Liet. matem. rink., T. 41, spec. nr., 538–541

Population projection: a parametric approach

Remigijus LAPINSKAS, Ramunė VERIKAITĖ (VU)
e-mail: remigijus.lapinskas@maf.vu.lt

Population projection is one of the central issues in demography. It is clear that the development of a (female) population is defined by the number of (female) newborns (which is described by a period age-specific fertility rate $f(x,t)$, $x = 15,\ldots,49$ – by definition, this is an average number of girl children born by one x-years old female during the tth year), by the mortality of its members (which is described by the survivorship function $l(x,t)$, $x = 0,1,\ldots,90+$ – by definition, this is a probability that a (female) newborn of a tth year synthetic cohort will survive till x years), and by migration (we do not investigate its effects here).

Let us start with the age-specific fertility curve. Many papers are published (see, for example, [1, 2, 3]), where one or another parametrization of the curve is proposed. We were looking for a simple and "natural" parametrization, i.e., such that its coefficients "well" (for example, linearly) depend on the total fertility rate *TFR*; here $TFR = TFR(t) = \sum_{x=15}^{49} f(x,t)$. Our initial intention was to separately parameterize each fertility function $f_k(x,t)$, $(k = 1,2,3,4+)$ and then to sum these expressions. However, it seems that approximation errors pool up and the sum presents a poor fit to $f(x,t)$ [5]. On the other hand, despite the fact that the *TFR* constantly diminishes, its ratios with TFR_k remain almost invariable (see Fig. 2). Therefore, we interprete $f(x,t)$ as a sum with a time-invariant structure and look for its direct approximation. Among many variants, one of the most succesful was the following four-parameters approximation: the

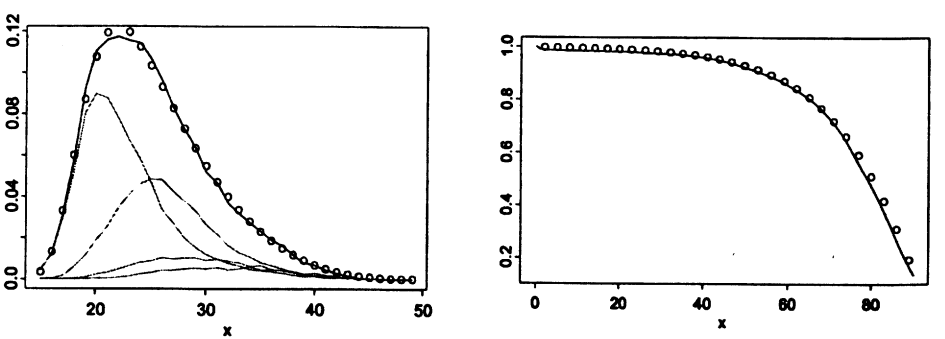

Fig. 1. The Lithuanian total age-specific fertility function $f(x,1995)$ together with those for the 1*st*, 2*nd*, 3*rd* and 4+*th* children $f_k(x,1995)$, $k = 1,2,3,4+$ (left) and the Lithuanian female survivorship function $l(x,1995)$ (right).

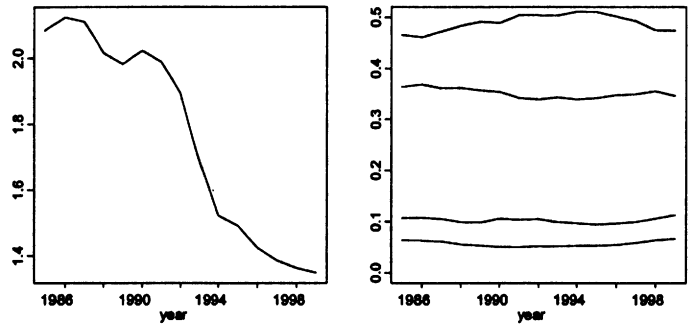

Fig. 2. *TFR* (left) and the ratios $TFR_k/TFR(k = 1, 2, 3, 4+)$ (right).

age vector $x = (15, 16, \ldots, 49)$ was transformed to $xx = (1/36, 2/36, \ldots, 35/36)$ and the fertility function expressed as

$$f(xx, t) = a(1 - xx)^b \exp(-c(1 - xx)^d). \qquad (*)$$

The four unknown coefficients, a, b, c, and d, were found by the use of the method of nonlinear regression. For example, for the year 1995, we have $a = 326.3590, b = 3.0299$, $c = 6.5915$, and $d = 13.5486$, the true *TFR* being 1.491 and the one estimated from the fit 1.493 (for the fit itself, see the left-hand side of Fig. 1, where it is denoted by circles). What is important, the dependence of these coefficients on *TFR* is almost linear (see Fig. 3 based on the 1993–1999 data).

The linear regression procedure gives the following formulas: $a = -128.4717 + 0.3013 \cdot TFR, b = 1.0136 + 0.0013 \cdot TFR$ etc (see [5]).

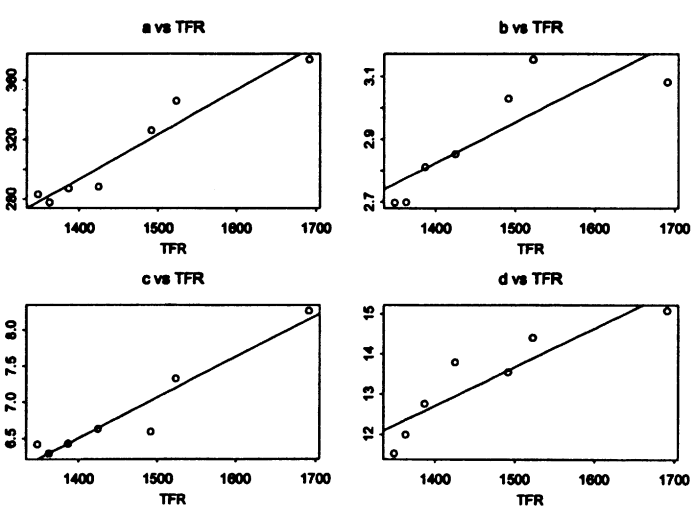

Fig. 3. Nonlinear regression (*): dependence of the coefficients a, b, c, and d on *TFR*.

When using the cohort-component method for population forecast, one has to know the fertility curves for coming years. We have already proved that it is enough to know only the future values of *TFR*, which can be obtained, for example, by using expert estimations or fitting a, say, ARIMA model to the time series $x_t = TFR(t)$, $t = 1960, \ldots, 1999$. We found that this time series is quite satisfactory described by ARIMA(1, 1, 0) model, specifically, if the symbol $\mathrm{d}x_t$ is used to denote the differenced process $x_t - x_{t-1}$, then $\mathrm{d}x_t - \overline{\mathrm{d}x_t} = 0.4468(\mathrm{d}x_{t-1} - \overline{\mathrm{d}x_t}) + \varepsilon_t$, where $\overline{\mathrm{d}x_t} = -0.032$ is the mean of $\mathrm{d}x_t$ and ε_t is the white noise with standard deviation 0.0576. We can still simplify x_t as

$$x_t = -0.018 + 1.4469x_{t-1} - 0.4469x_{t-2} + \varepsilon_t. \qquad (**)$$

Fig. 4 reveals gloomy prospects: if $TFR(t)$ follows this trend, Lithuania will be doomed to extinction.

Now we shall discuss a parameterization and forecast for the survivorship function $l(x,t)$. Traditionally, these attempts are connected with the names of B. Gompertz, W. K. Makeham or W. Brass [1]. In [4], even 45 variants of approximation were proposed with the number of parameters fluctuating from two to six. We shall examine one more possibility. The differenced survivorship function $d\tilde{l}(x,t) = \tilde{l}(x,t) - \tilde{l}(x-1,t)$ (here $\tilde{l}(x,t) = 1 - l(x,t)$) for sufficiently big values of x is quite similar to the gamma density function (see Fig. 5; it is well known that the mortality of newborn children follows another pattern), therefore we shall approximate the function $l(x,t)$ itself by a simple two-parameter gamma cumulative distribution function: $l(x,t) = \frac{1}{\Gamma(r)}\Gamma_{s(93,5-x)}(r)$ (here $\Gamma_z(r) = \int_0^z t^{r-1}e^{-t}dt$ is the incomplete gamma function). We found the two unknown parameters, r and s, by the methods of nonlinear regression. For example, for the year 1995, $r = 0.0734$ and $s = 1.3513$, the true life expectancy $e_0 = 75.2355$, the one given by the fit equals 75.4963, whereas the fit itself is presented on the right side of Fig. 1

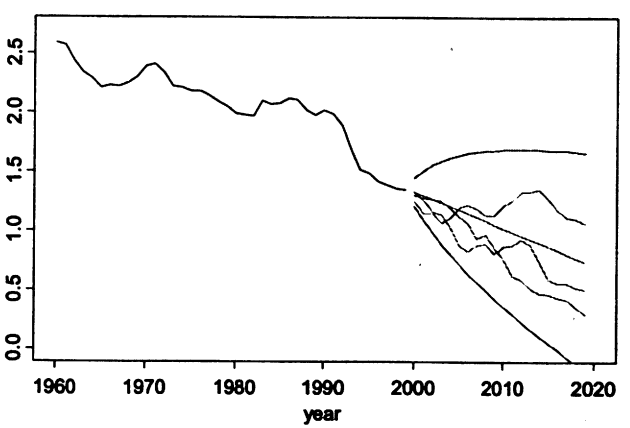

Fig. 4. Time series x_t, $t = 1960, \ldots, 1999$ and three possible scenarios of its development for the years 2000–2020 together with the mean of (**) and its 95% confidence interval.

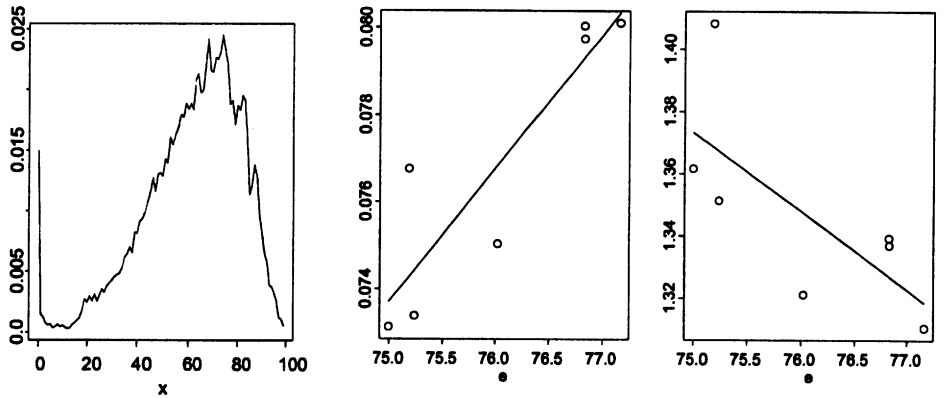

Fig. 5. Function $\widetilde{dl}(x, 1995)$ (left), the regression lines for $r : r = -0.1556 + 0.0031e_0$ (middle) and for $s : s = 3.2668 - 0.0252e_0$ (right).

(by circles). What is important, the dependence of r and s on e_0 (according to the World Bank forecast, e_0 in Lithuania ought to increase by one year every 10 years) is almost linear (see two right-most plots in Fig. 5; our procedure was based on the data of the period 1993 to 1999). The coefficients r and s were estimated by the formulas of linear regression: $r = -0.1556 + 0.0031e_0$, $s = 3.2668 - 0.0252e_0$.

References

[1] C. Newell, *Methods and Models in Demography*, Guilford Press, N.Y. (1988).

[2] P.A. Thompson, W.R. Bell, J.F. Long, R.B. Miller, Multivariate time series projections of parameterized age-specific fertility rates, *Journal of the American Statistical Association*, **84**(407), 689–699 (1989).

[3] A.E. Raftery, S.M. Lewis, A. Aghajanian, M.J. Kahn, `http://www.stat.washington.edu/www/research/online/1994/evenhist.ps` (1994).

[4] R. Senkuvienė, Fitting parametrical models to Lithuanian cause-specific mortality data, *Liet. matem. rink.*, **40** (spec.nr.), 486–490 (2000).

[5] R. Verikaitė, *Some Forecasting Methods in Demography* (Master Thesis), Faculty of Mathematics and Informatics, Vilnius University (2001).

Demografinė prognozė: parametrinis metodas

R. Lapinskas, R. Verikaitė

Populiacijos prognozei naudojant kohortos komponenčių metodą, būtina žinoti ateities vaisingumo ir mirtingumo funkcijų išraiškas. Šiame darbe pasiūlytos kelios šių funkcijų parametrinės formulės ir ištirta jas nusakančių koeficientų priklausomybė nuo suminio vaisingumo ir, atitinkamai, nuo prognozuojamos gyvenimo trukmės. Taip pat sudarytas Lietuvos suminio vaisingumo 1960–1999 metų ARIMA (1, 1, 0) modelis.