A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 255, 2011

Tomasz Jurkiewicz[*]

MULTIDIMENSIONAL SMOOTHING IN TABLES OF FERTILITY RATES

Abstract. One of the basic measures of women's fertility is a partial coefficient of the general fertility of women having *p*-th baby. Estimation of fertility rates can be done on the basis of official records. However, migration processes cause that obtaining corresponding estimates for regional data tends to be more difficult. Estimation of fertility rates is carried out for a large number of dimensions. As a result, a national sample which consists of several hundred women provides relatively small domains of women born in a particular year and living in a given province. The distribution of fertility in a cohort of women over time tends to be highly positively skewed with smooth shape. A sample survey will cause disturbances in the obtained distribution because of the sampling process. For this reason, the need for smoothing the distribution arises. The purpose of this study is to test possibilities of using some smoothing methods and to evaluate the impact of smoothing treatment on the efficiency of estimation.

Key words: small domain estimation, fertility rates, multidimensional smoothing.

I. INTRODUCTION

Poland possesses poor statistical data related to women's fertility in the period between the First and Second World Wars. Also, for the long period after 1945 statistical files of marriages, live births and similar variables were insufficient for performing age, period and cohort analyses. First opportunity of reconstructing fertility rates in real cohorts was created by a Women Fertility Survey, a part of the 1970 National Census. Unfortunately, computational difficulties and methodological mistakes which were committed caused that Polish demographers did not exploit all possibilities in order to evaluate cohort fertility for Polish regions, see J. Paradysz (1990, p. 26–30). Those times principles and methods of small area estimation were not well known yet. Two further Fertility Surveys, however, based on smaller samples were conducted in years 1988 and 2002, as parts of the National Polish Censuses. The first of these surveys was elaborated in the CSO by J. Paradysz (1992) but only on the

[*] PhD, Department of Statistics, University of Gdańsk, jurkiewicz@wzr.ug.edu.pl

national level. Nowadays, progress in development of small area methodology and in computational processing of data, supported by growing experience of statisticians make it possible to create cohort fertility rates for particular regions.

One of the basic statistical measure of women's fertility is age-specific fertility rates (second category) according to the number of children. The formula is given by:

$$F(x;p) = \frac{U(x;p)}{K(x)} \tag{1}$$

where $U(x;p)$ is a number of children of *p-th* order, born by mother aged x and $K(x)$ is a number of women aged x at 30 June. Fertility rates tend to be presented as a sequence of tables for the following years of birth (see Table 1). They can also be calculated and presented for particular provinces.

Table 1. A typical table of fertility rates for the first birth given by mothers (an example)

Mother year of birth (Cohort)	Age of mother (in years)					
	...	21	22	23	24	...
...	...	...	...	...	...	...
1970	...	$F_{1970}(21;1)$	$F_{1970}(22;1)$	$F_{1970}(23;1)$	$F_{1970}(23;1)$	...
1971	...	$F_{1971}(21;1)$	$F_{1971}(22;1)$	$F_{1971}(23;1)$	$F_{1971}(23;1)$	...
1972	...	$F_{1972}(21;1)$	$F_{1972}(22;1)$	$F_{1972}(23;1)$	$F_{1972}(23;1)$	...
...	...	...	...	...	...	...

From 1974 estimation of fertility rates in Poland can be carried out on the basis of official population registers. Data about the number of live births are regarded to be reliable and precise due to obligatory registration of infant births. A more complex problem is the evaluation of the number of women in particular age intervals for given regions, because of migration movements within the population. Also a reconstruction of fertility rates for the past periods can only be made owing to the surveys which accompanied the national censuses. One of the practical difficulties in random sample surveys consists in constructing fertility rates which tend to be multidimensional, presenting very detailed description of the examined sample. It results in small frequencies of women born in a given year and living in a particular area, even if the sample size reaches several thousands. The problem of estimating fertility rates tables is one of the issues of small-domain statistics. However, unlike the conventional methods it is estimated the entire array of parameters. The author presents the

view that applications of smoothing methods for small domain estimation should be regarded as model supported approach rather than model-based approach.

The distribution of women's cohort fertility rates by time is positive skew and has usually smooth shape. There are still some natural disturbances, which are a derivatives of e.g. war periods, crises or changes in government's policy. Similarly the distributions in other dimensions have rather smooth shape. Distributions obtained in a sampling survey are often disturbed due to sampling mechanism. For this reason use of smoothing methods to minimize the influence of sampling error is necessary. The purpose of this paper is to verify possibilities of applying smoothing methods to smooth fertility tables and to evaluate the influence of smoothing on effectiveness of estimation.

II. METHODS OF MULTIDIMENSIONAL SMOOTHING

There are many methods for one-dimensional smoothing. Yet because age-specific fertility rates are displayed in tables with many dimensions, a multidimensional smoothing is to applied. Methods based on fitting theoretical distributions are not good for this case due to appearance of natural disturbances in fertility tables.

One of the possible solution is to apply generalized linear array model (GLAM) for multidimensional smoothing, see Currie et al. (2006). This method is an extension of the B-spline[1] approach for smoothing data to multidimensional grids with roughness penalties (P-spline). The optimum smoothness in GLAM is attained by minimizing weighted sum of squares of two elements. The first element is, analogous to OLS method, sum of squares of differences between original and smoothed values. The second element is the sum of squares of d-order differences between regression coefficients for splines. The straightforward application of the penalized Fisher scoring algorithm to solve this equation cause computational difficulties. Therefore, author applies a novel algorithm proposed by Eilers et al. (2006).

Additionally, another method of iterative smoothing of fertility tables (OPT) was used. Its main idea is to minimize a given loss function. In this paper a weighted sum of three components was used as the loss function. The first component involves the mean value of squared differences between the values prior to smoothing and after smoothing. The second component is the average difference between smoothed fertility rates in succeeding (neighbouring) cohorts

[1] B-spline method is based on replacement the real values of the independent variable X with values defined by splines, i.e. properly joined polynomials. Smoothed values are estimated as theoretical values from regression model, which splines appear as independent variable in, see Eilers and Marx (1996).

(i.e. the mean difference between the succeeding rows of the table). The third component consists of the similar average difference of fertility rates, however, in this case between the succeeding columns, rather than rows. Weights for all three components were set arbitrary in the applied loss function.

As a result of GLAM and OPT, the array with smoothed data is obtained. This method still doesn't provide compatibility weighted sum of conditional distributions with marginal distributions. An additional correction of the results is needed. In this paper a method of raking, known from its applications in analysis of sample data is used. The essence of this method involves repeated alternating multiplication of particular elements of rows and columns by proper correction coefficients.

III. DATA SOURCE AND PROCEDURE OF THE MONTE CARLO ANALYSIS

Efficiency of the applied small area techniques was evaluated on the basis of statistical data representing the results of a sample survey which accompanied the 2002 National Census. Due to numerical complexity of smoothing methods the analysis covered only a part of the population: 10 cohorts of women born in the period 1926–1935 who gave birth of their first baby at the age ranging from 17 to 36 years. The size of the table was 10x20.

The actual table of fertility rates estimated upon the sample survey was not smooth due to random effects. Therefore, the purpose of the first step was to obtain smoothed tables of fertility rates. It was achieved by applying GLAM method, as the first variant, and OPT method as the second one. The resulted tables determined the assumed distribution of the population. Population related to the first variant (popul_1) was better smoothed than the population corresponding to variant 2 (popul_2).

In each of 1000 simulation experiments a simple random sample of size n was generated. Four different sample sizes were applied: 5.000, 10.000, 20.000, and 50.000. The smallest size reflects the typical sample size for a large Poland's province, whereas the largest one corresponds to the size of the investigated cohorts in the survey from year 2002.

On the basis of a generated sample a table of fertility rates was estimated. Then the table was subject to smoothing by GLAM and OPT methods and to correcting by "raking" method.

For evaluating the efficiency of the smoothing a relative measure based on the square root of the mean square error (RRMSE) was applied. Additionally, the magnitude value of the relative bias (ARB) was used:

$$RRMSE_d = \frac{\sqrt{\frac{1}{1000}\sum_{i=1}^{1000}\left(\widetilde{F}_{i,d} - F_d\right)^2}}{F_d} \cdot 100 \tag{2}$$

$$ARB_d = \left| \frac{\frac{1}{1000}\sum_{i=1}^{1000}\left(\widetilde{F}_{i,d} - F_d\right)}{F_d} \right| \cdot 100 \tag{3}$$

where: d indicates the cell of the table, F_d stands for the actual fertility coefficient in the population, and $\widetilde{F}_{i,d}$ denotes the smoothed coefficient in ith iteration.

IV. RESULTS OF THE STUDY

Figure 1 shows one of possible distributions of fertility rates obtained in a single iteration of direct estimation without smoothing (DT), with smoothing by GLAM (GLAM), with smoothing by GLAM and corrected by raking, and similarly by OPT (respectively).

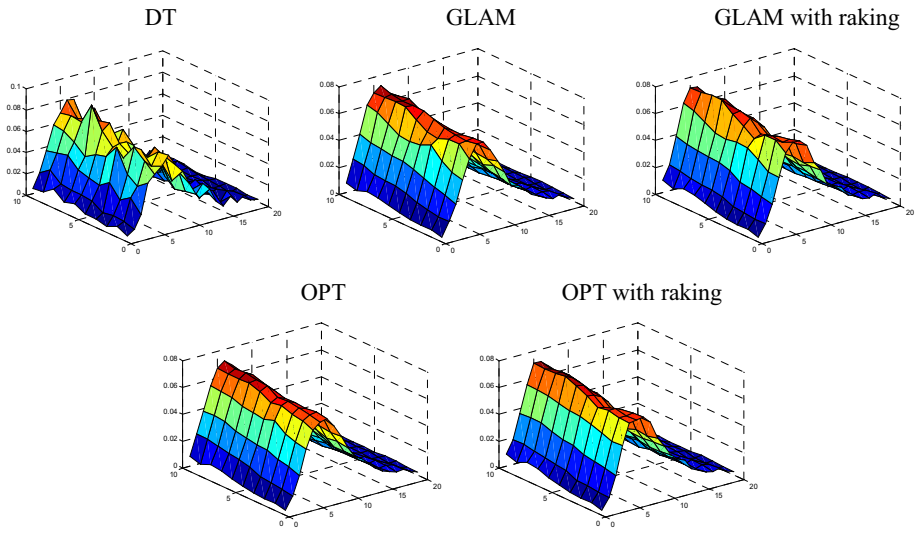

Figure 1. The distribution of fertility rates (an example)

Source: own study.

The large number of estimated coefficients (200 in each iteration) caused that the results of efficiency measures are presented as summary statistics, rather than individual characteristics. Table 2 presents some major quantiles and other characteristics in the distribution of RRMSE for the sample size 10.000, corresponding to popul_1.

Table 2. Summary statistics of the distribution of RRMSE for particular methods of smoothing and the sample size n = 10.000

Statistic	GLAM	GLAM with raking	OPT	OPT with raking	DT
min	4,04	5,63	3,32	4,45	10,43
1st deciles	5,07	6,28	4,20	5,53	12,36
1st quartile	6,32	7,56	5,55	7,03	14,62
median	9,80	11,30	8,91	10,70	23,01
3rd quartile	16,73	19,15	17,34	19,73	39,34
9th deciles	25,55	28,02	26,93	28,68	49,01
max	94,85	94,92	87,20	86,40	104,55
mean	13,58	15,23	13,30	14,97	27,89
std. dev.	12,27	12,19	12,15	12,00	16,37
skew	3,55	3,30	3,01	2,74	1,44

Source: own study.

One can easily conclude that smoothing contributes essentially to the precision of estimation, the errors tend to be reduced by half for each applied smoothing method. OPT seems to yield smaller errors compared with GLAM. The procedure of correction by raking led to a decrease in a relative bias, however it slightly deteriorated the efficiency of estimation measured by RRMSE. There are reasons to assume that if data on marginal distributions were more reliable (e.g. obtained from official registers), the procedure of correction could even improve the estimation results.

Similar results have been obtained for the other, less smoothed population. The level of RRMSE for non- smoothed estimates was nearly the same as for the first population. However, the application of smoothing resulted in a sharp decrease of the error, though not as large as for the former population.

In Table 3 the mean and the median for RRMSE corresponding to four sample sizes drawn from population popul_1 are given. For the direct estimation the error is proportional to the reciprocal of the square root of the sample size. An application of smoothing methods changes this relation, and cause that a decrease in the sample size results in an increase in the efficiency of estimation compared with the direct estimation. For 50.000 sample units the reduction of

the error owing to an application of smoothing methods reaches 30%, but for 5.000 sample units it attains as much as 50%.

It is also worth noting that an increase in error due to the "raking" correction applied is about several per cent for large samples, and a dozen for small samples. Similar results have been obtained for the latter, less smoothed population.

Table 3. Impact of the sample size on RRMSE for particular methods of smoothing

Sample size	GLAM	GLAM with raking	OPT	OPT with raking	DT
the mean					
5.000	17,98	20,61	17,40	20,07	39,48
10.000	13,58	15,23	13,30	14,97	27,89
20.000	10,63	11,48	10,68	11,55	19,67
50.000	8,28	8,43	8,65	8,84	12,43
the median					
5.000	13,19	15,77	11,34	14,36	31,65
10.000	9,80	11,30	8,91	10,70	23,01
20.000	7,11	8,23	7,21	8,15	15,73
50.000	5,15	5,62	5,76	5,98	10,02

Source: own study.

V. CONCLUSIONS

An application of multivariate smoothing techniques results in an essential reduction of the error of estimation, especially for small sample sizes. This approach seems to be one of reasonable alternatives to the problem of estimation fertility rates, especially in spatial dimensions. Author's own research (presented on SAE-2009 Conference in Elche) focused on applications of small domain estimators confirms the gain in efficiency of estimation in this approach. Smoothing by GLAM is slightly less efficient than by OPT, however less numerically complicated. GLAM does not require an arbitrary setting of weights for the loss function. OPT, on the other hand, allows for some additional criteria and some prior information to be incorporated, e.g. information about natural disturbances in women's fecundity. Both methods enable the researcher to carry out the smoothing in more than two dimensions. For instance, the smoothing may take into account information about similarities among the considered regions or provinces.

BIBLIOGRAPHY

Currie, I.D., Durban, M., Eilers, P.H.C. (2006), *Generalized linear array models with applications to multidimensional smoothing*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), Vol. 68, No. 2, April (259–280).

Eilers P.H.C., Curie I.D., Durban M. (2006), *Fast and compact smoothing on large multidimensional grids*, Computational Statistics & Data Analysis, 50 (61–76).

Eilers P.H.C., Marx B.D. (1996), *Flexible smoothing with B-splines and penalties*, Statistical Science, Vol. 11, No. 2, (89–121).

Paradysz J. (1990), *Women Fertility In Poland: Methodological And Cognitive Study*. SGPiS, Warsaw. In Polish.

Paradysz J. (1992), *Women Fertility In Poland*. Central Statistical Office, Warsaw. In Polish. Wieczorkowski R., Zieliński R. (1997) *Random number computer generators*, WNT, Warsaw. In Polish.

Tomasz Jurkiewicz

WIELOWYMIAROWE WYGŁADZANIE TABLIC WSPÓŁCZYNNIKÓW PŁODNOŚCI

Jednym z podstawowych mierników płodności kobiet jest cząstkowy współczynnik (drugiej kategorii) płodności ogólnej kobiet rodzących dziecko kolejności *p*. Szacowanie współczynników płodności możliwe jest na podstawie bieżącej rejestracji ruchu naturalnego ludności. Większym problemem, z uwagi na ruchy wędrówkowe ludności, jest jednak uzyskanie oszacowań dla danych regionalnych oraz dla kohort, które nie były objęte jeszcze rejestracją bieżącą. Estymacja współczynników płodności odbywa się dla wielu szczegółowych przekrojów. Sprawia to, że nawet przy kilkusettysięcznej ogólnopolskiej próbie współczynnik wyznaczany jest dla stosunkowo niewielkiej domeny kobiet urodzonych w danym roku i zamieszkujących w określonym województwie. Rozkład płodności kohorty kobiet w czasie jest rozkładem o silnej prawostronnej asymetrii i o zazwyczaj gładkim przebiegu. W wyniku badania reprezentacyjnego uzyskany rozkład będzie jednak zaburzany na skutek losowego mechanizmu doboru jednostek. Tym samym pojawia się konieczność wygładzenia rozkładu tak, aby zniwelować wpływ składnika losowego w ostatecznym obrazie rozkładu płodności. Celem pracy jest weryfikacja możliwości zastosowania metod wygładzania dla tablic płodności i ocena wpływu wygładzania na efektywność estymacji.