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

Richard Wen and Claus Rinner
Department of Geography and Environmental Studies, Ryerson University

Objectives

This research experimented with random forest models on OpenStreetMap (OSM) data to:

1. Detect outlier geographic objects based on nearest neighbour distances and geometry
2. Determine the influence of variables for the discovered pattern

1.0 Introduction

OpenStreetMap (OSM) is an online platform that enables volunteers to contribute massive quantities of data by creating geographic objects with user assigned tags (Figure 1). These tags are structured as key-value pairs that enable machine learning algorithms to support user editing by identifying outlier objects and discovering patterns. This results in lowered search times for errorneous objects and user knowledge enhancements.

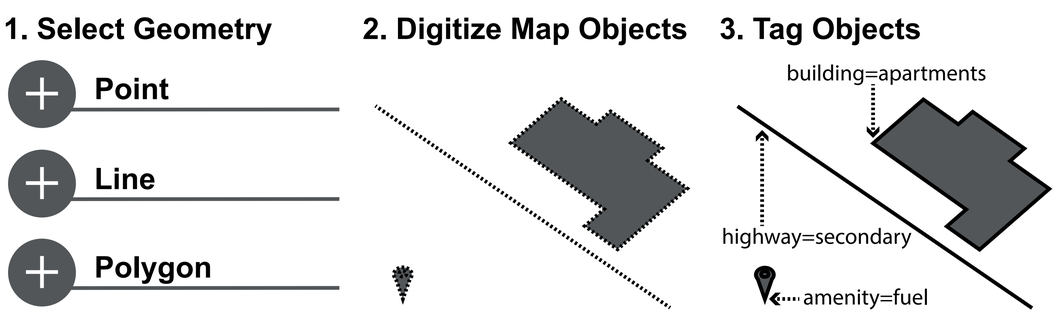

Figure 1. OpenStreetMap Contribution Process

2.0 Data and Methods

The data was downloaded from Mapzen Metro Extracts (Mapzen, 2016) and processed with an automated workflow programmed in Python (Figure 2). The random forest algorithm uses an implementation from the Scikit-learn Python library (Pedregosa et al., 2011).

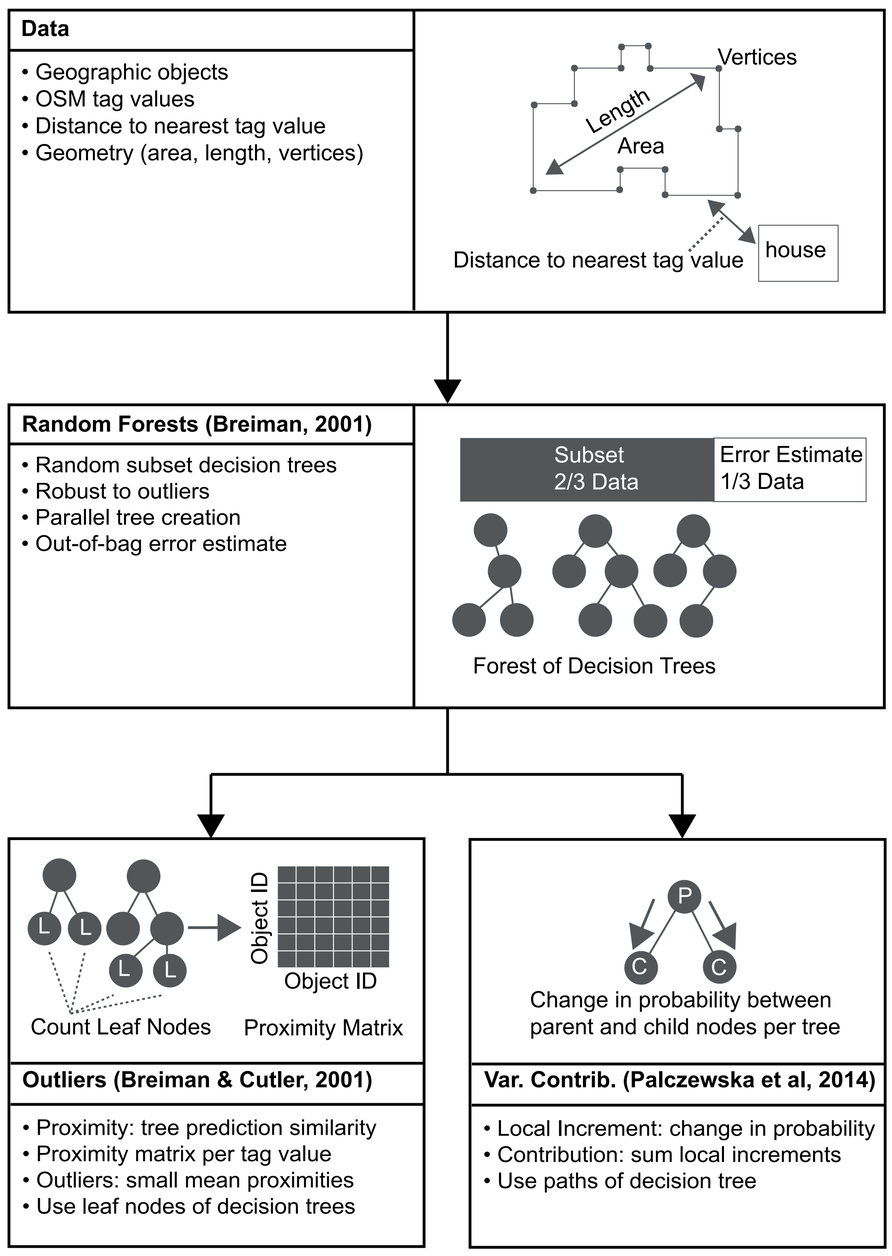

Figure 2. Automated Python Workflow

3.0 Results

A random forest model with 128 trees was used to detect outliers in a selection of Toronto, Ontario, Mapzen data (Table 1). Hospital objects from the Sunnybrook wings (Figure 3) and several historical school objects (Figure 4) were detected as outliers with the top 5 contributing variables using the median as an example.

Table 1. Selected Mapzen OSM Data for Toronto, Ontario

Category	Geometry	Number of Objects
Aero Ways	Line	438
Amenities	Point	1507
Places	Point	760
Roads	Line	46,812
Transport Areas	Polygon	72
Transport Points	Point	21,309

Figure 3. Sunnybrook Hospital Outliers

Figure 4. Historical School Outliers

4.0 Conclusion

User experience improvements using random forest models could:

- Encourage participation in volunteered geographic data collection
- Improve the knowledge of inexperienced contributors

Outlier detection and variable contributions can:

- Lower search times for irregular objects
- Improve the interpretability of data patterns in OSM

To potentially lower the error of the random forest model, it is suggested to explore satellite imagery and user editing histories.

qrd.by/l4j68g

References

- Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32
- Breiman, L., & Cutler, A. (2004). Random Forests. Retrieved March 23, 2016, from Random Forests: https://www.stat.berkeley.edu/~breiman/
- Mapzen. (2016, March 10). Metro Extracts. Retrieved from Mapzen: https://mapzen.com/data/metro-extracts
- Palczewska, A., Palczewski, J., Robinson, R. M., & Neagu, D. (2014). Interpreting random forest classification models using a feature contribution method. Integration of Reusable Systems, 193-218.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 2825-2830.

Acknowledgements
We would like to thank the Geothink Social Sciences and Humanities Research Council (SSHRC) Partnership Grant for the funding provided during the duration of this research. Map data copyrighted OpenStreetMap contributors and available from http://www.openstreetmap.org

Contact
Richard Wen (rwen@ryerson.ca)
Claus Rinner (crinner@ryerson.ca)