Simple but Flexible Stochastic Population Forecasts based on Conditional Expert Opinions

Francesco C. Billari, Rebecca Graziani and Eugenio Melilli
Carlo F. Dondena Center for Research on Social Dynamics,
Department of Decision Sciences and IGIER
Università Bocconi, Milan, Italy

October 14, 2007

Abstract

In this paper we build a formal framework for developing expert-based stochastic population forecasts starting from standard scenario-based population projections. More specifically, we formalize a procedure in which the full probability distribution of forecasts is derived from 1) an initial scenario (which might be based on actual scenarios as available from standard deterministic projections); 2) expert opinions on developments conditional to the realization of high-low scenario. In the final version of the paper a full example will be developed.

(THIS IS AN EXTENDED ABSTRACT PREPARED FOR SUBMISSION TO EPC 2008-NOT TO BE CITED OR QUOTED)

1 Introduction

In recent years, stochastic population forecasting has received a great attention by researchers, albeit it has not yet influenced most official forecasting

agencies (see, e.g., Booth, 2006). The possibility to link stochastic population forecasting to classical scenario-based population projection might be a key to the future use of this approach in official population forecasting.

In stochastic population forecasting, two main directions have been explored. The first one is based on time series models and it develops stochastic population forecasts by estimating model parameters on the basis of past data (see, for a description and use of this approach, Lee and Tuljapurkar,1994). The second approach, known as "random scenario", is based on the use of expert opinions in the definition of probability distributions for the future values of vital rates and on the subsequent production of probabilistically coherent population forecasts that use projected rates in the traditional cohort-component framework. Lutz (1996) underlines that the use of expert opinions allows to take into account behavioral theories on the future of population.

Some contributions in the literature have been devoted to discuss the positive and negative aspects of both approaches and to compare them. In this extended abstract we only refer to some specific remarks. Random scenario is, as observed by many demographers (see, for instance, the comments by Goldstein, 2004, Lutz and Goldstein, 2004), more appealing to official forecasters, due to its simplicity, its traditional scenario-based framework and the direct involvement of experts. Moreover, smooth trajectories generated by random scenarios seem more suitable to adequately represent future rates trends than the often irregular paths deriving from stochastic processes based on classical time series models. See, for comments and results on the latter aspect, Tuljapurkar et al. (2004). On the other hand, existing random scenario methods, being based on trajectories obtained by the (generally linear) interpolation of a starting (known) and a final (random) rate values, are characterized by a structure of variance and correlation that is not very flexible. In particular, when the interpolation is linear, serial correlations by definition are all equal to one.

In this paper we present a proposal that lies in the framework of the random scenario approach and that aims at making more flexible the stochastic process generated, retaining at the same time the simplicity of the procedure and the immediacy of expert opinions on which the definition of the process is based. Roughly speaking, the population forecasting method we propose proceeds through a series of subsequent expert-based conditional evaluations on vital rates, given the values of the rates at some previous time points. Formally, by imposing for each rate, a gaussian joint distribution of its values at some string of equally spaced time points (i.e., splitting the projection

period), the probability distribution of the rate vector itself is completely specified. Finally, by interpolation, the process is extended over the whole forecast span. In our proposal, serial correlations and variances derive from simple and easily interpretable expert opinion and, at the same time, they show a greater flexibility than in random scenario approaches that have been developed so far.

In this extended abstract, Section 2 is devoted to a description of the proposed method. In Section 3, some simple examples are illustrated in order to highlight the main characteristics of the procedure and to compare it with other stochastic forecasting methods. The paper will contain a full-fledged example based on existing scenarios for Italy and on conditional expert opinions starting from these scenarios.

2 The Proposal

Consider a population forecasting time horizon of length T, and denote by $t_0 = 0$ the starting point of this time horizon. Population projections for the period $[0, T]$ will be functions of vital rates for the same period. In this proposal, we limit ourselves to the case of independence between vital rates, and treat migration rates as deterministic quantities. Under these latter assumptions, by generating (independent) stochastic scenarios for age specific fertility and mortality (or survival) rates over the projection period $[0, T]$, one obtains population forecasts for the same period.

Our method aims at producing independent random processes for age-specific fertility rates $F_{t,j}$ and age-specific mortality rates $D_{t,j}$, where in both cases $j = 1, 2, \ldots, n$ denotes the age class. Age-specific mortality rates can be obtained, as often done, also starting from life expectancy at birth and using a specific transformation. Similarly, age-specific fertility rates can be obtained by e.g. a combination of the total fertility rate and of the mean age at (first) birth.

We first explain the characteristics of our proposal by using a general (i.e. not age-specific) rate R_t, and describe the main steps towards the definition of the whole process $(R_t)_{t\in[0,T]}$. Such process is obtained in the following way: chosen k equally spaced time points $t_1, t_2, \ldots, t_k = T$, a (gaussian) random vector $(R_{t_1}, R_{t_2}, \ldots, R_{t_k})$ is defined. The parameters of this random vector are fixed on the basis of expert opinion, according to the procedure we describe below. Then, given the vector $(R_{t_1}, R_{t_2}, \ldots, R_{t_k})$, an interpolation

yields R_t for each $t \in [0, T]$.

In order to simplify notation and highlight the main aspects of the method, we fix $k = 2$. As a first step, experts are asked to express low, medium and high scenarios for the rate R_t at time point t_1. These quantities are used to define the expected value μ_1 and the variance σ^2 for the random variable R_{t_1}. We might think of these quantities as the one provided by official forecasting agencies. More precisely, denoting by μ_1, L_1 and U_1 respectively medium, low and high scenarios for R_{t_1}, we set $E(R_{t_1}) = \mu_1$ and $Var(R_{t_1}) = \sigma^2$, where σ^2 is chosen so that L_1 and U_1 are, respectively, the lower and upper bounds of an interval covering a probability $\gamma \in (0, 1)$ for the normal random variable R_{t_1}. Moreover, we assume that the conditional variance of R_{t_2} given R_{t_1} is equal to σ^2; in this way the variance of R_t is an increasing function of t (depending, of course, also on the correlation between R_{t_1} and R_{t_2}).

Second, experts elicit conditional forecasts for R_{t_2} given low and high scenarios for R_{t_1}. That is, experts give values $\mu_2(L_1)$ and $\mu_2(U_1)$, representing conditional guesses for the vital rate R_{t_2} at time t_2 given that R_{t_1} is, respectively, at its lowest or highest level. These quantities are interpreted as conditional expectations $E(R_{t_2}|R_{t_1} = L_1)$ and $E(R_{t_2}|R_{t_1} = U_1)$.

Finally, the hypothesis of joint normality for the random vector (R_{t_1}, R_{t_2}) and an extension to other time points $t \in [0, T]$ by (for instance linear or quadratic) interpolation completely determine the distribution of the stochastic process $(R_t)_{t\in[0,T]}$, as requested. In order to obtain age specific vital rates, i.e. fertility and mortality (or survival) age specific rates, we use well-known procedures or models, such as model life tables or the model proposed in Tuljapurkar et al. (2004).

3 A Simple Example and Discussion

In order to show some results obtained with our method and highlight the main differences between this and other approaches we use (as a first example) an oversimplified (i.e., without age structure) population renewal model as done in Tuljapurkar et al.(2004). In this simple example, values determined by expert's opinions are chosen so to make input quantities as similar as possible to those used in the cited paper.

Let us consider a projection span of T years and suppose the renewal

population equation to be simply

$$N_{t+1} = R_{t+1} N_t \qquad t \in [0, T], \tag{1}$$

where N_t and R_t denote respectively population size and growth factor at time t, being N_0 and R_0 (known) values at starting time $t = 0$. We refer to the procedure described in the previous section and consider a partition of the projection span in $k = 2$ sub-periods, determined by time points $t_0 = 0$, t_1 and $t_2 = T$; in the example, $t_1 = 25$ and $t_2 = 50$, $N_0 = 6$ and $R_0 = 1.028$. Experts are asked to elicit the quantities concerning the growth factor R_t described in Section 2; the values chosen are

- $\mu_1 = E(R_{t_1}) = 0.524$;
- $L_1 = 0.519$ and $U_1 = 0.529$. With a covering probability γ equal to 0.9, we obtain a variance $\sigma^2 = 9.225 \cdot 10^{-6}$ for R_{25};
- $\mu_2(L_1) = E(R_{t_2}|R_{t_1} = L_1) = 0.012$ and $\mu_2(U_1) = E(R_{t_2}|R_{t_1} = U_1) = 0.032$;

Using these values, we have completely characterized a gaussian random vector (R_{t_1}, R_{t_2}) with mean vector $(0.524, 0.022)$, variances $9.225 \cdot 10^{-6}$ and $4.612 \cdot 10^{-5}$ and correlation coefficient 0.894. For $t \in (0, 25)$, by linear interpolation,

$$R_t = \frac{1}{25} \left(t R_{25} + (25 - t) R_0 \right);$$

for $t \in (25, 50)$,

$$R_t = \frac{1}{25} \left((t - 25) R_{50} + (50 - t) R_{25} \right).$$

Of course both $Var(R_t)$ as well as $corr(R_s, R_t)$ can be computed for each $s, t \in [0, 50]$. We report below a few of these values:

$$Var(R_t) = \begin{cases} 1.5t^2 \cdot 10^{-8} & \text{if } t \in (0, 25); \\ 2.95 \cdot 10^{-8} t^2 - 2.95 \cdot 10^{-6} t + 1.84 \cdot 10^{-5} & \text{if } t \in (25, 50) \end{cases}$$

and

$$corr(R_s, R_t) = \begin{cases} 1 & \text{if } s, t \in (0, 25); \\ \frac{0.96t}{\sqrt{1.845t^2 - 46.05t + 575}} & \text{if } s \in (0, 25) \text{ and } t \in (25, 50); \\ \frac{1.845st - 23.025s - 23.025t + 57.5}{\sqrt{1.845t^2 - 46.05t + 575}\sqrt{1.845s^2 - 46.05s + 575}} & \text{if } s, t \in (25, 50). \end{cases}$$

It is useful to notice the (obvious) fact that the variance of the rate R_t is (quadratically) increasing in t and observe the expression of the correlation between R_s and R_t as a function of s and t; remember that in the traditional stochastic scenario model the correlations are all equal to one. Having defined the whole process for the rate R_t, using (1) we can of course obtain the corresponding distribution of the population process N_t; this is better described and compared with those obtained by other methods by means of simulation of trajectories from it. Results will be given in the full paper.

Let us conclude with a brief discussion of the proposed method and a comparison of it with other existing projection models, using the numerical results in the example. As a first remark, we stress that the proposed projection procedure has a very simple structure and implementation, it can immediately be interpreted and used by official forecasters and, above all, solicits expert's opinion in a direct and simple, even if unusual, way. Referring to this latest aspect, we believe that asking an expert for conditional values of rates (as it is done in the proposed method) is easier than directly soliciting correlation coefficients or variances. On the other hand, with respect to the classical random scenarios with a single final random variable R_T, this procedure appears more flexible and rich, not imposing serial correlations equal to 1 and letting more choices in the trend of the projection over time.

References

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

[2] Goldstein, J.R. (2004), "Simpler Probabilistic Population Forecasts: Making Scenarios Work", *International Statistical Review*, 72, 1, 93-106.

[3] Lutz, W. (Ed.) (1996). *The future population of the world: What can we assume today?* (Revised ed.) London: Earthscan.

[4] Lutz, W. and Goldstein, J.R. (2004), "Introduction: How to Deal with Uncertainty in Population Forecasting ?", *International Statistical Review*, 72, 1 1-4.

[5] Lee, R.D. and Tuljapurkar, S.(1994), "Stochastic Population Forecasts for the United States: Beyond high, Medium, and Low", *Journal of the American Statistical Association*, 89, 428, 1175-1189.

[6] Tuljapurkar, S., Lee, R.D. and Li, Q. (2004) "Random Scenario Forecast Versus Stochastic Forecasts" ,*International Statistical Review*, 72, 2, 185-199.