AppliedMath

Article

Visualization and Analysis of Three-Way Data Using Accumulated Concept Graphs

Manabu Ichino [1,*], Kadri Umbleja [2] and Hiroyuki Yaguchi [1]

[1] School of Science and Engineering, Tokyo Denki University, Hatoyama, Saitama 350-0394, Japan; yaguchi@mail.dendai.ac.jp
[2] Department of Computer Systems, Tallinn University of Technology, Ehitajate Tee 5, 19086 Tallinn, Estonia; kadri.umbleja@taltech.ee
* Correspondence: ichino@mail.dendai.ac.jp

Abstract: This paper introduces the Accumulated Concept Graph (ACG), a visualization tool based on the quantile method designed to analyze three-way data, including distributional data. Such data often have complex structures that make it difficult to identify patterns using conventional visualization techniques. The ACG represents each object with one or more monotonic line graphs. As a result, the three-way data are visualized as a set of parallel monotonic line graphs that never intersect. This non-intersecting property allows for the easy identification of both macroscopic and microscopic patterns within the data. We demonstrate the usefulness of ACGs and principal component analysis in the analysis of real three-way datasets.

Keywords: three-way data; distributional data; the quantile method; parallel monotone line graphs; visualization; microscopic property; macroscopic property; PCA

Citation: Ichino, M.; Umbleja, K.; Yaguchi, H. Visualization and Analysis of Three-Way Data Using Accumulated Concept Graphs. *AppliedMath* **2024**, *4*, 1162–1180. https://doi.org/10.3390/appliedmath4030062

Academic Editor: Alexander Ayriyan

Received: 28 June 2024
Revised: 19 August 2024
Accepted: 26 August 2024
Published: 9 September 2024

1. Introduction

Visual perception is a powerful tool for humans, aiding in the recognition of complex patterns within data. We can identify details and patterns more easily through visual input than by analyzing large volumes of numeric data. This ability is a crucial factor in visual data mining and is particularly beneficial during the exploratory data analysis phase, where little is known about the data or the patterns within them [1]. Numerous ideas have been proposed and examined within traditional data analysis (e.g., [1–5]).

The visualization of multi-dimensional data is a complex procedure, even more so than traditional data visualization, yet it is essential for a comprehensive understanding of the data. Symbolic data, a type of multi-dimensional data, allow for the aggregation of large datasets (including big data that cannot be analyzed using classical approaches), reducing them to a more compact format and thus enabling researchers to analyze and process the data. The most thorough overview of the uniqueness, benefits, and available approaches for symbolic data remains [6]. Due to the complexity of this problem, many different ideas have been explored over the years (e.g., [7–18]). The most influential research in the field of symbolic data visualization has likely been led by Noirhomme-Fraiture, who has authored multiple works on symbolic data visualization, using a radial chart-based "zoomstar" approach to describe various features and different data types (e.g., [8,9]). Visualization of multiway (including three-way) data has also been covered in works such as [16–18].

The main drawback of the currently available approaches is that they can only handle a limited number of features, require a very specific type of symbolic data, or are unable to support datasets with varying types of symbolic data. Most of the focus has been on visualizing the results of principal component analysis or clustering. However, detailed visualizations designed for exploratory data analysis or human perception, which would allow for the identification of both microscopic and macroscopic details in the data, have been largely overlooked.

The aim of this paper is to propose a solution that overcomes these shortcomings by using a quantile method to analyze distributional symbolic data with Accumulated Concept Graphs (ACGs) ([19]). The main advantage of distributional symbolic data is that any other type of symbolic data can be transformed to have a distributional format. The proposed approach enables readers to observe both macroscopic and microscopic properties of the underlying aggregated data. In Section 2, we describe three ACGs for the Hardwood dataset [20,21], which is a three-way dataset organizing ten hardwoods. Each hardwood is composed of seven quantile vectors and described by eight features. We define the Quantile Vectors ACG (QV-ACG), the Feature-Wise ACG (FW-ACG), and the Total ACG (T-ACG) for the given numerical data: (10 objects) × (7 quantile vectors) × (8 features). The QV-ACG represents each object using a set of seven monotonically increasing line graphs, obtained by accumulating eight (0–1)-normalized feature values for each quantile vector. The FW-ACG represents each object with a set of eight monotonically increasing line graphs, obtained by accumulating seven (0–1)-normalized quantile values. By further accumulating the FW-ACG, we obtain the T-ACG. In the T-ACG, each line graph describing the hardwood is obtained by accumulating the eight feature-wise line graphs in a given order. We describe various macroscopic and microscopic properties of the Hardwood dataset using the ACGs, along with the results of the PCA.

In Section 3, we examine the Prefecture Profile Data I [22] dataset, another three-way dataset. This dataset represents Japan's 47 prefectures based on the number of people employed in 10 different job categories over six distinct time periods. By applying the quantile method with ACGs and PCA, we visualize and analyze this numerical data, structured as (47 objects) × (6 quantile vectors) × (10 features).

In Section 4, we analyze another three-way dataset: Prefecture Profile Data II [22]. This dataset represents the 47 prefectures using seven different types of numerical data tables. By applying the quantile method, we transform this general three-way data into the following format: (47 objects) × (5 quantile vectors) × (11 features). We then visualize and analyze the transformed data by combining ACGs with the quantile method of PCA

Section 5 summarizes the major collaborative properties of ACGs and PCA for the given three-way datasets.

2. Analysis of Distributional Data (Hardwood Data)

2.1. Accumulated Concept Graphs for the Hardwood Data

As a typical three-way data problem, we selected the following ten hardwoods based on five species from the US Geological Survey [20]:

Acer East, Acer West; Alnus East, Alnus West; Fraxinus East, Fraxinus West;
Juglans East, Juglans West; and Quercus East, Quercus West.

Table 1 lists the eight histogram-valued features describing the selected hardwoods. For example, Acer East is described by seven quantile vectors, corresponding to quantile values for 0, 10, 25, 50, 75, 90, and 100 percentiles. The quantile vector QV 4, associated with the 50th percentile, has feature values such as ANNT = 9.2, JANT = −5.1, JULT = 22.2, and so on.

Table 1. Features for Hardwood data.

Feature	Description
F1	Annual Temperature (ANNT) (°C)
F2	January Temperature (JANT) (°C)
F3	July Temperature (JULT) (°C)
F4	Annual Precipitation (ANNP) (mm)
F5	January Precipitation (JANP) (mm)
F6	July Precipitation (JULP) (mm)
F7	Growing Degree Days on 5 °C base × 1000 (GDC5)
F8	Moisture Index (MITM)

In Table 2, the quantile values for each feature satisfy the monotonicity property, meaning that the seven quantile vectors follow a consistent vector order:

$$\text{QV1} \leq \text{QV2} \leq \cdots \leq \text{QV7}. \tag{1}$$

Table 2. Acer East, represented with 7 quantile vectors by 8 feature values.

QV	ANNT	JANT	JULT	ANNP	JANP	JULP	GDC5	MITM
1	−2.30	−24.6	11.5	415	10	56	0.5	0.62
2	0.60	−18.3	16.6	720	23	77	1.2	0.89
3	3.80	−12.3	18.2	835	40	89	1.5	0.94
4	9.20	−5.10	22.2	1010	69	100	2.5	0.97
5	14.4	2.30	25.8	1200	96	113	3.6	0.99
6	17.9	7.90	27.3	1355	127	135	4.8	0.99
7	23.8	18.9	28.8	1630	166	22	6.8	1.00

We apply (0–1) normalization to the quantile values of the ten hardwoods for each feature. It is important to note that (0–1) normalization is essential to ensure consistent data with unitless numbers. Table 3 shows the results for Acer East.

Table 3. Acer East, represented with 7 quantile vectors by 8 (0–1)-normalized feature values.

QV	ANNT	JANT	JULT	ANNP	JANP	JULP	GDC5	MITM
1	0.25	0.11	0.16	0.07	0.01	0.12	0.05	0.59
2	0.32	0.22	0.36	0.14	0.03	0.17	0.13	0.88
3	0.41	0.33	0.42	0.16	0.06	0.20	0.17	0.93
4	0.54	0.45	0.57	0.20	0.10	0.22	0.29	0.97
5	0.68	0.58	0.70	0.24	0.14	0.25	0.42	0.99
6	0.76	0.68	0.76	0.28	0.19	0.30	0.56	0.99
7	0.91	0.87	0.81	0.34	0.25	0.49	0.80	1.00

Let x_{ij}, $j = 1, 2, \ldots, 8$; $i = 1, 2, \ldots, 7$, be the (0–1)-normalized quantile values in Table 3. Then, the accumulated quantile values y_{ij}, $j = 1, 2, \ldots, 8$, for the quantile vector QV_i, $i = 1, 2, \ldots, 7$, are given by:

$$y_{ij} = x_{ij} + y_{i(j-1)}, j = 1, 2, \ldots, 8; i = 1, 2, \ldots, 7, \tag{2}$$

where we assume that $y_{i0} = 0$ for all i. We accumulate the (0–1)-normalized feature values for each quantile vector and obtain the result in Table 4 for Acer East.

Table 4. Representation of Acer East by 7 accumulated quantile vectors.

QV	ANNT	JANT	JULT	ANNP	JANP	JULP	GDC5	MITM
1	0.25	0.36	0.53	0.60	0.61	0.74	0.78	1.37
2	0.32	0.55	0.90	1.04	1.07	1.24	1.37	2.25
3	0.41	0.73	1.15	1.31	1.37	1.57	1.73	2.67
4	0.54	0.99	1.56	1.76	1.86	2.08	2.37	3.34
5	0.68	1.26	1.96	2.20	2.34	2.59	3.01	4.00
6	0.76	1.44	2.20	2.48	2.67	2.96	3.52	4.51
7	0.91	1.79	2.60	2.93	3.18	3.67	4.47	5.47

We organized the data into a long Excel column by vertically arranging the QV1–QV7 values from Table 4, with one space inserted between different quantile vectors. The QV-ACG was then generated using the scatter plot command, as shown in Figure 1. By swapping the positions of the quantile vectors and features, we create the FW-ACG for Acer East, shown in Figure 2.

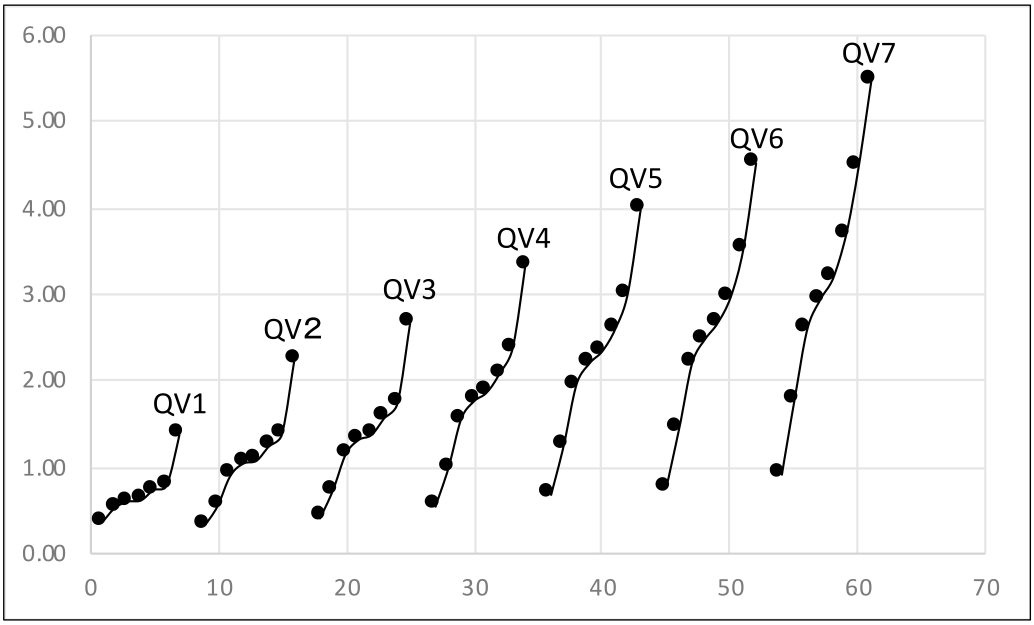


Figure 1. The Quantile Vectors ACG (QV-ACG) for Acer East.

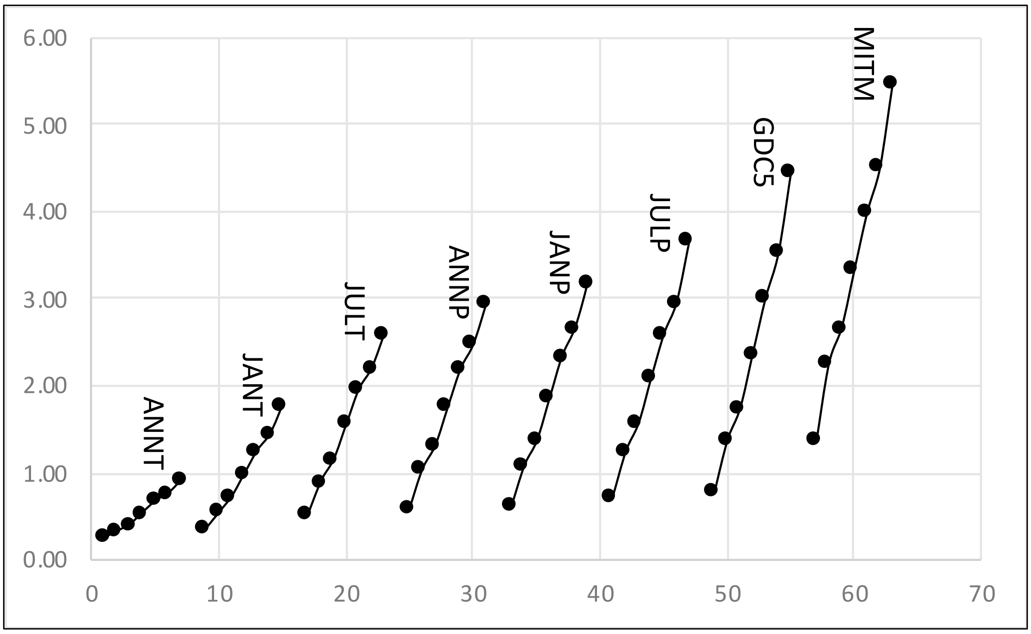


Figure 2. The Feature-Wise ACG (FW-ACG) for Acer East.

The following important observations should be noted:

1. The resulting line graphs display the accumulated sizes of the quantile vectors for the given set of features, and they do not intersect within the ACGs.
2. The shapes of the line graphs in the QV-ACG change when the order of features is altered. However, their overall lengths, or accumulated sizes, remain unchanged.

We generated the QV-ACG (Figure 3) and FW-ACG (Figure 4) for the Hardwood data, with two spaces placed between different hardwoods in the Excel column. The Total ACG (T-ACG) was obtained in Figure 5 by further accumulating the line graphs for each hardwood in the FW-ACG.

The following insights can be drawn from these ACGs:

1. Among the hardwoods, Alnus West and Quercus West have the largest line graph sizes for the 7th quantile in the QV-ACG, the 8th line graph in the FW-ACG, and the T-ACG (macroscopic property).
2. In the T-ACG, East and West hardwoods show similarities in their final segments, with the West hardwoods being larger than the East hardwoods.
3. In the QV-ACGs for each species, the difference between East and West is mainly influenced by the 7th quantile vector. Removing the 7th quantile line graphs from Figure 3 significantly reduces the differences between East and West for each species.
4. The lower portions of the graphs in both the QV-ACG and FW-ACG display similar shapes for each species.
5. In both the QV-ACG and FW-ACG, the East hardwoods exhibit general similarity to each other, with the exception of Alnus East.
6. In the QV-ACG and FW-ACG, Fraxinus West, Juglans West, and Quercus West display a general similarity, whereas Acer West and Alnus West also exhibit comparable characteristics.

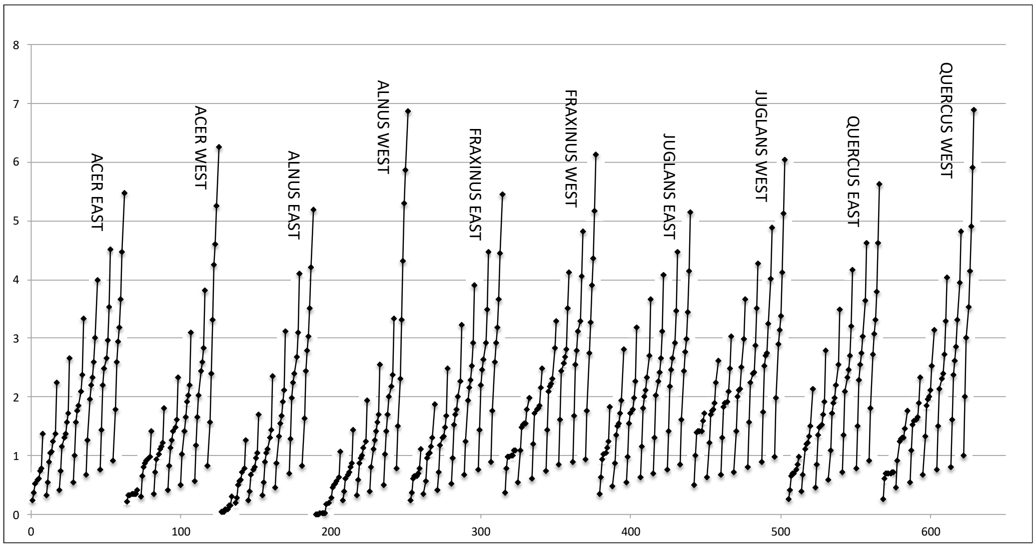


Figure 3. The QV-ACG for the Hardwood data.

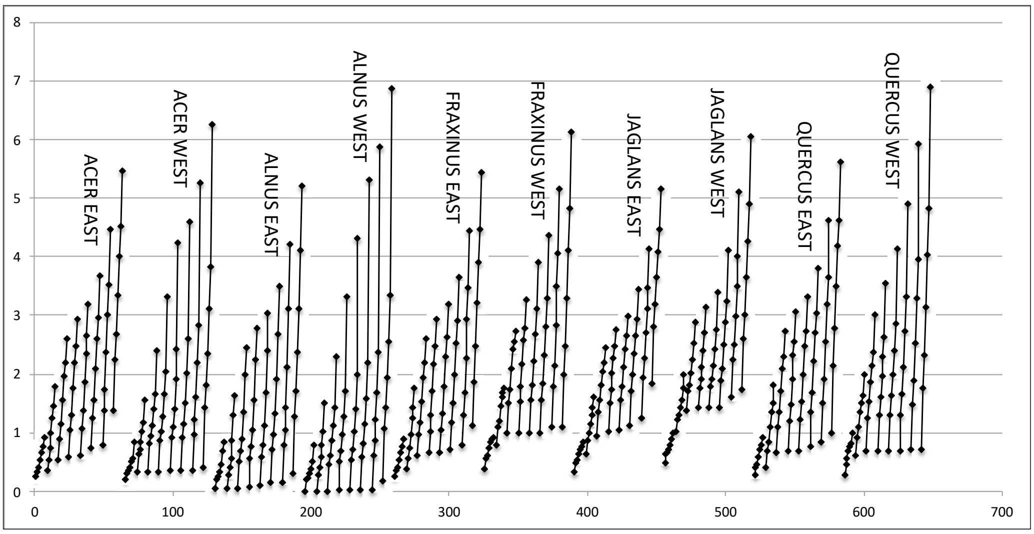


Figure 4. The FW-ACG for the Hardwood data.

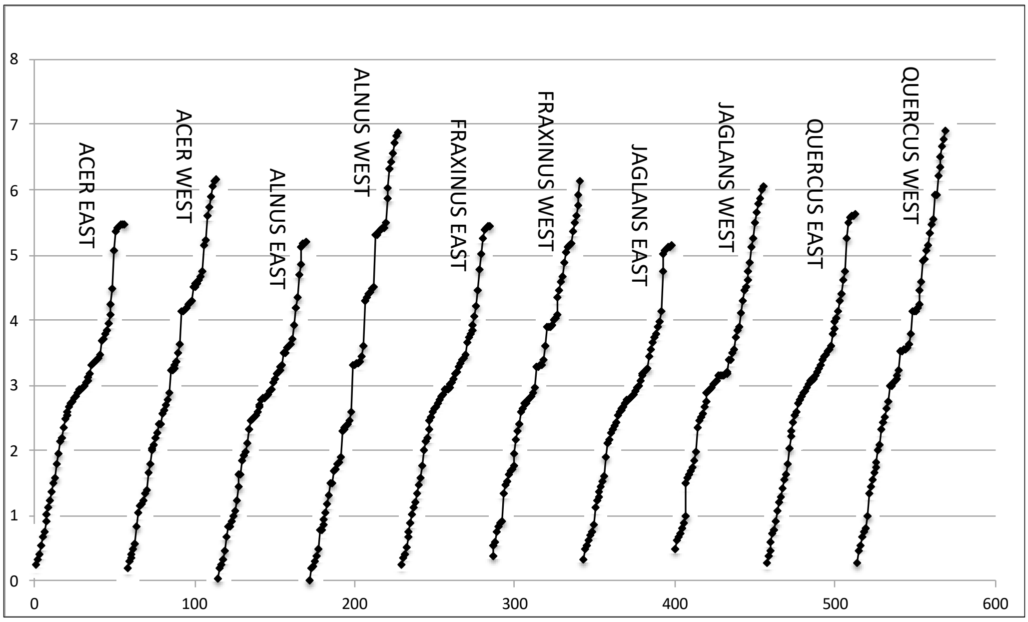


Figure 5. The T-ACG for the Hardwood data.

Figure 6 displays a scatter plot of the Hardwood data, utilizing the minimum value of QV1 and the maximum value of QV7 for each hardwood. This plot effectively demonstrates the macroscopic properties of the Hardwood data, as uncovered by the ACGs.

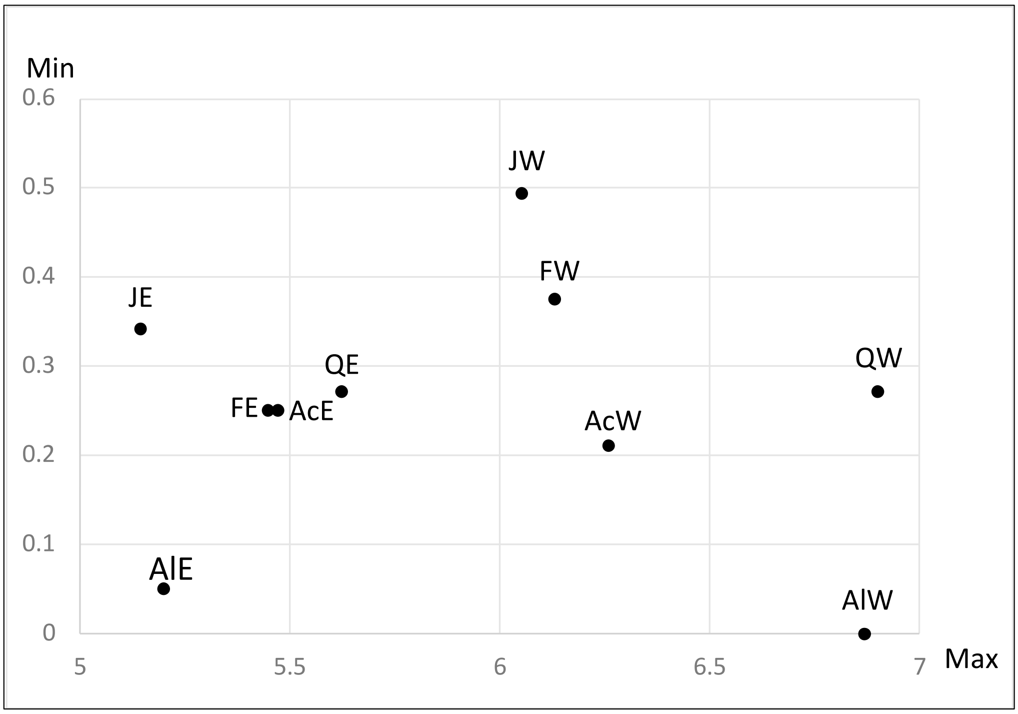


Figure 6. Scatter plot by the minimum value of QV1 and the maximum value of QV7.

2.2. *Principal Component Analysis of the Hardwood Data*

Principal component analysis (PCA) is a valuable tool for visualizing the relationships between objects in factor planes defined by the principal components. In the PCA for the Hardwood data [21], we calculate Spearman's rank order correlation matrix from the (10 $\times$ 7) 8-dimensional quantile vectors, with the results shown in Table 5. In this table, the first principal component (Pc1) represents a size factor, with similar weights assigned to all eight features, and has a significantly high contribution ratio. The second principal component (Pc2) represents a shape factor, distinguishing two groups: precipitation and moisture index; temperature and growing degree days. Figure 7 illustrates the positions of the eight features, clearly separating them into the two groups: (precipitation and moisture) and (temperature and growing degree days).

Table 5. Principal components for the Hardwood data.

Spearman	Pc1	Pc2
Eigenvalues	6.691	0.909
Contribution (%)	83.635	11.357
Eigenvectors	**Pc1**	**Pc2**
ANNT	0.362	−0.363
JANT	0.346	−0.427
JULT	0.372	−0.208
ANNP	0.359	0.369
JANP	0.337	0.365
JULP	0.352	0.170
GDC5	0.365	−0.331
MITM	0.335	0.484

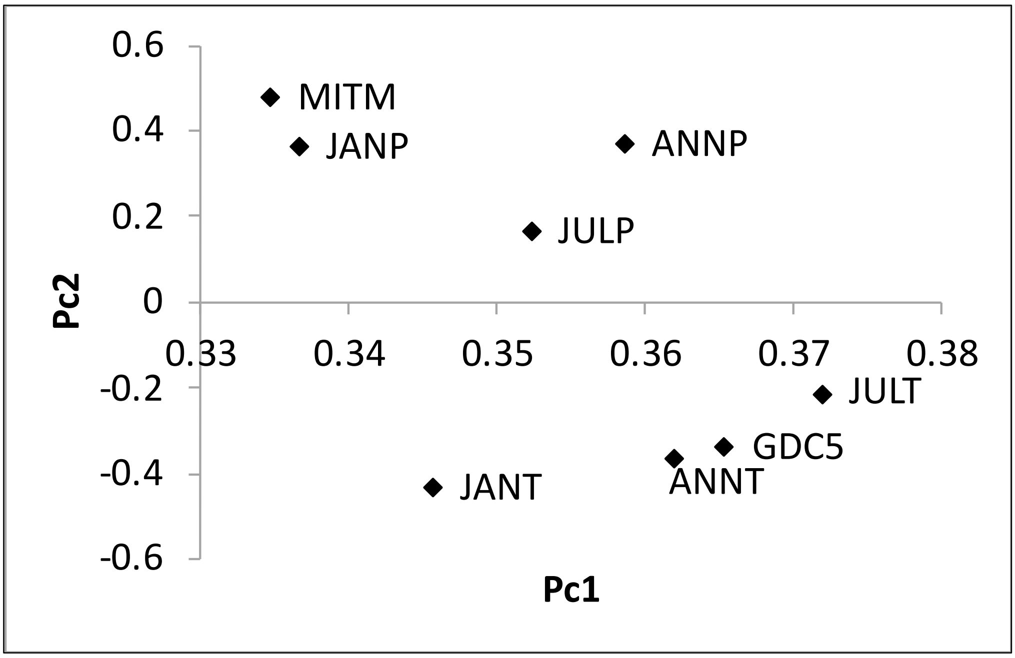


Figure 7. Scatter plot by the eigenvectors.

Figure 8 shows the results of the PCA, where each hardwood is represented by six connected lines spanning from the minimum to the maximum quantile vector. Three distinct groups are evident: (Acer West, Alnus West), (East Hardwoods), and (Fraxinus West, Juglans West, Quercus West). Alnus West and Quercus West are the largest. Additionally, the minimum and maximum quantile vectors effectively highlight the similarities and differences between hardwoods, as identified by the ACGs. The final arrow lines, connecting the 90th and 100th percentile quantile vectors, are particularly long for the West hardwoods, a characteristic that is also observed in both the QV-ACG and FW-ACG.

This highlights that visualizations using ACGs and the quantile method of PCA are effective tools for gaining insights into the data.

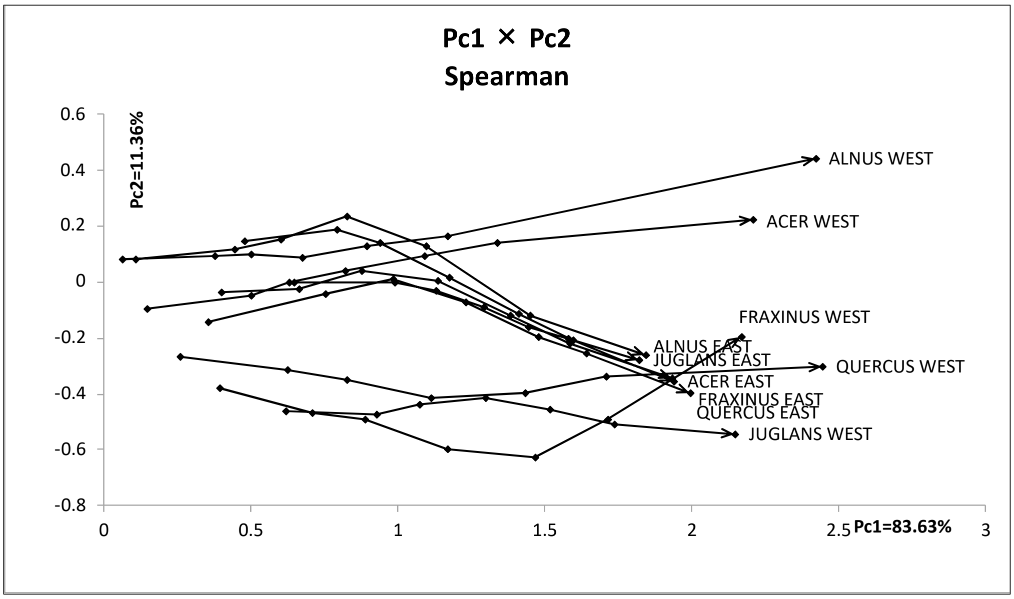


Figure 8. The result of PCA for the Hardwood data.

3. Analysis of Periodically Summarized Multiple Data Tables (Prefecture Profile Data I)

3.1. Accumulated Concept Graphs for the Prefecture Profile Data I

We have n periodically summarized data tables with the same structure, where each table consists of N objects described by d features. This results in a three-way data table in the form of $n \times N \times d$.

In the Prefecture Profile Data I [22] dataset, the number of people employed in ten different job categories, as shown in Table 6, was recorded across 47 prefectures of Japan, from Hokkaido to Okinawa, for the years 1980, 1985, 1990, 1995, 2000, and 2005. We represent these data as (47 objects) × (6 quantile vectors) × (10 features). As part of our analysis, Tables 7 and 8 present a summary of the (0–1)-normalized quantile vectors and the accumulated quantile vectors for Hokkaido, respectively.

Table 6. Ten job types in Prefecture Profile Data I [22].

Feature	Description
F1	Professional skills
F2	Management jobs
F3	Office works
F4	Sales jobs
F5	Service jobs
F6	Security jobs
F7	Agricultural forestry and fisheries
F8	Transportation and communication
F9	Industrial process work
F10	Unclassified jobs

Table 7. The (0–1)-normalized quantile vectors for Hokkaido for Prefecture Profile Data I.

Hokkaido	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
1980	0.040	0.070	0.054	0.064	0.055	0.059	0.129	0.087	0.134	0.002
1985	0.096	0.134	0.121	0.137	0.119	0.128	0.279	0.185	0.286	0.009
1990	0.165	0.204	0.198	0.216	0.189	0.202	0.408	0.287	0.448	0.025
1995	0.243	0.280	0.279	0.303	0.269	0.285	0.532	0.395	0.608	0.041
2000	0.327	0.331	0.360	0.392	0.358	0.377	0.642	0.501	0.764	0.082
2005	0.413	0.373	0.442	0.476	0.455	0.480	0.744	0.601	0.920	0.180

Table 8. The accumulated quantile vectors for Hokkaido for Prefecture Profile Data I.

Hokkaido	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
1980	0.040	0.110	0.164	0.228	0.282	0.341	0.47	0.557	0.691	0.693
1985	0.096	0.23	0.351	0.488	0.607	0.735	1.015	1.200	1.486	1.495
1990	0.165	0.369	0.567	0.783	0.972	1.174	1.582	1.869	2.317	2.342
1995	0.243	0.524	0.802	1.106	1.374	1.659	2.192	2.587	3.195	3.236
2000	0.327	0.658	1.018	1.410	1.768	2.145	2.787	3.288	4.052	4.134
2005	0.413	0.786	1.228	1.703	2.158	2.638	3.382	3.983	4.903	5.083

Figure 9 presents the QV-ACG, where each prefecture is represented by six monotone line graphs, with a spacing of one unit between each line graph and a gap of two units between different prefectures. Similarly, Figure 10 shows the FW-ACG, where each prefecture is depicted by ten monotone line graphs, with the same spacing arrangement.

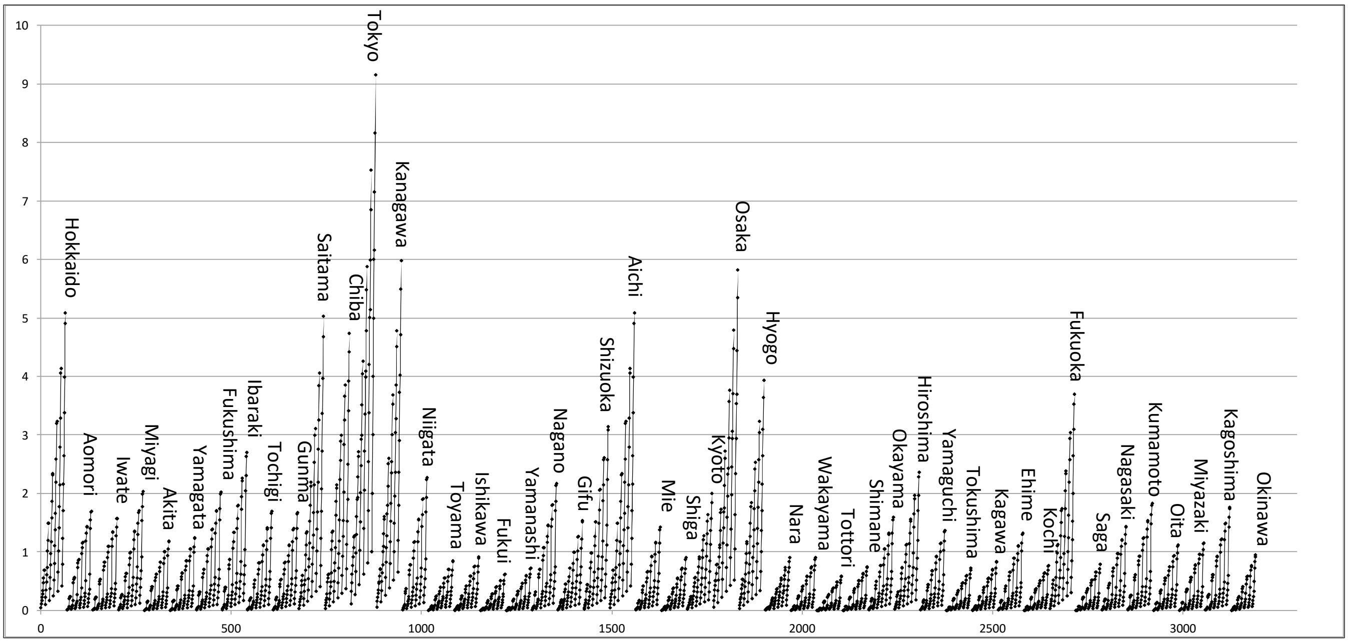


Figure 9. The QV-ACG for Prefecture Profile Data I.

From the QV-ACG and FW-ACG, we can observe the following:

1. Tokyo is the largest prefecture, while Tottori is the smallest. The length of the line graphs primarily reflects the population size of each prefecture.
2. By analyzing the patterns in the QV-ACGs and FW-ACGs, it is easy to identify similar prefectures. For example, Aomori and Iwate, Akita and Yamagata, Tochigi and Gunma, and Toyama and Ishikawa share similar patterns. Additionally, it is straightforward to distinguish between rural and urban areas.
3. As a microscopic observation, Tokyo, Kanagawa, and Osaka have significantly higher numbers of people employed in unclassified jobs compared to other prefectures.
4. In 1980, many rural prefectures, such as Akita and Aomori, show very low starting values in the first six positions in the line graph (corresponding to service jobs), while the last four positions (related to production jobs) display noticeably higher values.
5. In many cases, such as in Kochi, Iwate, and Tochigi, the line graphs for service jobs remain short. In contrast, prefectures like Ibaraki and Chiba show a significant increase in the length of line graphs for service and management jobs in later years.

It is important to emphasize that visualizing data through ACGs allows us to effectively capture both the macroscopic and microscopic similarities and differences between prefectures in two-dimensional figures.

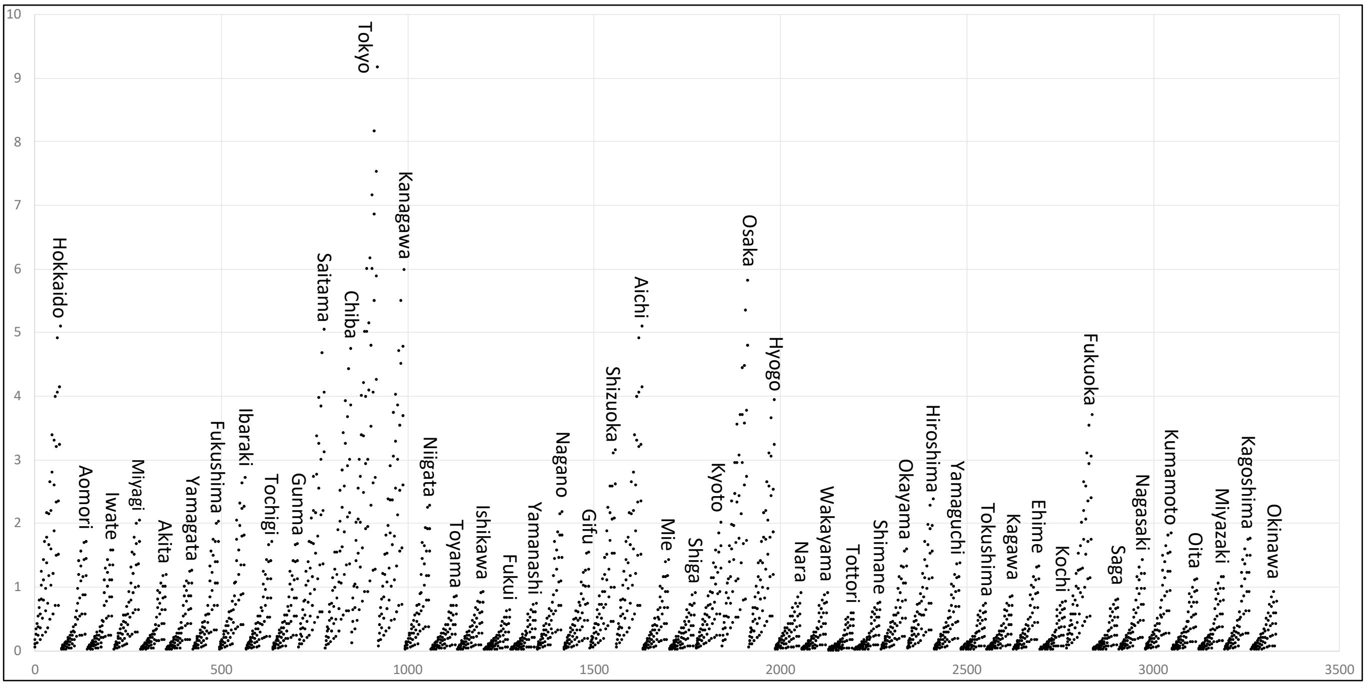


Figure 10. The FW-ACG for Prefecture Profile Data I.

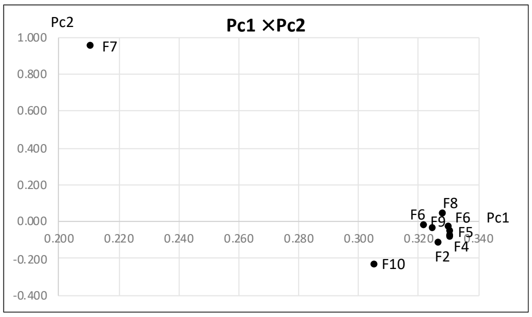
3.2. Principal Component Analysis for the Prefecture Profile Data I

We derive the Spearman's rank–order correlation matrix from the (47 × 6) 10-dimensional quantile vectors. Table 9 presents the resulting principal components. The first principal component, Pc1, represents the size factor, with a very high contribution ratio. In the second principal component, Pc2, F7 (agriculture, forestry, and fishery) has a notably large positive value. Similarly, in the third principal component, Pc3, F10 (unclassified jobs) shows a significantly large positive value. Figure 11a,b illustrate the relationships among the ten features, as represented by pairs of eigenvectors (Pc1, Pc2) and (Pc1, Pc3), respectively.

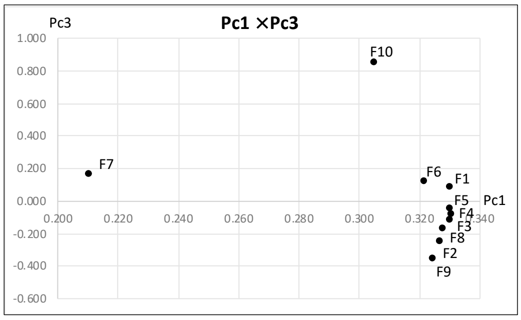
Table 9. The principal components for Prefecture Profile Data I.

Spearman	Pc1	Pc2	Pc3
Eigenvalues	9.035	0.651	0.161
Contribution (%)	90.354	6.514	1.612
Eigenvectors	**Pc1**	**Pc2**	**Pc3**
F1	0.331	−0.057	0.077
F2	0.327	−0.119	−0.256
F3	0.331	−0.087	−0.121
F4	0.331	−0.078	−0.094
F5	0.331	−0.030	−0.055
F6	0.322	−0.026	0.112
F7	0.211	0.954	0.152
F8	0.328	0.041	0.180
F9	0.325	−0.039	−0.363
F10	0.306	−0.233	0.838

Figures 12 and 13 show the results of the PCA. In these figures, the zoomed-in results are obtained by removing ten large prefectures from Tokyo to Shizuoka. We have the following facts.

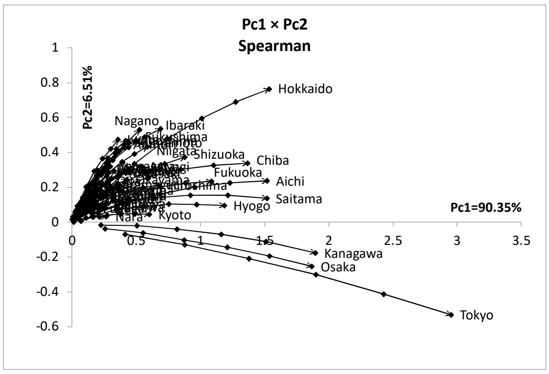


(a) Plot by (Pc1, Pc2)

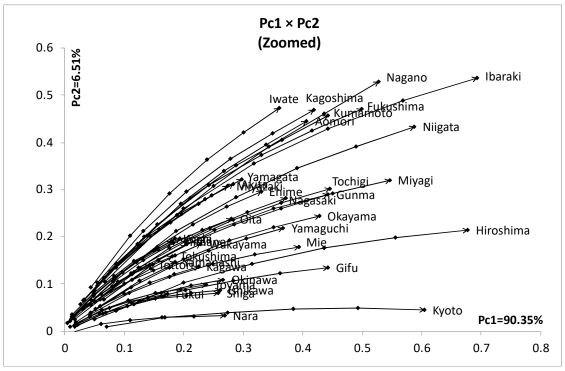


(b) Plot by (Pc1, Pc3)

Figure 11. Scatter plots by (Pc1, Pc2) and (Pc1, Pc3) for Prefecture Profile Data I.

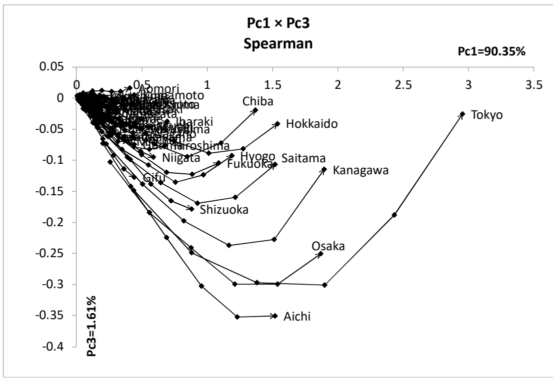


(a) 47 prefectures

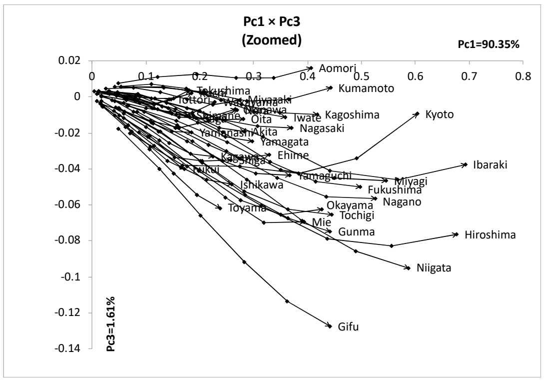


(b) 37 prefectures

Figure 12. Results of PCA in the factor plane by (Pc1, Pc2) for Prefecture Profile Data I.



(a) 47 prefectures



(b) 37 prefectures

Figure 13. Results of PCA in the factor plane by (Pc1, Pc3) for Prefecture Profile Data I.

1. In 1980, with the exceptions of Tokyo, Kanagawa, and Osaka, other prefectures existed in a narrow region in the factor planes. As time goes on, each prefecture spreads in their respective direction in the factor planes.
2. In the factor plane by (Pc1, Pc2), many prefectures grow up with addition of other job types to F7, i.e., agriculture, forestry, and fishery.
3. In the factor plane by (Pc1, Pc3), nine large prefectures from Tokyo to Fukuoka are affected by job type F10, i.e., unclassified jobs. In the zoomed-in factor plane, Aomori, Kumamoto, Kyoto, Ibaraki, and Hiroshima show the same tendency.

These findings, obtained through the PCA, may be useful in understanding the Prefecture Profile Data I together with the QV-ACG data, shown in Figure 9, and the FW-ACG data, shown in Figure 10.

4. Analysis of Multiple Different Sized Data (Prefecture Profile Data II)

4.1. Total ACG of the Prefecture Profile Data II

In the Prefecture Profile Data II [22] dataset, seven different data tables, summarized in Table 10, describe the profiles of 47 Japanese prefectures in 2010. Each feature is represented by a numerical value drawn from a set of possible feature values.

Table 10. Seven different data tables for Prefecture Profile Data II.

Table	Feature	Description
1	Number of employed persons (19 job types, 1000 people)	1. Agriculture and forestry; 2. Fisheries; 3. Mining and quarrying of stone and gravel; 4. Construction; 5. Manufacturing; 6. Electricity; gas, head supply, and water; 7. Information and communications; 8. Transport and postal activities; 9. Wholesale and retail trade; 10. Finance and insurance; 11. Real estate and goods rental and leasing; 12. Scientific research, professional, and technical services; 13. Accommodations, eating, and drinking services; 14. Living-related and personal services and amusement services; 15. Education, learning support; 16. Medical, healthcare, and welfare; 17. Compound services; 18. Services (not elsewhere classified); 19. Government, expect elsewhere classified.
2	Nominal GDP	(JPY 10 billion)
3	Temperature	1. Minimum temperature; 2. Maximum temperature
4	Area	(Square kilometer)
5	People in 18 age categories	1. [0, 4], 2. [5, 9], 3. [10, 14], 4. [15, 19], 5. [20, 24], 6. [25, 29], 7. [30, 34], 8.[35, 39], 9. [40, 44], 10. [45, 49], 11. [50, 54], 12. [55, 59], 13. [60, 64], 14. [65, 69], 15. [70, 74], 16. [75, 79], 17. [80, 84], 18. [85-]
6	Number of people	1. Birth; 2. Death; 3. Marriage; 4. Divorce.
7	Number of households	Private household: 1. Number of households; 2. Number of household members. Industrial households: 3. Number of households; 4. Number of household members.

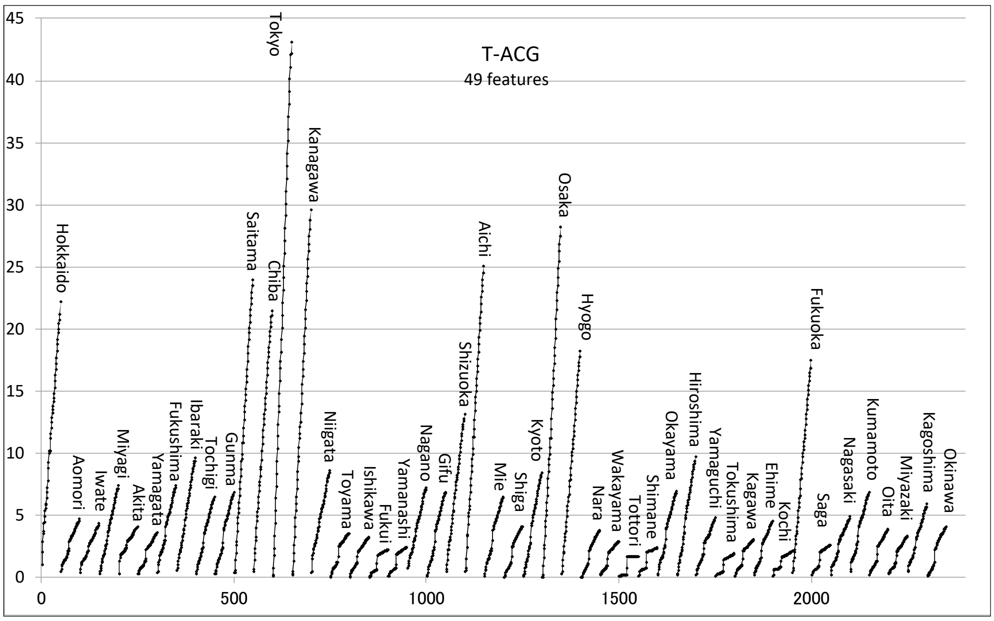


Figure 14. The T-ACG for Prefecture Profile Data II.

By merging the seven data tables from Table 10 into one large table, we create a two-way dataset with a size of (47 prefectures) × (49 features). The Total ACG (T-ACG) shown in Figure 14 is obtained by accumulating the (0–1)-normalized 49 features. From this figure, we can observe the following insights:

1. The largest ten prefectures are consistent with those identified in the QV-ACG and FW-ACG for Prefecture Profile Data I, while Tottori remains the smallest.

2. Macroscopically similar prefectures include: (Aomori and Iwate), (Tochigi and Gunma), (Toyama and Ishikawa), (Fukui and Yamanashi), (Gifu and Mie), (Okayama and Kumamoto), and (Shiga, Nara, and Oita).

Figure 15 presents a scatterplot of the 47 prefectures, using two values: F1 (agriculture and forestry) and the total accumulated value, F49. This figure effectively illustrates the macroscopic properties of the 47 prefectures as revealed by the T-ACG in Figure 14.

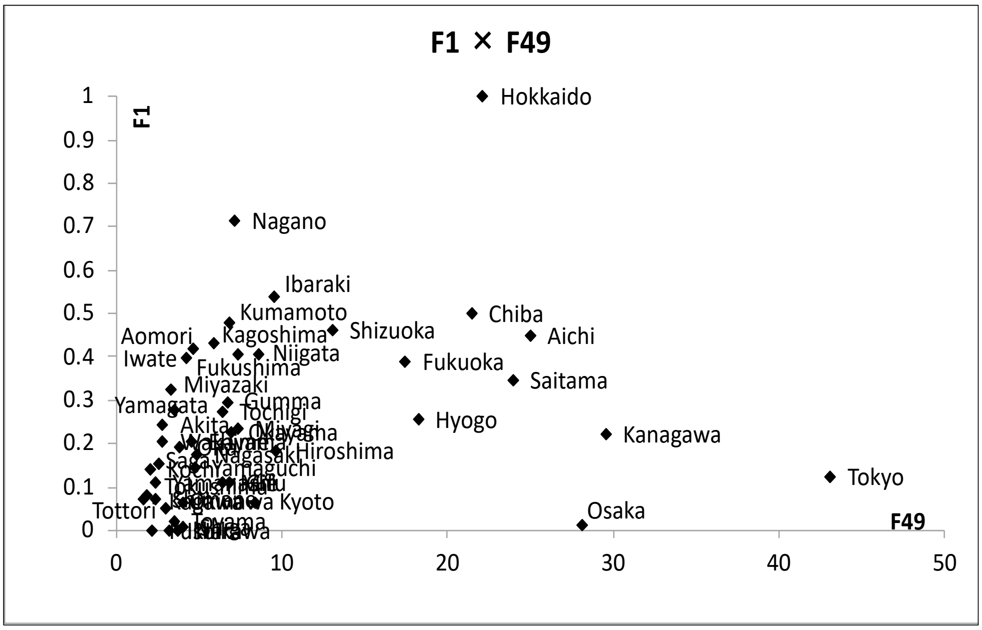


Figure 15. Plot by the minimum and the maximum values of the T-ACGs.

Further findings using the QV-ACG, FW-ACG, and PCA are discussed in the next section.

4.2. The Quantile Method of the ACGs for the Prefecture Profile Data II

In the analysis of the Hardwood data and Prefecture Profile Data I, the quantile methods effectively detected the microscopic patterns within the datasets. To extend this approach, we combined the seven tables from Table 10 into a single table, as shown in Table 11, where each of the eleven features is represented by five quantile values. In Table 11, features F4, F5, F9, F10, and F11 include a value labeled as "Dummy", which we assume to be zero. Consequently, our dataset takes the following form: (47 prefectures) × (5 quantile vectors) × (11 features).

Table 11. Eleven features described by five quantile values.

Feature	Description
F1	1. Agriculture and forestry; 2. Fisheries; 3. Mining and quarrying of stone and graves; 4. Construction; 5. Manufacturing
F2	1. Electricity, gas, heat supply, and water; 2. Information and communications; 3. Transport and postal activities; 4. Wholesale and retail trade; 4. Finance and insurance;
F3	1. Real estate and goods rental and leasing; 2. Scientific research, professional, and technical services; 3. Accommodations, eating, and drinking services; 4. Living-related and personal services and amusement services; 5. Education, learning support;
F4	1. Medical, healthcare, and welfare; 2. Compound services; 3. Services (not elsewhere classified); 4. Government, expect elsewhere classified; 5. Dummy
F5	1. Nominal GDP; 2. Minimum temperature; 3. Maximum temperature; 4. Area; 5. Dummy.
F6	1. [0, 4], 2. [5, 9], 3. [10, 14], 4. [15, 19], 5. [20, 24]
F7	1. [25, 29], 2. [30, 34], 3. [35, 39], 4. [40, 44], 5. [45, 49]
F8	1. [50, 54], 2. [55, 59], 3. [60, 64], 4. [65, 69], 5. [70, 74]
F9	1. [75, 79], 2. [80, 84], 3. [85-], 4. Dummy, 5. Dummy.
F10	1. Birth; 2. Death; 3. Marriage; 4. Divorce; 5. Dummy.
F11	Private household: 1. Number of households; 2. Number of household members. Industrial households: 3. Number of households; 4. Number of household members; 5. Dummy.

For each of the eleven features, we calculated the (0–1)-normalized quantile values for the 47 prefectures. Given that the features have different units, it is essential to emphasize the importance of (0–1)-normalization. Tables 12 and 13 display the (0–1)-normalized quantile vectors and the accumulated quantile vectors for Hokkaido, respectively. Figures 16 and 17 show the QV-ACG and FW-ACG for Prefecture Profile Data II, respectively.

Table 12. The (0–1) normalized quantile vectors for Hokkaido for Prefecture Profile Data II.

QV	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11
1	0.270	0.124	0.043	0.166	0.081	0.075	0.059	0.089	0.164	0.086	0.117
2	0.540	0.144	0.078	0.455	0.081	0.159	0.120	0.196	0.349	0.210	0.243
3	0.776	0.247	0.164	0.580	0.081	0.250	0.148	0.293	0.548	0.282	0.259
4	0.946	0.334	0.248	0.806	0.516	0.334	0.251	0.383	0.548	0.396	0.587
5	1.000	0.374	0.331	0.806	0.516	0.398	0.321	0.476	0.548	0.396	0.587

Table 13. The accumulated quantile vectors for Hokkaido for Prefecture Profile Data II.

QV	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11
1	0.270	0.394	0.437	0.603	0.684	0.759	0.818	0.907	1.070	1.156	1.273
2	0.540	0.684	0.763	1.217	1.298	1.458	1.578	1.774	2.123	2.333	2.576
3	0.776	1.023	1.187	1.767	1.848	2.097	2.281	2.574	3.122	3.405	3.663
4	0.946	1.280	1.528	2.334	2.850	3.184	3.434	3.817	4.365	4.761	5.349
5	1.000	1.374	1.705	2.511	3.027	3.425	3.746	4.222	4.770	5.167	5.754

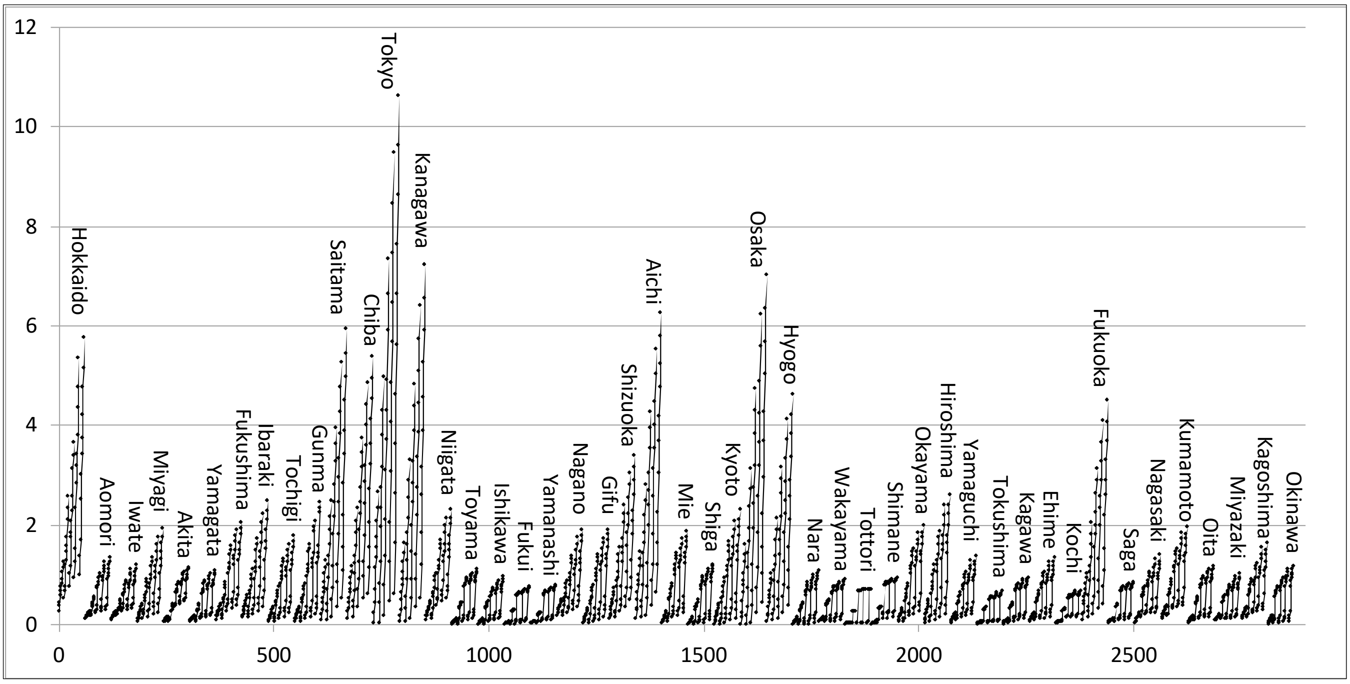


Figure 16. The QV-ACG for the Prefecture Profile Data II.

The following insights can be drawn from the QV-ACG and FW-ACG:

1. The ten largest prefectures remain in the same order in terms of macroscopic size, consistent with the T-ACG: Tokyo > Kanagawa > Osaka > Aichi > Saitama > Hokkaido > Chiba > Hyogo > Fukuoka > Shizuoka.
2. The smallest prefecture is not Tottori but Tokushima.

3. Some prefecture pairs that appeared similar in the T-ACG become dissimilar in the QV-ACG and FW-ACG. For instance, Akita and Yamagata differ in features F1~F4, Tochigi and Gunma differ in F9~F11, and Toyama and Ishikawa differ in F4~F11.

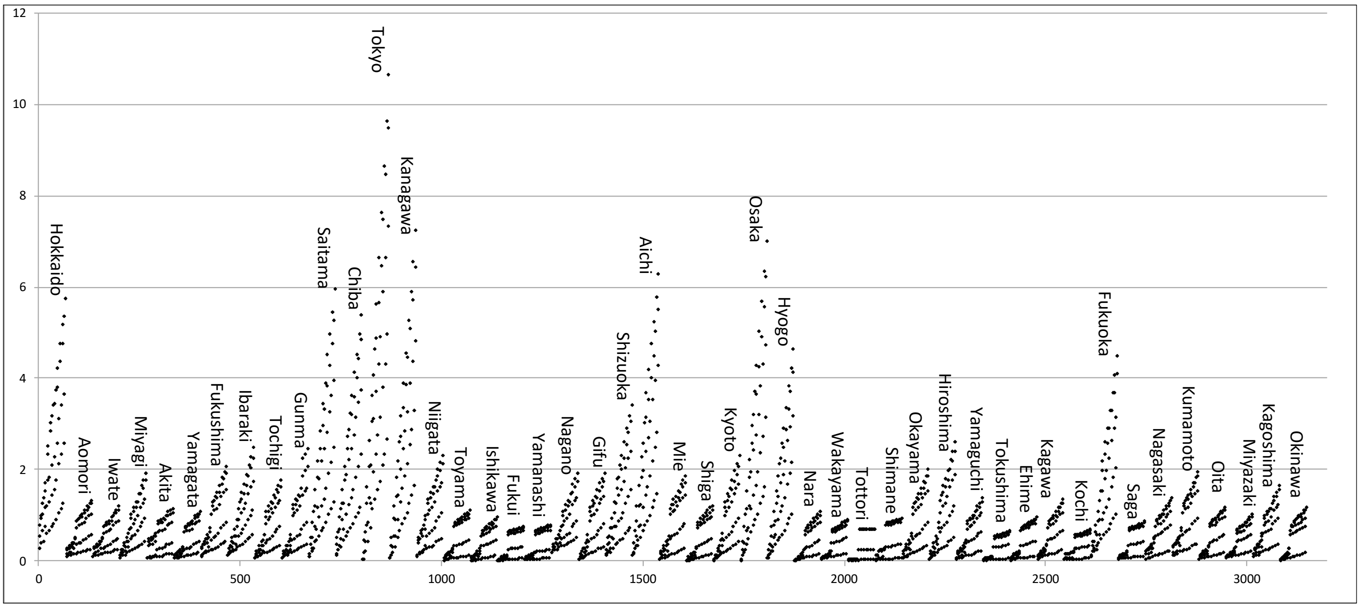


Figure 17. The FW-ACG for the Prefecture Profile Data II.

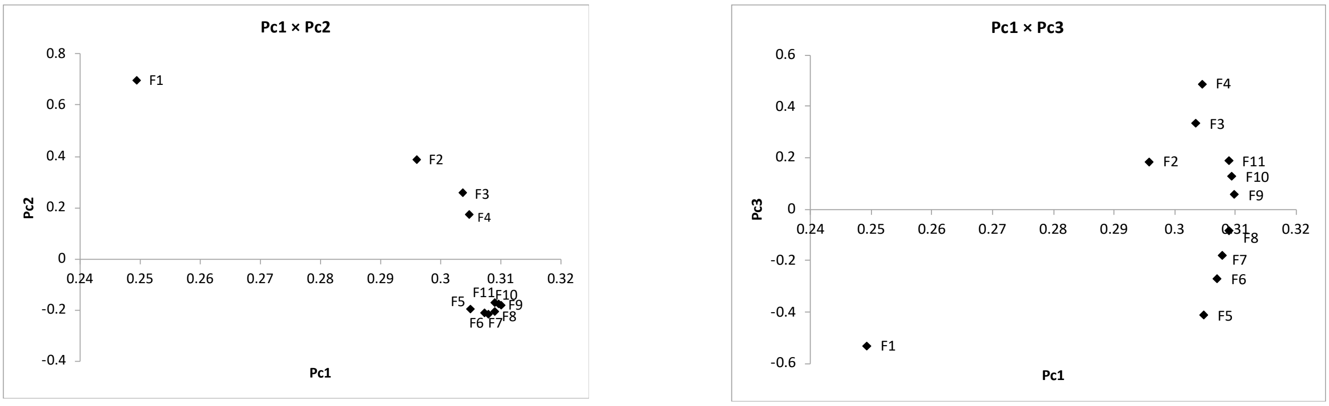
Overall, the QV-ACG and FW-ACG provide more detailed information than the T-ACG in the analysis of Prefecture Profile Data II.

4.3. The Quantile Method of PCA for the Prefecture Profile Data II

We calculate the Spearman's rank–order correlation matrix from the (47 $\times$ 5) 11-dimensional quantile vectors. Table 14 presents the resulting principal components. In this table, the first principal component (Pc1) represents the size factor, with a notably high contribution ratio. In Pc2, features F1~F4 have positive values, with F1 carrying a significant weight, while features F5~F11 show relatively negative weights, as illustrated in Figure 18a. In Pc3, F4 has a substantial positive weight, while F1 and F5 exhibit large negative weights. The remaining features span a broad range of values, as shown in Figure 18b.

Table 14. The principal components for the Prefecture Profile Data II.

Spearman	**Pc1**	**Pc2**	**Pc3**
Eigenvalues	10.155	0.660	0.146
Contribution (%)	92.322	6.003	1.325
Eigenvectors	**Pc1**	**Pc2**	**Pc3**
F1	0.249	0.699	−0.529
F2	0.296	0.387	0.187
F3	0.304	0.256	0.334
F4	0.305	0.172	0.487
F5	0.305	−0.196	−0.411
F6	0.307	−0.211	−0.270
F7	0.308	−0.217	−0.179
F8	0.309	−0.207	−0.082
F9	0.310	−0.179	0.057
F10	0.310	−0.176	0.128
F11	0.309	−0.172	0.188



(**a**) Plot by (Pc1, Pc2) (**b**) Plot by (Pc1, Pc3)

Figure 18. The FW-ACG for the Prefecture Profile Data II.

Figures 19 and 20 are the results of the quantile method of PCA in the factor planes (Pc1, Pc2) and (Pc1, Pc3). Figure 21 is the zoomed-in result in the factor plane (Pc1, PC3).

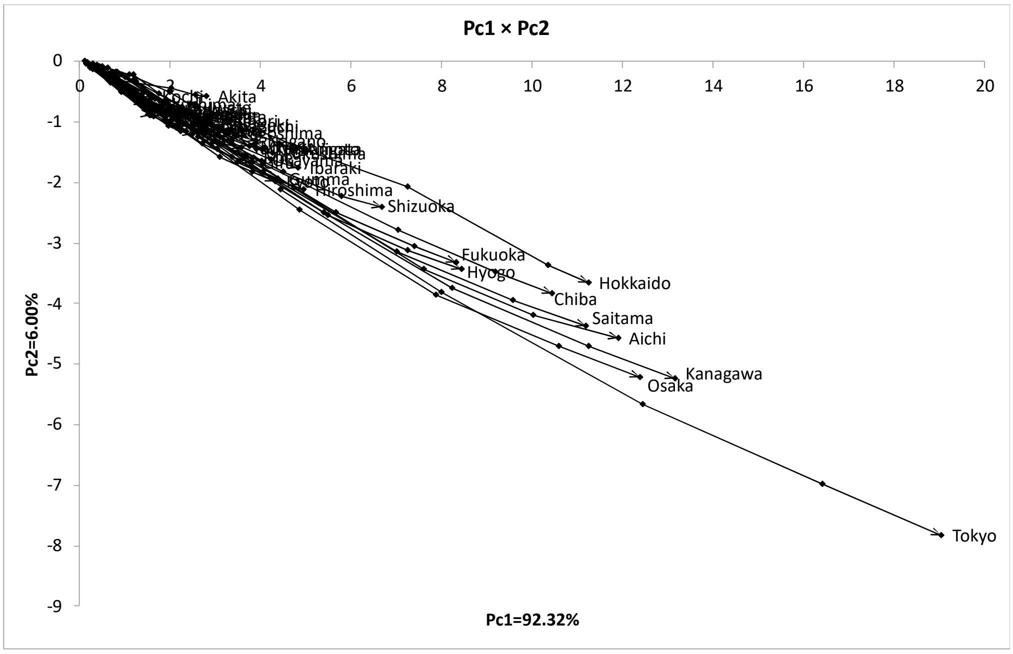


Figure 19. The FW-ACG for the Prefecture Profile Data II.

The following observations have been made:

1. The ten largest prefectures in the factor planes (Pc1, Pc2) and (Pc1, Pc3) align with the results from the QV-ACG and FW-ACG analyses.
2. Many prefectures are concentrated in a narrow region of the factor plane (Pc1, Pc2).
3. Tottori is isolated from other prefectures in the plane (Pc1, Pc3), where QV3~QV5 almost overlap.
4. In the zoomed-in factor plane, most prefectures show an upward trend, except Tottori. Notably, Gunma displays significant movement in the final portion due to feature F9, while Tokushima exhibits the smallest size. Among similar prefecture pairs, Okayama and Kumamoto trace comparable curves.

It is important to note that the results from both PCA and ACGs complement each other, enhancing the analysis and understanding of the three-way data.

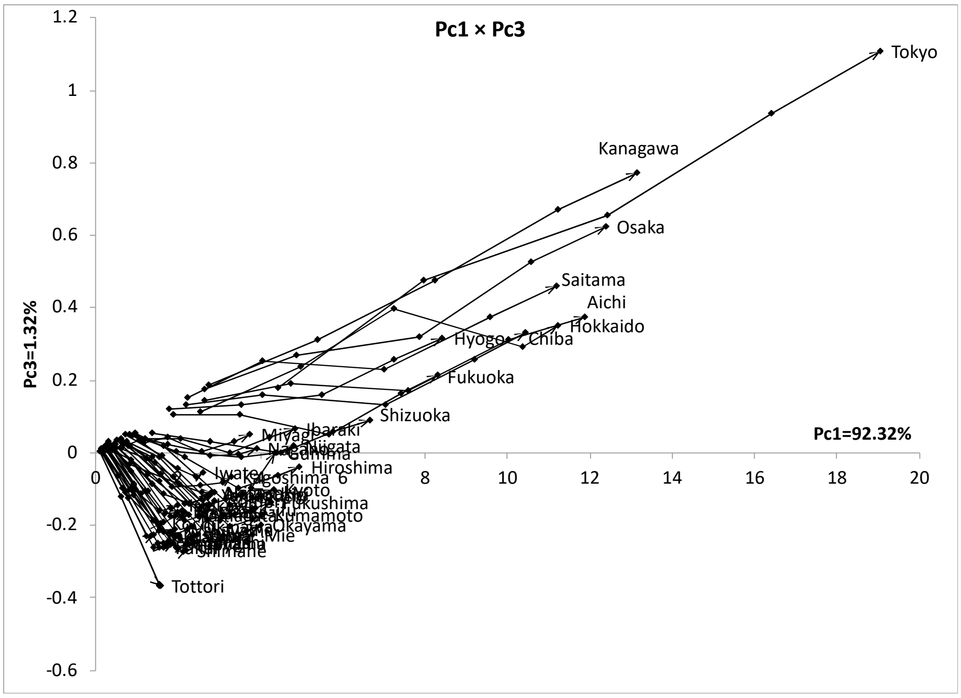


Figure 20. The FW-ACG for the Prefecture Profile Data II.

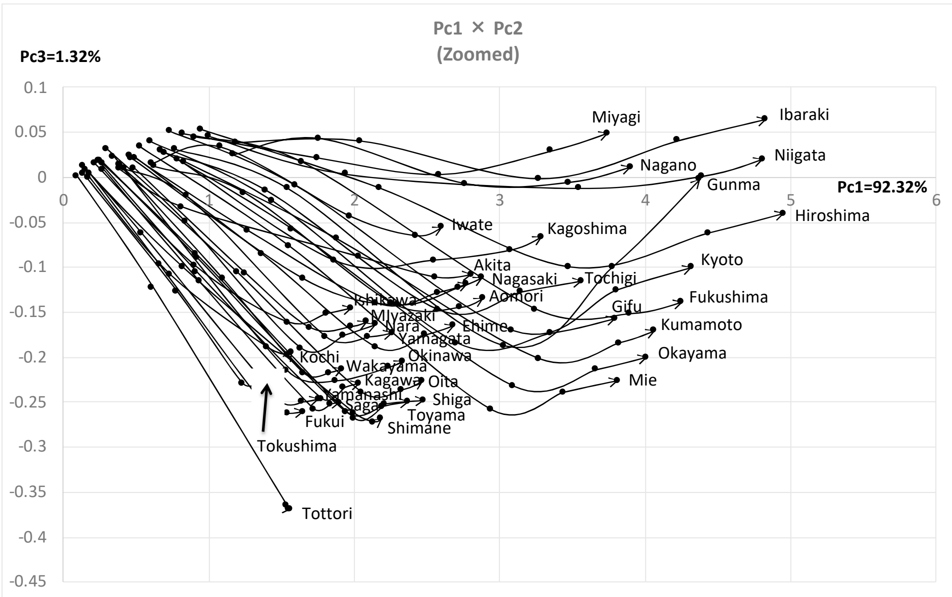


Figure 21. The FW-ACG for the Prefecture Profile Data II.

5. Discussion

This paper demonstrates the effectiveness of Accumulated Concept Graphs (ACGs) and their collaborative use with the quantile method of PCA for analyzing three-way datasets. The advantages of ACGs, as illustrated by the examples, are as follows:

1. **Universal Approach with Data Specificity:** ACGs are versatile and can be applied to various types of symbolic data while still capturing detailed microscopic properties within unaggregated data.
2. **Simplicity:** Transforming three-way data into a distributional format is computationally efficient, making it suitable for large datasets.

3. **Microscopic and Macroscopic Properties:** ACGs highlight macroscopic differences between objects through the total lengths of line graphs, while microscopic differences are revealed through the local shapes created by accumulated values at specific points.
4. **Outlier Detection:** Unlike traditional visualizations like parallel coordinates or radar charts, ACGs feature parallel monotone line graphs that never intersect, making it easier to detect outliers.
5. **Enabling Classical Analysis on Symbolic Data:** ACGs and the transformation of symbolic data into a distributional format allow classical methods like PCA to be applied to symbolic data, which would otherwise be computationally demanding or impractical.

Additionally, ACGs can be easily created using scatter plots in Excel, and the proposed methods can be extended to more complex symbolic data. Future work could explore how ACGs can enhance other areas of data analysis, such as clustering, especially considering the computational challenges of existing clustering methods for symbolic data.

In conclusion, this paper proposes a visualization technique using simple line graphs, termed Accumulated Concept Graphs, for three-way and symbolic data. This approach enables the visualization of both macroscopic and microscopic details embedded in the data. The primary contribution of the method is its ability to provide a simple yet comprehensive visual overview of complex relationships within the dataset. By facilitating exploratory data analysis through visual interpretation, the proposed method aids analysts in making informed decisions about further analyses. Furthermore, this method can be applied to datasets with intricate internal structures that are difficult to visualize using currently available techniques.

Author Contributions: Conceptualization and methodology, M.I. and H.Y.; software and validation, K.U.; original draft presentation, M.I. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by JSPS KAKWNHI (Grants-in-Aid for Scientific Research) Grant Number 253000268. Part of this work has been conducted under JSPS International Research Fellow program.

Data Availability Statement: The data presented in this study are openly available in the websites of the references [21,22].

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Franconeri, S.L.; Padilla, L.M.; Shah, P.; Zacks, J.M.; Hullman, J. The Science of Visual Data Communication: What Works. *Psychol. Sci. Public Interest* **2021**, *22*, 110–161. [CrossRef] [PubMed]
2. Healy, K. *Data Visualization: A Practical Introduction*; Princeton University Press: Princeton, NJ, USA, 2018.
3. Keim, D.A. Information Visualization and Visual Data Mining. *IEEE Trans. Vis. Comput. Graph.* **2002**, *8*, 1–8. [CrossRef]
4. Tukey, J.W. *Exploratory Data Analysis*; Addison-Wesley Publishing Company: San Francisco, CA, USA, 1977; Volume 2.
5. Tukey, P. *Graphical Methods for Data Analysis*; Wadsworth: Belmont, CA, USA, 1983.
6. Billard, L.; Diday, E. *Symbolic Data Analysis: Conceptual Statistics and Data Mining*; Wiley & Sons: Hoboken, NJ, USA, 2006.
7. Khalid, Z.M.; Zeebaree, S.R. Big Data Analysis for Data Visualization: A Review. *Int. J. Sci. Bus.* **2021**, *5*, 64–75.
8. Elhamdadi, H.; Stefkovics, A.; Beyer, J.; Moerth, E.; Pfister, H.; Bearfield, C.X.; Nobre, C. Vistrust: A Multidimensional Framework and Empirical Study of Trust in Data Visualizations. *IEEE Trans. Vis. Comput. Graph.* **2023**, *30*, 348–358. [CrossRef] [PubMed]
9. Noirhomme-Fraiture, M. Visualization of Large Data Sets: The Zoomstar Solution. *Int. Electron. J. Symb. Data Anal.* **2002**, 26–39. Available online: https://www.researchgate.net/publication/228615915_Visualization_of_large_data_sets_The_Zoom_Star_solution (accessed on 25 August 2024).
10. Diday, E.; Noirhomme-Fraiture, M. *Symbolic Data Analysis and the SODAS Software*; Wiley & Sons: Hoboken, NJ, USA, 2008.
11. Verde, R.; Lechevallier, Y.; Chavent, M. Symbolic Clustering Interpretation and Visualization. *Electron. J. Symb. Data Anal.* **2003**, *1*, 1.
12. Cui, W.; Strazdins, G.; Wang, H. Visual Analysis of Multidimensional Big Data: A Scalable Lightweight Bundling Method for Parallel Coordinates. *IEEE Trans. Big Data* **2021**, *9*, 106–117. [CrossRef]
13. Su, E.C.Y.; Wu, H.M. Dimension Reduction and Visualization of Multiple Time Series Data: A Symbolic Data Analysis Approach. *Comput. Stat.* **2024**, *39*, 1937–1969. [CrossRef]

14. Pelka, M. Outlier Identification for Symbolic Data with the Application of the DBSCAN Algorithm. In *Conference of the Section on Classification and Data Analysis of the Polish Statistical Association*; Springer International Publishing: Cham, Switzerland, 2021; pp. 53–62.
15. Umbleja, K.; Ichino, M.; Yaguchi, H. Improving Symbolic Data Visualization for Pattern Recognition and Knowledge Discovery. *Vis. Inform.* **2020**, *4*, 23–31. [CrossRef]
16. Nguyen, H.; Rosen, P.; Wang, B. Visual Exploration of Multiway Dependencies in Multivariate Data. In *SIGGRAPH ASIA 2016 Symposium on Visualization*; ACM: New York, NY, USA, 2016.
17. Sarkar, D. *Lattice: Multivariate Data Visualization with R*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2008.
18. Todorov, V.; Di Palma, M.; Gallo, M. Robust Tools for Visualization in Three-way Analysis. In Proceedings of the International Conference on Robust Statistics, Leuven, Belgium, 2–6 July 2018.
19. Ichino, M. The Quantile Method for Symbolic Principal Component Analysis. *Stat. Anal. Data Min. ASA Data Sci. J.* **2011**, *4*, 184–198. [CrossRef]
20. Ichino, M.; Britto, P. The Data Accumulation Graph (DAG) to Visualize Multidimensional Symbolic Data. In Proceedings of the Workshop in Symbolic Data Analysis, Taipei, Taiwan, 14–16 June 2014.
21. US Geological Survey. Tables of Histogram Data: Climate-Vegetation Atlas of North America. 2013. Available online: http://pubs.usgs.gov/pp/p1650-b/datatables/hgtable.xls (accessed on 2 October 2018).
22. National Statistics Center Japan. Statistics of Japan. 2018. Available online: https://www.e-stat.go.jp/en (accessed on 8 October 2018).