INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, 12(6), 790-802, December 2017.

Time Series Clustering Based on Singularity

D. Chang, Y.F. Ma, X.L. Ding

Dan Chang, Yunfang Ma*, Xueling Ding
School of Economic and Management,
Beijing Jiaotong University,
Beijing, 100044, China
dchang@bjtu.edu.cn
*Corresponding author: yunfangm@bjtu.edu.cn
dingxueli@boe.com.cn

Abstract: With relevant theories on time series clustering, the thesis makes research into similarity clustering process of time series from the perspective of singularity and proposes the time series clustering based on singularity applying K-means and DBScan clustering algorithms according to the shortage of traditional clustering algorithm. In accordance with the general clustering process of time series, time series clustering based on singularity and K-means are made respectively to get different clustering results and make a comparison, thus proving that similarity clustering research of time series from the perspective of singularity can better find out people's concern on time series.

Keywords: time series, clustering, singularity, DBScan, Kmeans.

1 Introduction

Time series is a high dimensional data type, a random variable which is strictly arranged by time order and correlated with each other. As an important data type in the economic field, it plays a significant role in people's analysis on market trend and their decision-making. As some time series data increase at the magnitude of several billion per day or even per minute, how to find the data correlation in time series and analyze such huge data with timely and fast response, thus figuring out similar or regular changing pattern, tendency, mutation with obvious change and distribution of outliers, has become an increasingly important and challenging hot topic.

During the last decades, a societal focus on the work of university faculty as a measure of return on the public's investment in higher education stimulated a reevaluation of how faculty performance ought to be measured and assessed. The development of workable assessment systems is difficult largely due to the fact that the value of assessment is often controversial:Clustering analysis is one of the important tasks in data mining. It is a process in which data set is divided into several groups or classes, making the data objects in the same group or class have high similarity and those in different groups or classes different. Differing from traditional clustering analysis, data processed in the clustering analysis of time series are changing over time with features of high-dimension, complexity, dynamism and high noise, which are easy to reach large-scale. Variable clustering approach usually is used to handle the high-dimensional variable [12]. Time series clustering refers to cluster time series with similar change into one class, and time series in different classes have obvious different changes [6]. How to cluster the mass time series which is closely related to daily life has aroused concern among many scholars in the field of data mining. Most of current classic similarity analysis is based on mathematical model, which considering the overall features of time series. However, for the mass data and the special time series with high-dimension, high noise and high complexity, traditional similarity searching based on mathematical model does not consider that user requirement and concern vary in different situations in real life, which means the importance of singularity in time series differs for different users, in spite of the shortening processing time caused by relative technology and more accurate results.

There are many definitions for singularity without a formal and standard one that is generally accepted. The definition in the thesis is the data that is obviously diverged from most sample data, namely data points different from most sample data which do not meet the general model of sample data. That is to say the data points taking up very small data volume in the clustering result.

The thesis aims to analyze through time series data mining algorithm with theories related to clustering such as DBScan and Kmeans, figuring out the existing problems in traditional time series data mining clustering algorithm, confirming the factors which should be comprehensively analyzed during the time series clustering analysis such as the influence of major events or accidental events on time series, and finishing the time series clustering research based on singularity. The case study at the end of the thesis demonstrates the application of time series clustering proposed by the thesis in detailed container port transportation.

2 Literature review

2.1 Time series

Basic Concept of Time Series

Time series is a series composed of data that change over time, which is also called dynamic series. Time series metrics or features that can be used for time series classification or regression analysis [7]. Time series are used in statistics, pattern recognition, econometrics, mathematical finance, weather forecasting, intelligent transport and trajectory forecasting [14], container shipping freight rate forecasting [8], etc. Different from static data, time series is a data object of high complexity and high noise that describes the changing process of things over time. Time series exists widely in various fields of daily life. For example, the grain yield of a certain place changes every year; stock price keeps fluctuating over time; the traffic flow of a certain road changes in different time periods.

Data of time series change over time. According to this feature, every data unit of time series is abstracted to a binary array (t, x) composed of time and corresponding time value, in which t represents time variable and x represents data variable.

Similarity of time series

Suppose that there are time series Q and time series C, $Q = \{q_1, q_2, ...q_n\}$, $C = \{c_1, c_2, ...c_n\}$, if the distance between Q and C satisfies $\text{dist}(Q, C) \leq \epsilon$ (given similar threshold value which is used to adjust the level of similarity), Q and C is similar. Besides, dist(Q,C) is a distance function, and the most typical one is Euclidean distance. However, Euclidean distance is not sensitive to noise data and can hardly recognize time series with displacement or stretch.

Similarity based on Euclidean distance regards time series as points in multi-dimensional space and measure the similarity of time series with the distance between points. Because of the advantages of fast calculation and low complexity, Euclidean distance becomes the most commonly used measurement method. For example, for time series Q and time series C of the same length, $Q = \{q_1, q_2, ...q_n\}$, $C = \{c_1, c_2, ...c_n\}$, then $Q, C \in R_n$; Q and C are regarded as two points q_i and c_i in N-dimensional space, and the distance between them is:

$$d = \sqrt{\sum_{k=1}^{n} (q_2 - c_k)^2}$$

2.2 Time series clustering algorithm

Data in real life have features such as huge amount, high-dimension and high noise etc., especially for time series data. And data are expanding and becoming more complicated [1]. So far there is no clustering algorithm that can be applied to data of any type. Many scholars have built many clustering algorithms in order to solve the clustering for multi-type data.

K-means algorithm

K-means has been successfully used in various areas, such as market segmentation, computer vision, geostatistics, astronomy and agriculture [4]. K-means by default considers that cluster comprises all objects within close distance and the ultimate goal is to obtain a compact and independent cluster. In a complete iteration, K-means will calculate the distance between the cluster center and every remaining object and distribute the object to the nearest cluster. Iteration is finished after all data objects are detected, and then new clustering center is figured out.

Description of detailed algorithm is as follows:

Given data set $X = \{x_1, x_2, x_3, ...x_n\}$, in which data object x_i has m-dimension variable, x_i=$\{x_{i1}, x_{i2}, x_{i3}, ...x_{im}\}$. K represents the number of needed cluster; data set X is divided into K classes; according to the requirement of high similarity inside the cluster and low similarity between clusters, square-sum-of-error criteria function E should be minimum in order to obtain the optimal clustering effect. E represents the sum of distance between each object and its located cluster center [5].

It can be seen from the experimental study of square-sum-of-error function E that when data objects in a data set are intensive and the differentiation between types is obvious, the criteria function of square sum of error is more effective, thus enabling K-means to get a better clustering effect. The algorithm has converged when the assignments no longer change. There is no guarantee that the optimum is found using this algorithm [3].

DBScan algorithm

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a spatial clustering algorithm based on density [2]. It requires two parameters: distance (radius of the detection region) and the minimum sample points (minPts). DBSCAN can be used with any distance function [2] [10] (as well as similarity functions or other predicates) [9].

Description of detailed algorithm [10]:

(1)Search for object p that is not detected in the data set. If p is not processed, then the number of object in the field is no less than MinPts. Build new cluster C, and add all objects in C to candidate set N;

(2)Detect object q that is not processed in candidate set N. If the number of object in its neighborhood is greater than or equal to MinPts, then add these objects to N; if q does not belong to any cluster, add q to C;

(3)Repeat 2); keep Detecting objects remaining unprocessed in N until N is empty. Repeat 1) 3) until all objects are processed. Clustering process is demonstrated in (Fig. 1).

DBScan defines cluster as the biggest set of points that are density-reachable, which means dividing areas of high density into clusters, so it can recognize the noise point, which is different from partition clustering and hierarchical clustering. Compared with K-means clustering, it does not need to know K in advance, and can find out cluster of any shape.

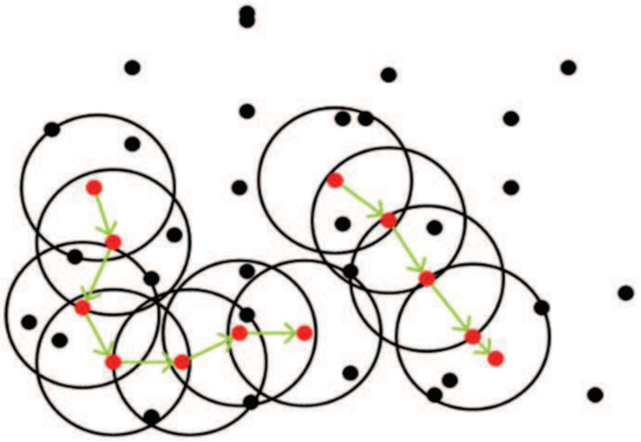

Figure 1: DBScan clustering process diagram

Advantages and disadvantages of the algorithm

As a mountain-climbing search algorithm, K-means algorithm has certain disadvantages: (1) K should be provided externally and the accuracy of K is closely related to the clustering result, but it is not easy to confirm K, which forms one of the disadvantages of K-means. (2) The clustering result is related to the initial center. If the choice of initial center fails, then it is impossible to get ideal clustering result. (3) K-means should keep iterating on classification adjustment and calculating new cluster until meeting the condition. Therefore, the time cost of K-means algorithm is huge for mass data. In spite of many disadvantages, K-means algorithm become one of the most commonly used algorithm in clustering research because of the features of simplicity, high intelligibility, high convergence rate and high scalability. Obviously there are many clustering algorithms. According to the brief introduction and analysis of the two algorithms above we can obtain the comparison result of the two clustering algorithms as showed in (Tab. 1).

Table 1: Comparison of the DBScan and K-means algorithms

Feature / Algorithm	K-means	DBScan
Data type	Symbolic type and numeric type	Numeric type
High dimension	Normal	Normal
Sensitivity to data order	Normal	Sensitive
Sensitivity to noise	Sensitive	Insensitive
Scalability	Good	Good
Efficiency	Normal	Normal
Algorithm implementation feature	Mathematical description; easy to understand and implement; repeated scan for the optimal result; but sensitive to the initial condition; number of clusters confirmed by man; all data put in internal storage.	One-time scan; able to recognize noise; clustering result independent of parameter; not very ideal clustering effect for data object with uneven density; not high degree of support for high-dimension data.

The above two algorithms for similarity measurement figure out the series which is similar to given series through various relative technology at certain target efficiency. Target efficiency means that the implement of similarity analysis algorithm is of high efficiency and low complexity. Most of current classic similarity analysis is based on mathematical model, which considers the overall features of time series. However, for the mass data and the special time series with high-dimension, high noise and high complexity, traditional similarity searching based on mathematical model does not consider that user requirement and concern vary in different situations in real life, in spite of the shortening processing time caused by relative technology and more accurate results. For example, people pay more attention to the place, time and probability of earthquake in the earthquake prediction; investors put special emphasis on the fluctuation of stock price in the stock market forecast. Therefore, it is essential to introduce the influence of key events such as major events and accidental events on time series during the time series analysis.

3 Clustering algorithm based on singularity

In light of the shortage of traditional algorithm that it does not consider that user requirement and concern vary in different situations in real life, the thesis made further optimization of the traditional algorithm by taking different concerns, namely singularity, of users in real life into consideration.

3.1 Basic thought of singularity clustering

Based on the distance-based K-means algorithm, cluster total sample data N to get K class; then with the density-based DBScan clustering algorithm, successively get the data volume m_i in every class of K_i, and then figure out $\rho_i = m_i/N$, given threshold value ϵ (defined according to user requirement):

When $\rho_i \geq \epsilon$,i type is density-reachable at ϵ;

When $\rho\_i <\epsilon$, i type is not density-reachable at ϵ. Then i is regarded as the data class diverged from most data sample, namely singular class, which refers to the data class including singularity.

3.2 Clustering based on singularity

Current time series clustering is made with every series as a whole. For example, if we cluster n time series with current time series clustering algorithm, n time series are regarded as data objects while clustering. What we consider is the whole similarity of n time series, but ignoring the possible similarity of data objects at certain time periods. Therefore, the thesis takes the data of n time series at the whole time period $(t_i - t_j, i \in N, j \in N)$ as data object, which means making similarity clustering with $n * (j - i)$ data objects, thus not only considering the similarity showed by different ports at the same year, but also comprehensively considering the similarity demonstrated by different ports at different years and that by the same port at different years.

Description of detailed process:

(1) Data labeling. Label data after preprocessing. Data of n at t is expressed as (t, n).

(2) DBScan clustering. Choose radius r as density in DBScan algorithm and get K class results.

(3) Decide discrete class, namely singular class. Figure out data volume m_i in every K_i successively and then figure out $\rho_i = m_i/N$; given threshold value ϵ. When $\rho_i \geq \epsilon$, i class is

density reachable at ϵ; when $\rho_i<\epsilon$, i class is density unreachable, then it is called the data class that deviates from most data sample, namely singular class.

(4) Combine with expert advice and accidental events to get reasonable clustering result.

3.3 Singularity detection

The important feature of singularity in time series carries important information. For example, according to the data from Census and Statistics Department of Hong Kong, several scholars summarized some irregular events through text mining technology based on webpage, and then integrated these events through Delphi Technique combining expert advice, figured out the influence direction of these events on the throughput respectively, and searched on International Maritime Information Website for the events that influenced throughput of container in recent years, thus drawing the influence direction of irregular events on port throughput as showed in (Tab. 2), which can be used to judge the influence on throughput when irregular events happen. Of course, data in the table should be updated and adjusted according to the actual situation in order to keep flexibility.

Table 2: Effect of irregular events on port throughput

Class	Important Event	Direction
Industry Factor	Container cost reduction	Up
	Container vacancy rate decline	Up
	Barge cost reduction	Up
Economic Factor	Economic crisis	Down
	Foreign exchange rate	Uncertain
	Crude oil price rise/Fall	Down/Up
Related Policy	Pallet carrier cross-border license fee reduction	Up
	Wharf occupation rate decline	Up
Natural Disaster	Typhoon/tsunami/earthquake	Down
Regional Competition	Port merger	Up
	Operating efficiency rise of other ports	Down
Political Factor	Port labor strike	Down
	Direct Flight for cross-strait trade	Down
	Terrorist attack	Down
Other Factor	Wharf construction	Up
	Berth number increase	Up
	Shipping company and wharf company reach related agreement	Uncertain

4 Experimental comparison-taking China's container port clustering as an example

Ports have become a significant strategic resource for the development of national economy and regional economy in China. They act as the nodes connecting water transport with land transport and play pivotal roles in logistics network [13]. As the development of port cities and hinterland economy relies on the development of ports, a series of research on ports is of great significance for not only the development of hinterland cities, but also the development of ports themselves [11].

4.1 Choice of representative ports

The thesis takes the clustering research of China's container ports as an example, during which the choice of container ports data is the basis for research. Coastal ports mainly spread at coastal areas such as Yangtze River delta, Pearl River delta, Shandong Peninsula and west coast of Bohai Sea etc., which are the optimal objects of clustering research. Ports at Yangtze River delta mainly include Shanghai Port and Ningbo-Zhoushan Port, froming a container transportation system consisting of ports such as Lianyungang Port; with an emphasis on Ningbo-Zhoushan Port and Lianyungang Port, imported mineral, handling of crude oil and transfer system of ports of Shanghai, Suzhou, Zhenjiang and Nanjing etc. should be developed correspondingly; Shanghai Port and Ningbo-Zhoushan Port form the major coal transfer system. Therefore, the thesis chooses Shanghai Port, Ningbo-Zhoushan Port and Liangyungang Port as the representatives of ports at Yangtze River delta.

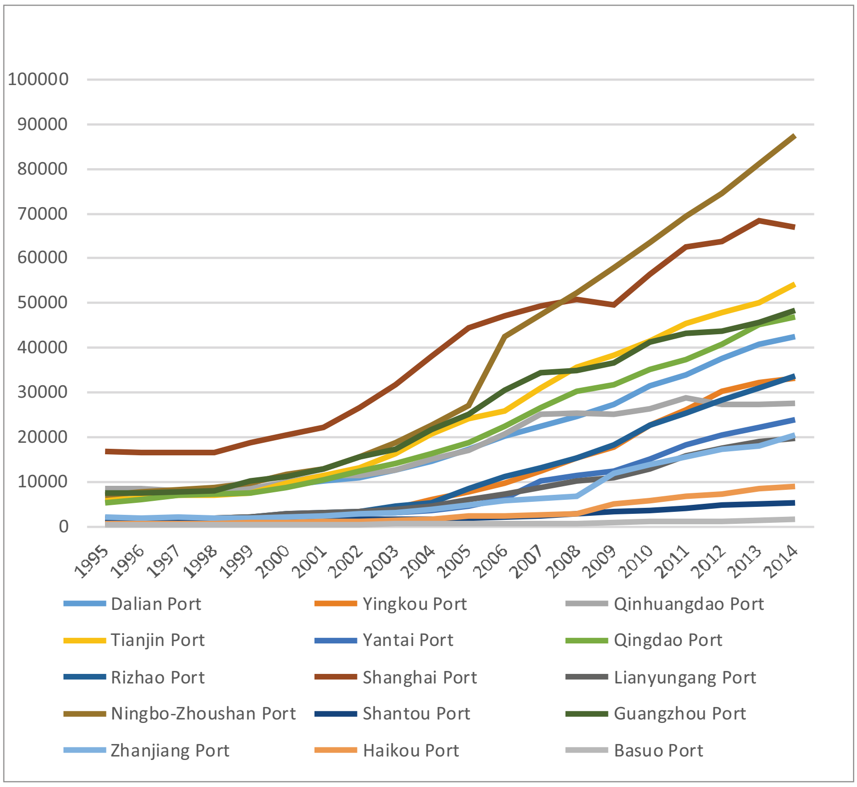

Figure 2: Comparison of the DBScan and K-means algorithms 1995- 2014 throughput of coastal container port above designated size in China

First of all, we choose throughput of 15 coastal container ports above designated size in China from 1995- 2014 as the time series data to analyze different variation trend demonstrated by different ports. Without any process on the original data, it can be seen from (Fig. 2) that although all ports largely demonstrate an upward trend as time goes by, some ports show different changing trend at certain stages. For example, from 2008 to 2009, Shanghai Port and Qinhuangdao Port showed downward trend while other ports went upward. Therefore, we assume that whether there was a big event in 2008 which had great influence on only Shanghai Port and Qinhuangdao Port while having no influence or little influence on other ports. In order to better describe various features of different time series at different stages, indicators representing tendency of every stage are clustered as attributes. Therefore, the thesis takes throughput growth rate as the attribute of clustering analysis calculation.

4.2 Ports clustering based on singularity

The time series clustering algorithm that can recognize singularity in 3.3 is adopted in this part in order to figure out the influence of accidental events on every port in port market.

Current time series clustering is all made with each series as a whole. For example, if we use current time series algorithm to cluster ports, it is made with 15 time series as data objects. Therefore, what is considered is the overall similarity of 15 time series, but ignoring the possible similarity of data objects at certain stage. The thesis takes the throughput data of every port from 1995 to 2014 as data objects. Taking the port data in each year as one time series data set, so total 19 years between 1995 to 2014 with 15 ports lead to $19 * 15 = 285$ data sets. To make similarity clustering, not only considering the similarity showed by different ports at the same year, but also comprehensively considering the similarity demonstrated by different ports at different years and that by the same port at different years.

Ports clustering based on singularity clustering process of throughput time series

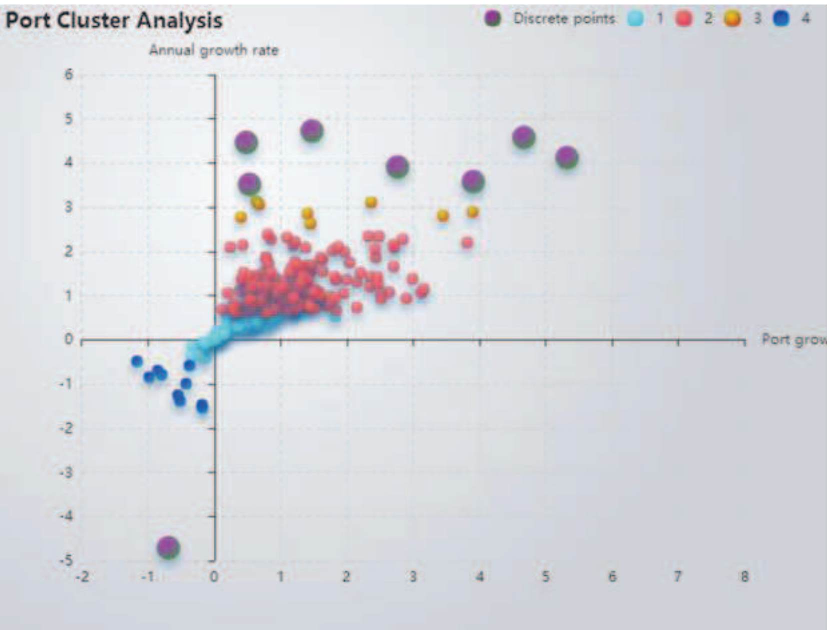

Figure 3: Cluster distribution of coastal container ports above designated size in China

According to the detailed process in 3.2, the discrete class results figured out are as follows (please note that the number for this year demonstrates the situation of last year):

[1.474297, 4.726063, 100, "1996-Yingkou Port"],
[0.482091, 4.467265, 100, "1998- Yingkou Port"],
[3.906951, 3.579445, 100, "2007-Yantai Port"],
[0.530989, 3.517523, 100, "1998-Lianyungang Port"],
[-0.693107, -4.709108, 100, "1998-Zhanjiang Port"],
[5.322229, 4.122826, 100, "2009-Zhanjiang Port"],
[4.670751, 4.580626, 100, "2009-Haikou Port"],
[2.764757, 3.911730, 100, "1997-Basuo Port"].

Clustering effect picture is showed in (Fig. 3).

Singularity detection of throughput time series

Take Ningbo-Zhoushan Port as an example, as showed in the table. It demonstrated a continuous upward trend on the whole, but changed greatly in 2006. We have found that the reason is the merger of Ningbo Port and Zhoushan Port, and based on the exploration of related records and expert interviews we have also found that it is the period of financial crisis. (Tab. 3) demonstrates the final turning point formed according to singularity detection, expert experience and major events taking place during the period. As mentioned before, throughput of different ports have different responses to a given event. We can see that in 1998 and 2009, some ports showed distinct turning trend because of the financial crisis, while others kept steady tendency.

Table 3: Turning point of Chinese port market

Time	Event or Major Change in Policy
1997	National key construction project: finishing the infrastructure of coastal ports
Latter half of 1997-1998	Asian financial crisis: leading to financial crisis in Southeast Asia
2006	Ningbo Port and Zhoushan Port merged as Ningbo-Zhoushan Port
2008	Shanghai financial crisis
2009	Administrative Regulations of Port Operation
2013	Official establishment of China (Shanghai) free trade pilot zone

According to the irregular influential factors and the clustering bubble diagram, singularity is figured out through comprehensive analysis of social events or expert experience.

(1) 1996-Yingkou Port. Yingkou Zhongyuan International Container Wharf Limited Liability Company, a container wharf company jointly operated and managed by Yingkou Port Office and China Ocean Shipping Group Company, was set up. The company owns a wharf coastline of 309 meters, a container yard of 150,000 square meters, 2 container unloading bridges, a CFS warehouse of 3000 square meters and other related infrastructure. It can berth one 15,000-tons container ship or two 5000-tons container ships, enabling Yingkou Port achieved container quantity of 14,900 standard containers in 1996.

(2) 1997-Basuo Port. Major coastal ports construction and operation, a national key construction project was finished in 1997 with 22 berths were completed and put into use, among which there are one in 200,000-tones mineral transfer wharf second-stage project in Beilun harbor district of Ningbo Port, five in the second-stage project in Bayuquan harbor district of Yingkou Port, six in the second-stage project in west basin of Yantai Port, three in the second-stage project in front bay of Qingdao Port, and one in the first-stage project in Xiahai of Zhanjiang Port. Throughput of most ports showed an obvious growing trend in this year.

(3) Because of the financial crisis in Southeast Asia from the latter half of 1997 to 1998, most ports showed a distinct deviation, especially Zhanjiang Port, Yingkou Port, Lianyungang Port, which experienced dramatic decline in terms of growth rate, indicating the relatively great influence of the financial crisis on the three ports compared with other ports.

(4) China's container transportation kept high-speed increase in 1999 with an obvious trend of concentrating on major hub ports, thus forming Yangtze River delta regional ports with Shanghai Port being the center, Pearl River delta regional ports represented by Shenzhen including

Shenzhen Port, Shantou Port, Guangzhou Port and Haikou Port, as well as circum-Bohai-Sea regional ports including Dalian Port, Qingdao Port and Yantai Port.

(5) 2007-Yantai Port. Based on repeated port integration, 30 new port projects were constructed in Yantai Port in 2007 with four already existing harbor districts including West district, Penglai district and Longkou district, leading to a throughput of over 100 million tons in 2007.

(6) The international financial crisis resulted in an obvious decline in terms of growth rate of ports, which was closely related to the insufficient supply of goods caused by the financial crisis.

(7) In 2009, ports largely stayed in the condition of insufficient supply of goods because of the international financial crisis in 2008. Although the throughput of most ports kept increasing, the growth rate showed a downward trend. Facing such situation, related administrative department in Zhanjiang and port companies jointly kept going forward by striving for supply of goods and took them to the port, thus leading to a throughput of over 100 million tons again in 2009, exceeding the expected target. In addition, *Some Opinions Concerning Promoting the Construction and Development of Hainan International Tourism Island* put forward by the State Council clearly proposed "implementing supporting policy for business related to international shipping, improving supporting policy for the development of modern logistics industry, forging a shipping hub and logistics center facing Southeast Asia while relying on the hinterland of southern part of China". Therefore, Haikou Port smoothly went through the crisis in 2009 caused by the financial crisis in 2008 and showed a distinct upward trend.

(8) In addition, it can be seen from the growth curve of throughput that the throughput of Shanghai Port reached its max at 2014. Throughput transferred from Shanghai Port to Taicang Port actually indicated the fact that the throughput growth space of Shainghai Port was increasingly narrow.

4.3 Port clustering based on K-means

In order to verify that the port clustering based on singularity can better detect the influence of accidental events on different port throughput time series, and to contrast the deficiency of current frequently-used K-means clustering algorithm, K-means clustering algorithm is adopted to cluster the port throughput time series again, thus making a comparison with the time series based on singularity.

After standardization of original data, it can be seen in (Fig. 4) that the throughput changing trend of the fifteen ports can be largely divided into three classes: fluctuate increase, steady increase, and accelerated increase. To be more specific, it can be divided into five classes. Therefore, we assume it is divided into five classes and choose K-means clustering with the max $K = 5$ in the SPSS statistical analysis software.

(Tab. 4) demonstrates the cluster class and the members in every class. According to the table, the first class: Ningbo-Zhoushan Port; the second class: Shanghai Port; the third class: Yingkou Port, Qinghuangdao Port, Rizhao Port; the forth class: Guangzhou Port, Tianjin Port, Qingdao Port, Dalian Port; the fifth class: Shantou Port, Lianyungang Port, Zhanjiang Port, Yantai Port, Haikou Port, Basuo Port.

4.4 Contrastive analysis of clustering results

It can be seen from the two clustering results in 4.2 and 4.3 that the clustering results obtained through two clustering algorithms are quite different.

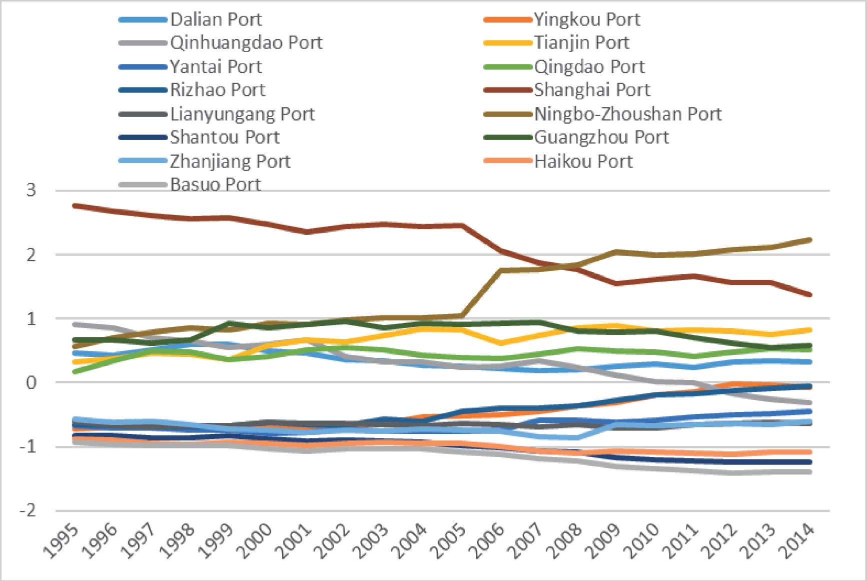

Figure 4: Trends of throughput of major ports along the Coast of China in 1995-2014 (after standardization)

Table 4: Cluster members

No.	Throughput of Port (Unit: ten thousand tons)	Cluster	Distance
1	Dalian Port	4	18180.566
2	Yingkou Port	3	11999.398
3	Qinhuangdao Port	3	22396.025
4	Tianjin Port	4	14222.241
5	Yantai Port	5	22071.565
6	Qingdao Port	4	6685.459
7	Rizhao Port	3	10612.377
8	Shanghai Port	2	.000
9	Lianyungang Port	5	15698.608
10	Ningbo-Zhoushan Port	1	.000
11	Shantou Port	5	16305.687
12	Guangzhou Port	4	12746.441
13	Zhanjiang Port	5	14322.937
14	Haikou Port	5	10183.355
15	Basuo Port	5	25108.685

The clustering result based on singularity is as follows:

"1996-Yingkou Port", "1998-Yingkou Port", "2007-Yantai Port", "1998-Lianyungang Port", "1998-Zhanjiang Port", "2009-Zhanjiang Port", "2009-Haikou Port", "1997-Basuo Port".

K-means clustering results are as follows:

The first class: Ningbo-Zhoushan Port;

The second class: Shanghai Port;

The third class: Yingkou Port, Qinhuangdao Port, Rizhao Port;

The forth class: Guangzhou Port, Tianjin Port, Qingdao Port, Dalian Port;

The fifth class: Shantou Port, Lianyungang Port, Zhanjiang Port, Yantai Port, Haikou Port, Basuo Port.

It can be seen that K-means algorithm can detect the similarity of overall feature variation trend of time series, but is not good at detecting different features of every time series at different stages and does not take into account different demands and concerns of users in real life under different situations. Times series clustering based on singularity just complement such aspect, which not only considers the similarity of different time series at the same time, but also fully considers the possible similarity of different time series at different times and that of a certain time series at different times; that is to say to take different demands and concerns of users in real life under different situations into consideration.

5 Conclusion

The thesis proposes the time series clustering based on singularity according to the shortage of traditional clustering algorithm. As users have different demands and concerns in real life under different situations, researches into similarity clustering process of time series are carried forward from the perspective of singularity, which make choice among trend indicators of time series at every stage and optimize the original data. Different clustering results are obtained through time series clustering based on singularity and K-means respectively. By the comparison of the clustering results, it can be figured out that time series similarity clustering research from the perspective of singularity can better find out the important point of time series.

However, during the singularity detection part, there are some subjective factors to some extent in only the event exploration in relative records and expert interview. If such factors could be quantified, it would be more convincing. Finally, as for the time series that are not involved in the research except ports, the clustering algorithm in the thesis can be adopted to figure out the sensitivity of time series to different major events, thus making precautions to the predicted events with the discrimination theory and prediction model.

Acknowledgment

This work was partially supported by Beijing Logistics Informatics Research Base.

Bibliography

[1] Barlas P., Heavey C., Dagkakis G. (2015); An Open Source Tool for Automated Input Data in Simulation, *International Journal of Simulation Modelling*, 14(4), 596–608, 2015.

[2] Ester M., Kriegel H.P., Xu X. (1996); A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise, *International Conference on Knowledge Discovery and Data Mining*, 226-231, 1996.

[3] Hartigan J.A., Wong M.A. (1979); Algorithm AS 136: A K-Means Clustering Algorithm, *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1), 100–108, 1979.

[4] Honarkhah M., Caers J. (2010); Stochastic Simulation of Patterns Using Distance-Based Pattern Modeling, *Mathematical Geosciences*, 42(5), 487–517, 2010.

[5] Hsu L.C. (2003); Applying the Grey prediction model to the global integrated circuit industry, *Technological Forecasting & Social Change*, 70, 563–574, 2003.

[6] Liao T.W. (2005); Clustering of time series data: a survey, *Pattern Recognition*, 38, 1857–1874, 2005.

[7] Mormann F., Andrzejak R.G., Elger C.E., Lehnertz K. (2007); Seizure prediction: the long and winding road, *Brain*, 130(2), 314–333, 2007.

[8] Munim Z.H., Schramm H.J. (2017); Forecasting container shipping freight rates for the Far East-Northern Europe trade lane, *Maritime Economics* & *Logistics*, 19(1), 106–125, 2017.

[9] Sander J., Ester M., Kriegel H.-P.; Xu X. (1998); Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications, *Mining and Knowledge Discoverye*, 2(2), 169-194, 1998.

[10] Schubert E., Sander J., Ester M., Kriegel H.P., Xu X. (2017); DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN, *ACM Trans. Database Syst.*, 42(3), 19:1–19:21, 2017.

[11] Tang M., Gong D., Liu S., Zhang H. (2016); Applying Multi-Phase Particle Swarm Optimization to Solve Bulk Cargo Port Scheduling Problem, *Advances in Production Engineering and Management*, 11(4), 299-310,2016.

[12] Yang K.W., Zhang P.L., Ge B.F. (2015); A Variables Clustering Based Differential Evolution Algorithm to Solve Production Planning Problem, *International Journal of Simulation Modelling*, 14(3), 525–538, 2015.

[13] Zhou X. (2015); Competition or Cooperation: a Simulation of the Price Strategy of Ports, *International Journal of Simulation Modelling*, 14(3), 463–474, 2015.

[14] Zissis D., Xidias E., Lekkas D. (2015); Real-time vessel behavior prediction, *Evolving Systems*, 7, 1-12, 2015.