MAKSWELL
MAKing Sustainable development and WELL-being frameworks work for policy analysis

www.makswell.eu

Horizon 2020 - Research and Innovation Framework Programme
Call: H2020-SC6-CO-CREATION-2017
Coordination and support actions (Coordinating actions)

Grant Agreement Number 770643

Work Package 4
Multivariate time series methodology to be applied to well-being indicators and SDGs

Deliverable 4.3

Report on survey discontinuities

October 2019

Destatis, HSCO, Istat, Statistics Netherlands, Pisa University, Southampton University, Trier University

This project has received funding from the European Union's Horizon 2020 research and innovation programme.

MAKSWELL
MAKing Sustainable development and WELL-being frameworks work for policy analysis

Deliverable D4.3

Report on survey discontinuities

Authors

Statistics Netherlands:
Jan van den Brakel and Sabine Krieg

Southampton University:
Paul Smith and Nikos Tzavidis

Trier University:
Florian Ertz and Ralf Münnich

MAKSWELL
MAKing Sustainable development and WELL-being frameworks work for policy analysis

Summary

The MAKSWELL project was set up to help strengthen the use of evidence and information on well-being and sustainability for policy-making in the EU, as the political attention to well-being and sustainability indicators has also been increasing in recent years. Traditionally sample surveys provide the primary source of data that are used for measurement frameworks for well-being and sustainability. However, survey organisations and national statistical institutes frequently review and occasionally change their approaches to collecting survey data. Although such changes are motivated by the need for a more efficient approach to allocating resources, they can lead to breaks in the series of published estimates known as discontinuities. This report presents approaches to estimating and adjusting for discontinuities in survey data. Alternative methodological approaches are presented, some of which assume the availability of experimental or pilot data under the new design, whilst other approaches make use of time series methods for quantifying discontinuities and hence not requiring the use of pilot data. The methods we present are illustrated using real data from the UK and the Netherlands. The targets of estimation are defined both at national (aggregate) level and subnational (domain) levels. The report presents estimation and inference both under the design-based and model-based frameworks. The methodological tools are general and therefore transferable to other survey settings.

1. Introduction **1**

2. Methods for estimating discontinuities: A review **3**

2.1. Introduction 3

2.2. Parallel run 3

2.2.1. Model variants and extensions 8

2.3. Structural time series model 9

2.4. Combining a parallel run with a time series 11

2.5. Adjustment methods 12

3. Data sources and applications in the presence of a parallel run: The case of surveys in Wales **13**

3.1. Survey data from Wales 13

3.2. Application to the Welsh Surveys 15

4. Data sources and applications using structural time series with no parallel run: The UK International Passenger Survey **17**

4.1. UK data from the International Passenger Survey 17

4.2. Application to the UK International Passenger Survey 17

5. Data sources and applications using a structural time series and a parallel run: The Dutch Consumer Survey **22**

5.1. The Dutch Consumer Survey 22

5.1.1. Parallel run 22

5.1.2. Backcast method 24

5.1.3. STM 25

5.1.4. Combination 28

5.1.5. Backcasting 29

6. Summary **31**

1. Introduction

Surveys are susceptible to a wide range of different types of errors Groves (2004). There has been a traditional, theoretical focus on sampling errors that is, on how much an estimate is likely to differ from the true value because we use a sample rather than observing the whole population. But in recent years there has been a bigger emphasis on errors arising from the use of questionnaires and measurement errors, which cover a range of types of issues, from non-response and processing (including data entry and scanning) to data editing and a range of other processes. Many surveys maintain consistent methodologies over quite long periods and one of the drivers behind this decision is to keep these errors as consistent as possible between different instances of the survey. This means that estimates of change will be approximately unbiased and this feature is of some significance for users of surveys who therefore prefer to avoid changes (Van den Brakel et al., 2008). It follows that changes to survey procedures may have an effect on estimates, and it is these types of effects which are normally called discontinuities. This may be due to the fact that the mode has changed from face-to-face/use of CAPI to the use of the web and it is known that the presence of an interviewer can give rise to satisficing i.e. giving socially acceptable rather than factually correct answers to questions, particularly where these are about perceived sensitive topics. In some cases, questions from multiple surveys are put together in a single questionnaire. This is done to make a logical structure and flow for the questions, but there are still risks that there will be question order or context effects and respondents' answers to particular questions may be affected by the other questions that have already been asked in the interview, particularly the most recent ones. A difference in such effects will give rise to a discontinuity and it is not usually apparent from the survey data which answer is closest to the true one. Additional re-interview studies or matches to administrative data are required to address this.

The aim of this report is to present research relevant to the estimation of survey discontinuities both when a parallel pilot survey under the new design is available and when such a survey is not available. Emphasis is placed on presenting applications using real data from the Netherlands and the UK. Although the case studies we present are specific to survey data from these two countries, the same methodologies can be applied to data from other countries including data used for estimating SDGs and other indicators. A novel feature in this report is that in addition to methodology for estimating discontinuities at national level, we further present methodology appropriate for estimating discontinuities at sub-national (domain) levels. This is important as in many applications interest is in estimating indicators at disaggregated levels of geography. The methodology we describe in this report focusses on changes in survey designs and might be adapted for situations where surveys are replaced by big data sources. Using big data (alternative data sources) instead of survey data is a topic we cover in other deliverables of the MAKSWELL project and may also result in discontinuities. However, measuring and adjusting for these discontinuities requires careful thinking and possibly developing new methodology and hence this is not covered as part of this report.

The structure of the report is as follows. In Section 2 we present an extensive review of methods for estimating discontinuities and adjusting for discontinuities, a topic that in our view requires ad-

ditional research. In particular, a working definition of a discontinuity is provided and three cases are presented, a) methodologies when data from a parallel run (pilot) under the new survey design are available, b) time-series methodologies when data from a parallel run are not available and c) time-series methodologies when data from a parallel run are available. This section further describes design-based and model-based methodology for estimating discontinuities at domain level. In Section 3 we present results from the application of methods when data from a pilot survey (parallel run) under the new design are available. The data we use in this case come from the UK (surveys in Wales) and interest is in estimating discontinuities both at aggregate and domain levels. In Section 4 we present results from the application of structural time series methods but no data from a parallel run are available. The data we use for illustrating the methods in this case come from the UK international passenger survey. In section 5 we compare results for discontinuities in the Dutch Consumer Survey, based on a parallel run only, a structural time series model without a parallel run only, and a combination of a parallel run and a structural time series model. The last section summarises the key findings and areas for future research.

2. Methods for estimating discontinuities: A review

2.1. Introduction

The full analysis of discontinuities in the change from an original set of surveys to the new survey(s) is likely to be a long process of several stages as evidence accumulates on the new design. In order to assist users and build confidence in the new survey design, it is important to make and publish early estimates of discontinuities. Assessing survey discontinuities is a challenging problem (Van den Brakel, 2008, Bollineni-Balabay et al., 2016). The best way to obtain good estimates of the discontinuities in a survey is to run an experiment embedded within the survey and use the results to estimate the discontinuity. In this way it is possible to have most control of the design and therefore to tailor the design to the properties needed (accuracy, cost etc). Designing an experiment embedded within a survey can be challenging. More typically, survey organisations use a pilot test of the new design in parallel to the original survey(s). In some of the case studies we present as part of this report we use a pilot survey (under the new design) in conjunction with information from the original surveys to produce estimates of the difference between estimates from the new survey (approximated by the pilot) and estimates from the old survey(s). The pilot test of the new design is not an embedded experiment, but does have similar properties which allow the estimation of differences between the old and new survey implementations. In the absence of a parallel run, recent literature has proposed the use of structural time series models as an alternative approach for estimating discontinuities. A combined approach that uses data from a parallel run with a structural time series model is also possible and will be outlined in this report.

2.2. Parallel run

As described in Van den Brakel et al. (2019), the most straightforward approach to estimate the size of the discontinuity is to collect data under the old and new survey designs alongside each other for some period. This is referred to as parallel data collection or parallel run. A parallel run is preferably designed as a randomized experiment, where the sampling units from a probability sample are randomized over the current and alternative survey designs such that the subsamples can be considered as the treatment groups in an experiment. To maximize the precision of a randomized experiment embedded in a probability sample, the structure of the sample design can be used to identify potential control variables for the experimental design. Instead of directly randomizing sampling units over treatments according to a Completely Randomized Design, Randomized Block Designs (RBD) can be used. In an RBD sampling units are randomized over the treatments within homogeneous groups or blocks. This eliminates the variation between the blocks from the variance of the treatment effects, similar to the concept of stratified sampling in sampling theory. Potential block variables are obviously sampling structures like strata, primary sampling units, clusters and interviewers. For details see Fienberg and Tanur (1987, 1988, 1989), Van den Brakel and Renssen (1998, 2005), ?? (bra). Another important part of the design of an experiment is to choose the minimum required sample size of the parallel run. Therefore an advance decision is required about the minimum detectable discontinuity that should result in a rejection of the null hypothesis that a discontinuity equals zero at a pre-specified significance level (i.e. the probability that the null hypothesis is true but incorrectly rejected) and power (i.e. the probability that alternative hypothesis is true and the null hypothesis is

indeed rejected). For details of the computation of the required size of the parallel run see Van den Brakel et al. (2019). A disadvantage of a parallel run is that extra cost is required for additional data collection. Obtaining sufficiently precise estimates for the discontinuities often requires large sample sizes. Designing and conducting an experiment for parallel data collection that accurately measures the discontinuities due to the changeover also significantly increases the complexity of the fieldwork operation. In addition, a parallel run is used to estimate the discontinuity in the level of the series. Discontinuities in the seasonal pattern are also possible, but the estimation of such discontinuities would require an unrealistically long parallel run.

The standard literature for design and analysis of experiments applies model-based inference procedures for the analysis of experiments. In this case estimates for the discontinuities are obtained from the estimated treatment effects of a linear model underlying an appropriate ANOVA for the applied experimental design. A drawback of this approach is that the sample design is ignored, which might result in biased estimates for the discontinuitiesin the case of sample designs with unequal inclusion probabilities, as well as incorrect variance estimates if for example stratification or clustering is ignored. For the analysis of experiments embedded in sample surveys Van den Brakel and Renssen (1998, 2005), ?? (bra) developed a design-based inference procedure that accounts for the sample design as well as the superimposition of the applied experimental design on the sampling design. Denoting by $\hat{\theta}^*$ an estimate of a parameter of interest under the original (old) survey and by $\hat{\theta}$ an estimate using the pilot (new design) data, an estimator of the discontinuity at national (aggregate) level, $\hat{\Delta}$, can be defined as follows,

$$\hat{\Delta} = \hat{\theta} - \hat{\theta}^*. \tag{2.1}$$

For the purposes of the applications we consider in this report, we focus on the estimation of means and proportions. Hence, in the case of using a complex survey design, an estimator of the finite population average, θ, of a random variable y is given by the well-known Horvitz-Thompson (HT) or Hajek estimator of the mean,

$$\hat{\theta} = \Big(\sum_{i=1}^{n} y_i/\pi_i\Big)/\Big(\sum_{i=1}^{n} 1/\pi_i\Big), \tag{2.2}$$

where π_i is the corresponding sample inclusion probability of unit $i = 1, ..., n$. Point and variance estimates of $\hat{\Delta}$ can be derived using standard survey estimation techniques that account for the possibly complex survey design of the surveys. In finite population sampling it is customary to try and increase the accuracy of the HT estimator if suitable auxiliary information that is correlated with the outcome is available. This can be achieved by means of the generalized regression estimator (GREG), see (Särndal et al., 1992). The GREG estimator is defined as follows,

$$\hat{\theta}^{GREG} = \hat{\theta} + \hat{\beta}^T(\bar{X} - \hat{\bar{x}}) \tag{2.3}$$

where $\bar{X}$ are the populations means of the auxiliary variables and $\hat{\bar{x}}$ are the estimated (using the HT or Hajek estimators) means of the auxiliary variables.

The problem becomes somewhat more complex when interest is in measuring possible discontinuities at small domain level, a problem that in our experience survey organisations are likely to be interested in. Extending the definition above to include the estimation of discontinuities at domain level i, we derive the following estimator,

$$\hat{\Delta}_i = \hat{\theta}_i - \hat{\theta}_i^* \tag{2.4}$$

Estimating the variance of the discontinuity is important.Analytical expressions for variance estimation for the HT/Hajek and the GREG are given in Van den Brakel and Renssen (2005). However, the sample size of the pilot survey is usually smaller than that of the original survey(s). This means that the estimates, particularly for subgroups of the population (domains) are likely to have large sampling variances. Since the pilot survey has a small sample size and large variance, the estimated difference will also have a large variance, and we won't have much evidence to say whether there has been a discontinuity. In order to improve our ability to say whether the difference is important, we need a way to reduce the variance, possibly by using statistical models. The approach is to fit models to the observed estimates using other variables, which are related to them but also measured with greater accuracy. These are typically administrative data, but survey data may also be used. For the purposes of the work we present here we need a working definition of a discontinuity. In one sense all of the differences between the estimate from the pilot and the corresponding estimate from the original survey are estimates of discontinuities. However, a more realistic definition would be that there is a consistent difference. This is challenging to assess with the pilot data as it is likely to have only a limited set of pilot run observations, so it is not directly possible to say whether any observed difference will be consistent. The estimates generally have quite large variances relative to their size, and therefore a formal hypothesis test will not be informative. This leads to a subject-matter based definition for discontinuity of 5 percentage points, but we note that quality measures for the estimates of discontinuities are needed so that these can be interpreted appropriately according to context. For the applications presented in this report discontinuities are evaluated at different levels as outlined below.

1. National (aggregate) discontinuity. The first level is an evaluation of the national level discontinuities for a range of important variables. These estimates are supplemented by an approximation to the survey design for calculating the variances of the estimates. The estimators that we use are motivated under the design-based estimation framework.

2. Design-based domain-specific discontinuity. The second level is the calculation of design-based discontinuity estimates for particular domains based on the difference between the design-based estimate from the pilot survey and the design-based estimates from the original surveys. The variance is conservatively calculated as the sum of the independent variances. The estimate is tailored specifically for the domain of interest, but because of the small sample sizes, the variances

may be quite large. Nevertheless, as we will see, there are instances where the confidence interval for the estimated discontinuity does not include zero.

3. Model-based domain-specific discontinuity. In this case we use small area models in particular, the area-level Fay-Herriot model (Fay and Herriot, 1979), which are fitted to direct estimates for domains accounting for their estimated variances, and which use other data sources as predictors. The principle of model-based estimation is to combine a biased but accurate estimator derived from the model with an unbiased but variable estimate derived from the survey. So, the outcome will be an estimator with a small mean squared error (MSE). That is, the outcome will be closer to the unobserved true value on average than the original unbiased estimator. Hence, estimates calculated using this approach balance a small amount of bias resulting from the use of data from multiple groups (domains) against a reduction in variance in such a way so as to make the mean squared error of the estimates as small as possible.

The classical approach to producing estimates for small domains from sparse data is to use a small area model, incorporating information from a data source that is well correlated with the variable of interest. In this case we have the estimates from the original surveys, $\hat{\theta}^*$, which should be very strongly correlated with the measurement under the new design, $\hat{\theta}$. Our general approach for estimating domain-specific discontinuities is to use the area-level model by Fay and Herriot (1979) as a way to incorporate this information, following the approach in Van den Brakel et al. (2016). Under this approach we model the direct estimate for a domain i under the new concept for the variable of interest, $\hat{\theta}_i$, as follows,

$$\begin{aligned} \hat{\theta}_i &= \theta_i + \epsilon_i \\ \theta_i &= X_i^T \beta + u_i \end{aligned} \tag{2.5}$$

The first part in equation (2.5) defines the sampling model whereas the second part defines the linking model. In equation (2.5) X_i is a vector of auxiliary variables defined at domain level (that can include the estimates from the original survey, $\hat{\theta}^*$), β is a vector of model parameters to be estimated, u_i is a domain random effect and ϵ_i is the sampling variance of the direct estimates obtained with the data from the parallel run. The two errors are assumed to be independent, $u_i \sim N(0, \sigma_u^2)$, $\epsilon_i \sim N(0, \psi_i)$ and ψ_i is assumed to be known. The empirical best predictor (EBLUP) of θ under the Fay-Herriot model is defined as follows,

$$\hat{\theta}_i^{EBLUP} = \hat{\gamma}_i \hat{\theta}_i + (1 - \hat{\gamma}_i) X_i^T \hat{\beta}, \tag{2.6}$$

where $\gamma_i = \frac{\sigma_u^2}{\sigma_u^2 + \psi_i}$ is the shrinkage factor. The mean squared error (MSE) of $\hat{\theta}_i^{EBLUP}$ can be estimated by using either the analytic estimator proposed by Prasad and Rao (1990) or computer intensive methods for example, the parametric bootstrap under the FH model. In its simplest form the parametric bootstrap consists of generating bootstrap samples (b) using the assumptions of the FH model and the estimated, with the original sample, parameters $\hat{\phi} = (\hat{\beta}, \hat{\sigma}_u^2, \hat{\psi}_i)$, computing new true area parameters

$\theta_i(\hat{\phi})$ and computing new EBLUP estimates using the generated bootstrap samples, $\hat{\theta}_i^{EBLUP,(b)}(\hat{\phi}^{(b)})$. Using a total of B bootstrap samples, the estimated MSE of $\hat{\theta}_i^{EBLUP}$ is then computed using

$$\widehat{MSE}(\hat{\theta}_i^{EBLUP}) = \frac{1}{B}\sum_{n=1}^{B}(\hat{\theta}_i^{EBLUP,(b)}(\hat{\phi}^{(b)}) - \theta_i(\hat{\phi}))^2$$

An estimator of the discontinuity is then defined as follows,

$$\hat{\Delta}_i^M = \hat{\theta}_i^{EBLUP} - \hat{\theta}_i^*, \tag{2.7}$$

where the superscript M is used to denote that a model is used in this case. Estimating the MSE of $\hat{\Delta}_i^M$ requires careful consideration. In the simplest case, we can assume that $\hat{\theta}_i^{EBLUP}$ and $\hat{\theta}_i^*$ are independent and therefore the MSE of $\hat{\Delta}_i^M$ can be estimated by the sum of the estimated MSE of $\hat{\theta}_i^{EBLUP}$ and the estimated design variance of $\hat{\theta}_i^*$. This is not an unreasonable assumption given that the pilot survey and the original survey are independent and assume that the probability of having overlapping units between the two surveys is small. Nevertheless, there is a situation where an additional covariance term needs to be accounted for when estimating the MSE of $\hat{\Delta}_i^M$. This is when the design-based estimates under the old design $\hat{\theta}_i^*$ are used as auxiliary variables in the FH model. This is a reasonable strategy for selecting model covariates since it is reasonable to expect a high correlation between the direct estimates from the old and the new surveys. A solution to this problem is proposed by Van den Brakel et al. (2016).

Let us now provide some additional comment on the FH model. The need to assume that ψ_i is known is because in the case of area-level models we work with area/domain-level data with one data point per domain available. Hence, given the available data, it is not possible to estimate ψ_i. In the simplest case ψ_i is estimated by using survey micro-data or is supplied by the data provider (if no survey micro-data are available) and then treated as fixed in the model. This is the approach we follow in the applications we present in this report. Alternative methods that account for the uncertainty in estimating ψ_i have been proposed in the literature using for example a hierarchical Bayes framework (You and Chapman, 2016) and can be implemented for example by using R and OpenBUGS. It is important to produce good estimates of ψ_i from the regular survey accounting for the complex design features, as these affect the weight given to the direct and indirect estimates under the FH model equation (2.5). The strength of the area-level model is that it has a very simple formulation, and can be fitted using aggregate (area-level) data. Access to aggregate data is easier than access to unit-level and this is an added advantage of the area-level model. The model relies on there being enough small areas to estimate the heterogeneity via random effects. In the applications we present in this report this should be fine for areas such as local authority districts, however, for other domains such as local health boards the number of groups is rather small. We will therefore consider whether including the random effects in the model is of added value, or the use of a direct estimator is sufficient. The data requirements for estimating these models are modest - only the area-level estimates from the old and parallel (new) surveys plus their sampling errors taking into account the complex survey designs are needed. Additional variables at the same geographical level (e.g. population estimates, possibly

within age-sex subgroups) may be useful as additional explanatory variables in the model. In order to assess whether this is an important difference, we would like to be able to combine the model and sampling variances, noting that they are not independent. An estimator for the variance of $\hat{\Delta}_i^M$ is given by Van den Brakel et al. (2016), and we propose to follow their approach in the applications. The modelling of the data is undertaken with the R statistical software package.

2.2.1. Model variants and extensions

Alternative approaches to modelling are also possible. One suggestion is model the discontinuity i.e. the difference between the two estimates directly. There is a debate over whether this is practical, on one side holding that it is more logical and methodologically simpler to estimate directly what you are interested in, and on the other considering that there may be no good predictors for directly modelling the discontinuity. In contrast, the estimates from the old survey would be good predictors for modelling the estimates under the new design. A further approach is to use a multivariate area-level (Fay-Herriot) model to jointly model the direct estimates under the new and old designs. An alternative approach would be to use a fully Bayesian approach to model directly the discontinuity and also to account for the uncertainty in the estimation of the sampling variance. The advantage in this case is that we can propagate the uncertainty from various sources, hence making the estimation of the mean squared error of the discontinuity estimates easier to quantify. Another aspect of the models we use in the applications presented in this report is that they do not account for the measurement error in the auxiliary variables coming from the old survey(s). Clearly, this is a strong assumption. However, approaches to accounting for the presence of measurement error in the case of the Fay-Herriot model having been considered in the literature and can be implemented with the data available (see Ybarra and Lohr (2008), Bell et al. (2019)). Assessing the uncertainty of the estimated discontinuities is also important if we were to attempt to make adjustments to the estimates. One pragmatic approach is to use the 5 percentage points rule as the threshold for defining significant discontinuities. However, using hard boundaries is always problematic, and we are suggesting that it is more important to assess the discontinuity estimates with respect to their variances than to have a hard boundary rule. In addition, important differences are likely to be of different sizes in different variables. It is nevertheless useful to have an initial position against which to assess estimates of discontinuities, so one can use the 5 percentage points rule as a guide.

From the topics listed, we will focus on the extension of the Fay-Herriot model to account for covariate measurement error. This is important as in the case of our application we did not have access to population (Census or register) data. Hence, the only source of model covariates were survey data, which are subject to sampling error. Following Ybarra and Lohr (2008) we assume that the X_i are fixed unknown quantities hence we consider a functional measurement error area-level model. In this case, the EBLUP is defined as follows,

$$\hat{\theta}_i^{EBLUP-ME} = \hat{\gamma}_i^{ME}\hat{\theta}_i + (1 - \hat{\gamma}_i^{ME})X_i^T\hat{\beta}, \tag{2.8}$$

where the superscript ME indicates that we are working under the Fay-Herriot model that accounts for covariate measurement error, $\gamma_i^{ME} = \frac{\sigma_u^2+\beta^T V_i\beta}{\sigma_u^2+\beta^T V_i\beta+\psi_i}$ and V_i is the matrix of estimated sampling

variances of the survey-based covariates X_i.

An estimator of the discontinuity in this case is defined by,

$$\hat{\Delta}_i^{M-ME} = \hat{\theta}_i^{EBLUP-ME} - \hat{\theta}_i^*. \tag{2.9}$$

2.3. Structural time series model

When there is no parallel run, a time series model can be used to estimate the discontinuity. This is called the intervention approach. The intervention approach with state-space models is originally proposed by Harvey and Durbin (1986) to estimate the effect of seat belt legislation on British road casualties. **?** and Van den Brakel and Roels (2010) applied this approach to estimate discontinuities induced by a redesign of a sample survey process. See Durbin and Koopman (2012) for a general introduction to structural time series models. In the simplest case, a univariate model is sufficient. We start with the description of the structural time series model. In the structural time series model, a time series $y_t, \quad t = 1, \ldots, T$ is modelled as the sum of different components. We consider the following components:

1. Trend (L_t)
2. Seasonal (S_t)
3. Regression component ($\boldsymbol{\beta}^{'}\mathbf{x}_t$), here used to model the discontinuity
4. White noise (I_t)

In a more general model, cycles and AR- or MA-components can be added, and a regression component can be used to take an auxiliary series into account. This is not considered here. So we write:

$$y_t = L_t + S_t + \boldsymbol{\beta}^{'}\mathbf{x}_t + I_t, \quad t = 1, \ldots, T. \tag{2.10}$$

In the literature, different models for the trend are described. Here we use the smooth trend model, which is defined by

$$\begin{aligned} L_t &= L_{t-1} + R_{t-1} \\ R_t &= R_{t-1} + \eta_t \end{aligned}$$

$$E(\eta_t) = 0$$

$$cov(\eta_t, \eta_{t'}) = \begin{cases} \sigma_\eta^2 & \text{if } t = t' \\ 0 & \text{if } t \neq t' \end{cases}$$

where L_t is called the level and R_t can be interpreted as the slope. This model often results in quite

stable trend patterns.

For the seasonal, the trigonometric model is used in this paper. The seasonal pattern is described with a set of $\frac{J}{2}$ harmonics, with J the number of periods in a year ($J = 12$ for monthly figures).

$$S_t = \sum_{j=1}^{J/2} \gamma_{j,t}$$

$$\gamma_{j,t} = \gamma_{j,t-1} \cos\left(\frac{\pi\ j}{J/2}\right) + \gamma^*_{j,t-1} \sin\left(\frac{\pi\ j}{J/2}\right) + \omega_{j,t}$$

$$\gamma^*_{j,t} = \gamma^*_{j,t-1} \cos\left(\frac{\pi\ j}{J/2}\right) - \gamma_{j,t-1} \sin\left(\frac{\pi\ j}{J/2}\right) + \omega^*_{j,t}$$

$$j = 1, \ldots, J/2$$

$$E(\omega_{j,t}) = E(\omega^*_{j,t}) = 0$$

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

The last harmonic is reduced to $\gamma_{6,t} = -\gamma_{6,t-1}$ and $\gamma^*_{6,t}$ is not required (for monthly figures). In a more general model, it can be assumed that each harmonic has its own variance $\sigma^2_{\omega,j}$.

Then the regression component is used to model the discontinuity, i.e. $x_t = 0$ for all periods t before the discontinuity, and $x_t = 1$ from the moment the discontinuity occurs. In the general case, more than one discontinuity may occur at different points in time. Then, more than one of such components can be added. Under the assumption that the sum of the trend L_t and seasonal S_t correctly models the evolution of the population variable, the regression coefficient β can be interpreted as the systematic effect of the redesign on the level of the series.

In a cross-sectional survey, the white noise parameter I_t is the sum of unexplained noise in the population parameter and the sample error, as these two noise variables cannot be distinguished. I_t is modelled as

$$E(I_t) = 0$$

$$cov(I_t, I_{t'}) = \begin{cases} \sigma^2_I & \text{if } t = t' \\ 0 & \text{if } t \neq t' \end{cases}$$

With this formula, it is assumed that the variance of the sample error is constant over time, and especially, that is does not change due to the redesign. When this assumption is not appropriate, changes in this variance can be modelled as well. This will be discussed in the next section.

This model can be applied to estimate discontinuities in series which are caused by redesigns. Sometimes, it is necessary to apply a tailor-made model to take specific properties of the series and the redesign into account. In the case of the Consumer Confidence Survey (CS) a small change of this

model is necessary, as will be discussed in the next section.

As pointed out in the literature (for example Van den Brakel and Krieg, 2015) it is assumed with step intervention βx_t that the redesign only has a systematic effect on the level of the series. Alternative interventions, e.g. for the slope or the seasonal components are also possible, see Durbin and Koopman (2012) and Van den Brakel and Roels (2010). Note that a discontinuity in the seasonal pattern can only be estimated after some years of data collection under the new design.

The estimate of the discontinuity can be improved if auxiliary information is available. This auxiliary information should be a related series which is not affected by the discontinuity. Then, a multivariate time series model can be used, where both the target series and the auxiliary series are the input. The auxiliary series is modelled as a sum of trend, seasonal and white noise. By modelling a correlation between the slope disturbances of both series, the auxiliary series is used to improve the estimates of the trend of the target series and therefore also of the discontinuity. It is also possible to take the auxiliary series into account by a regression component.

The general way to fit a structural time series model is to express the model in the so-called state-space representation and apply the Kalman filter to obtain optimal estimates for the state variables, see e.g. Durbin and Koopman (2012). Estimates for state variables for period t based on the information available up to and including period t are referred to as the filtered estimates. The filtered estimates of past state vectors can be updated, if new data become available. This procedure is referred to as smoothing and results in smoothed estimates that are based on the completely observed time series. The analysis is conducted with software developed in OxMetrics in combination with the subroutines of SsfPack 3.0, see Doornik (2009) and Koopman et al. (2008). The non-stationary variables are initialised with a diffuse prior, i.e. the expectation of the initial states are equal to zero and the initial covariance matrix of the states is diagonal with large diagonal elements. The white noise is stationary and therefore initialised with a proper prior. The initial value is equal to zero and the variance is equal to the hyperparameter σ_I^2.

In Ssfpack 3.0 an exact diffuse log-likelihood function is obtained with the procedure proposed by Koopman (1997). Maximum likelihood estimates for the hyperparameters, i.e. the variance components of the stochastic processes for the state variables are obtained using a numerical optimization procedure (BFGS algorithm, Doornik, 2009). To avoid negative variance estimates, the log-transformed variances are estimated.

2.4. Combining a parallel run with a time series

The two methods described above, the parallel run and the structural time series approach, can be combined. This can be done with exactly the same structural time series model as before, but now with the estimate of the discontinuity included through an exact initialization of the Kalman filter. When no parallel run is conducted, the variable β is initialized with a diffuse prior. This means that no information about the size of the variable is available. With a parallel run, there is a point estimate from the parallel run, together with the variance of this estimate. This information is used to initialize the discontinuity with an exact prior. When the model is estimated, this prior information is improved

using the information of the development of the series before and after the parallel run. This approach is interesting when a short parallel run is carried out and the estimates based on this parallel run are not sufficiently accurate.

2.5. Adjustment methods

After estimating the discontinuities, it has to be decided whether and how the series are corrected, in order to avoid that real developments are confounded with the discontinuities. One possibility is to publish the uncorrected series together with the estimates of the discontinuity. For some users, however, consistent series are necessary, and with this choice the correction is left to them. It is also possible to publish corrected series. Then the series from the past can be adjusted to the level under the new design, which is called backcasting. It is also possible to adjust the observations under the new design to the level of the series before the changeover. This works similarly as backcasting and is not discussed here. Backcasting methods are often based on synthetic approaches that use naive numerical adjustments that rely on the strong assumption that the observed discontinuities are time-invariant. Additive adjustments simply subtract the difference between the level under the old and the new design from the series observed before the changeover to make it comparable with the observations under the new design. This assumes that the adjustment is independent of the value of the series to be adjusted. Ratio adjustments multiply the series observed before the changeover with the ratio of the levels under the new and the old designs and assume that the adjustment is proportional to the level of the observed series. This approach can be useful to make appropriate adjustments for variables that cannot take negative values. Both backcasting methods depend on strong assumptions, and the choice between them depend on which assumptions seems more plausible, and on the necessity to avoid impossible values in the adjusted series. When both additive and ratio adjustment are not plausible, other backcasting methods can be developed according to the assumptions about the development of the discontinuity in the past.

3. Data sources and applications in the presence of a parallel run: The case of surveys in Wales

3.1. Survey data from Wales

We illustrate the methodology for estimating survey discontinuities when data from a parallel run are available. In particular, the Welsh Government (WG) has reviewed the way in which social surveys are conducted in Wales, and for a range of reasons, including value for money grounds (for more details of the options and reasons for choosing among them see Welsh Government, 2014), instituted a new National Survey (NSn) from 2016. The NSn collects information previously collected in five other surveys, the [old] National Survey (NSo), the Welsh Health Survey (WHS), the Active Adults Survey (AAS), the Arts in Wales Survey (AWS) and the Welsh Outdoor Recreation Survey (WORS). The NSn has a longer questionnaire which allows many (but not all) of the questions from the original surveys to be included, though the exact pattern of timing for the inclusion of particular questions has not yet been worked out in all instances.

The process of agreeing the NSn involved consultations with the customers of the existing surveys and negotiations to ensure that the new structure is best able to meet their needs and that they are happy to support it. It is important, however, to demonstrate to the users that the methodology for the new survey is appropriate and to produce estimates of the change in the series (discontinuity) caused by the change from one design to another. The NSn is similar in concept to the NSo, but has some differences in the design intended to make it as statistically efficient as possible. It follows the NSo in using a rotating design which covers all regions in Wales over a year in a regular pattern, such that any aggregation of a contiguous year of sample cases forms an unclustered sample, with associated benefits of lower sampling variance relative to clustered designs. Shorter periods remain clustered. Some differential sampling by Local Authorities (LAs) is included to ensure that there are sufficient cases to form the basis of estimates for LAs in Wales (22 in total) and Welsh Health Boards (7 in total). The work we present in this application is important because the Wellbeing of Future Generations Act 2015 (Welsh Government 2015), which applies to all public bodies in Wales, requires decisions to be made with regard to a wellbeing model, using evidence from national indicators and between a quarter and a third of these indicators come from the NSn, so the accuracy and credibility of the survey are particularly important to support this policy. The characteristics of the five original surveys are summarized in Table 3.1. They cover a wide range of topics and have a combined annual sample size of 40,900; the sample size for the NSn is 12,000.

Table 3.1: Design features of the original surveys

Survey	Frequency and last instance	Average interview length (mins)	Sample Size	Design
National Survey	Annual, continuous 2014-15	25	14.500	Face-to-face
Welsh Health Survey	Annual, continuous 2015	25	15.000	Primarily self-completion
Active Adults Survey	Every two years 2014	20	8.000	Face-to-face
Arts in Wales Survey	Every five years 2015	15	7.000	Face-to-face
Welsh Outdoor Recreation Survey	Every three years 2014	15	6400	Telephone interview

The data from the last instances of these component surveys were made available in a secure setting at the University of Southampton. The National Survey (NSo) was taken from the version deposited at the ESRC Data Archive, the Welsh Outdoor Recreation Survey microdata are freely available in anonymised format and the other three surveys were provided directly under specific agreements for processing in a secure facility. All the datasets are provided with weights, which compensate for the sampling (using sampling weights) and non-response (using either or both of a non-response propensity model adjustment and a calibration adjustment for known population totals by age-sex and local authority). These weights were used in the modelling and variance calculation procedures.

For the purposes of estimating discontinuities the new survey is represented in this report by the National Survey pilot (parallel run), and its design was as similar as possible to the new survey design. The main differences are that it was clustered for fieldwork efficiency, though still designed to provide reasonably precise estimates at national level, and that sampling was more differential by LA so that approximately equal interview numbers were achieved in each local authority. The sample was drawn from the postcode address file and was stratified by LA. Fieldwork took place closely after the end of fieldwork for the final waves of the 2012-15 National Survey, the Active Adults Survey, and the Welsh Outdoor Recreation Survey and in parallel with the final waves of the Welsh Health Survey and the Arts in Wales Survey. The median survey length in the parallel run was approximately 45 minutes, with substantial subsampling and a computer assisted self interviewing (CASI) module covering potentially sensitive topics (mainly ones taken from the current Welsh Health Survey). The data from the parallel run were made available under specific agreements for processing in a secure facility at the University of Southampton.

Interest in our applications was in estimating discontinuities both at national and domain levels with domains defined by local authorities and health board districts also cross-classified by demographic groups for example, age by gender groups. For estimating discontinuities we are making the following assumptions:

1. There is no material difference between the properties of the estimates from the pilot (parallel) run and the properties of the estimates from the NSn, except for those that are due to sampling variation. That is, that the pilot survey is the same as the new National Survey, just with a reduced sample size. This is a reasonable assumption since the structure of the survey is the same and since the one real difference is accounted for approximately in the calculation of the standard errors.

2. The impact of the timing differences between the last instances of the five original surveys and the pilot on estimation are negligible. In some cases, the surveys overlapped with the pilot survey. The Active Adults Survey and Welsh Outdoor Recreation Survey were one year in advance of the parallel run.

3. There are no differences in the responses due to seasonality caused by the limited time period of the pilot (a six-month survey period) and the whole-year versions of the original surveys.

In Section 3.2 survey discontinuities are estimated for a range of variables and domains for four of the original surveys. However, due to the sensitive nature of the variables, the results for this application are presented in an anonymised form, i.e. we do not disclose the names of the variables, the surveys and the domains. Instead, we report the levels of discontinuities and corresponding confidence intervals.

3.2. Application to the Welsh Surveys

Table 3.2 presents ranges of national (aggregate) discontinuities for different variables in four of the original Welsh surveys estimated by using the pilot data under the new design. These are ranges of point estimates and not confidence intervals. Our analyses aims at estimating discontinuities both at national and domain levels. Estimates at national (aggregate) level are obtained using equation (2.1). We conclude that for a range of surveys and variables we consistently estimate negative discontinuities (in particular for survey 1). A number of variables also show discontinuities that are larger than the nominal five percentage points threshold. These results indicate that an adjustment for continuity may be needed.

A more detailed picture is provided by estimating discontinuities at domain level (different definitions of domains are used for creating the various plots). Figures 3.1 and 3.2 serve a twofold purpose. The left-hand side figures present estimated discontinuities for different domains using three estimation approaches namely, a direct domain estimator (in this case we use the Horvitz-Thompson (HT) estimator, denoted by DiscHT) and two model-based estimators under the Fay-Herriot model (DiscFH) and the Fay-Herriot model that accounts for covariate measurement error (DiscME). We note that the model-based estimates of discontinuities are more stable than direct estimates. This may be expected since direct estimation may rely on small domain-specific sample sizes. Secondly, we observe that in most cases the model-based estimates that account for covariate measurement error are closer to the direct estimates than the model-based that do not account for covariate measurement error. This is also something we may expect since accounting for the uncertainty in the covariates may increase the weight we give to the direct estimator. The right-hand side figures show 95% confidence intervals constructed by using the estimated variance of the HT estimator and the Prasad and Rao (1990) analytic estimator in the case of the estimates produced under the Fay-Herriot model. In this case we assume independence and therefore ignore the covariance term proposed by Van den Brakel et al. (2016). Confidence intervals for the domain estimates under the FH measurement error model are not presented as computation in this case requires the use of computer intensive methods e.g Jackknife (see Ybarra and Lohr (2008)). As expected, we notice that model-based estimates have narrower confidence intervals compared to direct estimates that are computed using only domain specific data. These plots also show the presence of what we would classify as significant discontinuities for specific domains (see for example the plot for variable 2).

Table 3.2: Ranges of estimated national discontinuities in 4 of the original surveys in Wales

Surveys	Ranges of discontinuities
Survey 1	-0.108, -0.058
Survey 2	-0.111, 0.166
Survey 3	-0.082, 0.110
Survey 4	-0.152, 0.184

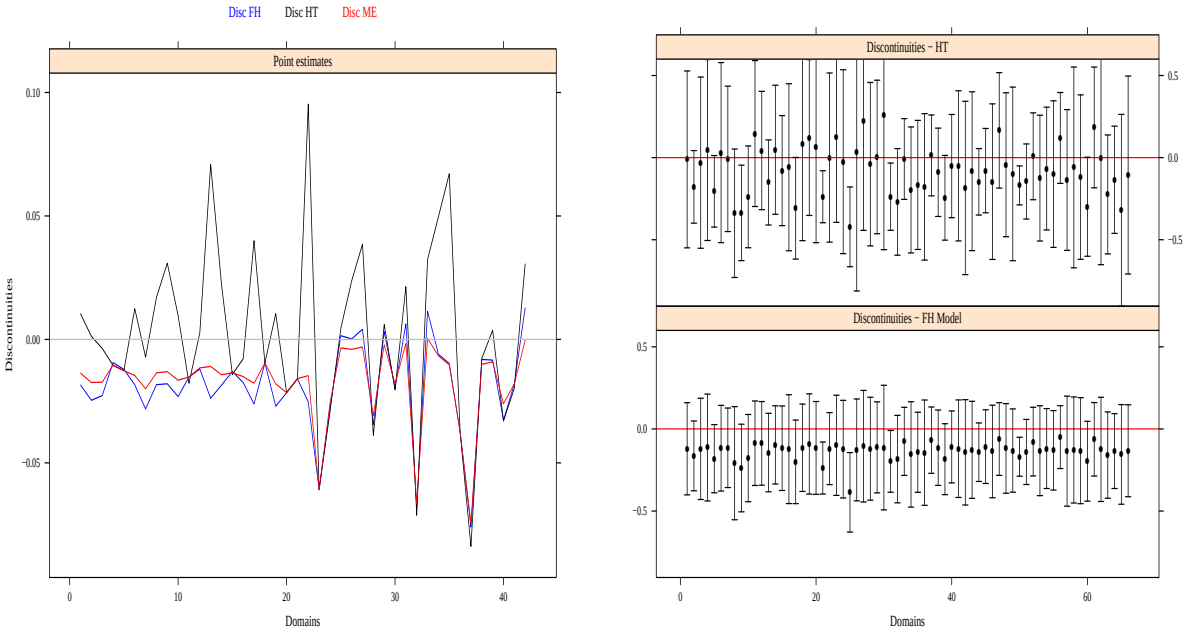

Figure 3.1: Left figure: Estimated discontinuities (using different estimators) for variable 1 in different domains. Right figure: 95% confidence intervals of direct and model-based (FH) estimates. The two plots present results for different definitions of domains.

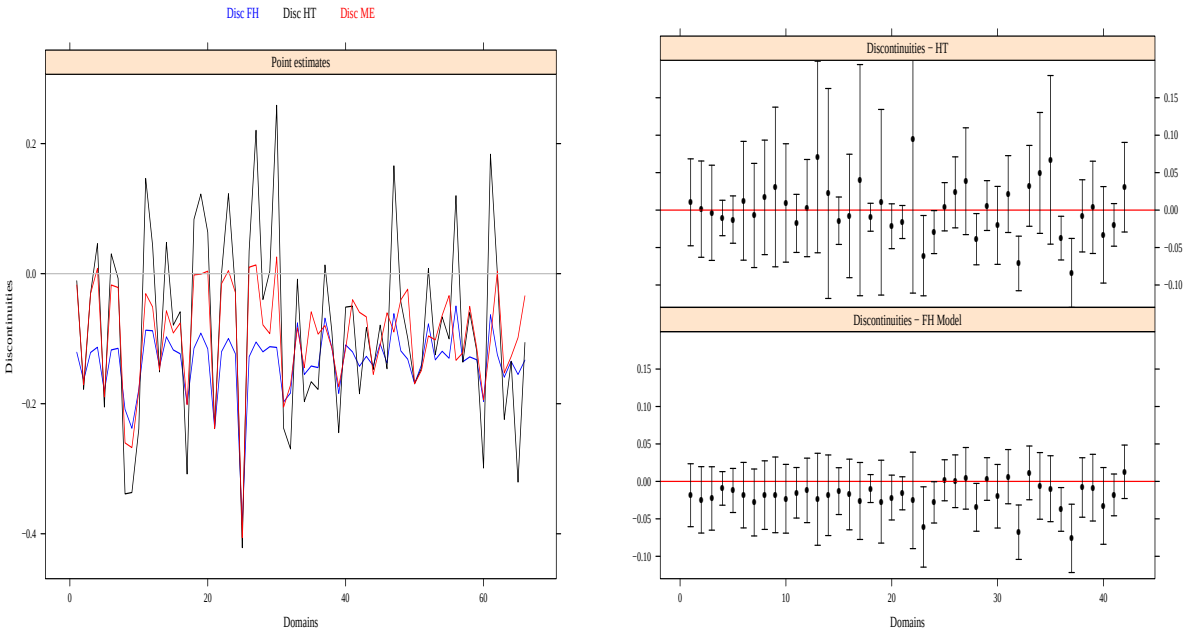

Figure 3.2: Left figure: Estimated discontinuities (using different estimators) for variable 2 in different domains. Right figure: 95% confidence intervals of direct and model-based estimates. The two plots present results for different definitions of domains.

4. Data sources and applications using structural time series with no parallel run: The UK International Passenger Survey

4.1. UK data from the International Passenger Survey

An additional source of data we will use in one of our applications is the UK International Passenger Survey (IPS). The IPS interviews people as they enter or leave the United Kingdom through ports, airports and the channel tunnel (the land border between Northern Ireland and Ireland is not covered). Interviewers attempt an interview with every k^{th} traveller, in some cases screening them to find out if they are a migrant (in which case they get a series of questions appropriate to migrants), and in a subset of cases going on to ask questions on expenditure. The interview has been designed to be short and to be flexibly administered by interviewers to maximise response. Data were collected on paper questionnaires and then keyed (mostly) on site via a laptop into a bespoke software tool (Blaise), then transmitted to the UK Office for National Statistics (ONS) via a secure connection, until 2017-18. ONS changed the data collection mode to use tablet-administered questionnaires, which have benefits from validation, coding etc at the point of data capture. Changing the mode of collection and the associated changes in the questionnaire and editing procedures may result in a discontinuity in the IPS estimates and a potential step change in the key IPS time series of travel, tourism and migration. As we have already discussed in this report, a discontinuity is defined as a change in an estimate that results from a change in the collection approach and is not a change due to sampling error or a real change due to a change in the environment. Such a discontinuity needs to be measured, controlled and understood in order that the IPS time series before and after the change can be compared accurately. The ONS decided to make the change by a gradual roll-out to the various interviewing locations, which made the transition operationally feasible, because different interviewer teams operate in each location. This presents less information than a situation with an embedded experiment, where the treatments (different modes) can be randomised at some level, or a parallel run, but still provides a way to estimate the parameters of the transition with a state space model. The standard approach to dealing with possible discontinuities in time series resulting from changes to field procedures involves an embedded experiment in the survey, starting with a small experiment run alongside the standard survey procedure. If there is no evidence of a major change from this, then it is extended, and finally the new method is rolled out with a small part retaining the original method. ONS's assessment of the operational considerations in introducing the change to the IPS was that the randomisation of cases, interviewers or shifts would introduce too much disruption and therefore risk to the quality of the outputs. There were also requirements for the roll-out of training for interviewers that made a staged transition interviewer team by team (where a team may cover a single site or a group of sites) the most tractable implementation approach. This makes it more challenging to estimate the specific effect of any discontinuity.

4.2. Application to the UK International Passenger Survey

In this section we analyse the data we introduced in Section 4.1. For this application discontinuities are estimated using the Structural Time Series approach that we presented earlier in this report in Section 2.3. One modification, however, is required. Covariate x_t in equation (2.10) now represents

Table 4.1: Definitions and abbreviations of key IPS variables

Inflow variable	Definition	Outflow variable	Definition
svisukres	Number of overseas visits by UK residents	svisosres	Number of visits to the UK by overseas residents
sexpukres	Expenditure abroad by UK residents in pounds	sexposres	Expenditure in the UK by overseas residents in pounds
smigosar	Overseas residents migrating to the UK	smigukdep	UK residents migrating abroad
sflowarr	Total arrival passenger flow	sflowdep	Total departure passenger flow
sflowarrn	Arrival passenger flow excluding flow from Channel Islands and Isle of Man	sflowdepn	Departure passenger flow excluding flow from Channel Islands and Isle of Man

an indicator explanatory variable which takes the value 0 before the discontinuity is introduced, then an increasing positive value, which is the fraction of sampling units observed under the new design, as the rollout progresses, and then the value 1 once the rollout is completed and for all subsequent periods.

The models are expressed in state-space form and the Kalman filter is used to estimate the state variables (Durbin and Koopman, 2012). The models have been implemented in OxMetrics (Doornik, 2009) in combination with SsfPack (Koopman et al., 2008). The Kalman filter works by producing initial estimates based on the first data points, and then updating these as more information accumulates. In practice this often means that early estimates are very large (or small) with very large variances, but that as data accumulate they settle to a more stable level and variance. As a discontinuity is introduced, we would therefore expect early estimates to be far from the truth (particularly as early in rollout there is little information on which to base an estimate since few ports will be using tablets), but to converge to a more stable estimate as further data accumulate.

Table 4.1 presents the definitions and abbreviations of key IPS variables. Figure 4.1 shows the estimated discontinuity for svisosres. In this case the estimate seems to be close to stabilising, although it is hard to say what will happen when additional data points are added. The estimated discontinuity (in millions of people) in Jun 2018 is -244 ± 150, which is significantly different from zero, but not very accurately estimated. By contrast, Figure 4.2 shows the estimated discontinuity for sexpukres, and here there is no sign that the estimate has stabilised yet. The latest month's estimate is still quite different from the previous month, and the estimated confidence intervals are wide. In both of these examples, the behaviour of the trend components of the models have not changed as a result of the addition of the latest data. This suggests that there has been no detectable effect of Brexit, or possibly that some of the Brexit effect has been picked up in the estimate of the discontinuity.

Table 4.2 summarises the estimates of discontinuities for inflow variables at June 2018. Since these are filtered estimates, the latest estimate is the best as it uses all the information in the series. However, it may be that all the accumulated information is not yet sufficient to stabilise the estimate, and we have made a subjective judgement based on the evolution of the series about whether they have converged; only further data will demonstrate whether this is correct, so this may not be the most helpful information. However, it seems clear that making adjustments from the current values will be risky in the current situation, where none of the inflow variables seems to have converged.

Table 4.3 provides similar summary information on the outflow variables. Among these variables, there are some instances where there appears to be some move towards convergence in our subjective judgement. It is also interesting that one of the variables shows a discontinuity significantly different

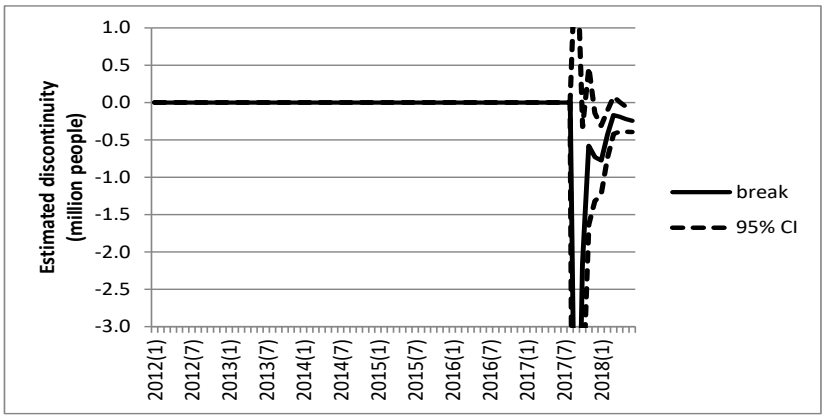

Figure 4.1: Estimated discontinuity and its 95% confidence interval for svisosres

Table 4.2: Summary statistics on estimates of discontinuities at June 2018 for inflow variables

Variable	Units	Estimated discontinuity - June 2018	Estimated se	Sig different from zero? ($\alpha = 0.05$)	Approx. converged? (subjective judgement)
svisukres	th	45	234	N	N
sexpukres	£M	-109	189	N	N
smigosar	no.	-8804	4793	N	N
sflowarr	th	-217	195	N	N
sflowarrn	th	-202	197	N	N

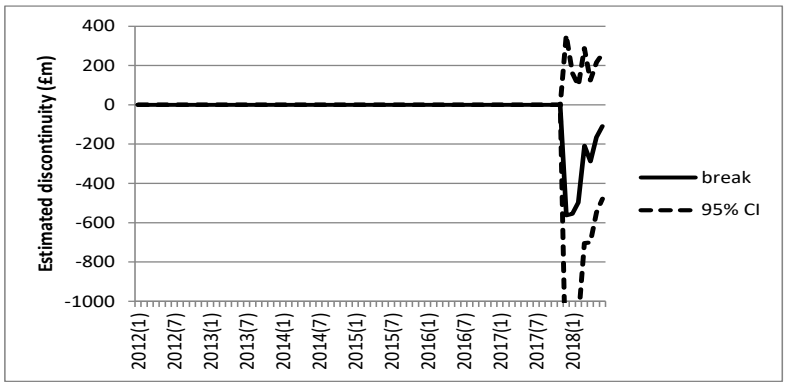

Figure 4.2: Estimated discontinuity and its 95% confidence interval for sexpukres

Table 4.3: Summary statistics on estimates of discontinuities at the latest data point (June 2018) for outflow variables

Variable	Units	Estimated discontinuity - June 2018	Estimated se	Sig different from zero? ($\alpha = 0.05$)	Approx. converged? (subjective judgement)
svisosres	th	-243	77	Y	Y
sexposres	£M	-169	121	N	Y
smigukdep	no.	-5733	3070	N	Y
sflowdep	th	44	255	N	N
sflowdepn	th	40	257	N	N

from zero. This discontinuity may therefore be real. However, we have made several comparison tests simultaneously, so finding one significant result by chance is less surprising, and we could interpret this as providing only mild evidence for a real effect.

Most of the estimates of discontinuities are not significantly different from zero. Nevertheless, some of the discontinuities are large, and the effects on estimated numbers of migrants are relevant to users. The discontinuity in expenditure by overseas residents is also large enough to affect the interpretation of the series. Almost all of the discontinuity estimates are negative, which means that the measurement made with tablets is lower than the previous paper-based measurement. This seems to contradict the initial indications from the pilot study, which were that the tablets were better at capturing expenditure, which was therefore higher in the new mode. The pilot used a small sample, however, and may not be a strong indicator of direction. If the indications of direction of the discontinuity from the pilot were correct, it is possible that the size of the discontinuity is at least in part confounded with changes in migration and expenditure patterns influenced by changing exchange rates and uncertainty over Brexit. A final adjustment still requires further time periods to ensure that the estimated discontinuity has indeed stabilised, and therefore revised series based on these revised adjustments should be expected in the future.

5. Data sources and applications using a structural time series and a parallel run: The Dutch Consumer Survey

5.1. The Dutch Consumer Survey

5.1.1. Parallel run

In the first three months of 2017, a parallel run took place for the Dutch CS. Detailed results of this parallel run are shown in Table 5.1 for one variable (economic situation of the last 12 months). The results are quite similar over the three months. In that period, most of the respondents were positive about the economic situation. We see that the percentage for "a little better" increases substantially under the new design. The percentage for "a little worse" increases too, but much less. On the other hand, the percentages for the neutral answers ("same" and "don't know") and for "a lot better" and "a lot worse" decrease.

We assume that part of these changes can be explained by the changes in the questionnaire. Under the old design, the respondent had to choose between better, same or worse, and when the situation was changed only a little, the answer better or worse probably did not feel appropriate, not knowing that "a little better/worse" are possible answers. Therefore, the respondents chose one of the neutral options. On the other hand, under the new design the respondent chooses one of the the options "a little", and in the period of the parallel run, mostly "a little better". We do not have an explanation for the fact that the options "a lot better" and "a lot worse" are chosen less often under the new design.

Table 5.2 shows the mean differences for the percentages for three other variables. The results are similar for the economic situation in the last 12 months (Table 5.1) with an increase in the percentages of "a little better" and "a little worse" and a decrease in the neutral options. The decrease in the percentages for "a lot better" and "a lot worse" is smaller for these three variables.

Table 5.3 summarizes the estimates for the discontinuities in $p_{i,+}$, $p_{i,-}$ and y_i for all 8 indicators. We see that the question about major purchases, the only question where the questionnaire has not been changed, is the only question where the discontinuity for the positive answers is negative, and smaller

Table 5.1: Results of the parallel run, economic situation in the last 12 months

	January			February			March			
	old	new	diff.	old	new	diff.	old	new	diff.	mean diff.
a lot better	14.1	7.0	-7.1	14.9	7.8	-7.1	16.4	9.9	-6.5	-6.9
a little better	31.3	51.1	19.8	31.1	50.0	18.9	32.4	48.6	16.2	18.3
same	35.7	25.6	-10.1	35.0	25.8	-9.2	34.0	25.7	-8.3	-9.2
a little worse	7.1	10.0	2.9	6.0	10.4	4.4	6.6	8.4	1.8	3.0
a lot worse	6.1	3.4	-2.7	6.9	3.1	-3.8	5.8	3.2	-2.6	-3.1
don't know	5.8	2.9	-2.9	6.1	2.9	-3.2	4.8	4.2	-0.6	-2.2

Table 5.2: Results of the parallel run, mean differences in the percentages over three months, for three variables

	Econ. next 12 months	Fin. last 12 months	Fin. next 12 months
a lot better	-0.8	-2.2	0
a little better	17.6	10.9	11.1
same	-7.7	-11.6	-11.3
a little worse	1.8	7.6	4.8
a lot worse	-1.7	-4.4	-1.7
don't know	-9.2	-0.3	-2.8

Table 5.3: Results of the parallel run, estimates of the discontinuities

	positive answers	SE	negative answers	SE	difference
Econ. last 12 months	11.4	1.3	-0.1	0.9	11.5
Econ. next 12 months	16.8	1.2	0.1	0.8	16.7
Fin. last 12 months	8.7	1.0	3.2	1.1	5.5
Fin. next 12 months	11.1	1.0	3.1	0.9	8.0
Major purchases	-4.7	1.2	0.1	0.8	-4.8
Economic climate	14.1	-	0.0	-	14.1
Willingness to buy	5.0	-	2.1	-	2.9
Consumer confidence	8.7	-	1.3	-	7.4

than for the other questions. As the focus of this paper is on the first 5 indicators, the standard errors of the last three indicators are not shown. In the continuation of this section, other estimates for these discontinuities will be computed, and we will compare both the point estimates and the standard errors.

5.1.2. Backcast method

The results of the parallel run suggest that the respondents tend to choose the "a little" answer options more often and the neutral answer options less often under the new design. During the parallel run, which was a period of positive consumer confidence, especially the option "a little better" was chosen more often. It seems likely that in a period of negative consumer confidence, the option "a little worse" would be chosen more often. This means that in periods of positive consumer confidence, the difference between positive and negative answers should be larger under the new design, whereas in periods of negative consumer confidence, this difference should be smaller. Therefore both the additive and the ratio adjustment methods are not plausible. It seems better to apply the correction on the percentages $p_{i,+}$ and $p_{i,-}$ instead of y_i. The percentages have to be between 0 and 100. The correction is applied in such a way that the correction becomes smaller when the percentage is close to these limits. This way, the probability is smaller that the corrected percentage is outside the interval [0,100]. This is realized by making the correction proportional to the population variances of the percentages. Let p^o denote the percentage under the old design. The corrected percentage $\tilde{p}^o_{i,t,s}$ is computed as:

$$\tilde{p}^o_{i,t,s} = \hat{p}^o_{i,t,s} + \beta_{i,s}\frac{\hat{p}^o_{i,t,s}(100-\hat{p}^o_{i,t,s})}{\hat{p}^o_{i,\tau,s}(100-\hat{p}^o_{i,\tau,s})}, \quad s=+,- \tag{5.1}$$

for all periods t before the changeover, $i = 1,...,5$ the 5 relevant questions of the CS. $\hat{p}^o_{i,t,s}$ is the estimated percentage under the old design and τ is the period of the parallel run. Dividing by the population variance of this period (mean of the three months) makes sure that the corrected percentages for these three months are close to the values observed under the new design during the parallel run. The values $\beta_{i,s}$ are the estimates for the discontinuity for the positive and negative answers, for example the values shown in Table 5.3, or other estimates based on the structural time series model, as will be discussed later.

Note that the size of the correction in formula (5.1) depends on the percentage $\hat{p}^o_{i,\tau,s}$ under the old design during the parallel run. When $\hat{p}^o_{i,\tau,s}$ is very small, a small change of the estimate of the discontinuity $\beta_{i,s}$ could have a large effect on the correction. For example, with $\hat{p}^o_{i,\tau,s} = 1$, $\hat{p}^o_{i,t,s} = 50$, $\tilde{p}^o_{i,t,s} = 62.6$ when the estimate of the discontinuity is $\beta_{i,s} = 0.5$, but $\tilde{p}^o_{i,t,s} = 75.4$ when the estimate of the discontinuity is $\beta_{i,s} = 1$. Given the standard errors of the estimates of the discontinuity, this shows that the accuracy of the discontinuity estimates would be unsufficient for reliable corrections in this case. In the application at Statistics Netherlands, $10 \leq \hat{p}^o_{i,\tau,s} \leq 90$ for the 5 considered variables, which makes the corrections less sensitive to small estimation errors.

Note furthermore that in extreme cases, with small values of $\hat{p}^o_{i,\tau,s}$, it is theoretically possible that the corrected percentage $\tilde{p}^o_{i,t,s}$ is not inside the possible range between 0 and 100. Then it is obvious that the assumptions under (5.1) does not hold, and another correction method should be developed. for example an alanysis based on a log ratio transformation.

Note that this backcast method is based on strong assumptions about what causes the discontinuity and about what the effects of the changes would be in times of a negative consumer confidence. The same correction method is also applied for the fifth question, where the questionnaire is not changed. The advantage of this method is that it is very unlikely that the corrected values are outside the

possible interval.

5.1.3. STM

For the proposed backcast method, estimates for the discontinuities of the percentages are needed. Instead of the model of Section 2.3, a multivariate model is used, where the three percentages of the positive, negative and neutral answers are modelled at the same time and it is ensured that the sum of the estimates of the discontinuities is equal to 100.

We use the following model

$$\mathbf{p}_t = \mathbf{L}_t + \mathbf{S}_t + \boldsymbol{\beta}' \mathbf{x}_t + \mathbf{I}_t, \quad t = 1, \ldots, T.$$

with $\mathbf{p}_t = (\hat{p}_{t,+}, \hat{p}_{t,0}, \hat{p}_{t,-})'$ direct estimates for the percentages, $\mathbf{L}_t = (L_{t,+}, L_{t,0}, L_{t,-})'$ the trend component, $\mathbf{S}_t = (S_{t,+}, S_{t,0}, S_{t,-})'$ the seasonal component, $\boldsymbol{\beta} = (\beta_+, \beta_0, \beta_-)'$ estimates for the discontinuities, $\mathbf{x}_t = (x_t, x_t, x_t)'$ the intervention variable, $\mathbf{I}_t = (I_{t,+}, I_{t,0}, I_{t,-})'$ the noise component.

The variables $L_{t,s}$, $S_{t,s}$, $s = +, 0, -$ are modelled as described in Section 2.3. The variable x_t is the same for the three series, as the discontinuity occurs at the same time. The restriction that β_+, β_0 and β_- add up to zero is enforced with the following transition equations in the state-space model:

$$\beta_{t,+} = \beta_{t-1,+}$$

$$\beta_{t,-} = \beta_{t-1,-}$$

$$\beta_{t,0} = -\beta_{t-1,+} - \beta_{t-1,-}$$

The subscript t indicate the notation of the transition equations. As there is no disturbance term, $\boldsymbol{\beta}$ is still time independent.

The white noise parameters represent the sum of noise due to sampling errors and noise in the population parameter. The sampling error depends on the sample size, which is approximately constant over time, and the percentage $\hat{p}_{t,s}$, $s = +, 0, -$ itself, as the variance of the direct estimate $\hat{p}_{t,s}$ is $var(\hat{p}_{t,s}) = \frac{\hat{p}_{t,s}(100-\hat{p}_{t,s})}{n}$ with $n \approx 1000$ the sample size. To take this into account, we model

$$E(I_{t,s}) = 0$$

$$cov(I_{t,s}, I_{t',s}) = \begin{cases} \hat{p}_{t,s}(100 - \hat{p}_{t,s})\sigma^2_{I,s} & \text{if } t = t' \\ 0 & \text{if } t \neq t' \end{cases}$$

When the estimate of $\sigma^2_{I,s}$ is around $\frac{1}{1000}$, the noise in the series is explained by the sampling error. This is the case for some of the series considered here. For some other series, the estimate of $\sigma^2_{I,s}$ is much larger (around $\frac{1}{300}$). This means that in these series there is substantial noise in the population parameter.

Table 5.4: Results STM with diffuse prior, estimates discontinuities, based on data until June 2019

	positive answers	SE	negative answers	SE
Econ. last 12 months	10.0	2.9	0.5	2.7
Econ. next 12 months	19.7	3.4	-0.3	3.3
Fin. last 12 months	9.7	1.2	2.2	1.5
Fin next 12 months	12.1	1.2	4.7	1.5
Major purchases	-6.6	1.8	1.4	1.6

In this application, there is no reason to assume that the variance of the noise parameter changes due to the redesign. In other applications, this could be the case. Then it is possible to take this into account by modelling different hyperparameters for the periods before and after the changeover.

With such a multivariate model, it is also possible to model correlations between the model parameters, for example between the disturbance terms of the slopes. This is not applied in this case.

This model is applied on series for the 5 questions of the CS. The series start in April 1986, and up to and including March 2017, the estimates are based on the old design. Estimates from April 2017 - June 2019 are based on the new design.

The estimates for the discontinuities which are found with this model are shown in Table 5.4. Here, the information from the parallel run is not used. The estimates are computed 2 years and three months after the redesign, with the data up to and including June 2019. The point estimates are comparable to the direct estimates based on the parallel run (Table 5.3) whereas the standard errors are substantially larger. These large standard errors are caused by properties of the CS-series. The series are quite flexible, therefore model predictions and estimates of the discontinuities are not very accurate. In other applications, it is possible that the model estimates of the discontinuities are more accurate than the direct estimates based on the parallel run (with the same size of the parallel run), see Van den Brakel et al. (2019) for an example.

The estimates of the discontinuities improve when more data gets available. Table 5.4 shows the most accurate estimates which can be computed in July 2019. In the first months after the changeover, less data were available, and less accurate figures could be computed. Figure 5.1 shows the estimates of the discontinuities based on the data up to and including April 2017 (first point on the x-axis) until June 2019 (last point on the x-axis). So the figures show how the estimates of the discontinuities evolve when more data is available (economic situation over the last 12 months is used as an example). It can be seen that the estimates are in the right order of magnitude from the beginning. Nevertheless there are some visible changes in the first 6 months. In this periods, the standard error of the estimates decreases substantially. After about 6 months, the point estimates and the standard errors are stable.

It can be expected that the estimates of the discontinuities improve when more data gets available also in other applications. Often, a period of a year (when monthly data is used) is sufficient to get close to the final estimate. However, it is possible in other applications that the estimate after a few

months is not yet reliable.

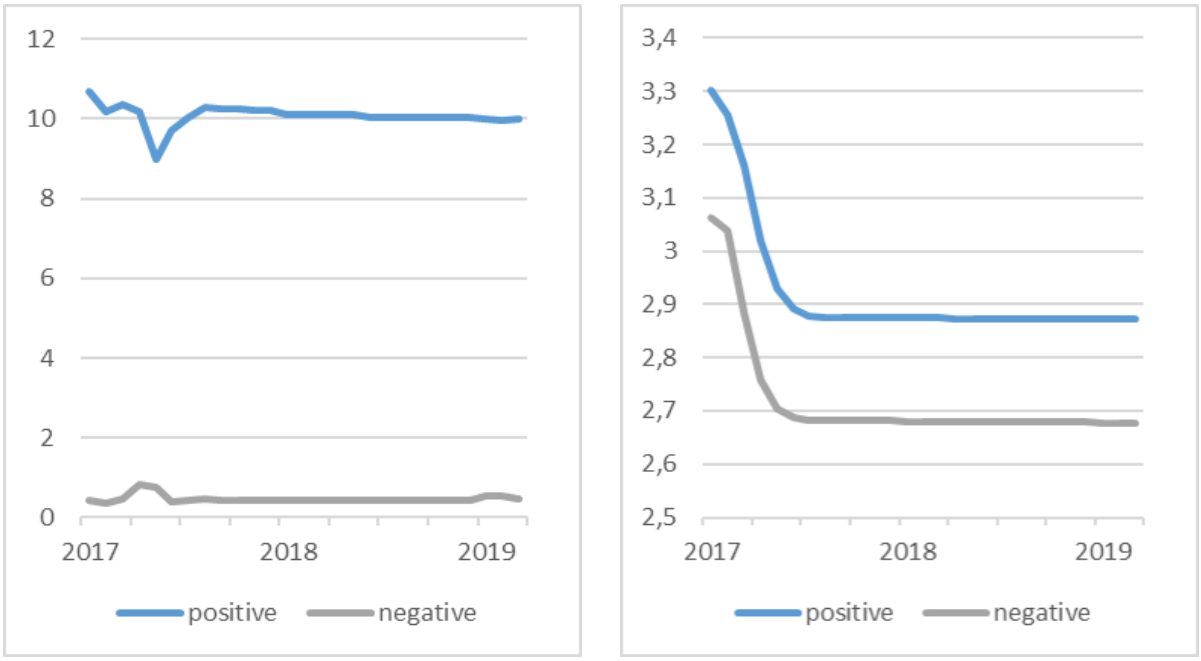

Figure 5.1: Development of point estimates (left panel) and standard errors (right panel) discontinuities with diffuse prior, economic situation last 12 months

Table 5.5: Results STM with exact prior, estimates discontinuities, based on data until June 2019

	positive answers	SE	negative answers	SE
Econ. last 12 months	11.2	1.1	-0.03	0.8
Econ. next 12 months	17.3	1.1	0.2	0.7
Fin. last 12 months	9.1	0.8	2.9	0.9
Fin next 12 months	11.7	0.7	3.6	0.8
Major purchases	-5.1	1.0	0.3	0.7

5.1.4. Combination

When the results of the parallel run are used as an exact prior in the structural time series model, the accuracy of the estimates is improved. The results are shown in Table 5.5. There are some small changes, compared to the estimates based on the parallel run only (Table 5.3), and the accuracy is slightly improved (compare with Table 5.3).

Figure 5.2 shows the development of the estimates of the discontinuities based on the data up to and including April 2017 (first point on the x-axis) until June 2019 (last point on the x-axis). Again, economic situation over the last 12 months is used as an example. Similar as in the situation with diffuse prior, there are some visible changes in the first 6 months, and in this period, the standard error of the estimates decreases. The changes are, however, much smaller, since the information from the parallel run is included.

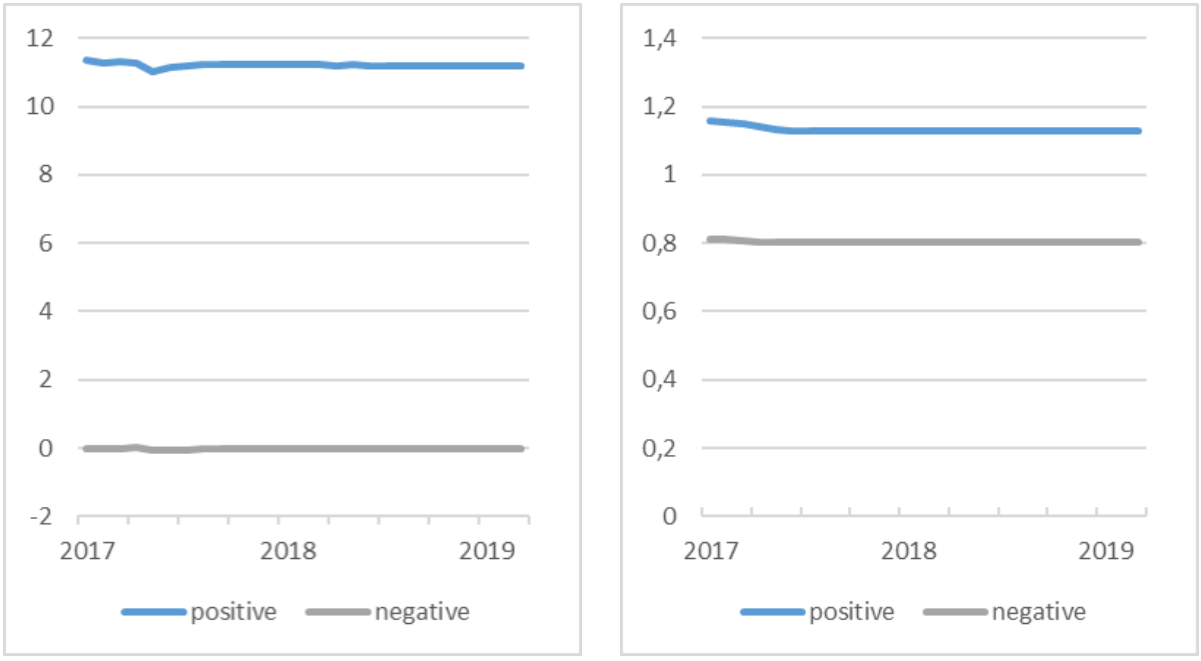

Figure 5.2: Development of point estimates (left panel) and standard errors (right panel) discontinuities with exact prior, economic situation last 12 months

5.1.5. Backcasting

Now, the backcast method described in Section 5.1.2 is applied. Figure 5.4 compares the original series under the old design and the corrected series for the percentages positive and negative answers. Again, economic situation over the last 12 months is used as an example. The discontinuities based on both the parallel run and the time series models (Table 5.5) are used here. The data under the new design is added. We see that the correction is larger for the positive answers than for the negative answers. This is since the discontinuity for the positive answers is much larger. We see furthermore that the correction is larger in periods where the percentage of positive answers is large. This is because of the chosen correction method. This correction of the percentages results in a correction of the differences, which is shown in Figure 5.4.

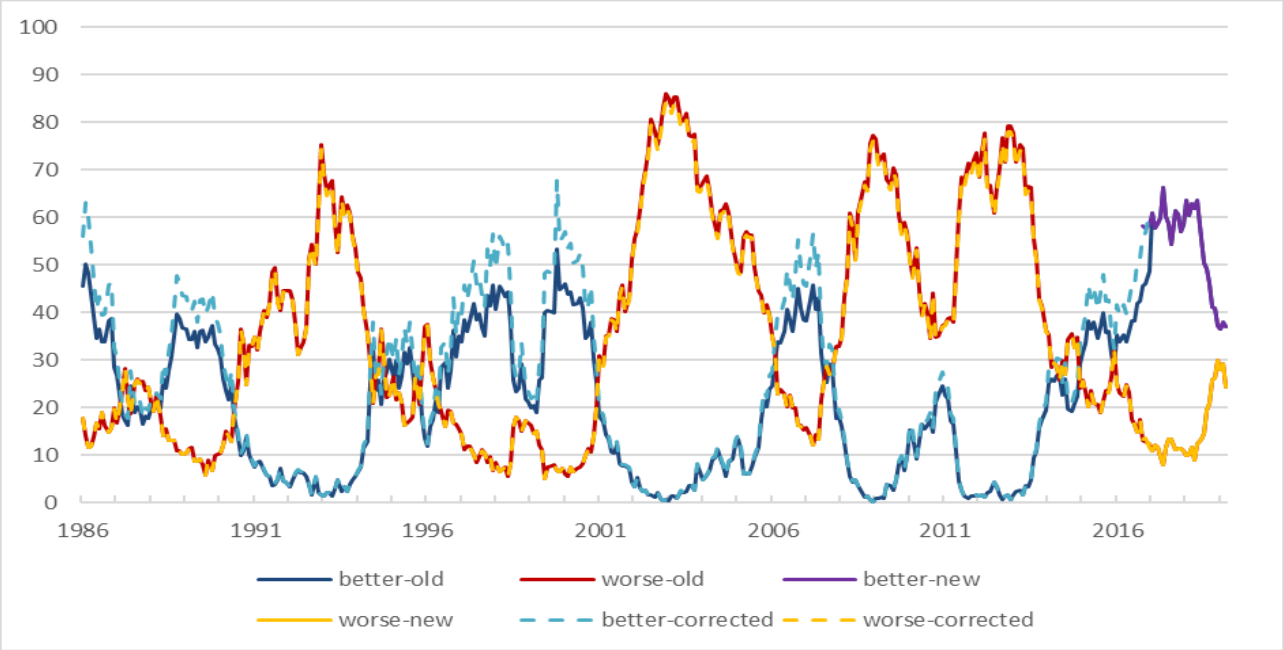

Figure 5.3: Estimates of percentages positive, negative under old design, new design, and old design corrected for discontinuity, economic situation last 12 months

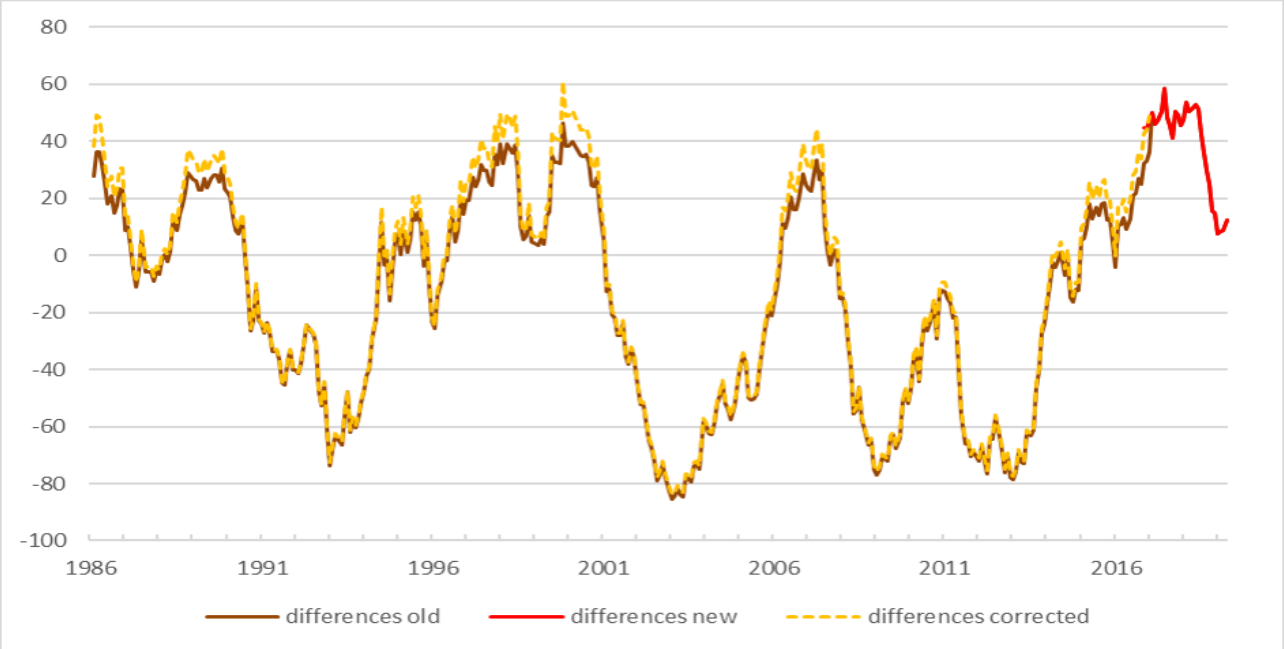

Figure 5.4: Estimates of differences under old design, new design, and old design corrected for discontinuity, economic situation last 12 months

6. Summary

When a survey is redesigned, it is likely that discontinuities in survey estimates occur. It is important that discontinuities are quantified, in order to separate the real development of the target parameter and changes due to the redesign of the survey process. When the series are corrected for the discontinuity, the interpretation of the changes is straightforward. In this report, different methods to quantify the discontinuity are discussed and applied to data from the UK and the Netherlands. The first method is based on a parallel run, the second one on the structural time series model and the third on a combination of both methods under which information from the parallel run and the time series can be combined to improve the estimates for the discontinuities. In the application to the Welsh survey data discontinuities are estimated both at aggregate and domain levels using design-based and model-based estimators. The results show that discontinuities are present in some surveys. In addition, model-based estimators improve the accuracy of the estimated discontinuities. In the application to the Dutch Consumer Survey, the estimates based on the structural time series model are less accurate, since the series are quite flexible. As a result the white noise of the population parameter is of the same order as the size of the sampling error. The variance of direct estimates of the consumer confidence only account for the sampling error and ignore the variance of the population parameter white noise. The time series model accounts for both sources of uncertainty and can be regarded as a more realistic measure of uncertainty, see Van den Brakel et al. (2017) for a detailed motivation. Different correction methods are discussed. All methods rely on strong assumptions about how the new design would have affected the estimates in the past. For the Dutch Consumer Survey, a tailor-made correction method is applied, as the assumptions of standard synthetic methods are not plausible. This method is applied to correct the series observed before the redesign to the level obtained with the new survey process. This corrected series is published as an official series of the Dutch Consumer Confidence.

This report does not resolve a number of open research questions. To start with additional research in model-based methods for estimating discontinuities is needed. This includes (a) jointly accounting for the fact that the sampling variance in the FH model is estimated and for covariate measurement error and (b) exploring model variations for example, the multivariate specification and the direct modelling of a discontinuity. Additional research is also needed in developing benchmarking techniques for discontinuities. Benchmarking is used in domain estimation to ensure that domain estimates are consistent with aggregate estimates at national level. However, how benchmarking can be applied when estimating and adjusting for discontinuities is currently an open research problem. Finally, as mentioned earlier in this report, the methods we present are currently applicable when working with survey data. Using new forms of data may also result in discontinuities. However, how to measure and adjust for discontinuities in this case is also an area that requires new research.

Bibliography

Bell, W., H. Chung, G. Datta, and C. Franco (2019). Measurement error in small area estimation: Functional versus structural versus naiÌˆve models. *Survey Methodology 45*(1), 61–80.

Bollineni-Balabay, O., J. Van den Brakel, and F. Palm (2016). 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 179*, 377–402.

Doornik, J. (2009). *An Object-oriented Matrix Programming Language Ox 6*. Timberlake, London.

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

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

Fienberg, S. E. and J. M. Tanur (1987). Experimental and sampling structures: Parallels divering and meeting. *International Statistical Review 55*, 75–96.

Fienberg, S. E. and J. M. Tanur (1988). From the inside out and the outside in: combing experimental and sampling structures. *Canadian Journal of Statistics 16*, 135–151.

Fienberg, S. E. and J. M. Tanur (1989). Combing cognitive and statisitcal approaches to survey design. *Science 243*, 1017–1022.

Groves, R. (2004). *Survey errors and survey costs*. John Wiley and Sons.

Harvey, A. and J. Durbin (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 149*, 187–227.

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

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

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

Särndal, C.-E., B. Swensson, and J. Wretman (1992). *Model assisted survey sampling*. Springer series in statistics. New York, NY, US: Springer-Verlag Publishing.

Van den Brakel, J. (2008). Design-based analysis of experiments with applications in the dutch labour force survey. *Journal of the Royal Statistical Society, Series A 171*, 581–613.

Van den Brakel, J., B. Buelens, and H. Boonstra (2016). Small area estimation to quantify discontinuities in sample surveys. *Journal of the Royal Statistical Society Series A 179*, 229–250.

Van den Brakel, J. and S. Krieg (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 R. H. Renssen (1998). Design and analysis of experiments embedded in sample surveys. *Journal of Official Statistics 14*, 277–295.

Van den Brakel, J. and R. H. Renssen (2005). Analysis of experiments embedded in complex sampling designs. *Survey Methodology 31*, 23–40.

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

Van den Brakel, J., P. Smith, and S. Compton (2008). Quality procedures for survey transitions - experiments, time series and discontinuities. *Survey Research Methods 2*, 123–141.

Van den Brakel, J., E. Söhler, P. Daas, and B. Buelens (2017). Social media as a data source for official statistics; the dutch consumer confidence index. *Survey Methodology 43*(2), 183–210.

Van den Brakel, J., X. Zhang, and S. Tam (2019). Measuring discontinuities in time series obtained with repeated sample surveys. *International Statistical Review To appear*.

Ybarra, L. M. R. and S. Lohr (2008). Small area estimation when auxiliary information is measured with error. *Biometrika 95*(4), 919–931.

You, Y. and B. Chapman (2016). Small area estimation using area level models and estimated sampling variances. *Survey Methodology 32*, 97–103.