Imputering av körsträckor

21/3 2024

Petter Ehn Wingårdh
Masteruppsats VT2023
Stockholms Universitet
Handledare: Dan Hedlin

Bakgrund

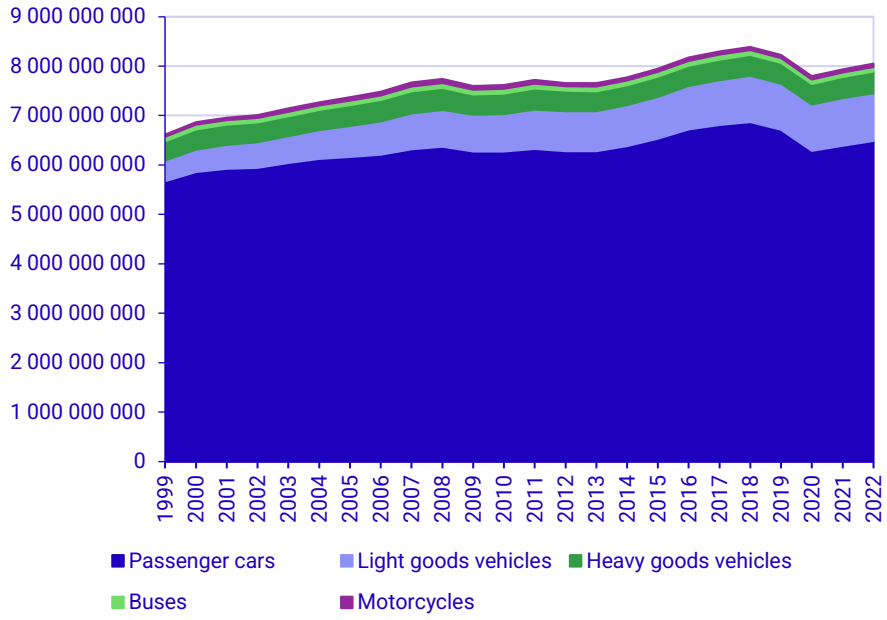
Körsträckor med svenskregistrerade fordon

- Officiell statistik som tas fram av SCB på uppdrag av Trafikanalys.

Statistikens användningsområden:

- Följa utvecklingen av totala och genomsnittliga körsträckor över tid
- Analysera övergången från traditionella bränslen till andra drivmedel
- Input till FoU

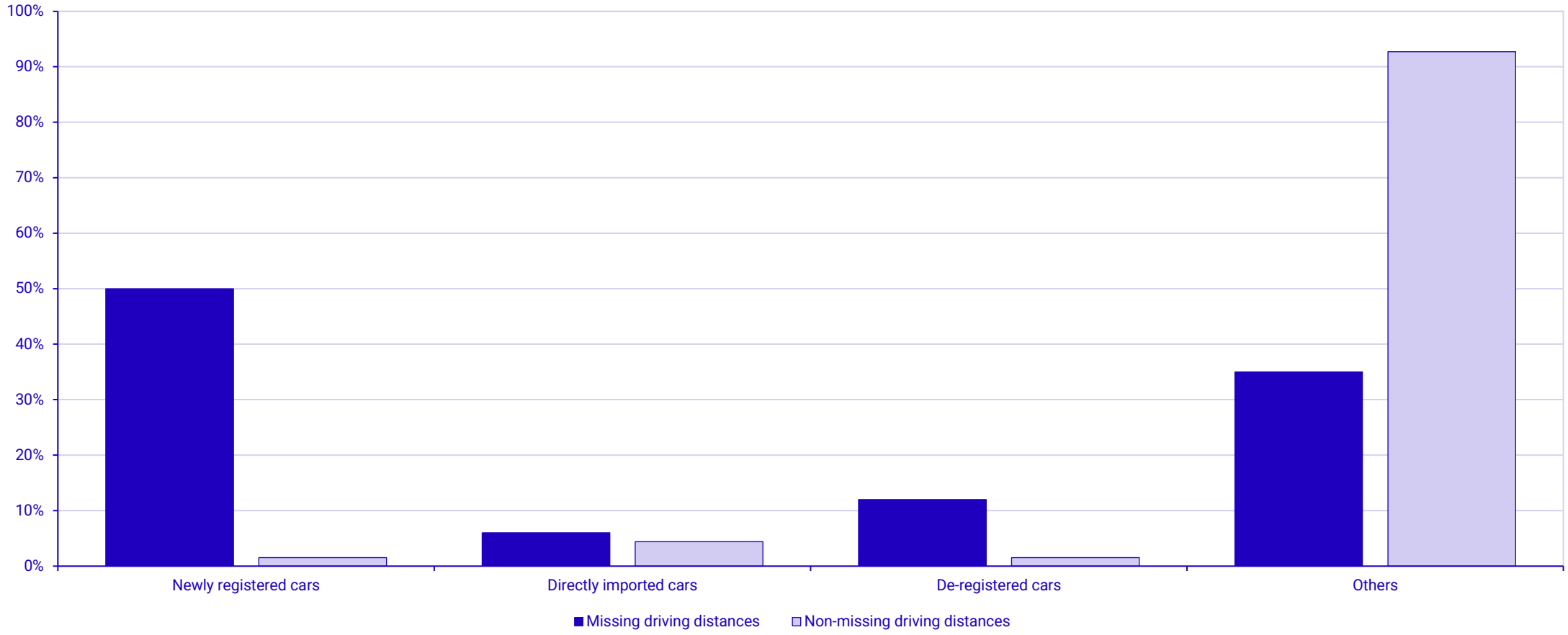
Figure 1
Driving distances (10 kilometres) by vehicle type for the years 1999-2021



Bakgrund

- Data från Transportstyrelsens vägtrafikregister
- Mätarställningsdata från besiktningar används för att beräkna körsträckor
- 28-37 procent saknade körsträckevärden 2018-2022

Vilka bilar saknar körsträckevärden?



Nuvarande metod: Geometrisk medelvärdesimputering

Imputeringsgrupper baserade på:

- Om bilen är leasad eller inte
- Drivmedel
- Bilens totalvikt (fem intervall)
- Bilägarens ålder (under eller över 60)
- Bilens ålder (5-årsintervall)

Problem:

- Minskad information vid kategorisering av kontinuerliga variabler
- Underhåll av grupperingarna
- Alla imputerade bilar i en grupp får samma värde
- Ingen dokumentation som motiverar geometriska medelvärdet

Data

- Data för 2018
- Nyregistrerade bilar exkluderas från analysen
- 3,7 miljoner bilar återstår varav 18 procent saknar körsträckevärden.

Explanatory variable	Number of categories
Leased	2
Fuel type	8
Gender of owner	2
Class II passenger car (Motorhome)	2
Curb weight	continuous
Model age	continuous
Age of owner	continuous
Directly imported	2
Registration status	3
Engine power in KW	continuous
“Super-Green car”	2
Car body type	5
Municipality classification	7
Number of days in traffic in the reference year	continuous

Imputeringsmetoder

Tweedie-GLM

Random forest

Full dataset	Observed driving distance	Training	Validation	Test
3 676 029	2 986 233	1 477 785	761 339	747 109

Table 1. Breakdown of the full dataset: total number of cars, number of cars with observed driving distances, and number of cars in the training, validation, and test datasets.

Utvärderingsmetod

Inget facit för bilarna som saknar körsträckor

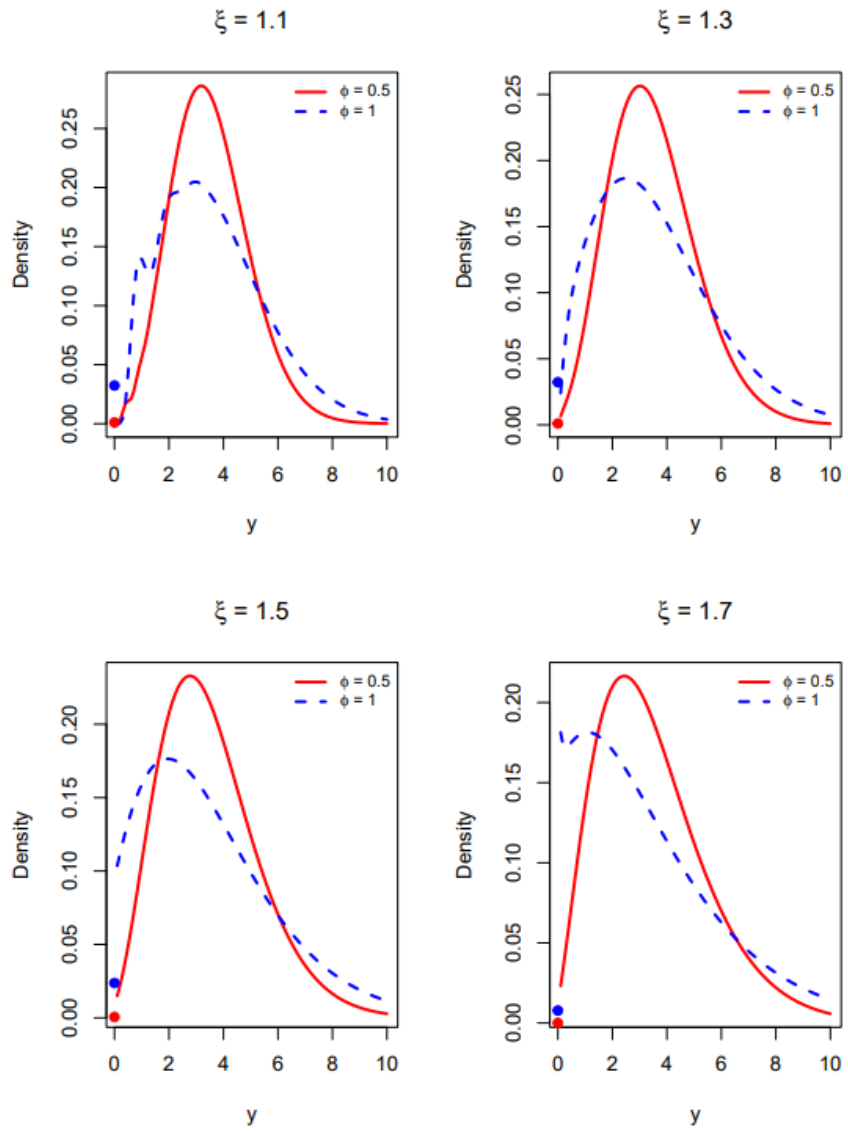
$\Rightarrow$ Vi delar slumpmässigt upp data i flera delar och predikterar körsträckor för “test-data”.

GLM

- Generalisering av multipel linjär regression
  - Inget normalfördelningsantagande
  - Förhållandet mellan väntevärdet för studievariabeln och den linjära kombinationen av prediktorerna behöver inte vara linjärt.

Tweedie GLM

- Körsträckor är positivt kontinuerliga (men kan eventuellt anta värdet 0), med icke-konstant varians.



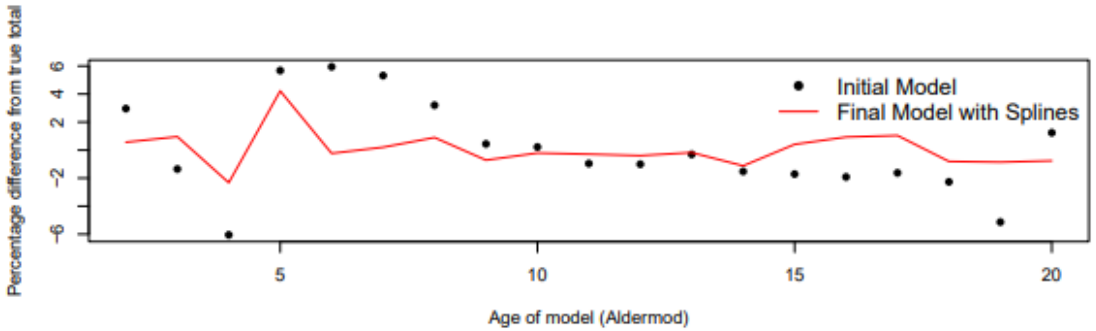
Val av prediktorer

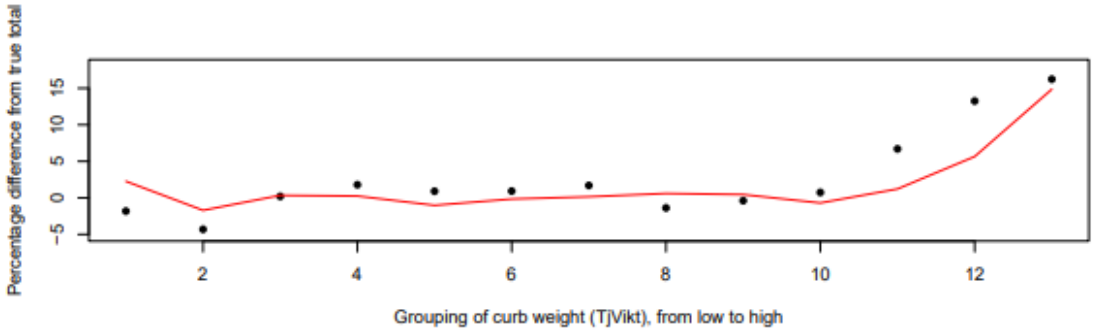
Modeller som tränats på träningsdata och utvärderats på valideringsdata

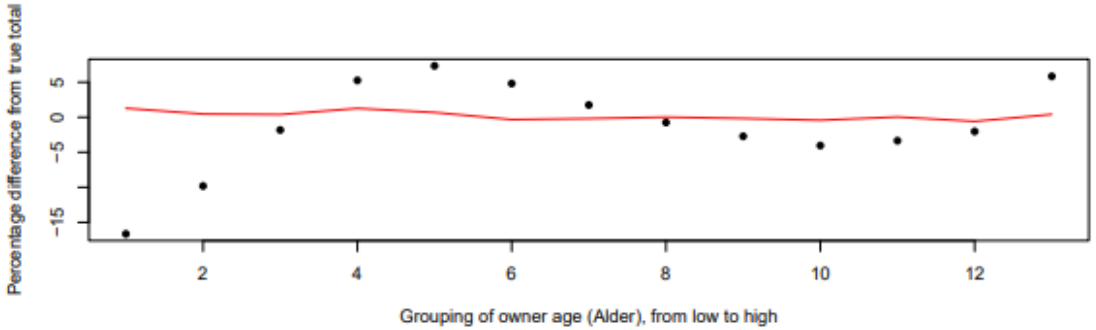
Fixed Predictors	Candidate Predictors
Aldermod	DirImp
Leas	Status
Drivmedel	log(Effekt1)
log(TjVikt)	Alder
Kon	Aldersq
Fordonsslagsklass	Supermiljöbil
	Komgrupp_SALAR
	Kaross
	Interaction of log(Effekt1) and log(TjVikt)
	Interaction of Fordonsslagsklass and log(TjVikt)

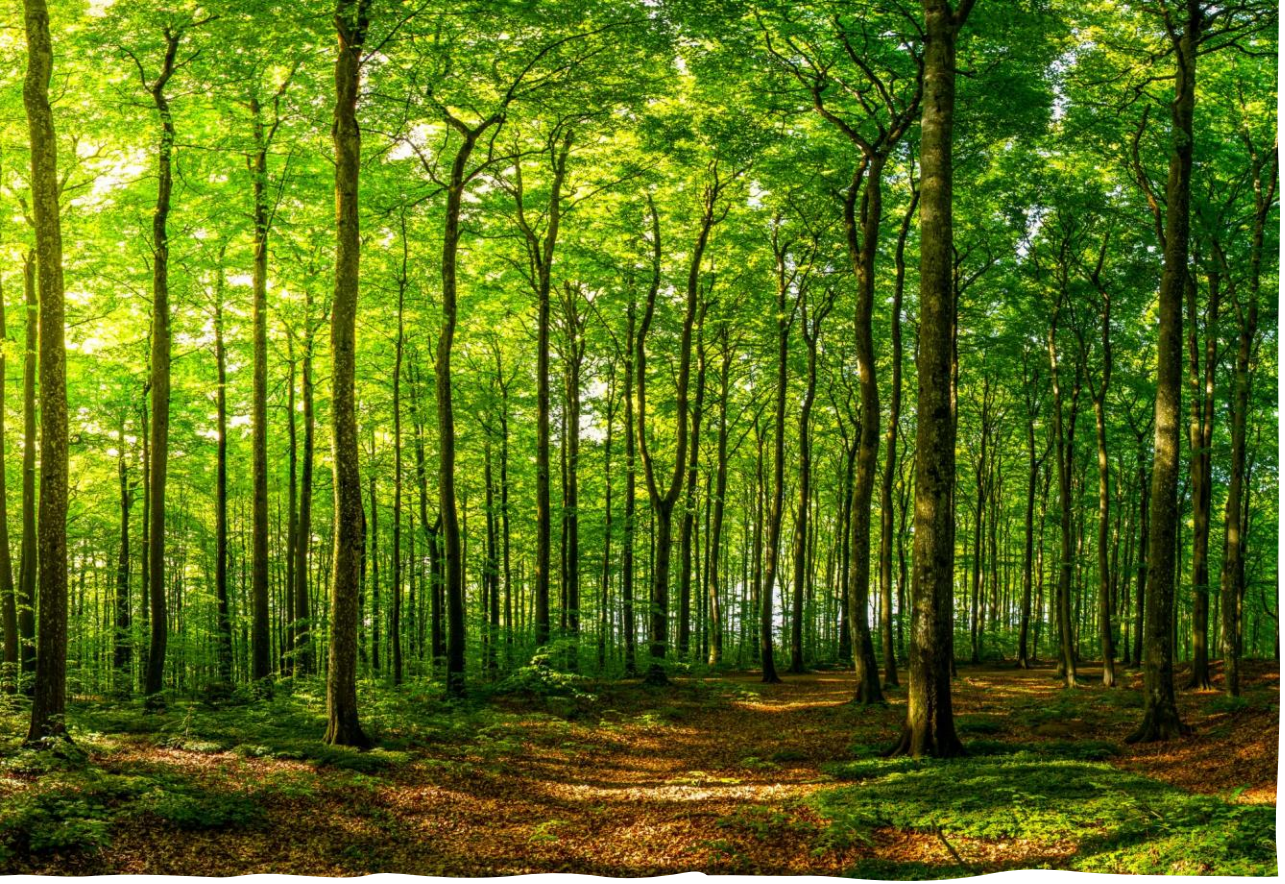
Table 3. Fixed and candidate predictors for initial variable selection. The possible combinations result in 576 models to evaluate.

Val av prediktorer, forts.









Random forest

- Bygger på regressionsträd
- Liknar på sätt och vis dagens metod, men automatiserat och optimerat

Resultat

Imputation model	Grand total	Relative difference to actual grand total (%)
Actual test driving distances	902 020 696	
Geometric mean imputations	759 826 766	-15.8
Arithmetic mean imputations	906 546 614	0.5
GLM imputations	903 283 801	0.1
Random forest imputations	904 649 307	0.3

Table 5. Grand total driving distance for the cars in the test data, and estimated grand totals based on the imputation methods.

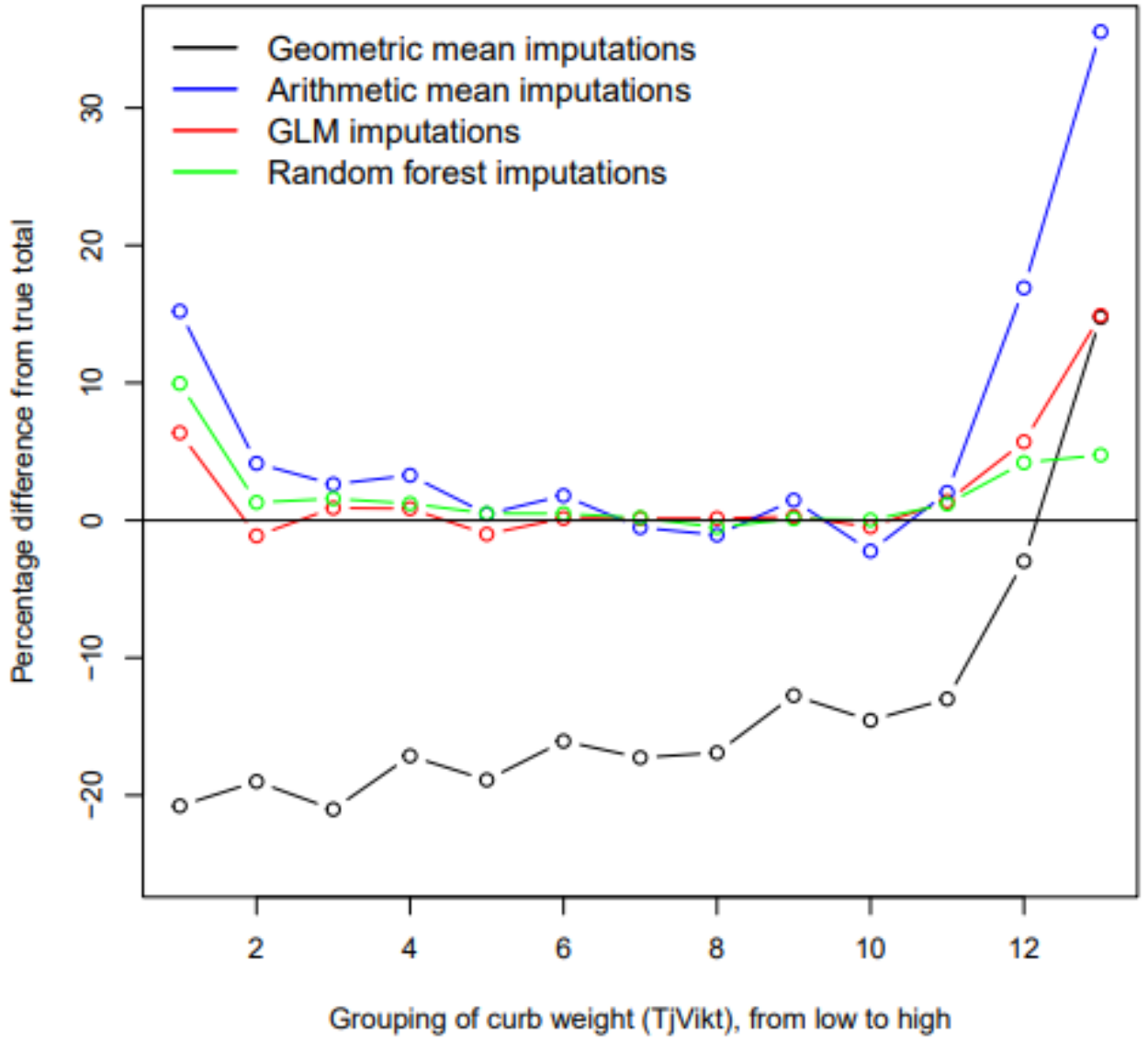
Imputation model	Test data RMSE
Geometric mean imputations	2.182098
Arithmetic mean imputations	2.102085
GLM imputations	2.052011
Random forest imputations	2.019789

Table 6. Test RMSE for the imputation methods.

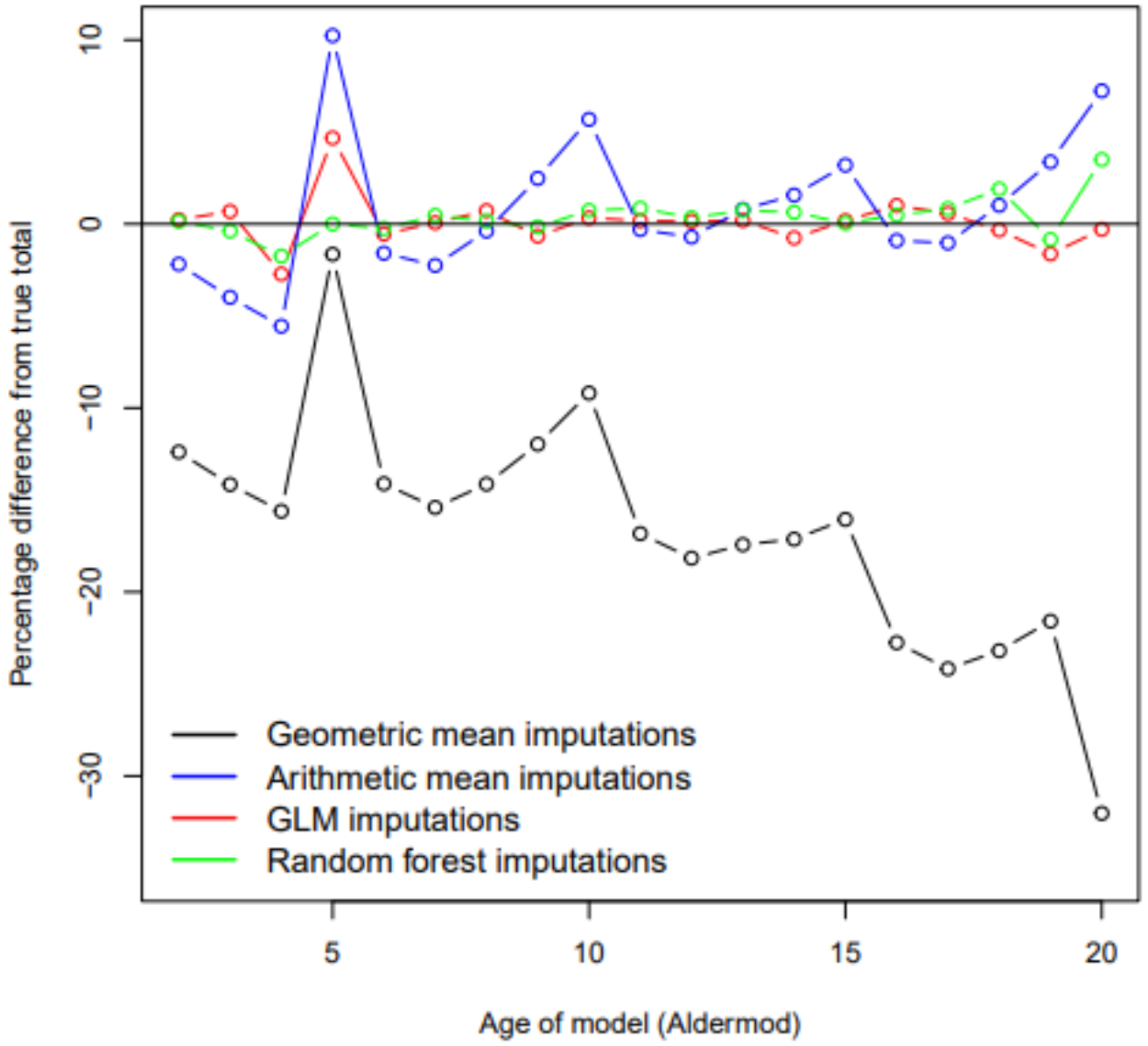
Imputation model	Combined evaluation metric
Geometric mean imputations	0.5154451
Arithmetic mean imputations	0.1008333
GLM imputations	0.02496143
Random forest imputations	0.02370557

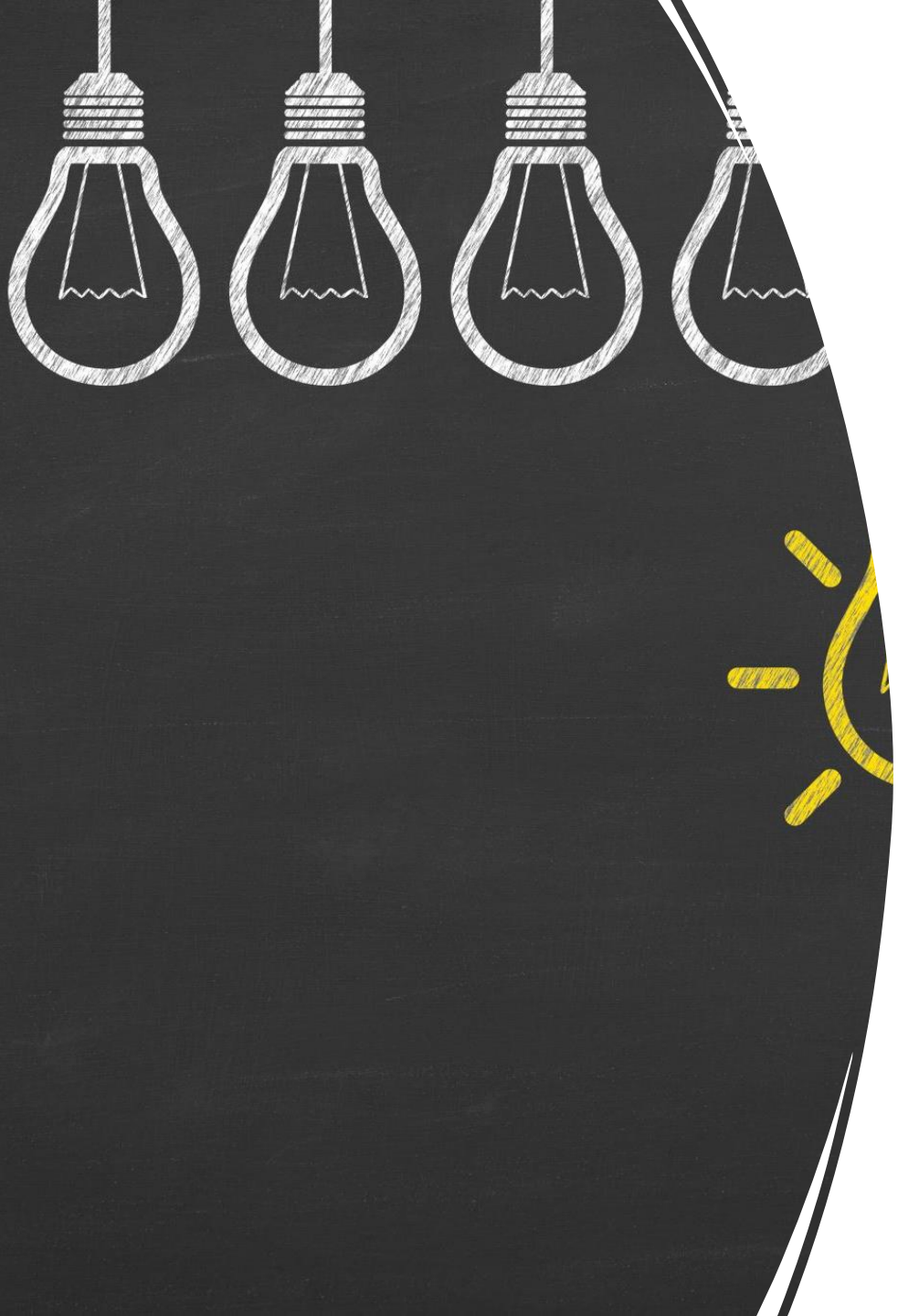
Table 7. Combined evaluation metric M, for the imputation methods.

Resultat


Geometric mean imputations
Arithmetic mean imputations
GLM imputations
Random forest imputations
Percentage difference from true total
30
20
10
0
−10
−20
2
4
6
8
10
12
Grouping of curb weight (TjVikt), from low to high

Resultat


Percentage difference from true total
10
0
−10
−20
−30
5
10
15
20
Age of model (Aldermod)
Geometric mean imputations
Arithmetic mean imputations
GLM imputations
Random forest imputations



Diskussion

- Resultaten bygger på antagandet om att bilarna som saknar värden gör detta slumpmässigt.
- Att gå från nuvarande metod till Random forest eller Tweedie GLM skulle båda vara bra alternativ.
  - GLM är kanske enklare att tolka
  - Random forest kanske enklare att använda
- Osäkerhet i skattningar
  - Multiple imputation
  - Inspel från SCB:s vetenskapliga råd: Se det som ett estimeringsproblem istället
- Framtida studier: Nya bilar

Frågor?