UNIVERSITEIT TWENTE.

Data Science [201400174]

Study: Master *Computer Science*
Course year 2017/2018, Quarter 1B

DATE
December 5, 2017

EXCERPT

Project 3: Transport domain [TRANSPORT]

TOPIC TEACHERS	MAIN TEACHER
Chintan Amrit	Maurice van Keulen
Mena Habib	
Maurice van Keulen	
Mannes Poel	

PROJECT OWNERS
Chintan Amrit
Karin Groothuis-Oudshoorn
Mena Habib
Maurice van Keulen
Mannes Poel
Stefano Schivo
Luc Wismans

Project 3

Project 3: Transport domain [TRANSPORT]

3.1 Introduction

Project owner: Luc Wismans
Primary topic: DPV or SEMI
Possibly combinable with TS as the data can be viewed as a time series.

Transportation is about moving of people and goods from A to B. Being able to transport people and goods is a prerequisite for economic growth and the consequence of the separation of production and consumption. There are various means of transportation (e.g. car, train or bicycle) being serviced by private companies or public authorities. Although transportation brings utility it also comes with a cost. Unwanted side effects (i.e. externalities) are caused by transportation affecting for instance the air quality, climate, safety and noise. Furthermore, there is a difference between the user needs and resulting behaviour and the societal needs and desired behaviour. Simple examples show that individuals pursuing their own objectives (e.g. shortest travel time from A to B) does not result in the optimal situation for society as a whole (e.g. minimal total delay in the system).

Road authorities are always working on improving the transport system balancing the societal objectives related to economic growth, minimizing externalities and user needs (i.e. sustainability). For this purpose they can adapt the system taking hard measures (infrastructural changes including deployment of intelligent transport systems) or influence the system providing services like traveller information. In most cases the infrastructure is owned by and a responsibility for governmental authorities as were the services provided on these networks. In the case of services these were at least controlled by the government (e.g. public transportation and provision of information). However, the past few years there is a shift of services provided by private parties not only because governmental authorities allow them to do so (providing data to such parties), but also as a result of an increase in data availability as well as ICT technologies not necessarily for transportation purposes deployed by private companies. Loop detector data, GPS, GSM, Bluetooth, WiFi, camera, smart card data, AVL and dedicated smartphone apps are examples of the sources capable of providing data of interest for transportation. Accurate maps/ topology of networks are needed to be able to map these data sources, offering the opportunity to connect and interpret this data for transportation purposes. Other spatial and temporal types of data like, socio economic data, points of interests, weather, deployment of measures, time tables and lines of PT might be of interest because of correlations with traffic conditions.

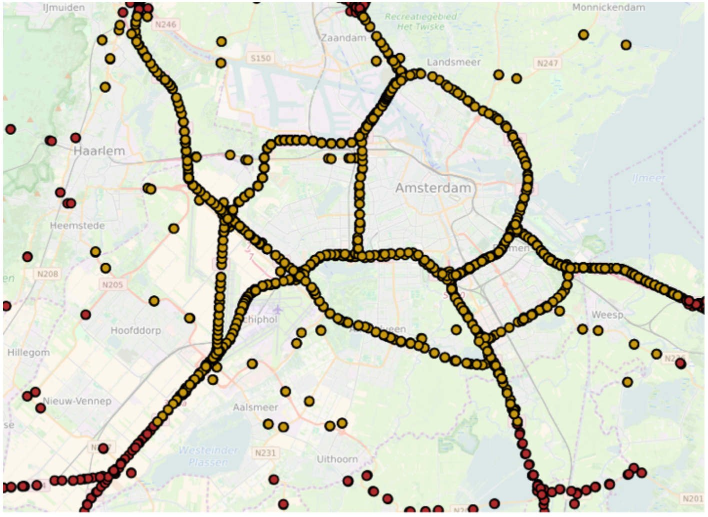
This data is obviously of interest for governmental authorities and private parties, because it allows to improve the decision support information for themselves or their customers. The enormous increase of data availability opens opportunities to better understand the current transport system (e.g. what are the traffic conditions, where are problems and when do they occur, and how do traffic conditions change as a result of construction works), to monitor the transport system (e.g. route choice effects of measures taken), as well as to improve predictions of the future (e.g. what will be the traffic conditions in the coming hour, what will be the travel time from A to B tomorrow during rush hour, etc.). Furthermore, it is useful to have some knowledge on GIS software packages like the open software QGis.

3.2 Description of data set

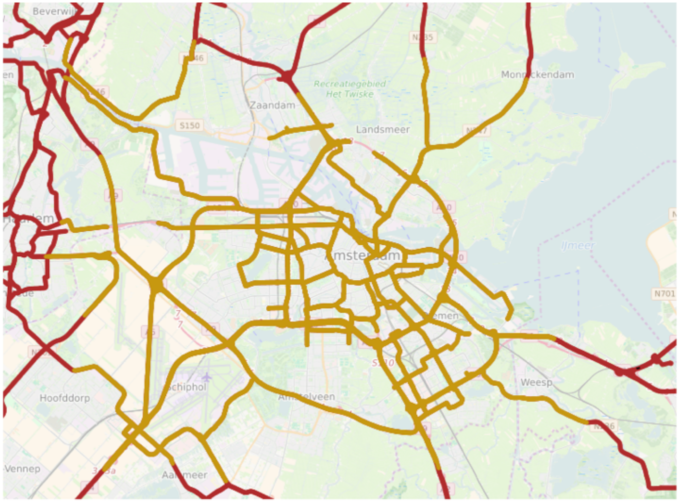
For this domain there are several data sources available. All data from NDW and BE-mobile is for the same time period June 1st 2016 till July 18th 2016 (for (part of the) Amsterdam region.

1. Data delivered by NDW containing

    - Flows and speeds from loop detectors, 1 minute aggregates (CSV-files)



    - Travel times of predefined routes (CSV-files)

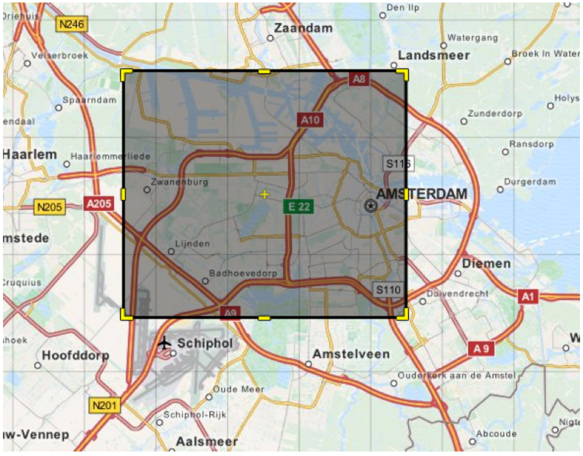


    - Status information on occurrence, whole of Netherlands (XML),
        - traffic measures,

  - opening bridges,
  - road works,
  - traffic jams
  - traffic reports
  - status opening rush hour lanes

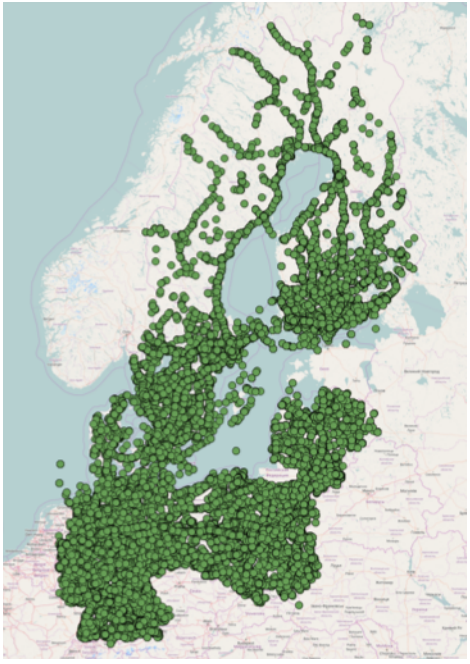
2. Data delivered by BE-mobile

   - Tripdata within bounding box part of A10 west
   - Important: Use of this data requires you to sign an agreement regarding the use.



3. CDR data Estonia

   - Outbound Call Detail Records of Estonians in 7 other countries (roaming data). Already connected to mast locations using OpenCellid.



3.2.1 Description NDW data speed, flow and traveltime

Extensive descriptions (in Dutch) of data can be found in

- NDWInterfacebeschrijvingversie2.2 (1).pdf
- handleiding.pdf

Datasets:

1. Datasets containing all measurements:
   - Filename speeds: utwente snelheden groot amsterdam.zip
   - Filename flows: utwente intensiteiten groot amsterdam.zip
   - Filename travel times: utwente reistijden groot amsterdam _20160916T115957_197.zip
2. Datasets containing metadata:
   - Filename speeds: utwente snelheden groot amsterdam 1 dag met metadata_20160916T105028_197.zip
   - Filename flows: utwente intensiteiten groot amsterdam 1 dag met metadata (2)_20160916T104708_197.zip
   - Filename travel times: utwente reistijden groot amsterdam 1 dag met metadata_20160916T103803_197.zip

Files within zip-files: Complete dataset is separated in several files Csv-files: "," separates fields

The raw data used by NDW are 1-minute aggregates. This means that the available datasets contain the raw data of NDW. As a result several fields are not filled, because these are used when higher aggregates would be delivered by NDW (e.g. numberOfInputValuesused, standerdDeviation and dataError).

The datasets containing all measurements need to be combined with the datasets containing metadata to add locational data (i.e. to be able to connect the measurements to the location of measurement). The metadata contains the data of 1 day containing all measurement locations (and additionally more information of possible interest, like lane number and lat lon position of measurement) and the measurements for one day, the dataset containing all measurements contain the measurements of all days between June 1st 2016 till July 18th 2016. Furthermore, if there is data available per lane and/or data per vehicle class (e.g. cars and trucks), these measurements for the same location and same minute will be presented in a separate line. For this purpose you also need to combine the complete dataset with the metadata file.

Important attribute fields are:

- MeasurementSiteReference and index: to connect the metadata information to all measurements
- PeriodStart and periodEnd indicating the timeinterval of measurement
- avgVehicleFlow: measurement of flow
- avgVehicleSpeed: measurement of speed
- avgTravelTime: measurement of traveltime
- measurementSide: side of road i.e. eastbound or westbound
- specificLane: lane nr, number starts left
- specificVehicleCharacteristics: vehicle class
- startLocatieForDisplayLat and startLocatieForDisplayLong: lat-lon location of measurement (or for travel time starting point) based on ETRS89 system, which is the same as WSG84
- generatedSiteName: description of location (unfortunately in Dutch)
- Specific for traveltime
- computationMethod: method used to compute the average measurement
- measurementEquipmentUsed: sensor used to measure, Dutch description (e.g. loop detector in Dutch "Lus")
- startLocatieForDisplayLat and startLocatieForDisplayLong: lat-lon location of starting point section of measurement
- eindLocatieForDisplayLat and eindLocatieForDisplayLong: lat-lon location of end point section of measurement
- lengthAffected: length of section

3.2.2 Description Status information

Available information on data description can be found in DATEX-II Dutch Profile 2015-2a (NP2015-2a).pdf, which is in English.

Datasets

- ActualTraffic.zip
- Bridge.zip
- ONDA.zip
- RoadMaintenance.zip
- SRTI.zip
- TrafficInfo.zip

Datasets contain data for days between June 1st 2016 till July 18th 2016. However, in this case the available data for the Dutch network.

3.2.3 Description of BE-mobile data

The Be-mobile data contains trip data of individual vehicles on the Amsterdam region network for all days between June 1st 2016 till July 18th 2016. Note that these are floating car data of a sample of vehicles equipped with devices providing these information. Further note that there is a possible bias in this sample (i.e. it is not a random sample of all vehicles driving within the Amsterdam region network).

Two datasets:

1. Trip data:
   - Archive 1.zip
   - Archive 2.zip
   - Archive 3.zip
2. SegmentId information
   - staticDataFiltered.xlsx

Files within the zipfiles are numbered, these numbers do not have any meaning. The csv files contain the following information:

- anonimized vehicle ID
- time stamp: YYYYMMDDHHMMSS
- segmentID
- traveltime (ms)
- covered distance (mm), not neceserally the same as segment length. Gps positions were mapped on segments and between two positions the route is determined. End of route can be placed at certain position on a segment.

staticDataFiltered.xlsx contains background information on segments and contains following information:

- SegmentID
- BeginLongitude and BeginLatidude: lat lon position starting point segment
- EndLongitude and EndLatitude: lat lon position end point segment
- OptimalTTMs: freeflow travel time segment in ms
- Lengthmm: length of segment in mm

3.2.4 Description of CDR data

Filename: 1week_outbound_data_extended.csv

Attributes:

- pos_time: date and time measurement
- usr_id: unique user identification number,
- mcc: country code
- lac: locational area code

- cell_id: identification number of mast (based on openCellid.org),
- lon, lat: location of measurement based on location mast according to openCellid.org
- type: reason of connection with mast:
  - HDR (Header Record, all types)
  - MOC (Mobile Originating Call, outgoing mobile call)
  - MTC (Mobile Terminating Call, incoming mobile call)
  - SMMO (Mobile Originating SMS Event, outgoing SMS)
  - SMMT (Mobile Terminating SMS Event, incoming SMS)
  - Data (GPRS/UMTS event, outgoing data session)
  - TRL (Trailer Record, all types)

3.3 Description of challenge

We provide you with four example challenges you could work on in your project.

1. State estimation The data provided for Amsterdam for the same time period are different sources which can be used to estimate the traffic conditions (speeds and flows) on the entire network. NDW data provides speed and flow measurements on locations based on all passing vehicles. NDW also provides travel times on routes and the Be-mobile floating car data provides spatio-temporal information for all locations based on a proxy of all vehicles. The challenge is to use this data to provide the spatio-temporal state estimates (e.g. speeds and flows for segments for every minute) as complete as possible. Also other state variables like routechoice/turnfractions or demand can be of interest.

2. Prediction The data provided for Amsterdam can be used to build a prediction module which predicts future (can be the next minutes or the next day) traffic states or travel times

3. Influence of roadworks on traffic conditions The provided NDW data also contains information on for example traffic measures, road works and bridge openings. This challenge is about analysing the impact of these aspects on the traffic conditions. Are there correlations and can we derive knowledge which we can use for future decision.

4. Reconstructing routes The outbound CDR records of Estonians can be used to analyse their trips. The challenge is to reconstruct the trips and routes of people. Is it possible to derive the mode of transport?

For all challenges it is required to connect the available data with a network and provide visualizations. First steps could be:

- Analysing and understanding of data set, e.g. by making figures of measurements for a specific location or specific user for a day or multiple days, computing averages, checking plausibility, determine whether there is data missing, etc
- Visualize locations of measurements
- Select suitable part of network for case study and determine what data is available for this part and which is not
- Connect data with a road network, visualizing locations
- Determine desired outcome and possible ways to compute this
- ...

3.4 Tips and suggestions

3.4.1 Creating new geographic maps for use in Tableau

If you’d want to show data using a geographic map and the available maps in Tableau do not suffice, then you can also import your own from *shapefiles* using a Geographic Information System (GIS) such as QGIS or ArcGIS. A shapefile is a special file that stores the polygones that make up a map.

See `http://onlinehelp.tableau.com/v10.1/pro/desktop/en-us/help.htm#maps_shapefiles.html%3FTocPath%3DDesign%2520Views%2520and%2520Analyze%2520Data%7CBuild%2520and%2520Use%2520Maps%7C_____1` for details.