Thematic Geo-visualization for Social Data Representation in Special Region of Yogyakarta

Sudaryatno*[a], Totok Wahyu Wibowo[a], Zulfa Nur'Aini 'Afifah[a], Shafiera Rosa El-Yasha[a], Achmad Rofi'i[a]
[a]Department of Geography Information Science, Universitas Gadjah Mada, Bulaksumur, Sleman, Yogyakarta 55281, Indonesia

ABSTRACT

Data related to social activities in Indonesia mostly used statistical data. Statistics for large numbers of social will make it difficult to interpret and analyze because it consists of many columns and rows with each value. Geo-visualization is a visualization of data represented in a geographic coordinate system. Social statistics could be visualized to ease the process of spatial analysis data that considers a spatial surface of the earth. The study area is in the Special Region of Yogyakarta. This study aims to (1) Select, test and find out color symbol scheme most effective classification method for choropleth mapping of Demographic Map, (2) Mapping happiness profile of population using small area estimation method, (3) Analyzing tourist trends based on Instagram data using space-time cube visualization. Secondary data used are population and happiness, while primary data uses social media data for tourist visualization. Geo-visualization of population and happiness used the choropleth method. In social media geo-visualization for tourists using space-time cube geo-visualization with hexagonal tessellation cells. The results obtained are population maps with the best classification scheme, happiness maps at different scale levels, and tourist map using space-time cube in Yogyakarta Special Region.

Keywords: Demography, Geo-visualization, Space-time cube, Social Data

1. INTRODUCTION

Indonesia is a broad country which has various ethnic groups and cultures. Community social activities would be varied according to their culture. A pattern of social activities has an impact on economic activities especially on the income sector for daily essentials. In Indonesia, data about social use of statistical data. Statistics for large numbers of social data will make it difficult to interpret and analyze because it consists of many columns and rows with each value.

Geo-visualization is data visualization represented in a geographic coordinate system. Social statistics could be visualized to ease the process of spatial analysis data that considers the spatial surface of the earth. The geo-visualization of social data has its primacy in studies that could be integrated with the physical surface of the earth. This supports the geographic idea that claims the physical surface of the earth affects social activities and vice versa. This research shows three different things in representing social data.

Requirements in delivering information in the form of maps are higher in various fields of studies. All field of studies requires effective data presentation to make a decision. Good decision making based on maps requires great understanding however not all fields of studies are familiar with map usage. Classes of data classification results is one of the support to help in understanding map usage.

The mapping method is constantly improving along with using social statistic to balance the lack of mapping which carried out conventionally using census and survey data. Survey and census obtained from a specific range of duration and usually presented in a province or regency/city level with a classification of urban and rural areas. Comprehensive data presentation up until sub-district level or rural/urban village is not possible since it will need time and cost. On the other hand, the mapping of social conditions in the society needed for the governments' programs right on target. Elbers et al (2003) uses its primacy of census and survey data to create a prediction model. This model aimed to estimate the value of the smallest geographic area which accounted for statistically. This method called the Small Area Estimation.

*sudaryatno@ugm.ac.id

Sixth Geoinformation Science Symposium, edited by Sandy Budi Wibowo, Andi B. Rimba
Stuart Phinn, Ammar A. Aziz, Proc. of SPIE, Vol. 11311, 113110H · © 2019 SPIE
CCC code: 0277-786X/19/$21 · doi: 10.1117/12.2547320

Tourists' data extraction using big data Instagram is one of the ways to find out tourist spatial patterns effectively and efficiently as well as able to scope all areas in real-time. People's tendency to share photos on holidays in Indonesia is many. Hexagonal tessellation is a geometry iteration to present overlapping data points by generalizing according to the scales. Areas that have similar near-by points were clearly through its aggregate of hexagonal tessellation. However, hexagonal tessellation is a discreet visual tessellation therefore, a variation of hexagonal sizes will directly influence tourists' geo-visualization. Hexagonal tessellation has yet explored in mapping activities despite it could be good for geo-visualization (Birch, et al 2007). Hexagonal tessellation geo-visualization used in tourists' density analysis associated with tourists' attractions however for spatial-temporal analysis is difficult because it involves time dimension (Kang, et al 2018). Space-time cube geo-visualization is data spatial representation which offers three dimensions of geo-visualization which are spatial (x, y) and time dimension (z). Hence, tourists' geo-visualization using space-time cube will make the analysis easier based on spatial and temporal.

2. METHODOLOGY

2.1 Classification Scheme Testing

Data that will be visualized in the choropleth map is population density of Special Region of Yogyakarta obtained from Statistics Indonesia in Special Region of Yogyakarta province. Supporting data to ease visualizing the Special Region of Yogyakarta province is in the form of shape-file format obtained from Geospatial Information Agency.

Steps of data classifying are by deciding the types of data to adjust the data classification method, limiting the total classes by using Sturgess method, where n is the total data (similarity 1) and limiting the total classes by deciding the lowest classes limit and followed by other limits which are the continuation of the lowest class limits. Limits of total classes have a condition where the decision of interval classes based on the limits of classes taken, interval classes cannot be repeated and all interval classes were filled.

Classification scheme testing uses some classification method. Classification method used are constant interval classification, arithmetic classification, geometric classification, quintile classification, standard deviation classification and dispersal graph classification method.

$$\text{Total Class} = 1 + 3{,}3\log(n) \tag{1}$$

2.2 Small Area Estimation

Estimation calculation carried out on the urban and rural provincial level. This is because census data only available in village aggregate level (Podes 2018) and happiness survey (SPTK 2017) sample schemes made to represent data up to rural and urban provincial level only. Estimation model equation (beta model) by Elbers et al (2003) developed using the following equation (2). GLS regression calculations needed because it uses OLS to get "β" in equation 2.

$$\ln Y_{ch} = X_{ch}\,\beta + \mu_{ch} \tag{2}$$

Where $\ln Y_{ch}$ = logarithm of overall satisfaction life scale variable village-h cluster-c, c = subscript for province cluster, h = subscript for rural and urban village-h on cluster-c, X_{ch} = household characteristic village-h cluster-c, μ_{ch} = vector of disturbance. Location variables is filled with locational variance ($\mu_{ch} = \eta_c + \varepsilon_{ch}$) where η_c is error level in cluster term and εch is error level in the household term.

$$\ln \tilde{y}_{ch} = x_{ch}\,\tilde{\beta} + \tilde{\eta}\;\; c + \tilde{\varepsilon}_{ch} \qquad \text{where } \beta \sim N(\hat{\beta}, \hat{\Sigma}_{\beta}) \tag{3}$$

The last step is to simulate the model to census data with bootstrap a hundred times to obtain estimation value and standard error. The simulation model defined as in equation (3). To obtain the result in a smaller area, census data must have a complete hierarchy ID to the area needed. Variables have to be available in both data and measured with the same method. Further analysis of the variable definition and statistic comparison required on each variable between census and survey data. For further explanation, see Elbers et al (2003). Explanatory variables that contribute significantly selected by a stepwise regression method. The happiness threshold of overall life satisfaction scale (range 1-10) set to score 5.

Census and survey data compiled based on the hierarchy ID (table 1). Hierarchy ID consists of cluster-ID and household ID. The household ID is a unique ID for each household sample. The household ID used as a truncation ID in PovMap 2.0 software. In this research, we use FGT0 to decide spatial profiles. FGT0 shows a percentage population that has a happiness level below threshold. Estimation results on the model classified into two classes i.e. Unhappy and Happy. Classification using the geometrical interval method. The classification model used to produce a profile map and to decide the spatial characteristics of happiness profiles in the Special Region of Yogyakarta.

Table 1. Hierarchy ID Model.

Hierarchy for Urban		Hierarchy for Rural	
ID	Description	ID	Description
13400000000	Yogyakarta Province	23400000000	Yogyakarta Province
13471000000	Yogyakarta City	23401000000	Kulon Progo Regency
13471010000	Gedongtengen Sub-district	23401010000	Temon Sub-district
13471010001	Pringgokusuman Village	23401010001	Temon Kulon Village

Source: Village Master File 2010 and Author' interpretation

2.3 Space-time Cube Geo-visualization

Space-time cube geo-visualization as a geo-visualization that accommodates spatial-temporal data. Hence, the data needed to cover three dimensions are spatial and temporal. In this research, data Instagram used. The process of obtaining the data uses a web-based application called Netlytic. These data go through preprocessing to produce the data without noise. The preprocessing process is a process of erasing iteration data and advertisements which text keywords. This process uses QGIS software. Data that went through preprocessing will then be visualized using a space-time cube to see the distribution pattern based on time tendency. Geo-visualization created using ArcGIS 10.5 application with space-time cube tool Geo-visualization able to see the time tendency and distribution which shows time pattern of each bin through temporal pattern classification results in two dimensions. The duration range is from 28 August 2018 until 11 December 2018. The interval bin used is weekly. Weekly time duration based on context and data availability. The context in tourism has different visitation of tourists each day. Ideally, tourists visit on the weekends which produces a balanced trend of minimal time which is suitable weekly. In the context of data availability, the data obtained are 3 months old, meaning time possibilities are daily, weekly and hourly. Based on two considerations, a suitable interval bin is weekly whereas the tessellation size used is 1 Km. This size is assumed that tourists' spot in the Special Region of Yogyakarta area is not big enough with the size of 1 Km.

3. RESULT AND DISCUSSION

3.1 Classification Scheme Testing

The most effective classification method tested based on the proportion test. All of the total population arranged based on the amount. The value of population density is a result of processing by using the total population divided based on the Special Region of Yogyakarta area of the administration data vector. The population density value of each sub-district has a unit of "population/km2". The population density value of each sub-district is then arranged from the lowest to highest to make it easier for classification.

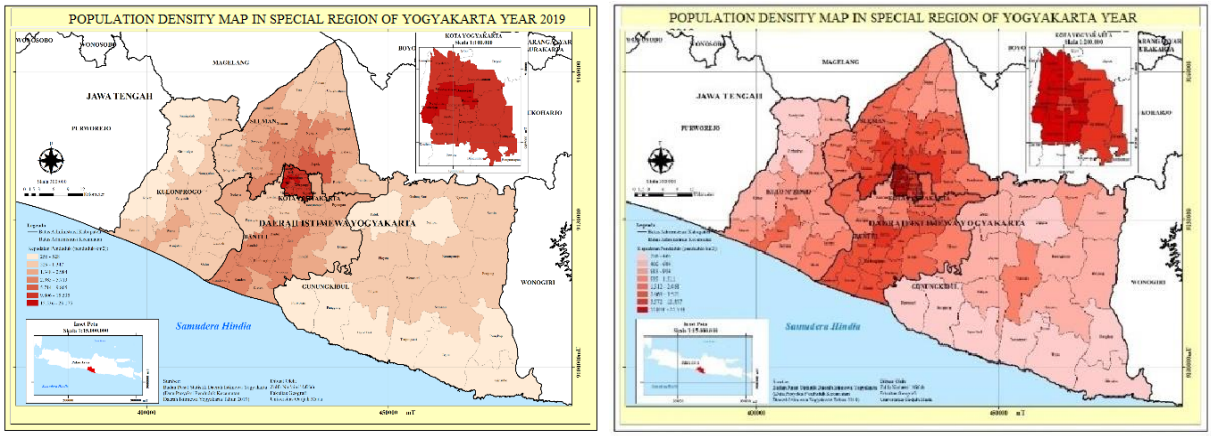
Classification methods used to simplify the sub-districts level with a similar population density value. Each classification methods have different lower and upper limits in each classification class. This caused members of each classification results classes different. Result obtained from the proportion test in Table 2. Based on the values obtained, the classification method with the lowest proportion value is Arithmetic Interval with 0.26. The lowest proportion value shows that the classification method can represent the classified data well and suitable with classification classes where each population density value of each sub-districts should be. Whereas the highest value is the Quantile classification method with 0.30. Based on the value, it was expected for the population density value to be well-distributed spatially hence it can be seen to be more equal. Figure 1 is a result of mapping with color symbol schemes which are similarly used in the arithmetic interval and quantile classification method as a comparison spatially.

Table 2. Proportion Testing Result of Classification Scheme.

Result of Proportion Testing	Constant Interval	Arithmetic Interval	Geometric Interval	Quantile	Standard Deviation	Dispersal Graph	Most Effective Classification Method According to The Lowest Proportion Value
Total	22.60	20.35	20.71	23.74	21.85	22.85	
Average	0.29	0.26	0.27	0.30	0.28	0.29	Arithmetic Interval: 0.26

Source: Author' calculation

Arithmetic Interval was able to cover seven classes of population density in the Special Region of Yogyakarta because quintile covers eight classes of population density. Introduction of population density classes with total classes above 5 should use a maximal of seven classes and have to still be attempted for mapping to get a result of total odd classes. Total odd classes are represented in the map to ease map users in knowing population density classes. If there are seven classes, then there is one middle class with three upper classes and three lower classes therefore, it will be easier for map users to compare than eight classes on the map. Additionally, spatial distribution of population density in arithmetic interval is visually more equal and easier to compare between classes than quintile.



(a) (b)

Figure 1. Population Density Map in Special Region of Yogyakarta using Arithmetic Interval Classification Scheme (a) and Quantile Classification Scheme (b).

3.2 Small Area Estimation

Statistics Indonesia (BPS) released the happiness index of the Special Region of Yogyakarta (DIY) in the year 2017 with a percentage of 72.93% and the eighth place as the happiest province in Indonesia. Publication of happiness index is only presented in the rural and urban province level. The small area estimation method by Elbers et al (2003) is used to map out the 2017s happiness index in the Special Region of Yogyakarta Province until the rural level by using a calculation model of province level. This calculation involves two significant house-hold variables for the urban estimation model and six significant house-hold variables for the rural estimation model.

Estimation calculation differentiated into two which are urban province estimation and rural province estimation (Elbers et al, 2003; Suryahadi et al, 2015). Table 3 shows part of the calculation results of small area estimation from province to rural level. The estimation result is in the form of FGT0 (headcount index) and standard error on each scale. Despite the result of standard error is small along with low hierarchy, this is not fully correct. This is caused by a total of census data used on simulation step that needs to consider significant variable regression result on a model, as well as total and survey samples distribution of Special Region of Yogyakarta Province which determined by BPS, are not enough to processed to rural level.

Table 3. Estimation Result of Small Area Estimation Calculation.

Urban			Rural		
ID	FGT0 (%)	SE	ID	FGT0 (%)	SE
13400000000	0.07	0.25	23400000000	1.29	0.0071
13471000000	0.07	0.38	23401000000	1.6	0.0152
13471010000	0.67	0.0469	23401010000	0.53	0.0182
13471010001	0	0	23401010001	0	0

Source: Author' interpretation

Figure 2 shows the maps as a result of small area estimation calculations on various map scales. Map visualization follows the result of estimation calculation which is separated according to urban and rural classification. Classification using the geometric interval classification method separated into two classes. Based on the estimation of each scale and province-level until rural show pattern of urban area dominating percentage classification of people living under the scale of life satisfaction is lesser compared to the rural area. Regency/city level estimation seen based on visualization, have not much difference compared to province level. This is due to urban and rural classification in the level of regency/city is not much different from the province level. Rural level estimation shows detailed disperse pattern of which village included in the first or second class.

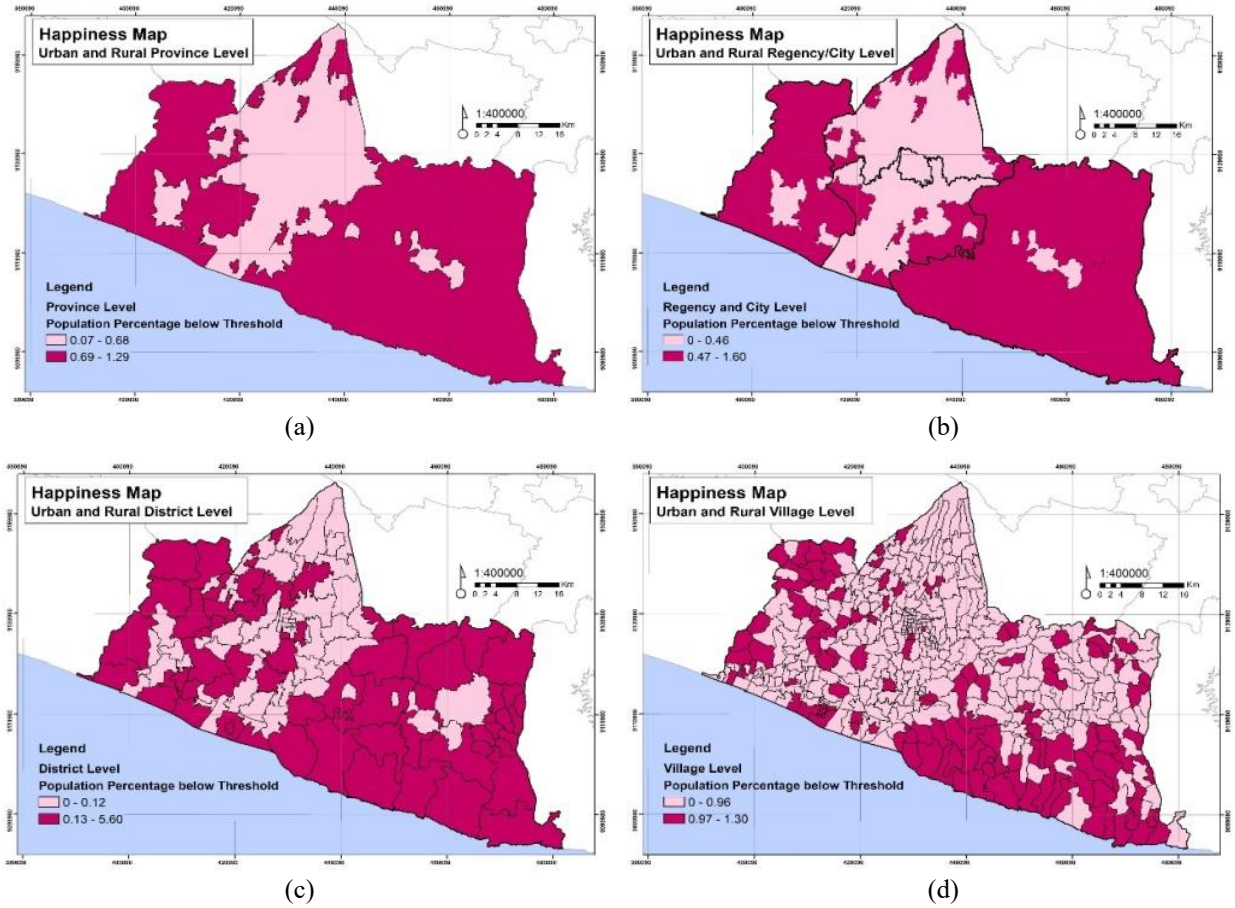


Figure 2. Urban and Rural Happiness Map based on Individual Proportion below threshold (FGT0) in (a) Province level (b) Regency and City level (c) District level (d) Village level.

3.3 Space-time Cube Geo-visualization

Result of big data Instagram geo-visualization for tourists shows a few numbers of tourists' attractions which are significant statistically. This condition shows that not all tourists share their photos on Instagram. The result of tourists is different from tourists in general. The border of tourists on the research is tourists which visited and shared content on Instagram. Therefore, a tourist attraction on the research is different from the most popular tourist attraction base on Instagram data. The result of tourist attraction from Instagram data extraction is something that people find good to share on Instagram (Instagramable). Figure 3 shows the tourists map using space-time cube geo-visualization. Three space-time cube patterns that are detected are a consecutive hotspot, sporadic hotspot, and no pattern. Tourists' attractions classification using space-time cube with the most distribution is no pattern. This is because not all tourists share their photos on Instagram. Hotspot identified on tourist maps using space-time cube geo-visualization is a consecutive hotspot and sporadic hotspot. Consecutive hotspot is a hotspot that happened at the end of the bin and no other previous hotspot while sporadic hotspot is an irregular hotspot or random pattern.

Based on the spatial-temporal map of Figure 3, sporadic hotspot pattern is in Malioboro area. Whereas consecutive hotspot patterns are seen in three locations which are Malioboro, Depok and The Lost World Castle. Malioboro is a tourist attraction area highly visited due to the various tourist spots such as shopping centers, historical building and clean cities with local culture. The Lost World Castle is a tourists' attraction that offers the beauty of flowers and castles. Consecutive hotspot patterns in Depok, if analyzed further is a shop building. This means that the account is an advertisement account which has not been filtered perfectly during the preprocessing process. As a result, tourist geo-visualization using space-time cube based on Instagram data cannot be filtered perfectly.

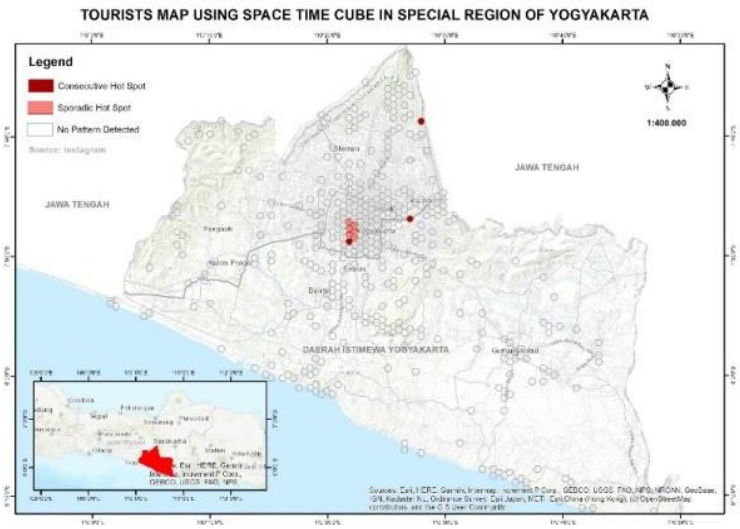


Figure 3. Tourists' Map using Space-time Cube.

Based on the overall statistic space-time cube calculation, the time tendency resulted each week caused the amount of tourists to increase. Trend-z values of 8.7113. The total time steps 15 bin with weekly interval, counted from 28 August 2018 until 11 December 2018 or equals to 106 days. Malioboro is an area with significant hotspot based on space-time cube geo-visualization. Hotspot phenomena in Depok Village in Table 3.1 which shows the space-time cube statistic calculation. Depok Village has a tessellation code 162 with a total sum_value of 27 posts. Malioboro has a tessellation code 197 with sum_value of 27 posts. The Lost World Caste has a tessellation code 31 with a sum_value of 39 posts. sum_value comparison of tourists' attractions with hotspot patterns has a high amount, except for code 162 and 197. The smallest value is 0 post and a maximum value of 24 and 14 posts. This means that many phenomena of no visitors in one week based on big data Instagram. Total tessellation identified is 404. However, only 9 tessellations identified as a hotspot. This shows tourists' mapping using Instagram data only covers Instagram users who share their photos in social media and share spatial information and does not cover older tourists and children.

Table 4. Result of Space-time Cube Calculation.

Objectid	Pattern	Perc_Hot	Trend_Z	Trend_P	Sum_Value	Min_Value	Max_Value
162	Consecutive Hot Spot	1.081	2.802	0.005	27.000	0.000	24.000
197	Consecutive Hot Spot	8.108	8.767	0.000	27.000	0.000	14.000
153	Sporadic Hot Spot	9.189	7.745	0.000	99.000	0.000	18.000
169	Sporadic Hot Spot	9.189	9.254	0.000	755.000	0.000	100.000
186	Sporadic Hot Spot	9.189	9.250	0.000	1684.000	0.000	215.000

Source: Author' calculation

4. CONCLUSION

The geo-visualization of social data in the Special Region of Yogyakarta shows arithmetic interval classification method is the most effective classification method based on proportion test because it produces an overall lowest proportion of 0.26. Estimation of each scales in the province to village levels shows urban area pattern dominate percentage classification of the population living under life satisfaction scale is lower compared to the rural area. Space-time cube geo-visualization using Instagram data can be a solution in tourist mapping especially in a wide area. Tourists in the Special Region of Yogyakarta have increased, especially in Malioboro and The Lost World Castle in 3 months.

REFERENCES

[1] Al-Ghamdi, Ali M., "Optimising the Selection of a Number of Choropleth Map Classes," Lecture Notes in Geoinformation and Cartography: Thematic Cartography for the Society, 61-78, (2014).

[2] Badan Pusat Statistik, [Indeks Kebahagiaan 2017], CV. Dharmaputra, Jakarta, ISBN 978-602-438-145-5, (2017).

[3] Barudin, Fitriyani, I. A., & Indriati, D., "Statistik Profil Wisatawan Nusantara Tahun 2016," Jakarta Pusat, Badan Pusat Statistik, (2016).

[4] Birch, C. P., Oom, S. P., & Beecham, J. A., "Rectangular and Hexagonal Grids Used for Observation, Experiment, and Simulation in Ecology," Elsevier, 347 - 359 (2007).

[5] Boy, J. D., & Uitermark, J., "Reassembling the city through Instagram," Transactions of the Institute of British Geographers, 612-624, (2017).

[6] Dharmawan, R. D., Suharyadi, & Farda, N. M., "Geovisualization Using Hexagonal Tessellation for Spatiotemporal Earthquake Data Analysis in Indonesia," Soft Computing in Data Science, 177-187, (2017).

[7] Elbers, Chris., Lanjouw, Jean O., Lanjouw, Peter., "Micro-Level Estimation of Poverty and Inequality," Econometrica, 71, (1), 355-364. https://doi.org/10.1111/1468-0262.00399 (2003).

[8] ESRI, "Space-time pattern mining," Retrieved from ArcGIS Pro: http://pro.arcgis.com/en/pro-app/tool-reference/space-time-pattern-mining/GUID-A9E88279-7DB2-4DFF-B096-D265513E98D2-web.png (6 May 2019)

[9] Harun, M., & Syam, H. M., "Pengaruh Penggunaan Media Sosial Instagram terhadap Perubahan Sikap Tujuan Wisata Liburan Mahasiswa Universitas Syiah Kuala," Jurnal Ilmiah Mahasiswa FISIP Unsyiah, 3 (2), (2018).

[10] Instagram, "Tourism", Retrieved from Instagram: https://www.instagram.com/explore/tags/wisata/ (6 October 2018).

[11] Kang, Y., Cho, N., & Son, S., "Spatiotemporal Characteristics of Elderly Population's Traffic Accidents in Seoul Using Space-Time Cube and Space-Time Kernel Density Estimation," PLOS ONE (2018).

[12] Kraak, Menno-Jan dan Ormelling, Ferjan. [Kartografi dan Visualisasi Data Geospasial], Yogyakarta, Gadjah Mada University Press, (2007).

[13] Kurniati, Erna dan Rahardjo, Noorhadi., "Evaluasi Metode Klasifikasi Dalam Pembuatan Peta Kepadatan Penduduk DIY Dengan Permukaan Statistik Dan Uji Proporsi," Jurnal Bumi Indonesia, (2012).

[14] Makhabel, B., Mishra, P., Danneman, N., & Heimann, R., "R: Mining Spatial, Text, Web, and Social Media Data," Birmingham, Packt Publishing Ltd, (2017).

[15] Rahardjo, Noorhadi., “Penggunaan Metode Permukaan Statistik untuk Penyusunan Peta Kepadatan Penduduk di Daerah Istimewa Yogyakarta,” Yogyakarta, Fakultas Geografi Universitas Gadjah Mada, Thesis, (1984)
[16] Rahayu, Theresia Puji., “The Determinants of Happiness in Indonesia,” Mediterranian Journal of Social Sciences vol 7(2). DOI 10.5901/mjss.2016.v7n2p393 (2016).
[17] Suryahadi, Asep., Wenefrida Widyanti., Daniel Perwira, et al., “Developing a Poverty Map for Indonesia: An Initiatory Work in Three Provinces,” Technical Report, SMERU Research Institute (2003).
[18] Wei, Ran, et al., “An integrated classification scheme for mapping estimates and errors of estimation from the American Community Survey,” Computers, Environment and Urban Systems, 95-103.
[19] Vu, H. Q., Li, G., Law, R., & Ye, B. H., “Exploring the travel behaviors of inbound tourists to Hong Kong using,” Tourism Management, 222-232, (2013).
[20] Yuliasih, Eko., Susanto, Irwan., “Determining Poverty Map Using Small Area Estimation Method,” Seminar Nasional Matematika 2010, Proceeding, Sebelas Maret University (2010).
[21] Zhao, Qinghua., Lanjouw, Peter., “Using PovMap2 a User's Guide,” World Bank, http://iresearch.worldbank.org/PovMap/PovMap2/PovMap2Manual.pdf.