ICIC Express Letters
Part B: Applications
Volume **14**, Number **5**, May **2023**

ICIC International ©2023 ISSN 2185-2766
pp. **475–482**

DATA FUSION BY WEB-GIS TO OPEN DATA FOR EVIDENCE-BASED POLICY MAKING

Towa Nagase[1,*], Antonio Oliveira Nzinga Rene[2] and Koji Okuhara[2]

[1]Department of Electrical and Computer Engineering
Graduate School of Engineering
[2]Faculty of Engineering
Toyama Prefectural University
5180 Kurokawa, Imizu, Toyama 939-0398, Japan
{ rene; okuhara }@pu-toyama.ac.jp
*Corresponding author: u255013@st.pu-toyama.ac.jp

Received October 2022; accepted January 2023

Abstract. *We propose a method to help solve the issue of subject complexity in policymaking by conducting several data analyses using a wide variety of open data 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.** Researchers in various research fields have published literature on Evidence-Based Policymaking (EBPM), including a discussion of EBPM initiatives in Japan [1] and a book that systematically reviews the most critical evidence [2].

However, many policies in local governments are still episode-based decision-making, a face-to-face processing response to problems brought to administrative agencies by residents [3]. 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 easier to identify which factors in the surrounding environment influence a problem after it becomes a problem.

One tool provided to assist in the analysis of data to promote EBPM is RESAS [4], which the Cabinet Office of Japan operates. However, the primary function of RESAS is to publish data, and the data analysis function provides nothing more than simple graphing of data. Therefore, developing a system to analyze data and visualize the results is necessary.

The present study proposes a method to help solve the issue of subject complexity in policymaking by conducting several data analyses using a wide variety of open data 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) [5, 6] 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 from this evaluation were considered critical factors for extensive analysis employing Data Envelopment Analysis (DEA), a method with several applications [7].

Additionally, an application to visualize the results using a Geographic Information System (GIS) [8] and to enable visual superimposition on the same platform as data used in

DOI: 10.24507/icicelb.14.05.475

the DEA with geographic characteristics was developed, thereby helping the government to obtain new policy knowledge.

The study explains the relationship between EBPM and ICT, GIS's characteristics, and its advantages. Then, it proposes a method for data analysis applying causal search and DEA to open data to solve the problems mentioned above, a method for effectively presenting results using GIS, and a method for superimposing data. Finally, we demonstrate its validity through numerical experiments. Following the Introduction, the remainder of the paper is structured as follows. Section 2 describes the application of data and aspects related to GIS. In Section 3, the study presents an overview of conventional analysis methods, while Section 4 presents the proposed method of the study. Section 5 presents a numerical example and a discussion of the results, and Section 6 concludes the paper.

2. Data Application and GIS Data Fusion.

2.1. ICT and data utilization in local administration.

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 enormous burdens on the person in charge, making it challenging to perform them manually.

Therefore, applying EBPM to a wide range of policies is challenging, especially in local governments, from the staffing viewpoint. For these reasons, 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 ICT knowledge by holding workshops throughout the agency. GIS is an example of a technology currently in constant use in each local government [9].

2.2. Data fusion by Web-GIS.

As mentioned previously, one can use the advantage of GIS to analyze geospatial data in a sophisticated manner, process data in various ways and visualize them on a map through superimposition and speed up decisions on geospatial data initially tricky to understand. Hence, it is possible to identify issues that have yet to be brought to the surface by visualization of single data alone and, conversely, to discover solutions to problems in seemingly unrelated fields.

Previous studies dealing with integrating research results from many fields and others show that the usefulness of data superimposition on the same platform in GIS can produce knowledge in complex problems unlimited to a single cause.

Therefore, GIS is very effective in supporting policymaking through data fusion, which is the goal of this study. Thus, 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 the 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 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. The model derives a causal graph, generally formulated as in Equation (1) below, with the following assumptions [10].

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

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 [11] to evaluate the performance of a set of organizations in a given field. An organization here is a Decision-Making Unit (DMU) that converts several types of inputs into several 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 institutions, companies, and financial institutions in several fields around the world [12, 13], and various models have been published to date, including basic models such as CCR and BCC.

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

When calculating the evaluation values, weights assigned to each input and output have constraint formulas based on the inputs and outputs of other DMUs. DEA uses linear programming problems to optimize the weights of input and output. Two principles rule the constraints in the CCR model, namely,

- 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 principles, the CCR [14] model can be formulated as a linear programming problem as follows.

<CCR model>

$$\text{maximize} \left\{ \frac{u^T y_o}{v^T x_o} : -v^T X + u^T Y \leq 1, u \geq 0, v \geq 0 \right\} \tag{2}$$

where $\frac{u^T y_o}{v^T x_o} = z$ with z representing rating value for the DMU in question, v and u are virtual weights for the DMU's inputs and virtual weights for the DMU's outputs, respectively. x_o and y_o are target DMU's virtual inputs and target DMU's virtual outputs, respectively; X is virtual inputs for each DMU and Y is virtual outputs for each DMU.

4. Proposed Method.

4.1. Data scraping and data analysis with causal discovery and DEA.

In the proposed method, we employ causal discovery and DEA to analyze data collected from RESAS-API [15] and the National Land Numerical Data Download [16] and stored in a database.

Tables 1, 2, and 3 show the data items used in the database. The attributes are divided into three main categories, depending on whether the data contains location, numerical, or geographical information. Numerical data with geographic information is associated with location data on a one-to-one basis.

Table 1. Numerical data with geographic information

Data item	Units
Number of facilities [airport]	place
Number of facilities [industrial park]	place
Number of facilities [park]	place
Number of facilities [roadside station]	place
Number of facilities [school]	place

Table 2. Location data

Data item	Units
Number of facilities [airport]	latitude, longitude
Number of facilities [industrial park]	latitude, longitude
Number of facilities [park]	latitude, longitude
Number of facilities [roadside station]	latitude, longitude
Number of facilities [school]	latitude, longitude

The data used in the analysis of this study have different units and a wide range of value magnitudes. Therefore, as shown in Table 3, we normalize the data using the following methods [17].

<robust Z-score>

$$\iota = \frac{x - median(x)}{NIQR} \tag{3}$$

<normalization>

$$\iota' = \frac{\iota + \max|\iota|}{2\max|\iota|} \tag{4}$$

The proposed method performs a causal discovery using Direct-LiNGAM on data. Tables 1 and 2 show data used in the causal. A single causal discovery is employed to automatically narrow to only data potentially causally related to the policy target. In doing so, we allocate inputs and outputs simultaneously in DEA.

The role of causal search in this study is to refine data representing the input and output of DEA. Therefore, 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.

Since increasing or decreasing 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 increasing or decreasing its value, it was appropriate to consider only the elements affecting data when considering the target of the policy.

Table 3. Numerical data without geographic information

Data item	Units
Abandoned land rate	%
Abandoned land area	ridge
Agricultural output	10 million yen
Labor productivity	None
Number of companies	companies
Number of employees	employees
Financial expenses	%
Public welfare expenses	%
Hygiene expenses	%
Agriculture, forestry and fisheries expenses	%
Commercial and industrial expenses	%
Public works expenses	%
Police and firefighting expenses	%
Education or school expenses	%
Public debt expenses	%
Labor expenses	%
Price of agricultural land	yen/m^2
Price of commercial land	yen/m^2
Price of residential land	yen/m^2
Price of forestland	yen/m^2
Price of apartment house	yen/m^2
Average age of the employed agricultural population	age
Average age of farm managers	age
Income from forestry work	10 thousand yen
Income from forest products sales	10 thousand yen
Sea catch sales	10 thousand yen
Shipment of manufactured goods	10 thousand yen
Annual sales of merchandise	million yen
Corporate inhabitant tax per capita	thousand yen
Local taxes per capita	thousand yen
Property tax per capita	thousand yen
Amount of value added	10 thousand yen
Number of offices	offices
Total population	population
Elderly population	%
Working age population	%
Juvenile population	%

4.2. **System development of data fusion using Web-GIS.** We created EBPM-GIS, a GIS application that provides feedback on the results of the proposed method. Figure 1 shows an example of implementation. 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. In addition, the colors of the icons for the target and reference set municipalities were changed, that is, red for the target and blue for the reference set.

EBPM-GIS marks all the municipalities used in the data analysis. Consequently, the screen displays several markers. However, a separate layer is implemented for each arrow mentioned earlier to improve the visibility and processing speed. We can switch between

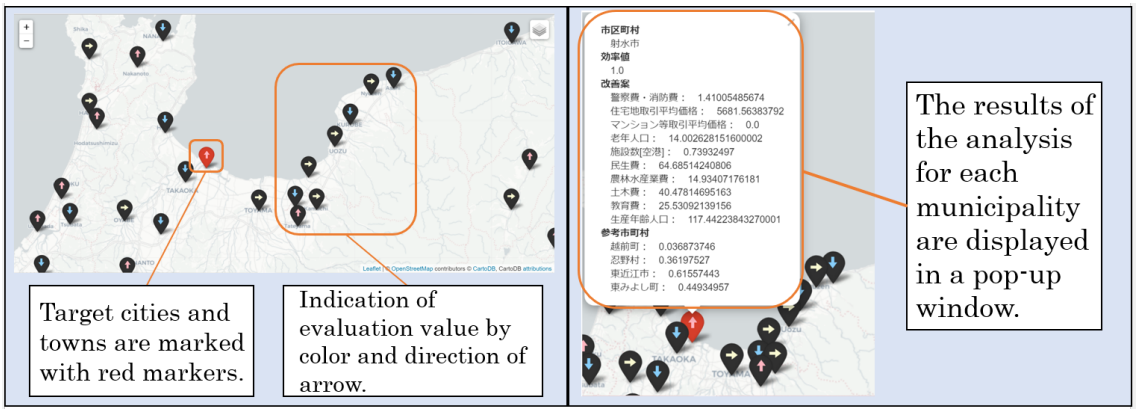

FIGURE 1. EBPM-GIS implementation (Japanese version)

each layer using the layer control in the upper right corner of the screen. The initial screen displays all arrows.

Suppose the causal discovery shows a causal relationship between the target and geographic data. In that case, the system displays all the locations of the facility distributed throughout the country through 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 a model case of a local government facing a problem. Therefore, this city represents the target municipality.

One of the city's most serious problems is its declining population [18]. The city's population peaked in 2005 and has been declining yearly. Mainly, 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 this problem using the proposed method defining the most relevant data item in the database, the "juvenile population [percentage]", as the target.

5.2. Experimental results and discussion.

In the causal discovery part, Direct-LiNGAM is used to find causal relationships from data on the database. The "juvenile population" refers to a data item for which an increase in value is considered necessary to solve the problem of an aging society with fewer children. It represents the percentage of the total population in each municipality composed of juveniles (under 15 years old).

Table 4 shows the names of the items in the path coefficient matrix derived by Direct-LiNGAM for which the path coefficient for the "juvenile population" is non-zero and their path coefficient. 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 most significant path coefficient, and three outputs: working age population, number of facilities [airports], and education expenses.

The database used in this experiment includes population percentages in three categories: the elderly, working age, and juvenile populations. Since these data interact with each other due to their proportions, it is reasonable to assume that the old and working

Table 4. Direct-LiNGAM results for "juvenile population"

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

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 exact reason for allocating education expenses to output can be considered. 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, police and firefighting expenses, and the number of firms allocated to input. Therefore, one of the study's significant findings is that these items can be derived through analysis and lead to new findings.

Data items in Table 4 with positive values represent the outputs, and those with negative values are the inputs, respectively. The evaluation of DEA over the 1851 municipalities in Japan as DMUs yielded 0.85 for the city of Imizu. Table 5 shows the list of cities forming the reference set, and the reference set includes four municipalities. Among these cities, Maibara City in Shiga Prefecture is considered similar in scale to Imizu City, so the issue may be solved by referring to this city in the future.

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

6. **Conclusions.** This study proposed a method for collecting and analyzing policy decision-making data to support EBPM at the municipal level. Firstly, by using an unspecified large number of open data, we took measures to prevent the collected data from being biased according to the target of the policy. We then selected data that had a causal relationship with the target by LiNGAM. Secondly, among the data for which causal relationships were observed, data with paths to the target data were divided into two groups, focusing on the paths' positivity and negativity, which allowed us to define 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 the CCR model. Finally, we visualized the results and performed data fusion by creating an EBPM-GIS. Future issues

include expanding data in the database and deepening models and analysis methods in causal search and DEA.

Furthermore, the proposed method can address several other problems, such as optimal administrative budget allocation. Therefore, in addition to the examples in Section 5, further accuracy improvement can be expected when addressing other issues.

Due to the complexity of the problem in policymaking, the main issue in the present study, the more diverse and voluminous the data used in the analysis, the more meaningful the results will be in solving problems. Hence, data from the database treated in the study needs to be significantly increased when used for actual policy making. In addition, allowing the theory of causal discovery to be adjusted to fit each problem better would make the research more fit for society.

REFERENCES

[1] T. Nakaizumi, Trends in evidence based policy making (EBPM) in the U.K. and challenges in introducing EBPM in Japan, *Annual Report of the Research Institute of Economics and Business Administration*, vol.41, pp.3-9, 2019 (in Japanese).

[2] M. Ii and A. Igarashi, *The New Economics of Health Care: The Costs and Benefits of Health Care*, Nippon Hyoron Sha Co., Ltd., 2019 (in Japanese).

[3] G. Riheng, Toward further promotion of EBPM in government: Current status and recommendations, *NRI Public Management Review*, vol.196, 2019 (in Japanese).

[4] Government of Japan, *Regional Economy Society Analyzing System (RESAS)*, 2015, https://resas.go.jp/#/13/13101, Accessed on Nov. 1, 2022 (in Japanese).

[5] S. Shimizu, T. Inazumi and Y. 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.

[6] S. Salehkaleyber, A. Ghassami, N. Kiyavash and K. Zhang, Learning linear non-Gaussian causal models in the presence of latent variables, *Journal of Machine Learning Research*, vol.21, pp.1-24, 2020.

[7] P. K. Donthula, R. Nellutla and V. V. Haragopal, Measuring the technical efficiency in agriculture farming through CCR model by data envelopment analysis, *IOSR Journal of Mathematics (IOSR-JM)*, vol.17, pp.1-8, 2021.

[8] Ministry of Land and Infrastructure of Japan, *What Is GIS*, Geospatial Information Authority of Japan, https://www.gsi.go.jp/GIS/whatisgis.html, Accessed on Nov. 1, 2022 (in Japanese).

[9] Ministry of Land and Infrastructure of Japan, *Examples of Utilization of Fundamental Geospatial Data*, Geospatial Information Authority of Japan, https://www.gsi.go.jp/common/000062939, Accessed on Nov. 1, 2022 (in Japanese).

[10] Dentsu Digital Tech Blog, *I Ran LiNGAM, a Statistical Causal Search Method, on Google Colab*, 2020, https://note.com/dd_techblog/n/nc8302f55c775, Accessed on Nov. 1, 2022 (in Japanese).

[11] T. Sueyoshi, DEA: Business efficiency analysis method, *The Operations Research Society of Japan*, Asakura Publishing Co., Ltd., 2001 (in Japanese).

[12] H. Fuji, Y. Fu and R. Kobayashi, A proposal for hometown tax strategy by data envelopment analysis –Case study of the hometown tax in K City–, *J. Jpn. Ind. Manage. Assoc.*, vol.71, no.4, pp.149-172, 2021 (in Japanese).

[13] C.-Y. Kao, M.-T. Phung, C.-P. Cheng and W.-H. Chung, Modeling a three-stage network data envelopment analysis model – A case of efficiency analysis for exchange traded funds, *International Journal of Innovative Computing, Information and Control*, vol.17, no.5, pp.1653-1666, 2021.

[14] W. W. Cooper, L. M. Seiford and K. Tone, The basic CCR model, in *Data Envelopment Analysis*, New York, NY, Springer, 2007, https://doi.org/10.1007/978-0-387-45283-8_2, Accessed on Nov. 16, 2022.

[15] Government of Japan, *RESAS-API*, https://opendata.resas-portal.go.jp/docs/api/v1/index.html, Accessed on Nov. 1, 2022 (in Japanese).

[16] Ministry of Land and Infrastructure of Japan, *National Land Information Download Service*, Geospatial Information Authority of Japan, https://nlftp.mlit.go.jp/ksj/, Accessed on Nov. 1, 2022 (in Japanese).

[17] T. Hobo, Development of certified reference materials accomplished by the Japan Society for Analytical Chemistry, *Bunseki Kagaku*, vol.57, no.6, pp.363-392, 2008 (in Japanese).

[18] Imizu City Hall, *Comprehensive Strategy – City of Imizu*, https://www.city.imizu.toyama.jp/appupload/EDIT/054/054185, Accessed on Nov. 1, 2022 (in Japanese).