Structural time series modelling of the monthly unemployment rate in a rotating panel

Jan van den Brakel and Sabine Krieg

Discussion paper (09031)

Statistics Netherlands

The Hague/Heerlen, 2009

Explanation of symbols

.	= data not available
*	= provisional figure
x	= publication prohibited (confidential figure)
–	= nil or less than half of unit concerned
–	= (between two figures) inclusive
0 (0,0)	= less than half of unit concerned
blank	= not applicable
2007–2008	= 2007 to 2008 inclusive
2007/2008	= average of 2007 up to and including 2008
2007/'08	= crop year, financial year, school year etc. beginning in 2007 and ending in 2008
2005/'06–2007/'08	= crop year, financial year, etc. 2005/'06 to 2007/'08 inclusive

Due to rounding, some totals may not correspond with the sum of the separate figures.

Publisher
Statistics Netherlands
Henri Faasdreef 312
2492 JP The Hague

Prepress
Statistics Netherlands - Grafimedia

Cover
TelDesign, Rotterdam

Information
Telephone +31 88 570 70 70
Telefax +31 70 337 59 94
Via contact form: www.cbs.nl/information

Where to order
E-mail: verkoop@cbs.nl
Telefax +31 45 570 62 68

Internet
www.cbs.nl

ISSN: 1572-0314

6008309000 X-10

Structural time series modelling of the monthly unemployment rate in a rotating panel

Jan van den Brakel and Sabine Krieg

Summary: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.

Keywords: Small Area Estimation, Rotation Group Bias, Survey Errors.

1. Introduction

The Dutch Labour Force Survey (LFS) is based on a rotating panel design. Each month a sample of addresses is drawn and data are collected by means of computer assisted personal interviewing of the residing households. The sampled households are re-interviewed by telephone four times at quarterly intervals. The estimation procedure of this survey is based on the generalized regression (GREG) estimator, developed by Särndal e.a. (1992).

Due to the following properties, GREG estimators are very attractive to produce official releases in a regular production environment and are therefore widely applied by national statistical institutes. First, GREG estimators are approximately design-unbiased, which provides a form of robustness in the case of large sample sizes. These estimators are derived from a linear regression model that specifies the relationship between the values of a certain target parameter and a set of auxiliary variables for which the totals in the finite target population are known. If this linear regression model explains the variation of the target variable reasonably well, then this might reduce the design variance as well as the bias due to selective nonresponse, Särndal and Swenson (1987), Bethlehem (1988), and Särndal and Lundström (2005). Model misspecification, on the other hand, might result in an increase of the design variance but the point estimates remain approximately design unbiased. Second, GREG estimators are often used to produce one set of weights for the estimation of all target parameters of a multi-purpose sample survey. This is not only convenient but also enforces consistency between the marginal totals of different publication tables.

There are two major problems with the rotating panel design of the LFS and the way that the GREG estimator is applied in the estimation procedure. First, there are substantial systematic differences between the subsequent waves of the panel due to mode- and panel effects. This is a well-known problem for rotating panel designs, and is in the literature referred to as rotation group bias (RGB), see Bailar (1975). In the LFS, the level of the unemployment rate in the subsequent waves is substantially smaller compared to the first wave. There are also systematic differences between the seasonal effects of the subsequent waves.

A second problem is that the monthly sample size of the LFS is too small to rely on the GREG estimator to produce official statistics about the monthly employment and unemployment. GREG estimators have a relatively large design variance in the case of small sample sizes. Therefore, in the LFS, each month the samples observed in the preceding three months are used to estimate quarterly figures about the labour market situation. The major drawback of this approach is that the real monthly seasonal pattern in the unemployment rate is smoothed out and thus biased. Also structural changes in the unemployment appear delayed in the series of the published figures.

Since the monthly sample size is too small to apply design-based or direct survey estimators, model-based estimation procedures might be used to produce sufficiently reliable statistics. In the case of continuously conducted surveys, a structural time series model can be applied to use information from preceding samples to improve the accuracy of the estimates. This model can be extended to account for the RGB and the autocorrelation (AC) between the different panels of the LFS. This approach makes efficient use of the rotating panel design of the LFS in estimating monthly figures about the labour market, and is originally proposed by Pfeffermann (1991) and Pfeffermann e.a. (1998). These techniques are applied in this paper to estimate the monthly unemployment rate of the LFS. Other references to authors that apply time series models to develop estimates for periodic surveys are Scott and Smith (1974), Scott e.a. (1977), Tam (1987), Binder and Dick (1989, 1990), Bell and Hillmer (1990), Tiller (1992), Rao and Yu (1994), Pfeffermann and Burck (1990), Pfeffermann and Bleuer (1993), Pfeffermann and Tiller (2006), Harvey and Chung (2000), and Feder (2001).

Composite estimators can be considered as an alternative. They are developed under the traditional design-based approach, to use information observed in previous periods from periodic surveys with a rotating panel design, to improve the precision of level and change estimates. Some key references to composite estimators are Hansen e.a. (1953), Rao and Graham (1964), Gurney and Daly (1965), Cantwell (1990), Singh (1996), Gambino e.a. (2001), Singh e.a. (2001) and Fuller and Rao (2001).

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

2. The Dutch Labour Force Survey

2.1 Sample design

The objective of the Dutch LFS is to provide reliable information about the labour market. Each month a sample of addresses is selected from which during the data collection households are identified that can be regarded as the ultimate sampling units. The target population of the LFS consists of the non-institutionalised population aged 15 years and over residing in the Netherlands. The sampling frame is a list of all known occupied addresses in the Netherlands, which is derived from the municipal basic registration of population data. 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. All households residing at an address, up to a maximum of three, are included in the sample (in the Netherlands, there is generally one household per address). 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 October 1999, the LFS changed from a continuous survey to a rotating panel design. In the first wave, data are collected by means of computer assisted personal interviewing (CAPI). For all members of the selected households, demographic variables are observed. For the target variables only persons aged 15 years and over are interviewed. When a household member cannot be contacted, proxy interviewing is allowed by members of the same household. Households, in which one or more of the selected persons do not respond for themselves or in a proxy interview, are treated as nonresponding households. The respondents aged 15 through 64 years are re-interviewed four times at quarterly intervals 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. Proxy interviewing is also allowed during these re-interviews. The monthly gross sample size averaged about 8000 addresses commencing the moment that the LFS changed to a rotating panel design. The monthly sample size gradually declined to about 6500 addresses in 2008. During this period about 65% completely responding households are obtained.

2.2 Rotation group bias

The rotating panel design, described in Section 2.1, results in systematic differences between the estimates of the unemployment rate of the successive waves in one time period. In the literature, this phenomenon is known as RGB, see e.g. Bailar (1975), Kumar and Lee (1983) and Pfeffermann (1991). The RGB in the LFS results in a systematic underestimation of the level of the unemployment rate in the CATI waves but also in systematic differences between the seasonal patterns. The RGB is a consequence of the following strongly confounded factors:

- Selective nonresponse between the subsequent waves, i.e. panel attrition.
- Systematic differences between the populations that are reached with the CAPI and CATI modes. It is anticipated that these differences are relatively small, since telephone numbers are asked during the first interview. As a result, secret numbers and cell-phone numbers are also called.
- Mode-effects, i.e. systematic differences in the data due to the fact that the interviews are conducted by telephone instead of face to face. Under the CAPI mode the interview speed is lower, respondents are more engaged with the interview and are more likely to exert the required cognitive effort to answer questions carefully. Also less socially desirable answers are obtained under the CAPI mode due to the personal contact with the interviewer. As a result, less measurement errors are expected under the CAPI mode (Holbrook e.a., 2003, and Roberts, 2007). Van den Brakel (2008) describes an experiment where the CAPI and CATI data collection modes are compared in the first wave of the LFS. It follows that the estimated unemployment rate is significantly smaller under the CATI mode.
- The fraction of proxy interviews is larger under the CATI mode (Van den Brakel, 2008). This might result in an increased amount of measurement errors.
- Effects due to differences between the CAPI questionnaire and the CATI questionnaire. The CATI questionnaire is a strongly condensed version of the CAPI questionnaire since the re-interviews focus on changes in the labour market position of the respondents.
- Panel effects, i.e. systematic changes in the behaviour of the respondents in the panel. For example, questions about activities to find a job in the first wave might increase the search activities of the unemployed respondents in the panel. Respondents might also adjust their answers in the subsequent waves systematically, since they learn how to keep the routing through the questionnaire as short as possible.

It is assumed that the estimates based on the first wave are the most reliable, since CAPI generally results in a higher data quality and the first wave does not suffer from the panel effects mentioned above. In order to minimize the effects of the RGB, the second, third, fourth and fifth waves are currently calibrated to the first wave as will be described in Section 2.3.

2.3 Regular estimation procedure

Target parameters about the employment and unemployment are defined as population totals or as ratios of two population totals. The unemployment rate, which is investigated in this paper, is defined as the ratio of the total unemployment and the total labour force. This population parameter is estimated as the ratio of the GREG estimates for the total unemployed labour force and the total labour force.

The monthly sample size of the LFS is too small to publish reliable monthly figures using the GREG estimator. Therefore each month estimates about the employment and unemployment for the preceding three months are published. The rotation scheme, described in Section 2.1, implies that the sample data obtained with the five waves in three successive months are based on unique households obtained from 15 independent monthly samples.

In an attempt to correct for the RGB, a rather laborious weighting procedure is used in the regular estimation procedure. The most important steps are summarized here. First, the inclusion probabilities are derived, which reflect the sampling design described above as well as the different response rates between geographical regions. Subsequently, the inclusion weights of each CATI wave are calibrated with the GREG estimator to the labour force status observed in the first wave. In the next step, the calibrated weights of the CATI waves and the inclusion weights of the CAPI wave are used as the design or starting weights of the GREG estimator, using a weighting scheme that is based on a combination of different socio-demographic classifications. The integrated method for weighting persons and families of Lemaître and Dufour (1987) is applied to obtain equal weights for persons belonging to the same household. Finally, a bounding algorithm proposed by Huang and Fuller (1978) is applied to avoid negative weights. This estimation procedure is conducted with the software package Bascula (Nieuwenbroek and Boonstra, 2002).

Since this weighting procedure hardly corrects for the RGB, an additional rigid correction is applied. For the most important parameters the ratio between the estimates based on CAPI only and the estimates based on all waves is computed using the data of 12 preceding quarters. Estimates for the preceding three months are multiplied by this ratio to correct for RGB.

2.4 Monthly GREG estimates based on monthly data

In Section 3, a structural time series model is developed to estimate the monthly unemployment rate. The input data for this time series model are the GREG estimates for the monthly unemployment rate using the monthly sample data of the separate waves. Let θ_t denote the true but unknown unemployment rate for month *t*. Now Y_t^{t-j} denotes the GREG estimate of the unemployment rate of month *t*, based on the sample which entered the panel in month *t-j*. For the period of January 2001 until December 2008 each month five independent GREG estimates for the same parameter θ_t are produced, using the five separate waves that are observed each month, i.e. Y_t^{t-j} for *j* = 0, 3, 6, 9, 12. These estimates are defined as

$$Y_t^{t-j} = \frac{t_{y,t}^{t-j}}{t_{z,t}^{t-j}}, \tag{2.1}$$

with $t_{y,t}^{t-j}$ and $t_{z,t}^{t-j}$ the GREG estimates for the unemployed labour force and the labour force at time *t*, based on the sample that entered the panel at *t-j*.

The separate monthly waves are weighted with a reduced version of the weighting scheme that is applied in the regular weighting procedure for the quarterly figures. The estimates based on the CATI data are not adjusted to correct for RGB, since a multivariate time series model is applied to correct for this bias.

The variance of (2.1) can be estimated with

$$\mathrm{var}(Y_t^{t-j}) = \frac{1}{(t_{z,t}^{t-j})^2} \sum_{h=1}^{H} \frac{n_{h,t}^{t-j}}{n_{h,t}^{t-j} - 1} \left(\sum_{k=1}^{n_{h,t}^{t-j}} (w_k e_{k,t}^{t-j})^2 - \frac{1}{n_{h,t}^{t-j}} \left(\sum_{k=1}^{n_{h,t}^{t-j}} w_k e_{k,t}^{t-j} \right)^2 \right), \tag{2.2}$$

with $e_{k,t}^{t-j} = \sum_{l=1}^{m_k} (y_{kl,t}^{t-j} - \mathbf{x}_{kl}^T \mathbf{b}_y) - Y_t^{t-j} (z_{kl,t}^{t-j} - \mathbf{x}_{kl}^T \mathbf{b}_z)$.

Here $y_{kl,t}^{t-j}$ is a binary variable taking value one if the *l*-th person belonging to the *k*-th household that entered the sample at time *t-j* belongs to the unemployed labour force at time *t* and zero otherwise, $z_{kl,t}^{t-j}$ a binary variable taking value one if the *l*-th person of the *k*-th household belongs to the labour force at time *t* and zero otherwise, $\mathbf{x}_{kl}$ a vector with the auxiliary information of the *l*-th person belonging to the *k*-th household used in the weighting scheme of the GREG estimator, $\mathbf{b}_y$ and $\mathbf{b}_z$ the regression coefficient of the regression function of $y_{kl,t}^{t-j}$ respectively $z_{kl,t}^{t-j}$ on $\mathbf{x}_{kl}$, w_k the regression weight of household *k*, $n_{h,t}^{t-j}$ the number of completely responding households of stratum *h=1,…H*, at time *t* of the sample that entered the panel at *t-j*, and m_k the number of persons aged 15 years and over belonging to the *k*-th household. Recall from Section 2.3 that persons belonging to the same household have equal weights due to the application of the integrated method for weighting persons and families of Lemaître and Dufour (1987). Formula (2.2) is the variance estimation procedure implemented in Bascula to approximate the variance of the ratio of two GREG estimators.

The estimates for the monthly unemployment rate obtained with the structural time series approach will be compared in Section 5.3 with monthly estimates based on the GREG estimator using the data observed in the five waves. For this comparison a slightly simplified version of the procedure described in Section 2.3 is applied to combine the data observed in the different waves to obtain monthly GREG estimates. First, a GREG estimate Y_t is computed using the data observed in the five waves using the same weighting procedure used in the regular production process to estimate quarterly figures, see Section 2.3. The weighting scheme is slightly simplified because less data are available. Subsequently a correction factor based on the preceding three years is computed as:

$$c_t = \frac{\sum_{j=0}^{35} Y_{t-j}^{t-j}}{\sum_{j=0}^{35} Y_{t-j}} . \tag{2.3}$$

Finally, the corrected estimate is computed:

$$Y_t^c = c_t Y_t \ . \tag{2.4}$$

Because the series start at January 2001, c_t can be computed from December 2003. To get a corrected GREG estimate for all months, $c_{December\,2003}$ is used in formula (2.4) for the periods preceding December 2003. The variance of (2.4) is approximated by $\mathrm{var}(Y_t^c) = c_t^2\ \mathrm{var}(Y_t),$ where $\mathrm{var}(Y_t),$ is computed with formula (2.2), using the data of all waves accordingly.

3. Time series model

Direct estimators, like the Horvitz-Thompson estimator or the GREG estimator, assume that the monthly unemployment rate q_t is a fixed but unknown population parameter. Under this design-based approach, an estimator for q_t for cross-sectional surveys only 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, but not in the case of repeatedly conducted cross-sectional designs. Scott and Smith (1974) proposed to consider the population parameter q_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 q_t, even in the case of non-overlapping sample surveys.

Recall from Section 2.4 that Y_t^{t-j} denotes the GREG estimator for q_t based on the panel observed at time *t*, which entered the survey for the first time at *t-j*. Due to the applied rotation pattern, each month a vector $\mathbf{Y}_t = (Y_t^t\ Y_t^{t-3}\ Y_t^{t-6}\ Y_t^{t-9}\ Y_t^{t-12})^T$ is observed. According to Pfeffermann (1991), this vector can be modelled as

$$\mathbf{Y}_t = \mathbf{1}_5 q_t + \boldsymbol{\lambda}_t + \boldsymbol{\gamma}_t + \mathbf{e}_t , \tag{3.1}$$

with $\mathbf{1}_5$ a five dimensional vector with each element equal to one, $\boldsymbol{\lambda}_t = (/_t^0\ /_t^3\ /_t^6\ /_t^9\ /_t^{12})^T$ and $\boldsymbol{\gamma}_t = (g_t^0\ g_t^3\ g_t^6\ g_t^9\ g_t^{12})^T$ vectors with time dependent components that account for the RGB in the trend and the RGB in 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. Time series models for the different components in (3.1), i.e. the population parameter q_t, the RGB for the trend $\boldsymbol{\lambda}_t$, the RGB for the seasonal patterns $\boldsymbol{\gamma}_t$, and the survey errors $\mathbf{e}_t$, are developed in Sections 3.1 through 3.3.

3.1 Time series model for the population parameter

With a structural time series model, the population parameter θ_t in (3.1) can be decomposed in a trend component, a seasonal component, and an irregular component, i.e.:

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

where L_t denotes a stochastic trend component, S_t a stochastic seasonal component, and ε_t the irregular component. For the stochastic trend component the so-called local linear trend model is used, which is defined by the following set of equations:

$$\begin{aligned}
&L_t = L_{t-1} + R_{t-1} + \eta_{L,t},\\
&R_t = R_{t-1} + \eta_{R,t},\\
&E(\eta_{L,t}) = 0, \quad Cov(\eta_{L,t}, \eta_{L,t'}) = \begin{cases} \sigma_L^2 & \text{if} \quad t = t' \\ 0 & \text{if} \quad t \neq t' \end{cases}\\
&E(\eta_{R,t}) = 0, \quad Cov(\eta_{R,t}, \eta_{R,t'}) = \begin{cases} \sigma_R^2 & \text{if} \quad t = t' \\ 0 & \text{if} \quad t \neq t'. \end{cases}
\end{aligned} \tag{3.3}$$

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

$$S_t = \sum_{l=1}^{6} S_{l,t}, \tag{3.4}$$

where

$$\begin{aligned}
&S_{l,t} = S_{l,t-1}\cos(h_l) + S_{l,t-1}^{*}\sin(h_l) + \omega_{l,t}\\
&S_{l,t}^{*} = S_{l,t-1}^{*}\cos(h_l) - S_{l,t-1}\sin(h_l) + \omega_{l,t}^{*}, \quad l = 1,\ldots,6,\\
&h_l = \frac{\pi l}{6}, \quad l = 1,\ldots,6,\\
&E(\omega_{l,t}) = E(\omega_{l,t}^{*}) = 0,\\
&Cov(\omega_{l,t}, \omega_{l',t'}) = Cov(\omega_{l,t}^{*}, \omega_{l',t'}^{*}) = \begin{cases} \sigma_\omega^2 & \text{if} \quad l = l' \text{ and } t = t' \\ 0 & \text{if} \quad l \neq l' \text{ or } t \neq t' \end{cases},\\
&Cov(\omega_{l,t}, \omega_{l,t}^{*}) = 0 \text{ for all } l \text{ and } t.
\end{aligned} \tag{3.5}$$

Since the trend and seasonal components are modelled as stochastic processes, the values of these model parameters are allowed to change gradually over time. The irregular component ε_t contains the unexplained variation and is modelled as a white noise process:

$$E(\varepsilon_t) = 0, \quad Cov(\varepsilon_t, \varepsilon_{t'}) = \begin{cases} \sigma_\varepsilon^2 & \text{if} \quad t = t' \\ 0 & \text{if} \quad t \neq t'. \end{cases} \tag{3.6}$$

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 in (3.1) 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 an unbiased estimate for θ_t is obtained with the first wave, which is observed by CAPI, i.e. Y_t^t. This implies that the first component of $\boldsymbol{\lambda}_t$ and $\boldsymbol{\gamma}_t$ equals zero. Now $\boldsymbol{\lambda}_t$ measures the systematic differences in the trend of the second, third, fourth and fifth wave with respect to the first wave. The components of $\boldsymbol{\lambda}_t$ are defined as:

$$\lambda_t^0 = 0, \quad \lambda_t^j = \lambda_{t-1}^j + \eta_{\lambda,j,t}, \quad j = 3, 6, 9, 12, \tag{3.7}$$

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

Furthermore $\boldsymbol{\gamma}_t$ measures the systematic differences in the seasonal components with respect to the first wave. This implies that $\gamma_t^0 = 0$. The other components of $\boldsymbol{\gamma}_t$ are defined as trigonometric functions, which are of the form of (3.5). The variance of the disturbances of the seasonal components are assumed to be equal for all waves and is denoted by σ_γ^2.

To borrow information across the panel waves, the RGB for the trend as well as the RGB for the seasonal components are preferably modelled as time invariant components, i.e. $\sigma_\lambda^2 = \sigma_\gamma^2 = 0$. As a kind of model diagnostic, the model initially allows for time dependent components. Preferably the maximum likelihood estimates for σ_λ^2 and σ_γ^2 tend to zero, which is the case in this application. If this is not the case, it might be possible to allow for separate time independent RGB components for different time intervals.

3.3 Time series model for the survey errors

Finally a time series model for the survey errors in (3.1) is developed, which uses the direct estimates for the variance and AC's for the survey errors of the different panels as prior information. From (3.1) it follows that the survey errors for the first wave are defined as $e_t^t = Y_t^t - \theta_t$. For the second, third, fourth and fifth wave, they are defined as $e_t^{t-j} = Y_t^{t-j} - \theta_t - \lambda_t^j - \gamma_t^j$, for j = 3, 6, 9, 12.

Direct estimates for the variances of the survey errors for the separate panels are obtained with (2.2). These estimates are smoothed by modelling the variance estimates for the separate panels with a linear regression model $Var(Y_t^{t-j}) = b_0^j + b_1^j (Y_t^{t-j} / n_t^{t-j}) + error$, where n_t^{t-j} denotes the sample size at time t of the sample that entered the panel at t-j. Since the variance of a fraction is modelled, it is considered to add $(Y_t^{t-j})^2 / n_t^{t-j}$ as an explanatory variable. The unemployment rate, however, is always smaller than 10 percent. Therefore, this quadratic term is negligible and does not improve the model.

The rotating panel design implies sample overlap with panels observed in the past. The sample of the first wave enters the panel for the first time at time *t,* so there is no sample overlap with panels observed in the past. Consequently, the survey errors of the first wave, e_t^t, are not correlated with survey errors in the past. The survey error of the second wave, i.e. e_t^{t-3}, is correlated with the survey error of the first wave that entered the panel three months earlier, i.e. e_{t-3}^{t-3}. In a similar way, the survey error of the third wave, i.e. e_t^{t-6}, is correlated with e_{t-3}^{t-6} and e_{t-6}^{t-6}. The survey error of the fourth wave, i.e. e_t^{t-9}, is correlated with e_{t-3}^{t-9}, e_{t-6}^{t-9} and e_{t-9}^{t-9}. Finally, the survey error of the fifth wave, i.e. e_t^{t-12}, is correlated with e_{t-3}^{t-12}, e_{t-6}^{t-12}, e_{t-9}^{t-12} and e_{t-12}^{t-12}.

The AC's between the survey errors of the subsequent waves are estimated using the approach proposed by Pfeffermann et al. (1998). Since the real survey errors cannot be observed directly, this approach starts with calculating the autocovariances for the pseudo survey errors, which are defined as $(Y_t^{t-j} - \overline{Y}_t)$, where $\overline{Y}_t$ denotes the average of the five panel estimates Y_t^{t-j} at time *t*. The autocovariances of the pseudo survey errors for a separate wave are influenced by the autocovariances of the real survey errors of the other waves, since the pseudo survey errors are defined as the deviation of a panel estimate with the average of all panel estimates obtained at time *t*. Equation (4) of Pfeffermann et al. (1998) specifies the relation between the autocovariances of the pseudo survey errors and the real survey errors. From this equation, it follows that the autocovariances of the real survey errors can be derived from the autocovariances of the pseudo survey errors by $\boldsymbol{\varphi}_k = \mathbf{F}^{-1}\mathbf{C}_k$, with $\mathbf{C}_k$ a vector containing the five autocovariances of the pseudo survey errors at lag *k*, $\boldsymbol{\varphi}_k$ a vector containing the five autocovariances of the survey errors at lag *k*, and $\mathbf{F}$ a $M \times M$ dimensional matrix where the diagonal elements equal $(M - 1/M)^2$ and the off-diagonal elements $(1/M)^2$. Here *M* denotes the number of waves of the panel design (*M*=5 in this application). The AC's and the partial autocorrelations (PAC) of the survey errors of the subsequent waves are given in Table 3.1.

Table 3.1: Correlations and partial autocorrelations for the survey errors of the separate panels.

wave		lag			
		1	2	3	4
1	AC	-0.029	0.264	0.022	0.230
	PAC	-0.029	0.263	0.038	0.175
2	AC	<u>0.291</u>	0.135	0.035	-0.250
	PAC	<u>0.291</u>	0.054	-0.020	-0.287
3	AC	<u>0.240</u>	<u>0.120</u>	0.087	0.219
	PAC	<u>0.240</u>	<u>0.066</u>	0.047	0.194
4	AC	<u>0.442</u>	<u>0.253</u>	<u>0.122</u>	0.156
	PAC	<u>0.442</u>	<u>0.072</u>	<u>-0.016</u>	0.115
5	AC	<u>0.249</u>	<u>0.298</u>	<u>-0.183</u>	<u>0.127</u>
	PAC	<u>0.249</u>	<u>0.252</u>	<u>-0.344</u>	<u>0.218</u>
Mean*	AC	0.306	0.224	-0.030	0.127
	PAC	0.306	0.144	-0.150	0.162

Underlined AC's and PAC's refer to waves with sample overlap
**): Means are based on the waves with sample overlap.*

The standard errors of the estimated AC's equal $1/\sqrt{T}$, where T denotes the number of observations. This implies that correlations with an absolute value larger than 0.21 are significantly different from zero at a 5% significance level. The lags in Table 3.1 refer to three months periods, so lag one equals a time lag of three months, lag two a time lag of six months, etc.

The AC's in Table 3.1, which are based on overlapping samples, are underlined. The AC's for the overlapping samples are positive as might be expected. An exception is the AC at lag three for the fifth wave, which has a negative value. This correlation, however, is not significantly different from zero. The AC's for lag one of the overlapping samples are all significantly different from zero. For lag two, the AC's of the overlapping samples are significantly different from zero for the fourth and the fifth wave, but not for the third wave. The AC's that are based on non-overlapping samples are sometimes unexpectedly large, e.g. lag two and four of the first wave and lag four of the third wave. The AC for lag four of the second wave, on the other hand, has a surprisingly large negative value.

Pfeffermann et al. (1998) also report large positive AC's for lags with non overlapping samples. In their case this can be explained since samples are replaced in small geographical regions. In the Dutch LFS sample replacement takes place at the national level. There is no good explanation why the AC's for the non overlapping samples are sometimes small and sometimes take significant positive as well as negative values. To obtain more stable estimates, the AC's are averaged over the waves which are based on overlapping samples. This implies that the mean AC for lag one is obtained as the average of the AC over the second, third, fourth and fifth wave, etc. The values are reported in the last two rows of Table 3.1. The standard errors of the PAC's of order p+1 and higher for an AR(p) equal $1/\sqrt{T}$, Box and Jenkins (1970). This implies that the PAC's are not significantly different from zero for lags two and higher if an AR(1) model with a correlation coefficient of 0.306 is assumed to capture the AC of the survey errors for the second, third, fourth and fifth wave.

The direct estimates for the variance and covariance structure of the survey errors are combined in the time series model using the following general form of the survey error model $e_t^{t-j} = k_t^{t-j}\tilde{e}_t^{t-j}$ where $k_t^{t-j} = \sqrt{Var(Y_t^{t-j})}$, see Binder and Dick (1990). This allows for non homogeneous variance in the survey errors, that arise e.g. due to the gradually decreasing sample size over the last decade.

Since the first wave is uncorrelated with survey errors obtained in the past, it is assumed that $\tilde{e}_t^t$ is white noise with $E(\tilde{e}_t^t) = 0$ and $Var(\tilde{e}_t^t) = 1$. As a result, the variance of the survey error equals $Var(e_t^t) = (k_t^t)^2$, which is equal to the direct estimate of the variance of the GREG estimate for the first wave. For the second, third, fourth and fifth wave, it is assumed that $\tilde{e}_t^{t-j} = \rho\,\tilde{e}_{t-3}^{t-j} + v_t^{t-j}$, with ρ =0.306, and

$$E(\nu_t^{t-j}) = 0, \quad Cov(\nu_t^{t-j}, \nu_{t'}^{t'-j}) = \begin{cases} \sigma_\nu^2 & \text{if} \quad t = t' \\ 0 & \text{if} \quad t \neq t'. \end{cases}$$

Since $\tilde{e}_t^{t-j}$ is an AR(1) process, $Var(\tilde{e}_t^{t-j}) = \sigma_\nu^2/(1-\rho^2)$. To enforce that $Var(e_t^{t-j})$ equals the direct estimate for the variance of the GREG estimate, it follows that $\sigma_\nu^2 = (1-\rho^2)$.

3.4 Final time series model for the monthly unemployment rate

The time series model for the vector with GREG estimates $\mathbf{Y}_t$ is obtained by inserting the different components developed in Sections 3.1 through 3.3 into (3.1). This model uses the five monthly GREG estimates as input data to obtain model-based estimates for the monthly unemployment rate. The component for the population parameter θ_t in (3.2), developed in Section 3.1, takes advantages of sample information observed in the past to improve the precision of the estimated monthly unemployment rate. The components for the RGB, developed in Section 3.2, account for the systematic differences between the five monthly GREG estimates to avoid that the estimated monthly unemployment rate is incurred with this bias. The component for the survey errors, developed in Section 3.3, accounts for the AC between the five GREG estimates that are based on the same sample, observed with quarterly intervals.

In summary, a time series model is developed to estimate monthly unemployment rates, that makes optimal use of the available sample information from preceding periods to improve the GREG estimates for the monthly unemployment rate. Furthermore, the model accounts for the rotating panel design of the LFS. Although this approach is model-based, it accounts for the complexity of the survey design of the LFS, since the GREG estimates are used as input data.

4. State space representation

The time series model for the five monthly GREG estimates developed in Section 3 can be analysed with the Kalman filter. Therefore it has to be expressed in the so-called 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, which is sometimes also called the signal equation, specifies how the observations depend on a linear combination of unobserved state variables, e.g. trend, seasonal, explanatory variables, RGB and the survey errors and can be expressed as

$$\mathbf{Y}_t = \mathbf{Z}_t \boldsymbol{\alpha}_t + \mathbf{1}_5 e_t. \tag{4.1}$$

Here $\boldsymbol{\alpha}_t$ denotes the state vector with unobservable state variables, $\mathbf{Z}_t$ a known design matrix that specifies the linear relationship between the observations and the elements of the state vector, and $\mathbf{1}_5$ a five dimensional vector with each element equal to one. The transition equation, which is sometimes also referred to as the system 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

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

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

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

$$\mathbf{Y}_t = (Y_t^t \;\; Y_t^{t-3} \;\; Y_t^{t-6} \;\; Y_t^{t-9} \;\; Y_t^{t-12})^T ,$$

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

$$\boldsymbol{\alpha}_t^{\theta} = (L_t \;\; R_t \;\; S_{1t} \; S_{1t}^* \ldots S_{5t} \; S_{5t}^* \;\; S_{6t}) , \quad \boldsymbol{\alpha}_t^{\lambda} = (\lambda_t^3 \;\; \lambda_t^6 \;\; \lambda_t^9 \;\; \lambda_t^{12} \;\;) ,$$

$$\boldsymbol{\alpha}_t^{\gamma} = (\boldsymbol{\gamma}_t^3 \;\; \boldsymbol{\gamma}_t^6 \, \boldsymbol{\gamma}_t^9 \, \boldsymbol{\gamma}_t^{12}) , \; \boldsymbol{\gamma}_t^j = (\gamma_{1t}^j \;\; \gamma_{1t}^{j*} \ldots \gamma_{5t}^j \; \gamma_{5t}^{j*} \; \gamma_{6t}^j) , \, j\text{=3,6,9,12},$$

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

$$\mathbf{Z}_t = (\mathbf{Z}^{\theta} \;\; \mathbf{Z}^{\lambda} \;\; \mathbf{Z}^{\gamma} \;\; \mathbf{Z}_t^{e} \, \mathbf{O}_{5\times 8}) ,$$

$$\mathbf{Z}^{\theta} = \mathbf{1}_5 \otimes (1\;0\;1\;0\;1\;0\;1\;0\;1\;0\;1\;0\;1) , \qquad \mathbf{Z}^{\lambda} = \begin{pmatrix} \mathbf{0}_4^T \\ \mathbf{I}_4 \end{pmatrix},$$

$$\mathbf{Z}^{\gamma} = \begin{pmatrix} \mathbf{0}_4^T \\ \mathbf{I}_4 \end{pmatrix} \otimes (1\;0\;1\;0\;1\;0\;1\;0\;1\;0\;1) , \;\; \mathbf{Z}_t^{e} = \mathrm{Diag}(k_t^t \; k_t^{t-3} \; k_t^{t-6} \; k_t^{t-9} \; k_t^{t-12}) ,$$

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

$$\mathbf{T}^L = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \;\; \mathbf{T}^S = \mathrm{Blockdiag}(\mathbf{C}_1 \; \ldots \; \mathbf{C}_5 \;\; -1) ,$$

$$\mathbf{C}_l = \begin{pmatrix} \cos(h_l) & \sin(h_l) \\ -\sin(h_l) & \cos(h_l) \end{pmatrix}, \quad h_l = \frac{l\,\pi}{6} , \, l = 1, \ldots, 5,$$

$$\mathbf{T}^{\lambda} = \mathbf{I}_4, \qquad \mathbf{T}^{\gamma} = \mathbf{I}_4 \otimes \mathbf{T}^{S}, \qquad \mathbf{T}^{e} = \begin{pmatrix} \mathbf{0}_4^T & 0 & \mathbf{0}_4^T & \mathbf{0}_4^T \\ \mathbf{O}_{4\times4} & \mathbf{0}_4 & \rho\,\mathbf{I}_4 & \mathbf{O}_{4\times4} \\ \mathbf{O}_{4\times4} & \mathbf{0}_4 & \mathbf{O}_{4\times4} & \mathbf{I}_4 \\ \mathbf{I}_4 & \mathbf{0}_4 & \mathbf{O}_{4\times4} & \mathbf{O}_{4\times4} \end{pmatrix},$$

$$\boldsymbol{\eta}_t = (\boldsymbol{\eta}_t^{\theta}\boldsymbol{\eta}_t^{\lambda}\boldsymbol{\eta}_t^{\gamma}\boldsymbol{\eta}_t^{e})^T, \quad \boldsymbol{\eta}_t^{\theta} = (\eta_{L,t}\ \eta_{R,t}\ \omega_{1,t}\ \omega_{1t}^{*}\ \dots\ \omega_{5,t}\ \omega_{5t}^{*}\ \omega_{6t}),$$

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

$$\boldsymbol{\omega}_t^{j} = (\omega_{1,t}^{j}\ \omega_{1t}^{j*}\ \dots\ \omega_{5t}^{j}\ \omega_{5t}^{j*}\ \omega_{6t}^{j}), \qquad \boldsymbol{\eta}_t^{e} = (\nu_t^{t}\ \nu_t^{t-3}\ \nu_t^{t-6}\ \nu_t^{t-9}\ \nu_t^{t-12}\ \mathbf{0}_8^T),$$

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

$$\mathbf{Q}^{\theta} = \text{Diag}(\sigma_L^2\ \ \sigma_R^2\ \ [\sigma_{\omega}^2\,\mathbf{1}_{11}^T]), \quad \mathbf{Q}^{\lambda} = \sigma_{\lambda}^2\mathbf{I}_4,$$

$$\mathbf{Q}^{\gamma} = \sigma_{\gamma}^2\mathbf{I}_{44}, \qquad \mathbf{Q}^{e} = \text{Diag}\left(1\ \ (1-\rho^2)\mathbf{1}_4^T\ \ \mathbf{0}_8^T\right).$$

Here **1** denotes a vector with each element equal to one and **I** the identity matrix. The subscripts for **0, 1, O,** and **I** specify the dimensions of the vectors and the matrices. As explained in Section 3.3, the standard errors of the survey errors (k_t^{t-j}) and the AC between the survey errors (ρ) are known in advance and are used as prior information in the state-space model.

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, as described in Section 3.3.

The transitional relations for the first five entries of $\boldsymbol{\alpha}_t^{e}$ follow from the assumed white noise process for the survey errors of the first wave and the AR(1) model for the other waves. 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 $t \neq t'$. See Pfeffermann (1991) for details.

Under the assumption of normally distributed error terms, 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 and results in smoothed estimates that are based on the completely observed time series. So the smoothed estimate for the state vector for period *t* also accounts for the information made available after time period *t*. 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.

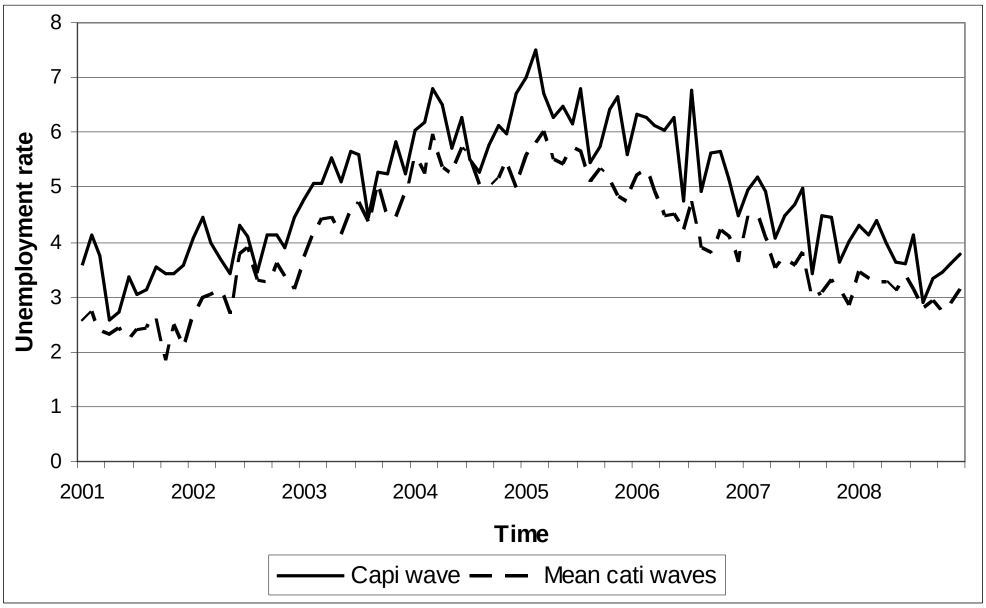
The analysis is conducted with software developed in Ox in combination with the subroutines of SsfPack 3.0, see Doornik (1998) and Koopman e.a. (2008). All state variables are non-stationary with the exception of the survey errors. The non-stationary variables are initialised with a diffuse prior, i.e. the expectation of the initial states are equal to zero and the initial covariance matrix of the states is diagonal with large diagonal elements. The survey errors are stationary and therefore initialised with a proper prior. The initial values for the survey errors are equal to zero and the covariance matrix is available from the model developed for the survey errors in Section 3.3. In Ssfpack 3.0 an exact diffuse log-likelihood function is obtained with the procedure proposed by Koopman (1997).

5. Results

5.1 Preliminary analyses

With the GREG estimator monthly estimates for the unemployment rate are obtained for each wave as described in Section 2.4. In Figure 5.1 the unemployment rate based on the CAPI wave is compared with the average of the four CATI waves. The graph shows that the unemployment rate observed with the first wave is systematically higher than for the other four waves.

Figure 5.1: RGB monthly unemployment rate based on GREG estimates



The five time series obtained with the different waves are modelled with the time series model proposed in Sections 3 and 4. Preliminary analyses indicate that 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

$$\boldsymbol{\alpha}_t^{\gamma} = (\gamma_{1t}\ \gamma_{1t}^{*} \ldots \gamma_{5t}\ \gamma_{5t}^{*}\ \gamma_{6t}), \quad \mathbf{Z}^{\gamma} = (0\ 0\ 1\ 1\ 1)^T \otimes (1\ 0 1\ 0\ 1\ 0\ 1\ 0\ 1\ 0\ 1),$$
$$\mathbf{T}^{\gamma} = \mathbf{T}^{S},\ \boldsymbol{\eta}_t^{\gamma} = (\omega_{1,t}\ \omega_{1t}^{*}\ \ldots\ \omega_{5t}\ \omega_{5t}^{*}\ \omega_{6t}),\ \mathbf{Q}^{\gamma} = \sigma_{\gamma}^2 \mathbf{I}_{11}.$$

5.2 Estimation results for the time series model

Maximum likelihood estimates for the hyperparameters, i.e. the variance components of the stochastic processes for the state variables are obtained using a numerical optimization procedure (BFGS algorithm, Doornik, 1998). To avoid negative variance estimates, the log-transformed variances are estimated. The maximum likelihood estimates for the log-transformed variance of the level of the trend (σ_L^2), the seasonal component (σ_ω^2), the RGB of the trend (σ_γ^2) and the RGB of the seasonals tend (σ_λ^2) to large negative values with extremely large standard errors. These variance components are therefore put to zero in the final model. The estimation results for the remaining hyperparameters are presented in Table 5.1.

Table 5.1: Maximum likelihood estimates hyperparameters

Hyperparameter		Ln-transformed variance comp.		Variance components		
		Estimate	St. error	Estimate	95% conf. interval	
					Lower b.	Upper b.
Slope	(σ_R^2)	-17.226	0.549	0.182E-3	0.106E-3	0.311E-3
Irregular comp.	(σ_ε^2)	-13.480	0.482	1.183E-3	0.737E-3	1.897E-3

The smoothed Kalman filter estimates for the unemployment rate θ_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 local linear trend model simplified to a smooth trend model since $\sigma_L^2 = 0$. The trend component is time dependent since the maximum likelihood estimate of the hyperparameter for the slope is positive (see Table 5.1). The seasonal component is also time independent, since $\sigma_\omega^2 = 0$. Therefore the estimated seasonal effects obtained with the trigonometric form are exactly the same as the results obtained with the well known dummy variable seasonal model. 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

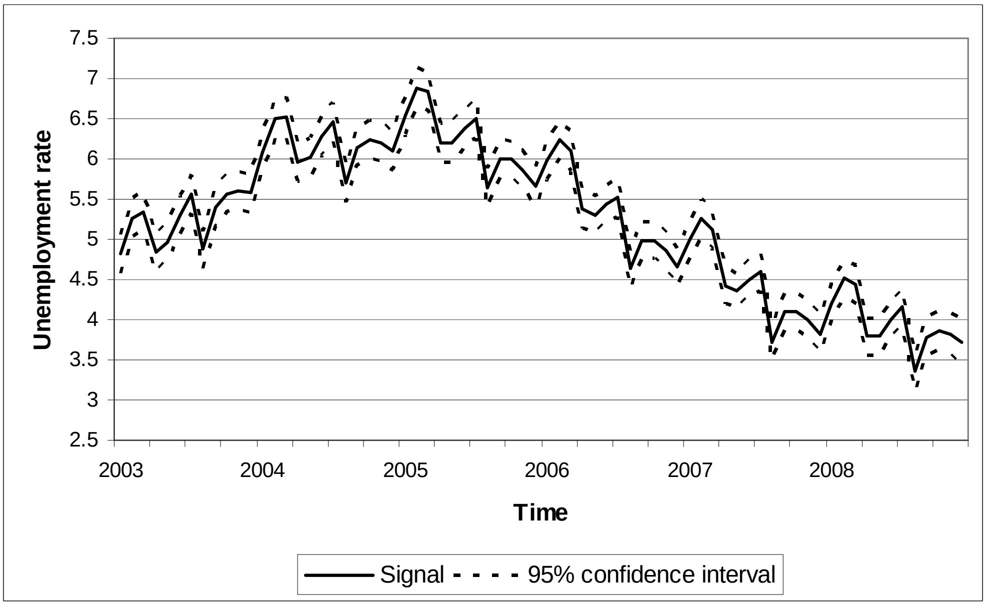


Figure 5.3 Smoothed Kalman filter estimates for the trend of the monthly unemployment rate

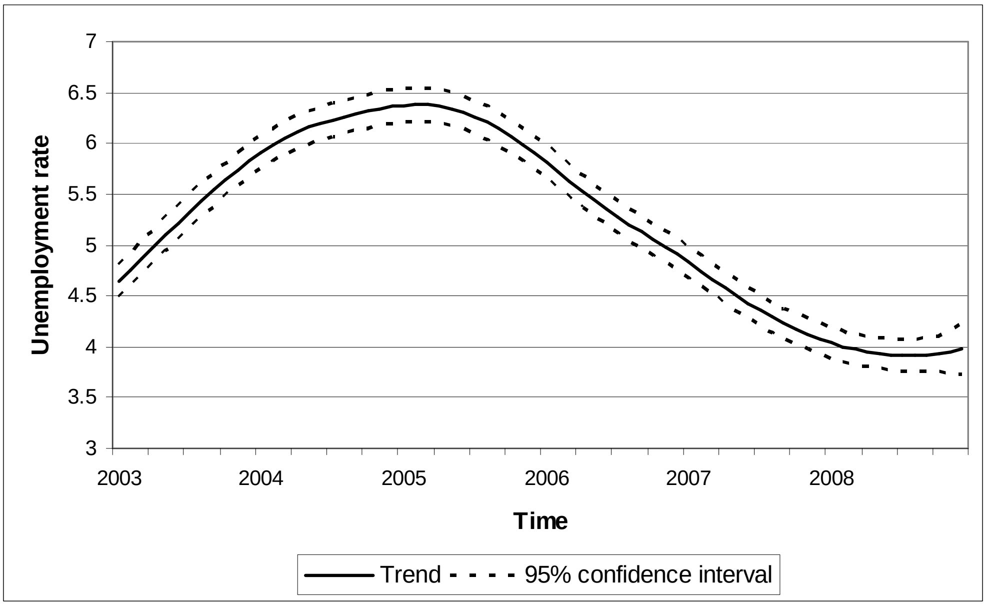
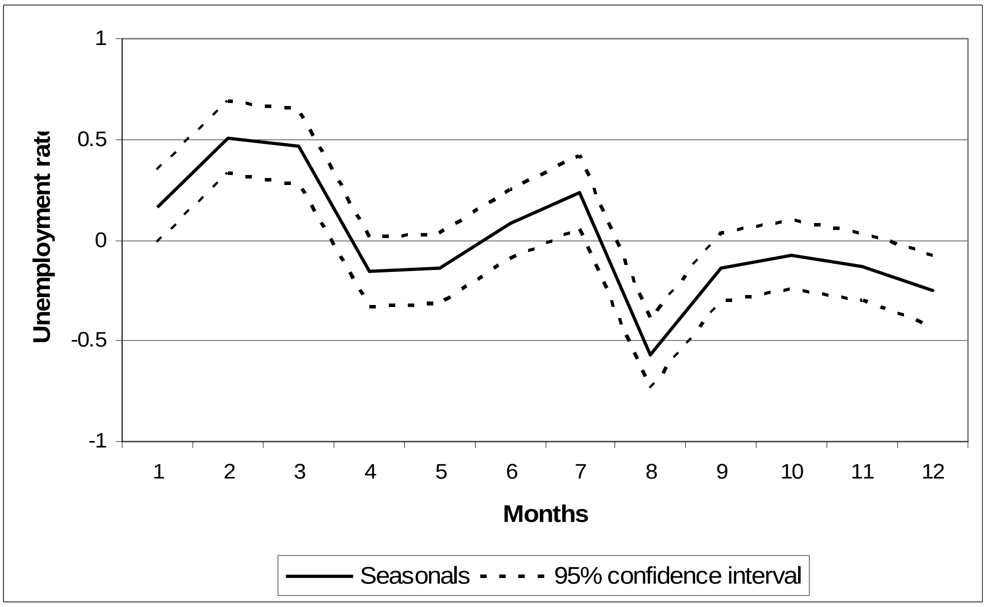


Figure 5.4 Smoothed Kalman filter estimates for the seasonal effect of the monthly unemployment rate



The Kalman filter estimates for the RGB of the trend are time independent. The smoothed Kalman filter estimates for the RGB are given in Table 5.2. The model beautifully detects a slightly increasing bias in the trend 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	-0.75	0.04
3	-0.86	0.04
4	-0.96	0.05
5	-1.10	0.05

An interesting empirical result of this application is the finding of the seasonality in the RGB. The Kalman filter estimates for the RGB of the seasonal effects are also time independent. Therefore, a sequence of likelihood ratio tests is conducted to reach the finally selected model and to test whether the seasonality effects in the RGB of this model are jointly significantly different form zero. Consider the following nested models:

- M1: separate and fixed RGB in the seasonality for wave two, three, four and five
- M2: equal to M1 where the RGB in the seasonality of wave two is equal to zero
- M3: equal to M2 with equal RGB in the seasonality of wave three four and five
- M4: RGB in the seasonality of wave two, three, four and five is equal zero

The results of the likelihood ratio tests of this sequence of models are specified in Table (5.3).

Table 5.3: Likelihood-ratio tests for RGB in seasonality

Model	Log likelihood	Null hypothesis	Likh. ratio stat.	D.f.	p-value
M1	1592.9				
M2	1585.5	M2 = M1	14.7	11	0.19568
M3	1573.7	M3 = M2	23.7	22	0.36422
M4	1559.9	M4 = M3	27.6	11	0.00373

Testing the hypothesis that M2 equals M1 shows that the seasonality of the second wave is not significantly different from the first wave. Testing the hypothesis that M3 equals M2 shows that the RGB in the seasonality of the third, fourth and fifth wave are not significantly different. Testing the hypothesis that M4 equals M3 shows that the RGB of seasonal effects in last three waves are jointly significantly different from zero.

The smoothed Kalman filter estimates for the RGB of the seasonal effects for wave three, four and five are given in Figure 5.5. The smoothed Kalman filter estimates of the seasonal effects are compared with the smoothed estimates for the RGB of the seasonal effects in Figure 5.6.

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

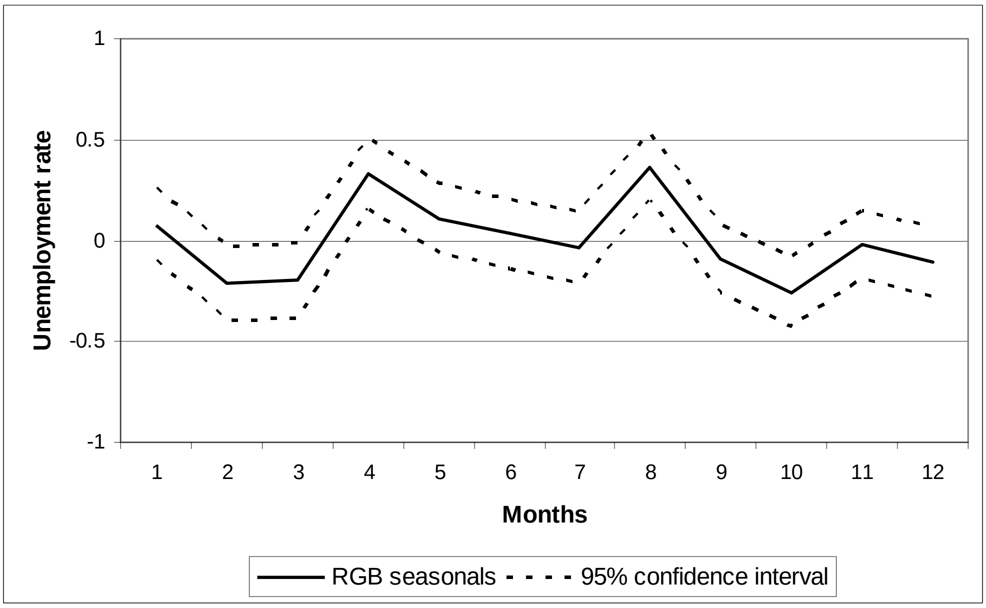
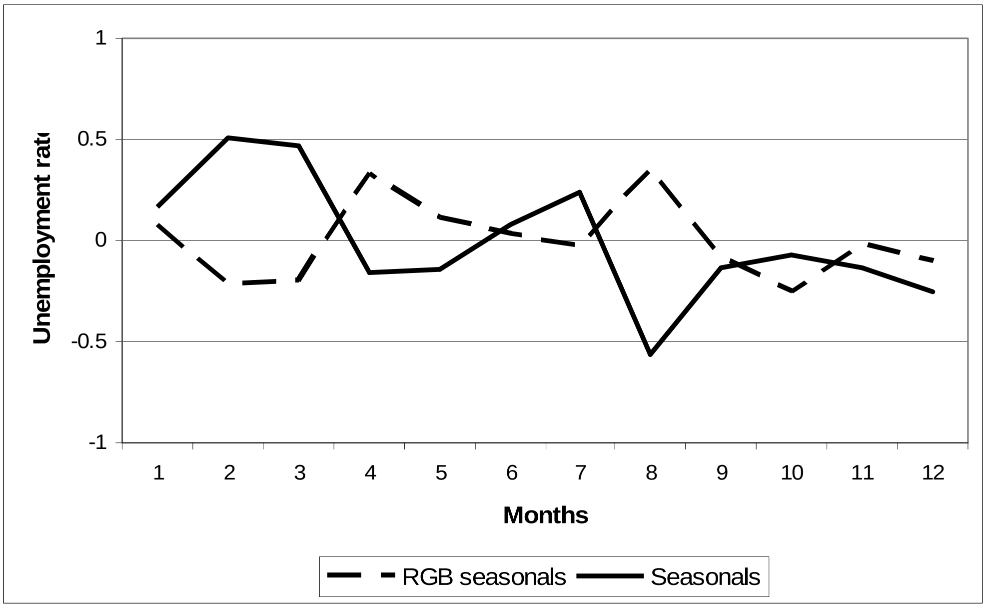


Figure 5.6: Comparison of smoothed Kalman filter estimates for the RGB of the seasonal effects in the third, fourth and fifth wave and the seasonal effects in 2008



It follows from Figure 5.5 that the seasonal effects in particularly February, March, April, August and October in the third, fourth and fifth wave are significantly different from the first and the second wave. Figure 5.6 shows that the RGB in the seasonal effects largely nullifies the seasonal effects in these months. The seasonal effects in the last three waves are, apparently, less pronounced than in the first two waves. The different factors that contribute to the RGB in both the trend and the seasonal patterns are summarised in Section 2.2.

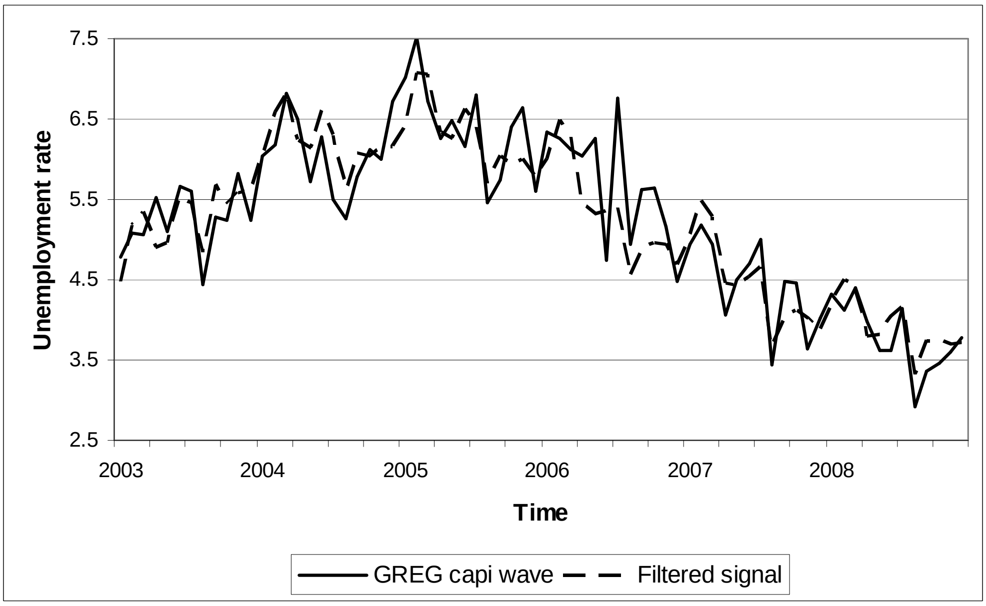
5.3 Comparison with GREG estimates

In this section the monthly GREG estimates for the unemployment rate and their standard errors are compared with the filtered model estimates. The filtered estimates are used since they are based on the complete set of information that would be available in the regular production process to produce a model-based estimate for the monthly unemployment rate for month *t*.

The GREG estimates based on the CAPI wave for the monthly unemployment rates are compared with the filtered model estimates in Figure 5.7. 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 series of the GREG estimates based on the CAPI wave. This is enforced with the assumption that the model parameters for the RGB for the first wave are zero

(Section 3.2). This implies that the CATI waves are benchmarked to the outcomes of the first wave.

Figure 5.7: Filtered estimates and GREG estimates based on the CAPI wave for the 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. The GREG estimates for the monthly unemployment rates based on the five waves, using formula (2.4), are compared with the filtered estimates in Figure 5.8. Both estimates for the monthly unemployment rate follow the same level, since they are both benchmarked to the outcomes of the first wave. The GREG estimator is benchmarked in a rather rigid way using ratio (2.3), which is assumed to be constant in advance over a period of three years. The filtered estimates are benchmarked in a more subtle way through the explicit modelling of the trend and the seasonality in the RGB. The seasonality in the RGB indicates that the assumption of a constant RGB is not tenable. The monthly GREG estimates based on all waves are also compared with the GREG estimates based on the CAPI wave in Figure 5.9.

The ratio correction applied in formula (2.4) to the GREG estimates based on all waves removes the RGB in the trend, but does not correct for the RGB in the seasonal patterns. This follows from Figure 5.8 and 5.9. The series of the GREG estimates based on all waves follows the same level as the GREG estimates based on the CAPI wave (Figure 5.9). There are, however, subtle differences between the filtered estimates and the GREG estimates based on all waves (Figures 5.8). They partially arise because some of the dips and peaks in the GREG estimates are considered as survey errors by the time series model but they are also the result of systematic differences in the seasonal patterns between the subsequent waves. The model estimates in February and March are for example larger in 2003, 2005 and 2006, and smaller in August in 2004, 2005 and 2006.

Figure 5.8: Filtered estimates and GREG estimates based on all waves for the monthly unemployment rate

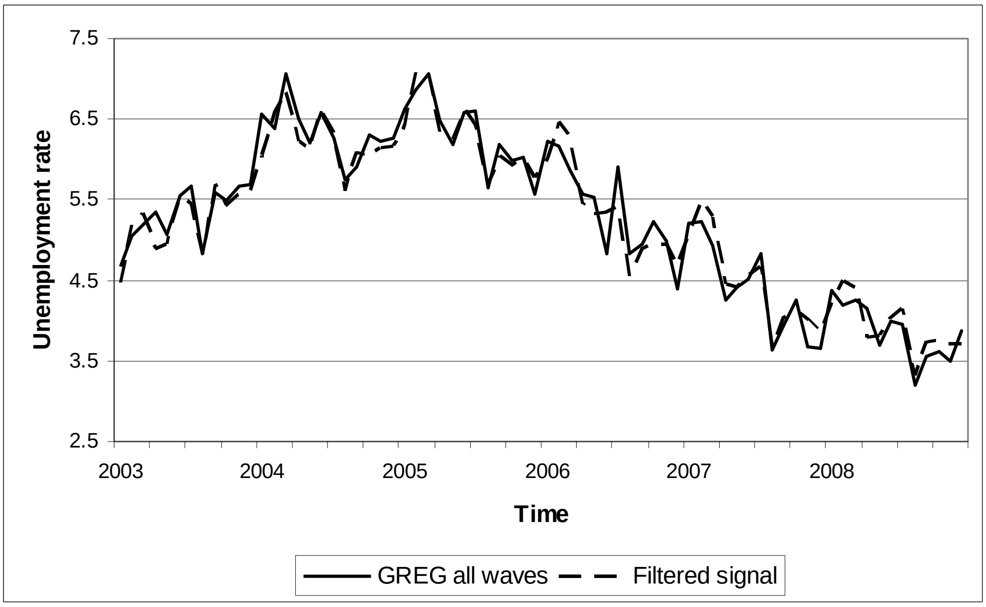
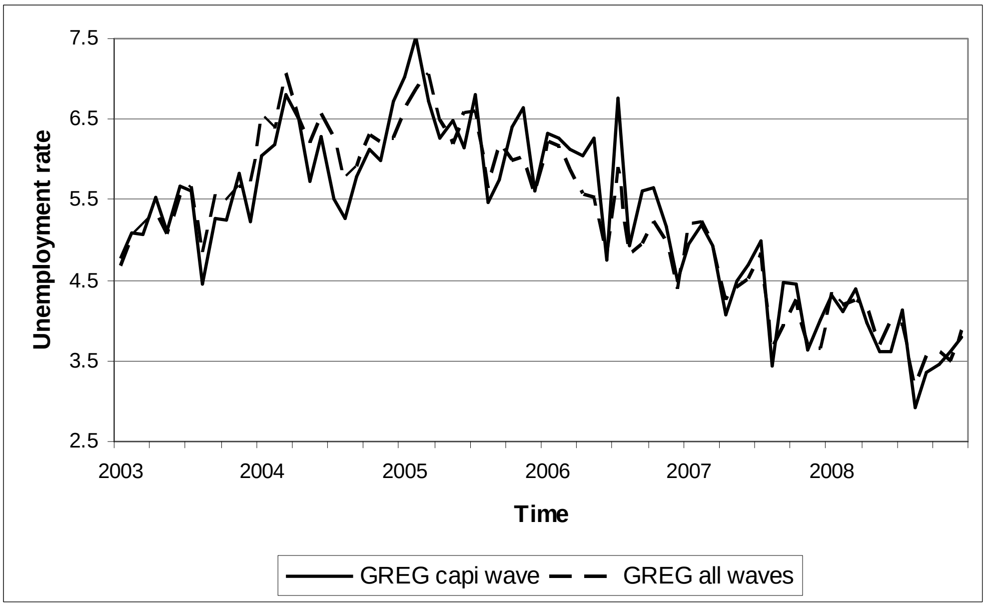
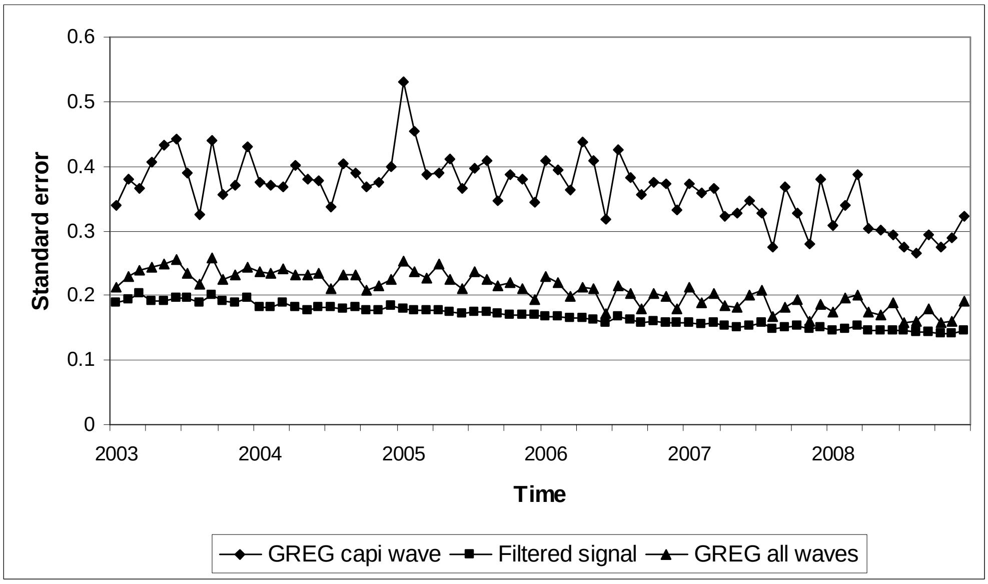


Figure 5.9: GREG estimates based on the CAPI wave and based on all waves for the monthly unemployment rate



The standard errors of the monthly GREG estimates based on all waves, the CAPI wave and the filtered estimates are compared with each other in Figure 5.10. The standard errors for the GREG estimates are computed as described in Section 2.4. Standard errors of the filtered estimates are obtained by the standard recursion formulas of the Kalman filter, see Harvey (1989) or Durbin and Koopman (2001). The Kalman filter recursion assumes that the fitted state space model is the truth. As a result the standard errors for the filtered estimates do not reflect the additional variation induced by the use of likelihood estimates for the variance components in the state space model and are therefore too optimistic.

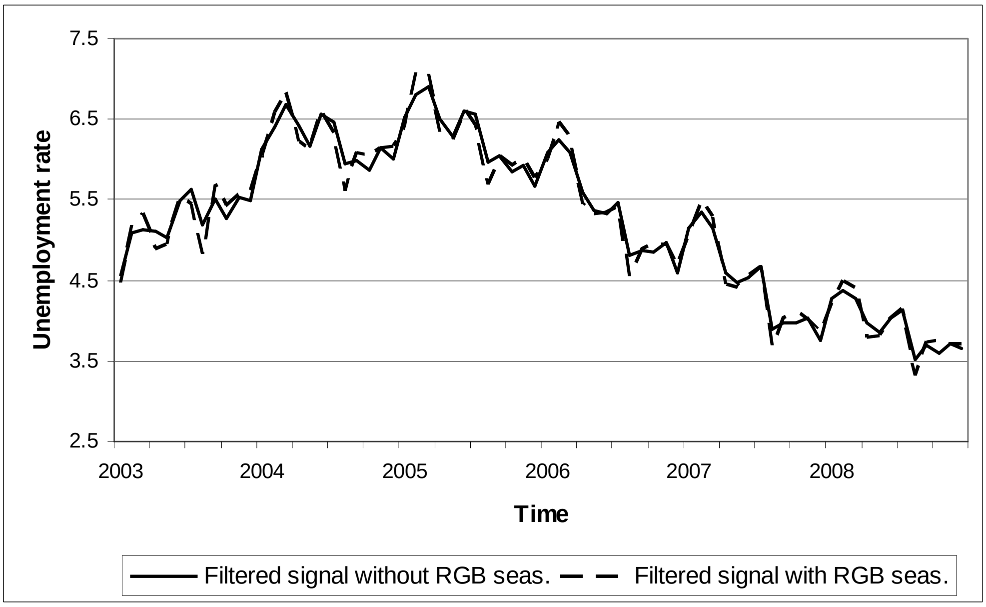
Figure 5.10: Standard errors of the GREG and filtered estimates for the monthly unemployment rate



As expected, the standard errors of the GREG estimates based on all waves are smaller than the standard errors of the GREG estimates based on the CAPI wave, since they are based on more data. The standard errors of the filtered estimates are smaller than the GREG estimates based on all waves, since the time series model uses additional sample information from preceding periods. The standard errors of the filtered estimates are slightly but continuously decreasing during the period 2003 to 2008. This indicates that the filter is picking up new information if new data becomes available. Smaller standard errors for the filtered estimates might be expected if more data become available.

The size and complexity of the applied time series model, is large compared to the length of the series available to fit the model. The final model that is applied to a five dimensional series which is monthly observed during a period of eight years contains 41 state variables. Therefore it is worthwhile to consider more parsimonious models, which might reduce the standard errors of the filtered estimates. Furthermore, the GREG estimate contains a bias since the RGB contains a seasonal effect, which is not reflected by its standard error. Therefore, 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 obtained with a time series model that accounts for the RGB in the trend 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 further improve the precision of the time series model estimates and to illustrate the variance reduction by borrowing strength over time. 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 state-space model. 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 5.11.

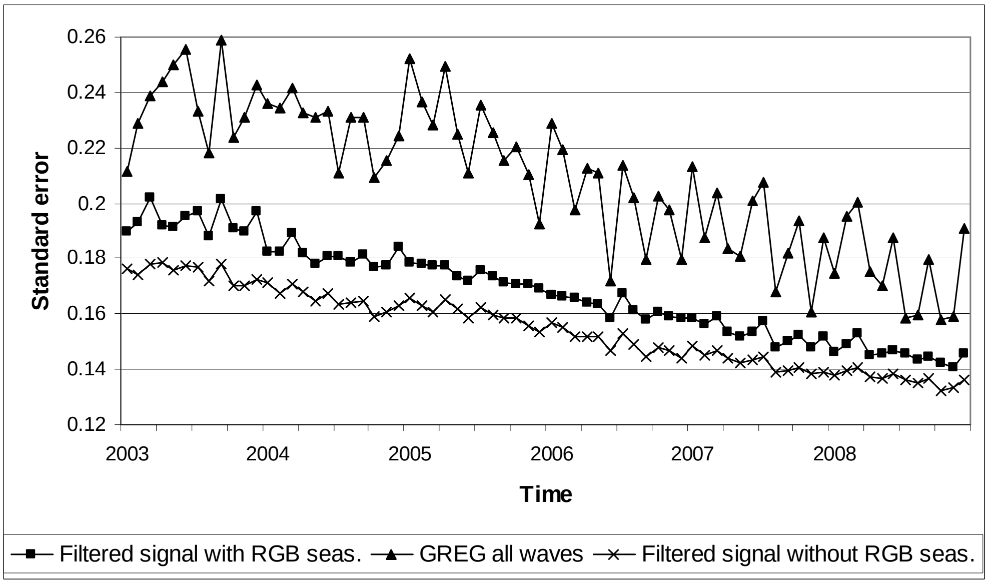
Figures 5.11: Filtered estimates of the monthly unemployment rate for two different time series models



The model without a component for the RGB of the seasonal effects assumes a seasonal effect for the population parameter θ_t that is based on an average of the seasonal effects of the five waves. The absolute values of the seasonal effects in February, March, and August are smaller under the simplified model, resulting in a lower estimate for the monthly unemployment rate in February and March and a larger estimate in August. This results in a more pronounced seasonal pattern in the filtered series obtained with the complete model.

The standard errors of the filtered estimates obtained with the two time series models and the standard errors of the GREG estimates using all waves are compared in Figure 5.12. 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 based on the first wave are unbiased, the time series model that accounts for the RGB in the seasonal effects is preferred, since it removes the bias in the seasonal pattern.

Figure 5.12: Standard errors of the GREG estimates based on all waves and filtered estimates for two different time series models for the monthly unemployment rate



6. Discussion

In this paper a multivariate structural time series model is applied to the monthly data of the 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. Compared with the GREG estimator, which is currently applied in the regular LFS, the time series model results in a substantial increase of the accuracy of the estimates of the unemployment rate. Firstly, the model explicitly estimates the RGB in the trend and the seasonal patterns between the first CAPI wave and the four subsequent CATI waves. As a result, the estimates for the unemployment rates are not incurred with this bias. Secondly, the time series model borrows strength from data observed in preceding periods via the assumed model for the population parameter and the AC between the survey errors of the different panels.

The RGB induced by the rotating panel design is substantial. The bias in the trend results in an underestimation of the unemployment rate in the subsequent waves and its magnitude slightly decreases from -0.8 percent points in the second wave to -1.1 percent points in the fifth wave. The seasonal patterns of the first two waves and the last three waves are also significantly different, since the seasonal pattern in the last three waves is less pronounced.

A parsimonious time series model that accounts for the RGB in the trend but not for the RGB in the seasonal pattern, results in a further reduction of the standard error of the filtered estimates. This, however, results in a biased seasonal pattern in the monthly estimates of the unemployment rates. Since the standard errors of the filtered estimates obtained under this parsimonious model do not reflect this bias, a time series model that accounts for both the RGB in the trend and the seasonal pattern is preferred.

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 data collection mode, reducing nonresponse, optimizing the weighting scheme, etc. Based on external information about the bias in the different waves, the restrictions for the RGB 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. Model misspecification might result in severely biased estimates. This bias is not reflected in the standard errors of the Kalman filter estimates. Extensive model selection and evaluation is therefore required for each separate target variable. This hampers a straightforward application of such estimation techniques, since there is generally limited time available for the analysis phase of the regular production process of official releases.

There is, on the other hand, a case for having official 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. The RGB observed under the rotating panel design of the LFS clearly illustrates the existence of non-sampling errors such as measurement errors and panel attrition. Therefore the traditional concepts that observations obtained from sampling units are true fixed values observed without error and that the respondents can be considered as a representative probability sample from the target population, generally assumed in design-based sampling theory, are not tenable under such designs. The application of direct estimators in the case of measurement errors and selective panel attrition will result in severely biased estimates if no additional corrections are made, like e.g. the ratio correction that is applied in the regular estimation procedure of the Dutch LFS. In the regular estimation procedure a ratio correction is applied to the GREG estimates, which is based on the implicit model assumption that the bias is constant over a period of three years. The time series model applied in this paper can be used to produce estimates that are corrected for the bias introduced by these non-sampling errors in a more advanced way.

This estimation procedure is also applicable in situations where small sample sizes result in unacceptable large standard errors. Small sample sizes arise if official statistics are required for small domains or for short data collection periods like the monthly unemployment figures in the LFS. Most surveys conducted by national statistical institutes operate continuously in time and are based on cross-sectional or rotating panel designs. Consequently, estimation procedures based on time series models that use sample information observed in preceding periods are particularly interesting.

The time series model yields estimates for the trend and seasonal components of the population parameter. Seasonally adjusted parameter estimates and their estimation errors are therefore obtained as a by-product of this estimation procedure. Another major advantage is that this approach accounts for the AC in the survey errors due to the rotating panel design. Pfeffermann, e.a. (1998) show that ignoring these AC, for example with the Henderson filters in X-11-ARIMA (Findley e.a. 1998), results in spurious trend estimates.

The model can be improved in several ways. Information about registered unemployment and related variables, available in the register of the Office for Employment and Income, can be used as auxiliary variables in the models. If longer series become available, an additional cyclic component might be required to capture economic fluctuations. Another possible improvement is detection and modelling of outliers. Furthermore the model needs to be extended to estimate monthly unemployment rates for different domains using sample information collected in the past as well as cross-sectional data from other small areas, using the approach proposed by Pfeffermann and Burck (1990) and Pfeffermann and Tiller (2006).

Acknowledgement

The authors would like to thank Professor D. Pfeffermann and Professor S.J. Koopman for their valuable advice during this project as well as the Associate Editor and the referees for giving constructive comments on earlier drafts of this paper.

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.

Bell, W.R., and S.C. Hillmer (1990). The time series approach to estimation of periodic surveys. *Survey Methodology*, 16, pp. 195-215.

Bethlehem, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4, pp. 251-260.

Binder, D.A., and J.P. Dick (1989). Modeling and estimation for repeated surveys. *Survey Methodology*, 15, pp. 29-45.

Binder, D.A., and J.P. Dick (1990). A method for the analysis of seasonal ARIMA models. *Survey Methodology*, 16, pp. 239-253.

Box, G.E.P., and G.W.M. Jenkins (1970). *Time series analysis - forecasting and control*. San Francisco: Holden-Day.

Brakel, J.A. van den (2008). Design-based analysis of embedded experiments with applications in the Dutch labour force survey. *Journal of the Royal Statistical Society, Series A*, 171, pp. 581-613.

Cantwell, P.J. (1990). Variance formulae for composite estimators in rotating designs. *Survey Methodology*, 16, pp. 153-163.

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.

Feder, M. (2001). Time series analysis of repeated surveys: the state-space approach. *Statistica Neerlandica*, 55, pp. 182-199.

Findley, D.F., B.C. Monsell, W.R. Bell, M.C. Otto, and B.C. Chen (1998). New capabilities and methods of the X-12-ARIMA seasonal adjustment program. *Journal of Business and Economic Statistics*, 16, pp. 127-176 (with Discussion).

Fuller, W.A., and J.N.K. Rao (2001). A regression composite estimator with application to the Canadian labour force survey. *Survey Methodology*, 27, pp. 45-51.

Gambino, J., B. Kennedy, and M.P. Singh (2001). Regression composite estimation for the Canadian labour force survey: evaluation and implementation. *Survey Methodology*, 27, pp. 65-74.

Gurney, M., and J.F. Daly (1965). A multivariate approach to estimation in periodic sample surveys. *Proceedings of the Section on Social Statistics*, American Statistical Association, pp. 242-257.

Hansen, M.H., W.N. Hurwitz, and W.G. Meadow (1953). *Sample survey methods and theory, 2*. New York: Wiley.

Harvey, A.C. (1989). *Forecasting, structural time series models and the Kalman filter.* Cambridge: Cambridge University Press.

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

Holbrook, A.L., M.C. Green, and J.A. Krosnick (2003). Telephone versus face-to-face interviewing of national probability samples with long questionnaires. *Public Opinion Quarterly*, 67, pp. 79-125.

Huang, E.T., and W.A. Fuller (1978). Nonnegative regression estimation for survey data, *Proceedings of the Section on Social Statistics*, American Statistical Association, pp. 300-303.

Koopman, S.J. (1997). Exact initial Kalman filtering and smoothing for non-stationary time series models. *Journal of the American Statistical Association*, 92, pp.1630-1638.

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

Kumar, S., and H. Lee (1983). Evaluation of composite estimation for the Canadian labour force survey. *Survey Methodology*, 9, pp. 178-201.

Lemaître, G., and J. Dufour (1987). An integrated method for weighting persons and families. *Survey Methodology*, 13, pp. 199-207.

Nieuwenbroek, N., and H.J. Boonstra (2002). Bascula 4.0 reference manual, BPA nr: 279-02-TMO, Statistics Netherlands, Heerlen.

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., and S.R. Bleuer (1993). Robust joint modelling of labour force series of small areas. *Survey Methodology*, 19, pp. 149-163.

Pfeffermann, D., and L. Burck (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, 16, pp. 217-237.

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.

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

Rao, J.N.K., and J.E. Graham (1964). Rotating designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 59, pp. 492-509.

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

Roberts, C. (2007). Mixing modes of data collection in surveys: A methodological review. Review paper, NCRM/008, National Centre for Research Methods, City University London.

Särndal, C.-E., and S. Lundström (2005). Estimation in surveys with nonresponse. New York: Wiley.

Särndal, C.E., and B. Swensson (1987). A general view of estimation for two phases of selection with application to two-phase sampling and nonresponse. *International Statistical Review*, 55, pp. 279-294.

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.

Scott, A.J., T.M.F. Smith, and R.G. Jones (1977). The application of time series methods to the analysis of repeated surveys. *International Statistical Review*, 45, pp. 13-28.

Singh, A.C. (1996). Combining information in survey sampling by modified regression. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 120-129.

Singh, A.C., B. Kennedy, and S. Wu (2001). Regression composite estimation for the Canadian labour force survey with a rotating panel design. *Survey Methodology*, 27, pp. 33-44.

Tam, S.M. (1987). Analysis of repeated surveys using a dynamic linear model. *International Statistical Review*, 55, pp. 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, pp. 149-166.