OVERVIEW

WIREs COMPUTATIONAL STATISTICS　WILEY

Echelon analysis and its software for spatial lattice data

Koji Kurihara [ID] │ **Fumio Ishioka** [ID]

Graduate School of Environmental and
Life Science, Okayama University,
Okayama, Japan

Correspondence
Koji Kurihara, Graduate School of
Environmental and Life Science,
Okayama University, Okayama, Japan.
Email: kurihara@ems.okayama-u.ac.jp

Edited by: Yuichi Mori, Commissioning
Editor and David W. Scott, Review Editor
and Co-Editor-in-Chief

Abstract

In this study, we explore the use of echelon analysis and its software named EcheScan for spatial lattice data. EcheScan is developed as a web application via an internet browser in R language and Shiny server for echelon analysis. The technique of echelon is proposed to analyze the topological structure for spatial lattice data. The echelon tree provides a dendrogram representation. Regional features, such as hierarchical spatial data structure and hotspots clusters, are shown in an echelon dendrogram. In addition, we introduce the conception of echelon with the values and neighbors for lattice data. We also explain the use of EcheScan for one- and two-dimensional regular lattice data. Furthermore, coronavirus disease 2019 death data corresponding to 50 US states are illustrated using EcheScan as an example of geospatial lattice data.

This article is categorized under:
　Statistical Learning and Exploratory Methods of the Data Sciences >
　　Exploratory Data Analysis
　Statistical Learning and Exploratory Methods of the Data Sciences >
　　Clustering and Classification
　Data: Types and Structure > Image and Spatial Data

KEYWORDS
echelon analysis, hierarchical structure, R language and shiny, spatial lattice data, web application

1 │ INTRODUCTION

It is essential to visualize data and analyze uncertainties and variations in data objectively and scientifically for various information obtained in each field. Recently, such data are often obtained as spatial data with spatial position information, and it is challenging to use specialized knowledge to build a model since the model must accurately represent the mechanism inherent in the data and possess its theoretical properties (Cressie, 1993, Chap. 1; Cressie & Chan, 1989). The spatial statistical analysis assumes the relevance of the position or spatial arrangement of an object to be analyzed and is interested in the independence of the observed values.

　Lattice data are observations from random processes observed over a collection of spatial regions. The main task for lattice data is to build a spatial model based on spatial autocorrelation, spatial regression, and so on. Before modeling the spatial components of grid data, it is crucial to investigate the spatial structure of the region of interest. There are

several approaches to detect the spatial clusters and topological hierarchal structure for lattice data. For example, spatial scan statistics (Kulldorff, 1997) based on likelihood ratio (LR) are typically used to detect spatial clusters (hotspots). Some approaches have been proposed to detect hotspots with higher LR (Ishioka et al., 2019; Kurihara, 2003; Myers et al., 2006; Patil & Taillie, 2004; Tango & Takahashi, 2005, 2012).

Echelon analysis method was proposed as a technique for finding the topological structure of two-dimensional spatial data (Myers et al., 1997). In synoptic regional monitoring of a surface variable, echelons provide an objective approach to prospecting for areas of potential concern. Numerous studies have been conducted on the basis of echelon analysis. Myers et al. (1999) and Kurihara et al. (2000) conducted studies on image and remote sensing data. Kurihara (2004) proposed procedures to divide lattice data in view of the classification of echelons. To perform echelon analysis, Kurihara et al. (2020) developed algorithms and EcheScan software for echelon analysis and its applications. Kajinishi et al. (2019) explained the usage of EcheScan software. Hotspot studies of spatial data using spatial scan statistics and echelon analysis employ the following kinds of data: multivariate spatial data (Kurihara et al., 2006), three-dimensional spatiotemporal data (Ishioka et al., 2007; Ishioka & Kurihara, 2012), DNA data (Tomita et al., 2008), suicide data (Tomita et al., 2015), and radiation monitoring post data (Ishioka & Kurihara, 2021).

In this study, we present a procedure to obtain the elements for the echelon of peaks and foundations given the values of lattice and its neighbors. Since space is limited, we will not handle hotspot detection issues (Kurihara et al., 2020). In addition, we also explore EcheScan to help researchers use these technologies easily. The rest of this paper is organized as follows. Section 2 discusses the procedure of the echelon technique for spatial lattice data based on spatial values and neighbors, consisting of peaks and foundations with a hierarchical structure. Section 3 explains EcheScan software for the echelon technique in detail. Section 3.2 explores the usages of two types of lattice data and the coronavirus disease 2019 (COVID-19) death data corresponding to 50 states in the United States using EcheScan software. Finally, Section 4 presents the summary and discussion of our findings.

2 | ECHELONS FOR SPATIAL LATTICE DATA

Spatial data can be classified into three types according to its structure: geostatistical, spatial point pattern, and lattice data (Cressie, 1993, Chap. 1). We focus on lattice data over a fixed subset D of the d-dimensional Euclidean space $\mathbb{R}^d$. The subset D is a collection of spatial sites at which data are obtained using neighborhood information. Using a graphical illustration, the vertices are connected by edges. The neighborhood information is usually based on sharing a common boundary between lattices. Some neighborhood systems are based on the distance between two lattices. In areal data analysis, measurements are aggregated over the spatial area and points **s** are the geographical regions represented as a network in an adjacency graph. Two types of lattice data are available: regular and irregular lattice data. A regular lattice has a geometric structure whose locations are regularly spaced, whereas an irregular lattice has a structure whose locations are not obvious from their geometry and do not have a predictable pattern.

Let **H**(**s**) be a random quantity at spatial location $\mathbf{s} \in \mathbb{R}^d$ and a realization is denoted {**h**(**s**): $\mathbf{s} \in D$}. In a one-dimensional case, lattice data are expressed as $x \in \mathbb{R}^1$ on a one-dimensional line and the value $h(x)$. For example, we have NL regularly divided lattice (interval) data x_i over $s_i = l_1(i) = [i-0.5, i+0.5]$, $i = 1, 2, ..., NL$. The set $D_1 = \{x_i \mid i = 1, 2, ..., NL\}$ is also identified using the index for a real seat $x_i = i$.

The neighbor of one-dimensional lattice data D_1 is given by

$$NB(i) = \begin{cases} \{i+1\}, & i=1 \\ \{i-1, \ i+1\}, & 1<i<NL \\ \{i-1\}, & i=NL \end{cases} \tag{1}$$

for $i = 1, 2, ..., NL$. Table 1 shows an example for $NL = 23$ intervals in the order of A to W, their values $h(i)$ and their neighbors $NB(i)$.

Figure 1 shows nine structured subsets based on the same topological area of cross sectional lattice data. These same topological parts with numbers are called echelons (Myers et al., 1997). Echelons 1–6 are the peaks of cross-sectional data, Echelon 7 is the foundation of peaks (1, 2), Echelon 8 is the foundation of peaks (3, 4) and foundation 7, and

TABLE 1 The index (ID) named A to W, its value $h(i)$ and its neighbors $NB(i)$ for intervals ($i = 1, 2, ..., NL$), $NL = 23$

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
ID	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
$h(i)$	1	2	1	2	3	4	5	4	3	4	3	4	6	4	5	7	5	4	3	2	1	3	1
$NB(i)$	B	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
		C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
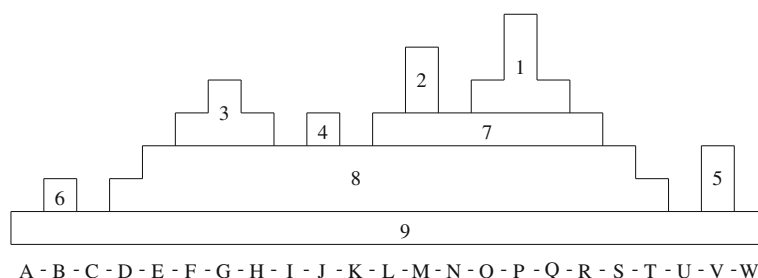
FIGURE 1 Nine structured subsets based on the same topological area for Table 1
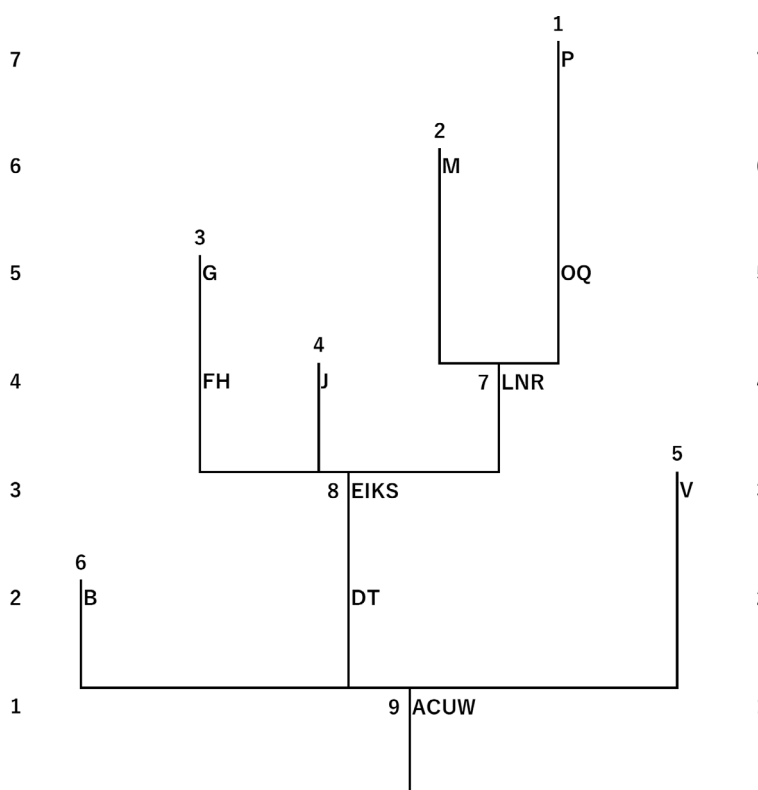
FIGURE 2 Hierarchical relation of echelons for Table 1

Echelon 9 is the foundation of peaks (5, 6) and foundation 8. The Echelon 9 is also called the root. Table 1 data are classified by six peaks, EN (1) = {P, O, Q}, EN (2) = {M}, EN (3) = {G, F, H}, EN (4) = {J}, EN (5) = {V}, EN (6) = {B}, two foundations, EN (7) = {L, N, R}, EN (8) = {E, I, K, S, D, T}, and root in the sense of an echelon: EN (9) = {A, C, U, W}. The relationship among the echelons is expressed as 9(8(7(1 2) 3 4) 5 6) using the numbers of the echelon. Figure 2 shows the topological structure of Table 1 given by the echelon dendrogram.

3 | A WEB APPLICATION FOR ECHELON ANALYSIS

3.1 | Overview

This section introduces *EcheScan*, an application that can execute the echelon analysis on the web (https://fishi.ems.okayama-u.ac.jp/echescan). EcheScan (version 0.2.1 at the time of writing this article) makes it possible to perform echelon analysis on arbitrary data prepared by a user and output its result as a dendrogram. Since this application is

TABLE 2 Required files for echelon analysis using EcheScan

File type	Purpose
Neighborhood information	Provide the labels and neighbors of each lattice.
Univariate file	Provide the values (h) of each lattice.

FIGURE 3 The neighbor information file (`Table1nb.txt`; left) and the univariate file (`Table1h.txt`; right) for the one-dimensional case introduced in Table 1
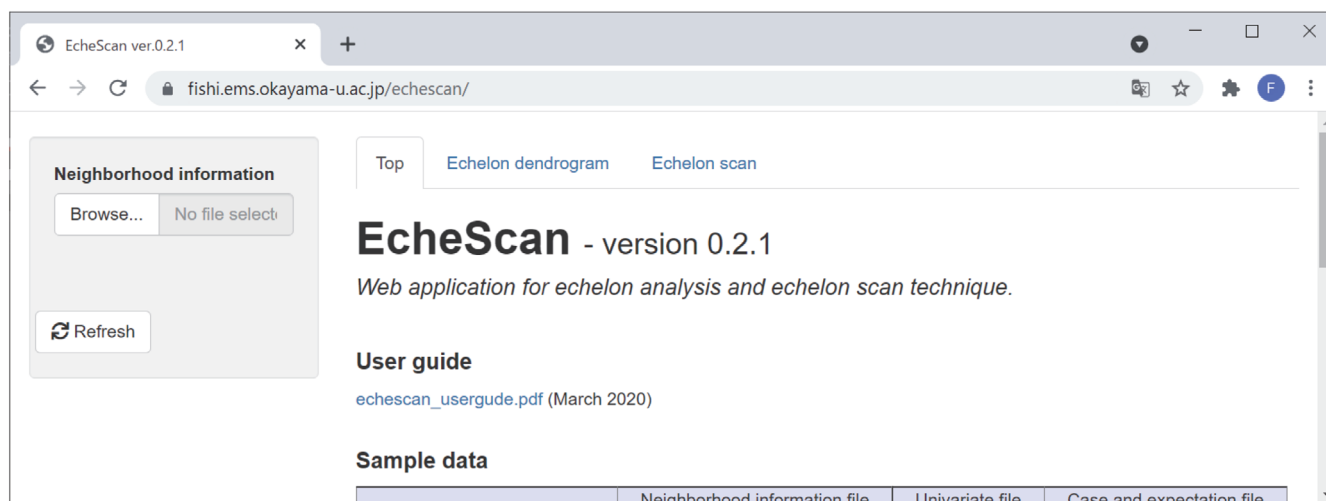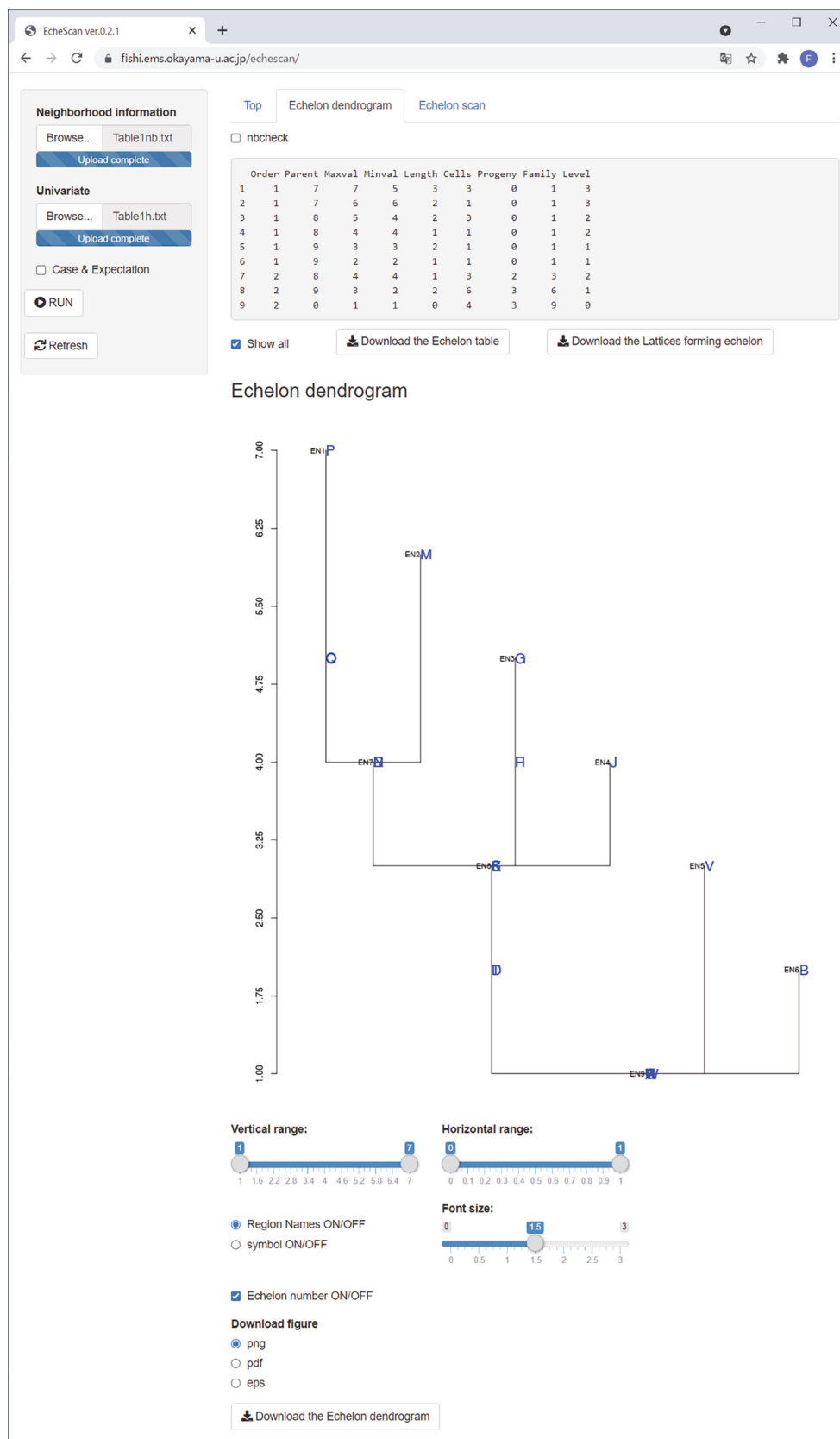

FIGURE 4 Start screen of the EcheScan

FIGURE 5   Execution result of echelon analysis for the one-dimensional case introduced in Table 1

developed in R language and Shiny server, users need only a web browser and do not need any other special environment. It works with any general web browser. The results presented in this paper use Google Chrome browser. The files required to perform echelon analysis on this application are shown in Table 2. For details on the format of each file, refer to the "User guide" on the application site. Each input file is allowed up to 5 MB.

3.2 | Execution examples of EcheScan

3.2.1 | One-dimensional lattice

First, we show an example using EcheScan for the one-dimensional case introduced in Table 1. To run the application, we prepare the two types of files shown in Table 2. Figure 3 (`Table1nb.txt`; left) shows the neighborhood information file for the data in Table 1, with the ID of each lattice in the first column, and the next columns specify the lattices that are neighbor to the first column by using the numbers of corresponding line. Here the line numbers are assigned by A = 1, B = 2, ..., W = 23. In the `Table1nb.txt`, for example, the lattice ID "B" (the second line) is a neighbor to the first line (A) and the third line (C). Figure 3 (`Table1h.txt`; right) shows the data (h) in Table 1 describe data by one vertical column.

TABLE 3 Digital values for a 5-by-5 lattice

	A	B	C	D	E
1	24	12	8	15	3
2	10	1	14	22	5
3	21	13	19	23	25
4	16	4	2	11	17
5	6	20	9	18	7

```
A1   2    6
A2   1    3    7
A3   2    4    8
A4   3    5    9
A5   4   10
B1   1    7   11
B2   2    6    8   12
B3   3    7    9   13
B4   4    8   10   14
B5   5    9   15
C1   6   12   16
C2   7   11   13   17
C3   8   12   14   18
C4   9   13   15   19
C5  10   14   20
D1  11   17   21
D2  12   16   18   22
D3  13   17   19   23
D4  14   18   20   24
D5  15   19   25
E1  16   22
E2  17   21   23
E3  18   22   24
E4  19   23   25
E5  20   24
```

```
24
10
21
16
6
12
1
13
4
20
8
14
19
2
9
15
22
23
11
18
3
5
25
17
7
```

FIGURE 6 Neighborhood information file (`Table3nb.txt`; left) and univariate file (`Table3h.txt`; right) for the 5-by-5 lattice data in Table 3
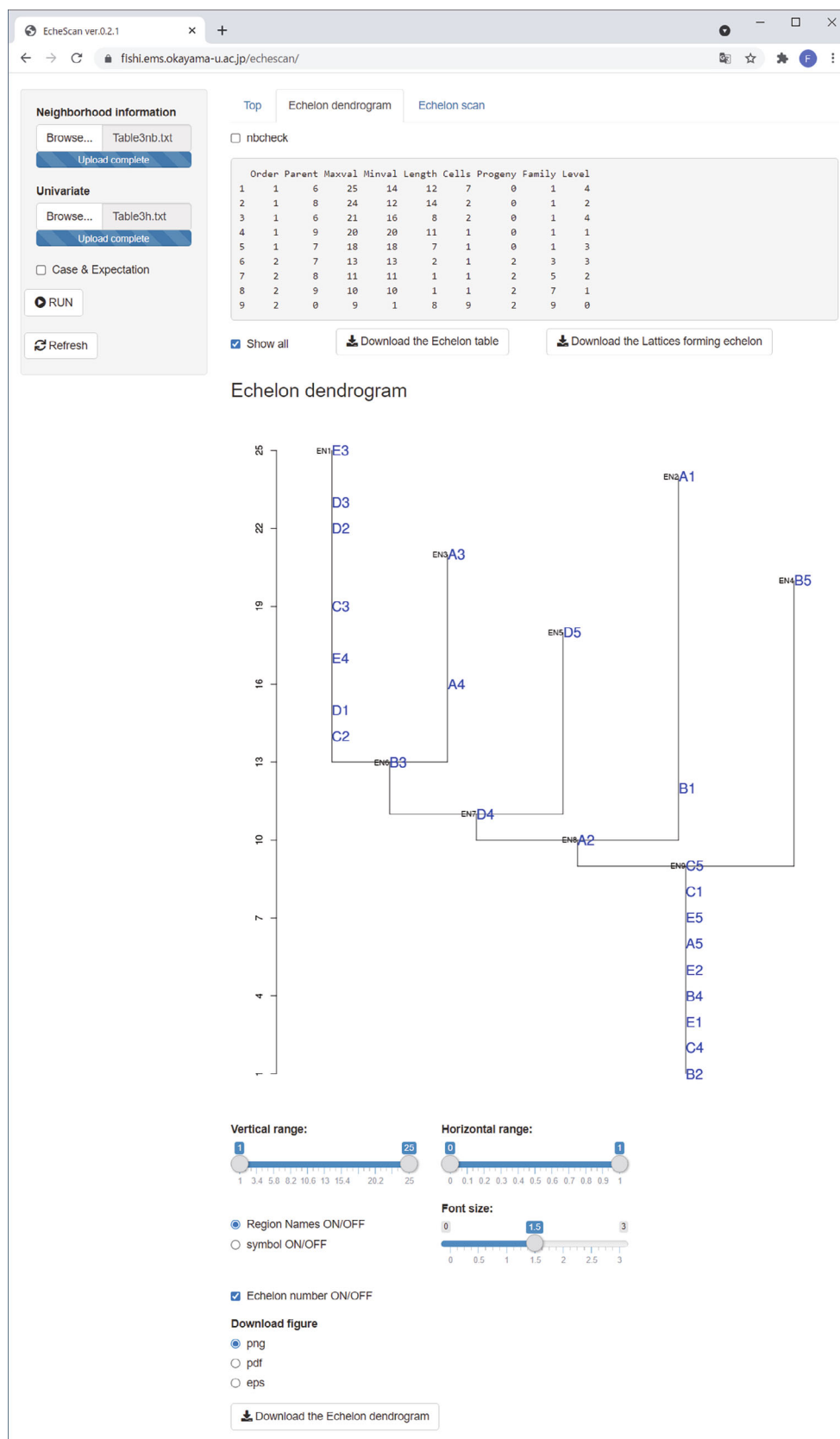
FIGURE 7 Execution result of echelon analysis for the 5-by-5 lattice data in Table 3

TABLE 4 The names of states in USA, their neighbors, COVID-19 deaths per 100k and echelon number

State name	Code	Neighbors of the states							Deaths	*EN(i)*
Alabama	AL	FL	GA	MS	TN				259	2
Alaska	AK	(WA)							63	1
Arizona	AZ	CA	CO	NM	NV	UT			264	3
Arkansas	AR	LA	MS	MO	OK	TN	TX		241	9
California	CA	AZ	NV	OR	(HI)				169	9
Colorado	CO	AZ	KS	NE	NM	OK	UT	WY	132	9
Connecticut	CT	MA	NY	RI					236	1
Delaware	DE	MD	NJ	PA					195	1
Florida	FL	AL	GA						231	2
Georgia	GA	AL	FL	NC	SC	TN			224	2
Hawaii	HI	(CA)							47	9
Idaho	ID	MT	NV	OR	UT	WA	WY		138	9
Illinois	IL	IA	IN	KY	MO	WI			212	5
Indiana	IN	IL	KY	MI	OH				220	5
Iowa	IA	IL	MN	MO	NE	SD	WI		201	7
Kansas	KS	CO	MO	NE	OK				198	6
Kentucky	KY	IL	IN	MO	OH	TN	VA	WV	179	9
Louisiana	LA	AR	MS	TX					281	2
Maine	ME	NH							72	9
Maryland	MD	DE	PA	VA	WV				168	9
Massachusetts	MA	CT	NH	NY	RI	VT			266	1
Michigan	MI	IN	OH	WI					219	5
Minnesota	MN	IA	ND	SD	WI				140	9
Mississippi	MS	AL	AR	LA	TN				299	2
Missouri	MO	AR	IA	IL	KS	KY	NE	OK	TN	184	8
Montana	MT	ID	ND	SD	WY				172	9
Nebraska	NE	CO	IA	KS	MO	SD	WY		122	9
Nevada	NV	AZ	CA	ID	OR	UT			218	3
New Hampshire	NH	MA	ME	VT					106	9
New Jersey	NJ	DE	NY	PA					305	1
New Mexico	NM	AZ	CO	OK	TX	UT			220	3
New York	NY	CT	MA	NJ	PA	VT			288	1
North Carolina	NC	GA	SC	TN	VA				144	9
North Dakota	ND	MN	MT	SD					210	4
Ohio	OH	IN	KY	MI	PA	WV			181	9
Oklahoma	OK	AR	CO	KS	MO	NM	TX		207	6
Oregon	OR	CA	ID	NV	WA				81	9
Pennsylvania	PA	DE	MD	NJ	NY	OH	WV		223	1
Rhode Island	RI	CT	MA						94	9
South Carolina	SC	GA	NC						217	2
South Dakota	SD	IA	MN	MT	ND	NE	WY		236	4
Tennessee	TN	AL	AR	GA	KY	MO	MS	NC	VA	203	6
Texas	TX	AR	LA	NM	OK				208	6

TABLE 4 (Continued)

State name	Code	Neighbors of the states						Deaths	EN(i)
Utah	UT	AZ	CO	ID	NM	NV	WY	85	9
Vermont	VT	MA	NH	NY				46	9
Virginia	VA	KY	MD	NC	TN	WV		141	9
Washington	WA	ID	OR	(AK)				91	9
West Virginia	WV	KY	MD	OH	PA	VA		179	9
Wisconsin	WI	IA	IL	MI	MN			148	9
Wyoming	WY	CO	ID	MT	NE	SD	UT	152	9

Figure 4 shows the start screen of the EcheScan. First, we click the [Browse ...] of "Neighborhood information" on the left side of the screen and select `Table1nb.txt`. If there is no error in the loaded file, the "Univariate" panel appears; then, we select `Table1h.txt` from [Browse ...]. Next, when we click on [Run], and echelon analysis is executed, and the result is displayed on the [Echelon dendrogram] tab (Figure 5). The table shown at the top of Figure 5 displays a summary of each echelon proposed by Myers et al. (1997) and Kurihara et al. (2000).

3.2.2 | 5-by-5 lattice

Two-dimensional lattice data such as remote sensing data are pixels of digital values over the m-by-n regularly spaced lattice area for $i = 1, 2, ..., m, j = 1, 2, ..., n$.

$$l_2(i,j) = \left\{ \left(x_i, y_j \right) \middle| i - 0.5 \leq x_i \leq i + 0.5, \ j - 0.5 \leq y_j \leq j + 0.5 \right\}. \tag{2}$$

This lattice also specifies an identifying features using areal seat $(x_i, y_j) = (i, j)$.

$$D_{22} = \{(i,j) | i = 1, 2, ..., m, \ j = 1, 2, ..., n\}. \tag{3}$$

The neighbors of cell $l_2(i, j)$ are represented as follows for $\mathcal{X} = \{1, 2, ..., m\}$ and $\mathcal{Y} = \{1, 2, ..., n\}$.

$$NB(l_2(i,j)) = \{(k,l) | k \in (\{i-1, i+1\} \cap \mathcal{X}), \ l \in (\{j-1, j+1\} \cap \mathcal{Y})\}. \tag{4}$$

Algorithms 1 and 2 are available for two-dimensional regular lattice data by applying $l_2(i,j)$ to $l_1((j-1) * m + i)$ for an m-by-n regularly spaced lattice, $i = 1, 2, ..., m, j = 1, 2, ..., n$.

We apply the EcheScan to the 5-by-5 lattice data in Table 3. The neighborhood information file (`Table3nb.txt`) and univariate file (`Table3h.txt`) we prepared are shown in Figure 6. As in the case of one-dimensional lattice data, we move `Table3nb.txt` to "Neighborhood information" and `Table3h.txt` to "Univariate." By clicking on [Run], the result of echelon table and dendrogram display on the [Echelon dendrogram] tab (Figure 7).

3.2.3 | Geospatial lattice

Finally, we introduce an example of applying EcheScan to geospatial data where data $(h(D_i))$ are given for each region $D_i, i = 1, 2, ..., n$. Regional features are mainly investigated over lattice regions, such as the counties in a state, ZIP code zone, and the states in the United States. Echelons can be made for geospatial data as well using the neighbors $NB(D_i)$ for each spatial region. Table 4 shows the number of COVID-19 deaths per 100,000 as of September 12, 2021 corresponding to 50 states in the United States. The COVID-19 deaths we used were published on the website "Tracking
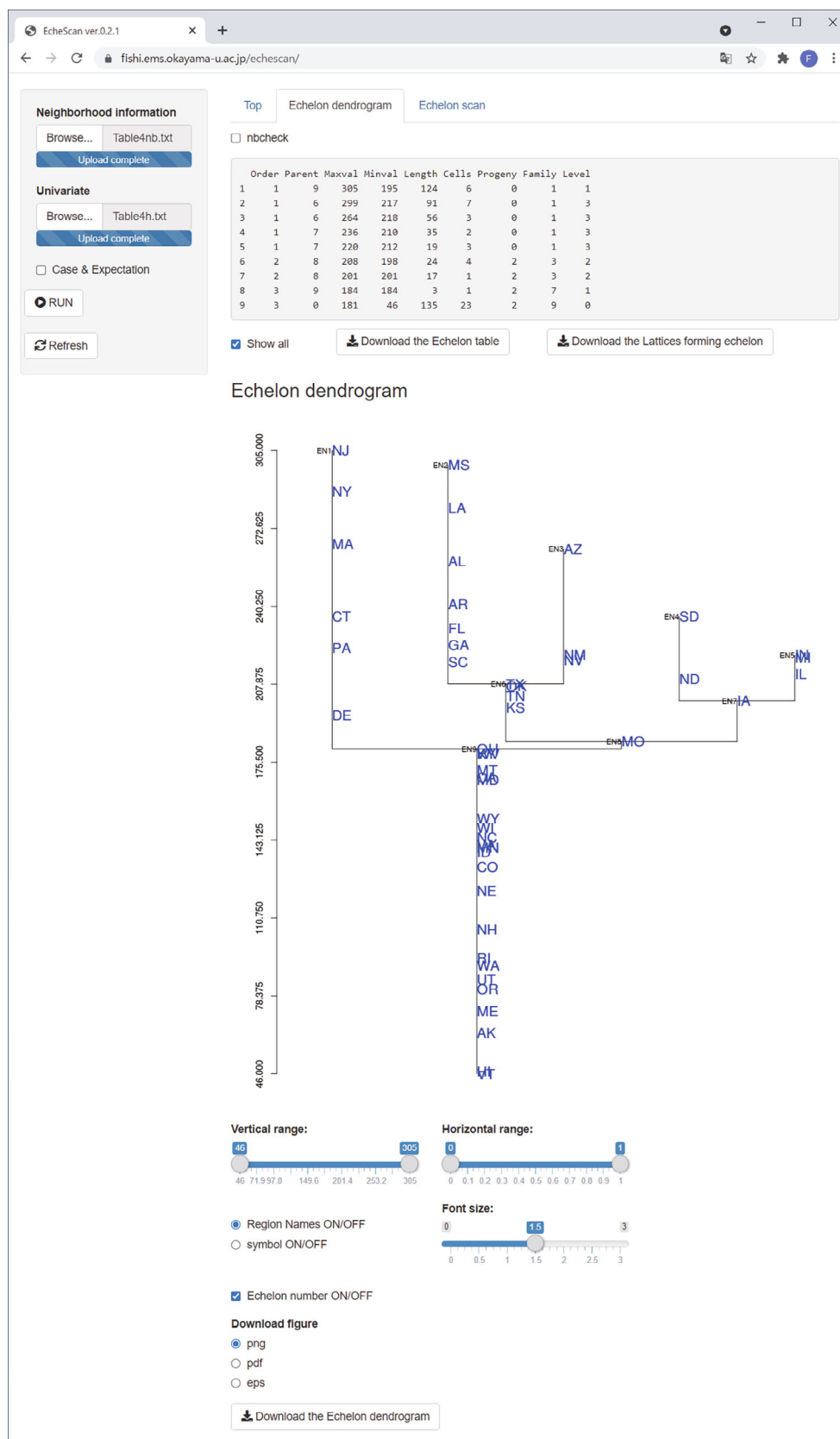
FIGURE 8   Dendrogram of COVID-19 deaths per 100,000 as of September 12, 2021

coronavirus vaccinations and outbreaks in the U.S." (https://graphics.reuters.com/HEALTH-CORONAVIRUS/USA-TRENDS/dgkvlgkrkpb/), and the sources of these data were state and local governments, health authorities and Reuters reporting. These data were taken within lattice areas of states with irregular spaces. The neighborhood information is also shown in this table. Here states that share boundaries are defined as neighbors. Although Alaska and Hawaii are isolated states, we assume that they are connected to Washington and California, respectively.

Figure 8 shows the dendrogram for the data given in Table 4. There were three broad peaks of the COVID-19 deaths in the United States. The first peak primarily comprised mid-Atlantic states: New Jersey, New York, Massachusetts, Connecticut, Pennsylvania, and Delaware. The second peak primarily comprised southeastern states: Mississippi, Louisiana, Alabama, Arkansas, Florida, Georgia, and South Carolina. The third peak primarily comprised southwestern states: Arizona, New Mexico, and Nevada. This graphical representation is useful for a surface understanding of specific regions.

4 | DISCUSSION AND CONCLUDING REMARKS

The echelon analysis was developed to systematically and objectively visualize the topological hierarchical structure of spatial data by dividing the spatial position based on the value of the data on the surface. An echelon aggregates the values on the surface of an entire region with the same topological structure based on the regions having high and low values of observation data with position information and expresses the regions hierarchically. The most relevant advantage of echelon analysis is that the structure can be hierarchically expressed for spatial data given neighborhood information, regardless of dimension. Echelon analysis can visually describe spatial structures, similar to using histograms, box plots, stem-and-leaf plots, scatter plots, and scatter plot matrices to examine the distribution and structure of data. In addition, there are many possibilities for more advanced spatiotemporal data analysis, such as detecting very high values (hotspots) and extending to multidimensional and multivariate spatiotemporal data. In this study, we explore the core of echelon analysis and its software named EcheScan. This study will help in performing echelon analysis and forming echelon project teams and users.

CONFLICT OF INTEREST
The authors have declared no conflicts of interest for this article.

DATA AVAILABILITY STATEMENT
Data sharing is not applicable to this article as no new data were created or analyzed in this study.

AUTHOR CONTRIBUTIONS
Koji Kurihara: Conceptualization (lead); methodology (lead). **Fumio Ishioka:** Software (lead).

ORCID
Koji Kurihara https://orcid.org/0000-0002-2652-3630
Fumio Ishioka https://orcid.org/0000-0002-1715-9574

RELATED WIREs ARTICLE
Clustering of multivariate geostatistical data

REFERENCES
Cressie, N. (1993). *Statistics for spatial data*. Wiley.
Cressie, N., & Chan, N. H. (1989). Spatial modelling of regional variables. *Journal of the American Statistical Association*, *84*, 393–401.
Ishioka, F., Kawahara, J., Mizuta, M., Minato, S., & Kurihara, K. (2019). Evaluation of hotspot cluster detection using spatial scan statistic based on exact counting. *Japanese Journal of Statistics and Data Science*, *2*(1), 241–262.
Ishioka, F., & Kurihara, K. (2012). Detection of spatial clustering using echelon scan. In A. Colubi, E. J. Kontoghorghes, K. Fokianos, & G. Gonzalez-Rodriguez. (Eds.), *Proceedings of the 20th international conference on computational statistics (COMPSTAT2012)* (pp. 341–352). Physica-Verlag.

Ishioka, F., & Kurihara, K. (2021). Detection of space-time cluster for radiation monitoring post data: An application of echelon scan method to point referenced data. In T. Imaizumi, H. Kestler, M. Mizuta, A. Nakayama & C. Welch, (Eds.), *Studies in classification, data analysis, and knowledge organization*. Springer.

Ishioka, F., Kurihara, K., Suito, H., Horikawa, Y., & Ono, Y. (2007). Detection of hotspots for 3-dimensional spatial data and its application to environmental pollution data. *Journal of Environmental Science for Sustainable Society*, *1*, 15–24.

Kajinishi, S., Ishioka, F., & Kurihara, K. (2019). *EcheScan v0.2.0: Web application for echelon analysis and echelon scan technique*. https://fishi.ems.okayama-u.ac.jp/EcheScan

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics: Theory and Methods*, *26*(6), 1481–1496.

Kurihara, K. (2003). The detection of hotspots based on the hierarchical spatial structure. *Bulletin of the Computational Statistics of Japan*, *15*(2), 171–183.

Kurihara, K. (2004). Classification of geospatial lattice data and their graphical representation. In D. Banks, L. House, F. R. McMorris, P. Arabie, & W. Gaul, (Eds.), *Classification, clustering, and data mining applications* (pp. 251–258). Springer.

Kurihara, K., Ishioka, F., & Kajinishi, S. (2020). Spatial and temporal clustering based on the echelon scan technique and software analysis. *Japanese Journal of Statistics and Data Science*, *3*(1), 313–332.

Kurihara, K., Ishioka, F., & Moon, S. (2006). Detection of hotspots on spatial data by using principal component analysis. *Journal of the Korean Data Analysis Society*, *8*(2), 447–458.

Kurihara, K., Myers, W. L., & Patil, G. P. (2000). Echelon analysis of the relationship between population and land cover patterns based on remote sensing data. *Community Ecology*, *1*, 103–122.

Myers, W. L., Kurihara, K., Patil, G. P., & Vraney, R. (2006). Finding upper level sets in cellular surface data using echelons and SaTScan. *Environmental and Ecological Statistics*, *13*(4), 379–390.

Myers, W. L., Patil, G. P., & Joly, K. (1997). Echelon approach to areas of concern in synoptic regional monitoring. *Environmental and Ecological Statistics*, *4*, 131–152.

Myers, W. L., Patil, G. P., & Taillie, C. (1999). Conceptualizing pattern analysis of spectral change relative to ecosystem status. *Ecosystem Health*, *5*, 285–293.

Patil, G. P., & Taillie, C. (2004). Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological Statistics*, *11*(2), 183–197.

Tango, T., & Takahashi, K. (2005). A flexible spatial scan statistic for detecting clusters. *International Journal of Health Geographics*, *4*, 11.

Tango, T., & Takahashi, K. (2012). A flexible spatial scan statistic with a restricted likelihood ratio for detecting disease clusters. *Statistics in Medicine*, *31*(30), 4207–4218.

Tomita, M., Hatsumichi, M., & Kurihara, K. (2008). Identify LD blocks based on hierarchical spatial data. *Computational Statistics & Data Analysis*, *52*(4), 1806–1820.

Tomita, M., Kubota, T., & Ishioka, F. (2015). Spatial clustering properties in the temporal variation of suicide rates/numbers among Japanese citizens: A comprehensive comparison and discussion. *PLoS One*, *10*(7), 1–11.

How to cite this article: Kurihara, K., & Ishioka, F. (2023). Echelon analysis and its software for spatial lattice data. *Wiley Interdisciplinary Reviews: Computational Statistics*, *15*(1), e1579. https://doi.org/10.1002/wics.1579