Evaluating Floating Car Data Quality for Knowledge Discovery

Vítor Cerqueira$^1$, Luís Moreira-Matias$^2$, and Jihed Khiari$^2$

U.Porto, Portugal$^1$

NEC Laboratories Europe, Germany$^2$

September, 2017
Outline

1. Introduction
   - Motivation
   - Objectives

2. Methodology
   - Statistical indicator set

3. Experimental Study
   - Case studies
   - Results

4. Conclusions
Outline

1 Introduction
   • Motivation
   • Objectives

2 Methodology
   • Statistical indicator set

3 Experimental Study
   • Case studies
   • Results

4 Conclusions
Motivation

Increasing availability of GPS-enabled vehicles and Floating Car Data (FCD) sources – Most fleets will be published as open data

Data quality issues

- GPS devices with different protocols (3G/4G);
- Road/urban topology (e.g. narrow streets, high buildings)

Costs of using unreliable FCD

- Suboptimal traffic control actions;
- Excessive storage usage;
- Excessive computational power usage.
Main goal

Develop an automated analysis of FCD sources
- Efficiently discard unreliable ones

Approach

- Set of statistical indicators
  - parametrizable, yet interpretable and with standard output
- Spatio-temporal coverage, and veracity of FCD
Main goal

Develop an automated analysis of FCD sources
- Efficiently discard unreliable ones

Approach

- Set of statistical indicators
  - parametrizable, yet interpretable and with standard output
- Spatio-temporal coverage, and veracity of FCD

Current state of the art

- Mostly about GPS accuracy analysis
- Used as pre-processing step, without a standard and comparable output.
Outline

1. Introduction
   - Motivation
   - Objectives

2. Methodology
   - Statistical indicator set

3. Experimental Study
   - Case studies
   - Results

4. Conclusions
Granularity - frequency of transmission of GPS traces

- Relevant for tracking vehicles, and tasks such as map-matching and congestion prediction
Macro temporal coverage

- Relevant for demand forecasting tasks
Micro temporal coverage

Also relevant for demand forecasting tasks, specifically in-day seasonality: Rush hours, demand peaks due to special events (e.g. soccer match)
Some parts of a town are more relevant than others
- Road density;
Spatial coverage

- Some parts of a town are more relevant than others
  - Road density;
  - Proximity to landmarks;
Spatial coverage

- Some parts of a town are more relevant than others
  - Road density;
  - Proximity to landmarks;
  - Neighborhood
Missing data

- Large time intervals without GPS transmission
  - How many gaps exist per trip, on average;
  - How are they typically spread across a trip
Reliability

- Analysis of objectivity of the dataset;
  - Reachability;
  - Activity
Accuract

- Measuring the accuracy of GPS devices
Outline

1 Introduction
   - Motivation
   - Objectives

2 Methodology
   - Statistical indicator set

3 Experimental Study
   - Case studies
   - Results

4 Conclusions
● **Timestamp, Vehicle Id, Trip Id, Latitude, and Longitude**

**Table:** Nanjing & San Francisco taxi fleets

<table>
<thead>
<tr>
<th></th>
<th>Nanjing</th>
<th>San Francisco</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. GPS traces</td>
<td>18 million</td>
<td>11 million</td>
</tr>
<tr>
<td>No. Trips</td>
<td>432.899</td>
<td>959.025</td>
</tr>
<tr>
<td>No. vehicles</td>
<td>7648</td>
<td>536</td>
</tr>
<tr>
<td>Timespan</td>
<td>1 day</td>
<td>23 days</td>
</tr>
<tr>
<td>Median Trip Duration</td>
<td>10 min</td>
<td>7.5 min</td>
</tr>
</tbody>
</table>

● **Data mining tasks:**
  ▶ Travel time prediction;
  ▶ Origin-destination flow estimation
**Table:** Results from the Data Mining Experiment I

<table>
<thead>
<tr>
<th>Error</th>
<th>Nanjing</th>
<th>San Francisco</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>1619.34</td>
<td>624.96</td>
</tr>
<tr>
<td>MAE</td>
<td>700.7</td>
<td>455.2</td>
</tr>
</tbody>
</table>

**Figure:** MSE bias-variance decomposition for DM-II
Results

- **Macro Temporal Coverage** shows the greatest discrepancy:
  - Driven by the timespans of the datasets
  - Suggests possible generalization issues in Nanjing, corroborated by the results of the data mining tasks.

Figure: Methodology results for Nanjing and San Francisco
Results

- **Spatial Coverage** difference is also noticeable:
  - Nanjing’s cover more ground in less time, but with a much larger fleet;
  - San Francisco’s value suggests that the fleet operates only in a small part of the town.

*Figure:* Methodology results for Nanjing and San Francisco
Figure: Methodology results for Nanjing and San Francisco

Results

- **Missing data** is comparable:
  - Nanjing shows a better granularity;
  - But it also has an higher no. of missing observations per trip, on average.
Results

- **Accuracy** is higher for San Francisco:
  - Average error of 40m and 15m, for Nanjing and San Francisco;
Outline

1. Introduction
   - Motivation
   - Objectives

2. Methodology
   - Statistical indicator set

3. Experimental Study
   - Case studies
   - Results

4. Conclusions
Results

- Link between the results of the methodology and the results of the data mining tasks
  - a way to reason about the differences in performance of the same model under different scenarios
- an automated approach without conducting any manual data mining procedure

Applicability

- Nanjing ×
- San Francisco ✓
Potential deployment

Market value

- Emerging data market
  - A standardized evaluation framework is helpful to define the market value of FCD feeds
Potential deployment

Market value

- Emerging data market
  - A standardized evaluation framework is helpful to define the market value of FCD feeds

Troubleshooting

- Reliability of FCD
  - Detecting malfunctions on the sensors
Potential deployment

Market value
- Emerging data market
  - A standardized evaluation framework is helpful to define the market value of FCD feeds

Troubleshooting
- Reliability of FCD
  - Detecting malfunctions on the sensors

Saving communicational broadband
- Reducing the storage requirements of typical TMC visualization systems
  - Periodically analysis of the data broadcasted by the input sources
- Reducing computational demand
  - for short-term inference of the traffic status using data mining
Thank you.