Graph-based Inference from Non-probability Road Sensor Data

Jonas Klingwort[1,2] Bart Buelens[3] Joep Burger[2] Rainer Schnell[1]

[1]University of Duisburg-Essen

[2]Statistics Netherlands

[3]VITO

NPSO Innovatiedag
Den Haag, 26 November 2019

Statistics Netherlands

UNIVERSITÄT DUISBURG ESSEN

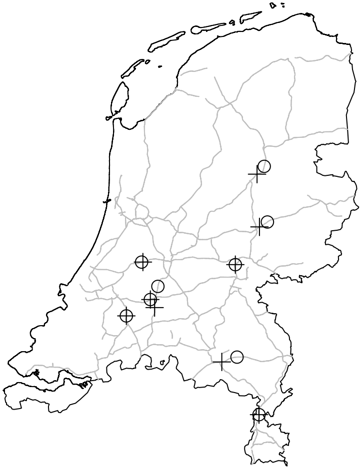
Introduction

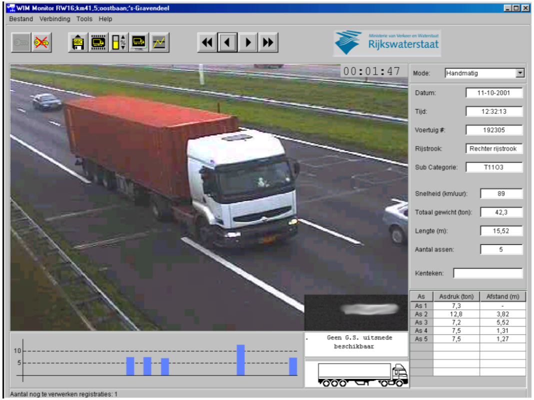
- Non-probability based sensor data is becoming increasingly popular in official statistics.
- Empirical studies required to assess the usefulness and applicability of such datasets.
- Such an assessment is possible when a survey and sensor independently measure an identical target variable.
- Dutch Road Freight Transport Survey provides an annual estimate of transported shipment weights.
- Real-time data from road sensor network could provide faster/cheaper estimates and reduce response burden.
- Can road sensor data replace survey-based estimates of truck days and transported shipment weights?

Data – Sensor

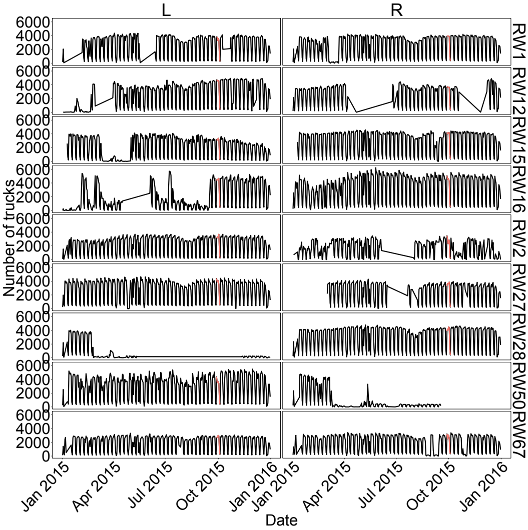
- Weigh-in-motion road sensor data of 2015 ($n = 36$ million).
- Purpose: detect and enforce penalties on overloaded trucks.
- Dynamic measurement of the weight for each passing truck at 18 sensor stations.
- Measurements: photograph of the front/rear license plate, total weight, axles pressure, and truck classification.
- Weight of the entire unit (truck, trailer, and shipment) measured.
- Observations can be linked on a micro-level to vehicle and enterprise register using license plate and time-stamp as unique identifier.
- Transported shipment weight is total weight measured by sensor minus empty truck and trailer weights linked from register.

Road sensor network – Stations & Measurement

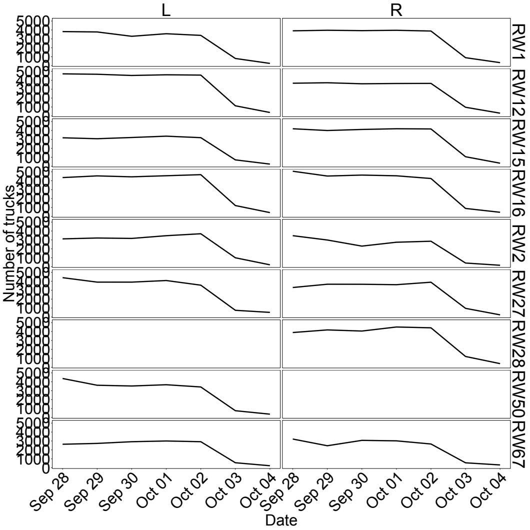




Road sensor network – Daily counts



Road sensor network – Daily counts (week)



Road network – Graph

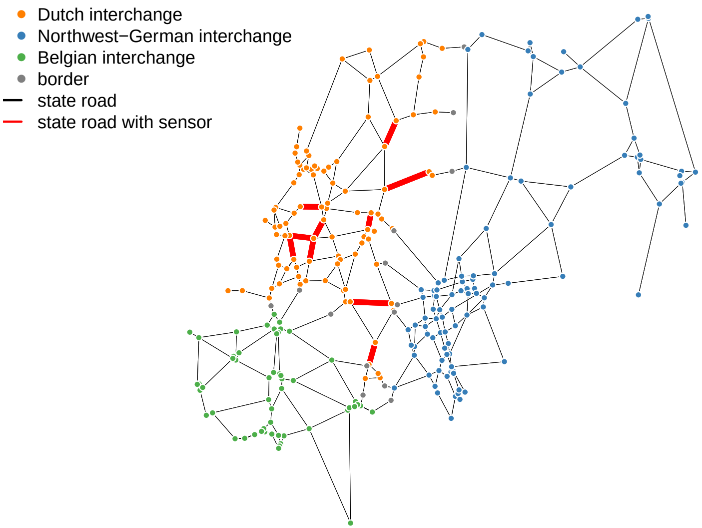
- Directed graph contains the state road network of the Netherlands, Belgium, and Northwest Germany (Lower Saxony, Bremen and Northrhine-Westphalia).
- Web scraping used to gather data from wiki on road networks (`https://www.wegenwiki.nl/Nederland`).

- Graph consists of 108 vertices, 284 edges.
- 6 computed features for vertices (incoming / outgoing).

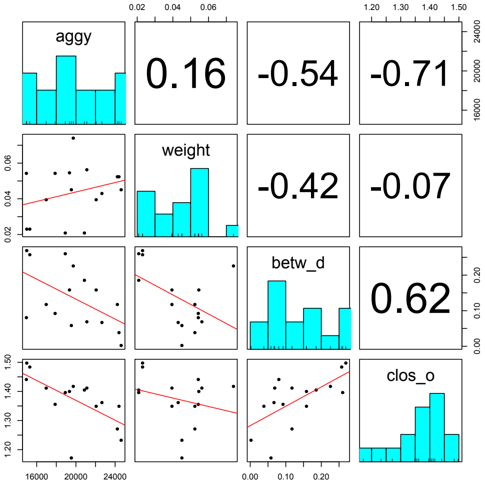
Road network – Vertex features

- Degree: number of incoming and outgoing edges.
- Strength: total weight of incoming and outgoing edges.
- Betweenness: number of shortest paths passing through.
- Closeness: inverse of the average length of the shortest paths to/from all other vertices.
- Vulnerability: loss in efficiency when excluding vertex.
- Cluster Coefficient: probability that adjacent vertices are connected.
- Weight: inverse haversine distance (great circle distance).

Road network – Graph of traffic junctions (vertices) and highways (edges)



Road sensor network – Data exploration (week)



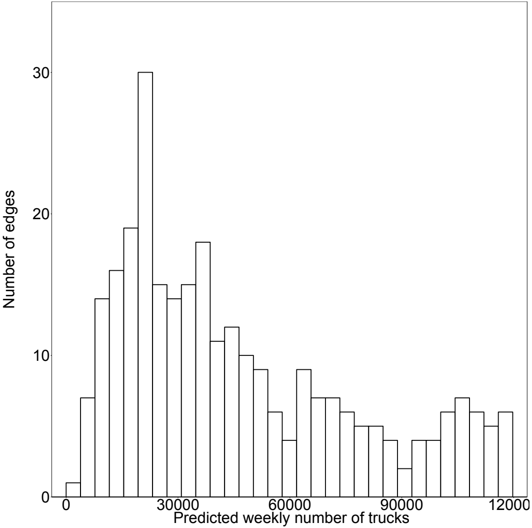
Method

- Truck $i \in \mathcal{U}$, with set $\mathcal{U}$ being the population of N trucks in the vehicle register.
- Road state network is represented by a weighted directed graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ consisting of set $\mathcal{V}(\mathcal{G})$ of V vertices (traffic junctions) and a set $\mathcal{E}(\mathcal{G})$ of $E = V(V-1)$ edges (state roads).
- The graph is represented by a $V \times V$ adjacency matrix W with w_{od} the weight of the edge from origin $o \in \mathcal{V}(\mathcal{G})$ to destination $d \in \mathcal{V}(\mathcal{G})$.
- Weight takes the strength of the connections into account and is currently the inverse edge length (km^{-1}).
- WIM station $s \in \mathcal{S}$, where set $\mathcal{S}$ is a non-probability sample of $|\mathcal{S}| = n$ stations from $\mathcal{E}(\mathcal{G})$.

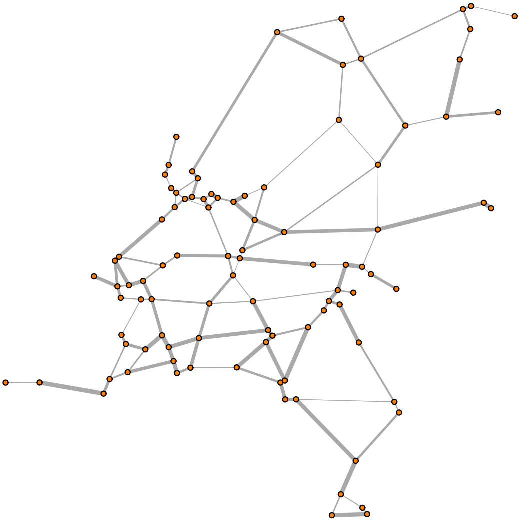
Method

- X, an $N \times k$ matrix with features about trucks from the vehicle register (e.g. number of wheels, horsepower) and about truck owners from the business register (e.g. economic activity, size class).
- Z, a $V \times l$ matrix with features about vertices.
- Model the probability of detecting a truck between origin o and destination d as a function of X, W, and Z using a GLM.
- The modeled probabilities are multiplied with the number of trucks registered in the vehicle register constituting the study population to derive the edge counts.
- Goal: Correct for the absence of sensors on most edges, and for the selectivity in their presence.

Modeled counts



Road network – Weighted graph of traffic junctions and highways



Future research

- Add more features (e.g. traffic intensity, weather, regional characteristics).
- Include more weeks.
- Account for temporal dependency and correct sensor errors using time series modeling.

Questions

- How to incorporate spatial dependency?
- Suggestions for further features?