Discussion Paper

Social media as a data source for official statistics; the Dutch Consumer Confidence Index

2016 | 01

Jan van den Brakel
Emily Söhler
Piet Daas
Bart Buelens

Social media as a data source for official statistics; the Dutch Consumer Confidence Index

Jan van den Brakel, Emily Söhler, Piet Daas and Bart Buelens

Abstract

In this paper the question is addressed how alternative data sources, such as administrative and social media data, can be used in the production of official statistics. Since most surveys at national statistical institutes are conducted repeatedly over time, a multivariate structural time series modelling approach is proposed to model the series observed by a repeated surveys with related series obtained from such alternative data sources. Generally, this improves the precision of the direct survey estimates by using sample information observed in preceding periods and information from related auxiliary series. The concept of cointegration is applied to address the question to which extent the alternative series represent the same phenomena as the series observed with the repeated survey. The methodology is applied to the Dutch Consumer Confidence Survey and a sentiment index derived from social media by Daas and Puts (2014).

Keywords:
Big data, Design-based inference, model-based inference, structural time series modelling, cointegration

Index

1. Introduction

National statistical institutes traditionally use probability sampling in combination with design-based or model-assisted inference for the production of official statistics. The concept of random probability sampling has been developed mainly on the basis of the work of Bowley (1926), Neyman (1934) and Hansen and Hurwitz (1943). See for example Cochran (1977) or Särndal et al. (1992) for an extensive introduction in sampling theory. This is a widely accepted approach, since it is based on a sound mathematical theory that shows how under the right combination of a random sample design and estimator, valid statistical inference can be made about large finite populations based on relative small samples. In addition, the amount of uncertainty by relying on small samples can be quantified through the variance of the estimators.

There is persistent pressure on national statistical institutes to reduce administration costs and response burden. In addition, declining response rates stimulate the search for alternative sources of statistical information. This could be accomplished by using administrative data like tax registers, or other large data sets – so called big data - that are generated as a by-product of processes not directly related to statistical production purposes. Examples of these include time and location of network activity available from mobile phone companies, social media messages from Twitter and Facebook and internet search behaviour from Google Trends. A common problem with this type of data sources is that the process that generates the data is unknown and likely selective with respect to the intended target population. A challenging problem in this context is to use this data for the production of official statistics that are representative of the target population. There is no randomized sampling design that facilitates the generalization of conclusions and results obtained with the available data to an intended larger target population. Hence, extracting statistically relevant information from these sources is a challenging task (Daas and Puts, 2014a).

Baker et al. (2010) address the problem of using non-probability samples and mention the possibility of applying design-based inference procedures to correct for selection bias. Buelens, Burger and Van den Brakel (2015) explore the possibility of using statistical machine learning algorithms to correct for selection bias. Instead of replacing survey data for administrative data or big data, these sources can also be used to improve the accuracy of survey data in model-based inference procedures. Marchetti et al. (2015) and Blumenstock, Cadamuro and On (2015) used big data as a source of auxiliary information for cross-sectional small area estimation models.

Many surveys conducted by national statistical institutes are conducted repeatedly. In this paper a multivariate structural time series modelling approach is applied to combine the series obtained by a repeated survey with series from alternative data sources. This serves several purposes. First, a model based estimation procedure based on a time series model increases the precision of the direct estimates by using the temporal correlation between the direct estimates in the separate editions of the survey. The use of time series modelling with the aim of improving the precision of survey data has been considered by many authors dating back to Blight and Scott (1973). Second, extending the time series model with an auxiliary series allows to model the correlation between the unobserved components of the structural time series models, e.g. trend and seasonal components. If the model detects strong positive correlations between these components, then this might further increase the

precision of the time series estimates for the sample survey. Harvey and Chung (2000) propose a time series model for the Labour Force Survey in the UK extended with a series of claimant counts. Third, the concept of cointegration in the context of multivariate state space models can be used to evaluate to which extent both series are identical. If the trend components of two observed series are cointegrated, then both series are driven by one underlying common trend. Although clearly weaker compared to the theory underlying probability sampling, it can be argued that if an auxiliary series is cointegrated with the series of the survey, they represent the same underlying stochastic process.

The Dutch Consumer Confidence Survey (CCS) is a monthly survey based on approximately 1000 respondents with the purpose of measuring the sentiment of the Dutch population about the economic climate by means of the so-called Consumer Confidence Index (CCI). Daas and Puts (2014b) developed a sentiment index, independently of the CCS, that is derived from social media platforms that was found to mimic the CCI very well. This index is referred to as the Social Media Index (SMI). In this paper the aforementioned multivariate structural time series modelling approach is applied to both series in an attempt to improve the precision of the CCI.

In Section 2, the survey design of the CCS and the estimation procedure for the CCI is described. The approach followed by Daas and Puts (2014b) to construct a sentiment index from social media platforms is also described. In section 3 a structural time series model for the CCI series and SMI series is proposed. Results obtained with the model are presented in Section 4. The paper concludes with a discussion in Section 5.

2. Data

2.1 Dutch Consumer Confidence Survey

The Consumer Confidence Index (CCI) is based on a monthly survey, called the Consumer Confidence Survey (CCS), and measures the opinion of households residing in the Netherlands about the economic climate in general and their own financial situation. The CCS is a continuous survey. Each month a self-weighted sample of approximately 2500 households is drawn by stratified two-stage sampling from a sample frame derived from the Dutch Municipal Register. Households for which a known telephone number is available are contacted by an interviewer who completes the questionnaire by computer assisted telephone interviewing during the first ten working days of the month. On average a net sample of about 1000 responding households is obtained.

The CCI is based on five questions that can be answered positively, neutral or negatively. The questions refer to the economic or financial situation in the last 12 month or the respondents expectations in the future 12 months. Let $P_{1,t}^q$, $P_{2,t}^q$, and $P_{3,t}^q$, denote the percentage of respondents that answered question $q = 1, \dots, 5$, in month t positively, neutral or negatively, respectively. Now the CCI is defined as the difference between the percentage of positive and negative respondents, averaged over the five questions:

$$I_t = \frac{1}{Q}\sum_{q=1}^{Q}\left(P_{1,t}^q - P_{3,t}^q\right). \quad (1)$$

Since the sample is self-weighted, and no auxiliary information is used in the estimation procedure, the percentages are estimated with the sample mean, i.e.

$$P_{j,t}^q = \frac{100}{n_t}\sum_{i=1}^{n}\delta_{i,j,t}^q\,, \quad (2)$$

for question $q = 1, \dots, 5$, and answer category $j = 1, 2, 3$. In (2) n_t is the net sample size in month t, and $\delta_{i,j,t}^q$ is a dummy indicator that is equal to one if respondent i chose category j to question q. Assuming simple random sampling without replacement for the households, Van den Brakel (2002) proved that the variance of (1) can be estimated by

$$\begin{aligned} Var(I_t) = {} & \frac{1}{Q^2}\sum_{q=1}^{Q}\left[Var\left(P_{1,t}^q\right) + Var\left(P_{3,t}^q\right)\right] - \frac{2}{Q^2}\sum_{q=1}^{Q}\sum_{q'=1}^{Q} Cov(P_{1,t}^q, P_{3,t}^{q'}) \\ & + \frac{1}{Q^2}\sum_{q=1}^{Q}\sum_{q'\neq q}^{Q}\left[Cov\left(P_{1,t}^q, P_{1,t}^{q'}\right) + Cov\left(P_{3,t}^q, P_{3,t}^{q'}\right)\right], \end{aligned} \quad (3)$$

with

$$Var\left(P_{j,t}^q\right) = \frac{1}{n_t}P_{j,t}^q(100 - P_{j,t}^q),$$

$$Cov\left(P_{j,t}^q, P_{j,t}^{q'}\right) = \frac{1}{n_t}(P_{jj,t}^{qq'} - P_{j,t}^q P_{j,t}^{q'}),$$

$$Cov\left(P_{j,t}^q, P_{j',t}^{q'}\right) = \frac{1}{n_t}(P_{jj',t}^{qq'} - P_{j,t}^q P_{j',t}^{q'}),$$

$$Cov\left(P_{j,t}^q, P_{j',t}^q\right) = -\frac{1}{n_t}P_{j,t}^q, P_{j',t}^q,$$

$$P_{jj',t}^{qq'} = \frac{100}{n_t}\sum_{i=1}^{n}\delta_{i,j,t}^q\delta_{i,j',t}^{q'}.$$

Figure 1 shows the CCI with a 95% confidence interval calculated using the approach described in this section, observed during the period December 2000 through March 2015. In October 2013 the official publication of the CCI is missing.

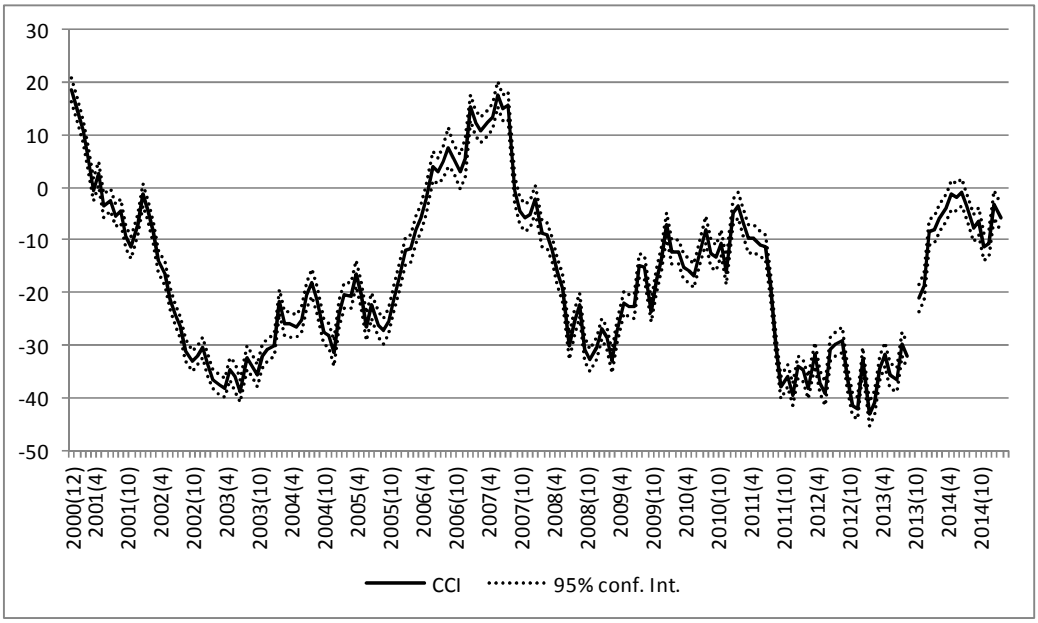


Figure 1: Consumer confidence index (CCI) with a 95% confidence interval

2.2 Social media sentiment

In an attempt to reduce administration costs and response burden, Daas and Puts (2014b) developed a sentiment index from social media sources that could be used as an alternative indicator for the CCI. They used messages posted on the most popular social media platforms in the Netherlands, written in the Dutch language. These messages are classified as containing positive, neutral, or negative messages using a variant of sentence-level based classification (Pang and Lee, 2008). An index is calculated by taking the difference between the percentage of positive and negative messages.

Combinations of all Facebook and Twitter messages with and without certain filters on phrases were compared with the CCI. The combination of all publicly available Facebook messages together with filtered Twitter messages containing personal pronouns had the highest correlation with the CCI. The Twitter messages had to be filtered due to the fact that a lot of Twitter messages are not very informative. See Daas and Puts (2014b) for further details. In their research Daas and Puts (2014b) also found that major changes in the behaviour of the public on social media, such as those caused by huge events and changes in the number of messages posted on each platform, have a disturbing effect on the series. The final indicator proposed is the average of the sentiment in the Facebook and Twitter messages during each period.

In Figure 2 the Social Media Index (SMI) is compared with the CCI for the period June 2010 until March 2015. Both series are clearly on a different level but show a more or less similar evolution. During the presented period, the CCI is always negative, while

the SMI is always positive. Many factors are responsible for this difference since the CCI is based on a survey where data collection is conducted by telephone and the SMI is based on classifying messages on Twitter and Facebook. The interesting question is to which extent the evolution of both series is similar.

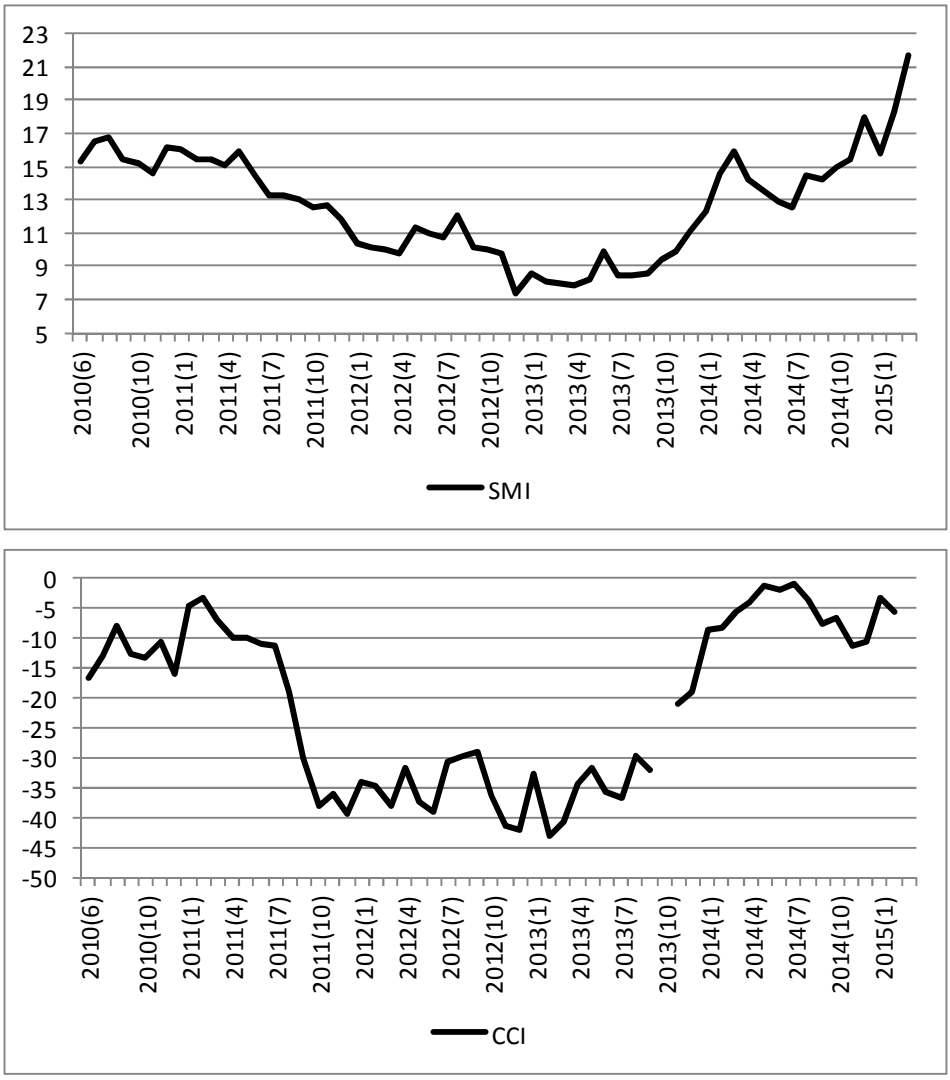


Figure 2 comparison of the Social media index (SMI, upper panel) with the Consumer confidence index (CCI, lower panel).

3. Structural time series modeling of the CCI and the SMI

In this section univariate and bivariate structural time series models for the CCI and SMI are developed. With a structural time series model a series is decomposed in a trend component, seasonal component, other cyclic components, regression component and an irregular component. For each component a stochastic model is assumed. This allows the trend, seasonal, and cyclic component but also the regression coefficients to be time dependent. If necessary ARMA components can be added to capture the autocorrelation in the series beyond these structural components. See Harvey (1989) or Durbin and Koopman (2012) for details about structural time series modelling.

The question addressed in this paper is to which extent the SMI follows a similar pattern as the CCI such that the SMI can be used in the estimation procedure of the CCI or, in the most extreme case, even can replace the CCI. This question is addressed by developing a bivariate structural time series model for the CCI and the SMI and modeling the correlation between the disturbance terms of the different components of the structural time series model for both series. The concept of cointegration is used to investigate to which extent the unobserved components of both series are driven by common factors. If e.g. the trends of both series are driven by one underlying common trend an argument can be made that the SMI represents similar evolution of sentiment feelings compared to the CCI. Alternatively, the SMI can be used as an auxiliary series in a model based estimation procedure for the CCI.

3.1 Univariate model CCI series

As a first step a univariate time series model for the CCI series is proposed. With the design-based approach described in Section 2.1 the sample information observed in each separate month is used to obtain an estimate for the CCI in that month. A drawback of this approach is that information observed in preceding periods is not used to obtain more accurate estimates for the CCI. In survey methodology, time series models are frequently applied to develop estimates for periodic surveys. Blight and Scott (1973) and Scott and Smith (1974) proposed to regard the unknown population parameters as a realization of a stochastic process that can be described with a time series model. This introduces relationships between the estimated population parameters at different time points in the case of non-overlapping as well as overlapping samples. The explicit modelling of this relationship between these survey estimates with a time series model can be used to combine sample information observed in the past to improve the precision of estimates obtained with periodic surveys. Some key references to authors that applied the time series approach to repeated survey data to improve the efficiency of survey estimates are Scott et al. (1977), Tam (1987), Binder and Dick (1989, 1990), Bell and Hillmer (1990), Tiller (1992), Rao and Yu (1994), Pfeffermann and Burck (1990), Pfeffermann (1991), Pfeffermann and Bleuer (1993), Pfeffermann et al. (1998), Pfeffermann and Tiller (2006), Harvey and Chung (2000), Feder (2001), Harvey and Chung (2000) and Van den Brakel and Krieg (2009, 2015).

Developing a time series model for survey estimates observed with a periodic survey starts with a model, which states that the survey estimate can be decomposed in the value of the population variable and a sampling error:

$$I_t = \theta_t + e_t, \tag{4}$$

where θ_t denote the real CCI in month t under a complete enumeration of the target population and e_t the sampling error.

The CCI is observed at a monthly frequency. Therefore, as a first step, the series of the finite population parameter can be decomposed in a stochastic trend, seasonal component to model systematic deviations from the trend within a year, and a white noise component for the remaining unexplained variation. These considerations lead to the following model for the series of the finite population parameter:

$$\theta_t = L_t + S_t + \xi_t, \tag{5}$$

where L_t denotes a stochastic trend, S_t a stochastic seasonal component and ξ_t the unexplained variation of the finite population parameter. Inserting (5) into measurement model (4) gives

$$I_t = L_t + S_t + \xi_t + e_t, \tag{6}$$

In a cross-sectional survey it is difficult to separate the sampling error from the white noise of the population parameter. Therefore both components are combined in one disturbance term

$$\upsilon_t = \xi_t + e_t. \tag{7}$$

It is assumed that $E(\upsilon_t) = 0$. To allow for nonhomogeneous variance in the sampling errors it is assumed that the variance of υ_t is proportional to the sampling variance of I_t, i.e.

$$Var(\upsilon_t) = Var(I_t)\sigma_\upsilon^2 \quad , \tag{8}$$

where $Var(I_t)$ is defined by (3) and is used as a-priori information in the time series model.

An extensive model selection showed that a smooth trend model is the most appropriate model to capture the trend and the economic cycle in the CCI series. The smooth trend model is defined as (Durbin and Koopman, 2012):

$$\begin{aligned} L_t &= L_{t-1} + R_t, \\ R_t &= R_{t-1} + \eta_t \quad , \\ \mathrm{E}(\eta_t) &= 0, \\ Cov(\eta_t, \eta_{t'}) &= \begin{cases} \sigma_\eta^2 & \text{if} \quad t = t' \\ 0 & \text{if} \quad t \neq t' \end{cases}. \end{aligned} \tag{9}$$

The seasonal component is modelled with a trigonometric model, which is defined as (Durbin and Koopman, 2012):

$$S_t = \sum_{j=1}^{6} \gamma_{jt}, \tag{10}$$

with

$$\gamma_{jt} = \gamma_{jt-1}\cos(\lambda_j) + \tilde{\gamma}_{jt-1}\sin(\lambda_j) + \omega_{jt},$$

$$\tilde{\gamma}_{jt} = -\gamma_{jt-1}\sin(\lambda_j) + \tilde{\gamma}_{jt-1}\cos(\lambda_j) + \tilde{\omega}_{jt}.$$

Here λ_j denotes the frequency of the different cycles in radians and is defined as

$$\lambda_j = \frac{2\pi j}{12}, \text{ for } j=1,\ldots,6.$$

For the disturbance terms, it is assumed that

$$E(\omega_{jt}) = 0, E(\tilde{\omega}_{jt}) = 0,$$

$$Cov(\omega_t, \omega_{t'}) = \begin{cases} \sigma_\omega^2 & \text{if} \quad t = t' \\ 0 & \text{if} \quad t \neq t' \end{cases}.$$

For reasons of parsimony, the same variance structure is assumed with the same hyperparameter for $\tilde{\omega}_{jt}$. Furthermore it is assumed that ω_t and $\tilde{\omega}_t$ are uncorrelated. To model unexplained serial correlation in the residuals v_t, AR and MA components can be added to the structural time series model (6). In this application there were no indications that such components are required. The model selection procedure however indicated that two level interventions are needed to model sudden jumps in the series. The first one is due the financial crisis in September 2008, and the second one is due to the economic downturn in September of 2011. Finally an outlier is required for September 2007. These considerations lead to the following model for the observed CCI series

$$I_t = L_t + S_t + \beta^{07}\delta_t^{07} + \beta^{08}\delta_t^{08} + \beta^{11}\delta_t^{11} + v_t, \qquad (11)$$

with

$$\delta_t^{07} = \begin{cases} 1 & \text{if} \quad t = 2007(9) \\ 0 & \text{if} \quad t \neq 2007(9) \end{cases}, \quad \delta_t^{08} = \begin{cases} 1 & \text{if} \quad t \geq 2008(9) \\ 0 & \text{if} \quad t < 2008(9) \end{cases},$$

$$\delta_t^{11} = \begin{cases} 1 & \text{if} \quad t \geq 2011(9) \\ 0 & \text{if} \quad t < 2011(9) \end{cases},$$

and β^x the corresponding regression coefficients.

3.2 Bivariate model CCI and SMI series

The next step is to combine the univariate model for the CCI with the series for the SMI. The main purpose of combining the series of the CCI with the series of the SMI is to investigate to which extent both series are cointegrated. The most straightforward approach to combine the series of the CCI with the series of the SMI is to extend (11) with an additional regression component, say βX_t, where X_t is the observed series of the SMI and β the corresponding regression component. This approach is not useful in this application since investigating whether both series are cointegrated requires a bivariate model that allows for correlation between slope disturbances of the trend and disturbances of the seasonal components. An additional drawback of adding the SMI series as a regression component in (11) is that the auxiliary series will partially explain the trend and seasonal effect in I_t leaving only a residual trend and seasonal. This hampers the estimation of a trend or a seasonal component for the CCI. The trend could be considered as an alternative for a seasonal adjusted release for the CCI.

Before combining CCI and SMI in a bivariate model, a univariate model for the SMI is developed with the purpose to better understand the behaviour of this series. An extensive model selection procedure indicated that the observed series for the SMI can

be modelled with a smoothed trend and a white noise component for the unexplained variation. No significant seasonal component is established. These considerations led to a bivariate model for the CCI and SMI where the CCI contains a trend and a seasonal component and the SMI a trend component. The disturbance terms of the trend of both series are correlated. Since the series for the SMI is available from June 2010, the model for the CCI also contains the last intervention for September 2011, but not the outlier in September 2007 and the intervention in September 2008. As a result the following bivariate model is obtained:

$$\begin{pmatrix} I_t \\ X_t \end{pmatrix} = \begin{pmatrix} L_t^I \\ L_t^X \end{pmatrix} + \begin{pmatrix} S_t^I \\ \end{pmatrix} + \begin{pmatrix} \beta^{11}\delta_t^{11} \\ \end{pmatrix} + \begin{pmatrix} v_t^I \\ v_t^X \end{pmatrix}, \quad (12)$$

with L_t^I and L_t^X the smooth trend model as defined in (9) with covariance structure

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

$$Cov(\eta_t^I, \eta_{t'}^X) = \begin{cases} \sigma_{\eta I}\sigma_{\eta X}\rho_\eta & \text{if} \quad t = t' \\ 0 & \text{if} \quad t \neq t' \end{cases}.$$

In the last expression ρ_η denotes the correlation between the slope disturbances of the CCI and SMI. Furthermore, S_t^I is the seasonal effect defined by (10) and δ_t^{11} the intervention for September 2011 with β^{11} the corresponding regression coefficient. Finally v_t^I and v_t^X are the disturbance terms for the CCI and SMI series and are defined as:

$$E(v_t^I) = E(v_t^X) = 0,$$

$$Cov(v_t^I, v_{t'}^I) = \begin{cases} Var(I_t)\sigma_{vI}^2 & \text{if} \quad t = t' \\ 0 & \text{if} \quad t \neq t' \end{cases}, \; Cov(v_t^X, v_{t'}^X) = \begin{cases} \sigma_{vX}^2 & \text{if} \quad t = t' \\ 0 & \text{if} \quad t \neq t' \end{cases},$$

$Cov(v_t^I, v_{t'}^X) = 0$ for all t and t'.

If the model detects a strong correlation between the trends of the CCI and the SMI, then the trends of both series will develop into the same direction more or less simultaneously. In this case the additional information from the SMI series will result in an increased precision of the estimates of the CCI figures. In the case of strong correlation between the disturbances of the trends, i.e. if $\rho_\eta \to 1$, the trends are said to be cointegrated. In that case there is one underlying common trend that drives the evolution of the trends of the two observed series. To see this, it is noted that the covariance matrix of the slope disturbances is implemented as a Choleski decomposition:

$$cov\begin{pmatrix} \eta_t^I \\ \eta_t^X \end{pmatrix} = \begin{pmatrix} \sigma_{\eta I}^2 & \sigma_{\eta I}\sigma_{\eta X}\rho_\eta \\ \sigma_{\eta I}\sigma_{\eta X}\rho_\eta & \sigma_{\eta X}^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ a & 1 \end{pmatrix}\begin{pmatrix} d_1 & 0 \\ 0 & d_2 \end{pmatrix}\begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix}. \quad (13)$$

Instead of estimating $\sigma_{\eta I}^2$, $\sigma_{\eta X}^2$, and ρ_η , parameters d_1, d_2, and a are estimated. If $d_2 \to 0$, it follows that $\rho_\eta \to 1$. In that case the covariance matrix of the slope disturbances is of reduced rank and both trends are driven by one common trend. This implies that the slope disturbances of both series simultaneously move up or down and that the slope disturbances of the SMI can be perfectly predicted from slope disturbances of the CCI by $\eta_t^X = a\eta_t^I$. Furthermore, the slope for the SMI series can be expressed as a linear combination of the slope for the CCI series as $R_t^X = aR_t^I + \bar{R}$. Similarly the trend for the SMI series can be expressed as a linear combination of the

trend for the CCI series as $L_t^X = aL_t^I + \bar{L} + \bar{R}t$. Note that $\bar{R}$ and $\bar{L}$ are constants that are derived from the estimated states at the last two time periods of the series.

Cointegration increases the precision of the estimated trend and signal of the CCI series, allows for formulating more parsimonious models, but could also be seen as an argument to replace the CCI series by the SMI series since both series are driven by and represent the same common trend. For a more detailed discussion about cointegration in the context of state-space modelling, see Koopman, Harvey, Shephard and Doornik (2009, sections 6.4 and 9.1).

3.3 Estimation of structural time series models

The general way to analyse a structural time series model, is to express it in the so-called state space representation and apply the Kalman filter to obtain optimal estimates for the state variables, see e.g. Durbin and Koopman (2012). The software for the analysis and estimation of the time series models is developed in Ox in combination with the subroutines of SsfPack 3.0, see Doornik (2009) and Koopman, Shephard and Doornik (2008).

All state variables are non-stationary and initialised with a diffuse prior, i.e. the expectation of the initial states are equal to zero and the initial covariance matrix of the states is diagonal with large diagonal elements. In Ssfpack 3.0 an exact diffuse log-likelihood function is obtained with the procedure proposed by Koopman (1997). Maximum likelihood estimates for the hyperparameters, i.e. the variance components of the stochastic processes for the state variables are obtained using a numerical optimization procedure (BFGS algorithm, Doornik, 1998). To avoid negative variance estimates, the log-transformed variances are estimated. More technical details about the analysis of state-space models can be found in Harvey (1989) or Durbin and Koopman (2001).

Under the assumption of normally distributed disturbance terms, the Kalman filter can be applied to obtain optimal estimates for the state variables, see e.g. Durbin and Koopman (2012). The Kalman filter assumes that the variance and covariance terms are known in advance and are often referred to as hyperparameters. In practise these hyperparameters are not known and are therefore substituted with their ML estimates. Estimates for state variables for period *t* based on the information available up to and including period *t* are referred to as the *filtered estimates*. They are obtained with the Kalman filter where the ML estimates for the hyperparameters are based on the complete time series. The filtered estimates of past state vectors can be updated, if new data become available. This procedure is referred to as smoothing and results in *smoothed estimates* that are based on the complete time series.

Standard errors of the Kalman filter estimates do not reflect the additional uncertainty of using the ML estimates for the unknown hyperparameters. Therefore the estimates of the standard errors are too optimistic.

4. Results

4.1 Univariate model CCI series

The univariate analysis is based on model (11) from Section 3.1 applied to the series of the CCI obtained from December 2000 until March 2015. In Table 1 the ML estimates for the hyperparameters of the model are specified.

In the upper panel of Figure 3, the smoothed trend plus interventions are compared with the direct estimates for the CCI. In the lower panel of Figure 3, the smoothed signal, defined as trend plus interventions plus seasonals, are compared with the direct estimates for the CCI. In the series of the smoothed trend and interventions, the seasonal effect, the white noise of the population parameter and the sampling error is removed from the original series. It follows from Figure 3 that with the time series model a more stable estimate for the CCI can be obtained. The filtered trend plus interventions is compared with the smoothed estimates in Figure 4. This filtered series approximate what would be obtained in the production of official statistics if no revisions would be published. It follows that even in this case a considerable part of the high frequency variation and seasonal fluctuations can be removed.

Figure 5 shows the smoothed seasonal pattern of the CCI series. It follows from the upper panel that the pattern is time invariant. In the lower panel the seasonal effects for the months in 2014 are shown. There are clear significant negative effects in the October, November and December and clear positive effects in January and August. Given the definition of the consumer confidence index and the way it is operationalized in the questionnaire, it is remarkable that there is a clear significant seasonal pattern. All questions refer to the respondents financial and economic situation over the last 12 month or the expectations for the future 12 months. If respondents would interpret the questions as intended, a pronounced seasonal pattern would not occur. This is an indication that answers given by the respondents are clearly driven by a much shorter emotion, which is among other things subject to seasonal fluctuations.

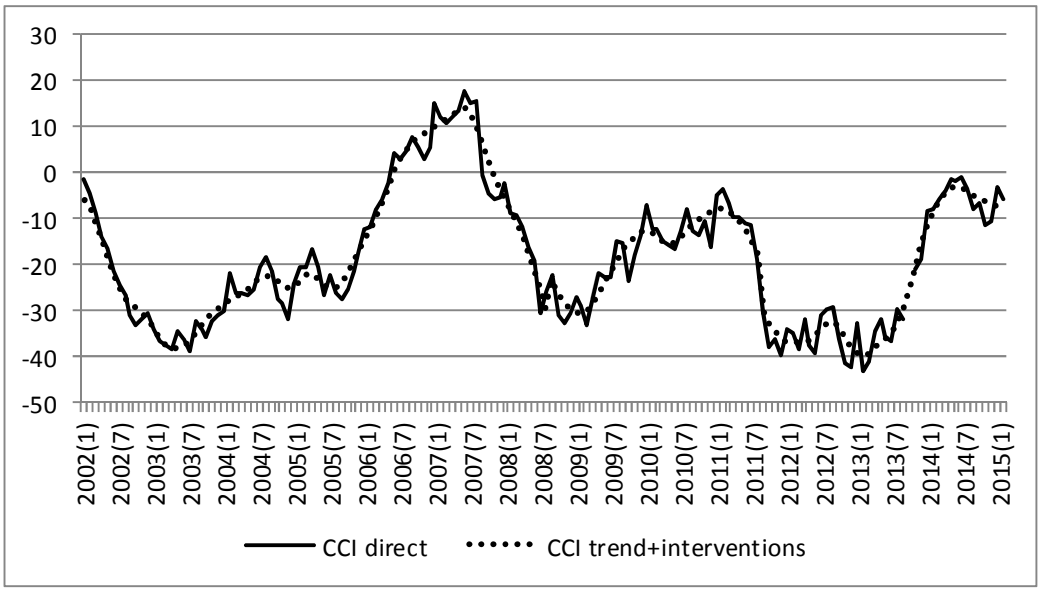
In Figure 6 the standard error of the direct estimates for the CCI are compared with the standard errors of the filtered and smoothed trend plus interventions. The spikes in the standard error of the filtered and smoothed estimates are the result of the intervention variables and the missing observation in 2013. If at a certain point in time an intervention variable is activated, a new regression coefficient has to be estimated. This results in additional uncertainty in the model estimates, and shows up as a sudden peak in the standard error of the filtered and smoothed trend. In 2013 one observation is missing, which also results in additional uncertainty since the state space model produces a prediction for this missing value.

The standard errors of the smoothed estimates are slightly larger than the standard errors of the direct estimates. The standard errors of the filtered estimates are considerably larger than the standard errors of the direct estimates. This is a remarkable result. Filtered and smoothed estimates based on the time series model are based on a considerably larger set of information since sample information from preceding periods (in the case of filtered estimates) or the entire series (in the case of smoothed estimates) are used to obtain an optimal estimate for the monthly CCI. The direct estimates, on the other hand, are based on the observed sample in that

particular month only. Most applications where structural time series models are applied as a form of small area estimation, result in substantive reductions of the standard error compared to the direct estimates, see e.g. Van den Brakel and Krieg (2009, 2015) and Bollineni-Balabay, van den Brakel and Palm (2015).

Standard deviation	ML estimate
Trend (σ_η)	1.18
Seasonal (σ_ω)	0.0025
Measurement equation (σ_υ)	2.01

Table 1: Maximum Likelihood estimates hyperparameters univariate model CCI



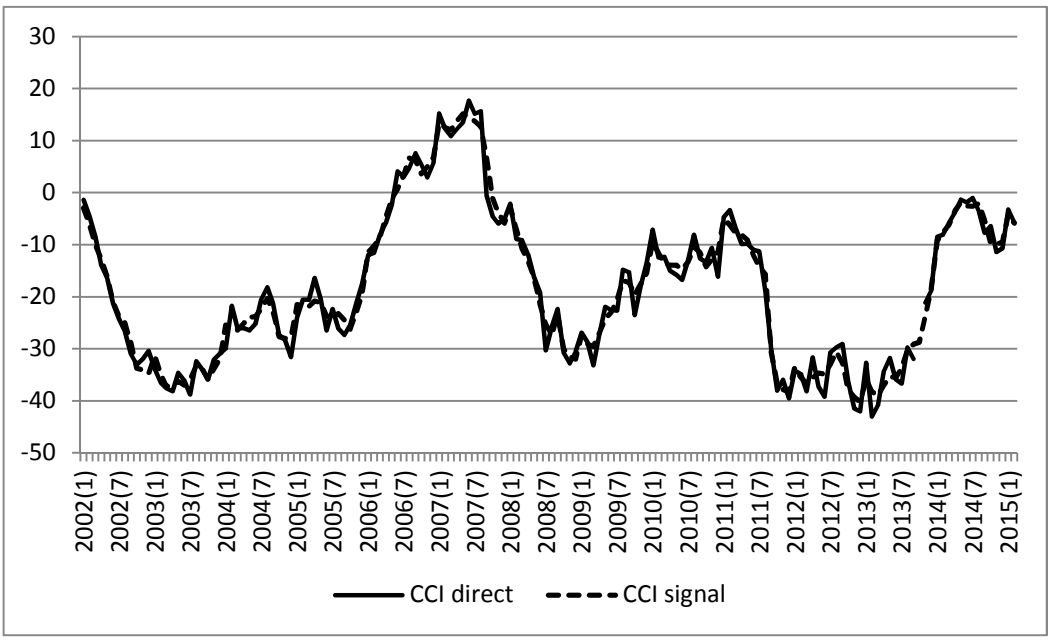


Figure 3: smoothed trend plus interventions compared with direct estimates CCI (upper panel) and smoothed signal compared with direct estimator (lower panel)

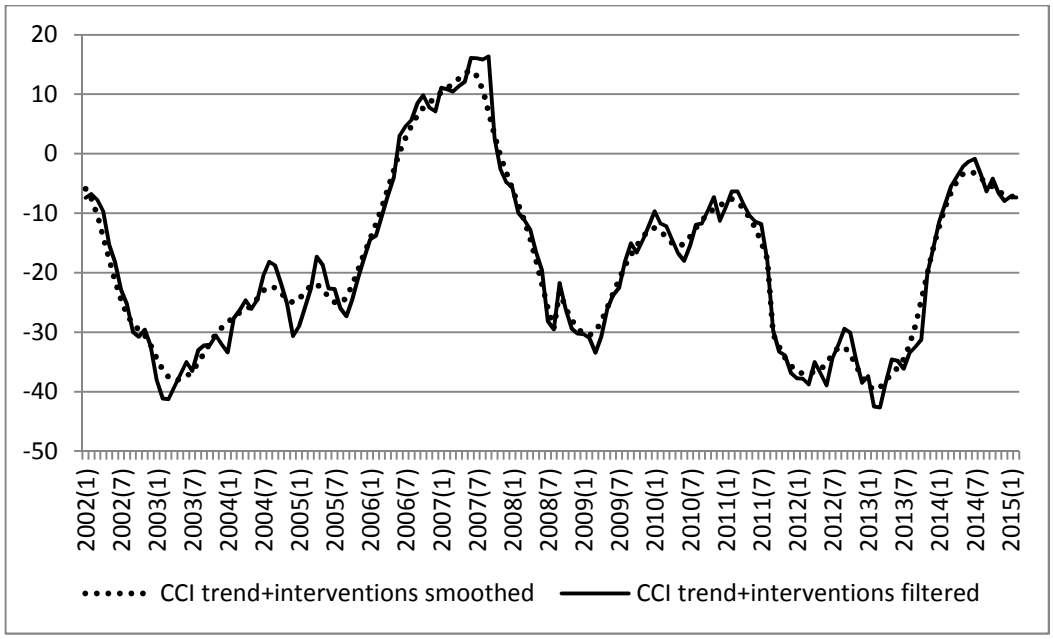
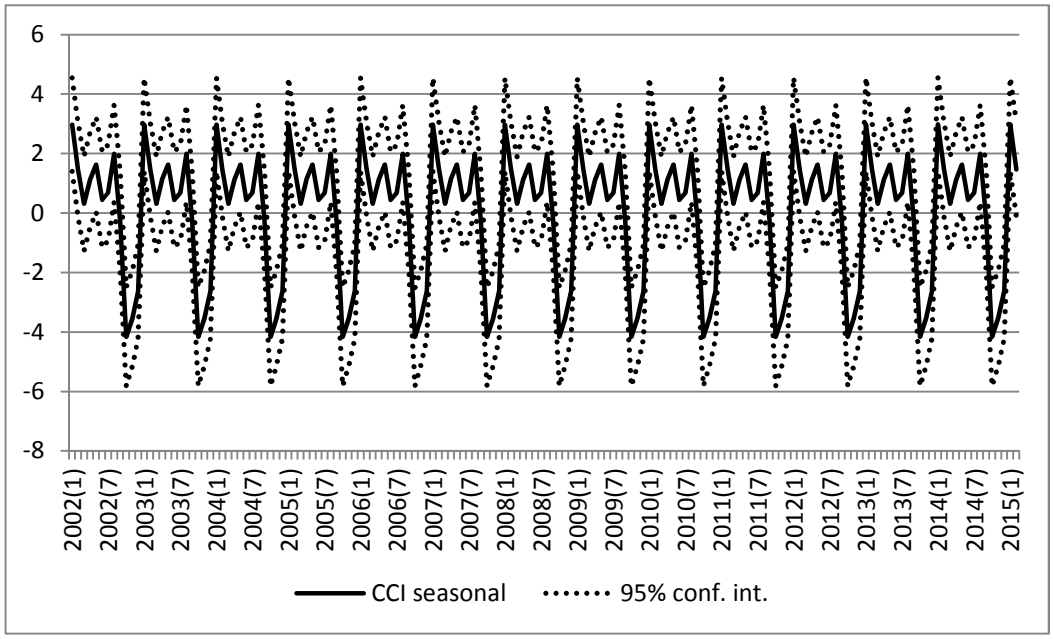


Figure 4: Filtered trend plus interventions compared with smoothed trend plus interventions CCI



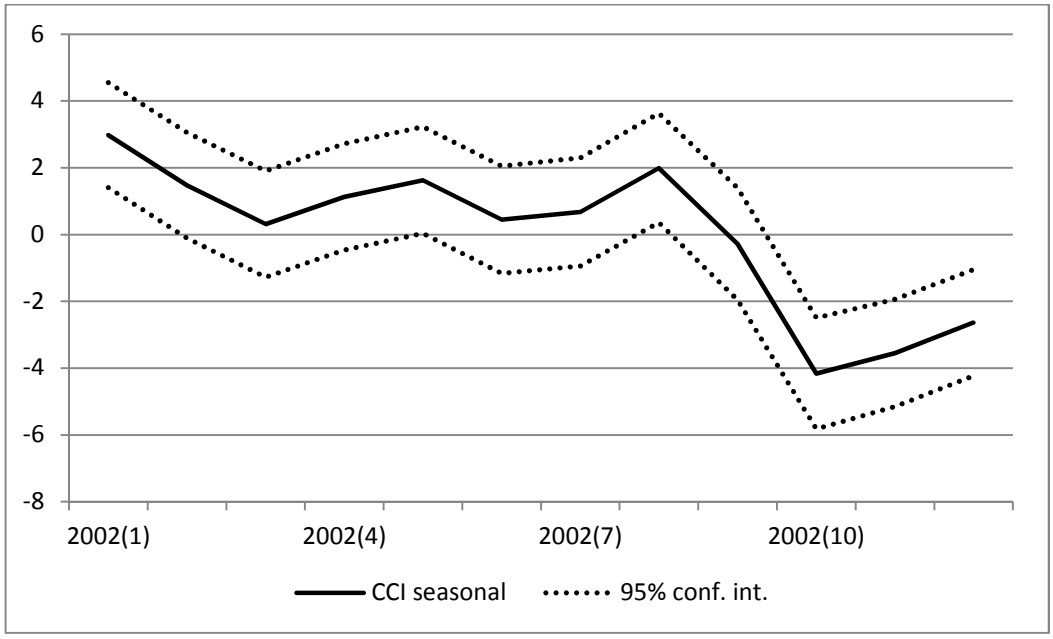


Figure 5: Smoothed seasonal pattern CCI for the complete series (upper panel) and the months of 2014 (lower panel).

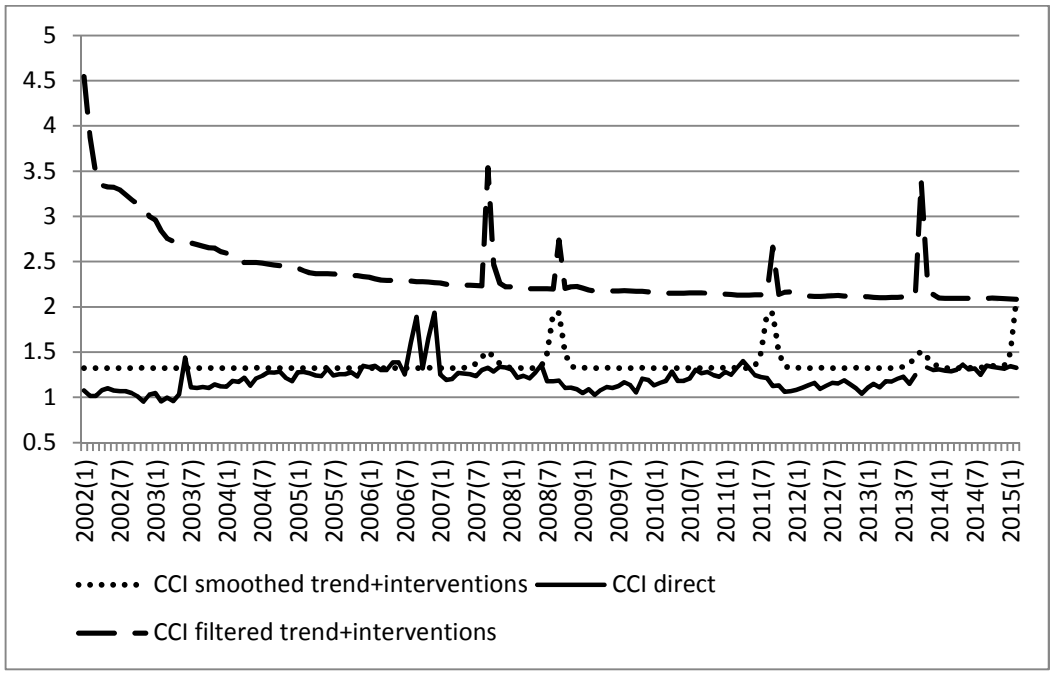


Figure 6: Standard error smoothed and filtered trend plus interventions compared with direct estimates CCI

The reason that in this application a time series modelling approach results in standard errors for filtered and smoothed times series model estimates that are larger than the standard errors of the direct estimates is a result of a large white noise component in the real population value of the CCI. Recall from Section 3.1 that the disturbance term of (11) contains two components; the sampling error and the unexplained high frequency variation of the real population value, as expressed by (7). If the sampling error dominates the variance of the disturbance term and modelling the variance of

the disturbance term proportional to the variance of the direct estimates, as expressed by (8), would result in a maximum likelihood estimate for σ_v that is approximately equal to one. From Table 1 it follows that σ_v is equal to 2. This is a strong indication that the variance of the white noise component in the true population variable exceeds the variance of the sampling error. The direct estimator for the CCI derived in Section 2 considers the CCI in each particular month as a fixed but unknown variable. The variance of the direct estimator only measures the uncertainty since a small sample instead of the entire population is observed to estimate the CCI. It does not measure the high frequency variation of the population value over time. This explains why the time series modelling approach does not result in a reduction of the standard error of the estimated CCI. A time series model is, however, still useful to filter a more stable long term trend from the high frequency variation in the population parameter and the sampling error.

4.2 Bivariate model for CCI and SMI series

In this section, the bivariate model (12) proposed in Section 3.2 is applied to the series of the CCI and SMI, which are available from June 2010 until March 2015. Maximum likelihood estimates for the hyperparameters are specified in Table 2. The model detects a strong positive correlation of about 0.92 between the slope disturbances of the CCI and the SMI. There is, however, no indication that both trends are cointegrated and share one common trend. A likelihood ratio test is applied to further investigate the significance of the correlation between the slope disturbances in the bivariate model. If the correlation parameter is set to zero, the log likelihood drops from -229.9 to -233.9. The p-value of the corresponding likelihood ratio test equals 0.0047, indicating that the correlation between the trends of both series is clearly significantly different from zero and should not be removed from the bivariate model. If the correlation parameter is set equal to one (by choosing d_2 in (13) equal to zero), the log likelihood drops from -229.9 to -242.1. The p-value of the corresponding likelihood ratio test with one degree of freedom equals zero, indicating that the trends are not cointegrated.

Standard deviation	ML estimate
Trend CCI ($\sigma_{\eta I}$)	1.24
Seasonal CCI (σ_{ω})	7.5E-6
Trend SMI ($\sigma_{\eta X}$)	0.25
Measurement equation CCI (σ_{vI})	2.28
Measurement equation SMI (σ_{vX})	0.86
Correlation trend CCI and SMI (ρ_{η})	0.92

Table 2: Maximum Likelihood estimates hyperparameters bivariate model CCI and SMI

Figure 7 compares the smoothed estimates for the slope of the CCI (x axis) and SMI (y axis) under the model without correlation, the model with an ML estimate for the correlation ($\rho_{\eta} = 0.92$) and the common trend model with $\rho_{\eta} = 1.0$. The model with uncorrelated slopes shows a clearly positive correlation between the slopes if both series are estimated independently (left panel Figure 7). This is picked up by the model

that allows for correlation (mid panel Figure 7). There is however a clear deviation between the slopes of both series, which can be seen if the cross-plot of the model with a correlation estimated with ML (mid panel Figure 7) is compared with the cross-plot of a common factor model (right-panel Figure 7).

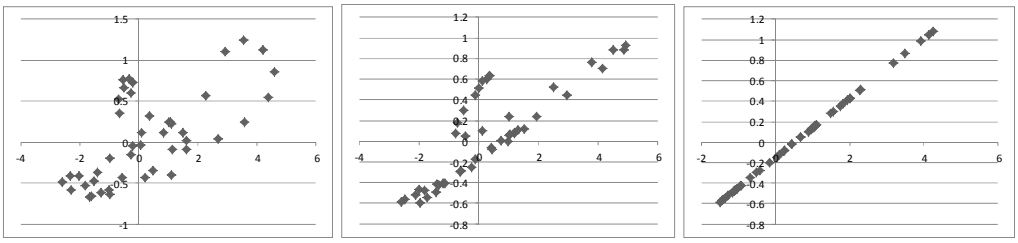


Figure 7: Cross-plot smoothed slopes CCI (x-axis) and SMI (y-axis) for a model without correlation (left panel), correlation estimated with ML (mid panel) and correlation set equal to one (right panel)

Figure 8 compares the observed SMI series with the smoothed trend obtained under the bivariate model. Figure 9 compares the direct estimates for the CCI series with the smoothed trend plus intervention under the univariate model and the bivariate model. As follows from Figure 9, the level and evolution of the smoothed estimates for the CCI series are almost identical under the univariate and bivariate model.

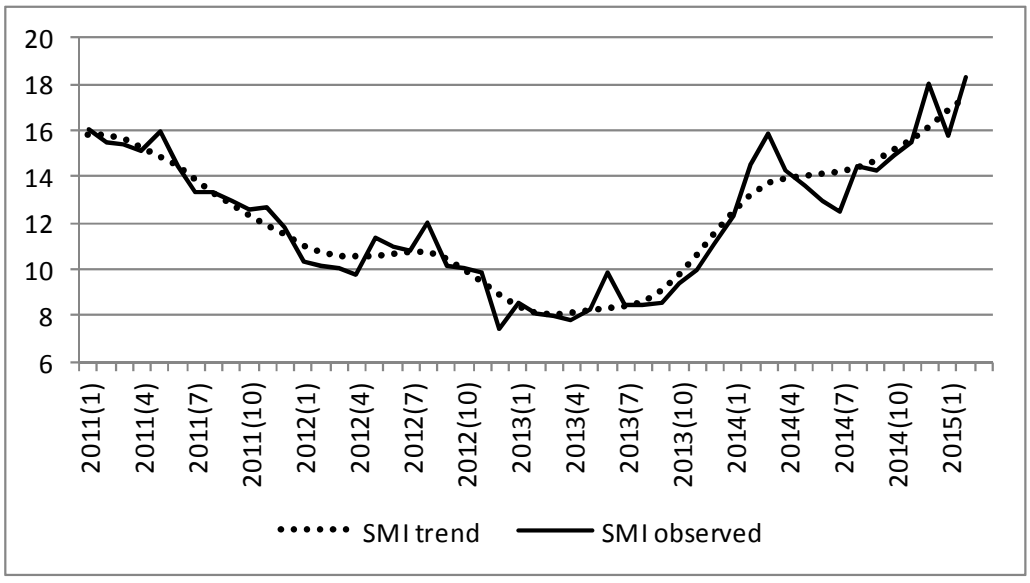


Figure 8: Observed series and smoothed trend SMI

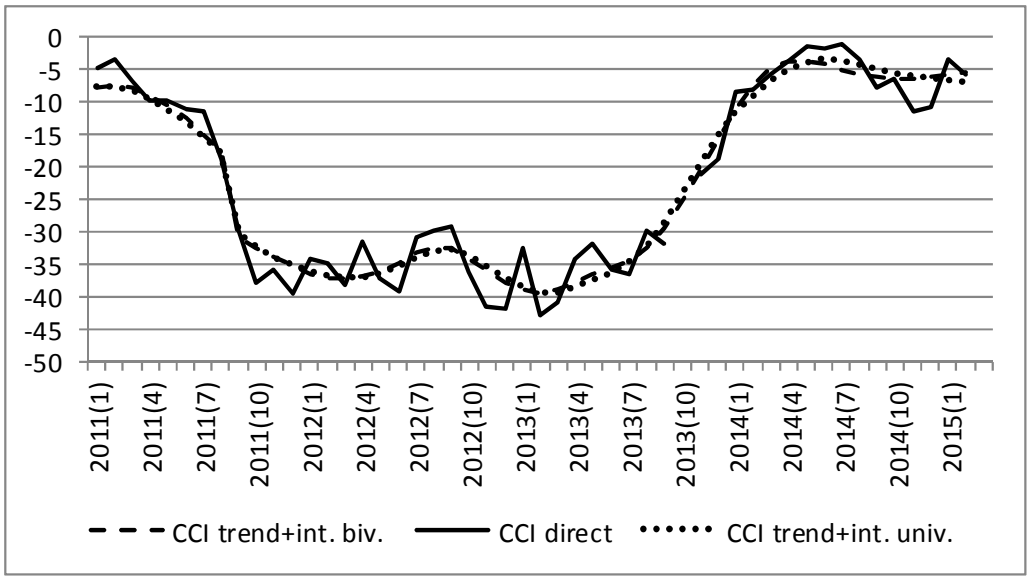


Figure 9: CCI comparison of the direct estimates and smoothed trend plus intervention under the bivariate and univariate model for CCI

Figure 10 compares the standard errors of the direct estimates for the CCI series with the smoothed trend plus intervention under the univariate model and the bivariate model. For a fair comparison, the results for the univariate model and bivariate model are based on series of equal length. Therefore the univariate model is re-estimated with the series from June 2010 until March 2015. As follows from Figure 10, the standard error under the bivariate model is slightly smaller compared to the standard error under the univariate model if both models are applied to series of equal length, as expected given the strong and significant positive correlation between the trend disturbance terms of both series. If, however, the univariate model is applied to the series available from December 2000, then the standard errors for the smoothed estimates under the univariate model are slightly smaller compared to the bivariate model as follows from Figure 11.

In conclusion it follows that the bivariate model detects a strong correlation between the CCI and SMI series. Using the SMI series as an auxiliary series slightly improves the precision of the model based estimates for the CCI. Since the series of the CCI is nine years longer than the SMI series, the increased precision obtained with the auxiliary series is compensated in the univariate model with the additional information in the CCI series available before 2010.

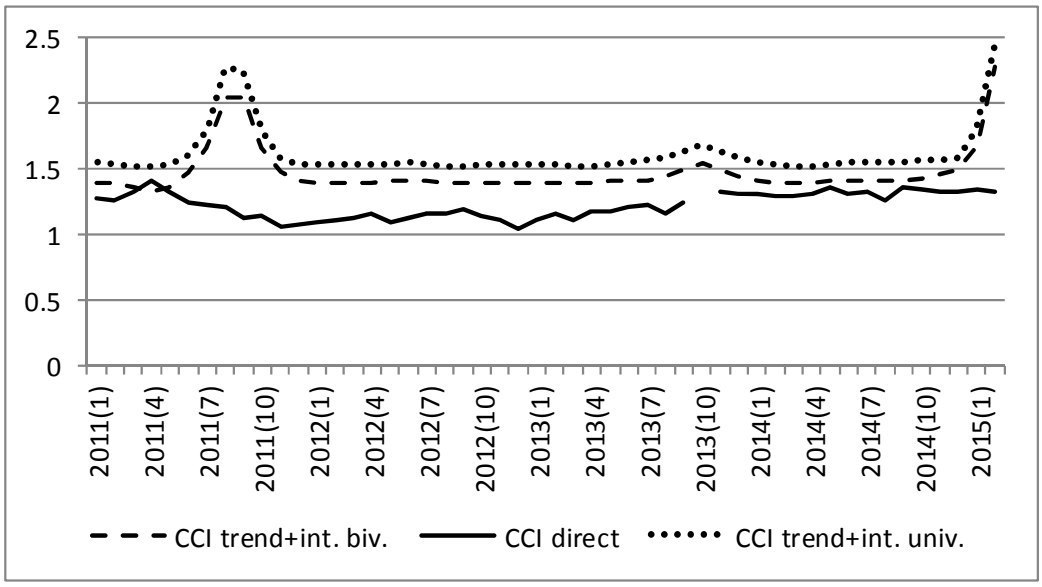


Figure 10: CCI comparison of standard errors direct estimates and smoothed trend plus intervention under the bivariate and univariate model for CCI if both models are applied to a series of equal length (June 2010-March 2015)

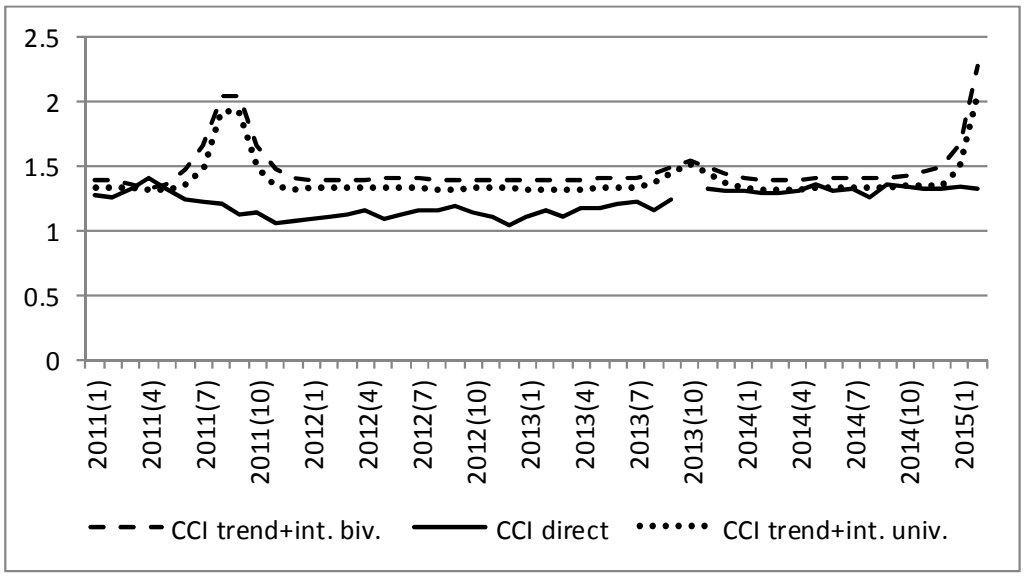


Figure 11: CCI comparison of standard errors direct estimates and smoothed trend plus intervention under the bivariate and univariate model for CCI if the univariate model is applied to the complete CCI series (December 2000)

5. Discussion

For decades national statistical institutes relied on probability sampling in the production of official statistics. This approach is based on a sound theory to draw valid statistical inference for large finite target populations based on relatively small random samples. Over the last decades more and more alternative data sources, such as administrative and big data, have become available and the question is raised how to use these data sources in the production of official statistics. An important question is how results obtained with these sources can be generalized to an intended finite target population. Since the data generating process is generally unknown, it is not obvious how to draw valid inference with such data sources.

In this paper the question is addressed how administrative and big data sources can be used in the production of official statistics. In the most extreme approach survey data are replaced by related alternative data sources, running the risk of introducing e.g. selection bias. Since most surveys are conducted repeatedly, a time series modelling approach is proposed to investigate to which extent related alternative data sources reflect a similar evolution compared to the series obtained with a repeated survey. With a multivariate state space model, the correlation between the underlying unobserved components of both series can be modelled. In the case that components of the time series model are cointegrated there are strong indications that both data sources are driven by the same underlying factor. This could be used as an argument that an alternative source can replace existing surveys since they reflect the same evolution of a process, generally at a different level.

The theory underlying probability sampling for finite population inference is stronger than reliance on the concept of cointegration. Series obtained from social media are selected by maximizing the correlation with the series from the sample survey. There is no guarantee that this correlation is based on true causality and that the correlation will remain to exist in the future. Sampling theory, in contrast, provides a rigid mathematical theory showing that under a correct sampling strategy, i.e. the right combination of a probability sample with an approximately design-unbiased estimator, results in valid statistical inference for intended target populations. There are of course also issues with survey sampling. For exampling the continuously declining response rates undermine the validity of this approach. Another problem is mode related measurement bias, which makes data obtained in mixed mode surveys less comparable.

Even in the case of cointegrated series, an extensive model evaluation, e.g. by some form of cross validation, will be required to assure that the alternative data source is a valid replacement. See in this context also Eichler (2013) for a discussion about the use of Granger causality for causal inference in multiple time series data. Instead of replacing a repeated survey for related data sources, they can be used in a multivariate time series modelling approach as an auxiliary series to improve the precision of the direct estimates obtained with a repeated survey. The time series model applied in this paper, initially proposed by Harvey and Chung (2000), is a generic approach for a model-based estimation procedure for repeated surveys.

In the application to the CCI, the time series modelling approach does not decrease the variance of the direct estimator. The reason is that the standard error of the time

series model reflects the sampling error and the white noise of the population parameter. The standard error of the direct estimator only reflects the sampling error. In the case of the CCI the variance component of the white noise of the population parameter is as large as the variance of the sampling error. The state space approach is still useful for producing official figures of the CCI, since it filters a more stable trend of the respondents opinion about the economic climate from the observed series of direct estimates. Using the SMI as an auxiliary series in a bivariate state space model slightly reduces the standard error of the model estimates of the CCI. However, since the available series of the SMI is relative short, the reduction obtained with this auxiliary series does not outweigh the loss of information in the CCI series that is observed in the period before the SMI became available. From this, it can be concluded that the CCI and SMI measure different phenomena. However since both series reflect a similar evolution and social media is rapidly available, the SMI could be an interesting index to get an indication of the sentiment of the Dutch population in near real time.

6. References

Baker R., Brick, J.M., Bates, N.A., Battaglia, M., Couper, M.P., Dever, J.A., Gile, K.J. and Tourangeau, R. (2013). Summary Report of the AAPOR Task Force on Non-probability Sampling. *Journal of Survey Statistics and Methodology* 1: 90–143, first published online September 26, 2013 doi:10.1093/jssam/smt008.

Bell, W.R., and Hillmer, S.C. (1990). The time series approach to estimation of periodic surveys. *Survey Methodology*, 16, pp. 195-215.

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

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

Blight, B.J.N., and Scott, A.J. (1973). A stochastic model for repeated surveys. *Journal of the Royal Statistical Society*, *Series B*, 35, 61-66.

Blumenstock, J., Cadamuro, G. and On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, 350, pp. 1073-1076.

Bollineni-Balabay, O. Brakel, J.A. van den and Palm, F. (2015). Multivariate state-space approach to variance reduction in series with level and variance breaks due to sampling redesigns, accepted for publication in Journal of the Royal Statistical Society, A series.

Bowley, A.L. (1926). Measurement of the precision attained in sampling. *Bulletin de l' Institute International de Statistique* 22**,** Supplement to Book 1: 6–62.

Buelens, B., J. Burger, and J. van den Brakel. Predictive inference for non-probability samples: a simulation study. Discussion paper 2015-13, Statistics Netherlands, Heerlen.

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

Daas, P. and Puts, M. (2014a). Big data as a source of statistical information. *The Survey Statistician,* 69, pp. 22-31.

Daas, P. and M. Puts (2014b). Social Media Sentiment and Consumer Confidence. European Central Bank Statistics Paper Series no. 5, Frankfurt Germany.

Daas, P., Roos, M., Van de Ven, M., and Neroni, J. (2012). Twitter as a potential data source for statistics. Discussion paper 2012-21, Statistics Netherlands, Heerlen.

Doornik, J.A. (2009). An Object-oriented Matrix Programming Language Ox 6. London: Timberlake Consultants Press.

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

Eichler, M. (2013). Causal inference with multiple time series: principles and problems. *Philosophical transactions of the Royal Statistical Society A*, 371, issue 1997.

Feder, M. (2001). Time series analysis of repeated surveys: the state-space approach. *Statistica Neerlandica*, 55, pp. 182-199.

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

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

Hansen, M.H. and Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics* 14: 333–362.

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

Koopman, S.J., Shephard, N. and Doornik, J.A. (2008). *SsfPack 3.0: Statistical algorithms for models in state space form*, London: Timberlake Consultants Press.

Koopman, S.J. (1997). Exact initial Kalman filtering and smoothing for non-stationary time series models. *Journal of the American Statistical Association*, 92, 1630-1638.

Lind, J.T. (2005). Repeated surveys and the Kalman filter. *Econometrics Journal,* 8, 418-427.

Marchetti, S., Giusti, C., Pratesi, M., Salvati, N., Giannotti, F., Perdreschi, D., Rinzivillo, S., Pappalardo, L., and Gabrielli, L. (2015). Small area model-based estimators using Big data sources. Journal of Official Statistics, 31, 263-281.

Pang, B., and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2, 1-135.

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

Pfeffermann, D., and Bleuer, S.R. (1993). Robust joint modelling of labour force series of small areas. *Survey Methodology*, 19, pp. 149-163.

Pfeffermann, D., and Burck, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, 16, pp. 217-237.

Pfeffermann, D., Feder, M., and Signorelli, D. (1998). Estimation of autocorrelations of survey errors with application to trend estimation in small areas. *Journal of Business & Economic Statistics*, 16, pp. 339-348.

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

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society* 97: 558–625.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling.* Springer-Verlag.

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

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

Tam, S.M. (1989). Analysis of repeated surveys using a dynamic linear model. *International Statistical Review*, 55, 63-73.

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

Van den Brakel, J.A. (2002). Varianties voor het Consumenten Conjunctuur Onderzoek en een experiment naar de effecten van een herziene vragenlijst. Centraal Bureau voor de Statistiek, Heerlen, Interne nota, BPA nr. 2002-02-TMO.

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

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

Explanation of symbols

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

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

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

Prepress
Studio BCO, Den Haag

Design
Edenspiekermann

Information
Telephone +31 88 570 7070
Via contact form: www.cbs.nl/information