# Modelling Rotation Group Bias and Survey Errors in the Dutch Labour Force Survey

**Jan van den Brakel**, Sabine Krieg
Department of Statistical Methods, Statistics Netherlands, Heerlen, Netherlands

## Abstract

In this paper a multivariate structural time series model is described that accounts for the panel design of the Dutch Labour Force Survey and is applied to estimate monthly unemployment rates. Compared to the generalized regression estimator, this approach results in a substantial increase of the accuracy due to a reduction of the standard error and the explicit modelling of the bias between the subsequent waves.

## 1. Introduction

The Dutch Labour Force Survey (LFS) is based on a rotating panel design. A major problem with such panels is that systematic differences occur between the subsequent waves due to mode and panel effects, which is known as rotation group bias (RGB). The estimation procedure of the LFS is based on the generalized regression (GREG) estimator. These estimators are widely applied by national statistical institutes since they are always approximately design unbiased. They have, however, relatively large design variances in the case of small sample sizes. The monthly sample size of the Dutch LFS is too small to produce reliable figures about employment and unemployment with the GREG estimator. Therefore each month the samples observed in the preceding three months are used to estimate moving averages about the labour market situation.

Since the monthly sample sizes are too small to apply direct survey estimators, model-based estimation procedures can be used to produce sufficiently reliable statistics. For rotating panel designs, Pfeffermann (1991) and Pfeffermann et al. (1998) proposed a structural time series model to borrow information or strength from preceding samples to improve the accuracy of the estimates and to account for the RGB as well as the autocorrelation between the different panels. This approach is applied to the unemployment rate of the Dutch LFS in this paper, which is defined as the ratio of the total unemployment and the total labour force.

In section 2, the survey design of the Dutch LFS is summarised. A structural time series model that accounts for the rotating panel design of the LFS is described in sections 3 and 4. The analysis results are detailed in section 5. Some general remarks are made in section 6.

## 2. The Dutch Labour Force Survey

The LFS is based on a stratified two-stage cluster design of addresses. Strata are formed by geographical regions. Municipalities are considered as primary sampling units and addresses as secondary sampling units. Since most target parameters of the LFS concern people aged 15 through 64 years, addresses with only persons aged 65 years and over are undersampled. In the first wave, data are collected by means of computer assisted personal interviewing (CAPI). The respondents aged 15 through 64 years are re-interviewed four times at quarterly intervals. In these four subsequent waves, data are collected by means of computer assisted telephone interviewing (CATI). During these re-interviews a condensed questionnaire is applied to establish changes in the labour market position of the respondents.

This rotating panel design results in systematic differences in the estimates of the unemployment rate between the successive waves in one time period. These differences are a consequence of panel effects due to systematic changes in the behaviour of the respondents in the panel, panel attrition, mode effects and differences between the CAPI and CATI questionnaires. Due to these factors, the estimates based on the subsequent panels are biased, which is known in the literature as rotation group bias (RGB), see e.g. Bailar (1975) and Pfeffermann (1991). This RGB in the Dutch LFS results in a systematic underestimation of the unemployment rate in the CATI waves and systematic differences in the seasonal effects.

The weighting procedure of the LFS is based on the GREG estimator. The inclusion probabilities reflect the sampling design as well as the different response rates between geographical regions. The weighting scheme is based on a combination of different social-demographical categorical variables. To correct for panel attrition, the inclusion weights of each CATI-wave of the sample data are calibrated with the GREG estimator to the labour force status (10 classes) crossed with age (3 classes) observed in the first wave. In the next step, the calibrated weights of the four CATI waves and the inclusion weights of the CAPI wave are used as the starting weights in the GREG estimator.

The calibration of the CATI waves to the labour force status of the CAPI wave hardly corrects for the RGB. Therefore, an additional rigid correction is applied to the GREG estimate obtained with the five waves. The ratio between the unemployment rate based on CAPI only and the estimates based on all waves is computed using the data of the 12 preceding quarters. Estimates for the preceding three months are multiplied by this ratio to correct for rotation group bias. See Van den Brakel and Krieg (2007) for details.

## 3. Time series model

Let $\theta_t$ denote the population parameter of interest, i.e. the true unemployment rate, at time *t*. Direct estimators, like the GREG estimator, assume that $\theta_t$ is a fixed but unknown parameter. Under this design-based approach, an estimator for $\theta_t$ for cross sectional surveys uses the data observed at time *t*. Data from the past are only used in the case of partially overlapping samples in a panel design. Scott and Smith (1974) proposed to consider the population parameter $\theta_t$ as a realization of a stochastic process that can be described with a time series model. Under this assumption, data observed in preceding periods *t*-1, *t*-2,..., can be used to improve the estimator for $\theta_t$, even in the case of non-overlapping sample surveys.

As a result of the rotating panel design of the Dutch LFS, each month five independent samples are observed to estimate the population parameter $\theta_t$. Let $\mathbf{Y}_t = (Y_t^t\ Y_t^{t-3}\ Y_t^{t-6}\ Y_t^{t-9}\ Y_t^{t-12})^T$ denote a vector containing the five GREG estimates $Y_t^{t-j}$ for $\theta_t$ based on the panel observed at time *t*, which entered the survey for the first time at *t-j*. The $Y_t^{t-j}$ are based on a reduced version of the regular weighting scheme for the quarterly figures without using the correction for the RGB described in section 2. This vector can be modelled as (Pfeffermann, 1991)

$$\mathbf{Y}_t = \mathbf{1}_5\theta_t + \boldsymbol{\lambda}_t + \boldsymbol{\gamma}_t + \mathbf{e}_t, \qquad (3.1)$$

with $\mathbf{1}_5$ a five dimensional vector with each element equal to one, $\boldsymbol{\lambda}_t = (\lambda_t^0\ \lambda_t^3\ \lambda_t^6\ \lambda_t^9\ \lambda_t^{12})^T$ and $\boldsymbol{\gamma}_t = (\gamma_t^0\ \gamma_t^3\ \gamma_t^6\ \gamma_t^9\ \gamma_t^{12})^T$ vectors with time dependent components that account for the RGB of the trend and the seasonal components respectively, and $\mathbf{e}_t = (e_t^t\ e_t^{t-3}\ e_t^{t-6}\ e_t^{t-9}\ e_t^{t-12})^T$ the corresponding survey errors for each panel estimate.

### 3.1 Time series model for the population parameter

The population parameter $\theta_t$ is modelled with the basic structural time series model, i.e.:

$$\theta_t = L_t + S_t + \varepsilon_t, \qquad (3.2)$$

where $L_t$ denotes a stochastic trend component, $S_t$ a stochastic seasonal component, and $\varepsilon_t$ the irregular component which contains the unexplained variation that is modelled as white noise. The stochastic trend is modelled as a smooth trend model, which is defined by:

$$L_t = L_{t-1} + R_{t-1},\ R_t = R_{t-1} + \eta_{R,t},\ E(\eta_{R,t}) = 0,$$

$$Cov(\eta_{R,t}, \eta_{R,t'}) = \begin{cases} \sigma_R^2 & \text{if} \quad t = t' \\ 0 & \text{if} \quad t \neq t'. \end{cases} \qquad (3.3)$$

The parameters $L_t$ and $R_t$ are referred to as the trend and the slope parameter respectively. The seasonal component is modelled as

$$\sum_{j=0}^{11} S_{t-j} = \eta_{S,t}, \quad E(\eta_{S,t}) = 0,$$

$$Cov(\eta_{S,t}, \eta_{S,t'}) = \begin{cases} \sigma_S^2 & \text{if} \quad t = t' \\ 0 & \text{if} \quad t \neq t'. \end{cases} \qquad (3.4)$$

### 3.2 Time series model for rotation group bias

The systematic differences between the trend and the seasonal components of the subsequent waves are modelled with $\boldsymbol{\lambda}_t$ and $\boldsymbol{\gamma}_t$. Additional restrictions for the elements of both vectors are required to identify model (3.1). Here it is assumed that the most accurate estimate for $\theta_t$ is obtained with the first wave, which is observed by CAPI. This implies that the first components of $\boldsymbol{\lambda}_t$ and $\boldsymbol{\gamma}_t$ equal zero. Now $\boldsymbol{\lambda}_t$ measures the time dependent differences in the low frequency variation with respect to the first wave. The components of $\boldsymbol{\lambda}_t$ are defined as:

$$\lambda_t^0 = 0, \quad \lambda_t^j = \lambda_{t-1}^j + \eta_{\lambda,j,t},\ j{=}3,6,9,12, \qquad (3.5)$$

$$E(\eta_{\lambda,j,t}) = 0,$$

$$Cov(\eta_{\lambda,j,t}, \eta_{\lambda,j',t'}) = \begin{cases} \sigma_\lambda^2 & : \quad t = t' \text{ and } j = j' \\ 0 & : \quad t \neq t' \text{ or } j \neq j'. \end{cases}$$

$\boldsymbol{\gamma}_t$ measures the systematic differences in the seasonal components with respect to the first wave. The components of $\boldsymbol{\gamma}_t$ are defined as

$$\gamma_t^0 = 0, \quad \sum_{l=0}^{11} \gamma_{t-l}^j = \eta_{\gamma,j,t},\ j{=}3,6,9,12, \qquad (\ 3.6)$$

$$E(\eta_{\gamma,j,t}) = 0,$$

$$Cov(\eta_{\gamma,j,t}, \eta_{\gamma,j',t'}) = \begin{cases} \sigma_\gamma^2 & : \quad t = t' \text{ and } j = j' \\ 0 & : \quad t \neq t' \text{ or } j \neq j'. \end{cases}$$

The variance components of the random walks in (3.5) and the seasonal components in (3.6) are assumed to be equal for all waves.

### 3.3 Time series model for survey errors

A consequence of the rotating panel design is that the survey errors in the subsequent time periods are correlated. To account for this autocorrelation, the dependency between the survey errors of the panel observed at the last time and previous occasions is modelled with the following autoregressive relationship:

$$\begin{aligned}
e_t^t &= \eta_{e,0,t}, \\
e_t^{t-3} &= \rho_1 e_{t-3}^{t-3} + \eta_{e,3,t}, \\
e_t^{t-6} &= \rho_1 e_{t-3}^{t-6} + \rho_2 e_{t-6}^{t-6} + \eta_{e,6,t}, \\
e_t^{t-9} &= \rho_1 e_{t-3}^{t-9} + \rho_2 e_{t-6}^{t-9} + \eta_{e,9,t}, \\
e_t^{t-12} &= \rho_1 e_{t-3}^{t-12} + \rho_2 e_{t-6}^{t-12} + \eta_{e,12,t}, \\
E(\eta_{e,j,t}) &= 0, \\
Cov(\eta_{e,j,t}, \eta_{e,j',t'}) &= \begin{cases} \dfrac{\sigma_{e,j}^2}{n_t^{t-j}} & : \ t = t' \text{ and } j = j' \\ 0 & : \ t \neq t' \text{ or } j \neq j'. \end{cases}
\end{aligned} \tag{3.7}$$

Here $n_t^{t-j}$ denotes the net number of respondents in the survey at time *t* that entered the panel at time *t-j*, and $\rho$ the autocorrelation coefficients of the AR model.

## 4. State-space representation

The model proposed in the preceding section can be analysed by means of the Kalman filter. To this end, the model is expressed in state-space representation, see Harvey (1989) or Durbin and Koopman (2001). A state-space model consists of a measurement equation and a transition equation. The measurement equation specifies how the observations depend on a linear combination of unobserved state variables, e.g. trend, seasonal, RGB and the survey errors. Thus

$$\mathbf{Y}_t = \mathbf{Z}\boldsymbol{\alpha}_t. \tag{4.1}$$

Here $\boldsymbol{\alpha}_t$ denotes the state vector with unobservable state variables, and $\mathbf{Z}$ a known design matrix that specifies the linear relationship between the observations and the elements of the state vector.

The transition equation specifies how the state vector evolves in time:

$$\boldsymbol{\alpha}_t = \mathbf{T}\boldsymbol{\alpha}_{t-1} + \boldsymbol{\eta}_t, \tag{4.2}$$

with

$$E(\boldsymbol{\eta}_t) = \mathbf{0},$$

$$Cov(\boldsymbol{\eta}_t, \boldsymbol{\eta}_{t'}) = \begin{cases} \mathbf{Q}_t & \text{if} \quad t = t' \\ \mathbf{O} & \text{if} \quad t \neq t'. \end{cases}$$

Here **0** and **O** denote a vector respectively a matrix with each element zero. The state-space representation of the model proposed in section 3 is obtained with (4.1) and (4.2) by taking

$$\boldsymbol{\alpha}_t = (\boldsymbol{\alpha}_t^\theta \ \boldsymbol{\alpha}_t^\lambda \ \boldsymbol{\alpha}_t^\gamma \ \boldsymbol{\alpha}_t^e)^T,$$

$$\boldsymbol{\alpha}_t^\theta = (L_t \ R_t \ S_t \ldots S_{t-10}), \ \boldsymbol{\alpha}_t^\lambda = (\lambda_t^3 \ \lambda_t^6 \ \lambda_t^9 \ \lambda_t^{12}),$$

$$\boldsymbol{\alpha}_t^\gamma = (\gamma_t^3 \ldots \gamma_{t-10}^3 \ \gamma_t^6 \ldots \gamma_{t-10}^6 \ \gamma_t^9 \ldots \gamma_{t-10}^9 \ \gamma_t^{12} \ldots \gamma_{t-10}^{12}),$$

$$\boldsymbol{\alpha}_t^e = (e_t^t \ e_t^{t-3} \ e_t^{t-6} \ e_t^{t-9} \ e_t^{t-12} \ e_{t-2}^{t-2} \ e_{t-2}^{t-5} \ e_{t-2}^{t-8} \ e_{t-2}^{t-11} \ e_{t-1}^{t-1} \ e_{t-1}^{t-4} \ e_{t-1}^{t-7} \ e_{t-1}^{t-10} \ e_{t-5}^{t-5} \ e_{t-5}^{t-8} \ e_{t-5}^{t-11} \ e_{t-4}^{t-4} \ e_{t-4}^{t-7} \ e_{t-4}^{t-10} \ e_{t-3}^{t-3} \ e_{t-3}^{t-6} \ e_{t-3}^{t-9}),$$

$$\mathbf{Z} = (\mathbf{Z}^\theta \ \mathbf{Z}^\lambda \ \mathbf{Z}^\gamma \ \mathbf{Z}^e),$$

$$\mathbf{Z}^\theta = \left(\mathbf{1}_5 \ \mathbf{0}_5 \ \mathbf{1}_5 \ \mathbf{O}_{5\times10}\right), \ \mathbf{Z}^e = \left(\mathbf{I}_5 \ \mathbf{O}_{5\times17}\right),$$

$$\mathbf{Z}^\lambda = \begin{pmatrix} \mathbf{0}_4^T \\ \mathbf{I}_4 \end{pmatrix}, \quad \mathbf{Z}^\gamma = \begin{pmatrix} \mathbf{0}_4^T \\ \mathbf{I}_4 \end{pmatrix} \otimes (1 \ \mathbf{0}_{10}^T),$$

$$\mathbf{T} = \text{Blockdiag}(\mathbf{T}^L \ \mathbf{T}^S \ \mathbf{T}^\lambda \ \mathbf{T}^\gamma \ \mathbf{T}^e),$$

$$\mathbf{T}^L = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{T}^S = \begin{pmatrix} -\mathbf{1}_{10}^T & -1 \\ \mathbf{I}_{10} & \mathbf{0}_{10} \end{pmatrix},$$

$$\mathbf{T}^\lambda = \mathbf{I}_4, \quad \mathbf{T}^\gamma = \mathbf{I}_4 \otimes \mathbf{T}^S,$$

$$\mathbf{T}^e = \begin{pmatrix} \mathbf{T}^{e11} & \mathbf{T}^{e12} \\ \mathbf{T}^{e21} & \mathbf{T}^{e22} \end{pmatrix}, \mathbf{T}^{e22} = \begin{pmatrix} \mathbf{O}_{6\times3} & \mathbf{I}_{6\times6} \\ \mathbf{O}_{3\times3} & \mathbf{O}_{3\times6} \end{pmatrix},$$

$$\mathbf{T}^{e11} = \begin{pmatrix} \mathbf{0}_4^T & 0 & \mathbf{0}_4^T & \mathbf{0}_4^T \\ \mathbf{O}_{4\times4} & \mathbf{0}_4 & \rho_1\mathbf{I}_{4\times4} & \mathbf{O}_{4\times4} \\ \mathbf{O}_{4\times4} & \mathbf{0}_4 & \mathbf{O}_{4\times4} & \mathbf{I}_{4\times4} \\ \mathbf{I}_{4\times4} & \mathbf{0}_4 & \mathbf{O}_{4\times4} & \mathbf{O}_{4\times4} \end{pmatrix},$$

$$\mathbf{T}^{e12} = \begin{pmatrix} \mathbf{O}_{2\times3} & \mathbf{O}_{2\times6} \\ \rho_2\mathbf{I}_{3\times3} & \mathbf{O}_{3\times6} \\ \mathbf{O}_{8\times3} & \mathbf{O}_{8\times6} \end{pmatrix},$$

$$\mathbf{T}^{e21} = \begin{pmatrix} \mathbf{O}_{3\times5} & \mathbf{O}_{3\times3} & \mathbf{O}_{3\times5} \\ \mathbf{O}_{3\times5} & \mathbf{O}_{3\times3} & \mathbf{O}_{3\times5} \\ \mathbf{O}_{3\times5} & \mathbf{I}_{3\times3} & \mathbf{O}_{3\times5} \end{pmatrix},$$

$$\boldsymbol{\eta}_t = (\boldsymbol{\eta}_t^\theta \ \boldsymbol{\eta}_t^\lambda \ \boldsymbol{\eta}_t^\gamma \ \boldsymbol{\eta}_t^e)^T,$$

$$\boldsymbol{\eta}_t^\theta = (0 \ \eta_{R,t} \ \eta_{S,t} \ \mathbf{0}_{10}^T),$$

$$\boldsymbol{\eta}_t^\lambda = (\eta_{\lambda,3,t} \ \eta_{\lambda,6,t} \ \eta_{\lambda,9,t} \ \eta_{\lambda,12,t}),$$

$$\boldsymbol{\eta}_t^\gamma = (\eta_{\gamma,3,t} \ \eta_{\gamma,6,t} \ \eta_{\gamma,9,t} \ \eta_{\gamma,12,t}) \otimes (1 \ \mathbf{0}_{10}^T),$$

$$\boldsymbol{\eta}_t^e = (\eta_{e,0,t} \ \eta_{e,3,t} \ \eta_{e,6,t} \ \eta_{e,9,t} \ \eta_{e,12,t} \ \mathbf{0}_{17}^T),$$

$$\mathbf{Q}_t = \text{Blockdiag}(\mathbf{Q}^\theta \ \mathbf{Q}^\lambda \ \mathbf{Q}^\gamma \ \mathbf{Q}_t^e),$$

$$\mathbf{Q}^\theta = \text{Diag}(0 \ \sigma_R^2 \ \sigma_S^2 \ \mathbf{0}_{10}^T), \ \mathbf{Q}^\lambda = \sigma_\lambda^2 \mathbf{I}_4,$$

$$\mathbf{Q}^{\gamma} = \mathrm{Diag}[\sigma_{\gamma}^{2}\mathbf{1}_{4}^{T} \otimes (1\ \mathbf{0}_{10}^{T})]\ ,$$

$$\mathbf{Q}_{t}^{e} = \mathrm{Diag}\left(\frac{\sigma_{e,0}^{2}}{n_{t}^{t}}\ \frac{\sigma_{e,3}^{2}}{n_{t}^{t-3}}\ \frac{\sigma_{e,6}^{2}}{n_{t}^{t-6}}\ \frac{\sigma_{e,9}^{2}}{n_{t}^{t-9}}\ \frac{\sigma_{e,12}^{2}}{n_{t}^{t-12}}\ \mathbf{0}_{17}^{T}\right).$$

Here $\mathbf{1}$ denotes a vector with each element one and $\mathbf{I}$ the identity matrix. The subscripts for $\mathbf{0}$, $\mathbf{1}$, $\mathbf{O}$, and $\mathbf{I}$ specify the dimensions of the vectors and the matrices.

Generally the measurement equation (4.1) also has an irregular term. The Kalman filter assumes that the disturbances of the measurement equations at different time periods are uncorrelated. This assumption is not met if the survey errors of the panel are incorporated in the irregular terms of the measurement equation. Therefore the survey errors are incorporated as unobserved components in the state vector and the dependency between the survey errors is explicitly modelled in the transition equation. In this application the irregular term of the population parameter in equation (3.2) is denominated by the survey errors which are already incorporated in the state vector. Assuming an additional irregular term in the measurement equation would result in identification problems. Therefore the irregular component of the measurement equation has been dropped.

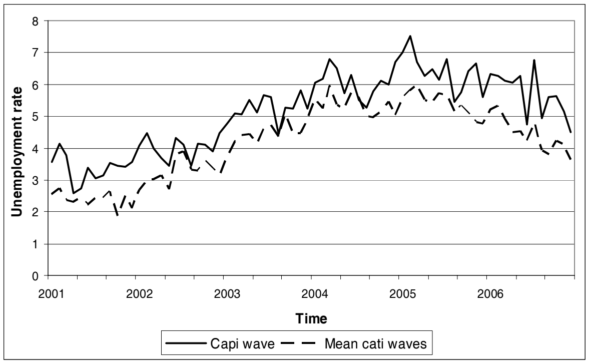
The transitional relationship of the survey errors is explained by Van den Brakel (2005). The transitional relations for the first five entries of $\boldsymbol{\alpha}_{t}^{e}$ follow from (3.7). The remaining elements are included to have the same elements in $\boldsymbol{\alpha}_{t}^{e}$ and $\boldsymbol{\alpha}_{t-1}^{e}$ with a time shift of 1 and to assure that the vector $\boldsymbol{\eta}_{t}^{e}$ is independent of past state vectors. This last property is required since the Kalman filter assumes that $Cov(\boldsymbol{\eta}_{t}, \boldsymbol{\eta}_{t'}) = \mathbf{O}$ for *t*≠ *t'*.

After having expressed the model in state-space form, the Kalman filter can be applied to obtain optimal estimates for the state vector $\boldsymbol{\alpha}_{t}$. 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. The filtered estimates of past state vectors can be updated, if new data become available. This procedure is referred to as smoothing. In this paper, the Kalman filter estimates for the state variables are smoothed with the fixed interval smoother. See Harvey (1989) or Durbin and Koopman (2002) for technical details.

## 5 Results

With the GREG estimator monthly estimates for the unemployment rate are obtained for each wave. In Figure 5.1 the unemployment rate based on the first wave is compared with the average of the four CATI waves. It follows that the unemployment rate observed with the first wave is systematically higher than the other four waves.

*Figure 5.1: RGB monthly unemployment rate based on GREG estimates*



### 5.1 Estimation results for the time series model

The five time series obtained with the different waves are modelled with the time series model proposed in sections 3 and 4. The analysis is conducted with software developed in Ox in combination with the subroutines of SsfPack (beta 3), see Doornik (1998) and Koopman et al. (1999). Note that a more recent version of Ssfpack is used than the 2.2 version described in Koopman et al. (1999). Version 3 is very appropriate for the estimation of complex multivariate structural time series models.

Preliminary analyses indicate that the model proposed in sections 3 and 4 can be simplified. The estimates for the RGB of the seasonal effects in the second wave are not significantly different from zero and the RGB for the seasonal effects of the third, fourth and fifth wave are not significantly different from each other. Therefore the model is simplified by taking $\boldsymbol{\alpha}_{t}^{\gamma} = (\gamma_{t} \ldots \gamma_{t-10})^{T}$ , $\mathbf{T}^{\gamma} = \mathbf{T}^{S}$ , $\mathbf{Z}^{\gamma} = (00111)^{T} \otimes (1\mathbf{0}_{10}^{T})$ , $\boldsymbol{\eta}_{t}^{\gamma} = (\eta_{\gamma}\ \mathbf{0}_{10}^{T})$ , $\mathbf{Q}^{\gamma} = \mathrm{Diag}[\sigma_{\gamma}^{2}\ \mathbf{0}_{10}^{T}]$. Furthermore the estimate for the AR parameter of lag 2 tends to zero, so the model for the survey errors (3.7) is simplified to an AR(1) model. This saves one hyperparameter $(\rho_{2})$ and nine state variables $(e_{t-5}^{t-5}\ e_{t-5}^{t-8}\ e_{t-5}^{t-11}\ e_{t-4}^{t-4}\ e_{t-4}^{t-7}\ e_{t-4}^{t-10}\ e_{t-3}^{t-3}\ e_{t-3}^{t-6}\ e_{t-3}^{t-9})$. The state space equations are simplified accordingly.

Maximum likelihood estimates for the hyperparameters, i.e. the variance components of the stochastic processes for the state variables $\sigma_{R}^{2}, \sigma_{S}^{2}, \sigma_{\lambda}^{2}, \sigma_{\gamma}^{2}, \sigma_{e,0}^{2}, \sigma_{e,3}^{2}, \sigma_{e,6}^{2}, \sigma_{e,9}^{2}, \sigma_{e,12}^{2}$, and the AR parameter between the survey errors $\rho_{1}$, are obtained using a numerical optimization procedure. The results are presented in Table 5.1.

*Table 5.1: Maximum likelihood estimates hyperparameters*

| Hyperparameter | estimate |
|---|---|
| Slope (smooth trend) | 0.000192 |
| Seasonal | 0.000325 |
| RGB trend | 0.000000 |
| RGB seasonal | 0.000000 |
| Survey error wave 1 | 0.343 |
| Survey error wave 2 | 0.256 |
| Survey error wave 3 | 0.293 |
| Survey error wave 4 | 0.336 |
| Survey error wave 5 | 0.278 |
| First order auto regression survey error | 0.2 |

The smoothed Kalman filter estimates for the unemployment rate $\theta_t$ are given in Figure 5.2. These are the estimates for the monthly unemployment rate, based on the smooth trend model and a seasonal component, corrected for the RGB between the five GREG estimates. The trend and the seasonal component are time dependent since the maximum likelihood estimates of the corresponding hyperparameter are positive (see Table 5.1). The smoothed Kalman filter estimates for the trend and the seasonal component are plotted in Figures 5.3 and 5.4 respectively.

*Figure 5.2 Smoothed Kalman filter estimates for the monthly unemployment rate*

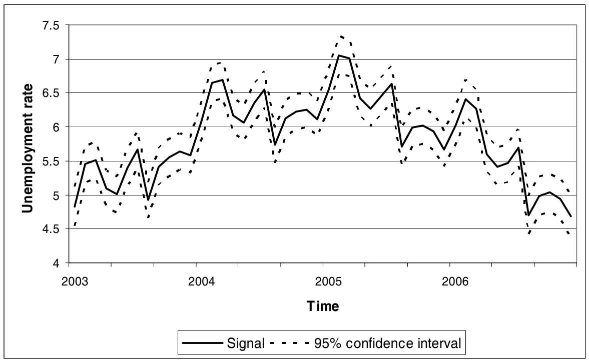


*Figure 5.3 Smoothed Kalman filter estimates for the trend of the monthly unemployment rate*

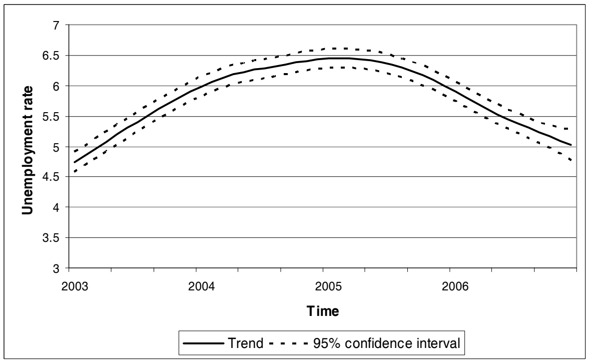
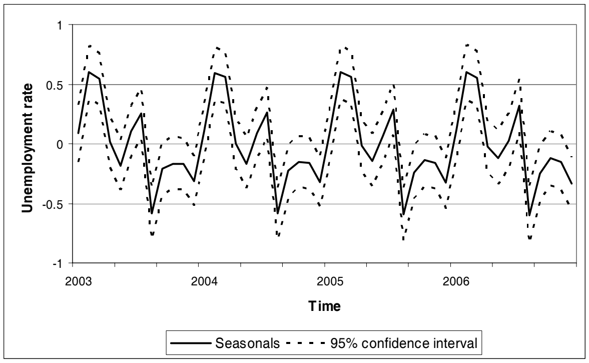


*Figure 5.4 Smoothed Kalman filter estimates for the seasonal effect of the monthly unemployment rate*



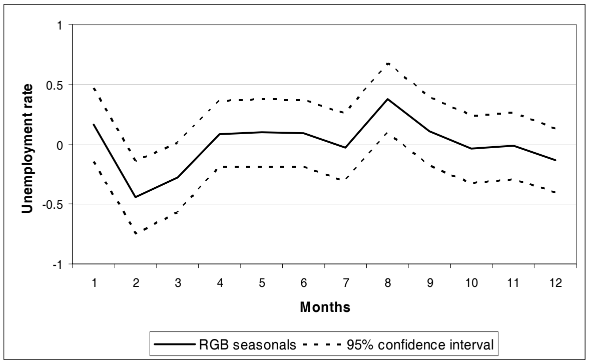
The Kalman filter estimates for the RGB of the trend are time independent since the maximum likelihood estimate of the corresponding hyperparameter tends to zero (see Table 5.1). The smoothed Kalman filter estimates for the RGB are given in Table 5.2. The model beautifully detects a slightly increasing bias in the low frequency variation of the subsequent waves. The estimates for the RGB of the four CATI waves are significantly different from zero.

*Table 5.2 Smoothed Kalman filter estimates RGB trend*

| Wave | RGB | St. error |
|---|---|---|
| 2 | -0.77 | 0.06 |
| 3 | -0.89 | 0.07 |
| 4 | -0.94 | 0.08 |
| 5 | -1.11 | 0.07 |

The Kalman filter estimates for the RGB of the seasonal effects are also time independent since the maximum likelihood estimate of the corresponding hyperparameter tends to zero (see Table 5.1). The smoothed Kalman filter estimates are given in Figure 5.5.

*Figure 5.5: Smoothed Kalman filter estimates for the RGB of the seasonal effects in the third, fourth and fifth wave*



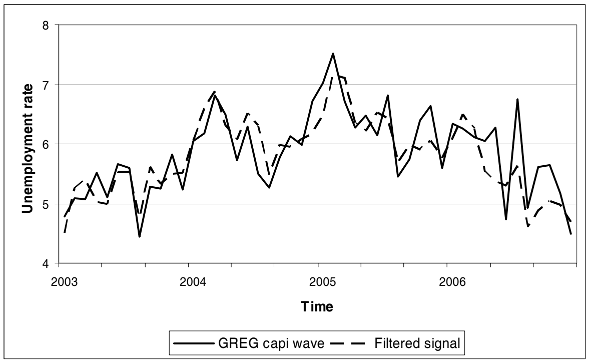
It follows that the seasonal effects in February and August in the third, fourth and fifth wave are significantly different from the first and the second wave. Comparing Figures 5.4 and 5.5 shows that the

RGB in the seasonal effects arises since the seasonal effects in the last three waves are less pronounced than in the first two waves.

### 5.2 Comparison with GREG estimates

The GREG estimates based on the CAPI wave for the monthly unemployment rates are compared with the filtered estimates obtained with the time series model in Figure 5.6. Some of the peaks and dips in the series of the GREG estimates are partially considered as survey errors under the structural time series model and flattened out in the filtered estimates for the series. Some of these peaks and dips are preserved since they are considered as seasonal effects under the time series model. It also follows that the filtered estimates are corrected for the RGB since the filtered series is at the same level as the GREG series for the CAPI wave.

*Figure 5.6: Filtered estimates and GREG estimates CAPI wave for monthly unemployment rate*



The procedure applied in the regular estimation procedure of the LFS, to combine the CATI and the CAPI waves, is also used to estimate monthly unemployment figures. As described in section 2, the four CATI waves are calibrated to the employment status in the CAPI wave. The GREG estimator is used to estimate the monthly unemployment rates using the five different waves. Finally a correction factor based on the preceding 36 months is used to remove the RGB. These corrected GREG estimates based on the data observed in the five waves are compared with the filtered estimates obtained with the time series model in Figure 5.7. The monthly GREG estimates based on all waves are also compared with the GREG estimates based on the CAPI wave in Figure 5.8.

The ratio correction applied to the GREG estimate based on all waves removes the RGB in the low frequency variation between the subsequent waves, but does not correct for the RGB in the seasonal patterns. This follows from figure 5.7 and 5.8. The series of the GREG estimates based on all waves follows the same level as the GREG estimates based on the CAPI wave (Figure 5.8). There are, however, subtle differences between the filtered estimates obtained with the time series model and the GREG estimate based on all waves (Figures 5.7). They are partially the result of systematic differences in the seasonal patterns between the subsequent waves. Moreover they arise because some of the dips and peaks in the GREG estimates are considered as survey errors by the time series model.

*Figure 5.7: Filtered estimates and GREG estimates all waves for monthly unemployment rate*

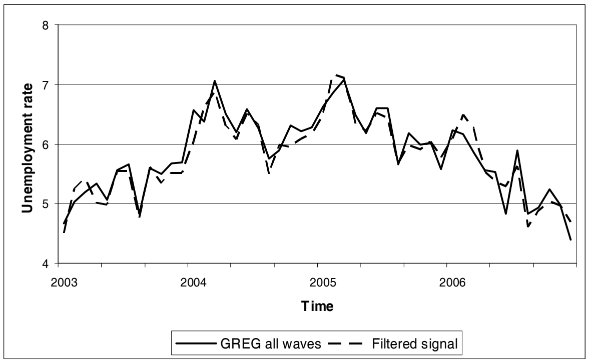
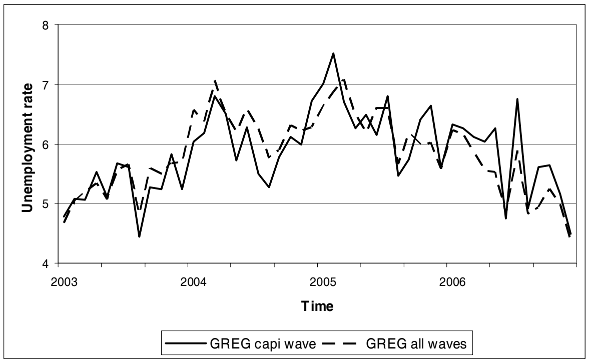


*Figure 5.8: GREG estimates CAPI wave and all waves for monthly unemployment rate*

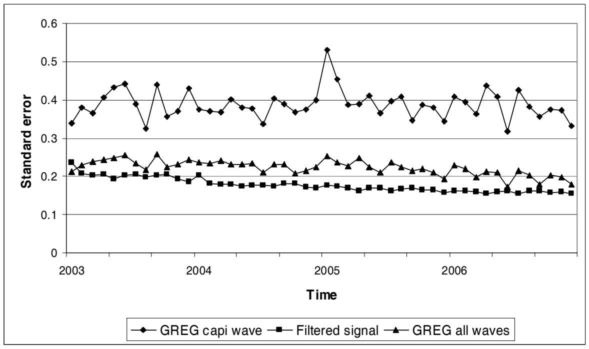


The standard errors for the monthly GREG estimates based on all waves, the CAPI wave and the filtered estimates are compared with each other in Figure 5.9. The standard errors for the GREG estimates are based on the variance of the ratio of two GREG estimators, see e.g. Särndal et al. (1992), formula 7.13.10. See Van den Brakel and Krieg (2007) for technical details of the variance approximation used to account for the calibration of the CATI waves to the CAPI wave and the applied ratio correction described in section 2.

As expected, the standard errors of the GREG estimates based on all waves are smaller than the GREG estimates based on the CAPI wave, since they are based on more data. The standard errors of the filtered estimates obtained with the time series model are smaller than the GREG estimates based on all waves, since the time series model uses additional sample information from preceding periods.

A smaller than expected difference between the standard error of the GREG estimates based on all waves and the filtered time series model estimates was found. This can be explained by the size of the time series model, which is large compared to the length of the series available to fit the model (41 state variables applied to a five dimensional series monthly observed during a period of six years). Smaller standard errors for the filtered estimates might be expected if more data becomes available. Another important aspect is that the GREG estimates are corrected for the RGB in the low frequency variation only. The time series model, on the other hand, accounts for the RGB in the low frequency variation and the seasonal patterns and the standard errors reflect the complexity of the applied model.

*Figure 5.9: Standard errors GREG and filtered estimates monthly unemployment rate*



The efficiency obtained by borrowing sample information from the past by relying on a time series model is illustrated more clearly if the standard error of the GREG estimates using all waves is compared with the standard error of the monthly estimates obtained with a time series model that accounts for the RGB in the low frequency variation only. Therefore a time series model without a component for the RGB in the seasonal pattern is applied to the data in an attempt to improve the precision of the time series model estimates. This implies that $\boldsymbol{\alpha}_t^{\gamma}$, $\mathbf{Z}^{\gamma}$, $\mathbf{T}^{\gamma}$, $\boldsymbol{\eta}_t^{\gamma}$, and $\mathbf{Q}^{\gamma}$ are deleted from the model in state space representation as described in section 4. The filtered estimates for the monthly unemployment rates based on a model with and without a component for the RGB in the seasonal pattern are compared in Figure 5.10.

The model without a component for the RGB of the seasonal effects finds an average seasonal effect for the population parameter $\theta_t$. As a result the absolute values of the seasonal effects in February and August are smaller under the simplified model, resulting in a lower estimate for the monthly unemployment rate in February and a larger estimate in August.

The standard errors for the filtered estimates obtained with the two time series models and the GREG estimator using all waves are compared in Figure 5.11. The standard error of the filtered estimates of the simplified time series model is substantially smaller than the standard error of the GREG estimates using all waves. This is the increase in precision that is obtained by using the sample information from preceding periods through the time series model. The simplification of the time series model by ignoring the RGB for the seasonal effects, results in a reduction of the standard error at the cost of an increased bias in the seasonal effects. Under the model assumption that the estimates for the monthly unemployment rates obtained with the data observed in the first wave are unbiased, the time series model that accounts for the RGB in the seasonal patterns is preferred, since it removes some bias in the seasonal pattern.

*Figures 5.10: Filtered estimates monthly unemployment rate for two different time series models*

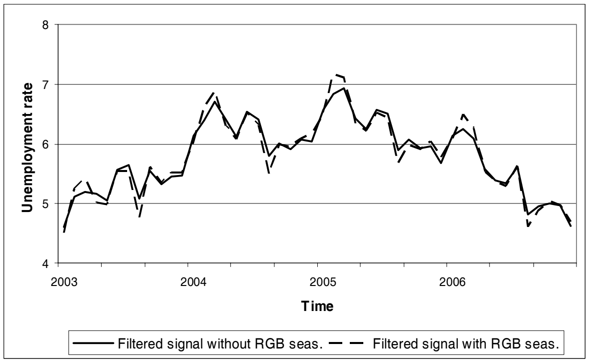
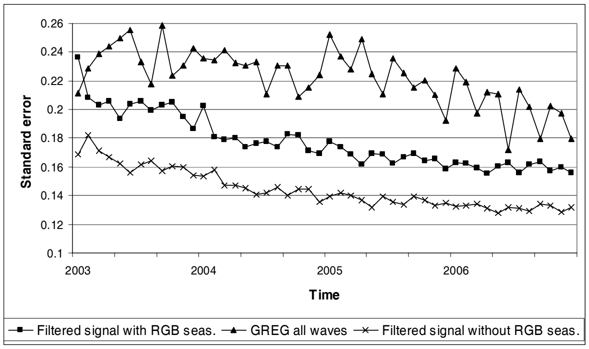


*Figure 5.11: Standard errors monthly estimates unemployment rate GREG for all waves and filtered estimates for two different time series models*



## 6. Discussion and conclusions

In this paper a multivariate structural time series model is applied to the monthly data of the Dutch LFS that accounts for the rotating panel design of this survey. This approach is initially proposed by Pfeffermann (1991) and extended in this paper with a component that models systematic differences in the seasonal effects between the subsequent waves. With this time series model a substantial increase of the accuracy of the monthly estimates for the unemployment rate is obtained. First, the model explicitly estimates the RGB in the low frequency variation and the seasonal patterns between the first CAPI-wave and the four subsequent CATI-waves. As a result, estimates for the unemployment rates are corrected for this RGB. Second, the time series model borrows strength from data observed in preceding periods via the assumed model for the population parameter and the autocorrelation between the survey errors of the different panels.

The RGB induced by the rotating panel design is substantial. The bias in the low frequency variation results in an underestimation of the unemployment rate in the subsequent waves and its magnitude slightly increases from -0.8 percent points in the second wave to -1.1 percent points in the fifth wave. The seasonal effect in February is about 0.5 percent points too small and in August 0.4 percent points too large in the third, fourth and fifth wave compared to the first two waves. This results in less pronounced seasonal effects in the last three waves.

The estimation procedure of the regular LFS is based on the GREG estimator. In this procedure the estimates for the unemployment rate are corrected with the ratio between the unemployment rate based on CAPI only and the estimates based on all waves, using the data of 12 preceding quarters. This ratio corrects for the RGB in the low frequency variation but not for the RGB in the seasonal patterns. Compared with the currently applied estimation procedure, the time series model improves the accuracy of the estimates of the unemployment rate, since it reduces the standard error and gives, under the assumption that the data obtained in the first wave are not biased, better corrections for the RGB.

The time series model is identified by adopting a restriction for the RGB parameters which assumes that the first wave is observed without bias. This implies that the estimates based on the first wave are used to benchmark the subsequent waves. If this restriction is used, then an all out effort in each part of the statistical process is required to reduce possible bias in the first wave, e.g. by using the most appropriate mode, reducing non response, optimizing the weighting scheme, etc. Based on external information about the bias in the different waves, the restriction for the rotation group bias components might be adjusted.

The time series approach explored in this paper is appropriate to produce model-based estimates for monthly unemployment figures. Statistics Netherlands, however, is generally rather reserved in the application of model-based estimation procedures for the production of official statistics. There is, on the other hand, a case for having official time series that are based on model-based procedures with appropriate methodology and quality descriptions for situations where direct estimators do not result in sufficiently reliable estimates. For example under rotating panel designs where measurement errors result in severely biased estimates or in the case of small domains or short data collection periods, where small sample sizes result in large standard errors for direct estimators.

## Acknowledgement

The authors whish to thank professor Pfeffermann for his advice during this research. The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands.

## References

Bailar, B.A. (1975). The Effects of Rotation Group Bias on Estimates from Panel Surveys. *Journal of the American Statistical Association*, 70, pp. 23-30.

Brakel, J.A. van den (2005). Small Area Estimators for the Dutch Labour Force Survey using Structural Time Series Models. Unpublished research paper, BPA nr: TMO-R&D-2005-05-02-JBRL, Statistics Netherlands, Heerlen.

Brakel, J.A. van den and S. Krieg (2007). Modelling Rotation Group Bias and Survey Errors in the Dutch Labour Force Survey. Unpublished research paper, BPA nr: DMH-R&D-2007-01-25-JBRL, Statistics Netherlands, Heerlen.

Doornik, J.A. (1998). *Object-Oriented Matrix Programming using Ox 2.0.* London: Timberlake Consultants Press.

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

Harvey, A.C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter.* Cambridge University Press, Cambridge.

Koopman, S.J., N. Shephard and J.A. Doornik (1999). Statistical Algorithms for Models in State Space using SsfPack 2.2. *Econometrics Journal*, 2, pp. 113-166.

Pfeffermann, D. (1991). Estimation and Seasonal Adjustment of Population Means Using Data from Repeated Surveys. *Journal of Business & Economic Statistics*, 9, pp. 163-175.

Pfeffermann, D., M. Feder and D. Signorelli (1998). Estimation of Autocorrelations of Survey Errors with Application to Trend Estimation in Small Areas. *Journal of Business & Economic Statistics*, 16, pp. 339-348.

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

Scott, A.J. and T.M.F. Smith (1974). Analysis of Repeated Surveys using Time Series Methods. *Journal of the American Statistical Association*, 69, pp. 674-678.