Discussion Paper

Multilevel time series modeling of mobility trends - Final Report

Harm Jan Boonstra, Jan van den Brakel and Sumonkanti Das

October 30, 2019

Abstract

This report describes the time series models developed for the mobility trend estimation project carried out by Statistics Netherlands in collaboration with KiM/Rijkswaterstaat. First, direct estimates along with standard error estimates are obtained for each year in the period 1999-2017 from the microdata of the Dutch Travel Survey for a detailed cross-classification by person characteristics sex and age class and trip leg characteristics mode and purpose. Consequently, these direct estimates are smoothed by modeling them using multilevel time series models that account for influential outliers as well as for the redesigns of the survey within the time span considered. Two target variables are modeled in this way: the number of trip legs per person per day and the distance traveled per trip leg. The models are specified in a hierarchical Bayesian framework and estimated using a Markov Chain Monte Carlo simulation method. From the model outputs smooth trend estimates can be computed at various aggregation levels for the mean number of trip legs per person per day and the mean distance traveled per trip leg, as well as for derived quantities such as the mean distance per person per day. We discuss the model building and evaluation processes as well as the results based on the fitted models.

1 Introduction

The Dutch Travel Survey (DTS) is a long-standing annual survey on mobility of residents of the Netherlands. It is carried out by Statistics Netherlands (CBS) and important users of the data are Rijkswaterstaat and The Netherlands Institute for Transport Policy Analysis (KiM, Kennisinstituut voor Mobiliteitsbeleid), both part of the Ministry of Infrastructure and Water Management.

Since 1985, the DTS survey has undergone several redesigns. The redesigns in 1999, 2004 and 2010 caused major discontinuities in the time series of estimates on mobility. In 2004 the design actually remained largely unchanged, but it was transferred to another agency for implementation, causing several changes in the observed series. For brevity, however, we will mostly also refer to this transition as a 'redesign'.

For users of mobility estimates the changes due to redesigns are very inconvenient as they hamper the temporal comparability. For the redesign of 1999 direct information was available on the sizes of the discontinuities, based on a parallel conducted pilot study. This has been used to correct the series of estimates prior to 1999 to the level of the estimates under the new design. For the redesigns of 2004 and 2010 such parallel studies have not been carried out, so in order to estimate the discontinuities a time series model is needed, see van den Brakel et al. (2017). The time series models developed in the current trend estimation project aim to account for the discontinuities due to the redesigns, such that reliable series of trend estimates are obtained with good comparability over time. For the latest redesign of the DTS in 2018 these considerations also apply, but accounting for the resulting discontinuities will be handled in a future project, as for this moment only a single year's data is available under the latest design, which is not enough for reliable estimation of the discontinuities in the absence of a period of parallel conducted surveys under both designs.

Another issue that is addressed in the trend estimation project is due to the fact that estimates are desired for a breakdown into many domains, meaning that for each domain determined by both person characteristics (sex and age class) and trip characteristics (purpose and transportation mode) reliable time series of estimates are to be produced. So not only the discontinuities in each of these series should be accounted for, but in addition the amount of data directly relevant to an estimation domain in a specific year is often so small that direct estimates are very noisy and unreliable. The time series of such direct estimates display a lot of volatility caused by the large variances. The time series models developed are able to reduce the noise and obtain smoother series of estimates by 'borrowing strength' over time as well as over multiple domains. Here the 'borrowing of strength' over domains is brought about by using multilevel time series models with random effects for several levels defining the domains. Within the field of official statistics, the framework of using models to smooth estimates over domains of interest is known as small area estimation, see Rao and Molina (2015) for an overview.

The overall purpose of the mobility trends project has been described by the initiators of the project (KiM, Rijkswaterstaat, CBS) as 'Development of a statistical methodology that can derive reliable trend estimates from OVG-MON-OViN-ODiN sample data for the most prevalent mobility data and that deals in a robust way with discontinuities due to redesigns of the survey process and sample noise.' Here OVG, MON, OViN and ODiN refer to the various names used for the DTS during periods with different survey designs. To achieve the purpose as described, time series multilevel models are employed to fit the input data, consisting of direct estimates and estimated standard errors compiled from the DTS survey data. The resulting trend estimates are used by KiM for example in their publication 'Mobiliteitsbeeld' containing actual figures, trends, and expectations about mobility in the Netherlands. The trend estimates will also be published on Statistics Netherlands' publication database StatLine along with the regular annual output based on the DTS.

The two target variables that are modeled using time series multilevel models are:

– number of trip legs per person per day (pppd)
– distance per trip leg (in hectometers)

A trip for a certain purpose may consist of several trip legs characterized by different transportation modes. Estimates are computed for domains defined by a cross-classification of some or all of the following classification variables:

– sex (male, female)
– ageclass (0-5, 6-11, 12-17, 18-29, 30-39, 40-49, 50-59, 60-69, 70+)
– purpose (work, shopping, education, other)
– mode (car driver, car passenger, train, BTM (bus/tram/metro), cycling, walking, other)

The purpose category 'other' includes family and social visits, recreation, as well as business visits. Combining these categories already reduces some discontinuities associated with the 2004 and 2010 redesigns, see e.g. Willems and van den Brakel (2015). The mode category 'other' includes for example motorcycle, boat and skates. The time series multilevel models are defined at the most detailed level, corresponding to the full cross-classification of sex, ageclass, purpose and mode, giving rise to $2 \times 9 \times 4 \times 7 = 504$ estimates for a particular year, although a few of them such as car-driving children are structurally zero. We use the series of direct estimates starting in 1999 to fit the time series multilevel models. The direct estimates of 1999 through 2017 constitute a time series of 19 years, which appears to be sufficient to fit the time series multilevel models.

In Bollineni-Balabay et al. (2017) both structural time series models and multilevel time series models were used to estimate trends for mobility (in particular distance travelled) by purpose and mode. There it was found that the differences between the two modeling frameworks were generally small. In Boonstra and van den Brakel (2016) it was found that multilevel time series models in a hierarchical Bayesian formulation have some advantages in terms of flexibility and computational efficiency. Therefore we chose to use the Bayesian multilevel model formulation for this application. This formulation is also closer to an earlier approach used by KiM (Wüst, 2017).

In Bollineni-Balabay et al. (2017) discontinuities were modeled as fixed effects. In the current project a different classification into domains is used, including breakdowns by sex and age, and modeling all discontinuities as fixed effects would now result in overestimated discontinuities. To reduce the risk of overestimated discontinuities and overfitting in general, we model many effects including discontinuities associated with the design transitions as random effects instead. In particular, a regularization method that employs non-normally distributed random effects is used that suppresses noisy model coefficients and at the same time allows large effects that are sufficiently supported by the data. Outliers in the input direct estimates are also modeled, either by adopting a sampling distribution with broader than normal tails or by modeling them explicitly as additional random effects, which are subsequently removed from the trend estimates.

The remainder of this report is organized as follows. Section 2 describes the data sources used including a brief overview of the different redesigns the DTS has undergone. In Section 3 the computation of direct estimates and variance estimates from the DTS survey data is discussed, along with transformations of direct estimates and the Generalized Variance Function approach for smoothing the variance estimates, which both improve model fitting. Section 4 describes the hierarchical Bayesian time series multilevel modeling framework. The models selected for trip legs and distance are presented in Section 5, along with a brief discussion of the model building process. Section 6 provides a discussion of the trend estimates based on the estimated models, and model evaluation results are given in Section 7. The paper concludes with a discussion in Section 8, and figures of selected results are displayed in the appendix.

2 Data sources

The DTS is an annual survey that attempts to measure the travel behaviour of the Dutch population. Each year, a sample is drawn with sampling units being defined either as households (before 2010), or persons (since 2010). The variables of interest considered in this study are the number of trip legs and the distance traveled. Direct estimates for these quantities can be obtained using the survey weights that are computed for each year's response data. The survey weights reduce the bias due to non-response, and the estimates based on them correspond to the general regression (GREG) estimator (see e.g. Särndal et al. (1992)).

The DTS started in 1978, and originally was known under the (Dutch) name Onderzoek Verplaatsingsgedrag (OVG). It started off as a face-to-face household survey where every household member 12 years in age or older was asked to report his/her mobility for two days. In 1985 the first large redesign took place. Interview modes changed to telephone and postal, and respondents reported their mobility of one day. This redesign led to discontinuities in the annual series of some of the statistics based on OVG. In the period until 2003 the survey was conducted by Statistics Netherlands. In 1994 the sample size of the DTS was substantially increased and from that year children under 12 years old have also been included in the surveyed population of interest. In 1999, the DTS went through the second major redesign that featured some response motivation and follow-up measures. In preparation to this redesign a pilot based on the new design was conducted in 1998 in parallel with the survey under the old design. Based on the parallel surveys, correction factors were computed to correct the 1985-1998 OVG to the level of the new OVG. In 2004, the data collection for the survey was transferred to another agency. The survey design remained largely unchanged except for smaller sample sizes and some methodological changes. This 2004 transition also gave rise to discontinuities in some of the series, notably those disaggregated by purpose. The DTS during the period from 2004 until the next major redesign in 2010 is referred to as MON (Mobiliteitsonderzoek Nederland). Since 2010 the DTS has been conducted by Statistics Netherlands again. In 2010 the survey changed to a person survey, and a sequential mixed mode design with face-to-face, telephone and web modes was established. This changeover led to sizeable discontinuities in many series. The years 2010 to 2017 constitute the OViN (Onderzoek Verplaatsingen in Nederland) period of the DTS. Finally, starting from 2018 another design is in place, named ODiN (Onderweg in Nederland), in which only the web mode remains, and substantial changes to the questionnaire have been carried through. Another change is that in ODiN children between 0 and 5 are no longer observed. For more information on the history of the DTS and the changes made by the redesigns, we refer to Konen and Molnár (2007), Molnár (2007) and Willems and van den Brakel (2015).

Even though the DTS dates back to 1978, for the present project we use DTS data starting from 1999, the first year of the new OVG survey. This turns out to be sufficient for the purpose of obtaining reliable trends over the last 15 years or so. We have considered using also data from the period 1994-1998 but eventually decided not to, since it would not outweigh the additional effort required for modeling the large discontinuities arising from the 1999 redesign. The total time span considered is $T = 17$ years, covering the years 1999-2017. This corresponds to the years of the (new) OVG design, the MON design, and the OViN design. Currently, it is too early to use the first ODiN data available.

The DTS considers only mobility within the Netherlands. In this project we are primarily interested in regular mobility, which means the mobility excluding holiday mobility (both domestic and abroad) and professional transportation mobility. Therefore, trips with purpose holiday and professional transportation are removed from the survey data. Unfortunately, this selection cannot be carried out completely consistently over the years, and the extent to which such non-regular mobility can be removed may give rise to some discontinuities. In any case, professional transportation trips are present in the DTS datasets in all years, and we have removed these trips. For the OViN years the data contain a small number of trips for children under the age of 12 with purpose 'work', and we have changed this to purpose 'other'. Flight trips are also removed from the data, because they lead to some unstable estimates of distance traveled for mode 'other'. Also, in ODiN flights are no longer reported. For now, data about all age classes are used. In the future it might be decided to drop the data for young children aged 0-5 years, since ODiN no longer observes this category.

It is considered important that mobility trend estimates based on the DTS are in line with external data sources on mobility. Such information has been used in a plausibility analysis, and a few external sources have also been considered for use as auxiliary information in the time series models used for trend estimation. One such source is a time series of annual total passenger train kilometers based on passenger surveys run by the Dutch railways NS. These series also include data on train rides by other private companies active in the Netherlands. Another relevant data source is the annual series of road intensities, compiled by Statistics Netherlands from road induction loop data. In addition, there is a time series of registered kilometers driven by cars from Nationale Autopas (NAP). This series includes kilometers driven abroad but nevertheless it is a potential covariate in the time series models based on the DTS. Finally, annual figures for a set of weather characteristics coming from the Royal Netherlands Meteorological Institute (KNMI) have been used in the model development.

3 Direct estimates

Basing a time series model for mobility trends directly on the microdata from all years would require a very complex model that must account for non-response, different aggregation levels of interest, discontinuities, time trends, etc, all at once, which would be computationally intractable. Instead we follow a two-step estimation procedure often used in small area estimation. In the first step, estimates and variance estimates of the target variables are obtained directly from each year's microdata, at the aggregation level of interest. Here we can make use of the existing survey weights, accounting for the sampling design and non-response. In the second step these so-called 'direct estimates' serve as input for a time series model, which can be used to compute smoothed estimates of mobility accounting for possible discontinuities caused by the redesigns. This section outlines the computation of the direct estimates from the OVG-MON-OViN survey data. For additional details, see Boonstra et al. (2018).

The direct estimates are computed for all years from 1999 until 2017 for trip legs pppd and distance per trip leg. This results in two tables of 504 series of direct estimates at the most detailed breakdown level considered.

3.1 Point estimates

Point estimates are readily computed using the existing survey weights. First consider the number of trip legs, and let r_i denote the number of trip legs reported by person i for the surveyed day. The average number of trip legs pppd is then estimated by

$$\hat{R} = \frac{\sum_{i \in s} w_i f_i r_i}{\sum_{i \in s} w_i f_i}, \qquad (1)$$

where the sums run over respondents, w_i are person weights satisfying $\sum_{i \in s} w_i = N$ with N the total population size, and f_i is a so-called vacation factor. The latter take values slightly less than 1, and are used to account for vacation mobility. The vacation factors are based on estimates obtained from the CVO (Continu Vakantieonderzoek) survey. They can be derived from the 'trip weights' v_i as

$$f_i = \frac{v_i}{D w_i}, \qquad (2)$$

where D is the number of days in a year. The estimates (1) can be written more compactly in terms of the trip weights as

$$\hat{R} = \frac{\sum_{i \in s} v_i r_i}{\sum_{i \in s} v_i}. \qquad (3)$$

The vacation factors have been used for official publications based on OVG, MON and OViN. For ODiN it is envisaged that the vacation factors will be integrated in the person weights w_i by using estimated population totals from CVO in the weighting scheme directly.

For the second target variable of interest, distance, we estimate the average distance per trip leg by

$$\hat{A} = \frac{\sum_{i\in s} w_i f_i a_i}{\sum_{i\in s} w_i f_i r_i} = \frac{\sum_{i\in s} v_i a_i}{\sum_{i\in s} v_i r_i}, \tag{4}$$

where a_i is the total distance for person i for all trip legs.

For estimates by mode and/or purpose, each particular category defines specific variables r and a referring only to the trip legs in that category, so that equations (1) and (4) still apply. For (further) subdivisions with regard to the person characteristics sex and ageclass, it is convenient to introduce a dummy variable δ_i for each combination of sex and ageclass, being 1 if person i belongs to this group and 0 otherwise, and then write instead of (1) and (4),

$$\begin{aligned}\hat{R} &= \frac{\sum_{i\in s} w_i f_i \delta_i r_i}{\sum_{i\in s} w_i f_i \delta_i}, \\ \hat{A} &= \frac{\sum_{i\in s} w_i f_i \delta_i a_i}{\sum_{i\in s} w_i f_i \delta_i r_i}.\end{aligned} \tag{5}$$

By using δ_i also in the denominator of $\hat{R}$, we obtain estimates of the means per sex, ageclass combination. Note that the denominator of $\hat{R}$ does not depend on any selection of purpose or mode.

As mentioned in the Introduction, at the most detailed level, each target variable gives rise to a set of 504 estimates per year, corresponding to the full cross-classification of person characterics sex and age class and trip characteristics purpose and mode. Some of the 504 domains are, however, non-existent. We refer to these domains as structural zeros, since the number of trips in these domains is zero by definition. This concerns the following domains: age 0-5 and 6-11 in combination with mode car driver or purpose work and age 12-17 mode car driver before 2011. Starting from 2011 it is possible to drive a car from age 17, and this can be seen in the data. Distances per trip leg corresponding to structural zero trip legs are undefined, and therefore missing in the set of direct estimates. Other occasional zeros for trip legs and missings for distance per trip leg occur in some years for 'rare domains' such as education for the elderly. These accidental zeros and missings will be filled in by the predictions based on the time series models.

3.2 Variance estimates

For variance estimation we distinguish between person surveys (OViN) and household surveys (OVG, MON). For the latter, the household is the unit of sampling. Observations from persons from the same household cannot be regarded as independent. For example, distances traveled by young children and their parents are often correlated, depending on purpose and mode. Variance estimates should account for the dependence between persons clustered within households.

First write estimates (1) and (4) in the general form

$$\hat{Y} = \frac{\sum_{i\in s} w_i y_i}{\sum_{i\in s} w_i z_i}, \tag{6}$$

which is a ratio of two population total estimates based on person weights w_i. For the average number of trip legs pppd, $y_i = f_i r_i$ and $z_i = f_i$; for the average distance per trip leg, $y_i = f_i a_i$ and $z_i = f_i r_i$.

Basic estimates of the sampling variances of $\hat{Y}$ that ignore the variation of the weights, finite population corrections and the variance of the denominator, are given by

$$v_0(\hat{Y}) = \frac{1}{(\sum_{i\in s} w_i z_i)^2} \frac{N^2}{n} S^2(y)\,, \tag{7}$$

where n is the number of respondents, $S^2(y) = \frac{1}{n-1}\sum_{i\in s}(y_i - \bar{y})^2$ is the sample variance of y, with $\bar{y} = \frac{1}{n}\sum_{i\in s} y_i$ the sample mean of y.

These variance estimates are improved by taking into account (1) the variance of the denominator, (2) the variance inflation due to variation of the weights (Särndal et al., 1989), and (3) the variance reducing effect of some covariates used for stratification or weighting. The variance estimates incorporating all three improvements are computed as (see e.g. Särndal et al. (1992))

$$\begin{aligned} v(\hat{Y}) &= \frac{n}{\left(\sum_{i\in s} w_i z_i\right)^2} S^2(we)\,, \\ e_i &= e_i^y - \hat{Y} e_i^z\,, \\ e_i^y &= y_i - x_i' \hat{\beta}^y\,, \\ \hat{\beta}^y &= \left(\sum_{i\in s} x_i x_i'/u_i\right)^{-1} \sum_{i\in s} x_i y_i/u_i\,, \\ e_i^z &= z_i - x_i' \hat{\beta}^z\,, \\ \hat{\beta}^z &= \left(\sum_{i\in s} x_i x_i'/u_i\right)^{-1} \sum_{i\in s} x_i z_i/u_i\,. \end{aligned} \tag{8}$$

Here $S^2(we)$ is the sampling variance of $w_i e_i$, where e_i are generalized residuals, defined in terms of regression residuals e_i^y for y and e_i^z for z. The regressions are based on vectors of covariates x_i and a positive variance factor u_i. For the persons survey case we use $u_i = 1$.

For the regressions defining the residuals in (8), the following covariate model is used:

$$hhsize + province + sex * ageclass + urbanisation + month + weekday + fuel$$

in which $hhsize$ is the number of persons in a household, $ageclass$ is as defined in the Introduction, $urbanisation$ is the degree of urbanisation of the residential municipality in 5 classes, $month$ is the survey month, $weekday$ the day in the week the response refers to, and $fuel$ is the fuel type of the car used by the respondent in three classes: petrol, other or none if the respondent doesn't use a car. These covariates represent an important subset of variables that have been used for stratification and weighting of the survey data over the years.

The variance formula (8) can be used for any variables y and z in (6) so it applies to all estimates by any combination of trip characteristics purpose, mode and person characteristics sex and ageclass.

We have compared the simple variance estimates computed with (7) with the refined ones based on (8), and observed that the differences are mostly modest but not generally negligible. The most important refinement turns out to be the variance inflation due to the variation of weights. This clearly increases the variance estimates for domain estimates based on widely varying weights.

For the years before 2010, when the surveys were conducted as household surveys, the same formulas can be used, with the understanding that in that case the unit index i refers to households. In that case y_i, z_i refer to weighted household totals, the weights w_i to the average of the person weights within a household, and x_i to household totals of the weighting covariates. The regression variances u_i are taken equal to the household size, and n in (8) becomes the number of responding households. For details we refer to Boonstra et al. (2018). We also refer to Boonstra et al. (2018) for a complete set of plots of the direct estimates and their standard errors for trip legs and distance at all aggregation levels. The graphs show that for 'common domains' such as purpose work for age classes 30-39, 40-49, standard errors are stable and rather small. For rare domains the standard errors are on average much larger, and, like the point estimates, volatile and sometimes missing. Boonstra et al. (2018) also contains a short discussion about the covariances/correlations between the direct estimates within each year. Most of these cross-sectional correlations are small, but there are some large positive and negative ones. The largest positive correlations occur between estimates for modes that are often combined in a single trip, like walking and train, while most negative correlations occur between modes that are rarely combined such as car driver and cycling. Furthermore, in OVG/MON years, there are some more positive correlations induced by the household clustering, for example between estimates for parents (car driver) and children (car passenger) and purpose shopping or other. The effect of the cross-sectional correlations on the (standard errors of) the trend estimates was tested using a simple multilevel time series model and found there to be quite small. However, due to computational problems we have not been able to use the full correlation matrices of the input estimates in the finally selected time series models, although we expect to see only a small effect there as well.

3.3 Transformations of input series

The direct estimates and standard errors of the number of trip legs and the distances serve as input for the multilevel time series models used to obtain more smooth and robust trend series. We started out using the direct estimates and their (smoothed) standard errors directly as input to the multilevel time series models, but it does not fit the considered class of models well, and related problems showed up with convergence of the MCMC simulations. The problems lie with the widely varying scales of the data, i.e. the large (relative) differences between typical numbers of trip legs or distance among the domains. Another related issue is that the point estimates and standard errors display strong dependence which is not accounted for by the models. To remedy these problems, the input series are transformed.

Let $\hat{Y}_{it}$ denote the direct estimate for year t and domain i of the number of trip legs or distance. We have considered three different transformations for both target variables:

- logarithmic transformation: $\hat{Y}_{it} \to \log(\hat{Y}_{it} + \varepsilon)$, where ε is a small number necessary because some direct estimates for the number of trip legs are zero. Standard errors for the transformed data are approximated by Taylor linearisation: $se(\hat{Y}_{it}) \to se(\hat{Y}_{it})/(\hat{Y}_{it} + \varepsilon)$.
- square root transformation: $\hat{Y}_{it} \to \sqrt{\hat{Y}_{it}}$. A Taylor linearisation yields approximated standard errors $se(\hat{Y}_{it}) \to se(\hat{Y}_{it})/(2\sqrt{\hat{Y}_{it}})$. Note that these standard errors are

undefined for domains without observed trips (zero point estimate and standard error), but this is no problem as they will be imputed using a Generalized Variance Function (GVF) smoothing model, as described below.

– a linear transformation by centering and scaling the direct estimates. In particular we centered the estimates around their mean for each age class, mode, purpose combination, and scaled them by the inverse standard deviation of the direct estimates also for each age class, mode, purpose combination. Standard errors are transformed using the same scale factors as the point estimates.

All three transformations gave rise to improved model performance regarding MCMC simulation converge. To choose among these transformations, we considered the issue of variance stabilization, see e.g. Sakia (1992). Stabilizing the variances means that a transformation is chosen so as to remove the dependence between direct point estimates and their standard errors. This simplifies the ensuing modeling task, as we model the data using normal distributions[1)], for which there is no inherent relation between mean and variance. Figures displaying these dependencies for trip legs and distance can be found in the model building reports Boonstra et al. (2019c) and Boonstra et al. (2019b), respectively.

For number of trip legs the direct point estimates and standard errors display a strong positive dependence, which is largely removed by a square root transformation. In that case the log transformation is too strong, as it results in negative correlations. For distance, the dependence is even stronger and only the log transformation is able to remove most of the dependence between the direct point and variance estimates. Not surprisingly, the linear scaling and centering transformations are not capable of removing the dependence between the point and variance estimates.

So for number of trip legs we use the square root transformed direct estimates as input for a multilevel time series model. The smooth trends are then obtained by back-transforming the model output to the original scale. In the same way, the log-transformed direct estimates for distance per trip leg are input for a multilevel time series model, and the model output is exponentiated in order to get predictions at the original scale. There was no need to add a small ε value, since the observed distances are never smaller than 1 hectometer.

3.4 Smoothing the standard errors of the direct estimates

The time series models considered regard the (transformed) direct point estimates as noisy estimates of a true underlying signal. However, the accompanying variance estimates are largely treated as fixed and given quantities by the model. As the variance estimates can be very noisy due to the detailed estimation level, it is wise to smooth them before using them in the model. That way they better reflect the uncertainty of the direct estimates. The most obvious defect of the estimated standard errors is that they are zero in case of zero or one contributing sampling unit.[2)] This is correct for the structural zero domains, but it does not correctly reflect the uncertainty about the accidental zero estimates for number of trip legs. For distance, the most problematic estimates are the zero variance estimates in case of a single contributing sampling unit.

1) We also consider Student-t distributions, which can be interpreted as scale-mixtures of normal distributions.

2) This means that there is at most one unit (person or household) in a specific sex, age class domain, who reported a trip leg for a specific purpose, mode combination.

If there are no contributing sampling units for a certain domain then the direct distance estimate is treated as missing.

The models considered for smoothing the variance estimates are simple regression models relating the variance estimates to a few predictors such as sample size, design effects, and point estimates. Such models are known as Generalized Variance Function (GVF) models in the literature, see Wolter (2007), Chapter 7. In Boonstra et al. (2018) GVF smoothing models were applied to the standard errors of the untransformed direct estimates, but it turns out to be better to apply them after transforming the estimates. Otherwise the standard errors for domains with just a few observed trips can become unreasonably large due to the linearization approximation. Therefore we apply the GVF models as described in Boonstra et al. (2018) to the transformed standard error estimates. The predictions from this model are used as (smoothed) standard errors accompanying the transformed direct estimates as input for the time series multilevel models. In particular, the GVF model yields reasonable standard errors for the domains with no observed trips.

Let $\hat{Y}_{tijkl}$ now denote either the sqrt-transformed direct estimates for trip legs or the log-transformed estimates for distance, for year t, sex i, age class j, purpose k and mode l. For both target variables we use the same GVF smoothing model

$$\log se(\hat{Y}_{tijkl}) = \alpha + \beta \log \tilde{Y}_{tijkl} + \gamma \log(m_{tijkl} + 1) + \delta \log(\text{deff}_{tijkl}) + \epsilon_{tijkl}\,, \quad (9)$$

where m_{tijkl} is the number of sampling units (households or persons, depending on the survey year) contributing to domain (i, j, k, l) in year t, and

$$\text{deff}_{tijkl} = 1 + \frac{var(w)_{tijkl}}{\bar{w}^2_{tijkl}}\,, \quad (10)$$

is the design effect of the survey weights, in which the second term is the squared coefficient of variation of the weights of the contributing units to a specific year and domain.[3)] This factor accounts for the variance inflation due to the variation of the weights. Since we cannot trust the direct estimates for very small m_{tijkl}, the $\tilde{Y}_{tijkl}$ on the right hand side of (9) are simple smoothed estimates

$$\begin{aligned} \tilde{Y}_{tijkl} &= \lambda_{tijkl}\hat{Y}_{tijkl} + (1 - \lambda_{tijkl})\hat{Y}_{..jkl}\,, \\ \lambda_{tijkl} &= \frac{m_{tijkl}}{m_{tijkl} + 1}\,, \end{aligned} \quad (11)$$

where $\hat{Y}_{..jkl}$ denotes the mean of $\hat{Y}_{tijkl}$ over the years and sexes. For $m_{tijkl} = 0$ this replaces the estimate by the mean over year and sex for the same age class, purpose and mode. For $m_{tijkl} = 1$ the average of this mean and the estimate itself is used, and for large m_{tijkl} essentially the original point estimate is used.

The regression errors ϵ_{tijkl} are assumed to be independent and normally distributed with a common variance parameter σ^2. The GVF models are fitted to the positive standard errors of the transformed direct estimates. Summaries of the estimated model coefficients for trip legs and distance are given in Tables 3.1 and 3.2. The predicted (smoothed) standard errors based on the fitted models are

$$se_{\text{pred}}(\hat{Y}_{tijkl}) = \exp\left(\hat{\alpha} + \hat{\beta} \log \tilde{Y}_{tijkl} + \hat{\gamma} \log(m_{tijkl} + 1) + \hat{\delta} \log(\text{deff}_{tijkl}) + \hat{\sigma}^2/2\right)\,, \quad (12)$$

3) In case of 0 or 1 contributing units we have defined deff to equal 1.

predictor	coefficient	estimate	se
1	α	-0.701	0.012
$\log \tilde{Y}_{tijkl}$	β	0.941	0.003
$\log(m_{tijkl} + 1)$	γ	-0.497	0.002
$\log(\text{deff}_{tijkl})$	δ	0.420	0.007

Table 3.1 Estimated coefficients of the GVF model (9) for number of trip legs.

predictor	coefficient	estimate	se
1	α	-0.905	0.022
$\log \tilde{Y}_{tijkl}$	β	0.197	0.012
$\log(m_{tijkl} + 1)$	γ	-0.356	0.002
$\log(\text{deff}_{tijkl})$	δ	0.196	0.027

Table 3.2 Estimated coefficients of the GVF model (9) for distance per trip leg.

where $\hat{\sigma}$ is 0.11 for trip legs and 0.42 for distance. The R-squared model fit measures for both models are quite high: 0.90 for trip legs and 0.70 for distance. Note that the exponential back-transformation in (12) includes a bias correction, which in this case has only a small effect.

Figures 3.1 and 3.2 show the transformed direct estimates and their smoothed and non-smoothed standard errors in the for the relatively rare trip domain of purpose Education and age class 50-59. The two types of standard errors are displayed as approximate 95% confidence bands. The pink bands correspond to the original, non-smoothed, standard errors, whereas the blue bands correspond to the smoothed standard errors. From these example plots it appears that the predicted, i.e. smoothed, standard errors replace the original zero standard estimates by reasonable values. The same holds true for standard errors based on very few contributing units. Otherwise there is some modest smoothing of large outlying standard errors in some cases.

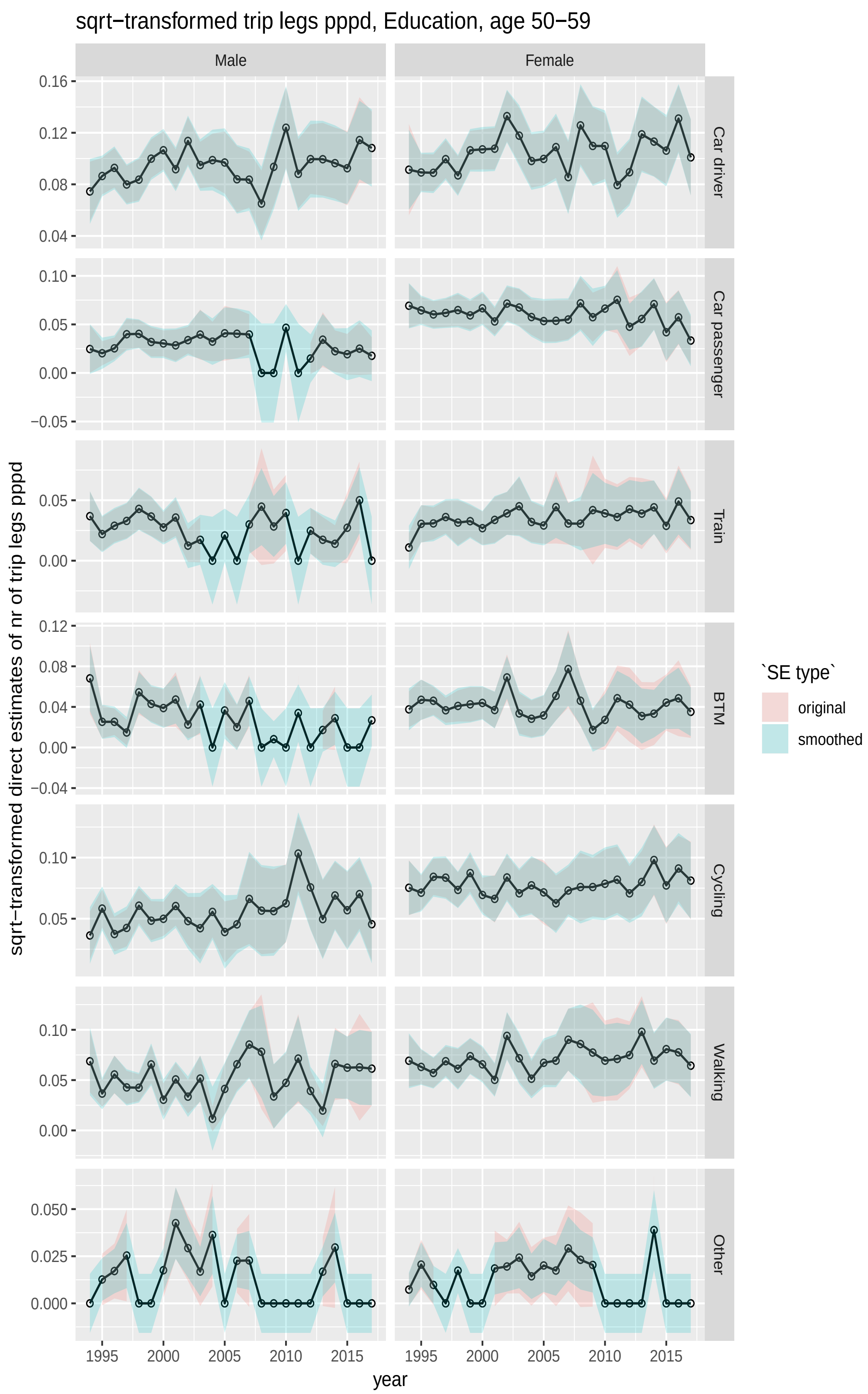

Figure 3.1 Transformed direct estimates and approximate 95% intervals based on smoothed and non-smoothed standard error estimates.

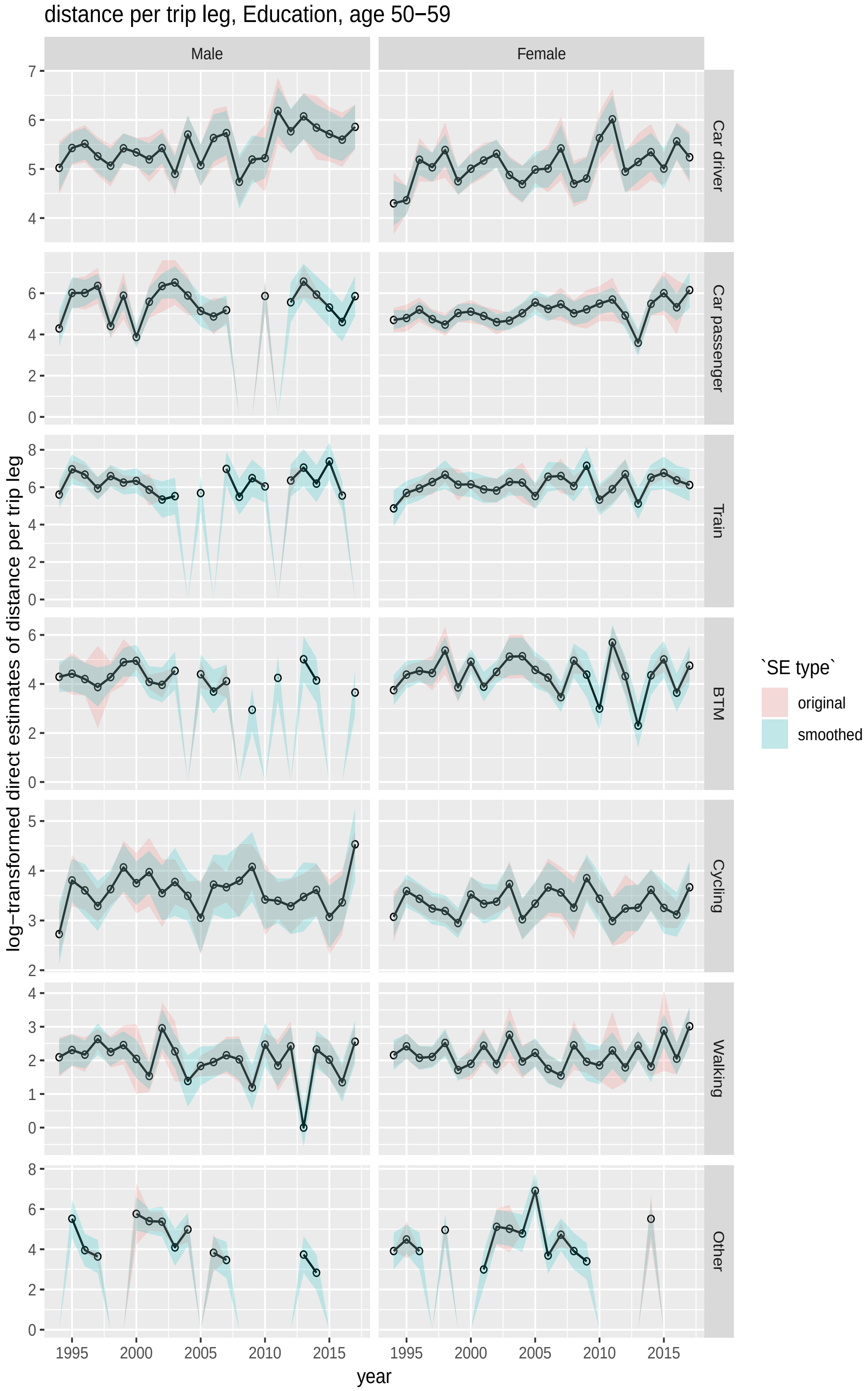

Figure 3.2 Transformed direct estimates and approximate 95% intervals based on smoothed and non-smoothed standard error estimates.

4 Time series multilevel modeling

The time series multilevel models considered are extensions of the popular basic area level model proposed by Fay and Herriot (1979). The models are defined at the most detailed level, i.e. the full cross-classification of sex, ageclass, purpose, mode and year. For convenience let us now denote by $\hat{Y}_{it}$ the transformed direct estimates for either trip legs or distance in year t and domain i. Here domain i refers to a particular combination of sex, ageclass, purpose and mode, so that i runs from 1 to $M_d = 504$ and t from 1 to T corresponding to the years 1999 to 2017. We further combine these estimates into a vector $\hat{Y} = (\hat{Y}_{11}, \dots \hat{Y}_{M_d 1}, \dots \hat{Y}_{1T}, \dots \hat{Y}_{M_d T})'$. Note that $\hat{Y}$ is a vector of dimension $M = M_d T$. Structural zero domains are not modeled, and it is implicitly understood that they are removed from all expressions. This means that the number of modeled initial estimates is reduced from $M = M_d T = 504 \times 19 = 9576$ to a total of 8720. For distance per trip leg there are in addition some domains without initial estimates due to the (accidental) absence of observed trips. The total number of available distance estimates is 8336. For both target variables model estimates are eventually produced for all 8720 non-structurally-zero domains.

4.1 Model structure

The multilevel models considered take the general linear additive form

$$\hat{Y} = X\beta + \sum_{\alpha} Z^{(\alpha)} v^{(\alpha)} + e\,, \tag{13}$$

where X is a $M \times p$ design matrix for a p-vector of fixed effects β, and the $Z^{(\alpha)}$ are $M \times q^{(\alpha)}$ design matrices for $q^{(\alpha)}$-dimensional random effect vectors $v^{(\alpha)}$. Here the sum over α runs over several possible random effect terms at different levels, such as transportation mode and purpose smooth trends, white noise at the most detailed level of the M domains, etc. This is explained in more detail below. The sampling errors $e = (e_{11}, \dots, e_{M_d 1}, \dots e_{M_d T})'$ are taken to be normally distributed as

$$e \sim N(0, \Sigma) \tag{14}$$

where $\Sigma = \bigoplus_{t=1}^{T} \Sigma_t$ with Σ_t the covariance matrix for the transformed direct estimates observed in year t. We have tried a simple model that takes (smoothed) estimated covariances for the input estimates into account, but eventually we take Σ_t and therefore Σ to be diagonal.

Equations (13) and (14) define the likelihood function

$$p\left(\hat{Y} | \eta, \Sigma\right) = N\left(\hat{Y} | \eta, \Sigma\right)\,, \tag{15}$$

where $\eta = X\beta + \sum_{\alpha} Z^{(\alpha)} v^{(\alpha)}$, called the linear predictor. A Student-t distribution for the sampling errors in (14) has been considered instead of the normal distribution to give smaller weight to more outlying observations. This is a traditional approach for handling outliers in Bayesian regression, see e.g. West (1984). We allow the degrees of freedom parameter of the Student-t distribution to be inferred from the data. It has been assigned a $\text{Gamma}(2, 0.1)$ prior distribution, which was recommended as a default prior in Juárez and Steel (2010).

The fixed effect part of η contains an intercept and main effects and possibly the second-order interactions for linear trends, discontinuities and the breakdown variables sex, age, purpose and mode. The vector β of fixed effects is assigned a normal prior

$p(\beta) = N(0, 100I)$, which is very weakly informative as a standard error of 10 is very large relative to the scales of the transformed direct estimates and the covariates used.

The second term on the right hand side of (13) consists of a sum of contributions to the linear predictor by random effects or varying coefficient terms. The random effect vectors $v^{(\alpha)}$ for different α are assumed to be independent, but the components within a vector $v^{(\alpha)}$ are possibly correlated to accommodate temporal or cross-sectional correlation. To describe the general model for each vector $v^{(\alpha)}$ of random effects, we suppress superscript α in what follows for notational convenience.

Each random effects vector v is assumed to be distributed as

$$v \sim N(0, A \otimes V), \tag{16}$$

where V and A are $d \times d$ and $l \times l$ covariance matrices, respectively, and $A \otimes V$ denotes the Kronecker product of A with V. The total length of v is $q = dl$, and these coefficients may be thought of as corresponding to d effects allowed to vary over l levels of a factor variable, e.g. purpose effects ($d = 4$) varying over time ($l = 19$ years). The covariance matrix A describes the covariance structure among the levels of the factor variable, and is assumed to be known. Instead of covariance matrices, precision matrices $Q_A = A^{-1}$ are actually used, because of computational efficiency (Rue and Held, 2005). The covariance matrix V for the d varying effects is parameterized in one of three different ways:

- an unstructured, i.e. fully parameterized covariance matrix
- a diagonal matrix with unequal diagonal elements
- a diagonal matrix with equal diagonal elements

The following priors are used for the parameters in the covariance matrix V:

- In the case of an unstructured covariance matrix the scaled-inverse Wishart prior is used as proposed in O'Malley and Zaslavsky (2008) and recommended by Gelman and Hill (2007).
- In the case of a diagonal matrix with equal or unequal diagonal elements, half-Cauchy priors are used for the standard deviations. Gelman (2006) demonstrates that these priors are better default priors than the more common inverse gamma priors for the variances.

The following random effect structures are considered in the model selection procedure:

- Random intercepts for the M_d domains obtained by the full cross clasification of age, gender, purpose and mode. In this case $A = I_{M_d}$ and V is a scalar variance parameter, and the corresponding design matrix is the $M \times M_d$ indicator matrix for domains. This can be extended to a vector of random domain intercepts, random slopes for linear time effects and discontinuities due to the redesigns in 2004 and 2010. In that case V is a 4×4 covariance matrix, parameterized by variance parameters for the intercepts, linear time slopes and the coeffcients for the level interventions, and possibly six correlation parameters. Similar random intercept and slope terms varying over the categories of the cross-classification of a subset of 2 or 3 of the 4 classification variables have also been considered.
- Random effects that account for outliers. The data for some years appear to be of lesser quality. This is the case for example for data on the number of trip legs in 2009. In order to deal with such less reliable estimates, random effects can be used to absorb some of the larger deviations in such years. The corresponding effects are removed from the trend prediction. This is an alternative to the use of fat-tailed sampling distributions such as the Student-t distribution for dealing with outliers.

- Random walks or smooth trends at aggregated domain levels (e.g. purpose by mode). See Rue and Held (2005) for the specification of the precision matrix Q_A for first and more smooth second order random walks. A full covariance matrix for the trend innovations can be considered to allow for cross-sectional besides temporal correlations, or a diagonal matrix with equal or different variance parameters to allow for temporal correlations only.
- White noise. In order to allow for random unexplained variation, white noise at the most detailed domain-by-year level can be included. In this case $A = I_M$ and V a scalar variance parameter, and the design matrix is $Z = I_M$.

We also investigate generalisations of (16) to non-normal distributions of random effects. Relevant references are Carter and Kohn (1996) in the state space modeling context, Datta and Lahiri (1995), Fabrizi and Trivisano (2010) and Tang et al. (2018) in the small area estimation context, and Lang et al. (2002) and Brezger et al. (2007) in the context of more general structured additive regression models. In particular, the following distributions are considered for various random effect terms:

- Student-t-distributed random effects
- Random effects with a so-called horseshoe prior (Carvalho et al., 2010).
- Random effects distributed according to the Laplace distribution. This corresponds to a Bayesian version of the popular lasso shrinkage, see (Tibshirani, 1996; Park and Casella, 2008).

These alternative distributions have fatter tails allowing for occasional large effects. The Laplace and particularly the horseshoe distribution have the additional property that they shrink noisy effects more strongly towards zero.

4.2 Model estimation

The models are fitted using Markov Chain Monte Carlo (MCMC) sampling, in particular the Gibbs sampler (Geman and Geman, 1984; Gelfand and Smith, 1990). See Boonstra and van den Brakel (2018) for a specification of the full conditional distributions. The models are run in R (R Core Team, 2015) using package `mcmcsae` (Boonstra, 2018) developed at Statistics Netherlands. The Gibbs sampler is run in parallel for three independent chains with randomly generated starting values. In the model building stage 1000 iterations are used, in addition to a 'burn-in' period of 250 iterations. This was sufficient for reasonably stable Monte Carlo estimates of the model parameters and trend predictions. For the selected model we use a longer run of 1000 burn-in plus 10000 iterations of which the draws of every fifth iteration are stored. This leaves $3 * 2000 = 6000$ draws to compute estimates and standard errors. The convergence of the MCMC simulation is assessed using trace and autocorrelation plots as well as the Gelman-Rubin potential scale reduction factor (Gelman and Rubin, 1992), which diagnoses the mixing of the chains. For the longer simulation of the selected model all model parameters and model predictions have potential scale reduction factors below 1.02 and sufficient effective numbers of independent draws.

Many models of the form (13) have been fitted to the data. For the comparison of models using the same input data we use the Widely Applicable Information Criterion or Watanabe-Akaike Information Criterion (WAIC) (Watanabe, 2010, 2013) and the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002). We also compare the models graphically by their model fits and trend predictions at various aggregation levels.

5 Model building, selected models, and model prediction

The direct estimates for all 504 domains over the 1999-2017 period along with their standard errors serve as input data for the time series models. The variables defining the domains and the years have been used in the model development in many ways, e.g. using different interactions of various orders. Some additional covariates have been constructed in order to model the discontinuities between 2003 and 2004 and between 2009 and 2010, as well as to reduce the influence of some lesser quality 2009 input estimates. For the MON level break in 2004, a variable br_mon is introduced taking values 1 for the years 2004-2009 and 0 outside that range. For the OViN level break in 2010, likewise a variable br_ovin is defined, taking values 1 for the years 2010 and later and 0 otherwise. As it turned out, a slight modification of this variable was necessary in order not to introduce artificial level breaks in the age 12-17, car driver domains, which are structurally zero domains before 2011. Also, for year 2009 a dummy variable $dummy\_2009$ has been created being only 1 when year equals 2009 and 0 otherwise. These variables have been used in different interactions, as fixed or random effects in the model. The year variable is also used quantitatively to define linear time trends, and for that purpose we use a scaled and centered version denoted $yr.c$.

Some other covariates extracted from other sources like Statistics Netherlands' Statline and KNMI meteorological annual reports[4)] have also been used as candidate covariates in the model development. As for example, a weather variable $snowdays$ representing the number of snow days by year is used in the final trip-legs model and an administrative variable km_NAP representing annual registered car kilometers collected from Nationale Autopas (NAP) is used in the distance model.

In the following two sub-sections, time series models developed for the number of trip legs and the distance per trip leg are discussed. Following that, it is described how the target trend estimates are derived from the developed time series models. The models are expressed as time series multilevel models in a hierarchical Bayesian framework and fit using a Markov Chain Monte Carlo (MCMC) simulation method, as described in Section 4.

5.1 Time series multilevel model for the number of trip legs

As described in Section 3, we model the square-root-transformed direct estimates of the number of trip legs pppd, using the corresponding transformed and GVF-smoothed standard errors to define the variance matrix Σ of the sampling errors.

The model parameters in (13) are separated in fixed and random effects. After extensive examination of different models, the following fixed effects components are included in the finally selected model:

$$sex * ageclass + purpose * mode + mode * snowdays \tag{17}$$

Terms like $sex * ageclass$ in (17) include both main and interaction effects. Inclusion of only main effects resulted in underfitting, while inclusion of higher than second-order

4) We thank Hans Wüst for the suggestion to consider weather variables and for collecting these figures.

interactions resulted in overfitting. Third and higher order interactions are instead modeled as random effect terms, see Table 5.1, which lists the selected random effects terms. The interaction between purpose and mode is modeled as a fixed effect because of the large differences in level between the purpose, mode combinations. The term $mode * snowdays$ was added because it (slightly) improves the model information criteria, and because of the small but plausible effect it has on some trend predictions for walking, cycling and car passenger trip legs.

For the random effects part of the model, the model selection procedure involved choosing between different prior distributions. First, if multiple varying effects are modeled then there is a choice between scalar, diagonal or full covariance matrix V in (16). Second, for variation over time it makes sense to choose a matrix A with appropriate correlation structure. We consider both first and second order random walks corresponding to local level and smooth trends. Finally, the normal distribution in (16) can be modified to other distributions with different behaviour by the introduction of additional local scale parameters. We tried Student-t, Laplace and horseshoe priors, as mentioned in Section 4. These alternative distributions have fatter tails allowing for occasional large outlying effects. The Laplace and particularly the horseshoe distribution have the additional property that they shrink effects not strongly supported by the data more towards zero.

Model Component	Formula V	Variance Structure	Factor A	Prior	Number of Effects
V_2009	$dummy\_2009$	scalar	$sex * ageclass * purpose * mode$	horseshoe	504
V_BR	$1 + yr.c + br\_mon\_SO + br\_ovin$	full	$sex * ageclass * purpose * mode$	Laplace	1764
RW2AMM	$ageclass * purpose * mode$	scalar	RW2(yr)	normal	4788
RW2MM	$purpose * mode$	diagonal	RW2(yr)	normal	532
WN	1	scalar	$sex * ageclass * purpose * mode * yr$	normal	9576

Table 5.1 Summary of the random effect components for the selected time series multilevel model. The second and third columns refer to the varying effects with covariance matrix V in (16), whereas the fourth and fifth columns refer to the factor variable associated with A in (16). The last column contains the total number of random effects for each term.

To reduce the influence of outliers, student-t distributed sampling errors were attempted at first. This approach did not work well for trip legs, perhaps because the outliers are quite specific and concentrated mostly in 2009 and specific domains. This led us to include random effects for the dummy variable $dummy\_2009$ at the domain level, resulting in a significant improvement of model fit (WAIC, DIC) and clearly visible reduction of the influence of several 2009 outliers. The resulting model term is named 'V_2009' in Table 5.1. It uses a horseshoe prior distribution, which improved the model fit as it is better able to accommodate some of the large outliers.

The random effects component 'V_BR' includes MON and OViN level break random effects, random intercepts, and random linear time trends, varying over all domains (the cross-classification of sex, ageclass, purpose and mode). The full covariance structure resulted in the best model improvement. The full covariance matix V in (16) is a 4 x 4 matrix parameterised in terms of standard deviation and correlation parameters. Since random MON break effects for the purposes work and education resulted in somewhat artificial and implausible trend estimates, it was decided to model MON breaks only for the purposes shopping and other. Therefore, we introduced a variable br_mon_SO, equal to br_mon for purposes shopping and other and zero for purposes work and education. Shopping and other are the purposes with the largest OViN breaks as well. Using a Laplace prior distribution for 'V_BR' further improved the model fit.

Two smooth time trend components at purpose × mode and ageclass × purpose × mode aggregation levels are included in the final model. These terms are named 'RW2MM' and 'RW2AMM' respectively in Table 5.1. The best model fit was obtained using a diagonal variance for 'RW2MM' and a scalar variance for the more detailed 'RW2AMM' component. The different values found for the variance components of the 'RW2MM' components indicate large diifferences in degrees of smoothness of the various series. The 'RW2AMM' component can be interpreted as a correction to the 'RW2MM' trends, allowing for some differences between age classes. The contribution of the 'RW2AMM' effects is indeed generally of a smaller size than that of the 'RW2MM' effects.

Finally, a white noise term named 'WN' in Table 5.1 was added in the final model to capture unstructured variation over all levels of all factors. This 'WN' component accounts for more or less random variation of the true average number of trip legs pppd over the domains and the years.

5.2 Time series multilevel model for the distance per trip leg

For distance we model the log-transformed direct estimates of distance per trip leg, using the corresponding transformed and GVF-smoothed standard errors discussed in Section 3 to define the variance matrix Σ of the sampling errors. The use of Student-t distributed sampling errors in this case succeeds in reducing the influence of outliers sufficiently. The degrees of freedom parameter of the Student-t distribution is assigned a weakly informative prior and is inferred from the data.

Similar to the model for number of trip legs pppd, only main effects and second order interaction effects are used in the fixed effects part of the selected model. In the finally selected model the following fixed effects components are included:

$$\begin{aligned} &sex * ageclass + purpose * mode + yr.c * mode + \\ &\quad + mode\_walking : br\_ovin + mode\_cardriver : logratio\_km\_NAP \end{aligned} \tag{18}$$

Here the term $yr.c * mode$ represents linear time trends by mode. The variables $mode\_cardriver$ and $mode\_walking$ are indicator variables for transportation modes car driver and walking. The term $mode\_walking : br\_ovin$ represents a single OViN break fixed effect for mode walking. Among the covariates extracted from external sources, the the year-by-year differences in the log of registered car kilometers (km_NAP), denoted by $logratio\_km\_NAP$, is used in combination with mode car driver as it as it slightly improves model fit and gives rise to some small but plausible changes in the trend estimates for mode car driver.

The other effects, including higher order interactions, are modeled as random effects, and the selected terms are shown in Table 5.2. As in the model development for number of trip legs, the combination of intercepts, linear time trends (slopes) and both level breaks varying over all domains, represented by the term 'V_BR' in Table 5.2, yields a reasonable fit. A full covariance matrix among the random intercepts, the linear time trends and the two break variables br_mon and br_ovin also works best here. A Laplace prior distribution for 'V_BR' enables to accommodate some of the larger breaks quite well.

Several time trend components at different levels have been considered for inclusion, but adding multiple such trend components showed signs of overfitting. Inclusion of a single smooth trend, i.e. second order random walk, for each mode, purpose combination with a variance parameter depending only on mode turned out to work best. This term is named RW2M in Table 5.2. Finally, a white noise term, WN in Table 5.2, is added to capture remaining unstructured variation over all levels.

Model Component	Formula **V**	Variance Structure	Factor **A**	Prior	Number of Effects
V_BR	$1 + yr.c + br\_mon + br\_ovin$	unstructured	$sex * ageclass * purpose * mode$	Laplace	2016
RW2M	$mode$	diagonal	$purpose$*RW2(yr)	normal	532
WN	1	scalar	$sex * ageclass * purpose * mode * yr$	normal	9576

Table 5.2 Summary of the random effect components for the selected time series multilevel model. The second and third columns refer to the varying effects with covariance matrix V in (16), whereas the fourth and fifth columns refer to the factor variable associated with A in (16). The last column contains the total number of random effects for each term.

5.3 Trend estimation and derived estimates

The trend estimates of main interest are computed based on the MCMC simulation results as follows. First, simulation vectors of model linear predictions are formed, i.e.

$$\eta^{(r)} = X\beta^{(r)} + \sum_{\alpha} Z^{(\alpha)} v^{(\alpha,r)}\,, \tag{19}$$

where superscript r indexes the retained MCMC draws, and each $\eta^{(r)}$ is of dimension M. Consequently, the level break effects are removed or added, depending on the choice of benchmark level. We choose the OViN level as the benchmark level, as it corresponds to the most recent period considered. Given the way the level break dummies are coded, it means that we need to add all OViN break effects to the predictions referring to the OVG and MON years, and in addition need to remove the MON effects from the predictions referring to the MON years. Also, if present, the dummy effects for outliers are removed. We note that the survey errors e in (13) are already absent from the linear predictor (19). The simulation vectors of linear predictors thus obtained are

$$\tilde{\eta}^{(r)} = \tilde{X}\beta^{(r)} + \sum_{\alpha} \tilde{Z}^{(\alpha)} v^{(\alpha,r)}\,, \tag{20}$$

where $\tilde{X}$ and $\tilde{Z}^{(\alpha)}$ are modified design matrices that accomplish the stated correction for level breaks and possibly outlier effects. Back-transformation of these vectors to the original scale yields the MCMC approximation to the posterior distribution of the trends. For the square root transformation as used for modeling the number of trip legs pppd, the back-transformation amounts to

$$\theta^{(r)} = (\tilde{\eta}^{(r)})^2 + (se(\hat{Y}_{it}^{\text{sqrt}}))^2\,, \tag{21}$$

The second term on the right hand side accomplishes a (relatively small) bias correction using the transformed and smoothed standard errors for number of trip legs. The bias correction stems from the fact that the design expectation of the direct estimates can be written as

$$E(\hat{Y}) = E((\hat{Y}^{\text{sqrt}})^2) = E((\eta + e^{\text{sqrt}})^2) = \eta^2 + 2\eta E(e^{\text{sqrt}}) + E((e^{\text{sqrt}})^2) = \eta^2 + var(e^{\text{sqrt}}), \tag{22}$$

where e^{sqrt} is the vector of sampling errors after transformation, assumed to be normally distributed with standard errors $se(\hat{Y}_{it}^{\text{sqrt}})$.

For the log transformation, as used in modeling distance per trip leg, back-transforming $\tilde{\eta}^{(r)}$ to the original scale yields the MCMC approximation to the posterior distribution of the distance trends. The exponential back-transformation including bias correction is

$$\theta^{(r)} = e^{\tilde{\eta}^{(r)} + \text{se}(\hat{Y}^{\text{log}})/2}. \tag{23}$$

The bias correction is added to largely correct a small negative bias induced by the log transformation, see for example Fabrizi et al. (2018).

The means over the MCMC draws $\theta^{(r)}$ are used as trend estimates, whereas the standard deviations over the draws serve as standard error estimates.

Recall that η and θ are vector quantities with components for all year-domain combinations. We have computed the trends at the most detailed level, but we also have computed aggregates over several combinations of the domain characteristics. Aggregation of distance per trip leg involves the number of trip legs, and so requires combining the MCMC output for both target variables. By multiplying the distance per trip leg results by the number of trip leg pppd results we obtain the results for distance pppd. Aggregation amounts to simple summation over trip characteristics purpose and mode, and to population weighted averaging over person characteristics sex and ageclass. Inference for other derived quantities like total number of trip legs per day and total distance per day at different aggregation levels can also be readily conducted using the simulation results for the two modeled target variables.

6 Results

The appendices contain several figures showing trend estimates at different aggregation levels based on the selected models for the number of trip legs pppd and the distance per trip leg described in Section 5. The black lines in these figures correspond to the series of direct estimates, the red lines to the model fit based on all model components, i.e. the back-transformation of (19), and the green lines to the trend series (21) or (23). A complete set of time series plots for the number of trip legs pppd, distance per trip leg, and distance pppd at different aggregation levels, including the most detailed level, are given in Boonstra et al. (2019c) and Boonstra et al. (2019b). In the following two sub-sections, the trend estimates obtained for number of trip legs and distance are discussed and illustrated.

6.1 Trip legs

Trend estimates of total number of trip legs per day at overall, purpose and mode levels are shown in Figures A.1, A.2, and A.3 respectively in Appendix A. It is noted that these figures are based on the (aggregation of the) trends for number of trip legs pppd estimated at the most detailed level. A selected set of plots for the number of trip legs pppd is given in Appendix B.

The plots show that the estimated trends for trip legs are hardly affected by any MON breaks. This has partly be enforced by excluding MON breaks for the purposes work and education in the random effects model component 'V_BR'. Since the OVG and MON designs are largely the same except that they were carried out by different data collection organisations, no large MON breaks are anticipated anyway. However, at a more detailed level, some MON discontinuities are clearly present for purposes shopping and other, as is shown in Figure B.5, for the 0-5 age group. It shows a possible exchange in classification of purposes shopping and other for young children.

In contrast, some OViN breaks are quite large. Overall, the OViN level for trip leg number is lower than the MON and OVG levels, as is clear from Figure B.1. This is the case for modes car driver, train and cycling, as illustrated by Figure B.3. For mode walking there is also a large jump downwards, but it happens in 2011, a year after the start of OViN. A full explanation of this jump is lacking, but it seems that the weather at least plays a partial role. The year 2010 was a year with a rather extreme amount of snow days, and under such circumstances it is expected that more walking trips are made, e.g. as an alternative to cycling. It appears strange, however, that the trend level before 2010 is not much lower than that of 2010. The small OViN break that is estimated here is actually due to inclusion of $snowdays$ as a covariate. Without it, there would have been even a larger jump in 2011.

There are some noteworthy differences in discontinuities between men and women trend lines, particularly for 30-39 and 40-49 age groups, see e.g. Figure B.7. In these particular cases the differences in the levels of the direct estimates between men and women are much larger during the OViN period. As these differences are most probably due to measurement errors in OViN data, this is a drawback of basing the trend estimates on the OViN level.

Since the trends are defined at the level of OViN in the current model, the outcomes during the MON and OVG period are corrected for the discontinuities induced by the redesigns in the past. It implies that due to the uncertainty of the estimated discontinuities the standard errors for the trend estimates in the OVG and MON period are larger compared to the OViN period. At an aggregated level, the standard errors of the trend estimates are even larger than the uncertainty of the direct estimates. See for example Figure B.1 for estimates at the overall level and Figures B.2 and B.3 for estimates by purpose and mode.

The 2009 outlier effects are pronounced for some domains, notably for young children and purposes shopping and 'other', as illustrated in Figure B.5). There is a clear exchange between both purposes for the young children. These effects have been captured by the random effect term 'V_2009' of the selected model for trip legs. The trend lines show that these outliers are indeed neutralized by excluding the 'V_2009' effects.

Tables 6.1 and 6.2 list the posterior means and standard errors of several variance components of the trip leg model. It is to be noted from Table 6.1 that the random intercepts over the domains are negatively correlated with the OViN break effects and random linear time trends. Table 6.2 shows that the scales of the second order random walks by purpose and mode in the 'RW2MM' model component are very diverse. The largest scales are seen for cycling and walking for purpose 'other'. This difference in volatility by purpose and mode is also visible in the trends, as shown in Figure B.4.

Especially the high volatility of the series for cycling and purpose 'other' is apparent, which may possibly be caused by weather effects. This series closely follows the direct estimates series due to the relatively small standard errors of the latter. Other domains have more smooth trends despite volatile direct estimates, as for example train for purpose 'other'. As the direct standard error estimates are much larger in this case, the model chooses a smoother trend series. Still other domains, such as car driver for purpose work, show smooth series for both direct estimates and trend predictions.

Compared to the fixed effects, 'V_BR' and 'RW2MM' components, it was found that the more detailed trend component 'RW2AMM' has a smaller contribution to the linear predictor, and is smoother than most 'RW2MM' components. The white noise term 'WN' was also found to have only a modest contribution to the trend estimates.

A positive trend in the number of trip legs can be observed for the 12-17 age group as car driver after the 2011 change in law, see Figure B.6. Though this domain is very small and the time period of available data is short, the fitted model seems to work satisfactorily here.

	Intercept	br_mon_SO	br_ovin	yr.c
Intercept	14.81 (0.73)	13.90 (9.90)	-41.80 (5.96)	-15.00 (6.89)
br_mon_SO		1.80 (0.18)	13.40 (11.28)	21.80 (11.39)
br_ovin			3.49 (0.23)	-11.40 (7.96)
yr.c				1.77 (0.11)

Table 6.1 Estimated standard deviations and correlations ($\times 100$) for the 'V_BR' component

	Car Driver	Car Passenger	Train	BTM	Cycling	Walking	Other
Work	3.66 (1.97)	3.00 (4.81)	2.16 (1.90)	1.19 (1.13)	1.23 (1.31)	6.77 (3.06)	1.82 (1.54)
Shopping	2.68 (1.46)	4.28 (2.08)	0.86 (0.95)	1.35 (1.66)	1.76 (1.67)	3.11 (4.07)	1.32 (1.14)
Education	0.86 (0.93)	1.85 (2.91)	0.67 (0.68)	2.89 (2.81)	5.18 (3.52)	3.14 (1.78)	2.80 (1.70)
Other	7.14 (3.64)	5.72 (2.72)	1.16 (1.19)	1.89 (1.14)	23.75 (7.07)	13.69 (4.17)	1.84 (2.17)

Table 6.2 Estimated standard deviations ($\times 1000$) with their standard errors in parentheses under the diagonal covariance matrix for the 'RW2MM' component

6.2 Distance

The trends of total distance per day at overall, purpose and mode levels are shown in Figures C.1, C.2, and C.3 respectively in Appendix C. The corresponding plots for the trends of distance pppd and distance per trip leg are given in Appendix D and Appendix E respectively.

The direct estimates of distance per trip leg are rather volatile, even at the most aggregated level, see Figure E.1. Due to the nature of the distance variable compared to number of trip legs, the former series are generally more affected by outliers than the trip leg series. The Student-t distribution with a posterior mean degrees of freedom of about 4 seems to account for outliers quite well, but it is harder to detect fine changes in the true underlying distance trends. Consequently, in order to avoid overfitting the model for distance is more parsimonious than that for the number of trip legs. One exception is that the distance model includes a fixed OViN break effect for mode walking. This effect was required to capture the very pronounced discontinuity in 2010 for mode walking[5)], as shown in Figure E.3.

For the distance model, some parameter estimates (posterior means and standard errors) are listed in Tables 6.3 and 6.4. The 'V_BR' component containing varying coefficients by domain for intercept, slope, and MON and OViN breaks, shows, as in the trip legs model, negative correlation among the intercepts and OViN breaks, see Table 6.3.

Table 6.4 shows that the mode-dependent scales of the smooth trend components by mode and purpose, as represented by model component 'RW2M', are quite diverse. Here, the components are most volatile for BTM, walking and, especially, 'other'. As is the case for the trip leg model, the white noise component makes only a relatively small contribution to the trends.

	Intercept	br_mon	br_ovin	yr.c
Intercept	26.40 (1.45)	-0.44 (26.06)	-22.26 (11.50)	18.08 (12.65)
br_mon		1.32 (0.89)	8.80 (30.73)	12.22 (35.28)
br_ovin			11.10 (1.51)	9.95 (22.63)
yr.c				3.98 (0.65)

Table 6.3 Estimated standard deviations and correlations ($\times 100$) for the 'V_BR' component.

Car Driver	Car Passenger	Train	BTM	Cycling	Walking	Other
1.77 (1.53)	3.92 (2.10)	10.13 (5.42)	20.68 (7.41)	7.47 (2.74)	26.59 (13.75)	33.73 (12.31)

Table 6.4 Estimated standard deviations ($\times 1000$) with their standard errors in parentheses under the diagonal covariance matrix for the 'RW2M' component.

7 Model assessment

As mentioned before, model selection was largely based on the WAIC, DIC, and graphical comparisons of their model fits and trend predictions at various aggregation levels. To evaluate how adequate the models fit the time series of direct estimates more formally, posterior predictive checks has been applied to the finally selected models. Furthermore it has been tested to what extent the residuals of the time series are independently and identically distributed by testing for normality or whether the residuals follow a t-distribution, depending on the assumed distributions in the likelihood. Also formal tests have been conducted for heteroscedasticity and remaining serial correlation in the

5) The effect is most probably due to the fact that in OViN walks are more often classified as single tours instead of consisting of a go and return trip.

residuals. Finally the bias in the model predictions, the variance reduction of the model predictions compared to the initial direct estimates, and the size of revisions if new data become available.

In this section, results are presented for bias and variance reduction along with the revision analysis. For the details of the posterior predictive checks and the various tests applied to the standardized residuals, see Boonstra et al. (2019a)

7.1 Bias and variance reduction

Two discrepancy measures are defined to evaluate and compare the time series multilevel models. The first measure is the Relative Bias (RB) which expresses the differences between model estimates and direct estimates, as percentage of the latter. For a given model, the RB_{it} for a domain is defined as

$$RB_{it} = \frac{(\hat{\theta}_{it} - \hat{Y}_{it})}{\hat{Y}_{it}} \times 100\% \,. \tag{24}$$

with $\hat{\theta}_{it}$ the model prediction and $\hat{Y}_{it}$ the direct estimate for domain i and year t. This benchmark measure shows for each domain i how much the model-based estimates deviate from the direct estimates. The discrepancies should not be too large as one may expect that the direct estimates on average are close to the true average number of trip legs or distances. The second discrepancy measure is the Relative Reduction of the Standard Errors (RRSE) and measures the percentages of reduction in standard error of the model-based compared to the direct estimates, i.e.,

$$RRSE_{it} = 100\% \times (se(\hat{Y}_{it}) - se(\hat{\theta}_{it}))/se(\hat{Y}_{it}) \,. \tag{25}$$

Both measures are evaluated at different aggregation levels. The distributions of the measures are presented in terms of the minimum value, 1st quartile, median, mean, 3rd quartile and maximum value. The measures are presented for the following aggregation levels:

- Yearly estimates at the highest aggregation level (19 estimates, where (24) and (25) are averaged over all domains)
- Yearly estimates for the mode categories (19 by 7 estimates, where (24) and (25) are averaged over all domains that belong to a mode category)
- Yearly estimates for the purpose categories (19 by 4 estimates)
- Yearly estimates for the cross-classification of mode and purpose (19 by 7 by 4 estimates)
- Yearly estimates for the cross-classification of mode, purpose, gender and age-class (19 by 7 by 4 by 2 by 9 estimates)

7.2 Revision analysis

The models for the number of trip legs and distances are fitted to time series of different lengths, starting with series observed up until 2010 and then adding a year sequentially. Directly after the change-over from MON to OViN, this gives an impression how many observations under the new design are required until stable estimates for the discontinuities are obtained. For the last three years of the observed series, the discontinuity estimates have converged to a stable value. For this period, this revision analysis gives an impression how stable the prediction for the last year is and how large its revision is when a new observation becomes available. Results are presented for the most aggregated level, the categories for mode and motive.

7.3 Trip legs

The distributions of the RB_{it} (24) and $RRSE_{it}$ (25) for the different aggregation levels specified in Section 7.1, are provided in Tables 7.1 and 7.2, respectively. The bias at the highest aggregation level is negligible (average of -0.14%) and gradually increases to an average of -1.16% at the most detailed level. The reduction of the variance is the smallest at the highest aggregation level (25% on average) and gradually increases with the level of detail to 46.9% at the most detailed level.

	Variable	Min.	1st.Qu.	Median	Mean	3rd.Qu.	Max.
1	Year	-1.16	-0.51	-0.03	-0.14	0.27	0.54
2	Motive	-0.35	-0.34	-0.21	-0.21	-0.08	-0.06
3	Mode	-0.77	-0.48	-0.25	-0.31	-0.10	-0.01
4	Motive and Mode	-1.54	-0.57	-0.28	-0.37	-0.05	0.11
5	Sex, Ageclass, Motive, and mode	-64.43	-1.08	-0.17	-1.16	0.30	50.29

Table 7.1 Summary statistics of mean relative bias (in %) at different aggregation levels for the SAE estimates of mean number of trip legs pppd

	Variable	Min.	1st.Qu.	Median	Mean	3rd.Qu.	Max.
1	Year	11.72	21.75	27.54	25.08	28.83	36.71
2	Motive	19.65	28.54	33.13	31.18	35.77	38.82
3	Mode	22.16	24.24	27.31	29.81	36.23	38.25
4	Motive and Mode	10.30	30.19	36.70	33.94	39.16	43.26
5	Sex, Ageclass, Motive, and mode	33.74	43.10	45.63	46.90	48.20	95.32

Table 7.2 Summary statistics of relative reduction of standard errors (in %) at different aggregation levels for the standard errors of the SAE estimates of mean number of trip legs pppd

The size of the revisions is evaluated graphically at the highest aggregation level in Figures 7.1 and 7.2. Figure 7.1 illustrates the revisions if directly after the change-over from MON to OVIN trends are published at the OVIN level for time series observed up until 2010 and then adding one year at a time. This is a so-called in real time analysis for periods 2010 to 2017. Similar figures for the purpose and mode categories are provided in Boonstra et al. (2019a).

Figure 7.2 illustrates the revisions if trends are published at the OVG level in real time for the period directly after the implementation of the OViN, i.e. 2010 to 2017. Similar figures for the purpose and mode categories are provided in Boonstra et al. (2019a).

In real time, estimates for the discontinuity due to the change-over from MON to OViN are presented in Figure 7.3. This figure illustrates that it takes about 4 years before the estimate for the discontinuity converges to a stable value. The updates for the OViN discontinuities during the years directly after the implementation of OViN are one factor that causes the revisions of the trends visible in 7.1 and 7.2.

Figure 7.1 in which the signals are estimated at the OViN level, shows large revisions of the signals during the MON period (2004-2009). If signals are estimated at the OViN level, the estimate for the OViN discontinuity is added to the signal at all time periods. During the MON period, the estimate for the MON discontinuity is removed. The revisions of the OViN discontinuities in the period 2010 until 2014, also influence the estimates for the MON discontinuities. The revisions of the signals during the OVG period (1999-2003) are purely the effect of the revision of the estimated OViN discontinuity. The updates of the trend levels in the OVG period in Figure 7.1 indeed agree with the revision of the estimated OViN discontinuity in Figure 7.3. These findings are confirmed with the in real time analyses at the level of the mode and purpose categories.

It appears that the revisions of the signals over the entire time period are smaller if the signals are estimated on the OVG level. These findings are confirmed with the in real time analyses at the level of the purpose and mode categories in Figures 7.1 and 7.2. The most important conclusion from the revision analysis at the OViN and OVG level is that the dummies that model the different redesigns are coded, must adapted to the publication level of the signals. In this application there is a dummy for the MON break, and a dummy for the OViN. These are regression variables that equal one during the MON period and zero elsewhere and equal one during the OViN period and zero elsewhere (Boonstra et al., 2019c,b). If the desired level of publication is OViN it might be preferable to take OViN as the baseline level; i.e. use one dummy that equals one during the OVG period and zero elsewhere and one dummy that equals one during the MON period and zero elsewhere. It appears to be advisable to choose the publication level as the baseline level. This, however, requires additional research. It is also not clear to what extent this empirical finding can be generalized to the implementation of ODiN.

As mentioned before, it takes about 4 years until stable estimates for the discontinuities are obtained (Figure 7.3). In the period after 2014 the revisions are very small. This is an indication that once stable estimates for the discontinuities are obtained, the revisions of the model predictions are small. This might suggest that, once sufficient data are available after the last redesign, no revision policy is required.

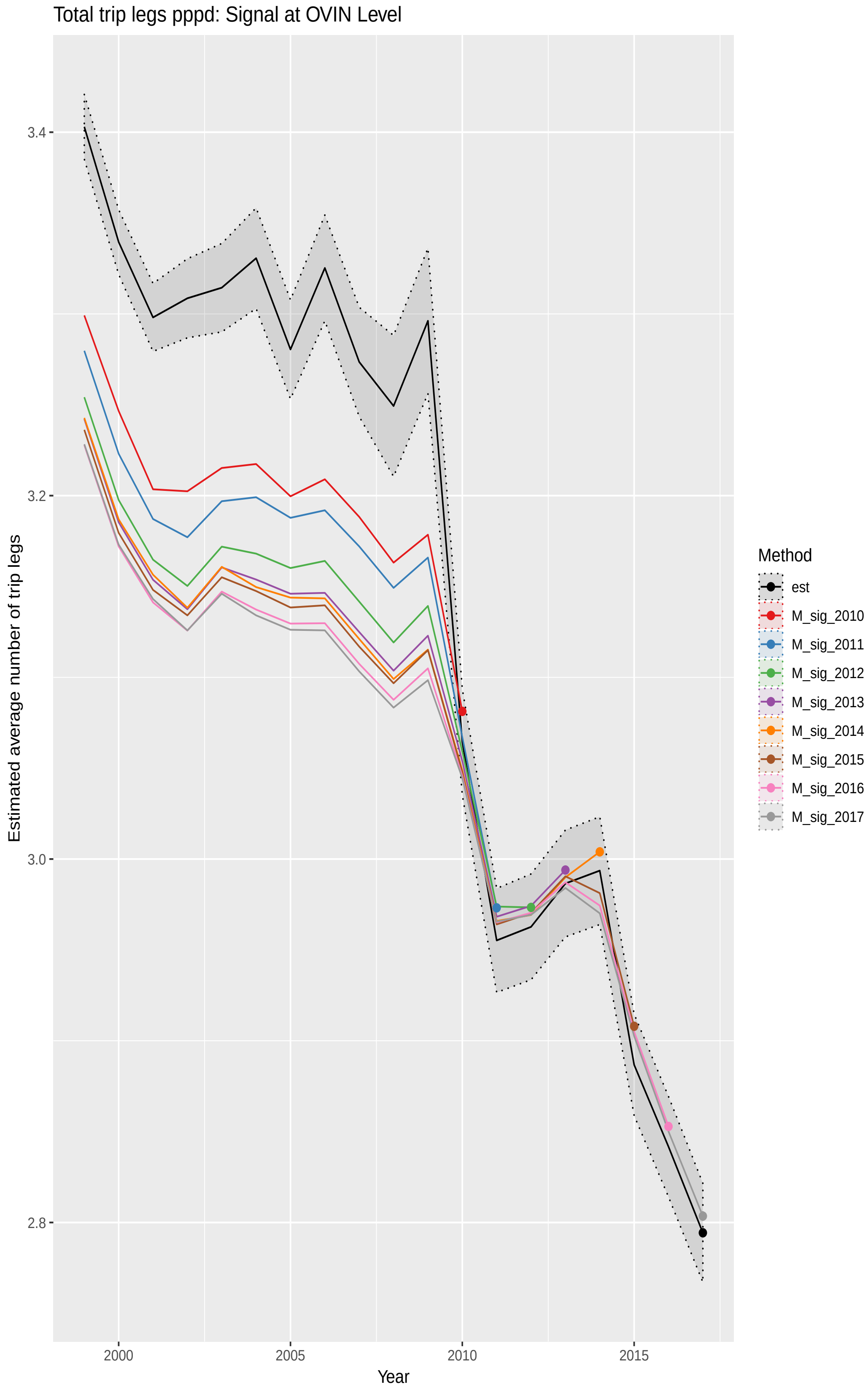

Figure 7.1 Signal of mean number of trip legs for overall level measured at OVIN level

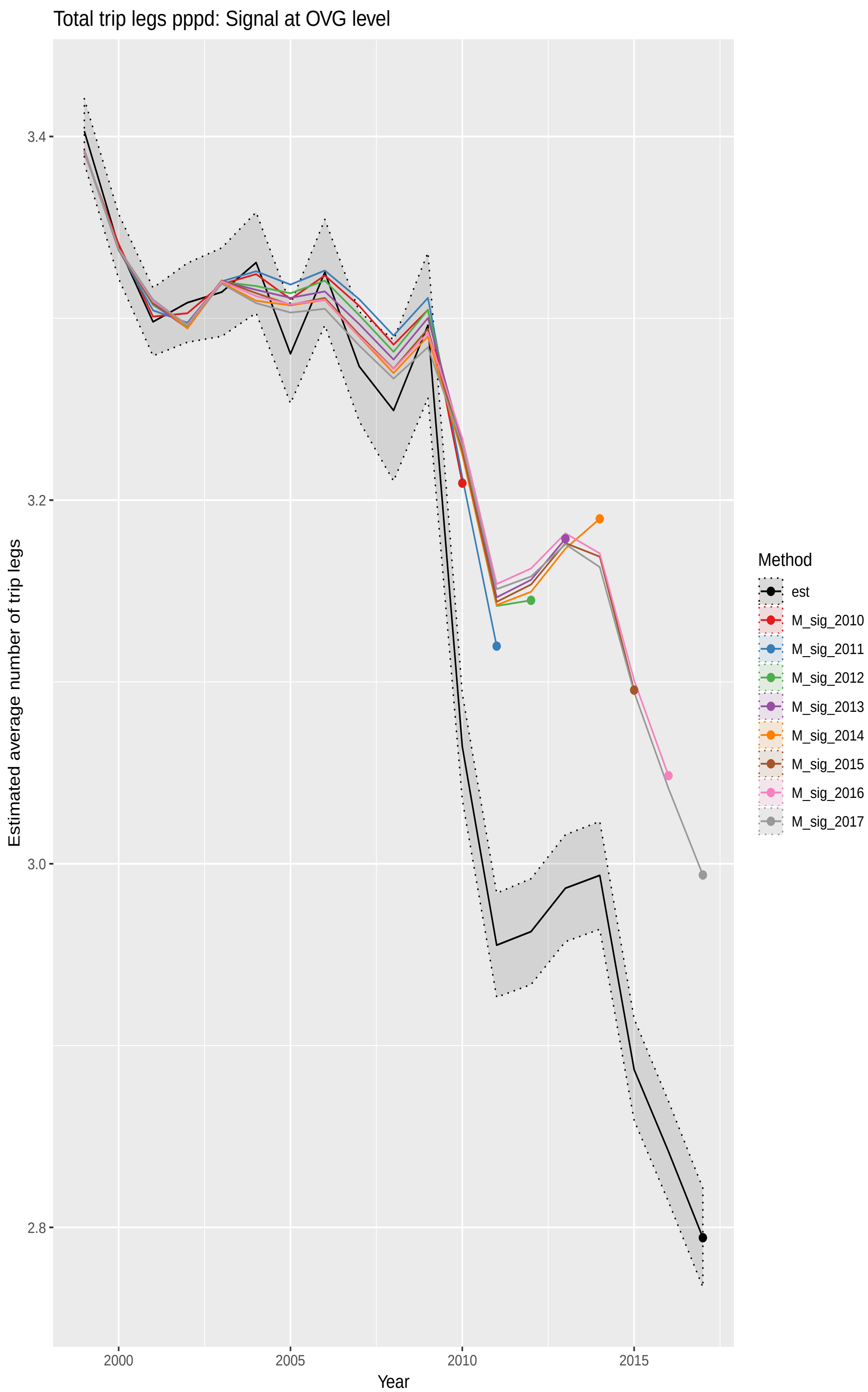

Figure 7.2 Signal of mean number of trip legs for overall level measured at OVG level

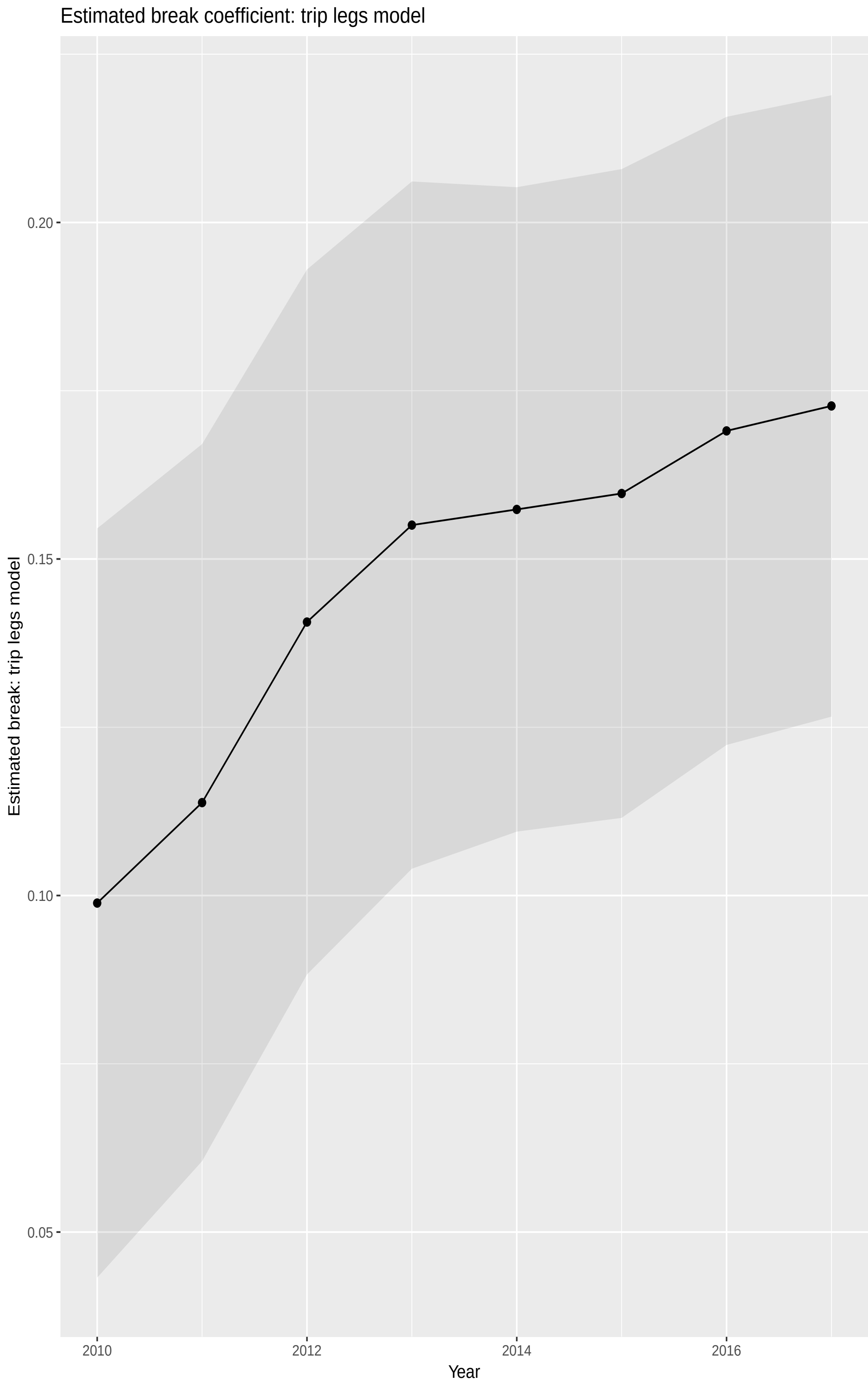

Figure 7.3 In real time estimates from 2010 to 2017 for the discontinuity for number of trip legs at the overall level with approximate 95% confidence interval

7.4 Distance

The distributions of the RB_{it} (24) and $RRSE_{it}$ (25) for the different aggregation levels specified in Section 7.1, are provided in Tables 7.3 and 7.4, respectively. The bias at the highest aggregation level is negligible (average of -0.08%) and gradually increases to an average of -3.15% at the most detailed level. The reduction of the variance is the smallest at the highest aggregation level (36.9% on average) and gradually increases with the level of detail to 58.7% at the most detailed level.

	Variable	Min.	1st.Qu.	Median	Mean	3rd.Qu.	Max.
1	Year	-3.04	-0.61	-0.11	-0.08	0.35	2.92
2	Motive	-0.95	-0.74	-0.36	-0.39	-0.01	0.11
3	Mode	-1.93	-0.37	-0.28	-0.35	-0.05	0.57
4	Motive and Mode	-11.22	-1.60	-0.36	-1.22	-0.03	1.27
5	Sex, Ageclass, Motive, and mode	-57.83	-3.94	-0.58	-3.15	0.37	68.46

Table 7.3 Summary statistics of mean relative bias (in %) at different aggregation levels for the SAE estimates of mean distance per trip legs pppd

	Variable	Min.	1st.Qu.	Median	Mean	3rd.Qu.	Max.
1	Year	14.76	31.67	39.63	36.58	42.26	44.84
2	Motive	36.30	37.16	37.79	39.12	39.75	44.60
3	Mode	6.18	30.61	38.94	35.16	45.64	48.49
4	Motive and Mode	10.42	35.86	44.48	41.61	51.98	58.79
5	Sex, Ageclass, Motive, and mode	-72.52	55.69	60.20	58.70	64.09	73.17

Table 7.4 Summary statistics of relative reduction of standard errors (in %) at different aggregation levels for the standard errors of the SAE estimates of mean distance per trip legs pppd

The size of the revisions is evaluated graphically at the highest aggregation level in Figures 7.4 and 7.5. Figure 7.4 illustrates the revisions estimated in real time for the signals at the OViN level for the period 2010 to 2017. Similar figures for the purpose and mode categories are provided in Boonstra et al. (2019a).

Figure 7.5 illustrates the revisions if trends are published at the OVG level in real time analysis for the period directly after the implementation of the OViN, i.e. 2010 to 2017. Similar figures for the purpose and mode categories are provided in Boonstra et al. (2019a).

In real time, estimates for the discontinuity due to the change-over from MON to OViN are presented in Figure 7.6. This figure illustrates that it is not significantly different from zero.

The revisions of the signals visible in Figures 7.4 and 7.5 are most likely caused by the revisions of the predicted number of trip legs, which are used in the aggregation of the average trip distances. As in the case of the number of trip legs, the revisions for the distance are larger if the predictions are at the OViN level during the OVG and MON period. For the OViN period, it appears that the revisions are slightly larger if the predictions are at the OVG level.

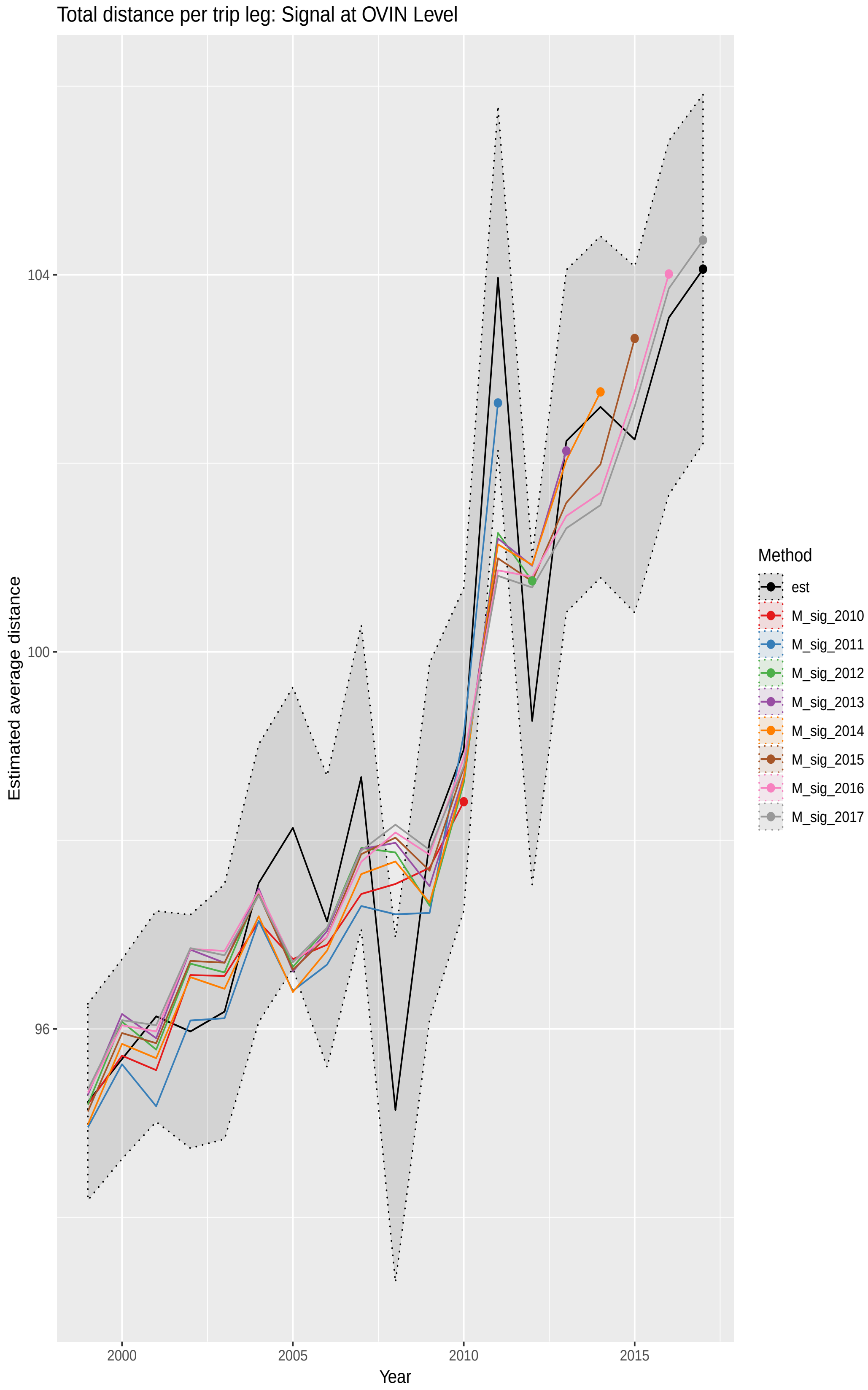

Figure 7.4 Signal of average distance of trip legs for overall level measured at OVIN level

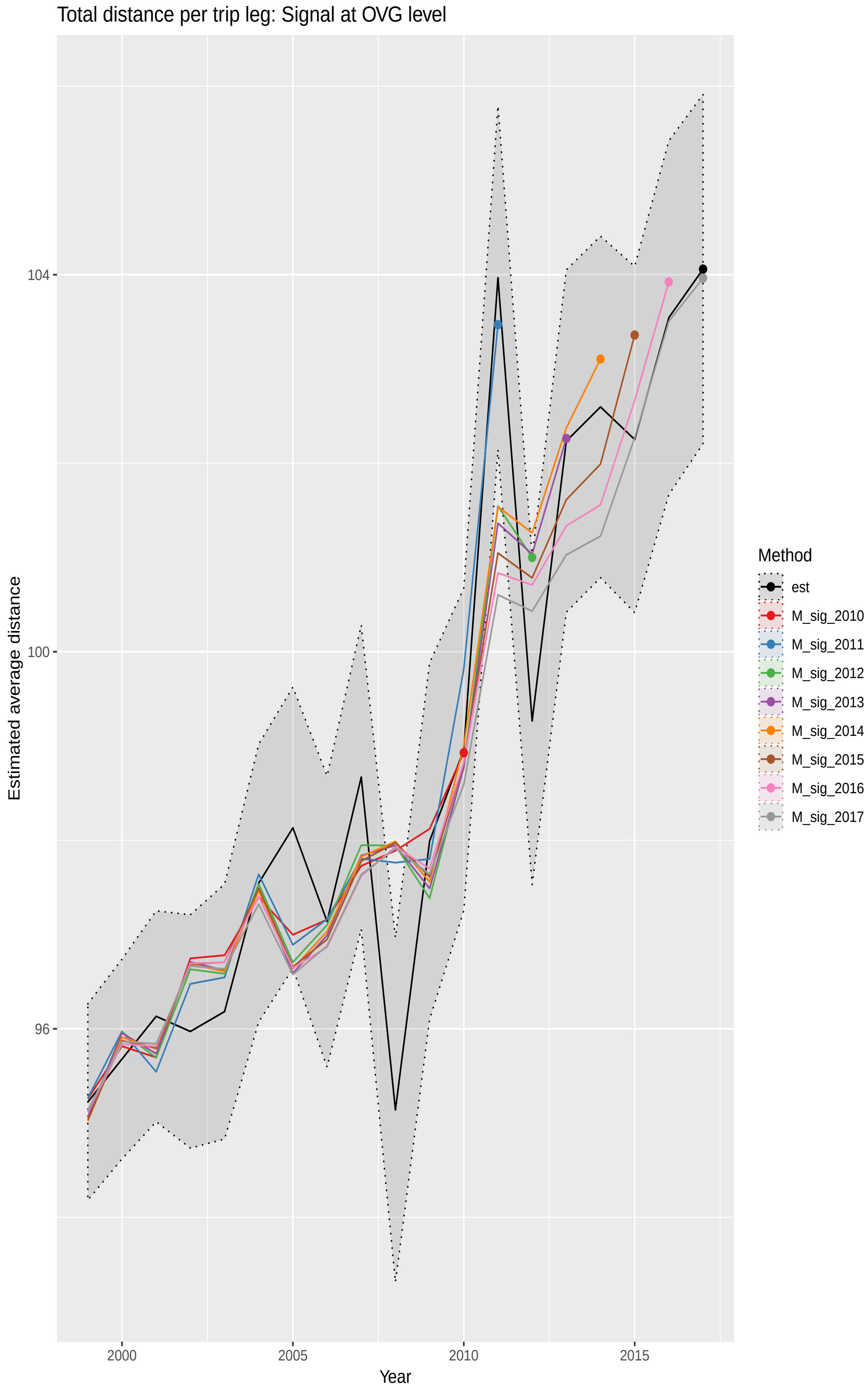

Figure 7.5 Signal of average distance of trip legs for overall level measured at OVG level

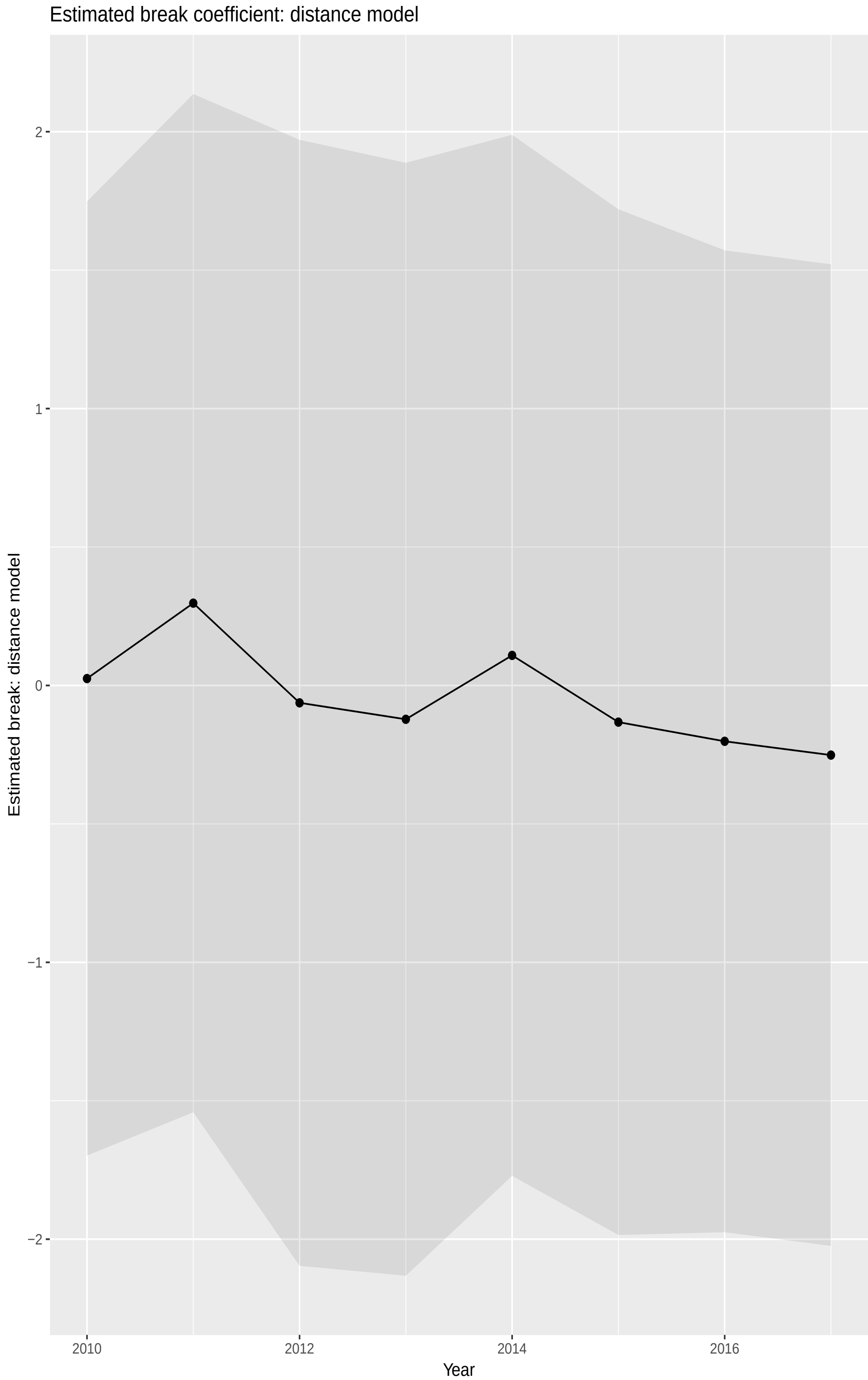

Figure 7.6 In real time estimates from 2010 to 2017 for the discontinuity for average distance of trip legs at the overall level with 95% confidence interval

8 Discussion

In this paper two models are developed for estimating trends for mobility indicators based on the Dutch Travel Survey (DTS). Two target variables are considered: the average number of trip legs per person per day and the average distance per trip leg for domains that are defined by the cross classification of age, gender, purpose, and transportation mode on a yearly frequency.

In a first stage direct estimates are compiled from the DTS data as well as their standard errors using the general regression estimator. The direct estimates are the input for the time series models and are first transformed to better meet normality assumptions. For the number of trip legs a square-root transformation is used and for distance a log transformation. The standard errors are also transformed and subsequently smoothed with a generalized variance function model.

The resulting direct estimates at the level of the aforementioned cross classification are used as input for multilevel time series models, which are fitted using MCMC simulations. The models account for discontinuities due to two redesigns that occurred due to change-over from OVG to MON in 2004 and the change-over from MON to OViN in 2010. Discontinuities are predominantly modeled as random effects. DTS time series are influenced by outliers. The model for trip legs contains random effects to model the most dominant outliers in 2009, while the model for distances assumes a Student-t distribution for the sampling errors. Other important model components are smooth trends at different levels of the cross classification variables. Several auxiliary annual time series are used as covariates: number of snow days for the trip legs model and registered car kilometers for the distance model. For random effects several non-normal priors are considered as a stronger form of regularization.

Model predictions at different aggregation levels of the cross-classification variables are obtained by aggregating the model predictions at the most detailed level. In addition trends at the OViN level are derived by accounting for the MON and OViN discontinuities. Finally model predictions and trends for total number of trip legs per day, distance per person per day, and total distance per day at different aggregation levels are derived from the two fitted models.

Model selection is predominantly based on model information criteria (WAIC and DIC), graphical comparisons in combination with subject matter knowledge about transportation in the Netherlands. The adequacy of the finally selected models is confirmed with a set of model diagnostics. Posterior predictive checks confirm a good model fit for most of the series, since there are no signs of bias, overfitting or underfitting in the majority of the time series. A graphical comparison of the direct estimates with the simulated posterior predictive distributions show that the direct estimates mostly fit well with the simulated distributions.

The standardized residuals of the time series obey the assumed standard normal distribution in the case of trip legs and Student-t distribution in the case of distance. The relative bias statistics illustrate negligible amounts of bias in the model predictions at higher aggregation levels, while the precision of the predictions is improved with 25% to 50% in the case of trip legs and 36% to 60% in the case of distance from the highest aggregation level to the most detailed level.

In real time analysis of the discontinuities due to the implementation of OViN at high aggregation levels illustrate that stable estimates for the discontinuities are obtained in about 4 years in the case of trip legs, while for distance no significant discontinuities are observed. In real time analysis of the signals illustrate that signals are revised since the estimated discontinuities for MON as well as OViN change, directly after the implementation of OViN. It appears that the revisions of the signal predictions are most stable if the survey design that is chosen as publication level is used as the base line in the definitions of the dummy indicators used to model the discontinuities. Once the discontinuity estimates are sufficiently stable, the revisions of the signal predictions are small.

Topics for further research are:

- accounting for correlations between the direct estimates.
- modeling trends for volatile domains. At the moment trends of number of trip legs for cycling and walking both for other purposes are very volatile and might be improved using different type of model components.
- parameterizing of the level interventions. Additional research is required to find the best parameterization of the level interventions for discontinuities. This is necessary to better understand whether it is advisable to change the trend estimates from the OViN to ODiN level, directly after the change-over to ODiN in 2018 or to wait until stable estimates for ODiN discontinuities are obtained.
- updating the trend estimates. The models need to be extended to account for the change-over to ODiN in 2018 for updating the trend estimates.

Acknowledgement

We wish to thank Rob Willems, Marc Smeets, Coen van Heukelingen, Remko Smit, Jan van der Waard and Hans Wüst for their comments and suggestions that helped improve this report.

References

Bollineni-Balabay, O., J. van den Brakel, F. Palm, and H. J. Boonstra (2017). Multilevel hierarchical bayesian versus state space approach in time series small area estimation: the dutch travel survey. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 180*(4), 1281–1308.

Boonstra, H. J. (2018). *mcmcsae: MCMC Small Area Estimation*. R package version 0.9.

Boonstra, H. J. and J. van den Brakel (2016). Estimation of level and change for unemployment using multilevel and structural time series models. Technical Report 201610, https://www.cbs.nl/nl-nl/achtergrond/2016/37/estimation-of-level-and-change-for-unemployment, Statistics Netherlands.

Boonstra, H. J. and J. van den Brakel (2018). Hierarchical bayesian time series multilevel models for consistent small area estimates at different frequencies and regional levels. Statistics Netherlands discussion paper, December 4, 2018.

Boonstra, H. J., J. van den Brakel, and S. Das (2018). Computing input estimates for time series modeling of mobility trends. CBS report.

Boonstra, H. J., J. van den Brakel, and S. Das (2019a). Multi-level time series modeling of mobility trends; model evaluation. CBS report.

Boonstra, H. J., J. van den Brakel, and S. Das (2019b). Multi-level time series modeling of mobility trends; results for distance. CBS report.

Boonstra, H. J., J. van den Brakel, and S. Das (2019c). Multi-level time series modeling of mobility trends; results for trip legs. CBS report.

Brezger, A., L. Fahrmeir, and A. Hennerfeind (2007). Adaptive gaussian markov random fields with applications in human brain mapping. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 56*(3), 327–345.

Carter, C. K. and R. Kohn (1996). Markov chain monte carlo in conditionally gaussian state space models. *Biometrika 83*(3), 589–601.

Carvalho, C. M., N. G. Polson, and J. G. Scott (2010). The horseshoe estimator for sparse signals. *Biometrika 97*(2), 465–480.

Datta, G. S. and P. Lahiri (1995). Robust hierarchical bayes estimation of small area characteristics in the presence of covariates and outliers. *Journal of Multivariate Analysis 54*(2), 310–328.

Fabrizi, E., M. R. Ferrante, and C. Trivisano (2018). Bayesian small area estimation for skewed business survey variables. *Journal of the Royal Statistical Society Series C 67*(4), 861–879.

Fabrizi, E. and C. Trivisano (2010). Robust linear mixed models for small area estimation. *Journal of Statistical Planning and Inference 140*(2), 433–443.

Fay, R. and R. Herriot (1979). Estimates of income for small places: An application of james-stein procedures to census data. *Journal of the American Statistical Association 74*(366), 269–277.

Gelfand, A. and A. Smith (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association 85*, 398–409.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis 1*(3), 515–533.

Gelman, A. and J. Hill (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, A. and D. Rubin (1992). Inference from iterative simulation using multiple sequences. *Statistical Science 7*(4), 457–472.

Geman, S. and D. Geman (1984). Stochastic relaxation, gibbs distributions and the bayesian restoration of images. *IEEE Trans. Pattn Anal. Mach. Intell. 6*, 721–741.

Juárez, M. A. and M. F. J. Steel (2010). Model-based clustering of non-gaussian panel data based on skew-t distributions. *Journal of Business and Economic Statistics 28*(1), 52–66.

Konen, R. and H. Molnár (2007). Onderzoek verplaatsingsgedrag - methodologische beschrijving. Statistics Netherlands.

Lang, S., E.-M. Fronk, and L. Fahrmeir (2002). Function estimation with locally adaptive dynamic models. *Computational Statistics 17*(4), 479–500.

Molnár, H. (2007). Mobiliteitsonderzoek nederland - methodologische beschrijving. Statistics Netherlands.

O'Malley, A. and A. Zaslavsky (2008). Domain-level covariance analysis for multilevel survey data with structured nonresponse. *Journal of the American Statistical Association 103*(484), 1405–1418.

Park, T. and G. Casella (2008). The bayesian lasso. *Journal of the American Statistical Association 103*(482), 681–686.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rao, J. and I. Molina (2015). *Small Area Estimation*. Wiley-Interscience.

Rue, H. and L. Held (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman and Hall/CRC.

Sakia, R. (1992). The box-cox transformation technique: a review. *The Statistician 41*, 169–178.

Särndal, C.-E., B. Swensson, and J. Wretman (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika 76*, 527–537.

Särndal, C.-E., B. Swensson, and J. Wretman (1992). *Model Assisted Survey Sampling*. Springer.

Spiegelhalter, D., N. Best, B. Carlin, and A. van der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B 64*(4), 583–639.

Tang, X., M. Ghosh, N. S. Ha, and J. Sedransk (2018). Modeling random effects using global–local shrinkage priors in small area estimation. *Journal of the American Statistical Association*, 1–14.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

van den Brakel, J., G. Griffiths, T. Surzhina, P. Wise, J. Blanchard, X. Zhang, and O. Honchar (2017). A framework for measuring the impact of transitions in official statistics. ABS research paper 1351.055.158, Australian Bureau of Statistics.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research 11*, 3571–3594.

Watanabe, S. (2013). A widely applicable bayesian information criterion. *Journal of Machine Learning Research 14*, 867–897.

West, M. (1984). Outlier models and prior distributions in bayesian linear regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 431–439.

Willems, R. and J. van den Brakel (2015). Methodebreukcorrectie ovin. PPM 210514/12, Statistics Netherlands.

Wolter, K. (2007). *Introduction to Variance Estimation*. Springer.

Wüst, H. (2017). Kim-methodiek voor trendanalyses op basis van ovg-mon-ovin-gegevens. Werkdocument 9 februari 2017, Kennisinstituut voor Mobiliteitsbeleid, Den Haag.

Appendix

A Total number of trip legs per day

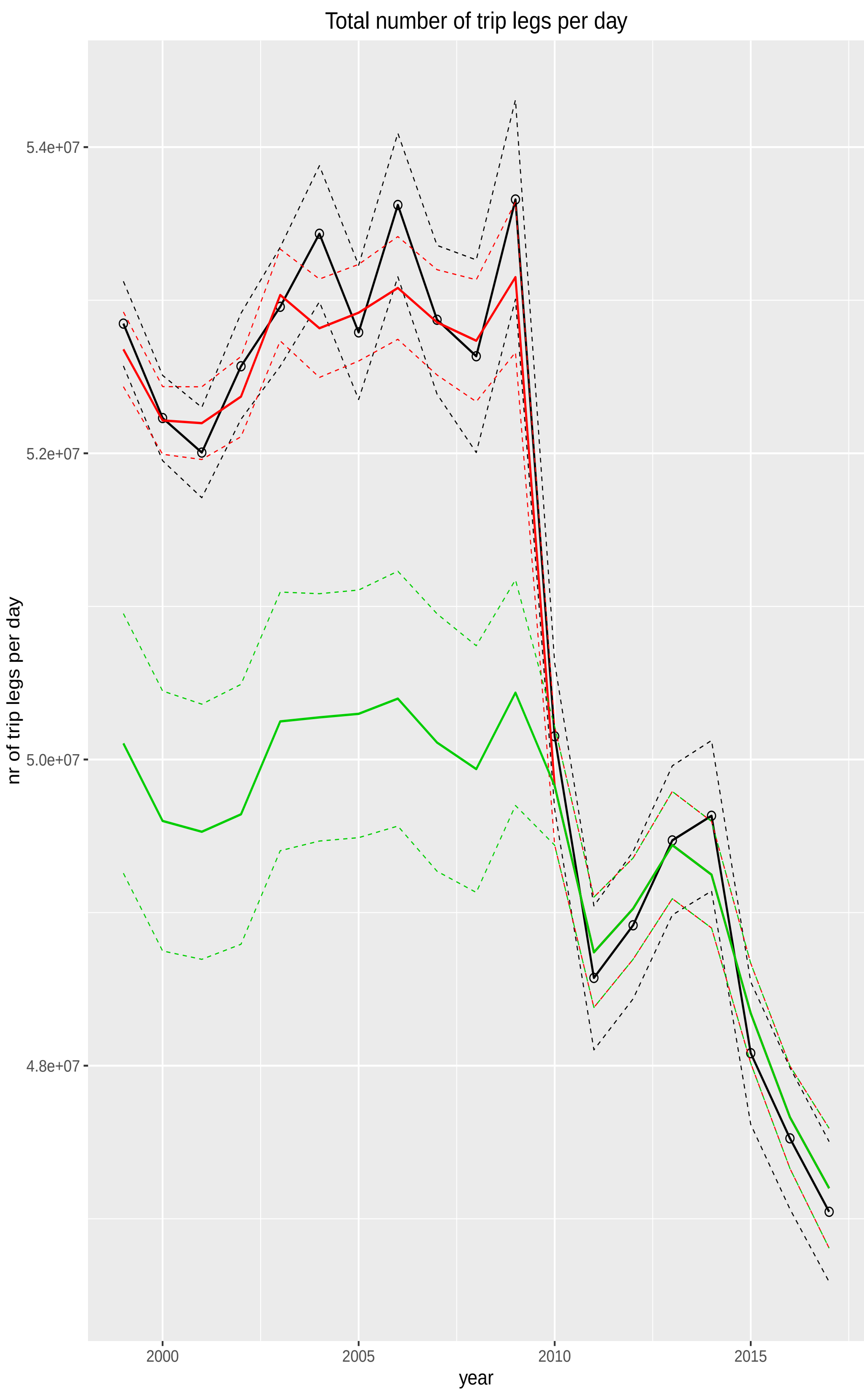

Figure A.1 Direct estimates (black), model fit (red) and trend estimates (green) with approximate 95% intervals.

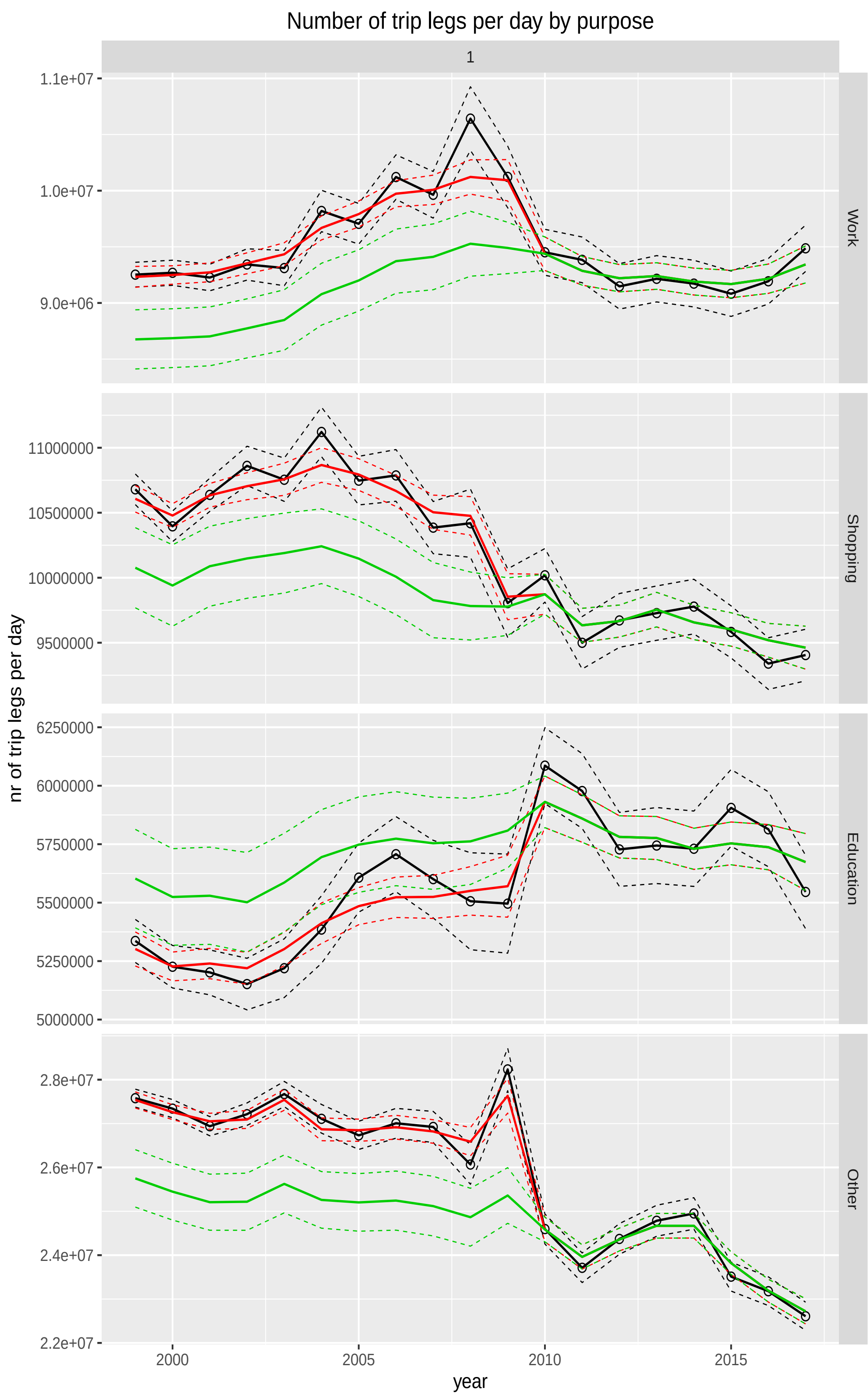

Figure A.2 Direct estimates (black), model fit (red) and trend estimates (green) with approximate 95% intervals.

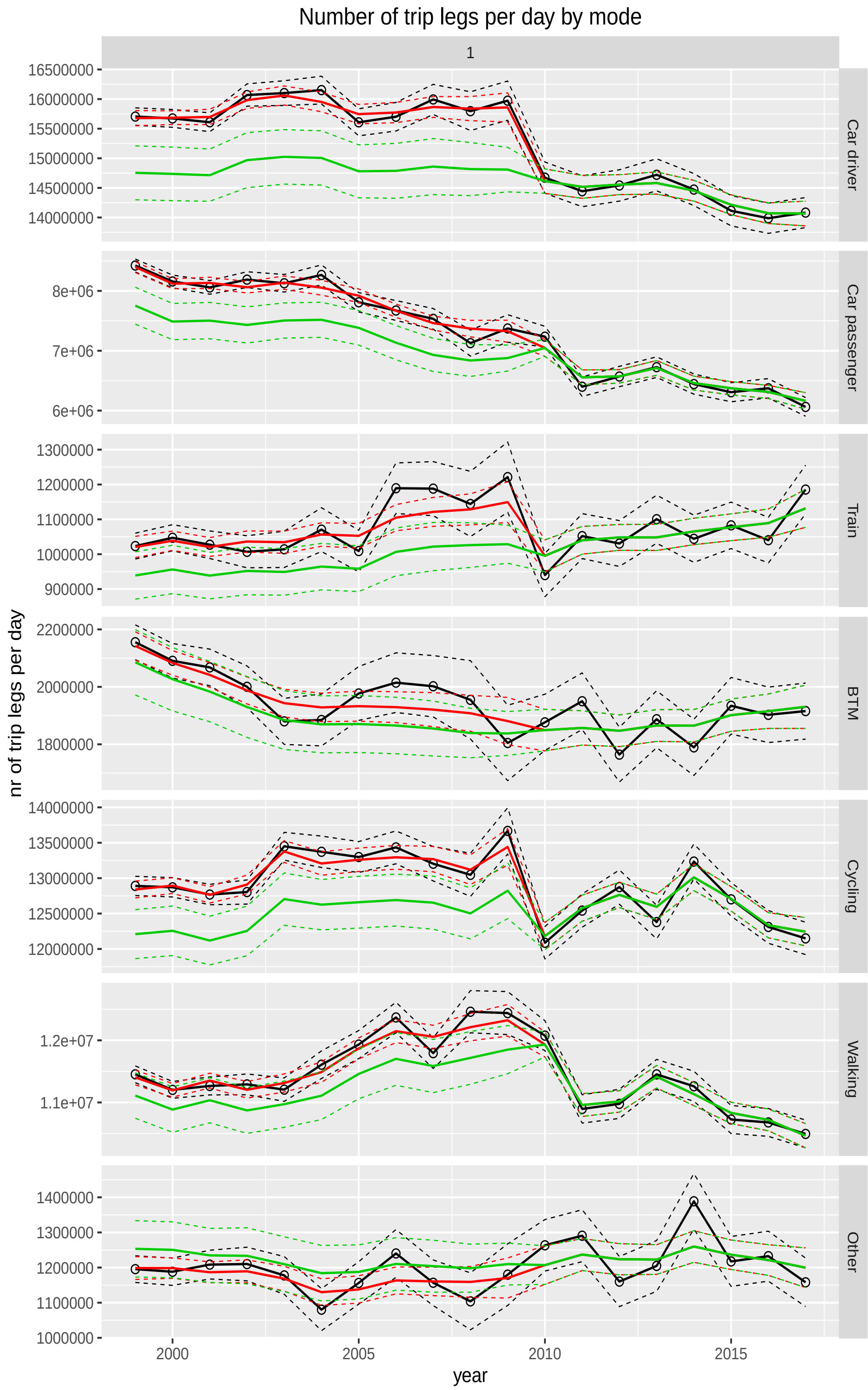

Figure A.3 Direct estimates (black), model fit (red) and trend estimates (green) with approximate 95% intervals.

B Number of trip legs per person per day

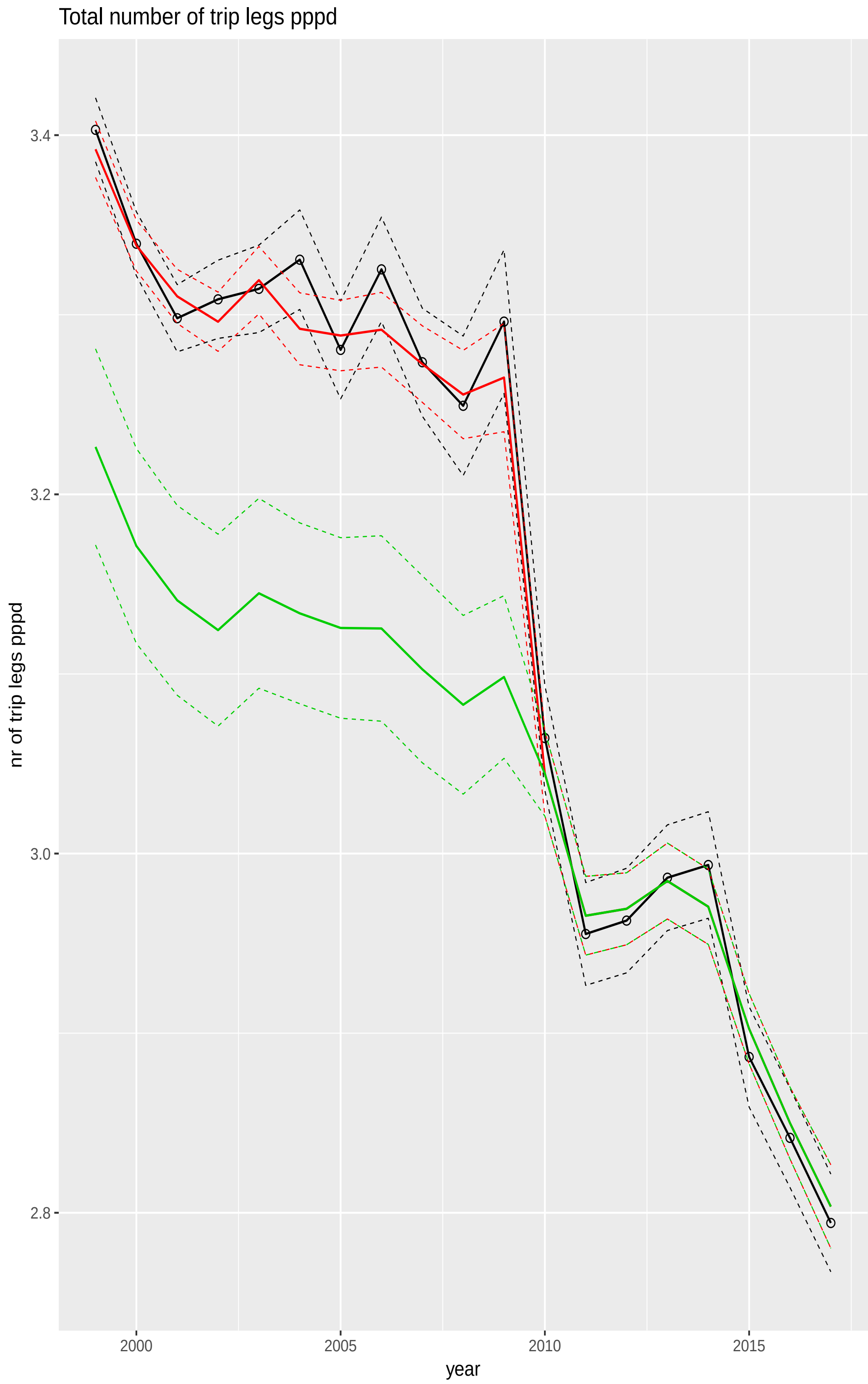

Figure B.1 Direct estimates (black), model fit (red) and trend estimates (green) with approximate 95% intervals.

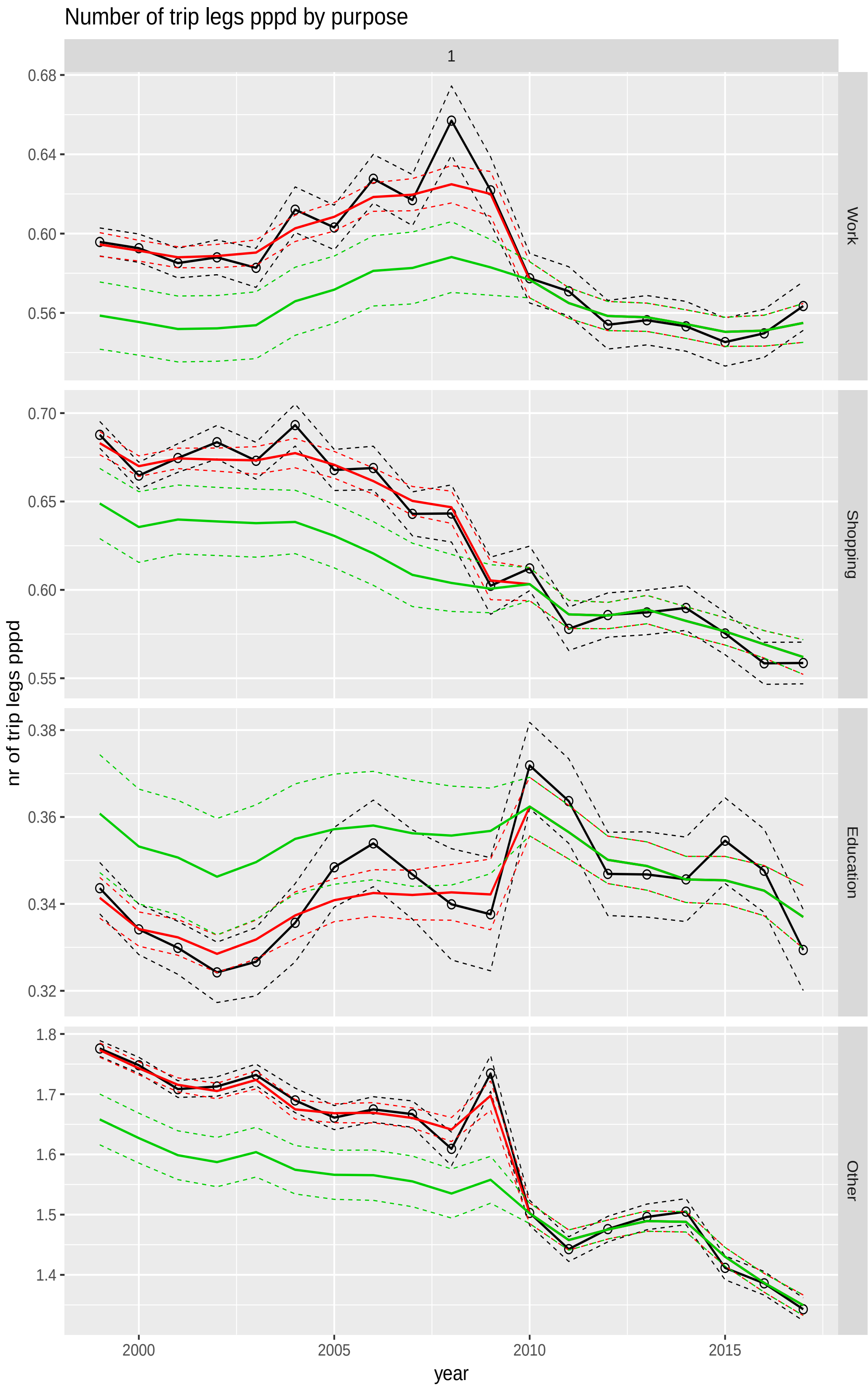

Figure B.2 Direct estimates (black), model fit (red) and trend estimates (green) with approximate 95% intervals.

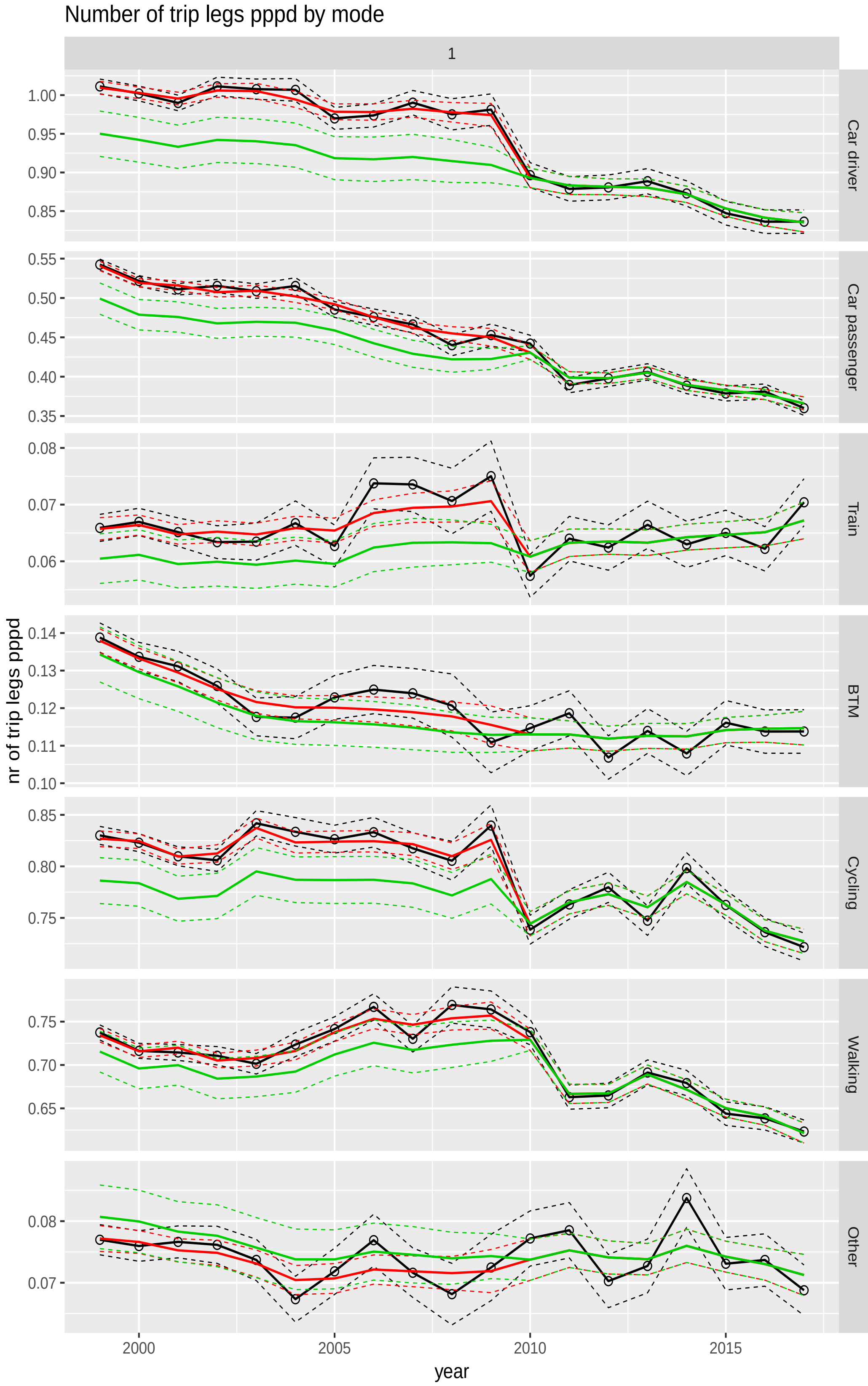

Figure B.3 Direct estimates (black), model fit (red) and trend estimates (green) with approximate 95% intervals.

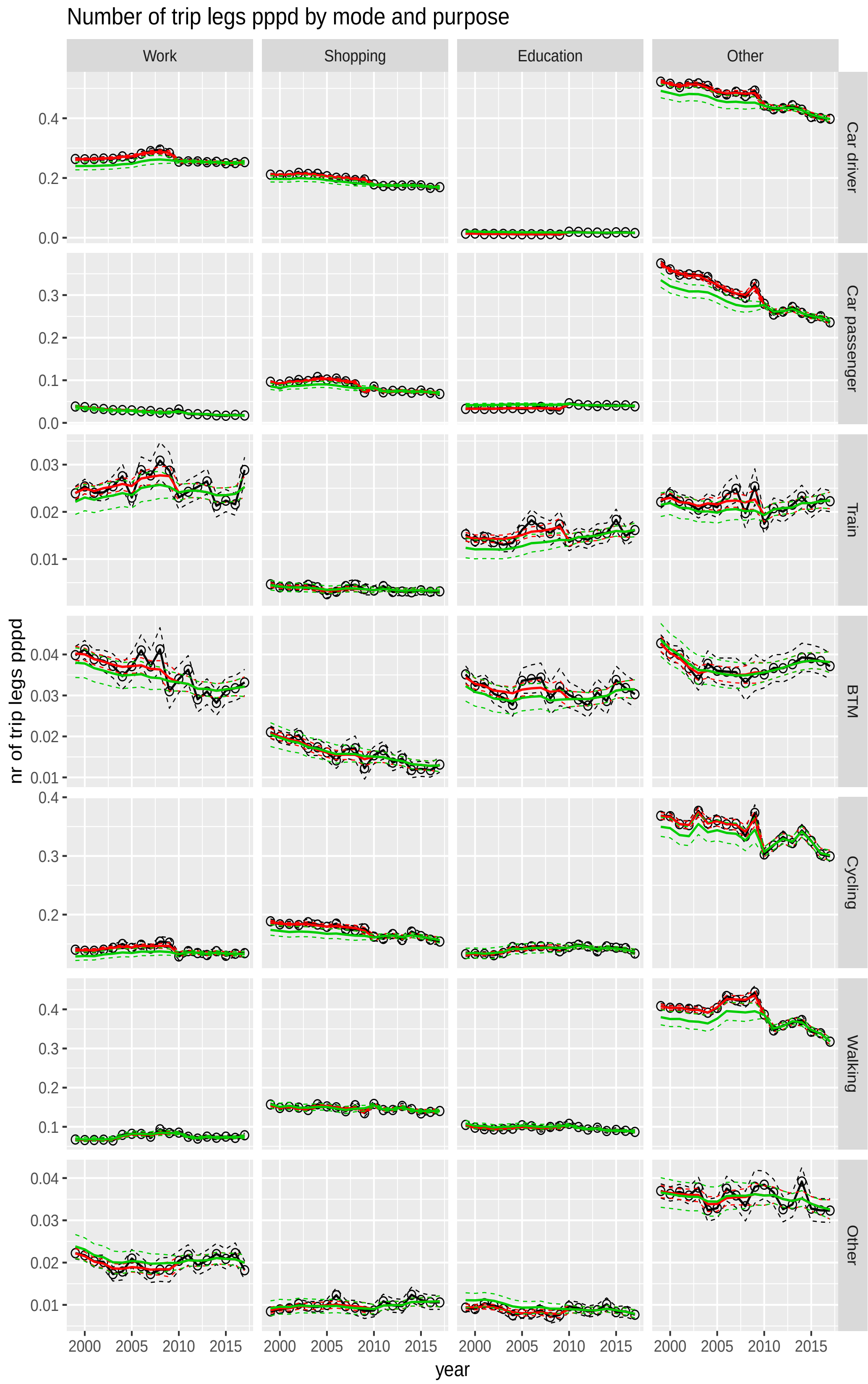

Figure B.4 Direct estimates (black), model fit (red) and trend estimates (green) with approximate 95% intervals.

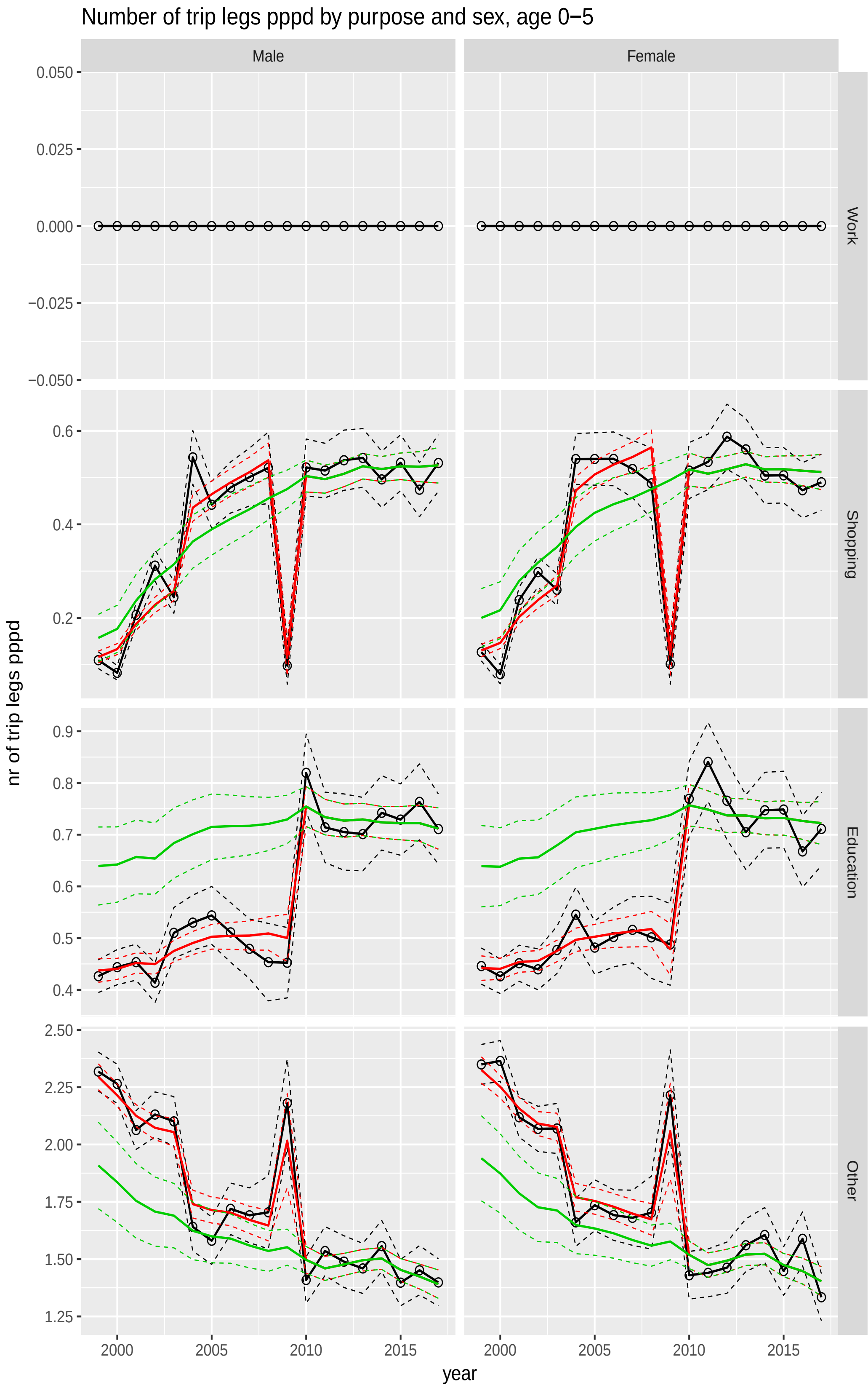

Figure B.5 Direct estimates (black), model fit (red) and trend estimates (green) with approximate 95% intervals.

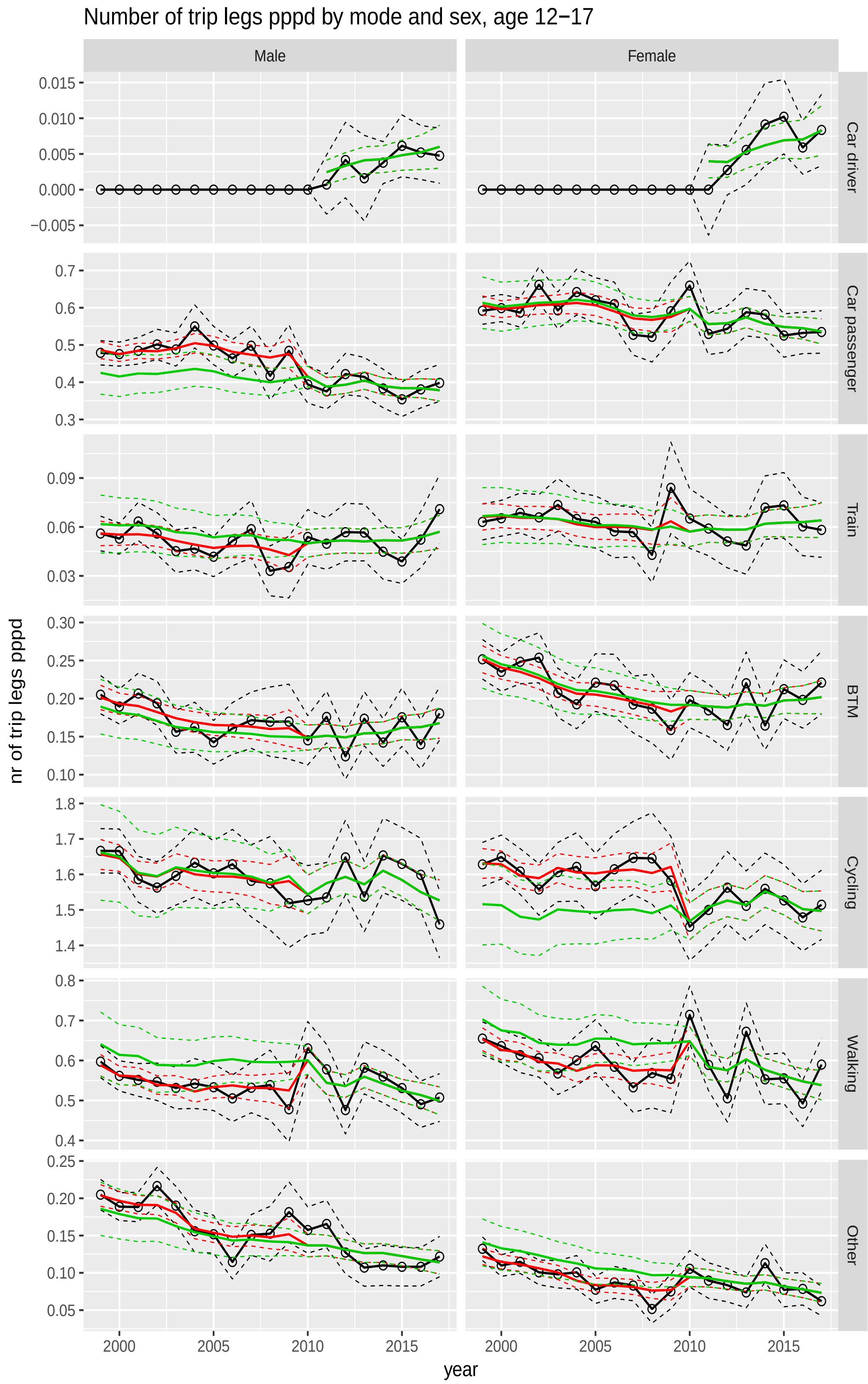

Figure B.6 Direct estimates (black), model fit (red) and trend estimates (green) with approximate 95% intervals.

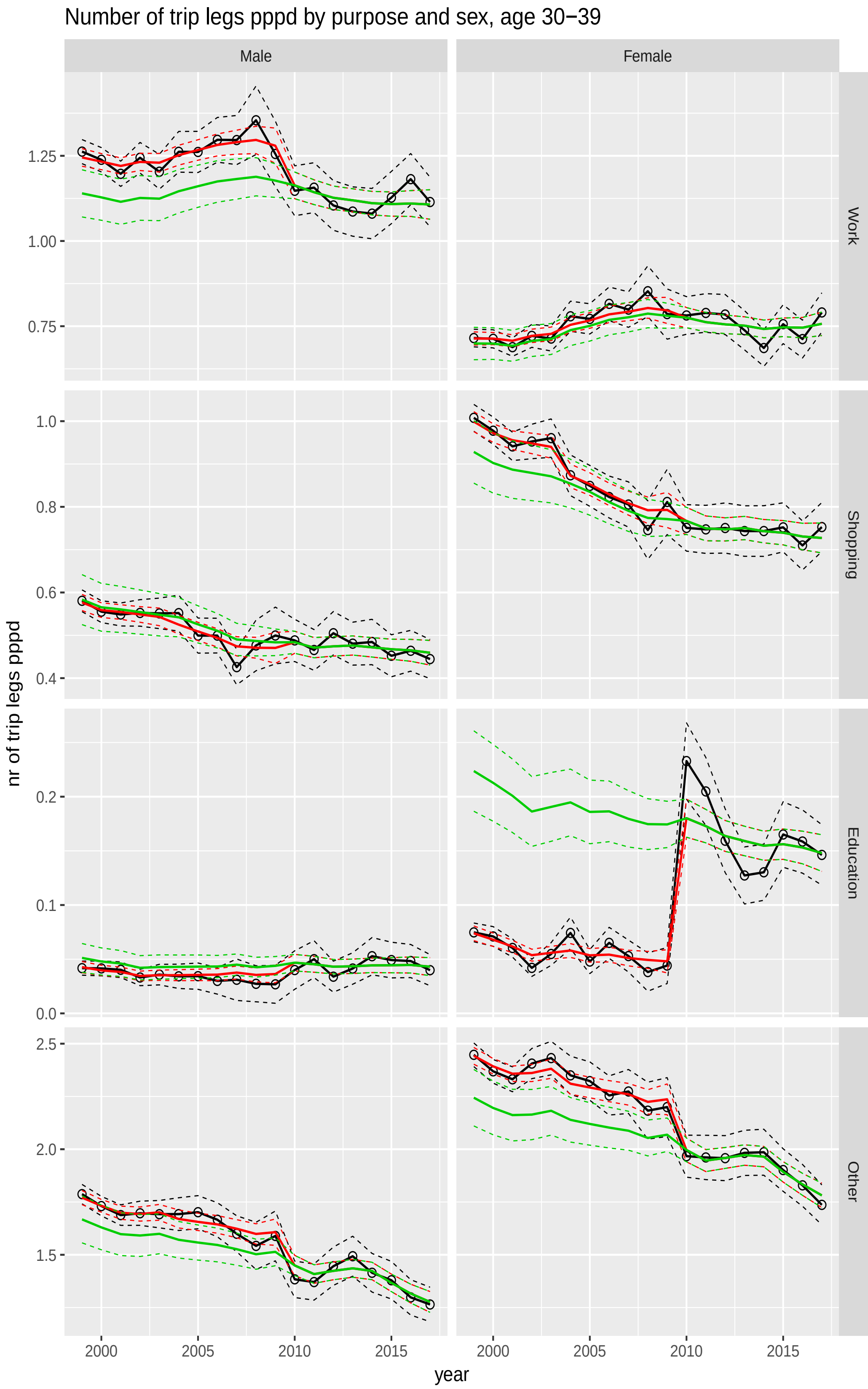

Figure B.7 Direct estimates (black), model fit (red) and trend estimates (green) with approximate 95% intervals.

C Total distance per day

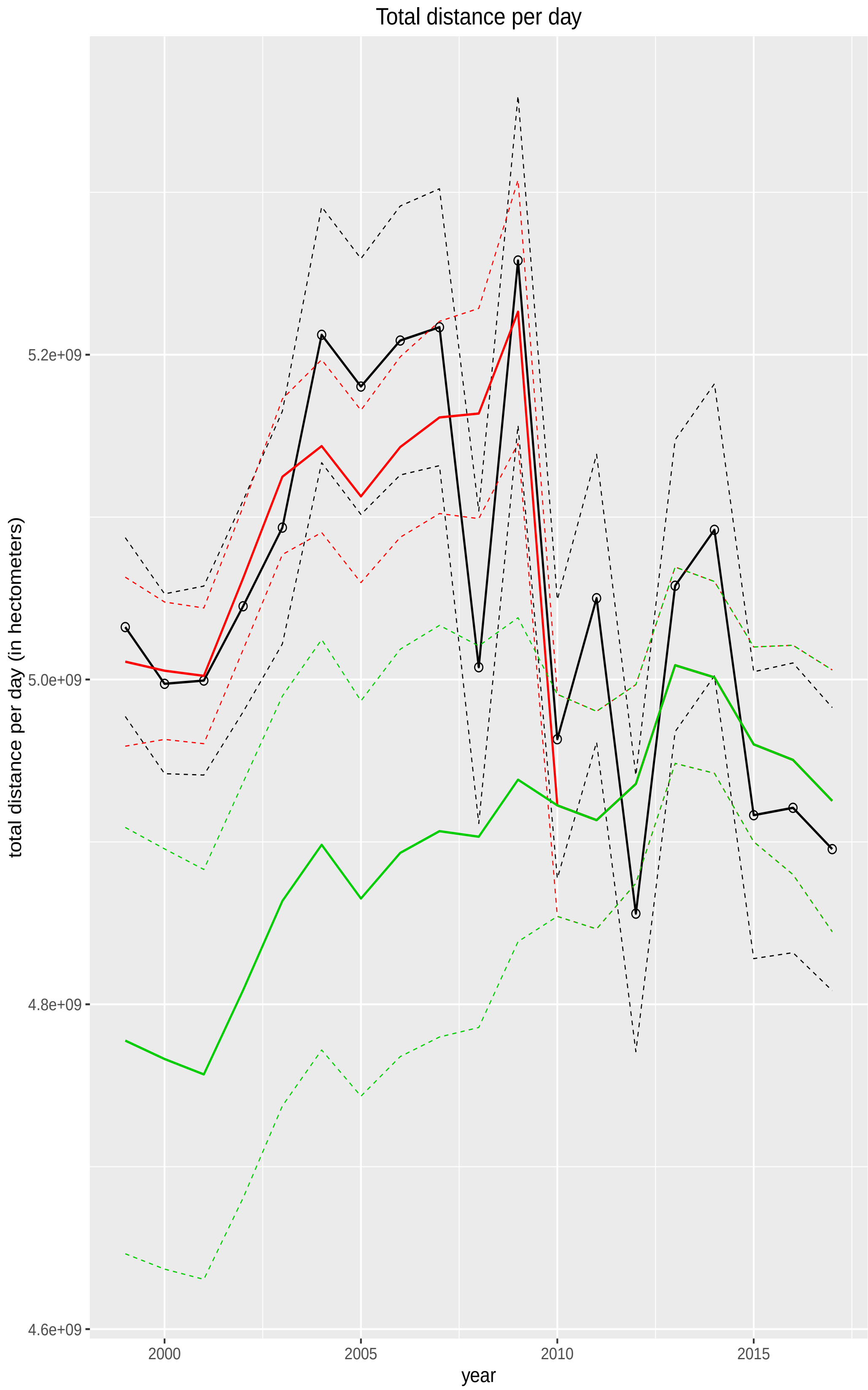

Figure C.1 Direct estimates (black), model fit (red) and trend estimates (green) with approximate 95% intervals.

Figure C.2 Direct estimates (black), model fit (red) and trend estimates (green) with approximate 95% intervals.

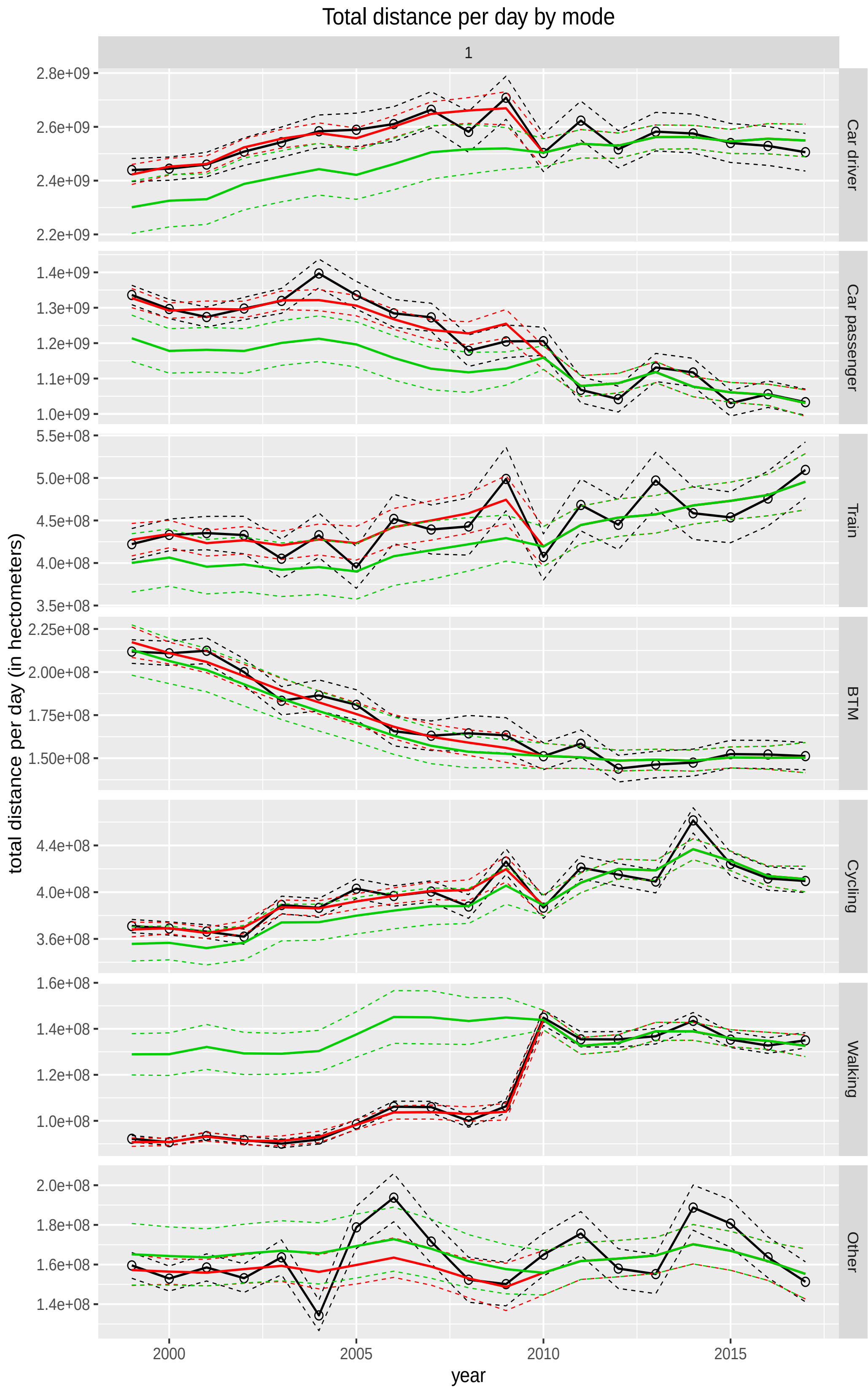

Figure C.3 Direct estimates (black), model fit (red) and trend estimates (green) with approximate 95% intervals.

D Total distance per person per day

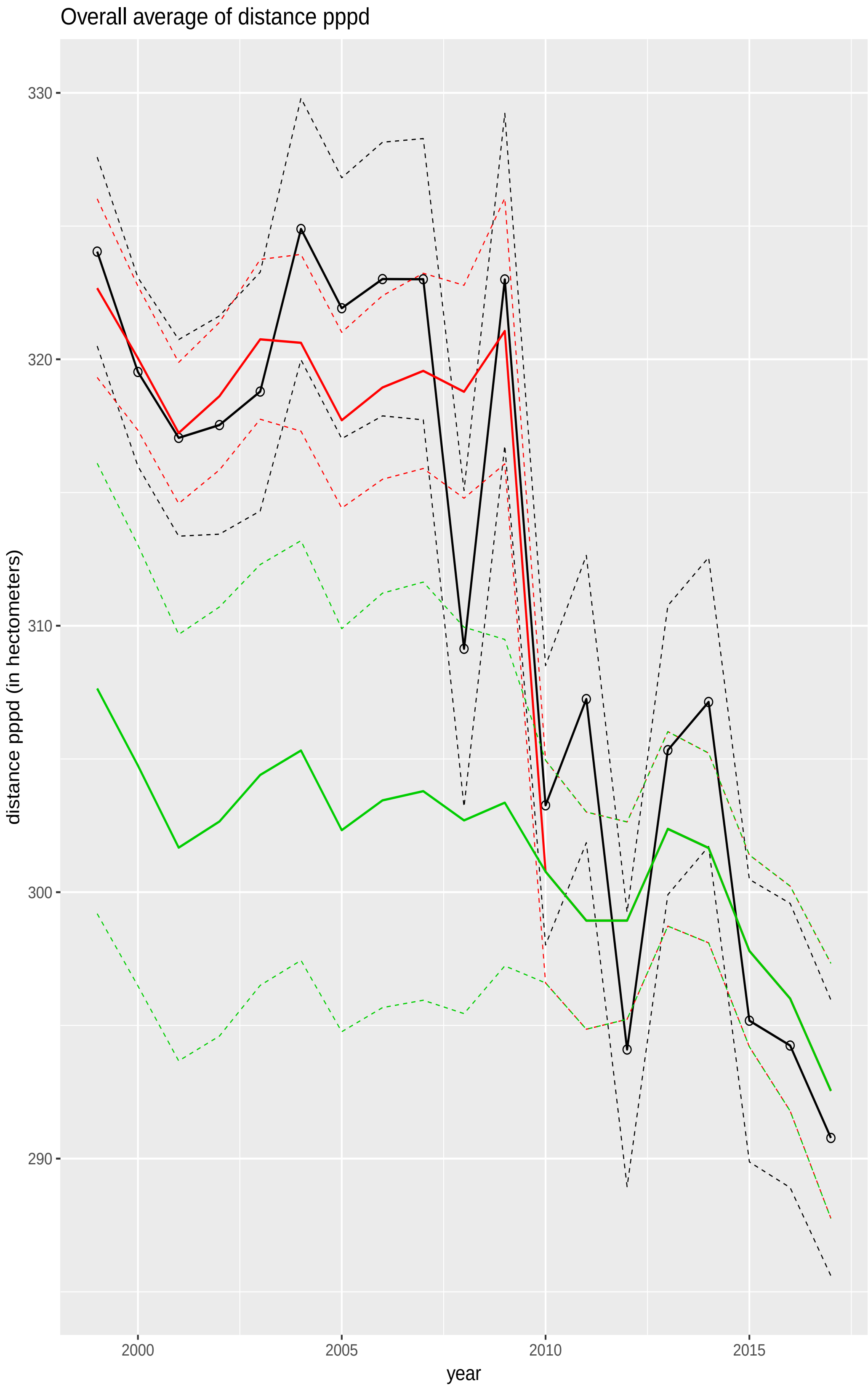

Figure D.1 Direct estimates (black), model fit (red) and trend estimates (green) with approximate 95% intervals.

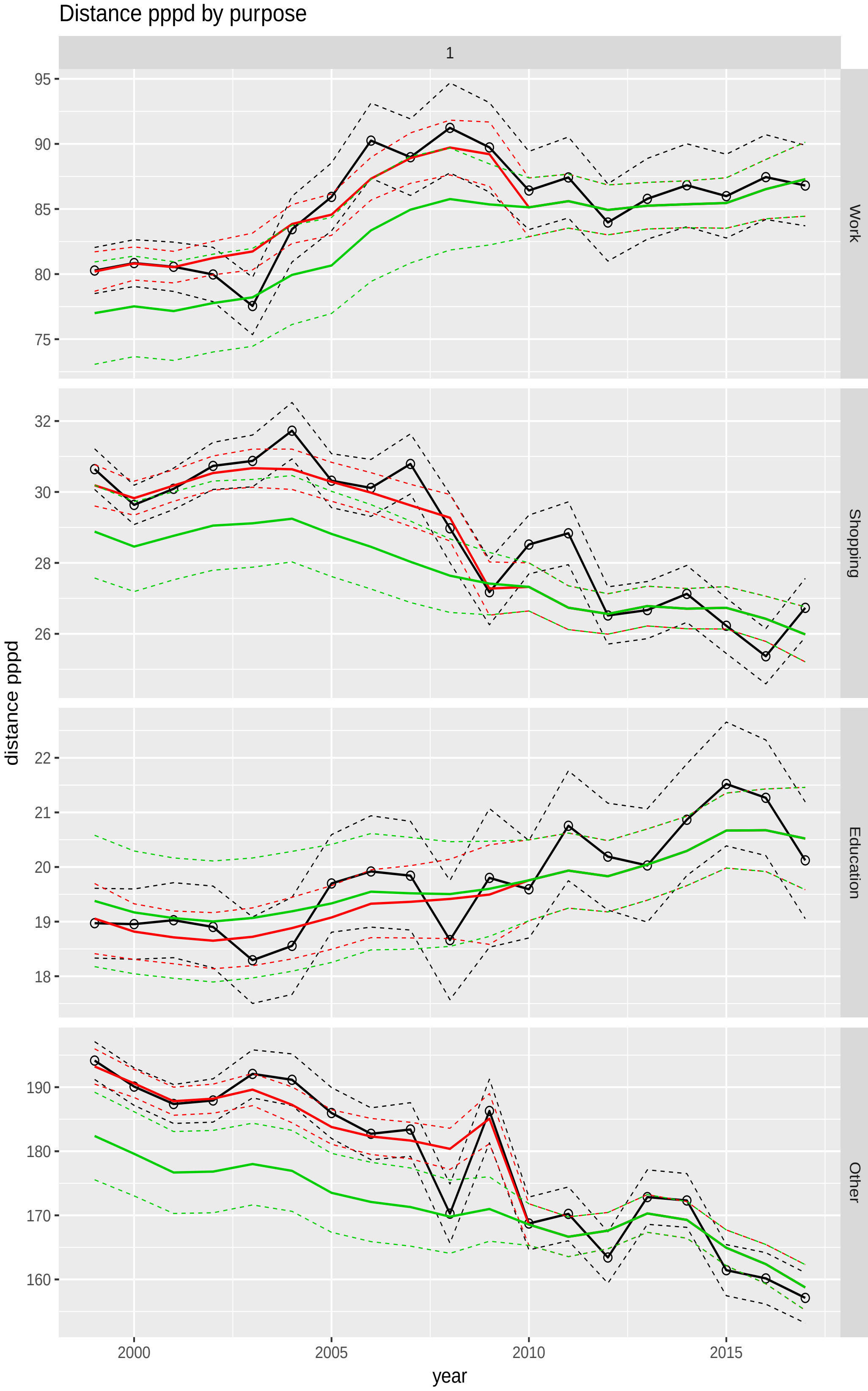

Figure D.2 Direct estimates (black), model fit (red) and trend estimates (green) with approximate 95% intervals.

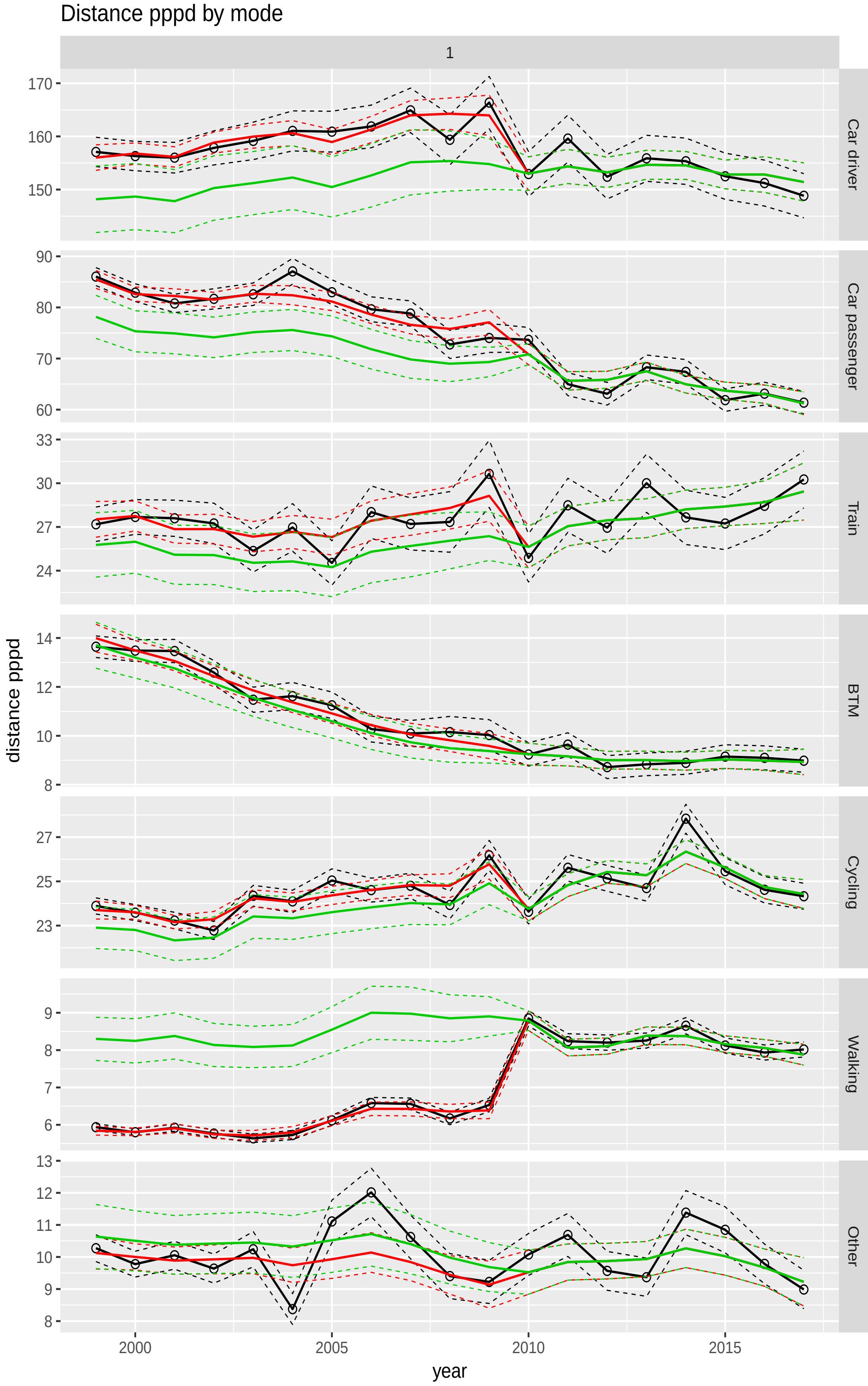

Figure D.3 Direct estimates (black), model fit (red) and trend estimates (green) with approximate 95% intervals.

E Distance per trip leg

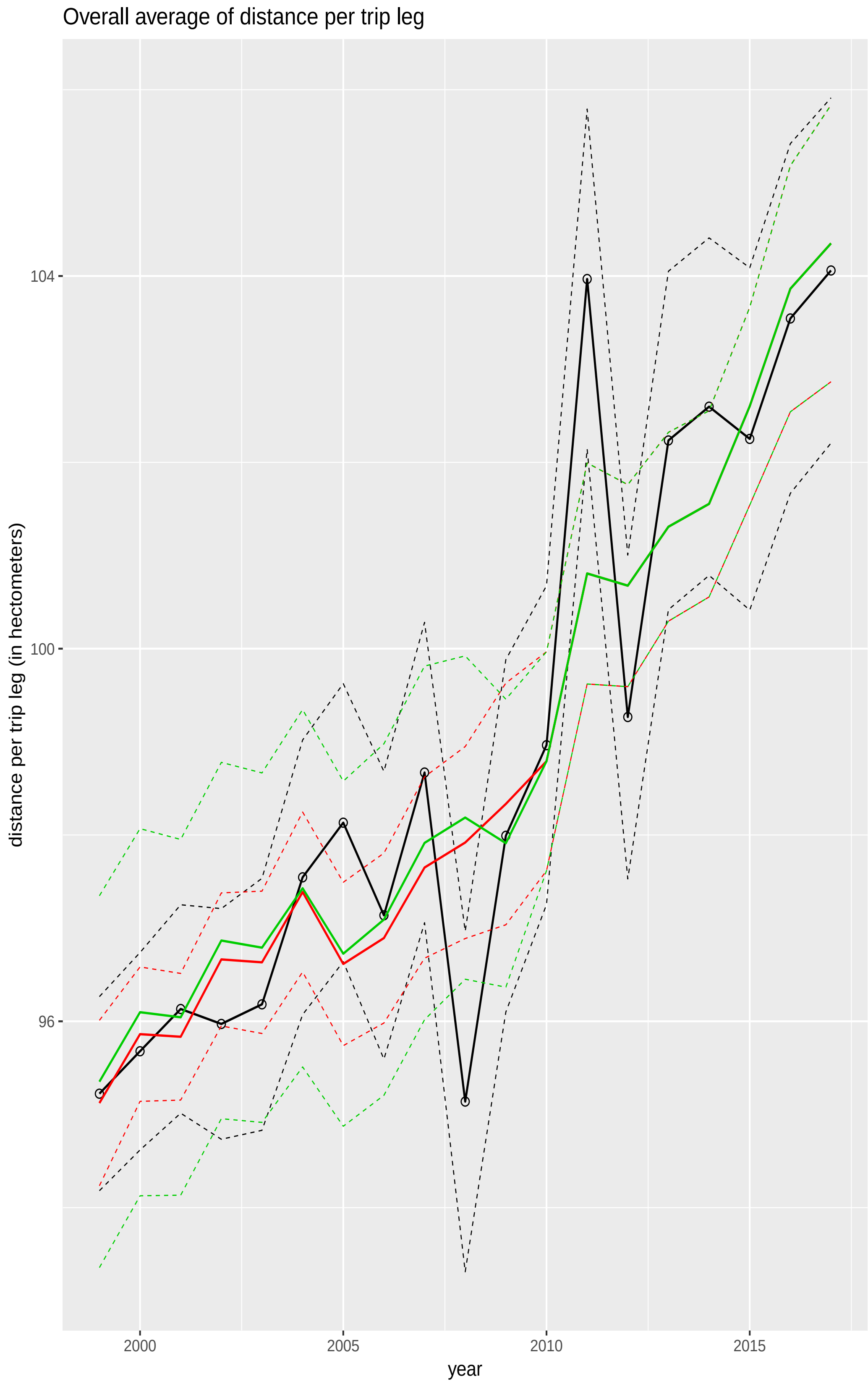

Figure E.1 Direct estimates (black), model fit (red) and trend estimates (green) with approximate 95% intervals.

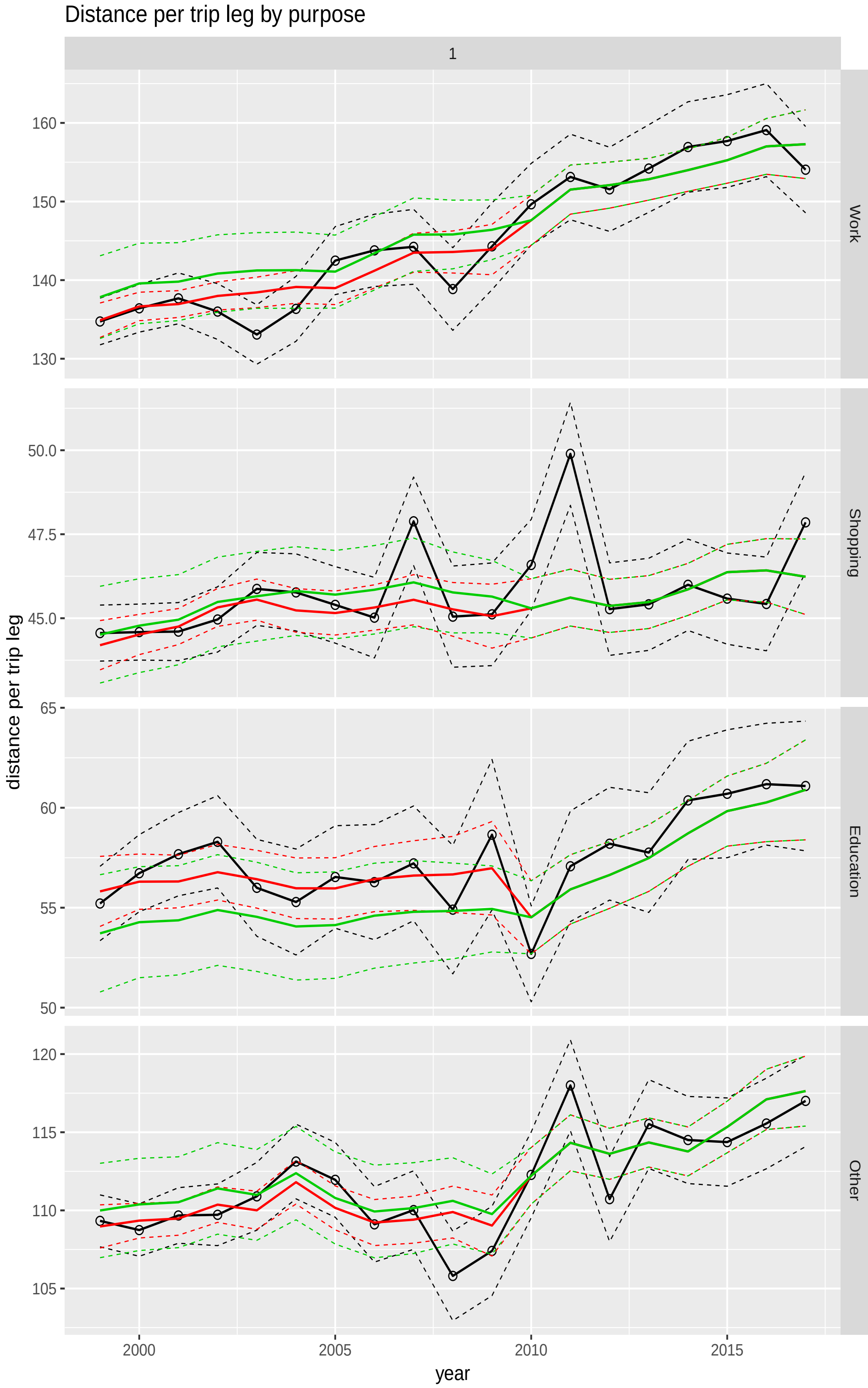

Figure E.2 Direct estimates (black), model fit (red) and trend estimates (green) with approximate 95% intervals.

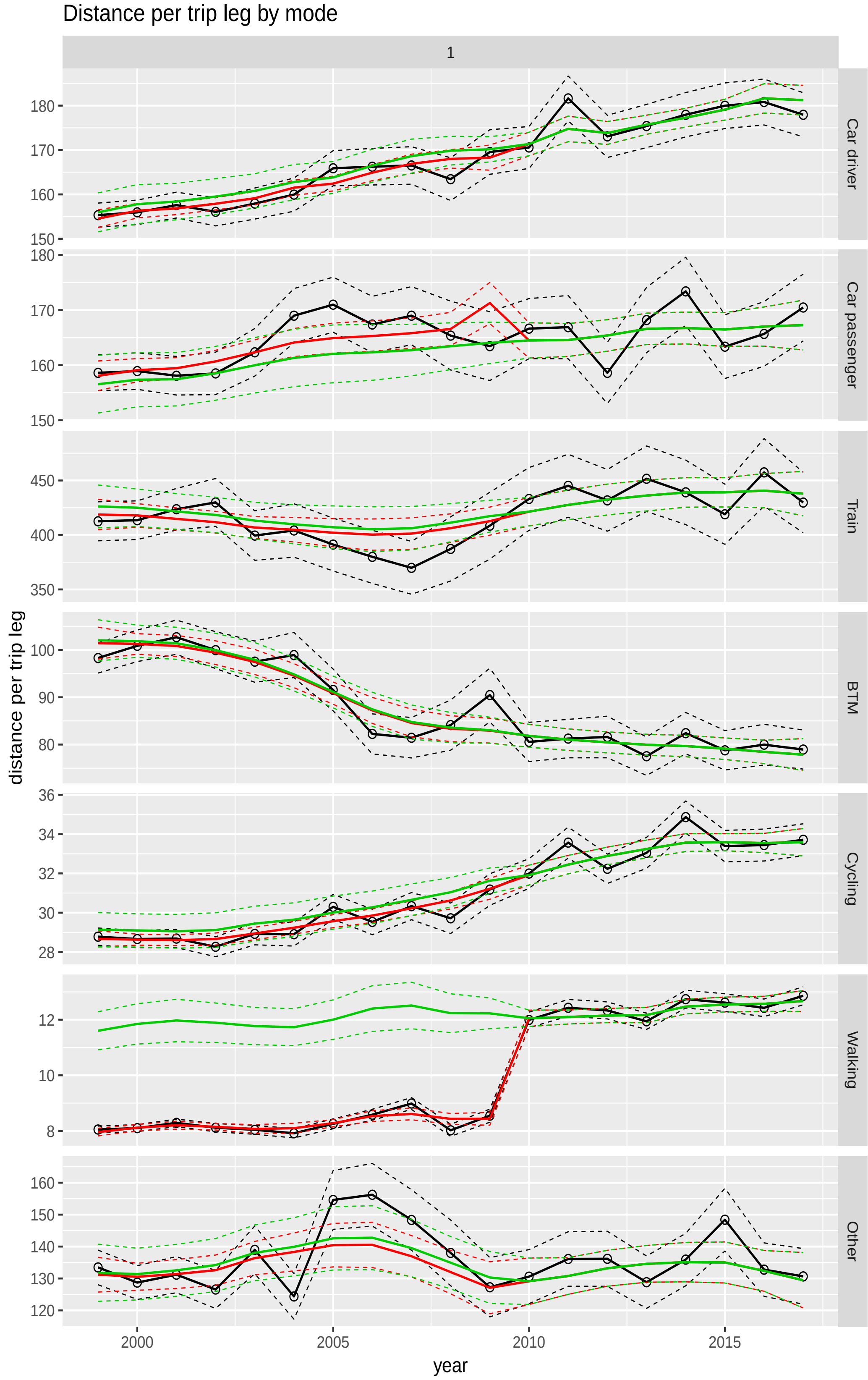

Figure E.3 Direct estimates (black), model fit (red) and trend estimates (green) with approximate 95% intervals.

Colophon

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

Prepress
Statistics Netherlands, Grafimedia

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contact form: www.cbs.nl/information