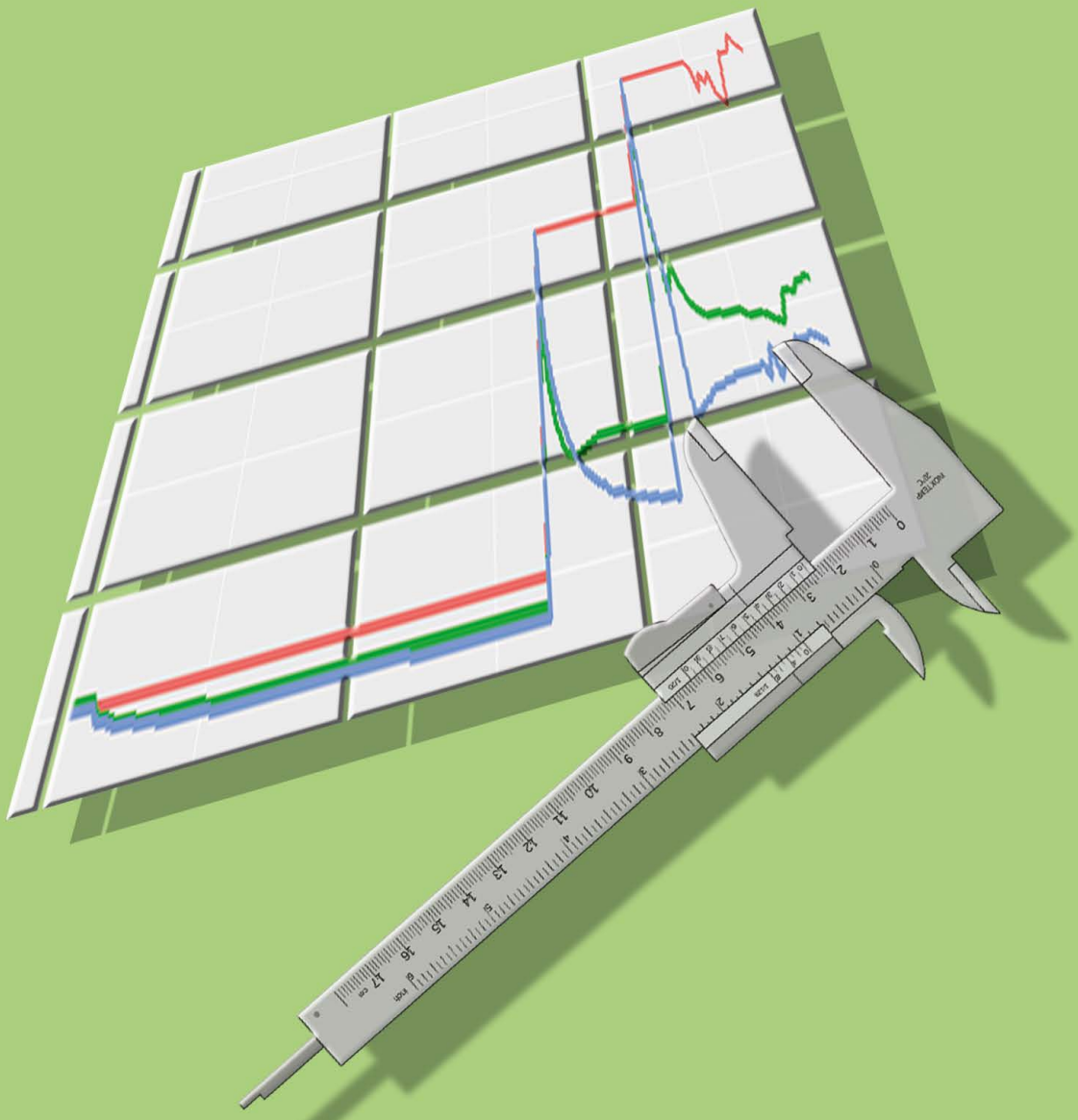
Time series modelling in repeatedly conducted sample surveys

Time Series Modelling in Repeatedly Conducted Sample Surveys



Oksana Balabay

Cover design and LaTeX typeset by the author.

Published by Universitaire Pers Maastricht
ISBN: 978 94 6159 550 8
Printed in the Netherlands by Datawyse Maastricht

Time Series Modelling in Repeatedly Conducted Sample Surveys

DISSERTATION

to obtain the degree of Doctor at
Maastricht University,
on the authority of the Rector Magnificus,
Prof. dr. L.L.G. Soete,
in accordance with the decision of the Board of Deans,
to be defended in public
on Wednesday, 11 May, 2016 at 10.00 o'clock

by

Oksana Balabay

Supervisors:

Prof. Dr. J.A. van den Brakel

Prof. Dr. F.C. Palm

Assessment Committee:

Prof. Dr. J.R.Y.J. Urbain (Chair)

Prof. Dr. D. Pfeffermann (Government Statistician of Israel; University of Southampton; Hebrew University of Jerusalem)

Prof. Dr. S.J. Koopman (VU University Amsterdam)

Prof. Dr. P.A.P.M.H. Mohnen (UNU-MERIT; University of Maastricht)

This research was financially supported by Statistics Netherlands (Het Centraal Bureau voor de Statistiek (CBS)).

To my husband, Vikram

నా భర్త, విక్రమ్ కు

Acknowledgements

As I am writing these lines, I realize that it has been four years and four days since I started my PhD project. 1 Dec 2011 is not a very usual date to start doing one's PhD. When I look back, I shudder at the thought that I actually was within a whisker of missing this chance. Being a master student from Sweden, I was not really familiar with the Dutch academic world. I knew I wanted to do a PhD, but all my applications for Dutch PhD programmes that had started on 1 Sep got me nowhere. So I started looking for a job in the corporate world. When I saw a PhD position at Maastricht University (UM), I said to my husband Vikram: "Look, there is another PhD opening. Nice! But – alas – the required background is Econometrics or Mathematics (while my major at university was Economics)". To tell the truth, I ended up at the UM thanks to Vikram because he encouraged me to apply despite my "unsuitable" profile. It is funny to recall how my interviews for this PhD position and for another job in Amsterdam (let us call it N) were almost overlapping. When I was on the train on my way to Maastricht for my first interview at the UM, I had a telephone interview with company N. In Maastricht, I met with my future supervisors, Franz Palm and Jan van den Brakel, as well as Wim de Witte (R.I.P.), for the first time. I felt their firm handshakes, but I was not sure about the outcome. In the meantime, I went to Ukraine to see my new-born nephew, Nikolai, and then, upon my return, went to my first face-to-face interview with N in Amsterdam, right from the airport, carrying a big suitcase. In mid-November, I was invited again to both N and the UM on the same day. At the UM, it was my last interview and I was accepted. One day later, I got a call from N who invited me to come for a final interview with the management where employment conditions and other formalities would be discussed. When I said that I was not considering the job any longer because I had got a PhD position, the voice on the phone sounded very surprised. I had to reiterate my final choice. Now I am happy they did not call me earlier, because my PhD position turned

out to be perfect for me – both in terms of content and supervision. I would like to express my sincere gratitude to Jan van den Brakel and Franz Palm for their expertise and constant support. Thank you for always being there for me, for your trust and confidence in me.

I would like to thank the reading committee - Jean-Pierre Urbain, Danny Pfeffermann, Siem Jan Koopman and Pierre Mohnen, - as well as Ralf Münnich, for their efforts related to reading and commenting on this thesis. I would also like to thank Ralf Münnich and Jean-Pierre Urbain for their courses that I had the privilege to attend and for sharing their vast knowledge with me.

Further, a big thanks goes to my colleagues at Statistics Netherlands (the CBS), especially to Harm Jan Boonstra, Sabine Krieg, Rob Willems, Joeri Roels, Marco Puts, Mathijs Jacobs, Bart Buelens, Frank de Pol, Sofie de Broe, Marij Amkreutz-Verbruggen, Paul Ras, Wilfred Zentjens, Joep Burger, Ricco Konen, Carin Zwaneveld, Hermine Molnar-in 't Veld, May Offermans, Martijn Tennekes, Elke Moons, Deirdre Giesen, Jeremy Visschers, Barry Schouten, Frank Pijpers (the list goes on) who helped me navigate in the ocean of the CBS-data, gave me timely advice, provided technical support and/or revised my papers; last but not least – my sport companions: Sabine Krieg, Nadine Wesselius and Hille Wesseling. Regrettably, I cannot mention everyone, but I would like to say that all of you are so friendly and always willing to help. Thanks for that! A special thanks goes to those who substantially facilitated my research activity by taking most of the administrative work upon themselves: Karin van den Boorn, Yolanda Paulissen, and Brigitte Defesche. I would also like to thank Miranda Gubbels for assisting me with my international travel grant, and of course, Prof. Dr. Jonas Andersson for giving me a warm welcome in Bergen, Norway, where I spent a few wonderful and productive months as a guest researcher. Olga P., Kristine T., Matthias D. and Daniel S. – thank you all for helping the whole thing go smoothly and even making it fun (yes, work should also be fun, even though "Qualität kommt aus Qual" and "Pressure creates diamonds"). I am also truly thankful to Daniel Lewis for proof-reading my first paper.

I am also thankful to my colleagues from the SBE, especially guys from the fourth floor: my office-mate Anne B. (thanks for proof-reading my Dutch summary!), Elnaz, Angie, Andrey K. (for always being there to help at the right time with the

right thing), Hanno R. (for discussions about econometrics and music), Seher F. (for our one-on-one piano evenings), Marc S. (for his “skiing for dummies” lesson), Anna Z. (for her cute souvenirs from all over the world and keeping me company at Andrej W. and Hanna M.’s wedding in Belarus), and, of course, Andrej W. himself for his constant support (including but not limited to his gracious act of placing his water filter, fridge and Chinese tea at my disposal, as well as taking care of my piano and some other things). I also thank my university/CBS friend Irina Rud for our "gezellige" trips to the CBS. I am also sending my appreciation to Nalan Basturk, Denis de Crombrugghe and Stephan Smeekes for their approachability and for their courses I had the pleasure and honour to attend.

My dear friends also deserve a token of gratitude as a minimum because they make this world a better place and made my PhD years a better time: Hanna M., Liza T., Oxana S. and Guney C., Katya and Stas D., Nina P., Irina S., Alexey N., Konstantin S.. I would also like to thank my friends from Kamerkoor Maastricht for unforgettable concerts and Thursday nights. Special thanks are reserved for f. Alexander Galaka and m. Martha (Blanca) Smits for their support in the most difficult times and for great Sundays.

I cordially thank my paranymphs – Anne Balter and Liza Schuivens-Tsinoeva for accepting the challenge of defending me before my opponents, if needed.

I would like to express my utmost gratitude to my family: my sister Marina (for improving my English writing skills) and my brother-in-law Dima (just for being such a cool brother-in-law), my cute niece Lena and nephew Nikolai, my mum and dad, my parents-in-law, my sister-in-law Sreelatha and her little son Arnav. I am so happy I can always count on your support and love. It goes without saying that my dearest husband Vikram deserves the lion’s share of my personal achievements by giving me all the freedom to do my PhD and even by letting me exchange Groningen for Maastricht. Without you, all of this would not have been possible.

I owe so much to all of you,
Oksana Bollineni-Balabay

Somewhere over the neutral waters of the Black Sea, December 2015

Contents

List of Acronyms

AA - asymptotic approximation
AR - absolute revision
DGP - data generating process
DLFS - the Dutch Labour Force survey
DRTS - the Dutch Road Transportation survey
DTS - the Dutch Transportation survey
GREG - the general regression (estimator)
HB - Hierarchical Bayesian
HR - Hire-and-Reward
HT - Horvitz-Thompson (estimator)
KF - Kalman filter
ML - maximum-likelihood
MSE - the mean square error
NSTR - Nomenclature uniforme des marchandises pour les Statistiques de Transport, Revisée
OA - Own-Account
PT1 - Pfeffermann-Tiller parametric bootstrap
PT2 - Pfeffermann-Tiller non-parametric bootstrap
RGB - rotation group bias
RR - relative revision
RR1 - Rodriguez-Ruiz parametric bootstrap
RR2 - Rodriguez-Ruiz non-parametric bootstrap
SAE - small area estimation
SE - standard error
STS - structural time series

Chapter 1

Introduction

To begin with, it is worth mentioning the principal driving force behind this research: a strive for high-quality official statistical figures. The challenge here is to improve the quality without resorting to trivial sample size increases. Nowadays, when budgets of national statistical institutes (NSIs) are shrinking and when the society is becoming ever more averse to administrative burden they experience from their NSIs, increasing the sample size is not a feasible option, despite the availability of new information technologies for interviewing and collecting information. Thus, the need for new methods, stretching beyond the traditional survey methodology, is more than evident. Time series modelling becomes an apparent choice for surveys that are repeatedly conducted over time.

This thesis has attempted to investigate and to work out some important and, I believe, useful aspects at the intersection of survey methodology and time series econometrics. These two fields have for decades been developing independently. Some attempts to introduce time series modelling to official statistics were made long ago in academic world, e.g., by Scott and Smith (1974), Binder and Dick (1989), Binder and Dick (1990), Pfeffermann (1991) and Tiller (1992). However, it was only at the beginning of the 21st century that the time series model-based approach was used in the official production of statistical figures. But to the best of my knowledge, only two NSIs resort to this approach so far: the US Bureau of Labor Statistics (Tiller (1992), Pfeffermann and Tiller (2006)) and Statistics Netherlands (Van den Brakel and Krieg (2009a)).

For decades the prevailing opinion at NSIs was that official statistics should be free from model assumptions. NSIs are reluctant to apply model-based techniques in the production of official statistics, mainly because model misspecification easily results in severely biased estimates. Hence, careful model evaluation and selection are required, which are an additional laborious stage in the production process, particularly for multipurpose surveys where separate models are required for different variables. The knowledge transfer from academia to NSIs is limited as well, because it often requires routine work from academics, as well as a certain degree of expertise in survey methodology.

NSIs usually apply design-based estimation procedures known from classical sampling theory in order to compile official statistics. Under these procedures, one of many possible probability samples is drawn from a finite target population. This concept is known as repeated sampling or randomization. Under this concept, statistical modelling plays a minor role, and (approximately) design-unbiased estimates for population parameters of interest are derived from the sampled data and optionally auxiliary data (in case of model-assisted estimation, where inference remains design-based). Well-known examples of design-based estimators are the Horvitz-Thompson (HT) estimator and general regression (GREG) estimator (Särndal et al. (1992), Cochran (1977)). The main advantage of these methods is that they are unbiased and perform reasonably well given the sample size is sufficiently large. Small sample sizes, however, result in unacceptably large standard errors. Sometimes, it is not at all possible to obtain design-based estimates for small areas with a zero sample size allocation. Another major drawback of design-based estimation procedures appears when the underlying survey process changes, rendering figures incomparable over time. Rao (2011) elaborates on the power and limitations of design-based and model-based methods in a cross-sectional dimension. The objective of this dissertation was to develop suitable time series models for several data sets for repeatedly conducted surveys with several survey redesigns. Apart from that, the thesis is aimed at illustrating how time and space dimension can be exploited in repeatedly conducted surveys. The purpose of this is two-fold: improving the quality of official statistical figures in terms of precision, and making these official figures comparable over time in case the survey has been redesigned. I elaborate on these two aspects below.

The first problem - small sample sizes - places the present work in the field of what is known as small area estimation (SAE). The problem of SAE emerges when reliable estimates for small areas cannot be produced relying solely on the survey design due to an insufficient number of sampled units. Either the number of units sampled in the domain of interest is too small, or the domain is defined post hoc by the outcome for the variable of interest that has a rare occurrence. An increasing demand for small area statistics boosted the development of numerous SAE techniques. Until lately, most of these were meant for cross-sectional data, where the so-called *strength over space* could be borrowed. This implies that sample information from other similar domains could be used to improve the design-based estimate for a small area of interest. Sometimes, an area may not be sampled at all, in which case regression synthetic estimators (Gonzalez (1973)) produce figures using some auxiliary variables. Composite estimators may be used to combine direct estimates with their synthetic counterparts. This is often done in the framework of the well known Fay-Herriot model (Fay and Herriot (1979)) at the area level, or of the model of Battese et al. (1988) at the unit level. The literature on these methods is extensive, see, e.g., Rao and Molina (2015), Ghosh and Lahiri (1987), Prasad and Rao (1990), Pfeffermann (2013), Pfeffermann (2002).

While small area estimates may be improved by borrowing strength over space, adding the time dimension is found to offer a huge potential for improvement. It is worth emphasising that at every time point t, design-based methods use only information collected at time t for estimate production, neglecting all the vast amount of information collected prior to time t. *Borrowing strength over time* can be realised with the help of a time series model by using sampling information accumulated over time in the respective, as well as in other domains. In this framework, the unknown population parameter of interest can be viewed as a realisation of a stochastic process that evolves over time. Time series techniques are powerful tools in producing more reliable estimates for repeatedly conducted surveys, both for small and non-small areas (see, e.g., Scott et al. (1977), Lind (2005), Blight and Scott (1973), Abraham and Vijayan (1992), Datta et al. (1999)). Combining the spatial and time dimension allows us to account for cross-sectional (contemporaneous) correlation between domains, which is particularly applicable to official statistics as most surveys are conducted repeatedly over time (see Pfeffermann and Burck (1990), Pfeffermann and Bleuer (1993), Krieg and Van den Brakel (2012), Bollineni-Balabay et al. (2016a) for applications). Being able to borrow strength

over time and domain space, such models can extract the so-called signal by removing a great part of the sampling noise from design-based point-estimates. As a result, standard errors (SEs) of such model-based point-estimates are usually substantially lower than the design SEs.

The second major problem with design-based methods is that from time to time, a survey process must be amended due to quality or budget issues. This results in level breaks, which are sometimes called discontinuities, as well as in the survey error variance breaks. To make pre- and post-redesign figures comparable, the redesign effect must be quantified. One way to do this is to compare the old and new approach in a parallel run. If this is not feasible, e.g. because of budget limitations, a time series model can be applied that disentangles the evolution of the series' pattern from the systematic effect of the redesign. Level break estimation is illustrated in Harvey and Durbin (1986), Van den Brakel and Roels (2010), as well as in Bollineni-Balabay et al. (2016a).

The last paper mentioned above is based on Chapters 2 and 3 of this thesis. This chapter deals with the small sample sizes problem in the Dutch Road Transportation Survey (DRTS), as well as with multiple level and variance breaks caused by survey redesigns. These three problems can simultaneously be solved with a structural time series (STS) model.

The STS approach should not be mistakenly taken for structural econometric modelling; the term "structural" rather stems from decomposing the time series into a number of unobserved components (e.g., trend, seasonal, regression, and irregular components). The STS approach is akin to the better known and widely spread autoregressive integrated moving average (ARIMA, see Box et al. (2011)) approach in that both generate predictions by exponentially weighting past observations. First works on STS modelling can be traced back to the 1960s (Muth (1960), Theil and Wage (1964), Harrison (1967)), and the Kalman filter mostly used for computing estimates of the STS model components was also introduced to the public quite early by Kalman (1960). Nevertheless, the ARIMA approach gained more importance throughout the history of time series econometrics, mainly due to its computational simplicity. See more on history of unobserved component modelling in Harvey and Durbin (1986). Harvey et al. (1998) note that STS models are a better alternative to standard ARIMA techniques, especially when

data is messy. STS models also offer an easier interpretation for the components of interest and are more flexible when it comes to complex structures in time and space dimensions. The key is putting a STS model into a state space form (hence another name "state space models"), whereafter the hyperparameter maximum-likelihood (ML) estimation and estimation of the unobserved components (state variables) can be done with the help of the Kalman filter. The Kalman filter is also known as a powerful device for dealing with missing and mixed-frequency data. Extensive theory on STS models together with applied aspects on model diagnostics, Kalman filtering, non-linear and non-Gaussian state space models can be found in Harvey (1989) and Durbin and Koopman (2012).

Chapter 2 presents how a real-life small area application can largely benefit from the STS model-based approach. The DRTS features a break-down into nine domains, whose design-based estimates have unacceptably large variances, as well as multiple level and variance breaks. Univariate STS framework has been extended to a multivariate one with cointegrated trends, where the terms "univariate" and "multivariate" in the STS context refer to models designed for one and several observed time series, respectively. While univariate modelling offers variance reduction through borrowing strength over time, detection and modelling of common factors and/or cross-correlations between series within the multivariate framework offers an additional variance reduction through borrowing strength over space.

As the DRTS time series are published both at the national level and at the underlying domain level, the question of how the aggregated series, i.e. the sum of the domain series, should be treated within the state-space approach is also addressed in Chapter 2 of this thesis. If the direct estimates at the aggregated level were sufficiently precise, they could be directly used as benchmarks for the domain model estimates, which is frequently done in SAE. This alternative way of borrowing strength over space also provides a form of built-in robustness against model misspecification, see Pfeffermann and Burck (1990), Pfeffermann and Bleuer (1993) and Pfeffermann and Tiller (2006). Durbin and Quenneville (1997) exploit aggregation over the time dimension. However, the high volatility of the DRTS design estimates at the national level suggests that they themselves could benefit from the STS approach. We were moved to write Chapter 3 in order to answer the question about whether or not to include the aggregated series into a multivariate model for domains, where the state variables of the aggregated level series

are set equal to the sum of the corresponding domain state variables. Including the aggregated series into the model may seem to be appealing because a series at a higher aggregation level usually has a better signal-to-noise ratio, compared to that of the underlying domains. On the other hand, it can be argued that no additional information would enter the model in this case. Chapter 3 sheds light on this question by investigating the performance of the Kalman filter under true and estimated hyperparameters.

Estimation of STS models presented in Chapters 2 and 3 rests on the assumption of known hyperparameters. In practice, these are estimated (e.g. with the ML estimator) and treated as known in the Kalman filter recursions. It is worth noting that other commonly applied SAE procedures (mostly based on multilevel models) also contain unknown hyperparameters that have to be estimated, which translates into larger standard errors around the domain predictions. If this uncertainty (here and further referred to as the hyperparameter uncertainty) is not taken into account, the MSE estimates of the quantities of interest become negatively biased. Within the multilevel model estimation framework (the empirical best linear unbiased predictor (EBLUP) or the hierarchical Bayesian (HB) approach), it is very common to take the hyperparameter uncertainty into account, see Rao (2003), Ch.6-7, 10. STS models, in turn, are not as widely used in SAE as multilevel models. As hyperparameter estimates are treated by the Kalman filter as known, the MSEs of the state variables are estimated with a negative bias due to neglecting uncertainty around these underlying hyperparameters. For this reason, the gains from the STS technique in terms of reduced variance estimates have to be treated without undue optimism. Applications that give evidence for substantial advantages of STS models over the design-based approach treat the estimated model hyperparameters as known, see, e.g., Krieg and Van den Brakel (2012), Van den Brakel and Krieg (2009a), Pfeffermann and Bleuer (1993), Tiller (1992). Chapter 4 is devoted to MSE approximation methods in a STS model which since 2010 has been employed by Statistics Netherlands in its Dutch Labour Force survey (DLFS) for production of official monthly labour force figures (this STS model is further referred to as the DLFS model). The literature offers several ways to account for the hyperparameter uncertainty in STS models. Among those methods considered in the Chapter are the asymptotic approximation developed by Hamilton (1986), as well as parametric and non-parametric bootstrapping approaches developed by Pfeffermann and Tiller (2005) and Rodriguez and Ruiz (2012). The DLFS model

acts as the data generation process (DGP) in an extensive Monte-Carlo study. This simulation study is set up to explore the performance of the afore-mentioned MSE approximation methods in a real life application, as well as in order to verify the claim of Rodriguez and Ruiz (2012) about the superiority of their method over the bootstrap of Pfeffermann and Tiller (2005) in a more complex model. The Monte-Carlo study also shows how the model can be checked for possible over-specification.

Chapter 5 compares the frequentist-based STS approach facilitated by the Kalman filter with a multilevel time series model estimated within the hierarchical Bayesian (HB) approach. The SAE application chosen for this purpose is the Dutch Travel Survey (DTS) featuring unacceptably unstable estimates in its more than 600 small domains. The reasons for that are small sample sizes at the provincial level and several survey redesigns, which make the official figures hard to compare over time. The multilevel model presented in this Chapter is a time series extensions of the Fay-Herriot model. Similar extensions can be found in Rao and Yu (1994), Datta et al. (1999), and You (2008) (see Rao and Molina (2015) for an overview). Although STS models could also be estimated in the fully Bayesian way, sequential updating in models like the DTS one would be computationally very expensive even with modern computers. However, it is not the main reason for the chosen comparison. As has been mentioned above, most practitioners, including Statistics Netherlands, estimate STS models in the frequentist framework. Therefore, the authors are aiming to find out whether the quality of the STS model-based estimates produced by frequentist estimation routines is good enough compared to the that of the full Bayesian multilevel framework. Both modelling approaches are made comparable in terms of the same pooling dimension, as well as in that no spatial correlation is assumed across domains. Apart from assuming nearly identical likelihood functions (differing only in the way the trend model is specified) for the two approaches, several other modifications - not applicable to the conventional STS approach - are explored within the multilevel framework: among other things, allowing for random effects. Rationales behind multivariate against univariate modelling dimensions are also addressed. The multilevel time series and STS approaches are compared in terms of adequacy of model-based point-estimates for the trend and for multiple level breaks due to the survey redesigns, as well as in terms of gain in precision that can be reached within these two mo-

delling frameworks.

Another novelty presented in Chapter 5 lies in exploiting the STS approach to refine the multilevel one in the case of unreliable variance estimates of the design-based estimator (further referred to as design-based variance estimates). These are used as input in multilevel models and are treated as the true known sampling error variances. Chapter 5 shows how volatility, missing values, as well as a possible bias in design-based variance estimates can be alleviated.

Conclusions to the thesis are contained in Chapter 6. It presents the main findings of the present work, as well as discusses possible paths for future research in this field.

Chapter 2

Multivariate State Space Approach to Variance Reduction in Series with Level and Variance Breaks due to Survey Redesigns

This chapter illustrates the power of the structural time series (STS) approach in a real life official statistical application - the Dutch Road transportation survey. Statistics Netherlands applies a design-based estimation procedure to produce official figures for this survey. Frequent survey redesigns caused discontinuities in these series, which obstructs the comparability of figures over time. Reductions of the sample size and changes in the sample design resulted in breaks in the sampling error variances, with sampling errors becoming unacceptably large in the recent part of the series. In this chapter, both problems are addressed and solved simultaneously. Discontinuities and small sample sizes are accounted for using a multivariate STS model that borrows strength over time and space. The present chapter illustrates an increased precision when one moves from univariate models to a multivariate one. This increase is especially significant in the most recent period when sample sizes get smaller, with the design-based standard errors being reduced by 40 to 70 percent with the model-based approach.[1]

[1]This chapter is based on the paper Bollineni-Balabay et al. (2016a) published in *Journal of the Royal Statistical Society: Series A*. The authors thank Dr. Sabine Krieg and Rob Willems (Statistics Netherlands), as well as the unknown reviewers and Associate Editor for valuable comments on earlier versions of this paper.

2.1 Introduction

Surveys repeatedly conducted by National Statistical Institutes (NSIs) aim at constructing consistent time series that reflect the evolution of the phenomena of interest. NSIs usually apply design-based estimation procedures to compile official statistics. Well known examples are the Horvitz-Thompson estimator and general regression estimator (Särndal et al. (1992)). If the sample size is sufficiently large and the underlying survey process remains unchanged, the design-based approach indeed results in reliable figures that are comparable over time. A major drawback of design-based estimation procedures is that small sample sizes result in unacceptably large standard errors. This is usually the case for estimates at lower aggregation levels (Rao and Molina (2015)). Another common problem in official statistics results from frequent survey redesigns. From time to time, a survey process has to be amended for quality or efficiency reasons, which often renders outcomes incomparable over time (Van den Brakel and Roels (2010)). This paper focuses on the two above-mentioned problems - survey redesigns and small sample sizes in repeated survey sampling.

Survey redesigns generally have systematic effects on the outcomes. This can be reflected by level breaks, sometimes called discontinuities, as well as by variance breaks. In order to make pre- and post-redesign figures comparable, the redesign effect must be quantified. One way to do it is to compare the old and new approach in a parallel run. If this is not feasible, e.g. due to budget limitations, a time series model can be applied that disentangles the evolution of the series' pattern from the systematic effect of the redesign. Structural time series models can be exploited to model different forms of discontinuities. As an example, see Harvey and Durbin (1986) where a state space intervention model is applied to account for level shifts. Redesigns can also affect the variance of the direct estimates, both directly (through changes in the sample size) and indirectly. Regarding indirect effects of redesigns, one could think of some changes in the data collection process. These can affect the variance of measurement errors for individuals, which translates into a change in the variance of the direct estimator. If the design variances are available from the micro-data, then they can be used as prior information in the time series model. This will automatically account for shock-effects in the sampling error and for other forms of heteroscedasticity induced by the survey redesigns (see, e.g., Binder and Dick (1990); Durbin and Quenneville (1997)). If

the design variances are not available, a time series model must account for this heteroscedasticity by making the model variance time-dependent.

Small sample size is another issue NSIs frequently have to deal with when striving for reliable estimates for subpopulations or domains. The problem with detailed subdivisions of the population is that sample sizes are often too small to apply design-based estimators for production of sufficiently precise figures. In this case, a model is needed to increase the precision of a domain estimate with the help of the sample information observed in preceding periods or in other domains. This kind of technique is often referred to as small area estimation (SAE), see Rao and Molina (2015).

Structural time series models are frequently applied in SAE. With a structural time series model, the sampling and measurement errors can be filtered out from the time series of direct estimates to obtain a more reliable series – the signal. The unobserved components underlying the signal, such as trend, seasonal and regression components, benefit from the sample information accumulated in the past. This is sometimes called borrowing strength over time. Further improvement in domain estimates can be obtained in a multivariate setting by modelling the correlation between (some) unobserved components of different domains (Pfeffermann and Burck (1990), Pfeffermann and Bleuer (1993), Krieg and Van den Brakel (2012)). This is usually referred to as borrowing strength over space. Some correlated auxiliary series can also be used to borrow information from. Harvey and Chung (2000), for instance, jointly modelled the series from the UK labour force survey with a series of claimant counts to improve the precision of the former series estimates. Pfeffermann and Tiller (2006) proposed to benchmark time series estimates of domains to sufficiently precise direct estimates at the national level as an alternative method to borrow strength over space. This method also provides robustness against model misspecification.

Effective structural time series modelling can improve the accuracy of time series published by NSIs that rely on the traditional design-based approach from sampling theory. Harvey and Chung (2000) provide an illustrative example for the Labour Force Survey in the UK. NSIs are nevertheless still reluctant to apply these techniques in the production of official statistics, mainly because model misspecification easily results in severely biased estimates. Hence, careful model

evaluation and selection are required, which is an additional laborious stage in the production process, particularly for multipurpose surveys where separate models are required for different variables. To our knowledge, only two governmental statistical institutes use a state-space model in the production of their official figures: Statistics Netherlands with their Dutch Labour Force Survey model (Van den Brakel and Krieg (2009a)), and the U.S. Bureau of Labor Statistics (Tiller (1992), Pfeffermann and Tiller (2006)).

The Dutch Road Transportation Survey (the DRTS) is a long-standing survey conducted by Statistics Netherlands to produce reliable annual and quarterly figures on freight transportation. This survey has been redesigned several times, resulting in multiple level breaks in the series, as well as in some design variance breaks visible in the published figures. Moreover, due to reduced budgets, the DRTS faces decreasing sample sizes, which results in an increasing loss of precision in the direct estimates. In this paper, multivariate structural time series modelling is applied to solve both problems simultaneously, which makes the application interesting from both practical and academic points of view. An additional complication that is addressed in this application is that sampling units can belong to more than one domain, resulting in additional correlation between the domain estimates.

As its main contribution, this paper provides substantially improved survey estimates using a multivariate state-space approach where multiple survey redesigns and other survey process changes are treated as exogenous events and are modelled as level and variance breaks. The model also allows for common domain trends and for contemporaneous inter-domain correlation in the sampling errors. As the DRTS time series are published both at the national level and at the underlying domain level, the question of how the aggregated series should be treated within the state-space approach is also addressed.

Section 2.2 describes the data and its major discontinuities induced by survey redesigns. Section 2.3 focuses on the structural time series models employed in the DRTS. Estimation results are presented in Section 2.4. Section 2.5 summarises the main findings and offers some possible further improvements.

2.2 The Dutch Road Transportation Survey

The DRTS survey measures freight transportation in terms of tons, kilometres and ton-kilometres. The target variables are constructed for international and domestic segments separately. Further, these variables are divided into Hire-and-Reward (HR) and Own-Account (OA) categories according to whether or not transportation is carried out at the cost of the vehicle owner. The present study analyses domestic OA road freight transportation carried out by vehicles registered at the Dutch Admission Authority for Vehicles. These series are measured in thousands of tons on a quarterly basis from 1976(1) until 2010(4) (where numbers in brackets denote quarters), and are divided into nine categories according to the so-called NSTR-classification (Nomenclature uniforme des marchandises pour les Statistiques de Transport, Revisée). This classification is based on the type of goods transported and includes ten categories (short names used hereinafter are given in brackets): 0. agricultural products and live animals (agriculture); 1. foodstuff and animal fodder (food); 2. solid mineral fuels; 3. petroleum oils and petroleum; 4. ores, metal scrap, roasted iron pyrites (ores); 5. iron, steel and non-ferrous metals (including intermediates) (metals); 6. crude and manufactured minerals, building materials (minerals); 7. fertilisers; 8. chemicals; 9. vehicles, machinery and other goods (including cargo) (other goods). The enumeration in the present paper begins from 1, with NSTR 2 and 3 being combined in domain 3 (oil). The analysis is therefore based on nine target variable series which are called domains in this paper. Since some vehicles transport goods from different categories, they may appear in more than one domain. As a result, an additional correlation arises between such domains. This makes this application different from the traditional situation where sampling units belong to one domain only.

The estimation procedure of this survey is based on the Horvitz-Thompson (HT) estimator (Horvitz and Thompson (1952), Narain (1951)). This is a design-based estimator that expands the observations by weights obtained as inverse inclusion probabilities of the sampled units. The HT point estimates of the own-account domestic transportation series, which are officially published by Statistics Netherlands (StatLine.cbs.nl), are shown in Fig. 2.2.1.

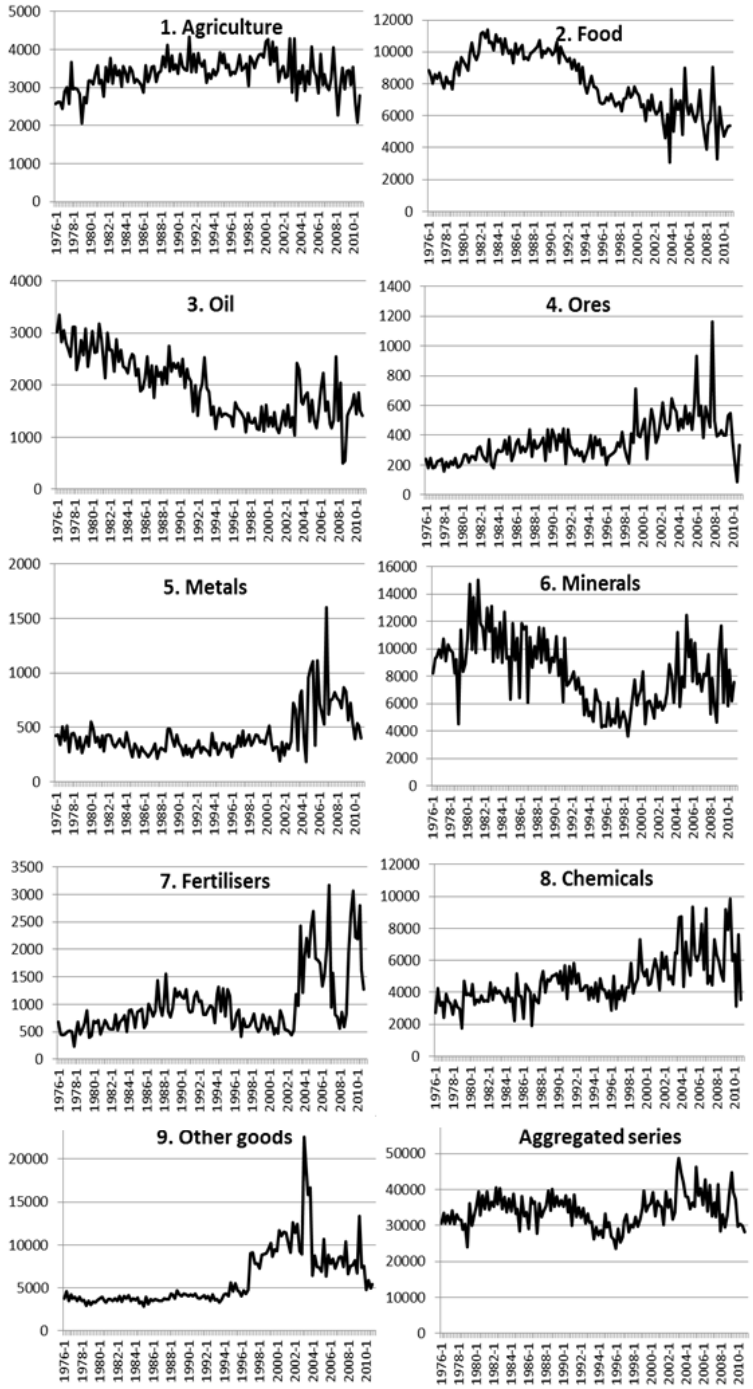


Figure 2.2.1: Horvitz-Thompson estimates of the own account domestic transportation series, 1976(1)-2010(4), in kilotons.

Over the years, the DRTS has undergone a number of methodological changes. Various amendments, more and less significant, were particularly frequent during the last decade. They caused breaks both in the level of the series and in their design variances. Only the most important changes that might have caused discontinuities are mentioned here.

Until 2003, the DRTS was based on a one-stage stratified sample of vehicles (sampling units). Companies reported on a three-day transportation performance for the selected vehicles using paper questionnaires. From 2003 to 2007, a stratified two-stage sampling design was applied with the company as a primary sampling unit, and the vehicle as a secondary sampling unit. Apart from that, several survey process changes took place in 2003, such as the introduction of a new questionnaire and the transition from a three-day to week-wise reporting period. These and other changes introduced in 2003 had a considerable effect on the design variances (which is visible in every series, except domain 4) and the level of the series point estimates (see domain 7).

Since 2008, vehicles have been drawn directly from the Dutch Admission Authority for Vehicles on a quarterly basis using a stratified sample. Since then, stratification has been based on the vehicle type and load capacity, as well as on the total load capacity of the company owning the vehicle, and on the industry branch the company belongs to. Vehicles drawn at the beginning of every quarter are further clustered by their owners. The latter are, in turn, randomly assigned to different weeks in a quarter. This means that the sample design still has a two-stage sampling structure. Vehicle transportation performance is still observed during one week. The net-sample size in the most recent decennium has been fluctuating between 9000 and 12000 vehicles per quarter. Certain ambiguities arose in the late 2000's regarding the classification of shipments into the OA and HR categories. This caused a shift mainly in domestic transporters from the former to the latter category.

In 1994, small vehicles with a load capacity less than 1.5 tons were excluded from the OA series. From 1997 to 2002, these vehicles, most of which are vans, were included in the survey again. This caused a considerable upward shift in the level of domain 9. Another artefact in the series of domain 9 is a huge peak in 2003. Since 2003, vans, together with special vehicles, have been monitored using

a separate smaller questionnaire. Starting from 2009, they are no longer surveyed. Since then, total weight transported by these vehicles has been estimated using register information from another national authority called "Nationale AutoPas", which contains the number of kilometres the vehicle has covered each time it visits a vehicle service station. This observation method is obviously less precise, so the inclusion of van freight adds uncertainty to the series. Therefore, Statistics Netherlands decided to publish two different series: one that includes van freights, and the other that does not. However, due to problems in 2003 related to the identification of vans and other types of small vehicles, the series without van freights is available only from 2004, leaving the series with the peak for the whole duration of 2003.

In different domains, the above-mentioned survey changes are reflected to a different extent, but as is clearly visible, every domain except for domain 4 exhibits an increase in variation in 2003 (see Fig. 2.2.1). This can be explained by decreasing sample sizes over time and by the above-mentioned changes in the survey design. The survey design variances are very high in the last decennium of the time series, which results in large fluctuations in the point-estimates. In order to produce more stable estimates, a multivariate structural time series model is developed for the DRTS in the next section.

2.3 Structural Time Series Models and Methods Employed

2.3.1 Structural Time Series Model Specification

Structural time series models decompose a series into a number of unobserved components, e.g., trend, seasonal, other cyclical and regression components (see Durbin and Koopman (2012)). If separate DRTS series at the domain or national level are modelled individually, i.e. in a univariate setting, the variance reduction comes from borrowing information over time. If the domains are jointly modelled in a multivariate model, it is possible to borrow information both over time and domain space. These models must allow for heteroscedasticity in the survey errors, as well as for the level breaks mentioned in Section 2.2.

Let $y_{i,t,d}$ denote a target variable observation at sampling unit i at time t in domain d. Generally, survey observations are subject to measurement errors, which gives rise to the following measurement error model: $y_{i,t,d} = u_{i,t,d} + m_{i,d} + w_{i,t,d}$. Here, $u_{i,t,d}$ is the true but not directly observable parameter of interest at the unit level at time t in domain d; $m_{i,d}$ is a systematic measurement error at the unit level, and $w_{i,t,d}$ is a random measurement error. The systematic measurement error $m_{i,d}$ depends on the design of the survey process and is assumed to be time independent as long as the underlying survey process remains unchanged. At the domain level, the unknown population parameter of interest at time t is denoted as $u_{t,d}$. In this application, the parameter of interest is the total number of tons transported in domain d: $u_{t,d} = \sum_{i \epsilon U_d} u_{i,t,d}$, where U_d denotes the population of vehicles in domain d.

Let $\hat{Y}_{t,d}$ denote an HT estimate of the unknown population total. Each of these HT estimates is based on the sample observed in period t. Together these estimates constitute an input series for the time series model and are expressed as $\hat{Y}_{t,d} = \theta_{t,d} + e_{t,d}$, where $e_{t,d}$ is the total effect of a sampling error and random measurement errors $w_{i,t,d}$, and $\theta_{t,d}$ is the parameter of interest obtained when the whole population is surveyed under a particular survey design. In this way, $\theta_{t,d} = u_{t,d} + m_d$ is the sum of the true population parameter $u_{t,d}$ and a measurement bias m_d of the survey design at the domain level. In what follows, $\theta_{t,d}$ will be referred to as the population parameter. These two terms cannot be separated from each other with the survey data at hand. Durbin and Quenneville (1997) illustrate how measurement bias can be estimated if exact or more accurate benchmarks are available, e.g., some bias-free auxiliary series or annual data.

As long as the survey process remains unchanged, the population parameter $\theta_{t,d} = u_{t,d} + m_d$ can be decomposed with a structural time series model into a stochastic trend $L_{t,d}$, a stochastic seasonal component $\gamma_{t,d}$, and an irregular term $\varepsilon_{t,d}$, so that $\theta_{t,d} = L_{t,d} + \gamma_{t,d} + \varepsilon_{t,d}$. Inserting this structural time series model into the measurement error model gives $\hat{Y}_{t,d} = L_{t,d} + \gamma_{t,d} + \varepsilon_{t,d} + e_{t,d}$. Unlike in a panel, where the irregular term $\varepsilon_{t,d}$ and the sampling error $e_{t,d}$ can follow different models, it is not possible to separate the two in a cross-sectional survey like the DRTS. Therefore, they are combined into one composite error term $\nu_{t,d} = \varepsilon_{t,d} + e_{t,d}$.

As described in the previous section, the multiple survey redesigns resulted in several discontinuities and in heteroscedasticity in the survey errors. First consider a single redesign resulting in a discontinuity in domain d. Then the measurement error m_d from the aforementioned measurement error model shifts to another level, say m'_d. Assuming that the difference $m_d - m'_d \equiv \beta_d$ is constant over time, it can be modelled with a level intervention using a dummy regressor. This gives rise to the following representation of the observed series: $\hat{Y}_{t,d} = L_{t,d} + \gamma_{t,d} + \delta_{t,d}\beta_d + \nu_{t,d}$, where $\delta_{t,d}$ is a dummy regressor that is equal to one for the time points when the intervention is effective and to zero otherwise (see Table 2.4.1 for different types of level interventions in this application). Then the regression coefficient β_d, being the difference between the measurement errors under two different survey designs, can be interpreted as a discontinuity induced by a single survey redesign. In general, K_d redesigns in domain d result in K_d discontinuities which can be modelled with K_d intervention variables in the structural time series model. These considerations imply the following model for the domains:

$$\hat{Y}_{t,d} = L_{t,d} + \gamma_{t,d} + \delta_{t,d,1}\beta_{d,1} + ... + \delta_{t,d,K_d}\beta_{d,K_d} + \nu_{t,d},\ d \in \{1, ..., D\}. \quad (2.3.1)$$

Another consequence of the redesigns is that they affected the variance of the HT estimates, which resulted in heteroscedasticity in the composite error terms $\nu_{t,d}$. One way to account for this heteroscedasticity could be to use the variance of the HT estimates, say $\hat{\sigma}^2_{HT,t,d}$, as prior information in the model by defining $\sigma^2_{\nu,t,d} = \hat{\sigma}^2_{HT,t,d} + \sigma^2_{\varepsilon,d}$, where $\sigma^2_{\nu,t,d}$ and $\sigma^2_{\varepsilon,d}$ denote the variances of $\nu_{t,d}$ and $\varepsilon_{t,d}$, respectively. Another approach, proposed by Binder and Dick (1990), as well as by Durbin and Quenneville (1997), suggests that $\sigma_{\nu,t,d}$ could be modelled proportionally to the standard error of the HT estimator, i.e. $\hat{\sigma}_{HT,t,d}$. In this application, the variances of the HT estimator are unfortunately not available, and the microdata for their calculation is available only for the most recent years. It would be possible to model variances $\sigma^2_{\nu,t,d}$ proportional to the sample sizes, but these are not available for the major time span of the series either. Therefore, the variance $\sigma^2_{\nu,t,d}$ of the composite error term is modelled as time-varying by allowing different $\sigma^2_{\nu,t,d}$-values for a few sub-periods specific for each domain.

Each of the D models in (2.3.1) uses sample information available over time for the respective domain by means of the trend and seasonal components in order to improve the precision of the $\theta_{t,d}$-estimates. Stacking the univariate models for the

D domains and modelling the correlation between the trend and seasonal disturbances of different domains can further improve the precision of the $\theta_{t,d}$-estimates with the help of sample information from other domains.

In the present application, the so-called *smooth trend model* was chosen to model the trends $L_{t,d}$, $d \in \{1, ..., D\}$. This model is well-known in the econometric literature for its reasonable flexibility and parsimony (Durbin and Koopman (2012), Ch. 3; Harvey (2001)). The smooth trend model (or the integrated random walk) can be defined by the following two equations for series d:

$$\begin{aligned} L_t &= L_{t-1} + R_{t-1}, \\ R_t &= R_{t-1} + \eta_{R,t}, \end{aligned} \tag{2.3.2}$$

where the state variables L_t and R_t are the level and slope of the d-th series, respectively. In the multivariate setting, it is natural to assume that the D trends are correlated because they describe similar processes. Therefore, the following covariance matrix structure is assumed for the normally distributed slope disturbances:

$$Cov(\eta_{R,t,d}, \eta_{R,t',d'}) = \begin{cases} \sigma^2_{\eta_R,d} & \text{if } t = t' \text{ and } d = d', \\ \varsigma_{\eta_R,d,d'} & \text{if } t = t' \text{ and } d \neq d',\ d \in \{1, ..., D\} \\ 0 & \text{if } t \neq t'. \end{cases}$$

For the seasonal component $\gamma_{t,d}$, a trigonometric model is assumed (see Hannan et al. (1970), Koopman et al. (2008), Koopman et al. (2009), Harvey (1989)). This model is widely applied in econometric time series modelling. If s denotes the number of seasons, then the model is defined as:

$$\gamma_{t,d} = \sum_{j=1}^{s/2} \gamma_{t,d,j}, d \in \{1, ..., D\},$$

where $s = 4$ for quarterly data, and thus the seasonal component consists of two harmonics $\gamma_{t,d,j}$, of which the first one is generated by two stochastic variables:

$$\begin{aligned} \gamma_{t,d,1} &= cos(\pi/2)\gamma_{t-1,d,1} + sin(\pi/2)\gamma^*_{t-1,d,1} + \omega_{t,d,1}, \\ \gamma^*_{t,d,1} &= -sin(\pi/2)\gamma_{t-1,d,1} + cos(\pi/2)\gamma^*_{t-1,d,1} + \omega^*_{t,d,1}, d \in \{1, ..., D\}, \end{aligned}$$

where disturbances $\omega_{t,d,1}$ and $\omega^*_{t,d,1}$ are uncorrelated. The last harmonic, in this case the second one, is always generated by only one stochastic variable:

$$\gamma_{t,d,2} = -\gamma_{t-1,d,2} + \omega_{t,d,2}, d \in \{1, ..., D\}.$$

It is assumed that the seasonal disturbance terms of each domain are normally, identically and independently distributed with the variance:

$$\sigma^2_{\omega,d,1} = \sigma^2_{\omega^*,d,1} = \sigma^2_{\omega,d,2} \equiv \sigma^2_{\omega,d},\ d \in \{1, ..., D\}.$$

The equality of the two harmonics' hyperparameters has been assumed for the sake of model parsimony and tested in univariate models using the likelihood ratio test. It turns out that the null hypothesis for the equality of the two variances could not be rejected at any reasonable significance level.

In this application, the covariances between corresponding seasonal disturbances of different domains are set to zero, because modelling these covariances resulted in a very small variance reduction. This is due to the fact that, first of all, the seasonal effect size has turned out to be relatively small compared to the magnitude of the corresponding trends, and, secondly, most of the seasonal component hyperparameters appeared to be not significantly different from zero (domains 1, 3, 4, 5, 7, 9).

It should be noted that the composite error terms $\nu_{t,d}$ in this application have non-zero covariances because certain sampling units may transport goods of several freight categories, and thus appear in different domains. The time-dependent covariance matrix of the composite error terms is denoted by $\boldsymbol{R_t} = E(\boldsymbol{\nu_t}\boldsymbol{\nu_t'})$ whose elements are assumed to have a zero-expectation and to be normally distributed with the following properties:

$$Cov(\nu_{t,d}, \nu_{t',d'}) = \begin{cases} \sigma^2_{\nu,t,d} & \text{if } t = t' \text{ and } d = d', \\ \varsigma_{\nu,t,d,d'} & \text{if } t = t' \text{ and } d \neq d',\ d \in \{1, ..., D\} \\ 0 & \text{if } t \neq t'. \end{cases} \tag{2.3.3}$$

Note that vector $\boldsymbol{\nu_t}$ is also independent of the disturbance terms of the state variables described above. Further, recall that $\sigma^2_{\nu,t,d}$s are modelled as time-dependent to allow for heteroscedasticity in the survey errors. To allow for time-dependent

covariances without increasing the number of hyperparameters, the correlations between the composite error terms of different domains are assumed to be time-invariant. First, the composite error terms and their variance hyperparameters are estimated from a model where the covariances between these error terms are restricted to zero. Next, these error term estimates are standardized as $\hat{\nu}^{St.}_{t,d} = \hat{\nu}_{t,d}/\hat{\sigma}_{\nu,t,d}$ and the correlations between the $\hat{\nu}^{St.}_{t,d}$-terms are calculated. Correlations $\hat{\rho}_{\nu,d,d'} = \hat{Corr}(\hat{\nu}^{St.}_{t,d}, \hat{\nu}^{St.}_{t,d'})$ are calculated over the whole time span. This information is used as input at the next estimation stage in order to approximate the time-varying covariance terms in (2.3.3) as: $\varsigma_{\nu,t,d,d'} = \hat{\sigma}_{\nu,t,d}\hat{\sigma}_{\nu,t,d'}\hat{\rho}_{\nu,d,d'}$.

The obtained model is compared with a model where the covariances between the domains are assumed to be equal to zero. In the present application, it has been found that point-estimates are not affected when the model is augmented with the time-invariant correlations. Only a few domains experience a variance reduction in the state variable estimates, but this effect is negligible. These results are not presented in this paper. For simplicity and with little loss in precision of the state estimates, we proceed by restricting the covariances in question to zero.

To summarize, a time series model for domain estimates has been obtained in order to produce more reliable indicators for the evolution of the parameters of interest. First of all, this model increases the effective sample size for a particular period and domain by using sample information from other time periods and domains. Secondly, sudden measurement error shocks caused by different redesigns are accounted for by explicitly modelling such level changes with intervention variables. Depending on whether or not a survey modification or a change in the population of interest (as in domain 9) are viewed as an improvement, level interventions can be considered to be part of the signal estimates. This paper mainly focuses on the analysis of the trend and signal estimates, the signal being defined as the sum of the trend, seasonal, and level interventions. To obtain series that are not interrupted by survey redesigns, the signal estimates can be corrected using the estimates of the accompanying discontinuities.

2.3.2 Model estimation setting

Structural time series models are generally put in a state-space form and analysed with the Kalman filter, see, e.g., Harvey (1989) or Durbin and Koopman (2012). The state-space form is presented in the Appendix. The model proposed in Sub-

section 2.3.1 contains non-stationary state variables and time-invariant regression coefficients. The state variables are initialised with a diffuse state vector to which the exact initial Kalman filter is applied as in Koopman (1997).

The estimation of the slope disturbance covariance matrix, say $\boldsymbol{Q_R}$, that is a part of the state disturbance covariance matrix $\boldsymbol{Q}$ (see Appendix) is carried out through the Cholesky decomposition of the form $\boldsymbol{Q_R = ADA'}$, where $\boldsymbol{A}$ is a lower triangular matrix of orthonormalised eigenvectors with ones on the diagonal, and $\boldsymbol{D}$ is a diagonal matrix of eigenvalues. The Cholesky decomposition ensures that the maximum likelihood estimate of the matrix $\boldsymbol{Q_R}$ is positive-(semi)definite. In the case of strongly correlated slope disturbances, the trends of the domains can be cointegrated. This means that the non-stationary trends of the D domains (see the trend model (2.3.2)) are driven by fewer than D underlying non-stationary stochastic trends. If an eigenvalue of matrix $\boldsymbol{D}$ is equal to zero, the corresponding domain's stochastic part of the trend can be expressed as a linear combination of the other domains' stochastic trends. Detecting and modelling cointegrated stochastic trends allows us to formulate more parsimonious models resulting in a higher estimation efficiency, and can provide an interpretation of the relationship between the domains. The concept of cointegration and the related testing procedures in the context of state space models are presented, e.g., in chapter 8 of Harvey (1989), Nyblom and Harvey (2001) and Koopman et al. (2008).

The Kalman filter assumes that the hyperparameters in matrices $\boldsymbol{Q}$ and $\boldsymbol{R_t}$, i.e. the state and measurement equation disturbance (co)variances, are known, but this is generally not the case. These hyperparameters are replaced by their maximum-likelihood estimates. The numerical procedure used to solve this nonlinear optimisation problem is the Broyden-Fletcher-Goldfarb-Shanno method (MaxBFGS in the OxMetrics package). The analysis is conducted with OxMetrics.5 (Doornik (2007)) in combination with *SsfPack 3.0* package (Koopman et al. (1999) and Koopman et al. (2008)). The variances of the Kalman filter estimates reported in this paper ignore the additional uncertainty of using maximum likelihood estimates for the hyperparameters instead of their true values. The variance hyperparameters are estimated on the log-scale to avoid negative variance estimates.

The Kalman filter produces what are called *filtered* estimates, which are the optimal state variable estimates for period t on the basis of information accumulated

up to and including period t. These estimates can be improved by various smoothing algorithms, where information is pooled over the entire time span. Smoothed estimates are usually treated as the most realistic ones, since they are based on the entire set of information available. This is true when the focus is on the analysis of unobserved components of the time series model. In the context of this paper, however, structural time series models are used for production purposes in official statistics. In this case, the focus is on filtered, rather than smoothed estimates, since the former ones approximate the conditional expectation of the variable given information up to (and including) time t, and therefore better reflect what can be obtained with this modelling approach in the real production process. However, filtered estimates do not fully imitate the real-time production of official estimates, since the hyperparameter maximum-likelihood estimates in this case are still based on the whole length of the time series. Therefore, we will concentrate on the so-called *concurrent* estimates. Concurrent estimates for period t are based on the Kalman filter, with the hyperparameter maximum-likelihood estimates also being based on the information available up to and including period t. This is sometimes called real-time analysis, since concurrent estimates exactly reflect the real production process outcomes and therefore are even more realistic compared to filtered estimates.

2.3.3 Aggregated Series Estimation

In the DRTS, the design estimates of the aggregated series are not sufficiently precise due to the sample size reduction and sample design modifications, particularly in the second half of the time period under consideration. Fig. 2.2.1 shows that the variability of the observed series does indeed increase after 2003. If the direct estimates at the aggregated level were sufficiently precise, they could be used directly as benchmarks for the domain model estimates, which is frequently done in SAE, Pfeffermann et al. (2014).

The precision of the aggregated series can be improved using a separate univariate model. However, this will inevitably result in differences between figures published at the national level and the sum of published domain model estimates. The method of Lagrange multipliers could be considered as an additional step to solve the problem, but this method requires the availability of covariance estimates between the point-estimates of each of the domains and of the aggregated series. Another approach is to derive the estimates at the aggregated level as a

linear combination of domain estimates from a D-dimensional multivariate model. Which approach is more efficient in terms of signal variances, is an empirical question. Next chapter elaborates on why modelling the aggregated series (that is exactly equal to the sum of the underlying domains) jointly with these domains in a $(D+1)$-dimensional setting is not worth pursuing.

Thus, the choice for obtaining aggregated figures lies between deriving them from a D-dimensional model and modelling the aggregated series univariately. Which approach is going to be chosen by an NSI for official figure production, depends on the accuracy of the aggregated figures obtained under both approaches, and will be made clear based on the results in the next section.

2.4 Model selection and estimation results

2.4.1 Univariate models for nine domains and national level series

The univariate analysis for each of the nine domains is a special case of the nine-dimensional model presented in Subsection 2.3.1 with no correlations among the slope disturbance terms. The univariate analysis is conducted for several purposes. First of all, it allows one to determine necessary interventions for the level and for the variance of the measurement equation term. Secondly, it is of great interest to compare the performance of univariate models with that of multidimensional models in terms of adequacy of point-estimates and variance reduction in the signals and state variables.

As regards the model choice for the trend, the preference has been given to the smooth trend model, as in (2.3.2). This model is a special case of the local linear trend model, whose level equation also has a stochastic term. The absence of this term makes the trend less volatile. The local linear trend model has also been tested and proved to produce very volatile trends, indicating that this model tends to overfit the data. For these reasons, the smooth trend model is preferred. An additional (third) "acceleration" component (see Harvey (1989), Ch.6.1.5 for the quadratic trend model) has also been tested and found to have no added value.

Table 2.4.1 provides a compact summary of level and variance interventions implemented in the univariate models. Not every survey adjustment undertaken by Statistics Netherlands requires an intervention parameter in the model. While some domains are visibly affected by a certain intervention, this effect may not be observable in other domains. To select parsimonious models, we model only those interventions that have empirically been found significant in the univariate models.

Table 2.4.1: Level and variance breaks in the series modelled in a univariate setting

Series	Level interventions (time points where $\delta_{t,k} = 1$)	Sub-periods for which different variance values apply
1. Agriculture	-	1976(1)-2002(4); 2003(1)-2010(4)
2. Food	-	1976(1)-2002(4); 2003(1)-2010(4)
3. Oil	2008(3)-(4)	1976(1)-1993(4); 1994(1)-2002(4); 2003(1)-2010(4)
4. Ores	-	1976(1)-1996(4); 1997(1)-2010(4)
5. Metals	-	1976(1)-2002(4); 2003(1)-2006(4); 2007(1)-2010(4)
6. Minerals	-	1976(1)-1991(4); 1992(1)-2002(4); 2003(1)-2010(4)
7. Fertilisers	2003(1)-2010(4), 2007(1)-2008(4)	1976(1)-2002(4); 2003(1)-2010(4)
8. Chemicals	-	1976(1)-2002(4); 2003(1)-2010(4)
9. Other goods	1997(1)-2002(4), 2003(1)-(4)	1976(1)-1996(4); 1997(1)-2002(4); 2003(1)-2010(4)
Aggregated series	2003(1)-(4)	1976(1)-1984(4); 1985(1)-1987(4); 1988(1)-1993(4); 1994(1)-2002(4); 2003(1)-2010(4)

Numbers in ( ) denote quarters.

An overview of the finally selected level interventions is given in the second column of Table 2.4.1. The time points refer to the periods where the dummy variables

$\delta_{t,k}$ are equal to one. The 2003 level increase in domain 7 (fertilisers), mentioned in Section 2.2, has been modelled as a level shift with an intervention variable that equals one until the end of the series. This domain also suffered an unexpected dip during eight quarters of 2007-2008. An analysis of other statistical figures related to this goods category did not reveal any particular factor that could have caused real changes in this domain of the size reflected in Fig. 2.2.1. Therefore, a separate dummy regressor is introduced that is equal to one during the eight quarters in question. The same holds for domain 3 where a dummy equals one for two quarters 2008(3)-(4)[2]. Domain 9 has two breaks attributed to changes in the population of interest, as described in Section 2.2. In the aggregated series, the only level break identified as significant in a univariate setting is the one for the four quarters of 2003, while the multivariate setting suggests that all the level interventions, including the insignificant ones, enter the aggregated series when it is derived as the sum of the domain estimates.

The significance of variance breaks for the measurement equation disturbance terms was tested using the likelihood ratio test relying on standard asymptotics, as all assumed unit root restrictions are imposed in the estimation process. This test suggests that four variance breaks for the measurement equation disturbances should be modelled for the aggregated series in a univariate setting, namely, four hyperparameters for five different periods of time, since the second and fourth period share the same hyperparameter. An overview of time periods that have separate variance hyperparameters is given in the third column of Table 2.4.1.

Some seasonal hyperparameters have turned out to be close to zero with vast standard errors, and were therefore removed from the model. Likelihood ratio tests suggest that only domains 2, 6 and 8 need a stochastic term for their seasonal component.

The selection of univariate models was based on the likelihood ratio test, as well as on three tests on the normality and independence of standardised innovations: the Doornik-Hansen normality test (see Doornik and Hansen (2008)), the Durbin-Watson test of first-order serial correlation, and a two-sided F-test for heteroscedasticity. Table 2.A.1 contains maximum likelihood estimates of the hyperparameters as well as the model evaluation measures of the finally selected

[2]Here and further quarters in ( )

univariate models for the ten series. It should be noted that several time series fail to satisfy the assumption about normality and independence of their standardised innovations. These are domain 3 with marginally positively autocorrelated standardised innovations; domains 4, 9 and the aggregated series violating normality, of which domain 9 also exhibits heteroscedasticity. However, given the extreme erratic pattern of the series (see Fig. 2.2.1), the obtained model fits can be viewed as satisfactory.

Filtered point-estimates of the signal and its components from the univariate setting are nearly identical to those from a multivariate model discussed in Subsection 2.4.2. Therefore, they are depicted only once and can be found in Fig. 2.4.4 and 2.4.5 for the most interesting domains that are representative of the others. Signals are presented with the level breaks included. The latter, in turn, are presented separately as an indication of the size of a certain discontinuity (see Fig. 2.4.6 for two breaks in domain 9). The standard errors of the filtered signals are depicted in Fig. 2.4.7, where peaking standard errors reflect an additional uncertainty brought about by the inclusion of a level break. It takes about one year for the standard errors to decay and stabilize around a new level (see domain 9 and the aggregate series in Fig. 2.4.7).

2.4.2 Multivariate model for the nine domains

While the implementation of univariate structural time series models makes it possible to borrow strength over time, multivariate models also benefit from information available over space. The present multivariate model is based on the level and variance breaks described for the nine univariate models in Table 2.4.1. These breaks remain significant in the multivariate setting as well. The multivariate model allows for non-zero covariances among the slope disturbance terms of different domains, and thus for common trends (presented in the next subsection), which significantly increases the maximum value of the likelihood function. While the point estimates remain almost unchanged, the variance of the signals gets considerably reduced (details are addressed in Subsection 2.4.4). Another improvement brought about by modelling the correlation among the slope disturbances is the absence of the serial correlation that was present in the standardised innovations of the third domain's univariate model (results are presented in Table 2.A.2 in the Appendix).

Insignificant eigenvalues of the slope disturbance covariance matrix are removed step-wise until the number of common trends is identified. Modelling the slope covariances has led to insignificant variances of the seasonal stochastic terms. However, three of them (those of domains 6, 8 and 9), still being close to zero, turned out to be one order of magnitude larger than the others. The question at this point is how to proceed with the common trend model: by either first removing insignificant seasonal disturbances and then removing eigenvalues that are close to zero, or the other way around. These two approaches lead to different models. First removing insignificant seasonal hyperparameters leads to a model with six common trends and with two significant seasonal hyperparameters (domains 6 and 8) that are excessively large. If insignificant eigenvalues of the slope disturbance covariance matrix are removed at the first stage, the model is featured by only five common trends and by four significant seasonal disturbance variances (domains 2, 3, 6 and 8) whose magnitudes are comparable to those in univariate models. Five turned out to be the minimum number of common trends required. Taking into account the large dimension of the model, the Bayesian Information Criterion (BIC), instead of the Akaike Information Criterion, is used to select between the two models (Durbin and Koopman (2012), Ch. 7.4). It turns out that the model with five common trends is superior to the model with six common trends. The resulting zero-eigenvalues are those of domains 4, 7, 8, and 9.

It should be noted that the whole data set is used to determine the number of non-stationary common trends. It can be expected that adding level breaks to account for discontinuities might affect the number of common trends before and after a redesign. In fact, a concurrent estimation of the complete nine-dimensional model, i.e. with all the seasonal hyperparameters estimated, suggests that the number of common trends varied between five and six up through the first quarter of 2010, whereafter it remained equal to five until the end of the series.

The selected model with five common trends results in noticeably lower signal variances for domains 4 and 8 compared to the model with six common trends. The variance of the other corresponding domains across the two models is almost the same. The selected model's diagnostics based on standardised innovations, as well as the maximum-likelihood hyperparameter estimates along with their asymmetric confidence intervals, can be found in the supplementary file to this article. The asymmetry in the confidence intervals arises from the fact that the hyperpa-

rameters are estimated on a log-scale. The Fisher information matrix is used to estimate the asymptotic standard errors of the log-transformed hyperparameters, whereafter the confidence interval bounds are transformed back to the original scale. Point-estimates obtained from the nine-dimensional and univariate models are virtually the same. The variance comparison will be illustrated in Subsection 2.4.4.

2.4.3 Interpretation of Common Trends

The previous subsection demonstrates that there are five stochastic trends that drive the development of all the nine domains. The trend equation could be expressed in the following form:

$$L_{t,d} = L_{0,d} + R_{0,d}t + \sum_{j=1}^{t-1}\sum_{i=1}^{j}\eta_{R,i,d}, \tag{2.4.1}$$

where $L_{0,d}$ is an intercept and $R_{0,d}$ is the slope coefficient of the deterministic part in the trend (Koopman et al. (2009), Ch. 6.4.4 and 9.1.4.2). The double partial sum of ηs is the stochastic part of the trend. The implication of the common trend model is that, for each of these D trends, this partial sum can be expressed as a linear combination of a smaller number of common trends. The system can also be written in a matrix form:

$$\boldsymbol{L_t} = \boldsymbol{\Theta L_t^\dagger} + \boldsymbol{L_0} + \boldsymbol{R_0}t, \tag{2.4.2}$$

where $\boldsymbol{\Theta}$ is a $D \times m$ matrix of factor loadings, m being the number of common trends; vector $\boldsymbol{L_t^\dagger}$ contains time-specific estimates for the m common stochastic trends. The first m elements of vectors $\boldsymbol{L_0}$ and $\boldsymbol{R_0}$ are zeros, and the remaining entries are $L_{0,d}$ and $R_{0,d}t$ as described in (2.4.1).

There are different ways to construct common trends and factor loading matrices. One option is to take m orthogonal common factors, for which the factor loading matrix $\boldsymbol{\Theta}$ is equal to the lower triangular matrix $\boldsymbol{A}$ from the Cholesky decomposition of the covariance matrix $\boldsymbol{Q_R}$. Then, the first common trend will be equal to the trend of the first domain, while the other trends will be expressed as linear combinations of the identified common trends. Another option is a factor rotation that makes the common trends equal to the trends of the m domains that have

non-zero eigenvalues. This factor rotation turns the upper part of the $\boldsymbol{\Theta}$ matrix into an m-dimensional identity matrix and results in correlated common trends. This approach shows how the remaining $D-m$ trends depend on the m extracted trends of the domains with non-zero eigenvalues. After the order of the series is slightly changed (in the following order: 1, 2, 3, 5, 6; 4, 7, 8, 9), i.e. all the trends with zero-eigenvalues are kept as the last ones in the system, the factor loadings in the lower $(D-m)\times m$ part of the modified matrix $\boldsymbol{\Theta}^*$ can be obtained from the eigenvector matrix of the covariance matrix $\boldsymbol{Q_R}$, as shown in Ch. 6.4.1 of Koopman et al. (2009). It results in the following factor loading matrix:

$$\boldsymbol{\Theta}^* = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0.31 & -0.08 & 0.35 & -0.74 & 0.03 \\ 0.40 & 0.29 & 1.92 & -0.16 & -0.34 \\ 2.36 & -0.31 & 3.63 & -1.25 & -0.04 \\ 4.08 & -1.86 & 1.10 & -6.46 & 0.89 \end{pmatrix}.$$

The common trends correspond to the trends of domains 1, 2, 3, 5 and 6. Each of the zero-eigenvalue trends 4, 7, 8 and 9 is defined by a linear combination of the first five stochastic trends, with the coefficients from the corresponding rows of $\boldsymbol{\Theta}^*$, and by an intercept and a deterministic time trend, as in (2.4.2). The estimates of the intercepts and deterministic trend coefficients are given in Table 2.4.2.

Not only does the common trend specification help to reduce the number of estimated hyperparameters, but it is also useful in the analysis of the relationship between the trends of different domains. This subsection dwells upon smoothed trends (rather than upon filtered or concurrent ones), since the focus of interest lies in this relationship, for which it is best to use all the available information. In this application, the fixed-interval smoother is used (Durbin and Koopman (2012), Ch. 4.3.1).

Table 2.4.2: Intercept and Slope Coefficients for the Common Trend Model

Domain	4	7	8	9
$L_{0,d}$	-977.8	-5381.7	-10510.3	-1252.9
$R_{0,d}$	4.8	20.8	50.9	15.6

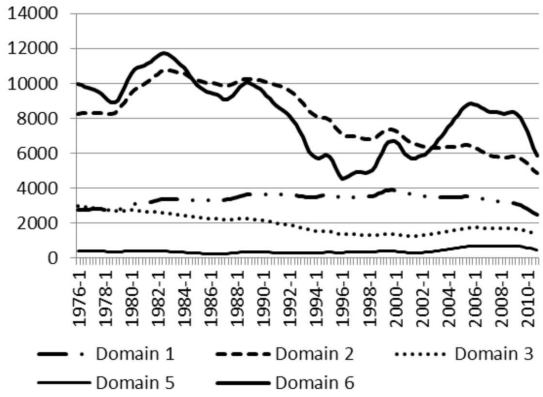


Figure 2.4.1: Smoothed common correlated factors, in kilotons.

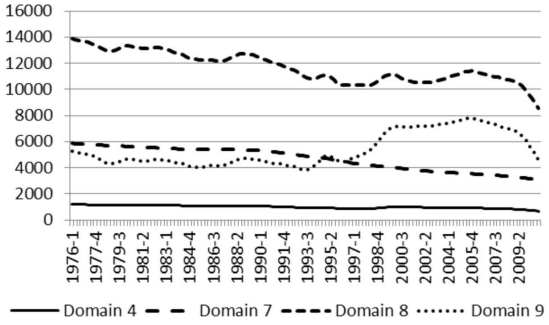


Figure 2.4.2: Smoothed zero-eigenvalue trends as a combination of common factors, excluding the intercept and deterministic trend, in kilotons.

Fig. 2.4.1 displays the smoothed correlated common factors that are equal to the extracted smoothed trends of domains 1, 2, 3, 5 and 6. The smoothed trends of the domains with zero-eigenvalues can be found in Fig. 2.4.2. The two figures along with the matrix , give an insight as to how much each of the common factors contributes to the trend of the remaining domains (4, 7, 8 and 9). For instance,

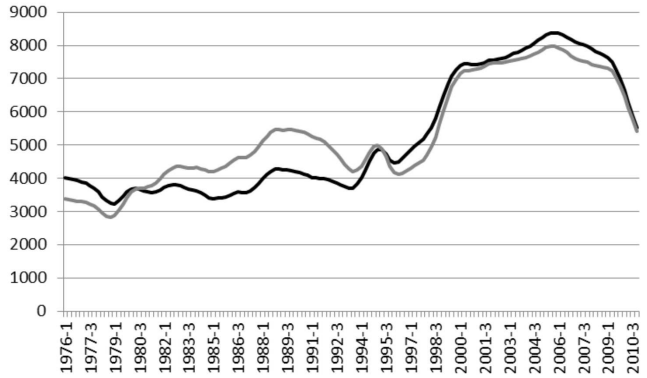


Figure 2.4.3: Smoothed trends of domain 9 (black line) and domain 4, scaled by a factor of 15 (grey line).

the trends of the large-scaled domains 8 (chemicals) and 9 (other goods) exhibit a very similar behaviour (Fig. 2.4.3). They also resemble the trend of domain 6 (minerals) in terms of local extrema. Indeed, domain 9 has a large positive factor loading that corresponds to the common factor of domain 6. Although the value 0.89 is smaller than the factor loading for the common factor of domain 1 (4.08), the common factor of domain 6 drives the development of the trend of domain 9 to a larger extent, since the magnitude of domain 6 (in terms of tons transported) is much larger than that of domain 1 (see Fig. 2.2.1 for the domain magnitude). Unlike in domain 9, the trend of domain 8 has a small negative factor loading on common factor 6. Still, trends 8 and 9 behave in a similar way. This is confirmed by the similarity of their factor loadings on the rest of the common factors. Taking into consideration the large scale of domains 6, 8 and 9 in terms of tons transported, the similarities between them seem quite plausible, since these domains are very likely to reflect overall developments in the economy.

As for the zero-eigenvalue trend of domain 4 (ores), it is correlated to a large degree with the trend of domain 5 (metals), which is confirmed by the largest factor loading for this common trend in an absolute value (-0.74). This gives empirical evidence to the fact that two separate stochastic factors are redundant to explain the variation of these two similar goods categories. The negative sign in the factor loading could be explained by a substitution effect between these domains.

Further, the trends of domains 4 and 9 are strikingly similar, as Fig. 2.4.3 shows. Here, the trend of domain 4 is magnified with a factor of 15, which allows us to superimpose it on the graph of trend 9. This similarity is also supported by the factor loadings of these two trends: positive and negative factor loadings correspond to each other, and so do their largest (in absolute terms) values (e.g., -0.74 and -6.46).

2.4.4 Estimation results

This subsection compares the estimation results obtained with the univariate and nine-dimensional models. For the aggregated series, estimates based on the univariate model are compared with estimates derived from the nine-dimensional model, as described in Subsection 2.3.3. Fig. 2.4.4 shows filtered and concurrent signal estimates of the national level series and the three most interesting domains. Fig. 2.4.6 shows some of the level interventions employed in the model. The trend estimates are presented in Fig. 2.4.5. The trend of domain 4 closely resembles the one of domain 9 and therefore is not presented here. Point-estimates in the two graphs are obtained from the nine-dimensional model, but they are basically the same for the univariate models. Moving backwards in calculating concurrent estimates, one comes to a point in time when the model needs to be vastly re-specified, i.e. certain level and variance breaks have to be removed, and common factors possibly re-identified. Therefore, concurrent estimates are calculated starting from 2007(1).The estimates for the periods before 2007(1), depicted in one line together with these concurrent estimates are, in fact, filtered estimates obtained by means of hyperparameters based on the information available up to and including 2006(4). The whole series is referred to as *concurrent estimates*. The difference between the filtered and concurrent point-estimates of the three domains becomes quite salient as early as in late 90's/early 2000's, especially in domain 5. This might be explained by the frequent changes introduced into the survey during this period. In domain 5, for instance, an increased variation due to the redesign of 2003 is accounted for by the variance break in the measurement equation error term. However, the gradually rising level of point-estimates, caused by this redesign, cannot be remedied by modelling the aforementioned variance break. Allowing for a variance break in the trend is likely to result in a model that better describes the visible pattern of the series. It is, however, difficult to motivate that a survey redesign results in a variance break in the trend component of the population parameter. Therefore, with a time-constant trend hyperparameter,

the concurrent estimates continue to follow the same pattern as they did before 2003, far into the more volatile post-redesign period. In the last couple of years, the concurrent and filtered estimates converge.

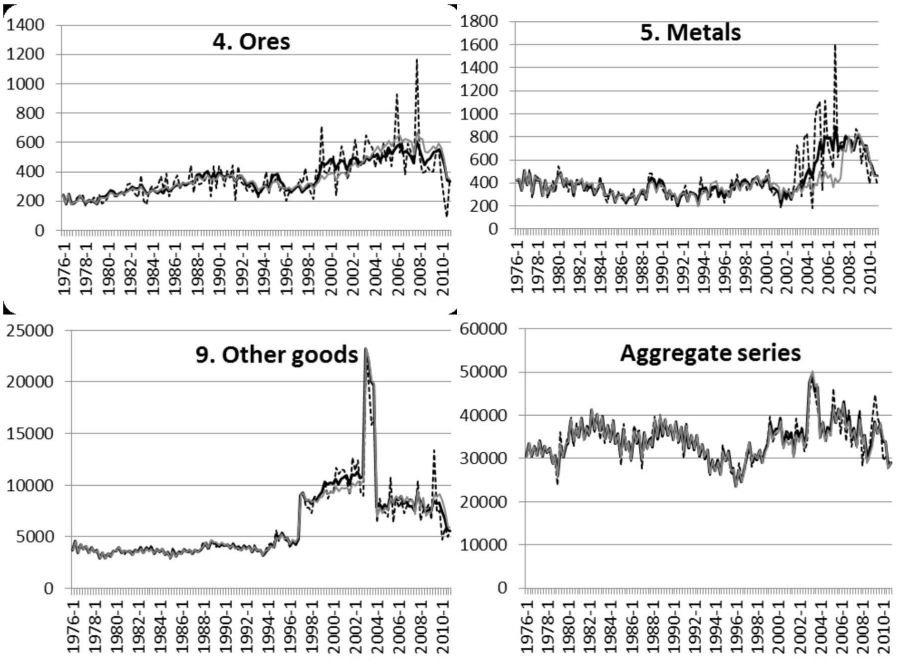


Figure 2.4.4: Dutch own-account road transportation, in kilotons: Horvitz-Thompson (dashed line) and the nine-dimensional model-based filtered (black solid line) and concurrent (from 2007(1), grey solid line) estimates of the signal.

Fig. 2.4.7 illustrates a reduction in the signal variance estimates when one moves from the univariate models to the multivariate one. The standard errors of the filtered signals are depicted in black, and those of the concurrent estimates in grey. The variance of the filtered and concurrent signal estimates of the aggregated series coming from the nine-dimensional model for the domains are calculated as the sum of domain variances and the corresponding covariances. The level interventions for 2007-2008 in domains 3 and 7 are very short and are not included in concurrent estimation for a better real-time imitation, since observations at the end of a series are unlikely to be identified as outliers. Apart from that, the variance break in domain 5 is included only in 2008(3), when the change in the variance becomes sufficiently pronounced.

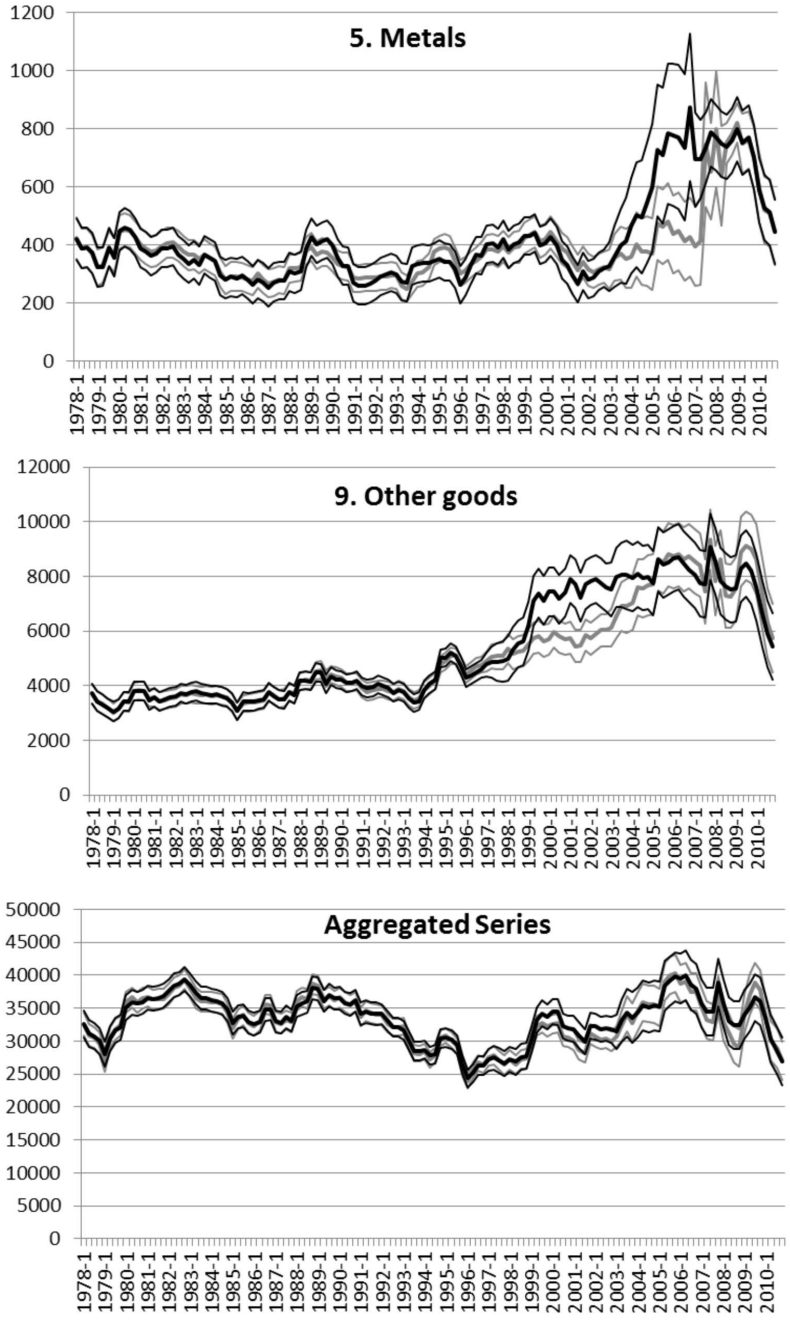


Figure 2.4.5: Dutch own-account road transportation, in kilotons: the nine-dimensional model-based filtered estimates of the trend (thick black line) and their 95%-confidence interval bands (thin black lines); concurrent estimates and 95%-confidence interval bands from 2007(1) in grey (thick and thin line, respectively).

As can be seen, the multivariate nine-dimensional model, as well as the estimates for the aggregated series derived therefrom, outperform the univariate models, especially in the last part of the time span. The peaks in the filtered estimate standard errors are caused by intervention variables that model discontinuities in the level of domain 9 and thus also appear in the aggregated series.

To analyse the increase in precision obtained with this modelling approach, standard errors of the model predictions must be compared with the standard errors of the direct estimates. As was mentioned in Subsection 2.3.1, the standard errors of direct estimates are not available for this survey. The micro-data is available for the last few years, so it can be used to approximate the variances of the HT estimates. Design variances have been approximated assuming a two-stage stratified sampling design and sampling with replacement in the first stage, with the companies as the primary sampling units (PSU) and the vehicles as the secondary sampling units. This estimator is commonly used for complex two- or multi-stage sample designs, see Särndal et al. (1992), Ch. 4.6 (equation (4.6.2)). An additional complication is that vehicles belonging to the same owner were initially drawn from different strata. This led to a situation where certain companies occurred simultaneously in more than one stratum. Therefore, the strata have been collapsed so that the new stratification scheme is based only on the economy branches the PSUs belong to, without differentiating between the vehicles' characteristics. The design variances have been approximated only for 2008(1). In Table 2.4.3, the standard errors of the HT estimator and the concurrent estimates are compared for 2008(1). In general, the standard errors of the HT estimator are much larger than the standard errors of the concurrent estimates. An exception is domain 2. One explanation is that the approximation of the standard errors assuming a stratified two-stage sampling design only captures variation over space, but not over time (recall that PSUs are assigned randomly to a certain week in a quarter). Another explanation is that the design variance estimates themselves are subject to uncertainty and thus fluctuate over time. As a result, the real design variance could accidentally be underestimated in one period or domain.

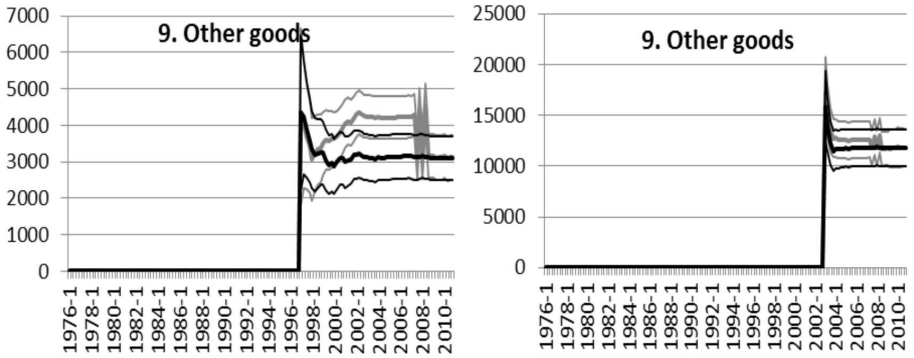


Figure 2.4.6: Filtered (in black) and concurrent (from 2007(1) to the end, in grey) nine-dimensional model-based estimates and their 95%-confidence interval bands for discontinuities in domain 9, in kilotons. Left: level shift 1997(1)-2002(4); right: level shift 2003(1)-(4).

Table 2.4.3: Standard errors of the concurrent model-based signal estimates, Horvitz-Thompson (HT) standard error estimates, and approximated HT standard error estimates; in kilotons, 2008(1).

Domains	Univariate model	Nine-dimensional model	HT estimator	Approximated HT estimator $\hat{\sigma}_{\nu,t,d}$
1. Agriculture	156	162	529	405
2. Food	648	520	481	1382
3. Oil	158	125	451	447
4. Ores	57	49	163	136
5. Metals	58	58	259	311**
6. Minerals	938	875	1063	1388
7. Fertilisers*	204	196	226	681
8. Chemicals	765	488	1105	1586
9. Other goods	816	618	723	1625
Aggregated series	2065	1892	2361	3152

* Concurrent estimates of this domain in 2008(1) did not include the level intervention for 2007(1)-2008(4).
** In the course of concurrent estimation of this domain, the variance break in 2008(1) was not included, as the change in the variance had not yet been sufficiently pronounced

Under the assumption that the measurement equation error term is dominated by the sampling error, the model variance estimates ($\hat{\sigma}_{\nu,t,d}$) of this error term are a better proxy for the variance estimates of the HT estimator. These maximum likelihood model estimates from the nine-dimensional model are presented in the last column of Table 2.4.3. Such a design variance approximation approach is

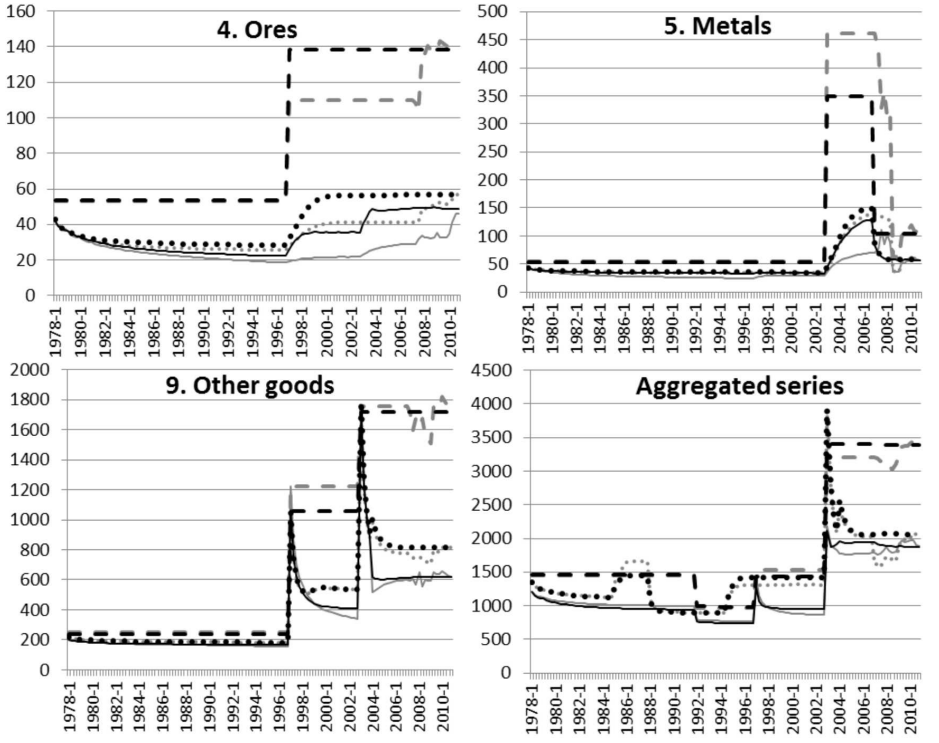


Figure 2.4.7: Hyperparameter estimates for the standard errors (in kilotons) of the measurement equation error term from the nine-dimensional model (black dashed line) and the standard errors of the filtered signal estimates obtained from different state-space models: univariate model (black dotted line), nine-dimensional model (thin black line); standard errors of the concurrent estimates from 2007(1) are depicted in the same style but in grey.

advocated by an empirical finding of Krieg and Van den Brakel (2012), where the standard deviation of the measurement equation error term is defined as the product of a hyperparameter and the standard error of the direct estimates. The maximum likelihood estimates of this hyperparameter are nearly equal to one for all the domains considered in their work. Fig. 2.4.7 presents the standard errors of the measurement equation error term and standard errors of the filtered and concurrent estimates of the signal. One can see that the model-based approach offers a considerable variance reduction as compared to the HT-estimator. This reduction is particularly salient in the most recent period when effective sample sizes get smaller. The standard errors of the HT estimates of small domains, such

as domains 4 and 5, are reduced by 70 to 80 percent, whereas those of large-scale series, like domain 9 or the aggregated series, are reduced by 50 percent in the latter part of the time period (after 1997).

The filtered and concurrent estimates that have been considered so far, illustrate what can be obtained with this model-based approach when the sample information observed until period t is used to produce optimal estimates for period t. An issue with this estimation procedure in production is that these estimates can be improved if new information becomes available after period t. Depending on the size of the adjustments, it might be necessary for an NSI to consider a revision strategy. Besides the variance of the filtered signal, it is interesting to analyse the variance of revisions $(\hat{l}^d_{t|t+k} - \hat{l}^d_{t|t})$, where $\hat{l}^d_{t|t+k}$ denotes a revised signal estimate of domain d for period t using the information available at time $(t+k)$, whereas $\hat{l}^d_{t|t}$ denotes a filtered estimate at time t. Revisions may be quite significant as it is usually difficult to produce a reliable estimate at the end of the series. This issue is discussed in Orphanides and Van Norden (2002) or Planas et al. (2013), and is closely related to the problem of revision strategy for seasonally adjusted figures. Many NSIs continuously revise their official releases of seasonally adjusted series, since estimates of seasonal effects can be improved as new information becomes available. For a revision strategy, it is important to choose the best revision horizon. Large revisions indicate that a certain revision strategy might be required. When revisions are small, it may be more convenient to leave the initially published figures unchanged.

To illustrate the size of revisions in the DRTS, the revised signal estimates at four different horizons (one-, two-, four-, and eight-quarters) are plotted together with 95-percent confidence intervals of the filtered estimates in Fig. 2.4.8 (domain 5 is chosen as an example typical of the rest of the series). These revised signal estimates are calculated starting from 1988(3), namely, for the last 89, 88, 86 and 82 quarters of the sample for one-, two-, four-, and eight-quarter revisions, respectively. The Kalman filter is run conditionally on the hyperparameter set estimated on the basis of the complete sample. The revisions at all the above-mentioned horizons remain within the confidence interval bands of the filtered estimates.

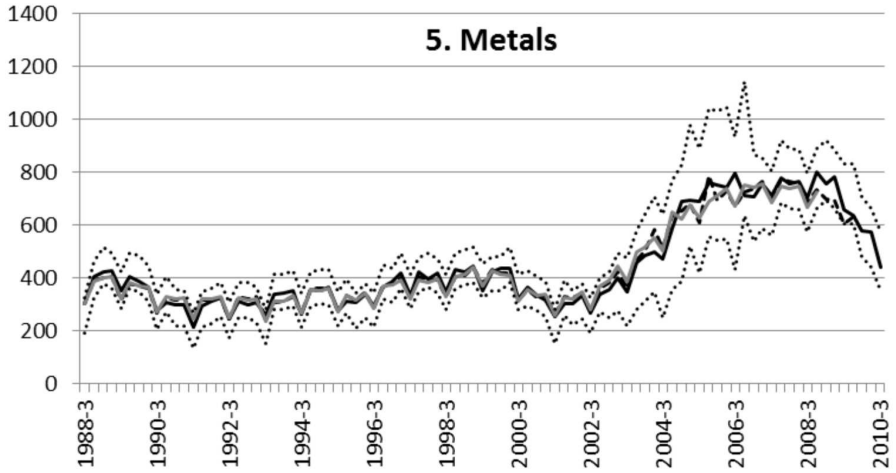


Figure 2.4.8: The filtered estimates' confidence interval bands (dotted lines) and revised signal estimates in between: after 1 quarter (black solid line), 4 quarters (black dashed line), and 8 quarters (grey solid line).

Table 2.4.4 presents the sample mean of absolute revisions (AR) in absolute values:

$$MAR_k = \frac{1}{89-k} \sum_{t=51}^{140-k} |\hat{l}^d_{t|t+k} - \hat{l}^d_{t|t}|$$

Sample means of relative revisions (RR) in absolute values are defined as:

$$MRR_k = \frac{100\%}{89-k} \sum_{t=51}^{140-k} |\hat{l}^d_{t|t+k} - \hat{l}^d_{t|t}|/\hat{l}^d_{t|t}$$

and presented in Table 2.4.5.

The MRR mostly remain under 5 percent for all the revision horizons. The small domains 5 and 7 are exceptions with revisions occasionally exceeding 7 percent depending on the revision horizon. For the choice of the most appropriate revision horizon, it is important to note that MAR and MRR clearly increase with the revision horizon, but at a decreasing rate. Namely, the increment in the mean revisions of the domains is the highest when one moves from no revision to one- and two-quarter horizons. After the second quarter, little is changed by subsequent

Table 2.4.4: Sample mean and standard deviation of the signals' absolute revisions after k quarters and the average standard error of the filtered signals based on 1988(3)-2010(4), kilotons.

Series	MAR_k, kilotons				$SD(AR_k)$, kilotons				$\frac{1}{89}\sum_{t=51}^{140}\sqrt{\widehat{MSE}(\hat{l}^d_{t\|t})}$
	k=1	k=2	k=4	k=8	k=1	k=2	k=4	k=8	
1	46	60	72	79	63	81	89	95	124
2	122	151	162	188	158	199	209	257	327
3	26	35	53	60	34	45	67	72	100
4	15	20	20	18	21	27	28	22	33
5	23	26	31	30	34	37	41	39	50
6	197	244	299	362	266	333	406	491	601
7	43	60	79	84	75	89	120	150	130
8	123	161	171	163	176	220	238	207	335
9	219	313	334	297	360	552	615	591	437
Aggregated	630	831	857	855	862	1081	1121	1101	1255

Table 2.4.5: Relative revisions of the signals after k quarters in percentage: sample mean and standard deviation over time.

Series	MRR_k, %				$SD(RR_k)$, %			
	k=1	k=2	k=4	k=8	k=1	k=2	k=4	k=8
1	1.37	1.78	2.09	2.27	1.92	2.47	2.68	2.77
2	1.83	2.31	2.43	2.85	2.51	3.31	3.49	4.50
3	1.81	2.38	3.45	3.91	2.81	3.48	4.61	4.97
4	3.61	4.58	4.87	4.54	4.84	5.98	6.19	5.54
5	4.94	5.93	7.07	7.23	6.48	7.45	8.62	8.83
6	2.88	3.53	4.24	5.14	3.90	4.88	5.62	6.82
7	4.02	5.43	7.22	7.95	6.65	7.62	10.26	13.01
8	2.25	2.89	3.09	3.13	3.06	3.69	3.98	3.86
9	2.69	3.62	3.79	3.34	3.73	4.80	5.12	4.59
Aggregated	0.02	0.02	0.02	0.02	0.02	0.03	0.03	0.03

revisions. This suggests that two-quarter revisions are worth considering. As for the aggregated series, the MRR is as small as 0.02 percent for all the revision horizons, indicating that the initially estimated aggregated series is quite reliable. However, if the domain estimates are subject to revision, so will the aggregated series estimates be.

Assuming that the difference between the filtered signal estimates $\hat{l}^d_{t|t}$ and smoothed estimates $\hat{l}^d_{t|t+k}$ is stationary and independently distributed, the ARs' sample standard deviations can act as a proxy for the volatility measure of absolute revisions.

These sample estimates are given in Table 2.4.4. The *AR*s' sample standard deviations can also be compared to filtered signal standard errors. Although the uncertainty of the filtered signals has been found to be different in several subperiods, an average of these standard errors (over 1988(3)-2010(4)) is presented in the last column of Table 2.4.4 for indicative purposes. As can be seen, the standard deviation of absolute revisions never exceeds the average standard error of the filtered signal estimates, with an exception of domain 9.

The sample standard deviations of the *RR*s are presented in the last part of Table 2.4.5 to provide an indication of how far the revisions can reach in relative terms. The standard deviations of the *RR*s and *AR*s show that the eight-quarter revisions are more volatile in all the domains, compared to revisions at the other three horizons, just as expected. As for cross-sectional differences, the signal estimates of domain 4, 5 and 7 are the least stable, with the *RR*s' standard deviations reaching sometimes 7 percent, whereas the aggregated series filtered estimates do not seem to be much affected by revisions. The three above-mentioned domains are the smallest ones and feature a highly volatile pattern of design estimates.

2.5 Discussion

This paper presents an application of univariate and multivariate structural time series models to the domestic own-account segment of the Dutch Road Transportation Survey (DRTS). This is a longstanding repeated survey conducted with the purpose to produce reliable series of survey estimates that are comparable over time. Two problems are solved simultaneously with the time series modelling approach. The first problem is frequent survey redesigns that have led to several level shifts in the direct estimates of this survey. So-called discontinuities hamper the comparability of the published figures over time. Secondly, several survey modifications have reduced the effective sample size, which resulted in variance breaks and gradually increasing variances. This made series of the direct survey estimates too imprecise and excessively volatile.

The DRTS can be improved by developing a multivariate time series model that accounts for level and variance breaks and improves the series precision by borrowing strength over time and space. Our findings suggest that a multivariate model for the domains outperforms the univariate setting. Both models account

for discontinuities, but the multivariate one features a larger variance reduction and better model evaluation measures. The standard errors of the design-based estimator can be reduced by 40 to 60 percent in large domains and/or in the aggregated series in the period following the most recent major survey redesign. When it comes to small domains, the reduction in the standard errors can reach 70 or 80 percent.

The national level estimates are derived from the multivariate model developed for the domains. In the case of the DRTS, this approach resulted in lower signal variances of the aggregated series compared to the ones obtained from a univariate model that was applied to the aggregated series itself. This gives evidence to the fact that the multivariate model for the domains is well specified, and that accounting for survey modifications at a lower aggregation (domain) level produces better outcomes for the aggregated series. An additional advantage is that time series estimates at the domain and aggregated level are consistent by definition.

The analysis of revisions shows that concurrent estimates can be improved with the information that becomes available in the next two quarters. The model estimates are, however, not largely affected by revisions. In extreme cases of small volatile domains with level breaks, the standard deviation of relative revisions may exceed 10 percent, whereas that of the aggregated series remains as low as 0.03 percent even at the eight-quarter revision horizon. In the present case, a two-quarter horizon may be considered, since only minor corrections are observed at longer revision horizons.

An additional advantage of the structural time series modelling approach is that it offers a breakdown of the signal into the trend, seasonal and intervention components. Seasonally adjusted series are therefore obtained as a by-product.

The technique presented here can be applied to any small area estimation problem, where a survey is repeatedly conducted and suffers from small sample sizes, as well as certain side effects of redesigns. Further improvements, specific to this application, can be derived with the help of information from other segments covered by this survey: hire-and-reward and international transportation. Another potential improvement could arise from modelling the variance structure of the measurement equation error term with the help of design-variance estimates for

the time points where micro-data is available. If these are not available, as is the case here, the measurement equation error variances can be made time dependent by defining separate hyperparameters for several time-periods. However, this approach will be limited in its application to the data from the past. This is due to the fact that, once a break in the variance occurs, a certain number of observations is needed to estimate another hyperparameter for the new sub-period, while the figures have to be produced and published on a continuous basis. When a sufficient number of observations become available after the break, the model has to be adjusted, which might require a revision of the published figures. This can be avoided if design variances are produced along with the point-estimates.

2.A State Space Representation, Hyperparameter Estimates and Model Diagnostics

This appendix presents the state space form of the multivariate structural time series model for D domains, as defined in Subsection 2.3.1. The multivariate model (2.3.1) can be written in the following vector notation:

$$\hat{\boldsymbol{Y}}_{\boldsymbol{t}} = \boldsymbol{L_t} + \boldsymbol{\gamma_t} + \boldsymbol{x_{t,1}}\beta_1 + ... + \boldsymbol{x_{t,K}}\beta_K + \boldsymbol{\nu_t}, \tag{2.A.1}$$

where all vectors have dimension D. In this notation, $K = \sum_{d=1}^{D} K_d$ stands for the total number of level interventions in the multivariate model. A distinctive feature of this application is that level breaks do not generally occur in all the domains. Therefore, the regression coefficients β_k are meant for a specific intervention and are indexed from 1 to K (hence the omission of the subscript d). Some domains may have several β_k, as shown in (2.3.1), while others may have no (significant) discontinuities. Each of the vectors $\boldsymbol{x_{t,k}} = (0, ...0, \delta_{t,k}, 0, ...0)$, being D-dimensional, contains $D-1$ zeroes and a dummy variable $\delta_{t,k}$. The position of $\delta_{t,k}$ in vector $\boldsymbol{x_{t,k}}$ is defined by the number of the domain the intervention is effective for. Every intervention starts at a particular point in time and lasts for a particular number of time points, during which the intensity of the intervention remains constant and domain-specific.

The state space representation of (2.A.1) consists of a measurement equation and a transition equation. The measurement or signal equation: $\hat{\boldsymbol{Y}}_{\boldsymbol{t}} = \boldsymbol{Z_t}\boldsymbol{\alpha_t} + \boldsymbol{\nu_t}$ reflects the relation between the observed design estimates $\hat{\boldsymbol{Y}}_{\boldsymbol{t}}$ and the vector $\boldsymbol{\alpha_t}$ with unobserved state variables, through a time-dependent design matrix $\boldsymbol{Z_t}$. In this case, $\boldsymbol{\alpha_t}$ contains the level, slope, $(s-1)$ seasonal harmonics per each series, and the β_k-regression components: $\boldsymbol{\alpha_t} = (\boldsymbol{\alpha_t^L}\ \boldsymbol{\alpha_t^\gamma}\ \boldsymbol{\alpha_t^\beta})$, where $\boldsymbol{\alpha_t^L} = (L_{t,1}\, R_{t,1} ... L_{t,D}\, R_{t,D})$, $\boldsymbol{\alpha_t^\gamma} = (\gamma_{t,1,1}\ \gamma_{t,1,1^*}\ \gamma_{t,1,2} ...\ \gamma_{t,D,1}\ \gamma_{t,D,1^*}\ \gamma_{t,D,2})$, and $\boldsymbol{\alpha^\beta} = (\beta_1 ... \beta_K)$. The time-dependent design matrix is defined as:

$$\boldsymbol{Z_t} = [\boldsymbol{I_{[D]}} \otimes (1\ 0)\ \ \boldsymbol{I_{[D]}} \otimes (1\ 0\ 1)\ \boldsymbol{Z_{\beta,t}}],$$

where $\boldsymbol{I_{[p]}}$ is a p-dimensional identity matrix, and $\boldsymbol{Z_{\beta,t}}$ is a matrix consisting of $K = 5$ column vectors $\boldsymbol{x_{t,k}}$:

$$\boldsymbol{Z_{\beta,t}} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & \delta_{t,5} \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \delta_{t,1} & \delta_{t,2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \delta_{t,3} & \delta_{t,4} & 0 \end{pmatrix}.$$

Furthermore, $\boldsymbol{\nu_t}$ denotes a vector with composite error terms that are assumed to be normally distributed with a zero expectation and the symmetric contemporary covariance matrix $\boldsymbol{R_t} = E(\boldsymbol{\nu_t}\boldsymbol{\nu_t'})$. This matrix is diagonal assuming that domains are independent: $\boldsymbol{R_t} = Diag(\sigma^2_{\nu,t,1}\, \sigma^2_{\nu,t,2} ... \sigma^2_{\nu,t,D})$.

The transition equation describes how each state variable evolves over time through a time-invariant design matrix $\boldsymbol{T}$: $\boldsymbol{\alpha_{t+1}} = \boldsymbol{T}\boldsymbol{\alpha_t} + \boldsymbol{\eta_{t+1}}$. Here,

$$\boldsymbol{T} = Blockdiag[\boldsymbol{T_L}\boldsymbol{T_\gamma}\boldsymbol{T_\beta}]$$

with $\boldsymbol{T_L} = \boldsymbol{I_{[D]}} \otimes \boldsymbol{T_{LR}}$, $\boldsymbol{T_\gamma} = \boldsymbol{I_{[D]}} \otimes \boldsymbol{H}$, and $\boldsymbol{T_\beta} = \boldsymbol{I_{[K]}}$, where $\boldsymbol{T_{LR}} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, $\otimes$ is the Kronecker product which multiplies every element of the first matrix by the second matrix, and H is the design matrix for the seasonal component's harmonics, with s denoting the number of seasons:

$$\boldsymbol{H} = \begin{pmatrix} cos(\frac{2\pi}{s}) & sin(\frac{2\pi}{s}) & 0 \\ -sin(\frac{2\pi}{s}) & cos(\frac{2\pi}{s}) & 0 \\ 0 & 0 & -1 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix}. \tag{2.A.2}$$

The transition matrix $\boldsymbol{T_\beta}$ for the dummy regression coefficients contains a K-dimensional identity matrix $\boldsymbol{I_{[I]}}$. The structure of the $\boldsymbol{T_\beta}$-matrix, together with a zero-variance of the corresponding state stochastic terms, means that the level intervention coefficients are constant over time $\beta_{k,t} = \beta_{k,t-1}$, $k = 1, ..., K$, $t = 2, ..., T$.

Vector $\boldsymbol{\eta_t}$ is assumed to be a zero-expectation vector of normally, identically and serially independently distributed state disturbances with the symmetric contem-

porary covariance matrix $\boldsymbol{Q} = E(\boldsymbol{\eta_t}\boldsymbol{\eta_t'})$. No correlation is assumed between the $\boldsymbol{\eta_t}$- and $\boldsymbol{\nu_t}$-terms.

The state noise covariance matrix $\boldsymbol{Q} = E(\boldsymbol{\eta_t}\boldsymbol{\eta_t'}) = Blockdiag[\boldsymbol{Q_l}\boldsymbol{Q_\gamma}\boldsymbol{Q_\beta}]$ consists of the following parts:

$$\boldsymbol{Q_L} = \begin{pmatrix} 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & \sigma^2_{\eta_R,1} & 0 & \varsigma_{\eta_R,1,2} & \dots & 0 & \varsigma_{\eta_R,1,D} \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & \varsigma_{\eta_R,2,1} & 0 & \sigma^2_{\eta_R,2} & \dots & 0 & \varsigma_{\eta_R,2,D} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & \varsigma_{\eta_R,D,1} & 0 & \varsigma_{\eta_R,D,2} & \dots & 0 & \sigma^2_{\eta_R,D} \end{pmatrix}, \tag{2.A.3}$$

$$\boldsymbol{Q_\gamma} = Blockdiag[\boldsymbol{\Gamma_1}...\boldsymbol{\Gamma_D}],\ \boldsymbol{\Gamma_d} = \begin{pmatrix} \sigma^2_{\omega,d,1} & 0 & 0 \\ 0 & \sigma^2_{\omega^*,d,1} & 0 \\ 0 & 0 & \sigma^2_{\omega,d,2} \end{pmatrix}, d \,\epsilon\, \{1,..,D\},$$

$\boldsymbol{Q_\beta} = \boldsymbol{0_{[K\times K]}}$, since the regression coefficients are modelled as time-independent.

Table 2.A.1: Univariate models: hyperparameter estimates (in kilotons) and model diagnostics based on standardised innovations

Series	1	2	3	4	5	6	7	8	9	Total
$\sigma_{R,d}$	3 (1; 11)	41 (26; 64)	6 (2; 15)	2 (1; 4)	7 (4; 11)	83 (39; 174)	4 (2; 7)	26 (13; 50)	48 (25; 92)	201 (121; 332)
$\sigma_{\omega,d}$	-	15 (4; 59)	-	-	-	66 (33; 133)	-	26 (9; 71)	-	-
$\sigma_{\nu,t,d}$ *)	281 (243; 325); 436 (339; 562)	406 (347; 475); 1370 (1060; 1770)	244 (203; 294); 164 (126; 213); 387 (297; 504)	54 (46; 64); 140 (115; 170)	54 (47; 63); 357 (247; 514); 102 (68.9; 151)	1230 (989; 1530); 720 (536; 966); 1650 (1240; 2180)	155 (135; 178); 541 (417; 701)	553 (472; 647); 1880 (1450; 2430)	267 (219; 325); 1050 (774; 1430); 1790 (1370; 2330)	1st: 1690 (1290; 2220); 2nd, 4th: 2410 (1920; 3040); 3rd: 1310 (944; 1820); 5th: 3910 (2990; 5110)
N	0.970 [0.616]	0.634 [0.728]	0.596 [0.742]	9.974 [0.007]†	2.057 [0.358]	2.578 [0.276]	1.048 [0.592]	3.229 [0.199]	47.399 [0.000]†	8.292 [0.016]*
Skewness	-0.202	-0.161	0.151	0.451	0.221	0.303	0.203	0.153	2.076	0.614
Excess kurtosis	-0.114	-0.170	-0.036	1.379	0.323	-0.193	0.028	0.542	10.743	0.503
$\sigma_{St.in.}$	0.982	1.009	1.014	1.010	0.997	1.015	1.016	1.016	1.273	1.047
DW	1.731	1.940	1.639*	2.166	2.024	1.706	1.868	1.846	2.098	1.725
H	1.287	0.932	1.220	1.656	1.005	0.908	1.323	1.276	3.867†	0.987

* Significant at the 5% level. † Significant at the 1% level.

Asymmetric confidence intervals in brackets. The asymmetry arises from the fact that the hyperparameters are estimated on a log scale. The Fisher information matrix is applied for estimating the asymptotic standard errors of the log-transformed hyperparameters, whereafter the confidence interval bounds are transformed back to the normal scale.

N: Doornik-Hansen normality test (chi-square test); p-values in [ ];

$\sigma_{St.in.}$: standard deviation of standardized innovations;

DW: Durbin-Watson test statistic for serial correlation;

H: F-test for heteroscedasticity

*) Multiple rows in this field stand for the separate variance values in different sub-periods according to Table 2.4.1. The dates defining the sub-periods can be found in the same table.

Table 2.A.2: Nine-dimensional model: hyperparameter estimates (in kilotons) and model diagnostics based on standardised innovations

Series	1	2	3	4	5	6	7	8	9
Zero-restrictions on eigenvalues	-	-	-	Yes	-	-	Yes	Yes	Yes
$\sigma_{R,d}$	24 (12; 48)	75 (17; 247)	11 (3; 53)	8 (2; 27)	10 (3; 41)	152 (20; 549)	8 (5; 50)	75 (13; 273)	89 (23; 309)
$\sigma_{\omega,d}$	-	17 (6; 54)	4 (0; 125)	-	-	70 (36; 135)	-	26 (10; 70)	-
$\sigma_{\nu,t,d}$ *)	256 (221; 296); 430 (331; 558)	387 (330; 455); 1350 (1044; 1740)	236 (196; 285); 161 (124; 208); 382 (295; 493)	54 (46; 63); 138 (114; 168)	54 (46; 62); 348 (243; 500); 103 (70; 153)	1220 (998; 1500); 582 (435; 779); 1660 (1240; 2210)	152 (131; 176); 537 (405; 711)	; 538 (462; 627); 1830 (1420; 2360)	239 (199; 288); 1060 (784; 1430); 1720 (1330; 2230)
N	1.121 [0.571]	0.512 [0.774]	1.210 [0.546]	18.439 [0.000]†	0.660 [0.719]	3.162 [0.206]	1.461 [0.482]	1.685 [0.431]	58.133 [0.000]†
Skewness	-0.105	-0.060	0.223	0.736	0.117	0.338	0.248	0.231	2.379
Excess kurtosis	0.205	0.076	0.019	2.569	0.072	-0.161	-0.079	0.207	12.530
$\sigma_{St.in.}$	0.968	1.055	1.012	1.049	1.000	1.039	1.019	1.046	1.336
DW	2.078	1.952	1.746	2.129	2.051	1.798	1.932	1.955	2.242
H	1.289	0.997	1.262	1.161	1.091	0.903	1.258	1.268	3.481†

* Significant at the 5% level. † Significant at the 1% level.

Asymmetric confidence intervals in brackets. The asymmetry arises from the fact that the hyperparameters are estimated on a log scale. The Fisher information matrix is applied for estimating the asymptotic standard errors of the log-transformed hyperparameters, whereafter the confidence interval bounds are transformed back to the normal scale.

N: Doornik-Hansen normality test (chi-square test); p-values in [ ];

$\sigma_{St.in.}$: standard deviation of standardized innovations;

DW: Durbin-Watson test statistic for serial correlation;

H: F-test for heteroscedasticity

*) Multiple rows in this field stand for the separate variance values in different sub-periods according to Table 2.4.1. The dates defining the sub-periods can be found in the same table.

Chapter 3

Aggregated Series estimation in a Multivariate State Space Model: the DRTS case

Issues concerning modelling the aggregated series in general, as well as in the case of the DRTS, have been brought to this chapter because they deserve a more careful discussion. The major issue is whether or not an aggregated series of interest should be modelled jointly with underlying D domains. A D-dimensional model has already been presented in the previous chapter. It may seem useful to include the aggregated series into the model because the signal-to-noise ratio of such a series must be much better compared to the underlying domains. However, one may fairly argue that D- and $(D+1)$-dimensional models must have identical outcomes, since no additional information enters the model. This is indeed so when the Kalman filter operates under the true hyperparameter values. A small simulation in this chapter sheds light on differences in the Kalman filter performance under true and estimated hyperparameters. As soon as the true hyperparameter values are replaced by their ML estimates, differences emerge both between point- and variance estimates produced by the two models. The results of the simulation explain differences in estimation results obtained from the seemingly equivalent D- and $(D+1)$-dimensional DRTS models.[1]

[1]This chapter is based on the supplement to the paper Bollineni-Balabay et al. (2016a) published in *Journal of the Royal Statistical Society: Series A*. The authors thank Dr. Sabine Krieg and Rob Willems (Statistics Netherlands), as well as the unknown reviewers and Associate Editor for valuable comments on earlier versions of this paper.

3.1 Introduction

In surveys like the DRTS, where estimates are produced at an aggregated (national) and domain level, a question may arise as to whether all of these series have to be included into the model. If the aggregated series' design-based estimates are sufficiently precise, they can be included into the model and be treated as external benchmarks for the sum of the domain predictions in order to secure the model against misspecification (as in Pfeffermann and Tiller (2006), Pfeffermann et al. (2014)). However, if the aggregated series is not sufficiently precise, as in the case of the DRTS, it cannot be used as a benchmarking series and itself would benefit from signal extraction. The choice between modelling it in a univariate setting and deriving it from a multivariate model for D domains has already been discussed in Subsection 2.3.3.

It may seem useful to jointly model the domains and aggregated series in one $(D+1)$-dimensional setting with the restriction that the sums over the domain state variables and over disturbance terms are equal to the corresponding states and disturbances of the aggregated series. The state space representation of such a model is given in Subsection 3.2.1. In this way, the aggregated series with a better signal-to-noise ratio would enter the multivariate model, and the sum of the domain model estimates would be equal to the aggregated series model estimates at each point in time. However, as no additional information would enter the model in this case, the $(D+1)$-dimensional approach must be identical to the D-dimensional one. Subsection 3.2.2 presents a simulation that confirms that the D- and $(D+1)$-dimensional approaches produce identical outcomes, though only if the true hyperparameter values are used in the Kalman filter. Both models, though based on estimated hyperparameters are also applied to the DRTS series, and the outcomes are compared in Subsection 3.2.3. Section 3.3 briefly discusses different possible approaches to modelling the aggregated series, as well as restates the main result of the simulation performed in this chapter.

3.2 Multivariate model for D domains extended with an aggregated series

3.2.1 State space representation

This subsection presents a $(D+1)$-dimensional model applicable to the DRTS series. The components are described in Subsection 2.3.1. The input vector in this case is extended to $\hat{\boldsymbol{Y}}_{\boldsymbol{t}} = (\hat{Y}_{t,1}...\hat{Y}_{t,D}\hat{Y}_{t,Tot})'$, where $\hat{Y}_{t,Tot} = \sum_{d=1}^{D} \hat{Y}_{t,d}$ are the aggregated series design estimates. The trend and seasonal components of the state vector are extended similarly to the $\hat{\boldsymbol{Y}}_{\boldsymbol{t}}$-vector, whereas the vector with level interventions remains unchanged. Then the state vector $\boldsymbol{\alpha_t}$ has the following structure:

$$\begin{aligned}
&\boldsymbol{\alpha_t} = (\boldsymbol{\alpha_t^L}\ \boldsymbol{\alpha_t^\gamma}\ \boldsymbol{\alpha_t^\beta})', \\
&\boldsymbol{\alpha_t^L} = (L_{t,1}\ R_{t,1}...\ L_{t,D}\ R_{t,D}\ L_{t,Tot}\ R_{t,Tot}), \\
&\boldsymbol{\alpha_t^\gamma} = (\boldsymbol{1'_{D+1}} \otimes \boldsymbol{\gamma_{t,d}}),\ d \in \{1, ..., D+1\},
\end{aligned}$$

where

$$\begin{aligned}
&\boldsymbol{\gamma_{t,d}} = (\gamma_{t,1,d}\ \gamma^*_{t,1,d}\cdots\gamma_{t,(s/2-1),d}\ \gamma^*_{t,s/2-1),d}\ \gamma_{t,s/2,d}), \\
&\boldsymbol{\alpha^\beta} = (\beta_1...\beta_K).
\end{aligned}$$

$\boldsymbol{1'_p}$ is p-dimensional horizontal vector of ones, $\otimes$ is the Kronecker product which multiplies every element of the first matrix by the second matrix, and s is the number of seasons. The inclusion of an aggregated series which is exactly equal to the sum over the domain estimates implies that each state variable of the aggregated series is restricted to the sum of the corresponding state variables of all the domains. These restrictions are incorporated in the design matrix $\boldsymbol{T}$ of the transition equation $\boldsymbol{\alpha_{t+1}} = \boldsymbol{T\alpha_t} + \boldsymbol{\eta_{t+1}}$. The transition matrix is defined as $\boldsymbol{T} = Blockdiag[\boldsymbol{T_L T_\gamma T_\beta}]$ with

$$\begin{aligned}
\boldsymbol{T_L} &= \begin{pmatrix} \boldsymbol{I_{[D]}} \otimes \boldsymbol{T_{LR}} & \boldsymbol{0_{[2D\times 2]}} \\ \boldsymbol{1'_{[D]}} \otimes \boldsymbol{T_{LR}} & \boldsymbol{0_{[2\times 2]}} \end{pmatrix}, \\
\boldsymbol{T_\gamma} &= \begin{pmatrix} \boldsymbol{I_{[D]}} \otimes \boldsymbol{H} & \boldsymbol{0_{[(s-1)D\times(s-1)]}} \\ \boldsymbol{1'_{[D]}} \otimes \boldsymbol{H} & \boldsymbol{0_{[(s-1)\times(s-1)]}} \end{pmatrix}, \\
\boldsymbol{T_\beta} &= \boldsymbol{I_{[K]}},
\end{aligned}$$

where $\boldsymbol{T_{LR}} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, and $\boldsymbol{I_{[p]}}$ is a p-dimensional identity matrix, and $\boldsymbol{0_{[m \times n]}}$ denotes a $m \times n$ matrix. In this application, the number of seasons is $s = 4$, which implies three seasonal stochastic variables per domain and thus a three-dimensional design matrix $\boldsymbol{H}$ as in (2.A.2). The design matrix $\boldsymbol{Z_t}$ for the measurement equation $\hat{\boldsymbol{Y}}_{\boldsymbol{t}} = \boldsymbol{Z_t \alpha_t} + \boldsymbol{\nu_t}$ has the following form:

$$\boldsymbol{Z_t} = [\boldsymbol{I_{[D+1]}} \otimes (1 \ 0) \ \boldsymbol{I_{[D+1]}} \otimes (1 \ 0 \ 1) \ \boldsymbol{Z_{\beta,t}}],$$

where $\boldsymbol{Z_{\beta,t}}$ consists of K vertical vectors $\boldsymbol{x_{t,k}}$ extended with one element in order to include each of the level interventions contained in vector $\boldsymbol{\alpha_t^{\beta}}$ into the aggregated signal. As described in Subsection 2.4.1, five level interventions are modelled for the DRTS, which results in:

$$\boldsymbol{Z_{\beta,t}} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & \delta_{t,5} \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \delta_{t,1} & \delta_{t,2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \delta_{t,3} & \delta_{t,4} & 0 \\ \delta_{t,1} & \delta_{t,2} & \delta_{t,3} & \delta_{t,4} & \delta_{t,5} \end{pmatrix}.$$

This implies that the magnitude of the level interventions in the aggregated series is equal to the cumulative magnitude of those in the domains at each point in time.

Further, proper restrictions have to be imposed on the covariance matrix of state disturbances $\boldsymbol{Q} = E(\boldsymbol{\eta_t \eta_t'}) = Blockdiag[\boldsymbol{Q_l Q_\gamma Q_\beta}]$. Elements of $\boldsymbol{Q}$ are defined below. These restrictions imply that the covariance between the slope disturbances of the aggregated series and those of the d-th domain must be equal to the sum of covariances between the stochastic terms of the domain in question and the other series:

$$Cov(\eta_{R,t,d}, \eta_{R,t,Tot}) = \sum_{d'=1}^{D} Cov(\eta_{R,t,d}, \eta_{R,t,Tot}), \ d \in \{1, ..., D\}. \tag{3.2.1}$$

For the slope disturbance variance of the aggregated series, the following will hold:

$$\sigma^2_{R,D+1} = \sum_{d=1}^{D} \sigma^2_{R,d} + 2 \sum_{d'=1}^{D} \sum_{d=d'+1}^{D} \varsigma_{\eta R,d,d'}. \tag{3.2.2}$$

These restrictions translate into the following matrix:

$$\boldsymbol{Q_L} = \boldsymbol{G Q_R G'},$$

where $\boldsymbol{G} = \begin{pmatrix} \boldsymbol{I}_{[D]} \otimes \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \\ \boldsymbol{1}'_{[D]} \otimes \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \end{pmatrix}$,
and $\boldsymbol{Q_R}$ is a $[2D \times 2D]$ covariance matrix of the D domains' slope disturbance terms as in (2.A.3).

If covariances between the seasonal disturbances are restricted to zero, restrictions on the covariance matrix of harmonic disturbances are reduced to a much simpler form than those for the slope disturbances:

$$Cov(\omega_{t,d,j}, \omega_{t,D+1,j}) = Cov(\omega^*_{t,d,j}, \omega^*_{t,D+1,j}) = \sigma^2_{\omega,d},\ j \in \{1, ..., [s/2]\},\ d \in \{1, ..., D\},$$

$$\sigma^2_{\omega,D+1} = \sum_{d=1}^{D} \sigma^2_{\omega,d}.$$

These restrictions have the following state-space representation:

$$\boldsymbol{Q_\gamma} = \begin{pmatrix} \boldsymbol{\Gamma_1} & \boldsymbol{O} & \boldsymbol{O} & \dots & \boldsymbol{\Gamma_1} \\ \boldsymbol{O} & \boldsymbol{\Gamma_2} & \boldsymbol{O} & \dots & \boldsymbol{\Gamma_2} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{\Gamma_3} & \dots & \boldsymbol{\Gamma_3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Gamma_1} & \boldsymbol{\Gamma_2} & \boldsymbol{\Gamma_3} & \dots & \sum_{d=1}^{D} \boldsymbol{\Gamma_d} \end{pmatrix},$$

$$\boldsymbol{\Gamma_d} = Diag(\sigma^2_{\omega,d,1}\ \sigma^2_{\omega^*,d,1}\ \sigma^2_{\omega,d,2}),\ d \in \{1, ..., D\},$$

where $\sigma^2_{\omega,d,j}$, $\sigma^2_{\omega^*,d,j}$ are the disturbance variances corresponding to the j-th seasonal harmonic of domain d, and $\boldsymbol{O}$ is a 3×3 zero-matrix.

The level interventions are modelled as time-invariant, so their covariance matrix is a $K \times K$ zero-matrix: $\boldsymbol{Q_\beta = O_{[K \times K]}}$.

Assuming that the covariances between the domain survey errors are equal to zero, the restrictions on the covariance matrix $\boldsymbol{R_t} = E(\boldsymbol{\nu_t \nu_t'})$ of the measurement equation disturbances will have a simplified form:

$$\begin{aligned} &\varsigma_{\nu,t,d,Tot} = \sigma^2_{\nu,t,d},\ d \in \{1, ..., D\}, \\ &\sigma^2_{\nu,t,Tot} = \sum_{d=1}^{D} \sigma^2_{\nu,t,d}. \end{aligned} \tag{3.2.3}$$

Matrix $\boldsymbol{R_t}$ has the following state space form:

$$\boldsymbol{R_t} = \begin{pmatrix} \sigma^2_{\nu,t,1} & 0 & 0 & \dots & \sigma^2_{\nu,t,1} \\ 0 & \sigma^2_{\nu,t,2} & 0 & \dots & \sigma^2_{\nu,t,2} \\ 0 & 0 & \sigma^2_{\nu,t,3} & \dots & \sigma^2_{\nu,t,3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma^2_{\nu,t,1} & \sigma^2_{\nu,t,2} & \sigma^2_{\nu,t,3} & \dots & \sum_{d=1}^{D} \sigma^2_{\nu,t,d} \end{pmatrix},$$

where $\sigma^2_{\nu,t,d}$ stands for the measurement equation error term variance of domain d at time t, $d \in \{1, ..., D\}$.

3.2.2 Simulation

This simulation is done in order to check whether and how outcomes from a $(D+1)$-dimensional model with appropriate restrictions imposed differ from those obtained from a D-dimensional model. The model for this simulation is built similarly to the DRTS model, although in a slightly simplified way. It contains only three domains and their aggregate. Within this simulation, one set of series is generated for 200 time-points using the *SsfPack* function *SsfRecursion* (Koopman et al. (2008), Ch.4.1), with hyperparameters of a magnitude similar to that in the DRTS. The trend model is the same smooth-trend model, but the seasonal component is removed for simplicity. Domains 2 and 3 feature level breaks that also enter the aggregated series. The level break in domain 2 constitutes 20 percent of the trend level at the time point preceding the break. This level intervention lasts for only 50 time points, which makes it similar to a change in the population of interest in domain 9 of the DRTS. Domain 3 experiences a 70 percent level

increase that lasts until the end of the series.

Restrictions given in equations (3.2.1)-(3.2.3) obviously render the Kalman filter recursions singular. The issue here is, however, not the singularity (since the univariate approach to multivariate Kalman filtering, which is implemented in *SsfPack* package, does allow for singularity in the system matrices (see Koopman and Durbin (2000)), but rather the fact that no additional information enters the model. Therefore, it is expected that this approach is identical to a D-dimensional multivariate model, from which the aggregated series estimates can also be derived.

A comparison of the two models has been conducted twice, using two different sets of hyperparameter values. First, the Kalman filter in the D- and $(D+1)$-dimensional models was run under the hyperparameters that had been used to generate the series. Under these true hyperparameter values, the point- and variance-estimates of the state variables turned out to be exactly the same for the two models. The four generated series are presented in Fig. 3.2.1, along with their filtered signal point-estimates. The variance-estimates are not presented because they are equal to each other.

In the other comparison of the two models, ML estimates are used instead of the hyperparameter true values in the Kalman filter recursions. In most cases, evaluation of the likelihood function in *SsfPack* fails due to the singularity problem, so the set of the hyperparameter ML estimates for running the Kalman filter under the $(D+1)$-dimensional model would have to be taken from the D-dimensional model, which is also done in this simulation. Filtered signals of the two models based on the estimated hyperparameters are almost identical to those based on the true hyperparameter values. The signal standard errors of the D- and $(D+1)$-dimensional models are presented in Fig. 3.2.2. The figure shows that, as soon as the hyperparameters used by the Kalman filter deviate from their true values, the signal variances of the two models deviate from each other. In this simulation, the model is quite simple, and the pattern of the input series seems to be much more stable compared to the series in the DRTS, so the hyperparameters are estimated quite accurately (see Table 3.2.1 below). Therefore, the differences between the signal variances of the two models are minute (see Fig. 3.2.2). These differences are so small that they can only be seen if zoomed in (note the scale of the y-axis). For this reason, only the last twenty time points are presented here. The reader

should not be misled by the seemingly steep decline in the standard errors. This decrease is very subtle and is part of the wavy behaviour inherent in these standard errors.

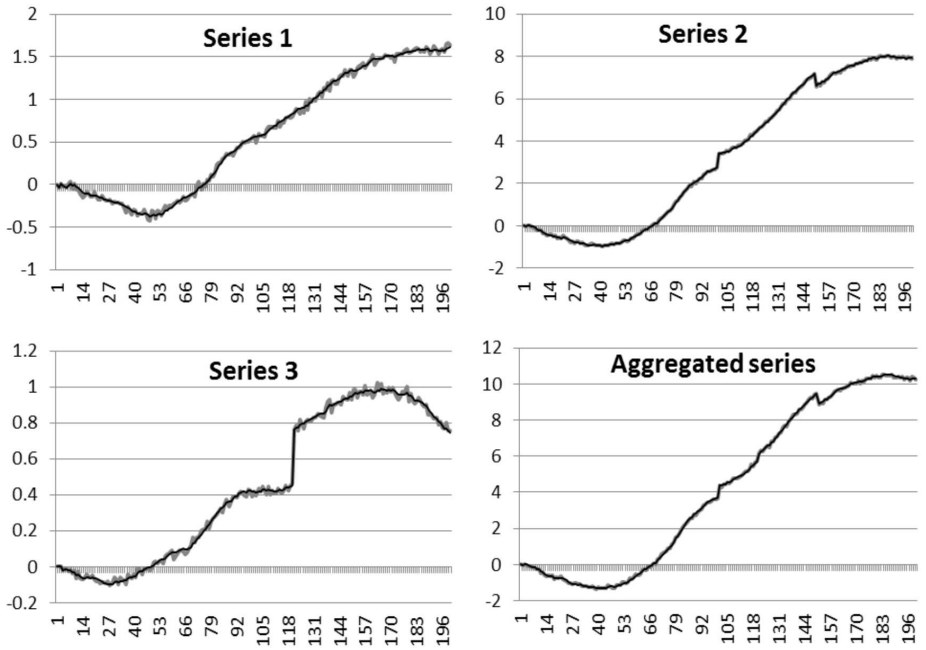


Figure 3.2.1: Simulated series (grey line) and their filtered signal estimates (black line) based on the true hyperparameter values in the Kalman filter recursions.

In order to see which model variance estimates are better, one can plot them together with the true variances, the latter being obtained from a model that uses the true hyperparameter values for the Kalman filtering (not presented here due to the scale problem). In the case of this simulation, there is no consistent pattern in which one or the other model over- or underestimates the true signal variances.

The simulation suggest that, at least in the case when the true hyperparameters are known, the D- and $(D+1)$-dimensional models produce identical outcomes. Therefore, the more parsimonious D-dimensional model is preferable.

Table 3.2.1: True hyperparameter values and their ML estimates.

	$ln(\sigma^2_{\eta_R,1})$	$ln(\sigma^2_{\eta_R,2})$	$ln(\sigma^2_{\eta_R,3})$
True	-12.00	-11.00	-15.00
Estimated	-12.30	-11.10	-15.30
	$\varsigma_{\eta_R,1,2}$	$\varsigma_{\eta_R,1,3}$	$\varsigma_{\eta_R,2,3}$
True	3.00	0.30	0.20
Estimated	3.24	0.29	0.24
	$ln(\sigma^2_{\nu,1})$	$ln(\sigma^2_{\nu,2})$	$ln(\sigma^2_{\nu,3})$
True	-7.00	-6.50	-8.00
Estimated	-6.82	-6.34	-8.01

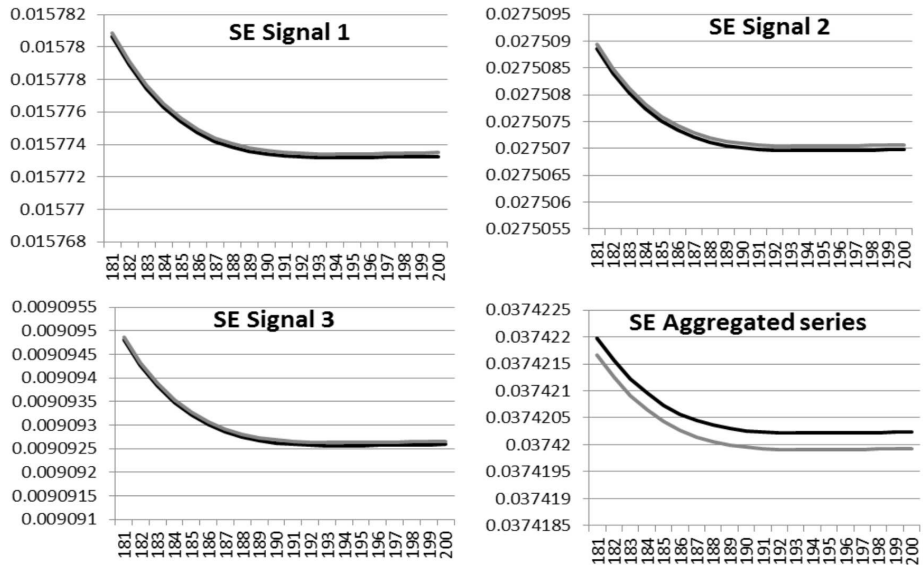


Figure 3.2.2: Standard error estimates of the filtered signals coming from the three-dimensional model for the domains (grey line) and the four-dimensional model that includes the aggregated series (black line); the hyperparameter ML estimates used instead of their true values.

3.2.3 Application to the DRTS

This subsection illustrates results from the $(D+1)$-dimensional model applied to the DRTS series. In particular, the signal variance estimates are compared to those from the D-dimensional model presented in the main body of this article.

The ML estimates for the hyperparameters of the ten-dimensional model are given in Table 3.2.2. Since no additional information enters the model when the aggregated series is included in the model, the hyperparameter ML estimates are expected to be equal or at least close to those in the nine-dimensional setting. This is indeed the case for the measurement equation error term variances and their asymmetric confidence intervals (compare Tables 2.A.2 and 3.2.2). An exception in this application is a nearly zero estimate of the mid-period measurement equation error term variance of domain 6 in the ten-dimensional setting. It turns out that the numerical optimization procedure in this model persistently ends up in a local optimum. Since the two models - with and without the aggregated series - are asymptotically equivalent, the hyperparameter ML estimates obtained from the nine-dimensional model (in Table 2.A.2) can be treated as given in the ten-dimensional setting. Whether one follows this approach or re-estimates the hyperparameters in the ten-dimensional setting, results in very similar point and variance estimates of the signals (except for the mid-periods in domains 6 and 9). The aim of this subsection, however, is to illustrate differences that emerge in the Kalman filter procedure carried out under the same set of hyperparameter estimates in both model settings. Therefore, the results presented below are obtained from the two models on the basis of hyperparameters estimated under the nine-dimensional model.

As has been mentioned in the previous subsection, the pattern of the DRTS series is much more erratic than that of the series simulated. Also, the model structure is more complex in that the model dimension is larger (with 9 domains), and in that a seasonal component enters the model. These problems could have been responsible for visible differences between the point-estimates of the two model settings (in domains 4, 5 and 9, see Fig. 3.2.3). The differences between the signal standard errors are also quite considerable, as shown in Fig. 3.2.4. Moreover, the signal standard errors from the ten-dimensional model are consistently lower than those from the nine-dimensional one. However, without true hyperparameter values at hand, it would be groundless to affirm that the ten-dimensional model estimates are better. In view of the simulation results, the more parsimonious D-dimensional model may be considered, from which the aggregated series estimates can be derived.

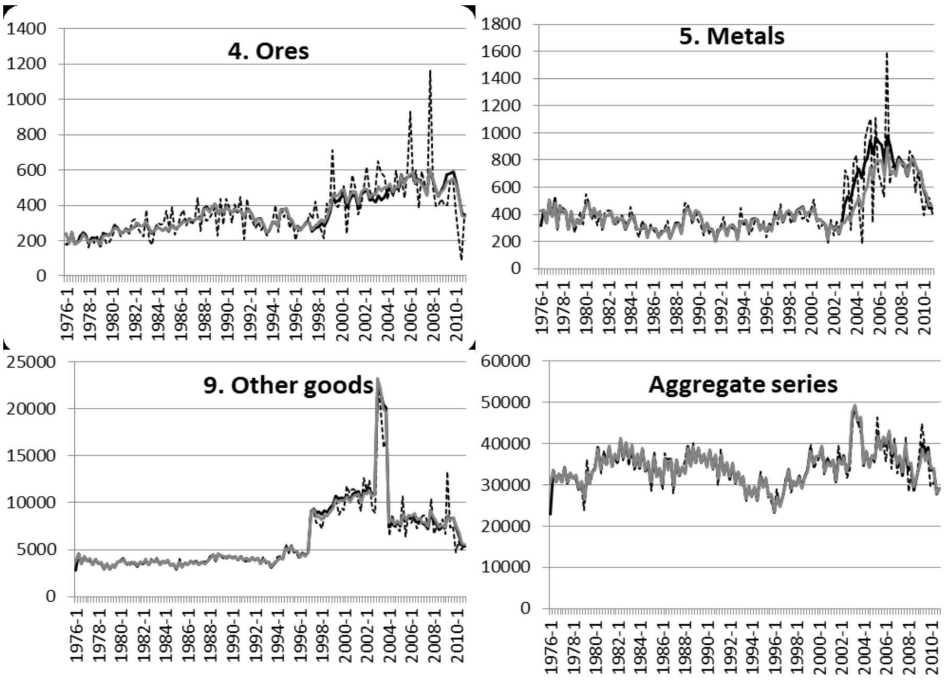


Figure 3.2.3: DRTS design estimates (dashed line) and filtered signal point-estimates from the nine-dimensional model (grey line) and from the ten-dimensional model that uses hyperparameter estimates from the nine-dimensional model (black line), in kilotons.

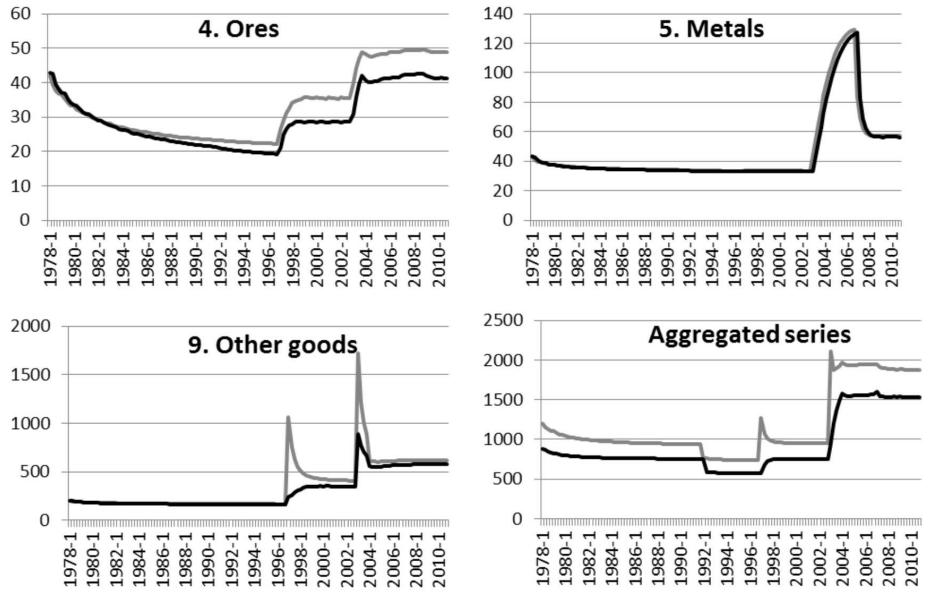


Figure 3.2.4: Standard errors of the filtered signal estimates obtained from the DRTS models: the nine-dimensional model (grey line), and ten-dimensional model that uses hyperparameter estimates from the nine-dimensional model (black line), in kilotons.

Table 3.2.2: Ten-dimensional model including the aggregated series: hyperparameter estimates (in kilotons) and model diagnostics based on standardised innovations

Series	1	2	3	4	5	6	7	8	9	Total***
Zero-restrictions on eigenvalues	-	-	-	Yes	-	-	Yes	Yes	Yes	-
$\sigma_{R,d}$	47 (23; 94)	168 (43; 507)	33 (3; 125)	15 (3; 49)	15 (3; 55)	403 (79; 1330)	22 (4; 117)	153 (35; 501)	155 (37; 502)	989 (856; 3212)
$\sigma_{\omega,d}$	-	21 (7; 61)	-	-	-	397 (306; 514)	-	34 (9; 124)	-	399 (306; 532)
$\sigma_{\nu,t,d}$ **)	258 (224; 298); 406 (312; 528)	373 (317; 439); 1340 (1050; 1710)	232 (196; 275); 157 (121; 202); 389 (300; 505)	56 (47; 65); 140 (114; 172)	66 (57; 77); 328 (231; 465); 99 (66; 148)	276 (65; 1170); 5 (0; $6.3{\cdot}10^{18*}$)); 1130 (740; 1720)	148 (128; 171); 502 (387; 651)	480 (403; 571); 1710 (1350; 2180)	224 (185; 271); 1350 (1020; 1780); 1510 (1160; 1980)	804 (642; 1471); 755 (638; $6.31{\cdot}10^{18*}$)); 735 (619; $6.31{\cdot}10^{18*}$)); 1525 (1183; $6.31{\cdot}10^{18*}$)); 2999 (2282; 3970); 2982 (2271; 3945)
N	0.494 [0.781]	1.333 [0.514]	1.874 [0.392]	16.242 [0.000]†	1.852 [0.396]	2.372 [0.306]	1.237 [0.539]	5.927 [0.052]	79.592 [0.000]†	12.351 [0.002]†
Skewness	-0.012	-0.156	0.284	0.851	0.051	0.308	0.188	0.454	3.026	0.787
Excess kurtosis	0.086	0.217	0.039	2.421	0.355	0.225	0.150	0.827	19.756	1.123
$\sigma_{St.in.}$	0.991	1.102	1.043	1.041	0.986	1.058	1.038	1.099	1.474	1.230
DW	2.171	2.113	1.818	2.158	1.714	2.180	2.048	2.267	2.277	2.401*
H	1.126	0.850	1.044	0.939	1.552	0.554	1.314	1.170	3.516†	0.853

* Significant at the 5% level. † Significant at the 1% level.

Asymmetric confidence intervals in brackets. The asymmetry arises from the fact that the hyperparameters are estimated on a log scale. The Fisher information matrix is applied for estimating the asymptotic standard errors of the log-transformed hyperparameters, whereafter the confidence interval bounds are transformed back to the normal scale.

N: Doornik-Hansen normality test (chi-square test); p-values in [ ];
$\sigma_{St.in.}$: standard deviation of standardized innovations;
DW: Durbin-Watson test statistic for serial correlation;
H: F-test for heteroscedasticity

*) The large upper margins are due to a failure to invert the Hessian when the corresponding hyperparameter is estimated close to zero.
**) Multiple rows in this field stand for the separate variance values in different sub-periods according to Table 2.4.1. The dates defining the sub-periods can be found in the same table. There are six different variance values for the national level series due to an overlap between sub-periods that stand for distinctive variances of different domains.
***) Hyperparameters of the aggregated series and their confidence intervals are not estimated but derived from domain hyperparameters.

Table 3.2.3: Ten-dimensional model including the aggregated series, where the Kalman filter uses the hyperparameter estimates from the nine-dimensional model: model diagnostics based on standardised innovations

Series	1	2	3	4	5	6	7	8	9	Total
N	1.697 [0.428]	0.480 [0.787]	1.591 [0.451]	16.202 [0.000]†	0.731 [0.694]	3.566 [0.168]	1.846 [0.397]	2.830 [0.243]	47.093 [0.000]†	13.044 [0.002]†
Skewness	-0.074	-0.095	0.261	0.690	0.118	0.379	0.279	0.351	2.064	0.791
Excess kurtosis	0.326	0.036	0.005	2.265	0.095	-0.029	-0.065	0.073	10.630	0.973
$\sigma_{St.in.}$	0.973	1.095	1.025	1.084	1.018	1.103	1.042	1.096	1.404	1.358
DW	2.115	1.972	1.769	2.049	2.044	1.948	1.832	1.917	2.186	2.109
H	1.254	0.951	1.203	1.142	1.094	0.896	1.474	1.227	3.136†	1.026

* Significant at the 5% level. † Significant at the 1% level.

N: Doornik-Hansen normality test (chi-square test); p-values in [ ];

$\sigma_{St.in.}$: standard deviation of standardized innovations;

DW: Durbin-Watson test statistic for serial correlation;

H: F-test for heteroscedasticity

3.3 Discussion

This chapter addresses the problem of modelling an aggregated series when its design-based estimates are not sufficiently precise, such that they should undergo signal extraction. The previous chapter has already discussed two possible options. One is developing a univariate state space model for the aggregated series. Another option is deriving the estimates at the aggregated level as a linear combination of the domain estimates from a D-dimensional multivariate model (or optionally from D univariate models, which is likely to result in larger variances, though). Which approach is more efficient in terms of signal variances, is an empirical question. As shown in Chapter 2, deriving the aggregated series model-based estimates from a $D-$dimensional model for domains in the case of the DRTS produces estimates with lower signal variances compared to when a univariate model is fitted to the aggregated series itself. If the consistency between estimates at the aggregated and domain level is important, only the $D-$dimensional approach is relevant.

The $(D+1)$-dimensional modelling approach has also been addressed. It may seem useful to jointly model the aggregated series with the underlying domains because it has considerably smaller sampling errors. With all the necessary restrictions on the state variables and the covariance matrices of their disturbance terms, such approach should be equivalent to a D-dimensional one. The small simulation with the known DGP shows that the outcomes from the D- and $(D+1)$-dimensional models are indeed identical, but only if the Kalman filter is equipped with the true hyperparameter values. Using hyperparameter estimates that deviate from the true hyperparameter values results in differences between the outcomes of the two models. The simulation has shown that the estimated signal variances may deviate from the true signal variances in both positive and negative directions. Therefore, there is no reason why one should resort to the more complex $(D+1)$-dimensional model setting.

One may be still tempted to include the aggregated series into the model and omit any other domain series in order to stay away from singularity which appears if the aggregated series is indeed the sum of the underlying domain series. Model-based estimates of the omitted series could then be derived from the estimates of the aggregated series and the rest of the domain series. This may seem like an appealing approach because a series at a higher aggregation level usually has a

better signal-to-noise ratio, compared to that of the underlying domains. However, it is better to abstain from such practice. First of all, it may be very hard to take the covariances between the domain and aggregated series into account when estimating time-varying hyperparameters in the observation equation. Ignoring this covariance structure will lead to overestimated signal variances of the series which has been omitted. Secondly, relatively minute under- or overestimates of the aggregated series' signal may translate into huge misestimates in the domain that has been omitted from the model. This is particularly true in the presence of level breaks. While some level breaks may not be identifiable in the aggregated series, they may be present in the underlying domains. Obtaining a decent level break estimate in a domain whose model-based estimates have been derived from the aggregated series' and remaining domains' model-based estimates would then be questionable. Apart from that, unmodelled level breaks in the aggregated series may result in an excessive flexibility of its stochastic trend, i.e. in an overestimated hyperparameter for the trend disturbance variance, which translates into a positive bias in the trend variances. Therefore, the reader should be reminded that survey modifications in the aggregated series can be better accounted for at a lower aggregation (domain) level.

Chapter 4

Accounting for Hyperparameter Uncertainty in State Space Models: the Case of the Dutch Labour Force Survey

Structural time series (STS) models, like the ones presented in Chapter 2, are known as a powerful technique for variance reduction in the framework of small area estimation (SAE) of repeatedly conducted surveys. Such models, however, contain unknown hyperparameters that have to be estimated before the Kalman filter can be launched to estimate state variables of the model. If the uncertainty around these hyperparameters is not taken into account, the variance estimates for state variables and signals obtained with the time series model become negatively biased, particularly in short time series. Statistics Netherlands currently uses a STS model for the Dutch Labour Force survey (DLFS) to obtain smaller variances for the unemployed labour force, compared to those of the design-based estimator. In order to account for the negative bias in the DLFS model variance estimates, several estimation approaches known in the literature are considered. Their performance ie compared in an extensive Monte-Carlo study, and the best approach is established. [1]

[1]This chapter is based on the paper Bollineni-Balabay et al. (2015) published by Statistics Netherlands as a discussion paper. The authors thank Dr. Harm Jan Boonstra and Dr. Frank Pijpers (Statistics Netherlands) for thorough reading and valuable comments on a draft of this paper.

4.1 Introduction

Monthly figures on the labour force produced by national statistical institutes (NSIs) are important economic indicators. These indicators are generally based on Labour Force Surveys (LFS). Most NSIs apply a rotating panel design in their LFS, but the sample size is hardly ever large enough even at the national level for producing sufficiently precise monthly figures about the labour force based on design-based estimators known from classical sampling theory (Särndal et al. (1992), Cochran (1977)). For the national level of the DLFS, the coefficient of variation between 2003 and 2010 is about 4 percent, which is quite good but not found to be sufficient by this survey's users. In such cases, statistical modelling can be used to improve the effective sample size of domains by borrowing information from preceding periods or other domains. Such techniques are often referred to as small area estimation (SAE), see Rao and Molina (2015). Repeatedly conducted surveys in particular have a potential for improvement within the framework of structural time series (STS) or multilevel time series models.

Commonly applied SAE procedures are based on multilevel models that are estimated with the empirical best linear unbiased prediction (EBLUP) or hierarchical Bayesian (HB) approaches. Similarly to STS models, such models usually contain unknown hyperparameters that have to be estimated, which translates into larger standard errors around the domain predictions. If this uncertainty (here and further in this paper referred to as the hyperparameter uncertainty) is not taken into account, the mean square error (MSE) estimates of the quantities of interest become negatively biased. Within the framework of the EBLUP and HB approaches, it is very common to take the hyperparameter uncertainty into account, see Rao (2003), Ch.6-7, 10.

STS models are not as widely used in SAE as multilevel models. The Kalman filter that is usually applied to analyse STS models ignores the hyperparameter uncertainty, and therefore produces negatively-biased MSE estimates. For this reason, the apparent gains from the STS technique in terms of reduced variance estimates have to be treated without undue optimism. Applications that give evidence for substantial advantages of STS models over the design-based approach treat estimated model hyperparameters as known, see, e.g., Krieg and Van den Brakel (2012), Van den Brakel and Krieg (2009a), Pfeffermann and Bleuer (1993),

Tiller (1992).

At Statistics Netherlands, a multivariate STS model is used to produce official monthly labour force figures. This model was originally proposed by Pfeffermann (1991) and is referred to as the DLFS (Dutch Labour Force survey) model in this paper. It uses sample information from preceding time periods and accounts for different features of the rotating panel design, such as the so-called rotation group bias (RGB) and autocorrelation in the survey errors. In this way, sufficiently precise monthly estimates for the unemployed labour force are obtained (see Van den Brakel and Krieg (2009a)). The paper, however, just like the other aforementioned studies, does not account for the hyperparameter uncertainty in the estimated MSEs.

This paper attempts to find the best MSE estimation method for the DLFS model. The literature offers several ways to account for the hyperparameter uncertainty in STS models: asymptotic approximation, bootstrapping and the full Bayesian approach. Among those approaches considered in this paper are the asymptotic approximation developed by Hamilton (1986), as well as parametric and non-parametric bootstrapping approaches developed by Pfeffermann and Tiller (2005) and Rodriguez and Ruiz (2012). These methods are applied to the DLFS model to see whether the hyperparameter uncertainty matters in terms of increased MSEs of the quantities of interest in this real life application.

This paper presents an extended Monte-Carlo simulation study, where the DLFS model acts as the data generation process. The contribution of the paper is threefold. First of all, it establishes the best out of the proposed approaches to the MSE estimation for the DLFS and offers a more realistic evaluation of the variance reduction obtained with the STS model compared to the design-based approach. Secondly, this Monte-Carlo study refutes the claim of Rodriguez and Ruiz (2012) about the superiority of their method over the bootstrap of Pfeffermann and Tiller (2005) in a more complex model. Finally, it is shown how the Monte Carlo simulation can be used to check for model overspecification.

The paper is structured as follows. Section 4.2 contains a description of the DLFS and the model currently used by Statistics Netherlands. Section 4.3 reviews the above-mentioned approaches to the MSE estimation. Some details on the simula-

tion and bootstrap setup specific for the DLFS are given in Section 4.4. Results of the simulation study are presented in Section 4.5. Section 4.6 concludes.

4.2 The LFS model

The DLFS is based on a five wave rotating panel design since October 1999. Every month, a sample of addresses is drawn according to a stratified two-stage sample design. All households residing on an address are included in the sample. Households are interviewed five times with a three-month gap between the interviews. Respondents of the first wave are interviewed by means of computer assisted personal interviewing. In the subsequent waves, computer assisted telephone interviewing applies with a shorter version of the questionnaire. For more details on the sample design, see Van den Brakel and Krieg (2009a). The production of monthly estimates of the total unemployed labor force starts with the general regression (GREG) estimator followed by a time series modelling approach. The GREG estimator can be found in (6.4.1) of Särndal et al. (1992), with the target variable $y_{i,k,t}$ being a binary response variable that is equal to 1 if the i-th person in the k-th household is unemployed, and zero otherwise. J auxiliary variables $x_{i,k,j}$ are (with the number of categories in brackets): ethnic origin (8), gender (2), age(21), civil status(2) and geographic region(44). The method of Lemaître and Dufour (1987) is used to make the weights for all persons within one household equal to each other.

Rotating panel designs are known to yield systematic differences between the estimates of their subsequent waves. This phenomenon is usually referred to as the rotation group bias (RGB), see, e.g., Bailar (1975), Kumar and Lee (1983), or Pfeffermann (1991). Common reasons behind the RGB are panel attrition, panel-effects, and differences in questionnaires and modes used in the subsequent waves. In the case of the DLFS, the first wave estimates are assumed to be most reliable, with the subsequent waves systematically underestimating the unemployed labour force numbers, which is reflected by a negative RGB in the level. Apart from the RGB, another problem with the LFS is small monthly sample sizes. With the net sample size of about 6500 persons in the first wave on average, the GREG estimates cannot produce sufficiently reliable unemployment figures on a monthly basis even at the national level. Both problems can be solved with the help of a STS model, which is currently used in the production of official LFS figures. Such

a model accounts for the RGB, as well as makes use of the information accumulated over time to produce point-estimates with smaller standard errors.

In a STS model, the series of design-based estimates is decomposed into several unobserved components, whereupon the so-called signal - a more reliable series of point-estimates - can be obtained. The signal is usually extracted with the Kalman filter. The filter removes a great part of the population and sampling noise from the GREG-estimates and produces point- and variance estimates for the signal and its unobserved components (trend, seasonal etc.). As a result, not only are these signal point-estimates less volatile (see Fig. 4.2.1), their standard errors are usually substantially lower than the design-based standard errors (in the case of the DLFS, 24 percent smaller). Apart from that, a STS model can extract the RGB pattern from the GREG series in case of a rotating panel design.

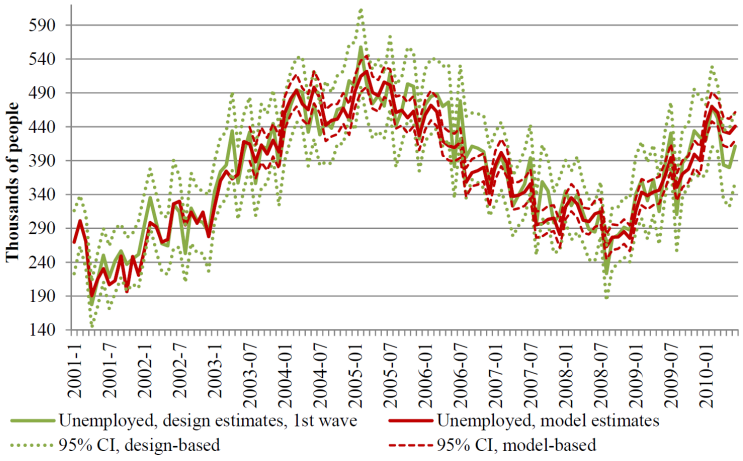


Figure 4.2.1: Numbers of unemployed in the Netherlands on a monthly basis: design- and model-based estimates, together with their confidence intervals

Let Y_t^{t-j} denote the GREG estimate of the total number of the unemployed labour force in month t. Five such estimates are obtained per month, each of them being respectively based on the sample that entered the survey in month $t-j$, $j = \{0, 3, 6, 9, 12\}$. Then, the five-dimensional vector $\boldsymbol{Y_t} = (Y_t^t\ Y_t^{t-3} Y_t^{t-6}\ Y_t^{t-9}\ Y_t^{t-12})'$ denotes the GREG estimates of the total number of the unemployed labour force for the five DLFS waves observed at time t. The term "wave" in this context means

a sample of households that enter the LFS panel at time t and leave the panel 12 months after five interviews. Based on Pfeffermann (1991), Van den Brakel and Krieg (2009a) developed the following model for the five-dimensional vector of GREG estimates:

$$\boldsymbol{Y_t} = \mathbf{1_5}\xi_t + \boldsymbol{\lambda_t} + \boldsymbol{e_t},$$

where $\mathbf{1_5}$ is a five-dimensional column vector of ones, $\boldsymbol{\lambda_t}$ is a vector containing the RGBs, and $\boldsymbol{e_t}$ is a vector with the survey errors that are correlated with their counterparts from previous waves (the structure will be presented later). It is assumed that the true population parameter is a scalar $\xi_t = L_t + \gamma_t + \varepsilon_t$, which is the sum of a stochastic trend L_t, a stochastic seasonal component γ_t, and an irregular component $\varepsilon_t \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$.

For the stochastic trend L_t, the so-called smooth-trend model is assumed:

$$L_t = L_{t-1} + R_{t-1},$$
$$R_t = R_{t-1} + \eta_{R,t},$$

where L_t and R_t represent the level and slope of the true population parameter, respectively, with the slope disturbance term being distributed as: $\eta_{R,t} \overset{iid}{\sim} N(0, \sigma_R^2)$.

In the case of monthly data, the seasonal component γ_t can be decomposed into six stochastic harmonics:

$$\gamma_t = \sum_{l=1}^{6} \gamma_{t,l},$$

where each of these six harmonics follows the process:

$$\gamma_{t,l} = cos(h_l)\gamma_{t-1,l} + sin(h_l)\gamma^*_{t-1,l} + \omega_{t,l},$$
$$\gamma^*_{t,l} = -sin(h_l)\gamma_{t-1,l} + cos(h_l)\gamma^*_{t-1,l} + \omega^*_{t,l},$$

with $h_l = \frac{\pi l}{6}$ being the l-th seasonal frequency, $l = \{1, ...6\}$. The zero-expectation stochastic terms $\omega_{t,l}$ and $\omega^*_{t,l}$ are assumed to be normally and independently distributed and to share the same variance within and across all the harmonics, such

that:

$$Cov(\omega_{t,l}, \omega_{t',l'}) = Cov(\omega^*_{t,l}, \omega^*_{t',l'}) = \begin{cases} \sigma^2_\omega & \text{if } l = l' \text{ and } t = t', \\ 0 & \text{if } l \neq l' \text{ or } t \neq t', \end{cases}$$

$$Cov(\omega_{t,l}, \omega^*_{t,l}) = 0 \text{ for all } l \text{ and } t.$$

Since the estimates of the first wave are assumed to be RGB-free (as motivated in Van den Brakel and Krieg (2009a)), the RGB vector for the five waves can be written in the following form: $\boldsymbol{\lambda_t} = (0\ \lambda_t^{t-3} \lambda_t^{t-6} \lambda_t^{t-9} \lambda_t^{t-12})'$. The RGB variables for the last four waves are time-dependent and are modelled as a random walk process:

$$\lambda_t^{t-j} = \lambda_{t-1}^{t-j} + \eta_{\lambda,t}^{t-j}, j = \{3, 6, 9, 12\}.$$

The RGB disturbances are not correlated across different waves and are normally distributed: $\eta_{\lambda,t}^{t-j} \overset{iid}{\sim} (0, \sigma^2_\lambda)$, with equal variances in all the four waves.

The last components in the model are the survey errors. Their variances are meant to be the GREG design-based variance estimates available from the micro-data:

$$\widehat{Var}(\hat{Y}_t^{t-j}) = \sum_{h=1}^{H} \frac{n_{h,t}}{n_{h,t} - 1} \left(\sum_{k=1}^{n_{h,t}} (w_{k,t} e_{k,t})^2 - \frac{1}{n_{h,t}} \left(\sum_{k=1}^{n_{h,t}} w_{k,t} e_{k,t} \right)^2 \right), \quad (4.2.1)$$

where the GREG residuals are $e_{k,t} = \sum_{i=1}^{n_{k,t}} (y_{i,k,t} - \boldsymbol{x'_{i,k,t}} \boldsymbol{\beta_t})$, with $y_{i,k,t}$ being a binary variable that is equal to 1 if the i-th person in the k-th household is unemployed, and zero otherwise. $x_{i,k,t}$ is a vector with auxiliary information on the i-th person in the k-th household at time t, vector $\boldsymbol{\beta_t}$ contains regression coefficients for period t, $w_{k,t}$ are (calibrated) weights for household k at time t, $n_{h,t}$ is the number of households in stratum h (with H being the total number of strata), and $n_{k,t}$ is the number of persons aged 15 or above in the k-th household.

In order to account for heterogeneity caused by differences in the survey design and sample sizes over time, the variance estimates from (4.2.1) have to be used as prior input into the model. To allow for flexibility in the process of numerical optimization, the survey errors can be modelled as $e_t^{t-j} = \tilde{e}_t^{t-j} \sqrt{\widehat{Var}(\hat{Y}_t^{t-j})}$, with $Var(\tilde{e}_t^{t-j}) \approx 1$. This formulation was proposed by Binder and Dick (1990). Since

the samples of different waves, starting from wave 2, are based on the same people as the preceding wave samples, there will be correlation between the survey errors. To account for this correlation, the survey error terms in Van den Brakel and Krieg (2009a) are modelled as an AR(1) process:

$$\tilde{e}_t^t = \nu_t^t;$$
$$\tilde{e}_t^{t-j} = \rho \tilde{e}_{t-3}^{t-j} + \nu_t^{t-j}, \;\; \nu_t^{t-j} \overset{iid}{\sim} N(0, \sigma^2_{\nu_t^{t-j}}), \; j = \{3, 6, 9, 12\}.$$

The survey error autoregressive parameter ρ is common to all the four waves and is estimated from the input data using a procedure developed by Pfeffermann et al. (1998). In this application $\hat{\rho} = 0.208$, see Van den Brakel and Krieg (2009b). This estimate is used as a prior input into the model, together with the design-based standard error estimates of the five waves. Since the variance of the product $\sqrt{\widehat{Var}(\hat{Y}_t^{t-j})}\tilde{e}_t^{t-j}$ is meant to be close to the variance estimate of the GREG estimator, $Var(\tilde{e}_t^{t-j}) \approx 1$. For the first wave, $Var(\tilde{e}_t^t) = \sigma^2_{\nu_t^t} \approx 1$, and for the four subsequent waves, with an AR(1) process assumed, $Var(\tilde{e}_t^{t-j}) \approx \sigma^2_{\nu_t^{t-j}}/(1-\rho^2) \approx 1, \quad j = \{3, 6, 9, 12\}$. Five different hyperparameters $\sigma^2_{\nu_t^{t-j}}$ are assigned to the five waves, and the estimated $Var(\tilde{e}_t^{t-j})$ have indeed turned out to be close to unity.

Linear structural time series models with unobserved components are usually analysed with the Kalman filter after putting them into a state space form:

$$\boldsymbol{\alpha_{t+1}} = \boldsymbol{T_{t+1}}\boldsymbol{\alpha_t} + \boldsymbol{\eta_t}, \tag{4.2.2}$$
$$\boldsymbol{Y_t} = \boldsymbol{Z_t}\boldsymbol{\alpha_t}. \tag{4.2.3}$$

Here, $\boldsymbol{T_t}$ and $\boldsymbol{Z_t}$ are known design matrices of the transition and measurement equations, respectively. The transition equation (4.2.2) describes how the state vector evolves over time, whereas the measurement equation (4.2.3) reflects the linear relationship between the observations and the state vector. The autoregressive structure of survey errors in the rotating panel design can be taken into account if the errors $\tilde{e}_t^{t-j}$ are modelled as state variables. Here, the population parameter error term ε_t is also modelled as a state variable. These disturbance terms thus disappear from the measurement equation (4.2.3) and move to the state vector $\boldsymbol{\alpha_t}$. The zero-expectation vector $\boldsymbol{\eta_t}$ contains identically and independently distributed disturbance terms. The design-based standard errors $\sqrt{Var(\hat{Y}_t^{t-j})}$ become time-varying elements of the matrix $\boldsymbol{Z_t}$ and are from now on denoted as

$z_t^{t-j}, j = \{0, 3, 6, 9, 12\}$. More details on the state-space form of the DLFS model are presented in Appendix 4.A.

Collecting the variables mentioned in this section produces the following state vector (with the survey error terms indexed differently for the state-space form):

$$\begin{aligned}
&\boldsymbol{\alpha_t} = (\boldsymbol{\alpha_t^{\xi}}\ \boldsymbol{\alpha_t^{\lambda}}\ \boldsymbol{\alpha_t^{e}})', \text{ where}\\
&\boldsymbol{\alpha_t^{\xi}} = (L_t\ R_t\ \gamma_{t,1}\ \gamma_{t,1}^*...\gamma_{t,5}\ \gamma_{t,5}^*\ \gamma_{t,6}\ \varepsilon_t),\\
&\boldsymbol{\alpha_t^{\lambda}} = (\lambda_t^{t-3}\lambda_t^{t-6}\lambda_t^{t-9}\lambda_t^{t-12}),\\
&\boldsymbol{\alpha_t^{e}} = (\tilde{e}_t^t\ \tilde{e}_t^{t-3}\ \tilde{e}_t^{t-6}\ \tilde{e}_t^{t-9}\ \tilde{e}_t^{t-12}\ \tilde{e}_{t-2}^{t-2}\ \tilde{e}_{t-2}^{t-5}\ \tilde{e}_{t-2}^{t-8}\ \tilde{e}_{t-2}^{t-11}\ \tilde{e}_{t-1}^{t-1}\ \tilde{e}_{t-1}^{t-4}\ \tilde{e}_{t-1}^{t-7}\ \tilde{e}_{t-1}^{t-10}).
\end{aligned}$$

All the state variables are initialised with a diffuse prior, except for the five survey error components $\tilde{e}_t^{t-j}, j = \{0, 3, 6, 9, 12\}$ and the population white noise ε_t. These stationary state variables are initialised with zeros and with the initial variances taken equal to unity.

The disturbance variances, together with the autocorrelation parameter ρ, are collected in a hyperparameter vector called $\boldsymbol{\theta} = (\sigma_R^2\ \sigma_\omega^2\ \sigma_\varepsilon^2\ \sigma_\lambda^2\ \sigma_{\nu^1}^2\ \sigma_{\nu^2}^2\ \sigma_{\nu^3}^2\ \sigma_{\nu^4}^2\ \sigma_{\nu^5}^2\ \rho)$, where the superscripts {1,...5} stand for the numbers of the waves (note that the variance of the last eight state variables is zero, see Appendix 4.A). Hyperparameter ρ is going to be estimated as in Pfeffermann et al. (1998) from the input data, whereafter the disturbance variance hyperparameters are estimated by the quasi-maximum likelihood method, treating $\hat{\rho}$-estimates as given. The vector with the hyperparameter estimates is denoted as $\hat{\boldsymbol{\theta}}$.

In the middle of 2010, the DLFS was subject to a survey transition that resulted in substantial discontinuities, which required an extension of the model. Therefore, the present study covers the time span from January 2001 until June 2010, in order to avoid unnecessary model complications that are not particularly relevant for this MSE estimation Monte-Carlo study.

Numerical analysis of this paper is conducted with OxMetrics 5 (Doornik (2007)) in combination with *SsfPack 3.0* package (Koopman et al. (2008)).

4.3 Review of MSE Estimation Approaches

State variables in structural time series models are usually extracted by the Kalman filter according to the following recursions:

$$\hat{\boldsymbol{\alpha}}_{t|t} = \hat{\boldsymbol{T}}_t \hat{\boldsymbol{\alpha}}_{t-1|t-1} + \boldsymbol{K}_t \boldsymbol{\epsilon}_t,$$
$$\boldsymbol{P}_{t|t} = \boldsymbol{P}_{t|t-1} - \boldsymbol{P}_{t|t-1} \boldsymbol{Z}_t' \boldsymbol{K}_t',$$
$$\boldsymbol{P}_{t|t-1} = \hat{\boldsymbol{T}}_t \boldsymbol{P}_{t-1|t-1} \hat{\boldsymbol{T}}_t' + \hat{\boldsymbol{Q}}_t,$$

where $'$ stands for a transpose, $\hat{\boldsymbol{\alpha}}_{t|t}$ and $\boldsymbol{P}_{t|t} = \boldsymbol{E}_t[(\hat{\boldsymbol{\alpha}}_{t|t} - \boldsymbol{\alpha}_t)(\hat{\boldsymbol{\alpha}}_{t|t} - \boldsymbol{\alpha}_t)']$ denote the conditional mean of the state vector and its MSE, respectively, extracted by the Kalman filter based on information available up to and including time t. This kind of estimates are usually referred to as *filtered* estimates. Matrix $\boldsymbol{P}_{t|t-1}$ is the predicted state covariance matrix:
$\boldsymbol{P}_{t|t-1} = \boldsymbol{E}_t[(\hat{\boldsymbol{\alpha}}_{t|t-1} - \boldsymbol{\alpha}_t)(\hat{\boldsymbol{\alpha}}_{t|t-1} - \boldsymbol{\alpha}_t)']$,
where *predicted* implies that the estimates for period t are based on information up to and including time $t-1$. $\boldsymbol{K}_t$ is the so-called Kalman gain: $\boldsymbol{K}_t = \boldsymbol{P}_{t|t-1} \boldsymbol{Z}_t' \boldsymbol{F}_t^{-1}$, $\boldsymbol{\epsilon}_t = \boldsymbol{Y}_t - \boldsymbol{Z}_t \hat{\boldsymbol{\alpha}}_{t|t-1}$ are innovations and $\boldsymbol{F}_t$ is the innovation covariance matrix: $\boldsymbol{F}_t = \boldsymbol{Z}_t \boldsymbol{P}_{t|t-1} \boldsymbol{Z}_t'$ (note that the covariance matrix of the measurement equation error terms is absent here because those error terms have been placed in the state vector). Here, the covariance matrix $\boldsymbol{Q}$ and transition matrix $\boldsymbol{T}$ are time-invariant, the former being a diagonal matrix with the state disturbance variance hyperparameters, and the latter containing the autoregressive parameter ρ.

The MSE extracted by the Kalman filter conditionally on the information up to and including time t is:

$$\boldsymbol{P}_{t|t} = \boldsymbol{E}_t[(\hat{\boldsymbol{\alpha}}_{t|t}(\boldsymbol{\theta}) - \boldsymbol{\alpha}_t)(\hat{\boldsymbol{\alpha}}_{t|t}(\boldsymbol{\theta}) - \boldsymbol{\alpha}_t)'], \tag{4.3.1}$$

where the expectation is taken with respect to the distribution of the state vector at time t, provided this expectation exists. In practice, however, the true hyperparameter vector is replaced by its estimates $\hat{\boldsymbol{\theta}}$ in the Kalman filter recursions. Then, the MSE in (4.3.1) is no longer the true MSE and is called "naive" as it does not incorporate the uncertainty around the $\hat{\boldsymbol{\theta}}$-estimates. The true MSE then becomes:

$$\boldsymbol{MSE}_{t|t} = \boldsymbol{E}_t[(\hat{\boldsymbol{\alpha}}_{t|t}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\alpha}_t)(\hat{\boldsymbol{\alpha}}_{t|t}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\alpha}_t)'],$$

which is larger than the MSE in (4.3.1) and can be decomposed as the sum of the filter uncertainty and parameter uncertainty:

$$\begin{aligned} \boldsymbol{MSE}_{t|t} =& \boldsymbol{E}_t[(\hat{\boldsymbol{\alpha}}_{t|t}(\boldsymbol{\theta}) - \boldsymbol{\alpha}_t)(\hat{\boldsymbol{\alpha}}_{t|t}(\boldsymbol{\theta}) - \boldsymbol{\alpha}_t)'] + \\ & \boldsymbol{E}_t[(\hat{\boldsymbol{\alpha}}_{t|t}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\alpha}}_{t|t}(\boldsymbol{\theta}))(\hat{\boldsymbol{\alpha}}_{t|t}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\alpha}}_{t|t}(\boldsymbol{\theta}))']. \end{aligned} \tag{4.3.2}$$

The first term, the filter uncertainty, is estimated by the naive MSE-estimates $\boldsymbol{P}_{t|t}$ delivered by the Kalman filter. Estimation of the second term, the parameter uncertainty, requires some additional effort. The literature on MSE estimation proposes two main approaches: asymptotic approximation and bootstrapping. Bootstrapping can be performed in a parametric or non-parametric way. A few remarks have to be mentioned about these methods in the context of STS models and of the DLFS model specifically.

If a STS model contains non-stationary components, as is the case with the DLFS model, a special procedure is required for bootstrap samples to be drawn conditionally on the given dataset $\{\boldsymbol{Y_1}, ..., \boldsymbol{Y_T}\}$. The simulation smoother algorithm developed by Durbin and Koopman (2002) can be exploited to generate conditional draws for the trend, seasonal and the RGB state variables. At the first step of this algorithm, the state variables and the data series are generated unconditionally on the original data set, either parametrically or non-parametrically. The generated series, say $\{\boldsymbol{Y_1^{b,\dagger}}, ..., \boldsymbol{Y_T^{b,\dagger}}\}$, will surely diverge from the dataset they have been bootstrapped from. They must therefore undergo some kind of correction. The state variables are corrected by the magnitude of the smoothed mean of their counterparts extracted from a "correction" dataset, which constitutes differences between the original data $\{\boldsymbol{Y_1}, ..., \boldsymbol{Y_T}\}$ and the unconditional bootstrap dataset $\{\boldsymbol{Y_1^{b,\dagger}}, ..., \boldsymbol{Y_T^{b,\dagger}}\}$, as described in Koopman et al. (2008), Ch.8.4.2. The survey errors generated as described in either parametric or non-parametric unconditional state recursion, do not need any adjustments as they constitute (autocorrelated) noise.

For the parametric bootstrap, the necessary disturbances for state recursions (4.2.2) and (4.2.3) are drawn from their joint conditional multivariate normal density $\boldsymbol{\eta_t} \overset{iid}{\sim} MN(\boldsymbol{0}, \boldsymbol{\Omega})$. Non-parametric bootstrap has an advantage of not depending on any particular assumption about this joint distribution. In this case, a bootstrap sample of standardised innovations $\{\boldsymbol{\epsilon_{r+1}^{b,St}}, ..., \boldsymbol{\epsilon_T^{b,St}}\}$ is obtained by sam-

pling with replacement from $\{\boldsymbol{\epsilon}_{r+1}^{St}(\hat{\boldsymbol{\theta}}), ..., \boldsymbol{\epsilon}_{T}^{St}(\hat{\boldsymbol{\theta}})\}$ where $\boldsymbol{\epsilon}_{t}^{St} = \boldsymbol{F}_{t}^{-1/2}(\hat{\boldsymbol{\theta}})\boldsymbol{\epsilon}_{t}(\hat{\boldsymbol{\theta}})$ are standardized innovations from the Kalman filter recursions based on the original ML estimates of the hyperparameters. For the DLFS model, a burn-in period r of 30 time-points is chosen. This choice is motivated by the number of state variables that are initialised with a diffuse prior, which is 25; additional 5 time points are skipped to have a burn-in period of two and a half years, and also in order to be on a safe-side. A bootstrap observation set $\{\boldsymbol{Y}_{1}^{b}, ..., \boldsymbol{Y}_{T}^{b}\}$ is then constructed by running the so-called innovation form of the Kalman filter:

$$\hat{\boldsymbol{\alpha}}_{t|t}^{b} = \hat{\boldsymbol{\alpha}}_{t|t-1}^{b} + \boldsymbol{K}_{t}(\hat{\boldsymbol{\theta}})\boldsymbol{F}_{t}^{1/2}(\hat{\boldsymbol{\theta}})\boldsymbol{\epsilon}_{t}^{b,St},$$
$$\hat{\boldsymbol{Y}}_{t}^{b} = \boldsymbol{Z}_{t}\hat{\boldsymbol{\alpha}}_{t|t-1}^{b} + \boldsymbol{F}_{t}^{1/2}(\hat{\boldsymbol{\theta}})\boldsymbol{\epsilon}_{t}^{b,St}, t = d+1, ..., T.$$

Note, that the univariate version of a multivariate Kalman filter, suggested by Koopman and Durbin (2000), is computationally more efficient. This version is implemented in *SsfPack* package that is used for this work. In this case, observations $Y_{t,d}$ in $\boldsymbol{Y_t}$ are reexpressed in a way that they can be treated as univariate time series. Then, for each of these series, the Kalman gain $\boldsymbol{K_{t,d}}$ is a vector, whereas $F_{t,d}$ is a scalar.

The following sections contain a brief presentation of the asymptotic approach, as well as the recent Rodriguez and Ruiz (2012) bootstrap approaches (hereafter referred to as the RR bootstrap) and Pfeffermann and Tiller (2005) (hereafter the PT bootstrap) bootstrap approaches.

4.3.1 Rodriguez and Ruiz Bootstrapping Approach

Rodriguez and Ruiz (2012) developed their bootstrap method for MSE estimation conditional on the data. Bootstrapping can be done both parametrically and non-parametrically, following the steps below:

1. Estimate the model and obtain all the hyperparameter estimates $\hat{\boldsymbol{\theta}}$.

2. Generate a bootstrap sample $\{\boldsymbol{Y}_{1}^{b}, ..., \boldsymbol{Y}_{T}^{b}\}$ using $\hat{\boldsymbol{\theta}}$, either parametrically or non-parametrically, as described before in paragraph 5 of the introduction to this section. If the model is non-stationary, the overall pattern of the bootstrap sample has to be brought in accordance with the pattern of the

original sample with the help of the simulation smoother, as described in the fourth paragraph of the introduction to Section 4.3.

3. The bootstrap dataset $\{\boldsymbol{Y_1^b}, ..., \boldsymbol{Y_T^b}\}$ is used to obtain both the survey error autocorrelation parameter $\hat{\rho}^b$ estimates and bootstrap ML estimates $\hat{\boldsymbol{\theta}}^{\boldsymbol{b}}_{\boldsymbol{ML}}$. Thereafter, the Kalman filter is launched using the original series $\{\boldsymbol{Y_1}, ..., \boldsymbol{Y_T}\}$ and the newly-estimated $\hat{\boldsymbol{\theta}}^{\boldsymbol{b}}$, which produces $\hat{\boldsymbol{\alpha}}_{\boldsymbol{t|t}}(\hat{\boldsymbol{\theta}}^{\boldsymbol{b}})$ and $\boldsymbol{P}_{\boldsymbol{t|t}}(\hat{\boldsymbol{\theta}}^{\boldsymbol{b}})$.

4. Then, steps 2-3 are repeated B times.

5. Having obtained B bootstrap replicates, the MSE can be estimated in the following way:

$$\widehat{\boldsymbol{MSE}}^{\boldsymbol{RR}}_{\boldsymbol{t|t}} = \frac{\boldsymbol{1}}{\boldsymbol{B}} \sum_{\boldsymbol{b=1}}^{\boldsymbol{B}} \boldsymbol{P}_{\boldsymbol{t|t}}(\hat{\boldsymbol{\theta}}^{\boldsymbol{b}}) + \frac{\boldsymbol{1}}{\boldsymbol{B}} \sum_{\boldsymbol{b=1}}^{\boldsymbol{B}} [\hat{\boldsymbol{\alpha}}_{\boldsymbol{t|t}}(\hat{\boldsymbol{\theta}}^{\boldsymbol{b}}) - \bar{\boldsymbol{\alpha}}_{\boldsymbol{t|t}}][\hat{\boldsymbol{\alpha}}_{\boldsymbol{t|t}}(\hat{\boldsymbol{\theta}}^{\boldsymbol{b}}) - \bar{\boldsymbol{\alpha}}_{\boldsymbol{t|t}}]', \tag{4.3.3}$$

where $\bar{\boldsymbol{\alpha}}_{\boldsymbol{t|t}} = \frac{\boldsymbol{1}}{\boldsymbol{B}} \sum_{\boldsymbol{b=1}}^{\boldsymbol{B}} \hat{\boldsymbol{\alpha}}_{\boldsymbol{t|t}}(\hat{\boldsymbol{\theta}}^{\boldsymbol{b}})$.

Equation (4.3.3) is applied both for the parametric and non-parametric bootstrap MSE-estimators (denoted further as MSE^{RR1} and MSE^{RR2}, respectively).

4.3.2 Pfeffermann and Tiller Bootstrapping Approach

The bootstrap developed by Pfeffermann and Tiller (2005) differs from the one described in the previous subsection in that expectations of the squared error loss elements in (4.3.2) are taken unconditionally on the data, whereas the Rodriguez and Ruiz (2012) approach conditions on the original dataset $\{\boldsymbol{Y_1}, ..., \boldsymbol{Y_T}\}$. However, application of the simulation smoother that makes bootstrap series conditional on the data (as described the beginning of this section) justifies the comparison between the bootstrap methods of Rodriguez and Ruiz (2012) and Pfeffermann and Tiller (2005). Unlike Rodriguez and Ruiz (2012), Pfeffermann and Tiller (2005) drop the terms that are of order $O(1/T^2)$ - a property that is theoretically proven in Pfeffermann and Tiller (2005), Appendix C. Using results in Hall and Martin (1988), they show that the true MSE, being

$$\boldsymbol{MSE}_{\boldsymbol{t|t}} = \boldsymbol{P}_{\boldsymbol{t|t}}(\boldsymbol{\theta}) + \boldsymbol{E}[(\hat{\boldsymbol{\alpha}}_{\boldsymbol{t|t}}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\alpha}}_{\boldsymbol{t|t}}(\boldsymbol{\theta}))(\hat{\boldsymbol{\alpha}}_{\boldsymbol{t|t}}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\alpha}}_{\boldsymbol{t|t}}(\boldsymbol{\theta}))'], \tag{4.3.4}$$

can be estimated by its bootstrap analogues as follows:

$$\boldsymbol{P_{t|t}(\theta) = 2P_{t|t}(\hat{\theta}) - \frac{1}{B}\sum_{b=1}^{B} P_{t|t}(\hat{\theta}^b) + O\left(1/T^2\right)}, \tag{4.3.5}$$

$$\begin{aligned} &\boldsymbol{E}[(\hat{\alpha}_{t|t}(\hat{\theta}) - \hat{\alpha}_{t|t}(\theta))(\hat{\alpha}_{t|t}(\hat{\theta}) - \hat{\alpha}_{t|t}(\theta))'] = \\ &\frac{1}{B}\sum_{b=1}^{B}[\hat{\alpha}^b_{t|t}(\hat{\theta}^b) - \hat{\alpha}^b_{t|t}(\hat{\theta})][\hat{\alpha}^b_{t|t}(\hat{\theta}^b) - \hat{\alpha}^b_{t|t}(\hat{\theta})]' + O\left(1/T^2\right). \end{aligned} \tag{4.3.6}$$

Equations (4.3.5) and (4.3.6) correspond to the first and the second terms of equation (4.3.2), respectively. The resulting MSE-estimator below has a bias of order $O\left(1/T^2\right)$:

$$\begin{aligned} \widehat{\boldsymbol{MSE}}^{\boldsymbol{PT}}_{t|t} =& 2\boldsymbol{P}_{t|t}(\hat{\theta}) - \frac{1}{B}\sum_{b=1}^{B} \boldsymbol{P}_{t|t}(\hat{\theta}^b) + \\ &+ \frac{1}{B}\sum_{b=1}^{B}[\hat{\alpha}^b_{t|t}(\hat{\theta}^b) - \hat{\alpha}^b_{t|t}(\hat{\theta})][\hat{\alpha}^b_{t|t}(\hat{\theta}^b) - \hat{\alpha}^b_{t|t}(\hat{\theta})]'. \end{aligned} \tag{4.3.7}$$

Equation (4.3.7) is applied both for the parametric and non-parametric bootstrap MSE-estimators (denoted further as MSE^{PT1} and MSE^{PT2}, respectively). MSE-calculation in (4.3.7) requires two Kalman filter runs for every bootstrap series. In the first run, $\hat{\boldsymbol{\alpha}}^{\boldsymbol{b}}_{\boldsymbol{t|t}}(\hat{\boldsymbol{\theta}}^{\boldsymbol{b}})$ is estimated from the bootstrap data set $\{\boldsymbol{Y}^{\boldsymbol{b}}_{\boldsymbol{1}}, ..., \boldsymbol{Y}^{\boldsymbol{b}}_{\boldsymbol{T}}\}$ and the bootstrap parameters $\hat{\boldsymbol{\theta}}^{\boldsymbol{b}}$. In this run, $\boldsymbol{P}_{\boldsymbol{t|t}}(\hat{\boldsymbol{\theta}}^{\boldsymbol{b}})$ can also be obtained based on $\hat{\boldsymbol{\theta}}^{\boldsymbol{b}}$ and irrespective of the data. The second Kalman filter run is needed to produce the state estimates $\hat{\boldsymbol{\alpha}}^{\boldsymbol{b}}_{\boldsymbol{t|t}}(\hat{\boldsymbol{\theta}})$ based on data set $\{\boldsymbol{Y}^{\boldsymbol{b}}_{\boldsymbol{1}}, ..., \boldsymbol{Y}^{\boldsymbol{b}}_{\boldsymbol{T}}\}$ and $\hat{\boldsymbol{\theta}}$-estimates that were obtained from the original dataset. The bootstrap procedure is summarized below:

1. Estimate the model using the original dataset and obtain the hyperparameter vector estimates $\hat{\boldsymbol{\theta}}$. Apart from that, save the "naive" MSE estimates $\boldsymbol{P}_{\boldsymbol{t|t}}(\hat{\boldsymbol{\theta}})$ for future use in (4.3.7).

2. Using either the parametric or non-parametric method, generate a bootstrap sample $\{\boldsymbol{Y}^{\boldsymbol{b}}_{\boldsymbol{1}}, ..., \boldsymbol{Y}^{\boldsymbol{b}}_{\boldsymbol{T}}\}$, conditional on the observed sample if the model is non-stationary, as described in paragraphs 4 and 5 of the introduction to Section 4.3.

3. Estimate bootstrap hyperparameter estimates $\hat{\boldsymbol{\theta}}^{\boldsymbol{b}}$ from the newly generated bootstrap dataset. Run the Kalman filter once to get $\hat{\boldsymbol{\alpha}}^{\boldsymbol{b}}_{\boldsymbol{t|t}}(\hat{\boldsymbol{\theta}}^{\boldsymbol{b}})$ and $\boldsymbol{P}_{\boldsymbol{t|t}}(\hat{\boldsymbol{\theta}}^{\boldsymbol{b}})$, and another time to obtain $\hat{\boldsymbol{\alpha}}^{\boldsymbol{b}}_{\boldsymbol{t|t}}(\hat{\boldsymbol{\theta}})$, as described under (4.3.7).

4. Repeat steps 2-3 B times.

5. Get the MSE estimate using (4.3.7).

Pfeffermann and Tiller (2005) note that, in the case of the parametric bootstrap, the second Kalman filter run can be avoided because the true state vector is generated (and thus known) for every bootstrap series. Thus, the state estimates $\hat{\boldsymbol{\alpha}}^{\boldsymbol{b}}_{\boldsymbol{t|t}}(\hat{\boldsymbol{\theta}})$ in (4.3.7) can be replaced by the true vector $\boldsymbol{\alpha}^{\boldsymbol{b}}_{\boldsymbol{t}}$ to obtain the following MSE estimator:

$$\widehat{\boldsymbol{MSE}}^{\boldsymbol{PT1}}_{\boldsymbol{t|t}} = \boldsymbol{P}_{\boldsymbol{t|t}}(\hat{\boldsymbol{\theta}}) - \frac{\boldsymbol{1}}{\boldsymbol{B}}\sum_{\boldsymbol{b=1}}^{\boldsymbol{B}} \boldsymbol{P}_{\boldsymbol{t|t}}(\hat{\boldsymbol{\theta}}^{\boldsymbol{b}}) + \frac{\boldsymbol{1}}{\boldsymbol{B}}\sum_{\boldsymbol{b=1}}^{\boldsymbol{B}} [\hat{\boldsymbol{\alpha}}^{\boldsymbol{b}}_{\boldsymbol{t|t}}(\hat{\boldsymbol{\theta}}^{\boldsymbol{b}}) - \boldsymbol{\alpha}^{\boldsymbol{b}}_{\boldsymbol{t}}][\hat{\boldsymbol{\alpha}}^{\boldsymbol{b}}_{\boldsymbol{t|t}}(\hat{\boldsymbol{\theta}}^{\boldsymbol{b}}) - \boldsymbol{\alpha}^{\boldsymbol{b}}_{\boldsymbol{t}}]'. \tag{4.3.8}$$

In this formulation, there is only one $\boldsymbol{P}_{\boldsymbol{t|t}}(\hat{\boldsymbol{\theta}})$ in the right-hand side of (4.3.8). This is due to the fact that the new term $\boldsymbol{E}_{\boldsymbol{B}}[\hat{\boldsymbol{\alpha}}^{\boldsymbol{b}}_{\boldsymbol{t|t}}(\hat{\boldsymbol{\theta}}^{\boldsymbol{b}}) - \boldsymbol{\alpha}^{\boldsymbol{b}}_{\boldsymbol{t}}][\hat{\boldsymbol{\alpha}}^{\boldsymbol{b}}_{\boldsymbol{t|t}}(\hat{\boldsymbol{\theta}}^{\boldsymbol{b}}) - \boldsymbol{\alpha}^{\boldsymbol{b}}_{\boldsymbol{t}}]'$, corresponding to the last term in the right-hand side of (4.3.8), can itself be decomposed, in the same fashion as in (4.3.4), into the parameter uncertainty $\boldsymbol{E}_{\boldsymbol{B}}[\hat{\boldsymbol{\alpha}}^{\boldsymbol{b}}_{\boldsymbol{t|t}}(\hat{\boldsymbol{\theta}}^{\boldsymbol{b}}) - \hat{\boldsymbol{\alpha}}^{\boldsymbol{b}}_{\boldsymbol{t|t}}(\hat{\boldsymbol{\theta}})][\hat{\boldsymbol{\alpha}}^{\boldsymbol{b}}_{\boldsymbol{t|t}}(\hat{\boldsymbol{\theta}}^{\boldsymbol{b}}) - \hat{\boldsymbol{\alpha}}^{\boldsymbol{b}}_{\boldsymbol{t|t}}(\hat{\boldsymbol{\theta}})]'$ and the filter uncertainty term $\boldsymbol{P}^{\boldsymbol{b}}_{\boldsymbol{t|t}}(\hat{\boldsymbol{\theta}}) = \boldsymbol{E}[\hat{\boldsymbol{\alpha}}^{\boldsymbol{b}}_{\boldsymbol{t|t}}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\alpha}^{\boldsymbol{b}}_{\boldsymbol{t}}][\hat{\boldsymbol{\alpha}}^{\boldsymbol{b}}_{\boldsymbol{t|t}}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\alpha}^{\boldsymbol{b}}_{\boldsymbol{t}}]'$, $\hat{\boldsymbol{\theta}}$ being the true parameter vector the bootstrap state variables $\boldsymbol{\alpha}^{\boldsymbol{b}}_{\boldsymbol{t}}$ are generated with. However, the bootstrap average term $\frac{1}{B}\sum_{b=1}^{B}[\hat{\boldsymbol{\alpha}}^{\boldsymbol{b}}_{\boldsymbol{t|t}}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\alpha}^{\boldsymbol{b}}_{\boldsymbol{t}}][\hat{\boldsymbol{\alpha}}^{\boldsymbol{b}}_{\boldsymbol{t|t}}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\alpha}^{\boldsymbol{b}}_{\boldsymbol{t}}]'$ replacing $\boldsymbol{P}_{\boldsymbol{t|t}}(\hat{\boldsymbol{\theta}})$ may need much more bootstrap iterations to converge. Further, one should be aware of the fact that this simplified method may result in an additional bias if the normality assumption about the model error terms is violated. Then, the term $\boldsymbol{E}_{\boldsymbol{B}}[\hat{\boldsymbol{\alpha}}^{\boldsymbol{b}}_{\boldsymbol{t|t}}(\hat{\boldsymbol{\theta}}^{\boldsymbol{b}}) - \boldsymbol{\alpha}^{\boldsymbol{b}}_{\boldsymbol{t}}][\hat{\boldsymbol{\alpha}}^{\boldsymbol{b}}_{\boldsymbol{t|t}}(\hat{\boldsymbol{\theta}}^{\boldsymbol{b}}) - \boldsymbol{\alpha}^{\boldsymbol{b}}_{\boldsymbol{t}}]'$ will also contain a non-zero expectation of the cross-terms: $\boldsymbol{E}\{[\hat{\boldsymbol{\alpha}}^{\boldsymbol{b}}_{\boldsymbol{t|t}}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\alpha}^{\boldsymbol{b}}_{\boldsymbol{t}}][\hat{\boldsymbol{\alpha}}^{\boldsymbol{b}}_{\boldsymbol{t|t}}(\hat{\boldsymbol{\theta}}^{\boldsymbol{b}}) - \hat{\boldsymbol{\alpha}}^{\boldsymbol{b}}_{\boldsymbol{t|t}}(\hat{\boldsymbol{\theta}})]\}$. In this application, the influence of non-zero cross-term bootstrap averages has turned out to be of a negligible importance, but the bootstrap average
$\frac{\boldsymbol{1}}{\boldsymbol{B}}\sum_{\boldsymbol{b=1}}^{\boldsymbol{B}}[\hat{\boldsymbol{\alpha}}^{\boldsymbol{b}}_{\boldsymbol{t|t}}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\alpha}^{\boldsymbol{b}}_{\boldsymbol{t}}][\hat{\boldsymbol{\alpha}}^{\boldsymbol{b}}_{\boldsymbol{t|t}}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\alpha}^{\boldsymbol{b}}_{\boldsymbol{t}}]'$
exhibited large deviations from the term it was meant to replace. This may be explained by the fact that the true Kalman filter MSE in (4.3.1) can be obtained from simulated series if the distribution of the state-vector is sufficiently dispersed. When bootstrapping non-stationary models, however, the bootstrap se-

ries are forced to follow the pattern of the underlying original series, as it has been mentioned in the description of the simulation smoother algorithm. Therefore, the replacement of $\boldsymbol{P_{t|t}(\hat{\theta})}$ with the term $\frac{1}{B}\sum_{b=1}^{B}[\hat{\alpha}_{t|t}^{b}(\hat{\theta}) - \alpha_t^b][\hat{\alpha}_{t|t}^{b}(\hat{\theta}) - \alpha_t^b]'$ in (4.3.8) is not equivalent. For this reason, both parametric (denoted as PT1) and non-parametric (PT2) bootstraps in this application rely on the estimator in (4.3.7).

A few words have to be said about the role of the simulation smoother of Durbin and Koopman (2002). We suggest it should be used at the bootstrap series generation step. Without it, the bootstrap hyperparameter distribution obtained from the unconditionally generated series in a non-stationary model could be very different from what it should be for a particular realisation of the data at hand. For a discussion on conditional and unconditional MSE estimation, see Ansley and Kohn (1986). At least in the case of the DLFS, omitting the simulation smoother step resulted in bootstrap hyperparameter distributions having a much wider range than that in distributions obtained with the help of the simulation smoother. Moreover, such (unconditional) bootstrap distributions for the DLFS are centered around values that are much larger than hyperparameter values the series have been generated with. This results in an excessively large bootstrap average $\frac{1}{B}\sum_{b=1}^{B}\boldsymbol{P_{t|t}(\hat{\theta}^b)}$ (relatively to $\boldsymbol{P_{t|t}(\hat{\theta})}$) and, subsequently, in MSE-estimates that are even lower than the naive ones. The term $\frac{1}{B}\sum_{b=1}^{B}[\hat{\alpha}_{t|t}^{b}(\hat{\theta}^b) - \hat{\alpha}_{t|t}^{b}(\hat{\theta})][\hat{\alpha}_{t|t}^{b}(\hat{\theta}^b) - \hat{\alpha}_{t|t}^{b}(\hat{\theta})]'$ also becomes very unstable over time and excessively large compared to when the simulation smoother is used, but that does not cure the negative bias obtained from (4.3.7) without the simulation smoother.

4.3.3 Asymptotic Approximation

An asymptotic approximation (AA) to the true MSE in equation (4.3.2) was developed by Hamilton (1986) and can be expressed as an expectation over the hyperparameter joint asymptotic distribution $\pi(\boldsymbol{\theta}|\boldsymbol{Y})$ conditional on the given dataset $\boldsymbol{Y} \equiv \{\boldsymbol{Y_1}, ..., \boldsymbol{Y_T}\}$. In this application, a part of the hyperparameter vector that is estimated with the ML-method (denoted as $\boldsymbol{\theta_{ML}}$), depends on the value of the autoregressive parameter ρ. Therefore, the hyperparameter joint distribution has the following form: $\pi(\boldsymbol{\theta}|\boldsymbol{Y}) = \pi(\rho|\boldsymbol{Y})\pi(\boldsymbol{\theta_{ML}}|\rho, \boldsymbol{Y})$. The MSE is approximated as

follows:

$$\begin{aligned}\boldsymbol{MSE_{t|t}} =& \boldsymbol{E_{\pi(\theta|Y)}}[\boldsymbol{P_{t|t}(\theta, Y)}]+ \\ & \boldsymbol{E_{\pi(\theta|Y)}}\left\{\boldsymbol{E_t}[(\hat{\boldsymbol{\alpha}}_{\boldsymbol{t|t}}(\boldsymbol{\theta, Y}) - \hat{\boldsymbol{\alpha}}_{\boldsymbol{t|t}}(\boldsymbol{Y}))(\hat{\boldsymbol{\alpha}}_{\boldsymbol{t|t}}(\boldsymbol{\theta, Y}) - \hat{\boldsymbol{\alpha}}_{\boldsymbol{t|t}}(\boldsymbol{Y}))']\right\},\end{aligned} \tag{4.3.9}$$

where $\boldsymbol{E_{\pi(\theta|Y)}}$ is an expectation taken over the hyperparameter joint asymptotic distribution $\pi(\boldsymbol{\theta|Y})$, and $\boldsymbol{E_t}$ is an expectation in the sense provided under (4.3.1). $\hat{\boldsymbol{\alpha}}_{\boldsymbol{t|t}}\boldsymbol{(Y)}$ are the state vector estimates when the hyperparameters are not known.

Distribution $N(\hat{\rho}, Var(\hat{\rho}))$ is taken to denote ρ's asymptotic distribution $\pi(\rho|\boldsymbol{Y})$, from which random ρ-realisations are drawn. From equation (3) in Bartlett (1946), and using the fact that the autoregressive coefficient in an AR(1) process is equal to the correlation estimate for lag 1, the variance estimator of ρ becomes: $Var(\hat{\rho}) \approx 1 - \hat{\rho}^2)1/T$. In the case of the DLFS, where $\hat{\rho} = 0.208$, this means that $\widehat{Var}(\hat{\rho}) \approx 0.96(1/T)$. Taking into account that the standard error is used for making draws from the asymptotic distribution, and that the square root is a concave function, the sample standard deviation would be an underestimate. Therefore, making ρ-draws by means of $1/\sqrt{T}$ as the asymptotic distribution's standard deviation would be a reasonable choice.

After a ρ-value is drawn from $\pi(\rho|\boldsymbol{Y})$, the other hyperparameters are re-estimated to obtain $\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}|\rho$ and the information matrix $\hat{\boldsymbol{I}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}|\rho_{\pi(\rho|\boldsymbol{Y})})$. Finally, a $\boldsymbol{\theta_{ML}}$-draw is made from distribution
$\sqrt{T}(\boldsymbol{\theta_{ML}} - \hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}|\rho_{\pi(\rho|\boldsymbol{Y})}) \sim MN(\boldsymbol{0}, T\hat{\boldsymbol{I}}^{-\boldsymbol{1}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}|\rho_{\pi(\rho|\boldsymbol{Y})}))$.
The Kalman filter is run again using ρ- and $\boldsymbol{\theta_{ML}}$-realisations to obtain the state estimates $\hat{\boldsymbol{\alpha}}_{\boldsymbol{t|t}}(\boldsymbol{\theta}_{\boldsymbol{\pi(\theta|Y)}}, \boldsymbol{Y})$ and their MSEs $\boldsymbol{P_{t|t}}(\boldsymbol{\theta}_{\boldsymbol{\pi(\theta|Y)}})$ ($\boldsymbol{\theta}_{\boldsymbol{\pi(\theta|Y)}}$-draws are further denoted as $\boldsymbol{\theta^a}$). The procedure is repeated until B $\boldsymbol{\theta^a}$-draws are obtained, whereafter (4.3.9) is obtained by averaging the necessary quantities over B iterations. If all the hyperparameters of the model are estimated within the ML-procedure, B draws can be made directly from
$\sqrt{T}(\boldsymbol{\theta}_{\boldsymbol{\pi(\theta|Y)}} - \hat{\boldsymbol{\theta}}) \sim MN(\boldsymbol{0}, T\hat{\boldsymbol{I}}^{-\boldsymbol{1}}(\hat{\boldsymbol{\theta}}))$.

The first term in (4.3.9) can be approximated by the average value of the Kalman filter variance $\boldsymbol{P_{t|t}}$ across B realizations of the hyperparameter vector, and the second term by the sampling variance of the state vector expectation. An asymp-

totic approximation to the true MSE could therefore be obtained in the following way:

$$\begin{aligned}\hat{\boldsymbol{MSE}}^{\boldsymbol{AA}}_{\boldsymbol{t|t}} =& \frac{1}{B}\sum_{\boldsymbol{a=1}}^{\boldsymbol{B}} \boldsymbol{P_{t|t}}(\boldsymbol{\theta^a}) + \\ &+ \frac{1}{B}\sum_{\boldsymbol{a=1}}^{\boldsymbol{B}} [\hat{\boldsymbol{\alpha}}_{\boldsymbol{t|t}}(\boldsymbol{\theta^a}, \boldsymbol{Y}) - \bar{\hat{\boldsymbol{\alpha}}}_{\boldsymbol{t|t}}][\hat{\boldsymbol{\alpha}}_{\boldsymbol{t|t}}(\boldsymbol{\theta^a}, \boldsymbol{Y}) - \bar{\hat{\boldsymbol{\alpha}}}_{\boldsymbol{t|t}}]'.\end{aligned} \tag{4.3.10}$$

where $\boldsymbol{\theta^a}$ is the a-th draw from the $\pi(\boldsymbol{\theta}|\boldsymbol{Y})$ asymptotic distribution. As Hamilton (1986) suggests, the sample average $\bar{\hat{\boldsymbol{\alpha}}}_{\boldsymbol{t|t}} = \frac{\boldsymbol{1}}{\boldsymbol{B}} \sum_{\boldsymbol{a=1}}^{\boldsymbol{B}} \hat{\boldsymbol{\alpha}}_{\boldsymbol{t|t}}(\boldsymbol{\theta}^{\boldsymbol{a}}_{\boldsymbol{\pi(\theta|Y)}}, \boldsymbol{Y})$ can replace $\hat{\boldsymbol{\alpha}}_{\boldsymbol{t|t}}(\boldsymbol{Y})$ in (4.3.9). Obviously, this MSE-estimator is entirely based on the asymptotic normality assumption about the hyperparameter vector estimator. Apart from that, this approach usually produces significant biases if the series is not of a sufficient length, in which case the asymptotic distribution would fail to approximate the finite (usually skewed) distribution of maximum-likelihood estimates.

Another problem with the asymptotic approach can appear if some of the hyperparameters are estimated to be close to zero. This can happen to the initial model estimates or during the procedure itself, e.g., due to certain extreme ρ-draws. In these cases, the asymptotic variance of such hyperparameters will be too large, which will inflate the MSE-estimates of the signal and its unobserved components. It may as well lead to a failure in inverting the information matrix for the hyperparameter vector.

4.4 The DLFS-specific Simulation and Bootstrap Setup for True MSE Estimation

The central data generating process of the present simulation study is the DLFS model. The performance of the five MSE estimation methods is examined on series of the original length from the DLFS survey (114 monthly time points from 2001(1) until 2010(6)), as well as on shorter series of length 80 months and longer ones of length 200. For each of these series lengths, a Monte-Carlo experiment is set up where multiple series (1000) are simulated on the basis of the LFS model used by Statistics Netherlands. MSEs for each of these series are estimated based on $B = 300$ bootstrap series; for asymptotic approximation, however, at least $B = 500$ draws turned out to be needed. This number has been found sufficient

for the approximated MSEs to converge. MSEs delivered by the five methods and averaged over the 1000 simulations are compared to MSE-averages produced by the "naive" Kalman filter. However, for the latter MSE estimates to converge to a certain average value, far more than 1000 simulations turned out to be needed (namely, at least 10000).

The hyperparameter maximum-likelihood estimates of the DLFS model, as well as the Yule-Walker estimate of the survey error autoregressive parameter ρ, are used to generate artificial data series $\hat{\boldsymbol{Y}}_{\boldsymbol{t}}^{\boldsymbol{s}}, s = 1, ...S$. Since the system is non-stationary, generating series $\boldsymbol{\alpha}_{\boldsymbol{t}}^{\boldsymbol{s}}$ unconditionally on the true data is going to result in negative or implausibly large numbers of the unemployed. In order to avoid an excessively large number of series with negative values, the unconditional recursion of the state variables is started with their smoothed estimates at one of the highest points the original process has ever reached within the available sample. Further, the first r time points are discarded in order to prevent that the series start at the same time-point (r being chosen equal to 30). With an assumption that unemployment in the Netherlands will not exceed 15 percent, the simulation data set is restricted to positive series below the upper bound of 1 mln of unemployed (this value comprised about 15 percent of the Dutch labour force in 2010). Keeping the artificial series below the upper bound is also done in order not to extrapolate outside of the original data range when simulating the design-based standard errors z_t^{t-j}, which are the time-dependent elements of the design-matrix $\boldsymbol{Z_t}$.

It is easy to guess that every series of simulated GREG point-estimates needs its own series of simulated design-based variance estimates, z_t^{t-j}s, that would depend on the corresponding numbers of the unemployed. That z_t^{t-j}s really depend on the number of the unemployed, can be seen from well-known formula for the variance estimator of the population total of a binary response variable (see, e.g., Särndal et al. (1992), (3.3.14)):

$$\widehat{Var}(\hat{Y}_t) = N_t^2 \, (1 - n_t/N_t) \, p_t(1 - p_t)/(n_t - 1), \tag{4.4.1}$$

where N_t is the population size in period t, n_t is the sample size, and p_t is the sample estimate of the unemployment rate Y_t/N_t. Taking logs on both sides of (4.4.1), rearranging the terms and assuming that unemployment p_t does not reach high values, which allows to neglect the term $ln(1 - p_t)$, produces the following

approximation function for the design-based variance estimator:

$$ln(\widehat{Var}(\hat{Y}_t)) \approx \alpha ln\,(n_t/N_t) + \beta ln(\hat{Y}_t), \tag{4.4.2}$$

where $\widehat{Var}(\hat{Y}_t)$ from (4.2.1) are used in the left-hand side to estimate the relationship between the variance estimates of the GREG estimator and the GREG point-estimates of the numbers of unemployed. This study simulates numbers of unemployed conditional on the same population and sample sizes as those observed in reality between 2001(1) and 2010(6). This allows to avoid simulating sample sizes in a study for longer series ($T = 200$). Instead, the information carried by the term $ln\,(n_t/N_t)$ can be represented by an intercept for the variances of the first wave. Starting from wave 2, each design-based variance, denoted as $(z_t^{t-j})^2$, $j = \{3, 6, 9, 12\}$, is highly dependent on the design-based variance of the preceding wave, since both are based on nearly the same group of people and sample size. The sample size decreases by approximately 10 percent in each subsequent wave due to panel attrition. Further, the signal $l_t^{t-j} = L_t + \gamma_t + \lambda_t^{t-j}$ can act as a proxy for $\hat{Y}_t^{t-j}$ in (4.4.2), since both the signal and the direct estimate contain information on the level of unemployed and the design effect (in this case, the RGB). The design-based standard errors z_t^{t-j} can be derived from the following equations for log-variances:

$$\begin{aligned} &ln[(z_t^{t-j})^2] = c + \beta^j ln(l_t^{t-j}) + \epsilon_t^{t-j}, j = 0; \\ &ln[(z_t^{t-j})^2] = \rho^j ln[(z_{t-3}^{t-j})^2] + \beta^j ln(l_t^{t-j}) + \epsilon_t^{t-j}, \epsilon_t^{t-j} \sim N(0, (\sigma_\epsilon^j)^2), j = \{3, 6, 9, 12\}. \end{aligned} \tag{4.4.3}$$

The regression coefficients in (4.4.3) are time-invariant. The superscripts are used to denote the wave these coefficients belong to. The coefficient estimates are presented in Table 4.4.1. They are obtained using the original design-based variances for the five panel waves and the extracted filtered signal estimates coming from the DLFS model. For series longer than the original series, the design-based standard errors $\boldsymbol{z_t^s}$ have to be simulated as well according to the same process.

Equations (4.2.2)-(4.2.3) and the estimated parameter vector $\hat{\boldsymbol{\theta}}$ are used to generate S=1000 series of artificial data. State disturbances (remember survey errors are also modelled as state variables) are randomly drawn from their joint normal distribution $N(\mathbf{0}, \boldsymbol{\Omega}(\hat{\boldsymbol{\theta}}))$, and series are generated unconditionally on the true

Table 4.4.1: Regression estimates for the design-based standard error process

	$j=0$	$j=3$	$j=6$	$j=9$	$j=12$
$\hat{c}$	12.219	-	-	-	-
$\hat{\beta}^j$	0.630	0.244	0.354	0.414	0.413
$\hat{\rho}^j$	-	0.859	0.786	0.749	0.751
$\hat{\sigma}_\epsilon^j$	0.202	0.265	0.228	0.225	0.267

data. Within each simulation, first the trend, seasonal and RGB components are simulated and summed up to comprise the wave-signals $l_{t,s}^{t-j}, j = \{0,3,6,9,12\}$. These are used to generate the design-based standard errors $z_{t,s}^{t-j}$ according to the process described in (4.4.3). As soon as an artificial data set is generated, ρ is re-estimated and saved as $\hat{\rho}_s$, whereafter the hyperparameter quasi-ML estimates are obtained. These are stored in $\hat{\boldsymbol{\theta}}_{\boldsymbol{s}}$ and used by the Kalman filter to produce the state estimates $\hat{\boldsymbol{\alpha}}_{\boldsymbol{t,s}}$. Both $\hat{\rho}_s$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{s}}$ are used to generate bootstrap samples. Note that the same set of design-based standard errors $\boldsymbol{z_{t,s}}$ is used to both generate and estimate bootstrap series within simulation s.

In order to obtain the true MSEs, the DLFS model is simulated a large number of times $M = 50000$, each of these replications being restricted to the same limits as before, i.e. between zero and 1 mln unemployed. The true MSE is calculated in the following way using the true state vector $\boldsymbol{\alpha_{m,t}}$ values known for every simulation m:

$$\boldsymbol{MSE_t^{true}} = \frac{\mathbf{1}}{\boldsymbol{M}} \sum_{\boldsymbol{m=1}}^{\boldsymbol{M}} [(\hat{\boldsymbol{\alpha}}_{\boldsymbol{m,t}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{m}}) - \boldsymbol{\alpha}_{\boldsymbol{m,t}})(\hat{\boldsymbol{\alpha}}_{\boldsymbol{m,t}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{m}}) - \boldsymbol{\alpha}_{\boldsymbol{m,t}})'].$$

The true MSE of the signal is calculated in the same way by using the true wave-signal values $\boldsymbol{l_{m,t}}$.

4.5 MSE Estimation Approaches with Application to the DLFS

The focus of this simulation study is the true MSEs of the trend and of the population signal. The latter consists of the trend and seasonal components and is therefore equal to the signal of the first wave. Apart from the MSE estimation, this simulation also checks the DLFS model for misspecification. Formal diagnostic

tests for normality, homogeneity and independence of the standardised innovations do not detect any problems with the DLFS model. However, asymptotic confidence intervals of the hyperparameters, together with their distribution across a large number of the DLFS model replications give a clear indication that the model tends to be overspecified in a sense that some components may have to be modelled as time-invariant as their hyperparameters tend to zero. Therefore, this study considers four models that differ in terms of the number of hyperparameters to be estimated with the ML method. The most complete model - Model 1 - is the one currently in use at Statistics Netherlands, but with the white noise component ε_t removed from the true population parameter ξ_t. This component has turned out to have an implausibly large variance and disturbed estimation of other marginally significant hyperparameters (the seasonal and RGB disturbance variances) in the case of the DLFS. In order to avoid this instability, this irregular component ε_t can be removed from the model. This formulation implies that the population parameter ξ_t does not suffer from any unusual irregularities that cannot be picked up by the stochastic structure of the trend and seasonal components. This assumption can be advocated by a relative rigidity of labour markets. Alterations of unemployment levels are usually gradual and therefore must be largely incorporated into the stochastic trend movements. The other three models are special cases of Model 1, thus with ε_t component removed (see Table 4.5.1).

Table 4.5.1: Hyperparameters estimated in the four versions of the DLFS model; the disturbance variances estimated on a log-scale

Models	Description	Parameters estimated
M1	complete model	$\rho, \sigma_R^2, \sigma_\omega^2, \sigma_\lambda^2, \sigma_{\nu_1}^2, \sigma_{\nu_2}^2, \sigma_{\nu_3}^2, \sigma_{\nu_4}^2, \sigma_{\nu_5}^2$
M2	seasonal time-independent	$\rho, \sigma_R^2, \sigma_\lambda^2, \sigma_{\nu_1}^2, \sigma_{\nu_2}^2, \sigma_{\nu_3}^2, \sigma_{\nu_4}^2, \sigma_{\nu_5}^2$
M3	RGB time-independent	$\rho, \sigma_R^2, \sigma_\omega^2, \sigma_{\nu_1}^2, \sigma_{\nu_2}^2, \sigma_{\nu_3}^2, \sigma_{\nu_4}^2, \sigma_{\nu_5}^2$
M4	seasonal, RGB fixed	$\rho, \sigma_R^2, \sigma_{\nu_1}^2, \sigma_{\nu_2}^2, \sigma_{\nu_3}^2, \sigma_{\nu_4}^2, \sigma_{\nu_5}^2$

Note that the disturbance variances are estimated on a log-scale in order to avoid negative estimates. The rationale behind studying the other three models becomes clear after inspecting the hyperparameter distribution of Model 1 after a large number of replications. The simulation has shown that the stochastic term variances of the seasonal and, in particular, RGB components are often estimated to be close to zero. This causes bi-modality in the distribution of these variance estimates with a significant mass concentrated around zero. Apart from that, an

attempt to estimate both of the hyperparameters, as in Model 1, seems to bring about certain instability to the hyperparameter estimates, such that even normality in $ln(\hat{\sigma}^2_{\nu_3}), ln(\hat{\sigma}^2_{\nu_4})$ and $ln(\hat{\sigma}^2_{\nu_5})$ (where indices 3, 4, and 5 stand for the numbers of the waves) is severely violated with extreme outliers and/or a huge kurtosis (see Fig. 4.B.1 in Appendix 4.B, where the x-axis is extended due to the outliers). Making the seasonal component time-invariant, as in Model 2, hardly changes the situation for the slope and RGB hyperparameters. Instead, it may even be seen as suboptimal due to more extreme outliers and excess kurtosis in the distribution of all the five survey error hyperparameters (Fig. 4.B.2). By contrast, under both models where the RGB-component is fixed over time (Models 3 and 4), all hyperparameters corresponding to the survey error component have turned out to be normally distributed, see Fig. 4.B.3 and Fig. 4.B.4. Under Model 3, distributions are still skewed for the slope and seasonal components (skewness of -0.88 and -0.72, and excess kurtosis of 2.56 and 1.61, respectively). Fixing σ^2_ω to zero under Model 4 results in only a marginal improvement: the distribution of $ln(\hat{\sigma}^2_R)$ is negatively skewed (-0.81) with an excess kurtosis of 1.76.

This simulation evidence suggests that the preference in modelling the DLFS series may be given to the more parsimonious Model 3, where only the RGB disturbance variance is set equal to zero. This hyperparameter is however retained for production purposes at Statistics Netherlands to secure the model robustness against sudden changes in the underlying process.

The distribution of the survey error autoregressive parameter ρ across the 1000 simulated series does not seem to be affected by model reformulations and ranges between 0 and 40 percent when $T = 114$, which is in line with the approximation of its asymptotic distribution mentioned in Subsection 4.3.3. The range is slightly wider when $T = 80$ and narrower when $T = 200$. The simulation procedure described in the previous section and the analysis of bootstrap methods that follows is performed separately for all the four models.

The performance of the Kalman filter at estimated parameter values, as well as of the five estimation methods mentioned in Section 4.3 is evaluated with the help of the MSE relative bias. First, the filtered MSE estimates from (4.3.10), (4.3.3), (4.3.7), and (4.3.8) are averaged over 1000 simulations, and the Kalman filter MSE estimates over 10000 simulations, as mentioned at the beginning of Section

4. These averaged MSE estimates for Model 3 (except for AA for the reason that will become clear soon) are depicted in Fig. 4.5.1, 4.5.2 and 4.5.3 for three different series lengths, skipping the first $d = 30$ time points of the sample. This is the time needed for the diffuse part of the state covariance matrix to decay (see Koopman (1997) for initialisation of non-stationary state variables). The percentage relative bias is calculated as $RB_t^f = \left(\overline{MSE}_{t|t}^f / MSE_{t|t}^{true} - 1\right) \cdot 100\%$, where f defines a particular estimation method. The core variables of interest for users of the DLFS are the signal and trend. The percentage relative MSE biases averaged over time (skipping the first $d = 30$ time points) for the signal, the trend and seasonal components are presented in Tables 4.5.2, 4.5.3, and 4.5.4. Note that all the analysis is based on filtered, rather than smoothed estimates, because filtered estimates better mimic the process of official figures production.

The AA-method turned out to be inapplicable to the models with marginally significant hyperparameters. When some of the hyperparameters are estimated close to zero, the matrix $\boldsymbol{I^{-1}(\hat{\theta}_{ML}|\rho^a))}$ is numerically either singular, leading to a failure in the procedure, or nearly singular. In the latter case, the asymptotic variance becomes excessively large and thus not reliable. Taking this into account, the AA-method could only be considered for Model 4. As expected, the method performs poorly in short series, with positive biases of about 15 percent. The performance for $T = 114$ and $T = 200$ is comparable to that of the PT1-bootstrap, but significantly worse than the PT2 performance.

The simulation results for $T = 80$ suggest that, when averaged over time (starting from $t = 30$), the relative bias of the signal MSE obtained with the Kalman filter ranges between -3.2 and -2.1 percent for the four models considered in the paper. This bias tends to decrease as the series length increases. The KF-biases are quite small for the case of $T = 200$, such that none of the estimation methods offers a smaller bias in absolute terms. One could still apply the best estimation method with positive biases in order to get a range of values containing the true MSE.

What one immediately sees is negative biases for the RR-bootstrap and positive ones for the PT-method. Against the claim of Rodriguez and Ruiz (2012) that their approach has better finite sample properties compared to the approach of Pfeffermann and Tiller (2005), the case of the DLFS suggests that the RR-estimates, both parametric and non-parametric ones, are even more negatively biased than

the uncorrected KF-estimates across all the models and series lengths (except for RR2 in Model 1, $T = 80$ and $T = 114$). The PT-methods have never produced negative biases for the DLFS, which makes these methods conservative and thus safe. While the PT-bootstrap is proven to have satisfactory asymptotic properties in Pfeffermann and Tiller (2005), Rodriguez and Ruiz (2012) illustrate the superiority of their method in small samples based on a simple model (a random walk plus noise). The present simulation study reveals that the RR-method may not behave well in more complex applications.

For series of lengths $T = 114$ and $T = 80$, positive biases produced by the PT2-bootstrap may slightly exceed the KF-biases in absolute values for models with insignificant hyperparameters (Models 1 and 2). In the more stable models (Models 3 and 4), the positive biases are smaller than the KF negative biases in absolute values.

The signal MSE of Model 3, which could be considered a better option for production of official DLFS figures, is best estimated by the PT2 approach, with relative biases of 1.4 and 1.9 percent for $T = 80$ and $T = 114$, respectively. The PT2-bootstrap also seems to be the best method for $T = 200$, but, as has been said, the negative KF biases are already quite small for series of this length.

Note that for both the PT- and RR-bootstraps, the absolute values of relative biases are smaller in the case of the non-parametric approaches, compared to their parametric counterparts. The superiority of the non-parametric approach over the parametric one can be explained by the disturbed normality of the error distribution in the models.

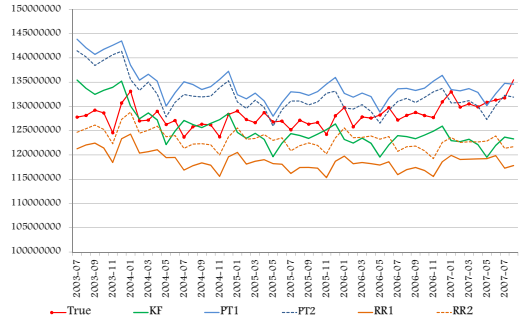


Figure 4.5.1: Signal MSE comparison for Model 3, T=80 months

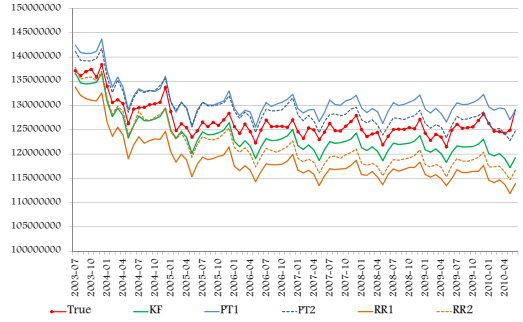


Figure 4.5.2: Signal MSE comparison for Model 3, T=114 months

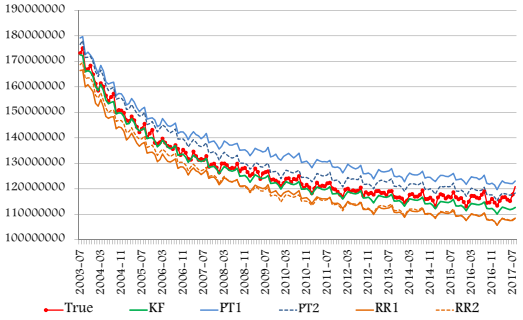


Figure 4.5.3: Signal MSE comparison for Model 3, T=200 months

Table 4.5.2: MSE relative bias averaged over time (d=30) in the DLFS model, percent, T=80

	Signal				Trend				Seasonal			
Models	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
KF	-3.0	-3.2	-2.1	-2.2	-3.5	-3.8	-2.5	-2.5	8.8	2.5	2.9	2.4
AA	N/A	N/A	N/A	14.9	N/A	N/A	N/A	15.0	N/A	N/A	N/A	14.9
PT1	8.6	6.7	4.9	6.2	10.6	8.9	7.1	8.4	20.8	10.7	10.3	11.1
PT2	**4.8**	**3.7**	**1.4**	**2.1**	4.8	4.9	2.1	2.3	17.3	8.2	6.9	7.1
RR1	-7.2	-9.0	-7.3	-7.2	-9.6	-11.2	-9.6	-9.5	-3.8	-9.0	-6.7	-6.6
RR2	6.7	-3.5	-3.9	-3.7	5.3	-4.1	-4.6	-4.3	18.6	-4.7	-4.1	-4.8

Table 4.5.3: MSE relative bias averaged over time (d=30) in the DLFS model, percent, T=114

	Signal				Trend				Seasonal			
Models	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
KF	-2.1	-2.6	-2.4	-2.2	-2.3	-2.7	-2.4	-2.3	2.5	-3.2	-3.1	-2.6
AA	N/A	N/A	N/A	5.2	N/A	N/A	N/A	4.1	N/A	N/A	N/A	12.5
PT1	8.1	5.7	3.3	5.5	10.0	7.9	5.2	7.6	4.9	1.4	1.4	0.3
PT2	**2.2**	**3.2**	**1.9**	**1.5**	3.3	4.3	3.1	2.8	1.2	-2.0	1.0	0.6
RR1	-8.3	-7.8	-6.4	-6.5	-10.7	-9.9	-8.7	-8.9	-3.1	-7.2	-5.5	-5.6
RR2	-1.1	-6.0	-3.9	-3.5	-3.0	-7.6	-5.5	-5.0	7.3	-5.9	-3.2	-3.0

Table 4.5.4: MSE relative bias averaged over time (d=30) in the DLFS model, percent, T=200

	Signal				Trend				Seasonal			
Models	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
KF	-1.3	-1.6	-1.3	-1.3	-1.7	-1.8	-1.6	-1.6	3.8	-1.7	-1.6	-1.6
AA	N/A	N/A	N/A	5.9	N/A	N/A	N/A	5.6	N/A	N/A	N/A	5.6
PT1	**6.3**	6.2	6.3	5.5	7.5	7.7	7.8	7.1	10.8	2.6	3.0	3.0
PT2	6.8	**4.0**	**3.0**	**2.3**	7.6	4.9	4.2	3.6	12.5	2.1	1.3	0.6
RR1	-8.0	-8.0	-4.9	-5.9	-10.0	-9.9	-6.8	-7.1	-1.1	-5.3	-3.8	-3.9
RR2	-5.1	-5.6	-4.5	-5.0	-7.0	-7.4	-6.0	-6.4	3.6	-3.1	-3.3	-3.9

In order to see if the STS model-based approach still offers some reduction in the design-based variance estimates after correcting for hyperparameter uncertainty, percentage reductions in the standard errors of the GREG estimates are presented in Table 4.5.5. These reductions come from applying the DLFS model to the GREG estimates without correcting for hyperparameter uncertainty (KF), as well as after this uncertainty has been taken into account with the help of the five different methods. Note that the RGB and seasonal hyperparameter estimates obtained from the original DLFS data set are quite small. Therefore, there are no noticeable differences between the point-estimates of the four models. However, these hyperparameter estimates are not small enough to cause the problems mentioned before with regard to the AA-approach, so the results for this approximation method are reported in Table 4.5.5 as well. It may first seem that the AA-method accounts for the parameter uncertainty best of all other methods, as it exhibits the smallest reduction in the GREG standard error estimates. However, keeping in mind that the AA may on average have some severe positive biases, especially in the case of small hyperparameters, one should feel more secure with the PT2 non-parametric approach that offers about 22 percent reduction in the estimated GREG standard errors. This means that the model-based approach even after accounting for parameter uncertainty offers a significant variance reduction compared with the traditional design-based approach.

Table 4.5.5: Reductions in the GREG SE estimates of the DLFS, averaged over time (d=30), and percentage increase in the model SE after applying the MSE correction in parentheses

	Model 1	Model 2	Model 3	Model 4
KF	-24.1	-24.1	-24.5	-24.5
AA	-18.8 (6.9)	-19.0 (6.7)	-19.1 (7.1)	-19.5 (6.6)
PT1	-20.1 (5.2)	-20.1 (5.2)	-21.1 (4.6)	-21.2 (4.4)
PT2	-22.9 (1.6)	-21.2 (3.8)	-22.2 (3.1)	-22.5 (2.6)
RR1	-26.5 (-3.2)	-26.6 (-3.4)	-26.5 (-2.7)	-26.5 (-2.7)
RR2	-24.0 (-0.1)	-25.4 (-1.8)	-25.6 (-1.4)	-25.7 (-1.6)

4.6 Conclusion

Most applications of small area estimation procedures in the literature are based on multilevel models. In the framework of this approach, it is common practice to account for the hyperparameter uncertainty in estimated MSEs. The literature on

structural time series models applied in the context of SAE is still rather limited, with most applications ignoring the hyperparameter uncertainty when computing the MSEs of small area predictions. This renders MSE estimates negatively biased when series are not long enough, which may be a serious issue when it comes to such important economic indicators as unemployment.

The literature offers several procedures to correct for the negative bias in the MSE estimates produced by STS models. The present work aimed at establishing the best estimation approach to the true MSE of a small area estimation approach applied to the DLFS for official production of estimated numbers of the unemployed in the Netherlands.

A simulation study conducted for this purpose reveals that the asymptotic approximation is not applicable to cases with hyperparameters close to zero due to failures when inverting the information matrix of the hyperparameter estimates. The simulation results suggest that the non-parametric bootstraps, being free of normality assumptions about the error distribution, perform better than their parametric counterparts in both Pfeffermann and Tiller (2005) and Rodriguez and Ruiz (2012) methods. A more important finding, however, is that the Pfeffermann and Tiller (2005) bootstrap approaches with their positive biases consistently outperform the respective approaches of Rodriguez and Ruiz (2012), where the biases are generally negative and larger than those of the Kalman filter in absolute terms. This is contrary to the claim of Rodriguez and Ruiz (2012) about the superiority of their method in short time series. Apparently, their findings are purely heuristic and are based on a simple model simulation (random walk plus noise), while Pfeffermann and Tiller (2005) prove that their bootstrapping approach produces MSE estimates with a bias of a correct order. Hence, the PT-methods should be considered for other survey data too, despite the fact that these methods may occasionally be outperformed by the RR-methods.

Another result of this simulation study has revealed that it might be worth considering a more restricted version of the DLFS model, with the variance of the RGB component and of the population parameter noise set equal to zero. For this model, the relative bias of the signal MSE produced by the Kalman filter can be reduced from about -2.4 to 1.9 percent with the non-parametric Pfeffermann and Tiller (2005) bootstrap approach. Even with this slightly positive bias, the

standard errors of the GREG estimates are reduced by about 22 percent. For the DLFS application, the bias in the Kalman filter MSE estimates is relatively small, therefore it may be deemed sufficient to rely on these naive MSE estimates for publication purposes.

4.A Some Details on the State Space Form of the DLFS Model

The measurement equation design matrix is a composite of four matrices: $\boldsymbol{Z_t} = (\boldsymbol{Z^{\xi}}\ \boldsymbol{Z^{\lambda}}\ \boldsymbol{Z_t^e}\ \boldsymbol{O_{5\times 8}})$, where $\boldsymbol{O_{5\times 8}}$ denotes a null-matrix of a dimension specified in the subscript, $\boldsymbol{Z_t^{\xi}} = [\mathbf{1_5} \otimes (1\,0\,1\,0\,1\,0\,1\,0\,1\,0\,1\,0\,1\,1)]$ selects the level L_t, six seasonal harmonics $\gamma_{t,1}, ...\gamma_{t,6}$ and the population parameter white noise ε_t for each of the five waves,
$\boldsymbol{Z_t^{\lambda}} = \begin{pmatrix} \mathbf{0_4'} \\ \boldsymbol{I_4} \end{pmatrix}$ selects the four RGB components for the corresponding waves, and
$\boldsymbol{Z_t^e} = Diag(z_t^t\, z_t^{t-3} z_t^{t-6} z_t^{t-9} z_t^{t-12})$ multiplies the five survey error components $\tilde{e}_t^{t-j}$ with the design-based standard errors. Vectors $\mathbf{1_5}$ and $\mathbf{0_4'}$ denote a vertical vector of ones and a horizontal vector of zeros, respectively, of a dimension specified in the subscript.

The transition matrix $\boldsymbol{T}$ is specified below:

$$\boldsymbol{T} = Blockdiag(\boldsymbol{T^L}\ \boldsymbol{T^{\gamma}}\ 0\ \boldsymbol{T^{\lambda}}\ \boldsymbol{T^e}),$$

$$\text{where } \boldsymbol{T^L} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \boldsymbol{T^{\gamma}} = Blockdiag(\boldsymbol{C_1 ... C_5}\ -1),$$

$$\boldsymbol{C_l} = \begin{pmatrix} cos(\frac{l\pi}{6}) & sin(\frac{l\pi}{6}) \\ -sin(\frac{l\pi}{6}) & cos(\frac{l\pi}{6}) \end{pmatrix}, l = \{1, ..., 5\},$$

$$\boldsymbol{T^{\lambda}} = \boldsymbol{I_4},$$

$$\boldsymbol{T^e} = \begin{pmatrix} \mathbf{0_4'} & 0 & \mathbf{0_4'} & \mathbf{0_4'} \\ \boldsymbol{O_{4\times 4}} & \mathbf{0_4} & \rho\boldsymbol{I_4} & \boldsymbol{O_{4\times 4}} \\ \boldsymbol{O_{4\times 4}} & \mathbf{0_4} & \boldsymbol{O_{4\times 4}} & \boldsymbol{I_4} \\ \boldsymbol{I_4} & \mathbf{0_4} & \boldsymbol{O_{4\times 4}} & \boldsymbol{O_{4\times 4}} \end{pmatrix}.$$

Vector $\boldsymbol{\eta_t}$ contains stochastic terms of the state vector $\boldsymbol{\alpha_t}$:

$$\boldsymbol{\eta_t} = (0,\ \eta_{R,t},\ \boldsymbol{\eta_{\gamma}},\ \varepsilon_t,\ \boldsymbol{\eta_{\lambda}},\ \nu_t^t, \nu_t^{t-3}, \nu_t^{t-6}, \nu_t^{t-9}, \nu_t^{t-12},\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0)',$$

where $\boldsymbol{\eta_\gamma} = (\omega_{t,1}, \omega^*_{t,1}, \omega_{t,2}, \omega^*_{t,2}, ..., \omega_{t,5}, \omega^*_{t,5}, \omega_{t,6})$, $\boldsymbol{\eta_\lambda} = (\eta^{t-3}_{\lambda,t}, \eta^{t-6}_{\lambda,t}, \eta^{t-9}_{\lambda,t}, \eta^{t-12}_{\lambda,t})$. The covariance matrix of the state stochastic terms is diagonal:

$$\boldsymbol{\Omega} = Blockdiag(0\, \sigma^2_R\, [\sigma^2_\omega \mathbf{1'_{11}}]\, \sigma^2_\varepsilon\, \boldsymbol{\Omega^\lambda}\, \boldsymbol{\Omega^e}),$$
$$\boldsymbol{\Omega^\lambda} = \sigma^2_\lambda \boldsymbol{I_4},$$
$$\boldsymbol{\Omega^e} = Diag(\sigma^2_{\nu^1}\, \sigma^2_{\nu^2}\, \sigma^2_{\nu^3}\, \sigma^2_{\nu^4}\, \sigma^2_{\nu^5}\, \mathbf{0'_8}).$$

In the case of the LFS, all the hyperparameters estimated with the ML-method are contained in the $\boldsymbol{\Omega}$-matrix, whereas the hyperparameter ρ in the transition matrix $\boldsymbol{T}$.

4.B Simulated Density Functions of the Hyperparameters under the Four Versions of the DLFS Model

This appendix presents the hyperparameter density functions obtained from simulations where the four versions of the DLFS model act as the data generating process. The x-axes depict variance hyperparameters on a log-scale, while the y-axes stand for frequencies. The x-axis may be extended due to outliers.

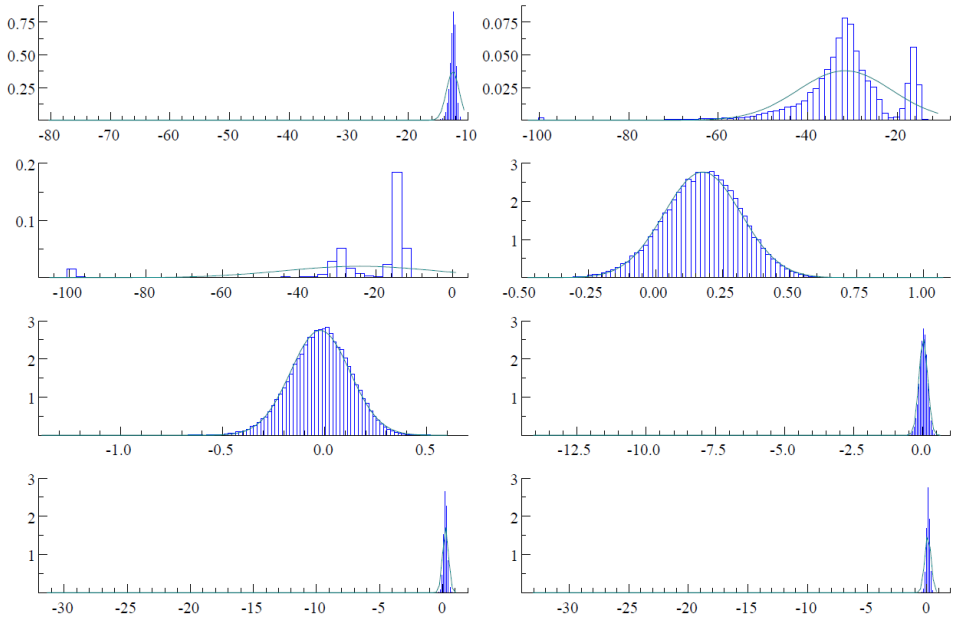


Figure 4.B.1: Hyperparameter distribution under the complete DLFS model (Model 1), left to right on the x-axes: $ln(\hat{\sigma}^2_R), ln(\hat{\sigma}^2_\gamma), ln(\hat{\sigma}^2_\lambda), ln(\hat{\sigma}^2_{\nu_1}), ln(\hat{\sigma}^2_{\nu_2}), ln(\hat{\sigma}^2_{\nu_3}), ln(\hat{\sigma}^2_{\nu_4}), ln(\hat{\sigma}^2_{\nu_5})$; the normal density with the same mean and variance superimposed; 50000 simulations, T=114.

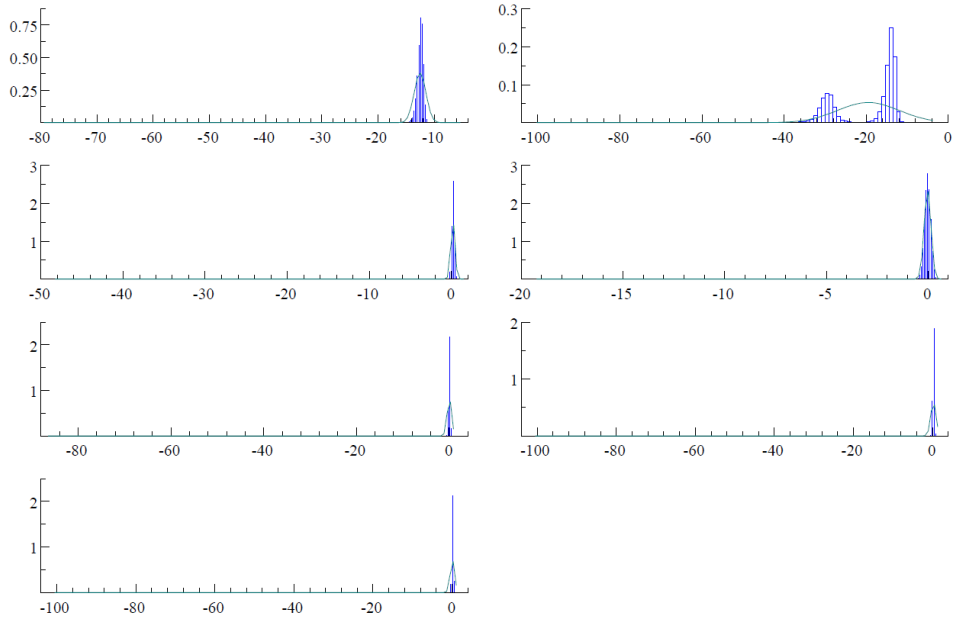


Figure 4.B.2: Hyperparameter distribution under Model 2, left to right on the x-axes: $ln(\hat{\sigma}^2_R), ln(\hat{\sigma}^2_\lambda), ln(\hat{\sigma}^2_{\nu_1}), ln(\hat{\sigma}^2_{\nu_2}), ln(\hat{\sigma}^2_{\nu_3}), ln(\hat{\sigma}^2_{\nu_4}), ln(\hat{\sigma}^2_{\nu_5})$; the normal density with the same mean and variance superimposed; 50000 simulations, T=114.

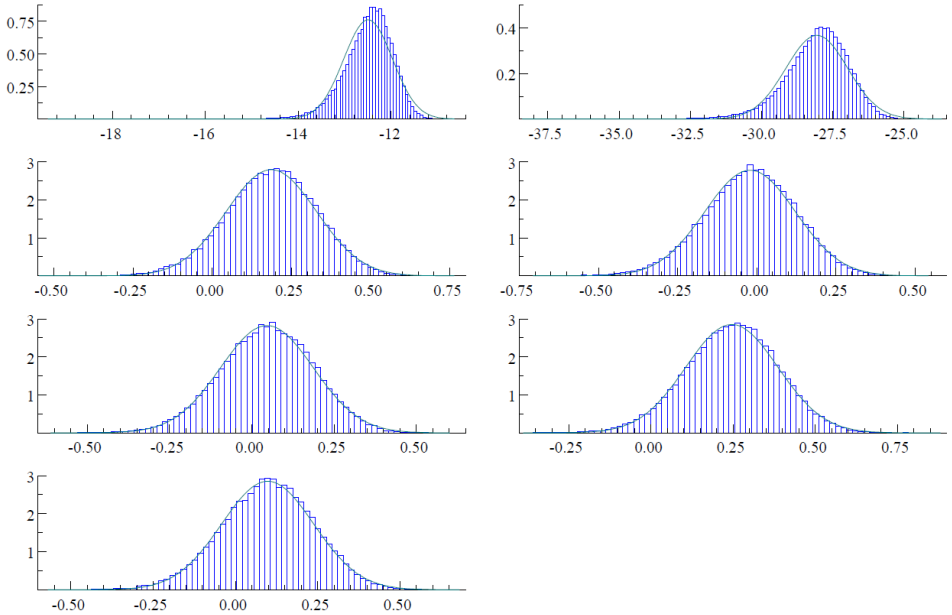


Figure 4.B.3: Hyperparameter distribution under Model 3, left to right on the x-axes: $ln(\hat{\sigma}^2_R), ln(\hat{\sigma}^2_\gamma), ln(\hat{\sigma}^2_{\nu_1}), ln(\hat{\sigma}^2_{\nu_2}), ln(\hat{\sigma}^2_{\nu_3}), ln(\hat{\sigma}^2_{\nu_4}), ln(\hat{\sigma}^2_{\nu_5})$; the normal density with the same mean and variance superimposed; 50000 simulations, T=114.

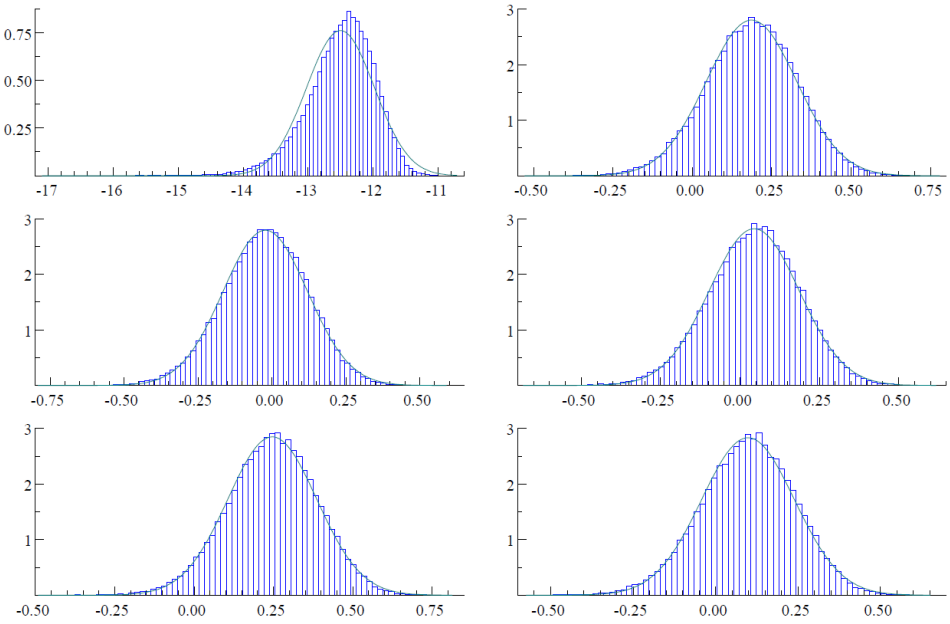


Figure 4.B.4: Hyperparameter distribution under Model 4, left to right on the x-axes: $ln(\hat{\sigma}^2_R), ln(\hat{\sigma}^2_{\nu_1}), ln(\hat{\sigma}^2_{\nu_2}), ln(\hat{\sigma}^2_{\nu_3}), ln(\hat{\sigma}^2_{\nu_4}), ln(\hat{\sigma}^2_{\nu_5})$; the normal density with the same mean and variance superimposed; 50000 simulations, T=114.

Chapter 5

Multilevel Hierarchical Bayesian vs. State Space Approach in Time Series Small Area Estimation: the Dutch Travel Survey

This chapter compares state space models (estimated with the Kalman filter with a frequentist approach to hyperparameter estimation), with multilevel time series models (based on the hierarchical Bayesian framework). The application chosen is the Dutch Travel Survey featuring small sample sizes and discontinuities caused by the survey redesigns. Both modelling approaches deliver similar point- and variance estimates. Slight differences in model-based variance estimates appear mostly in small-scaled domains and are due to neglecting uncertainty around the hyperparameter estimates in the state space models, and to a lesser extent due to skewness in the posterior distributions of the parameters of interest. The results suggest that the reduction in design-based standard errors with the hierarchical Bayesian approach is over 50% at the provincial level, and over 30% at the national level.[1]

[1]This chapter is based on the paper Bollineni-Balabay et al. (2016b) published by Statistics Netherlands as a discussion paper. The authors thank Dr. Bart Buelens and Dr. Barry Schouten (Statistics Netherlands) for valuable comments on an earlier draft of this paper.

5.1 Introduction

The problem of small area estimation (SAE) emerges when reliable estimates for small areas cannot be produced relying solely on a design-based inference procedure due to an insufficient number of sampled units. Either the number of units sampled in the domain of interest is too small, or the variable of interest has a rare occurrence, in cases when the break-up into domains is defined post hoc, after the survey has been done. An increasing demand for small area statistics boosted the development of numerous SAE techniques. Most of these are meant for cross-sectional data, where the so-called *strength over space* could be borrowed. This implies that sample information from other similar domains could be used to improve the design estimate for a small area of interest. Sometimes, an area may not be sampled at all, in which case regression synthetic estimators (Gonzalez (1973)) produce figures using some auxiliary variables. Composite estimators may be used to combine direct estimates with their synthetic counterparts. This is often done in the framework of the Fay-Herriot model (Fay and Herriot (1979)), which is estimated using the empirical best linear unbiased predictor (EBLUP) or hierarchical Bayesian (HB) approach. The literature on these methods is extensive, see, e.g., Rao and Molina (2015), Ghosh and Lahiri (1987), Prasad and Rao (1990).

While small area estimates may be improved by borrowing strength over space, adding the time dimension further offers a huge potential for improvement. *Borrowing strength over time* is realised by using sampling information accumulated over time in the respective, as well as in other domains. Time series techniques, such as multilevel times series and structural (unobserved component) time series (STS) models, are powerful tools in producing more reliable estimates for repeatedly conducted surveys, both for small and non-small areas. Thanks to their ability to borrow strength over time and domain space, such models can extract the so-called signal by removing a great part of the sampling noise from design-based point-estimates. As a result, standard errors (SEs) of such model-based point-estimates are usually substantially lower than the design SEs. Furthermore, time series models can help compare figures prior to and after survey redesigns by modelling level breaks. Time series extensions of the Fay-Herriot model can be found in Rao and Yu (1994), Datta et al. (1999), and You (2008) (see Rao and Molina (2015) for an overview). For small area applications modelled with the STS techniques, see Pfeffermann (1991), Pfeffermann and Bleuer (1993), Pf-

effermann and Tiller (2006), Pfeffermann et al. (2014), Krieg and Van den Brakel (2012) and Bollineni-Balabay et al. (2016a). Level break estimation within the STS framework is illustrated in Van den Brakel and Roels (2010) and Van den Brakel and Krieg (2015).

With this paper, the authors would like to contribute to empirical evidence for and practical considerations in time series SAE. This study compares two different time series techniques in a real life time series small area application of Statistics Netherlands - the Dutch Travel Survey (DTS). This survey features unacceptably volatile estimates in its more than 600 small domains. The reasons for that are small sample sizes at the provincial level and several survey redesigns, which make the official figures hard to compare over time. Therefore, this survey could largely benefit from the time series model-based approach in producing official statistical figures. The two time series approaches considered in this paper are multilevel time series modelling within the hierarchical Bayesian (HB) framework and the frequentist-based STS approach facilitated by the Kalman filter. The initial intention was to check if specifications with random regression coefficients and/or random intercepts, like in multilevel models, perform better than STS models where such components are conventionally modelled as fixed effects. Further, another model is considered within the multilevel Bayesian framework - a model with a likelihood function nearly identical to that of the STS model (and differing only in the way the trend model is specified). This comparison is performed in order to check if the quality of the STS model-based estimates produced by frequentist estimation routines is good enough compared to the full Bayesian multilevel framework, as most practitioners, including Statistics Netherlands, estimate STS models in the frequentist framework.

All model specifications within both modelling approaches are made comparable in terms of the same pooling dimension for the stochastic trend dynamics, as well as in that no spatial correlation is assumed across domains. Rationales behind multivariate against univariate modelling dimensions are also addressed. The multilevel time series and STS approaches are compared in terms of adequacy of model-based point-estimates for the trend and for multiple level breaks due to the survey redesigns, as well as in terms of gain in precision that can be reached within these two modelling frameworks. It is well known, though, that variance estimates often get a negative bias in the frequentist STS models due to neglected hyperparame-

ter uncertainty (see, e.g., Pfeffermann and Tiller (2005), Bollineni-Balabay et al. (2015)), so a direct comparison of variance estimates obtained from HB multilevel and frequentist STS models is not completely fair. In this work, we quantify the uncertainty around hyperparameters to make the comparison of the two models fairer.

Another novelty of this paper lies in exploiting the STS approach to refine the multilevel one in the case of unreliable or missing variance estimates of the design estimator (further referred to as design-based variance estimates). These are used as input in multilevel models and are treated as the true known sampling error variances. This paper shows how volatility, missing values, as well as a possible bias in design-based variance estimates can be alleviated.

Section 5.2 provides a brief description of the DTS. Section 5.3 is devoted to the multilevel model setting. It presents the model itself, as well as estimation details for the HB approach in two subsections. Section 5.4 presents the STS model and has the same structure as Section 5.3. Section 5.5 explains how unreliable design-based variance estimates can be improved using the STS technique. Sections 5.6 and 5.7 present empirical results for the DTS at the provincial and national level, respectively. In each of these sections, the multivariate trend structure is examined with regard to possible pooling across one or more factor variables. Further, the performance of the multilevel HB model is compared to that of the STS model. Section 8 contains conclusions and a discussion.

5.2 Dutch Travel survey

The DTS is a stratified two-stage survey that attempts to measure the travel behavior of the Dutch population. Every year, a cross-sectional sample is drawn with sampling units being defined either as households (before 2010), or persons (since 2010). The variable of interest considered in this study is the number of kilometers per person per day (km-pppd) covered per transport modality i and motive j either at the national, or provincial level. The total number of provinces is $P = 12$, and each province p is broken into $J = 7$ motives, and each motive j is broken into $I = 8$ transport modalities (see Table 5.2.1).

Table 5.2.1: Domain classification of the DTS along three factors

Provinces	Mobility motives ("Mot")	Transport modalities ("Mod")
1. Groningen	1. Work	1. Car driver
2. Friesland	2. Business	2. Car passenger
3. Drenthe	3. Visits	3. Train
4. Overijssel	4. Shopping	4. Bus/tram/metro
5. Flevoland	5. Education	5. Scooter/moped
6. Gelderland	6. Recreative	6. Bicycle
7. Utrecht	7. Other motives	7. Walking
8. Noord-Holland		8. Other modalities
9. Zuid-Holland		
10. Zeeland		
11. Noord-Brabant		
12. Limburg		

Design estimates for km-pppd are obtained with the general regression (GREG) estimator (equation (6.4.1) in Särndal et al. (1992)) in the following way: $Y_{(p),j,i,t} = t^Y_{(p),j,i,t}/c_t N_{(p),t}$, where $t^Y_{(p),j,i,t}$ is the GREG-estimator for the total number of kilometers covered in the whole country (or in province p) within a year for a certain transport modality/motive intersection, $N_{(p),t}$ is the population size in the country (for provincial level, in a province p), and c_t is the number of days in a year (365 or 366). Point- and variance GREG-estimates are produced on an annual basis and for different break-downs into domains. The time span considered in this paper is $T = 29$, covering the years 1985-2013. Since 1985, the survey has undergone several redesigns. Until 1994, the population of interest consisted of residents aged 12 years and older. Since 1994, children under 12 years old have also been included in the population of interest. In 1995, the sample became six times as large as before (a net-sample size increase from 10 000 to about 65 000 households). Since 2003-2004, however, net sample sizes have been considerably lower constituting slightly more than 20 000 households (about 43 000 persons between 2010 and 2013). In 1999, the DTS went through the second major redesign that featured some response-motivation and follow-up measures. The published series were corrected for this level break within Statistics Netherlands. In 2004, the data collection for the survey was transferred to another agency. Finally, the survey was redesigned again in 2010 and since then has been conducted by Statistics Netherlands. These two redesigns caused major discontinuities in the time series and have not been corrected for so far.

The model described in the subsequent section will be applied at two aggregation levels of the DTS survey. Firstly, time series of national averages of distances covered, i.e. km-pppd, can be obtained, thus forming an intersection of modalities and motives. Secondly, one can go one level lower where the target variable is defined as time series of provincial averages, i.e. km-pppd at the intersection of all the three above-mentioned factors.

5.3 Multilevel time series approach applied to the DTS

5.3.1 Model specification

Application of multilevel models is common practice in small area estimation (SAE). The model presented in this section is a time series extension of the well-known Fay-Herriot model (Fay and Herriot (1979)) for a cross-section. Similar time series extensions for data contaminated with survey errors have already been mentioned in the introduction to this paper.

In the model below, small areas (domains) are indexed with $m \in \{1, ..., M\}$, and the number of years is T. At the national level, $M = I \times J = 56$; at the provincial level, $M = I \times J \times P = 672$. The GREG-estimates $Y_{(p),j,i,t}$ are contained in a M-dimensional vector $\boldsymbol{Y_t}$ that can be expressed as the sum of signal $\boldsymbol{\theta_t}$ and survey errors contained in vector $\boldsymbol{e_t}$:

$$\boldsymbol{Y_t} = \boldsymbol{\theta_t} + \boldsymbol{e_t}, \quad \boldsymbol{e_t} \overset{id}{\sim} N(\boldsymbol{0}, \boldsymbol{\Phi_t}), \quad t \in \{1, ..., T\}, \tag{5.3.1}$$

where the signal $\boldsymbol{\theta_t}$ consists of the true population parameter and level breaks (discontinuities) that start when the survey faces redesigns and that are constant over time under a particular survey design. Level breaks can be kept or removed from the signal, depending on whether or not a certain survey modification is viewed as an improvement. Survey errors $\boldsymbol{e_t}$ are normally and independently distributed over time with a M-dimensional covariance matrix $\boldsymbol{\Phi_t}$. Although sampled units may occur in several domains at the same time (e.g., in domains "Work-Bicycle" and "Work-Train"), correlations between such domains are likely to be negligible and are therefore assumed to be zero, which renders the $\boldsymbol{\Phi_t}$-matrix diagonal with design-based variances for every domain m. These variances are generally un-

known, so design-based variance estimates $\widehat{Var}(Y_{m,t})$ (Särndal et al. (1992)) are used instead and are treated as the true known variances.

Here and further in this paper, vectors and matrices are printed in bold. Vectors with a subscript t or m are M- or T-dimensional, respectively. No correlation is assumed between different stochastic terms in this paper, unless explicitly mentioned.

The signal development over time is described by a combination of a stochastic and deterministic trend together with level breaks, as mentioned above. Potential cyclical movements in the economy are covered by the stochastic trend, hence the absence of the cyclical component. There is no seasonal component because the data is annual. The following time-series model is assumed for the M-dimensional vector of signals $\boldsymbol{\theta_t}$:

$$\begin{aligned}\boldsymbol{\theta_t} =& c\boldsymbol{\iota} + \kappa t\boldsymbol{\iota} + \delta^{RE}_{1,t}(\beta^0_1\boldsymbol{\iota} + \boldsymbol{\beta^{RE}_1}) + ... + \delta^{RE}_{K_{RE},t}(\beta^0_{K_{RE}}\boldsymbol{\iota} + \boldsymbol{\beta^{RE}_{K_{RE}}}) + \boldsymbol{\nu} + \boldsymbol{\kappa} t + \\ & \delta^{FE}_{1,t}\boldsymbol{\beta^{FE}_1} + ... + \delta^{FE}_{K_{FE},t}\boldsymbol{\beta^{FE}_{K_{FE}}} + \boldsymbol{u_t}, \quad t \,\epsilon\, \{1, ..., T\},\end{aligned} \tag{5.3.2}$$

where c and κ are the overall intercept and linear time trend coefficient, respectively, $\boldsymbol{\iota}$ is an M-dimensional vector of ones, $\boldsymbol{\nu} = (\nu_1, ..., \nu_M)'$ are independent domain effects that serve as random intercepts in the model, $\boldsymbol{\kappa} = (\kappa_1, ..., \kappa_M)'$ are domain-specific random linear time trend coefficients; K_{RE} and K_{FE} denote numbers of level interventions modelled as random or fixed effects, respectively; $\boldsymbol{\beta^{RE}_k}$ and $\boldsymbol{\beta^{FE}_k}$ are vectors with level break coefficients assumed to be random or fixed effects, respectively. Scalars β^0_k represent the mean of the k-th random level break coefficient across domains; $\delta^{RE}_{k,t}$ and $\delta^{FE}_{k,t}$ are indicator variables for level breaks modelled as random and fixed effects, respectively. These indicators switch from zero to one at the moment of a survey redesign. With an entry of a new dummy variable, all previous dummies remain equal to one. Note that scalar terms c, κ, β^0_k are estimated as fixed effects. They are contained in a vector, say $\boldsymbol{\beta}$, along with $\boldsymbol{\beta^{FE}_k}$.

Terms in $\boldsymbol{u_t} = (u_{1,t}, ..., u_{M,t})'$ are domain-by-time effects modelled as integrated random walks. Stationarity tests and a visual inspection of the DTS series suggests a non-stationary model for the trend $\boldsymbol{u_t}$. Therefore, one could assume either a

random walk ($I(1)$), or the local linear trend model, or an integrated random walk ($I(2)$, the so-called smooth-trend model, as in Harvey (2001)). Here, the latter formulation is preferred due to its well-known flexibility and parsimony (Durbin and Koopman (2012), Ch. 3, Harvey (2001)). This application, as well as the one in Bollineni-Balabay et al. (2016a), provides evidence for data overfitting when the other two specifications of the stochastic trends were applied. Therefore, only the smooth-trend model is considered in this paper. The disturbance terms $\boldsymbol{\epsilon_t}$ of the $\boldsymbol{u_t}$-terms are assumed to be normally, identically and independently distributed over time and across domains:

$$\begin{aligned} &\boldsymbol{u_t} = \boldsymbol{u_{t-1}} + \boldsymbol{r_{t-1}}, \\ &\boldsymbol{r_t} = \boldsymbol{r_{t-1}} + \boldsymbol{\epsilon_t}, \quad \boldsymbol{\epsilon_t} \overset{iid}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma_u}), \quad t \,\epsilon\, \{1, ..., T\}, \end{aligned} \tag{5.3.3}$$

where $\boldsymbol{\Sigma_u}$ is a diagonal IJ- or PIJ-dimensional matrix at the national or provincial level, respectively.

The following identifiability constraints are imposed on the stochastic trend components $u_{m,t}$ to insure that, after accounting for the level break interventions, all the deterministic time variation in signal $\theta_{m,t}$ is accounted for by the linear trend $(\kappa + \kappa_m)t$, and the stochastic time variation around this deterministic trend and the remaining time-average level $c + \nu_m$ is accounted for solely by $u_{m,t}$:

$$\sum_{t=1}^{T} u_{m,t} = 0, \quad \sum_{t=1}^{T} u_{m,t} t = 0. \tag{5.3.4}$$

Since there is no reason to assume that the stochastic trends have similar dynamics across either motives, or transport modalities, as will be verified in Section 5.7.1, every diagonal element of a IJ-dimensional square matrix $\boldsymbol{\Sigma_u}$ is assigned a unique value at the national level. As for the provincial level, one could assume that the trend disturbances for motive j and transport modality i from P provinces come from the same distribution. Then matrix $\boldsymbol{\Sigma_u}$ will consist of P block replicas of a IJ-dimensional covariance matrix $\boldsymbol{\Sigma_{[IJ]}}$ for motive-modality intersections:

$$\boldsymbol{\Sigma_u} = \boldsymbol{I_{[P]}} \otimes \boldsymbol{\Sigma_{[IJ]}}. \tag{5.3.5}$$

Whether or not this assumption is feasible, will be verified in Subsection 5.6.1.

The area-specific terms ν_m, κ_m and $\beta^{RE}_{m,k}$ are assumed to share the same variance across domains and to be normally and independently distributed over the domain space and between each other. By construction, they are distributed around zero due to the presence of the overall elements c, κ, and β^0_k:

$$\begin{aligned}
&\boldsymbol{\nu} \overset{iid}{\sim} N(\mathbf{0}, \sigma^2_\nu \boldsymbol{I}),\\
&\boldsymbol{\kappa} \overset{iid}{\sim} N(\mathbf{0}, \sigma^2_\kappa \boldsymbol{I}),\\
&\boldsymbol{\beta^{RE}_k} \overset{iid}{\sim} N(\mathbf{0}, \sigma^2_{\beta^{RE}_k} \boldsymbol{I}),
\end{aligned} \tag{5.3.6}$$

where $\mathbf{0}$ and $\boldsymbol{I}$ denote an M-dimensional vector and identity matrix, respectively. Such model specification allows us to draw terms ν_m, κ_m and $\beta^{RE}_{m,k}$ from distributions centred around zero, which will be necessary in the Bayesian estimation, but it makes these terms, as well as their means weakly identifiable. The sums $(c + \nu_m)$, $(\kappa + \kappa_m)$ and $(\beta^0_k + \beta^{RE}_{m,k})$ are, however, well identifiable and constitute random intercepts, random linear trend slopes and random level break coefficients, respectively. Note that $\boldsymbol{\beta^{RE}_k}$-coefficients can be estimated as fixed effects by setting their variance $\sigma^2_{\beta^{RE}_k}$ equal to infinity. There is always a trade-off between regression coefficients' bias and variance when a random or fixed effects assumption is chosen, see Clark and Linzer (2015) for a discussion.

Model (5.3.2) assumes that area-specific random terms ν_m originate from one distribution (the same applies to κ_m and $\beta^{RE}_{m,k}$). It is, however, quite plausible that differences exist between means and variances of random terms belonging to different motives, transport modalities or provinces. If the pattern of the series for a certain intersection of the motives with transport modalities suggests that provinces do not differ much from each other, the model could be restricted to have only two, instead of three additional terms per random component. Then the remaining random effect variation around the motive and modality means would be described by an M-dimensional random-effect vector. All this gives rise

to the following model, which is an extension of Model (5.3.2):

$$\begin{aligned}\boldsymbol{\theta_t} =& c\boldsymbol{\iota} + \boldsymbol{\iota_{[M/I]}} \otimes \boldsymbol{\nu_{Mod}} + \boldsymbol{\nu_{Mot}} \otimes \boldsymbol{\iota_{[M/J]}} + \boldsymbol{\nu} \\ &+ \kappa t\boldsymbol{\iota} + \boldsymbol{\iota_{[M/I]}} \otimes \boldsymbol{\kappa_{Mod}} t + \boldsymbol{\kappa_{Mot}} t \otimes \boldsymbol{\iota_{[M/J]}} + \boldsymbol{\kappa} t \\ &+ \delta^{RE}_{1,t}(\beta^0_1\boldsymbol{\iota} + \boldsymbol{\iota_{[M/I]}} \otimes \boldsymbol{\beta^{RE}_{1,Mod}} + \boldsymbol{\beta^{RE}_{1,Mot}} \otimes \boldsymbol{\iota_{[M/J]}} + \boldsymbol{\beta^{RE}_1}) + \\ &... \\ &+ \delta^{RE}_{K_{RE},t}(\beta^0_{K_{RE}}\boldsymbol{\iota} + \boldsymbol{\iota_{[M/I]}} \otimes \boldsymbol{\beta^{RE}_{K_{RE},Mod}} + \boldsymbol{\beta^{RE}_{K_{RE},Mot}} \otimes \boldsymbol{\iota_{[M/J]}} + \boldsymbol{\beta^{RE}_{K_{RE}}}) + \\ &+ \delta^{FE}_{1,t}\boldsymbol{\beta^{FE}_1} + ... + \delta^{FE}_{K_{FE},t}\boldsymbol{\beta^{FE}_{K_{FE}}} + \boldsymbol{u_t}, \quad t \,\epsilon\, \{1, ..., T\},\end{aligned} \tag{5.3.7}$$

where each of the ν, κ, β^{RE}-terms sub-indexed with Mod contains random term expectations taken across motives and provinces for each of I transport modalities and is distributed as

$$\begin{aligned}\boldsymbol{\nu_{Mod}} &\overset{iid}{\sim} N(\mathbf{0}, \sigma^2_{\nu_{Mod}}\boldsymbol{I}), \\ \boldsymbol{\kappa_{Mod}} &\overset{iid}{\sim} N(\mathbf{0}, \sigma^2_{\kappa_{Mod}}\boldsymbol{I}), \\ \boldsymbol{\beta^{RE}_{Mod}} &\overset{iid}{\sim} N(\mathbf{0}, \sigma^2_{\beta^{RE}_{Mod}}\boldsymbol{I}).\end{aligned} \tag{5.3.8}$$

The same applies to random terms labelled with Mot. These are J-dimensional. The random terms $\boldsymbol{\nu, \kappa, \beta^{RE}_k}$ are M-dimensional and take care of the variation around their respective motive and modality means.

Note that the signal $\theta_{m,t}$ has no area-by-time white noise in the multilevel models presented in this paper. It is assumed that all the stochastic variation in $\theta_{m,t}$ over time is picked up by the stochastic structure of the trend. In fact, allowing for an additional noise in the signal resulted in data overfit in many, especially small domains. The absence of the additional noise term makes it furthermore easier to compare the multilevel approach with the structural time series one.

Special cases of models (5.3.2) and (5.3.7) could also be considered. For instance, all level breaks could be modelled as fixed effects. Further, if distribution assumptions about the random terms ν_m and κ_m do not seem to hold, one could consider modelling these terms as fixed effects. In the latter case, estimation at the national level will be reduced to a univariate setting.

5.3.2 Estimation details on the multilevel model

The multilevel modelling technique constitutes a composite estimator, i.e. a combination of the design and synthetic estimators. The model in (5.3.1)-(5.3.2) can be estimated with the EBLUP (empirical best unbiased predictor), or within the hierarchical Bayesian (HB) approach. For this application, the HB approach with non-informative priors is chosen that relies on the Gibbs-sampler pre-programmed in the *mcmcsae* R-package (Boonstra (2015)). The posterior means of the signals $\theta_{m,t}$ are taken to be domain point-estimates, and the posterior standard deviations of $\theta_{m,t}$s serve as their measures of uncertainty. In this paper, these posterior standard deviations are referred to as SEs for brevity.

Random terms ν_m, κ_m and $\beta^{RE}_{m,k}$ are independent and assigned normal priors as in (5.3.6) within the Bayesian estimation framework. The random walk term has a normal prior distribution as well: $\boldsymbol{u_m} \overset{iid}{\sim} N(\mathbf{0}, \sigma^2_{u,m}\boldsymbol{A})$, $\boldsymbol{A}$ being a T-dimensional covariance matrix. In case when the initial u_{m1} in all domains are assigned a non-informative prior, matrix $\boldsymbol{A}$ becomes unbounded. However, the limit of the precision matrix $\boldsymbol{A^{-1}}$ used in the Gibbs sampler is well-defined, see, e.g., Rue and Held (2005). Apart from that, the precision matrix $\boldsymbol{A^{-1}}$ is sparse (i.e., mostly contains zeros and, in this case, has five diagonals), which makes the Gibbs sampler efficient (Rue and Held (2005)). The variances of the random-effect components are assigned inverse chi-squared priors with degrees of freedom v and a scale parameter s^2: $\sigma^2_\nu \sim Inv-\chi^2(v_\nu, s^2_\nu)$, $\sigma^2_\kappa \sim Inv-\chi^2(v_\kappa, s^2_\kappa)$, $\sigma^2_{\beta^{RE}} \sim Inv-\chi^2(v_{\beta_{RE}}, s^2_{\beta^{RE}})$, $\sigma^2_{u,m} \sim Inv-\chi^2(v_{u,m}, s^2_{u,m})$. Vague priors on these variance parameters would require small values (close to zero) for the hyperparameter v.

If variances of the stochastic terms ν_m, κ_m etc. are small (relative to the design-based variances), the convergence of the Gibbs sampler could be very slow. Apart from that, as Gelman et al. (2008) and Polson et al. (2012) suggest, the inverse-gamma (or inverse chi-squared) parametrization for variance parameters is often not robust and should be replaced by distributions for standard deviation parameters from the folded noncentral t-family, of which the half-Cauchy distribution is a special case. In order to solve both problems, Gelman et al. (2008) suggest applying a redundant multiplicative parametrisation. In this paper, this parameter expansion is applied to any stochastic term components for which non-informative priors are chosen: $\nu_m = \xi_\nu \tilde{\nu}_m, \kappa_m = \xi_\kappa \tilde{\kappa}_m$, etc., where ξ-terms are multiplicative scalar parameters, and the terms with a tilde are distributed like in (5.3.6) and

(5.3.3), but with variances $\tilde{\sigma}^2_\nu, \tilde{\sigma}^2_\kappa$, etc., for which inverse-Gamma (or inverse chi-squared) priors can be chosen. Such parametrization is used for standard deviation parameters that are expressed as: $\sigma_\nu = |\xi_\nu|\tilde{\sigma}_\nu, \sigma_\kappa = |\xi_\kappa|\tilde{\sigma}_\kappa$, etc. The ξ-terms are independent of each other and are assigned normal priors:

$$\xi_\nu \sim N(\alpha_\nu, \gamma_\nu), \quad \xi_\kappa \sim N(\alpha_\kappa, \gamma_\kappa), \quad \xi_{\beta^{RE}} \sim N(\alpha_{\beta^{RE}}, \gamma_{\beta^{RE}}), \quad \xi_u \sim N(\alpha_u, \gamma_u). \tag{5.3.9}$$

Setting $\alpha = 0$ (and $\gamma = 1$, without loss of generality) implies that priors on the standard deviation parameters come from the half t-family, see Gelman (2006). Setting $\alpha = 1$ and $\gamma = 0$ is equivalent to the original (non-expanded) parametrization of the model. Combining $\xi \sim N(0, 1)$ with 1 degree of freedom in $\tilde{\sigma}^2 \sim Inv - \chi^2(1, s^2)$ results in a half-Cauchy prior for parameter σ. For numerical reasons, the scale parameters s^2 in this application are restricted to the standard deviation of the variable of interest in vector $\boldsymbol{Y}$.

The overall intercept c, linear trend coefficient κ and expectations β^0_k of random level break coefficients are estimated as regression coefficients and are contained in $\boldsymbol{\beta}$, along with vectors $\boldsymbol{\beta^{FE}_k}$. The prior for all the regression coefficients is vague (normal with mean $\boldsymbol{\beta_0 = 0}$ and a large variance $\boldsymbol{\Omega_{\beta_0}}$).

Denoting the parameter vector by $\boldsymbol{\psi}$:
$\boldsymbol{\psi} = (\boldsymbol{\nu', \kappa', \beta^{RE'}, u', \beta'}, \sigma^2_\nu, \sigma^2_\kappa, \sigma^2_{\beta^{RE}}, \sigma^2_u, \xi_\nu, \xi_\kappa, \xi_{\beta^{RE}}, \xi_u)'$,
the likelihood function can be written as:

$$p(\boldsymbol{Y}|\boldsymbol{\psi}) = N_Y(\xi_\nu \tilde{\boldsymbol{\nu}} \otimes \boldsymbol{\iota_{[T]}} + \xi_\kappa \tilde{\boldsymbol{\kappa}} \otimes \boldsymbol{t} + \boldsymbol{X^{RE}\beta^{RE}} + \boldsymbol{X\beta} + \xi_u \tilde{\boldsymbol{u}}, \boldsymbol{\Phi}), \tag{5.3.10}$$

where $\boldsymbol{t}$ denotes a vertical vector with time indicators $(0, 1, ..., T-1)'$, $\boldsymbol{\iota_{[T]}}$ is a T-dimensional column vector of ones, $\boldsymbol{\Phi}$ is a $[MT \times MT]$ matrix, matrix $\boldsymbol{X^{RE}}$ is $[MT \times MK_{RE}]$ and contains dummy regressors for the vector with random level break effects $\boldsymbol{\beta^{RE}} = (\boldsymbol{\beta^{RE'}_1}, ..., \boldsymbol{\beta^{RE'}_{K_{RE}}})'$, and $\boldsymbol{X}$ is a $[MT \times (2 + K_{RE} + d^{FE})]$-dimensional matrix, d^{FE} being the number of level break coefficients modelled as fixed effects. The dimension of matrix $\boldsymbol{X}$ is such due to the presence of the overall effects c, κ, and K_{RE} β^0_ks in the vector $\boldsymbol{\beta}$, as well as due to the presence of the d^{FE}-dimensional vector $\boldsymbol{\beta^{FE}}$ of fixed level break coefficients.

The parameters in the prior distribution below are assumed to be mutually independently distributed. The joint prior is then a product of each parameter's marginal prior distribution:

$$p(\boldsymbol{\psi}) = \left[\prod_{l=\nu,\kappa,\beta^{RE}} \left[\prod_{m=1}^{M} N_l(0,\tilde{\sigma}_l^2)\right]\right]\left[\prod_{m=1}^{M} N_u(\mathbf{0},\tilde{\sigma}_{u,m}^2\boldsymbol{A})\right]\times \left[\prod_{\substack{l=\nu,\kappa,\\ \beta^{RE},u}} Inv-\chi^2_{\sigma_l^2}(v_l,s_l^2)\right]\left[\prod_{\substack{l=\nu,\kappa,\\ \beta^{RE},u}} N_{\xi_l}(\alpha_l,\gamma_l)\right] N_\beta(\boldsymbol{\beta_0},\boldsymbol{\Omega_{\beta_0}}). \tag{5.3.11}$$

Then the posterior density of the parameter vector $\boldsymbol{\psi}$ is proportional to the following joint density: $p(\boldsymbol{\psi}|\boldsymbol{Y}) \propto p(\boldsymbol{\psi})p(\boldsymbol{Y}|\boldsymbol{\psi})$. See Appendix 5.A for each parameter's conditional posterior distribution used in the Gibbs-sampler.

5.4 Structural time series (STS) unobserved component modelling in the case of the DTS

5.4.1 STS model specification

The general theory on STS models is presented in Durbin and Koopman (2012) and Harvey (1989). Similarly to the multilevel framework presented in Section 5.3.1, the series of the DTS design estimates $Y_{m,t}$ in a STS model can be decomposed into signal $\theta_{m,t}$ and the survey error component, as in (5.3.1). The $\theta_{m,t}$-term, in turn, is decomposed into several unobserved components :

$$\theta_{m,t} = L_{m,t} + \delta_{1,t}\beta_{m,1} + ... + \delta_{K,t}\beta_{m,K} + \varepsilon_{m,t}, \quad t \,\epsilon\, \{1,...,T\}, \tag{5.4.1}$$

where $L_{m,t}$ is the stochastic trend component, K is the total number of level breaks in domain m, and $\delta_{k,t}$ is an indicator variable for the k-th level-break regression coefficient $\beta_{m,k}$. The population parameter error term is $\varepsilon_{m,t} \overset{iid}{\sim} N(0,\sigma_\varepsilon^2)$. In cross-sectional surveys like the DTS, it is difficult to separate this term from the sampling error $e_{m,t}$, especially if the variance of $\varepsilon_{m,t}$ is small relatively to the sampling variance. Therefore, the two terms are combined into one composite error term $v_{m,t}$ that is assumed to be largely dominated by the sampling error. In order to incorporate the design-based variance estimates $\widehat{Var}(Y_{m,t})$ in a STS model, the composite error term $v_{m,t}$ can be modelled as $\tilde{e}_{m,t}\sqrt{\widehat{Var}(Y_{m,t})}, \tilde{e}_{m,t} \overset{iid}{\sim} N(0,\sigma^2_{\tilde{e}_m})$.

For the variance of this product to be close to the design-based variance estimate $\widehat{Var}(Y_{m,t})$, $\sigma^2_{\tilde{e}_m}$ should be close to unity. Deviations from unity should correct for a possible under- or overestimation of design-based variance estimates. In this way, STS models feature more flexibility, unlike the multilevel model presented in Subsection 5.3.1, where the survey error variance estimates, used as prior input into the model, are assumed to be true and are thus fixed.

For one domain m, the model in (5.4.1) is referred to as a univariate STS model. If several domains have to be estimated simultaneously so that they can borrow cross-sectional sample information from each other, univariate models can be stacked under each other to constitute a multivariate STS model. In this application, no spatial correlation is assumed, just like for the multilevel model presented in the previous section.

The obtained multivariate STS model largely resembles the multilevel one in (5.3.1), with each domain being represented by the following equation:

$$Y_{m,t} = L_{m,t} + \delta_{1,t}\beta_{m,1} + ... + \delta_{K,t}\beta_{m,K} + \tilde{e}_{m,t}\sqrt{\widehat{Var}(Y_{m,t})}, \quad t \in \{1, ..., T\}. \quad (5.4.2)$$

As mentioned in Section 5.3.1, the smooth-trend model is assumed for the trend in the multilevel setting. Here, the local linear trend model and a random walk with a drift both resulted in a severe data overfit (see Harvey (2001) or Durbin and Koopman (2012) for different trend model specifications). Therefore, the same trend model is assumed for the structural time series (STS) model setting as in (5.3.3), though without the identification constraints for the multilevel setting mentioned in (5.3.4). The trend component $L_{m,t}$, apart from the random-walk component $u_{m,t}$ with its deterministic part $t(\kappa + \kappa_m)$, also implicitly contains an intercept. However, this intercept is defined solely on the basis of the corresponding domain's input series $Y_{m,t}$ and therefore would be equivalent to intercept $c + \nu_m$ in (5.3.2) only if the latter is modelled as a fixed, rather than random effect. The covariance matrix of the disturbance terms belonging to the trend $L_{m,t}$ is diagonal as in (5.3.5).

5.4.2 STS model estimation details

Linear structural time series models with unobserved components are usually put into a state-space form and analysed with the Kalman filter. First, the model hyperparameters (here, $\sigma^2_{u,m}$s and $\sigma^2_{\tilde{e}_m}$s) are estimated using the maximum-likelihood (ML) approach by iteratively running the Kalman filter. The hyperparameters are set equal to their ML- estimates and are treated as known, whereafter state variables ($L_{m,t}, \beta_{m,k}$s) can be extracted by the Kalman filter recursions. These recursions are initialised with diffuse priors for non-stationary variables (see Koopman (1997)). One has to be aware of the fact, that the mean square error (MSE) estimates of the state variables produced by the Kalman filter are negatively biased, since the uncertainty around the hyperparameter estimates is not taken into account. See Pfeffermann and Tiller (2005) for the true MSE estimation approaches, as well as Bollineni-Balabay et al. (2015) for a simulation-based comparison of different approaches existing in the literature. The STS models presented in this paper are estimated in OxMetrics 7 (Doornik (2007)) in combination with the *SsfPack 3.0* package (Koopman et al. (2008)). One could think of a full Bayesian approach to STS model estimation with prior distributions for the model hyperparameters instead of hyperparameter ML estimates. However, the computational capacity required for sequential updating in models like the DTS one would make this approach very challenging. Apart from that, since most practitioners, including Statistics Netherlands, estimate STS models in the frequentist framework, the authors aim to find whether the quality of frequentist STS model-based estimates is sufficiently good compared to that when estimation is carried out within the full Bayesian multilevel framework.

5.5 Tackling unreliability and missing values in design-based variance estimates of the DTS

Usually, unreliability in design-based variance estimates are mitigated by pooling them across similar domains (in this application, similarity is only observed across provinces, as will be clear from the next section). In time series, alternatively, this could be done with the help of the Kalman filter.

A problem with the DTS is that the design-based variance estimates are missing in 2004-2009. For 2010-2013, they are only available at the intersections of provinces

with modalities and of modalities with motives at the national level. The missingness in all domains in 2004-2009 can be imputed with the help of the Kalman filter, as shown a bit later in this subsection. As for the variances missing only at certain intersections at the end of the sample (2010-2013), these may be approximated by using the variances available at the other intersections. The following approach is applied in order to approximate the standard errors ($SE(Y_{m,t})$) at the three-dimensional intersection. First, a factor is defined that should reflect an inflation of the design-based variance in modality Mod_i when switching from the national to the provincial level:

$$F_{Mod_i,t} = SE(Y_{Mod_i,p,t})/SE(Y_{Mod_i,t}).$$

This factor is used to inflate the national level standard error of the GREG estimator (further referred to as the design-based standard error) to the provincial level for a certain intersection of motives with modalities:

$$SE(Y_{Mod_i,Mot_j,p,t}) = SE(Y_{Mod_i,Mot_j,t}) \cdot F_{Mod_i,t}. \tag{5.5.1}$$

The performance of this method is verified using the 2003 data and is strikingly good for every province for the domain intersection in the left panel of Fig. 5.5.1 (here, relative margins of error are plotted, i.e. the margins of error at the 95% confidence level divided by the point-estimate). In most other domain intersections, the approximation performs similarly. The right panel shows how the approximation performs in the worst case.

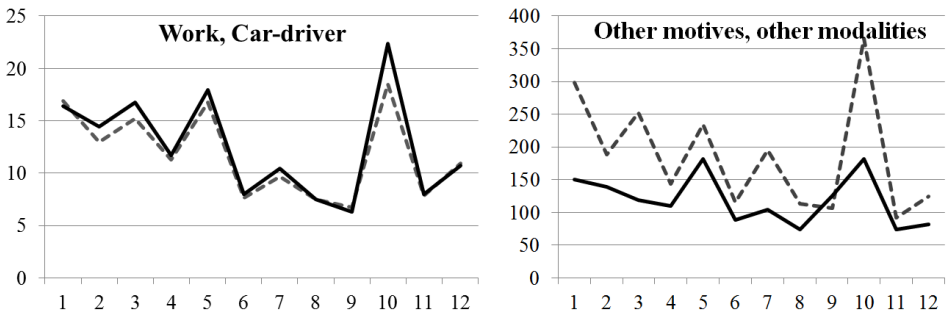


Figure 5.5.1: Original (solid line) and approximated (dashed line) relative margins of error at the 95% confidence level for 12 provinces, 2003, percent

The Kalman filter (KF) is applied to impute the missing design-based standard error estimates and to smooth out sampling fluctuations. Therefore, *smoothed* trends of the design-based standard error estimates from the following univariate STS model are used as input information in multilevel and STS models for the variable of interest, i.e. km-pppd:

$$\begin{aligned} SE(Y_{m,t}) &= L^{SE}_{m,t} + \varepsilon_{m,t}, \quad \varepsilon_{m,t} \overset{iid}{\sim} N(0, \sigma^2_{\varepsilon,m}), \\ L^{SE}_{m,t} &= L^{SE}_{m,t-1} + R^{SE}_{m,t-1}, \\ R^{SE}_{m,t} &= R^{SE}_{m,t-1} + \eta^{SE}_{m,t}, \quad \eta^{SE}_{m,t} \overset{iid}{\sim} N(0, \sigma^2_{\eta^{SE}_m}), \quad t \in \{1, ..., T\}, \end{aligned} \tag{5.5.2}$$

where $L^{SE}_{m,t}$ and $R^{SE}_{m,t}$ are the level and slope of the trend for the design-based standard errors, respectively. Some of these smoothed SE estimates are plotted against their original counterparts in Fig. 5.5.2.

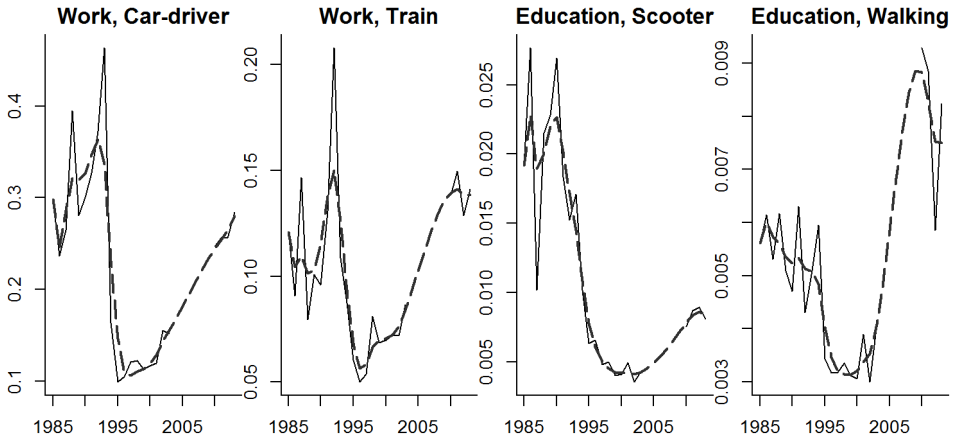


Figure 5.5.2: Original (solid line) and smoothed (dashed line) design-based standard errors, Zuid-Holland.

If these smoothed estimates are used as an input in multilevel models (and thus treated as the known true standard errors of the design estimator), some domains still experience too volatile multilevel model point-estimates (see Fig.5.5.3, solid lines). This occurs only in domains where the number of km-pppd is small (further referred to as small-scale domains, e.g., "Walking", "Bus/Tram/Metro") and makes one suspect that the design-based variance estimates are biased in these domains. For this reason, one could consider using the smoothed design-based standard errors scaled within the univariate STS framework. This means that the

sampling error variances would be represented both in the STS and multilevel models by the product of the scaling parameter $\sigma^2_{\tilde{e}_m}$ and the smoothed design-based variance estimates $\widehat{Var}(Y_{m,t})$. Fig. 5.5.3 shows point-estimates resulting from the two approaches applied to the same multilevel model. The differences in other domains are negligible, because most of these scaling parameters are concentrated around unity, as Fig. 5.5.4 shows. Further in this paper, all the multilevel analysis will be based on the second approach which applies scaled smoothed design-based variance estimates. Not only does it prevent overfitting in domains whose design-based variance estimates are not reliable, but it also eliminates another factor responsible for differences between the outcomes from multilevel and STS models, because in this way both models use design-based variance estimates corrected in the same way.

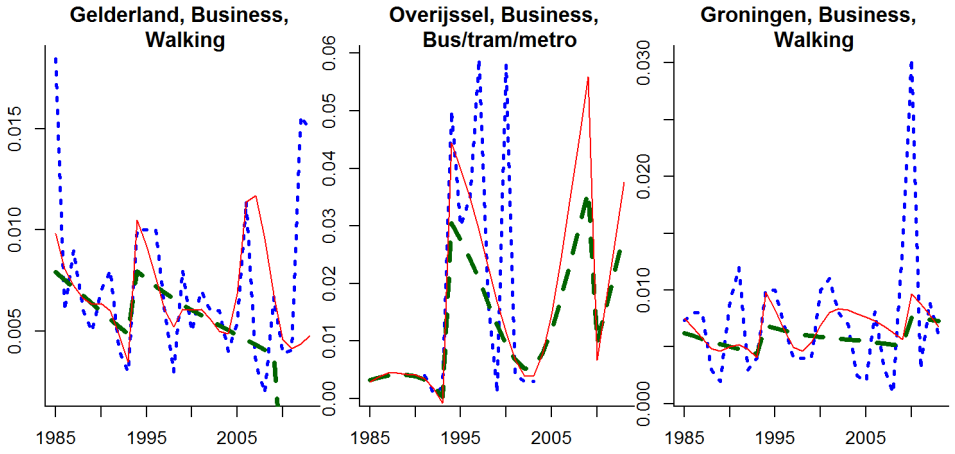


Figure 5.5.3: Design-based point-estimates (blue dotted lines) and multilevel model posterior means that use smoothed design SE estimates (red solid line) and smoothed design SE estimates scaled within the univariate STS framework (green dashed line).

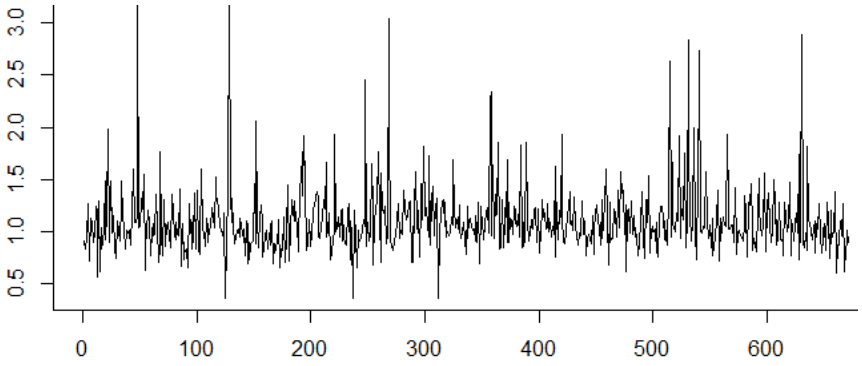


Figure 5.5.4: Scaling factor for design-based standard errors in univariate STS models.

5.6 The DTS at the provincial level

Before we estimate both models, a careful inspection for outliers is needed. The trend disturbance variance $\sigma^2_{u,m}$ may get overestimated due to one single outlier, especially if the design-based variance of this outlying point is negatively biased. Excessively large trend disturbance variances would cause model point-estimates to overfit the data. In order to eliminate the effect of outliers, we inspect estimation results from univariate STS models for each domain and assign a large value to the design-based variance of every outlying point. There are five outliers in total at the provincial level, which belong to small-scale domains involving "Scooter", "Train" or "Other modalities". These outliers are treated in the same way in the multilevel and STS models.

5.6.1 The multivariate variance structure of stochastic trends

Small areas (domains) at the provincial level are formed by an intersection of 12 provinces with 8 modalities and 7 motives ($M = 672$). Subsection 5.3.1 presented a general variance structure of the trend disturbances. To see if some domains with certain irregularities in the data can benefit from other domains by means of pooling by either provinces, motives or modalities, it is best to compare the trend disturbance variance estimates $\hat{\sigma}_{u,p,j,i}$ obtained from univariate STS models (5.4.2). By looking at the top panel of Fig. 5.6.1 with box-plots for these $\sigma_{u,p,j,i}$-estimates, it becomes clear that they are very heterogeneously distributed across

the modalities. With heterogeneity less pronounced across the motives (the lower panel), there is no systematic way in which some motives resemble the others, and the existing differences are still too big for the trend disturbance variances to be pooled across the motives.

Viewing provinces as panel units with responses on $\sigma_{u,p,j,i}$ for motives and modalities provides an opportunity to use larger amounts of sample information to estimate the hyperparameters. This would make the hyperparameters estimates more reliable if the processes in provinces are similar indeed. In order to check that, one could construct scatter- or barplots of $\sigma_{u,p,j,i}$-estimates from univariate STS models for the twelve provinces for every intersection of the motives and transport modalities (find the scatterplot and barplots in Fig. 5.6.2 and 5.B.1, respectively). These plots reveal that $\sigma_{u,p,j,i}$-estimates of provinces 2 (Friesland) and 5 (Flevoland) often exhibit very different values from the rest of the provinces - at eight and seventeen Mot/Mod intersections, respectively. Large $\sigma_{u,p,j,i}$-values in province 5 (e.g., in Mod 1/Mot 1-3,6, Mod 2/Mot 3, Mod 4/Mot 1) are mostly caused by a bigger scale of the series, i.e. with more km-pppd being covered in this province compared to the other provinces. As for Province 2 (e.g., in Mod 2/Mot 1, Mod 5/Mot 1, Mod 8/Mot 6), large $\sigma_{u,p,j,i}$-values are caused by either a bigger scale, or a more volatile pattern of the series. $\sigma_{u,p,j,i}$-estimates of province 1 exhibit the largest values (compared to the rest of the provinces) for about five times, and those of province 10 and 12 - four and three times, respectively.

It is worthwhile to observe the scale of the latter provinces' series (1, 10 and 12): it hardly ever exceeds the scale of the rest of the provinces, and is sometimes even smaller. Therefore, these provinces would be better off if pooled together with the other ones. Further, one can see that the trend disturbance variances in many cases tend to take on very small values when estimated within the STS approach. It means that the trend in such cases resembles a straight line. Keeping in mind that the ML estimator tends to underestimate hyperparameters when their distribution is right-skewed, it would also be desirable to estimate such variance hyperparameters in a pool with other domains (provinces in this case). Thus, we abstain from pooling provinces 2 and 5 with the other ten. Therefore, the twelve provinces are divided into a cluster of two and a cluster of ten provinces, within which variances are pooled. This applies to both the multilevel and STS analysis.

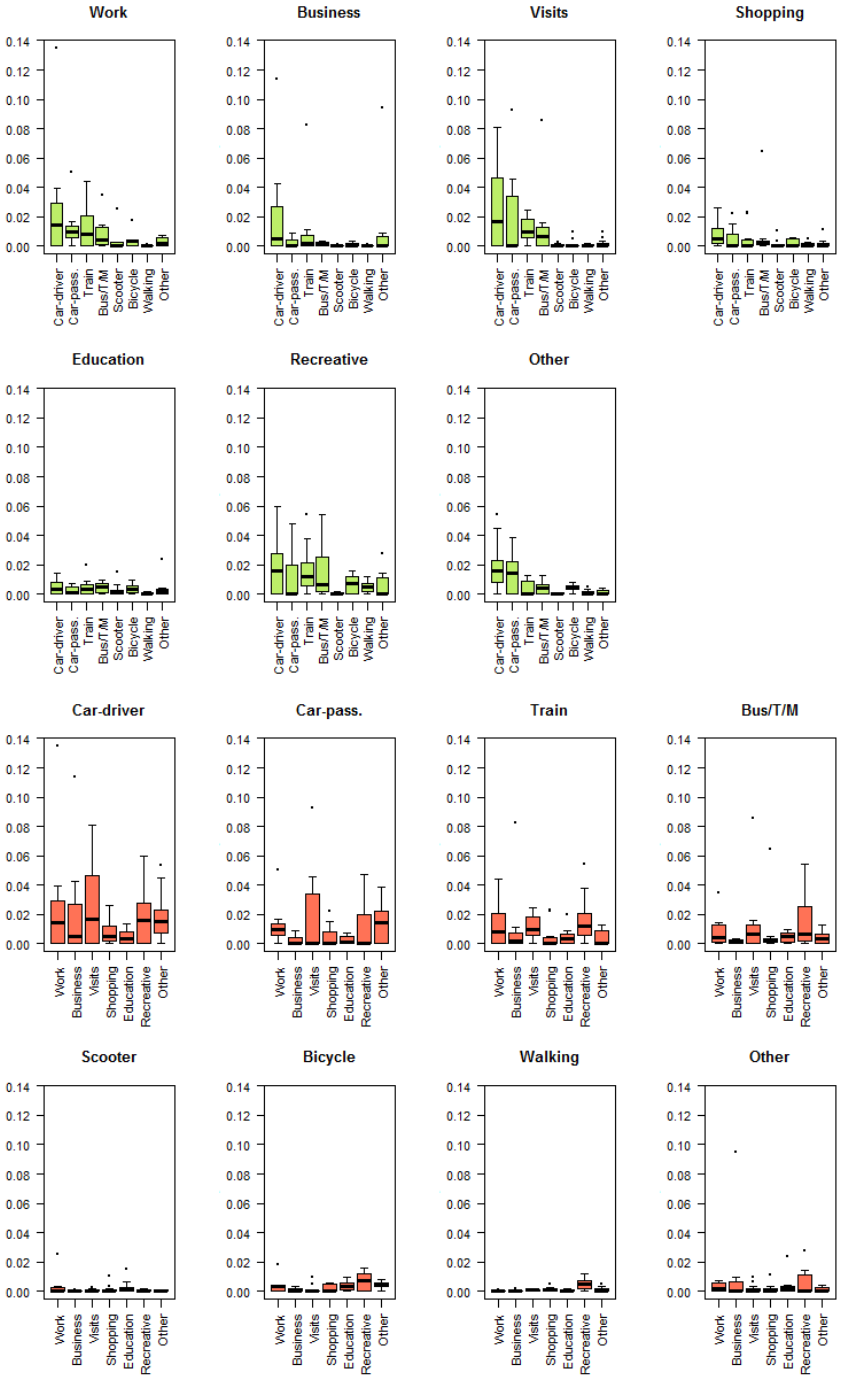


Figure 5.6.1: Box-plots for provincial σ_us from univariate STS models (separate dots correspond to different provinces): modalities within motives (upper panel) and motives within modalities (lower panel)

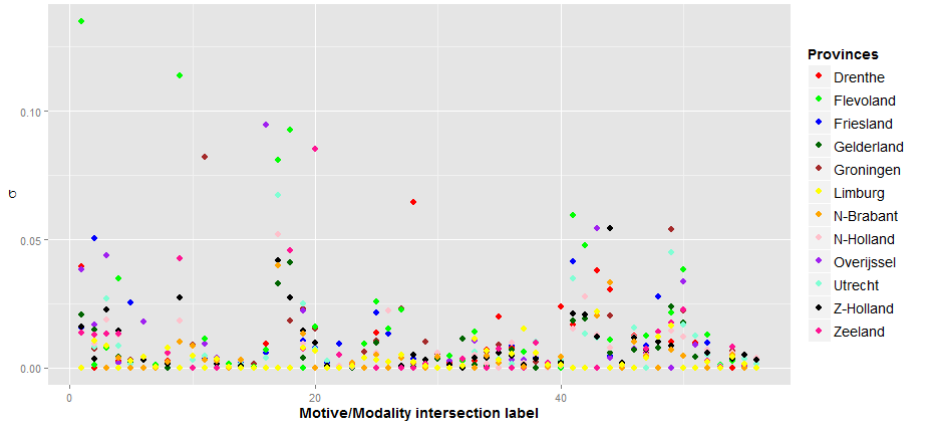


Figure 5.6.2: Provincial σ_u from univariate STS models for all 56 intersections of motives and transport modalities.

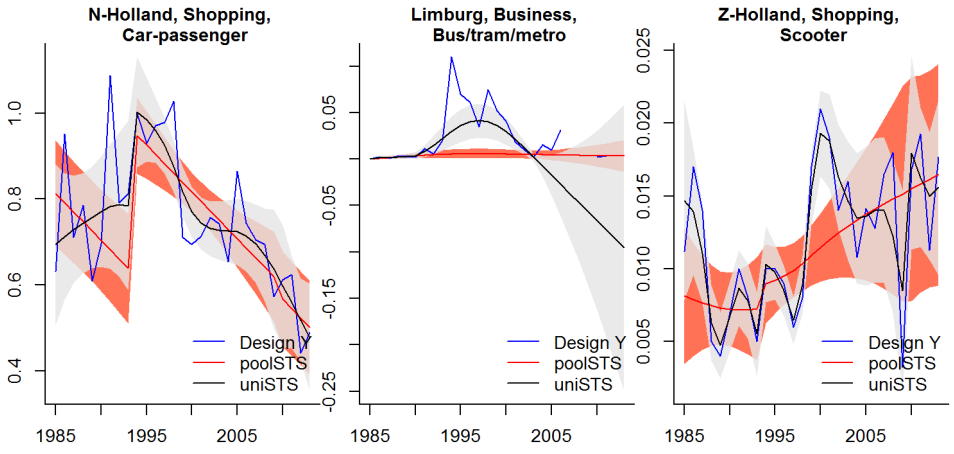


Figure 5.6.3: STS model-based point-estimates obtained from univariate STS models (black line) and from a multivariate STS model where σ_u^2s are pooled across provinces (coral line); the colour of the confidence intervals corresponds to the colour of the line; the design-based point-estimates in blue.

The DTS point-estimates are very similar when comparing the univariate and multivariate STS settings with each other. There are, however, a few domains that exhibit significant differences (depicted in Fig. 5.6.3). It is worth considering the consequences that pooling may have on model estimates. If one province turns out to be very different from the others it has been pooled with, then an underestimate of the stochastic trend variance can force the model-based point-estimates to

take strange slopes, for example, when κ_ms are not sufficiently scattered due to a small σ^2_κ (as, e.g., in Noord-Holland/Shopping/Car-passenger). The advantage of pooling $\sigma_{u,m}$ across provinces is that certain idiosyncracies occurring only in one province can be eliminated. If there is, for instance, no real factor behind the sudden surge in point-estimates of the Limburg/Business/Bus-Tram-Metro domain, pooling could be considered in order to get rid of such irregularities. Further, excessive volatility in the univariate model point-estimates of one province due to underestimation of design-based variances in this province can be overcome by borrowing information from other provinces for an identical intersection of motives and transport modalities (as in Zuid-Holland/Shopping/Scooter).

5.6.2 Multilevel model estimation results

Several model formulations have been considered for the multilevel framework. These formulations have already been mentioned at the end of Section 5.3.1 and are summarised in Table 5.6.1. Model selection is based on the deviance information criterion (DIC) (see Spiegelhalter et al. (2002)) which is a generalisation of the well-known AIC for hierarchical models. Adequacy of point- and variance estimates is also taken into account as an informal criterion. Table 5.6.1 also contains information on the minimum sufficient number of iterations, burn-in iterations, as well as the thinning interval h. The latter means that only every h-th draw from the parameter posterior distribution is saved in the MCMC chain. Thinning helps overcome memory constraints when a chain is strongly autocorrelated (see Gelfand and Smith (1990)). Further, in order to help diagnose whether the stationary distribution of a parameter has been reached, more than one chain is needed, each chain starting with draws from an overdispersed distribution (Gelman and Rubin (1992)). Three parallel chains are constructed in this application for each parameter. We use the Gelman-Rubin convergence statistic for multiple chains (R-hat) to be able to judge about the sufficient length of a burn-in period and about whether the chains have mixed well, which allows us to arrive at a reasonable estimate of the posterior distribution.

HB-bRE stands for a model with random effects as in (5.3.1)-(5.3.2). HB-bFE differs from HB-bRE in that all level break coefficients are estimated as fixed effects. HB-FE differs from HB-bFE in that ν_m and κ_m terms are estimated as fixed effects too. Different model modifications of the kind in (5.3.7) have been tried

Table 5.6.1: Multilevel models considered for the DTS at the provincial level

Model labels	Year and domain indication for level breaks	Iterations	Burn-in	Thinning	DIC
HB-bRE	β^{RE}: 1994, 2010 for all m; β^{FE}: 2004 for m in Mot 6,7	75000	50000	50	-43668
HB-bFE	β^{FE}: 1994, 2010 for all m; 2004 for m in Mot 6,7	55000	30000	50	-43606
HB-FE	β^{FE}: 1994, 2010 for all m; 2004 for m in Mot 6,7; $(c+\nu_m)$, $(\kappa+\kappa_m)$ as FE	5000	1000	10	-43859

for HB-bRE and HB-bFE models. However, these modifications either resulted in a numerical failure (in model HB-bRE variations), or failed to secure multiple chain convergence in some variance parameters ($\sigma^2_{\kappa_{Mod}}, \sigma^2_{\nu_{Mot}}$) in model HB-bFE variations. The only model modification that could be fitted successfully was the HB-bFE model with an additional $\boldsymbol{\nu_{Mod}}$-term. However, this specification did not result in model improvement, with the DIC value being even higher than that of the basic HB-bFE model.

Fig. 5.6.4 shows point-estimates for the three multilevel models mentioned in Table 5.6.1. While most domains get comparable point-estimates from the three models, a combination of "Car-driver" with "Work" and "Business" motives features some most striking problems inherent to models HB-bRE and HB-bFE. It is worth mentioning that the 1994-break is the largest exactly at the intersection of modality "Car-driver" with motive "Busines". Only five provinces are depicted, but the results are similar for the rest of the provinces. As the second row of Fig. 5.6.4 demonstrates, the model with level breaks modelled as random effects is incapable of fitting the level breaks for the year 1994. Neither of the HB-bRE modifications can cure this problem (some of them even result in less adequate level break estimates for 1994). The level break for the years 2010-2013 is not captured by the HB-bRE model in these domains either. Apparently, the assumptions about common normal distributions for β^{RE}_ms are violated in the case of the DTS. The first row of Fig. 5.6.4 suggests that the assumptions about common normal distributions for κ_ms and/or for ν_ms are often not applicable either.

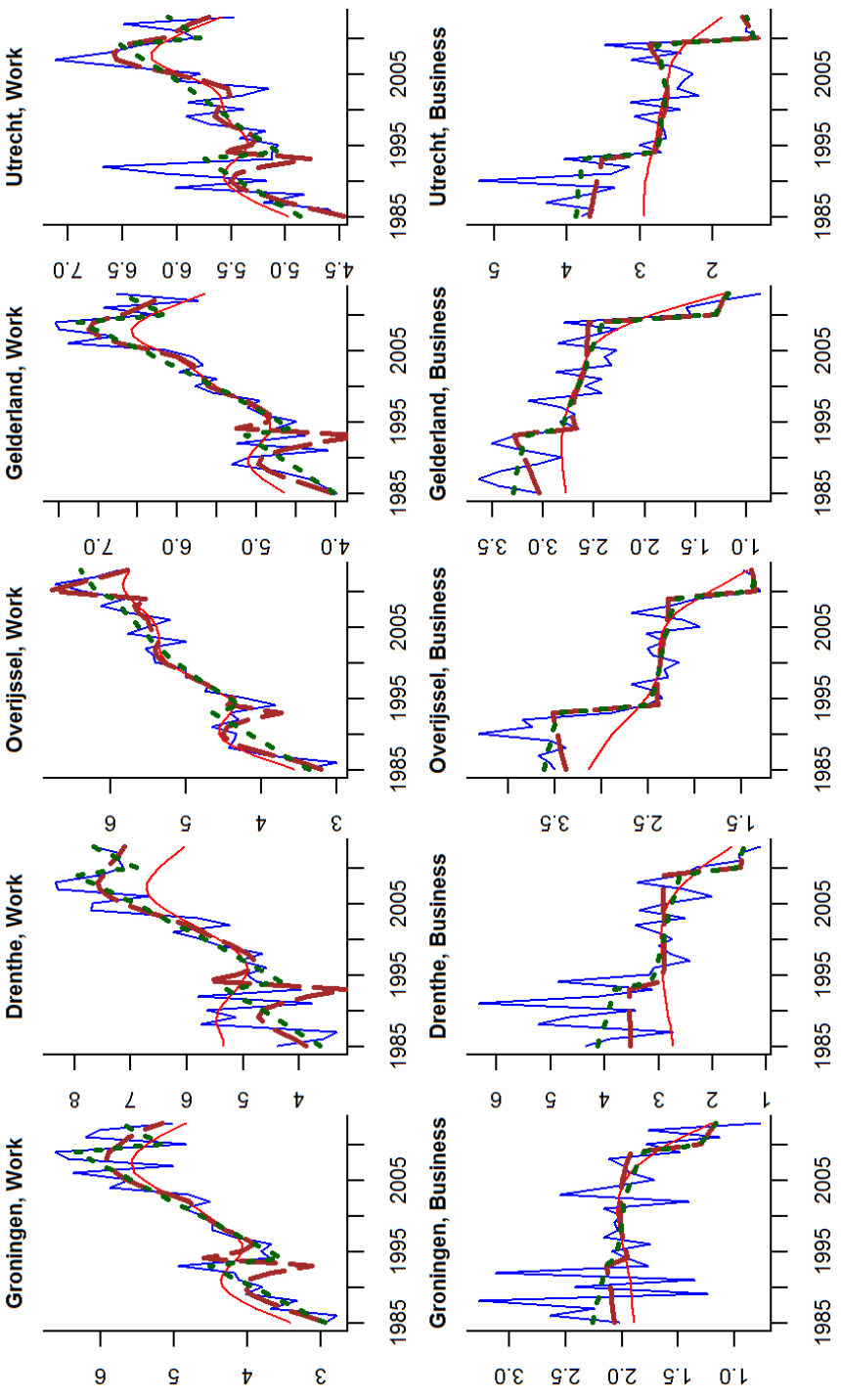


Figure 5.6.4: Mod_1 (Car-passenger) point-estimates produced by the design-based estimator (blue solid line) and HB-bRE (red solid line), HB-bFE (brown dashed line), HB-FE (green dotted line) models

The model HB-bFE exhibits a greatly improved fit. However, the fit for the years 1985-1994 is not optimal for the "Work"-motive (see the first row). Apart from that, the level of the signal in "Business" seems to be too low (see the second row of Fig. 5.6.4), suggesting that the common normal distributions imposed on the random intercepts do not allow the ν_m-terms to take sufficiently large values. Finally, the fit of the model with only fixed effects (HB-FE) seems to be most adequate. This model also gets the lowest DIC-value, see Table 5.6.1.

5.6.3 Multilevel and STS model estimation results compared

Not only does the multilevel model HB-FE turn out to be the best one out of the multilevel models considered in this paper, but it also offers a more straightforward comparison with STS models, since intercepts and regression coefficients in the latter type of models are in fact fixed effects. However, the performance of the two modelling approaches in terms of signal (trend etc.) variance estimates cannot be fairly compared at this stage. There are at least two reasons for that. We will try to quantify the effect of each of them shortly.

The first reason is the fact that the true (unknown) hyperparameters of a STS model (in this case, σ_u^2 and $\tilde{\sigma}_e^2$) are replaced by their maximum likelihood estimates and are treated as known. Within the HB approach, the uncertainty about these hyperparameter estimates is summarised by the variance of their posterior distribution. The signal MSE bias in STS models due to hyperparameter uncertainty can be negligible if the distribution of the ML estimator is symmetric and well concentrated around its mean. But, if the series is short or if the hyperparameters are close to their boundary values, the ML estimator distributions of such hyperparameters (and their posteriors too) are not symmetric. In this case, the uncertainty around the hyperparameter estimates could be very large, resulting in a considerable signal MSE bias in STS models. Therefore, lower signal variance estimates delivered by an STS model, in comparison with posterior signal variances, do not necessarily mean that the STS approach is superior to the HB one. In fact, the extent to which HB model-based SE estimates exceed those from a comparable STS model gives an indication for the scale of the MSE negative bias in such STS model. The negative bias in STS models can be accounted for by the bootstrap method of Pfeffermann and Tiller (2005). Chapter 4 shows why this method is superior to other existing methods for MSE estimation in STS models.

Another source of differences when comparing the two modelling approaches may appear when the signal posterior distributions are skewed, generally producing larger posterior standard deviations for domain predictors than the signal standard errors produced by a linear STS model. In fact, if the normality assumption about the data/disturbances is strongly violated, linear STS models like the ones considered in this paper should not be applied.

Point-estimates produced by the HB-FE and multivariate STS model are depicted in Fig. 5.6.5, together with the 95% confidence intervals from the STS model superimposed on the credible intervals from the HB-FE model. The latter are symmetric as they are constructed using the posterior standard deviations, rather than being quantile-based. In most domains, the posterior distribution of the domain predictors is symmetric. It is sometimes slightly skewed in small-scale domains. This is where differences in point- and variance estimates become visible. The STS model-based point-estimates tend to be smoother (as in the lower panel of Fig. 5.6.5), since they are based on smaller (close to the boundary space) values of the trend disturbance variances. The HB model-based point-estimates in such domains, in turn, stem from a set of draws from a heavy right-tailed distribution of the trend variance, which results in more flexible HB-based trends. Further, the differences in point-estimates from the STS and HB models are also partially due to the fact that posterior means of $\theta_{m,t}$, rather than medians, are taken as domain m predictors.

The signal standard error estimates produced by the two approaches could be compared and summarised in terms of relative deviation (RD) of the STS-derived signal SEs from the HB posterior standard deviations:

$$RD_{HB,t}^{STS} = \frac{SE_t^{STS} - SE_t^{HB}}{SE_t^{HB}} \cdot 100\%. \tag{5.6.1}$$

For the multivariate setting, the overall average of this measure across time and domains is equal to -9.4%, with a median of -3.1%. This itself does not imply serious differences between the outcomes of the two approaches, but shows that the distribution of the RD-terms is very skewed, with the mean being pulled to the left by extreme SE-differences in small-scale domains.

It would be interesting to check if there is a tendency for the HB-based SEs to exceed the STS-based ones in small-scale domains, since such domains are suspected to feature skewed posteriors of their signals. By "scale" here is meant a time average of the domain's point-estimates produced by the multivariate STS model. In this way, the scale represents both the number of km-pppd and the effective sample size: the scarcity of respondents belonging to a certain motive/modality intersection translates into low per-person figures for kilometers travelled for that intersection. The 672 provincial domain numbers sorted by scale in descending order are plotted on the x-axis of Fig. 5.6.6. It is clearly visible that, as the domain scale decreases, the STS signal standard errors tend to deviate more from the HB-FE posterior standard deviations in a negative direction. Most of the extreme deviations (those up to 200%) occur around level interventions. There, the HB-based signals obtain larger uncertainty than the STS-based ones.

Now we try to quantify the effect of hyperparameter uncertainty on the differences between the HB-FE-based and STS-based SEs. For that, an identical HB-FE model has been estimated (referred to as HB-FE-ML), for which informative priors on the trend disturbance variances have been set with a large number of degrees of freedom and the scale parameter taken equal to the ML estimates from the STS multivariate model. This effectively makes the hyperparameter values in the multilevel model equal to the ML-estimates from the STS model. Relative deviations as in (5.6.1) have been calculated for the HB-FE and HB-FE-ML posterior standard deviations. Their average value suggests that skewness in the signal posterior distributions can be blamed for only -1.6% out of the above-mentioned -9.4% reduction/underestimation in the HB posterior standard deviations by the STS approach. The remaining -7.8% are due to the hyperparameter uncertainty around the trend disturbance variances, not accounted for in the STS approach. The negative bias in the STS-based variance estimates can also be accounted for by the bootstrap method of Pfeffermann and Tiller (2005) presented in Chapter 4.

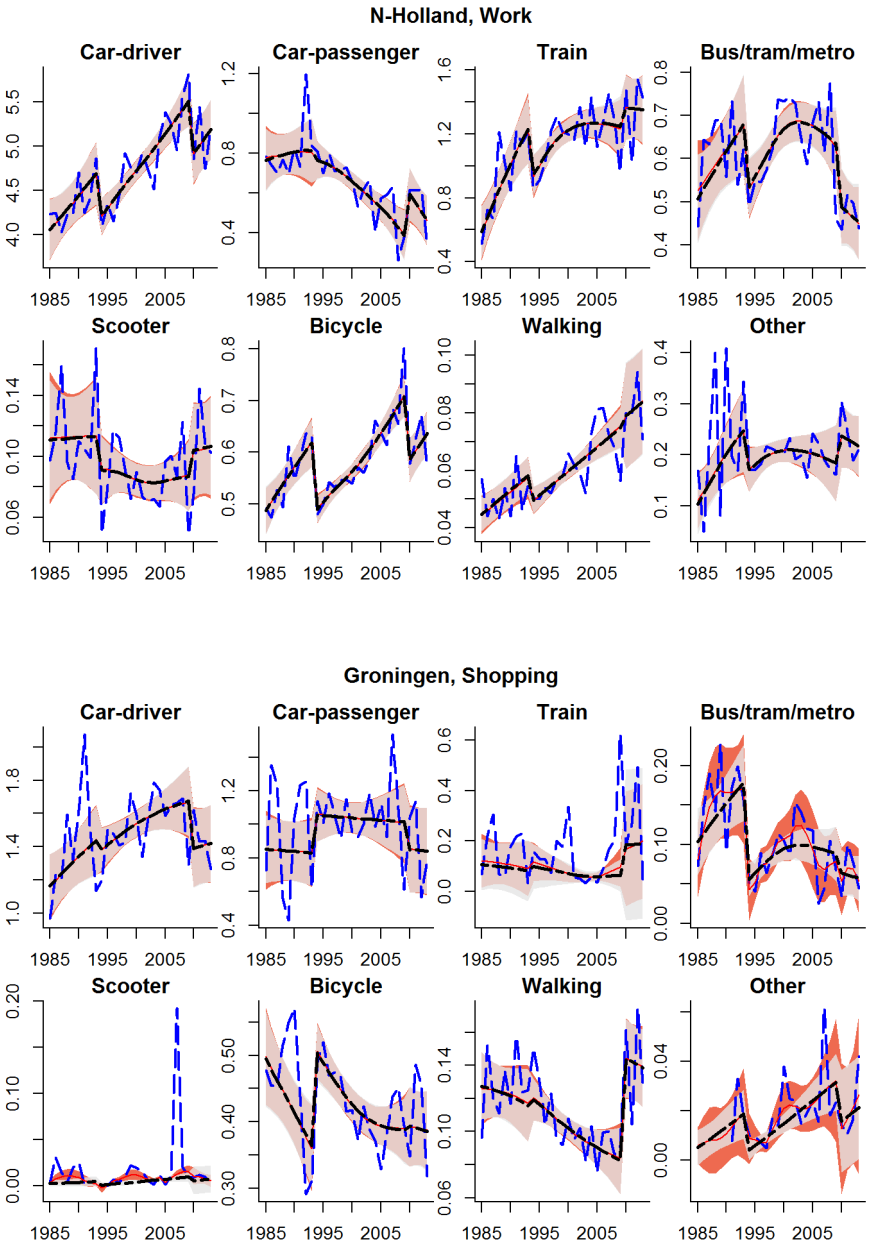


Figure 5.6.5: The provincial level design point-estimates (blue dashed line), point-estimates from the multivariate STS model (black dashed) and posterior means from the HB-FE model (red solid line), km-pppd; 95% confidence intervals from the STS model superimposed on the credible intervals from the HB-FE model, the colour corresponds to the colour of the point-estimates

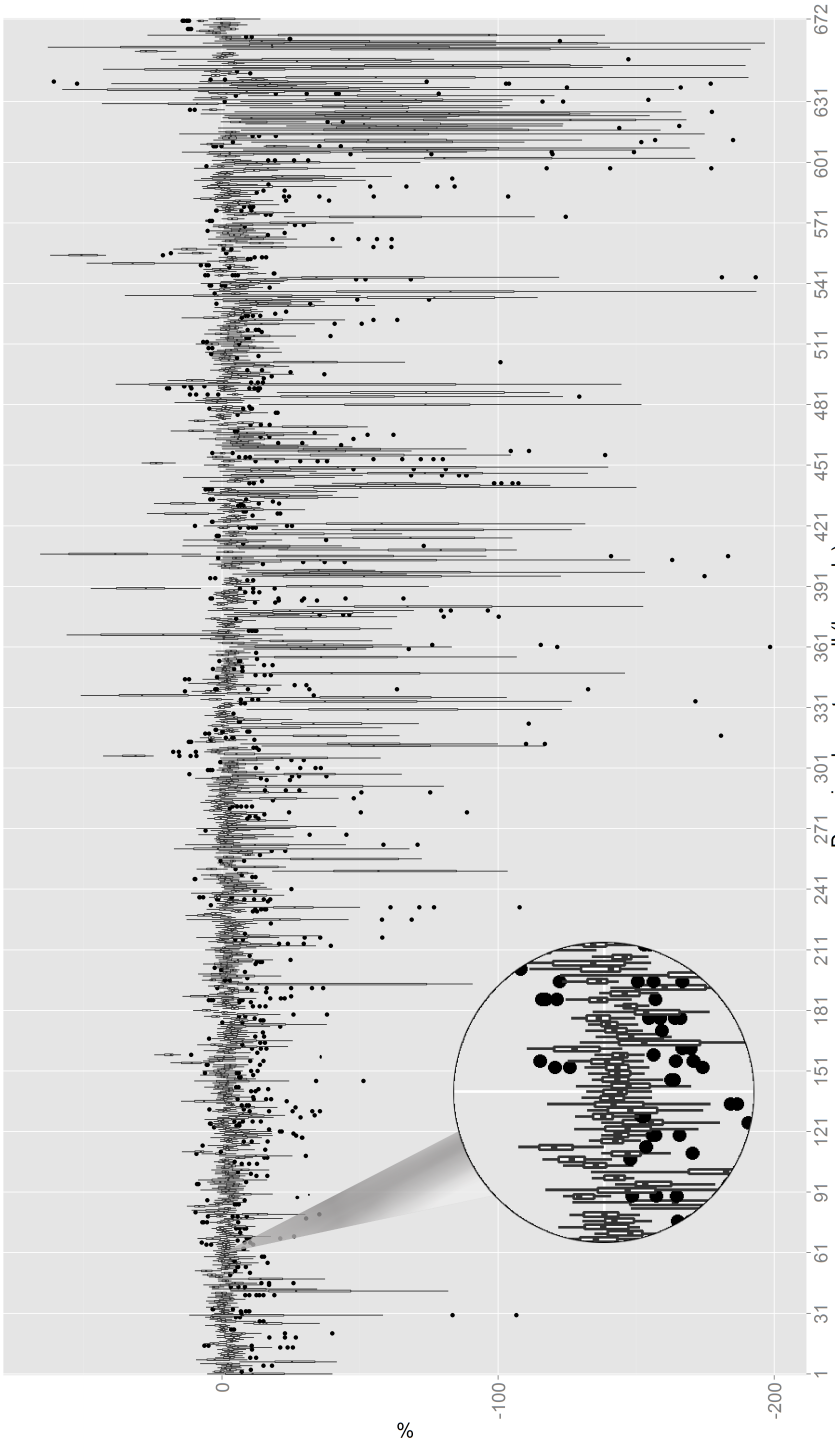


Figure 5.6.6: Box-plots for the provincial level $RD^{STS}_{HB,t}$ - relative deviation of signal SEs produced in the multivariate STS model from signal SEs produced by the HB-FE model, % (dots correspond to different time points).

It is of a particular interest how much reduction in design-based standard errors can be obtained by time series modelling technique. The HB-FE model offers a 51% reduction in design-based standard errors at the provincial level on average, with a median of 54%. For the STS multivariate model, where provinces are pooled as described in Subsection 5.6.1, the mean and median percentage reduction are slightly bigger - 54% and 57%, respectively. These and the above-mentioned figures indicate that the HB-FE and STS approaches deliver very similar results, with sizeable differences appearing mostly in small domains due to neglected hyperparameter uncertainty in STS models.

5.7 The DTS at the national level

5.7.1 The multivariate variance structure of stochastic trends

The number of domains at the national level is $M = 56$, defined by an intersection of 8 transport modalities with 7 motives. As already described in Subsection 5.6.1, one can look at box-plots with $\hat{\sigma}_{u,j,i}$s obtained from univariate STS models in order to see if these trend disturbance variance estimates can be pooled across motives and/or modalities. Similarly to the provincial level, $\sigma_{u,j,i}$s do not exhibit resemblance either across motives, or (and in particular) across the modalities, as can be seen in Fig. 5.7.1. The seeming resemblance between some of the box-plots (e.g., Mod_5-Mod_8) disappears when zooming in the plots. Endowing each domain with its own $\sigma_{u,j,i}$ at the national level limits the STS approach to the univariate setting. Within the multilevel approach, a multivariate structure hinges on the assumption of common distributions for the random effects in the HB-bRE and HB-bFE models. The HB-FE model at the national level constitutes a set of univariate HB time-series models.

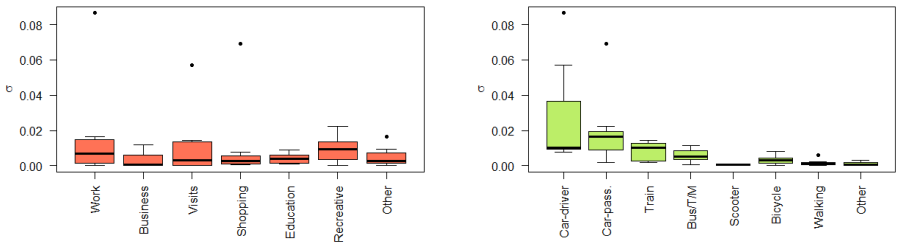


Figure 5.7.1: σ_u from univariate state space models at the national level (dots correspond to the transport modalities in the left panel and to the motives in the right one)

5.7.2 Multilevel and STS model estimation results compared

For the national level series, the same multivariate model variations described in Subsection 5.6.2 are considered. According to Table 5.7.1, the HB-FE model has the lowest DIC-value, but the differences among the three HB model modifications are minute. As Fig. 5.7.2 shows, the HB-bRE model does not experience any difficulties with the fit, as was the case at the provincial level. Point-estimates from the HB-bFE, HB-FE and STS models almost coincide. Appendix 5.C presents point-estimates and their credible/confidence intervals for the HB-FE and univariate STS models, as well as for the design-based estimator. The striking similarity between the HB-FE and STS model-based point- and variance estimates is less strong in small-scale domains, with the STS point-estimates being slightly smoother (e.g., in "Visits/Bicycle", "Visits/Other modalities", "Recreative/Bicycle", "Recreative/Other modalities").

For the comparison of the two approaches in terms of SEs, the reader is referred to equation (5.6.1). The overall average of $RD^{STS}_{HB,t}$-terms across time and domains at the national level is equal to -10.9%, and the median equals -9.1%. This negative mean is clearly visible in Fig. 5.7.3.

Table 5.7.1: Multilevel models considered for the DTS at the national level

Model labels	Iterations	Burn-in	Thinning	DIC
HB-bRE	75000	50000	50	-8114
HB-bFE	55000	30000	50	-8166
HB-FE	5000	1000	10	-8189

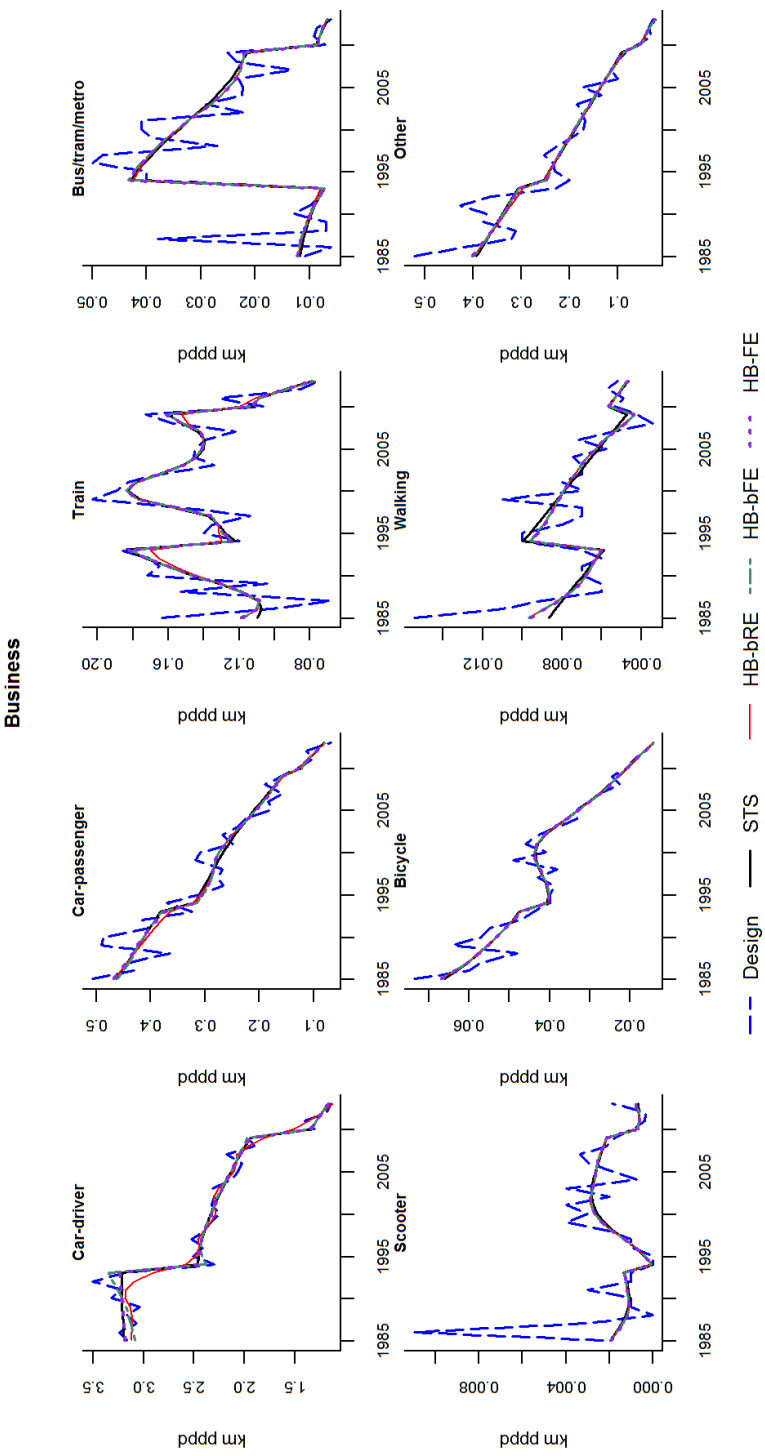


Figure 5.7.2: National level design- and model-based point-estimates for Mot 2, km-pppd.

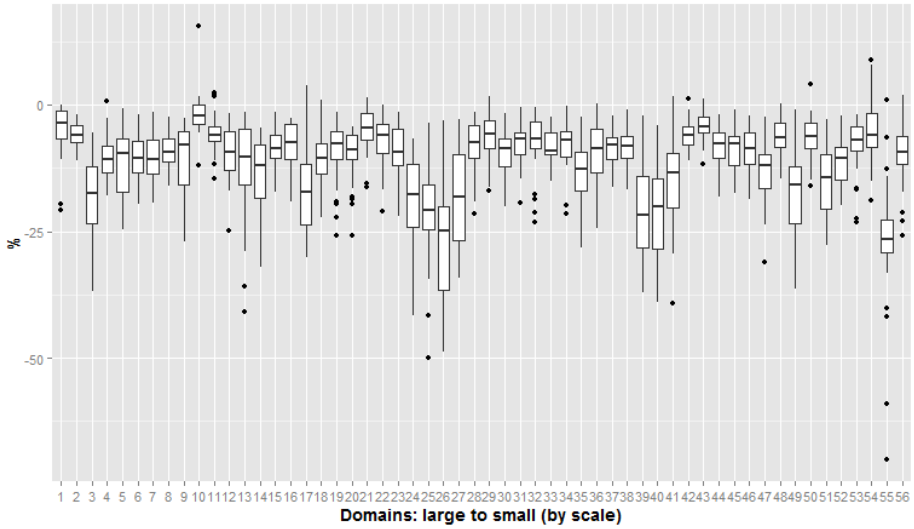


Figure 5.7.3: The national level $RD^{STS}_{HB,t}$ - relative deviation of STS SEs from HB-FE SEs for signals,% (dots correspond to different time points).

Though much less pronounced than at the provincial level, some association of low RD-terms with the decreasing domain scale can still be seen at the national level (Fig. 5.7.3). Indeed, extreme RD-values of more than 20% occur in several small domains (e.g., "Visits/Other modalities", "Visits/Bicycle"). Some of the largest domains still feature low RD-values (e.g., "Visits/Car-passenger", the third one from left). An inspection of the point-estimates of these domains in Appendix 5.C shows that the low $RD^{STS}_{HB,t}$-values correspond to those domains where the signal point-estimates from the HB models are visibly more volatile than those produced by STS models. As with the provincial data, a similar HB-FE-ML model has been estimated at the national level. It turns out that the difference between the HB- and STS-based SEs is almost entirely due to the hyperparameter uncertainty around the trend disturbance variances (an average -10.1% out of the above-mentioned -10.9%).

Mean reduction in the design-based standard errors with the model-based approach is less than at the provincial level, but still quite appreciable with the overall average of 31.7%, and the median of 34.3% for the HB-FE model, (39% and 41%, respectively, for the STS approach).

5.8 Discussion

Time series models are well known for their power in reducing design-based variances and in making point-estimates more stable, be it multilevel or structural time series models. Apart from that, time series techniques are indispensable when it comes to estimating level breaks due to survey redesigns. This paper aimed at establishing which of these two modelling approaches should be preferred.

The multilevel model estimated with the hierarchical Bayesian approach (HB model) is a time-series extension of the Fay-Herriot model. Apart from featuring hyperpriors for its parameters, it differs from the conventional STS model (in the sense of Harvey (1989), Durbin and Koopman (2012)) in that time-invariant effects (e.g. intercepts, level break or linear trend coefficients) can be treated as random. In this application, however, assumptions about random effects sharing the same variance across domains turned out to be invalid at the provincial level. Therefore, several HB-model variations have been considered where (some of) these random components are modeled as fixed ones. Not only does the fixed-effect specification of the HB-model (HB-FE) make the comparison of the HB and STS approaches more straightforward, but it is also the only specification that provides an adequate fit to the DTS data.

The comparison between the STS and the HB-FE model in terms of estimated signal variances is still not completely fair. First of all, this is due to the fact that the true (unknown) hyperparameters of a STS model are replaced by their maximum likelihood estimates and are treated as known. To account for this additional uncertainty, one would in addition have to resort to bootstrapping techniques, see Pfeffermann and Tiller (2005) and Chapter 4 of this thesis. Within the HB-FE approach, the uncertainty around hyperparameter estimates is summarised by the variance of their posterior distribution. Therefore, lower signal variance estimates delivered by an STS model do not necessarily mean that the STS approach is superior to the HB-FE one. In fact, the extent to which HB-FE model-based variance estimates exceed those from a comparable STS model gives us an indication for the scale of the negative bias in the signal MSEs of such STS model. This bias could be particularly large if variance hyperparameters are close to their boundary values, with heavy right tails in the posterior distributions/distributions of the ML estimator. This is the case in many small-scale domains of the DTS,

which makes the hyperparameter uncertainty the primary source of differences between the two modelling approaches in the case of the DTS. Secondly, another source of differences is skewed signal posterior distributions that produce larger posterior standard deviations for domain predictors than the signal standard errors produced by a linear STS model relying on the assumption of normality in the data/disturbances. In the DTS, slight skewness in the signal posterior is observed at the provincial level, mainly in those domains whose trend disturbance variance posteriors are skewed enough to feed some degree of asymmetry through to the signal posterior.

It turns out that both point- and variance model-based estimates produced with the STS and HB-FE techniques are very similar. Differences become visible mainly in small domains. Standard errors produced by the STS model are smaller than posterior standard deviations from the HB-FE model by 9.4% on average (across time and domains) at the provincial level (median 3.1%), and by 10.9% at the national level (median 9.1%). At the national level of the DTS, skewness turned out to have a negligible effect on the standard error difference between the two approaches, but at the provincial level it is responsible for about 1.5% out of the above-mentioned 9.4% of the HB-FE model-based posterior standard deviations on average. The rest 7.9% is due to the hyperparameter uncertainty unaccounted for within the STS approach. For these above-mentioned reasons, one should be aware of negative biases in frequentist STS-based variances in short time series or series that feature small variance hyperparameters. In such cases, the negative biases should be accounted for by means of an additional procedure, such as the bootstrap of Pfeffermann and Tiller (2005) studied in Chapter 4.

As an important by-product, the results of this paper give an idea about how much reduction in design-based standard errors can be obtained by these time series modelling techniques. The mean reduction in design-based standard errors with the HB-FE model is 51% at the provincial level, with the median of 54%. In other words, in order to reduce the design-based variance to this extent, one would have to increase the sample size more than four-fold (conditional on the point-estimates). For the STS model, the mean and median percentage reduction in the design-based standard errors is slightly bigger - 54% and 57%, respectively, - due to the reasons already mentioned above. Mean reduction in design-based standard errors at the national level is smaller than at the provincial one, but is

still quite appreciable with the overall average of 31.7% and the median of 34.3% for the HB-FE model, and 39% and 41%, respectively, for the STS model.

Another aspect we look at is unreliable design-based variance estimates. First of all, these estimates are subject to sampling volatility. With the help of a simple STS model, design-based standard errors can be smoothed with the Kalman filter. These smoothed standard error estimates are further used as input in the multilevel and STS models. Secondly, design-based variance estimates could be biased in the case of small domains. If the bias is negative, for instance, then treating these estimates as the true variances in a multilevel setting results in model overfit by putting too much weight on the design-based estimates. We suggest using the STS univariate analysis to scale the design-based variance estimates in the right direction for further use in multilevel models. The uncertainty around design-based variance estimates could also be taken into account by imposing a prior distribution on them, as in You and Chapman (2006). The comparative analysis of the two time series modelling approaches, however, should not be affected much by the way the design-based variance estimates are treated in this paper. The reported reduction in design-based standard errors is conditional on these approximated design-based variance estimates.

The techniques presented here can be used nearly in any repeatedly conducted SAE application, especially if a survey suffers from discontinuities due to redesigns. Unlike in the application considered in Bollineni-Balabay et al. (2016a), the availability of design-based variance estimates here makes it possible to continuously apply this model-based approach in the production of official statistical figures. Accounting for each new survey redesign will, however, be possible only with a delay of at least one period, unless a parallel run is being carried out to get an idea about the extent of the new level break. In addition, in the first time periods after a survey redesign, estimated figures are likely to undergo substantial revisions, as soon as new data become available under the new design. Yet, this problem is of a temporary nature and there does not seem to exist any other solution except for a parallel run that increases the survey expenses. It is hardly possible to make a guess about the effect of the coming redesign based on the past level breaks caused by other redesigns. See Van den Brakel and Krieg (2015) for an example where a parallel run is conducted to obtain design-based estimates for discontinuities. These estimates are then used as a priori information in a structural time series

model to avoid the problem of revisions.

In this paper, the time dimension has been exploited for variance and volatility reduction in the point-estimates, as well as for level break estimation. As for the spatial dimension, so far it has been used only for getting rid of some idiosyncracies through pooling trend variances across provinces. However, it can also be used as another source of variance reduction in model-based estimates by exploiting spatial correlations between domains belonging to different provinces, e.g., by allowing for f common stochastic trends shared by more than f domain trends, the way it was done in Chapter 2, as well as in Krieg and Van den Brakel (2012) and Van den Brakel and Krieg (2016). In the case of the DTS, this approach seems to be worth exploring, since the pattern of trends belonging to a certain motive-modality intersection is very similar among the twelve provinces of the Netherlands.

5.A Full conditional distributions for the Gibbs-sampler

The Gibbs sampler was first described by Geman and Geman (1984). Here we present conditional posterior densities for each parameter b in
$\boldsymbol{\psi} = (\boldsymbol{\nu'}, \boldsymbol{\kappa'}, \boldsymbol{\beta^{RE'}}, \boldsymbol{u'}, \boldsymbol{\beta'}, \sigma^2_\nu, \sigma^2_\kappa, \sigma^2_{\beta^{RE}}, \sigma^2_u, \xi_\nu, \xi_\kappa, \xi_{\beta^{RE}}, \xi_u)'$
for Model (5.3.2). Let $\boldsymbol{\psi^{(-b)}}$ denote the parameter vector where element b is deleted. Within the Gibbs sampler, the b-th parameter values are drawn conditionally on the data $\boldsymbol{Y}$ and the rest of the parameters $\boldsymbol{\psi^{(-b)}}$.

The conditional posterior density for parameters $c, \kappa, \beta^0_1, ..., \beta^0_{K^{RE}}$ and $\boldsymbol{\beta^{FE}}$ contained in vector $\boldsymbol{\beta}$ originate from the product of the densities that contain these parameters:

$$p(\boldsymbol{\beta}|\boldsymbol{\psi}^{(-\boldsymbol{\beta})}, \boldsymbol{Y}) \propto N_\beta(\boldsymbol{\beta_0}, \boldsymbol{\Omega}_{\boldsymbol{\beta_0}}) \times N_Y(\xi_\nu \tilde{\boldsymbol{\nu}} \otimes \boldsymbol{\iota}_{[\boldsymbol{T}]} + \xi_\kappa \tilde{\boldsymbol{\kappa}} \otimes \boldsymbol{t} + \xi_{\beta^{RE}} \boldsymbol{X^{RE}} \tilde{\boldsymbol{\beta}}^{\boldsymbol{RE}} + \boldsymbol{X\beta} + \xi_u \tilde{\boldsymbol{u}}, \boldsymbol{\Phi}), \tag{5.A.1}$$

which, using the results on conjugate priors in Gelman et al. (2014), turns into a normal density $N(\boldsymbol{\mu_\beta}, \boldsymbol{\Omega_\beta})$ with the following mean and variance:

$$\begin{aligned}
\boldsymbol{\mu_\beta} =& (\boldsymbol{X'\Phi^{-1}X} + \boldsymbol{\Omega^{-1}_{\beta_0}})^{-1} \times \\
& \left[\boldsymbol{X'\Phi^{-1}}(\boldsymbol{Y} - \xi_\nu \tilde{\boldsymbol{\nu}} \otimes \boldsymbol{\iota}_{[\boldsymbol{T}]} - \xi_\kappa \tilde{\boldsymbol{\kappa}} \otimes \boldsymbol{t} - \xi_{\beta^{RE}} \boldsymbol{X^{RE}} \tilde{\boldsymbol{\beta}}^{\boldsymbol{RE}} - \xi_u \tilde{\boldsymbol{u}}) + \boldsymbol{\Omega^{-1}_{\beta_0}} \boldsymbol{\beta_0}\right], \\
\boldsymbol{\Omega_\beta} =& (\boldsymbol{X'\Phi^{-1}X} + \boldsymbol{\Omega^{-1}_{\beta_0}})^{-1}.
\end{aligned}$$

For vector and matrix dimensions, refer to the description under the likelihood function presentation in (5.3.10).

The conditional posterior of the M-dimensional vector of scaled area effects $\tilde{\nu}_m$ is:

$$\begin{aligned}
p(\tilde{\boldsymbol{\nu}}|\boldsymbol{\psi}^{(-\tilde{\nu})}, \boldsymbol{Y}) =& N_{\tilde{\nu}}(\boldsymbol{\mu}_{\tilde{\boldsymbol{\nu}}}, \boldsymbol{\Omega}_{\tilde{\boldsymbol{\nu}}}) \propto N_{\tilde{\nu}}(\mathbf{0}_{[M]}, \tilde{\sigma}_\nu^2 \boldsymbol{I}_{[M]}) \times \\
& N_Y(\xi_\nu \tilde{\boldsymbol{\nu}} \otimes \boldsymbol{\iota}_{[\boldsymbol{T}]} + \xi_\kappa \tilde{\boldsymbol{\kappa}} \otimes \boldsymbol{t} + \xi_{\beta^{RE}} \boldsymbol{X}^{\boldsymbol{RE}} \tilde{\boldsymbol{\beta}}^{\boldsymbol{RE}} + \boldsymbol{X}\boldsymbol{\beta} + \xi_u \tilde{\boldsymbol{u}}, \boldsymbol{\Phi}), \\
\boldsymbol{\mu}_{\tilde{\boldsymbol{\nu}}} =& (\xi_\nu^2 \boldsymbol{I}_{[M]} \otimes \boldsymbol{\iota}'_{[\boldsymbol{T}]} \boldsymbol{\Phi}^{-\mathbf{1}} \boldsymbol{I}_{[M]} \otimes \boldsymbol{\iota}_{[\boldsymbol{T}]} + 1/\tilde{\sigma}_\nu^2 \boldsymbol{I}_{[M]})^{-1} \times \\
& \xi_\nu^2 \boldsymbol{I}_{[M]} \otimes \boldsymbol{\iota}'_{[\boldsymbol{T}]} \boldsymbol{\Phi}^{-\mathbf{1}} (\boldsymbol{Y} - \xi_\kappa \tilde{\boldsymbol{\kappa}} \otimes \boldsymbol{t} - \xi_{\beta^{RE}} \boldsymbol{X}^{\boldsymbol{RE}} \tilde{\boldsymbol{\beta}}^{\boldsymbol{RE}} - \boldsymbol{X}\boldsymbol{\beta} - \xi_u \tilde{\boldsymbol{u}}), \\
\boldsymbol{\Omega}_{\tilde{\boldsymbol{\nu}}} =& (\xi_\nu^2 \boldsymbol{I}_{[M]} \otimes \boldsymbol{\iota}'_{[\boldsymbol{T}]} \boldsymbol{\Phi}^{-\mathbf{1}} \boldsymbol{I}_{[M]} \otimes \boldsymbol{\iota}_{[\boldsymbol{T}]} + 1/\tilde{\sigma}_\nu^2 \boldsymbol{I}_{[M]})^{-1}.
\end{aligned} \tag{5.A.2}$$

A similar expression is valid for the linear trend random effects $\tilde{\kappa}_m$:

$$\begin{aligned}
p(\tilde{\boldsymbol{\kappa}}|\boldsymbol{\psi}^{(-\tilde{\kappa})}, \boldsymbol{Y}) =& N_{\tilde{\kappa}}(\boldsymbol{\mu}_{\tilde{\boldsymbol{\kappa}}}, \boldsymbol{\Omega}_{\tilde{\boldsymbol{\kappa}}}) \propto N_{\tilde{\kappa}}(\mathbf{0}_{[M]}, \tilde{\sigma}_\kappa^2 \boldsymbol{I}_{[M]}) \times \\
& N_Y(\xi_\nu \tilde{\boldsymbol{\nu}} \otimes \boldsymbol{\iota}_{[\boldsymbol{T}]} + \xi_\kappa \tilde{\boldsymbol{\kappa}} \otimes \boldsymbol{t} + \xi_{\beta^{RE}} \boldsymbol{X}^{\boldsymbol{RE}} \tilde{\boldsymbol{\beta}}^{\boldsymbol{RE}} + \boldsymbol{X}\boldsymbol{\beta} + \xi_u \tilde{\boldsymbol{u}}, \boldsymbol{\Phi}), \\
\boldsymbol{\mu}_{\tilde{\boldsymbol{\kappa}}} =& (\xi_\kappa^2 \boldsymbol{I}_{[M]} \otimes \boldsymbol{t}' \boldsymbol{\Phi}^{-\mathbf{1}} \boldsymbol{I}_{[M]} \otimes \boldsymbol{t} + 1/\tilde{\sigma}_\kappa^2 \boldsymbol{I}_{[M]})^{-1} \times \\
& \xi_\kappa^2 \boldsymbol{I}_{[M]} \otimes \boldsymbol{t}' \boldsymbol{\Phi}^{-\mathbf{1}} (\boldsymbol{Y} - \xi_\nu \tilde{\boldsymbol{\nu}} \otimes \boldsymbol{\iota}_{[\boldsymbol{T}]} - \xi_{\beta^{RE}} \boldsymbol{X}^{\boldsymbol{RE}} \tilde{\boldsymbol{\beta}}^{\boldsymbol{RE}} - \boldsymbol{X}\boldsymbol{\beta} - \xi_u \tilde{\boldsymbol{u}}), \\
\boldsymbol{\Omega}_{\tilde{\boldsymbol{\kappa}}} =& (\xi_\kappa^2 \boldsymbol{I}_{[M]} \otimes \boldsymbol{t}' \boldsymbol{\Phi}^{-\mathbf{1}} \boldsymbol{I}_{[M]} \otimes \boldsymbol{t} + 1/\tilde{\sigma}_\kappa^2 \boldsymbol{I}_{[M]})^{-1},
\end{aligned} \tag{5.A.3}$$

where $\boldsymbol{t}$ denotes a vertical vector with time indicators $(0, 1, ..., T-1)'$. The same logic applies to random level break coefficients $\tilde{\boldsymbol{\beta}}_{\boldsymbol{k}}^{\boldsymbol{RE}}$ (inter alia, vectors $\boldsymbol{t}$ are replaced with T-dimensional vectors with dummy regressors for the level break k).

The conditional distribution for the T-dimensional scaled stochastic trend random terms $\tilde{\boldsymbol{u}}_{\boldsymbol{m}}$ is based on the data and other terms that are related to domain m:

$$\begin{aligned}
p(\tilde{\boldsymbol{u}}_{\boldsymbol{m}}|\boldsymbol{\psi}^{(-\tilde{\boldsymbol{u}}_{\boldsymbol{m}})}, \boldsymbol{Y}) =& N_{\tilde{u}}(\boldsymbol{\mu}_{\tilde{\boldsymbol{u}}_{\boldsymbol{m}}}, \boldsymbol{\Omega}_{\tilde{\boldsymbol{u}}_{\boldsymbol{m}}}) \propto N_{\tilde{u}}(\mathbf{0}_{[\boldsymbol{T}]}, \tilde{\sigma}_{u,m}^2 \boldsymbol{A}) \times \\
& N_Y(\xi_\nu \tilde{\nu}_m \boldsymbol{\iota}_{[\boldsymbol{T}]} + \xi_\kappa \tilde{\kappa}_m \boldsymbol{t} + \xi_{\beta^{RE}} \boldsymbol{X}_{\boldsymbol{m}}^{\boldsymbol{RE}} \tilde{\boldsymbol{\beta}}_{\boldsymbol{m}}^{\boldsymbol{RE}} + \\
& \boldsymbol{X}_{\boldsymbol{m}} \boldsymbol{\beta}_{\boldsymbol{m}} + \xi_u \tilde{\boldsymbol{u}}_{\boldsymbol{m}}, \boldsymbol{\Phi}_{\boldsymbol{m}}), \\
\boldsymbol{\mu}_{\tilde{\boldsymbol{u}}_{\boldsymbol{m}}} =& (\xi_u^2 \boldsymbol{\Phi}_{\boldsymbol{m}}^{-\mathbf{1}} + \boldsymbol{A}^{-\mathbf{1}}/\tilde{\sigma}_{u,m}^2)^{-1} \times \\
& \xi_u^2 \boldsymbol{\Phi}_{\boldsymbol{m}}^{-\mathbf{1}} (\boldsymbol{Y}_{\boldsymbol{m}} - \xi_\nu \tilde{\nu}_m \boldsymbol{\iota}_{[\boldsymbol{T}]} - \xi_\kappa \tilde{\kappa}_m \boldsymbol{t} - \xi_{\beta^{RE}} \boldsymbol{X}_{\boldsymbol{m}}^{\boldsymbol{RE}} \tilde{\boldsymbol{\beta}}_{\boldsymbol{m}}^{\boldsymbol{RE}} - \boldsymbol{X}_{\boldsymbol{m}} \boldsymbol{\beta}_{\boldsymbol{m}}), \\
\boldsymbol{\Omega}_{\tilde{\boldsymbol{u}}_{\boldsymbol{m}}} =& (\xi_u^2 \boldsymbol{\Phi}_{\boldsymbol{m}}^{-\mathbf{1}} + \boldsymbol{A}^{-\mathbf{1}}/\tilde{\sigma}_{u,m}^2)^{-1},
\end{aligned} \tag{5.A.4}$$

where vector $\boldsymbol{Y_m}$ and matrix $\boldsymbol{\Phi_m}$ are T-dimensional, matrix $\boldsymbol{X_m^{RE}}$ is $[T \times K_{RE}]$ and contains indicator variables for the scaled random level break coefficients $\tilde{\boldsymbol{\beta}}_{\boldsymbol{m}}^{\boldsymbol{RE}}$ in domain m, $\tilde{\boldsymbol{\beta}}_{\boldsymbol{m}}^{\boldsymbol{RE}}$ being $[K_{RE} \times 1]$. The matrix with fixed effects regressors $\boldsymbol{X_m}$ is build by the same logic, i.e. contains regressors only applicable to domain m and is therefore $[T \times (2 + K_{RE} + d_m^{FE})]$, d_m^{FE} being the number of level breaks modelled as fixed effects in domain m. $\boldsymbol{\beta_m}$ contains the overall effects $c, \kappa, \beta_1^0, ...\beta_{K_{RE}}^0$ along with d_m^{FE} fixed level break coefficients. Drawing from (5.A.4) results in unconstrained draws of the stochastic trend terms. Due to restrictions presented in (5.3.4), these unconstrained draws should be adjusted to $\left(\boldsymbol{I_{[T]}} - \boldsymbol{\Omega_{\tilde{u}} R(R'\Omega_{\tilde{u}} R)^{-1} R'}\right) \tilde{\boldsymbol{u}}_{\boldsymbol{m}}$, where $\boldsymbol{R} = (\boldsymbol{\iota_{[T]}}, \boldsymbol{t_{[T]}})$ is a $[T \times 2]$-matrix, and $\boldsymbol{t} = (1, ..., T)'$, according to Rue and Held (2005).

Variance components of random effects are drawn from the $Inv - \chi^2$ density:

$$\begin{aligned} p(\tilde{\sigma}_\nu^2 | \boldsymbol{\psi^{(-\tilde{\sigma}_\nu^2)}}, \boldsymbol{Y}) &\propto N_{\tilde{\nu}}(\boldsymbol{0_{[M]}}, \tilde{\sigma}_\nu^2 \boldsymbol{I_{[M]}}) \times Inv - \chi^2_{\tilde{\sigma}_\nu^2}(v_\nu, s_\nu^2), \\ p(\tilde{\sigma}_\nu^2 | \boldsymbol{\psi^{(-\tilde{\sigma}_\nu^2)}}, \boldsymbol{Y}) &= Inv - \chi^2_{\tilde{\sigma}_\nu^2}(v_\nu + M, \frac{v_\nu s_\nu^2 + \sum_m \tilde{\nu}_m^2}{v_\nu + M}). \end{aligned} \tag{5.A.5}$$

The same goes for the other random effects, except for the stochastic trend terms (e.g., for random linear trend effects $\tilde{\kappa}_m$, terms v_ν, s_ν^2 and $\sum_m \tilde{\nu}_m^2$ would be replaced by v_κ, s_κ^2 and $\sum_m \tilde{\kappa}_m^2$, respectively).

If every domain m is assigned a unique value for its stochastic trend variance (as is the case at the national level of the DTS), then $\tilde{\sigma}_{u,m}^2$ is drawn from the following conditional:

$$\begin{aligned} p(\tilde{\sigma}_{u,m}^2 | \boldsymbol{\psi^{(-\tilde{\sigma}_{u,m}^2)}}, \boldsymbol{Y}) &\propto Inv - \chi^2_{\tilde{\sigma}_{u,m}^2}(v_u, s_u^2) N_{\tilde{u}_m}(\boldsymbol{0_{[T]}}, \tilde{\sigma}_{u,m}^2 \boldsymbol{A}), \\ p(\tilde{\sigma}_{u,m}^2 | \boldsymbol{\psi^{(-\tilde{\sigma}_{u,m}^2)}}, \boldsymbol{Y}) &= Inv - \chi^2_{\tilde{\sigma}_{u,m}^2}(v_u + T - 2, \frac{v_u s_u^2 + \tilde{\boldsymbol{u}}_{\boldsymbol{m}}' \boldsymbol{A^{-1}} \tilde{\boldsymbol{u}}_{\boldsymbol{m}}}{v_u + T - 2}), \end{aligned} \tag{5.A.6}$$

where 2 is subtracted from T due to the two restrictions for the integrated random walk model.

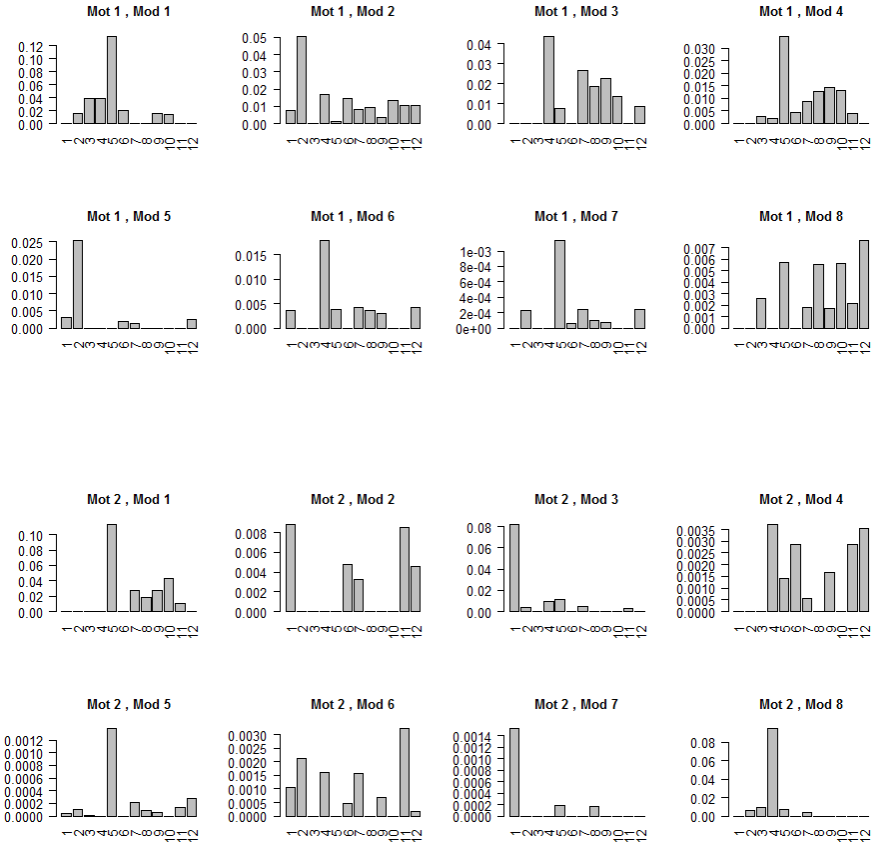
The scaling ξ_ν-effects are drawn from the following distribution:

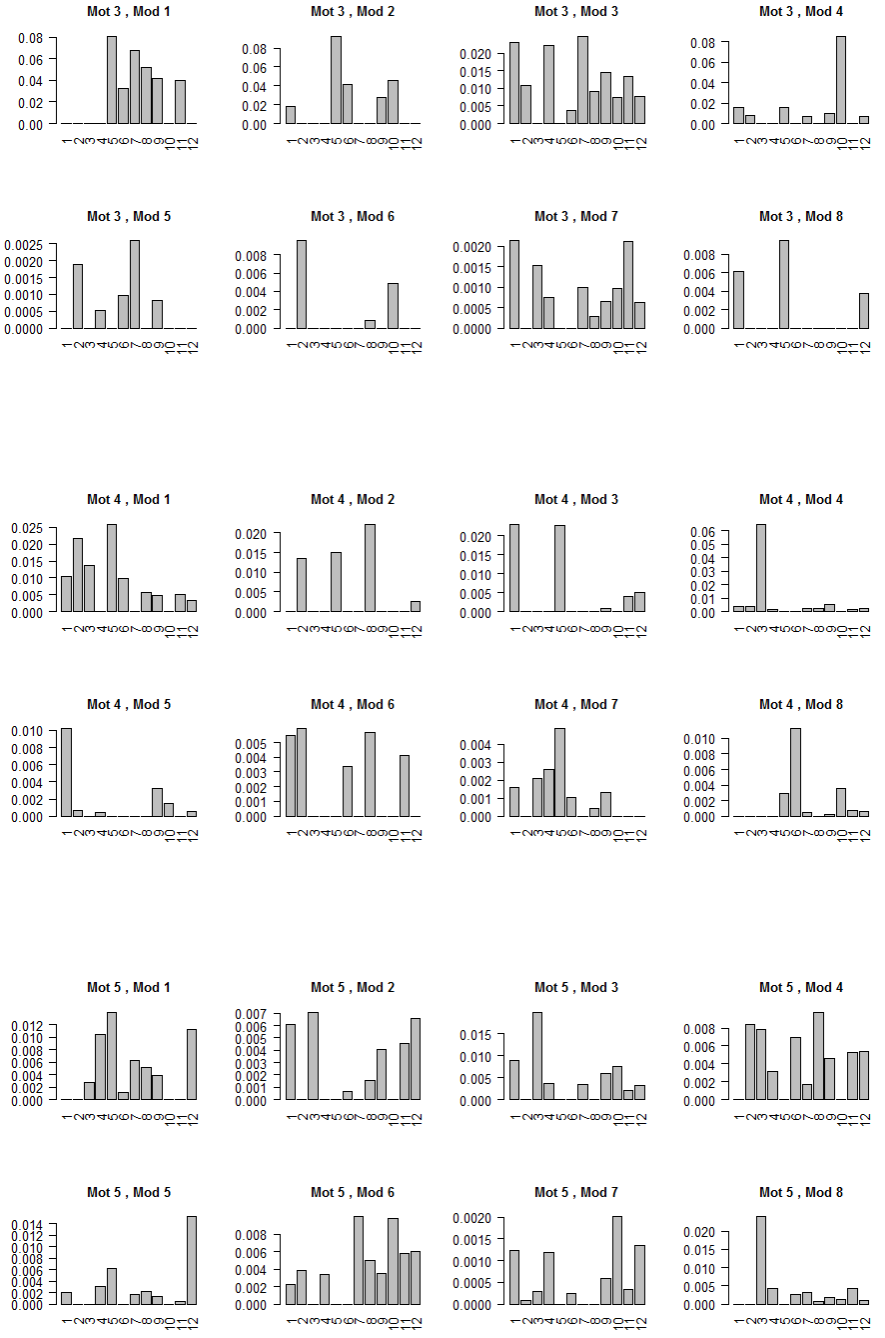
$$\begin{aligned}
p(\xi_\nu|\boldsymbol{\psi}^{(-\boldsymbol{\xi_\nu})}, \boldsymbol{Y}) =& N_{\xi_\nu}(\mu_{\xi_\nu}, \omega_{\xi_\nu}) \propto N_{\xi_\nu}(\alpha_\nu, \gamma_\nu)\times \\
& N_Y(\xi_\nu \tilde{\boldsymbol{\nu}} \otimes \boldsymbol{\iota}_{[\boldsymbol{T}]} + \xi_\kappa \tilde{\boldsymbol{\kappa}} \otimes \boldsymbol{t} + \xi_{\beta^{RE}} \boldsymbol{X}^{\boldsymbol{RE}} \tilde{\boldsymbol{\beta}}^{\boldsymbol{RE}} + \boldsymbol{X\beta} + \xi_u \tilde{\boldsymbol{u}}, \boldsymbol{\Phi}), \\
\mu_{\xi_\nu} =& (\tilde{\boldsymbol{\nu}}' \boldsymbol{I}_{[M]} \otimes \boldsymbol{\iota}'_{[\boldsymbol{T}]} \boldsymbol{\Phi}^{-\mathbf{1}} \boldsymbol{I}_{[M]} \otimes \boldsymbol{\iota}_{[\boldsymbol{T}]} \tilde{\boldsymbol{\nu}} + 1/\gamma_\nu)^{-1} \times \\
& \tilde{\boldsymbol{\nu}}' \boldsymbol{I}_{[M]} \otimes \boldsymbol{\iota}'_{[\boldsymbol{T}]} \boldsymbol{\Phi}^{-\mathbf{1}} (\boldsymbol{Y} - \xi_\kappa \tilde{\boldsymbol{\kappa}} \otimes \boldsymbol{t} - \xi_{\beta^{RE}} \boldsymbol{X}^{\boldsymbol{RE}} \tilde{\boldsymbol{\beta}}^{\boldsymbol{RE}} - \boldsymbol{X\beta} - \xi_u \tilde{\boldsymbol{u}}), \\
\omega_{\xi_\nu} =& (\tilde{\boldsymbol{\nu}}' \boldsymbol{I}_{[M]} \otimes \boldsymbol{\iota}'_{[\boldsymbol{T}]} \boldsymbol{\Phi}^{-\mathbf{1}} \boldsymbol{I}_{[M]} \otimes \boldsymbol{\iota}_{[\boldsymbol{T}]} \tilde{\boldsymbol{\nu}} + 1/\gamma_\nu)^{-1}.
\end{aligned} \tag{5.A.7}$$

The same logic applies to the scaling parameters of the rest of the random terms.

As for scaling the stochastic trend terms, the number of unique ξ_u-parameters can be made equal to the number of unique variances on the main diagonal of $\boldsymbol{\Sigma_u}$. At the provincial level of the DTS, for instance, the number of unique stochastic trend variances, and thus ξ_u-parameters, is $I \times J$, producing a IJ-dimensional $\boldsymbol{\xi_u}$ vector. In this case, $\xi_u \tilde{\boldsymbol{u}}$ terms in every expression of this appendix should be substituted with $\{\tilde{\boldsymbol{u}}'_{[\boldsymbol{PJIT}]}[\boldsymbol{I}_{[\boldsymbol{P}]} \otimes \boldsymbol{diag}(\boldsymbol{\xi}_{[\boldsymbol{IJ}]}) \otimes \boldsymbol{I}_{[\boldsymbol{T}]}]\}'$.

5.B Homogeneity analysis of σ_us




Mot 3 , Mod 1
Mot 3 , Mod 2
Mot 3 , Mod 3
Mot 3 , Mod 4
Mot 3 , Mod 5
Mot 3 , Mod 6
Mot 3 , Mod 7
Mot 3 , Mod 8
Mot 4 , Mod 1
Mot 4 , Mod 2
Mot 4 , Mod 3
Mot 4 , Mod 4
Mot 4 , Mod 5
Mot 4 , Mod 6
Mot 4 , Mod 7
Mot 4 , Mod 8
Mot 5 , Mod 1
Mot 5 , Mod 2
Mot 5 , Mod 3
Mot 5 , Mod 4
Mot 5 , Mod 5
Mot 5 , Mod 6
Mot 5 , Mod 7
Mot 5 , Mod 8

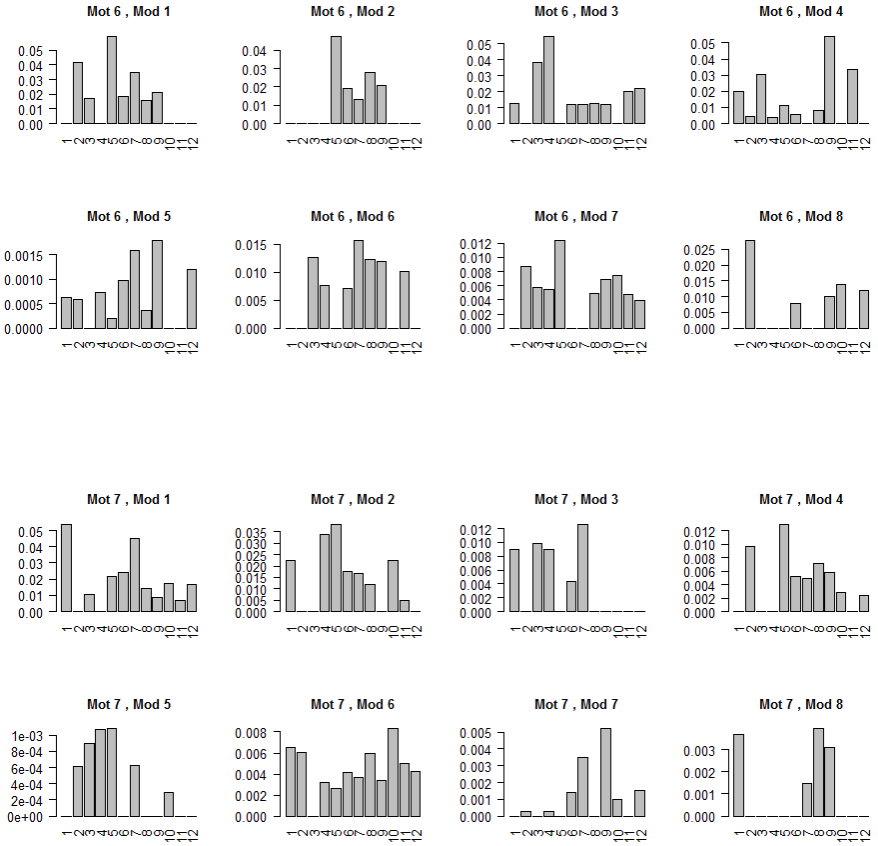
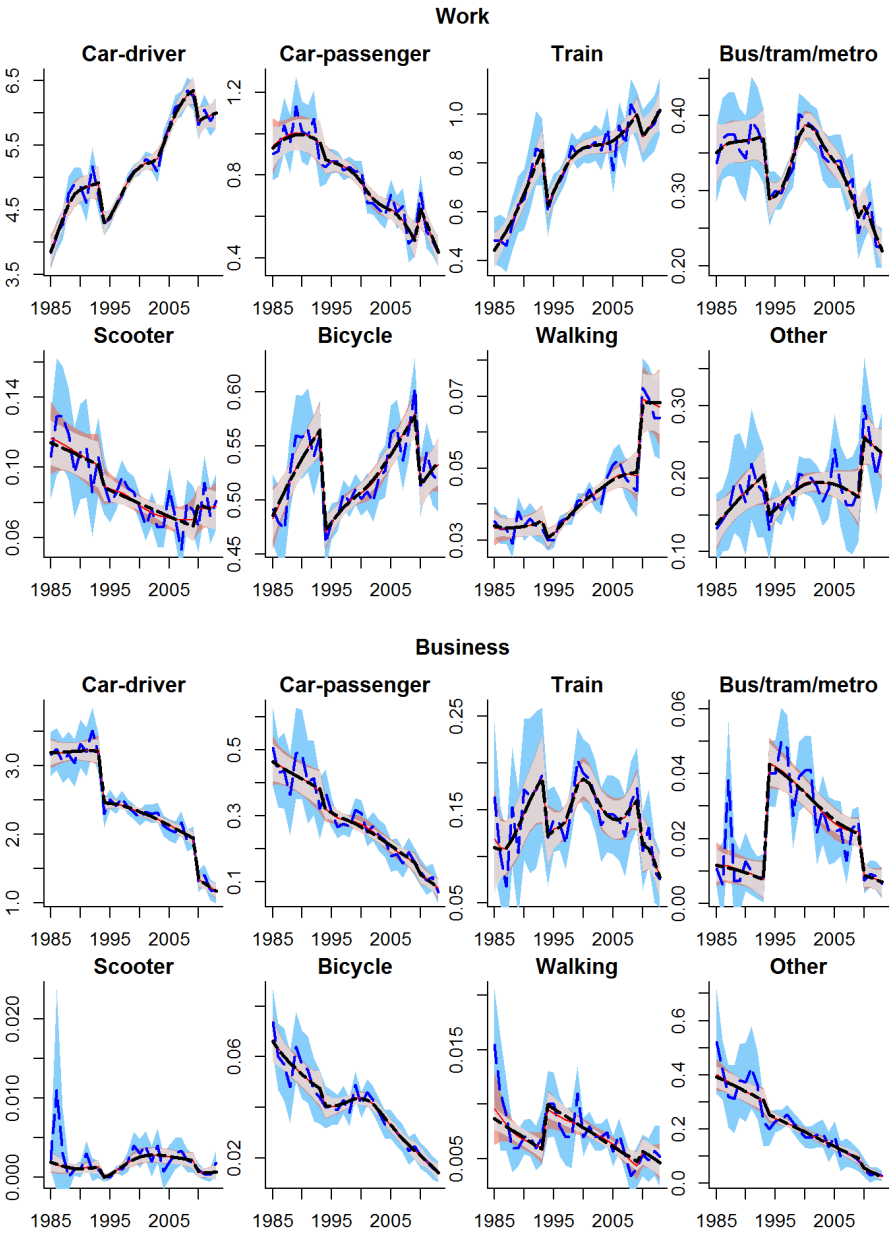
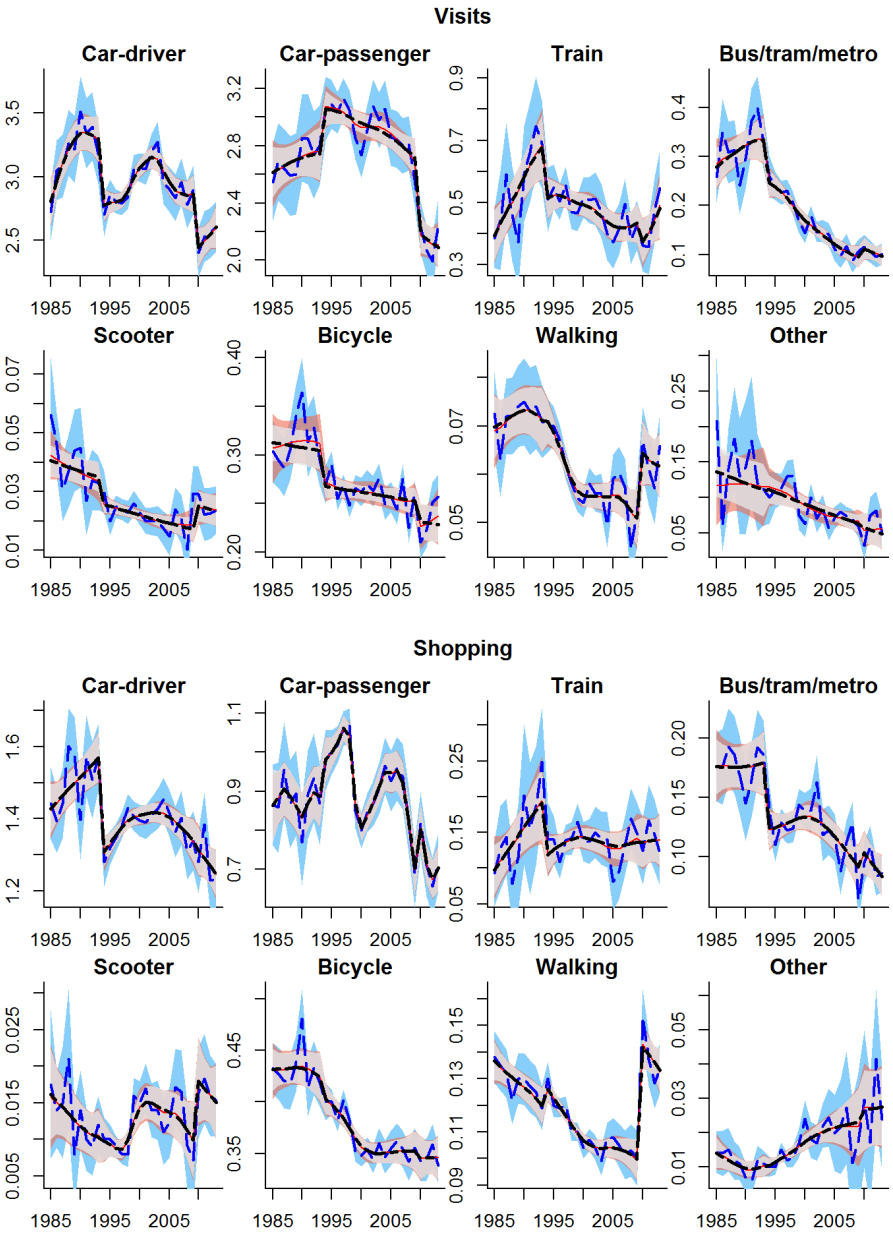
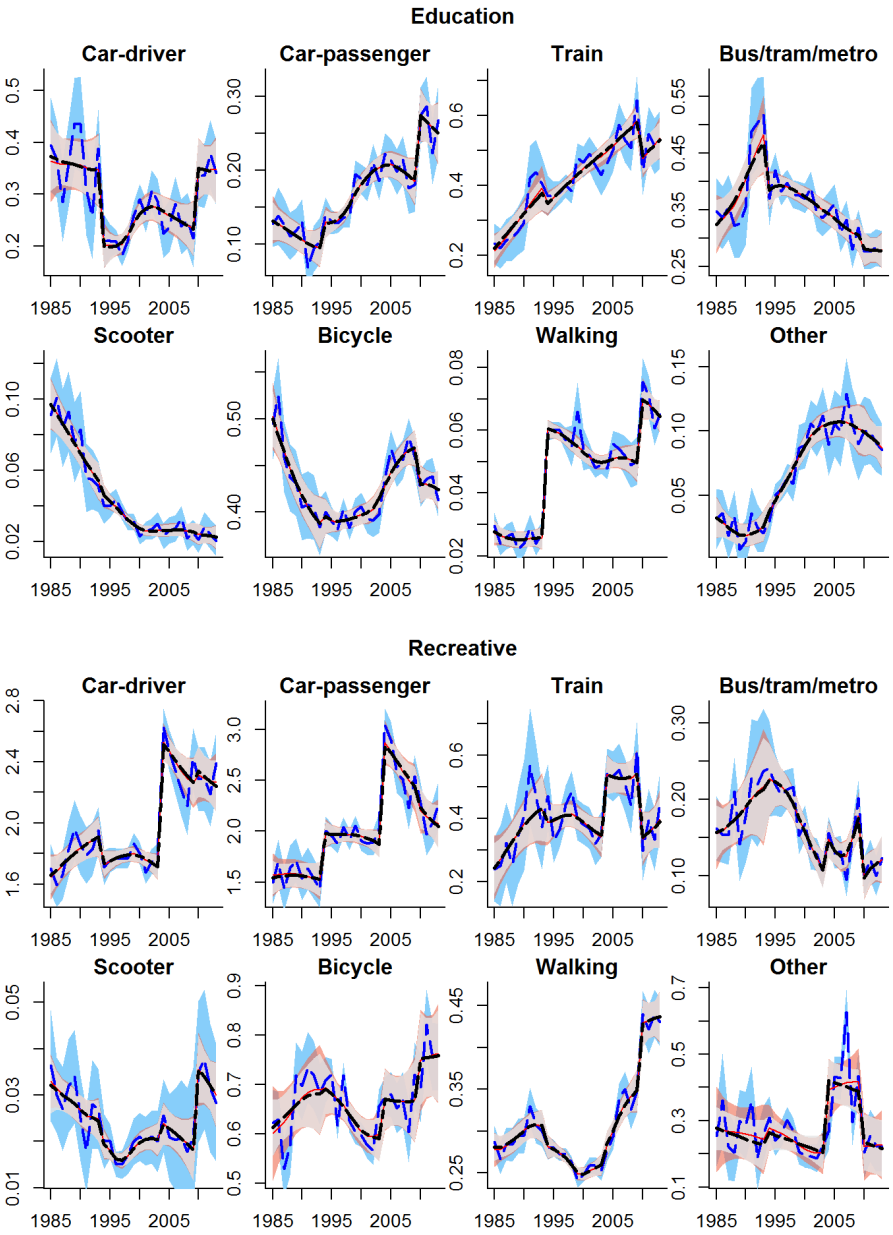


Figure 5.B.1: Provincial σ_u-estimates from univariate STS models for every intersection of motives and transport modalities.

5.C National level estimation results




Visits
Car-driver
Car-passenger
Train
Bus/tram/metro
Scooter
Bicycle
Walking
Other
Shopping
Car-driver
Car-passenger
Train
Bus/tram/metro
Scooter
Bicycle
Walking
Other
1985
1995
2005


Education
Car-driver
Car-passenger
Train
Bus/tram/metro
Scooter
Bicycle
Walking
Other
Recreative
Car-driver
Car-passenger
Train
Bus/tram/metro
Scooter
Bicycle
Walking
Other
1985
1995
2005

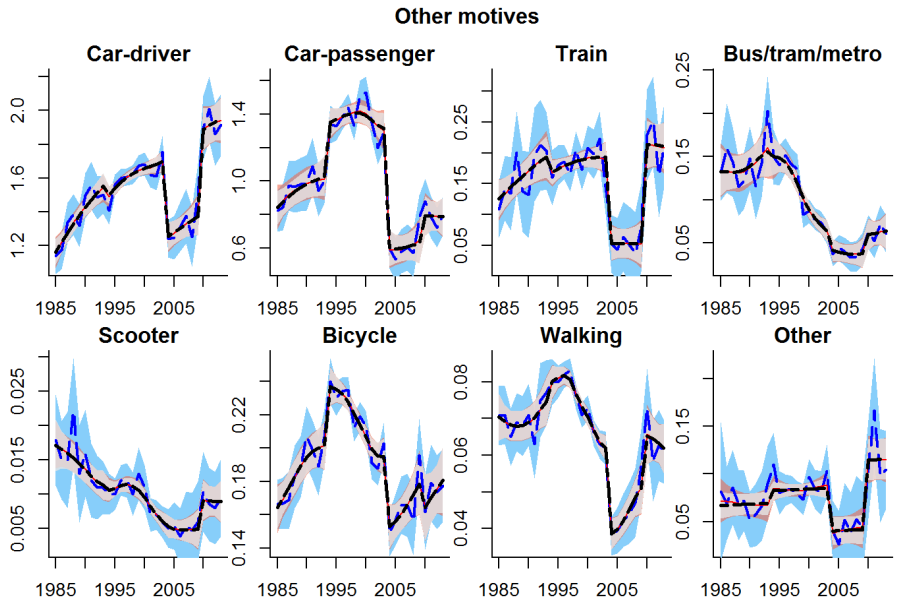


Figure 5.C.1: The national level design-based point-estimates (thin blue dashed line), estimates from the multilevel HB (solid red) and multivariate STS models (thick black dashed); the colour of the 95% confidence intervals corresponds to the colour of the point-estimates

Chapter 6

Conclusions

This last chapter is intended to provide the reader with conclusions to the present thesis. It is not my intention to repeat what has already been said at the end of each chapter, so I will rather mention the most essential concluding remarks.

This thesis dealt with time series SAE applications in three surveys repeatedly conducted at Statistics Netherlands. The main intention of each of these applications was to illustrate how time series techniques can be applied to survey data in order to solve two problems simultaneously: accounting for small sample sizes that make design-based figures unreliable, and for discontinuities that appear due to survey redesigns and hamper the comparability of published figures over time.

Chapter 2 presents both a univariate and multivariate STS approach to the DRTS. Several modifications of this survey have caused level breaks and reduced the effective sample size, which resulted in variance breaks and increasing variances of the design-based estimator. The latter made series of the design-based survey estimates too imprecise and excessively volatile. This chapter shows that standard errors of the design-based estimator can be reduced by 40 to 60 percent in large domains and/or in the aggregated series in the period following the most recent major survey redesign. When it comes to small domains, the reduction in the standard errors can reach 70 or 80 percent. Most of this variance reduction comes from exploiting the time dimension of the data. The model-based variances are further reduced by exploiting cointegration in the domain series, i.e. by modelling common stochastic factors. The above-mentioned models also account for level breaks.

In surveys like the DRTS, where estimates are produced at an aggregated (national) and domain level, a question may arise as to whether all of these series have to be included into the model. If the aggregated series' design-based estimates are sufficiently precise, they can be included into the model and be treated as external benchmarks for the sum of the domain predictions in order to secure the model against misspecification (as in Pfeffermann and Tiller (2006), Pfeffermann et al. (2014)). However, if the aggregated series is also subject to signal extraction, including it into the model by making its state variables equal to the sum of the respective domain state variables would be redundant. One could as well derive the aggregated series' point- and variance model-based estimates as a linear combination of the underlying domain (point- and covariance) estimates from a $D-$dimensional model for the domains. Chapter 3 shows, however, that both point- and variance estimates from a $D-$ and $(D+1)-$dimensional model may differ, but only under unknown (and estimated) hyperparameters in the Kalman filter recursions. If the true hyperparameters are used by the Kalman filter, the two models produce identical outcomes, as should be expected. Therefore, a more parsimonious $D-$dimensional model should be considered, from which the aggregated series estimates can be derived. One may still be tempted to include the aggregated series, which has a better signal-to-noise ratio, into the model and omit any other domain series in order to stay away from the covariance matrix singularity which appears if the aggregated series is indeed the sum of the underlying domain series. However, this approach has several drawbacks. Firstly, relatively minute under- or overestimates in the aggregated series may translate into huge misestimates in the domain that has been omitted from the model. Survey modifications in the aggregated series can be also better accounted for at a lower aggregation (domain) level. Secondly, if covariances between the domain and aggregated series are not duly taken into account when estimating time-varying hyperparameters in the observation equation, the signal variances of the series which has been omitted are likely to be overestimated.

Design-based variance breaks described above, and thus breaks in the variance of the measurement equation error term, can easily be taken into account if the design-based variance estimates are available and fed into the model. Such prior information is, however, absent in the application presented in Chapter 2. As a solution to this problem, the chapter suggests an approach where average design-based variance values would have to be estimated as hyperparameters for several

time-periods between survey redesigns. There will be difficulties, though, if another variance break occurs at the end of the time series. Therefore, it is important that NSIs produce and save information on design-based variance estimates. The LFS model presented in Chapter 4 is fitted by Van den Brakel and Krieg (2009a) by using design-based variance estimates as prior information. The two DTS time series models presented in Chapter 5 also use design-based variance estimates as prior information, though with many missing values in the series of these variance estimates.

This thesis also casts some light/shadow on theoretical methods for MSE estimation in STS models. The problem with the MSE estimation within the STS framework is that the hyperparameter uncertainty is usually ignored, which is also the case with the Dutch LFS model used by Statistics Netherlands for production of official monthly labour force figures. This renders MSE estimates negatively biased when series are not long enough, which may be a serious issue when it comes to such important economic indicators as unemployment. The performance of the existing MSE estimation methods has been studied in Chapter 4 using the DLFS model as the data generation process in a Monte-Carlo study. This study reveals that the asymptotic approximation as in Hamilton (1986) is not applicable to cases with hyperparameters close to zero due to failures when inverting the information matrix of the hyperparameter estimates. The simulation results also suggest that the non-parametric bootstraps, being free of normality assumptions about the error distribution, perform better than their parametric counterparts in both Pfeffermann and Tiller (2005) and Rodriguez and Ruiz (2012) methods. A more important finding, however, is that the Pfeffermann and Tiller (2005) bootstrap approaches with their positive biases consistently outperform the respective approaches of Rodriguez and Ruiz (2012), where the biases are generally negative and larger than those of the Kalman filter in absolute terms. This is contrary to the claim of Rodriguez and Ruiz (2012) about the superiority of their method in short time series. Hence, the PT-methods, being theoretically sound unlike their RR-counterparts, should be considered for other survey data too, despite the fact that the former methods may occasionally be overperformed by the latter ones. It is also important that the simulation smoother correction procedure of Durbin and Koopman (2002) is used in both bootstraps when constructing non-stationary bootstrap series. Another finding described in Subsection 4.3.2 may be important for practitioners wishing to estimate the MSE with the help of a bootstrap. The

shortcut suggested by Pfeffermann and Tiller (2005) for their parametric bootstrap in equation (9) of their article should not be used for non-stationary processes. Instead, the parametric PT bootstrap should be implemented in the same way as the non-parametric one, differing only in the way the bootstrap series are generated.

Chapter 4 also shows that simulating a model can be a good tool to check the model for overspecification by constructing the hyperparameters' density functions. Such procedure suggests that it might be worth considering a more restricted version of the DLFS model, by restricting one of its hyperparameters to zero. The Kalman filter-based MSEs in both variants of the DLFS model do not raise high concerns with biases slightly higher than -2%, based on nine and a half years. With more than sixteen years, the biases are expected to be around -1.3%, which does not constitute a problem for the model-based approach in the production of official figures.

Apart from STS models, time series extensions of multilevel models are also frequently used in repeatedly conducted surveys. Chapter 5 is called to compare an area-level (Fay-Herriot) multilevel time series model with an STS one. Here, the multilevel model has been estimated with the Hierarchical Bayesian approach (HB model). Similarly to STS models, multilevel models contain unknown hyperparameters, for which hyperpriors are adopted in the Bayesian setting. In this way, the uncertainty around the hyperparameter estimates is summarised by the variance of their posterior distribution. Apart from this, the multilevel model differs from the conventional STS one in that some or all of the time-invariant effects are treated as random. In the case of the DTS, however, assumptions about random effects sharing the same variance across domains turned out to be inappropriate at the provincial level, so another model modification (HB-FE model) has been considered where these random components have been replaced by fixed ones. Not only does the HB-FE model make the comparison of the multilevel and STS approaches more straightforward, but it also turned out to be the only specification that provides an adequate fit to the data at the provincial level.

The comparison between the STS and the HB-FE model in terms of estimated signal variances is still not completely fair due to the ignored hyperparameter uncertainty in the STS model. This comparison rather gives an idea about the scale of the MSE negative bias in the STS model. This bias could be particularly

large if variance hyperparameters are close to their boundary values, with heavy right tails in the posterior distributions/distributions of the ML estimator. This is the case in many small-scale domains of the DTS, which makes the hyperparameter uncertainty the primary source of differences between the two modelling approaches in the case of the DTS. Secondly, another source of differences may appear when the signal posterior distributions are skewed, thus producing larger posterior standard deviations for domain predictors than the signal standard errors produced by a linear STS model relying on the assumption of normality in the data/disturbances. In the DTS, slight skewness in the signal posterior is observed at the provincial level, mainly in those domains whose trend disturbance variance posteriors are skewed enough to feed some degree of asymmetry through to the signal posterior. At the national level of the DTS, skewness turned out to have a negligible effect.

It turns out that both point- and variance model-based estimates produced with the STS and HB-FE techniques are very similar, with differences becoming visible mainly in small domains. At the provincial level, out of the 9.4% average difference between standard errors of the STS and HB-FE model, the hyperparameter uncertainty seemed to be responsible for about 7.9%, and signal skeweness - for the remaining 1.5% difference. The averages are taken across time and domains. Quantification of the hyperparameter uncertainty has been performed within the HB multilevel framework by imposing informative priors on the trend disturbance variances with a large number of degrees of freedom and with the scale parameter taken equal to the ML estimates from the STS multivariate model. For the above-mentioned reasons, one should be aware of the limitations of the frequentist-based STS approach in short time series or series that feature small variance hyperparameters. In such cases, the negative biases in STS-based variances should be accounted for by means of an additional procedure, such as the bootstrap of Pfeffermann and Tiller (2005) studied in Chapter 4.

When the standard errors of design-based estimators are available, like in the case of the DTS, analysing their reduction through the time-series model-based methods becomes even more straightforward compared to when the design-based variances are approximated with the variance hyperparameter of the measurement equation error term (the way it has been done in Chapter 2 of this thesis). The HB-FE model offers an above-50%-average reduction in the design-based standard

errors at the provincial level and above 30% at the national level (the figures for the STS model look slightly better due to the reasons already mentioned above). In other words, in order to reduce the design-based variances to the level of the model-based variances at the provincial level, one would have to increase the sample size more than four-fold, and more than twice at the national level (conditional on the point-estimates). Therefore, time series techniques should be adopted at NSIs as soon as the time series becomes sufficiently long to make a switch to the model-based approach in production of official statistical figures.

To reach the best results, it may be worth combining the multilevel and STS modelling approaches. The multilevel approach presented here relies on the STS one in several respects. Missing values in the design-based variances that are used as input information in both multilevel and STS models are not the only problem that can be easily tackled with the help of simple STS models. Another problem is that even those design-based variance estimates that are available may be unreliable. First of all, they are subject to sampling volatility, so the Kalman filter can be applied to smooth them for further use as input information in the multilevel and STS models. Apart from that, missing design-based variance estimates can be imputed with the help of the Kalman filter. Secondly, design-based variance estimates could be very unstable in the case of small domains. In short times series, it could happen that all of these variance estimates have a bias of the same sign. If the bias is negative, for instance, then treating these estimates as the true variances in a multilevel setting results in model overfit by putting too much weight on the design-based estimates. A suggestion is made in Chapter 5 to use STS models in order to scale the design-based variance estimates in the right direction before they are used in multilevel models.

In Chapter 5, the time dimension has been exploited for variance and volatility reduction in the point-estimates, as well as for level break estimation. As for the spatial dimension, so far it has been used only for getting rid of some idiosyncracies through pooling trend variances across provinces. However, it can also be used as another source of variance reduction in model-based estimates, the way it is done in the case of the DRTS, as described in Chapter 2. Spatial correlations between domains belonging to different provinces can be exploited, e.g., by allowing for f common stochastic trends shared by more than f domain trends. In the case of the DTS, this approach seems to be worth exploring, since the pattern of

trends belonging to a certain motive-modality intersection is very similar among the twelve provinces of the Netherlands.

Time series techniques presented in this thesis are, of course, not exhaustive. Further, STS models of the form presented here apply only to linear Gaussian processes. When dealing with non-linear transition and/or observation functions, one would have to resort to the extended Kalman filter. For non-Gaussian processes (e.g., processes for count data or with t-distributed disturbances in the case of outliers), one has to consider conjugate filters that hinge on simulation techniques (e.g., Markov Chain Monte Carlo methods, importance sampling). As for the multilevel models, the Gibbs sampler considered in Chapter 5 is particularly useful for high dimensional problems and can be easily applied to normally distributed data. In case of non-standard posteriors, though, other resampling methods have to be applied.

Many official statistical applications, however, can be largely improved with the help of the time series techniques considered in this thesis and do not require the more sophisticated methods mentioned above. What they do require from NSIs is certain degree of expertise and availability of design-based estimates for a decent number of time periods, so that a time-series model can be set up. NSIs should be made aware of the gains in terms of comparability of figures over time together with reduced variances that do not require sample size increases. The latter must be of a particular importance at the time when most NSIs have to deal with reducing budgets. Hopefully, this thesis provides practitioners with useful guidelines, and survey methodologists with sufficient evidence about the great potential of the time series techniques in repeatedly conducted surveys.

Bibliography

Abraham, B. and Vijayan, K. (1992). Time series analysis for repeated surveys. *Communications in Statistics-Simulation and Computation*, 21(3):893–908.

Ansley, C. F. and Kohn, R. (1986). Prediction mean squared error for state space models with estimated parameters. *Biometrika*, 73(2):467–473.

Bailar, B. (1975). The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70:23–30.

Bartlett, M. S. (1946). On the theoretical specification and sampling properties of autocorrelated time-series. *Supplement to the Journal of the Royal Statistical Society*, 8(1):27–41.

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83(401):28–36.

Binder, D. and Dick, J. (1989). Modelling and estimation for repeated surveys. *Survey Methodology*, 15(1):29–45.

Binder, D. and Dick, J. (1990). A method for the analysis of seasonal ARIMA models. *Survey Methodology*, 16:239–253.

Blight, B. and Scott, A. (1973). A stochastic model for repeated surveys. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 61–66.

Bollineni-Balabay, O., Van den Brakel, J., and Palm, F. (2015). Accounting for hyperparameter uncertainty in small area estimation based on state-space models. *Discussion paper, Statistics Netherlands, Heerlen.* `http://www.cbs.nl/NR/rdonlyres/B2E54E0F-C791-4665-9830-95EDF4D5A198/0/2015accountinghyperparameteruncertaintyinsmallsarea.pdf`.

Bollineni-Balabay, O., Van den Brakel, J., and Palm, F. (2016a). Multivariate state space approach to variance reduction in series with level and variance breaks due to survey redesigns. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 179(2):377–402.

Bollineni-Balabay, O., Van den Brakel, J., Palm, F., and Boonstra, H. J. (2016b). Multilevel hierarchical bayesian vs. state space approach in time series small area estimation: the dutch travel survey. *Discussion paper, Statistics Netherlands, Heerlen.* `http://www.cbs.nl/NR/rdonlyres/17C87FF3-885E-4F9A-B4B0-64BBCCBC7A12/0/2016DP03multilevelhierarchical.pdf`.

Boonstra, H. J. (2015). *mcmcsae: MCMC small area estimation.* R package version 0.5.

Box, G., Jenkins, G., and Reinsel, G. (2011). *Time Series Analysis: Forecasting and Control*, volume 734. John Wiley & Sons.

Clark, T. S. and Linzer, D. A. (2015). Should I use fixed or random effects? *Political Science Research and Methods*, 3:399–408.

Cochran, W. (1977). *Sampling Techniques.* New York, Wiley and Sons.

Datta, G. S., Lahiri, P., Maiti, T., and Lu, K. L. (1999). Hierarchical bayes estimation of unemployment rates for the states of the us. *Journal of the American Statistical Association*, 94(448):1074–1082.

Doornik, J. (2007). *An Object-Oriented Matrix Programming Language Ox 5.* Timberlake Consultants Press, London.

Doornik, J. and Hansen, H. (2008). An omnibus test for univariate and multivariate normality. *Oxford Bulletin of Economics and Statistics*, 70(s1):927–939.

Durbin, J. and Koopman, S. J. (2002). A simple and efficient simulation smoother for state space time series analysis. *Biometrika*, 89:603–615.

Durbin, J. and Koopman, S. J. (2012). *Time Series Analysis by State Space Methods.* Number 38. Oxford University Press.

Durbin, J. and Quenneville, B. (1997). Benchmarking by state space models. *International Statistical Review*, 65(1):23–48.

EUROSTAT (2014). ESS handbook for quality reports - 2014 edition. http://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/KS-GQ-15-003.

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

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–534.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2014). *Bayesian Data Analysis*, volume 2. Taylor & Francis.

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

Gelman, A., Van Dyk, D. A., Huang, Z., and Boscardin, J. W. (2008). Using redundant parameterizations to fit hierarchical models. *Journal of Computational and Graphical Statistics*, 17(1).

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):721–741.

Ghosh, M. and Lahiri, P. (1987). Robust empirical bayes estimation of means from stratified samples. *Journal of the American Statistical Association*, 82(400):1153–1162.

Gonzalez, M. (1973). Use and evaluation of synthetic estimates. *Proceedings of the Social Statistics Section, American Statistical Association*, pages 33–36.

Hall, P. and Martin, M. (1988). On bootstrap resampling and iteration. *Biometrika*, 75:661–671.

Hamilton, J. (1986). A standard error for the estimated state vector of a state-space model. *Journal of Econometrics*, 33:387–397.

Hannan, E. J., Terrell, R., and Tuckwell, N. (1970). The seasonal adjustment of economic time series. *International Economic Review*, pages 24–52.

Harrison, P. (1967). Exponential smoothing and short-term sales forecasting. *Management Science*, 13(11):821–842.

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, Cambridge.

Harvey, A. C. (2001). Trend analysis. *Encyclopedia of Environmetrics*.

Harvey, A. C. and Chung, C. H. (2000). Estimating the underlying change in unemployment in the uk. *Journal of the Royal Statistical Society: Series A*, 163(3):303–309.

Harvey, A. C. and Durbin, J. (1986). The effects of seat belt legislation on british road casualties: A case study in structural time series modelling. *Journal of the Royal Statistical Society. Series A (General)*, pages 187–227.

Harvey, A. C., Koopman, S. J., and Penzer, J. (1998). Messy time series: a unified approach. *Advances in Econometrics*, 13:103–144.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering*, 82(1):35–45.

Koopman, S. J. (1997). Exact initial kalman filtering and smoothing for nonstationary time series models. *Journal of the American Statistical Association*, 92(440):1630–1638.

Koopman, S. J. and Durbin, J. (2000). Fast filtering and smoothing for multivariate state space models. *Journal of Time Series Analysis*, 21:281–296.

Koopman, S. J., Harvey, A. C., Doornik, J., and Shephard, N. (2009). *STAMP: Structural Time Series Analyser, Modeller and Predictor*. Timberlake Consultants Press, London.

Koopman, S. J., Shephard, N., and Doornik, J. (1999). Statistical algorithms for models in state space form using ssfpack 2.2. *Econometrics Journal*, 2:113–66.

Koopman, S. J., Shephard, N., and Doornik, J. (2008). *SsfPack 3.0: Statistical Algorithms for Models in State Space Form.* Timberlake Consultants Press, London.

Krieg, S. and Van den Brakel, J. (2012). Estimation of the monthly unemployment rate for six domains through structural time series modelling with cointegrated trends. *Computational Statistics & Data Analysis*, 56(10):2918–2933.

Kumar, S. and Lee, H. (1983). Evaluation of composite estimation for the Canadian labour force survey. *Survey Methodology*, 9:178–201.

Lemaître, G. and Dufour, J. (1987). An integrated method for weighting persons and families. *Survey Methodology*, 13(2):199–207.

Lind, J. T. (2005). Repeated surveys and the kalman filter. *The Econometrics Journal*, 8(3):418–427.

Muth, J. F. (1960). Optimal properties of exponentially weighted forecasts. *Journal of the American Statistical Association*, 55(290):299–306.

Narain, R. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3:169–174.

Nyblom, J. and Harvey, A. C. (2001). Testing against smooth stochastic trends. *Journal of Applied Econometrics*, 16(3):415–429.

Orphanides, A. and Van Norden, S. (2002). The unreliability of output-gap estimates in real time. *Review of Economics and Statistics*, 84(4):569–583.

Pfeffermann, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business and Economic Statistics*, 9:163–175.

Pfeffermann, D. (2002). Small area estimation: New developments and directions. *International Statistical Review/Revue Internationale de Statistique*, pages 125–143.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, 28:40–68.

Pfeffermann, D. and Bleuer, S. (1993). Robust joint modelling of labour force series of small areas. *Survey Methodology*, 19:149–163.

Pfeffermann, D. and Burck, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, 16:217–237.

Pfeffermann, D., Feder, M., and Signorelli, D. (1998). Estimation of autocorrelations of survey errors with application to trend estimation in small areas. *Journal of Business and Economic Statistics*, 16:339–348.

Pfeffermann, D., Sikov, A., and Tiller, R. (2014). Single- and two-stage cross-sectional and time series benchmarking procedures for small area estimation. *TEST*, 23(4):631–666.

Pfeffermann, D. and Tiller, R. (2005). Bootstrap approximation to prediction MSE for state-space models with estimated parameters. *Journal of Time Series Analysis*, 26:893–916.

Pfeffermann, D. and Tiller, R. (2006). Small-area estimation with state-space models subject to benchmark constraints. *Journal of the American Statistical Association*, 101(476):1387–1397.

Planas, C., Roeger, W., and Rossi, A. (2013). The information content of capacity utilization for detrending total factor productivity. *Journal of Economic Dynamics and Control*, 37(3):577–590.

Polson, N. G., Scott, J. G., et al. (2012). On the half-cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4):887–902.

Prasad, N. and Rao, J. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American statistical Association*, 85(409):163–171.

Rao, J. (2003). *Small Area Estimation*. John Wiley & Sons.

Rao, J. (2011). Impact of frequentist and bayesian methods on survey sampling practice: A selective appraisal. *Statistical Science*, pages 240–256.

Rao, J. and Molina, I. (2015). *Small Area Estimation*. John Wiley & Sons.

Rao, J. and Yu, M. (1994). Small-area estimation by combining time-series and cross-sectional data. *Canadian Journal of Statistics*, 22(4):511–528.

Rodriguez, A. and Ruiz, E. (2012). Bootstrap prediction mean squared errors of unobserved states based on the Kalman filter with estimated parameters. *Computational Statistics and Data Analysis*, 56:62–74.

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

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

Scott, A. and Smith, T. (1974). Analysis of repeated surveys using time series methods. *Journal of the American Statistical Association*, 69(347):674–678.

Scott, A., Smith, T., and Jones, R. (1977). The application of time series methods to the analysis of repeated surveys. *International Statistical Review/Revue Internationale de Statistique*, pages 13–28.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Theil, H. and Wage, S. (1964). Some observations on adaptive forecasting. *Management Science*, 10(2):198–206.

Tiller, R. (1992). Time series modelling of sample survey data from the US current population survey. *Journal of Official Statistics*, 8:149–166.

Van den Brakel, J. and Krieg, S. (2009a). Estimation of the monthly unemployment rate through structural time series modelling in a rotating panel design. *Survey Methodology*, 16:177–190.

Van den Brakel, J. and Krieg, S. (2009b). Structural time series modelling of the monthly unemployment rate in a rotating panel design. *Discussion paper (09031), Statistics Netherlands, Heerlen*. `http://www.cbs.nl/NR/rdonlyres/DF041949-9046-45ED-A673-72787ECCE778/0/200931x10pub.pdf`.

Van den Brakel, J. and Krieg, S. (2015). Dealing with small sample sizes, rotation group bias and discontinuities in a rotating panel design. *Survey Methodology*, 41:267–296.

Van den Brakel, J. and Krieg, S. (2016). Small area estimation with state space common factor models for rotating panels. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

Van den Brakel, J. and Roels, J. (2010). Intervention analysis with state-space models to estimate discontinuities due to a survey redesign. *Annals of Applied Statistics*, 4:1105–1138.

You, Y. (2008). An integrated modeling approach to unemployment rate estimation for sub-provincial areas of canada. *Survey Methodology*, 34(1):19–27.

You, Y. and Chapman, B. (2006). Small area estimation using area level models and estimated sampling variances. *Survey Methodology*, 32(1):97.

Nederlandse samenvatting

Kort samengevat is het doel van dit proefschrift het verbeteren van de kwaliteit van officiële statistieken. Kwaliteit van statistische informatie heeft meerdere facetten. Dit werk richt zich op twee van de door EUROSTAT gedefinieerde kwaliteitskenmerken: precisie van geproduceerde cijfers en vergelijkbaarheid daarvan door de tijd heen.

Statistische bureaus voeren steekproefonderzoeken uit om sociaal-economische ontwikkelingen in onze samenleving zo nauwkeurig mogelijk te beschrijven. In een steekproef onderzoek wordt aan de hand van een kansmechanisme een klein deel van de doelpopulatie geselecteerd. Aan de hand van de in de steekproef getrokken elementen worden vervolgens schattingen gemaakt voor de onbekende populatie grootheden. Omdat slechts een fractie van de doelpopulatie wordt waargenomen ontstaat er altijd een afwijking tussen de schatting en de echte (maar onbekende) waarde van de populatiegrootheid. Deze onzekerheid wordt gemeten met de zogenaamde variantie van de steekproefschattingen. Deze variantie meet in hoeverre puntschattingen over alle mogelijke kanssteekproeven variëren, die volgens het steekproefontwerp zouden kunnen worden getrokken. Gebruikelijke schattingsmethodieken zijn voornamelijk gebaseerd op het kansmechanisme waarmee de steekproef uit de doelpopulatie is getrokken. Statistische modellen spelen hierbij geen of slechts in zeer beperkte mate een rol. Deze methoden worden daarom vaak design-gebaseerde schattingsmethoden genoemd. Het voordeel van deze methoden is dat ze niet gevoelig zijn voor veronderstellingen van een gekozen statistisch model. Bij steekproeven van voldoende omvang resulteren deze schattingsmethoden in uitkomsten die in verwachting overeen komen met de echte populatiewaarden. Dit soort eigenschappen heeft er in het verleden voor gezorgd dat het gebruik van design-gebaseerde schattingsmethoden bij statistische bureaus erg populair is ge-

worden.

Het nadeel van design-gebaseerde schattingsmethoden is dat bij kleine steekproeven de varianties van de schatters onacceptabel groot worden. Dit probleem ontstaat indien schattingen voor gedetailleerde uitsplitsingen moeten worden gemaakt, bijvoorbeeld een uitsplitsing van een variabele naar de afzonderlijke gemeenten. Deelpopulaties waarvoor schattingen gemaakt moeten worden, worden vaak kortweg aangeduid als domeinen. Indien de steekproefomvang van dergelijke domeinen te klein is om met design-gebaseerde methoden voldoende betrouwbare schattingen te maken, spreekt men ook wel over kleine domeinen. Soms komt het voor dat een domein helemaal geen steekproefmassa krijgt, terwijl er toch een schatting voor is vereist. Een triviale manier om de betrouwbaarheid van deze cijfers te verbeteren is het verhogen van de steekproefomvang in de domeinen waarvoor schattingen gemaakt moeten worden. Meestal is dit niet realistisch omdat dit resulteert in onacceptabel hoge kosten voor nationale statistische bureau's en in een te hoge enquêtedruk voor respondenten. Het probleem van kleine steekproefomvangen gaf in de afgelopen decennia aanleiding tot een nieuw onderzoeksgebied – small area estimation. Dit refereert naar een klasse van schattingsmethoden waarbij, via een statistisch model, de effectieve steekproefomvang van een afzonderlijk domein wordt vergroot met steekproefinformatie waargenomen in andere domeinen of voorgaande waarnemingsperioden. Bij dergelijke schattingsmethoden spelen statistische modellen een prominente rol en worden daarom ook wel modelgebaseerde schattingsmethoden genoemd.

Naast steekproeffouten, is er altijd sprake van (systematische) meetfouten bij de dataverzameling. Deze hangen af van de manier waarop respondenten worden benaderd en de manier waarop vragenlijsten worden ontworpen. Een herontwerp van de enquête resulteert daarom vaak in een verandering van deze systematische fouten en in een sprong (breuk) in de trend van de tijdreeks. Als gevolg daarvan, zijn cijfers vóór en na een herontwerp niet meer vergelijkbaar.

Een manier om beide problemen op te lossen is om tijdreeksen, waargenomen via herhaaldelijk uitgevoerde kansteekproeven, te modelleren met behulp van tijdreeksmodellen. In dit proefschrift zijn zijn structurele (state space) tijdreeksmodellen en multilevel tijdreeksmodellen gehanteerd. Hoofdstuk 2 illustreert hoe univariate (d.w.z. één tijdreeksmodel per afzonderlijk domein) en multivariate (ge-

zamenlijk modeleren van meerdere tijdreeksen) modellen kunnen worden gebruikt in een officiële statistische toepassing (namelijk, Enquête Wegvervoer van Nederland (EWN)) om variantie te reduceren en trendbreuken te schatten. De EWN toepassing laat zien dat het grootste deel van variantiereductie afkomstig is van steekproefinformatie uit voorgaande perioden (in univariate modellen). Een extra variantiereductie wordt bereikt door gebruik te maken van steekproefinformatie uit andere domeinen, door tijdreeksen van verschillende domeinen te modeleren in een multivariaat model. Bij de kleinste domeinen van de EWN worden standaardfouten tot wel 70 of 80 procent gereduceerd. Bij de grote domeinen en bij de reeks op nationaal niveau wordt een reductie van 40-60 procent bereikt.

In hoofdstuk 3 wordt nader ingegaan op de verschillende manieren waarop tijdreeksen op nationaal niveau kunnen worden geschat. Het uitgangspunt hierbij is dat met behulp van een tijdreeksmodel preciezere schattingen moeten worden verkregen maar dat deze schattingen consistent moeten zijn met de som van onderliggende domeinen. Een mogelijkheid is om de reeks op nationaal niveau aan het model toe te voegen of met deze reeks een andere domeinreeks te vervangen, omdat reeksen op een hoger aggregatieniveau een betere signaal-ruisverhouding hebben. Beargumenteerd wordt dat dit niet efficiënt is, vooral als het weggelaten domein trendbreuken heeft. De reden daarvoor is dat, bij het afleiden van modelschattingen voor het (kleine) weggelaten domein, verwaarloosbare misschattingen op geaggregeerd niveau tot enorme misschattingen op domeinniveau kunnen leiden. Voor de EWN wordt aanbevolen om schattingen op nationaal niveau af te leiden uit het multivariaat model voor de onderliggende domeinen.

State-space modellen worden vaak geanalyseerd met het Kalman filter. Een probleem bij het toepassen van de standaard Kalman filter recursie is dat de bijbehorende standaardfouten geen rekening houden met het feit dat de onbekende hyperparameters in het state-space model zijn vervangen door hun maximum likelihood schattingen. In de praktijk wordt de onzekerheid rond deze hyperparameter schattingen vaak genegeerd, wat resulteert in een onderschatting (of een negatieve vertekening) in de standaardfouten van de kleinedomeinschattingen op basis van het state-space model. Het gebruik van te optimistische betrouwbaarheidsintervallen bij belangrijke sociaal-economische indicatoren, zoals bijvoorbeeld werkloosheid, bemoeilijkt een correcte interpretatie van de resultaten voor bijvoorbeeld beleidsmakers. Het Centraal Bureau voor de Statistiek (CBS) maakt gebruik van

een state-space model voor het maken van officiële maandcijfers over de werkloze en werkzame beroepsbevolking. Standaardfouten zijn gebaseerd op de standaard Kalman filter recursie. In hoofdstuk 4 wordt onderzocht hoe groot de onderschatting is in de onzekerheid van maandcijfers over de werkloze beroepsbevolking in de Nederlandse Enquête Beroepsbevolking (EBB) ten gevolge van het vervangen van de onbekende hyperparameters door hun maximum likelihood schatters. Daartoe is een simulatie studie opgezet waarbij het tijdreeksmodel van het CBS gebruikt wordt als het data genererend proces.

Het blijkt dat de vertekening in de standaardfout bij deze toepassing niet groot is en verwaarloosbaar wordt naarmate de reeks langer wordt. De simulatie laat ook zien hoe het model beter gespecificeerd kan worden. Tenslotte, geeft de simulatie in hoofdstuk 4 nieuwe inzichten in de verschillende methoden die in de literatuur bekend zijn om rekening te houden met de onzekerheid van maximum likelihood schattingen voor de hyperparameters in state space modellen.

Zoals hierboven vermeld, zijn multilevel tijdreeksmodellen een alternatieve optie voor herhaaldelijk uitgevoerde enquêtes. De hiërarchische (volledige) Bayesiaanse (HB) aanpak, die vrij eenvoudig in het kader van multilevel modellen is te implementeren, is een generieke alternatieve aanpak om rekening te houden met alle onzekerheid, ook rond hyperparameters. In hoofdstuk 5 worden de state space en multilevel tijdreekstechnieken vergeleken op basis van het Onderzoek Verplaatsingen in Nederland (OViN). Dit onderzoek heeft last van methodebreuken ten gevolge van veranderingen in de onderliggende enquête en kleine steekproefomvangen. Het doel van het onderzoek beschreven in Hoofdstuk 5 is om na te gaan of beide methoden vergelijkbare uitkomsten leveren, en om de negatieve vertekening in de gemiddelde standaardfouten in state space modellen te kwantificeren. Het blijkt dat punt- en variantieschattingen afkomstig uit beide tijdreekstechnieken vergelijkbaar zijn. Verschillen verschijnen eerder in de geschatte varianties dan in de puntschattingen en komen alleen in kleinschalige domeinen of in domeinen met beperkt flexibele stochastische trends voor, omdat de variantie hyperparameters van de trends daar dicht bij de grens van de parameterruimte liggen. Beide methoden bieden een behoorlijke reductie in design-gebaseerde standaardfouten. Voor het OViN worden standaardfouten op nationaal niveau met circa 30 procent gereduceerd, uitgaande van het HB multilevel model. Op provinciaal niveau is de reductie circa 50 procent. Om dezelfde reductie in de standaardfouten van

de design-gebaseerde schattingsmethoden te bewerkstelligen zou de steekproefomvang meer dan verdubbeld moeten worden op nationaal niveau en meer dan vier keer zo groot moeten worden op provinciaal niveau. In het geval van OViN kan de negatieve vertekening in standaardfouten oplopen tot 8 procent op provinciaal niveau en tot meer dan 10 procent op nationaal niveau indien de standaard Kalman filter recursies gebruikt worden.

In state space modellen en multilevel modellen worden standaardfouten van de design-gebaseerde schattingen als prior informatie in het model gebruikt. Vaak zijn deze erg onbetrouwbaar en moeten daarom gesmoothd worden over de domeinen en over de tijd. In hoofdstuk 5 wordt een nieuwe manier voorgesteld om met ontbrekende en/of onbetrouwbare design-gebaseerde variantieschattingen om te gaan.

Vrijwel elke enquête die herhaaldelijk wordt uitgevoerd kan profiteren van tijdreeksmodellen beschreven in dit proefschrift. Deze beschreven technieken maken het mogelijk om de precisie van design-gebaseerde schattingen te verbeteren. Daarnaast kan rekening worden gehouden met systematische effecten op de uitkomsten die ontstaan door het veranderingen in het onderliggende enquêteproces. Dit laatste voorkomt dat cijfers voor en na een herontwerp niet meer vergelijkbaar zijn. Het gebruik van dergelijke model-gebaseerde schattingsmethode is in de productie van officiële statistieken nog vrij beperkt. Twee voorbeelden van nationale statistische bureau's die state space tijdreeksmodellen gebruiken in de productie van officiële cijfers zijn het Bureau of Labor Statistics in de Verenigde Staten en het CBS in Nederland. Hopelijk maakt dit proefschrift meer officiële statistici bewust van de potentie van state space en multilevel tijdreeksmodellen en biedt het richtlijnen voor enquête beoefenaars.

Valorisation

The well-being of a country is defined to a large extend by good governing and planning, which, in turn, is not possible without reliable quantitative and qualitative data. Collecting, summarizing and releasing such data about the country's state of affairs is usually laid over the shoulders of the country's national statistical institute (in the Netherlands, it is Statistics Netherlands, or Centraal Bureau voor de Statistiek (CBS)). If to be described in just a couple of words, the matter of this thesis is improving the quality of official statistical figures. The EUROSTAT defines several quality dimensions (EUROSTAT (2014)), of which two are considered in this work: 1. accuracy and reliability and 2. coherence and comparability.

The first problem arises because observing the whole population, i.e. census, can be very costly or even impossible and is thus hardly ever done in statistical practice. Instead, a sample is drawn that should be representative of the whole population. It is obvious that the point-estimate for the quantity of interest (e.g., for the true average or total) based on this sample is not exactly equal to the true value in the population, and drawing another sample would result in a different estimate. This variability of point-estimates across all possible samples is summarized by the (design-based) variance or standard error. The bigger the sample size is, the closer we expect the estimate to be to the true value, and hence, the smaller the variance is. High design-based variances can be detected in time series with the naked eye, as they result in excessively volatile point-estimates.

Sometimes, estimates for small areas or domains are required (and the demand on these has been growing in recent years), where the sample size may be so small (could be even zero) that the estimated variance is unacceptably large. Think, for instance, of an estimated unemployment rate of $3(\pm 4)\%$. Such a huge uncertainty around the point-estimate tells us virtually nothing about the existing

unemployment rate. But even if it is estimated more precisely up to ±1%, it is still not precise enough, since a two-percent uncertainty around the unemployment rate makes it difficult for policy makers to decide on their fiscal and monetary instruments. The confidence interval bands of ±0.1%, as one usually sees in unemployment statistics produced by EUROSTAT, are achievable but require interviewing much more people. This means an increase in public money expenditures due to additional interviewing, as well as in the total administrative burden respondents experience by participating in the survey. The problem with comparability stems from the fact that surveys sooner or later get redesigned due to cost of quality considerations. A change in a way data are collected renders the figures before and after a redesign incomparable over time. As one can see, NSIs produce estimates contaminated with different kinds of errors, while being a respectable source of data for policy makers or non-governmental researchers who usually treat these estimates as the true data in their econometric models or other kinds of analytics. Therefore, it is of utmost importance that data produced by NSIs are of best possible quality. Nowadays, when NSIs' budgets are shrinking and when the society is becoming ever more averse to administrative burden they experience from their NSIs, increasing the sample size is not the best of even feasible option, so statisticians have to look for more sophisticated methods to meet both constraints. In fact, NSIs dispose of lots of information that can be used in improving newly-produced figures without resorting to trivial sample size increases. It is obvious that information on the same quantities of interest from the past or from similar domains can be of a great use, but is usually discarded, though. With this dissertation, I have tried to raise the awareness about the potential of time series modelling (namely, state space and multilevel time series) techniques that address both problems of small sample sizes and incomparability. Committing to these techniques would put the production of official statistics on the rails of model-based, rather than design-based approach. Most NSIs in the world are, however, conservative in this respect and prefer relying on traditional (assumption-free) methods based on the survey design in production of their figures. This is understandable, because a bias-free implementation of new methods requires a careful model selection, while the penetration of ideas about the true power of these methods among official statisticians is not high. The knowledge transfer from academia to NSIs is limited because it often requires routine work from academics, as well as certain degree of expertise in survey methodology. In order to fill this gap, Statistics Netherlands tries to strengthen its collaboration

with academia by establishing professorships and funding PhD research projects, like the present one. This dissertation is aimed at illustrating how information accumulated over time and space can be exploited in repeatedly conducted surveys to improve the quality of official statistical figures in terms of precision and comparability over time. As this research is conducted thanks to and in the interests of Statistics Netherlands, the objective has been to develop suitable time series models for several data sets for repeatedly conducted surveys with several survey redesigns. Variance reduction runs as a red thread through the whole dissertation, and in Chapters 2 and 5 it is the central issue. Chapter 2 applies the state space model-based approach to the Dutch Road Transportation survey that features several survey redesigns and insufficiently large sample sizes. It is shown that most variance reduction comes from borrowing information over time, i.e. even when domains are modelled individually. Jointly modelling all the domains in one model by exploiting the correlation between them, i.e. borrowing information over space, further reduces the variance of the newly produced time series. The resulting standard errors of the series are eventually 40 to 70 percent smaller than the design-based ones, with smaller domains experiencing more variance reduction compared to the larger ones. The coherence between estimated domain series, on the one hand, and the series that constitutes the sum thereof, on the other hand, is addressed in Chapter 3 based on the same survey. State space models are usually estimated with the help of the well-known to engineers Kalman filter. The problem with these models, despite their indisputable power, is that in practice the Kalman filter relies on estimated hyperparameters, rather than on their true values. Since there must always be a certain amount of uncertainty around any estimate, it should be taken into account; otherwise, the estimated variances of the newly produced series will be negatively biased (underestimated). The literature knows several methods to correct for the bias in variances estimated with the Kalman filter. However, before any newly developed theory can be applied in the regular production process, its value added and flaws must be verified. Usually, the performance of new methods is illustrated with simple examples where the superiority of the method is salient and unambiguous. The same methods may not, however, perform as well in real-life examples where different factors play a role. Therefore, one of the chapters in this dissertation (Chapter 4) is devoted to a comparative simulation study of several existing variance approximation methods in state space models. For this purpose, a real-world complex application is chosen: the Dutch Labour Force survey (DLFS) model that is used

by Statistics Netherlands in the production of official statistical figures on the unemployed labour force. Apart from casting some light/shadow on these variance approximation methods, this study also suggests that simulating a model can be a good tool to check the model for overspecification. Such procedure suggests that it might be worth considering a more restricted version of the DLFS model. Biases in the Kalman filter-based variances in both variants of the DLFS model do not raise high concerns and become negligible as the time progresses, which makes the state space approach safe and attractive for use in the production of official figures. Another way to account for the uncertainty around (unknown) estimated hyperparameters is a full (hierarchical) Bayesian paradigm. Chapter 5 compares the state space approach with the multilevel Hierarchical Bayesian (HB) one in another official statistical application – the Dutch Transportation survey (DTS). It turns out that outcomes from the two approaches are quite comparable, with differences becoming visible mainly in small-scaled domains and applicable mainly to variance-, rather than point-estimates. Unlike in the DLFS, negative biases in standard error estimates produced by the Kalman filter are considerable in the case of the DTS reaching almost 8% at the provincial level and more than 10% at the national one. The more conservative standard error estimates from the multilevel HB model still offer a significant reduction in the design-based standard errors in the DTS: above 50% at the provincial level, and over 30% at the national level, averaged over the domains and time. In other words, in order to reduce the true variance within the design-based framework to the extent the time series techniques do, one would have to increase the sample size more than four-fold for the provincial level, and more than twice for the national level (conditional on the point-estimates). Chapter 5 also presents an approach for dealing with unreliable and missing design-based variance estimates that are used as input information in both multilevel and state space models. Many official statistical applications can be largely improved with the help of the time series techniques considered in this dissertation. At Statistics Netherlands, the US Census Bureau and the UK Office for National Statistics, certain attempts have been made to move further than the traditional sampling theory. As mentioned above, Statistics Netherlands is already producing official figures for the Labour Force survey based on the state space model. For the DTS presented in Chapter 5, certain attempts are being made to switch to the model-based times series approach in the production of official figures, as this survey is going to redesigned again in 2018. However, much more needs to be done in popularizing various modelling techniques among official

statisticians. The applications presented in this dissertation have demonstrated how information accumulated over time and optionally domain space can be successfully used to increase the precision of survey estimates without resorting to (expensive) sample size increases, as well as in order to restore the comparability of figures before and after survey redesigns. The former must be of a particular importance at the time when most NSIs have to deal with reducing budgets. Examples of successful implementation of novel modelling methods and dissemination of these results is therefore crucial for advancement of official statistics. Hopefully, this dissertation provides practitioners with useful guidelines, and survey methodologists with sufficient evidence about the great potential of the time series techniques in repeatedly conducted surveys.

Curriculum Vitae

Oksana Balabay was born on April 1, 1988 in Poltava, Ukraine. After finishing Kupyansk grammar school n.1 in 2005, she got her bachelor's and master's degrees in Banking from V.N. Karazin Kharkiv National University, Ukraine. In 2011, she gained her MSc in Business and Economics from Uppsala Univeristy, Sweden. In December of the same year, she started a PhD programme under the guidance of Prof. Dr. Jan van den Brakel and Prof. Dr. Franz Palm. It was a common project of Statistics Netherlands, where Oksana is working now, and the Department of Quantitative Economics of the Maastricht University School of Business and Economics. Oksana presented her works at several international conferences: the 2013 and 2015 New Techniques and Technologies for Statistics (NTTS) biannual conference organised by Eurostat in Brussels, the First Asian ISI Satellite Meeting on Small Area Estimation (SAE 2013, Bangkok, Thailand), the First Latin American ISI Satellite Meeting on Small Area Estimation (SAE 2015, Santiago, Chile), and the International Work-Conference on Time Series (ITISE 2014, Granada, Spain). Her works can be found in international academic journals, as well as on Statistics Netherlands' official website.

Oksana Balabay started her PhD programme in Dec 2011 under the guidance of Prof. Dr. Jan van den Brakel and Prof. Dr. Franz Palm. The objective of this thesis was to develop suitable time series models for repeatedly conducted surveys that suffer from unreliability due to small sample sizes and survey redesigns. Variance reduction and modelling level breaks caused by survey redesigns run as a red thread through the whole dissertation. Chapter 2 presents a comparison between the univariate state space modelling approach and the multivariate one with common factors in the Dutch Road Transportation survey. Chapter 3 ponders how an aggregated series should be modelled to remain coherent with its underlying domains. Chapter 4 presents a comparison of (mainly) bootstrapping approaches to the MSE estimation in state space models that treat estimated hyperparameters as known (based on the Dutch Labour Force survey). Chapter 5 is a comparative study of the state space approach based on the Kalman filter and of the multilevel Hierarchical Bayesian approach (applied to the Dutch Transportation survey). The Valorization section is meant for out-of-the-field readers, and the Dutch summary for Dutch speakers.

"Education is the most powerful weapon which you can use to change the world."

"He uses statistics as a drunken man uses lamp posts - for support rather than for illumination."

"You can't fix by analysis what you bungled by design."