Statistical Inference on Mobile Phone Network Data

European Forum for Geography and Statistics (EFGS 2018)

Martijn Tennekes
October 16-18, 2018
Predecessors of Mobile Phones

Car telephone system

Walkie-talkie
Why are mobile phones called ‘cell-phones’?

The target area is chopped into small cells such that each cell is covered by a cell tower.

Advantages:

- Close proximity to antennas -> small batteries
- Communication frequencies can be reused without disturbance from other antennas
**Type of antennas**

- **Cell tower**
  - 3 antennas, each covering 120º
  - Coverage up to 40 km

- **Rooftop cell site**
  - Coverage up to 40 km

- **Small cell**
  - Coverage up to 2 km

- **Indoor cell**
  - Coverage 200 m
Simplified cell-plan...

- Cell site (BTS)
- Cell antenna
- Cell coverage area
## Mobile phone generations

<table>
<thead>
<tr>
<th>Generation / description</th>
<th>Year of introduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>0G Mobile radio telephone, used in car telephones.</td>
<td>1940’s - 1970’s</td>
</tr>
<tr>
<td>1G Mobile analog telecommunications.</td>
<td>1981</td>
</tr>
<tr>
<td>2G Global System for Mobile Telecommunications (GSM) standard.</td>
<td>1991</td>
</tr>
<tr>
<td>Digital encryption used. Introduction of SMS and MMS messages.</td>
<td></td>
</tr>
<tr>
<td>3G Universal Mobile Telecommunications Service (UMTS) and CDMA2000 standards.</td>
<td>2001</td>
</tr>
<tr>
<td>Introduction of mobile internet. 10 Mb/s</td>
<td></td>
</tr>
<tr>
<td>4G Mobile broadband data, including voice over data. Enabling video conferencing and</td>
<td>2008</td>
</tr>
<tr>
<td>cloud computing. Download rates:</td>
<td></td>
</tr>
<tr>
<td>100 Mb/s at high mobility (cars/trains)</td>
<td></td>
</tr>
<tr>
<td>1 Gb/s at low mobility (pedestrians)</td>
<td></td>
</tr>
<tr>
<td>5G High speed mobile internet. Probably around 10Gb/s.</td>
<td>2020</td>
</tr>
</tbody>
</table>
Signaling data / Call Detail Records

**Signaling data**
- 100 variables, e.g.
  - Antenna id (geolocation)
  - Time/date
  - Country
  - Provider
  - Type of event
- Hundreds of records per device per day (4G)

**Mobile phone usage**
- Call (incl. being called)
- SMS (send and receive)
- Data (continuous logging)

**Events trigged by movements**, e.g. handovers from one area to another.

**Call Detail Records (CDR)**
- Used for billing
- Every provider should have them
Applications for Official Statistics

1. **Day Time Population**: the number of people in a certain region at a certain time. Useful for visitor counts during events, infrastructure planning, emergency management.

2. **Tourism statistics**: what places do they visit, where do they overnight, where do they come from?

3. **Commuting patterns**: where do people live and work? How and when do they commute?

4. **Urban planning / smart city**: what trips do people make in urban areas? By what mode of transport?

5. **Social networking**: who is connected to whom?

6. **Natural disasters**: what are the migration flows over time?
How to determine geolocation?

Voronoi tessellation

Cell tower

Area for which it is the nearest cell tower
A device using the cell tower is supposed to be somewhere in this polygon (uniform distribution)

Assumptions:
• All antennas are omnidirectional
• Areas do not overlap
Taking overlap into account

Bayesian approach

\[ Pr(g|a) = \frac{Pr(a|g) Pr(g)}{Pr(a)} \]

where \( g \) is a grid cell and \( a \) an antenna

- \( Pr(g) \) specifies a prior probability that a device is in grid cell \( g \)
- \( Pr(a) \) serves as a normalization constant
- \( Pr(a|g) \) is the likelihood, which can be defined as:

\[
Pr(a|g) = \begin{cases} 
0 & \text{if grid cell } g \text{ is not covered by } a \\
\frac{s(g, a)}{\sum_{a' \in B(a')} s(g, a')} & \text{if grid cell } g \text{ is covered by } a
\end{cases}
\]

where \( s(g, a) \) the (relative) signal strength of antenna \( a \) in grid cell \( g \) and \( B(a') \) is the set of grid cells covered by \( a' \)
Signal strength is complex in reality...

Radiation plots for a specific antenna:

Beam (simplified) for which signal strength is good

Signal delta (dBm)

Angle w.r.t. main direction
**R package mobloc**

- Developed at Statistics Netherlands, as part of the ESSnet Big Data (WP 5)
- Process:
  1. Setup signal strength model (interactive tool)
  2. Check cell-plan data
  3. Create coverage area per antenna (polygons)
  4. Calculate the relative signal strength and likelihood per grid cell and antenna.
  5. Check results (interactive tool)
- https://github.com/MobilePhoneESSnetBigData/mobloc
Signal strength configuration

**Antenna data**

**Analysis plots**
mobloc

Elevation data

Dots are antennas
mobloc

Voronoi tessellation

Needed to approximate coverage area polygon sizes (see next slide)
mobloc

Coverage area polygons

Voronoi area sizes are used as a proxy
mobloc

Cell Inspection Tool

Signal strength (dBm)

yellow = high
red = low
mobloc

Cell Inspection Tool

Relative signal strength

yellow = low
red = high
Cell Inspection Tool

Probability (likelihood)

yellow = low
red = high
From location to estimates

1. Deriving home location (needed because signaling data / CDR does not contain customer data. Method: take antenna(s) where a device is during nights (proxy) level: neighbourhood or municipality

2. Aggregate likelihood values per time frame (e.g. one hour) per device

3. Data cube:

4. Calibrate with population registers and education registers.
Further research

• Use prior information. Currently, the prior $Pr(g)$ is assumed to be uniform. More realistic would be to use the following sources:
  • Land use. People are more likely present in buildings than in grass fields. Also roads, railway tracks have a higher probability.
  • Cell-plan. Mobile Network Operators place the antennas in order to optimize the mobile communication service. Therefore, the cell-plan contains valuable information about where people are expected to be.

• Use of additional network data, e.g. timing advance, best area maps

• Improve the location by looking at the sequence of connected antennas, e.g. with MCMC