

Estimation of Discontinuities from the Introduction of Tablet Based Data Collection on the International Passenger Survey

Journal of Official Statistics

2024, Vol. 40(4) 748–782

© The Author(s) 2024

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/0282423X241275836

journals.sagepub.com/home/jof



Paul A. Smith¹ , Jan van den Brakel^{2,3}
and Giles Horsfield⁴

Abstract

The International Passenger Survey (IPS) is undertaken by the Office for National Statistics to measure tourism flows and tourist expenditure, and international migration. It is interviewer-administered, and the questionnaire instrument was changed in 2017 to 2018 from a paper questionnaire (completed by the interviewer) to an electronic questionnaire administered with a tablet. For operational reasons no parallel run was possible, but the new questionnaire was rolled out progressively to sampling locations. This phased introduction supported the estimation of the effects of the new questionnaire on the main outputs from the survey. We describe initial simulations designed to estimate the power of the phased introduction approach to detect important difference in the IPS outputs, and analyses of the survey estimates at different stages up to the end of 2019 using state-space models, to estimate the discontinuities in the survey outputs. We make an assessment of the effectiveness of the overall approach.

Keywords

discontinuity, state-space models, migration, tourist expenditure, border survey

¹S3RI and Department of Social Statistics & Demography, University of Southampton, Southampton, UK

²Quantitative Economics, Maastricht University, Maastricht, The Netherlands

³Department of Research & Development, Statistics Netherlands (CBS), Heerlen, The Netherlands

⁴Office for National Statistics, Newport, UK

Corresponding author:

Paul A. Smith, Department of Social Statistics & Demography, University of Southampton, Highfield, Southampton SO17 1BJ, UK.

Email: p.a.smith@soton.ac.uk



I. Introduction

Border surveys are one strategy for gathering information on tourism and visits to a country (see Rideng and Christensen 2004, and Frenç 2016 for summaries of border survey methodologies). The International Passenger Survey (IPS) is the United Kingdom (UK)'s border survey, and interviews people as they enter or leave the UK through ports, airports and the channel tunnel (the land border between Northern Ireland and Ireland is not covered). The sample design has two stages. In the first stage a stratified sample of time slots (from a population of morning and afternoon/evening slots at different airports/ports, covering 362 days of the year) is selected. Teams of interviewers work at the selected times and places, and there is a counting line, with interviews attempted with every k th traveler (for a suitable choice of k which may vary by stratum) crossing the line.

The IPS has been running since 1961, but the sample has been adjusted several times, most notably in 2009 when it was reallocated across a wider range of ports (ONS 2009). A detailed description of the history and methodology of the IPS is available in ONS (2014). The IPS has several purposes, including the measurement of tourism, the measurement of expenditures by visitors to the UK and by UK residents abroad, and the measurement of international migration. Migration filter shifts, designed to increase the number of migrants in the IPS sample, were introduced in 1980 to 1981. A very short questionnaire was used on these shifts, designed to identify migrants, who then received additional questions. Migration filter shifts were integrated with the main shifts in the 2009 redesign, but reintroduced in 2016 to increase the precision of migration estimates (White 2018). The IPS was suspended during the COVID-19 pandemic from March to December 2020, and has been replaced as a source of migration statistics by an approach based on administrative data and models (Rogers et al. 2021). A modified design was introduced in 2024. In this paper we focus on the changes to the IPS in 2017 to 2018, when migration was still a major topic of interest in the IPS.

The IPS interview is designed to be short and to be flexibly administered by interviewers to maximize response. Until 2017 data were collected on paper questionnaires (completed by the interviewer, except for some foreign language self-completion questionnaires (Pendry 2000)), and then mostly entered on-site using a laptop and a bespoke data capture tool programmed in Blaise (Statistics Netherlands 2002). The data could then be transmitted via a secure connection to a central database for further processing.

The ONS developed an electronic questionnaire for the IPS, to be administered using tablets. This was expected to have a range of benefits in the efficiency of the interviewers, and in the quality of the responses to the questions in the interview. We present these changes in more detail in Section 2 below.

Changing the data collection, and the associated changes in the questionnaire and editing procedures, had the potential to induce a discontinuity in the key IPS estimates for travel, tourism, and migration. Here a discontinuity is defined as a change in an estimate that results from a change in the collection approach and is not a change due to sampling variation or to a real evolution in the time series

(Van den Brakel et al. 2008). Such a discontinuity needs to be measured, controlled, and understood in order that the IPS time series before and after the change can be compared accurately. For example, if collecting the data on tablets results in an increase in expenditure by overseas visitors, perhaps because responders can view the questions in their own language, the increase due to the change in the data collection needs to be measured. It is not a real change in expenditure and when comparing the series before and after the change an adjustment is needed to produce valid estimates of the real period-to-period changes of the variables of interest.

The recommended approach to dealing with possible discontinuities in time series resulting from changes to field procedures involves an embedded experiment in the survey, starting with a small experiment run with the new approach alongside the standard survey procedure. This is often referred to as a parallel run. If there is no evidence of a major change from this, then the sample with the new approach can be extended, and finally the new method is rolled out with only a small part retaining the original method (Van den Brakel et al. 2008, 2021). However, this kind of experimental design within an existing survey can be operationally challenging because of the need for interviewers to use different questionnaires and procedures simultaneously. The ONS therefore decided to make the change by a gradual roll-out to the various interviewing locations, which made the transition operationally feasible, because different interviewer teams operate in each location. This presents less information than a situation with an embedded experiment, where the treatments (different modes) can be randomized at some level, or a parallel run, but still provides a way to estimate the parameters of the transition with a state-space model. This kind of approach to analyzing transitions without a parallel run has been considered by Van den Brakel et al. (2020).

The purpose of this paper is to describe how a new field work strategy can be implemented in an ongoing survey in a situation where there is no capacity for parallel data collection. The new design is gradually phased in. To estimate the discontinuities in the key variables of the survey, a time series modeling approach is applied, where the effect of the redesign on the outcomes is modeled with a generalized version of a level intervention. This is achieved with an auxiliary variable that gradually changes from zero, for the period before the start of the implementation, to one, after complete implementation of the new design. During the transition period, the auxiliary variable reflects the proportion of the variables covered by the new design. A major drawback of this approach is that there is no control over the precision of the discontinuity estimates and that the initial estimates directly after the change-over are unstable. To manage this additional risk, a simulation prior to the start of the change-over is conducted to assess with what precision discontinuities can be observed and how many observations under the new survey design are required before stable estimates for the discontinuities are obtained. Additional risks, like confounding of the discontinuity estimates with unexpected events like Brexit, are discussed.

The paper is structured as follows. Section 2 summarizes the development of the tablet-based data collection, and the additional procedures and functionality that it offered. Section 3 describes the idealized measurement of discontinuities, and the constraints operating in the IPS field work which led to the roll-out. It sets out a

state-space model for the evolution of the IPS series, and then uses the model in a simulation to assess the power of the rollout procedure to identify effects of given size. Section 4 describes the results of the implementation of the new questionnaire, and the estimated discontinuities as the time series developed. In Section 5 we discuss the findings and identify some lessons for the future.

2. Developing Tablet Data Collection for the IPS

The IPS has a number of different (but related) questionnaire instruments. There are arrivals and departures versions, relating to travelers entering or leaving the UK respectively, and trailers (additional survey instruments) for specific categories of travelers, including migrants, students, and employees. There are minor differences in the questionnaires to accommodate the different modes of travel (air, ferries, channel tunnel). For the purposes of the analyses in Sections 3 and 4 we assume that the mode of travel does not have an important effect separate from the direction of travel.

In order to move from paper to electronic capture, the questionnaire was redesigned to operate with the tablet screen (the early stages of this process are described in Benedikt (2015)). As part of this process the question wording and formatting was reviewed, to ensure it presented well on the screens. The routing was also considered; although the questionnaire is short, the different trailers make the routing quite complicated. One of the benefits of the change to an electronic questionnaire is that errors in routing were eliminated. The final design had one question per screen, and “there is evidence that respondents relate better to the ‘one-question-per-screen’ layout of the tablet, where they can see the questions in writing more easily themselves” (ONS 2018).

The new questionnaire allowed lookup tables for codes (e.g., for purpose of visit) to be automated; this was mainly a benefit for less experienced interviewers as experienced interviewers knew most codes automatically. The tablet questionnaire also implemented instant switching between languages in the display, based on the flag of the country as an indicator, which was much faster than the paper-based equivalent. This facilitated self-completion by travelers who did not speak English sufficiently well to undertake an interview. Some features could also be used to advantage, including the use of edit checks within the questionnaire (though many of these were treated as soft edits which could be overridden, in order to allow the interview to proceed quickly when necessary). The questionnaire could also be easily updated.

As a result of using the tablet questionnaire, a number of other changes to the survey process were also expected. The main changes include:

- the exercise of entering the data collected at the site would no longer be required. This process previously allowed for some quality assurance of the data close to the point of data collection;
- the post data collection editing (off-site by an editing team) was adjusted, because of the checks introduced in the questionnaire.

The new questionnaire was implemented in a limited pilot study before being included in the rollout. The pilot involved running several shifts at each of selected key survey sites over a two to three week period. This was sufficient to show qualitatively that the tablets were viable. The data collected suggested that the tablet questionnaire was better at capturing expenditure (largely because of the easier availability of questions in alternative languages in the tablet questionnaire), which would therefore be higher in the new mode.

There is a large body of literature on mode effects, see for example, Dillman and Christian (2005), de Leeuw (2005, 2008), Couper (2011), Dillman et al. (2014), and Schouten et al. (2022). Zhang et al. (2021) discuss and compare different onsite electronic survey data collection methods. Hassler et al. (2018) compared cost, completion times, and percent completion of electronic tablet to paper-based questionnaires administered onsite. Leisher (2014) compared tablet-based and paper-based survey data collection in terms of response rates, data collection costs, and completion time. Ravert et al. (2015) compared the equivalence of response obtained with paper based versus table-based questionnaires and reported only minor differences. Fanning and McAuley (2014) report non-significant effects between both modes in an experiment in a Health Survey questionnaire. Kusumoto et al. (2017) observed no difference in completion speed between paper and pencil and tablet surveys. Tourangeau et al. (2017) analyzed difference in measurement errors between surveys conducted on smartphones, tablets, and laptop devices. Although the results from these experiments cannot be generalized to the IPS we expect no large differences between the old and new approach, since both data collection approaches are based on interviewer administered data collection modes. Potential differences might be induced by the differences in the questionnaire.

Particularly in the context of mixed-mode surveys there is an increasing amount of literature on methods that attempt to correct and adjust for mode-effects. A regression modeling approach is proposed by Suzer-Gurtekin (2013). Imputation methods are proposed by Kolenikov and Kennedy (2014), while Vannieuwenhuyze (2014) and Klausch et al. (2017) propose adjustment methods based on re-interview designs. These methods focus on equalizing mode effects in mixed-mode designs and are not directly applicable in a change-over from one uni-mode design to another uni-mode design where the questionnaire changed simultaneously.

3. Planning for and Dealing with Changes in Survey Procedures

3.1. Randomization or Deterministic Transition

The best conditions for the estimation of a discontinuity are to use an embedded experiment with the new and old approaches as the treatments (Van den Brakel et al. 2021). This effectively gives a parallel run of the new and old methods, and allows the independent estimation of the discontinuity and the real evolution of the time series. There are several levels at which randomization of treatments could be

undertaken in the IPS. These levels and the expected sample sizes from a 10% treatment group are:

- (a) by interview, $n \approx 2,500$
- (b) by interviewer shifts, $n \approx 100$
- (c) by interviewers and flows (arrivals/departures), $n \approx 40$
- (d) by interviewers, $n \approx 20$
- (e) by site, $n \approx 5$.

The effective sample size for (b) to (e) would however be reduced by the clustering of observations within the experimental units, and therefore the power of these designs to detect a difference would generally be low. ONS's assessment of the operational considerations in introducing the changes to the IPS was that the randomization of cases, interviewers, or shifts would introduce too much disruption to the fieldwork and therefore risk the quality of the outputs. There was also a requirement to progressively roll-out training for interviewers that made a staged transition team by team (where a team of interviewers may cover a single site or a group of sites) the most practical implementation approach. The inflow and outflow questionnaires have some differences, so interview teams were trained first on the new outflow questionnaire, which was implemented first, then later on the inflow questionnaire. Therefore the rollout patterns are different on the two sets of variables. This meant that a parallel collection on both methods would not be available, and methods based on the availability of parallel run data (e.g., Van den Brakel et al. 2021) were effectively ruled out.

Without a parallel run the estimate of the discontinuity is confounded with the normal evolution of the time series, but by making assumptions about that evolution, the discontinuity can be estimated (Van den Brakel et al. 2008, 2020). The staged transition across IPS sites gives an increasing coverage of the main IPS variables by the tablet questionnaire (the variable abbreviations used here and their definitions are given in Table 1). The proportion of each variable which is moved to tablet data collection at each stage of the process is shown in Table 2; these are different for different variables because the characteristics of passengers vary by airport.

Before the change to the tablet questionnaire was implemented, an indication of the power of the analysis to detect a discontinuity in the different series was required, as part of the communication with users about the expected effects of the transition. The power can be assessed by simulation using a suitable model for the evolution of the monthly IPS estimates (Van den Brakel et al. 2020). So we first present a model for the IPS variables in Subsection 3.2, then return to the assessment of power in Subsection 3.3.

3.2. *The Model*

The monthly estimates of the IPS output variables were used to develop a structural time series model. The model used to represent the evolution of the IPS series y_t is

Table 1. IPS Variable Abbreviations and Their Definitions.

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

Table 2. Estimated Proportions of IPS Variables Covered by the Staged Transition to the Tablet Questionnaire.

Inflow variables							
Date	svisukres	sexpukres	smigosar	sflowarr	sflowarrn	narr	ncases
1/11/2017	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1/12/2017	0.512	0.476	0.354	0.477	0.475	0.393	0.385
1/1/2018	0.512	0.476	0.354	0.477	0.475	0.393	0.385
1/2/2018	0.679	0.591	0.504	0.663	0.663	0.552	0.524
1/3/2018	0.853	0.773	0.633	0.795	0.792	0.677	0.670
1/4/2018	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Outflow variables							
Date	svisosres	sexposres	smigukdep	sflowdep	sflowdepn	ndep	ncases
1/8/2017	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1/9/2017	0.032	0.025	0.029	0.043	0.043	0.047	0.044
1/10/2017	0.151	0.112	0.092	0.131	0.132	0.113	0.113
1/11/2017	0.227	0.177	0.177	0.239	0.241	0.208	0.223
1/12/2017	0.409	0.362	0.332	0.478	0.474	0.378	0.385
1/1/2018	0.409	0.362	0.332	0.478	0.474	0.378	0.385
1/2/2018	0.688	0.522	0.476	0.668	0.666	0.502	0.524
1/3/2018	0.763	0.593	0.628	0.801	0.798	0.665	0.670
1/4/2018	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Note. The dates reflect the actual rollout pattern and the patterns are different for different variables because the characteristics of passengers vary by airport. Each month is shown, although there were no changes to the patterns between December and January. narr and ndep are proportions by the number of arrivals and departures respectively, and ncases gives the proportions by the number of sample cases. Other variable names are defined in Table 1.

$$y_t = L_t + S_t + \beta x_t + \varepsilon_t \quad (1)$$

where L_t is a trend component which depends on the previous level of the trend and the previous difference R_{t-1} , which in turn is modeled as a random walk, that is,

$$\begin{aligned} L_t &= L_{t-1} + R_{t-1}, \\ R_t &= R_{t-1} + \eta_t, \\ E(\eta_t) &= 0, \quad \text{Cov}(\eta_t, \eta_{t'}) = \begin{cases} \sigma_\eta^2 & \text{if } t = t' \\ 0 & \text{if } t \neq t'. \end{cases} \end{aligned}$$

Note that the level of the trend can implicitly contain a deterministic level, that is, $L_t = L_{t-1} + R_{t-1} + \mu$, with μ an intercept. This trend model is a quadratic trend that involves second order differencing, since $\eta_t = L_{t+1} - 2L_t + L_{t-1}$. Deterministic features, like an intercept or a linear trend, which might be present in the data, cannot be identified due to the implicit second order differencing. These features are, however, preserved by the Kalman filter and smoothing algorithms. Furthermore, S_t is a trigonometric seasonal component:

$$S_t = \sum_{l=1}^6 S_{l,t},$$

with

$$\begin{aligned} S_{l,t} &= S_{l,t-1} \cos(h_l) + S_{l,t-1}^* \sin(h_l) + \omega_{l,t}, \\ S_{l,t}^* &= S_{l,t-1}^* \cos(h_l) - S_{l,t-1} \sin(h_l) + \omega_{l,t}^*, \quad l = 1, \dots, 6, \\ h_l &= \frac{\pi l}{6}, \quad l = 1, \dots, 6, \\ E(\omega_{l,t}) &= E(\omega_{l,t}^*) = 0, \\ \text{Cov}(\omega_{l,t}, \omega_{l',t'}) &= \text{Cov}(\omega_{l,t}^*, \omega_{l',t'}^*) = \begin{cases} \sigma_\omega^2 & \text{if } l = l' \text{ and } t = t' \\ 0 & \text{if } l \neq l' \text{ or } t \neq t', \end{cases} \\ \text{Cov}(\omega_{l,t}, \omega_{l,t}^*) &= 0 \quad \forall l \text{ and } t, \end{aligned}$$

and ε_t is the measurement error, modeled as a white noise:

$$E(\varepsilon_t) = 0, \quad \text{Cov}(\varepsilon_t, \varepsilon_{t'}) = \begin{cases} \sigma_\varepsilon^2 & \text{if } t = t' \\ 0 & \text{if } t \neq t'. \end{cases}$$

Finally, x_t represents an indicator explanatory variable which takes the value 0 before the discontinuity is introduced, then an increasing positive value (as in Table 2) as the rollout progresses, and then the value 1 once rollout is completed and for all subsequent periods. It can be interpreted as a generalized version of a level intervention variable. The fitted parameter β will therefore contain the estimate of the discontinuity.

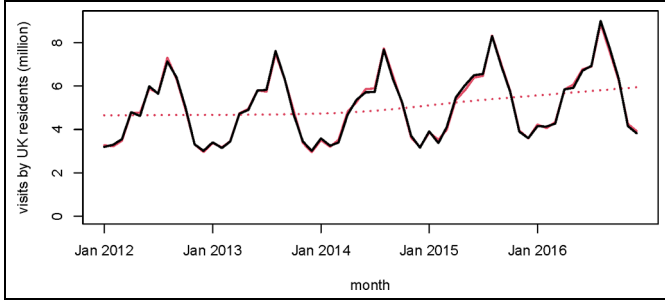


Figure 1. Original series of estimates of visits abroad by UK residents *visukres* (black), the Kalman smoother estimates of the signal from the state-space model (solid red), and the trend from the model (dotted red).

For the trend, the so-called smooth trend model is chosen. This is a popular trend model in econometric time series modeling (Durbin and Koopman 2012, Ch.3). See Subsection 3.4 for a more extended motivation and a comparison with the local level trend model. Also the trigonometric seasonal is a standard model in econometric time series modeling (Durbin and Koopman 2012, Ch.3) and is an appropriate specification to model a seasonal pattern as visible in Figure 1. The level intervention approach to estimate the effect of an intervention was originally proposed by Harvey and Durbin (1986) to estimate the effect of seatbelt legislation on road casualties. Van den Brakel et al. (2008, 2020, 2022) and Van den Brakel & Roels (2010) used this approach to estimate discontinuities in repeated sample surveys induced by a redesign of the survey process. A recent similar application is for example, Hungnes et al. (2024).

An alternative approach is to construct time series for each airport separately, combine them in one multivariate model and model the effect of the transition in each series with a level intervention. Let $\hat{y}_t^{(i)}$ denote the population estimate for the entire UK for the particular series based on observations from airport $i = 1, \dots, n$. To obtain meaningful input series, a rescaling of the observations from each airport to the same level, that is, national level, is required. Then a multivariate model for the discontinuity at the national level could be defined as

$$\begin{pmatrix} \hat{y}_t^{(1)} \\ \vdots \\ \hat{y}_t^{(n)} \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} (L_t + S_t + I_t) + \begin{pmatrix} \lambda_t^{(1)} \\ \vdots \\ \lambda_t^{(n)} \end{pmatrix} + \begin{pmatrix} \delta_t^{(1)} \\ \vdots \\ \delta_t^{(n)} \end{pmatrix} \beta + \begin{pmatrix} e_t^{(1)} \\ \vdots \\ e_t^{(n)} \end{pmatrix}. \quad (2)$$

Here L_t , S_t , and I_t denote the trend, seasonal, and population white noise for the variable of interest at the national level. Deviations of the series of each airport from the national level are modeled with random walks $\lambda_t^{(i)} = \lambda_{t-1}^{(i)} + \xi_t^{(i)}$, for $i = 1, \dots, n$. To identify this component, the variables must obey the restriction

$$\sum_{i=1}^n \lambda_t^{(i)} = 0. \quad (3)$$

The $\delta_t^{(i)}$ are dummy indicators that switch from zero to one at the moment that airport i changes to the new tablets for data collection. Similar to Equation (1) β is the regression coefficient that can be interpreted as an approximation of the discontinuity for the variable at the national level. Finally $e_t^{(i)}$ are the sampling errors of the input series which are normally and independently distributed, that is, $e_t^{(i)} \sim N(0, \text{var}(\hat{y}_t^{(i)}))$, where $\text{var}(\hat{y}_t^{(i)})$ is estimated from the sample data. We were unable to estimate this model because the series for the individual airports were not available. Note that the model in Equation (2) is very similar to the multivariate state space model proposed by Van den Brakel and Krieg (2015) and Hungnes et al. (2024) for modeling discontinuities in a rotating panel design. The question is whether this model reliably estimates the discontinuities because there are large differences in passenger flows between the different airports. A large airport like Heathrow has totally different flows from, say, a small airport like Bristol.

To avoid the rescaling of the input series to the national level, a Seemingly Unrelated Time Series Equation (SUTSE) model can be considered. In this case $\hat{y}_t^{(i)}$ denote the observations from airport $i = 1, \dots, n$. Each series has its own trend, seasonal, and population white noise component. The vector with random walks $\lambda_t^{(i)}$ that models differences between airports is unnecessary. Furthermore each series has its own level intervention $\delta_t^{(i)}\beta^{(i)}$ and its own sampling error. An estimate for the discontinuity at the national level is obtained as the sum over the discontinuities of the separate airports, that is, $\beta = \sum_{i=1}^n \beta^{(i)}$. See Harvey (1989, Ch. 8) for more details of SUTSE models.

The models are expressed in state-space form, and the Kalman filter is used to estimate the state variables (Durbin and Koopman 2012). The state space representation distinguishes between state variables and hyperparameters. The state variables define the trend (L_t and R_t), seasonal ($S_{t,l}$ and $S_{t,l}^*$) and regression coefficient of the level intervention (β). The hyperparameters define the dynamics of the processes for the state variables, which are the variance components of the state disturbance terms (σ_η^2 and σ_ω^2) and the variance of the measurement errors (σ_ε^2). The state space representations of models in Equations (1) and (2) are defined in Appendix A. The models were implemented in OxMetrics (Doornik 2009) in combination with SsfPack (Koopman et al. 2008).

The Kalman filter is a recursive algorithm that starts at the beginning of the series and provides for each period t optimal estimates and standard errors for the state variables based on the time series observed until period t . These are referred to as the Kalman filter estimates. The Kalman filter estimates for each period t , can be updated with the information that became available after period t . This procedure is called smoothing and is based on a recursive algorithm that starts with the last observation of the observed series and updates the filtered estimates, including their standard errors, for the state variables of all preceding periods. These are referred to as the Kalman smoother estimates. Under the assumption that the disturbance terms and the initial state vector are normally distributed, the Kalman filter provides optimal estimates in the sense that they minimize the mean squared

error. If the normality assumption doesn't hold, the Kalman filter is still an optimal estimator in the sense that it minimizes the mean squared error within the class of all linear estimators (Harvey 1989, Subsection 3.2). The stated normality assumption implies that the one-step-ahead prediction errors are normally and independently distributed. This is evaluated by testing the standardized one-step-ahead prediction errors for (1) heteroscedasticity using an F -test, (2) normality using a Bowman-Shenton test, and (3) autocorrelation using a Ljung-Box test (Durbin and Koopman 2012, Subsection 2.12).

To start the Kalman filter, initial values for the state variables as well as values for the hyperparameters are required. Equation (1) contains non-stationary state variables, which are initialized with a diffuse initialization. This implies that the initial values of all state variables are equal to zero with a diagonal covariance matrix with diagonal elements diverging to ∞ . The exact initial solution for the Kalman filter with diffuse initial conditions, proposed by Koopman (1997), is used. With this diffuse initialization of the Kalman filter the first d observations are required to construct a proper prior for the Kalman filter, where d equals the number of non-stationary state variables of the state space model. For this reason, the Kalman filter estimates for the first d time periods are ignored in the analysis and also in the model evaluation of the one-step-ahead prediction errors.

The values of the hyperparameters are also unknown. They are estimated by means of maximum likelihood. The likelihood function is obtained by the so-called prediction-error decomposition (Harvey 1989, Subsection 3.4). The likelihood function is optimized by repeatedly running the Kalman filter in a numerical optimization procedure using MaxBFGS (Doornik 2009). Since the hyperparameters are variances, which cannot take negative values, they are estimated on the log-scale. The starting values for the hyperparameters in the optimization procedure are equal to $\ln(0.1) + 0.5\ln(\hat{\sigma}^2)$, with $\hat{\sigma}^2 = 1/(T-d) \sum_{t=d+1}^T (v_t^2/\text{var}(v_t))$ and v_t the one-step-ahead prediction errors obtained by evaluating the likelihood function with hyperparameters taken equal to $\ln(0.1)$ (Koopman et al. 2008, Ch.5). This generally results in reasonably good starting values. To minimize the risk of finding a local maximum, the optimization procedure is also started with different starting values. Under the model in Equation (1) the MaxBFGS always converged to the same maximum likelihood estimates for the hyperparameters. The unknown values of the hyperparameters of the state-space model are replaced by their maximum likelihood estimates in the Kalman filter. The standard errors of the Kalman filter estimates for the state variables do not reflect the additional uncertainty of replacing the true values of the hyperparameters by their maximum likelihood estimates, which is the common approach in state-space time series analysis.

We expected that the IPS, which is not designed to produce estimates for single months for most variables would be rather volatile, but the models were surprisingly well-behaved. Figure 1 shows an example from the modeling of the number of trips by UK residents, measured on the passenger inflow, *svisukres*. The black line shows the original data, and the red solid line the Kalman smoother estimates for the trend and seasonal of the model, Equation (1), with $\beta=0$ because there is no discontinuity in the original series. The observed and fitted series are almost

coincident over most of the plot (although there are of course small differences). The fitted trend component is relatively smooth. This suggests that the model has a good a chance to detect a discontinuity if one is present, but we still need to account for the variability in the model in making an assessment.

3.3. Power Assessment

The basic strategy then is to take a period of the IPS equal in length to the proposed rollout, to assume a level of discontinuity (consistently across sites within a flow), and to introduce this discontinuity to the series according to the pattern of the rollout. This creates an adjusted series which is used as the input to a model which includes the rollout pattern x_t , and an estimate is made of the size of the discontinuity and its variance. We expect early estimates to be far from the truth (as early in rollout few ports will be using the tablet questionnaire and there is little information on which to base an estimate of the discontinuity), but to converge to a more stable estimate as further information on the size of estimates with the tablet collection accumulates. Even beyond the rollout period, additional information about the size of the discontinuity is obtained as the parameters of the model are affected by new observations. This allows us to assess the size of discontinuity which is likely to be detectable (i.e., the power of such an analysis) and the time required to obtain a stable estimate for the discontinuity.

We examine one variable, the number of visits by overseas residents, *svisosres* (which is measured on the outflow), in detail to demonstrate the approach that has been followed.

The first step is to introduce a discontinuity (as a percentage of the mean of the series from January 2012 until December 2016) into the existing series. Figure 2 shows the original series and the discontinuity, which is phased in over eight months in accordance with the rollout plan on the outflow. We show the actual periods of the data on the x axis, though the real time periods are not important for this simulation. The new (red) series from Figure 2 now contains the discontinuity, but also the sampling error from the IPS in the periods used, which is expected to obscure the discontinuity. Equation (1) is fitted to this new series, now with β free and to be estimated. The resulting series of estimates of the discontinuity is shown in Figure 3. The black line is the real value of the simulated discontinuity during and after the phase in period, that is, βx_t . The real value of the discontinuity, that is, β , is the level of the horizontal black line after October 2015. The red line shows the Kalman filter estimates for β . They are based on the information observed at the different points in time and show how the filtered estimates are updated if a new observation becomes available and how it finally converge to a stable estimate.

The first thirty-six months of the series are not shown in Figure 3, as no discontinuity is expected by the model so nothing happens—although this period does allow the parameters of the other components of the Kalman filter to stabilize from their starting values. The initial erratic behavior of the discontinuity estimate is clearly seen, and some of the early estimates of the discontinuity do not contain the

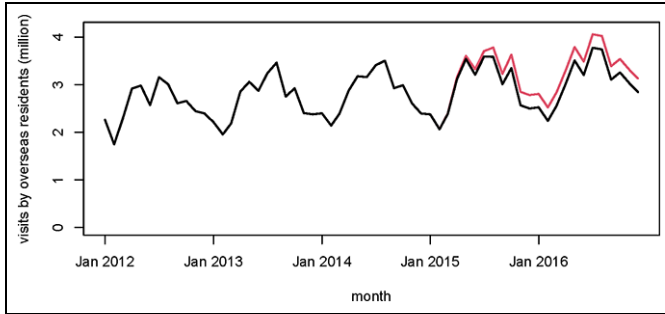


Figure 2. Original (old) series of estimates of the number of visits to the UK by overseas residents *svisosres* (black), and the new series (red) after a 10% discontinuity has been phased in over eight months from March 2015.

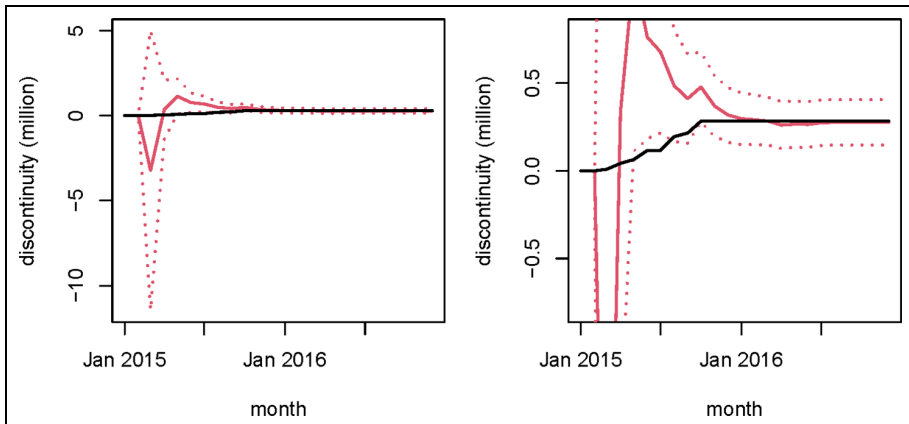


Figure 3. In black the “real” discontinuity in the series during and after completion of the phase-in period (i.e., βx_t) for a 10% discontinuity introduced to the number of visits to the UK by overseas residents *svisosres* beginning in March 2015. In red the Kalman filter estimate for β during and after completion of the phase-in period. The dashed lines show the estimated 95% confidence interval for the estimated discontinuity. The left plot shows the whole range of the series and the right plot shows more detail as the estimated discontinuity converges.

“real” series within the confidence interval. By the time the rollout is completed in October 2015 (the eighth month of the rollout) the estimate is much improved and the confidence interval much narrower (although in this particular example the “real” change is only just in the interval at this stage). Around January 2016 the estimate has essentially converged on the correct value, although the confidence interval continues to get slightly smaller until September 2016.

This example shows the situation with a discontinuity of a particular size in a particular month. The same discontinuity appearing at a different point in the evolution of the series may have a different impact on the results, so in another experiment the same rollout pattern is applied starting from September 2014. This produces qualitatively the same pattern, although there are some differences of detail. In this case the estimated discontinuity does not stabilize until around July 2015 (the eleventh month after the rollout). We also used the same procedures with a -5% discontinuity, again introduced at two separate time points. Table 3 summarizes the effects of these procedures in these four example cases.

The evolution of the estimated standard errors are almost the same regardless of the size of the discontinuity or the month in which it is introduced. So for this variable we expect to have the power to detect a change of approx. 200,000 at the end of the roll-out; 150,000 four months after the end of roll-out; and 130,000 ten months after the end of roll-out, as follows from the standard errors reported in Table 3. This is just enough in this case to detect a 5% change.

The example of *svisosres* works quite well, but this is not always the case. Figure 4 shows *sflowarrn* with a $+5\%$ discontinuity. Here the estimated discontinuity does not converge toward the real (induced) one, but instead to a lower value which eventually leaves the true discontinuity more or less outside the confidence interval.

Table 4 gives an overall summary of the approximate minimum detectable effects at the end of roll-out, end of roll-out + four months, and end of roll-out + ten months for variables on both flows, based on the estimated standard errors from the simulations.

It can be seen that only the largest discontinuities in the migrant flows are expected to be detectable, and that expenditure differences less than 10% are not expected to be detectable. But on other person-based flows discontinuities of around 5% will generally be detectable.

3.4. Simulation with Different Trend Models

In this subsection the choice for a smooth trend model is motivated for L_t in Equation (1). Alternative model choices are the local linear trend model, which has a disturbance term for both the level (L_t) and the slope (R_t) and the local level trend model that assumes a random walk for the level without a slope parameter ($L_t = L_{t-1} + \zeta_t$, with ζ_t a normally and independently distributed disturbance term). The smooth trend model is well known in the econometric literature for its reasonable flexibility and parsimony (Durbin and Koopman 2012, Ch. 3). It results in more smooth trend estimates compared to the other two trend models.

A statistical argument for a choice between the smooth trend model and the local level model is the order of integration of the observed series. The smooth trend model assumes a second order of integration, $I(2)$, for the observed series while the local level model assumes a first order of integration, that is, $I(1)$. An Augmented Dickey-Fuller (ADF) test rejected the null hypothesis for all series that there is a second order unit root (all p -values smaller than 1%). This would motivate the choice for a local level model instead of a smooth trend model.

Table 3. Summary Statistics for Example Discontinuities of +10% and -5% Applied to the IPS *svsisres* Series at Two Different Times.

Discontinuity (%)	Start month	Estimate made after month	Value		Relative (%)		
			Actual discontinuity	Estimated discontinuity	Estimated standard error of discontinuity	Estimated discontinuity	1.96 × Estimated standard error of discontinuity
+10	33	40 (+8)	282,613	141,434	106,134	5.0	7.4
		44 (+12)	282,613	314,082	74,219	11.1	5.1
		50 (+18)	282,613	302,084	66,271	10.7	4.6
+10	39	46 (+8)	282,613	477,927	101,620	16.9	7.0
		50 (+12)	282,613	291,029	72,012	10.3	5.0
		56 (+18)	282,613	276,366	66,076	9.8	4.6
-5	33	40 (+8)	-141,307	-282,486	106,135	-10.0	7.4
		44 (+12)	-141,307	-109,837	74,220	-3.9	5.1
		50 (+18)	-141,307	-121,835	66,272	-4.3	4.6
-5	39	46 (+8)	-141,307	54,007	101,620	1.9	7.0
		50 (+12)	-141,307	-132,891	72,012	-4.7	5.0
		56 (+18)	-141,307	-147,554	66,076	-5.2	4.6

Note. Relative values are relative to the average of the series over all considered periods. Note that under "value" the SE is given, while under "relative" 1.96 × SE (half the width of the confidence interval) is given.

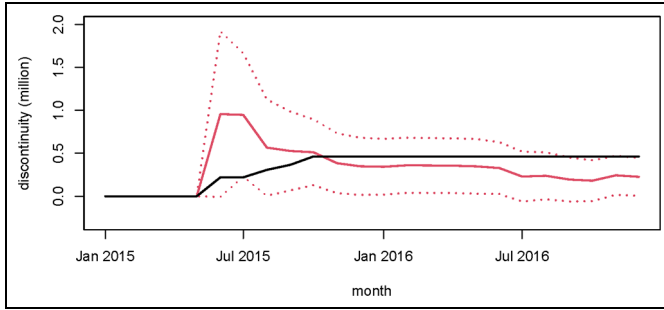


Figure 4. In black the “real” discontinuity in the series during and after completion of the phase-in period (i.e., βx_t) for a 5% discontinuity introduced to the arrival passenger flow excluding flows from the Channel Islands and Isle of Man *sflowarrn* beginning in June 2015. In red the Kalman filter estimate for β during and after completion of the phase-in period. The dashed lines show the estimated 95% confidence interval for the estimated discontinuity.

Applying a local level model to the IPS series results in more volatile trend estimates compared to the smooth trend model. This affects the estimates for the discontinuities. The smoother the trend the greater the influence of observations that are further away from the time of transition on the discontinuity estimates. With the local level trend model, more weight is given to the observations directly before and after the transition. With the smooth trend model observations further away from the transition also contribute to the discontinuity estimates.

Since the results for the discontinuity estimate depend on the choice for the trend model, a simulation was conducted. The setup is similar to the simulation discussed in Subsection 3.3. For the series observed between January 2012 and December 2016, a discontinuity of 10% of the level of the series in 2016 is added to the series, proportional to the roll-out in Table 2, assuming that the roll-out starts in January 2014. In a next step, the discontinuities are estimated with the model in Equation (1), once with a smooth trend model for L_t and once with a local level model for L_t . Results are presented in Table 5. Comparing the actual discontinuities with the estimated discontinuities shows that the estimates under the smooth trend model are much closer to the true values than the estimates under the local level model. Also the standard errors of the discontinuity estimates under the smooth trend model are much smaller compared to the local level model.

Finally the model assumptions of both state space models are tested by evaluating to what extent the one-step-ahead prediction errors meet the assumption that they are normally and independently distributed. To this end the following tests (see Durbin and Koopman 2012, Subsection 2.12 for an overview) are applied to the standardized innovations: (1) F -test for heteroscedasticity, (2) Bowman-Shenton test for normality (Bowman and Shenton 1975), and (3) Ljung-Box test (Ljung and Box 1978) for serial auto correlation up to lag 12. Results are included in Appendix B and indicate some deviation from the normality assumption for about half of the series. Based on these model diagnostics, however, there is no preference for one of the two trend models.

Table 4. Approximate Minimum Detectable Effects (Rounded to the Nearest ½% Below 10, or Nearest 1% Above It) for Analyzed Variables, at the End of the Roll-Out Period, and Four and Ten Months After It.

Inflow variables		Minimum detectable effect (%)		
		Months after end of rollout		
		0	4	10
Number of overseas visits by UK residents	svisukres	9	8	8
Expenditure abroad by UK residents in £	sexpukres	13	12	11
Overseas residents migrating to the UK	smigosar	37	29	28
Total arrival passenger flow	sflowarr	4½	4	3
Arrival passenger flow excl. Channel Islands & Isle of Man	sflowarrn	4½	4	3½
Outflow variables		Minimum detectable effect (%)		
		Months after end of rollout		
		0	4	10
Number of visits to the UK by overseas residents	svisosres	7	5	4½
Expenditure in the UK by overseas residents in £	sexposres	15	12	11
UK residents migrating abroad	smigukdep	29	20	19
Total departure passenger flow	sflowdep	5½	4½	4½
Departure passenger flow excl. Channel Islands & Isle of Man	sflowdepn	5½	4½	4½

Based on these considerations the smooth trend model is used in Equation (1). Even though the ADF test clearly rejects the null hypothesis that the series contain a second order unit root, the simulation clearly indicates that the smooth trend model is more appropriate for estimating discontinuities. In addition the Ljung-Box test statistics in Tables B1 to B3, provide comprehensive evidence that the chosen smooth trend model is adequate. If the data were I(1) and the smooth trend model was misspecified, the one-step-ahead prediction errors should be over-differenced and thus autocorrelated. This is, however, ruled out by the test results of the Ljung-Box test for almost all series. An interesting interpretation, proposed by a constructive reviewer, is that the structural break overshadows the persistent characteristics of the series. If for example, the break has a strong signal whereas the I(2) trend is comparably smooth, then the ADF test may reject even in the presence of an I(2) trend. This could be tested with an ADF test that is robust to structural breaks, for example, with a slightly modified version of the GLS-type ADF test by Elliott et al. (1996). Since the simulation results and diagnostic tests are already convincing, this is left for further research.

The maximum likelihood estimates for the hyperparameters, with their standard errors for the finally selected model are included in Appendix C.

4. Evaluation of Discontinuities

There was a natural desire to make an assessment of the discontinuity as quickly as possible after the tablet questionnaire was in use, in order to inform users about

Table 5. Simulation Results for Discontinuity Estimates with a Smooth Trend Model and Local Level Trend Model.

	Actual disc.		Smooth trend		Local level trend	
	Estimated discontinuity	Abs. diff.	Estimated discontinuity	Abs. diff.	Estimated discontinuity	Abs. diff.
Inflow variables						
Number of overseas visits by UK residents	506	178	684	170	791	238
Expenditure abroad by UK residents in £	300,330	20,827	321,157	146,700	452,366	264,853
Overseas residents migrating to the UK	4	1	3	5	6	4
Total arrival passenger flow	938	7	945	127	1,227	265
Arrival passenger flow excl. Channel Islands & Isle of Man	925	65	990	129	1,257	254
Outflow variables						
Number of visits to the UK by overseas residents	283	12	295	66	444	144
Expenditure in the UK by overseas residents in £	171,141	21,110	192,251	94,948	236,700	101,594
UK residents migrating abroad	3	1	2	3	3	1
Total departure passenger flow	937	8	945	165	1,381	325
Departure passenger flow excl. Channel Islands & Isle of Man	924	46	970	165	1,398	321

Note. All values in units of 1,000.

the impacts of the change on the time series of estimates. The published estimates were accompanied by warnings that the quality of estimates of change would be reduced during and after the rollout of the new questionnaire, but there was pressure from users of the statistical outputs for more certainty in how the estimates could be used. The evidence from Table 4 is that the longer the elapsed period after the rollout, the better the estimation of the discontinuity becomes (though with reducing benefits of additional months).

This led to several assessments of the size and importance of the discontinuities on the main IPS output variables during the period after the rollout. We give some examples of each of these below, starting with an assessment after two months in Subsection 4.1. The analysis periods were only chosen after the rollout, so did not correspond exactly with those chosen in the power analysis.

Recall from Section 3 that without a parallel run, the estimation of the discontinuity relies on some assumptions about the stable evolution of the underlying time series of estimates. The period of rollout was however affected by changes to traveler and migrant behavior driven by the period of uncertainty over Brexit. The Brexit referendum was in June 2016, but the two-year transition period was coming to an end just after the rollout, so there was an unusual amount of migration and some changes to tourism in expectation. So a priori we might expect that the model will not be as effective, since the actual changes are affected by Brexit, and these might obscure the effect of the discontinuity. We return to this topic in Section 5.

Since some of the key IPS variables are monetary and many of the values being estimated are large, we also considered that the variance could increase with the estimate, which would suggest that a log transformation would be needed to stabilize the variance. We therefore applied the same models to log-transformed data, though in most of the variables analyzed this did not provide a substantial improvement.

4.1. Early Estimation of Discontinuities

An initial analysis used data up to June 2018, which covered the roll-out period (September 2017–April 2018) and (since the roll-out was essentially completed by early April) three months afterward. The minimum detectable effects would be expected to be between the zero- and four-month columns in Table 4 if the series behaved as in the test data. Figure 5 (left) shows the estimated discontinuity for *svissosres*. In this case the estimate seems to be close to stabilizing, although it is hard to say what will happen when additional data points are added. The estimated discontinuity (in millions of people) in June 2018 is -0.244 ± 0.153 —significantly different from zero, but not very accurately estimated. This discontinuity is around 7% (% discontinuity values are calculated relative to the estimate of the trend at the given time throughout. They are therefore not affected by seasonal variations), which is around the smallest detectable difference according to the earlier power analysis. This was the only variable where the estimated discontinuity was significantly different from zero at this stage.

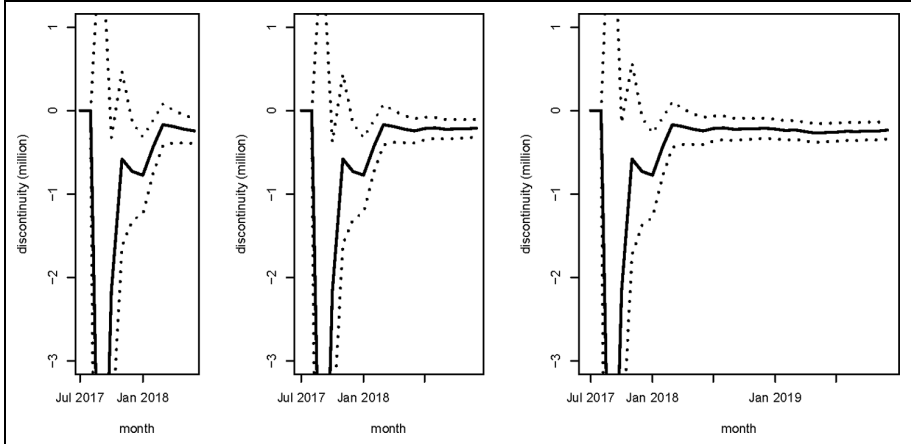


Figure 5. Estimated discontinuity (millions of visits) and its 95% confidence interval for the number of visits to the UK by overseas residents *svisosres*; (left) based on data up to June 2018; (center) based on data up to December 2018; (right) based on data up to December 2019. Periods before the rollout begins have estimated discontinuity of zero and are not shown before July 2017.

By contrast, Figure 6 shows the estimated discontinuity for *sexpukres*, and here there is no sign that the estimate has stabilized yet. The latest month’s estimate is still quite different from the previous month, and the estimated confidence intervals are wide. The discontinuity is around 3%, considerably smaller than the minimum detectable difference from the power analysis. The apparent lack of stabilization may therefore result only from the inability of the model to detect a discontinuity of this size with the current design.

In both of these examples, the behavior of the trend component of the models has not changed as a result of the addition of the latest data. This suggests that there has been no detectable effect of Brexit, or possibly that some of the Brexit effect has been picked up in the estimate of the discontinuity.

Across all the variables considered, most of the estimates of discontinuities are not significantly different from zero at this stage, and are smaller than the anticipated minimum detectable effects from Table 4. Nevertheless, some of the estimated discontinuities are large, up to 20%, and the effects on (for example) the estimated numbers of migrants would be relevant to users.

Almost all of the discontinuity estimate are negative, which means that the measurement made with tablets is lower than the previous paper-based measurement. This seems to contradict the initial indications from the pilot study, which were that the tablets were better at capturing expenditure, which was therefore higher in the new mode. The pilot used a small sample, however, and results from it may not be a strong indicator of direction of the discontinuity. If the indications of direction of the discontinuity from the pilot were correct, it is possible that the size of the

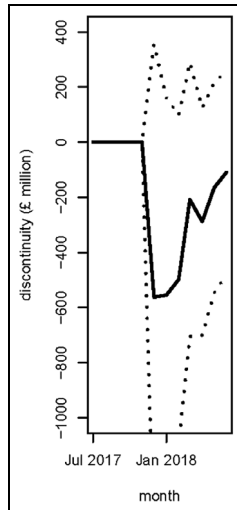


Figure 6. Estimated discontinuity (£m) and its 95% confidence interval for expenditure abroad by UK residents *sexpukres*, based on data up to June 2018.

discontinuity is at least in part confounded with changes in migration and expenditure patterns influenced by changing exchange rates and uncertainty over Brexit.

4.2. Estimation as the Basis for Deciding on an Adjustment

The second assessment used data up to December 2018, covering the roll-out and eight months afterward. This was the main “live” evaluation to make a judgment about whether to make a formal adjustment to the IPS estimates, since it was felt that users could not wait longer for an official assessment. The minimum detectable effects would be expected to be close to the 10-month columns in Table 4 if the series behaved in the same way as in the test data.

In Figure 5 (left panel) we subjectively assessed that the estimated discontinuity for *svisosres* was close to stabilizing. With the additional data to December 2018 we can see the evolution of this series (Figure 5, center panel).

The estimated discontinuity (in millions of people) estimated with data up to December 2018 is -0.212 ± 0.107 , slightly smaller than the discontinuity estimated using the earlier data only, and with the variance halved. The estimated discontinuity for this variable is significantly different from zero, and the discontinuity is around 6%, which is larger than the smallest detectable difference at this stage according to the earlier power analysis (Table 2).

Most of the series had estimated discontinuities which had stabilized over the period considered. *Svisukres* shows different behavior however (Figure 7, left panel).

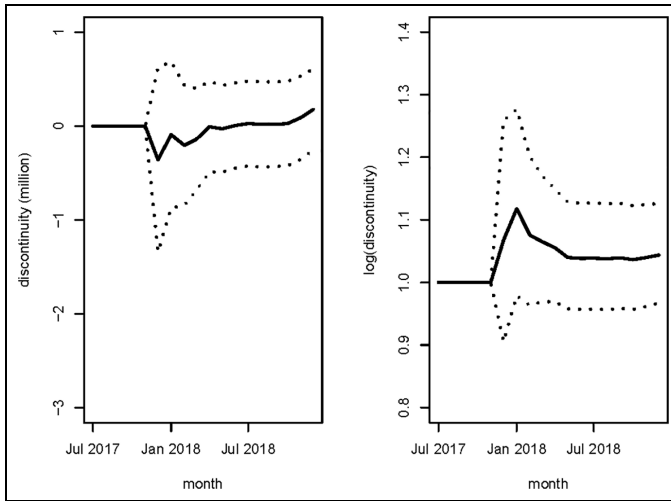


Figure 7. Estimated discontinuity (millions of visits) and its 95% confidence interval for the number of overseas visits by UK residents *svisukres* (left) and $\log(\text{svisukres})$ (right), based on data up to December 2018.

This clearly has not stabilized, and the estimated discontinuity has a wide confidence interval. We tried the log transformation for this variable, and estimated the discontinuity in the transformed variable. This gives us the filtered estimates in Figure 7 (right panel), which have stabilized, an instance where the log transformation is clearly helpful. The log transformation implies a multiplicative model. The fitted parameter can be interpreted as a proportional discontinuity in the series. The power was not assessed on the log-transformed data, so we cannot see whether the effects are in the range that was expected to be identifiable.

In all the series considered using data up to December 2018, the behavior of the trend components of the models continues to be unchanged. This suggests that either there has been no detectable effect of Brexit on the trend, or possibly that some of the Brexit effect has been picked up in the estimate of the discontinuity. It is not possible to disentangle these alternative explanations further with the collected data.

Only two of the estimated discontinuities are significantly different from zero (*svisosres* and *sexposres*), and only these two estimates are larger than the anticipated minimum detectable effects in Table 4. Where the estimated discontinuity is not detectable (note that this is not the same as saying that no discontinuity is present, just that we have not been able to detect it with the statistical power of the fitted model), making an adjustment would involve an assumption about the stability or evolution of the discontinuity estimate outside the period of the analysis. Combined with the uncertainty in estimating the discontinuity, an adjustment would therefore not improve the quality of the series of estimates.

The two series with estimated discontinuities significantly different from zero are more problematic. First, we actually make assessments for ten variables (though two pairs of variables are so similar (Table 1) that there are probably only eight independent tests). A Bonferroni type correction to the significance level would mean that neither discontinuity would continue to be significant. Second, we are concerned that some of the actual evolution in the series due to behavior changes induced by the approaching Brexit deadline have been incorporated in the estimated discontinuity. For these two reasons it was decided that no adjustment was warranted in these two series either.

4.3. Retrospective Evaluation

It was also possible to revisit the series up to December 2019, including the rollout and a further twenty months. This is in fact almost the longest period that can be available for evaluation, since the IPS series were strongly disrupted from late March 2020 by the COVID-19 pandemic. Any adjustment to the state-space model from Subsection 3.2 to make the trend sufficiently responsive to include this period would automatically mean that the trend at the time of the discontinuity was not affected by the new data, so nothing would be gained from adding anything further.

The extra year of data makes almost no difference to the conclusions drawn at the time a decision on adjustment was made. Two of the estimated discontinuities are significantly different from zero, and *svisukres* continues to be the only series where the log transformation leads to a substantial improvement in the stability. The evolution of the discontinuity estimates is shown for *svisosres* and $\log(\textit{svisukres})$ in Figures 5 (right panel) and 8 respectively.

Table 6 shows the estimated effect of the change to tablet data collection in the IPS for all the considered variables and their standard errors. Only *svisosres* has a discontinuity which is significantly different from zero, and this is also the only variable where the estimated discontinuity is close to the minimum detectable effect from Table 4.

We did not further extend the time series with the observations that became available after December 2019. As a result of the COVID 19 crisis, international passenger traffic virtually ground to a halt, resulting in a huge disruption of the time series and a sudden misspecification of the model in Equation (1). There are several ways to account for these shocks in the model, Equation (1). One approach is to increase the flexibility of the trend component by making the variance of the slope disturbance terms time varying, see Van den Brakel et al. (2022) for details. In addition a major adjustment of the seasonal component would be necessary. The consequence of such interventions is that data observed after the start of the corona crisis do not add additional information to the estimate of the discontinuities. On top of that it was already established that the series observed until December 2019 provided enough information to obtain stable estimates for the discontinuities.

In Subsection 4.2 we saw that log transformation of the data was potentially useful for one variable. We therefore examined the performance of the models with log

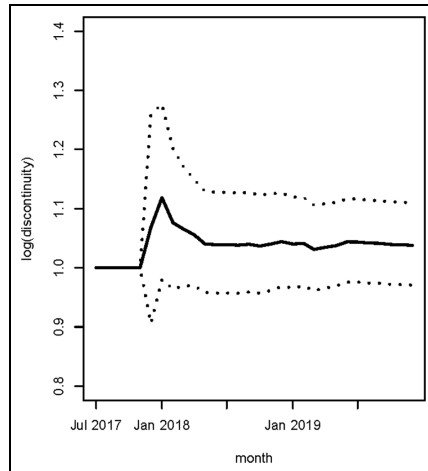


Figure 8. Estimated discontinuity and its 95% confidence interval for the log of the number of overseas visits by UK residents, $\log(svisukres)$, using data up to December 2019.

transformations for all the variables in the transition, and the results of these are presented in Table 7. With the log transformation the models meet the normality assumption of the state space model better (see Appendix B, Table B3); nevertheless we prefer the models on the original scale as the interpretation of the discontinuity is more straightforward.

5. Discussion

The changes to the IPS field procedures could only be managed with a phased transition from the old to the new tablet-based questionnaire, which did not allow for a randomization of the two questionnaires as treatments in an embedded experiment. A drawback of the time series modeling approach is that there is no control over the precision and the size of the discontinuities that can be observed, which increases the risk that substantial discontinuities cannot be assessed. To assess the size of the discontinuities that can be detected with a time series modeling approach before the start of the transition to a new survey design, a simulation is proposed. If such a simulation indicates that the time series modeling approach is insufficient to detect discontinuities that are of importance of the data users, then the precision can be improved by combining the time series modeling approach with a (small) parallel run. The direct estimates for the discontinuities and their standard errors obtained with the parallel run can be used in the time series model through an exact initialization of the Kalman filter. The data observed before and after the change over to the new design will further improve the discontinuity estimates from the parallel run and result in a final estimate that will be more precise and that converges more quickly to a stable estimate than the time series model estimates based on a diffuse initialization of the Kalman filter in absence of a parallel run. As

Table 6. Final Estimates of the Discontinuity Parameters for the IPS Variables, and Their Standard Errors, Twenty Months After the Completion of the Tablets Rollout. The Final Column, Sim SE Gives the Standard Errors Expected from the Simulations Described in Section 3.3.

Inflow variables		Estimate	SE	Sim SE
Number of overseas visits by UK residents (000s)	svisukres	159.1	188.8	172.9
Expenditure abroad by UK residents (£m)	sexpukres	210.4	163.2	144.1
Overseas residents migrating to the UK (000s)	smigosar	8.9	5.3	5.5
Total arrival passenger flow (000s)	sflowarr	232.5	161.9	118.2
Arrival passenger flow excl. Channel Islands & Isle of Man (000s)	sflowarrn	255.6	163.5	128.9
Outflow variables		Estimate	SE	Sim SE
Number of visits to the UK by overseas residents (000s)	svisosres	-233.9	54.2	66.1
Expenditure in the UK by overseas residents (£m)	sexposres	-231.1	143.2	70.6
UK residents migrating abroad (000s)	smigukdep	-3.1	2.6	2.4
Total departure passenger flow (000s)	sflowdep	321.5	253.8	165.7
Departure passenger flow excl. Channel Islands & Isle of Man (000s)	sflowdepn	335.1	253.2	164.8

shown in Van den Brakel et al. (2020) it is possible to assess through simulations what precision can be obtained with the time series modeling approach in combination with parallel runs of different lengths.

Van den Brakel et al. (2020, Figure 3) demonstrate the impacts of different designs of a parallel run on the size of detectable effects, but it is not known whether the pattern they observe is generalizable. In the IPS example, the transition is rather lumpy because of the disproportionate size of the flow of passengers through Heathrow airport, and therefore the discontinuity is not well estimated, and the power to detect changes with this rollout pattern is rather low.

In another simulation, the performance of the discontinuity estimates under a model with a smooth trend model and a local level trend model were compared. An Augmented Dickey Fuller test clearly rejects the null hypothesis that the input series have a second order level of integration, which supports the choice of a local level trend model. The simulation, nevertheless, shows that estimates for the discontinuities under the smooth trend model are much closer to the true values assumed in the simulation than under the local level trend model. We anticipate that this is because of the more volatile behavior of the trend under the local level trend model. This implies that, compared to the smooth trend model, the discontinuity estimates are more based on observations close to the period of the introduction of the tablets while observations further away from this period have less influence.

Table 7. Final Estimates of the Log Discontinuity Parameters from the Models of the Log-Transformed IPS Variables Twenty Months After the Completion of the Tablets Rollout.

Inflow variables		Estimate of log discontinuity	SE
Number of overseas visits by UK residents (000s)	svisukres	0.037	0.034
Expenditure abroad by UK residents (£m)	sexpukres	0.022	0.046
Overseas residents migrating to the UK (000s)	smigosar	-0.169	0.107
Total arrival passenger flow (000s)	sflowarr	0.017	0.017
Arrival passenger flow excl. Channel Islands & Isle of Man (000s)	sflowarrn	0.019	0.017
Outflow variables		Estimate	SE
Number of visits to the UK by overseas residents (000s)	svisosres	-0.083	0.018
Expenditure in the UK by overseas residents (£m)	sexposres	-0.128	0.073
UK residents migrating abroad (000s)	smigukdep	-0.094	0.094
Total departure passenger flow (000s)	sflowdep	0.028	0.023
Departure passenger flow excl. Channel Islands & Isle of Man (000s)	sflowdepn	0.029	0.023

The estimation of the discontinuity in real time is a classical example of a trade-off of timeliness and accuracy. When the roll-out was complete there was already pressure from users for an estimate of the effect of the new questionnaire, but at this stage it could only be estimated very imprecisely. It took some further build-up of the time series after the transition before the effect was reasonably estimated.

When there is a transition without a randomization in a parallel run we must always require the implicit assumption that the evolution of the underlying series continues undisturbed. For the transition in the IPS this assumption was not met, because of the effects of the transition following the Brexit referendum. It was not really practical to foresee all of these effects at the time the questionnaire was being introduced, but the effect was to include some of the real change in the estimate of the discontinuity (i.e., the real change and the discontinuity were partially confounded), which made it more difficult to assess whether any change was real. Even with this effect, however, most of the estimated discontinuities were smaller than the minimum detectable effects. As a result, no adjustment was made to the series on account of the discontinuities. Users were kept in touch with the expected effects of the change of questionnaire, and warned about the additional uncertainty arising around the transition period. But in the end no adjustment was made. Nevertheless this is an interesting case study of how to plan and execute a survey transition in the case where no parallel run is possible, a situation which arises quite frequently because of the difficulty of expanding the field force to deal with parallel data collection. The confounding of the discontinuity with some changes in the real evolution of the time series is a salutary lesson to avoid periods of predictable change in

introducing a new method. Such periods, however, often cannot be predicted. If estimation of a discontinuity is critical, it may be necessary to postpone a change. But even that may not be practical because of the cost of waiting and the unpredictability of a period of stability. The only way to retain some control in periods of change is to do a parallel run.

Appendix A

State Space Representations

The state space representations of the models in Section 3 are defined by a measurement equation and a transition equation. The measurement equation defines how the observed series depends on the unobserved state variables, which are collected in a vector α_t . The transition equation describes how the state variables evolve from period $t-1$ to t .

The state space representation for the model in Equation (1) with the smooth trend is defined by the measurement equation:

$$\begin{aligned} \hat{y}_t &= z_t \alpha_t + \varepsilon_t, \\ z_t &= (1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, x_t), \\ \alpha_t &= (L_t, R_t, S_{t,1}, S_{t,1}^*, S_{t,2}, S_{t,2}^*, S_{t,3}, S_{t,3}^*, S_{t,4}, S_{t,4}^*, S_{t,5}, S_{t,5}^*, S_{t,6}, \beta)' , \\ \varepsilon_t &\sim N(0, \sigma_\varepsilon^2), \end{aligned}$$

and transition equation:

$$\begin{aligned} \alpha_t &= T \alpha_{t-1} + \eta_t, \\ T &= T^L \oplus T^S \oplus 1, \\ T^L &= \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \\ T^S &= C_1 \oplus C_2 \oplus C_3 \oplus C_4 \oplus C_5 \oplus 1, \\ C_l &= \begin{pmatrix} \cos(h_l) & \sin(h_l) \\ -\sin(h_l) & \cos(h_l) \end{pmatrix}, \quad h_l = \frac{\pi l}{6}, \quad l = 1, \dots, 5, \\ \eta_t &\sim N(\mathbf{0}_{[14]}, \Sigma), \\ \Sigma &= 0 \oplus \sigma_\eta^2 \oplus \sigma_\omega^2 \oplus I_{[11]} \oplus 0. \end{aligned} \tag{A.1}$$

Here \oplus denotes the direct sum that defines a (block) diagonal matrix, $\mathbf{0}_{[p]}$ a p dimensional column vector with each element equal to zero, and $I_{[p]}$ the p dimensional identity matrix. Note that z_t and T are known design matrices, which follow from the model specification of the state variables.

The state space representation for the model in Equation (2) with the smooth trend is defined by the measurement equation:

$$\begin{aligned}\hat{\mathbf{y}}_t &= \mathbf{Z}_t \boldsymbol{\alpha}_t, \\ \hat{\mathbf{y}}_t &= \left(\hat{y}_t^{(1)}, \dots, \hat{y}_t^{(n)} \right)', \\ \mathbf{Z}_t &= \left(\mathbf{Z}^\theta \ \mathbf{Z}^\lambda \ \mathbf{Z}_t^\beta \ \mathbf{Z}_t^e \right), \\ \mathbf{Z}^\theta &= (1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 1) \otimes \mathbf{1}_{[n]}, \\ \mathbf{Z}_t^\beta &= \left(\delta_t^{(1)}, \dots, \delta_t^{(n)} \right)', \ \mathbf{Z}_t^e = \mathbf{I}_{[n]}. \\ \boldsymbol{\alpha}_t &= \left(\boldsymbol{\alpha}_t^\theta \ \boldsymbol{\alpha}_t^\lambda \ \boldsymbol{\alpha}_t^\beta \ \boldsymbol{\alpha}_t^e \right)', \\ \boldsymbol{\alpha}_t^\theta &= \left(L_t, R_t, S_t, 1, S_{t,1}^*, S_{t,2}^*, S_{t,3}^*, S_{t,4}^*, S_{t,5}^*, S_{t,6}^*, I_t \right), \\ \boldsymbol{\alpha}_t^\beta &= \beta, \ \boldsymbol{\alpha}_t^e = \left(e_t^{(1)}, \dots, e_t^{(n)} \right),\end{aligned}$$

with \otimes the Kronecker product. There is no measurement error in the measurement equation, since all series share one population white noise term. This error term is therefore included in the state vector. The transition equation for the model in Equation (2) is defined as (A.1) with

$$\begin{aligned}\mathbf{T} &= \mathbf{T}^\theta \oplus \mathbf{T}^\lambda \oplus \mathbf{1} \oplus \mathbf{T}^e, \\ \mathbf{T}^\theta &= \mathbf{T}^L \oplus \mathbf{T}^S \oplus \mathbf{0}, \ \mathbf{T}^e = \mathbf{O}_{[n \times n]}, \\ \boldsymbol{\eta}_t &\sim \mathcal{N}(0_{[15+2n]}, \boldsymbol{\Sigma}), \\ \boldsymbol{\Sigma} &= \boldsymbol{\Sigma}^\theta \oplus \boldsymbol{\Sigma}^\lambda \oplus \mathbf{0} \oplus \boldsymbol{\Sigma}^e, \\ \boldsymbol{\Sigma}^\theta &= \mathbf{0} \oplus \sigma_\eta^2 \oplus \sigma_\eta^2 \mathbf{I}_{[11]} \oplus \sigma_\varepsilon^2, \ \boldsymbol{\Sigma}^e = \oplus_{i=1}^n \sqrt{\text{var}(\hat{y}_t^{(i)})},\end{aligned}$$

with $\mathbf{O}_{[n \times n]}$ an $n \times n$ matrix with each element equal to zero and $\mathbf{1}_{[p]}$ a p dimensional column vector with each element equal to 1. Furthermore it is understood that $\text{var}(\hat{y}_t^{(i)})$ can be estimated from the survey data.

Finally $\boldsymbol{\alpha}_t^\lambda$, \mathbf{Z}^λ , and \mathbf{T}^λ need to be specified such that $\boldsymbol{\alpha}_t^\lambda$ obeys the restriction in Equation (3). This can be achieved in different ways. Hungnes et al. (2024) propose to impose restriction in Equation (3) in the measurement equation by taking

$$\begin{aligned}\mathbf{Z}^\lambda &= \begin{pmatrix} \mathbf{I}_{[n-1]} \\ -\mathbf{1}'_{[n-1]} \end{pmatrix} \left[\begin{pmatrix} \mathbf{I}_{[n-1]} \\ -\mathbf{1}'_{[n-1]} \end{pmatrix}' \quad \begin{pmatrix} \mathbf{I}_{[n-1]} \\ -\mathbf{1}'_{[n-1]} \end{pmatrix} \right]^{-1/2}, \\ \boldsymbol{\alpha}_t^\lambda &= (\lambda_t^{(1)}, \dots, \lambda_t^{(n-1)}), \ \mathbf{T}^\lambda = \mathbf{I}_{[n-1]} \text{ and } \boldsymbol{\Sigma}^\lambda = \sigma_\lambda^2 \mathbf{I}_{[n-1]}.\end{aligned}$$

A constructive and supportive referee noted that the same restriction can be imposed in the transition equation by taking $\boldsymbol{\alpha}_t^\lambda = \mathbf{T}^\lambda \boldsymbol{\alpha}_{t-1}^\lambda + \mathbf{R}^\lambda \boldsymbol{\eta}_t^\lambda$ with

$$R^\lambda = \begin{pmatrix} I_{[n-1]} \\ -\mathbf{1}'_{[n-1]} \end{pmatrix} \left[\begin{pmatrix} I_{[n-1]} \\ -\mathbf{1}'_{[n-1]} \end{pmatrix}' \begin{pmatrix} I_{[n-1]} \\ -\mathbf{1}'_{[n-1]} \end{pmatrix} \right]^{-1/2}$$

$$Z^\lambda = T^\lambda = I_{[n]}, \alpha_t^\lambda = (\lambda_t^{(1)}, \dots, \lambda_t^{(n)}) \text{ and } \Sigma^\lambda = \sigma_\lambda^2 I_{[n-1]}.$$

This implies that the vector with state disturbance terms in Equation (A.1) is pre-multiplied with a selection matrix \mathbf{R} , that is, $\mathbf{R}\eta_t$, with $\mathbf{R} = I_{[14]} \oplus \mathbf{R}^\lambda \oplus 1 \oplus I_{[n]}$. Finally the method proposed by Doran (1992) to impose a restriction on the state variables can be used. According to this method the measurement equation is extended with the restriction as follows:

$$\begin{pmatrix} \hat{y}_t \\ 0 \end{pmatrix} = \begin{pmatrix} Z_t \\ r \end{pmatrix} \alpha_t + \mathbf{0}_{[n+1]},$$

with $Z^\lambda = T^\lambda = I_{[n]}$, $\alpha_t^\lambda = (\lambda_t^{(1)}, \dots, \lambda_t^{(n)})$. In the transition equation we now have $\Sigma^\lambda = \sigma_\lambda^2 I_{[n]}$. The restriction entails that the sum over the elements of α_t^λ equals zero which implies that $r = (r^\theta \ r^\lambda \ r^\beta \ r^e)'$ with $r^\theta = \mathbf{0}'_{[14]}$, $r^\lambda = \mathbf{1}'_{[n]}$, $r^\beta = 0$, and $r^e = \mathbf{0}'_{[n]}$.

Appendix B

Model Evaluation Tests

See Durbin and Koopman (2012, 38–9) for an overview of the test definitions.

Table B1. Diagnostics for the Model in Equation (1) with the Smooth Trend Model.

Variable	F-test for heteroscedasticity		Bowman-Shenton test for normality		Ljung Box test for autocorrelation (lag 12)	
	F_{24}^{24}	p-Value	$\chi^2_{(2)}$	p-Value	$\chi^2_{(12)}$	p-Value
svisukres	1.169	.705	0.085	.959	29.861	.003
svisosres	2.673	.019	0.808	.668	14.575	.266
sexpukres	2.283	.048	0.231	.891	13.949	.304
sexposres	2.176	.063	0.761	.684	12.419	.413
smigosar	2.449	.033	12.902	.002	2.630	.998
smigukdep	1.735	.184	2.792	.248	3.886	.985
sflowarr	2.292	.047	12.040	.002	11.873	.456
sflowdep	2.068	.081	15.227	.001	8.096	.778
sflowarrn	1.814	.152	12.375	.002	12.625	.397
sflowdepn	1.971	.103	15.166	.001	8.802	.720
narr	1.445	.373	1.486	.476	13.173	.357
ndep	1.330	.490	1.783	.410	18.069	.114

Table B2. Diagnostics for the Model in Equation (1) with the Local Level Trend Model.

Variable	F-test for heteroscedasticity		Bowman-Shenton test for normality		Ljung Box test for autocorrelation (lag 12)	
	F_{24}^{24}	p-Value	$\chi_{(2)}^2$	p-Value	$\chi_{(12)}^2$	p-Value
svisukres	0.621	.250	0.838	.658	8.168	.772
svisosres	2.071	.081	0.628	.731	11.787	.463
sexpukres	2.681	.019	1.196	.550	10.829	.544
sexposres	2.272	.048	0.549	.760	14.413	.275
smigosar	2.605	.023	12.432	.002	2.250	.999
smigukdep	1.585	.266	5.038	.081	3.230	.994
sflowarr	1.655	.225	5.952	.051	11.951	.450
sflowdep	1.364	.453	14.391	.001	10.485	.574
sflowarrn	1.344	.474	8.116	.017	12.314	.421
sflowdepn	1.325	.496	14.261	.001	10.717	.553
narr	1.157	.724	1.728	.421	7.489	.824
ndep	1.314	.509	1.390	.499	18.870	.092

Table B3. Diagnostics for the Model in Equation (1) with the Smooth Trend Model and with Log-Transformed Data.

Variable	F-test for heteroscedasticity		Bowman-Shenton test for normality		Ljung Box test for autocorrelation (lag 12)	
	F_{24}^{24}	p-Value	$\chi_{(2)}^2$	p-Value	$\chi_{(12)}^2$	p-Value
svisukres	0.498	.094	2.050	.359	14.369	.278
svisosres	2.764	.016	1.904	.386	17.162	.144
sexpukres	1.315	.507	5.159	.076	4.419	.975
sexposres	1.784	.164	0.081	.961	16.961	.151
smigosar	2.236	.054	0.819	.664	7.646	.812
smigukdep	1.262	.573	1.053	.591	10.486	.573
sflowarr	1.472	.350	2.021	.364	11.742	.467
sflowdep	1.418	.399	10.270	.006	11.376	.497
sflowarrn	1.500	.327	1.959	.376	12.559	.402
sflowdepn	1.398	.418	10.724	.004	11.956	.449
narr	1.303	.522	0.664	.717	14.521	.269
ndep	1.769	.170	0.553	.758	24.869	.016

Appendix C

Hyperparameter Estimates

The hyperparameters of the model in Equation (1) are, as explained in Subsection 3.2, estimated on the log scale. Let $\hat{\sigma}_{ML}^2$ denote the Maximum Likelihood estimate for hyperparameter σ^2 . The variance of $\hat{\sigma}_{ML}^2$ is derived from the Fisher information

Table C1. Hyperparameter Estimates for the Model in Equation (1) with the Smooth Trend Model.

Variable	Slope		Seasonal		Measurement error	
	$\hat{\sigma}_\eta$	$se(\hat{\sigma}_\eta)$	$\hat{\sigma}_\omega$	$se(\hat{\sigma}_\omega)$	$\hat{\sigma}_\varepsilon$	$se(\hat{\sigma}_\varepsilon)$
svisukres	6,711	2,799	0	—	208,019	17,055
svisosres	0.18	29.00	0	—	140,409	10,964
sexpukres	7.442E6	3.355E6	1.561E7	2.932E6	1.257E8	2.189E7
sexposres	6.483E6	2.328E6	3.304E6	2.891E6	1.087E8	1.373E7
smigosar	84.54	66.20	503.05	156.43	7,297.75	940.46
smigukdep	32.74	24.27	0.01	1.83	3,901.49	311.83
sflowarr	6,536	2,454	19,029	3,613	130,371	26,650
sflowdep	7,085	2,622	2.59	885.80	276,685	22,270
sflowarrn	7,259	2,826	18,736	3,443	123,856	26,278
sflowdepn	7,143	2,708	0	—	273,888	22,043
narr	108.87	29.11	0.05	6.46	425.62	41.20
ndep	120.57	30.93	0.06	20.18	465.05	44.41

Note. The point estimates for the variance of the seasonal disturbance terms for *svisukres*, *svisosres*, and *sflowdepn* tend to zero with very large standard errors. For these variables the seasonal component is modeled as a time invariant effect, that is, $\hat{\sigma}_\omega$ is set to zero.

matrix and is denoted as $V(\tilde{\sigma}_{ML}^2)$. The back-transformed point estimates for the standard errors on the normal scale are given by

$$\hat{\sigma} = \sqrt{\exp(\tilde{\sigma}_{ML}^2)}.$$

A first order Taylor approximation for the back-transformed standard error of $\hat{\sigma}$ is obtained by

$$se(\hat{\sigma}) = 0.5 * \sqrt{\exp(\tilde{\sigma}_{ML}^2)} \sqrt{V(\tilde{\sigma}_{ML}^2)}.$$

The standard errors with their standard errors for Equation (1) with the smooth trend model are summarized in Table C1.

Acknowledgements


The authors would like to thank three anonymous referees and an Associate Editor for careful reading and providing useful comments on a previous draft of the manuscript. The views expressed in this paper are those of the authors and do not reflect the policies of the Office for National Statistics or Statistics Netherlands.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Part of the work of PAS and JvdB was funded by the

Office for National Statistics under contracts ITT PU-16/0031-5.001 and ITT PU-16/0031-6.009.

ORCID iD

Paul A. Smith  <https://orcid.org/0000-0001-5337-2746>

References

- Benedikt, L. 2015. "Developing the UK International Passenger Survey in Blaise 5 on Tablet Computer." *16th International Blaise Users Conference, Beijing, April 13–15*. https://www.blaiseusers.org/2015/papers/06_B%20IBUC_2015_IPSOntablet.pdf (accessed April 7, 2023).
- Bowman, K. O., and L. R. Shenton. 1975. "Omnibus Test Contours for Departures from Normality Based on $\sqrt{b_1}$ and b_2 ." *Biometrika* 62: 243–50. DOI: <https://doi.org/10.1093/biomet/62.2.243>.
- Couper, M. P. 2011. "The Future of Modes of Data Collection." *Public Opinion Quarterly* 75: 889–908. DOI: <https://doi.org/10.1093/poq/nfr046>.
- De Leeuw, E. 2005. "To Mix or Not to Mix Data Collection Modes in Surveys." *Journal of Official Statistics* 21 (2): 233–55. DOI: <https://www.scb.se/dokumentation/statistiska-metoder/JOS-archiv/>.
- De Leeuw, E. 2008. "Choosing the Method of Data Collection." In *International Handbook of Survey Methodology*, edited by E. D. de Leeuw, J. J. Hox, and D. A. Dillman, 113–35. European Association of Methodology (EAM) Methodology Series. New York, NY: Routledge, Taylor and Francis.
- Dillman, D. A., and L. M. Christian. 2005. "Survey Mode as a Source of Instability in Responses Across Surveys." *Field Methods* 17 (1): 30–52. DOI: <https://doi.org/10.1177/1525822X04269550>.
- Dillman, D. A., J. D. Smyth, and L. M. Christian. 2014. *Internet, Phone, Mail and Mixed-Mode Surveys: The Tailored Design Method*. New York, NY: Wiley and Sons.
- Doornik, J. A. 2009. *An Object-Oriented Matrix Programming Language Ox 6*. London: Timberlake Consultants Press.
- Doran, H. E. 1992. "Constraining Kalman Filter and Smoothing Estimates to Satisfy Time Varying Restrictions." *Review of Economics and Statistics* 74: 568–72. DOI: <https://doi.org/10.2307/2109505>.
- Durbin, J., and S. J. Koopman. 2012. *Time Series Analysis by State Space Methods*. 2nd ed. Oxford: Oxford University Press. DOI: <https://doi.org/10.1093/acprof:oso/9780199641178.001.0001>.
- Elliott, G., T. J. Rothenberg, and J. H. Stock. 1996. "Efficient Tests for an Autoregressive Unit Root." *Econometrica* 64: 813–36. DOI: <https://doi.org/10.2307/2171846>.
- Fanning, J., and E. McAuley. 2014. "A Comparison of Tablet Computer and Paper Based Questionnaires in Health Aging Research." *JMIR Research Protocols* 3 (3): e3291. DOI: <https://doi.org/10.2196/resprot.3291>.
- Frenč, C. 2016. "Measuring Tourism at the Border: A Critical Analysis of the Icelandic Context." *Scandinavian Journal of Hospitality and Tourism* 16 (Sup1): 87–97. DOI: <https://doi.org/10.1080/15022250.2016.1244597>.
- Harvey, A. C. 1989. *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press. DOI: <https://doi.org/10.1017/CBO9781107049994>.

- Harvey, A. C., and J. Durbin. 1986. "The Effects of Seatbelt Legislation on British Road Casualties: A Case Study in Structural Time Series Modelling." *Journal of the Royal Statistical Society: Series A (General)* 149: 187–227. DOI: <https://doi.org/10.2307/2981553>.
- Hassler, K., J. P. Kelly, and T. L. Serfass. 2018. "Comparing the Efficacy of Electronic-Tablet to Paper-Based Surveys for On-Site Survey Administration." *International Journal of Social Research Methodology* 21 (4): 487–97. DOI: <https://doi.org/10.1080/13645579.2018.1432403>.
- Hungnes, H., T. Skjerpen, J. I. Hamre, X. C. Jansen, D. Q. Pham, and O. Sandvik. 2024. "Structural Break in the Norwegian Labor Force Survey Due to a Redesign During a Pandemic." *Journal of Official Statistics* 40: 122–60. DOI: <https://doi.org/10.1177/0282423X241235267>.
- Klausch, L. T., B. Schouten, B. Buelens, and J. A. Van den Brakel. 2017. "Adjusting Measurement Bias in Sequential Mixed-Mode Surveys Using Re-Interview Data." *Journal of Survey Statistics and Methodology* 5 (4): 409–32. DOI: <https://doi.org/10.1093/jssam/smx022>.
- Kolenikov, S., and C. Kennedy. 2014. "Evaluating Three Approaches to Statistically Adjust for Mode Effects." *Journal of Survey Statistics and Methodology* 2 (2): 126–58. DOI: <https://doi.org/10.1093/jssam/smu004>.
- Koopman, S. J. 1997. "Exact Initial Kalman Filtering and Smoothing for Non-Stationary Time Series Models." *Journal of the American Statistical Association* 92 (440): 1630–8. DOI: <https://doi.org/10.1080/01621459.1997.10473685>.
- Koopman, S. J., N. Shephard, and J. A. Doornik. 2008. *SsfPack 3.0: Statistical Algorithms for Models in State Space Form*. London: Timberlake Consultants Press.
- Kusumoto, Y., Y. Kita, S. Kusaka, Y. Hiyama, T. Junko, T. Kutsuna, H. Kameda, S. Aida, M. Umeda, and T. Takahashi. 2017. "Difference Between Tablet Modes and Paper Questionnaire Methods of Conducting a Survey with Community-Dwelling Elderly." *The Journal of Physical Therapy Science* 29: 2100–102. DOI: <https://doi.org/10.1589/jpts.29.2100>.
- Leisher, C. 2014. "A Comparison of Tablet-Based and Paper-Based Survey Data Collection in Conservation Projects." *Social Sciences* 3: 264–71. DOI: <https://doi.org/10.3390/socsci3020264>.
- Ljung, G. M., and G. E. Box. 1978. "On a Measure of Lack of Fit in Time Series Models." *Biometrika* 65: 297–303. DOI: <https://doi.org/10.1093/biomet/65.2.297>.
- ONS. 2009. "Port Survey Review Stage Two Final Technical Report." <https://webarchive.nationalarchives.gov.uk/ukgwa/20160105160709/http://www.ons.gov.uk/ons/guide-method/method-quality/imps/msi-programme/working-groups/entry-and-exit-working-group/port-survey-review-stage-two-final-technical-report.pdf> (accessed April 7, 2023).
- ONS. 2014. "International Passenger Survey – Overseas Travel and Tourism. User Guide (Volume 1): Background and Methodology." <https://webarchive.nationalarchives.gov.uk/ukgwa/20160105160709/http://www.ons.gov.uk/ons/guide-method/method-quality/specific/travel-and-transport-methodology/international-passenger-survey-methodology/ips-user-guide-volume-1--background--methodology.pdf> (accessed April 7, 2023).
- ONS. 2018. "Report on International Migration Data Sources: July 2018." <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/articles/reportonthecomplexityandqualityofinternationalmigrationstatistics/july2018> (accessed April 7, 2023).
- Pendry, E. 2000. "The Development of Self-Completion Questionnaires for Non-English Speakers on the International Passenger Survey." *Survey Methodology Bulletin* 47: 27–32.

- Ravert, R. D., J. Gomez-Scott, and M. B. Donnellan. 2015. "Equivalency of Paper Versus Tablet Computer Survey Data." *Educational Researcher* 44: 308–10. DOI: <https://doi.org/10.3102/0013189X15592845>.
- Rideng, A., and P. Christensen. 2004. "En Route Surveys – Some Methodological Issues." *Scandinavian Journal of Hospitality and Tourism* 4: 242–58. DOI: <https://doi.org/10.1080/15022250410003807>.
- Rogers, N., L. Blackwell, D. Elliott, A. Large, S. Ridden, and M. Wu. 2021. "Using Statistical Modelling to Estimate UK International Migration." Working Paper Series No. 23, ONS. <https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/usingstatisticalmodellingtonestimateukinternationalmigration> (accessed April 7, 2023).
- Schouten, B., J. A. Van den Brakel, B. Buelens, D. Giesen, A. Luiten, and V. Meertens. 2022. *Mixed-Mode Official Surveys; Design and Analysis*. New York, NY: Chapman and Hall.
- Statistics Netherlands. 2002. *Blaise Developers Guide*. Heerlen: Statistics Netherlands. <https://blaise.com>.
- Suzer-Gurtekin, T. 2013. "Investigating the Bias Properties of Alternative Statistical Inference Methods in Sequential Mixed-Mode Surveys." PhD thesis, University of Michigan.
- Tourangeau, R., A. Maitland, G. Rivero, H. Sun, D. Williams, and T. Yan. 2017. "Web Surveys by Smart Phone and Tablets: Effects on Survey Responses." *Public Opinion Quarterly* 81: 896–929. DOI: <https://doi.org/10.1093/poq/nfx035>.
- Van den Brakel, J. A., and S. Krieg. 2015. "Dealing with Small Sample Sizes, Rotation Group Bias and Discontinuities in a Rotating Panel Design." *Survey Methodology* 41: 267–96.
- Van den Brakel, J. A., and J. Roels. 2010. "Intervention Analysis with State-Space Models to Estimate Discontinuities Due to a Survey Redesign." *Annals of Applied Statistics* 4: 1105–138. DOI: <https://doi.org/10.1214/09-AOAS305>
- Van den Brakel, J. A., P. A. Smith, and S. Compton. 2008. "Quality Procedures for Survey Transitions – Experiments, Time Series and Discontinuities." *Survey Research Methods* 2: 123–41. DOI: <https://doi.org/10.18148/srm/2008.v2i3.68>.
- Van den Brakel, J. A., P. A. Smith, D. Elliott, S. Krieg, T. Schmid, and N. Tzavidis. 2021. "Assessing Discontinuities and Rotation Group Bias in Rotating Panel Designs." In *Advances in Longitudinal Survey Methodology*, edited by P. Lynn, 399–423. Hoboken, NJ: Wiley. DOI: <https://doi.org/10.1002/9781119376965.ch16>.
- Van den Brakel, J. A., M. Souren, and S. Krieg. 2022. "Estimating Monthly Labour Force Figures During the COVID-19 Pandemic in the Netherlands." *Journal of the Royal Statistical Society Series A: Statistics in Society* 185: 1560–83. DOI: <https://doi.org/10.1111/rssa.12869>.
- Van den Brakel, J. A., X. Zhang, and S.-M. Tam. 2020. "Measuring Discontinuities in Time Series Obtained with Repeated Sample Surveys." *International Statistical Review* 88: 155–75. DOI: <https://doi.org/10.1111/insr.12347>.
- Vannieuwenhuyze, J. T. 2014. "On the Relative Advantage of Mixed-Mode Versus Single-Mode Surveys." *Survey Research Methods* 8: 31–42. DOI: <https://doi.org/10.18148/srm/2014.v8i1.5500>.
- White, N. 2018. "International Passenger Survey: Quality Information in Relation to Migration Flows." <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/methodologies/internationalpassengersurveyqualityinformationinrelationtomigrationflows> (accessed April 7, 2023).

Zhang, H., L. Groshong, S. W. Stanis, and M. Morgan. 2021. "Comparing Onsite Electronic Survey Distribution Methods." *Annals of Tourism Research* 87: 102997. DOI: <https://doi.org/10.1016/j.annals2020.102997>.

Received: September 11, 2023

Accepted: July 29, 2024