This is a pre-print of the Accepted Manuscript (AM).
This article has been accepted for publication in International Journal of Geographical Information Science, published by Taylor & Francis.
Refer to the publisher's Version of Record (VOR) for the authoritative version - Rajesh Chittor Sundaram Et al. (2020), Can you fixme? An intrinsic classification of contributor-identified spatial data issues using topic models, International Journal of Geographical Information Science,
DOI: 10.1080/13658816.2021.1893323

Can you fixme? An intrinsic classification of contributor-identified spatial data issues using topic models

Rajesh Chittor Sundaram[a], Elham Naghizade[a], Renata Borovica-Gajic[a] and Martin Tomko[a]

[a]Faculty of Engineering and Information Technology (Department of Infrastructure Engineering), The University of Melbourne, Parkville, VIC 3010, Australia;

ARTICLE HISTORY
Compiled February 9, 2021

ABSTRACT
Assessing OpenStreetMap (OSM) data quality against authoritative data sources may not always be viable. This is primarily because of the multi-dimensional nature and heterogeneity of the maps, yet the activity is pivotal for targeted data cleansing and quality enhancement undertakings in these data sets. A salient facet of OSM, allowing contributors to flag *potential* problems encountered during the mapping process, is the `FIXME` tag. In this article, we examine and discuss OSM data quality through the vast expanse of issues (knowledge) documented via `FIXME`. We present a classification and analysis of these quality issues, exposed as topic models and grounded in the ISO-19157 standard, across USA and Australia. Regional distributions of these topics are further qualitatively analyzed to ascertain the variation of key issues in OSM. We also present a comparison of the intrinsic issue classification against those identified in an issue corpus of an authoritative map data source. Due to the considerable heterogeneity in user mapping and reporting, OSM issue detection and classification remains problematic. This research presents a flexible and intrinsic data-mining approach, linking established ISO data quality standards to OSM issue categorization. Our work, thus informs the development of automated error correction methods for VGI datasets.

KEYWORDS
VGI; OpenStreetMap; Topic Modelling; LDA; L-LDA; Latent Labeled Dirichlet Allocation; Spatial data quality; FIXME; fixme; Text Mining.

CONTACT: Rajesh, Chittor Sundaram. Email: mr.iyer@hotmail.com

1. Introduction

Volunteered Geographic Information (VGI) (Goodchild 2007) refers to geographic information created by citizens who often are not data collection experts, and are untrained in the formal process of mapping. OpenStreetMap (OSM) is the most popular VGI initiative to collect and distribute open spatial data of the world, that has grown to about 6.5 million contributors at present[1]. OSM data curation is often undertaken by users with heterogeneous expertise, interpreting vague and informal guidelines. The importance of VGI initiatives in general and OSM in particular cannot be over emphasized, considering the freedom and flexibility offered in the mapping. Volunteer mappers responsively collect information about features of interest to the users themselves, leading to multi-faceted data sets reflecting the needs of diverse user communities. Thus, OSM not only caters to the needs of the general user community (by supporting common infrastructure networks such as road and street networks), but also supports special interest groups' mapping needs (maps such as wheelchair routing, drinking water sources and humanitarian relief[2]). While the support for such diverse content is a strength of OSM (discussed across diverse areas such as national security (Papapesios *et al.* 2018), improving accessibility (Zipf *et al.* 2016) and disaster monitoring (Auer *et al.* 2018)), the usability and reliability of its data quality have also been questioned (Salk *et al.* 2016, Hashemi and Ali Abbaspour 2015). OSM mandates manual content creation and curation by users (using armchair and outdoor mapping)[3]. The contributors, however, have vastly varying skills, and are not supported by strict data collection protocols. This, in turn, often leads to non-homogenous and dirty data, not immediately fit for many mapping purposes.

Dirty data is a term used to collectively indicate issues with data such as non-standardized representations of data (Williams 1997), data outliers (Hawkins *et al.* 2002), missing data or incorrect records and values (Simoudis *et al.* 1995), and duplicates (Hernández and Stolfo 1998). Dirty data are not *analysis-ready*. Dirty spatial data may lead to significant and potentially harmful consequences, including liability issues (Blatt 2015). While nuanced strategies to address specific kinds of errors in spatial data exist, the first step towards an effective error mitigation strategy requires the ability to gain an overview of the kinds of errors affecting the data set. OSM is one of the main data sources for critical map services[4] (Arsanjani *et al.* 2015, Corcoran *et al.* 2013) and thereby, it is important to understand its data quality. Given the expansive diversity prevalent in OSM, it is not always viable to assess its data quality against authoritative and reference data sources, a grounded and popular approach in the research community. This is primarily because - (i) comparative approaches with authoritative data are bound by the granularity at which the reference data set is available (e.g., general purpose maps such as street networks) and (ii) authoritative data for a specific mapping or activity domain to make the comparison is rarely available. (e.g., special purpose maps such as marine and shipping networks). In addition, data heterogeneity in OSM is a particular, desirable feature of OSM to which data quality assessment should adapt. Approaching OSM with a fixed ontology of data quality measures may only partially cover issues experienced, as we show here. An ontology of alternative data grounding (including intrinsic indicators) have also been recently

[1] https://www.openstreetmap.org/stats/data_stats.html
[2] https://wiki.openstreetmap.org/wiki/Mapping_projects
[3] https://wiki.openstreetmap.org/wiki/Pick_your_mapping_technique
[4] http://wiki.openstreetmap.org/wiki/List_of_OSM-based_services

discussed by (Mocnik *et al.* 2018).

In this paper, we have focused on understanding the data quality issues in OSM, by harnessing a valuable intrinsic source of knowledge about OSM data quality that remains largely untapped - The tag `FIXME/fixme`[5] (further only capitalized) is used by OSM contributors to highlight potential issues with a given mapped feature encountered during the mapping or curating process. `FIXME` can either serve as a mapper's note to self, or a request for input from peer mappers. Currently, there are over 1.5 million[6] `FIXME` issues documented. While the potential of this latent knowledge about OSM issues held within the `FIXME` tag corpus is extensive, the free-form nature of text presents challenges to the efficient extraction of knowledge, that can serve as a pre-cursor to addressing the challenges of dirty data in OSM. How can knowledge captured in `FIXME` tags be transformed into useful and actionable insights about OSM data quality? We hypothesize that, latent knowledge present within the unstructured text corpus of `FIXME` tags can be effectively mined using Topic Modelling (Hofmann 1999, Blei and Lafferty 2007), to reveal valuable knowledge about OSM data quality, in alignment with standard spatial data quality indicators. Topic Modeling (TM) is a statistical modeling technique, built on the insight that document semantics are primarily governed by latent variables. The hidden semantic structures within a document corpus can be discovered through these *abstract topics* (Lau *et al.* 2010).

The ISO-19157 (ISO 2013) spatial data quality standard was devised to serve professional mapping by trained professionals in mind, following established data collection and curation protocols. While previously applied for the assessment of VGI data quality (Fonte *et al.* 2017), it is unclear how well user-reported OSM data issues align with the typology of errors captured in the standard. We explore the classification and alignment of `FIXME` derived topic models with the ISO-19157 indicators of spatial data quality. We thus contribute to the understanding of the spectrum of user identified data problems, by linking them with standardized spatial quality indicators. Furthermore, this understanding enables to gauge the applicability and effectiveness of standardized data quality indicators to capture the full spectrum of data issues in OSM as a whole - a contested question.

Specifically, in this research article, we make the following contributions:

(1) Proposing a generic yet flexible framework to study and understand OSM data quality issues reported in `FIXME` tags by using the ISO-19157 quality indicators. These indicators serve as the background knowledge to drive a semi-supervised learning process using topic models;
(2) Providing a thorough qualitative analysis of our proposed framework across two different study areas, by exploring the distinct distribution of topics popular within the OSM contributor community;
(3) Conducting a comparative analysis to discover the commonalities and distinctions across the topic distributions using the OSM `FIXME` tags and an authoritative spatial data set.

Our model closely aligns with the *data mining* approaches to VGI quality assessment discussed in (Senaratne *et al.* 2017). The usefulness of this approach cannot be over emphasized, especially in scenarios where authoritative data is not available or useful for making objective comparisons, and being independent of geographic, crowdsourcing or social approaches to quality evaluation (Goodchild and Li 2012). On similar

[5]https://wiki.openstreetmap.org/wiki/Key:fixme
[6]https://taginfo.openstreetmap.org/search?q=fixme

lines, (Keßler and de Groot 2013) propose a trust-based approach modelled on data provenance, also enabling to assess OSM data quality without reference to a ground truth dataset. Finally, we show that our model is not constrained in applicability to VGI alone. This is illustrated by applying it in parallel to OSM and to an authoritative map data set for the same geography (VicMap, for Victoria, Australia). We are thus able to reveal and discuss the possible sources for the distinct patterns of issues found in VGI vs. authoritative datasets.

2. Related work

To make informed decisions about the fitness-for-purpose of a spatial data set, the ability to measure and report spatial data quality is necessary (Boin and Hunter 2008). In this paper, we propose to use the rich, yet noisy textual data available via the `FIXME` tag in the OSM data sets, to assess OSM data quality. To this aim, we consider the ISO-19157 Data Quality Standards (ISO 2013), as our quality indicators. We use a semi-supervised topic modelling approach called Labeled Latent Dirichlet Allocation (L-LDA) (Ramage *et al.* 2009), that uses these quality indicators as labels and make inferences about the quality of the data set, given the `FIXME` tags in that data set. In this section, we discuss the ISO-19157 data quality standards and OSM data quality issues that have been presented in the literature with respect to these standards, and then present an overview of the state-of-the-art in topic modeling, focusing in the context of short textual data, e.g., tweets, and quality monitoring.

2.1. *ISO-19157 data quality standard and OSM data quality*

The ISO 19157 standard establishes the principles for describing the quality of geographic data. Its authors defined data quality elements to describe distinct aspects of geographic data, primarily from the perspective of traditional, centralized mapping efforts typical for mapping agencies. Initial research assessing OSM data quality was primarily comparative, contrasting OSM quality along one of the ISO-19157 data quality elements with external authoritative data sources. Here, we provide the definitions of ISO-19157 data quality elements across the main classification (Level 1) and sub-classification levels (Level 2) in Table 1, and how these have been investigated in OSM-related research. While our study focuses on the indicators shown in bold in Table 1, however, we present all the Level 1 and 2 classifications for completeness.

Completeness *measures the rate of omissions (data absent from the data set) /commissions (excess data present in the data set) for spatial features.* Completeness is a key parameter of fitness for use, and an early focus area of OSM data quality research. Ciepłuch *et al.* (2010) discussed the completeness of OSM in reference to Google / Bing Maps as an authoritative source. Similarly, Zielstra and Zipf (2010) and Ludwig *et al.* (2011) assessed the completeness of the OSM Germany road network against TeleAtlas and NAVTEQ reference data sets, and Zielstra and Hochmair (2011) assessed the completeness of the USA road network against the TIGER data set. Fan *et al.* (2014) evaluated the completeness of OSM building footprints against the German Authority Topographic Cartographic Information System, while Jackson *et al.* (2013) assessed the coverage of school buildings against the US Department of Education's data. These studies generally conclude that the completeness of OSM data is good, but variable.

Thematic Accuracy *pertains to the accuracy of quantitative attributes, correct-*

ness of non-quantitative attributes, and of the classification of map features. Girres and Touya (2010) and Mooney and Corcoran (2012) note that OSM map features are generally weakly annotated, highlighting this to be a pressing issue in the data set. While Girres and Touya (2010) use a reference data set, the analysis of Mooney and Corcoran (2012) is confined to contents of the OSM data sets of Ireland, United Kingdom, Germany and Austria. Ludwig *et al.* (2011) and Neis *et al.* (2012) highlight the issue of OSM attributes incompleteness, as a hurdle to solving advanced GIS problems. Ballatore and Bertolotto (2011) and Ballatore *et al.* (2013) similarly comment on the poor state of OSM semantics. A subsequent analysis by Vandecasteele and Devillers (2015) notes that the high semantic heterogeneity in VGI data sets presents challenges to broader adoption, which is supported by the findings of Davidovic *et al.* (2016), discussing semantic issues in OSM data for over 40 cities of the world.

Positional Accuracy is *defined as the accuracy of the position of features within a spatial reference system.* It includes the sub-categories of Absolute, Relative, and Gridded data accuracy). Of note here is the *Absolute, or External Accuracy* (defined as the closeness of reported coordinate values to values accepted as or being true), which has been the focus of extensive research in OSM quality analysis. While Haklay (2010) analyzed the positional accuracy of UK OSM linear features (e.g., roads and walkways) by comparing them with Ordnance Survey data sets, others undertook similar assessments for Germany (Zielstra and Zipf 2010, Helbich and Amelunxen 2012) and against the French National Mapping Agency's authoritative data (Girres and Touya 2010).

Logical Consistency *measures the degree of adherence to the logical rules of the data structure, attribution and their relationships.* While data in-homogeneity ensuing from the lack of a strict logical schema leads to OSM data quality problems (Majic *et al.* 2017), it is impossible to quantify the level of disagreement with a *defined* schema in OSM. This similarly applies to the sub-category *Topological Consistency (defined as the correctness of the explicitly encoded topological relationships between map features)* capturing qualitative relationship issues, such as between road features (Will 2014, Barron *et al.* 2013, Corcoran *et al.* 2010, Girres and Touya 2010, Neis *et al.* 2012, Majic *et al.* 2019).

Temporal Quality is *defined as the quality of temporal attributes and temporal relationships of spatial features in the data set.* The Temporal Quality of OSM data set is not reflecting a single, time-stamped mapping of the state of the world, as the data set is updated continuously. The ISO definition of temporal quality thus, does not apply to OSM. To address this, recent studies started proposing novel, intrinsic measures capturing distinct OSM data quality issues, such as local data set maturity (Maguire and Tomko 2017).

Usability Element *is primarily driven by the user requirements to ascertain the fitness of use of a spatial data, set to serve a given purpose.* The ISO standard recommends the incorporation of specific user requirements together with the above dimensions to evaluate the usability of a data set.

The vast and latent source of knowledge within `FIXME` has thus far been little studied. Our analysis handles the dossier of pre-eminent mapping challenges documented in `FIXME` by OSM contributors, as a fundamental indicator of OSM pain points, and hence its data quality. In addition, our study represents a first in terms of (i) harnessing this latent knowledge to broadly understand OSM's *'State of the Map'* and (ii)

present a unified, fine-grained analysis of OSM data quality issues, by considering all applicable measures of spatial data quality indicators from the ISO framework.

2.2. *Latent topic mining*

Unstructured text is a valuable source of information for knowledge discovery, and facilitate new data-driven hypotheses through the application of Natural Language Processing (NLP) techniques (Basole *et al.* 2013). In particular, probabilistic knowledge discovery models such as Topic Models (Lafferty and Blei 2009, Griffiths and Steyvers 2004) are effective in understanding the semantic structure of an unstructured document corpus based on a Bayesian analysis of the underlying text (Stevens *et al.* (2012)). Latent Dirichlet Allocation (LDA) (Blei *et al.* 2003) is an unsupervised probabilistic topic modelling algorithm that has been widely used to identify topics from a text corpus. The intuition behind LDA is that each document in a corpus can be modelled as an underlying distribution of latent topics, and each topic can be modelled as an underlying distribution of words. In the context of geo-textual data, geo-tagged tweets from Twitter have been used in Lansley and Longley (2016) to expose behavioral classifications of user trends in a localized geography, while a similar classification experiment based on an aggregated corpus of twitter messages is presented by (Hong and Davison 2010). In contrast, the usefulness of LDA and topic models to mine knowledge from VGI data sets with a spatial footprint, have been discussed using locations (or places). In addition to LDA being a central tenet towards addressing place name disambiguation (Ju *et al.* 2016), the algorithm has also been effective in discovering unique characteristics of places and mapping places with a unique thematic distinction, using natural language data from VGI travel blogs (Adams and McKenzie 2013).

While topic models generated from LDA are useful for exploratory analysis revealing broad textual patterns in document collections, they may expose topics that are not meaningful for a specific analysis (Chang *et al.* 2009), making the interpretation difficult (Ramage *et al.* 2011). This is a problem common to all unsupervised methods. Furthermore, topics in documents may not be clearly differentiated or cover multiple topics. This highly affects short documents, including OSM `FIXME` issues. Indeed, the applicability of LDA for the analysis of shorter text has been questioned (Andrienko *et al.* 2013), with an alternative approach proposed, such as linking authors to topics Steyvers *et al.* (2004). This is, however, not applicable here, as issues highlighted by individual OSM contributors are heterogeneous, as opposed to a homogeneous group of inter-related topics as occurring in traditional documents Ertl *et al.* (2012).

Considering the limitations and shortcomings of LDA when handling short and noisy textual data (Jin *et al.* 2011, Chen *et al.* 2014), that can otherwise serve a key role in revealing cardinal quality facets (Rahimi *et al.* 2020), we propose to harness the effectiveness of L-LDA in mining latent knowledge from a corpus of short texts (Kataria and Agarwal 2015, Lingad *et al.* 2013, Ritter *et al.* 2011, Ramage *et al.* 2010), in order to assess OSM data quality, using short texts in the `FIXME` corpus.

Table 1.: ISO 19157 - Geographic information data quality measures. ‡ - FIXME discussion not applicable, † - FIXME discussion (Level 1), § - FIXME discussion across one or more indicators (Level 2)

ID	Classification (Level 1)	Sub Classification (Level 2)	Description
1	Completeness (**CO**) (†)		Indicates presence or absence of features, their attributes and relationships.
		Omission	Indicates data absent from the data set.
		Commission	Indicates excess data present in the data set.
2	Thematic Accuracy (§)		Accuracy of Quantitative Attributes, correctness of non-quantitative attributes, and the classification of features and their relationships
		Non-Quantitative Attribute Correctness **(NQAC)**	Measure of whether a non-quantitative attribute is correct.
		Quantitative Attribute Accuracy **(QAA)**	Closeness of the value of a quantitative attribute to a value accepted as or known to be true.
		Classification Correctness	Comparison of the classes assigned to features or their attributes to a university of discourse (ground truth).
3	Positional Accuracy **(PA)** (†)	Absolute External Positional Accuracy	Closeness of reported coordinate values, to values accepted as, or being true.
		Relative Internal Positional Accuracy	Closeness of relative position of features in the data set to their respective relative positions accepted as, or being true.
		Gridded Data Positional Accuracy	Closeness of gridded data spatial positional values to values accepted as, or being true.
4	Logical Consistency (‡)		Degree of adherence to logical rules of data structure, attribution, and relationships
		Conceptual Consistency	Adherence to rules of conceptual schema
		Domain Consistency	Adherence of values to value domains
		Format Consistency	Degree to which data is stored in accordance with the physical structure of the data set
		Topological Consistency	Correctness of the explicitly encoded topological characteristics of the data set
5	Temporal Quality (‡)		Quality of temporal attributes, and temporal relationships of features
		Accuracy of a Time Measurement	Closeness of time measurements to values accepted as, or known to be true
		Temporal Consistency	Correctness of the order of events
		Temporal Validity	Validity of data with respect to time
6	Usability Element (‡)		Based on user requirements. Aggregation of quality measures for usability evaluation.

3. Preliminary concepts

3.1. *OSM tagging practice*

OSM map features have free-form text tags associated with them. Each of these key-value pair describe a specific attribute of the map feature[7]. Each feature must have at least a single tag, but there is no upper limit. The OSM tagging practice in loosely governed by an informal set of guidelines[8]. The application of OSM tagging practice(s) has been extensively discussed by Mooney and Corcoran (2012).

The tag `FIXME` allows OSM contributors to mark objects that require further attention. A related OSM tag called `NOTES`[9] allows comments from both OSM contributors as well as anonymous visitors to be added to indicate general map issues. `NOTES` are not directly tagged to any OSM object. Since anonymous comments can also be posted, a review of a sample subset of `NOTES` data exposed information not relating to map issues (such as general discussions and marketing information). `NOTES` data is not considered for our analysis. A typical `key-value` pair for a `FIXME` tag is shown below.

```
{fixme=Please specify unambiguous opening hours in the proper syntax
 in 24 hour format. "5-11" -- could be 05:00-11:00, 05:00-23:00, or
 17:00-23:00, delivery=no, addr:state=ME, smoking=no, cuisine=pizza,
 addr:street=Iron Point Rd, takeaway=yes, addr:postcode=04853,
 capacity=50, addr:city=North Haven, addr:country=US}
```

In this example, the OSM user indicates that the opening hours of a spatial feature (restaurant) are to be corrected. `FIXME` tags serve as a call for guidance or to request additional mapping input from other mappers. Some common issues seen in `FIXME` tags are requests for resurveys when dealing with objects with unconfirmed coordinates, untagged nodes (objects with missing street names) and missing attributes (e.g., building address). The free text nature of `FIXME` tags makes it difficult to automatically categorize and analyze the knowledge about OSM data quality issues.

3.2. *Labelled latent dirichlet allocation (L-LDA)*

An alternative to unsupervised topic modelling is offered through fine tuning of topics in a semi-supervised approach, enabling to focus on the discovery of only a pre-selected set of topics (e.g., ISO-19157 data quality elements as in Table 1). A semi-supervised refinement of LDA is known as the Labelled Latent Dirichlet Allocation (L-LDA) algorithm (Ramage *et al.* 2009). This is achieved by providing the algorithm with a set of seed words (labels) that are believed to be optimally representative of the topics in the document corpus. A semi-supervised approach thus constrains the algorithm's multinomial topic distribution by the seed words, and thus leads to the identification of topics aligned with the scope of the analysis. Figure 1 shows a graphical representation of the L-LDA algorithm.

Formally, the objective of L-LDA is similar to LDA, with the main difference being that, while LDA generates the multinomial mixture distribution θ^d, for each document d, considering all K topics, the L-LDA algorithm constrains θ to only those topics that correspond to the document's observed labels ($\Lambda_d = (l_1, l_2, l_3, \ldots, l_K)$, and each

[7] https://wiki.openstreetmap.org/wiki/Tags
[8] http://wiki.openstreetmap.org/wiki/Map_Features
[9] https://wiki.openstreetmap.org/wiki/Notes

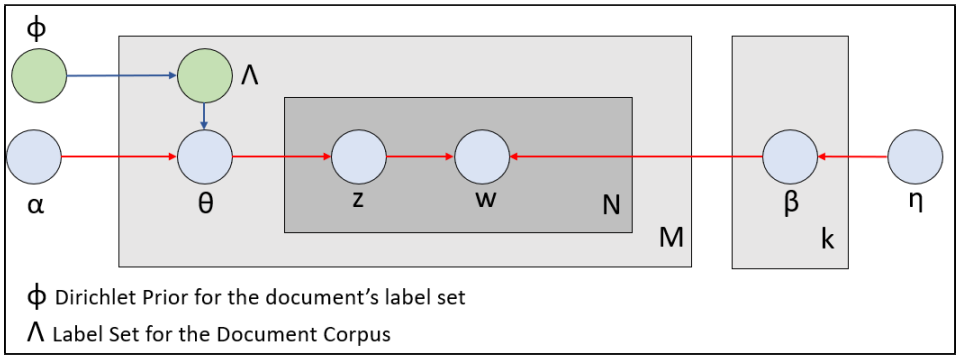

Figure 1.: Graphical Model of Labelled Latent Dirichlet Allocation.

$l_k \in \{0, 1\}$). The dependency of θ on a combination of α and Λ is indicated in Figure 1.

Both LDA and L-LDA require the setting of two hyper-parameters, also known as Dirichlet priors, as mentioned below.

α - Controls the mixture of topics for any given document. Low α values indicate the topic distribution samples to be near the topics, i.e., a document could possibly be associated with only one topic. A value of 1 for α indicates that a distribution sample of documents associated with one topic has the same likelihood as a sample of documents associated with an even mixture of all the documents, or a distributions sample that is something in-between. Higher values of $\alpha(> 1)$ tend to make the distribution samples more uniform with an even mixture of all the samples.

β - Controls the distribution of words per topic. A lower value of β result in topics with less words, while a higher value will result in topics represented by more words.

4. Methodology

Here, we categorise map issues tagged by OSM contributors to the broad ISO-19157 categories. The collection of issues captured by OSM `FIXME` tags can be considered a document corpus. We model the content of topics in `FIXME` tag descriptions by training a semi-supervised L-LDA model based on a labelled document corpus, where each document relates to one or more standardized indicators of spatial data quality.

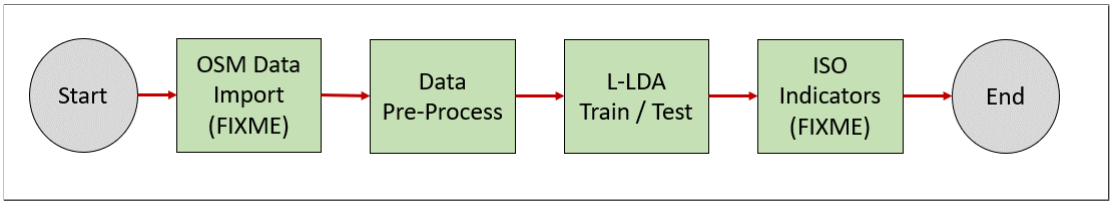

Figure 2.: Methodology

4.1. *Data sets*

In this work we use two data sets, namely OSM VGI data set and VicMap data set for the state of Victoria (Australia), details of which are presented in the following.

4.1.1. OSM data

For the analysis, we use OSM data for North America and Australia, downloaded from Geofabrik[10]. The data sets are current as on March 2019. The OSM data is imported into a PostGIS database for analysis, accounting for the consideration of only those spatial entities that have the `FIXME/fixme` tag populated (both capitalized and non-capitalized). In addition, since our objective is to derive insights from OSM contributor community (and avoid bias in the analysis), `FIXME` tags from bulk import of data (such as the National Hydrography data in USA, contributing to about 77% of the data) are not considered. In addition, `FIXME` feature tags are filtered to remove non-English tags, using a language detection framework[11]. The final OSM `FIXME` North America corpus has about 64068 records, and the Australia (Victoria) corpus has about 8700 records. The average length of a `FIXME` record is about five words.

4.1.2. OSM FIXME Corpus selection for L-LDA

The L-LDA model has been trained and evaluated across each of the four census areas of USA separately. The `FIXME` corpus has about 6644 records in the NE region. Out of this, we selected uniformly at random two sets of records, each set with 1000 `FIXME` records. The first set of 1000 records is used to train the L-LDA algorithm for the NE region, while the second set is then used to apply the trained L-LDA model for the NE region. We train and evaluate the L-LDA algorithm for the remaining three census areas of USA by following a similar approach and thus create training and test data sets (see Section 5.2). The OSM `FIXME` corpus has about 6075 records in the MW region, 9898 in the SO region, and 11,650 records in the WE region. Similarly, we have created training and validation data sets for OSM Australia (Victoria).

4.1.3. Vicmap Data

The Department of Environment, Land, Water and Planning is the custodian of authoritative map data for the state of Victoria (Australia)[12]. We compare Vicmap issue reports to the issues reported in `FIXME` tags from OSM. The Vicmap issue reports data set is current as of March 2019, detailing data quality issues reported by authorized users from local governments, utilities and other specialist users of authoritative map data. Similar to the process done with Victoria data set from OSM, we select uniformly at random 1000 records each for the training and testing phases to compare the distribution of topics between OSM and authoritative data sets.

4.2. *ISO quality measures as labels*

As discussed in Section 2.1, the ISO 19157 has identified six measures to determine the quality of a data set. Here we discuss how we adapt these measures to derive labels that will be eventually used in training our L-LDA model.

Completeness: Existing studies consider this indicator holistically, and do not differentiate between missing or excess data. We adopt a similar approach, and treat data completeness issues in its entirety. As we observe later, `FIXME` issues are still dominated by missing data reports, i.e., omissions in the data set.

[10] https://download.geofabrik.de/
[11] https://pypi.org/project/langdetect/
[12] https://www2.delwp.vic.gov.au/

Thematic Accuracy: Current studies discuss this quality indicator across sub-categories *Non-Quantitative Attribute Correctness,* and *Quantitative Attribute Accuracy.* Considering the free form nature of OSM tagging, attribute issues with OSM geometries (including issues reported in `FIXME`) would closely fit with these two classifications (e.g., *Street Name* of a building representing a non-quantitative attribute v/s *Speed Limit* of a road segment representing a quantitative attribute). The OSM tagging framework and data model constrain quality assessments for sub-category *Classification Correctness*, and hence not considered in our model.

Positional Accuracy: Considering the data sources for OSM curation[13] (including authorized aerial and satellite imagery), positional accuracy issues documented in `FIXME` are consistent with the ISO sub-classification of *Absolute Accuracy.* Therefore, our discussion on positional accuracy (PA) only refers to *Absolute Accuracy.*

Logical Consistency: Since OSM lacks an enforced logical schema, logical consistency is not strictly definable, and hence not considered as a label in this paper.

Temporal Quality: The dynamic nature of the OSM data does not allow to adapt this data quality measure for assessment and hence, it is not discussed in this paper.

Usability Element: This quality indicator represents an aggregated dimension of the previous indicators and hence not presented separately in our discussion.

In summary, we further investigate how to conduct an intrinsic analysis of the data quality issues documented using `FIXME` tags, across four major classifications: Completeness (CO), Non-Quantitative Attribute Correctness (NQAC), Quantitative Attribute Accuracy (QAA) and Positional Accuracy (PA). These serve as the labels for our topic model. When labelling the training data for L-LDA (a manual and expert judgement driven process), we have confirmed our arguments constraining the applicability of the analysis, to these four indicators, to be valid for the data set.

4.3. *Application of L-LDA*

Consider a set of all `FIXME` issues as the main document corpus. These issues can either be considered for an entire country, or a specific region (e.g., a census area in the United States of America). Considering the Australian data corpus, $D_{AUS} = \{\mathbf{w}_{aus\_1}, \mathbf{w}_{aus\_2}, \mathbf{w}_{aus\_3}, \dots \mathbf{w}_{aus\_m}\}$ would represent the document corpus in the English Language, after pre-processing (mentioned in Section 6). For each document $\mathbf{w}$ belonging to D_{AUS}, let $w = (w_1, w_2, w_3, \dots w_N)$, be the word indices representing the document, and (w_N) representing the length of the document. Let V represent the total size of the vocabulary from the `FIXME` corpus. Let K represent the total number of unique labels, each representing a specific ISO-19157 spatial data quality indicator. For our experiments and discussion, $K = \{\Omega_{PA}, \Omega_{CO}, \Omega_{NQAC}, \Omega_{QAA}\}$, would also represent the number of topics (ISO Quality elements) for the L-LDA algorithm. Finally, let $\wedge_d = (l_1, l_2, l_3, \dots l_K)$ represent a list of binary presence/absence indicators for each spatial data quality indicator K, such that each $w_i \in \{1, \dots, V\}$ and each $l_k \in \{0, 1\}$. Given these inputs, our objective is to draw the multinomial topic distributions over vocabulary β_k for each topic k, from a Dirichlet prior η. Additionally, the multinomial mixture distribution $\theta_{(d)}$, generated over all K topics for each document d from a Dirichlet prior α, is to be restricted to only those topics, corresponding to the labels in $\wedge_{(d)}$. The majority of OSM `FIXME` tags are brief texts indicative of a primary

[13] https://wiki.openstreetmap.org/wiki/Potential_Datasources

ISO quality dimension (such as the completeness of a line geometry or the positional accuracy of a point geometry), with recurring common terms used. Hence, the hyper parameters in our experiments were set heuristically with values less than 1.

The process outlined in this section is generic, and the current analysis/discussion can be extended to represent any data quality standard or any corpus of quality reports or issues. This is done by modifying K to represent the unique labels for the new quality framework, and the subsequent re-training of the L-LDA algorithm. The symbols and terminologies mentioned in the problem statement is grounded in the L-LDA model, and documented in the Appendix.

4.3.1. Learning and inference

The underpinnings of LDA and L-LDA relies on the principle of posterior inference, a process where learning posterior distributions of latent variables is ascertained from the observed data. Since posterior inference distributions are intractable to compute, approximate inference techniques such as Gibbs sampling (Griffiths and Steyvers 2004) are used. Since in LDA and L-LDA the latent '*document-topic*' and '*topic-word*' distributions can be inferred using only the '*topic-index*' assignments, a simpler process, referred to as a collapsed Gibbs sampler, that samples only from the '*topic-index*' assignments is sufficient. Once the topic multinomials are learned from the training set, we perform an inference on test documents using the same sampling process. From the output of the L-LDA algorithm, we rank the user specific labels in their order of relevance to each document to determine an optimal ISO data quality indicator for each `FIXME` issue, and thus achieve a final categorization for the test data set.

5. Experimental analysis

In this section we detail the experimental results of the conducted study, focusing on two distinct geographical regions, USA and the state of Victoria, Australia. The results for USA have been stratified across the 4 census regions. The availability of an authoritative data set for Victoria facilitated in comparing the nature of the distributions across different quality indicators, between VGI and reference data sets.

5.1. *Experimental setup*

For the analysis, OSM Data was imported into a PostgreSQL database having PostGIS enabled. In addition, data pre-processing tasks and the L-LDA main algorithm was implemented using Python 3.7. An implementation of the L-LDA algorithm from (Hong 2019) has been customized to suit our needs. All experiments were conducted on a cloud-based server with 8 virtual CPU's, 32 GB RAM, and a 5TB disk, running Ubuntu Linux 16.04 LTS operating system.

5.2. *L-LDA Model Evaluation*

An ISO-19157 quality indicator label have been manually assigned to FIXME documents in all training data sets. A primary labelling has been performed by the first author using expert judgement when applying ISO category definitions, and cross verified by one of the co-authors. Labelling disagreements were resolved by consensus.

For each training data set (one for each census area in USA, and one for Victoria, Australia), the L-LDA model accuracy was evaluated using a 5-fold cross validation (Arlot and Celisse 2010), in each partition of the training data. The mean accuracy of the learned models for each region (and all indicators) are highly consistent: for NE region 81.7%, WE region 84.2%, SO region 87.6%, and MW region 88.7%. For OSM Australia (Victoria), the overall accuracy of the model is 92.2%.

The core objective of the training is to closely align the model learning with the ISO quality indicators, thereby making recall, a priority to our objective. The recall for individual ISO quality indicators in NE are (PA: 84.7%, NQAC: 74.4%, CO: 82.2%, QAA: 98.6%, common_topic: 72.4%). Similarly, in WE region, we see (PA: 82.6%, NQAC: 52.5%, CO: 86.7%, QAA: 61.5%, common_topic: 95.9%). Furthermore, the recall for SO region are (PA: 85.6%, NQAC: 51.9%, CO: 94.1%, QAA: 89.3%, common_topic: 78.8%), and for the MW region are (PA: 92.6%, NQAC: 59.5%, CO: 93.1%, QAA: 81.7%, common_topic: 87.1%). For OSM Australia (Victoria), the recall was observed as (PA: 87.3%, NQAC: 94.6%, CO: 84.9%, QAA: 83.3%, common_topic: 95.6%).

5.3. *Experimental results*

The distribution of issues based on ISO-19157 quality indicators is shown in Figure 3 (USA census areas) and Figure 8 (OSM vs. VicMap, Australia). Table 3 shows messages documented in both VGI data (OSM) and authoritative data (VicMap) sets, as an indicator of different styles adopted by contributors of data issues. Table 2 shows the most frequent words associated with the ISO-19157 quality indicator topics learned by L-LDA across the two geographies, for the OSM data set. The labels for the model training, comprising a sub set of applicable ISO quality indicators, comprises the main topics against which the `FIXME` issues would be categorized by L-LDA.

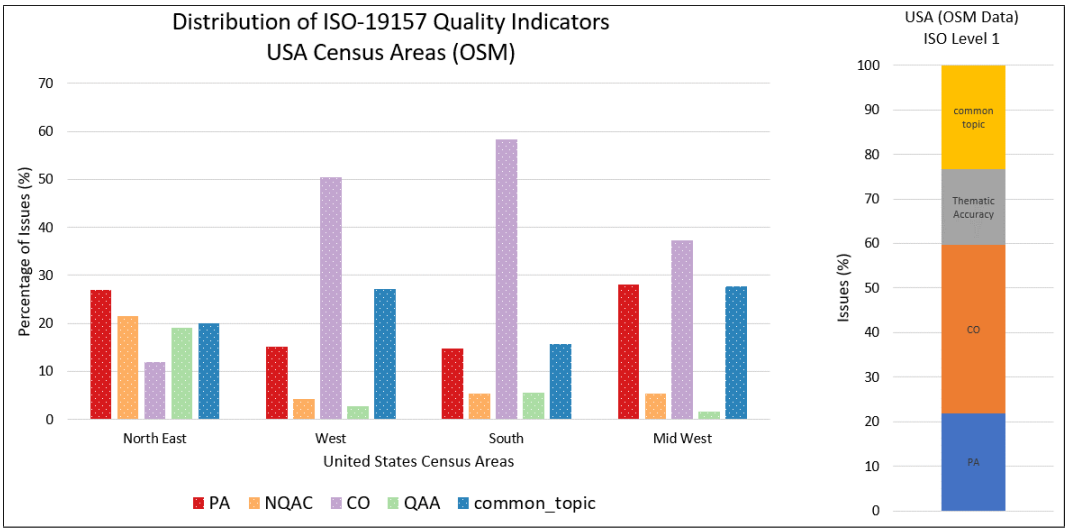

Figure 3.: FIXME Issues across USA Census Areas and Complete USA (ISO Level 1)

Generally, we observe that the distribution of issues is very different in the NE region of the United States compared to the WE and SO regions. A major percentage of issues is related to Completeness in all census areas except the NE, where the distribution of issues across the categories is more evenly spread with the least issues being reported

for Completeness. Positional Accuracy issues reported in `FIXME` are prominent across the NE and MW regions accounting for about 27% of the total issues. While these issues are not as frequent in the WE and SO regions, they still account for about 15% of the total. Thematic Accuracy issues (represented collectively using the categories Non Quantitative Attribute Completeness (NQAC) and Quantitative Attribute Accuracy (QAA) in Figures 3 and 8) account for a major share of issues in the NE, representing about 40% of total issues. This category accounts for about 7% of total issues for WE and MW and marginally higher (about 11%) in SO states.

Data issues related to Completeness (CO) account for a majority of issues across all census regions except the NE (SO: 50%, MW: 37%). This could possibly indicate that OSM data are more mature in the NE census region of USA, and relate to the number of active OSM mappers concentrating in this densely populated, more urbanized region. For example, the OSM edit density map utility discussed in Anderson (2016) visualizes the focus areas of OSM mappers and shows that OSM contributors are more focused around the NE census area.

Finally, we report L-LDA results where `FIXME` documents could not be tagged against any of the ISO-19157 quality indicators as `common_topic`. This occurs when L-LDA is not able to associate the `FIXME` document specifically with any of the indicators that it was trained with. This is mainly due to these issues being represented by a group of one or more words that are not significantly associated with any single ISO quality indicator. This may be related to varying styles of documenting an issue (due to the free form nature of OSM tagging) using `FIXME`, but also to the ambiguity in the vocabulary used. L-LDA then fails to associate the topic with sufficient certainty to a single label. Uncategorized issues represent over 20% of the issues across all the regions except SO. Based on the heterogeneity in user language and styles adopted for discussing map issues, our model has been able to categorize over 80% of the total issues analyzed as belonging to one or more of the standardized ISO data quality indicators. The content of the `common_topic` set needs to be investigated further, as these issues could potentially represent important quality concerns of interest to the OSM mapping community, but possibly not well captured by ISO indicators. Examples of these un-categorized messages from the US data set are below:

```
[PA]:   This should be moved to the actual building where it is located.
[QAA]:  somewhere around here I saw a 35 mph sign (around 1062 outbound)
        need to find where it changes and map it
[NQAC]: All I know is the name from the sign. What is kind of business?
[CO]:   construction in Bing, empty in USGS
```

5.3.1. FIXME distribution across the USA

Figures 4, 5, 6, and 7 capture the distribution of `FIXME` issues per attributed ISO quality indicator across each census area of the USA. As expected, we observe that OSM contributor mapping activities (and hence the number of issues documented) are dominant in the more populous states within each of the census areas. For example, over 90% of PA issues in the NE region are reported for New York and Pennsylvania, both states with high population densities[14]. Similar patterns are manifest in other census areas, such as Illinois/Indiana in MW and California/Washington in WE regions. As shown in Figure 3, more populous states in NE suffer less of issues relating

[14]https://www.census.gov/library/publications/2011/compendia/statab/131ed/population.html

Table 2.: LLDA Topics and Terms revealed from ISO-19157 Spatial Quality Indicators

ID	LLDA Algorithm Topic	Top Terms United States Census Areas North, South, Mid West and West	Top Terms Australia
1	Completeness	continue, footprint, outline, trace, track, line	continue, trail, path, connects
2	Non Quantitative Attribute Correctness	address, street, name, slipway, type, capacity	name, restrictions, type, classification, building=yes, landuse=forest
3	Quantitative Attribute Accuracy	range, speed, time, closing, housenumber, addr:housenumber, limit, opening_hours, closing, elevation	maxspeed
4	Positional Accuracy	location, approximate, verify, position, survey, placement, resurvey	location, approximate, verify, position, survey, estimated, imagery

to Completeness (CO). To balance these significant distribution effects and maintain some comparability, in our maps the proportions are reported per each census area.

5.3.2. FIXME distribution in Victoria, Australia (OSM and VicMap)

A similar analysis was undertaken using the OSM dataset of Victoria, Australia. While Thematic Accuracy issues account for a majority of total issues, at about 53%, issues across Positional Accuracy and Completeness seems to be evenly distributed at about 12% and 14% respectively. The challenges of L-LDA-based FIXME classification by ISO quality indicators found in the US OSM dataset exhibit a similar pattern in the Australian data set. About 19% of the `FIXME` issues cannot be assigned to one of the ISO quality indicators.

Yet, this is not a problem of the L-LDA model, but of the applicability of the ISO quality indicators to a VGI dataset. Furthermore, the generality of the L-LDA model's applicability noted in Section 1 is demonstrated by analyzing the distribution of quality issues in VicMap, the official general-purpose map data set for Victoria, Australia (Figure 8). Completeness data issues are dominant in VicMap, contributing to about 66% of the total reported issues, while Positional Accuracy issues are consistent those observed in OSM at about 7%. While Thematic Accuracy issues are less than half of those reported in OSM (at about 25%), issues across different sub-classifications are commonly observed in both the data sets.

This large fraction of *common_topic* issues does not manifest in the VicMap authoritative data set. The standardized tools, restricted set of authorized reporters, and established protocols for reporting issues that enable the quality assurance process in VicMap leads to only about 2% of the issues not attributable to an ISO standard quality indicator (Figure 8). VicMap, as any authoritative dataset, is a product of a regulated, rigid top-down mapping process, for which these quality indicators have not only been designed, but along which errors are reported, too.

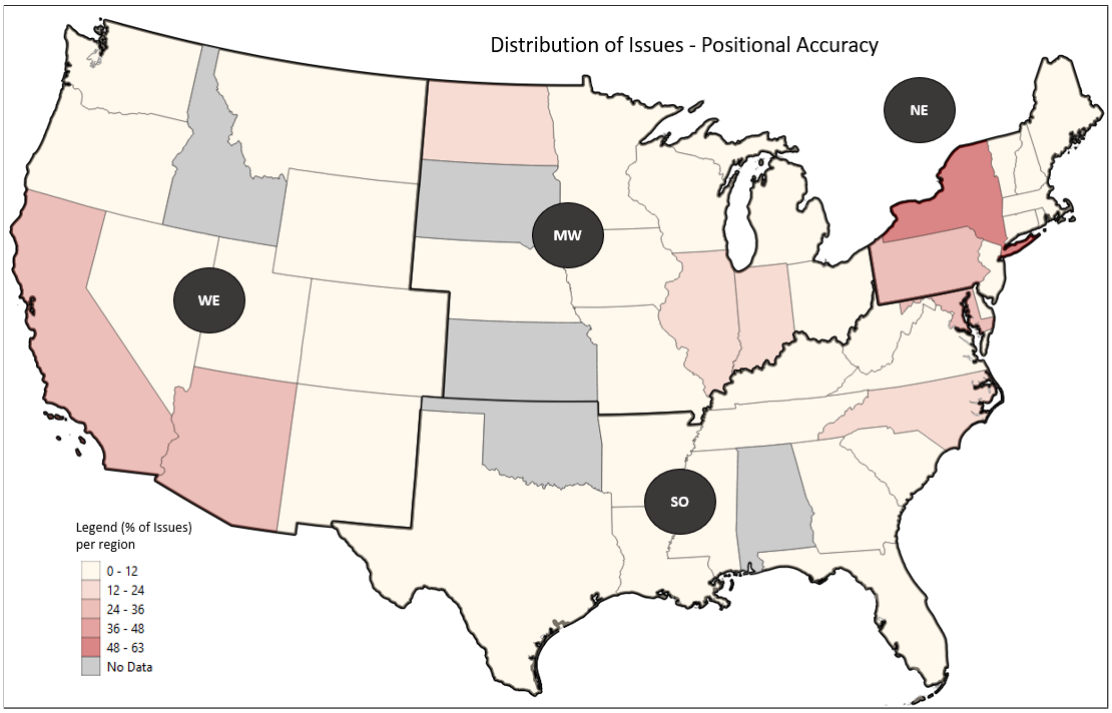

Figure 4.: Distribution of FIXME Issues, USA census regions - ISO Quality Indicator: Positional Accuracy (PA)

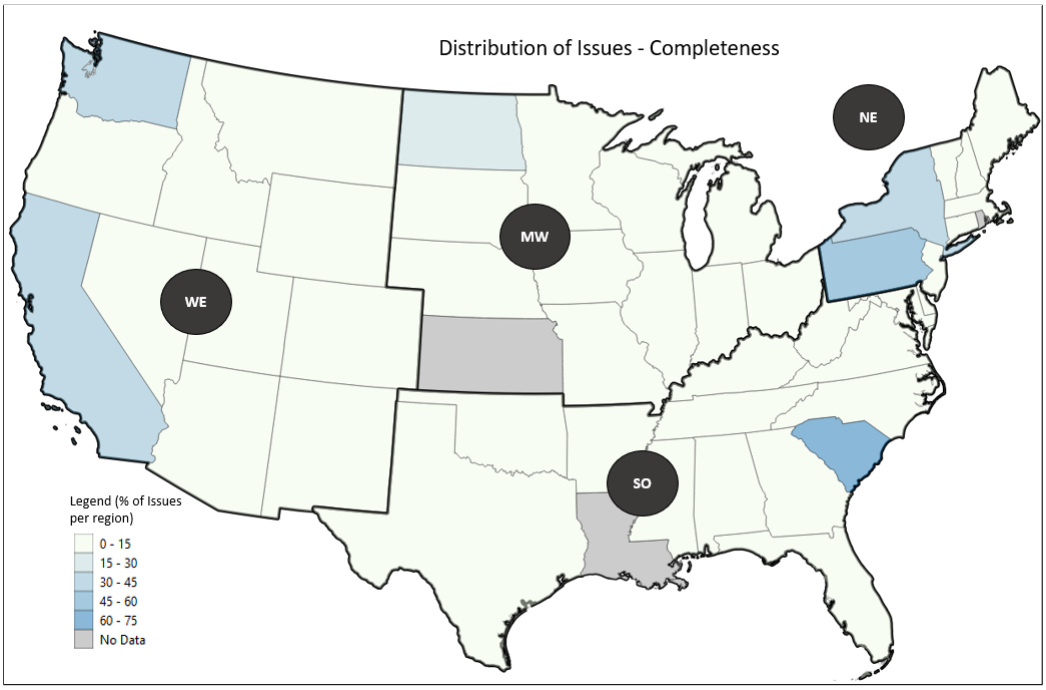

Figure 5.: Distribution of FIXME Issues, USA census regions - ISO Quality Indicator: Completeness (CO)

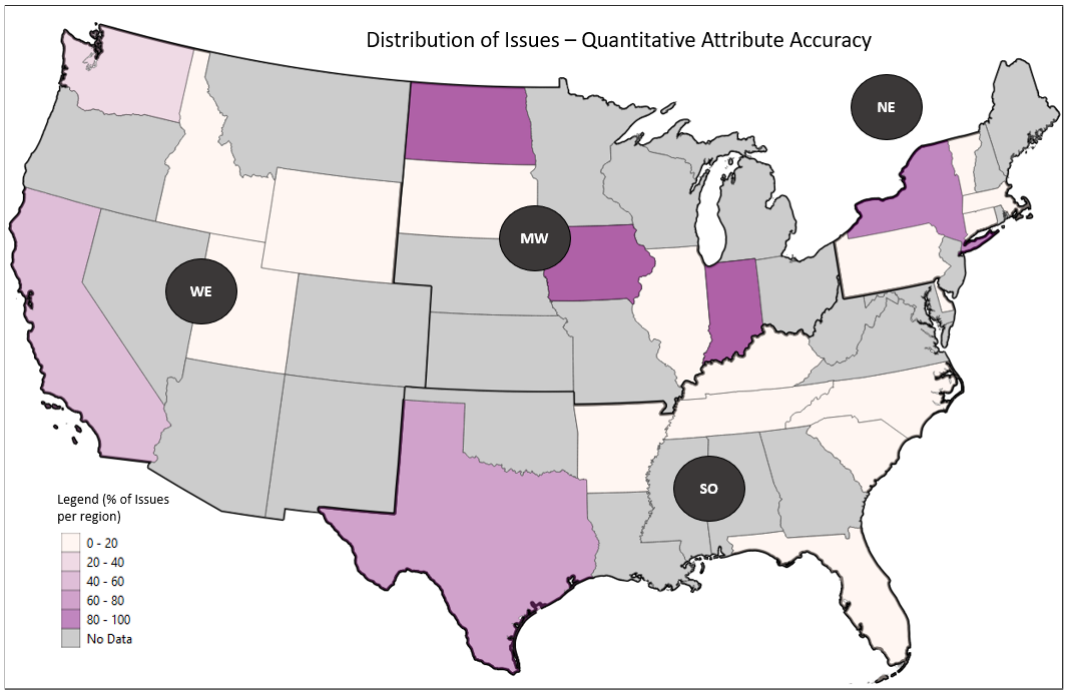

Figure 6.: Distribution of FIXME Issues, USA census regions - ISO Quality Indicator: Quantitative Attribute Accuracy (QAA)

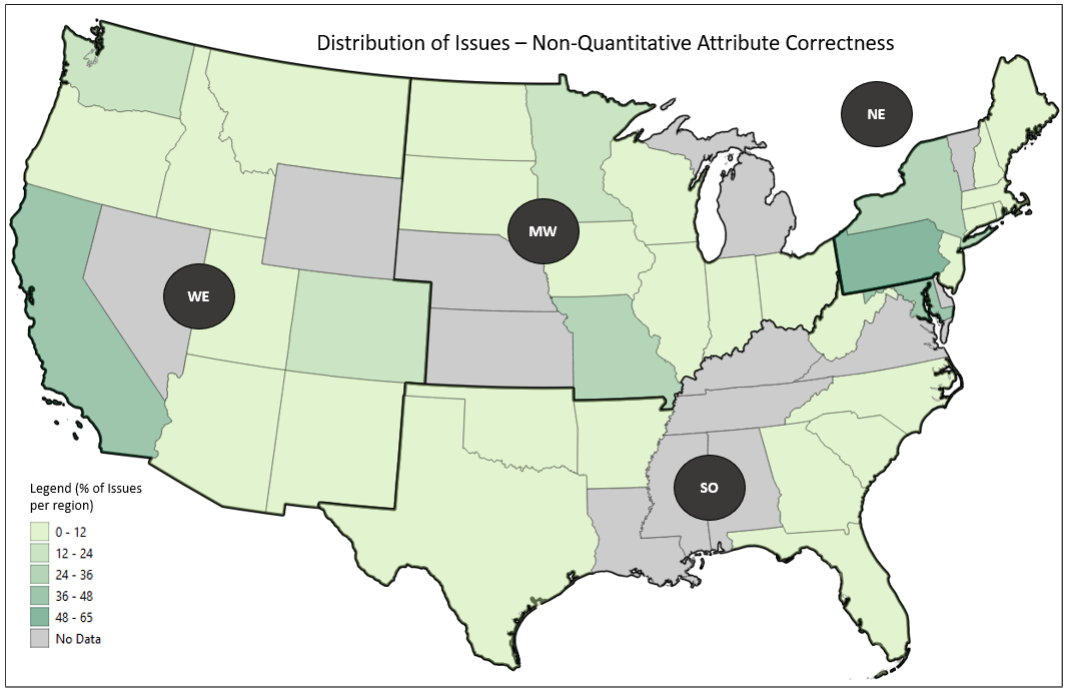

Figure 7.: Distribution of FIXME Issues, USA census regions - ISO Quality Indicator: Non-Quantitative Attribute Correctness (NQAC)

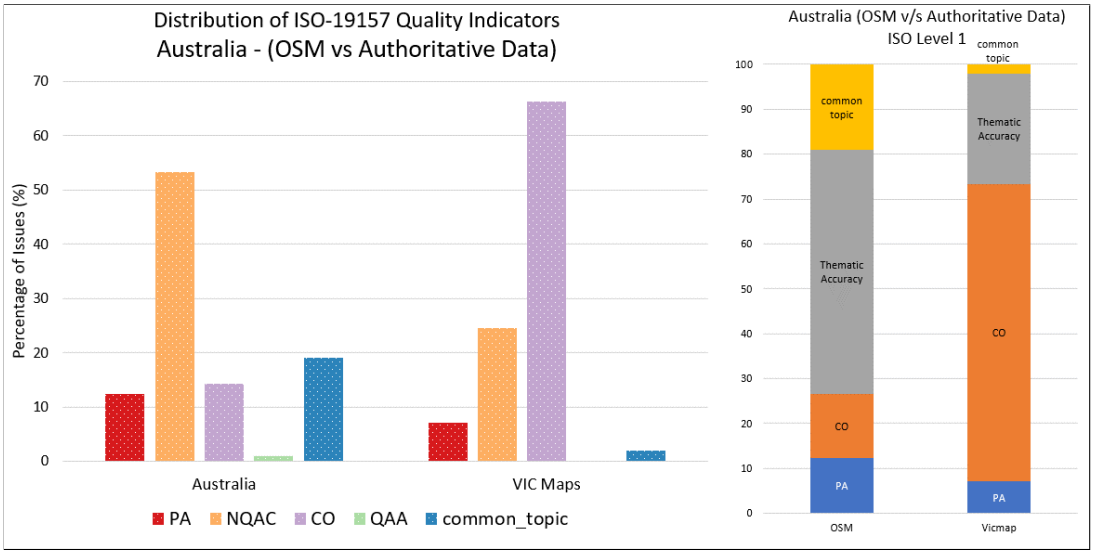

Figure 8.: Australia Data Quality Issues - OSM v/s Reference Data Set - A comparison

6. Findings and discussion

The distribution of data quality issues across two geographical regions (Figure 3 and 8) exposes a new perspective in understanding the data quality of OSM. From our observations, the distribution of OSM data issues, as seen through ISO data quality indicators, exhibits both similarities and differences with the reviewed studies (see Section 2). Thematic Accuracy issues (represented collectively by Non-Quantitative Attribute Completeness (NQAC) and Quantitative Attribute Accuracy (QAA)) is generally poor in OSM and an impediment to the usability of OSM data. Our findings confirm these observations: about 40% of the total data quality issues reported in the North-East Census Area of USA belong collectively to Thematic Accuracy issues (at about 21.5% and 19% respectively). Even though the overall percentage of such issues are lesser for the Mid-Western and the Western regions, they still account for about 7% of the overall issues reported. Combined, issues in these categories represent about 11% of all issues in the Southern states – still a substantial fraction.

Contrary to previous research about Completeness and Positional Accuracy of OSM (assessed as comparable to, or exceeding the quality of reference data sets), we find that more than half of the total issues reported by OSM contributors across the Western and Southern regions of USA belong to this category, at about 50.5% and 58% respectively. These issues are also prevalent in the Mid-Western region, accounting for about 38% of the issues reported. Positional Accuracy issues are dominant in the North-East and Mid-Western regions (27% and 28%, respectively), but present a lower fraction of all issues (15%) in the Western and Southern census regions (Figure 3).

Comparative approaches to data quality assessment of OSM are bound by the granularity at which the reference data set captures data about the world. In contrast, our observations here are based on reports by contributors with diverse mapping interests. This is important, since OSM can often be more detailed than a reference data set, with the volunteer mappers having the freedom and flexibility to cater to diverse mapping needs of local communities. Thus, our observations do not categorically dispute the validity of past research discussions, but rather provide another facet about the

Table 3.: Sample Illustrations: Messages from OSM and VicMap

ID	ISO Quality Indicator	OSM FIXME	VicMap
1	Completeness	complete road continues further road continues further track continues further	missing land parcel missing section of a road missing a property
2	Non Quantitative Attribute Completeness	unknown type of barrier add name add type check name	add address error in VicMap road name incorrect address parcel description incorrect
3	Quantitative Attribute Accuracy	resurvey housenumber speed limit	no_sufficient_data
4	Positional Accuracy	road inaccurate improve accuracy location approximate alignment is an estimate	parcel shape is incorrect problem with road alignment move address point in map boundary incorrect in VicMap

data quality issues in OSM. While the support for multi-faceted data collection is a strength of OSM, it can also contribute to data usability issues across multiple ISO quality indicators, as we discuss now.

6.1. *Completeness*

A map can never be complete, as details can be continuously altered as the world changes. Thus, completeness fully depends on purpose and level of detail at which a feature type (e.g., road network) is collected. These are typically fixed in traditional data collection, but unrestricted in OSM. The ability to collect nuanced data sets in OSM is thus responsible for a high percentage of issues reported under the Completeness data quality indicator.

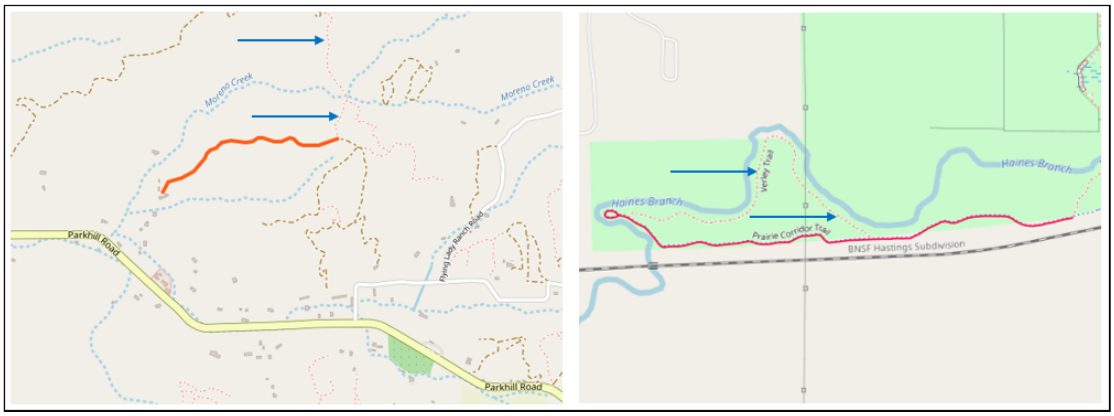

Figure 9.: FIXME - Completeness Issue. United States data set

A commonly used terms that indicates completeness issues in `FIXME` tags is *Con-*

tinue. This indicates that the mapping of a spatial feature (often line geometries) is incomplete. It acts as a note to the current OSM contributor or other future contributors to resume the mapping of this feature later. Examples of such issues from the United States data set are shown in Figure 9. Left of Figure 9 shows that after mapping a segment of pedestrian footpath (partial solid segment shown in orange), the OSM contributor has highlighted this segment as incomplete, to indicate to peer mappers the need to complete the mapping of the footpath. Similarly, in right of Figure 9, the mapper indicates that *Prairie Corridor Trail* needs to be completed.

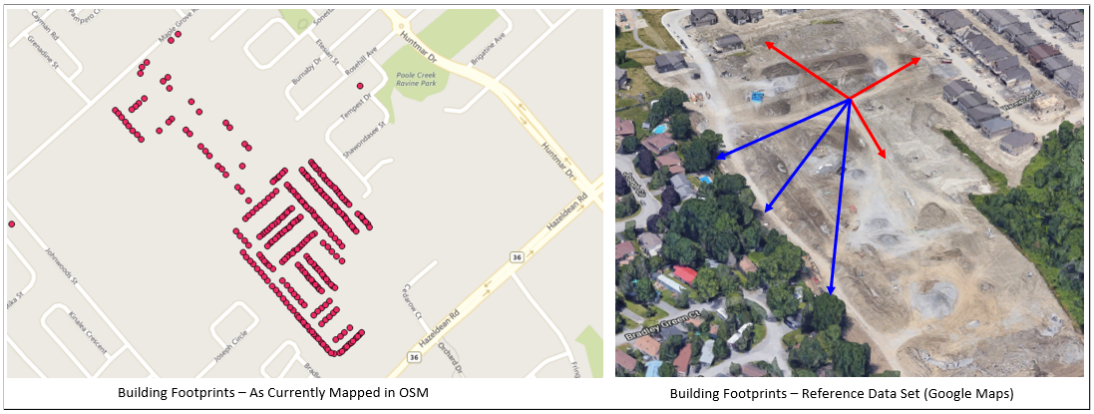

Figure 10.: Topic : Completeness (Building Footprint) USA data set

The term *'footprint'* has also been commonly used by OSM contributors to represent features, whose future changes can be assessed with a reasonable degree of confidence. In Figure 10, a group of buildings in a given neighbourhood has been represented as approximate point geometries. This can happen when the spatial feature is missing in a reference source (such as a satellite imagery) when creating the map features, but the current mapper is confident of its presence in the future (e.g., houses coming up in a new housing estate on a currently vacant plot, see Figure 10 right, indicated with red arrow). An approximate representation is given initially (group of red points, Figure 10 left), with a note to revisit once the data is in the reference source. An approximate representation (group of points as shown in Figure 10) can also occur when the reference ground truth (e.g., satellite imagery) is not legible, for instance due to cloud cover (shown with blue arrows in Figure 10 right).

6.2. *Positional accuracy*

From our results, prominent terms used by OSM contributors to indicate *Positional Accuracy* issues are *location, approximate, verify, position, placement*. These terms are used across the study areas to primarily target issues with point features. These issues could be related to the reference source (such as bad satellite imagery) or the mapping style of the contributor, irrespective of the community recommendations. For example, in Figure 11, the mapper indicates approximate point positions for homes recently constructed in a new residential estate, with a note mentioning that the positions are approximate and need to be possibly represented as polygon geometries (see (Maguire and Tomko 2017) and OSM Wiki guidelines[15]). A lack of well-defined quality assurance

[15]https://wiki.openstreetmap.org/wiki/Main_Page

process in OSM is responsible for a high occurrence of positional accuracy issues in our results, and somehow contradicts the optimistic findings of recent research studies.

Figure 11.: FIXME - Positional Accuracy Issue - USA data set

6.3. *Non quantitative attribute completeness*

Prominent terms that indicate issues with the completeness of attributes are *name, classification*, and *restrictions*. Notably in this category, in the USA data set the term *name* is predominantly associated with point features (e.g, requests for completing building attributes), while it is more frequently used to indicate issues with linear features in the Australian data set. Figure 12, left, shows an example where an OSM contributor indicates that the street networks (in red) need the name and the classification to be re-evaluated (near Griffith, NSW, Australia).

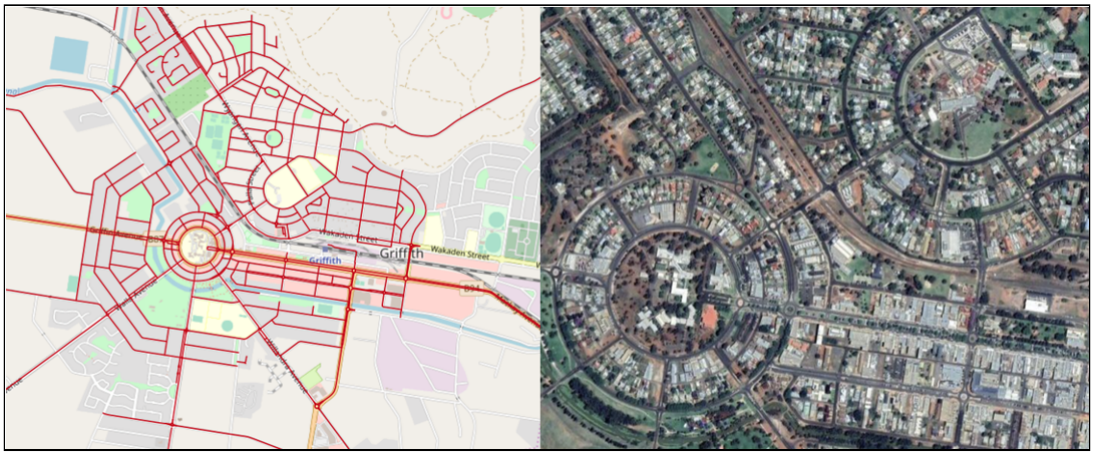

Figure 12.: FIXME - Non Quantitative Attribute Completeness - Australia data set

6.4. *Quantitative attribute accuracy*

Quantitative Attribute Accuracy (QAA) is frequently associated with building features in the USA data set. The term *'housenumber'* of building features occurs prominently

in descriptions of issues, primarily indicating buildings that have missing house numbers in address keys (*key:addr*). For road network elements, the most frequent term is *'speed limit'*, indicating missing maximum speed (*key:maxspeed*) for road segments and also to indicate issues with speed limit signage at street intersections (Figure 13).

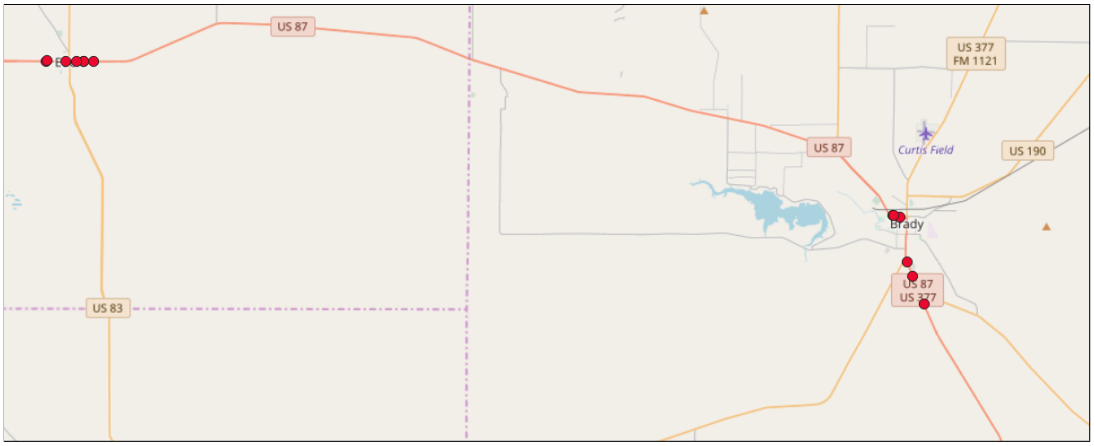

Figure 13.: Topic : Quantitative Attribute Accuracy - USA data set

7. Limitations of the study

While providing novel insights about the distribution of OSM data quality issues, the results of our study have certain limitations. First, while there are over 1.5 million `FIXME` issues documented, the study exclusively analyzes issues captured in English. Hence, the current error classification should not be taken as being representative of the entire OSM data set as a whole. In addition, the study does not explore if there is a relation between the quality of reference data sources[16] from which OSM is continually mapped and the errors reported in `FIXME` tag. Moreover, the vastly different styles of documenting map issues by contributors (20% of the reported issues are not attributable to an ISO quality indicators), could also influence the distribution.

While many `FIXME` issues can be discussed with a specific ISO spatial quality indicator, Figure 14 illustrates a significant challenge stemming from the practices of OSM mappers. Consider the point entities (Figure 14, left) that have a `FIXME` tag mentioning their *housenumber* attribute missing. These point features also represent approximate positions of existing homes (to be later mapped as polygons, as per community guidelines). Hence, while the OSM contributor has explicitly mentioned the issue with respect to an address attribute, the issue is equally a Positional Accuracy issue. Thus, a subset of the issues may relate to multiple quality concerns, although the OSM contributors highlighted only one area. In our experiments, the L-LDA algorithm has been trained only based on the category documented by the OSM contributor (here, a Quantitative Attribute Accuracy issue). While the capability of L-LDA to model multi-topic scenarios can address some of these scenarios, the impediment of curating the training data set (there is no simple way to ascertain if the documented issues represent one or more categories, purely based on their description) would make this a

[16] http://wiki.openstreetmap.org/wiki/Potential_Datasources

significant challenge. The quality of the L-LDA results are fundamentally conditioned on the quality of the manually labelled training dataset. A production model applicable globally would require a comprehensive multi-annotator effort, with stringently monitored inter-annotator agreement and controlled for training dataset bias (here, demonstrated by controlling across regional subsets).

Figure 14.: FIXME - Multiple Quality Indicators - USA data set

8. Conclusion and Future Work

Over the last two decades, the growth and adoption of open source geospatial data such as OSM, has been preeminent not only in its adoption across diverse domains, but also driving numerous open source geospatial software applications (Mobasheri *et al.* 2020). Given these actualities, analyzing OSM data quality not only continues to be a focus area for research (Jacobs and Mitchell 2020), but also serves as an antecedent to research focusing on improving OSM data quality (Chittor Sundaram *et al.* 2020). Our study presented a new intrinsic approach for classifying errors documented by OSM contributors aligned with the ISO 19157 framework. The study present insights into commonly observed data quality documenting practices of OSM contributors. Our method is not dependent on data external to OSM – typically difficult to access and expensive to procure. Instead, it relies on the hidden knowledge in one specific OSM feature tag `FIXME`, as a key indicator for the classification. While individuals, governments and commercial companies have already begun putting OSM data to use, this study should help in understanding key areas that warrant further attention towards improving OSM data quality.

An analysis of the underlying reasons for the issues documented and their correction is beyond the scope of this paper. Erroneous map data lead to adverse consequences[17] [18]. Understanding the different types of map errors is an important first step towards formulating nuanced strategies for their rectification through automated means. By analyzing large unstructured text corpora with latent knowledge about OSM data quality issues using topic models for two prominent geographical regions of the world,

[17]https://www.propublica.org/article/using-outdated-data-fema-is-wrongly-placing-homeowners-in-flood-zones

[18]http://www.news.com.au/technology/online/relying-on-google-maps-got-this-woman-stranded-in-the-grand-canyon-for-five-days/news-story/5c0e6b404f6d24ecab7bd857ca12c722

we exposed salient areas of OSM data quality requiring attention. We show that these differ to the issues documented in authoritative data, where trained users report issues in a less ambiguous manner. Our research thereby contributes to the understanding of major map data issues, and informs approaches to their rectification, thus enabling enhanced VGI data quality.

Our research and the current approach highlights the potential to facilitate future work on more comprehensive knowledge discovery initiatives, using the rich content and vernacular miscellany of OSM FIXME issues, including those not captured in English. The distributions of issues may differ based on the underlying idiosyncrasies of the dataset, the priorities of local mappers, and the cultural and linguistic environment. The distribution of FIXME issues may well vary by geographical areas, size of the mapping community, and their core mapping interests. An analysis of user tagging from under-represented regions in OSM (e.g., Africa[19]) may provide new insights into OSM data quality issues. Furthermore, issues reported in different languages may manifest highly distinct patterns. The model would need to be retrained on new corpora. Also, further analysis of issues not currently associated to an ISO quality indicator enables the discovery of contributor paradigms or new categories of quality issues closely tied to VGI data. These un-categorized issues (*common_topic*) often represent reports that are better interpreted in their semantic context (Section 5.3). Analyzing and addressing these un-categorized issues may also inform the development of new OSM quality assurance tools and improve the data quality.

9. Acknowledgements

This research was supported by a grant from Australian Research Council (ARC DP170100153). We also wish to extend our thanks to the reviewers for their valuable comments, that helped in improving this research article.

10. Data availability statement

The data and code for the algorithm supporting the findings of this study are available at the following link - `https://figshare.com/s/d22df058cb18c0eef245`

References

Adams, B. and McKenzie, G., 2013. *Inferring thematic places from spatially referenced natural language descriptions*. Dordrecht: Springer Netherlands, 201–221. Available from: `https://doi.org/10.1007/978-94-007-4587-2_12`.

Anderson, J., 2016. OpenStreetMap Contribution Analysis - A research collaboration with Mapbox. Available from: `http://mapbox.github.io/osm-analysis-collab/#about`.

Andrienko, G.L., *et al.*, 2013. Thematic Patterns in Georeferenced Tweets through Space-Time Visual Analytics. *Computing in Science and Engineering*, 15 (3), 72–82. Available from: `http://dblp.uni-trier.de/db/journals/cse/cse15.html#AndrienkoABEFJT13`.

Arlot, S. and Celisse, A., 2010. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4, 40–79.

[19] http://resultmaps.neis-one.org/unmapped#2/23.7/3.5

Arsanjani, J.J., *et al.*, 2015. The emergence and evolution of openstreetmap: a cellular automata approach. *International Journal of Digital Earth*, 8 (1), 76–90. Available from: https://doi.org/10.1080/17538947.2013.847125.

Auer, M., *et al.*, 2018. Towards Using the Potential of OpenStreetMap History for Disaster Activation Monitoring. *In*: *15th International Conference on Information Systems for Crisis Response and Management*. 317–325.

Ballatore, A. and Bertolotto, M., 2011. Semantically enriching vgi in support of implicit feedback analysis. *In*: K. Tanaka, P. Fröhlich and K.S. Kim, eds. *Web and Wireless Geographical Information Systems*, Berlin, Heidelberg. Springer Berlin Heidelberg, 78–93.

Ballatore, A., Bertolotto, M., and Wilson, D.C., 2013. Geographic knowledge extraction and semantic similarity in openstreetmap. *Knowledge and Information Systems*, 37 (1), 61–81. Available from: https://doi.org/10.1007/s10115-012-0571-0.

Barron, C., Neis, P., and Zipf, A., 2013. A comprehensive framework for intrinsic openstreetmap quality analysis. *Transactions in GIS*, 18 (6), 877–895. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/tgis.12073.

Basole, R.C., Seuss, C.D., and Rouse, W.B., 2013. It innovation adoption by enterprises: Knowledge discovery through text analytics. *Decision Support Systems*, 54 (2), 1044 – 1054. Available from: http://www.sciencedirect.com/science/article/pii/S0167923612002849.

Blatt, A.J., 2015. The Benefits and Risks of Volunteered Geographic Information. *Journal of Map & Geography Libraries*, 11 (1), 99–104. Available from: https://doi.org/10.1080/15420353.2015.1009609.

Blei, D.M. and Lafferty, J.D., 2007. A correlated topic model of Science. *Annals of Applied Statistics*, 1 (1), 17–35. Available from: https://doi.org/10.1214/07-AOAS114.

Blei, D.M., Ng, A.Y., and Jordan, M.I., 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3, 993–1022. Available from: http://dl.acm.org/citation.cfm?id=944919.944937.

Boin, A.T. and Hunter, G.J., 2008. *What Communicates Quality to the Spatial Data Consumer?* CRC Press, 140–147. Available from: https://www.crcpress.com/Quality-Aspects-in-Spatial-Data-Mining/Stein-Shi-Bijker/p/book/9781420069266.

Chang, J., *et al.*, 2009. Reading tea leaves: How humans interpret topic models. *In*: *Proceedings of the 22Nd International Conference on Neural Information Processing Systems*, NIPS'09, USA. Curran Associates Inc., 288–296. Available from: http://dl.acm.org/citation.cfm?id=2984093.2984126.

Chen, C., *et al.*, 2014. Making recommendations on microblogs through topic modeling. *In*: Z. Huang, C. Liu, J. He and G. Huang, eds. *Web Information Systems Engineering – WISE 2013 Workshops*, Berlin, Heidelberg. Springer, 252–265.

Chittor Sundaram, R., *et al.*, 2020. Harnessing spatio-temporal patterns in data for nominal attribute imputation. *Transactions in GIS*, 24 (4), 1001–1032. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/tgis.12617.

Ciepłuch, B., *et al.*, 2010. Comparison of the accuracy of openstreetmap for ireland with google maps and bing maps. *Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resuorces and Enviromental Sciences 20-23rd July 2010*, 337–340. Available from: http://eprints.maynoothuniversity.ie/2476/.

Corcoran, P., Mooney, P., and Bertolotto, M., 2013. Analysing the growth of openstreetmap networks. *Spatial Statistics*, 3, 21 – 32. Available from: http://www.sciencedirect.com/science/article/pii/S2211675313000031.

Corcoran, P., Mooney, P., and Winstanley, A.C., 2010. Topological consistent generalization of openstreetmap. *In*: *GISRUK 2010: GIS Research UK 18th Annual Conference*. 353–358. Research presented in this paper was part-funded by a Strategic Research Cluster grant (07/SRC/I1168) from Science Foundation Ireland under the National Development Plan., Available from: http://eprints.maynoothuniversity.ie/4918/.

Davidovic, N., *et al.*, 2016. Tagging in Volunteered Geographic Information: An Analysis of

Tagging Practices for Cities and Urban Regions in OpenStreetMap. *ISPRS International Journal of Geo-Information*, 5 (12). Available from: http://www.mdpi.com/2220-9964/5/12/232.

Ertl, T., *et al.*, 2012. Spatiotemporal Social Media Analytics for Abnormal Event Detection and Examination Using Seasonal-trend Decomposition. *In*: *Proceedings of the 2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, VAST '12, Washington, DC, USA. IEEE Computer Society, 143–152. Available from: http://dx.doi.org/10.1109/VAST.2012.6400557.

Fan, H., *et al.*, 2014. Quality assessment for building footprints data on openstreetmap. *International Journal of Geographical Information Science*, 28 (4), 700–719. Available from: https://doi.org/10.1080/13658816.2013.867495.

Fonte, C.C., *et al.*, 2017. *Assessing vgi data quality*. Ubiquity Press, 137–164. Available from: http://www.jstor.org/stable/j.ctv3t5qzc.10.

Girres, J.F. and Touya, G., 2010. Quality Assessment of the French OpenStreetMap Dataset. *Transactions in GIS*, 14 (4), 435–459. Available from: http://dx.doi.org/10.1111/j.1467-9671.2010.01203.x.

Goodchild, M.F., 2007. Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0. 2, 24–32.

Goodchild, M.F. and Li, L., 2012. Assuring the quality of volunteered geographic information. *Spatial Statistics*, 1, 110 – 120. Available from: http://www.sciencedirect.com/science/article/pii/S2211675312000097.

Griffiths, T.L. and Steyvers, M., 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101 (suppl 1), 5228–5235. Available from: https://www.pnas.org/content/101/suppl_1/5228.

Haklay, M., 2010. How Good is Volunteered Geographic Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets for London and the Rest of England. 37, 682–703.

Hashemi, P. and Ali Abbaspour, R., 2015. *Assessment of Logical Consistency in OpenStreetMap Based on the Spatial Similarity Concept*. Cham: Springer International Publishing, 19–36. Available from: https://doi.org/10.1007/978-3-319-14280-7_2.

Hawkins, S., *et al.*, 2002. Outlier detection using replicator neural networks. *In*: *Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery*, DaWaK 2000, London, UK, UK. Springer-Verlag, 170–180. Available from: http://dl.acm.org/citation.cfm?id=646111.679466.

Helbich, M. and Amelunxen, C., 2012. Comparative Spatial Analysis of Positional Accuracy of OpenStreetMap and Proprietary Geodata. *Proceedings of GI Forum*, 24–33.

Hernández, M.A. and Stolfo, S.J., 1998. Real-world data is dirty: Data cleansing and the merge/purge problem. *Data Mining and Knowledge Discovery*, 2 (1), 9–37. Available from: https://doi.org/10.1023/A:1009761603038.

Hofmann, T., 1999. Probabilistic Latent Semantic Indexing. *In*: *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 99, New York, NY, USA. Association for Computing Machinery, 5057. Available from: https://doi.org/10.1145/312624.312649.

Hong, J., 2019. Implementation of L-LDA algorithm with Python. Available from: https://github.com/JoeZJH/Labeled-LDA-Python.

Hong, L. and Davison, B.D., 2010. Empirical study of topic modeling in twitter. *In*: *Proceedings of the First Workshop on Social Media Analytics*, SOMA 2010, New York, NY, USA. Association for Computing Machinery, 8088. Available from: https://doi.org/10.1145/1964858.1964870.

ISO, 2013. ISO 19157:2013: Geographic information – Data quality. Available from: https://www.iso.org/standard/32575.html.

Jackson, S.P., *et al.*, 2013. Assessing completeness and spatial error of features in volunteered geographic information. *ISPRS International Journal of Geo-Information*, 2 (2), 507–530. Available from: http://www.mdpi.com/2220-9964/2/2/507.

Jacobs, K.T. and Mitchell, S.W., 2020. OpenStreetMap quality assessment using unsupervised machine learning methods. *Transactions in GIS*, 24 (5), 1280–1298. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/tgis.12680.

Jin, O., *et al.*, 2011. Transferring Topical Knowledge from Auxiliary Long Texts for Short Text Clustering. *In*: *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM 2011, New York, NY, USA. Association for Computing Machinery, 775784. Available from: https://doi.org/10.1145/2063576.2063689.

Ju, Y., *et al.*, 2016. Things and Strings: Improving Place Name Disambiguation from Short Texts by Combining Entity Co-Occurrence with Topic Modeling. *In*: E. Blomqvist, P. Ciancarini, F. Poggi and F. Vitali, eds. *Knowledge Engineering and Knowledge Management*, Cham, Switzerland. Springer International Publishing, 353–367.

Kataria, S. and Agarwal, A., 2015. Supervised Topic Models for Microblog Classification. *In*: *2015 IEEE International Conference on Data Mining*, Nov. IEEE Press, 793–798.

Keßler, C. and de Groot, R.T.A., 2013. *Trust as a proxy measure for the quality of volunteered geographic information in the case of openstreetmap*. Springer International Publishing, 21–37. Available from: https://doi.org/10.1007/978-3-319-00615-4_2.

Lafferty, J.D. and Blei, D.M., 2009. Topic Models. *In*: A.N. Srivastava and M. Sahami, eds. *Text mining - classification, clustering and applications*. Boca Raton, FL, USA: Chapman and Hall/ CRC Press, Ch. 10, 71–94.

Lansley, G. and Longley, P.A., 2016. The geography of twitter topics in london. *Computers, Environment and Urban Systems*, 58, 85 – 96. Available from: http://www.sciencedirect.com/science/article/pii/S0198971516300394.

Lau, J.H., *et al.*, 2010. Best topic word selection for topic labelling. *In*: *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, Stroudsburg, PA, USA. Association for Computational Linguistics, 605–613. Available from: http://dl.acm.org/citation.cfm?id=1944566.1944635.

Lingad, J., Karimi, S., and Yin, J., 2013. Location Extraction from Disaster-Related Microblogs. *In*: *Proceedings of the 22nd International Conference on World Wide Web*, New York, NY, USA. Association for Computing Machinery, 10171020. Available from: https://doi.org/10.1145/2487788.2488108.

Ludwig, I., Voss, A., and Krause-Traudes, M., 2011. A comparison of the street networks of Navteq and OSM in Germany. *In*: S. Geertman, W. Reinhardt and F. Toppen, eds. *Advancing geoinformation science for a changing world*. Berlin, Heidelberg: Springer, 65–84. Available from: https://doi.org/10.1007/978-3-642-19789-5_4.

Maguire, S. and Tomko, M., 2017. Ripe for the picking? dataset maturity assessment based on temporal dynamics of feature definitions. *International Journal of Geographical Information Science*, 31 (7), 1334–1358.

Majic, I., *et al.*, 2019. Discovery of topological constraints on spatial object classes using a refined topological model. *Journal of Spatial Information Science*.

Majic, I., Winter, S., and Tomko, M., 2017. Finding Equivalent Keys in Openstreetmap: Semantic Similarity Computation Based on Extensional Definitions. *In*: *Proceedings of the 1st Workshop on Artificial Intelligence and Deep Learning for Geographic Knowledge Discovery*, GeoAI '17, New York, NY, USA. ACM, 24–32. Available from: http://doi.acm.org/10.1145/3149808.3149813.

Mobasheri, A., *et al.*, 2020. Highlighting recent trends in open source geospatial science and software. *Transactions in GIS*, 24 (5), 1141–1146. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/tgis.12703.

Mocnik, F.B., *et al.*, 2018. A grounding-based ontology of data quality measures. *Journal of Spatial Information Science*, 16, 1 – 25. Available from: https://www.josis.org/index.php/josis/article/viewArticle/360.

Mooney, P. and Corcoran, P., 2012. The annotation process in openstreetmap. *Transactions in GIS*, 16 (4), 561–579. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9671.2012.01306.x.

Neis, P., Zielstra, D., and Zipf, A., 2012. The street network evolution of crowdsourced maps:

OpenStreetMap in Germany 2007–2011. *Future Internet*, 4 (1), 1–21. Available from: http://www.mdpi.com/1999-5903/4/1/1.

Papapesios, N., *et al.*, 2018. Exploring the use of crowdsourced geographic information in defence: challenges and opportunities. *Journal of Geographical Systems*, 133 – 160. Available from: https://link.springer.com/article/10.1007/s10109-018-0282-5.

Rahimi, M.M., *et al.*, 2020. Service quality monitoring in confined spaces through mining Twitter data. *Journal of Spatial Information Science*.

Ramage, D., Dumais, S., and Liebling, D., 2010. Characterizing microblogs with topic models. *In*: *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, May, Menlo Park, CA, USA. AAAI Press.

Ramage, D., *et al.*, 2009. Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-labeled Corpora. *In*: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1*, EMNLP '09, Stroudsburg, PA, USA. Association for Computational Linguistics, 248–256. Available from: http://dl.acm.org/citation.cfm?id=1699510.1699543.

Ramage, D., Manning, C.D., and Dumais, S., 2011. Partially labeled topic models for interpretable text mining. *In*: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 2011, New York, NY, USA. ACM, 457–465. Available from: http://doi.acm.org/10.1145/2020408.2020481.

Ritter, A., *et al.*, 2011. Named Entity Recognition in Tweets: An Experimental Study. *In*: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP 2011, USA. Association for Computational Linguistics, 15241534.

Salk, C.F., *et al.*, 2016. Assessing quality of volunteer crowdsourcing contributions: lessons from the Cropland Capture game. *International Journal of Digital Earth*, 9 (4), 410–426. Available from: https://doi.org/10.1080/17538947.2015.1039609.

Senaratne, H., *et al.*, 2017. A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science*, 31 (1), 139–167. Available from: https://doi.org/10.1080/13658816.2016.1189556.

Simoudis, E., Livezey, B., and Kerber, R., 1995. Using recon for data cleaning. *In*: *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, KDD'95. AAAI Press, 282–287. Available from: http://dl.acm.org/citation.cfm?id=3001335.3001382.

Stevens, K., *et al.*, 2012. Exploring topic coherence over many models and many topics. *In*: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, Stroudsburg, PA, USA. Association for Computational Linguistics, 952–961. Available from: http://dl.acm.org/citation.cfm?id=2390948.2391052.

Steyvers, M., *et al.*, 2004. Probabilistic Author-topic Models for Information Discovery. *In*: *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, New York, NY, USA. ACM, 306–315. Available from: http://doi.acm.org/10.1145/1014052.1014087.

Vandecasteele, A. and Devillers, R., 2015. *Improving volunteered geographic information quality using a tag recommender system: The case of openstreetmap*. Cham: Springer International Publishing, 59–80. Available from: https://doi.org/10.1007/978-3-319-14280-7_4.

Will, J., 2014. Development of an automated matching algorithm to assess the quality of the openstreetmap road network : a case study in gteborg, sweden. Student Paper, Available from: https://lup.lub.lu.se/student-papers/search/publication/4464336.

Williams, J., 1997. Tools for traveling data. *DBMS*, 10 (7), 69–76. Available from: http://dl.acm.org/citation.cfm?id=258019.258027.

Zielstra, D. and Hochmair, H.H., 2011. Comparative Study of Pedestrian Accessibility to Transit Stations Using Free and Proprietary Network Data. *Transportation Research Record*, 2217 (1), 145–152. Available from: https://doi.org/10.3141/2217-18.

Zielstra, D. and Zipf, A., 2010. A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany.

Zipf, A., *et al.*, 2016. *Crowdsourcing for individual needs the case of routing and navigation for mobility-impaired persons.* Ubiquity Press, London, 325–338.

11. Appendix

11.1. *LDA and L-LDA notation / terminology*

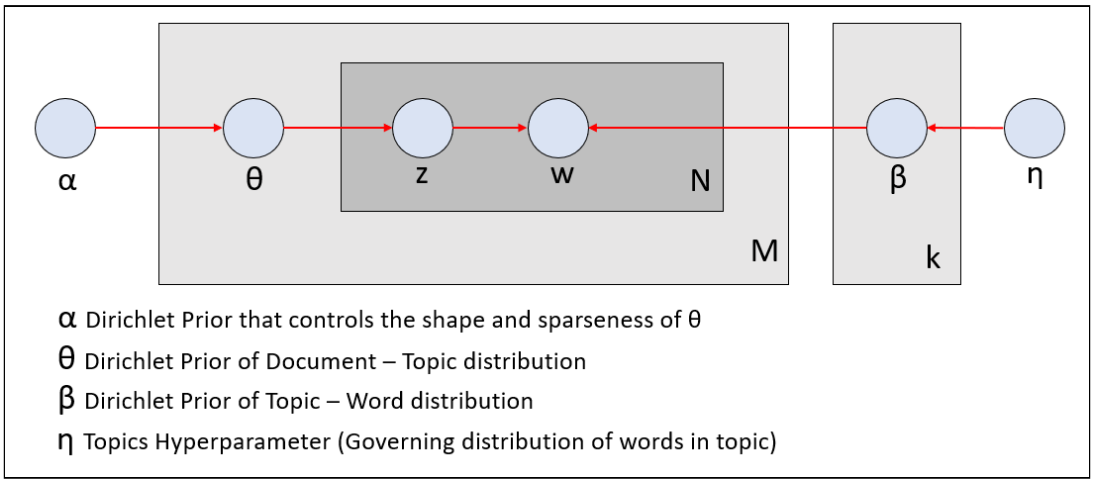

Figure 15.: Graphical Model of Latent Dirichlet Allocation

Formally, the objective of LDA (Blei *et al.* 2003) can be summarized as:

$$P(\theta_{1:M}, \mathbf{z}_{1:M}, \beta_{1:\mathbf{k}}|D, \alpha_{1:M}, \eta_{1:\mathbf{k}})$$

Given D, (the document corpus from a collection of M documents, with each document having N words and each word having been generated by one topic of out the given $\mathbf{k}$ topics), we determine the intractable joint posterior probabilities (that reveals the hidden topic structure) of the following:

- θ - The document topic distribution
- $\mathbf{z}$ - N topics for each document
- β - The word topic distribution, using the parameters
- α - The parameter vector for each document
- η - The parameter vector for each topic

To better understand the algorithms (discussed in Section 3.2) around which the findings are presented in this research article, a formal introduction of the notations and terminologies used in the algorithms, is presented herewith:

(1) **Vocabulary** - Set of words from a given language. V is the size of the vocabulary (as number of words);
(2) **Word** - A single distinct text element associated with a meaning, and member of the vocabulary $\{1, \ldots, V\}$. Words are represented as one hot encoded unit vectors, such that the v^{th} word in the vocabulary is represented by a unit vector w (such that $w^v = 1$ and $w^u = 0, u \neq v$)
(3) **Document** (d) - Represents a sequence of words. $\mathbf{w}_d = \{w_1, w_2, w_3, \ldots, w_{Nd}\}$. N is the total number of words in d.

(4) **Document Corpus** (D) - Represents a collection of documents ($D = \{\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3, \ldots, \mathbf{w}_M\}$, where M is the total number of documents in D.
(5) **Topic** - A subject matter discussed within a document. In LDA and L-LDA, a topic is modelled as a distribution of words.
(6) **k** - Total finite count of topics that a document can belong to.
(7) **z** - Represents a topic from the given set of **k** topics.
(8) **label** - An indicator of one or more themes discussed within a document (e.g., QAA for a `FIXME` document, having a Quantitative Attribute Accuracy issue, such as an incorrect value for the speed limit of a road segment).
(9) **K** - Total number of unique labels in the corpus. If the L-LDA algorithm will be trained with 4 topics, K can be represented as $K = \{\Omega_1, \Omega_2, \Omega_3, \Omega_4\}$.