Journal of the Korean Data & Information Science Society
2014, **25**(6), 1531–1538

http://dx.doi.org/10.7465/jkdi.2014.25.6.1531
한국데이터정보과학회지

Visualization and interpretation of cancer data using linked micromap plots

Se Jin Park[1] · Jeong Yong Ahn[2]

[1]Department of Statistics, Chonbuk National University
[2]Department of Statistics (Institute of Applied Statistics), Chonbuk National University

Received 26 August 2014, revised 25 September 2014, accepted 5 October 2014

Abstract

The causes of cancer are diverse, complex, and only partially understood. Many factors including health behaviors, socioeconomic environments and geographical locations can directly damage genes or combine with existing genetic faults within cells to cause cancerous mutations. Collecting the cancer data and reporting the statistics, therefore, are important to help identify health trends and establish normal health changes in geographical areas. In this article, we analyzed cancer data and demonstrated how spatial patterns of the age-standardized rate and health indicators can be examined visually and simultaneously using linked micromap plots. As a result of data analysis, the age-standardized rate has positive correlativity with thyroid and breast cancer, but the rate has negative correlativity with smoking and drinking. In addition, the regions with high age-standardized rate are located in southwest and the areas of high population density while the standardized mortality ratio is higher in southwest and northeast where there are lots of rural areas.

Keywords: Age-standardized rate, data visualization, health indicators, linked micromap plots.

1. Introduction

Cancer has been the leading cause of death in Korea since 1983 and is associated with the largest disease burden (Yoon *et al.*, 2007). More than 190,000 new cancer cases are diagnosed annually in Korea, and one in four deaths results from cancer (Jung *et al.*, 2012). According to GLOBOCAN 2012 (Forman and Bray, 2013), an estimated 14.1 million new cancer cases and 8.2 million cancer-related deaths occurred globally in 2012, compared with 12.7 million and 7.6 million, respectively, in 2008. Prevalence estimates for 2012 show that there were 32.6 million people (over the age of 15 years) alive who had had a cancer diagnosed in the previous five years.

The causes of cancer are diverse, complex, and only partially understood. Many things are known to increase the risk of cancer including health behaviors, socioeconomic factors, geographical location and so on (Anand *et al.*, 2008). These factors can directly damage genes

[1] Graduate student, Department of Statistics, Chonbuk National University, Jeonbuk 561-756, Korea.
[2] Corresponding author: Professor, Department of Statistics, Chonbuk National University, Jeonbuk 561-756, Korea. E-mail: jyahn@jbnu.ac.kr

or combine with existing genetic faults within cells to cause cancerous mutations (Vogelstein and Kinzler, 2002). Collecting the cancer data and reporting the statistics, therefore, are important to help identify health trends and establish normal health changes in geographical areas.

In many cases, the health data tend to depend on tables to disseminate the information. The presentation of the data in tabular form, however, is uninformative from an interpretative standpoint. It may be difficult and frustrating for a reader to observe trends, relationships, and anomalies that may be present in the data (Gebreab *et al.*, 2008). An alternative to solve the problems is the conversion of tabular data into a visual and ordered context.

In this study, we introduce an approach for visualizing cancer data and describe how spatial patterns of cancer incidence rate and the health indicators can be examined visually and simultaneously using the approach called linked micromap plots. To explore the potential utility of the plots, we analyze and visualize cancer data of Korea in 2010. First, we grasp the relationship between the incidence rate and health indicators through statistical data analysis, and select some indicators that affect the rate. Second, we explore the geographic patterns and the relationships between the variables in the health data using linked micromap plots.

2. Related works

Over the last decade, many researchers have developed many improvements to make statistical graphics and data visualization techniques more accessible to the general public. These improvements include making statistical summaries more visual and providing more information at the same time (Symanzik and Carr, 2008), and they were further practiced in many researches (Cho, 2012; Han *et al.*, 2012; Wong *et al.*, 2012; Lee *et al.*, 2013; Ha and Noh, 2013; Cho, 2014).

As data visualization techniques have been used as an important tool to gain insights into data sets, statistical maps are frequently used in the fields of health research. Showing geographically distributed health related data in a choropleth map, a type of statistical map, using sequential color scheme is quite useful for data driven knowledge discovery (Maceachren *et al.*, 1998). Edsall (2003) created a geospatial data exploration system, including a choropleth map, a parallel coordinate plot, and a scatterplot, to analyze health statistics data. He argued that the multidimensional nature of health statistics and their analysis called for the integrated approach for geovisualization. Choropleth map, however, has many problems and limitations such as it is difficult to show more than one variable in a map (Harris, 1999).

To solve the problems, Carr and Pierson (1996) developed the linked micromap (LM) plots. LM plots can display small maps with boxplots, dotplots, and other statistical graphs. Chen *et al.* (2006) introduced web-based interactive LM plots to display cancer statistics as an application and implementation example to present the plots. Gebreab *et al.* (2008) demonstrated the use of LM plots for the display of geographically indexed oral cleft occurrence, and Pickle and Carr (2010) designed a LM plot to visualize the spatial patterns of US cancer data. These LM plots are implemented to disseminate health data for public use in an easily interpretable format.

3. Data and methods

3.1. Data sources

To analyze and visualize the cancer data, we collected the health data for the period of 2010 from the NCC (National Cancer Center), and aggregated the data by 16 regional local governments of South Korea. The data set consists of 16 indicators/variables including region name. Table 3.1 lists the indicators used in this study.

Table 3.1 Variables of data set

Categories		Indicators
Cancer indicators	Incidence rates	Age-standardized rate (ASR)
	Incidence rates by sites	ASR Stomach, ASR Colon and rectum, ASR Liver, ASR Lung, ASR Breast, ASR Thyroid
Health determinants	Mortality ratio	Standardized mortality ratio (SMR)
	Health behaviors	Smoking, Binge drinking, Walking activity, Cancer screening
	Socioeconomic factors	Aged people, Single household, Poverty
Geographical locations		Region name

3.2. Visualization technique

The graphical visualization technique presented in this study is referred to as LM plot, a type of statistical maps. LM plots are graphics that link statistical information to an organized set of small maps in order to explore and communicate patterns in the outcome variable, geographic locations and the associations among them (Carr and Pickle, 2010).

LM plots have four key features (Carr and Pierson, 1996). The first feature is three or more sequence panels in parallel linked by location. The types of the panel are micromap, label, and statistical summary panels. The second feature is sorting the geographic subregions based on the statistical variables of interest. Sorting improves perception between consecutive panels from the top to the bottom of the display. The third feature is the partitioning of the regions into perceptual groups of size five or less to allow the viewer's attention to focus on explicit areas at a time. The last feature is color and location that links corresponding elements within the parallel sequence panels. These features of LM plots make it possible to identify specific geographic patterns in the data that are often lost with other types of graphs and maps.

4. Data analysis and visualizing

4.1. Data analysis

The most commonly used indicator in cancer data analysis is the age-standardized rate (ASR). The ASR is a summary measure of the rate that a given population would have if it had a standard age structure. Standardization is necessary when comparing several populations that differ with respect to age because age has a powerful influence on the risk of cancer.

The ASR is a weighted mean of the age-specific rates and the weights are given by population distribution of a standard population. It is also expressed per 100,000. The ASR

is affected to a greater or lesser extent by many factors. Health behaviors or lifestyle factors have often been cited as the major determinants of cancer. More recently, differences in health outcomes by socioeconomic position have been recognized as a persisting and perhaps even increasing public health problem.

Table 4.1 Correlation coefficient between the ASR and the variables

Categories	Indicators	Correlation coefficient
Cancer sites	Stomach	0.348
	Colon and rectum	0.454
	Liver	-0.182
	Lung	0.015
	Thyroid	0.952**
	Breast	0.537*
Health behaviors	Smoking	-0.723**
	Binge drinking	-0.782**
	Walking activity	0.569*
	Cancer screening	0.119
Socioeconomic factors	Aged people	-0.323
	Single household	-0.264
	Poverty	-0.146

*p <0.05. **p <0.01

Table 4.1 shows the correlation coefficients between the ASR and the variables used in this study. Thyroid, breast and walking activity are positively correlated with the ASR while smoking and drinking are negatively. Note that the ASR is a measure standardized by age.

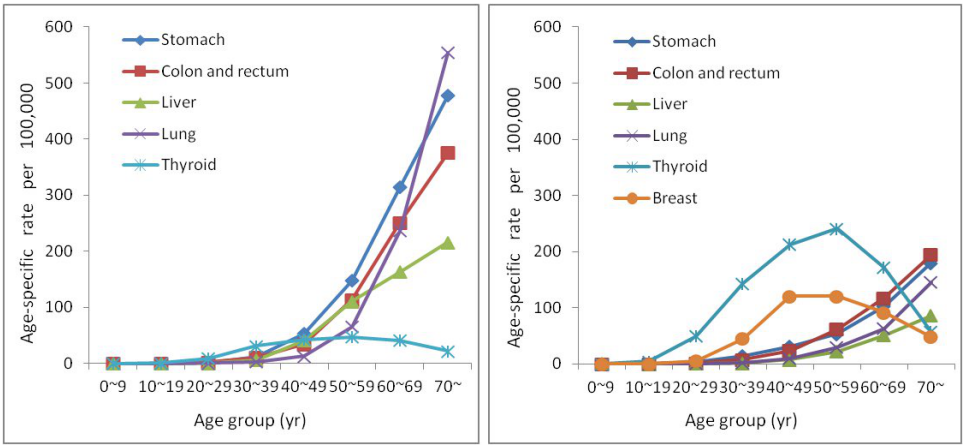

Figure 4.1 ASR of major cancers (left: male, right: female)

Figure 4.1 shows the ASR for selected cancers in males and females during 2010 in South Korea. The figure shows that the incidences of lung, stomach, colorectal, and liver cancers increased gradually with age. Incidences of breast and thyroid cancers in females were highest in those in their 40s and 50s, respectively, and leveled off thereafter.

4.2. Visualizing data using linked micromap plots

The LM plots provide spatial patterns and the relationships of variables while linking regional names to their locations on a map and to the estimates represented in statistical panels. The patterns and relationships are often lost with other types of graphs and data analysis.

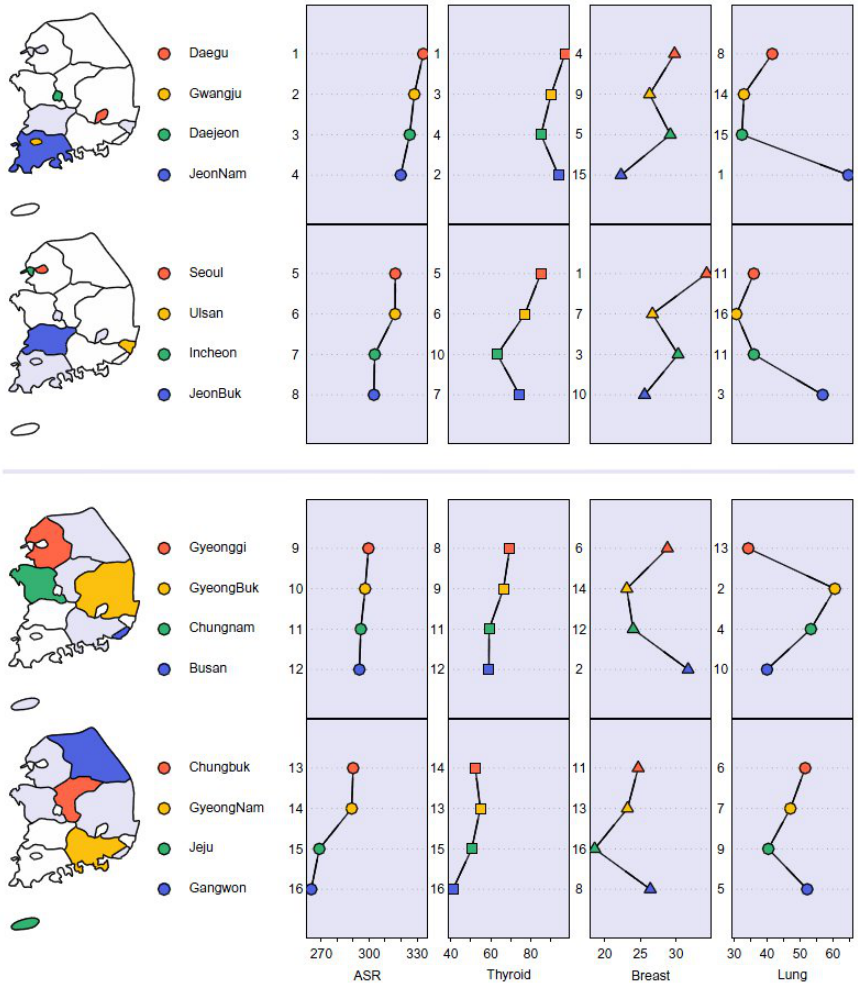

Figure 4.2 Total ASR and ASR by cites

To represent statistical information in LM plots, we developed an extended version of the function, LinkedMicroMap(), developed in our previous studies (Park and Ahn, 2013; Han *et al.*, 2014). In the extended version, we reformed the function and added some options to establish various statistical plots. The program modules can be downloaded on author's web site (http://stat_park.blog.me).

Figure 4.2 shows the LM plot to the relationships between the ASR and three sites of cancer, thyroid, breast and lung, for the 16 regional local governments of South Korea. The regions in the figure are sorted by the ASR from largest to smallest. The ASR and thyroid cancer have high correlativity while the ASR and lung cancer have no correlativity. The figure also provides a viewer with a quick overview of any spatial patterns presented in the ASR. The regions with high ASR are located in southwest and the areas of high population density.

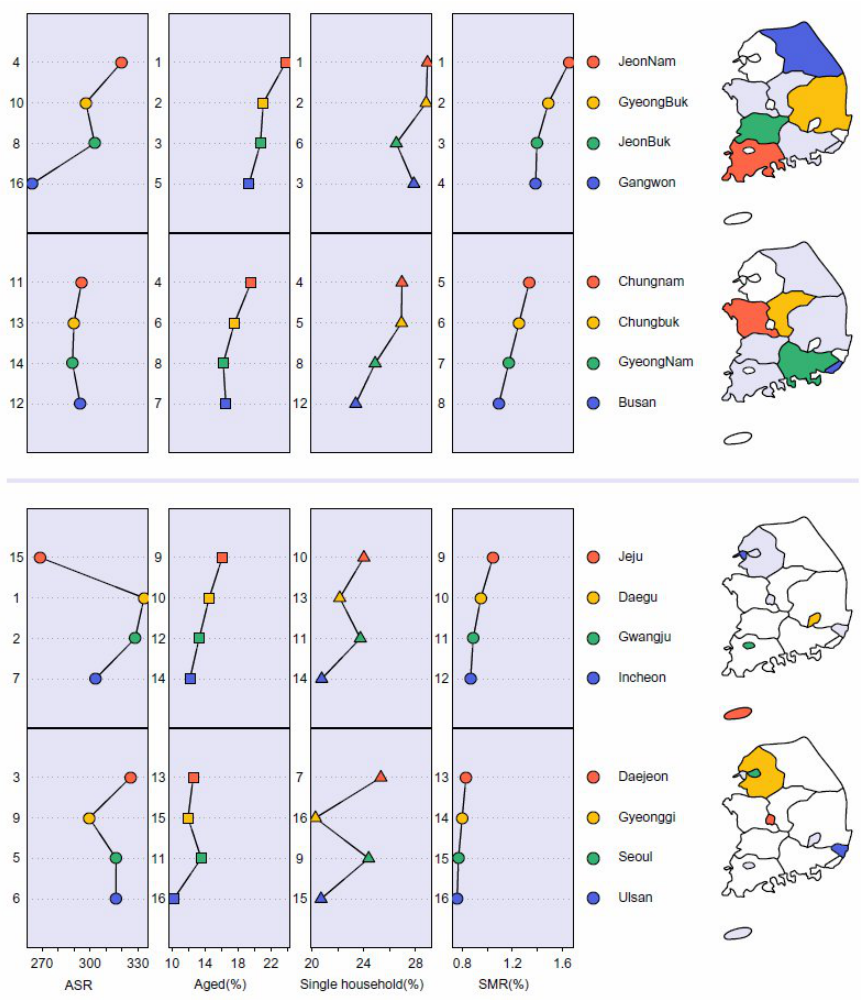

Figure 4.3 ASR and socioeconomic factors

Figure 4.3 shows the LM plot to explore the spatial patterns of the ASR and socioeconomic factors. The regions in the figure are sorted by the standardized mortality ratio (SMR) from largest to smallest. The SMR is the ratio of the observed number of deaths to the expected number of deaths. In Figure 4.3, the SMR and two variables (aged people and

single household) have high correlativity while the SMR and ASR have no correlativity. The regions with high SMR are located in southwest and northeast where there are lots of rural areas.

5. Conclusion

Increasing concern in health and advances in medical technology have lengthened the lives of people. However, there is still a distinct difference between the regions and socioeconomic factors of the regions have been affecting fitness levels. In this study, we analyzed the health data and described how spatial patterns of the ASR and the health indicators can be examined visually and simultaneously using LM plots.

Linked micromaps link statistical graphics and small maps by color across the rows in a tabular format. More information is provided by this connection than by either the graphic or the map alone, providing a geographic context for the statistics. This design has been used successfully to explore data interactively, such as by sorting the rows or selecting variables to display, and also in static form to communicate data or analytic results.

The results of data analysis show that the ASR has positive correlativity with thyroid and breast cancer, while high negative with smoking and drinking. In addition, some spatial patterns are explored from the LM plots. The plots has shown that the regions with high ASR are located in southwest and the areas of high population density, while the SMR is higher in southwest and northeast where there are lots of rural areas. With the results of this study, we can present and visualize more enriched statistical information and help users understand a great variety of geographically referenced data.

References

Anand, P., Kunnumakkara, A. B., Kunnumakara, A. B., Sundaram, C., Harikumar, K. B., Tharakan, S. T., Lai, O. S., Sung, B. and Aggarwal, B. B. (2008). Cancer is a preventable disease that requires major lifestyle changes. *Pharmaceutical Research*, **25**, 2097-2116.

Carr, D. B. and Pickle, L. W. (2010). *Visualizing data patterns with micromaps*, Chapman and Hall/CRC, FL.

Carr, D. B. and Pierson, S. M. (1996). Emphasizing statistical summaries and showing spatial context with micromaps. *Statistical Computing and Graphics Newsletter*, **7**, 16-23.

Chen, J. X., Carr, D. B., Wechsler, H. and Pan, Z. (2006). Interactive visualization of multivariate statistical data. *The International Journal of Virtual Reality*, **5**, 67-73.

Cho, J. S. (2012). Inflow and outflow analysis of double majors using social network analysis. *Journal of the Korean Data & Information Science Society*, **23**, 693-701.

Cho, J. S. (2014). Analysis of employee's characteristic using data visualization. *Journal of the Korean Data & Information Science Society*, **25**, 727-736.

Edsall, R. (2003). Design and usability of an enhanced geographic information system for exploration of multivariate health statistics. *Professional Geographer*, **55**, 605-619.

Forman, D. and Bray, F. (2013). *GLOBOCAN 2012: Cancer incidence and mortality worldwide*, available from http://globocan.iarc.fr.

Gebreab, S., Gillies, R. R., Munger, R, G. and Symanzik, J. (2008). Visualization and interpretation of birth defects data using linked micromap plots. *Birth Defects Research (Part A)*, **82**, 110-119.

Ha, I. D. and Noh, M. S. (2013). A visualizing method for investigating individual frailties using frailtyHL R-package. *Journal of the Korean Data & Information Science Society*, **24**, 931-940.

Han, K. S., Park, S. J., Mun, G. S., Choi, S. H., Symanzik, J., Gebreab, S. and Ahn, J. Y. (2014). Linked micromaps for the visualization of geographically referenced data. *ICIC Express Letters*, **8**, 443-448.

Han, K. Y., Park, S. J. and Ahn, J. Y. (2012). Development of a R function for visualizing statistical information on Google static maps. *Journal of the Korean Data & Information Science Society*, **23**, 971-981.

Harris, R. L. (1999). *Information graphics - A comprehensive illustrated reference*, Oxford University Press, New York.

Jung, K. W., Park, S., Kong, H. J., Won, Y. J., Lee, J. Y., Seo, H. G. and Lee, J. S. (2012). Cancer statistics in Korea: incidence, mortality, survival, and prevalence in 2009. *Cancer Research and Treatment*, **44**, 11-24.

Lee, J. Y., Bae, J. Y., Lee, J. M., Oh, D. Y. and Lee, S. W. (2013). Major gene interactions effect identification on the quality of Hanwoo by radial graph. *Journal of the Korean Data & Information Science Society*, **24**, 151-159.

Maceachren, A. M., Brewer, C. A. and Pickle L. W. (1998). Visualizing georeferenced data: Representing reliability of health statistics. *Environment and Planning A*, **30**, 1547-1561.

Park, S. J. and Ahn, J. Y. (2013). Visualizing statistical data using linked micromap plots. *Journal of The Korean Official Statistics*, **18**, 111-127.

Pickle, L. W. and Carr, D. B. (2010). Visualizing health data with micromaps. *Spatial and Spatio-temporal Epidemiology*, **1**, 143-150.

Symanzik, J. and Carr, D. B. (2008). Interactive linked micromap plots for the display of geographically referenced statistical data. In *Handbook of Data Visualization*, edited by C. Chen, W. Hardle & A. Unwin, Springer, Heidelberg, Berlin.

Vogelstein, B. and Kinzler, K. W. (2002). *The genetic basis of human cancer*, McGraw-Hill, New York.

Wong, D. H., Ramadass, S. and Chai, K. (2012). Towards an adaptive framework for network data visualization, *ICIC Express Letters*, **6**, 425-430.

Yoon, S. J., Bae, S. C., Lee, S. I., Chang, H., Jo, H. S., Sung J. H., Park, J. H., Lee, J. Y. and Shin, Y. (2007). Measuring the burden of disease in Korea. *Journal of Korean Medical Science*, **22**, 518-523.