Towards Official Tourism Statistics - Machine Learning for Processing Signalling Data

Marc Ponsen, Yvonne Gootzen, Marco Puts, Martijn Tennekes, Edwin de Jonge, Shan Shah, May Offermans

March 15, 2019

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 four steps: geolocation, feature extraction, machine learning and aggregation.

Our geolocation algorithm processes information about an antenna network into probabilities that a connecting device is located in a given location.

In the feature extraction step, all observed antennas and their geolocation probabilities are analysed for a given time period. Multi-class particle filters utilise the temporal component of these observations. The output of these models is twofold: trajectory estimation and transportation mode detection. Trajectories describe an estimated location of a device over time. Transportation mode models are used for particle class definitions. The survival rates of each particle class is interpreted as the similarity between the actual movement and a theoretical transportation mode definition.

Extracted features for foreign devices are used as input for an unsupervised machine learning clustering algorithm. The resulting clusters are used to distinguish potentially new tourism developments. This allows for data-driven definitions, rather than analysis influenced by cluster definitions based on foreknowledge.

Detailed pre-aggregation analysis result in more accurate anonymised tourism statistics.