International Institute for Applied Systems Analysis
Schlossplatz 1 • A-2361 Laxenburg • Austria
Telephone: (+43 2236) 807 342 • Fax: (+43 2236) 71313
E-mail: publications@iiasa.ac.at • Internet: www.iiasa.ac.at

Interim Report IR-99-063

Variances of Population Projections: Comparison of Two Approaches

Dietmar Bauer (Dietmar.Bauer@tuwien.ac.at)
Gustav Feichtinger (or@e119ws1.tuwien.ac.at)
Wolfgang Lutz (lutz@iiasa.ac.at)
Warren Sanderson (wsanderson@datalab2.sbs.sunysb.edu)

Approved by

Gordon J. MacDonald (macdon@iiasa.ac.at)
Director

November 8, 1999

Contents

Abstract

There has been a recent upsurge of interest in probabilistic population projections. Two methods have been suggested in the literature for forecasting the inputs into those projections: (1) a random lines (RL) approach, and (2) a simple autoregressive approach (AR(1)). The purpose of this paper is to study analytically the differences produced by the two methods. We do this in the context of a model of variability in population growth rates. Two cases are considered: One where there is no population age structure and one where there is one. In both, we find that the variance using the AR(1) approach is initially smaller than that of the RL approach, but that over time the variance using the RL approach grows more rapidly leading to an instant of time when the variances are equal.

About the Authors

Dietmar Bauer and Gustav Feichtinger are from the Institute for Econometrics, Operations Research and System Theory, Technical University of Vienna, Argentinierstrasse 8, A-1040 Vienna, Austria. Professor Feichtinger is also a part-time Research Scholar at IIASA.

Warren Sanderson is from the Departments of Economics and History, State University of New York at Stony Brook, Stony Brook, NY 11794-4384, U.S.A. When this paper was written he was also a Research Scholar with the Population Project at IIASA.

Wolfgang Lutz is the Leader of the Population Project at IIASA.

Variances of Population Projections: Comparison of Two Approaches

Dietmar Bauer, Gustav Feichtinger, Wolfgang Lutz, Warren Sanderson

Probabilistic population projections have recently been receiving more attention. The reason for this seems to lie in a dissatisfaction with the conventional approach of producing high and low variants in addition to a medium variant. There seem to be three main points of criticism of this approach as, e.g., represented by the UN population projections published every two years.

One often criticized aspect lies in the fact that the variants only modify fertility assumptions and do not take account of the uncertainty in future mortality and migration trends. Especially with respect to future life expectancy there seems to be a greater sense of uncertainty today than some years ago in both high and low mortality countries. The HIV/AIDS pandemic and the return of some infectious diseases, together with environmental changes that may impact health and basic subsistence in certain parts of the world, have made declining life expectancy a sad reality in some countries and a real possibility for the future in a larger number of countries. On the other hand, a strong scientific dispute about the upper limit of the human life span makes it an open question, whether in some countries, by the middle of the next century, life expectancy can reach 95 years or will stop around 85 years. Hence there is a sense that these great uncertainties should also be reflected in the population projections.

Another problem has to do with the regional aggregation of high and low variants. As defined by the UN, "low and high fertility variants are usually thought to bracket the probable range of future population change for each country." Whatever is meant by a "probable range," it is going to be a very different kind of range. If the strong additional assumption of perfect correlation among countries is made, e.g., in the case of the global high variant, all countries in the world will simultaneously experience the upper end of the range. By orders of magnitude, this should be less likely than for a given country to experience the upper variant.

This problem leads to the final and most general point: there are essentially two alternative ways to deal with the issue of possible deviations from the medium or main variant, which is also considered to be the best guess: One is to go for scenarios which demonstrate the effect of a specific consistent set of assumptions without attempting to make any statement about the probability of the specific combinations. Such not-impossible scenarios can have important educational functions but are not really a good planning tool because they do not tell the user whether the case described is of such probability that it should be taken seriously, or whether it can be disregarded as a very low probability event, such as a major meteorite hitting the earth soon. The alternative is to go for explicitly probabilistic projections. To the users who are interested in information beyond a best guess variant, such probabilistic projections can give important additional information, especially in cases that involve a cost function. For instance, in the case of the reform of a pension system, a deviation from the expected path by 20% may mean the collapse of the system, while a deviation of 5% could still be absorbed. Clearly, a policy maker will be interested whether a 20% deviation is something he will have to actually worry about, or whether it is an extremely low probability event. The traditional high (or low) variant is difficult to interpret. A consistent probabilistic framework seems to be the only meaningful way to deal with this issue.

Among a larger number of proposals on how to design probabilistic projections (Keyfitz 1981; Stoto 1983; Ahlburg and Land 1992; Pflaumer 1992; Lee and Tuljapurkar 1994; Alho 1990; Lutz et al. 1996, 1997) there seem to be two main approaches that have been translated into empirical applications for specific countries. One is based mainly on time series analysis and uses autoregressive time series models (AR(1)). The other is closely related to the traditional expert-based variants approach which typically assumes linear trends between the starting value and the value assumed in the target year. If the values in the target year are assumed to have a certain distribution, this technique may be labelled a probabilistic random lines approach (RL). An example of the first approach is the work by Lee and Tuljapurkar (1994), which focuses on the US. The latter approach has been undertaken by Lutz et al. (1997) for all world regions, and recently Lutz and Scherbov (1997) for Austria.

Because of the different kinds of data required and the different ways of deriving and using expert judgment, the two approaches cannot be directly compared. Since for many countries in the world no reliable demographic time series data exist and the application of any time series model requires a certain number of expert choices, there never can be a fully "objective" way to obtain the future variance of the process. Empirical sensitivity analyses of some of the assumptions involved in the different models have been given elsewhere (Lutz and Scherbov 1997).

An unresolved analytical question is whether the random lines approach systematically underestimates the variance of demographic variables of interest as compared to the AR(1) approach. Lee (1999, p. 172) conjectures with respect to this approach that it "could not exactly generate the probability distribution for the age structure (dependency ratios, for example) or for any other measure that depends on the shapes of vital rate trajectories." This is a crucial question to be resolved before generally recommending this approach to statistical agencies. This specific question is the focus of this paper.

The paper is organized as follows. In Section 1, the univariate case of an aggregated population is considered. Section 2 deals with the case of an age-structured population with only one fertile age group. In Section 3 we then compare the analytical results to the findings from simulations. Section 4 concludes the paper.

1. The Univariate Case

In this section, we will consider the following problem: Let the population size at time t be denoted with P_t. The population size at time 0, P_0 , is measured and thus given as a deterministic variable. We will assume, that population growth can be described by a model of the following form:

$$P_{t+1} = e^{r_t} P_t$$

Here r_t denotes the growth rate, which is assumed to be a random variable. Again the growth rate at time 0, r_0, is assumed to be given and deterministic. The population projection is now performed using the following general scheme:

1. Experts are asked for their guesses of three values of r_T, for some projection horizon T: The estimated mean plus two values specifying the uncertainty of the guess, e.g., the 5% and the 95% quantiles.
2. The stochastic process (r_t), $1 \le t \le T$ is modeled by one of the two approaches described in 1.1 and 1.2 below.
3. A number of realizations corresponding to the stochastic process derived in step 2 are generated.
4. The population size is calculated for each of the realizations of step 3.
5. The distribution of the population projection is evaluated and its quantiles are plotted.

The two approaches, which are compared in this paper can be described as follows.

1.1. The Random Lines Approach

In the random lines approach, r_t is assumed to be a random variable, which can be decomposed into two components: $r_t = r_t^d + r_t^s$. Here r_t^d denotes the deterministic part of the random variable, which can be calculated from the mean r_T^d, given by the experts, and the initial growth rate r_0 as follows:

$$\mathrm{r}_t^{\mathrm{d}} = \frac{\mathrm{t}}{\mathrm{T}} \mathrm{r}_{\mathrm{T}}^{\mathrm{d}} + \frac{\mathrm{T-t}}{\mathrm{T}} r_0$$

The stochastic part r_t^s is derived from the experts' guess in the following way: the two quantities provided by the experts, corresponding to the uncertainty of the guess of the mean, can be used to specify a distribution (e.g., a normal distribution), which represents this uncertainty. Denote this random variable with r_T^s. Then r_t^s is assumed to be $t/T\, r_T^s$. This corresponds to Figure 1.

Note that for the main result, we do not impose the assumption that the distribution of r_T^s is normal. Also note that the deterministic part r_T^d can show any behavior provided the values at time 0 and at time T coincide with the corresponding values r_0^d and r_T^d. Both assumptions, that the normal distribution is used and that the deterministic part of the growth rate is linear, are made only for the sake of notational simplicity. They are not crucial for the analysis, as will be clear from the discussion.

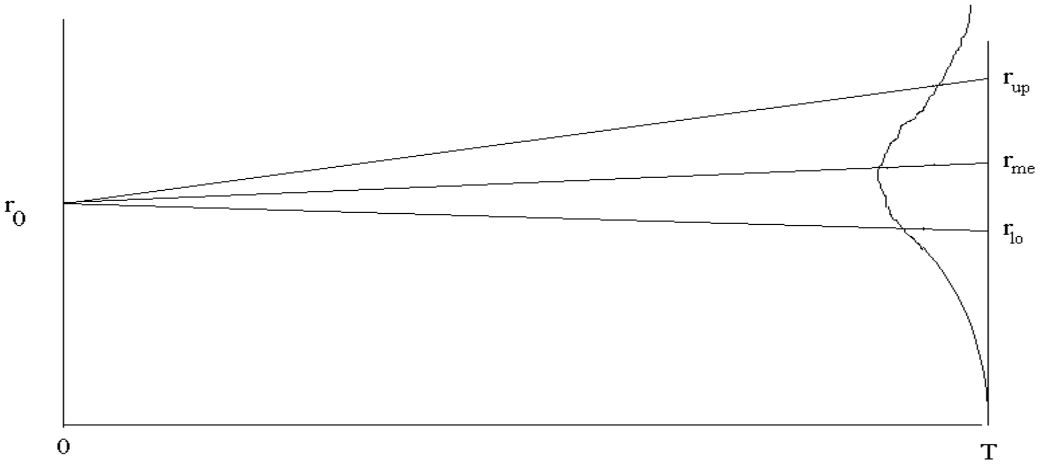

Figure 1. Scheme for the random lines approach. r_0 is measured, r_{me} is the mean given by the experts. r_{lo} and r_{up} are the upper and the lower points given by the experts. These two points are used to construct the distribution of the growth rate at time *T*.

The variance of r_T^s is given by the experts as indicated above and will be denoted with V_T^2. Note that due to the assumed model of the population size, the following equation holds true:

$$\log(P_{t+1}) = \log(P_t) + r_t = \sum_{j=1}^{t} r_j + \log(P_0)$$

Thus the logarithm of P_{t+1} can be written as the sum of two terms: $log(P_0)$ and the sum of r_j. Since P_0 is assumed to be deterministic, the variance contribution of this term is equal to zero. Thus we have to investigate the variance of the first term. Recall, that r_j was decomposed into a deterministic and a stochastic part. The variance is due to the stochastic part only and can be written as:

$$Var(\sum_{j=1}^{t} r_j^s) = Var(\sum_{j=1}^{t} \frac{t}{T} r_T^s) = (\sum_{j=1}^{t} \frac{t}{T})^2 Var(r_T^s) = \frac{t^2(t+1)^2}{4T^2} V_T^2$$

Note that for $t=T$ this variance is quadratic in *T*.

1.2. The AR(1) Approach

This approach uses a different model for the stochastic part of the process (r_t). Again the growth rate is decomposed into a deterministic part and a stochastic part: $r_t = r_t^d + r_t^s$. But now the stochastic part is modeled as an AR(1) process, i.e., the following equation holds true:

$$r_t^s = \alpha r_{t-1}^s + e_t$$

Here (e_t) denotes a white noise process, i.e., e_t is a sequence of independently identical distributed random variables with mean zero and variance V_e^2. Thus $E(e_t\, e_s^T) = 0$ for $s \neq t$ and $E(e_t^2) = V_e^2$ where *E* denotes the expectation. $|\alpha| < 1$ is a real number, the autoregression coefficient. A typical realization of an AR(1) process with autoregression coefficient α = 0.85 can be seen in Figure 2.

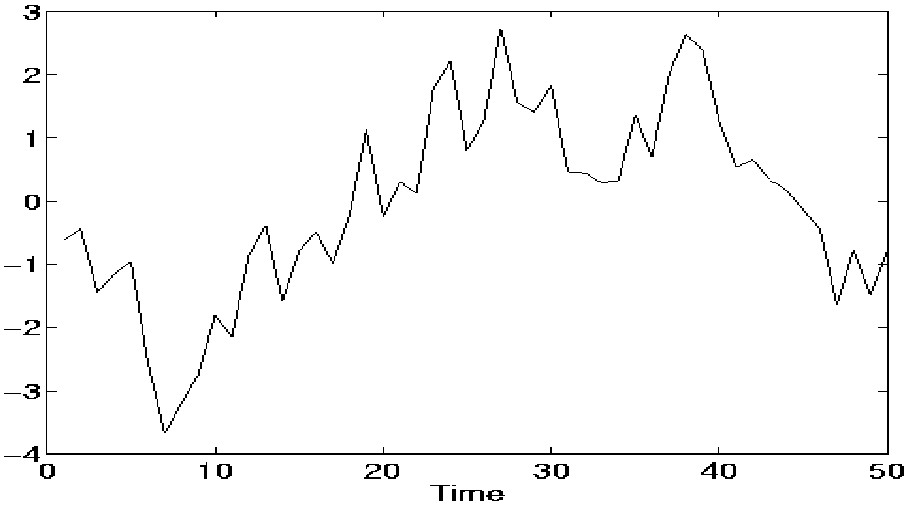

Figure 2. One example of a trajectory of an AR(1) process, T=50, α =0.85.

Note that the choice of α introduces another parameter, which has to be somehow specified. Two possibilities are to estimate it from past observations, or to use another expert's guess. However, as will be clear from the results below, the choice of α does not influence the results obtained in this paper (qualitatively). We will thus assume that α has already been specified. Using the recursion defining the autoregression, we obtain: $r_t^s = \sum_{j=0}^{t-1} \alpha^j e_{t-j}$. The variance of the white noise process is derived from the variance of r_T^s. The restriction, that r_T^s has variance V_T^2 leads to an equation defining the variance of the noise e_t. Let the variance of the noise be denoted with V_e^2. Then the autoregressive model for r_t^s implies:

$$V_T^2 = Var(\sum_{j=0}^{T-1}\alpha^j e_{T-j}) = \sum_{j=0}^{T-1}\sum_{i=0}^{T-1}\alpha^{i+j}E(e_{T-i}e_{T-j}) = (\sum_{j=0}^{T-1}\alpha^{2j})V_e^2 = \frac{1-\alpha^{2T}}{1-\alpha^2}V_e^2$$

The second equality sign holds due to the white noise property. This equation defines for fixed α the variance V_e^2. Note that for T large, the variance of e_t will be close to $(1-\alpha^2)V_T^2$ and thus will not tend to zero, as is the case for the random lines approach. Analogously to the random lines approach, the variance of the logarithm of the population size can be calculated. Again, the variance is totally due to the sum of the stochastic parts of r_t. Thus the variance can be written as:

$$Var(\log(P_t)) = Var(\sum_{j=1}^{t} r_j^s) = Var(\sum_{j=1}^{t}\sum_{i=0}^{j-1}\alpha^i e_{j-i}) = Var(\sum_{s=1}^{t} e_s \sum_{j=0}^{t-s}\alpha^j) = V_e^2 \sum_{s=1}^{t}(\frac{1-\alpha^{t-s+1}}{1-\alpha})^2$$

Inserting the formula obtained for the noise variance V_e^2 into this expression gives:

$$Var(\log(P_t)) = \frac{2\alpha^{1+t} - \alpha^{2(t+1)} + 2\alpha^{2+t} - 2\alpha + t - \alpha^2 t - \alpha^2}{(1-\alpha^2)(1-\alpha^{2T})}V_T^2$$

Note that this expression for $t=T$ and large T is close to $T\ V_T^2/(1-\alpha^{2T})$ and thus grows only linearly in the projection horizon.

Using this framework, we are able to state the following theorem, which compares the two variances for the logarithms of the population size obtained above.

1.3. Theorem 1

For the logarithm of the population size the following statements are true:

- For $t=1$ the variance using the AR(1) approach is not smaller than the variance corresponding to the random lines approach, irrespective of the value of $|\alpha|<1$. If $T>1$, strict inequality holds.
- For $t=T$, there exists a T_0, such that the variance corresponding to the random lines approach is greater than the variance corresponding to the AR(1) approach for $T>T_0$.
- For $0<\alpha<1$ there exists at most one value t, where the variances of both approaches are equal.

Proof: At $t=1$, the variance of the logarithm of the population size corresponding to the random lines approach is easily seen to be equal to V_T^2/T^2. For the AR(1) approach, the variance of the logarithm is equal to $V_e^2=V_T^2\,(1-\alpha^2)/(1-\alpha^{2T})$, the variance of the white noise sequence. For $T=1$ we have equality, and also for $\alpha = 0$. The variance from the AR(1) approach is strictly greater for $T>1$. Next note, that $1-\alpha^{2T} = (1+\alpha^2+^4+...+^{2(T-1)})(1-\alpha^2)$ and thus the variance of the AR(1) approach is monotonically decreasing with increasing $|\alpha|$. For $|\alpha| \rightarrow 1$ the variance tends to $1/T$, as can be seen by de l'Hopitals rule. Since for $T>1$, $1/T > 1/T^2$ holds, the first part of the theorem is proven.

To show the second part, note that for T tending to infinity, the variance of the logarithm of the population size at $t=T$ corresponding to the random lines approach increases quadratically in T, whereas the variance corresponding to the AR(1) approach only increases linearly. Thus from a certain T_0 onwards, the variance corresponding to the random lines approach will be greater.

In order to show the last point, we first assume without restriction of generality that $V_T^2 =1$. Thus the variance corresponding to the random lines approach will be equal to $t^2(t+1)^2/4T^2$, which clearly is convex in t. For the AR(1) approach the numerator of the expression for the variance is equal to $(2\alpha^{1+t}-\alpha^{2(t+1)}+2\alpha^{2+t}-2\alpha+t-\alpha^2 t-\alpha^2)$, which will be shown to be non-negative and convex in t. First we will prove the convexity. The derivative of this expression with respect to t is equal to $1-\alpha^2 + 2\alpha^{1+t} \log \alpha-2\alpha^{2(t+1)} \log \alpha+2\alpha^{2+t} \log \alpha$. This is easily seen to be positive for $t=0$. Its derivative with respect to t is equal to $2\,(\log \alpha)^2\,(\alpha^{1+t} + \alpha^{2+t} - 2\alpha^{2(t+1)})$, which is positive due to $|\alpha|<1$. This shows convexity. Non-negativity for general t follows from non-negativity for $t=1$ and positivity of the derivative with respect to t.

Thus we have proven that both variances are convex functions, where at $t=1$ the variance corresponding to the AR(1) approach is bigger, whereas for T large enough, the variance corresponding to the random lines approach is bigger. Straightforward but cumbersome evaluations show that for any intersection point the first derivative of $t^2(t+1)^2/(4T^2)$ is greater than the maximal derivative corresponding to the AR(1) approach for any $0<\alpha <1$. This completes the proof.

Note that the theorem uses only second order properties and is not confined to any assumption on the distribution of r_T^s, except for the existence of second order moments. Also it is robust with respect to α, i.e., it holds for any value of α. However, the point where the two variances intersect depends on the actual value of the autoregressive parameter. In Figure 3, the difference of the variance obtained from the random lines approach minus the variance obtained by the AR(1) approach for V_T^2 =1 is plotted for various values of *T*. Figure 4 plots the difference of both variances for *T*=60 and typical values of α. It can be seen, that with increasing α the intersection point also increases.

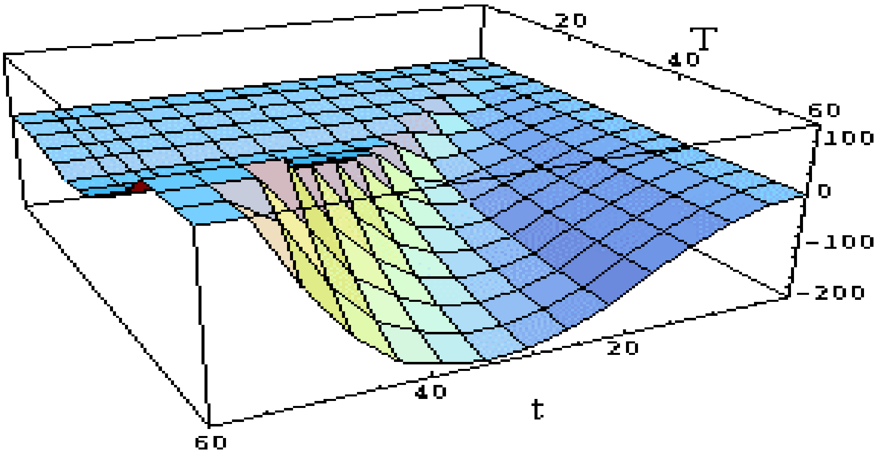

Figure 3. Difference of the variance of random lines approach and AR(1) approach. α=0.85,0≤t≤T≤60.

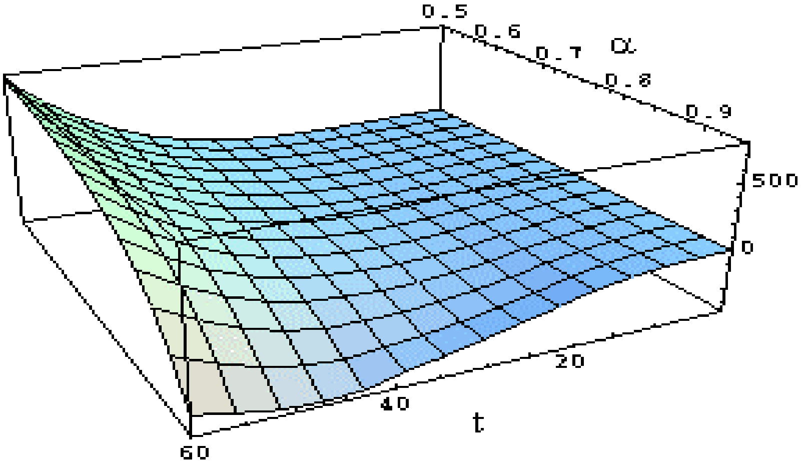

Figure 4. The variance of the random lines approach - variance of the AR(1) approach. *T=60,0≤t≤60,0.5≤α≤1.*

In the last step, we will use these results to obtain results for the distribution of the population size rather than its logarithm. In order to make the analysis easy, it will be assumed that the distribution of r_T^s is normal, which then implies that the distribution of the logarithms of the population size will be normal. However, it is obvious how to generalize these results to other classes of distributions. In this case, the smaller variances translate to smaller symmetric simulated confidence regions for *(P_t)*. Here with simulated confidence regions, we denote the interval which is obtained from the simulated distributions of the population sizes at different times *t* by the interval $[t_a, t_{1-a}]$, where t_x denotes the simulated *x*-quantile. Since the logarithm is a monotonic function, the results of Theorem 1 also hold for the population sizes.

1.4. Theorem 2

The simulated confidence regions for the population size for the AR(1) approach for $t=1$ are larger than the simulated confidence regions for the random lines approach. There exists a T_0, such that for $T > T_0$, the simulated confidence regions are larger for the random lines approach. There exists at most one intersection point where the simulated confidence regions have the same length.

Thus we have found that for the first couple of decades the AR(1) approach leads to a higher variability in the projections, whereas in the long run the converse is true. Higher variability has to be understood in the sense that the confidence regions obtained from the simulations using one approach include the corresponding intervals obtained from the other approach. Therefore in every statement like “the population size will be between p_{lo} and p_{hi} in the year T with at least 95% confidence” drawn from the random lines approach would also be supported from a corresponding AR(1) approach (where the probability would be estimated to be larger), whereas this is not true for the first periods.

2. The Multivariate Case

In this section, we will try to generalize the findings of the last section to a more general model including age structure. Denote the age structure of the population at time *t* with $P_{t,i}$, where *t* refers to the considered period and the index *i* indicates the different classes. These quantities are the components of the vector of the age structure. We will assume that the age structure evolves correspondingly to the following model.

$$P_{t+1} = \begin{bmatrix} 0 & & & f(t) & & \\ m_1(t) & 0 & & & & \\ & m_2(t) & & & & \\ & & \cdot & & & \\ & & & m_f(t) & & \\ & & & & \cdot & \\ & & & & & .\;\; 0 \end{bmatrix} P_t$$

Here $m_i(t)$ denotes the mortality rate and $f(t)$ denotes the fertility rate. The most restricting feature of this model lies in the fact that we assume only one fertile age group. Denoting the big matrix with $L(t)$, it is easy to see that the elements $P_{t,i}$ are just products of the mortality and fertility rates times the initial condition. These products can be transformed into sums in analogy to the univariate model by taking logarithms.

We will investigate two different situations: a) the distributions of the various rates are assumed to be uncorrelated; and b) the rates are assumed to be perfectly correlated with not necessarily equal variances. Note that by doing so, we restrict the experts’ guess. In principle, the experts could be free to choose the complete distribution of the final rates by, e.g., specifying the logarithms of the rates to be normally distributed with arbitrary variance-covariance matrix, which would require (number of classes)^2 guesses for the variance. We reduced this degree of freedom in two ways: The assumption of independent models for the rates is equivalent to the covariance matrix being diagonal, whereas the other assumption is equivalent to the idea that there is just one variable that drives the shifts in the rates. It is straightforward to generalize the analysis to the case where there are several groups of rates where the different groups are described by uncorrelated processes, where each group is driven by just one random variable.

Let $F(t)$ denote the vector $(m_1(t),m_2(t),...,m_{j-1}(t),f(t))$ and let $lF(t) = log(F(t))$ denote the vector of logarithms of the various rates. The population projection is now performed using the same general scheme as in the univariate case.

1. Experts are asked for their guesses of three values for every component of $F(t)$ for some projection horizon T: the estimated mean plus two values specifying the uncertainty of the guess, e.g., the 5% and the 95% quantiles.
2. The stochastic process $(lF(t)),\ 1 \leq t \leq T$ is modeled by one of the following two approaches.
3. A number of realizations corresponding to the stochastic process derived in step 2 are generated.
4. The population size is calculated for each of the realizations of step 3.
5. The distribution of the population projection is evaluated and quantiles can be plotted.

The following analysis will describe the properties in the multivariate setting of the two approaches used in the univariate case in more detail.

2.1. The Random Lines Approach

The generalization of the random lines approach to the multivariate case is straightforward. The vector is $lF(t)=t/T\ lF(T) + (T-t)/T\ lF(0)$, where $lF(0)$ denotes the observed value at time t=0 and $lF(T)=lF(T)^s + lF(T)^d$. Here $lF(T)^s$ denotes the random variable, specified by the experts, and $lF(T)^d$ denotes the mean, which is also given by the experts. In both cases, the variance of the vector $lF(t)$ can easily be seen to be equal to $t^2/T^2\ V_T^2$. Here V_T^2 denotes the covariance matrix of $lF(T)^s$. The quantity of interest is the age structure. For notational simplicity we will only give the argument for i=0 and $t=fk$ for some integer k. In this case, $P_{0,fk} = \prod_{j=0}^{k-1}\left(f(t-jf-1)\prod_{i=1}^{f} m_i(t-jf-i-1)\right)$.

Thus the logarithm is equal to

$$\log P_{0,fk} = \sum_{j=0}^{k-1} \log f(t-jf-1) + \sum_{i=1}^{f}\sum_{j=0}^{k-1} \log m_i(t-jf-i-1) .$$

From this equation, we observe, that the essential terms are the same as in the univariate case, namely sums of components of the random processes. Therefore, we conclude that the basic structure of the univariate and the multivariate problems are quite the same.

2.2. The AR(1) Approach

For the AR(1) approach first note that we have to deal with a more complicated model structure, since the autoregressive coefficient is no longer a scalar, but a matrix: *lF(t) = AlF(t-1)+ E(t)*, where *E E(t) = 0 and E* $E(t)E(t)^T = V_e^2$ is a *j* times *j* matrix. From the white noise properties of *E(t)*, i.e., $E\ E(t)E(s)^T = 0,\ s \neq t$, we obtain the variance V_T^2 of *lF(t)* as $V_T^2 = \sum_{j=0}^{T-1} A^j V_e^2 (A^j)'$. In order to solve this equation with respect to the entries in V_e^2, we have to impose some constraints on the various matrices. One possible approach would be to restrict V_T^2 and *A* to be diagonal: This means, that all processes are scalar AR(1) processes which are uncorrelated. In this case it is straightforward, how to apply the results of the last section to obtain the relevant variances. Another possibility is to restrict *A* to be proportional to the identity matrix, i.e., $A= \alpha I$ and V_T^2 to be arbitrary. In this case, the sum reduces to a scalar sum times V_e^2. In this case we obtain $V_e^2 = V_T^2\ (1-\alpha^2)/(1-\alpha^{2T})$.

Using the discussion above, we are able to state the main theorem of this section.

2.3. Theorem 3

Using either the assumption of uncorrelated mortality and fertility rates or the assumption of perfectly correlated rates, for the logarithm of the vector of the age structure the following statements are true:

- For *t=1* the variance using the AR(1) approach is not smaller than the variance corresponding to the random lines approach, irrespective of the value of $|\alpha|<1$. If *T*>1, strict inequality holds.
- For *t=T*, there exists a T_0, such that the variance corresponding to the random lines approach is greater than the variance corresponding to the AR(1) approach for $T>T_0$.
- If the distribution of *lF(T)* is normal, then the simulated symmetric confidence regions for the random lines approach will be smaller for *t*=1 than the simulated symmetric confidence regions obtained from the AR(1) approach. There exists a T_0 such that for $T>T_0$, the simulated symmetric confidence regions for the random lines approach will be larger.

Proof: The proof of the first point is obvious from the result for the univariate case. The proof of the second point uses the representation of the logarithm of the age structure as given by:

$$\log P_{0,fk} = \sum_{j=0}^{k-1} \log f(t-jf-1) + \sum_{i=1}^{f} \sum_{j=0}^{k-1} \log m_i(t-jf-i-1)$$

Each of the sums is similar to the sum used in the univariate case. The time index is shifted *f* periods instead of 1 period. However, this leads only to a different interpretation: For the random lines approach, the only difference lies in the fact, that the time horizon *T* has to be adjusted to *T/f* and in the AR(1) approach the autoregressive coefficient α is replaced by α^f , the variance is equal to $V_e^2 \sum_{j=0}^{f-1} \alpha^{2j}$. However, this does not change the fact, that for the random lines approach the variance of the sum tends essentially quadratically to infinity as *T* goes to infinity, whereas for the AR(1) approach the variance essentially only grows linearly. Thus, for each of the sums in the above equation we have obtained the required result. For the case that the rates are modeled independently, we have shown the second point. In the case of correlated rates, note that we assumed that there is only one driving variable. Thus, we may rewrite the above equation as a constant times a sum of the driving variable. Since the variance of the driving variable will be larger for the random lines approach for *T* tending to infinity due to the result for the univariate case, the second point is also proven in this case.

Finally, as in the univariate case, it is possible to translate these results to results concerning the simulated symmetric confidence regions, provided the distribution of *lF(T)* is assumed to be normal. This completes the proof.

3. A Simulation Study

In this section the results obtained in Section 1 will be demonstrated using a small simulation study. The data is taken from Lutz and Scherbov (1997). Recall the model of Section 2: The total population size (in the present case of Austria) is modeled as $P_t = P_{t-1}e^{r_t}$. As a starting point the population size in 1995 was chosen. The expert choice in our setup is given by the projections of Lutz and Scherbov (1997): The logarithm of their mean scenario for 2050 was linearily interpolated, which led to a constant annual growth rate of approximately r^d=-4.2325*10^{-4}. This completed the specification of the deterministic part r_t^d. The random part r_t^s was generated with either the random lines approach or an AR(1) model, where the variance V_T^2 was selected, so that (somewhat arbitrarily) the results of the AR(1) approach matched the data at T=2050. This led to a standard deviation of approx. V_T =0.004. Figure 5 visualizes the resulting distributions of the population size. It can be seen clearly, that the random lines approach leads to small variability in the first couple of decades, whereas for T=2050 the random lines approach clearly leads to larger confidence regions.

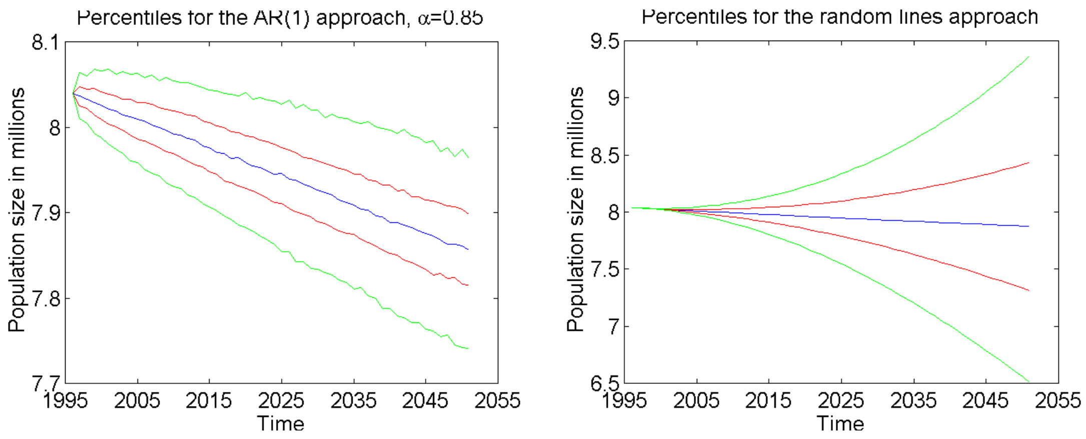

Figure 5. These two plots show the simulated 5%, 25%, 50%, 75%, 95% percentiles respectively. The left plot shows the results for the AR(1) approach using α=0.85. The right plot shows the results for the random lines approach.

This can also be seen from Figure 6, which documents the sample mean and standard deviation of the population sizes obtained from the two approaches. All graphs are drawn on the basis of 1000 simulated trajectories. Note, that the estimated mean is not equal for the two approaches, since only the mean of the logarithm is identical. Also note, that the variance plots show the approx. linear increase of the variance in the AR(1) approach, whereas the corresponding plot for the random lines approach shows the higher curvature. Clearly the results given in Theorem 1 can be verified in these plots. Also note the dependence of the intersection point on the value of α: The higher the value of α, the earlier the two curves intersect.

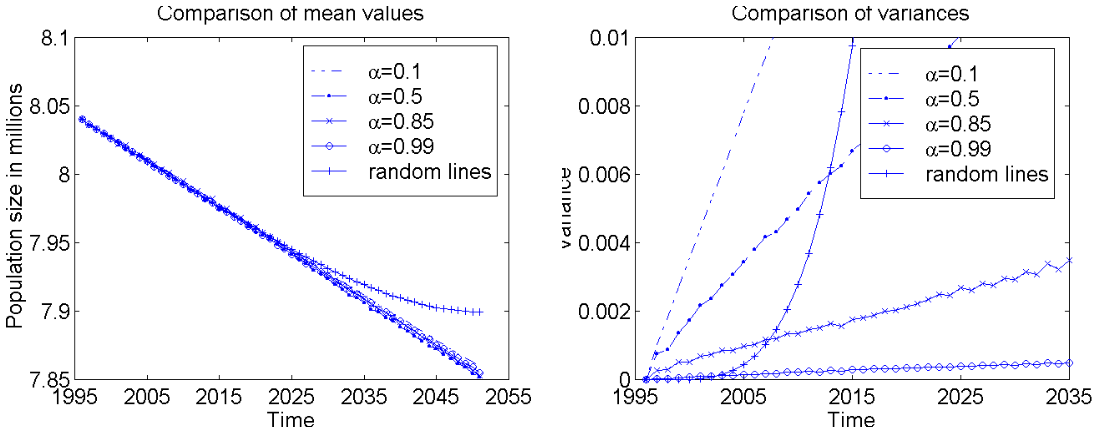

Figure 6. Sample mean and sample standard deviation of the population size. The left picture shows a comparison of the sample mean of the AR(1) approach for various values of α and the random lines approach. The right picture shows the comparison corresponding to the standard deviation.

4. Concluding Remarks

The aim of this study was to compare the variance of the autoregressive approach (AR(1)) with those of the random lines (RL) method. In Theorem 1 we have shown that only for small time periods does the variance of the AR(1) approach exceed the variance corresponding to the RL approach. For sufficiently large time periods, however, the variance of the RL approach is always greater than the variance of the AR(1) method.

It should be noted that this result proves that the conjecture that the RL approach leads to a systematic underestimation of the variance of the future population size is not correct. As is exhibited in Theorem 3, this result remains essentially valid for an age-structured model with only one fertile age group. While these results can be derived *analytically*, we see no possibility to deal with the general case of an age-structured Leslie model in a similar way. The reason for that is the difficulty to derive an expression for the variance of the sum of the products resulting from multiplying the first row of a (general) Leslie matrix with the age-structured stock vector. Further work would be necessary to derive similar results. Simulations recently carried out by Lutz and Scherbov (1997), using Austrian data for single year age groups to study the same research question, give very similar results of higher variances for using the RL approach. We, therefore, have good reason to assume that the properties proven here also hold for the multivariate case with more than one fertile age group.

References

Ahlburg, D.A. and K.C. Land. 1992. Population forecasting: Guest editors' introduction. *International Journal of Forecasting* 8(3): 289-299.

Alho, J.M. 1990. Stochastic methods in population forecasting. *International Journal of Forecasting* 6: 521-530.

Keyfitz, Nathan. 1981. The limits of population forecasting. *Population and Development Review* 7(4): 579-593.

Lee, Ronald. 1999. Probabilistic approaches to population forecasting. Pages 156-190 in Wolfgang Lutz, James W. Vaupel, and Dennis A. Ahlburg, eds. *Frontiers of Population Forecasting.* A Supplement to Vol. 24, 1998, *Population Development Review*. New York: Population Council.

Lee, R.D. and S. Tuljapurkar. 1994. Stochastic population forecasts for the United States: Beyond high, medium, and low. *Journal of the American Statistical Association* 89(428): 1175-1189.

Lutz, Wolfgang, Warren Sanderson, and Sergei Scherbov. 1997. Doubling of world population unlikely. *Nature* 387: 803-805.

Lutz, Wolfgang, Warren Sanderson, and Sergei Scherbov. 1996. Probabilistic population projections based on expert opinion. Pages 397-428 in Wolfgang Lutz, ed. *The Future Population of the World. What Can We Assume Today?* Revised Edition. London: Earthscan.

Lutz, Wolfgang and Sergei Scherbov. 1997. Sensitivity Analysis of Expert-Based Probabilistic Population Projections in the Case of Austria. Interim Report IR-97-48. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Pflaumer, P. 1992. Forecasting US population totals with the US Box-Jenkins approach. *International Journal of Forecasting* 8(3): 329-338.

Stoto, M. 1983. The accuracy of population projections. *Journal of the American Statistical Association* 78(381): 13-20.