Journal of the Korean Data & Information Science Society
2009, **20**(5), 919–924

한국데이터정보과학회지

Correspondence analysis for studying association between geography and cancer[†]

Joon Jin Song[1] · Pingjian Yu[2] · Yuan Ren[3] · Ming-Hua Chung[4]

[134]Department of Mathematical Sciences, University of Arkansas
[2]Department of Industrial Engineering, University of Arkansas

Received 12 July 2009, revised 3 September 2009, accepted 7 September 2009

Abstract

Geographical location carries information such as demography, local economy, environment, and life styles, which could be the sources of cancer occurrence. Analyzing geographical location associated with cancer occurrence can be instructive to physicians, patients, and health administrators regarding resource allocation, expenditures, prophylaxis and treatments. In this paper, we explored the correspondence relationship between geographical locations and mortality rates of the cancers using correspondence analysis and illustrated the approach with the mortality rates of the top 10 cancers in the 75 counties in Arkansas from 2001 to 2005. Geographical variations with respect to the mortality rates of cancers are evaluated across Arkansas counties. Based on the contingency table, correspondence analysis model is developed and the simple indices which indicate the degree to which the regions and the cancers affect each other are calculated. Quantitative results are visualized and mapped in two-dimensional graphs.

Keywords: Cancer, correspondence analysis, exploratory spatial data analysis, geography.

1. Introduction

Cancer registry data carries plenty of information such as the association between cancers (Liu *et al.*, 1998; Martín *et al.*, 1998; Matikainen *et al.*, 2001), the association between regions (Rushton *et al.*, 2004; Haselkorn *et al.*, 2005), and the association between cancers and regions (Rosenberg *et al.*, 1999). Exploring these associations can help in the study, diagnosis, control, and prevention of cancers.

[†] This research was partially supported by Arkansas Biosciences Institute (ABI).
[1] Corresponding author: Assistant Professor, Department of Mathematical Sciences, University of Arkansas, Fayetteville, AR 72701, USA. E-mail: jjsong@uark.edu
[2] Graduate student, Department of Industrial Engineering, University of Arkansas, Fayetteville, AR 72701, USA.
[3] Graduate student, Department of Mathematical Sciences, University of Arkansas, Fayetteville, AR 72701, USA.
[4] Graduate student, Department of Mathematical Sciences, University of Arkansas, Fayetteville, AR 72701, USA.

Correspondence analysis is an exploratory technique to analyze two-way or multi-way tables containing some measure of correspondence between the row variables and the column variables. Correspondence analysis provides simple indices that show the weights between the row and the column variables, based on which we can tell which column variables are more important in the row variables of the table or vice versa. This technique has been applied in cancer data analysis. Palli *et al.* (2001) applied correspondence analysis to the gastric cancer data to investigate the association between patient age and dietary pattern. Hans *et al.* (2000) used the technique to investigate the association between the survival years after surgery for breast cancer and patient characteristics, such as patient age, clinical stage, lymph nodes status and etc.

The objective of this paper is to discover the correspondence relationship between geographical locations and mortality rates of the cancers using correspondence analysis. To illustrate this approach, a part of Arkansas Cancer Restry is used, containing the mortality rates of the top 10 cancers in 75 counties in Arkansas, United States, from 2000 to 2005.

2. Correspondence analysis

2.1. Test of independence

Before measuring the association between the row and column variables in corresponding analysis, χ^2-test is commonly used to test if there is an association between the row variables and the column variables. The null hypothesis is that the row and column variables are independent. The test statistic is given by

$$t = \sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij} - E_{ij})^2/E_{ij},$$

where $E_{ij} = (x_{i\cdot}x_{\cdot j})/x_{\cdot\cdot}$, $x_{i\cdot} = \sum_j x_{ij}$, $x_{\cdot j} = \sum_i x_{ij}$, and $x_{\cdot\cdot} = \sum_{i,j} x_{ij}$. This statistic follows χ^2 distribution with $(n-1)(p-1)$ degree of freedom. If null hypothesis of independence is rejected, it is suspected that there is some association between the row and column variables.

2.2. χ^2 decomposition

Correspondence analysis decomposes a measure of association, typically the total χ^2 value used in the test of independence. For the purpose, singular value decomposition (SVD) is applied to the $(n \times p)$ matrix $C = [c_{ij}]$. Assume X is a contingency table, and x_{ij} is an element in X. The matrix C is given by

$$c_{ij} = (x_{ij} - E_{ij})/E_{ij}^{1/2},$$

a measure of the weighted departure between the observed value and the theoretical values under independence. Then, SVD factorizes C by $\Gamma\Lambda\Delta^T$, where Γ are eigenvectors of CC^T, Δ are eigenvectors of C^TC, $\Lambda = Diag(\lambda_1^{1/2}, \cdots, \lambda_R^{1/2})$, $\lambda_1^{1/2} \geq \cdots \geq \lambda_R^{1/2}$ are eigenvalues of CC^T, and $R = rank(C) \leq \min(n-1, p-1)$. Then, C can be rewritten by

$$c_{ij} = \sum_{k=1}^{R} \lambda_k^{1/2}\gamma_{ik}\delta_{jk},$$

where γ_{ik} and δ_{jk} are the elements of Γ and Δ, respectively.

2.3. Projections for graphical displays

From previous result, duality relations can be derived. Let $\delta_k = C^T\gamma_k/\sqrt{\lambda_k}$, $\gamma_k = C\delta_k/\sqrt{\lambda_k}$, $A = Diag(x_{i\cdot})$, and $B = Diag(x_{\cdot j})$ for $k = 1, \cdots, R$. The projections (r_k and s_k) for graphical display are given by

$$r_k = A^{-1/2}C\delta_k = \sqrt{\lambda_k}A^{-1/2}\gamma_k$$

and

$$s_k = B^{-1/2}C^T\gamma_k = \sqrt{\lambda_k}A^{-1/2}\delta_k.$$

The duality relations for r_k and s_k are

$$r_k = \frac{1}{\sqrt{\lambda_k}}A^{-1/2}CB^{1/2}s_k = \sqrt{\frac{x_{\cdot\cdot}}{\lambda_k}}A^{-1}Xs_k$$

and

$$s_k = \frac{1}{\sqrt{\lambda_k}}B^{-1/2}C^TA^{1/2}r_k = \sqrt{\frac{x_{\cdot\cdot}}{\lambda_k}}B^{-1}X^Tr_k.$$

Choosing $k = 2$ can form a 2-dimension graph to display the correspondence relationship. r_k and s_k represent the k^{th} indices for the row variables and for the column variables, respectively.

2.4. Interpretations

In the graph based on the projections in the previous section, the nearness of two row variables on the graph indicates a similar profile in these two row variables. "Profile" here refers to the conditional frequency distribution of a row variable. On the other hand, the dissimilarity is observed when two row variables are far apart from each other. Further distance between two row variables on the graph reflects more dissimilarity. The distance of two column variables is also interpreted in this way. Meanwhile, short distance between a row variable and a column variable indicates that the two variables have a particularly important weight in each other, which means the row variable is observed frequently in the column variable (and vice versa). In contrast to this, if a row variable is located far from a column variable, hardly any this row variable can be observed in this column variable (and vice versa). These conclusions are particularly true when the points are far away from origin, which is the average of the factors r_k and s_k. Hence, a profile similar to average is shown when a point is projected close to the origin.

3. Application: Arkansas cancer data

To perform correspondence analysis, a contingency table is constructed with Arkansas cancer data. The row variable is county and column variable is cancer. In this data set, the values of row variables are from 1 to 75, which represent 75 counties in Arkansas, while the column variables are the top 10 cancers with the highest age-adjusted death rate in

Arkansas. The cancers are prostate, female breast, lung and bronchus, colon and rectum, urinary bladder, non-hodgkin lymphoma, corpus uteri, kidney and renal pelvis, melanoma of the skin and ovary.

Firstly, a χ^2-test for independence is conducted, resulting in very small p-value, 0.0004998. This small p-value indicates that there is a suspected correspondence relationship between county and cancer. By using the formula in previous section, the 2-dimension coordinates $(r_k,\ s_k)$ for each county and cancer are computed. r_1 and r_2 are the first and second coordinate values for the row variable (counties). Similarly, s_1 and s_2 are the first and second coordinate values for the column variable (cancers). To clearly demonstrate the relationship between rows (counties) and columns (cancers), counties and cancers are projected to a two-dimensional display in Figure 3.1. Note that this display is based on the relative position of the points in terms of the weights corresponding to the column and the row. The

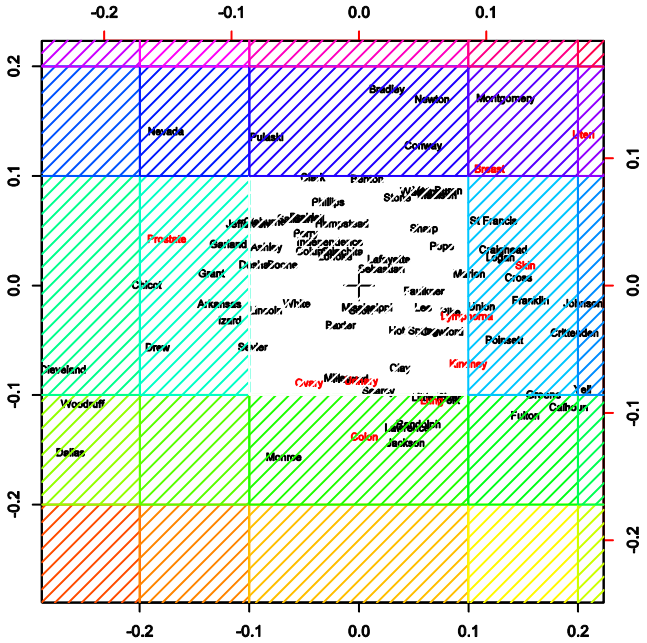

Figure 3.1 Biplot of correspondence between county and cancer.

proximity of two counties (or cancers) indicates a similar profile in these two counties (or cancers), which means the likelihood of occurrence of a particular cancer is similar to contiguous cancer(s), or the likelihood of occurrence of cancers in a particular county is similar to contiguous county(s). Oppositely, when the two counties (or cancers) are far apart from each other, the dissimilarity of profiles is shown. The proximity of a particular county to a particular cancer indicates that the county (cancer) has a particularly important weight in the cancer (county). The closer a cancer and a county are located, the higher risk of the cancer happens in this county. In contrast to this, a county that is quite distant from a particular cancer indicates that there are fewer observations of this cancer in this county (and vice versa). In order to see the result more efficiently, biplot is used to divide this plot into 25 areas which represented by 25 colors. Specifically, we use color “white” to represent

the counties and cancers that have no significant correspondence relations. According to the criteria of interpretation, if two counties are close to each other, they have the positive correlation. In the following graph, two close counties will also share the same color. If two counties are far away in the opposite direction, they have a negative correlation. The distances between origin and spots show how strong the relationships are. The further a spot is away from origin, the stronger the relationship is.

As expected, there is an association between the counties and the occurrence of cancers. This relationship can be interpreted in terms of similar distance and direction from the origin. In particular, in contrast to other cancers, corpus uteri cancer is far away from other cancer and all counties, which indicates that this cancer has different profile than other cancers and no county is significantly more risky to have this cancer than other counties. In the sky-blue area (between 0.1 and 0.2 horizontally and -0.1 and 0.1 vertically), melanoma of the skin cancer dominates. Meanwhile, counties observed within this area represent that they have similar profile and have high risk of occurring the cancer I.

To be clearer, the relationship is also mapped in Figure 3.2, which geographically shows correspondence between counties and cancers. For example, as indicated in Figure 3.1, a spa-

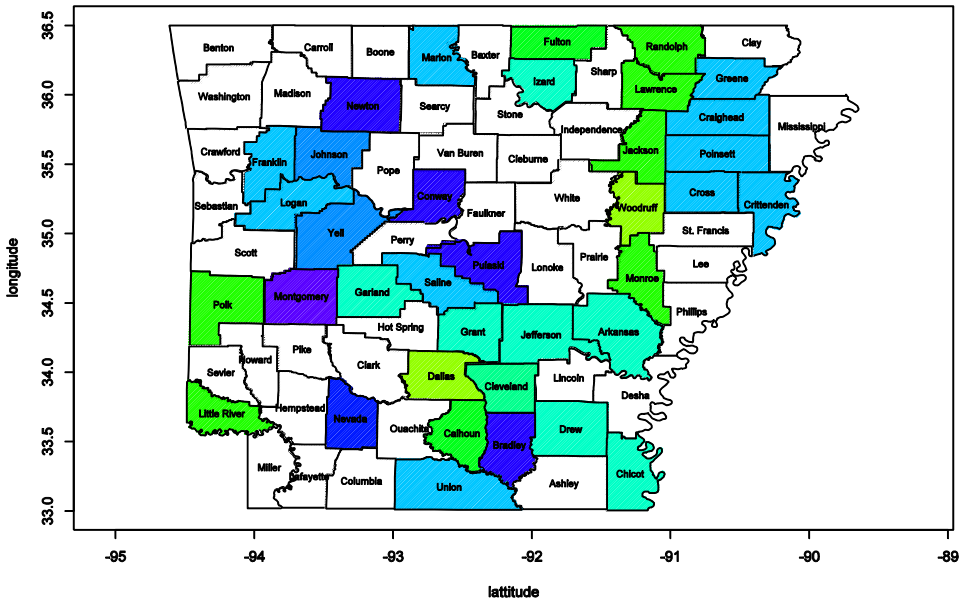

Figure 3.2 Classification of correspondence between counties and top 10 cancers in Arkansas.

tial cluster (sky-blue colored) is shown, consisting of greene, craighead, cross, poinsett, and crittenden counties which geographically share their borders in the northeast of Arkansas. Right beside the northeast part, there is a series of counties having a high association with lung and bronchus cancer as well as colon and rectum cancer. In the southeast part of Arkansas, a group of counties have a higher association with prostate cancer than other counties.

4. Conclusion and discussion

In this paper, we propose to use correspondence analysis for studying the association between geography and cancer, and illustrate this approach using Arkansas cancer data. The main idea of correspondence analysis is to develop simple indices that can show the relationship between the row and the column variables. In the analysis, we discover some association between counties and cancers in Arkansas using the proposed approach. As a future study, this study can be extended to investigate the association with accounting for more potential risk factors, such as demography, economics, industry structure, and pollutions.

References

Hans, D., Ojasoo, T. and Doré, J. (2000). Deaths from breast cancer: Tackling multidimensionality and non-linearity by correspondence analysis. *The Journal of Steroid Biochemistry and Molecular Biology*, **74**, 195-202.

Haselkorn, T., Whittemore, A. S. and Lilienfeld, D. E. (2005). Incidence of small bowel cancer in the United States and worldwide: Geographic, temporal, and racial differences. *Cancer Causes & Control*, **16**, 781-787.

Liu, L., Deapen, D. and Bernstein, L. (1998). Socioeconomic status and cancers of the female breast and reproductive organs: A comparison across racial/ethnic populations in Los Angeles county, California (United States). *Cancer Causes & Control*, **9**, 369-380.

Martín, A. A., Galán, Y. H., Rodríguez, A. J., Graupera, M., Lorenzo-Luaces, P., Fernández, L. M., Camacho, R. and Lezcano, M. (1998). The Cuban national cancer registry: 1986-1990. *European Journal of Epidemiology*, **14**, 287-297.

Matikainen, M. P., Pukkala, E., Schleutker, J., Tammela, T. L., Koivisto, P., Sankila, R. and Kallioniemi, O. (2001). Relatives of prostate cancer patients have an increased risk of prostate and stomach cancers: a population-based, cancer registry study in Finland. *Cancer Causes & Control*, **12**, 223-230.

Palli, D., Russo, A. and Decarli, A. (2001). Dietary patterns, nutrient intake and gastric cancer in a high-risk area of Italy. *Cancer Causes & Control*, **12**, 163-172.

Rosenberg, M. S., Sokal, R. R., Oden N. L. and DiGiovanni, D. (1999). Spatial autocorrelation of cancer in Western Europe. *European Journal of Epidemiology*, **15**, 15-22.

Rushton, G., Peleg, I., Banerjee, A., Smith, G. and West, M. (2004). Analyzing geographic patterns of disease incidence: Rates of late-stage colorectal cancer in Iowa. *Journal of Medical Systems*, **28**, 223-236.