SPATIAL EVOLUTION OF OPENSTREETMAP DATASET IN TURKEY

M. Zia\(^{a}\), D. Z. Seker\(^{a}\), Z. Cakir\(^{b}\)

\(^{a}\) ITU, Civil Engineering Faculty, 34469 Maslak Istanbul, Turkey - mohammed.zia33@gmail.com
\(^{b}\) ITU, Faculty of Mines, 34469 Maslak Istanbul, Turkey

KEY WORDS: OpenStreetMap, spatial evolution, time-series

ABSTRACT:

Large amount of research work has already been done regarding many aspects of OpenStreetMap (OSM) dataset in recent years for developed countries and major world cities. On the other hand, limited work is present in scientific literature for developing or underdeveloped ones, because of poor data coverage. In presented study it has been demonstrated how Turkey-OSM dataset has spatially evolved in an 8 year time span (2007-2015) throughout the country. It is observed that there is an east-west spatial biasedness in OSM features density across the country. Population density and literacy level are found to be the two main governing factors controlling this spatial trend. Future research paradigms may involve considering contributors involvement and commenting about dataset health.

1. INTRODUCTION

1.1 Introduction

Volunteered Geographic Information (VGI) ((Goodchild, 2009a), (Li and Qian, 2010), (Haklay, 2009)) or Crowdsourcing Geographic Data ((Heipke, 2010), (Dodge and Kitchin, 2013)) has come into picture since last decade because of easy geo-data creation and upload by human-beings who behave like a sensor (Goodchild, 2007), to add up to the geographic information which was generally being collected and stored by National Mapping Agencies and other private GIS companies (Elwood, 2008). This technology has allowed even the amateur mappers with limited mapping experience to collect, map, and upload geo-data with tangible options of any place to cloud-servers (Wikipedia-OSMTags, 2015). A famous live VGI example is OpenStreetMap (OSM) project (Haklay and Weber, 2008) which started in 2004 with the goal to generate a free and editable street map of the world; in addition to other similar projects like WikiMapia, Wikiloc, Foursquare, Google Map Maker etcetera but with varied visions. This project has recently gained huge fame because of large volume data (mapped by varied geo-data producers, AKA Neo-Geographers ((Goodchild, 2009b), (Haklay et al., 2008)), because of limited editing restrictions), data heterogeneity, abundance, and open access; and thus has attracted immense interest from researchers of various domains (Zhao et al., 2015).

This study has tried to answer how Turkey-OSM dataset has spatially evolved from 2007 to 2015, considering the importance which has already been discussed in many researches in the past ((Zielstra and Alexander, 2010), (Hagenauer and Helbich, 2012), (Corcoran and Mooney, 2013), (Corcoran et al., 2013), (Zhang et al., 2015), (Kuhn, 2007)) like the reconstruction and growth rate estimation of live/future VGI projects.

1.2 Previous Work

By the end of 2015, OSM dataset has already had a huge amount of geo-tagged global data in the form of around 5 billion GPS points, 3 billion nodes, 3 billion ways, and 4 million relations (OpenStreetMap, 2016); contributed by around 2.5 million registered users worldwide. Deep delve on estimating OSM dataset’s completion has always been restricted by rigid licensing policy, limited usage, confined availability, and big pricing of governmental/proprietary geo-data sources which act as reference dataset ((Estima and Painho, 2013), (Haklay, 2009), (Leeuw et al., 2011), (Zielstra and Alexander, 2010)). Nonetheless, recent years have come across some famous contributions about this into scientific literature (for example, Germany-OSM street network data by (Neis et al., 2011) using proprietary dataset, USA-OSM bicycle trail and lane data by (Hochmair et al., 2015) using data from local planning agencies, and USA-OSM street network data by (Zielstra et al., 2013) using TIGER/Line data (Willis, 2008)).

1.3 OSM Effort in Turkey

According to authors, no recognised study considering Turkey-OSM dataset’s spatial evolution is present in online scientific literature. Carrying out current study, thus, is necessary to understand VGI responses at higher resolution (spatial and temporal) in Turkey ((Zhao et al., 2015), (Leeuw et al., 2011), (Zhang et al., 2015)), which has considerably rich OSM dataset (17 million points, 1.3 million edges AKA lines, and 0.4 million polygons (PlanetOSM, 2016)), and which will, eventually, add comments to help expand the same.

2. DATA PROCESSING AND METHODOLOGY

Huge interoperability of OSM geo-data, by having ranging data sources (Full Planet dump file (PlanetOSM, 2016), Geofabrik downloads (Geofabrik, 2015), Overpass (Overpass-API, 2016) API) and formats (ESRI-Shapefiles +.shp, Extensible Markup Language (XML) +.osm, Protocolbuffer Binary Format +.pbf), are considered to be another reason for its increase in popularity. Full Planet dump file (sized approximately 67 GB and 1.5 TB when compressed and uncompressed, respectively), dated September 02, 2015 (last stable history release at the time of processing), was downloaded from (PlanetOSM, 2016) which contained complete OSM database including edit-history from 2007 to September 2015. The file is in human-readable XML format which consists of three basic data elements/features i.e. node (point), way (polyline and polygon), and relation (combination of nodes, ways and/or other relations defining a particular structure), attributed with tags in a key-value structure of free format text fields (OSMElements, 2015).
OSM XML file can be processed by a large number of command line tools such as osmosis (Java application for reading/writing databases) (Wikipedia-Osmosis, 2015), osmium (multipurpose tool for data interoperability and time-series analysis) (Osmium-Tool, 2016), osm2pgsql (tool to convert XML data to PostGIS-enabled PostgreSQL databases) (Wikipedia-Osm2pgsql, 2015), osm2postgresql (to simplify rendering with QGIS and other GIS, web servers) (Wikipedia-Osm2pgsql, 2012), osm2pgsqlouting (PgRouting-Osm2pgouting, 2016) (to import data file into pgRouting databases) etcetera. However, because of being structured to work on latest data version of a given area for specific task these packages are not suitable for current processing. Instead, osmium based open-source osm-history-splitter packages are not suitable for current processing. Instead, osm-history-splitter packages are not suitable for current processing. Instead, osm-history-splitter (Corcoran and Zielstra, 2015) has also reported similar growth rate for both the number of nodes and Edges for Beijing, China because of restructuring of Odbl license and hyped OSM usage in various mapping applications (Zielstra et al., 2013). Although, the sudden increase in data generation activity is exponential, it is not equal for all the provinces as it is a function of the number of active mappers in the area (Zhao et al., 2015). A closer look at the graph (especially for provinces with high nodes density) demonstrates that the exponential curve itself is a partial exponential-step curve (a step-wise growing exponential curve). This is because time-span between September, 2007 and April, 2007 every year has observed lesser nodes edit through different mapping events as compared to between April, 2007 and September, 2007, because of low tourism and out-door activity in winter. However, this observation is merely visual.

The time intervals used for feature classification are as follows: between April 01, 2004 to April 01, 2007, between April 01, 2007 to September 01, 2007, between September 01, 2007 to April 01, 2008 and so on, till between April 01, 2015 to September 01, 2015; making 18 such intervals (April and September months are selected to divide a year into two halves, one portraying summer (tourism and out-door activity) and another winter (arm-chair mapping activity)).

It is not linearly possible to compare two primitive geometry features, i.e. point, edge, and polygon, when the goal behind it is to determine contribution effort (figure (1)). Creation of a point, a line, and a polygon does not involve similar effort from contributors side, thus making them mutually incomparable in this perspective; similar explanation is valid while comparing lines ((Zhao et al., 2015), (Corcoran et al., 2013), (Strano et al., 2015), (Haklay, 2009)) and polygons of varied length and area, respectively. Finally, to neutralize the geographic effect of different provincial area on nodes count for each province, it was divided by respective area.

Figure 2 shows the time-series evolution of nodes density for 81 provinces of Turkey constituting Points(all) between 2007-2015. It can be observed that data before April, 2007 is absent in the dump file as all graphs are merging towards x-axis as one traces back in time. It was an expected observation as the object history feature in OSM project in 2007 (OSM-API, 2015) was introduced by Editing API v0.5, meaning no history data before that. Between 2007 and 2012, the curves are following a gradual rise (figure (2)), which shows limited involvement by dormant mappers because of limited flexibility in editing by old OSM license (OSM-License, 2012). This was then observed by an exponential growth in 2012 ((Zhao et al., 2015) has also reported similar growth rate for both the number of nodes and Edges for Beijing, China) because of restructuring of Odbl license and hyped OSM usage in various mapping applications (Zielstra et al., 2013). Although, the sudden increase in data generation activity is exponential, it is not equal for all the provinces as it is a function of the number of active mappers in the area (Zhao et al., 2015). A closer look at the graph (especially for provinces with high nodes density) demonstrates that the exponential curve itself is a partial exponential-step curve (a step-wise growing exponential curve). This is because time-span between September, 2007 and April, 2007 every year has observed lesser nodes edit through different mapping events as compared to between April, 2007 and September, 2007, because of low tourism and out-door activity in winter. However, this observation is merely visual.

Another observation of interest is regarding nodes density value across the country. Figure (2) also shows nodes density map at 2009, 2012, and 2015 time-slices. At each profile it can be seen that the eastern and south-eastern part of the country is less densified as compared to the western and south-western ((Zhang et al., 2015) has also reported spatial distribution of China’s OSM road network in a similar fashion), with some outliers. Socio-economic factors can be attributed for this nodes density’s spatial spread. Literacy Level, Population Density, Tourism Activity, Internet Usage, and Human Development Index have also showed similar density pattern which was observed in figure (2). Similar pattern has been observed for line and polygon features as well. The road length density throughout the country is constant, depicting that road itself is not scarce rather OSM Edges features are non uniformly mapped by contributors ((Corcoran and Zielstra, 2015) has also reported spatial distribution of China’s OSM road network in a similar fashion), with some outliers. Socio-economic factors can be attributed for this nodes density’s spatial spread. Literacy Level, Population Density, Tourism Activity, Internet Usage, and Human Development Index have also showed similar density pattern which was observed in figure (2). Similar pattern has been observed for line and polygon features as well. The road length density throughout the country is constant, depicting that road itself is not scarce rather OSM Edges features are non uniformly mapped by contributors ((Corcoran and Zielstra, 2015) has also reported spatial distribution of China’s OSM road network in a similar fashion), with some outliers. Socio-economic factors can be attributed for this nodes density’s spatial spread. Literacy Level, Population Density, Tourism Activity, Internet Usage, and Human Development Index have also showed similar density pattern which was observed in figure (2).
Mooney, 2013), (Zhao et al., 2015), and (Corcoran et al., 2013) have also studied OSM road network evolution. Some provinces in the eastern and south-eastern part have shown nodes density spikes, thus acting as an outlier because of having active Senior mappers there. One such region is Batman province, red box in figure (2), which shows high nodes frequency, especially in 2015, because of a mapper (student of Batman University) who is responsible for approximately 4 and 3% of whole Turkey’s OSM contribution for Points and Edges, respectively. Since this mapper is currently a university student, no spikes are present for this province for earlier years.

4. CONCLUSIONS

This study presents an spatial evolution analysis of Turkey-OSM dataset in an eight year time span (2007-2015). It has been observed that the dump file does not have any history information before 2007 because of the absence of object history feature in Editing API v0.4 or earlier. In figure (2) the curves are horizontal between 2007 to 2012, which shows a period of immobility in contribution activity, however, there is an exponential rise after year 2012 because of the change in OSM license from Creative Commons Attribution-ShareAlike 2.0 to Odbl. After 2012, the curves are following a partial exponential-step function because of less contribution activities in winter season. The spatial analysis has revealed that there is a spatial biasedness from west to east of the country towards the evolution of dataset at any given point in time (figure (2)), with some exceptional provinces. Provinces along the Mediterranean sea (western and south-western part) have experienced more nodes density at selected time-slices (2009, 2012, and 2015) as compared to eastern and south-eastern part of the country which were always under-developed. This pattern in nodes density is believed to be the consequence of socio-economic factors, i.e. Literacy Level, Population Density, Tourism Activity, Internet Usage, and Human Development Index, of the region. Similar observations have been found for edges and polygons feature as well.

ACKNOWLEDGEMENTS

The authors would like to thank The Scientific and Technological Research Council of Turkey (TÜBİTAK, URL: https://www.tubitak.gov.tr/en) for supporting financially under the Graduate Scholarship Programme for International Students (2215 Program).

REFERENCES


Kuhn, W., 2007. Volunteered geographic information and giscience. NCGIA and Vespucci Workshop on Volunteered Geographic Information; Santa Barbara.


Kuhn, W., 2007. Volunteered geographic information and giscience. NCGIA and Vespucci Workshop on Volunteered Geographic Information; Santa Barbara.


