Outlier Detection in OpenStreetMap Data using the Random Forest Algorithm

Richard Wen, Claus Rinner

Department of Geography and Environmental Studies, Ryerson University
350 Victoria St., Toronto, Ontario, M5B 2K3, Canada
Email: {rwen, crinner}@ryerson.ca

Abstract

OpenStreetMap (OSM) data consist of digitized geographic objects with semantic tags assigned by the volunteer contributors. The tags describe the geographic objects in a way that is understandable by both humans and computers. The variability in contributor behaviour creates reliability concerns for the tagging quality of OSM data. The detection of irregular contributions may improve OSM data quality and editing tools. This research applies the random forest algorithm on geospatial variables in order to detect outliers without ground-truth reference data to direct human inspection. An application to OSM data for Toronto, Ontario, was effective in revealing abnormal amenity tagging of school and hospital objects.

1. Introduction

OpenStreetMap (OSM) is an online platform enabling registered volunteers to contribute geospatial data by digitizing point-, line-, or polygon-shaped geographic objects and annotating them with tags referring to common feature classes such as roads and restaurants (Haklay 2008). OSM tags are semantically structured as key-value pairs, where the key refers to a broad class of geographic objects and the value details the specific geographic object being tagged (Ballatore *et al*. 2013). Examples of tags are *amenity=school*, *highway=residential*, and *building=house*.

The open and flexible nature of OSM tagging leads to varying contribution behaviour by different communities (Mooney *et al*. 2010). The varying contribution behaviour creates concerns about the quality of OSM data and the community standards of OSM tagging. Quality control and corrections rely heavily on human interaction, which raises additional questions on the reliability of OSM data. Finally, the experience of the volunteer contributor has an effect on the tagging quality of each geographic object as experienced contributors are more familiar with the tagging norms of the area being edited. Although OSM is an effective and efficient platform for generating masses of geospatial data, it is plagued by reliability, quality, and completeness issues.

The aim of this paper is to examine the ability of an automated machine learning algorithm, the random forest algorithm, to support manual human inspection and minimize bias in OSM data editing. The use of an automated algorithm improves the detection of abnormal tagging behaviour, avoids the bias of human judgement, and reduces the time required to search through masses of tagged geographic objects. A combination of human knowledge and experience with the logical accuracy of machines could improve OSM tagging quality and standards, and enable the development of advanced editing tools.

2. Data and Methods

OSM data for the City of Toronto, Ontario, were downloaded from Mapzen Metro Extracts in the form of a GEOJSON file (Mapzen 2016). A GEOJSON file contains one or more spatial objects described by geometry types and properties in a key/value data structure (Butler *et al*. 2008). The OSM key category datasets amenities, places, transport areas, aero ways, transport points, and roads were selected to be used from the downloaded data. The selected data consisted of 70,535 geographic objects in the City of Toronto detailed in Table 1. The majority of geographic objects resided in the transport points and roads datasets. The data were projected from a geographic coordinate system (WGS 1984) into a planar coordinate system (NAD 1983 UTM Zone 17 North) for geometric calculations. A tag value is referred to as a tag in this paper.

Table 1. OpenStreetMap Data for City of Toronto, Ontario from Mapzen.

Key Category Dataset	Tag Values Available	Geometry Type	Count
Amenities	fire_station, fuel, hospital, library, police, school, townhall, university	Point	1507
Places	city, county, hamlet, locality, neighbourhood, suburb, town, village	Point	760
Transport Areas	aerodrome, apron, helipad, platform, station, terminal	Polygon	72
Aero Ways	runway, taxiway	Line	438
Transport Points	aerodrome, bus_stop, crossing, gate, halt, helipad, level_crossing, motorway_junction, station, subway_entrance, terminal, tram_stop, turning_circle	Point	21,309
Roads	disused, monorail, motorway, motorway_link, preserved, primary, primary_link, rail, secondary, secondary_link, subway, tertiary, tertiary_link, tram, trunk, trunk_link	Line	46,812

The methods first required the extraction of geospatially meaningful variables to describe the geometric characteristics and spatial relationships of each geographic object. The first set of geospatial variables were extracted by utilizing the geometric structures in the data. Area, length, and the number of vertices for each geographic object in the data were extracted as columns. These geospatial variables describe the geographic objects in terms of geometric characteristics such as size (area) and geometric complexity (vertices). Representative coordinates were also extracted in the form of the x- and y-coordinates of the centre of each object. These representative coordinates allowed the random forest model to utilize spatial patterns if they existed. The second set of geospatial variables were the distances to the nearest uniquely tagged geographic object for each individual geographic object. A column was created for each of the available tag values. These columns were referred to as the Distance to the Nearest Neighbour Tag (DNNT). An example of the DNNT concept is seen in Figure 1.

Redundant variables were removed by examining each variable for a high correlation (less than -0.7 and greater than 0.7) to another variable and removing the other highly correlated variables in a specified order. The order arranged the area, length and vertices first, followed by sorting the DNNT variables by their tag frequency. The result of the extracted geospatial variables after removing redundant variables is referred to as the input data in this paper.

Figure 1. Distance to the Nearest Amenity Tag for a Police Station Object.

Several random forest models were run on the input data to classify the tag value of geographic objects. A random forest consists of a number of decision trees built on subsamples of approximately two-thirds of the input data (Breiman 2001). The other one-third of the subsamples are used to calculate an out-of-bag error estimate by aggregating the predictive scores (Liaw and Wiener 2002). Each random forest model used balanced tag weights, penalizing misclassification of minority tags, to adjust for tag frequency imbalances in the data (Chen *et al*. 2004). A number of maximum split variables equal to the square root of the number of variables in the input data were used for each decision tree in the random forest models. Three random forests models were constructed to optimize the number of decision trees using 64, 96, and 128 decision trees as suggested by Oshiro *et al*. (2012) to determine the model with the lowest out-of-bag error estimate. The selected model with the lowest out-of-bag error estimate is referred to as the Tree Optimized Random Forest (TORF) model in this paper.

The TORF model was used to determine outliers in the input data by calculating proximity measures between two geographic objects (Louppe 2014) to produce proximity matrices for geographic objects inside each tag followed by calculating outlier measures (eq. 1) according to Breiman and Cutler (2004) for each geographic object.

$$outlier(n_c) = \frac{N}{\sum_{k_c}^{K}[proximity(n_c, k_c)]^2} \quad \text{(eq. 1)}$$

where n_c is a sample instance of tag c, k_c is all other sample instances of tag c, K is the total number of k_c, and N is the total number of n samples. The outlier measures were then normalized by subtracting every outlier value for n instances of each tag c by the median of all outlier measures inside the same tag c, and dividing by the absolute deviation from the median.

A geographic object was suspected of being an outlier if its normalized outlier measure was greater than 10.

The interpretation of outliers was enhanced by using local variable contribution increments (eq. 2) to calculate variable contributions (eq. 3) according to Palczewska *et al* (2014) for the outlier tags. The variable contributions were ranked to find the most influential variables.

$$LI_f^c = \begin{cases} Y_{mean}^c - Y_{mean}^p & if\ split\ of\ p\ is\ for\ f \\ 0 & otherwise \end{cases} \quad \text{(eq. 2)}$$

where LI is the local variable contribution increment, f is a variable, c is the child node, p is the parent node, Y_{mean}^c is the fraction of training samples in a child node, and Y_{mean}^p is the fraction of training samples in a parent node.

$$FC_i^f = \frac{1}{T}\sum_{t=1}^{T} FC_{i,t}^f \quad \text{(eq. 3)}$$

where FC_i^f is the variable contribution of a training sample, $FC_{i,t}^f$ is the sum of local variable specific contribution increments, f is a variable, T is the total number of trees in the forest, t is a tree in the forest, and i is a training sample.

3. Results

The TORF model was obtained from 128 trees, which provided the lowest out-of-bag error of 0.162 compared to 0.166 and 0.167 for 96 and 64 trees respectively. The schools in Figure 2 and the hospitals in Figure 3 had normalized outlier measures above 10. Closer inspection of the schools revealed that the schools are historical and were far away from bus stops. The hospitals were individual wings of Sunnybrook hospital, which were further away from secondary roads than normal.

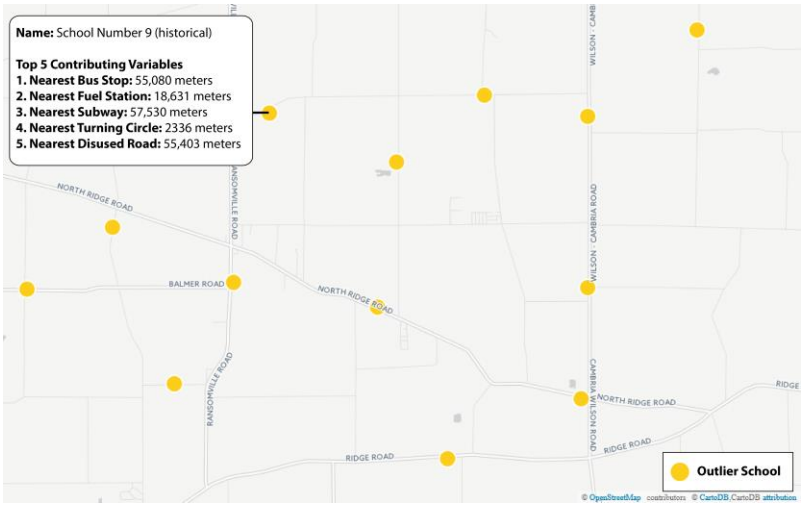

Figure 2. Detected Tag Outliers of Value School

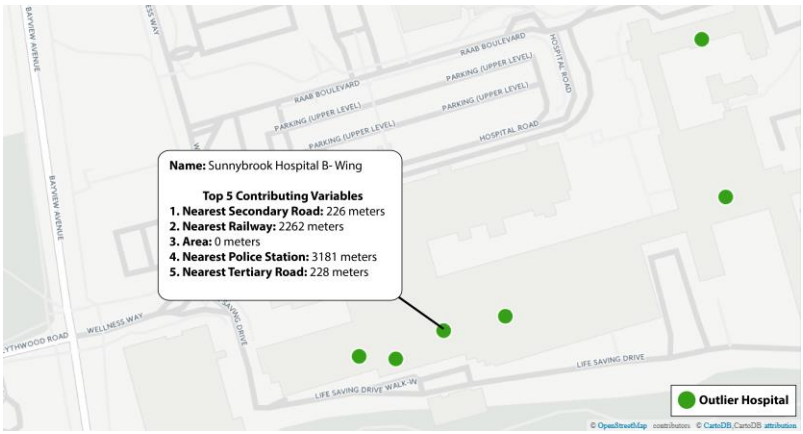

Figure 3. Detected Tag Outliers of Value Hospital

4. Conclusion

The use of random forests for outlier detection has the potential to support manual data cleaning efforts, where the discovery of potential outliers that may yield insight into the community tagging standards and errors in a study area. The outlier interpretation was also enhanced by the variable contributions, which may provide reasons for abnormal tags. However, satellite imagery and user editing history were not used, although they may have a significant impact on outlier detection. Only nearest neighbour objects were used, other spatial relations such as distance buffers should be tested in the future. Adding raster data, temporal data, and a variety of spatial relations to the random forest model could further improve the outlier detection of OSM tags.

Acknowledgements

To be added after review.

References

Ballatore A, Bertolotto M and Wilson DC, 2013, Geographic knowledge extraction and semantic similarity in OpenStreetMap, *Knowledge and Information Systems*, 37(1):61-81.

Breiman L, 2001, Random Forests, *Machine Learning*, 45(1):5-32.

Breiman L, Cutler A, 2004, Random Forests, Retrieved from: https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#outliers

Butler H, Daly M, Doyle A, Gillies S, Schaub T and Schmidt C, 2008, The GeoJSON Format Specification, Retrieved from: https://datatracker.ietf.org/doc/draft-ietf-geojson/

Chen C, Liaw A, and Breiman L, 2004, Using random forest to learn imbalanced data, University of California, Berkeley.

Haklay MM, 2008, OpenStreeMap: User-Generated Street Maps, *Pervasive Computing*, 12-18.

Liaw A and Wiener M, 2002, Classification and Regression by randomForest, *R News*, 18-22.

Louppe G, 2014, Understanding Random Forests: From Theory to Practice, University of Liege, Belgium.

Mapzen, 2016, Metro Extracts, Retrieved from: https://mapzen.com/data/metro-extracts/

Mooney P, Corcoran P and Winstantly AC, 2010, Towards Quality Metrics for OpenStreetMap, *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, New York, USA, 514-517.

Oshiro, TM, Perez PS and Augusto J, 2012, How many trees in a random forest?, *Machine Learning and Data Mining in Pattern Recognition*, 7376: 154-168.

Palczewska A, Palczewski J, Robinson RM and Neagu D, 2014, Interpreting random forest classification models using a feature contribution method. *Integration of Reusable Systems*, 26:193-218.