Decision Making System for Clustering of Spread Curves

Dr. Anatoliy Antonov, Ventsislav Nikolov
Eurorirsk Systems Ltd., 31, General Kiselov Str., 9002 Varna, Bulgaria
e-mail: Antonov at euroriskSystems dot com, vnikolov at euroriskSystems dot com

Abstract: *In the paper a decision-making system for reducing a big number of multivariate financial spread curves is presented. The core of the system performs clustering and generating of synthetic curves with the same dimensionality. The synthetic curves are generated for using instead of the real curves for other time-consuming calculations which are hard or impossible to be done because of the huge amount of information. The system is based on self-organizing map with modified Kohonen learning rule which allows different priorities for curves to be assigned. The solution is demonstrated for a set of financial spread curves with specified maturities number, the results are shown graphically and analyzed.*

Key words: *Decision Making, Clustering, Self-organizing Map, Spread Curves*

INTRODUCTION

Given is a set of economical spread curves of different issuers (countries, concerns) for a given time period. A spread curve can be understood as an insurance contract. Credit default swap spread curves show insurance against the possibility that an issuer might get into financial trouble and cause money losing on the bond position [3]. Spread curves, similar to other instruments, have maturities of different periods. Maturities for periods smaller than one year are designated as money market (MM) and those for periods equal or larger than one year as capital market (CM).

Examples of actual spread curves on 23.07.2008 are given in table 1. Each curve is specified in a row by its name coding and percentages for different maturities. Curves having different or other maturity points can be interpolated and mapped to selected common maturity points. Every spread curve has also a weight which default value is 100%.

The objective of analysis is reducing the amount of information by grouping of spread curves with similar historical behaviour. This should be made with giving an account of curves weights. A weight interpretation involves number of equal curves. For example, if curve "A" has weight 100% and curve "B" has 400% then it is the same as there are one curve "A" and four curves "B". In addition to clustering a synthetic spread curve for every cluster should be generated. This synthetic curve will represent all other curves in the cluster by best fit and stored into curve manager for later use instead of the real curves. The synthetic spread curve should have synthetic historic development derived from historic development of all spread curves included into cluster.

MOTIVATION

In the real situations the system should reduce a large set (over 1000) of separate individual spread curves to a small set (40 – 50) of synthetic curves. The main reason for solving this task is using of synthetic cluster curves instead of the real curves for time consuming operations, such as correlations estimation with as minimal error as possible. Moreover, the synthetic curves can be used as benchmarks for different financial tasks as pricing and risk estimation of spread based instruments such as credit default swap (CDS) on spread or on index and of spread risks within financial instruments having relevant credit risk exposure such as bonds, loans and etc.

Table 1

	23.07.08	MM	CM	CM	CM	CM	CM
Num	**Maturity(Years)**	**0,5**	**1**	**2**	**3**	**5**	**10**
1	FORTUM-AEUR-MM	0,1439%	0,2293%	0,3047%	0,3686%	0,4647%	0,5828%
2	SRBIA-AEUR-CR	1,2863%	1,2500%	1,7880%	2,1133%	2,6375%	3,0123%
3	UKRAIN-AUSD-CR	1,7606%	1,9218%	2,7834%	3,2418%	3,8589%	4,5407%
4	ITALY-AEUR-CR	0,1022%	0,1167%	0,1896%	0,2727%	0,3965%	0,4980%
5	SLOVEN-AEUR-CR	0,0376%	0,1044%	0,1351%	0,1584%	0,2338%	0,3622%
6	CZECH-AEUR-CR	0,1295%	0,1471%	0,2220%	0,2583%	0,3627%	0,4925%
7	TURKEY-AUSD-CR	0,8137%	0,9853%	1,6326%	2,1247%	2,7918%	3,4798%
8	ROMANI-AUSD-CR	0,5478%	0,5594%	1,1125%	1,3618%	1,7735%	2,0718%
9	POLAND-AUSD-CR	0,0883%	0,1608%	0,2432%	0,3576%	0,5238%	0,6876%
10	PEUGOT-AEUR-MM	0,6865%	0,8433%	1,0406%	1,2925%	1,6265%	1,7360%
11	JPM-CUSD-MR	1,0987%	1,0272%	1,1421%	1,2384%	1,3570%	1,3909%
12	DANBNK-AEUR-MM	0,1830%	0,2635%	0,3686%	0,4481%	0,5610%	0,6098%
13	GS-AUSD-MR	1,1586%	1,1238%	1,1868%	1,2137%	1,2638%	1,2969%
14	CROATI-AEUR-CR	0,1554%	0,2741%	0,4423%	0,5602%	0,8274%	1,0806%
15	ESPSAN-CEUR-MM	0,7725%	0,9577%	1,2941%	1,5505%	1,9120%	1,9371%
16	SUEDZU-AEUR-MM	0,6148%	0,7388%	0,9829%	1,2075%	1,5032%	1,6203%
17	LEH-AUSD-MR	6,7011%	6,8482%	5,4481%	4,5185%	3,2897%	2,6982%
18	HELLAS-CEUR-MM	4,6793%	4,9875%	8,3839%	9,9521%	10,6632%	10,6198%
19	REPHUN-AUSD-CR	0,2793%	0,3050%	0,5168%	0,7809%	1,1361%	1,3992%
20	BGARIA-AUSD-CR	0,3123%	0,4825%	0,8722%	1,1222%	1,5268%	1,8284%

The similarity in behaviour of spread curves is investigated not only for actual rates, but also for all historical rates so it is possible to put the current curve having high current spread rates or specific current rates per maturity into cluster including curves with lower spread rates. This may be occurred when the curve set has different actual rates for current date, but similar historic development and behaviour.

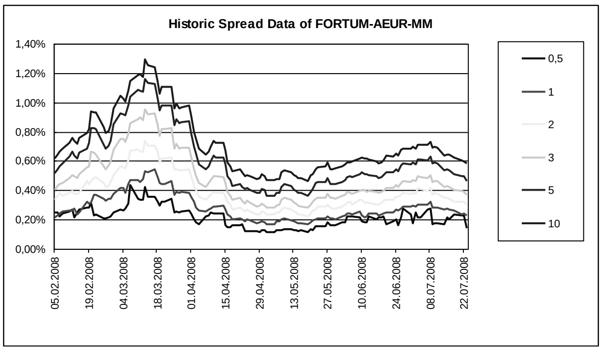


Fig.1 Historic development of spread curve FORTUM-AEUR-MM for last 121 days

The historic development of all seven maturities of spread curve for curve name FORTUM-AEUR-MM for last 121 days is shown on fig.1. As a result of analysis this curve will be included into cluster that includes similar curve behaviour and rates. All seven maturities of each curve are analyzed together as an indivisible unity.

The core of the decision-making system for solving this problem is based on modified unsupervised trained neural network. In addition to the initial data some parameters like curves for clustering, weights, number of clusters, etc. should be specified.

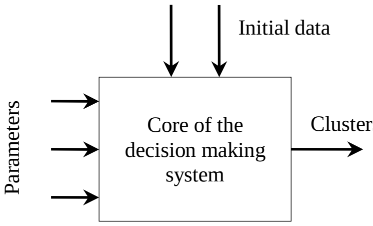


Fig.2 Core of the system

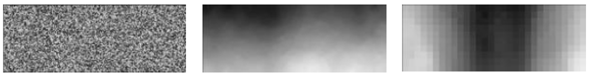
METHODOLOGY

There are many well known clustering techniques [2] [5]. Many comparative analyses have been made between them. Different authors have given different conclusions about statistical results of the comparative analysis. Some authors suggest one approach as better, other author opposite for the same approaches. One of the most commonly used k-means method does not involve mutual influence between different cluster centers. ISODATA finds optimal number of clusters but it requires many parameters which should be precisely adjusted for good results to be obtained. Fuzzy clustering is a good technique for decision making about elements which are on the border between two clusters, but some new information provided by the membership values is applied also for all elements and it must be interpreted [5]. In current solution clustering is made by self-organizing map as a kind of processing model useful for different data operations as clustering and feature extraction. Nevertheless, the self-organizing map allows modification of the general algorithm in order weights for the elements which have to be clustered to be applied.

On fig.3 11100 random generated three elements data vectors (a) are mapped to a lattice of 300 units (c). That is these 11100 vectors have been compressed to 300 (30x10 lattice) vectors. Each colored unit represents more than one color. Thus, the colours represented by a single unit belong to the same cluster. This is the main idea of clustering with using of self-organizing map. The number of output vectors could be set by the user to be smaller than the input so the degree of compression may be much greater.

The number of input units is equal to the number of values in a spread curve and the number of output units is equal to the number of clusters that must be obtained. The Kohonen layer acts as a classifier where all similar input spread curves, namely those belonging to the same cluster are mapped to the same output unit. When the number of real data curves is large and the number of units in the output grid is small then real spread curves are compressed and their representation can be extracted in compressed form by the weights of the connections between input and output units of the self-organizing map. The prototype vector of a grid unit representing given cluster defines the centre of this cluster and this centre is used as representative of the whole set of curves in the cluster. The most important characteristics of the curves in the cluster are as these of the codebook which correspond to the output unit that represents this cluster. Thus the more clusters the more characteristics are preserved from the original data. If the number of clusters is small, then there could be situation some of the characteristics change and/or disappear from the original data but always good approximate values are found. Using the names of spread curves in the mapping process makes possible the set of curves mapped to a given cluster to be identified.

There are two main specificities for implemented systems which affect the architecture of self-organizing map. First, the input is two-dimensional data. The first dimension is historical dates and the second is maturities of the spread curves. Second, the output is one-dimensional grid. Thus, the self-organizing map is created as a structure with two-dimensional input and one-dimensional output. In clustering tasks when mapping ability of the self-organizing map is not essential, one dimensional output layer shows better results compared to two dimensional grids because of the smaller tension exerted to every output unit by the neighbouring units in the case of matrix configuration. This tension restricts the ability of self-organizing map to adapt to the distribution of the initial data [4]. This means that horizontal or vertical size of the output layer is one unit and the other dimension is equal to the number of clusters. That is the reason the self-organizing map in the solution to be realized with one dimensional output.



(a) (b) (c)

Fig.3 11100 random generated colours
(a) mapped to 11100 cells, (b) without compression and (c) to 300 cells

Clustering algorithm for the self-organizing map is as follows:

- Spread curves for clustering are loaded from database.
- Prototype vectors are initialized with random values. Because of random initialization if clustering is performed two or more times with the same conditions, clusters could have the same curves distribution but different order.
- Initial values for learning rate and neigbourhood radius are assigned.
- In each epoch:
  - Self-organizing map is fed by all values of a curve.
  - The output unit which has closest prototype to the curve is winner (best matching unit) and the curve name is written in its curves list.
  - All nodes in the neigbourhood are adjusted to be closer to the current curve according to the modified Kohonen learning rule, where p is a priority of the curve:

$$W(t+1) = W(t) + (1-(1-\theta(t)L(t))^{p})(S(t)-W(t))$$

  - The steps in this epoch are repeated for all curves selected for clustering.
  - New neighbourhood radius is calculated:

$$\sigma(t) = \sigma_0 e^{-\frac{t}{\lambda}}, \text{ t = 1, 2, ... , n}$$

where:

λ is time constant

σ0 is initial radius calculated as a function of the grid size

σ_0=g(x, y)

If neighbourhood radius is zero then the self-organizing map performs clustering like k-means algorithm [1].

- The new learning rate is calculated:

$$L(t)=L_0e^{-\frac{t}{n-t}}$$

where:

L_0 is initial learning rate value

t is the current epoch number

n is the total number of epochs

- Results are obtained and organized for visualization and saving.

Each unit in the input gets a single value from a curve. In addition to the prototype vector, each unit in the output grid contains a list of curves identifiers which are mapped to it. A whole curve is passed to the self-organizing map at a given time. This curve is mapped to one unit of the output grid, the learning rule is applied to the winner and all units in the neigbourhood and then the identifier of the curve is assigned to the winner. This operation is repeated for the whole set of curves and for specified by the user number of iterations. When at a given iteration the winning neuron for a curve is different than the winner for this curve in the past, then the old unit winner deletes the identifier of the curve and the new winner writes in itself the identifier of the curve (fig.4). Thus only one output unit represents a curve and the clustering is correct performed.

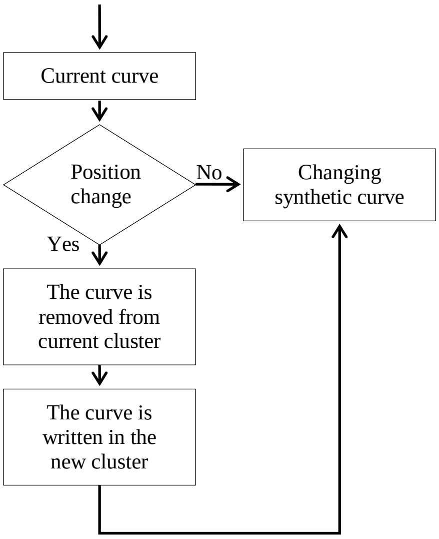

Current curve
Position change
No
Changing synthetic curve
Yes
The curve is removed from current cluster
The curve is written in the new cluster

Fig.4 Decision for the position of a curve

Through the process of clustering the core of the system decides where to put current curve as a best solution according to its current state. Because there are many curves which should be considered in this way, every curve makes corrections to the current state. In the next epoch when the curves pass again a curve which has been classified in a given cluster may match with another cluster. Then the curve should change its position.

Generally speaking the training of self-organizing map can be divided into two main stages. In the first stage a rough changing of the state is done because the neighbourhood radius is large and influences between clusters are large too. In the second stage more precisely adjustment is made. In this stage neighbourhood radius is shrunk to the size of one output unit. The second stage lasts two or more times longer compared to the first stage and often curves do not change their cluster which remains the same to the end of the training. This allows reducing the time of training using different optimization techniques.

RESULTS ANALYSIS

The historical development of all curves is processed by the system. The similarity and main component analysis is performed by learning the historical dependencies of curves rates. The resulted grouping of the spread curves into clusters is shown in table 2.

Table 2

	23.07.08	MM	CM	CM	CM	CM	CM	Std Deviation
Num	**Maturity(Years)**	**0,5**	**1**	**2**	**3**	**5**	**10**	
Cluster 1								0,1725%
1	FORTUM-AEUR-MM	0,14%	0,23%	0,30%	0,37%	0,46%	0,58%	
4	ITALY-AEUR-CR	0,10%	0,12%	0,19%	0,27%	0,40%	0,50%	
5	SLOVEN-AEUR-CR	0,04%	0,10%	0,14%	0,16%	0,23%	0,36%	
6	CZECH-AEUR-CR	0,13%	0,15%	0,22%	0,26%	0,36%	0,49%	
9	POLAND-AUSD-CR	0,09%	0,16%	0,24%	0,36%	0,52%	0,69%	
12	DANBNK-AEUR-MM	0,18%	0,26%	0,37%	0,45%	0,56%	0,61%	
14	CROATI-AEUR-CR	0,16%	0,27%	0,44%	0,56%	0,83%	1,08%	
	Cluster Spread	**0,12%**	**0,19%**	**0,27%**	**0,35%**	**0,48%**	**0,62%**	
Cluster 2								0,30%
8	ROMANI-AUSD-CR	0,55%	0,56%	1,11%	1,36%	1,77%	2,07%	
13	GS-AUSD-MR	1,16%	1,12%	1,19%	1,21%	1,26%	1,30%	
10	PEUGOT-AEUR-MM	0,69%	0,84%	1,04%	1,29%	1,63%	1,74%	
11	JPM-CUSD-MR	1,10%	1,03%	1,14%	1,24%	1,36%	1,39%	
15	ESPSAN-CEUR-MM	0,77%	0,96%	1,29%	1,55%	1,91%	1,94%	
16	SUEDZU-AEUR-MM	0,61%	0,74%	0,98%	1,21%	1,50%	1,62%	
19	REPHUN-AUSD-CR	0,28%	0,31%	0,52%	0,78%	1,14%	1,40%	
20	BGARIA-AUSD-CR	0,31%	0,48%	0,87%	1,12%	1,53%	1,83%	
	Cluster Spread	**0,68%**	**0,75%**	**1,02%**	**1,22%**	**1,51%**	**1,66%**	
Cluster 3								0,80%
2	SRBIA-AEUR-CR	1,29%	1,25%	1,79%	2,11%	2,64%	3,01%	
3	UKRAIN-AUSD-CR	1,76%	1,92%	2,78%	3,24%	3,86%	4,54%	
7	TURKEY-AUSD-CR	0,81%	0,99%	1,63%	2,12%	2,79%	3,48%	
17	LEH-AUSD-MR	6,70%	6,85%	5,45%	4,52%	3,29%	2,70%	
	Cluster Spread	**2,64%**	**2,75%**	**2,91%**	**3,00%**	**3,14%**	**3,43%**	
Cluster 4								0,00%
18	HELLAS-CEUR-MM	4,68%	4,99%	8,38%	9,95%	10,66%	10,62%	
	Cluster Spread	**4,68%**	**4,99%**	**8,38%**	**9,95%**	**10,66%**	**10,62%**	

Table 2 shows 4 clusters with included original spread curves. Following results are shown:

- The list of spread curves within every cluster
- The actual rates for every spread curve for standard different maturities (0.5, 1, 2, 3, 5 and 10 years)
- The actual rates for synthetic spread curve (called Cluster Spread in Table 2) for every cluster
- A standard deviation of curves within cluster giving a measure for fitting.

The decision-making system generates not only rates for actual synthetic curves, but also for whole history period. This involves:

- Every synthetic curve can be stored on same way as real curves into database having curve head data and historic rates per standard maturity.
- An additional database table should hold the list of represented real curves within the cluster.
- These synthetic curves are selected and used as benchmarks for other calculations instead of real curves.

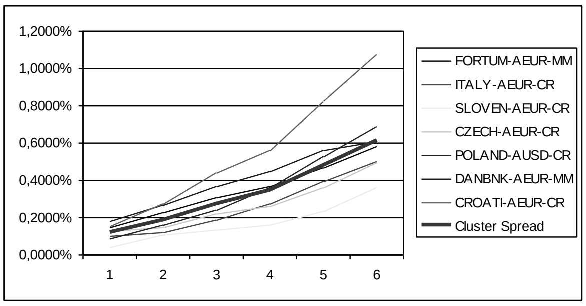


Fig.5 Chart of actual spread curves in Cluster 1 on 23.07.2008

The historic development of synthetic curve (thick, called Cluster) and of real curves for Cluster 2 within all 121 business days for 6 months, 1, 5 and 10 years is represented on Figures 7 to 10.

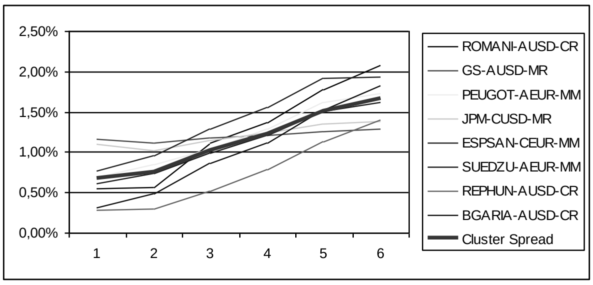


Fig.6 Chart of actual spread curves in Cluster 2 on 23.07.2008

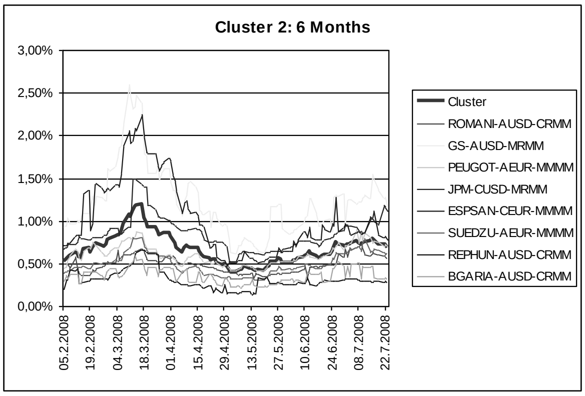


Fig.7 Historic development of 6 months spread curves in Cluster 2 for last 121 days

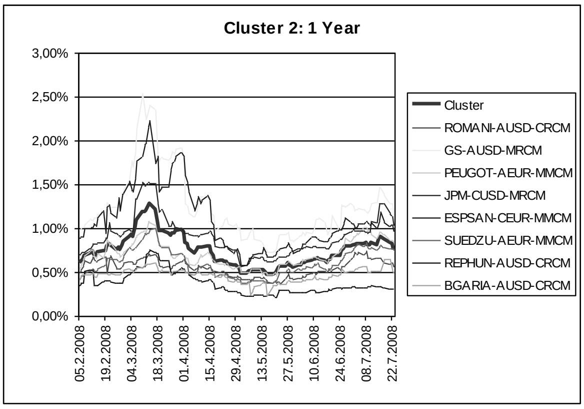


Fig.8 Historic development of 1 year spread curves in Cluster 2 for last 121 days

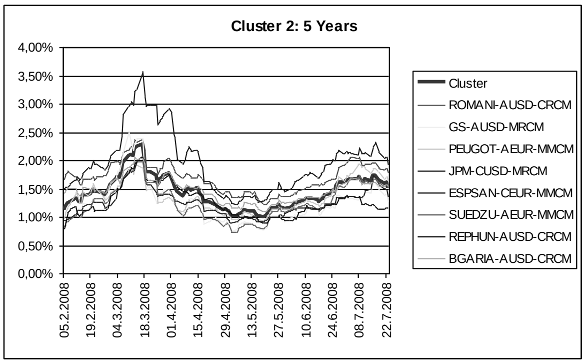


Fig.9 Historic development of 5 years spread curves in Cluster 2 for last 121 days

Specific is the volatility reduction of real spread curve GS-AUSD-MRCR for 5 and 10 years. The spread curve ESPSAN-CEUR-MMCR is permanently higher as all other curves and as the synthetic curve. This can be explained by:

- the small chosen number of clusters where curves are forced for inclusion into inappropriate clusters or
- problematic credit standing of the issuer represented by this curve

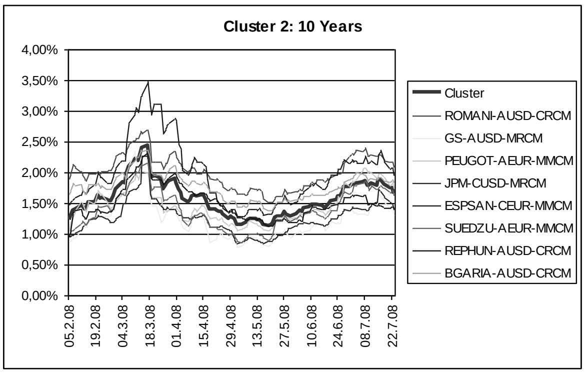


Fig.10 Historic development of 10 years spread curves in Cluster 2 for last 121 days

CONCLUSIONS AND FUTURE WORK

Conducted experiments and the analysis of the results show that the deviance of the curves in each cluster is small enough and using of generated by the decision making system two-dimensional data can be successfully used instead of the real data in situations where data amount can be a problem.

The future work is related to combination of generated data with other calculations such as correlation estimation, prediction, risk estimation, etc.

REFERENCES

1. Jost Schatzmann. Using self-organizing maps to visualize clusters and trends in multidimensional datasets. Imperial College London, 2003.
2. И. Д. Мандель. Кластерный анализ. Финансы и статистика, Москва, 1988.
3. http://richnewman.wordpress.com/2007/12/09/a-beginners-guide-to-credit-default-swaps
4. Fernando Bação, Victor Lobo, Marco Painho. Self-organizing Maps as Substitutes for K-Means Clustering. Springer-Verlag Berlin Heidelberg. V.S.Sunderam et al. (Eds.): ICCS 2005, LNCS 3516.
5. Anil K. Jain, Richard C. Dubes. Algorithms for Clustering Data. Prentice-Hall, New Jersey, 1998.