International Journal of Architectural, Civil and Construction Sciences
ISSN: 2415-1734
Vol:11, No:4, 2017

Forthcoming Big Data on Smart Buildings and Cities: An Experimental Study on Correlations among Urban Data

Yu-Mi Song, Sung-Ah Kim, Dongyoun Shin

***Abstract*—**Cities are complex systems of diverse and inter-tangled activities. These activities and their complex interrelationships create diverse urban phenomena. And such urban phenomena have considerable influences on the lives of citizens. This research aimed to develop a method to reveal the causes and effects among diverse urban elements in order to enable better understanding of urban activities and, therefrom, to make better urban planning strategies. Specifically, this study was conducted to solve a data-recommendation problem found on a Korean public data homepage. First, a correlation analysis was conducted to find the correlations among random urban data. Then, based on the results of that correlation analysis, the weighted data network of each urban data was provided to people. It is expected that the weights of urban data thereby obtained will provide us with insights into cities and show us how diverse urban activities influence each other and induce feedback.

***Keywords*—**Big data, correlation analysis, data recommendation system, urban data network.

I. Introduction

CITIES are made up of many elements including people, buildings, natural environments, infrastructure, and others. Activities in cities have complex relationships and generate a diversity of urban phenomena that influence the lives of citizens positively or negatively. Diverse urban phenomena occur via the clear causal relationships among urban elements [1]; multiple elements influence each other within a complex network. The relations among multiple elements can have hidden relations that people cannot perceive. And it is very difficult to reveal such relations of urban elements intuitively. Therefore it is a significant challenge to find the hidden relations among urban elements. One key to solving this problem is big data. Many urban phenomena occurring in cities are stored as data, which are often offered free to citizens. By analysis of urban data, people can understand the influences and complex relations among urban phenomena. If methods to reveal the complex relationships among diverse urban elements can be derived, urban activities will be better understood, and therefore also, urban planning strategies will become possible.

Y.-M. Song is with the Department of Convergence Engineering for Future City, Sungkyunkwan University, Suwon, Republic of Korea (e-mail: hanimyu@skku.edu).

S.-A. Kim is with the Department of Architecture, Sungkyunkwan University, Suwon, Republic of Korea (e-mail: sakim@skk.edu).

D. Shin is with the BTU, Department of Architectural Engineering, Sungkyunkwan University, Suwon, Republic of Korea (corresponding author, phone: +82 1063745354; e-mail: dongyoun79@gmail.com).

II. State of the Art of Urban Big Data

Many studies have applied big data to cities. Urban data are collected through sensors installed on a person's devices or in cities. In this way, the embedding of computers in the fabric of urban life has given rise to the notion of the 'smart city' wherein municipal functioning is supported by massive datasets [2]. Certainly, there is a role for big data in efforts to improve sustainability and overall living standards in cities. Thus, the current availability of smart devices that generate large heterogeneous datasets and the smart applications that offer seamless connections among various objects and individuals have been researched in efforts to realize the dream of the smart city. To this end too, comprehensive and in-depth analyses of state-of-art technologies have been performed, and suitable big-data structures have been presented [3]. Other research has explored how big data can be useful in urban planning by formalizing the planning process as a general computational problem. They resolved the dilemma between the need for planning and its impossibility in detail by recognizing that cities are foremost self-organizing social networks embedded in space and enabled by infrastructure and services of cities [4]. Another recent paper proposed a combined IoT-based system for smart city development and urban planning based on big-data analytics. The proposed system consists of eight steps that begin with data generation and end in decision making. In that study, the system was tested and evaluated with respect to efficiency measures in terms of throughput and processing time [5]. However, cities are complex spaces with integrated social, cultural, and technical aspects; utilizing of big data in such contexts, therefore, is problematic. One relevant study in this regard compared two countries in discussing issues of big data as they relate to urban research. It looked at the process of change of datasets in urban research and explored opportunities for big data. Then, it discussed the issues and solutions for the two cities [6]. Another study conducted systematic research in an attempt to understand the complex relations among urban data. It then provided, based on the concept of data vitalization, a data-correlation framework within which the correlation characteristics of data can be understood and expressed via a data-correlation diagram [7].

III. Research Questions

Even though various studies seeking to understand urban complexities by means of urban big data have been undertaken,

it is still a very challenging task. In this context, the present study pursued an experiment, specifically an urban big-data analysis, by analyzing the correlations among real urban data in Seoul. When a user obtains the necessary data from a Korean government website [8], the associated data are recommended and provided [Fig. 1]. Currently, such recommendations and the associated data network among urban big data are not generated by data mining but rather by simple categorization of a data source department. Therefore, the current recommended data are based only on the tag entered by the site administrator when the data are uploaded. This is a simple recommendation method that displays data that are similar to the searched data. But this kind of recommendation is meaningless, because it does not take into account the data correlations. So, sometimes there is a problem in that data is recommended that actually are not relevant.

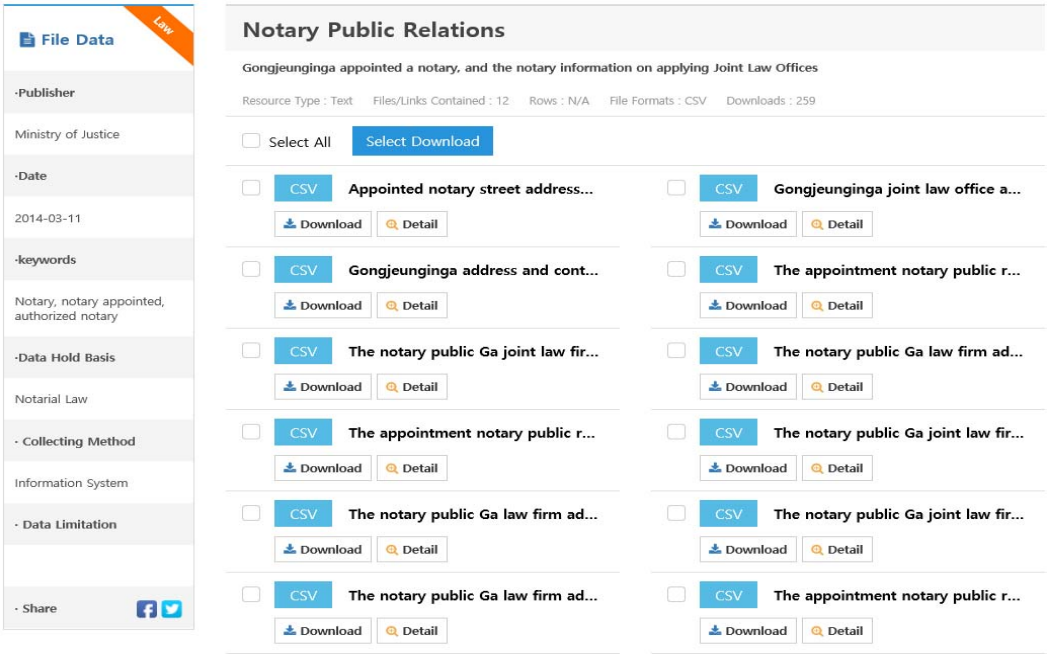


Fig. 1 Current data recommendations of Korean public data website

IV. Research Method

This study aimed to solve the problem of the data recommendations found on the Korean public data homepage referenced above and to find a means of recommending more relevant data to people.

First, a correlation analysis was performed to find the correlations among random urban data. It was proved possible to understand highly relevant data and to recommend more relevant data to people. Based on this analysis, a data-connection map was drawn. Then, the results of the correlation analysis could be used to provide insight into urban activities. The importance of any data can be determined based on the correlations among data collected in a city. This makes it easy to understand the effects of data in the areas in which it is collected. The weights of data are defined by correlation analysis. As a result, it becomes easier for citizens to understand huge and complex urban networks with their manifold ordered relations.

Data recommendations and weight derivations via correlation analysis of urban big data are expected to stimulate the interest of citizens to improve their urban environment and to provide clear insight for urban planning purposes.

This research used public data provided by the Seoul Metropolitan Government [9]. Korean municipal-administrative districts are in the following order: ‘do’, ‘si’, ‘gu’, ‘dong.’ ‘Do’ is a global concept and ‘dong’ is a local concept. The City of Seoul has 25 gu. Approximately 30 datasets for each gu were randomly extracted. This means that the data could have certain relations, but that these were not affected in the data selection. Then, a correlation analysis was performed to find the relations among the diverse urban datasets. The purpose of this study was not actually to focus on finding the causal relationships among certain urban phenomena but rather to find the associations among the diverse urban datasets. In this context, correlation analysis is very effective, as it can identify relationships among data that are not easily predicted. RapidMiner Studio (RM) was used as the analysis tool in the present study. RM [10] is effective in any visual design environment for rapidly building complete predictive analytic workflows with various machine-learning algorithms. RM has a simple UI (user interface), and so even non-experts have the advantage of being able to easily analyze data by application of complicated machine-learning algorithms.

V. Results of Correlation Analysis

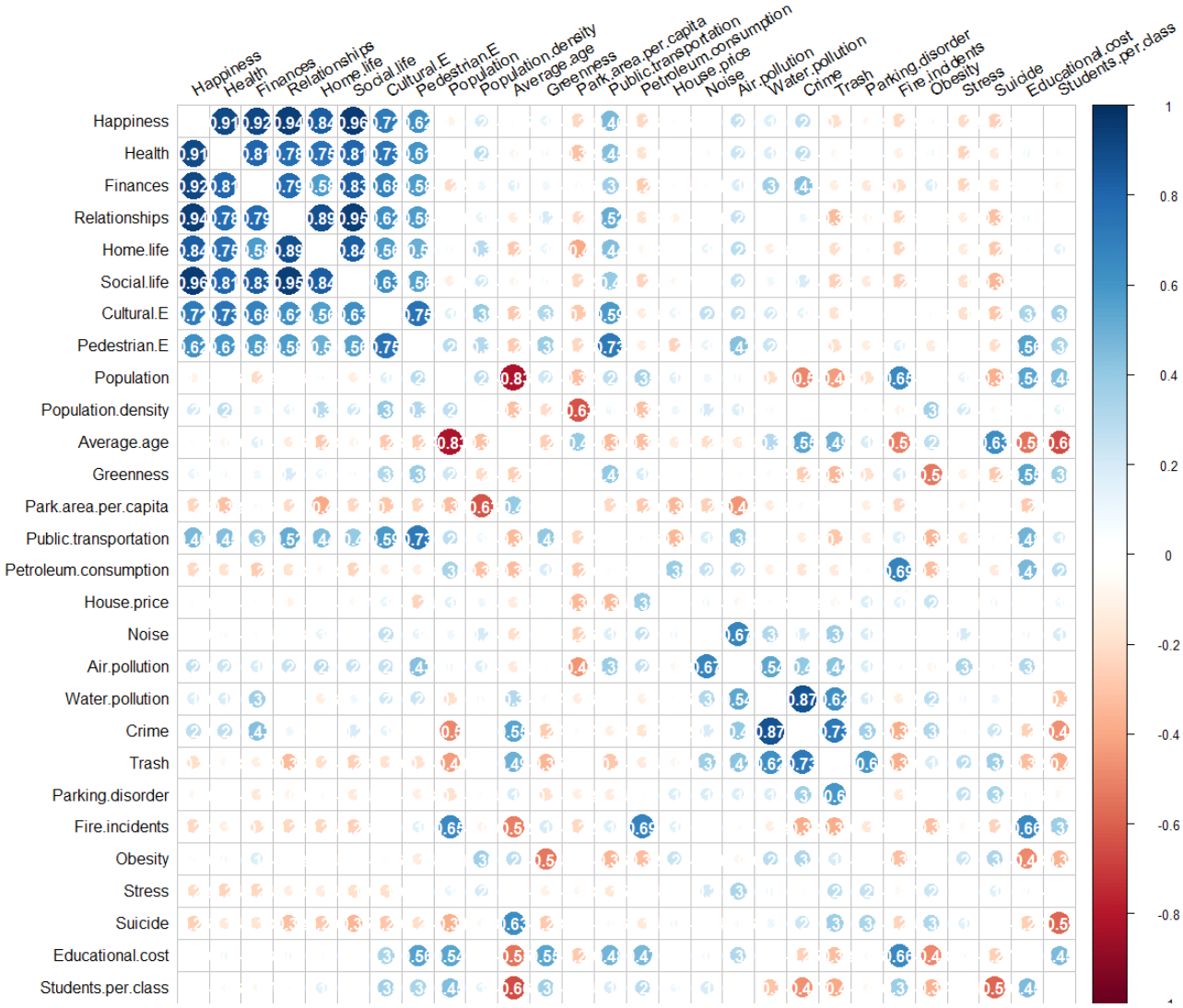
A. Correlation Analysis

The correlation analysis reveals the degrees of association

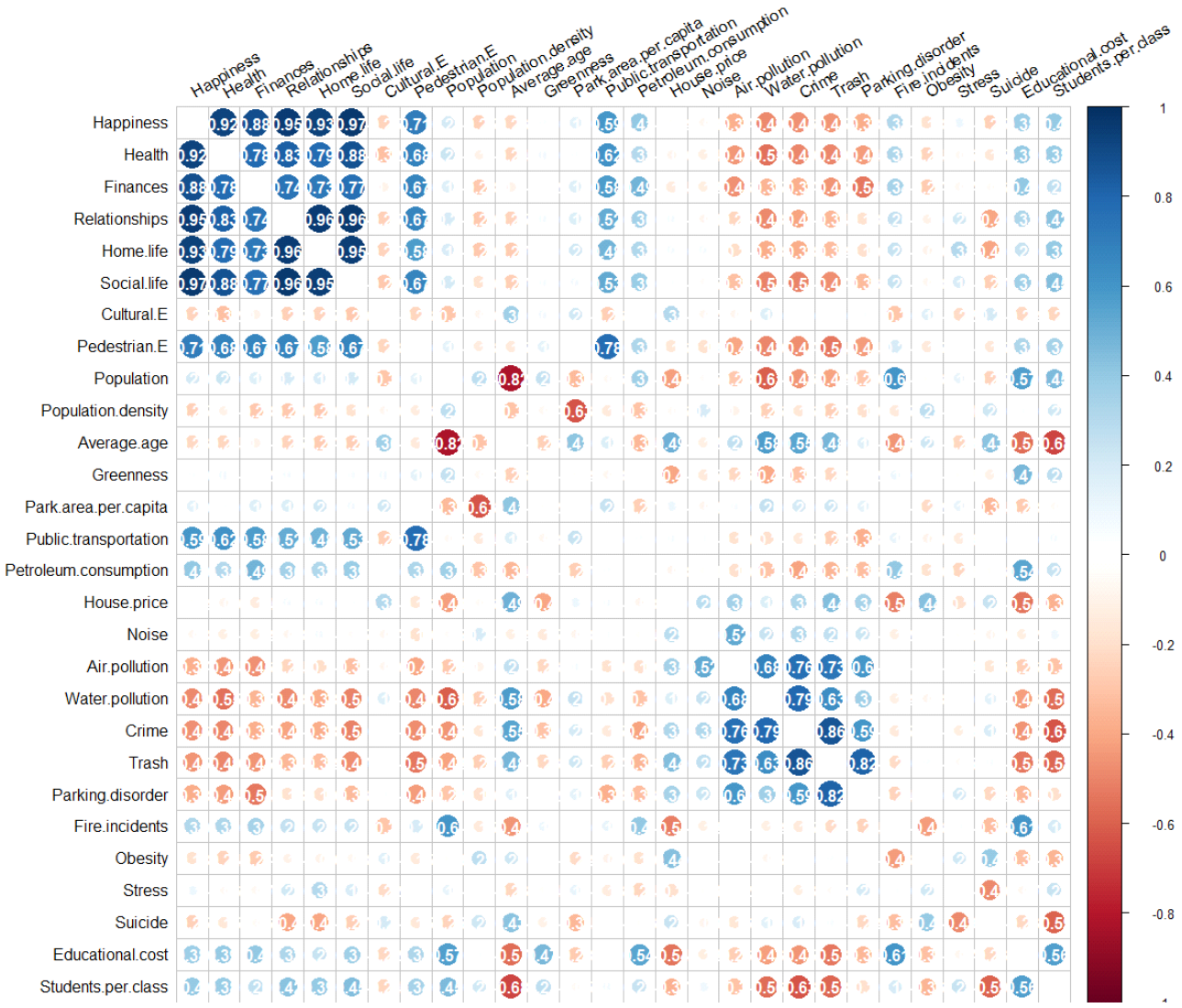
among the data via a correlation coefficient. The correlation coefficient quantifies the strength of association between a pair of variables from –1 to 1. A negative number indicates a negative correlation, and a positive number indicates a positive correlation. A higher absolute correlation coefficient value (-1 or 1) indicates higher correlation, and a lower correlation coefficient value (0) indicate lower correlation. In this study, a correlation analysis of 28 Seoul datasets was conducted [Table I].

TABLE I
THE LIST OF 28 DATA OF SEOUL CITY

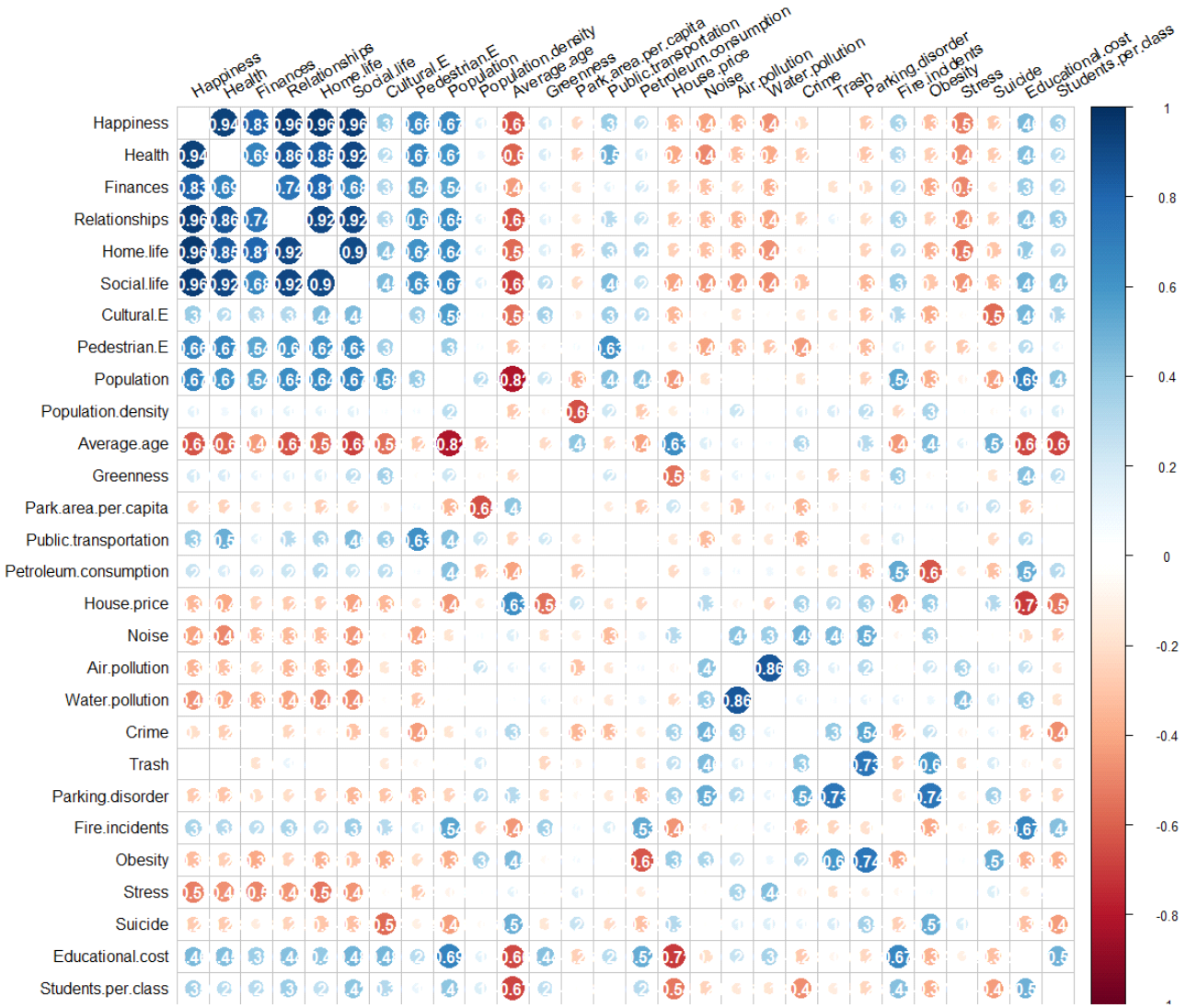
Happiness	Petroleum consumption
Health	House price
Finances	Noise
Relationship (with surrounding people)	Air pollution
Home life (satisfaction)	Water pollution
Social life (satisfaction)	Crime
Cultural E(nvironment satisfaction)	Trash
Pedestrian E(nvironment satisfaction)	Parking disorder
Population	Fire incidents
Population density	Obesity
Average age	Stress
Greenness	Suicide
Park area per capita	Educational cost
Public transportation (satisfaction)	Students per class



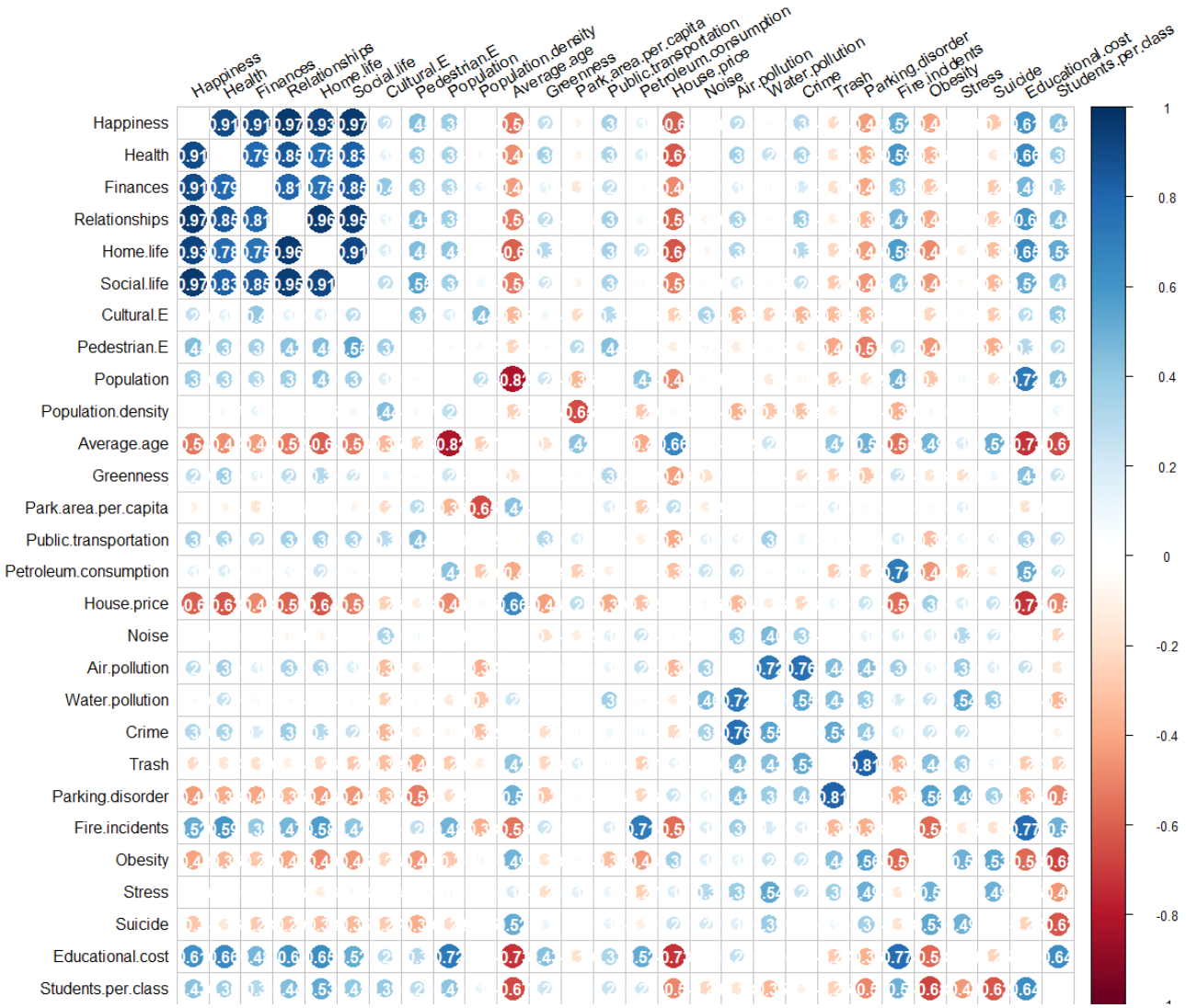
(a)


Happiness
Health
Finances
Relationships
Home.life
Social.life
Cultural.E
Pedestrian.E
Population
Population.density
Average.age
Greenness
Park.area.per.capita
Public.transportation
Petroleum.consumption
House.price
Noise
Air.pollution
Water.pollution
Crime
Trash
Parking.disorder
Fire.incidents
Obesity
Stress
Suicide
Educational.cost
Students.per.class
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

(b)


Happiness
Health
Finances
Relationships
Home.life
Social.life
Cultural.E
Pedestrian.E
Population
Population.density
Average.age
Greenness
Park.area.per.capita
Public.transportation
Petroleum.consumption
House.price
Noise
Air.pollution
Water.pollution
Crime
Trash
Parking.disorder
Fire.incidents
Obesity
Stress
Suicide
Educational.cost
Students.per.class
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

(c)



(d)

Fig. 2 Results of correlation analysis of Seoul data: (a) 2011, (b) 2012, (c) 2013, (d) 2014

Fig. 2 shows the results of the correlation analysis of Seoul data. The correlation between data is significant with a p-value of 0.01, when the correlation coefficient exceeds the absolute value 0.5. The correlation between data is significant with a p-value 0.05, when the correlation coefficient exceeds the absolute value 0.4. But the correlation between data is not significant, when the correlation coefficient is less than the absolute value 0.4. Generally, it is assumed that when the correlation coefficient exceeds the absolute value 0.7, it has a strong correlation. Looking at the relationships among urban data of Seoul within 4 years, the positive correlation data is stronger than the negative correlation data. Data showing strong correlations are as follows [Fig. 3].

The relationship between data changes each year. However, data with strong correlations has not changed significantly. The data with strong relationships for 4 years are as follows: Happiness-Health, Happiness-Finances, Happiness-Home life, Happiness-Social life, Health-Relationships, Health-Home life, Health-Social life, Finances-Relationships, Relationships-Home life, Relationships-Social life, Home life-Social life, Population-Average age. The data with weak relationships are as follows: Pedestrian E-Public transportation, Water pollution-Crime, Crime-Trash, and so on. The data with newly increased relationships include Trash-Parking disorder, House price-Educational cost, and others.

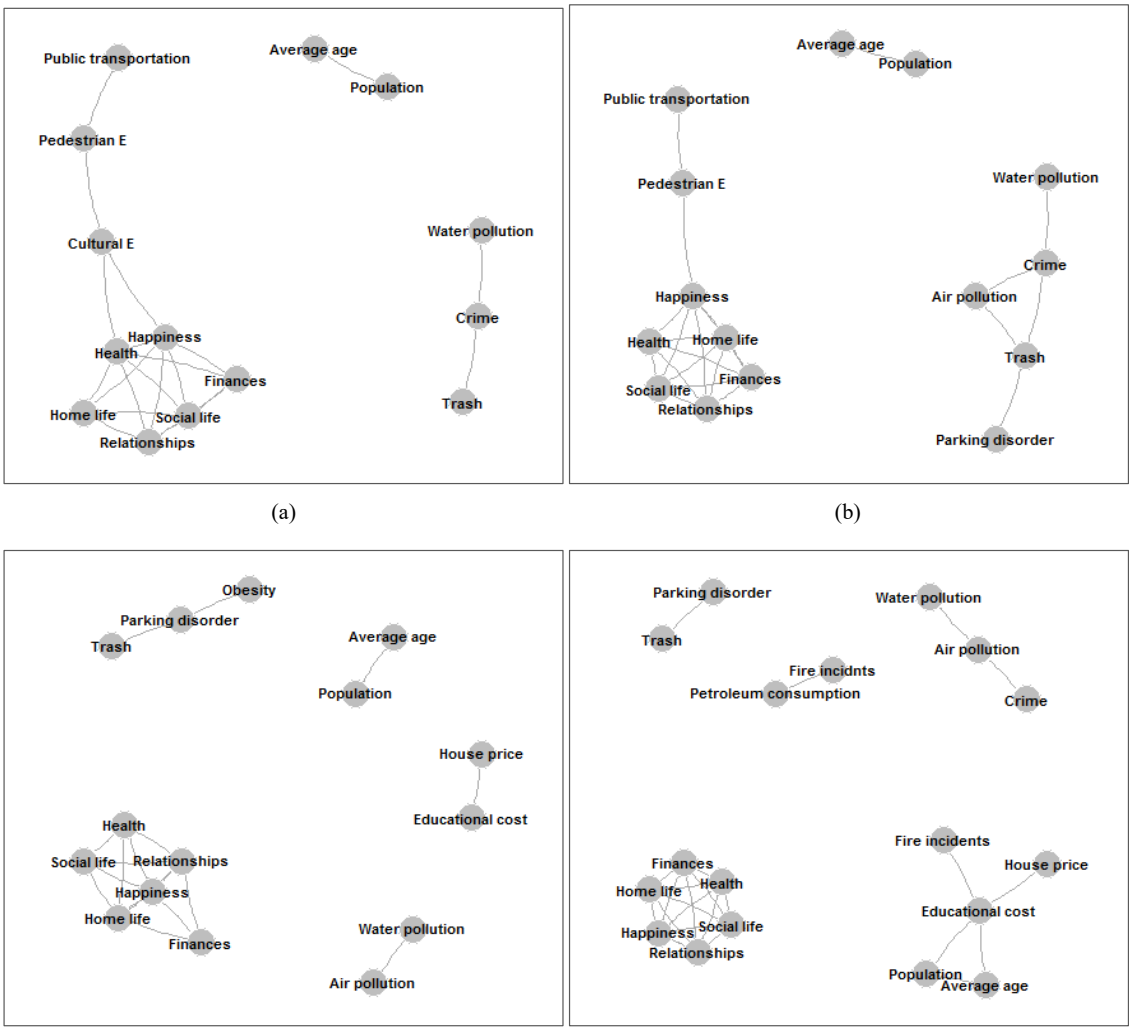


Fig. 3 Data networks showing strong correlations among data: (a) 2011, (b) 2012, (c) 2013, (d) 2014

B. Weights by Correlation Analysis

Looking at the weights of the data, it becomes clear what is important in the given area. The weights of data of the Seoul area were calculated by correlation analysis. Two operators were used to determine the weights. Each operator calculated a general correlation weight and a squared correlation weight. The four weights were then averaged, and the final weight of urban data was calculated [Fig. 4].

Looking at Fig. 5, the weights of the data for each year have similar curves. However, there are cases in which they change dramatically. The weights of population, house price, and the like, are usually high; on the other hand, the weights of cultural E, greenness, noise, and the like, are usually low. The weights of house price, noise and obesity, meanwhile, tend to increase gradually, and the weights of fire incidents, public transportation and students per class show a tendency to decrease gradually. The weights of cultural E, health and social life show wide change ranges. There are also data showing decreasing trends that are increasing formerly, such as the weights of finances, social life and cultural E.

The final weight for Seoul was normalized with a value of 0-1 [Fig. 6]. The weight of each data for four years was ranked. The weights of average age, population and house price were higher, while those of air pollution, noise and greenness were lower.

VI. Discussion

A. Suggested Data-Connection Map

In the current Korean public data homepage, the data are recommended based on keyword tags. Such simple "categorizing" has clear limitations in that it cannot show the hidden correlations among diverse urban datasets. This research therefore examined more advanced correlation-finding methods by using the RM application, and presented the results in the form of a data network. This can trigger the interest of citizens by providing, in the forms of displayed

relevant datasets, insight into both the target information and related information.

All generated city data are analyzed by correlation analysis. However, when recommending data, only the data with high absolute values of correlation coefficients are provided. For example, data are provided in sets of ten: the positive correlation data of the top five quantities and the negative correlation data of the bottom five quantities. In the many cases, the citizens take no interest in the direction of correlation, whether positive or negative. However, some people, such as students, researchers and policymakers, want to know all of the analysis results, including the direction of correlation. So, it is better to express the recommended data in different ways by using colors, shapes, and other means, in order to provide clearer, more accurate and more usable information. An example of representation based on correlation analysis of data for 2011 is provided in Fig. 7.

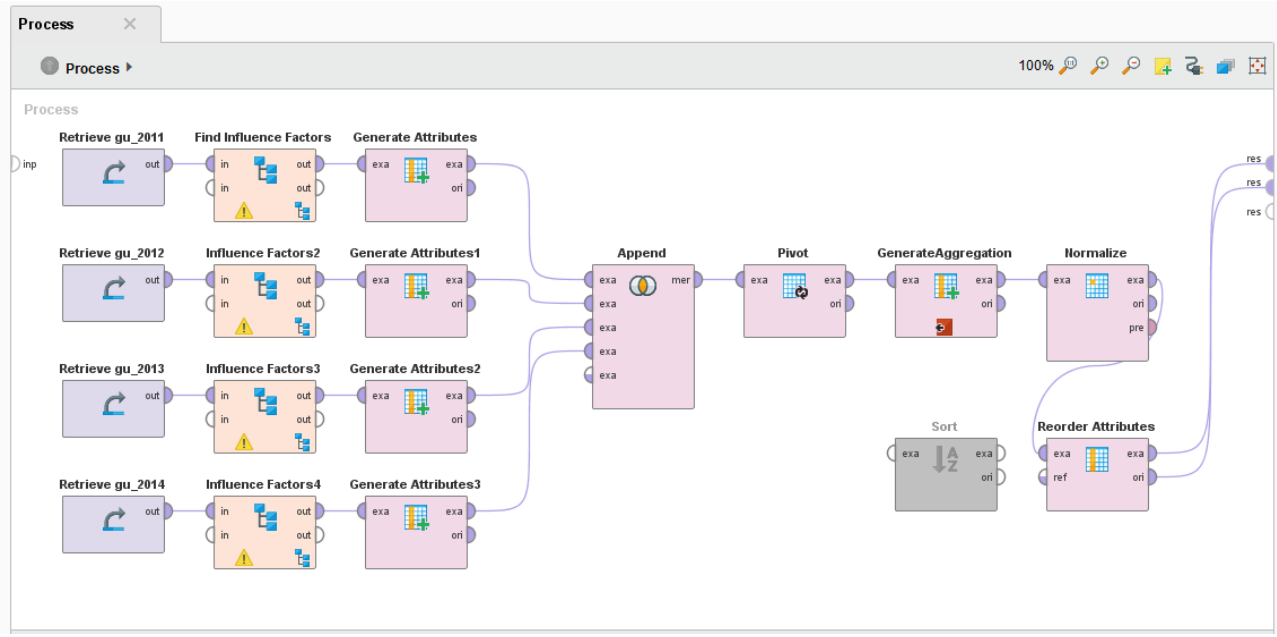


Fig. 4 Calculation of weights of urban data using RM

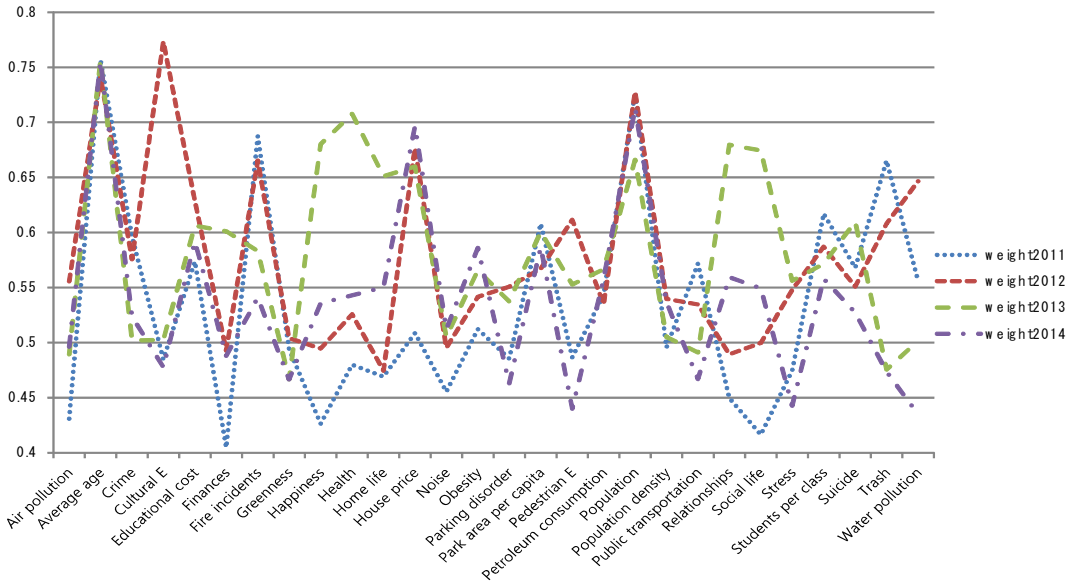


Fig. 5 Weights of urban data each year

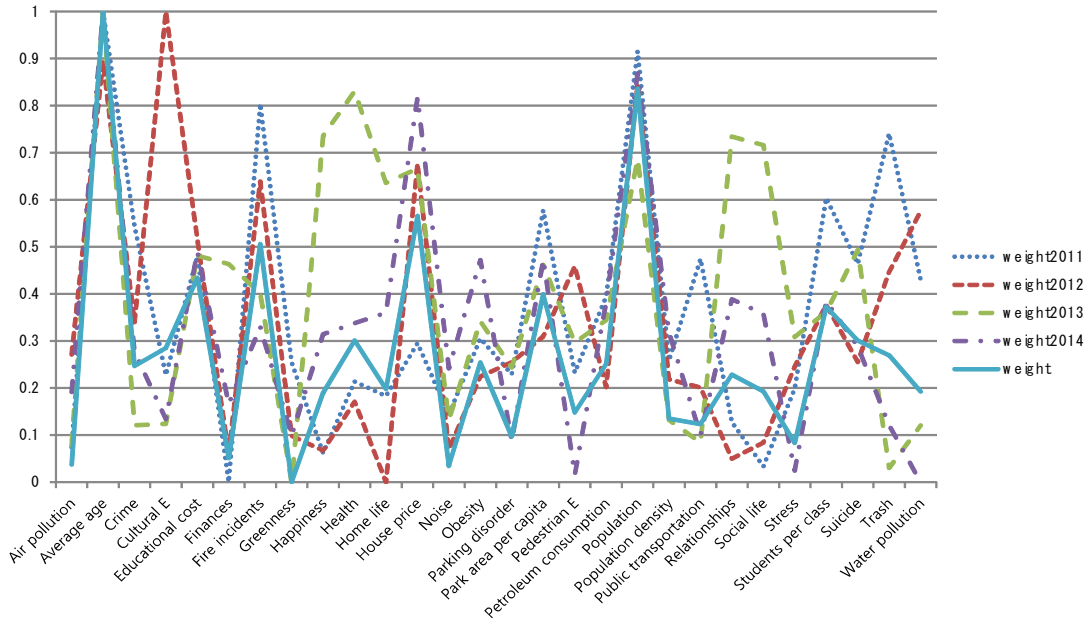
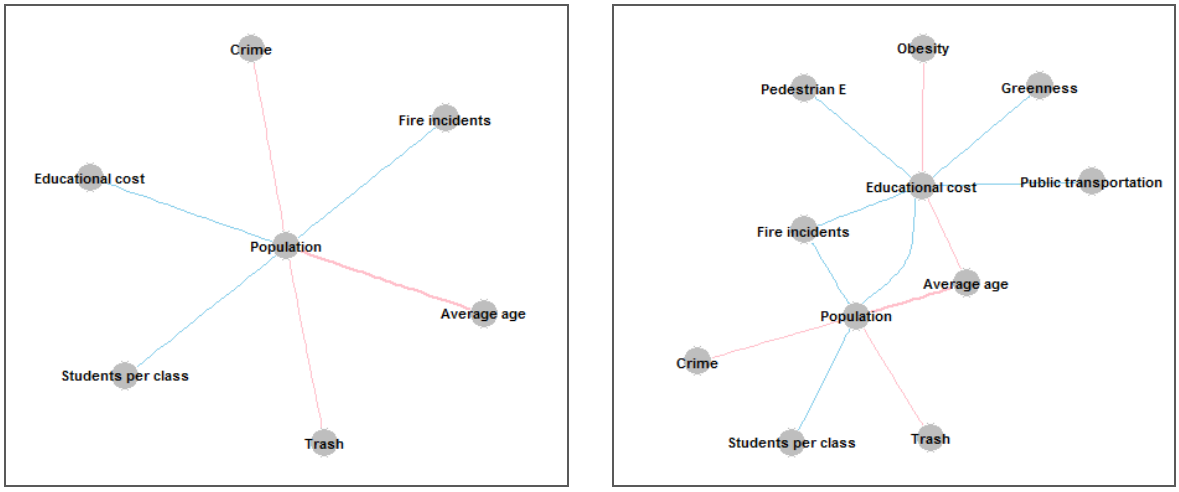


Fig. 6 Weights of Seoul data



(a) (b)

Fig. 6 Data networks of (a) primary relations and (b) secondary relations

B. Utilization of Weight by Correlation Analysis

The weight by correlation analysis can be utilized as a measure of the extent of influence the urban data has on the characteristics of the area. The weight by correlation is appropriate to show the influence of urban data, because it takes into account the relationships with other data. Urban data is greatly affected by changes in people's behavior. Therefore, it is expected that weights will change, each year, with the interests of people and according to government policy. For example, the weight for 2012 was shown to be quite different from the weights for other years. It can be inferred that the interests of the people and the policies of the government changed considerably during the Korean presidential election in 2012.

Based on the weight of 4 years, the characteristics of the weight of Seoul are analyzed. Seoul is the capital of Korea and has a large population. The population seems to play a role as an important element, because the characteristics of the city are functions of citizens' activities. For this reason, average age also is important data, since it determines the age at which economic activity can be done. Uniquely valuable data in this regard are housing prices, which have the third highest weight. It is speculated that changes in real estate policy for economic recovery are influential in people's lives. On the other hand, natural environmental changes, such as air pollution, noise, and greenness, have no significant influence on changes in people's perceptions. In this way, weights enable us to find the identity of the region. Also, weights can help to set directions for a

city's future development.

Correlation analysis of data allows, by application of machine learning, predictions of correlations with future data. Using machine-learning technology with current data, people can anticipate future weights; forecasted future weights, in turn, can be used as indispensable references for improvement of urban environments, policy decisions, and so on.

VII. Conclusion

In this paper, a correlation-analysis-based method for recommendation of related urban data was proposed. This will improve the data recommendation system of Korean public data homepages. In the present study, a correlation analysis was conducted on approximately 30 urban datasets. Based on the results, a data-connection map was drawn with recommended data and differentiated with colors and figures according to the positive and negative relationships. Then, the determined weight values were utilized for understanding the city. The weight of each data was affected by political and social changes. The weights and created networks of the urban data provide us with deeper insights by presenting urban characteristics and identities. It is expected that future work will result in advanced machine-learning algorithms that enable automation of this analysis process for support of users' decision-making.

Acknowledgment

This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIP) (No. 2016R1C1B2013424). This work was financially supported also by the Korean Ministry of Land, Infrastructure and Transport (MOLIT) as 「U-City Master and Doctor Course Grant Program」.

References

[1] F. E. Horton and D. R. Reynolds, "Effects of urban spatial structure on individual behavior", *Economic geography, Perspectives on Urban Spatial Systems*, vol.47 no.1, 1971, pp.36-48.
[2] M. Batty, "Urban Informatics and Big Data", ESRC (The UK Economic and Social Research Council) Cities Expert Group, 2013, unpublished.
[3] I. A. T. Hashem, V. Chang, N. B. Anuar, K. Adewole, I. Yaqoob, A. Gani, E. Ahmed, and H. Chiroma, "The role of big data in smart city", *International Journal of Information Management* 36, 2016, pp.748-758.
[4] L. M. A. Bettencourt, "The Uses of BIg Data in Cities", *SFI Working Paper*, Santafe Institute, 2013, pp.1-20.
[5] M. M. Rathore, A. Ahmad, A. Paul, and S. Rho, "Urban planning and building smart cities based on the Internet of Things using Big Data analytics", *Computer Networks* 101, 2016, pp.63–80.
[6] J. Cheng, N. Gould, and C. Jin, "Big Data for urban studies: opportunities and challenges: a comparative perspective", 2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress, *IEEE Computer Society*, 2016, pp.1229-1234.
[7] Y. Zhang, X. Tang, B. Du, W. Liu, J. Pu, and Y. Chen (2013), "Correlation Feature of Big Data in Smart Cities", H. Gao et al. (Eds.): DASFAA 2016 Workshops, *LNCS* 9645, pp. 223–237.
[8] Open Data Portal, http://www.data.go.kr/main.do (accessed Sep., 2016).
[9] Seoul open data plaza, http://data.seoul.go.kr/ (accessed Sep., 2016).
[10] Rapidminer, https://rapidminer.com/ (accessed Dec., 2016).