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

Oksana Bollineni-Balabay, Jan van den Brakel, Franz Palm
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:

\[
\hat{Y}_{d,t} = \theta_{d,t} + e_{d,t}
\]  

(1)

\(\theta\) is the true value of the population variable, 
\(e_t\) is a sampling error.

\[
\theta_{t,d} = L_{t,d} + \gamma_{t,d} + x'_{t,d}\beta_d + \alpha_{t,d} \varepsilon_{t,d}
\]

(2)

\(
\hat{Y}_{t,d} = L_{t,d} + \gamma_{t,d} + x'_{t,d}\beta_d + \varepsilon_{t,d} + e_{t,d}
\)

(3)

\(L_{t,d}\) - trend component;

\(\gamma_{t,d}\) - seasonal component;

\(x_{t,d}\) - K (dummy) regressors;

\(\beta_d\) - 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

<table>
<thead>
<tr>
<th>NSTR 0</th>
<th>Level interventions</th>
<th>Number of breaks in $\sigma^2_{v,t,d}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>NSTR 1</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>NSTR 2/3</td>
<td>2008(3)-(4)</td>
<td>2</td>
</tr>
<tr>
<td>NSTR 4</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>NSTR 5</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>NSTR 6</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>NSTR 7</td>
<td>2003(1)-2010(4), 2007(1)-2008(4)</td>
<td>1</td>
</tr>
<tr>
<td>NSTR 8</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>NSTR 9</td>
<td>1997(1)-2002(4), 2003(1)-(4)</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>2003(1)-(4)</td>
<td>4</td>
</tr>
</tbody>
</table>
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

- 9-dimensional Model
- 10-dimensional Model
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

\[ \hat{Y}_t = L_t + \gamma_t + x_{t,1}\beta_1 + \ldots + x_{t,5}\beta_5 + \nu_t \]

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

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

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

- reveals a relationship between the domains;
- model parsimony;
- variance reduction.
Dependent trends: NSTRs 4, 7, 8 and 9

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

National level series

- Univariate model
- Nine-dimensional model
- Model with the aggregated series included
SE of Filtered Signal Estimates; 1000 tons

- **NSTR 4**
  - Univariate model
  - Nine-dimensional model
  - Model with the aggregate series included

- **NSTR 5**
  - Univariate model
  - Nine-dimensional model
  - Model with the aggregate series included

- **NSTR 7**
  - Univariate model
  - Nine-dimensional model
  - Model with the aggregate series included

- **NSTR 9**
  - Univariate model
  - Nine-dimensional model
  - Model with the aggregate series included
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.
  ⇒ The 10-dimensional model with the aggregate series outperforms all the other models.
Thank you!