Data mining: a tool for detecting cyclical disturbances in supply networks

A A Afify[1], **S S Dimov**[1]*, **M Naim**[2], **V Valeva**[1], and **V Shukla**[2]

[1]The Manufacturing Engineering Centre, Cardiff University Innovative Manufacturing Research Centre, Cardiff, UK
[2]Logistics Systems Dynamics Group, Cardiff University Innovative Manufacturing Research Centre, Cardiff, UK

The manuscript was received on 23 March 2007 and was accepted after revision for publication on 16 August 2007.

DOI: 10.1243/09544054JEM879

Abstract: Disturbances in supply chains may be either exogenous or endogenous. The ability automatically to detect, diagnose, and distinguish between the causes of disturbances is of prime importance to decision makers in order to avoid uncertainty. The spectral principal component analysis (SPCA) technique has been utilized to distinguish between real and rogue disturbances in a steel supply network. The data set used was collected from four different business units in the network and consists of 43 variables; each is described by 72 data points. The present paper will utilize the same data set to test an alternative approach to SPCA in detecting the disturbances. The new approach employs statistical data pre-processing, clustering, and classification learning techniques to analyse the supply network data. In particular, the incremental *k*-means clustering and the RULES-6 classification rule-learning algorithms, developed by the present authors' team, have been applied to identify important patterns in the data set. Results show that the proposed approach has the capability automatically to detect and characterize network-wide cyclical disturbances and generate hypotheses about their root cause.

Keywords: data mining, supply chains, uncertainty, time series, spectral analysis, incremental *k*-means clustering, rule induction

1 INTRODUCTION

A supply network is a system whose constituent parts include material suppliers, production facilities, distribution services, and customers linked together by the feedforward flow of materials and the feedback flow of information [**1–3**]. The system is aimed at matching supply with demand. As such, it should maximize value in terms of satisfying final customer needs in terms of quality, delivery time, availability, and total costs.

A major inhibitor to supply network value delivery is uncertainty [**4–11**]. Uncertainties in supply networks may be described as variance and/or perturbations in the information and material flows, which lead to increased total logistics costs. Production companies in a supply network may either aim to track the variations and thus increase their on-costs, or buffer themselves against such variations via the use of inventory, thereby increasing the risk from stock holding and obsolescence costs. Importantly, such uncertainties may propagate through the supply network [**12**, **13**].

Davis [**12**] classified these uncertainties into three principal categories:

(a) production uncertainties associated with a company's own manufacturing process such as equipment breakdown, operator absenteeism, or material yield losses;
(b) supplier uncertainties related to late deliveries, incomplete orders, or poor-quality products;
(c) customer uncertainties resulting from changes in customer schedules or delivery specifications.

Of particular interest to this research is the detection of uncertainties known as 'rogue seasonalities' which represent cyclical disturbances with periods of few months and which should not be present in supply networks. These are seasonal demand patterns that are induced by the internal processes themselves and not by any external disturbances. Forrester [**14**] highlighted internal decision structures, such as inventory control

*Corresponding author: Manufacturing Engineering Centre, Cardiff University, Queen's Building, The Parade, Newport Road, Cardiff CF24 3AA, UK. email: Dimov@cardiff.ac.uk policies, as the primary cause of such 'rogue seasonality'. More recently, the term 'bullwhip' was used to describe this phenomenon [**15**, **16**]. This is an important area of research for improving the supply network's performance. In particular, the research efforts are focused on developing methods for automated detection of rogue seasonality and also diagnostic tools for identifying its root causes.

Thornhill and Naim [**17**] proposed a data-driven technique known as spectral principal component analysis (SPCA) to identify rogue seasonality by detecting and characterizing cyclical disturbances in a supply network. Principal component analysis (PCA) is an analytical procedure that reduces the dimensionality of the data set by transforming a number of possibly correlated features of the data into a smaller number of uncorrelated features called principal components (PCs). The first few PCs are usually taken for further analysis because they capture the main characteristics of the original data set. In contrast, the last few PCs are often considered only as representative of the residual 'noise' in the data. SPCA performs a PCA analysis on the power spectra (see section 3.1) rather than on the time series data. Detailed descriptions of PCA can be found in reference [**18**].

The empirical study by Thornhill and Naim [**17**] demonstrated the capability of the SPCA technique in identifying seasonal endogenous and exogenous disturbances in a steel industry supply network. Using this technique, consultants or managers can effectively perform a unified statistical analysis to associate groups of variables with a disturbance, and reach conclusions about the root cause of the latter. However, further work is required to automate the clustering process and increase its accuracy. In addition, as was stated by Thornhill and Naim, there is a need to compare SPCA against other techniques with a potential for providing a greater clarity in grouping the variables.

The current paper discusses an alternative data-mining-based approach that automates the identification of network cyclical characteristics. The aim is to develop a method for extracting meaningful signatures from time series data that characterize the dynamical behaviour of a complex supply network. By applying this method the ultimate goal is to be able to differentiate between exogenous induced dynamics, such as customer orders, and rogue seasonality induced by production scheduling systems.

The paper is organized as follows. Section 2 presents an overview of data mining techniques. Section 3 describes the data mining approach for identifying the network cyclical characteristics. Then, section 4 outlines the case study used in this research to illustrate the proposed data mining approach. Empirical results are reported in section 5. Section 6 summarizes the capabilities of the proposed data mining approach.

2 OVERVIEW OF DATA MINING

Recently, the field of data mining, or knowledge discovery in databases (KDD), has seen a great deal of interest from both academia and industry [**19–30**]. Data mining is an interdisciplinary field that attracts researchers with interests in databases, information systems, statistics, machine learning, artificial intelligence, and pattern recognition, with the aim of processing data into knowledge bases for better decision making. It includes all the processing steps in identifying important patterns and relationships in rich databases. There are three main steps in data mining, namely, data preparation, data modelling, and post-processing and model evaluation, as shown in Fig. 1 [**31–34**].

High-quality data are a prerequisite for applying any data mining technique. Prior to data modelling, the data need to be prepared. The objective at this stage is two-fold: to convert the data in a format required by the data mining algorithms and also to expose as much information as possible to data modelling. Data preparation includes such steps as data selection,

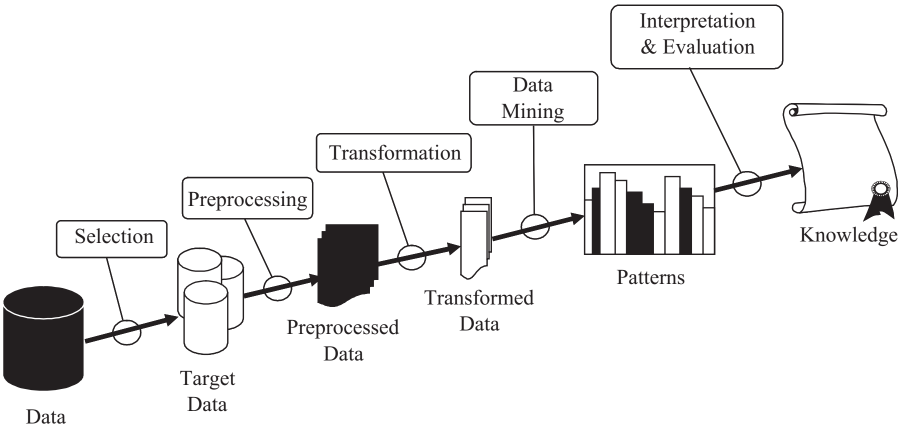


Fig. 1 Data mining process

data cleansing, and data transformation. First, the data might need to be extracted from different sources and then integrated to obtain a data set. After this initial compiling of the data, data cleansing needs to take place. This includes many activities such as identifying and removing the duplicate data objects, and replacing or deleting the missing values. After consistent and clean data sets have been formed, data sampling and feature selection techniques are usually employed to reduce the data, thus speeding up the data mining process. Data often contain a mixture of categorical and continuous-valued attributes, and therefore continuous-valued attributes may have to be discretized first [**35**]. Data preparation is the most time-consuming stage in the whole data mining process. While the focus of data mining research has been on developing different modelling techniques, the data preparation process still cannot be fully automated. There is even a lack of integrated tools to support this important task [**36**].

In any data mining application it is very important to select the most appropriate algorithms for processing the available data sets. There are many different kinds of algorithms, such as those for clustering and classification. Clustering techniques are concerned with partitioning of data sets into several homogeneous clusters. These techniques assign a large number of data objects to a relatively small number of groups so that data objects in a group share the same properties while, in different groups, they are dissimilar. Classification learning employs a set of pre-categorized data objects to develop a model that can be used to classify new data objects from the same population or to provide a better understanding of the data objects' characteristics. Clustering can be used as a pre-processing step to identify groups of related data that can be further explored [**37**]. For example, rather than focusing on each data point in the database, the data points can be clustered first, and then each cluster can be summarized and represented by its characteristics, such as its mean and standard deviation. Thus, any subsequent analysis can focus on such a compressed representation of the data. Finally, if it is required, classification learning algorithms can be applied to these clusters to discover patterns within them.

The result of the data mining process needs to be analysed carefully. First, post-processing may be required. The models created by many data mining algorithms are not easily understandable to human beings, e.g. they do not always generate 'if–then' rules that use original data attributes. Thus, these algorithms lack the ability to explain their results. Even by applying techniques that are capable of generating information in an understandable form, the volume of the generated information may be such that it is unusable without additional post-processing or visualization. Therefore, post-processing techniques, e.g. rule merging and rule filtering, are usually employed. It is also important to equip data mining tools with a friendly interface to help users to distil valuable patterns. Second, the generated model should be validated against the domain knowledge. The strengths and weaknesses of the model should be explained in the problem context. Furthermore, the model needs to be tested using test data or cross validation [**38**].

3 PROPOSED DATA MINING APPROACH

In this research it is proposed to identify cyclical characteristics of supply networks by applying data mining techniques. This approach includes the following steps.

1. Preparing and transforming data into a format suitable for data mining algorithms.
2. Using clustering techniques to discover groups of related data.
3. Applying classification learning techniques to produce a compact description of the discovered groups in step 2.
4. Identifying the exogenous and endogenous cyclical characteristics of the network from the discovered groups and their description.

This section discusses the techniques applied in this research to realize the first three steps of the proposed approach. Step 4 is usually performed by an expert who makes a decision based on the structured information provided to them.

3.1 Data preparation

To prepare the data for further processing, supply network time series data are transformed into the frequency domain using the discrete Fourier transform (DFT). The Fourier transform is a well-known technique [**39**, **40**] that decomposes a time series into a linear combination of sinusoids at different frequencies. This representation is especially useful in time series data mining because of its ability to reduce the dimensionality of the data (a long time series can be represented by a few sinusoids) [**41**]. Another key advantage arises from its invariance to time delays or phase shifts in the data and its robustness in handling missing data points [**42**]. In this research the fast Fourier transform (FFT) algorithm is used for the transformation.

To prepare the available data for further analysis the following pre-processing steps are suggested.

1. Transform the time series data into a frequency domain using Fourier transformation techniques:

   (a) mean-centre the data to allow the analysis to focus on its deviation from the mean;

(b) calculate the power spectra using the FFT technique;
(c) normalize the power spectra to the unit power in order the sum of the spectral power in each frequency channel to be equal to one;
(d) filter the data (if required) by removing the low frequency features and thus highlight subtle effects.

2. Organize the frequency components into a data object format that is suitable for data mining algorithms.

3.2 Data modelling

This section discusses the data mining techniques that were used to analyse the supply network data. The analysis is performed in two stages. First, the data points, data objects, are grouped into several homogeneous clusters, and then a compact description of each group is generated.

3.2.1 Clustering techniques

Many clustering techniques have been proposed over the years from different research disciplines [**43–49**]; k-means is one of the best known and commonly used clustering algorithms. The algorithm forms k clusters that are represented by the mean value of the data points belonging to each of them. This is an iterative process that searches for a division of data objects into k clusters to minimize the sum of Euclidean distances between each object and its closest cluster centre.

The k-means algorithm is relatively scalable and efficient in clustering large data sets because its computational complexity grows linearly with the number of data objects. However, it is sensitive to the initial selection of cluster centres and requires the number of clusters k to be specified before the clustering process starts. Pham *et al.* [**50–52**] have improved the algorithm to address many of its deficiencies. In particular, a new version called incremental k-means was introduced to reduce the dependence of the k-means algorithm on the initialization of cluster centres [**50**]. To validate the robustness of the new algorithm it has been tested on a number of artificial and real data sets. The results showed clearly that incremental k-means consistently outperforms the original algorithm [**50**]. Therefore, this algorithm was applied in this research to search for interesting and natural clusters in the supply network data.

The incremental k-means algorithm is summarized in Fig. 2. The algorithm starts initially with one cluster, with the number of clusters k being incremented by 1 at each step thereafter. With each increase of k, a new cluster centre is inserted into the cluster with the highest distortion, and the objects are reassigned to clusters until the centres stop 'moving'. The process is repeated until k reaches the specified number of clusters.

3.2.2 Classification learning techniques

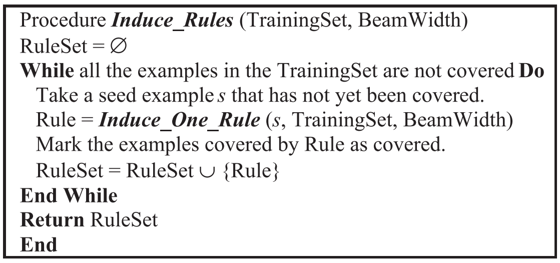
Among the various classification learning techniques developed, inductive learning may be the most commonly used in real-world applications [**53**]. The inductive learning techniques are relatively fast compared to other techniques. Another advantage is that they are simpler and the models that they generate are easier to understand.

In this study a simple inductive learning algorithm called RULES-6 (RULe Extraction System – version 6) [**54**] is used to produce a compact description of clustering results. RULES-6 extracts a set of classification rules from a collection of examples, each belonging to one of a number of given classes. The examples together with their associated classes constitute the set of training examples from which the algorithm generates the rules. Every example is described in terms of a fixed set of attributes, each with its own set of possible values.

In RULES-6, an attribute–value pair constitutes a condition. If the number of attributes is N_a, a rule may contain between one and N_a conditions, each of which must be a different attribute–value pair. Only

```
Assign $k = 1$.
Phase 1: Normal training
  Step 1: If $k = 1$, choose an arbitrary point for a cluster centre.
      If $k > 1$, insert the centre of the new cluster in the cluster with the greatest
      distortion.
  Step 2: Assign each object in the training set to the closest cluster and update its
      centre.
  Step 3: If the cluster centre does not move, go to Phase 2.
      Else, go to Phase 1, Step 2.
Phase 2: Increasing the number of clusters
  If $k$ is smaller than a specified value, increase $k$ by 1 and go to Phase 1 – Step 1.
     Else, stop.
```

Fig. 2 A pseudo-code description of the incremental k-means algorithm



```
Procedure Induce_Rules (TrainingSet, BeamWidth)
RuleSet = $\emptyset$
While all the examples in the TrainingSet are not covered Do
  Take a seed example $s$ that has not yet been covered.
  Rule = Induce_One_Rule ($s$, TrainingSet, BeamWidth)
  Mark the examples covered by Rule as covered.
  RuleSet = RuleSet $\cup$ {Rule}
End While
Return RuleSet
End
```

Fig. 3 A pseudo-code description of the RULES-6 algorithm

conjunction of conditions is permitted in a rule, and therefore the attributes must all be different if the rule comprises more than one condition.

RULES-6 works in an iterative fashion. In each iteration, it takes a 'seed' example not covered by previously created rules to form a new rule. Having found a rule, RULES-6 marks those examples that are covered by it and appends the new rule to its rule set. The algorithm stops when all examples in the training set are covered. A pseudo-code description of RULES-6 is given in Fig. 3.

To form a rule, the procedure Induce_One_Rule performs a general-to-specific beam search for a rule that optimizes a given quality criterion. It starts with the most general rule and specializes it in steps considering only conditions that can be formed from the selected seed example. The aim of specialization is to construct a rule that covers as many examples from the target class and as few examples from the other classes as possible, while ensuring that the seed example remains covered. As a consequence, simpler rules that are not consistent, but are more generic, can be formed. RULES-6 uses effective search-space pruning rules to avoid useless specializations, and terminates any non-productive search during the rule formation. This substantially increases the efficiency of the learning process. A detailed description of the Induce_One_Rule procedure can be found in reference [**54**].

RULES-6 deals with continuous-valued attributes using a pre-processing discretization method [**55**]. With this method, the range of each attribute is split into a number of smaller intervals that are then regarded as nominal values.

4 DEMONSTRATION CASE STUDY

To illustrate the data mining approach proposed in this research a case study from the steel sector is used. The proposed approach was applied on an analytical research data set that was collected from a steel industry supply network. The data set is the same as that used by Thornhill and Naim [**17**] in their empirical study. This is done intentionally to ensure ease in comparison between the SPCA technique proposed by them and the data mining approach discussed in the present paper.

4.1 Network description

The steel industry supply network consists of four autonomous business units as shown in Fig. 4. The 'steel works' unit manufactures a wide range of products that are used as raw materials by the three 'mills' units. The mills units then produce an even wider range of products, which are sold to customers who include stock holders as well as end users. The analytical data used were extracted from the companies' management information systems and were made available in spreadsheet form. To check their consistency, both analytical techniques and opinion-based methods were applied. The data include such variables as flowrates of customer orders, production output, despatch and receipts, and inventory levels such as finished goods, raw materials, and order books. The time series are available monthly and cover a period of six years. The variables to be analysed are listed in Table 1 and are referred to as 'tags'. Their meaning should mostly be self-evident from their descriptions in the table. The order books (tags 29–31) are the orders that have been accepted by a company but have not yet been met. Occasionally, when product is available in stock it is despatched immediately.

4.2 Data pre-processing steps

The available data comprise monthly time series data for 72 months and encompass each of the sources of uncertainty defined by Davis [**12**]. This subsection illustrates how the data pre-processing steps of the proposed approach can be implemented to prepare the supply network data for further processing by data mining algorithms.

First, the FFT algorithm available within the MATLAB software is used to calculate the power spectra. Figure 5 shows the time series and the power spectra of the pre-processed supply network data. The horizontal axis of the spectra is a normalized frequency axis and represents the sampling frequency. Since the data were sampled monthly, a spectral peak for example at 0.25 on the frequency axis, will correspond to a cycle in the time series with a periodicity of 1/0.25, or 4 months. The spectra stop at 0.5 on the frequency axis because the Nyquist sampling theorem requires a sinusoidal signal to be sampled at least twice per cycle [**40**].

The following characteristic points can be observed in Fig. 5 [**1**].

1. The variability in some of the time series such as tags 16, 32, and 33 appears visually to be random,

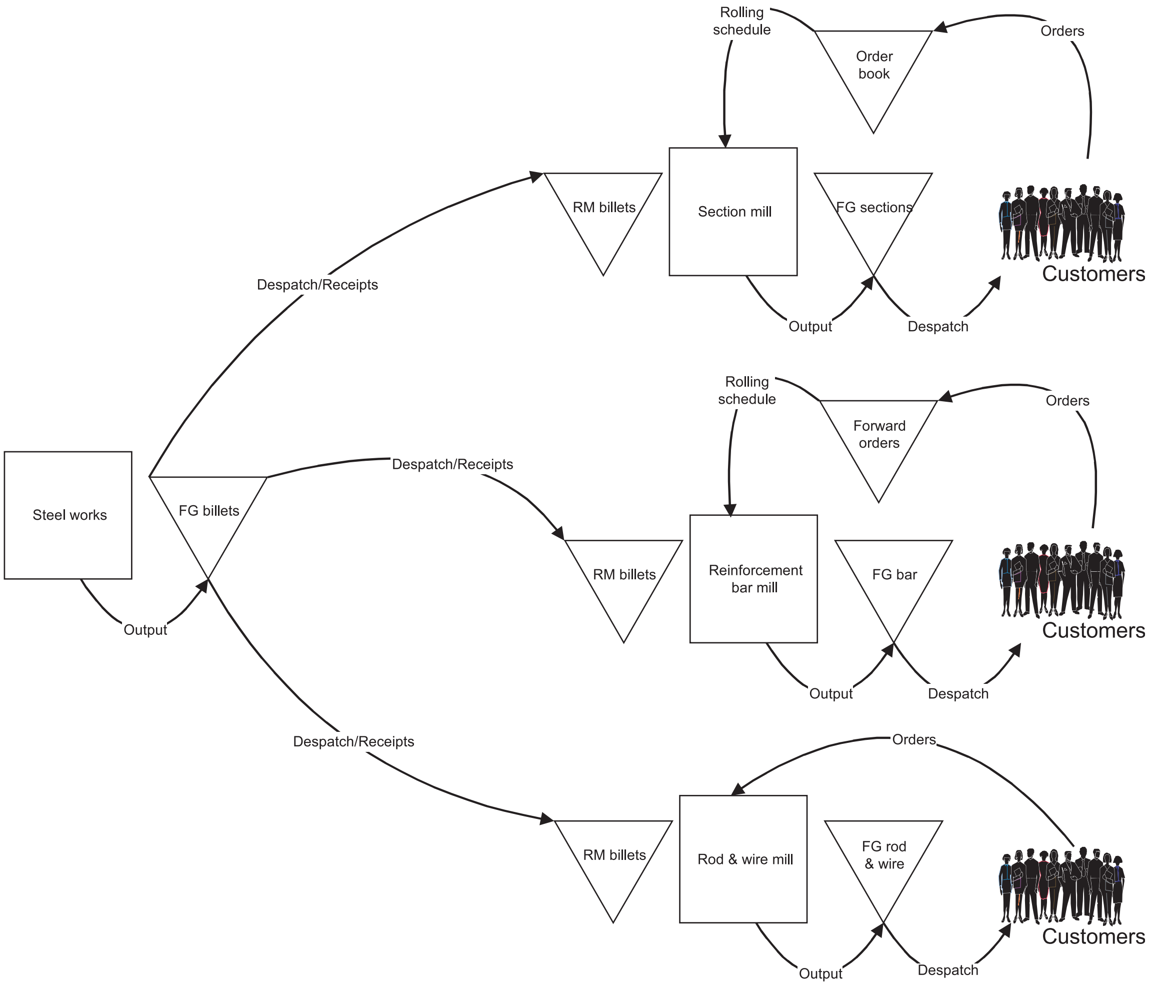


Fig. 4 Supply network under study

while others such as 11, 12, and 13 have a more structured repeating pattern. Peaks in the spectra show that in many cases the variability indeed has a structure.

2. There are prominent spectral peaks at about 0.17, 0.25, and 0.34 on the frequency axis corresponding to oscillatory features in some time series, with periods of about 6, 4, and 3 months. These are the features of interest in this paper because they indicate the seasonality whose origin is to be determined.
3. Some tags show low-frequency spectral features at and below 0.08 on the frequency axis. These correspond to features with periods of more than one year. These trends correspond to long-term effects such as gradual changes in demand because of the long-term economic cycle, and will not be considered further in the present paper because the focus here is on seasonality. However, the modelling techniques will need to distinguish these long-term cycles from the shorter cycles of interest.
4. The data sets were not all complete. Missing data appear as zero values, e.g. in tags 38–43. The main effect of a series of zeros is to distort the low-frequency part of the spectrum.

Second, the spectra are converted into a data set that can be further processed by data mining algorithms. Each tag is considered as an object, and its corresponding frequency coefficients resulting from the Fourier transformation are taken as its attributes. Table 2 shows the structure of the data set created in this way. The next subsection demonstrates how the data mining algorithms are applied on these pre-processed data.

4.3 Data modelling techniques

The supply networks are usually multivariable dynamic systems, in which some of the variables have common behaviour and may be correlated. Therefore, the incremental *k*-means clustering algorithm is applied in this research to process supply networks'

Table 1 The variables of the steel supply network data

Tag No.	Description
Bar mill	
1	Production
2	Shifts worked
3	Output per shift
4	Despatches, home
5	Despatches, export
6	Despatches, total
7	Home orders
8	Export orders
9	Total orders
10	Receipts from FG
11	Receipts, other
12	Receipts, total
13	Billet stocks, total
14	Total stocks
Section mill	
15	Production
16	Shifts worked
17	Output per shift
18	Despatches, home
19	Despatches, export
20	Despatches, total
21	Home orders
22	Export orders
23	Total orders
24	Receipts from FG
25	Receipts, other
26	Receipts, total
27	Billet stocks, total
28	Total stocks
Bar/section mill	
29	Order book, bar mill
30	Order book, section mill
31	Order book, total
Rod mill	
32	Production
33	Shifts worked
34	Output per shift
35	Billet stocks, total
36	Total stocks
37	Total orders
FG billets	
38	Total stocks
39	Despatches, rod
40	Despatches, bar/section
41	Despatches, other
42	Despatches, total
43	Production

data, and thus identify groups of variables with similar profiles. However, the data are pre-processed before clustering, as described in section 4.2. The distortion error was used to evaluate the clustering results [**50**], with a lower value of this measure indicative of better quality of clustering.

The incremental k-means algorithm requires users to specify a number of parameters, namely, the number of clusters and the termination conditions for stopping the clustering process. To find a satisfactory clustering result, a number of iterations were conducted where the algorithm was executed with different values of k, the number of clusters. In this work, the *optimal* number of clusters was considered to be in the range of 1 to 15. The clustering process could be stopped by specifying termination conditions such as a predefined number of iterations and the percentage reduction of the distortion errors in one iteration being smaller than a given value ε. In this work, these two termination criteria were used. In particular, the maximum number of iterations was set to 50 and ε to 10^{-7} to stop the search process when one of these conditions is satisfied.

To assist users in interpreting the clustering results, the RULES-6 algorithm was applied on the discovered clusters. In particular, this inductive learning algorithm was used to extract if–then rules from the data objects forming each cluster. These rules provide a comprehensive insight into the data and can be used to support future decision making. It should be noted that when clustering the pre-processed supply network data, labels such as $G_1, G_2, \ldots, G_n$ are automatically assigned to the discovered clusters and thus the training set for RULES-6 is created.

Two criteria were used to evaluate the performance of the RULES-6 algorithm: the accuracy for the training data set and the complexity of the rule set. RULES-6 has a number of parameters whose values determine the quality of the induced rule sets. In this research the default settings were used [**54**].

5 RESULTS AND DISCUSSION

This section discusses the results of applying the proposed data mining approach on the steel supply network data.

5.1 Clustering results

5.1.1 Full spectra analysis

Figure 6 shows the clustering results obtained when the incremental k-means algorithm is applied to the full spectra of the steel supply network data. Eleven distinct clusters were created. This subsection describes their composition and the important features captured by these clusters. In particular, by analysing the results the following observations could be made.

1. Tags in clusters 1 and 2 have two or more of the distinct spectral features noted earlier at 0.17, 0.25, and 0.34 on the frequency axis, corresponding to cyclical features in the time series with periodicity of about 6, 4, and 3 months, respectively. Thus, by applying the proposed clustering technique, cyclical disturbances in the supply network can be detected. Clusters 1 and 2 feature tags from the bar mill and the section mill, respectively. Both clusters indicate that an exogenous cyclical disturbance is apparent in the endogenous variables.

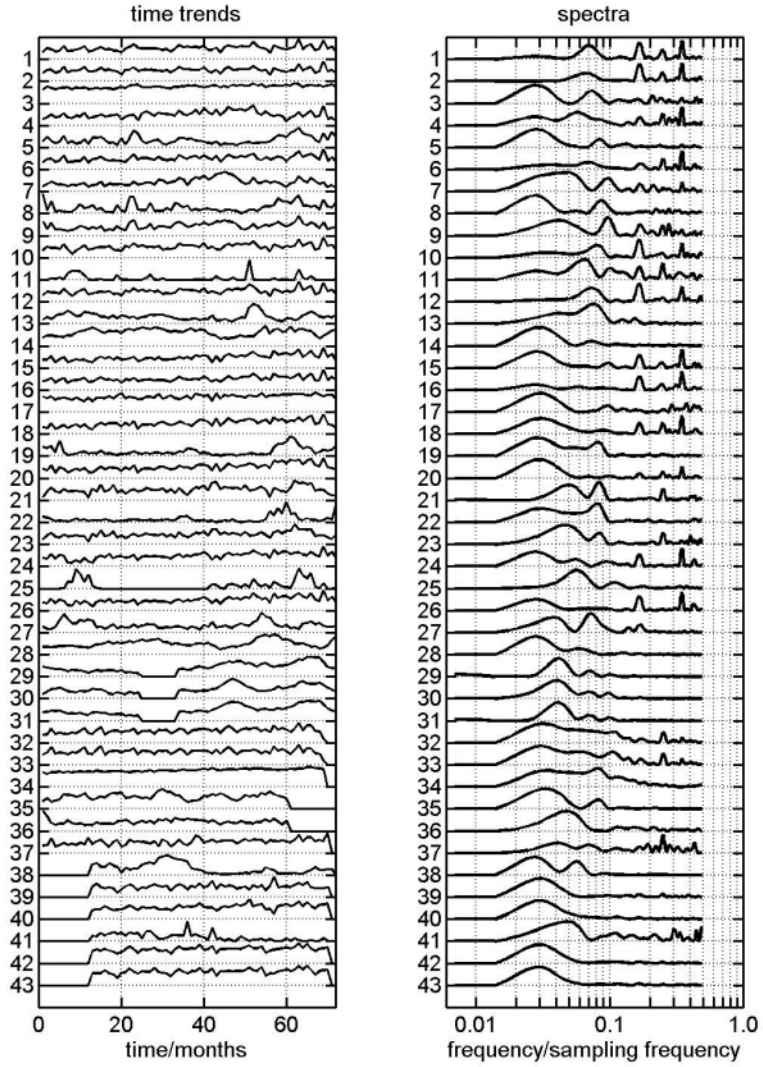
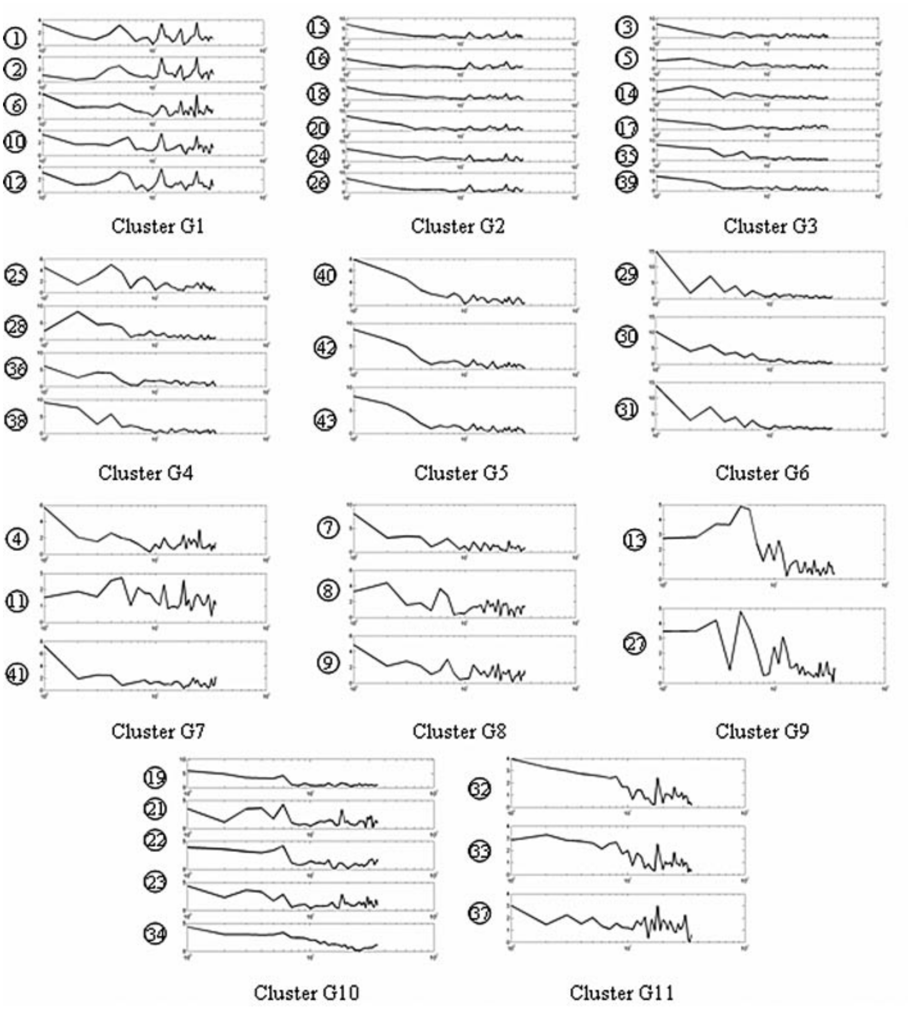


Fig. 5 Time series and spectra of the supply network data

Table 2 The structure of the training set generated by applying DFT on the supply network data

	Discrete Fourier transform (DFT) coefficients				
	F_1	F_2	F_3	...	F_{35}
Tag_1	1.46	0.88	1.72		1.95
Tag_2	0.32	0.57	2.14		1.47
Tag_3	3.51	1.66	0.6		2.25
⋮					
Tag_{43}	5.15	2.99	1.37		3.44

2. Tags in clusters 3, 4, and 5 are characterized by very low-frequency features in the spectrum. Therefore, these tags with long-term, non-stationary trends that reflect long-term changes were grouped together. Clusters 3 and 4 indicate frequencies that appear in all three channels of the network, which are difficult to reconcile from a managerial perspective. For example, it is not obvious why a frequency in tag 17 (section mill's output per shift) should also appear in the rod mill's stock level (tag 36). Cluster 5 has grouped together despatches to the bar mill (tag 40) with total despatches to all three mills (tag 42) and the steel mill's production output (tag 43).
3. Tags in cluster 6 are the order book variables and are clustered together because they have a distinct peak at 0.041 on the frequency axis, a periodicity of about 24 months. This indicates long-term shifts in customer demand due to economic factors external to the supply network.
4. Tags in clusters 7 to 11 are characterized by multiple features over the whole frequency range and thus show a random behaviour in the time domain. Cluster 7 is difficult to explain. Cluster 8 has grouped together all the order tags for the rod mill. Cluster 9 indicates a relationship between raw material stock in the section mill and the bar mill. Managerially this can be explained, as both



Cluster G1 Cluster G2 Cluster G3

Cluster G4 Cluster G5 Cluster G6

Cluster G7 Cluster G8 Cluster G9

Cluster G10 Cluster G11

Fig. 6 Incremental *k*-means clustering results for the full spectra of the supply network data

mills share a common inventory and production control system. In Cluster 10, the grouping of tags 19, 21–23 (section mill) is logical, although their relationship with the rod mill's output per shift (tag 34) is not easily explained. Cluster 11 groups three tags in the rod mill which are intuitively related.

However, owing to the dominance of low-frequency features there is a danger that the interesting and relevant behaviour in the high-frequency range above 0.1 on the frequency axis is not evaluated properly during the clustering process. Thus, cyclical disturbances present in these tags could remain 'hidden' by long-term deviations. To avoid this, the clustering could be conducted on data sets with suppressed low-frequency features.

5.1.2 *Filtered spectra analysis*

An additional analysis was conducted on the same data set. However, this time its low-frequency features were suppressed. The filtering step mentioned in the data preparation procedure was used to remove spectral features with periodicity of 8 months or longer, in particular the features below 0.125 on the frequency axis. This allows the clustering algorithm to focus on grouping the tags based on their high-frequency behaviour and thus identify cyclical disturbances with periodicities of 6, 4, and 3 months, in particular with spectral peaks at about 0.17, 0.25, and 0.34 on the frequency axis, respectively.

Figure 7 shows the clustering results when the incremental *k*-means algorithm was applied to the filtered spectra of the supply network data. Nine distinct clusters were created. The features captured by these clusters are discussed below, together with the advantages of using the filtered spectra. In particular, the following observations could be made by analysing the results.

1. Although it is difficult to explain the grouping of tag 11 with the other tags, tags in clusters 1 and 2 contain a peak at 0.25 on the frequency axis, indicating

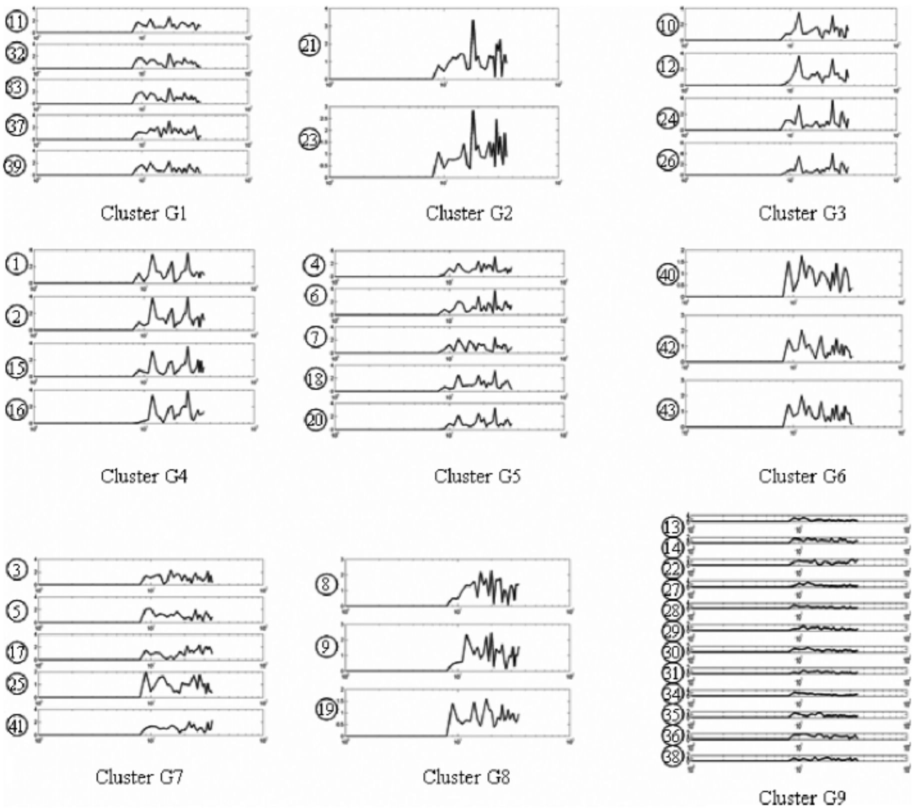


Fig. 7 Incremental *k*-means clustering results for the filtered spectra of the supply network data

a 4 month cycle, but no other salient features. The 4 month cycle is present in the orders of the section and rod mills, i.e. tags 23 and 37. Since the orders are independent inputs to the network it could be concluded that the 4 month cycle is an externally induced disturbance.

2. Tags in cluster 3 have distinct peaks at 0.17 and 0.34 on the frequency axis: 6 month and 3 month cycles, respectively; however, there is no evidence of any 4 month cycle. In particular, the 3 and 6 month cycles are present in the bar and section mills, i.e. tags 10 and 24. They operate an order book and therefore make products for stock that are not necessarily synchronized with order arrivals.
3. Tags in clusters 4, 5, and 6 are dominated by 3, 4, and 6 month cycles. The spectra of tags 2 and 16 (the shifts worked), show 3 and 6 months' features especially strongly – spectral peaks at 0.17 and 0.34 on the frequency axis, respectively. The same cycles also appear for tags 1 and 15, and tag 40, the production and the FG billets' despatches of bar and section mills, respectively. It is worth remembering from section 5.1.1 that the section and rod mills have a common inventory and production control system.
4. Tags in clusters 7, 8, and 9 have spectral content across the whole range and are thus random in the time domain.

Tag 4 is a good example of the benefit that an analysis of the filtered spectra could offer. Based on the full spectra analysis, it was categorized as having multiple features across the frequency range and appeared in a cluster difficult to justify. However, now, after suppressing the low-frequency behaviour, it is clear that this tag has three distinct spectral peaks, and thus shows 6, 4, and 3 months' cyclical behaviour and appears in a distinctive explainable cluster.

The following conclusions can be drawn from the above analysis.

1. The filtering leads to more distinctive and more easily explainable clusters.
2. The 3 and 6 months' cyclical disturbances are rogue seasonalities caused by the bar and section mills' production planning that propagate to material requests from the steel works. In addition, the 3 and 6 months' cycles have an impact on customers because disturbances at the same periodicity are present in the despatches, tags 6 and 20.
3. There are no 3 and 6 months' cycles evident in the rod mill. This mill is a make-to-order factory. The production and shifts-worked variables in the rod mill, tags 32, 33, and 37, all have a 4 month cycle with a spectral peak at 0.25 on the frequency axis. It could be concluded that the production in the rod mill is responsive to customer orders because the periodicity of the shift work, tag 33, is the

R1: IF { |**F18 = 0**| AND |F20 = 1| AND |F25 = 1| AND |F29 = 1| AND |F30 = 1| } THEN Class = G1

R2: IF { |**F10 = 0**| AND |**F18 = 0**| } THEN Class = G1

R3: IF { |**F29 = 0**| } THEN Class = G2

R4: IF { |F1 = 1| AND |F2 = 1| AND |F3 = 1| AND |F4 = 1| AND |F5 = 1| AND |F6 = 1| AND |F7 = 1| AND
|F8 = 1| AND |F9 = 1| AND |F10 = 1| AND |F11 = 1| AND |**F12 = 0**| AND |F13 = 1| AND
|F14 = 1| AND |F15 = 1| AND |F16 = 1| AND |F17 = 1| AND |F18 = 1| AND |F19 = 1| AND
|F20 = 1| AND |F21 = 1| AND |F22 = 1| AND |F23 = 1| AND |F24 = 1| AND |**F25 = 0**| AND
|F26 = 1| AND |F27 = 1| AND |F28 = 1| AND |F29 = 1| AND |F30 = 1| AND |F31 = 1| AND
|F32 = 1| AND |F33 = 1| AND |F34 = 1| AND |F35 = 1| } THEN Class = G3

R5: IF { |**F12 = 0**| AND |F18 = 1| AND |**F31 = 0**| } THEN Class = G3

R6: IF { |F1 = 1| AND |F2 = 1| AND |F3 = 1| AND |F4 = 1| AND |F5 = 1| AND |F6 = 1| AND |F7 = 1| AND
|F8 = 1| AND |F9 = 1| AND |F10 = 1| AND |F11 = 1| AND |**F12 = 0**| AND |F13 = 1| AND
|F14 = 1| AND |F15 = 1| AND |F16 = 1| AND |F17 = 1| AND |F18 = 0| AND |F19 = 1| AND
|F20 = 1| AND |F21 = 1| AND |F22 = 1| AND |F23 = 1| AND |F24 = 1| AND |**F25 = 0**| AND
|F26 = 1| AND |F27 = 1| AND |F28 = 1| AND |F29 = 1| AND |F30 = 1| AND |F31 = 1| AND
|F32 = 1| AND |F33 = 1| AND |F34 = 1| AND |F35 = 1| } THEN Class = G4

R7: IF { |**F20 = 0**| AND |**F25 = 0**| } THEN Class = G5

R8: IF { |**F18 = 0**| AND |**F25 = 0**| AND |**F31 = 0**| } THEN Class = G5

R9: IF { |**F12 = 0**| AND |**F15 = 0**| AND |**F25 = 0**| } THEN Class = G5

R10: IF { |**F12 = 0**| AND |**F14 = 0**| } THEN Class = G6

R11: IF { |**F9 = 0**| AND |**F10 = 1**| AND |**F12 = 0**| AND |**F25 = 0**| } THEN Class = G6

R12: IF { |F14 = 1| AND |**F32 = 0**| } THEN Class = G7

R13: IF { |**F17 = 0**| } THEN Class = G7

R14: IF { |**F27 = 0**| } THEN Class = G7

R15: IF { |F12 = 1| AND |**F20 = 0**| } THEN Class = G8

R16: IF { |F9 = 1| AND |**F20 = 0**| AND |F25 = 1| } THEN Class = G8

R17: IF { |**F9 = 0**| AND |F10 = 1| AND |F12 = 1| } THEN Class = G8

R18: IF { |**F11 = 0**| } THEN Class = G9

R19: IF { |**F9 = 0**| AND |**F20 = 0**| } THEN Class = G9

R20: IF { |F25 = 1| AND |**F30 = 0**| } THEN Class = G9

R21: IF { |**F13 = 0**| AND |F24 = 1| } THEN Class = G9

R22: IF { |**F9 = 0**| AND |F25 = 1| AND |F27 = 1| } THEN Class = G9

Fig. 8 The rule-based description of the clustering results in Fig. 7

same. The material receipts from the steel works to the rod mill, tag 39, also show a 4 month cycle.

5.2 Cluster descriptions

The RULES-6 algorithm was applied on the clustering results obtained for the filtered spectra of the steel supply network data. Figure 8 displays the produced set of rules to describe the clusters shown in Fig. 7. The features F_1, F_2, …, F_n are the DFT coefficients as described in Table 2. It should be noted that in order to focus the analysis on the spectral peaks, the original DFT values, F_i, are transformed to binary values such that

$$F_i = \begin{cases} 0 & \text{if } F_i - \overline{F} > 1.0{*}S \\ 1 & \text{otherwise} \end{cases} \tag{1}$$

where $\bar{F}$ and S are the arithmetic mean and standard deviation, respectively, of the DFT values for each tag.

It is clear from Fig. 8 that the number of rules generated is significantly lower than the number of data objects associated with the discovered clusters. Also, the number of features describing each cluster of tags is drastically reduced. The generated rule set is a compressed description of the clustering results that could be used to identify salient features of data objects, the tags. For example, as was already stated, the spectral peaks of the tags are the features of interest because they indicated seasonalities whose origin should be determined. Therefore, it is important to automate the process of identifying such distinctive spectral peaks and thus help consultants and managers in analysing the clustering results. This could be achieved by focusing only on conditions in the rules that identify spectral peaks for a given cluster.

During the rule-forming process only peaks that help one cluster to be distinguished from another are considered, owing to their 'high information content' in the context of the problem domain. If this is carried out for the rule set in Fig. 8, Table 3 is formed, which includes the important spectral peaks for each cluster. It is not difficult to see that this table depicts the characteristic points observed in Fig. 5. In addition, this table provides information about the coverage of each rule (the number of tags covered by each rule) that could be used to rank the rules. In particular, the rules that cover a higher number of tags are more general, with a higher 'weight' in regards to cluster descriptions, than those formed for fewer data objects. Thus, the users of the clustering results should focus first on analysing the spectral peaks identified by more generic rules.

This is only one example showing how the compressed descriptions of clustering results could be used to analyse the supply network data. Depending on the objectives of the analysis carried out, the rule sets generated using RULES-6 or other inductive learning algorithms could be transferred in other formats to depict better an important behaviour of supply networks.

Table 3 Important spectral peaks used in forming the rules

			Characteristic spectral points															
			F9	F10	F11	F12	F13	F14	F15	F17	F18	F20	F25	F27	F29	F30	F31	F32
Rule	Cluster	Covered tags	0.13	0.14	0.15	0.17	0.18	0.19	0.21	0.24	0.25	0.28	0.35	0.38	0.40	0.42	0.43	0.44
1	G1	11, 37, 39									x							
2	G1	32, 33, 39		x							x							
3	G2	21, 23													x			
4	G3	10, 12, 15, 16				x							x					
5	G3	24, 26				x											x	
6	G4	1, 2, 18, 20				x					x		x					
7	G5	4										x	x					
8	G5	6									x		x				x	
9	G5	7				x			x				x					
10	G6	40, 43				x		x										
11	G6	40, 42, 43	x			x							x					
12	G7	3, 41																x
13	G7	5								x								
14	G7	17, 25												x				
15	G8	8										x						
16	G8	8, 9										x						
17	G8	19	x															
18	G9	13, 29, 31, 34, 36			x													
19	G9	33	x									x						
20	G9	22, 38														x		
21	G9	27, 28, 34, 35					x											
22	G9	13, 14, 28, 30, 34, 35, 36	x															

6 SUMMARY AND CONCLUSIONS

The current paper introduces a new approach that employs statistical data pre-processing, clustering, and classification learning techniques to analyse the time series data of supply networks. The FFT technique is first used to produce the power spectra of the variables in the data set by projecting the time series data into the frequency domain. A key advantage is that power spectra are invariant to time delays or phase shifts in the data and are robust to missing data points. The use of power spectra rather than the time trends has proved useful in characterizing the underlying modes of behaviour in the data set.

The proposed approach has the capability automatically to detect and characterize network-wide cyclical disturbances in a complex supply network. The selected clustering algorithm, the incremental *k*-means algorithm, has the capability to discover interesting groupings of data objects, from which general descriptions can be derived by applying inductive learning techniques, in particular the RULES-6 algorithm. From the clusters and their descriptions, external and internal disturbances could be distinguished easily, although not all clusters can be fully explained from an intuitive managerial perspective.

To illustrate the work of the proposed approach it was applied on steel supply network data. The full and filtered spectra of this data set were used to analyse the behaviour of the network. In the case of the filtered spectra the low-frequency features were removed in order to focus the analysis on high-frequency components that were more important in studying the network's operational behaviour. Then, the incremental *k*-means clustering algorithm was applied to the transformed data in order to identify groups of variables with a common behaviour. After that, the RULES-6 classification learning technique was used to produce a compact description for the discovered groups. Finally, the produced groups and their descriptions were used to identify automatically the important spectral peaks of the network variables that were the features of interest in detecting uncertainties such as rogue seasonalities.

Figure 9 shows the process charts outlining the technical phase of the SPCA technique as described by Thornhill and Naim [**17**] and the data mining approach proposed in this research. The data preparation in both cases is identical, thus ensuring that the clustering methods are applied on the same data objects. However, by applying the SPCA approach, the clusters and their characteristics have to be identified visually. In essence, SPCA is only a technique for reducing data dimensionality by forming new features that are combinations of the initial ones. Using these new features, the data can be represented graphically in two or three dimensions, and groups can be distinguished by visual inspection. Thus, this is a manual clustering method that relies on the expert in capturing the clusters' characteristics and structure. In contrast, the data mining approach proposed in this research automates this process. In particular, it

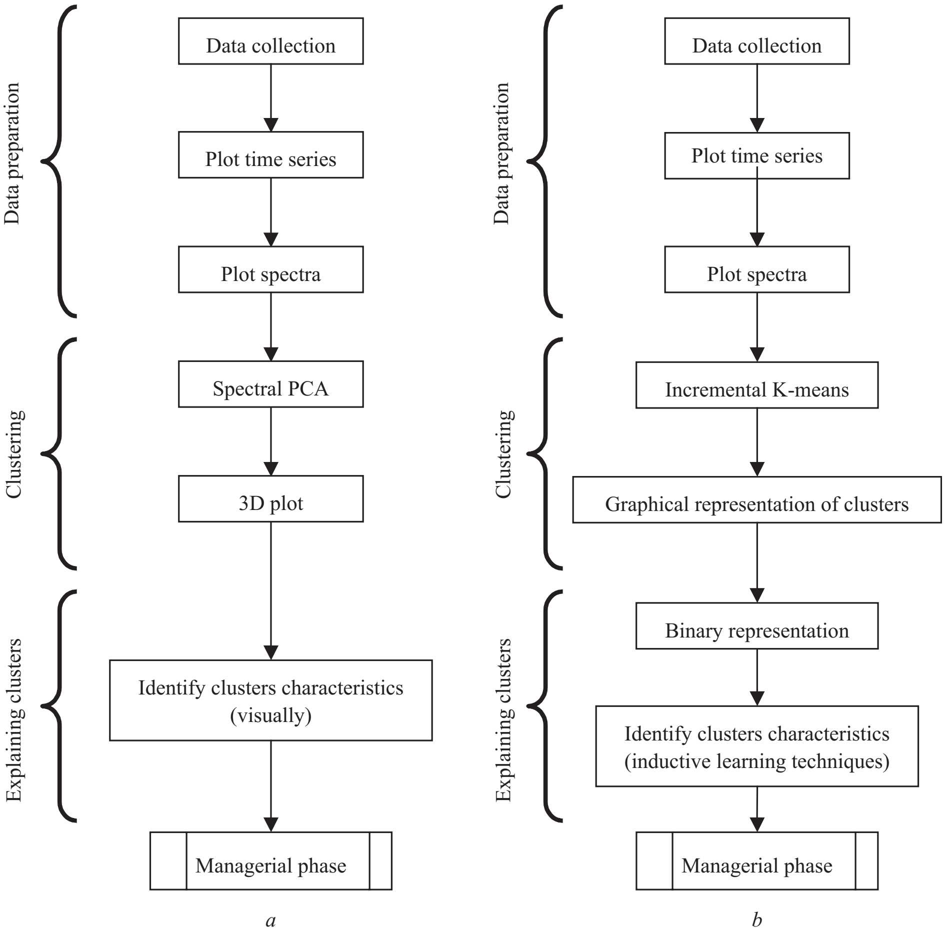


Fig. 9 Comparison of the technical phases of SPCA and data mining approaches

enables rogue seasonality in the data to be detected quickly, consistently, and rigorously, and thus allows managers to spend more time on interpreting the results than on visual and numerical analysis. This is particularly relevant in view of the proliferation of radio frequency identification devices and advances in information and communication (ICT) technologies, which have significantly increased the scale of data collection.

Further testing and comparison of the approach herein, and that by Thornhill and Naim [**17**], is required. This can be undertaken via simulation modelling of a supply network with known structure, parameters, and variables. Stochastic and deterministic disturbances may then be induced in one part of the network with known likely consequences for other variables or parameters. Knowing these causal relationships will lead to a more robust experimental design for testing the approaches.

ACKNOWLEDGEMENTS

This work was carried out within the project 'Mass-Customised Collaborative Logistics for Sustainable Manufacture' (McCLOSM), undertaken as part of the Cardiff University Innovative Manufacturing Research Centre (CUIMRC) and funded by the Engineering and Physical Sciences Research Council (EPSRC) with EPSRC award reference GR/S75505/01. The authors are truly grateful to Professor Nina Thornhill from the Department of Electronic and Electrical Engineering, University College London, for her help in the data preparation.

REFERENCES

1 **Stevens, G.** Integrating the supply chain. *Int. J. Phys. Distribution Mater. Mgmt*, 1989, **19**(8), 3–8.
2 **Pontrandolfo, P., Gosavi, A., Okogbaa, O. G.,** and **Das, T. K.** Global supply chain management: a reinforcement learning approach. *Int. J. Prod. Res.*, 2002, **40**(6), 1299–1317.
3 **Piramuthu, S.** Machine learning for dynamic multi-product supply chain formation. *Expert Systems with Applic.*, 2005, **29**, 985–990.
4 **Mason-Jones, R.** and **Towill, D. R.** Shrinking the supply chain uncertainty circle. *IOM Control*, 1998, **24**(7), 17–22.
5 **Wilding, R.** The supply chain complexity triangle: uncertainty in the supply chain. *Int. J. Phys. Distribution Logistics Mgmt*, 1998, **28**(8), 599–616.

6 **Applequist, G. E., Pekny, J. F.,** and **Reklaitis, G. V.** Risk and uncertainty in managing chemical manufacturing supply chains. *Comput. Chem. Engng*, 2000, **24**, 2211–2222.

7 **Van der Vorst, J. G. A. J.** and **Beulens, A. J. M.** Identifying sources of uncertainty to generate supply chain redesign strategies. *Int. J. Phys. Distribution Logistics Mgmt*, 2002, **32**(6), 409–430.

8 **Kouvelis, P.** and **Milner, J. M.** Supply chain capacity and outsourcing decisions: the dynamic interplay of demand and uncertainty. *IEE Trans.*, 2002, **34**(8), 717–728.

9 **Blackhurst, J., Wu, T.,** and **O'grady, P.** Network-based approach to modelling uncertainty in a supply chain. *Int. J. Prod. Res.*, 2004, **42**(8), 1639–1658.

10 **Childerhouse, P.** and **Towill, D. R.** Reducing uncertainty in European supply chains. *J. Mfg Technol. Mgmt*, 2004, **15**(7), 585–598.

11 **Chan, F. T. S.** and **Chan, H. K.** A simulation study with quantity flexibility in a supply chain subjected to uncertainties. *Int. J. Integr. Mfg*, 2006, **19**(2), 148–160.

12 **Davis, T.** Effective supply chain management. *Sloan Mgmt Rev.*, 1993, **34**(4), 35–46.

13 **Blackhurst, J., Craighead, C. W., Elkins, D.,** and **Handfield, R. B.** An empirically derived agenda of critical research issues for managing supply-chain disruptions. *Int. J. Prod. Res.*, 2005, **43**(19), 4067–4081.

14 **Forrester, J. W.** *Industrial dynamics*, 1961 (MIT Press, Cambridge, Massachusetts, USA).

15 **Lee, H. L., Padmanabhan, V.,** and **Whang, S.** Information distortion in a supply chain: the bullwhip effect. *Mgmt Sci.*, 1997, **43**(4), 546–558.

16 **Wang, J., Jia, J.,** and **Takahashi, K.** A study on the impact of uncertain factors on information distortion in supply chains. *Prod. Planning Control*, 2005, **16**(1), 2–11.

17 **Thornhill, N.** and **Naim, M. M.** An exploratory study to identify rogue seasonality in a steel company's supply network using spectral principal component analysis. *Euro. J. Op. Res.*, 2006, **172**, 146–162.

18 **Jolliffe, I. T.** *Principal component analysis*, 2nd edition, 2002 (Springer-Verlag, New York, USA).

19 **Srinivasan, M.** and **Moon, Y. B.** A comprehensive clustering algorithm for strategic analysis of supply chain networks. *Computers Ind. Engng*, 1999, **36**, 615–633.

20 **Braha, D.** *Data mining for design and manufacturing: methods and applications*, 2001 (Kluwer Academic Publishers, Boston, Massachusetts, USA).

21 **Rygielski, C., Wang, J.-C.,** and **Yen, D. C.** Data mining techniques for customer relationship management. *Technol. Soc.*, 2002, **24**, 483–502.

22 **Monostori, L.** AI and machine learning techniques for managing complexity, changes and uncertainties in manufacturing. *Engng Applic. Artif. Intell.*, 2003, **16**(3), 227–291.

23 **Symeonidis, A. L., Kehagias, D. D.,** and **Mitkas, P. A.** Intelligent policy recommendations on enterprise resource planning by the use of agent technology and data mining techniques. *Expert Systems with Applic.*, 2003, **25**, 589–602.

24 **Caniato, F., Kalchschmidt, M., Ronchi, S., Verganti, R.,** and **Zotteri, G.** Clustering customers to forecast demand. *Prod. Planning Control*, 2005, **16**(1), 32–43.

25 **Chen, M.-C.** and **Wu, H.-P.** An association-based clustering approach to order batching considering customer demand patterns. *Omega*, 2005, **33**, 333–343.

26 **Chen, M.-C., Huang, C.-L., Chen, K.-Y.,** and **Wu, H.-P.** Aggregation of orders in distribution centers using data mining. *Expert Systems with Applic.*, 2005, **28**, 453–460.

27 **Pham, D. T.** and **Afify, A. A.** Machine-learning techniques and their applications in manufacturing. *Proc. ImechE, Part B: J. Engineering Manufacture*, 2005, **219** (B5), 395–412.

28 **Harding, J. A., Shahbaz, M., Srinivas, S.,** and **Kusiak, A.** Data mining in manufacturing: a review. *J. Mfg Sci. Engng*, 2006, **128**, 969–976.

29 **Shahbaz, M., Srinivas, S., Harding, J. A.,** and **Turner, M.** Product design and manufacturing process improvement using association rules. *Proc. IMechE, Part B: J. Engineering Manufacture*, 2006, **220**(B2), 243–254.

30 **Kusiak, A.** and **Smith, M.** Data mining in design of products and production systems. *A. Rev. Control*, 2007, **31**(1), 147–156.

31 **Fayyad, U. M., Piatetsky-Shapiro, G.,** and **Smyth, P.** From data mining to knowledge discovery: an overview. In *Advances in knowledge discovery and data mining* (Eds U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy), 1996, pp. 1–36 (AAAI Press, Menlo Park, California).

32 **Witten, I. H.** and **Frank, E.** *Data mining: practical machine learning tools and techniques with Java implementations*, 2000 (Morgan Kaufmann Publishers, San Francisco, California, USA).

33 **Han, J.** and **Kamber, M.** *Data mining: concepts and techniques*, 2001 (Morgan Kaufmann Publishers, San Francisco, California, USA).

34 **Klösgen, W.** and **Żytkow, J. M.** *Handbook of data mining and knowledge discovery*, 2002 (Oxford University Press, New York, USA).

35 **Pham, D. T.** and **Afify, A. A.** On-line discretisation of continuous-valued attributes in rule induction. *Proc. IMechE, Part C: J. Mechanical Engineering Science*, 2005, **219**(C8), 829–842.

36 **Pyle, D.** *Data preparation for data mining*, 1999 (Morgan Kaufmann Publishers, San Francisco, California, USA).

37 **Jain, A. K., Murty, M. N.,** and **Flynn, P. J.** Data clustering: a review. *ACM Computing Survey*, 1999, **31**(3), 264–323.

38 **Devijver, P. A.** and **Kittler, J.** *Pattern recognition: a statistical approach*, 1982 (Prentice Hall, Englewood Cliffs, New Jersey and London, UK).

39 **Koopmans, L. H.** *The spectral analysis of time series*, 2nd edition, 1995 (Academic Press, San Diego, California, USA).

40 **Chatfield, C.** *The analysis of time series: an introduction*, 6th edition, 2004 (Chapman and Hall, London, UK).

41 **Agrawal, R., Faloutsos, C.,** and **Swami, A.** Efficient similarity search in sequence databases. In Proceedings of the 4th International Conference on *Foundations of data organization and algorithms*, Evanston, Illinois, USA, 1993, pp. 69–84.

42 **Thornhill, N. F., Shah, S. L., Huang, B.,** and **Vishnubhotla, A.** Spectral principal components analysis of dynamic process data. *Control Engng Practice*, 2002, **10**, 833–846.

43 **Michaud, P.** Clustering techniques. *Future Generation Computer Systems*, 1997, **13**, 135–147.

44 **Zaït, M.** and **Messatfa, H.** A comparative study of clustering methods. *Future Generation Computer Systems*, 1997, **13**, 149–159.

45 **Meilă, M.** and **Heckerman, D.** An experimental comparison of model-based clustering methods. *Machine Learning*, 2001, **42**, 9–29.

46 **Grabmeier, J.** and **Rudolph, A.** Techniques of cluster algorithms in data mining. *Data Mining and Knowledge Discovery*, 2002, **6**, 303–360.

47 **Kaufman, L.** and **Rousseeuw, P. J.** *Finding groups in data: an introduction to cluster analysis*, 2nd edition, 2005 (John Wiley, New Jersey, USA).

48 **Liao, T. W.** Clustering of time series data – a survey. *Pattern Recognition*, 2005, **38**, 1857–1874.

49 **Li, T.** A unified view on clustering binary data. *Machine Learning*, 2006, **62**, 199–215.

50 **Pham, D. T., Dimov, S. S.,** and **Nguyen, C. D.** An incremental *k*-means algorithm. *Proc. Instn Mech. Engrs, Part C: J. Mechanical Engineering Science*, 2004, **218** (C7), 783–795.

51 **Pham, D. T., Dimov, S. S.,** and **Nguyen, C. D.** A two-phase *k*-means algorithm for large datasets. *Proc. Instn Mech. Engrs, Part C: J. Mechanical Engineering Science*, 2004, **218**(C10), 1269–1273.

52 **Pham, D. T., Dimov, S. S.,** and **Nguyen, C. D.** Selection of *k* in *k*-means clustering. *Proc. Instn Mech. Engrs, Part C: J. Mechanical Engineering Science*, 2005, **215**(C1), 103–119.

53 **Pham, D. T.** and **Afify, A. A.** Machine learning: techniques and trends. In Proceedings of the 9th International Workshop on *Systems, signals and image processing* (IWSSIP-02), Manchester Town Hall, UK, 2002, pp. 12–36 (World Scientific Publishing Co. Ltd, London).

54 **Pham, D. T.** and **Afify, A. A.** RULES-6: a simple rule induction algorithm for handling large data sets. *Proc. IMechE, Part C: J. Mechanical Engineering Science*, 2005, **219**(C10), 1119–1137.

55 **Fayyad, U. M.** and **Irani, K. B.** Multi-interval discretization of continuous-valued attributes for classification. In Proceedings of the 13th International Joint Conference on *Artificial intelligence*, Chambery, France, 1993, pp. 1022–1027.

APPENDIX

Notation

DFT	discrete Fourier transform
FFT	fast Fourier transform
$\bar{F}$	the arithmetic mean of the F_i values
F_i	the *i*th DFT coefficients for each tag
G_i	the *i*th labels assigned to the discovered clusters in order to create the training set for the RULES-6 algorithm
KDD	knowledge discovery in databases
PCs	principal components
PCA	principal component analysis
RULES-6	Rule extraction system – version 6
S	the standard deviation of the F_i values
SPCA	spectral principal component analysis