Developing big data analytics for socioeconomic and land use characterization

Huina Mao, Liane Russell Fellow
maoh@ornl.gov
Computer Science and Engineering Division
Oak Ridge National Laboratory

**INTRODUCTION**

- Measuring socioeconomic conditions at a fine level of spatial granularity and in a timely manner has become a priority task for governments for the reasons:
  - Identifying the poor and targeting interventions to reduce poverty.
  - Evaluating the effectiveness of policy interventions taken towards the poor.
- Traditional measurement based on surveys: expensive; time-consuming; low frequency.
- The socioeconomic statistics of low-income developing countries is very scarce due to the high cost of data collection.
- Existing work uses satellite imagery to identify the informal (e.g. slums) and formal settlements based on building and street characteristics. However, physical features are not enough to capture socioeconomic conditions.
- **Goal of this project:** study the relation between social/behavioral features derived from mobile phone communication data (i.e. Call Detail Records) and socioeconomic conditions. In addition, we study the temporal spatial patterns of call activities and land uses.

**DATA SOURCES**

**Call Detail Records (CDR)**
- Senegal: 01/01/2013 to 12/31/2013 hourly CDR error 1,666 cell phone towers
- Data Format: from site_ID to site_ID, Rcalls, duration
- **Call flow from tower to tower** (see Fig. a)
- Construct a social communication network (nodes: towers; edges: call links between two towers; edge weight: call volume)

**Socioeconomic data**
- **Demographic and Health Survey (DHS) Wealth index**
- Wealth index: 1 (poorest) to 5 (richest)
- 7,902 households are surveyed; 391 clusters formed in Senegal.
- The distribution of 391 clusters is shown in Fig. b; the histogram of DHS Wealth Index is in Fig. c. 88% of clusters with wealth index <= 4.

**FEATURE EXTRACTION**

**Call Activity Features**
- Total number of calls
- Number of incoming, outgoing, self-connecting calls
- Introspection rate (i.e. self-calls/total calls)
- Outgoing call ratio (i.e. outgoing calls/incoming calls)

**Network Features**
- Centrality (i.e. PageRank)
- In-degree
- Out-degree
- Social diversity: entropy of communication links

**SOCIOECONOMIC PREDICTION**

**LINEAR REGRESSION**
- Dataset: 70% (241 obs) for training dataset; 30% (103 obs) for testing dataset
- Model: Ordinary Least Square (OLS) Linear Regression
- Outputs: call activity and network features derived from CDR.
- Output: 1-5 at continuous scale.
- Results: Table 1 and Fig. g, Regression error metrics: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE)
- Pearson correlation between true and predicted values is: 0.77.

**CLASSIFICATION**
- **Convert continuous DHS to 3 classes:** rich, middle, poor.
- **Dataset:** 70% from each class for training; 30% from each class for testing.
- Model: Support Vector Machines
- Output: 3 classes (rich, middle, poor).
- Results: Table 2. The overall classification accuracy is 73%. The precision for the rich class achieves 100%.

**Table 1. Regression error metrics.**

<table>
<thead>
<tr>
<th>Features</th>
<th>MAE</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network features</td>
<td>0.75</td>
<td>0.87</td>
</tr>
<tr>
<td>Activity features</td>
<td>0.01</td>
<td>0.75</td>
</tr>
<tr>
<td>Network + Activity features</td>
<td>0.50</td>
<td>0.74</td>
</tr>
</tbody>
</table>

**Table 2. Classification accuracy report.**

<table>
<thead>
<tr>
<th>Classes</th>
<th>precision</th>
<th>recall</th>
<th>f1-score</th>
<th>num_observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>poor</td>
<td>0.69</td>
<td>0.74</td>
<td>0.71</td>
<td>42</td>
</tr>
<tr>
<td>middle</td>
<td>0.71</td>
<td>0.71</td>
<td>0.71</td>
<td>49</td>
</tr>
<tr>
<td>rich</td>
<td>1.00</td>
<td>0.77</td>
<td>0.87</td>
<td>13</td>
</tr>
<tr>
<td>avg / total</td>
<td>0.74</td>
<td>0.73</td>
<td>0.73</td>
<td>104</td>
</tr>
</tbody>
</table>

**CONCLUSION AND FUTURE WORK**

- We extracted call activity and social network features from mobile phone communication data, and have found the strong predictive power to the socioeconomic indicator obtained from surveys (i.e. DHS Wealth index).
- **Our method can measure socioeconomic conditions at a high resolution, 1km by 1km.**
- Monitoring socioeconomic conditions in a timely manner is important for identifying the poor, targeting interventions to reduce poverty and income inequality.
- Mobile phone call activities may also reveal land uses, and may track land use changes over time, which complements traditional urban analysis methods. Future work includes to evaluate the land use results.