Accurate Pollutant Modeling and Mapping: Applying Machine Learning to Participatory Sensing and Urban Topology Data

Axel Schulz\(^1\), Jakob Karolus\(^1\), Frederik Janssen\(^2\), **Immanuel Schweizer\(^1\)**

\(^1\)Telecooperation Lab \(^2\)Knowledge Engineering Lab

Technische Universität Darmstadt
Mobile Sensing
Participatory Sensing
Lärmausweisung Hessen 2007
Stadtbereich Frankfurt am Main
Hauptverkehrsstraßen und sonstige Straßen

Stufenpegel L_eq in dB(A)

0 - 79 dB(A)
30 - 59 dB(A)
30 - 59 dB(A)
50 - 79 dB(A)
50 - 79 dB(A)
70 - 90 dB(A)
70 - 90 dB(A)

Schallschutzeinrichtungen

Maßstäbe: 1:50,000

Datum: 05.06.2007

Datenquelle: Stadt Frankfurt am Main, Verwaltung der Stadt Frankfurt am Main
Model sparse environmental data
Related Work

• Dispersion Models
  • Government maps
  • Impact of industrial pollution

• (Land-use) Regression
  • Environmental science
  • Mobile Sensor Data [1]

Why classification?

• Why regression?
  • Labels vs. numbers

• Interpretability

• Proof-of-concept

IEEE Netsys 2015
System architecture

Sensing and Preprocessing

- Noisemap
- Application
- Calibration
Sensing and Preprocessing

• Noisemap
  • Noise pollution

• More than 3,500 installations
  • ~500 active installations
  • ~30 active users / day
  • ~1,000,000 data points

• Calibration [1]
• Incentive systems [2]

Noise Data

![Bar chart showing noise data by sound level in dB.](chart.png)
Sensor Data

• Weather Data
  • Crawls from DWD

• Street Data
  • OpenStreetMap

• Building Data
  • LinkedGeoData
Building / Street Data

• # of buildings of a certain type (e.g., PlaceOfWorship)

• Distance to closest building of a certain type

• Closest street
Feature Generation

Initial Dataset
- Noisemap
  - Instances of noise data
    - Attributes:
      - Timestamp
      - Sound value in dB
      - GPS coordinates
      - ID of sensor

Feature Generation
- OpenStreetMap
  - Data File
    - Extracting OSM information on nearby streets
      - Attributes:
        - Type of street
        - Number of Lanes
        - Surface
        - Max. Speed

- LinkedGeoData
  - Object Data (RDF) SPARQL
    - Extracting information on nearby buildings
      - Attributes:
        - Numbers for specific types
        - Distance to nearest of a type

- WeatherData
  - Data File
    - Extracting weather information in the area
      - Attributes:
        - Temperature
        - Precipitation
        - Solidity ratio
        - Humidity
        - Wind velocity
        - Sunshine duration
        - Barometric pressure

- Additional Data
  - Adding additional information
    - Attributes:
      - Time
      - Calendar week

Classification
- Classification
  - Selected attributes
  - Attribute values
Classifier

• WEKA

• Decision Tree

• C4.5 (J48)
  • Missing values
  • Numeric values
  • Pruning
  • Interpretability
Results

<table>
<thead>
<tr>
<th>Feature</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sound Level and Time (B)</td>
<td>48.9%</td>
</tr>
<tr>
<td>B + Calendar Week</td>
<td>59.8%</td>
</tr>
<tr>
<td>B + SensorId</td>
<td>64.9%</td>
</tr>
<tr>
<td>B + Weather</td>
<td>70.5%</td>
</tr>
<tr>
<td>B + Building Type Distance</td>
<td>72.2%</td>
</tr>
<tr>
<td>B + Building Types Count</td>
<td>72.2%</td>
</tr>
<tr>
<td>B + Nearby Streets</td>
<td>72.8%</td>
</tr>
<tr>
<td>All Features Combined</td>
<td>80.9%</td>
</tr>
</tbody>
</table>
Confusion Matrix

<table>
<thead>
<tr>
<th>Is (in dB)</th>
<th>&lt;40</th>
<th>40-50</th>
<th>50-60</th>
<th>60-70</th>
<th>70-80</th>
<th>&gt;80</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;40</td>
<td>82.4</td>
<td>11.1</td>
<td>2.7</td>
<td>2.6</td>
<td>1.0</td>
<td>0.2</td>
</tr>
<tr>
<td>40-50</td>
<td>16.2</td>
<td>52.0</td>
<td>21.8</td>
<td>5.3</td>
<td>2.5</td>
<td>2.2</td>
</tr>
<tr>
<td>50-60</td>
<td>1.7</td>
<td>6.6</td>
<td>65.0</td>
<td>23.8</td>
<td>1.7</td>
<td>1.3</td>
</tr>
<tr>
<td>60-70</td>
<td>0.3</td>
<td>0.5</td>
<td>5.5</td>
<td>85.3</td>
<td>8.2</td>
<td>0.3</td>
</tr>
<tr>
<td>70-80</td>
<td>0.2</td>
<td>0.5</td>
<td>0.7</td>
<td>13.8</td>
<td>82.6</td>
<td>2.2</td>
</tr>
<tr>
<td>&gt;80</td>
<td>0.2</td>
<td>1.0</td>
<td>0.7</td>
<td>1.7</td>
<td>11.6</td>
<td>84.7</td>
</tr>
</tbody>
</table>
Visualization

MINI: Mashup for Identifying Noisy Infrastructure

<table>
<thead>
<tr>
<th>Label</th>
<th>DirectType</th>
<th>ID</th>
<th>Distance</th>
<th>Lat</th>
<th>Lon</th>
</tr>
</thead>
<tbody>
<tr>
<td>64</td>
<td>Clothes</td>
<td>674042174</td>
<td>0.1408760970350991</td>
<td>43.6099</td>
<td>3.87598</td>
</tr>
<tr>
<td>7AC Vodka Bar</td>
<td>Bar</td>
<td>724181321</td>
<td>0.02600641515436863</td>
<td>43.6114</td>
<td>3.87607</td>
</tr>
<tr>
<td>Acteur Sud</td>
<td>EstateAgent</td>
<td>1624095366</td>
<td>0.9708825882562285</td>
<td>43.6197</td>
<td>3.87322</td>
</tr>
<tr>
<td>Affat Voyages - Avant Départ Voyages</td>
<td>TravelAgency</td>
<td>724181332</td>
<td>0.02085876864007248</td>
<td>43.6114</td>
<td>3.87599</td>
</tr>
<tr>
<td>Agnès B</td>
<td>Clothes</td>
<td>712142809</td>
<td>0.137105877083844</td>
<td>43.6109</td>
<td>3.87432</td>
</tr>
<tr>
<td>Agnès Soronelles</td>
<td>Hairdress Shop</td>
<td>674061440</td>
<td>0.8815341739664681</td>
<td>43.6119</td>
<td>3.86506</td>
</tr>
<tr>
<td>Alimentation des Arceaux</td>
<td>Grocery</td>
<td>674061432</td>
<td>0.806286644494976</td>
<td>43.6119</td>
<td>3.86599</td>
</tr>
<tr>
<td>Animates Charles II Berards</td>
<td>Furniture</td>
<td>712158725</td>
<td>0.0870463757512446</td>
<td>43.6114</td>
<td>3.87491</td>
</tr>
</tbody>
</table>
Conclusion

• Heaps of sparse environmental data

• Models
  • Classification
  • Regression

• Urban Topology Model is critical

• Transferability
Questions?