Enhancing Data Classification Quality of Volunteered Geographic Information

Towards guided human-centered data classification

Ahmed Loai Ali

Dissertation
zur Erlangung des Grades eines
Doktors der Ingenieurwissenschaften
– Dr.-Ing. –

Vorgelegt im
Fachbereich 3 (Mathematik & Informatik)
Universität Bremen

September 2016

Die vorliegende Doktorarbeit wurde in der Zeit von April 2013 bis Juli 2016 in der Arbeitsgruppe Cognitive Systems am Bremen Spatial Cognition Center der Universität Bremen erstellt.

1. Gutachter: Professor Christian Freksa, Ph.D. (Universität Bremen)
2. Gutachter: Professor Dr. Alexander Zipf. (Universität Heidelberg)
Datum des Promotionskolloquiums: 23.09.2016.

Abstract

Geographic data is one of the fundamental components of any Geographic Information System (GIS). Nowadays, the utility of GIS becomes part of everyday life activities, such as searching for a destination, planning a trip, looking for weather information, etc. Without a reliable data source, systems will not provide guaranteed services. In the past, geographic data was collected and processed exclusively by experts and professionals. However, the ubiquity of advanced technology results in the evolution of *Volunteered Geographic Information* (VGI), when the geographic data is collected and produced by the general public. These changes influence the availability of geographic data, when common people can work together to collect geographic data and produce maps. This particular trend is known as collaborative mapping. In collaborative mapping, the general public shares an online platform to collect, manipulate, and update information about geographic features. OpenStreetMap (OSM) is a prominent example of a collaborative mapping project, which aims to produce a free world map editable and accessible by anyone.

During the last decade, VGI has expanded based on the power of crowdsourcing. The involvement of the public in data collection raises great concern about the resulting data quality. There exist various perspectives of geographic data quality; this dissertation focuses particularly on the quality of data classification (i.e., thematic accuracy). In professional data collection, data is classified based on quantitative and/or qualitative observations. According to a pre-defined classification model, which is usually constructed by experts, data is assigned to appropriate classes. In contrast, in most collaborative mapping projects data classification is mainly based on individuals' cognition. Through online platforms, contributors collect information about geographic features and transform their perceptions into classified entities. In VGI projects, the contributors mostly have limited experience in geography and cartography. Therefore, the acquired data may have a questionable classification quality.

This dissertation investigates the challenges of data classification in VGI-based mapping projects (i.e., collaborative mapping projects). In particular, it lists the challenges relevant to the evolution of VGI as well as to the characteristics of geographic data. Furthermore, this work proposes a guiding approach to enhance the data classification quality in such projects. The proposed approach is based on the following premises: (i) the availability of large amounts of data, which fosters applying machine learning techniques to extract useful knowledge, (ii) utilization of the extracted knowledge to guide contributors to appropriate data classification, (iii) the humanitarian spirit of contributors to provide precise data, when they are supported by a guidance system, and (iv) the power of crowdsourcing in data collection as well as in ensuring the data quality.

This cumulative dissertation consists of five peer-reviewed publications in international conference proceedings and international journals. The publications divide the dissertation into three parts: the *first* part presents a comprehensive literature review about the relevant previous work of VGI quality assurance procedures (Chapter 2), the *second* part studies the foundations of the approach (Chapters 3-4), and the *third* part discusses the proposed approach and provides a validation example for implementing the approach (Chapters 5-6). Furthermore, Chapter 1 presents an overview about the research questions and the adapted research methodology, while Chapter 7 concludes the findings and summarizes the contributions.

The proposed approach is validated through empirical studies and an implemented web application. The findings reveal the feasibility of the proposed approach. The output shows that applying the proposed approach results in enhanced data classification quality. Furthermore, the research highlights the demands for intuitive data collection and data interpretation approaches adequate to VGI-based mapping projects. An interaction data collection approach is required to guide the contributors toward enhanced data quality, while an intuitive data interpretation approach is needed to derive more precise information from rich VGI resources.

Zusammenfassung

Geographische Daten sind eine der wichtigsten Komponenten eines jeden Geoinformationssystems (GIS). Heutzutage werden GIS zunehmend in Alltagssituationen eingesetzt z.B. für die Suche eines Zielorts, die Planung einer Route, zum Abruf von Wettervorhersagen usw. Ohne eine zuverlässige Datenquelle funktionieren solche Dienste jedoch nicht wie beabsichtigt. In der Vergangenheit wurden geographische Daten ausschließlich von Fachleuten und Experten gesammelt und verarbeitet. Inzwischen ist es durch die Verbreitung fortschrittlicher Technologien möglich, dass im Rahmen von *Volunteered Geographic Information* (VGI; deutsch: "freiwillig erhobene geographische Informationen") geographische Daten von der Öffentlichkeit gesammelt und erstellt werden. Diese Veränderung beeinflusst die Verfügbarkeit von geographischen Daten, da beim VGI eine Gruppe von Menschen zusammen arbeiten kann, um Daten zu sammeln und Karten zu erstellen. Dieser Trend wird auch als kollaboratives Mapping bezeichnet. Beim kollaborativen Mapping wird eine Online Plattform verwendet, um Daten über geographische Eigenschaften zu sammeln, zu bearbeiten und zu aktualisieren. Ein prominenter Vertreter des kollaborativen Mappings ist OpenStreetMap (OSM), welches das Ziel verfolgt, eine für jeden frei verfügbare und bearbeitbare Karte der Welt bereit zu stellen.

Während des letzten Jahrzehnts hat sich VGI mit Hilfe des Crowdsourcing weiterverbreitet. Die Beteiligung der Öffentlichkeit an der Erhebung geographischer Daten führt jedoch zu Bedenken bezüglich der Datenqualität. Es existieren unterschiedliche Betrachtungsweisen hinsichtlich der Qualität von geographischen Daten. Diese Dissertation beschäftigt sich hauptsächlich mit der Qualität der Datenklassifizierung (d.h. thematische Genauigkeit). In der professionellen Datenerhebung werden die Daten nach quantitativen und/oder qualitativen Kriterien klassifiziert. Mit Hilfe eines vordefinierten Klassifikationsmodells, welches von Experten erstellt wird, werden die Daten geeigneten Klassen zugeordnet. Im Gegensatz dazu basiert die Datenklassifikation im Rahmen der meisten kollaborativen Mapping Projekte auf dem subjektiven Eindruck der beitragenden Individuen. Über Online-Plattformen werden Informationen über geografische Objekte gesammelt indem die Mitwirkenden ihre subjektiven Eindrücke in klassifizierte Einträge umwandeln. In VGI-Projekten haben die Mitwirkenden in der Regel begrenzte Erfahrung in der Domäne Geographie und Kartographie. Als Folge ist eine hohe Qualität der resultierenden Klassifikationen nicht gewährleistet.

Diese Dissertation befasst sich mit den Herausforderungen die im Rahmen der Klassifizierung von Daten in VGI basierten Mapping Projekten (z.B. kollaborative Mapping Projekte) auftreten. Insbesondere werden die Herausforderungen, welche für die Entstehung von VGI und die Charakteristik von geographischen Daten relevant sind aufgelistet. Darüber hinaus wird ein Ansatz zur Verbesserung der Klassifizierung von Daten in solchen Projekten vorgestellt. Der vorgeschlagene Ansatz macht sich die folgenden Eigenschaften von VGI zu nutze: i) die Verfügbarkeit großer Datenmengen, durch die die Anwendung von Techniken des Maschinenlernens möglich wird, (ii) die Verwendung von extrahiertem Wissen zur Unterstützung einer korrekten Datenklassifikation, iii) die Bereitschaft und das Streben der Mitwirkenden präzise Daten zur Verfügung zu Stellen, wenn Sie von einem Leitsystem unterstützt werden, und iv) der Nutzen des Crowdsourcing sowohl für die Datenerfassung als auch für die Prüfung der Datenqualität.

Diese kumulative Dissertation besteht aus fünf begutachteten Artikeln, welche in internationalen Konferenzberichten bzw. internationalen Fachzeitschriften veröffentlicht wurden. Die Veröffentlichungen gliedern die Dissertation in drei Teile. Der erste Teil enthält eine umfassende Literaturübersicht existierender Arbeiten zu Methoden der Qualitätssicherung in VGI (Kapitel 2). Der zweite Teil prüft die Grundlagen des vorgeschlagenen Ansatzes (Kapitel 3-4) und der dritte Teil diskutiert den vorgeschlagenen Ansatz und stellt ein Beispiel zur Validierung der Implementierung des Ansatzes vor (Kapitel 5-6). Des weiteren gibt Kapitel 1 einen Überblick über die wissenschaftliche Fragestellung und die verwendete Methodik, während Kapitel 7 die Ergebnisse und den wissenschaftlichen Beitrag der Arbeit zusammenfasst.

Der vorgeschlagene Ansatz wird durch eine Reihe von empirischen Untersuchungen sowie durch eine eigene dafür entwickelte Web-Anwendung validiert. Die Validierungsergebnisse sprechen für die Durchführbarkeit des vorgeschlagenen Ansatzes und weisen nach, dass mit dem Ansatz die Qualität der Klassifikation verbessert werden kann. Ferner verdeutlicht die gesamte Dissertation den Bedarf von VGI-basierten kollaborativen Mapping Projekten an interaktiven Datenerfassungsschnittstellen und intuitiven Ansätzen zur Dateninterpretation. Während für die Dateneingabe eine intelligente Benutzerschnittstelle erforderlich ist, um die Beiträge der Anwender qualitativ zu verbessern, ist eine intuitive Dateninterpretation notwendig, um den Informationsgewinn – ungeachtet der Qualität der bereitgestellten Daten – zu erhöhen.

To Soul of my Parents ...

To my *Wife* and my *Daughters* ...

Acknowledgements

Thank GOD for blessing, supporting, and giving me patience during this work and in all moments. Firstly, I have to admit that I am lucky to work under the supervision of Prof. Christian Freksa. No words could describe great supports that I got from him during this period. Thank you, Christian, for everything you did for me. Through your continuous encouragement, I realized the real meaning of the German word "Doctorvater". Thank you for giving me freedom to select my research topic, for your supportive words in hard moments, for your comments and discussions, and for your efforts to secure my future career. Under your supervision, I learned how to enjoy doing research ("It is fun!").

Special acknowledge for Prof. Alexander Zipf. The output of your research and my short visit to your productive group influenced my research perspectives significantly. Your feedback and discussions substantially enhance my research directions. Thank you in advance for reviewing this dissertation.

To Falko Schmid, you are my tutor and my closest friend here in Germany. Our scientific discussions acted as a kick-off of my research career. Your friendly advices, suggestions, and instructions helped me a lot in research, as well as in evolving my life in Germany. Thank you for our informal gatherings.

To all co-authors, reviewers and editors of my publications, Thank you for your constructive discussions, comments, and suggestions. You helped me a lot and supported maturity of this work.

I am grateful to the German Academic Exchange Services (DAAD), the Egyptian Ministry of Higher Education (MoHE), and Faculty of Computer and Information (FCI) at Assiut University for giving me this great opportunity to pursue my research in Germany. Without this scholarship, I would not be able to develop my academic skills and experience. Thanks to University of Bremen and the COST Action IC1203 ENERGIC, supported by COST (European Cooperation in Science and Technology) for funding my participations in various scientific events.

To CoSy group, with you, I have never had the feeling of being a foreigner. Your friendly discussions, during graduate seminars, changed my views of doing science. Thanks to the people by name: Zoe, Jasper, Frank, Thomas, Holger, Julia, and Ana-Maria. To Gracia and Alexander, without you, several administrative and technical issues were not be solved. Rami, although I missed you in CoSy, I will not forget to acknowledge you. Thanks to everyone was in CoSy: former researchers, principal investigators, and visitors.

My Classmates and Friends, either in Egypt, in Germany, or elsewhere, our true friendship would be the best forever, Thank You, for keeping in contact ✉ over distance, for sharing with me good and hard moments, and for spending a good time ☕ with me over the years. You are part of this work. Although pages are too limited to mention your names and your supports, I would never forget how great You are. Thanks !

My large family: Brothers (Haitham & Ehab), Parents-in-law (Prof. Sameeh & Mrs. Nagah), and Brothers-in-law (Mohamed & Mostafa & Seif). I appreciate how much efforts and sacrifices you did/are doing for me: your continuous encouragement, your prays for me, your management of my issues, ... more and more. You are an essential part of my success. I am lucky to have you all in my life. May GOD bless you.

Last but not least, no text could express my gratefulness to my little Family. My wife, ♡ ***Nour*** ♡, your endless love is the main source of my happiness. Nothing would be done at all without you in my life. Your heart paved my way of success. I believe in "*Behind every successful Man is a great Woman*". You are a legend: wife, friend, mother, computer scientist, and more. My beloved awesome daughters, ***Halla & Habiba***, your hugs and smiles ☺☺, everyday, were recovering and removing my pains. Thank you all for being patient during this long trip, for sharing with me homesickness, for your true love, and for your endless sacrifices. Many Thanks !

Finally, to my Parents who passed away in 2015, I missed you a lot ☹☹. Although you did not reach the accomplishment of this work, your unforgettable encouragements and sacrifices were always my motives to success. Without you behind me, this work would never be complete. May GOD bless your Souls.

Contents

List of Figures

List of Tables

Abbreviations

AC	**A**ssociation **C**lassification
AUC	**A**rea **U**der the ROC **C**urve
CBR	**C**ase **B**ased **R**easoning
CV	**C**ross **V**alidation
DE-9IM	**D**imensionally **E**xtended nine-**I**ntersection **M**odel
EL	**E**ager **L**earning
G&G	**G**rass&**G**reen
GIS	**G**eographic **I**nformation **S**ystem
GNSS	**G**lobal **N**avigation **S**atellite **S**ystem
GPS	**G**lobal **P**ositioning **S**ystem
ICT	**I**nformation and **C**ommunication **T**echnologies
ISO	**I**nternational **S**tandardization **O**rganization
KNN	**K**-**N**earest **N**eighbours
LBS	**L**ocation **B**ased **S**ervices
LC	**L**and **C**over
LCA	**L**aten **C**lass **A**nalysis
LL	**L**azy **L**earning
LU	**L**and **U**se
NN	**N**eural **N**etwork
OSicM	**O**pen**Scie**nce**M**ap
OSM	**O**pen**S**treet**M**ap
POI	**P**oint **O**f **I**nterests
PTV	**P**ossibilistic **T**ruth **V**alue
UGC	**U**ser **G**enerated **C**ontent
UTM	**U**niversal **T**ransverse **M**ercator
SDI	**S**patial **D**ata **I**nfrastructure
ROC	**R**eceiver **O**peration **C**haracteristics
SVM	**S**upport **V**ector **M**achine
VGI	**V**olunteered **G**eographic **I**nformation

Chapter 1

Introduction

This dissertation focuses on Volunteered Geographic Information (VGI), which results from collaborative mapping activities. In particular, this research investigates the potential utility of VGI as a complementary data source for land use and land cover mapping. However, there are several challenges to ensure the data quality. This dissertation gives an overview of the quality assurance procedures of VGI. It investigates, with more focus, the quality of data classification. Furthermore, the dissertation proposes a guided-classification approach to enhance data classification quality in VGI resources. This chapter presents an overview of VGI evolving, discusses the challenges of quality assurance in VGI, and highlights the scope of this research. In addition, the chapter includes the research foundations and methodologies, the organization of the following chapters, and a list of the contributions.

1.1 Motivation

Digital forms of geographic information date to the earliest Geographic Information System (GIS) in the 1960s. In a broad sense, digital geographic data is one of the five fundamental elements of any GIS application (DeMers, 2009). Without data, GIS applications would not be able to provide reliable services. Over the past decades, the availability of digital geographic data has changed dramatically. In the past, users had to wait – sometimes years – for mapping agencies and large organizations to produce digital maps. The process was sophisticated, time consuming, and costly. Moreover, the fields of geographic data collection and map production were exclusively reserved for cartographers and well-trained experts (Cowen, 2008). Nowadays, the advanced web technologies (e.g., Web 2.0 (O'Reilly, 2009)) and the ubiquity of location sensing devices (e.g., smartphones) empower ordinary users to take part in the process of geographic data

production. By exploiting information and communication technologies (ICT), users are no longer only passive receivers of data, but they turned to be active data producers as well. In 2006, the rise of geo-crowdsourcing – as a bottom-up paradigm of geographic data collection – results in producing various formats of geographic content in addition to changing the conventional ways of mapping (Howe, 2006). It supports the evolving of a special kind of user-generated content (UGC), which has been known as *Volunteered Geographic Information* (VGI). The term has been coined by Goodchild (2007), who described it as a phenomenon, when humans act as sensors to collect geographic data ("citizen as sensors"). In VGI projects, ordinary users utilize online platforms to produce different forms of content associated with geographic information implicitly or explicitly. Among others, Wikipedia[1], Wikimapia[2], OpenStreetMap[3] (OSM), Flickr[4], Twitter[5], Google Map Maker[6] and Foursquare[7] are examples of platforms that generate various formats of VGI. According to the contributors' intentions, VGI is classified as *Active* or *Passive*. It is also classified according to the geographic contents into *Aspatial* (e.g., wikipedia) or *Georeferenced* (e.g., OSM) (See et al., 2016). This dissertation focuses exclusively on *Active/Georeferenced* VGI that results from collaborative mapping (Mac Gillavry, 2006), when users intentionally participate in the process of mapping geographic features.

During the last decade, VGI has evolved in a dramatic fashion to be utliized as an individual or a complementary data source in various GIS applications (Heipke, 2010). Moreover, researchers argue about its potential role, as a fundamental component, in spatial data infrastructure (SDI) (McDougall, 2009; Cooper et al., 2011), and consequently, in developing reliable GIS applications.

Although most VGI projects do not have standard procedures to ensure the quality of the resulting data, the data acts as a data source in various applications, such as land use and land cover mapping (Fritz et al., 2012; Arsanjani et al., 2015; Vaz and Jokar Arsanjani, 2015), crisis management (Goodchild and Glennon, 2010; Zook et al., 2010; Roche et al., 2013), demographic studies (Chow et al., 2012; Chow, 2013), urban planning (Foth et al., 2009), map provision (Haklay and Weber, 2008), environmental monitoring (Gouveia and Fonseca, 2008), and numerous applications of location-based services (LBS) (Savelyev et al., 2011; Thatcher, 2013).

[1] www.wikipedia.org
[2] www.wikimapia.org
[3] www.openstreetmap.org
[4] www.flickr.com
[5] www.twitter.com
[6] www.google.com/mapmaker
[7] www.foursquare.com

Hence, different applications can be developed based on various formats of VGI. Each individual type of application requires particular concerns regarding data quality assurance. The technical and non-technical foundations of VGI are presented in Section 1.1.1, while Section 1.1.2 discusses the VGI quality assurance. Moreover, Section 1.1.3 focuses on the data classification problem, whereas scenarios of appropriate and inappropriate data classification are illustrated in Section 1.1.4.

1.1.1 Foundations of VGI

VGI has evolved adopting the success of Wikipedia, when anyone with an access to the Internet could be able to provide information, however, the information here is related to geographic locations. The birth of VGI is based on the development of technologies that empower users to produce geographic content: (1) *Georeferencing*: when users are enabled to assign spatial coordinates to data using global coordinate systems like Universal Transverse Mercator (UTM), (2) *Geotagging*: a standardized format of assigning geographic information to content, (3) *Global Navigation Satellite System (GNSS)*: when global geospatial positioning technology is provided to the public without further restrictions, and (4) *Broadband communication*: the high capacity Internet connections, which are now available to most households particularly in developed countries (Goodchild, 2007).

Regarding non-technical foundations, the power of VGI comes from the local knowledge of contributors (Heipke, 2010). For example, when we visit a new place, we most likely ask local people for the location of a particular place or the route for a certain destination. Hence, VGI projects have been developed to promote people to contribute their local geographic knowledge to develop rich geographic content. In such content, the data is provided by public individuals, regardless of their background and their geographic experience. This fact raises research concerns about the resulting data quality.

1.1.2 Quality assurance of VGI

The International Organization for Standardization[8] (ISO) has developed standards for geographic information in (ISO/TC 211)[9] (Østensen and Smits, 2002; ISO, 2009). In particular, ISO/TS 19113 includes principles that describe the geographic data quality and specifications. They proposed five basic measures for geographic data quality: *positional accuracy, completeness, lineage, logical consistency*, and *thematic accuracy.*

[8] www.iso.org
[9] www.isotc211.org

In addition, other measures of *semantic accuracy* and *temporal information* have been developed thereafter (Shi et al., 2003; Guptill and Morrison, 2013).

On the professional level, mapping agencies follow the ISO standard procedures to ensure the data quality. The quality assurance procedures are divided into pre- and post-procedures; when the pre-procedures describe the standards that should be followed during data acquisition and compilation processes; and the post-procedures describe the validation and the documentation processes for the developed data. The entire procedures are usually attached to the data source as meta-data for the assessment of the purpose of use.

In contrast in VGI projects, the data is acquired through crowdsourcing following no quality assurance procedures. Therefore, the resulting data possess a questionable quality. This stimulates the researchers to develop procedures to ensure and assess the resulting data quality. Goodchild and Li (2012) proposed three intuitive approaches to ensure the quality of VGI: crowdsourcing, social, and geographic approaches.

- crowdsourcing: data quality is ensured through the "wisdom of crowds", when a group of people might be able to accomplish a solution that experts might not able to do (Surowiecki, 2005). By following Linus's Law "given enough eyes, all bugs are shallow", a large group of people will be able to validate and correct the contributions of each other (Raymond, 1999).
- social: in this approach data quality is ensured by analyzing the characteristics of its producers, where the communication between producers generates a reputation indicator of the data quality. This approach simulates the hierarchical formal structure in professional organizations, but in a voluntary structure. In VGI, different contributors play different roles; when some are used to add new content and others are interested in validating the content.
- geographic: the approach follows Tobler's Law "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970). The second inclusion of the law points to the consistency between content and its geographic context. For example, an image with a content about "Brandenburg Gate" in Berlin, Germany and a geolocation of Jakarta, Indonesia should be detected as an outlier (see Figure 2.1). Furthermore, geographic rules could be set up to ensure data integrity.

In general, VGI is assessed by following either extrinsic or intrinsic approaches. In extrinsic approaches, the data is matched and compared with a reference data source – when the latter is available – to determine a particular measure of data quality (e.g.,

completeness). In intrinsic approaches, the data is evaluated by analyzing its internal characteristics (i.e., looking for the meta-data) to find out indicators of the data quality. Several research in literature follow the approaches of Goodchild and Li to assess the VGI intrinsically. Trustworthiness (Bishr and Kuhn, 2007; Keßler and Groot, 2013), credibility (Flanagin and Metzger, 2008), fitness of use (Barron et al., 2014), and reputations (Bishr and Kuhn, 2013; D'Antonio et al., 2014) are examples of intuitive indicators that have raised as qualitative indicators of data quality. To assess measures and indicators of various formats of VGI, researchers applied different methods: from direct methods like matching and comparison (Haklay, 2010; Ludwig et al., 2011; Dorn et al., 2015) through statistical methods (Foody et al., 2015; Sparks et al., 2015) to machine learning (Huang et al., 2010; Castillo et al., 2011).

1.1.3 Data classification in VGI: the case of OSM

This dissertation addresses quality from the perspective of data classification (i.e. thematic accuracy). We utilized OSM data, as an example of VGI mapping project. OSM is the most prominent VGI-based mapping project, which aims to develop a free digital world map that is editable and obtainable by anyone (Bennett, 2010). Regarding data classification, the OSM project provides suggestions and recommendations on the project Wiki pages[10]. These recommendations describe the appropriate ways of mapping (e.g., delineating) and classifying different geographic features, even in different geographic locations or cultures. These recommendations are based on discussions between local mapping communities.

In OSM data, each entity is classified by means of tags; when a tag has the format of $Key = Value$; the Key describes the classification perspective (e.g., landuse, highway, building, etc.), while the $Value$ describes a specific class (e.g., "forest" (landuse), "primary" (highway), "public" (building), etc.). There is no limitation on the number of tags that describe each entity (Bennett, 2010; Mooney and Corcoran, 2012b). According to the scope of this dissertation, all tags related to land use and land cover features are provided in Appendix A.

In VGI mapping, the data classification is related to human cognition, when contributors interpret their qualitative/quantitative observations into classes, aligned with the provided recommendations. Moreover, in most projects, there are neither standard procedures nor integrity checking mechanisms to ensure data quality. Therefore, the resulting data inherits a problematic data classification (Mooney and Corcoran, 2012b).

[10] http://wiki.openstreetmap.org/wiki/Map_Features

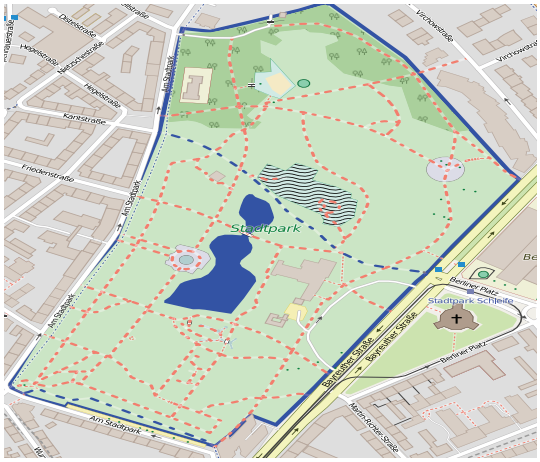
1.1.4 Scenarios of appropriate and inappropriate data classification

In this dissertation, we are concerned with the quality of data classification in VGI mapping projects. In such projects, limited contributors' experience, ambiguous definitions of geographic features, and flexible contribution mechanisms might lead to problematic data classification. This section presents examples of what we call "appropriate" and "inappropriate" classification. In this dissertation, classification appropriateness is defined with respect to land use and land cover features; *appropriate classification* of an entity must reflect its internal and external characteristics. In addition, it should be consistent with its geographic context and indicates the potential utilization of the entity.

Figure 1.1 illustrates examples of an appropriate classification, where the target entities are outlined with blue colour in Figures 1.1a and 1.1c. They are both classified

(A) OSM map of "Volkspark Marienberg", Nuremberg, Germany.

(B) Google map of "Volkspark Marienberg", Nuremberg, Germany.



(C) OSM map of "Stadtpark", Nuremberg, Germany.

(D) Google map of "Stadtpark", Nuremberg, Germany.

FIGURE 1.1: Examples of appropriate classification. The entities are outlined by blue colour in the OSM map (on the left hand), and they are presented also visually from Google maps (on the right hand).

on the OSM project by the tag of "leisure=park". Figures 1.1b and 1.1d represent the corresponding satellite views of these entities, respectively. By analyzing both entities visually, we can realize the typical amusement and entertainment characteristics of parks. They are paved with footways (red dotted lines) and cycleways (blue dotted lines), include water bodies (blue areas), are located in a residential area (like in "Stadtpark") or surrounded by other grass-related features (like in "Volkspark Marienberg"), and might contain a playground (waved area), cafe, restaurant, and/or sport areas.

(A) OSM map of "Leipziger Platz", Nuremberg, Germany.

(B) Google map of "Leipziger Platz", Nuremberg, Germany.

FIGURE 1.2: Example of inappropriate classification. The target entities are highlighted with red colour on 1.2a, while 1.2b shows the entities visually from Google maps.

In contrast, Figure 1.2 illustrates an example of inappropriate classification of the same type of feature. The indicated entities are outlined by red colour on Figure 1.2a. They are classified on the OSM project by the tag "leisure=park". Figure 1.2b shows the satellite view of the indicated entities from Google maps. By inspecting the entities visually, we can realize that they are located at the corner of a public square called "Leipziger Platz" (dotted area on the OSM); they include only a limited number of sparse trees; and they do not reflect any amusement or entertainment characteristics of parks.

Another example in Figure 1.3 illustrates the problem of conceptually overlapping classes. The figure shows three entities in the OSM data: 1) blue; 2) green; and 3) red outlined entities. The first entity (blue outline) has the name "Kontumazgarten", while it is classified as "park". It has similar characteristics as the second entity (green outline) with a slight difference: it includes playgrounds; however, the latter is classified as "meadow". The entity's classification shows the conceptual overlap between the classes "park", "garden", and "meadow". The third entity (red outline) shows another example of inappropriate classification; the indicated entity is classified as "park", while it is a small grass area (in comparison to the entities in Figure 1.1) and it is located in a backyard of a "hospital" and a "parkhaus".

(A) Overlapping classes at the OSM project: green ("meadow"), blue and red ("park")

(B) Visual view of the identical location from Google maps.

FIGURE 1.3: Example of problematic classification due to conceptual overlapping classes between "park", "garden", and "meadow" classes.

The previous examples indicate the problematic data classification in VGI mapping projects. On the one hand, contributors' preferences play a major role in classifying the data. All examples are chosen from the same city (Nuremberg, Germany), to show different individual perceptions within the same mapping community. On the other hand, the non-rigid boundaries between similar classes might result in conceptually overlapping classes. Therefore, a given entity could plausibly belong to multiple classes with various degrees of appropriateness.

1.2 Research Focus and Questions

The motto of this research is: "exploiting VGI to develop reliable GIS applications requires ensuring data quality". Despite the technologies facilitate the production of massive data, the data quality is still a matter of concern regardless of the data quantity.

Figure 1.4 illustrates the conceptual framework of VGI from the production to the utilization. The highlighted part of the figure indicates the focus of this dissertation; from the bottom, the framework starts with the contributors (i.e., the power of any VGI project), who are utilizing different platforms to generate various formats of VGI. Different formats are used to support numerous kinds of applications, and hence, each data format requires particular procedures of quality assurance. Quality assurance is an intermediate layer that links the data production and the effective data utilization; when a quality assurance procedure consists of *approaches, methods,* and *measures/indicators* (see Chapter 2).

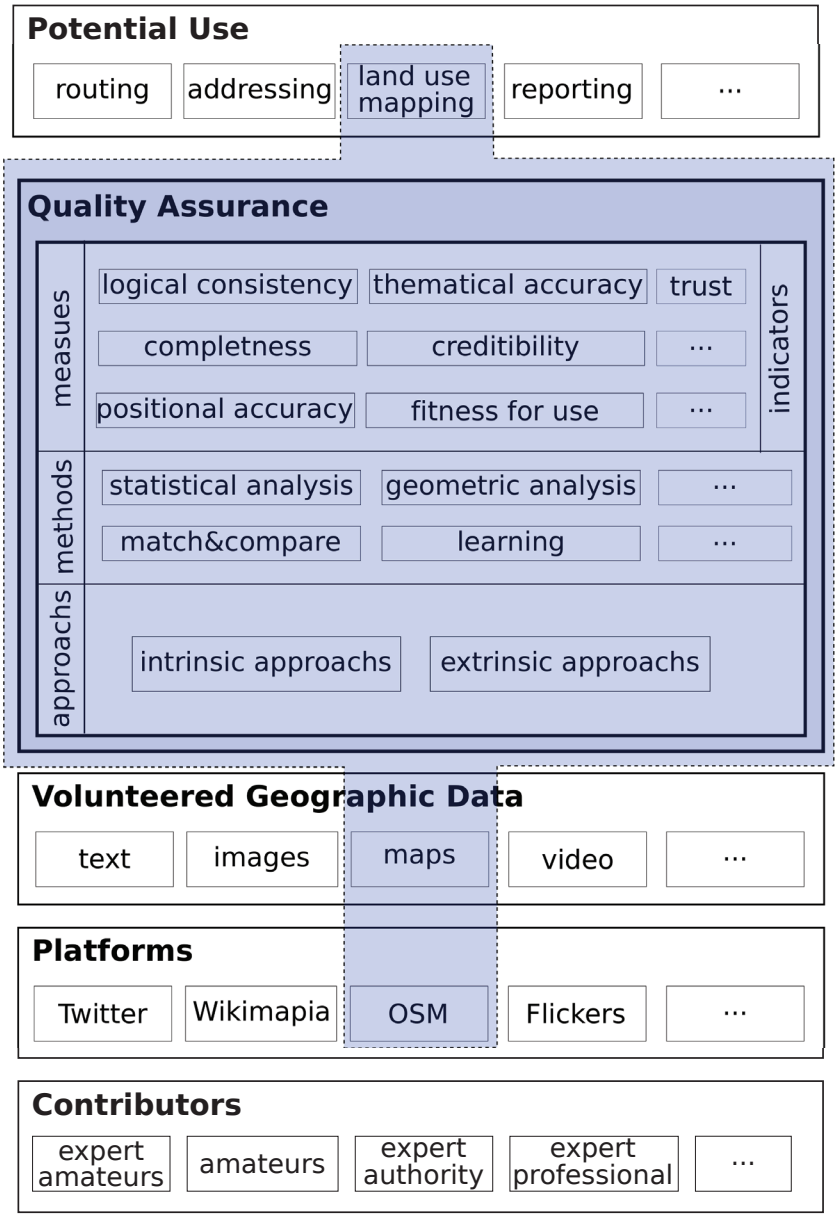


FIGURE 1.4: Conceptual framework of various formates of VGI from production to utilization, with the focus of this dissertation highlighted.

From a broad perspective, with more focus on map-based VGI the dissertation addresses the question:

Q1. *What are appropriate quality assurance procedures for VGI mapping projects?*

From a particular point of view, the dissertation investigates the exploitation of VGI as a complementary data source for land use and land cover thematic maps. Several publications emphasize the applicability of VGI as a potential data source for these features (Mooney et al., 2010; Hagenauer and Helbich, 2012; Arsanjani et al., 2015; Dorn et al., 2015). Nevertheless, the classification of these features in general poses

several challenges related to human cognition (Ahlqvist, 2012). Thus, the dissertation focuses on VGI mapping projects to answer the question:

Q2. *What are the challenges of data classification in VGI mapping projects?*

In VGI, contributors' cognition determines the data classification. For example, whether an areal water body is classified as "lake", "pond", or "reservoir"; and if a land parcel covered by grass and mixed vegetation is classified as "park", "garden", or "forest"; the classification depends on individual perception. Humans perceive geographic features differently, and consequently, they interpret their observations in different ways. This fact stimulates the idea of enhancing data classification by developing a guiding approach. From a cognitive perspective, humans might be able to provide appropriate data classification, whenever they are guided. Whether an entity is classified in an appropriate or inappropriate way is related to quantitative and qualitative observations. The lack of contributors' experience, particularly of the non-experts, might lead to misinterpretation of observations, and hence, inappropriate classification. Hence, with the availability of a large amount of data in the OSM project, the dissertation answers the question of:

Q3. *Can we learn the distinct characteristics (observations) of a specific geographic feature from VGI?*

If so,

Q4. *How can we use extracted knowledge to detect outliers and to guide contributors towards the most appropriate classification?*

In VGI-based mapping projects, guiding the contributors might conflict with their flexibility, and hence, influence their motivations negatively. Otherwise, their local knowledge is the fundamental source of information. Therefore, this work involves contributors in enhancing data classification quality by proposing human-centered guiding (i.e., recommendation) approach. The dissertation addresses the questions:

Q5. *What is the proper way to involve contributors in enhancing data classification quality?*

Q6. *How can we guide contributors intuitively and preserve their flexibility?*

And finally, the dissertation answers the question:

Q7. *Would the proposed approach enhance data classification quality?*

1.3 Research Foundations and Methodologies

This dissertation is based on the following foundations:

- Since VGI evolution in 2007, different quality assurance procedures have been developed to cope with this paradigm of geographic data collection. There is a consensus in the research field regarding the adequacy of intrinsic data quality assurance approaches to the characteristics of VGI.
- Among other projects, the OSM project – in most parts of the world – has massive amounts of data with a remarkable quality, particularly in urban areas of developed countries (e.g., Germany, the UK, and USA).
- In VGI mapping projects, although humans are eager to provide data, they lack guidance and aiding tools. Thus, the resulting data sources are rich regarding the content, but limited regarding the quality.
- The availability of rich VGI resources facilitate applying machine learning techniques to extract useful knowledge. This utility can be exploited to enhance data classification quality.

To address the presented research questions based on the aforementioned foundations, I adopt the following methodologies:

- Review previous related research of VGI quality assurance with a particular concern on VGI-mapping.

As exemplification, during this research I targeted the OSM project and the resulting data to investigate the utility of VGI as a potential data source of land use and land cover maps. The objectives are to:

- Study and understand the data classification of various geographic features to highlight the challenges of the process.
- Exploit the availability of data to apply machine learning methodologies to tackle data classification quality.
- Adopt the idea of developing human-centered guiding approach to improve the quality of data classification.
- Take advantage of crowdsourcing by employing voluntary contributors in the process of data classification enhancement as well as in data collection process.
- Conduct empirical studies to check the feasibility of the proposed approach.

1.4 Dissertation Output and Contributions

In Chapter 2, we review the quality assurance procedures related to various kinds of VGI. In this work, we conduct a survey to investigate the quality regarding image-based, text-based, and map-based VGI. This survey studies the previous related research of VGI quality assurance starting from evolving of the term in 2007 until the middle of 2015. The procedures described in the literature are classified according to the proposed approaches, the utilized methods, the quality measures/indicators and the VGI formats. According to the 56 papers reviewed, there exist 17 different measures/indicators that can be used to assess the data quality. The survey includes 30 methods that have been developed to assess VGI. The methods are grouped according to the proposed approaches into: crowdsourcing, social, geographic, and data mining. The review highlights the promising role of data mining in assessing, as well as in enhancing the VGI quality.

With focus on map-based VGI, we investigate various geographic features to understand the challenges of data classification. Different geographic features follow various structures of data classification; the features either follow a strict hierarchical structure (e.g., administrative boundary) or they follow a loose structure (e.g., land use). Whatever, the elementary step to tackle data classification is to be able to detect potentially problematic classified data (i.e., outliers). Due to the availability of large amounts of data, we examine the feasibility of applying machine learning, particularly data mining techniques, to detect outliers.

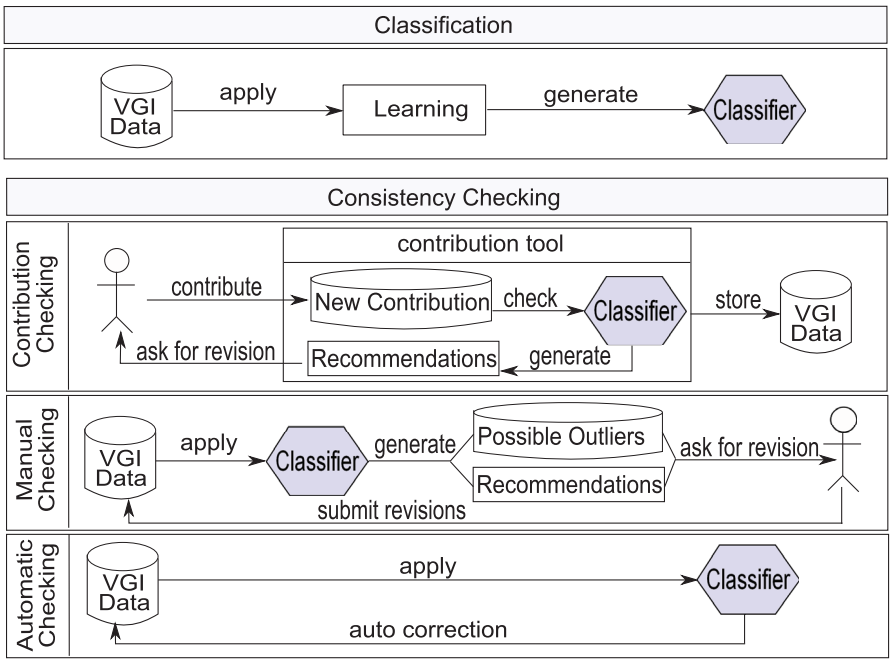


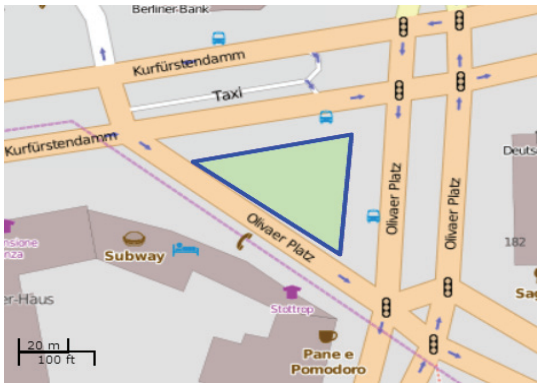
FIGURE 1.5: The proposed approaches to ensure data classification quality in VGI mapping projects.

Chapters 3 and 4 present the learning-based approach to tackle the data classification. Figure 1.5 illustrates the proposed approach. According to the figure, the approach consists of two phases: *Classification* and *Consistency Checking*.

In brief, the first phase aims to learn the characteristics that probably distinguish and/or describe a specific geographic feature. These characteristics might be quantitative (based on measures) or qualitative (relative to the context) and significantly identify this specific feature. The objective is to develop a model (i.e., classifier) that will be able to detect the problematic data, as well as, suggest the most proper classification of a given data. In the second phase, the approach proposes three scenarios to employ the developed model; 1) Contribution Checking: when the developed model can be encoded in an editing tool to detect the outliers and to suggest recommendation on the fly at contribution time, 2) Manual Checking: when the model is applied directly to an existing data set, it acts to present the potential outliers associated with recommendations for crowdsourcing validation, and 3) Automatic Checking: when the model is able to justify the recommendations then an auto correction might be possible. In the first and second scenarios, the contributors have a potential role in validating the data classification by accepting or rejecting the recommendations.

We conduct empirical studies to check the validation of the proposed approach. We check the classification of administrative boundaries as an example of the strict hierarchical structure. Moreover, we analyze the classification of some grass-related features as an example of the loose classification structure. With more focus on the latter kind of classification, we apply machine learning methodologies to distinguish classes based on quantitative characteristics (e.g., area by m^2) and qualitative characteristics (e.g., topological characteristics). Figure 1.6 illustrates samples of the detected potential outliers.

(A) An entity is classified as "grass", however the entity is surrounded by residential houses and contains amusement facilities. It can be classified appropriately as "garden".



(B) An entity is classified as "park", while the entity does not include any amusement characteristics. Thus, it is recommended to be classified generally as "grass".

FIGURE 1.6: Samples of the potential outliers detected by the proposed approach.

The presented examples illustrate how the qualitative characteristics can be exploited to distinguish similar classes. In Figure 1.6a, the given entity is surrounded by houses, contains some amusement facilities, and is paved by footways. It was generally classified as "grass", while a more appropriate classification for this entity might be "garden". In contrast in Figure 1.6b, the given entity is relatively small, is surrounded by roundabouts, and contains no facilities. Therefore, it does not have the typical characteristics of "park", while it might be more appropriately classified as "grass". In Chapter 4, findings of an empirical study indicate participants' agreement on the detected outliers. Besides, they show disagreement of participants on absolute classification of the presented entities.

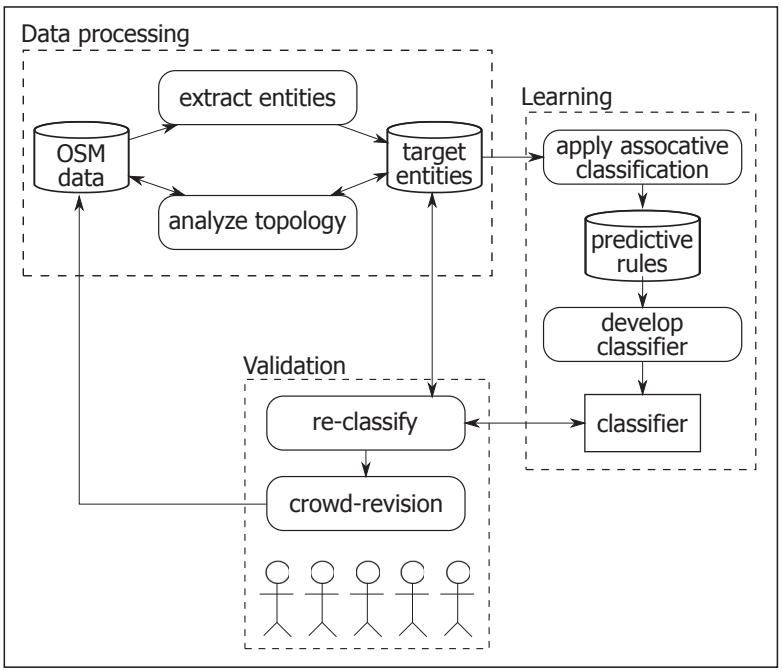


FIGURE 1.7: The proposed rule-guided classification approach.

Afterwards, the learning-based approach is refined to develop a guiding classification approach. In Chapter 5, the rule-based guided classification approach is proposed. Figure 1.7 illustrates the conceptual structure of the proposed approach. This approach aims to develop a guiding system (i.e., recommendation system) that presents the most appropriate classes of a given entity. The approach exclusively investigates the qualitative characteristics to distinguish between related classes. According to the figure, the approach consists of three phases: data processing, learning, and validation. The target entities of particular features are topologically checked with their context, to find out the distinct characteristics that identify each feature. During the learning phase, we applied the associative classification data mining technique. The extracted characteristics are encoded as a set of predictive rules. These rules are organized into the classifier and act to rank the potential classes of a given entity based on the matched rules. In this approach, the validation phase is needed to double check the presented recommendations.

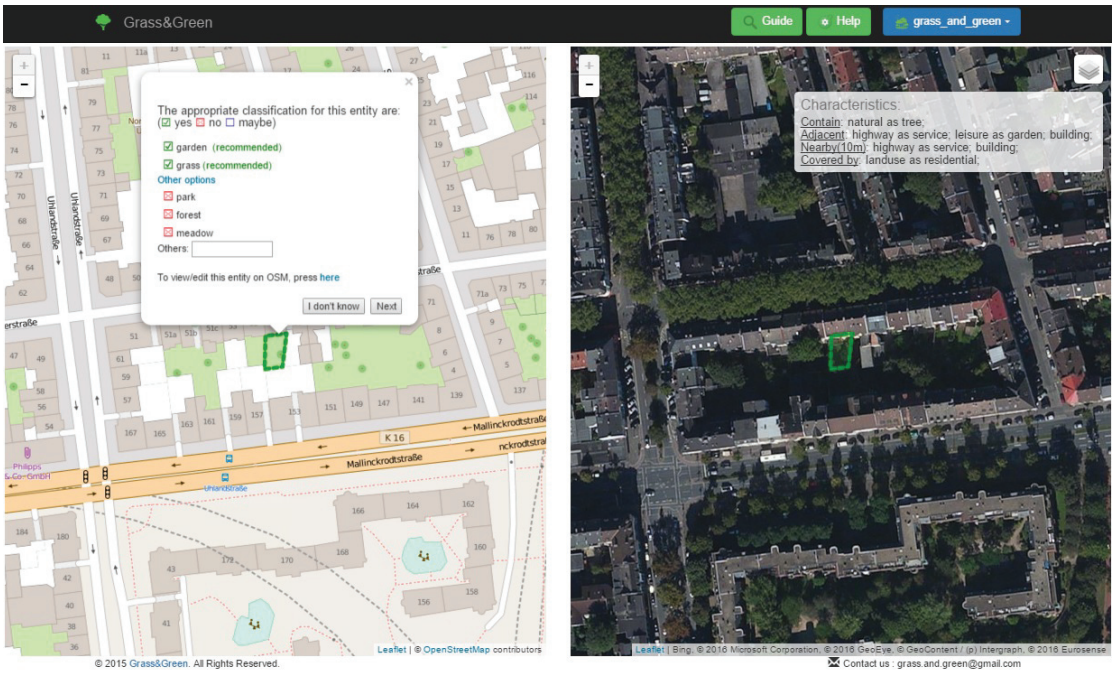


FIGURE 1.8: **Grass&Green**: a recommended classification application for some grass-related features.

We propose crowdsourcing validation to determine the applicability of the recommended classes. In Chapter 6, we exemplify the proposed approach on some grass-related classes. We utilize the OSM data set of Germany and tackle the classes "park", "garden", "meadow", "forest", and the public class of "grass". Although the classes may be conceptually overlapping, there exist fine details that distinguish each individual class. For example, the "park" class points to places, where people can have amusements and perform leisure activities (e.g., walking, jogging). The "garden" class might imply the same, but it is usually cultivated with plants. Otherwise, the woody plants and the context might distinguish between the classes "meadow" and "forest". We developed a web application for crowdsourcing validation. Figure 1.8 shows the general user interface of the developed application, which is called **Grass&Green** (http://opensciencemap.org/quality/).

The application presents a set of entities associated with their most appropriate classes. Afterwards, the crowds are invited to validate the proposed recommendation. In a duration of four months, about 90% of crowd participants agreed on the presented recommendation. The detailed analysis reveals the potential enhancement of data classification quality, when participants follow a specific guide line. The findings demonstrate the significance of the proposed approach and the feasibility of exploiting the qualitative characteristics to distinguish similar features. Participants encourage to apply the proposed approach on different classes and in different locations.

To summarize, the contributions of this dissertation are:

C1. Presenting a literature review about quality assurance procedures regarding different formats of VGI.

C2. Summarizing the data classification challenges in VGI mapping projects.

C3. Confirming the significance of learning from crowdsourcing data, under certain circumstances.

C4. Proposing a human-centered guided classification approach for VGI mapping projects.

C5. Developing an intuitive application to enhance the data classification quality of some grass-related features in the OSM project.

C6. Encouraging the role of crowdsourcing in the process of data collection as well as in the procedures of quality assurance.

1.5 Dissertation Outline and Formatting

Figure 1.9 illustrates organization of the dissertation with respect to the publications and the contributions.

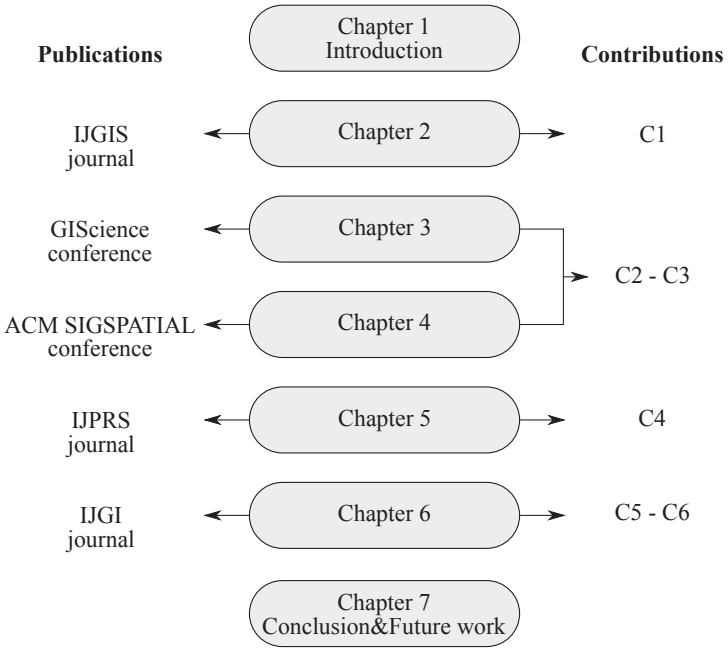


FIGURE 1.9: Dissertation outline with respect to contributions (right side) and the publications (left side).

Table 1.1 lists the publications and their status at the time of submitting this dissertation.

Chapter	Publication	Status
2	Hansi Senaratne, Amin Mobasheri, **Ahmed Loai Ali**, Cristina Capineri and Mordechai (Muki) Haklay (2016). "A Review of Volunteered Geographic Information Quality Assessment Methods". In: *International Journal of Geographical Information Science (IJGIS)*, pp. 1–29.	Published online
3	**Ahmed Loai Ali** and Falko Schmid (2014). "Data quality assurance for Volunteered Geographic Information". In: *Geographic Information Science: 8th International Conference, GIScience 2014, Vienna, Austria, September 24-26, 2014. Proceedings*. Ed. by Matt Duckham, Edzer Pebesma, Kathleen Stewart, and Andrew U. Frank. Cham: Springer International Publishing, pp. 126–141.	Published
4	**Ahmed Loai Ali**, Falko Schmid, Rami Al-Salman, and Tomi Kauppinen (2014). "Ambiguity and plausibility: managing classification quality in Volunteered Geographic Information". In: *Proc. of the 22nd ACM SIGSPATIAL International Conf. on Advances in Geographic Information Systems*. Ed. by Yan Huang, Markus Schneider, Michael Gertz, John Krumm, and Jagan Sankaranarayanan. New York, NY, USA: ACM, pp. 143–152.	Published
5	**Ahmed Loai Ali**, Zoe Falomir, Falko Schmid, and Christian Freksa (2016). "Rule-Guided Human Classification of Volunteered Geographic Information". In: *ISPRS Journal of Photogrammetry and Remote Sensing*. ISSN: 0924-2716.	Published online
6	**Ahmed Loai Ali**, Nuttha Sirilertworakul, Alexander Zipf, and Amin Mobasheri (2016). "Guided Classification System for Overlapping Classes in OpenStreetMap". In: *ISPRS International Journal of Geo-Information (IJGI)* 5.6, p. 87. ISSN: 2220-9964.	Published online

TABLE 1.1: List of accumulated publications and their status at the time of submitting this dissertation.

Formatting Consistency

To preserve a consistent structure of this dissertation, we adapted the original publications as follows:

- In Chapter 2,
  - Tables 2.1 and 2.2 are reoriented and rescaled to fit the document format.
  - The word "crowd-sourcing" is spelled "crowdsourcing", to be consistent within the entire document.
- In Chapters 4 and 5
  - The original publications are modified from two-column format to single-column format. In addition, the figures are rescaled to fit the modified format.

- In Chapter 6,
  - For consistency, the endnotes in the original publication are modified into footnotes.
  - Tables format are modified to be consistent with the entire dissertation.

Note:
Please, cite the original publications when referring to any content within Chapters 2 – 6.

Bibliography

Ahlqvist, O. (2012). "Semantic Issues in Land-Cover Analysis: representation, analysis, and visualization". In: *Remote Sensing of Land Use and Land Cover: principles and applications*. Ed. by C. P. Giri. CRC Press, pp. 25–35.

Arsanjani, J. J., P. Mooney, A. Zipf, and A. Schauss (2015). "Quality Assessment of the Contributed Land Use Information from OpenStreetMap Versus Authoritative Datasets". In: *OpenStreetMap in GIScience: experiences, research, and applications*. Ed. by J. J. Arsanjani, A. Zipf, P. Mooney, and M. Helbich. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 37–58.

Barron, C., P. Neis, and A. Zipf (2014). "A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis". In: *Transactions in GIS* 18.6, pp. 877–895.

Bennett, J. (2010). *OpenStreetMap*. Otlon, Birmingham, UK: Packt Publishing Ltd. Chap. 4, p. 61. ISBN: 978-1-847197-50-4.

Bishr, M. and W. Kuhn (2007). "Geospatial Information Bottom-Up: A Matter of Trust and Semantics". In: *The European Information Society: Leading the Way with Geo-information*. Ed. by S. I. Fabrikant and M. Wachowicz. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 365–387. ISBN: 978-3-540-72385-1.

Bishr, M. and W. Kuhn (2013). "Trust and Reputation Models for Quality Assessment of Human Sensor Observations". In: *Spatial Information Theory: 11th International Conference, COSIT 2013, Scarborough, UK, September 2-6, 2013. Proceedings*. Ed. by T. Tenbrink, J. Stell, A. Galton, and Z. Wood. Cham: Springer International Publishing, pp. 53–73. ISBN: 978-3-319-01790-7.

Castillo, C., M. Mendoza, and B. Poblete (2011). "Information Credibility on Twitter". In: *Proceedings of the 20th International Conference on World Wide Web*. WWW '11. New York, NY, USA: ACM, pp. 675–684. ISBN: 978-1-4503-0632-4.

Chow, T. E. (2013). "We Know Who You Are and We Know Where You Live: A Research Agenda for Web Demographics". In: *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*. Ed. by D. Sui, S. Elwood, and M. F. Goodchild. Dordrecht: Springer Netherlands, pp. 265–285. ISBN: 978-94-007-4587-2.

Chow, T. E., Y. Lin, N. T. Huynh, and J. M. Davis (2012). "Using web demographics to model population change of Vietnamese-Americans in Texas between 2000 and 2009". In: *GeoJournal* 77.1, pp. 119–134.

Cooper, A. K., P. Rapant, J. Hjelmager, D. Laurent, A. Iwaniak, S. Coetzee, H. Moellering, and U. Düren (2011). "Extending the formal model of a spatial data infrastructure to include Volunteered Geographical Information". In: *Proc. of the 25th International Cartographic Conference*. Ed. by A. Ruas. ICC 2011. Paris, France.

Cowen, D (2008). "The availability of geographic data: the current technical and institutional environment". In: *The handbook of geographic information science*. Ed. by J. P. Wilson and A. S. Fotheringham. Wiley-Blackwell, pp. 11–34.

D'Antonio, F., P. Fogliaroni, and T. Kauppinen (2014). "VGI Edit History Reveals Data Trustworthiness and User Reputation". In: *Connecting a Digital Europe Through Location and Place*. Ed. by J. Huerta, S. Schade, and C. Granell. Springer-Verlag.

DeMers, M. N. (2009). *Fundamentals of Geographic Information Systems*. 4th. USA: John Wiley & Sons. Chap. 1, p. 20. ISBN: 978-0-470-12906-7.

Dorn, H., T. Törnros, and A. Zipf (2015). "Quality Evaluation of VGI Using Authoritative Data–A Comparison with Land Use Data in Southern Germany". In: *ISPRS International Journal of Geo-Information* 4.3, pp. 1657–1671.

Flanagin, A. J. and M. J. Metzger (2008). "The credibility of Volunteered Geographic Information". In: *GeoJournal* 72.3, pp. 137–148.

Foody, G. M., L. See, S. Fritz, M. van der Velde, C. Perger, C. Schill, D. S. Boyd, and A. Comber (2015). "Accurate Attribute Mapping from Volunteered Geographic Information: Issues of Volunteer Quantity and Quality". In: *The Cartographic Journal* 52.4, pp. 336–344.

Foth, M., B. Bajracharya, R. Brown, and G. Hearn (2009). "The Second Life of urban planning? Using NeoGeography tools for community engagement". In: *Journal of Location Based Services* 3.2, pp. 97–117.

Fritz, S., I. McCallum, C. Schill, C. Perger, L. See, D. Schepaschenko, M. Van der Velde, F. Kraxner, and M. Obersteiner (2012). "Geo-Wiki: An online platform for improving global land cover". In: *Environmental Modelling & Software* 31, pp. 110–123.

Goodchild, M. F. (2007). "Citizens as sensors: the world of volunteered geography". In: *GeoJournal* 69.4, pp. 211–221.

Goodchild, M. F. and J. A. Glennon (2010). "Crowdsourcing geographic information for disaster response: a research frontier". In: *International Journal of Digital Earth* 3.3, pp. 231–241.

Goodchild, M. F. and L. Li (2012). "Assuring the quality of Volunteered Geographic Information". In: *Spatial statistics* 1, pp. 110–120.

Gouveia, C. and A. Fonseca (2008). "New approaches to environmental monitoring: the use of ICT to explore Volunteered Geographic Information". In: *GeoJournal* 72.3-4, pp. 185–197.

Guptill, S. C. and J. L. Morrison, eds. (2013). *Elements of Spatial Data Quality*. Oxford, UK: Elsevier Science.

Hagenauer, J. and M. Helbich (2012). "Mining urban land-use patterns from Volunteered Geographic Information by means of genetic algorithms and artificial neural networks". In: *International Journal of Geographical Information Science* 26.6, pp. 963–982.

Haklay, M. (2010). "How good is Volunteered Geographic Information? A comparative study of OpenStreetMap and Ordnance Survey datasets". In: *Environment and planning. B Planning & design* 37.4, p. 682.

Haklay, M. and P. Weber (2008). "OpenStreetMap: user-generated street maps". In: *IEEE Pervasive Computing* 7.4, pp. 12–18. ISSN: 1536-1268.

Heipke, C. (2010). "Crowdsourcing geospatial data". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 65.6, pp. 550–557.

Howe, J. (2006). "The rise of crowdsourcing". In: *Wired magazine* 14.6, pp. 1–4.

Huang, K. L., S. S. Kanhere, and W. Hu (2010). "Are You Contributing Trustworthy Data?: the case for a reputation system in participatory sensing". In: *Geospatial Thinking*. MSWIM '10. New York, NY, USA: ACM, pp. 14–22. ISBN: 978-1-4503-0274-6.

ISO (2009). *Standards Guide ISO/TC 211 Geographic Information/Geomatics*. A manual guide. Online published by International Organization for Standardization (ISO).

Keßler, C. and R. T. A. de Groot (2013). "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*. Ed. by D. Vandenbroucke, B. Bucher, and J. Crompvoets. Springer-Verlag, pp. 21–37.

Ludwig, I., A. Voss, and M. Krause-Traudes (2011). "A Comparison of the Street Networks of Navteq and OSM in Germany". In: *Advancing Geoinformation Science for a Changing World*. Ed. by W. Geertman Stanand Reinhardt and F. Toppen. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 65–84. ISBN: 978-3-642-19789-5.

Mac Gillavry, E. (2006). "Collaborative mapping and GIS: an alternative geographic information framework". In: *Collaborative Geographic Information Systems*. Ed. by S. Balram and S. Dragicevic. Hershey, PA, USA: Idea Group Publishing, pp. 103–119.

McDougall, K. (2009). "The potential of citizen volunteered spatial information for building SDI". In: *Proc. of 11th World Conference on Spatial Data Infrastructure Convergence: Building SDI Bridges to Address Global Challenges*. Rotterdam, The Netherlands: GSDI Association Press.

Mooney, P. and P. Corcoran (2012b). "The annotation process in OpenStreetMap". In: *Transactions in GIS* 16.4, pp. 561–579.

Mooney, P., P. Corcoran, and A. Winstanley (2010). "A study of data representation of natural features in OpenStreetMap". In: *Proceedings of the 6th GIScience International Conference on Geographic Information Science, Zurich, Switzerland, September 14-17, 2010*. Vol. 150, pp. 150–156.

O'Reilly, T. (2009). *What is Web 2.0*. O'Reilly Media, Inc.

Østensen, O. M. and P. C. Smits (2002). "ISO/TC211: Standardisation of geographic information and geo-informatics". In: *Geoscience and Remote Sensing Symposium, 2002. IGARSS'02. 2002 IEEE International*. Vol. 1. IEEE, pp. 261–263.

Raymond, E. (1999). "The cathedral and the bazaar". In: *Knowledge, Technology & Policy* 12.3, pp. 23–49.

Roche, S., E. Propeck-Zimmermann, and B. Mericskay (2013). "GeoWeb and crisis management: Issues and perspectives of Volunteered Geographic Information". In: *GeoJournal* 78.1, pp. 21–40.

Savelyev, A., S. Xu, K. Janowicz, C. Mülligann, J. Thatcher, and W. Luo (2011). "Volunteered geographic services: developing a linked data driven location-based service". In: *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Spatial Semantics and Ontologies*. ACM, pp. 25–31.

See, L., P. Mooney, G. Foody, L. Bastin, A. Comber, J. Estima, S. Fritz, N. Kerle, B. Jiang, M. Laakso, et al. (2016). "Crowdsourcing, Citizen Science or Volunteered Geographic Information? The Current State of Crowdsourced Geographic Information". In: *ISPRS International Journal of Geo-Information* 5.5, p. 55.

Shi, W., P. Fisher, and M. F. Goodchild, eds. (2003). *Spatial data quality*. CRC Press.

Sparks, K., A. Klippel, J. O. Wallgrün, and D. Mark (2015). "Citizen Science Land Cover Classification Based on Ground and Aerial Imagery". In: *Spatial Information Theory: 12th International Conference, COSIT 2015, Santa Fe, NM, USA, October 12-16, 2015, Proceedings*. Ed. by I. S. Fabrikant, M. Raubal, M. Bertolotto, C. Davies, S. Freundschuh, and S. Bell. Cham: Springer International Publishing, pp. 289–305. ISBN: 978-3-319-23374-1.

Surowiecki, J. (2005). *The wisdom of crowds*. US: Anchor.

Thatcher, J. (2013). "From Volunteered Geographic Information to Volunteered Geographic Services". In: *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*. Ed. by D. Sui, S. Elwood, and M. F. Goodchild. Dordrecht: Springer Netherlands, pp. 161–173. ISBN: 978-94-007-4587-2.

Tobler, W. R. (1970). "A computer movie simulating urban growth in the Detroit region". In: *Economic geography* 46, pp. 234–240.

Vaz, E. and J. Jokar Arsanjani (2015). "Crowdsourced mapping of land use in urban dense environments: An assessment of Toronto". In: *The Canadian Geographer/Le Géographe canadien* 59.2, pp. 246–255.

Zook, M., M. Graham, T. Shelton, and S. Gorman (2010). "Volunteered Geographic Information and crowdsourcing disaster relief: a case study of the Haitian earthquake". In: *World Medical & Health Policy* 2.2, pp. 7–33. ISSN: 1948-4682.

Chapter 2

A Review of Volunteered Geographic Information Quality Assessment Methods

Authors:

Hansi Senaratne, Amin Mobasheri, **Ahmed Loai Ali**, Cristina Capineri and Mordechai (Muki) Haklay.

Journal:

International Journal of Geographical Information Science (IJGIS).

Citation:

Hansi Senaratne, Amin Mobasheri, Ahmed Loai Ali, Cristina Capineri and Mordechai (Muki) Haklay (2016). "A Review of Volunteered Geographic Information Quality Assessment Methods". In: *International Journal of Geographical Information Science*. pp. 1–29.

Status:

This article is published and available online since 31 May 2016.

Contribution Statement:

This is collaborative work in the period from October 2014 to April 2016. Hansi was concerned with image-based and text-based VGI. Amin and me were responsible for the map-based VGI. I contributed in surveying various quality assurance procedures related to map-based VGI. I was involved strongly in classifying the quality assurance procedures as well as in writing the manuscript and in responding to reviewers' comments. Muki and Christina contributed in discussing and reviewing the manuscript.

Abstract:

With the ubiquity of advanced web technologies and location-sensing hand held devices, citizens regardless of their knowledge or expertise, are able to produce spatial information. The phenomena is known as Volunteered Geographic Information (VGI). During the last decade VGI has been used as a data source supporting a wide range of services such as environmental monitoring, events reporting, human movement analysis, disaster management etc. However, these volunteer contributed data also come with varying *quality*. Reasons for this are: data is produced by heterogeneous contributors, using various technologies and tools, having different level of details and precision, serving heterogeneous purposes, and a lack of gatekeepers. Crowdsourcing, social, and geographic approaches have been proposed and later followed to develop appropriate methods to assess the quality measures and indicators of VGI. In this paper, we review various quality measures and indicators for selected types of VGI, and existing quality assessment methods. As an outcome, the paper presents a classification of VGI with current methods utilized to assess the quality of selected types of VGI. Through these findings we introduce data mining as an additional approach for quality handling in VGI.

Keywords:

Geographic Information Systems; Volunteered Geographic Information; Spatial Data Quality; Spatial Data Applications.

2.1 Introduction

Volunteered Geographic Information (VGI) is where citizens, often untrained, and regardless of their expertise and background create geographic information on dedicated web platforms (Goodchild, 2007), e.g., OpenStreetMap[1] (OSM), Wikimapia[2], Google MyMaps[3], Map Insight[4] and Flickr[5]. In a typology of VGI, the works of Antoniou et al. (2010) and Craglia et al. (2012) classified VGI based on the type of explicit/implicit geography being captured and the type of explicit/implicit volunteering. In explicit-VGI, contributors are mainly focused on mapping activities. Thus, the contributor explicitly annotates the data with geographic contents (e.g., geometries in OSM, Wikimapia, or Google). Data that is implicitly associated with a geographic location could be any kind of media: text, image, or video referring to or associated with a specific geographic location. For example, geo-tagged microblogs (e.g., Tweets), geo-tagged images from Flicker,

[1] http://www.openstreetmap.org
[2] http://www.wikimapia.org
[3] https://www.google.com/maps/mm
[4] http://www.mapsharetool.com/external-iframe/external.jsp
[5] http://www.flickr.com

or Wikipedia articles that refer to geographic locations. Craglia et al. (2012) further elaborated that for each type of implicit/explicit geography and volunteering there are potentially different approaches for assessing the *quality*.

Due to the increased potential and use of VGI (as demonstrated in the works of Chunara et al. (2012), Sakaki et al. (2010), Fuchs et al. (2013), MacEachren et al. (2011), Liu et al. (2008), McDougall (2009), Bulearca and Bulearca (2010), and Jacob et al. (2009)), it becomes increasingly important to be aware of the quality of VGI, in order to derive accurate information and decisions. Due to a lack of standardization, quality in VGI has shown to vary across heterogeneous data sources (text, image, maps etc.). For example, as seen in Figure 2.1 a photo of the famous tourist site the Brandenburg Gate in Berlin is incorrectly geo-tagged in Jakarta, Indonesia on the photo sharing platform Flickr. On the other hand OSM has also shown heterogeneity in coverage between different places (Haklay, 2010). These trigger a variable quality in VGI. This can be explained by the fact that humans perceive and express geographic regions and spatial relations imprecisely, and in terms of vague concepts (Montello et al., 2003). This vagueness in human conceptualization of location is due not only to the fact that geographic entities are continuous in nature, but also due to the quality and limitations of spatial knowledge (Hollenstein and Purves, 2014).

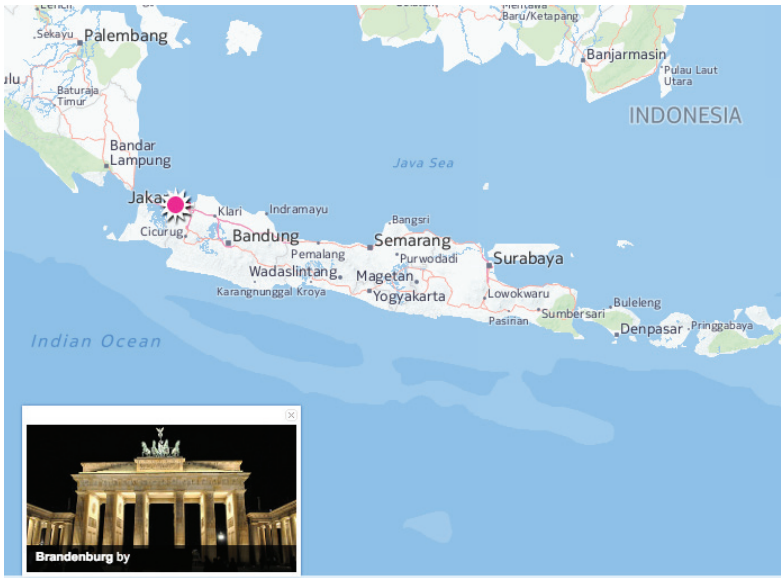


FIGURE 2.1: A photo of the Brandenburg Gate in Berlin is incorrectly geotagged in Jakarta, Indonesia on the popular photo sharing platform Flickr.

Providing reliable services or extraction of useful information require data with a fitness-for-use quality standard. Incorrect (as seen in Figure 2.1) or malicious geographic annotations could be minimized in place of appropriate quality indicators and measures for these various VGI contributions.

Goodchild and Li (2012) have discussed three approaches for assuring the quality of VGI: crowdsourcing (the involvement of a group to validate and correct errors that have been made by an individual contributor), social approaches (trusted individuals who have made themselves a good reputation with their contributions to VGI can for example act as gatekeepers to maintain and control the quality of other VGI contributions), and geographic approaches (use of laws and knowledge from geography, such as Tobler's first law to assess the quality). Many works have developed methods to asses the quality of VGI based on these approaches.

In this paper we present an extensive review of the existing methods in the state-of-the-art to assess the quality of map, image, and text based VGI. As an outcome of the review we identify *data mining* as one more stand alone approach to assess VGI quality by utilizing computational processes for discovering patterns and learning purely from data, irrespective of the laws and knowledge from geography, and independent from social or crowd-sourced approaches. Extending the spectrum of approaches will sprout more quality assessment methods in the future, especially for VGI types that have not been extensively researched so far. To the best of our knowledge surveys on existing methods have not been done so far. This review provides an overview of methods that have been built based on theories and discussions in the literature. Furthermore, this survey gives the reader a glimpse to the practical applicability of all identified approaches. The remainder of this paper unfolds as follows: In Section 2.2, we describe the different quality measures and indicators for VGI. In Section 2.3, we describe the main types of VGI that we consider for our survey, and in Section 2.4, we describe the methodology that was followed for the selection of literature for this survey. Section 2.5 summarizes the findings of the survey, and Section 2.6 discusses the limitations and future research perspectives. Lastly we conclude our findings in Section 2.7.

2.2 Measures and Indicators for VGI Quality

Quality of VGI can be described by quality *measures* and quality *indicators* (Antoniou and Skopeliti, 2015). Quality measures, mainly adhering to the ISO principles and guidelines refer to those elements that can be used to ascertain the discrepancy between the contributed spatial data and the ground truth (e.g., completeness of data) mainly by comparing to authoritative data. When authoritative data is no longer usable for comparisons, and the established measures become no longer adequate to assess the quality of VGI, researchers have explored more intrinsic ways to assess VGI quality by looking into other proxies for quality measures. These are called quality indicators, that rely on various participation biases, contributor expertise or the lack of it, background, etc., that influence the quality of VGI, but cannot be directly measured (Antoniou and

Skopeliti, 2015). In the following these quality measures and indicators are described in detail. The review of quality assessment methods in Section 2.5 is based on these various quality measures and indicators.

2.2.1 Quality measures for VGI

ISO[6] (International Standardization Organization) defined geographic information quality as *totality of characteristics of a product that bear on its ability to satisfy stated and implied needs.* ISO/TC 211[7] (Technical Committee) developed a set of international standards that define the measures of geographic information quality (standard 19138, as part of the meatadata standard 19115). These quantitative quality measures are: completeness, consistency, positional accuracy, temporal accuracy and thematic accuracy.

Completeness describes the relationship between the represented objects and their conceptualizations. This can be measured as the absence of data (errors of omission) and presence of excess data (errors of commission). Consistency is the coherence in the data structures of the digitized spatial data. The errors resulting from the lack of it are indicated by (i) conceptual consistency, (ii) domain consistency, (iii) format consistency, and (iv) topological consistency. Accuracy refers to the degree of closeness between a measurement of a quantity and the accepted true value of that quantity, and it is in the form of positional accuracy, temporal accuracy and thematic accuracy. Positional accuracy is indicated by (i) absolute or external accuracy, (ii) relative or internal accuracy, (iii) gridded data position accuracy. Thematic accuracy is indicated by (i) classification correctness, (ii) non-quantitative attribute correctness, (iii) quantitative attribute accuracy. In both cases, the discrepancies can be numerically estimated. Temporal accuracy is indicated by (i) accuracy of a time measurement: correctness of the temporal references of an item, (ii) temporal consistency: correctness of ordered events or sequences, (iii) temporal validity: validity of data with regard to time.

2.2.2 Quality indicators for VGI

As part of the ISO standards, geographic information quality can be further assessed through qualitative quality indicators such as the purpose, usage, and lineage. These indicators are mainly used to express the quality overview for the data. Purpose describes the intended usage of the dataset. Usage describes the application(s) in which the dataset has been utilized. Lineage describes the history of a dataset from collection,

[6] http://www.iso.org/iso/home/standards.htm
[7] http://www.isotc211.org/

acquisition to compilation and derivation to its form at the time of use (Van Oort and Bregt, 2005; Hoyle, 2001; Guinée, 2002). In addition, where ISO standardised measures and indicators are not applicable, we have found in the literature more abstract quality indicators to imply the quality of VGI. These are: trustworthiness, credibility, text content quality, vagueness, local knowledge, experience, recognition, reputation. Trustworthiness is a receiver judgment based on subjective characteristics such as reliability or trust (good ratings on the creations, and the higher frequency of usage of these creations indicate this trustworthiness) (Flanagin and Metzger, 2008). In assessing the credibility of VGI, the source of information plays a crucial role, as it is what credibility is primarily based upon. However, this is not straight forward. Due to the non-authoritative nature of VGI, the source maybe unavailable, concealed, or missing (this is avoided by gatekeepers in authoritative data). Credibility was defined by Hovland et al. (1953) as the *believability of a source or message, which comprises primarily two dimensions, the trustworthiness (as explained above), and expertise.* Expertise contains objective characteristics such as accuracy, authority, competence, or source credentials (Flanagin and Metzger, 2008). Therefore, in assessing the credibility of data as a quality indicator one needs to consider factors that attribute to the trustworthiness and expertise. Metadata about the origin of VGI can provide a foundation for the source credentials of VGI (Frew, 2007). Text content quality (mostly applicable for text-based VGI) describes the quality of text data by the use of text features such as the text length, structure, style, readability, revision history, topical similarity, the use of technical terminology etc. Vagueness is the ambiguity with which the data is captured (e.g., vagueness caused by low resolutions) (De Longueville et al., 2010). Local knowledge is the contributors' familiarity to the geographic surroundings that she/he is implicitly or explicitly mapping. Experience is the involvement of a contributor with the VGI platform that she/he contributes to. This can be expressed by the time that the contributor has been registered with the VGI portal, number of GPS tracks contributed (for example in OSM) or the number of features added and edited, or the amount of participation in online forums to discuss the data (Van Exel et al., 2010). Recognition is the acknowledgement given to a contributor based on tokens achieved (for example in gamified VGI platforms), and the reviewing of their contributions among their peers (Van Exel et al., 2010). Maué (2007) described reputation as a tool to ensure the validity of VGI. Reputation is said to be assessed by, for example the history of past interactions that are happening between collaborators. Resnick et al. (2000) described contributors' abilities and dispositions as features where this reputation can be based upon. Maué (2007) further argue that similar to the eBay rating system[8], the created geographic features on various VGI platforms can be rated, tagged, discussed, and annotated, which affects the data contributor's reputation value.

[8] http://ebay.about.com/od/gettingstarted/a/gs_feed.htm

2.3 Map, Image, and Text based VGI: Definitions and Quality Issues

The effective utilization of VGI is strongly associated with data quality, and this varies depending primarily on the type of VGI, the way data is collected on the different VGI platforms, and the context of usage. The following sections describe the selected forms of VGI: 1) *map*, 2) *image*, and 3)*text*, their uses, and how data quality issues arise. These three types of VGI are chosen based on the methods that are used to capture the data (maps: as gps points and traces, image: as photos, text: as plain text), and because they are the most popular forms of VGI currently used. This section further lays the ground work to understand the subsequent section on various quality measures and indicators, and quality assessment methods used for these three types of VGI.

2.3.1 Map-based VGI

Map-based VGI concerns all VGI sources that include geometries as points, lines and polygons, the basic elements to design a map. Among others, OSM, Wikimapia, Google Map Maker, and Map Insight are examples of map-based VGI projects. However, OSM is the most prominent project due to the following reasons: (i) It aims to develop a free map of the world accessible and obtainable for everyone; (ii) It has millions of registered contributors; (iii) It has active mapper communities in many locations; and (iv) It provides free and flexible contribution mechanisms for data (useful for map provision, routing, planning, geo-visualization, point of interests (POI) search etc.). Thus, during the rest of the article we will discuss OSM as an example for map-based VGI. As in most VGI projects, the spatial dimension of OSM data is annotated in the form of nodes, lines, or polygons with latitude/longitude referencing, and attributes are annotated by tags in the form of key-value pairs. Each tag describes a specific geographic entity from different perspectives. There are no restrictions to the usage of these tags: endless combinations are possible, and the contributors are free to choose the tags they deem appropriate. Nevertheless, OSM provides a set of recommendations of accepted key-value pairs, and if the contributors want their contributions to become a part of the map, they need to follow the agreed-upon standards. This open classification scheme can lead to misclassification and reduction in data quality. Map-based VGI is commonly used for purposes like navigation and POI search. For these purposes the positional accuracy and the topological consistency of the entities are as important as their abstract locations. The other dimension is the attribute accuracy, where the annotations associated with an entity should reflect its characteristics without conflicts (e.g., for road tags, $oneway = true$ and $twoway = true$). In OSM, the loose contribution mechanisms result in problematic classifications that influence the attribute accuracy. In addition to

accuracy, providing reliable services is affected by data completeness; features, attribute, and model completeness. Whether a map includes all the required features, whether a feature is annotated with a complete set of attributes, and if the model is able to answer all possible queries, all these points are related to the completeness quality measure. Especially due to the lack of ground-truth data for comparison, assessing VGI completeness still raises some challenges.

2.3.2 Image-based VGI

Image-based VGI is mostly produced implicitly within portals such as Flickr, Panoramio, Instagram etc., where contributors take pictures of a particular geographic object or surrounding with cameras, smart phones, or any hand held device, and attach a geospatial reference to it. These objects/surroundings can be spatially referenced either by giving geographic coordinates and/or user-assigned geospatial descriptions of these photographs in the form of textual labels. These photo sharing websites have several uses such as environmental monitoring (Fuchs et al., 2013), pedestrian navigation (Robinson et al., 2012), event and human trajectory analysis (Andrienko et al., 2009), for creating geographical gazetteers (Popescu et al., 2008), or even to complement institutional data sources in your locality (Milholland and Pultar, 2013).

Tagging an image is a means of adding metadata to the content in the form of specific keywords to describe the content (Golder and Huberman, 2006), or in the form of geographic coordinates (Geotagging) to identify the location linked to the image content (Valli and Hannay, 2010). There exist several approaches to geotag an image: record the geographic location with the use of an external GPS device, with an in-built GPS (in many of the modern digital cameras, smart phones), or manually positioning the photo on a map interface.

Not only the GPS precision and accuracy errors resulting from various devices, but also other factors influence the quality of image-based VGI. For example, instead of stating the position from where the photo was taken (photographer position) some contributors tend to geotag the photo with the position of the photo content, which could be several kilometers away from where the photo originated causing positional accuracy issues (as also discussed in Keßler et al. (2009)). This is a problem when we want to utilize these photos for example in human trajectory analysis. Furthermore, due to the lack of sufficient spatial knowledge contributors sometimes incorrectly geotag their photographs (Figure 2.1), also in lower geographic resolutions (in case of Flickr, some contributors do not zoom enough to the street level, instead they zoom up to country or city level to geotag their photos). Or some contributors geotag and textually label random irrelevant photos for actual events, causing the users to doubt the trustworthiness of the content.

Such content are not fit for use for tasks such as disaster management, environmental monitoring, or pedestrian navigation. Citizen Science Projects such as GeoTag-X[9] have in place machine learning and crowdsourcing methods to discover unauthentic material and clean them.

2.3.3 Text-based VGI

Text-based VGI (typically microblogs) is mostly produced implicitly on portals such as Twitter, Reddit or various Blogs, where people contribute geographic information in the form of text by using smart phones, PCs, or any hand held devices. Twitter for example is used as an information foraging source (MacEachren et al., 2011), in journalism to disseminate data to the public in near real-time basis (O'Connor, 2009; Castillo et al., 2011), detect disease spreading (Chunara et al., 2012), event detection (Bosch et al., 2013), and for gaining insights on social interaction behavior (Huberman et al., 2008) or trajectories of people (Andrienko et al., 2013; Senaratne et al., 2014).

In text-based VGI, the spatial reference can be either in the text, where the contributor refers to a place-name (e.g., 'Lady Gaga is performing in New York today'), or the spatial reference can be the geotag where the tweet is originating from. While some people contribute meaningful information most others use these mediums to express personal opinions, moods, or for malicious aims such as bullying or trolling to harass other users. Gupta and Kumaraguru (2012) conducted a study to investigate how much information is credible and therefore useful, and how much information is spam, on Twitter. They found that 14% of Tweets collected for event analysis were spam, while 30% of the Tweets contained situational awareness information, out of which only 17% of the total tweets contained credible situational awareness information. Such spam makes it difficult to derive useful information that could be of interest for the above named use-cases. Therefore quality analysis of these data is important to filter out the useful information, and disregard the rest. Other than the inherent GPS errors in devices, a bigger role for quality issues is played by the contributor herself/himself based on the information she/he provides. Also due to the lack of spatial knowledge of some contributors the location is incorrectly specified, and at times at a low resolution (in the Twitter interface on PCs the contributor can specify the location not only at the city level, but also at a more coarse state level). Sometimes if the contributor is writing about an event that takes place a few hundred kilometers away from her position, she would geotag her content with the location of the event rather than her position; Or the other way around. A summary of quality assessment methods for these VGI types is presented in Section 2.5.

[9] http://geotagx.org/

2.4 The Literature Review Methodology

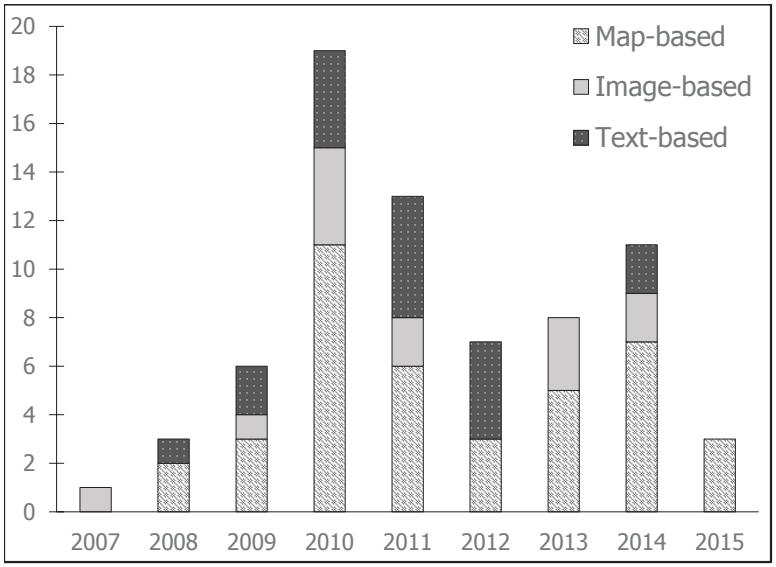


FIGURE 2.2: Distribution of the surveyed papers.

This review provides an overview of the state-of-the-art methods to assess the quality of selected types of VGI. To achieve this goal we breakdown our review in to three categories. Firstly, we show how the topic of quality assessment within map, image, and text VGI has evolved over the years since the birth of VGI in 2007 until the time of writing this article (mid of 2015). Secondly, the reviewed papers are classified according to the type of quality measure or indicator that is assessed within each of the papers. Thirdly, all the quality measures and indicators that are addressed within each of the reviewed papers are classified with the different methods utilized to assess them.

We used the following strategy to select the literature for our review. We used Google Scholar to search for papers that include the following terms in their title or abstract: *data quality assessment*, *methods and techniques*, *uncertainty*, *volunteered geographic information*, *map*, *microblog*, *photo*. This query resulted in 425 research papers. We sorted the search results according to the Google Scholar relevance ranking[10]. This relevance ranking follows a combined ranking algorithm that contains a weighting for the full text of each article, author of article, publisher, and how often the article has been cited in other scholarly articles. We refined our collection of papers by filtering out the papers according to the following criteria: (1) papers were published from 2007; (2) papers should describe quality assessment methods, or techniques, or tools; and (3) a latest paper was selected when multiple versions of similar methods were available from the same research group. Citizen Science research studies are not considered in this review. As such, we selected 56 papers in total: out of which 33 of them discuss quality assurance methods

[10] https://scholar.google.com/scholar/about.html

for map-based VGI, 10 on text-based VGI, 6 on image-based VGI, and 8 on all three types of VGI.

Figure 2.2 shows the distribution of the reviewed papers for VGI quality assessment methods. Evidently, the publication of papers on this topic gained momentum in 2010, for the most part papers discuss methods for map-based VGI.

2.5 Existing Methods for Assessing the Quality of VGI

We have reviewed state-of-the-art methods to assess various quality measures and indicators of VGI. Within this review, a method is considered to be a systematic procedure that is followed to assess the quality measures and quality indicators. For example, comparing with satellite imagery is a method to assess the positional accuracy of maps. The found methods have been mostly conceptually implemented for a particular usecase. These methods have been reviewed mainly based on the type of VGI, the quality measures and indicators supported, and the approaches followed to develop the method.

2.5.1 Distribution of selected literature

Out of the 56 papers that we reviewed, 40 papers discuss methods on assessing the quality of map-based VGI, in most cases taking OSM data as the VGI source. 18 papers introduce methods for text-based VGI taking mainly Twitter, Wikipedia, and Yahoo! answers as the VGI source. 13 papers introduce methods for image-based VGI taking Flickr and Panoramio as their VGI source. In reference to Craglia et al. (2012) typology of VGI with the reviewed papers, most quality assessment work is done on explicit VGI and lesser amount of work is done on implicit VGI, although implicit VGI due to its very nature has more concerns regarding its quality.

2.5.2 Type of quality measures, indicators, and their associated methods

We have found 17 quality measures and indicators (7 measures and 10 indicators) that are addressed within the 56 papers we surveyed. In Table 2.1 we have classified these surveyed papers according to the type of quality measures and indicators and the type of VGI. We found that papers particularly focusing on map-based VGI are clearly using only ISO standardized measures for quality assessment, whereas text-based VGI have been assessed only on the credibility, text content quality, and vagueness. Image-based VGI have been assessed in several papers on the positional/thematic accuracy, credibility, vagueness, experience, recognition, and reputation.

	Quality measures and indicators																
Papers	Positional accuracy	Thematic accuracy	Topological consistency	Completness	Temporal accuracy	Geometric accuracy	Semantic accuracy	Lineage	Usage	Credibility	Trust	Content quality	Vaguness	Local Knowledge	Experince	Recognition	Reputation
Jacobs et al. (2007)	•																
Agichtein et al. (2008)												◇					
Schmitz et al. (2008)			⋆														
Mummidi and Krumm (2008)		⋆															
Hasan Dalip et al. (2009)												◇					
Kounadi (2009)	⋆																
Ather (2009)	⋆			⋆													
De Longueville et al. (2010)													⋈				
Bishr and Janowicz (2010)											⋈						
Mendoza et al. (2010)										◇							
Haklay (2010)	⋆			⋆													
Ciepłuch et al. (2010)	⋆			⋆													
Corcoran et al. (2010)			⋆														
Girres and Touya (2010)		⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆								
Haklay et al. (2010)	⋆																
Poser and Dransch (2010)		⋈															
Brando and Bucher (2010)	⋈	⋈	⋈	⋈	⋈												
Huang et al. (2010)											⋈						
De Tré et al. (2010)	⋆	⋆															
Al-Bakri and Fairbairn (2010)	⋆																
Van Exel et al. (2010)															⋈	⋈	⋈
Ciepłuch et al. (2011)			⋆														
Neis et al. (2011)			⋆														
Codescu et al. (2011)		⋆															
Castillo et al. (2011)										◇							
Becker et al. (2011)												◇					
Canini et al. (2011)										◇							
Ostermann and Spinsanti (2011)										⋈							
Keßler et al. (2011)								⋆									

O'Donovan et al. (2012)										◇							
Kang et al. (2012)										◇							
Gupta and Kumaraguru (2012)										◇							
Morris et al. (2012)										◇							
Helbich et al. (2012)	⋆																
Mooney and Corcoran (2012b)		⋆															
Koukoletsos et al. (2012)				⋆													
Keßler and Groot (2013)		⋆	⋆	⋆													
Senaratne et al. (2013)	•									•							
Zielstra and Hochmair (2013)	•																
Canavosio-Zuzelski et al. (2013)	⋆																
Hecht et al. (2013)				⋆													
Vandecasteele and Devillers (2013)		⋆															
Jackson et al. (2013)	⋆			⋆													
Foody et al. (2015)		•															
Barron et al. (2014)			⋆	⋆													
Siebritz (2014)			⋆														
Wang et al. (2014)			⋆														
Fan et al. (2014)	⋆			⋆													
Tenney (2014)	⋆			⋆													
Ali et al. (2014)		⋆															
Bordogna et al. (2014)													•◇		•	•	•
Forghani and Delavar (2014)			⋈														
Hollenstein and Purves (2014)	•																
Arsanjani et al. (2015)		⋆															
Vandecasteele and Devillers (2015)							⋆										
Hashemi and Abbaspour (2015)			⋆														

TABLE 2.1: Classification of the reviewed papers according to the quality measures and indicators. ⋆ = map-based, • = image-based, and ◇ = text-based VGI. While ⋈ = all types of VGI.

	Type of approaches and methods																																
	Geographic													Social								Crowdsourcing	Data mining										
	Compare with reference data	Line of sight	Formal specifications	Semantic consistency check	Geometrical analysis	Intrinsic data check	Integrity constraints	Automatic tag recommendation	Geographic proximity	Time between observation	Automatic scale capturing	Geographic familiarity	Manual inspection	Manual inspection/annotation	Manual annotation	Compare with previous evaluation	Linguistic decision making	Meta-data analysis	Tokens achieved, peer reviewing	Applying Linus law	Possibillstic truth value	Cluster analysis	Latent class analysis	Correlation statistics	Automatic detection of outliers	Regression analysis	Supervised classification	Feature classification	Provenance vocabulary	Heuristic metrics/fuzzy logic
Positional accuracy	⋈	•	⋈										•	•						⋆	⋆											
Thematic accuracy	⋈		⋈	⋆											⋆							⋆	•						⋆			
Topological consistency	⋈		⋈	⋆	⋆	⋆	⋆																						⋆	⋈		
Completness	⋈		⋈			⋆		⋆																					⋆			
Temporal accuracy																							⋆									
Geometric accuracy	⋆																															
Semantic accuracy	⋆							⋆																								
Lineage													⋆																⋆			
Usage																⋆																
Credibility	⋈	•																									◇	◇				
Trust									⋈	⋈														⋈								
Content quality																	• ◇								◇			◇				
Vagueness											⋈				⋈																	
Local knowledge												⋈																				
Experience																	• ◇	⋈														
Recognition																	• ◇		⋈													
Reputation																	• ◇															

TABLE 2.2: Quality measures and indicators are classified according to the type of methods to assess them, and the types of VGI. Methods are further classified according to the quality assessment approaches. ⋆ = map-based, • = image-based, and ◇ = text-based VGI, while ⋈ = all types of VGI.

Within these 56 papers we came across 30 methods to assess these quality measures and indicators. These quality measures/indicators gather previously discussed spatial data quality elements in the literature, but also extends the previous categorizations such as Thomson et al. (2005), to include further spatial data quality indicators such as reputation, text content quality, or experience.

A classification of the VGI quality measures and indicators according to the type of quality assessment methods and the type of VGI used in the respective applications is presented in Table 2.2. The sparse cells in the matrix indicate the quality measures/indicators that have not been explored excessively. We have further classified these methods according to the approach categorization by Goodchild and Li (2012). In addition to their categorization, we have also found methods based on the data mining approach.

2.5.2.1 Quality assessment in Map-based VGI

Positional Accuracy

In the works of Kounadi (2009), Ather (2009), Haklay (2010), Ciepłuch et al. (2010), Al-Bakri and Fairbairn (2010), Zandbergen et al. (2011), Helbich et al. (2012), Jackson et al. (2013), Fan et al. (2014), Tenney (2014), Brando and Bucher (2010), and Al-Bakri and Fairbairn (2010), authors employ officially gathered reference datasets to assess the positional accuracy of map-based VGI (mostly OSM data) by comparison. The comparison with reference data method has been further employed for the assessment of thematic accuracy (Girres and Touya, 2010; Poser and Dransch, 2010; Kounadi, 2009; Brando and Bucher, 2010; Arsanjani et al., 2015), completeness (Haklay, 2010; Ciepłuch et al., 2010; Kounadi, 2009; Ather, 2009; Ciepłuch et al., 2011; Hecht et al., 2013; Jackson et al., 2013; Fan et al., 2014; Tenney, 2014; Brando and Bucher, 2010), geometric accuracy (Girres and Touya, 2010). For geometric accuracy OSM objects of same structure were manually matched. This manual approach was preferred over an automated approach to avoid any processing errors.

Haklay et al. (2010) applied the Linus Law and found out that higher the number of contributors on a given spatial unit on OSM, higher the quality. This study shows that comparison to reference datasets isn't the only way to assess the quality of OSM data as done in many use-cases.

De Tré et al. (2010) uses a Possibilistic Truth Value (PTV) as a normalized possibility distribution to determine the uncertainty of the POIs being co-located. The uncertainty regarding the positioning of a POI is primarily caused by the imprecision with which the POI are positioned on the map interface. The proposed technique further semantically checks and compares the closely located POIs. Their method helps to identify redundant

VGI, and fuse the redundancies together. Furthermore, this approach has been applied to also assess the thematic accuracy of map-based VGI.

In a rather different approach, Canavosio-Zuzelski et al. (2013) perform a photogrammetric approach for assessing the positional accuracy of OSM road features using stereo imagery and a vector adjustment model. Their method applies analytical measurement principles to compute accurate real world geo-locations of OSM road vectors. The proposed approach was tested on several urban gridded city streets from the OSM database with the results showing that the post adjusted shape points improved positional accuracy by 86%. Furthermore, the vector adjustment was able to recover 95% of the actual positional displacement present in the database.

Brando and Bucher (2010) present a generic framework to manage the quality of ISO standardized quality indicators by using formal specifications and reference datasets. Formal specifications facilitate the assurance of quality in three manners with means of integrity constraints: i) support on-the-fly consistency checking, ii) comparison to external reference data, and iii) reconcile concurrent editions of data. However, due to a lack of proof of concept the practical applicability of this approach is difficult to conceive.

Topological Consistency

The topological consistency in OSM data is assessed mainly on intrinsic data checks to detect and alleviate problems occurring through for example overlapping features or overshoots and undershoots in the data (also known as dangles where start and end point of two different lines should meet but do not, due to bad practices in digitization). The authors Schmitz et al. (2008), Neis et al. (2011), Barron et al. (2014), and Siebritz (2014) have demonstrated that for each of these measures a separate topology integrity rule can be designed and applied.

Further, based on the definition of planar and non-planar topological properties Corcoran et al. (2010) and Da Silva and Wu (2007) have used geometrical analysis methods to assess the topological consistency of the OSM data. In another work, the concept of spatial similarity in multi-representations have been employed in order to perform both extrinsic and intrinsic quality analysis (Hashemi and Abbaspour, 2015). The authors discuss that their method could be efficiently applied to VGI data for the purpose of vandalism detection. Other studies have also focused on evaluating the topological consistency of OSM data with a focus on road network infrastructures (Will, 2014). In Wang et al. (2014) and Girres and Touya (2010) the authors have used the Dimensional Extended nine-Intersection Model (DE-9IM) in order to compute the qualitative spatial relation between road objects in OSM. This method and model allows them to check for topological inconsistencies and be able to locate the junctions of roads in order to, for example generate expected road signs.

Thematic Accuracy and Semantic Accuracy

Mooney and Corcoran (2012b) points out that most errors in OSM are caused by manual annotation by contributors who sometimes misspell the feature values. Addressing this issue, Codescu et al. (2011), Vandecasteele and Devillers (2013), and Ali et al. (2014) have developed semantic similarity matching methods, which automatically assess the contributor annotation of features in OSM according to the semantic meaning of such features. In the work of Girres and Touya (2010), they found semantic errors were mainly due to the mis-specification of roads. For example: roads that were classified as 'secondary' in the reference dataset were classified as 'residential', or 'tertiary' by contributors in OSM data. The reasons for these inaccuracies as seen by authors are the lack of a standardized classification, looseness for contributors to enter tags and values that are not present in the OSM specification, lack of naming regulations w.r.t. for example capitalization or prefixes. The authors emphasize the need for standardized specifications to improve semantic and attribute accuracy of OSM data.

Furthermore, in regard to semantic accuracy of map-based VGI, Vandecasteele and Devillers (2015) introduced a tag recommender system for OSM data which aims to improve the semantic quality of tags. OSMantic is a plugin for the Java OpenStreetMap editor which automatically suggests relevant tags to contributors during the editing process. Mummidi and Krumm (2008) use clustering methods on Microsoft's Live Search Maps[11] to group user contributed pushpins of POIs that are annotated with text. Frequent text phrases that appear in one cluster but infrequently in other clusters help to increase the confidence that the particular text phrase describes a POI.

Completeness

Koukoletsos et al. (2012) propose to use a feature-based automated matching method for linear data using reference datasets. Barron et al. (2014) and Girres and Touya (2010) use intrinsic data checks to record the statistics of the number of objects, attributes, and values, thereby keeping track of all omissions and commissions to the database.

Temporal Accuracy

Very few works exist to assess the temporal accuracy. We reviewed the works of Girres and Touya (2010) where they use statistics to observe the correlations of the number of contributors to the mean capture date, and to the mean version of the capture object in order to assess how many objects are updated. Their results show a linear increase of the mean date, and the mean version of captured object in relation to the number of contributors in the chosen geographic area. Concluding results show higher the number of contributors, more recent the objects were, and the more up-to-date the objects were.

[11] http://maps.live.com

Lineage, Usage, Purpose

In Keßler et al. (2011), following a data oriented approach with a focus on the origins of specific data items, their provenance vocabulary explicitly shows the lineage of data features of any online data. They base their provenance approach on Hartig (2009) on 'provenance information in the web of data'. Their approach allows them to classify OSM features according to recurring editing and co-editing patterns. To keep track of the data lineage Girres and Touya (2010) urge the need for moderators who has control over screening the contributions (as in Wikipedia) for necessary source information. They further analyze the usage of data by comparing the limitations that were observed in previous evaluations of map-based VGI.

As a generic approach to assess ISO standardized quality indicators, (Keßler and Groot, 2013) propose Trust as a proxy to measure the topological consistency, thematic accuracy, and completeness in these map data based on data provenance, a method which relies on trust indicators as opposed to ground truth data.

2.5.2.2 Quality assessment in Image-based VGI

Positional Accuracy and credibility

Jacobs et al. (2007) explored the varying positional accuracy of photos by matching photos with ancillary satellite imagery. They localize cameras based on satellite imagery that correlates with the camera images taken at a known time. Their approach helps where it is important to know the accurate location of the photographer instead of the target object. Zielstra and Hochmair (2013) on the other hand compared the geotagged positions of photos to the manually corrected camera position based on the image content. Their results indicate better positional accuracy for Panoramio photos compared to Flickr photos. Hollenstein and Purves (2014) assessed the positional accuracy of such photos by manually inspecting these photos for their correspondence between the tagged geographic label and geotagged position. Senaratne et al. (2013) assessed the positional accuracy of Flickr photos by computing a line of sight between the camera position and the target position based on in-between surface elevation data. They further manually inspected the geographic label against the geographic location. The results are used as a reference of quality for contributor and photo features of Flickr, and thereby used to derive credibility indicators.

Thematic Accuracy

Foody et al. (2015) use Geowiki as the data source, where it contains a series of satellite imagery. Volunteered contributors were given the task to label the land use categories in these satellite imagery from a pre-defined set of labels. The accuracy of the labeling was

assessed through conducting a latent class analysis (LCA). LCA allows the analyst to derive an accuracy measurement of the classification when there are no reference datasets available to compare with. The authors further emphasize that this method can be applied to image-based VGI. Further, their approach characterizes the volunteers based on the accuracy of their labels of land use classes. This helps to ultimately determine the volunteer quality.

On a related work, Zhang and Kosecka (2006) used feature-based geometric matching using the image recognition software SIFT (Lindeberg, 2012) to localize sample photos in urban environments. Although their work was not based on VGI, this is a potential approach to solve quality related issues within image-based VGI.

2.5.2.3 Quality assessment in Text-based VGI

Quality of text-based VGI has been mainly assessed through the credibility of such data based on contributor, text, and content features, and through the text content quality.

Credibility

Relating to a social approach of quality analysis, Mendoza et al. (2010) found out that rumors on Twitter tend to be more questioned by the Twitter community during an emergency situation. They further indicate that the Twitter community acts as a collaborative filter of information.

Castillo et al. (2011) employed users on mechanical turk[12] to classify pre-classified 'newsworthy events' and 'informal discussions' on Twitter according to several classes of credibility (i. almost certainly true, ii. likely to be false, ..). This is then used in a supervised classification to evaluate which Tweets belong to these different classes of credibility. This helped the authors to derive credibility indicators. The user features such as average status count or the number of followers among others were found to be the top ranked user-based credibility features.

The work of Gupta and Kumaraguru (2012) is similar to Castillo et al. (2011), and follows a supervised feature classification PageRank like method to propagate the credibility on a network of Twitter events. They use event graph-based optimization to enhance the trust analysis at each iteration that updates the credibility scores. A credible entity (node) links with a higher weight to more credible entities than to non-credible ones. Their approach is similar to that of Castillo et al. (2011), but the authors proposed a new technique to re-rank the Tweets based on a Pseudo Relevance Feedback.

[12] https://www.mturk.com

Canini et al. (2011) divided credibility into implicit and explicit credibility. Implicit credibility is the perceived credibility of Twitter contributors, and is assessed by Twitter users by evaluating an external data source together with the Tweeters content topicality and its relevance to the context, and social status (follower/status counts). Explicit credibility is evaluated by ranking Tweeters (Twitter contributors) on a scale from 1 to 5 based on their trustworthiness. End result is a ranking recommendation system on whom to follow on Twitter regarding a particular topic.

O'Donovan et al. (2012) provided an analysis of the distribution of credibility features in four different contexts in the Twitter network: diversity of topics, credibility, chain length and dyadic pairs. The results of their analysis say that the usefulness of credibility features depends on the context in question. Thus the presence of a credibility feature alone is not good enough to evaluate the credibility of the context, but rather a particular combination of different credibility features that are 'suitable' for the context in question.

Morris et al. (2012) designed a pilot study with participants (with no technical background) to extract a list of features that are useful to make their credibility judgments. Finally to run the survey, the authors sent the survey to a sample of Twitter users in which they were asked to assess how each feature impacts their credibility judgment on a five-point scale. Their findings indicate that features such as verified author expertise, re-tweets from someone you trust, or author is someone you follow have higher credibility impact. These features differ somewhat to the features extracted through the supervised classification of Castillo et al. (2011). These features were further ranked according to the amount of attention received by Twitter users.

Kang et al. (2012) defined three different credibility prediction models and studied how each model performs in terms of credibility classification of Twitter messages. These are: (1) social model, (2) content-based model, and (3) hybrid model (based on different combinations of the two previous models). The social model relies on a weighted combination of credibility indicators from the underlying social network (e.g., re-tweets, no. of followers). The content-based model identifies patterns and tweet properties that leads to positive reactions such as re-tweeting or positive user ratings, by using a probabilistic language-based approach. Most of these content-based features are taken from Castillo et al. (2011). The main results from the paper indicate that the social model outperformed all other models in terms of predication accuracy, and that including more features in the predication task doesn't mean a better predication accuracy.

Text Content Quality

Agichtein et al. (2008) describes a generic method for all text-based social media data.They use three inputs for a feature classifier to determine the content quality: (1) textual features (e.g., word n-grams up to length 5 that appears in the text more than 3 times,

semantic features such as punctuations, typos, readability measures, avg. no. of syllables per word, entropy of word lengths, grammarticality), (2) user relationships (between users and items, uses intuition such as good answers are given by good answerers, and vote for other good answerers), (3) usage statistics (no. of clicks on an item, dwell time on content).

Becker et al. (2011) use a two tier approach for the quality analysis of text-based Twitter data in an event analysis context. To identify the events, they first cluster tweets using an online clustering framework. Subsequently, they use three centrality based approaches to identify messages in the clusters that have high textual quality, strong relevance, and are useful. These approaches are: (1) centroid similarity approach that calculates the cosine similarity of the 'tf-idf' statistic of words, (2) degree centrality methods which represents each cluster message as a node in a graph, and two nodes are connected with an edge when their cosine similarity exceeds a predetermined threshold, (3) LexRank approach distributes the centrality value of nodes to its neighbors, and top messages in a cluster are chosen according to their LexRank value.

Hasan Dalip et al. (2009) on the other hand used text length, structure, style readability, revision history, and social network as indicators of text content quality in Wikipedia articles. They further use regression analysis to combine various such weighed quality values into a single quality value, that represents an overall aggregated quality metric for text content quality.

Bordogna et al. (2014) measured the validity of text data by measuring the number of words, proportion of correctly spelled words, language intelligibility, diffusion of words, and the presence of technical terms as indicators of text content quality. They further explored quality indicators such as experience, recognition and reputation to determine the quality of VGI.

2.5.2.4 Generic approaches

As a generic method for all VGI Forghani and Delavar (2014) propose a new quality metric for the assessment of topological consistency by employing heuristic metrics such as minimum bounding geometry area and directional distribution (Standard Deviational Ellipse). Van Exel et al. (2010) propose to use contributor related quality indicators such as local knowledge (e.g., spatial familiarity), experience (e.g., amount of contributions), and recognition (e.g., tokens achieved). A conceptual workflow for automatically assessing the quality of VGI in crisis management scenarios was proposed by Ostermann and Spinsanti (2011). VGI is cross-referenced with other VGI types, and institutional ancillary data that are spatially and temporally close. However, in a realistic implementation

this combination of different VGI data types for cross referencing is a challenging task due to their heterogeneity. Bishr and Janowicz (2010) proposed to use trust together with reputation as a proxy measure for VGI quality, and established the spatial and temporal dimensions of trust. They assert that shorter geographic proximity of VGI observations provide more accurate information as opposed to higher geographic proximity VGI observations (implying that *locals know better*, *the proximate spectator sees more*). On a temporal perspective of trust, they further claim that trust in some VGI develop and decay over time, and that the observation time of an event has an affect on the trust we endow in one's observation. Furthermore, to assess the trust of VGI Huang et al. (2010) developed a method to detect outliers in the contributed data. De Longueville et al. (2010) proposed two methods to assess the vagueness in VGI. (1) contributor encodes the vagueness of their contributed spatial data in a 0 - 5 scale (e.g., 5 = it's exactly there, 0 = I don't know where it is. (2) the second type is system created vagueness that is assessed through automatically capturing the scale at which VGI is produced. VGI produced in lower scales is classified as more vague.

Table 2.2 shows a summary matrix of all quality measures and indicators observed in the literature review, with various methods that can be applied to assess these quality measures/indicators. Following this matrix we can learn which methods can be applied to solve various quality issues within map, text and image-based VGI. However, this should be followed with caution, as we present here only what we discovered through the literature review, and the presented methods could be applied beyond our discovery, and therefore need to be further explored.

2.6 Discussion and Future Research Perspectives in VGI Quality

VGI is available with tremendous amounts through various platforms, and it is crucial to have methods to ensure the quality of these VGI. The vast amount of data and the heterogeneous characteristics of utilization make the traditional comparison with reference data sets no longer viable in every application scenario (also due to the lack of access to reference data). Based on such characteristics, Goodchild and Li (2012) propose three approaches to ensure the quality of VGI: (1) crowd-sourced, (2) social, and (3) geographic. As seen in Table 2.2, 20 of the methods we have discovered in the literature fall in to geographic, social, or crowd-sourced approaches. Furthermore, 10 of the methods we discovered fall in to an additional approach: 4) data mining, that helps to assess VGI quality by discovering patterns and learning purely from the data. Data mining can be used as a stand-alone approach, completely independent of the laws and knowledge of geography, and independent from social or crowd-sourced approaches to assess the quality of VGI. For example, the possibilistic truth value method is used to assess

the positional uncertainty of POIs based only on the possibility distribution. Similarly, outlier detection, cluster analysis, regression analysis, or correlation statistics methods can be used to assess the data quality by purely discovering and learning data patterns, irrespective of the laws and knowledge from geography. The supervised learning, and feature classification methods that are used to assess the quality of text based VGI use text, message, and user features to train the classifier. These two machine learning methods we found in the literature once again work irrespective of the laws and knowledge from geography. Therefore, we believe these methods deserve to be represented under an additional approach to assess VGI quality.

We have classified the found methods according to these 4 approaches based on the description of the methods in the literature. By this discovery, we aim to extend Goodchild and Li (2012)'s classification through this survey.

While most methods have been utilized to assess the positional accuracy, thematic accuracy, and topological consistency, fewer methods tackle the rest of the quality measures and indicators we review such as the completeness, temporal accuracy or vagueness. Future work should focus also on other potential approaches to handle quality measures and indicators. Different VGI platforms should clearly communicate to the contributors and the consumers, as to what kind of data that one could contribute. The more precise this is, the more comprehensive it is to the contributor on what is expected in terms of data. As also stated by Antoniou et al. (2010), explicit VGI gives a loosely coupled specification(s) of what volunteers can contribute. If these specifications are more rigid the future of VGI can expect higher quality information, although it may be a compromise with lesser contributions. This may further vary depending on the task at hand.

Lower population density positively correlates with fewer number of contributions, thus affecting data completeness or positional accuracy (Neis et al., 2013; Haklay, 2010; Girres and Touya, 2010; Mullen et al., 2014). However, more research needs to be done regarding this issue. Hence, a step further in this direction is to derive the socio-economic impacts on OSM data quality. As presented in section 5.2., there have been a number of studies and empirical research performed on the subject of OSM quality. Nevertheless, a solid framework for assessing OSM data is far from being established, let alone a framework of quality measurement for specific application domains. The limitation is that existing measures and indicators (described by ISO) are not inclusive enough to evaluate OSM data. This is mainly because the nature of OSM (and VGI in general) is fundamentally different to what geospatial experts have dealt with so far. Therefore, we argue that there are still research gaps when defining quality measures/indicators and proposing methods to calculate these measures/indicators. In addition, only few studies have been conducted to explore and analyze the differences in quality requirements for different

application domains. Therefore, as a recommendation for future research in this topic, we suggest to develop a systematic framework that provides methods and measures to evaluate the fitness for purpose of each VGI type. This would need to not only focus on the analysis of data itself, but also explore the social factors which are the driving forces behind public contributions, and thus considerably affect the quality. For example, one could define a mathematical model based on OSM intrinsic data indicators (e.g., number of contributors, number of edits, etc.) to estimate the quality (e.g., completeness) of data without having reference data in hand. This would enrich and complete the new paradigm of intrinsic quality evaluation, which by far has received less attention by the research community, compared to the common extrinsic quality evaluation: i.e., comparison with reference data.

The utilization of text and image-based VGI still mostly depend on the geo-tagged content. However, the sparse geo-tagged content of these two VGI types in most cases represent only a minority of the data. Therefore, generalization based on VGI is still limited and need further demographic studies.

Gamification has become a popular way to involve people to contribute spatial data (Geograph, Foursquare[13], Ingress[14] are some examples). Such gamification approaches have increased participation as well as spatial coverage (Antoniou and Schlieder, 2014; Antoniou et al., 2010). Due to the clear incentives of this data collection approach (going high up in rankings, collecting badges etc.) this popular method can be used to control the process of collecting more accurate data by incorporating data quality concepts (Yanenko and Schlieder, 2014). One way to do that would be to give a ranking to the contributor based on the quality of their collected data. Revealing such rankings of their peers would further encourage the contributors to pay more attention to the quality of their data (peer pressure).

As encouragement mechanisms are required to motivate people to contribute, we should also research methods to make contributors aware of the importance of quality, and secondly to involve the contributors and consumers to maintain the quality of the VGI contents. This can be achieved for example by collaboratively doing quality checks on the data. Such collaborative efforts are presently actively done in OSM, but rather inadvertently done on Flickr or Twitter. As evident from the review, image and text-based VGI have been given far less attention to its quality as compared to map-based VGI. We see this as mainly due to the complexity of the image and text data types. Comments and discussions associated with image and text contents might be one way to ensure the contribution while systematic analysis of these resources is not a trivial

[13] https://foursquare.com/
[14] https://www.ingress.com/

process. Our understanding is that quality assurance methods for text and image-based VGI are still on the phase of experimentation, and therefore need more attention in order to standardize these methods in to practice. This is crucial because more and more text and image-based VGI are being utilized in various applications. Furthermore, the works of Sacha et al. (2014), where they introduce a framework that integrates trust and other various quality indicators in a knowledge generation process within the visual analytics paradigm can be adapted in future research to assess and visually analyze quality of VGI. Their framework allows the user to comprehend the associated quality at each step of knowledge generation, and also express their confidence in the findings and insights gained by externalizing their thoughts. This facilitates the user to comprehend the provided quality of data as well as the perceived quality.

As further evident from this review, there is no holy grail that could solve all types of quality issues in VGI. We should be aware of the heterogeneity of these data, and be informed of the existing state-of-the-art to resolve many of the quality issues of VGI, and their limitations. Addressing these limitations and thereby improving the existing methods already paves for new contributions on this topic that should be recognized as valid scientific contributions in the VGI community.

2.7 Conclusions

In this review of VGI quality, we have taken a critical look at the quality issues within map, image, and text VGI types. The heterogeneity of these VGI types give rise to varying quality issues that need to be dealt with varying quality measures and indicators, and varying methods. As a result of this review, we have summarized the literature in to a list of 30 methods that can be used to assess one or more of the 17 quality measures and indicators that we have come across in the literature for map, image, and text-based VGI respectively. This review further shows the following: 1) a majority of reviewed papers focus on assessing map-based VGI. 2) Though implicit VGI (e.g., text-based Twitter or image-based Flickr) has higher quality concerns in comparison to explicit VGI (e.g., map-based OSM), such explicit VGI has received significantly higher attention to resolve quality issues, compared to implicit VGI. The review shows the increasing utilization of implicit VGI for geospatial research. Therefore, more efforts should be in place to resolve quality issues within these implicit VGI. 3) Mostly ISO standardized quality measures have been used to assess the quality of map-based VGI (OSM). Text-based VGI have been assessed on the credibility, vagueness, and the content quality. Image-based VGI have been assessed on the positional/thematic accuracy, credibility, vagueness, experience, recognition, and reputation. A logical explanation for this is that ISO standardized measures are most often assessed through comparative analysis with ground truth data.

For the explicit VGI (e.g., OSM) we can easily realize which ground truth data to look for. However for implicit VGI, it is not straight forward to realize which ground truth data to look for, therefore comparative analysis is not always possible (e.g., topological consistency, or thematic accuracy cannot be directly assessed, as we need to derive the topology or the thematic attributes from the VGI in an additional data processing step). These implicit VGI are further enriched with contributor sentiments and contextual information. Therefore ISO standardized measures alone are not enough to assess the quality of implicit VGI. This explains the use of indicators such as reputation, trust, credibility, vagueness, experience, recognition, or local knowledge as quality indicators. A lack of standardization of these more abstract quality indicators is a reason why fewer works exist for image and text-based VGI. In addition, the implicit nature of the geography that is contributed in most of these VGI is yet another reason for the insufficiency of quality assessment methods for text and image-based VGI. 4) we have classified the quality assessment methods according to the crowd-sourced, geographic, and social approaches as introduced by Goodchild and Li (2012). We have further discovered data mining as an additional approach in the literature that extends Goodchild and Li (2012)'s classification.

Acknowledgments

This work has been partly funded by the SPP programme under grant agreement no. 1335 (ViAMoD), the European Union's Seventh Framework Programme under grant agreement no. 612096 (CAP4Access), and the German Academic Exchange Service (DAAD). We thank particularly Tobias Schreck, Alexander Zipf, Mohamed Bakillah, and Hongchao Fan for their valuable discussions on the topic.

Bibliography

Agichtein, E., C. Castillo, D. Donato, A. Gionis, and G. Mishne (2008). "Finding high-quality content in social media". In: *Proceedings of the 2008 International Conference on Web Search and Data Mining*. ACM, pp. 183–194.

Al-Bakri, M. and D. Fairbairn (2010). "Assessing the accuracy of Crowdsourced data and its integration with official spatial data sets". In: *Proceedings of Accuracy 2010 Symposium*. Leicester, UK, pp. 317–320.

Ali, A. L., F. Schmid, R. Al-Salman, and T. Kauppinen (2014). "Ambiguity and plausibility: managing classification quality in Volunteered Geographic Information". In: *Proc. of the 22nd ACM SIGSPATIAL International Conf. on Advances in Geographic Information Systems*. Ed. by Y. Huang, M. Schneider, M. Gertz, J. Krumm, and J. Sankaranarayanan. New York, NY, USA: ACM, pp. 143–152. ISBN: 978-1-4503-3131-9.

Andrienko, G., N. Andrienko, P. Bak, S. Kisilevich, and D. Keim (2009). "Analysis of community-contributed space-and time-referenced data (example of flickr and panoramio photos)". In: *Proceedings of IEEE Symposium on Visual Analytics Science and Technology (VAST 2009)*. IEEE, pp. 213–214.

Andrienko, G., N. Andrienko, H. Bosch, T. Ertl, G. Fuchs, P. Jankowski, and D. Thom (2013). "Thematic patterns in georeferenced tweets through space-time visual analytics". In: *Computing in Science and Engineering* 15.3, pp. 72–82.

Antoniou, V and A Skopeliti (2015). "Measures and Indicators of VGI Quality: an Overview". In: *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* 1, pp. 345–351.

Antoniou, V. and C. Schlieder (2014). "Participation Patterns, VGI and Gamification". In: *Connecting a Digital Europe Through Location and Place*. Ed. by J. Huerta, S. Schade, and C. Granell. Springer-Verlag, pp. 3–6.

Antoniou, V., J. Morley, and M. Haklay (2010). "Web 2.0 geotagged photos: Assessing the spatial dimension of the phenomenon". In: *Geomatica* 64.1, pp. 99–110.

Arsanjani, J. J., P. Mooney, A. Zipf, and A. Schauss (2015). "Quality Assessment of the Contributed Land Use Information from OpenStreetMap Versus Authoritative Datasets". In: *OpenStreetMap in GIScience: experiences, research, and applications*. Ed. by J. J. Arsanjani, A. Zipf, P. Mooney, and M. Helbich. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 37–58.

Ather, A. (2009). "A quality analysis of OpenStreetMap data". MA thesis. University College of London, London, UK.

Barron, C., P. Neis, and A. Zipf (2014). "A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis". In: *Transactions in GIS* 18.6, pp. 877–895.

Becker, H., M. Naaman, and L. Gravano (2011). "Selecting Quality Twitter Content for Events". In: *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*. Ed. by L. A. Adamic, R. A. Baeza-Yates, and S. Counts. Vol. 11. The AAAI Press.

Bishr, M. and K. Janowicz (2010). "Can we trust information?-the case of Volunteered Geographic Information". In: *Towards Digital Earth Search Discover and Share Geospatial Data Workshop at Future Internet Symposium, volume*. Vol. 640.

Bordogna, G., P. Carrara, L. Criscuolo, M. Pepe, and A. Rampini (2014). "A linguistic decision making approach to assess the quality of volunteer geographic information for citizen science". In: *Information Sciences* 258, pp. 312–327.

Bosch, H., D. Thom, F. Heimerl, E. Puttmann, S. Koch, R. Kruger, M. Worner, and T. Ertl (2013). "Scatterblogs2: Real-time monitoring of microblog messages through user-guided filtering". In: *Visualization and Computer Graphics, IEEE Transactions on* 19.12, pp. 2022–2031.

Brando, C. and B. Bucher (2010). "Quality in user generated spatial content: A matter of specifications". In: *Geospatial Thinking*. Ed. by M. Painho, M. Yasmina Santos, and H. Pundt. Springer-Verlag. Springer-Verlag, pp. 11–14.

Bulearca, M. and S. Bulearca (2010). "Twitter: a viable marketing tool for SMEs". In: *Global Business and Management Research* 2.4, pp. 296–309. ISSN: 1947-5667.

Canavosio-Zuzelski, R., P. Agouris, and P. Doucette (2013). "A photogrammetric approach for assessing positional accuracy of OpenStreetMap© roads". In: *ISPRS International Journal of Geo-Information* 2.2, pp. 276–301.

Canini, K. R., B. Suh, and P. L. Pirolli (2011). "Finding credible information sources in social networks based on content and social structure". In: *3rd International Conference on Social Computing (SocialCom), Privacy, Security, Risk and Trust (PASSAT)*. IEEE, pp. 1–8.

Castillo, C., M. Mendoza, and B. Poblete (2011). "Information Credibility on Twitter". In: *Proceedings of the 20th International Conference on World Wide Web*. WWW '11. New York, NY, USA: ACM, pp. 675–684. ISBN: 978-1-4503-0632-4.

Chunara, R., J. R. Andrews, and J. S. Brownstein (2012). "Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak". In: *The American Journal of Tropical Medicine and Hygiene* 86.1, pp. 39–45.

Ciepłuch, B., R. Jacob, P. Mooney, and A. Winstanley (2010). "Comparison of the accuracy of OpenStreetMap for Ireland with Google Maps and Bing Maps". In: *Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences 20-23rd July 2010*. University of Leicester, p. 337.

Ciepłuch, B., P. Mooney, R. Jacob, J. Zheng, and A Winstanely (2011). "Assessing the quality of open spatial data for mobile location-based services research and applications". In: *Archives of Photogrammetry, Cartography and Remote Sensing* 22, pp. 105–116. ISSN: 20832214.

Codescu, M., G. Horsinka, O. Kutz, T. Mossakowski, and R. Rau (2011). "OSMonto – An Ontology of OpenStreetMap Tags". In: *Proceedings of the SOTM-EU 2011 : 1st State of the Map - Europe Conference*. Ed. by M. Schmidt and G. Gartner, pp. 55 –65.

Corcoran, P., P. Mooney, and A. Winstanley (2010). "Topological Consistent Generalization of OpenStreetMap". In: *Proceedings of GISRUK 2010: GIS Research UK 18th Annual Conference*. London, UK: Maynooth University.

Craglia, M, F Ostermann, and L. Spinsanti (2012). "Digital Earth from vision to practice: making sense of citizen-generated content". In: *International Journal of Digital Earth* 5.5, pp. 398–416.

Da Silva, A. C. and S.-T. Wu (2007). "Consistent handling of linear features in polyline simplification". In: *Advances in Geoinformatics*. Springer, pp. 1–17.

De Longueville, B., N. Ostländer, and C. Keskitalo (2010). "Addressing vagueness in Volunteered Geographic Information (VGI)–A case study". In: *International Journal of Spatial Data Infrastructures Research* 5, pp. 1725–0463.

De Tré, G., A. Bronselaer, T. Matthé, N. Van de Weghe, and P. De Maeyer (2010). "Consistently Handling Geographical User Data". In: *Information Processing and Management of Uncertainty in Knowledge-Based Systems, Applications*. Springer, pp. 85–94.

Fan, H., A. Zipf, Q. Fu, and P. Neis (2014). "Quality assessment for building footprints data on OpenStreetMap". In: *International Journal of Geographical Information Science* 28.4, pp. 700–719.

Flanagin, A. J. and M. J. Metzger (2008). "The credibility of Volunteered Geographic Information". In: *GeoJournal* 72.3, pp. 137–148.

Foody, G. M., L. See, S. Fritz, M. van der Velde, C. Perger, C. Schill, D. S. Boyd, and A. Comber (2015). "Accurate Attribute Mapping from Volunteered Geographic Information: Issues of Volunteer Quantity and Quality". In: *The Cartographic Journal* 52.4, pp. 336–344.

Forghani, M. and M. R. Delavar (2014). "A quality study of the OpenStreetMap dataset for Tehran". In: *ISPRS International Journal of Geo-Information* 3.2, pp. 750–763.

Frew, J. (2007). "Provenance and Volunteered Geographic Information". In: *Retrieved March* 10, p. 2008.

Fuchs, G., N. Andrienko, G. Andrienko, S. Bothe, and H. Stange (2013). "Tracing the German centennial flood in the stream of tweets: first lessons learned". In: *Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*. ACM, pp. 31–38.

Girres, J.-F. and G. Touya (2010). "Quality assessment of the French OpenStreetMap dataset". In: *Transactions in GIS* 14.4, pp. 435–459.

Golder, S. A. and B. A. Huberman (2006). "Usage patterns of collaborative tagging systems". In: *Journal of information science* 32.2, pp. 198–208.

Goodchild, M. F. (2007). "Citizens as sensors: the world of volunteered geography". In: *GeoJournal* 69.4, pp. 211–221.

Goodchild, M. F. and L. Li (2012). "Assuring the quality of Volunteered Geographic Information". In: *Spatial statistics* 1, pp. 110–120.

Guinée, J. B. (2002). "Handbook on life cycle assessment operational guide to the ISO standards". In: *The International Journal of Life Cycle Assessment* 7.5, pp. 311–313.

Gupta, A. and P. Kumaraguru (2012). "Credibility ranking of tweets during high impact events". In: *Proceedings of the 1st Workshop on Privacy and Security in Online Social Media*. ACM, p. 2.

Haklay, M. (2010). "How good is Volunteered Geographic Information? A comparative study of OpenStreetMap and Ordnance Survey datasets". In: *Environment and planning. B Planning & design* 37.4, p. 682.

Haklay, M., S. Basiouka, V. Antoniou, and A. Ather (2010). "How many volunteers does it take to map an area well? The validity of Linus' law to Volunteered Geographic Information". In: *The Cartographic Journal* 47.4, pp. 315–322.

Hartig, O. (2009). "Provenance Information in the Web of Data." In: *LDOW* 538.

Hasan Dalip, D., M. André Gonçalves, M. Cristo, and P. Calado (2009). "Automatic quality assessment of content created collaboratively by web communities: a case study of wikipedia". In: *Proceedings of the 9th ACM/IEEE-CS joint Conference on Digital libraries*. ACM, pp. 295–304.

Hashemi, P. and R. A. Abbaspour (2015). "Assessment of Logical Consistency in OpenStreetMap Based on the Spatial Similarity Concept". In: *OpenStreetMap in GIScience: experiences, research, and applications*. Ed. by J. J. Arsanjani, A. Zipf, P. Mooney, and M. Helbich. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 19–36.

Hecht, R., C. Kunze, and S. Hahmann (2013). "Measuring completeness of building footprints in OpenStreetMap over space and time". In: *ISPRS International Journal of Geo-Information* 2.4, pp. 1066–1091.

Helbich, M., C. Amelunxen, P. Neis, and A. Zipf (2012). "Comparative spatial analysis of positional accuracy of OpenStreetMap and proprietary geodata". In: *Proceedings of GI_Forum 2012*, pp. 24–33.

Hollenstein, L. and R. Purves (2014). "Exploring place through user-generated content: Using Flickr tags to describe city cores". In: *Journal of Spatial Information Science* 1, pp. 21–48.

Hovland, C. I., I. L. Janis, and H. H. Kelley (1953). *Communication and persuasion; psychological studies of opinion change*. Yale University Press. Chap. Communication and persuasion; psychological studies of opinion change.

Hoyle, D. (2001). *ISO 9000: Quality Systems Handbook*. 4th ed. Oxford, UK: Butterworth-Heinemann.

Huang, K. L., S. S. Kanhere, and W. Hu (2010). "Are You Contributing Trustworthy Data?: the case for a reputation system in participatory sensing". In: *Geospatial Thinking*. MSWIM '10. New York, NY, USA: ACM, pp. 14–22. ISBN: 978-1-4503-0274-6.

Huberman, B. A., D. M. Romero, and F. Wu (2008). "Social networks that matter: Twitter under the microscope". In: *First Monday:* 14.1.

Jackson, S. P., W. Mullen, P. Agouris, A. Crooks, A. Croitoru, and A. Stefanidis (2013). "Assessing completeness and spatial error of features in Volunteered Geographic Information". In: *ISPRS International Journal of Geo-Information* 2.2, pp. 507–530.

Jacob, R., J. Zheng, B. Ciepłuch, P. Mooney, and A. C. Winstanley (2009). "Campus guidance system for international conferences based on OpenStreetMap". In: *Web and Wireless Geographical Information Systems*. Springer, pp. 187–198.

Jacobs, N., S. Satkin, N. Roman, R. Speyer, and R. Pless (2007). "Geolocating static cameras". In: *IEEE 11th International Conference on Computer Vision (ICCV 2007)*. IEEE, pp. 1–6.

Kang, B., J. O'Donovan, and T. Höllerer (2012). "Modeling topic specific credibility on twitter". In: *Proceedings of the ACM International Conference on Intelligent User Interfaces*. ACM, pp. 179–188.

Keßler, C. and R. T. A. de Groot (2013). "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*. Ed. by D. Vandenbroucke, B. Bucher, and J. Crompvoets. Springer-Verlag, pp. 21–37.

Keßler, C., P. Maué, J. T. Heuer, and T. Bartoschek (2009). "Bottom-up gazetteers: Learning from the implicit semantics of geotags". In: *GeoSpatial semantics*. Springer, pp. 83–102.

Keßler, C., J. Trame, and T. Kauppinen (2011). "Tracking editing processes in Volunteered Geographic Information: the case of OpenStreetMap". In: *Proceedings of Workshop on Identifying objects, processes and events in spatio-temporally distributed data (IOPE), Conference on Spatial Information Theory (COSIT 2011)*. Vol. 12.

Koukoletsos, T., M. Haklay, and C. Ellul (2012). "Assessing data completeness of VGI through an automated matching procedure for linear data". In: *Transactions in GIS* 16.4, pp. 477–498.

Kounadi, O. (2009). "Assessing the quality of OpenStreetMap data". MA thesis. University College of London (UCL), Department of Civil, Environmental And Geomatic Engineering.

Lindeberg, T. (2012). "Scale invariant feature transform". In: *Scholarpedia* 7.5, p. 10491.

Liu, S. B., L. Palen, J. Sutton, A. L. Hughes, and S. Vieweg (2008). "In search of the bigger picture: The emergent role of on-line photo sharing in times of disaster". In: *Proceedings of the Information Systems for Crisis Response and Management Conference (ISCRAM)*.

MacEachren, A. M., A. Jaiswal, A. C. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang, and J. Blanford (2011). "Senseplace2: Geotwitter analytics support for situational awareness". In: *IEEE Conference on Visual Analytics Science and Technology (VAST), 2011*. IEEE, pp. 181–190.

Maué, P. (2007). "Reputation as tool to ensure validity of VGI". In: *Proceedings of Workshop on Volunteered Geographic Information*. University of California, Santa Barbara.

McDougall, K. (2009). "The potential of citizen volunteered spatial information for building SDI". In: *Proc. of 11th World Conference on Spatial Data Infrastructure Convergence: Building SDI Bridges to Address Global Challenges*. Rotterdam, The Netherlands: GSDI Association Press.

Mendoza, M., B. Poblete, and C. Castillo (2010). "Twitter Under Crisis: Can we trust what we RT?" In: *Proceedings of the first workshop on social media analytics*. ACM, pp. 71–79.

Milholland, N. and E. Pultar (2013). "The San Francisco public art map application: using VGI and social media to complement institutional data sources". In: *Proceedings of the 1st ACM SIGSPATIAL International Workshop on MapInteraction*. ACM, pp. 48–53.

Montello, D. R., M. F. Goodchild, J. Gottsegen, and P. Fohl (2003). "Where's downtown?: Behavioral methods for determining referents of vague spatial queries". In: *Spatial Cognition & Computation* 3.2-3, pp. 185–204.

Mooney, P. and P. Corcoran (2012b). "The annotation process in OpenStreetMap". In: *Transactions in GIS* 16.4, pp. 561–579.

Morris, M. R., S. Counts, A. Roseway, A. Hoff, and J. Schwarz (2012). "Tweeting is believing?: understanding microblog credibility perceptions". In: *Proceedings of the ACM Conference on Computer Supported Cooperative Work, 2012*. ACM, pp. 441–450.

Mullen, W. F., S. P. Jackson, A. Croitoru, A. Crooks, A. Stefanidis, and P. Agouris (2014). "Assessing the impact of demographic characteristics on spatial error in Volunteered Geographic Information features". In: *GeoJournal*, pp. 1–19.

Mummidi, L. N. and J. Krumm (2008). "Discovering points of interest from users? map annotations". In: *GeoJournal* 72.3-4, pp. 215–227.

Neis, P., D. Zielstra, and A. Zipf (2011). "The street network evolution of crowdsourced maps: OpenStreetMap in Germany 2007–2011". In: *Future Internet* 4.1, pp. 1–21.

Neis, P., D. Zielstra, and A. Zipf (2013). "Comparison of Volunteered Geographic Information Data Contributions and Community Development for Selected World Regions". In: *Future Internet* 5.2, pp. 282–300.

O'Connor, R. (2009). *GLOBAL: Facebook and Twitter 'reshaping journalism as we know it'*. Retrieved from: http://kauri.aut.ac.nz:8080/dspace/handle/123456789/1839.

O'Donovan, J., B. Kang, G. Meyer, T. Hollerer, and S. Adalii (2012). "Credibility in context: An analysis of feature distributions in twitter". In: *4th International Conference on Social Computing (SocialCom),Privacy, Security, Risk and Trust (PASSAT)*. IEEE, pp. 293–301.

Ostermann, F. O. and L. Spinsanti (2011). "A conceptual workflow for automatically assessing the quality of Volunteered Geographic Information for crisis management". In: *Advancing Geoinformation Science for a Changing World*. Ed. by S. Geertman, W. Reinhardt, and F. Toppen. Springer-Verlag.

Popescu, A., G. Grefenstette, and P. A. Moëllic (2008). "Gazetiki: automatic creation of a geographical gazetteer". In: *Proceedings of the 8th ACM/IEEE-CS joint Conference on Digital libraries*. ACM, pp. 85–93.

Poser, K. and D. Dransch (2010). "Volunteered geographic information for disaster management with application to rapid flood damage estimation". In: *Geomatica* 64.1, pp. 89–98.

Resnick, P., K. Kuwabara, R. Zeckhauser, and E. Friedman (2000). "Reputation systems". In: *Communications of the ACM* 43.12, pp. 45–48. ISSN: 0001-0782.

Robinson, S., M. Jones, J. Williamson, R. Murray-Smith, P. Eslambolchilar, and M. Lindborg (2012). "Navigation your way: from spontaneous independent exploration to dynamic social journeys". In: *Personal and Ubiquitous Computing* 16.8, pp. 973–985.

Sacha, D., H. Senaratne, B. C. Kwon, and D. A. Keim (2014). "Uncertainty Propagation and Trust Building in Visual Analytics". In: *IEEE VIS 2014 - Provenance for Sensemaking Workshop (poster paper)*.

Sakaki, T., M. Okazaki, and Y. Matsuo (2010). "Earthquake shakes Twitter users: real-time event detection by social sensors". In: *Proceedings of the 19th International Conference on World Wide Web*. WWW '10. ACM, pp. 851–860.

Schmitz, S., P. Neis, and A. Zipf (2008). "New applications based on collaborative geodata–the case of routing". In: *Proceedings of XXVIII INCA International Congress on Collaborative Mapping and Space Technology*.

Senaratne, H., A. Bröring, and T. Schreck (2013). "Using Reverse Viewshed Analysis to Assess the Location Correctness of Visually Generated VGI". In: *Transactions in GIS* 17.3, pp. 369–386.

Senaratne, H., A. Bröring, T. Schreck, and D. Lehle (2014). "Moving on Twitter: Using Episodic Hotspot and Drift Analysis to Detect and Characterise Spatial Trajectories".

In: *Proceedings of the 7th ACM SIGSPATIAL International Workshop on Location-Based Social Networks (LBSN 2014)*.

Siebritz, L.-A. (2014). "Assessing the accuracy of OpenStreetMap data in south africa for the purpose of integrating it with authoritative data". MA thesis. University of Cape Town.

Tenney, M. (2014). "Quality Evaluations on Canadian OpenStreetMap Data". In: *Proceedings of Conference on Spatial Knowledge and Information*. Canada.

Thomson, J., E. Hetzler, A. MacEachren, M. Gahegan, and M. Pavel (2005). "A typology for visualizing uncertainty". In: *Electronic Imaging 2005*. International Society for Optics and Photonics, pp. 146–157.

Valli, C. and P. Hannay (2010). "Geotagging Where Cyberspace Comes to Your Place." In: *Security and Management*, pp. 627–632.

Van Exel, M., E. Dias, and S. Fruijtier (2010). "The impact of crowdsourcing on spatial data quality indicators". In: *(Extended abstract) in the 6th International Conference on Geographic Information Science (GIScience), Zurich, Switzerland, September 14-17, 2010*.

Van Oort, P. and A. Bregt (2005). "Do Users Ignore Spatial Data Quality? A Decision-Theoretic Perspective". In: *Risk analysis* 25.6, pp. 1599–1610.

Vandecasteele, A. and R. Devillers (2013). "Improving Volunteered Geographic Data Quality Using Semantic Similarity Measurements". In: *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 1.1, pp. 143–148.

Vandecasteele, A. and R. Devillers (2015). "Improving Volunteered Geographic Information Quality Using a Tag Recommender System: The Case of OpenStreetMap". In: *OpenStreetMap in GIScience: experiences, research, and applications*. Ed. by J. J. Arsanjani, A. Zipf, P. Mooney, and M. Helbich. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 59–80.

Wang, D., Z. Huang, Q. Liu, X. Zhang, D. Xu, Z. Wang, N. Li, J. Zhang, and D. Zhang (2014). "Using Semantic Technology for Consistency Checking of Road Signs". In: *Web Information Systems Engineering–WISE 2013 Workshops*. Springer, pp. 11–22.

Will, J. (2014). "Development of an automated matching algorithm to assess the quality of the OpenStreetMap road network: a case study in Göteborg, Sweden". MA thesis. Lund University, Sweden.

Yanenko, O. and C. Schlieder (2014). "Game principles for enhancing the quality of user-generated data collections". In: *Workshop of Geogames and Geoplay in the 17th AGILE Conference on Geographic Information Science,* pp. 1–5.

Zandbergen, P. A., D. A. Ignizio, and K. E. Lenzer (2011). "Positional accuracy of TIGER 2000 and 2009 road networks". In: *Transactions in GIS* 15.4, pp. 495–519.

Zhang, W. and J. Kosecka (2006). "Image based localization in urban environments". In: *3D Data Processing, Visualization, and Transmission, Third International Symposium on*. IEEE, pp. 33–40.

Zielstra, D. and H. H. Hochmair (2013). "Positional accuracy analysis of Flickr and Panoramio images for selected world regions". In: *Journal of Spatial Science* 58.2, pp. 251–273.

Chapter 3

Data Quality Assurance for Volunteered Geographic Information

Authors:

Ahmed Loai Ali and Falko Schmid.

Conference:

The 8th International Conference on Geographic Information Science (GIScience 2014).

Citation:

Ahmed Loai Ali and Falko Schmid (2014). "Data quality assurance for Volunteered Geographic Information". In: *Geographic Information Science: 8th International Conference, GIScience 2014, Vienna, Austria, September 24-26, 2014. Proceedings.* Ed. by Matt Duckham, Edzer Pebesma, Kathleen Stewart, and Andrew U. Frank. Cham: Springer International Publishing, pp. 126–141. ISBN: 978-3-319-11593-1.

Contribution Statement:

The idea was initiated through the discussion between the authors. I developed the proposed approach, implemented the analysis and performed the empirical studies. Furthermore, I contributed strongly in conducting the learning task and analyzing the results. Falko contributed in developing the approach, in discussing the results, and in editing the text.

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.

3.1 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) (Goodchild, 2007), 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 (Coleman et al., 2009; Feick and Roche, 2010; Zook et al., 2010).

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 foster the generation of data of mixed quality. Whatever a body of water is a pond or a

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

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 (Girres and Touya, 2010; Ludwig et al., 2011; Neis et al., 2011). 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.

3.2 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 (Goodchild and Li, 2012). 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 (Haklay, 2010; Ludwig et al., 2011) 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 (Haklay, 2010; Neis et al., 2013).

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 (Mooney and Corcoran, 2012b). Amongst others, semantic inconsistency is one of the essential problems of VGI data quality (Elwood et al., 2012). In (Mülligann et al., 2011) and (Vandecasteele and Devillers, 2013) 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 (Mülligann et al., 2011; Vandecasteele and Devillers, 2013), as well as by ontological reasoning about the contributed information (e.g., (Schmid et al., 2012)). Another approach for tackling quality issues is the development of appropriate interfaces for the data generation and submission. In (Schmid et al., 2013a; Schmid et al., 2013b) 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 (Goodchild and Li, 2012):

- *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 (Arteaga, 2013).

- *Social measures*: this approach focuses on the assessment of the contributors themselves as a proxy measure to the quality of their contributions. (Haklay, 2010; Ludwig et al., 2011) use the number of contributors as a measure for data quality, (Neis and Zipf, 2012) analyzes the individual activity, (Mooney and Corcoran, 2012b) investigates positive and negative edits, (Barron et al., 2014) 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" (Tobler, 1970).

3.3 Managing Quality of VGI Data

A big challenge for VGI is the quality management of the contributed data because of its multidimensional heterogeneity (e.g.,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 6.2). 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 3.1 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., (Devogele et al., 1998).

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

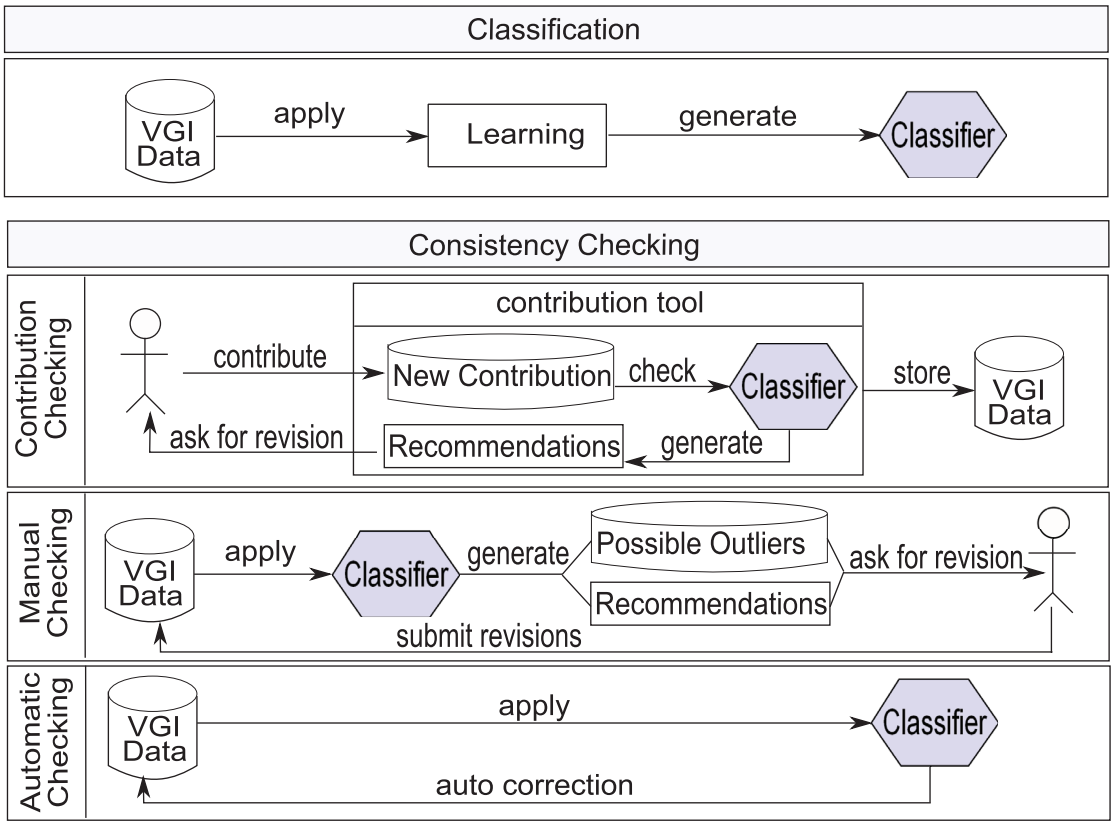


FIGURE 3.1: Proposed approaches to ensure VGI quality, see Section 3.3 for a detailed description.

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.

3.4 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 6.2). 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*.

3.5 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):

$$\forall u \in U_i \; where \; 1 \leq i \leq 11 \tag{3.1}$$

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:

$$\forall u_a \in U_{i>1}, \exists u_b \in U_{j>i} : u_a \subset u_b \tag{3.2}$$

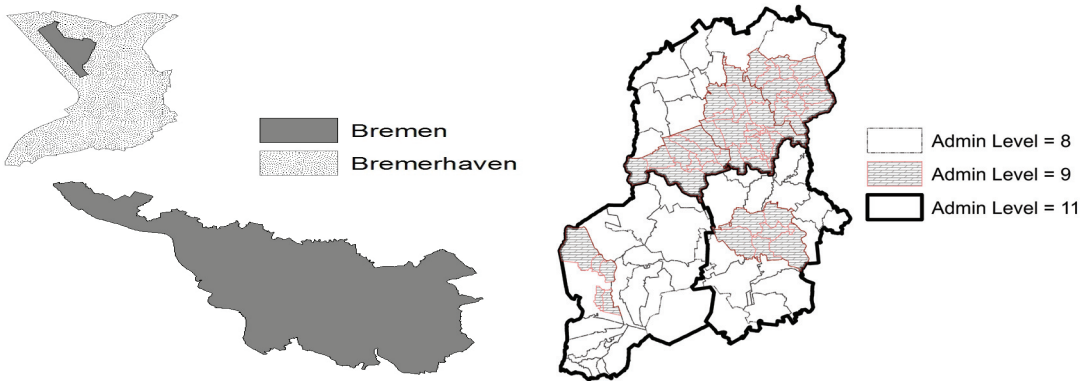
Administrative units on one level can share borders but do *not intersect* each other:

$$\forall U_j, U_k \subset U_i : U_j \cap U_k = \emptyset \tag{3.3}$$

[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 3.2a.
- *Inconsistency*: hierarchical inconsistency occurs when entities of higher administrative units are contained in units of lower levels or the same level. See Figure 3.2b
- *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.

FIGURE 3.2: 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.

3.5.1 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 3.3. We identified 14,842 duplications, 9,305 inconsistencies and 263 incorrect values.

Figure 3.2a illustrates an example for *duplication*: a part of the administrative unit representing Bremen city, is part of another unit representing Bremerhaven city. Figure 3.2b 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 3.3 would draw the attention of the contributor towards potential errors.

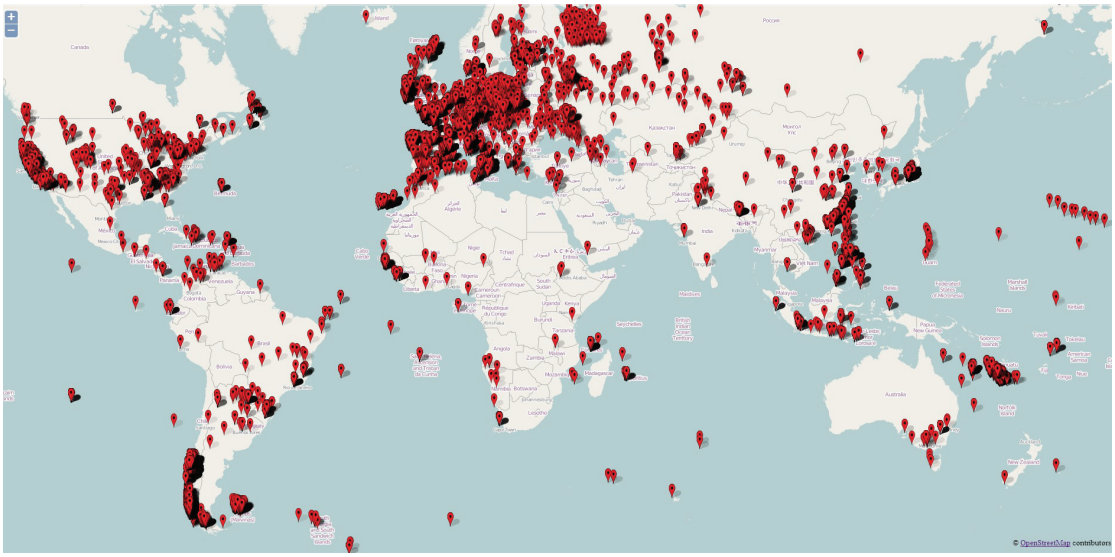


FIGURE 3.3: Distribution of potentially incorrect hierarchical classification of administrative units.

3.6 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 analysing 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 3.6 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*.

3.6.1 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 3.1, 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* (Bishop, 2006; Han et al., 2011). 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 (Bishop, 2006; Han et al., 2011).

3.6.2 Experiments and setup

As described previously, we focus on classification plausability 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 (Haklay,

2010; Ludwig et al., 2011), OSM data is of acceptable quality in Germany and the UK. In our study we use data downloaded on December 20th, 2013.

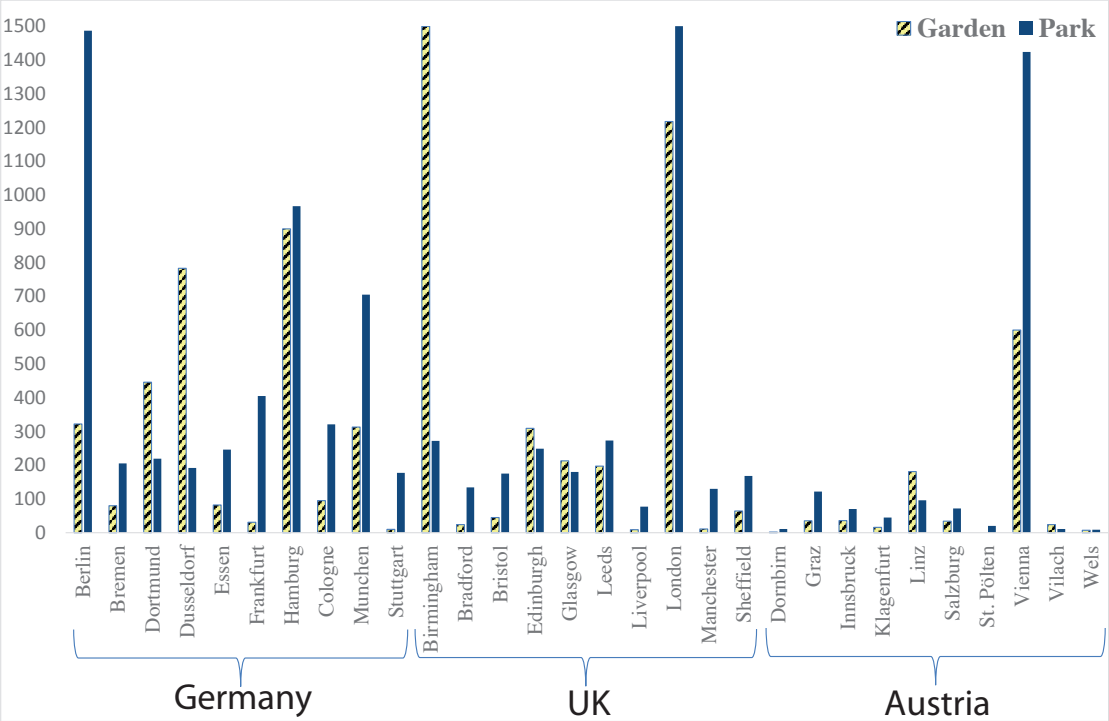
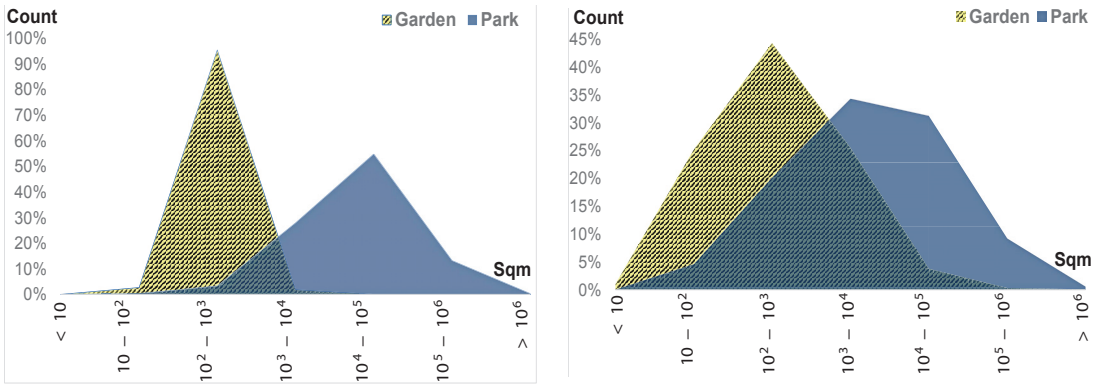


FIGURE 3.4: Number of Parks and Gardens within the selected data set.

We selected data from the ten densest (population/area) cities of each country. Figure 3.4 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 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.



(A) Birmingham. (B) London.

FIGURE 3.5: Distribution of parks and gardens areas in London and Birmingham.

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 3.5a). There are also places with a less clear separation, e.g., the case of London (see Figure 3.5b), 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 3.3, can be fed by such a classifier. Figure 3.6 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.

3.6.2.1 Feature selection

The areas of each class have a specific distribution in each city. Figure 3.6 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 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.

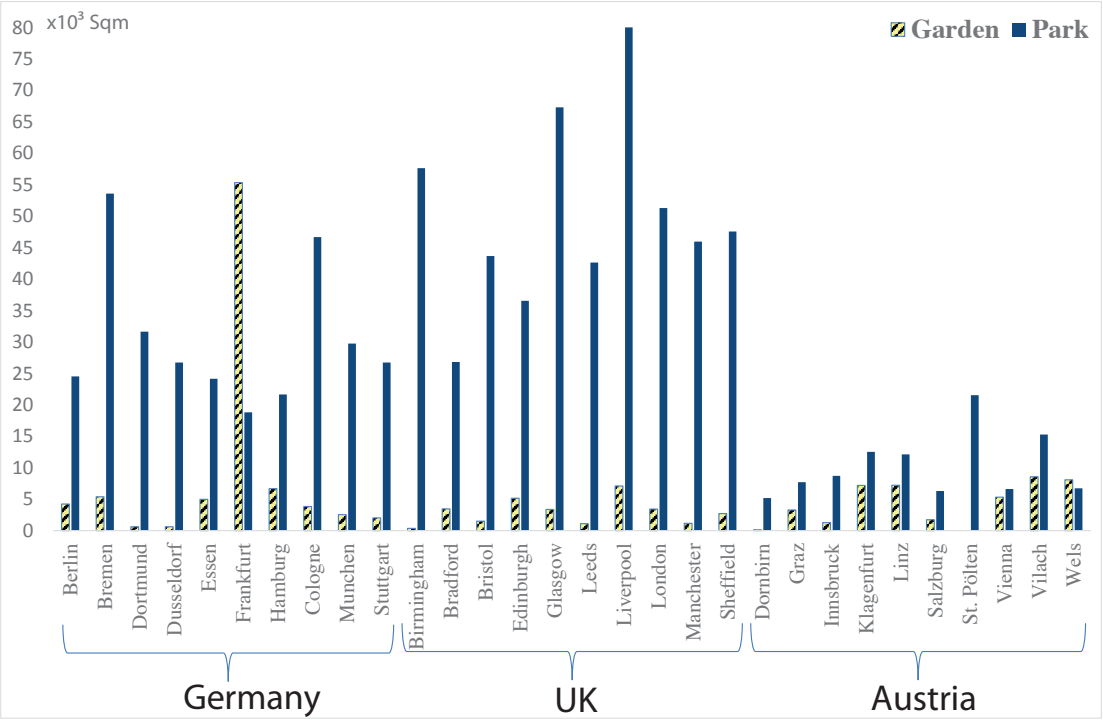


FIGURE 3.6: Mean area size of parks and gardens for the selected data set.

3.6.2.2 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 neighbours (KNN) and case based reasoning (CBR) are examples of lazy learning (Bishop, 2006; Han et al., 2011). 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 (Cover and Hart, 1967; Witten and Frank, 2005) 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 neighbours. The similarity between the unclassified entity and the training set is calculated by a similarity measure, such as euclidean distance.

	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 3.1: Classification accuracy for parks and gardens of cities in Germany.

3.6.2.3 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) (Kohavi et al., 1995) 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 classifier is measured as the average of developed classifiers (Kohavi et al., 1995). We build classifiers for each city in a country. The results can be inspected in Tables 3.1, 3.2 and 3.3. 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).

3.6.2.4 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

[3] 5 and 10 are recommended values for K

	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

TABLE 3.2: Classification accuracy for parks and gardens of cities in the UK.

	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

TABLE 3.3: Classification accuracy for parks and gardens of cities in Austria.

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 (Bishop, 2006; Han et al., 2011). 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 3.2). 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 performance. Good classifiers should have AUC value between 0.5 and 1 (Fawcett, 2006). Tables 3.1, 3.2, and 3.3 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.

3.6.2.5 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 3.1 and 3.2). 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 3.3) 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 3.4 and Table 3.3).

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 3.5) 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 entitiy 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 3.3, 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.

3.7 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 hierachical 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.

Acknowledgements

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.

Bibliography

Arteaga, M. G. (2013). "Historical map polygon and feature extractor". In: *Proceedings of ACM MapInteract, 1st International Workshop on Map Interaction*. Ed. by F. Schmid and C. Kray. ACM, pp. 66–71.

Barron, C., P. Neis, and A. Zipf (2014). "A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis". In: *Transactions in GIS* 18.6, pp. 877–895.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc. ISBN: 0387310738.

Coleman, D. J., Y. Georgiadou, J. Labonte, et al. (2009). "Volunteered Geographic Information: the nature and motivation of produsers". In: *International Journal of Spatial Data Infrastructures Research* 4.1, pp. 332–358.

Cover, T. and P. Hart (1967). "Nearest Neighbor pattern classification". In: *IEEE Transactions on Information Theory* 13.1, pp. 21–27.

Devogele, T., C. Parent, and S. Spaccapietra (1998). "On spatial database integration". In: *International Journal of Geographical Information Science* 12.4, pp. 335–352.

Elwood, S., M. F. Goodchild, and D. Z. Sui (2012). "Researching Volunteered Geographic Information: Spatial data, geographic research, and new social practice". In: *Annals of the Association of American Geographers* 102.3, pp. 571–590.

Fawcett, T. (2006). "An introduction to ROC analysis". In: *Pattern recognition letters* 27.8, pp. 861–874.

Feick, R. and S. Roche (2010). "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.

Girres, J.-F. and G. Touya (2010). "Quality assessment of the French OpenStreetMap dataset". In: *Transactions in GIS* 14.4, pp. 435–459.

Goodchild, M. F. (2007). "Citizens as sensors: the world of volunteered geography". In: *GeoJournal* 69.4, pp. 211–221.

Goodchild, M. F. and L. Li (2012). "Assuring the quality of Volunteered Geographic Information". In: *Spatial statistics* 1, pp. 110–120.

Haklay, M. (2010). "How good is Volunteered Geographic Information? A comparative study of OpenStreetMap and Ordnance Survey datasets". In: *Environment and planning. B Planning & design* 37.4, p. 682.

Han, J., M. Kamber, and J. Pei (2011). *Data Mining: Concepts and Techniques*. 3rd. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. ISBN: 0123814790, 9780123814791.

Kohavi, R. et al. (1995). "A study of Cross-validation and Bootstrap for accuracy estimation and model selection". In: *IJCAI*. Vol. 14. 2, pp. 1137–1145.

Ludwig, I., A. Voss, and M. Krause-Traudes (2011). "A Comparison of the Street Networks of Navteq and OSM in Germany". In: *Advancing Geoinformation Science for a Changing World*. Ed. by W. Geertman Stanand Reinhardt and F. Toppen. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 65–84. ISBN: 978-3-642-19789-5.

Mooney, P. and P. Corcoran (2012b). "The annotation process in OpenStreetMap". In: *Transactions in GIS* 16.4, pp. 561–579.

Mülligann, C., K. Janowicz, M. Ye, and W.-C. Lee (2011). "Analyzing the spatial-semantic interaction of points of interest in Volunteered Geographic Information". In: *Spatial information theory*. Springer, pp. 350–370.

Neis, P. and A. Zipf (2012). "Analyzing the contributor activity of a Volunteered Geographic Information project: the case of OpenStreetMap". In: *ISPRS International Journal of Geo-Information* 1.2, pp. 146–165.

Neis, P., D. Zielstra, and A. Zipf (2011). "The street network evolution of crowdsourced maps: OpenStreetMap in Germany 2007–2011". In: *Future Internet* 4.1, pp. 1–21.

Neis, P., D. Zielstra, and A. Zipf (2013). "Comparison of Volunteered Geographic Information Data Contributions and Community Development for Selected World Regions". In: *Future Internet* 5.2, pp. 282–300.

Schmid, F., O. Kutz, L. Frommberger, T. Kauppinen, and C. Cai (2012). "Intuitive and natural interfaces for geospatial data classification". In: *Workshop on place-related knowledge acquisition research (P-KAR), Kloster Seeon, Germany*, p. 26.

Schmid, F., L. Frommberger, C. Cai, and F. Dylla (2013a). "Lowering the barrier: How the What-You-See-Is-What-You-Map paradigm enables people to contribute Volunteered Geographic Information". In: *Proc. of the 4th Annual Symposium on Computing for Development*. ACM. Cape Town, South Africa, pp. 8–18.

Schmid, F., L. Frommberger, C. Cai, and C. Freksa (2013b). "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*. Ed. by D. Vandenbroucke, B. Bucher, and J. Crompvoets. Springer-Verlag, pp. 3–19.

Tobler, W. R. (1970). "A computer movie simulating urban growth in the Detroit region". In: *Economic geography* 46, pp. 234–240.

Vandecasteele, A. and R. Devillers (2013). "Improving Volunteered Geographic Data Quality Using Semantic Similarity Measurements". In: *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 1.1, pp. 143–148.

Witten, I. H. and E. Frank (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd ed. San Francisco: Morgan Kaufmann, p. 119.

Zook, M., M. Graham, T. Shelton, and S. Gorman (2010). "Volunteered Geographic Information and crowdsourcing disaster relief: a case study of the Haitian earthquake". In: *World Medical & Health Policy* 2.2, pp. 7–33. ISSN: 1948-4682.

Chapter 4

Ambiguity and Plausibility: Managing Classification Quality in Volunteered Geographic Information

Authors:

Ahmed Loai Ali, Falko Schmid, Rami Al-Salman and Tomi Kauppinen.

Conference:

The 22nd ACM SIGSPATIAL International Conference on Advanced Geographic Information Systems (SIGSPATIAL 2014).

Citation:

Ahmed Loai Ali, Falko Schmid, Rami Al-Salman, and Tomi Kauppinen (2014). "Ambiguity and plausibility: managing classification quality in Volunteered Geographic Information". In: *Proceedings of the 22nd ACM SIGSPATIAL International Conf. on Advances in Geographic Information Systems*. Ed. by Yan Huang, Markus Schneider, Michael Gertz, John Krumm, and Jagan Sankaranarayanan. New York, NY, USA: ACM, pp. 143–152. ISBN: 978-1-4503-3131-9.

Contribution Statement:

Large part of the idea and the proposed approach have been developed by me, Falko contributed by discussing the results and by giving suggestions to improve the proposed approach. Rami suggested the utilization of the topological model. Tomi supported in developing the study. I wrote the major part of the text, while the others provided reviews and comments that substantially improved the text.

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.

Keywords:

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

4.1 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) (Goodchild, 2007). 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) (Mooney and Corcoran, 2011).

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 (Haklay and Weber, 2008). OSM has no restriction about the spatial data to be contributed, and its rich data set enables numerous

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

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 (Flanagin and Metzger, 2008) 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 (Goodchild and Li, 2012).

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.

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.

4.2 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 (Elwood et al., 2012; Flanagin and Metzger, 2008).

Quality of VGI data has various perspectives and notions: completeness, positional accuracy, attribute consistency, logical consistency, and lineage (Devillers et al., 2010; Goodchild and Li, 2012). 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 (Goodchild, 2007; Goodchild and Li, 2012)). In Girres and Touya (2010), Haklay (2010), and Ludwig et al. (2011) the authors compare OSM data to reference data, in Haklay (2010) and Ludwig et al. (2011) 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 (Haklay, 2010; Neis et al., 2013). 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 (Keßler and Groot, 2013).
- *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., Barron et al. (2014) presenting 25 methods to assess VGI quality without the need for authoritative data. The methods are focused around "fitness for purpose" approach. In Keßler et al. (2011) and Neis and Zipf (2012) the authors analyze intrinsic information, such as tracking edits history,

and contributor's reputation analysis. In D'Antonio et al. (2014) and Keßler and Groot (2013) the authors use trustworthiness as a proxy to assess the quality. Mooney and Corcoran (2012b) 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 (Mooney and Corcoran, 2012b). Amongst others, semantic inconsistency is one of the essential problems of VGI data quality (Elwood et al., 2012): for instance, different classes represent the same geographic phenomena (*synonymy*), or one class describes different geographic phenomena (*polysemy*). In Mülligann et al. (2011) and Vandecasteele and Devillers (2013) 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 (Mülligann et al., 2011; Vandecasteele and Devillers, 2013). Another approach for tackling quality issues is the development of appropriate interfaces for the data generation and submission. In Schmid et al. (2013b) and Schmid et al. (2013a) the authors demonstrate that task-specific interfaces support the generation of high quality data even under difficult conditions.

4.3 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 (Devillers et al., 2010; Grira et al., 2010). Particularly in VGI, contributors are often non-experts 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 (Flanagin and Metzger, 2008). 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.

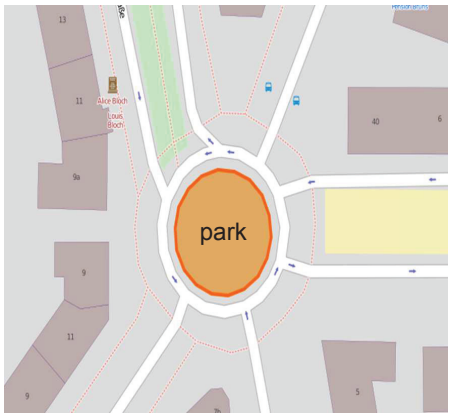


FIGURE 4.1: Inappropriate Classification: a "park" placed in a roundabout.

From other perspective, when mappers have local knowledge they contribute based on their personal perspectives (Neis and Zipf, 2012), 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 4.1 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".

4.3.1 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.

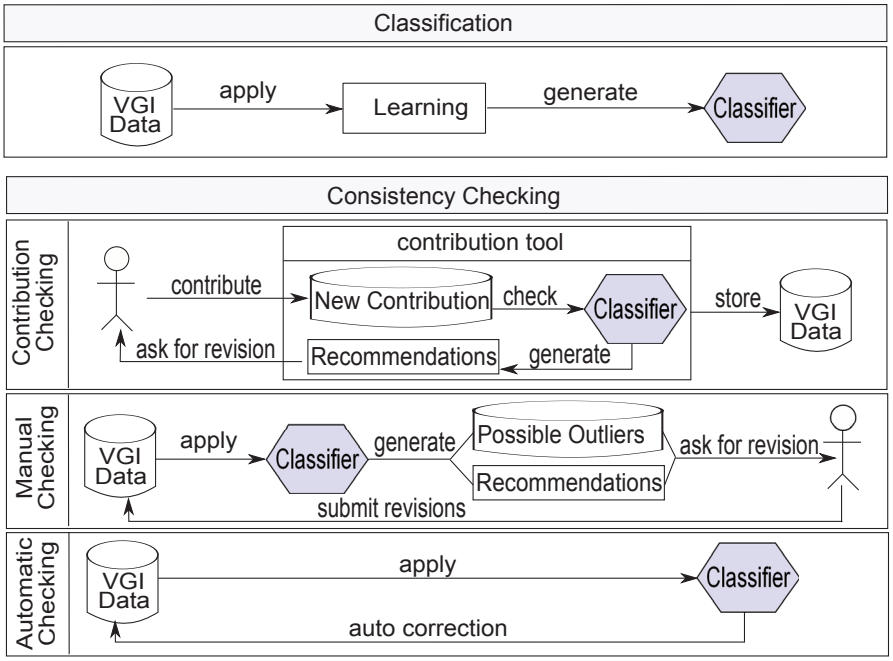


FIGURE 4.2: Learning-based approach to tackle classification plausibility.

4.4 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 4.2 illustrates the approach of using learning for quality assurance as introduced in Ali and Schmid (2014). 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 (Haklay, 2010; Ludwig et al., 2011); 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],

[2] http://maproulette.org/

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 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).

4.4.1 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 (Tobler, 1970). 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.

4.5 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.

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

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*".

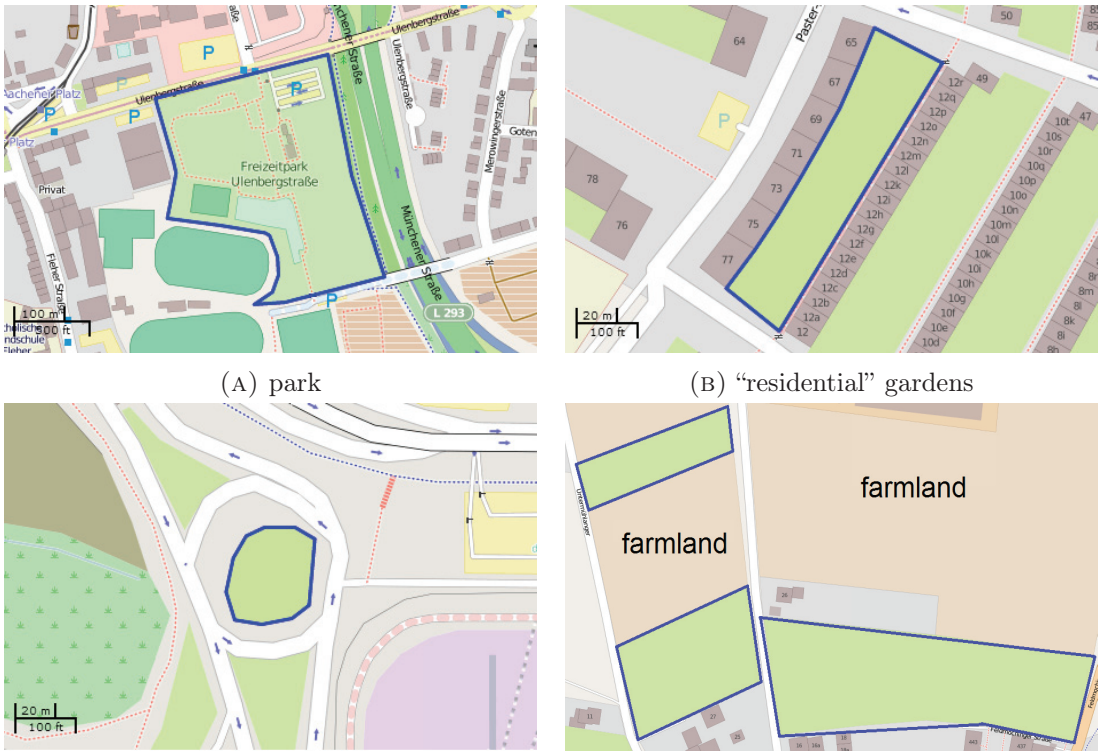
4.5.1 Selection of classification properties

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 4.3 shows typical entities of interest. Figure 4.3a depicts a "park" containing a playground, sport center, and paths. Figure 4.3b illustrates a *residential* "garden" surrounded by *residential* houses. Figure 4.3c shows a typical "grass" entity not containing other infrastructural entities and usually surrounded by or meet roads. Figure 4.3d shows "meadow" entities next to farmland and not containing other infrastructural entities.

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 Ali and Schmid (2014), we developed classifiers based on geometric properties to distinguish between entities of the classes "park" and "garden". This property is also observable in Figure 4.3: parks are usually larger than gardens. However, building classifiers for multiple classes requires the analysis of more

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



(A) park (B) "residential" gardens

(C) grass (D) meadow

FIGURE 4.3: Samples of typical entities of interest.

properties, as size of entities can be similar, but their characteristics might be fundamentally different.

4.5.1.1 Geometric properties: size

Some entities are classifiable by considering their size. Figure 4.4 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.

4.5.1.2 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 4.3 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

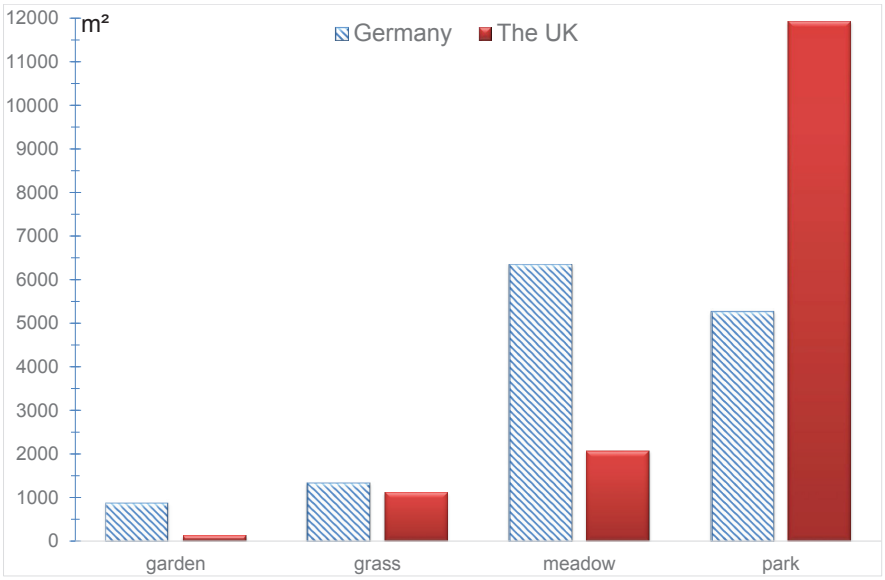


FIGURE 4.4: Average areas for the classes "garden", "grass", "meadow", "park" in Germany and the UK.

"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) (Egenhofer, 1995). As depicted in Figure 4.5, the 9IM distinguishes eight topological relations holding between two regions: *equal*, *disjoint*, *meet*, *overlap*, *contains*, *covers*, *inside*, and *coveredBy*.

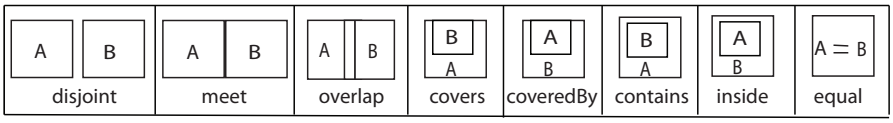


FIGURE 4.5: The eight distinct topological relations of the 9-intersection model.

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.

4.5.1.3 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 4.5.1.1) and spatial context properties (Sections 4.5.1.2 and 4.5.1.3) 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.

4.5.2 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 (Bishop, 2006).

4.5.2.1 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 (Bishop, 2006; Han et al., 2011). 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 (Cover and Hart, 1967; Witten and Frank, 2005) 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.

4.5.2.2 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 (Bishop, 2006). 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 (Bishop, 2006; Han et al., 2011). A reason can be an unbalanced population of the training or the test set. For example, Figure 4.6 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 (Fawcett, 2006; Witten and Frank, 2005). 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 Fawcett (2006), good classifiers should have AUC value between 0.5 and 1.0.

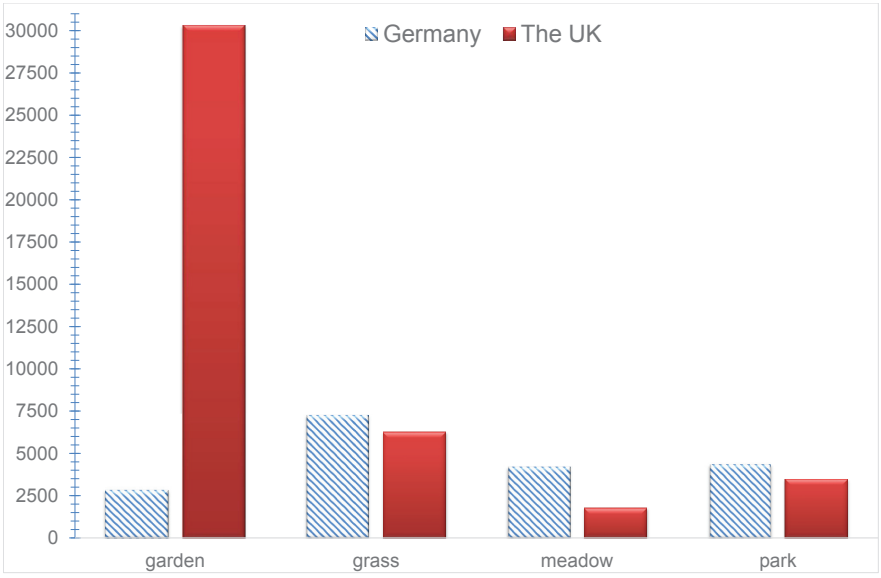


FIGURE 4.6: Number of "garden", "grass" ,"meadow" ,"park" entities in Germany and the UK.

4.6 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 (Ludwig et al., 2011), and (Haklay and Weber, 2008), OSM data for Germany and the UK is of acceptable quality.

4.6.1 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 (Mooney and Corcoran, 2012b; Neis and Zipf, 2012), 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 (Neis and Zipf, 2012): *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.*

Table 4.1 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.

	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 4.1: Extracted data from Germany and the UK.

4.6.2 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 4.2 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 4.2: LBM classifiers performance of data extracted from Germany (GER) and the UK.

From Table 4.2, 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 4.6 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 4.3.1, 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 4.3 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 4.3: TBM classifiers performance of data extracted from Germany (GER) and the UK.

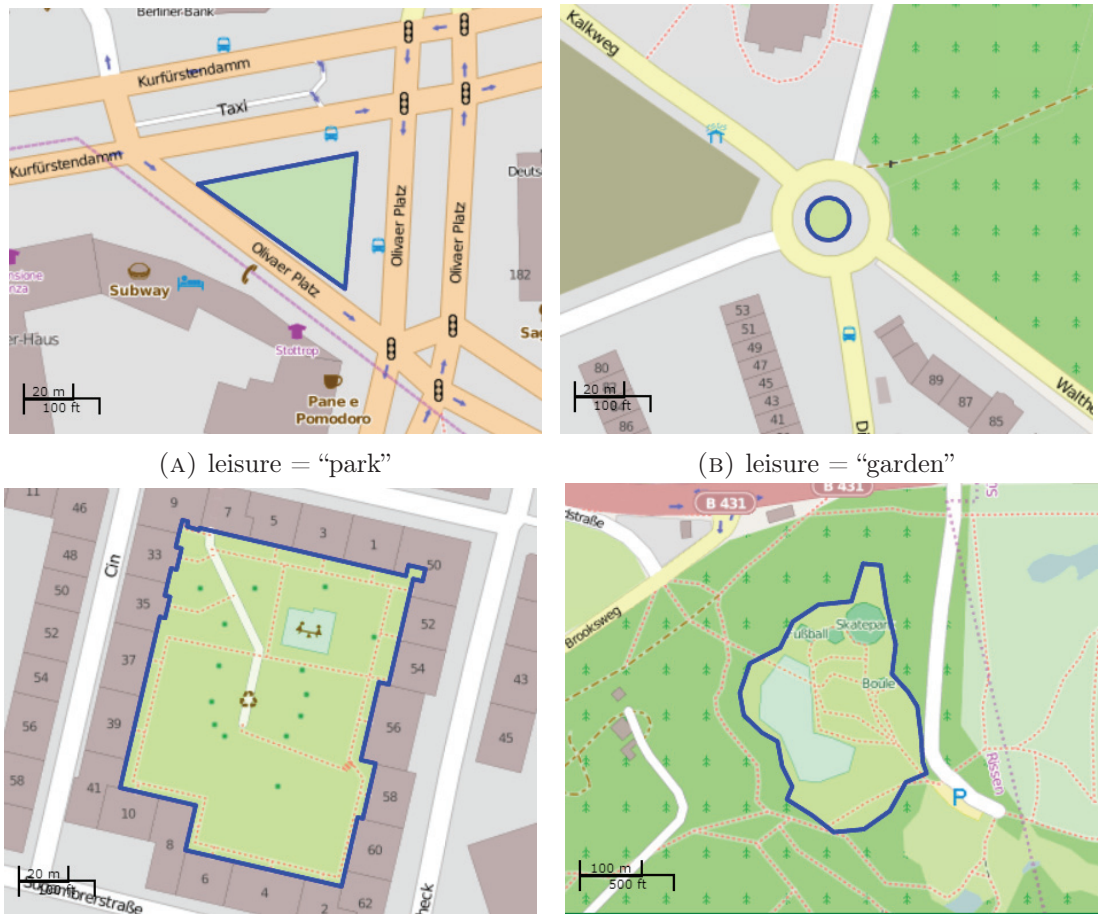
Table 4.3 conveys that the classifiers of *TBM* have higher performance than the classifiers of *LBM*. According to the table, the classifier based on the data set of Germany has an average performance with accuracy of 76.8% and AUC equal to 0.85, whereas the classifier based on the UK data set has an average performance by 89.0% accuracy and AUC equal 0.92.

4.6.3 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.

Figure 4.7 shows a sample of detected entities with potentially inappropriate classification. Figures 4.7a and 4.7b 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 4.7c and 4.7d 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 4.7c is surrounded by houses and contains a playground. The entity in Figure 4.7d 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



(A) leisure = "park"

(B) leisure = "garden"

(C) landuse = "grass"

(D) landuse = "meadow"

FIGURE 4.7: Samples of entities with potentially inappropriate classification.

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

4.7 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

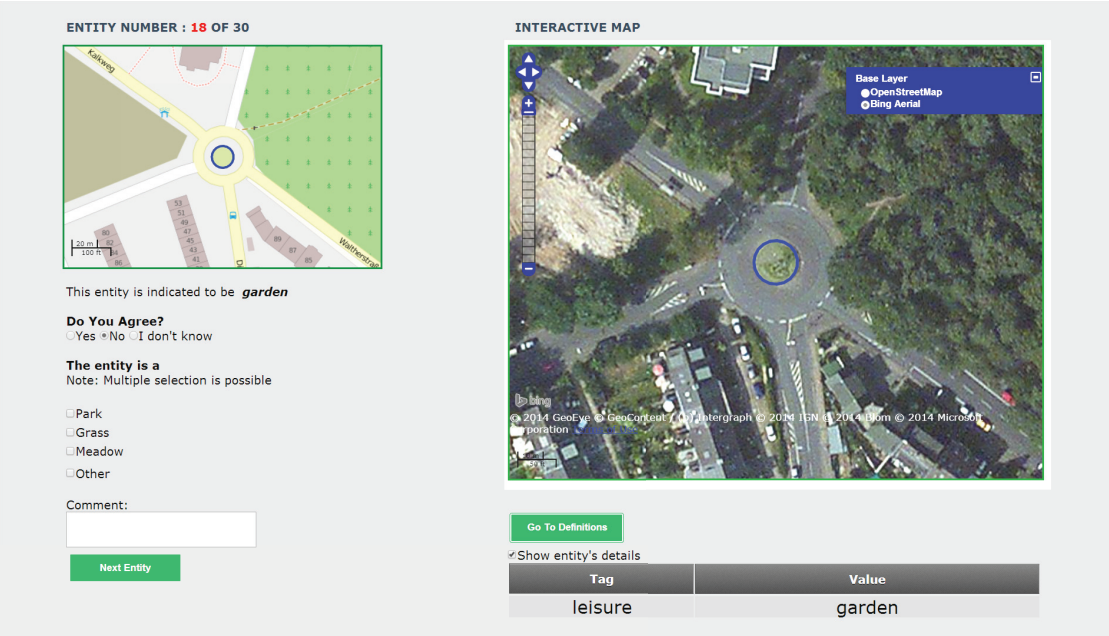


FIGURE 4.8: A snapshot from the website of the study.

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 4.8 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 (Light, 1971) 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

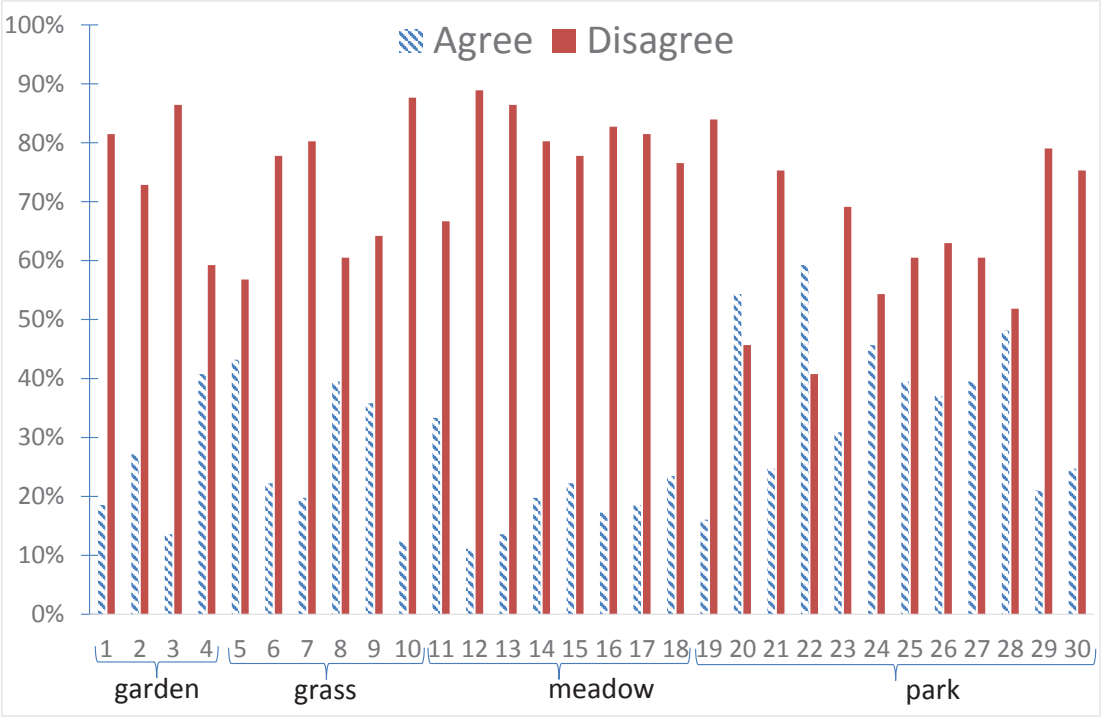


FIGURE 4.9: The percentage of total agreement and disagreement of the participants on the current classifications per entity.

range from 0.01 to 1.0 is divided into slight, fair, moderate, and substantial agreements (Viera, Garrett, et al., 2005).

Light's Kappa for all 81 participants was 0.176, meaning thus *a slight agreement*. We analyzed 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 of the 81 participants about the current classifications of entities. Figure 4.9 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".

4.7.1 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 4.5.1 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.

4.8 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.

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 Geo-Crowd. 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.

Bibliography

Ali, A. L. and F. Schmid (2014). "Data quality assurance for Volunteered Geographic Information". In: *Geographic Information Science: 8th International Conference, GIScience 2014, Vienna, Austria, September 24-26, 2014. Proceedings*. Ed. by M. Duckham, E. Pebesma, K. Stewart, and A. U. Frank. Cham: Springer International Publishing, pp. 126–141. ISBN: 978-3-319-11593-1.

Barron, C., P. Neis, and A. Zipf (2014). "A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis". In: *Transactions in GIS* 18.6, pp. 877–895.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc. ISBN: 0387310738.

Cover, T. and P. Hart (1967). "Nearest Neighbor pattern classification". In: *IEEE Transactions on Information Theory* 13.1, pp. 21–27.

D'Antonio, F., P. Fogliaroni, and T. Kauppinen (2014). "VGI Edit History Reveals Data Trustworthiness and User Reputation". In: *Connecting a Digital Europe Through Location and Place*. Ed. by J. Huerta, S. Schade, and C. Granell. Springer-Verlag.

Devillers, R., A. Stein, Y. Bédard, N. Chrisman, P. Fisher, and W. Shi (2010). "Thirty years of research on spatial data quality: achievements, failures, and opportunities". In: *Transactions in GIS* 14.4, pp. 387–400.

Egenhofer, M. J. (1995). "On the Equivalence of Topological Relations". In: *International Journal of Geographical Information Systems* 9, pp. 133–152.

Elwood, S., M. F. Goodchild, and D. Z. Sui (2012). "Researching Volunteered Geographic Information: Spatial data, geographic research, and new social practice". In: *Annals of the Association of American Geographers* 102.3, pp. 571–590.

Fawcett, T. (2006). "An introduction to ROC analysis". In: *Pattern recognition letters* 27.8, pp. 861–874.

Flanagin, A. J. and M. J. Metzger (2008). "The credibility of Volunteered Geographic Information". In: *GeoJournal* 72.3, pp. 137–148.

Girres, J.-F. and G. Touya (2010). "Quality assessment of the French OpenStreetMap dataset". In: *Transactions in GIS* 14.4, pp. 435–459.

Goodchild, M. F. (2007). "Citizens as sensors: the world of volunteered geography". In: *GeoJournal* 69.4, pp. 211–221.

Goodchild, M. F. and L. Li (2012). "Assuring the quality of Volunteered Geographic Information". In: *Spatial statistics* 1, pp. 110–120.

Grira, J., Y. Bédard, and S Roche (2010). "Spatial data uncertainty in the VGI world: Going from consumer to producer". In: *Geomatica* 64.1, pp. 61–72.

Haklay, M. (2010). "How good is Volunteered Geographic Information? A comparative study of OpenStreetMap and Ordnance Survey datasets". In: *Environment and planning. B Planning & design* 37.4, p. 682.

Haklay, M. and P. Weber (2008). "OpenStreetMap: user-generated street maps". In: *IEEE Pervasive Computing* 7.4, pp. 12–18. ISSN: 1536-1268.

Han, J., M. Kamber, and J. Pei (2011). *Data Mining: Concepts and Techniques*. 3rd. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. ISBN: 0123814790, 9780123814791.

Keßler, C. and R. T. A. de Groot (2013). "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*. Ed. by D. Vandenbroucke, B. Bucher, and J. Crompvoets. Springer-Verlag, pp. 21–37.

Keßler, C., J. Trame, and T. Kauppinen (2011). "Tracking editing processes in Volunteered Geographic Information: the case of OpenStreetMap". In: *Proceedings of Workshop on Identifying objects, processes and events in spatio-temporally distributed data (IOPE), Conference on Spatial Information Theory (COSIT 2011)*. Vol. 12.

Light, R. (1971). "Measures of response agreement for qualitative data: Some generalizations and alternatives". In: *Psychological Bulletin* 76, pp. 365–377.

Ludwig, I., A. Voss, and M. Krause-Traudes (2011). "A Comparison of the Street Networks of Navteq and OSM in Germany". In: *Advancing Geoinformation Science for a Changing World*. Ed. by W. Geertman Stanand Reinhardt and F. Toppen. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 65–84. ISBN: 978-3-642-19789-5.

Mooney, P. and P. Corcoran (2011). "Can Volunteered Geographic Information Be a Participant in eEnvironment and SDI?" In: *Environmental Software Systems. Frameworks of eEnvironment*. Springer, pp. 115–122.

Mooney, P. and P. Corcoran (2012b). "The annotation process in OpenStreetMap". In: *Transactions in GIS* 16.4, pp. 561–579.

Mülligann, C., K. Janowicz, M. Ye, and W.-C. Lee (2011). "Analyzing the spatial-semantic interaction of points of interest in Volunteered Geographic Information". In: *Spatial information theory*. Springer, pp. 350–370.

Neis, P. and A. Zipf (2012). "Analyzing the contributor activity of a Volunteered Geographic Information project: the case of OpenStreetMap". In: *ISPRS International Journal of Geo-Information* 1.2, pp. 146–165.

Neis, P., D. Zielstra, and A. Zipf (2013). "Comparison of Volunteered Geographic Information Data Contributions and Community Development for Selected World Regions". In: *Future Internet* 5.2, pp. 282–300.

Schmid, F., L. Frommberger, C. Cai, and F. Dylla (2013a). "Lowering the barrier: How the What-You-See-Is-What-You-Map paradigm enables people to contribute Volunteered Geographic Information". In: *Proc. of the 4th Annual Symposium on Computing for Development*. ACM. Cape Town, South Africa, pp. 8–18.

Schmid, F., L. Frommberger, C. Cai, and C. Freksa (2013b). "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*. Ed. by D. Vandenbroucke, B. Bucher, and J. Crompvoets. Springer-Verlag, pp. 3–19.

Tobler, W. R. (1970). "A computer movie simulating urban growth in the Detroit region". In: *Economic geography* 46, pp. 234–240.

Vandecasteele, A. and R. Devillers (2013). "Improving Volunteered Geographic Data Quality Using Semantic Similarity Measurements". In: *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 1.1, pp. 143–148.

Viera, A. J., J. M. Garrett, et al. (2005). "Understanding interobserver agreement: the kappa statistic". In: *Fam Med* 37.5, pp. 360–363.

Witten, I. H. and E. Frank (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd ed. San Francisco: Morgan Kaufmann, p. 119.

Chapter 5

Rule-Guided Human Classification of Volunteered Geographic Information

Authors:

Ahmed Loai Ali, Zoe Falomir, Falko Schmid, and Christian Freksa.

Journal:

ISPRS Journal of Photogrammetry and Remote Sensing.

Citation:

Ahmed Loai Ali, Zoe Falomir, Falko Schmid, and Christian Freksa (2016). In: *ISPRS International Journal of Photogrammetry and Remote Sensing*, ISSN:0924-2716.

Status:

The article is submitted on Januray 2016 and accepted with minor revisions on April 2016. On June 2016, the article is finally accepted and published online.

Contribution Statement:

This article is an extended version of research published in the International Symposium of Spatial Data Quality (ISSDQ-2015). I contributed the main idea of this work, by developing a rule-based guiding system for VGI mapping projects. I involved strongly in applying learning algorithms and in conducting the empirical study. Zoe and Falko contributed in discussing the approach and the findings. Christian contributed by reviewing and discussing the manuscript.

Abstract:

During the last decade, web technologies and location sensing devices have evolved generating a form of crowdsourcing known as Volunteered Geographic Information (VGI). VGI acted as a platform of spatial data collection, in particular, when a group of public participants are involved in collaborative mapping activities: they work together to collect, share, and use information about geographic features. VGI exploits participants' local knowledge to produce rich data sources. However, the resulting data inherits problematic data classification. In VGI projects, the challenges of data classification are due to the following: i) data is likely prone to subjective classification, ii) remote contributions and flexible contribution mechanisms in most projects, and iii) the uncertainty of spatial data and non-strict definitions of geographic features. These facts lead to various forms of problematic classification: inconsistent, incomplete, and imprecise data classification. This research addresses classification appropriateness. Whether the classification of an entity is appropriate or inappropriate is related to quantitative and/or qualitative observations. These observations – in most cases – may be not recognizable particularly for non-expert participants. Hence, in this paper, the problem is tackled by developing a rule-guided classification approach. This approach exploits data mining techniques of Association Classification (AC) to extract descriptive (qualitative) rules of specific geographic features. The rules are extracted based on the investigation of qualitative topological relations between target features and their context. Afterwards, the extracted rules are used to develop a recommendation system able to guide participants to the most appropriate classification. The approach proposes two scenarios to guide participants towards enhancing the quality of data classification. An empirical study is conducted to investigate the classification of grass-related features like *forest, garden, park*, and *meadow*. The findings of this study indicate the feasibility of the proposed approach.

Keywords:

Volunteered Geographic Information (VGI), Spatial Data Quality, Spatial Data Mining, Classification, Topology, Qualitative Spatial Reasoning

5.1 Introduction

The advanced technologies of Web 2.0, geo-tagging, geo-referencing, Global Navigation Satellite System (GNSS), and broadband communication enable the public to generate spatial content known as User Generated Geographic Content (UGGC) (Goodchild, 2008). They empower ordinary citizens to participate in mapping activities producing geo-spatial content, such activities were formerly conducted by mapping agencies and

professional organizations. This trend results in evolving a form of crowdsourcing data known as Volunteered Geographic Information (VGI) (Goodchild, 2007). In this research, we are concerned with the form of VGI, when a group of participants collaboratively work to collect, share, update, and use information about geographic features. Among others, OpenStreetMap[1](OSM), Google Map Maker[2] and Wikimapia[3] are examples of collaborative mapping projects which aim to produce a digital map of the world. During the last decade, VGI played a significant role in the GIScience community. Various applications and services have been developed based on VGI data sources including – but not limited to – environmental monitoring (Gouveia and Fonseca, 2008), crisis management (Roche et al., 2013), urban planning (Foth et al., 2009; Song and Sun, 2010), land use mapping (Arsanjani et al., 2015), and mapping provision (Haklay and Weber, 2008). Moreover, VGI acted as a means of geographic data collection and as a complementary component of spatial data infrastructure (SDI)(McDougall, 2009).

However, the dramatic increase of VGI triggers questions about the resulting data quality (Flanagin and Metzger, 2008; Elwood et al., 2012). Among other things, the lack of detailed information about data quality and the difficulty of applying the conventional spatial quality measures are key reasons behind its questionable quality (Flanagin and Metzger, 2008; Elwood et al., 2012). Generally, multiple measures are used to describe the quality of spatial data from different perspectives, such as completeness, positional accuracy, thematic accuracy, logical consistency, and lineage. However, this paper addresses the quality from the perspective of data classification.

In a VGI context, the classification of data faces various challenges. On one hand, a large amount of data is contributed by arm-chair participants based on their local knowledge. This remote contribution method results in imprecise classification. On the other hand, human observations generate subjective data classification. Whether a water body is classified in VGI as *pond* or *lake*, depends on the participant's perceptions. In contrast, in the professional field, a strictly defined classification model is developed by experts in advance, and then data is classified according to measures and observations in comparison to the predefined model. Hence, remote contributions and subjective perceptions, among other reasons, produce problematic data classification, and consequently, difficulties for data integration and utilization.

For example, Figure 5.1 shows one of the common interfaces (iD editor) of OSM project,

[1] www.openstreetmap.org
[2] www.google.com/mapmaker
[3] www.wikimapia.org

where participants can edit geographic features using the appropriate geometric representation (point, line, or polygon) by tracking over satellite images provided by Bing[4]. Afterwards, they describe (classify) the sketched entity using tags (see Section 5.3.2).

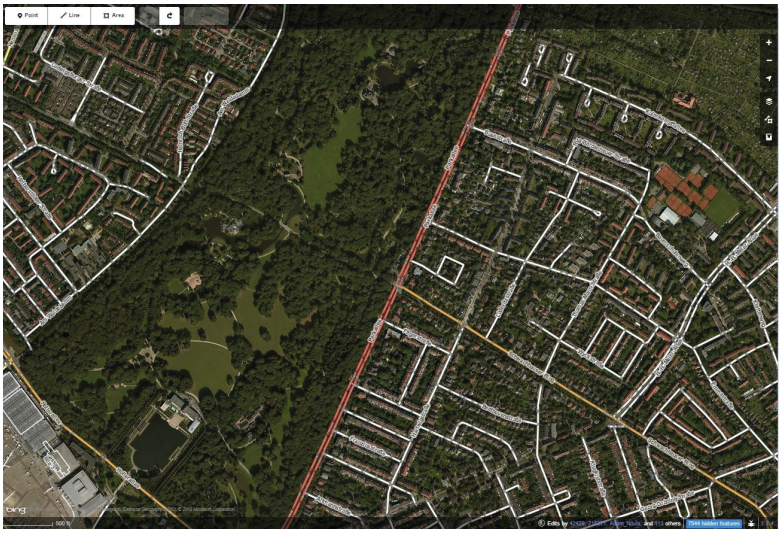


FIGURE 5.1: Example of an editing interface in OSM project (iD editor).

Whether this piece of land covered by grass – in the middle of Figure 5.1 – is classified as *park, garden, meadow*, or generally *grass*, is not strictly defined. The human-centered classification generates multiple acceptable class labels with higher or lower degrees of appropriateness. The given entity can be recognized by a participant as *park*, even if it has been classified by others as *garden* or *forest*. The most appropriate classification of an entity is related to qualitative and/or quantitative observations. Small difference in observations might lead to different classification. These differences might be not recognizable by voluntary participants. Hence, this paper presents a rule-guided classification approach to tackle the classification problems of VGI.

The proposed approach exploits the dramatic increase of VGI towards enhancing data classification. It consists of two phases: *Learning* and *Guiding* phases. During the *Learning* phase, the task is to learn the qualitative characteristics that distinguish among similar classes. This task exclusively investigates qualitative topological characteristics of specific classes. The extracted characteristics are formulated into descriptive qualitative rules able to guide the participants towards the most appropriate classification. The *Guiding* phase presents two scenarios for applying the generated guidance and recommendations.

To validate the proposed approach, an empirical study has been conducted addressing the classification of grass-related features. Classes of *forest, garden, grass, meadow, park*, and *wood* are selected for the study. The classification of these features represents

[4] www.bing.com/maps

a challenge; they are commonly covered by grass, although each class has its unique features. For example, the classes *park* and *garden* have entertainment characteristics, *forest* and *wood* are usually covered with trees or other woody vegetation, *meadow* has agriculture characteristics, etc. The findings indicate the feasibility of the proposed approach. Specifically, the developed system is able to unambiguously classify some of the target classes, while other classes still have poor classification accuracy.

This paper is organized as follows. Section 5.2 presents a review of VGI data quality. Section 5.3 gives insight into the fundamental reasons behind the problematic classification of VGI. Section 5.4 presents an overview of the qualitative spatial reasoning field, which provides intuitive and well-defined semantics from spatial quantitative data. Section 5.5 presents the proposed approach and its phases and Section 5.6 describes the empirical study carried out. Section 5.7 envisions the application of the presented approach in emerging GIS trends. The last section concludes the findings and points to future work.

5.2 Issues of VGI Data Quality

In VGI, humans are the fundamental source of data. Particularly in collaborative mapping projects, participants record their observations by collecting, updating, and sharing information about geographic features. VGI employs participants' local knowledge and their willingness to contribute in order to produce rich spatial data sources (Goodchild, 2007). But the quality of the resulting data is questionable. With increasing utilization of VGI in GIScience research, data quality becomes a concern of highest priority (Flanagin and Metzger, 2008; Elwood et al., 2012). Various methods to assess data intrinsically or extrinsically can be found in the literature (Section 5.2.1), also methodologies/approaches to improve data quality (Section 5.2.2), whereas there is only a limited number of research that addresses data classification problems (Section 5.2.3).

5.2.1 Extrinsic and intrinsic data assessment

Generally, VGI is evaluated by following either extrinsic or intrinsic procedures. In the extrinsic procedure, with the availability of ground-truth data, the VGI data set is compared with a comparable ground-truth data source. Girres and Touya (2010), Haklay (2010), Neis et al. (2011), and Jackson et al. (2013) compared OSM data against ground-truth data sources in France, UK, Germany, and USA, respectively. They emphasized the quality of VGI data particularly in urban areas. In Hecht and Stephens (2014), authors found that VGI data quality decreases with increased distance from urban areas.

In the intrinsic procedure, comparable data sources are not available. The data is assessed by analyzing its intrinsic properties like participants' mapping activities, data development, and participants' reputation. Goodchild and Li (2012) presented three dimensions that could be followed to ensure VGI quality intrinsically: the crowdsourcing, social, and geographic dimensions. Bishr and Kuhn (2007), Keßler et al. (2011), Neis et al. (2011), Mooney and Corcoran (2012a), and Barron et al. (2014) assessed VGI data intrinsically. They analyzed meta-data of VGI like contributors' mapping activities and reputation, editing history of entities, etc. Neis et al. (2013) compared the development of contributors' communities in different cities around the world indicating the relation between the communities and data quality. Moreover, the nature of VGI results in new intrinsic measures of data quality like fitness of use and conceptual quality. Barron et al. (2014) developed 25 intrinsic measures that fit specific purposes of use. Ballatore and Zipf (2015) proposed a framework that assesses VGI conceptually.

5.2.2 Towards enhanced data quality

In an attempt to improve data quality, Pourabdollah et al. (2013) conflated VGI data with authoritative data to enrich the data source. Vandecasteele and Devillers (2013) provided a semantic solution to guide contributors during the editing process aiming to improve the semantic data quality. Moreover, Schmid et al. (2013a) argued a task-specific interface approach toward acquiring higher data quality even in harsh conditions.

In previous research, we presented the approach of guided classification in (Ali and Schmid, 2014) and then we enhanced it to detect problematic classifications of VGI (Ali et al., 2014). The introduction of rule-guided classification was originally presented in Ali et al. (2015), and the current paper extends this work to discuss all aspects and complications of this approach in more detail.

5.2.3 Data classification in VGI

Regarding the problematic data classification of VGI, Sparks et al. (2015) highlighted the ability of volunteers to give precise classification of land cover features given different sources of information like aerial and ground-based photos. Klippel et al. (2015) addressed the influence of cultural, linguistic, and regional factors on the classification consistency of VGI and concluded the need for statistical grouping methods that allow to identify relevant semantic contexts. Foody et al. (2015) assessed the classification quality of VGI with a reference to the contributors and the data that they provided using a statistical model (e.g., Laten Class Model). Arsanjani et al. (2015) conducted a comprehensive assessment of Land Cover and Land Use (LC/LU) classification on OSM

data sets. Arsanjani et al. (2015) and Dorn et al. (2015) concluded the promising of OSM as a source of LC/LU maps with an acceptable level of classification quality, and completeness as well.

To our knowledge, a limited number of research focuses on improving the data quality by guiding the participants. On one hand, researchers argue that developing data with limited quality is better than having no data at all; on the other hand, other researches find that free contribution mechanisms encourage participants to express what they actually observe, generating multi-dimensional data sources. However, in the present research, we aim to adapt contribution mechanisms to guide participants through an implicit approach as well as to support multiple classification for overlapping feature categories.

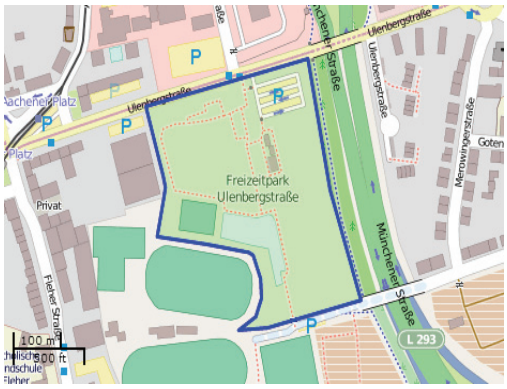
5.3 The Problematic Data Classification in VGI

In general, the uncertainty of spatial data results in different formats of errors. Based on whether a geographic feature is well or poor defined, errors are classified into ambiguity, vagueness, and probabilistic errors (Fisher, 1999). Moreover, most geographic features are not strictly defined. These facts lead to problematic data classification in VGI. In particular, a single geographic feature can be described by multiple acceptable labels, with various degrees of accuracy. This can be conceptualized by overlapping categories, for example between *park* and *garden*, *lake* and *pond*, or *swamp* and *marsh*. However, the characteristics of a geographic feature could be exploited to distinguish between these overlapping categories. In addition, the flexibility of classification mechanisms and the absence of integrity checking can result in heterogeneous data classification. In most VGI projects, contributors are heterogeneous, i.e., they have various levels of geographic and cartographic knowledge, and come from diverse cultures and educational backgrounds. These issues generate human-centered classification of data. Whether a piece of grass-covered land is classified as *park* or *meadow*, is highly determined by participants' perception. While in fact, the appropriate classification of a feature is related to quantitative observations and/or qualitative measures of its context.

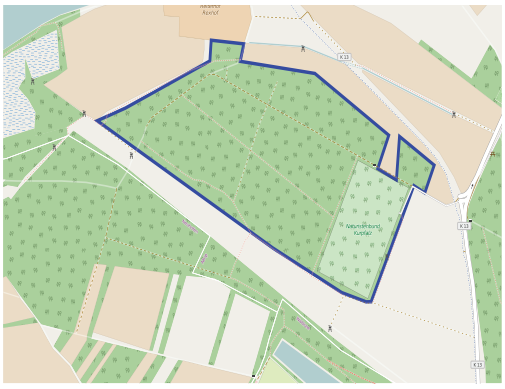
The following sections discuss the classification appropriateness (Section 5.3.1) and the classification ambiguity exemplified in grass-related features (Section 5.3.2).

5.3.1 Appropriate and inappropriate classification

In this paper, an *appropriate classification* is defined as assigning a given entity a class label that highly reflects its intrinsic and extrinsic characteristics and matches its geographic context. Figures 5.2 and 5.3 illustrate the terms of appropriate and inappropriate classification, respectively.



(A) entity classified as *park*; it contains playground/sport centers, is paved by footpaths, and located within an urban area.

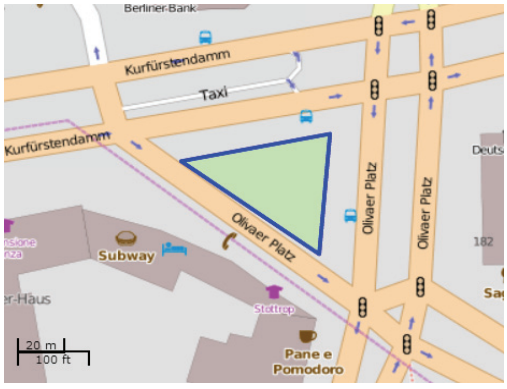


(B) entity classified as *forest*; it contains woody plants, is paved by tracks, and located beside farmland.

FIGURE 5.2: Examples of *appropriate* classification.

In Figure 5.2a, the selected entity contains some amusement facilities such as a playground, sport centers, and accessibility for walking. This entity is classified as *park*, which typically reflects the characteristics of the entity. While in Figure 5.2b, the selected entity is classified as *forest*, since it is covered by woody plants and is located in non-urban area next to farmland. Here, *park* and *forest* class labels are examples of appropriate classification with respect to the context and characteristics of the entities.

In contrast, in Figure 5.3, the selected entities represent small pieces of land covered by grass. In Figure 5.3a, the entity contains no infrastructure and is located beside road connections. This entity is misclassified as *park*, because it lacks amusement and entertainment characteristics. The same appears in Figure 5.3b, the selected entity is located within a school and does not have the proper characteristics of a *forest*. Note that, in these scenarios, both entities are misclassified as *park* and *forest*, respectively.



(A) the entity is misclassified as *park*; it contains no infrastructure, and located between roundabouts.

(B) the entity is misclassified as *forest*; it is located within a school, and relatively surrounded by non-forest characteristics.

FIGURE 5.3: Examples of *inappropriate* classification.

However, the entities may be classified appropriately as *grass*, the label that describes their general characteristics of land cover.

Hence, as indicated in the examples, the qualitative characteristics significantly influence the classification appropriateness. Thus, we exploit these characteristics to guide the participants towards an appropriate data classification.

5.3.2 Grass-related classification ambiguity

As a case study, we address the classification of grass-covered land. A piece of land covered by grass could be classified as *garden*, *forest*, *park*, *meadow*, or even generally as *grass*. These classes represent a sample among other potential classes (e.g., *recreation ground*, *scrubs*). Our previous study (Ali et al., 2014) demonstrates how contributors are unlikely to agree between themselves on a certain class for a given set of entities. The participants of the study typically reflect the nature of VGI contributors: diversity of age, gender, culture, education, and geographic knowledge. The findings indicate the following: (i) difficulties in classifying such entities; (ii) a massive need for having multiple classes for some entities; and (iii) the demand for a guided classification approach. During remote classification, it is difficult, even for experts, to recognize the intrinsic characteristics of an entity to assign the most appropriate class. Thus recommendations and guides are both required particularly for non-expert contributors, which represent the majority in VGI projects.

We utilize OSM data, as a prominent example of a VGI project. In OSM, the classification is done by means of tags in form of $key = value$, where the key represents a classification perspective and the $value$ represents a class of that perspective. For example, tag $leisure = park$ the key $leisure$ is associated with the set of entities that are used for entertainment purposes, while $park$ represents one class label between others like *garden*, *pitch*, *recreation*, etc. There are no restrictions on the number of tags that are associated with an entity; each entity could be related to no tags or several tags with arbitrary combinations of tags (Mooney and Corcoran, 2012b). At the same time, OSM provides only recommendations of tags based on discussions between mapper communities. However, most contributors do not spend enough time to check the given recommendations. Moreover, particularly for non-experts, some recommendations might be conceptually misinterpreted (e.g., *wood* or *forest* and *landuse* or *landcover*).

5.4 Qualitative Spatial Reasoning and Geospatial Information

Qualitative Spatial Representation and Reasoning (QSR) (Guesgen, 1989; Cohn and Renz, 2007; Ligozat, 2011) deals with modeling and reasoning about properties of *space* (i.e. topology, location, direction, proximity, geometry, intersection, etc). QSR models avoid the high computational cost of managing all the quantitative information which can be gathered from space; instead, they identify the qualitative spatial relations/properties which are important for a particular problem. These relations are usually modeled as disjoint but continuous, so that they can identify the important changes taking place in space (i.e. *North*, *West*, *South*, *East*), and in this way, reason about it more intuitively. Maintaining the consistency in space and time are the basics in qualitative reasoning when solving spatial and temporal problems. And for that, the evolution of relations between continuous conceptually neighbouring situations (Freksa, 1991) is studied.

QSR models can deal with imprecise and incomplete data on a symbolic level since qualitative labels (i.e. *close, far, in, touching*) include already a margin for uncertainty and can be defined even if part of the numerical data is not known. Moreover, QSR models help in human-machine interaction because they align human cognitive concepts with numerical perception of computational systems. Another advantage of a description based on qualitative relations is also that semantics can be assigned to them by means of logics and ontologies.

QSR has been successfully applied to many areas such as robotics (Falomir et al., 2013b; Wolter et al., 2011), computer vision (Falomir et al., 2011; Cohn et al., 2006), ambient intelligence (Bhatt and Dylla, 2009; Falomir and Olteţeanu, 2015), shape recognition (Falomir et al., 2013a), architecture and design (Richter et al., 2010; Bhatt and Freksa, 2015), etc. Specifically GIS has been the field in which most QSR models – for example RCC-8 (Randell et al., 1992), 9-Intersection model (9IM)(Egenhofer, 1995) – have found a direct application when investigating: topological changes in space (Egenhofer and Al-Taha, 1992), and in sensor networks (Jiang and Worboys, 2008), topological relations between multi-holed regions (Vasardani and Egenhofer, 2009), the extraction of qualitative spatial relations between recognized places from natural language place descriptions (Khan et al., 2013; Vasardani et al., 2013), the generation of narratives to explain spatio-temporal dynamics (Bhatt and Wallgrün, 2014), spatial query solving and retrieval (Fogliaroni, 2013; Al-Salman, 2014), the alignment of sketch and metric maps (Schwering et al., 2014), etc.

In this paper, qualitative topological relations between pairs of entities are investigated to understand the qualitative characteristics of target features. Based on the first law of geography (Tobler, 1970): “Everything is related to everything else, but near things are

more related than distant things", we need to find the frequent relations between entities that uniquely distinguish each class. For example, a *park* typically contains playgrounds, sport centers, pathways, etc., whereas a *meadow* contains less infrastructure; also a *park* is probably located within or near urban areas, whereas *meadow* is typically located near farms and rural areas, etc. We apply the 9IM (Egenhofer, 1995) to investigate qualitative topological relations between pairs of entities. As shown in Figure 5.4, 9IM describes topological relations between pairs of entities as: disjoint, meet, overlap, covers, covered by, contains, inside, and equal. Basically, geographic features are represented by means of point, line, and polygon data elements. In this work, target classes are usually represented by polygons. Thus, we consider all mutual topological relations between polygon and other data elements; *polygon-point, polygon-line and polygon-polygon.*

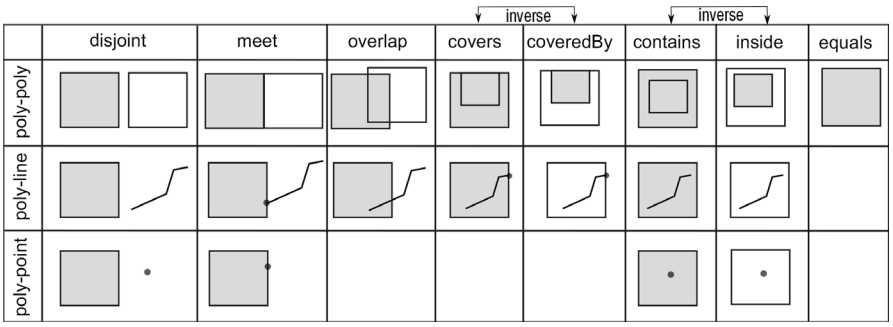


FIGURE 5.4: The 8 topological relations in the 9-Intersection Model by Egenhofer (1995).

At Figure 5.4, let us assume that the gray entities represent the target entities, then the relevant relations to consider are: disjoint, meet, overlap, contains, and covers. Regarding the disjoint relation we analyze entities within a distance from 5 to 10 meters from target entities. Particularly, the disjoint relation gives insight about the external geographic context, while the others represent the relations resulting from the intersections of the interiors and boundaries of entities. Note that inside, covers, and equals relations are not considered because: (a) inside and covers are inverse relations of contains and covered by, respectively; and (b) the equal relation rarely occurs and does not add useful information for this analysis.

5.5 The Proposed Rule-Guided Classification Approach

The proposed approach is aimed to improve the quality of data classification in VGI by guiding the participants during the classification process. Through this guidance we aim to obtain data of homogeneous and appropriate classification. Figure 5.5 illustrates the proposed approach, which consists of two phases: *Learning phase* (Section 5.5.1) and *Guiding phase* (Section 5.5.2).

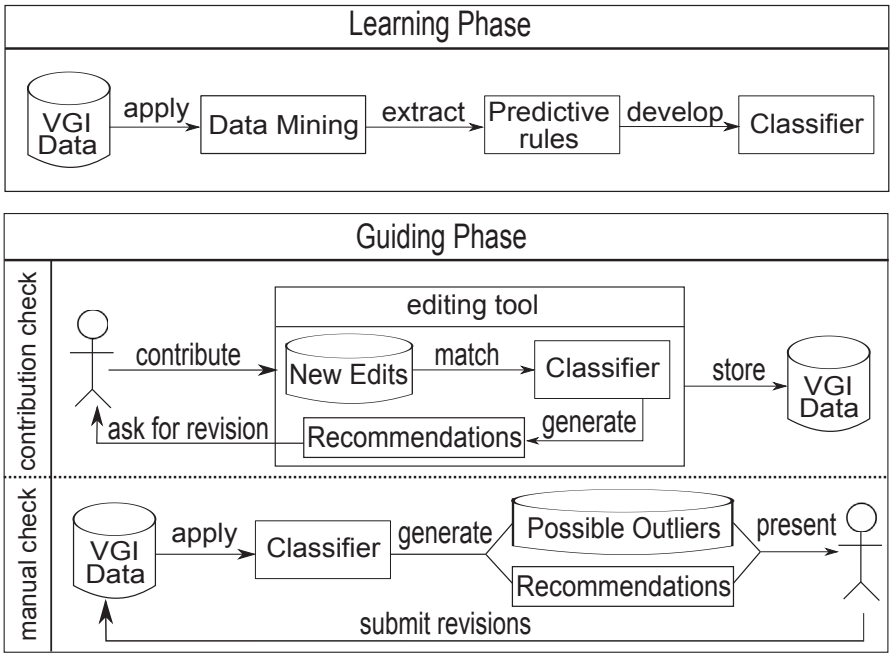


FIGURE 5.5: Rule-guided classification approach.

5.5.1 Learning phase

The objective of the *Learning phase* is mining the VGI data source to extract a set of individual characteristics of specific geographic features. These characteristics are extracted by analyzing the qualitative topological relations of the target features. The characteristics are formulated into predictive rules with the format:

$$head \leftarrow body \tag{5.1}$$

where the body describes the qualitative topological characteristics of a geographic feature and the head points to the recommended classification. The combination of rules will be able to describe a specific feature. Afterwards, the extracted rules are organized into a rule-based classifier, which consequently would be able to recommend the most appropriate classification for a given set of characteristics of a specific geographic feature.

During the learning process, we excursively investigate the qualitative topological characteristics of features to understand the geographic context of target features. We take into account the locality principle: we assume that at country level a certain geographic feature should have the same characteristics. For example, learning the characteristics of *forest* features in China and applying the developed classifier in Germany may not make sense. Thus, we maintain the locality principle in the Learning and Guiding phases, as well.

5.5.1.1 Data mining process

This process aims to find frequent patterns (in this case topological relations) involved between target classes and other geographic features; e.g., *park* contains playground, sport center, *garden* meet residential houses, fences, etc. According to the OSM tagging method, we consider each combination of *key* = *value* as a new feature type. For example, *leisure* = *playground* and *leisure* = *sport* are two different geographic features. We encode them as `leisure_playground` and `leisure_sport` respectively and relate each new feature with a unique identifier (ID) in an indexed file. The analysis includes the common map features that are suggested by the OSM project on its Wiki page[5]. Due to the free contribution mechanism of the OSM project, the analysis results in more than 1,000 unique features, after filtering. The mining process works to extract atomic rules in form of rule (5.1), which is translated into:

$$Class(X, C) \leftarrow R(X, F) \tag{5.2}$$

where X represents a target entity, C is the predicted class and $C \in \{park,\ meadow,\ ...\}$, R is one of the topological relations where $R \in \{\mathsf{contains}, \mathsf{meet}, ...\}$ and F represents the set of frequent features that is mostly involved in a relation R with entities of class C.

To extract such rules, we apply the Apriori algorithm (Agrawal, Srikant, et al., 1994) which is one of the most common data mining algorithms initially developed to extract frequent item sets and to learn association rules from a transactional database (Witten and Frank, 2005). In this work, we particularly use a class association rule mining task, when rules have a predefined class (e.g., *park*) as their consequences (left side at rules (5.1) and (5.2)). Extracting interesting rules among a large number of possibilities requires setting up some constraint thresholds: support (*supp*) and confidence (*conf*) are two commonly used constraint thresholds for extracting and evaluating interesting rules, as follows:

- **Support:** used to filter interesting patterns. It is defined as the percentage of entities that hold the body description. For example, *supp* (contains(X, [1, 15])) = 20%, means 20% of the entire entities contain playground (where 1=`leisure_playground`) and footways (where 15=`highway_footway`) features.
- **Confidence:** used to evaluate extracted rules. It is equal to the percentage of entities that hold the body description and consequently the head. e.g., *conf* ($Class(X,\ park)$ ← contains(X, [1, 15])) = 80%, implies 80% of the entities hold the rule: body associated with class *park*.

[5] http://wiki.openstreetmap.org/wiki/Map_Features

5.5.1.2 Classifier development

The main idea of association rule mining has been adapted to solve other problems, such as classification, resulting in *Associative Classification* (AC) mining field; It is one branch of data mining that combines two mining fields, associating rule mining and classification, to build a classifier based on a set of predictive association rule (Thabtah, 2007). Generally, developing such a classifier based on a set of predictive rules consists of 4 steps:

Step 1: Find all interesting class association rules from a data set, using the *supp* threshold;

Step 2: Filter the extracted rules into a set of predictive association rules, based on the *conf* threshold;

Step 3: Encode the rules into a rule-based classifier; then

Step 4: Evaluate the classifier performance on a test data set.

In geographic contexts, anything could be possible. For example, a building may be located in a desert, a highway might cross a residential area or a public park, etc. Moreover, in VGI projects there exist unlimited unique features (see Section 5.5.1.1). Thus, we set a *support* threshold to 1% and we consider as frequent those patterns that occur with a frequency higher than 1%. This threshold is used due to: (i) the lower frequencies are considered as rare patterns and might have no significant influence on the classification; and (ii) from a rational perspective, considering these rare patterns might lead to biased classification. During this learning process, we are mining to extract atomic rules per topological relation per class. The extracted rules represent the output of Step 1.

However, due to the uncertainty of spatial data, the extracted rules themselves represent a challenge at Steps 2 and 3. The aim at Step 2 is to filter and organize the extracted rules into a set of predictive association rules for developing the classifier in Step 3. Hence, the difficulties come from the following points: (a) Step 1 results in rules of identical bodies associated with different heads (classes); (b) during Step 2, the higher the *confidence* threshold for filtering interesting rules, the higher the possibility to dismiss useful information; (c) due to overlapping classes (see Section 5.3.2), an entity could plausibly belong to more than one class; and (d) due to geographic context, an entity could match with several atomic rules associated with different heads (classes). In summary: How should we classify? By the majority of rules or by rules of higher *confidence*? In this paper, we considered the most *appropriate classification* of a given entity to be one that best reflects its qualitative characteristics.

5.5.2 Guiding phase

During the *Guiding phase*, the aim is to enhance the classification quality of VGI by applying the developed classifier. This approach presents two guiding scenarios for applying the classifier:

First, contribution checking scenario, when the classifier is implemented in an editing tool. At contribution time, the tool informs the participant about a potential classification problem, based on the classifier. The editor provides the participant with recommendations. Afterwards, the participant considers the guidance provided and responds with correction (if required).

Second, manual checking scenario, when the classifier is applied directly on an existing data set. The classifier points out entities with problematic classification, which do not match any provided recommendation. The classifier generates the problematic entities combined with the generated recommendations. Afterwards, both are presented for assessment and correction (if required).

In both scenarios, we do not restrict participants to the given recommendations. However, we provide them with flexible guidance, which probably might lead to indirect data enrichment, for example, when participants add more information to satisfy recommendations (if they find additional appropriate classifications).

5.6 Experimentation and Results

To evaluate the presented approach, we performed an empirical study. The study checked the ability of the developed classifier to distinguish between similar classes. We used the OSM data set of Germany. The reasons for this choice were: (i) in Germany there are active mapper communities, particularly in urban areas; (ii) no authoritative bulks of data are imported; so the data set still reflects the voluntary nature; and (iii) several studies concluded that the quality of OSM data in Germany is higher than that of other places (Zielstra and Zipf, 2010; Ludwig et al., 2011; Neis et al., 2013).

In our previous study Ali et al. (2015), we utilized the German data set dated December 2013, while in the present study, we use an updated version of May 2015. Following the methodology described, we extracted all entities that are represented by polygons and are classified as *forest, garden, grass, meadow, park*, or *wood*. The entities are extracted from the ten most densly populated cities in Germany[6] to ensure active mapper communities and acceptable quality levels. These cities are: Berlin, Bremen, Cologne, Dortmund,

[6] http://en.wikipedia.org/wiki/List_of_cities_in_Germany_by_population

Düsseldorf, Essen, Frankfurt, Hamburg, Munich, and Stuttgart. This data set consists of 23,567 entities as follows: 3,590 *forest*, 3,025 *garden*, 7,188 *grass*, 4,038 *meadow*, 4,298 *park*, and 1,428 *wood* entities. We processed each entity individually by analyzing the topological relations between pairs of entities within its geographic context. Each entity is described by a set of topological relations with other surrounding features and is assigned to a specific class.

This section provides a detailed description of the learning process (Section 5.6.1), the classification hypotheses (Section 5.6.2), and the classification process (Section 5.6.3) followed in the experimentation. Then, the results obtained are explained (Section 5.6.4) and their validation is described (Section 5.6.5).

5.6.1 Learning process

In the learning process, we applied the Apriori algorithm (Agrawal, Srikant, et al., 1994) to investigate the frequent topological relations that describe each class. During the topological analysis, we adapted our previous study in Ali et al. (2015) to handle the imprecise editing in VGI. We considered entities in a distance up to 1 meter from the boundary of a target entity as they are on the boundary, and hence, they fulfill the topological meet relation. Moreover, we checked the disjoint relation within distances of 5 and 10 meters to have insights into various geographic scopes. We used a *support* threshold of 1% to find the interesting patterns. Each topological relation is processed individually with a given class to generate a set of predictive qualitative rules of that class. The rules represent the output of Step 1 (see Section 5.5.1.2). Table 5.1 shows a snapshot of the extracted rules.

Rule	*supp.*	*conf.*
$Class(X, grass) \leftarrow \mathsf{disjoint}_{10m}(X, [13, 40, 45, 57])$	1%	99%
$Class(X, park) \leftarrow \mathsf{contains}(X, [1, 15, 22, 27, 36])$	1%	98%
$Class(X, garden) \leftarrow \mathsf{meet}(X, [27, 42, 78, 235])$	2%	98%
$Class(X, park) \leftarrow \mathsf{contains}(X, [1])$	23%	88%
$Class(X, garden) \leftarrow \mathsf{overlap}(X, [43])$	1%	78%
$Class(X, park) \leftarrow \mathsf{overlap}(X, [15])$	57%	52%
$Class(X, grass) \leftarrow \mathsf{contains}(X, [nothing])$	77%	34%
$Class(X, meadow) \leftarrow \mathsf{contains}(X, [nothing])$	75%	20%
$Class(X, park) \leftarrow \mathsf{contains}(X, [nothing])$	37%	9%
$Class(X, forest) \leftarrow \mathsf{contains}(X, [nothing])$	78%	7%

where
1=`leisure_playground`
6=`highway_residential`
13=`route_bus`
15=`highway_footway`
21=`sport_soccer`
22=`leisure_pitch`
27=`building`
36=`highway_steps`
40=`route_road`
42=`highway_service`
43=`building_residential`
45=`highway_cycleway`
57=`landuse_grass`
78=`barrier_fence`
89=`nature_water`
181=`highway_track`
235=`leisure_garden`

TABLE 5.1: Samples of the extracted qualitative descriptive rules.

In Table 5.1, the 1^{st} rule describes the case when a *grass* entity is located beside public roads and cycle ways for decoration purposes; the 2^{nd} rule points to the probable enclosure of *park* entities to leisure facilities and footways; the 3^{rd} and 5^{th} rules sketch the scene of (residential) *garden* entities, when they are located adjacent to houses, fences, service ways, and other *garden* entities and overlapping with (residential) houses; the 4^{th} rule, emphasizes the absolute relation between the playground facilities and *park* entities; and the 6^{th} rule partially identifies the logical connection between the interior and exterior of *park* entities, by means of a footway.

From another specific view, an example of duplicated rules are shown in the last four rules of $\mathsf{contains}(X, [nothing])$, while the various values of *confidence* threshold raise a conceptual classification issue: when an entity contains *nothing* it is more likely classified as *grass* or *meadow* than as *park* or *forest*.

As indicated in Table 5.2, we extracted 4,425 rules describing the classes as follows: 1,246 describe *forest*, 216 describe *garden*, 659 describe *grass*, 441 describe *meadow*, 1,468 describe *park*, and 395 describe *wood*. The rules have a wide range of *confidence* threshold: 1,235 out of 4,425 rules have a *confidence* $\geq 50\%$, while the others have a descending *confidence* to less than 1%. Otherwise, the constraint thresholds and the rules are distributed differently among classes and topological relations as well.

	forest	*garden*	*grass*	*meadow*	*park*	*wood*	**Total**
contains	45	8	7	13	468	17	558
coveredBy	9	8	16	12	9	13	67
$\mathsf{disjoint}_{5m}$	161	28	100	64	115	50	518
$\mathsf{disjoint}_{10m}$	679	106	470	241	618	180	2,294
meet	130	55	36	84	116	54	475
overlap	222	11	30	27	142	81	513
Total	1,246	216	659	441	1,468	395	**4,425**

TABLE 5.2: The distribution of rules per classes per relations.

5.6.2 Classification hypotheses

As shown and mentioned previously, the rules resulting from the learning process represent a challenge for developing the classifier. The classification of entities based on a single topological relation or a rule of the highest *conf.* may be biased. For example, classification of *park* entities depending on the contain relation, or classification of forest entities based on the meet relation might lead to biased classification. However, other significant conceptual rules like $\mathsf{contains}(X, [nothing])$ can be exploited as a filter for the

classification, at least to reduce the number of plausible alternatives. To overcome these challenges, we tested the following hypotheses:

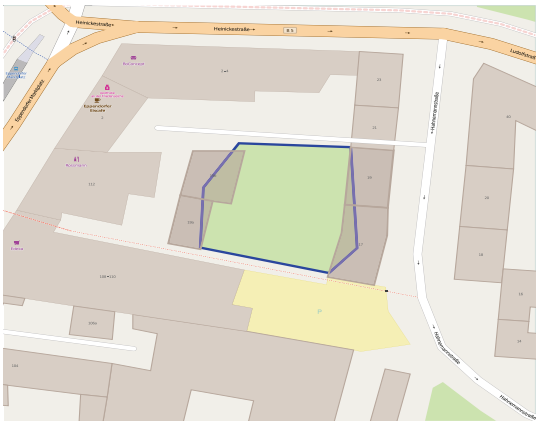
- **Pruning**:
  1. filtering based on the *confidence* threshold: the classification is done once by considering the entire set of extracted rules and once by exploiting the rules with $conf \geq 50\%$.
  2. $\mathsf{disjoint}_{5m} \subset \mathsf{disjoint}_{10m}$, then we check applying both relations together or applying each relation individually.
- **Ranking 1^{st} and 2^{nd} recommendations**: During the classification process, we consider the 1^{st} and 2^{nd} recommended classes given by the predictive rules; when each entity is matched against the entire developed rules, the maximum *conf* per class determines 1^{st} and 2^{nd} recommended classes.
- **Classification assumptions**: Due to an unbalanced distribution of rules, we consider only rules with maximum *conf* per class to define 1^{st} and 2^{nd} potential classes.

Based on the procedure of data selection, we assumed that a large fraction of the data set has an acceptable classification quality. Therefore, we depend on the classification accuracy as a measure to evaluate the proposed hypotheses. Here, the classification accuracy implies that the percentage of corrected classified entities that have been assigned a class label match with one of the 1^{st} or 2^{nd} recommended classes.

5.6.3 Classification process

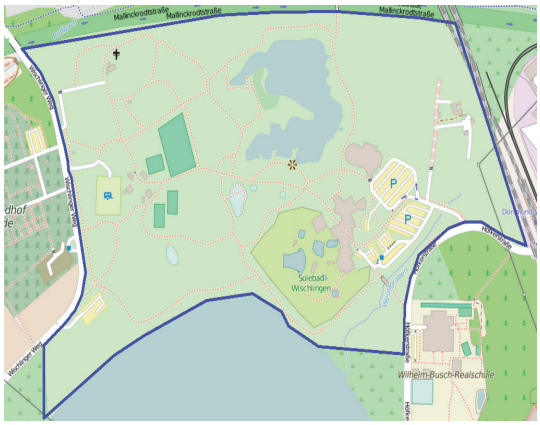
During this process, each entity is classified with respect to the matched qualitative rules. For example, the given entities in Figure 5.6a and 5.6b illustrate the classification process; they show entities and their corresponding samples of the matched rules. In Figure 5.6a, the entity matches 136 rules: 25 *park*, 25 *forest*, 24 *grass*, 21 *wood*, 22 *meadow*, 19 *garden*, while the entity in Figure 5.6b matches 401 rules: 232 *park*, 132 *forest*, 25 *grass*, 8 *meadow*, 2 *wood*, and 2 *garden*.

Following the illustrations and the information in the previous Figures, the entity in Figure 5.6a can be described as: overlaps residential buildings, meet buildings, contains *nothing*, and disjoint within 10 m to a service/foot roads (highway). While, the other entity in Figure 5.6b can be described as: contains a nature water body/playground/sport center/footways, meet residential/footway/services roads (highway), and overlap other forest areas.



Rule	*conf.*
$Class(X, garden) \leftarrow \mathsf{overlap}(X, [43])$	78%
$Class(X, garden) \leftarrow \mathsf{meet}(X, [27])$	39%
$Class(X, grass) \leftarrow \mathsf{disjoint}_{10m}(X, [15, 42])$	39%
$Class(X, grass) \leftarrow \mathsf{contains}(X, [nothing])$	34%
$Class(X, park) \leftarrow \mathsf{disjoint}_{10m}(X, [15, 42])$	24%

(A) entity with osm_id = 82449147 and sample of the matched rules corresponding to the entity



Rule	*conf.*
$Class(X, park) \leftarrow \mathsf{contains}(X, [1, 15, 27, 89])$	94%
$Class(X, park) \leftarrow \mathsf{contains}(X, [1, 15, 21, 22])$	83%
$Class(X, park) \leftarrow \mathsf{meet}(X, [6, 15])$	70%
$Class(X, park) \leftarrow \mathsf{meet}(X, [6, 15, 42])$	55%
$Class(X, forest) \leftarrow \mathsf{overlap}(X, [42, 181])$	43%

(B) entity with osm_id = 25422214 and sample of the matched rules corresponding to the entity

FIGURE 5.6: Examples of the classification process; see Table 5.1 for the indications of the ID's.

Considering exclusively the highest *confidence* threshold, the entities in Figure 5.6a and 5.6b have recommended class labels as *garden* and *park*, respectively. However, to consider 1^{st} and 2^{nd} recommendations, we looked into the maximum *confidence* per class. The entity in Figure 5.6a is classified as *garden* (1^{st} recommendation) and as *grass* (2^{nd} recommendation), while the other entity in Figure 5.6b is classified as *park* (1^{st} recommendation) and as *forest* (2^{nd} recommendation).

5.6.4 Results

We used the classification accuracy measure to judge on the proposed hypotheses in Section 5.6.2. In this work, the *classification accuracy* implies the compatibility between our recommendations and the presented classification on the OSM data.

First, in the case of filtering based on the $conf \geq 50\%$, the classifier provided poor performance since 25% of the entities did not match any rule and the classification accuracy was 55%. The reason for that is that, although the approach extracted several meaningful qualitative rules, which are identical to the textual recommendations given on OSM project to some extent, the filter led to missing valuable information embedded in rules with low *confidence* threshold. For example, we extracted the following rule: $Class(X, park) \leftarrow \mathsf{meet}(X, [highway\_footway])$, with a *confidence* threshold of 38%;

The rule is identically defined in OSM Wiki recommendations[7], in the description of how to map a *park* feature. Second, when comparing the $\mathsf{disjoint}_{5m}$ and $\mathsf{disjoint}_{10m}$ relations, the highest performance is obtained when the disjoint relation within 10 metres ($\mathsf{disjoint}_{10m}$) is applied, producing 72.5% classification accuracy. Thus, in further analysis, we considered only the rules of $\mathsf{disjoint}_{10m}$ relation and avoid the 518 rules generated from the $\mathsf{disjoint}_{5m}$ relation.

The classification accuracy per class is shown in Figure 5.7. According to this Figure, *grass*, *garden*, and *forest* have higher classification accuracies, 92%, 84%, and 70% respectively, while *meadow* and *park* have moderated accuracies of 62%. However, the *wood* class has a noticeable lower classification accuracy of 16%.

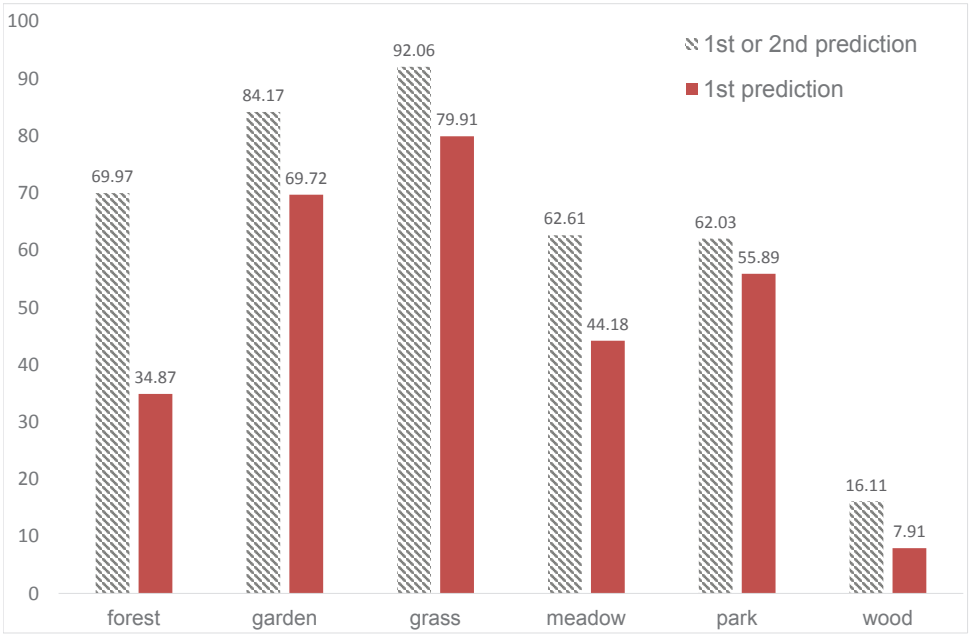


FIGURE 5.7: Classification accuracies per class

We summarize our explanation of these results as follows:

1. The higher classification accuracies obtained for the classes *grass*, *garden*, and *forest* are due to: the *grass* class, is a general class that could describe the land cover of all of these entities, while the characteristics of the second and third classes are well recognizable and they might be well mapped particularly in urban areas (cities); The *garden* entities which are contributed within city boundaries are mostly (residential) *garden* that have unique characteristics; and the *forest* entities can be recognized by heavy coverage of woody trees.

2. The moderate classification accuracy obtained for the *meadow* class can be due to the limited occurrence of *meadow* entities within city boundaries. Moreover, the

[7] http://wiki.openstreetmap.org/wiki/Tag:leisure%3Dpark

concept of *meadow* might not be identically received by participants. For example, for some participants, *meadow* is an open grass area, that is artificially created, and mostly contains multiple wildlife, while for others, it is a place where hay and pasture are growing and used for the purpose of feeding animals.

3. The moderate classification accuracy obtained for the *park* class is due to: 37% of entities contain no infrastructure (see Table 5.1, the 9^{th} rule *support*), and hence, they might represent either a problematic conceptual classification or incomplete mapped entities.

4. The lower classification accuracy obtained for the *wood* class is due to the limited number of entities in the training data set and the multiple classification recommendations that are presented at the OSM Wiki[8] for *forest* and *wood* classes.

In addition, it is important to note that we are dealing with a VGI data source and therefore, some features might be better mapped than others. It could also happen that the training data set contains incorrectly classified entities. However, we assumed the correctness of a large fraction of data.

5.6.5 Validation

Due to the unavailability of ground-truth data for the selected features, we adopted the following ways to validate the findings.

FIGURE 5.8: Appropriately classified entities as *forest* matched the recommendations: 1^{st} *forest*, 2^{nd} *meadow*.

FIGURE 5.9: Inappropriately classified entities as *park*, while the recommendations are: 1^{st} *garden*, 2^{nd} *grass*.

First, the results were visually examined to check the plausibility of the proposed recommendations. Figure 5.8 and 5.9 illustrate examples of detected appropriately and

[8] http://wiki.openstreetmap.org/wiki/Tag:natural%3Dwood

inappropriately classified entities, with respect to the generated recommendations. In Figure 5.8, the entities are appropriately classified as *forest*, matching the recommendations: *forest* (1^{st}) and *meadow* (2^{nd}). The entities are located within a meadow area and near highways and farmland areas (i.e., non-urban area). The entities might look sparse and smaller in size than the common *forest* entities. However, in the OSM project, there are no restrictions on specific definitions of the features, but community agreements control the data classification. In Figure 5.9, the entities are misclassified as *park* and the recommendations provided are *garden* (1^{st}) and *grass* (2^{nd}). The reasons behind these recommendations are that the entities contain nothing and are located adjacent to houses. The findings indicate that applying the proposed classifier and following the given recommendations will potentially enhance the classification quality.

Second, we exploited intrinsic properties, like number of tags, version, mapper's reputation, etc., to extract a data set for the validation process. For example, we extract all entities that are named *park* and are tagged *leisure=park* as a validation data set. The extraction done from the entire data set of Germany resulted in 1,856 *park* entities. We applied the developed classifier on the extracted entities. The results show that 90% of the entities are correctly classified by the 1^{st} recommendation, while 98.5% of the entities are correctly classified by the 1^{st} or 2^{nd} recommendation. The validation reflects the classifier efficiency in distinguishing a specific class based on learning its intrinsic and extrinsic characteristics. Hence, applying the classifier on the entire set of *park* entities of Germany would point out the inappropriately classified entities. The problematic classification might be relevant to incomplete mapping of an area or to an incorrect mapping attitude of participants; the classification could be improved by applying the proposed classifier.

Third, we compare the results of our previous study in Ali et al. (2015) with the current study. Both studies utilized data sets of the same features from the same location, but with different dates: December 2013 and May 2015, respectively. Due to the dynamic nature of spatial data generally and VGI data particularly, we detected that 6% of the entities of 2013 have been deleted, whereas, a larger fraction of 94% remain in the data set of 2015. Among the remaining fraction, 96% of the entities are still in the target classes, while only 4% of the entities are updated to other related classes like *scrubs*, *recreation*, and *construction*. In some sense, the remaining of a large fraction of data for 18 months may indicate the conceptual quality of these entities. During our analysis of the remaining fraction, we found that a promising percentage of 8% of entities have been updated according to our recommendations, without our interference or intention of guidance. The findings encourage us to implement a crowdsourcing revision scenario to check the classification of these features.

5.7 Discussion

For VGI to become an everyday common practice for human beings across all sectors, an *intuitive* geographic information capture, exchange, and reasoning is needed (Nittel et al., 2015). More natural interfaces between people and their smartphones, cars, etc., which enable for example, dialogue communication, would increase the quality of VGI. Therefore, the resulting improvement of VGI may help in disambiguating vague place descriptions (Jones et al., 2008).

Research towards *spatial cognition engineering* (Richter et al., 2015) is carried out for developing more cognitive interfaces/systems so that devices adapt to users, instead of forcing users to accommodate to devices. Regular users are not system designers and are often not experts in the field. Spatial intelligence is needed for normal humans and more specifically when they arrive at new places. GIS and VGI can help humans to improve this spatial intelligence. Spatial cognition studies can also help to improve GIS and the way VGI is captured and presented to users towards augmenting their spatial intelligence for example in wayfinding and decision taking.

Developing *intuitive data capture interfaces*, is one solution, among others, to enhance the resulting data. Particularly, the data quality can be improved, when the interfaces support the ability to interpret visual information more easily (in a cognitively adequate manner), and provide volunteered participants with better feedback. In the VGI context, participants usually use free and flexible mapping tools. A guiding mechanism may motivate them without affecting their flexibility to choose and decide about the entities.

Furthermore, in the *next-generation of GIS* systems, the automated extraction of high-level entities (i.e. objects, properties, processes, etc.) from remote sensing images is envisioned, as the advances in Geographic Object-Based Image Analysis (GEOBIA) highlight (Arvor et al., 2013; Blaschke et al., 2014). Therefore, high-level image descriptions would lead towards more intuitive GIS user interfaces, which will enable higher precision and higher quality of data.For example, if the houses and the grass from the remote sensing image shown in Figure 5.10 were identified automatically, then the selected entity to be classified would follow the qualitative descriptions of *within a residential area* and *grass-related class*, and consequently, that would reduce the classification space and generate appropriate recommendations. Conversely, VGI observations contributed by local volunteers may help GEOBIA systems to improve their remote sensing image classifiers, for example in situations where coarse resolution cells (e.g., 1 km^2) might not differentiate open patches, paths, and roads inside a forested area. Finally, when volunteered participants are guided by intuitive interfaces and quality assurance mechanisms are developed

to ensure the data quality of VGI, we envision a very fruitful collaboration of the GEOBIA + VGI research fields. In particular, the integration of experts' perception (who observe the geographic features from remote sensing images captured by satellites) with local participants' observations (who contribute geographic information based on their local knowledge about places where they live) may produce richer data sources with a higher data quality.

The *advantages* of the proposed approach are; first, it is grounded on investigation of the topological relations, hence, it could be applied on different types of geographic features (e.g., water body features); and second, with our assumption of "identical entities should be classified similarly within the same Country"; the approach could be used to enhance the data classification in non-urban areas of a Country that has rich data sets in urban areas.

Otherwise, the proposed approaches has some *limitations*, since developing the classifier requires large amounts of data with a certain level of quality. Regarding the availability of large amounts of data, it is related to data mining. As most algorithms of data mining act effectively with large training data sets. Moreover, we assumed locality of data classification. For example, learning the rules form data of Germany and applying them on data of India might be inappropriate. Thus, in case of unavailability of data, applying the extracted information at completely different geographic boundaries (i.e., different cultures), might result in inappropriate classification. Regarding the quality of the utilized data, we assumed that the OSM data in urban areas are ensured by crowdsourcing and social approaches; In particular, there exist active mapping communities in Germany. However, applying the approach on data of another location requires careful investigation of the utilized data quality.

5.8 Conclusions and Future Work

With the increasing development of VGI data sources, the demand for high data quality rises with high priority. Nowadays, VGI is a data source that supports diversity of applications and services in GIScience research. In the present paper, we are mainly concerned with VGI resulting from collaborative mapping, where public participants work together to map geographic features. The uncertainty of spatial data, human-centred classification, and flexible contribution mechanisms result in data of problematic classification. However, consistent and precise feature classification is required towards effective utilization of the resulting data.

From a professional perspective, spatial data classification is carried out based on field observations and physical measurements, with respect to a pre-defined classification model.

In contrast, in VGI projects, public participants are eager to record their individual observations classifying the data based on their local knowledge and personal cognitions. This gap stimulates the idea of guided classification in VGI projects. The proposed guidance aims to drive appropriate data classification by imitating professional data classification methods.

In this paper, we proposed a rule-based guided classification approach for VGI projects. The approach exploits QSR as well as VGI to learn qualitative characteristics of specific geographic features. We addressed the classification of grass-related features using OSM data of Germany. We developed a classifier able to distinguish between *forest, garden, grass, park, meadow*, and *wood* entities. We applied data mining functions to extract qualitative rules describing the target features. Afterwards, we encoded these rules into a classifier, which was able to successfully distinguish between the entities.

The findings reveal the feasibility of the proposed approach. The developed classifier was able to detect and provide appropriate recommendations for problematically classified entities. According to the extracted entities, the classifier showed that 72.5% of the entities have an appropriate classification. The results pointed to problematic classification of 8%, 16%, and 30% of *grass, garden*, and *forest* entities, respectively, and 34% of *park* and *meadow* entities. The findings indicated the noticeable problematic classification of *wood* entities. Three methodologies were adopted to validate the findings: i) checking a sample of entities visually; ii) using a test data set; and iii) analyzing the temporal classification evolving of entities. These validations emphasized the promising results of the proposed approach.

In the discussion, we argued the role of intuitive interfaces to enhance the data quality of VGI. Tackling the classification of the target features, we started to implement a web application that will present our recommendations to crowdsourcing revisions. This application is called `Grass&Green` and it implements the manual checking scenario proposed in this paper under the concept of crowdsourcing revision. The application has the following objectives:

1. presenting our generated recommendations to the community;
2. checking the validity of the proposed approach;
3. measuring the participants' satisfaction towards the guiding approach; and
4. improving the data classification of these features.

Figure 5.10 shows the interface of the `Grass&Green` application. On the right hand, we intend to present the entities associated with their qualitative description above a

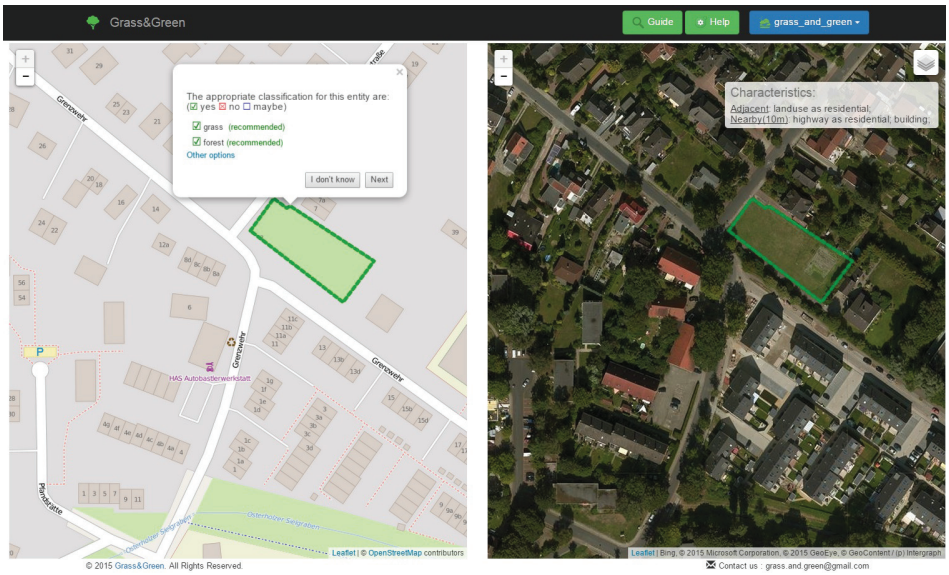


FIGURE 5.10: The interface of the `Grass&Green` web application.

satellite image. On the left hand, the generated recommendations are provided with a flexibility to accept/reject the recommendation and make further updates to express the appropriate classification. Moreover, textual and visual descriptions of the features will be provided for participants in a dedicated menu (Guide). An adapting mechanism to attract the crowds to participate in this application is still under study. Other formats of crowdsourcing like social media, and discussion blogs will be exploited to announce this application.

We discussed the potential integration of enhanced VGI and GEOBIA towards producing more rich and precise geographic data sets. Furthermore, additional investigations are required to evaluate the extracted rules. In future work, we intend to implement the *Guiding* phase and measure the classification improvements based on the provided recommendations. We plan to study the OSM ontology, e.g., OSMonto (Codescu et al., 2011), to determine whether the semantic distance between the ontological concepts could solve the ambiguity between similar classes.

Acknowledgment

This work is partially funded by the German Academic Exchange Service (DAAD), the Bremen Spatial Cognition Center (BSCC), the European Marie Curie project COGNITIVE-AMI and the University of Bremen (project *Cognitive Qualitative Descriptions and Applications* – CogQDA).

Bibliography

Agrawal, R., R. Srikant, et al. (1994). "Fast algorithms for mining association rules". In: *Proc. 20th int. conf. very large data bases, VLDB*. Vol. 1215, pp. 487–499.

Al-Salman, R. (2014). "Qualitative Spatial Query Processing: Towards Cognitive Geographic Information Systems". Supervised by Prof. Christian Freksa (University of Bremen) and Prof. Christian Jensen (Aalborg University). PhD thesis. University of Bremen.

Ali, A. L. and F. Schmid (2014). "Data quality assurance for Volunteered Geographic Information". In: *Geographic Information Science: 8th International Conference, GIScience 2014, Vienna, Austria, September 24-26, 2014. Proceedings*. Ed. by M. Duckham, E. Pebesma, K. Stewart, and A. U. Frank. Cham: Springer International Publishing, pp. 126–141. ISBN: 978-3-319-11593-1.

Ali, A. L., F. Schmid, R. Al-Salman, and T. Kauppinen (2014). "Ambiguity and plausibility: managing classification quality in Volunteered Geographic Information". In: *Proc. of the 22nd ACM SIGSPATIAL International Conf. on Advances in Geographic Information Systems*. Ed. by Y. Huang, M. Schneider, M. Gertz, J. Krumm, and J. Sankaranarayanan. New York, NY, USA: ACM, pp. 143–152. ISBN: 978-1-4503-3131-9.

Ali, A. L., F. Schmid, Z. Falomir, and C. Freksa (2015). "Towards Rule-Guided Classification for Volunteered Geographic Information". In: *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* II-3/W5, pp. 211–217.

Arsanjani, J. J., P. Mooney, A. Zipf, and A. Schauss (2015). "Quality Assessment of the Contributed Land Use Information from OpenStreetMap Versus Authoritative Datasets". In: *OpenStreetMap in GIScience: experiences, research, and applications*. Ed. by J. J. Arsanjani, A. Zipf, P. Mooney, and M. Helbich. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 37–58.

Arvor, D., L. Durieux, S. Andres, and M.-A. Laporte (2013). "Advances in Geographic Object-Based Image Analysis with ontologies: A review of main contributions and limitations from a remote sensing perspective". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 82, pp. 125 –137. ISSN: 0924-2716.

Ballatore, A. and A. Zipf (2015). "A Conceptual Quality Framework for Volunteered Geographic Information". In: *Spatial Information Theory: 12th International Conference, COSIT 2015, Santa Fe, NM, USA, October 12-16, 2015, Proceedings*. Ed. by S. Fabrikant, M. Raubal, C. Bertolotto M.and Davies, S. Freundschuh, and S. Bell. Santa Fe, NM, USA, pp. 89–107.

Barron, C., P. Neis, and A. Zipf (2014). "A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis". In: *Transactions in GIS* 18.6, pp. 877–895.

Bhatt, M. and F. Dylla (2009). "A Qualitative Model of Dynamic Scene Analysis and Interpretation in Ambient Intelligence Systems". In: *I. J. Robotics and Automation* 24.3.

Bhatt, M. and C. Freksa (2015). "Spatial Computing for Design—an Artificial Intelligence Perspective". In: *Studying Visual and Spatial Reasoning for Design Creativity*. Ed. by J. S. Gero. Springer, pp. 109–127. ISBN: 978-94-017-9296-7.

Bhatt, M. and J. O. Wallgrün (2014). "Geospatial Narratives and Their Spatio-Temporal Dynamics: Commonsense Reasoning for High-Level Analyses in Geographic Information Systems". In: *ISPRS International Journal of Geo-Information* 3.1, p. 166. ISSN: 2220-9964.

Bishr, M. and W. Kuhn (2007). "Geospatial Information Bottom-Up: A Matter of Trust and Semantics". In: *The European Information Society: Leading the Way with Geo-information*. Ed. by S. I. Fabrikant and M. Wachowicz. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 365–387. ISBN: 978-3-540-72385-1.

Blaschke, T., G. J. Hay, M. Kelly, S. Lang, P. Hofmann, E. Addink, R. Q. Feitosa, F. van der Meer, H. van der Werff, F. van Coillie, and D. Tiede (2014). "Geographic Object-Based Image Analysis – Towards a new paradigm". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 87, pp. 180 –191. ISSN: 0924-2716.

Codescu, M., G. Horsinka, O. Kutz, T. Mossakowski, and R. Rau (2011). "OSMonto – An Ontology of OpenStreetMap Tags". In: *Proceedings of the SOTM-EU 2011 : 1st State of the Map - Europe Conference*. Ed. by M. Schmidt and G. Gartner, pp. 55 –65.

Cohn, A., D. Hogg, B. Bennett, V. Devin, A. Galata, D. Magee, C. Needham, and P. Santos (2006). "Cognitive Vision: Integrating Symbolic Qualitative Representations with Computer Vision". In: *Cognitive Vision Systems: sampling the spectrum of approaches*. Ed. by H. I. Christensen and H.-H. Nagel. Vol. 3948. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 221–246. ISBN: 978-3-540-33971-7.

Cohn, A. G. and J. Renz (2007). *Qualitative Spatial Reasoning, Handbook of Knowledge Representation*. Ed. by V. L. F. Harmelen and B. Porter. Wiley-ISTE, London: Elsevier.

Dorn, H., T. Törnros, and A. Zipf (2015). "Quality Evaluation of VGI Using Authoritative Data–A Comparison with Land Use Data in Southern Germany". In: *ISPRS International Journal of Geo-Information* 4.3, pp. 1657–1671.

Egenhofer, M. J. and K. K. Al-Taha (1992). "Reasoning about Gradual Changes of Topological Relationships". In: *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space: Proc. of the International Conference GIS*. Ed. by A. U. Frank, I. Campari, and U. Formentini. Berlin, Heidelberg: Springer, pp. 196–219.

Egenhofer, M. J. (1995). "On the Equivalence of Topological Relations". In: *International Journal of Geographical Information Systems* 9, pp. 133–152.

Elwood, S., M. F. Goodchild, and D. Z. Sui (2012). "Researching Volunteered Geographic Information: Spatial data, geographic research, and new social practice". In: *Annals of the Association of American Geographers* 102.3, pp. 571–590.

Falomir, Z. and A.-M. Olteţeanu (2015). "Logics based on Qualitative Descriptors for Scene Understanding". In: *Neurocomputing* 161, pp. 3–16. ISSN: 0925-2312.

Falomir, Z., E. Jiménez-Ruiz, M. T. Escrig, and L. Museros (2011). "Describing Images using Qualitative Models and Description Logics". In: *Spatial Cognition and Computation* 11.1, pp. 45–74.

Falomir, Z., L. Gonzalez-Abril, L. Museros, and J. Ortega (2013a). "Measures of Similarity between Objects from a Qualitative Shape Description". In: *Spatial Cognition and Computation* 13, pp. 181–218.

Falomir, Z., L. Museros, V. Castelló, and L. Gonzalez-Abril (2013b). "Qualitative Distances and Qualitative Image Descriptions for Representing Indoor Scenes in Robotics". In: *Pattern Recognition Letters* 38, pp. 731–743.

Fisher, P. F. (1999). "Models of uncertainty in spatial data". In: *Geographical information systems* 1, pp. 191–205.

Flanagin, A. J. and M. J. Metzger (2008). "The credibility of Volunteered Geographic Information". In: *GeoJournal* 72.3, pp. 137–148.

Fogliaroni, P. (2013). *Qualitative Spatial Configuration Queries. Towards Next Generation Access Methods for GIS*. Vol. 9. Dissertations in Geographic Information Science. ISBN 978-1614992486. IOS Press.

Foody, G. M., L. See, S. Fritz, M. van der Velde, C. Perger, C. Schill, D. S. Boyd, and A. Comber (2015). "Accurate Attribute Mapping from Volunteered Geographic Information: Issues of Volunteer Quantity and Quality". In: *The Cartographic Journal* 52.4, pp. 336–344.

Foth, M., B. Bajracharya, R. Brown, and G. Hearn (2009). "The Second Life of urban planning? Using NeoGeography tools for community engagement". In: *Journal of Location Based Services* 3.2, pp. 97–117.

Freksa, C. (1991). "Conceptual neighborhood and its role in temporal and spatial reasoning". In: *Proceedings of the IMACS Workshop on Decision Support Systems and Qualitative Reasoning*. Ed. by M. G. Singh and L. Travé-Massuyès. North-Holland, Amsterdam, pp. 181–187.

Girres, J.-F. and G. Touya (2010). "Quality assessment of the French OpenStreetMap dataset". In: *Transactions in GIS* 14.4, pp. 435–459.

Goodchild, M. F. (2007). "Citizens as sensors: the world of volunteered geography". In: *GeoJournal* 69.4, pp. 211–221.

Goodchild, M. F. (2008). "Assertion and authority: the science of user-generated geographic content". In: *Proc. of the Colloquium for Andrew U. Frank's 60th Birthday, Department of Geoinformation and Cartography*. Citeseer.

Goodchild, M. F. and L. Li (2012). "Assuring the quality of Volunteered Geographic Information". In: *Spatial statistics* 1, pp. 110–120.

Gouveia, C. and A. Fonseca (2008). "New approaches to environmental monitoring: the use of ICT to explore Volunteered Geographic Information". In: *GeoJournal* 72.3-4, pp. 185–197.

Guesgen, H. W. (1989). *Spatial Reasoning Based on Allen's Temporal Logic*. Tech. rep. International Computer Science Institute.

Haklay, M. (2010). "How good is Volunteered Geographic Information? A comparative study of OpenStreetMap and Ordnance Survey datasets". In: *Environment and planning. B Planning & design* 37.4, p. 682.

Haklay, M. and P. Weber (2008). "OpenStreetMap: user-generated street maps". In: *IEEE Pervasive Computing* 7.4, pp. 12–18. ISSN: 1536-1268.

Hecht, B. and M. Stephens (2014). "A Tale of Cities: Urban Biases in Volunteered Geographic Information". In: *Proceedings of the 8th International Conference on Weblogs and Social Media (ICWSM)*. Michigan, USA.

Jackson, S. P., W. Mullen, P. Agouris, A. Crooks, A. Croitoru, and A. Stefanidis (2013). "Assessing completeness and spatial error of features in Volunteered Geographic Information". In: *ISPRS International Journal of Geo-Information* 2.2, pp. 507–530.

Jiang, J. and M. Worboys (2008). "Detecting Basic Topological Changes in Sensor Networks by Local Aggregation". In: *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. GIS '08. Irvine, California: ACM, 4:1–4:10. ISBN: 978-1-60558-323-5.

Jones, C. B., R. S. Purves, P. D. Clough, and H. Joho (2008). "Modelling vague places with knowledge from the Web". In: *International Journal of Geographical Information Science* 22.10, pp. 1045–1065.

Keßler, C., J. Trame, and T. Kauppinen (2011). "Tracking editing processes in Volunteered Geographic Information: the case of OpenStreetMap". In: *Proceedings of Workshop on Identifying objects, processes and events in spatio-temporally distributed data (IOPE), Conference on Spatial Information Theory (COSIT 2011)*. Vol. 12.

Khan, A., M. Vasardani, and S. Winter (2013). "Extracting Spatial Information From Place Descriptions". In: *Proceedings of The First ACM SIGSPATIAL International Workshop on Computational Models of Place*. COMP '13. Orlando FL, USA: ACM, 62:62–62:69. ISBN: 978-1-4503-2535-6.

Klippel, A., K. Sparks, and J. O. Wallgrün (2015). "PITFALLS AND POTENTIALS OF CROWD SCIENCE: A META-ANALYSIS OF CONTEXTUAL INFLUENCES". In: *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* II-3/W5, pp. 325–331.

Ligozat, G. (2011). *Qualitative Spatial and Temporal Reasoning*. Wiley-ISTE, London: MIT Press. ISBN: 978-1-84821-252-7.

Ludwig, I., A. Voss, and M. Krause-Traudes (2011). "A Comparison of the Street Networks of Navteq and OSM in Germany". In: *Advancing Geoinformation Science for a Changing World*. Ed. by W. Geertman Stanand Reinhardt and F. Toppen. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 65–84. ISBN: 978-3-642-19789-5.

McDougall, K. (2009). "The potential of citizen volunteered spatial information for building SDI". In: *Proc. of 11th World Conference on Spatial Data Infrastructure Convergence: Building SDI Bridges to Address Global Challenges*. Rotterdam, The Netherlands: GSDI Association Press.

Mooney, P. and P. Corcoran (2012a). "Characteristics of heavily edited objects in OpenStreetMap". In: *Future Internet* 4.1, pp. 285–305.

Mooney, P. and P. Corcoran (2012b). "The annotation process in OpenStreetMap". In: *Transactions in GIS* 16.4, pp. 561–579.

Neis, P., D. Zielstra, and A. Zipf (2011). "The street network evolution of crowdsourced maps: OpenStreetMap in Germany 2007–2011". In: *Future Internet* 4.1, pp. 1–21.

Neis, P., D. Zielstra, and A. Zipf (2013). "Comparison of Volunteered Geographic Information Data Contributions and Community Development for Selected World Regions". In: *Future Internet* 5.2, pp. 282–300.

Nittel, S., L. Bodum, K. C. Clarke, M. Gould, P. Raposo, J. Sharma, and M. Vasardani (2015). "Emerging Technological Trends likely to Affect GIScience in the Next 20 Years". In: *Advancing Geographic Information Science: The Past and Next Twenty Years*. Ed. by H. Onsrud and W. Kuhn. Global Spatial Data Infrastructure Association (GSDI).

Pourabdollah, A., J. Morley, S. Feldman, and M. Jackson (2013). "Towards an authoritative OpenStreetMap: conflating OSM and OS OpenData national maps' road network". In: *ISPRS International Journal of Geo-Information* 2.3, pp. 704–728.

Randell, D. A., Z. Cui, and A. Cohn (1992). "A Spatial Logic Based on Regions and Connection". In: *Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning*. Ed. by B. Nebel, C. Rich, and W. Swartout. San Mateo, California: Morgan Kaufmann, pp. 165–176.

Richter, K.-F., M. Tomko, and A. Coltekin (2015). "Are We There Yet? Spatial Cognitive Engineering for Situated Human-Computer Interaction". In: *Proceedings of Workshop on Cognitive engineering for spatial information processes: From user interfaces to model-driven design, Conference on Spatial Information Theory (COSIT 2015)*. Ed. by S. Bertel, P. Kiefer, A. Klippel, S. Scheider, and T. Thrash.

Richter, K., B. Weber, B. Bojduj, and S. Bertel (2010). "Supporting the designer's and the user's perspectives in computer-aided architectural design". In: *Advanced Engineering Informatics* 24.2, pp. 180–187.

Roche, S., E. Propeck-Zimmermann, and B. Mericskay (2013). "GeoWeb and crisis management: Issues and perspectives of Volunteered Geographic Information". In: *GeoJournal* 78.1, pp. 21–40.

Schmid, F., L. Frommberger, C. Cai, and F. Dylla (2013a). "Lowering the barrier: How the What-You-See-Is-What-You-Map paradigm enables people to contribute Volunteered Geographic Information". In: *Proc. of the 4th Annual Symposium on Computing for Development*. ACM. Cape Town, South Africa, pp. 8–18.

Schwering, A., J. Wang, M. Chipofya, S. Jan, R. Li, and K. Broelemann (2014). "SketchMapia: Qualitative Representations for the Alignment of Sketch and Metric Maps". In: *Spatial Cognition & Computation* 14.3, pp. 220–254.

Song, W. and G. Sun (2010). "The role of mobile Volunteered Geographic Information in urban management". In: *Proc. of the 18th International Conference on Geoinformatics*. IEEE, pp. 1–5.

Sparks, K., A. Klippel, J. O. Wallgrün, and D. Mark (2015). "Citizen Science Land Cover Classification Based on Ground and Aerial Imagery". In: *Spatial Information Theory: 12th International Conference, COSIT 2015, Santa Fe, NM, USA, October 12-16, 2015, Proceedings*. Ed. by I. S. Fabrikant, M. Raubal, M. Bertolotto, C. Davies, S. Freundschuh, and S. Bell. Cham: Springer International Publishing, pp. 289–305. ISBN: 978-3-319-23374-1.

Thabtah, F. (2007). "A review of associative classification mining". In: *The Knowledge Engineering Review* 22.01, pp. 37–65.

Tobler, W. R. (1970). "A computer movie simulating urban growth in the Detroit region". In: *Economic geography* 46, pp. 234–240.

Vandecasteele, A. and R. Devillers (2013). "Improving Volunteered Geographic Data Quality Using Semantic Similarity Measurements". In: *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 1.1, pp. 143–148.

Vasardani, M. and M. Egenhofer (2009). "Comparing Relations with a Multi-holed Region". In: *Spatial Information Theory: 9th International Conference, COSIT 2009 Aber Wrach́, France, September 21-25, 2009 Proceedings*. Ed. by K. S. Hornsby, C. Claramunt, M. Denis, and G. Ligozat. Vol. 5756. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 159–176. ISBN: 978-3-642-03832-7.

Vasardani, M., S. Winter, and K.-F. Richter (2013). "Locating place names from place descriptions". In: *International Journal of Geographical Information Science* 27.12, pp. 2509–2532.

Witten, I. H. and E. Frank (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd ed. San Francisco: Morgan Kaufmann, p. 119.

Wolter, D., F. Dylla, and A. Kreutzmann (2011). "Rule-Compliant Navigation with Qualitative Spatial Reasoning". In: *Robotic Sailing*. Springer, pp. 141–155.

Zielstra, D. and A. Zipf (2010). "Quantitative studies on the data quality of OpenStreetMap in Germany". In: *Proc. of the 6th International Conference on Geographic Information Science (GIScience), Zurich, Switzerland, September 14-17, 2010*. Zurich, Switzerland, pp. 20–26.

Chapter 6

Guided Classification System for Overlapping Classes in OpenStreetMap

Authors:

Ahmed Loai Ali, Nuttha Sirilertworakul, Alexander Zipf, and Amin Mobasheri.

Journal:

ISPRS International Journal of Geographic Information.

Citation:

Ahmed Loai Ali, Nuttha Sirilertworakul, Alexander Zipf, and Amin Mobasheri (2016). "Guided classification system for overlapping classes in OpenStreetMap". In: *ISPRS International Journal of Geo-Information. 5.6. p.87. ISSN: 2220-9964.*

Status:

The article is published and available online since 07 June 2016.

Contribution Statement:

This work shows the validation of my proposed approach. The web site was developed by Nuttha based on my previous empirical study. I was involved in processing the data sets, in applying the proposed approach, and in writing the manuscript. Moreover, I analyzed the results in collaboration with Amin, while Alexander contributed in discussing the results and providing a proof-reading that substantially improved the text.

Abstract:

The increased development of Volunteered Geographic Information (VGI) and its potential role in GIScience studies raise questions about the resulting data quality. Several studies address VGI quality from various perspectives like completeness, positional accuracy, consistency, etc. They mostly have consensus on the heterogeneity of data quality. The problem may be due to the lack of standard procedures for data collection and absence of quality control feedback for voluntary participants. In our research, we are concerned with data quality from the classification perspective. Particularly in VGI-mapping projects, the limited expertise of participants and the non-strict definition of geographic features lead to conceptual overlapping classes, where an entity could plausibly belong to multiple classes; e.g., *lake* or *pond*, *park* or *garden*, *marsh* or *swamp*, etc. Usually there exist quantitative and/or qualitative characteristics that distinguish between classes. Nevertheless, these characteristics might not be recognizable for non-expert participants. In previous work, we developed the rule-guided classification approach that guides participants to the most appropriate classes. As exemplification, we tackle the conceptual overlapping of some grass-related classes. For a given data set, our approach present the most highly recommended classes for each entity. In this paper, we present the validation of our approach. We implement a web-based application called `Grass&Green` that presents recommendations for crowdsourcing validation. The findings show the applicability of the proposed approach. In four months, the application attracted 212 participants from more than 35 countries, who checked 2,865 entities. The results indicate that 89% of the contributions fully/partially agree with our recommendations. We then carried out a detailed analysis that demonstrates the potential of this enhanced data classification. This research encourages the development of customized applications that target a particular geographic feature.

Keywords:

Volunteered Geographic Information (VGI); Classification; Spatial Data Quality; OpenStreetMap (OSM)

6.1 Introduction

Web and information revolutions, the increased availability of location sensing devices, and the advanced communication technologies facilitate the evolution of free geographic content, which is known as Volunteered Geographic Information (VGI) (Goodchild, 2007). In particular, we are concerned with the VGI format, in which the public participates in mapping processes regardless of their prior geographic experience. In the past,

these processes were performed exclusively by cartographers at mapping agencies and in specialized organizations. Among others, OpenStreetMap[1] (OSM), Wikimapia[2], and Google Map Maker[3]are examples of VGI-based mapping projects. With the expansion of crowdsourcing, participants have developed a tremendous amount of free geographic data that have been utilized in various applications. For example, VGI acts as a potential data source for applications of environmental mapping (Gouveia and Fonseca, 2008; Mooney and Corcoran, 2011), crisis management (Roche et al., 2013; Zook et al., 2010), urban planning (Foth et al., 2009; Mooney et al., 2011), map provision (Haklay and Weber, 2008), and location-based services (LBS) (Mooney et al., 2011; Savelyev et al., 2011). However, in each application, the data quality is an issue of high concern. Several studies have concluded that the quality of VGI is heterogeneous Elwood et al., 2012. This finding impacts on the utility of VGI as a complementary source or as an alternative to authoritative data sources (Ali and Schmid, 2014; Devillers et al., 2010; Goodchild and Li, 2012; Goodchild, 2008).

In general, VGI — as spatial data — has multiple measures of data quality such as: completeness, lineage, logical consistency, positional accuracy, and semantic (attribute) accuracy (Guptill and Morrison, 2013). In our research, we are concerned with the attribute accuracy. In particular, we investigate data quality from the viewpoint of classification, i.e., whether a piece of land covered by grass is being classified as *park, garden,* or *forest*; if an areal water body belongs to the *lake, pond* or *reservoir* class, etc. In VGI projects, data classification is mainly based on participants' cognition. On one hand, the appropriate classification depends on quantitative (e.g., size, area) and/or qualitative (e.g., context) characteristics. However, these characteristics, which distinguish between classes, might not be observed by participants. In addition, the non-standard data collection procedures and the limited expertise of participants may result in heterogeneous data classification. On the other hand, the non-strict definition of geographic features leads — in some cases — to conceptual overlapping classes. Thus, a given entity may be classified as *lake* or *pond*, *park* or *garden*, *marsh* or *swamp* and it could plausibly belong to multiple classes, but only small details might distinguish between the most appropriate class (Ali et al., 2015; Ali et al., 2016).

To tackle the aforementioned problems, we propose the rule-guided classification approach in our previous work Ali et al. (2015) and Ali et al. (2016). The approach learns the distinct qualitative characteristics of specific classes and encodes them into predictive rules. Afterwards, the extracted rules are organized into a classifier that acts to guide

[1] http://openstreetmap.org/
[2] http://www.wikimapia.org/
[3] https://www.google.com/mapmaker

the participants towards the most appropriate classes. In this paper, we propose crowdsourcing validation as one of many possible implementation scenarios of our approach. In this scenario, we present a set of entities associated with our recommended classes to the crowd for the purpose of validation.

In this paper, we present the `Grass&Green` application (http://www.opensciencemap.org/quality): a web-app that addresses the conceptual overlapping challenge of some grass-related classes. We utilized the data from the OSM project, particularly the data set of Germany. However, the results were presented to the entire OSM mappers as well as public participants. We selected the classes of *garden, grass, forest, park*, and *meadow* as an exemplification of the conceptual overlapping problem. The choice is based on the following reasons: i) In the utilized data set, they are the most common grass-related classes within city boundaries (our geographic scope of research); and ii) For non-experts, there exists conceptual overlapping between these classes, since they are related to the global concept of grass, but with finer differences. We launched the application to validate our previous work in Ali et al. (2015) and Ali et al. (2016). The participants were allowed to express their agreement/disagreement with the recommended classes. In addition, the participants were encouraged to send us feedback and comments. We announced the application on OSM diaries[4] and other social media blogs. In four months, the application attracted 212 participants from more than 35 countries. During this period, the participants checked 2,865 entities. The findings indicate the applicability of the proposed approach. Around 89% of the contributions are fully/partially in agreement with our recommended classes. Moreover, the detailed investigation of the results demonstrates the enhanced classification of the target entities. We received positive feedback from participants, which encourages the expansion of the application of the proposed approach to different locations. Moreover, the findings of this work motivate the development of more customized applications that handle a particular geographic feature in order to enhance the data quality of voluntary geographic data sets.

This paper is organized as follows. Section 6.2 provides an overview about related works. The reasons for problematic data classification in VGI projects, including subjective classification, participant heterogeneity, and conceptual overlapping classes are discussed in Section 6.3. A summary of our proposed approach is provided in Section 6.4. The `Grass&Green` application is presented in Section 6.5 including: the description, the conceptual architecture, and the announcement methodologies. Section 6.6 illustrates the results from various perspectives. A vision of the proposed approach with respect to enhancing data quality is provided in Section 6.7. Section 6.8 concludes the paper and highlights some future research directions.

[4] http://www.openstreetmap.org/diary

6.2 Related Work

With the increased availability of VGI sources, the resulting data quality has been raised as an issue of high concern in GIScience (Goodchild, 2008; Elwood et al., 2012; Devillers et al., 2010). Most research has targeted the OSM project, as the most prominent VGI mapping project. The project aims to develop a free world digital map editable and obtainable by everyone (Haklay and Weber, 2008). Currently, OSM data covers most of the world and the project has more than 2,500,000 registered users at 10^{th} April 2016 according to OSMstats[5] website. Several research studies have addressed the quality from various perspectives like the assessment of the resulting data (Section 6.2.1) and the development of approaches and methodologies to enhance the data quality (Section 6.2.2). Other research has focussed on data classification in user-generated geographic contents (Section 6.2.3).

6.2.1 VGI quality assessment

Generally, geo-spatial data are assessed either by comparison with an authoritative data source or by analyzing the intrinsic properties of the data. The assessment is carried out based on the standard spatial data quality measures developed in ISO/TC[6] 211 (Østensen and Smits, 2002). The OSM data are compared with the authoritative data in the UK, Germany, Canada and France (Haklay, 2010; Ludwig et al., 2011; Arsanjani et al., 2015; Dorn et al., 2015; Vaz and Jokar Arsanjani, 2015; Girres and Touya, 2010). With the evolution of VGI, authors in Goodchild and Li (2012) argue that there are three dimensions in assessing VGI data: crowdsourcing, social, and geographic dimensions. Hence, the intrinsic properties of data like contributors' reputation, editing history, and data evolution have been analyzed to assess data quality (Flanagin and Metzger, 2008; Bishr and Kuhn, 2007; Neis and Zipf, 2012; Neis et al., 2011; Keßler and Groot, 2013; Keßler et al., 2011; D'Antonio et al., 2014; Neis et al., 2013). Researchers have investigated different quality measures like positional accuracy, completeness, and thematic accuracy with respect to various geographic features like road networks, buildings, and land use features. Another perspective of quality assessment has been presented in Ballatore and Zipf (2015), where the data quality is associated with the purpose of use. In Barron et al. (2014), the authors presented a framework to assess the data quality conceptually.

Most of the research concludes that VGI is a potentially valuable data source, particularly in urban places (Hecht and Stephens, 2014). Nevertheless, they mostly agree on the

[5] http://osmstats.neis-one.org/
[6] http://www.isotc211.org/

heterogeneous quality of the data with respect to various quality measures (Goodchild and Li, 2012; Devillers et al., 2010).

6.2.2 VGI quality enhancement: approaches & methods

Several economic and cultural factors influence data quality in VGI-mapping projects (Quattrone et al., 2014; Neis et al., 2013). To our knowledge, there are only a limited number of research studies concerned with enhancing the data quality in VGI-based mapping projects.

In Schmid et al. (2012) and Schmid et al. (2013a), the authors argue that intuitive human interfaces can play a role in producing data of high quality. The work in Pourabdollah et al. (2013) encourages conflating OSM and authoritative data to develop an integrated open data source while Vandecasteele and Devillers (2013) present a semantic solution that aids the contributors during the editing process toward enhanced data quality, to overcome cross-cultural and multi-language problems. Moreover, Ali et al. (2014) and Ali and Schmid (2014) discussed the utilization of learning to enhance the data classification of VGI projects. In Ali et al. (2015) and Ali et al. (2016), we presented the rule-guided classification approach, which acted to generate recommended classes to improve the classification quality. As an alternative, "Gamification" has been presented as another method for enhancing VGI quality (Yanenko and Schlieder, 2014).

For the OSM project in particular, OSMRec[7] is presented in Karagiannakis et al. (2015); it is an editor plugin tool for automatic annotation of spatial entities in the OSM project. In addition, OSM Inspector[8], KeepRight[9], MapRoulette[10], and MapDust[11] are examples, among others, of web-applications that have been developed to enhance the data quality of the project. These applications have been either customized for a particular feature in a particular location like NOVAM[12], which manages bus stop features in the UK, or they have been developed generally for multiple features in various locations. These applications encourage the role of participants to enhance data quality through crowdsourcing revision.

[7] https://github.com/GeoKnow/OSMRec
[8] http://tools.geofabrik.de/osmi/
[9] http://keepright.ipax.at/
[10] http://maproulette.org/
[11] http://www.mapdust.com/
[12] http://b3e.net/novam/

6.2.3 Human-centered data classification

Other research has focus particularly on the data classification in user generated geo-spatial content. In VGI, the data classification is human-centered; the data are classified based on individual perceptions rather than on a pre-defined model as is the case in professional data classification. The authors in Fisher (1999) presented different forms of spatial data uncertainty, which influence the classification precision and granularity. In Ali et al. (2014), the authors analyzed the plausible and ambiguous classification in VGI. Nevertheless, the research in Sparks et al. (2015) concludes the ability of the public to precisely classify land cover features when they are provided with aerial and ground photos. The work of Klippel et al. (2015) studied cultural, linguistics, and regional influences on the data classification while the authors of Foody et al. (2015) investigate the classification quality of land use and land cover features in VGI with respect to the contributors and the provided data.

The authors of Fritz et al. (2012) have developed Geo-Wiki[13] (a crowdsourcing web-application) to validate and enhance the classification of global land cover data. Geo-Wiki also aims to develop a hybrid global land cover map from different data sources, where the authoritative data sources are enhanced with open sources and the power of crowdsourcing is used for validation.

In Mooney and Corcoran (2012b), the authors studied the annotation process in the OSM project. They identified the problem of using OSM data taxonomy and its impacts on data classification. From a particular point of view, the cross-cultural nature of the OSM project results in heterogenous data classification of identical geographic features, and hence, limited use of the data. However, semantic solutions have been used to overcome this problem (Ballatore et al., 2013; Baglatzi et al., 2012).

Nevertheless, the research in Arsanjani et al. (2015) and Vaz and Jokar Arsanjani (2015) has assessed the classification accuracy of land use and land cover features in the OSM project. They highlighted the remarkable data quality and the potential utilization of VGI as a complementary data source of these features.

6.3 Beyond Data Classification in VGI Projects: the case of OpenStreetMap

Several research studies have emphasized the significance of VGI sources. However, they also highlight their problematic data classification: in most applications, imprecise

[13] http://www.geo-wiki.org/

data classification results in either incorrect or incomplete results. How are the data classified? Do the data follow a strict classification model? How could we verify the data classification? At which granularity level is the data classification complete? All of these are critical issues that will impact on the effective utilization of VGI sources. Thus, this section gives an insight into the classification challenges in VGI projects. In this paper, we analyzed the OSM data. The impacts of the contribution mechanism and the utilized data models on data quality are presented in Section 6.3.1. In any VGI projects, participants play a major role in the data collection process. Thus, the OSM communities and their influence on data classification are addressed in Section 6.3.2 whereas Section 6.3.3 discusses general difficulties of geographic data classification.

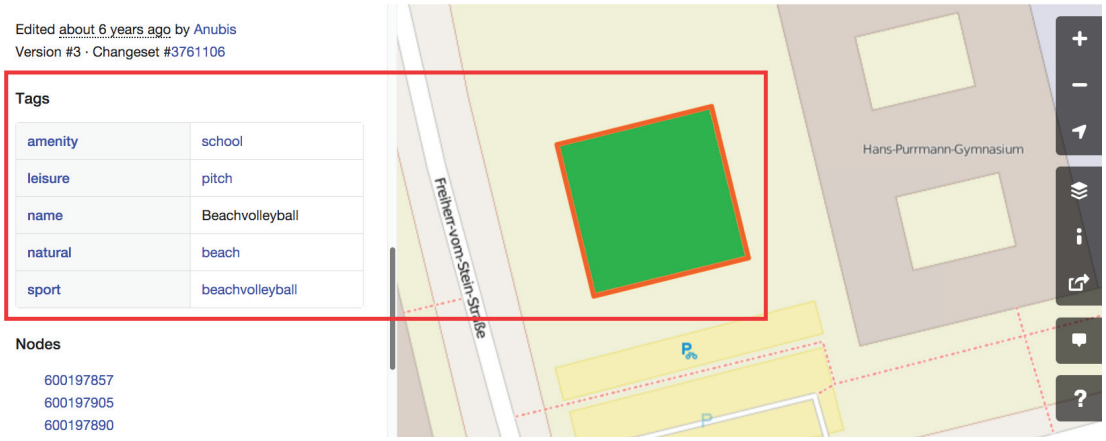


FIGURE 6.1: An example of problematic classification in the OSM project: the highlighted entity is classified as *pitch, school*, and *beach*, while it is actually a beach volleyball playground in a school.

6.3.1 Classification by tags (`key` = `value`)

In OSM, the contributions are performed by participants as follows: the participants delineate geographic features from provided satellite images (e.g., Bing aerial images), by using one of the OSM editors (e.g., iD editor). The features are represented as entities using the appropriate data models: point (0-D features), way (linear features), and relation (complex features). Afterwards, the participants are free to describe and classify the contributed entity by means of tags; when a tag has the format of `key` = `value`, the `key` describes the classification perspective and the `value` is the class label. For example, the tag of `natural` = `water` describes the natural coverage of an entity as a water body, while an additional tag, e.g., `water` = `lake`, is required to express the precise classification.

The OSM project presents the recommended tags and appropriate ways of mapping various geographic features on its Wiki pages[14]. However, the lack of integrity checking mechanisms and the complete free contribution mechanisms result in problematic classification. For example, an entity could be assigned no tags or infinite tags and even the repetition of tags is possible, e.g., `natural` = `water` and `natural_1` = `sand`. Although these flexible mechanisms allow participants to initiate new classes, they generate various challenges during data processing and cleaning. Figure 6.1 illustrates a problematic classification example, when the indicated entity is assigned to conflicting classes.

6.3.2 Subjective classification

VGI mapping projects are run by the power of crowds. The contributions come from the local knowledge of participants. They are free to translate their observation into an annotated geographic feature with description/categorization/classification. As humans interpret the observations differently, they may perceive the geographic features differently; a given entity might be classified as a *restaurant* by a participant, but it may be categorized by others as a *cafe*; whether a water body is large enough to be classified as a *lake* or small enough to be appropriately classified as a *pond*; these classifications depend on rational and individual aspects. This fact leads to subjective classification.

In the OSM project, participants have unequal mapping and cartographic experience; they come from different cultures; and they have various educational backgrounds and interests. Thus, the heterogeneous participants boost the problematic classification. Incomplete and inconsistent classification are examples of the problems related to subjective classification.

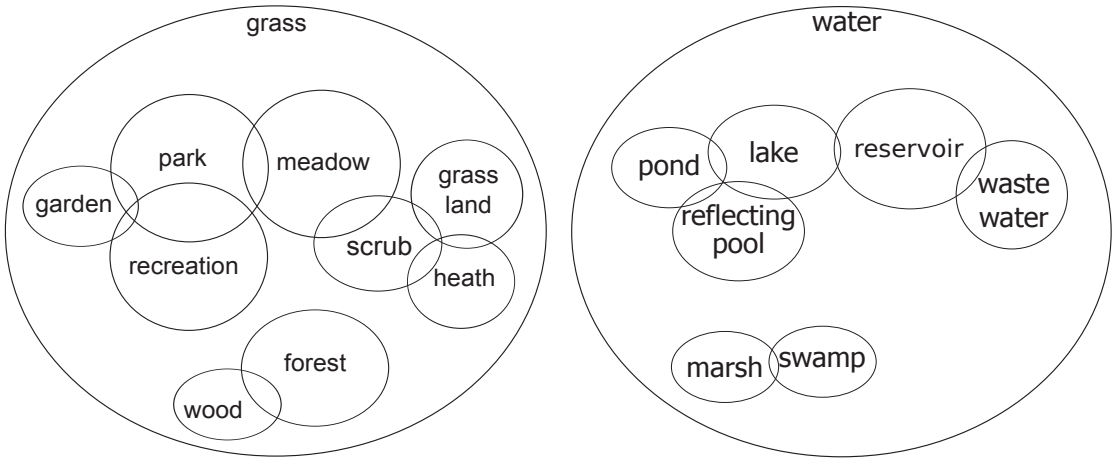
- Incomplete classification: the limited local knowledge of a participant or the unclear perceived observation from the provided satellite images impacts on the classification granularity. In a pilot study on the OSM data set of Germany (May 2015), we found 225,933 entities related to water body classes. Only 20% out of these entities have further finer classes like *lake, pond, waste water, reservoir,* etc. We detected about 10,520,418 unclassified building entities, which have a coarser classification as *building* while other entities of *building* are classified into finer classes like *residential, service, public, industrial, house,* etc.
- Inconsistent classification: when participants interpret a given feature differently, they assign it to conflicting classes or an ambiguous class. During our investigations, we found out some entities are assigned to conflicting classes; some entities

[14] http://wiki.openstreetmap.org/wiki/Map_Features

are classified as *meadow* (i.e. grass land) and *wetland* (i.e. water body). Figure 6.1 illustrates a clear example of the classification inconsistency, when the given entity is classified by the *pitch, school*, and *beach* classes.

6.3.3 Conceptual overlapping classes

In general, spatial data are prone to various forms of uncertainty: probability, vagueness, and ambiguity. The problem might be related to whether a geographic feature is well or poorly defined (Fisher, 1999). In Comber et al. (2006) and Grira et al. (2010), the authors link the uncertainty of the spatial data with the VGI quality. In particular, poor definitions lead to crisp boundaries between similar classes. Thus, a particular entity could plausibly belong to multiple overlapping classes with various degrees of accuracy. Nevertheless, there are usually qualitative and/or quantitative characteristics that could distinguish between these classes.



(A) Overlapping grass-related classes. (B) Overlapping water-related classes.

FIGURE 6.2: Conceptual overlapping classes due to the given descriptions in the OSM Wiki.

Among others, the features of water bodies, grass-related, and wetland are examples of features with non-strict definitions, and hence, they include overlapping classes. Figure 6.2 illustrates the conceptual overlapping classes within grass-related and water body features, with respect to the recommendations given in the OSM Wiki. Table **??** describes the mapping between the OSM tags and their corresponding classes. In the OSM project, a single class could be described by various tags; however, we investigate the most common tagging. The overlapping between classes in the figure is based on sharing a particular concept or common characteristics. Moreover, the size of overlapping indicates the degree of conceptual similarity.

For example, the *park*, *recreation*, and *garden* are overlapping classes in Figure 6.2a: they share the characteristics of being used for entertainment and amusement. The classes

OSM tag	Class	OSM tag	Class
`landuse` = `grass` or `landcover` = `grass`	grass	`natural` = `wood` or `wood` = `yes`	wood
`leisure` = `park`	park	`natural` = `water`	water
`leisure` = `garden`	garden	`natural` = `water` `water` = `lake`	lake
`landuse` = `recreation ground`	recreation	`natural` = `water` `water` = `pond`	pond
`landuse` = `meadow`	meadow	`natural` = `water` `water` = `reflecting_pool`	reflecting pool
`natural` = `scrub`	scrub	`natural` = `water` `water` = `reservoir`	reservoir
`natural` = `grassland`	grassland	`natural` = `water` `water` = `wastewater`	waste water
`natural` = `heath`	heath	`natural` = `wetland` `wetland` = `swamp`	swamp
`landuse` = `forest`	forest	`natural` = `wetland` `wetland` = `marsh`	marsh

TABLE 6.1: Mapping between OSM tags and some of grass-related and water-related overlapping classes.

of *park*, *garden* are classified by the `leisure` key, while the *recreation* class is described by the `landuse` key. However, the *recreation* entities are most likely related to certain activities (e.g., sport, or social activities), the *garden* entities are more cultivated with flowers and plants than others, and the *park* entities are in general larger than garden and recreation and might include both of them as well. Figure 6.2b shows another example of overlapping classes related to water body features. When a water body is stagnant and natural, it could be classified as *lake* (if it is large) or as *pond* (if it is small), but when it is man-made it would be more appropriately classified as *reservoir*. Other classes such as *marsh* and *swamp* are both describing the land area that is saturated with water, either permanently or seasonally. In the OSM data, they are both described by the `wetland` key. Only the type of vegetation distinguishes between the classes: *swamp* when woody vegetation and *marsh* when non-woody vegetation and open habitats.

The previous discussions summarize the reasons behind the problematic classification in VGI projects; Sections 6.3.1 and 6.3.2 argue the problem from the nature of VGI projects, while Section 6.3.3 discusses the problem from the perspective of spatial data uncertainty. These classification problems impact not only on the data quality, but they also limit the development of general applications, e.g., global rendering and visualizing applications. Moreover, the problematic data quality will determine the utility of VGI sources for particular types of application.

6.4 Rule-Guided Classification Approach

In Ali et al. (2015) and Ali et al. (2016), we tackled the classification by developing the rule-guided classification approach. In VGI projects, participant conceptualization of geographic features impacts on the data classification. From a human cognitive perspective, people are likely to investigate the qualitative characteristics of a given feature in order to classify it appropriately. Moreover, humans implicitly contrast between similar classes to infer a certain class instead of others. For example, we contrast between *park* and *forest* classes by looking into the coverage of trees, the availability of amusement and entertainment facilities, and the accessibility for pedestrians. Hence, our approach exploits the qualitative characteristics and comparison to distinguish between similar classes. For particular entities of overlapping classes, we apply a machine learning mechanism to extract the distinct qualitative topological characteristics that identify each class. These characteristics are formulated and organized to develop a classifier. Then, the approach employs the developed classifier to re-classify the entities and presents them again for crowdsourcing validation. In this approach, we assume that identical entities should be classified similarly within the same country (i.e., localized classification). Thus, learning from data of India and applying the extracted knowledge on data of Germany might lead to another problematic classification, due to different cultures and concepts. For further details see Ali et al. (2016).

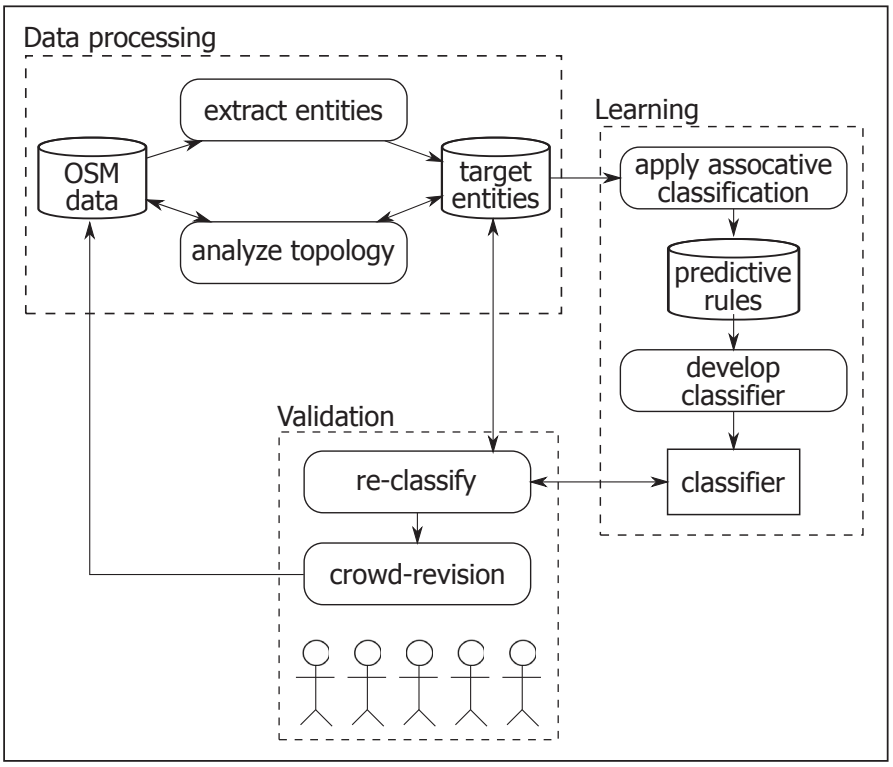


FIGURE 6.3: Conceptual structure of the rule-guided classification approach.

Figure 6.3 illustrates the conceptual structure of the rule-guided classification approach. For exemplification, we demonstrate the approach on a case study. We utilize the OSM data set of Germany and target the classification of some grass-related classes: *grass, garden, forest, park*, and *meadow*. The choice of the Germany data set is due to the following reasons: a) in Germany, there exists an active mappers community on the OSM project; b) several studies confirmed the high quality of data, particularly in the urban areas; and c) there is no large bulk import of data. Figure 6.3 divides the approach into three phases: data processing, learning, and validation phases.

1. **Data processing**:
   From the OSM data set of Germany, we extracted the entities of target classes. The entities are extracted from the most densely populated cities to ensure data of high quality. We are concerned with the areal entities. Thus, to understand the qualitative characteristics of the classes, we topologically checked each individual entity. We developed an automatic algorithm using the 9-Intersection Model (9IM) to perform the investigation (Egenhofer and Al-Taha, 1992). This investigation aims to find out the common topological relations between pairs of entities; these relations are potentially useful to distinguish between similar classes. For example, find the relation between pairs of entity (E_1, E_2), when E_1 represents the target feature (e.g., *park* entity) and E_2 is another kind of nearby feature to E_1 (e.g., playground, water bodies, etc.).

2. **Learning**:
   The target of the learning phase is developing a classifier able to potentially distinguish between similar classes. We apply an associative classification Thabtah, 2007 data mining mechanism to perform the learning task. This mining approach utilizes the association rule to construct the classification system (Thabtah, 2007). First, we extract a set of predictive rules that describe each class, and then these rules were ranked and organized into the classifier. During the classification process, a given entity is matched against the entire extracted set of rules. The matched rules are ranked in descending order based on their confidence measures. Due to the overlapping problem (see Section 6.3), the developed classifier is configured to give the two most appropriate classes instead of picking out a single class.

3. **Validation**:
   Due to the nature of VGI, the proposed approach exploits crowdsourcing to validate the classification. The entities are re-classified using the developed classifier. Afterwards, they are presented to the public again for the purpose of revising the recommended classes. The validation phase has multiple functionalities: a)

enhance/ensure the target entities' classification by crowdsourcing revision, b) understand the public conception of target classes, and c) find out the response of participants to the provided recommendations.

The first and second phases are presented with more details in a previous work Ali et al. (2015) and Ali et al. (2016) while, this paper focuses on the third phase, where the implementation of the validation phase is presented in the next section.

6.5 `Grass&Green`: Customized Quality Assurance Application

As a validation of the rule-guided classification approach, we developed a web application called `Grass&Green`[15]. We adopted a web-based architecture to reach a broad number of participants. The application has been launched since August 2015, and targeted at public participants and OSM mappers as well. The application is hosted on an Ubuntu[16] server as a sub-branch of the OpenScienceMap[17] (OScieM) project.

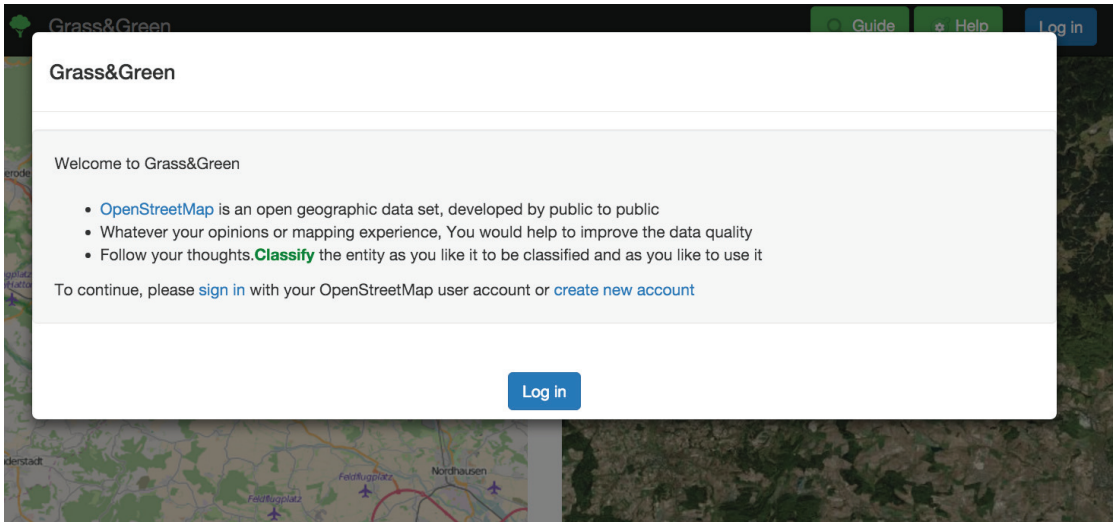


FIGURE 6.4: Application instructions and the OSM user login options.

The application description is presented in Section 6.5.1. Section 6.5.2 demonstrates the application architecture and its components while the utilized channels to attract participants are discussed in Section 6.5.3.

[15] http://www.opensciencemap.org/quality/
[16] http://www.ubuntu.com/server
[17] http://www.opensciencemap.org/

6.5.1 Application description

Figures 6.4, 6.5, and 6.6 illustrate the user interface (UI) of the application. The interface usability and ease of use are of our concern to achieve the application objectives and to simulate the nature of VGI projects as well. Before logging in, `Grass&Green` presents the instructions for use to the participant. As we contribute directly to the OSM project, participants must have an OSM user account. The application allows non-OSM users to register for an account (see Figure 6.4).

For non-expert participants, the application has a menu called "Guide" that introduces the class descriptions. The descriptions are provided visually and as text from multiple sources: Wikipedia, OSM Wiki, and WordNet[18] (see Figure 6.5).

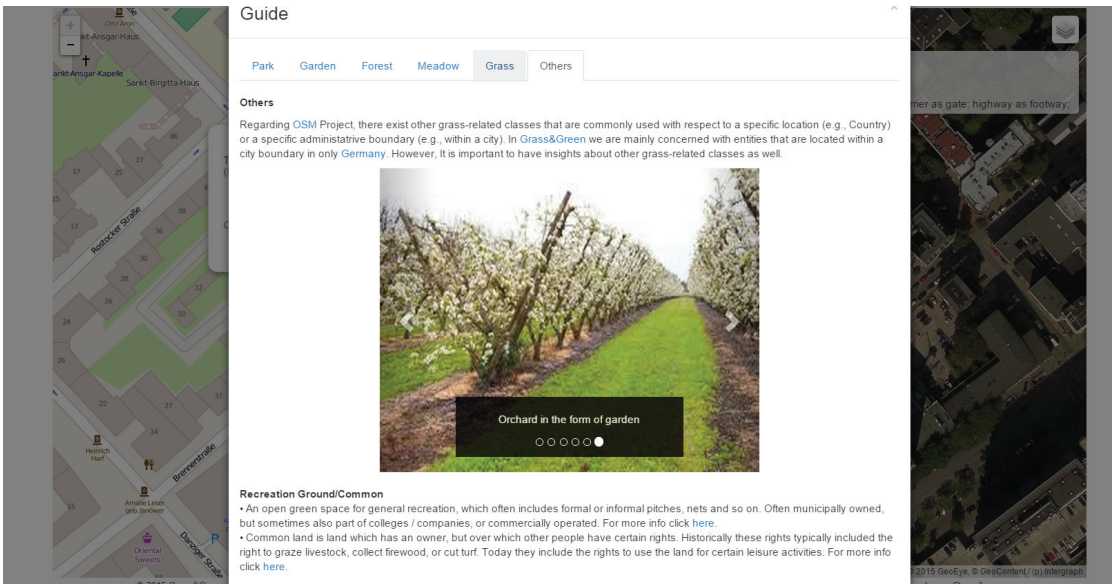


FIGURE 6.5: Textual and visual descriptions of target classes.

After login, the application shows the entities to the participant randomly. Figure 6.6 shows the simple interface of the revision process. On the right hand side, the given entity is outlined and overlapped with Bing satellite images, which is an aerial image provider. In addition, the topological qualitative descriptions of the entity are provided as text. For example, the given entity in Figure 6.6 CONTAINS trees, ADJACENT to a building, a garden, and a service way, and COVERED BY a residential area. On the left hand side, the entity is outlined and overlapped with the OSM base map. Over the entity, a pop-up message shows the recommended classes (marked as recommended) and the other classes as well. The validation is flexible, similar to the contribution mechanism of the OSM project; the participant could select between “yes”, “no”, and “maybe” options from the provided classes. The participant could deselect our recommendations and select other

[18] https://wordnet.princeton.edu/

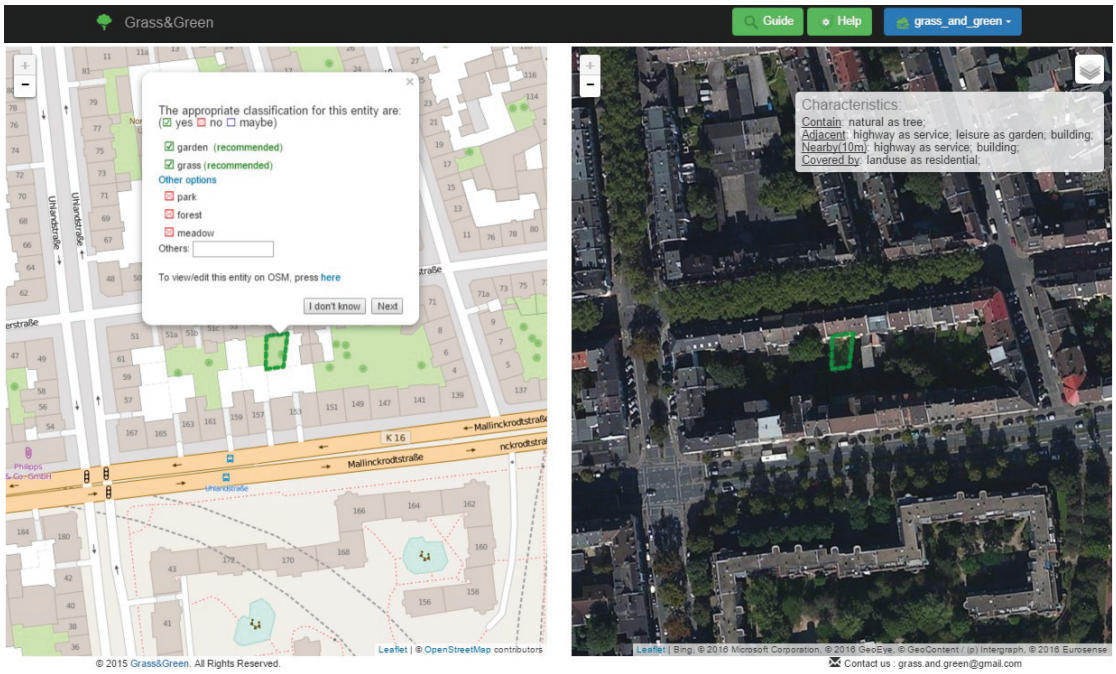


FIGURE 6.6: Validation interface for the presented entities.

classes or add a new class (if required). More options are provided for the participant like view and edit the entity directly through the OSM project interfaces. In both maps, a zoom in/out option is provided to enable the participants to explore the geographical context.

Furthermore, the "Help" menu provides participants with the instructions at anytime if required. At the bottom, a contact e-mail address is given for further feedback and comments from interested participants. At any point, participants are allowed to logout or simply close the application to exit the validation process.

6.5.2 Application architecture

As a web-based application, `Grass&Green` consists of front-end and back-end components; the front-end components control the usability and the visualization in the UI like the leaflet[19] component, the Bootstrap[20] framework, and the JQuery[21] library while the back-end components are responsible for performing efficient and reliable communications among application layers. Figure 6.7 shows how the application is composed of three layers: interface layer, data layer, and external layer.

[19] http://leafletjs.com/
[20] http://getbootstrap.com/
[21] https://jquery.com/

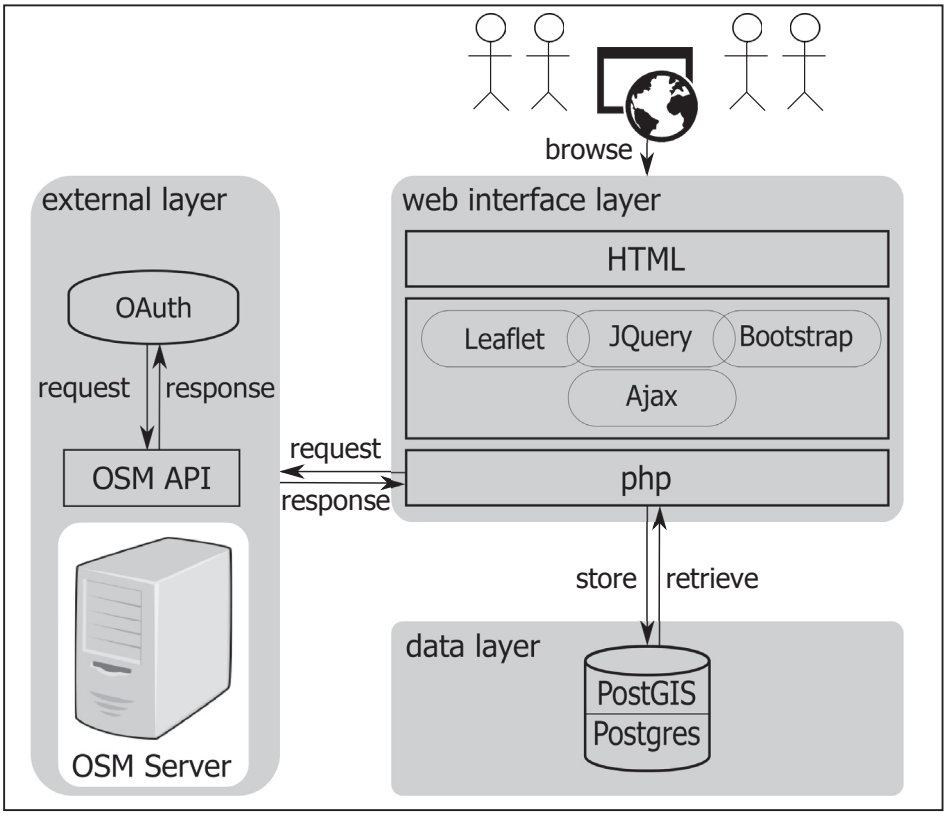


FIGURE 6.7: The `Grass&Green` application structure.

Using any internet browser, the participants can access the interface layer. First, the participants login to the application using the authorization open standard of OAuth[22], which allows them to connect to a third party website — in this case the OSM project — in a secure way without exposing their password. After successful login, the interface layer, by means of Ajax and php, starts to call the data from the data layer for the validation process. By means of php functions, the application controls the validation results and participant contributions. The data layer contains the data set developed by the proposed approach in Ali et al. (2016). In the data set, each entity is associated with its topological qualitative characteristics, its geometry, and two recommended classes. The data set is stored in a Postgres data base with postGIS extension to handle the geometry of entities. As an external layer, the OSM server is accessed through the OSM Application Program Interface (API). We used the OSM user account as a reference to participant experience and their geographic origin. During the validation, participants have options to edit/view the presented entities by OSM editors/viewers. In addition, the interface layer calls the OSM API to update the entities after the validation process.

[22] http://oauth.net/

6.5.3 Announcement methods and target participants

Participants are the power of any VGI project. Thus, attracting and encouraging participants to contribute is one of the deployment challenges. The aim is to attract a large number of participants: OSM mappers and public participants as well. We have exploited the power of the crowd to attract participants using the following channels:

- OSM diaries:
  We announced the launch and the objectives of the application locally to the OSM mappers through the project diaries[23]. The OSM diaries are public to every one.
- Social Media:
  We developed two pages for the project: one on Twitter[24] and the other on Facebook[25] to use the power of social media to attract public participants. We infrequently sent news of the application and thanked the participants on the project pages.
- Others:
  Mailing lists and paper-based flyers are also utilized to target other researchers and students as well.

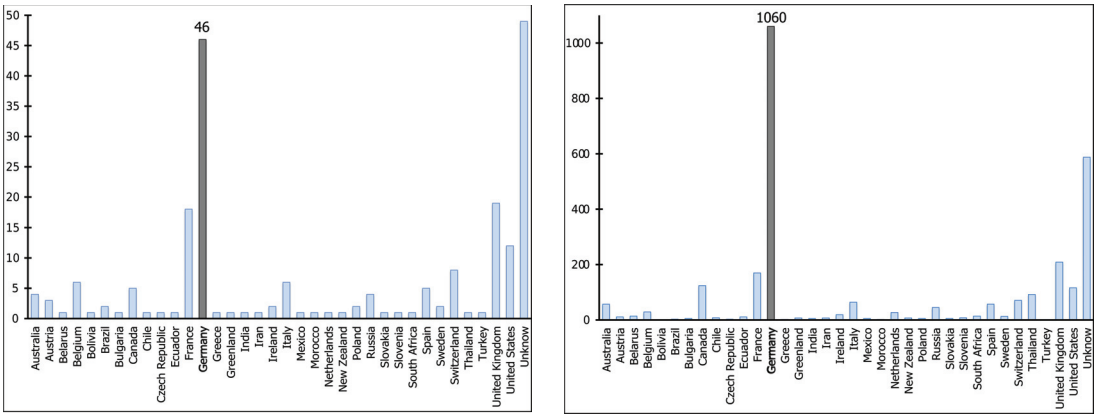
6.6 Results

In this section, we discuss the results that have been obtained by the application from various perspectives: participant and contribution patterns (Section 6.6.1), the participant responses to recommendations (Section 6.6.2), and the potential enhanced data classification (Section 6.6.3). In addition, we analyzed the participant feedback as well (Section 6.6.4). The presented results represent the contributions over a four month period from 28^{th} August to 28^{th} December 2015.

6.6.1 Participant and contribution patterns

Taking into account that we used simple declaration approaches, Figures 6.8, 6.9, and 6.10 give insight into the patterns of participants and contributions. The application attracted 212 participants: 163 participants have a known origin of location from 35 different countries while the others are from unknown locations. Figure 6.8a shows that

[23] https://www.openstreetmap.org/user/grass_and_green/diary
[24] https://twitter.com/grass_and_green
[25] https://www.facebook.com/grassANDgreen/



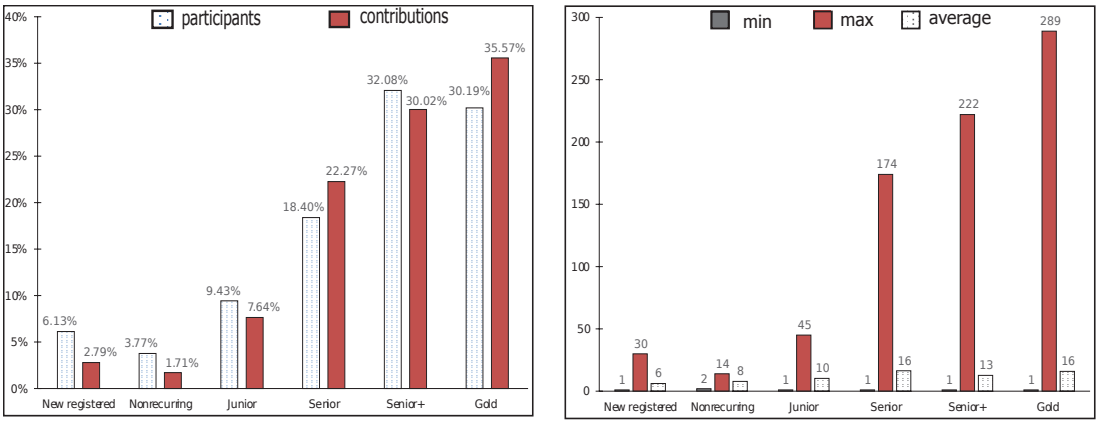
(A) The distribution of participant geo-origins.

(B) Contributions relative to participant geo-origins.

FIGURE 6.8: Participant and contribution patterns with respect to the participant geographic origins.

46 (about 28%) out of 163 participants are from Germany. In addition, the participants examined the classification of 2,865 entities; 1,060 out of these entities have been checked by participants related to Germany, as shown in Figure 6.8b, which is relevant to the data set used here. The rest of the entities have been checked by participants from different locations.

On the other hand, the participants have various levels of familiarity with the OSM project, and consequently, distinct levels of contributions. We use the OSM mapper categorization schema proposed in Neis and Zipf, 2012 to group the participants, as shown in Figure 6.9:



(A) Distribution of participants and contributions per group.

(B) Participant concerns per group.

FIGURE 6.9: Participants and contributions relative to participant experience.

Figure 6.9a shows the distribution of participants and contributions per group as follows: 30.19% *Gold* (changesets[26] >= 2000), 32.08% *Senior*$^+$ (500 <= changesets < 2000), 18.4% *Senior* (100 <= changesets < 500), 9.43% *Junior* (10 <= changesets < 100), 3.77% *Nonrecurring* (1 < changesets < 10), and 6.13% *New registered* (changesets <= 1). In `Grass&Green`, about 65% of contributions are from *Senior*$^+$ and *Gold* mappers, which adds reliability to the obtained results. Figure 6.9b shows the minimum and maximum contributions of participants per group, in addition to the average contributions per participant. This figure indicates that the more experience and familiarity of a participant with the OSM project, the more they are concerned and contribute. Figure 6.9b shows that the participants from *Gold*, *Senior*$^+$, *Senior*, and *Junior* groups examined on average between 11-16 entities/participant, while participants from *Nonrecurring* and *New registered* groups checked on average between 6-8 entities/participant. The finding shows some extreme concerns of individual contributions of 289, 222, and 174 entities from participants belonging to *Gold*, *Senior*, and *Senior*$^+$ groups, respectively.

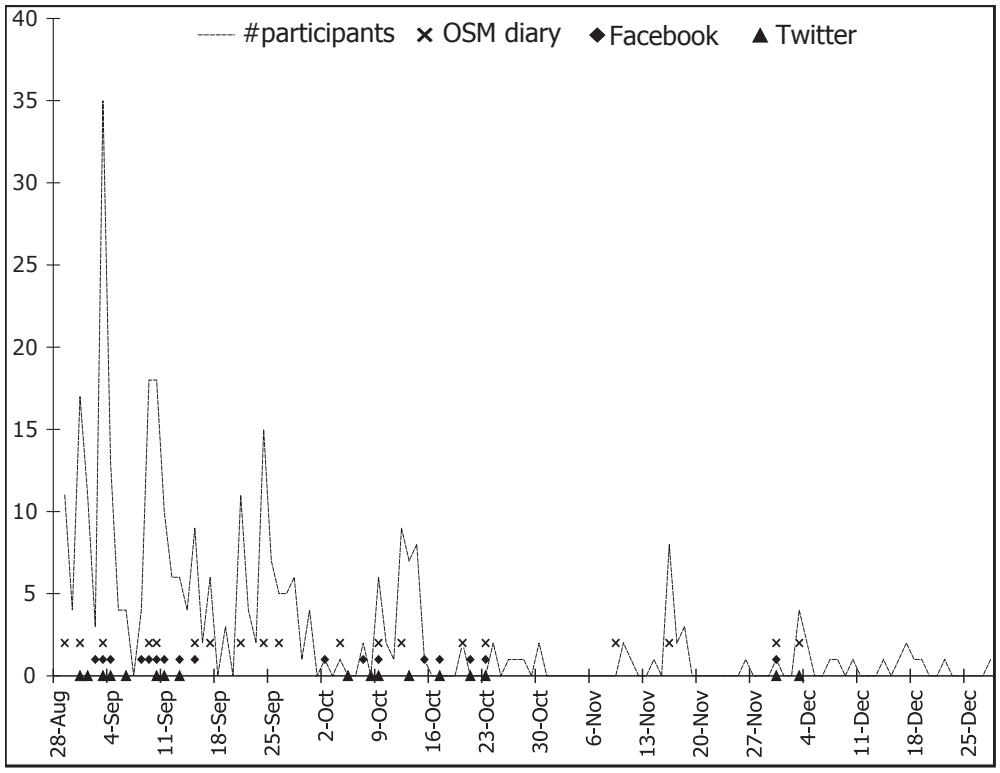


FIGURE 6.10: Numbers of participants per days relative to the announcement methods.

Figure 6.10 shows the contribution patterns relative to the utilized announcement methods. After two weeks, the number of participants are mostly less than ten per day. The figure shows that the number of participants decreases with time and increases with using an attraction method, particularly the OSM diaries.

[26]Changeset: is the number of changes the OSM user done including add, delete, and update operations

6.6.2 Participant responses

The participants checked 2,865 entities. During the validation, the participant may select the "I do not know" option, when they are not confident about a certain classification. For 586 entities we received the "I do not know" option, when the variances between classes were not recognized by the participants. In these cases, the entities have not been updated on the OSM project and have been excluded from our analysis as well. For the rest of the 2,279 entities, we received a participant's opinion. As explained before (see Section 6.5.1), the participant has complete flexibility to adapt our recommended classes resulting in three levels of participant agreement:

- **Complete agreement**: when a participant agrees with both of the recommended classes and marks them with the "yes" option.
- **Partial agreement**: when a participant agrees with only one of the recommended classes and marks the other with a "no" or "maybe" option.
- **Disagreement**: when a participant does not agree with any of the recommended classes and marks them both with a "no" or "maybe" option.

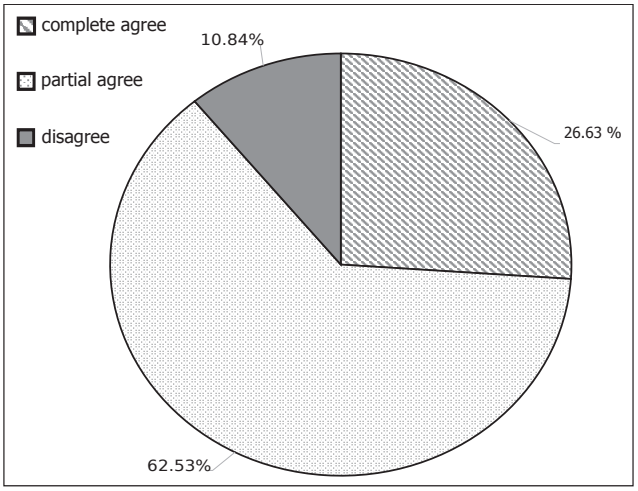


FIGURE 6.11: Participant agreement with the recommended classes.

Figure 6.11 shows the agreement of the participants with the recommended classes as follows: 10.84% disagree, 26.89% completely agree, and 62.53% partially agree. We can conclude that about 89% of the participants have complete/partial agreement with the recommended classes. The findings indicate the success of the developed classifier to distinguish between the target classes. Furthermore, the responses and the participation implies the feasibility of the proposed approach as well.

6.6.3 Enhanced data classification quality

To understand the influence of our approach on data classification quality, we analyzed the contributions in more detail. We examined the classification of entities before and after the validation with respect to the recommended classes. Tables 6.2 and 6.3 give two different views of the results.

entities/class before validation	participants' response	previous class in recommendation	previous class *not* in recommendation	acceptance percentage
412 entities (garden)	yes/maybe	**261**	11	75.9%
	no	88	**52**	
1,136 entities (grass)	yes/maybe	**942**	24	89.2%
	no	98	**72**	
731 entities (park)	yes/maybe	**426**	41	85.2%
	no	67	**197**	
Total 2,279 entities				85.5%

TABLE 6.2: Entities classified before and after the validation with respect to the recommended classes and participant opinions.

Table 6.2 compares the classification of entities before and after the validation with respect to the recommended classes and participant opinions. During the indicated period, participants validated 2,279 entities; these entities were classified previously as follows: 412 *garden*, 1,136 *grass*, and 731 *park*. In the analysis, we investigate whether the previous classification is recommended or not by our approach. From a cognitive view, in this analysis we consider a "maybe" answer to be closer to "yes" than to "no". The findings indicate that the participants accepted 75.9%, 89.2%, and 85.2% of the recommendations of the *garden*, *grass*, and *park* entities, respectively. The participants confirmed the classification of a large portion of the presented entities, as well as correcting other potential misclassified entities (bold numbers in 3rd and 4th columns of Table 6.2). In general, they accepted about 85.5% of the provided recommendations.

classes	in recommended classes	participants response	
		yes/maybe	no
forest	748	**184**	564
garden	753	443	310
grass	1970	1605	365
park	747	542	205
meadow	340	**106**	234

TABLE 6.3: Classes with respect to recommendations and participant responses after the validation.

In another analysis, Table 6.3 gives insight into the classes with respect to the recommendations and participant opinions after the validation process. During the validation

process, the *forest* class was recommended for 748 entities either as 1st or 2nd recommendations. For 184 out of the 748 entities, participants agreed on the potential recommended classes when the *forest* class was not previously assigned to any of the presented entities; the same occurred with the *meadow* class (bold numbers in Table 6.3). Furthermore, entities that have potentially accepted classes of *garden, grass,* and *park* are more than the presented entities per each class as shown in comparison with Table 6.2. On one hand, the finding may indicate the potential correction of misclassified entities. On the other hand, the overall results in Table 6.3 proved the conceptual overlapping classification and demonstrate the plausibility of multiple classes as indicated in Figure 6.12.

Through manual investigation, we detected cases when entities can strongly belong to various classes. According to participant validations, we found numerous entities with two valid classes; among others, 37 entities as *park/forest*, 24 entities as *park/garden*, and 2 as *park/meadow*. Figure 6.12 illustrates some of these examples when the given entity in Figure 6.12a is located within a forest area and adjacent to a farmyard. However, the entity contains a playground (i.e., entertainment facility) and is paved by footways (dashed red lines). Thus, it is recommended and validated to be classified as *park/meadow* while the presented entities in Figures 6.12b and 6.12c are recommended and validated as *park/forest*; they are partially covered by heavy trees and woody plants (dark green areas). In addition, they contain water bodies (outlined by a blue line), and cycle ways (dashed blue lines).

Figure 6.13 illustrates visually the potential of the enhanced data classification. The figure shows three scenarios of contributions: confirmation, correction, and ignorance. Figure 6.13a presents the confirmation scenario, when the indicated entity is classified as *park*. The approach suggests *park* and *grass* as recommended classes. During the

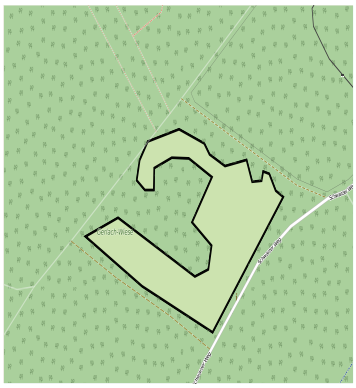
(A) An entity is validated to be classified as *park/meadow*.

(B) An entity is validated to be classified as *park/forest*.

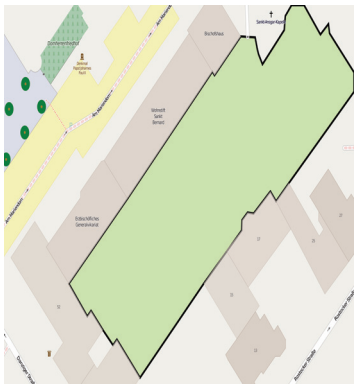
(C) An entity is validated to be classified as *park/forest*.

FIGURE 6.12: Visual illustrations of entities that plausibly belong to conceptual overlapping classes. The given entities (outlined by black lines) are validated by the participants.

(A) A participant followed our recommendation and confirmed the entity classification as *park*.



(B) A participant followed our recommendation and corrected the entity classification from *park* to *meadow*.



(C) A participant ignored our recommended *garden* class, and misclassified the entity as *meadow*.

FIGURE 6.13: Visual investigation of participant contributions compared to the provided recommendations by our approach and the resulting enhanced data classification.

validation, a participant selected only the *park* class. Figure 6.13b shows the correction scenario, when the given entity is classified as *park* and the approach recommends *meadow* and *grass* classes. During the validation, a participant classified it as a *meadow*. Figure 6.13c illustrates the ignorance scenario, when the indicated entity is classified as *grass*. The approach recommends *garden* and *grass* classes. However, a participant decided to classify it as *meadow*, which was an inappropriate choice.

In the first scenario, the given entity has leisure characteristics and the participant followed our recommendations and confirmed its classification as *park*. The entity in the second scenario contains no other features, is located within a forest area, and has a name "Gerlach-Wiese" [27]; it was classified as *park*, but a participant followed our recommendations and updated it to *meadow*. In the third scenario, the entity is surrounded by buildings and has a higher probability of being a *garden*, according to our recommendations. However, the participant classified it as *meadow*, which was an inappropriate class. The last scenario does not enhance the data classification, but it reflects individual perceptions. This scenario could also happen when our recommendations are wrong or do not reflect reality. In such cases, multiple validations could be the proper solution.

6.6.4 Participant feedback

Participants were allowed to contact us giving their comments and feedback either by e-mail or by commenting on our posts. We received both positive and negative feedback as well. Regarding the positive feedback, participants showed respect and encouraged us by different statements like: "great service, plans to expand?", "If you plan to include

[27] wiese (German) = meadow (English)

Belgium, you'll see very strange stuff", "just perfect. thank you", "It's a good subject indeed!", etc. On the contrary some people sent us negative or improvement feedback like: "Your questions will produce a very strong response bias", "referring to Wikipedia and definitions from the dictionaries is completely wrong since OSM does not use natural language to describe objects", "To be able to use this tool correctly, there should be clear consensus on exact meaning", etc. We thank all the participants for their contributions and feedback. The entire feedback will be considered to extend the application.

6.7 Discussion

In the past, mapping was an exclusive task of cartographers and well-trained individuals. Nevertheless, the errors and the accuracy of maps was an issue of concern even in professional production. In reality there is no accurate map due to geographic data ambiguity and temporal developments of data (Crone, 1966; Goodchild and Gopal, 1989; Goodchild, 1993). With the availability of new technologies, VGI has become a potential source of geographic data. In particular, VGI facilitates the mapping process, when the public takes part in the process of data collection. However, in VGI other factors influence the resulting data accuracy such as: the heterogeneous characteristics of the participants, the lack of expertise, and the flexible contribution mechanisms. In particular, most VGI sources have inherent issues such as problematic data classification that is either inconsistent or incomplete. To provide reliable services requires data of guaranteed quality. The concept of Volunteered Geographic Services (VGS) has been introduced in Thatcher (2013). However, there still exists a need for reliable data sources (Parker et al., 2013).

VGI is based on the power of crowdsourcing. From our perspective, in order to exploit the crowd to provide valuable information, participants should be guided and/or well educated regarding the required data quality. Thus, we proposed the rule-guided classification approach in Ali et al. (2015) and Ali et al. (2016). The approach aims to fill the gap between the need for flexible contribution mechanisms, the uncertainty of spatial data, and the various participant perceptions. With the increase in the evolution of VGI sources, machine learning, particularly data mining, can play a vital role in ensuring data quality. In our approach, we applied data mining mechanisms to develop a classifier that can distinguish between similar classes. Afterwards, the developed classifier is utilized to guide the participants towards more accurate classification.

To enhance the data quality, the use of crowdsourcing is one possibility, which has been previously encouraged as one dimension to ensure the data quality (Goodchild and Li, 2012). In this paper, we encourage exploiting the crowds, but in a guided manner. In

crowdsourcing, participants are willing to contribute. However, they generally do not care about the target goal. For example, we tracked the participant interactions during their contribution in `Grass&Green` to find out whether they carefully investigated the provided descriptions or not. We found out that only 80 out of the 212 participants checked the given descriptions in the "Guide" menu. The same situation occurs in the OSM project where most of the participants contribute without spending enough time to read the provided suggestions and recommendations on the OSM Wiki pages.

The application presented in this paper shows the feasibility of the proposed approach. In addition, it encourages the development of customized applications for a particular geographic feature. For example, regarding the OSM project, several applications and services have been developed to check and enhance road networks in various locations. Consequently, OSM provides more reliable and precise information about roads than authoritative data sources in some locations. In `Grass&Green`, we developed a simple application to verify our approach. The few perceived drawbacks could be tackled by intelligent modules. Developing intuitive and interactive interfaces for VGI-based mapping projects would be one possibility to overcome the classification challenges. For example, by negotiation or by exemplification, an intelligent interface might be able to drive the participants towards more precise and finer classification.

From a cognitive perspective, understanding human perception of geographic features is required, because they are the engine of VGI mapping projects. The diversity of participants' cultures and interests have dual functionality: enrich the data source and ensure the data quality. In `Grass&Green`, we coped with participants' diversities by focusing on the concepts and investigating the qualitative representation of the classes. Thus, we utilized classes definitions and descriptions from Wikipedia and dictionaries. Cognitive acquisition techniques and adequate data representation are also required to encourage participants to produce more accurate data. Moreover, the classification problems could be tackled by employing geo-spatial ontology. The need for geo-spatial ontology has been previously discussed for better understanding of space and building more efficient GIS applications (Frank, 1997).

The developed approach is grounded in strong foundations, and thus it can be configured to other geographic features and other locations as well. First, the approach is based on the topological investigation of target features with respect to their context. Therefore, it can be applied to any other areal geographic features (e.g., water body features). Second, the approach is built upon the assumption of localized classification. Thus, within a particular country the approach may be used to enrich the data classification in non-urban areas, after learning from the data of urban areas, if the latter are available. In contrast, the approach has some limitations as well. Firstly, the classifier is dependent on

the availability of *large amounts of data* in order to extract reliable knowledge. Secondly, learning from data with problematic quality may trigger uncertainty in the developed classifier, and hence, a careful investigation of the utilized training data quality is needed.

6.8 Conclusion

VGI can act as a complementary data source for authoritative data and a significant element in a geo-spatial data infrastructure. Nevertheless, heterogeneous data quality limits the utility of this promising resource. In particular, this research tackles the problematic classification of VGI, where the data classification depends on individual preferences and perceptions. In a previous work, we developed the rule-guided classification approach that exploits machine learning mechanisms to handle the classification challenges in VGI projects. The approach utilizes the data availability to learn the distinct characteristics that can help to distinguish between similar classes. The learned characteristics were used afterwards to develop a classifier, which was able to distinguish between similar classes. The classifier is developed to guide the participants towards most appropriate classification.

As a validation of the approach, we developed a web-based application called `Grass&Green`. The application addresses the overlapping classes of some grass-related entities. For a given data set, the application applied the rule-guided classification and presented the recommended classes for public validations. The findings indicate the feasibility of the proposed approach and the success of the application as well. Using simple announcement methods, we attracted the attention of 212 participants from more than 35 different cultural backgrounds. About 89% of the contributions agree with our recommendations. Analysing the contributions shows a potential enhancement of data classification. Participant feedback has encouraged the application of our approach to other data sets. The results stimulate the development of more customized applications to ensure the classification quality of a particular feature. In future works, we intend to design cognitive and interactive data acquisition mechanisms. In addition, we would like to exploit the nature of VGI and the participants in order to develop more intuitive data interpretation.

Acknowledgments:

We gratefully acknowledge the German Academic Exchange Service (DAAD) and the host research group at the Bremen Spatial Cognition Center (BSCC). Moreover, we would like to thank the CapacityLab at the University of Bremen for facilitating a student internship for the second author. Thanks to all the participants for their contributions and feedback to the developed application.

Bibliography

Ali, A. L. and F. Schmid (2014). "Data quality assurance for Volunteered Geographic Information". In: *Geographic Information Science: 8th International Conference, GIScience 2014, Vienna, Austria, September 24-26, 2014. Proceedings*. Ed. by M. Duckham, E. Pebesma, K. Stewart, and A. U. Frank. Cham: Springer International Publishing, pp. 126–141. ISBN: 978-3-319-11593-1.

Ali, A. L., F. Schmid, R. Al-Salman, and T. Kauppinen (2014). "Ambiguity and plausibility: managing classification quality in Volunteered Geographic Information". In: *Proc. of the 22nd ACM SIGSPATIAL International Conf. on Advances in Geographic Information Systems*. Ed. by Y. Huang, M. Schneider, M. Gertz, J. Krumm, and J. Sankaranarayanan. New York, NY, USA: ACM, pp. 143–152. ISBN: 978-1-4503-3131-9.

Ali, A. L., F. Schmid, Z. Falomir, and C. Freksa (2015). "Towards Rule-Guided Classification for Volunteered Geographic Information". In: *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* II-3/W5, pp. 211–217.

Ali, A. L., Z. Falomir, F. Schmid, and C. Freksa (2016). "Rule-guided human classification of Volunteered Geographic Information". In: *ISPRS Journal of Photogrammetry and Remote Sensing*. ISSN: 0924-2716.

Arsanjani, J. J., P. Mooney, A. Zipf, and A. Schauss (2015). "Quality Assessment of the Contributed Land Use Information from OpenStreetMap Versus Authoritative Datasets". In: *OpenStreetMap in GIScience: experiences, research, and applications*. Ed. by J. J. Arsanjani, A. Zipf, P. Mooney, and M. Helbich. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 37–58.

Baglatzi, A., M. Kokla, and M. Kavouras (2012). "Semantifying OpenStreetMap". In: *Proceedings of the 5th International Terra Cognita Workshop*. Citeseer, pp. 39–50.

Ballatore, A. and A. Zipf (2015). "A Conceptual Quality Framework for Volunteered Geographic Information". In: *Spatial Information Theory: 12th International Conference, COSIT 2015, Santa Fe, NM, USA, October 12-16, 2015, Proceedings*. Ed. by S. Fabrikant, M. Raubal, C. Bertolotto M.and Davies, S. Freundschuh, and S. Bell. Santa Fe, NM, USA, pp. 89–107.

Ballatore, A., M. Bertolotto, and D. C. Wilson (2013). "Geographic knowledge extraction and semantic similarity in OpenStreetMap". In: *Knowledge and Information Systems* 37.1, pp. 61–81.

Barron, C., P. Neis, and A. Zipf (2014). "A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis". In: *Transactions in GIS* 18.6, pp. 877–895.

Bishr, M. and W. Kuhn (2007). "Geospatial Information Bottom-Up: A Matter of Trust and Semantics". In: *The European Information Society: Leading the Way with Geo-information*. Ed. by S. I. Fabrikant and M. Wachowicz. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 365–387. ISBN: 978-3-540-72385-1.

Comber, A. J., P. Fisher, F. Harvey, M. Gahegan, and R. Wadsworth (2006). "Using metadata to link uncertainty and data quality assessments". In: *Proc. of the 12th Inter. Symposium on Spatial Data Handling*. Springer, pp. 279–292.

Crone, G. R. (1966). *Maps and their makers: an introduction to the history of cartography*. Hutchinson.

D'Antonio, F., P. Fogliaroni, and T. Kauppinen (2014). "VGI Edit History Reveals Data Trustworthiness and User Reputation". In: *Connecting a Digital Europe Through Location and Place*. Ed. by J. Huerta, S. Schade, and C. Granell. Springer-Verlag.

Devillers, R., A. Stein, Y. Bédard, N. Chrisman, P. Fisher, and W. Shi (2010). "Thirty years of research on spatial data quality: achievements, failures, and opportunities". In: *Transactions in GIS* 14.4, pp. 387–400.

Dorn, H., T. Törnros, and A. Zipf (2015). "Quality Evaluation of VGI Using Authoritative Data–A Comparison with Land Use Data in Southern Germany". In: *ISPRS International Journal of Geo-Information* 4.3, pp. 1657–1671.

Egenhofer, M. J. and K. K. Al-Taha (1992). "Reasoning about Gradual Changes of Topological Relationships". In: *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space: Proc. of the International Conference GIS*. Ed. by A. U. Frank, I. Campari, and U. Formentini. Berlin, Heidelberg: Springer, pp. 196–219.

Elwood, S., M. F. Goodchild, and D. Z. Sui (2012). "Researching Volunteered Geographic Information: Spatial data, geographic research, and new social practice". In: *Annals of the Association of American Geographers* 102.3, pp. 571–590.

Fisher, P. F. (1999). "Models of uncertainty in spatial data". In: *Geographical information systems* 1, pp. 191–205.

Flanagin, A. J. and M. J. Metzger (2008). "The credibility of Volunteered Geographic Information". In: *GeoJournal* 72.3, pp. 137–148.

Foody, G. M., L. See, S. Fritz, M. van der Velde, C. Perger, C. Schill, D. S. Boyd, and A. Comber (2015). "Accurate Attribute Mapping from Volunteered Geographic Information: Issues of Volunteer Quantity and Quality". In: *The Cartographic Journal* 52.4, pp. 336–344.

Foth, M., B. Bajracharya, R. Brown, and G. Hearn (2009). "The Second Life of urban planning? Using NeoGeography tools for community engagement". In: *Journal of Location Based Services* 3.2, pp. 97–117.

Frank, A. U. (1997). "Spatial ontology: a geographical information point of view". In: *Spatial and temporal reasoning*. Springer, pp. 135–153.

Fritz, S., I. McCallum, C. Schill, C. Perger, L. See, D. Schepaschenko, M. Van der Velde, F. Kraxner, and M. Obersteiner (2012). "Geo-Wiki: An online platform for improving global land cover". In: *Environmental Modelling & Software* 31, pp. 110–123.

Girres, J.-F. and G. Touya (2010). "Quality assessment of the French OpenStreetMap dataset". In: *Transactions in GIS* 14.4, pp. 435–459.

Goodchild, M. F. (1993). "Data models and data quality: problems and prospects". In: *Environmental modeling with GIS*. Ed. by M. F. Goodchild, B. O. Parks, and L. T. Steyaert. Oxford, UK: Oxford University Press, pp. 94–103.

Goodchild, M. F. (2007). "Citizens as sensors: the world of volunteered geography". In: *GeoJournal* 69.4, pp. 211–221.

Goodchild, M. F. (2008). "Assertion and authority: the science of user-generated geographic content". In: *Proc. of the Colloquium for Andrew U. Frank's 60th Birthday, Department of Geoinformation and Cartography*. Citeseer.

Goodchild, M. F. and S. Gopal (1989). *The accuracy of spatial databases*. CRC Press.

Goodchild, M. F. and L. Li (2012). "Assuring the quality of Volunteered Geographic Information". In: *Spatial statistics* 1, pp. 110–120.

Gouveia, C. and A. Fonseca (2008). "New approaches to environmental monitoring: the use of ICT to explore Volunteered Geographic Information". In: *GeoJournal* 72.3-4, pp. 185–197.

Grira, J., Y. Bédard, and S Roche (2010). "Spatial data uncertainty in the VGI world: Going from consumer to producer". In: *Geomatica* 64.1, pp. 61–72.

Guptill, S. C. and J. L. Morrison, eds. (2013). *Elements of Spatial Data Quality*. Oxford, UK: Elsevier Science.

Haklay, M. (2010). "How good is Volunteered Geographic Information? A comparative study of OpenStreetMap and Ordnance Survey datasets". In: *Environment and planning. B Planning & design* 37.4, p. 682.

Haklay, M. and P. Weber (2008). "OpenStreetMap: user-generated street maps". In: *IEEE Pervasive Computing* 7.4, pp. 12–18. ISSN: 1536-1268.

Hecht, B. and M. Stephens (2014). "A Tale of Cities: Urban Biases in Volunteered Geographic Information". In: *Proceedings of the 8th International Conference on Weblogs and Social Media (ICWSM)*. Michigan, USA.

Karagiannakis, N., G. Giannopoulos, D. Skoutas, and S. Athanasiou (2015). "OSMRec Tool for Automatic Recommendation of Categories on Spatial Entities in OpenStreetMap". In: *Proc. of the 9th ACM Conf. on Recommender Systems*. New York, NY, USA: ACM, pp. 337–338.

Keßler, C. and R. T. A. de Groot (2013). "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*. Ed. by D. Vandenbroucke, B. Bucher, and J. Crompvoets. Springer-Verlag, pp. 21–37.

Keßler, C., J. Trame, and T. Kauppinen (2011). "Tracking editing processes in Volunteered Geographic Information: the case of OpenStreetMap". In: *Proceedings of Workshop on Identifying objects, processes and events in spatio-temporally distributed data (IOPE), Conference on Spatial Information Theory (COSIT 2011)*. Vol. 12.

Klippel, A., K. Sparks, and J. O. Wallgrün (2015). "PITFALLS AND POTENTIALS OF CROWD SCIENCE: A META-ANALYSIS OF CONTEXTUAL INFLUENCES". In: *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* II-3/W5, pp. 325–331.

Ludwig, I., A. Voss, and M. Krause-Traudes (2011). "A Comparison of the Street Networks of Navteq and OSM in Germany". In: *Advancing Geoinformation Science for a Changing World*. Ed. by W. Geertman Stanand Reinhardt and F. Toppen. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 65–84. ISBN: 978-3-642-19789-5.

Mooney, P. and P. Corcoran (2011). "Can Volunteered Geographic Information Be a Participant in eEnvironment and SDI?" In: *Environmental Software Systems. Frameworks of eEnvironment*. Springer, pp. 115–122.

Mooney, P. and P. Corcoran (2012b). "The annotation process in OpenStreetMap". In: *Transactions in GIS* 16.4, pp. 561–579.

Mooney, P., H. Sun, and L. Yan (2011). "VGI as a Dynamically Updated Data Source in Location-Based Services in Urban Environments". In: *Proceedings of the 2nd International Workshop in Ubiquitous Crowdsourcing (UbiCrowd'11)*. Beijing, China.

Neis, P. and A. Zipf (2012). "Analyzing the contributor activity of a Volunteered Geographic Information project: the case of OpenStreetMap". In: *ISPRS International Journal of Geo-Information* 1.2, pp. 146–165.

Neis, P., D. Zielstra, and A. Zipf (2011). "The street network evolution of crowdsourced maps: OpenStreetMap in Germany 2007–2011". In: *Future Internet* 4.1, pp. 1–21.

Neis, P., D. Zielstra, and A. Zipf (2013). "Comparison of Volunteered Geographic Information Data Contributions and Community Development for Selected World Regions". In: *Future Internet* 5.2, pp. 282–300.

Østensen, O. M. and P. C. Smits (2002). "ISO/TC211: Standardisation of geographic information and geo-informatics". In: *Geoscience and Remote Sensing Symposium, 2002. IGARSS'02. 2002 IEEE International*. Vol. 1. IEEE, pp. 261–263.

Parker, C. J., A. May, V. Mitchell, and A. Burrows (2013). "Capturing volunteered information for inclusive service design: potential benefits and challenges". In: *The Design Journal* 16.2, pp. 197–218.

Pourabdollah, A., J. Morley, S. Feldman, and M. Jackson (2013). "Towards an authoritative OpenStreetMap: conflating OSM and OS OpenData national maps' road network". In: *ISPRS International Journal of Geo-Information* 2.3, pp. 704–728.

Quattrone, G., A. Mashhadi, and L. Capra (2014). "Mind the map: the impact of culture and economic affluence on crowd-mapping behaviours". In: *Proceedings of the 17th*

ACM Conference on Computer supported cooperative work & social computing. ACM, pp. 934–944.

Roche, S., E. Propeck-Zimmermann, and B. Mericskay (2013). "GeoWeb and crisis management: Issues and perspectives of Volunteered Geographic Information". In: *GeoJournal* 78.1, pp. 21–40.

Savelyev, A., S. Xu, K. Janowicz, C. Mülligann, J. Thatcher, and W. Luo (2011). "Volunteered geographic services: developing a linked data driven location-based service". In: *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Spatial Semantics and Ontologies*. ACM, pp. 25–31.

Schmid, F., O. Kutz, L. Frommberger, T. Kauppinen, and C. Cai (2012). "Intuitive and natural interfaces for geospatial data classification". In: *Workshop on place-related knowledge acquisition research (P-KAR), Kloster Seeon, Germany*, p. 26.

Schmid, F., L. Frommberger, C. Cai, and F. Dylla (2013a). "Lowering the barrier: How the What-You-See-Is-What-You-Map paradigm enables people to contribute Volunteered Geographic Information". In: *Proc. of the 4th Annual Symposium on Computing for Development*. ACM. Cape Town, South Africa, pp. 8–18.

Sparks, K., A. Klippel, J. O. Wallgrün, and D. Mark (2015). "Citizen Science Land Cover Classification Based on Ground and Aerial Imagery". In: *Spatial Information Theory: 12th International Conference, COSIT 2015, Santa Fe, NM, USA, October 12-16, 2015, Proceedings*. Ed. by I. S. Fabrikant, M. Raubal, M. Bertolotto, C. Davies, S. Freundschuh, and S. Bell. Cham: Springer International Publishing, pp. 289–305. ISBN: 978-3-319-23374-1.

Thabtah, F. (2007). "A review of associative classification mining". In: *The Knowledge Engineering Review* 22.01, pp. 37–65.

Thatcher, J. (2013). "From Volunteered Geographic Information to Volunteered Geographic Services". In: *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*. Ed. by D. Sui, S. Elwood, and M. F. Goodchild. Dordrecht: Springer Netherlands, pp. 161–173. ISBN: 978-94-007-4587-2.

Vandecasteele, A. and R. Devillers (2013). "Improving Volunteered Geographic Data Quality Using Semantic Similarity Measurements". In: *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 1.1, pp. 143–148.

Vaz, E. and J. Jokar Arsanjani (2015). "Crowdsourced mapping of land use in urban dense environments: An assessment of Toronto". In: *The Canadian Geographer/Le Géographe canadien* 59.2, pp. 246–255.

Yanenko, O. and C. Schlieder (2014). "Game principles for enhancing the quality of user-generated data collections". In: *Workshop of Geogames and Geoplay in the 17th AGILE Conference on Geographic Information Science,* pp. 1–5.

Zook, M., M. Graham, T. Shelton, and S. Gorman (2010). "Volunteered Geographic Information and crowdsourcing disaster relief: a case study of the Haitian earthquake". In: *World Medical & Health Policy* 2.2, pp. 7–33. ISSN: 1948-4682.

Chapter 7

Conclusions and Future Work

The presented research reveals various issues regarding the quality of VGI. In particular, this dissertation focuses on the quality of map-based VGI and tackles the challenges of human-centered data classification. During this research, we studied the potential causes behind problematic data classification in map-based VGI. Furthermore, we tackled the problem by developing a guiding approach to cope with the evolution of VGI. To conclude, this chapter summarizes the answers to the presented research questions within the scope of the contributions of this research (see Chapter 1). Furthermore, it highlights envisions of future research directions.

7.1 Discussions and Conclusions

GIS is a particular kind of information system, which facilitates manipulation and processing of data related to properties on or near the Earth's surface. In GIS, the data links a specific property and an associated geo-location. Geographic data has various formats, either records associated with locations (e.g., census and socioeconomic data) or records describe recognizable geographic features (e.g., natural features, road networks, buildings). Over the past years, GIS had a reputation of being difficult to use and geographic data collection and processing were exclusively reserved for professionals and well-trained individuals. However, the utility of ICT fosters significant changes not only in the usability and accessibility of GIS, but in the way data is collected and processed. Nowadays, everyone with access to the Internet is able to: i) participate in mapping and collecting geographic data, ii) use open-source GIS software, iii) perform geographic data processing, and iv) access, utilize, and reproduce various formats of open-geographic content. Therefore, the dilemma in GIScience research has changed from *How to collect*

and produce geographic data? to *How to guarantee the quality and effectively utilize the resulting data?*

Concerns of data quality rise to the highest priority, when millions of volunteers around the world collaboratively act to collect, update, and use information about geographic features. Different quality assurance procedures are presented in Chapter 2 (see **C1**). According to the literature, the data quality is ensured probably by following either extrinsic or intrinsic approaches. In the extrinsic approaches, the data is matched and compared with a reference data set, while in the intrinsic approaches the data is analyzed to find out an indirect signal of data quality. With the limited availability of reference data sets, the intrinsic approaches assess the contributors' reputation, analyze the contribution pattern, and check for credibility and trustworthiness to ensure data quality. With the availability of large amounts of data, data mining techniques arise as a promising method to ensure the data quality.

In this dissertation, data quality is addressed from the perspective of classification. We examined the problematic classification of various areal geographic features in Chapters 3-4. The findings show that a limited number of geographic features follows a strict structure of data classification (e.g., hierarchical classification). For these features the data can be checked against the constraints to ensure the integrity. However, most geographic features follow context-based classification; when the classification of a given feature is related to its characteristics and its geographic context. In this research several concepts have been introduced:

- *Classification Ambiguity*: entities could belong to several classes due to conceptual overlap of the classes.
- *Classification Plausibility*: entities have a high compatibility with a specific class rather than other possible classes.
- *Appropriate Classification*: the class that strongly reflects the intrinsic and extrinsic characteristics of a given entity and is consistent with its geographic context. Furthermore, it has the highest compatibility among other similar classes as well.
- *Wrong/Inappropriate Classification*: the class that describes an entity inappropriately. This class might be in conflict with entity characteristics and its geographic context.

In the context of VGI, the term *Wrong Classification* is not recommended unless the classification is completely incorrect, such as classifying a "residential area" as a "forest". In VGI, the classification is based on human perception, and hence, the terms of *Appropriate Classification* and *Inappropriate Classification* are more adequate.

Mapping and collecting information about the land use was of great concern since the earliest GIS. By the middle of the 1960s, the first operational GIS had been developed in Canada to collect and store information about the land use (Foresman, 1998). Land use (LU) and land cover (LC) data are complementary: LC indicates the physical type of land in a particular area, while LU determines the appropriate utilization of a particular area by humans (Fisher et al., 2005). From a cognitive perspective, humans need categorical data to build memories, process experience, and communicate knowledge (Rosch, 1978). However, categorical classification of LU/LC represents a challenge due to the following reasons: i) classes might lead to binary treatments and loss of information, ii) there are no standard measures to distinguish various classes, iii) a particular land might be used differently by humans, and iv) due to the non-strict definition of most geographic features, there exists conceptual overlap between similar classes (Ahlqvist and Ban, 2007; Ahlqvist, 2012). The challenges are doubled in VGI, as the data classification is based on rational perspectives with no integrity checking mechanisms. Moreover, participants are mostly volunteers; they are not well-trained and they might be not interested in geography or cartography at all. Therefore, LU/LC data resulting from VGI projects comes with an inherently problematic classification and requires careful investigation before utilization (see **C2**).

Hence, this dissertation proposes a guided classification approach in Chapters 4-5 to exploit the leverage of VGI to produce data of enhanced quality. The approach employs the availability of data to learn the characteristics of particular geographic features. Various characteristics are used to distinguish between classes: quantitative and qualitative characteristics. In Chapter 5, a rule-based guided classification approach is presented. In this approach, qualitative spatial reasoning (QSR) is adopted to extract the distinct qualitative characteristics of particular geographic features. The extracted characteristics are formulated as associative prediction rules and are utilized to develop a classifier. Afterwards, the developed classifier acts to generate recommendations and guides the participants to the most proper classes for a given entity. The findings indicate the capabilities of the developed classifier to distinguish between similar classes. Findings of empirical studies show the agreement of the participants on the recommended classes (see **C3-C4**).

The rule-based guided classification approach is practically implemented in Chapter 6. As exemplification, the approach is applied to some grass-related features. These kinds of features present challenges for classification. Although they share the general vegetation characteristics, there exist fine details that may identify each individual geographic feature. Otherwise, quantitative measures and qualitative observations are usually able to distinguish an appropriate and inappropriate classification. The entities “park”, “garden”, “forest”, “meadow”, and “grass” are extracted from the OSM data set of Germany.

Based on the proposed approach, we developed a classifier that is able to distinguish these features. Thereafter, the web application `Grass&Green` was developed to present the entities associated with our recommended classes for crowdsourcing validation. The validation process showed three major findings: 1) the feasibility to learn the qualitative characteristic of a specific geographic feature from VGI, 2) the significant role of crowdsourcing in enhancing the data classification as well as in data collection, and 3) the potential enhancement of data classification when volunteers are supported by guidance (see **C5-C6**). Users of the application strongly agreed on a large fraction of the presented recommendations. They provide feedback to improve and extend the developed application as well.

In general, there is no absolute accurate geographic map (Goodchild, 1993); as any map is likely a model or a generalization of reality, it might contain a certain level of inaccuracy. In geographic maps, thematic inaccuracy might be due to either measurement problems, or problems related to definition and classification granularity. In VGI-based mapping particularly, thematic accuracy is problematic due to contributors' rational preferences and their limited knowledge and experience. Therefore, this dissertation proposes a human-centered guided classification approach to tackle thematic inaccuracy of the resulting data.

7.2 Future Directions

The potential of VGI in mapping activities will increase with the expansion of geoinformation technologies. Nevertheless, merit of the resulting data might be limited as long as there are no adequate procedures to ensure the data quality. Thus, the next sections highlight some research directions related to VGI quality assessment and enhancement approaches (Sections 7.2.1 and 7.2.2). Otherwise, toward effective utilization of data, further research is required to develop intuitive data interpretation methodologies that are able to handle the questionable data (Section 7.2.3). Furthermore, the extension of the `Grass&Green` application is discussed in Section 7.2.4, as a part of my future work.

7.2.1 Data quality: an assessment approach

Regarding data quality assessment, several research projects have started to look for characterizing VGI quality by finding a way to characterize data providers (Maué, 2007; Flanagin and Metzger, 2008; Foody et al., 2015). They adopted the *social* approach that has been developed by Goodchild and Li (2012).

To assess data providers, one idea is to implement a reputation system that evaluates the interaction between community members. Reputation systems are well known in other applications such as e-commerce, where service providers, services, and goods are associated with scores indicating their quality (Resnick et al., 2000; Jøsang et al., 2007). In e-commerce applications, these scores are calculated conventionally based on collected feedback and opinions provided by both sides of a commercial transaction. In VGI mapping projects, the challenge is that there is no direct relation or feedback between the contributors; they only share a platform to contribute geographic data. However, they share editing the same entities in a collaborative manner. Hence, tracking the editing history for a specific set of entities can be exploited to construct the interaction network among contributors, and consequently, to calculate their reputation scores. In previous work of Keßler et al. (2011) and Keßler and Groot (2013), the authors categorized the editing actions into positive and negative feedback. This work can be extended to develop a reputation system for VGI. The system will act to rank the contributors, and hence, to assess resulting data quality.

7.2.2 Data quality: an enhancement approach

This dissertation argues the usefulness of a guidance approach to enhance data classification quality in VGI-based mapping projects. There are various ways of guiding that might be adequate to such projects. For example, guiding by asking a series of questions, guiding by illustrative examples, or guiding by comparison. In this dissertation, we applied the classical type of guiding, where contributors are informed about recommended classes among potential alternatives. From a cognitive perspective, comparison or exemplification might be adequate ways to get precise information from contributors. When guiding is provided in an adequate manner, it will have dual functionality: first, it will probably result in enhanced data quality; second, it acts to raise the contributors' experience by learning. Further studies might be required to find the proper guiding methodology, which should not hinder the contributors to express their observations.

In GIScience, the "Gamification" approach has evolved originally as a means of data collection (e.g., Towns Conquer (Castellote et al., 2013)). However, research started to employ it as a tool for enhancing the data quality (Yanenko and Schlieder, 2014) (e.g., Cropland Capture (See et al., 2014)). In this research, it has been shown that a good motivation methodology is required to keep contributors active in validating the presented recommendation (see Figure 6.10). Further research is required to develop a game-based guiding system for VGI mapping. Here, the challenge is how to preserve the properties of VGI mapping (e.g., flexibility), games (e.g., motivation), and guiding systems (e.g., accuracy).

7.2.3 Intuitive data interpretation

The next generation of GIS must address the quality of VGI. Due to the importance of VGI as a source of information, intelligent data interpretation should be developed to handle the uncertainty of data. In cases of conceptual overlap between classes, binary classification leads to loss of information. However, the overlap can be interpreted in such a way that is result in more precise information.

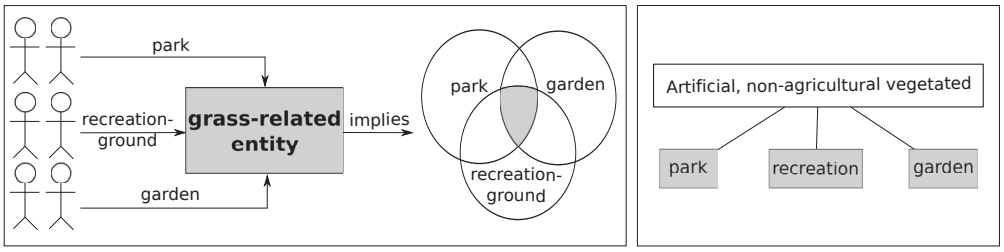


FIGURE 7.1: Intelligent data interpretation of overlapping classes.

For example, Figure 7.1 depicts a case when an entity could be classified differently by contributors: "park", "garden", and "recreation-ground". Every classification might be based on an individual observation. Therefore, the entity might belong to all of these classes with near levels of appropriateness. In this case, these overlapping classes might imply that this entity is "park–like a garden with a recreation facilities". From another view, the overlapping classes emphasize the broad level of classification (i.e. artificial, non-agricultural vegetated, not agricultural area, not forest, and not water body). The challenge here is the difficulty of data interpretation, when the assigned classes are not overlapping. Further research is needed to develop such intuitive data interpretation approaches.

7.2.4 Extension of `Grass&Green`

The users of the `Grass&Green` application encouraged extending of the application. The extension can be achieved by applying the approach either on different locations or on different geographic features. The OSM project requires particular research that focuses on improving the data classification. The improvement denotes checking for the classification integrity (i.e., horizontal view) and increasing the level of classification granularity (i.e., vertical view). In some applications coarse classification suffices while in others finer levels of detail are required. For example, during disaster situations, planning for evacuation requires knowing the type of building. Thus, developing customized applications to enhance the data classification in OSM will increase the utility of the resulting data. Further research is needed to keep contributors motivated towards guaranteed data quality.

Bibliography

Ahlqvist, O. (2012). "Semantic Issues in Land-Cover Analysis: representation, analysis, and visualization". In: *Remote Sensing of Land Use and Land Cover: principles and applications*. Ed. by C. P. Giri. CRC Press, pp. 25–35.

Ahlqvist, O. and H. Ban (2007). "Categorical Measurement Semantics: a new second space for geography". In: *Geography Compass* 1.3, pp. 536–555. ISSN: 1749-8198.

Castellote, J., J. Huerta, J. Pescador, and M. Brown (2013). "Towns Conquer: Developing a linked data Gamified application to collect geographical names (vernacular names/toponyms)". In: *Geographic Information Science at the Heart of Europe*. Ed. by D. Vandenbroucke, B. Bucher, and J. Crompvoets. Springer-Verlag.

Fisher, P., A. J. Comber, and R. Wadsworth (2005). "Land use and Land cover: contradiction or complement". In: *Re-presenting GIS*. Ed. by P. Fisher and D. J. Unwin. The Atrium, Southern Gate, UK: John Wiley & Sons, pp. 85–98.

Flanagin, A. J. and M. J. Metzger (2008). "The credibility of Volunteered Geographic Information". In: *GeoJournal* 72.3, pp. 137–148.

Foody, G. M., L. See, S. Fritz, M. van der Velde, C. Perger, C. Schill, D. S. Boyd, and A. Comber (2015). "Accurate Attribute Mapping from Volunteered Geographic Information: Issues of Volunteer Quantity and Quality". In: *The Cartographic Journal* 52.4, pp. 336–344.

Foresman, T. W. (1998). *The history of geographic information systems: perspectives from the pioneers*. Prentice Hall.

Goodchild, M. F. (1993). "Data models and data quality: problems and prospects". In: *Environmental modeling with GIS*. Ed. by M. F. Goodchild, B. O. Parks, and L. T. Steyaert. Oxford, UK: Oxford University Press, pp. 94–103.

Goodchild, M. F. and L. Li (2012). "Assuring the quality of Volunteered Geographic Information". In: *Spatial statistics* 1, pp. 110–120.

Jøsang, A., R. Ismail, and C. Boyd (2007). "A survey of trust and reputation systems for online service provision". In: *Decision support systems* 43.2, pp. 618–644.

Keßler, C. and R. T. A. de Groot (2013). "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*. Ed. by D. Vandenbroucke, B. Bucher, and J. Crompvoets. Springer-Verlag, pp. 21–37.

Keßler, C., J. Trame, and T. Kauppinen (2011). "Tracking editing processes in Volunteered Geographic Information: the case of OpenStreetMap". In: *Proceedings of Workshop on Identifying objects, processes and events in spatio-temporally distributed data (IOPE), Conference on Spatial Information Theory (COSIT 2011)*. Vol. 12.

Maué, P. (2007). "Reputation as tool to ensure validity of VGI". In: *Proceedings of Workshop on Volunteered Geographic Information*. University of California, Santa Barbara.

Resnick, P., K. Kuwabara, R. Zeckhauser, and E. Friedman (2000). "Reputation systems". In: *Communications of the ACM* 43.12, pp. 45–48. ISSN: 0001-0782.

Rosch, E. (1978). "Principles of categorization". In: *Cognition and categorization*. Ed. by E. Rosch and B. B. Lloyd. Hillsdale, NJ, USA: Lawrence Erlbaum, pp. 27–48.

See, L., T. Sturn, S. Fritz, I. McCallum, and C. Salk (2014). "Cropland capture: a gaming approach to improve global land cover". In: *Connecting a Digital Europe Through Location and Place*. Ed. by J. Huerta, S. Schade, and C. Granell. Springer-Verlag.

Yanenko, O. and C. Schlieder (2014). "Game principles for enhancing the quality of user-generated data collections". In: *Workshop of Geogames and Geoplay in the 17th AGILE Conference on Geographic Information Science,* pp. 1–5.

Appendix

Appendix A

OpenStreetMap *landuse* related tags

This appendix lists most common OSM tags, which are related to land use and land cover mapping. The tags represent various level of classification granularity: the broarder level (e.g., "grass", "water") and the finer level (e.g., "garden", "lake"). These tags are developed based on discussions among mapper communities. They are described in more detail on the OSM Wiki pages.

Key	Value	Comments and Remarks
landuse	allotments	A piece of land given over to local residents for growing vegetables and flowers.
landuse	basin	An area of water body that drains into a river.
landuse	brownfield	Describes land scheduled for new development where old buildings have been demolished and cleared
landuse	cemetery	Place for burials. You can add *religion* = *value*. Smaller places (e.g. with a church nearby) may use *amenity* = *grave_yard*.
landuse	commercial	Predominantly offices, business parks, etc.
landuse	farmland	An area of farmland used for tillage and pasture (animals, crops, vegetables, flowers, fruit growing).
landuse	farmyard	An area of land with farm buildings like farmhouse, dwellings, farmsteads, sheds, stables, barns, equipment sheds, feed bunkers, etc. plus the open space in between them and the shrubbery/trees around them.
landuse	forest	Managed forest or woodland plantation.
landuse	grass	For areas covered with grass. Consider finder tags when more information are available.
landuse	industrial	Predominantly workshops, factories or warehouses.

landuse	meadow	An area of land primarily vegetated by grass and other non-woody plants, usually mowed for making hay.
landuse	railway	Area for railway use, generally off-limits to the general public.
landuse	recreation_ground	An open green space for general recreation, which may include pitches, nets and so on, usually municipal but possibly also private to colleges or companies.
landuse	reservoir	Stores water, may be covered or uncovered
landuse	residential	Predominantly houses or apartment buildings
landuse	retail	Predominantly shops
landuse	village_green	An area of common land, usually grass, in the centre of a village
leisure	garden	Place where flowers and other plants are grown in a decorative and structured manner or for scientific purposes.
leisure	golf_course	The outline of a golf course. The node form may be used to place an icon within the course. This tag implies sport=golf.
leisure	marina	For mooring leisure yachts and motor boats.
leisure	nature_reserve	Protected area of importance for wildlife, flora, fauna or features of geological or other special interest.
leisure	park	Open, green area for recreation, usually municipal.
leisure	pitch	E.g. a field for playing football/soccer, cricket, baseball sports, and skate parks. To describe which kinds of sport(s) use sport=*.
natural	wood	Woodland where timber production does not dominate use.
natural	scrub	Uncultivated land covered with bushes or stunted trees.
natural	heath	Bare lower lying uncultivated land with bushes but little or no tree cover.
natural	sand	Ground coverage of mostly silica particles, with no or very sparse vegetation.
natural	water	General tags for all kinds of water: Lakes, pond, etc.
natural	wetland	Waterlogged area.

TABLE A.1: List of OSM tags related to land use and land cover mapping.

Appendix B

Other Publications

During this research I have participated as (co)author in the following publications:

- Ahmed Loai Ali (2016). "Tackling the thematic accuracy of areal features in OpenStreetMap". In: *European Handbook of Crowdsourced Geographic Information*, Ed. by Capineri, C. , Haklay, M., Huang, H., Antoniou, V., Kettunen, J., Ostermann, F. and Purves, R., pp. 113–129. London: Ubiquity Press. [In press].
- Ahmed Loai Ali, Falko Schmid, Zoe Falomir, and Christian Freksa (2015). "Towards Rule-Guided Classification for Volunteered Geographic Information". In: *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences II-3/W5*, pp. 211–217.
- Amin Mobasheri, Alexander Zipf, Lukas Loos, and Ahmed Loai Ali. "Open Geospatial Data Quality Assessment; A Case Study of OpenStreetMap Data Completeness for Specialized Routing Services". In: *ISPRS International Journal of Geo-Information*. [Under review].

Declaration of Authorship

I, Ahmed Loai Ali, declare that this thesis titled, 'Enhancing Data Classification Quality of Volunteered Geographic Information' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____