Generaliseren uit niet-kanssteekproeven

Joep Burger, Bart Buelens, Jan van de Brakel

Statistics Netherlands

NPSO, 2 May 2017, Antwerpen

Disruptive technologies

Big picture

UBER

Transport

Light

Spotify

Music

Disruptive technologies

Big picture

Official statistics

Big data

- Human-sourced data
  - social media
  - internet search
- Process-mediated data
  - scanners
  - electronic funds transfers
- Machine-generated data
  - GPS
  - sensors

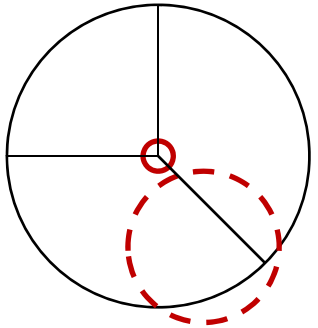
UNECE 2013

Potential

- Timelier
- Higher frequency
- More detail
- Higher precision (≠ more accurate)
- Lower measurement bias
- Cheaper
- Less burden

Challenges

- Representation
  - big data sample ≠ population of interest
  - non-probability samples
- Others
  - measurement (data ≠ information)
  - processing
  - privacy
  - continuity



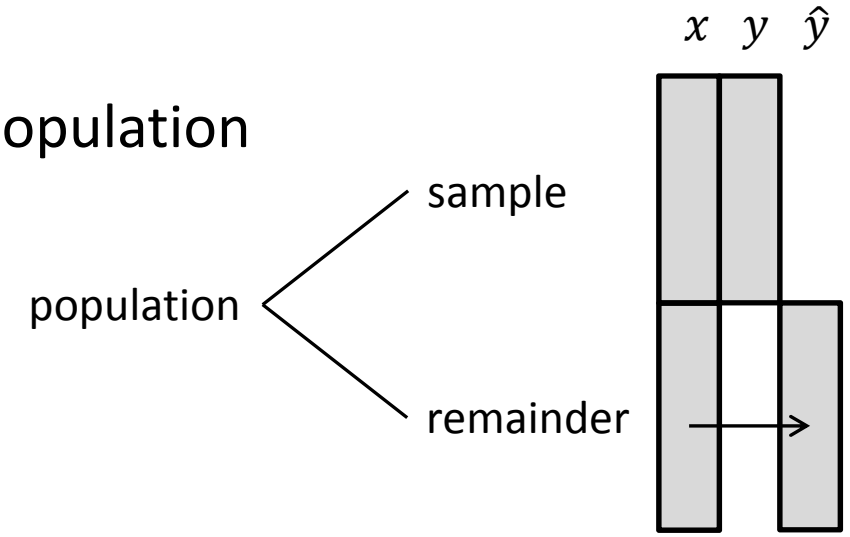
Street Bump App

- Automatic pothole mapping

- Selection bias
  - wealthier, younger neighborhoods

INPS

- Inference: generalize from sample to population
- Predict missing values
  - Pseudo-design-based
  - Model-based
  - Algorithmic
- Auxiliary information
  - Known for all units in the population



Formal

- Population quantity of interest

$$G_Y = G(y_i) \text{ for } i = 1, \dots, N$$

- y values known for sample *S*, unknown for remainder *R*
- Prediction estimator

$$\hat{G}_Y = G(\{y_i\}_{i \in S} \cup \{\hat{y}_i\}_{i \in R})$$

with

$$\hat{y}_i = F(x_i, \hat{\beta}) \text{ for } i \in R$$

$$\hat{\beta} = P_F(x_i, y_i) \text{ for } i \in S$$

- Variance through bootstrapping

Methods

- Sample mean (SAM)
- Pseudo-design-based (PDB)
- Generalized linear model (GLM)
- k-Nearest neighbors (KNN)
- Artificial neural network (ANN)
- Regression tree (RTR)
- Support vector machine (SVM)

Sample mean (SAM)

- Mean of observed units

$$\hat{y}_j = \bar{y} = \frac{1}{n}\sum_{i \in S} y_i$$

Pseudo-design-based (PDB)

- Mean of observed units within stratum
  - Use auxiliary variables

$$\hat{y}_j = \bar{y}_h = \frac{1}{n_h} \sum_{i \in h \cap S} y_i \; \forall j \in h$$

Generalized linear model (GLM)

- Generalized combination of auxiliary variables

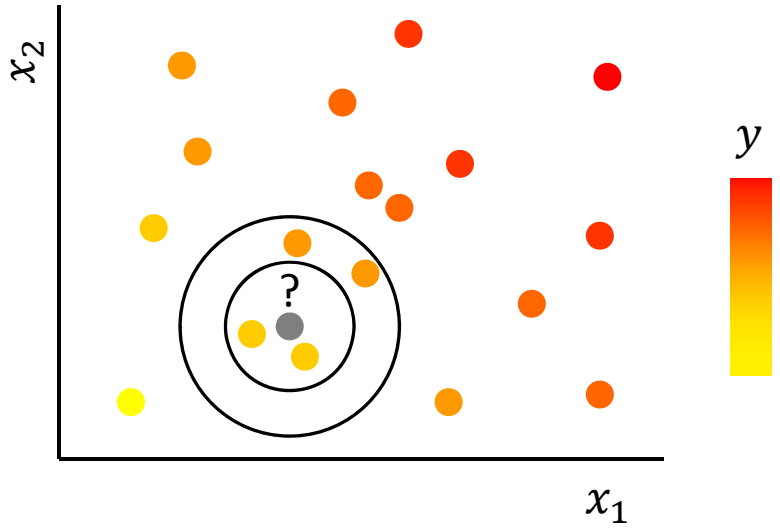
$$g(E(Y_i)) = \beta x_i$$

$$\hat{y}_j = g^{-1}(\hat{\beta} x_j)$$

k-Nearest neighbors (KNN)

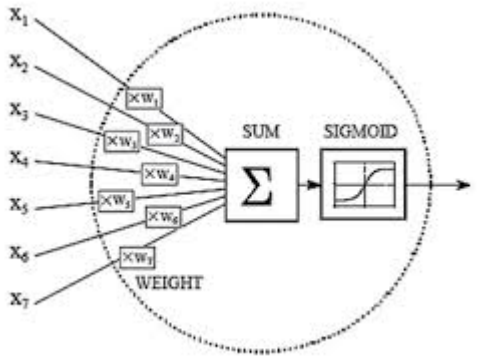
- Mean of k observed units closest in x space
  - Distance measure

$$\hat{y}_j = \frac{1}{k} \sum_{a \in A_j} y_a$$

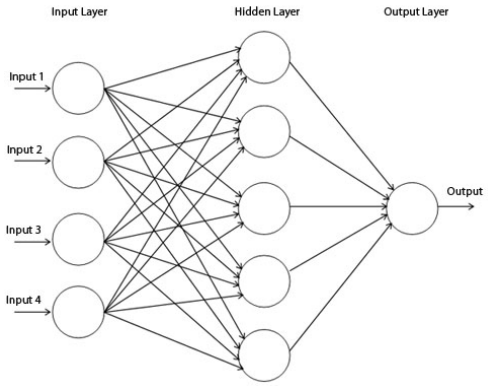


Artificial neural network (ANN)

- Artificial neuron



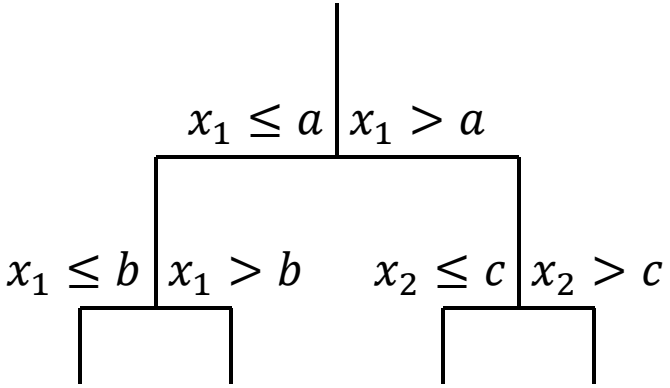
- Network of artificial neurons



$$\hat{y}_j = ANN(x_j, \hat{\beta})$$

Regression tree (RTR)

- Construct binary tree
  - Maximize between variance
  - Stopping criterion



- Mean of observed units within leaf

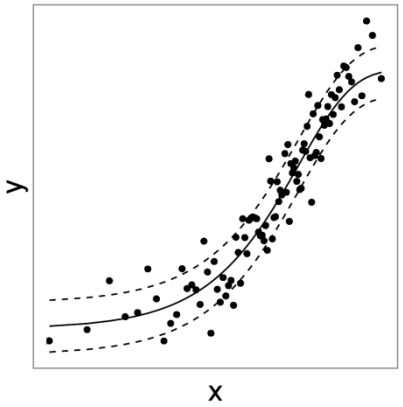
$$\hat{y}_j = RTR\left(x_j, \hat{\beta}\right) = \frac{1}{n_\lambda} \sum_{k \in \lambda \cap S} y_k$$

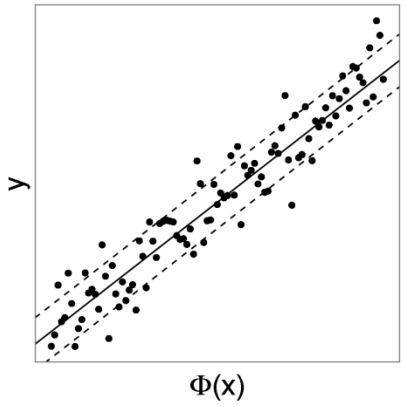
- Algorithmic version of PDB

Support vector machine (SVM)

- Linear in $\mathbb{R}^{M>N}$ $(\Phi(x))$
- Learn in $\mathbb{R}^{N}$(Kernel trick)

Original space $\mathbb{R}^{N}$ Higher-dimensional space $\mathbb{R}^{M}$





$$K(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle_{M>N}$$

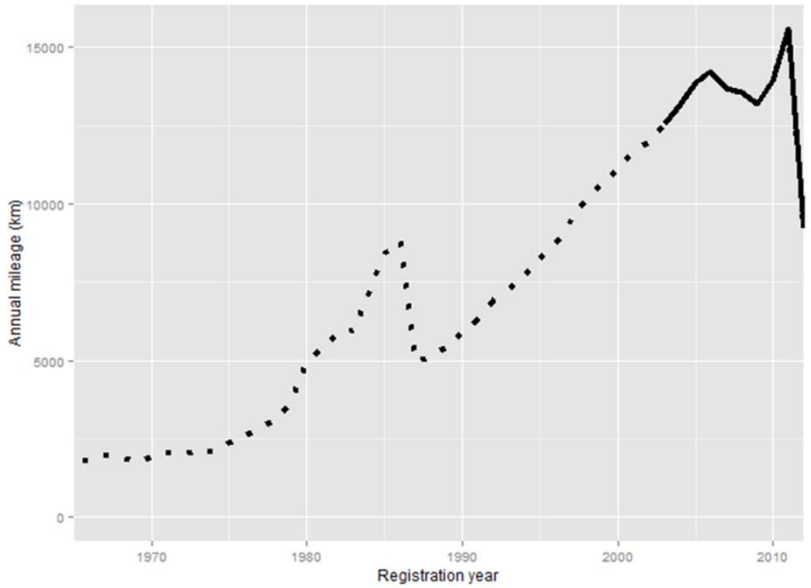
$$\hat{y}_j = \sum_i \hat{\alpha}_i K(x_i, x_j)$$

Case study

- Online Kilometer Registration
  - 6,7 mln privately owned cars
  - Mileage readings → annual mileage
- Auxiliary variables
  - Registration year
  - Weight
  - Fuel type
  - Owner's age

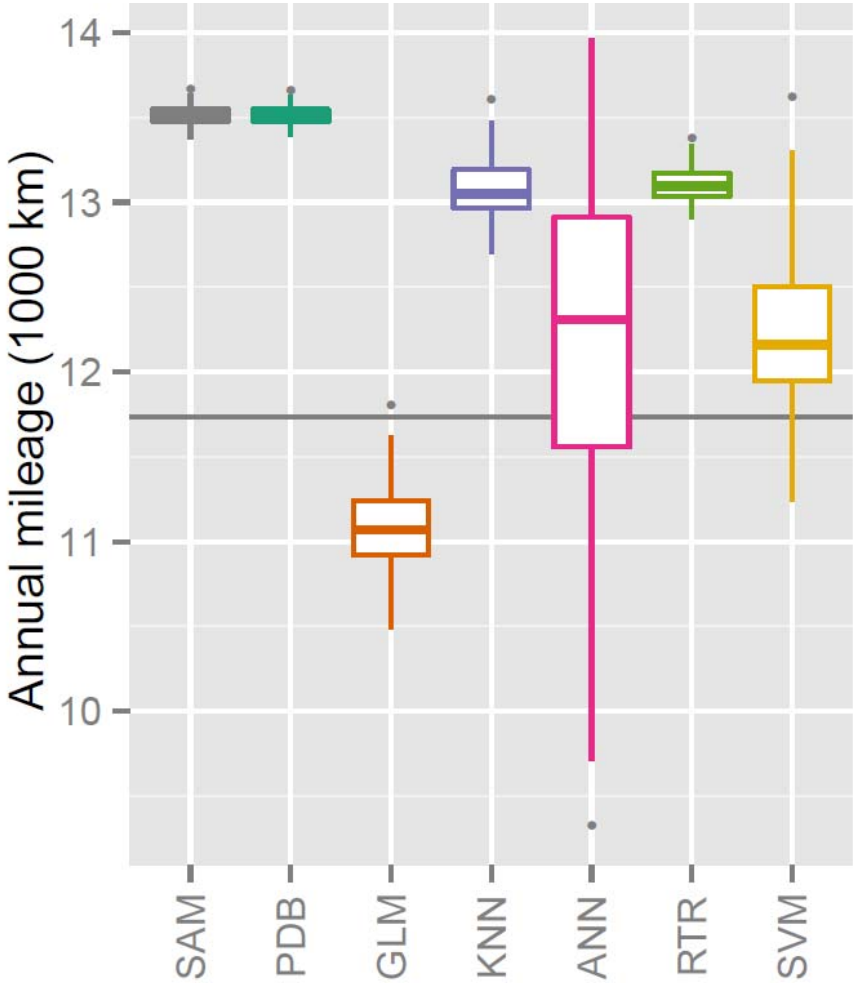
Non-probability sample

- Only data about young cars

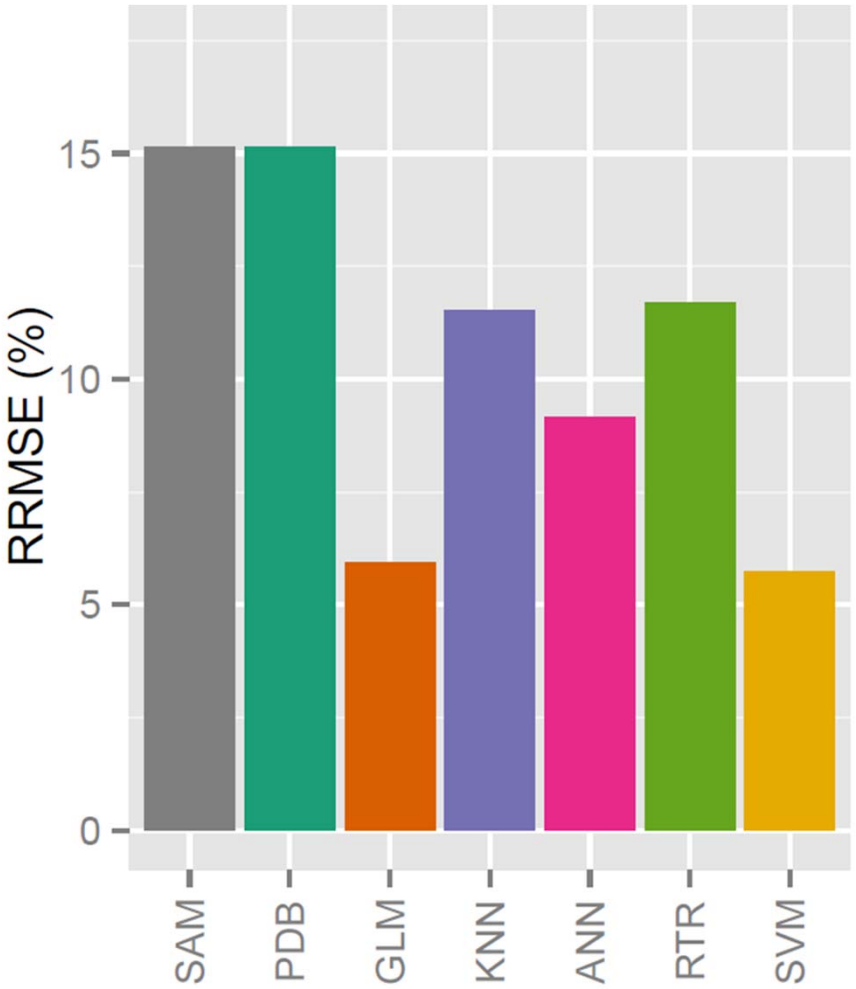


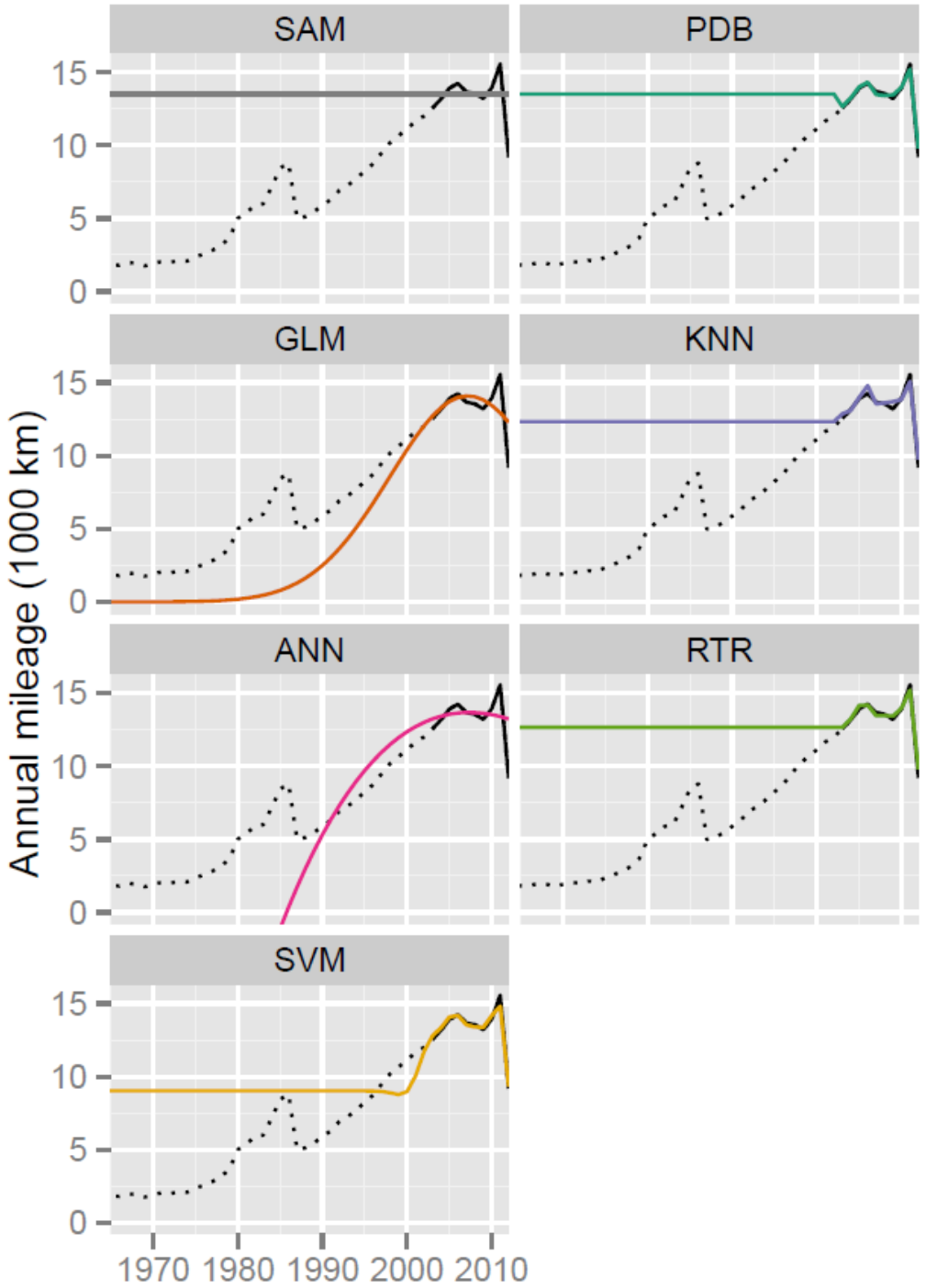
- Inference: $\hat{G}_Y = G(\{y_i\}_{i \in S} \cup \{\hat{y}_i\}_{i \in R})$

Inference


Annual mileage (1000 km)
14
13
12
11
10
SAM
PDB
GLM
KNN
ANN
RTR
SVM

Accuracy




SAM
PDB
GLM
KNN
ANN
RTR
SVM
Annual mileage (1000 km)
0
5
10
15
1970
1980
1990
2000
2010

Conclusions

- Both PS and NPS may suffer from selection bias
- Beyond pseudo-design-based methods
  - Model-based
  - Algorithmic
- (Many, continuous) auxiliary variables crucial
  - Registers
  - Paradata
  - Profiling

Working paper

https://www.cbs.nl/nl-nl/achtergrond/2015/44/predictive-inference-for-non-probability-samples

Contact

j.burger@cbs.nl