Applying spatial epidemiology to hematological disease using R: a guide for hematologists and oncologists

Kei Kohno1
Hiroto Narimatsu2
Katsumi Otani2
Ri Sho2
Yosuke Shiono1
Ikuko Suzuki1
Yuichi Kato1
Akira Fukao2
Takeo Kato1

1Department of Neurology, Hematology, Metabolism, Endocrinology, and Diabetology, Yamagata University School of Medicine, 2Department of Public Health, Yamagata University Graduate School of Medicine, Yamagata, Japan

Correspondence: Hiroto Narimatsu
Department of Public Health, Yamagata University Graduate School of Medicine, 2-2-2 Iida-nishi, Yamagata, 990-9585, Japan
Tel +81 02 3628 5260
Fax +81 02 3628 5261
Email hiroto-narimatsu@umin.or.jp

Abstract: “Spatial statistics” is an academic field that deals with the statistical analysis of spatial data, and has been applied to econometrics and various other policy fields. These methods are easily applied by hematologists and oncologists using better and much less expensive software. To encourage physicians to use these methods, this review introduces the methods and demonstrates the analyses using R and FleXScan, which can be freely downloaded from the website, with sample data. It is demonstrated that spatial analysis can be used by physicians to analyze hematological diseases. In addition, applying the technique presented to the investigation of patient prognoses may enable generation of data that are also useful for solving health policy-related problems, such as the optimal distribution of medical resources.

Keywords: leukemia, malignant lymphoma, Tango’s index, spatial regression model

Introduction

“Spatial statistics” is an academic field that deals with the statistical analysis of spatial data. In the field of epidemiology, Snow created a cholera map in the 19th century with the goal of extracting the spatial unevenness in the distribution of cholera patients in an outbreak in London, and he used it as the basis for establishing measures for preventing cholera. This is a formulation of what today is called spatial clustering, and its modern applications have been developed as spatial epidemiology, directed toward analyzing risk assessment for infectious diseases and various other diseases.1 “Spatial statistics” has also been applied to many fields, including econometrics and various other policy fields.2

Implementing spatial statistics requires a statistics package for the use of special statistical techniques, but in recent years, R Software3 and FleXScan software,4 which are statistical packages for spatial statistics, have become available free of charge to all. It is also essential to use a graphic information system (GIS). A GIS is a construct for linking text, numbers, images, or the like to a map, creating a reproduction on a computer, and integrating, analyzing, or making an easy to understand map representation of various forms of information from locations and positions; it has been widely used in the fields of disaster management and in business settings. To use a GIS, there is not only commercial software, such as ArcGIS (ESRI; Redlands, CA, USA), but also free software, such as the Quantum GIS (QGIS Development Team; Quantum GIS Geographic Information System. Open Source Geospatial Foundation Project, http://qgis.osgeo.org),5 and environments have been set up for clinicians to allow them to conduct spatial epidemiological research.

Regional clustering can help elucidate the etiology of hematological and oncological diseases, such as adult T-cell leukemia.6 The study of regional clustering is expected to lead to the identification of risk factors and a better understanding of the pathology
of these diseases. Since the uneven distribution of diseases is thought to be dependent also on the availability of medical services aimed at the proper diagnosis of hematological diseases, spatial analysis of hematological diseases would also be useful in the field of health policy.7,8

Yamagata Prefecture, which is located about 300 km north of Tokyo with a population of about 1.2 million, boasts a regional cancer registry of the highest precision in Japan, and it is one of the few prefectures where the incidence of cancer can be comprehensively understood. Therefore, this information was used to implement spatial analysis of hematological diseases with a spatial statistics package as a guide to hematologists and oncologists. To encourage physicians to use these methods, this review introduces the methods and demonstrates the analyses using R and FleXScan with sample data.

Software used for statistical analysis

R version 2.14.2 (R Foundation for Statistical Computing, Vienna, Austria) and the packages “spdep”, “Dcluster”, and “classInt” were used. R can be downloaded from the website.3 FleXScan software version 3.1 (FleXScan; National Institute of Public Health, Tokyo, Japan) was used to conduct global clustering tests using Tango’s index.9 The users’ guide can also be downloaded from the website.4

For regression analysis in an econometric model,7 the incidences of diseases in each municipality and the number of hospitals that employ full-time hematologists were shown. These data were collected from interviews with hematology physicians and from the hospitals’ websites.

The age-adjusted disease incidence was calculated using the 1985 model population of Japan10 and the 2008 model population of Yamagata Prefecture.11 The detailed method of spatial analysis using R has been described elsewhere.7,8

Data used for analysis

The data related to hematological malignant diseases including malignant lymphoma, leukemia, and multiple myeloma between 2000 and 2008 were provided by the cancer registry of Yamagata Prefecture. The data included type of disease, date of onset of disease, age, sex, and the cities where the patients lived. The cancer registry in Yamagata Prefecture is of sufficient quality; in 2008, rates of death certificate notification and death certificate only were 18.5% and 5.9%, respectively.12 The data from the registry are included in the IARC (International Agency for Research on Cancer) Scientific Publications entitled “Cancer Incidence in Five Continents”.13

Preparing datasets: first step

As the first step, the data set must be prepared in a “csv file”. Microsoft Excel® (Microsoft; Redmond, WA, USA) is used to prepare a table including the following data as columns: the names of regions or their identifications, the x and y coordinates on a plane rectangular coordinate system,
longitude and latitude, the population, incidences of diseases, and the explanatory variable.

The example dataset is shown in Figure 1 and Table 1. It includes the names of the municipalities in Yamagata Prefecture as names and regions, the x, y coordinates of the municipalities on a plane rectangular coordinate system, the longitudes and latitudes of the municipalities, the population, and the incidences of diseases. As the explanatory variable, the number of doctors in the municipalities was included. The age-adjusted disease incidence was used; in Figure 1, it was calculated using the 1985 model population of Japan and the 2008 model population of Yamagata Prefecture. This dataset was saved as a “csv file” (“blood.csv” in this review).

### Preparing for the analysis: second step

These are the instructions that were used for the analysis:

- **Go to the Excel Save menu**
- **Save your worksheet file as a “csv file” (“blood.csv”) in R work directory (the work directory can be set using the preference menu of R)**
- **Close Excel**
- **Start R by double clicking on the desktop icon**
- **R shows the symbol, then expects input commands**
- **Select “Packages” from the main menu, select “Install package(s),” choose a CRAN (Comprehensive R Archive Network; <http://cran.r-project.org>) site, and select the “spdep” and “DCluster” packages to download and install.**
Preparing data frame for analysis

Type:
>library(spdep)
>library(DCluster)
>blood_OE=data.frame(Observed=blood$cases)
>blood_OE=cbind(blood_OE,Expected=round(blood$population_1000*sum(blood$cases)/sum(blood$population_1000), digits=1),x=blood$easting, y=blood$snothng)
>achisq.stat(blood_OE, lambda=1)

Output:
$T
[1] 62.3842
$df
[1] 34
$pvalue
[1] 0.00218706

Preparing coordinate data for analysis

Type:
> blood_OE=cbind(blood_OE, x=blood$easting, y=blood$snothng)
> coords=as.matrix(blood_OE[,c("x","y")])
> dlist=dnearneigh(coords, 0, Inf)
> dlist=include.self(dlist)
> dlist.d=nbdists(dlist, coords)
> col.W.tango=nb2listw(dlist, glist=lapply(dlist.d, function(x){exp(x)}), style="C")
> a=tango.test(Observed~offset(log(Expected)), blood_OE, model="poisson", R=100, list=col.W.tango, zero.policy=TRUE)

Output:
Tango's test of global clustering
Type of boots: parametric
Model used when sampling: Poisson
Number of simulations: 100
Statistic: 0.0007150427
P-value: 0.01980198

Figure 2 Instructions for Pearson’s chi-squared test and Tango’s test using R with the “spdep” and “Dcluster” packages.

Conducting the analysis using R: third step

The instructions for spatial analysis with Pearson’s chi-squared test and Tango’s test using R are shown in Figure 2. Tango’s test indicates the presence of disease clustering in hematological diseases.

FleXScan is another useful tool for spatial analysis detecting disease clustering. The results of global clustering tests using Tango’s index by FleXScan are shown in Figure 3. Instructions are available in the users’ guide, which can be downloaded from the website. A map of Yamagata Prefecture can be downloaded from the website of freemap (http://www.freemap.jp).

The impact of medical supply on disease incidence can be examined by spatial regression analysis using R with the package “spdep”. Using spatial data, whether the disease incidence as an objective variable has a relationship to the explanatory variables can be tested. The instructions are shown in Figure 4. The detailed information relating to spatial statistics and the method of spatial analysis using R have been described elsewhere.7,8,14

Usefulness of spatial statistics in hematology and oncology

In this review, spatial statistical analysis was implemented in the field of hematology using the latest techniques. All of the
tools used are available free of charge. It was demonstrated that hematology/oncology physicians can implement such an analysis in various settings using these tools to compile the data. One of the advantages of the technique used is that hypotheses on spatial clustering can be tested. This technique enables a spatial statistics investigation of disease clustering, whereas in the past, such clustering could only be estimated visually by plotting the disease incidence. This method is useful in that it enables scientific validation of the clinical impressions of patient clustering that clinicians often glean through daily clinical practice.

The present analysis showed that, when adjusted for age, clustering of hematological malignancies in Yamagata Prefecture

**Figure 3** Disease cluster analysis by Tango’s index using crude and age-adjusted disease incidences by region of Yamagata Prefecture.

Notes: Crude (A) and age-adjusted disease incidences using the 1985 model population of Japan (B) and the 2008 population of Yamagata Prefecture (C), by region of Yamagata Prefecture. Disease clusters using crude incidences are shown for Tsuruoka, Sakata, Obanazawa, Mogami, Funagata, Marumagawa, Okura, Mikawa, Shonai, and Uza (P=0.048). Disease clusters using age-adjusted disease incidences and the 1985 model population of Japan are shown for Yamagata, Kaminoyama, and Takahata (P=0.001). Disease clusters using the age-adjusted disease incidences and the 2008 population of Yamagata Prefecture are shown for Kaminoyama (P=0.001). Points and lines indicate municipalities and their contiguous areas, respectively. Disease clusters are shown by black dots with red lines.

**Figure 4** Instructions for spatial auto-regression analysis using R with the package “spdep”.

```r
Type: Library(spdep)
Coords <- matrix(0, nrow=length(blood$case_ageadj), ncol=2)
Coords[,1] <- blood$longitude
Coords[,2] <- blood$latitude
lph.tri.nb <- tri2nb(coords)
lph.lag <- lagsarlm(blood$case_ageadj~number_Dr, data=blood, nb2listw(lph.tri.nb, Style="W"))
Summary(lph.lag)
```

Output:
```
Call: lagsarlm(formula = blood$case_ageadj ~ number_Dr, data = blood, listw = nb2listw(ph.tri.nb, style = "W"))
Residuals:
       Min      1Q  Median      3Q     Max
-13.9143064 -2.4598636   0.0087099  2.4144788 11.0247369
Type: lag
Coefficients: (asymptotic standard errors)

                      Estimate Std. Error    z value  Pr(>|z|)
(Intercept)   22.98859     6.75187   3.4048   0.0006622
number_Dr  0.18274      0.32700   0.5589   0.5762635
```

**Figure 4** Instructions for spatial auto-regression analysis using R with the package “spdep”. 
showed significant accumulation of disease in Yamagata City and its environs. However, in interpreting this result, consideration must be given to the role of health care providers. Specifically, care for hematological malignancies is highly specialized, and diagnosis is difficult in medically underserved regions, such as residential areas that are far from a hospital that has a specialist physician, and there is concern that the incidence of disease might be underestimated. Even this point can be assessed with the technique of spatial analysis presented. Although the present data show that the number of hematologists in a municipality is not a factor clearly related to incidence, it would be possible to assess for each disease a variety of different variables other than the number of specialist physicians in the area, such as the number of hospitals or the number of outpatient visits to specialist hematological departments for each municipality.

A method for analyzing the method of spatial clustering of hematological malignancies is shown. Although the present analysis was performed at the municipality level, it would also be possible to use GIS data of even smaller districts, and an even more detailed spatial epidemiological analysis is also possible. However, the comprehensive acquisition of cancer information is also limited in that it is only possible to obtain data in places with a highly precise cancer registry such as Yamagata Prefecture. Even this, however, will be solved by the expansion of the cancer registration system.

The etiology of most hematological diseases has not been elucidated. Investigation of these epidemiological aspects may potentially contribute to a better understanding of the etiology of these diseases. In addition, applying the technique presented to the investigation of patient prognoses may enable generation of data that are also useful for solving health policy-related problems, such as the optimal distribution of medical resources.

Acknowledgments
This work was supported by the Institute for Regional Innovation, Yamagata University.

The authors are grateful to Professor Hiroshi Suzuki (Niigata Seiryo University, Niigata, Japan) for critical reading of the manuscript and providing useful discussion. The authors are also grateful to Hidenori Sato (Yamagata University) for support of spatial analysis using R.

Disclosure
The authors declare that they have no conflict of interest in this work.

References
8. Web page of Data Sciences for the Resilient Society. (Home page on the Internet). Available from: http://web.sfc.keio.ac.jp/~manuz/index.php?%E6%8F%B4%E6%95%85%EF%BC%9A%E5%9F%8E%E5%BC%8F%E5%9C%8B%E5%BE%8F%E5%9C%8B. Accessed February 20, 2013.