until a query is made to the system. K-nearest neighbours (KNN) and case based reasoning (CBR) are examples of lazy learning [3, 17]. In OSM a set of pre-classified entities is already stored, and the classification process is performed at arrival of a new entity. The new entity is classified based on similarity to the existing entities. Hence, we use the lazy learning paradigm to develop the classifier.

In particular, we use KNN [4, 32] for building a classifier KNN classifies entities based on the closest training examples. An unclassified entity is classified by checking the K nearest classified neighbours. The similarity between the unclassified entity and the entities stored in training dataset is calculated by euclidean distance.

5.2.2 Classifier validation

The aim of the validation process is to check the classifier's generalization ability. Thus, several test sets are applied on the same classifier to determine its performance. There exists more than one measure to determine a classifier performance, however, depending on just one measure could introduce bias [3]. We use two measures to assess the classifier performance: the accuracy and the area under the Receiver Operation Characteristics (ROC) curve.

The accuracy measure of a classifier is the percentage of correctly classified entities on a given test set. In some cases accuracies are biased due to overfitting or underfitting [3, 17]. A reason can be an unbalanced population of the training or the test set. For example, Figure 6 shows the majority of "garden" entities, in the UK, over the others. This phenomena can influence the classifier performance. Thus, we utilize more than one measure to assess the resulting classifiers. The (ROC) curve is a useful measure to assess the performance of a classifier [9, 32]. In particular the Area Under the ROC Curve (AUC) is a useful measure to evaluate a classifier. The closer the value of AUC is to 1.0, the higher its performance. According to [9], good classifiers should have AUC value between 0.5 and 1.0.

6. EMPIRICAL STUDY

To evaluate our approach and the derived classifiers, we performed an empirical study. We used OSM data of Germany and the UK. According to [21], and [16], OSM data for Germany and the UK is of acceptable quality.

6.1 Data Preprocessing

We do not have a reference data set to assess the classifier performance. I.e., to set up training and test data for the classifiers we need to identify a subset of the OSM data which is of sufficient quality. It has been shown that mapping activities of individual contributors and the frequency of edits are good indicators for quality [23, 26], thus we selected entities with a high number of edits and contributed by trustworthy users.

In OSM, every edit is stored as new *version* of the edited entity. Additionally, a collection of all edits of a particular contributor over 24 hours are stored in a *changeset*. For each entity we stored the last version number and the contributor ID. The contributors themselves are categorized based on the work in [26]: *New registered* (1 changeset), *Non-recurring* (up to 10 changesets), *Junior* (up to 100 changesets), *Senior* (up to 500 changesets), *Senior*$^+$ (up to 2000 changesets), *Gold* (more than 2000 changesets).

The data we used was extracted from OSM on December 2nd, 2013. During the development of our classifiers, we maintained the locality of each country by developing different classifiers for both regions: we used the data of the ten most densest cities (population/city area) of both countries. The data of the most densest cities was selected to ensure a data with active contributor communities and hence data of sufficient quality. In Germany, we utilized data of *Berlin, Bremen, Cologne, Dortmund, Dusseldorf, Essen, Frankfurt, Hamburg, Munich, and Stuttgart.* As well in the UK we utilized data of *Birmingham, Bradford, Bristol, Edinburgh, Glasgow, Leeds, Liverpool, London, Manchester, and Sheffield.*

	Germany	**The UK**
Entities of the ten most densest cities (D)	19,088	41,822
Entities of active mappers ($D_{top\_mappers}$)	14,736	38,186
Entities with freq. edits ($D_{top\_versions}$)	2,080	854

Table 1: Extracted data from Germany and the UK.

Table 1 summarizes the facts of the extracted data of Germany and the UK. In developing the classifiers we utilized the data of the ten most densest cities (D). From D, we extracted two data sets for the classifiers validation process: $D_{top\_mappers}$ and $D_{top\_versions}$. $D_{top\_mappers}$ contains entities of highly active mappers (Senior^{+} and Gold mappers), while $D_{top\_versions}$ contains frequently edited entities with more than five versions.

6.2 Classifier Learning

In order to learn our classifiers efficiently, we extracted multiple data sets for the training and validation process. We developed classifiers based on two different models: Label-Based Model (*LBM*) and Tag-Based Model (*TBM*).

In *LBM*, we trained the classifiers to distinguish between the four classes. We utilized D in training the classifiers. Afterwards, the classifiers are validated using D, $D_{top\_mappers}$, and $D_{top\_versions}$. Table 2 shows the results of the classifiers performances measures; accuracy (Acc.) and AUC.

	D		$D_{top\_mappers}$		$D_{top\_versions}$	
	Acc.	AUC	Acc.	AUC	Acc.	AUC
GER	60.4 %	0.85	64.7 %	0.86	67.8 %	0.86
UK	88.3 %	0.98	92.0 %	0.99	75.2 %	0.84

Table 2: LBM classifiers performance of data extracted from Germany (GER) and the UK.

From Table 2, we calculate the average performance of the classifiers for each country. The classifier for Germany has an average accuracy of 64.3%, and AUC equal 0.85. The UK classifier has a higher performance: it has an average performance with an accuracy of 85.1% and AUC equal 0.93.

The unbalanced data in *LBM* has an influence on the performance of the classifiers (see Figure 6 for details). Additionally, the four classes represent two pairs of entities belonging to two different tags (*leisure* = "*value*" and *landuse* = "*value*"). As discussed in Section 3.1, selecting a proper tag is crucial for a plausible classification. Hence, we developed the *TBM* classifiers that distinguish between two tags: *leisure* = "*value*" and *landuse* = "*value*". In the *TBM*, both "park" and "garden" entities belong to the *leisure* key, whereas "grass" and "meadow" entities belong to the *landuse* key. However, the opposite usage indicates a potentially inappropriate classification. In the classifiers development, we followed the same methodology and used the same data sets as in *LBM*. Table 3 illustrates the classifiers performance measures.

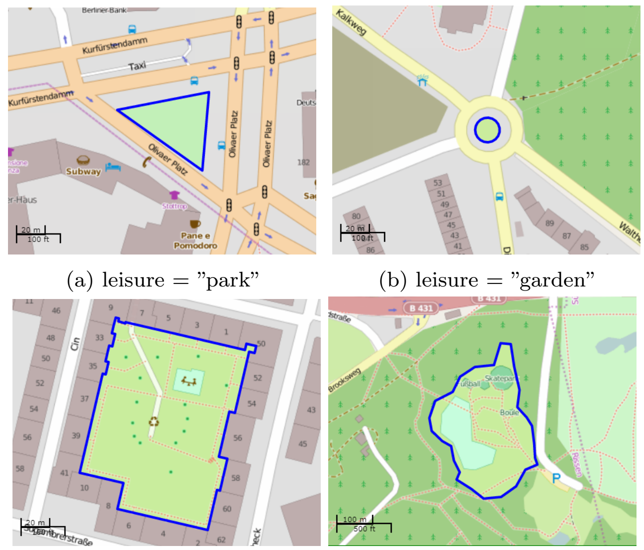
	D		$D_{top\_mappers}$		$D_{top\_versions}$	
	Acc.	AUC	Acc.	AUC	Acc.	AUC
GER	78.4 %	0.85	79.0 %	0.86	73.0 %	0.80
UK	92.2 %	0.97	93.6 %	0.97	81.4 %	0.84

Table 3: TBM classifiers performance of data extracted from Germany (GER) and the UK.

Table 3 conveys that the classifiers of *TBM* have higher performance than the classifiers of *LBM*. According to the table, the classifier based on the data set of Germany has an average performance with accuracy of 76.8% and AUC equal to 0.85, whereas the classifier based on the UK data set has an average performance by 89.0% accuracy and AUC equal 0.92.

6.3 Discussion

In this work, we applied the developed classifiers of *TBM* to check the integrity of the target entities of Germany and the UK. According to the results, the comparison between the classifiers of *LBM* and *TBM* shows that the AUC measures are nearly the same in both models. However, the accuracy measures indicate a higher performance of *TBM* classifiers.



(a) leisure = "park" (b) leisure = "garden"

(c) landuse = "grass" (d) landuse = "meadow"

Figure 7: Samples of entities with potentially inappropriate classification.

Figure 7 shows a sample of detected entities with potentially inappropriate classification. Figures 7a and 7b show entities belonging to the *leisure* tag and classified as "park" and "garden" respectively. The selected examples illustrate that the entities do not show the properties of leisure-related entities. They are relatively small and do not have any kind of infrastructure to be either a "park" nor a "garden". In both cases, the appropriate classification of the entities is most likely "grass". Whereas the entities of Figure 7c and 7d are tagged with *landuse*. They are classified as "grass" and "meadow" respectively. When inspecting the properties of these entities, their current classifications seem to be inappropriate. The entity in Figure 7c is surrounded by houses and contains a playground. The entity in Figure 7d contains a large playground and some entities tagged with *sport*="value". Both of them are relatively large and also have footpaths, i.e., the entities are more likely leisure-related entities. These examples show the validity of the proposed classifiers.

In order to understand which kinds of entities the OSM community consider as problematic, we also downloaded the OSM data concerning the period from December 2nd, 2013 to June 2nd, 2014 (about 6 months). We particularly checked the data for the updated entities, i.e. where the OSM tag (e.g. *leisure* = "*park*") was changed or the entity was completely deleted. We also used the *TBM* classifier to check the integrity of the updated data. Using the updated data of Germany, the classifier identified 23% of 6,568 updated entities to be potentially inappropriate classified.

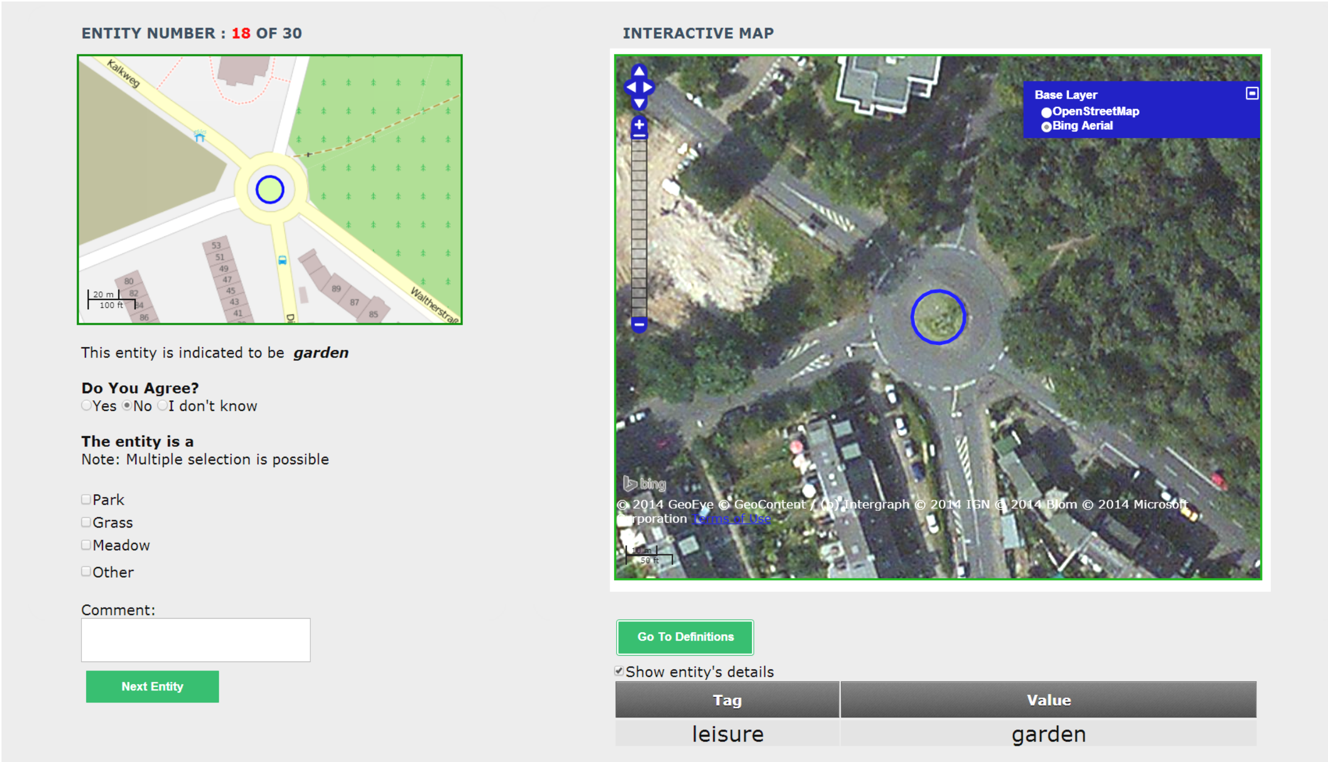


Figure 8: A snapshot from the website of the study.

However, when applied to data of the UK, the classifier identified 60% of 310 updated entities to have potentially inappropriate classifications.

7. EXPERIMENTAL EVALUATION

In order to evaluate our approach, we designed a web-based user study with anonymous participants. The aim of the study was to measure the intra-user agreement of the participants on a set of 30 entities. All entities were detected by LBM and TBM classifiers to have potentially inappropriate classifications.

The study consisted of two phases: *learning* and *evaluation*. In the *learning* phase, we introduced to the participants the OSM recommendations of the four target classes (i.e. tags). Additionally, we displayed them also recommendations of other classes, that are conceptually related. The participants were asked to provide their OSM experience, age, gender, and mother tongue. In the *evaluation* phase we showed all the participants the same set of 30 classified entities; 4 "garden", 6 "grass", 8 "meadow", and 12 "park" entities.

For each entity, the participants were firstly asked about their agreement or disagreement with the current classification. In case of disagreement, the participants were allowed to select from different options to classify the entity. Figure 8 depicts a snapshot from the study website. The left side displays the investigated entity and the opinion of the participant. At the right side the participant was allowed to check the entity's context via an aerial image or on OSM maps. Participants were also allowed to check the recommendations of classes at any point of the study, and also to check other tags used to describe the given entity.

In total we had 157 participants to the experiment. Out of these 115 participants finished the study. Together 81 participants gave complete assessments of all entities (it was possible to skip entities), and thus we considered this group for the analysis. Together there were 65 males and 16 females. 24 of the participants had no knowledge about OSM, 17 were beginners, 21 had moderate knowledge, and 19 considered themselves as experts. The average age of the participants was 27 years and they had more than 10 different mother languages.

In order to evaluate the results, we used Light's Kappa for m raters [20] to measure the intra-user agreement of the participants. Kappa value of 1.0 means maximum agreement and the values ≤ 0 mean less than chance agreement. Moreover, the range from 0.01 to 1.0 is divided into slight, fair, moderate, and substantial agreements [31].

Light's Kappa for all 81 participants was 0.176, meaning thus *a slight agreement*. We analysed the intra-user agreements also per subgroups. To create the subgroups we considered different levels of expertise about OSM project by participants (no knowledge, beginner, and expert). Participants with expert knowledge about OSM had somewhat higher intra-user agreement — 0.21 (*fair agreement*) — than participants with limited or no knowledge — 0.19 and 0.15 (*slight agreement*), accordingly. We also grouped the intra-user agreements data to entity types (garden, park, meadow, grass). This provided not much difference, except for somewhat higher intra-user agreement (0.26) concerning "meadow" entities and accordingly lower concerning "park" entities (0.09).

We also analyzed the experiment results by investigating entities individually. For each entity, we counted the different opinions and checked the agreement or disagreement

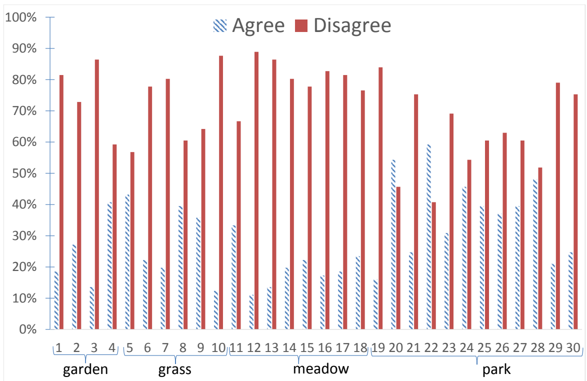


Figure 9: The percentage of total agreement and disagreement of the participants on the current classifications per entity.

of the 81 participants about the current classifications of entities. Figure 9 shows the results as percentages of the participants' agreement and disagreement per entity. This reveals that the participants had in a substantial amount of cases a disagreement with the current classifications of entities. However, there are small differences: "park" for instance was found in more cases an acceptable categorization than, say, "meadow".

7.1 Discussion

These findings clearly show that the participants of the study substantially disagreed with the current classification of the entities. This is a strong support for the classifiers we developed and for the method in general. This means that, we were able to identify controversial entities within the OSM data set by a combination of analytically and statistically derived properties (see Section 5.1 for details). However, the participants also largely disagreed among themselves even when they are supported by materials like maps and class descriptions. Participants also gave comments such as *"Needs further investigation/survey", "not sure" and "difficult to see"*, which all suggesting to further study classification mechanisms of VGI projects. Especially the remote annotation of satellite imagery by contributors not familiar with a region can be problematic: if an entity is not clearly recognizable on the image and the contributor is not fully aware of the recommendations — the resulting data might not be of sufficient quality. One way of avoiding this is the explicit integration of local contributors in the validation process. In OSM this is a common practice, however, coupling the results of automatic approaches as proposed in this paper with local contributors requires new communication infrastructures and modalities within VGI projects.

8. CONCLUSIONS

In this work, we presented a novel approach to address a facet of data quality in Volunteered Geographic Information (VGI): classification ambiguity and plausibility. In many cases geographic features can belong to multiple classes, depending on the motivation, viewpoint, or conceptualization of the individual contributor. However, in many cases the classification is just not correct and needs to be fixed. We developed an approach based on machine learning from VGI data itself, thus without the need for reference data. In this work, "park", "garden", "grass", and "meadow" entities are selected reflecting the ambiguous classification of entities. We tackle the classification ambiguity problem by learning properties and characteristics of representative entities within the dataset. We utilize geometric and contextual geographic properties to build classifiers based on a carefully selected subset of the OSM dataset.

The developed classifier was able to detect obviously inappropriate classified entities. To validate the classifier beyond computational measures, we conducted a user study. In this study, our participants were asked to revise the classification of 30 detected entities. If they disagreed with the current tagging (e.g. "park") they had a chance to propose another tagging (e.g. "garden"). The result of our study showed that the participants disagreed with the actual classification but also disagreed amongst themselves. This result is a strong indicator for the feasibility of our classifiers: they detect controversial entities, which is the original purpose of our approach. The output of the classifiers can be presented to volunteers and validated by their knowledge.

However, the generation of classifiers is still a rather manual task: one has to identify a set of potentially ambiguous entities, and define their discriminating properties in form of classification rules. In our future work we will focus on the automatic detection of ambiguous classes and the characteristic properties.

9. ACKNOWLEDGMENTS

We gratefully acknowledge supports provided by the German Academic Exchange Service (DAAD), the German Research Foundation (DFG) via the Transregional Collaborative Research Center Spatial Cognition SFB/TR8, as well as the European Commission Initial Training Network GeoCrowd. We thank the European Union via grant agreement FP7-PEOPLE-2011-IRSES 295269. We also want to thank all participants of our study for donating their valuable time.

10. REFERENCES

[1] A. L. Ali and F. Schmid. Data quality assurance for Volunteered Geographic Information. In *Proceedings of the eighth International Conference on Geographic Information Science*, pages 126–141. Springer International Publishing Switzerland, 2014.

[2] C. Barron, P. Neis, and A. Zipf. A comprehensive framework for intrinsic OpenStreetMap quality analysis. *Transactions in GIS*, 18, 2014.

[3] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[4] T. Cover and P. Hart. Nearest Neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

[5] F. D'Antonio, P. Fogliaroni, and T. Kauppinen. VGI edit history reveals data trustworthiness and user reputation. In *Proceedings of the 17th AGILE Conference on Geographic Information Science,*

Connecting a Digital Europe through Location and Place, Castellon, Spain, June 2014.

[6] R. Devillers, A. Stein, Y. Bédard, N. Chrisman, P. Fisher, and W. Shi. Thirty years of research on spatial data quality: achievements, failures, and opportunities. *Transactions in GIS*, 14(4):387–400, 2010.

[7] M. J. Egenhofer. On the equivalence of topological relations. *International Journal of Geographical Information Systems*, 9:133–152, 1995.

[8] S. Elwood, M. F. Goodchild, and D. Z. Sui. Researching Volunteered Geographic Information: Spatial data, geographic research, and new social practice. *Annals of the Association of American Geographers*, 102(3):571–590, 2012.

[9] T. Fawcett. An introduction to ROC analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[10] A. J. Flanagin and M. J. Metzger. The credibility of Volunteered Geographic Information. *GeoJournal*, 72(3-4):137–148, 2008.

[11] J.-F. Girres and G. Touya. Quality assessment of the french OpenStreetMap dataset. *Transactions in GIS*, 14(4):435–459, 2010.

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

[13] M. F. Goodchild and L. Li. Assuring the quality of Volunteered Geographic Information. *Spatial statistics*, 1:110–120, 2012.

[14] J. Grira, Y. Bédard, and S. Roche. Spatial data uncertainty in the VGI world: Going from consumer to producer. *Geomatica*, 64(1):61–72, 2010.

[15] M. Haklay. How good is Volunteered Geographic Information? a comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and planning. B, Planning & design*, 37(4):682, 2010.

[16] M. Haklay and P. Weber. OpenStreetMap: User-generated street maps. *Pervasive Computing, IEEE*, 7(4):12–18, 2008.

[17] J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2011.

[18] C. Keßler and R. T. A. de Groot. Trust as a proxy measure for the quality of Volunteered Geographic Information in the case of OpenStreetMap. In *Geographic Information Science at the Heart of Europe*, pages 21–37. Springer, 2013.

[19] C. Keßler, J. Trame, and T. Kauppinen. Tracking editing processes in Volunteered Geographic Information: The case of OpenStreetMap. In *Identifying objects, processes and events in spatio-temporally distributed data (IOPE), workshop at conference on spatial information theory*, volume 12, 2011.

[20] R. Light. Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin*, 76:365–377, 1971.

[21] I. Ludwig, A. Voss, and M. Krause-Traudes. A comparison of the street networks of Navteq and OSM in Germany. In *Advancing Geoinformation Science for a Changing World*, pages 65–84. Springer, 2011.

[22] P. Mooney and P. Corcoran. Can Volunteered Geographic Information be a participant in eEnvironment and SDI? In *Environmental Software Systems. Frameworks of eEnvironment*, pages 115–122. Springer, 2011.

[23] P. Mooney and P. Corcoran. The annotation process in OpenStreetMap. *Transactions in GIS*, 16(4):561–579, 2012.

[24] C. Mülligann, K. Janowicz, M. Ye, and W.-C. Lee. Analyzing the spatial-semantic interaction of points of interest in Volunteered Geographic Information. In *Spatial information theory*, pages 350–370. Springer, 2011.

[25] P. Neis, D. Zielstra, and A. Zipf. Comparison of Volunteered Geographic Information data contributions and community development for selected world regions. *Future Internet*, 5(2):282–300, 2013.

[26] P. Neis and A. Zipf. Analyzing the contributor activity of a Volunteered Geographic Information project: The case of OpenStreetMap. *ISPRS International Journal of Geo-Information*, 1(2):146–165, 2012.

[27] F. Schmid, L. Frommberger, C. Cai, and F. Dylla. Lowering the barrier: How the What-You-See-Is-What-You-Map paradigm enables people to contribute volunteered geographic information. In *Proceedings of the 4th Annual Symposium on Computing for Development*, page 8. ACM, 2013.

[28] F. Schmid, L. Frommberger, C. Cai, and C. Freksa. What you see is what you map: Geometry-preserving micro-mapping for smaller geographic objects with mapit. In *Geographic Information Science at the Heart of Europe*, pages 3–19. Springer International Publishing, 2013.

[29] W. R. Tobler. A computer movie simulating urban growth in the detroit region. *Economic geography*, 46:234–240, 1970.

[30] A. Vandecasteele and R. Devillers. Improving volunteered geographic data quality using semantic similarity measurements. *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 1(1):143–148, 2013.

[31] A. J. Viera, J. M. Garrett, et al. Understanding interobserver agreement: the kappa statistic. *Fam Med*, 37(5):360–363, 2005.

[32] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, 2 edition, 2005.