ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 216, 2008

*Krystyna Pruska**

SIMILARITIES AMONG SMALL AREA RELATIVE FREQUENCY DISTRIBUTIONS IN SMALL AREA SYNTHETIC ESTIMATION

ABSTRACT. In the paper the estimation precision for the small area mean is considered for six synthetic estimators. For three estimators auxiliary data dealing with the whole population are used and for other estimators – data dealing with some groups of similar small areas. The group of similar small areas is determined on the basis of relative frequency distribution for the ratio of the considered variable and auxiliary variable for stratified sampling design without replacement. The obtained results are compared with the results of analogous experiments in which the group of similar small areas is determined with usage of the notation of series.

Key words: small area estimation, similarity of small areas, simulation experiments.

I. INTRODUCTION

In small area statistics information about the whole population, or parts of it, is used for estimating subpopulation characteristics. An application of auxiliary variables in construction of small area estimators can lead to better estimates.

In case of small area estimation the choice of auxiliary variables and auxiliary data can be connected with the choice of a group of small areas which are similar to one considered small area.

This study was undertaken in order to propose the measures of similarity between small areas and to compare a precision of estimates in synthetic estimation using proposed measures.

II. SYNTHETIC ESTIMATORS OF SMALL AREA PARAMETERS

In small area statistics there are different possibilities of using variables and data auxiliary (see Cz. Bracha (1996), J. Paradysz (1998), J. Kordos (1999), Cz. Domański,

* Ph.D., Associate Professor, Chair of Statistical Methods, University of Łódź.

K. Pruska (2001), J. N. K. Rao (2003), E. Gołata (2004)). We can consider synthetic estimators. These estimators are constructed with an assumption of a similarity between the distinguished small area and the whole population, or its part. We assume that the distributions of the considered variables, or some relations between them, in small area and in the whole population, or its part, are similar.

We assume that the considered population is divided into G strata. In the population H small areas are distinguished. Let Y and X denote considered and auxiliary variables in the population and small areas. Let $\overline{Y}_{h.}$ be the mean of Y for h-th small area where $h = 1,..., H$.

We take into account six estimators of parameter $\overline{Y}_{h.}$ for $h = 1,..., H$:

$$T_1^{(h)} = \frac{1}{N_{h.}} \sum_{g=1}^{G} Y_{.g}^{*} \frac{X_{hg}}{X_{.g}^{*}}, \quad (1)$$

$$T_2^{(h)} = \frac{1}{N_{h.}} X_{h.} \beta^{*}, \quad (2)$$

$$T_3^{(h)} = \frac{1}{N_{h.}} [y_{h.} + (X_{h.} - x_{h.}) \beta^{*}], \quad (3)$$

$$T_4^{(h)} = \frac{1}{N_{h.}} \sum_{g=1}^{G} Y_{U_h g}^{*} \frac{X_{hg}}{X_{U_h g}^{*}}, \quad (4)$$

$$T_5^{(h)} = \frac{1}{N_{h.}} X_{h.} \beta_{U_h}^{*}, \quad (5)$$

$$T_6^{(h)} = \frac{1}{N_{h.}} [y_{h.} + (X_{h.} - x_{h.}) \beta_{U_h}^{*}], \quad (6)$$

where

$$Y_{.g}^{*} = \frac{N_{.g}}{n_{.g}} \sum_{h=1}^{H} \sum_{i=1}^{n_{hg}} y_{hgi}, \quad (7)$$

$$X_{.g}^{*} = \frac{N_{.g}}{n_{.g}} \sum_{h=1}^{H} \sum_{i=1}^{n_{hg}} x_{hgi}, \quad (8)$$

$$Y_{U_h g}^{*} = \frac{N_{U_h g}}{n_{U_h g}} \sum_{j \in U_h} \sum_{i=1}^{n_{hg}} y_{jgi}, \quad (9)$$

$$X^{*}_{U_h g} = \frac{N_{U_h g}}{n_{U_h g}} \sum_{j \in U_h} \sum_{i=1}^{n_{hg}} x_{jgi}, \tag{10}$$

$N_{h.}$ – number of population elements which belong into h-th small area,

$N_{.g}$ - number of population elements which belong into g-th stratum,

$N_{U_h g}$ – number of population elements which belong into the group U_h of small areas and g-th stratum,

$n_{h.}$ – number of population sample elements which belong into h-th small area,

$n_{.g}$ – number of population sample elements which belong into g-th stratum,

$n_{U_h g}$ – number of population sample elements which belong into the group U_h of small areas and g-th stratum,

X_{hg} – total value of X for h-th small area and g-th stratum,

$X_{h.}$ – total value of X for h-th small area,

$x_{h.}$ – sample total value of X for h-th small area,

x_{hgi} – value of variable X for i-th sample element for h-th small area and g-th stratum,

$y_{h.}$ – sample total value of Y for h-th small area,

y_{hgi} – value of variable Y for i-th sample element for h-th small area and g-th stratum,

β^{*} – regression parameter for linear regression of variable Y with respect to X determined on the basis of population sample,

$\beta^{*}_{U_h}$ – regression parameter for linear regression of variable Y with respect to X determined on the basis of sample for group U_h of similar small areas.

The estimators $T_1^{(h)}$ and $T_4^{(h)}$ are ratio-synthetic estimators and $T_2^{(h)}$, $T_3^{(h)}$, $T_5^{(h)}$ and $T_6^{(h)}$ are regression synthetic estimators (see Cz. Domański, K. Pruska (2001), E. Gołata (2004)). The values of estimators $T_1^{(h)}$, $T_2^{(h)}$, $T_3^{(h)}$ are determined with usage of estimates of totals for the considered and auxiliary variables for population strata. The values of estimators $T_4^{(h)}$, $T_5^{(h)}$ and $T_6^{(h)}$ are calculated analogously but this time for strata of group U_h of similar small areas.

III. SELECTION OF SMALL AREAS SIMILAR TO GIVEN SMALL AREA

In case of synthetic estimation we can consider a problem of determining a group of small areas similar to given small area. The similarity means that assumptions, which are taken in synthetic estimation, are satisfied.

We can apply the cluster analysis for determining similar small areas (see E. Gołata (2004)). In paper written by K. Pruska (2006) there is proposed a method of selection of similar small areas with the usage of a number of series in sequences of ratios of the considered variable values and auxiliary variable values. We determine the sequence of values y_{hgi}/x_{hgi} for two small areas (for example $h = k,m$) and $g = 1,..., G,$ and $i = 1,..., n_{h.}$. Next, we calculate the number of series of ratios of values from k-th small area and m-th small area (the ratios for one small area create one type of symbol and the ratios for another small area create another type of symbol). We divide the obtained value of series, which is denoted by l_{km}, by a total number of observations from k-th and m-th small areas. We take that the measure of similarity between k-th and m-th small areas is of the form:

$$s_{km} = l_{km}/(n_{k.} + n_{m.}) \text{ for } k \neq m \tag{11}$$

$$s_{km} = 1 \text{ for } k = m. \tag{12}$$

Values s_{km} belong to interval (0; 1]. If value s_{km} is near to 1 then k-th and m-th small areas are similar. We determine a group of small areas similar to the considered small area choosing the subpopulations for which the value of similarity measure is near to 1. Information about this group can be used for calculating the values of estimators: $T_4^{(h)}$, $T_5^{(h)}$, $T_6^{(h)}$.

Another approach to determining similar small areas is to use a measure of similarity between relative frequency distributions for the ratio of considered variable and auxiliary variable for two subpopulations. Let us denote the measure by W_{ij} for i-th and j-th small areas. We determine the small areas for which value of measure W_{ij} for fixed i and $j = 1,...,H$ is greater than the assumed value which is denoted by w_0 (for example $w_0 = 0.5$). Next, we include them in the group of small areas similar to fixed i-th small area. The measure W_{ij} is of the form:

$$W_{ij} = \sum_{l=1}^{k_{ij}} \min\{w_{il}, w_{jl}\}, \tag{13}$$

where k_{ij} is the number of classes to which the sum of sets of values of the ratio of the considered variable and auxiliary variable for i-th and j-th small areas is divided, and w_{il}, w_{jl} are proportions for l-th class ($w_{il} = n_{il}/n_{i.}$, $w_{jl} = n_{jl}/n_{j.}$ where n_{il}, n_{jl} are the numbers of values of the considered ratios which belong into l-th class).

IV. MONTE CARLO ANALYSIS OF PRECISION OF ESTIMATES IN SYNTHETIC ESTIMATION

The conducted Monte Carlo analysis deals with the precision of mean estimates for small area which are obtained on the basis of estimators: $T_1^{(h)}$, $T_2^{(h)}$, $T_3^{(h)}$, $T_4^{(h)}$, $T_5^{(h)}$, $T_6^{(h)}$.

In order to conduct the Monte Carlo experiments some populations were created. Each population consisted of 50 000 pairs of numbers and was divided into G strata and H small areas where $G = 5$ and $H = 10$. Each stratum consisted of 10 000 pairs of numbers and each small area consisted of 5 000 pairs of numbers.

Elements of each population were determined in the following way:

– for each small area (it means for h-th small area, $h = 1,...,H$) 5 000 numbers were generated according to fixed distribution (see Table 1.); the numbers are denoted by c_{hj}, $h = 1,...,10$, $j = 1,...,5\,000$;

– on the basis of numbers c_{hj}, $h=1,...,10$, $j=1,...,5\,000$, there were determined the numbers x_{hgi} according to the formula: $x_{hgi} = c_{h,i+(g-1)1000} + \varepsilon_{hgi}$ where $h=1,...,10$, $g=1,...,5$, $i=1,...,1000$ and ε_{hgi} for $g=1,...,5$, $i=1,...,1000$ were generated according to normal distribution $N(a*g, 1/g)$ where $a = 0.1$ or $a = 0.5$ (see Table 1);

– on the basis of numbers x_{hgi} there were determined numbers y_{hgi} according to formula $y_{hgi} = f(x_{hgi}) + \xi_{hgi}$ where ξ_{hgi} for $h=1,...,10$, $g=1,...,5$, $i=1,...,1000$ were generated according to normal distribution $N(0; b)$where $b = 0.1$ or $b = 2$ and the forms of function f are presented in Table1.;

– there were determined the pairs of numbers (y_{hgi}, x_{hgi}) where $h=1,...,10$, $g=1,...,5$ and $i=1,...,1000$; these pairs created the population for which the Monte Carlo experiments were conducted.

In this way 9 populations were created and the means for them were determined.

Next, 500-element sample was drawn from each population. The stratified sampling design without replacement was applied and 100 elements (the pairs of numbers) were drawn from each stratum. The sample for small area was created from these elements of population sample which belong to the small area.

On the basis of the obtained samples there were determined the values of estimators: $T_1^{(h)}$, $T_2^{(h)}$, $T_3^{(h)}$, $T_4^{(h)}$, $T_5^{(h)}$, $T_6^{(h)}$ for small areas with numbers $h = 1, 7, 10$, and the error of estimation was calculated for them. In case of estimators $T_4^{(h)}$, $T_5^{(h)}$, $T_6^{(h)}$ the group U_h consisted of the small areas for which the measure of similarity (13) was greater than fixed value w_0. There were considered the following variants: $w_0 = 0.25$; 0.5; 0.75. We also investigated the case in which the group U_h consisted of 5 most similar small areas according to measure (11)–(12).

Table 1

Population variants in simulation experiments

Denotation of population variant[a)]	Distribution[b)] for generating of c_{hj} for h = 1,10; j = 1,...,5000	Relation between values y_{hgi} and x_{hgi} for g = 1,...,5; i = 1,...,1000
CHI02 (I, II, III)	χ^2_{h+9}	$y_{hgi} = 1.2\, x_{hgi} + \xi_{hgi}$ for h = 1, 2 $y_{hgi} = 1.4\, x_{hgi} + \xi_{hgi}$ for h = 3,...,10
CHI05 (I, II, III)		$y_{hgi} = 1.2\, x_{hgi} + \xi_{hgi}$ for h = 1,...,5 $y_{hgi} = 1.4\, x_{hgi} + \xi_{hgi}$ for h = 6,...,10
CHI10 (I, II, III)		$y_{hgi} = 1.2\, x_{hgi} + \xi_{hgi}$ for h = 1,...,10

[a)]Variant I: values ε_{hgi} were generated according to normal distribution $N(0.1g;1/g)$ and values ξ_{hgi} - to normal distribution $N(0;0.1)$) for h=1,...,10; g=1,...,5; i = 1,...,1000; variant II: values ε_{hgi} were generated according to normal distribution $N(0.1g;1/g)$ and values ξ_{hgi} - to normal distribution $N(0;2.0)$) for h=1,...,10; g=1,...,5; i = 1,...,1000; variant III: values ε_{hgi} were generated according to normal distribution $N(0.5\, g;1/g)$ and values ξ_{hgi} - to normal distribution $N(0;2.0)$) for h=1,...,10; g=1,...,5; i=1,...,1000.

[b)] Symbol χ^2_k denotes the chi-square distribution with k degrees of freedom and symbol $N(\mu,\sigma)$ – normal distribution with expectation μ and standard deviation σ.

Source: own assumptions.

The samples were drawn 1000 times for each populations and the error of estimation for each estimator was determined according to formula:

$$RMSE_k^{(h)} = \frac{\sqrt{\frac{1}{1000}\sum_{i=1}^{1000}(T_{ki}^{(h)} - \overline{Y}_{h.})^2}}{\overline{Y}_{h.}}, \tag{14}$$

where $T_{ki}^{(h)}$ denotes value of estimator $T_k^{(h)}$ for i-th repetition, i=1,...,1000, h = 1, 7, 10, k=1,...,6.

The results of calculations are presented in Tables 2-4. We can see that the errors of estimation (RMSE) are smaller (for the same type of estimators) in the majority of cases for four variants of determining the group of similar small areas U_h in comparison with estimation using information about the whole population if we use information about the group in case different relations of considered variables in small areas. The method of using the measure of similarity (11)-(12) to determine the group U_h , consisting of 5 similar small areas from among 10 small areas, seems to lead to more precise estimates in synthetic estimation than the other considered methods. In the conducted analysis the asymmetric distributions of variable Y and X were considered.

Table 2

RMSE in experiments for estimators $T_1^{(h)}$, $T_2^{(h)}$, $T_3^{(h)}$

Population variant		Number of small area (h)	Estimator $T_1^{(h)}$	Estimator $T_2^{(h)}$	Estimator $T_3^{(h)}$
CHI02	I	1	0.1424	0.1863	0.0115
		7	0.0209	0.0173	0.0010
		10	0.0211	0.0172	0.0009
	II	1	0.1443	0.1906	0.0250
		7	0.0182	0.0246	0.0125
		10	0.0223	0.0208	0.0109
	III	1	0.1434	0.1936	0.0224
		7	0.0190	0.0270	0.0116
		10	0.0227	0.0233	0.0097
CHI05	I	1	0.0975	0.1767	0.0109
		7	0.0594	0.0119	0.0008
		10	0.0596	0.0117	0.0007
	II	1	0.0994	0.1811	0.0247
		7	0.0565	0.0189	0.0125
		10	0.0605	0.0161	0.0103
	III	1	0.0981	0.1859	0.0223
		7	0.0576	0.0221	0.0116
		10	0.0613	0.0190	0.0097
CHI10	I	1	0.0003	0.0006	0.0011
		7	0.0003	0.0007	0.0007
		10	0.0003	0.0006	0.0006
	II	1	0.0052	0.0128	0.0218
		7	0.0064	0.0138	0.0146
		10	0.0051	0.0123	0.0120
	III	1	0.0048	0.0126	0.0196
		7	0.0060	0.0135	0.0136
		10	0.0047	0.0122	0.0113

Source: own calculations.

Table 3

RMSE in experiments for measure of similarity (11)-(12) and card $U_h = 5$ or for measure of similarity (13) and $w_0 = 0.25$

Population variant		Number of small area (h)	Estimator					
			$T_4^{(h)}$	$T_5^{(h)}$	$T_6^{(h)}$	$T_4^{(h)}$	$T_5^{(h)}$	$T_6^{(h)}$
			measure of similarity (11)-(12) and card $U_h = 5$			measure of similarity (13) and $w_0 = 0.25$		
CHI02	I	1	0.1093	0.1848	0.0113	0.0008	0.0018	0.0011
		7	0.0003	0.0008	0.0006	0.0003	0.0007	0.0006
		10	0.0003	0.0007	0.0005	0.0002	0.0006	0.0005
	II	1	0.1096	0.1755	0.0245	0.1442	0.1905	0.0249
		7	0.0072	0.0170	0.0125	0.0182	0.0246	0.0125
		10	0.0058	0.0149	0.0103	0.0222	0.0208	0.0103
	III	1	0.1083	0.1813	0.0220	0.1432	0.1935	0.0224
		7	0.0066	0.0162	0.0116	0.0189	0.0269	0.0116
		10	0.0055	0.0149	0.0097	0.0226	0.0230	0.0097
CHI05	I	1	0.0004	0.0010	0.0011	0.0004	0.0010	0.0011
		7	0.0003	0.0008	0.0006	0.0003	0.0008	0.0006
		10	0.0003	0.0008	0.0005	0.0003	0.0008	0.0005
	II	1	0.0063	0.0313	0.0222	0.0993	0.1809	0.0247
		7	0.0033	0.0080	0.0125	0.0564	0.0190	0.0125
		10	0.0018	0.0070	0.0103	0.0600	0.0163	0.0103
	III	1	0.0177	0.0435	0.0158	0.0978	0.1855	0.0223
		7	0.0093	0.0181	0.0091	0.0574	0.0224	0.0116
		10	0.0069	0.0157	0.0078	0.0604	0.0196	0.0097
CHI10	I	1	0.0004	0.0010	0.0011	0.0003	0.0006	0.0011
		7	0.0004	0.0010	0.0007	0.0003	0.0007	0.0007
		10	0.0003	0.0009	0.0006	0.0003	0.0006	0.0006
	II	1	0.0085	0.0194	0.0218	0.0052	0.0128	0.0218
		7	0.0079	0.0195	0.0146	0.0064	0.0138	0.0146
		10	0.0069	0.0173	0.0120	0.0051	0.0123	0.0120
	III	1	0.0076	0.0189	0.0195	0.0048	0.0126	0.0196
		7	0.0076	0.0184	0.0135	0.0060	0.0135	0.0136
		10	0.0065	0.0173	0.0113	0.0047	0.0122	0.0113

Source: own calculations.

Table 4

RMSE in experiments for measure of similarity (13) and $w_0 = 0.5$ or $w_0 = 0.75$

Population variant		Number of small area (h)	Estimator					
			$T_4^{(h)}$	$T_5^{(h)}$	$T_6^{(h)}$	$T_4^{(h)}$	$T_5^{(h)}$	$T_6^{(h)}$
			$w_0 = 0.5$			$w_0 = 0.75$		
CHI02	I	1	0.0008	0.0018	0.0011	0.0009	0.0021	0.0011
		7	0.0003	0.0007	0.0006	0.0003	0.0008	0.0006
		10	0.0002	0.0006	0.0005	0.0003	0.0008	0.0005
	II	1	0.1168	0.1701	0.0243	0.1442	0.1905	0.0220
		7	0.0123	0.0198	0.0125	0.0182	0.0246	0.0125
		10	0.0108	0.0145	0.0109	0.0222	0.0208	0.0103
	III	1	0.1027	0.1597	0.0215	0.0243	0.0488	0.0197
		7	0.0105	0.0189	0.0116	0.0062	0.0157	0.0116
		10	0.0087	0.0138	0.0097	0.0058	0.0159	0.0097
CHI05	I	1	0.0004	0.0010	0.0011	0.0006	0.0013	0.0011
		7	0.0003	0.0008	0.0006	0.0004	0.0010	0.0006
		10	0.0003	0.0008	0.0005	0.0003	0.0010	0.0005
	II	1	0.0644	0.1261	0.0233	0.0147	0.0297	0.0218
		7	0.0332	0.0250	0.0125	0.0076	0.0199	0.0125
		10	0.0234	0.0207	0.0103	0.0069	0.0191	0.0103
	III	1	0.0530	0.1093	0.0207	0.0154	0.0332	0.0195
		7	0.0263	0.0262	0.0116	0.0076	0.0200	0.0116
		10	0.0172	0.0197	0.0097	0.0071	0.0193	0.0097
CHI10	I	1	0.0003	0.0006	0.0011	0.0005	0.0010	0.0011
		7	0.0003	0.0007	0.0007	0.0004	0.0008	0.0007
		10	0.0002	0.0006	0.0006	0.0003	0.0008	0.0006
	II	1	0.0052	0.0128	0.0218	0.0102	0.0207	0.0218
		7	0.0064	0.0138	0.0146	0.0072	0.0162	0.0146
		10	0.0051	0.0123	0.0120	0.0070	0.0170	0.0120
	III	1	0.0048	0.0127	0.0196	0.0089	0.0216	0.0196
		7	0.0060	0.0135	0.0136	0.0068	0.0157	0.0136
		10	0.0047	0.0122	0.0113	0.0064	0.0167	0.0113

Source: own calculations.

V. CONCLUSIONS

The results of the conducted Monte Carlo experiments allow us to conclude that the use of information about a group of small areas similar to the distinguished small area according to the proposed measures of similarity, can lead to better estimates of mean than the using of information about the whole population. Similar conclusions arise from other analyses conducted by the author and dealing with synthetic estimation.

REFERENCES

Bracha Cz. (1996), *Teoretyczne podstawy metody reprezentacyjnej*, Wydawnictwo Naukowe PWN, Warszawa.

Domański Cz., Pruska K. (2001), *Metody statystyki małych obszarów*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

Gołata E. (2004), *Estymacja pośrednia bezrobocia na lokalnym rynku pracy*, Wydawnictwo Akademii Ekonomicznej w Poznaniu, Poznań.

Kordos J. (1999), Problemy estymacji danych dla małych obszarów, *Wiadomości Statystyczne* 1, 85–101.

Paradysz J. (1998), Small Area Statistics in Poland. First Experiences and Application Possibilities, *Statistics in Transition*, Vol.3, No. 5, 1003–1015.

Pruska K. (2006), Dobór danych pomocniczych w badaniach małych obszarów, *Wiadomości Statystyczne* 7 i 8, 23–34.

Rao J. N. K. (2003), *Small Area Estimation*, John Wiley & Sons, New Jersey.

Krystyna Pruska

PODOBIEŃSTWO STRUKTUR MAŁYCH OBSZARÓW W ESTYMACJI SYNTETYCZNEJ DLA MAŁYCH OBSZARÓW

W pracy rozpatrywana jest dokładność estymacji średniej dla małego obszaru w przypadku zastosowania sześciu estymatorów syntetycznych. Do wyznaczenia wartości trzech spośród nich wykorzystane są dane pomocnicze dotyczące całej populacji, a w pozostałych przypadkach – dane dotyczące grupy małych obszarów podobnych do rozpatrywanego. Grupę małych obszarów podobnych do danego wyznaczono, wykorzystując wskaźnik podobieństwa struktur odpowiadający ilorazowi zmiennej badanej i zmiennej pomocniczej w małym obszarze. Następnie przeprowadzono analizę Monte Carlo, w której dokonano porównania dokładności oszacowań średnich dla małych obszarów dla rozpatrywanych estymatorów w przypadku warstwowego losowania zależnego. Porównano też wyniki z wynikami analogicznego badania, w którym grupę podobnych małych obszarów wyznaczono, wykorzystując pojęcie serii.