Towards Official Tourism Statistics – Machine Learning for Processing Signalling Data

Yvonne Gootzen, Marc Ponsen, Sigrid van Hoek, Shan Shah,
Marco Puts, Martijn Tennekes, Edwin de Jonge, May Offermans
Statistics Netherlands, Center for Big Data Statistics

INTRODUCTION

Statistics Netherlands is working on the next step towards official tourism statistics based on anonymised mobile phone signalling data. The privacy preserving process consists of multiple steps, detailed in this poster. This project is funded by Eurostat. Our geolocation algorithm enhances every event in the signalling data with a home- and current location variable. The table below shows these data of mobile device users in an aggregated form. Counts below 15 are omitted to preserve privacy.

Home location	Current location	Time	Count
Germany	Maastricht	09-07-2019 17:00	240

The table above allows for day time population statistics[1], but lacks detail for tourism statistics. Do observed devices in Maastricht belong to tourists, either foreign or domestic? Are they living there or visiting for work? This level of detail is required to create official tourism statistics. To achieve this, we aim to break this table down for different groups of tourism by introducing additional steps in the analysis. All events of one person are summarized to a single entry in the *cluster representation dataset*. Machine learning techniques are applied to distinguish types of tourism from signalling data, based on extracted features.

Behavior	Count
Border commuter	120
Day trip	105
Multi day holiday	15

GEOLOCATION

Properties of cells in the cellular network are translated to a signal strength model. After translating signal strength (dBm) to signal dominance (relative) a probability is determined of connection with a certain cell given a location. A Bayesian framework and prior assumptions allow for calculations of the probability of a location, given a connection event (Figure 1).

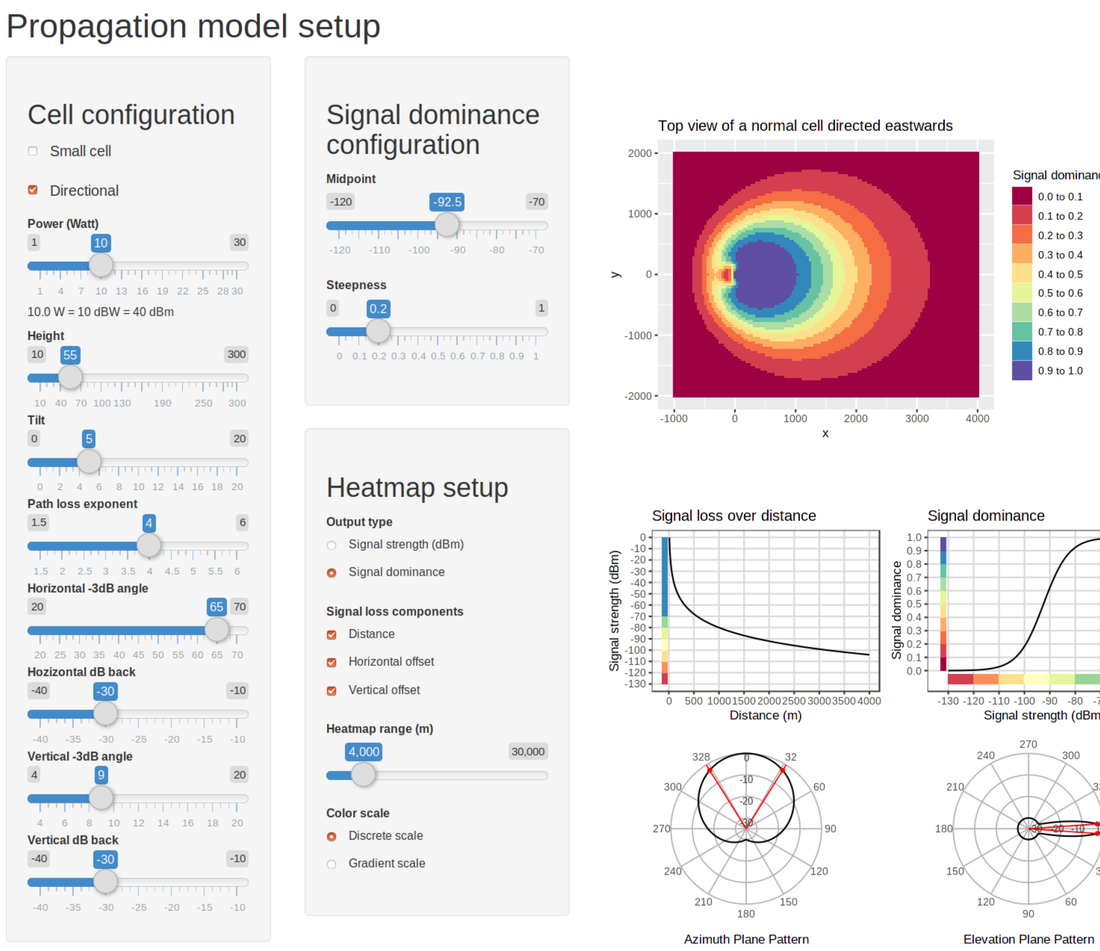

Figure 1. Geolocation algorithm and underlying signal strength model.

FEATURE EXTRACTION

Based on estimated geolocation for an event, we define the following features in an attempt to extract as much relevant behavioral information:

- Distance to border
- Municipality
- Part of day (morning, afternoon, evening, night)
- Distance to point of interest (e.g. touristic attraction)
- Cell size (small: inside, large: outside)
- Distance to highway / railroad

These event-based features are then summarized for every device ID, resulting in a compact representation which is processed by the cluster algorithm.

[1] https://dashboards.cbs.nl/v1/dtp/

ADVANCED GEOLOCATION

The current geolocation algorithm returns a probability distribution over likely locations of an event, independent from all other events. The combination of the events of one device for an entire day is expected to contain more information than separate events. This can in turn be used to estimate the travelled path more precisely (Figure 2). The particle filter method combines timely noisy observations to not only enhanced location estimates per event, but also potentially interesting features such as velocity and mode of transportation.

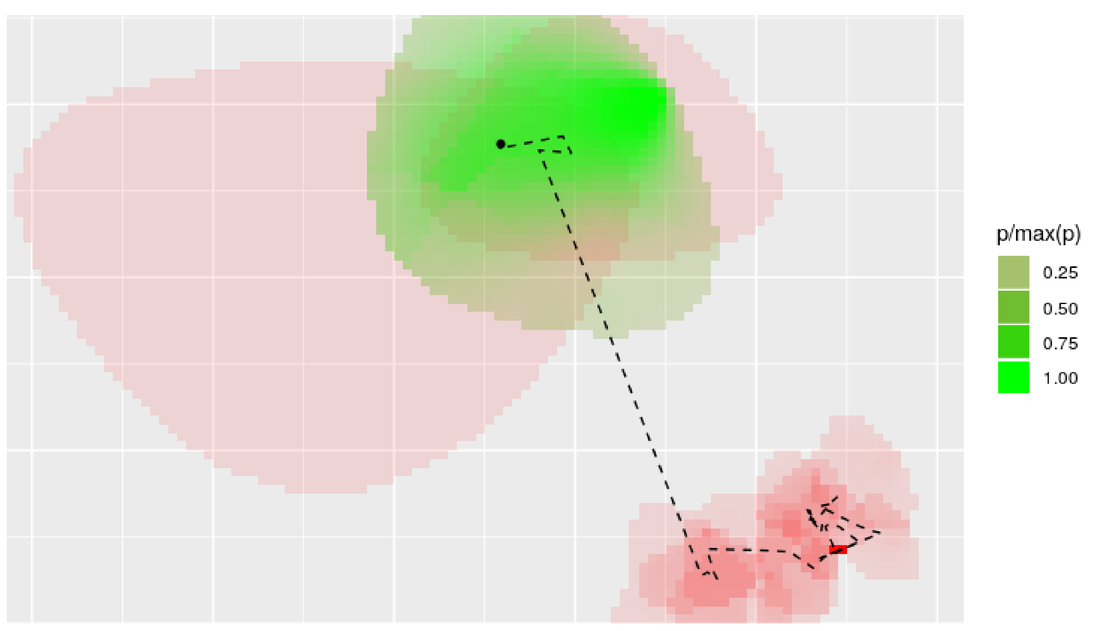

Figure 2. Expected travelled path by the particle filter, based on consecutive location distributions of single events from the geolocation algorithm.

MACHINE LEARNING

Based on the extracted features, we apply algorithms such as K-means in an attempt to distinguish different clusters. A simulated dataset is generated with predefined tourist behaviour to measure performance and validate cluster interpretability. The ground truth of the simulated data allows for a comparison in performance of different machine learning methods.

RESULTS

Unsupervised machine learning returns clusters without interpretable labels. A dashboard was designed to help domain experts at Statistics Netherlands in validating and labelling these clusters (Figure 3).

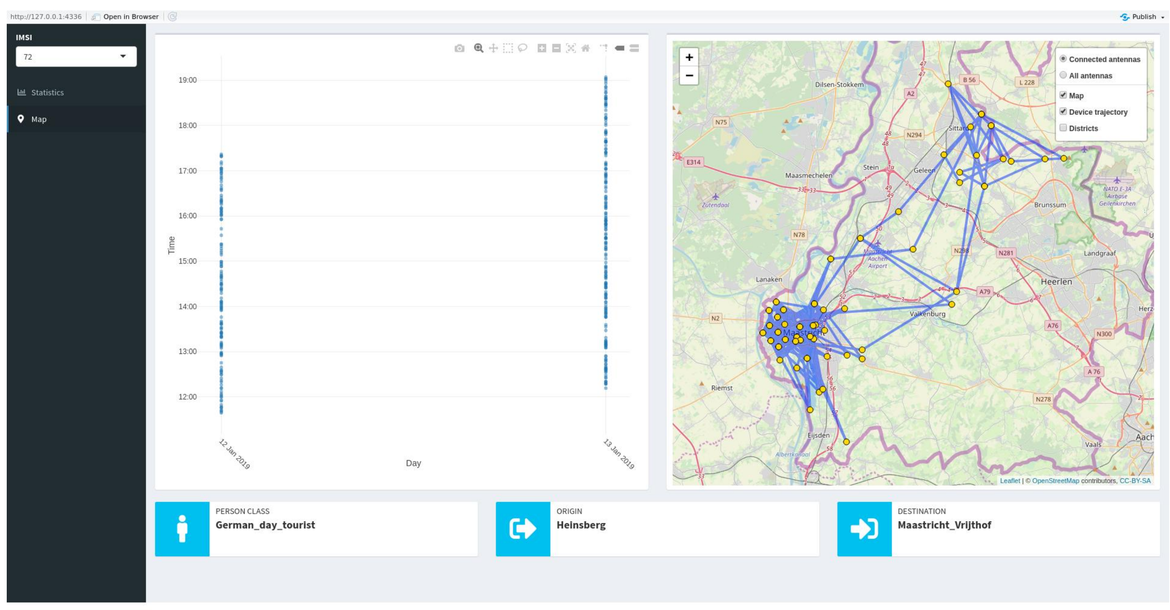

Figure 3. Visualization of simulated data, classified in the group of German day tourists. In this case, with home location Heinsberg and work location Maastricht.

FUTURE WORK

The ultimate goal is to apply clustering to real-life signaling data of a Dutch mobile phone provider. Since this dataset lacks the ground truth of its synthetic counterpart, validation will be sought in particular via man – machine interaction. The feedback loop between domain expert and algorithm should facilitate the creation of logical clusters, and in the end the creation of official tourism statistics.

Center for Big Data Statistics

eurostat