DOI: 10.1111/tgis.12617

RESEARCH ARTICLE

Transactions in GIS WILEY

Harnessing spatio-temporal patterns in data for nominal attribute imputation

Rajesh Chittor Sundaram[1] | Elham Naghizade[1] | Renata Borovica-Gajic[2] | Martin Tomko[1]

[1]Department of Infrastructure Engineering, The University of Melbourne, Parkville, VIC 3010, Australia

[2]Department of Computing and Information Systems, The University of Melbourne, Parkville, VIC 3010, Australia

Correspondence
Rajesh Chittor Sundaram, Department of Infrastructure Engineering, The University of Melbourne, Parkville, VIC 3010, Australia.
Email: c.s@unimelb.edu.au

Funding information
Australian Research Council, Grant/Award Number: ARC DP170100153

Abstract

Missing data in Volunteered Geographic Information (VGI) are an unavoidable consequence of data collection by non-experts, guided by only vague and informal mapping guidelines. While various Missing Value Imputation (MVI) techniques have been proposed as data cleansing strategies, they have primarily targeted numerical data attributes in non-spatial databases. There remains a significant gap in methods for imputing nominal attribute values (e.g., *Street Name*) in map databases. Here, we present an imputation algorithm called the Membership Imputation Algorithm (MIA), targeting spatial databases and enabling imputation of nominal values in spatially referenced records. By targeting membership classes of spatial objects, MIA harnesses spatio-temporal characteristics of data and proposes efficient heuristics to impute the class name (i.e., a membership). Experimental results show that the proposed algorithm is able to impute the membership with high levels of accuracy (over 94%) when assigning *Street Name(s)*, across highly diverse regional contexts. MIA is effective in challenging spatial contexts such as street intersections. Our research serves as a first step in highlighting the effectiveness of spatio-temporal measures as a key driver for nominal imputation techniques.

KEYWORDS

Spatial Database, Spatial Data, Missing Value Imputation, Spatio-Temporal Proximity

1 | INTRODUCTION

Many real world data sets are *dirty* (Prasad et al., 2011). The term *dirty data* refers to data sets with issues such as missing or incorrect records or values (Simoudis, Livezey, & Kerber, 1995), non-standard representations (Williams, 1997), outliers (Hawkins, He, Williams, & Baxter, 2002), and duplicate values (Hernández & Stolfo, 1998). Even databases created with strict regulatory requirements have considerable amounts of missing data (Kurgan, Cios, Sontag, & Accurso, 2005). Volunteered Geographic Information (Goodchild, 2007; VGI) databases are created by non-experts, untrained in the formal process of map data collection and curation. VGI databases are severely impacted by dirty data, due to the mandated manual data entry by data contributors with widely varying skills, often without a local knowledge of the mapped area (i.e., remote online mappers), the absence of data collection protocols, informal quality assurance processes, and the use of instruments with varying levels of precision by on-the-ground mappers. As an example, OpenStreetMap (OSM) (https://www.openstreetmap.org.), the most prominent VGI data source, is heavily impacted by map features with incomplete attribute data (Davidovic, Mooney, Stoimenov, & Minghini, 2016). This is a general issue prominent in databases without a strict schema or data definition rules. In OSM, the free tagging system allows the contributors to use an unlimited number of attributes to describe a map feature. This free-form nature of tagging, coupled with a lack of adherence to community guidelines (https://wiki.openstreetmap.org/wiki/Tagging.), results in considerable missing data for features.

There are multiple VGI data curation projects to create a free map of the world. Wikimapia (https://www.https://wikimapia.org) is a multilingual collaborative VGI initiative and similar to OSM in areas of commonality, such as the volunteer nature of the data curation process guided by general mapping guidelines, and free access to geo-data using a public access framework. However, the data diversity and volume of data in Wikimapia is much smaller compared to OSM. For example, Wikimapia has a user community of about 2.6 million users worldwide and about 29 million places created (https://wikimapia.org/#lang=en&lat=-37.813900&lon=144.963400&z=12&m=w&show=/stats/action_stats>fstat=2&period=2&year=2019&month=9), whereas OSM has a much higher data volume at about 6,090 million data elements, contributed by about 5.69 million contributors (https://www.openstreetmap.org/stats/data_stats.html). In addition, it is reported that in spite of the benefits of a crowdsourced model, Wikimapia is prone to erroneous content, and making corrections is a challenging process (Goodchild & Glennon, 2010). Such errors have contributed to a decline in its popularity (Ballatore & Arsanjani, 2019). While other popular crowdsourced offerings such as Yandex Maps exist, there are restrictions on the use of data and the service offerings are not completely free (https://tech.yandex.com/maps/commercial/). Finally, the significant user base of the OSM contributing community, coupled with vague mapping guidelines, makes the challenge of dirty data more pronounced in VGI data sources such as OSM. In light of these observations, we focus on OSM data sets in the remainder of this article.

The rapid growth in OSM data creation and use over the last decade (Corcoran, Mooney, & Bertolotto, 2013) has raised questions about its reliability and fitness for use (Hashemi & Ali Abbaspour, 2015; Maguire & Tomko, 2017). In particular, address attributes (e.g., *Street Name*, *Street Number*, *City*, *State*, *and Postal Code*) are critical location-based service enablers. Missing values for these attributes severely impact critical location and geo-coding services (e.g., navigation systems and emergency response, public health, and crime analyses systems). Reliable Missing Value Imputation (MVI) techniques are thus fundamental enablers for location-based decision-making (Zandbergen, 2008) using VGI.

The applicability and effectiveness of imputation techniques vary with the nature of the missing data, broadly classified as Missing Completely at Random (MCAR) or Missing Not at Random (MNAR) (Rubin, 1976). From one perspective, missing attributes in OSM can be considered as MCAR, due to the nature of VGI data creation, wherein missing data results from the weak adherence of volunteer mappers to the (vague) mapping guidelines. Alternatively, missing attributes can also be perceived as MNAR, when missing data are the result of a batch import process or the creators' lack of knowledge of the local geography. Nevertheless, there is no standardized process to ascertain the category of missing data for arbitrary data sets, and a straightforward application of

MCAR or MNAR techniques is therefore difficult. Methods such as Mean Imputation (Scheffer, 2002) commonly produce biased value estimates with MCAR data and no universal methods support data imputation for MNAR data (Little, 1992; Vach, 1994).

Furthermore, while most imputation techniques are largely focused on *ordinal*, *interval*, and *ratio* values as defined by Stevens's *measurement scales* (Stevens, 1946), recent efforts have attempted to impute missing *nominal* attributes (Josse & Husson, 2016) by analyzing the correlation between attributes in a multivariate data set. Such techniques are, however, not directly applicable to map databases since: (a) nominal attributes in OSM are sparse; (b) nominal attributes are functionally dependent on spatial properties, rather than on other non-spatial attributes (e.g., the geometry of a free-standing residential building in OSM drives the nominal attribute *building=detached*, as opposed to other non-spatial attributes such as building *name* or *height*); and therefore (c) the remodeling of attribute features as multivariate data is challenging. This is despite the fact that missing nominal attributes (e.g., a *Street Name* in an address) impact a large proportion of database records (including in OSM, where a well-mapped country like Switzerland has about 1.84 million building footprints (28%, as of March 2018) without *Street Name* attribution). In this article, we hypothesize that unique spatio-temporal characteristics of spatial data (such as spatial proximity measures between objects, as well as temporal variation in spatial characteristics) can facilitate MVI in spatial data sets. Our hypothesis is rooted in Tobler's first law of geography, which states that "Everything is related to everything else, but near things are more related than distant things" (Tobler, 1970).

We formalize our problem as a membership imputation, where a set of spatial entities belong to a Membership Class (MC), and the aim is to impute the class name for those entities that are not a member of any existing class. We demonstrate this on the OSM Associated Street Relation (ASR), a MC to associate heterogeneous types of map features with a street (such as schools and residential buildings). We propose the Membership Imputation Algorithm (MIA), which imputes the nominal attributes of an OSM relation (here, the ASR membership) for any map feature, by evaluating the spatial and temporal proximity of the neighboring map features that already belong to an existing relation (here, the ASR). The cornerstone of MIA is based upon the principle of nearest neighbor analysis (Cover & Hart, 1967), where similarities in attribute values across a set of neighborhood elements in close proximity and belonging to different membership classes drives the imputation framework. This approach was found to be effective in imputing missing nominal values for map features (e.g., *Street Name* as a part of address information for a residential apartment), further discussed in Section 2.2. MIA achieves an imputation accuracy of almost 94% in general cases and 97% in rural areas, along with an accuracy of over 88% at street intersections. Moreover, MIA performs best at small spatial distances, meaning that it has a low computational cost. In this article we:

1. Propose an algorithm called MIA to impute missing nominal values in spatial databases;
2. Propose distinct heuristics based on different measures of spatial and temporal proximity, and analyze the sensitivity of each approach in our algorithm;
3. Evaluate MIA's performance on a non-relational VGI data set (OSM) across multiple distinct geographical regions.

Section 2 outlines the current research and development in the field of data cleansing and its relevance to VGI data sets. We then formulate the main problem for membership imputation using MIA in Section 3.3. This is followed by an overview of our proposed approach in Section 4, where we first introduce the various proximity measures that are available in spatial data sets, followed by a discussion on the contrivance of the algorithm for different proximity measures. We then proceed to explain the operational logic behind MIA in Section 4.1, which is then followed by an introduction to the baseline algorithms that are used to compare and contrast the effectiveness of MIA in Section 4.2. Section 5 presents the core components of our experimental evaluation such as the main experimental setup (Section 5.1), the ground truth data set (Section 5.2), and the performance of MIA against the baseline algorithms (Section 5.3.1) and across different intersection types (Section 5.3.2). Section 5.5 discusses the sensitivity analysis of the algorithm across three key dimensions of spatial buffer (Section 5.5.1), land use distribution (Section 5.5.2), and spatio-temporal metrics (5.5.3). The experimental evaluation section

concludes with a discussion of the performance of MIA across different geographical regions in the world (Section 5.5.4). We conclude our work in Section 6 and discuss potential areas for future work in Section 7.

2 | RELATED WORK

The impetus for MIA stems from two key focus areas of research: data cleansing techniques and tools in relational database management systems (RDBMS) with a focus on MVI (Section 2.1); and the assessment of VGI data quality against reference data sets (Section 2.2). Our research extends principles of MVI in relational databases to address missing attributes in a spatial (VGI) database, an area overlooked by the research community.

2.1 | RDBMS data cleansing

Figure 1 shows standardized approaches to data cleansing, as consolidated from research and error classification mechanisms in this area (Chu, Ilyas, & Paolo, 2013; Chu et al., 2015; Ilyas & Chu, 2015; Kim, Choi, Hong, Kim, & Lee, 2003) and algorithmic approaches to data repair discussed in Chu, Ilyas, Krishnan, and Wang (2016). We focus primarily on rule-based techniques for MVI.

2.1.1 | Data cleansing frameworks

A variety of tools, frameworks, and algorithms have been devised to support the usually highly manual processes of data cleansing. We now examine their applicability to spatial data in particular, to identify suitable baselines for the assessment of the proposed Membership Imputation Algorithm.

Galhardas, Florescu, Shasha, Simon, and Saita (2001) proposed a declarative data cleansing framework effective in creating standardized representations of data by duplicate merging. This is similar to the extensible frameworks AJAX (Galhardas, Florescu, Shasha, & Simon, 2000) and TAILOR (Elfeky, Verykios, & Elmagarmid, 2002) that address data cleansing through data consolidation and record linkages. Similarly, an open source entity matching and record linkage system (Christen, 2008) has been found to be effective in performing data de-duplication. Kandel, Paepcke, Hellerstein, and Heer (2011), discuss a new data cleansing solution called WRANGLER, modeled around iterative data transformations. Dallachiesa et al. (2013), discuss an extensible data cleansing platform, NADEEF, to handle

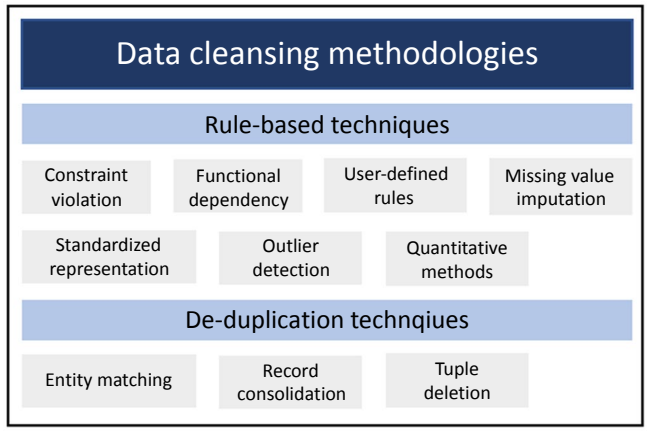


FIGURE 1 RDBMS data cleansing: An overview

the detection and repair of constraint violations in databases. In addition, various heuristic methods have been discussed to handle functional dependencies, conditional functional dependencies (Bohannon, Fan, Flaster, & Rastogi, 2005; Fan, Geerts, Jia, & Kementsietsidis, 2008) and constraint-based cleansing (Arenas, Bertossi, & Chomicki, 1999; Beskales, Ilyas, & Golab, 2010; Kolahi & Lakshmanan, 2009). Data cleansing through functional dependencies using knowledge bases has been implemented in the KATARA data cleansing system (Chu et al., 2015).

While each of the data cleansing frameworks and tools discussed above has been assessed for its generality and efficiency using various data sets, there has been no discussion of the applicability of these platforms for performing an MVI.

2.1.2 | Data imputation

Two commonly used data cleansing strategies to deal with missing data are to ignore (remove) or impute (fill) missing values (Farhangfar, Kurgan, & Pedrycz, 2007). Missing value removal discards records that have missing values and is applicable when a small amount of data is missing. MVI is routinely performed using statistics and rule-based algorithms (Cong, Fan, Geerts, Jia, & Ma, 2007; Müller & Freytag, 2005; Rahm & Do, 2000; Sung, Li, & Sun, 2002). While techniques such as regression, mean/median, hot deck and multiple imputation have been assessed for their effectiveness on epidemiological and medical databases (Engels & Diehr, 2003; Shrive, Stuart, Quan, & Ghali, 2006), alternative methodologies such as cold deck and substitution imputation techniques are found to be effective with survey databases (Lakshminarayan, Harp, & Samad, 1999; Song & Shepperd, 2007; Sarle, 1998). While the ERACER data cleansing framework handles missing values and outliers (Mayfield, Neville, & Prabhakar, 2010), the imputations are for numerical values only. Recently, machine learning algorithms have been found to have a superior imputation accuracy in addressing MVI (Farhangfar, Kurgan, & Pedrycz, 2004; Zhang, 2000).

None of these frameworks discuss their extensibility to spatial data sets. Identifying data quality rules for spatio-temporal data sets presents a challenge due to the implicit nature of spatio-temporal dependencies and constraints. This is pronounced in the case of VGI data, where the lack of a relational model makes the identification of rules linking attribute data and geometrical information difficult.

We also note that data imputation from external data sets (data integration) is not treated here. Data integration can be used as a source of data cleansing only if a second data set, of assumed higher quality, is available. Related to the scenario studied in this article, while reverse geo-coding (the process of converting geographic coordinates to addresses and place names) may be a useful mechanism to assist in nominal attribute imputations (such as missing address information for spatial entities), the data integration approaches may not generalize well, and would suffer from limitations of coverage (here, geographic) of the external data sets.

While MVI packages such as missMDA (Josse & Husson, 2016) support nominal data imputation in generic databases, representations of spatial features in VGI data sets are predominantly driven by spatial properties such as positional accuracy—how close the coordinate descriptions of spatial features are compared to their actual location in reality (such as the position of a building in a university campus represented in the VGI data set, as compared to its actual position on the "ground") – and much less based on the attribute accuracy of the same feature (i.e., how thoroughly are the attributes such as the *Height* and *Street Name* of this building represented in the data set). In other words, attribute features are only used to impart additional information about a spatial feature, and they themselves are not drivers (independent variables) that determine the representation of this feature (dependent variable) in VGI data sets. Therefore, the approaches highlighted in such MVI packages are not suitable for performing a nominal imputation with spatial attributes as the driving variables. Furthermore, missing value removal is not appropriate for spatial data sets because a wealth of information is conveyed by the attributes and geometry of map features. Moreover, these features could be a part of a larger relation (e.g., a building is a part of a bigger university campus) and removing data could invalidate the structure of these logical groupings. Yet, space presents useful heuristics that can inform and drive MVI tasks. MIA exploits the spatial and temporal

proximity measures to perform a nominal attribute imputation, without having an explicit dependency on external frameworks that come with additional financial costs. This is a key area of contribution from our research and is discussed further in Sections 3–6.

2.2 | VGI data quality indicators

VGI data quality is expressed in terms of quality indicators and measures (Vyron & Andriani, 2015). While quality measures are usually based on International Standardization Organization (ISO) guidelines for measuring the discrepancy between data and ground truth (usually comparing VGI data with authoritative data sets such as from governmental mapping agencies), quality indicators are intrinsic and cannot be directly measured against ground truth (such as VGI contributor expertise). ISO 19157 (ISO, 2013) quantitative measures assess the quality on multiple dimensions such as positional accuracy, completeness, topological consistency, and semantic accuracy of spatial data. Qualitative indicators of spatial data discuss the purpose, usage, and lineage of data sets (Van Oort & Bregt, 2005).

Positional Accuracy (PA) is defined as the closeness of the coordinate values reported for a map feature to values being accepted as true. Extensive research in spatial sciences has been devoted to positional accuracy. In VGI (and hence OSM), this is usually assessed by comparing VGI data with official reference data sets (Al-Bakri & Fairbairn, 2012; Fan, Zipf, Fu, & Neis, 2014; Haklay, 2010; Jackson et al., 2013; Neis, Zielstra, & Zipf, 2012).

Completeness (CO) measures the rate of omissions/commissions of features in the data set. Completeness is one of the primary foci of spatial data quality measurement, and one of the first measures to be studied in OSM data quality research, primarily with a focus on road networks (Ludwig, Voss, & Krause-Traudes, 2011; Zielstra & Hochmair, 2011; Zielstra & Zipf, 2010). Yet the measurement of completeness requires a reference data set and generally does not apply to the rate of missing attributes, but only of entire features. With MIA, we only update the attributes of existing features. As such, no new features (i.e., buildings) are generated, and thus the completeness of the data set is not altered.

Topological Consistency (TC) relates to the logical consistency of spatial data, defined as the correctness of the explicitly encoded topological relationships between map features. While ISO 19157 explicitly focuses on point-curve connections, self-intersections, slivers and self-overlaps of geometries, topological consistency (a constraint satisfaction problem) (Mackworth, 1977) may relate to a broader set of issues, for example in the relationships between road features (Barron, Neis, & Zipf, 2013; Corcoran, Mooney, & Winstanley, 2010; Girres & Touya, 2010; Neis et al., 2012; Will, 2014) or features in complexes scenes (Lewis, Dube, & Egenhofer, 2013), or between vague features (Du, Qin, Wang, & Ma, 2008).

Semantic Accuracy (SA) pertains to the accuracy of attribute values attached to map features. However, this does not cover semantic completeness, which is of primary interest here. Originating from annotation activities of OSM contributors, OSM map features are annotated with weak adherence to mapping guidelines. Girres and Touya (2010), Mooney, Corcoran, and Winstanley (2010) and Mooney and Corcoran (2012) note this as a pressing issue. This is similar to the findings of Davidovic et al. (2016), who report the overall compliance for OSM tagging to be generally average to poor, based on a study on tagging practices in 40 cities worldwide. This is supported in Neis et al. (2012), where it is observed that approximately 16% of the streets did not have a *Street Name* in the OSM Germany data set. A similar observation is discussed in Girres and Touya (2010) for the OSM French data set. While current research shows that VGI data are sufficiently accurate and consistent—often comparable or even exceeding the Positional Accuracy and Topological Consistency of proprietary reference data sets, attribute completeness in VGI data sets is an area that warrants further attention. This is supported by Neis et al. (2012), who explicitly note that OSM data can substantially increase their usability if missing attribute information can be addressed. Furthermore, Ludwig et al. (2011) highlight the issue of incompleteness of OSM attributes that could otherwise help in solving advanced spatial analysis problems.

Missing data have a detrimental impact on decision-making. They affect data set usability and, more significantly, impact the conclusions drawn from the data set (Graham, 2009). In the remainder of this article, we discuss how spatio-temporal measures associated with VGI data can be effectively utilized to implement MVI strategies

for nominal attributes. We illustrate our nominal imputation algorithm MIA, with a specific case study of missing street names (key: "*addr:street*") in OSM. We believe this method can be generalized to other types of nominal value imputation in (non-relational) spatial databases such as Infrastructure Networks.

3 | PROBLEM FORMULATION

3.1 | OSM concepts

Three fundamental objects of the OSM data model are nodes, ways, and relations (https://wiki.openstreetmap.org/wiki/Elements). Objects can have tags associated with them as key–value pairs to describe the attributes of the object (such as the type of a restaurant). An object must have a minimum of one tag, but there is no upper limit. The tagging practice in OSM is based on an informal rule book and guidelines, as highlighted in the OSM Map Features wiki (http://wiki.openstreetmap.org/wiki/Map_Features). One such group of tags in OSM is collectively referred to as *key:addr* keys. These tags are used to provide address information for buildings and facilities that are mapped in OSM as per community guidelines (https://wiki.openstreetmap.org/wiki/Addresses). An important element of *key:addr* keys is the *Street Name*, represented using the key–value pair *addr:street=**.

Relation: This is a data element in OSM that describes logical and geographical relationships between map features in geographic proximity. It consists of one or more tags and an ordered list of one or more nodes, ways and (other) relations as members. The aim of relations is thus to represent tightly associated and spatially clustered items, by grouping their members in a membership class.

Associated Street Relation (ASR) (https://wiki.openstreetmap.org/wiki/Relation:associatedStreet): This is one of the most used relations in OSM. It provides an explicit link between map features (indicated using the tag *addr:housenumber*) and the street to which it belongs (tag *addr:street*), based on geographic proximity. Figure 2 shows ASR elements for a street in Switzerland.

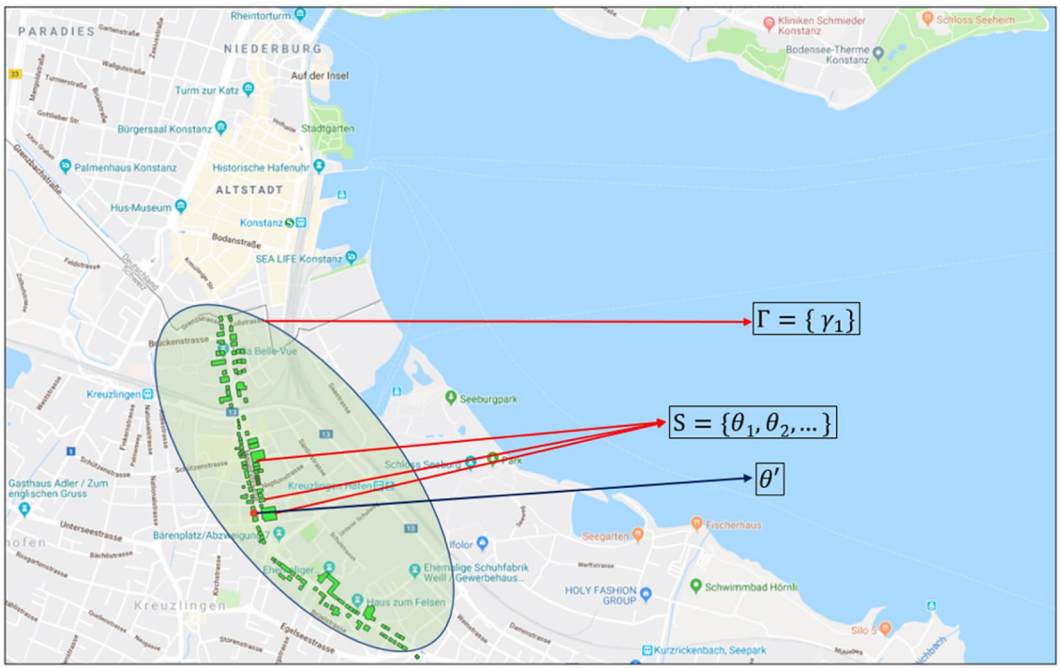


FIGURE 2 An example of an ASR and its related elements

Changeset (https://wiki.openstreetmap.org/wiki/Changeset): Represents a collective set of operations that a single OSM user does on map features (e.g., additions, tagging updates, and deletions), usually constrained to a geographic proximity, over a short period of time, (single edit session). A changeset is thus a suitable construct to identify and analyze relationships between map features, with an assumed logical interdependence.

Although we formalize and evaluate MIA using OSM data, our approach can be utilized for imputations in other spatio-temporal data sets, if either the above-mentioned concepts exist in that data set or the data can be transformed into a similar representation.

3.2 | Spatial analysis concepts

3.2.1 | Spatio-temporal data

This is a term used here to denote information that identifies the geographic location and extent of features, anchored to a state at a point in time. While often applied to moving objects, in the context of MIA, the term only refers to the temporally tagged states of mapped spatial features. In particular, we are interested in features mapped at about the same time (or within short time intervals), as further discussed in Section 4.

3.2.2 | Spatial buffer

A buffer delimits a neighborhood area (buffer zone) that is within a specified distance of a real-world map feature (a candidate for membership imputation, also referred to as a *seed*, as explained in Section 3.3). In the context of MIA, a buffer zone helps the algorithm to filter down the neighborhood elements from diverse ASR membership classes, to only those that exist within a specified distance from the *seed*. Differing buffer zones result in varying allotments of neighborhood elements from different ASR membership classes to be considered for the imputation. The buffer zone varies between 0 and 1,000 m in our experiments.

3.2.3 | Distance proximity measures

Euclidean distance (shortest straight-line distance between two geometries) is employed as the spatial distance measure throughout this article. It is the most straightforward, generic application of Tobler's law (see Section 4), without additional assumptions or data dependencies. Network distance (distance along street segments such as Manhattan distance) are natural refinements, but rely heavily on underpinning data and their quality (such as the existence of attributes, informing about the directedness of the street network and proper geometry nodding). Section 2.2 discusses the completeness of OSM with respect to street networks being generally very good, but with poor non-quantitative attribute completeness. In addition, OSM data have generally poor quality for routing, with the exception of a few regions, primarily in Europe (Neis et al., 2012; Schmitz, Pascal, & Alexander, 2008). We therefore aimed to design MIA with the least amount of data reliance as possible. Users of MIA are free to refine the algorithm whenever the data or geographical context allows it.

3.3 | Problem statement

We aim to impute the membership of a map feature (denoted by the seed θ') to an ASR membership class, based on the spatio-temporal proximity between the seed and its neighbors (thereby deriving its *Street Name*). Having a

set of ASRs as our membership classes, $\Gamma = \{\gamma_1, \gamma_2, \gamma_3, \ldots, \gamma_p\}$, we assume the map features are decomposed into two sets: map features with known ASRs, $S = \{\theta_1, \theta_2, \theta_3, \ldots, \theta_n\}$, each being mapped to a single relation in Γ (i.e. $\forall\theta_i \in S, \exists\gamma_j \in \Gamma$, where $\theta_i \rightarrow \gamma_j$), and seeds, $S' = \{\theta'_1, \theta'_2, \theta'_3, \ldots, \theta'_m\}$, such that $S' \cap S = \emptyset$. Note that there can be a subset in S that is mapped to a unique relation (membership class) in Γ through a mapping function f, that is, $f: \hat{S} \rightarrow \hat{\Gamma}$, where $\hat{S} \subset S$, $|\hat{S}| \geq 1$, and $\hat{\Gamma} \subset \Gamma$, $|\hat{\Gamma}| = 1$. We also assume that S' is spatially and/or temporally dependent on S, hence, $\forall\theta'_i$, $1 \leq i \leq m$, we can find a set Φ_i where $\Phi_i \subset S$ and Φ_i is in the *neighborhood* of θ'_i. Our aim is then to impute the missing ASR for elements in S', using each element's neighborhood set, that is, having the Φ_i, and their corresponding mapping in ASR, we aim to find the best relation in Γ and map the seeds to them. In Section 4 we first discuss different heuristics to find the neighborhood set for each element in S', and then detail our algorithm to find the best candidate in Γ to impute the missing ASR for the seeds.

Elements mentioned in the problem statement are shown in Figure 2. An ASR from Switzerland is shown in the blue oval. Map features (green) belong to this ASR. The red feature indicates the seed θ'. The seed's membership is imputed by evaluating its spatial and temporal proximity measures from its neighbors (map features in green).

Sample attributes in OSM for a map feature are shown below:

```
A = {"addr:postcode"=>"8280","name"=>"Hauptstrasse", "type"=>"associatedStreet",
     "addr:country"=>"CH", "addr:city"=>"Kreuzlingen"}
```
(1)

The neighborhood of θ' is shown in Figure 3. From this, the ASR of θ' can be imputed as one of the membership class $\Gamma =$ ("Marktstrasse", "Hauptstrasse", "Löwenschanz", "Parkstrasse", "Sandbreitestrasse").

4 | PROPOSED APPROACH

MIA operationalizes the common heuristic known as *Tobler's law* (Tobler, 1970). The law deliberately leaves out the details of what *nearness* or *proximity* means. We explore diverse measures of spatial and temporal proximity for MIA, and demonstrate that while the heuristic is universally valid, individual choices of proximity measures significantly alter the performance of the algorithm. We consider the following three proximity measures:

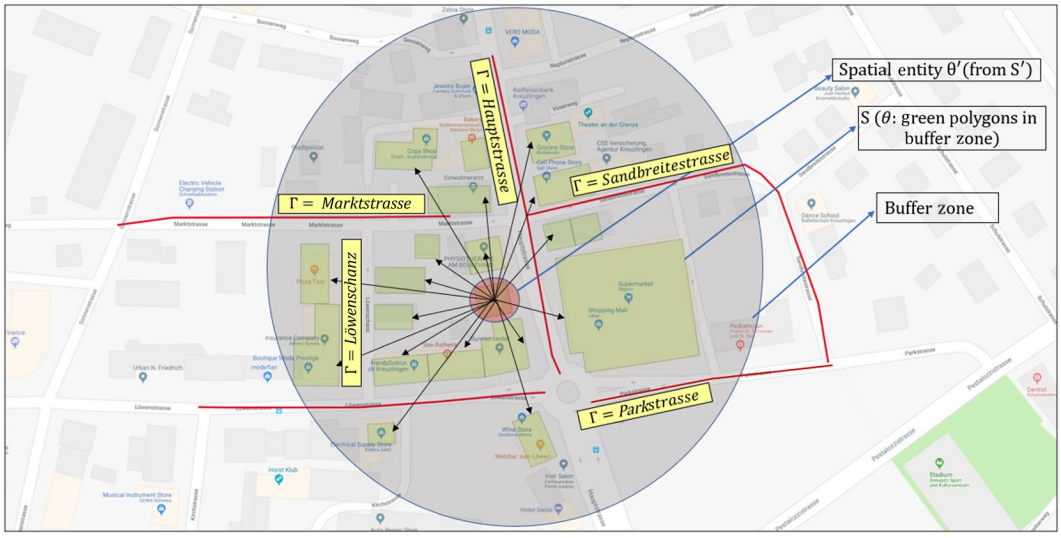


FIGURE 3 Imputation of an ASR for a map feature θ'

Spatial Proximity Measure (SPM): This is a measure of spatial proximity between a seed map feature and its neighboring map features based on Euclidean distance. The SPM is a simple and natural choice of proximity in spatial data. For example, among multiple map features (houses or apartment complexes) created by different contributors, map features closer to the given street can be considered as being more related, compared to map features clustered around other streets.

Temporal Proximity Measure (TPM): This evaluates the closeness in time of the seed's creation in relation to the creation of other map features in the data set (in other words, all neighbors created in the same changeset as the seed). Temporal proximity is not a measure unique to spatial data. For example, considering multiple map features (houses or apartment complexes) created by different contributors, map features inserted or altered at a similar time (e.g., part of the same changeset) can be considered as likely to be more closely related, compared to map features inserted or altered at a different point in time.

In our evaluation of MIA's performance, two variations of the TPM are used:

1. *Relative Temporal Proximity (RTP).* Only the temporal closeness of map features that have been mapped to an ASR (although possibly different ASRs) is considered by MIA. The distribution of neighbors across different ASRs determines the best fit for the seed's membership imputation.
2. *Absolute Temporal Proximity (ATP).* The temporal proximity of all neighborhood entities (irrespective of their membership in an ASR) is considered by MIA. In this scenario, there can be a combination of neighbors that belong to an ASR and those that do not belong to an ASR. As an example, spatial features such as drains and canals, which are not defined as belonging to any ASR, may very well be in temporal proximity to the seed, by virtue of having been created at around the same time as the seed. This can be in addition to other spatial features such as buildings and parks, which are defined as being a part of an ASR. The distribution of neighbors across these categories (belonging versus not belonging to an ASR) determines the best fit for the seed's membership imputation.

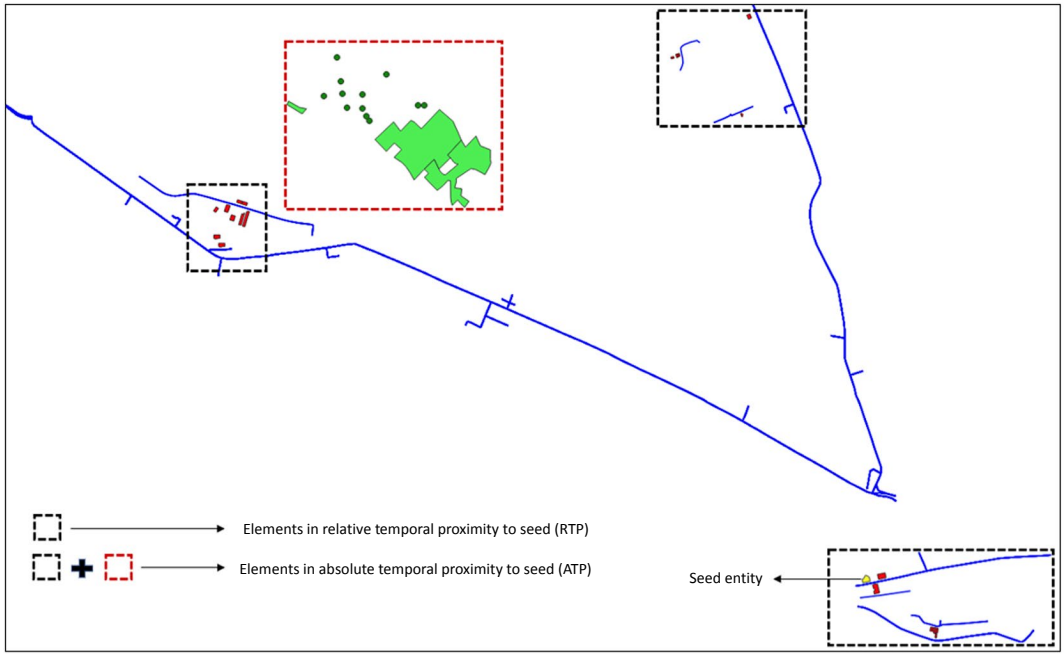


FIGURE 4 Illustration of temporal proximity measures

With reference to the above proximity definitions, a *neighborhood* in the context of temporal proximity (RTP or ATP) represents every spatial feature that was created at around the same time as the seed element, excluding itself, without accounting for a spatial distance filter. This is further explained in Section 4.1.2. Figure 4 illustrates both the temporal proximity measures from a single changeset in OSM. Neighborhood entities that only belong to ASRs (black rectangular regions) are candidates for RTP evaluation. All neighborhood entities (combination of red rectangular region with elements such as trees, parks and benches in addition to black rectangular regions) represent candidates for ATP evaluation.

Spatio-Temporal Proximity Measure (SPTPM): This is a combination of the SPM and RTP. MIA evaluates the spatial proximity on the neighborhood entities first, and then accounts for their closeness in time to the seed (temporal proximity), when evaluating the membership.

Considering the nature of the spatial data elements in the data set (such as residential buildings and apartment complexes) and the core objective of the algorithm, MIA accounts for the "Disjoint" and "Touches" topological relations (Clementini, Di Felice, & van Oosterom, 1993) in its framework, in conjunction with the proximity measures.

4.1 | The Membership Imputation Algorithm

MIA takes a set of seed entities whose ASR membership is to be imputed, a set of map features within a specified maximum search distance (here, a buffer specified in meters) and a flag indicating the variant of the proximity measure as input parameters (SPM, TPM, or SPTPM). The imputation methodology for each of these proximity measures is discussed further in Sections 4.1.1.–4.1.3, respectively. The examples used in the illustration are taken from the OSM Swiss data set, with MIA methods implemented in PostGIS (https://postgis.net/docs/reference.html). The algorithm returns the set of seeds along with their imputed ASR membership values.

4.1.1 | MIA with Spatial Proximity Measure

In this variant, MIA performs a *k*-nearest neighbors (*k*NN) search from the set of map features that are within the maximum spatial proximity search distance, as defined by the buffer zone input parameter. In accordance with Tobler's law, the nearest neighbors are ranked by inverse distance based on their proximity (i.e., closer neighbors rank higher than distant neighbors) using the inverse distance weighed (IDW) metric. In addition, MIA accounts for the total number of neighbors belonging to a given ASR, in order to determine an Adaptive Inverse Weighed Distance (AIDW) for a given neighborhood element. An integration of the normalized AIDW scores for each neighborhood element of an ASR determines the final AIDW score for the ASR. This is given by

$$aidw_j = \sum_{i=1}^{n} \frac{n}{d(\theta', \theta_i^j)}, \tag{2}$$

where θ_i^j, $1 \le i \le n$, belongs to the subset of neighbors that are associated with the jth ASR, that is, γ_j, and $d(\cdot)$ computes the Euclidean distance between two points.

Consider an example where a seed (whose neighbors are shown in Table 1) is mapped to the ASR (e.g., Γ = "Konradstrasse") in OSM. The neighbors are represented by their identifiers (Rank) in Table 1, each row representing a nearest neighbor to the respective seed, ranked by distance (the *D*[m] and IDW columns). The nearest neighbor (NN-1) to the seed is about 1 m away and its IDW score is 1.00. Similarly, the second nearest neighbor (NN-2) is about 12.24 m away with an IDW of 0.081. It can be seen that the 10 neighbors belong to four explicit ASR membership classes (Γ = "Konradstrasse" has four neighbors, Γ = "Neuwiesnstrasse" has three neighbors, etc.).

TABLE 1 Illustration of MIA spatial proximity measure

Rank	ASR	*D* [m]	IDW	*k*NN	AIDW	Total score
NN-1	Konradstrasse	1.00	1.00		4.00	
NN-2	Konradstrasse	12.248	0.081		0.326	
NN-4	Konradstrasse	27.221	0.036	4	0.146	4.556
NN-5	Konradstrasse	48.155	0.020		0.083	
NN-6	Neuwiesenstrasse	69.173	0.014		0.043	
NN-7	Neuwiesenstrasse	77.670	0.012	3	0.038	0.112
NN-9	Neuwiesenstrasse	98.811	0.010		0.030	
NN-8	Rudolfstrasse	92.85	0.010		0.021	
NN-10	Rudolfstrasse	108.694	0.009	2	0.018	0.039
NN-3	Salstrasse	19.965	0.050	1	0.055	0.050

The IDW score for each neighbor is weighed using the total elements in the ASR that the neighbor belongs to. This is shown in the AIDW column. The final AIDW score for each ASR is shown in the Total score column. MIA imputes the membership of the seed to the ASR that has the maximum AIDW score. In this example, MIA imputes the membership of the seed as "Konradstrasse"—this matches the ASR membership in the ground truth OSM data for the seed. From the example, even though NN-3 is a closer neighbor to the seed, there is low support to select it, as MIA uses a combination of distance and total neighbors to evaluate the best fit for the membership. The imputation outcome of MIA for an SPM is indicated using the function *SPM_Score()* in the algorithm presented in Section 4.1.4.

4.1.2 | MIA with Temporal Proximity Measure

In this scenario, all neighbors in temporal proximity to the seed are considered. Since this measure is independent of a search buffer around the seed, there can be many such neighbors, all originating from the same changeset as the seed. This is more pronounced in the ATP case. Table 2 shows an example from Swiss OSM data set for both variants of temporal proximity (Figure 4), for one seed.

The membership imputation in this scenario is primarily driven by the Temporal Weight (TW) of each ASR. Temporal Weight is defined as the measure of influence expended by any given ASR that is in temporal proximity to the seed, in determining its final membership association. Specifically, the temporal weight of each ASR is computed as the ratio of the number of neighbors in the given Associated Street Relation (ASR (*k*)) to the total number of neighbors in temporal proximity to the seed. This is shown in the TW column of Table 2. With reference to the temporal *neighborhood* defined in Section 4 and considering the example of RTP in Table 2, a seed that has been mapped to the ASR *Stadthausstrasse*, is shown to have elements from eight different ASRs that were created at about the same time as the seed. Seven of these neighborhood elements belong to the ASR *Holzlegistrasse*, and there are a total of 20 neighborhood elements in relative temporal proximity to the seed. Hence, the RTP of ASR *Holzlegistrasse* with respect to the seed is 7/20 = 0.35. In the absence of a spatial distance filter in this metric, the membership of the seed will be imputed as the ASR with the highest score for RTP. Considering the RTP example as shown in Table 2 (left), the ASR of the seed will be imputed as *Holzlegistrasse*, which is incorrect.

The key difference between ATP and RTP is that, for the former, a unique category, "Map Features Outside ASR" (the first row in the ATP section of Table 2), captures the ratio of neighbors in temporal proximity to the seed, which do not belong to any ASR (in addition to neighbors belonging to ASRs). Considering the example of ATP in Table 2 (right), a seed that has been mapped to the ASR *Stadthausstrasse*, is shown to have elements from 8 different ASRs that were created at about the same time as the seed. More importantly, it can be seen that there

TABLE 2 Illustration of MIA temporal proximity measure: (left) relative temporal proximity; (right) absolute temporal proximity

Actual ASR of seed	Neighbor ASR	*k*NN	TW
Stadthausstrasse	Am Schützenweiher	3	0.15
Stadthausstrasse	Hainbuchenweg	1	0.05
Stadthausstrasse	Holzlegistrasse	7	0.35
Stadthausstrasse	Schaffhauserstrasse	1	0.05
Stadthausstrasse	Stadthausstrasse	2	0.1
Stadthausstrasse	Steinberggasse	4	0.2
Stadthausstrasse	Technikumstrasse	1	0.05
Stadthausstrasse	Untertor	1	0.05
(RTP)	Relative temporal proximity		

Actual ASR of seed	Neighbor ASR	*k*NN	TW
Stadthausstrasse	Map features outside ASR	19	0.487
Stadthausstrasse	Am Schützenweiher	3	0.077
Stadthausstrasse	Hainbuchenweg	1	0.025
Stadthausstrasse	Holzlegistrasse	7	0.179
Stadthausstrasse	Schaffhauserstrasse	1	0.025
Stadthausstrasse	Stadthausstrasse	2	0.051
Stadthausstrasse	Steinberggasse	4	0.102
Stadthausstrasse	Technikumstrasse	1	0.025
Stadthausstrasse	Untertor	1	0.025
(ATP)	Absolute temporal proximity		

are 19 map features that do not belong to any ASR, but are in temporal proximity to the seed (spatial features such as benches, trees, and vineyards discussed previously in Section 4, shown inside the red rectangular region in Figure 4). In the absence of a spatial distance filter in this metric, the membership of the seed will be imputed to an ASR, if an ASR has the highest score for ATP among all the neighborhood elements. In contrast, if elements that do not belong to any ASR constitute the highest number of neighbors around the seed, the algorithm will not be able to impute a membership for the seed with any ASR from the neighborhood elements. This is shown for the ATP section of Table 2, wherein there are 19 map features that do not belong to any ASR, represented as "Map Features Outside ASR", thereby giving an ATP score of 19/39 = 0.487 (highest among all the neighbors). The inability of the algorithm to assign an ASR membership to the seed is considered as an incorrect imputation of the membership. This is because the seed is already known to belong to the ASR *Stadthausstrasse* in the ground truth data set (shown in the ATP section of Table 2). Since we are considering every neighbor without any spatial distance filter, the seed could be associated with an ASR with a large spatial distance, even though it is in temporal proximity to the seed. In other words, the application of TPM for the imputation is usually only effective when used in conjunction with the SPM.

Finally, the first row for the RTP section in Table 2 has been intentionally left blank. This row, only relevant and applicable to ATP, represents the categorization of elements that do not belong to any ASR (denoted by "Map Features Outside ASR" in Table 2) and its influence on the membership imputation. The remaining *Neighborhood* elements belonging to one or more ASRs are the same for both temporal proximity measures.

4.1.3 | MIA with Spatio-Temporal Proximity Measure

While the imputation accuracy of MIA is high across various buffer zones for the SPM (see Section 5 for sensitivity analysis), there are scenarios where it does not succeed in finding the correct ASR. For instance, in Table 3, by using the SPM, MIA selects the ASR of the seed as "Bahnhofplatz", instead of "Stadthausstrasse" (according to ground truth OSM data). In such scenarios, MIA's performance can be enhanced by applying an additional TPM in conjunction with the SPM. In doing so, MIA considers the percentage of neighbors from each ASR that are in temporal proximity to the seed, in order to assign a temporal weight to the ASR among the neighborhood elements.

The algorithm evaluates the TPM only after the evaluation of the SPM among the neighborhood elements for a given ASR. The product of the AIDW score (determined by applying the pure SPM filter on the neighborhood elements and shown in the Total column in Table 3) and the temporal weight of each ASR with the seed (shown in the TW column in Table 3) determines the final SPTPM score for each ASR (shown in the rightmost column in Table 3). For example, in Table 3, none of the neighbors belonging to ASR "Bahnhofplatz" (from changesets CS-1, CS-3, CS-6, CS-7, and CS-8) are in temporal proximity to the seed (changeset SEED-CS). Hence, the temporal weight of ASR "Bahnhofplatz" with relation to the seed is 0. Even though the elements of ASR "Bahnhofplatz" are in spatial proximity to the seed, since their temporal weight in relation to the seed is 0, their final SPTPM score is also 0. Similarly, one out of the three neighbors belonging to ASR "Stadthausstrasse" is in temporal proximity to the seed, leading to the temporal weight of this ASR being 0.333 and its final SPTPM score of 0.167. ASRs with no neighbors in temporal proximity to the seed will have a temporal weight of 0 (so as the final SPTPM score) and hence cannot be selected as a possible candidate for a membership imputation by MIA. The outcome of MIA for the SPTPM is indicated using the *SPTPM_Score()*" function in the algorithm presented in Section 4.1.4.

Finally, in the scenario where none of the neighbors are in temporal proximity with the seed, MIA falls back on pure SPM-based AIDW of the neighbors as the final score (even if the algorithm is being evaluated explicitly for the SPTPM), in order to determine the membership for the seed. Thus, in the example in Table 3, MIA imputes the membership of seed as "Stadthausstrasse" (which is identical to ground truth OSM data).

TABLE 3 Illustration of MIA spatio-temporal proximity measure

ID	ASR	D [m]	IDW	*k*NN	AIDW	Total	Changeset	Status	TW	SPTPM Score
NN-2	Bahnhofplatz	8.217	0.012		0.604		CS-1	Mismatch		
NN-3	Bahnhofplatz	19.549	0.051		0.255		CS-3	Mismatch		
NN-6	Bahnhofplatz	47.332	0.021	5	0.105	1.104	CS-6	Mismatch	0	0
NN-7	Bahnhofplatz	67.927	0.014		0.073		CS-7	Mismatch		
NN-8	Bahnhofplatz	77.353	0.012		0.064		CS-8	Mismatch		
NN-1	Stadthausstrasse	7.406	0.135		0.405		CS-1	Mismatch		
NN-10	Stadthausstrasse	87.468	0.111		0.034		CS-10	Mismatch		
NN-5	Stadthausstrasse	46.778	0.021	3	0.064	0.503	SEED-CS	Match	0.333	0.167
NN-4	Untertor	22.811	0.043		0.087		CS-4	Mismatch		
NN-9	Untertor	79.96	0.012	2	0.025	0.112	SEED-CS	Match	0.5	0.056

4.1.4 | Algorithm logic overview

This section summarizes the logic flow in MIA as highlighted in Algorithm 1. MIA executed for the SPTPM gets the *k* nearest neighbors around the seed, based on the buffer zone (line 1). Each of these neighbors are grouped by the ASR that they belong to (line 2). For each ASR in the group, MIA computes the SPM score along with the SPTPM score using the neighbors present in the ASR (line 4). Finally, the algorithm checks if there is a value for the SPTPM score. If present (thereby indicating that these neighborhood entities have a spatial proximity as well as temporal proximity with the seed), MIA returns the ASR that has the maximum SPTPM score (line 8) from the ASR groups of the neighbors. In the absence of a maximum value for SPTPM score (i.e., there are no neighbors in temporal proximity to the seed and they are only related by spatial proximity), the algorithm returns the maximum SPM score from the ASR group (line 10). The membership for the seed is imputed with the ASR returned by MIA. This process is repeated for every seed in the input data set. It should be noted that the set of neighborhood elements for MIA is always determined based on the buffer zone (query radius) around the seed. The algorithm has been executed for varying buffer zones between 0 and 1,000 m in order to analyze and evaluate the sensitivity of the algorithm with respect to the imputation accuracy.

Algorithm 1: MIA - Spatio-Temporal Proximity Measure (SPTPM)

Input: S, A, Γ, θ' and Buffer Zone R (in meters)

Output: Associated Street Relation γ for seed θ'

1 $N \leftarrow$ Buffer_Filter(θ', R, S) : $N \subset S$

2 NG $\leftarrow$ Group_By_ASR(N, Γ)

3 **foreach** *ng IN NG* **do**

4 $\{ng'\} \leftarrow$ SPM_Score(ng, θ')

$\{ng''\} \leftarrow$ SPTPM_Score(ng, θ')

5 $\gamma_{spm} \leftarrow$ ASR_Name(max($\{ng'\}, \Gamma$) : $\gamma_{spm} \subset \Gamma$

6 $\gamma_{sptpm} \leftarrow$ ASR_Name(max($\{ng''\}, \Gamma$) : $\gamma_{sptpm} \subset \Gamma$

7 **if** γ_{sptpm} *IS NOT NULL* **then**

8 **return** γ_{sptpm}

9 **else**

10 **return** γ_{spm}

11

4.2 | Baseline imputation algorithms

Five baseline algorithms, operationalizations of Tobler's law, are developed to assess the performance of MIA:

1. **Nearest Neighbor Street Name (NNSN)** – For a given seed, NNSN identifies the seed's nearest neighbor map feature by Euclidean distance. The seed's *Street Name* is imputed as the value of the nearest neighbor's *Street Name* key.
2. **Nearest Neighbor Street Name through Majority Voting (NNSN MV)** – For a given seed, NNSN MV identifies its nearest neighbors (using Euclidean distance) for a default buffer zone. Majority voting (Dymitr & Bogdan, 2005; Kotsiantis, 2007) is a decision rule that selects a winner from one or more alternatives, based on the majority of elements in each of them. For this baseline, majority voting is performed among the street names of all the nearest neighbors, to ascertain the street name with the maximum number of neighbors. This street name is declared as the winner. In the scenario of one or more street names having the same number of neighbors, the

(a) MIA 1mt (b) MIA 5mts (c) MIA 10mts

FIGURE 5 MIA: imputations using neighbors around the seed

majority voting procedure considers one of these street names at random, to declare a winner. The seed's *Street Name* key is then imputed as the *Street Name* of the winner. Finally, if a majority of the neighbors around the seed do not have a street name assigned (due to the semantic data quality issues in VGI data discussed in Section 2.2), the algorithm will not be able to assign a *Street Name* to the seed, and hence the imputation scenario for this seed is not considered to be successful.

3. **Nearest Neighbor Associated Street Name (NNASN)** – For a given seed, NNASN identifies the ASR of the seed's nearest neighbor map feature (by Euclidean distance) and imputes the seed's membership to this relation. The seed's *Street Name* would be imputed as the *Name* of the ASR (i.e., by the virtue of the higher-order OSM relation).
4. **Nearest Neighbor Associated Street Name through Majority Voting (NNASN MV)** – For a given seed, NNASN MV identifies the ASRs of the seed's nearest neighbors (by Euclidean distance). Majority voting is performed among the ASRs of all the nearest neighbors, to ascertain the associated street with the maximum number of neighbors. This ASR is declared the winner and the seed's membership is imputed to this relation. The seed's *Street Name* would then be derived as the *Name* of this ASR (i.e., by virtue of the higher-order OSM relation). In the scenario of one or more associated street names having the same number of neighbors, the majority voting procedure considers one of these associated street names at random, to declare a winner.
5. **Nearest Associated Street Name (NAS)** – For a given seed, NAS identifies the nearest ASR to the seed (and the corresponding street map feature), determined by minimum Euclidean distance between the nearest vertices of the street's geometry and the seed's geometry. The seed's membership is imputed to this ASR. The seed's *Street Name* would then be derived as the *Name* of this ASR (i.e., by the virtue of the higher-order OSM relation).

MIA significantly extends the baseline algorithms by harnessing the spatial and temporal metrics of a set of neighborhood entities, instead of a single nearest neighbor of a seed (or a simple majority vote from a set of neighborhood entities, which do not account for the spatial and temporal properties of the data). In doing this, MIA addresses the issues of equidistant neighbors belonging to different streets. Consider Figure 5a, where the seed has two touching neighbors, each belonging to a different ASR membership class (*Strittackerstrasse* and *Kernstrasse*). This issue frequently occurs in dense urban neighborhoods (such as buildings in an apartment complex). MIA addresses the challenge by varying the buffer size and the number of neighbors *k*, as shown in Figures 5b and c. From Figure 5b, when doing an ASR membership imputation for the seed (highlighted in red), even though there are two neighbors touching the seed (and belonging to different ASRs *Strittackerstrasse and Kernstrasse*), MIA can impute the membership as *Kernstrasse* with higher confidence due to the presence of two other neighborhood entities that belong to ASR *Kernstrasse*.

Land use category	CLC classification nomenclature (level 1)
Urban	1 – Artificial surfaces
	2 – Agricultural areas
Rural	3 – Forest and semi-natural areas
	4 – Wetlands
	5 – Water bodies

FIGURE 6 MIA categories versus CLC nomenclature

5 | EXPERIMENTAL EVALUATION

In this section we evaluate the performance of MIA against the baseline algorithms, along with a sensitivity analysis of MIA to the input parameters.

5.1 | Experimental setup

We report performance results across three test regions: Switzerland, Great Britain, and France. We also show the results in contexts of particular importance: at street intersections, and across urban and rural environments.

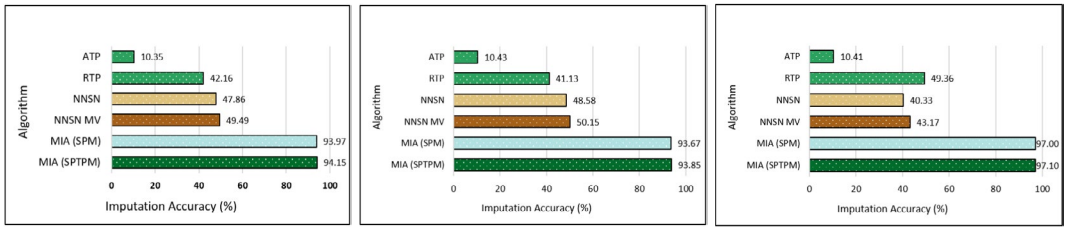
Hardware. All experiments were conducted on a cloud server with eight virtual CPUs, 32 GB RAM, and a 5 TB disk, running on the Ubuntu Linux 16.04 LTS operating system.

OSM data. All experiments use OSM data (Switzerland (291 MB), Great Britain (967 MB), and France (3.4 GB)). French and British data are current as of December 2018 and Swiss data are current as of March 2018. Data were downloaded from Geofabrik (https://download.geofabrik.de/) and loaded into a PostGIS database for analysis with MIA. The OSM Swiss data set contains 3,036 ASRs with about 66,200 map features. The OSM Great Britain data set contains 5,270 ASRs with about 144,527 map features. The OSM France data set contains 11,612 ASRs with about 169,249 map features.

Land cover data. Copernicus is a European system for Earth monitoring. The CORINE Land Cover (CLC) Inventory for 2012 (https://land.copernicus.eu/pan-european/corine-land-cover) is a pan-European land cover data set, used here to stratify OSM data by environments, between urban and rural, based on the coarse CLC Level 1 classification (Figure 6).

5.2 | Ground truth data set

OSM data, where the correct value of each seed's ASR is known in advance, are the ground truth for all of our experiments. In order to assess the extensibility and generality of MIA across independent data sets, a fivefold cross-validation (Arlot & Celisse, 2010) is performed across different partitions of the main data set. In each iteration, we consider a 20% subset as the test data, by explicitly removing the ASR membership values. MIA's imputation of the seed's ASR is compared against the actual known values of the ASR from OSM. MIA's imputation is then averaged for the cross-validation runs, to assess the overall accuracy. In addition, MIA's performance has been assessed and presented with reference to entities near street intersections. These spatial features, referred to as "Intersection Entities" in the remainder of this article, have also been curated using OSM data. They have been identified and filtered as spatial features present in the vicinity of street intersections, where at least two or more streets meet. In our experiments, we consider a default buffer size of 100 m to identify map features around intersections.



(a) All Entities (b) All Entities (URBAN) (c) All Entities (RURAL)

FIGURE 7 MIA versus Baseline 1 and 2 (Nearest Neighbor Street Name (NNSN) and Nearest Neighbor Street Name through Majority Voting (NNSN MV)), All Entities

5.3 | Experimental results

The results of the algorithm are presented across two broad categories. The *All Entities* category comprises all seeds in the data set (such as all buildings belonging to a street). The *Intersection Entities* category focuses on the performance of MIA, specifically with respect to seeds present at street intersections (e.g., buildings at street corners). Intersection entities are analyzed separately because the neighborhood of entities at intersections presents a unique challenge, due to neighbors being distributed across multiple ASR membership classes, in comparison to a neighborhood around a given street, a pattern dominating the overall data set.

5.3.1 | MIA versus Baseline Imputation Algorithms

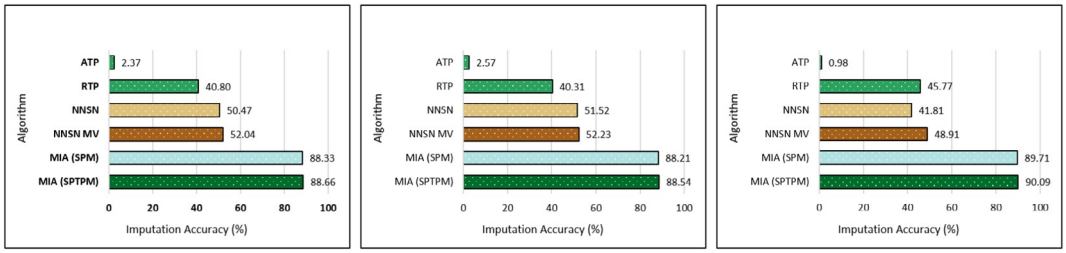
The MIA results are presented as the average accuracy of the algorithm in a buffer zone of 0–15 m, executed in spatio-temporal proximity mode. The default buffer zone for the algorithm was empirically learnt, and the accuracy was observed to be optimal among all the buffer zones used in our experiments (varying between 0 and 1,000 m). The results are then compared with the five baseline algorithms.

NNSN and NNSN MV Baseline (All Entities)

The results for this scenario are shown in Figure 7. In NNSN baseline, imputing a missing street name for the seed equates to imputing the street name of its nearest neighbor. Using this approach, we see that the accuracy of NNSN baseline is very low at about 48% (Figure 7a). The low levels of accuracy with the NNSN baseline are primarily due to the issues of semantic accuracy and missing attributes of map features in VGI data sets, discussed in Section 2.2.

Many of the map features miss street names themselves and thus are not effective in the NNSN imputation. In addition, a common scenario shown in Figure 5a depicts the ambiguity of imputation of the street names for the seed (in red) based on the street names of its immediate neighbors where both touch the seed and both are associated with different street names. A clear decision is thus impossible with the NNSN baseline algorithm.

The NNSN MV baseline algorithm performs a majority voting procedure to ascertain the most commonly occurring street name from the neighborhood elements. This approach is also not immune to the challenges of missing attribute data and semantic accuracy of VGI data sets, seen previously with the results of the NNSN baseline. In other words, missing street names for multiple neighbors is going to manifest itself when carrying out a majority vote, where the baseline algorithm will not be able to impute the street name of the seed. This is because the majority voting algorithm considers the set of neighbors with missing street names as a common group and declares them the winner if they have a majority vote count. In the absence of a street name for this group of elements, the street name of the seed cannot be imputed. This scenario is considered a miss when evaluating the



(a) Intersections (b) Intersections (URBAN) (c) Intersections (RURAL)

FIGURE 8 MIA versus Baseline 1 and 2 (Nearest Neighbor Street Name (NNSN) and Nearest Neighbor Street Name through Majority Voting (NNSN MV)), Intersection Entities

accuracy of the baseline. Secondly, in the presence of two distinct groups among the neighborhood elements with an equal representation in both groups (one group of neighbors having a street name and one group of neighbors not having a street name), the group of neighbors without a street name can have an undue influence on the random selection process of the winner, in the case of a tie, over a large data set. Therefore, it is not surprising to see that the NNSN MV baseline algorithm performs only marginally better in comparison to the NNSN baseline, with an imputation accuracy of about 49.5%.

MIA significantly outperforms both the NNSN and the NNSN MV baselines with a high accuracy of about 94.15% when imputing the seed's membership, as shown in Figure 7a. In addition, Figures 7b and c present the accuracy of MIA across urban and rural environments, respectively. The performance is slightly higher at about 97% for rural environments. This could be related to the highly urban bias of the OSM data set, with very few ASRs in rural areas.

NNSN and NNSN MV Baseline (Intersection Entities)

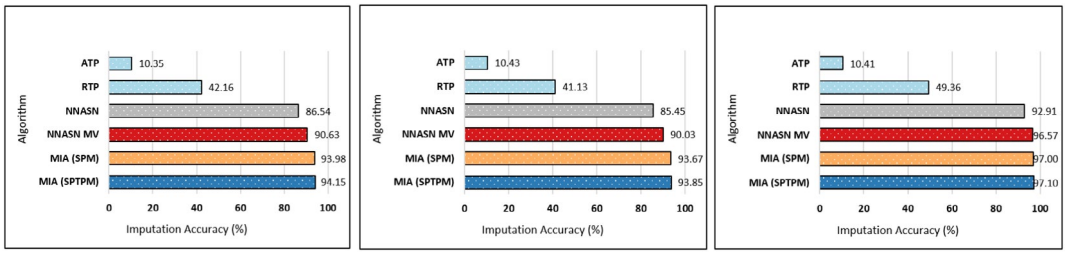
Intersection entities are not immune to the problem of missing attributes (Figure 8). Figure 8a shows that the accuracy of the NNSN baseline is slightly higher (at about 50%) at intersections, and similar in accuracy for "All Entities" (Figure 7a). In line with the observations made for "All Entities", the NNSN MV algorithm performs marginally better at intersections than its NNSN counterpart, with the accuracy being about 52%.

In comparison to the baselines, MIA delivers an average accuracy of about 88.67% for "Intersection Entities". The imputation accuracy is slightly lower compared to "All Entities", primarily due to the challenges of multiple ASRs in the vicinity of street intersections in dense urban landscapes. Overall, we see that MIA executed using a spatio-temporal proximity measure, considerably outperforms both NNSN baselines, with the average accuracy levels of the membership imputation varying between 88% and 94%, considering both normal and intersection entities.

NNASN and NNASN MV Baseline (All Entities)

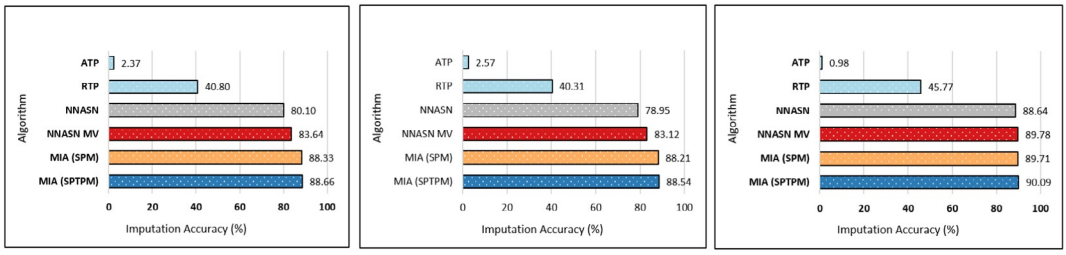
The results for this baseline are shown in Figure 9. In this approach, by the virtue of belonging to the same ASR as its nearest neighbor, the *Street Name* for the seed is implicitly derived from the higher order ASR name. The completeness of attributes (such as the presence or absence of "Street Name" or other such attributes used to describe a spatial entity's *Address*) do not drive the creation of an ASR. In other words, the relation is governed by spatial proximity and not semantic completeness. Therefore, it is not surprising that the NNASN baseline performs well and its accuracy is high.

In contrast to the missing street names in NNSN MV baseline having an adverse impact in the majority voting procedure, the mechanism of a well-defined ASR helps in making a beneficial contribution towards the majority voting procedure, when considering the NNASN MV baseline. In dense urban neighborhoods, where the nearest



(a) All Entities (b) All Entities (URBAN) (c) All Entities (RURAL)

FIGURE 9 MIA versus Baseline 3 and 4 (Nearest Neighbor Associated Street Name (NNASN) and Nearest Neighbor Associated Street Name through Majority Voting (NNASN MV)), All Entities



(a) Intersections (b) Intersections (URBAN) (c) Intersections (RURAL)

FIGURE 10 MIA versus Baseline 3 and 4 (Nearest Neighbor Associated Street Name (NNASN) and Nearest Neighbor Associated Street Name through Majority Voting (NNASN MV)), Intersection Entities

neighbor to a seed may be from a different ASR as compared to the seed (and thereby influencing the NNASN baseline algorithm to make a wrong choice when considering only the nearest neighbor), a majority vote among the ASRs of all the neighbors is a better option. This is evident from the accuracy of the NNASN MV baseline algorithm, which, not surprisingly, is slightly higher than the NNASN baseline, at about 90.6%.

While the accuracy of the two NNASN baseline algorithms vary between 86% to 90% (Figure 9a), MIA outperforms both these baselines, with its average accuracy exceeding by over 4% at the higher end of the baselines, at about 94.15%. Figure 9b,c present the accuracy of MIA across urban and rural environments respectively, showing patterns similar to the NNSN baseline algorithms, due to the nature of the data set.

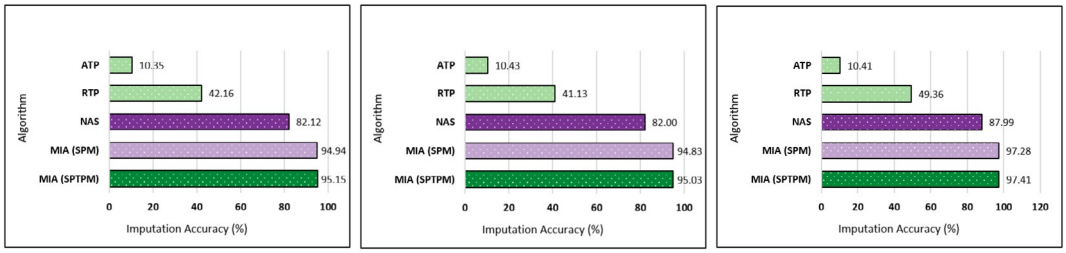
NNASN and NNASN MV Baseline (Intersection Entities)

While the NNASN baseline for intersections is high at about 80% (Figure 10), and the NNASN MV baseline performing marginally better at about 83.6%, MIA's imputation accuracy is still about 5% higher than the both the NNASN baseline algorithms at about 88.66%. The strength of MIA is evident when considering the challenges of multiple ASRs around street intersections, as shown in Figure 10a.

Overall, we can conclude that while the design principles of spatial proximity in ASRs contribute to good baselines for both scenarios ("All Entities" and "Intersection Entities"), MIA executed using a spatio-temporal proximity measure outperforms both the NNASN baseline algorithms with accuracy levels varying between 88 and 94% and a minimum improvement of about 4%.

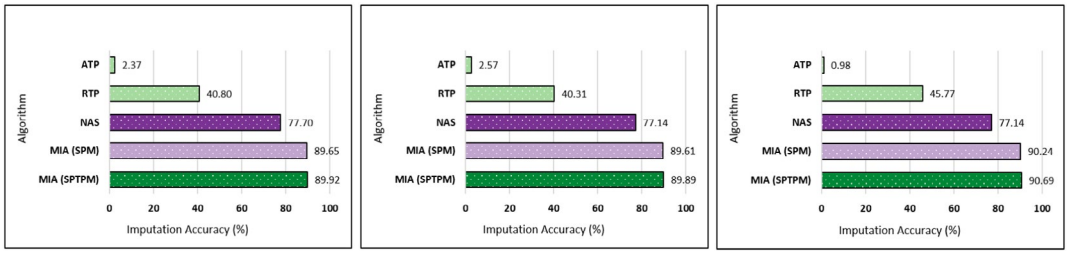
NAS Baseline (All Entities)

The NNASN baseline can also be extended by directly considering the underlying associated street, instead of the associated street of the seed's nearest neighbor (or a majority vote from the ASRs of the nearest neighbors).



(a) All Entities (b) All Entities (URBAN) (c) All Entities (RURAL)

FIGURE 11 MIA versus Baseline 5 (Nearest Associated Street (NAS)), All Entities



(a) Intersections (b) Intersections (URBAN) (c) Intersections (RURAL)

FIGURE 12 MIA versus Baseline 5 (Nearest Associated Street (NAS)), Intersection Entities

In this approach, imputing a missing *Street Name* for the seed can be addressed as imputing the membership of the seed to its nearest ASR. In line with the previous NNASN baselines, the street name for the spatial entity is implicitly derived from the higher-order membership. Exhibiting a similar behavior in terms of the high accuracy levels as compared to the NNASN baselines, the current NAS baseline performs generally well at about 82%, as shown in Figure 11. Still, it is worthwhile to note that, MIA's accuracy is much higher exceeding the NAS baseline by 13% at about 95% (Figure 11a). Figures 11b and c present the accuracy of MIA across urban and rural environments, respectively.

NAS Baseline (Intersection Entities)

The NAS baseline for intersection entities is about 77%, but lower than for all entities, for challenges of multiple ASR memberships and neighborhood elements present in the vicinity of intersections (Figure 12). MIA's accuracy is about 12% higher than the NAS baseline, at about 90% as shown in Figure 12a.

Summary

From the results presented and discussed for the five baseline algorithms above, we see that MIA's imputation accuracy is superior. Even accounting for strong the baselines that are primarily a result of well-formed ASRs in the input data set, MIA's accuracy outperforms consistently in all scenarios. Furthermore, none of the data cleansing tools reviewed in Section 2.1.2 have discussed their imputation capabilities and support for spatial data sets. While this scenario presents a unique challenge of not being able to compare MIA's performance against existing industry standard tools, our approach that effectively exploits the spatio-temporal characteristics of VGI data, shows the strong performance and potential of MIA against a variety of baselines. Finally, it serves as a promising direction for the future development of spatial data cleansing tools, frameworks and algorithms.

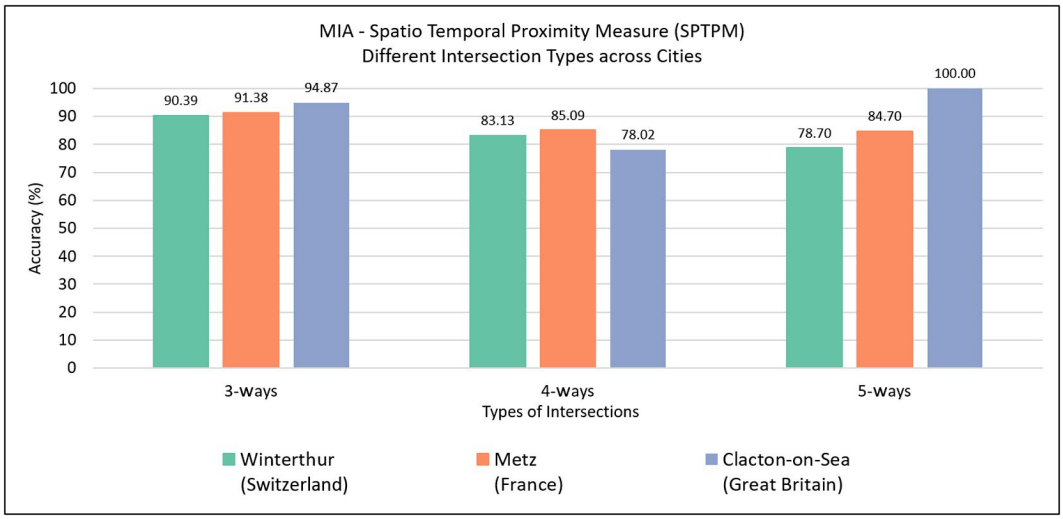


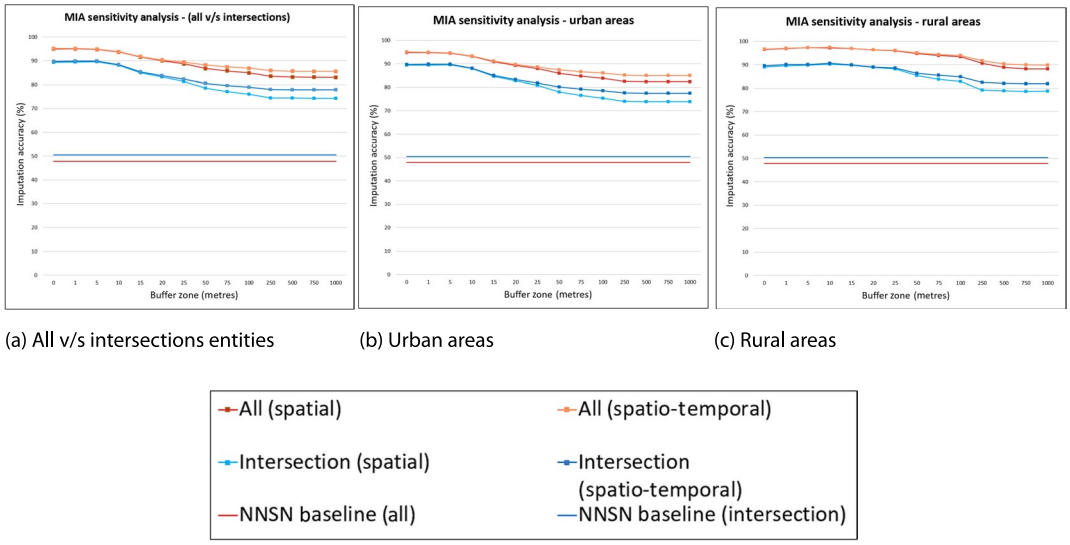
FIGURE 13 MIA: performance based on different types of intersections

5.3.2 | MIA: Street intersection types

The design of street networks is closely dependent on the natural geographical constraints in the area of their presence. Hence, not all street networks are designed as simple grid structures, thereby accounting for diverse street intersection patterns such as *n*-way intersections (three- or four-way intersections) and crossroad intersections (such as Y-intersections). The complexity of these polymorphic street intersections is a driving factor governing the distribution of their neighborhood spatial entities, thus influencing the accuracy of MIA's framework. Figure 13 analyzes the performance of MIA in the context of diverse street intersection types.

We undertake this exercise considering different street intersection types for three cities, one from each geographical region discussed in the article. The analysis takes into account OSM ways, which are normal road segments (ways that can be traversed by both cars and pedestrians). The road intersection types for the three cities have been obtained from the "Intersections Framework" of Fogliaroni, Bucher, Jankovic, & Giannopoulos (2018) which makes worldwide road intersection data available for *At-Grade* level roads, using OSM data. At-Grade intersections are more applicable in the context of an ASR, as opposed to *Grade-Separated* intersections (Wolhuter, 2015) which are mostly used to represent highways and motorways.

The data distribution statistics for intersections indicate that they almost entirely represent three- and four-way intersections. These two intersection categories account for over 99% of the roads for our three cities. This pattern is consistent with the results reported in Fogliaroni et al. (2018) for many other cities in the world. MIA's accuracy is well above 90% for three-way intersections across all three cities (Figure 13). Considering the distribution of spatial entities across more complex four-way intersections, the accuracy is lower than for three-way intersections, but still higher at over 83% for cities in Swiss and France, and marginally lower by 2% at about 78% for Great Britain. The results for five-way intersections should be treated with caution. Even though the results look encouraging, with a minimum accuracy of about 78% for Winterthur, the data set contains only six intersections that are five-way. The scenario is slightly better for Metz in France, but still the data set only accounts only for about 44 five-way intersections. The data set is negligible (only a couple of five-way intersections) for Clacton-on-Sea in Great Britain. But overall, the imputation accuracy for MIA, even accounting for the challenges of different street intersection types, is significantly higher than the accuracy of the baseline algorithms discussed in Section 5.3.1.



(a) All v/s intersections entities (b) Urban areas (c) Rural areas

FIGURE 14 MIA sensitivity analysis, all environments, urban, and rural

5.4 | Time complexity of MIA

The time complexity of MIA is primarily governed by the (i) total seed entities (n) for imputation with a membership class; and (ii) maximum number of neighborhood entities (m) in the largest buffer zone among all the seed entities ($m = \max\{m_1, m_2, m_3, \ldots, m_n\}$, where, given $1 \le i \le n$, m_i is the total number of neighbors in a buffer zone of 1,000 m for the i^{th} seed). The neighborhood entities are determined through PostGIS functions that use the underlying R-Tree spatial indices (Guttman, 1984) and exhibit a time complexity of $O(\log m)$. Overall, MIA has been found to exhibit a linear logarithmic time complexity $O(n \log m)$ for the imputation, considering all the seed elements used in our test data set. Furthermore, since MIA's optimal performance has been established to occur in lower buffer zones (discussed in Section 5.3), the scalability of the algorithm stands up well for the imputation tasks.

5.5 | Sensitivity analysis of MIA

MIA has been further analyzed to assess its sensitivity to different values for the buffer zone input parameter. The analysis has been performed for buffer zones varying between 0 m (neighbors touching the seed) and 1,000 m around the seed, and the results are discussed below.

5.5.1 | The impact of spatial buffer size

When MIA is executed using an SPTPM for "All Entities" in the data set, the algorithm is found to execute optimally in a buffer zone of 0–15 m, where the average accuracy is about 94.15% (Figure 14a). While the accuracy decreases with an increasing buffer zone (as predicted by Tobler's law), it remains healthy and above 85% for the largest buffer zones. More importantly, the lowest accuracy of about 85.5% for a buffer zone of 1,000 m still outperforms the NNSN baseline (at about 48%). Even considering a pure SPM, MIA outperforms the NNSN baseline with an accuracy ranging between 83 and 95%. Figure 14a also shows that while the impact of considering the temporal proximity filter in the algorithm for an SPTPM is marginal at about 0.3% for lower buffer zones (varying

between 0 and 15 m), the importance of the temporal proximity filter becomes apparent for larger buffer zones (varying between 15 and 1,000 m), where the temporal dimension adds an additional accuracy boost of about 2.5% in comparison to a pure SPM.

We also explored the sensitivity of MIA at street intersections with the same buffer zone variation (Figure 14a). Considering the challenges of multiple ASRs present in small buffer zones around street intersections and the distribution of neighborhood entities across multiple such relations, the accuracy of MIA is slightly lower around street intersections compared to "All Entities". Using an SPM, the optimal performance of MIA is observed in a buffer zone of 0–15 m, where the average accuracy is about 88.3%. Considering an SPTPM, the accuracy improvement is seen to be marginally higher at about 88.66%. However, the accuracy holds up in larger buffer zones at about 78% for a buffer of 1,000 m, as compared to 74% for a pure SPM. This is consistent with the behavior for "All Entities", where the impact of temporal dimension is significant for larger buffer zones, contributing to an additional accuracy boost of about 4% in comparison to a pure SPM. The lowest accuracy of MIA with an SPTPM is about 78% and superior to the NNSN baseline of about 48%.

5.5.2 | Urban versus rural regions

The spatial buffer size sensitivity analysis of MIA for all map features is also discussed by stratifying the results across urban and rural regions. Figures 14b and c show the accuracy of MIA across Urban and Rural regions. The results are quite similar to the accuracies achieved for the overall data set. Considering "all entities" and urban areas under an SPM, the accuracy varies between 82 and 95%. It improves to between 85 and 95% when the algorithm is executed for an SPTPM. The accuracy is slightly higher in rural settings and varies between 88 and 96% for an SPM, and between 90 and 96% for an SPTPM. An enhanced accuracy in a rural setting is primarily caused by a low percentage of seeds, fewer neighborhood entities, and ASRs in a rural setting. While the data used for the sensitivity analysis (Switzerland OSM data) have ASRs spread across both urban and rural regions, the data are highly urban biased.

The algorithm's behavior is similar for intersection entities as well. From Figures 14b and c, it can be seen that the accuracy levels in rural settings are slightly higher, compared to the accuracy levels for urban landscapes. The accuracy under an SPTPM is seen to vary between 78 and 90% in urban settings, as compared to Rural settings where it varies between 82 and 90%. Overall, the accuracy of MIA is superior at all buffer zone sizes for both urban and rural areas, compared to the NNSN baseline of about 48%.

5.5.3 | The impact of spatial and temporal metrics

The contribution of each input metric has been analyzed separately and holistically. The results are then compared with results from the baseline algorithms to discuss the accuracy of MIA, as seen through the lens of spatio-temporal metrics being the key drivers for the imputation.

Temporal proximity measures

With reference to Figures 7, 9, and 11, it can be seen that when MIA is executed under ATP, the accuracy is very low at about 10%. This is consistent across urban and rural landscapes. This is because ATP considers all neighbors in temporal proximity to the seed. Most of these neighbors would not be a part of any membership class in general, and an ASR membership class in particular. A majority of the neighbors not belonging to a membership class would drive the algorithm to disassociate the seed from any ASR membership. As an example, if there are 100 neighbors in temporal proximity to the seed and 50 of these are not a part of any ASR (e.g., map features such as trees, canals, or ditches), the algorithm would be influenced by this large group and keep the seed independent of

any membership. The challenges at intersections are more pronounced, with the accuracy of ATP for *Intersection Entities* being very low, at about 2%. This is consistent across urban and rural areas, as evident from the results in Figures 7, 9, and 11. Lastly, another reason for why a pure TPM may underperform is the fact that within the same changeset, the data may suffer from systematic errors brought in by a single mapper, thereby violating the MNAR assumption. This is mediated in the RTP measure for the algorithm.

The results of MIA when executed under RTP are much higher, at about 42%, but much lower than the majority voting NNSN and NNASN baseline algorithms (of about 49 and 90%, respectively) for the imputation (Figures 7, 9, and 11). The accuracy of MIA for RTP is much higher than for ATP because RTP considers only those neighbors in temporal proximity of the seed, with the additional condition that these neighbors are also associated with an ASR membership class. The accuracy is still low (40%) primarily because, pure temporal proximity can have neighbors belonging to ASRs that are quite far from the seed and still influence the overall results. As an example, if there are 100 neighbors belonging to different ASRs and 50 of these belong to an ASR that is very far off from the seed, the algorithm will still be influenced by this distant street and impute the membership of the seed to this distant relation (noting that there are no constraints of spatial proximity when using a pure TPM). In addition, MIA executed using a pure TPM is computationally expensive. This is because the number of neighbors to a given seed can be very high (observed to be in the thousands for well-mapped countries such as Switzerland and Great Britain). This is more pronounced with ATP because all neighbors are considered.

Spatial proximity measures

Alternatively, as shown in Figures 7, 9, and 11, when MIA is executed using an SPM, the accuracy is at about 94% in general for all map features. The results are consistent across urban and rural areas, with MIA performing slightly better in rural settings with an accuracy of about 97%. MIA's accuracy is much better (by about 8–10%), even when compared with NNASN and NNSN baselines (which themselves are good at about 86 and 82%, respectively). While the accuracy of MIA drops a little for *Intersection Entities*, it still healthy at about 90%. The accuracy is consistent without much variation across urban and rural areas. The SPM is the majority contributor towards MIA's accuracy and is not computationally expensive. The high accuracy is primarily related to the observations underpinning Tobler's law, and to closer entities having more influence on the seed than farther entities. Furthermore, the performance of the spatial filter underpinning the SPM is trivially assisted by a filter-refine approach using spatial indices.

Spatio-temporal proximity measure

The best of MIA is observed when the algorithm uses a combination of proximity heuristics (SPM and TPM in unison). All comparisons of MIA with the baseline algorithms (Figures 7–12) show that MIA achieves a high accuracy of about 95%, considering all map features. The variation over urban and rural regions is small, but the imputation for rural landscapes is marginally higher at about 97% for all entities. Even with the challenges of street intersections, MIA's accuracy still achieves about 90% with no significant variation across urban and rural regions. MIA's accuracy in spatio-temporal proximity mode consistently outperforms all the five baselines discussed in Section 4.2.

From the results we can conclude that spatial proximity drives the imputation accuracy and the temporal proximity enhances the results. While MIA modeled purely on the TPM underperforms and is computationally expensive, it boosts the accuracy of MIA when used in conjunction with the SPM. There is little or no change to the computational complexity (primarily due to the subset of neighborhood elements on which the TPM acts, having already been filtered by the SPM). MIA with both spatial and temporal proximity heuristics (SPTPM) consistently provides the best results in all our experiments.

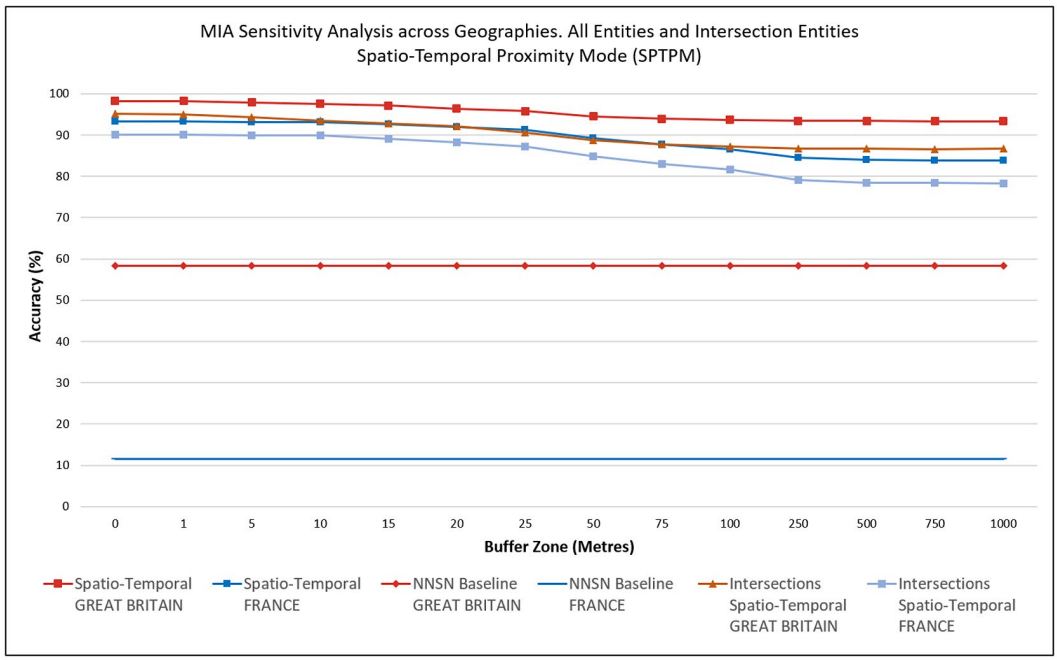


FIGURE 15 MIA across different geographies

5.5.4 | MIA across geographies

MIA's effectiveness and applicability across different geographies was assessed using two additional data sets from OSM, namely Great Britain and France. The assessment was undertaken using the optimal SPTPM variant of the algorithm. The results are shown in Figure 15. The accuracy of MIA for Great Britain varies between 97 and 98.29% at lower buffer intervals of 0–15 m and is about 4% higher as compared to the Swiss data set for the same buffer zone. MIA is consistent in its behavior with reference to its high accuracy across all the buffer zones, maintaining an imputation accuracy of about 93.3% for a 1,000 m buffer. This is significantly higher than the NNSN baseline (about 58.4%). The results exhibit a similar behavior for the France OSM data set, with the imputation accuracy varying between 92.5 and 93.3% at smaller buffer zones and about 84% for the largest buffer zone (similar to the Swiss OSM data set). This is a significant improvement compared to the France NNSN baseline (about 11.47%). The high accuracy of imputation is also observed with intersection entities in the data set. The results validate the algorithm's generic applicability across different geographies.

6 | CONCLUSIONS

Missing Value Imputation techniques have been researched and discussed extensively in the scientific and database community. While imputation techniques are specific to the nature of the problem and the expected outcome of such analysis, they have so far been discussed only in the context of numerical data sets. Furthermore, to the best of our knowledge, there has been no research and discussion of the effectiveness of imputation techniques either for spatial data sets, leveraging the spatio-temporal characteristics unique to these data, or for nominal data in general. Our algorithm has demonstrated high levels of accuracy when imputing nominal values for spatial data sets. Our research serves as the first step in exploring and harnessing the unique spatio-temporal

characteristics of spatial data for nominal imputations, and thus provides a significant contribution and direction to spatial data cleansing techniques.

Data quality issues that drive location-based services such as geocoding systems remain a major impediment to critical services such as emergency response and public utility services. This is a pressing issue in both the global North, as well as in the global South, in countries such as Uganda, Ghana, and Costa Rica (Leslie, 2012; Matthews, 2016; Tamale, 2014). MIA, through its effective imputation support for nominal values (illustrated using the case study of an OSM address attribute, *Street Name*), strives to address an important need in VGI data cleansing, beyond pure error detection. In future work, we will test the effectiveness of harnessing spatio-temporal characteristics of data in the MIA framework to impute other types of values, such as ordinal values (e.g., street numbers) in VGI data.

7 | FUTURE WORK

The effectiveness of spatial characteristics and measures in addressing data cleansing challenges has been illustrated using MIA, with a focus on missing value imputation. These measures have also served to address spatial data integration challenges, as illustrated for a case study in Majic, Winter, and Tomko (2017). Other areas, discussed in Section 2.1, are also significant pain points, such as Entity Matching (Hernández & Stolfo, 1998; Rahm & Do, 2000). The challenges are more pronounced for entities with limited data quality and consistency, more applicable to VGI data. Several Entity Matching frameworks have been discussed in Köpcke and Rahm (2010), and it is evident that the dialectic mechanisms behind these frameworks hinge on semantic entity types that are purely attribute oriented, and do not handle spatial data. As a future work, the richness of spatial characteristics, coupled with spatial reasoning (Cohn & Renz, 2008) could be exploited to identify and address Entity Matching issues in VGI data.

Finally, the effectiveness of MIA has been assessed on a data set created from crisp and determinate spatial objects, wherein the curation of data elements is driven on an implicit assumption of being able to precisely determine the extent and boundary of regions, the position of point geometries, and the placement of lines. In addition, the advantage of our method is that it is independent of feature engineering (where the shape of the geometry could play a major role in determining the final result). Hence the algorithm would generalize much better. In the future it would also be interesting to test the applicability of the imputation framework on spatial entities where the extent and boundaries are indeterminate (such as boundaries of forests and vegetation areas), also referred to as vague spatial data (Pauly & Schneider, 2010; Wang & Hall, 1996).

CONFLICT OF INTEREST

The authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaux; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

ORCID

Rajesh Chittor Sundaram http://orcid.org/0000-0003-1388-1401
Elham Naghizade http://orcid.org/0000-0001-7640-4624
Renata Borovica-Gajic http://orcid.org/0000-0003-3503-4123
Martin Tomko http://orcid.org/0000-0002-5736-4679

REFERENCES

Al-Bakri, M., & Fairbairn, D. (2012). Assessing similarity matching for possible integration of feature classifications of geospatial data from official and informal sources. *International Journal of Geographical Information Science*, *26*, 1437–1456.

Arenas, M., Bertossi, L., & Chomicki, J. (1999). Consistent query answers in inconsistent databases. In *Proceedings of the 18th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, Philadelphia, PA (pp. 68–79). New York, NY: ACM.

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

Ballatore, A., & Arsanjani, J. J. (2019). Placing Wikimapia: An exploratory analysis. *International Journal of Geographical Information Science*, *33*, 1633–1650.

Barron, C., Neis, P., & Zipf, A. (2013). A comprehensive framework for intrinsic OpenStreetMap quality analysis. *Transactions in GIS*, *18*, 877–895.

Beskales, G., Ilyas, I. F., & Golab, L. (2010). Sampling the repairs of functional dependency violations under hard constraints. *Proceedings of VLDB Endowment*, *3*, 197–207.

Bohannon, P., Fan, W., Flaster, M., & Rastogi, R. (2005). A cost-based model and effective heuristic for repairing constraints by value modification. In *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, Baltimore, MD (pp. 143–154). New York, NY: ACM.

Christen, P. (2008). Febrl—An open source data cleaning, deduplication and record linkage system with a graphical user interface. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Las Vegas, NV (pp. 1065–1068). New York, NY: ACM.

Chu, X., Ilyas, I. F., Krishnan, S., & Wang, J. (2016). Data cleaning: Overview and emerging challenges. In *Proceedings of the 2016 International Conference on Management of Data*, San Francisco, CA (pp. 2201–2206). New York, NY: ACM.

Chu, X., Ilyas, I. F. and Paolo, P. (2013). Holistic data cleaning: Putting violations into context. In *Proceedings of the 29th IEEE International Conference on Data Engineering*, Brisbane, Australia (pp. 458–469). Piscataway, NJ: IEEE.

Chu, X., Morcos, J., Ilyas, I. F., Ouzzani, M., Papotti, P., Tang, N., & Ye, Y. (2015). KATARA: A data cleaning system powered by knowledge bases and crowdsourcing. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, Melbourne, Australia (pp. 1247–1261). New York, NY: ACM.

Clementini, E., Di Felice, P., & van Oosterom, P. (1993). A small set of formal topological relationships suitable for end-user interaction. In D. Abel & B. Chin Ooi (Eds.), *Advances in spatial databases* (pp. 277–295). Berlin, Germany: Springer.

Cohn, A. G., & Renz, J. (2008). Qualitative spatial representation and reasoning. In F. van Harmelen, V. Lifschitz, & B. Porter (Eds.), *Handbook of knowledge representation* (pp. 551–596). Amsterdam, the Netherlands: Elsevier.

Cong, G., Fan, W., Geerts, F., Jia, X., & Ma, S. (2007) Improving data quality: Consistency and accuracy. In *Proceedings of the 33rd International Conference on Very Large Data Bases*, Vienna, Austria (pp. 315–326). New York, NY: ACM.

Corcoran, P., Mooney, P., & Bertolotto, M. (2013). Analysing the growth of OpenStreetMap networks. *Spatial Statistics*, *3*, 21–32.

Corcoran, P., Mooney, P., & Winstanley, A. C. (2010) Topological consistent generalization of OpenStreetMap. In M. Haklay, J. Morley, & H. Rahemtulla (Eds.), *Proceedings of the 18th Annual GIS Research UK Conference*, London (pp. 353–358). London, UK: University College London.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*, 21–27.

Dallachiesa, M., Ebaid, A., Eldawy, A., Elmagarmid, A., Ilyas, I. F., Ouzzani, M., & Tang, N. (2013). NADEEF: A commodity data cleaning system. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, New York, NY (pp. 541–552). New York, NY: ACM.

Davidovic, N., Mooney, P., Stoimenov, L., & Minghini, M. (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), 232.

Du, S., Qin, Q., Wang, Q., & Ma, H. (2008). Reasoning about topological relations between regions with broad boundaries. *International Journal of Approximate Reasoning*, *47*, 219–232.

Dymitr, R., & Bogdan, G. (2005). Classifier selection for majority voting. *Information Fusion*, *6*, 63–81.

Elfeky, M. G., Verykios, V. S., & Elmagarmid, A. K. (2002). TAILOR: A record linkage tool box. In *Proceedings of the 18th International Conference on Data Engineering*, San Jose, CA (pp. 17–28). Piscataway, NJ: IEEE.

Engels, J. M., & Diehr, P. (2003). Imputation of missing longitudinal data: A comparison of methods. *Journal of Clinical Epidemiology*, *56*, 968–976.

Fan, H., Zipf, A., Fu, Q., & Neis, P. (2014). Quality assessment for building footprints data on OpenStreetMap. *International Journal of Geographical Information Science*, *28*, 700–719.

Fan, W., Geerts, F., Jia, X., & Kementsietsidis, A. (2008). Conditional functional dependencies for capturing data inconsistencies. *ACM Transactions on Database Systems*, *33*(2), 6.

Farhangfar, A., Kurgan, L. A., & Pedrycz, W. (2004). Experimental analysis of methods for imputation of missing values in databases. *Proceedings of SPIE*, *5421*, 172–182.

Farhangfar, A., Kurgan, L. A., & Pedrycz, W. (2007). A novel framework for imputation of missing values in databases. *IEEE Transactions on Systems, Man, & Cybernetics—Part A: Systems & Humans*, *37*(5), 692–709.

Fogliaroni, P., Bucher, D., Jankovic, N., & Giannopoulos, I. (2018) Intersections of our world. In *Proceedings of the 10th International Conference on Geographic Information Science*, Melbourne, Australia (pp. 3:1–3:15). Wadern, Germany: Dagstuhl Publishing.

Galhardas, H., Florescu, D., Shasha, D., & Simon, E. (2000). AJAX: An extensible data cleaning tool. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, Dallas, TX (p. 590). New York, NY: ACM.

Galhardas, H., Florescu, D., Shasha, D., Simon, E., & Saita, C.-A. (2001). Declarative data cleaning: Language, model, and algorithms. In P. M. G. Apers, P. Atzeni, S. A. Ceri, S. Paraboschi, K. Ramamohanarao, & R. T. Snodgrass (Eds.), *Proceedings of the 27th International Conference on Very Large Data Bases*, Rome, Italy (pp. 371–380). San Francisco, CA: Morgan Kaufmann.

Girres, J.-F., & Touya, G. (2010). Quality assessment of the French OpenStreetMap dataset. *Transactions in GIS*, *14*, 435–459.

Goodchild, M. F. (2007). Citizens as voluntary sensors: Spatial data infrastructure in the World of Web 2.0. *International Journal of Spatial Data Infrastructures Research*, *2*, 24–32.

Goodchild, M. F., & Glennon, J. A. (2010). Crowdsourcing geographic information for disaster response: A research frontier. *International Journal of Digital Earth*, *3*, 231–241.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549–576.

Guttman, A. (1984). R-trees: A dynamic index structure for spatial searching. *ACM SIGMOD Record*, *14*, 47–57.

Haklay, M. (2010). How good is OpenStreetMap information? A comparative study of OpenStreetMap and Ordnance Survey datasets for London and the rest of England. *Environment & Planning B*, *37*, 682–703.

Hashemi, P., & Ali Abbaspour, R. (2015). Assessment of logical consistency in OpenStreetMap based on the spatial similarity concept. In J. Jokar Arsanjani, A. Zipf, P. Mooney, & M. Helbich (Eds.), *OpenStreetMap in GIScience* (Lecture Notes in Geoinformation and Cartography, pp. 19–36). Cham, Switzerland: Springer.

Hawkins, S., He, H., Williams, G. J., & Baxter, R. A. (2002). Outlier detection using replicator neural networks. In Y. Kambayashi & W. Winiwarter (Eds.), *Proceedings of the Fourth International Conference on Data Warehousing and Knowledge Discovery*, London, UK (pp. 170–180). Berlin, Germany: Springer.

Hernández, M. A., & Stolfo, S. J. (1998). Real-world data is dirty: Data cleansing and the merge/purge problem. *Data Mining & Knowledge Discovery*, *2*, 9–37.

Ilyas, I. F., & Chu, X. (2015). Trends in cleaning relational data: Consistency and deduplication. *Foundations & Trends in Databases*, *5*, 281–393.

ISO. (2013). *ISO 19157:2013: Geographic information—Data quality standard*. Geneva, Switzerland: International Organization for Standardization.

Jackson, S. P., Mullen, W., Agouris, P., Crooks, A., Croitoru, A., & Stefanidis, A. (2013). Assessing completeness and spatial error of features in volunteered geographic information. *ISPRS International Journal of Geo-Information*, *2*, 507–530.

Josse, J., & Husson, F. (2016). missMDA: A package for handling missing values in multivariate data analysis. *Journal of Statistical Software*, *70*, 1–31.

Kandel, S., Paepcke, A., Hellerstein, J., & Heer, J. (2011). Wrangler: Interactive visual specification of data transformation scripts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Vancouver, Canada (pp. 3363–3372). New York, NY: ACM.

Kim, W., Choi, B.-J., Hong, E.-K., Kim, S.-K., & Lee, D. (2003). A taxonomy of dirty data. *Data Mining & Knowledge Discovery*, *7*, 81–99.

Kolahi, S., & Lakshmanan, L. V. S. (2009). On approximating optimum repairs for functional dependency violations. In *Proceedings of the 12th International Conference on Database Theory*, St. Petersburg, Russia (pp. 53–62). New York, NY: ACM.

Köpcke, H. and Rahm, E. (2010) Frameworks for entity matching: A comparison. *Data & Knowledge Engineering*, *69*, 197–210.

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. In I. Maglogiannis, K. Karpouzis, M. Wallace, & J. Soldatos (Eds.), *Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real World AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies* (pp. 3–24). Amsterdam, the Netherlands: IOS Press.

Kurgan, L. A., Cios, K. J., Sontag, M. K., & Accurso, F. J. (2005). Mining the cystic fibrosis data. In J. Zurada & M. Kantardzic (Eds.), *Next generation of data-mining applications* (pp. 415–444). Piscataway, NJ: IEEE Press.

Lakshminarayan, K., Harp, S. A., & Samad, T. (1999). Imputation of missing data in industrial databases. *Applied Intelligence*, *11*, 259–275.

Leslie, J. (2012) *When getting directions, it helps to know where the fig tree was: Costa Rica addresses its lack of street names*. Retrieved from https://www.wsj.com/articles/SB10001424052702304870304577489094121477570

Lewis, J. A., Dube, M. P., & Egenhofer, M. J. (2013). The topology of spatial scenes in $\mathbb{R}^2$. In T. Tenbrink, J. Stell, A. Galton, & Z. Wood (Eds.), *Spatial information theory: 11th International Conference, COSIT 2013*, Scarborough, UK (Lecture Notes in Computer Science, Vol. *8116*, pp. 494–515). Cham, Switzerland: Springer.

Little, R. J. A. (1992). Regression with missing X's: A review. *Journal of the American Statistical Association*, *87*, 1227–1237.

Ludwig, I., Voss, A., & Krause-Traudes, M. (2011). A comparison of the street networks of Navteq and OSM in Germany. In S. Geertman, W. Reinhardt, & F. Toppen (Eds.), *Advancing geoinformation science for a changing world* (Lecture Notes in Geoinformation and Cartography, pp. 65–84). Berlin, Germany: Springer.

Mackworth, A. K. (1977). Consistency in networks of relations. *Artificial Intelligence*, *8*, 99–118.

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

Majic, I., Winter, S., & Tomko, M. (2017). Finding equivalent keys in OpenStreetMap: Semantic similarity computation based on extensional definitions. In *Proceedings of the First Workshop on Artificial Intelligence and Deep Learning for Geographic Knowledge Discovery*, Los Angeles, CA (pp. 24–32). New York, NY: ACM.

Matthews, C. (2016) Finding your way in a country without street addresses. *BBC News*, February 1. Retrieved from https://www.bbc.com/news/world-africa-35385636

Mayfield, C., Neville, J., & Prabhakar, S. (2010). ERACER: A database approach for statistical inference and data cleaning. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, Indianapolis, IN (pp. 75–86). New York, NY: ACM.

Mooney, P., & Corcoran, P. (2012). The annotation process in OpenStreetMap. *Transactions in GIS*, *16*, 561–579.

Mooney, P., Corcoran, P., & Winstanley, A. C. (2010). Towards quality metrics for OpenStreetMap. In *Proceedings of the 18th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, San Jose, CA (pp. 514–517). New York, NY: ACM.

Müller, H., & Freytag, J. (2005). *Problems, methods, and challenges in comprehensive data cleansing*. Berlin, Germany: Humboldt University of Berlin.

Neis, P., Zielstra, D., & Zipf, A. (2012). The street network evolution of crowdsourced maps: OpenStreetMap in Germany 2007–2011. *Future Internet*, *4*, 1–21.

Pauly, A., & Schneider, M. (2010). VASA: An algebra for vague spatial data in databases. *Information Systems*, *35*, 111–138.

Prasad, K. H., Faruquie, T. A., Joshi, S., Chaturvedi, S., Subramaniam, L. V., & Mohania, M. (2011). Data cleansing techniques for large enterprise datasets. In *Proceedings of the Annual SRII Global Conference*, San Jose, CA (pp. 135–144). Piscataway, NJ: IEEE.

Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin*, *23*, 3–13.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592.

Sarle, W. S. (1998). Prediction with missing inputs. *Proceedings of JCIS*, *2*, 399–402.

Scheffer, J. (2002). Dealing with missing data. *Biometrika*, *3*, 153–160.

Schmitz, S., Pascal, N., & Alexander, Z. (2008). New applications based on collaborative geodata: The case of routing. In *Proceedings of the 28th INCA International Congress on Collaborative Mapping and Space Technology*, Gandhinagar, India. Hyderabad, India: Indian National Cartographic Association.

Shrive, F. M., Stuart, H., Quan, H., & Ghali, W. A. (2006). Dealing with missing data in a multi-question depression scale: A comparison of imputation methods. *BMC Medical Research Methodology*, *6*, 57.

Simoudis, E., Livezey, B., & Kerber, R. (1995). Using recon for data cleaning. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, Montreal, Canada (pp. 282–287). Menlo Park, CA: AAAI.

Song, Q., & Shepperd, M. (2007). Missing data imputation techniques. *International Journal of Business Intelligence & Data Mining*, *2*, 261–291.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 677–680.

Sung, S. Y., Li, Z., & Sun, P. (2002). A fast filtering scheme for large database cleansing. In *Proceedings of the 11th International Conference on Information and Knowledge Management*, McLean, VA (pp. 76–83). New York, NY: ACM.

Tamale, A. K. (2014). *Street naming and address is essential for Kampala City*. Retrieved from https://www.monitor.co.ug/OpEd/Commentary/Kampala-requires-functional-street-addressingsystem/689364-4340444-9vue3uz/index.html-naming-and-address-is-essential-for-kampala-city/

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

Vach, W. (1994). *Logistic regression with missing values in the covariates* (Lecture Notes in Statistics, Vol. *85*). New York, NY: Springer.

Van Oort, P., & Bregt, A. (2005). Do users ignore spatial data quality? A decision-theoretic perspective. *Risk Analysis*, *25*, 1599–1610.

Vyron, A., & Andriani, S. (2015). Measures and indicators of VGI quality: An overview. *ISPRS Annals of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, *2*, 345–351.

Wang, F., & Hall, G. B. (1996). Fuzzy representation of geographical boundaries in GIS. *International Journal of Geographical Information Systems*, *10*, 573–590.

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

Williams, J. (1997). Tools for traveling data, DBMS and internet systems. *DBMS Magazine*, *10*(7), 69–76.

Wolhuter, K. M. (2015). *Geometric design of roads handbook*. Boca Raton, FL: CRC Press.

Zandbergen, P. A. (2008). A comparison of address point, parcel and street geocoding techniques. *Computers, Environment & Urban Systems*, *32*, 214–232.

Zhang, W. (2000). Association-based multiple imputation in multivariate datasets: A summary. In *Proceedings of the 16th International Conference on Data Engineering*, San Diego, CA. Piscataway, NJ: IEEE.

Zielstra, D., & Hochmair, H. H. (2011). Comparative study of pedestrian accessibility to transit stations using free and proprietary network data. *Transportation Research Record*, *2217*, 145–152.

Zielstra, D., & Zipf, A. (2010). A comparative study of proprietary geodata and volunteered geographic information for Germany. In M. Painho, M. Y. Santos, & H. Pundt (Eds.), *Geospatial thinking: Proceedings of the 13th AGILE International Conference on Geographic Information Science*, Guimarães, Portugal. AGILE. ISBN 9789892019536.

How to cite this article: Chittor Sundaram R, Naghizade E, Borovica-Gajic R, Tomko M. Harnessing spatio-temporal patterns in data for nominal attribute imputation. *Transactions in GIS*. 2020;00:1–32. https://doi.org/10.1111/tgis.12617