

The 2008 VAST Contest — The Paraiso Manifesto

Mini-Challenge 3 — Cell Phone Calls on Isla del Sueño

Nyenrode Business Universiteit — Eric Melse

SUMMARY

Abstract—This submission presents Spectral Map Analysis (SMA), a data analysis and visualization method that increases the number of dimensions to graph in a biplot from three to six with the CIELAB colorimetric space. My aim is to reveal the internal structure of the VAST 2008 Mini-Challenge 3 data in an unbiased way. The visual analysis follows the suggestion that caller ID 200 might be 'Ferdinando Catalano' and searches for possible structural properties in the cell phone call records that cover the ten-day period in June 2006 and which might disclose the Catalano/Vidro social structure. Spectramap is used for biplot visual analysis of the complete set of data as well as daily subsets and matrix product tables. Color coding of higher decomposed dimensions reveal cluster-like properties of both callers and receivers. Of those, receiver ID 14 profiles markedly in the color coded components of the matrix product tables. Therefore, my best guess is that ID 14 is David Vidro. Caller ID's that profile in the Spectramap biplot match near perfect with calling and receiving records of ID 14. Furthermore, all calling and receiving ID's of ID 200 were cross referenced with those that are also present in the Spectramap biplot cluster. Of these, either ID 283 or 395 are, in my view, most likely of Juan Vidro or Jorge Vidro. Estaban Catalano is most likely ID 382 when I follow the suggestion that Ferdinando is calling him most frequently. However, the next possible alternative offered by Spectramap visual analytics is ID 281 because of an indirect link via ID 42 to the clustered ID 251. As a result, either ID 281 or ID 382 is a new, hitherto unknown member of the Catalano/Vidro social structure. Visual analysis points at ID 281 and the cell phone records show that it is part of an indirect connection between ID's 14 (David) and 200 (Ferdinando) via ID's 42 and 251 (the latter one is member of the cluster). In my view, it is possible that these three ID's all are unknown members of the Catalano/Vidro social structure. Spectramap visual analysis does not offer any other clear association of ID's within the Catalano/Vidro social structure.

Index Terms—Spectral Map Analysis, Biplot, color coding, CIElab, Spectramap.

1 INTRODUCTION

This submission presents a new analysis and visualization method that increases the number of dimensions to graph in a biplot from three to six by using the CIE-LAB colorimetric space. Decomposition clusters in this case a distinct group of caller ID's in which, after further investigation of the cell phone records, enable the identification of the Catalano/Vidro social structure.

2 SPECTRAL MAP ANALYSIS

Lewi proposed Spectral Map Analysis (SMA), as an extension of Principal Components Analysis (PCA) to compute biplots that allow the graphing of calibrated ratio axis between any pair of table items [1][2][3][4][5]. Greenacre has rediscovered the method and uses the term Log Ratio Analysis (LRA) [3]. SMA is a method similar to PCA but especially suitable for the (graphical) analysis of contrast from log ratios [5][6]. SMA has the advantage that the decomposition translates all ratio relations into a single geometric solution of contrast [3][5]. Briefly, SMA involves the logarithmic transformation and factorial decomposition. Instead of only column-centering, as is done with PCA [3][6][7], double-centering is applied to the original data table. This can be thought of as a simultaneous correction for differences in size between columns or objects (the cell phone call receivers) and for differences in importance between rows or measurements (the cell phone callers). Additionally the row- and column-items can be weighted with respect to their relative importance, e.g. by taking weights proportional to the marginal sums of rows and columns. Lewi asserts ([4], 41-51) that it is natural to apply double-centering to two-way tabulations, such as counts of transactions

(like in this case the cell phone records). Usually, two or three decomposed factors (components) are plotted with Cartesian XYZ metrics into a so-called 2D biplot, or 3D data space (assuming that the decomposition renders three or more factors), called Spectramap™ [12]. This visualizes table objects (by loadings of the table columns) and subjects (by scores of the table rows). Objects and subjects are usually symbolized by squares and circles on the 2D biplot and by cubes and globes in the 3D data space (but other symbols are used as well). The objective is to find possible associations between the table items and their contrast [5]. The method is of general use and was applied in pharmaceutical research, competitive positioning and financial analysis [9][10].

3 COLOR CODING

Current color coding algorithms either tag objects with a category color or some form of color scale using the digital RGB additive color model. However, these approaches exclude color coding with a colorimetric system that could facilitate the analysis of the (possible) relation between table objects and subjects. This submission shows that cell phone ID's from callers and receivers, which might be part of the Catalano/Vidro social structure, can be identified from a cluster that is found not only by their location in the three structural components of the data space but also by their color components. In this case, the proposed visual analytics method demonstrates that the next three components render meaningful relations by color coding.

4 DATA ANALYSIS & VISUALIZATION

From the 9833 cell phone call records, 4 were excluded from analysis because these have negative *duration*

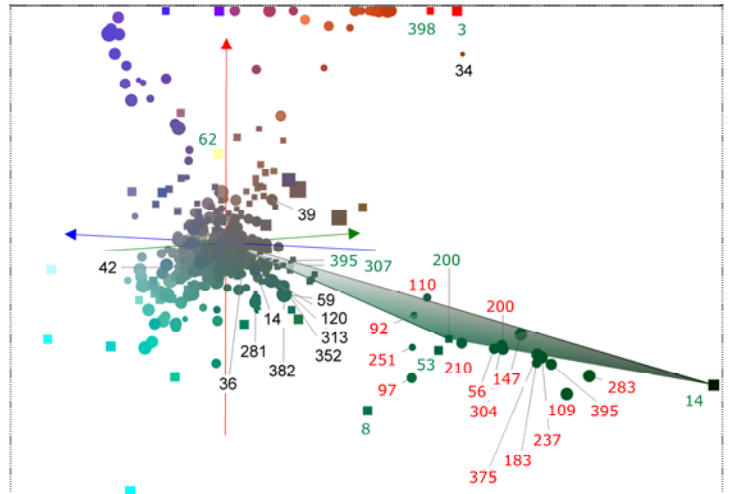
time: R7245, R6450, R4221 and R0822. The remaining records were counted and transformed to a so-called *caller-receiver symmetric product matrix* computed by the total number of calls of each combination of caller-receiver cell phone ID's (i.e. a 400x400 matrix). Data analysis showed that two cell phones never received any calls and these were excluded from the final table: receiver ID's 65 and 108 (i.e. 400x398 matrix). Note that the median value of each row (caller) and column (receiver) of the matrix has a median value of zero calls but that the mean value differs for each ID.

5 DECOMPOSITION

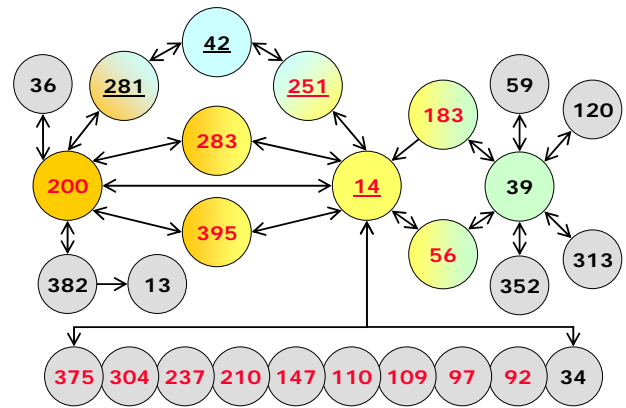
The objective of SMA decomposition is to reduce the number of 'dimensions' of a data table while retaining as much as possible of the variation present, assuming that it has more than 3 columns and rows [5]. The principal reason from a visual analytics perspective to do this is the fact that not more than 3 dimensions can be used to draw and calibrate axis scales of the Cartesian XYZ system for the purpose of graphing [11]. The Eigen value by component is reported in the table below in Panel A by factor and cumulatively by three factors. The scree plot graphs the variance expressed by each factor (component). In it, the red lines show the 'break off' point between structural components and the colorimetric and/or grey components (noise).

6 SOCIAL STRUCTURE

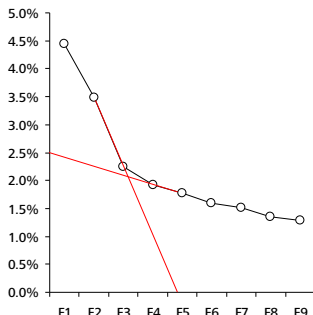
Various biplots were analyzed but one summarizes my findings rather well. If the 4th, 5th and 6th component is used to draw the XYZ data space and as well code the colors of all items (abL), a distinct set of ID's appears as a cluster on the biplot (top figure) [12]. The labels of the potential members of the C./V. social structure are in red. Note that symbol of the callers is a circle and the square is that of the receivers. Observe that receiver ID 200 (Ferdinando) associates markedly with ID 14 (their projection axis to the plot centre has the same angle), and, hence, my conclusion is that this could be David Vidro. The individual cell phone records match near perfect with the red ID's in the Spectramap biplot (table right). This hardly is accidental. Inspection of the records of ID 200 shows that the likely ID's of Juan or Jorge Vidro are 283 and 395 as these matches in the biplot and the cell records. From this, a network can easily be established (middle figure). Finally, note that 'dead end' ID's in the cell phone records are not positioned in the cluster (black ID's), but that an indirect connection is traceable via ID 251.



Spectramap™ identifies the potential ID's in a cluster (456-456).



C./V. social structure: network (red ID's from Spectramap™ cluster).



	Panel A	Panel B
	Cartesian xyz biplot	Color coding
	eigenvalue	cumulative
Factor 1	0.044	
Factor 2	0.035	
Factor 3	0.022	0.102
Factor 4	0.019	
Factor 5	0.018	
Factor 6	0.016	0.053
		0.155

SMA factor (component) decomposition.

Left: scree plot. Right: variance explained & graphed.

		Panel A			
SMA		Caller 14		Receiver 14	
contrast		R by ID		C by ID	
ID	#	ID	#	ID	#
R 14	2	237	4		109
R 200	6	283	3		283
	34	375	3	34	2
	56	56	2	56	5
C 92	92	97	2	92	4
C 97	97	2	1	97	5
C 109	109	6	1	109	18
C 110	110	34	1	110	4
C 147	147	92	1	147	8
C 183	183	109	1	183	10
C 200	200	110	1	200	9
C 210	210	147	1	210	6
C 237	237	200	1	237	16
C 251	251	210	1	251	5
C 283	283	251	1	283	18
C 304	304	304	1	304	10
C 375	375	395	1	375	15
C 395	395			395	11

Panel B			
ID	ID #	ID #	ID #
C 395	R 200 17	R 14 11	
C 382	R 200 16	R 13 13	
C 283	R 200 18	R 14 13	
C 281	R 200 7	R 42 4	
C 42	R 281 2	R 251 3	

Panel C			
ID	ID #	ID #	ID #
C 200	R 14 9		
	R 382 6		
	R 281 5		
	R 283 4		
	R 395 3		
	R 36 1		

C = Caller (globes), R = Receiver (cubes).

Cell Phone Records of Catalano/Vidro social structure.

7 REFERENCES

- [1] Paul J. Lewi, "Spectral mapping, a technique for classifying biological activity profiles of chemical compounds," *Arzneimittel-Forschung / Drug Research*, 1976, 26, pp. 1295–1300.
- [2] Paul J. Lewi, "Spectral mapping, a personal and historical account of an adventure in multivariate data analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 77, no. 1-2, 2005, pp. 215-223.
- [3] Michael Greenacre and Paul J. Lewi, "Distributional equivalence and sub compositional coherence in the analysis of contingency tables, ratio-scale measurements and compositional data," Economics Working Paper 908, Department of Economics and Business, Universitat Pompeu Fabra, March 2008. In press: *Journal of Classification*.
- [4] Paul J. Lewi, *Multivariate data analysis in industrial practice*. Research Studies Press, John Wiley & Sons Ltd, Chichester etc., 1982, ISBN 0471104663.
- [5] Paul J. Lewi, "Spectral map analysis. Factorial analysis of contrasts, especially from log ratios," *Chemometrics and Intelligent Laboratory Systems*, vol. 5, no. 2, 1989, pp. 105-116.
- [6] I.T. Jolliffe, *Principal Component Analysis*. Springer-Verlag, New York, 2002, ISBN 9780387954424.
- [7] F. Cuesta Sánchez, Paul J. Lewi and D. L. Massart, 'Effect of different preprocessing methods for principal component analysis applied to the composition of mixtures Detection of impurities in HPLC—DAD.' *Chemometrics and Intelligent Laboratory Systems*, vol. 25, no. 2, 1994, pp. 157-177.
- [8] Michael Greenacre, "Dynamic Graphics of Parametrically Linked Multivariate Methods Used in Compositional Data Analysis," 2008. Electronic copy available at: <http://ssrn.com/abstract=1124810>.
- [9] Eric Melse, "What Color is your Balance Sheet?," *Balance Sheet*, vol. 12, no. 4, 2004, pp. 17-32.
- [10] Paul J. Lewi, Spectral mapping of drug-test specificities with extensions to the classification of receptor proteins and the correlation of activity spectra. Vrije Universiteit Brussel, Brussel, 1995, dissertation.
- [11] J. C. Gower and David J. Hand, *Biplots*. London, 1995, Chapman & Hall, ISBN 9780412716300.
- [12] Spectramap™ is the name of a trademarked computer program and the graphical display produced by it. Spectral Map Analysis (SMA) is the corresponding mathematical method of multivariate data analysis, it is abbreviated as spectral mapping.

8 ACKNOWLEDGEMENTS

The author wish to thank the helpful assistance of Prof. Dr. P.J. Lewi during the analysis.
This work was supported by a NRG grant from the Nyenrode Research Institute, Breukelen.