Survey Research Methods (2008)
Vol.2 , No.3 , pp. 123-141
ISSN 1864-3361
http://www.surveymethods.org

Quality procedures for survey transitions - experiments, time series and discontinuities

Jan A. van den Brakel
Statistics Netherlands

Paul A. Smith
Office for National Statistics, Newport

Simon Compton
Office for National Statistics, London

To maintain uninterrupted time series, surveys conducted by national statistical institutes are often kept unchanged as long as possible. When a change is proposed to improve the methods, it may affect the continuity of these series. It is important to minimise the impact so as to minimise the inconvenience for users. In this paper we set out the steps in an orderly transition, provide practical guidance on how to minimise discontinuities, and review methods for dealing with discontinuities if they arise so as to maintain a consistently-estimated series.
Keywords: backcasting, design and analysis of experiments, survey sampling, synthetic adjustments, time series analysis

1 Introduction

Many surveys run by official statistical organisations are continuous, and a significant aspect of their value comes from that continuity, sometimes over very long periods. Methods, procedures and definitions applied in the survey gradually become outdated, which makes change and improvement inevitable from time to time. This, however, may affect the continuity of the time series of survey outputs and make the statistics less suitable for users' needs; in extreme cases this may lead to poor policies or decision-making (for one example see Chambers, Weale and Youll 2000). Therefore it is important to minimise the impact of such changes, and to maximise the utility of the information for the users. Consultation with users and the presentation of findings and results need to be considered throughout the transition. In this paper we examine the statistical aspects of planning and implementing a transition between methods in a survey, identify guidelines for this process, and apply them in some example survey changes.

In an ideal transition process, the new approach is tested to determine what its effect will be, before it is fully adopted as part of the regular survey. In cases where the underlying data remain the same, the differences can be investigated by recalculation, for example the introduction of new editing, imputation or estimation methods. Also a new economic activity classification system in business surveys will generally result in discontinuities in time series, which can be quantified using the same data with the addition of the new classification. In these cases it may be appropriate on cost grounds to evaluate the difference on a subsample and use this information to make inferences about the overall change (see for example Clogg et al. 1991).

Where collection or capture procedures are affected, however, the data are not consistent, and in these cases we need an alternative approach to obtain information from which to assess the impact of the change. A natural way to evaluate the effect of the change in approach is to conduct a field experiment where the regular and new survey design are run concurrently. This allows us to estimate the main survey parameters under both survey designs and to test whether these estimates are significantly different. A field experiment also provides a safe method of transition, since the new approach is conducted with a full-scale sample before its formal acceptance and implementation. Finally we have the implementation step and the need to estimate the discontinuity in a production situation, and to use this estimate to produce consistent series, for example by using a backcasting procedure.

The focus of this paper is twofold - first to describe the methods appropriate to quantify the effect of a survey redesign, particularly those for designing an experiment in a survey context, and the constraints under which this type of experiment can be carried out. We use examples from social surveys to illustrate the approaches, although they are generally applicable to business surveys too, with appropriate modifications for the characteristic differences in the sampling structure of those surveys (Rivière 2002). We also consider situations where an embedded experiment may not be practical, and what can be achieved in these situations.

Second, we examine a range of approaches for dealing with discontinuities which have been detected as part of a change to the survey. Many of these approaches are analytical and require technical development, but we also cover practical measures designed to ensure that the users are not surprised by the changes, and know what to do to compensate for differences in series within their own systems. From this we deduce best practice for safely introducing appropriate methods, and dealing with their effects.

Contact information: Jan A. van den Brakel, Statistics Netherlands, Kloosterweg 1, 6412 CN Heerlen, Netherlands, e-mail: jbrl@cbs.nl

The structure of the paper expands these two parts: in section 2 three examples of survey redesigns are discussed. An overview of different methods that can be used to quantify the effect of a survey redesign is given in section 3. One of these methods is to conduct a parallel run by means of an embedded experiment. The different aspects of this approach are further detailed in sections 4 and 5. In section 4 the methods for testing, the use of significance and power measures, what can be deduced from tests, and some aspects of design and analysis of experiments embedded in ongoing sample surveys are reviewed. Section 5 discusses the situation in which a full-scale experiment is not possible. In these situations it is important to maximise the opportunities for understanding and assessing potential sources of discontinuity from any piloting or field trials which may be taking place. Two generic backcasting procedures for joining series together are discussed in section 6 - a synthetic approach, and time series methods. A numerical example is worked out in section 7. In section 8 some general principles are set out for keeping the quality as high as possible during transitions in surveys, based on the discussions in earlier sections.

2 Examples

2.1 Dutch National Travel Survey

The Dutch National Travel Survey is a household survey. From 1985-1998 households were telephoned to collect household level information. Subsequently each household member was asked to keep a record of all the journeys for one day in journey diaries, which were sent by mail. Under this survey design, the response rates gradually dropped to about 55%. To improve response rates, the National Travel Survey was redesigned in 1998. To collect data, paper questionnaires were sent by mail (Paper-and-pencil interviewing - PAPI). Households receive a household questionnaire and journey diaries, which were substantially simplified compared to the old questionnaires. Since the response rates for PAPI surveys are generally low, all households were contacted by telephone immediately after sending the questionnaires to motivate them to complete the questionnaires. The interviewers also assisted the household members with the completion of the questionnaires, or followed up incorrect or incomplete questionnaires. If households did not respond, they were contacted by telephone, or reminders were sent by mail.

In 1998, the regular and the new designs were conducted in parallel for one complete year. The objective of this experiment was twofold. First, to test whether it was possible to use this new design on a large scale in Statistics Netherlands' fieldwork organization. The success of this new design depended strongly on the capability of the fieldwork organization to keep close contact with the sampled households to motivate them to participate in the survey. For a continuously conducted survey with an average monthly sample size of 13,000 addresses it was not obvious in advance that this was tenable. Second, this experiment was used to quantify discontinuities in the time series of the main parameters of the National Travel Survey due to this redesign.

In 1998, the monthly sample size was 14,650 addresses. The monthly sample was randomly divided into two subsamples according to a completely randomized design. It would have been more efficient if a randomized block design was applied with strata as the block variable, see section 4. In the first two quarters, the monthly subsample size assigned to the regular design amounted to 13,000 addresses, and the remaining 1,650 addresses were assigned to the new approach. During the last two quarters, the size of the subsample assigned to the new design was gradually raised to 13,000 addresses while the size of the subsample assigned to the regular design was gradually reduced to 1,650 addresses. During this year enough experience was obtained to change safely to this new design in 1999. With the new design a response rate of more then 70% has been achieved.

2.2 Dutch Security Monitor

In this example two surveys, the Permanent Survey on Living Conditions and the Population Police Monitor, are integrated into one new survey, the Security Monitor. The Permanent Survey on Living Conditions is a module-based integrated survey combining various themes concerning living conditions and quality of life. This survey has been conducted by Statistics Netherlands since 1997. One of the modules is used to publish figures about justice and crime victimisation, and is called the Justice and Security module (JSM). Parallel to this survey, the Population Police Monitor (PPM) has been conducted since 1993 under the auspices of the Ministry of Justice and the Ministry of Interior and Kingdom Relations to publish figures about police performance, security perception and crime victimisation. There was pressure to produce consistent figures about the overlapping themes of both surveys and to reduce response burden and costs, so it was planned in 2004 that the JSM module of the Permanent Survey on Living Conditions and the PPM would be replaced by the Dutch Security Monitor (SM), which would be conducted by Statistics Netherlands.

The PPM was a telephone interview survey of persons aged 15 years or older with a non-secret permanent telephone connection. From 1993-2001, it was conducted biannually and from 2001-2006 annually in the first quarter of the year. The sample size of the PPM varied between 25,000 and 52,000 persons. The JSM and the SM both use persons aged 15 years or older as the target population. In the JSM, interviewers visited all the sampled persons at home and administered the questionnaire in a face-to-face interview (CAPI). This was a continuously conducted survey with a yearly net sample size of about 10,000 persons. The data collection of the SM is based on a mixed mode design. Persons with a non-secret permanent telephone connection are interviewed by telephone (CATI), and other persons are interviewed face-to-face. The data collection of the SM is also conducted in the first quarter of the year.

In the changeover from the JSM to the SM, the questionnaire and the context of the survey changed since questions from the JSM are skipped and new questions from the PPM are added. The data collection period changed from a survey

that is continuously conducted throughout the year to the first quarter of the year. The data collection mode changed from a uni-mode design via CAPI to a mixed mode design via CAPI and CATI. The conversion from the PPM to the SM implied major modifications in the questionnaire and context of the survey. The target population changed from persons aged 15 years or older with a non-secret permanent telephone connection to the entire population of all persons aged 15 years and over. The data collection mode changed from CATI to a mixed mode design via CAPI and CATI.

In the first quarter of 2005 a two-treatment experiment was conducted to test the effect of the changeover from the PPM to the SM on the most important parameter estimates that originate from the PPM. A net sample size of about 52,500 persons was observed under the PPM and 5,500 persons under the SM. Different parameters in the survey process changed in this redesign. A consequence of the two-treatment experiment is that all factors that changed in the survey design are confounded. As a result one can only estimate the net effect of all the factors that changed simultaneously. In section 7.1, one of the most important parameters that originates from the PPM is analysed, satisfaction with police performance. It is measured as the fraction of respondents that have had contact with the police during the 12 months prior to the interview that were satisfied with police performance.

For budgetary reasons, the JSM stopped at the end of 2004. This hampers a direct comparison between parameter estimates of the JSM and the SM based on data observed in the first quarter of 2005. Time series forecasts of the JSM variables were made as the best possible substitute. In section 7.2, a set of crime victimization parameters that originates from the JSM are analysed. These are the mean number of total offences against Dutch inhabitants during the 12 months prior to the interview and its breakdown over the categories violence, property, and vandalism offences.

2.3 Census test in England & Wales

A Census Test took place in 2007, to provide evidence which will be used for decision-making for the population census in England and Wales in 2011. Similar tests are used by several NSIs (for example the US, Canada, New Zealand) to examine the effects of different approaches ahead of the full population census. The target of the 2007 test in England & Wales was to examine the effect of different treatments for delivery (hand delivery or postal delivery) and the effect of the inclusion of a question on income, on the response rate. The delivery method testing took place within five strata, defined by expected response based on a model derived from 2001 Census data; the strata were formed by uneven division of the *national* range of predicted responses, with one stratum covering the 2% of areas with the lowest predicted responses. 100,000 households were covered, divided equally between the five strata, but some additional control of variation between selected areas was built in through stratification on area characteristics. Generalising from the results of the Census Test is complicated because it is not possible to replicate the Census conditions for the test - the Census is compulsory, but the test is only voluntary, and this means that they have very different response rates. The test also takes place in only a restricted subset of regions (Local Authorities in this case) which, although they exhibit a range of characteristics, are chosen purposively. This means that the experiment will not give quantitative estimates of the expected change in outcomes for the full Census, but will provide circumstantial evidence which is then available alongside other evidence for making an appropriate decision on which delivery method to choose, and whether or not to include an income question.

Constraints on the significance (5%) and power (95%) for detecting a 2% difference in response were specified at the beginning of the design stage, but it proved impossible to meet these for separate tests within the five strata within the resources available. By testing separately in the five strata it is hoped to identify whether benefits are realisable by having different strategies for delivery in different strata. Although the treatments are applied to small areas (the primary sampling units (PSU's) within strata for the delivery method treatment), we quote the sample sizes in terms of number of households, since the PSU's vary in size.

3 Quantifying the effect of a survey redesign

It is well known that adjustments in the survey process can affect survey characteristics such as response bias and therefore have a systematic effect on the parameter estimates of a sample survey. When an ongoing survey is changed, it is not clear whether a change in the series is a result of a real development or is induced by the redesign. Even if no change in the series is observed, it is still possible that a real development could be nullified by an opposite redesign effect.

A general way to avoid confounding the autonomous development with redesign effects is to conduct an experiment embedded in the ongoing survey, where the regular and new approaches are run concurrently for some period. In an embedded experiment, the sample is randomly divided into two (or more) subsamples according to an experimental design. In survey literature, such experiments are also referred to as split-ballot designs or interpenetrating subsampling, and date back to Mahalanobis (1946), but see also Fellegi (1964), Cochran (1977 section 13.15), Hartley and Rao (1978), and Fienberg and Tanur (1987, 1988, 1989). Under this approach, the subsamples can be considered as probability samples from the target population. Therefore estimates of the target parameters under the different treatments can be obtained to compare the effects of the redesign and test whether these parameter estimates are significantly different.

Experiments are particularly appropriate if the sample data observed under the regular and the new approach are not consistent. This is for example the case if the redesign affects the questionnaire or the data collection procedure. In other situations it might be possible to quantify the effect of a redesign through recalculation using the sample data obtained

under the regular approach, possibly completed with some additional variables. For example the effect of a new classification system can be quantified using the data observed under the regular survey, completed with a domain indicator variable that is based on the new classification system. In this case the difference induced by the introduction of a new classification system can be analysed with the standard sampling theory for domain estimators, see for example Särndal et al. (1992 Section 5.8 and 5.9). Also the effect of new coding procedures might be established using a multiple imputation approach to add the new variable to the regular survey (Clogg et al. 1991).

Related to the experimental approach is two-phase sampling (or double sampling) which can be used to adjust for measurement error or some other types of differences in surveys. The approach is generally to have a large sample, typically of a type of measurement that is cheap but not accurate. From this large sample a smaller subsample is drawn with a type of measurement that is expensive but accurate. Subsequently, the correlation between these variables is exploited to improve the precision of the accurate measurement with the sample size of the cheap measurement. There are numerous examples of this technique - for introductory reviews see Biemer and Stokes (1991), Groves (1989, Chapter 7) and Särndal et al. (1992, Chapter 9). In our context this could translate to a large sample on the existing methodology and a small subsample on the new methodology. The major limitation of this approach is that in the small subsample, data are required under both the regular and the new survey. This might be feasible in the case of new classifications or coding procedures, for example, but will generally not be feasible in the case of different data collection procedures or questionnaires.

Another major advantage of quantifying the effect of a redesign through recalculation or conducting an experiment is that it provides a safe method of transition from a regular to a new design. If the new design turns out to be a failure, the data obtained under the regular design can still be used for publication purposes. This reduces the risk that there is a period for which no reliable figures are available. In the example of section 2.2, the experiment demonstrated that the new design resulted in a discontinuity in the parameter "satisfaction with police performance" of about 10%. This was a reason for one of the main users, the Ministry of Interior and Kingdom Relations, to continue the PPM in 2006.

Finally time series models, which will be developed in section 6.2, can be used to quantify the effect of a redesign. Time series models are appropriate to join series together, particularly if there are sufficient observations available under the new approach. This approach is also a second best option to quantify discontinuities if a parallel run cannot be conducted, for example because of budget constraints, as in the case of the change over from the Dutch JSM to the SM in section 2.2. Timeliness is the main drawback of this approach, since the effect of the redesign is estimated more accurately as more data on the new design become available.

4 Field experiments for evaluating survey changes

Randomized experiments are typically undertaken under a clearly specified protocol, which sets out in advance what is to be tested, decision rules for the test outcomes, the procedures to be followed and the analysis to be undertaken. The key decisions which need to be set out when an experiment (whether or not part of a sample survey) is set up are:

- clear definitions of and a decision about the number of treatment factors and treatment levels
- clear specification of the hypotheses about the main effects and interactions between the different treatment factors that need to be analysed
- dependent variables (parameters for which hypotheses about treatment effects are tested)
- the differences between the parameter estimates, i.e. the main effects and their interactions, that at least should result in a rejection of the null hypothesis of no treatment effects
- the power and significance levels to test these hypotheses
- experimental design (randomisation of sampling units over the treatments and level of randomisation)
- decisions concerning the use of the field staff in the data collection of the experiment
- minimum required sample size
- the method of analysis including a decision whether a design-based or model-based approach is applied

This results in the specification of the hypotheses to be tested. The typical approach in design and analysis of experiments is to pre-specify and quantify the objective of the experiment to avoid unnecessary post hoc analysis. A general framework and practical guidelines for this process of planning and conducting experiments are given by Robinson (2000).

Before a large scale field experiment is planned to test hypotheses about discontinuities, the survey process of the redesign must be definite. This implies that pilots to test a new approach strategy or questionnaire must precede field experiments that are aimed to test differences in the target parameters due to the survey redesign. It is perilous to combine both purposes in the same experiment. The results of the experiment might indicate that the new survey process must be adjusted. In this stage of a survey redesign, however, there is often no time and budget to conduct a new large scale field experiment to investigate discontinuities of the revised survey process.

The most straightforward approach is to split the sample into subsamples by means of a completely randomized design (CRD). Generally this is not the most efficient design available. The power of an experiment might be improved by using sampling structures such as strata, clusters or interviewers as block variables in a randomized block design (RBD) (Fienberg and Tanur 1987, 1988). Unrestricted randomization by means of a CRD might also result in practical complications, like long travelling distances for interviewers. This can be avoided by using small geographical regions as a block variable.

In the case of clustering it might be unattractive to randomize ultimate sampling units over the treatments. There might be practical objections to assigning respondents that belong to the same household or that are interviewed by the same interviewer, to different treatments in the experiment. In such situations we can consider randomising clusters of sampling units over the treatments, at the cost of reduced power. See Van den Brakel (2008) for a detailed discussion.

The field staff also requires special attention in the planning and design stage of an experiment. To draw conclusions that can be generalised to a situation were the new approach is implemented as a standard, it is advisable to use the entire or a representative sample of the field staff. Newly recruited staff, on the other hand, might be precluded for this reason. It is also advisable to provide sufficient training, to ensure that the field staff has sufficient experience with the data collection under the new approach. One might also anticipate that the data collected under the new approach in the first period of the experiment cannot be used in the analysis, since the interviewers must adapt to or gain sufficient experience with the new methods.

From a statistical point of view it is attractive to use interviewers as the block variable in an RBD, since this removes the interviewer variance component from the analysis of the experiment. A major drawback is that this implies that each interviewer has to collect data under both the regular and the new methodology, which might give rise to confusion. If it is decided that interviewers are assigned to one treatment only, then this must be done randomly to avoid one of the treatments being systematically favoured with experienced interviewers or handicapped with newly recruited staff. See Van den Brakel and Renssen (1998) and Van den Brakel (2008) for more details about issues concerning the field staff in embedded experiments.

In each application the right trade-off between the number of treatments in one experiment and the accompanying practical problems must be established carefully. Users generally expect that the effect of each separate factor that has varied in the survey process can be quantified. This generally requires a factorial design, which is difficult to apply in the fieldwork of a survey process, since the number of treatment combinations grows rapidly. One solution is to confound higher order interactions with blocks or to apply fractional factorial designs, see for example Montgomery (2001). Confounding is a design technique for arranging a complete factorial experiment in blocks, where the number of treatment combinations within a block is smaller than the number of treatment combinations of the factorial experiment. This requires that certain treatment effects, generally higher order interactions, are indistinguishable from blocks. Such design techniques might be used if interviewers are blocks and it is necessary to reduce the number of treatment combinations assigned to each interviewer. In fractional factorial designs, the number of treatment combinations is reduced by running only a fraction of the complete factorial experiment. Again this implies that higher order interactions are not distinguishable from each other. These designs, however, are highly balanced and generally hard to combine with the fieldwork restrictions encountered in the daily practise of survey sampling. In practice it is usually necessary to combine the factors that changed into one treatment and test the total effect against the standard alternative in a two-treatment experiment. This implies that the effects of all factors in the experiment are confounded and cannot be separately estimated.

Another consideration is the minimum required sample size. An indication is required about the size of the treatment effects that should at least result in a rejection of the null hypothesis at prespecified levels of significance and power. Based on these, the minimum subsample sizes can be determined by an appropriate power calculation, see for example Montgomery (2001). In survey sampling minimum sample size requirements are generally based on significance level requirements only (as in Cochran 1977, chapter 4). Therefore, as an example, we give expressions for the minimum sample size in the case of a two-treatment experiment in appendix A.

If we treat the Census test example of section 2.3 as a survey experiment, we have two treatments for delivery and two treatments for the inclusion or not of an income question, in a fully factorial design. The level at which differences in response are required are different for the two treatments. For the delivery method treatment, discrimination within each stratum of the five stratum breakdown is needed, to give information from which to choose delivery methods in different types of areas. This therefore provides the tightest constraint. The original objective, for a detectable difference of 2% within each of the five detailed strata for the delivery method would have required a PSU sample size sufficient to give 120,000 households in each stratum (using formula (17) in appendix A with equal subsample sizes). This 600,000 household sample size is however completely impractical.

For the income question test, only a single difference over all strata is required. The constraints (2% difference, 95% power and 5% significance) for the income question test would have been met by a test of 200,000 households. With a sample size of 200,000 households over all strata, a difference of 4.4% for the delivery method test within each stratum could be detected. In the event resource constraints meant that the test was only of 100,000 households.

In the example of section 2.2, the sample size assigned to the regular sample, i.e. the PPM, was fixed in advance. The net sample size of 5,500 persons for the experimental group, i.e. the subsample assigned to the SM, was determined using formula (16) in appendix A, requiring an overall significance level of 5% and a power of 90% to detect a difference of five percent points in the parameter satisfaction with police performance and a difference of three in the mean number of offences.

A design-based analysis procedure for experiments embedded in sample surveys designed as CRD's or RBD's that account for the sampling design and the weighting procedure of the ongoing survey is proposed by Van den Brakel and Van Berkel (2002), Van den Brakel and Renssen (1998, 2005) and Van den Brakel (2008). In their approach the Horvitz-Thompson estimator and the generalized regression estimator are applied to derive approximately design unbiased esti-

mators for the population parameters observed under the different treatments of the experiment. Furthermore, an approximately design unbiased estimator for the covariance matrix of the contrasts between the parameter estimates is derived. This gives rise to a design-based Wald- or t-statistic to test whether finite population parameter estimates observed under different treatments or survey implementations are significantly different. These analysis procedures are implemented in a software package, called X-tool, which is available as a component of the Blaise survey processing software, developed by Statistics Netherlands (Statistics Netherlands 2002).

5 Practical restrictions of field experiments

For many reasons, but often including resource constraints as in the Census Test example described above, it will not always be possible to achieve the constraints of significance and power simultaneously. In these cases we would normally expect to relax one of these, and often it is the power which is adjusted. The risk in testing a difference on a low power is that an observed difference can be found not to be significant, but a noticeable discontinuity can still be found after implementation of the change in the regular survey. This is particularly important if a cheaper approach is tested which might result in an increased response bias. The mismatch between the aim and the resources may, nevertheless, be too great. There are several alternatives in such situations.

(a) Increase the effective sample size by removing sample design constraints such as clustering and select an efficient experimental design. As mentioned in section 4, it is efficient to use homogeneous groups of sampling units as a block variable since such designs increase the power of an experiment. Fieldwork restrictions might make it attractive to randomize clusters of sampling units over the different treatments. For example all sampling units that belong to the same household or that are assigned to the same interviewer. This, however, will increase the variance of the treatment effects and can be avoided by randomizing the ultimate sampling units instead of clusters of sampling units over the treatments. In the Census Test example, the treatments are applied to whole postcodes. This reduces the effective sample size but has important operational benefits. Systems to deliver questionnaires to whole postcodes are much simpler than those to deliver questionnaires by hand to some addresses and by post to others.

(b) If there is insufficient field capacity, consider changing the experiment from a one-off to a parallel run which can be managed over a longer period.

(c) If no large differences are expected, one might consider using the data obtained under the alternative treatments for the regular publication. In this case it is advisable to assign relatively small fractions of the sample to the alternative treatments and conduct the experiment over a longer period to achieve the required sample size. If it turns out that the differences are too large to use the data obtained under the alternative treatments for the regular publication, then the loss of accuracy in the regular figures remains limited. In this situation the experiment can be terminated sooner, since a smaller sample size is needed than was anticipated in advance.

(d) Restricting the experiment to the most important research question(s). In example 2.2 the data collection mode changed from a uni-mode to a mixed mode approach. From this an additional research question arises, namely to quantify the effects of the two data collection modes (telephone and face-to-face interviewing) in the SM. This should confirm that the data obtained under different data collection modes within the same sample are comparable in order to preclude problems with data integrity. This requires, however, that a randomly selected part of the sampling units with a non-secret permanent telephone connection are assigned to the CAPI mode. As a result, the effective sample size to quantify the effect of collecting data under the survey design of the SM compared to the PPM or the Permanent Survey on Living Conditions on the most important parameters would be reduced. Therefore it was decided not to test this additional hypothesis and assume that the data obtained under both modes can be combined in the generalized regression estimator. The generalized regression estimator of the SM accounts for different amounts of nonresponse bias under both modes since the weighting model contains variables that stratify the population into subpopulations with and without a non-secret permanent telephone connection crossed with region and age classifications. This estimation procedure will, however, not correct for different amounts of response bias between data collection modes.

(e) Assume that the discontinuities observed at the national level hold for subpopulations. In example 2.2 a regional analysis comparable with the precision of the regular survey at the national level was out of the question. The main objective of the PPM, however, is to estimate figures about police performance at a regional level for 25 separate police districts. Figures for these 25 police regions are based on sample sizes that vary between 1,000 and 2,500 respondents. Therefore it was decided to analyse mode effects at the national level and assume that the observed differences also held at regional levels. That is, it is assumed that there is no interaction between region and treatment. This hypothesis could not be rejected in this particular application. The problem with this approach is that in these situations the experiment does not have sufficient power to detect these interactions. Under this assumption a reasonable precision for the analysis of discontinuities for these regional figures was achievable in spite of the relatively small sample size of the experimental group. In section 7.1 we will apply this idea on the observed differences of the parameter satisfaction with police performance.

(f) Undertake the experiment, and analyse it to infer which parameters have the largest effect on the estimates, with less regard for whether this effect is significant. If the factors detected in this way corroborate conceptions based on experience, then it may well be valid to take the evidence such as it is from the experiment and the experience together in determining which approach to adopt. We would

still expect this strategy to be better than deciding only from experience what to do and needing to deal with any impacts afterwards, and this is the strategy being adopted in the UK Census test, where non-significant differences in delivery methods may be noted within each of the strata, but if they have some correlation with the stratum characteristics or with other evidence, they may give quite good evidence for the impact of those methods.

6 Implementation of changes and dealing with discontinuities

There are several ways to deal with observed discontinuities. A conservative approach is to quantify the discontinuities only for the period in which both approaches are run concurrently (without extrapolation). This implies that the autonomous development in the series is separated from the effect of the redesign on the parameter estimates for this period only. One example of this approach is for population censuses, where the 'survey repeats' are typically so far apart that a single period of overlap is the most that can be achieved - see for example Clogg et al. (1991), where a sample of units is dual coded, and multiple imputation is used to show the impact of a change in classification. Providing such a one-time estimate of the change can in general be considered as a design-based and rather safe approach since the observed effects are not extrapolated beyond the period where both approaches were run concurrently. On the other hand, this generally does not meet the users' requirements, since they often desire uninterrupted series for policy evaluation.

6.1 Synthetic correction methods

Other methods, which meet the requirement of maintaining uninterrupted series, rely on a model to adjust the series for the observed difference beyond the period where both approaches are run in parallel. Some models are available to adjust the series observed under the regular design, to make them comparable with the figures obtained under the new design, and these are discussed below. These procedures are in this context also known as backcasting.

Let T denote the period where both approaches are run concurrently by means of an experiment. Furthermore $\hat{y}_{R,T}$ and $\hat{y}_{N,T}$ denote the design-based estimators for a parameter observed under the regular and the new design respectively at time T. The most straightforward approach is an additive adjustment of the series, which is obtained with

$$\tilde{y}_{N,t} = \hat{y}_{R,t} + (\hat{y}_{N,T} - \hat{y}_{R,T}) \equiv \hat{y}_{R,t} + \hat{\Delta}_T, \quad \text{for } t = 1, \ldots, T-1, \quad (1)$$

with $\hat{\Delta}_T = \hat{y}_{N,T} - \hat{y}_{R,T}$. Model (1) implies that the correction is independent of the value of $\hat{y}_{R,t}$. This might result in an adjusted series that takes values outside the admissible range of the parameter. To avoid (for example) negative values, a multiplicative correction might be preferred:

$$\tilde{y}_{N,t} = \hat{y}_{R,t} \frac{\hat{y}_{N,T}}{\hat{y}_{R,T}}, \text{ for } t = 1, \ldots, T-1. \quad (2)$$

This model assumes that the correction is proportional to the value of $\hat{y}_{R,t}$, which is often a more plausible assumption than the independence assumption required for an additive adjustment.

Both adjustments (1) and (2) may be inappropriate for certain parameters. For example fractions can only take values in the range [0,1]. Adjustment (2) can still result in adjusted parameter estimates that take values larger than one. For the series of the police performance in example 2.2 the following adjustment is proposed for fractions:

$$\tilde{y}_{N,t} = \hat{y}_{R,t} + \gamma \hat{\Delta}_T \, \delta(\hat{y}_{R,t}), \text{ for } t = 1, \ldots, T-1. \quad (3)$$

Here $\delta(\hat{y}_{R,t})$ is a damping factor that take values in the range [0,1] and is defined as a function of $\hat{y}_{R,t}$, such that $\delta(\hat{y}_{R,t}) = 1$ if $\hat{y}_{R,t} = 1/2$ and $\delta(\hat{y}_{R,t}) = 0$ if $\hat{y}_{R,t} = 1$ or 0. From all possible functions that satisfy these conditions, the following quadratic form is chosen:

$$\delta(\hat{y}_{R,t}) = 4\hat{y}_{R,t}(1 - \hat{y}_{R,t}). \quad (4)$$

Since $\hat{y}_{R,t}(1 - \hat{y}_{R,t})$ is the population variance of an estimated fraction, (4) has the attractive statistical interpretation that $\delta(\hat{y}_{R,t})$ is proportional to the population variance of $\hat{y}_{R,t}$. As a result, the extent of the adjustment of a parameter estimate with (3) depends on the precision of this parameter estimate. Small population variances for the parameter result in smaller adjustments. Large population variances result in larger adjustments, with a maximum at $\hat{y}_{R,t} = 1/2$. Finally γ is chosen such that the equality in (3) holds exactly at time T, i.e. $\hat{y}_{N,T} = \hat{y}_{R,T} + \gamma \hat{\Delta}_T \, \delta(\hat{y}_{R,T})$. Inserting $\gamma = 1/\delta(\hat{y}_{R,T})$ in (3) gives:

$$\tilde{y}_{N,t} = \hat{y}_{R,t} + \hat{\Delta}_T \frac{\hat{y}_{R,t}(1 - \hat{y}_{R,t})}{\hat{y}_{R,T}(1 - \hat{y}_{R,T})}, \quad (5)$$

showing that the correction is proportional to the ratio of the population variances of $\hat{y}_{R,t}$ and $\hat{y}_{R,T}$.

Another property, which makes (3) appropriate for adjusting fractions is that the size of the adjustment is symmetric around $\hat{y}_{R,t} = 0.5$. For example, the adjustment for $\hat{y}_{R,t} = 0.1$ is the same as $\hat{y}_{R,t} = 0.9$. This does not hold for (2) since the adjustment under this model is proportional to the value of $\hat{y}_{R,t}$. Variance approximations for series adjusted with models (1), (2) or (5) are given in appendix B.

As mentioned in section 5, it might be necessary to use the discontinuities observed at the national level to adjust series of parameter estimates for subpopulations. Let $\hat{y}^r_{R,t}$ be the estimate for the r-th subpopulation in period t, obtained under the regular design. An additive adjustment for this series under the new design with the discontinuity observed at the national level is obtained with

$$\tilde{y}^r_{N,t} = \hat{y}^r_{R,t} + (\hat{y}_{N,T} - \hat{y}_{R,T}) \equiv \hat{y}^r_{R,t} + \hat{\Delta}_T, \quad \text{for } t = 1, \ldots, T-1. \quad (6)$$

In a similar way, a multiplicative adjustment is obtained with

$$\tilde{y}^r_{N,t} = \hat{y}^r_{R,t} \frac{\hat{y}_{N,T}}{\hat{y}_{R,T}}, \text{ for } t = 1, \ldots, T-1. \quad (7)$$

Adjusting a series defined as fractions for subpopulations can be obtained with

$$\tilde{y}^r_{N,t} = \hat{y}^r_{R,t} + \gamma\,\hat{\Delta}_T\,\delta(\hat{y}^r_{R,t}) = \hat{y}^r_{R,t} + \hat{\Delta}_T\,\frac{\hat{y}^r_{R,t}(1-\hat{y}^r_{R,t})}{\hat{y}_{R,T}(1-\hat{y}_{R,T})}. \quad (8)$$

Variance approximations for (6), (7) and (8) are given in appendix B.

Adjusting series according to models (1) - (8) is a synthetic approach and will almost certainly result in biased estimates for the adjusted series, since strong model assumptions are used to extrapolate the observed difference outside the period that both survey approaches run in parallel, or that differences observed at the national level also hold for subpopulations. This bias is not reflected in the variance approximations that are derived in appnedix B. These model assumptions become more questionable as the time between the adjusted parameter (t) and the moment of conducting the experiment (T) increases. Moreover it is very hard to validate this assumption. Indeed in one recent example Soroka et al. (2006) demonstrated that recalculating a series using exact methods (an exact classification in their case) could show substantial differences compared with using a linking approach.

6.2 Consistency between adjusted series

Adjusting series according to (2), (5), (7) or (8) might give rise to consistency problems. In the example of section 2.1, discontinuities are quantified for total travelling distance and its breakdown over different subclasses. If such series are adjusted according to (2), there is no guarantee that the sum over the adjusted subclasses equals the adjusted total. The same problem arises if fractions are adjusted according to (5). After this adjustment, there is no guarantee that the fractions sum to one. Consistencies between adjusted parameter estimates can be restored with a linear restriction estimator. Let $\tilde{y}_{N,t} = (\tilde{y}_{N,t,1}, ..., \tilde{y}_{N,t,q})$ denote a q-vector containing the q adjusted parameter estimates for period t. These q parameters must obey a set of m linear restrictions. This problem comes down to minimizing $(\tilde{y}^*_{N,t} - \tilde{y}_{N,t})^T V^{-1}(\tilde{y}^*_{N,t} - \tilde{y}_{N,t})$, subject to the constraint $R\tilde{y}^*_{N,t} = c$, with V the covariance matrix of $\tilde{y}_{N,t}$ and R a $m \times q$ matrix that contains the linear combinations of $\tilde{y}^*_{N,t}$ that must satisfy the m restrictions of c. Minimizing the Lagrangian function

$$(\tilde{y}^*_{N,t} - \tilde{y}_{N,t})^T V^{-1}(\tilde{y}^*_{N,t} - \tilde{y}_{N,t}) - \lambda(R\tilde{y}_{N,t} - c)$$

with respect to $\tilde{y}^*_{N,t}$ and λ, gives

$$\tilde{y}^*_{N,t} = \tilde{y}_{N,t} + VR^T(RVR^T)^{-1}[c - R\tilde{y}_{N,t}], \quad (9)$$

with covariance matrix

$$V(\tilde{y}^*_{N,y}) = V - VR^T(RVR^T)^{-1}RV, \quad (10)$$

see for example Knottnerus (2002, chapter 12). This quadratic minimization approach is sometimes applied for balancing estimates for national accounts (Stone, Champernowne and Meade, 1942) and benchmarking monthly and quarterly figures to annual totals (Denton, 1971). As long as R does not contain redundant restrictions, RVR^T is of full rank, since V is positive semi definite.

In equation (9) the discrepancies $[c - R\tilde{y}_{N,t}]$ are distributed over the values of $\tilde{y}_{N,t}$ such that imprecise elements of $\tilde{y}_{N,t}$ receive larger adjustments than more precise elements of $\tilde{y}_{N,t}$. For example, let $\tilde{y}_{N,t} = (\tilde{y}_{N,t,+}, \tilde{y}_{N,t,1}, \tilde{y}_{N,t,2}, \tilde{y}_{N,t,3})$ denote a vector that contains adjusted estimates for total travelling distance ($\tilde{y}_{N,t,+}$) and its breakdown over the distance travelled by car ($\tilde{y}_{N,t,1}$), by public transportation ($\tilde{y}_{N,t,2}$), and others forms of transportation ($\tilde{y}_{N,t,3}$). It is required that $\tilde{y}_{N,t,+} = \tilde{y}_{N,t,1} + \tilde{y}_{N,t,2} + \tilde{y}_{N,t,3}$. If $R = (1, -1, -1, -1)$ and $c = (0)$, then the four elements are adjusted, such that the largest adjustments are attributed to the most imprecise figures. If it is required that the estimate for the total travelling distance remains unaffected, then take $\tilde{y}_{N,t} = (\tilde{y}_{N,t,1}, \tilde{y}_{N,t,2}, \tilde{y}_{N,t,3})$, $R = (1, 1, 1)$, and $c = (\tilde{y}_{N,t,+})$. One could consider avoiding this procedure by adjusting the three classes $\tilde{y}_{N,t,1}$, $\tilde{y}_{N,t,2}$, and $\tilde{y}_{N,t,3}$ separately and deriving the adjusted series for the total from the sum over the adjusted classes. This is, however, not an efficient procedure since the available information from the experiment about the observed difference of the total is not used. This approach also results in unnecessarily large standard errors for the adjusted series for the total, particularly if the total is estimated more precisely than the separate classes. Furthermore, it follows directly from (10) that the constrained estimator (9) has smaller variances than the separately adjusted series, since the restriction adds additional information.

Constraining the separately adjusted series is particularly important if they specify the distribution over a set of categories. Let $\tilde{y}_{N,t} = (\tilde{y}_{N,t,1}, \tilde{y}_{N,t,2}, \tilde{y}_{N,t,3}, \tilde{y}_{N,t,4}, \tilde{y}_{N,t,5})$ denote a vector containing five adjusted proportions, for example the fraction of persons that is 1) very satisfied, 2) satisfied, 3) not satisfied and not unsatisfied, 4) unsatisfied, and 5) very unsatisfied with police performance. It is required that the sum over the five categories is one, which can be achieved with (9), by taking $R = (1, 1, 1, 1, 1)$ and $c = (1)$.

With the experimental approach, two estimates for a parameter are obtained, $\hat{y}_{R,T}$ and $\hat{y}_{N,T}$. It might be attractive to combine both figures, in a way that accounts for the uncertainty of both parameter estimates. This can be accomplished with linear restriction estimator (9), where

$$\hat{y}_T = (\hat{y}_{N,T}, \hat{y}_{R,T})^T, \;\; c = 0, \;\; R = (1, -1).$$

Under the assumption that the new survey process results in less response bias, the difference between $\hat{y}_{R,T}$ and $\hat{y}_{N,T}$ can be used as an estimate for the response bias in the mean squared error of $\hat{y}_{R,T}$. This implies that the MSE for both estimates can be approximated by

$$V_1 = Var(\hat{y}_{N,T}), \;\; V_2 = Var(\hat{y}_{R,T}) + (\hat{y}_{N,T} - \hat{y}_{R,T})^2.$$

After some algebra it follows that the two components of $\hat{y}^*_T$ in (9) are both equal to:

$$\hat{y}^*_T = \frac{V_2\hat{y}_{N,T} + V_1\hat{y}_{R,T}}{V_1 + V_2}.$$

This comes down to the regular way of pooling two estimates using their accuracy measures as weights. The covariance between $\hat{y}_{N,T}$ and $\hat{y}_{R,T}$ is neglected, which is reasonable in many applications, since the variances are of order $(1/n)$ while the covariance is of order $(1/N)$, where n denotes the sample size and N the size of the target population.

6.3 Time series approach

As pointed out in section 3, time series models are particularly interesting to deal with discontinuities for different reasons. They are appropriate to join series together and they might be a second best option to quantify the effect of a redesign in situations where there is no budget available to conduct a parallel run. As the survey proceeds, more data under the new approach become available, which might be used to obtain better estimates for the discontinuity through time series modelling even in the case of a parallel run.

In section 6.3.1 time series models are developed for situations where there are no overlapping periods between the regular and the new methodology. In section 6.3.2 bivariate models are developed for the case where there is an overlap between the regular and new approach. The focus in both sections is on structural time series models, although ARIMA models might also be appropriate.

6.3.1 Time series models without overlapping periods

One possibility to account for discontinuities where the regular and new approach are not conducted in parallel during some period, is to model the moment that the survey is redesigned explicitly in a time series model. This is generally referred to as intervention analysis. This approach assumes that the time series model approximates the development of the indicator reasonably well and that there is no structural change in the trend or the seasonal component at the moment that the new survey is implemented. If a change in the real development of the indicator does coincide with the implementation of the new survey, then the model will wrongly assign this effect to the intervention variable which is intended to describe the redesign effect.

One possibility is to use an augmented ARIMA model, using for example, the REGARIMA tool in X12-ARIMA (see Findley et al. 1998) or TRAMO in TRAMO-SEATS (see Gomez and Maravall 2000). This approach includes a dummy variable that incorporates auxiliary information on the time and duration of the transition period from the regular to the new design.

Another approach is to adopt a structural time series model, where the series is decomposed into a trend, a seasonal component, a component predicted with explanatory variables, and an irregular component. Again the vector with explanatory variables contains at least a dummy variable that indicates the moment that the survey changed from the regular to the new design. This intervention variable measures a change in level. Other forms of intervention variables can be designed, for example to measure a change in slope (Durbin and Koopman 2001, section 3.2) or to measure a gradual change of level (Box and Tiao 1975). The standard (but not the only) way to proceed is to write this model in state-space form and obtain parameter estimates with the Kalman filter (see for example Harvey 1989, or Durbin and Koopman 2001). The parameter estimate for the intervention variable can be interpreted as the discontinuity in the series due to the survey redesign.

Adjusting series independently form each other according to univariate time series models will give rise to inconsistencies between the adjusted parameters. They can be restored using the linear restriction approach (9) proposed in section 6.2. Another possibility is to apply a multivariate time series model and augment this model with an additional restriction on the regression coefficients of the intervention parameters. Let $\hat{y}_t = (\hat{y}_{t,+}, \hat{y}_{t,1}, ..., \hat{y}_{t,K})$ denote a vector that contains $K+1$ estimates for the mean or the total of a parameter. The first component, $\hat{y}_{t,+}$, is broken down over K categories specified by the remaining estimates $\hat{y}_{t,k}$, $k = 1 ... K$. These $K+1$ variables are subjected to the restriction $\hat{y}_{t,+} = \sum_{k=1}^{K} \hat{y}_{t,k}$ for all t. For illustrative purposes, the $K+1$ series are assumed to be the realisation of a stochastic trend and an intervention variable only. The intervention variable δ_t equals zero during the period that the series is observed under the regular approach and equals one during the period where the series is observed under the new approach. The univariate structural time series model for the j-th component of $\hat{y}_t$ is defined as:

$$\hat{y}_{t,j} = L_{t,j} + \beta_j\delta_t + \varepsilon_{t,j}, \tag{11}$$

with $L_{t,j}$ a stochastic trend, β_j the time independent regression coefficients for the intervention variable which can be interpreted as the discontinuity in the j-th series due to the survey redesign and $\varepsilon_{t,j}$ an irregular component with $E(\varepsilon_{t,j}) = 0$ and $Cov(\varepsilon_{t,j}, \varepsilon_{t',j}) = \sigma^2_{\varepsilon j}$ if $t = t'$ and zero otherwise. The stochastic trend is modelled as

$$\begin{aligned} L_{t,j} &= L_{t-1,j} + R_{t-1,j} + \eta_{t,Lj}\ , \\ R_{t,j} &= R_{t-1,j} + \eta_{t,Rj}\ , \end{aligned} \tag{12}$$

with $L_{t,j}$ the stochastic level component and $R_{t,j}$ the stochastic slope component, and $\eta_{t,Lj}$ and $\eta_{t,Rj}$ irregular components. It is assumed that $E(\eta_{t,Lj}) = E(\eta_{t,Rj}) = 0$, $Cov(\eta_{t,Lj}, \eta_{t',Lj}) = \sigma^2_{Lj}$ if $t = t'$ and zero otherwise, $Cov(\eta_{t,Rj}, \eta_{t',Rj}) = \sigma^2_{Rj}$ if $t = t'$ and zero otherwise, and $Cov(\eta_{t,Lj}, \eta_{t',Rj}) = 0$ for all t and t'. The model can be extended with a seasonal component, explanatory variables and even an ARMA component to remove remaining autocorrelation from the residuals if necessary (see for example Durbin and Koopman 2001). These $K+1$ univariate models can be put in one multivariate model augmented with restriction $\hat{y}_{t,+} = \sum_{k=1}^{K} \hat{y}_{t,k}$ for all t. In state space representation this model reads as:

$$\hat{y}_t = Z_t\alpha_t + \varepsilon_t \tag{13}$$

$$\alpha_t = T\alpha_{t-1} + \eta_t \tag{14}$$

The measurement equation (13) is the multivariate extension of (11) and describes how the observed series depends on a vector of unobserved state variables α_t and a vector with disturbances ε_t. In this case the state variables are the level and slope components of the trend models and the regression coefficients of the intervention variables. The transition equation (14) describes how these state variables evolve in time. The vector η_t contains the disturbances of the assumed stochastic processes of the state variables. Modelling the $K+1$ series with separate stochastic trend models and intervention variables, implies that the matrices in (13) and (14) are given by

$$\alpha_t = (L_{t,+}, R_{t,+}, L_{t,1}, R_{t,1}, ..., L_{t,K}, R_{t,K}, \beta_+, \beta_1, ..., \beta_K)^T,$$

$$Z_t = (I_{K+1} \otimes (1,0) \mid \delta_t I_{K+1}),$$

$$T = Blockdiag(T_{tr}, T_{iv}),$$

$$T_{tr} = I_{K+1} \otimes \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},$$

$$T_{iv} = \begin{pmatrix} 0 & 1_K^T \\ 0_K & I_K \end{pmatrix}, \tag{15}$$

with 0_p a vector of order p with each element equal to zero, 1_p a vector of order p with each element equal to one, and I_p the $p \times p$ identity matrix. The disturbances vectors are defined as

$$\varepsilon_t = (\varepsilon_{t,+},\ \varepsilon_{t,1}, ..., \varepsilon_{t,K})^T,$$

$$\eta_t = (\eta_{t,L+},\ \eta_{t,R+},\ \eta_{t,L1}, \eta_{t,R1}, ..., \eta_{t,LK}, \eta_{t,RK}, 0_{K+1}^T)^T.$$

It is assumed that

$$E(\varepsilon_t) = 0_{K+1}, Cov(\varepsilon_t) = Diag(\sigma^2_{\varepsilon+}, \sigma^2_{\varepsilon1}, ..., \sigma^2_{\varepsilon K}),$$

$$E(\eta_t) = 0_{3(K+1)}, Cov(\eta_t) =$$
$$Diag(\sigma^2_{L+}, \sigma^2_{R+}, \sigma^2_{L1}, \sigma^2_{R1}, ..., \sigma^2_{LK}, \sigma^2_{RK}, 0_{K+1}).$$

Since the regression coefficients of the intervention variables are time independent, their accompanying irregular terms in η_t as well as their variances equal zero. Due to (15) these regression coefficients as well as their Kalman filter estimates obey the restriction $\beta_+ = \sum_{k=1}^K \beta_k$. Subsequently the time series after the moment of the survey transition can be adjusted for the estimated discontinuities with $\tilde{y}_{t,j} = \hat{y}_{t,j} - \hat{\beta}_j$. As an alternative, the series before the survey transition can be adjusted with $\tilde{y}_{t,j} = \hat{y}_{t,j} + \hat{\beta}_j$. Since the observed series and the estimated discontinuities obey the required consistencies, the adjusted series also does.

A slightly different type of constraint requires that K series add up to a constant, i.e. $\sum_{k=1}^K \hat{y}_{t,k} = c$. For example in the case of proportions, $c = 1$. In this case, the K regression coefficients of the intervention variables must obey the restriction $\sum_{k=1}^K \beta_k = 0$. This requires a K dimensional multivariate structural time series model defined analogous to (13) and (14), where (15) is replaced by

$$T_{iv} = \begin{pmatrix} I_{K-1} & 0_{K-1} \\ -1_{K-1}^T & 0 \end{pmatrix}.$$

6.3.2 Time series models for overlapping periods

If the period where the regular and new approach are conducted in parallel is sufficiently long, it might be efficient to construct a bivariate structural time series model for the series $\hat{y}_t = (\hat{y}_{R,t}, \hat{y}_{N,t})^T$. Both components of the vector are observed together only during the period that the regular and the new survey approach are conducted in parallel. During the period that the series is only observed under the regular approach, the values for $\hat{y}_{N,t}$ are missing observations. During the period that the series is only observed under the new survey design, the values for $\hat{y}_{R,t}$ are missing observations. Subsequently, one common trend component, seasonal component and regression model for explanatory variables is assumed for both series. The differences between the observations under the regular and the new approach are modelled with an intervention variable. For illustrative purposes only a stochastic trend and an intervention is assumed. This model can be put in state space representation with (13) and (14), taking

$$\alpha_t = (L_t, R_t, \beta)^T, Z = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \end{pmatrix},$$

$$T = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$Cov(\varepsilon_t) = diag(\sigma^2_{\varepsilon r}, \sigma^2_{\varepsilon n}), Cov(\eta_t) = diag(\sigma^2_L, \sigma^2_R, 0).$$

In the state vector, L_t is the level component and R_t the slope component of the stochastic trend model defined by (11) and (12) common to both series being modelled. This approach conceives the reconstruction of the time series as a missing observation problem. The Kalman filter and the fixed interval smoother can be applied to obtain forecasts for the missing values of the series under the regular approach and backcasts for the missing values of the series under the new approach (Durbin and Koopman, 2001 section 2.8 and 4.8). Numerical problems, however, might be expected if there are only a few paired observations.

7 Numerical example: the Dutch Security Monitor

In example 2.2 it was explained that the Dutch Security Monitor (SM) replaced two partially overlapping surveys, namely the Population Police Monitor (PPM) and the Justice and Security Module (JSM) of the Permanent Survey on Living Conditions. This redesign resulted in discontinuities in many parameters from the PPM and the JSM, and these are the results of several factors that changed simultaneously. The most important ones are:

- Increase in the response rate of the SM compared to the JSM and the PPM, which might result in different amounts of non-response bias. The response to the SM is 70%, while the response rates of the JSM varied between 55% and 60% and the PPM between 50% and 64%.
- Differences between sample frames. The PPM is based on a sample of persons aged 15 years and older with a non-secret permanent telephone connection. The SM and the JSM are based on samples of all persons aged 15 years and older. A substantial part of the discontinuities in the parameters that originate from the PPM can be explained since the PPM does not observe the subpopulation that does not have a non-secret permanent telephone number. Additional analyses showed that this results in an under-representation of young people and ethnic minorities in the PPM. This group also reports a higher rate of crime victimization and a more negative opinion about police performance.
- Differences in data collection modes. The PPM is a telephone based survey. In the JSM, data are collected in face-to-face interviews conducted with the respondents at home. In the SM a mixed mode design is used. Interviews are conducted by telephone if persons have a non-secret permanent telephone connection. For the remaining persons data are collected in face-to-face interviews conducted at the respondents' homes. Many references in the literature emphasize that different collection modes have systematic effects on the responses, see for example De Leeuw (2005), and Dillman and Christian (2005).
- Differences between data collection periods. Data collection for the PPM and the SM is conducted in the first quarter of the year, while the JSM is conducted continuously throughout the year. Real developments result in different means if the data refer to a complete calendar year instead of the first quarter of it. There are also small seasonal effects in the quarterly figures of the property and violent offences of the JSM. This explains a small part of the discontinuities in the parameters that originate from the JSM.
- Differences between questionnaire designs. There are differences between the routing, order and formulation of the questions and the answer categories in the questionnaires of the PPM, SM and JSM. These might have systematic effects on the outcomes of these surveys. One of the most important differences is that the questionnaires of the SM and the JSM use a bounded recall procedure for retrospective questions about crime victimization, see Sudman, Finn and Lannom (1984). This approach, which is intended to minimize the number of measurement errors in remembering events due to telescoping, is not used in the questionnaire of the PPM and might result in an overestimation of the number of events in the PPM.
- Differences between the contexts of the surveys. The PPM and the SM are introduced as surveys that are focussed on crime victimization and safety topics. The JSM is a module of a more general survey on living conditions. This might have a systematic selection effect on the respondents who decide to participate in the survey. Furthermore, in the PPM and the SM the attention of the respondent is completely focussed on one topic, contrary to the JSM. This might influence the effort made by the respondents to answer the questions as well as possible.

It is difficult to quantify the separate effects of these factors, since they are confounded in the chosen design (see section 2.2). In the following we discuss the discontinuities and the possible ways of dealing with them for one parameter of the PPM, satisfaction with police performance (in section 7.1), and four parameters of the JSM, covering a total and breakdown of offences against Dutch inhabitants (in section 7.2).

7.1 Population Police Monitor: Satisfaction with police performance

One of the most important parameters from the PPM is the fraction of the population which is satisfied with police performance during their last contact with the police. The experiment in 2005 demonstrated that the new design resulted in a difference in this parameter of about 9.4%. This was a reason for the Ministry of Interior and Kingdom Relations to continue the PPM in 2006, since they use the PPM outputs to evaluate police performance. As a result, the SM and the PPM were conducted in parallel for two years, which gives an excellent opportunity to test hypotheses about discontinuities and investigate the performance of the proposed adjustments discussed in section 6.1. In 2005 the SM was conducted on a small scale; the number of respondents of the SM and the PPM were 5,200 and 52,500 respectively. In 2006 both surveys were conducted on a full scale; the net sample sizes of the SM and PPM were 22000 and 25000 persons respectively. The analysis results concerning the discontinuities in satisfaction with police performance between both surveys are summarized in Table 1.

Since satisfaction with police performance is defined as a fraction, formula (5) is proposed to adjust the series based on the PPM from 1993 to 2005 for the observed difference with the SM. Based on the difference observed in 2005 and the estimate obtained from the PPM in 2006, a prediction for the SM in 2006 can be obtained using adjustment (5). This prediction can be confronted with the real estimate obtained from the SM in 2006. In a similar way, the difference ob-

Table 1: Analysis of discontinuities in "Satisfaction with police performance" (standard errors in brackets)

Year	SM		PPM		Difference		z	p-value
2005	52.38	(1.42)	61.75	(0.62)	9.37	(1.55)	6.05	0.000
2006	55.14	(0.78)	63.40	(0.65)	8.26	(1.02)	8.14	0.000

served in 2006 and the estimate of the PPM in 2005 can be used to make a prediction for the SM in 2005. This prediction can be confronted with the real estimate obtained from the SM in 2005. The results are summarized in Table 2.

The estimate for the discontinuity in 2006 is about one percentage point smaller than the estimate in 2005 (Table 1). Since the sample size for the SM in 2006 is four times larger, the most accurate estimate for the difference is obtained from the 2006 data. The differences between the predicted and estimated values for satisfaction with police performance via the SM are not significantly different form zero (Table 2).

The original series based on the PPM with a 95% confidence interval, and a prediction of the series under the SM, based on formula (5) with a 95% confidence interval based on formula (18) are plotted in Figure 1. The figure clearly illustrates that the adjusted series draws heavily on the assumption that the observed difference in 2006 is time invariant. Since the proportions observed under the PPM take values in a relatively small interval [62% - 68%], the damping factor (4) does not result in large differences in the adjustments between the years in this application.

Due to the relatively small sample size of the SM in 2005, no accurate direct estimates for parameters and discontinuities are available for the 25 police regions. Therefore it was initially planned to assume that the national level estimate for the discontinuity also held at the regional levels. Under this assumption, predictions for the SM at a regional level are obtained with (8). This approach assumes no interaction between region and treatment effects. With the data obtained in 2005 the hypothesis of no interaction could not be rejected in a logistic regression analysis. Based on the difference observed at the national level in 2005 and estimates obtained in 2006 with the PPM for the 25 police regions, predictions for these regions under the SM are obtained with (8). These predictions are confronted with the regional estimates observed with the SM in 2006 in Table 3. The difference between the predicted and the estimated value under the SM is significantly different from zero ($z > 1.96$) in only one region. This difference is not significant, if a multiple comparison procedure, like Bonferroni or the more powerful sequentially rejective multiple test proposed by Holm (1979), is applied.

It is also possible to quantify discontinuities with the sample of the SM and the PPM in 2006 where both surveys are full scale. In this year the differences in the separate regions can be estimated and model (5) can be applied to predict the outcome for the SM in 2005 within each separate region. The standard errors for these predictions take values between 5.3 and 5.9 with a mean of 5.5. If the difference between the SM and the PPM observed at the national level in 2006 is used to predict the outcomes for the SM at the regional level in 2005 with (8), then the standard errors of these predictions take values close to 3.2. This illustrates that we cannot expect more accurate predictions if both surveys are replicated on a full scale and the estimated differences within the regions are used for prediction in this application. With the assumption that the difference observed at the national level holds in each separate region, model (8) borrows strength from other regions to adjust the estimates at a regional level. This reduces the variance of the estimated difference and therefore of the adjusted series for each region considerably. If, on the other hand, this assumption does not hold it will introduce additional bias in the adjusted regional series. In this application there is, however, no evidence against the assumption that the difference observed at the national level holds in each separate region. First, the hypothesis of no interaction between regions and the treatment effect could not be rejected. Second, there is only one region in Table 3 where the difference between the SM prediction and the direct estimate is outside the 95% confidence interval.

7.2 Justice and Security Module of the Permanent Survey on Living Conditions: crime victimisation

Important crime victimization parameters that originate from the JSM are violence, property and vandalism offences against Dutch inhabitants. In this section we discuss the effect of the redesign on the estimated mean number of these three types of offences against Dutch inhabitants, and their total, as an example. Since the JSM and the SM were not conducted in parallel, no direct comparison of estimates obtained under both surveys is possible. In Figure 2 the time series for the mean number of violence, property, vandalism and the total number of offences observed under the JSM and the SM are plotted. This figure suggests that there is a clear discontinuity at the moment that the JSM is replaced by the SM in 2005 for the property and vandalism offences, and only a small discontinuity for the violence offences. Since the redesign appears to have increased the estimates of the three reported offences, there is also a clear discontinuity in the total number of offences.

One way to analyse the discontinuities in these crime victimization series is to model these series with the multivariate structural time series model (13) and (14). For each series, a separate smooth trend model is assumed, which is given by (12) where $\sigma^2_{L_j} = 0$. Since the violence, property and vandalism offences add up to the total offences, the regression coefficients of the corresponding intervention variables must obey the same restriction and therefore evolve in

Table 2: Predicted Satisfaction with police performance (standard errors in brackets)

Year	SM prediction		SM estimate		Difference		z	p-value
2005	53.40	(1.27)	52.38	(1.42)	1.02	(1.91)	0.54	0.589
2006	54.19	(1.71)	55.14	(0.78)	-0.95	(1.88)	-0.51	0.610

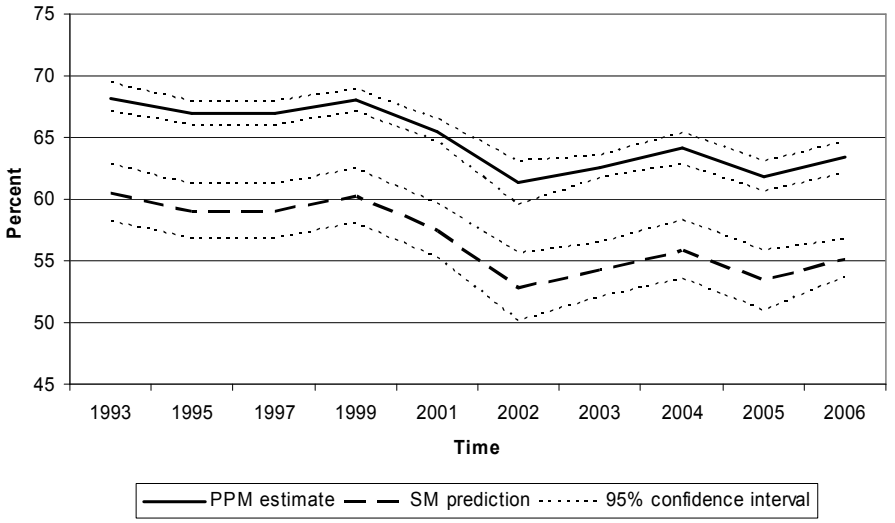

Figure 1. Time series for Satisfaction with police performance, original and adjusted for the discontinuity between the PPM and the SM observed in 2006

Table 3: Estimated and predicted values for Satisfaction with police performance for different police regions (standard errors in brackets)

Region	PPM estimate 2006		SM estimate 2006		SM predicted 2006		Difference of SM estimate from prediction		z
Amsterdam-Amstelland	62.4	(2.80)	52.5	(3.57)	53.1	(3.45)	-0.6	(4.96)	-0.12
Rotterdam-Rijnmond	60.4	(2.82)	46.9	(3.46)	50.9	(3.45)	-4.0	(4.88)	-0.82
Haaglanden	62.8	(2.79)	55.7	(3.32)	53.5	(3.45)	2.2	(4.78)	0.45
Utrecht	61.1	(2.81)	53.6	(3.40)	51.7	(3.45)	1.9	(4.84)	0.40
Midden- en West-Brabant	58.0	(2.85)	58.5	(3.48)	48.3	(3.44)	10.2	(4.89)	2.08
Hollands Midden	62.1	(2.80)	58.5	(3.34)	52.8	(3.45)	5.7	(4.80)	1.18
Kennemerland	63.6	(2.78)	55.9	(3.57)	54.4	(3.44)	1.5	(4.96)	0.30
Brabant-Zuid-Oost	62.8	(2.79)	53.4	(3.50)	53.5	(3.45)	-0.1	(4.91)	-0.03
Groningen	68.3	(2.69)	58.1	(3.71)	59.7	(3.40)	-1.6	(5.03)	-0.32
Limburg-Zuid	64.9	(2.76)	58.2	(3.67)	55.9	(3.44)	2.3	(5.03)	0.46
Gelderland-Midden	63.3	(2.78)	53.5	(3.75)	54.1	(3.44)	-0.6	(5.09)	-0.11
Zuid-Holland-Zuid	64.5	(2.76)	52.8	(3.39)	55.4	(3.44)	-2.6	(4.83)	-0.54
Twente	68.9	(2.67)	58.9	(3.64)	60.4	(3.39)	-1.5	(4.97)	-0.30
Noord- en Oost-Gelderland	64.0	(2.77)	57.4	(3.71)	54.9	(3.44)	2.5	(5.06)	0.50
Noord-Holland-Noord	64.2	(2.77)	54.4	(3.36)	55.1	(3.44)	-0.7	(4.81)	-0.14
Brabant-Noord	59.2	(2.84)	53.7	(3.79)	49.6	(3.44)	4.1	(5.12)	0.80
Gelderland-Zuid	65.7	(2.74)	54.6	(3.63)	56.8	(3.43)	-2.2	(4.99)	-0.43
Fryslân	65.3	(2.75)	54.7	(3.92)	56.3	(3.43)	-1.6	(5.21)	-0.31
IJsselland	61.4	(2.81)	54.5	(3.73)	52.0	(3.45)	2.5	(5.08)	0.49
Zaanstreek-Waterland	66.4	(2.73)	55.0	(3.85)	57.5	(3.42)	-2.5	(5.15)	-0.49
Gooi en Vechtstreek	64.3	(2.77)	61.5	(3.22)	55.2	(3.44)	6.3	(4.71)	1.34
Limburg-Noord	68.1	(2.69)	59.7	(3.58)	59.5	(3.40)	0.2	(4.94)	0.04
Flevoland	69.8	(2.65)	55.3	(3.15)	61.4	(3.37)	-6.1	(4.62)	-1.33
Drenthe	67.8	(2.70)	64.4	(3.73)	59.1	(3.41)	5.3	(5.05)	1.04
Zeeland	68.1	(2.69)	55.6	(3.74)	59.5	(3.40)	-3.9	(5.06)	-0.77

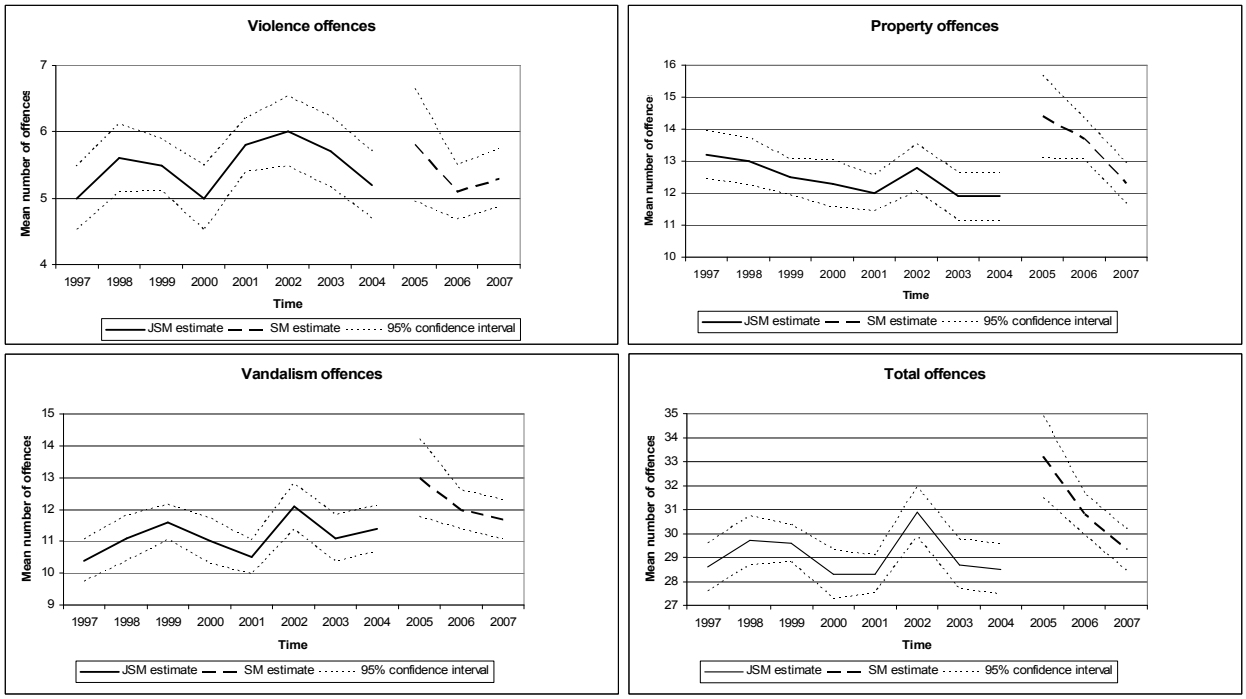

Figure 2. Mean number of offences against Dutch inhabitants (hundreds) during the 12 months prior to the interview, observed with the JSM (1997-2004) and the SM (2005-2007)

time according to (15). The annual net sample size of the JSM was about 10,000 respondents, while the number of respondents in the SM equals 5,200 in 2005, 22,000 in 2006, and 19,000 in 2007. To account for these large fluctuations in the annual sample size, it is assumed that the variance of the measurement equations is proportional to the sample size, i.e. $Var(\varepsilon_{t,j}) = \sigma^2_{\varepsilon,j} \big/ n_t$ where n_t denotes the number of respondents observed in the survey at time t.

With the Kalman filter, smoothed estimates for the trend parameters and the regression coefficients β_j are obtained using the fixed-interval smoother, see Harvey (1989) or Durbin and Koopman (2001) for details. The analysis was conducted with software developed in Ox in combination with the subroutines of SsfPack (beta 3.0) (Doornik 1998 and Koopman et al. 1999).

In Table 4 the results of six different analyses are summarised. Three of them are based on the data available up to and including 2006. The other three analyses are based on the complete series, including 2007. Comparison of the results obtained under equivalent models illustrates the size of the revision if an additional year becomes available to estimate the discontinuity induced by the redesign. The sizes of the revisions are substantial. The advantage is that the discontinuities are quantified more accurately if additional information becomes available. A concomitant drawback is that the estimated discontinuities three years after redesigning the survey are still subject to major revisions.

The analysis results in Table 4 also illustrate the effect of restricting the regression coefficients of the intervention variable of violence, property, and vandalism offences to add up to the regression coefficient of the intervention variable for the total offences. If the regression coefficients are not subject to this restriction, then (15) is replaced by the identity matrix. Under the unrestricted analysis, the consistency between the four series is seriously disturbed. It also follows that the restriction improves the precision of the regression estimates, particularly the estimate for the total offences. This can be explained since the restriction adds additional information to the model.

The estimated regression coefficients for property, vandalism and total offences are significantly different from zero. In the series of violence offences, however, no significant discontinuity can be established. To select the most parsimonious model, the intervention variable for violence offences is dropped, while the intervention variables for the property, vandalism and total offences still obey the consistency restriction. The estimated discontinuities obtained under this model, using the data up to 2007, i.e. the estimates in the final two columns of Table 4, are used to adjust the series. To eliminate the estimated discontinuity from the series, the estimates obtained with the SM in 2005 through 2007 can be corrected by $\tilde{y}_{t,j} = \hat{y}_{t,j} - \hat{\beta}_j$ to make them comparable with the outcomes under the design of the JSM. These corrected series are given in Figure 3. No adjustment is applied to the series of violence offences. As an alternative, the estimates obtained with the JSM can be corrected by $\tilde{y}_{t,j} = \hat{y}_{t,j} + \hat{\beta}_j$ to make them comparable with the outcomes under the SM.

Table 4: Smoothed Kalman filter estimate of the regression coefficient of the intervention variable (standard errors in brackets)

	estimate of intervention variable coefficient											
	2006 - four series				2007 - four series				2006 - three series		2007 - three series	
Variable	unrestricted		restricted		unrestricted		restricted		restricted		restricted	
Violence	0.44	(0.67)	0.52	(0.69)	0.14	(0.56)	-0.15	(0.51)	-	-	-	-
Property	2.21	(0.72)	2.43	(0.74)	3.17	(0.77)	3.49	(0.73)	2.52	(0.72)	3.40	(0.83)
Vandalism	1.22	(0.76)	1.69	(0.75)	1.33	(0.78)	1.51	(0.66)	1.54	(0.74)	1.26	(0.73)
Total	3.17	(1.32)	4.64	(0.97)	5.14	(1.57)	4.85	(0.94)	4.06	(0.98)	4.66	(0.90)

There is also a case to keep the intervention for the violence offences in the model. The best estimate for the discontinuity in this series may not be zero and if there is a real small discontinuity, although not significant, including it in the model may give a more satisfactory adjustment for the group of related series.

Most parameters about crime victimization that originate from the JSM showed an increase in the mean number of offences due to the transition to the SM. It appears that differences in the context of the survey and the questionnaire of the JSM and the SM are important explanations, even though these factors are confounded with other changes in the survey redesign. The questions about crime victimization in the JSM follow directly after a block of general questions about living conditions. In the questionnaire of the SM these questions are preceded by questions about feelings of insecurity, police performance and neighbourhood problems. Gibson et al. (1978) and Kalton and Schuman (1982) describe an experiment conducted in the National Crime Survey (USA) where a comparable modification in the questionnaires is tested. In this experiment the standard questionnaire about crime victimization is compared with an alternative questionnaire where questions about safety of the neighbourhood and police performance precede the questions about crime victimization. In this experiment a similar increase in the mean number of reported offences is observed due to the addition of a block of related questions. Adding a block of questions that are related to the retrospective crime victimization questions clearly affects the memory of respondents. It is unclear whether this results in fewer omissions or in an increase of telescoping errors.

8 General considerations and guidelines

Some of the range of issues which need to be considered when making a change to a long-running survey are discussed and illustrated with examples in this paper. In this section we discuss the practices which need to be followed to help ensure a smooth transition, and set out some general guidelines.

Quantifying the effect of a survey redesign is essential to avoid the confounding of real developments with the systematic effect induced by the redesign on the series of official statistics. Depending on the type of change and its place in the survey process, these differences can be quantified through an experiment where data under the new approach are collected from a separate probability sample, or by applying the new methods to the existing sample – or a subsample from it – to complete it with additional data, and then using standard methods from sampling theory like domain estimation or two-phase sampling. An intervention analysis through time series modelling can be considered as a second best alternative where there are insufficient resources to conduct an experiment.

Clear communication with the main users during the entire process of redesigning a survey is essential for the acceptance of a redesign. Users should be informed about plans for redesigning the survey and the possible consequences of discontinuities in the series. This should receive sufficient attention, since users are generally not (survey) statisticians and are mostly unaware of the sensitivity of survey estimates to changes in the design parameters of the underlying survey process. They should be involved in the experimental design stage where it is decided which differences should be observed in the experiment and which effects should be quantified. It is important that they have realistic expectations about the conclusions that can be drawn from the experiment. For example, the consequence of running the regular and new approach in parallel according to a two-treatment experiment is that the effects of all changes are confounded and that only the total effect of these changes is quantified. Users often expect a precision that approaches the accuracy of the figures at the national level of the regular survey. This requires a subsample size for the experimental group which equals the sample size of the ongoing survey. Power calculations can be helpful to illustrate the trade-off between costs and precision. In some cases users might finance an increased sample if they require more detailed or precise information about possible discontinuities. In the National Travel Survey example of section 2.1, the Ministry of Transport and Public Works partially financed the additional costs to run both surveys in parallel to quantify the effects of the redesign.

A synthetic estimation procedure is proposed to adjust series for observed discontinuities. Accuracy measures for these adjusted series are based on variance approximations but do not account for the bias, which arises from model misspecification. A time series approach is probably the most natural way to deal with discontinuities. A time series approach utilizes information across many samples of repeated

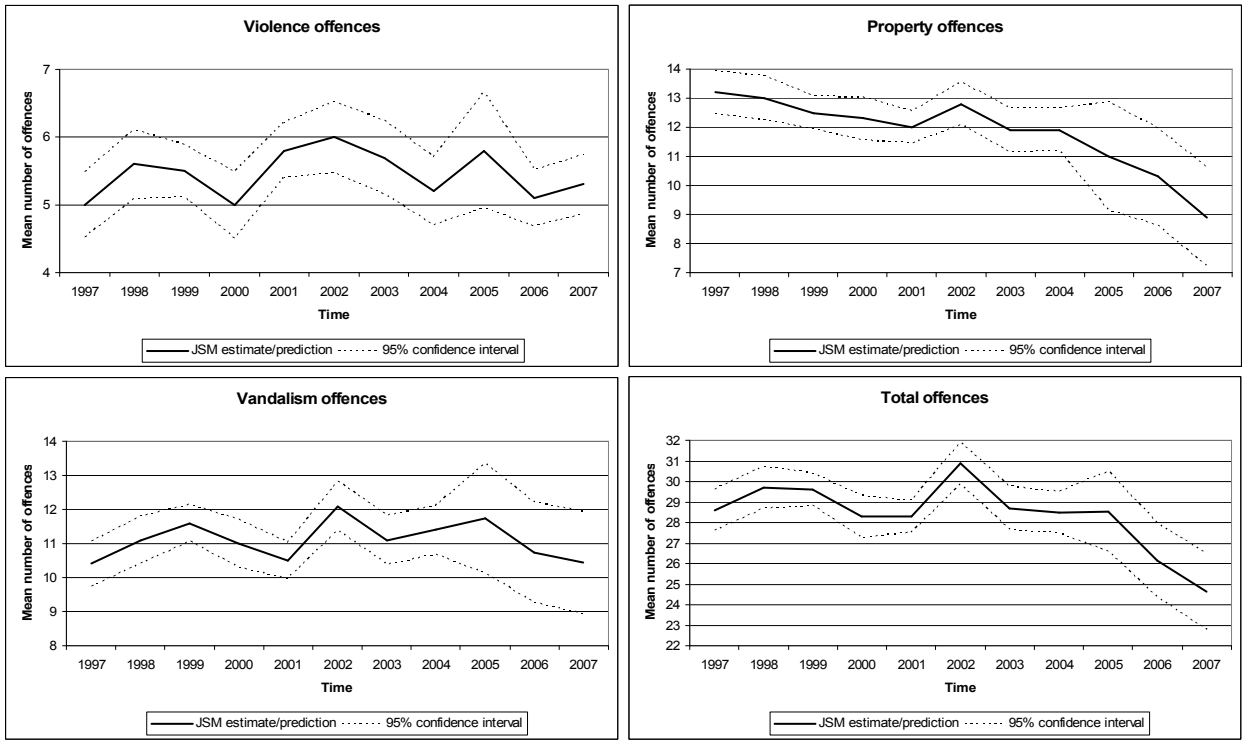

Figure 3. Mean number of offences against Dutch inhabitants (hundreds) during the 12 months prior to the interview, observed with the JSM (1997-2004) and the SM corrected for the estimated discontinuity (2005-2007)

surveys and is therefore appropriate to join series together. If available, auxiliary time series can be used to improve the model estimates for the discontinuity. Estimates are refined as more post-data become available, so a revision policy may be required.

Adjusting series according to synthetic estimation procedures or intervention analysis based on time series modelling is an appropriate tool for users to obtain adjusted estimates for the target parameters, for policy evaluation for example. Generally, national statistical institutes are rather reserved in the application of model-based estimation procedures for the production of official statistics. There is, however, a case for having an official series (with appropriate quality descriptions) rather than allowing each user to generate a slightly different version of the series for their own use.

From the foregoing discussion we can set out some general guidelines for making the quality of transitions in continuing surveys as high as possible, corresponding with the steps described in detail in the paper.

- Set up an appropriate mechanism for producing continuous series
  Once a potential for discontinuities has been identified, a strategy for producing a continuous series is needed. The best approach will depend on the particular situation of the survey change, but a variety of possibilities are described within this paper.
- Document the important parts of the development
  They can be used later when more information is available to make better revisions, and so that they can add to the core of knowledge of such developments.
- Test (or pilot) new approaches to determine their impact
  A formal test using an appropriate experimental method will give a statistical framework for the interpretation of the results which is valuable when discussing with users of the statistics. Otherwise pilot information can be used to make a judgement call, but this means that the quality across the change will not be quantifiable.
- Make inferences of the effect
  The outcome of the test must be analysed to infer the size of the discontinuity. This is relatively straightforward, if an experimental approach has been adopted or if the sample of the regular survey, completed with additional data, is used to calculate the outcomes under the new method. In the situation where there is no overlap between the regular and the new approach, or where an experimental approach has not been adopted, it may be possible to make appropriate inferences through time series methods
- Implement the change
  Make the survey changes, estimate the differences, preferably with the help of a parallel run or by recalculation, and implement the agreed approach for a continuous series to produce the required outputs
- Publish separate documentation of the redesign including:

- reasons for redesigning the survey including a detailed description of the regular and new design;
- revised results;
- estimates of discontinuities (possibly itemised if due to several changes, although the experiment may not be sufficient to provide this information);
- a description of the methodology employed to investigate and quantify discontinuities (experimental design, minimum sample size requirements, sampling techniques used to estimate the target parameters under the new approach from the sample available for the regular survey), as well as the methodology used to correct for discontinuities or advice for users on how to deal with them;
- descriptive interpretations and explanations of which factors contribute to the observed differences.

Acknowledgements

The authors thank the referees and the editors for careful reading and giving contructive comments on former drafts of the paper. The views expressed in this paper are those of the authors and do not necessarily reflect the policy of Statistics Netherlands or the Office for National Statistics.

Appendix A: Sample size determination for two-treatment experiments

Let u denote the population mean of the target parameter of interest. Testing hypotheses about response bias in the estimates of a finite population parameter due to different survey approaches, implies the existence of measurement errors. Therefore a measurement error model is required to link systematic differences between finite population parameters observed under different survey approaches, see Van den Brakel and Renssen (2005). It is assumed that the data obtained under the regular approach are a realisation of the model $y_{iR} = u_i + \beta_R + e_{iR} \equiv u_{iR} + e_{iR}$ and that the data obtained under the new approach can be modelled as $y_{iN} = u_i + \beta_N + e_{iN} \equiv u_{iN} + e_{iN}$. Here y_{iR} and y_{iN} denote the observations obtained from sampling unit i assigned to the regular or the new approach respectively, u_i the true intrinsic value of the target parameter of sampling unit i, β_R and β_N the bias (or treatment effect) induced by the regular and new approach respectively, and e_{iR} and e_{iN} measurement errors with expectations zero. Let $u_R = u + \beta_R$ and $u_N = u + \beta_N$ denote the population parameters observed under a complete enumeration of the finite population under the regular and the new survey approach and σ_R and σ_N the corresponding standard deviations. It is required that a pre-specified difference of $\Delta = u_R - u_N$ results in a rejection of the null-hypothesis of no treatment effect, i.e. $H_0 : u_R = u_N$, against an unspecified alternative that $H_1 : u_R \neq u_N$. Furthermore n_R and n_N denote the subsample sizes assigned to the regular and new surveys respectively. Finally α denotes the required significance level of the test and $(1 - \beta)$ the power. This implies that the probability that the null hypothesis is rejected if $u_R = u_N$ may not exceed α, and the probability that the null hypothesis is accepted given that $u_R \neq u_N$ may not exceed β. The sample sizes of the field experiments that are considered in this paper are generally sufficiently large to use a standard normal distribution to approximate the t-statistic to test the hypothesis of no treatment effects. In most practical situations there is no information about σ_N. Therefore it is assumed in general terms that σ_N is proportional to σ_R, i.e. $\sigma_N = k\sigma_R$. In most cases, k will typically be equal to one. If the new method, for example, is expected to reduce variance, then k might be taken smaller than one.

First consider an experiment where the subsample size of the regular survey is fixed in advance, since this subsample is used for the regular survey publication and must meet prespecified precision requirements. In this case the minimum sample size for the subsample assigned to the experimental group equals

$$\begin{aligned} n_N &= \frac{k^2\, n_R\, \hat{\sigma}_R^2 (Z_{(1-\alpha/2)} + Z_{(1-\beta)})^2}{\Delta^2 n_R - \hat{\sigma}_R^2 (Z_{(1-\alpha/2)} + Z_{(1-\beta)})^2} \quad , \\ k &= \sigma_N / \sigma_R \end{aligned} \tag{16}$$

where Z_γ denotes the γ -th percentile point of the standard normal distribution and $\hat{\sigma}_R$ is an estimator for the standard deviation under the regular survey.

Second consider an experiment where the sample size of the regular and the experimental groups are unknown, but there is a decision about the ratio between the subsample sizes of the regular and the experimental group, i.e. it is known that $n_N/n_R = f$. In this case, the minimum sample size can be determined as

$$\begin{aligned} n_R &= \frac{(k^2 + f)}{f} \frac{\hat{\sigma}_R^2 (Z_{(1-\alpha/2)} + Z_{(1-\beta)})^2}{\Delta^2} \quad . \\ n_N &= f n_R \end{aligned} \tag{17}$$

In the case of a specified alternative hypothesis, i.e. $u_R > u_N$ or $u_R < u_N$, $Z_{(1-\alpha/2)}$ is replaced by $Z_{(1-\alpha)}$ in (16) and (17).

Appendix B: Variance approximation of adjusted series

In this appendix variance approximations are given for series that are adjusted according to model (1), (2) and (5) and their corresponding models for subpopulation parameters (6), (7) and (8) described in section 6.1. The variance approximations are predominantly design-based, which implies that they are derived under the concept of repeatedly drawing samples with the finite population values held fixed. It is assumed that the survey estimates are based on repeated cross-sectional surveys. As a result the sample estimates for different time periods are uncorrelated, since they are based on independent samples. The approximations in this appendix do

not apply to panel designs, because they require appropriate correlations between sample estimates for different periods.

The variance approximations assume that the difference between the regular and the new approach is estimated from two separated probability samples, through an embedded experiment or parallel run. These subsample estimates are correlated, since they are based on samples drawn without replacement from a finite population. In general, these correlations are negligible since they are of order $1/N$, while the variances are of order $1/n$, where N and n denote the population and sample size respectively. Van den Brakel and Renssen (2005) and Van den Brakel (2008) describe conditions when the covariance terms in the variance of contrasts between sample estimates in embedded experiments vanish. If, however, the same sample is used to obtain parameter estimates under both the regular and the new approach (see section 3), then the correlation between these estimates cannot be neglected.

Since model (1) is linear and sample estimates for different time periods are based on separate independent samples, the variance is given by:

$$\hat{Var}(\tilde{y}_{N,t}) = \hat{Var}(\hat{y}_{R,t}) + \hat{Var}(\hat{\Delta}_T),$$

where $\hat{Var}(\hat{y}_{R,t})$ is the design variance of $\hat{y}_{R,t}$. Assuming that $\hat{y}_{N,T}$ and $\hat{y}_{R,T}$ are uncorrelated, it follows that $\hat{Var}(\hat{\Delta}_T) = \hat{Var}(\hat{y}_{R,T}) + \hat{Var}(\hat{y}_{N,T})$, where $\hat{Var}(\hat{y}_{N,T})$ and $\hat{Var}(\hat{y}_{R,T})$ are the design variances of $\hat{y}_{N,T}$ and $\hat{y}_{R,T}$ respectively. It is also possible to use formula (29) of Van den Brakel and Renssen (2005) to estimate the variance of the contrast between $\hat{y}_{N,T}$ and $\hat{y}_{R,T}$. In an equivalent way, the variance for (6) is given by

$$\hat{Var}(\tilde{y}^r_{N,t}) = \hat{Var}(\hat{y}^r_{R,t}) + \hat{Var}(\hat{\Delta}_T).$$

Models (2), (5), (7) and (8) are non-linear. Therefore, their variances are approximated by means of a Taylor linearization. To this end, (2) is expressed as a function of $(\hat{y}_{R,t}, \hat{y}_{R,T}, \hat{y}_{N,T})$ and linearised with a Taylor expansion about their real values in the finite population $(y_{R,t}, y_{R,T}, y_{N,T})$, truncated at the first order term. If sample estimates for different time periods are based on independent samples and if it is assumed that $\hat{y}_{N,T}$ and $\hat{y}_{R,T}$ are uncorrelated, then the variance for (2) can be approximated as:

$$\hat{Var}(\tilde{y}_{N,t}) \approx \left(\frac{\hat{y}_{N,T}}{\hat{y}_{R,T}}\right)^2 \hat{Var}(\hat{y}_{R,t}) + \left(\frac{\hat{y}_{R,t}}{\hat{y}_{R,T}}\right)^2 \hat{Var}(\hat{y}_{N,T}) + \left(\frac{\hat{y}_{N,T}}{\hat{y}_{R,T}}\frac{\hat{y}_{R,t}}{\hat{y}_{R,T}}\right)^2 \hat{Var}(\hat{y}_{R,T}).$$

In an equivalent way, an approximation to the variance for (7) is given by

$$\hat{Var}(\tilde{y}^r_{N,t}) \approx \left(\frac{\hat{y}_{N,T}}{\hat{y}_{R,T}}\right)^2 \hat{Var}(\hat{y}^r_{R,t}) + \left(\frac{\hat{y}^r_{R,t}}{\hat{y}_{R,T}}\right)^2 \hat{Var}(\hat{y}_{N,T}) + \left(\frac{\hat{y}_{N,T}}{\hat{y}_{R,T}}\frac{\hat{y}^r_{R,t}}{\hat{y}_{R,T}}\right)^2 \hat{Var}(\hat{y}_{R,T}).$$

If (5) is expressed as a function of $(\hat{y}_{R,t}, \hat{y}_{R,T}, \hat{y}_{N,T})$ and linearised around $(y_{R,t}, y_{R,T}, y_{N,T})$ by means of a first order Taylor linearization, then it can be shown that the variance of (5) can be approximated by

$$\hat{Var}(\tilde{y}_{N,t}) \approx \hat{c}_1^2\, \hat{Var}(\hat{y}_{R,t}) + \hat{c}_2^2 \hat{Var}(\hat{y}_{N,T}) + \hat{c}_3^2 \hat{Var}(\hat{y}_{R,T}), \quad (18)$$

where

$$\hat{c}_1 = 1 + \frac{\hat{\Delta}_T(1 - 2\hat{y}_{R,T})}{\hat{y}_{R,T}(1 - \hat{y}_{R,T})},$$

$$\hat{c}_2 = \frac{\hat{y}_{R,t}(1 - \hat{y}_{R,t})}{\hat{y}_{R,T}(1 - \hat{y}_{R,T})},$$

$$\hat{c}_3 = -\hat{y}_{R,t}(1-\hat{y}_{R,t})\left[\frac{\hat{y}_{N,T}(1 - 2\hat{y}_{R,T})}{\hat{y}^2_{R,T}(1 - \hat{y}_{R,T})^2} + \frac{1}{(1 - \hat{y}_{R,T})^2}\right].$$

Again it is assumed that $\hat{y}_{N,T}$ and $\hat{y}_{R,T}$ are uncorrelated and that sample estimates for different time periods are based on independent samples. An approximation for the variance of (8) is obtained in an equivalent way with

$$\hat{Var}(\tilde{y}^r_{N,t}) \approx (\hat{c}^r_1)^2\, \hat{Var}(\hat{y}^r_{R,t}) + (\hat{c}^r_2)^2 \hat{Var}(\hat{y}_{N,T}) + (\hat{c}^r_3)^2 \hat{Var}(\hat{y}_{R,T}),$$

where

$$\hat{c}^r_1 = 1 + \frac{\hat{\Delta}_T(1 - 2\hat{y}^r_{R,t})}{\hat{y}_{R,T}(1 - \hat{y}_{R,T})},$$

$$\hat{c}^r_2 = \frac{\hat{y}^r_{R,t}(1 - \hat{y}^r_{R,t})}{\hat{y}_{R,T}(1 - \hat{y}_{R,T})},$$

$$\hat{c}^r_3 = -\hat{y}^r_{R,t}(1-\hat{y}^r_{R,t})\left[\frac{\hat{y}_{N,T}(1 - 2\hat{y}_{R,T})}{\hat{y}^2_{R,T}(1 - \hat{y}_{R,T})^2} + \frac{1}{(1 - \hat{y}_{R,T})^2}\right].$$

References

Biemer, P., & Stokes, S. L. (1991). Approaches to the modeling of measurement error. In P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (chap. 24). Hoboken, New Jersey: Wiley.

Box, G. E. P., & Tiao, G. C. (1975). Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical Association*, *70*, 70-79.

Chambers, R., Weale, M., & Youll, R. (2000). The average earnings index. *The Economic Journal*, *110*, F100-F121.

Clogg, C. C., Rubin, D. B., Schenker, N., Schultz, B., & Weidman, L. (1991). Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression. *Journal of the American Statistical Association*, *86*, 68-78.

Cochran, W. G. (1977). *Sampling Techniques*. New York: Wiley & Sons.

De Leeuw, E. (2005). To mix or not to mix data collection modes in surveys. *Journal of Official Statistics*, *21*, 233-255.

Dillman, D. A., & Christian, L. M. (2005). Survey mode as a source of instability in responses across surveys. *Field Methods*, *17*, 30-52.

Doornik, J. A. (1998). *Object-oriented matrix programming using Ox 2.0*. London: Timberlake Consultants Press.

Durbin, J., & Koopman, S. J. (2001). *Time series analysis by state space methods*. Oxford: Oxford University Press.

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

Fienberg, S. E., & Tanur, J. M. (1987). Experimental and sampling structures: parallels diverging and meeting. *International Statistical Review*, *55*, 75-96.

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

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

Findley, D. F., Monsell, B. C., Bell, W. R., Otto, M. C., & Chen, B. C. (1998). New capabilities and methods of the X-12-ARIMA Seasonal Adjustment Program. *Journal of Business and Economic Statistics*, *16*, 127-176. (with Discussion)

Gibson, C. O., Shapiro, G. M., Murphy, L. R., & Stanko, G. J. (1978). Interaction of survey questions as it relates to interviewer-respondent bias. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, *1978*, 251-256.

Gomez, V., & Maravall, A. (2000). Automatic modeling methods for univariate series. In D. Pena, G. C. Tao, & R. S. Tsay (Eds.), *A course in time series* (chap. 7). New York: Wiley & Sons.

Groves, R. M. (1989). *Survey errors and survey costs*. New York: Wiley.

Hartley, H. O., & Rao, J. N. K. (1978). Estimation of nonsampling variance components in sample surveys. In N. Namboodiri (Ed.), *Survey sampling and measurement* (p. 35-43). New York: Academic Press.

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

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*, 65-70.

Kalton, G., & Schuman, H. (1982). The effect of the question on survey responses: a review. *Journal of the Royal Statistical Society, Series A*, *145*, 42-73.

Koopman, S. J., Shephard, N., & Doornik, J. A. (1999). Statistical algorithms for models in state space using SsfPack 2.2. *Econometrics Journal*, *2*, 113-166.

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

Montgomery, D. C. (2001). *Design and analysis of experiments* (5th ed.). New York: Wiley & Sons.

Rivière, P. (2002). What makes business statistics special? *International Statistical Review*, *70*, 145-159.

Robinson, G. K. (2000). *Practical strategies for experimenting*. New York: Wiley & Sons.

Särndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer Verlag.

Soroka, S. N., Wlezien, C., & McLean, I. (2006). Public expenditure in the UK: how measures matter. *Journal of the Royal Statistical Society*, *169*, 255-271. (Series A)

Statistics Netherlands. (2002). *Blaise developer's guide*. Heerlen: Statistics Netherlands. Available from `www.Blaise.com`

Sudman, S., Finn, A., & Lannom, L. (1984). The use of bounded recall procedures in single interviews. *Public Opinion Quarterly*, *48*, 520-524.

Van den Brakel, J. A. (2008). Design-based analysis of embedded experiments with applications in the Dutch Labour Force Survey. *Journal of the Royal Statistical Society, Series A*, *171*(3), 581-613.

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

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

Van den Brakel, J. A., & van Berkel, C. A. M. (2002). A design-based analysis procedure for twotreatment experiments embedded in sample surveys. *Journal of Official Statistics*, *18*, 217-231.