Journal of Blood Medicine

Dovepress
open access to scientific and medical research

Open Access Full Text Article

REVIEW

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

This article was published in the following Dove Press journal:
Journal of Blood Medicine
5 March 2014
Number of times this article has been viewed

Kei Kohno[1]
Hiroto Narimatsu[2]
Katsumi Otani[2]
Ri Sho[2]
Yosuke Shiono[1]
Ikuko Suzuki[1]
Yuichi Kato[1]
Akira Fukao[2]
Takeo Kato[1]

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

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 Software[3] 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

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.org

submit your manuscript | www.dovepress.com
Dovepress
http://dx.doi.org/10.2147/JBM.S57944

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 Japan[10] 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,

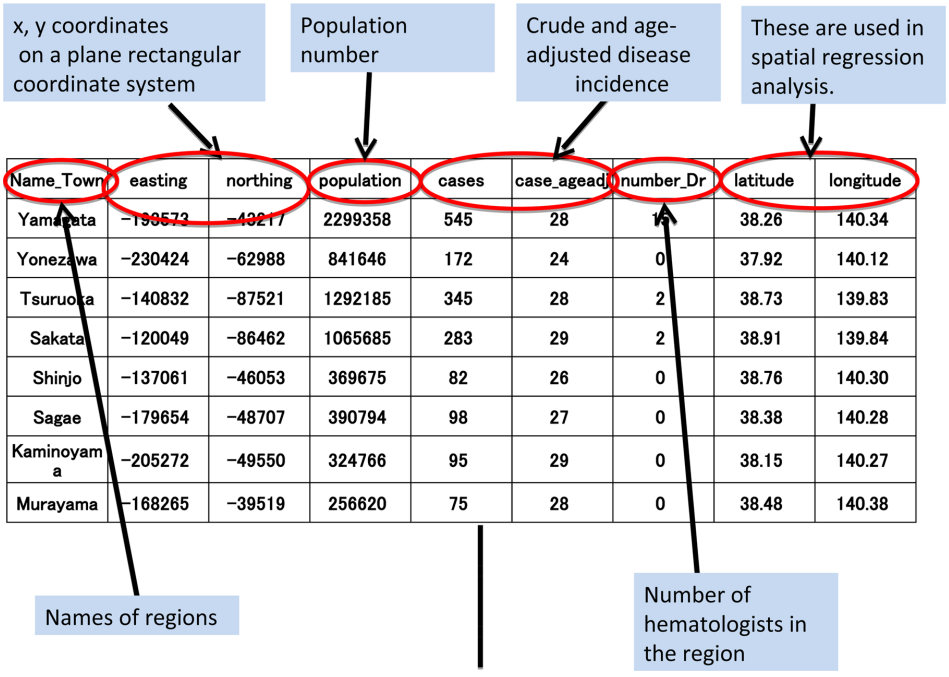

Figure 1 Preparing the dataset.

Table 1 Example data set for analysis using R in the style of "csv file" format.

Name_Town	Easting	Northing	Population	Cases	Case_ageadj	Number_Dr	Latitude	Longitude
Yamagata	−193573.0988	−43217.3042	2,299,358	545	28.389	15	38.25555556	140.3397222
Yonezawa	−230424.3634	−62988.36059	841,646	172	23.768	0	37.92222222	140.1169444
Tsuruoka	−140832.37	−87520.82925	1,292,185	345	28.114	2	38.72722222	139.8266667
Sakata	−120049.4896	−86461.856	1,065,685	283	28.612	2	38.91444444	139.8363889
Shinjo	−137061.3	−46052.9	369,675	82	25.932	0	38.76472222	140.3019444
Sagae	−179654.0025	−48707.31171	390,794	98	27.335	0	38.38111111	140.2761111
Kaminoyama	−205271.791	−49550.03961	324,766	95	29.032	0	38.14972222	140.2677778
Murayama	−168265.0176	−39519.26074	256,620	75	27.645	0	38.48361111	140.3805556
Nagai	−209778.9763	−69523.72955	279,688	59	21.521	0	38.10777778	140.0405556
Tendo	−181703.4383	−39801.64365	572,313	132	27.471	0	38.36222222	140.3783333
Higashine	−174039.9806	−38610.92744	410,802	107	30.146	0	38.43138889	140.3911111
Obanazawa	−155258.5607	−37248.51264	188,536	72	35.768	0	38.60083333	140.4058333
Nanyo	−215666.9414	−60110.67313	317,943	70	22.726	0	38.05527778	140.1483333
Yamanobe	−189765.6438	−49962.91601	138,887	32	23.648	0	38.28916667	140.2625
Nakayaka	−184893.6124	−48110.01265	112,686	25	24.301	0	38.33333333	140.2830556
Kahoku	−174560.9516	−45330.48574	188,542	33	17.634	0	38.42638889	140.3144444
Nishikawa	−174421.1478	−59933.33229	63,162	14	20.841	0	38.42666667	140.1477778
Asahimachi	−188584.8381	−60139.77135	78,745	26	28.92	0	38.29916667	140.1458333
Oe	−179566.3401	−54732.39917	90,530	26	26.658	0	38.38083333	140.2066667
Oishida	−155995.0525	−40128.01409	80,408	19	22.865	0	38.59388889	140.3727778
Kanayama	−123826.3723	−42855.32396	63,336	11	16.938	0	38.88333333	140.3394444
Mogami	−137786.8174	−27288.06881	98,093	27	28.044	0	38.75861111	140.5194444
Funagata	−145140.4643	−44670.79428	60,105	15	23.314	0	38.69166667	140.32
Mamurogawa	−126631.9239	−50427.91024	90,914	25	25.42	0	38.85777778	140.2525
Okuramura	−143684.0694	−52437.41507	38,206	9	22.35	0	38.70416667	140.2305556
Ayukawa	−133430.2608	−53099.1817	49,539	6	11.357	0	38.79611111	140.2216667
Tozawamura	−139928.2397	−59947.05624	54,411	11	19.27	0	38.73777778	140.1436111
Takahata	−221510.741	−56572.59365	236,178	67	30.49	0	38.00277778	140.1891667
Kawanishi	−221226.2899	−69153.53206	169,893	42	23.761	4	38.00444444	140.0458333
Ogunicho	−214607.455	−95645.93711	88,141	22	23.433	0	38.06138889	139.7433333
Shirataka	−201432.5264	−64381.17725	148,156	36	23.604	0	38.18305556	140.0986111
Iide	−216596.8133	−74224.76254	78,737	26	31.877	0	38.04583333	139.9875
Mikawa	−133372.6294	−85462.65233	70,855	23	32.516	0	38.79444444	139.8497222
Shonai	−127246.8596	−80613.04245	222,489	55	25.788	0	38.84972222	139.9047222
Yuza	−108994.8673	−80174.36181	154,019	59	36.389	0	39.01472222	139.9075

Notes: Data set includes the names of the municipalities in Yamagata: prefecture as names and regions including 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.
Abbreviations: ageadj, age-adjusted; Dr, doctor.

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[10] and the 2008 model population of Yamagata Prefecture.[11] 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.

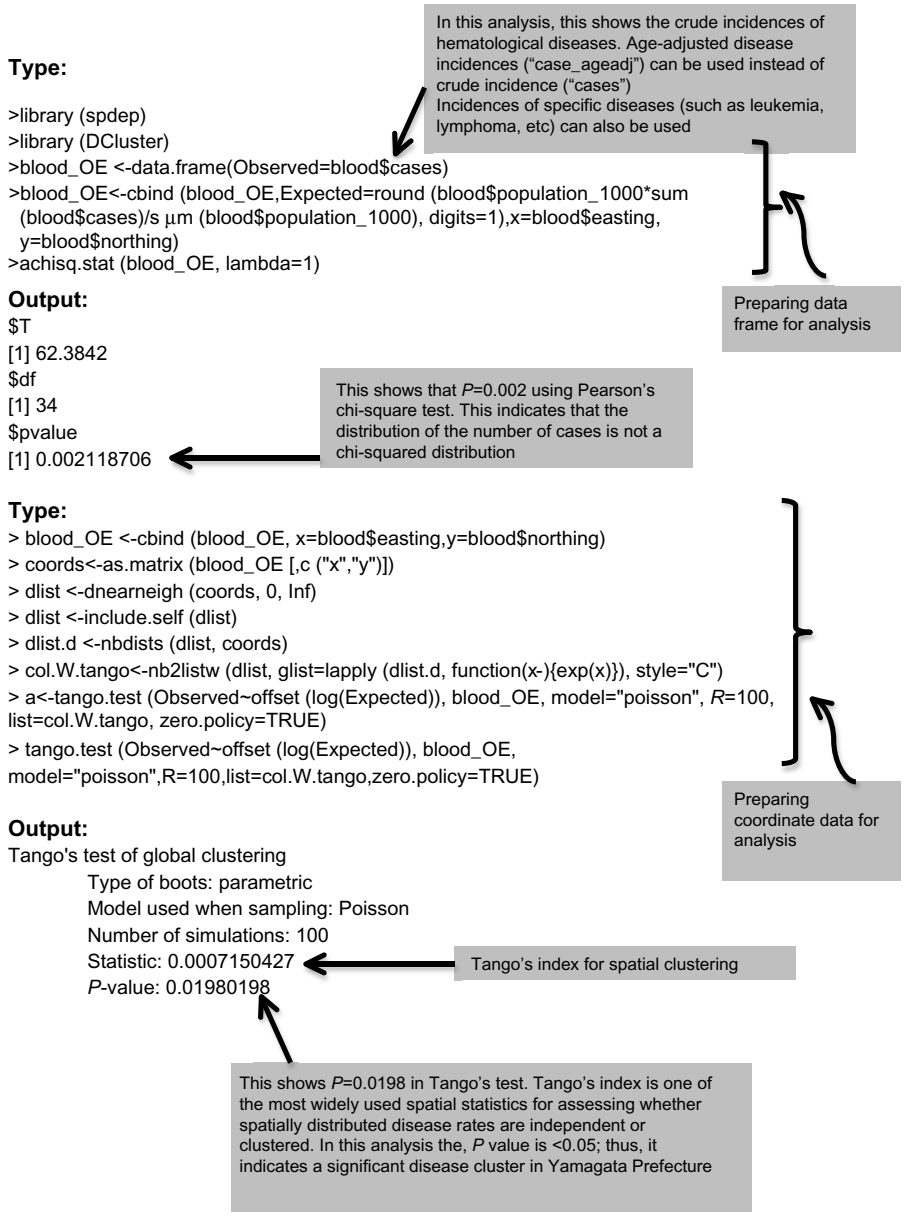

Figure 2 Instructions for Pearson's chi-squared test and Tango's test using R with the "spdep" and "Dcluster" packages.

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.[4] 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

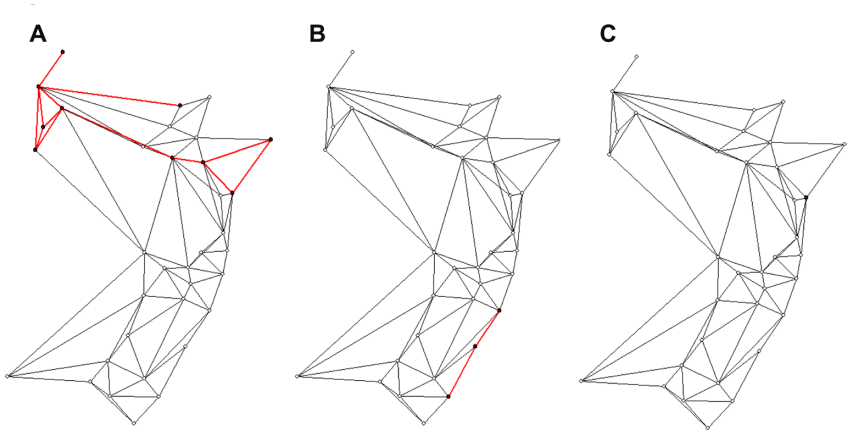

Figure 3 Disease cluster analysis by Tango's index using crude and age-adjusted disease incidences by region of Yamagata Prefecture.
Notes: Crude (**A**) and age-adjusted disease incidences using the 1985 model population of Japan (**B**) and the 2008 population of Yamagata Prefecture (**C**), by region of Yamagata Prefecture. Disease clusters using crude incidences are shown for Tsuruoka, Sakata, Obanazawa, Mogami, Funagata, Mamuragawa, 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.

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.[9] 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

Longitude and latitude are used in this analysis

```
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~num  ber_Dr, data=blood, nb2listw(lph.tri.nb,
Style="W"))
Summary (lph.lag)
```

Preparing adjacency matrix for analysis

Crude incidences ("cases") can also be used. Incidences of specific diseases (such as leukemia, lymphoma, etc.) can also be used

```
Output:
Call:lagsarlm(formula = blood$case_ageadj ~ number_Dr, data = blood,
     listw = nb2listw(lph.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
Rho: 0.098211, LR test value :0.11509, P-value: 0.73442
Asymptotic standard error :0.26125
     z-value: 0.37593, P-value:0.70697
Wald statistic: 0.14132, P-value:0.70697
Log likelihood: −106.0312 for lag model
ML residual variance (sigma squared): 25.008, (sigma: 5.0008)
Number of observations: 35
Number of parameters estimated: 4
AIC: 220.06, (AIC for lm: 218.18)
LM test for residual autocorrelation
Test value: 0.26683, P-value: 0.60546
```

In this analysis, whether disease incidence as an objective variable has relationships with explanatory variables spatially can be tested; the number of hematologists is not significantly related to disease onset in this spatial auto-regression analysis

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.[15,16] 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

1. Stevenson M, Stervens KB, Rogers DJ. *Spatial Analysis in Epidemiology*. 1st ed. Oxford, UK: Oxford University Press; 2008.
2. Diggle PJ, Ribeiro PJ. *Model-based Geostatistics*. New York, NY, USA: Springer; 2007.
3. The R Project for Statistical Computing. (Home page on the Internet). Available from: http://www.r-project.org. Accessed February 12, 2013.
4. National Institute of Public Health. (Home page on the Internet). Available from: http://www.niph.go.jp/soshiki/gijutsu/download/index.html. Accessed February 12, 2013.
5. Web page of The Quantum GIS project. (Home page on the Internet). Available from: http://www.qgis.org. Accessed February 12, 2013.
6. Takatsuki K. Adult T-cell leukemia. *Intern Med.* 1995;34(10):947–952.
7. Furuya T. [*Statistical Analysis of Spatial Data using R*]. Tokyo, Japan: Asakura Shoten; 2011. Japanese.
8. Web page of Data Sciences for the Resilient Society. (Home page on the Internet). Available from: http://web.sfc.keio.ac.jp/~maunz/wiki/index.php?%B6%F5%B4%D6%A5%C7%A1%BC%A5%BF%A4%CE%C5%FD%B7%D7%CA%AC%C0%CF. Accessed February 20, 2013.
9. Tango T. A class of tests for detecting 'general' and 'focused' clustering of rare diseases. *Stat Med.* 1995;14(21–22):2323–2334.
10. [The 1985 model population of Japan]. (Home page on the Internet). Available from: http://www.mhlw.go.jp/toukei/saikin/hw/jinkou/suii06/fuhyo.html. Accessed February 12, 2013. Japanese.
11. [Home page of Yamagata Prefectural Government]. Available from: http://www.pref.yamagata.jp/ou/kikakushinko/020052/tokei/jinkel.html. Accessed February 13, 2013. Japanese.
12. [Home page of the cancer registry in Yamagata Prefecture 2012]. Available from: https://http://www.pref.yamagata.jp/kenfuku/kenko/gan/7090005gantouroku.html. Accessed June 11, 2012. Japanese.
13. *Cancer Incidence in Five Continents*. Volume IX. IARC Scientific Publications No 160. Lyon, France: International Accreditation Recognition Council; 2007.
14. Tango T. *Statistical Methods for Disease Clustering*. New York, NY, USA: Springer; 2010.
15. [National Land Numerical Information Download Service Japan]. (Home page on the Internet). Available from: http://nlftp.mlit.go.jp/ksj/. Accessed March 12, 2013. Japanese.
16. [Portal Site of Official Statistics of Japan]. Available from: e-stat http://www.e-stat.go.jp/SG1/estat/eStatTopPortalE.do. Accessed March 12, 2013. Japanese.