Geographical Visualization of Rare Events

Roh, Hye-jung, Jeong, Jae-joon

Department of Geography, Sungshin Women's University.
249-1 Dongseon-dong 3-ga Seongbuk-gu, Seoul, Korea
E- mail: poulpican@hanmail.net, jeongjj@sungshin.ac.kr

ABSTRACT:

Maps contain and effectively visualize a number of spatial information. Advances in GIS enable researchers to analyze and represent spatial information through digital maps. Choropleth maps represent different quantities showing usually rates, percentages or densities. Generally, researchers make choropleth maps using raw rates. But, if the events are rare, raw rates cannot be sufficient in representing spatial phenomena. That is to say, if the population is large and events are rare, we cannot be sure that the raw rate is correct.

The objective of this study is to make choropleth maps by several rate calculation methods and compare them. We use three methods in choropleth mapping ; a raw rate, empirical Bayesian method, and spatial rate method which use prior probabilities. The experiments reveal that maps are somewhat different by used methods. We suggest that a raw rate method can not be an only way to make a rate map and researchers should choose an appropriate method for their objectives.

KEY WORDS: Choropleth map, rare events, geographic visualization, Bayesian method, density map

1. INTRODUCTION

Maps contain and effectively visualize a number of spatial information. Advances in GIS enable researchers to analyze and represent spatial information through digital maps. Choropleth maps are widely used to visualize the distribution of information collected for enumeration (Xiao.N et al. 2007). They represent different quantities showing usually rates, percentages or densities. Although, choropleth maps contain same cartographic data, they are could be different according to the number of class categories and visualization methods (i.e. same interval, standard deviation, same quantities, etc). But these methods are not changing rates but changing visualization methods.

Generally, researchers make choropleth maps using raw rates. When the attribute of interest is rate or proportion, exploratory mapping of the rates for displaying geographical variability is an obvious first step in any analysis. However, using the raw observed rates may be misleading, since the variability of such rates will be a function of the values of the 'population' to which they relate, and this may differ widely from area to area (Bailey and Gatrell, 1995). Also, if the events are rare, raw rates or proportions cannot be sufficient in representing spatial phenomena. That is to say, if the population is large and events are rare, we cannot be sure that the raw rate is correct.

Rare events happen in two cases that either 'population' or 'observation' is small. Rare rates are easily changed in case of rare events. The objective of this study is to make choropleth maps by several rate calculation methods and compare them. We use three methods in choropleth mapping a raw rate, empirical Bayesian method, and spatial rate method which use prior probabilities and compare them.

2. METHODOLOGY

2.1 Raw rate method

Raw rate (or proportions) method is usually used in choropleth mapping. Proportions are ratios of events over a population. With as the counts of events and as the population at risk in area i, the "raw rate" is the simple proportion:

$$r_i = y_i / n_i . \tag{1}$$

2.2 Empirical Bayesian method

Bayesian statistics is concerned with statistical estimation where prior knowledge or beliefs about parameters of interest are taken onto account when estimating their values, as well as observed data. Bayes theorem is used to derive the posterior distribution by combining the likelihood for data with the prior distribution. Our prior knowledge will be represented by considerations based on overall rate across all rate. And then we will use Bayesian techniques to modify what we observe in given area on the basis of this. The best Bayes estimates θ_i based on combining these prior distributions with the observed rates are given as followings:

$$\hat{\theta}_i = w_i r_i + (1 - w)\gamma_i \tag{2}$$

$$\text{where} \quad w_i = \frac{\phi_i}{(\phi_i + \gamma_i / n_i)} \tag{3}$$

$\hat{\theta}_i$ = best Bayes estimates
r_i = observed rate, $r_i = y_i / n_i$
γ_i = mean value
ϕ_i = variance
n_i = population in area i
y_i = observed value

Note that when the population is large, the second term in the denominator of (3) becomes near zero, and $w_i \rightarrow 1$, giving all the weight in (2) to the raw rate estimate. As r_i gets smaller, more and more weight is given to the second term in (2). The empirical Bayes approach (EB) consists of estimating the moments of the prior distribution from the data, rather than taking them as a ''prior'' in a pure.

The principle is referred to as shrinkage, in the sense that the raw rate is moved (shrunk) towards an overall mean, as an inverse function of the inherent variance. Rate smoothing or shrinkage is the procedure used to statistically adjust the estimate for the underlying risk in a given spatial unit, by borrowing strength from the information provided by the other spatial units (Anselin et al, 2004).

2.3 Spatial rate method

A spatial rate smoother is based on the notion of a spatial moving average or window average. $\hat{\theta}_i$ is computed for that unit together with a set of "reference" neighbors. An important practical consideration in the implementation of a spatial smoother is the size of the "window," or, the selection of the relevant neighbors. The spatially smoothed rate map is a choropleth map based on the ranking of the smoothed rate values. It emphasizes broader regional trends and removes some of the spatial detail from the original map. The total number of neighbors for each unit, is not constant and depends on the J_i contiguity structure.

$$\hat{\theta}_i = \frac{y_i + \sum_{J=1}^{J_i} y_j}{n_i + \sum_{J=1}^{J_i} n_j} \quad (4)$$

3. CASE STUDY: GEOGRAPHIC VISUALIZATION

Study data is one of communicable disease, Shigella, and study areas are Gwangju city, Jeollanam-do, and Jeollabuk-do, Korea. Shigella germs are found in the intestinal tract of infected people, and are spread by eating or drinking food or water contaminated by an infected person. It can also be spread by direct contact with an infected person. Communicable diseases appear regional difference according to weather conditions, humidity or temperature etc..

We calculate raw rate, empirical Bayesian rate and spatial rate and create choropleth maps using quantile visualization method. Quantile classification yields visually attractive maps because all of the classes have the same number of features. Table 1 and figure 2-4 show each calculated rate and choropleth maps.

Map and rates using empirical Bayesian rate represent influence of weight (see equation (3)). Raw rates in the region having close rates to mean and more population are only slightly changed to mean because weights are closer to 1. But, raw rates in the region having centric rates to mean and less population are changed much because weights are distant to 1.For instance, Jeollabuk-do Jangsu and Jeollabuk-do Wanju have same raw rates 0, but EB rate of Jangsu is higher than that of Wanju because the population of Wanju is bigger than that of Jangsu.

Also, spatial rate method shows more aggregate patterns because it uses the neighborhood rates (refer Figure 4). For example, Jeollanam-do Gangjin represents 3rd range in raw rate and empirical Bayesian rate while show 1st range (See figure 2-4). Because Spatial rates are calculated not only their population and observation but also neighbour's those (See equation (4) and figure (5)).

Table 1. Rate by calculating raw rate method, empirical Bayesian method, and spatial rate method

Distinct	Shigella	Population	Raw Rate	EB Rate	Spatial Rate
Gwangju Dong-gu	1	128,461	0.000008	0.000009	0.000026
Gwangju seo-gu	22	267,362	0.000082	0.000082	0.000028
Gwangju Nam-gu	5	237,478	0.000021	0.000021	0.000028
Gwangju Buk-gu	3	477,593	0.000006	0.000007	0.000027
Gwangju Gwangsan-gu	7	248,752	0.000028	0.000028	0.000028
Jeollabuk-do Jeonju	0	611,921	0.000000	0.000000	0.000000
Jeollabuk-do Gunsan	0	280,400	0.000000	0.000001	0.000000
Jeollabuk-do iksan	0	337,436	0.000000	0.000000	0.000000
Jeollabuk-do jeongeup	0	151,665	0.000000	0.000001	0.000010
Jeollabuk-do Namwon	0	104,704	0.000000	0.000001	0.000004
Jeollabuk-do Gimje	0	118,811	0.000000	0.000001	0.000000
Jeollabuk-do wanju	0	85,796	0.000000	0.000002	0.000000
Jeollabuk-do Jinan	0	33,752	0.000000	0.000004	0.000000
Jeollabuk-do Muju	0	31,813	0.000000	0.000004	0.000000
Jeollabuk-do Jangsu	0	30,207	0.000000	0.000005	0.000000
Jeollabuk-do Imsil	0	39,327	0.000000	0.000004	0.000000
Jeollabuk-do Sunchang	0	35,860	0.000000	0.000004	0.000000

Jeollabuk-do Gochang	0	76,219	0.000000	0.000002	0.000014
Jeollabuk-do Buan	0	77,620	0.000000	0.000002	0.000000
Jeollanam-do Mokpo	7	246,741	0.000028	0.000029	0.000195
Jeollanam-do Yeosu	14	326,942	0.000043	0.000043	0.000032
Jeollanam-do Suncheon	6	268,204	0.000022	0.000023	0.000027
Jeollanam-do Naju	68	110,501	0.000615	0.000593	0.000120
Jeollanam-do Gwangyang	4	138,267	0.000029	0.000029	0.000027
Jeollanam-do Damyang	0	56,806	0.000000	0.000003	0.000011
Jeollanam-do Gokseong	0	41,676	0.000000	0.000004	0.000014
Jeollanam-do Gurye	1	34,286	0.000029	0.000030	0.000019
Jeollanam-do Goheung	1	102,591	0.000010	0.000011	0.000031
Jeollanam-do Boseong	4	63,112	0.000063	0.000062	0.000019
Jeollanam-do Hwasun	0	78,234	0.000000	0.000002	0.000026
Jeollanam-do Jangheung	0	55,608	0.000000	0.000003	0.000030
Jeollanam-do Gangjin	1	51,238	0.000020	0.000021	0.000111
Jeollanam-do Haenam	23	100,867	0.000228	0.000220	0.000084
Jeollanam-do Yeongam	4	66,199	0.000060	0.000059	0.000172
Jeollanam-do Muan	14	71,341	0.000196	0.000187	0.000174
Jeollanam-do Hampyeong	1	46,929	0.000021	0.000023	0.000167
Jeollanam-do Yuonggwang	2	74,851	0.000027	0.000027	0.000064
Jeollanam-do Jangseong	4	56,647	0.000071	0.000068	0.000015
Jeollanam-do Wando	10	68,871	0.000145	0.000139	0.000092
Jeollanam-do Jindo	0	43,384	0.000000	0.000003	0.000082
Jeollanam-do Sinan	2	54,961	0.000036	0.000036	0.000089
SUM	204	5,533,433			
MEAN	5	134,962	0.000044	0.000044	0.000043

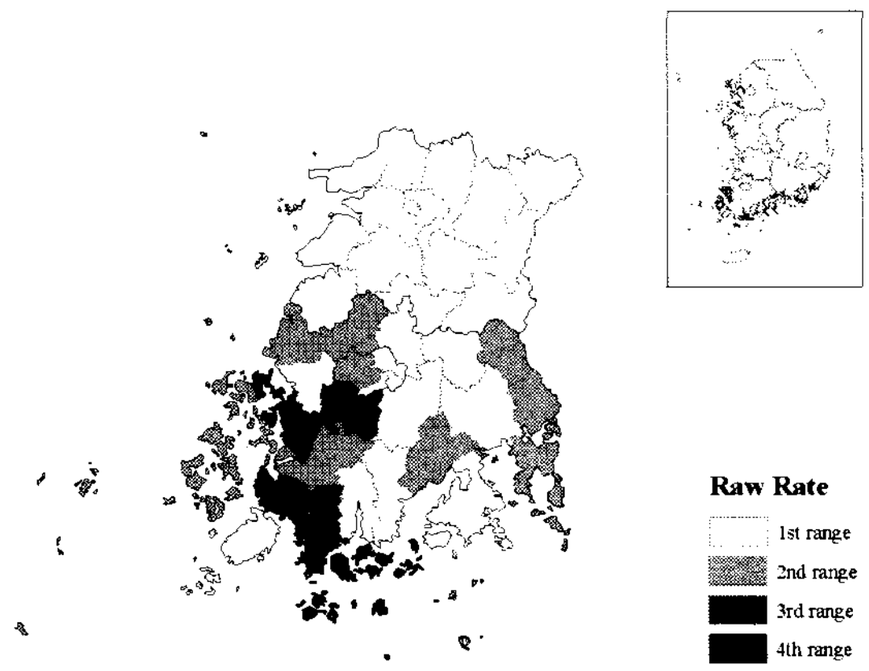

Figure 2. Raw rates method of *Shigella*

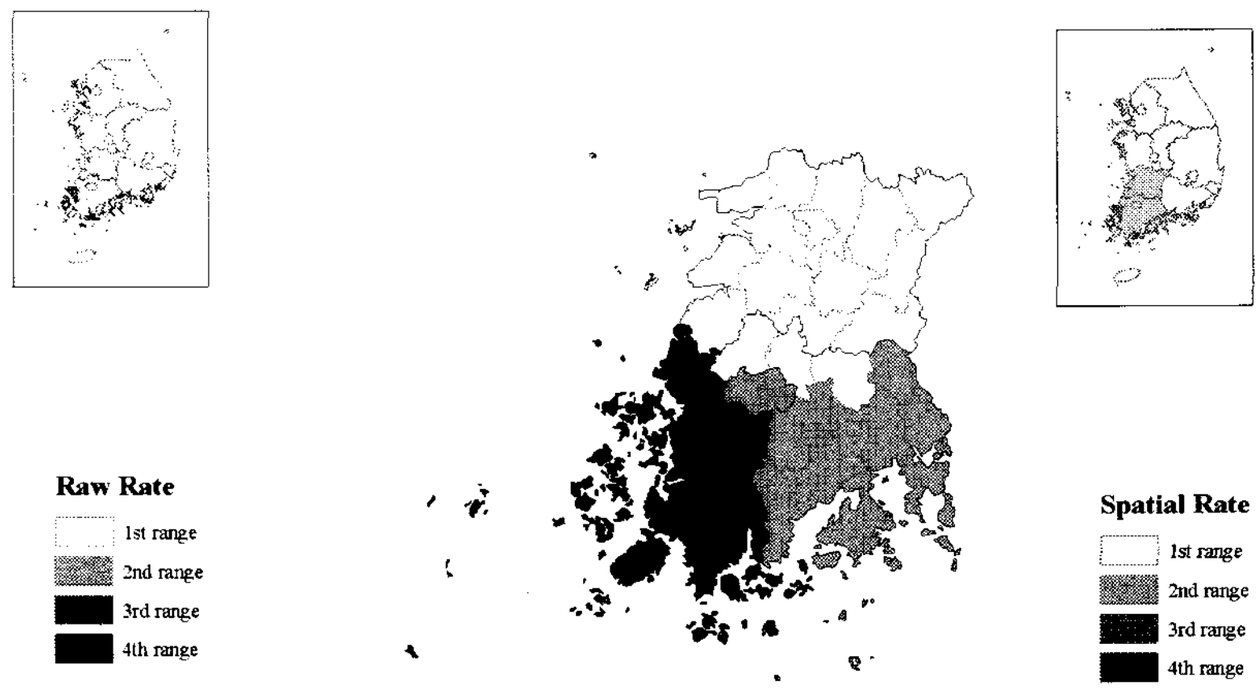

Figure 4. Spatial rate method of *Shigella*

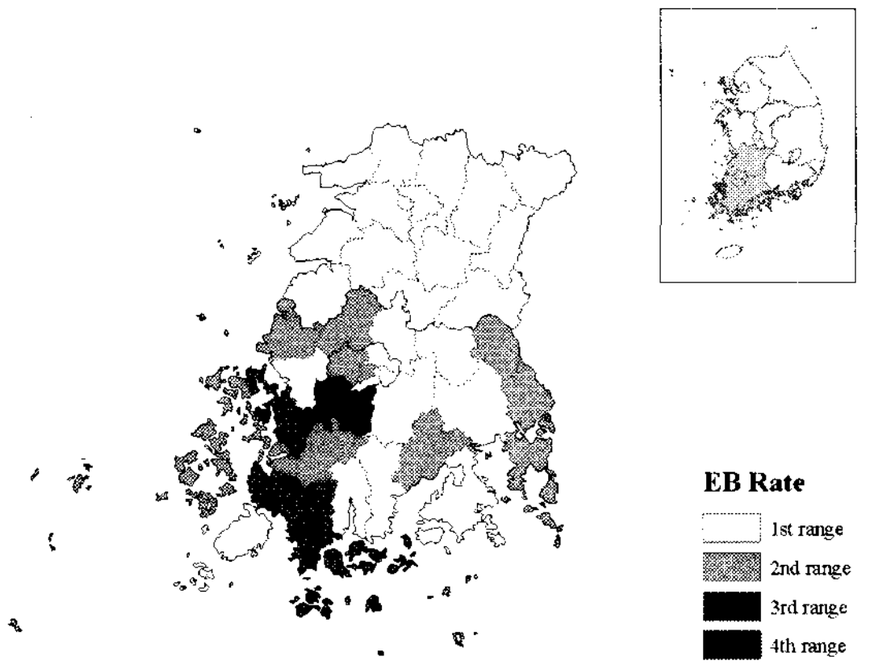

Figure 3. Empirical Bayesian method of *Shigella*

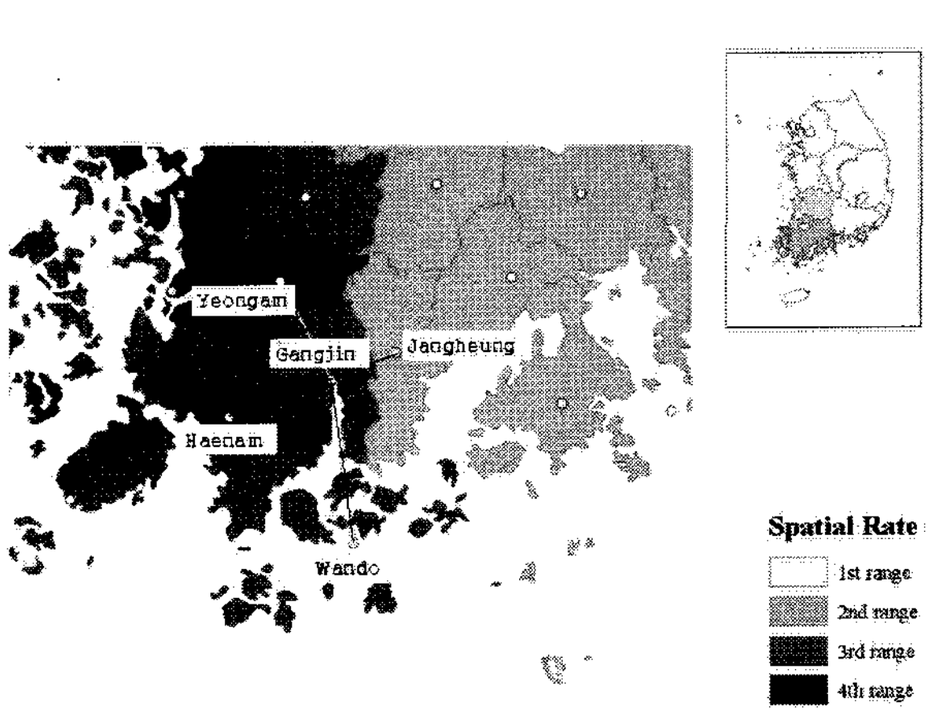

Figure 5. Example of calculating spatial rate method (Jeollanam-do Gangjin)

4. CONCLUSION

In our research, varying calculation methods for rare events, we made three types of ratio map. The experiments reveal that maps are somewhat different by used methods. We suggest that a raw rate method can not be an only way to make a rate map and researchers should choose an appropriate method for their objectives. It is important to realize that there is no best method. Rather, in an exploratory exercise, an assessment of sensitivity of the identified ''patterns'' to the choice of methods is an important consideration.

REFERENCES

Anselin, L., Kim, Y-W., and Syabri,I., 2004, Web-based analytical tools for the exploration of spatial data, *Journal of Geographical Systems*, 6:197-218.

Bailey, T. C. and Gatrell, A.C., 1995, *Inter-active Spatial Data Analysis*, John Wiley and Sons, New York, NY.

Kafadar. K., 1997, *Geographic Trend in Prostate Cancer Mortality: An Application of Spatial Smoother and the Need for Adjustment, ELSIVIER*, New York, NY, pp.35-44

Xiao, N., Calder, C.A., and Armstrong. C.A., 2007, Assessing the effect of attribute uncertainty on the robustness of choropleth map classification, *International Journal of Geographical Information Science*, 21(2), 121-144

Department of Health, New York State, USA, http://www.health.state.ny.us