Outlier Detection in OpenStreetMap Data using the Random Forest Algorithm and Variable Contributions

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 consists of digitized geographic objects with semantic tags assigned by volunteer contributors. These human and machine readable tags are edited manually and automatically to improve data quality. The structure of the tags allow machine learning algorithms to support user editing by learning to identify irregular objects and data patterns. This research experimented with a random forest algorithm on geospatial variables for geospatial outlier detection and knowledge discovery in OSM data without ground-truth reference data.

1. Introduction

OpenStreetMap (OSM) is a global crowdsourcing initiative and online platform enabling registered volunteers to contribute geospatial data by digitizing and annotating point-, line-, or polygon-shaped geographic objects with tags of common feature classes such as roads and restaurants (Haklay 2008). OSM tags are semantically structured as key-value pairs, where the key is a class of geographic objects and the value is a specific object being tagged (Ballatore *et al.* 2013). Examples of tags are *amenity=school*, *highway=residential*, and *building=house*. OSM data quality and quantity is dependent on user contributions (Mooney *et al.* 2010).

Users without basic knowledge of OSM often experience a moderate to steep learning curve which can result in inconsistent edits or discouraged participation. To enhance the accuracy of volunteered geographic data, previous research has examined "spatial-semantic interactions" (Mülligann *et al.* 2011) and "thematic signatures" (Adams and Janowicz 2015). The aim of the present research was to experiment with a random forest model for geospatial outlier detection and knowledge discovery to support user editing.

2. Data and Methods

OSM data for the City of Toronto, Ontario, were downloaded from Mapzen Metro Extracts (Mapzen 2016). The key datasets amenities, places, transport areas, aero ways, transport points, and roads were chosen to be used. The data consisted of 70,535 geographic objects in the City of Toronto detailed in Table 1. 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.

The methods required the extraction of geospatial variables for the random forest algorithm to learn from. Area, length, and the number of vertices for each geographic object in the data were extracted to learn geometric characteristics. Representative coordinates, x- and y-coordinates for the object centres, were extracted to learn locational patterns.

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

Keys	Tag Values	Geometry	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 Distances to the Nearest Neighbour Tag (DNNT) were used to learn spatial relationships (Figure 1). Redundant variables were automatically removed if a variable had a high correlation (< -0.7 and > 0.7) to another variable. The order arranged the area, length and vertices first, followed by sorting the DNNT variables by their tag frequency. The result of the geospatial variables after removing redundant variables is referred to as the input data in this paper.

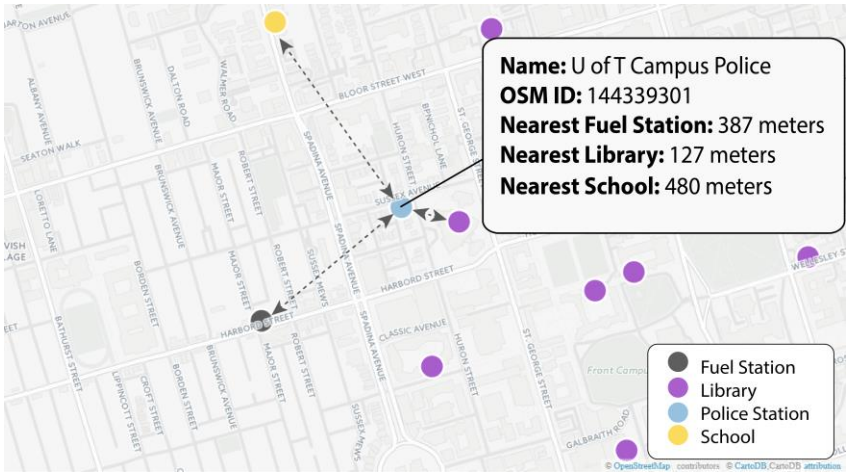

Figure 1. Distance to the Nearest Amenity Tag for a Police Station Object.

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 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 models. Three 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. Outliers were geographic objects that had irregular geospatial variable values. Proximity matrices between two geographic objects (Louppe 2014) were used to calculate outlier measures (eq. 1) according to Breiman and Cutler (2004).

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

where n_c is a sample of tag c, k_c is all other samples of tag c, K is the total number of k_c, and N is the total number of n_c. 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.

Local variable contribution increments (eq. 2) were used to calculate variable contributions (eq. 3) according to Palczewska *et al* (2014) for the outlier tags.

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

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. The variable contributions were ranked from highest to lowest values to determine the most influential variables.

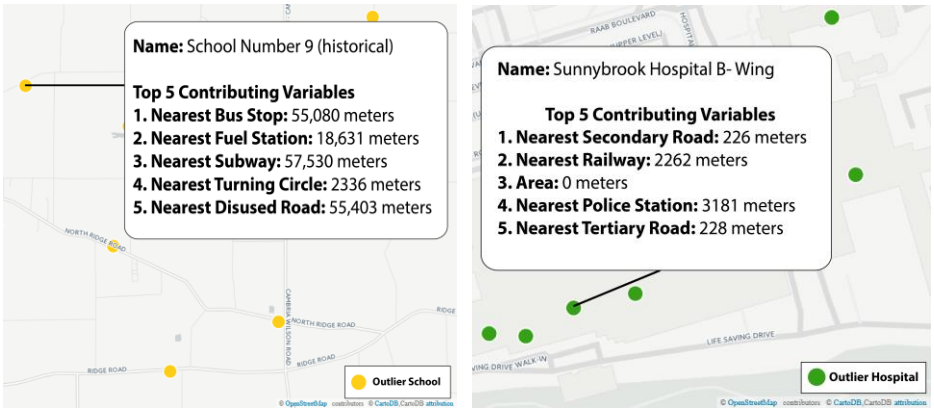

Figure 2. Detected Tag Outliers of Value School and Hospital.

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 and hospitals in Figure 2 had normalized outlier measures above 10. The schools were 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.

4. Limitations

Satellite imagery, which may improve model classification, and user editing history, which may improve outlier detection, were not used in this research. Users who digitize geographic objects often use satellite imagery as a reference. User histories can provide insight into user-specific editing behaviour and contribution quality of objects, but were not included in this approach. Other spatial relations such as distance buffers and multiple nearest neighbours should be experimented with to explore their effects on model classification.

5. Conclusion

The use of random forests for geospatial outlier detection and knowledge discovery can support user editing, which may improve OSM data quality and quantity. Potential objects that require editing and their influential characteristics were automatically obtained with the random forest model without requiring the geographic knowledge of experienced users. This approach could encourage participation in volunteered geographic data collection by lowering search times for irregular objects, and improve the geographic knowledge of inexperienced contributors by providing the most influential geospatial variables for each object.

References

Adams B and Janowicz K, 2015, Thematic Signatures for Cleansing and Enriching Place-Related Linked Data. *International Journal of Geographical Information Science*, 29(4):556-579

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

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, *Proc. of the 18th SIGSPATIAL International Conf. on Advances in Geographic Information Systems*, New York, USA, 514-517.

Mülligann C, Janowicz K, Ye M and Lee W-C, 2011, Analyzing the Spatial-Semantic Interaction of Points of Interest in Volunteered Geographic Information. In MJ Egenhofer, NA Giudice, R Moratz and MF Worboys (eds), *Conference on Spatial Information Theory (COSIT 2011)*, 350-370. Berlin: Springer

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.