Geospatial Big Data Applications

Peter Goetsch
peter.goetsch@cs.helsinki.fi
27 February 2017
Outline

1. What is geospatial big data
2. Why it is important
3. Case study: geospatial big data in transportation modeling
4. Conclusions
Geospatial Data

- “Geospatial data describe objects and things with relation to geographic space, often with location coordinates in a spatial referencing system” [1]
- Origins in Geographic Information Systems (GIS)
Geospatial Big Data

“Geospatial big data refers to spatial data sets exceeding capacity of current computing systems” [2]
Source: Lee et al. [2015]
Why is Geospatial Big Data Important?

- Example: Millennium Project
- Large potential to solve today’s problems
- Increasing exponentially in size

Source: Lee et al. [2015]
Use Case: Transportation
“Can we use geospatial big data to model travel patterns?”
Data Sources

• Call Detail Records: \(<\text{hash(id)}, \text{timestamp}, ~\text{GPS location}>\) of each call and some SMS

• Census Data: population and vehicular usage data

• Road Networks: GIS shapefiles describe roads (speed limit, lanes, classifications etc)
  • OpenStreetMap.org
System Constraints

1. Handle different regions of earth with different data quality and availability
2. Modular \(\rightarrow\) easy to update algorithms and other building blocks
3. Communicate results via online visualizations
Standardizing Data

- Utilize Google Protocol Buffers
- Convert raw data to standardized format
- Write new parser for new data sources
- Compact data storage

See https://developers.google.com/protocol-buffers/
Architecture

- Postgres extension PostGIS (relational DB)
- Origin-Destination Matrix Computation
- Route computation
- Visualization

Source: Colak et al. [2015]
Algorithms (1/2)

• Determine “stay” points – places where a user stays
• Map stay points to <home, work, other>
• Look at sequences of stay points to identify trips
  • Home ↔ work, home ↔ other, <work, other> ↔ <work, other>
• Estimate departure times

• Trip Purpose + Departure Time + Trip Start/End Points = Full Trip
Why it works

• “Recent work has found that individuals are predictable, unique, and slow to explore new places (González et al., 2008; Brockmann et al., 2006; de Montjoye et al., 2013; Song et al., 2010a,b; Candia et al., 2008; Calabrese et al., 2013). The availability of similar data nearly anywhere in the world has facilitated comparative studies that show many of these properties hold across the globe despite differences in culture, socioeconomic variables, and geography.” [3] (emphasis added)
Algorithms (2/2)

• Predict route
• Predict traffic demand
• Predict travel time

• Borrows algorithms from Traffic Engineering
Results

• Method works reasonably well

• Not thoroughly tested in many geographic regions

• Promising
Results (1/2)

Table 2
Trip tables estimates. Where possible, our results are compared to estimates made using travel surveys. For each city, we report the number of person trips in millions for a given purpose or time. Trip purposes include: home-based word (HBW), home-based other (HBO), and non-home-based (NHB). Trip periods include: 7–10 am (AM), 10 am–4 pm (MD), 4–7 pm (PM), and the rest of the day (RD). We note that the exact boundaries of the surveys do not exactly coincide with those used in our estimation so direct comparisons are not exact. In general, trip magnitudes align closely, with the exception of Rio de Janeiro, where the survey results report far too few trips, illustrating the difficulty of obtaining sensible measurements via certain techniques. No comparisons could be found for Porto.

<table>
<thead>
<tr>
<th>City</th>
<th>HBW</th>
<th>HBO</th>
<th>NHB</th>
<th>AM</th>
<th>MD</th>
<th>PM</th>
<th>RD</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boston</td>
<td>5.76</td>
<td>8.99</td>
<td>6.72</td>
<td>3.71</td>
<td>7.68</td>
<td>5.75</td>
<td>4.33</td>
<td>21.47</td>
</tr>
<tr>
<td>MHTS</td>
<td>3.22</td>
<td>12.83</td>
<td>9.49</td>
<td>5.32</td>
<td>8.87</td>
<td>8.20</td>
<td>3.15</td>
<td>25.54</td>
</tr>
<tr>
<td>SF Bay</td>
<td>4.07</td>
<td>10.05</td>
<td>7.04</td>
<td>4.47</td>
<td>7.81</td>
<td>5.35</td>
<td>3.53</td>
<td>21.16</td>
</tr>
<tr>
<td>BATS</td>
<td>4.60</td>
<td>11.54</td>
<td>4.66</td>
<td>4.18</td>
<td>6.90</td>
<td>4.22</td>
<td>3.00</td>
<td>20.80</td>
</tr>
<tr>
<td>Survey</td>
<td>2.06</td>
<td>–</td>
<td>–</td>
<td>1.31</td>
<td>1.19</td>
<td>1.24</td>
<td>–</td>
<td>3.74</td>
</tr>
<tr>
<td>Lisbon</td>
<td>1.08</td>
<td>2.01</td>
<td>1.21</td>
<td>0.79</td>
<td>1.67</td>
<td>1.26</td>
<td>0.58</td>
<td>4.30</td>
</tr>
<tr>
<td>Survey</td>
<td>0.61</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Porto</td>
<td>0.49</td>
<td>0.87</td>
<td>0.46</td>
<td>0.32</td>
<td>0.70</td>
<td>0.54</td>
<td>0.27</td>
<td>1.83</td>
</tr>
<tr>
<td>Survey</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

*Note that the Lisbon Survey only contains estimates of vehicle trips in millions.*

Source: Colak et al. [2015]
Results (2/2)

Fig. 3. Correlations between OD matrices produced by our system and those derived from travel surveys at the largest spatial aggregation of the two models. In Boston, this is town-to-town, in San Francisco, MTC superdistrict-to-super district, in Rio, census superdistrict-to-superdistrict, and in Lisbon, Freguesia-to-Freguesia. The larger of these area units (e.g. towns in Boston), the better our correlations, while correlations at the smallest aggregates (e.g. Freguesias in Portugal), correlations are lower. However, more work must be done to understand uncertainties in estimates provided by both models.

Source: Colak et al. [2015]
Visualization

• Simple query interface returning GeoJSON (http://geojson.org) documents
  • RFC 7946
    • Geographic data structures representation
• D3.js (https://d3js.org)
• Online mapping service, eg Google Maps
Visualization Example

Source: Colak et al. [2015]
Challenges

- Data cleaning and filtering
- Handling multi-modal trips / incorporating public transit
- Triangulation Accuracy → only uses SMS and call records
- Obvious time vs speed tradeoffs
Why is this Important?

• Engineering significance
• Near real-time traffic and infrastructure usage estimates
• No waiting 5-10 years for a survey
• Saves cities/countries time, resources and money
Conclusion

• Pluggable architecture allow for easy updating → update algorithms as they get better
• Some important details left out, eg computation methods
• Only going to become more accurate
• A good “first-pass”, but still work to be done
• Illustrates cross-disciplinary reach of Big Data
• One of many applications of Big Data
Sources - Papers

1. Geospatial Big Data Handling Theory and Methods: A Review and Research Challenges, Li et al. [2016]
2. Geospatial Big Data: Challenges and Opportunities, Lee et al. [2015]
3. The path most traveled: Travel demand estimation using big data resources, Colak et al. [2015]
Sources - Photos

3. https://www.tesla.com/sites/default/files/images/support/model-s-red@2x-png8.png
4. https://www.hsl.fi/sites/default/files/campaign-images/lij_matkustajat_3_white_0.jpg
5. http://verkkokauppa.hsl.fi/media/catalog/product/cache/1/image/9df78eab33525d08d6e5fb8d27136e95/m/a/matkakortti_1_4.jpg
7. https://www.openstreetmap.org/assets/osm_logo-bfddf8749f46ea7c92950a82eeb425045cddfc0725e8281799d0efddee45c425d.svg