Stochastic Population Forecasting based on a Combination of Experts Evaluations and accounting for Correlation of Demographic Components

Francesco Billari, Rebecca Graziani and Eugenio Melilli[1]

EXTENDED ABSTRACT

Abstract

We suggest a method for deriving expert based stochastic population forecasts, by combining evaluations of several experts and allowing for correlation among demographic components and among experts. Evaluations of experts are elicited resorting to the conditional method discussed in Billari et al. (2012) and are then combined resorting to the supra-Bayesian approach (Lindley, 1983) so to derive the joint forecast distribution of all summary indicators of the demographic change. In particular, the elicitation procedure makes it possible to elicit evaluations from experts not only on the future values of the indicators and on their expected variability, but also on the across time correlation of each indicator and on the correlation (at the same time and across time) between pairs of indicators. The central scenarios provided by the experts on future values of each summary indicator are treated as data and a likelihood function is specified by the analyst on the basis of all additional information provided by the experts, such likelihood been parametrized in terms of the unknown future values of the indicators. Therefore the posterior distribution, obtained on the basis of the Bayes theorem and updating the analyst prior opinions on the basis of the evaluations provided by the experts, can be used to describe the future probabilistic behavior of the vital indicators so to derive probabilistic population forecasts in the framework of the traditional cohort component model.

1 Introduction

Population forecasts are strongly requested both by public and private institutions, as main ingredients for long-range planning. Virtually all population forecasts are based on the cohort-component method, so that the forecast of the population reduces to the forecast of the three main components of the demographic change: fertility, mortality and migration. Traditionally official national and international agencies derive population projections in a deterministic way: in general three deterministic scenarios

[1] Francesco Billari, Department of Policy Analysis and Public Management, Carlo F. Dondena Center for Research on Social Dynamics and IGIER, Bocconi University, Milan, Italy; Rebecca Graziani, Department of Policy Analysis and Public Management and Carlo F. Dondena Center for Research on Social Dynamics, Bocconi University, Milan, Italy; Eugenio Melilli, Department of Decision Sciences Bocconi University, Milan, Italy.

are specified, low, medium and high scenarios, based on combinations of assumptions on vital indicators and separate forecasts are derived by applying the cohort-component method. In this way, uncertainty is not incorporated so that the expected accuracy of the forecasts cannot be assessed: prediction intervals for any population size or index of interest cannot be computed. Yet the high-low scenario interval is generally portrayed as containing likely future population sizes. In recent years stochastic (or probabilistic) population forecasting has, finally, received a great attention by researchers. In the literature on stochastic population forecasting, three main approaches have been developed (Keilman et al., 2002). The first approach is based on time series models: for each indicator time series models are fitted to past series of data and forecasts are obtained resorting to usual extrapolation techniques. The second approach is based on the extrapolation of empirical errors, with observed errors from historical forecasts used in the assessment of uncertainty in forecasts (e.g., Stoto, 1983). In particular Alho and Spencer (1997) proposed in this framework the so-called Scaled Model of Error, which was used for deriving stochastic population forecasts within the Uncertainty Population of Europe, (UPE) project. Finally, the third approach referred to as random scenario defines the probabilistic distribution of each vital rate on the basis of expert opinions. In Lutz et al. (1998), the forecast of a vital rate at a given future time T is assumed to be the realization of a random variable, having Gaussian distribution with parameters specified on the basis of expert opinions. For each time t in the forecasting interval $[0, T]$ the vital rate forecast is obtained by interpolation from the starting known and final random rate. In Billari et al. (2012) the full probability distribution of forecasts is specified on the basis of expert opinions on future developments, elicited conditional on the realization of high, central, low scenarios, in such a way to allow for not perfect correlation across time.

In this work we build on Billari et al. (2012) and suggest a method that makes it possible to derive stochastic population forecasts on the basis of a combination of experts evaluations and accounting for correlation across demographic components. Our proposal is described in the following section.

2 The Proposal

Since population forecasts by age and sex are obtained resorting to the standard cohort-component method, the first issue to address is how to derive the forecast distribution of the three fundamental components of the demographic change: mortality, fertility and migration. Our forecasting method is expert based, in the sense that population forecasts strongly rely on evaluations elicited from the experts. Different sources of information, if available, are mainly used to assess expert reliability and between-experts correlations. For this reason, in order to involve experts in a simplified and direct way, we consider standard indicators, summarizing the three components of the demographic change: Total Fertility Rate for the fertility component, Male and Female Life Expectancies for the mortality component, Immigration and Emigration. The joint forecast distribution of the demographic indicators is obtained, on the basis of a combination of evaluations elicited from experts. The novelty of our contribution displays in the way evaluations of experts are at first elicited and then combined: the evaluations are elicited resorting to the conditional method discussed in Billari et al. (2012) and then combined resorting to the so-called supra-Bayesian approach. The supra-Bayesian approach was introduced by Lindley in 1983 and used, among others, by Winkler (1981) and Gelfand et al (1995) to model and combine experts opinions; later, Roback and Givens (2001) apply it in the framework of deterministic simulation models. Such approach makes it possible to combine experts opinions on unknown features of a phenomenon within the formal framework provided by the bayesian approach to statistics, by assuming that such opinions are data. The analyst is therefore asked to specify a likelihood function, to be parametrized in terms of the unknown features. The posterior distribution, obtained by applying the Bayes theorem and updating the analyst prior opinions on the basis of the evaluations provided by the experts, can be used to describe the probabilistic behavior of the unknown quantities of interest.

In the following we describe how the joint forecast distribution of any two pair of indicators at two different time points can be derived according to our proposal. It is worth emphasizing that the method can be generalized so to consider more than two indicators and more than two time points, the only additional difficulty being related to the elicitation process, that in the case of several indicators and several time points becomes cumbersome. The entire joint distribution of the pair of indicators over the considered forecasting interval can be obtained resorting to interpolation techniques.

Let us consider two indicators, denoted by R_1 and R_2 and let $[t_0\ T]$ be the forecasting interval. Following Billari et al.(2010), we split the forecasting interval into two sub-intervals, by considering an inner point t_1 in $[t_0\ T]$ and we begin by deriving the distribution law of the vector of the values of the indicators R_1 and R_2 at time t_1 and t_2, that is $(R_{11}, R_{12}, R_{21}, R_{22})$, R_{ij} being the random variable associated with the value of R_i at time j, $i = 1, 2$ and $j = t_1, t_2$, where $t_2 = T$. Consider k experts and assume that they are asked for central scenarios of the rates R_1 and R_2 at times t_1 and t_2. Let us denote by $x_i = (x_{i1}, x_{i2})$ the vector of central evaluations provided by expert i, on the pair of indicators at times t_1 and t_2. Furthermore we can elicit, from each expert and for each single indicator, information on the marginal variability of the evaluations at t_1 and t_2 and on the across time correlation. Moreover, we can also elicit information on the correlation of the evaluations on the two indicators at the same time and across time.

Following a supra-Bayesian approach, the experts central scenarios $(x_1, x_2, \ldots, x_k)$ are treated as data and the analyst is asked to specify the likelihood $f(x_1, \ldots,\ x_k | R_{11}, R_{12}, R_{21}, R_{22})$. One possible and natural choice, can be a gaussian model, that is to assume that the sample vector $(x_1, \ldots, x_k)$ is the realization of a multivariate gaussian distribution of dimension $4k$, centered at $\mu = (\mu_1, \mu_2, \ldots, \mu_k)$, where $\mu_i = (R_{11}, R_{12}, R_{21}, R_{22})$ and with covariance matrix given by:

$$\begin{pmatrix} \Sigma_1 & \Sigma_{12} & \cdots & \Sigma_{1k} \\ \Sigma_{21} & \Sigma_2 & \cdots & \Sigma_{2k} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ \Sigma_{k1} & \Sigma_{k2} & \cdot & \Sigma_k \end{pmatrix},$$

where for $i = 1, \ldots, k$, Σ_i is the covariance matrix of the evaluations of expert i on the two indicators R_1 and R_2 at time t_1, t_2; while, for $i, j = 1, \ldots, k$ $i \neq j$, Σ_{ij} is made up by the covariances between the evaluations of pairs of different experts at the same time and across time. Note that $\Sigma_{ij} = \Sigma'_{ji}$. Some comments about such specification of the likelihood are needed. First, the choice of the normal distribution can be primarily motivated by mathematical convenience; indeed, using gaussian likelihood computation of posterior quantities is greatly simplified. This is in fact a standard choice, reasonable unless clear asymmetries are suspected on the distributions of the expert evaluations or significant associations among experts' different from linear are expected. Possible different choices for the likelihood are joint distributions not normal, but having normal marginals (in this case, copulae can be a suitable option) or having not normal marginals (for instance, multivariate Gamma densities can be considered). Second, by assuming that the mean vector is equal to $(R_{11}, R_{12}, R_{21}, R_{22})$, the analyst states that he expects the experts to be unbiased in their evaluations, excluding a *systematic* underestimation or overestimation.

The elements of the matrices Σ_i are specified by the analyst on the basis of the information available. In particular, for each expert and for each indicator the marginal variances of the evaluations at time t_1 and t_2 along with the covariances across time can be specified on the basis of the information elicited resorting the conditional method and the same is for the covariances at the same time and across time of the evaluations provided by a single expert on the two indicators. As for the remaining elements of Σ, the matrices Σ_{ij}, in order to avoid an overparametrization of the model, considering that $4k$ data are available, some assumptions on the structure of the covariance can be made. In particular, we can assume that the correlations between the evaluations of any pair of experts at a given time are all equal to ρ regardless the time, the pair of experts and the indicators considered, and that also the correlations across time are all the same, and set them equal to $\rho^{t_2 - t_1}$. Under such assumptions, the analyst has therefore to specify only the value of ρ in order to specify all the matrices Σ_{ij}. Such correlation can be determined by the analyst resorting to different sources of information, but in particular on the basis of an in-depth study of the scientific production of the experts.

Under the assumption of a flat non-informative prior on $(R_{11}, R_{12}, R_{21}, R_{22})$, the Bayes theorem makes it possible to derive the posterior distribution $\pi(R_{11}, R_{12}, R_{21}, R_{22}|x_1, \ldots, x_k)$, that can therefore be used as the forecast distribution of value of the indicators R_1 and R_2 at t_1 and t_2. The posterior distribution turns out to be gaussian distribution.

The suggested method will be applied so to derive stochastic forecasts of the components of the demographic change for Italy from 2010 to 2065, on the basis of evaluations collected by means of a questionnaire written and submitted by us to a group of Italian expert demographers. The results will be shown and discussed.

References

Alho, J.M. and Spencer, B.D. (1997) "The practical specification of the expected error of population forecasts", *Journal of Official Statistics*, **13**, 204–225.

Booth, H. (2006) "Demographic forecasting: 1980 to 2005 in review", *International Journal of Forecasting*, **22**, 547–581.

Billari, F.C., Graziani R. and Melilli, E. (2012) "Stochastic population forecasts based on conditional expert opinions", *Journal of the Royal Statistical Society A*, **175**, 2, 491–511.

Gelfand, A. E., Mallick, B. K. and Dey, D. K. (1995) "Modeling Expert Opinion Arising as Partial Probabilistic Specification", *Journal of the American Statistical Association*, **90**, 430, 598–604.

Keilman, N., Pham, D.Q. and Hetland, A. (2002) "Why population forecasts should be probabilistic - illustrated by the case of Norway", *Demographic Research*, **6**, 15, 409–454.

Land, K.C. (1986) "Methods for National Population Forecasts: A Review", *Journal of the American Statistical Association*, **81**, 396, 888–901.

Lindley, D. (1983) "Reconciliation of Probability Distributions", *Operations Research*, **31**, 5, 866–880.

Lutz, W., Sanderson, W.C. and Scherbov, S. (1997) "Doubling of world population unlikely", *Nature*, **387**, 803–805.

Lutz, W., Sanderson, W.C. and Scherbov, S. (1998) "Expert-Based Probabilistic Population Projections", *Population and Development Review*, **24**, 139–155.

Roback, P.J and Givens, G.H. (2001) "Supra-Bayesian pooling of priors linked by a deterministic simulation model", *Communications in Statistics – Simulation and Computation*, **30**, 3, 381, 447–476.

Stoto, M.A. (1983) "The Accuracy of Population Projections", *Journal of the American Statistical Association*, **78**, 381, 13–20.

Winkler, R. L. (1981) "Combining Probability Distribution from Dependent Information Sources", *Management Science*, **27**, 4, 479–488.