Scientific Research Publishing

Modern Economy, 2021, 12, 1593-1611
https://www.scirp.org/journal/me
ISSN Online: 2152-7261
ISSN Print: 2152-7245

Verification of Clustering Accuracy by Applying Direction-Based Method and Data Conversion

Young Rhee

Department of Industrial Engineering, Keimyung University, Daegu, South Korea
Email: yrhee@gw.kmu.ac.kr

How to cite this paper: Rhee, Y. (2021). Verification of Clustering Accuracy by Applying Direction-Based Method and Data Conversion. *Modern Economy, 12,* 1593-1611.
https://doi.org/10.4236/me.2021.1211080

Received: October 1, 2021
Accepted: November 5, 2021
Published: November 8, 2021

Abstract

In this study, the process of clustering into a group with similar characteristics is shown through pattern analysis, and the similarity is examined whether this group has homogeneity. A direction-based method for clustering the time series data is introduced, and the grouping process is carried out through the direction setting by up or down and the logical operations thereafter. The similarity is verified by comparing the parts within the group after clustering. For more effective verification, data conversion, a data homogenization process, is performed. MAD, MSE, MPSE and TS are reviewed as similarity indicators. For data such as time series data, MPSE and TS, which are scale-independent measurements, are recommended.

Keywords

Clustering, Direction Based Method, Data Conversion, Similarity

1. Introduction

Demand forecasting is one of the most important management sciences methodology applied in various areas of business. In fact, the selection of the demand forecasting methodology is determined according to the data type, the purpose of forecasting, the forecast period and data characteristics. Selecting the model that provides the best fit to historical data generally does not guaranteed a forecasting method that produces the best forecast of new data. The criteria that a desirable forecast model should have are presented as follows. The first criterion is to analyze the pattern exhibited by the data. Next is how much historical data is available. The last is the length of the forecast period, the time to the future for which the forecast should be made.

In a company with supplying hundreds of thousands of service parts, the demand forecasting is performed by considering it as time series data leading to the flow of the corresponding order. In this process, since predictive expeditiousness is as important as its accuracy, one way to solve this problem is to group time series data with certain rules. The basic step of this grouping methodology can be said to be pattern analysis, which mainly treats data as time related data (Fukunaga, 1990). It starts by identifying the characteristics of the data group to be classified, and by making objects belonging to the same group have similarities. And even for data whose grouping standard is ambiguous, data can be grouped together by adjusting its accuracy. Although different results are obtained depending on the classification criteria, the number of groups to be classified for the entire data cannot be known. And the number of similar groups may be limited by adjusting its predicting accuracy.

This study is motivated by the process of developing a demand forecasting engine for a finished product manufacturing company that requires tens of thousands of parts, and started with the idea of making dozens of groups of time series data of parts based on the certain criteria. The method of grouping quantitative data is related to the number of attributes the data have. Optimization of multiple attributes, not grouping, is mathematically solved as a multiple criterion problem, and an alternative is proposed and a method of mixing them is applied. A method for grouping data with multiple attributes is the MTS (Mahalanobis Taguchi System). The MTS method utilizes MD, which is the distance between an object and a special space of a data group, for grouping (Taguchi & Jugulum, 2002).

In this study, as a methodology for analyzing patterns with time-dependent data, a direction based method is introduced by using a multidimensional scale method. The direction based method is divided into the method applying up and down by the fixed level and the method applying up and down by the present level. The grouping or clustering process using this will be described in detail. Clustering is performed to help efficiently or quickly forecast service parts demand, and the accuracy of similar groups can be improved by applying the multidimensional factor method evaluated by pattern analysis. It is necessary to measure the similarity indicating that the data within the group are homogeneous after analyzing the pattern. However, it is not easy to analyze the accuracy even for a time-dependent data group with the similar pattern, since the significant differences between each group are observed. In the case where there exists a difference between groups, a method of simply correcting the data may also be considered. The similar pattern groups by applying multidimensional factor analysis applied can be a problem due to differences in means between groups when using the original data. Therefore, it is reasonable to verify the similarity between groups by data conversion that corrects the original data. The basic conversion method to measure the similarity between compared group is to match the mean between groups, that is, a method of applying a first-order moment. Another approach is to simultaneously match the mean and variance between groups, and it can be called data conversion using the first order moment and the second order moment.

Pattern analysis of time series data has been studied in various ways, but the direction-based method introduced in this study is a new approach to pattern analysis on time series data stream. To the best of my knowledge, the direction-based method model introduced has not yet been studied in the public literature. The paper is organized as follows. In Section 2, related studies are reviewed. A direction-based method for pattern analysis and data conversion methods are introduced in Section 3. Section 4 shows the procedure for testing the similarity of parts within a group. Finally, Section 5 gives concluding remarks.

2. Related Studies

Data classification is defined as the process of grouping data into categories based on the characteristics of the data so that the data can be used more efficiently. The criterion for classifying numerous data into several groups is related to the purpose of use, but provides clarity of data and speed of data analysis. In this section, applied statistics technologies for classifying data and clustering are introduced.

2.1. Pattern Analysis

Pattern analysis is called a process of mapping from pattern space to class membership space. That is, important features are extracted from data obtained from the outside, and accordingly, they are classified according to the standard pattern sample closest to the features and included in the representative group. The most important issues in accessing data related to time and space are how to recognize it, how to store time related information of patterns, and real-time processing of moving spatial patterns.

There are two types of pattern analysis: univariate analysis and multivariate analysis. Univariate analysis is an analysis method with one response variable and provides a basic framework for multivariate analysis. Multivariate analysis generally refers to a statistical method that analyzes two or more variables simultaneously. The basic criterion for dividing the analysis into univariate analysis or multivariate analysis depends on the number of response variables being considered for the analyzing objective, that is, the number of variables with stochastic responses. In addition, multivariate analysis does not change the basic framework of univariate analysis, such as paired-sample t-test, regression analysis, and multivariate ANOVA. And multivariate analysis can be divided into an area where the dimension is extended simply by increasing the number of response variables and an area where the properties of multivariate data cannot be considered in the category of univariate analysis. Analysis techniques belonging to the unique domain of multivariate analysis include factor analysis, discriminant analysis, cluster analysis, canonical correlation analysis, multidimensional scaling, and structural equation model. The field of grouping applying pattern analysis deals with the problem of estimating a density function in a multidimensional space and dividing the space into categories or class domains.

Generally, data grouping based on pattern analysis is used when a company tries to divide big market into several small markets or to analyze the economic structure within the market. Variables according to criteria are combined. In such a situation, it is possible to clarify the differences between groups by identifying the characteristics of the data group to be classified and by making objects belonging to the same group have similarities. However, according to the classification criteria, the number of groups to be classified for the entire data cannot be known in advance, and the absolute number of groups cannot be checked. The purpose of pattern analysis is to group data using several techniques. Groups are clustered by similarity criteria, and differences between groups can be made clear. Also, in the case of ambiguous data, it can be categorized and integrated into one.

In this study, a methodology for analyzing patterns of time series data is introduced, and grouping and clustering processes using this are also introduced. After grouping by similarity, in order to quickly forecast the time-dependent data such as service parts demand, it is to analyze the accuracy of a group composed of the same data pattern by factor analysis before clustering.

2.2. Factor Analysis

As one of the quantitative analysis methods, factor analysis is a statistical technique used to identify relationships or patterns between various variables and to condense or summarize the information possessed by variables into a small number of latent structures. Through factor analysis, the interrelationships of numerous variables can be analyzed, and the dimensions (potential factors) commonly measured by each variable can be identified and explained based on these relationships (Mirkin, 1987).

Factor analysis methods include a method of extracting inherent factors in consideration of the correlation between variables, a method of grouping similar variables together, and identifying the influence of the extracted factors.

Factor analysis is classified into two types according to the purpose of analysis. First, EFA (exploratory factor analysis) is an attempt to estimate factors in the absence of an existing factor model. Naturally, the factor model created through EFA cannot be presented convincingly to others because it has not been tested. Therefore, the researcher must go through CFA (confirmatory factor analysis) to check whether the model is really suitable and whether there are any areas to be improved or refined in the factor structure. Factor analysis is one of the analysis methods that follows the principle of parsimony. Since the factor extracted through factor analysis is the concept that best represents the characteristics of countless observations, it makes it possible to explain the phenomenon simply and clearly.

The basic assumptions of factor analysis are causality, linearity, and multivariate normality. Linearity means that the common factor and the dependent variable have a linear relationship. That is, it is assumed that there is no abrupt change in the magnitude of causality. Multivariate normality is that all dependent variables follow a multivariate normal distribution when determining goodness of fit.

2.3. MDS (Multidimensional Scaling)

Multidimensional analysis is a statistical technique for visualizing proximity between objects. Similar to cluster analysis, it is an analysis method in which variables are examined for objects and then similarity between entities is measured, and finally the entities are displayed as in a two-dimensional space (France & Carroll, 2011).

In the marketing areas. the MDS method is initially applied in product positioning and product design by expressing the relationship between objects on a single awareness map through property and similarity. Awareness requires that the measure of attribute values for a specific object should be greater than or equal to sequence data or interim data. Specially, the MDS method can be applied to various fields such as opportunity capture by identifying the characteristics of products that are important to consumers, market segmentation strategy by analyzing the position of the given products and competitor products, and advertising effectiveness measurement by developing the products preferred by consumers.

The Euclidean distance is mainly used as a method to measure the distance between entities in multidimensional scaling method. On the other hand, Mahalanobis distance is applied as an effective way to measure the distance between groups since this distance indicates the degree of dispersion of the entities considering the correlation between each attribute. The MDS method uses the fitness as a stress value to increase the accuracy of the relative distance between observation entities, and the S-stress as a non-fitness criterion for expressing each entity in space. MDS method is classified into quantitative MDS method and non-quantitative MDS method according to data characteristics. In the case of numerical data such as intervals or ratios, the former is used, and in the case of an ordinal scale, the latter is used.

2.4. Clustering

The result of pattern analysis is clustering into a representative group consisting of entities with similar characteristics. Clustering is an algorithm that groups entities with similar characteristics without prior information about the entity. In other words, by measuring the similarity between objects, the group is classified to form a cluster among the objects with the highest similarity. According to the degree of completion of clustering, the similarity between individuals within a group is maximized and the similarity between groups is minimized. For basic similarity measurement, Manhattan Distance, Euclidean Distance, Minkowski Distance, Mahalanobis Distance, etc. are applied in the clustering (Bindra & Mishra, 2017; Rhee, 2018).

The clustering method can be divided into a partitioning method, a hierarchical method, a density-based method, a grid-based method, and a model-based method. The partitioning method is an algorithm that classifies an entity into *k* partitioned groups representing clusters. After determining the number of clusters *k*, at least one entity must be included in each partitioned group, and each entity must belong to only one partitioned group. After that, the location of the cluster is redistributed by moving the entities from one group to another in order to divide the entities again. In this case, the distance between the objects is used as a criterion for moving an object, and the Euclidean distance is usually used. For the partition clustering, k-means and k-medoids are commonly used.

The hierarchical method is a method of classifying a given entity hierarchically and clustering it. This method is expressed in a tree structure and is divided into a bottom-up method and a top-down method. The bottom-up method is a method in which an entity is initially regarded as a kind of cluster, and the entities with the greatest similarity are grouped together to finally form a single cluster. Contrary to the bottom-up method, the top-down method regards all entities as single cluster and is continuously divided into smaller groups until all entities are distributed or the termination condition is met. This method has the advantage of not having to determine the number of clusters in advance like the partitioning method. However, this method has the disadvantage of having to decide at each stage whether to merge or split the cluster and at which stage to stop the cluster.

The density-based method is a method of forming clusters based on density. Density-based clustering is known as an efficient clustering method for objects with multidimensional and spatial properties. The DBSCAN algorithm is a commonly used algorithm for density-based methods. DBSCAN determines that the area is dense if there are more than the predefined minimum number of entities within a certain radius of the area where the specific entity is located. In this way, only entities located in the evaluated dense area form a cluster. Therefore, the noise can be effectively dealt with, but the results vary depending on the size of the radius and the minimum number of objects determined. The OPTICS (Ordering Points to Identify the Clustering Structure) algorithm has been proposed to overcome this problem.

The grid-based method has a structure in which a set of entities is divided into a finite number of cells, and all clustering operations are performed on this grid structure. The performance efficiency of this method is generally independent of the number of entities and dependent on the number of divided cells. Therefore, there is a problem in that the time efficiency decreases as the dimension increases and the number of cells increases. Well-known algorithms for the grid-based clustering include Statistical Information Grid and WaveClustering.

The model-based method is a method such that an optimal combination between a model and a given entity is ensured by assuming some mathematical models for each of the clustered groups. Regression models, neural networks, Markov, Hidden Markov Model, etc. are known to apply the model-based method. However, it is not suitable for describing the characteristics of time series data having continuous values, and it is difficult to express the data depicted by a large number of time series characteristics. To solve this problem, the Hidden Markov model is proposed

2.5. Similarity

Time series data consisting of continuous one-dimensional real numbers have been studied in various database applications such as data mining and data warehousing. However, in the recent complex business environment, a multidimensional data sequence is increasing in importance as much as one-dimensional time series data. As an example of a multidimensional data sequence, a video stream can be represented in a multidimensional space with properties such as color and texture. A multidimensional data sequence is divided into several segments, and each segment is represented by various properties. A similarity measure is defined for these segments. Using this measure, segments irrelevant to a given query sequence are primarily excluded from the search. Both the data sequence and the quality sequence are divided into segments, and query processing is performed based on comparing the characteristics of the data segment and the query segment without searching all data in the entire sequence.

The final step in demand forecasting is measuring the accuracy of the forecast, also called the similarity. The accuracy of the forecasting value is determined by the error, the difference between the actual value and the forecasted value, and of course, the smaller the error, the higher the accuracy. Methods for evaluating these errors are largely divided into the following five methods: MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error), MSE (Mean Squared Error), and TS (Tracking Signal).

$$\begin{aligned} \text{MSE} &= \frac{1}{n}\sum_{t=1}^{n}\left(A_t - F_t\right)^2 \\ \text{MAPE} &= \frac{1}{n}\sum_{t=1}^{n}\frac{\left|A_t - F_t\right|}{A_t} \\ \text{MAD} &= \frac{1}{n}\sum_{t=1}^{n}\left|A_t - F_t\right| \\ \text{TS} &= \sum_{t=1}^{n}\frac{A_t - F_t}{\text{MAD}} \end{aligned} \tag{1}$$

MAPE is mainly used for scales related to ratios, and it is difficult to understand underestimates or overestimates with respect to actual values. MSE and MAD have the same scale as actual data, so understanding is intuitive but sensitive to outliers. In TS, the denominator is MAD, that is, the average of the prediction deviation, and the numerator is the sum of errors, which is often applied in control charts. In the case of a continuous upward or downward trend, the TS index may be a bad index, and the index within ±4 is a good indicator. Therefore, MSE and MAD are scale-dependent indicators, and MAPE and TS are scale-independent indicators.

3. Clustering and Conversion

In this chapter, the direction-based method is introduced as one of the multidimensional factor analysis for analyzing the pattern. The results of pattern analysis are ultimately used for clustering. And a data conversion method for similarity verification is introduced.

3.1. Direction Based Method

The choice of clustering method depends on how similar groups can be formed. There are four approaches to clustering that group similar objects, and the advantages and disadvantages of each method are described in Section 2.3. Most of them are approaches to clustering objects of quantitative data. However, although time series data consists of quantitative data, it is reasonable to regard the data stream as a pattern rather than a quantitative data. Therefore, in this case, it seems appropriate to apply the pattern analysis method to analyze the data stream.

In this study, a kind of direction based method is proposed as a method for grouping time series data. To the best of my knowledge, these kinds of method have not as yet been studied in the open literature. The multidimensional factor analysis applied to time series data analysis starts with simply setting criteria using the mean and deviation. As a method similar to multi-factor analysis, a new method based on logical operation includes a fixed-level direction method and a moving-level direction method depends on the comparing criterion. In other words, as the basis for logical operation, the fixed level direction method is used the fixed value as the criterion, and the moving-level direction method is used the present value as the criterion. This is an analysis method that uses this criterion to determine the similarity of data streams by judging more or less compared to the currently given time series data. Each of these judgments becomes a value that is the basis for the logical operations. Therefore, the multidimensional factor analysis method and the fixed-level direction-based method can be said to be fixed-level direction-based methods, except that the reference point for comparison is different.

On the other hand, in the moving-level direction method, various methods can be applied by different comparison criteria. In simple terms, the comparison criterion can be set to the current data, and a value that is used as the basis for the logical operation is generated. The comparison criteria may differ depending on the time window like the moving average method in the forecasting method. In this case, this method is more often referred to as the moving average level direction method. The larger the time window, the more conservative the approach that is not influenced by the latest value. If the time window is very long, that is, if all historical time series data are considered, it is similar to the fixed-level direction method.

The direction-based method for grouping the time series data uses a characteristic used to quantify the qualitative data that is often applied in factor analysis.

The direction-based method for grouping the time series data is characterized by quantifying the qualitative data, which is often applied in factor analysis. This study is to identify the similarity of data patterns by examining the up or down patterns of data as an initial step for grouping time series data. The process of the direction-based method is divided into a stage of generating logical operation data and a stage of verifying data grouping.

Multidimensional factor analysis is used as the comparison criterion for the sum of the mean and deviation of the data for the relevant period. If the data to be compared is lower than this criterion, 0 is assigned, otherwise 1 is assigned. The fixed-level method uses the mean for the given period as a criterion, and if it is higher than the mean, 1 is specified, otherwise 0 is specified. In the fixed level method, a fixed value rather than the mean may be used as a comparison criterion. In this case, applying a value that deviates a lot from the mean may cause a drop of discriminability in the final logical operation. On the other hand, the moving level method is used based on the present value, not the mean. If the compared data value is larger than the given criterion, 1 is assigned, otherwise 0 is assigned. In this study, this process of converting the data stream to 0 or 1 is referred to as the up or down conversion. The algorithm for logical operation to find a clustering is as follows.

Initialization step: the preparation

- Designate representative part for grouping as a basis.
- Daily data on demand is summed into weekly data or monthly data.

Step 1: the up or down conversion

- Execute the up or down conversion by comparing all data, including the basis, to the criterion.
- For the criterion, the mean for the fixed level method, and the preceding data for the moving level method.

Step 2: the logical operation

- Execute logical operation between basis and data every month using up or down conversion result.

Step 3: the sum of the logical operation

- Sum the logically operated values of data for a given period.
- Pattern matching, if the sum is greater than the minimum level.

Step 4: Termination

- Go to Step 1 by retrieving the next data
- Until no data left.

In Step 3, pattern match means that the sum of logical operations of the data exceeds the minimum level. Therefore, the larger the minimum level, the smaller the number of pattern matched data. In this study, 9 is used as the level for pattern match, which means that the pattern for 9 out of 12 months is consistent. The direction-based method and logical operation are implemented using actual parts demand in 2018 of Daedong Engineering. The final result of these processes is applied to data grouping by determining the pattern match. The processes of the up or down conversion and the subsequent logical operation by the fixed-level method is shown in Table 1.

In Table 1, the results of the up or down conversion and the logical operation are shown based on the average. The first step, the up or down conversion process, assigns 0 if the monthly demand for certain part is less than the average annual demand for that part and 1 if greater. For example, in serial #11011, since the average is 12.3, 0 is given in January because it is less than the average, and 1 is given in February. The up or down conversion can be completed by continuing this process until the end of year. In the second step, the logical operation is carried out every month with the result of the up or down conversion of the basis and the part to be examined. The logical operations are always performed in comparison with the basis, which is the representative part of the group. For example, in January, the up or down conversion results of basis and serial #11011 are both 0, so the logical operation result of serial #11011 becomes 1 in January. However, in June, the up or down conversion result of basis is 0, and serial #11011 is 1, so the logical operation result of serial #11011 becomes 0 in June. If this process is continued, the logical operation is completed. This process can be applied to the moving level method, where the comparison criterion is not the annual average but the previous month's data, and the results shown in Table 2 can be obtained.

Table 1. Logical operation of the fixed level method.

Original data

Serial#	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Mean
basis	50	177	170	318	173	100	139	79	43	22	86	42	116.5
11011	0	28	20	39	16	13	15	8	0	3	2	3	12.3
32415	3	39	50	70	40	32	58	8	0	10	11	6	27.3
18264	21	45	34	48	19	11	30	25	11	3	19	37	24.3

Up or down conversion

Serial#	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
basis	0	1	1	1	1	0	1	0	0	0	0	0
11011	0	1	1	1	1	1	1	0	0	0	0	0
32415	0	1	1	1	1	1	1	0	0	0	0	0
18264	0	1	1	1	0	0	1	1	0	0	0	1

Logical operations (basis vs data)

Serial#	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Sum
11011	1	1	1	1	1	0	1	1	1	1	1	1	11
32415	1	1	1	1	1	0	0	1	1	1	1	1	10
18264	1	1	1	1	0	1	1	0	1	1	1	0	9

Table 2. Logical operation of the moving level method.

Original data													
Serial#	**Dec**	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
basis	25	50	177	170	318	173	100	139	79	43	22	86	42
11011	0	0	28	20	39	16	13	15	8	0	3	2	3
32415	1	3	39	50	70	40	32	58	8	0	10	11	6
18264	40	21	45	34	48	19	11	30	25	11	3	19	37

Up or down conversion												
Serial#	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
basis	1	1	0	1	0	0	1	0	0	0	1	0
11011	1	1	0	1	0	0	1	0	0	1	0	1
32415	1	1	1	1	0	0	1	0	0	1	1	0
18264	0	1	0	1	0	0	1	0	0	0	1	1

Logical operations (basis vs part)													
Serial#	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**	**Sum**
11011	1	1	1	1	1	1	1	1	1	0	0	0	9
32415	1	1	0	1	1	1	1	1	1	0	1	1	10
18264	0	1	1	1	1	1	1	1	1	1	1	0	10

As shown in Table 2, the data on December before January are given, and the up or down conversion is executed using this as a starting point. Compared with the results of the fixed-level method, the results of Table 2 show some differences in the sum of logical operations. However, the final pattern matching result remained the same.

3.2. Data Conversion

Data conversion is necessary to verify similarity within the grouped data set. Data conversion is originally derived from data transformation while maintaining data characteristics for the purpose of comparison between groups. For example, when it is suspected that different time-dependent data groups have the same characteristics, but there is no proper way to compare them, it is possible to analyze the inherent characteristics through data transformation. The most basic data conversion method is mean conversion), which matches the mean of the entire group. This is called the first-moment data conversion.

A more advanced data transformation method is to match the mean and variance, which are characteristics of a group. It is called the standard normal conversion because it matches the first and second moments of the data. The standard normal conversion is to use the characteristics of the standard normal distribution under the assumption that the data constituting the group form a normal distribution. Assuming that the random variable X has a normal distribution with mean μ and variance σ^2, the random variable Z of the probability distribution with mean 0 and variance 1 is said to follow the standard normal distribution. The relationship between random variables X and Z is $Z = \frac{X - \mu}{\sigma}$.

$$X_1 = \frac{\sigma_1}{\sigma_2}\left(X_2 - \mu_2\right) + \mu_1 \tag{2}$$

The relationship between the actual data X_1 and X_2 using the equivalence relationship between the random variables X and Y of the two groups with standard normal distribution is shown in (2). If the amount of parts A and B required to produce a finished product is in a proportionate form, the standard normal conversion can provide unexpectedly good results. As it were, the data streams of the two groups appear to be completely different types of data groups before the standard normal conversion, but after the conversion, it can be seen that they are not exactly the same, but have similar flows.

Since pattern matching has already been performed in Section 3.2, it is possible to identify the similarity of data streams in the data group only when the comparison criteria are the same. Therefore, the method of matching the mean and standard deviation of the two groups is meaningful. The standard normal conversion is largely divided into two stages. In the stage 1, the mean conversion that makes the basis equal to the mean of other data. In other words, it is a method of correcting the difference between the means of the two groups. In the stage 2, the standard normal conversion is executed by applying (2). In this case, if the converted value becomes a negative number, a work of compensating for the minimum negative value may be performed as an additional step. The method of replacing the data with 0 after the mean conversion, which is a negative number, is also applicable, but in this study, the method of correcting the minimum negative value is applied.

The results of the data conversion process step by step is shown in **Table 3**. The data used for data conversion are the parts turned out to be pattern match in Section 3.1 by the direction-based method. An additional step is performed since a negative number occurred in step 2.

Figure 1 shows the original demand stream and the demand stream after the standard normal conversion for 4 parts. In particular, if the monthly data for each part is compared in the original data stream, the difference in demand is so large that it is doubted whether it is a pattern match group. This shows a completely different demand stream before the standard normal transformation, even if it is classified as a group of similar patterns by applying the direction-based method. However, the data streams for all parts after the standard normal conversion show a similar pattern.

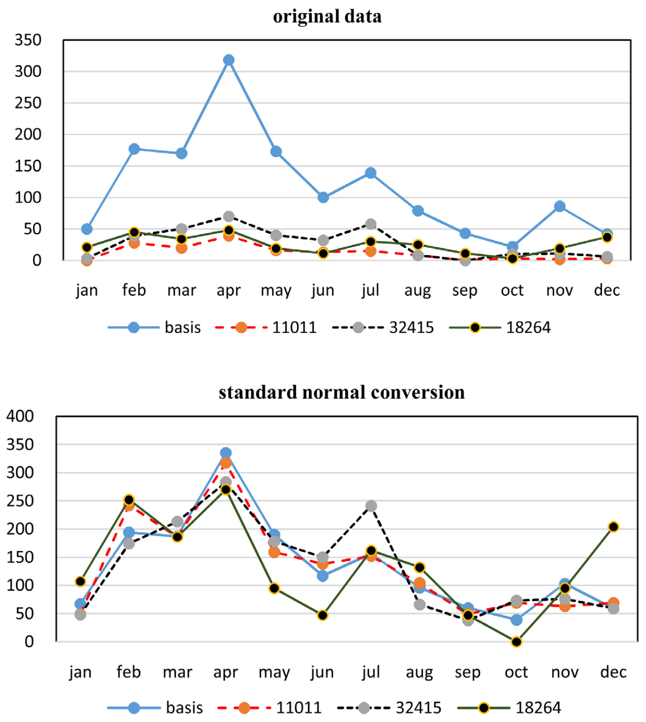

Figure 1. Before and after standard normal conversion.

Table 3. Standard normal conversion.

Step 1													
Serial#	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Mean
basis	50	177	170	318	173	100	139	79	43	22	86	42	117
11011	104	132	124	143	120	117	119	112	104	107	106	107	117
32415	92	128	139	159	129	121	147	97	89	99	100	95	117
18264	112	136	125	139	110	102	121	116	102	94	110	128	117

Step 2													
Serial#	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Mean
basis	50	177	170	318	173	100	139	79	43	22	86	42	117
11011	32	225	169	300	142	121	135	87	32	52	46	52	117
32415	31	157	196	266	161	133	224	49	21	56	59	42	117
18264	90	235	169	253	78	30	145	115	30	-17	78	187	117

Additional step														
Serial#	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Mean	Stdv
basis	67	194	187	335	190	117	156	96	60	39	103	59	134	84
11011	49	242	186	317	159	138	152	104	49	69	63	69	134	84
32415	48	174	213	283	178	150	241	66	38	73	76	59	134	84
18264	107	252	186	270	95	47	162	132	47	0	95	204	134	84

4. Similarity Verification

The clustering process and data conversion process presented in section 3 are shown by giving an actual example. In this section, the accuracy of each part within a group, that is, clustering accuracy, is analyzed by applying the method of analyzing the accuracy of demand forecasting. The similarity of data is determined by error in prediction, and of course, the smaller the error, the higher the accuracy. There are MAD, MAPE, MSE, and TS as methods for measuring accuracy, which are frequently used because it is easy to compare prediction errors between quantitative models and their reliability is high.

In demand forecasting, the deviation between the measured value and the forecast value is important to measure the forecasting accuracy. In order to apply this point to this study, it can be said that it is the same mechanism as the accuracy analysis of demand forecasting if the demand for parts is regarded as the actual value and the basis, the representative part as the actual value The direction-based method presented in this study only judged similar patterns as the same group based on the data stream, and the deviation of time series data of representative parts and comparative parts may be very large. Therefore, analyzing group similarity using real data may be erroneous. It is necessary to at least homogenize the group's data for more accurate comparison. For this reason, it is considered appropriate to conduct group similarity verification through the data conversion.

Table 4. Data conversion for the group with small demand and few parts.

Before conversion												
serial#	**Jan**	**Feb**	**May**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
26602	3.0	13.0	8.0	7.0	4.0	1.0	16.0	14.0	10.0	7.0	18.0	0.0
43101	3.0	4.0	1.0	0.0	5.0	0.0	42.0	32.0	12.0	13.0	46.0	2.0
41822	4.0	10.0	6.0	3.0	0.0	7.0	95.0	90.0	5.0	14.0	72.0	2.0
27511	5.0	3.0	3.0	1.0	0.0	0.0	13.0	11.0	10.0	5.0	37.0	19.0
A0631	2.0	7.0	2.0	0.0	3.0	20.0	35.0	37.0	28.0	17.0	50.0	6.0

After conversion												
serial#	**Jan**	**Feb**	**May**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
26602	3.0	13.0	8.0	7.0	4.0	1.0	16.0	14.0	10.0	7.0	18.0	0.0
43101	4.8	5.1	4.1	3.7	5.5	3.7	18.4	14.9	7.9	8.3	19.8	4.4
41822	4.9	5.8	5.2	4.7	4.2	5.4	19.5	18.7	5.0	6.5	15.8	4.6
27511	6.2	5.1	5.1	4.0	3.4	3.4	10.5	9.5	9.0	6.2	24.0	14.0
A0631	3.0	4.8	3.0	2.3	3.4	9.3	14.6	15.3	12.1	8.3	19.9	4.4

In this study, the similarity is analyzed for two types of groups in which each part is found to be a pattern match by the direction-based method. The first group is a group with relatively small monthly data and a small number of parts, and the second group is a group with large monthly data and a large number of parts. Table 4 shows the data before and after the standard normal conversion of a small group consisting of a small number of parts with small amounts of monthly data. For the standard normal conversion, serial #26602 is used as a basis.

The following is to show the data conversion process for a large amount of monthly demand and a group consisting of many parts. As seen in Table 5, the part in the first row is a part representing this group, and data conversion is performed based on this part. Although the first row is a representative part, the demand for the representative part was also modified in the process of correcting negative numbers in the conversion process.

As seen in Table 5, the large or small number of parts belonging to a group does not affect the data conversion process. However, in the case of large monthly data, most of them have large values even after the data conversion, which affects the scale-dependent indicator such as MSE and MAD.

Table 5. Data conversion for the group with many demand and many parts.

Before conversion												
serial#	Jan	Feb	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec
D100F	1	50	45	280	337	30	80	10	10	0	0	9
6001	8	23	2	156	54	3	35	9	19	8	57	27
72531	0	5	27	300	400	3	40	50	0	1	5	10
43212	7	5	3	119	103	75	97	15	66	31	20	30
43751	16	58	108	168	152	47	52	16	12	4	41	30
25124	5	16	6	67	54	21	37	50	14	15	35	13
85454	13	40	30	73	81	86	55	16	64	6	13	13
13431	6	47	90	102	139	49	52	29	32	22	25	12
43751	2	25	73	100	83	16	37	20	8	6	31	10
12142	7	59	26	111	120	21	30	7	9	0	2	22
43681	40	73	99	161	152	59	186	43	10	71	44	6
33634	12	32	60	85	72	10	83	35	6	31	9	6
30101	18	10	123	102	170	2	13	24	1	8	32	5
72061	7	12	48	65	46	6	31	40	23	8	19	4
50592	20	42	21	112	276	6	30	4	1	4	0	2

After conversion												
serial#	Jan	Feb	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec
D100F	70.3	119.3	114.3	349.3	406.3	99.3	149.3	79.3	79.3	69.3	69.3	78.3
6001	72.5	112.5	56.4	467.4	195.2	59.1	144.5	75.1	101.8	72.5	203.2	123.2
72531	80.3	84.6	103.4	337	422.6	82.9	114.5	123.1	80.3	81.2	84.6	88.9
43212	30.3	24.9	19.5	333.8	290.4	214.6	274.2	52	190.2	95.3	65.5	92.6
43751	51.6	138.9	242.8	367.6	334.3	116	126.4	51.6	43.3	26.6	103.5	80.7
25124	13.3	74.7	18.8	359.4	286.8	102.6	191.9	264.5	63.5	69.1	180.8	57.9
85454	33.2	137.1	98.6	264	294.8	314.1	194.8	44.7	229.4	6.2	33.2	33.2
13431	14.1	130.6	252.7	286.8	391.9	136.2	144.8	79.4	88	59.5	68.1	31.1
43751	28.5	108.2	274.6	368.1	309.2	77	149.8	90.9	49.3	42.4	129	56.2
12142	63.9	208.3	116.7	352.8	377.8	102.8	127.8	63.9	69.4	44.4	50	105.5
43681	65.6	129.3	179.5	299.2	281.8	102.3	347.5	71.4	7.7	125.5	73.4	0
33634	47.9	122.5	227.1	320.4	271.9	40.4	313	133.7	25.5	118.8	36.7	25.5
30101	91.1	75	303.2	260.8	398.1	58.8	81	103.3	56.8	70.9	119.4	64.9
72061	33.7	62.2	266.7	363.3	255.3	28.1	170.1	221.2	124.6	39.4	101.9	16.7
50592	107.1	138.6	108.5	238.9	473.9	87	121.4	84.2	79.9	84.2	78.4	81.3

As shown in Table 6, the similarity indicators MAD, MSE, MPSE and TS are analyzed. In a data stream with a small value like the first group, MAD and MSE show relatively good similarity results, but it cannot be considered as a good indicator for a data stream with a large value like the second group. The reason for this is that similarity errors may occur significantly in data with large values or in scale-dependent data with different comparison units since MAD and MSE are absolute indicators. It is considered that MAD and MSE are not suitable for similarity analysis of time series data. However, since MAPE and TS are scale independent measures, it can be said that they are suitable for similarity analysis of time series data. In Table 6, MPSE and TS indicators show relatively good results regardless of the number of parts constituting the group and the size of the data. The similarity is also analyzed for a part group consisting of pattern-mismatched parts. MAD and MSE are relatively decreased after the data conversion, but MPSE remained large.

Table 6. Similarity survey.

Before conversion				
serial #	**MAD**	**MSE**	**MAPE**	**TA**
43101	8.9	167.4	77.8	−6.62
41822	20.3	1257.8	193.8	−10.22
27511	6.2	77.2	55.5	−0.97
A0631	12.3	238.2	246.9	−8.59

After conversion				
serial #	MAD	MSE	MAPE	TA
43101	2.84	11.35	50.39	−6.62
41822	3.26	14.3	63.55	−10.22
27511	4.3	31.58	53.27	−0.97
A0631	3.29	18.13	93.37	−8.59

Before conversion				
serial #	MAD	MSE	MAPE	TA
6001	54.1	8713.8	118.7	8.34
72531	22.6	897.9	75.6	0.49
43212	56.9	7648.3	160.4	4.94
43751	41.8	4523.2	179.7	3.54
25124	60.3	11096.9	106.9	8.61
85454	55.3	9649.0	188.8	6.54
13431	47.3	6337.9	105.8	5.23
43751	49.6	8456.1	48.6	8.89
12142	41.5	6574.3	88.5	10.55
43681	58.8	6125.3	396.3	−1.56
33634	49.9	9251.8	142.6	8.23
30101	53.5	6183.3	209.8	6.43
72061	58.4	11407.6	122.7	9.3
50592	31.7	3017.0	201.5	10.55

After conversion				
serial #	MAD	MSE	MAPE	TA
6001	54.13	7001.25	39.81	8.34
72531	18.16	476.98	17.46	0.49
43212	65.25	6331.91	57.28	4.94
43751	36.65	2386.96	33.77	3.54
25124	58.8	6495.42	58.92	8.61
85454	71.37	8406.47	67.82	6.54
13431	32.59	2528.38	29.95	5.23
43751	41.84	3636.61	38.89	8.89
12142	20.96	945.33	20.18	10.55
43681	56.14	6345.4	45.01	−1.56

Continued

33634	63.96	6213.58	55.53	8.23
30101	47.59	4670.34	41.37	6.43
72061	67.87	7022.7	63.77	9.3
50592	26.04	1648.35	16.53	10.55

Special attention should be paid to the TS indicator, where a positive value indicates that demand is higher than expected. Negative numbers indicate that demand is lower than forecast. In general, it is known that if the TS index is within ±4, it is a very good similarity index. The results in Table 6 are shown to be good TS indicators regardless of before and after data conversion, which reflects that the clustering by applying the direction-based method is effective.

5. Conclusion

In this study, the process of similarity grouping by applying pattern analysis by factor analysis is shown, and the similarity is examined whether this group has homogeneity. A direction-based method for clustering the time series data is introduced, and the grouping process is carried out through the direction setting by up or down and the logical operations thereafter. This kind of method has not been studied as a grouping method for time series data.

After clustering, similarity is verified. For more accurate comparison, it is appropriate to verify similarity through data conversion, the process of data homogenization. For the similarity verification, MAD, MSE, MAPE, and TS, which are used to analyze the accuracy of demand forecasting, are applied. The similarity errors may occur significantly in data with large values or scale-dependent data with different comparison units since MAD and MSE are absolute indicators. On the other hand, MAPE and TS are scale-independent and relative indicators, so it can be said that these indicators are suitable for analyzing the similarity of time series data.

Even if classified as a pattern matching group, similarity may be inaccurate because amplification by data conversion may occur in a group in which parts with relatively little demand data and parts with large order data are mixed. In order to prevent this, the more effective results are expected if the data is pre-classified by the annual average of the time series data of all parts before grouping and then grouped. An extension of this study is to analyze clustering accuracy verification through comparison between the pattern-matching group and the non-pattern-matching group, and also to study a more accurate pattern analysis method before data conversion.

Acknowledgements

The present research has been conducted by the Sabbatical Research of Keimyung University in 2016.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

Bindra, K., & Mishra, A. (2017). A Detailed Study of Clustering Algorithms. *IEEE Xplore 6th International Conference on Reliability,* 370-376. https://doi.org/10.1109/ICRITO.2017.8342454

France, S. L., & Carroll, J. D. (2011). Two-Way Multidimensional Scaling: A Review. *IEEE Transactions on Systems, Man, and Cybernetics, 41,* 644-661. https://doi.org/10.1109/TSMCC.2010.2078502

Fukunaga, K. (1990). Introduction to Statistical Pattern Recognition (Computer Science and Scientific Computing Series). *Material,* 1-23. https://doi.org/10.1016/B978-0-08-047865-4.50007-7

Mirkin, B. G. (1987). Additive Clustering and Qualitative Factor Analysis Methods for Similarity Matrices. *Journal of Classification, 4,* 7-31. https://doi.org/10.1007/BF01890073

Rhee, Y. (2018). A Study on the Balanced Assignment of Allocating Large Group with Multiple Attributes into Subgroups. *American Journal of Industrial and Business Management, 8,* 1418-1432. https://doi.org/10.4236/ajibm.2018.86095

Taguchi, G., & Jugulum, R. (2002). *The Mahalanobis-Taguchi Strategy. A Pattern Technology System*. John Wiley and Sons. https://doi.org/10.1002/9780470172247