Data Quality Assurance for Volunteered Geographic Information

Ahmed Loai Ali[1,3] and Falko Schmid[1,2]

[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

Abstract. The availability of technology and tools enables the public to participate in the collection, contribution, editing, and usage of geographic information, a domain previously reserved for mapping agencies or companies. The data of Volunteered Geographic Information (VGI) systems, such as OpenStreetMap (OSM), is based on the availability of technology and participation of individuals. However, this combination also implies quality issues related to the data: some of the contributed entities can be assigned to wrong or implausible classes, due to individual interpretation of the submitted data, or due to misunderstanding about available classes. In this paper we propose two methods to check the integrity of VGI data with respect to hierarchical consistency and classification plausibility. These methods are based on constraint checking and machine learning methods. They can be used to check the validity of data during contribution or at a later stage for collaborative manual or automatic data correction.

1 Introduction

During the last decade, low-cost sensing devices like handheld GPS receivers or smartphones became available and accessible for many consumers. In the same period powerful open GIS software and web technologies have been developed. The availability of technology and tools enables the public to participate in the collection, contribution, editing, and usage of geographic information, a domain previously reserved for mapping agencies or large organizations. Volunteered Geographic Information (VGI) [1], the voluntary collection and contribution of geo-spatial data by interested individuals became a large and vital movement. VGI projects like OpenStreetMap[1] (OSM) result in large scale data sets of geographic data covering many parts of the world. This new way of geographic data production changed not only the way of data processing but also applications and services built on it [2–4].

There exist a huge number of services based on e.g., OSM data, such as map providers, trip advisers, navigation applications. Depending on the service, reliable data is necessary. However, without coordinated action, the experience and training of experts, and industrial grade sensing devices it is hard to guarantee data of homogeneous quality.

The absence of a clear classification system in, e.g., OSM, the ambiguous nature of spatial entities, and the large number of users with diverse motivations and backgrounds

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

M. Duckham et al. (Eds.): GIScience 2014, LNCS 8728, pp. 126–141, 2014.

foster the generation of data of mixed quality. Whatever a body of water is a pond or a lake, whatever a grassland is a meadow, natural reserve, a park, or a garden is not just a question of a proper, crisp definition, but also a question of perception, conceptualization, and cultural background. What is a pond somewhere, might be a lake in a different environment, a river might be a creek or a stream. In addition to rather conceptual issues, many contributed entities are incompletely classified or wrongly attributed due to the open and loose attributation mechanism in OSM. As a result, a significant amount of data is not correctly classified and can cause errors whenever they are addressed by algorithms, such as rendering, analysis, or routing. This situation triggers questions about the quality of VGI data, suitable mechanisms for guaranteeing and fostering high quality contributions, and correcting problematic data.

Hence, it becomes increasingly important to analyze the heterogeneous quality of VGI data. Several studies investigate the quality of VGI by applying geographic data quality measures, such as feature completeness, positional accuracy, and attribute consistency [5–7]. These approaches usually require using reference data sets to evaluate the VGI data. However, these data sets are in many cases not available.

In this paper we present two approaches for analyzing the quality of VGI data: one by constraint checking and one by machine learning, i.e., we are analyzing the available data only with respect to consistency and plausibility based on contributions themselves. The results can be used to re-classify existing data and to provide guidance and recommendations for contributors during the contribution process. Recommendations can be directly generated from the data source itself by analyzing the distribution of the contributed feature in the surrounding area, thus the locality of entitles is preserved and no global rules are applied to locally generated data.

2 Related Work

In VGI, contributors produce geographic information without necessarily being educated surveyors or cartographers. In open platforms such as OSM, 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. Hence, the assessment of VGI data quality became a focus in VGI related research.

Quality of VGI data has various perspectives and notions: completeness, positional accuracy, attribute consistency, logical consistency, and lineage [8]. The quality can be assessed by basically three different methods: comparison with respect to reference data, semantic analysis, and intrinsic data analysis.

One approach to assess the quality of VGI data is by means of a direct comparison with reference data collected with a certain quality standards. The challenge of this approach is to identify a robust mutual mapping function between the entities of both data sets. In [6, 9] 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 relative to others. They also show a tight relation between completeness and urbanization [9, 10].

Different aspects have influence on 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 the heterogeneous quality of VGI data [11]. Amongst others, semantic inconsistency is one of the essential problems of VGI data quality [12]. In [13] and [14] the authors present methods for improving the semantic consistency of VGI. The analysis of semantic similarity is applied to enhance the quality of VGI by suggesting tags and detecting outliers in existing data [13, 14], as well as by ontological reasoning about the contributed information (e.g., [15]). Another approach for tackling quality issues is the development of appropriate interfaces for the data generation and submission. In [16, 17] the authors demonstrate that task-specific interfaces support the generation of high quality data even under difficult conditions.

An alternative approach is evaluating the available data along three intrinsic dimensions [8]:

- *Crowdsourcing evaluation*: the quality of data can be evaluated manually by means of cooperative crowdsourcing techniques. In such an approach, the quality is ensured through checking and editing of objects by multiple contributors, e.g., by joint data cleaning with gamification methods [18].
- *Social measures*: this approach focuses on the assessment of the contributors themselves as a proxy measure to the quality of their contributions. [6, 9] use the number of contributors as a measure for data quality, [19] analyzes the individual activity, [11] investigates positive and negative edits, [20] is researching fitness-for-purpose of the contributed data.
- *Geographic context*: this approach is based on analyzing the geographic context of contributed entities. This approach relates to Tobler's first law of geography which states that "all things are related, but nearby things are more related than distant things" [21].

3 Managing Quality of VGI Data

A big challenge for VGI is the quality management of the contributed data because of its multidimensional heterogeneity (knowledge and education, motivation for contribution, and technical equipment). The problem requires the development of tools advising contributors during the entity creation process, but also to correct already existing data of questionable quality. Amongst others, quality problems can be general accuracy issues, geometric or topological constraint violations, hierarchical inconsistencies, and wrong or incomplete classification. In this work we focus on hierarchical inconsistencies and wrong or incomplete classification. Whenever we use the term "wrong" in our study we mean the assignment of a *potentially* wrong class or *tag* to the respective entity due to labeling ambiguity. "Wrong" entities will be detected by our classification and consistency checking algorithms. This is only an indicator for a potential conflict.

In the case of OSM, it is known that the data set contains large amounts of problematic data (e.g., see Section 2). On the other hand, we can assume that a significantly larger part of the data is of sufficient quality: the large amount of volunteers constantly improving the data set and the large number of commercial applications built on top of the data set are good indicators for it. Given that this rather unprovable statement is true, we can use the data itself for quality assessment by learning its properties and using the results as an input for the processes described in our approach.

Figure 1 describes the two phase approach: in the *Classification* phase, we can either apply machine learning algorithms to learn classifiers of the so far contributed data, or we can define classification constraints the data has to satisfy. Some of the before mentioned quality issues could be solved if at the point of data generation or contribution the integrity with existing data is checked. Depending on the potential problem to be addressed, different automatic approaches for satisfying inherent constraints are available, e.g., [22].

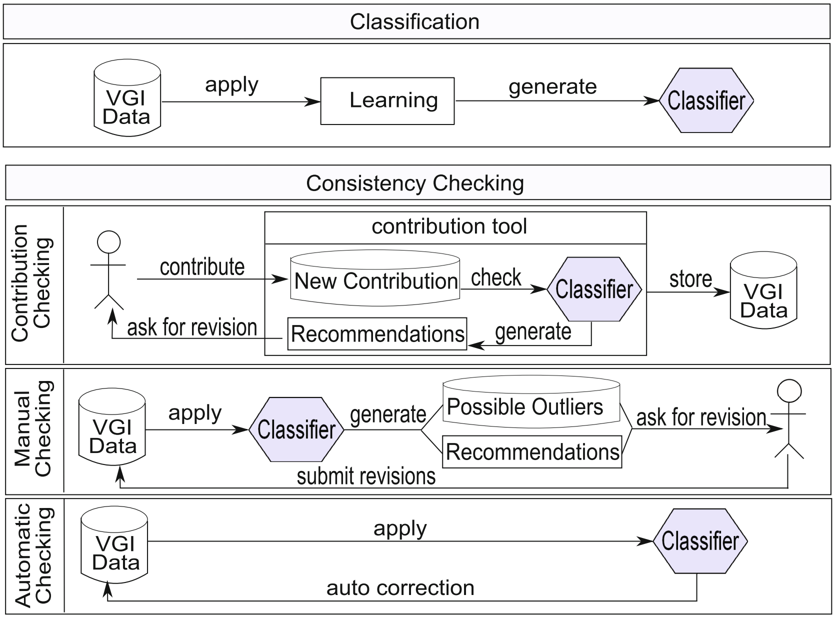


Fig. 1. Proposed approaches to ensure VGI quality, see Section 3 for a detailed description

Hence, in the *Consistency Checking* phase we propose three approaches for checking the consistency of the data: during *Contribution Checking* the contribution tool should inform users during the contribution process about potentially problematic data based on the generated classifier. Contributors can now consider the hints generated by the system about an object and can take actions to correct it if necessary. After contribution, the new data can be used to train the classifier again (if checking is based on an learning approach). *Manual Checking* should provide tools allowing the identification of problematic entities in the existing data set. They can be presented to volunteers for checking and correcting, ideally based on plausible suggestions. And finally, *Automatic Checking* can correct obviously wrong data automatically, if the correction can be computed without human assistance.

4 Tackling Areal Consistency and Classification Plausibility

The majority of data quality studies focus on point-like or linear geographic entities, such as points of interest or road networks (see Section 2). In this work we focus on quality issues related to areal entities, that is extended geometric entities. Our methods can be applied to entities of all possible scales, from very large administrative or natural entities to rather small ones like buildings or park benches.

The focus of our work is the quality of the *classification* of the contributed data. We are particularly interested in:

- *Hierarchical consistency* of administrative data: we check if administrative elements are used according to intrinsic, logical rules.
- *Classification plausibility* of areal entities: the correct classification of entities can be difficult, especially when contributors are not aware of potential conflicts due to similar concepts. Here we focus on ambiguity issues resulting from the availability of two or more possible classification options of entities (e.g., park vs. garden vs. grass).

Our study is build on OSM data. We will use notions typically used in the OSM tagging scheme, such as: *keys* and *values*.

5 Hierarchical Consistency Analysis

Administrative boundaries are political geographic entities with a strict inherent structure, such as *continents* consist of *countries*, *countries* consisting of *states* and *states* consisting of *districts*, etc. In OSM[2] administrative boundaries are defined as *subdivisions of areas/territories/jurisdictions recognized by governments or other organizations for administrative purposes. Administrative boundaries range from large groups of nation states right down to small administrative districts and suburbs, with an indication of this size/level of importance, given by tag "admin_level" which takes a value from 1 to 10*. However, as countries can have different administrative partitioning, some levels might not be applicable or the classification schema may not be sufficient. In this case it can be extended to 11 levels (e.g., in Germany and Netherlands).

Typically, administrative boundaries around administrative Units U are structured such that every administrative unit typically belongs to *one* administrative level of 1 to 11 (exceptions are, e.g., city states):

$$\forall u \in U_i \; where \; 1 \leq i \leq 11 \tag{1}$$

Each administrative unit where $i > 1$ is contained in an administrative unit of a higher level; all together the contained units *exhaustively* cover the territory of the containing unit:

$$\forall u_a \in U_{i>1}, \exists u_b \in U_{j>i} : u_a \subset u_b \tag{2}$$

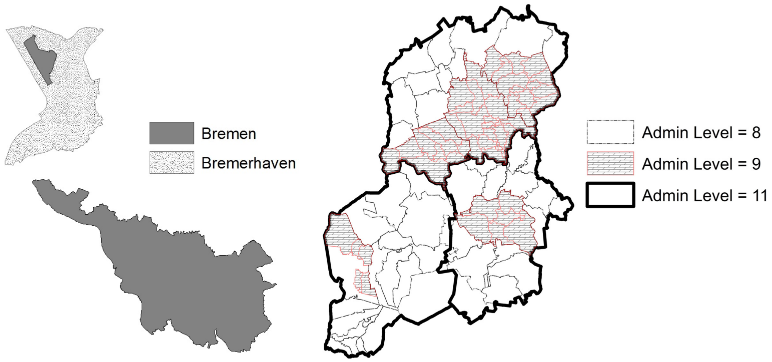
Administrative units on one level can share borders but do *not intersect* each other:

$$\forall U_j, U_k \subset U_i : U_j \cap U_k = \emptyset \tag{3}$$

[2] http://wiki.openstreetmap.org/wiki/Key:admin_level#admin_level

However, there are exceptions from this strict hierarchy, such as exclaves, enclaves, city states, or embassies. Still, the vast majority of administrative units follow a clear and exhaustive hierarchical ordering. This allows checking the integrity of the available administrative data in OSM by checking the following type of outliers:

- *Duplication*: in the case of duplication, entities belong to two or more different administrative units. See Figure 2(a).
- *Inconsistency*: hierarchical inconsistency occurs when entities of higher administrative units are contained in units of lower levels or the same level. See Figure 2(b)
- *Incorrect Values*: incorrect values occur throughout the OSM data set, probably due to the import from different classification schemes. Typically the value of *admin_level* tag is not a numerical value between 1–11.



(a) Duplication (b) Inconsistency

Fig. 2. Incorrect classification plausibility (Duplication & Inconsistency). In a) a part of Bremen city is within Bremerhaven, in b) units on level 11 contain elements of level 8 and 9.

5.1 Consistency Analysis Results and Discussion

We applied the consistency rules on the complete OSM data set downloaded at January 20th, 2014. At the time of analysis, the OSM data contained 259,667 geographic entities classified as administrative units (admin_level = *value*). 24,410 entities, thus about 10% of all administrative units contained problematic assignments, see Figure 3. We identified 14,842 duplications, 9,305 inconsistencies and 263 incorrect values.

Figure 2(a) illustrates an example for *duplication*: a part of the administrative unit representing Bremen city, is part of another unit representing Bremerhaven city. Figure 2(b) shows an instance of inconsistency: some administrative units of level 8 and 9 are contained by administrative units of level 11.

Of course, not all of the 24,410 detections represent wrong data, some cases already represent the mentioned special cases, some inconsistencies might be detected due to incomplete presence of administrative hierarchies. However, a plausibility check as sketched in Section 3 would draw the attention of the contributor towards potential errors.

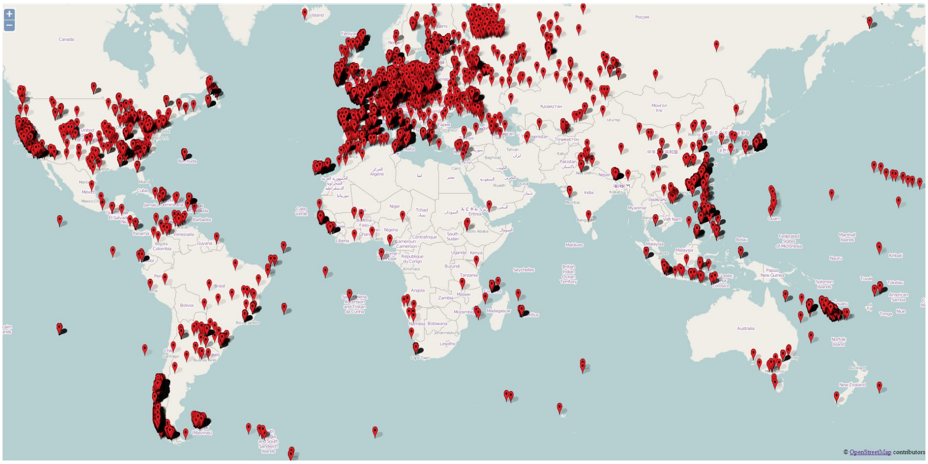


Fig. 3. Distribution of potentially incorrect hierarchical classification of administrative units

6 Classification Plausibility Analysis

When users contribute data to OSM, they have a large range of possibilities to classify the data. In some cases classifying entities is not straightforward; depending on the perspective of the contributor different possible classes may be applicable. A water body can still be a pond or already be a lake, the grass covered area can be a park, a garden, meadow or grassland. In many cases there is no definite answer, especially as in OSM there is no explicit classification system, just recommendations. However, utilizing spatial data requires homogeneous handling of data of identical concepts. Only if the same type of entities are identically classified, algorithms can access them properly for analysis, rendering, or reasoning. However, in many cases users contribute data with wrong classifications either due to conceptual ambiguity or due to a different understanding of the available concepts.

In this work we exemplify our approach on analyzing classification plausibility of entities, which are classified either as *park* or *garden*. We chose these classes as they are good examples for classification ambiguity: within OSM, parks and gardens lack a clear definition distinguishing them. Thus, contributions of these features mainly depend on individual conceptualizations. Many entities are obviously not correctly classified when we inspected them with a commonsense understanding of parks and gardens. Typically parks are public, accessible areas of a cultivated nature. Gardens, in contrast are typically private areas also featured with cultivated nature. However, one large difference of both entities is not only their infrastructural containments, but also their size: parks are usually significantly larger than gardens. As usual when it comes to geospatial reality, we can observe everything such as large public gardens or small parks. However, the vast majority of gardens and parks follow this vague classification (see Figure 6 for a support of this statement), especially relative to entities in their surrounding (parks and gardens can have significantly different dimensions in different areas of the world, usually correlated to the available territory in relation to the population). In the following we analyzed entities classified with the tags leisure=*park* and leisure=*garden*.

6.1 Classification Learning to Ensure VGI Quality

Due to the large amount of data in OSM, it is possible to apply machine learning techniques to tackle data quality issues. Machine learning algorithms can learn from existing data and extract implicit knowledge to build a classifier. Then such a classifier can be used for ensuring the quality as sketched in Figure 1, either during contribution or by applying on already existing data. In our approach learning the classifier on the contributed data is used to predict the correct class of an entity (i.e. park or garden in our example). This is done in two steps: a learning or training step, and a validation step.

In the first step our system learns a classifier based on the properties of pre-classified entities of a *training set* [23, 24]. In this work, the training set consists of entities representing parks and gardens, $D_{train} = (E_1, E_2, ..., E_n)$, where each Entity E is represented by a set of features (such as: size, location ...etc.) and is assigned to a class C (i.e. park or garden), $E = (F_1, F_2, ..., C)$. This step tries to identify a function, $f(E) = C$ to predict the class C of a given entity E.

In the second step the generated classifier is used for classification: we apply it on a test set to measure the accuracy of the classifier. The test set only contains entities not used for training. The classifier performance is evaluated according to classification accuracy on the test entities [23, 24].

6.2 Experiments and Setup

As described previously, we focus on classification plausibility in case of similarly applicable classes, in our case parks (leisure = *park*) and gardens (leisure = *garden*). We use data from Germany, the United Kingdom (UK), and Austria. According to [6, 9], OSM data is of acceptable quality in Germany and the UK. In our study we use data downloaded on December 20th, 2013.

We selected data from the ten densest (population/area) cities of each country. Figure 4 shows the selected cities and the present number of parks and gardens within each city. We decided to use cities as spatial units, as they define graspable spatial regions. In our experiments we follow the locality assumption of Tobler's first law of geography: different cities in the same country might have a closer understanding of parks and gardens than cities of different countries. Thus, it will be more likely to produce meaningful results if we apply a learned classifier from one city on the data of another city in the same country. Learning areal properties in Hong Kong and applying them on data of Perth/Australia might not be valid due to the size of the available territory. The same holds for the idea of learning *global* parameters for parks and gardens — spatial entities have a strong grounding in local culture and history of a particular country, applying global rules on local data will lead in many cases to wrong classifications due to different local concepts.

In the following we learned the classifiers of 10 cities per country, and applied them mutually to every other city. By assessing the classification accuracy, this method

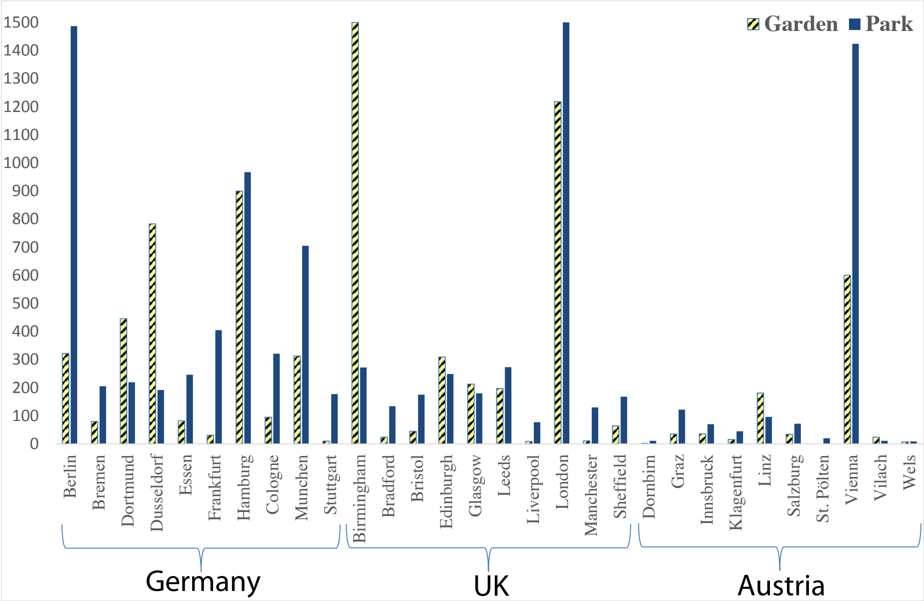
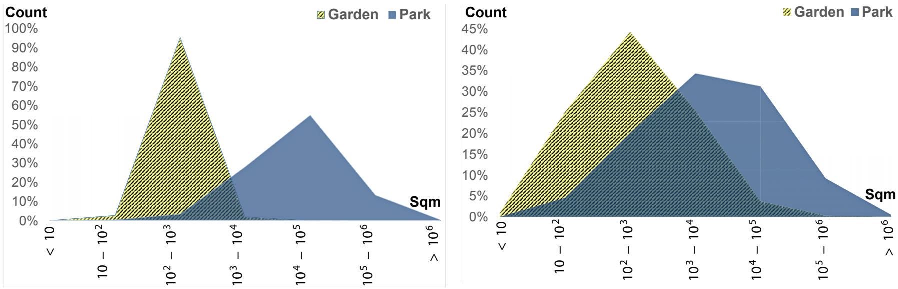


Fig. 4. Number of Parks and Gardens within the selected data set

allows identifying the most accurate classifiers for a city, and the identification of biased classifiers due to biased or ambiguous classification practices within specific cities.

In our study we applied a straightforward approach to distinguish between parks and gardens: we compared their size. Size is not probably enough to reliably distinguish between gardens and parks, especially if we consider other related classes such as meadows or grassland. When we have a closer look into how the classes are populated, we can see that the distribution can be rather clear, as it is, e.g., the case in Birmingham (see Figure 5(a)). There are also places with a less clear separation, e.g., the case of London (see Figure 5(b)), where parks and gardens seem to have a large conceptual overlap. However, our intention behind choosing the area is to detect incorrect classification at a very early point of contribution, when no other features are yet provided. Confronted with an "early-warning," users can reconsider the class they selected and modify it if required. However, especially a review of the existing data, as suggested in Section 3, can be fed by such a classifier. Figure 6 shows the mean areas of parks and gardens. It clearly shows that the areas per class are generally distinct and can be used to distinguish between entities of the two classes.

Feature Selection. The areas of each class have a specific distribution in each city. Figure 6 shows that parks are more likely to be large (i.e., tens of thousands to millions sqm), while gardens are more likely to cover rather smaller areas (i.e., a few sqm to a few thousands sqm). Although there are rare cases (i.e. Royal Botanic Gardens in the UK about one million sqm, however, they can be considered to be parks) corrupting



(a) Birmingham (b) London

Fig. 5. Distribution of parks and gardens areas in London and Birmingham

the distribution; the majority of entities follow a common distribution. This distribution might also be similar in other cities, even if the data does not reflect it. By learning these distributions, we can distinguish between parks and gardens, and apply the learned classifiers to other cities and check the existing data or to guide contributors during the contribution process.

Classifier Training. Building a classifier basically can be done 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. In LL, generalization beyond the training data is delayed until a query is made to the system. K-nearest neighbors (KNN) and case based reasoning (CBR) are examples of lazy learning [23, 24]. In OSM a set of pre-classified entities is already stored, and the classification process is performed on new entities at contribution time. The new entity is classified based on similarity to existing entities. Hence, it is a good idea to follow the lazy learning paradigm to develop a classifier.

We decided to use KNN [25, 26] for building a classifier. KNN classifies entities based on closest training examples. It works as follows: the unclassified entity is classified by checking the K nearest classified neighbors. The similarity between the unclassified entity and the training set is calculated by a similarity measure, such as Euclidean distance.

Classifier Validation. During the validation process we use independent data sets for training and testing or we use the same data set for mutually applied classifiers (with this method, we evaluate if a classifier from a different city can be applied to another city). In the latter case, we use *K-fold cross validation* (CV) [27] to show the validity of our classification. In CV a training set is divided into K disjointed equal sets, where each set has roughly the same class distribution. Then the classifier is trained K times[3], and each time a different set is used as a test set. Afterwards the performance of the

[3] 5 and 10 are recommended values for K.

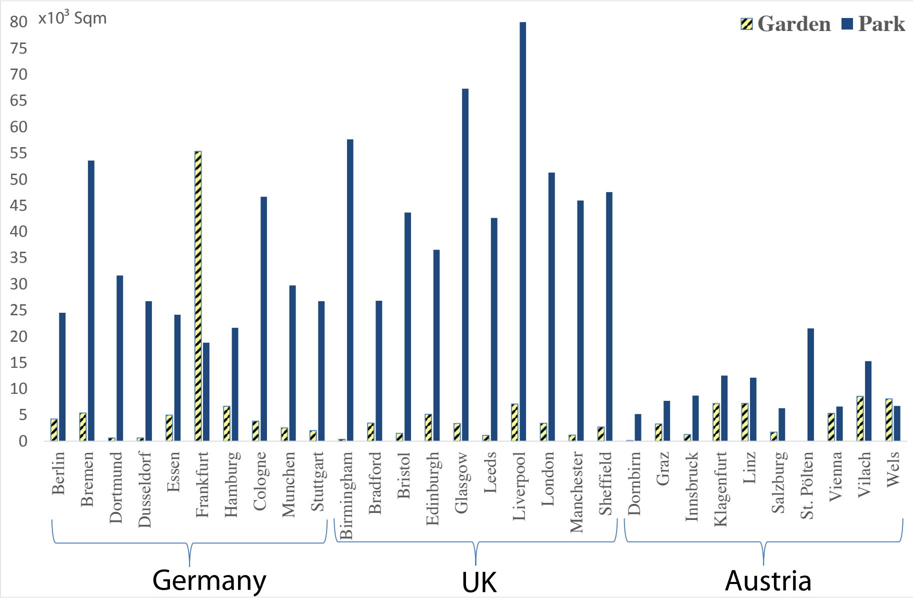


Fig. 6. Mean area size of parks and gardens for the selected data set

classifier is measured as the average of developed classifiers [27]. We build classifiers for each city in a country. The results can be inspected in Tables 1, 2 and 3. The rows of the tables represent the accuracies of different classifiers for the data of each city as a test set. These classifiers were generated based on the data of other cities as training sets and are represented in the columns. The last column "Class. Acc." shows the average classification accuracy of parks and gardens within each city based on the top three classifiers (italic red values).

Classifier Assessment. Depending on just one training and test set might result in biased classifiers. Furthermore, we aim to detect possible incorrect classifications based on the similarity between cities within the same country. Thus, we build mutual classifiers between cities at the same country. One challenge is to assess the classifier performance. The accuracy of a classifier applied on a given test set is expressed by the percentage of correctly classified entities (please see the next section for a deeper discussion on the measurability of the results). However, in some cases accuracies are biased due to overfitting or underfitting [23, 24]. A reason can be unbalanced population of the training or the test set. This happens for instance when the classifiers created from Liverpool or Manchester are applied on the Birmingham data (see Table 2). The Receiver Operation Characteristics (ROC) curve is a useful measure to asses the performance of classifiers. The ROC curve represents the relative trade-off between benefits and costs of the classifier. In particular the Area Under the ROC Curve (AUC) is a useful measure to asses a classifier. The closer the value of a AUC is to 1, the higher its

Table 1. Classification accuracy for parks and gardens of cities in Germany

	Training Set										
Test Set	Berlin	Bremen	Dortmund	Dusseldorf	Essen	Frankfurt	Hamburg	Cologne	Munchen	Stuttgart	**Class. Acc.**
Berlin	80.43	76.78	76.23	***72.25***	***74.07***	**82.03**	56.44	***79.38***	78.94	82.2	75.23
Bremen	***71.93***	72.28	70.18	***70.18***	69.12	72.28	59.30	***72.98***	71.23	71.93	71.70
Dortmund	54.14	55.79	***83.31***	82.26	***82.41***	**32.93**	76.84	***81.05***	76.84	32.93	82.26
Dusseldorf	43.59	59.08	85.74	***91.38***	***91.18***	**19.69**	86.36	***87.28***	78.26	19.69	89.95
Essen	77.44	71.95	***79.27***	79.88	***82.32***	**75.00**	66.16	***80.49***	78.35	75.00	80.69
Frankfurt	**89.68**	79.13	**75.00**	**62.39**	**65.37**	***92.66***	**47.94**	***78.67***	**78.21**	***92.89***	88.07
Hamburg	54.15	55.87	59.03	***61.27***	***61.76***	**51.69**	***61.06***	58.97	57.90	51.79	61.36
Cologne	78.13	79.09	81.49	***80.05***	***80.05***	**77.16**	66.35	80.53	***80.29***	77.16	80.13
Munchen	72.50	71.02	79.37	***77.90***	***79.17***	**69.16**	62.48	78.49	***78.88***	69.25	78.65
Stuttgart	***93.58***	74.33	***80.75***	65.24	67.38	94.65	***54.01***	74.33	**78.61**	94.65	76.11

Table 2. Classification accuracy for parks and gardens of cities in the UK

	Training Set										
Test Set	Birmingham	Bradford	Bristol	Edinburgh	Glasgow	Leeds	Liverpool	London	Manchester	Sheffield	**Class. Acc.**
Birmingham	99.73	0.99	70.03	92.65	***90.79***	***92.67***	**0.94**	69.27	**1.29**	***94.73***	92.73
Bradford	59.49	84.81	***73.42***	54.43	67.09	***70.25***	84.81	***74.68***	81.65	68.99	72.78
Bristol	***72.73***	79.55	78.64	67.27	75.91	79.09	***79.55***	76.82	79.55	***81.82***	78.03
Edinburgh	65.23	44.44	***59.14***	59.32	***63.26***	63.26	44.62	***59.50***	51.61	60.75	60.63
Glasgow	***74.30***	45.55	67.18	70.23	69.72	***73.03***	***45.80***	67.94	61.07	69.97	71.76
Leeds	***75.96***	57.87	72.34	70.43	***77.45***	75.96	58.09	73.40	58.94	***77.66***	77.02
Liverpool	***86.05***	89.53	88.37	80.23	***87.21***	***89.53***	89.53	87.21	89.53	90.70	87.60
London	68.26	64.88	72.51	66.77	***72.02***	72.22	65.05	***73.03***	68.12	***72.83***	72.63
Manchester	***67.38***	92.20	80.85	63.83	***73.05***	78.01	92.20	79.43	91.49	***79.43***	73.29
Sheffield	***71.55***	72.41	78.88	70.26	74.14	***77.59***	72.41	73.71	73.71	***78.02***	75.72

performance. Good classifiers should have AUC value between 0.5 and 1 [28]. Tables 1, 2, and 3 represent the accuracies of the generated classifiers, while AUC measures are dropped due to space restrictions. A combination of accuracy and AUC is used to determine the classification accuracy of parks and gardens for each city. We select the three top classifiers with the highest AUC measures (italic red values), and neglect biased classifiers with AUC less than or equal 0.5 (blue values). The classification accuracy is measured on the basis of the average accuracy.

Table 3. Classification accuracy for parks and gardens of cities in Austria

	Training Set										
Test Set	Dornbirn	Graz	Innsbruck	Klagenfurt	Linz	Salzburg	St. Pölten	Vienna	Vilach	Wels	**Class. Acc.**
Dornbirn	**100**	**84.62**	***84.62***	***84.62***	23.08	**53.85**	**84.62**	***76.92***	15.38	**76.92**	82.05
Graz	**63.06**	**77.71**	**64.33**	**77.71**	**31.85**	***68.15***	**77.71**	**74.52**	***35.03***	**60.51**	51.59
Innsbruck	80.19	**66.04**	***83.02***	***66.04***	***52.83***	**50.94**	**66.04**	66.98	47.17	**47.17**	67.30
Klagenfurt	**72.13**	***73.77***	70.49	**70.49**	**31.15**	**62.30**	**73.77**	***75.41***	***47.54***	**49.18**	65.57
Linz	41.52	**34.66**	***43.32***	34.66	***62.09***	**37.91**	**34.66**	***38.63***	61.01	**40.07**	48.01
Salzburg	56.60	***67.92***	**59.43**	**67.92**	**39.62**	***70.75***	**67.92**	**64.15**	***42.45***	**58.49**	60.38
St. Pölten	**100**	**100**	**100**	**100**	**25.00**	**80.00**	**100**	**95.00**	**30.00**	**55.00**	X
Vienna	59.39	**70.36**	***58.45***	***70.36***	38.93	**62.10**	**70.36**	***68.28***	37.50	**61.86**	65.69
Vilach	34.29	31.43	**34.29**	***31.43***	***68.57***	**48.57**	**31.43**	31.43	***77.14***	**22.86**	59.02
Wels	**56.25**	***56.25***	**56.25**	**56.25**	**31.25**	***56.25***	**56.25**	**50.00**	**50.00**	**37.50**	56.25

Results Discussion. Our results show that the cities in Germany and the UK have a classification accuracy from 70% to 90% for parks and gardens (see Tables 1 and 2). This means, according to our generated classifiers and their mutual application in other cities, about 10% to 30% of all analyzed entities within each city might be incorrectly classified. In Austria (see Table 3) we achieve poorer results. This might be due to the relative low number of entities in the available data set, or to already existing classification problems. In some of the cities, e.g., St. Pölten only one class of entities is available or predominant and causes the classifier to be highly biased and practically unusable (see Figure 4 and Table 3).

Of course, the classification results have to be interpreted with care. In none of the selected data sets, we had a qualified reference data set of known good quality. We selected the data sets as they were, and tried to identify two size classes within them: one for gardens and one for parks. In most cities we could identify good classifiers, however, their accuracies are not verifiable to full extend. As we have no clear ground truth, we cannot claim the correctness of the classifiers. With our approach we were able to identify a large set of entities worth looking at again. All samples we inspected showed clear evidence for entities that have been classified in an inappropriate way: "parks" around residential buildings in residential areas, as well as "gardens" with typical park facilities such as ways, playgrounds, or larger water bodies.

Although these samples were randomly chosen, they showed indicators for the validity of our approach. There are other evidences about that our results point in the right direction. In April 2014 we reviewed all entities that were detected as outliers in this paper. Of the originally 24,410 detected conflicts of the hierarchy consistency analysis (see Section 5) 10,635 entities had been already corrected or removed by the OSM community. Thus, in about 40% our approach pointed to entities identified as incorrect by crowdsourcing reviewers. The classification plausibility analysis resulted in 2,023 problematic entities in Germany, 2,516 in the UK, and 1,062 in Austria. About 8% of the German entities, 8% of the UK entities, and 11% of the Austrian entities have been revised since then. It is necessary to state that they have been revised without explicitly pointing to them. An appropriate infrastructure, e.g., a website or a gamified entity checker, can help to point users to the detected entities and revise them if necessary.

Also, the developed a very simple classifiers. If we want to successfully distinguish more than two classes, we need to consider more features than just size, thus we have to learn, e.g., typically contained or surrounding features of entities. By applying the approach as discussed in Section 3, we can select the detected entities and present them in a crowdsourcing manner to volunteers for inspection. The potentially re-classified entities could be used for rebuilding the classifier with clearer evidence.

7 Conclusion and Future Work

In this work we propose a new approach to manage the quality of VGI data during contribution, and on the existing data set manually or automatically. We presented two approaches to tackle VGI quality. We mainly focused on the problem of potentially wrong classifications that might lead to heterogeneous data quality. We developed two methods to tackle hierarchical consistency and classification issues based on ambiguity of potential entity classes.

With our first method, constraint based checking of hierarchical elements, we are able to detect all inconsistencies in the existing OpenStreetMap data set. With our second method, we can identify potentially wrong areal classifications in the OpenStreetMap data set by learning classifiers of different entity classes. The results show that we can identify a large number of existing problems in OSM data with both approaches. These detected conflicts could be presented to voluntary users to validate the entities' class, potentially based on suggestions generated along with it. For more complex classifiers being able to detect multiple possible classes, like, e.g., the "green areas" on a map (parks, gardens, meadow, grassland, scrub, etc.) we need to develop meaningful classifiers considering sets of features to be learned. We also need to think about appropriate ways to implement the proposed quality assurance methods, e.g., by means of gamification of user-based validation of the detect problematic data.

Acknowledgments. We gratefully acknowledge support provided by the German Academic Exchange Service (DAAD), as well as the German Research Foundation (DFG) via the Transregional Collaborative Research Center on Spatial Cognition SFB/TR8. Furthermore, we would like to thank the anonymous reviewers for their valuable comments.

References

1. Goodchild, M.F.: Citizens as sensors: the world of volunteered geography. GeoJournal 69(4), 211–221 (2007)
2. Coleman, D.J., Georgiadou, Y., Labonte, J., et al.: Volunteered geographic information: the nature and motivation of produsers. International Journal of Spatial Data Infrastructures Research 4(1), 332–358 (2009)
3. Feick, R., Roche, S.: Valuing volunteered geographic information (VGI): Opportunities and challenges arising from a new mode of GI use and production. In: Proceedings of the 2nd GEOValue Workshop, HafenCity University Hamburg, Germany, pp. 75–79 (2010)
4. Zook, M., Graham, M., Shelton, T., Gorman, S.: Volunteered geographic information and crowdsourcing disaster relief: a case study of the Haitian earthquake. World Medical & Health Policy 2(2), 7–33 (2010)
5. Girres, J.F., Touya, G.: Quality assessment of the french OpenStreetMap dataset. Transactions in GIS 14(4), 435–459 (2010)
6. Ludwig, I., Voss, A., Krause-Traudes, M.: A comparison of the street networks of Navteq and OSM in Germany. In: Advancing Geoinformation Science for a Changing World, pp. 65–84. Springer (2011)
7. Neis, P., Zielstra, D., Zipf, A.: The street network evolution of crowdsourced maps: OpenStreetMap in Germany 2007–2011. Future Internet 4(1), 1–21 (2011)
8. Goodchild, M.F., Li, L.: Assuring the quality of volunteered geographic information. Spatial Statistics 1, 110–120 (2012)
9. Haklay, M.: How good is volunteered geographical information? a comparative study of OpenStreetMap and Ordnance Survey datasets. Environment and Planning. B, Planning & Design 37(4), 682 (2010)
10. Neis, P., Zielstra, D., Zipf, A.: Comparison of volunteered geographic information data contributions and community development for selected world regions. Future Internet 5(2), 282–300 (2013)
11. Mooney, P., Corcoran, P.: The annotation process in OpenStreetMap. Transactions in GIS 16(4), 561–579 (2012)
12. Elwood, S., Goodchild, M.F., Sui, D.Z.: Researching volunteered geographic information: Spatial data, geographic research, and new social practice. Annals of the Association of American Geographers 102(3), 571–590 (2012)
13. Mülligann, C., Janowicz, K., Ye, M., Lee, W.C.: Analyzing the spatial-semantic interaction of points of interest in volunteered geographic information. In: Egenhofer, M., Giudice, N., Moratz, R., Worboys, M. (eds.) COSIT 2011. LNCS, vol. 6899, pp. 350–370. Springer, Heidelberg (2011)
14. Vandecasteele, A., Devillers, R.: 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)
15. Schmid, F., Kutz, O., Frommberger, L., Kauppinen, T., Cai, C.: Intuitive and natural interfaces for geospatial data classification. In: Workshop on Place-Related Knowledge Acquisition Research (P-KAR), Kloster Seeon, Germany (2012)
16. Schmid, F., Frommberger, L., Cai, C., Dylla, F.: 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, p. 8. ACM (2013)
17. Schmid, F., Frommberger, L., Cai, C., Freksa, C.: 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, pp. 3–19. Springer International Publishing (2013)

18. Arteaga, M.G.: Historical map polygon and feature extractor. In: Schmid, F., Kray, C. (eds.) Proceedings of ACM MapInteract, 1st International Workshop on Map Interaction. ACM (2013)
19. Neis, P., Zipf, A.: 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)
20. Barron, C., Neis, P., Zipf, A.: A comprehensive framework for intrinsic OpenStreetMap quality analysis. Transactions in GIS 18 (2014)
21. Tobler, W.R.: A computer movie simulating urban growth in the detroit region. Economic Geography 46, 234–240 (1970)
22. Devogele, T., Parent, C., Spaccapietra, S.: On spatial database integration. International Journal of Geographical Information Science 12(4), 335–352 (1998)
23. Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag New York, Inc., Secaucus (2006)
24. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques, 3rd edn. Morgan Kaufmann Publishers Inc., San Francisco (2011)
25. Cover, T., Hart, P.: Nearest Neighbor pattern classification. IEEE Transactions on Information Theory 13(1), 21–27 (1967)
26. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)
27. Kohavi, R., et al.: A study of Cross-Validation and Bootstrap for accuracy estimation and model selection. In: Proc. International Joint Conference on Artificial Intelligence (IJCAI), pp. 1137–1145 (1995)
28. Fawcett, T.: An introduction to ROC analysis. Pattern Recognition Letters 27(8), 861–874 (2006)