World Academy of Science, Engineering and Technology
International Journal of Medical and Health Sciences
Vol:11, No:5, 2017

Exploring the Spatial Characteristics of Mortality Map: A Statistical Area Perspective

Jung-Hong Hong, Jing-Cen Yang, Cai-Yu Ou

***Abstract*—**The analysis of geographic inequality heavily relies on the use of location-enabled statistical data and quantitative measures to present the spatial patterns of the selected phenomena and analyze their differences. To protect the privacy of individual instance and link to administrative units, point-based datasets are spatially aggregated to area-based statistical datasets, where only the overall status for the selected levels of spatial units is used for decision making. The partition of the spatial units thus has dominant influence on the outcomes of the analyzed results, well known as the Modifiable Areal Unit Problem (MAUP). A new spatial reference framework, the Taiwan Geographical Statistical Classification (TGSC), was recently introduced in Taiwan based on the spatial partition principles of homogeneous consideration of the number of population and households. Comparing to the outcomes of the traditional township units, TGSC provides additional levels of spatial units with finer granularity for presenting spatial phenomena and enables domain experts to select appropriate dissemination level for publishing statistical data. This paper compares the results of respectively using TGSC and township unit on the mortality data and examines the spatial characteristics of their outcomes. For the mortality data between the period of January 1st, 2008 and December 31st, 2010 of the Taitung County, the all-cause age-standardized death rate (ASDR) ranges from 571 to 1757 per 100,000 persons, whereas the 2nd dissemination area (TGSC) shows greater variation, ranged from 0 to 2222 per 100,000. The finer granularity of spatial units of TGSC clearly provides better outcomes for identifying and evaluating the geographic inequality and can be further analyzed with the statistical measures from other perspectives (e.g., population, area, environment.). The management and analysis of the statistical data referring to the TGSC in this research is strongly supported by the use of Geographic Information System (GIS) technology. An integrated workflow that consists of the tasks of the processing of death certificates, the geocoding of street address, the quality assurance of geocoded results, the automatic calculation of statistic measures, the standardized encoding of measures and the geo-visualization of statistical outcomes is developed. This paper also introduces a set of auxiliary measures from a geographic distribution perspective to further examine the hidden spatial characteristics of mortality data and justify the analyzed results. With the common statistical area framework like TGSC, the preliminary results demonstrate promising potential for developing a web-based statistical service that can effectively access domain statistical data and present the analyzed outcomes in meaningful ways to avoid wrong decision making.

***Keywords*—**Mortality map, spatial patterns, statistical area, variation.

Jung-Hong Hong, Jing-Cen Yang, and Cai-Yu Ou are with the Department of Geomatics, National Cheng Kung University, Tainan, Taiwan (e-mail: junghong@mail.ncku.edu.tw, jingcen1022@gmail.com, pumking23@gmail.com).

I. Introduction

WITH the fast growth of GIS technology over the last few decades, the modelling of geographic phenomena advances to a new era where the geographic inequality can be readily analyzed from a spatial perspective. Compared to the table formats, a map illustrating the variation of selected phenomena at different locations provides a simple and straightforward viewpoint for making correct decisions. For example, the well-known epidemiology study made by John Snow in the mid-19th century about the outbreak of cholera [1] showed the importance and even the necessity for introducing spatial consideration to the tasks at hands. For data with statistical property, it has a good practice to publish statistical maps along with the original statistical data for showing their spatial variation after the phenomena in reality can be appropriately modeled. Further analysis, e.g. clustering, hot spots, neighborhood, also becomes possible with the GIS technology. Especially with the recent development of internet technology, the sharing of spatial resources enables the researchers to easily access cross-discipline data and to explore new findings that were extremely difficult or almost impossible in the past.

Data serve as the core for the GIS; that is, regardless how many models and operations come with GIS, the outcomes are dominated by the data chosen. The geographic inequality can be conclusively identified via GIS if the data collected and analyzed correctly modeled the phenomena. For some sensitive applications, the location of individual instances is prohibited to be publicly released due to privacy concerns [5]. Aggregating such data to coverage-based data is an often adopted approach. Since the accurate location is not provided, the correct interpretation of such coverage-based data requires in-depth knowledge about how the coverage structure influences the presented outcomes [12]. In public health, mapping has been an important tool to explore spatial patterns of incident cases [9]. In mortality domains, in addition to the death certificates of individual cases, the mortality data are often presented by the values of quantitative measures (e.g., mortality rate) to the regions of interests (e.g., individual county) to examine their differences. A time-series of mortality statistical data helps to identify places with significant patterns that need further attention (e.g., continuously increasing, high mortality rate). As the data are established with respect to the areal units, choropleth map is a natural choice for geographically presenting the outcomes. Despite widely used, one major limitation for choropleth maps is the framework of spatial units to which the data are referred. To be more specific, a choropleth map merely presents the reality according to the selected spatially partitioned units, and the results may be rather different if a different set of spatial units is used. The abrupt changes across the boundary of two neighboring spatial units do not necessarily imply that such difference exits in reality. The choice of framework for spatial units is well known as the MAUP. Any attempt to interpret and analyze the coverage-based statistical maps without considering the impacts of MAUP is risky with unpredictable threats.

The development of different spatial reference frameworks makes the integrated analysis across a variety of disciplines a complicated task. It is therefore advantageous to have a common spatial framework endorsed by the cross-discipline stakeholders. In order to solve this problem, some countries have developed framework with smaller spatial units to address different application needs, e.g. the census tract of United States [4] and the census subdivision of Canada [13]. The Ministry of Interior in Taiwan launched a project in 2012 for developing a common hierarchical framework of spatial units for presenting and aggregating domain statistical data for long term analysis and policy making. Before TGSC was officially released in the year of 2012, the socio-economic data in Taiwan were mainly collected at the city/county and village/township level. While the boundary of city and county normally remains unchanged, the boundary of village and township may continuously change. Consequently, the statistical outcomes were either too general (based on larger spatial units like city and county) or too unstable (based on continuously changing spatial units). TGSC aims to serve as a robust and consistent framework for spatially partitioned units, such that the socio-economic data from the MOI will refer to a common spatial reference for integrated analysis, as well as to provide a way to link to other domains that are willing to transform or establish their data by complying to the TGSC.

Due to the MAUP consideration, the statistical data based on the TGSC have its unique characteristics and limitation for presenting the geographic inequality. In addition to exploration and comparison of the spatial characteristics of mortality data respectively based on the areal units of administrative units and TGSC, we further propose auxiliary measures for showing the hidden spatial characteristics for the distributed features. The remaining of the paper is organized as follows. Section II first introduces the design of TGSC, then discusses its spatial characteristic; Section III discusses the uses of GIS for data processing, management, and visualization; Section IV explores and analyzes the outcomes from different perspectives; and finally, Section V concludes our major findings and future research directions.

II. Spatial Characteristic

The development of thematic cartography has fostered the recognition of space, attribute, and time as the basic components of mapping [2]. Via the consideration of space, time, and theme (attributes), a variety of maps are produced for visually presenting the phenomena of geographic inequality to address domain needs. Serving as a common framework of spatial units for the whole territories, TGSC consists of seven levels of spatial units, where the statistical area serves as the bottom level of the whole architecture, and the other six levels of dissemination area (named from 1st to 6th) are aggregated from the level of statistical area in a hierarchical fashion. The 3rd, 4th, 5th, and 6th levels are further binding to the current levels of administration units. This implies that the TGSC, though recently introduced, can present statistical phenomena with finer granularity, as well as work with the current administrative units (e.g., city and county), so a time series analysis together with the historical statistical data is still possible. The spatial extent of an individual spatial unit at the upper level is aggregated by a number of spatial units in the lower level. Such hierarchical relationship enables domain researchers to generate and disseminate statistical results at the most appropriate levels by aggregating the results from the raw data or lower levels of statistical area.

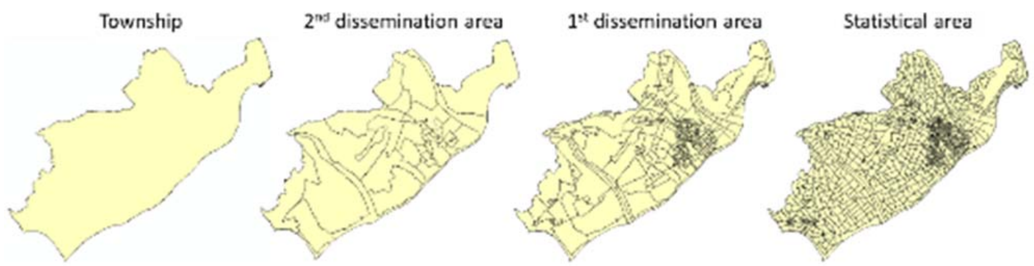


Fig. 1 TGSC framework

Based on a particular TGSC level of statistical data, theoretically one choropleth map can be generated to illustrate the geographic inequality. For example, Fig. 2 shows a portion of mortality map for the Taitung County. The color difference is used for showing the different levels of quantitative measured values associated with each statistical area. As the spatial partition principle of TGSC is mainly based on the homogeneity consideration on the number of population and households, the variation on the size of the statistical area is very significant. For example, the maximum number of people for a statistical area (the bottom level) is 450, so the statistical areas or dissemination areas in urban, rural, and mountain area are clearly different. Various types of natural or manmade features, such as road networks, ridges, and rivers, were selected as the boundary of statistical area. In general, TGSC provides a better granularity of spatial partition and enables the integration of cross-domain data, as well as the protection of the citizens' privacy.

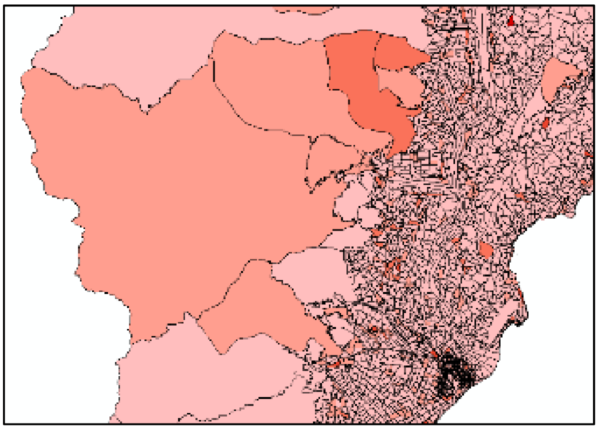


Fig. 2 Mortality rate map

The consideration of MAUP implies that the spatial partitions and their aggregations will affect the statistical data and the decisions made upon them. Two types of effects for MAUP have been extensively discussed in the past: the scale effect and the zoning effect [7]. The scale effect refers to the different consequences for the same source presented by spatial units of different spatial resolutions, while the zoning effect refers to the different outcomes due to the different choices of aggregations [6]. Both effects have impact on the presentation of the outcomes of the geographic distribution and the analysis of the geographic inequality. Several approaches were proposed to reduce the influences of MAUP, e.g., assessing the appropriate zoning or scaling system [8], designing homogeneous areal units based on the people's activities for optimizing the spatial distribution of an area-level exposure [11].

Using statistical area as the basic unit for presenting geographic distribution can protect the privacy of individual features, but unfortunately it also hides the details of geographic distribution of the features. For example, the total counts of the features in the two statistical areas in Fig. 3 are the same, the reading and interpretation on the two subfigures on the right, however, are obviously different from each other. The areal presentation thus fails to show the clustering phenomena, and there is no way for the users to gain such understanding with only the choropleth maps. Additional auxiliary measures are proposed in Section III to address this issue.

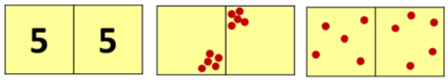


Fig. 3 Examples of point-based data distribution

III. GIS Perspective

A. Data Process Workflow

The major source of the mortality data is the death certificate issued for each mortality case. The street address information extracted from the death certificate provides the geo-reference to the location for individual mortality case. By submitting the street address to the national geocoding service from MOI, we acquire the coordinates and the identifiers for each mortality case. Due to the type error and the lack of invalid street address no longer used, some of the mismatched cases must be handled manually. The geocoded mortality cases are then assigned to the chosen levels of spatial units (administrative units, statistical area and dissemination area) with the point-in-polygon test. Following the workflow depicted in Fig. 4, two mortality datasets based on administrative units and 2nd level of dissemination area were generated.

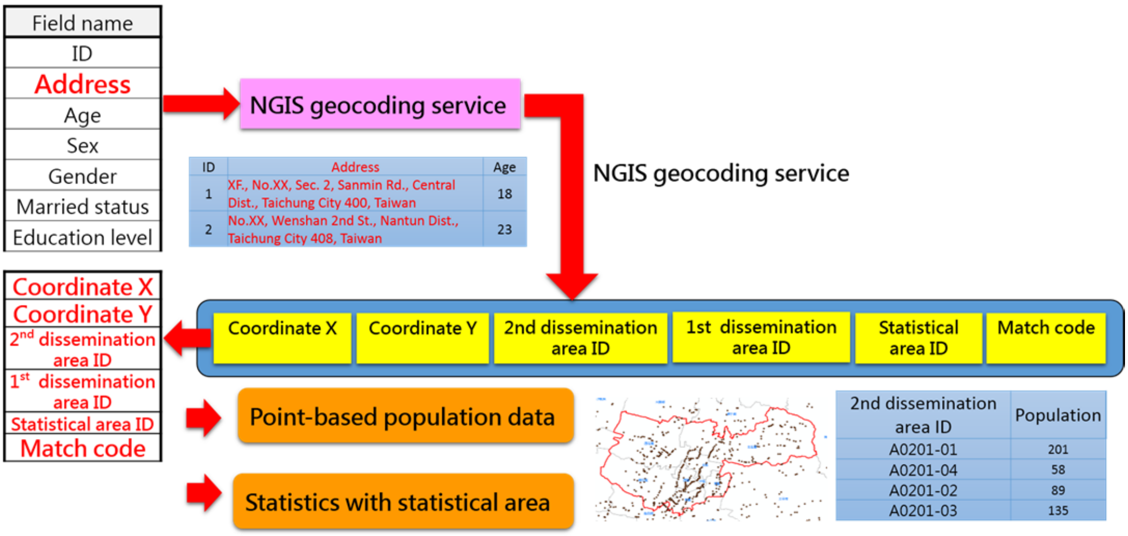


Fig. 4 Data process workflow-Geocoding service

B. Auxiliary Measures

In GIS, the attributes of the areal representation are often used to indicate the overall status within the specified region, e.g. the count of features and feature density. By doing so, the actual geographic distribution of features is hidden to the user and sometimes may therefore limit the analysis that users can execute. Without disclosing the location of individual features, we propose to use auxiliary measures to improve users' understanding about the features with the spatial units. First of all, we propose to use the geometric center of the features within the spatial unit as an approximate location of the features. Since there are many different forms of geographic distribution, we further use the spatial dispersion index proposed by Weng and Tsai in 2006 [10] to assess the degree of dispersion or clustering of features in the spatial units. The value of SDI is based on the area of statistical unit, comparative area of surface features, and distance to neighbor. The smaller the value is, the more clustered the features are. The combination of these two measures helps to identify the pattern where the features are clustered and located at a particular place within the spatial unit. For example, the features in the statistical unit shown in Fig. 5 are clustered at the corner of the spatial units. We can easily identify this situation with the geometric center of the spatial unit, the geometric center of the features, and the SDI value without the location of individual mortality cases. This implies that we can protect the privacy, and in the meanwhile, we can still provide meaningful information about the geographic distribution of the features.

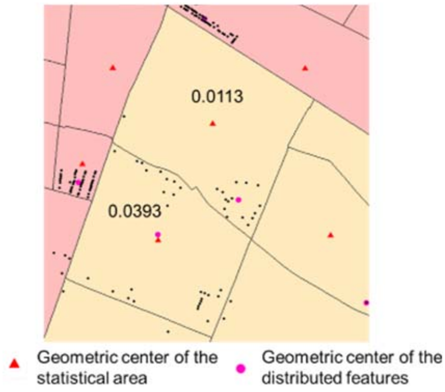


Fig. 5 Examples of auxiliary indicator

C. Different Levels of Spatial Unit

As shown in Fig. 6, with the same indicator, different scales of spatial partition would be totally different. With the most meticulous statistical area, small area is able to present the spatial pattern. Once the mortality rate data aggregated into larger levels of dissemination area, the geographic phenomena would be eliminated. From this example, it is obvious that the scale effect truly influence the result of spatial pattern.

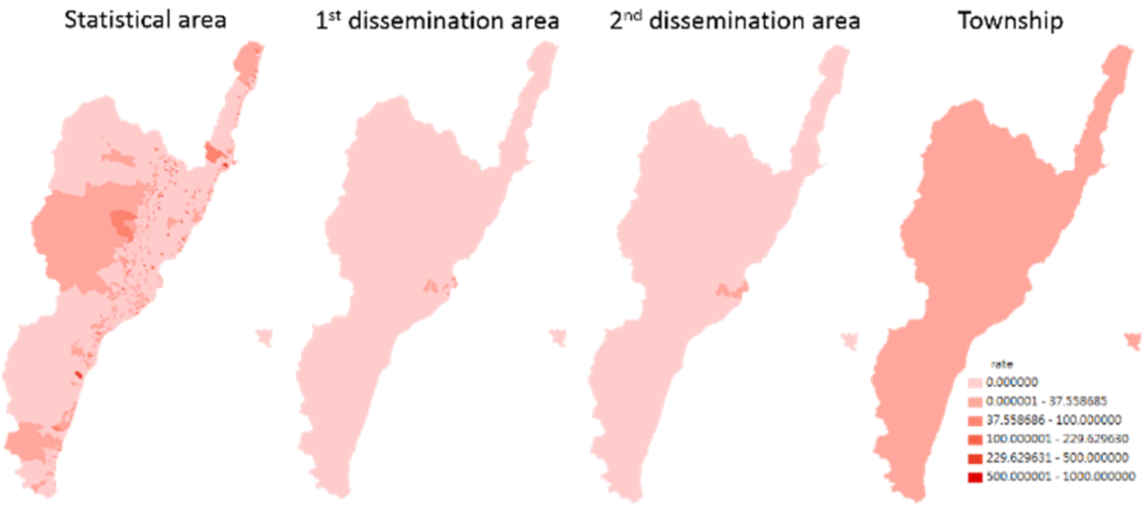
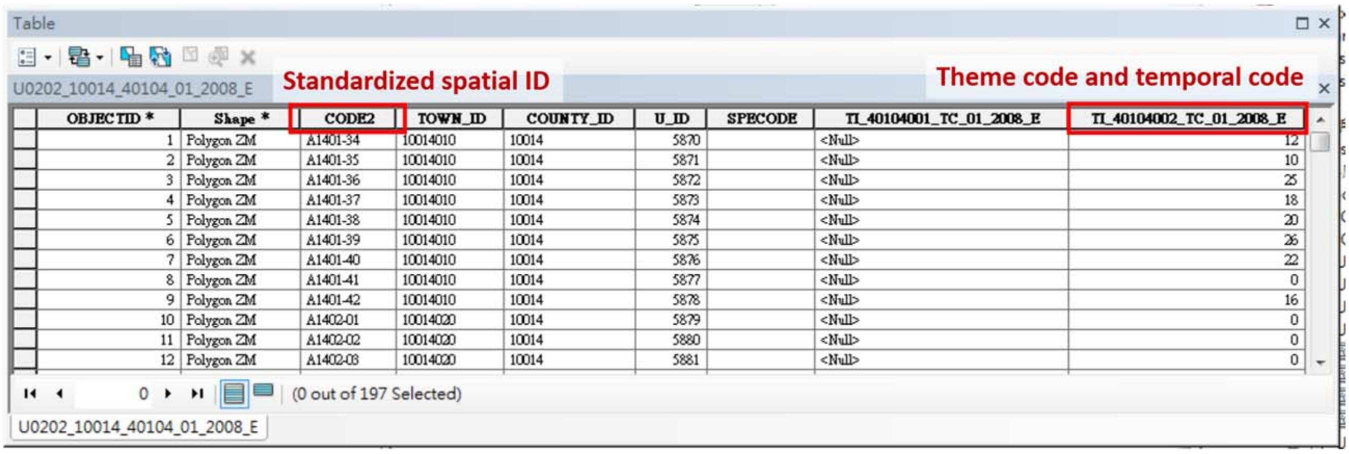


Fig. 6 Different scales of mortality rate map

D. Unique Identifier

The development of a standardized ID system enables to ensure the quality and consistent spatial data to all spatial data users [3]. To ensure the correct use, avoid wrong linking, and better management of statistical data, a standardized coding system for the cross-discipline statistical data is proposed. Every piece of information being used includes three aspects: space, theme, and time. This research designs standardized codes to manage the spatial statistical data for each component. As each row typically represents the quantitative measured results of a spatial unit at a certain time, every table by default has an attribute to record the standardized spatial ID, the unique identifier of TGSC spatial units in this research. All of the other attributes represent a particular quantitative measure for a given time. The naming of the attributes is the combination of two standardized codes, the theme measure code, and the temporal code. The temporal codes must be able to address different time scenarios for describing the temporal aspects of the data. While the temporal codes can be applied to any domains, the theme codes are normally domain dependent, and each domain can develop their own code system. Fig. 7 uses the "death statistics data in 2008" in this research as an example, the code of "TI "means time instant which is used to describe the time mode of statistical data. The code of "40104002" and "TC" represent the total count of death statistics data. The designing principle is to describe data theme as well as statistical method. The code of "01_2008" stands for the time resolution is in the year of 2008. Finally, the code of "E" indicates that the death statistics data are collected at the end day of 2008.



OBJECTID *	Shape *	CODE2	TOWN_ID	COUNTY_ID	U_ID	SPECODE	TI_40104001_TC_01_2008_E	TI_40104002_TC_01_2008_E
1	Polygon ZM	A1401-34	10014010	10014	5870		<Null>	12
2	Polygon ZM	A1401-35	10014010	10014	5871		<Null>	10
3	Polygon ZM	A1401-36	10014010	10014	5872		<Null>	25
4	Polygon ZM	A1401-37	10014010	10014	5873		<Null>	18
5	Polygon ZM	A1401-38	10014010	10014	5874		<Null>	20
6	Polygon ZM	A1401-39	10014010	10014	5875		<Null>	26
7	Polygon ZM	A1401-40	10014010	10014	5876		<Null>	22
8	Polygon ZM	A1401-41	10014010	10014	5877		<Null>	0
9	Polygon ZM	A1401-42	10014010	10014	5878		<Null>	16
10	Polygon ZM	A1402-01	10014020	10014	5879		<Null>	0
11	Polygon ZM	A1402-02	10014020	10014	5880		<Null>	0
12	Polygon ZM	A1402-03	10014020	10014	5881		<Null>	0

Fig. 7 Death statistics data in 2008-unique identifier

The crude mortality rate is calculated by number of death data in a year and population data at midyear. Its rule can be easily represented by the standardized codes in the following:

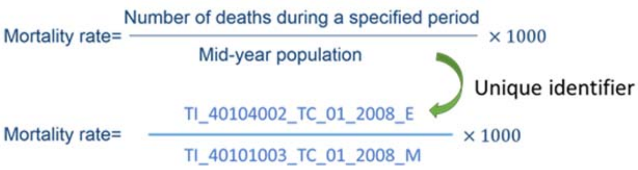


Fig. 8 Use standardized codes to represent mortality rate

After all the individual attributes have standardized codes, the time series data, the related measures, and the cross-discipline measures can all be easily handled and even automatically updated via the help of GIS.

IV. Result and Analysis

The county of Taitung is chosen as the test site in this research. Taitung County is located in the south-eastern Taiwan with a population of 206,302 in the year of 2008. Geographically speaking, the east side of the Taitung County is the Pacific Ocean, and the majority of the western part is covered by the mountain. Following the workflow discussed in Section III, the mortality data in a 3-year period (between January 1st, 2008 and December 31st, 2010) are created, and two datasets respectively based on the township and 2nd dissemination area of TGSC are also obtained.

Fig. 9 shows the map for all-cause ASDR in Taitung County. With its finer granularity, the results of the 2nd dissemination area clearly show more meticulous pattern. The ASDR value for the township-level spatial units is from 571 to 1757 per 100,000 persons, whereas that of the 2nd dissemination area shows greater variation, ranged from 0 to 2222 per 100,000. It is also obvious that some of the spatial units are identified as higher mortality rate, while no such phenomena are observed from the township data.

One of the major reasons for the difference between the township and 2nd dissemination area is clearly the spatial partition. The number of spatial units for the 2nd dissemination area is 12 times more than that of the township. Furthermore, the average area size of the township is 12 larger than that of the 2nd dissemination area. Overall, the average population and number of death of 2nd dissemination area are less than the township. Without any doubt, the spatial units of the 2nd dissemination area can provide a much more detailed description as far as the geographic distribution is concerned.

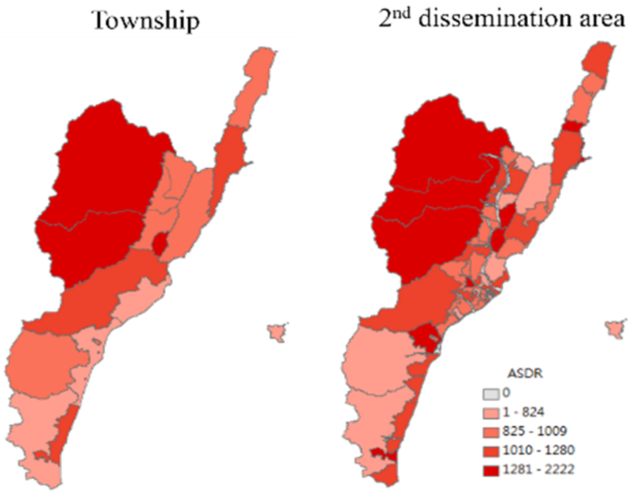


Fig. 9 ASDR of township and 2nd dissemination area

TABLE I
Summary of ASDR

Township		2nd dissemination area	
Min.	Max.	Min.	Max.
571	1757	0	2222

TABLE II
Total Count

Township	2nd dissemination area
16	197

TABLE III
Area

Township				2nd dissemination area			
Min	Max	AVG	SD	Min	Max	AVG	SD
162.9	8936.9	2233.6	2206.3	0.0002	6065.5	181.4	683.7

The ASDR for 2nd dissemination area is subcategorized into five different intervals using natural break operations. Table VI shows the summarized results for applying the interval sets to the two mortality datasets. Tables VII and VIII show the area, average of area, standard deviation of the area, and CV. A significant difference can be found in the results of level 3. The TGSC spatial units which belong to the level 3 appear to be smaller units with relatively homogeneous sizes. But, such phenomena can be found from the township data. However, there is an indicator which provides relative measure of dispersion namely coefficient of variation. The 2nd dissemination area shows higher coefficient of variation. The result of area size implies that the variance of 2nd dissemination area is much higher than township. Fig. 10 shows the spatial units with the highest level of ASDR values. While there are only two towns whose ASDR reaches the highest level, 13 TGSC spatial units pass the specified threshold. As shown in Fig. 10, the TGSC spatial units which belong to the level 5 identify a number of smaller spatial units with higher ASDR values. However, for the spatial units with higher ASDR value in mountain area, there is no significant difference. Fig. 11 shows the TGSC spatial units belonging to the level 3.

TABLE IV
POPULATION-FROM 2008 TO 2010

Township				2nd dissemination area			
Min	Max	AVG	SD	Min	Max	AVG	SD
9845	328285	43439	74684	0	11484	3225	3970

TABLE V
NUMBER OF DEATH-FROM 2008 TO 2010

Township				2nd dissemination area			
Min	Max	AVG	SD	Min	Max	AVG	SD
51	2328	382	530	0	220	31	42

TABLE VI
LEVELS OF ASDR

Level	ASDR	Total count		Percentage (%)	
		Township	2nd dissemination area	Township	2nd dissemination area
1	0	0	109	0	55
2	1-824	5	29	31	15
3	825-1009	6	29	38	15
4	1010-1280	3	20	19	10
5	1281-2222	2	10	13	5

TABLE VII
AREA SIZE-TOWNSHIP

Level	ASDR	SUM (km2)	AVG	SD	CV
1	0	-	-	-	-
2	1-824	6172.17	1234.43	1028.27	83.30
3	825-1009	9379.83	1563.31	1017.59	65.09
4	1010-1280	6494.13	2164.71	1569.80	72.52
5	1281-2222	13691.63	6845.82	2091.12	30.55

TABLE VIII
AREA SIZE-2ND DISSEMINATION AREA

Level	ASDR	SUM (km2)	AVG	SD	CV
1	0	580.99	5.33	16.48	309.19
2	1-824	8835.24	304.66	797.99	261.93
3	825-1009	3703.49	127.70	130.88	102.49
4	1010-1280	7727.23	386.36	625.10	161.79
5	1281-2222	14890.82	1489.08	2084.17	139.96

Although the TGSC of 2nd dissemination area shows more detail in smaller area, the reading based on choropleth map still limits the users' perception about the geographic distribution. Overlapping the geocoded mortality cases with the spatial coverage of the 2nd dissemination area shows that the geocoded mortality data have its own geographic distribution (Fig. 12) and it is almost impossible to envision such distribution by simply looking at the choropleth maps. The proposed auxiliary measures are thus used to provide more information about the distribution.

Because of the privacy protection policy, the location of mortality cases may not be presented to the users. For the spatial units of A1401-07 in Fig. 13, the mortality cases are mainly located at the western part of the spatial units. The geometric center of features presents a good indication to the approximate location of the features. Since its SDI value is 0.1734, we can generally conclude a clustering situation with respect to the whole spatial extent of the spatial unit. So even if the location of mortality cases is not shown, the geometric centers of features from the neighboring spatial units may be used for evaluating the geographic distribution of mortality cases.

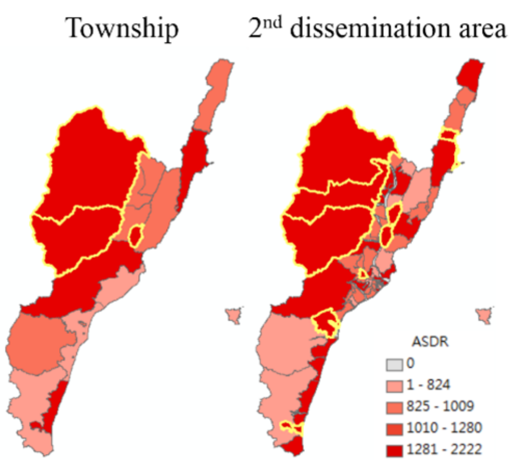


Fig. 10 Level 5 spatial units (township and TGSC)

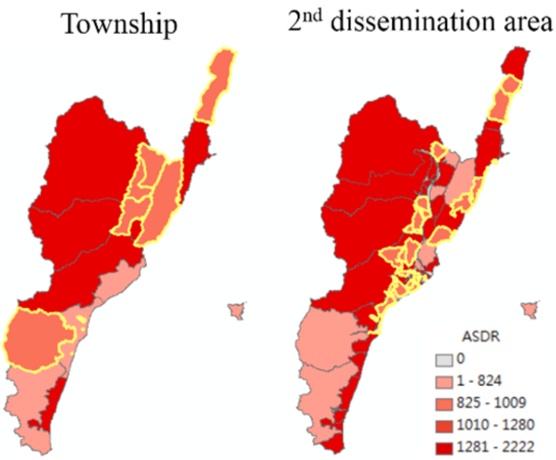


Fig. 11 Level 3 spatial units (township and TGSC)

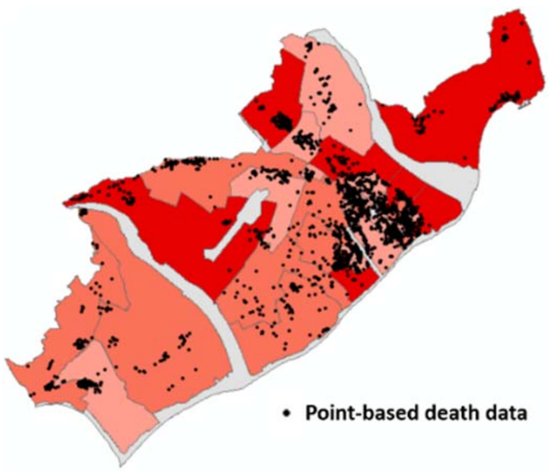


Fig. 12 Overlap point-based death data with 2nd dissemination area in Taitung city

V. Conclusion

The release of TGSC offers a new opportunity for integrating cross-discipline statistical data in Taiwan. More in-depth analysis is still necessary to explore the spatial characteristics of TGSC for making correct interpretation. For the ASDR data, the results based on townships and the 2nd dissemination area of TGSC show significant difference in terms of the values and the spatial distribution. The ASDR of township ranges from 571 to 1757 per 100,000 persons, whereas the 2nd dissemination area shows greater variation, ranged from 0 to 2222 per 100,000. The 2nd dissemination area cannot only show the variation of individual units, but it also identifies those places with zero population. This helps to provide a closer look to the real mortality situation in the region. The ASDR for 2nd dissemination area is subcategorized into five different intervals using natural break operations. From the result of area, average of area, standard deviation, and coefficient of variation, the TGSC spatial units belonging to the level 3 are smaller units with relatively homogeneous sizes. In addition, the TGSC spatial units which belong to the level 5 identify a number of smaller spatial units with higher ASDR values. However, for the spatial units with higher ASDR value in mountain area, no significant difference is found. In brief, the 2nd dissemination area presents more meticulous spatial distribution of ASDR than the results based on the township. In addition, the auxiliary measures help to identify spatial patterns that may be missing during aggregation. With the use of GIS, after the geocoded mortality cases are created, a whole series of datasets of statistical area and dissemination areas can be automatically created, managed, and distributed.

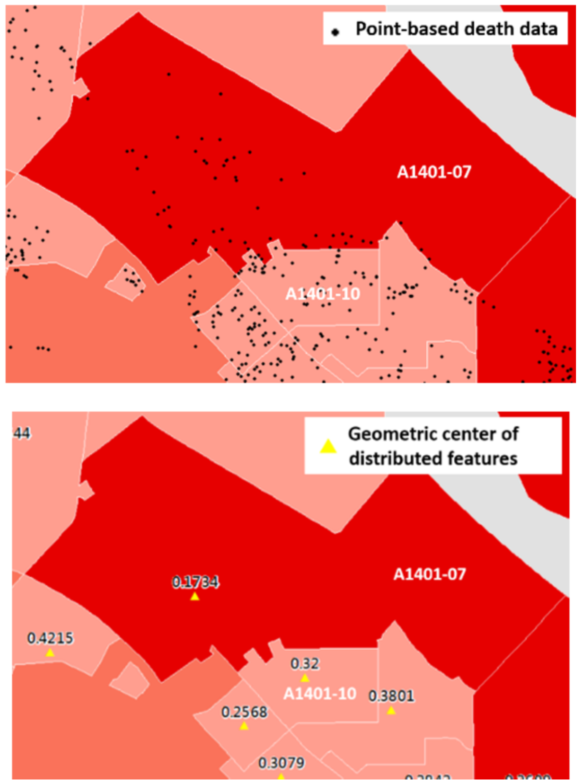


Fig. 13 Application of auxiliary indicator

Acknowledgment

This paper is partial result from the research project (MOST 104-2627-M-006-004) granted by the Ministry of Science and Technology in Taiwan.

References

[1] Brody, H., Rip, M. R., Vinten-Johansen, P., Paneth, N., & Rachman, S. (2000). Map-making and myth-making in Broad Street: the London cholera epidemic, 1854. The Lancet, 356(9223), 64-68.

[2] Chrisman, N. R. (1991). The error component in spatial data. Geographical information systems, *1*, 165-174.

[3] Cope, M. A., & Pincetl, S. (2014). Confronting Standards and Nomenclature in Spatial Data Infrastructures: A Case Study of Urban Los Angeles County Geospatial Water Management Data. *IJSDIR*, *9*, 36-58.

[4] Krieger, N., Waterman, P., Chen, J. T., Soobader, M. J., Subramanian, S. V., & Carson, R. (2002). Zip code caveat: bias due to spatiotemporal mismatches between zip codes and us census–defined geographic areas—the public health disparities geocoding project. American Journal of Public Health, 92(7), 1100-1102.

[5] Nelson, J. K., & Brewer, C. A. (2015). Evaluating data stability in aggregation structures across spatial scales: revisiting the modifiable areal unit problem. Cartography and Geographic Information Science, 1-16.

[6] Openshaw, S., & Taylor, P. J. (1979). A million or so correlation

coefficients: three experiments on the modifiable areal unit problem. Statistical applications in the spatial sciences, 21, 127-144.
[7] Openshaw, S. (1984). The modifiable areal unit problem. Concepts and Techniques in Modern Geography, 38(41).
[8] Openshaw, S. (1996). Developing GIS-relevant zone-based spatial analysis methods. Spatial analysis: modelling in a GIS environment, 55-73.
[9] Pickle, L. W., Mungiole, M., Jones, G. K., & White, A. A. (1999). Exploring spatial patterns of mortality: the new atlas of United States mortality. Statistics in Medicine, 18(23), 3211-3220.
[10] Pei-Wen Weng, & Bor-Wen Tsai (2006). Spatial Dispersion Index: Old Conception, New Formula, Journal of Taiwan Geographic Information Science, 4, 1-12.
[11] Riva, M., Gauvin, L., Apparicio, P., & Brodeur, J. M. (2009). Disentangling the relative influence of built and socioeconomic environments on walking: the contribution of areas homogenous along exposures of interest. Social Science & Medicine, 69(9), 1296–1305.
[12] Stafford, M; Duke-Wilson, O; Shelton, N. (2008). Small area inequalities in health: are we underestimating them? Soc. Sci. Med, 67, 891–899.
[13] Schuurman, N., Bell, N., Dunn, J. R., & Oliver, L. (2007). Deprivation indices, population health and geography: an evaluation of the spatial effectiveness of indices at multiple scales. Journal of urban health, 84(4), 591-603.