Computational Statistics and Data Analysis 56 (2012) 2918–2933

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Estimation of the monthly unemployment rate for six domains through structural time series modelling with cointegrated trends

Sabine Krieg[a,*], Jan A. van den Brakel[a,b]

[a] Statistics Netherlands, Division of Methodology and Quality, P.O. Box 4481, 6401CZ Heerlen, The Netherlands

[b] Maastricht University, Department of Quantitative Economics, P.O. Box 616, 6200 MD Maastricht, The Netherlands

ARTICLE INFO

Article history:
Received 14 January 2011
Received in revised form 1 February 2012
Accepted 9 February 2012
Available online 20 February 2012

Keywords:
Cointegration
Common factor models
Kalman filter
Small area estimation
State-space models

ABSTRACT

National statistical institutes generally apply design-based techniques like the generalized regression estimator to compile official statistics. These techniques, however, have relatively large design variances in the case of small sample sizes. In such cases, model based small area estimation techniques can be considered to improve the precision of the estimates. A multivariate structural time series model is developed and applied to obtain more precise estimates of the Dutch monthly unemployment rate for six domains. The model takes advantage of sample information from preceding time periods through an appropriate time series model and from other domains by modelling the correlation between the trend components of the time series models for the different domains. The trends of the six domains are cointegrated, which allows the use of a more parsimonious common factor model that is based on three common trends. Although the use of common factor models is a well known approach in econometrics, its application in the context of small area estimation is novel. The standard errors of the direct estimates of the monthly unemployment rates are more than halved with this approach.

1. Introduction

Generalized regression (GREG) estimators (Särndal et al., 1992) are widely applied by national statistical institutes to compile official statistics about a society. This class of estimators is approximately design unbiased and is derived from the probability structure of the sample design that is used to draw a sample from a finite target population and the available auxiliary information about this target population. Statistical modelling only plays a minor role in this design-based inference mode. GREG estimators generally have relatively large design variances in the case of small sample sizes. In these situations, statistical models can be used to improve the precision of these design-based or direct estimates, which is referred to as small area estimation. For a comprehensive overview of small area estimation techniques, see Rao (2003). A briefer but very nice overview is given by Pfeffermann (2002). New developments in small area estimation are reviewed in Pfeffermann (2010).

Small area estimation refers to a class of estimation procedures and techniques that explicitly rely on statistical models to take advantage of sample information that is observed in preceding periods or in other domains. The first approach is called borrowing strength over time and can be achieved by time series modelling. The second approach is called borrowing strength across domains, which is commonly achieved by multilevel modelling. Sometimes this approach is also called borrowing strength over space, since the domains are often geographical areas.

* Corresponding author. Tel.: +31 455706969; fax: +31 455706244.
E-mail addresses: skrg@cbs.nl (S. Krieg), jbrl@cbs.nl (J.A. van den Brakel).

0167-9473/$ – see front matter
doi:10.1016/j.csda.2012.02.008

Most surveys undertaken by national statistical institutes are conducted repeatedly over time and generally require estimates for various subpopulations. It is, therefore, efficient to apply models that borrow strength both across domains and over time. Pfeffermann and Bleuer (1993), Pfeffermann and Burck (1990) and Pfeffermann and Tiller (2006) therefore proposed a multivariate structural time series approach. With a structural time series model, a series is decomposed into several unobserved components, like a trend component, a seasonal component, other cyclical components, a regression component and an irregular component. These time series components are, in the context of small area estimation, used to borrow sampling information from preceding periods to improve the precision of survey estimates. Information across domains is borrowed by modelling correlations between model parameters or by benchmarking the domain estimates to an estimate at a high aggregation level. If the series of the different domains are strongly correlated, than the unobserved components in a multivariate structural time series model can be cointegrated. This results in more parsimonious models since the number of unobserved trends or seasonal components is smaller than the number of series that are included in the model. These so called common factor models are widely applied in econometrics and are applied in the context of small area estimation in this paper.

An alternative approach to borrow strength over time and across domains is developed by Rao and Yu (1994) who extended the multilevel model with random time effects, assuming an AR(1) model. Other key references to authors that applied the time series approach to repeated survey data to improve the efficiency of survey estimates are Blight and Scott (1973), Scott and Smith (1974), Scott et al. (1977), Smith (1978), Tam (1987), Binder and Dick (1989, 1990), Bell and Hillmer (1990), Pfeffermann (1991), Tiller (1992), Abraham and Vijayan (1992), Pfeffermann et al. (1998), Harvey and Chung (2000), Feder (2001), Silva and Smith (2001), Lind (2005), Van den Brakel et al. (2008) and Van den Brakel and Roels (2010).

In the present paper, a multivariate structural time series model is applied to the Dutch Labour Force Survey (LFS), to estimate monthly unemployment rates for six domains (men and women in three age classes). Until 1999 the LFS was a continuously conducted cross-sectional survey. In October 1999 this survey changed to a rotating panel design, where the respondents are interviewed five times at quarterly intervals. Based on Pfeffermann (1991), Van den Brakel and Krieg (2009) developed a multivariate model that accounts for the rotating panel design of the LFS to estimate monthly unemployment figures at the national level. Since June 2010 this multivariate model has been implemented to produce official monthly figures not only at the national level but also for six domains. This model does not take advantage of information across domains.

In this application, only the first wave of the panel design is used, since the main objective of the paper is to study the effects of borrowing strength across domains and over time in a structural time series model framework. It appears that the trends of the six domains are cointegrated and that a more parsimonious common factor model can be used. Other possible model improvements, like outlier modelling and simplification of the seasonal component, are also investigated. Some preliminary results were published in Boonstra et al. (2008).

Ignoring the rotating panel structure has the advantage that we can abandon the additional complications to account for the rotating panel design in the model and focus the discussion on the use of common factor models and model selection in the context of small area estimation. Furthermore, the length of the available series can be increased by five years, since the available cross-sectional data from 1996 until 2000 can be combined with the data observed in the first wave after 2000. Finally, the dimensions of the model would increase substantially if the panel structure were taken into account, which may result in numerical problems, especially when the length of the series is limited. The results obtained in this paper apply to cross-sectional surveys in general. It can also be anticipated that the results can be generalized to panel designs, but it requires additional research to handle the more complex model structures.

In Section 2 the LFS is summarized. In Section 3, multivariate structural time series models for the monthly unemployment rate for six domains are described. The estimation results of the models are presented in Section 4. The paper concludes with a discussion in Section 5.

2. The Dutch Labour Force Survey

The purpose of the Dutch Labour Force Survey (LFS) is to produce reliable information about the situation in the Dutch labour market. The target population of the LFS consists of the non-institutionalized population aged 15 years and over residing in the Netherlands. Each month a stratified two-stage sample of addresses is selected. Strata are formed by geographical regions. Municipalities are used 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.

Until September 1999, the LFS was a continuous survey with a monthly gross sample size of about 10,000 addresses. In October 1999 the LFS changed from a continuous survey to a rotating panel design. The monthly gross sample size of the first wave averaged about 8000 addresses at that point and gradually declined to about 6500 addresses in 2009.

During the period that the LFS was a cross-sectional survey and in the first wave of the panel design, data are collected by means of computer assisted personal interviewing (CAPI). Since the LFS has changed to a rotating panel design, the respondents are re-interviewed four times at quarterly intervals by means of computer assisted telephone interviewing (CATI). In July 2010, the LFS was redesigned substantially. Therefore the data available up to and including 2009 are used in this paper.

The estimation procedure of the LFS is based on the GREG estimator (Särndal et al., 1992). The inclusion probabilities reflect the sampling design described above as well as the different response rates between geographic regions. The