16th International Conference on Innovative Computing, Information and Control
September 15-16, 2022, onsite (Chongqing, China) and online
Volume **xx**, Number **0x**, X **xxxx** pp. 1–10

DATA FUSION THROUGH WEB-GIS VISUALIZATION USING OPEN DATA FOR EVIDENCE-BASED POLICY MAKING

Towa Nagase[1], Antonio Oliverira Nzinga Rene[1] and Koji Okuhara[1,*]

[1]School of Technology
Toyama Prefectural University
5180 Kurokawa, Imizu, Toyama, Japan
u255013@st.pu-toyama.ac.jp; *Corresponding author: okuhara@pu-toyama.ac.jp

Received April 2022; accepted July 2022

Abstract. *We propose a method to help solve the issue of subject complexity in policy making by conducting several data analyses using a wide variety of open data that exist in cyberspace in the form of data collected by governments and local governments and made available to the public. We will develop methods for analyzing open data using causal discovery and DEA, as well as methods for effectively presenting results using GIS and data superimposition.*
Keywords: EBPM, causal discovery, DEA, GIS, data-fusion

1. **Introduction.** EBPM is the concept of evidence-based decision making in policy. Currently, researchers in various fields of research have published literature on EBPM, including a discussion of EBPM initiatives in Japan [1] and a book that systematic reviews as the most important evidence [2].

However, many policy decisions in local governments still use episode-based decision making, which is a face-to-face processing response to problems brought to administrative agencies by residents.

One of these causes is the complexity of factors related to the issue that is the subject of the policy. In other words, it is difficult to identify which factors in the surrounding environment influence a problem before it becomes a problem..

We proposes a method to help solve the issue of subject complexity in policy making by conducting several data analyses using a wide variety of open data that exists in cyberspace in the form of data collected by governments and local governments and made available to the public in this research.

In our proposed method, we first collect multiple data from open data existing in cyberspace, regardless of the classification of the items. We then perform a causal search using a Linear non-Gaussian acyclic model (LiNGAM) [3] based on the data on the matter for which we want to make a policy, and extract data that have a causal relationship with the data on the matter for which we want to make a policy. The results are then used to perform Data Envelopment Analysis (DEA) [4].

In addition, we will develop an application that visualizes these results using a Geographic Information System (GIS) [5] and enables visual superimposition on the same platform as the data used in the DEA analysis that have geographic characteristics, thereby helping the government to obtain new policy knowledge.

In this paper, we present the significance of this study by explaining the relationship between EBPM and ICT, as well as the characteristics of GIS and its advantages in use. Then, we propose a method for applying causal search and DEA-based data analysis to open data to solve the aforementioned problems, a method for effectively presenting results using GIS, and a method for superimposing data. Finally, we demonstrate the

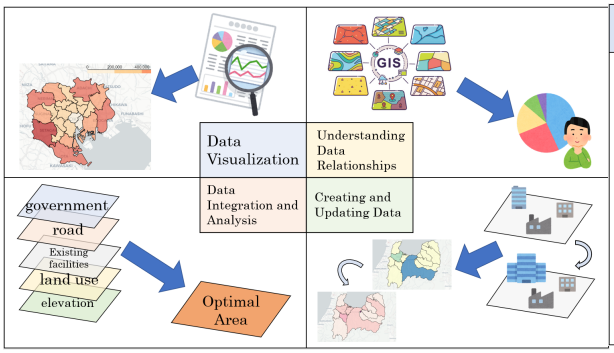

FIGURE 1. Four characteristics in GIS

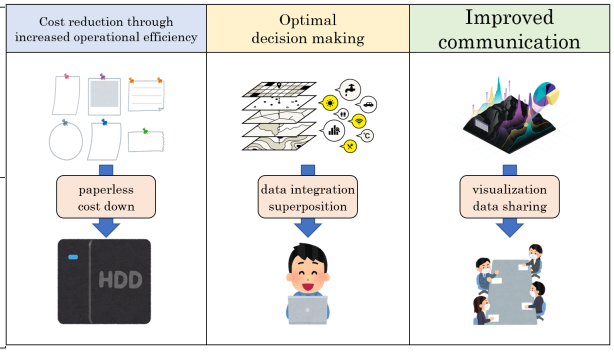

FIGURE 2. Three advantages of using GIS

validity of this study by conducting numerical experiments in which they are applied in practice.

2. Data Application and GIS Data Fusion.

2.1. ICT and Data Utilization in Local Administration. In order to apply effective EBPM to all policies, it is necessary to collect, store, and manage a large amount and variety of data, and to select, integrate, and analyze these data appropriately and quickly with a high degree of confidence. These are large burdens on the person in charge, making it difficult to perform them manually.

Therefore, it is difficult to apply EBPM to a wide range of policies, especially in local governments, from the viewpoint of manpower. For these reasons, the use of ICT is essential for collecting and analyzing appropriate evidence in EBPM.

In such cases, it may be necessary to provide a system that is easy to understand for local government employees, who generally have little contact with specialized ICT, or to foster knowledge of ICT by holding workshops throughout the agency. GIS is an example of a technology that is currently in constant use in each local government [6].

2.2. GIS Features and Benefits. GIS is a technology that enables advanced analysis and rapid decision-making of spatial data by comprehensively managing and processing data with location-related information (geospatial data) and visually displaying the data by linking geographic locations with the data.

Examples of geospatial data include thematic maps (land use maps, geological maps, hazard maps, etc.), urban planning maps, topographic maps, place name information, ledger information, statistical information, aerial photographs, and satellite images, which are used in diverse fields due to their versatility.

In addition, because GIS is superior in several respects, there are multiple advantages to using GIS to analyze data, the four features shown in Figure 1 and the three advantages shown in Figure 2 being the most representative of them [7].

2.3. Data Fusion by Web-GIS. As mentioned above, GIS has the advantage of being able to analyze geospatial data in a sophisticated manner by processing them in various ways and visualizing them on a map to speed up decisions on geospatial data that are difficult to understand as they are, as well as being able to superimpose visualized data on a map.

This makes it possible to identify issues that have not been brought to the surface by visualization of single data alone, and conversely, to discover solutions to problems in seemingly unrelated fields.

As seen in previous studies dealing with the integration of research results from multiple research fields and previous studies showing the usefulness of GIS superimposition, data superimposition on the same platform in GIS can be very useful for discovering new knowledge in complex problems that are not limited to a single cause of the problem.

Therefore, we believe that GIS will be very effective in supporting policy making through data fusion, which is the goal of this study. In this study we will perform data fusion in the form of a GIS overlay of geospatial data on the same platform with the results of our data-based analysis.

3. **Overview of conventional analysis methods.**

3.1. **Relationships between data through causal discovery.** Causal discovery is an unsupervised learning process that uses observed data to derive a causal graph (a structured representation of the degree of influence that each value has on each other in a set of observed data).

In recent years, research on causal search methods has become more active, and various models for causal search have been proposed. A typical example is LiNGAM, a semiparametric model based on independent principal component analysis that can be applied to non-time series data.

LiNGAM is a method for deriving a causal graph, which is generally formulated as in Equation 1 below, with the following assumptions [10].

$$x_i = \sum_{i \neq j} b_{ij} x_j + e_i \qquad i, j = 1, \cdots, p \tag{1}$$

1. The function connecting the exogenous and endogenous variables is a linear function. (An endogenous variable is the variable that is actually observed, and an exogenous variable is a variable other than the endogenous variable that is unknown for each of the endogenous variables.)
2. The distribution of the exogenous variables is non-Gaussian continuous.
3. Causal graphs are assumed to be acyclic.
4. The exogenous variables are assumed to be independent of each other.

LiNGAM estimates causal relationships among endogenous variables using the aforementioned algorithm, and several approaches have been proposed to date, depending on the difference in calculation methods used in the estimation. Typical examples include ICA-LiNGAM, an approach based on independent component analysis, and Direct-LiNGAM, an approach based on regression analysis and independence evaluation. In this study, we use Direct-LiNGAM.

3.2. **Derivation of efficiency values by DEA.** DEA is a non-parametric method developed by Charnes, Cooper and Rhodes in 1978 to evaluate the performance of a set of organizations in a given field. An organization here is a Decision Making Unit (DMU) that is involved in converting several types of inputs into several types of outputs in its activities. One of the advantages of analysis with DEA is the ability to handle data with multiple inputs and outputs.

Since its proposal in 1978, DEA has been actively studied and applied by research institutes around the world [11], and various models have been published to date, including the basic models such as CCR and BCC.

The basic idea of the DMU evaluation method in DEA is how many outputs are produced using how few inputs. Therefore, the evaluation value in DEA is obtained by dividing the sum of the outputs by the sum of the inputs after assigning weights to each input and output in the DMU of interest.

TABLE 1. 地理情報を持たない数値データ

データ項目	単位	データ項目	単位
耕作放棄地率	%	経営耕地面積	1 畝 / 経営体
農業産出額	千万円	労働生産性	なし
企業数	社	従業員数	人
歳出決算額 [総務費]	%	農地平均取引価格	円 $/m^2$
歳出決算額 [民生費]	%	商業用地平均取引価格	円 $/m^2$
歳出決算額 [衛生費]	%	住宅用地平均取引価格	円 $/m^2$
歳出決算額 [農林水産業費]	%	林地平均取引価格	円 $/m^2$
歳出決算額 [商工費]	%	マンション等平均取引価格	円 $/m^2$
歳出決算額 [土木費]	%	1 人あたりの地方税	千円
歳出決算額 [警察費・消防費]	%	製造品出荷額	万円
歳出決算額 [教育費]	%	事業所数	事業所
歳出決算額 [公債費]	%	総人口	人
歳出決算額 [労務費]	%	老年人口	%
歳出決算額 [その他（雑費）]	%	生産年齢	%
農業就業人口平均年齢	歳	年少人口	%
農業経営者平均年齢	歳	年間商品販売額	百万円
林作業請負収入	万円	海面漁獲物等販売額	万円
林産物販売金額	万円	付加価値額	万円
一人当たりの法人住民税	千円	1 人あたりの固定資産税	千円

The weights assigned to each input and output when calculating the evaluation value have constraint formulas based on the inputs and outputs of other DMUs, and the evaluation value in DEA can be calculated by solving a linear programming problem to optimize the input and output weights based on these formulas. The two constraints in the CCR model are as follows. The CCR model shown below is also used in this study.

- None of the evaluation values for all DMUs exceed 1.
- The weights for both input and output are greater than or equal to 0.

Based on these, the CCR model can be formulated as a linear programming problem as follows [12].

<CCR model>

$$\text{maximize} \qquad \frac{u^T y_o}{v^T x_o} = z \tag{2}$$

$$\text{subject to} \qquad -v^T X + u^T Y \leq 0 \tag{3}$$

$$u \geq 0 \tag{4}$$

$$v \geq 0 \tag{5}$$

4. **Proposed Method.**

4.1. **Data scraping and data analysis with causal discovery and DEA.** In the proposed method, data collected from the Web are stored in a database, and causal discovery and DEA are performed on them to analyze the data. Data were collected using RESAS [8] and National Land Numerical Data Download [9].

The data items used in the database are shown in Tables 1, データベースの例 2, and 3. Data attributes can be divided into three main categories, depending on whether the data contains location data, numerical data, or geographical information. Numerical data with geographic information is associated with location data on a one-to-one basis.

Table 2. 地理情報を持つ数値データ

データ項目	単位
施設数 [空港]	箇所
施設数 [工業団地]	箇所
施設数 [都市公園]	箇所
施設数 [道の駅]	箇所
施設数 [学校]	箇所

Table 3. 位置データ

データ項目	単位
施設位置 [空港]	経度・緯度
施設位置 [工業団地]	経度・緯度
施設位置 [都市公園]	経度・緯度
施設位置 [道の駅]	経度・緯度
施設位置 [学校]	経度・緯度

The data used in the analysis in this study have different units and a wide range of value magnitudes, as shown in Table 1, we normalize the data using the following methods [13].

<robust Z-score>

$$\iota = \frac{x - median(x)}{NIQR} \tag{6}$$

<normalization>

$$\iota' = \frac{\iota + \max|\iota|}{2\max|\iota|} \tag{7}$$

The proposed method performs a causal discovery using Direct-LiNGAM on these data. The data used in the causal search are all the data shown in Tables 1 and 2, and we use all of these data in a single causal discovery to automatically narrow down to only those data that are potentially causally related to the policy target. In doing so, we also simultaneously allocate the inputs and outputs in the DEA.

Of the causal graphs identified by causal discovery, only those with arrows pointing toward the data that are the target of the policy shall be treated as the object of analysis. This is because the role of causal search in this study is to refine the data that will be the input and output of DEA. Since an increase or decrease in the data at the beginning of the arrow in the results of the causal discovery will affect the data at the end of the arrow and increase or decrease its value, it was appropriate to consider only the data that affect the data when considering the target of the policy.

Since the data are positively correlated when the weights of the paths in the causal graph are positive and negatively correlated when the weights are negative, we considered it appropriate to use the data at the starting point in the positive case as the output of DEA and the data at the starting point in the negative case as the input of DEA. We use these results to perform DEA.

4.2. **System development of data fusion using Web-GIS.** We have created EBPM-GIS, a GIS application that provides feedback on the results of the proposed method. An example implementation is shown in Figure 3. The direction of the arrow on the icon is the magnitude of the evaluated value in DEA, with blue pointing downward for values less than 0.75, yellow pointing sideways for values between 0.75 and 0.90, and red pointing upward for values greater than 0.90. The colors of the icons for the target and reference set municipalities were also changed, with red for the target and blue for the reference set.

In EBPM-GIS, markers are placed for all the municipalities used in the data analysis, so a very large number of markers are displayed on the screen. For this reason, we implemented a separate layer for each of the aforementioned arrow types to improve visibility and processing speed. We can switch between each layer using the layer control in the upper right corner of the screen. In the initial screen, all arrows are displayed.

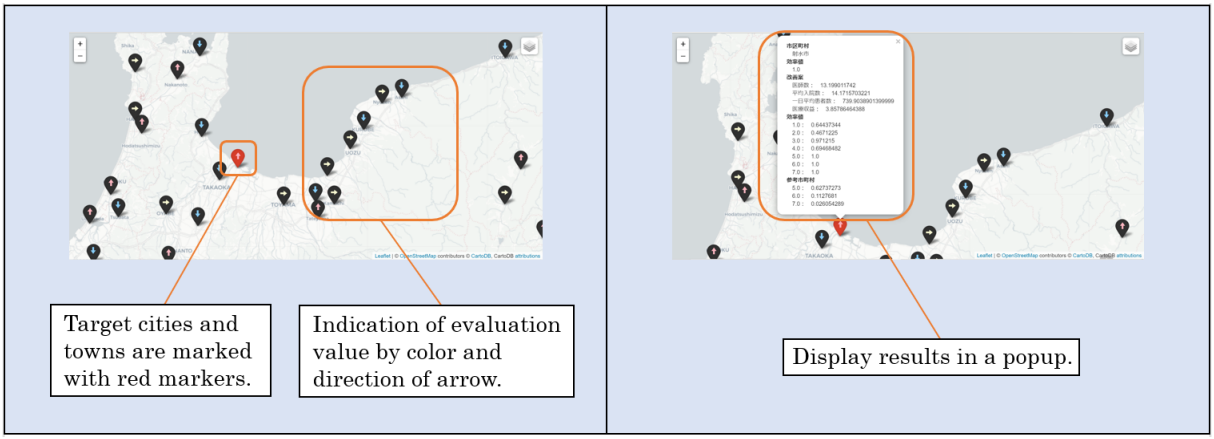

FIGURE 3. EBPM-GIS Implementation

If the causal discovery shows a causal relationship between the target data and the geographic data, the system displays all the locations of the facility distributed throughout the country by means of marker clusters. This marker cluster is also plotted on a separate layer, decoupled from the base map, just like the markers of the evaluation values, and we can show or hide it using the layer control. Users make policy decisions by switching and overlapping data in EBPM-GIS.

5. **Numerical Example.**

5.1. **Summary of Numerical Experiments.**

In this experiment, the city of Imizu in Toyama Prefecture, where the engineering campus of Toyama Prefectural University is located, is taken as a model case of a local government facing a problem.

One of the most serious problems facing the city is its declining population [14]. The city's population peaked in 2005 and has been declining every year since then. In particular, the decline in the younger population has been inversely proportional to the increase in the older population, making it difficult to halt the declining birthrate and aging population.

Therefore, as an evaluation of the effectiveness of the proposed method in this study, we will conduct a numerical experiment targeting "what policies should be implemented to halt the population decline of the city of Imizu in the future". We analyze the aforementioned problem using the proposed method, targeting the most relevant data item in the database, the "juvenile population [percentage]". We select the city of Imizu as the target municipality.

5.2. **Experimental Results and Discussion.**

In the causal discovery part, DirectLiNGAM is used to find causal relationships among the data in the database, targeting "juvenile population [percentage]," a data item for which an increase in its value is considered necessary to solve the problem of aging society with fewer children. "juvenile population [ratio]" is the percentage of the total population in each municipality that is composed of juveniles (under 15 years old).

Table 4 shows the names of the items in the path coefficient matrix derived by DirectLiNGAM for which the path coefficient for "juvenile population [percentage]" is non-zero and their path coefficients. The number of input and output items was six, including the elderly population, average transaction price of residential land, and police and fire expenses, from the one with the largest path coefficient, and three outputs: working-age population, number of facilities [airports], and education expenses.

Table 4. Direct-LiNGAM results for "Juvenile Population [Percentage]"

data item	path coefficient
Number of facilities [airports]	0.059
Number of companies	−0.006
hygiene expenses	−0.019
commercial and industrial expenses	−0.024
Police and firefighting expenses	−0.038
education or school expenses	0.017
Average transaction price of residential land	−0.043
working age population	0.249
elderly population	−0.559

Table 5. Municipalities belonging to the reference set

Municipalities belonging to the reference set	weight
Kawanishi Town, Higashiokitama-gun, Yamagata	0.268
Miyota Town, Kitasakuma-gun, Nagano	0.197
Maibara City, Shiga	0.108
Hino Town, Gamo-gun, Shiga	0.186

The database used in this experiment includes population percentages in three categories: the elderly population, the working-age population, and the juvenile population. Since these data interact with each other due to their proportions, it is reasonable to assume that the old and working-age populations show a causal relationship. The old-age population was assigned as input and the working-age population as output because the parental generation of the young population corresponds to the working-age population.

It is also easy to imagine that the transaction price of land for housing should be lower when considering the increase in the working-age population. The same reason for the allocation of education expenses to output can be considered, and it is thought that the youth population will increase when there is a secure environment for child rearing and education.

Conversely, it is difficult to imagine a direct causal relationship between the number of airport facilities and police and firefighting expenses, and the number of firms allocated to input, so one of the significant findings of this study is that these items can be derived through analysis and lead to new findings.

The DEA results for the city of Imizu were 0.85. The cities that form the reference set are listed in Table 5. The reference set included four municipalities. Among these cities, Maibara City in Shiga Prefecture is considered to be similar in scale to Imizu City, so the issue may be solved by referring to this city in the future.

6. **Conclusions.** In this study, we proposed a method for collecting and analyzing data for decision-making on policies in order to support EBPM in municipalities at the municipal level. First, by using an unspecified large number of open data, we prevented the data to be collected from being biased according to the target of the policy, and then selected data that had a causal relationship with the target by LiNGAM.

Second, among the data for which causal relationships were shown, the data with paths to the target data were divided into two groups, focusing on the positivity and negativity of the paths, and were used as the input and output of the DEA. We then evaluated and analyzed the current situation in the target municipalities by calculating the evaluation values for each municipality using CCR model. Finally, we visualized the results and performed data fusion by creating an EBPM-GIS

Future issues include the expansion of data in the database and the deepening of models and analysis methods in causal search and DEA. Due to the nature of the complexity of the problem in policy making, which is the issue addressed in this study, the more diverse and voluminous the set of data used in the analysis, the more meaningful the results will be in solving the problem. Therefore, we believe that the amount of data from the database treated in this study needs to be significantly increased when it is used for actual policy making.

REFERENCES

[1] 中泉拓也, "英国の EBPM（Evidence Based Policy Making）の動向と我が国への EBPM 導入の課題", 関東学院大学経済経営研究所年報, Vol. 41, pp. 3-9, 2019.
[2] 井伊雅子, 五十嵐中, "新医療の経済学：医療の費用と効果を考える", 日本評論社, 2019.
[3] Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, "DirectLiNGAM: A Direct Method for Learning a Linear Non-Gaussian Structural Equation Model", Journal of Machine Learning Research, Vol. 12, pp. 1225-1248, 2011.
[4] 末吉俊幸, "DEA-経営効率分析法-", 朝倉書店, 2001.
[5] 国土交通省国土地理院, "GIS とは", 閲覧日 2022-02-08, https://www.gsi.go.jp/GIS/whatisgis.html.
[6] 国土交通省国土地理院, "基盤地図情報の利活用事例集", 閲覧日 2022-02-08, https://www.gsi.go.jp/common/000062939.
[7] esri ジャパン," GIS（地理情報システム）とは" , 閲覧日 2022-02-08, https://www.esrij.com/getting-started/what-is-gis/.
[8]
[9]
[10] Dentsu Digital Tech Blog, "Google Colab で統計的因果探索手法 LiNGAM を動かしてみた", 閲覧日 2022-02-08, https://note.com/dd_techblog/n/nc8302f55c775.
[11] 藤井秀幸, 傅靖, 小林里佳子, "データ包絡分析を用いたふるさと納税の戦略提案-K 市のふるさと納税への適用事例-", 日本経営工学会論文誌, Vol. 71, No. 4, pp. 149-172, 2021.
[12] 刀根薫, "包絡分析法 DEA", 日本ファジィ学会誌, Vol. 8, No. 1, pp. 11-14, 1996.
[13] 保母敏行ほか, "日本分析学会における標準物質の開発", 日本分析化学会誌, vol. 57, No. 6, pp. 363-392, 2008.
[14] 射水市役所, "総合戦略-射水市", 閲覧日 2022-02-08, https://www.city.imizu.toyama.jp/appupload/EDIT/054/054185.