# OSM POI ANALYZER: A PLATFORM FOR ASSESSING POSITION OF POIs IN OPENSTREETMAP

A. Kashian [a,*] , A. Rajabifard [a], Y. Chen [a], K. F. Richter [b]
[a] Dept. of Infrastructure Engineering, University of Melbourne, Parkville, Australia – (alireza.kashian,abbas.r, yiqun.c)@unimelb.edu.au
[b] Dept. of Computing Science, Umeå University, Sweden – kai-florian.richter@umu.se

**KEY WORDS:** OpenStreetMap, Spatial Data Quality, Co-existence analysis, Co-location Pattern, Spatial Association Rule, POI

**ABSTRACT:**

In recent years, more and increased participation in Volunteered Geographical Information (VGI) projects provides enough data coverage for most places around the world for ordinary mapping and navigation purposes, however, the positional credibility of contributed data becomes more and more important to bring a long-term trust in VGI data. Today, it is hard to draw a definite traditional boundary between the authoritative map producers and the public map consumers and we observe that more and more volunteers are joining crowdsourcing activities for collecting geodata, which might result in higher rates of man-made mistakes in open map projects such as OpenStreetMap. While there are some methods for monitoring the accuracy and consistency of the created data, there is still a lack of advanced systems to automatically discover misplaced objects on the map. One feature type which is contributed daily to OSM is Point of Interest (POI). In order to understand how likely it is that a newly added POI represents a genuine real-world feature scientific means to calculate a probability of such a POI existing at that specific position is needed. This paper reports on a new analytic tool which dives into OSM data and finds co-existence patterns between one specific POI and its surrounding objects such as roads, parks and buildings. The platform uses a distance-based classification technique to find relationships among objects and tries to identify the high-frequency association patterns among each category of objects. Using such method, for each newly added POI, a probabilistic score would be generated, and the low scored POIs can be highlighted for editors for a manual check. The same scoring method can be used for existing registered POIs to check if they are located correctly. For a sample study, this paper reports on the evaluation of 800 pre-registered ATMs in Paris with associated scores to understand how outliers and fake entries could be detected automatically.

## 1. INTRODUCTION

The development of innovative analytical tools to extract useful information from existing geo-spatial datasets is crucial for most researchers to understand what is recorded under the skin of large crowdsourcing platforms such as OSM. Extracting patterns and mining the spatial data in open-source projects recently draws the attention of many researchers and led us to look for effective methods for investigating man-made mistakes and remove errors from existing databases. These methods also can help stopping vandalism and vicious activities which, for example, may add nonsense data to map datasets. Fortunately, without these automatic mechanisms, most of the geodata crowdsourcing platforms have already incorporated different control mechanisms within their editors for quality assurance and validation. For example, the JOSM editor informs a user prior to the upload if there are any intersecting geometries or duplicated elements. However, this type of check only stays on the level of topology validation without considering the spatial relationships between the newly added feature and its neighbouring features and therefore can be easily bypassed; moreover, instead of automatically refusing the new edits, it hands the submission decision to users without providing enough supporting information such as the likelihood of the misplacement of a POI (Neis et al., 2012).

In this work, we developed an analytic tool for assessing the position of POIs in OSM. The main design concept is based on Tobler's first law of geography, which claims that everything is related to everything else but nearby things are more related than distant things (Tobler, 1979). Based on this law, we aim to discover potential co-existence patterns among POIs and between POIs and other geographical features, such as roads and buildings, which are in close proximity to each other. For example, consider the relationship between gas stations and road segments. As we all know, vehicles need access to road structure to drive into gas stations. We would assume that whenever we find a gas station, it is highly likely to have a road segment nearby as well. Although classical data mining algorithms (Agrawal and Srikant, 1994) are often based on assumptions which violate Tobler's law (e.g. independent, identical distributions), in a spatial context nearby objects tend to affect each other in important ways rather than acting independently.

All processing in this analytic tool is based on distance and geometry types. There is no 'semantic' analysis in this system. It can help researchers and the OSM community explore the vast amount of data to find hidden relationships between two sets of POIs in one specific city. The platform also generates recommendation scores which can direct OSM editors whether a newly added POI is correctly positioned or not. The score is generated statistically based on similarities between the newly added POI and existing patterns for a similar category of POIs in the same city.

In the following sections, we briefly introduce the OSM project and discuss the definition and importance of POI data; then the discussion continues with the spatial association patterns among POIs and reviews some existing monitoring tools in OSM. Section 4 articulates the implementation details of the proposed platform and Section 5 discusses the system performance by validating the positions of ATMs in Paris. Some known issues and possible improvements are discussed at the end of this paper.

## 2. OPENSTREETMAP AND POIS

* Corresponding author

The OpenStreetMap (OSM) project started in 2004. Main web services and databases of OSM are saved and hosted on several servers located at University College London. "Building a global map" is stated to be the main aim of the project Data et al. (2012). OSM provides free access to all data as well as the history of changes for each individual object. A complete review of OSM's recent developments is available(Neis and Zielstra, 2014). For being eligible to submit data to the OSM project, contributors should register and create an account for themselves. In the OSM project, also members who have only registered recently can modify, add or even delete geographic objects in the OSM database right after passing the routines of getting registered. This method of data reception is in complete contrast to other VGI projects, for example, Google Map Maker (retired in March 2017), where the alterations made by new members are reviewed before being applied to the Google Maps

To understand the ecosystem of OSM, it is important to provide a brief introduction to the four main data elements in OSM project. This information is available from OSM's Wiki page as well.

**Node:** A node represents a specific point on the earth's surface defined by its latitude and longitude. Each node comprises at least an id number and a pair of coordinates. Nodes can be used to define standalone point features. For example, a node could represent a park bench or a water well. Nodes are also used to define the shape of a way. When used as points along ways, nodes usually have no tags, though some of them could. For example, 'highway=traffic signals' marks traffic signals on a road, and 'power=tower' represents a pylon along an electric power line. A node can be included as a member of a relation. A relation also may indicate a member's role: that is, a node's function in this particular set of related data elements.

**Way:** A way is an ordered list of between 2 and 2,000 nodes that define a polyline. Ways are used to represent linear features such as rivers and roads. Ways can also represent the boundaries of areas (solid polygons) such as buildings or forests. In this case, the way's first and last node will be the same. This is called a "closed way". Note that closed ways occasionally represent loops, such as roundabouts on highways, rather than solid areas. A way's tags must be examined to discover which it is. Areas with holes, or with boundaries of more than 2,000 nodes, cannot be represented by a single way. Instead, the feature will require a more complex multi-polygon relation data structure.

**Relation:** A relation is a multi-purpose data structure that documents a relationship between two or more data elements (nodes, ways, and/or other relations). Examples include:

- A route relation, which lists the ways that form a major (numbered) highway, a cycle route, or a bus route.
- A turn restriction that says you cannot turn from one way into another way.
- A multipolygon that describes an area (whose boundary is the 'outer way') with holes (the 'inner ways').

Thus, relations can have different meanings. Its tags define a relation's meaning. Typically, a relation will have a 'type' tag. A relation's other tags need to be interpreted in light of the type tag. A relation is primarily an ordered list of nodes, ways, or other relations. These objects are known as a relation's members. Each element can optionally have a role within a relation. For example, a turn restriction would have members with "from" and "to" roles, describing the particular turn that is forbidden. A single element such as a particular way may appear multiple times in a relation.

**Tag:** All types of data element (nodes, ways and relations) can have tags. Tags describe the meaning of the particular element to which they are attached. A tag consists of two free-format text fields; a 'key' and a 'value'. Both of these are Unicode strings of up to 255 characters. For example, 'highway=residential' defines the way as a road whose main function is to give access to people's homes.

There is no fixed dictionary of tags, but there are many conventions documented on OSM'a online wiki (starting with the Map Features page). Tag usage can be measured with the Taginfo application[2]. If there is more than one way to tag a given feature, it is probably best to use the most common approach. Moreover, the unrestricted use of key-value pairs for tagging features provides an excellent means of customized annotations suitable for thematic applications.

Many applications in the world use OSM data, while the data accuracy and reliability is always under question. As most of the contributors in OSM are not experts in the GIS field, answering the quality question is challenging (Hashemi and Abbaspour, 2015; Salk et al., 2016). Many researchers have analysed the quality of OSM (Amirian et al., 2015; Arsanjani et al., 2015; Fan et al., 2014; Helbich et al., 2012; Koukoletsos et al., 2012). Some of these quality studies focused on comparing OSM data with other reference data such Ordnance Survey (UK) or even commercial datasets such as Google and Here. The comparisons are in the domain of positional, temporal, and thematic accuracy including completeness of coverage.

Most geographic crowdsourcing initiatives have embedded solutions for collecting and managing point features that refer to a specific spot on the map (or on earth). These locations are known as points of interest (POI). There is a wide range of geographic objects that may be considered POIs. For example, post boxes may not be what come immediately to our mind, but for specific users and tasks they may well be important, therefore could be called POIs. Several map-centric applications such as Google Maps, local directory services such as Yelp and location-based social networks such as Foursquare extensively use POIs to design commercial services. Based on the application and context, the definition of POI varies. Determining a neighbourhood's accessibility is one of the domains where POIs show their importance. In recent years, researchers paid more attention to the analysis of distance between the user and different POIs, based on walking or cycling (Iacono et al., 2010)

The following list provides some examples of what may be considered POIs in general:

- Churches, hospitals, schools, town halls, distinctive buildings
- Post offices, shops, post boxes, telephone boxes
- Pubs (pub names are useful when navigating by map)
- Car parks and lay-bys (and whether free or not)
- Bus stations, metro stations, ferry terminals, airport terminals
- Speed cameras, police stations
- Tourist attractions

In this work, we are mostly interested in investigating the spatial relationships and co-existence patterns among these POIs.

[2] http://taginfo.openstreetmap.org

## 3. SPATIAL RELATIONSHIPS AMONG POIS AND EXISTING ANALYTICAL TOOLS

In this project, we treated the problem as a spatial data mining challenge and attempted to figure out whether POIs in short distance to each other have a strong relationship and whether dataset-wide co-existence patterns are discoverable. Exploring the spatial relationships of POIs can improve the quality of OSM data by checking existing POI objects or by enabling the user interface to block or advise on wrongly positioned POIs. Normally, OSM data quality can be assessed by two different methods: (1) using reference or authoritative resources to compare with existing OSM data; (2) establishing rules (manual or automatic) and checking incoming data with these rules to detect mistakes automatically. In the second method, the rules can be defined by the user or could be extracted from existing data. Our proposed tool checks the spatial association rules between POIs and surrounding objects using spatial queries from the existing OSM dataset. Due to the huge amount of geometric data, pre-processing of the data is necessary to provide real-time recommendations to users.

Working with geometric representations (points, lines, and regions) is often cumbersome and undesirable as running intersection queries inside geodatabases are usually computationally heavy. Thus, we converted and transformed data into easily query-able relational DBs which includes tables and fields describing every relationship between each object pair in numerical format. Another important aspect to take into consideration for pre-processing is looking at a proper spatial resolution which can have a direct impact on the strength of patterns that can be discovered in the datasets. Unusually most general patterns are more likely to be discovered at the lowest resolution/granularity level. Low resolution means a large spatial unit scale in this context. On the other hand, large support is more likely to exist at higher levels of resolution. To find patterns with strong support, a higher level of resolution and granularity is recommended.

Many developers and researchers have worked on different tools for monitoring and bug reporting in the OSM project. Each tool is developed based on specific needs of ordinary volunteers or advanced editors. Some of these tools help to identify bugs in existing data while others help to monitor and visualize the live creation of data in OSM. These tools are fundamentally developed for improving the quality of data and many of them are listed at the Wiki page of OSM under quality assurance tag.[3]

Below is a short review of tools for monitoring and analyzing OSM. This helps to shape a better perspective on what other groups of researchers and OSM patrons have explored in OSM dataset. We do not intend to compare them or even analyze them in details, and we only cover a few of them as a reference to readers.

WHODIDIT[4]: A tool to analyze the changesets in OSM. A group of changes created by a single user in a short period will shape a changeset.

OSM Relation Analyzer[5]: will analyze OSM relations for gaps. This analyzes will help the user to find errors, especially for route relations. Relations often get corrupted during general edits by inexperienced users.

OSMarelmon[6]: The OSM Relation Monitor checks relations periodically. It generates RSS feeds based on preprocessed relations.

OSM Route Manager[7]: This tool analyzes route relations and tries to find gaps between relations and exports GPX files.

NoName[8]: This tool highlights the roads that have no name in the database of OSM. This helps all editors to trace and find those unnamed roads and assign them with names.

NOVAM[9]: Displays bus stops in the UK in order to monitor and verify NaPTAN data import to OSM. NaPTAN is the UK official dataset for bus stops with around 350,000 public transport access points including bus stops, railway stations and tram stops.

OSM Inspector[10]: A comprehensive tool for inspecting OSM elements. The user can check for many incomplete elements to fix them later. Problems such as geometric inconsistencies (e.g., self-intersecting ways) are detectable. In addition, the user can find those tags which has no value or are unnamed.

Keep Right[11]: A good tool for detecting errors in the OSM project. Errors include missing tags, floating islands, motorways without ref and non-closed areas.

## 4. OSM POI ANALYZER

POIs are one of the main data elements in OSM which many commercial and free applications depend on. They are especially useful for end users to find nearby facilities and navigate to these places. The important role of POIs inspired us to investigate how to automatically validate and improve their positional credibility.

We devised a new analytic tool (http://openstreetmap.me) to understand the distance-based relationship between POIs and their surrounding objects. The platform dives into the OSM data and makes millions of spatial queries to compare each individual POI with its surrounding geographic objects. The generated results can help researchers and OSM editors evaluate the structure of the city and quantitatively gauge the acceptance likelihood of the position of existing or new POIs. The platform also generates a ranking score which declares the probability of a particular type of POI existing at the proposed position, which is measured against all other existing similar objects in the same city. For example, generally we expect a gas station service to be close to a road segment, or a ferry terminal to be close to a lake or a river. We will be able to discover this knowledge in a systematic way by checking all ferry terminals or all gas stations and then establishing their relationships with nearby objects, such as roads, rivers and lakes. In other words, access to interpretable and meaningful knowledge about our existing world is critical for finding meaningful relational patterns.

Different cities have different urban designs; thus, the results would vary from one location to another. This means the results we extract for each city should be unique for that city, though some patterns may be globally observed. In the developed tool,

[3] http://wiki.openstreetmap.org/wiki/Quality_assurance
[4] http://zverik.osm.rambler.ru/whodidit/
[5] http://ra.osmsurround.org/
[6] http://osmarelmon.won2.de/
[7] http://osmrm.openstreetmap.de/index.jsp
[8] http://qa.poole.ch/
[9] http://www.mappa-mercia.org/novam
[10] http://tools.geofabrik.de/osmi/
[11] http://keepright.at/

the focus is on the co-existence of pairs of geographic objects. This co-existence is checked between one POI type and the rest of geographic objects such as roads, rivers, jungles and parks. As an example, gas station and road segment are co-existing.

### 4.1 The Concept of Processing POIs

To simplify the processing of data and better understand the association rules among POIs, geographic features are divided into two separate categories in this study. The first category covers all geographic features, which are mainly used to visit for our daily tasks or we might live, work or shop there. Hospitals, houses, bus stops, monuments and police stations are examples of such geographic features. The second category refers to all other features which we use regularly as a means of transportation or we simply pass through them to go from point A to point B. Roads, rivers, lakes and forests are examples of this second category. We assigned two short terms to each category: RP (Reference Place) to the first group (most of the POIs are in this category) and TP (Transit Places) to the second group (which covers roads, lakes and jungles). In this definition, no feature could belong to both RP and TP category at the same time, thus they are mutually exclusive. To assign each class of objects to one of these categories, we prepared the list of all popular and commonly used tags from the Taginfo application. Then we manually processed the list and assigned each tag to one of our defined categories. For example, amenity:hospital and natural:tree are inserted in the RP set and highway:service and waterway:stream are placed in the TP set. In our list, more than 800 tags are processed. Having RP and TP groups, we are able to investigate meaningful relationships between both categories as well as relations between pairs inside each category. For example, we might see that most of the times, amenity:hospital (RP) is located close to highway:trunk (TP) or amenity:ferry_terminal (RP) is placed close to waterway:river (TP). Aside from the spatial relationships between the RT and TP sets, there are also interesting relations between members of the RP category. For example, in most cases, amenity:clinic is close to amenity:hospital or amenity:atm is close to amenity:bank. There are more relations yet to be discovered by using the analytic tool.

Another important issue for pre-processing the data was to bring the distance element into our calculations. Assume if we only have one playground in Melbourne city and it is 10km away from the nearest park, then it would be very hard to establish a meaningful relationship between this single playground and the park. But what if there were 200 playgrounds in Melbourne and most of them were located inside or were very close to parks and green area? This might indicate a reasonable pattern to deduct a relationship between park and playground. In a similar way we can logically deduct that most ATMs are close to major city roads and also close to some bank branches. This way we would say that something interesting is observable and that it is a repetitive pattern for the location of objects in one specific city. As discussed earlier, the distance between two objects is important in our processing unit. To convert continuous distance values into granular discrete values, we assign 15 circular doughnut shape clusters around each instance of POI objects. These 15 distance regions reflect Tobler's First Law of Geography, which states that nearby objects have a stronger relationship with each other than more distant ones. Figure 1 illustrates the concept of the doughnut clusters. The distance regions, which form doughnut rings, have different ranges. The first 5 doughnut regions have a 10 meters range, the 6$^{th}$ ring has a 50 meters range, and all the rest have a 100 meters range. Since the 15$^{th}$ region ranges from 900 meters to 1000 meters, objects beyond 1,000 meters are ignored in our processing. The 1,000 meters distance was selected as a cut-off point for reducing processing of unnecessary objects, which would likely have no or only insignificant relation with the POI that we inspect. Further work is needed to tell whether we miss important relationships using this cut-off point or not.

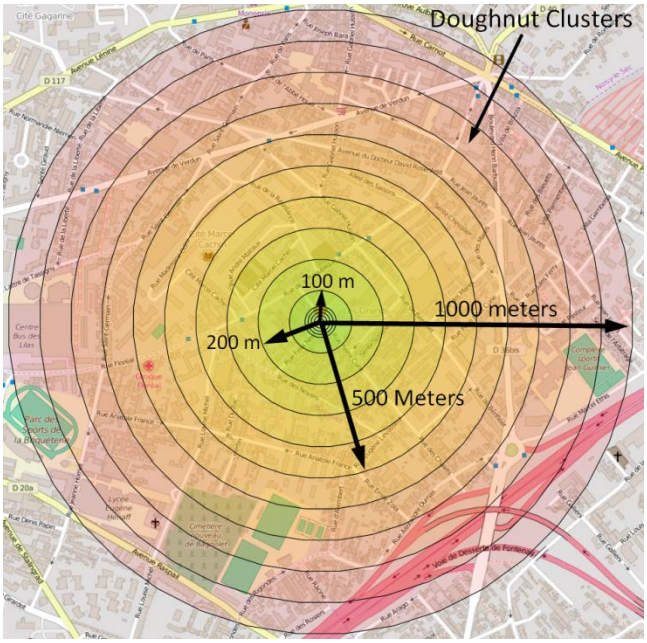

Figure 1. 15 doughnut clusters around each POI

### 4.2 Data Preparation Phase

The platform has been implemented in PHP and Python using a PostgreSQL database on a Debian cloud server. To prepare the data for processing, the following steps are taken:

1- We selected 5 sample cities which according to OSM statistics have an almost complete range of objects and the enrichment level of data in OSM was good enough for our purpose. The selected cities are Paris, Frankfurt, Melbourne, Madrid, and Vancouver. We exported OSM data via bbbike[12] and then imported it to our PostGIS database running on a remote server. The imported data includes all tags, nodes, ways, and relations.

2- We assigned all imported objects for each city to two different tables named as [city]_rp and [city]_tp which is based on the definitions introduced in the previous section.

3- A list of known tags for POIs in OSM was prepared. Some tags are officially listed by the OSM project as commonly accepted tags for specific features on the map. This selection of tags helped us identify the nodes which actually represent POIs as there are many nodes, which are only part of a way and do not represent a POI.

4- To reduce the amount of computation, only a subset of POI types was selected for further processing. The sample subset includes 22 manually selected types, such as amenity:ATM, amenity:bank, office:company, leisure:playground, and amenity:post_box. The list is shown in.

5- Table **1**. The selection criteria were based on those object types which were frequent and we expect to find them

[12] http://download.bbbike.org/osm/bbbike/

normally throughout the city as this enables us to consider all possible patterns in different locations. Therefore object types such as airports are not selected.

6- For each of the 22 POI types, all instances of existing objects were extracted using spatial queries. For example, 3,944 traffic_signals in Madrid and 1,890 amenity:bank in Paris were extracted.

7- For each instance of these objects (a single POI), 15 circular doughnut clusters were created and then spatial queries were run to identify any object (Way, Node), which intersected with each cluster area. The results were stored as raw data in a new table for each city. For example, for each amenity:bakery in Paris, an intersection query for each doughnut cluster is performed separately and if Paris has 2,000 bakeries, this results in 30,000 (2,000*15) queries.

8- As a last step, we ran queries to find how many times a specific object type is observed in each cluster. This information is again stored in a new table which we later use to find a confidence ratio for each association rule between the main object and the surrounding objects. As an example, this table can give us information about what percentage of amenity:ATM are within 50 meters distance of amenity:bank or what percentage of amenity:Ferry_Terminals are within 100 meters distance of a waterway:river in London?

### 4.3 How the OSM POI Analyzer Works

Using the OSM POI analyzer starts with the selection of a POI type and then clicking on the map to propose a location for a new instance of that type. It is important that the user zooms in enough to accurately select a correct position. Once the position is selected, the analyzer starts processing the request and it might take up to 10 seconds to respond. On the server side, the processing continues with queries to pre-processed data tables for the proposed object type and, simultaneously, it makes geometric queries to evaluate the new location for all nearby objects within all 15-cluster distances around the new position. The following steps are taken while processing the new POI:

Step 1: Analyzing nearby objects to check whether duplicate object(s) exists or not. If there are duplicates, the platform informs the user that it has found a similar object very close by. To find duplicates, we prepared a table for all tags to indicate what is the minimum possible distance between two objects within the same class. We filled up this table manually based on our own intuitions but this process can be developed further to find this minimum distance automatically. For example, for amenity:ferry_terminals we set the distance to 50 meters while for amenity:ATM it is set to 2 meters. In this case, if another ATM is registered in less than 2 meters distance of another ATM, then it is considered as a duplicate. This simple technique was fit for our primary purposes; however, other researchers have worked on techniques to compare duplicate objects using automatic conflation of geometries or attributes or measuring the auto-correlation of similar objects in the same class.(Blasby et al., 2002; Ching-Chien Chen, 2008; Grant McKenzie, 2013; Yuan and Tao, 1999)

Step 2: Similar to checking for duplicates, an evaluation of the geometric relationship between newly registered objects and surrounding objects is useful. In our work, we manually created a table to indicate the minimum buffer zone for each object type. For example, we set trunk:highway to a 20 meters buffer from its center line, therefore if new objects are registered inside this buffer area, we know it is very close to this road segment so we can report it back to the user. This information is helpful, for example, if we need to reject or give an alert about a registration of amenity:hospital very close to trunk:highway which is probably incorrect. As we later explain, we do not use this buffer measurement in our scoring system and we only report it as an extra signal to the user.

Step 3: Checking to see if the registered POI has a similar association with nearby objects through comparison with similar object types which are processed earlier (pre-processing engine). The result is a composite score which indicates the credibility of the proposed location for the new object.

Step 4: The system prepares different reports about the status of the registered point including a list of nearby objects within 20 meters range or a list of objects for which the newly registered object is located inside their buffer zone. All prepared reports are sent back to the client for further inspection by the user. One of the tables expresses the similarity values between what we observed for the new object and what we had observed before by processing all other similar objects in the database. These tables are extensively explained in section 5, which steps into an example for registering an ATM in Paris. Another feature of the report page is that the user can also customize the results by applying a combination of filters. One sample filter is shown in Figure 2 which can find all objects with co-existence confidence ratio bigger than 80 for all 15 clusters including the condition of all objects with similarity less than 40.


Figure 2. Sample filter to customize the reports in the platform

Table 1. List of 22 selected POIs - Tags

| Key | Value | Key | Value |
|---|---|---|---|
| Amenity | Atm | Amenity | Restaurant |
| Tourism | Artwork | Highway | Traffic_signals |
| Amenity | Bench | Shop | Clothes |
| Shop | Bakery | Leisure | Playground |
| Amenity | Recycling | Amenity | Bank |
| Shop | Hairdresser | Shop | Supermarket |
| Historic | Memorial | Highway | Bus_stop |
| Amenity | Bar | Amenity | Cafe |
| Amenity | Post_box | Amenity | Bicycle_parking |
| Amenity | Fast_food | Emergency | Fire_hydrant |
| Shop | Convenience | Office | Company |

## 5. SAMPLE ANALYSIS

This section presents sample results from the analysis of ATMs in Paris. In this analysis, a user tries to register a new ATM at a corner close to a central street in Paris. The platform reports that 803 ATMs exist in Paris. Figure 3 shows the spatial distribution of all ATMs with 1000 meters radius circle around them to indicate the scale of processing for each ATM. Later we review the reports which are generated by the platform.

The first result table shown in Table 2 reveals some basic information about the current position of the newly registered ATM. Based on the previously defined buffer size for each object type, all objects for which the new ATM is located inside

their buffer zone are reported. For example, an object named as "Rue Beaubourg" which is a secondary type street had 4 meters buffer zone and the ATM was inside this range. This table is only intended to provide a quick overview of nearby objects, without considering any other criteria.

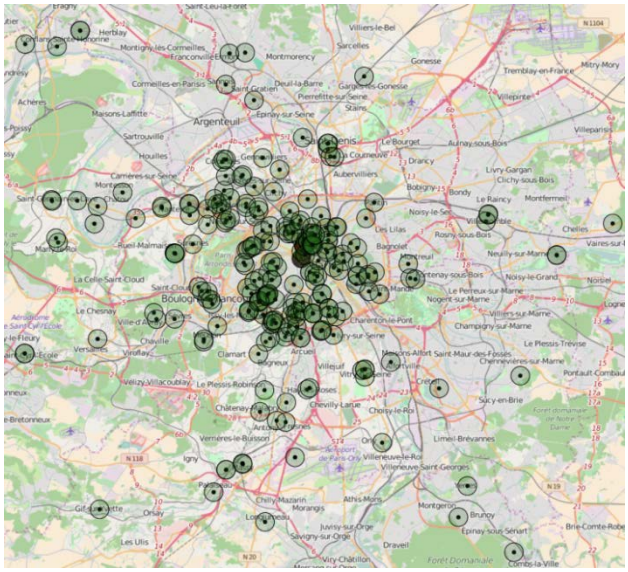

Figure 3. 803 ATM points and clusters in Paris

Another report is about any existing duplicate objects by searching for similar objects within the predefined buffer distance. In our scenario, no duplicates are found but if there were any duplicates, the tag information would be reported. Next will be three statistical analyses of all RP and TP objects that might link to ATMs in Paris. The first table reports all objects that exist around the currently registered point and are observed in our pre-processing as well. The second table represents all objects that do not exist around the newly registered point but were observed before around other ATMs in Paris. The third table lists all other objects, which exist around the currently registered point but are not observed around other ATMs in Paris. The union of all these three tables will cover all existing ATM in the database for Paris within 1000 meters radius of the registered point.

Table 2 List of objects with the new ATM in their buffer zone

| Name | Key | Value | Type | Buffer Size |
|---|---|---|---|---|
| Rue Beaubourg | highway | secondary | Line | 4 |
| Paris | boundary | administrative | Polygon | 20 |
| Paris | boundary | administrative | Polygon | 20 |
| Paris | boundary | administrative | Polygon | 20 |
| Le Marais | place | locality | Polygon | 20 |
| 4e Arrondissement | boundary | administrative | Polygon | 20 |
| Quartier Saint-Merri | boundary | administrative | Polygon | 20 |
| | boundary | political | Polygon | 20 |

As previously mentioned, these three tables are generally reporting statistical numbers about all 803 ATMs in Paris and they are not intended to compare any objects with the newly registered ATM. The numbers in these tables indicate the number of times this type of object has been discovered around each individual ATM in Paris for each specified cluster. For example, 1007 for building:yes in cluster 1 in Table 3 means 1007 buildings have been totally counted around 803 ATMs in Paris within 10 meters distance (cluster 1) which includes duplicate counts as well. As seen in Table 3, most of the ATMs in Paris have a strong co-existing connection with administrative and political boundaries, buildings, other ATMs, residential lands, banks, post_offices, residential roads, bus routes and pedestrian passages. Also, Table 4 shows information about all objects, which do not exist around the current ATM but exist around the other ATMs in Paris. Some of these objects are tagged as postal_code, residential buildings and streets, retail places, marketplace, apartments and fast_food. As seen in both Table 3 and Table 4, the manually sorted results are descending based on cluster 1. This sorting feature helps us identify those objects, which have a higher support rate in different clusters.

The next two tables show the query results for nearby objects.

Table 5 reports all objects which are polygon shaped and for which the newly registered ATM is located inside them. Table 6 also shows all TP objects within 500m distance of the newly registered ATM.

Table 3. List of objects which exists around current ATM and were observed before as well – only cluster 1 to 5

| Key | Value | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| boundary | administrative | 2736 | 2986 | 3197 | 3342 | 3440 |
| boundary | political | 1054 | 1154 | 1246 | 1298 | 1354 |
| building | yes | 1007 | 1621 | 2770 | 4090 | 5516 |
| amenity | atm | 883 | 42 | 50 | 76 | 60 |
| landuse | residential | 673 | 726 | 769 | 812 | 835 |
| amenity | bank | 259 | 206 | 78 | 69 | 76 |
| amenity | post_office | 86 | 111 | 71 | 56 | 54 |
| highway | residential | 79 | 325 | 526 | 705 | 871 |
| route | bus | 69 | 674 | 1193 | 1513 | 1732 |
| highway | footway | 63 | 151 | 254 | 337 | 439 |
| highway | pedestrian | 62 | 122 | 192 | 252 | 305 |
| natural | tree | 57 | 254 | 331 | 440 | 552 |
| man_made | surveillance | 43 | 20 | 14 | 20 | 15 |
| amenity | post_box | 37 | 47 | 28 | 29 | 36 |

Table 4. List of objects which do not exists around current ATM but were observed before

| Key | Value | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| boundary | postal_code | 14 | 14 | 14 | 14 | 17 |
| building | residential | 13 | 20 | 28 | 36 | 38 |
| landuse | retail | 10 | 12 | 13 | 15 | 15 |
| building | retail | 8 | 8 | 10 | 12 | 14 |
| building | apartments | 6 | 9 | 14 | 19 | 24 |
| building | office | 6 | 9 | 13 | 20 | 21 |
| building | train_station | 3 | 4 | 4 | 4 | 4 |
| building | commercial | 2 | 5 | 6 | 8 | 10 |
| emergency | phone | 2 | 1 | 1 | 0 | 4 |
| highway | track | 2 | 4 | 7 | 10 | 14 |
| landuse | commercial | 2 | 2 | 2 | 2 | 3 |
| aeroway | aerodrome | 1 | 1 | 1 | 1 | 1 |
| aeroway | terminal | 1 | 1 | 1 | 1 | 1 |
| barrier | cycle_barrier | 1 | 1 | 0 | 0 | 4 |
| barrier | turnstile | 1 | 2 | 1 | 3 | 9 |

Table 5. List of objects in which new ATM is located

| tag | value | name | type |
|---|---|---|---|
| boundary | administrative | Paris | Polygon |
| boundary | administrative | Paris | Polygon |
| boundary | administrative | Paris | Polygon |
| boundary | political | | Polygon |
| boundary | administrative | Quartier Saint-Germain-l'Auxerrois | Polygon |

| | | | |
|---|---|---|---|
| boundary | administrative | Ler Arrondissement | Polygon |
| tourism | museum | Le Louvre | Polygon |
| historic | castle | Le Louvre | Polygon |
| landuse | residential | Ler Arrondissement | Polygon |

Table 6. All TP objects in radius 500m of the registered ATM

| tag | value | distance | type |
|---|---|---|---|
| highway | footway | 8.8443 | Line |
| highway | primary | 21.9628 | Line |
| route | bus | 26.7752 | Line |
| route | bicycle | 26.7752 | Line |
| highway | tertiary | 29.5961 | Line |
| route | hiking | 30.6429 | Line |
| highway | service | 33.6827 | Line |
| leisure | park | 46.3868 | Polygon |
| Waterway | riverbank | 58.9076 | Polygon |
| barrier | retaining_wall | 140.4433 | Line |
| boundary | administrative | 148.6306 | Line |
| boundary | political | 148.6306 | Line |
| waterway | river | 154.1471 | Line |
| barrier | wall | 173.4831 | Polygon |
| natural | water | 173.4831 | Polygon |
| natural | wood | 209.485 | Polygon |
| highway | pedestrian | 249.8915 | Line |
| highway | cycleway | 256.3833 | Line |
| railway | rail | 263.7275 | Line |
| route | train | 263.7275 | Line |
| highway | residential | 290.8675 | Line |
| highway | primary_link | 311.0095 | Line |
| highway | secondary | 364.8004 | Line |
| highway | steps | 414.3831 | Point |

The main analytical tables will appear next in the result screen. Two major tables will cover information about TP and RP objects separately. As seen in Table 7, each table has three rows for each object. These rows are tagged as P, R and S respectively. The tags represent the aggregation of analytical values for each doughnut cluster, which helps us quantitatively measure the probability of positional acceptance of a proposed POI. Tag P refers to the concept of probability of co-existence which is the confidence ratio of association between ATMs and other nearby objects. As we mentioned earlier, the value of P is preprocessed in the database and is calculated for all 22 sample POIs for each cluster through the following formula:

(a) $P_{cluster(x)}(i) = \frac{total\ number\ of\ observation\ in\ cluster\ x}{total\ number\ of\ instances\ for\ object\ i}$

In this context, we refer to P as the probability of co-existence of two objects. We can call it the confidence or support of the association rule between these two objects. A higher confidence rate in each cluster means they might have a stronger connection in terms of their location and thus they could be highly correlated throughout the city. Tag R refers to the status of the registered point which is the binary test result for the existence of all objects around the newly proposed point in each cluster. It has only values 0 and 1, which is the result of an intersection test of each object with each cluster of the new POI. Tag S reveals the similarity between an ATM and the current object in each cluster. We use the following formula to calculate the distance of P and R and then convert it to a similarity metric:

(b) $Distance(i,j)_{Cluster(x)} = |(P - (R * 100)|$

(c) $Sim(i,j)_{cluster(x)} = 100 - Distance(i,j)_x$

Formula (b) represents the distance between object i and j in cluster x. Formula (c) converts the distance value to the similarity between object i and j for cluster x.
Note: we normalized R (which was 0 or 1) to 0 and 100 to make it comparable with P values.

For example, in Table 7 the P value for boundary:administrative (object X) for cluster 1 is 98.63 which means for 98.63% of all ATMs in Paris, at least one boundary: administrative is found within 10m distance of the ATM. The R value for all 15 clusters is 1, which means for the current registered ATM, at least one ATM was close or had intersected with boundary:administrative in all these clusters. The S value is the reverse of the distance value between P and R. As P represents the generic status of objects in relation to ATMs in Paris, and R represents the current value for registered ATM.

The vector of S values for 15 clusters will give us valuable information to assess whether the newly registered ATM is following the same patterns of existing ATMs in Paris. Table 7 represents a subset of P, R and S vectors accordingly. Due to limited space, only values for the first 5 clusters are shown in this paper.

Table 7 Partial result for analysis of TP objects

| Type | Key | Value | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|---|
| P | boundary | administrative | 98.63 | 98.38 | 98.51 | 98.51 | 98.38 |
| R | boundary | administrative | 1 | 1 | 1 | 1 | 1 |
| S | boundary | administrative | 98.63 | 98.38 | 98.51 | 98.51 | 98.38 |
| P | boundary | political | 95.39 | 95.27 | 95.52 | 95.64 | 95.52 |
| R | boundary | political | 1 | 1 | 1 | 1 | 1 |
| S | boundary | political | 95.39 | 95.27 | 95.52 | 95.64 | 95.52 |
| P | highway | residential | 9.34 | 32.5 | 45.7 | 54.67 | 61.39 |
| R | highway | residential | 0 | 1 | 1 | 1 | 1 |
| S | highway | residential | 90.66 | 32.5 | 45.7 | 54.67 | 61.39 |
| P | highway | pedestrian | 7.22 | 12.7 | 18.56 | 22.67 | 24.16 |
| R | highway | pedestrian | 0 | 0 | 1 | 1 | 1 |
| S | highway | pedestrian | 92.78 | 87.3 | 18.56 | 22.67 | 24.16 |
| P | highway | footway | 6.23 | 12.33 | 18.18 | 22.79 | 26.03 |
| R | highway | footway | 0 | 0 | 1 | 1 | 1 |
| S | highway | footway | 93.77 | 87.67 | 18.18 | 22.79 | 26.03 |

To demonstrate how the P value helps to determine data quality, we sorted the previous result table based on P for all RP objects. Table 8 shows the result of this sort. The sort is done on cluster 1. As seen, having cluster 1 as major selection criteria, some objects such as building:yes, landuse:residential, or amenity:bank have strong relationships with ATMs.

Table 8 Partial result for analysis of RP objects with high co-existence ratio with ATMs sorted on cluster 1 - only 5 clusters

| Key | Value | 1 | 2 | 3 | 4 | 5 | Sum | Avg |
|---|---|---|---|---|---|---|---|---|
| building | yes | 93.15 | 94.27 | 95.77 | 96.26 | 97.01 | 1473.49 | 98.233 |
| landuse | residential | 75.84 | 76.09 | 76.21 | 76.34 | 76.71 | 1227.01 | 81.801 |
| amenity | bank | 32.25 | 25.03 | 9.59 | 7.85 | 8.84 | 526.52 | 35.101 |
| amenity | post_office | 10.71 | 13.7 | 8.84 | 6.97 | 6.72 | 237.59 | 15.839 |
| natural | tree | 5.11 | 12.95 | 15.19 | 15.94 | 17.43 | 512.58 | 34.172 |
| amenity | post_box | 4.48 | 5.6 | 3.49 | 3.61 | 4.23 | 532.26 | 35.484 |
| man_made | surveillance | 4.36 | 1.87 | 1.37 | 1.99 | 1.62 | 213.21 | 14.214 |
| shop | mall | 2.49 | 2.74 | 2.74 | 2.86 | 3.36 | 87.67 | 5.845 |
| amenity | telephone | 1.99 | 3.86 | 3.36 | 4.48 | 3.61 | 443.07 | 29.538 |
| amenity | parking | 1.74 | 4.73 | 7.85 | 11.08 | 13.08 | 593.41 | 39.561 |
| building | residential | 1.62 | 1.74 | 1.87 | 2.37 | 2.24 | 134.63 | 8.975 |
| highway | crossing | 1.25 | 17.93 | 20.67 | 24.03 | 26.15 | 955.9 | 63.727 |
| landuse | retail | 1.25 | 1.49 | 1.62 | 1.87 | 1.87 | 35.36 | 2.357 |

The comparison of P and R which leads to S will help to generate a composite score to determine whether the proposed location for the new object is properly adjusted or not. While P gives us a general pattern of association between two object types in one specific city, S helps us evaluate the current object

with existing P patterns. One approach for generating a recommendation score is to use the S value for all objects in the city to generate scoring criteria. For example, the average of all S values would give us a good ranking score. If the new object has a high S average, we can infer that its position is highly acceptable. However, in this simple formula, all objects are equally treated in the average, while many of them might not add any value to the final score and the final score might be biased. For example, if none of the ATMs in Paris were close to any bakery shop, the P value would be 0. If the newly registered ATM was also far from the bakery, it has 100% similarity to the P value. This means the average score would be elevated just by bringing bakeries into the calculation.

Using the OSM POI analyser, there are many possibilities to explore in terms of relationships between two objects in different cities of the world. By changing the combination of filters for P and S value and recording the results for different new POIs in different cities, many different discoveries are possible, which are left to researchers to explore. Some interesting questions to explore are:

- Which object has the highest P value in relation to a tree?
- Does this object have the same pattern in all cities?
- Which objects have similar P values in all cities?
- Is it possible to reason automatically where the best location to register a new object type is?
- Which ATM in Paris has a lower ranking score compared to the 803 ATMs if we try to remove and re-register it again?

## 6. CONCLUSION

This paper introduced and explored the key features of an analytic system for OSM to discover the spatial association rules between different types of POIs and their surrounding objects. We believe that such analytical tools will shed light on understanding how our cities are built and inspire other researchers to explore more challenging problems to improve the quality of geodata in crowdsourcing projects. While the report covers essential information about the fundamental concepts of the system, it does not exhibit comparative results between sets of objects. A part of currently undergoing research focuses on techniques for POI feature extraction to use machine learning classifiers to predict best location for one certain type of POI. In addition, extending and completing the current recommendation engine can help to rank existing POIs in terms of positional credibility as a new module in OSM if adopted by the community. Our research trend is continuing with mining multiple association patterns to extract hidden and interesting knowledge behind the maps.

## ACKNOWLEDGMENTS

We would like to thank the Centre of Disaster Management and Public Safety (CDMPS) at the University of Melbourne to support this research which was included in the context of using VGI for designing disaster management online applications.

## REFERENCES

Agrawal, R., Srikant, R., 1994. Fast algorithms for mining association rules, Proc. 20th int. conf. very large data bases, VLDB, pp. 487-499.

Amirian, P., Basiri, A., Gales, G., Winstanley, A., McDonald, J., 2015. The next generation of navigational services using openstreetmap data: the integration of augmented reality and graph databases, OpenStreetMap in GIScience. Springer, pp. 211-228.

Arsanjani, J.J., Mooney, P., Zipf, A., Schauss, A., 2015. Quality assessment of the contributed land use information from OpenStreetMap versus authoritative datasets, OpenStreetMap in GIScience. Springer, pp. 37-58.

Blasby, D., Davis, M., Kim, D., Ramsey, P., 2002. GIS conflation using open source tools. The Jump Project. http://citeseerx.ist.psu.edu.

Ching-Chien Chen, C.A.K., 2008. Conflation of Geospatial Data, in: Shashi Shekhar, H.X. (Ed.), Encyclopedia of GIS. Springer US.

Data, S., Social, N., Elwood, S., Goodchild, M.F., Sui, D.Z., 2012. Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice. Annals of the Association of American Geographers 102, 571-590.

Fan, H., Zipf, A., Fu, Q., Neis, P., 2014. Quality assessment for building footprints data on OpenStreetMap. International Journal of Geographical Information Science 00, 1-20.

Grant McKenzie, K.J., Benjamin Adams, 2013. Weighted Multi-Attribute Matching of User-Generated Points of Interest, 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, pp. 440-443

Hashemi, P., Abbaspour, R.A., 2015. Assessment of logical consistency in OpenStreetMap based on the spatial similarity concept, OpenStreetMap in GIScience. Springer, pp. 19-36.

Helbich, M., Amelunxen, C., Neis, P., 2012. Comparative Spatial Analysis of Positional Accuracy of OpenStreetMap and Proprietary Geodata, Proceedings of GI_Forum 2012: Geovisualization, Society and Learning.

Iacono, M., Krizek, K.J., El-Geneidy, A., 2010. Measuring non-motorized accessibility: issues, alternatives, and execution. Journal of Transport Geography 18, 133-140.

Koukoletsos, T., Haklay, M., Ellul, C., 2012. Assessing Data Completeness of VGI through an Automated Matching Procedure for Linear Data. Transactions in GIS 16, 477-498.

Neis, P., Goetz, M., Zipf, A., 2012. Towards Automatic Vandalism Detection in OpenStreetMap. ISPRS International Journal of Geo-Information 1, 315-332.

Neis, P., Zielstra, D., 2014. Recent Developments and Future Trends in Volunteered Geographic Information Research: The Case of OpenStreetMap. Future Internet 6, 76-106.

Salk, C.F., Sturn, T., See, L., Fritz, S., Perger, C., 2016. Assessing quality of volunteer crowdsourcing contributions: Lessons from the Cropland Capture game. International Journal of Digital Earth 9, 410-426.

Tobler, W.R., 1979. Cellular geography, Philosophy in geography. Springer, pp. 379-386.

Yuan, S., Tao, C., 1999. Development of conflation components. Proceedings of Geoinformatics, Ann Arbor, 1-13.