DOI: 10.1111/tgis.12694

RESEARCH ARTICLE

Transactions in GIS WILEY

Regional variations of context-based association rules in OpenStreetMap

Christina Ludwig[1] | Sascha Fendrich[2] | Alexander Zipf[1,2]

[1]GIScience Research Group, Institute of Geography, Heidelberg University, Heidelberg, Germany

[2]Heidelberg Institute for Geoinformation Technology (HeiGIT), Heidelberg, Germany

Correspondence
Christina Ludwig, GIScience Research Group, Institute of Geography, Heidelberg University, Im Neuenheimer Feld 348, Heidelberg 69120, Germany.
Email: christina.ludwig@uni-heidelberg.de

Abstract

As a user-generated map of the whole world, OpenStreetMap (OSM) provides valuable information about the natural and built environment. However, the spatial heterogeneity of the data due to cultural differences and the spatially varying mapping process makes the extraction of reliable information difficult. This study investigates the variability of association rules extracted from OSM across different geographic regions and depending on different context variables, such as the number of OSM mappers. The focus of this study is the spatial co-occurrence of OSM tags mapped inside of parks within eight different cities. Without considering any context variable, most association rules were very region-specific without any rule being valid across all cities. Limiting the association rule analysis to parks based on specific context variables increased the number of rules which are applicable across multiple cities. Furthermore, additional region-specific association rules emerged. The most important context variables were found to be the number of features mapped inside the park, the number of tags and the park size. These results suggest that the mapping process has a significant influence on the emergence of association rules within user-generated data. Therefore, this subject needs further investigation to enable effective usage of OSM data across different cultural realms.

1 | INTRODUCTION

OpenStreetMap (OSM) is a collaborative mapping project which aims to create a digital map of the world. Anyone can contribute to the project by mapping different geospatial objects such as buildings, roads or points of interest (POIs). The data is licensed under the Open Data Commons Open Database license, which makes OSM one of the most comprehensive open data sources for information about the built and natural environment on a global scale.

However, applying algorithms on OSM data across different geographical regions is often not feasible in a meaningful way due to the spatial heterogeneity of the data. Obviously, to a large extent this is due to cultural or climatic differences which influence the natural environment and the way cities are built. But in the case of a collaborative mapping project with more than 1 million contributors (OpenStreetMap contributors, 2020a), the representation of certain places within OSM also depends on how the local mapping process unfolds. Distinguishing between culturally induced differences in the built environment and those that are due to locally varying mapping practices is not always possible, but it is crucial for a better understanding of the data set and its reliability.

Due to its nature as a collaborative mapping project, the extraction of information and the assessment of its reliability is difficult, since local mapping practices and overall mapping activity vary across regions (Neis, Zielstra, & Zipf, 2013). Within OSM, mappers describe the properties of an object using key–value pairs called tags (e.g., *highway=residential*). Mapping guidelines explaining the meaning and usage of these tags are openly discussed within the OSM community in forums. Once a new tag is accepted through a vote among the OSM mappers, it is added to the OSM Wiki (OpenStreetMap contributors, 2020b). Although OSM users are expected to adhere to these guidelines, deviations from them are frequently found in the data (Mocnik, Zipf, & Raifer, 2017). These might be due to misconceptions of inexperienced mappers or diverging conceptualizations of certain places between the local mappers and the OSM community's mainstream which is sometimes biased toward Western standards (Ludwig & Zipf, 2019). In other cases, newly introduced tags are often added to the documentation only after they have already been widely adopted and applied within the data by the OSM community (Mocnik, Zipf, & Raifer, 2017) Becoming aware of such implicit patterns introduced by the mapping process is necessary for a better understanding of the data set and to ensure the right conclusions are drawn from it. Therefore, the application of methods from knowledge discovery and databases (KDD; Piateski & Frawley, 1991) on OSM data has become an important research field.

In geospatial data, spatial co-occurrence patterns between certain types of objects are an important source of knowledge. They describe how the presence of certain object types influences the probability of occurrence of other object types. For example, a gas station is usually located next to a road. Negative co-occurrence patterns can be equally informative (e.g., benches are never located in open water). Several studies have applied methods from KDD such as association rule mining to discover spatial relationships within large databases (Bahrdt, Funke, Gelhausen, & Storandt, 2017). In the context of OSM, such co-occurrence rules may be used within data quality assessment by identifying logical inconsistencies (Mocnik et al., 2018) and have already been applied within tag recommendation systems (Kashian, Rajabifard, Richter, & Chen, 2019; Vandecasteele & Devillers, 2015). However, Kashian et al. (2019) have mentioned issues when mining association rules from OSM due to its spatial heterogeneity.

Mining association rules from heterogeneous data is difficult, since universally applicable rules are harder to detect due to the variability of the data. When working with global data sets, region-specific rules emerge due to the culturally shaped built environment. In the case of user-generated data, spatial heterogeneity is additionally often caused by locally varying mapping processes leading to data quality issues such as low levels of completeness or conceptual consistency (Ballatore & Zipf, 2015). In addition, the cultural background of the mappers will influence what is represented in the data. When using co-occurrence patterns for tag recommendation systems or data quality assessment, both cultural and data quality aspects are important to consider (Ali, Sirilertworakul, Zipf, & Mobasheri, 2016; Kashian et al., 2019). Yet, to date no study has investigated the variability of co-occurrence patterns in OSM data depending on cultural and map production contexts.

A case study for mining association rules from OSM on a global scale is the extraction of reliable information about public, urban green spaces such as parks and semi-natural areas. Due to their positive influence on the urban climate and well-being of city dwellers (Tost et al., 2019), information about their location and the amenities

they provide is very valuable both to urban planners and citizens. Still, comprehensive and open data sets on urban green spaces are generally scarce. Cities and municipalities usually only have information about the green spaces they own and maintain, but often lack information about privately owned but publicly accessible green spaces such as playgrounds within apartment blocks. OSM might be able to fill this void provided that reliable information can be extracted across different regions and data quality contexts.

Our study investigates the variance of association rules between OSM tags across multiple cities and depending on different context variables, such as the size of the park or the number of active OSM mappers. The analysis is focused on parks and the physical structures and amenities mapped within them such as paths, benches and playgrounds. Association rule mining is performed separately for eight different cities to identify region-specific as well as universally applicable patterns. The influences of different context variables on the strength of the association rules are investigated using context-based association rule mining.

The remainder of the article is structured as follows. Section 2 gives a brief summary of studies related to association rule mining in the context of OSM data. Section 3 gives a more detailed description of the OSM project and introduces the methods applied for context-based association rule analysis. Section 4 begins with an exploratory data analysis of the parks extracted from OSM. Then the results of the general association rule analysis using all parks are presented followed by the results of the context-based association rule analysis. A discussion follows in Section 5 and Section 6 concludes. The source code and data used to perform this analysis can be found at https://doi.org/10.5281/zenodo.4056680.

2 | RELATED WORK

Association rule mining was first introduced by Agrawal, Imieliński, and Swami (1993) for the purpose of discovering relationships between objects in large databases. Schmitz, Hotho, Jäschke, and Stumme (2006) were the first to apply it to the analysis of folksonomies for ontology learning and building tag recommender systems. The application of association rule mining in the spatial domain was first proposed by Koperski and Han (1995). The goal of this study was to identify pairs of objects which frequently appear close together or are connected through other spatial predicates such as topological or directional relationships. Since then, association rule mining has been further extended to the discovery of spatio-temporal relationships within several other studies (Mennis & Liu, 2005).

Within the context of OSM, there have been a few studies investigating co-occurrence patterns of OSM keys and tags. Davidovic, Mooney, Stoimenov, and Minghini (2016) analyzed the suggested tag combinations described within the OSM Wiki for compliance in the OSM data. They found that the tagging is often not complete and varies considerably across different cities. Such conceptual inconsistencies can cause substantial data quality issues (Ballatore & Zipf, 2015). Using association rule mining, Kinas (2018) identified association rules between OSM tags attached to the same feature and explored the meaning of exceptions to these rules in regard to different mapping practices.

There have also been studies exploring spatial co-occurrence patterns in OSM. Mülligann, Janowicz, Ye, and Lee (2011) developed a spatial-semantic interaction model to analyze the semantic and spatial co-occurrences of different feature types in OSM. Kashian et al. (2019) used an adapted form of spatial association rule mining to extract spatial coexistence patterns between different POI types in OSM. The results were integrated within a tag recommender system to quantify the plausibility of newly created POIs. Within this study, the authors also briefly mention issues related to the spatial heterogeneity of the data. They observed that association rules derived from OSM data differed considerably across space due to changes in urban design, varying levels of completeness or disparate tag usages, which is why they derived association rules for all cities separately. In addition, they observed pattern stability issues within cities with scarce data. A detailed analysis of these observations was, however, not given within this study.

This issue of spatial heterogeneity during spatial association rule mining has been addressed by Sha, Tan, and Bai (2015). They proposed a quadtree-based framework to mine localized association rules by partitioning the study region in multiple sub-areas for which association rules were derived separately. In this way, the spatial variability of the association rules became apparent and could be analyzed. Tang, Chen, and Hu (2008) proposed the context-based market basket analysis, which enables the derivation of association rules from transactional data of multiple stores for different regions and time periods. Shaheen, Shahbaz, and Guergachi (2013) included context variables such as air temperature into the association rule analysis and showed that context information does influence the accuracy of association rules.

3 | DATA AND METHODS

3.1 | OpenStreetMap

As already stated in Section 1, the properties of an OSM object are mapped using key–value pairs called tags (e.g., a residential street would be tagged as *highway=residential*). Each object in OSM may contain one or multiple tags with different keys (e.g., the name of a road can be mapped using an additional tag with the key *name*). The data structures used to represent the spatial properties of objects are nodes (point geometries) and ways (line or polygon geometries). A relation groups nodes, ways or other relations into a coherent object (e.g., a bus route is usually mapped as a relation containing several ways which are tagged with the key *highway* and may also include nodes representing the bus stops).

3.2 | Case study

The focus of this study is the analysis of association rules between OSM tags mapped within public urban green spaces. There are several tags in OSM to map different types of green spaces such as *leisure=park*, *leisure=garden* and *landuse=grass*. According to the OSM Wiki (OpenStreetMap contributors, 2020b), the tag *leisure=park* denotes vegetated areas for recreational use, the tag *leisure=garden* signifies planned spaces for the display, cultivation and enjoyment of plants and the tag *landuse=grass* is meant to be used for smaller areas with managed grass usually located along streets but without recreational services. Still, despite these definitions, the usage of those tags in the OSM data is not always consistent with these guidelines (Ali, Schmid, Al-Salman, & Kauppinen, 2014). Furthermore, the different types of green spaces will also yield different association rules, which would make regional differences between cities harder to detect when all of them are considered at once. Therefore, only one type of green space was considered within this study. The focus was put on features with the tag *leisure=park*, because they usually contain more objects than other green space types and therefore yield more statistically robust association rules. In the remainder of this article, the word "park" is used to refer to objects mapped in OSM using the tag *leisure=park* and not to actual parks in the real world unless stated otherwise.

Parks in OSM are usually mapped as polygonal representations using ways or relations. Inside them different physical structures (e.g., pathways, ponds) and amenities (e.g., benches, playgrounds) may be mapped as nodes, ways or relations. Each of these features will have one or more tags attached to them such as *highway=path* or *amenity=bench*. The aim of this case study is to identify association rules between OSM tags which frequently occur together inside a park independently of whether they are attached to the same OSM feature or not.

The regional variations of association rules between OSM tags within parks were analyzed by comparing eight cities located within different cultural realms (Table 1). These cities were selected to facilitate the analysis of regional differences at a national, continental and global level. In regard to data quality, only cities were selected which contain a sufficient number of parks in OSM to ensure meaningful and statistically robust results. The main

goal of the analysis was not to provide a comprehensive analysis of all parks within each city, but rather to take a large enough sample of parks from each city to facilitate regional comparisons. Therefore, the study area for each city was defined by querying the center coordinates of the city using the Nominatim geocoding service and placing a 20 by 20 km bounding box around it. This size was chosen because it ensures the extraction of a sufficient number of park features within each city to ensure statistically robust results from the association rule analysis.

The variability of association rules is also analyzed in regard to the following context variables:

Area: The size of the park in hectares.
Building density: The areal density of buildings within a 200 m buffer around the park, excluding the park itself.
Feature count: The number of features fully or partially located inside the park.
Days since creation: The number of days since the creation of the park in OSM.
Number of tags: The number of tags of the park.
Number of changes: The number of tag and geometry changes of the park.
Current version: The current version number of the park.
Inner user density: The density of users that have been active inside the park.
Inner user count: The number of users that have been active inside the park.
Outer user density: The density of users that have been active within a 200 m buffer around the park, excluding the park itself.
Outer user count: The number of users that have been active within a 200 m buffer around the park, excluding the park itself.
Random context variable: A random variable was generated to test the analysis for the significance of the results.

The area of a park was included within the context variables, since the size of a park might have an influence on the types of amenities it provides. Building density was added as an indicator for urbanity in order to investigate whether different association rules appear specifically within urban or rural areas. The remaining context variables are indicators for mapping activity in OSM and have already been applied within previous studies (Barron, Neis, & Zipf, 2014; Keßler & de Groot, 2013; Mooney & Corcoran, 2012; Neis et al., 2013).

During the context-based association rule analysis (see Section 3.5), subsets as small as 100 parks are created to derive context-based association rules. With such small sample sizes, however, occasionally strong association rules might appear just by chance. To account for this influence in the interpretation of the results, a random context variable based on a uniform distribution was generated and also included in the analysis.

TABLE 1 Study sites

City	Country	Number of parks in OSM
Dresden	Germany	467
Berlin	Germany	1,390
London	Great Britain	1,578
Tel Aviv	Israel	1,168
Tokyo	Japan	2,670
Osaka	Japan	1,022
New York	USA	893
Vancouver	Canada	552

3.3 | Data extraction

All data extraction for this study was performed using the OpenStreetMap History Database (OSHDB; Raifer et al., 2019). For each city, all features (including nodes, ways and relations) containing the tag *leisure=park* were extracted from the OSHDB on December 10, 2019 (Table 1). Subsequently, the OSM tags attached to nodes or ways located fully or partially inside each park were extracted. Relations were not considered in this step, because they usually spread over larger areas and do not belong to one specific park. The analysis was limited to frequently occurring OSM tags which specifically describe amenities and objects found within parks. Therefore, general tags such as *name* or *description* were not considered. Based on the retrieved OSM data of parks, the most frequently occurring OSM keys were identified. From this selection, tags containing the following keys were included in the analysis: *amenity*, *building*, *fitness_station*, *shop*, *highway*, *leisure*, *landuse*, *natural*, *playground*, *sport*, *surface*, *tourism*, *water* and *waterway*.

The context variables described in Section 3.2 were calculated for each park. Data extraction for this step was also done using the OSHDB. The user count and density variables were calculated for the period from November 1, 2007 until December 10, 2019. In order to reduce the computational burden, only single nodes (e.g., POIs) were considered to get an estimate of the user count and density. The user density variables are given as the number of users per square kilometer.

3.4 | Association rule analysis of OSM data

Association rules between OSM tags were derived using an adapted version of association rule mining as introduced by Agrawal, Imieliński, and Swami (1993). Originally, this method was meant to be applied to two-dimensional data structures consisting of a set of items I and a set of transactions T. Each transaction $t \in T$ contains a subset of the items in I. A popular use case of association rule mining is the market basket analysis, which aims to identify products that are frequently bought together by customers of a store. In this example, transactions represent purchases by customers and items represent products available in a store. Each transaction contains products which have been bought within the respective purchase. Within our case study, OSM tags represent items and parks represent transactions. So a park within which the tags *highway=footway*, *amenity=bench* and *bicycle=yes* occur will be represented as one transaction containing the tags t={*highway=footway*,*amenity=bench*,*bicycle=yes*}.

The first step in association rule mining is the discovery of frequent itemsets within all transactions. A frequent itemset is a set of items which occur frequently together in the same transaction. An itemset may contain multiple items or just one, in which case it is called a singleton itemset. In this study, itemsets consist of OSM tags which frequently appear together within parks. For mining frequent itemsets, we employ the commonly used Apriori algorithm introduced by Agrawal and Srikant (1994).

Based on the frequent itemsets, association rules are derived in the form of $X \rightarrow Y$, where X is an itemset called the antecedent and Y is an itemset called the consequent. For example, the rule {*amenity=cafe*}→{*amenity=toilet*} means that if the park contains a feature with the tag *amenity=cafe*, it is likely to contain a feature with the tag *amenity=toilet* as well. For better readability, the braces marking the itemsets of a rule will be omitted for the remainder of the article. The size of a rule is given by the overall number of items in the rule.

The strength and relevance of an association rule is described by different metrics, most importantly support and confidence. The support can be calculated for both rules and itemsets. It indicates how often a rule or itemset occurs within all transactions. For an itemset X the support is defined as:

$$\mathrm{supp}(X) = \frac{|t \in T; X \in t|}{|T|} \tag{1}$$

For a rule $X \rightarrow Y$, the support is equal to the support of the union of its itemsets X and Y:

$$\text{supp}(X \rightarrow Y) = \text{supp}(X \cup Y) = \frac{|t \in T; X \cup Y \in t|}{|T|} \tag{2}$$

The confidence value, sometimes also referred to as the strength of a rule, indicates how often the consequent tag occurs within transactions, where the antecedent tag is given. It is defined as:

$$\text{conf}(X \rightarrow Y) = \frac{\text{supp}(X \rightarrow Y)}{\text{supp}(X)} \tag{3}$$

As a third metric, the lift value was used in this study to identify interesting rules in the sense that the joint occurrence of the tags is not just due to chance but rather due to a genuine relationship. It is defined as:

$$\text{lift}(X \rightarrow Y) = \frac{\text{supp}(X \rightarrow Y)}{\text{supp}(X) \times \text{supp}(Y)} \tag{4}$$

A lift value of 1.0 means that the co-occurrence of the two itemsets or tags is only due to chance. When lift is greater than 1.0, the two tags occur more frequently together than would be the case if their occurrence were mutually independent. When lift is less than 1.0, the two tags occur less frequently together than would be the case if their occurrence were mutually independent.

3.5 | Context-based association rule analysis

The context-based association rule analysis was performed similarly to previous studies. The set of transactions representing parks was partitioned into multiple subsets based on the context variables. This analysis was performed separately for each context variable, so combinations of two or more context variables were not considered. The partitioning of the set of parks was done recursively, taking the respective median of the context variable of the subset as the threshold. In the first iteration, the whole data set was partitioned into two subsets using the median of the respective context variable as a threshold. In all following iterations, the resulting subsets were partitioned again using the median of the context variable of each subset. The iteration stopped as soon as a subset had reached the minimum number of 100 parks. This parameter was set in order to preserve statistical stability of the results. Subsequently, association rules were calculated separately for each subset and all rules were retained which exceeded the minimum support value of 0.05. So the rule of a subset was only retained if it occurred in at least 5% of all parks within the subset.

The variability of an association rule in dependence of a certain context variable was analyzed by comparing the confidence and lift values of the different subsets. For better visualization, the value range of each subset and its respective confidence and lift values were plotted together with a histogram of the context variable (e.g., Figure 4). Each line in the graph represents a rule derived from a subset. Its length represents the value range of the context variable indicating which parks were included in the subset. Its vertical location indicates the confidence or lift values of the rule derived from this subset.

To analyze the overall influence of the context variables on multiple rules within multiple cities a more comprehensive visualization was created. The Spearman correlation coefficient between the confidence values of all subsets and the mean value of the context variable of all subsets was calculated to quantify whether there was a continuous increase or decrease in confidence of the rule depending on the context variable. Association rules with correlation coefficients higher than 0.9 or lower than −0.9 and an absolute difference between the lowest

and highest confidence value of more than 0.2 were retained, while the remaining rules were excluded for not showing a clear dependence on the context variable.

4 | RESULTS

4.1 | Exploratory data analysis of parks in OSM

The number of parks and the number of tags they contain have a considerable influence on the stability of the association rules. Therefore, an exploratory data analysis was performed prior to association rule mining.

The number of parks extracted from OSM for each city varies considerably, from 467 in Dresden to 2,670 in Tokyo (Table 1). The distribution of park sizes is highly skewed in all cities, spanning from less than 0.1 ha to more than 100 ha (Figure 1). Tokyo seems to contain a particularly large number of very small parks, with 50% of them being smaller than 0.1 ha, while the median size of parks in Vancouver is around 1 ha. A highly skewed distribution is also visible for the number of features mapped within a park, ranging from none to hundreds of features for large parks (Figure 1). Tokyo (21%) and Osaka (22%) show the highest share of parks without any OSM features mapped inside, while this is true for only 2% of parks in Dresden. To some degree, these variations between cities are due to difference in urban design, such as the size of the parks, but some of them might also be influenced by the local mapping process.

Since 2007, the number of parks in OSM has been quite steadily increasing in most cities except for Berlin, where the number of parks decreased between 2014 and 2017 (Figure 2). This is mostly due to tag changes; for example, all animal enclosures within Berlin Zoo used to be tagged with *leisure=park*, which was later changed to *landuse=grassland*. Of all cities, the parks in Berlin have the highest mean version number, suggesting that the mapping process is to a large degree characterized by revisions instead of newly mapped features.

In contrast, Tokyo shows a very strong increase in the number of parks since 2018. The data does not contain any evidence of large data imports or signs of vandalism. Instead, the increasing number of active contributors, especially at the beginning of 2018, suggests an increased mapping activity in regard to parks by the community. This phenomenon could also be related to Pokémon Go players joining OSM as described by Juhász, Hochmair, Qiao, and Novack (2019). Of the newly created parks 64% are smaller than 0.1 ha, while this share is only 41% for parks created before 2018. Furthermore, 25% of the newly created parks do not contain any features, while this

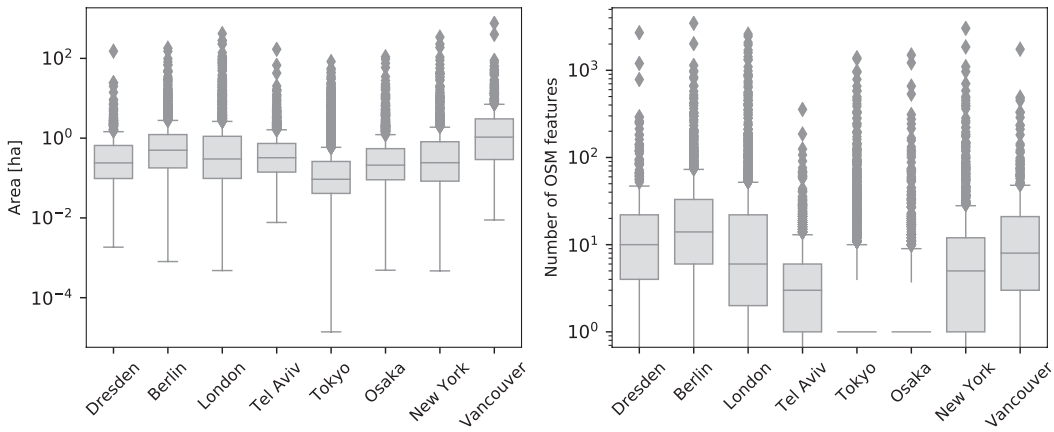

FIGURE 1 Boxplots of the size of park features (left) and the number of features within each park feature (right) extracted from OSM for each city

is true for only 17% among the older parks. Whether these parks do not contain any features in reality or whether they just have not been mapped yet cannot be fully clarified at this writing.

Generally, these results imply that the mapping process in regard to parks is still ongoing in most cities. The example of Tokyo shows that although the temporal evolution of the number of parks indicates a saturation status for most cities, this does not mean that all parks are completely mapped yet.

A correlation analysis between all context variables including parks from all cities was performed using the Spearman correlation coefficient, since most variables are not normally distributed (Table 2). Most correlation coefficients show slightly positive relationships, while those variables which are indicators for user activity (e.g., number of changes, version number, inner user count, feature count) show rather strong positive relationships between each other. They also correlate positively with the age of the park feature (days since creation), suggesting that the longer a park has existed in OSM, the more complete it is in OSM. The size of the park is also slightly positively correlated with the number of days since creation, suggesting that larger parks tend to be mapped earlier than smaller ones. Large parks also show more unique active users and higher numbers of features mapped inside them, which is probably due to both their physical structure (e.g., large parks generally provide more amenities than small ones) and the level of completeness, which is probably higher since more users are active inside them. Highest correlation coefficients appear between the user density and user count variables. However, the inner user density attains very high values for very small parks, leading to an unrealistic representation of user activity.

Comparing the frequencies of individual tags occurring within parks already gives an indication of the regional differences in the representation of parks in OSM (Table 3). The tag *highway=footway* is among the five most frequent tags within every city, although the frequency values vary considerably from 78% in Berlin to 14% in Osaka. This is an indication of a rather low level of completeness of paths mapped within parks in Osaka and, in fact, a comparison of randomly selected parks in OSM with aerial imagery revealed that many parks actually do contain footpaths in reality, which have not been mapped yet. Buildings and playgrounds are mapped quite frequently within parks in most cites, while benches appear especially often in German parks (Dresden and Berlin) and toilets are among the most frequently mapped objects within Japanese parks (Osaka and Tokyo).

4.2 | General association rule analysis

Within the first analysis, association rules were calculated for each city separately using all available parks (Figure 3). Rules were filtered using a minimum support value of 0.05, meaning that rules were only considered in

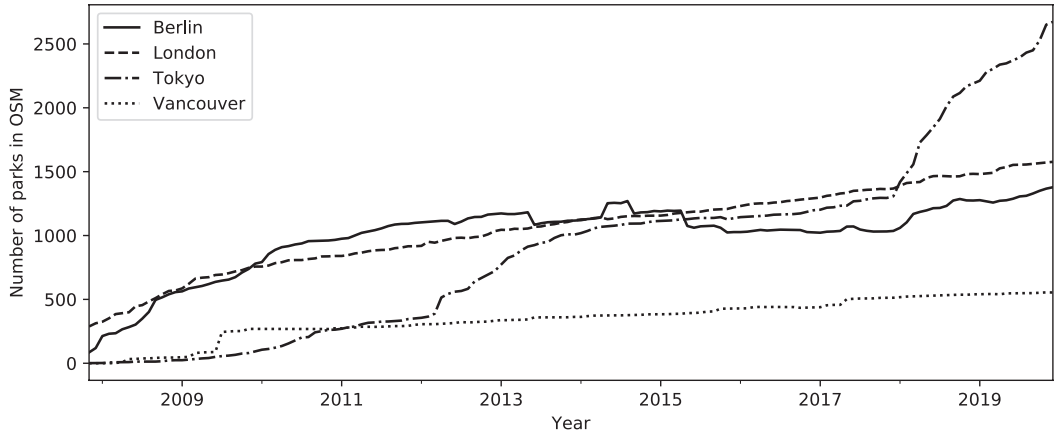

FIGURE 2 Temporal evolution of the number of features with the tag *leisure=park* from November 1, 2007 to December 10, 2019 for each city

TABLE 2 Spearman correlation coefficients between context variables of parks in all cities

	Area	Building density	Days since creation	Number of tags	Number of changes	Version number	Inner user count	Inner user density	Outer user count	Outer user density	Feature count
Area	**1.00**	−0.1	0.34	0.25	0.5	0.5	**0.53**	0.12	0.19	−0.12	**0.67**
Building density	−0.1	**1.00**	0.13	0.15	0.12	0.11	0.12	0.21	0.34	0.39	0.09
Days since creation	0.34	0.13	**1.00**	0.16	**0.6**	0.49	0.28	0.16	0.19	0.11	0.31
Number of tags	0.25	0.15	0.16	**1.00**	0.36	0.4	0.27	0.16	0.12	0.01	0.24
Number of changes	0.5	0.12	**0.6**	0.36	**1.00**	**0.84**	**0.51**	0.29	0.37	0.19	**0.56**
Version number	0.5	0.11	0.49	0.4	**0.84**	**1.00**	**0.51**	0.29	0.38	0.21	**0.57**
Inner user count	**0.53**	0.12	0.28	0.27	**0.51**	**0.51**	**1.00**	**0.82**	0.5	0.31	**0.75**
Inner user density	0.12	0.21	0.16	0.16	0.29	0.29	**0.82**	**1.00**	0.42	0.39	0.5
Outer user count	0.19	0.34	0.19	0.12	0.37	0.38	0.5	0.42	**1.00**	**0.91**	0.49
Outer user density	−0.12	0.39	0.11	0.01	0.19	0.21	0.31	0.39	**0.91**	**1.00**	0.26
Feature count	**0.67**	0.09	0.31	0.24	**0.56**	**0.57**	**0.75**	0.5	0.49	0.26	**1.00**

Note: Correlation coefficients above 0.5 are printed in bold.

TABLE 3 Frequency of OSM tags within parks

	Dresden	Berlin	London	Tel Aviv	Tokyo	Osaka	New York	Vancouver
highway=footway	**0.63**	**0.78**	**0.57**	**0.42**	**0.22**	**0.14**	**0.41**	**0.63**
landuse=residential	**0.6**	**0.58**	**0.52**	**0.31**	0.03	0.01	0.07	**0.28**
building=yes	0.19	0.26	**0.22**	**0.13**	**0.18**	**0.19**	**0.36**	**0.29**
leisure=playground	**0.23**	**0.34**	**0.26**	**0.15**	0.05	0.04	**0.21**	**0.36**
amenity=bench	**0.46**	**0.34**	0.18	0.01	0.06	0.04	0.08	0.17
leisure=pitch	0.08	0.21	0.19	0.04	0.05	**0.08**	**0.26**	**0.3**
highway=path	**0.24**	0.24	0.14	00.1	0.05	0.05	0.04	0.26
highway=steps	0.22	0.23	0.1	0.08	**0.09**	0.05	0.1	0.16
natural=tree	0.18	**0.27**	**0.27**	0.01	0.05	0.04	0.06	0.1
amenity=toilets	0.01	0.05	0.06	0.02	**0.21**	**0.08**	0.12	0.16
amenity=drinking_water	0	0.01	0.04	0.08	**0.12**	0.01	**0.21**	0.14
highway=residential	0.03	0.07	0.07	**0.11**	0.05	0.04	0.06	0.09
highway=unclassified	0	0.01	0.03	0	0.04	**0.07**	0.01	0.01

Note: The selection of tags includes the five most frequent tags of each city, which are highlighted in bold.

the evaluation if they occurred in at least 5% of all parks within a city. Interesting rules were identified by setting a minimum confidence and minimum lift value, which had to be attained in at least one city.

Overall, Osaka, Tel Aviv and Tokyo contain the fewest association rules. This is due to the fact that there are many parks mapped within these cities which contain few or no features (Figure 1). As a result, the minimum support threshold of 5% is rarely reached. While there does not seem to exist any plausible, real-world explanation for some of these rules—for example, in reality the presence of fine gravel does not necessarily imply that there must be a bench nearby (*surface=fine_gravel* → *amenity=bench*)—other rules seem to represent a genuine real-world relationship—for example, if there are steps, there should also be a path leading up to them. However, none of the rules seem to be universally applicable within all cities. The only rule coming close to this is *highway=steps* → *highway=footway* reaching confidence levels above 0.77 in all cities except for Osaka. The lift values for this rule vary between 1.26 in Berlin and 4.11 in Tokyo, suggesting that the presence of the tag *highway=steps* has a much higher impact on the probability of occurrence of the tag *highway=footway* in Tokyo than in other cities. Generally, most of the rules containing the tag *highway=footway* as the consequent show rather low lift values, since this tag generally occurs quite frequently in parks (Table 3). As a result, the probability that the tag *highway=footway* will appear together with any other OSM tag is high by default even if they were independent of each other. So even though this rule is very reliable in almost all cities, as indicated by the high confidence values, it provides a lot more additional information about the presence of the consequent tag in Tokyo, where the lift value is high. As an example, the presence of the tag *highway=steps* increases the probability of occurrence of the tag *highway=footway* from 78 to 99% in Berlin, but in Tokyo the probability increases more than four times from 22 to 91%. So even though a rule might be universally applicable, its relevance may change across cities. Whether this phenomenon is due to low levels of completeness of paths mapped within parks in Tokyo or due to the fact that a lot of the small parks in Tokyo actually do not contain any paths cannot be clarified at this writing.

There are several rules which seem to be both applicable and relevant but only within certain cities in OSM. This is the case for many rules which contain a tag with the key *sport* as the antecedent and the tag *leisure=pitch* as a consequent as well as for the rule *amenity=waste_basket* → *amenity=bench*. These rules are characterized by both high confidence and high lift values in all cities where the minimum support is attained, since the overall

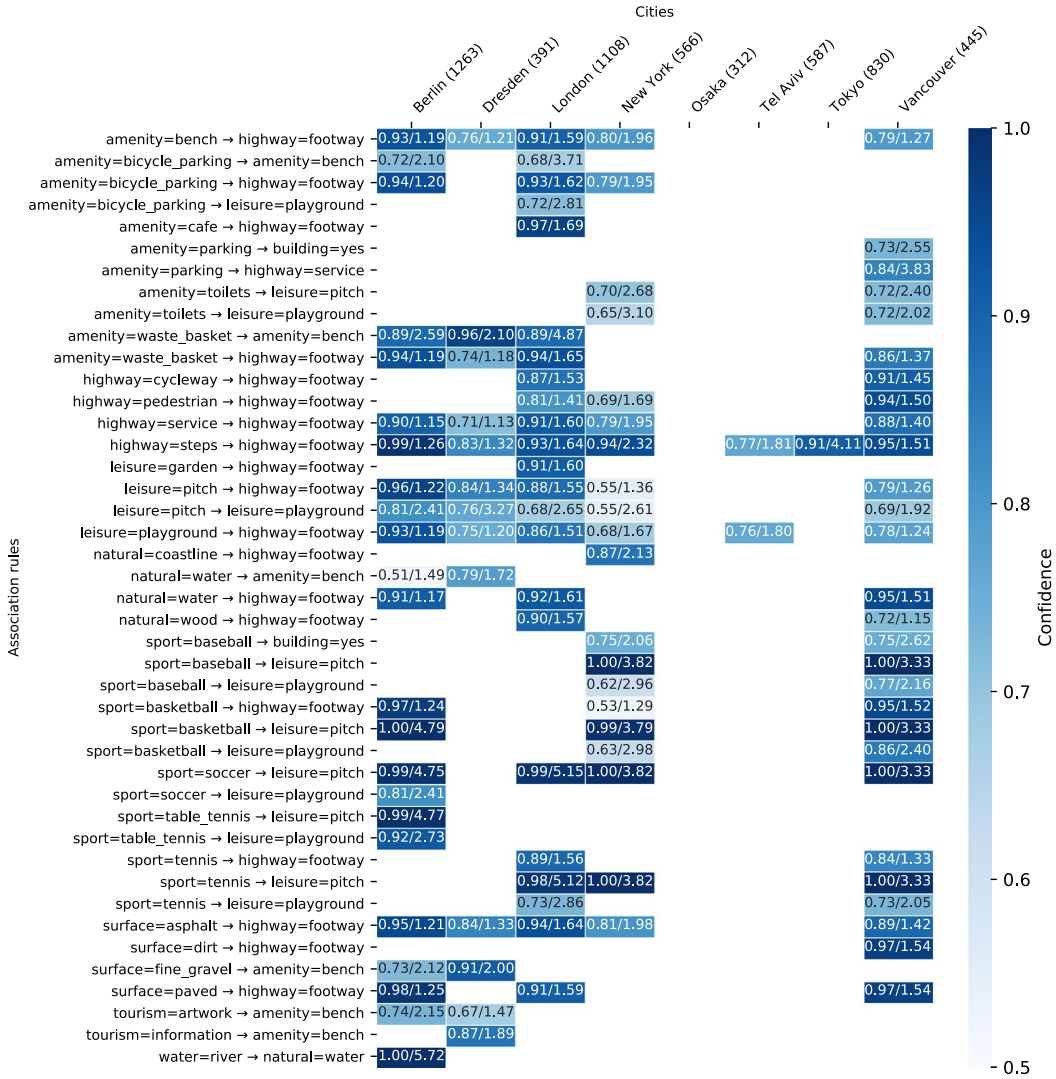

FIGURE 3 Selected association rules between OSM tags based on all parks for each city. The first value indicates confidence, the second value represents lift of the rule. Empty cells mark rules whose support is less than or equal to 0.05. Only rules are shown where confidence is 0.7 or greater and lift is 1.5 or greater in at least one city and confidence is 0.5 or greater and lift is 1.1 or greater in all cities where support is less than or equal to 0.05. The number of parks within each city mapped in OSM is given in parentheses next to their names. For better readability only rules of size 2 are shown

frequency of occurrence of these tags is quite low—for example, the tag *amenity=bench* only occurs in 34% of parks in Berlin (Table 3). All of them seem to represent plausible real-world relationships; however, they do not appear in the data within every city. Whether this is due to cultural influences (e.g., baseball is only played in parks in North American cities) or due to the state of the mapping process cannot be distinguished at this writing.

In addition to rules of size 2 (i.e., rules with one antecedent and one consequent tag) the occurrence of larger rules containing several antecedent and consequent tags was analyzed. Based on the minimum support value of 5%, rules with 3 or more items could be retrieved in all cities but Osaka and Tel Aviv. However, in most cases these large rules did not yield relevant new information in regard to the regional differences of association rules, since

the additional antecedent or consequent tags are often tags which occur very frequently by default (e.g. *highway=path*; see Table 3). For example, in Berlin the rule *amenity=waste_basket* → *amenity=bench* has a confidence of 0.89, while the rule *amenity=waste_basket*, *highway=path* → *amenity=bench* attains a confidence of 0.91. In other cases, similar but less pronounced patterns emerged in the large rules which were already visible in the rules of size 2. For example, in London the rule *sport=tennis* → *leisure=pitch* has a confidence of 0.98, while *sport=tennis* → *leisure=pitch*, *highway=footway* attains a confidence of 0.89. This redundancy in the results hampers the interpretability of the rules in regard to their regional differences. Therefore, the following context-based association rule analysis was limited to rules of size 2. For more details on the association rule analysis, including large rules, refer to the supplementary material and source code provided at https://doi.org/10.5281/zenodo.4056680.

4.3 | Context-based association rule analysis

Within the context-based association rule analysis, the variability of the association rules depending on the context variables was investigated (see Section 3.5). This analysis was again performed separately for each city. The biggest impact of the context variables on the confidence of the association rules can be observed in Berlin, London and Tokyo (Table 4). These cities show the highest number of rules which positively correlated with a change of some of the context variables (e.g., partitioning the parks in Berlin based on the number of tags of the park leads to 33 association rules gaining more than 0.2 in confidence). This is partly due to the fact that these cities contain the highest number of parks in OSM, so more subsets containing at least 100 parks can be generated than for other cities.

For most context variables, parks with high values of a certain context variable lead to increases in association rule strength (e.g., association rules gain in confidence if they are derived from parks with a large number of features mapped). The reverse is true for the inner and outer user density. These seem to yield more association rules with high confidence when user density is low. This seems implausible and might be due to the fact that very small parks attain extremely high user density values, which does not seem to be representative of the mapping activity, but rather an artifact of the small area of the park. The random variable does not show any clear positive or negative influence on the association rules, which supports the significance of the results for the other context variables.

An interesting and frequently occurring phenomenon can be observed when comparing the evolution of the confidence and lift values of a rule across different subsets of parks. As an example, Figure 4 shows association rules derived from different subsets of parks partitioned based on their number of tags for London. A steady increase in confidence along with a steady decrease in lift values can be observed. So, while the rule gains in reliability, the association seems to be increasingly due to chance instead of a strong co-occurrence pattern. As an example, when considering all parks the association rule indicates that the presence of the tag *leisure=pitch* increases the probability of occurrence of the tag *amenity=bench* by a factor of more than 2 (from 18 to 42%). But when only considering parks which contain at least five tags the probability of occurrence of the tag *amenity=bench* on its own is already at 59%. The presence of the tag *leisure=pitch* only increases by a factor of 0.25 to 74%. So the process which causes the co-occurrence of the tags seems to become weaker. Since no external reference data is available, it cannot be clarified at this writing whether this phenomenon is due to a real-world change in co-occurrence patterns within the parks of the subsets or whether this is driven by varying levels of completeness of the data. Still, since this phenomenon can be observed quite frequently across different rules, an influence of the mapping process is likely.

Overall, the most influential context variables on the variability of the association rules are the number of tags of a park feature, its area and the number of features mapped inside a park. The latter two variables are especially important in cities with many empty and small parks on OSM (Tel Aviv, Tokyo and Osaka). This seems plausible, since excluding empty parks from the analysis increases the support of all remaining items in the data set, so that

TABLE 4 Influence of context variables on the association rule strength

Index	Dresden	Berlin	London	Tel Aviv	Tokyo	Osaka	New York	Vancouver
Number of tags	0/0 (6)	**33/0 (1)**	**34/2 (1)**	0/0 (3)	**24/0 (2)**	**1/0 (2)**	**13/5 (1)**	0/0 (9)
Area	0/0 (6)	**21/1 (2)**	21/0 (3)	**2/0 (1)**	20/0 (3)	**2/0 (1)**	**9/0 (2)**	0/0 (9)
Feature count	**1/0 (2)**	**21/0 (2)**	**27/0 (2)**	0/0 (3)	**28/1 (1)**	0/0 (4)	0/0 (8)	1/0 (6)
Inner user count	**1/0 (2)**	17/0 (4)	20/2 (5)	0/0 (3)	7/0 (4)	0/0 (4)	5/0 (4)	3/0 (4)
Outer user count	**2/0 (1)**	7/2 (7)	14/3 (7)	0/1 (3)	7/0 (4)	0/0 (4)	1/0 (6)	3/0 (4)
Number of changes	0/0 (6)	14/2 (5)	16/1 (6)	**2/0 (1)**	6/0 (7)	0/0 (4)	**9/0 (2)**	**4/1 (2)**
Version number	0/0 (6)	14/1 (5)	21/1 (3)	0/0 (3)	7/0 (4)	**1/0 (2)**	3/1 (5)	1/0 (6)
Days since creation	**1/2 (2)**	0/3 (12)	8/0 (8)	0/0 (3)	4/0 (8)	0/0 (4)	1/0 (6)	**4/0 (2)**
Building density	0/0 (6)	1/3 (10)	5/5 (9)	0/0 (3)	1/3 (9)	0/0 (4)	0/2 (8)	**5/6 (1)**
Random	0/0 (6)	1/0 (10)	1/0 (11)	0/0 (3)	1/0 (9)	0/0 (4)	0/0 (8)	1/0 (6)
Inner user density	0/2 (6)	3/9 (8)	0/12 (12)	0/1 (3)	1/25 (9)	0/0 (4)	0/1 (8)	0/0 (9)
Outer user density	**1/1 (2)**	2/9 (9)	4/18 (10)	0/0 (3)	0/2 (12)	0/1 (4)	0/3 (8)	0/3 (9)

Note: Correlation coefficients above 0.5 are printed in bold.

Notes: The first number represents the number of rules which increase in strength with an increase of the context variable. The second number indicates the number of rules with a negative change. The number in parentheses is the rank of the context variable based on its influence on the association strength.

the minimum support threshold is met by more association rules. As a consequence, more region-specific association rules appear when considering only parks with at least three features such as the co-occurrence of the tags *amenity=toilet*, *amenity=drinking_water* and *amenity=clock* in Tokyo (Figure 5). In addition, more rules which are applicable across multiple cities are derived. The rule *highway=steps* → *highway=footway* is now valid within all cities, whereas before it was not valid in Osaka, since the minimum support threshold was not attained. Still, the lift value in Osaka is almost twice as high as in the other cities, suggesting a stronger dependence between the tags *highway=steps* and *highway=footway* within parks in this city. There could be two different reasons for this. First, parks in Osaka do contain fewer footways in reality, and if they do then they often appear together with steps. Secondly, footways and steps within parks are still mapped at a quite low level of completeness, so their true frequencies of occurrence are not yet correctly represented in OSM. Without a comprehensive data quality assessment, it cannot be conclusively clarified which of these scenarios is true.

The area of a park is highly positively correlated with the feature count, therefore it appears as an important context variable as well. Association rules were also derived separately for parks smaller and larger than 0.5 ha. Using this threshold, the data is partitioned into subsets which contain more than 100 parks each, so that stable rules can be derived.

When only deriving association rules for large parks, numerous rules appear which are valid across several cities (Figure 6). A good example of this are rules related to the OSM key *sport*. Compared to the initial association rules derived for all parks, it is now apparent that baseball is also played in parks in Tokyo and Osaka. This rule has been in the data set all along, but it was not apparent until the relevant context variables area or feature count were considered during the association rule analysis.

Another notable observation is that the lift values of rules containing the tag *highway=footway* are approaching 1.0 in some cities. This is due to the fact that the tag *highway=footway* is now so frequent in large parks (e.g., 94% in Berlin), that their co-occurrence with other tags is now mainly driven by chance. This confirms that the observation made for the association rule *leisure=pitch* → *amenity=bench* in Figure 4 is also valid for other rules.

Due to the high correlation between park size and the number of mapped features inside, it cannot be clarified whether this increase in the number of additional rules is due to the fact that large parks generally contain more

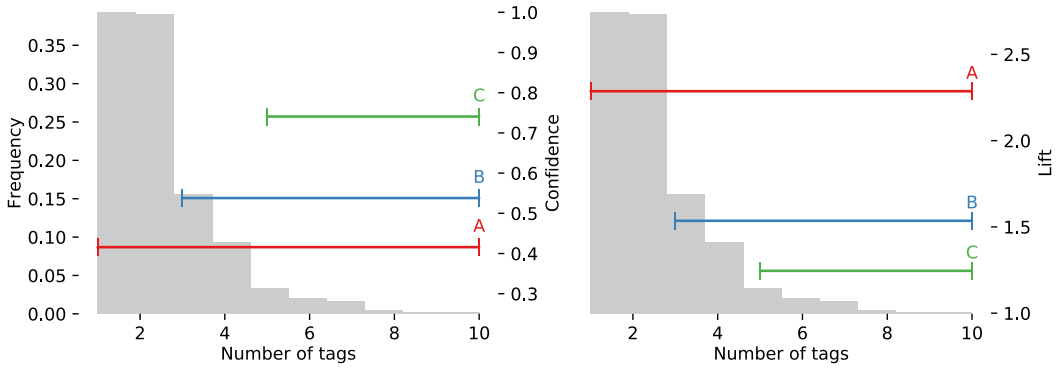

FIGURE 4 Confidence (left) and lift values (right) depending on the number of tags of the parks in London. The association rule *leisure=pitch* → *amenity=bench* was calculated for different subsets of parks (A,B,C) partitioned by the number of tags. The horizontal extent of the lines marks the value range of the subset (e.g., subset C contains only parks with at least five tags). Its vertical position indicates the confidence and lift values of the rule. The histogram shows the distribution of parks regarding their number of tags

objects or whether the level of completeness plays a role as well. However, when analyzing small parks containing at least three *amenity=toilet* and *amenity=clock* in Tokyo (Figure 7). This indicates that the impact of the feature count on the number of new association rules is not just due to the fact that a larger share of large parks is analyzed, but is to some degree also dependent on the level of completeness of the data.

In cities with a low number of empty parks, such as Berlin and London, the number of tags of a park feature seems to have the strongest impact on the association rule strength. In contrast to the feature count, this variable is less correlated with the size of the park. Therefore, it is another indication that the increase in confidence is to some degree connected to the level of completeness of the data. Still, real-world changes in the structure of the parks cannot be ruled out.

5 | DISCUSSION

As expected, our results showed that association rules within parks in OSM vary considerably across regions and only a few rules were detected which apply universally. Some of the cross-regional differences may be traced back to cultural differences. Parks are public spaces which are shaped by the local culture and therefore their physical structure and the amenities they provide differ across regions. So while in reality it is common for many small parks in Tokyo to contain toilets but rarely wastebaskets, the reverse is true for German cities. Such culturally influenced rules were also detected in our study (Figure 5) and are naturally only valid within a certain region.

In addition, the context-based association rule analysis showed that it is important to differentiate parks based on their size, since large parks are likely to provide different amenities than small ones. Therefore, many more rules emerged which apply in multiple cities, when only large parks were included in the analysis. Regarding different types of parks, it should also be noted that some of the variability in the association rules derived for different cities might also be due to varying tag usages for urban green spaces (Ali et al., 2014). While mappers in one city might prefer the tag *landuse=grass* for very small green spaces, similar green spaces might be tagged using *leisure=park* in other cities. The influence of this phenomenon on the association rules could not be answered in this study, since only features with the tag *leisure=park* were investigated.

The results of this study might also be of interest to the OSM community, since they give an indication of the regional variability of the representation of parks in OSM, which is not yet reflected in the OSM Wiki. Further association rule analyses on OSM tags within other kinds of green spaces (e.g., *landuse=grass*, *leisure=garden*) or

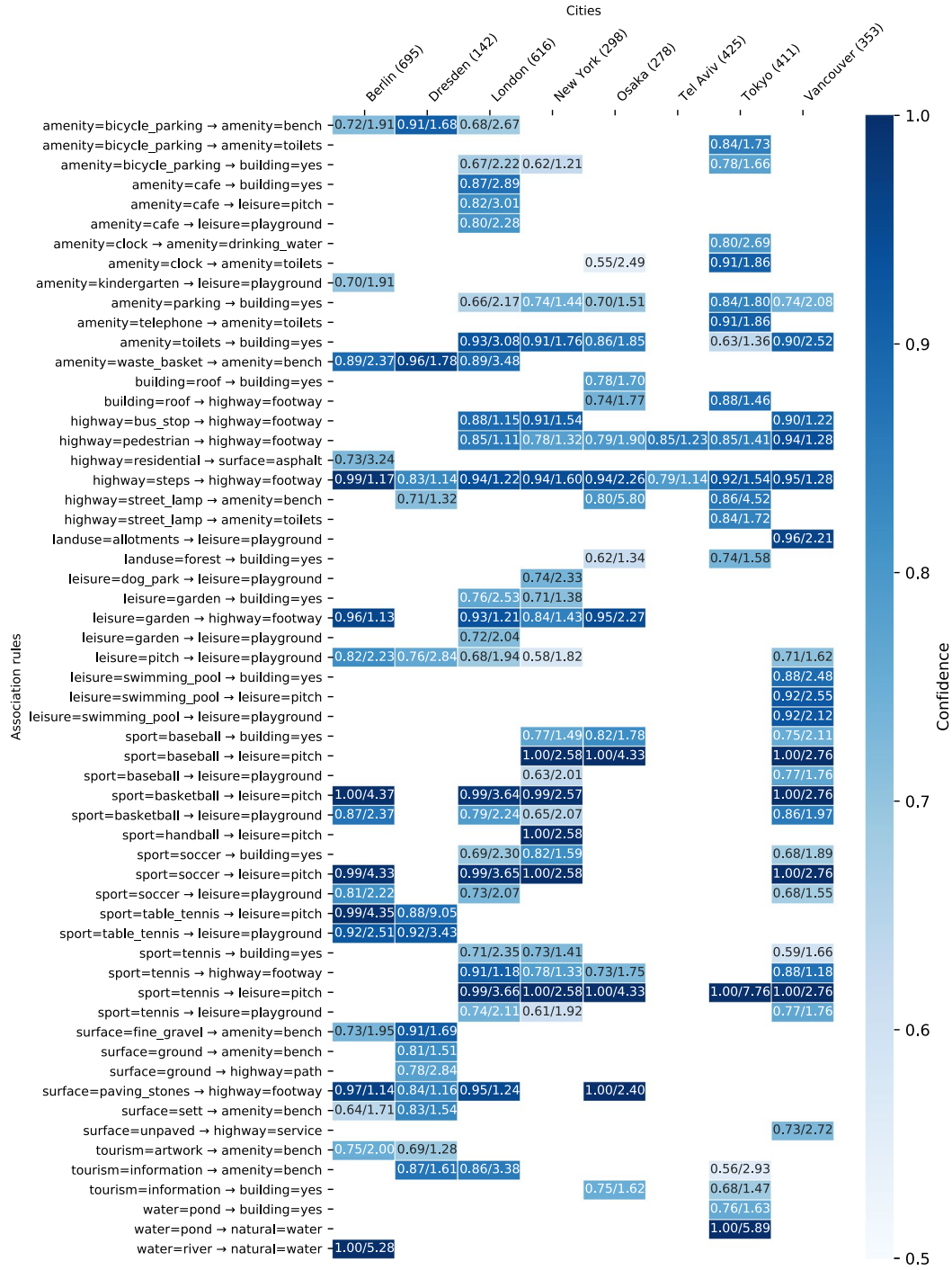

FIGURE 5 Selected association rules between OSM tags based on parks with at least three features mapped for each city. The first value indicates confidence, the second value represents lift of the rule. Empty cells mark rules whose support is less than or equal to 0.05. Only rules are shown where confidence is 0.7 or greater and lift is 1.5 or greater in at least one city and confidence is 0.5 or greater and lift is 1.1 or greater in all cities where support is less than or equal to 0.05. The number of selected parks within each city is given in parentheses next to their names

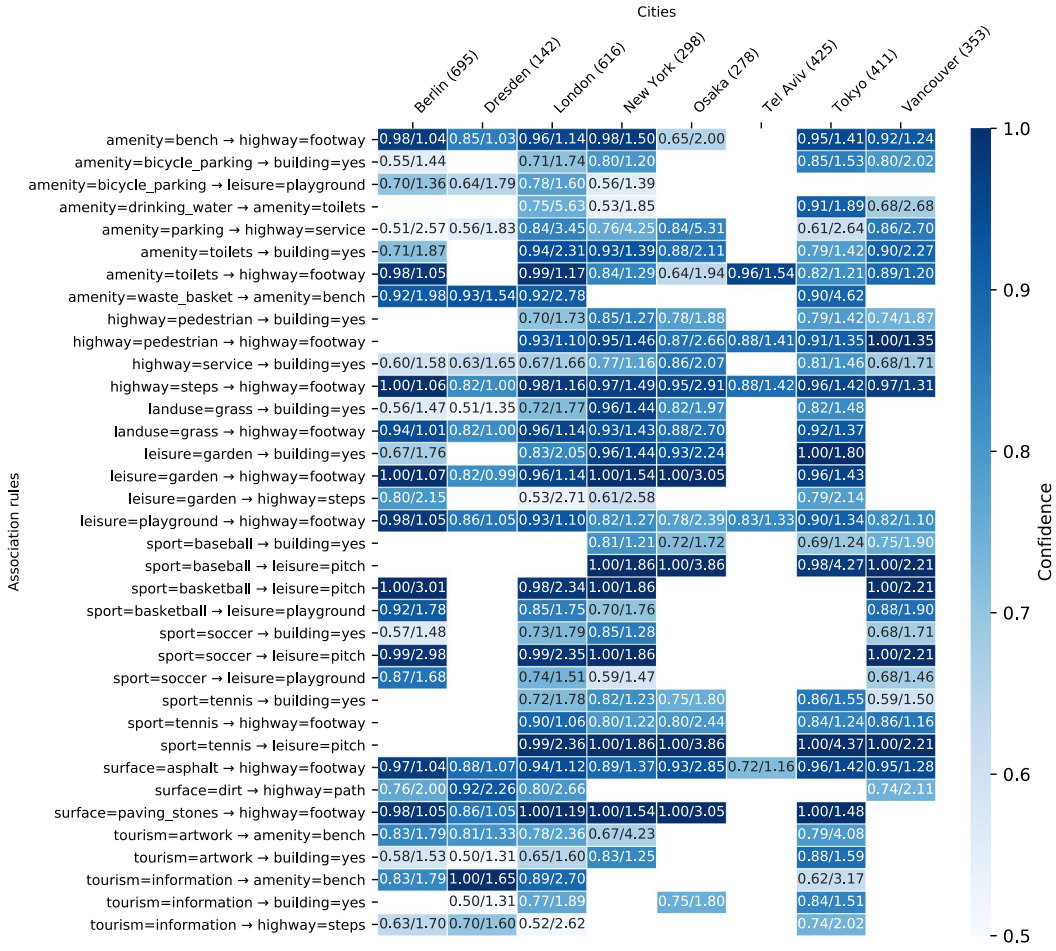

FIGURE 6 Selected association rules between OSM tags based on parks larger than 0.5 ha in OSM for each city. The first value indicates confidence, the second value represents lift of the rule. Empty cells mark rules whose support is less than or equal to 0.05. Only rules are shown where confidence is 0.7 or greater and lift is 1.5 or greater in at least one city and confidence is 0.5 or greater in all cities where support is less than or equal to 0.05. Only rules are shown which apply in at least four cities. The number of selected parks within each city is given in parentheses next to their names

other types of urban areas could yield additional insights into regional differences in the conceptualization and tag usages regarding urban green spaces in OSM. Performing the analysis on OSM keys instead of tags could be useful in this regard as well. The higher relative frequencies of OSM keys could yield more robust association rules; however, it needs to be kept in mind that OSM keys are less specific than OSM tags, which could lead to misinterpretation of the association rules derived.

Apart from cultural aspects, the results of the context-based association rule analysis suggest that the mapping process itself also influences the association rules. Several context variables which are connected to the amount of mapping activity in OSM led to an increase in the number of rules which exceeded the minimum support and confidence thresholds. This in turn led to a strong increase in the number of rules which were valid within multiple cities. As a result, more regional commonalities were detected. However, there are two aspects which need to be noted here. First, the size of the parks is positively correlated with many of the context variables describing mapping activity. Therefore, the distinction between how much of the change is solely due to the

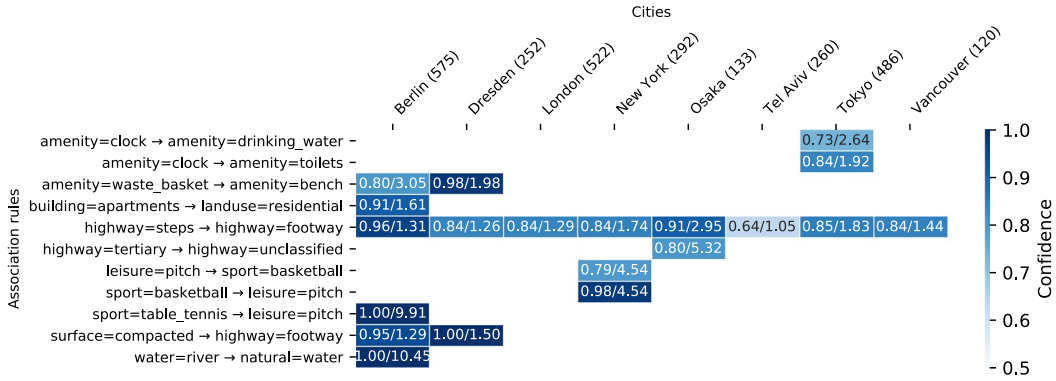

FIGURE 7 Selected association rules between OSM tags based on parks smaller than 0.5 ha in OSM with at least three features mapped inside for each city. The first value indicates confidence, the second value represents lift of the rule. Empty cells mark rules whose support is less than or equal to 0.05. Only rules are shown where confidence is 0.7 or greater and lift is 1.5 or greater in at least one city and confidence is 0.5 or greater in all cities where support is less than or equal to 0.05. The number of parks selected within each city is given in parentheses next to their names

increased mapping activity and how much due to a larger share of large parks being analyzed cannot be made at this point. Second, when excluding empty park features from the analysis, it is not possible to tell whether these parks truly do not contain any objects in reality or whether some objects exist but are yet to be mapped in OSM. Distinguishing between these two cases should be considered when interpreting the lift value, since it can lead to an over- or underestimation of the strength of co-occurrence of the respective tags.

The context-based association rule analysis also suggests that the relevance of a rule, quantified by its lift value, varies depending on the level of completeness of the data. Within the analysis, rules were detected which, based on common sense, should be applicable universally, (e.g., stairs marked by the tag *highway=steps* should always be connected to a path or a road). However, the analysis revealed that universal rules like this one show similarly high confidence values but different lift values across cities. The lift value for Osaka was twice as high as that for the other cities. Although it cannot be ruled out at this point that this is due to the fact that most parks in Osaka do not contain any paths in reality, it is still possible that this elevated lift value is due to the low level of completeness of footpaths mapped in Osaka. This indication is supported by the continuously decreasing lift values with increasing values of context variables related to the mapping activity (Section 4.3).

Mapping spatial objects in OSM usually does not happen randomly across space but is very much determined by the spatio-temporal distribution of mappers. Independently of whether they are mapping remotely or locally, they are probably more likely to map objects which are located closely together. If a person visits a park, they will probably map different kinds of amenities at the same time instead of only one specific type of object while randomly visiting several parks. As a result, two tags which actually occur independently of each other in reality might seem to show a strong co-occurrence relationship just because they always occur together within the few parks that have been fully mapped in OSM.

A clear distinction between which regional variations in association rules are due to cultural influences and which are due to the mapping process cannot be made at this point. A more detailed analysis of association rules within the mapping process itself (e.g., which kinds of objects are frequently mapped together) could enable a better understanding of whether differences in lift values are actually due to different physical structures in the real world or rather due to the mapping process. Further investigations into the influence of the mapping process on the association rules are necessary to reach a better understanding of the OSM data in a certain region, especially if this information is to be used within tag recommendation systems or for data quality assessments.

6 | CONCLUSIONS

In this study we explored how association rules derived from OpenStreetMap data vary across geographic regions and depending on different context variables. A context-based association rule analysis was conducted for eight cities to derive association rules between OSM tags occurring within parks mapped in OSM. Including all parks in the association rule analysis yielded mostly region-specific association rules and only few universally applicable ones. Limiting the association rule analysis to parks based on specific context variables increased the number of rules which are applicable across multiple cities. Some association rules showed high confidence values across all cities, but elevated lift values for cities with a low level of completeness. Furthermore, a connection between the lift value of a rule and context variables related to mapping activity was found. These results suggest that the mapping process has a significant influence on the emergence of association rules within user-generated data. This phenomenon is not yet sufficiently understood and should be further investigated to enable more effective usage of OSM data across different cultural realms.

ACKNOWLEDGMENT

The authors of this study were funded by the Federal Ministry of Transport and Digital Infrastructure (BMVI) within the mFUND research initiative and the Klaus Tschira Stiftung. Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Christina Ludwig https://orcid.org/0000-0003-4669-3298
Alexander Zipf https://orcid.org/0000-0003-4916-9838

REFERENCES

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, Washington, DC (pp. 207–216). New York, NY: ACM.

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. In J. Bocca, M. Jarke, & C. Zaniolo (Eds.), *Proceedings of the 20th International Conference on Very Large Data Bases: VLDB '94*, Santiago, Chile (pp. 487–499). San Francisco, CA: Morgan Kaufmann.

Ali, A. L., Schmid, F., Al-Salman, R., & Kauppinen, T. (2014). Ambiguity and plausibility: Managing classification quality in volunteered geographic Information. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, Dallas, TX (pp. 143–152). New York, NY: ACM.

Ali, A. L., Sirilertworakul, N., Zipf, A., & Mobasheri, A. (2016). Guided classification system for conceptual overlapping classes in OpenStreetMap. *ISPRS International Journal of Geo-Information*, *5*(6), 87.

Bahrdt, D., Funke, S., Gelhausen, R., & Storandt, S. (2017). Searching OSM Planet with context-aware spatial relations. In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, Redondo Beach, CA. New York, NY: ACM.

Ballatore, A., & Zipf, A. (2015). A conceptual quality framework for volunteered geographic information. In S. Fabrikant, M. Raubal, M. Bertolotto, C. Davies, S. Freundschuh, & S. Bell (Eds.), *Spatial information theory: COSIT 2015* (Lecture Notes in Computer Science, Vol. *9368*, pp. 89–107). Cham, Switzerland: Springer.

Barron, C., Neis, P., & Zipf, A. (2014). A comprehensive framework for intrinsic OpenStreetMap quality analysis. *Transactions in GIS*, *18*(6), 877–895.

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.

Juhász, L., Hochmair, H. H., Qiao, S., & Novack, T. (2019). Exploring the effects of Pokémon Go vandalism on OpenStreetMap. In M. Minghini, A. Y. Grinberger, P. Mooney, L. Juhász, & G. Yeboah (Eds.), *Proceedings of the*

Academic Track, *State of the Map 2019 Conference*, Heidelberg, Germany. Retrieved from https://zenodo.org/record/3405431#.X1SDQ-eSnGg

Kashian, A., Rajabifard, A., Richter, K.-F., & Chen, Y. (2019). Automatic analysis of positional plausibility for points of interest in OpenStreetMap using coexistence patterns. *International Journal of Geographical Information Science*, *33*(7), 1420–1443.

Keßler, C., & de Groot, R. T. A. (2013). Trust as a proxy measure for the quality of volunteered geographic information in the case of OpenStreetMap. In D. Vandenbroucke, B. Bucher, & J. Crompvoets (Eds.), *Geographic information science at the heart of Europe* (pp. 21–37). Cham, Switzerland: Springer.

Kinas, A. (2018). *Entwicklung eines Werkzeuges zur räumlichen Analyse von Attributassoziationen am Beispiel OpenStreetMap* (Unpublished BS thesis). Heidelberg University, Heidelberg, Germany.

Koperski, K., & Han, J. (1995). Discovery of spatial association rules in geographic information databases. In M. J. Egenhofer, & J. R. Herring (Eds.), *Advances in spatial databases: SSD 1995* (Lecture Notes in Computer Science, Vol. *951*, pp. 47–66). Berlin, Germany: Springer.

Ludwig, C., & Zipf, A. (2019). Exploring regional differences in the representation of urban green spaces in OpenStreetMap. In *Proceedings of the "Geographical and Cultural Aspects of Geo-Information: Issues and Solutions" AGILE 2019 Workshop*, Limassol, Cyprus (pp. 10–14).

Mennis, J., & Liu, J. W. (2005). Mining association rules in spatio-temporal data: An analysis of urban socioeconomic and land cover change. *Transactions in GIS*, *9*(1), 5–17.

Mocnik, F.-B., Mobasheri, A., Griesbaum, L., Eckle, M., Jacobs, C., & Klonner, C. (2018). A grounding-based ontology of data quality measures. *Journal of Spatial Information Science*, *16*, 1–25.

Mocnik, F.-B., Zipf, A., & Raifer, M. (2017). The OpenStreetMap folksonomy and its evolution. *Geo-spatial Information Science*, *20* (3), 219–230.

Mooney, P., & Corcoran, P. (2012). Characteristics of heavily edited objects in OpenStreetMap. *Future Internet*, *4*(1), 285–305.

Mülligann, C., Janowicz, K., Ye, M., & Lee, W.-C. (2011). Analyzing the spatial-semantic interaction of points of interest in volunteered geographic Information. In M. Egenhofer, N. Giudice, R. Moratz, & M. Worboys (Eds.), *Spatial information theory* (Lecture Notes in Computer Science, Vol. *6899*, pp. 350–370). Berlin, Germany: Springer.

Neis, P., Zielstra, D., & Zipf, A. (2013). Comparison of volunteered geographic Information data contributions and community development for selected World Regions. *Future Internet*, *5*(2), 282–300.

OpenStreetMap contributors. (2020a). *OSM Stats*. Retrieved from https://wiki.openstreetmap.org/wiki/Stats

OpenStreetMap contributors. (2020b). *OSM Wiki*. Retrieved from https://wiki.openstreetmap.org

Piateski, G., & Frawley, W. (1991). *Knowledge discovery in databases*. Cambridge, MA: MIT Press.

Raifer, M., Troilo, R., Kowatsch, F., Auer, M., Loos, L., Marx, S., & Zipf, A. (2019). OSHDB: A framework for spatio-temporal analysis of OpenStreetMap history data. *Open Geospatial Data, Software & Standards*, *4*(1), 3.

Schmitz, C., Hotho, A., Jäschke, R., & Stumme, G. (2006). Mining association rules in folksonomies. In V. Batagelj, H. H. Bock, A. Ferligoj, & A., Ziberna (Eds.), *Data science and classification: Studies in classification, data analysis, and knowledge organization* (pp. 261–270). Berlin, Germany: Springer.

Sha, Z., Tan, X., & Bai, Y. (2015). Localized spatial association: A case study for understanding vegetation successions in a typical grassland ecosystem. In F. Bian & Y. Xie (Eds.), *Geo-informatics in resource management and sustainable ecosystem: GRMSE 2014* (Communications in Computer and Information Science, Vol. *482*, pp. 33–45). Berlin, Germany: Springer.

Shaheen, M., Shahbaz, M., & Guergachi, A. (2013). Context based positive and negative spatio-temporal association rule mining. *Knowledge-Based Systems*, *37*, 261–273.

Tang, K., Chen, Y.-L., & Hu, H.-W. (2008). Context-based market basket analysis in a multiple-store environment. *Decision Support Systems*, *45*(1), 150–163.

Tost, H., Reichert, M., Braun, U., Reinhard, I., Peters, R., Lautenbach, S., … Meyer-Lindenberg, A. (2019). Neural correlates of individual differences in affective benefit of real-life urban green space exposure. *Nature Neuroscience*, *22*(9), 1389–1393.

Vandecasteele, A., & Devillers, R. (2015). Improving volunteered geographic information quality using a tag recommender system: The case of OpenStreetMap. In J. Jokar Arsanjani, A. Zipf, P. Mooney, & M. Helbich (Eds.), *OpenStreetMap in GIScience: Experiences, research, and applications* (Lecture Notes in Geoinformation and Cartography, pp. 59–80). Cham, Switzerland: Springer.

How to cite this article: Ludwig C, Fendrich S, Zipf A. Regional variations of context-based association rules in OpenStreetMap. *Transactions in GIS*. 2020;00:1–20. https://doi.org/10.1111/tgis.12694