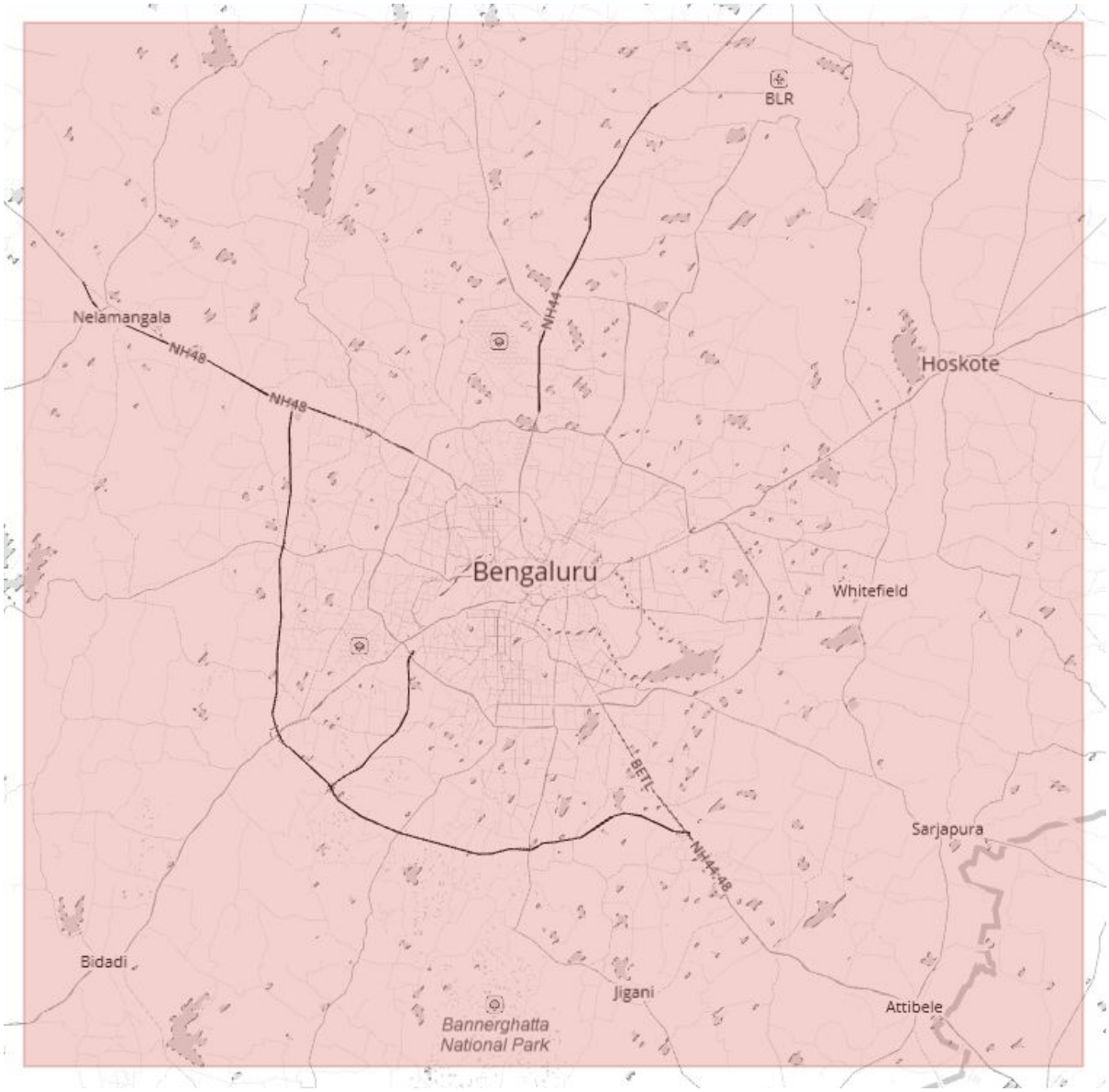
1. Introduction

OpenStreetMap is a collaborative project to create a free editable map of the world.[1] The aim of Open Street Maps project is to make maps freely available, by crowdsourcing the data population. This indirectly means that the data is not verified by an entity having commercial interest for validating the map data. Thus we end up with map data which can be inconsistent/incomplete. In countries like India, which does not have a large population of technology savvy individuals, the open data collection is generally low. This point is further validated in the project.

Note:Open Street Maps has been abbreviated as OSM in the report.

1.1 City region for this report

For this report the data was obtained from Mapgen metro extracts for Bengaluru. The compressed size of the extract is 37MB and the uncompressed size is 608MB.



1.2 Datastore for analysis

MongoDB has been used as the datastore for the preprocessed OSM data.

2. Cleaning the Data

The data in OSM has been added my several individuals who have in interest of making OSM maps better. Since many people are uploading data and there is no strict enforcement of a common data model, we can find many inconsistencies in the data model used to store data in OSM. This section discusses the data cleaning and data munging that was done before analysis can be done with the data.

2.1 City names in *addr:city* tag

City names had many distinct values, some of which are shown below. In *2014* the government official changed the name of the city from *Bangalore* to *Bengaluru*. Some entries in OSM are still referring to the city as *Bangalore* The many variations of bengaluru/bangalore are in bold below. Also note that some city tags had street address in them.

Abbigere, Abbigere Village, **Banglore**, Begur, **Benagluru, Bengaluru, Bengauru,** Bidadi, **Bngalore**, Budigere, Devanahalli Taluk, Kaggalahalli, Chandapura, Cheemasandra, Chelkere, Devara Bisnhalli, Bellandur, Devarachikkanahalli, **Benagalur**, Indiranagar, HRBR **bangaluru, banglore, banngalore,** btm layout, yelahanka

These could be simple typos, but it makes it difficult for performing search on the OSM database by city. These entries were cleaned by changing all the city names to Bengaluru and storing the other entries information in the tag as street address.

2.2 Phone numbers in *phone* tag

This is one field which is expected to have a high level of variation. People represent phone numbers in many forms. To keep the data consistent the following two formats were the final versions stored in the database.

+91-XXXXXXXXXX **10 digit mobile numbers**
+91-080-XXXXXXXX **8 digit landline/fixed-line numbers with city STD code *080***

The following phone number formats were converted to the above format.

Mobile number formats	Landline number formats
XXXXX XXXXX +91XXXXX XXXXX +91 XXXXX XXXXX +91XXXXX XXXXX	XXXXXXXX 080 XXXXXXXX +91 80 XXXXXXXX +91 80 XXXX XXXX

2.3 Postcode in *addr:postcode* tag

Postcode tags which have the postal code/zip code of the address of a location has 3 variations. The standard postcode in India has 6 digits and Bengaluru region has the postcode in the series 56XXXX[2] , with some regions in 57XXXX. The variations of the postcode are as follows:

Postcode Format	Comments
XXXXXX	No Issues
XXX XXX	Postcode split into 2 blocks of 3 digits. Indexing/ searching algorithms cannot understand the split postcode.
XX	Shorthand notation used for bangalore. The missing prefix for the postcode is 5600

---

3. Data Overview and Exploration

Basic details about the data are as follows.

Number of Documents:

```
>>>db.OSM.find().count()
3492364
```

Number of Nodes:

```
>>>db.OSM.find({'type' : 'node'}).count()
2839193
```

Number of Ways:

```
>>>db.OSM.find({'type' : 'way'}).count()
652203
```

Number of Relations:

```
>>>db.OSM.find({'type' : 'relation'}).count()
22
```

Number of users contributing to OSM data:

```
>>>len(db.OSM.distinct('created.user'))
1501
```

3.1 User contribution statistics:

```
Mean no. contribution by a user   : 2326.69
Median of contributions by a user : 7
```

```
db.OSM3.aggregate([{'$match' : {'created.user': {'$exists' : 1}}}, {'$group' : {'_id' : '$created.user', 'count' : {'$sum' : 1}}}, {'$sort' : {'count' : -1}}])
```

The above statistics show that the the data is positively skewed, with approximately half the users contributing 7 or less points of interest on OSM.

Note: The calculation was done by loading data into pandas.

3.2 Popular postcodes

```
>>> db.OSM.aggregate([{'$match' : {'addr.postcode': {'$exists':1}}},
                      {'$group' : {'_id' : '$addr.postcode', 'count' : {'$sum' : 1}}},
                      {'$sort' :{'count':-1}}, {'$limit' : 5}])
```

{u'_id': **u'560066'**, u'count': 105},
{u'_id': **u'560040'**, u'count': 97},
{u'_id': **u'560038'**, u'count': 85},
{u'_id': **u'560095'**, u'count': 62},
{u'_id': **u'560001'**, u'count': 60}

Postcode 560066 refers to the area of Whitefield, which is a technology hub and thus residents and people visiting this area would be tech savvy and would have familiarity with Open Street Maps.

3.3 Documents with names

In our complete OSM collection which has 3492364 documents, 27123 had name tags. It's good to see that off these 27123 documents which have name tags, 8608 have names in the regional language Kannada.

```
# Documents with name tag
>>> db.OSM.find({'name':{'$exists' : 1}}).count()
    27123
# Documents with name:kn tag for names Kannada
>>> db.OSM.find({'name.kn':{'$exists' : 1}}).count()
    8608
```

3.4 Amenities

Popular amenities

The top 5 amenities which are marked on OSM for Bengaluru are:

```
# Filter documents which have amenity tag, group them by amenity, sort them by count
>>> db.OSM.aggregate([{'$match' : {'amenity': {'$exists':1}}},
                      {'$group' : {'_id' : '$amenity', 'count' : {'$sum' : 1}}},
                      {'$sort' :{'count':-1}},
                      {'$limit' : 5}])]
[{u'_id': u'restaurant', u'count': 1248},
 {u'_id': u'place_of_worship', u'count': 913},
 {u'_id': u'atm', u'count': 666},
 {u'_id': u'bank', u'count': 651},
```

```
{u'_id': u'school', u'count': 623}]
```

The earliest entries for restaurant are for

```
        # Filter to get documents of restaurants with names, sort by created.timestamp
>>> db.OSM.aggregate([{'$match' : {'amenity' : 'restaurant', 'name' :
{'$exists' : 1}}},
 {'$sort' : {'created.timestamp' : 1}},
{'$limit' : 5},
{'$project' : {'name' : 1, 'created.timestamp' : 1}}])
[{u'_id': ObjectId('57bc6f97b54510d8a4fe14e1'),
  u'created': {u'timestamp': u'2008-05-28T11:10:58Z'},
  u'name': u'Kusum Delux Seafood'},
 {u'_id': ObjectId('57bc6f97b54510d8a4fe491f'),
  u'created': {u'timestamp': u'2008-12-07T04:14:59Z'},
  u'name': u'Kadamba Veg'},
 {u'_id': ObjectId('57bc6f98b54510d8a4fe534f'),
  u'created': {u'timestamp': u'2009-03-09T08:38:21Z'},
  u'name': u'Tashan'},
 {u'_id': ObjectId('57bc6f98b54510d8a4fe5a1d'),
  u'created': {u'timestamp': u'2009-05-07T12:13:44Z'},
  u'name': u'Hotel Garden'},
 {u'_id': ObjectId('57bc6f98b54510d8a4fe66c0'),
  u'created': {u'timestamp': u'2009-06-27T08:09:47Z'},
  u'name': u'Anand Sagar Punjabi Dhaba'}]
```

---

4 Missing Data

4.1 Schools in Bengaluru

According to the government survey for the year 2013 we have the information that Bangalore has 1613 private schools and 1207 government schools.[3] The OSM data shows 623 entries having *amenity* as *school.* Even by ignoring government schools, OSM has records only for approximately 39% of the schools.

```
>>> db.OSM3.find({'amenity' : 'school'}).count()
    623
```

4.2 Post Offices in Bengaluru

According to the India Post, the government postal service company, there are 201 post offices in bengaluru city.[4] OSM has record for 101 off these post offices.

```
>>> db.OSM3.find({'amenity' : 'post_office'}).count()
    101
```

5. Ideas to improve the dataset

In India, street addresses are quite complex, in comparison to US for example. The street address are generally represented in 2 or more lines. It would be interesting to evaluate the use of ML/AI techniques for splitting of street names into a form the postal department accepts.

Benefit: The address would look professionally maintained, can be used for sending physical mails.

Anticipated Problems: While it would be great if this process could be done automatically, Indian address have their own share of complexity that makes it difficult to do this automatically.

An AI/ML agent can be used to learn the names of Indian banks, to make the data about bank and ATMs consistent. For example, there are many variations for storing the name of the ATM operator. Some ATM entries just have the information in the name name of the ATM, while others have explicit entries for tags like 'operator' / 'fee'. While building such a bot, it's important to validate that the ML/AI agent is working correctly for all inputs.

Benefits: It would be easier for users to search for ATM/Bank locations for their own banks.

Anticipated Problems: There are many types of banks in India for which we might not get authentic data to automate the process of classifying the names of the banks. For example there are many co-operative banks which also have ATMs but the information for which is difficult to gather

Though the data is incomplete, OSM data can help us in conducting some basic analysis relevant to urban planning. We can find the distance between schools and the closest public transport transit. This should be low for enabling easy accessibility for schools. Bengaluru has many business clusters, having identical/related business shops all in the same road/locality. Studying the trend of historic map data can help us discover such new/emerging business clusters. While it this data can be used for exploration, since its not verified data, it can not be used for government work like policy decisions.

Benefits: OSM data can be very useful for policy designers, which can also be evaluated by the general public, and improve transparency in public policy decisions

Anticipated Problems: As seen in the report above, OSM data is lacking in many areas. Its difficult

To use OSM data in its current form for policy decisions. The current data should be used to push government officials to consider wider adoption for

a mapping platform which does not charge anything for the service.

References

[1] https://en.wikipedia.org/wiki/OpenStreetMap
[2] https://en.wikipedia.org/wiki/Postal_Index_Number
[3] https://data.gov.in/catalog/town-amenities-census-2011
[4] https://www.indiapost.gov.in/vas/pages/LocatePostoffices.aspx