Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

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


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

Release date: December 21, 2017

Statistics Canada Statistique Canada

Canada

## How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

**email at** STATCAN.infostats-infostats.STATCAN@canada.ca

**telephone,** from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-877-287-4369

**Depository Services Program**

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

## Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

## Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

## Standard table symbols

The following symbols are used in Statistics Canada publications:

. not available for any reference period
.. not available for a specific reference period
... not applicable
0 true zero or a value rounded to zero
$0^s$ value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
$^p$ preliminary
$^r$ revised
x suppressed to meet the confidentiality requirements of the *Statistics Act*
$^E$ use with caution
F too unreliable to be published
* significantly different from reference category ($p < 0.05$)

Published by authority of the Minister responsible for Statistics Canada



**An HTML version is also available.**

*Cette publication est aussi disponible en français*.

Survey Methodology, December 2017
Vol. 43, No. 2, pp. 183-210
Statistics Canada, Catalogue No. 12-001-X


# 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**[1]

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. This model also makes it possible to utilize the higher frequency of the social media to produce more precise estimates for the sample survey in real time at the moment that statistics for the social media become available but the sample data are not yet available. 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.

**Key Words:** Big data; Design-based inference; Model-based inference; Nowcasting; Structural time series modelling; Cointegration.

## 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, Swensson and Wretman (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

1. Jan van den Brakel, Statistics Netherlands, Methodology Department, Heerlen, The Netherlands and Maastricht University School of Business and Economics, Department of Quantitative Economics, The Netherlands. E-mail: ja.vandenbrakel@cbs.nl; Emily Söhler, Student Econometrics, Maastricht University; Piet Daas and Bart Buelens, Statistics Netherlands, Methodology Department, Heerlen, The Netherlands.

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, Brick, Bates, Battaglia, Couper, Dever, Gile and Tourangeau (2013) 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, Giusti, Pratesi, Salvati, Giannotti, Perdreschi, Rinzivillo, Pappalardo and Gabrielli (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. Harvey and Chung (2000) propose a time series model for the Labour Force Survey in the UK extended with a series of claimant counts. If such a model detects strong positive correlations between these components, then this might further increase the precision of the time series estimates for the sample survey. Indicators derived from social media are generally available at a higher frequency than related series obtained with periodic surveys. This allows to use this time series modelling approach to make early predictions for the survey outcomes in real time at the moment that the outcomes for the social media are available, but the survey data not yet. In this case, the social media are used as a form of nowcasting. 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. It can be argued that if an auxiliary series is cointegrated with the series of the survey, they represent the same underlying stochastic process. This could be used as an argument to motivate that a statistic measured with a big data source is representative for an intended target population. This is, however, more an empirical argument and not as strong as the theory underlying probability sampling, that proves that random sampling in combination with an (approximately) design-unbiased estimator results in representative statistics.

The Dutch Consumer Confidence Survey (CCS) is a monthly survey based on approximately 1,000 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. It is also illustrated how the SMI in this time series model can be used to make early predictions or nowcasts 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 2,500 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 1,000 responding households is obtained, which comes down to a response rate of about 40%. A major part of the nonresponse are households for which no known telephone number of a land-line connection is available. The response among households for which a known telephone number is available is about 60%.

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 $q = 1, \ldots, 5,$ 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:

$$I_t = \frac{1}{Q} \sum_{q=1}^{Q} \left( P_{1,t}^{q} - P_{3,t}^{q} \right). \tag{2.1}$$

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.,

$$P_{j,t}^{q} = \frac{100}{n_t} \sum_{i=1}^{n} \delta_{i,j,t}^{q}, \tag{2.2}$$

for question $q = 1, \ldots, 5,$ and answer category $j = 1, 2, 3.$ In (2.2) $n_t$ 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, it can be proved that the variance of (2.1) can be estimated by

$$\begin{aligned}\mathrm{Var}\left(I_t\right) &= \frac{1}{Q^2}\sum_{q=1}^{Q}\left[\mathrm{Var}\left(P_{1,t}^{q}\right)+\mathrm{Var}\left(P_{3,t}^{q}\right)\right]-\frac{2}{Q^2}\sum_{q=1}^{Q}\sum_{q'=1}^{Q}\mathrm{Cov}\left(P_{1,t}^{q},P_{3,t}^{q'}\right)\\ &+\frac{1}{Q^2}\sum_{q=1}^{Q}\sum_{q'\neq q}^{Q}\left[\mathrm{Cov}\left(P_{1,t}^{q},P_{1,t}^{q'}\right)+\mathrm{Cov}\left(P_{3,t}^{q},P_{3,t}^{q'}\right)\right],\end{aligned} \tag{2.3}$$

with

$$\mathrm{Var}\left(P_{j,t}^{q}\right)=\frac{1}{n_t}P_{j,t}^{q}\left(100-P_{j,t}^{q}\right),\qquad \mathrm{Cov}\left(P_{j,t}^{q},P_{j',t}^{q'}\right)=\frac{1}{n_t}\left(P_{jj',t}^{qq'}-P_{j,t}^{q}P_{j',t}^{q'}\right),$$

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

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

Figure 2.1 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.

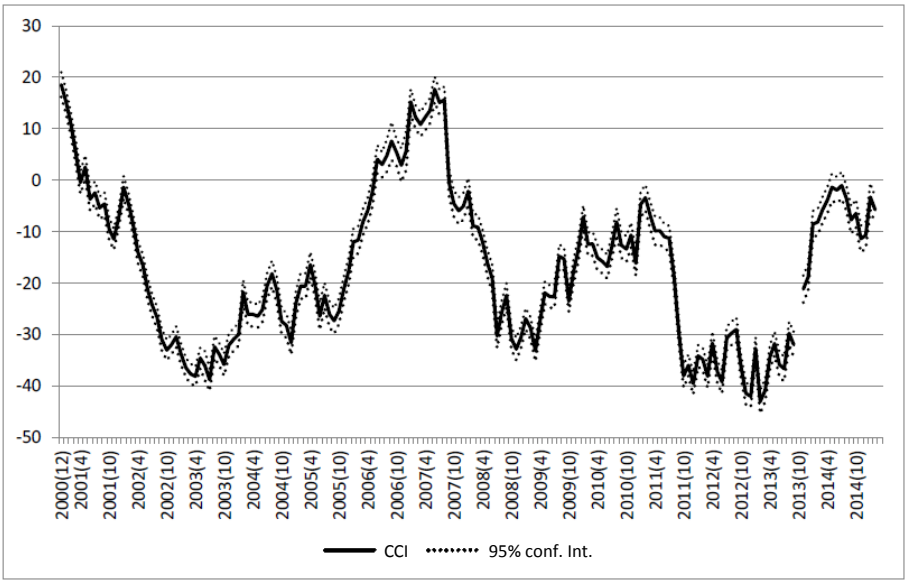


**Figure 2.1 Consumer confidence index (CCI) with a 95% confidence interval.**

## 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 2.2, 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. The size or amplitude of the movements of the CCI are also considerably larger compared to the SMI. 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.

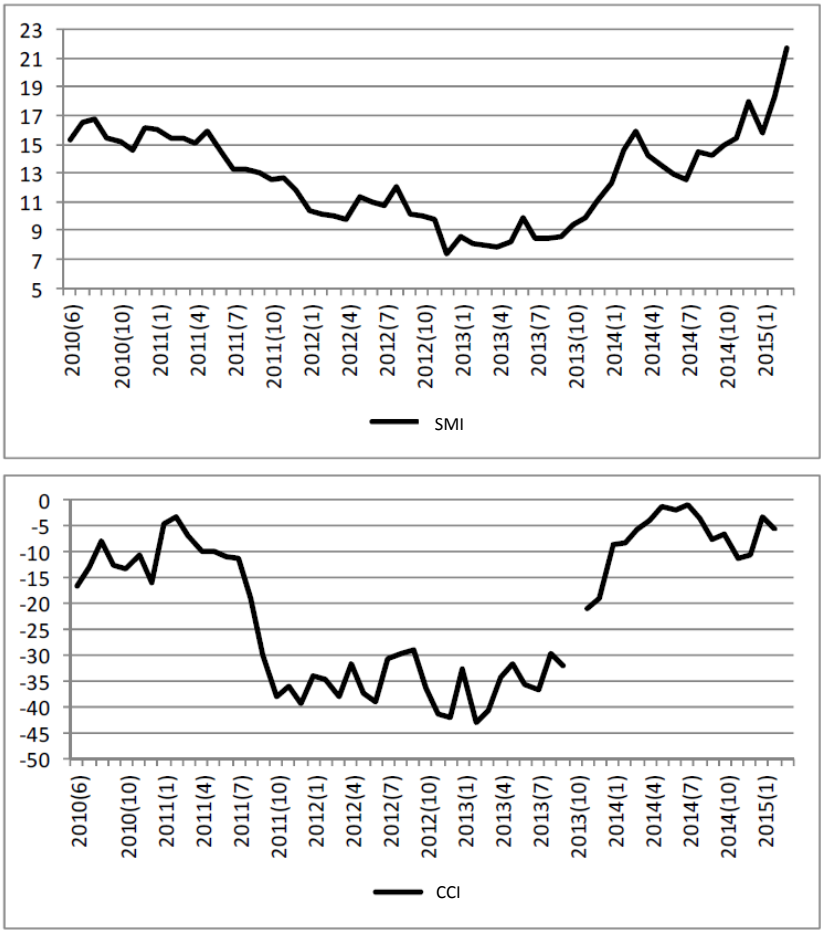


**Figure 2.2 Comparison of the Social media index (SMI, upper panel) with the Consumer confidence index (CCI, lower panel).**

## 2.3 Quality aspects of the CCI and the SMI

The accuracy of statistics are measured with its variance and bias. For simplicity we only distinguish between selection bias and measurement bias. The variance of survey sample statistics, like the CCI, depends on the sample size and will typically constitute a substantial part of the uncertainty of sample statistics. In big data sources the concept of sampling variance is meaningless since the data generating process is not a probability sample from a finite target population. Variance components of the model used to describe the assumed data generating process could be used as an accuracy measure instead. The model-based variance of statistics obtained with time series models applied to series obtained from internet or social media will always be positive depending of the volatility of the series, which predominantly depends on the frequency of the observed series and the dynamics of the stochastic process instead of the volume of the data.

The selection bias of sample survey statistics is approximately zero under complete response. In practice however, selection bias arises due to selective nonresponse, under coverage of the sample frame and to which extend with the field work strategy the target population is successfully reached. In the case of the CCI, only the population with a known telephone number of a land-line connection is reached and the response among this subpopulation is about 60%. The selection bias of big data sources is generally unknown. In this paper, we apply the concept of cointegration to evaluate to which extend the SMI measures the same concept as the CCI. Note, however, that in the case of cointegration, the SMI might reflect a similar nonresponse and coverage selection bias as the CCI. Baker et al. (2013) pointed out that there are similarities between selection bias in probability samples and the non-probabilistic approach followed with data sources like social media.

The measurement bias in sample statistics typically depends on the extend that the conceptual variables to be measured, are implemented in the questionnaire, but also on data collection mode and the quality of the interviewers. Problems with measurement bias in surveys arises, since measurements of the variables of interest are indirect in that respondents are asked to report about their behaviour, introducing all kind of measurement errors. In the case of the CCI the question can be raised to which extend respondents are capable to express their long-term confidence in the economy and to which extend it is influenced by short-term emotions. These problems do not arise with big data if they contain direct measurements of people behaviour. With an index derived from social media like the SMI the question can be raised to which extend it measures a similar concept as the CCI. In Subsection 2.2, it was already mentioned that major changes in the behaviour of the public on social media have a disturbing effect on the series. Particularly at the end of the series, a sudden change in behaviour on social media will be very hard to distinguish from a real turning point. For example, a Google-trend series on search related to vacancies might track an official series on unemployment. It does measure unemployment, however, search behaviour before the start of the financial crisis in 2009 might be completely different compared to the period directly after the financial crisis, invalidating the concept intended to be measured.

# 3 Structural time series modelling 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 autoregressive-moving-average (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 or in a nowcasting procedure to obtain more precise real time estimates.

## 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, Smith and Jones (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 Rubin-Bleuer (1993), Pfeffermann, Feder and Signorelli (1998), Pfeffermann and Tiller (2006), Harvey and Chung (2000), Feder (2001), Lind (2005) 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:

$$I_t = \theta_t + e_t, \tag{3.1}$$

where $\theta_t$ 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:

$$\theta_t = L_t + S_t + \xi_t, \tag{3.2}$$

where $L_t$ denotes a stochastic trend, $S_t$ a stochastic seasonal component and $\xi_t$ the unexplained variation of the finite population parameter. Inserting (3.2) into measurement model (3.1) gives

$$I_t = L_t + S_t + \xi_t + e_t. \tag{3.3}$$

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

$$\upsilon_t = \xi_t + e_t. \tag{3.4}$$

It is assumed that $E(\upsilon_t) = 0$ and $\text{Var}(\upsilon_t) = \sigma_\upsilon^2$. To allow for nonhomogeneous variance in the sampling errors, Binder and Dick (1990) proposed a measurement error where the disturbance terms $\upsilon_t$ are proportional to the standard errors of $I_t$, i.e.,

$$\upsilon_t = \sqrt{\text{Var}(I_t)}\ \tilde{\upsilon}_t, \tag{3.5}$$

with $E(\tilde{\upsilon}_t) = 0$, $\text{Var}(\tilde{\upsilon}_t) = \sigma_\upsilon^2$, and where $\text{Var}(I_t)$ is defined by (2.3) and is used as a priori information in the time series model. Such a model would be useful if the sampling error dominates the white noise in the population parameter. Initial analyses indicate that in this application the variance of the population white noise is substantial, invalidating (3.5) for this application. In addition, the variance of the sampling error in this application is constant over time. Therefore, it is decided to combine the sampling error with the population white noise and assume a constant variance over time. The question how to account for sampling variance is also an issue in seasonal adjustment variances (Pfeffermann and Sverchkov, 2014). Bell (2005) studied the contribution of the sampling variance in the variance of the estimation error of seasonally adjusted series and in the nonseasonal component. In the case of (rotating) panels, the sampling error can be separated from the population white noise. In cross-sectional repeated surveys, it is difficult to identify the separate components and therefore both terms are combined in one disturbance term that captures both the sampling variance and the unexplained variation of the population parameter.

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):

$$L_t = L_{t-1} + R_t,$$

$$R_t = R_{t-1} + \eta_t, \quad \mathrm{E}(\eta_t) = 0, \tag{3.6}$$

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

Adding a random component for the level in (3.6) improves the log-likelihood with five units but results in an overfit of the data in a sense that the smoothed signal almost exactly follows the observed series with a very small measurement error variance. A local level model (random level without a slope) improves the log-likelihood with three units but also intends to overfit the data.

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

$$S_t = \sum_{j=1}^{6} \gamma_{jt}, \tag{3.7}$$

with

$$\begin{aligned} \gamma_{jt} &= \gamma_{jt-1} \cos(\lambda_j) + \tilde{\gamma}_{jt-1} \sin(\lambda_j) + \omega_{jt}, \\ \tilde{\gamma}_{jt} &= -\gamma_{jt-1} \sin(\lambda_j) + \tilde{\gamma}_{jt-1} \cos(\lambda_j) + \tilde{\omega}_{jt}. \end{aligned}$$

Here $\lambda_j$ denotes the frequency of the different cycles in radians and is defined as

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

For the disturbance terms, it is assumed that

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

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

For reasons of parsimony, the same variance structure is assumed with the same hyperparameter for $\tilde{\omega}_{jt}$. Furthermore, it is assumed that $\omega_t$ and $\tilde{\omega}_t$ are uncorrelated.

After including the stochastic trend component (3.6) and seasonal component (3.7), no additional cycle components are required. The model selection procedure indicated that two level interventions are needed to model sudden jumps in the series. The first one is due to 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. Adding these three components increases the log-likelihood with 15 units. These considerations lead to the following model for the observed CCI series

$$I_t = L_t + S_t + \beta^{07}\delta_t^{07} + \beta^{08}\delta_t^{08} + \beta^{11}\delta_t^{11} + \upsilon_t, \tag{3.8}$$

with

$$\delta_t^{07} = \begin{cases} 1 & \text{if} \quad t = 2007(9) \\ 0 & \text{if} \quad t \neq 2007(9) \end{cases}, \qquad \delta_t^{08} = \begin{cases} 1 & \text{if} \quad t \geq 2008(9) \\ 0 & \text{if} \quad t < 2008(9) \end{cases}, \qquad \delta_t^{11} = \begin{cases} 1 & \text{if} \quad t \geq 2011(9) \\ 0 & \text{if} \quad t < 2011(9) \end{cases},$$

and $\beta^x$ the corresponding regression coefficients.

Finally, autoregressive (AR) and moving average (MA) components can be added to the structural time series model (3.8). In this application, there were no indications that such components are required, since there are no clear signs of remaining serial correlation in the standardized innovations. Adding an AR(1) or an MA(1) to (3.8) increases the log-likelihood with 5 and 4.5 units respectively. Adding second-order AR or MA models does not further improve the log-likelihood. Adding an ARMA(1,1) also does not further increases the log-likelihood. An AR(1) or MA(1) slightly improves the correlogram but also increases the standard error of the filtered smoothed signals. Therefore, model (3.8) was finally selected for the CCI series.

State space models assume that the disturbance terms are normally and independently distributed. These assumptions translate into the assumption that the innovations are normally and independently distributed. Table A.1 in the appendix contains an overview of goodness of fit statistics applied to the standardized innovations. The values for skewness, kurtosis and the Bowman-Shenton test do not indicate deviations from normality of the standardized innovations. The values for the Ljung-Box test and Durban-Watson test do not indicate serial correlations in the standardized innovations. This is also confirmed by a correlogram (not shown). In conclusion, these diagnostics indicate that (3.8) fits the series of the CCI reasonably well.

## 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. 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. A model selection procedure, similar to the one conducted for the CCI series in Subsection 3.1, indicated that the observed series for the SMI can be modelled with a smooth trend model and a white noise component for the unexplained variation. No significant seasonal component or business cycle is established. There are no signs for outliers or level shifts. AR(1) and MA(1) components are not included since there is no serial correlation in the standardized innovations. 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.

Tables A.2 and A.3 in the appendix contain an overview of goodness of fit statistics for the standardized innovations of the CCI and SMI respectively. There are no indications that the standardized innovations of both series deviate from a normal distributions. The null hypothesis of no serial correlation in the standardized innovations could not be rejected. The correlogram of the innovations for the SMI, however, show a non-significant seasonal pattern (not shown). The innovations of the SMI, also contain heteroscedasticity.

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:

$$\begin{pmatrix} I_t \\ X_t \end{pmatrix} = \begin{pmatrix} L_t^I \\ L_t^X \end{pmatrix} + \begin{pmatrix} S_t^I \\ 0 \end{pmatrix} + \begin{pmatrix} \beta^{11}\delta_t^{11} \\ 0 \end{pmatrix} + \begin{pmatrix} \upsilon_t^I \\ \upsilon_t^X \end{pmatrix}, \tag{3.9}$$

with $L_t^I$ and $L_t^X$ the smooth trend model as defined in (3.6) with covariance structure

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

In the last expression $\rho_\eta$ denotes the correlation between the slope disturbances of the CCI and SMI. Furthermore, $S_t^I$ is the seasonal effect defined by (3.7) and $\delta_t^{11}$ the intervention for September 2011 with $\beta^{11}$ the corresponding regression coefficient. Finally, $\upsilon_t^I$ and $\upsilon_t^X$ are the disturbance terms for the CCI and SMI series and are defined as:

$$\begin{aligned} E(\upsilon_t^I) &= E(\upsilon_t^X) = 0, \\ \text{Cov}\left(\upsilon_t^I, \upsilon_{t'}^I\right) &= \begin{cases} \sigma_{\upsilon I}^2 & \text{if} \quad t = t' \\ 0 & \text{if} \quad t \neq t' \end{cases}, \\ \text{Cov}\left(\upsilon_t^X, \upsilon_{t'}^X\right) &= \begin{cases} \sigma_{\upsilon X}^2 & \text{if} \quad t = t' \\ 0 & \text{if} \quad t \neq t' \end{cases}, \\ \text{Cov}\left(\upsilon_t^I, \upsilon_{t'}^X\right) &= 0 \text{ for all } t \text{ and } t'. \end{aligned}$$

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 singular value decomposition:

$$\text{cov}\begin{pmatrix} \eta_t^I \\ \eta_t^X \end{pmatrix} = \begin{pmatrix} \sigma_{\eta I}^2 & \sigma_{\eta I}\sigma_{\eta X}\rho_\eta \\ \sigma_{\eta I}\sigma_{\eta X}\rho_\eta & \sigma_{\eta X}^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ a & 1 \end{pmatrix}\begin{pmatrix} d_1 & 0 \\ 0 & d_2 \end{pmatrix}\begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix}. \tag{3.10}$$

Instead of estimating $\sigma_{\eta I}^2$, $\sigma_{\eta X}^2$, and $\rho_\eta$, 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 $R_t^X = aR_t^I + \overline{R}$. Similarly, the trend for the SMI series can be expressed as a linear combination of the trend for the CCI series as $L_t^X = aL_t^I + \overline{L} + \overline{R}t$. Note that $\overline{R}$ and $\overline{L}$ 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 (ML) estimates for the hyperparameters, i.e., the variance components of the stochastic processes for the state variables are obtained using a numerical optimization procedure (Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm, Doornik, 2009). 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 (2012).

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 (3.8) from Section 3.1 applied to the series of the CCI obtained from December 2000 until March 2015. In Table 4.1, the ML estimates for the hyperparameters of the model are specified.

**Table 4.1**
**Maximum Likelihood estimates hyperparameters univariate model CCI**

| Standard deviation | ML estimate |
|---|---|
| Trend $(\sigma_\eta)$ | 1.18 |
| Seasonal $(\sigma_\omega)$ | 0.0025 |
| Measurement equation $(\sigma_\upsilon)$ | 2.46 |

The average of the standard errors of the direct estimates for the CCI equals 1.21. The standard deviation of the disturbance terms of the measurement equation equals 2.46, as follows from Table 4.1. This illustrates that the population white noise dominates the variance of the measurement disturbance terms as mentioned by the choice of the variance structure for (3.4) in Section 3.1.

In the upper panel of Figure 4.1, the smoothed trend plus interventions are compared with the direct estimates for the CCI. In the lower panel of Figure 4.1, 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 are removed from the original series. It follows from Figure 4.1 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 4.2. This filtered series approximates 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. Both figures illustrate that the Kalman filter provides plausible smoothed but also filtered imputations for the missing observation in October 2013.

Figure 4.3 shows the smoothed seasonal pattern of the CCI series. Since the seasonal effects are almost time invariant, the effects are displayed for the 12 months of one year only. There are clear significant negative effects in October, November and December and clear positive effects in January and August. The intention of the CCI is to measure a long-term confidence of respondents, since all questions refer to the respondents financial and economic situation over the last 12 month or the expectations for the future 12 months. The clear significant seasonal pattern, however, indicates 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 4.4, 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.

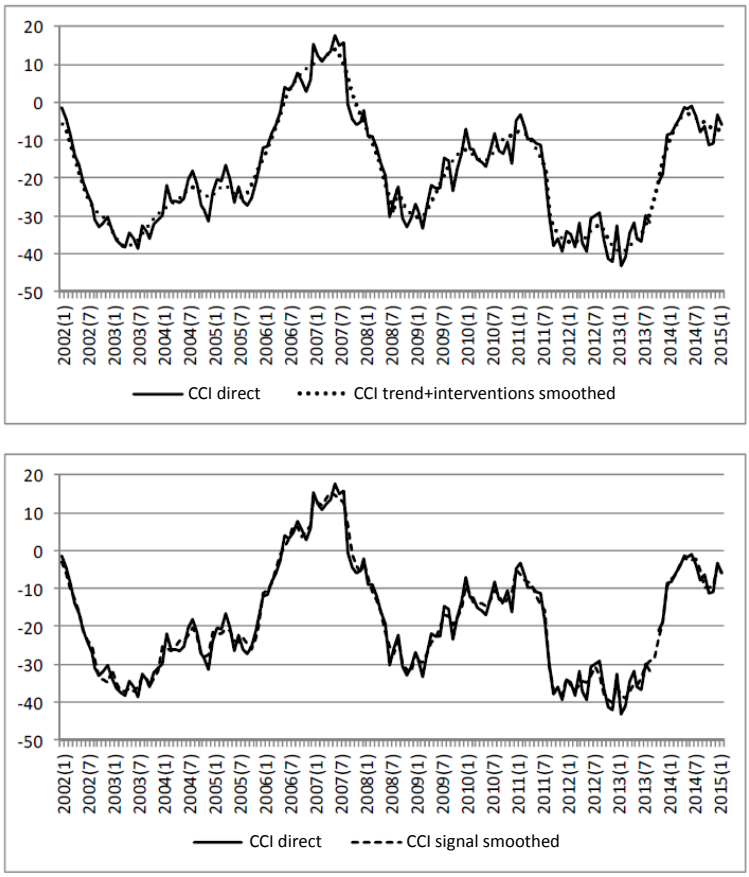


**Figure 4.1 Smoothed trend plus interventions compared with direct estimates CCI (upper panel) and smoothed signal (trend plus intervention plus seasonal) compared with direct estimates CCI (lower panel).**

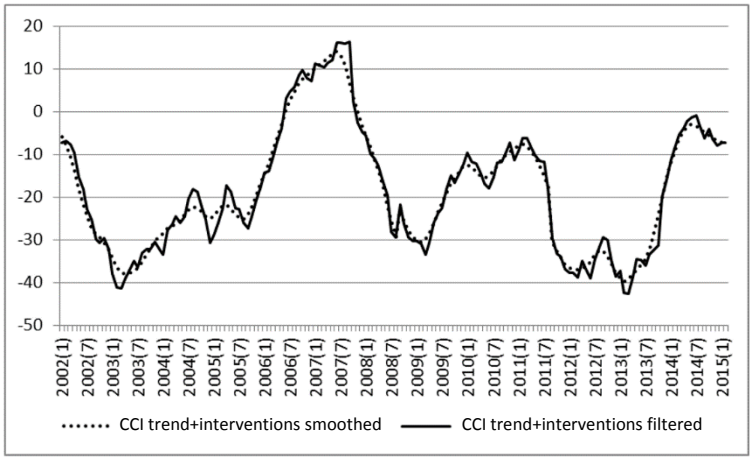


**Figure 4.2 Filtered trend plus interventions compared with smoothed trend plus interventions CCI.**

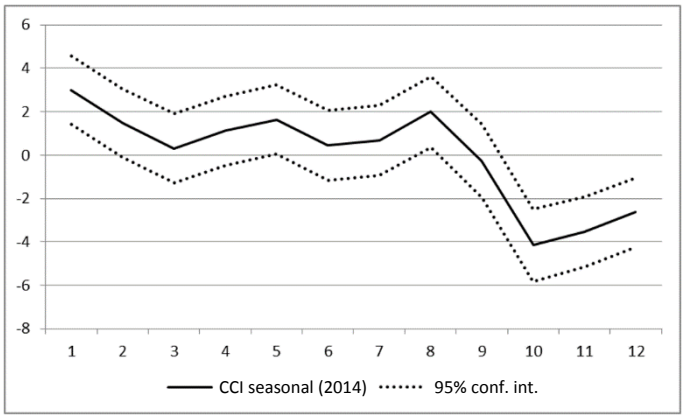


**Figure 4.3 Smoothed seasonal pattern CCI for 2014.**

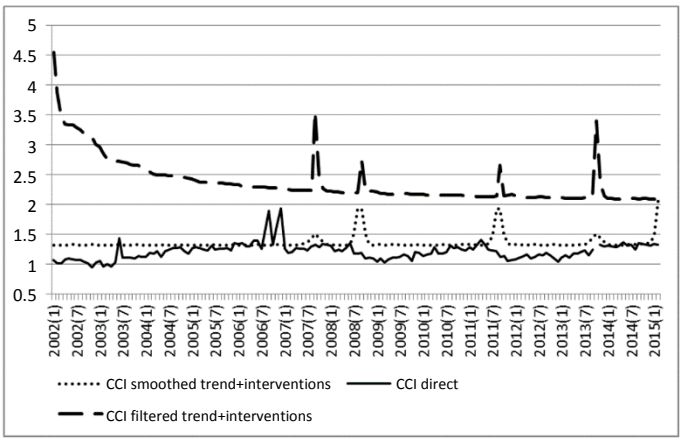


**Figure 4.4 Standard error smoothed and filtered trend plus interventions compared with direct estimates CCI.**

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, 2017).

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 (3.8) contains two components; the sampling error and the unexplained high-frequency variation of the real population value, as expressed by (3.4). Recall from Table 4.1 that $\sigma_\upsilon$ is

equal to 2.46 and is twice as large as the average value of the standard errors of the direct estimates. This is a strong indication that the variance of the white noise component in the true population variable is of the same order as 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.

Although the gain in precision of level estimates obtained with the time series model is limited, the estimates for the trend are more stable as follows from Figures 4.1 and 4.2. A time series model will therefore still be useful to filter a more stable long term trend from the high-frequency variation in the population parameter and the sampling error. Because the state variables of the trend component of subsequent periods will have a strong positive correlation, more gain from the time series modelling approach can be expected by focussing on month-to-month changes, see e.g., Harvey and Chung (2000). Filtered estimates for the month-to-month change of the CCI are defined as

$$\Delta_{t|t} = L_{t|t} - L_{t-1|t} + \beta^{08}_{t|t}\,\delta^{08}_{t} - \beta^{08}_{t-1|t}\,\delta^{08}_{t-1} + \beta^{11}_{t|t}\,\delta^{11}_{t} - \beta^{11}_{t-1|t}\,\delta^{11}_{t-1}, \tag{4.1}$$

where the notation $\Theta_{t|t'}$ stands for the estimate for state variable $\Theta$ for period $t$ given the data observed until period $t'$. The outlier in 2007(9) is, naturally, removed from the signal. Furthermore, the regression coefficients are time invariant. Therefore, $\beta^{x}_{t|t} = \beta^{x}_{t-1|t}$ for $x = 08$ and 11. Since $t = 2008(9)$ and $t = 2011(9)$ are the months that $\delta^{08}_{t}$ and $\delta^{11}_{t}$ change form value, expression (4.1) can be simplified to

$$\Delta_{t|t} = L_{t|t} - L_{t-1|t} + \beta^{08}_{t|t}\,\tilde{\delta}^{08}_{t} + \beta^{11}_{t|t}\,\tilde{\delta}^{11}_{t}, \tag{4.2}$$

with $\tilde{\delta}^{08}_{t} = 1$ if $t = 2008(9)$ and $\tilde{\delta}^{08}_{t} = 0$ for all other periods and $\tilde{\delta}^{11}_{t} = 1$ if $t = 2011(9)$ and $\tilde{\delta}^{11}_{t} = 0$ for all other periods. Smoothed estimates for the month-to-month change of the CCI are defined as

$$\Delta_{t|T} = L_{t|T} - L_{t-1|T} + \beta^{08}_{t|T}\,\tilde{\delta}^{08}_{t} + \beta^{11}_{t|T}\,\tilde{\delta}^{11}_{t}. \tag{4.3}$$

To compare the month-to-month changes based on (4.2) and (4.3) with the direct estimates, the smoothed seasonal effects in (3.8) are subtracted from the direct estimates. The standard errors of the direct estimates are not corrected for this adjustment.

Figure 4.5 compares the direct estimates for the month-to-month change with the smoothed estimates (upper panel) and the filtered estimates (middle panel) obtained with the time series model. The lower panel compares the standard errors of the smoothed, filtered and direct estimates. The filtered and in particular the smoothed estimates for month-to-month change have a more stable pattern compared to the direct estimates. This is also reflected by the standard errors. The strong positive correlations of the states of the trend component between subsequent periods results in standard errors for filtered and smoothed estimates of the month-to-month change that are clearly smaller compared to the direct estimator. Exceptions are the two periods where a level intervention is required. Introducing a level shift results for a short period in an increased level of uncertainty.

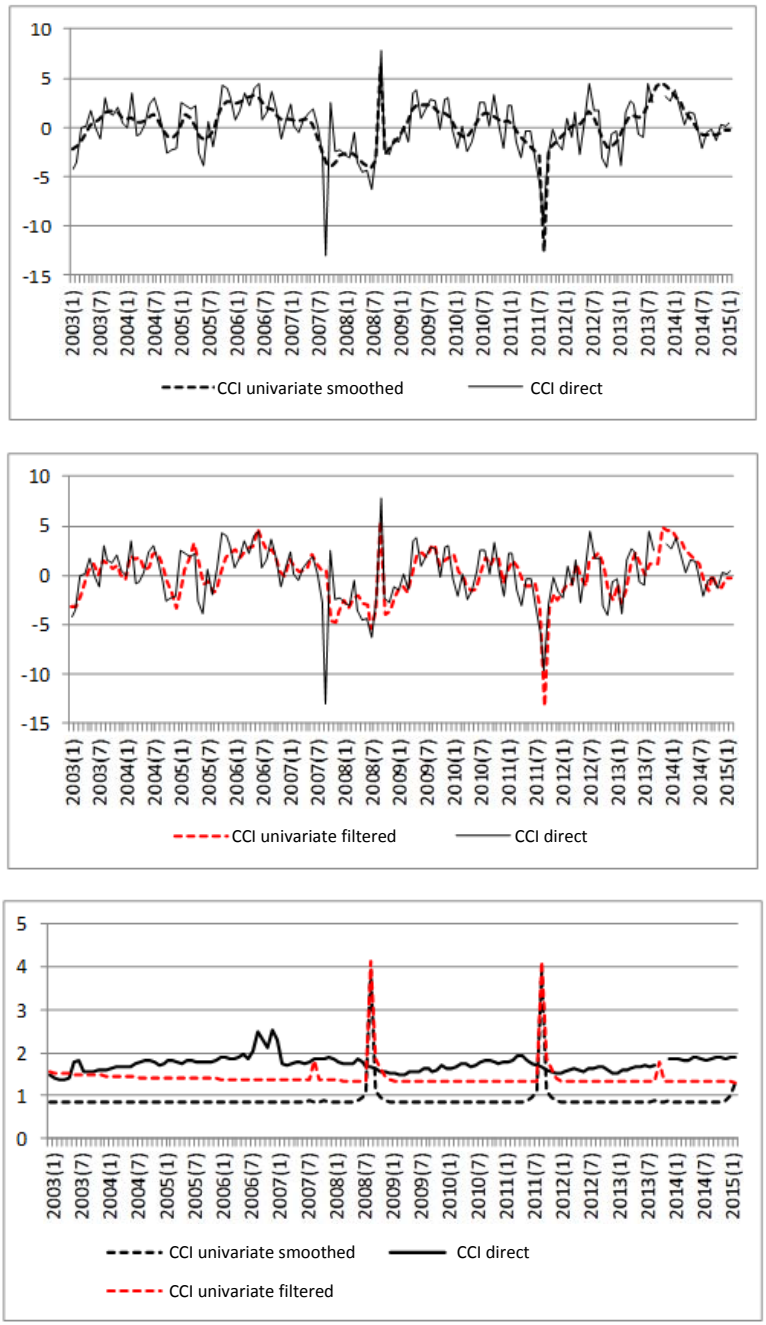


**Figure 4.5 Comparison month-to-month change univariate model and direct estimates. Upper panel: smoothed estimates, middle panel: filtered estimates, lower panel standard errors.**

The reduction in standard error is measured as the Mean Relative Difference in Standard Error (MRDSE), and is for filtered estimates defined as $\text{MRDSE} = 100/(T - t') * \sum_{t=t'}^{T} \left[\text{se}\left(\hat{\Delta}_t\right) - \text{se}\left(\Delta_{t|t}\right)\right] \Big/ \text{se}\left(\hat{\Delta}_t\right)$, with $\text{se}\left(\hat{\Delta}_t\right)$ the standard error for the direct estimate for the month-to-month change. The MRDSE for smoothed estimates is obtained by replacing $\text{se}\left(\Delta_{t|t}\right)$ for $\text{se}\left(\Delta_{t|T}\right)$. During the period observed from 2003(1), the MRDSE for smoothed estimates equals 47% and for the filtered estimates 17%.

## 4.2 Bivariate model for CCI and SMI series

In this section, the bivariate model (3.9) proposed in Section 3.2 is applied to the series of the CCI and SMI, which are available from June 2010 until March 2015. Note that the time series components for the CCI are re-estimated using the shorter series. Maximum likelihood estimates for the hyperparameters are specified in Table 4.2. 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 (3.10) 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.

**Table 4.2**
**Maximum Likelihood estimates hyperparameters bivariate model CCI and SMI**

| Standard deviation | ML estimate |
|---|---|
| Trend CCI $(\sigma_{\eta I})$ | 1.25 |
| Seasonal CCI $(\sigma_{\omega})$ | 7.5E-6 |
| Trend SMI $(\sigma_{\eta X})$ | 0.25 |
| Measurement equation CCI $(\sigma_{\upsilon I})$ | 2.68 |
| Measurement equation SMI $(\sigma_{\upsilon X})$ | 0.84 |
| Correlation trend CCI and SMI $(\rho_{\eta})$ | 0.92 |

Figure 4.6 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 4.6). This is picked up by the model that allows for correlation (mid panel Figure 4.6). 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 4.6) is compared with the cross-plot of a common factor model (right-panel Figure 4.6).

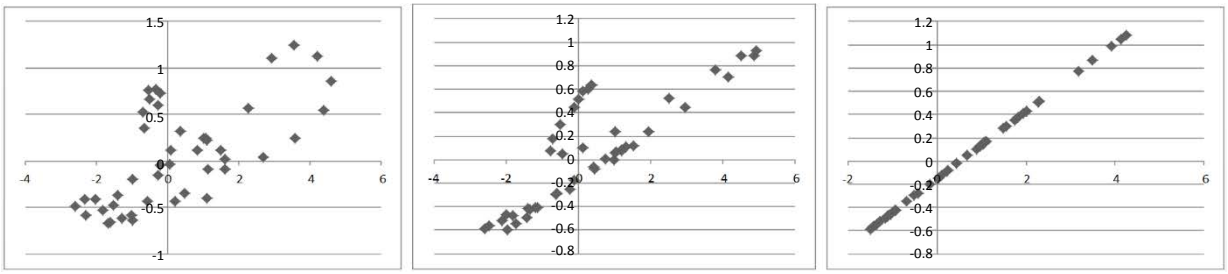


**Figure 4.6 Cross-plot smoothed slopes CCI (x-axis) and SMI (y-axis) for a model without correlation (left panel), correlation estimated with ML (mid panel) and correlation set equal to one (right panel).**

Figure 4.7 compares the observed SMI series with the smoothed trend obtained under the bivariate model. Figure 4.8 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 4.8, the level and evolution of the smoothed estimates for the CCI series are almost identical under the univariate and bivariate models.

Figure 4.9 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 4.9, 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 4.10.

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.

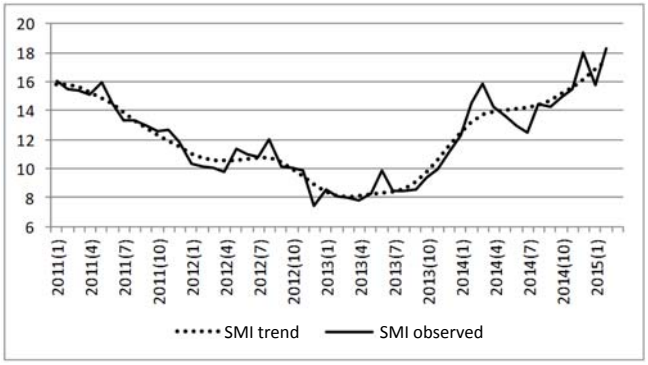


**Figure 4.7 Observed series and smoothed trend SMI.**

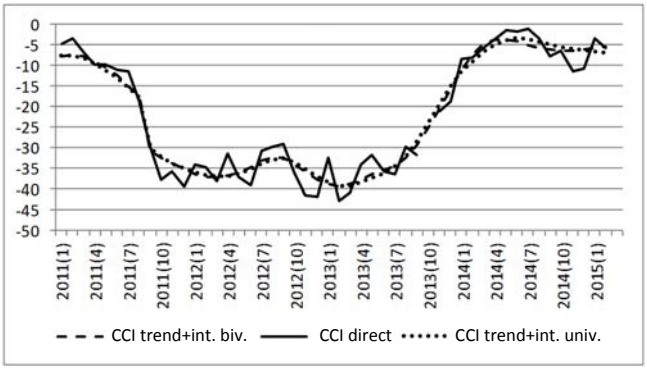


**Figure 4.8 CCI comparison of the direct estimates and smoothed trend plus intervention under the bivariate and univariate models for CCI.**

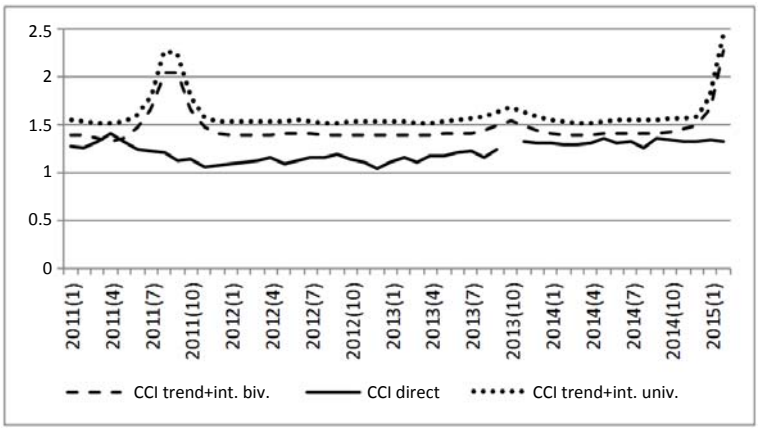


**Figure 4.9 CCI comparison of standard errors direct estimates and smoothed trend plus intervention under the bivariate and univariate models for CCI if both models are applied to a series of equal length (June 2010-March 2015).**

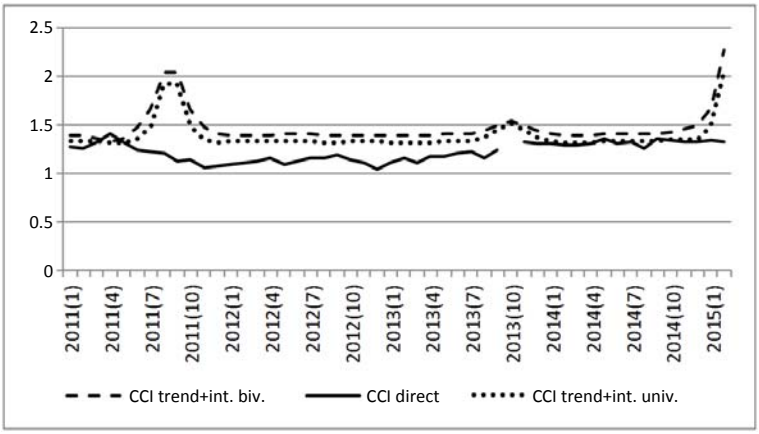


**Figure 4.10 CCI comparison of standard errors direct estimates and smoothed trend plus intervention under the bivariate and univariate models for CCI if the univariate model is applied to the complete CCI series (December 2000).**

The upper panel of Figure 4.11 compares the direct estimates for the month-to-month change with the smoothed estimates obtained with the univariate and bivariate time series models (both based on the series observed from June 2010). The lower panel compares the standard errors of these estimates. During the period observed from 2011(1), the MRDSE for smoothed estimates under the univariate model equals 39% and under the bivariate model 43%. The MRDSE for filtered estimates under the univariate model equals 7% and under the bivariate model 14%. As in the case of the univariate model, the time series modelling approach results in more stable and more precise estimates for the month-to-month change. The use of the SMI series slightly improves the precision of the month-to-month changes compared to the univariate model.

Once the direct estimate for the CCI for month $t$ becomes available, the additional value of the SMI series is limited to improve a time series estimate for the CCI for month $t$. A drawback of sample surveys, however, is that they generally are less timely compared to social media sources. The additional value of the SMI becomes more clear when the higher frequency of this series is used to produce early predictions or nowcasts for the CCI with the bivariate state space model. If during month $t$ or directly at the end of month $t$ a first early prediction for the CCI is required, the univariate model can only produce a one-step-ahead prediction. As soon as during month $t$ or at the end of month $t$ results for the SMI series become available, the bivariate model exploits the strong correlation between the series to make a more precise prediction for the CCI, already before the direct estimate for month $t$ becomes available.

To illustrate the additional value of the SMI in a nowcast procedure for the CCI, we compare in the upper panel of Figure 4.12, the one-step-ahead predictions for the trend plus intervention of the CCI series obtained with the univariate model with the estimate obtained with the bivariate model if the SMI for month $t$ is available but the direct estimate of the CCI is still missing. The smoothed estimates for the trend plus intervention of the CCI obtained with the univariate model are included as a benchmark. In the lower panel the standard errors of these three estimates are compared.

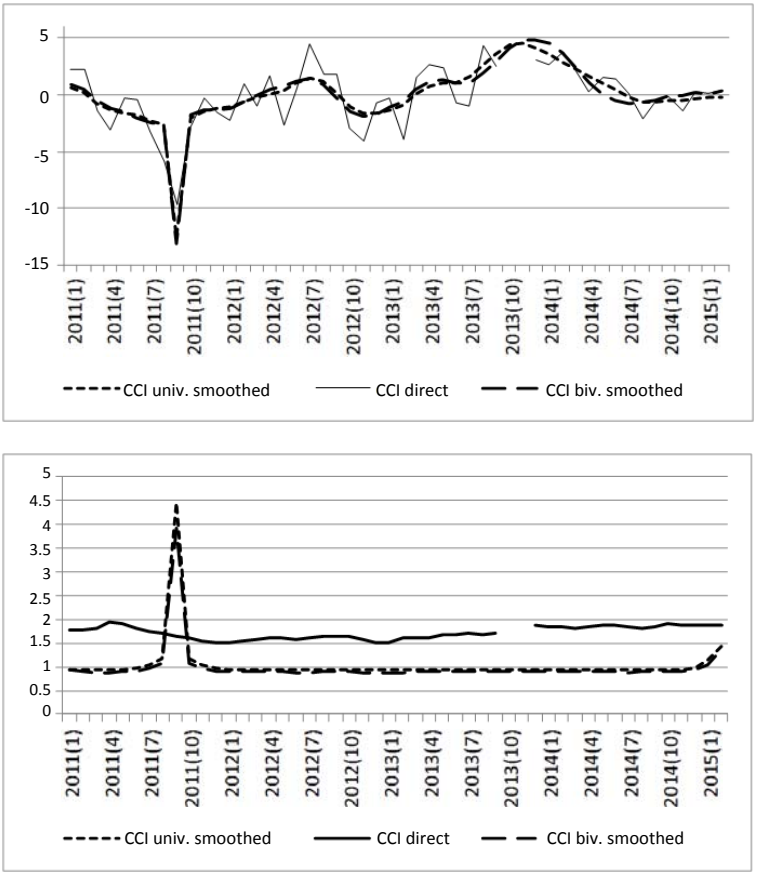


**Figure 4.11 Comparison month-to-month change bivariate model, univariate model and direct estimates. Upper panel: smoothed estimates, lower panel standard errors.**

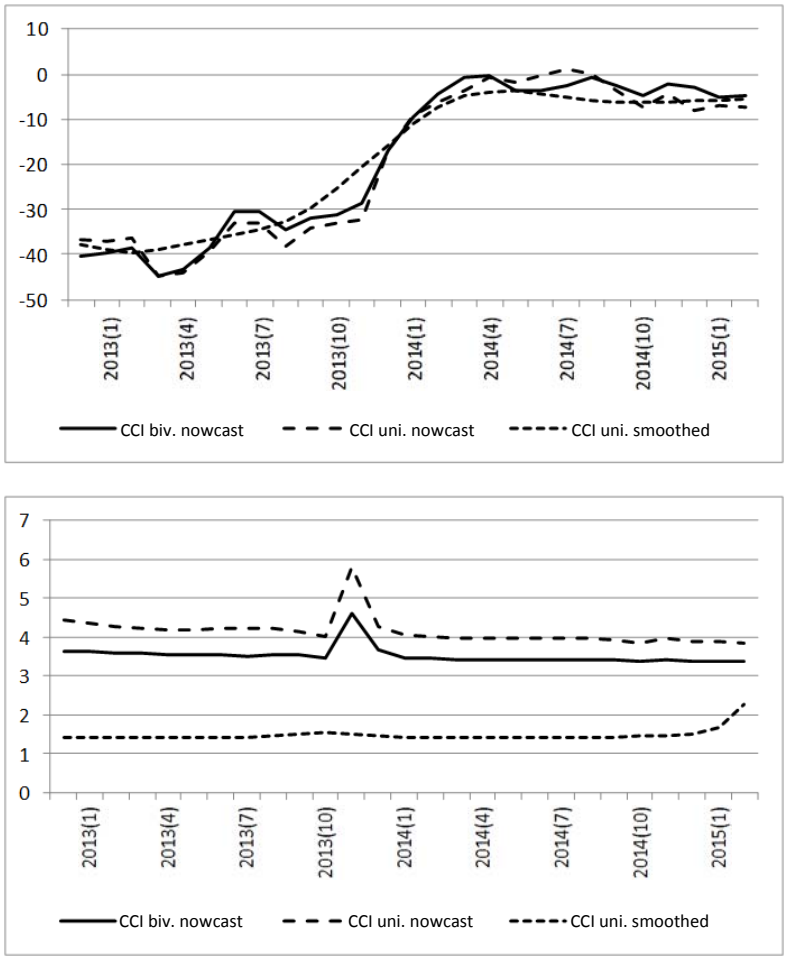


**Figure 4.12 Comparison estimates for trend plus intervention CCI series; one-step-ahead prediction univariate model (CCI uni. nowcast), bivariate model if the SMI for month *t* is available but the direct estimate of the CCI is missing (CCI biv. nowcast) and smoothed estimates with the univariate model (CCI uni. smoothed). Upper panel compares point estimates. Lower panel compares standard errors.**

If the smoothed estimates obtained with the univariate model are used as a benchmark, the Mean Absolute Relative Difference (MARD) between nowcasts and smoothed estimates is used as a measure for the size of the revision and is defined as $\text{MARD} = 100/(T - t') * \sum_{t=t'}^{T} \left|\theta_{t|T} - \theta_{t|t-1}\right| \Big/ \left|\theta_{t|T}\right|$, where $\theta_t = L_t + \beta^{11}\delta_t^{11}$ denotes the trend plus intervention of the CCI series. Based on the months observed from $t = 2013(1)$ the MARD for nowcasts obtained with the univariate model equals 35% and for the bivariate model 31%. This shows that the size of the revisions is a bit smaller and thus more stable with nowcasts for the CCI with the bivariate model. The difference in precision between the nowcasts obtained with the univariate model and the bivariate model are measured with the MRDSE and is in this case defined as $\text{MRDSE} = 100/(T - t') * \sum_{t=t'}^{T} \left[\text{se}\left(\theta_{t|t-1}^{\text{uni}}\right) - \text{se}\left(\theta_{t|t-1}^{\text{biv}}\right)\right] \Big/ \text{se}\left(\theta_{t|t-1}^{\text{biv}}\right)$. Based on the months observed from $t = 2013(1)$ the difference in precision of both nowcasts based on this MRDSE equals 17%. Figure 4.12 as well as the MARD and the MRDSE illustrate that the SMI improves the stability and precision of nowcasts for the CCI.

# 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 or Google Trends are selected by maximizing the correlation with the series from the sample survey and does not necessarily measure the same concept as the 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.

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 periodic 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 or period-to-period change of the direct estimates obtained with a periodic survey. Another important benefit with big data sources is to use the higher frequency of these data sources to make more precise early predictions or nowcasts if in real time the survey estimate is not yet available but the covariate is already available. 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 periodic surveys. There are of course also issues with survey sampling. For example, continuously declining response rates and data

collection modes that does not reach the intended target population result in selection bias either. In this case, cointegration with a related series derived from social media might be indication that there are similarities between the selection bias in the non-probabilistic big data sources and the non-response selection and coverage bias in a survey sample as pointed out by Baker et al. (2013).

In the application to the CCI, the time series modelling approach does not decrease the variance of the direct estimator if it is used for making level estimates. 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. The situation, however, becomes different if the time series model is used to estimate month-to-month change. The stable trend estimates are the result of a strong positive correlation between the trend estimates between subsequent periods. As a result the standard errors of month-to-month change obtained with the time series model are clearly smaller than those of the direct estimates. Standard errors of smoothed month-to-month changes are about 47% smaller than those of the direct estimates. Standard errors of the filtered estimates are about 17% smaller than the standard errors of the 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. However, since both series reflect a similar evolution and social media is rapidly available, the SMI proved to be useful as an auxiliary series in the bivariate model to produce more reliable nowcasts for the CCI in real time at the moment that the SMI becomes available but the CCI is not available yet. In this application the SMI reduces the standard errors of the CCI in a nowcasting procedure with about 17%.

The question can be raised whether the SMI in its current operationalization measures the same concept as the CCI attempts and how the full potentials of social media or other big data sources can be used to measure consumer confidence better than the current CCI and SMI. Instead of constructing a social media index by taking the difference between positive and negative classified messages, an SMI could be constructed by looking at the concepts of the questions used for the CCI. If for example consumer confidence is measured by the amount of purchases of expensive goods during the last 12 months, or with the tendency of households to buy expensive goods, social media indices should be constructed that measure internet search for such goods (cars, houses, white goods, etc.) as well as actual purchases of such goods during the previous months. The strong advantage of this approach is that now actual behaviour of households is measured directly, while a survey measures it indirectly inducing more measurement error. This might eventually result in cointegrated series that measure similar concepts and further improves or even replaces the CCI.

# Acknowledgements

The authors are grateful to the Associate Editor and the reviewers for careful reading of a former draft of this paper and providing constructive comments, which significantly improved the content of this paper. The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands.


# Appendix A

## Model diagnostics

**Table A.1**
**Univariate model (3.8) for CCI 172-24 obs**

| Diagnostic | Value | $p-$value | 95% conf. int. | |
|---|---|---|---|---|
| | | | L | U |
| Log-likelihood | -464 | | | |
| Mean std. innovations | 0.0152 | | | |
| Variance std. innovations | 1.0851 | | | |
| Skewness std. innovations | 0.0276 | | | |
| Kurtosis std. innovations | 2.8901 | | | |
| Bowman-Shenton test[1] on normality in the std. innovations | 0.0926 | 0.955 | | |
| Ljung-Box test[2] on serial correlation in std. innovations | 24.108 | 0.287 | | |
| Durban-Watson test[3] on serial correlation of std. innovations $(T = 148)$ | 2.082 | | 1.68 | 2.32 |
| $F-$test[4] on heteroscedasticity of std. innovations $(df_{\text{num}} = df_{\text{denom}} = 60)$ | 0.913 | | 0.60 | 1.67 |

**Table A.2**
**Bivariate model (3.9) for CCI 57-24 obs**

| Diagnostic | Value | $p-$value | 95% conf. int. | |
|---|---|---|---|---|
| | | | L | U |
| Log-likelihood | -230 | | | |
| Mean std. innovations | -0.0872 | | | |
| Variance std. innovations | 0.9777 | | | |
| Skewness std. innovations | 0.0982 | | | |
| Kurtosis std. innovations | 2.545 | | | |
| Bowman-Shenton test[1] on normality in the std. innovations | 0.3382 | 0.844 | | |
| Ljung-Box test[2] on serial correlation in std. innovations | 18.060 | 0.645 | | |
| Durban-Watson test[3] on serial correlation of std. innovations $(T = 33)$ | 2.133 | | 1.32 | 2.68 |
| $F-$test[4] on heteroscedasticity of std. innovations $(df_{\text{num}} = df_{\text{denom}} = 15)$ | 0.783 | | 0.35 | 2.86 |

1) Bowman-Shenton statistic: $\chi^2_2$ distribution.
2) Ljung-Box test statistic for serial correlation in the first 24 lags: $\chi^2_{21}$ distribution.
3) Durban-Watson test statistic approximated with $N(2, 4/T)$.
4) $F-$statistic: $F^{df_{\text{num}}}_{df_{\text{denom}}}$ distribution.

**Table A.3**
**Bivariate model (3.9) for SMI 57 -12 obs**

| Diagnostic | Value | $p$ – value | 95% conf. int. | |
|---|---|---|---|---|
| | | | L | U |
| Log-likelihood | -230 | | | |
| Mean std. innovations | 0.0954 | | | |
| Variance std. innovations | 1.0437 | | | |
| Skewness std. innovations | -0.1311 | | | |
| Kurtosis std. innovations | 2.5331 | | | |
| Bowman-Shenton test[1] on normality in the std. innovations | 0.5377 | 0.764 | | |
| Ljung-Box test[2] on serial correlation in std. innovations | 24.208 | 0.283 | | |
| Durban-Watson test[3] on serial correlation of std. innovations $(T = 45)$ | 2.028 | | 1.42 | 2.58 |
| $F-$test[4] on heteroscedasticity of std. innovations $(df_{\text{num}} = df_{\text{denom}} = 20)$ | 0.329 | | 0.41 | 2.46 |

1) Bowman-Shenton statistic: $\chi^2_2$ distribution.
2) Ljung-Box test statistic for serial correlation in the first 24 lags: $\chi^2_{21}$ distribution.
3) Durban-Watson test statistic approximated with $N(2, 4/T)$.
4) $F-$statistic: $F^{df_{\text{num}}}_{df_{\text{denom}}}$ distribution.

# 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. (2005). Some considerations of seasonal adjustment variances. Census Bureau. Paper available at https://www.census.gov/ts/papers/jsm2005wrb.pdf.

Bell, W.R., and Hillmer, S.C. (1990). The time series approach to estimation for repeated surveys. *Survey Methodology*, 16, 2, 195-215. Paper available at http://www.statcan.gc.ca/pub/12-001-x/1990002/article/14535-eng.pdf.

Binder, D.A., and Dick, J.P. (1989). Modelling and estimation for repeated surveys. *Survey Methodology*, 15, 1, 29-45. Paper available at http://www.statcan.gc.ca/pub/12-001-x/1989001/article/14579-eng.pdf.

Binder, D.A., and Dick, J.P. (1990). A method for the analysis of seasonal ARIMA models. *Survey Methodology*, 16, 2, 239-253. Paper available at http://www.statcan.gc.ca/pub/12-001-x/1990002/article/14533-eng.pdf.

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, 1073-1076.

Bollineni-Balabay, O., van den Brakel, J.A. 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, Series A*.

Bollineni-Balabay, O., van den Brakel, J.A. and Palm, F. (2017). State space time series modelling of the Dutch Labour Force Survey: Model selection and mean squared errors estimation. *Survey Methodology*, 43, 1, 41-67. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2017001/article/14819-eng.pdf.

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

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

Cochran, W. (1977). *Sampling Theory*. New York: John Wiley & Sons, Inc.

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

Daas, P., and Puts, M. (2014b). Social media sentiment and consumer confidence. European Central Bank Statistics paper series No. 5, Frankfurt Germany.

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

Durbin, J., and Koopman, S.J. (2012). *Time Series Analysis by State Space Methods, Second Edition*. 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, 182-199.

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

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.

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.

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

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.

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.

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, 163-175.

Pfeffermann, D., and Burck, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, 16, 2, 217-237. Paper available at http://www.statcan.gc.ca/pub/12-001-x/1990002/article/14534-eng.pdf.

Pfeffermann, D., and Rubin-Bleuer, S. (1993). Robust joint modelling of labour force series of small areas. *Survey Methodology*, 19, 2, 149-163. Paper available at http://www.statcan.gc.ca/pub/12-001-x/1993002/article/14458-eng.pdf.

Pfeffermann, D., and Sverchkov, M. (2014). Estimation of mean squared error of X-11-ARIMA and other estimators of time series components. *Journal of Official Statistics*, 30, 811-838.

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

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, 339-348.

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

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

Tam, S.-M. (1987). Analysis of repeated surveys using a dynamic linear model. *International Statistical Review/Revue Internationale de Statistique*, 55, 1, 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., and Krieg, S. (2009). Estimation of the monthly unemployment rate through structural time series modelling in a rotating panel design. *Survey Methodology*, 35, 2, 177-190. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2009002/article/11040-eng.pdf.

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*, 41, 2, 267-296. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2015002/article/14231-eng.pdf.