WP 35
26 April 2010

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

CONFERENCE OF EUROPEAN STATISTICIANS

Joint Eurostat/UNECE Work Session on Demographic Projections
(28-30 April 2010, Lisbon, Portugal)

Item 12 – Stochastic techniques for demographic projections

Combining Deterministic and Stochastic Population Projections[1]

Salvatore Bertino[2] and Eugenio Sonnino[3], University "La Sapienza", Italy,
and Giampaolo Lanzieri[4], Eurostat

1. INTRODUCTION

Population projections are sometimes requested for the assessment of the impact of population developments in the very long run. Such a long time horizon clearly undermines whatever meaning of forecast one would like to attribute to the projections results. Despite of that, users have the tendency to consider long-term projections not as one of the plausible developments of the population under given assumptions, but instead as an "almost sure" outcome. In order to raise the awareness of the users about the inherent uncertainty of population projections, the quantification of prediction intervals may play a very important role.

In conventional population projections exercises, this is usually done by producing results corresponding to different paths of evolution of the demographic components. Population projections that are the outcome of these alternative sets of assumptions are called variants. It is common practice of the official institutes producing projections to release several variants, often corresponding to more optimistic/pessimistic developments than those assumed in a baseline projection. However, the user is normally left with no hints about the higher or lower plausibility of the different variants. Certainly, strategies can be implemented by the users to make a rational and less costly choice among the available options (Duchêne and Wanner, 1999), but nevertheless the message on the uncertainty associated to the various variants is not straightforward. This is better accomplished in a stochastic framework, where probabilities can be associated to projections outcomes.

The quantification of the uncertainty can be considered the most important development in population forecasting in the past decade (Wilson and Rees, 2005). Four main approaches have been proposed in literature to quantify uncertainty in population projections: the first is based on the application of time series techniques directly on demographic components (like the total fertility rate) or on parameters of a mathematical expression of age-specific rates; the second approach analyses *ex-post* the errors of previous projections exercises to get information about future likely errors; the third builds upon the experts' judgement; the fourth method uses the output of micro-simulations to quantify the forecast intervals. Attempts to apply a combination of these methods have also been presented in literature (e.g., Keilman *et al.*, 2002). To the last of the four groups listed above belongs the method proposed by Bertino and Sonnino (2007), based on the simulation of birth, death, immigration and emigration point-event process, resulting from compounding independent Poisson processes. The same authors have also developed software for the implementation of their method (Bertino and Sonnino, 2010).

Despite of the richness of proposals in literature, probabilistic projections have still a number of shortcomings which somehow slow down the process of adoption of these methods by the official forecasters. For instance, besides the technical complexity of some of them, which may require a knowledge not always belonging to demographers (e.g., ARIMA models), subjectivity may still be present to a certain extent (e.g., in the choice of the base period or of the

---

[1] The views expressed in this paper are exclusively those of the authors and may not represent the views of the European Commission.
[2] University "La Sapienza", Department of Statistics, Probability and Applied Statistics, Rome. E-mail: salvatore.bertino@uniroma1.it
[3] University "La Sapienza", Department of Social, Economic, Actuarial and Demographic Studies, Rome. E-mail: eugenio.sonnino@uniroma1.it
[4] European Commission – Eurostat, Population Unit, Luxembourg. E-mail: giampaolo.lanzieri@ec.europa.eu

model for extrapolation), or the outcome may sometimes be implausible, or again giving such a wide range of variation to be useless for practical purposes. Further, sometimes it is not possible to incorporate demographic knowledge into the stochastic approach, knowledge which at the very end is the real element of strength of conventional cohort-component projections, nor visions about future long-term demographic developments can always be easily integrated. These limitations of the probabilistic methods may become as more relevant as more the time horizon of the projections is set in the far future. Therefore, the attempt to combine deterministic and stochastic projections answers to the need of preserving the advantages of the conventional approach, whilst quantifying the inherent uncertainty of projections.

This article focuses on the way to integrate deterministic projections within a stochastic framework as from the micro-simulation approach of Bertino and Sonnino (BS hereinafter). In Section 2, there is a description of the possibilities of linkage between the conventional deterministic projections and the stochastic BS method, including its consequences in terms of estimation of the variability and therefore of the forecast intervals. In Section 3 are reported some elements for discussion.

2. THE METHOD

In the conventional cohort-component approach, plausible demographic developments take the form of fertility, mortality and migration age-specific (and sex-specific, whether appropriate) rates assumed to occur along the time horizon of the population projections. The combination with a stochastic approach would entail that the information represented by the assumed "conventional" deterministic rates is somehow incorporated into that probabilistic method.

The BS method requires the definition of instantaneous rates for the Poisson processes. In its simplest version, it starts from the probabilities of the vital events *e* (fertility, mortality and migration) for each age *a* and (whether appropriate) sex *s* at the beginning and at the end of the projections period. Probabilities for intermediate years are obtained by linear interpolation and modified accordingly to a random walk. These probabilities are then transformed in instantaneous rates of Poisson processes by means of the following equation:

$$\lambda_e(a,s) = -\ln[1 - q_e(a,s)] \tag{1}$$

These instantaneous rates, assumed equal within each age-sex category, are modified after each event occurring in the population (for further details on the methodology, see Bertino and Sonnino, 2007). The BS method has thus an entry door for inputting demographic knowledge in its probabilistic framework. It is indeed sufficient to assume that the instantaneous rates for each age and sex group at the beginning of each year of projections are equal to the corresponding deterministic rate use in the conventional projections. Although the instantaneous rates are stochastically modified also between simulations, these deterministic rates may "drive" the stochastic process accordingly to defined long-term scenarios.

In the BS method, given *m* simulations, any statistic of interest *V* (e.g., size of age groups, number of live births, old age dependency ratio, etc.) can be estimated by:

$$\hat{\mu}_s = \frac{1}{m}\sum_{i=1}^{m} v_i \tag{2}$$

where $\hat{\mu}_s$ is the estimated average of the determinations v_i of the statistic of interest *V* from the stochastic simulations. Likewise, the variance $\hat{\sigma}_s^2$ of the statistic *V* can be estimated by:

$$\hat{\sigma}_s^2 = \frac{1}{m-1}\sum_{i=1}^{m}(v_i - \hat{\mu}_s)^2 \tag{3}$$

From (2) and (3) it is possible to derive an expression for the forecasts intervals. Assuming that the distribution of the statistic *V* is properly approximated by a Normal distribution $V \approx N(\mu, \sigma^2)$, whose average μ and variance σ^2 can be estimated respectively by (2) and (3), it can be shown that the forecast interval for *V* is:

$$Pr\left\{\hat{\mu}_s - T_{\alpha/2,m-1}\cdot\hat{\sigma}_s\cdot\sqrt{1+\frac{1}{m}} \le V \le \hat{\mu}_s + T_{\alpha/2,m-1}\cdot\hat{\sigma}_s\cdot\sqrt{1+\frac{1}{m}}\right\} = 1-\alpha \tag{4}$$

where $T_{\alpha/2,m-1}$ is the $100\cdot(1-\alpha/2)$ percentile of a random variable of Student with *m*-1 degrees of freedom. By using (4), it is thus possible to quantify the uncertainty for any statistic of interest that can be computed from the output of the stochastic simulations.

The possible uncertainty related to the deterministic assumptions can properly be taken into account in this framework as well. Such uncertainty in deterministic projections can basically be expressed by alternative values (or ranges of variation) of selected demographic indicators (e.g., total fertility rate). For instance, an expert could judge that, in a given year, the value of the total fertility rate (TFR) could be included in a range defined by the given TFR assumption ± 0.15 live births per woman. The expert opinion can then be used to calculate alternative sets of parameters to be used as input for the projections; e.g., two sets of age-specific fertility rates could be estimated such that the corresponding TFRs would be equal to the extremes of the range of variation expressed by the expert. This logic can be applied to fertility as well as to mortality and migration.

To incorporate the uncertainty expressed by the variants into the BS method, it is sufficient to replace the instantaneous rates of the stochastic processes with the alternative sets of corresponding parameters. Let us assume that k sets of parameters are produced for the deterministic projections (thus composed by k variants) and let $m_1, m_2, \ldots, m_k$ be the number of simulations carried out with each set of parameters. The ratio m_i/m, being $m=m_1+m_2+\ldots+m_k$, can in fact be considered a measure of the confidence attributed to the i-th variant. The forecast interval can again be calculated using the (4), and in this case its range can be expected to be wider due to the imputed variability of the parameters. The significance of the differences of the projections outcomes for the statistic of interest as from the alternative sets of parameters can be tested with the usual techniques of analysis of variance (ANOVA) on the k groups.

Let clarify with a simple example: three variants (Low, Medium and High) are produced in a framework of conventional deterministic projections. Let the Medium variant be the result of baseline values for fertility, mortality and migration; further, let the Low variant be the outcome of lower fertility and higher mortality (thus resulting in smaller population size), and vice versa for the High variant. This is a typical outcome of a conventional projections exercise. Let now suppose that the expert attributes 80% of confidence to the baseline assumptions and only 10% each to the alternative assumptions. Assuming that 100 simulations are carried out, 80 will use as input for the instantaneous rates of the BS method the conventional baseline rates, 10 will use those from the Low variant and 10 from the High variant. Focussing on the population size, it could be expected that the simulations using rates from the Low variant would produce on average lower population values than from the Medium variant, and vice versa for the High variant, thus increasing the variability of the outcomes.

An alternative way to incorporate the uncertainty related to the deterministic assumptions into a probabilistic framework can be developed after the computation of the deterministic projections. In this situation, the variability of any statistic of interest can be estimated from the variants of the conventional projections. However, it should be noted that no level of probability is associated to it: therefore, such variability has still a deterministic nature. Let thus $\hat{\mu}_d$ be the estimated average of the statistic of interest V calculated across the k projections variants:

$$\hat{\mu}_d = \frac{1}{k}\sum_{j=1}^{k} v_j \tag{5}$$

and $\hat{\sigma}_d^2$ the variance:

$$\hat{\sigma}_d^2 = \frac{1}{k}\sum_{j=1}^{k}\left(v_j - \hat{\mu}_d\right)^2 \tag{6}$$

The variance $\hat{\sigma}_d^2$ could also be directly expressed by the expert or the result of alternative methods of estimation. Further, the (5) could take the form of a weighted average, in case preference towards one or another variant is expressed:

$$\hat{\mu}'_d = \sum_{j=1}^{k} w_j \cdot v_j \quad ; \quad 0 \le w_j \le 1 \quad \forall j = 1,\ldots,k \tag{7}$$

Given the expression of deterministic uncertainty (6) related to a given statistic of interest, two approaches are possible. In the first, the two variances (deterministic and stochastic) are assumed to be independent and p is the weight attributed to the deterministic component. Then:

$$\hat{\sigma}_V^2 = p^2 \cdot \hat{\sigma}_d^2 + (1-p)^2 \cdot \hat{\sigma}_s^2 = (1-p)^2 \cdot \hat{\sigma}_s^2 \cdot \left[1 + \frac{p^2 \cdot \hat{\sigma}_d^2}{(1-p)^2 \cdot \hat{\sigma}_s^2}\right] \tag{8}$$

And the forecast interval is:

$$Pr\left\{\hat{\mu}_V - T_{\alpha/2,m-1}\cdot\hat{\sigma}_s\cdot\sqrt{(1-p)^2+\frac{p^2\cdot\hat{\sigma}_d^2}{\hat{\sigma}_s^2}} \le V \le \hat{\mu}_V + T_{\alpha/2,m-1}\cdot\hat{\sigma}_s\cdot\sqrt{(1-p)^2+\frac{p^2\cdot\hat{\sigma}_d^2}{\hat{\sigma}_s^2}}\right\} = 1-\alpha \quad (9)$$

being:

$$\hat{\mu}_V = p\cdot\hat{\mu}_d + (1-p)\cdot\hat{\mu}_s \quad (10)$$

In case equal confidence is given to the deterministic and to the stochastic component, the (9) becomes:

$$Pr\left\{\hat{\mu}_V - T_{\alpha/2,m-1}\cdot\frac{\hat{\sigma}_s}{2}\cdot\sqrt{1+\frac{\hat{\sigma}_d^2}{\hat{\sigma}_s^2}} \le V \le \hat{\mu}_V + T_{\alpha/2,m-1}\cdot\frac{\hat{\sigma}_s}{2}\cdot\sqrt{1+\frac{\hat{\sigma}_d^2}{\hat{\sigma}_s^2}}\right\} = 1-\alpha \quad (11)$$

An alternative method to incorporate deterministic uncertainty is based on the Bayesian approach. In this framework, the deterministic forecast and its measure of uncertainty is the initial distribution (before the experiment) of the statistic of interest, while the stochastic forecast is the outcome of an experiment of *m* independent trials of a simulation procedure. Assuming that the initial distribution is $N(\mu_d, \sigma_d^2)$ and the stochastic simulations are the outcome of a $N(\mu_V, \sigma_V^2)$, whose variance is estimated by $\hat{\sigma}_s^2$, the final forecast distribution is:

$$N\left[\frac{\mu_d\cdot\sigma_s^2 + m\cdot\mu_s\cdot\sigma_d^2}{\sigma_s^2+m\cdot\sigma_d^2}, \sigma_s^2\cdot\left(1+\frac{\sigma_d^2}{\sigma_s^2+m\cdot\sigma_d^2}\right)\right] \quad (12)$$

The (12) is valid if the parameter σ_s^2 is known. If this variance is estimated, then:

$$\frac{\mu - \dfrac{\mu_d\cdot\hat{\sigma}_s^2 + m\cdot\mu_s\cdot\sigma_d^2}{\hat{\sigma}_s^2+m\cdot\sigma_d^2}}{\sqrt{\hat{\sigma}_s^2\cdot\left(1+\dfrac{\sigma_d^2}{\hat{\sigma}_s^2+m\cdot\sigma_d^2}\right)}} \approx T_{\alpha/2,m-1} \quad (13)$$

from which:

$$\Pr\left\{\begin{array}{l}\dfrac{\hat{\mu}_d\cdot\hat{\sigma}_s^2 + m\cdot\hat{\mu}_s\cdot\hat{\sigma}_d^2}{\hat{\sigma}_s^2+m\cdot\hat{\sigma}_d^2} - T_{\alpha/2,m-1}\cdot\hat{\sigma}_s^2\cdot\sqrt{1+\dfrac{\hat{\sigma}_d^2}{\hat{\sigma}_s^2+m\cdot\hat{\sigma}_d^2}} \le V \\ \qquad \le \dfrac{\hat{\mu}_d\cdot\hat{\sigma}_s^2 + m\cdot\hat{\mu}_s\cdot\hat{\sigma}_d^2}{\hat{\sigma}_s^2+m\cdot\hat{\sigma}_d^2} + T_{\alpha/2,m-1}\cdot\hat{\sigma}_s^2\cdot\sqrt{1+\dfrac{\hat{\sigma}_d^2}{\hat{\sigma}_s^2+m\cdot\hat{\sigma}_d^2}}\end{array}\right\} = 1-\alpha \quad (14)$$

If the two variances (deterministic and stochastic) are equal, then the (14) becomes:

$$\Pr\left\{\frac{\hat{\mu}_d + m\cdot\hat{\mu}_s}{1+m} - T_{\alpha/2,m-1}\cdot\hat{\sigma}_s^2\cdot\sqrt{1+\frac{1}{1+m}} \le V \le \frac{\hat{\mu}_d + m\cdot\hat{\mu}_s}{1+m} + T_{\alpha/2,m-1}\cdot\hat{\sigma}_s^2\cdot\sqrt{1+\frac{1}{1+m}}\right\} = 1-\alpha \quad (15)$$

In this case, the deterministic value is considered as the outcome of a further simulation.

3. DISCUSSION

The importance of the quantification of uncertainty in population forecasting has been highlighted by several scholars. In a conventional projections exercise, uncertainty is usually expressed by means of variants; however, by doing so, no probability is associated to the various options, nor is the user driven in the choice of the most appropriate variant. Several methods for stochastic projections have then been proposed in literature, by which it is possible to obtain forecast intervals. On the other hand, the "demographic knowledge" incorporated in the assumption-setting for deterministic projections risks to be – at least partially - lost when a fully stochastic approach is undertaken. Various attempts are thus been made to link these two aspects of a projections exercise.

This article builds on the method for stochastic projections proposed by Bertino and Sonnino (2007) to develop the integration between deterministic and stochastic projections. The BS method is particularly suitable for this task because it is based on micro-simulations of the behaviour of a population, which gives great flexibility in the outputs that can be derived, and it allows inputting deterministic assumptions straight into the stochastic procedure. The link is straightforward, as it is sufficient to consider as input for the instantaneous rates of the point-event Poisson processes of the BS method the deterministic rates produced for the conventional projections.

In addition, methods are here proposed to include also the uncertainty that may have been formulated in the context of the deterministic exercise. In particular, two ways are proposed: the *ex-ante* approach directly incorporates in the simulation procedure the deterministic uncertainty, as expressed by alternative sets of deterministic rates; the *ex-post* method instead includes an estimate of the deterministic variability directly into the formulas for the computation of the forecasts intervals. This latter approach is also developed in a Bayesian framework.

The BS method allows the calculation of forecast intervals for any statistic of interest. This is of considerable importance for certain policy needs, where specific indicators are considered to be of particular relevance. Work is still ongoing to test empirically the proposed approach. In particular, an application is currently being developed on EUROPOP2008, a set of population projections for the European Union (for details about these projections, see Lanzieri 2009).

As conventional deterministic projections are still common practice in the national statistical offices, a possible strategy could then be based on the following steps: first, the deterministic exercise is carried out, including also the release of various variants; second, the deterministic projections are combined with the stochastic method, producing forecast intervals for the statistics of interest. This two-steps strategy is easy to implement and allows keeping the “control” on the assumed future demographic scenario, associating at the same time a probability level.

4. REFERENCES

Bertino, S. and E. Sonnino (2007): “A new technique for stochastic population projections”. Proceedings of the Eurostat-UNECE Workshop on Demographic Projections, Bucharest, 10-12 October 2007, 161-175.

Bertino, S. and E. Sonnino (2010): PRODEST – Proiezioni Demografiche Stocastiche, version 5.051.

Duchêne J. and P. Wanner (1999): “Uncertainty in demographic projections and its consequences for the user”. *Statistical Journal of the United Nations ECE*, 16:267-279.

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

Lanzieri G. (2009): "EUROPOP2008: a set of population projections for the European Union". Paper for the XXVI IUSSP International Population Conference, Marrakech, October 2009.

Wilson, T. and P. Rees (2005): "Recent Developments in Population Projections Methodology: A Review". *Population, Space and Place*, 11:337-360.