Multivariate State-Space Approach to Variance Reduction in Series with Level and Variance Breaks due to Sampling Redesigns

the Case of the Dutch Road Transportation Survey

Oksana Bollineni-Balabay Jan van den Brakel
Franz Palm

Maastricht University/Statistics Netherlands

1-4 Sep 2013, SAE, Bangkok

Outline

- Data
- State-Space Structural Modelling Approach
- Competing alternatives
    - Univariate Models
    - A 9-dimensional model
    - A 10-dimensional model
- Signal Variance Comparison
- Conclusions

The subject of study

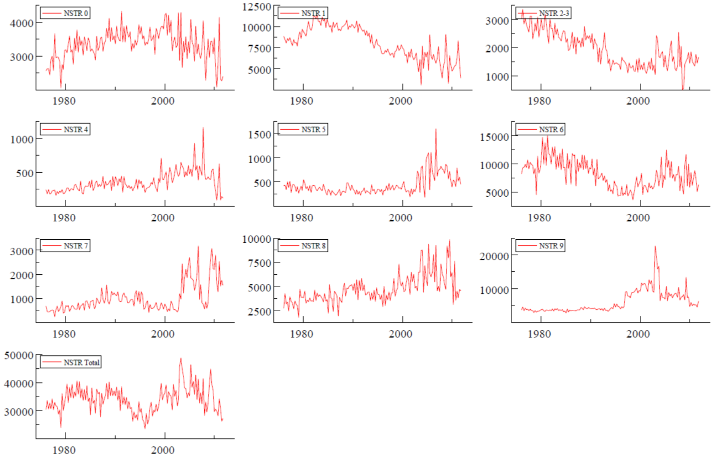
road freight transportation carried out by vehicles registered in the RDW

- domestic
- own-account
- measured in tons
- quarterly, since 1976
- subdivided into 9 NSTR-domains (*Nomenclature uniforme des marchandises pour les Statistiques de Transport, Revise*)

NSTR categories

NSTR 0: Agricultural products and live animals;
NSTR 1: Foodstuff and animal fodder;
NSTR 2/3: Solid mineral fuels; Petroleum oils and petroleum;
NSTR 4: Ores, metal scrap, roasted iron pyrites;
NSTR 5: Iron, steel and non-ferrous metals (including intermediates);
NSTR 6: Crude and manufactured minerals, building materials;
NSTR 7: Fertilizers;
NSTR 8: Chemicals;
NSTR 9: Vehicles, machinery and other goods (including cargo).

Horvitz-Thompson Estimates of the Own Account Transportation Series, 1000 tons



Major Amendments to the Survey design

- **up until 2003:** sampling unit in stratified sampling scheme - the vehicle;
- **2003-2007:** 2-stage stratified sampling design; PSU - the company;
- **from 2008:** back to 1-stage stratified sampling design; sampling unit - the vehicle;

- **decreasing sample sizes throughout the course of the survey**

Modelling Alternatives

- 10 univariate models;
- a 9-dimensional multivariate model $\Rightarrow$ get the total series by summing up the 9 series estimates;
- a 10-dimensional model: 9 domains and 1 national level series.

Decomposition into Unobserved Components

Horvitz-Thompson estimates:

$$\hat{Y}_{d,t} = \theta_{d,t} + e_{d,t} \quad (1)$$

θ is the true value of the population variable,
e_t is a sampling error.

$$\theta_{t,d} = \underbrace{L_{t,d} + \gamma_{t,d} + \mathbf{x}'_{\mathbf{t,d}}\boldsymbol{\beta}_{\mathbf{d}}}_{signal} + \underbrace{\varepsilon_{t,d}}_{irregular\ term} \quad (2)$$

$$\hat{Y}_{t,d} = \underbrace{L_{t,d} + \gamma_{t,d} + \mathbf{x}'_{\mathbf{t,d}}\boldsymbol{\beta}_{\mathbf{d}}}_{\boldsymbol{\alpha}_{\mathbf{t,d}}} + \underbrace{\varepsilon_{t,d} + e_{t,d}}_{\nu_{t,d}} \quad (3)$$

$L_{t,d}$ -trend component;
$\gamma_{t,d}$ - seasonal component;
$\mathbf{x}_{\mathbf{t,d}}$ - K (dummy)regressors;
$\boldsymbol{\beta}_{\mathbf{d}}$ - K regression coefficients.

Univariate Model Estimated

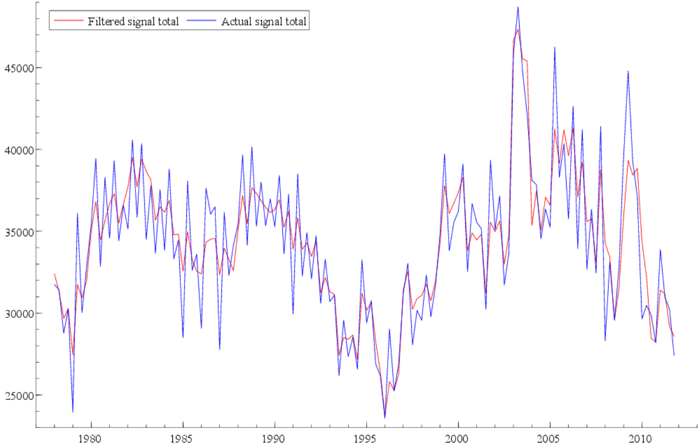
$$\hat{Y}_{t,d} = L_{t,d} + \gamma_{t,d} + x_{t,d}\beta_d + \nu_{t,d}$$

- point-estimates nearly identical to those in multivariate settings;
- variance estimates have a potential for improvement

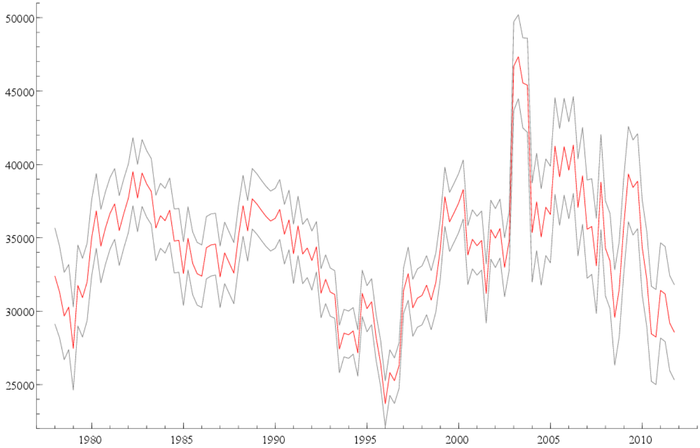
Level and Variance breaks

	Level interventions	Number of breaks in $\sigma^2_{\nu,t,d}$
NSTR 0	-	1
NSTR 1	-	1
NSTR 2/3	2008(3)-(4)	2
NSTR 4	-	1
NSTR 5	-	2
NSTR 6	-	2
NSTR 7	2003(1)-2010(4), 2007(1)-2008(4)	1
NSTR 8	-	1
NSTR 9	1997(1)-2002(4), 2003(1)-(4)	2
Total	2003(1)-(4)	4

Horvitz-Thompson vs. Filtered Signal Estimates of the Total Series from the Ten-Dimensional Model; 1000 tons

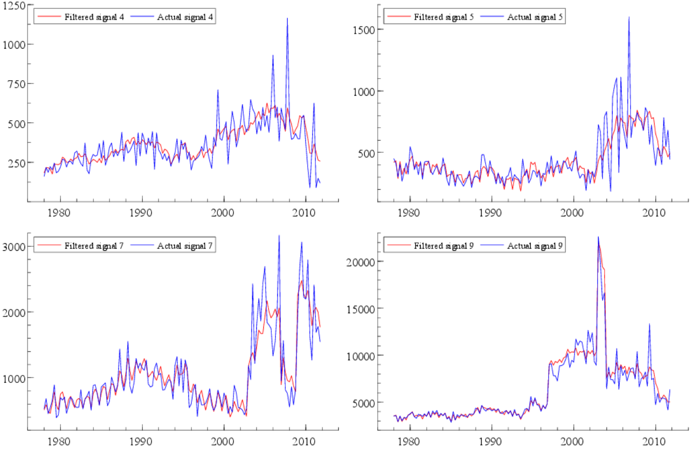


Filtered Signal Estimates of the Total Series from the Ten-Dimensional Model, 1000 tons

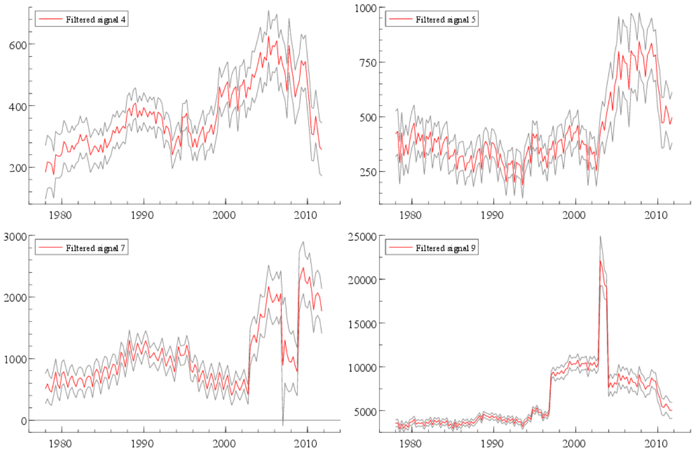


Filtered Trend Estimates of the Total Series from the Ten-Dimensional Model, 1000 tons

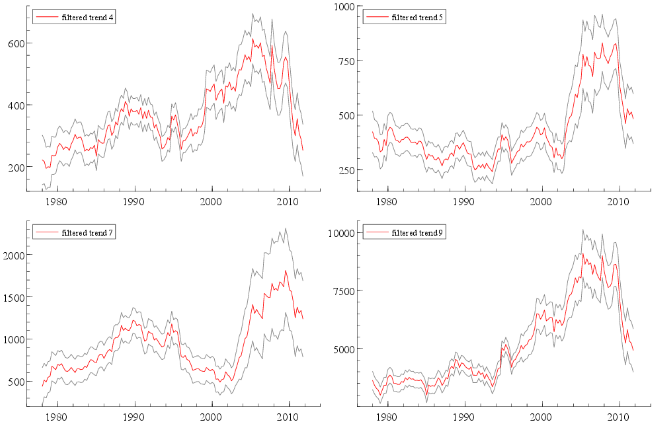
Filtered Signal Estimates from the Ten-Dimensional Model vs. Horvitz-Thompson Estimates, 1000 tons



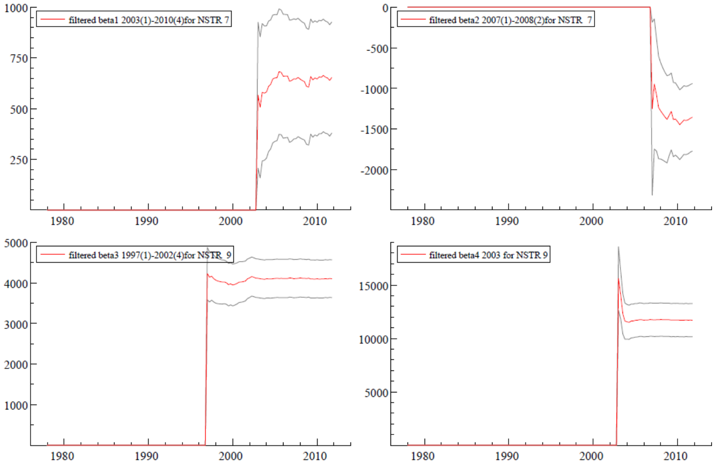
Filtered Signal Estimates from the Ten-Dimensional Model, 1000 tons



Filtered Trend Estimates from the Ten-Dimensional Model, 1000 tons



Filtered level break estimates from the ten-dimensional model (NSTR 7 and 9), 1000 tons



9-dimensional Model Estimated

$$\hat{\mathbf{Y}}_{\mathbf{t}} = \mathbf{L}_{\mathbf{t}} + \gamma_{\mathbf{t}} + \mathbf{x}_{\mathbf{t},\mathbf{1}}\beta_1 + \ldots + \mathbf{x}_{\mathbf{t},\mathbf{5}}\beta_5 + \boldsymbol{\nu}_{\mathbf{t}}$$

Cointegration concept implementation (common factor model):

- D trends are driven by $p < D$ stochastic factors;
- dependent trends expressed as a linear combination of the other trends.

Common Factor Model

Cointegration detection:

modelling covariances between the slope disturbances $\eta_{R,t,d}, \eta_{R,t,d'}$ through the Cholesky decomposition

$$\mathbf{Q_R} = \mathbf{E}(\eta_\mathbf{R} \eta'_\mathbf{R}) = \begin{pmatrix} Q_{11} & Q_{12} & Q_{13} & \cdots & Q_{19} \\ Q_{21} & Q_{22} & Q_{23} & \cdots & Q_{29} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Q_{91} & Q_{92} & Q_{93} & \cdots & Q_{99} \end{pmatrix} = \mathbf{ADA'}$$

an eigenvalue d_{ii} close to zero $\Rightarrow$ dependent trend

Common Factor Model

- reveals a relationship between the domains;
- model parsimony;
- variance reduction.

Common Factor Model

Dependent trends: NSTRs 4, 7, 8 and 9

$$\mathbf{D} = \begin{pmatrix} d_{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & d_{22} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & d_{33} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & d_{55} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & d_{66} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

10-dimensional Model

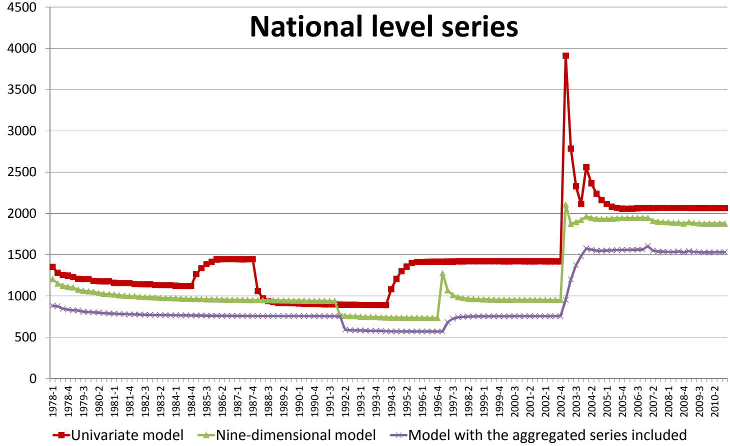
Why at all?

- secures $\sum_{d=1}^{9} \hat{Y}_{d,t} = \hat{Y}_{Total,t}$
- more efficient Kalman filtering $\Rightarrow$ reduced variance in estimates

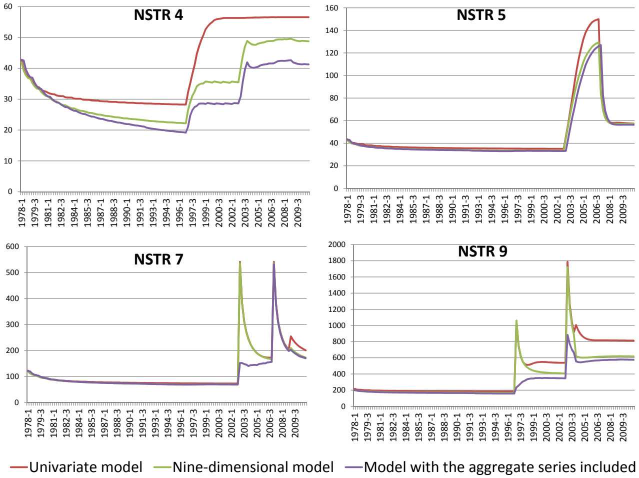
Additional complications/assumptions:

- proper restrictions on the structure of the covariance matrix of disturbance terms;
- composite error terms:
  - non-constant variances;
  - $\Rightarrow$ assumed: constant conditional correlation.
  - $\Rightarrow\Rightarrow$ for simplicity and with little loss in estimate precision, these covariances are set to zero.

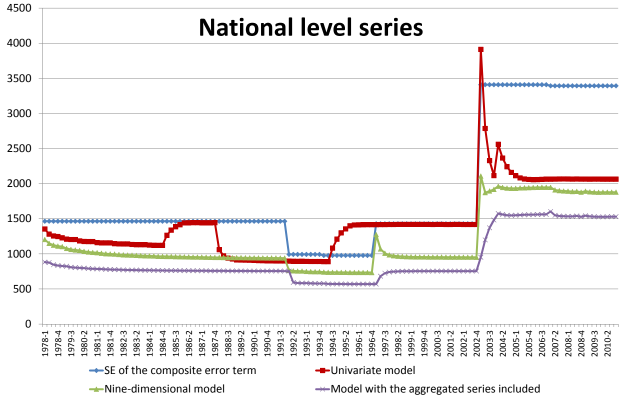
SE of Filtered Signal Estimates; 1000 tons



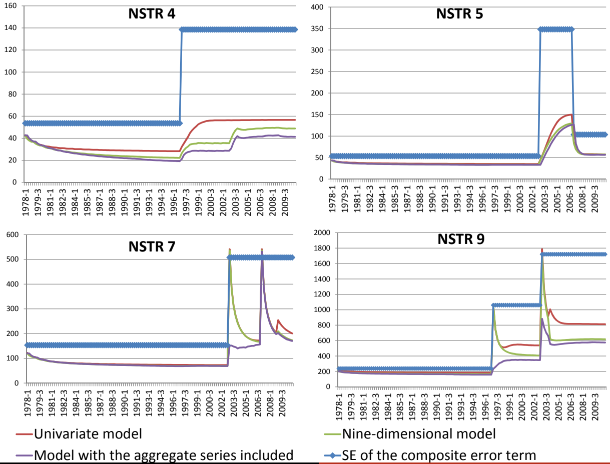
SE of Filtered Signal Estimates; 1000 tons



SE of Filtered Signal Estimates and of Measurement Equation Error Term; 1000 tons



SE of Filtered Signal Estimates and of Measurement Equation Error Term; 1000 tons



Conclusion

- Two problems solved simultaneously:
  - breaks (in the level and variance);
  - small sample sizes.
- The signal variance gets reduced when one moves from the univariate models to multivariate ones.
  $\Rightarrow$ The 10-dimensional model with the aggregate series outperforms all the other models.

Thank you!