Australian Bureau of Statistics

1351.0.55.158

Research paper

A framework for measuring the impact of transitions in official statistics

Australia

2017

www.abs.gov.au

Research paper

A framework for measuring the impact of transitions in official statistics

Jan van den Brakel[1], Greg Griffiths[2], Tatiana Surzhina[2], Phillip Wise[2], Jonathan Blanchard[2], Xichuan (Mark) Zhang[2], Oksana Honchar[2]

1. Department of Statistical Methods, Statistics Netherlands and Department of Quantitative Economics, Maastricht University School of Business and Economics. 2. Methodology Division, Australian Bureau of Statistics.

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30AM (CANBERRA TIME) FRI 30 JUNE 2017

ABS Catalogue No. 1351.0.55.158

Produced by the Australian Bureau of Statistics.

INQUIRIES

For further information about these and related statistics, contact the National Information and Referral Service on 1300 135 070.

CONTENTS

A FRAMEWORK FOR MEASURING THE IMPACT OF TRANSITIONS IN OFFICIAL STATISTICS

Jan van den Brakel[1], Greg Griffiths[2], Tatiana Surzhina[2], Phillip Wise[2], Jonathan Blanchard[2], Xichuan (Mark) Zhang[2], Oksana Honchar[2]

1. Department of Statistical Methods, Statistics Netherlands and Department of Quantitative Economics, Maastricht University School of Business and Economics. 2. Methodology Division, Australian Bureau of Statistics.

ABSTRACT

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 of such changes on the estimates for repeated surveys. 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.

1. INTRODUCTION

Official statistics produced by national statistical institutes (NSIs) are frequently based on censuses, 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 parameters of interest. In order to preserve the comparability of estimates, the underlying process of the survey is kept unchanged as long as possible. Note that we use the term survey in this paper to refer to the business process in place for the production of official statistics, which can be either based on a probability sample or a register. It is inevitable, however, that adjustment or redesign of this business process is needed from time to time, since the existing procedures become gradually out-dated or more cost effective methods are required.

Major redesign of the underlying business process, however, generally has systematic effects on the survey estimates, disturbing comparability with figures published in the past. To minimise 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 alterations in processing.

Quantifying discontinuities is only one aspect of a smooth change-over to a new survey process. Implementing a new survey process without disrupting the regular production of official statistics requires careful preparation and planning. It is essential that the new survey process is well defined before impact measurement starts or implementation for production can be considered. Pretesting different design options for a new survey process, or parts of it, by means of field experiments is an essential aspect of developing and establishing a new survey design.

Quantifying discontinuities is also only one aspect of quality to be considered in a survey transition. In addition, other standard quality measures, for example response indicators, variances of sample estimates and selectivity indicators, are important measures for obtaining insight into the quality of the new survey design. If there is a period of parallel data collection for impact measurement, then these indicators offer additional insight into the differences between the old and new survey process.

During the redesign of a survey process there are different roles for NSI management, statistical methodologists and other stakeholders (e.g. main data users). NSI management is responsible for making a decision on the nature and resourcing of impact measurement and the change implementation process. Stakeholders need to be informed about the redesign and risk of discontinuities in advance to minimise the impact on data users. Ideally key stakeholders are involved in the decisions concerning the new design and how the impact measurement should be executed. Finally, statistical methodologists advise on the design of the new survey process and design an impact measurement approach.

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 following options for measuring impact will be discussed:

- reprocessing data
- parallel data collection based on an embedded experiment
- time series methods
- small area estimation
- modelling measurement error at the unit level

The different methods presented in this paper illustrate that the choice of the most appropriate method depends on the:

- type of change in the survey process
- available budget for impact measurement
- importance of the indicators/statistics
- accepted level of risk for failing to detect discontinuities or for detecting them at an inadequate level of accuracy
- accepted level of risk for disturbing the regular survey process and official publications of the survey
- required timeliness of the impact estimates
- accepted amount of revisions

A bottom line approach is to accept that the period-to-period change after implementation of a redesign is differently biased, with a discontinuity at the time of change, and to make no further attempt to disentangle real change from this alteration in measurement bias. Throughout the paper, examples from major survey redesigns that have taken place in the past at Statistics Netherlands and the Australian Bureau of Statistics are used for illustrative purposes.

It is important to realise that the five aforementioned methods for impact measurement are not mutually exclusive. As will be seen in the presentation of these methods and the examples below, they can be combined and complement each other to some extent.

2. REPROCESSING DATA

If the micro data collected under the old and new approaches are consistent, then there is no need to collect additional data through an experimental setup. 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. In practice complications often arise since additional variables are often required for processing data under the new approach.

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. Although the micro data to compile short-term statistics and business statistics are completely compatible under both classification systems, a smooth change-over form the NACERev1.1 to the NACERev2.0 was still a major operation. 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 which have insufficient sample size.

3. PARALLEL DATA COLLECTION

Parallel data collection implies that the old and new approaches are conducted at the same time alongside each other for some period of time. Ideally this is based on a randomized experiment or a split-ballot design, but that is only possible if the sample design itself is not modified. A parallel data collection enables the construction of design-based estimates under both approaches for the same period. Differences between the estimates can be interpreted as an estimate of the systematic difference between approaches.

Parallel data collection can be considered when it is expected that the micro data will not be consistent under the old and new approaches. 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.

3.1 ADVANTAGES AND DISADVANTAGES

Conducting the old and new approaches in parallel over some duration has multiple advantages. If sufficient sample size can be allocated to the new approach, direct estimators can be constructed for the same parameter of interest under both approaches. The contrast between estimates under both approaches is then a direct estimate for the discontinuity. No model assumptions are required, unlike situations where, for example, time series methods are required. From one point of view, such an experimental setup 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.

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.

Ideally, parallel data collection is based on a randomized experiment. In the case of a model-based inference approach, randomisation protects against model misspecification. An analysis based on a randomized experiment will be more robust to model misspecification compared to data where there is no randomised experiment. In the latter case, stronger auxiliary information is required to control for selection bias or to separate real change over time from differences in measurement bias. For sufficiently large field experiments embedded in a probability sample, Van den Brakel and Renssen (2005) and Van den Brakel (2008, 2013) proposed a design-based inference

procedure. Their approach gives (approximately) design-unbiased estimates for treatment effects or discontinuities.

There are of course also disadvantages related to parallel data collection. The approach 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 fieldwork. More details are given in Subsection 3.3.

3.2 EXAMPLES OF SURVEY REDESIGNS WITH PARALLEL DATA COLLECTION

3.2.1 Examples of parallel runs at Statistics Netherlands

At Statistics Netherlands, parallel runs have been conducted frequently. An early example is the redesign of the 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 in this survey a complete redesign of the fieldwork and data collection method was introduced. The old and new designs were 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.

In 2010 and 2012 two major redesigns of the Labour Force Survey (LFS) took place. The Dutch Labour Force Survey is based on a rotating panel design. Each month a stratified two-stage sample of households is drawn. Sampled households are observed five times at quarterly intervals. These five subsequent observations are further referred to as waves. In the first redesign the data collection mode in the first wave changed from uni-mode Computer Assisted Personal Interviewing (CAPI) to a mixed mode CAPI and Computer Assisted Telephone Interviewing (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 2012 the data collection in the first wave changed to a sequential mixed mode design starting with web interviewing. In the follow-up phase, non-respondents were approached by CATI or CAPI. Also the questionnaire was adjusted again. On both occasions the first wave of the LFS was conducted in parallel at full sample size for a period of six months, while discontinuities in the remaining waves were estimated through the time series approach proposed in Section 4. See Van den Brakel and Krieg (2015) for details.

Finally, several redesigns of the Dutch Crime Victimization Survey took place in 2005 and 2008. During this period, the data collection mode changed from uni-mode CAPI 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 approach was conducted in parallel with the regular approach with a sample size of about 6500 and 19000 respondents respectively. 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 is sufficient to quantify discontinuities. Small area estimation was applied as an alternative, as explained in Section 5. Details can be found in Van den Brakel et al. (2008) and Van den Brakel et al. (2016).

3.2.2 Examples of parallel runs at the Australian Bureau of Statistics

An embedded experiment was used by the Australian Bureau of Statistics to establish the effect arising from the introduction of a new LFS questionnaire in 2001. The changes included a number of revised definitions to enhance the quality of the statistics and better align the survey with international standards, and methodological changes to reduce provider load such as introducing a new age category for persons aged 65 years and over. The statistical impact of the new questionnaire was investigated during a test from March to August 2000. The test was implemented in the form of a controlled experiment in which one-eighth of the sample (one rotation group) was assigned the new questionnaire, forming the treatment group, with the remaining seven rotation groups enumerated using the old questionnaire, forming the control group. The chosen rotation group differed each month and stayed in the LFS sample for a different number of months. Survey participants in the chosen rotation group were asked to respond via the new questionnaire for one month only. During their remaining time in the sample they responded using the old questionnaire. This was to ensure that the treatment groups included units which varied in terms of their length of stay in the sample, which was important in terms of separating the month in survey bias from other measurement biases.

Effects due to the definition changes were quantified by calculating percentage differences between estimates for key labour force estimates based on the old questionnaire and the new questionnaire. The estimates were calculated for each month of the experiment by 'reconstructing' responses from the old questionnaire using data collected via the new questionnaire. Effects due to different wording or order of the questions were analysed using the Best Linear Unbiased Estimator (BLUE) composite estimator (Bell, 2001). These tests found there was no evidence of a significant effect on the labour force estimates. However, the experiment did not have sufficient power to measure changes smaller than two standard errors of the monthly movement. The design allowed the ABS to monitor such impacts throughout the experiment which increased the potential for detecting undesirable effects and mitigating their impact.

In 2004, an embedded experiment was conducted to quantify the effect of Computer Assisted Interviewing (CAI) in the Australian LFS. In this case 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 CAI on the labour force estimates, maximise the sample size for statistical impact measurement and increase the time to respond to adverse effects of CAI. The effect of CAI on the labour force estimates was estimated using a model-assisted estimator in conjunction with the BLUE composite estimator. Three types of effects were quantified using separate estimators: an overall effect, an effect constant over time but different by subgroup, and an effect varying by subgroup and over time. This phase-in approach had the advantage of increasing power as the sample size grew over time, and could be implemented relatively quickly.

In 2008-2009 the Australian Bureau of Statistics' Retail Trade, Quarterly Business Indicators Survey (QBIS) and Capital Expenditure Survey (CapEx) underwent a substantial sample redesign for the purposes of implementing the new industry classification (ANZSIC06), expanding the scope to include small employing and non-employing business units, and using revised strata cut-offs based on Business Activity Statement (BAS) data. To manage statistical risk due these major changes, two samples were selected in parallel with an attempt to maximize the overlap between them. The old sample was based on the industry classification ANZSIC93 and included only employers (i.e. did not include non-employers). The old and new samples were implemented in parallel for two quarters, March and June 2009. From September 2009 the new sample design was exclusively used to produce the sample. At the same time, both QBIS and CapEx surveys moved to a more efficient estimation and imputation system, ABS-SF.

The new sample design, in particular the industry classification change, affected the comparability of published time series of business indicators. The impacts were managed by providing to users revised historical series which aligned past estimates with estimates based on the new sample design. The statistical impact was the difference between the parallel survey estimates and the estimates based on the old design for the same quarters, March and June 2009. The estimated impacts (i.e. the difference between level estimates) were backcast in the historical series which were produced on the ANZSIC06 basis using post-stratification. Backcasting allowed the Australian Bureau of Statistics to maximise the comparability of time series based on the new classification at the Australian broad industry and state levels.

As a part of the risk mitigation strategy, estimates produced by the new estimation and imputation system were compared with estimates produced by the old estimation method in order to investigate possible discrepancies between the two estimation systems. Where there were major discrepancies, causes were investigated and the code within ABS-SF was corrected to prevent discrepancies in future. For example, during the investigation it was found that the new system used values of completely enumerated units to impute values for sampled units. As a result, the estimation code was modified to distinguish completely enumerated units from sampled units during the imputation process.

3.3 KEY CONSIDERATIONS FOR THE DESIGN AND ANALYSIS OF EXPERIMENTS

3.3.1 Purpose of the experiment

Parallel data collection through an experimental setup requires careful planning and preparation. For brevity we will further refer to this approach as a 'parallel run'. The design of the new survey must be final before launching a large scale parallel run. The purpose of the parallel run is to establish the impact of the new approach. Although it sounds trivial, 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. In the event that the new survey process is found to be defective, the parallel run should cease and be reinitiated at a later date. 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, which maximises the power for estimating the impact as will be explained below. In general, it should be anticipated that the first period of the parallel run might not be usable since, for example, the field staff may have to adapt to the new approach.

Planning and designing an experiment requires key decisions on several themes. First of all, the purpose of the experiment should be stated as clearly and explicitly as possible. 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. These are competing purposes. 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. This design results the highest power for estimating a discontinuity, based on a fixed budget for the parallel run. 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.

Estimating effects of separate factors or just the overall effect is a crucial decision that has to be made at the start of a redesign. During the redesign of the Dutch LFS, it was decided in advance that the purpose of the parallel run was only to estimate the net effect of the redesign. After conducting the parallel run, an increase of 20% of the unemployed labour force was observed, resulting in questions about the extent to which underlying factors contributed to this increase - insights which might be useful to further improve the new survey process. The raising of these type of questions in an unplanned way, after finalizing a parallel run, should be avoided.

An alternative is to consider a factorial design with an unbalanced setup. This 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, very inefficient in terms of power to estimate contrasts. A better alternative is to conduct pilots to understand the effect of the different factors and make a final decision about the design of the new approach before starting a parallel run, as explained above.

An important aspect of designing a randomized experiment is to clearly specify in advance the hypotheses about the main effects and possible interactions to be tested, to avoid post-hoc analyses as much as possible. In the case of continuing surveys it might also be desirable to quantify the impact on the seasonal effects. This will dramatically increase the length and sample size of the parallel run and will almost always be infeasible in practical terms.

3.3.2 Experimental design considerations

Based on predefined decisions, power and minimum sample size calculations can be made. An advance decision is needed on the smallest size of impact that should result in a rejection of the null hypotheses of zero impact at a pre-specified significance and power level. If the budget and thus the sample size are fixed in advance, the minimum observable differences at pre-specified significance and power levels can be calculated. This provides an indication of what can be expected from the experiment.

In the design stage a choice of an appropriate experimental design that maximizes the power of the analysis of treatment effects must be made. The most straightforward way to design a parallel run is to draw two or more subsamples independently from each other which are assigned to one of the treatments of the experiment. The precision of an experiment, however, can be improved by embedding an experiment in a probability sample and using the framework of the sample design to identify potential control variables for the experimental design. Instead of directly randomizing sampling units over treatments according to a Completely Randomized Design (more or less similar to drawing independent samples), Randomized Block Designs (RBD) can be used. In an RBD the sampling units are first divided in more or less homogeneous blocks. The sampling units within each block are randomized over the treatments. This eliminates the variation between the blocks from the variance of the treatment effects. Potential block variables are sampling factors 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, clusters of respondents assigned to interviewers 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.

3.3.3 Fieldwork considerations

To design an experiment that accurately measures the difference between the old and new set up, it is also important to carefully plan how the experiment is implemented in the fieldwork and in particular how the field staff will conduct the data collection under the different treatments. 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 the experiment. If, on the other hand, interviewers can be assigned to one treatment only, it is also worthwhile to consider a double blind set-up. This implies that the interviewers are unaware of the fact that they are participating in an experiment since this might influence their normal behaviour, even unconsciously. Field staff restrictions might require that interviewers, including the clusters of respondents assigned to them, are randomised over the treatments, at the cost of reduced power of the experiment. If interviewers are assigned to one of the treatments only, then this allocation must be

done randomly. For example, assigning newly recruited field staff to one of the treatments only in order to save training costs would invalidate the outcomes of the parallel run.

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 in 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 where interviewers conducted more than one treatment. This gradually decreased as the field staff became more and more familiar with conducting fieldwork for embedded experiments.

3.3.4 Choice of an appropriate inference mode

Finally, the mode of inference used to analyse the experiment should be considered. 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 survey 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. 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, for example, 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. Van den Brakel (2008) showed how to test treatment effects for such ratios. In a standard model-based analysis it is not 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.

3.4 POWER CONSIDERATIONS

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. 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 unacceptably 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 randomisation of the ultimate sample units over the treatments.

The precision of an experiment can be increased by using the concept of RBD's to preclude, as much as possible, variation in estimated treatment effects as can be controlled for by the design of the experiment. As mentioned in Subsection 3.3, the sample design offers a framework for such control variables. For example sample strata, interviewers or clusters are potential block variables in randomized block design where possible.

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 were 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.

Another option is to reduce the sample size of the regular survey to allow an increase in 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 Van den Brakel et al. (2008) for details. 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 extremely 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. See Van den Brakel (2008) for details.

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 tests 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. Reliance on application of post-hoc analysis must be restricted as much as possible.

After making an all-out effort to optimize the design, one can undertake the experiment with the available budget and make reasonable assumptions for the interpretation of the results. With the transitions of the Dutch Crime Victimization Survey for example, there was budget for conducting the new approach with a sample size of 6000 respondents in parallel to the regular approach that has a sample size of about 19,000 respondents. In this survey the main publication domains are 25 police regions and official statistics for these domains are based on about 750 respondents. Under the assumption that discontinuities observed at the national level also hold for the underlying domains, this parallel run was sufficient to quantify impact of the redesign for these domains. See Van den Brakel

et al. (2008) for details. In order to relax this assumption, small area estimation procedures could have been applied as an alternative. See Van den Brakel et al. (2016) for details.

If the available budget for a parallel data collection is insufficient to meet the expected precision and power requirements, it can still be useful to undertake the parallel run at a smaller sample size. First of all, this reduces the risk of having a period without data for official publication purposes. Moreover the estimates for the discontinuities obtained with the parallel run can be further improved with a time series modelling approach as will be explained in Subsection 4.1.

A useful general framework and practical guidelines for planning and conducting experiments is given by Robinson (2000). An introduction for design and analysis of experiments is provided by Montgomery (2001). Advanced text books on design and analysis of experiments are Hinkelmann and Kempthorne (1994, 2007).

4. TIME SERIES ANALYSIS

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 survey estimates.

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. Generally these are a trend to model low frequency variation, a seasonal component to model cyclic deviations from the trend within a period of one year, a cycle to model economic or business cycles with a period typically longer than one year 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, $\bar{Y}_{k,t}$ denotes the value obtained if Θ_t is observed under a complete enumeration using treatment k for time period t, $\hat{\bar{Y}}_{k,t}$ is 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 $e_{k,t}$ denotes the sampling error. Note that it is assumed here that measurement bias is time independent. The population parameter is modelled with an appropriate set of components, i.e. a trend, seasonal, cycle, regression component and a white noise. After inserting the time series model for the population parameter into the measurement model, a time series model for the observed series of sample estimates is obtained. Assuming for the

population parameter a local linear trend or a smooth trend (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, the following model is obtained for the observed series $\hat{\bar{Y}}_{k,t} = L_t + S_t + I_t + \gamma_k + e_{k,t}$. In the case of (rotating) panel design, the white noise of the population parameter and the survey errors are identifiable, see Pfeffermann (1991) or Van den Brakel and Krieg (2009). In the case of cross-sectional surveys, both components are typically confounded and estimated as one term, say $v_{t,k} = I_t + e_{k,t}$. 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 Van den Brakel and Krieg (2009).

In the case where there is no period in which the old and new design of the survey are conducted in parallel, an appropriate time series model for the observed series can be constructed to model the real evolution of the population parameter of interest. Subsequently an intervention component is added to the model which models an intervention from 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 only using the survey data. 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: $\hat{\bar{Y}}_{k,t} = \tilde{L}_t + S_t + \beta\delta_t + v_{k,t}$.

This approach is proposed by Van den Brakel and Roels (2010) and is a direct application of the state space intervention model proposed by Harvey and Durbin (1986) for analysing the effects of seatbelt legislation on road casualties in the UK. 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 impact due to the change-over.

At the end of Subsection 3.3.1 mention was made that it might be desirable to quantify the impact on the seasonal effects. The only way to do this in a cost effective way is to model this with an intervention on the seasonal component, but this requires several years of observations before it can be estimated with sufficient reliability.

4.1 ADVANTAGES AND DISADVANTAGES

A time series modelling approach is appropriate if the micro data observed under the old and new approach are not consistent and if there is no budget for a parallel run. 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. Another advantage

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 analyse the statistical impact. The state space method has, in addition, the flexibility to combine information in the entire series with the information obtained with parallel data collection. 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 to construct an exact initialization of the Kalman filter, as pointed out by 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 $t = \tau, \tau+1, \ldots, \tau'$ 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 $t = 1, \ldots, \tau'$ and is missing for $t = \tau', \ldots, T$. Similarly $\hat{\tilde{Y}}_{new,t}$ is observed from $t = \tau, \ldots, T$ and missing for $t = 1, \ldots, \tau$. 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. This method was implemented in 2010 to produce official monthly figures about the labour force.

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 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 volatility 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.

In addition, it should be mentioned that with the time series modelling approach there is no control over the precision and size of the minimum observable impact. The minimum detectable difference depends 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. See subsection 4.2.2 for the details of simulations to obtain insight into the attainable precision with the time series approach in combination with small parallel runs.

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.

Note that all these risks 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.

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.

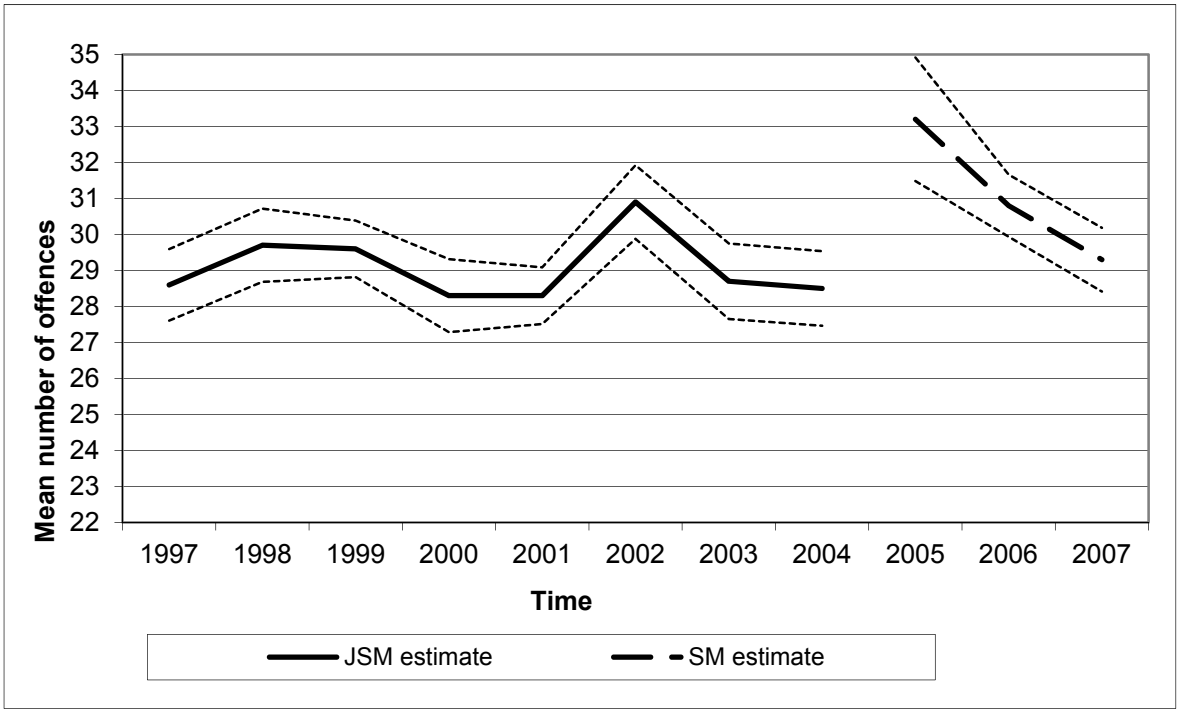
4.2 EXAMPLES

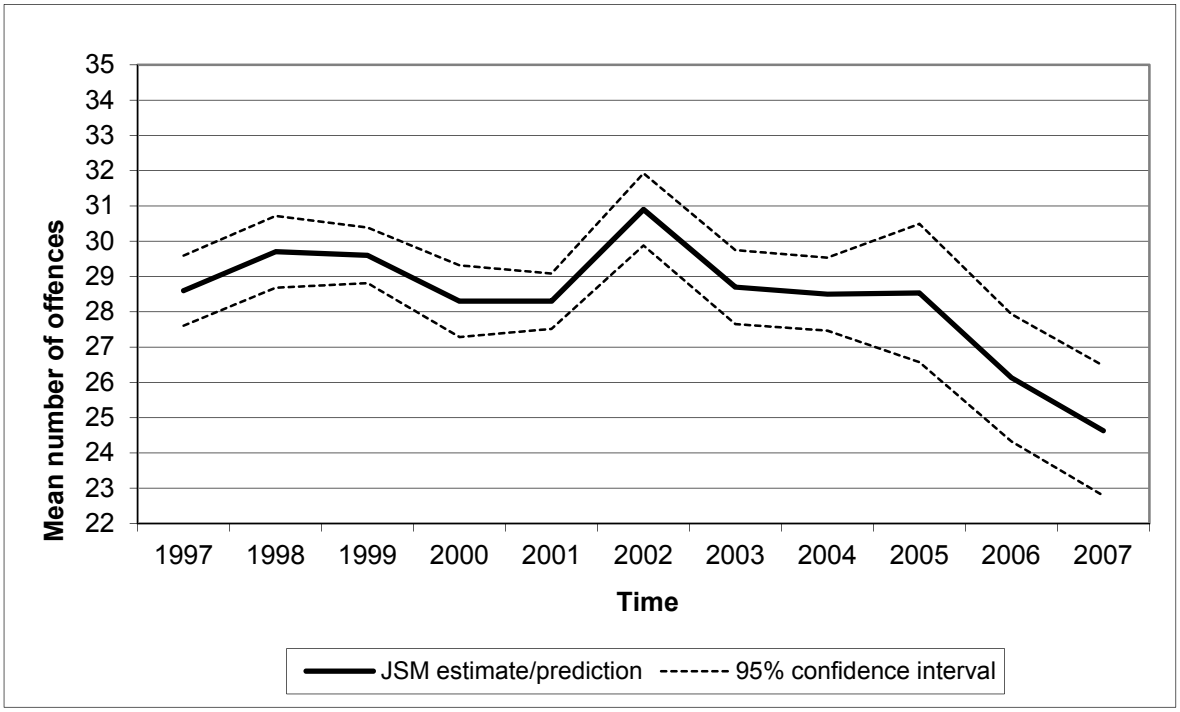
4.2.1 Applications of the time series approach at Statistics Netherlands

The time series approach was applied at Statistics Netherlands for the first time with the change-over form the Justice and Security Module of the Permanent Survey on Living Conditions to the Dutch Crime Victimisation Survey in 2005. At this time there was no budget for a parallel run, even at a reduced sample size. From 1997 through 2004 the Justice and Security module was used to estimate figures about the rates of different crimes committed against the Dutch inhabitants. The change-over to the Crime Victimization survey in 2005 resulted in a strong and unexpected increase in the estimates of total number of offences as shown in the top panel of Figure 1. To disentangle real evolution in the crime rates from the impact of the survey redesign, 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 the bottom panel of Figure 1. The choice to adjust observations after the change-over to the level before the change-over is arbitrary in this example. Estimates under the old approach before the change-over can be adjusted to the level of the new approach in a similar way. 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 applied frequently since then.

FIGURE 1: Top panel: Mean number of offences against Dutch inhabitants (hundreds) during the 12 months prior to the interview, observed with the JSM (1997-2004) and the SM (2005-2007). Bottom panel: Mean number of offences against Dutch inhabitants (hundreds) during the 12 months prior to the interview, observed with the JSM (1997-2004) and the SM corrected for the estimated discontinuity (2005-2007).





Another interesting example is 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 were estimated with a state space intervention model. In this application there were 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 was 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. An example of this is the redesign of the Dutch Labour Force Survey in 2010 and 2012. The Dutch LFS changed from a cross-sectional survey to a rotating panel in 2000. Each month a stratified two-stage sample of households entered the panel. These samples were observed five times with quarterly intervals. A consequence of this change-over was that the effects of rotation group bias (Bailar, 1975) became very visible in the figures of the LFS. To handle problems with sample sizes and rotation group bias a multivariate structural time series model for the production of official monthly figures on the Labour Force was implemented in 2010. The inputs for this model were five series of GREG estimates on a monthly frequency based on the five waves of the panel, following the approach proposed by Pfeffermann (1991). In this estimation procedure the time series model estimates for the population parameters were 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 assumed that the observations obtained in the first wave were the most reliable and required 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 therefore decided to allocate the available budget for a parallel run to the first wave only. 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 were modelled with the state space intervention approach. Unreliable estimates in these waves directly after the change-over did not impact the population parameter estimates because the population parameter estimates were 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).

4.2.2 Scenarios involving a major redesign of the Australian LFS

For 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. With the state space model, observations under the new design that become available after the change-over are used to further improve the precision of the discontinuities. 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 where eight monthly samples (rotations groups) are each observed over eight months each commencing the month after the rotation group before. For this simulation an eight dimensional state space model is developed for the series of GREG estimates from these eight waves, see Pfeffermann (1991) for details. This model is used to generate 100 replicates of the series of unemployed labour force, including discontinuities 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 for three different scenarios of parallel runs of reduced sample sizes, as summarized in Table 1. The standard errors in Table 1 refer to the statistical impact estimates obtained with the control sample, the treatment sample and the specified parallel run periods. Percentages refer to the sample size of the regular LFS.

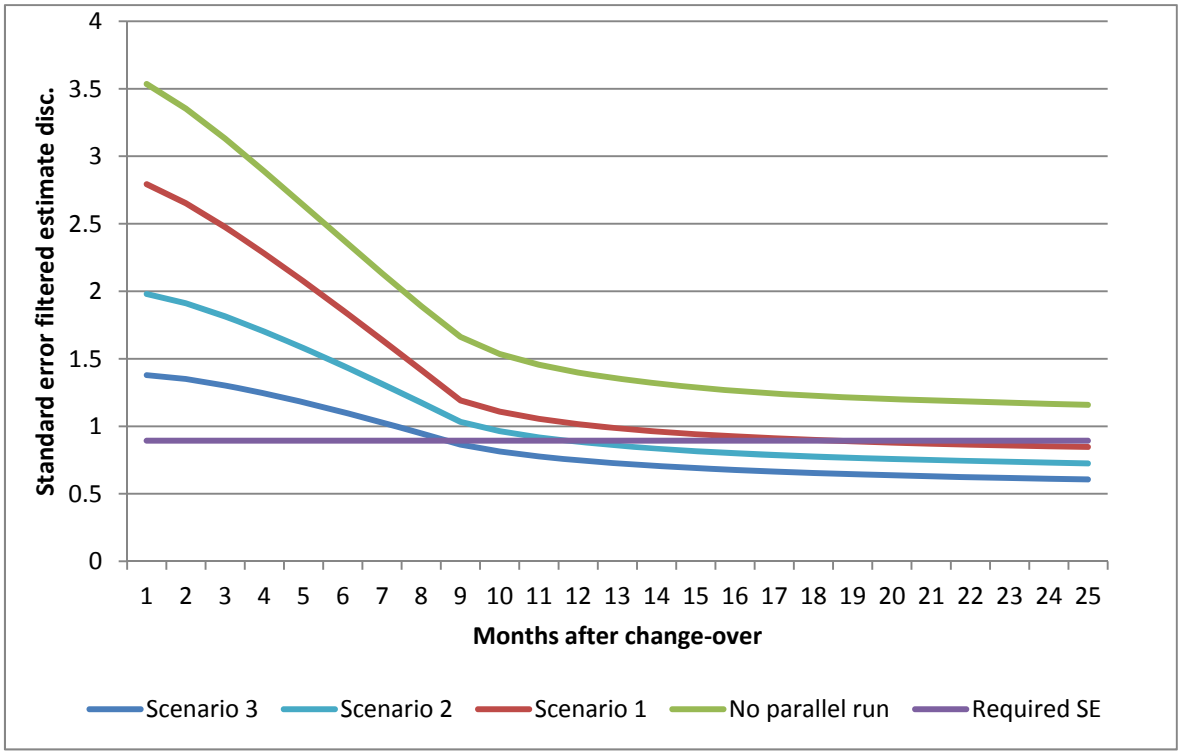
Figure 2 illustrates for the three scenarios the standard errors of the filtered estimates for different periods after the changes-overs. Standard errors for the filtered estimates for the scenario without a parallel run are also included. The horizontal line refers to the minimum required standard error of 0.9%.

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

Scenario	Standard error separate waves	Standard error aggregated	Sample size control sample	Sample size treatment sample	Parallel run period
1	7.9%	2.79%	100%	20%	18 months
2	5.6%	1.98%	100%	50%	12 months
3	3.9%	1.38%	100%	50%	18 months

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 Scenarios 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.

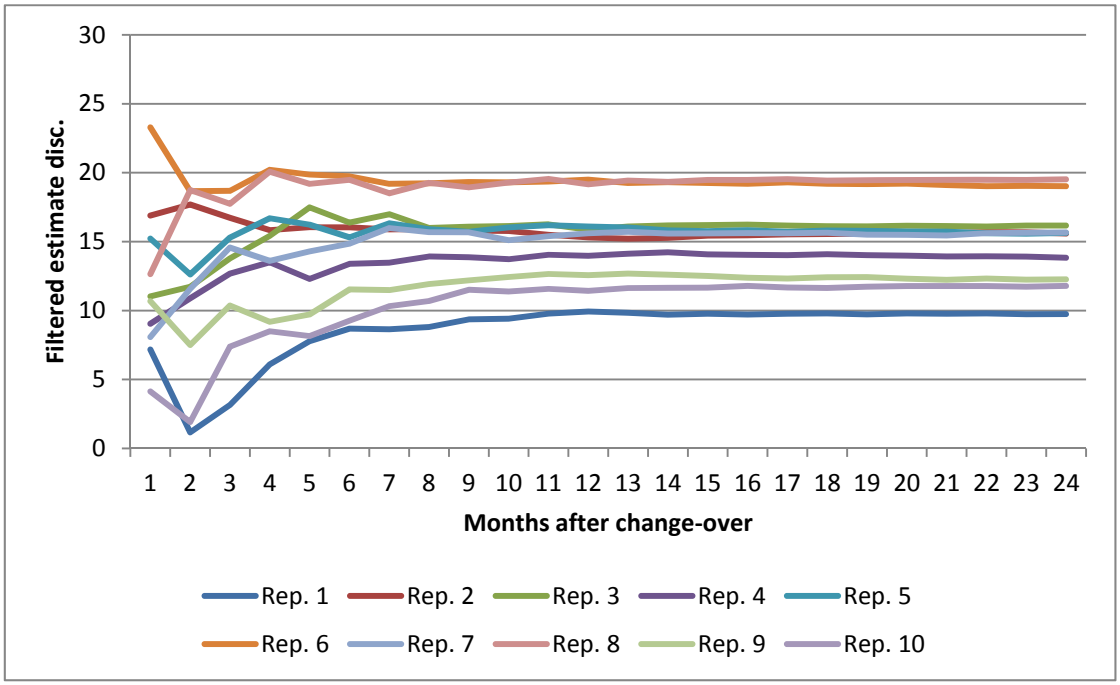
FIGURE 2: Standard errors filtered estimate discontinuity obtained with the time series model for different periods after the change-over for an exact initialisation of the Kalman filter with the scenarios from Table 1 and a diffuse initialisation (no parallel run). The horizontal line refers to the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

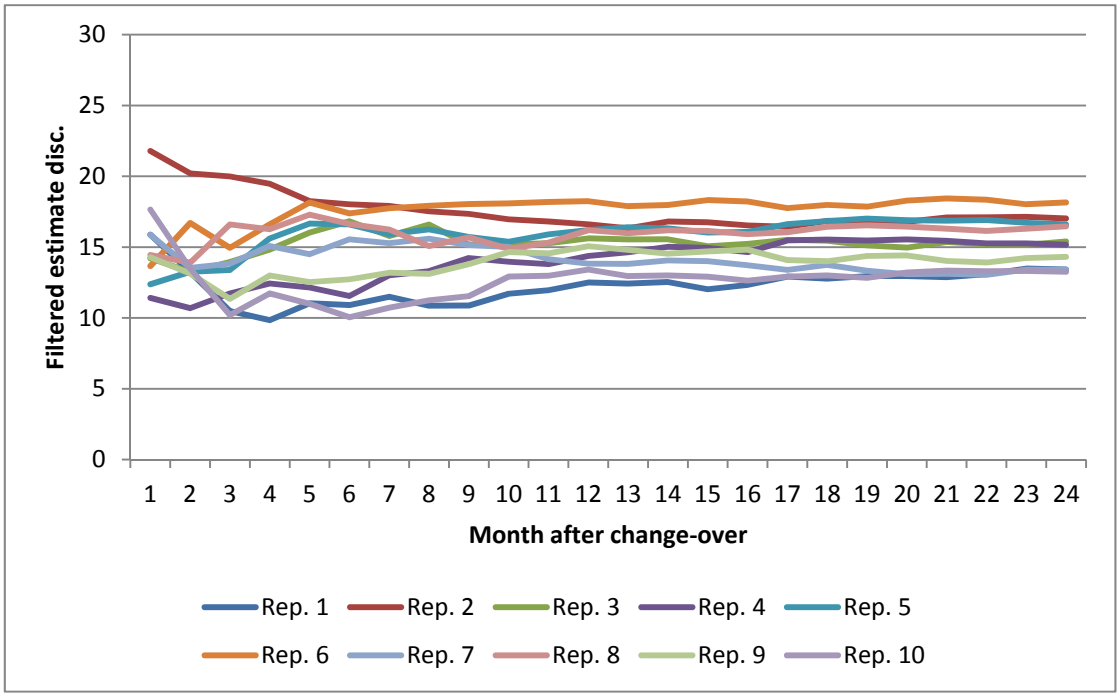


To illustrate the volatility of the impact estimates if there is no parallel run, the top panel of Figure 3 shows for each of 10 replicates how the time series model estimates the impact if more observations become available after the change-over in the first rotation group with an impact of 15%. The horizontal axis depicts the number of months observed under the new design after the change-over. As can been seen, it takes about 12 months before a stable estimate for the impact in a particular wave is obtained. The bottom 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 one wave obtained with the time series model for different periods after the change-over without a parallel run (top panel) and with a parallel run according to Scenario 3 (bottom panel). Real 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.

A parallel run of 12 months in combination with an intervention model for the seasonal component can be used to assess the statistical impact on the seasonal component in a similar way.

4.3 EXTENSIONS TO INCORPORATE AUXILIARY INFORMATION

The use of a state space intervention model for separating real change over time from processing 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 and common factor models can be applied to incorporate this auxiliary information to better separate real evolution in the population parameter from the discontinuities induced by the redesign. 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. This concept can be used to formulate more parsimonious models. See Koopman et al. (2009) for more details.

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 for 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.

5. SMALL AREA ESTIMATION

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.

As explained in Subsection 4.2, the change-over to the Crime Victimization survey in 2005 resulted in a strong unexpected increase in estimates of the total number of offences. The state space intervention approach successfully separates real developments from the measurement discontinuities induced by the redesign. This survey was redesigned again in 2008. To avoid the risks of a time series approach, summarized in Subsection 4.1, it was decided to explain the impact of redesigns in the figures of important surveys like the Crime Victimization Survey using a parallel run at a reduced scale.

In the previous example 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. 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 models to incorporate sample information from preceding periods, Rao and Yu (1994), Datta et al. (1999). Multivariate structural time series modelling is another approach to combine cross-sectional and temporal information, Pfeffermann and Burck (1990), Pfeffermann and Bleuer (1993). See e.g. Rao and Molina (2016) or Pfeffermann (2002, 2013) for an overview of small area estimation.

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, that borrows strength over space, 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 a 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 $\hat{\bar{Y}}_{d,reg}$, 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 are 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 a parallel run where the regular Crime Victimization Survey was conducted at a sample size of about 19,000 respondents and the new approach at a limited scale of 6,000 respondents. The most important output for this survey are crime figures for 25 police regions. These domains form the stratification variable in the sample design and about 750 respondents are observed in each domain. With a parallel run of about 6000 observations, direct estimates for the discontinuities were obtained at the national level. Under the assumption that these discontinuities hold for the underlying domains, this parallel run has sufficient power to detect discontinuities at the level of 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.

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:

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

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:

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

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 $\hat{\bar{Y}}_{d,new}$ and $\hat{\bar{Y}}_{d,reg}$. 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 within a model-based framework, and is currently being explored at Statistics Netherlands.

6. MODEL-BASED INFERENCE WITH GENERALIZED LINEAR MODELS

The Australian Bureau of Statistics explored (generalized) linear modelling of the observations obtained from the sampling units for the estimation of discontinuities, Griffiths et al. (2016). In the ideal case there is an experimental design embedded in the sample survey to randomize sampling units over the treatments so as to estimate discontinuities unbiasedly. In that case the main difference with the methods described in Section 3 is the mode of inference. The analysis procedures proposed by e.g. Van den Brakel and Renssen (2005) and Chipperfield and Bell (2010) for embedded experiments account for the sampling design and experimental design by applying a design-based mode of inference. The methods proposed in Griffiths et al. (2016) can be seen as a form of covariance analysis and have the flexibility to account for the sampling design by including variables that model sampling design features.

The advantage of applying a linear model for the analysis of a parallel run is that under the assumed model, the power of testing hypotheses about discontinuities will be stronger compared to the design-based mode of inference. With the linear model, hypotheses about differences in discontinuities between different subpopulations can be easily analysed by modelling interactions between treatments and socio-demographic auxiliary variables. In the design-based approach this would require testing differences in domain estimates which will have lower power.

An intermediate approach is to apply the small area estimation methods from Section 5, which can be seen as composite estimator of a design-based estimator and synthetic prediction from a linear model.

In the case of strong auxiliary information, the linear modelling approach might be considered to estimate discontinuities without a parallel run. This requires combining samples observed before and after the change-over including strong auxiliary information to correctly separate the real change over time of the population parameter from differences in measurement bias. In this situation the time series modelling approach described in Section 4 is probably a more powerful approach for separating real evolution from discontinuity induced by the survey redesign, particularly if appropriate auxiliary information is lacking. In this context it should be emphasized that a parallel run based on a randomized experiment is helpful to improve robustness against model misspecification.

An advantage of using a design-based mode of inference for the analysis of parallel runs is that it is more robust against model misspecification, since it is based on a probability sample drawn from the intended target population, a randomized experiment to assign the selected sampling units randomly to one of the treatments and a design-based inference approach used to construct estimators for the intended population parameters that are approximately design-unbiased with respect to the sample design and the experimental design. This comes at the cost of a reduced power or increased variance in the analysis of discontinuities.

A drawback of establishing the impact of a redesign through a generalized linear model is that the regression coefficients that measure the impact are not necessarily directly related to the population parameters of the survey. This applies in particular to parameters defined as totals or the ratio of two estimated population totals and complicates the interpretation of the results.

7. ADJUSTING SERIES

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 often seems less natural, although there are sometimes reasons to do this (see 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 $\hat{\bar{Y}}_{T,reg}$ and $\hat{\bar{Y}}_{T,new}$ denote the estimates obtained during a parallel run, say in period T, respectively under the survey approach 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 generally 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 satisfactorily. 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 discontinued 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 mentioned in Section 4, a structural time series model is used for the production of monthly figures about the labour force in the Netherlands. This model is also used in combination with a parallel run to account for discontinuities due to the major redesigns in 2010 and 2012. In 2010 Statistics Netherlands was faced with large budget cuts to 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, however, 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. In the meantime expertise with web interviewing in household samples was built with the intent of eventually changing to the final sequential mixed-mode design based on web interviewing, CATI and CAPI in 2012. During the period from 2010 to 2012, the time series model as described in Van den Brakel and Krieg (2015) 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. Also the series published before 2012 were backcast to the level of the new process. In this way data users where confronted only once with the side effects of the transition.

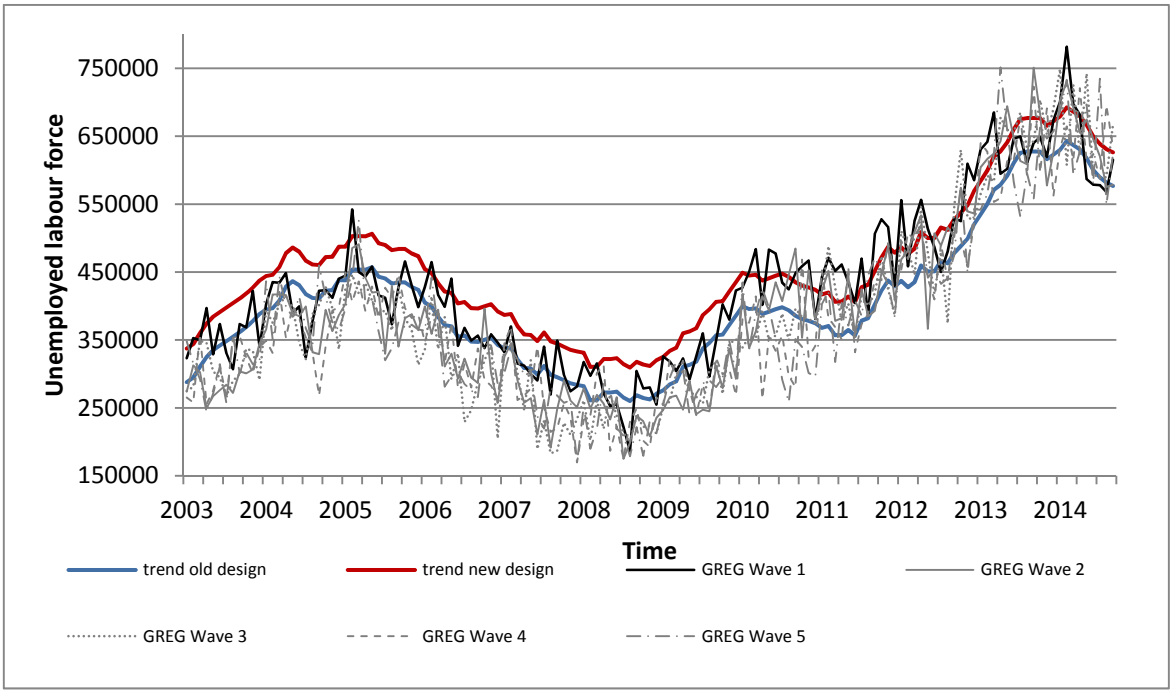
As an illustration, the top panel in Figure 4 shows results for the monthly figures of the unemployed labour force at the national level. The figure shows the five series of the GREG estimates observed in the five waves of the rotating panel, which are the input series for the model. The solid blue line 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 4 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.

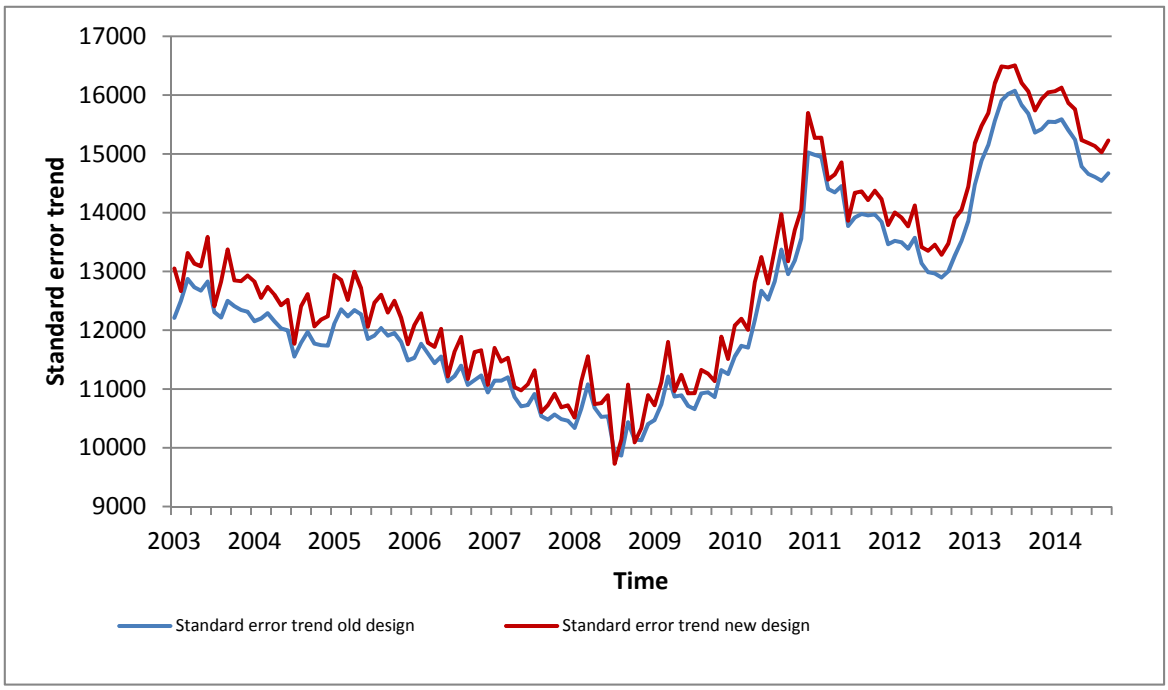
One of the consequences of these two redesigns is an increased level of uncertainty in the filtered estimates. The bottom panel of Figure 4 compares the standard errors of the filtered trend under the old and the new design.

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 which 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.

Section 5 explained that the transition from the Justice and Security Module in the Permanent Survey on Living Conditions to the Crime Victimization Survey in 2005 resulted in large unexpected increase of crime figures. Figure 1 illustrates how results under the Crime Victimization Survey were adjusted to the level of the series observed before the change-over with a state space intervention model. In 2008 another redesign was implemented. The experience in 2005 provided a reason to conduct a parallel run at a reduced sample size. The new design was conducted at the regular sample size, since it was directly used for publication purposes. The old design was conducted with a sample size of one third of the regular sample size. To maintain uninterrupted series with the past, these parallel runs were repeated in 2009, 2010, 2011 and 2012. Finally the small area estimation approach described in Section 5 was developed to produce estimates at the domain level under the old approach.

FIGURE 4: 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 of the filtered trend under the old and the new design.





8. DISCUSSION

Major redesigns and survey process transitions bring additional risk for the continuity of time series produced by national statistical institutes with repeated surveys. It is therefore important for a national statistical institute to have a statistical framework in place to manage the risks to official statistics from 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 recalculation. If no additional variables are required in the new process, use of recalculation 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 heavily increases the risk level of a redesign. In a worst-case scenario a period without regular official statistics is 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. 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. 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 for by conducting a parallel run at a small sample size and applying small area estimation methods to improve the precision of the impact estimates. Parallel runs can also be combined with 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.

Modelling of the observations is another way of assessing the impact. It can be regarded as a model-based mode of inference for a parallel run. It has the advantage that is has a stronger power compared to the other methods, in particular to test hypotheses about differences between discontinuities of different subpopulations. The method is however less robust to model misspecification and the model parameters that model the differences in measurement bias do not necessarily reflect the impact on the population parameters of interest from the survey, e.g. in the case of ratios. This method is in particular useful for analysing pilots that precede a parallel run to fine tune the final survey design and obtain quantitative insight into the effect of the underlying factors that are changed.

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. These methods generally depend on strong assumptions, therefore it may also be decided to just publish the estimated impacts as they occur during the period of the change-over.

REFERENCES

ABS (2009). Information Paper : ANZSIC 2006 Implementation in Retail Trade Statistics, July 2009, 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

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.

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.

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

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.

Koopman, S.J., A. Harvey, N. Shephard, and J.A. Doornik (2009). *STAMP 8.2*. London: Timberlake Consultants Press.

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 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 & 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.

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.

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. *CBS Discussion paper*, 2016/17, Statistics Netherlands, Heerlen.

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.

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.

FOR MORE INFORMATION . . .

www.abs.gov.au the ABS website is the best place for data from our publications and information about the ABS.

INFORMATION AND REFERRAL SERVICE

Our consultants can help you access the full range of information published by the ABS that is available free of charge from our website. Information tailored to your needs can also be requested as a 'user pays' service. Specialists are on hand to help you with analytical or methodological advice.

POST Client Services, ABS, GPO Box 796, Sydney NSW 2001

FAX 1300 135 211

EMAIL client.services@abs.gov.au

PHONE 1300 135 070

FREE ACCESS TO STATISTICS

All ABS statistics can be downloaded free of charge from the ABS web site.

WEB ADDRESS www.abs.gov.au

Produced by the Australian Bureau of Statistics