Seiler, Christian; Wohlrabe, Klaus

Working Paper

Archetypal scientists

CESifo Working Paper, No. 3990

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Seiler, Christian; Wohlrabe, Klaus (2012) : Archetypal scientists, CESifo Working Paper, No. 3990, Center for Economic Studies and ifo Institute (CESifo), Munich

This Version is available at:
https://hdl.handle.net/10419/67519

CESifo Working Papers
www.cesifo.org/wp

Archetypal Scientists

Christian Seiler
Klaus Wohlrabe

CESIFO WORKING PAPER NO. 3990
CATEGORY 12: EMPIRICAL AND THEORETICAL METHODS
NOVEMBER 2012

CESifo *Working Paper No. 3990*

Archetypal Scientists

Abstract

We introduce archetypal analysis as a tool to describe and categorize scientists. This approach identifies typical characteristics of extreme ('archetypal') values in a multivariate data set. These positive or negative contextual attributes can be allocated to each scientists under investigation. In our application, we use a sample of seven bibliometric indicators for 29,083 economists obtained from the RePEc database and identify six archetypes. These are mainly characterized by ratios of published work and citations. We discuss applications and limitations of this approach. Finally, we assign relative shares of the identified archetypes to each economist in our sample.

JEL-Code: A120, A140.

Keywords: archetypal analysis, classification, scientists, RePEc.

Christian Seiler
Ifo Institute – Leibniz-Institute for Economic Research at the University of Munich
Poschingerstraße 5
81679 Munich
Germany
seiler@ifo.de

Klaus Wohlrabe
Ifo Institute – Leibniz-Institute for Economic Research at the University of Munich
Poschingerstraße 5
81679 Munich
Germany
wohlrabe@ifo.de

1 Introduction

'Albert Einstein is one of the greatest scientists of all time' or 'Leonardo da Vinci was one of the most influential scientists' are typical characterizations in the media. Further keywords in the public associated with science are 'rocket scientists' or 'new wunderkind of science'. But how these characteristics can be traced back to quantitative measures remains unclear. They are often based (maybe implicitly) on an one–dimensional ordering of specific bibliometric measures, like citations, influence or published work. Such measures might also be unobserved or driven by a perceived quality of a certain type.

However, the evaluation of science or scientists is multidimensional and should include observable variables. But there is no uniquely defined strict order (and therefore no minimum and maximum) of such a space. Therefore, a dimension reduction of the collected statistics to an one–dimensional space – where a strict order exists – is often conducted. One can think of, e.g., multidimensional scaling and principal components analysis. The main drawback of these methods is the loss of information but it enables a simple ranking of scientists. A different approach is to build a ranking along each observed dimension and then aggregate these rankings.[1]

Archetypal analysis takes a different road for categorizing scientists: It aims to find a few, not necessarily observed, extreme observations in a multivariate data set such that all data points can be represented as convex combinations of these so–called archetypes or pure types. Therefore, the archetypes lie on the boundary of the data set, i.e. the convex hull. This kind of analysis was introduced by Cutler and Breiman (1994). Since then it has been used in a few applications, e.g., in university benchmarking (Porzio, Ragozini, and Vistocco, 2008), astrophysics (Chan, Mitchell, and Cram, 2003), or sports (Eugster, 2012).

Archetypes can be seen as data–driven extreme values. In science, these extreme values are the archetypal scientists which are outstanding – positively and/or negatively – in one or more of the collected bibliometric statistics. One can think of a scientist with only one paper, but this one attracts enormous number of citations, whereas another one needs many papers

[1] See Seiler and Wohlrabe (2012) for different aggregation approaches.

to achieve the same.

In this paper we introduce the archetypal analysis to bibliometric research. We illustrate the approach with data from the RePEc (Research Papers in Economics) database. This is currently one of the largest bibliometric databases in the field of economics. We extract archetypal economists based on bibliometric scores like citations, published work and number of downloads. Furthermore, we discuss some limitations of the archetypal analysis in practice.

The paper is organized as follows. In Section 2 we outline archetypal analysis and illustrate the approach. Furthermore, we compare it to other multidimensional reduction methods. In Section 3 we describe our data set and identify archetypal economists both for the full sample and a reduced sample containing top economists. Finally, we conclude.

2 Archetypal analysis

2.1 Methodology

Consider an $N \times m$ matrix $\boldsymbol{X}$ representing a multivariate data set with N observations (in our case scientists) and m attributes (e.g. works, citations, downloads, etc.). For a given number of archetypes k the algorithm finds the matrix $\boldsymbol{Z}$ by minimizing the residual sum of squares

$$RSS = \left\| \boldsymbol{X} - \boldsymbol{\alpha}\boldsymbol{Z}' \right\|_2 \tag{1}$$

with $\boldsymbol{\alpha}$ as the coefficients of the archetypes with dimensions $N \times k$ and $\|\cdot\|_2$ denotes an appropriate matrix norm (L^2 norm in this case). Equation (1) is minimized subject to the following constraints

$$\alpha_{ij} \geq 0 \text{ and } \sum_{j=1}^{k} \alpha_{ij} = 1$$

for $i = 1, \ldots, N$. It follows that the k archetypes are convex combinations of the data, i.e.

$$\boldsymbol{Z} = \boldsymbol{X}'\boldsymbol{\beta} \tag{2}$$

where $\boldsymbol{\beta}$ is an $N \times k$ matrix. Equation (2) is estimated with the following constraints

$$\beta_{ji} \geq 0 \text{ and } \sum_{i=1}^{k} \beta_{ji} = 1$$

for $j = 1, \ldots, k$. From equation (2) it can clearly be seen that the archetypes are convex combinations of the original data set $\boldsymbol{X}$. These two ingredients form the basis for the estimation algorithm: it alternates between finding the best $\boldsymbol{\alpha}$ for given archetypes $\boldsymbol{Z}$ and finding the best archetypes $\boldsymbol{Z}$ for given $\boldsymbol{\alpha}$. This approach as known as alternating least squares algorithm as described in Cutler and Breiman (1994).

This approach is similar to a principle component analysis, but relates the observations to the archetypes and not to the components. The estimation involves at each iteration the solution of convex least squares problems. A characteristic of this approach is that the archetypes lie on the boundary of the convex hull of the data by definition. No additional assumptions, e.g. a hypothesis for the distribution of the data, have to be made since archetypal analysis is a purely data–driven method. Given the size of the data set N, which defines the boundary of the convex hull, Cutler and Breiman (1994) showed: if $1 < k < N$, there are k archetypes on the boundary which minimize RSS; if $k = N$, the data points define the convex hull and it follows that RSS $= 0$ and if $k = 1$, the sample mean minimizes the RSS. Archetypal analysis allows to extract *more* archetypes than dimensions of the data set. This is in contrast to the principal components analysis, where the extracted factors cannot exceed the dimension of the data set.

There is no rule for the determination of the correct number of archetypes k for a given data set. One approach is use the 'elbow' criterion, i.e. a screeplot of the RSS. A 'flattening' of the curve suggests a value of k. For detailed explanations we refer to Cutler and Breiman (1994), Eugster and Leisch (2009) on numerical issues, stability, computational complexity and implementation in R (package `archetypes` version 2.1-0 by Eugster and Leisch, 2012), and Eugster and Leisch (2011) for robustness checks.

2.2 An illustrative example

We illustrate the approach with a simple example: We generate a 2–dimensional data set with 50 random draws for each of four different types of multivariate normal distributed variables, i.e. $(x_{(i,r)}, y_{(i,r)}) \sim N(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r), r = 1, \ldots, 4, i = 1, \ldots, 50$.[2] The upper left plot in Figure 1 shows the scatterplot of the data set. We now run an archetypal analysis as described above for $k = 1, \ldots, 5$ archetypes. The gray lines in Figure 1 describe the convex hull of the data. All observations are on or in this hull. We added an outlier at $(10, -1)'$ to demonstrate how the model reacts to extreme values. The red lines display the convex hull defined by k archetypes. For $k = 1$ and 2 we obtain trivial cases of a convex hull defined by a single point or line and is very uninformative. For $k \geq 3$ a 'real' convex hull is constructed. With increasing k, this hull approximates the 'perfect' convex hull. For $k = 4$ the two archetypes in the lower left corner are relatively close to each other. There seems to be no more additional information for more than three archetypes. The screeplot in Figure 2 confirms our assumption. A clear cut can be detected for $k = 3$ archetypes.

For the data set including the outlier one can see that the calculated archetypes are relatively similar. With the exception of the archetype in the lower right corner, all other archetypes are close the those from the original data set. In addition, the screeplot in the Figure 2 supports three archetypes as in the case without the outlier.

2.3 Relation to other data-mining techniques

Given a multivariate data set with N scientists and m characteristics (bibliometric measures), how is archetypal analysis related to other data-mining (reduction) techniques? The screeplot is also used within the a *principle components analysis* (PCA). The PCA is a pure dimension-reduction method to extract 'key factors' within a multivariate data set. These factors are supposed to represent underlying (unobserved) trends common to all variables. The maximum number of factors is given by m. For each extracted component, the individual share of

[2] In our example, we define $\boldsymbol{\mu}_1 = (1, 5)', \boldsymbol{\mu}_2 = (2, 1)', \boldsymbol{\mu}_3 = (1, 4)'$ and $\boldsymbol{\mu}_4 = (-3, 1)'$ and $\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}, \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}, \boldsymbol{\Sigma}_3 = \begin{pmatrix} 1 & 0.1 \\ 0.1 & 1 \end{pmatrix}$ and $\boldsymbol{\Sigma}_4 = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}$.

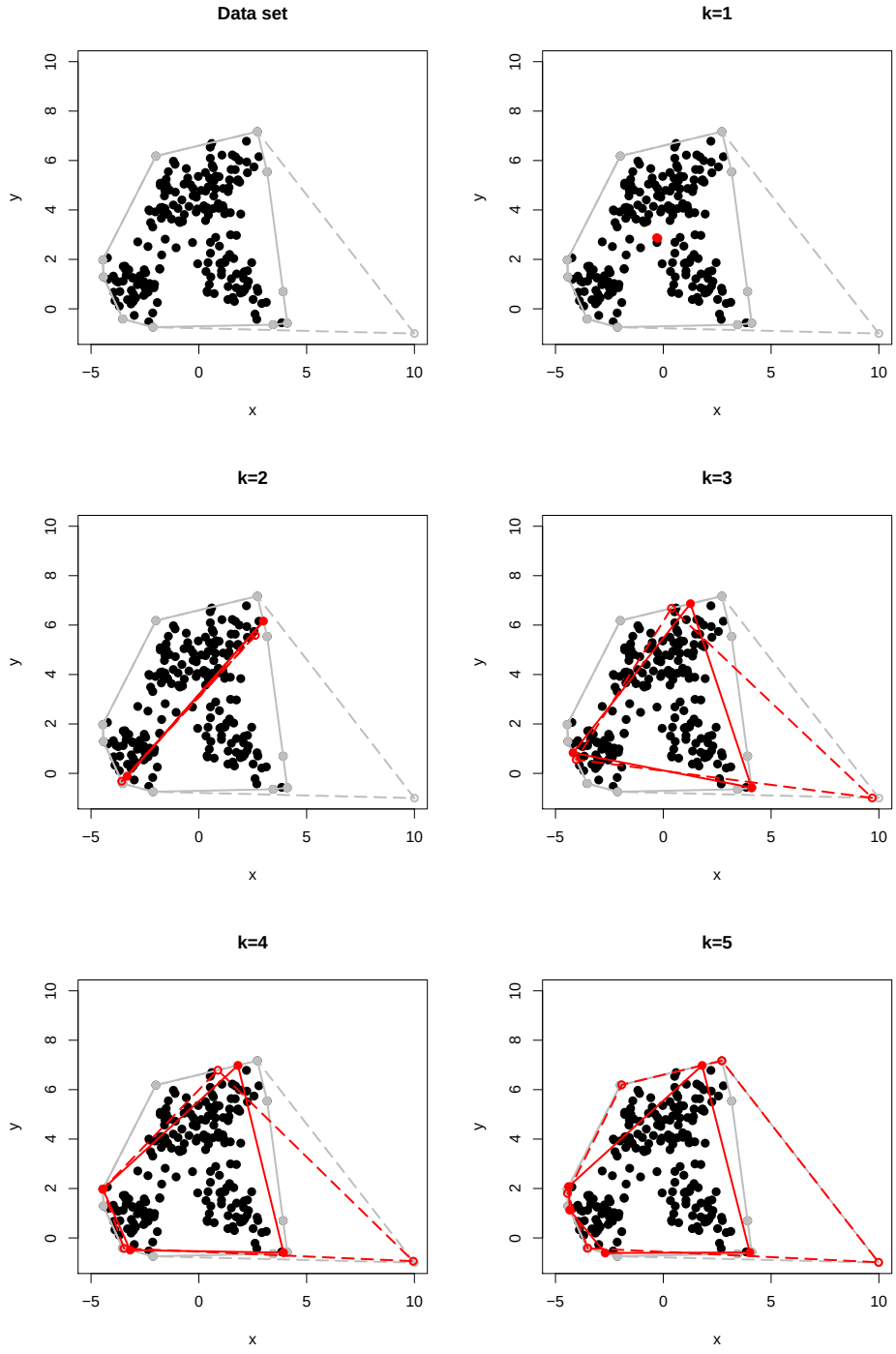

Figure 1: Observations and convex hulls from the archetypal analysis for the artificial data set. ● and ● denote observations, ○ the oulier, ● archetypes of the original data set and ○ archetypes of the data set including the outlier. — denotes the convex hull and — the hull spanned by the calculated archetypes of the original data set whereas - - - denotes the convex hull and - - - the hull spanned by the calculated archetypes of the data set including the outlier.

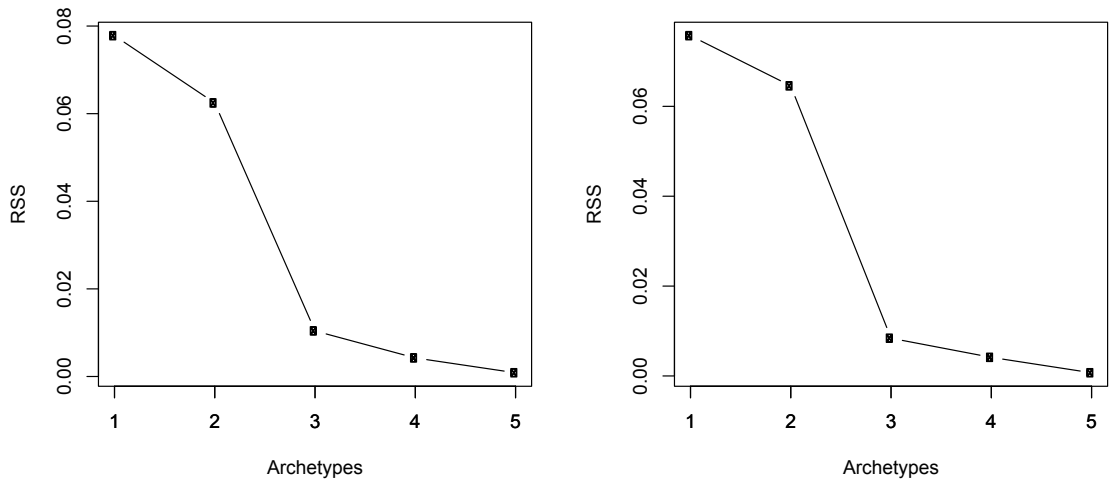

Figure 2: Screeplots for the artifical data set (left) and the data set excluding the outlier (right)

variables can be calculated, i.e. a relative importance of variables can be detected. However, it is not a classification approach. In bibliometric research PCA has been employed to classify determinants of research productivity. See for instance Ramesh Babu and Singh (1998), Costas and Bordons (2007), Franceschet (2009), Docampo (2011), or Ortega, Lopez-Romero, and Fernandez (2011). Seiler and Wohlrabe (2012) suggest a ranking aggregation approach using factor analysis.

Cluster analysis (CA) takes a different route to find structures within a large data set: CA aims to locate layers in a data cloud to separate observations into different clusters. It therefore provides the assignment of single observations to the identified clusters. For recent examples see Franceschet (2009), Osareh (2012), Neff and Corley (2009), or Zhang, Liu, Janssens, Liang, and Glänzel (2010).

Data envelopment analysis (DEA) relates some input measures to some respective output. It ask how efficient a scientist transforms some input measures (e.g. number of works) into output measures (e.g. citations). Thus, the space of m attributes must be divided into input and output variables. The DEA delivers a ranking with the most efficient scientists. The

results are not purely data–driven but by classification into input and output measures. See Halkos and Tzeremes (2011) for an applications and references therein.

3 A case study for economists

3.1 The data set

We illustrate archetypal analysis using a large data set of economists from the RePEc (Research Papers in Economics) website *www.repec.org*. In socio-economic sciences RePEc has become an essential source for the spread of knowledge and ranking of individual authors and academic institutions. It is the same data set as in Seiler and Wohlrabe (2012) consisting of $N = 29,083$ economists with $m = 31$ bibliometric indicators each. These can be grouped into four main categories: number of works, citations, journal pages and access statistics. Within these categories, variables are weighted with different quality measures, i.e. with various impact factors or number of authors. See Seiler and Wohlrabe (2012) for further details and descriptive statistics.

A natural way to start our analysis is to use all scores from all 31 categories for all registered authors. But there are two main objections: First, with such a large data set $(29,083 \times 31)$ the identification of archetypes becomes very burdensome as it needs a lot of computational power to run the programme. Furthermore, the analysis was highly sensitive to starting values and the optimization process, i.e. a robust analysis was not possible. Second, in case of identified archetypes we found that these are grouped in line with the main categories mentioned above. Thus, interpretation of the results would not be clear cut but rather tedious. The reason behind are the high correlations between the indicators.[3] In order to get robust interpretable results we decided to focus on four main categories mentioned above. In addition to the simple counts we include one quality–weighted index with exception of the number of downloads. Finally, we end up with seven indicators

- Published work (Number of distinct works and Number of distinct works weighted by

[3] See Table 6 in Seiler and Wohlrabe (2012).

simple impact factor), also includes working papers, books, software codes, and chapters.

- Number of citations (Number of overall citations, unweighted and weighted by the simple impact factor), represents the impact of an author.
- Number of pages (unweighted and weighted by the simple impact factor), accounts for the published articles.
- Access statistics (Number of downloads).

Thus, we represent in all main categories both the quantity and quality of each economist.[4]

3.2 Full sample results

Given the seven indicators for each economist it is not obvious how many archetypes are reasonable. We perform an analysis with up to 10 potential archetypes. In Figure 3 we show the corresponding screeplot. Using the elbow criterion it is not possible to extract a clear cut–off point, a potential flattening can be detected for 4, 6, and 8 archetypes. By using a two–dimensional data set it is easy to draw the convex hull (see Section 2) and to identify the corresponding archetypes. For three archetypes it becomes more difficult using a 3D–plot and for $k > 3$ archetypes it is impossible. An alternative are barplots which are a different graphical representation of the convex hull. In Figure 4 we show the barplots for four archetypes.[5] The height of each barplot denotes the share of the convex hull relative to the maximum in the category. The percentage numbers denote the respective percentiles. These are helpful for interpretations since they are not biased by potential outliers. On top of that we show the values defining the convex hull. The four archetypes can be interpreted as follows:

- Archetype 1 denotes the standard economist, given the other archetypes. He/she has lot of publications, but a smaller number of citations compared to archetypes 2 and 4. His/her citation–publication ratio (c/p–ratio) is roughly four.

[4] We repeated the following analysis with different quality measures, but the results stay qualitatively the same.

[5] *Dnb_ works* denotes 'Number of distinct works', *Sc_ works* 'Number of distinct works weighted by simple impact factor', *Nb_ cites* 'Number of citations', *D_ cites* 'Number of citations weighted by simple impact factor', *Nb_ pages* 'Number of pages' and *Sc_ pages* 'Number of pages weighted by the simple impact factor'.

- Archetype 2 is also described by many number of works as archetype 1, but these are published in rather low quality outlets on average. Furthermore, the c/p–ratio of about one is lower than in the previous archetype class. Additionally, economists in this group receive a lot of downloads.

- Archetype 3 represents the low–performer with only one publication and no corresponding citations.

- Archetype 4 denotes the top economists with large values in all categories. The c/p–ratio is by far higher, each publication is cited 62 times on average.

Figure 4 reveals how the barplots are influenced by the maximum value in each category. This most obvious for the number of works (first bar in each panel). Although the percentiles do not differ too much their respective heights do.

Now we add two further archetypes to the analysis. The results are shown in Figure 5. We find more differentiated results:

- Archetype 1 is similar to the first one from before. Economists are characterized by high–quality publications and moderate citation success (c/p–ratio: 13).

- Archetype 2 corresponds to economists with more published work compared to archetype one, but on average in lower ranked series. Furthermore the c/p–ratio is lower (about 2).

- Economists categorized by archetype 3 publish a lot of works, but mainly in working paper series since number of pages are relatively small. Both the number of citations and journal articles are below the tenth percentile.

- Archetype 5 represents economists with a high access statistics. In our data set, these are primarily economists with software components.

- Archetype 6 is the same as archetype 3 in the previous analysis representing the low–performer.

Going a step further by including another two archetypes provides rather small additional insights. In Figure 6 we show the barplots for the analysis with eight archetypes. It is obvious that many archetypes correspond with the ones presented in Figure 5. Archetype 5 and 7 are now quite similar. They are characterized by a lot of publications and a rather low c/p–ratio. Although the numbers differ, like the access statistics additional archetypes do not seem to be necessary.[6]

Given the aggregated analysis one is able to assign each economist to all archetypes. In practice, percentages of each archetypes are allocated to the scientists which add up to one. In Figure 7 we show a boxplot for all percentage shares for each of the six identified archetypes. It states that the majority of economists in our data set are characterized by archetype six as it represents the largest relative shares. The other archetypes have only few observations with large relative shares.

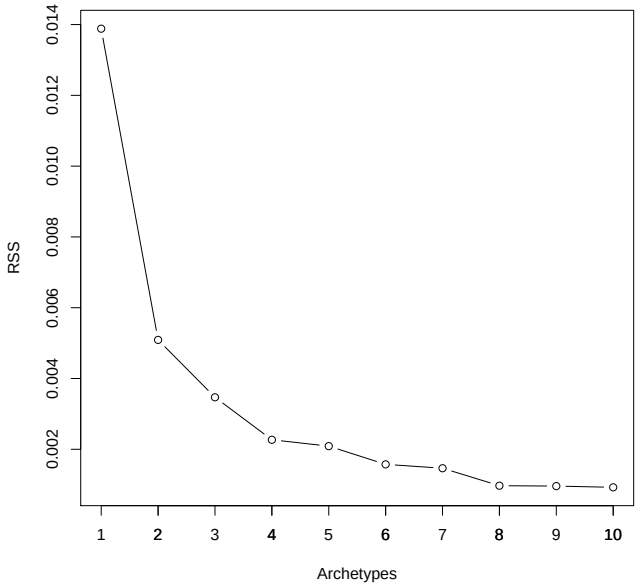

Figure 3: Screeplot for the full data set

[6]This is also confirmed by looking at higher number of archetypes not shown here. These results can be obtained from the authors upon request.

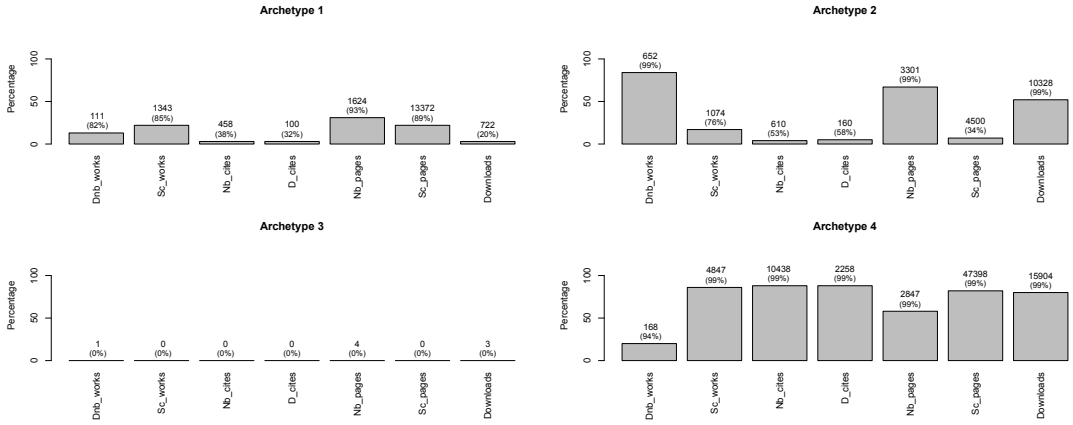

Figure 4: Barplots for 4 archetypes for the full data set

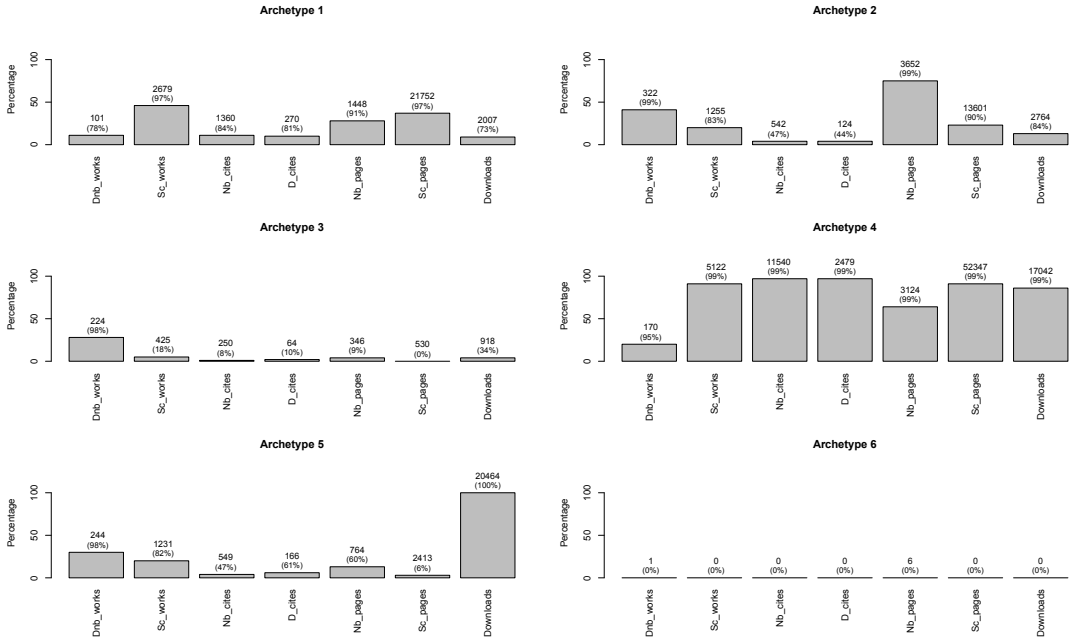

Figure 5: Barplots for 6 archetypes for the full data set

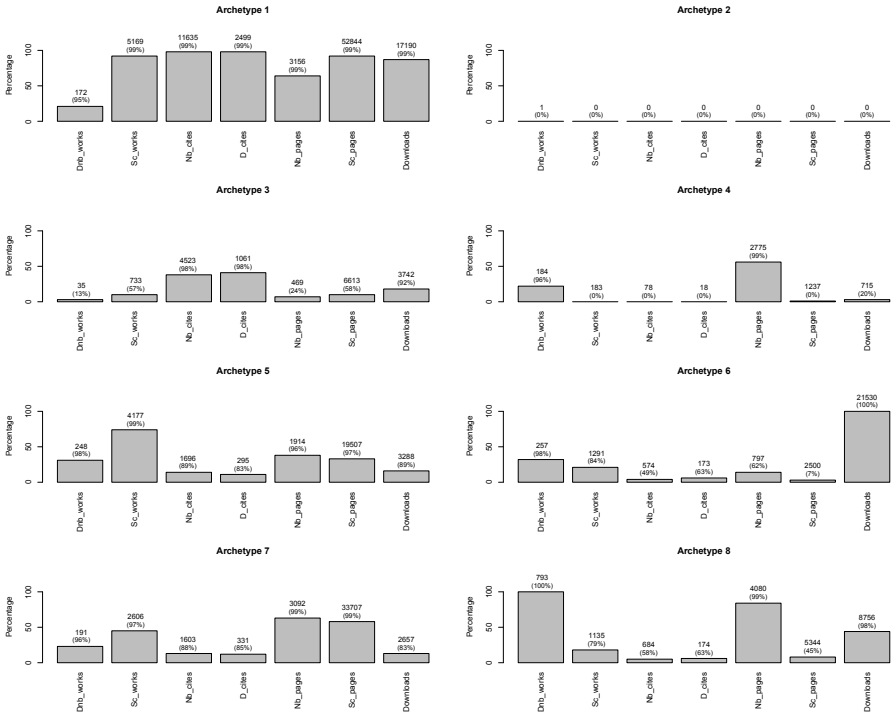

Figure 6: Barplots for 8 archetypes for the full data set

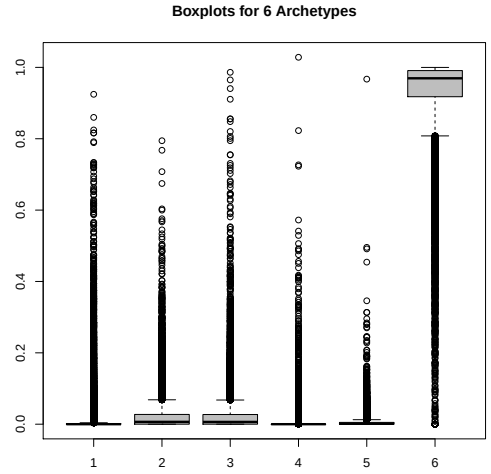

Figure 7: Distribution of relative shares across 6 archetypes

3.3 Results for the top economists

The results so far may give clear cut results, but they may be not too differentiated as the bandwidth over all variables is quite large. Furthermore, it might be interesting how scientists can be differentiated within specific groups, e.g. top scientists. Therefore we go a step further and include only the top 1,500 authors of the July 2011 ranking and run the archetypal analysis again. In Figure 8 (left panel) we plot the corresponding screeplot for a maximum of 10 archetypes. In this case no cut–off point is obvious. Therefore, it is recommended to start with a small number of archetypes. Then, each additional archetype should be checked if it provides further insights to the classification. In order to save space we report only our final selection of six archetypes. In Figure 9 we plot the corresponding barplots.

- Archetype 1 describes economists with a lot of working papers, which are published in rather high–ranked working paper series. But these have not been published in corresponding journals (yet). The c/p–ratio with roughly six is rather low.

- Economists characterized by archetype 2 are the overall top economists. Their performance values are within the highest percentiles for all variables.

- Archetype 3 is very close to the first one with similar values in almost all categories. The difference is that more journal articles are published.

- Archetype 4 represents economists with a high output level both for the number of overall works and articles in journals. But the quality–weighted journal output is rather low. Due to this high output level, also the access statistics show considerable high counts. Furthermore, the c/p–ratio below one is rather poor.

- Archetype 5 denotes economists which are located at the lower bound for each variable within the category of top economists.

- Archetype 6 is quite similar to the the second one with very high values in each category. But authors in this category have by far higher c/p–ratio (80) than economists described by archetype 2 (about 30).

In Section 2 we demonstrated that outliers may influence the location of archetypes. Seiler and Wohlrabe (2012) showed that many RePEc variables include outliers, especially the number of works category. How does our analysis change if we exclude the top values for each variable? In Figure 8 (right panel) we plot the corresponding screeplot. Although the curves are quite similar, seven archetypes might be appropriate as a flattening at this point can be observed. In order make comparisons and highlight the exclusion of outliers we plot the barplots for six archetypes in Figure 10. It is not surprising that the results differ. The relative heights of each barplot are not directly comparable as the respective maximum values changed. The following archetypes are similar in its interpretation

- Archetype 1 (Figure 10) and archetype 3 (Figure 9),
- Archetype 2 (Figure 10) and archetype 4 (Figure 9),
- Archetype 3 (Figure 10) and archetype 2 (Figure 9),
- Archetype 4 (Figure 10) and archetype 5 (Figure 9).

Although they show a similar pattern, their respective quantiles differ. Archetype 5 has no counterpart. Authors with high access statistics, i.e. many downloads, are described by this archetype. Archetype 6 denotes economists with a lot of publications of all kinds, but with comparable less journal articles. This analysis demonstrates that the interpretation of archetypes is (partially) driven by outliers. But this is not surprising as this type of analysis aims to find the extreme values, i.e. the convex hull, of a data set. Before the exclusion of outliers one has to keep in mind first that there is no formal definition of an outlier. And second, the exclusion of top scientists in each category may change the relative interpretation of the results. We prefer to leave the data set unchanged.

In Figure 11 we plot the boxplots (without the exclusion of outliers) for the distribution across all six archetypes. The dominating archetype is the fifths one, followed by the first and third one. As expected only few economists are described by more than 50% by the two top archetypes 2 and 6. Only one author is described best by the fourth archetype. This is an example how one extreme observation can shape an archetype. In Table 1 we report the top

five economists for each archetype based on the analysis with six archetypes, i.e. economists with the largest relative share for the respective archetype. A value of 1.00 denotes that the respective economists represents the identified archetype perfectly. This is the case for archetype 3 and 4, where Peter Nijkamp and Derek A. Neal hold the first position. In the other cases the identified archetypes are artificial, i.e. no economist in our sample fits this archetype to 100 percent. In these cases economists are a mixture of different types. The Nobel Prize winner Joseph E. Stiglitz is the top economist in archetype 2 which we considered as the best category. With a small gap the Nobel Prize winner James Heckman follows at the second place.

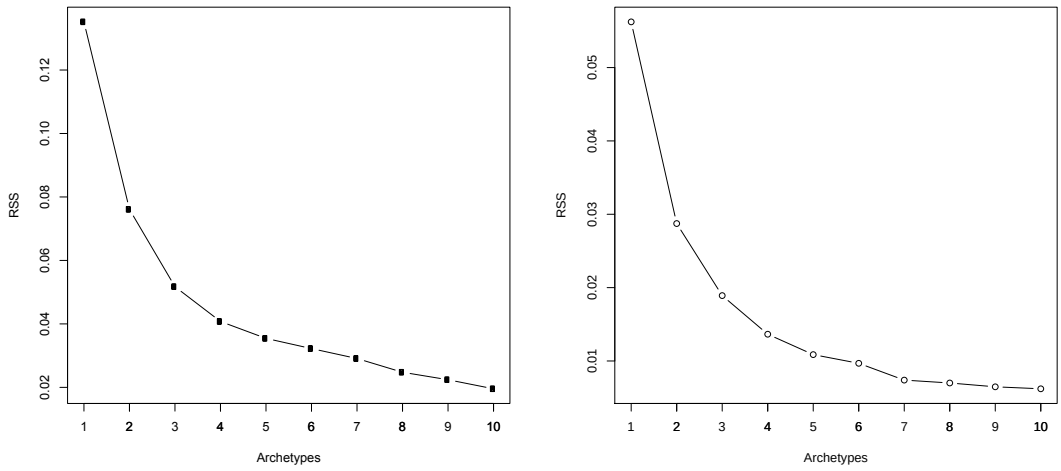

Figure 8: Screeplot for the top 1,500 economists with (left) and without (right) outliers

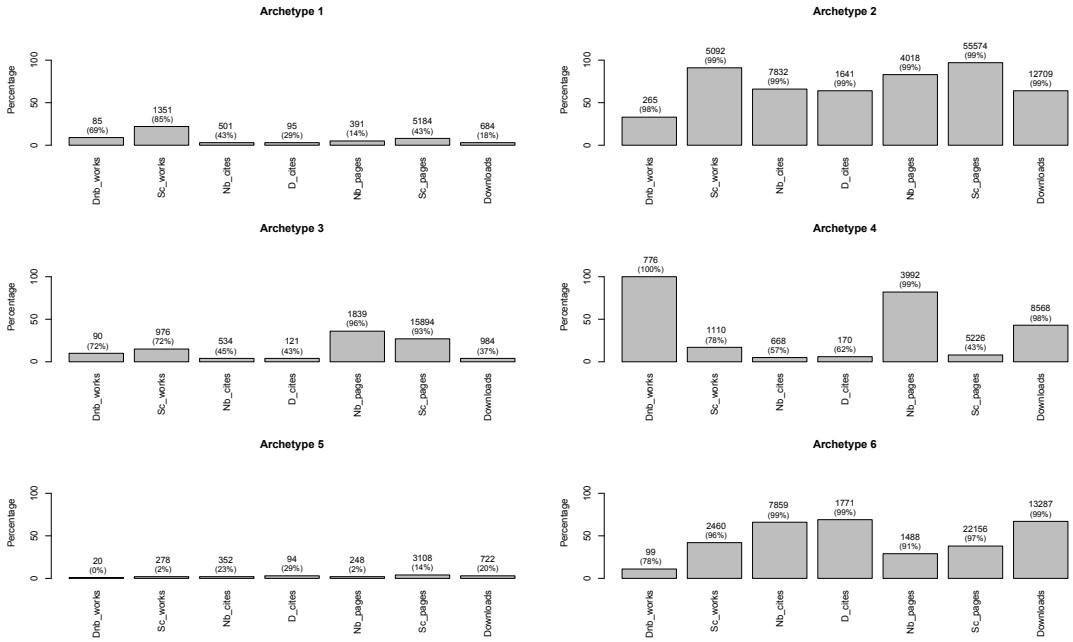

Figure 9: Barplots for 6 archetypes for the top 1,500 economists

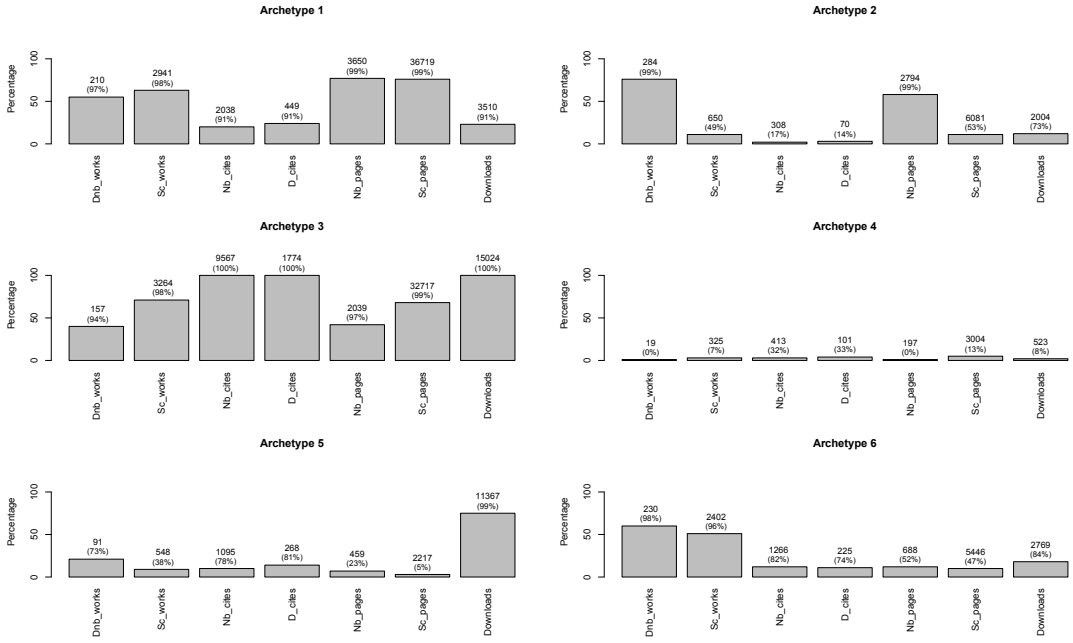

Figure 10: Barplots for 6 archetypes for the top 1,500 economists excluding outliers

Table 1: Representative economists for each archetype for the top 1,500 economists

Archetype	**1**	**2**	**3**	**4**	**5**	**6**
Roger H. Gordon	**0.94**	0.06	0.00	0.00	0.00	0.00
James R. Hines Jr.	**0.93**	0.07	0.00	0.00	0.00	0.00
Don Fullerton	**0.93**	0.03	0.00	0.04	0.00	0.00
Kala Krishna	**0.92**	0.00	0.07	0.01	0.00	0.00
James Bradford DeLong	**0.91**	0.04	0.00	0.00	0.00	0.05
Joseph E. Stiglitz	0.00	**0.93**	0.00	0.07	0.00	0.00
James J. Heckman	0.00	**0.92**	0.00	0.00	0.00	0.08
Andrei Shleifer	0.00	**0.80**	0.00	0.00	0.00	0.21
Jean Tirole	0.10	**0.75**	0.15	0.00	0.00	0.00
Daron Acemoglu	0.15	**0.75**	0.00	0.00	0.00	0.10
Anjan V. Thakor	0.00	0.00	**0.87**	0.04	0.00	0.09
Bruce D. Smith	0.00	0.15	**0.84**	0.01	0.00	0.00
Xavier Vives	0.00	0.04	**0.82**	0.11	0.00	0.03
Lung-Fei Lee	0.00	0.03	**0.78**	0.00	0.19	0.00
David E. M. Sappington	0.05	0.01	**0.77**	0.02	0.14	0.01
Peter Nijkamp	0.00	0.00	0.00	**1.00**	0.00	0.00
Piet Rietveld	0.00	0.00	0.20	**0.34**	0.47	0.00
Richard S.J. Tol	0.00	0.00	0.00	**0.34**	0.60	0.06
Bruno S. Frey	0.06	0.00	0.39	**0.31**	0.00	0.23
John Creedy	0.00	0.00	0.58	**0.31**	0.11	0.00
Derek A. Neal	0.00	0.00	0.00	0.00	**1.00**	0.00
Thomas Laubach	0.00	0.00	0.00	0.00	**0.99**	0.01
Avner Greif	0.00	0.00	0.00	0.00	**0.99**	0.01
Brent R. Moulton	0.00	0.00	0.00	0.00	**0.99**	0.01
Luca Benati	0.01	0.00	0.00	0.00	**0.99**	0.00
Mark L. Gertler	0.00	0.00	0.00	0.00	0.11	**0.89**
Ross Levine	0.00	0.00	0.00	0.08	0.09	**0.82**
Kenneth S. Rogoff	0.12	0.07	0.00	0.06	0.00	**0.74**
Robert W. Vishny	0.00	0.00	0.00	0.00	0.27	**0.73**
Robert J. Barro	0.00	0.28	0.00	0.00	0.00	**0.71**

This table reports economists representing the top five maximum relative shares for each archetype. The results are based on an archetypal analysis with $k = 6$ based on 7 indicators for the top 1,500 economists.

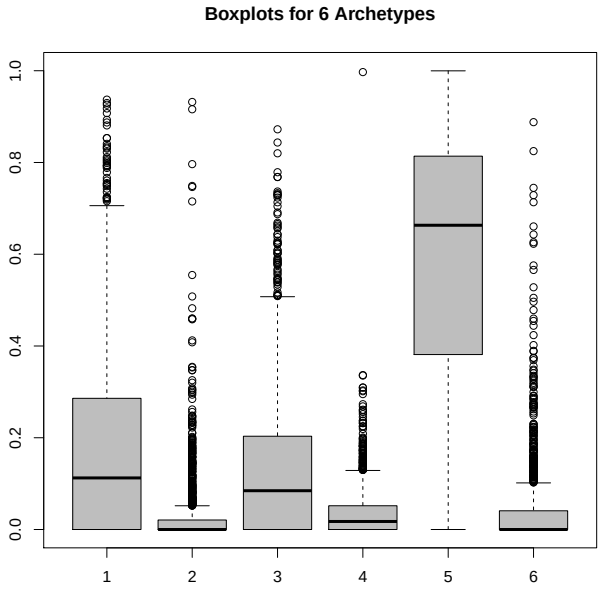

Figure 11: Distribution across 6 archetypes for the top 1,500 economists

4 Summary

In this paper we introduced archetypal analysis to bibliometric research. It allows to extract typical characteristics (archetypes) within a multivariate data set. We illustrated the approach with a large data set with almost 30,000 economists. For each economist we have seven bibliometric scores, spanning over various measures of (quality–weighted) number of published work, citations and access statistics. There is no formal criterion to judge the number of included archetypes. The screeplot gives a first hint how many archetypes are present in the data. Therefore, this choice is mainly driven by an analyis of the context. We identified six archetypes both for the full sample and a reduced sample of top economists. These are mainly characterized by different relationships of published work and citations. We demonstrated how the analysis and interpretation of archetypes can be driven by outliers but advocate to leave the data set unchanged. Future research could include further characteristics of science, as teaching, press relations, acquisition of grants, supervision of students, among others.

References

Chan, B., D. Mitchell, and L. Cram (2003): "Archetypal analysis of galaxy spectra," *Monthly Notices of the Royal Astronomical Society*, 338(3), 790–795.

Costas, R., and M. Bordons (2007): "The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level," *Journal of Informetrics*, 1(3), 193–203.

Cutler, A., and L. Breiman (1994): "Archetypal Analysis," *Technometrics*, 36(4), 338–347.

Docampo, D. (2011): "On using the Shanghai ranking to assess the research performance of university systems," *Scientometrics*, 86(1), 77–92.

Eugster, M. (2012): "Performance Profiles based on Archetypal Athletes," *International Journal of Performance Analysis in Sport*, 12(1), 166–187.

Eugster, M., and F. Leisch (2009): "From Spider-Man to Hero–Archetypal Analysis in R," *Journal of Statistical Software*, 30(8), 1–23.

——— (2011): "Weighted and robust archetypal analysis," *Computational Statistics & Data Analysis*, 55(3), 1215–1225.

Eugster, M. J. A., and F. Leisch (2012): *archetypes: Archetypal Analysis* –R-Package, Version 2.1-0.

Franceschet, M. (2009): "A cluster analysis of scholar and journal bibliometric indicators," *Journal of the American Society for Information Science and Technology*, 60(10), 1950–1964.

Halkos, G., and N. Tzeremes (2011): "Measuring economic journals citation efficiency: A data envelopment analysis approach," *Scientometrics*, 88(3), 979–1001.

Neff, M., and E. Corley (2009): "35 years and 160,000 articles: A bibliometric exploration of the evolution of ecology," *Scientometrics*, 80(3), 657–682.

Ortega, J., E. Lopez-Romero, and I. Fernandez (2011): "Multivariate approach to classify research institutes according to their outputs: The case of the CSIC's institutes," *Journal of Informetrics*, 5(3), 323–332.

Osareh, F. (2012): "The use and application of multivariate analysis techniques in bibliometric and scientometric studies," *International Journal of Information Science and Management*, 1(2), 59–71.

Porzio, G., G. Ragozini, and D. Vistocco (2008): "On the use of archetypes as benchmarks," *Applied Stochastic Models in Business and Industry*, 24(5), 419–437.

Ramesh Babu, A., and Y. Singh (1998): "Determinants of research productivity," *Scientometrics*, 43(3), 309–329.

Seiler, C., and K. Wohlrabe (2012): "Ranking economists on the basis of many indicators: An alternative approach using RePEc data," *Journal of Informetrics*, 6(2), 389–402.

Zhang, L., X. Liu, F. Janssens, L. Liang, and W. Glänzel (2010): "Subject clustering analysis based on ISI category classification," *Journal of Informetrics*, 4(2), 185–193.