CORPUS PUBLISHERS

Advance Research in Sciences (ARS)

ISSN: 2837-5777

Volume 3, Issue 2, 2025

Article Information

Received date : 15 August, 2025

Published date: 01 September, 2025

*Corresponding author

C.A.M. van Berkel, Statistics Netherlands, Division of Data Services, Research and Innovation, P.O. Box 4481, 6401 CZ, Heerlen, the Netherlands

DOI: 10.54026/ARS/1033

Key Words

Democracy; travel time; cyclists; Root Mean Square Error; travel distance

Short Communication

Predicting Travel Time and Distance for Statistics Netherlands Interviewers with Machine Learning

J.W.M. Jansen, S. van der Doef and C.A.M. van Berkel*

Statistics Netherlands, Division of Data Services, Research and Innovation, Heerlen, the Netherlands

Abstract

The usual observation strategy of Statistics Netherlands surveys is via the Internet with follow-ups for nonrespondents by telephone or face-to-face interviews. The face-to-face interviewers work from their homes and receive sample addresses every month. The approach strategy includes a maximum of six visits, evenly spread over the days and parts of the days of the month in question. The interviewers schedule their visits themselves. It is not obvious in advance how much travel time and distance are needed to complete an interviewer's work package. This paper provides a model to estimate travel time and distance of these work packages, applying machine learning techniques to interviewers' travel declarations. According to Mean Absolute Percentage Error, the best way to predict the travel distance is to use a Support Vector Regression model with a log-log plus one transformation. The log-log transformation ensures homoscedasticity. The plus one, is for the cyclists who have zero declarations, because they do not get a km-reimbursement. The explanatory variables in the model are road distance, distance to the ideal interviewer's residence, radius of the circle containing the addresses, number of addresses, urbanity of the interviewer's residence, means of the interviewer's transport, region, province and month. In order to predict travel time, according to Mean Absolute Percentage Error, it is also best to use a Support Vector Regression, this time with a log-log transformation. Again, the log-log transformation is used to remove heteroscedasticity from the model. Plus, one is not necessary here, as travel time is always accounted for. The same explanatory variables are used as in the model for travel distance, with road distance replaced by road time. Both road distance and time are determined by an offline server with routing.

Introduction

Statistics Netherlands provides reliable statistical information and data to produce insight into social issues, thus supporting the public debate, policy development and decision-making while contributing to prosperity, well-being and democracy, see www.CBS.nl. Various data sources are used for this purpose, including population registers, open data and survey data. Most surveys follow the mixed-mode strategy of Internet observation with follow-ups for nonrespondents via telephone or face-to-face interviews. This article is about a planning problem for face-to-face interviewers. The face-to-face interviewers work from their homes and receive sample addresses monthly. The approach strategy includes a maximum of six visits, evenly spread over the days and parts of the days of the month in question. As the interviewers schedule their visits themselves, and it is unknown whether sampled people are at home and are willing to participate in the survey, it is difficult to make statements in advance about travel time and distance needed for the monthly work. This complicates the match of required and available interviewer capacity. In this research several machine learning algorithms are examined to predict interviewers' travel declarations. Examples are Support Vector Regression [1], Linear Mixed Models [2]and Quantile Regression [3]. These algorithms are applied to the problem and the algorithm with the least mean absolute percentage error is selected. The article reads as follows. In section 2 the star distance model is presented. Section 3 provides the machine learning algorithms that have been studied. Section 4 deals with the selection method of explanatory variables and model parameters. Section 5 provides information on the validity and reliability of the chosen models. Sections 6 and 7 deal with limitations and recommendations.

Star distance model

A simple model for estimating travel distances is the so-called star distance model. The star distance of a work package is the sum of the straight-line distances, back and forth, between the interviewer's residence and the addresses in the work package. On an aggregate level, the star distance correlates with the actual distance driven. The sum of the star distances of all work packages multiplied by 1.04 turns out to be a good approximation of the total declared kilometers. On an individual level this does not apply. This can be solved by clustering star distances by size where the variance of declared kilometers within each cluster may not exceed a prescribed maximum. For a work package in a cluster, the median of previously declared kilometers of work packages in this cluster is taken. In the next paragraph the star distance model is compared with other machine learning models.

Machine learning model

The data from 2018 and 2019 were used to develop a new model. These data are randomly divided into a train set and a test set. A 10-fold cross validation was used to divide the data into 90% train data and 10% test data. Explanatory variables were selected with backward selection [4]. The following assumptions were made during the modelling process.

a) All interviewers fill in claims truthfully for the month in which the travel hours and kilometers were actually incurred.
b) All interviewers work a full month. If an interviewer cannot work the whole month, for example due to illness, the data of this interviewer are not included in the dataset.
c) The interviewers did not move throughout the period.
d) The interviewers work in the region where they live.
e) The interviewers travel by car or by bicycle. They can walk between two addresses that are close to each other.
f) If the interviewer walks or cycles, only the travel time counts. If the interviewer drives a car, the travel time and distance count.

g) Time and distance for parking the car are included in the travel time and distance.
h) The location of an address is determined by its six-digit postal code.
i) The urbanity of an interviewer's working area is equal to the urbanity of the postal code of the interviewer's home address.
j) If an interviewer received addresses from a colleague during the month, this work package is not included in the dataset.

Consultation of team interviewing provided insight into variables related to the time and distance required by interviewers to perform their duty. These explanatory variables are included in Appendix A. Interviewers came up with some variables that are not known in advance, such as the number of working hours per day and number of working days per week. Therefore, these could not be included in the models.

Table A1: Explanatory variables for machine learning algorithms

Description	Name	Type
Sum of all times over road from interviewer's residence to sample address, per month.	Road Time	Continuous
Sum of all times over road from interviewer's residence to sample address, per month.	Road Distance	Continuous
Address density of the working area in which the interviewer works	Urbanity	Categorical
		1 = very strong urban
		2 = strong urban
		3 = moderate urban
		4 = little urban
		5 = not urban
Means of transport of the interviewer	Bicycle	Categorical
		1 = almost never
		2 = depending on circumstances
		3 = only or mostly
Region where the interviewer works	Region	Categorical
		1 = North
		2 = West
		3 = South
Province of interviewer's residence	Province	Categorical:
		The twelve Dutch provinces
The month in which the interviewer worked	Month	Categorical:
		The twelve months of the year
The number of addresses assigned to the interviewer at the beginning of the month	Number Addresses	Integer
Radius of circle overlapping the addresses	Radius	Continuous
Sum of straight-line distances from interviewer's residence to sample addresses.	DistanceGeo	Continuous
Hemispheric distance from the ideal residential location of the interviewer for the addresses assigned in the month.	DistanceMid	Continuous

The values of the explanatory variables Urbanity, Bicycle, Region, Province, Month and Number-Addresses are taken from the Survey administration system. Values of the remaining variables of Table A1 are calculated using an offline server. In this way, routes can be determined free of charge and quickly, about 38 routes per second. Using a scraper is also an option. However, this involves costs and is more time-consuming. The offline server used in this research uses the map and routing files from [5] downloaded from [6] more than one of the variables Road Distance, Road Time and Distance Geo can be included in the models because of the high multicollinearity (VIF > 30) between these explanatory variables. At the end of June, the interviewers receive sample addresses for July and August. They are free to visit these addresses throughout the period July-August. In this way, the chance of contact during the holiday period is increased. When using the data, the two-months sample for July and August should be taken into account. The interviewers do account for their visits separately for July and August. It was decided to divide the addresses, travel distances and times in proportion to the number of visits per month over July and August. This resulted in the most realistic data without outliers.

To assess the forecast accuracy of the models, the Mean Absolute Percentage Error (MAPE) has been used.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|A_i - F_i|}{A_i} \times 100\%,$$

where A_i is the actual value, F_i is the forecast value, and n is the number of observations. So, the performance indicator MAPE is the mean of the absolute percentage errors of forecasts. The MAPE for the star distance model equals 29,04%. To beat this value, the following machine learning algorithms were applied in modelling travel distance and travel time:

a) Support Vector Regression (SVR) [1]
b) Linear Mixed Models (LMM) [2]
c) Quantile Regression (QR) [3]
d) Least Absolute Deviation Regression (LAD Regression) [7]
e) Linear Regression [8]
f) Decision Tree [9]
g) Elastic Net [10]

For each model, the selection of explanatory variables was made via backward selection in R. For predicting both travel distance and travel time, SVR performs best with MAPE=17.5% and 19.0% respectively. Worst are the predictions of the decision tree with MAPE=23.0% for travel distance and 22.8% for travel time. The performance and selection of explanatory variables per model for travel distance and travel time are presented in Tables B1, Tables B2 of Appendix B.

Table B1: Model, expanarory variables and MAPE for travel distance.

Model	Expanarory variables	MAPE
Support Vector Regression	Number Addresses, DistanceMid, Radius, Road Time, Region, Province, Month	17,49%
Quantile Regression	DistanceMid, Radius, Road Distance, U rbanity, Region, Province, Month	20,65%
Least Absolute Deviation Regression	Number Addresses, DistanceMid, Radius, Road Distance, Urbanity, Region, Month	20.81%
Linear Mixed Models	Fixed: Number Addresses, DistanceMid, Radius, Road Distance Random (in R notation): 1|Bicycle: Urbanity, Region Province	21,86%
Linear Regression	Number Addresses, DistanceMid, Radius, Road Time, Urbanity, Province, Month	22,09%
Elastic Net	Number Addresses, DistanceMid, Radius, Road Distance, Urbanity, Region, Province	22,12%
Decision Tree: 3-layers	Road Distance, Bicycle, Region	23,03%

Table B2: Model, expanarory variables and MAPE for travel time.

Model	Explanatory variables	MAPE
Support Vector Regression	Number Addresses, DistanceMid, Road Time, Bicycle, Urbanity, Province	19,01%
Quantile Regression	Number Addresses, DistanceMid, Radius, Road Time, Bicycle, Urbanity, Region, Province, Month	19,48%
Least Absolute Deviation Regression	Number Addresses, DistanceMid, Road Time, Bicycle, Urbanity, Region, Province, Month	20,79%
Linear Regression	Number Addresses, DistanceMid, Radius, Road Time, Bicycle, Urbanity, Province, Month	22.01%
Linear Mixed Models	Fixed: Number Addresses, DistanceMid, Radius, Road Time Random (R-notation): Bicycle: Urbanity, Province, Month	22,27%
Elastic Net	Number Addresses, DistanceMid, Radius, Road Time, Urbanity, Province, Month	22,32%
Decision Tree: 3-layers	Number Addresses, Road Time, Province	22,78%

Tables B1, Tables B2 show that SVR and QR perform best. For these models, the residuals were examined for heteroscedasticity. For both models, the residuals were found to be heteroscedastic. By applying Weighted Least Squares [11] this problem could not be solved. By using a log-log transformation [12,13] this could be done for the prediction of travel time. The performance of the models for travel time with log-log transformation is shown in Table C1 of Appendix C.

Table C1: Model, expanarory variables and MAPE for travel time, with log-log transformation

Model	Explanatory variables	MAPE
Support Vector Regression	Log (Number Addresses), log (DistanceMid), log (Road Time), Bicycle, Urbanity, Province	19,71%
Quantile Regression	Log (DistanceMid), log (Road Time), Bicycle, Urbanity, Province, Month	20,18%
Elastic Net	Log (DistanceMid), log (Road Time), Bicycle, Urbanity, Province, Region, Month	20,74%
Linear Mixed Models	Fixed: log (Distance Mid), log (Road Time) Random: Urbanity, Province, Region, Month	20,86%
Linear Regression	Log (DistanceMid), log (Road Time), Urbanity, Province, Region, Month	20,89%
Least Absolute Deviation Regression	Log (Number Addresses), log (DistanceMid), log (Road Time), Urbanity, Province, Month	21,26%
Decision Tree: 3-layers	log (Number Addresses), log (Road Time)	23,97%

Because there were zero values in travel distance, no log-log transformation could take place. These zero values are due to cyclists who do not receive a financial compensation for their travel distance. Therefore, it is necessary to look at the best way to deal with cyclists in the model for travel distance. Three possibilities were examined:

A. Remove cyclists from the training set.
B. Model cyclists and motorists separately.
C. Apply a log-log+1 transformation [12]

Removing cyclists from the training data results in over-prediction of travel distance for cyclists. This causes outliers in the residuals. Splitting the models is a better option. But according to MAPE, the log-log+1 transformation is best. The transformation not only resulted in homoscedasticity, but also in a reduction of the MAPE. The performance of the models for travel distance with log-log transformation is shown in Table D1 of Appendix D.

Table D1: Model, expanarory variables and MAPE for travel distance, with log-log+1 transformation.

Model	Explanatory variables	MAPE
Support Vector Regression	log (Number Addresses), log (Road Distance), log (Radius), Province, Urbanity, Bicycle	16,15%
Least Absolute Deviation Regression	log (Number Addresses), log (Road Distance), log (DistanceMid), Province, Region, Bicycle, Month	17,12%
Quantile Regression	log (Number Addresses), log (Road Distance), log (Radius), log (DistanceMid), region	18,02%
Linear Mixed Models	Fixed: log (Number Addresses), log (Road Distance) Random (in R-notation): Region, Bicycle: Urbanity	19,87%
Elastic Net	log (Number Addresses), log (Road Distance), log (Radius), log (DistanceMid), Bicycle	20,17%
Linear Regression	log (Number Addresses), log (Road Distance), Region, Bicycle	20,42%
Decision Tree: 3-layers	log (Road Distance), log (DistanceMid), Province, Bicycle, Month	26,02%

For predicting both travel distance and travel time, SVR performs better than QR. In the next section, the selection of variables and parameters in SVR is examined in more detail.

Selecting variables and parameters

In the SVR modelling process, the explanatory variables were selected by Recursive Feature Elimination (RFE) [14] RFE uses Root Mean Square Error (RMSE) as the performance indicator. In Figures 1,2. The RMSE is plotted against the number of variables n in the prediction model for travel distance and travel time, respectively. It follows that the RMSE does not decrease much for n greater than 10. After selecting the explanatory variables, a margin of tolerance and a cost parameter were inimized by applying grid search. Finally, the MAPE for predicting travel distance was 17.65% and for travel time 19.38%. So, the accuracy of both SVR models is greater than the accuracy of the star distance model.

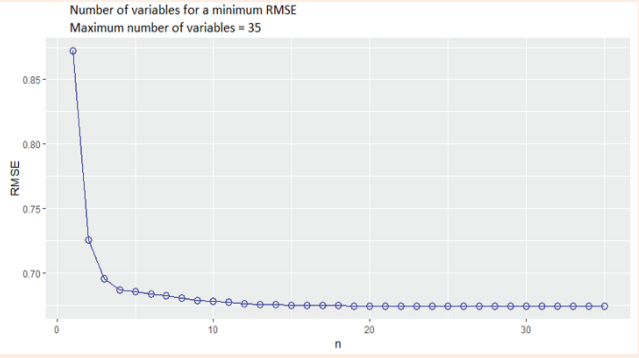

Figure 1: RFE for travel distance.

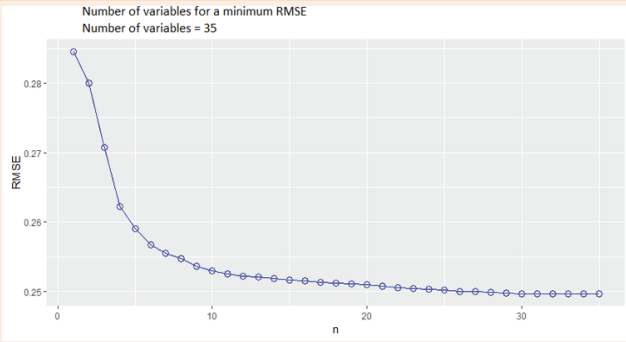

Figure 2: RFE for travel time.

Validity and reliability

To evaluate the validity and reliability of the SVR model, k-fold cross-validation [15] and sensitivity analysis were performed.

The results of the K-fold cross-validation are shown in Figures 3,4. For k=2, 3, ..., 10 the dataset was randomly divided into k groups. For each k, each unique group was used once as a test set and the remaining groups as a training set. The model was fitted to the training set and evaluated with the MAPE on the test set. These are the blue points in the figures. The red points are the averages of the blue points per k.

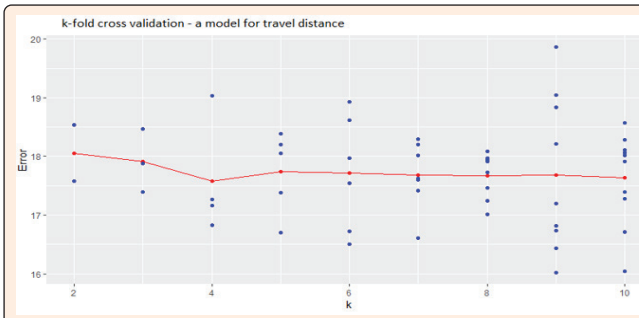

Figure 3: K-fold cross-validation for travel distance.

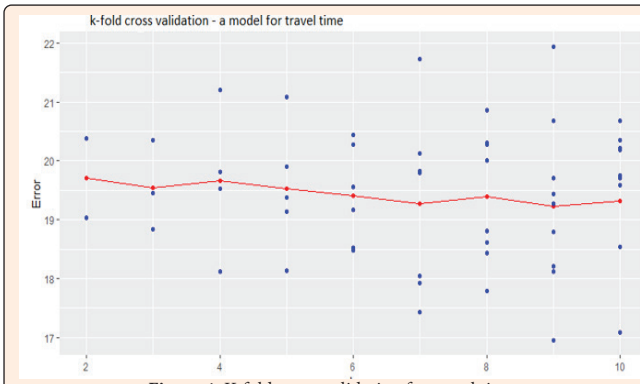

Figure 4: K-fold cross-validation for travel time.

It follows from the red points of Figures 3, 4 that the average performance of the models is fairly stable and that the spread of the blue points is fairly small. It can therefore be concluded that the models are valid [15]. By performing sensitivity analysis, it can be seen whether the models provide realistic forecasts [16,17]. Four scenarios were considered:

A. Regular situation.
B. Fewer interviewers are available. As a result, the interviewers get more addresses.
C. More interviewers are available. As a result, the interviewers get fewer addresses.
D. More interviewers go by bicycle. The motorists with the smallest radius of addresses have been changed to cyclists.

For each scenario, a mild and an extreme situation were considered. The models did not predict any negative values and did not predict very high values either. Despite extrapolation, the models performed reliably. The influence of the variables on the predicted value is shown in Figures 5, 6. Here 'value' equals the product of the coefficient associated with the variable and the support vector from the SVM model. Variables with negative influence or close to zero are not meaningful to include in the model.

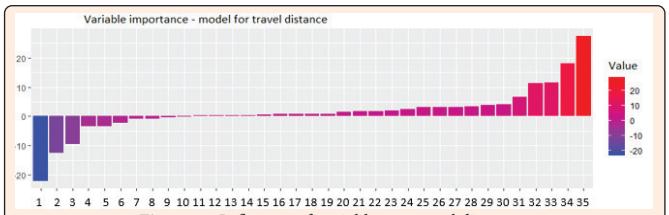

Figure 5: Influence of variables on travel distance.

The five most influential variables on travel distance are: log (Road Distance), log (Number Addresses), log (Radius), Cyclist: almost never, Province of North Holland.

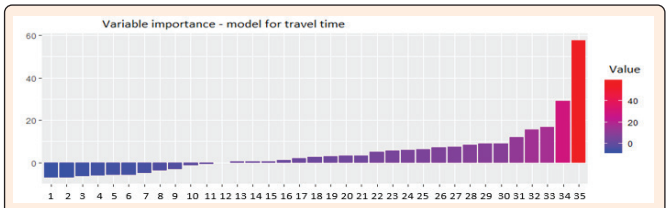

Figure 6: Influence of variables on travel time.

The five most influential variables on travel time are: log (Number Addresses), log (Road Distance), Province of Groningen, log (Distance Mid), Province of Utrecht. The influence of most variables is as expected. However, there are also variables that are not as expected. For example, it was thought beforehand that the province of North-Holland would have a negative influence, and Groningen would have a positive influence, because the addresses in North-Holland are closer to each other than in Groningen. However, when this was examined in more detail, it emerged that the interviewers in Groningen live close to their work area, while the interviewers in North-Holland live far from their work area. As a result, the interviewers in Groningen are expected to travel less than those in North Holland. As the models have been shown to be valid and reliable, they can be used to predict travel distance and travel time per interviewer.

Limitations

The research has limitations. This is due to the assumptions and choices made in the research. For example, the model is based on data in the Netherlands. The results cannot simply be copied for interviewing in another country. The method can be adopted and the models must be trained with input data from relevant interviewers. In addition, only declared distances and times were available as provided by the interviewers. Whether these data correspond to the actual values was not checked. Further, the dataset was unbalanced. It was not possible to obtain a balanced dataset, because there are regions without cyclists. Nevertheless, as we have seen, cyclists had to be included in the model. Finally, there are numerous possible approaches to solving prediction problems. These include cleansing the data, collecting the data, choosing and devising explanatory variables, and choosing machine learning algorithms.

Recommendations

Since the SVR model performs better than the star distance model, it is recommended to use the SVR model.

It is efficient to use an offline server to determine the routes. Using a scraper is more expensive and takes more time to determine all the routes.

The interviewers should submit their declarations accurately and on time to ensure a reliable data set.

References

1. Sayad S (ndb) (2020) Support vector machine-regression (SVR)
2. UCLA (nd) (2020) Introduction to linear mixed models
3. University of Virginia (nd) (2020) Getting started with quantile regression
4. NCSS (nd) (2020) Stepwise regression
5. OpenStreetMap (nd) (2020) OpenStreetMap
6. Geofabric GEOFABRIK (2020)
7. Koenker R, Basset G (1978) Regression quantiles in the econometric society (Red.) Econometrica pp. 33-50.
8. Zou KH, Tuncali K, Silverman SG (2003) Correlation and simple linear regression. Radiology 227(3): 617-622.
9. Sayad S (nda) (2020) Decision Tree-Regression
10. Wieringen WV (nd) (2020)Lasso regression
11. Frost J Heteroscedasticity in regression analysis.
12. Sahai H, Ageel MI (2012) Transformations to correct lack of homoscedasticity. In H. Sahai, M.I. Ageel (Red)The Analysis of Variance: Fixed, Random and Mixed Models pp. 110-113.
13. Benoit K (2011) Linear regression models with logarithmic transformations
14. Kuhn M (2019) Recursive feature elimination. In M. Kuhn (Red)The caret Package.
15. Brownlee J (2023) A gentle introduction to k-fold cross-validation
16. Institute of Business Forecasting & Planning (IBF)(nd) (2020) MAPE (Mean Absolute Percentage Error)
17. Swamidass PM (2000) MEAN ABSOLUTE PERCENTAGE ERROR (MAPE). In P.M. Swamidass (Red) Encyclopedia of Production and Manufacturing Management 30