2014-3

A New Method of Estimating Small Area Demographics Using Population Potential

Takashi Inoue

April 2014

Working Paper Series, Institute of Economic Research
Aoyama Gakuin University

A New Method of Estimating Small Area Demographics Using Population Potential

Takashi Inoue

Department of Public and Regional Economics

College of Economics

Aoyama Gakuin University

4-4-25 Shibuya, Shibuya-ku

Tokyo, 150-8366 JAPAN

Abstract

The purpose of this paper is to propose a new method of estimating small area demographics using population potential developed by Stewart (1947). The new method has the following features different from previous methods: first, it sufficiently utilizes position data; secondly, it does not require specific distribution for original data; thirdly, new estimators obtained through it can easily be calculated. This paper presented eight formulas to calculate new estimators and applied two formulas of them to actual data. The result of application shows that the new method is effective in smoothing similarly to a previous method.

1. Introduction

Since small area demographics are generally unstable, some methods of estimating true values of such data have been developed chiefly by statisticians and demographers (e.g. Tanba 1988). Most of previous methods basically take a form of smoothing by using data of adjacent small areas. These previous methods, however, have the following several problems: (1) do not sufficiently utilize position or coordinate data of adjacent areas; (2) can be used only under the specific conditions; (3) require complicated calculation procedures. In order to solve these problems, this paper proposes a new method of estimating small area demographics using population potential developed by Stewart (1947).

Population statistics used in previous studies are quotients of two data (mortality rate, elderly population rate, cohort change rate, child-woman ratio, proportion of workers by industry, and so on). For that reason, this paper also handles such quotients. Although small area demographics generally indicate statistics of areas smaller than municipalities, this paper handles not only such statistics but also statistics of areas similar or larger size of municipalities.

This paper presents the simplest previous method in Chapter 2, proposes a new method in Chapter 3, confirms the availability of a new method in Chapter 4 by estimating demographics from actual data using the method, and describes conclusion in Chapter 5.

2. The Simplest Previous Method

The simplest previous method is to calculate an average value of demographics of the neighborhood including an area to be an object of estimating. This method is, as it were, to take a spatial moving average, which corresponds to the simplest method of spatial filtering in image processing.

Let x_i be demographics of object area *i*, estimator $\dot{x}_i$ of x_i is expressed as the following equation:

$$\dot{x}_i = \frac{1}{n_i} \sum_{j \in V_i} x_j \qquad (1)$$

Here V_i is a set of area numbers selected according to a criterion such as to be adjoining to area *i*, to be located within a constant distance from area *i*, and to belong to a municipality common to area *i*, including *i* in any case; n_i is the number of elements of V_i. In a word, V_i means the neighborhood of area *i*.

Figure 1 shows a sample of the relationship between an object area and the neighborhood of it.

0	1	2	3	4	5	6	7	8	9
10	11	12	13	14	15	16	17	18	19
20	21	22	23	24	25	26	27	28	29
30	31	32	33	34	35	36	37	38	39
40	41	42	43	44	45	46	47	48	49
50	51	52	53	54	55	56	57	58	59
60	61	62	63	64	65	66	67	68	69
70	71	72	73	74	75	76	77	78	79
80	81	82	83	84	85	86	87	88	89
90	91	92	93	94	95	96	97	98	99

Figure 1 An Object Area and the Neighborhood of it

In Figure 1, each cell shows a small area; the dark-toned cell indicates an object area; the dark-toned and light-toned cells indicate the neighborhood; each integer means the area number. In this case, $n_i = 9$ and $V_i = \{i-11, i-10, i-9, i-1, i, i+1, i+9, i+10, i+11\}$.

3. Proposal of a New Method

After describing the fundamental idea of a new method in Section 3-1, this chapter introduces the concept of population potential in Section 3-2, which is crucially important in proposing the method. Furthermore this chapter derives basic (in Section 3-3) and applied (in Section 3-4) formulas from the above fundamental idea.

3-1. Fundamental Idea of a New Method

The reason why, when estimating demographics of an object area, we use statistics of its neighborhood is that there is a universal principle, namely, “the demographics approximate those of an area located more closely to the object area.” In addition, if it is the same distance from the object area, the demographics approximate those of a more populated area. A new method proposed in this paper is to estimate the true value from a weighted moving average based on the above two principles.

Consequently the weight becomes larger when closer and more populated, and just matches the concept of population potential. This is because population potential is crucially important in a new method.

3-2. Introduction of Population Potential

Population potential, which is developed by Stewart (1947), means the amount of population energy acting on an area from the neighborhood of it. Let e_{ij} be population energy acting on area i from area j and E_i be population potential of area i,

$$E_i = \sum_j e_{ij} \qquad (2)$$

The variable e_{ij} corresponds to the above-mentioned weight and it is given by equation (3).

$$e_{ij} = 2kp_i\sqrt{\frac{\pi}{a_i}} \quad (\text{if } j = i) \qquad e_{ij} = k\frac{p_j}{d_{ij}} \quad (\text{otherwise}) \quad (3)$$

Here *k* is a constant; p_i, and a_i indicate population and size of area *i*, respectively; d_{ij} denotes the distance between areas *i* and *j*.

3-3. Derivation of Basic Formula

This section derives a basic formula from equation (3). Let x_i be demographics of object area *i* and $\hat{x}_i$ be an estimator of x_i,

$$\hat{x}_i = \frac{\sum_{j\in V_i} e_{ij}x_j}{\sum_{j\in V_i} e_{ij}} = \frac{e_{ii}x_i + \sum_{j\in V_i, j\neq i} e_{ij}x_j}{e_{ii} + \sum_{j\in V_i, j\neq i} e_{ij}} \qquad (4)$$

Substituting equation (3) into equation (4), we obtain

$$\hat{x}_i = \frac{2p_i x_i\sqrt{\frac{\pi}{a_i}} + \sum_{j\in V_i, j\neq i}\frac{p_j x_j}{d_{ij}}}{2p_i\sqrt{\frac{\pi}{a_i}} + \sum_{j\in V_i, j\neq i}\frac{p_j}{d_{ij}}} \qquad (5)$$

Furthermore, assuming that $x_i = q_i / p_i$ (q_i is population of area *i*, differing from p_i , e.g. death population and elderly population), equation (5) leads equation (6).

$$\hat{x}_i = \frac{2q_i\sqrt{\frac{\pi}{a_i}} + \sum_{j\in V_i, j\neq i}\frac{q_j}{d_{ij}}}{2p_i\sqrt{\frac{\pi}{a_i}} + \sum_{j\in V_i, j\neq i}\frac{p_j}{d_{ij}}} \qquad (6)$$

Equation (6) is the basic formula of a new method.

3-4. Derivation of Applied Formulas

The basic formula needs a lot of distance data between all combinations of small areas and requires complicated numerical calculation even if obtaining such distance data. Thus this section derives three applied formulas by modifying the basic formula (equation (6)).

3-4-1. Applied Formula 1

First, we assume that 1) both of areas *i* and V_i take forms of a precise circle, 2) population density shows uniform distribution both in area *i* and in the donut-shaped area which generates by removing area *i* from area V_i (Figure 2).

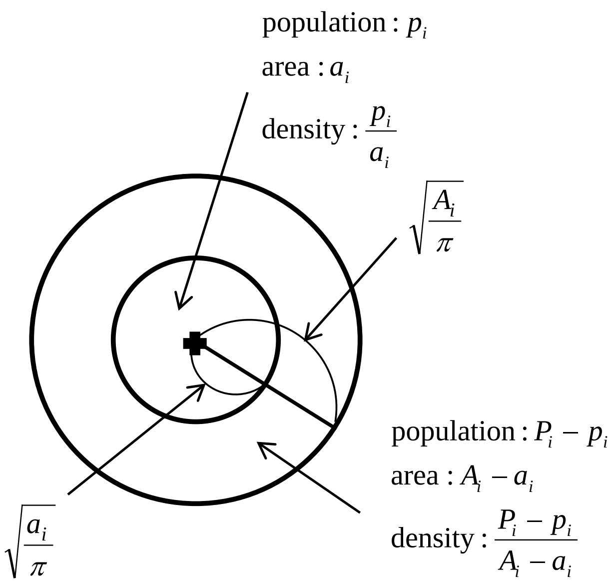

Figure 2 Positional Relationship between Area *i* and the Donut-shaped Area

Secondly, we define several variables on area V_i as the following equations: $P_i = \sum_{j \in V_i} p_j$, $Q_i = \sum_{j \in V_i} q_j$, $A_i = \sum_{j \in V_i} a_j$. At that time, population energy acting on the center of area *i* from area *i* and from the donut-shaped area are written in equations (7) and (8), respectively.

$$k \int_0^{\sqrt{\frac{a_i}{\pi}}} \frac{1}{r} \times \frac{p_i}{a_i} \times 2\pi r dr = 2kp_i \sqrt{\frac{\pi}{a_i}} \qquad (7)$$

$$k \int_{\sqrt{\frac{a_i}{\pi}}}^{\sqrt{\frac{A_i}{\pi}}} \frac{1}{r} \times \frac{P_i - p_i}{A_i - a_i} \times 2\pi r dr = 2k(P_i - p_i) \frac{\sqrt{\pi}}{\sqrt{A_i} + \sqrt{a_i}} \qquad (8)$$

Meanwhile, as demographics x_j of area *j* constituting the donut-shaped area become constant regardless of *j*, expressed as $(Q_i - q_i)/(P_i - p_i)$ (although being dependent on *i*), equation (4) is transformed into

$$\hat{x}_i = \frac{\sum_{j \in V_i} e_{ij} x_j}{\sum_{j \in V_i} e_{ij}} = \frac{e_{ii} \times \frac{q_i}{p_i} + \frac{Q_i - q_i}{P_i - p_i} \times \sum_{j \in V_i, j \neq i} e_{ij}}{e_{ii} + \sum_{j \in V_i, j \neq i} e_{ij}} \qquad (9)$$

Equations (7) and (8) mean the weights of moving averages, corresponding to e_{ii} and $\sum_{j \in V_i, j \neq i} e_{ij}$ in equation (9), respectively.

Therefore we can obtain the following equation:

$$\hat{x}_i = \frac{2kp_i \sqrt{\frac{\pi}{a_i}} \times \frac{q_i}{p_i} + 2k(P_i - p_i) \frac{\sqrt{\pi}}{\sqrt{A_i} + \sqrt{a_i}} \times \frac{Q_i - q_i}{P_i - p_i}}{2kp_i \sqrt{\frac{\pi}{a_i}} + 2k(P_i - p_i) \frac{\sqrt{\pi}}{\sqrt{A_i} + \sqrt{a_i}}} \qquad (10)$$

Equation (10) reduces to the applied formula 1 as follows:

$$\hat{x}_i = \frac{q_i\sqrt{A_i} + Q_i\sqrt{a_i}}{p_i\sqrt{A_i} + P_i\sqrt{a_i}} \qquad (11)$$

3-4-2. Applied Formula 2

The applied formula 1 requires size data of areas *i* and V_i, and those data are not always attached to population statistics. Thus we make an assumption such that population density of the neighborhood (area V_i including area *i*) shows a uniform distribution, that is, $p_i/a_i = P_i/A_i$, and then transpose equation (11) into the applied formula 2a (equation (12)).

$$\hat{x}_i = \frac{q_i\sqrt{P_i} + Q_i\sqrt{p_i}}{p_i\sqrt{P_i} + P_i\sqrt{p_i}} \qquad (12)$$

At that time, the equation (12) leads to the applied formula 2b (equation (13)).

$$\hat{x}_i = \frac{\sqrt{p_i}}{\sqrt{p_i} + \sqrt{P_i}} \cdot \frac{q_i}{p_i} + \frac{\sqrt{P_i}}{\sqrt{p_i} + \sqrt{P_i}} \cdot \frac{Q_i}{P_i} \qquad (13)$$

Because of $\frac{\sqrt{p_i}}{\sqrt{p_i} + \sqrt{P_i}} + \frac{\sqrt{P_i}}{\sqrt{p_i} + \sqrt{P_i}} = 1$, equation (13) means that estimator $\hat{x}_i$ becomes the weighted average of demographics q_i/p_i of area *i* and demographics Q_i/P_i of area V_i.

Furthermore, assuming that V_i is fixed regardless of *i* and expressed as V (e.g. V shows a municipality including all small areas), P_i and

Q_i become constant given by P ($=\sum_{i\in V} p_i$) and Q ($=\sum_{i\in V} q_i$), respectively, and Q/P becomes a weighted average of q_i/p_i. Accordingly, equation (13) is rewritten in the applied formula 2c (equation (14)).

$$\hat{x}_i = \frac{\sqrt{p_i}}{\sqrt{p_i}+\sqrt{P}} \cdot \frac{q_i}{p_i} + \frac{\sqrt{P}}{\sqrt{p_i}+\sqrt{P}} \cdot \frac{Q}{P} \qquad (14)$$

In equation (14), if q_i indicates elderly population of small area i, q_i/p_i and Q/P mean elderly population rates of area i and of the municipality including area i, respectively; estimator $\hat{x}_i$ becomes the weighted average of the two rates; and the weight is given by the ratio of the root of population to the sum of the roots of population. As a result, the applied formula 2c shows a numerical form similar to the Empirical Bayes Estimator (EBE) or Stein-type shrinkage Estimator (SE) (Shinozaki 1991), although the weight is changeable dependent on i, in difference from that of EBE and from that of SE.

3-4-3. Applied Formula 3

Another estimator can be obtained by reciprocal transformation of the estimator of p_i/q_i (= the inverse of q_i/p_i). The estimator of p_i/q_i is calculated similarly to that of p_i/q_i, if not p_i but q_i is considered as population energy. We can easily derive the another estimator from the applied formula 2 (equation (12)), by interchanging p_i and q_i as well as P_i and Q_i, and by taking the inverse of the formula. As a result, the estimator is calculated by the following equation, which gives applied formula 3a.

$$\hat{x}_i = \frac{q_i\sqrt{Q_i} + Q_i\sqrt{q_i}}{p_i\sqrt{Q_i} + P_i\sqrt{q_i}} \qquad (15)$$

Similarly to the applied formula 2a, equation (15) can be transposed into the two applied formulas 3b (equation (16)) and 3c (equation (17)).

$$\hat{x}_i = \frac{p_i\sqrt{Q_i}}{p_i\sqrt{Q_i} + P_i\sqrt{q_i}} \cdot \frac{q_i}{p_i} + \frac{P_i\sqrt{q_i}}{p_i\sqrt{Q_i} + P_i\sqrt{q_i}} \cdot \frac{Q_i}{P_i} \qquad (16)$$

$$\hat{x}_i = \frac{p_i\sqrt{Q}}{p_i\sqrt{Q} + P\sqrt{q_i}} \cdot \frac{q_i}{p_i} + \frac{P\sqrt{q_i}}{p_i\sqrt{Q} + P\sqrt{q_i}} \cdot \frac{Q}{P} \qquad (17)$$

Here as mentioned above, P and Q are given by $\sum_{i \in V} p_i$ and $\sum_{i \in V} q_i$, respectively.

4. Application of a New Method

This chapter examines the availability of a new method proposed in this paper by applying the method and previous one to the same data, and by comparing them to each other. Three kinds of estimators are used for examination. One is the Empirical Bayes Estimator (EBE) based on Poisson-Gamma model, which has been widely adopted in estimation of mortality (Tango 1988); the others are new estimators calculated by the applied formulas 2c and 3c. There are three large differences between EBE and new estimators. First, EBE does not utilize position data at all; nevertheless new estimators utilize such data. Secondly, EBE (in the case of Poisson-Gamma model) requires that population depends on Poisson distribution; nevertheless new estimators do not require such distribution. Thirdly, new estimators can more easily be calculated than EBE.

Data used for examination are SMR (standardized mortality ratio) on female death of gastric cancer in Saitama Prefecture by municipality in 1995-99, calculated by Kubokawa (2013). Table 1 shows the result of computation. In this table, four columns on the left-hand side are Kubokawa’s original data and the other two columns are data added newly to them. Expected death population means the number of dead persons estimated on assumption that the mortality of each municipality equals to that of whole Saitama. SMR is the ratio of actual death population to expected death population and the weighted average of SMR is just 100.0 in theory. All data of three columns (EBE, applied formula 2c, applied formula 3c) indicates estimators of SMR.

According to Table 1, every estimator is approaching more closely to 100.0 than SMR and therefore every type of estimation has a remarkable tendency to smooth the variation of SMR. This suggests that the new method is effective similarly to a previous method (i.e. the empirical Bayes method), at least in smoothing. The intensity of the smoothing, however, is much different among them. Estimation by the applied formula 2c smooths the variation of SMR more intensely than the empirical Bayes method. Estimation by the applied formula 3c smooths it most intensely and reduces the range of SMR to less than 4.0. Such strong smoothing has both an advantage and a disadvantage, although particularly estimation by the applied formula 3c gives us an impression of excessive shrinkage. The advantage is to avoid yielding an extraordinary value especially in the case that an estimator should be iteratively used like a cohort change rate in long-term population projection. The disadvantage is to possibly lose the true variation of original data. As regards such advantage/disadvantage, more mathematical discussion is beyond the scope of a working paper and further consideration will be needed.

Table 1 Female Mortality of Gastric Cancer in Saitama Prefecture by Municipality, 1995-99

municipality	death population	expected death population	SMR	EBE	applied formula 2c	applied formula 3c
Kawagoe	206	192.1	107.2	105	101.3	100.1
Kumagaya	136	102.7	132.4	112	104.5	100.4
Kawaguchi	253	242.8	104.2	104	100.8	100.1
Urawa	244	256.7	95.1	98	99.0	99.9
Oomiya	244	264.8	92.1	97	98.4	99.9
Gyouda	69	61.2	112.7	103	101.4	100.2
Tokorozawa	174	179.6	96.9	100	99.5	100.0
Honjyo	39	45.9	85.0	97	98.5	99.7
Higashimatsuyama	41	52.7	77.8	95	97.7	99.6
Kasukabe	121	105.5	114.7	106	102.1	100.2
Yono	58	48.3	120.1	104	102.0	100.3
Koshigawa	130	153.1	84.9	94	97.5	99.7
Hatogaya	46	35.2	130.7	105	102.6	100.4
Sakado	67	51.6	129.8	107	103.1	100.4
Satte	44	34.9	126.1	104	102.2	100.4
Yoshikawa	18	27.9	64.5	95	97.2	99.3
Ogose	15	10.8	138.9	102	101.9	100.5
Naguri	5	3.5	142.9	100	101.2	100.6
Namekawa	6	8.9	67.4	98	98.5	99.4
Ogawa	19	27	70.4	96	97.7	99.4
Kawashima	21	16.5	127.3	102	101.7	100.4
Yoshimi	28	15.3	183.0	106	104.9	101.0
Nagatoro	10	8	125.0	101	101.1	100.4
Okano	4	11.7	34.2	95	96.6	98.3
Ryoujin	1	3.2	31.3	98	98.1	98.1
Higashichichibu	6	4.4	136.4	100	101.2	100.5
Kamiizumi	5	1.5	333.3	101	104.5	102.0
Kamisato	25	18.3	136.6	103	102.3	100.5
Oosoto	11	6.3	174.6	102	102.9	100.9
Okabe	17	13.9	122.3	101	101.2	100.3
Kawamoto	15	9.1	164.8	103	103.0	100.8
Shiraoka	20	26.1	76.6	97	98.2	99.6
Shoubu	18	15.8	113.9	101	100.8	100.2

source: Kubokawa (2013) (partly revised)

5. Conclusion

This paper attempted to propose a new method of estimating small area demographics using population potential. The new method consists of eight formulas (basic formula and applied formulas 1, 2a, 2b, 2c, 3a, 3b, 3c). This paper applied two of those formulas and the Empirical Bayes method to actual data of SMR in order to compare them and to confirm the availability of the new method. The results of application are summarized as follows:

1) Both types of estimation by the two formulas are effective in smoothing similarly to the previous method;

2) estimation by the applied formula 2c smooths the variation of SMR more intensely than the previous method;

3) estimation by the applied formula 3c smooths it most intensely.

These facts enable us to conclude that the new method has both an advantage and a disadvantage: the former is to avoid yielding an extraordinary value especially if an estimator should be iteratively used; the latter is to possibly lose the true variation of original data. Further studies are needed to evaluate them mathematically.

Acknowledgements

This paper is a result of the study which received a financial support of the Japan Society for the Promotion of Science by grant-in-aid (grant number: 25370919). The outline of this paper was presented in the second eastern Japan regional meeting of Population Association of Japan by the author in 2014. The author deeply appreciates Mr. Masaru Nakada (MRI Research Associates) for technical advice on numerical calculation.

References

Kubokawa, T., “The Benchmark Problem on Small Area Estimation and Error Estimate of Constrained Empirical Bayes Estimation”, *CIRJE Research Report Series*, CIRJE-R-12, pp.143-165.

Shinozaki, N., 1991, “Stein-type Shrinkage Estimators and their Applications”, *Japanese Journal of Applied Statistics*, Vol.20, No.2, pp.59-76.

Stewart, J. Q., 1947, “Empirical Mathematical Rules concerning the Distribution and Equilibrium of Population”, *Geographical Review*, Vol. 37, No. 3, pp.461-485.

Tango, T., 1988, “Empirical Bayes Estimation for Mortality Indices: Applications to Disease Mapping”, *Japanese Journal of Applied Statistics*, Vol.17, No.2, pp.81-96.