Discussion Paper

Measuring discontinuities due to survey process redesigns

2017 | 13

Jan van den Brakel,
Xichuan (Mark) Zhang,
Siu-Ming Tam

Content

Summary

A key requirement of repeated surveys conducted by national statistical institutes is the comparability of estimates over time, resulting in uninterrupted time series describing the evolution of population parameters. This is often an argument to keep survey processes unchanged as long as possible. It is nevertheless inevitable that a survey process will need to be redesigned from time to time, for example, to improve or update methods or implement more cost effective data collection procedures. To avoid the implementation of a new survey process disturbing the comparability of estimates over time, it is important to quantify the impact on the estimates of a repeated survey. This paper presents a framework of statistical methods that can be used to measure the impact and manage the risk due to a survey process redesign.

Keywords

Survey sampling, Randomized experiments, Structural time series modelling, Small area estimation

Acknowledgement

The views expressed in this paper are those of the authors and do not necessarily reflect the views of the Australian Bureau of Statistics or Statistics Netherlands. The authors are thankful to the constructive input and comments received from Greg Griffiths, Tatiana Surzhina, Phillip Wise, Jonathan Blanchard, Oksana Honchar, Kristen Stone, Annette Kelly, and Harm Jan Boonstra.

1. Introduction

Official statistics produced by national statistical institutes (NSI) are based on probability samples or derived from administrative collections. Official statistics are typically published repeatedly with a monthly, quarterly or annual frequency with the purpose of building consistent time series that describe the evolution of population parameters of interest. In order to preserve the comparability of estimates, the underlying process of the survey is kept unchanged as long as possible. It is inevitable, however, that adjustment or redesign of this process is needed from time to time, since the existing procedures become gradually out-dated or more cost effective methods are required.

Survey samples contain besides sampling errors different sources of non-sampling errors that have a systematic effect on the outcomes of a survey. As long as the survey process is kept constant, this bias component is not visible. If, however, one or more components of the survey process are modified, the biases induced by these non-sampling errors are changed. Major redesign of the underlying survey process therefore generally has systematic effects on the survey estimates, disturbing comparability with figures published in the past. To minimize the impact for users it is therefore important that the effect of a redesign on the estimates of a survey can be quantified. This avoids confounding real change in the parameters of interest with changing measurement bias due to alteration in the survey process.

The purpose of this paper is to present a framework of methods that can be used to measure the impact of a survey transition and find the right balance between risk and costs during the implementation of the change-over. The method used to measure the impact will be determined by the type of change in the survey process. If the micro data collected under the old and new approaches are consistent, then the impact of a redesign can be assessed by processing sample data from the regular survey with the old and new process. This is typically the case if redesigns affect only the data processing part of a survey, for example, if new data editing methods, imputation methods or estimation approaches are implemented. Another example is the implementation of a new classification system. If it is expected that the micro data are not consistent under the old and new approaches, then it might be required to collect data under the old and new approach alongside of each other at the same time for some period of time. This is shortly referred to as parallel data collection. This typically occurs with modifications in the data collection phase of a survey process. For example the implementation of new questionnaires, data collection methods and field work strategies.

In the case of a sufficiently large parallel run, contrasts between direct or design-based sample estimates under the old and new approach can be used as estimates for the discontinuities, using design-based inference procedures for experiments embedded in probability samples, Van den Brakel (2008, 2013). In most cases the available budget does not allow to conduct a sufficiently large parallel run that meets

pre-specified precision and power requirements. In the most extreme case there is no parallel data collection at all. In that case time series models can be used to separate real period-to-period change from differences in bias due to a redesign of the process, for example with state-space intervention models. In many situations there is budget for a small parallel run. In that case model-based inference procedures, e.g. known from the realm of small area estimation, can be used to obtain precise as possible estimates for the discontinuities. This information can also be combined with state-space intervention models. In this case the information of the entirely observed series before and after the parallel run is used to further improve the estimates for the discontinuities.

The different methods presented in this paper illustrate that the choice of the most appropriate method depends on: 1) type of change in the survey process, 2) available budget for quantifying discontinuities, 3) importance of the indicators/statistics, 4) accepted level of risk for failing to detect or of detecting discontinuities at an inadequate level of accuracy, 5) accepted level of risk for disturbing the regular survey process and official publications of the survey, 6) required timeliness of the discontinuity estimates, and 7) the accepted amount of revisions.

Design and analysis of parallel runs combines the statistical methodology from sampling theory (Kish 1965, Cochran (1977), Hansen et al., 1953) with design and analysis of randomized experiments (Fisher, 1935, Cochran and Cox 1957, Scheffé 1959, Hinkelmann and Kemptorne 1994, 2005, and Searle 1971). This is not new, since embedding randomized experiments in sample surveys to estimate measurement error components dates back to Mahalanobis (1946), Fellegi (1964), Hartley and Rao (1978), and Fienberg and Tanur (1987, 1988, 1989) and is sometimes called interpenetrating subsampling or split-ballot designs. An interesting question is whether the design-based mode of inference from classical sampling theory or the model-based mode of inference from experimental design theory should be applied to analyse a parallel run. In the case of parallel runs with limited sample sizes model-based procedures known from small area estimation (Rao and Mollina, 2016 and Pfeffermann, 2002, 2013) are an appropriate alternative to design-based inference procedures (Van den Brakel and Renssen, 2005, Van den Brakel 2008, 2013, and Chipperfield and Bell, 2010) but require adaption to the specific context of parallel data collection.

The idea of improving survey estimates with times series models dates back to Scott and Smith (1974, 1977). Tam (1987), Tiller (1992), Rao and Yu (1994), Datta et al. (1999), Pfeffermann and Burck (1990), Pfeffermann and Bleuer (1993), Durbin and Quenneville (1997), Harvey and Chung (2000), and Pfeffermann and Tiller (2006) are some key references to authors that further elaborate on the idea of using time series models to improve the precision of survey estimates. In the context of this paper, time series models are useful to separate period-to-period change from differences in measurement bias induced by a redesign of the survey process.
The paper is structured as follows. In Section 2 some examples from major survey redesigns that have taken place in the past at Statistics Netherlands and the Australian Bureau of Statistics are discussed and are used in the rest of the paper for

illustrative purposes. In Section 2 it is also briefly illustrated how recalculation can be used to assess the impact of the change-over to a new classification system. In the rest of the paper, the focus is on methods for situations where the micro data under the old and new approach are not consistent. In section 3 methods for experimental designs embedded in sample surveys are reviewed. Section 4 continues with state-space intervention models including a discussion how a small parallel run can be combined with this method. Section 5 reviews small area estimation methods for the analysis of small sized parallel runs. After quantifying discontinuities, it might desirable to adjust series observed before the change-over to a new design. Different backcasting methods are reviewed in Section 6. The paper concludes with a discussion in Section 7.

2. Examples of survey redesigns

2.1 Implementation of a new classification system

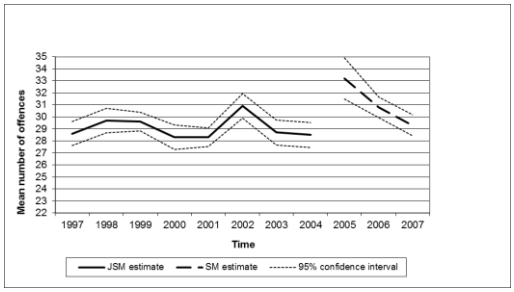
In 2008 a new classification system for economic activities (NACE) was introduced in the European Statistical System. The old economic classification (NACERev1.1) was replaced by the NACERev2.0. This is a typical example where the micro data used to compile short-term statistics and business statistics are completely compatible under both classification systems and the effect of the change-over can be estimated by reprocessing existing survey data. A smooth change-over form the NACEREv1.1 to the NACERev2.0 was still a major operation, since additional co-variates are required. The minimum requirement to quantify the effect of a new classification is that the units in the sample have indicators that specify to which domain an enterprise is classified under the old and new classification. Preferably, all enterprises in the sample frame are recoded according to both the old and new classification system, since this allows construction of more efficient domain estimators under both classifications. Finally to have sufficiently reliable estimates under both the old classification and the new classification, it might be necessary to reconsider the stratification scheme of the sample, the allocation of the sample over strata and possibly draw additional sample in the domains under the new classification with insufficient sample size. An extended discussion is provided by Van den Brakel (2010) and Smith and James (2017).

2.2 Redesigns of the Dutch Crime Victimization Survey

Until 2004, statistical information about crime victimization, unsafety feelings and satisfaction with police performance in the Netherlands was measured with the Justice and Security Module (JSM), which was one of the modules of the Permanent

Survey on Living Conditions (PSLC). This survey was based on a stratified two-stage sample of persons using Computer Assisted Personal Interviewing (CAPI) as data collection. Due to severe budget cuts in 2004, several modules of the PSLC were cancelled from the statistical program of Statistics Netherlands and the JSM continued as a new survey called the Crime Victimization Survey (CVS). To reduce administration costs, data collection changed to a mixed mode design of Computer Assisted Telephone Interviewing (CATI) and CAPI. In addition, the questionnaire was redesigned. At this time there was no budget for a parallel run, even at a reduced sample size. This change-over resulted in a strong and unexpected increase in the estimates of total number of offences as shown in Figure 1. These discontinuities came at a very inconvenient moment, since at that time the Ministry of Interior and Kingdom Relations had a strong police in place to reduce crime rates in the Netherlands.

Figure 1: Mean number of offences against Dutch inhabitants (hundreds) during the 12 months prior to the interview, observed with the JSM (1997-2004) and the CVS (2005-2007).



At this time, a time series modelling approach was applied at Statistics Netherlands for the first time to disentangle discontinuities from real period-to-period change. After the introduction in 2005, several redesigns of the Dutch CVS took place afterwards. To avoid difficult to explain differences in key figures of important surveys these redesigns were always accompanied with a small parallel run. The first redesign was already in 2008 where, the data collection mode changed from mixed-mode CAPI and CATI to a sequential mixed mode design based on web interviewing, Paper and Pencil Interviewing (PAPI), CAPI, and CATI. This also required adjustments in the questionnaire design. The new design was conducted with a sample size of about 6500 respondents in parallel with the regular design that is conducted at a sample size of 19000 respondents. The main publication domains for this survey are based on a breakdown of the Netherlands into 25 police regions. Assuming that discontinuities between police regions are similar, a parallel run with a size of one third of the regular sample size was considered to be sufficient to quantify discontinuities. Small area estimation was applied as an alternative, as explained in Section 5. To maintain uninterrupted series with the past, these parallel runs where repeated in 2009, 2010, 2011 and 2012, where design before the change-over was conducted at a sample size of about 6000 respondents yearly and the new design at the regular sample size of about 19000 respondents.

2.3 Redesigns of the Dutch Labour Force Survey

The Dutch Labour Force Survey (LFS) changed from a cross-sectional survey to a rotating panel in 2000. Each month a stratified two-stage sample of households enters the panel. These samples are observed five times with quarterly intervals. As a result each month data are collected in five independent samples that are observed for the 1st, 2nd, 3rd, 4th, and 5th time respectively and are further referred to as wave 1 through 5. A consequence of this change-over is that the effects of rotation group bias (RGB) (Bailar, 1975) became very visible in the figures of the LFS. In 2010 a multivariate structural time series model is implemented for the production of official monthly figures on the Labour Force as a form of small area estimation and to handle problems with RGB. The inputs for this model are five series of general regression (GREG) estimates (Särndal et al., 1992) on a monthly frequency based on the five waves of the panel, following the approach proposed by Pfeffermann (1991).

This model is also used in combination with a parallel run to account for discontinuities due to two major redesigns in 2010 and 2012. In 2010 Statistics Netherlands was faced with large budget cuts on the data collection. At that time it was foreseen that the data collection in the first wave of the LFS must change from uni-mode CAPI to a sequential mixed-mode starting with web interviewing and a follow up by CATI and CAPI to realize the required cutbacks. It was felt that there was not enough experience with web interviewing in household surveys to implement this sequential mixed-mode design directly in the LFS. It was therefore decided to change in 2010 from uni-mode CAPI to a mixed-mode design using CAPI and CATI. This also required a major revision of the questionnaire. To allow for CATI data collection in the first wave, the length of the questionnaire had to be reduced by moving blocks from the questionnaire in the first wave to the follow-up waves. In the meantime the sequential mixed mode design was built with the intent of eventually changing to the final sequential mixed-mode design based on web interviewing, CATI and CAPI in 2012. Also the questionnaire was adjusted again to allow for web interviewing. On both occasions the first wave of the DLFS was conducted in parallel at full sample size for a period of six months, while discontinuities in the remaining waves are estimated through the time series approach proposed in Section 4.

2.4 Redesigns of and planning for measuring discontinuities in the Australian Labour Force Survey

The Australian LFS is based on a rotating panel design where each month a multi-stage area sample of dwellings is selected. Households selected in these monthly samples are interviewed using face-to-face, phone or web form each month for eight consecutive months. As a result, each month data are collected in eight independent samples or waves. The estimation method used in the Australian LFS is based on a model-assisted estimator in conjunction with the Best Linear Unbiased Estimator (BLUE) composite estimator (Bell, 2001).

In 2004, an embedded experiment was conducted to quantify the effect of Computer Assisted Interviewing (CAI) in the Australian LFS. In this case the CAI was gradually introduced using a phase-in approach from October 2003 to August 2004. To quantify the statistical impact, 10% of the regular sample was assigned to the new approach and 90% to the regular approach by randomly assigning 10% of the interviewers (rather than dwellings) to the treatment group, with additional interviewers added in the treatment group in subsequent months. In February 2004, the fraction of interviewers assigned to the new survey was increased to 40%, in June to 70%, and 100% in August. This was to minimise the impact of the CAI on the labour force estimates, maximise the sample size for estimating discontinuities and increase the time to respond to adverse effects of the CAI. This phase-in approach had the advantage of increasing power as the sample size grew over time, and could be implemented relatively quickly.

In preparation of a major redesign of the Australian LFS, an eight dimensional structural time series model is developed, where series of GREG estimates based on the eight waves of the rotating panel design are the input for the model. Similar to the model for the Dutch LFS (Subsection 2.3) this model accounts for RGB and autocorrelation between survey errors due to panel overlap. The model is developed to evaluate different scenarios for estimating discontinuities. The rotating panel design in combination with the composite estimator has a smoothing effect on abrupt discontinuities in the separate waves. The implementation of a new design takes eight months before it is implemented in each of the eight waves. As a result discontinuities in the LFS figures are gradually introduced in a period of eight months. Second, the composite estimator combines past observations of GREG estimates at the level of separate waves, resulting in a further smoothing of possible discontinuities. To avoid this smoothing effect, discontinuities are modelled in eight dimensional time series model at the level of the separate waves. More details are provided in Subsection 4.4.

3. Design and analysis of experiments for estimating discontinuities

If it is expected that the micro data are not consistent under the old and new approach, then a straightforward and safe approach to quantify discontinuities is collect data under both approaches at the same time alongside each other for some period of time. Ideally this is based on a randomized experiment that can be embedded in the probability sample of the survey. This implies that experimental units are selected randomly from an intended target population using a probability sample, generally the sample design of the survey under consideration. Subsequently

the sample is randomized over different treatments according to an appropriate experimental design.

Embedding randomized experiments in probability samples generally increases the validity of the results. Theory of experimental design focusses on establishing the causality between difference in treatments and observed effects. Selecting the experimental units randomly using probability sampling enables the generalization of conclusions observed in an experiment to larger target populations. This is particularly important if experiments are conducted to obtain quantitative insight into the effect of a new survey process on the outcomes of a repeated survey. In such cases, a design-based inference procedure for estimating discontinuities in sample estimates for unknown population parameters naturally fits with the purpose of probability sampling to generalize conclusions to larger target populations. A design-based inference procedure for this type of experiments is proposed by Van den Brakel and Renssen (1998, 2005) and Van den Brakel (2008, 2013) and Chipperfield and Bell (2010). In the case of insufficiently large sample sizes, model-based approaches can be considered as an alternative. Randomized selection from a target population and randomized assignment to different treatments avoids making strong ignorability assumptions and makes results more robust for model miss-specification.

Embedding experiments in ongoing sample surveys is efficient since the regular survey serves as the control group in the experiment and is simultaneously used for the regular publication purposes of the survey. Another strong point of a parallel data collection is the low risk level for the regular publications during the change-over to the new design. This approach can avoid the risk of a period without data for regular publication should the new approach turn out to be a failure. Through a well-designed experiment the risk of failing to detect a discontinuity is minimized since the design of an experiment gives full control over the minimum detectable difference at a pre-specified significance and power level.

A further major advantage of parallel data collection is that it facilitates the production of timely estimates for impact measurements. Estimates for the discontinuity can be made directly after finalizing the field work. If the sample size meets the pre-specified precision requirements for estimating the discontinuities then there is no need for revisions of the estimated discontinuities when results of subsequent editions of the new survey become available. This in contrast to the use of time series methods, where there will be updates when future figures under the new approach become available.

A disadvantage of a parallel run is that it is not cost neutral since additional data collection is required. Obtaining sufficiently precise estimates for the discontinuities often requires sample sizes for the new approach that come close to the regular sample size. See Subsection 3.4 for details on power considerations. Designing and conducting an experiment for parallel data collection that accurately measures the discontinuities due to the change-over also significantly increases the complexity of the fieldwork organisation.

A parallel run requires a careful planning and preparation and key decisions on various themes. This will be detailed in the following subsections.

3.1 Purpose of the experiment

In planning a parallel run, a clear definition is needed about the treatments to be tested and the number of factors to be included in the experiment. It is crucial to decide whether the experiment is intended to estimate the difference between the old and new design as a whole or whether the purpose is to explain the effect of the underlying factors that changed. If the purpose is to estimate the net effect of the change-over, a two-sample experiment where the old and new approaches are compared is the most effective. If the purpose of the experiment is to explain the individual contributions of the factors that changed in the redesign, then a factorial design is required. This is at the cost of a reduced power for estimating the overall discontinuity or impact of the new design. This can be seen by noticing that the overall discontinuity is one of the interactions of a factorial setup, namely the contrast between the subsample where all factors are on the level of the old design and the subsample where all factors are on the level of the new design.

An alternative is to consider a factorial design with an unbalanced setup, which means that a major part of the sample size goes to the subsamples of the treatment combinations that define the old and new designs. Small sample sizes are assigned to all other treatment combinations. Unbalanced set ups are, however, less efficient in terms of precision and power to estimate contrasts.

The analysis of a parallel run must not be allowed to result in further modifications to the new survey process, since that will immediately outdate the results of the parallel run. This prohibition on tweaking the new process during the parallel run implies that smaller field experiments or pilots should precede a parallel run to fine tune the final design of the new survey process. These experiments also provide insight into the impact of the underlying factors that will be modified in the redesign. Once sufficient insight is obtained into the effects of these factors, a parallel run can be designed as a two-sample experiment to just measure the net impact, which maximises the power for estimating discontinuities.

In the case of continuing surveys it might also be desirable to quantify the impact on the seasonal effects. This will, however, significantly increase the length and sample size of the parallel run and will practically almost always be infeasible. In practice it is more realistic to estimate such effects with a time series modelling approach, see Section 4.

3.2 Design and field work considerations

It is the typical approach followed in design of randomized experiments to clearly specify in advance the hypotheses about the main effects and possible interactions to be tested. An advance decision is needed on the smallest size of the treatment effects or discontinuities that should result in a rejection of the null hypotheses of zero impact at a pre-specified significance and power level. This can be used to derive the minimum required sample size and gives full control over the minimum detectable differences in the experiment. By pre-specifying the hypotheses to be tested in the experiment, post-hoc analyses can be avoided to protect inflating the over-all significance level of tests on treatment effects.

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 allows eliminating 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 structure like strata, primary sampling units, clusters and interviewers. For details see Fienberg and Tanur (1987, 1988, 1989), Van den Brakel and Renssen (1998, 2005), and Van den Brakel (2008).

A decision related to the experimental design is the choice of the level of randomization. From a statistical point of view it is optimal to randomize the ultimate sampling units over the treatments. This results in the maximum number of degrees of freedom for variance estimation and thus optimizes the power of an experiment. Due to field work restrictions it might be necessary to randomize clusters of ultimate sampling units over the treatments. For example interviewers with the clusters of respondents assigned to them or all household members belonging to the same household. This, however, reduces the number of effective experimental units available for variance estimation and thus reduces the power of the experiment. An important question to address is whether an interviewer should conduct the different treatments in the experiment or if they can be assigned to one of the treatments only. In the first case interviewers can be used as the block variables in an RBD. If on the other hand interviewers can be assigned to one treatment only, it is also worthwhile to consider a double blind set-up, to avoid that the awareness of participating in an experiment might influence their normal behaviour, even unconsciously. The choice of whether interviewers should conduct different treatments or not finally depends on the type and number of treatments tested in the experiment and the experience of the field staff with conducting field experiments. Conducting different versions of questionnaires with subtle differences in wording is surely more complicated for interviewers than testing differences in advance letters. When embedded experiments were conducted at Statistics Netherlands for the first time at the end of the nineties, there was a strong resistance against designs were interviewers conduct more than one treatment. This gradually

decreased as the field staff became more and more familiar with conducting fieldwork for embedded experiments.

The available budget for impact measurement will always put restrictions on the maximum available sample size for a parallel run and thus for the power of detecting differences. Different options can be considered to optimise the power of the experiment. First of all attempts must be made to avoid the effective sample size of the experiment being reduced due to field work restrictions relating to difficulties randomizing the ultimate sample units over the treatments instead of clusters of sampling units as in the example of Subsection 2.4. If it is not possible to assign interviewers to more than one treatment combination, alternatives should be considered before choosing a design where clusters of sampling units assigned to the same interviewer are randomized over the treatments. In the case of computer assisted personal interviewing and a small sample size for the alternative treatment, randomizing interviewers over the treatments might result in unacceptable large travelling distances for the interviewers assigned to the alternative treatment. Van den Brakel and Van Berkel (2002) proposed an experiment where initially experimental regions are created by taking the union of two neighbouring interviewers. Then the sample units in these regions are randomized over the old and new approach as well as the two interviewers. This slightly increased the usual travelling distance for the interviewers but still allowed randomizing the ultimate sample units over the treatments.

One way to increase the precision of the parallel run is to increase the period of parallel data collection. If the regular survey used for official publication purposes is used as the control group, then the total number of observations of the units in the control group is automatically increased due to the extended duration of the parallel collection, and at no extra cost as these units was going to be observed as part of the regular survey anyway. If the original number of observations of the treatment group is now spread over the longer period, then the costs for the parallel run are not increased. This results in an unbalanced setup for the experiment. Although unbalanced designs are less optimal for estimating contrasts, this still increases the power of the experiment without increasing additional costs since the sample size of the regular survey increases.

Another option is to reduce the sample size of the regular survey to the benefit of increasing the sample size of the new approach. If this results in a more balanced allocation over the treatments, then the power of the experiment is increased at the cost of loss in precision for official publications. This approach was utilised by the parallel run of the Dutch National Travel Survey in 1998, see Subsection 3.4. Small area estimation techniques might also be considered to compensate for this loss in precision.

If a small impact is expected, consider using the data under the alternative approach for regular publication purposes. This, of course, increases the risk of introducing impact in the official publications. This risk might be manageable if the period of the parallel run is long and the allocation over the regular and new approach suitably

unbalanced. In the Dutch LFS this approach was applied to test the effect of a new advance letter on the response rates. Ten percent of the regular sample was allocated to the new approach but observations obtained under this group were still used for publication purposes (Van den Brakel, 2008).

Another way to optimize the power is to restrict the experiment to the most important research questions. This implies that the number of treatments and factors as well as the number of target variables that are analysed are restricted to a minimum. In the case of a parallel run, a two sample set up that only test the net effect of all underlying factors that changed has a larger power compared to a factorial set up, which is required if the main effects of the different factors also have to be explained. If the number of target variables is restricted, the loss of power as a consequence of applying simultaneous comparison methods is avoided as much as possible.

A useful general framework and practical guidelines for planning and conducting experiments is given by Robinson (2000). More details about practical issues in the design of embedded experiments and planning of the field work is given by Van den Brakel et al. (2008).

3.3 Mode of inference

Estimating systematic differences between finite population parameters observed under different survey implementations implies the existence of measurement errors. Regardless of the mode of inference, a measurement error model is required to explain systematic differences between a finite population parameter observed under different surveys implementations.

Let y_{ik} denote the observation obtained from respondent i assigned to survey approach or treatment k. One approach is to assume that responses obtained in an experiment can be modelled as $y_{ik} = \theta_i + \gamma_k + \varepsilon_{ik}$, with θ_i the true intrinsic value of the variable of interest of respondent i, γ_k a systematic treatment effect or measurement bias related to the k-th treatment or survey approach and ε_{ik} a random measurement error for respondent i observed under treatment k. Population means are defined as $\bar{Y}_k = \frac{1}{N}\sum_{i=1}^{N} y_{ik} \equiv \Theta + \gamma_k$. The random measurement error cancels out by taking the expectation over the measurement error model and assuming that the random measurement errors are zero in expectation. The true population parameter Θ cannot be observed due to measurement bias. Even in the case of a complete enumeration under the regular survey we observe, say, $\bar{Y}_{reg} = \Theta + \gamma_{reg}$. In the case of a probability sample, we obtain an approximately design-unbiased estimator for $\bar{Y}_{reg}$, say $\hat{\bar{Y}}_{reg}$. In a similar way, the population parameter under the new design is defined as $\bar{Y}_{new} = \Theta + \gamma_{new}$, with $\hat{\bar{Y}}_{new}$ an approximately design unbiased estimator based on observations obtained from a probability sample. Discontinuities are in fact the relative differences between the selection and measurement bias of two different survey implementations, i.e.

$\beta = \bar{Y}_{reg} - \bar{Y}_{new} = \gamma_{reg} - \gamma_{new}$, estimated using either a model-based or design-based inference mode as detailed below.

The standard literature for design and analysis of experiments applies model-based inference procedures for the analysis of experiments (Montgomery (2001) and Hinkelmann and Kempthorne (1994, 2007)). 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. For a one-way ANOVA, e.g. observations are assumed to be a realization of the linear model $y_{ik} = \alpha + \beta_k + \varepsilon_{ik}$, with y_{ik} the observation obtained from sampling unit i assigned to treatment k, α an intercept, β_k the treatment effects and ε_{ik} normally and independently distributed residuals. If α is identified as the sample mean under the control group (or the regular survey), then β_k can be interpreted as the discontinuities or relative measurement bias between the regular and new survey implementation. A drawback of this approach is that the sample design is ignored, which might result in biased estimates for the discontinuities if the sample design is not self-weighting, 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), Van den Brakel (2008, 2013) and Chipperfield and Bell (2010) 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. This approach starts with deriving an approximately design-based estimator for $\bar{Y}_k$, which is approximately design unbiased with respect to both the sampling and experimental design and an approximately design-unbiased estimator for the variance of the contrasts between $\hat{\bar{Y}}_k$, $k \in \{reg, new\}$. An estimate for the discontinuity is simply $\hat{\beta} = \hat{\bar{Y}}_{reg} - \hat{\bar{Y}}_{new}$. This gives rise to design-based Wald statistics in assessing hypotheses about $\hat{\beta}$. See Van den Brakel and Renssen (2005) and Van den Brakel (2008, 2013, 2016) for conditions where these design-based Wald statistics coincide with the F-tests of a standard model-based ANOVA.
The advantage of a design-based approach is that it accounts for the complexity of the sample design and facilitates the generalisation of the results observed in the experiment to the intended finite target population from which the sample is drawn. In addition it simplifies the interpretation of the results, since the estimated treatment effects or discontinuities are in terms of differences between the estimated population parameters as they are defined in the survey. Many parameters in sample surveys are defined as ratios of two estimated population totals and test on treatment effects for ratios is given by Van den Brakel (2008). In a standard model-based analysis it is not immediately clear how to estimate the impact on such parameters. An additional advantage of a design-based mode of inference is that it is more robust against model misspecification than standard model-based modes of inference, even if the data are obtained under a randomized experiment. Model-based procedures on the other hand will have stronger power, conditional on the underlying model assumptions holding.

Griffiths et al. (2016) explored generalized linear modelling of observations obtained from sampling units for the estimation of discontinuities. Adding strong auxiliary information to the model increases the precision of the analysis and might be considered as an alternative for using a formal randomized experiment for the parallel run or as an alternative for a parallel data collection at all. This requires, however, strong assumptions that observations are missing at random, Rosenbaum and Rubin (1983). See also Pfeffermann and Landsman (2011) for methods that attempt to avoid strong ignorability assumptions like latent variable models, and instrumental models. In addition Pfeffermann and Landsman (2011) propose an approach for observational studies that is based on a population model and a treatment selection model and apply a combined likelihood to avoid strong ignorability assumptions. Basing a parallel run on a randomized experiments has the advantage that randomisation protects against model misspecification and avoids using strong ignorability assumptions. An analysis based on a randomized experiment will therefore be more robust to model misspecification compared to data where there is no randomised experiment. For survey transitions without a period of parallel data collection, model based methods e.g. the time series modeling approach in Section 4 are better alternatives to separate discontinuities from period-to-period change.

3.4 Examples

At Statistics Netherlands and the ABS, parallel runs have been conducted frequently. Several examples are given in Subsections 2.2, 2.3, 2.4 and 3.2. Another early example is the redesign of the Dutch National Travel Survey in 1998. This is an annual cross sectional survey aimed to measure travel behaviour of the Dutch population. To improve response rates of this survey a complete redesign of the fieldwork and data collection method was introduced. The old and new design where conducted in parallel for a period of one year. The regular sample size amounted in that year to about 13,000 addresses per month. This sample size was increased to about 15,000 addresses per month. During the first two quarters of the year, each month 13,000 addresses were assigned to the old design and 2,000 to the new design. After six months it was decided to gradually increase the fraction of addresses assigned to the new approach and decrease the fraction assigned to the old approach. Observations obtained under the old approach served to compile official statistics about travel behaviour for that year. This approach resulted in a (acceptable) decrease of the regular sample size and allowed estimation of discontinuities. In this case, it was decided to gradually increase the sample size assigned to the new design and decrease the sample size of the regular design since this provided a low risk transformation to a new complex field work strategy. This came, however, at the cost of reduced precision for the regular publication, not only because the sample size was reduced but also because the variation between the design weights was increased to compensate for the large differences in inclusion probabilities between months.

4. Estimating discontinuities using state-space intervention models

If, due to lack of budget or field work capacity, the new survey process is implemented without parallel data collection, then the discontinuity might be estimated by fitting a structural time series model to the observed series. A structural time series model decomposes an observed series into several unobserved components like a trend, seasonal component, a cycle and regression components to account for auxiliary series. The remaining unexplained variation is modelled with a white noise component. Remaining serial autocorrelation beyond these components can be captured with Auto Regressive or Moving Average components. Trend, seasonal and cycle components are modelled with specific stochastic processes, which allow them to be time dependent and adapt gradually to changing dynamics in the observed series. For an introduction in structural time series modelling, see Harvey (1989) or Durbin and Koopman (2012).

In the case of modelling series observed with repeated sample surveys, a measurement model is required that describes the series of sample estimates as decomposed into a true, but unknown, finite population parameter and a sampling error. Using similar notation as in Section 3, this implies that $\hat{\bar{Y}}_{k,t} = \bar{Y}_{k,t} + e_{k,t} = \Theta_t + \gamma_k + e_{k,t}$, where Θ_t denotes the true population parameter for period t, $\bar{Y}_{k,t}$ the value obtained if Θ_t is observed under a complete enumeration using treatment k for time period t, $\hat{\bar{Y}}_{k,t}$ a design-based estimate for $\bar{Y}_{k,t}$, γ_k the measurement bias if the population parameter Θ_t is measured under the k-th treatment or survey approach, and $e_{k,t}$ the sampling error. Note that it is assumed that measurement bias is time independent. Assuming for the population parameter an appropriate trend model (say L_t), a dummy seasonal component or a trigonometric seasonal component (say S_t) and white noise (say I_t) for the unexplained variation and inserting it into the measurement model, the following model is obtained for the observed series $\hat{\bar{Y}}_{k,t} = \mathrm{L}_t + \mathrm{S}_t + \mathrm{I}_t + \gamma_k + e_{k,t}$. In the case of (rotating) panel designs, the white noise of the population parameter and the survey errors are identifiable, see Pfeffermann (1991). In the case of cross-sectional surveys, both components are confounded and estimated as one term, say $v_{t,k} = \mathrm{I}_t + e_{k,t}$. To account for heteroscedasticity due to changing sampling sizes or design over time, the variances of the direct estimates can be used as prior information in the variance of the measurement equation of the state space model, see Binder and Dick (1989, 1990) or Krieg and Van den Brakel (2012).

To separate period-to-period change from the discontinuity an intervention component is added to the model which describes the effect in the outcomes at the moment that the survey process changes from the old to the new design. The most straightforward approach is a level intervention which means that a dummy indicator,

say δ_t, is added to the model that changes from zero to one at the moment of the change-over to the new survey process. The regression coefficient of this intervention variable can be interpreted as the discontinuity or impact induced by the redesign of the survey process. Note that the measurement bias γ_k induced by a particular treatment or survey implementation cannot be observed using the survey data only. Similar to parallel runs and experiments, this approach only allows estimating the relative difference in measurement bias between the survey process before and after the change-over, i.e. $\beta = \gamma_k - \gamma_{k'}$ (where k' and k refer to the survey approach before and after the change-over respectively). The measurement bias ($\gamma_{k'}$) of the survey approach before the change-over will typically be absorbed in the trend component of the population parameter, i.e. $\tilde{L}_t = L_t + \gamma_{k'}$. These considerations finally result in the following time series model for the observed series: $\hat{\tilde{Y}}_{k,t} = \tilde{\mathrm{L}}_t + \mathrm{S}_t + \beta\delta_t + v_{k,t}$.

This state space intervention model is proposed by Harvey and Durbin (1986) for analysing the effects of seatbelt legislation on road casualties in the UK. Van den Brakel and Roels (2010) applied this to series obtained with repeated surveys to estimate discontinuities for situations where there is no parallel data collection. Other possible interventions due to survey redesigns, for example of the slope of the trend or the seasonal component, are discussed by Van den Brakel and Roels (2010). This approach relies on the assumption that the time series model for the population parameter models the real evolution correctly. All deviations from this evolution are interpreted as discontinuities due to the change-over. Recall from Subsection 3.1 that it might be desirable to quantify the effect of a redesign on the seasonal effects. A state-space intervention is particularly suitable to estimate discontinuities on seasonal effects in a cost effective way but might requires several years of observation before it can be estimated with sufficient reliability.

4.1 Advantages and disadvantages

A major advantage of the time series approach is that no additional data collection is required, which makes this approach very cost effective. In addition, the complications of embedding a parallel run in the daily field work of a national statistical institute are avoided. Other advantages of the time series modelling approach is that all available data under both the old and the new approach are used, since the entire observed series is used to estimate discontinuities. The state space method has, in addition, the flexibility to combine information in the entire series with the information obtained with parallel data collection. By casting the model in a state space form (Harvey, 1989), the parameters of the model can be estimated using the Kalman filter. For state variables without any a-priori information, typically a diffuse initialisation of the Kalman filter is used. i.e. the initial value for the state variables equal zero with a large value for their variances, expressing that this initial estimate is highly uncertain. Design-based estimates for discontinuities including their variances obtained with a parallel data collection, however, can be used in an exact initialization of the Kalman filter, Van den Brakel and Krieg (2015). Additional

information that becomes available after the implementation of the new survey is used to further improve the estimates for the discontinuities.

An alternative way of combining information from partial overlap is to define a separate series for the regular and new approach, say $\hat{\tilde{Y}}_{reg,t}$ and $\hat{\tilde{Y}}_{new,t}$. Let $t = \tau, \tau+1, \dots, \tau'$ denotes the period of overlap of both series, i.e. the period of the parallel run. This implies that $\hat{\tilde{Y}}_{reg,t}$ is observed from $t = 1, \dots, \tau'$ and is missing for $t = \tau'+1, \dots, T$. Similarly $\hat{\tilde{Y}}_{new,t}$ is observed from $t = \tau, \dots, T$ and missing for $t = 1, \dots, \tau-1$. These series can be combined in a bivariate model:

$$\begin{pmatrix} \hat{\tilde{Y}}_{reg,t} \\ \hat{\tilde{Y}}_{new,t} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \left(\tilde{\mathrm{L}}_t + \mathrm{S}_t \right) + \begin{pmatrix} 0 \\ \beta \end{pmatrix} + \begin{pmatrix} \upsilon_{reg,t} \\ \upsilon_{new,t} \end{pmatrix}.$$

The Kalman filter can handle missing values in the series in a way similarly to the EM-algorithm (Durbin and Koopman, 2012 section 2.8 and 4.8). Numerical problems, however, might be expected if there are only a few paired observations or if the number of observations under the new approach is short. Therefore the univariate model that uses the information from the parallel run through an exact initialization is a more parsimonious and computationally more efficient way of handling this problem.

Another strong advantage of the time series modelling approach is that it simultaneously solves other problems, for example small area problems and rotation group bias in rotating panels. Van den Brakel and Krieg (2015) developed a structural time series model for the Dutch Labour Force Survey that solves problems with small sample sizes for estimating Labour Force figures on a monthly frequency, rotation group bias and discontinuities due to two major redesigns in 2010 and 2012. Skipping a period of parallel data collection and relying on a time series model to estimate discontinuities has also several disadvantages and risks. The figures obtained under the first edition of the new survey are in fact disregarded, since with only one observation under the new approach this method comes down to modelling this observation as an outlier. Furthermore, estimates for the discontinuities will change if new observations under the new survey become available. As a consequence revisions must be accepted. The size of these revisions mainly depends on the dynamics of the trend component. As the trend component becomes more volatile only local observations before and after the change-over influence the level of the trend which reduces the size of revisions of the estimated discontinuities. See Van den Brakel and Roels (2010) for more details. As a result of these revisions, final estimates are not timely.

Implementing the change-over without a period of parallel data collection or pretesting implies increased risk levels during the change-over. If after the change-over, the new approach turns out to be a failure and it is decided to fall back on the old approach, then there is a period where no data are available for the production of official statistics. Another factor that contributes to an increased level of risk is that real developments and estimates for the discontinuities are confounded if the real evolution of the population parameter deviates from the assumed time series model. This situation can for example occur, if the change-over of the survey coincides with the start of the Global Financial Crisis. Finally there is no control over

the precision and size of the minimum observable impact with the time series modelling approach, resulting in an increased risk of failing to detect relevant discontinuities. The minimum detectable differences depend on the stochastic behaviour of the series. This in contrary to a parallel run designed as an embedded experiment, where power calculations offer full control over the minimum observable differences. Simulations give an indication of the minimum observable differences with the time series modelling approach as illustrated in Subsection 4.2. In the case of large numbers of publication domains the time series approach rapidly becomes complicated. Consistency restrictions between different aggregation levels can be imposed on the regression coefficients that model the discontinuities in multivariate structural time series models, Van den Brakel and Roels (2010). If the number of domains increases numerical problems can be expected.

4.2 Combining parallel run with time series modelling approach

The risks associated with a time series modelling approach are reduced if a parallel run with a reduced sample size is conducted. During this period the final decision about the change-over to the new design can be made. The estimates for the discontinuities can be used to initialize the Kalman filter to further improve the precision of these estimates with sample information that becomes available after the change-over. At the same time the amount of revision and the time required to obtain stable estimates compared to the situation without a parallel run will be reduced. Finally the assumption that the time series model correctly disentangles real developments from measurement bias is relaxed.

For different scenarios involving a major redesign of the Australian LFS, power calculations are being conducted to establish the required sample size of a parallel run to detect discontinuities. An initial requirement was that a difference of one standard error of the monthly unemployment labour force figures must be detected at 5% significance level with a power of 80%. One standard error at the national level equals 19,500 unemployed or 2.5% of the unemployed labour force. To observe a difference of 2.5% at a 5% significance level with a power of 80%, requires a parallel run that estimates a discontinuity with a standard error that is equal to or smaller than 0.9%. To achieve this precision with a parallel run, the regular and new survey must be conducted in parallel, both at the regular monthly sample size, for a period of 18 months. One option to reduce costs is to conduct the treatment group at a reduced sample size or for a shorter period and combine this information in a state space model through an exact initialization of the Kalman filter.

A simulation is conducted for different options to obtain an indication how long it takes to achieve the required precision. The Australian LFS is based on a rotating panel design. The eight dimensional state space model described in Subsection 2.4 is used for this simulation to generate 100 replicates of the series of unemployed labour force, for each wave separately. Discontinuities are added that differ between the waves (discontinuities simulated are at 15%, 5%, 2%, 0.5%, 0.01%, 0%, 0%, and -0.5% for the eight subsequent waves respectively). Simulations for the unemployed

labour force at the national level are conducted to illustrate the precision of the impact estimates obtained with the time series model approach without a parallel run and three different scenarios for parallel runs of reduced sample sizes as summarized in Table 1.

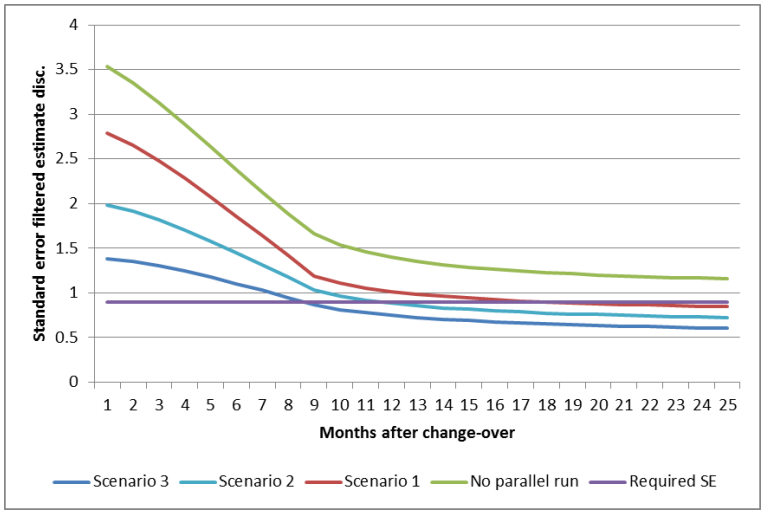
Table 1: Standard errors of discontinuities for different scenarios of parallel runs used in the simulation.

Scenario	Parallel run period	Sample size control sample*)	Sample size treatment sample*)	Standard error estimated discontinuity
1	18 months	100%	20%	2.79%
2	12 months	100%	50%	1.98%
3	18 months	100%	50%	1.38%

*): Percentages refer to the sample size of the regular LFS

Figure 2 illustrates for these scenarios the standard errors of the filtered estimates for different periods after the changes-overs. Standard errors for the filtered estimates for the scenario without a parallel run are also included. The horizontal line refers to the minimum required standard error of 0.9%. For the scenario without a parallel run the standard errors converges to a value of about 1.2, which implies that under this scenario the pre-specified precision requirement cannot be achieved. For scenario 1, 2 and 3 it takes respectively 18, 12 and 9 months after the change-over before the minimum required standard error of 0.9% is achieved.

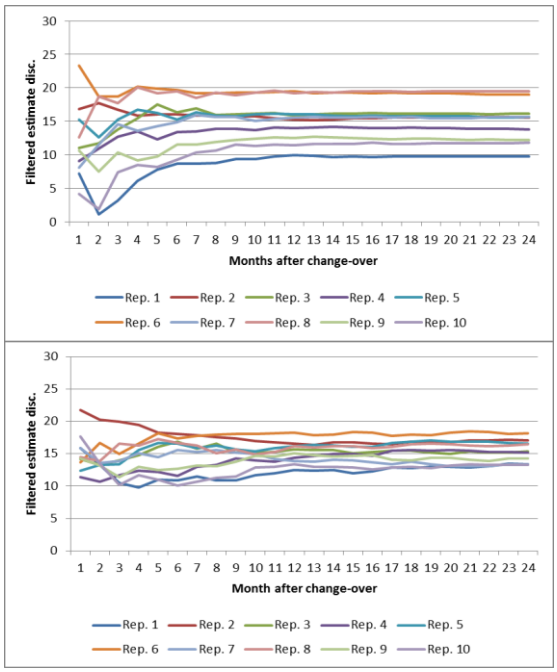
Figure 2: Standard errors of the filtered estimated discontinuity obtained with the time series model for different periods after the change-over for an exact initialisation of the Kalman filter with the scenarios from Table 1 and a diffuse initialisation (no parallel run). The horizontal line refers to the pre-specified minimum required standard of 0.9%



To illustrate the volatility of the impact estimates if there is no parallel run, the left panel of Figure 3 shows for each of 10 replicates how the time series model

estimates the simulated discontinuity if more observations become available after the change-over in the first rotation group with an assumed impact of 15%. As can been seen, it takes about 12 months before a stable estimate for the impact in a particular wave is obtained. The right panel contains similar estimates but now combined with the information obtained with a parallel run under scenario 3 in Table 1 above. The time series model further improves the impact estimates, and the volatility of the estimates directly after the change-over is clearly reduced.

Figure 3: Impact estimates for ten replicates of one wave obtained with the time series model for different periods after the change-over without a parallel run (left panel) and with a parallel run according to Scenario 3 (right panel). Assumed value of the discontinuity is 15%.



A consequence of improving the results of a relatively small parallel run with a time series model is that the initial estimates of the statistical impact obtained with the parallel run are likely to be revised after, for example, a period of 12 months. Through the simulation an estimate of the expected amount of revision is calculated under each of the three parallel run scenarios in Table 1 combined with a time series model after 12 months. As expected, the size of the revisions decreases with the sample size of the parallel run. The expected revision is about 5.8% under Scenario 1, 4% under Scenario 2 and 2.7% under Scenario 3.

4.3 Improving discontinuity estimates with auxiliary series

The use of a state space intervention model for separating real change over time from discontinuities without a parallel run relies on the assumption that the time series model for the population parameter correctly captures the real evolution. If

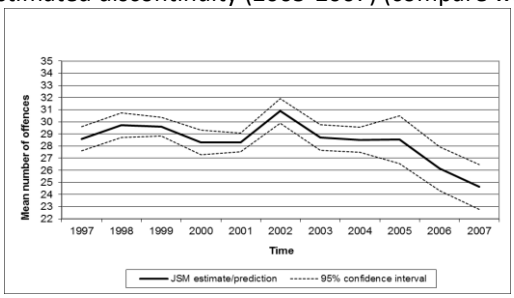
strongly related auxiliary series are available, seemingly unrelated time series equation (SUTSE) models can be applied to incorporate this auxiliary information to better separate real period-to-period change from discontinuities. The series observed with the survey are combined with the auxiliary series in a multivariate state space model. SUTSE models, model the correlation between the disturbance terms of the trend, seasonal, cycle or irregular components. In the case of strong correlation, the covariance matrices of the disturbance terms become of reduced rank, implying that the component of say K observed series are driven by less than K common components. This implies that the series are cointegrated.

Harvey and Chung (2000) combined a series obtained with the Labour Force Survey and claimant counts in the UK to improve the precision of estimates of change in the monthly unemployment figures. Van den Brakel and Krieg (2015) and Zhang and Honchar (2016) proposed to use auxiliary series to improve the robustness of the intervention state space approach against model misspecification. The auxiliary series indeed help to better separate the real evolution of the population parameter from discontinuities, particularly in the case of sudden changes in the evolution of the population parameter due to, for example, the Global Financial Crisis. These models, however, rely on the strong assumption that the correlation between the disturbances of both series is constant over the entire observed time period. If due to any reason the correlation between series gradually changes, the real evolution of the population parameter will be biased as well as the estimates for the discontinuities. Van den Brakel and Krieg (2016) proposed as an alternative a multivariate model that models the differences between the series with state variables that are time dependent, resulting in a model that is more robust for the assumption of time invariant correlations between the series.

4.4 Examples

In Subsection 2.2 the change-over from the Dutch JSM to the CVS in 2005 without any parallel data collection was introduced. To disentangle real evolution in the Dutch crime rates from the discontinuities, the state space intervention approach was introduced. The estimated discontinuity obtained with the time series model was used to construct an uninterrupted series by adjusting the estimates after the change-over to the level of the observed series before the change-over, as depicted in Figure 4. Adjusting observations after the change-over to the level before the change-over in this example reflects a preference and choice considered to be the best for users. Estimates under the old approach before the change-over can be adjusted to the level of the new approach in a similar way, referred to as backcasting (see Section 6).

Figure 4: Mean number of offences against Dutch inhabitants (hundreds) during the 12 months prior to the interview, observed with the JSM (1997-2004) and the CVS corrected for the estimated discontinuity (2005-2007) (compare with Figure 1).



Since this approach is cost effective and avoids the complications for the field staff for designing and conducting a parallel run, this method has been often used since then. For example in the redesign of the Dutch National Travel Survey in 2010. Starting in 2004 the field work for this survey was conducted by a marketing research bureau. In 2008 concerns about the data quality arose, which finally resulted in a termination of the contract with this bureau in 2009. As a result the data collection was transferred back to Statistics Netherlands from 2010. This was also an opportunity to redesign the sample and data collection process. Due to the conflict that resulted in terminating the contract with the marketing research bureau, a parallel run was not an option. In this case discontinuities are estimated with a state space intervention model. In this application there are two complicating factors. The first problem was the strongly reduced reliability of the estimates in the two years before the change-over. The finally selected time series model allows for a time varying variance structure of the measurement equation to reduce the influence of these years. The second problem is the large number of publication domains with consistency requirements. This problem was solved by applying high dimensional multivariate models where consistency constraints are applied to the regression coefficients of the interventions in the system equation of the state space model. Final estimates for the discontinuities were obtained using the data observed until 2014.

Other applications of the time series approach are the implementation of a sequential mixed-mode data collection using Web Interviewing (WI), CATI and CAPI and a new questionnaire in the Dutch Health Survey in 2010 and the Occupational Accident Monitor in 2015. In both cases the lack of budget for having a parallel run with sufficient power, even at the national level, was the decisive consideration in choosing a time series modelling approach.

Bollineni-Balabay, Van den Brakel and Palm (2016) applied a multivariate state space model to the domains of the Dutch Transportation Survey as a form of small area estimation and to account for discontinuities in the level as well as in the variances of the observed series. They avoid consistency problems by deriving estimates at the national level from the domain estimates obtained with a multivariate time series model. One approach to account for heteroscedasticity due to gradually changing sample sizes or modifications in the sampling design is to make the variance of the measurement equation in a state space model proportional to design variances of

the input series. In this case design variances are calculated from the micro data and used as a-priori information in the state space model. As an alternative, for example if no design variance estimates are available from the micro data, the variance of the disturbance terms of the measurement equation can be made time dependent by defining separate variance components for different periods.

Time series modelling and parallel runs can also be applied simultaneously as in the case of the Dutch LFS. In Subsection 2.3 a time series model for the rotating panel of the Dutch LFS is briefly introduced as a form of small area estimation and a method to account for RGB and discontinuities. The change-over from a cross-sectional design to a rotating panel design resulted in discontinuities due to introduction of RGB. In the time series model, mentioned in subsection 2.3, estimates for the population parameters are benchmarked to the level of the GREG estimates in the first wave to make publications comparable with the period before the change-over to the rotating panel design. This assumes that the observations obtained in the first wave are the most reliable and requires that, in particular, the data quality of the first wave should be as high as possible. With respect to the change-over to a new survey process in 2010 and 2012, it was recognized that discontinuities in the first wave must be estimated as precise as possible. It was therefore decided to allocate the available budget for a parallel run to the first wave only. For the same reason it was also decided to estimate the net effect of all factors that changed simultaneously in a two-sample experiment instead of quantifying the separate effects in a factorial design. Discontinuities in the follow up waves are modelled with the state space intervention approach. Unreliable estimates in these waves directly after the change-over do not impact the population parameter estimates because the population parameter estimates are benchmarked to the first wave estimates. This approach facilitated a smooth transition from the old to the new design without disturbing regular publication. Details can be found in Van den Brakel and Krieg (2015). Zhang and Honcar (2016) also provide an example of using auxiliary series to improve the estimates of discontinuities in the Australian LFS.

5. Estimating discontinuities for small areas

A parallel run analysed with a design-based mode of inference requires large sample sizes to observe differences that are comparable with the precision of the regular survey, with sufficient power. The advantage of this approach is that randomization through an embedded experimental design in combination with a design-based mode of inference provides a form of built-in robustness against model misspecification. The opposite approach is to have no period of parallel data collection and fully rely on a state space intervention model to separate discontinuities from real evolution of the population parameter of interest. In practice there is often a limited budget to conduct a parallel run at reduced sample size. This might require a model-based

inference procedure to obtain sufficiently precise estimates of discontinuities for small areas, using for example small area estimation methods. It can be seen as an intermediate case between a full scaled parallel run analysed with a design-based approach as described in Subsection 3.3 and a change-over without a parallel run where discontinuities are estimated by relying on a time series model.

As explained in Subsection 2.2, the change-over to the CVS in 2005 resulted in a strong unexpected increase in estimates of the total number of offences. The state space intervention approach was successfully applied to separate real developments from the discontinuities induced by the redesign at the national level. In the redesign of 2008 it was nevertheless decided that for important surveys like the CVS, the risks of a time series approach, as summarized in Subsection 4.1, should be avoided by estimating discontinuities using a parallel run at a reduced scale.

In the case of the CVS, the regular survey used for official publication purposes, was conducted at full scale while the alternative approach was conducted at a reduced sample size. It is also possible to reduce the sample size of the regular survey to partially finance the parallel run, for example both arms of the parallel run at 75% of the regular sample size. In these situations, small area estimation techniques can be put in place to compensate for the loss of precision in the regular publications and also to improve the precision of the estimates under the alternative approach, particularly for small areas. A wide range of small area estimation procedures are available in the literature to improve the effective sample size of the alternative survey, the reduced regular survey or both. These methods typically rely on explicit statistical models that use temporal information from preceding periods or cross-sectional information from other domains to improve the effective sample size for a specific domain and time period. Cross-sectional information is typically based on multilevel models, Fay and Herriot (1979) and Battese, Harter and Fuller (1988). These models can be extended to time series multilevel model to incorporate sample information from preceding periods, Rao and Yu (1994), Datta et al. (1999).

Subsection 4.1 has already explained how the structural time series modelling approach can combine temporal information from the entirely observed series with information obtained in a small parallel run with an exact initialisation of the Kalman filter to further improve the precision of the discontinuity estimates. If the sample size of the regular survey is reduced during the parallel run this approach can at the same time be used as a form of small area estimation, to obtain more precise model based estimates for the regular survey and to compensate for the loss of precision during the parallel run. This generally implies, however, that the mode of inference for the production of official figures changes from a design-based to a model-based approach.

Instead of time series methods, multilevel models can be considered to take advantage of cross-sectional information. These methods are relevant if estimates for discontinuities at disaggregated level are required. Small area estimation procedures strongly rely on correlated auxiliary information to borrow sample information from other domains. Applications in regular ongoing surveys use auxiliary information

available from other surveys, administrative sources or censuses. In the case of a parallel run where the regular sample is conducted on full sample size, reliable design-based estimates for the target variables are available for at least the planned domains, i.e. the domains for which the sample is designed to produce estimates with minimum precision requirements. These estimates are potentially strong auxiliary variables for use in a Fay-Herriot model to predict the variables under the alternative design, conducted at a lower sample size. Using similar notation as in Sections 3 and 4, let $\hat{\bar{Y}}_{d,new}$ and $\hat{\bar{Y}}_{d,reg}$ denote the direct estimate under the new approach and regular approach for domain d respectively. To improve the precision of the $\hat{\bar{Y}}_{d,new}$ these direct estimates are modelled in a Fay-Herriot multilevel model: $\hat{\bar{Y}}_{d,new} = \bar{Y}_{d,new} + e_{d,new} = \alpha' X_d + \upsilon_{d,new} + e_{d,new}$, with $e_{d,new}$ the sampling error, X_d a vector with auxiliary variables to explain the domain variables and α a vector with regression coefficients. Finally $\upsilon_{d,new}$ is the random component that models the unexplained variation between the domains. Obviously $\hat{\bar{Y}}_{d,reg}$ is highly correlated with $\hat{\bar{Y}}_{d,new}$ since it measures the same population parameter only using a different survey approach. Using $\hat{\bar{Y}}_{d,reg}$, among others, as auxiliary variables in the vector X_d leads to a Fay-Herriot model containing auxiliary information with error: $\hat{\bar{Y}}_{d,new} = \alpha' \hat{X}_d + \upsilon_{d,new} + e_{d,new}$. This is an application of Ybarra and Lohr (2008) who derived empirical best linear unbiased estimators, say $\tilde{\bar{Y}}_{d,new}$, for the Fay Herriot model using auxiliary variables observed with sampling error.

There are technical issues with the estimation of the variance of the impact or discontinuity. Let $Var(\hat{\bar{Y}}_{d,reg} - \tilde{\bar{Y}}_{d,new})$ denote the variance of the contrast of interest. Since $\tilde{\bar{Y}}_{d,new}$ uses $\hat{\bar{Y}}_{d,reg}$ or related estimates from the regular survey as auxiliary variables in the model to construct a small area prediction, there is a strong positive correlation between $\tilde{\bar{Y}}_{d,new}$ and $\hat{\bar{Y}}_{d,reg}$. Another issue is that $\hat{\bar{Y}}_{d,reg}$ and its variance is obtained through a design-based mode of inference, while $\tilde{\bar{Y}}_{d,new}$ including its measure of uncertainty is obtained through a model-based mode of inference. What inference mode should be chosen for the covariance between $\tilde{\bar{Y}}_{d,new}$ and $\hat{\bar{Y}}_{d,reg}$? Van den Brakel, Buelens and Boonstra (2016) proposed design-based estimators for the MSE of $\tilde{\bar{Y}}_{d,new}$ and the covariance between $\tilde{\bar{Y}}_{d,new}$ and $\hat{\bar{Y}}_{d,reg}$, resulting in design-based estimators for $Var(\hat{\bar{Y}}_{d,reg} - \tilde{\bar{Y}}_{d,new})$.

This approach was applied to estimate discontinuities in the parallel of the CVS in 2008, see subsection 2.2. With a parallel run of about 6,500 observations under the alternative approach and 19,000 observations under the regular approach, reliable direct estimates for the discontinuities were obtained at the national level. Under the assumption that these discontinuities holds for the underlying domains, this parallel run has sufficient power to detect discontinuities at the level of the 25 domains (police regions). To relax the strong assumption that discontinuities in all domains are equal to the discontinuity at the national level, the above described small area estimation approach was finally applied to quantify discontinuities for the police regions. Direct estimates from the regular survey provided strongly correlated auxiliary information in small area prediction models for the alternative approach and therefore improved the precision of the domain estimates under the alternative approach and thus the domain estimates of the discontinuities. Due to the strong

positive correlation between $\tilde{\hat{Y}}_{d,new}$ and $\hat{\hat{Y}}_{d,reg}$, the MSE of the discontinuities are further reduced.

Instead of using the aforementioned univariate Fay-Herriot model for the small scale sample for the new approach, it is also possible to model the direct estimates under the regular and new approach simultaneously in a bivariate Fay-Herriot model:

$$\begin{pmatrix} \hat{\hat{Y}}_{d,reg} \\ \hat{\hat{Y}}_{d,new} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \Theta_d + \begin{pmatrix} \gamma_{d,reg} \\ \gamma_{d,new} \end{pmatrix} + \begin{pmatrix} e_{d,reg} \\ e_{d,new} \end{pmatrix},$$

with Θ_d the unknown domain parameter, $\gamma_{d,reg}$ and $\gamma_{d,new}$ the measurement bias related to the regular and new survey approach respectively and $e_{d,reg}$ and $e_{d,new}$ the sampling errors of the new and regular sample. If the samples are drawn independently from each other, then the sampling errors can be assumed to be uncorrelated. The measurement bias $\gamma_{d,reg}$ and $\gamma_{d,new}$ cannot be observed, only the difference between them; i.e. $\beta_d = \gamma_{d,reg} - \gamma_{d,new}$. This gives rise to the following multivariate Fay-Herriot model:

$$\begin{pmatrix} \hat{\hat{Y}}_{d,reg} \\ \hat{\hat{Y}}_{d,new} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \bar{Y}_{d,reg} + \begin{pmatrix} \upsilon_{d,reg} \\ \upsilon_{d,new} \end{pmatrix} + \begin{pmatrix} e_{d,reg} \\ e_{d,new} \end{pmatrix}.$$

The differences between the predictions for the random domain effects $\upsilon_{d,reg}$ and $\upsilon_{d,new}$ can be used as an estimate for the discontinuity β_d. Assuming a full correlation matrix for $\upsilon_{d,reg}$ and $\upsilon_{d,new}$ allows for random effects for the discontinuities and models the correlation between $\hat{\hat{Y}}_{d,new}$ and $\hat{\hat{Y}}_{d,reg}$. The precision of the estimated discontinuities is improved by increasing the effective sample size within the domains with cross-sectional correlations. In addition a positive correlation between $\upsilon_{d,reg}$ and $\upsilon_{d,new}$ further decreases the standard error of the estimated discontinuities. This approach avoids the technical problems with variance estimation encountered under the univariate Fay Herriot approach, since it is completely casted in a model-based framework, and is currently being explored at Statistics Netherlands.

6. Adjusting time series for discontinuities to preserve comparability over time

After quantifying the impact of a redesign, the question can be raised of whether series observed in the past should be adjusted to the level of the new approach. This is often called backcasting. As an alternative, the series observed under the new approach can be adjusted to the level of the series observed before the change-over. This seems often less natural, although there are sometimes reasons to do this (see the example in Section 4.4 and the example below).

Backcasting methods are often based on synthetic approaches that rely on the strong assumption that the observed discontinuities are time invariant. Let $\hat{\hat{Y}}_{T,reg}$ and $\hat{\hat{Y}}_{T,new}$ denote the estimates obtained during a parallel run, say in period T, respectively

under the survey approach before and after the change-over. Additive adjustments simply subtract the contrast $(\hat{\bar{Y}}_{T,reg} - \hat{\bar{Y}}_{T,new})$ from the series observed before the change-over to make them 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 change-over with a factor $\hat{\bar{Y}}_{T,new}/\hat{\bar{Y}}_{T,reg}$ and assume that the adjustment is proportional to the level of the observed series. This can be useful to avoid adjusting variables that cannot take negative values outside their valid range. For proportions an adjustment can be made proportional to the population variance of the observed proportion to reduce the possibility that the adjusted series have values outside their admissible range. In this case, the adjusted series observed before the change-over is obtained by $\hat{\bar{Y}}_{t,reg} - (\hat{\bar{Y}}_{T,reg} - \hat{\bar{Y}}_{T,new})[\hat{\bar{Y}}_{t,reg}(100 - \hat{\bar{Y}}_{t,reg})/\hat{\bar{Y}}_{T,reg}(100 - \hat{\bar{Y}}_{T,reg})]$ with $\hat{\bar{Y}}_{t,reg}$, t=1,…, T-1, the values to be adjusted. Alternatively, the analysis of the discontinuities, including the adjustment, can be applied to the logratio transformed values (Aitchison, 1986). This transformation avoids adjusted values taking values outside their admissible range but can also result in extremely large adjustments, depending on the value of the proportions to be adjusted.

If a structural time series model is applied to the series to assess the impact, then adjusted series directly follow from the assumed model. A filtered or smoothed signal plus the intervention serves as a backcast series while a filtered or smoothed signal without the intervention results in an adjustment of the new approach to the level of the series observed before the change-over. It is also possible to use the estimates for the discontinuities with the time series model in the aforementioned synthetic method. For example, using the estimated level shift obtained with the time series model as input in a ratio adjustment or an adjustment for proportions can be considered. It is, however, preferable to build a time series model that implies the preferred adjustment, for example by applying a log transformation to the observed series to have a proportional adjustment or a log-ratio transformation for proportions, (Van den Brakel and Roels (2010)).

Literature on backcasting indices and deflated series is limited. The general accepted approach is to backcast the underlying series of the variables required to calculate indices or deflated series, for example turnover, deflation prices and the weights used to aggregate indices. In the next step the indices can be recalculated using the backcast input variables. See Smith and James (2017), Nolan et al. (2008) and James (2008) for details and issues with backcasting indices.

Applying methods that assume that the observed discontinuities are time invariant to backcast series over longer time intervals is often not realistic. The implementation of a new economic classification system is usually necessary since the structure of the economy gradually changes. As a result an existing classification becomes out-dated and cannot describe the structure of the economy satisfactory. Therefore the European statistical system changed from the NACE Rev. 1.1 to NACE Rev. 2 in 2010. Most European countries assessed the impact of this change-over by having a double coded register for one or two years in order to calculate business statistics under both classifications. The ABS applied the similar dual coding approach to handle the

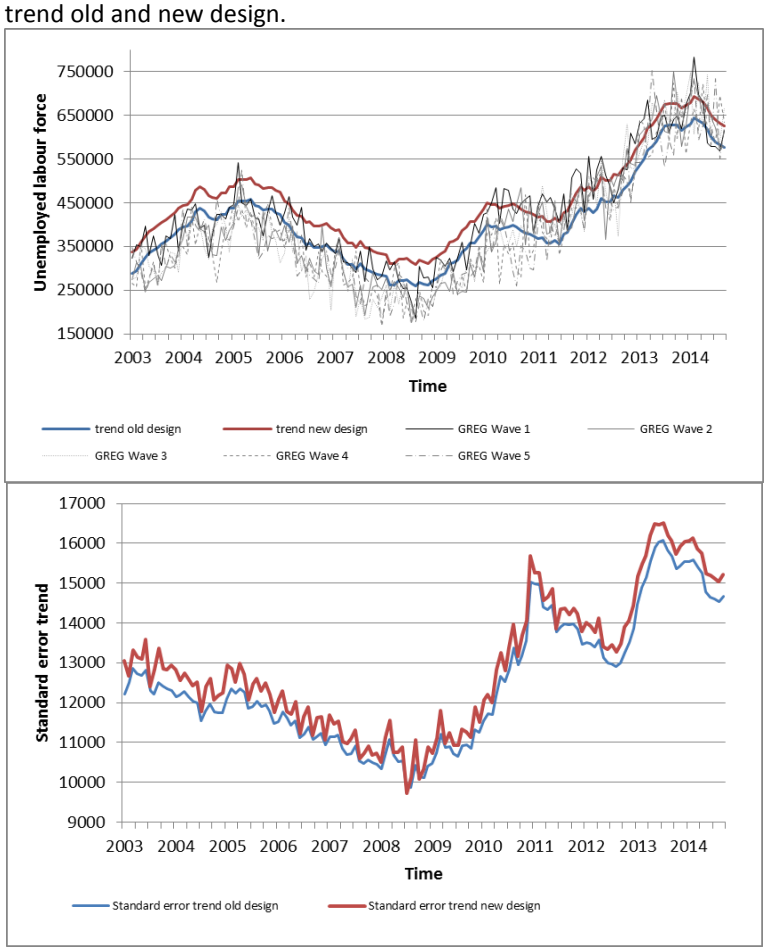
industry coding standard change from ANZSIC1993 to ANZSIC2006 in 2009. Backcasting in this context generally proceeds via use of transformation matrices which specify the distribution of enterprises over the categories of the new classification within each domain of the old classification. The assumption that these distributions are time independent is generally not tenable since the reason for the change-over was the changing structure of the economy. For backcasting purposes it might therefore also be useful to have at least the sample units from previous editions of the survey recoded according to the new classification or to use other external information that better allows for making more realistic backcasts. More technical details about the implementation of a new classification system can be found in Smith and James (2017) and Van den Brakel (2010), and ABS practices in Zhang (2002 and 2008) and ABS (2009) to handle the length of a backcast in form of exponential decay and retaining the seasonally adjustment movement, as well as how to backcast cross classifications of variables for which backcasts are already available. For example, producing backcasts for state by industry series while the discontinue factors were estimated and backcasts were performed only for the marginal estimates, i.e. state totals and industry category totals directly for ABS business surveys.

The aforementioned considerations might be a reason for a national statistical institute to only publish the discontinuities at the moment of the change-over. This is a safe approach, which avoids making unrealistic strong assumptions that discontinuities are time invariant but still useful since it avoids confounding real developments from discontinuities induced by the redesign for the period directly before and after the change-over. This approach, however, moves the problem of constructing a consistent time series to the data users. It can also be argued that the national statistical institute, as a collector of the data, has the best knowledge to produce adjusted uninterrupted series. The final decision also depends on the available information to quantify the discontinuities.

As explained for the Dutch LFS in Subsection 2.3, the change-over in 2010 from CAPI uni mode to a mixed-mode design using CAPI and CATI was the first step to the introduction of a sequential mixed-mode design based on web interviewing, CATI and CAPI in 2012. During the period from 2010 to 2012 the structural time series model, described in Subsection 2.3 and 4.4, was used to publish figures at the level of the old design before 2010. After implementing the final design in 2012, the official publication series changed to the new level of series produced by that final redesign. Under the strong assumption that discontinuities are time invariant, the series published before 2012 were also backcast to the level of the new process. In this way data users where confronted only once with the side effects of the transition.
As an illustration, the top panel in Figure 5 shows results for the monthly figures of the unemployed labour force at the national level. The figure shows the five series of the GREG estimates observed in the five waves of the rotating panel, which are the input series for the model. The solid blue line is the filtered trend benchmarked to the level of the first wave of the design applied before 2010. These trends are published until the implementation of the final design in 2012. Until 2010 the filtered trend is indeed equal to the level of the GREG series in the first wave, which is in the

upper bound of the band of the input series. According to the parallel run, the change-over to the intermediate design in 2010 resulted in an increase of the estimated unemployed labour force of 55,000 persons. Based on the second parallel run the final design implemented in 2012 resulted in an additional decrease of the estimated unemployed labour force with 2,000 persons. The net effect of the two redesigns is an increase of the estimated unemployed labour force of 53,000 people. Since the filtered trend at the level of the old design is corrected for the upward shocks in the input series, it drops after 2010 to the lower bound in the band of the input series. The solid red line in the top panel of Figure 5 is the filtered trend benchmarked to the level of the first wave of the design applied after 2012. Between 2010 and 2015 this trend is equal to the level of the GREG series in the first wave. The backcast trend before 2010 is high compared to the input series.

Figure 5: Monthly figures of the Dutch unemployed labour force at the national level. Top panel: Filtered trend under the old and new design compared with the observed GREG series in the five waves of the rotating panel. Bottom panel: standard errors filtered trend old and new design.



One of the consequences of these two redesigns is an increased level of uncertainty in the filtered estimates. The bottom panel of Figure 5 compares the standard errors of the filtered trend under the old and the new design. During the period between 2003 and 2009, the standard errors gradually decrease since more and more

information is accumulated over time which improves the precision of the filtered trend. In 2008 the standard errors start increasing, since the sample size is decreased in this period. In January 2010 the change-over form the old to the intermediate design starts with the implementation in the first wave. Since households are observed five times with quarterly intervals, the intermediate design is implemented in the second wave in April 2010, in the third wave in July 2010, in the fourth wave in October 2010 and the last wave in January 2011. During this period of five interventions, variables are introduced to model change-over, resulting in a strong increase of the standard error of the filtered trend. After finalizing the implementation, the standard error again gradually decreases as more information under the intermediate design becomes available. This improves the estimate of the intervention regression coefficients. As in 2012, when the implementation of the final design was introduced, five new interventions are added to model the shocks in the input series. This resulted in another increase of the standard error of the filtered trend.

7. Discussion

Major redesigns and survey process transitions bring risks for the continuity of time series produced from repeated surveys by national statistical institutes. It is therefore important for a national statistical institute to have a statistical framework in place to manage the risks to official statistics of the implementation of a redesign. In this paper a framework is proposed by pointing out the different methods available for quantifying the impact of a redesign on the estimates of a survey and the options to correct series for the observed differences in measurement bias.

The choice of method typically depends on the type of change in the survey process, the accepted level of risk, the required timeliness and accepted amount of revision of the impact estimates and the available budget for additional data collection. As pointed out in the paper, the different methods can be combined in a strategic transition design.

As far as the redesign concerns solely the data processing phase and the micro data under the old and new approach remain consistent, impact estimates can be obtained by reprocessing. If no additional variables are required in the new survey process, use of reprocessing is also possible to backcast series.

Parallel runs are typically appropriate if the data collection phase of a survey is changed. This approach has the advantage that it has a low risk of disturbing regular publications, and results in timely direct estimates for the impact on the publication if budget for a sufficiently large parallel run is available. This approach is typically useful for the most important image defining statistics of a national statistical institute. The opposite of a full parallel run is to directly change-over to the new design without having a period with overlap. In this case, state space intervention models

can be applied to separate the real evolution of the population parameter and the impact or discontinuities. This approach increases the risks in the production of official statistics. In a worst-case scenario a period without regular official statistics would be created if it is decided, after some period of time, to return to the old design. In addition, the estimates for the impact are revised as estimates of additional time points under the new design become available. As a result stable impact estimates are not timely. Finally the minimum observable discontinuities depend on the dynamics of the observed components and cannot be controlled by design, as in the case of planned randomized experiment. The strong advantage of this method is that the entire series is used to assess the impact, no additional costs for data collection are required and the complications of planning and conducting the field work of a parallel run are avoided, all of which makes the method extremely cost effective. The precision of the discontinuity estimates can be further improved by using SUTSE models where co-integrated series are available. This approach is typically useful for less important statistics which do not affect the image of a national statistical institute.

The major drawback of a parallel run, i.e. planning a large costly additional sample, can be compensated by conducting a parallel run at a small sample size and apply small area estimation methods to improve the precision of the impact estimates. Parallel runs can also be combined with the state space intervention methods by using the direct estimates for the discontinuities in the parallel run, including their variances, as a-priori information through an exact initialization of the Kalman filter. In this way, the information from a small parallel run is further complemented with the information available in the observed time series, in particular the additional information that comes available directly after the change-over. This illustrates that the methods are not mutually exclusive but can be combined and compensate each other's disadvantages to some extent.

Finally a publication strategy must be in place to communicate the impact of the redesign with the data users. Several methods are available to adjust the series for the observed differences. The choice of methods will also depend on assumptions of time invariance. As an alternative, it might also be decided to just publish the estimated discontinuities as they occur during the period of the change-over.

References

ABS (2009). Information Paper : ANZSIC 2006 Implementation in Retail Trade Statistics, July 2009, catalogue number 8501.0.55.006, http://www.abs.gov.au/ausstats/abs@.nsf/mf/8501.0.55.006

Aitchison, J. (1986). The statistical analysis of compositional data. London: Chapman and Hall.

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

Battese, G., R. Harter and W. Fuller (1988). An error-component model for prediction of county crop areas using survey and satellite data. Journal of the American Statistical Association, 83, pp. 28-36.

Bell, P. (2001). Comparison of Alternative Labour Force Survey Estimators. Survey Methodology, 27, pp. 53-63.

Binder, D.A., and J.P. Dick (1989). Modeling and estimation for repeated surveys. Survey Methodology, 15, pp. 29-45.

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

Bollineni-Balabay, O. Brakel, J.A. van den and Palm, F. (2016). Multivariate state-space approach to variance reduction in series with level and variance breaks due to sampling redesigns, Journal of the Royal Statistical Society, A series, vol 179, pp. 377-402.

Chipperfield, J. and P. Bell (2010). Embedded experiments in repeated and overlapping surveys. Journal of the Royal Statistical Society, A series, 173, pp. 51-66.

Cochran, W.G. (1977). Sampling theory, New York, Wiley and Sons.

Cochran, W.G. and Cox, G.M. (1957). Experimental Design, 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 U.S. Journal of the American Statistical Association, 94, pp. 1074-1082.

Durbin, J., and S.J. Koopman (2012). Time series analysis by state space methods. Oxford: Oxford University Press.

Durbin, J. and B. Quenneville (1997). Benchmarking by State Space models. International Statistical Review, 65, pp. 23-48.

Fay, R.E., and Herriot, R.A. (1979). Estimation of income for small places: an application of Jame-Stein procedures to census data. Journal of the American Statistical Association, 74, pp. 268-277.

Fellegi, I.P. (1964). Response variance and its estimation. Journal of the American Statistical Association, 59, 1016-1041.

Fienberg, S.E. and J.M. Tanur (1987). Experimental and Sampling Structures: Parallels diverging and meeting. International Statistical Review, 55, pp. 75-96.

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

Fienberg, S.E. and J.M. Tanur (1989). Combining cognitive and statistical approaches to survey design. Science, 243, pp. 1017-1022.

Fisher, R.A. (1935). The Design of Experiments. Oliver and Boyd. Edinburgh.

Griffiths, G., T. Surzhina, J. Blanchard, and P. Wise (2016). Exploring a frame work for unit level statistical impact measurement. Paper for the Australian Bureau of Statistics' Methodology Advisory Committee.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953). Sample Survey Methods and Theory, Vol I and II. New York : Wiley.

Hartley, H.O., and Rao, J.N.K. (1978). Estimation of nonsampling variance components in sample surveys. In Survey Sampling and Measurement, (Eds. N.K. Namboodiri). New York: Academic Press. 35-43.

Harvey, A.C. (1989). Forecasting, structural time series models and the Kalman filter. Cambridge: Cambridge University Press.

Harvey, A.C., and Chung, C.H. (2000). Estimating the underlying change in unemployment in the UK. Journal of the Royal Statistical Society, A series, vol.163, pp. 303-339.

Harvey, A.C., 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.

Hinkelmann, K. and O. Kempthorne (1994). Design and Analysis of experiments, Volume 1: introduction to experimental design. New York: Wiley & Sons.

Hinkelmann, K. and O. Kempthorne (2007). Design and Analysis of experiments, Volume 2: advanced experimental design. New York: Wiley & Sons.

James, G. (2008). Backcasting for use in short-term statistics. Interim report from the UK Office for National Statistics.

Kish, L. (1965). Survey Sampling, New York, Wiley and Sons.

Krieg, S. and J.A. van den Brakel (2012). Estimation of the monthly unemployment rate for six domains through structural time series modelling with cointegrated trends, Computational statistics and Data Analysis, 56, 2918-2933.

Mahalanobis, P.C. (1946). Recent experiments in statistical sampling in the Indian statistical institute. Journal of the Royal Statistical Society, 109, 325-370.

Montgomery, D.C. (2001). Design and Analysis of experiments. New York: Wiley & Sons.

Nolan, L., M.G. Sova, G. Brown, G. James and P. Lewis (2008). Backcasting for use in short-term statistics. Final report from the UK Office for National Statistics.

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

Pfeffermann, D. (2002). Small Area Estimation – New developments and directions. International Statistical Review, 70, pp. 125-143.

Pfeffermann, D. (2013). New important developments in small area estimation. Statistical Science, vol. 28, pp. 40-68.

Pfeffermann, D. and S.R. Bleuer (1993). Robust Joint Modelling of Labour Force Series of Small Areas. Survey Methodology, 19, pp. 149-163.

Pfeffermann, D. and L. Burck (1990). Robust Small Area Estimation Combining Time Series and Cross-Sectional Data. Survey Methodology, 16, pp. 217-237.

Pfeffermann, D. and Landsman, V. (2011). Are private schools better than public schools? Appraisal for Ireland by methods for observational studies. The Annals of Applied Statistics, 5, pp. 1726-1751, doi 10.1214/11-AOAS456.

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

Rao, J.N.K. and I. Molina (2016). Small Area Estimation. New York: Wiley en Sons.

Rao, J.N.K., and Yu, M. (1994). Small area estimation by combining time-series and cross-sectional data. The Canadian Journal of Statistics, 22, pp. 511-528.

Robinson, G.K. (2000). Practical strategies for experimenting. New York: Wiley & Sons.
Rosenbaum, P. and D.B. Rubin (1983). The central role of propensity score in observational studies for causal effects. Biometrika, 70, pp. 41-55.

Scheffé, H. (1959). The Analysis of Variance. New York: Wiley & Sons.

Scott, A.J., and T.M.F. Smith (1974). Analysis of repeated surveys using time series methods. Journal of the American Statistical Association, 69, pp. 674-678.

Scott, A.J., T.M.F. Smith, and R.G. Jones (1977). The application of time series methods to the analysis of repeated surveys. International Statistical Review, 45, pp. 13-28.

Searle, S.R. (1971), Linear Models. New York: Wiley & Sons.

Smith, P.A. and G. James (2017). Changing industrial classification to SIC (2007) at the UK Office for National Statistics. Journal of Official Statistics, 33, pp. 1-25.

Tam, S.M. (1987). Analysis of repeated surveys using a dynamic linear model. International Statistical Review, 55, pp. 63-73.

Tiller, R.B. (1992). Time series modelling of sample survey data from the U.S. current population survey. Journal of Official Statistics, 8, pp. 149-166.

Van den Brakel, J.A. (2008). Design-based analysis of experiments with applications in the Dutch Labour Force Survey. Journal of the Royal Statistical Society, Series A, 171, pp. 581-613.

Van den Brakel, J.A. (2010). Sampling and estimation techniques for the implementation of new classification systems: the change-over from NACE Rev. 1.1 to NACE Rev. 2 in business surveys. Survey Research Methods, 4, pp. 103-119.

Van den Brakel, J.A. (2013). Design-based analysis of factorial designs embedded in probability samples. Survey Methodology, vol. 39, pp. 323-349.

Van den Brakel, J.A. (2016). Design-based analysis of experiments embedded in probability samples Wiley chapter.

Van den Brakel, J.A. and S. Krieg, (2009). Estimation of the monthly unemployment rate through structural time series modelling in a rotating panel design. Survey Methodology, vol. 35, pp. 177-190.

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

Van den Brakel, J.A. and S. Krieg (2016). Small area estimation with state-space common factor models for rotating panels. Journal of the Royal Statistical Society A series. Vol. 179, pp. 763-791.

Van den Brakel, J.A. and R. Renssen (1998). Design and Analysis of Experiments Embedded in Sample Surveys. Journal of Official Statistics, 14, pp. 277-295.

Van den Brakel, J.A. and R. Renssen (2005). Analysis of Experiments Embedded in Complex Sample Designs. Survey Methodology, 31, pp. 23-40.
Van den Brakel and J. Roels (2010). Intervention analysis with state-space models to estimate discontinuities due to a survey redesign. Annals of Applied Statistics, vol. 4, pp. 1105-1138.

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

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

Van den Brakel, J.A. van den and C.A.M. van Berkel, (2002). A design-based analysis procedure for two-treatment experiments embedded in sample surveys, Journal of Official Statistics, vol. 18, no. 2, pp. 217-231.

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

Zhang, M (2002). Backcasting and seasonal adjustment models, ABS internal document.

Zhang, M (2008). Backcasting facility phase 2 development, ABS internal document.

Zhang, M. and O. Honchar (2016). Predicting survey estimates by states space models using multiple data sources. Paper for the Australian Bureau of Statistics' Methodology Advisory Committee.

Explanation of symbols

Empty cell	Figure not applicable
.	Figure is unknown, insufficiently reliable or confidential
*	Provisional figure
**	Revised provisional figure
2015–2016	2015 to 2016 inclusive
2015/2016	Average for 2015 to 2016 inclusive
2015/'16	Crop year, financial year, school year, etc., beginning in 2015 and ending in 2016
2013/'14–2015/'16	Crop year, financial year, etc., 2013/'14 to 2015/'16 inclusive

Due to rounding, some totals may not correspond to the sum of the separate figures.

Colofon

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

Prepress
Statistics Netherlands, Studio BCO

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contactform: www.cbsl.nl/information