Intensive Courses in the context of the Jean Monnet Chair:

Big data in official statistics

Block 3: Bivariate structural time series model for nowcasting

14 december 2018,

University of Pisa

Jan van den Brakel

Introduction

Purpose of this block:

Combining time series from repeated sample surveys with tme series form big data sources

Motivating example

Statistics Netherlands:

- Consumer confidence survey
- Sentiments index derived from social media platforms
- How to use this additional information?
  - Separate statistic
  - As an auxiliary series to improve accuracy and timeliness of the consumer confidence index

Consumer confidence survey

- Consumer Confidence Index (CCI)
- Monthly cross-sectional survey of 1000 respondents
- Stratified simple random sampling (self weighted)
- Computer assisted telephone interviewing
- CCI:
  - 5 questions to measure sentiment of the Dutch population about the economic climate (economic and financial situation last 12 months and expectations next 12 months)
  - $P^{+}_{q,t}, P^{0}_{q,t}, P^{-}_{q,t}$, $q = 1, ..., 5$

$$I_t = \frac{1}{5}\sum_{q=1}^{5}(P^{+}_{q,t} - P^{-}_{q,t})$$

  - Questions: economic and financial situation last 12 months and expectations next 12 months

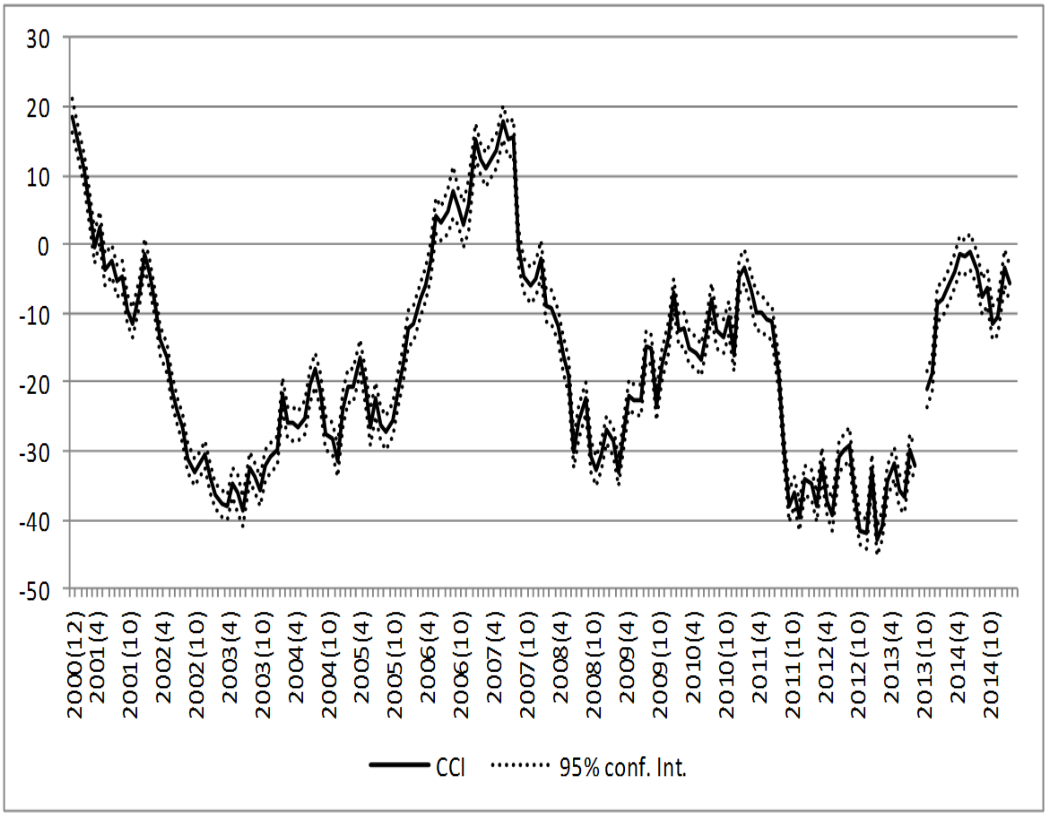


Figure 1: Consumer Confidence Index

Sentiment Index

Sentiment Index Social Media (SMI):

- Derived from Facebook and Twitter (Daas and Puts, 2014)
- Messages are classified as positive or negative
- SMI is the difference between the fraction of positive and negative messages
- High frequency, very timely, no response burden, cost effective

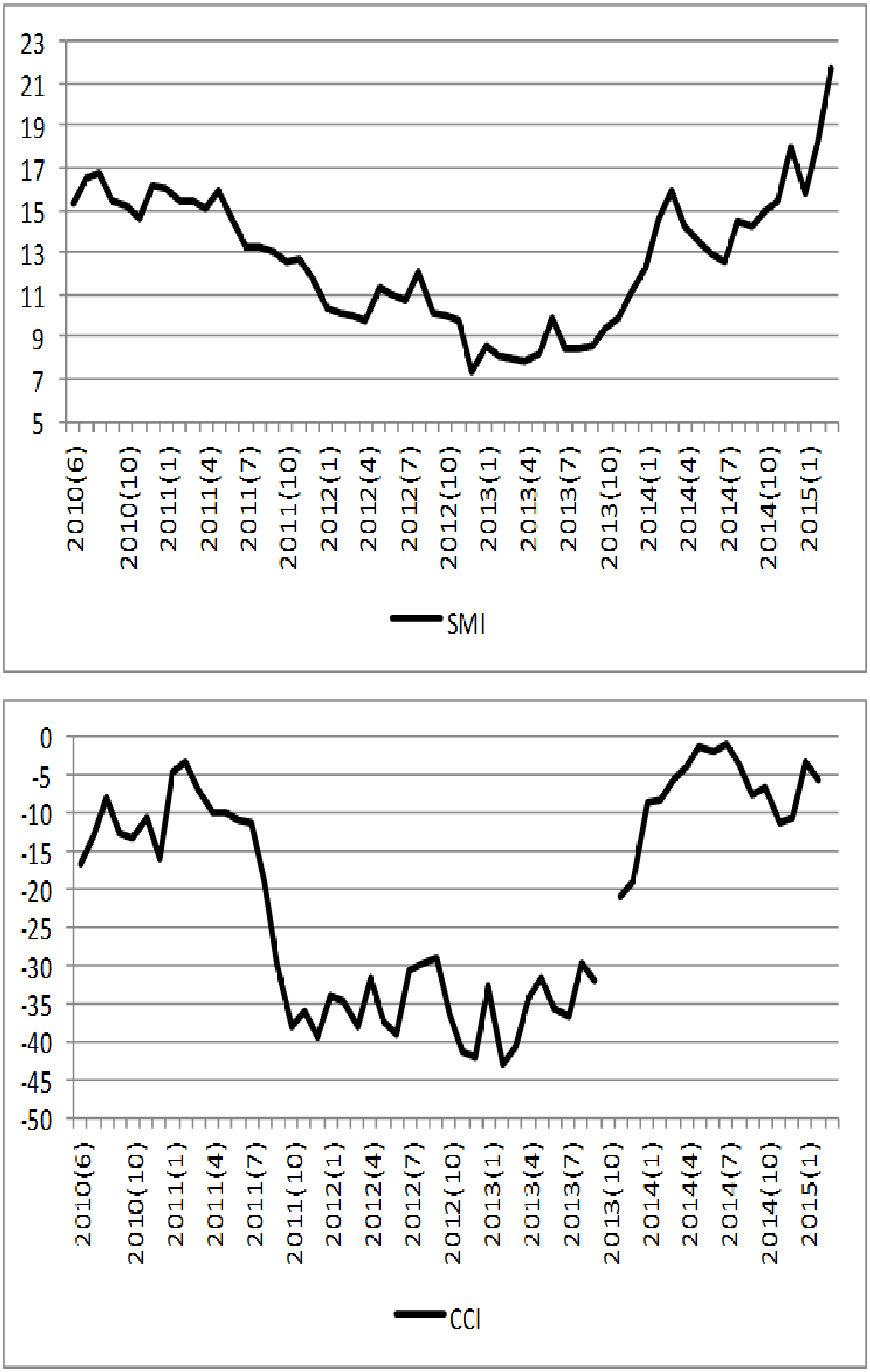


Figure 2: SMI (top) versus CCI (bottom)

Univariate STM CCI

- Measurement error model: $I_t = \theta_t + e_t$
  - I_t: sample estimate CCI
  - θ_t: population value CCI
  - e_t: sample error
- STM for population value: $\theta_t = L_t + S_t + \epsilon_t$
  - L_t: Smooth trend model
  - S_t: Trigonometric seasonal component
  - ϵ_t: population white noise
- STM observed series:

  $I_t = L_t + S_t + \epsilon_t + e_t \equiv L_t + S_t + \nu_t$
  - $\nu_t \simeq \mathcal{N}(0, \sigma_\nu^2)$
  - $Cov(\nu_t, \nu_{t'}) = 0$

- Final model CCI:

  $I_t = L_t + S_t + \beta\delta_t^{11} + \nu_t$

  δ_t models a level shift in 2011(9): economic downturn

  $\nu_t \simeq \mathcal{N}(0, \sigma_\nu^2)$

In case of heteroscedastic sampling errors:

- Time dependent variance structure: $\nu_t \simeq \mathcal{N}(0, Var(\nu_t))$
  - $Var(\nu_t) = Var(I_t)\sigma_\nu^2 \quad Cov(\nu_t, \nu_{t'}) = 0$
  - $Var(I_t)$: sample variance of I_t

Univariate STM SMI

- Final model SMI series 2010-2015:

  $X_t = L_t + \epsilon_t$

  - $\epsilon_t \simeq \mathcal{N}(0, \sigma_\epsilon^2)$
  - $Cov(\epsilon_t, \epsilon_{t'}) = 0$

- L_t: Smooth trend model
- Weak non-significant seasonal pattern
- No level shift required for 2011(9)

Bivariate time series model CCI and SMI

- $$\begin{pmatrix} I_t \\ X_t \end{pmatrix} = \begin{pmatrix} L_t^I \\ L_t^X \end{pmatrix} + \begin{pmatrix} S_t \\ - \end{pmatrix} + \begin{pmatrix} \beta^{11}\delta_t^{11} \\ - \end{pmatrix} + \begin{pmatrix} \nu_t^I \\ \epsilon_t^X \end{pmatrix}$$
- Trend:

$$L_t^I = L_{t-1}^I + R_{t-1}^I, \qquad L_t^X = L_{t-1}^X + R_{t-1}^X,$$

$$R_t^I = R_{t-1}^I + \eta_t^I, \qquad R_t^X = R_{t-1}^X + \eta_t^X,$$

$$\begin{pmatrix} \eta_t^I \\ \eta_t^X \end{pmatrix} \simeq \mathcal{N}\left(\mathbf{0}, \Sigma\right)$$

$$\Sigma = \begin{pmatrix} \sigma_{\eta_I}^2 & \rho_\eta \sigma_{\eta_I}\sigma_{\eta_X} \\ \rho_\eta \sigma_{\eta_I}\sigma_{\eta_X} & \sigma_{\eta_X}^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ a & 1 \end{pmatrix} \begin{pmatrix} d_1 & 0 \\ 0 & d_2 \end{pmatrix} \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix}$$

If $d_2 \to 0$ then $\rho_\eta \to 1$, and

$$\eta_t^X = a\eta_t^I, \quad R_t^X = aR_t^I + \bar{R}, \quad L_t^X = aL_t^I + \bar{L} + t\bar{R},$$

Strong correlation:

- More precise estimates for L_t^I and thus I_t
- $d_2 \rightarrow 0$: cointegration
- Trends of both series are driven by one common trend
- Harvey and Chung (2000)

Alternative model :

$$I_t = L_t + S_t + \beta\delta_t^{11} + \gamma X_t + \nu_t$$

Drawback:

- γX_t absorbs a main part of the trend and the seasonal effect
- L_t residual trend

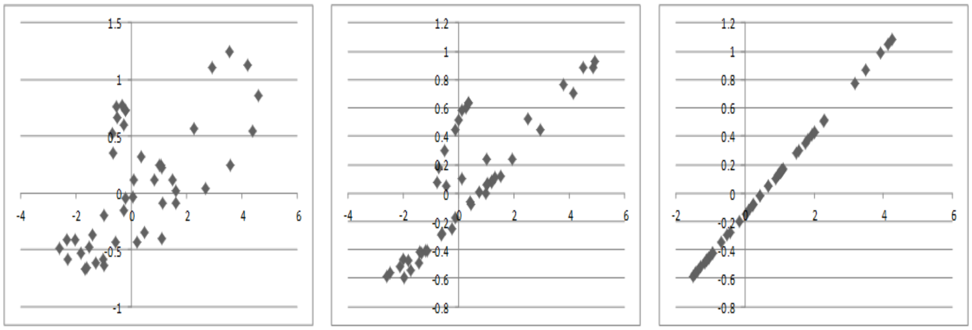
- Structural time series models expressed as state-space models
- Kalman filter to fit the model
- Maximum likelihood for hyperparameters
- Software: OxMetrics with SsfPack (Doornik, 2009; Koopman et al., 2008)

Results

Results hyperparameters

Maximum likelihood estimates hyperparameters		
Hyperparameter	Bivariate	Univariate
SD slope disturbances trend CCI	1.25	1.18
SD slope disturbances trend SMI	0.25	-
Correlation slope disturbances CCI,SMI	0.92	-
SD seasonal disturbances CCI	7.5E-6	0.0025
SD disturbances measurement eq. CCI	2.68	2.46
SD disturbances measurement eq. SMI	0.84	-
Average SE direct estimates CCI	1.21	

Results



Cross plots slope disturbances CCI (x axis) versus SMI (y axis)

Left: $\rho_\eta = 0$ (log likelihood: -234)

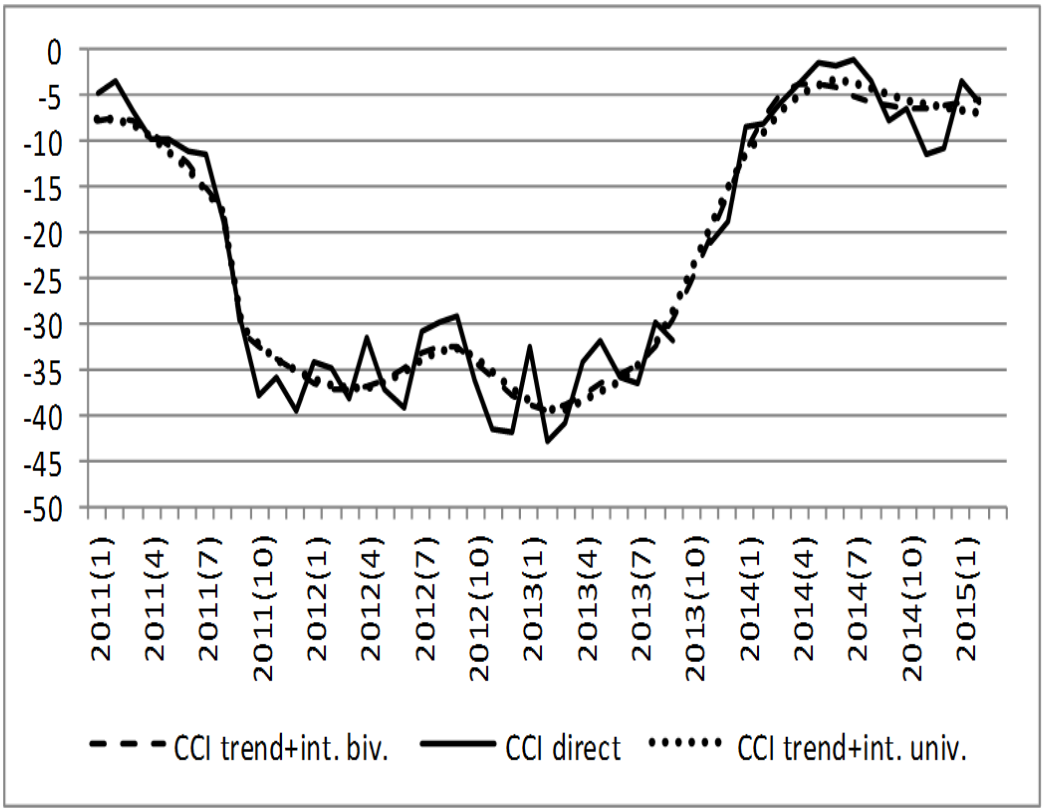
Middle: $\rho_\eta = 0.92$ (log likelihood: -230)

Right: $\rho_\eta = 1.0$ (log likelihood: -242)

p-value LR test on $H_0 : \rho = 0$: 0.0047

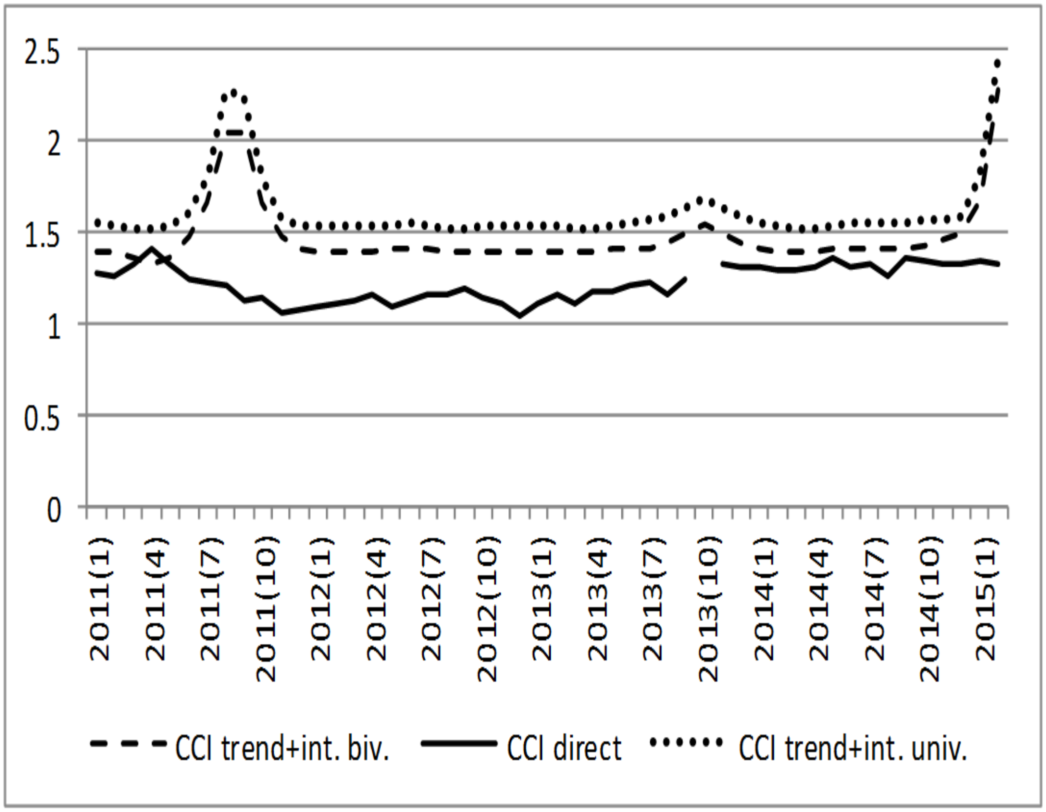
Results

Comparison signal estimates CCI (smoothed estimates)



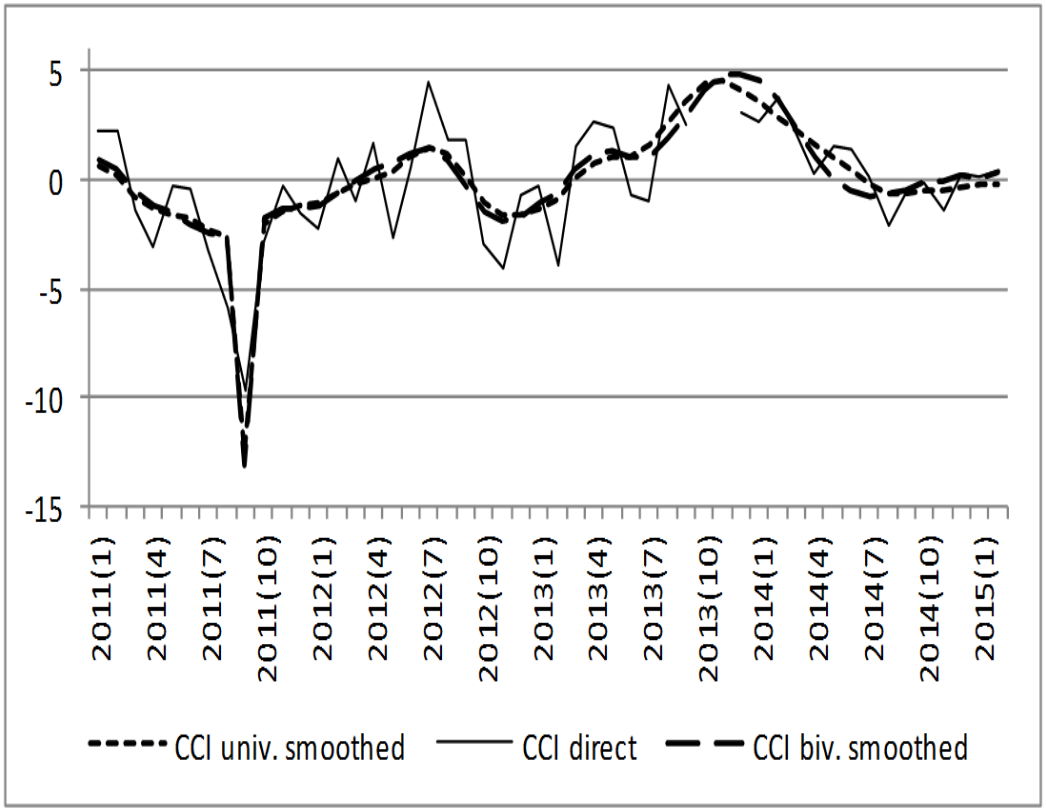
Results

Comparison standard errors of signal estimates CCI



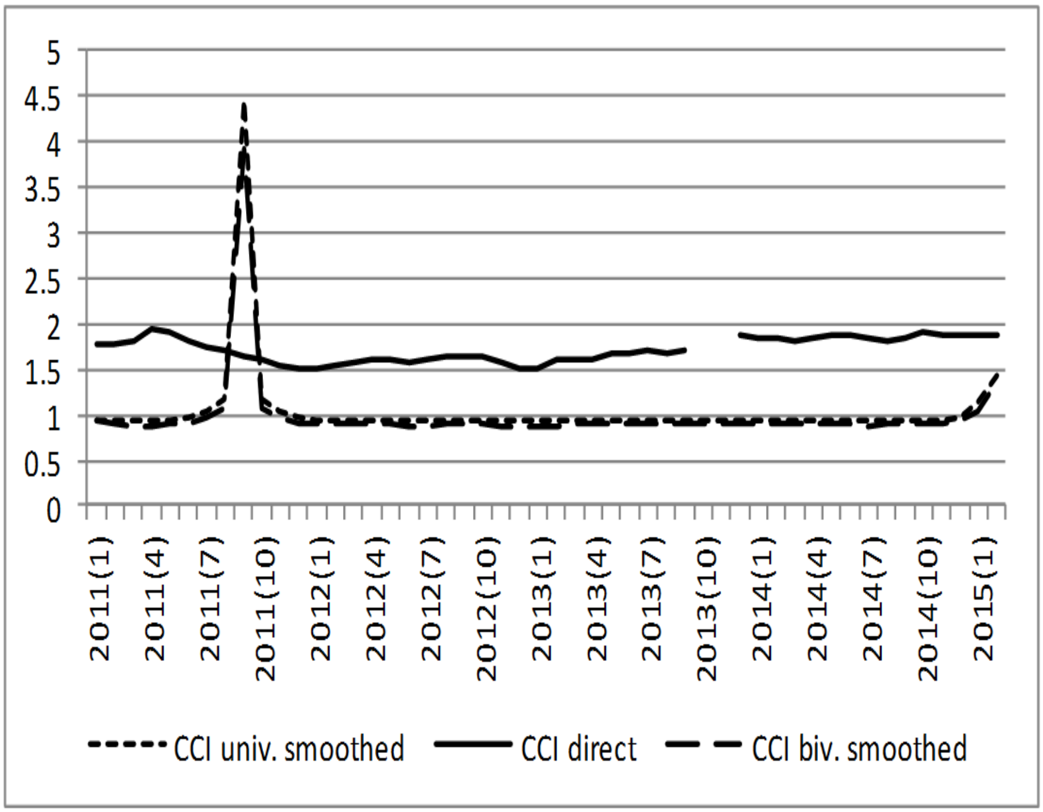
Results

Comparison estimates month-to-month change CCI



Results

Comparison standard errors month-to-month change CCI

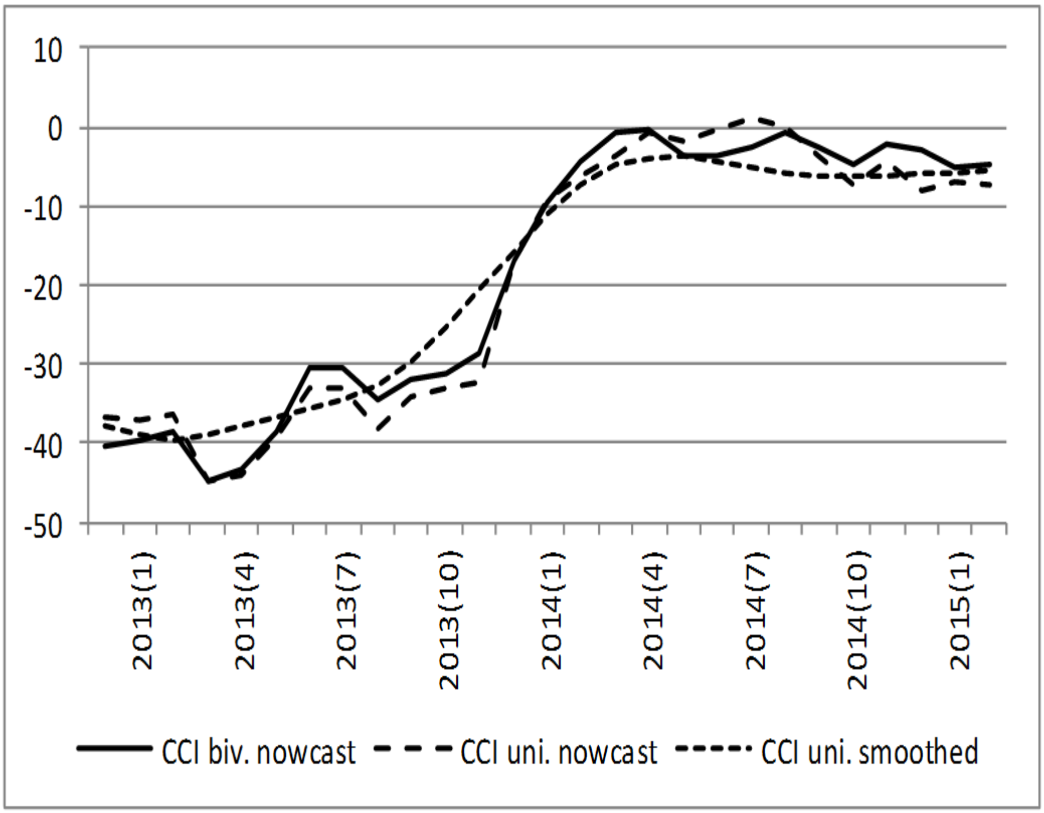


Nowcasting

- Sample surveys are less timely compared to big data sources
- More precise early estimates in real time when SMI is available, but CCI not yet
- Compare:
  - One-step-ahead forecast univeriate model CCI
  - Estimation with the bivariate model where for the last month CCI is missing
  - Benchmark: smoothed estimates univariate model

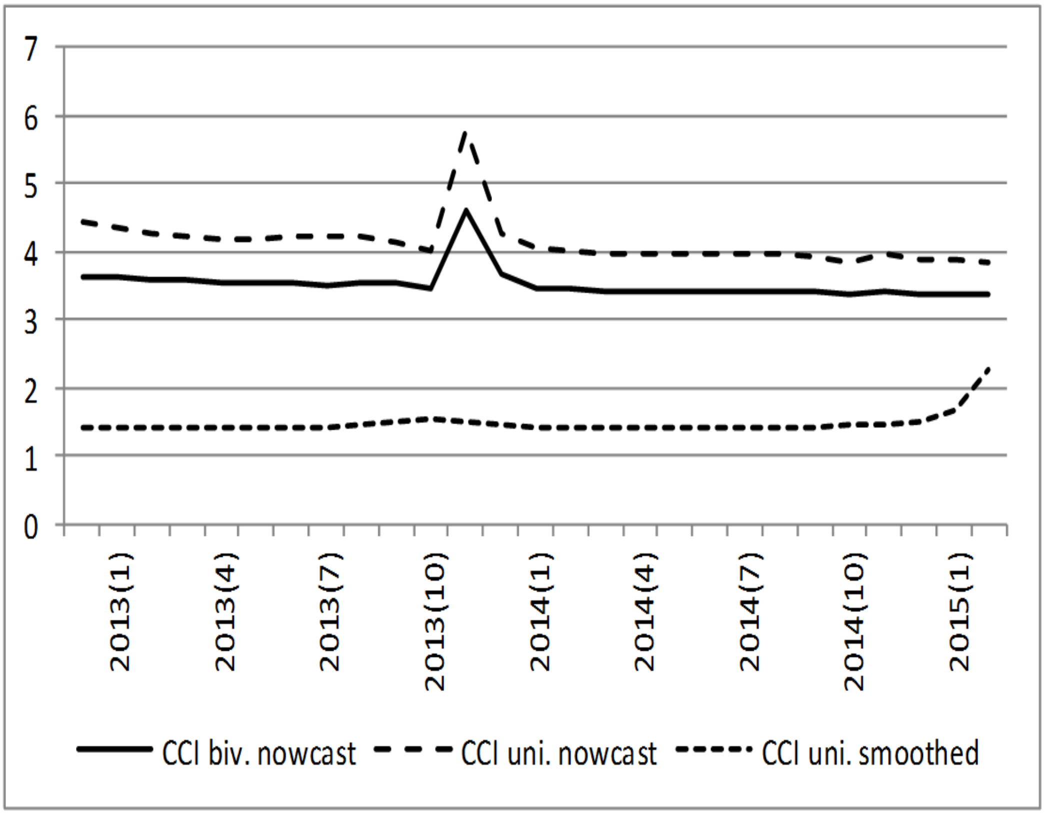
Results nowcasting

Comparison nowcasts bivariate and univariate model CCI



Results nowcasting

Comparison standard errors nowcasts bivariate and univariate model CCI



Discussion

- Official statistics
  - Repeated surveys
  - Time series models appropriate form of SAE
- Bivariate structural time series model
  - Combine series from repeated surveys with auxiliary series
  - Assess similarities between CCI and SMI
  - Improve precision of CCI estimates
  - Form of nowcasting to improve timeliness sample surveys
- Useful approach to borrow strength from auxiliary series and improve timeliness of survey samples
- Details: van den Brakel et al. (2017)

References

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

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

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

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

van den Brakel, J., Söhler, S., Daas, P., and Buelens, B.

(2017). Social media as a data source for official statistics; the Dutch Consumer Confidence Index. *Survey Methodology* 43, 183–210.