# Archetypal analysis and classical segmentation methods. Comparison of two approaches on financial data

**Urszula Grzybowska and Marek Karwański**
Department of Applied Mathematics
Warsaw University of Life Sciences

**Abstract**

*Macroeconomic analyses are, to a large extent, based on firm segmentation and creating homogeneous groups of entities. Thanks to this procedure one can estimate indicators (e.g., various Key Performance Indicators) with high accuracy and examine trends of the market. Unfortunately, the drawback is that segmentation algorithms based on distance measures model average values and are not suitable to analysing unusual events. Modelling in Archetypal Analysis is done in a different way. Both approaches are used e.g., in marketing research, where on one hand one looks for target groups and on the other hand introduces active techniques in form of trend makers. Archetypal Analysis was introduced in 1994 by Cutler and Breiman as a method that provides some kind of reference observations for given data. Archetypes are extreme observations, vertexes of convex hull of the data points obtained as a result of a two-stage nonlinear optimization. The aim of our research is to compare results of analyses done using both, segmentation methods and averaging forecasts and Archetypal Analysis for firms listed on WSE and described by financial indicators (KPI). The authors propose an approach that brings together advantages of both methods. In order to compare Archetypal Analysis and the approach based on segmentation methods, the authors used financial data.*

## Introduction

One of the objectives of multidimensional data analysis is distinguishing patterns or groups of similar objects. If observations under consideration are described by many attributes it may be very difficult or even impossible either to find patterns or distinguish clusters. There may be to many clusters or no clusters at all that makes any interpretation difficult. The reason is known as the curse of dimensionality. In case the observations are described by many attributes, i.e., they are points in multidimensional data space, their high dispersion prevents finding groups of similar objects. The points are sparse in the data space. In our research we have shown archetypes as points in the space obtained with different segmentation methods and visualized in 2-dimensional space of the initial set of objects. We have used classical dimension reduction methods such as **MDS** and **PCA.** Calculations were done for the set of 68 production companies traded on Warsaw Stock Exchange and described with financial indicators. We have also applied a new approach **PHATE** and an optimization method called **DEA**.

**Multidimensional scaling (MDS)** refers to a wide class of visualisation methods that aim to show the structure of a multidimensional data set in a low dimensional setting. It was introduced by Warren S. Torgerson in 1952.

**Principal Component Analysis (PCA)** is a linear method that is applied to reduce dimensionality of large data sets by transforming variables into a smaller set of new variables called principal components, which are linear combinations of the initial variables. Principal components represent the directions of the data that explain a maximal amount of variance.

## Archetypal Analysis

The aim of Archetypal Analysis is to find some representations of objects in a multidimensional data set that would provide reference for other observations. Archetypes are some extreme, not necessarily existing, observations. Each object can be represented as a convex combination of archetypes. Archetypal Analysis enables both visualization of objects in a low dimensional space and giving reference for other objects. It found widespread application in recent years.

**PHATE algorithm** is an affinity-preserving embedding of a multidimensional data set into a two or three- dimensional space that preserves local and global properties of the data structure. PHATE (Potential of Heat-diffusion for Affinity-based Trajectory Embedding) is a novel approach first proposed in 2017 by Moon et al. It was designed as an answer to an increasing need to visualize, explore and interpret high dimensional biological data, mainly genetical. The key idea is connected with application of Diffusion Maps to dimensionality reduction and data visualisation.

## Data Envelopment Analysis (DEA) as a method of ranking and clustering

DEA is a mathematical programming tool for evaluating the performance of a set of objects called Decision Making Units (DMU). The method gives an efficiency rating, i.e., a score $\theta$ for each DMU and an efficiency reference set (a peer group of objects that are efficient), which is a target for the inefficient DMUs. In most of many DEA models the DMUs with the efficiency score equal to 1 are called efficient.

## Data

We have compared described methods on a set of 68 production companies traded on Warsaw Stock Exchange (WSE). In analysis 1 we have used 6 financial indicators: Assets Turnover and Total Liabilities/Total Assets (Debt Ratio) as input indicators (indicators for which law values are preferred) and Return on Assets (ROA), Return on Equity (ROE), Current Ratio (CR), Operating profit margin (OPM) as output indicators (high values are preferred). In analysis 2 we have selected 21 financial indicators published in financial reports to describe the companies under consideration. The indicators were divided into four groups: profitability ratios, liquidity ratios, activity ratios and debt ratios.

## Method

We have performed DEA first for the set of 68 companies described with 6 financial indicators. In order to obtain division into homogeneous groups of companies we have performed the DEA algorithm to the whole set of DMUs. The efficient units with efficiency score 1 constitute the first homogeneous group. After removing all efficient units we applied DEA algorithm to the remaining set. This resulted in distinguishing the next group of units. The procedure was repeated until 6 groups of objects were found. The first group consists of the best 10 companies. For these companies ROA, ROE, CR and OPM values are high and DR and AT low. Next, we have found Archetypes for this set of companies using unitarized values of indicators. To visualize the observations and detect similarities between them we have used PHATE algorithm. The Archetypes have been visualized as objects in this space. Apart from Archetypes, DEA groups were visualized in this space. The same procedure has been repeated for the set of 68 companies described with 21 financial indicators. As a comparison we have also used visualization of objects in the space spanned by two components in PCA and in two dimensions of MDS.

The calculations were done in SAS (ver.9. 4), Python and R.

## Results

### Analysis 1

We have performed DEA first for the set of 68 companies described with 6 financial indicators and distinguished 6 groups of homogeneous objects. The first group consists of the best 10 companies. For these companies ROA, ROE, CR and OPM values are high and DR and AT low. Next, we have found 3 Archetypes for this set of companies using unitarized values of indicators. In Table 1 percentile values of each archetype were presented as percentiles of maximal values of financial indicators.

Table 1. The percentile value in an archetype as compared to the maximum value of the variable

| | AT | DR | ROA | ROE | CR | OPM |
|---|---|---|---|---|---|---|
| Archetype 1 | 91 | 21 | 10 | 9 | 72 | 13 |
| Archetype 2 | 18 | 91 | 40 | 66 | 21 | 47 |
| Archetype 3 | 43 | 25 | 99 | 97 | 96 | 97 |

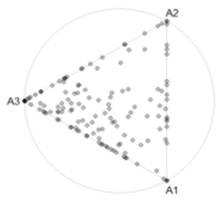


Figure 1. Simplex showing the archetypes as vertexes of the convex hull of the set of observations

Figure 2. displays the percentile plots bar plots (i.e., the percentile value in an archetype as compared to the data) of the three archetypal companies. We can describe archetypes referring to financial indicators.

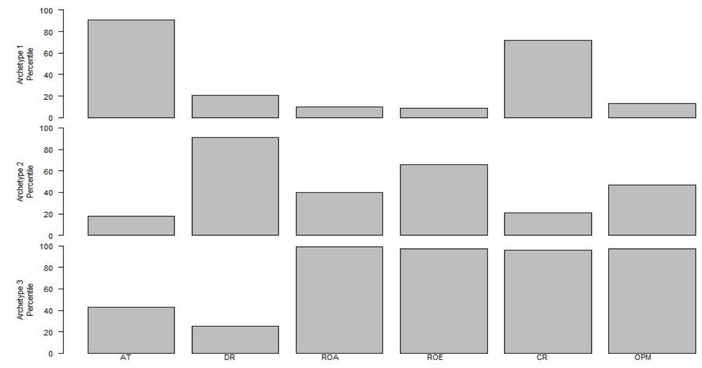


Figure 2. Bar plots showing percentile values of each archetype

### Archetype 1

Archetype one is a company with quite high value of AT, very low values of DR, ROA, ROE and OPM and moderate value of CR. There are two companies that represent this archetype: MOJ and TAURON. The percentiles of maximal values for these companies are given in Table 2. Both firms belong to DEA group 4.

Table 2. The percentiles of maximal values for firms representing archetype 1

| | AT | DR | ROA | ROE | CR | OPM |
|---|---|---|---|---|---|---|
| MOJ | 0.67 | 0.32 | 0.07 | 0.06 | 0.47 | 0.04 |
| TAURON | 0.77 | 0.32 | 0.20 | 0.17 | 0.15 | 0.02 |

Table 3. The percentiles of maximal values for firms representing archetype 2

| | AT | DR | ROA | ROE | CR | OPM |
|---|---|---|---|---|---|---|
| RAFAKO | 0.32 | 0.96 | 0.17 | 0.32 | 0.15 | 0.30 |
| SYNEKTIK | 0.27 | 0.98 | 0.16 | 0.33 | 0.17 | 0.24 |

### Archetype 2

Archetype 2 has low value of AT but quite high value of DR. The other values are moderate ranging from 20% to 66 %. This archetype is represented by RAFAKO and SYNEKTIK, both in DEA group 4.

### Archetype 3

This Archetype can be easily interpreted. There cannot be found an existing company that exactly matches the archetype, but there are two companies very close to it: AC with weight 0.95 and EKO_EXP with weight 0.94. This archetype has low values of AT and DR and high values of other indicators.

Table 4. The percentiles of maximal values for firms representing archetype 3

| | AT | DR | ROA | ROE | CR | OPM |
|---|---|---|---|---|---|---|
| AC | 0.23 | 0.43 | 1. | 1 | 0.52 | 0.71 |
| EKO_EXP | 0.45 | 0.20 | 0.70 | 0.55 | 1 | 1 |

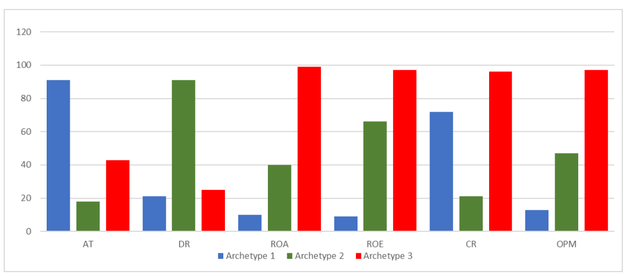


Figure 3. Comparison of three archetypes with respect to percentile values of indicators

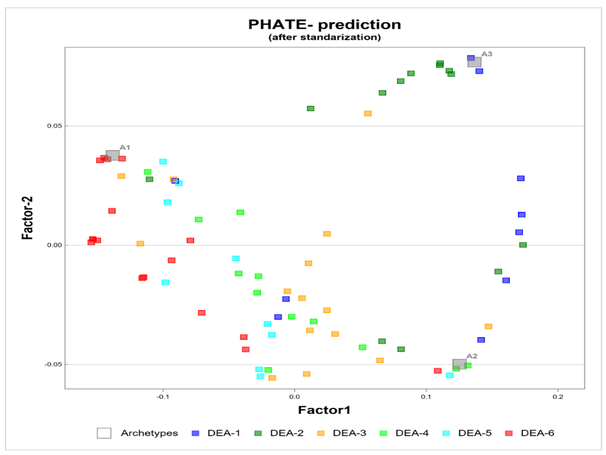


Figure 4. Archetypes in 2- dimensional space generated by PHATE to visualize objects.

PHATE algorithm has been applied to visualize object in 2-dimensional space. Figure 4 shows the result. PHATE confirms the relations that were discovered by archetypal analysis and properly captures the connections to DEA groups 1 and 2. Archetypes are not only extreme observations but representatives of certain groups of objects. PHATE provides good insight into data structure as it shows clusters of objects that are related or close with respect to DEA.

### Analysis 2

We have considered the same set of 68 companies described by 21 financial indicators. We have distinguished 3 archetypes.

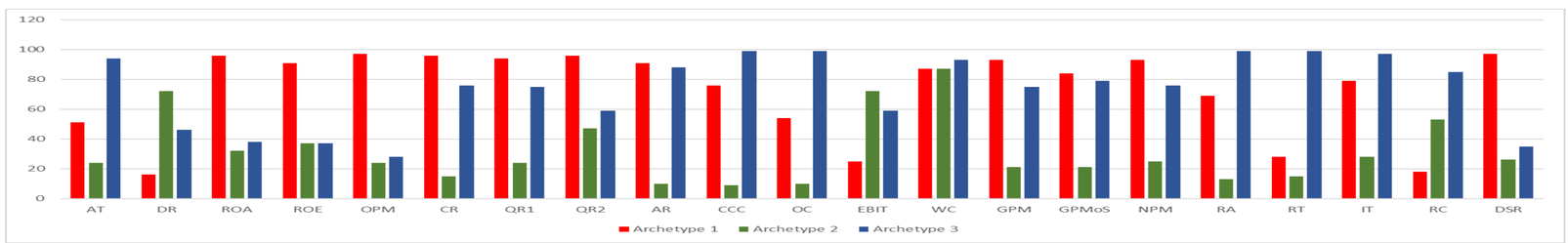


Figure 5. Comparison of three archetypes with respect to percentile values of indicators.

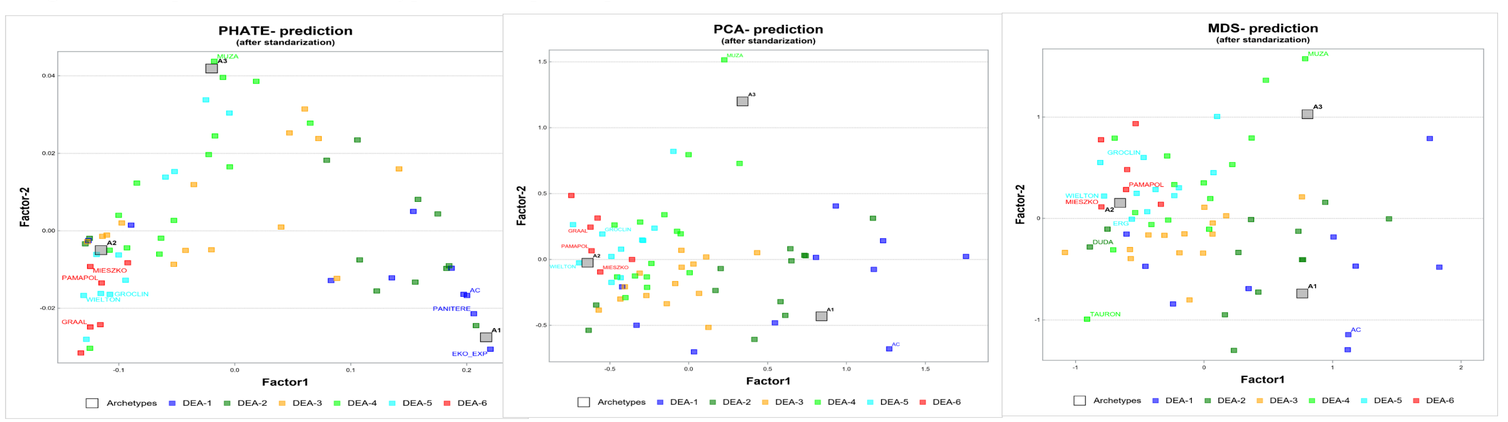


Figure 6. Archetypes in 2- dimensional space generated by PHATE( left) PCA (middle) MDS (right) to visualize objects. Archetypes and DEA groups are shown.

We have compared obtained results with PCA and MDS. As shown on Figure 6, PCA and MDS provide a dispersed visualization. Most objects are close to Archetype 2 and they represent DAE groups with low efficiency score. Archetypes 1 and 3 are not represented by any companies and are far away from any objects.

## CONCLUSIONS

Archetypal analysis supported by PHATE algorithm is a promising tool in multidimensional data analysis. It provides good insight into the data structure, captures similarities between objects and with help of archetypes produces some representations of data in low dimensional space. The results we have obtained show advantage of Archetypal analysis supported by PHATE over PCA and MDS.

## References

1. Cutler A, Breiman L. “Archetypal Analysis.” *Technometrics*, 36(4), 338–347, 1994.
2. Eugster MJA, Leisch F. “From Spider-Man to Hero – Archetypal Analysis in R.” *Journal of Statistical Software*, 30(8), 1–23, 2009. URL http://www.jstatsoft.org/v30/i08.
3. Hastie T, Tibshirani R., Friedman J. *The Elements of Statistical learning*. *Data Mining, Inference, and Prediction.* Springer Series in Statistics. Second Edition, 2009.
4. Keller S. M, Samarin M., Torres FA, Wieser M., Roth V. “Learning Extremal Representations with Deep Archetypal Analysis. ”International Journal of Computer Vision, 129, 805-820, 2020
5. Moon K. R., van Dijk D., Wang Z., Gigante S., et al. *PHATE: A Dimensionality Reduction Method for Visualizing Trajectory Structures in High-Dimensional Biological Data*, „2019.
6. Nadler B, Lafon S, Coifman R. R., Kevrekidis I, *Diffusion maps, spectral clustering and eigenfunctions of Fokker-Planck operators*, Advances in Neural Information. Processing Systems, pp. 955–962, 2005.
7. Nadler B, Lafon S, Coifman R R., Kevrekidis I, Diffusion maps, spectral clustering and reaction coordinates of dynamical systems, Applied and Computational Harmonic Analysis, vol. 21, no. 1, pp. 113–127, 2006.
8. de la Porte J., Herbst B. M., Herman W, van der Walt S. J, An Introduction to Diffusion Maps Conference: Proceedings of the 19th Symposium of the Pattern Recognition Association of South Africa (PRASA 2008) At: Cape Town, South Africa
9. Torgerson W. S., Multidimensional scaling. I. Theory and method, Psychometrika, volume 17, pages 401-419, 1952