New Data Sources and Inference Methods for Official Statistics

Jan van den Brakel[1,2]

[1] Maastricht University, The Netherlands

[2] Statistics Netherlands, The Netherlands
e-mail: jbrl@cbs.nl

Abstract

Official statistics, published by national statistical institutes (NSIs) are traditionally based on repeated probability sampling in combination with design-based inference methods. This is a widely applied approach by NSIs because of its low level of risk. Under this approach NSIs are indeed in control over the availability of the data and the inference methods do not depend on statistical model assumptions of which the validity is often difficult to verify. There is nevertheless a growing interest among NSI's to use registers or other large data sets that are generated as a by-product of processes not directly related to statistical production purposes. The purpose of this, is to reduce data collection costs and response burden, to improve timeliness or to refine the level of detail of official statistics.

Roughly spoken, there are two ways to use these so called non-probability data in the production of official statistics. One approach is to use them as a primary data source for the compilations of official statistics. This generally requires a high risk apatite for the NSI, since there is no control over the availability of the data. On top of that, model-based inference methods are required to correct for selectivity. Most of these methods are based on strong missing ad random assumptions.

A second approach, which requires the acceptance of an intermediate level of risk, is to combine non-probability data sources with sample data in a model-based inference approach. Although inference methods are model-based, the NSI is still in control over the availability of the survey data that are the primary data source under this approach. Since official statistics are predominantly based on repeated surveys, time series methods provide a natural framework.

In this paper it will be illustrated how multivariate state space models and multivariate multilevel time series models can be used as a form of small area estimation by modelling temporal and cross-sectional correlations between previous reference periods and other domains. Extensions to models that include related auxiliary series to further improve the precision of the predictions will be discussed. It will be shown how dynamic factors models can be used in this context to avoid high-dimensionality problems in the case of a large amount of auxiliary series. This easily occurs if data sources like google trends are considered as auxiliary series. A major limitation of standard linear state space models is they assumption that correlations between state variables are constant over time. To relax this assumption a novel non-linear state space model will be proposed. To this end, time varying state correlations are modelled with separate stochastic processes. It will be illustrated how these methods can be used to refine the level of detail and timeliness of official statistics with real life examples at Statistics Netherlands.

Keywords: Small area estimation, multivariate state space models, time varying state correlations, dynamic factor models, multilevel time series models.