Pattern-analysis of financial strategies of companies

S.A.Dzuba[1] D.V.Krylov[2]

[1,2]Far Eastern Federal University

XXI April International Academic Conference On Economic And Social Development

Outline

1. Clusterization Preparation
   - Short History Behind The Problem
   - Data Preparation
2. Clusterization
   - Choosing The Right k (number of clusters)
   - Clusterization Results ($k = 8$)

Outline

1. Clusterization Preparation
   - Short History Behind The Problem
   - Data Preparation
2. Clusterization
   - Choosing The Right k (number of clusters)
   - Clusterization Results ($k = 8$)

What Was Presented At The Last Conference

- The purpose of the previous work was to define a menu of companies' strategies and financial indicators that indicate them through clustering
- Number of observations was 1750. They were collected on the basis of Forbes Global 2000 listing
- The main problem we encountered was that the results of pam clustering were difficult to interpret. Also, visualization of the results through a set of scattering diagrams was too complex and confusing
- At the conference we were told to use the pattern-analysis method to solve these problems, **so we did**

Outline

1. Clusterization Preparation
   - Short History Behind The Problem
   - Data Preparation
2. Clusterization
   - Choosing The Right k (number of clusters)
   - Clusterization Results ($k = 8$)

Data

- We have a data set that consists of 2030 companies from the Forbes Global 2000 listing as of 2006-2018
- Total number of observations – 26390
- We are preparing the data in four steps

Steps

1. Missing data removal
2. Removal of outliers using a boxplot and violinplot
3. Data standartization with Z-Score
4. Removal of outliers using clustering methods with supervised pattern recognition

Parameters

(P) $EBIT\ Ratio = \frac{Operating\ Income}{Sales}$

(R) $Return\ on\ Capital\ Employed = \frac{Net\ Income + Interest\ Expenses}{S\&L + Equity}$

(L) $Leverage = \frac{S\&L\ Liabilities}{S\&L\ Liabilities + Equity}$

(G) $Goodwill\ Ratio = \frac{Goodwill}{Goodwill + Equity}$

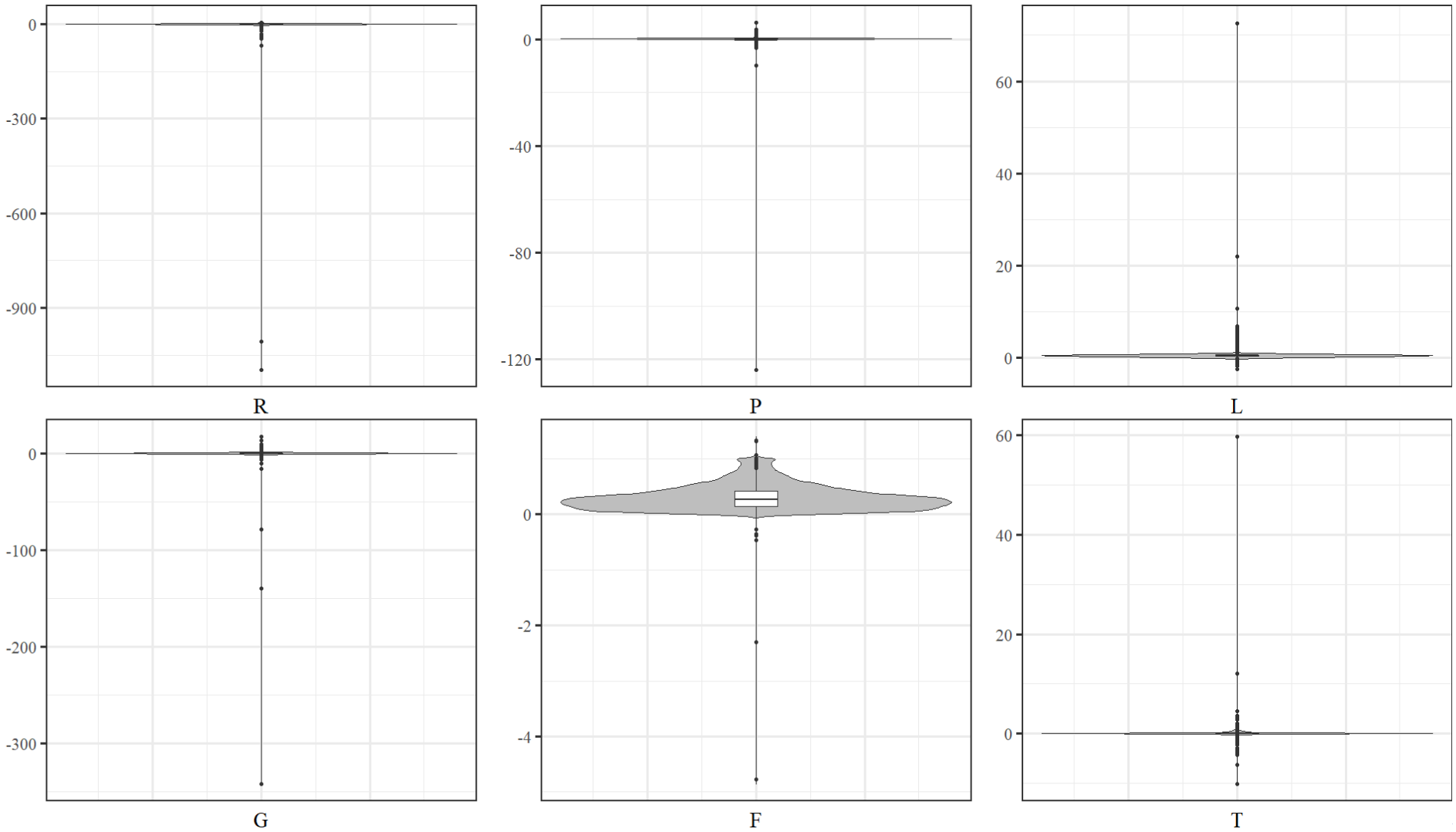
(F) $Financial\ Assets\ Ratio = 1 - \frac{Gross\ Fixed\ Assets - Accumulated\ Depreciation + Goodwill}{Non - Current\ Assets}$

(T) $Treasure\ Ratio = \frac{-Treasure\ stock}{Equity - Treasure\ stock}$

EBIT – Earnings Before Interest and Taxes
S&L Liabilities – Short-term Interest Bearing and Long–term Liabilities

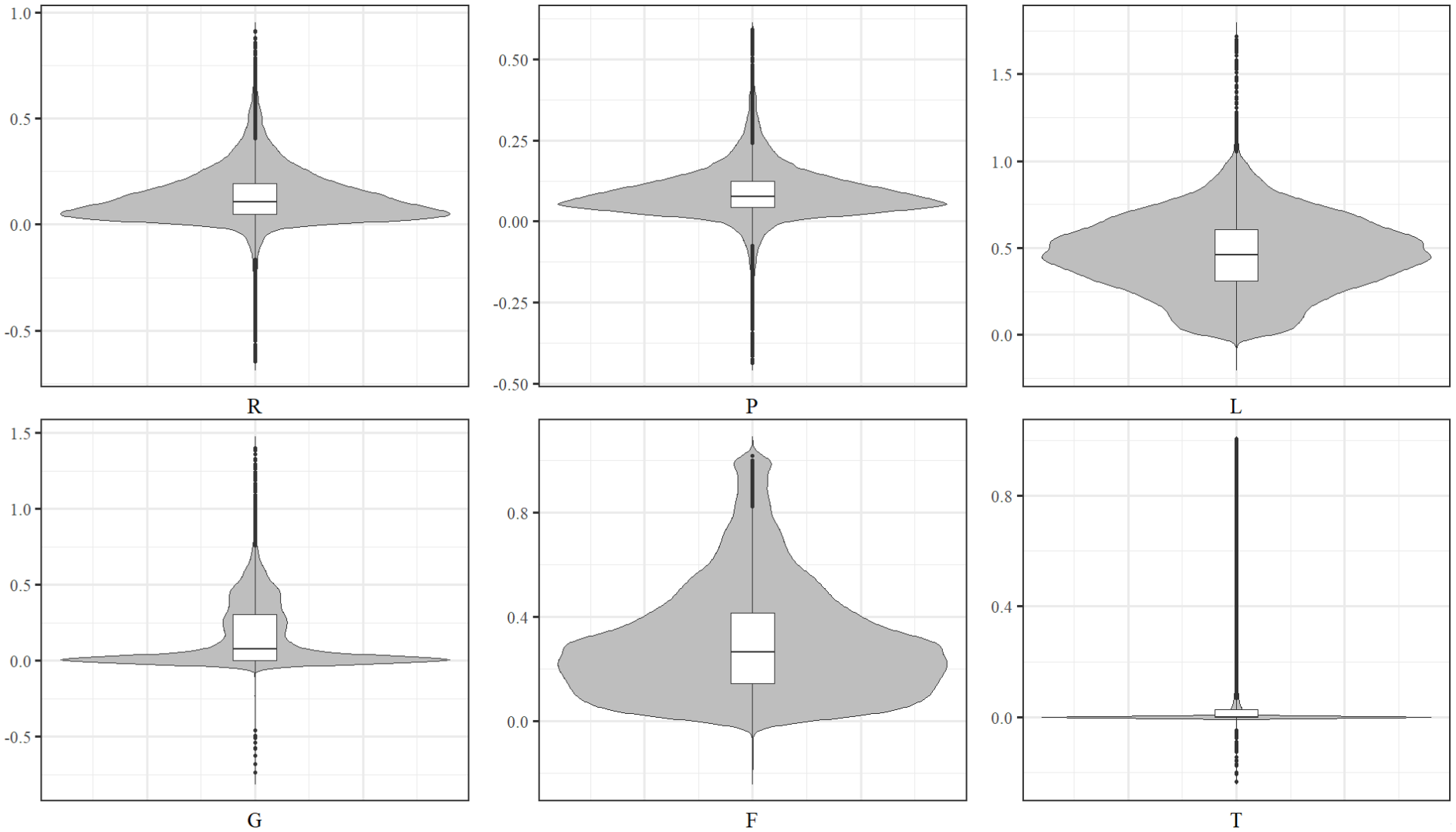
Removal Of Outliers Using A Boxplot And Violinplot

0 iteration



Removal Of Outliers Using A Boxplot And Violinplot

6 iteration, №Out = 2282

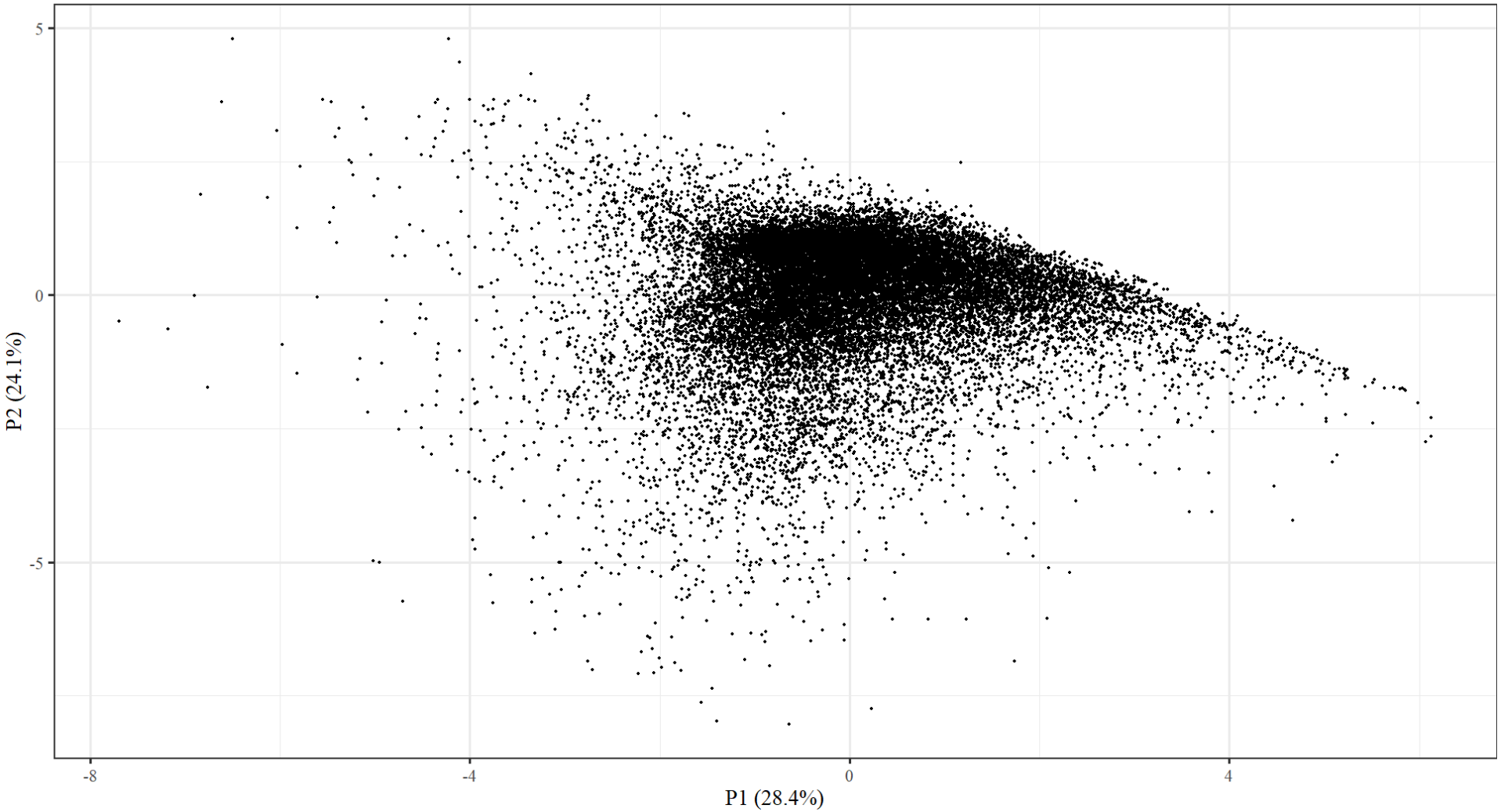


Data Standartization With Z-Score

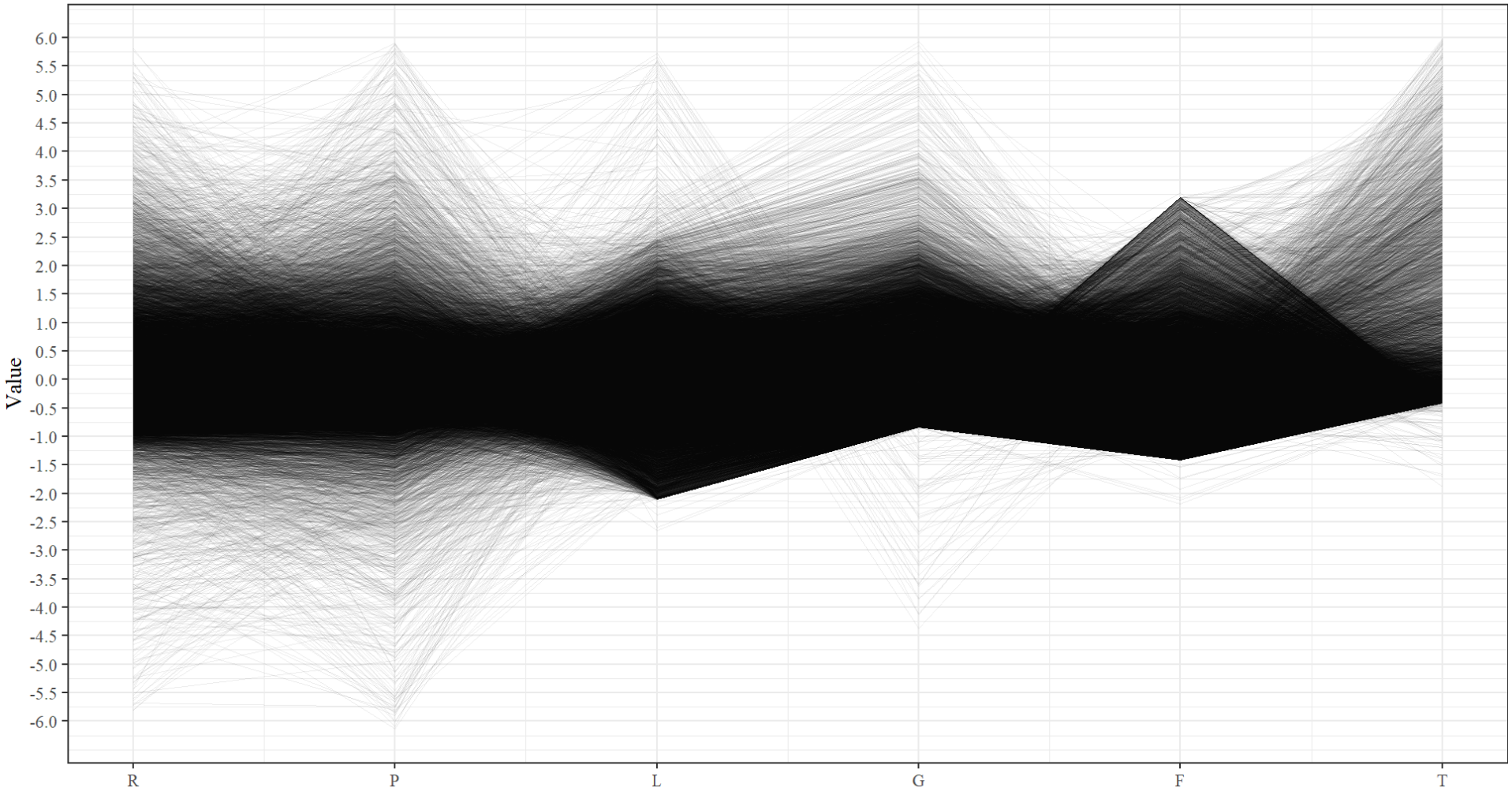
$$z_i^j = \frac{x_i^j - \mu^j}{\sigma^j}$$

- x_i^j is value of i company of j parameter
- $\mu^j$is average of j parameter vector
- $\sigma^j$is standard deviation of j parameter vector
- $i \in 1, 2, 3, ..., 23560, j \in 1, 2, 3, ..., 6$

Data In Principal Components Plot



Data In Parallel Coordinates



Removal Of Outliers Using Clustering Methods

- In this approach, we specifically specify a very big number of clusters. In this case, we took $k = 400$ (there's no formal rule for a certain number) (Loureiro et al., 2004)
- The classic algorithm of this approach defines a cluster an outlier if its size is below a certain critical value (e.g. 10)
- The main feature of this method is that observations with normal values of parameters tend to merge into large clusters even at $k = 400$

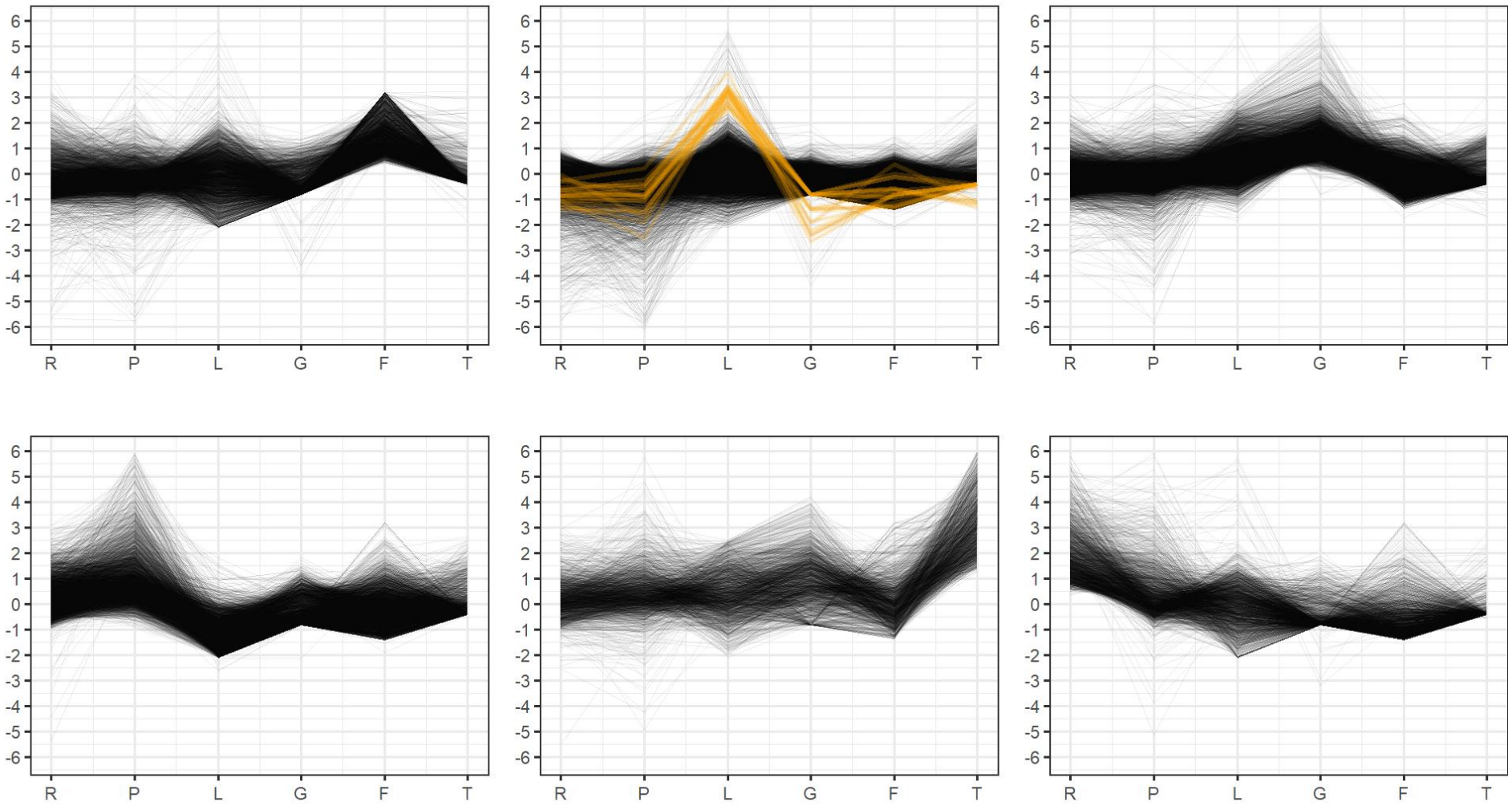
Removal Of Outliers Using Clustering Methods

- Unfortunately, this particular algorithm did not work well on our data
- Mainly due to the fact that individual companies were forming whole clusters, using data of their indicators for each year of the analyzed period: 2006-2018
- What we did to solve this problem is that we mixed this technique with a visual pattern-analysis method

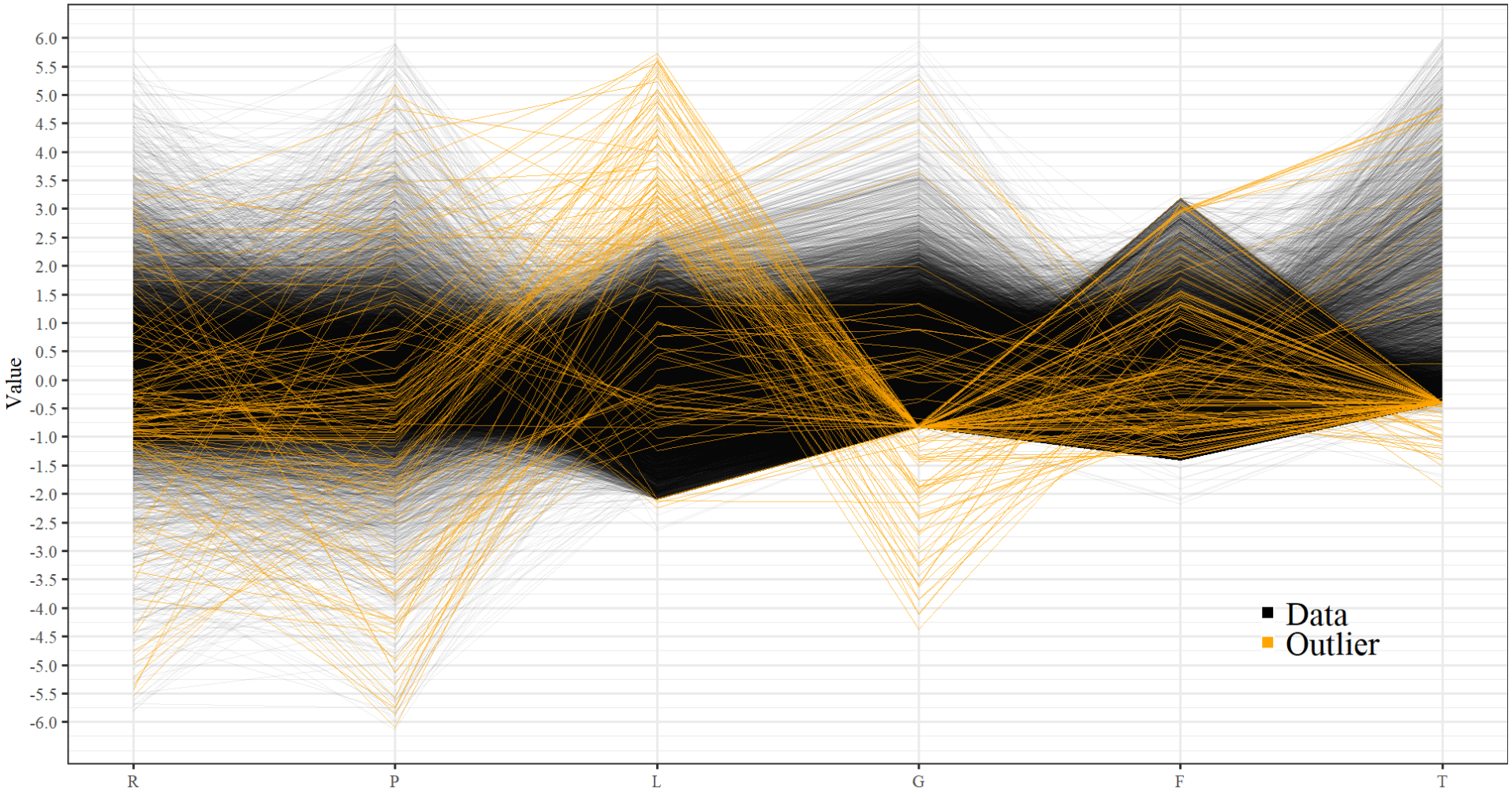
Removal Of Outliers Using Clustering Methods With Supervised Pattern Recognition

- In this combined method we perform an additional clustering procedure for the small k, we took $k = 6$ (the explanation on why we chose this particular number will be a little further)
- Then we draw all clustering results on parallel coordinates
- After that we visually compare each pattern from the results of $k = 400$ with the main six patterns
- If a pattern from the results of $k = 400$ is not similar in structure to any of the basic patterns, it is declared to be an outlier

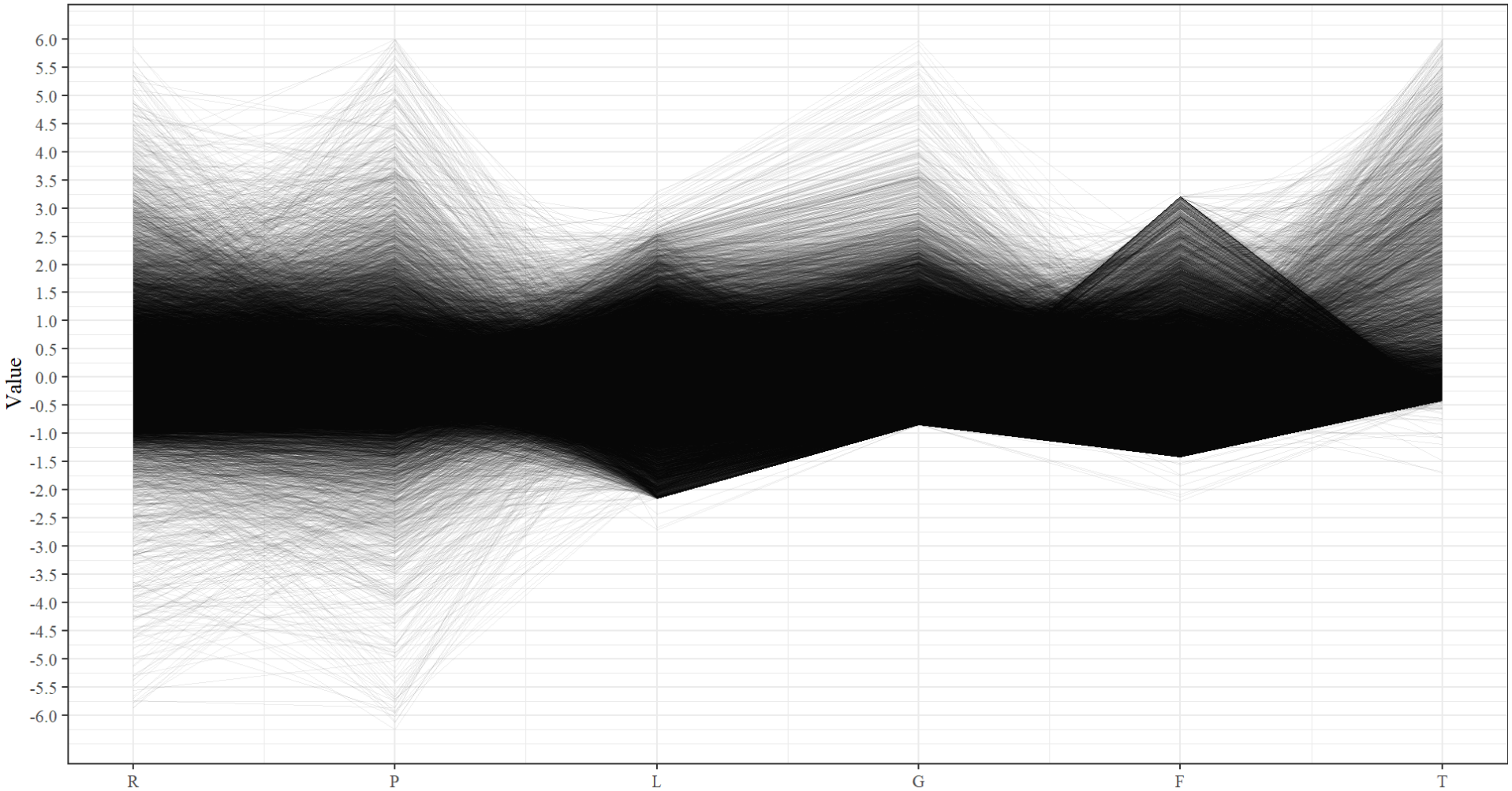
Example Of Determining An Outlier Using Described Method



All Found Outliers, №Out = 135



Data After Cleaning



Outline

1. Clusterization Preparation
   - Short History Behind The Problem
   - Data Preparation
2. Clusterization
   - Choosing The Right k (number of clusters)
   - Clusterization Results ($k = 8$)

Validation Metric

Silhouette Width (Peter J. Rousseeuw, 1987)

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}$$

where

$$a(i) = average(dist(i,j)), i,j \in C_i, i \neq j$$

$$b(i) = \min_{k \neq i}(average(dist(i,j))), i \in C_i, j \in C_k$$

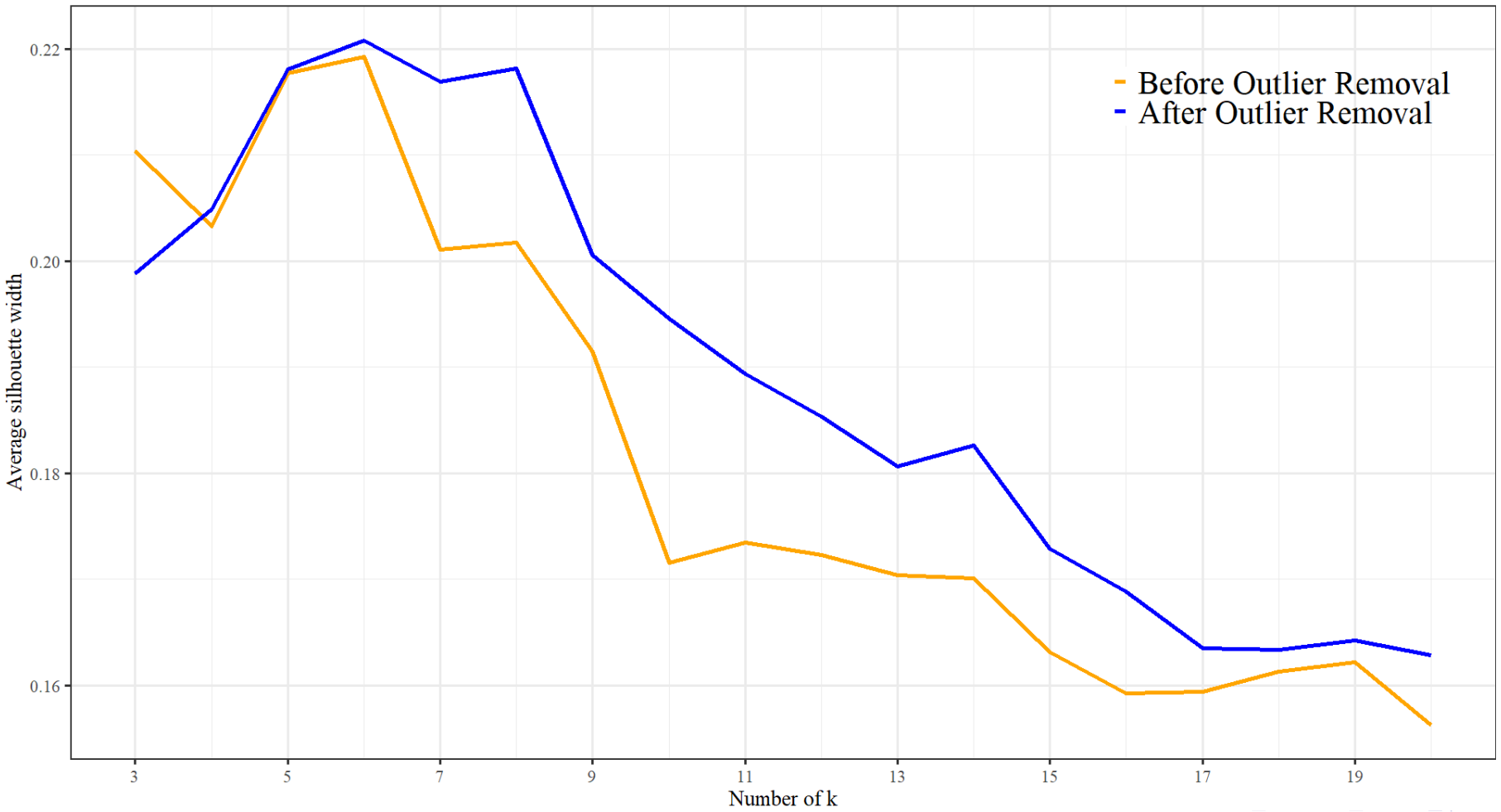
where C is an index-matched cluster

Distance Metric And Clustering Algorithm

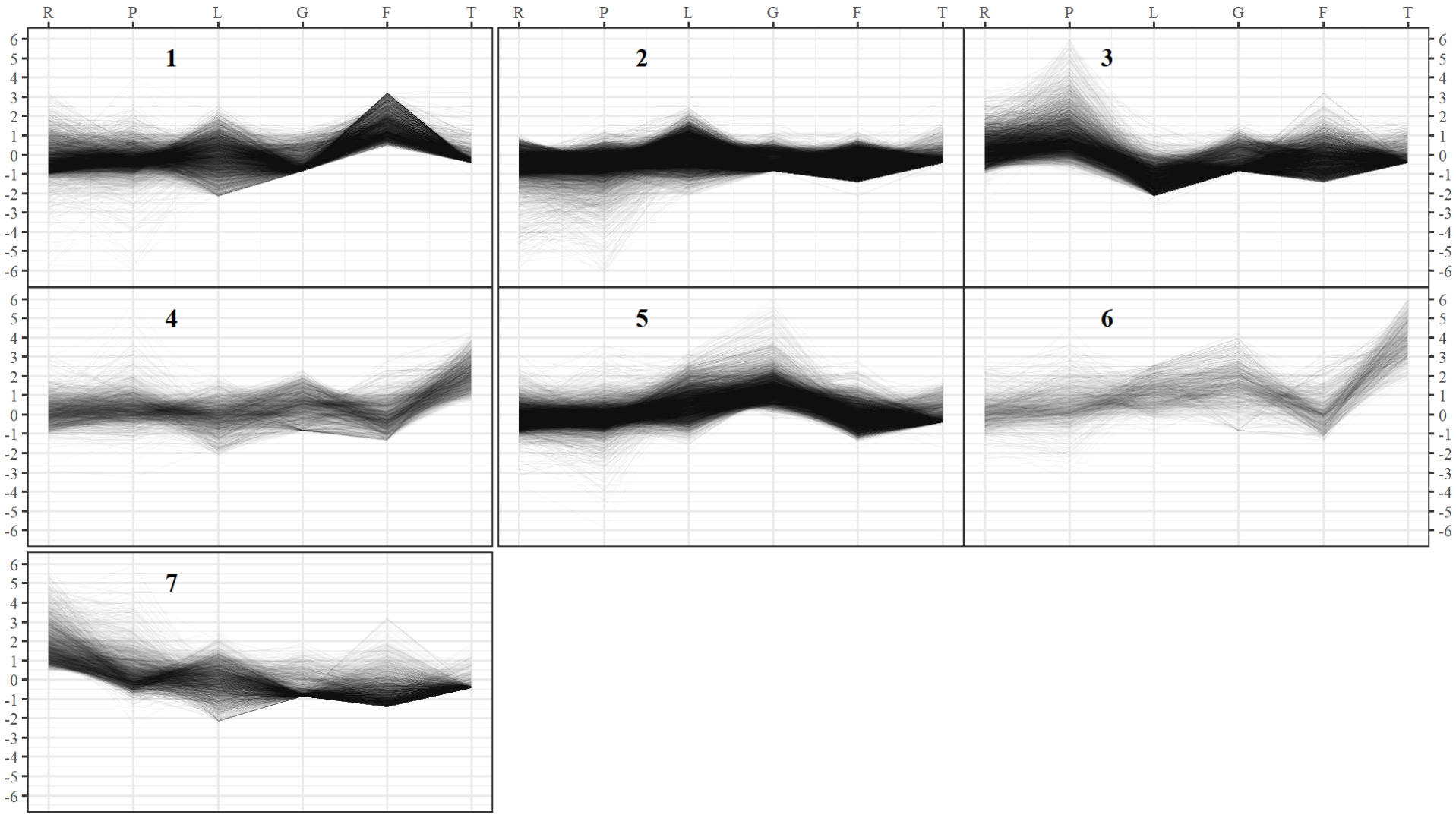
- We tried 2 distance metrics: manhattan and euclidean
- We also tested several clustering algorithms on our data: pam, hyerarchycal, k-means, hk-means (hybrid of k-means and hyerarchycal clustering)
- We stopped on a combination of the Euclidean metric and hk-means algorithm
- The final choice was determined by two factors: the average value of the Si metric and visual analysis of the resulting patterns

Average Silhouette Width Plot

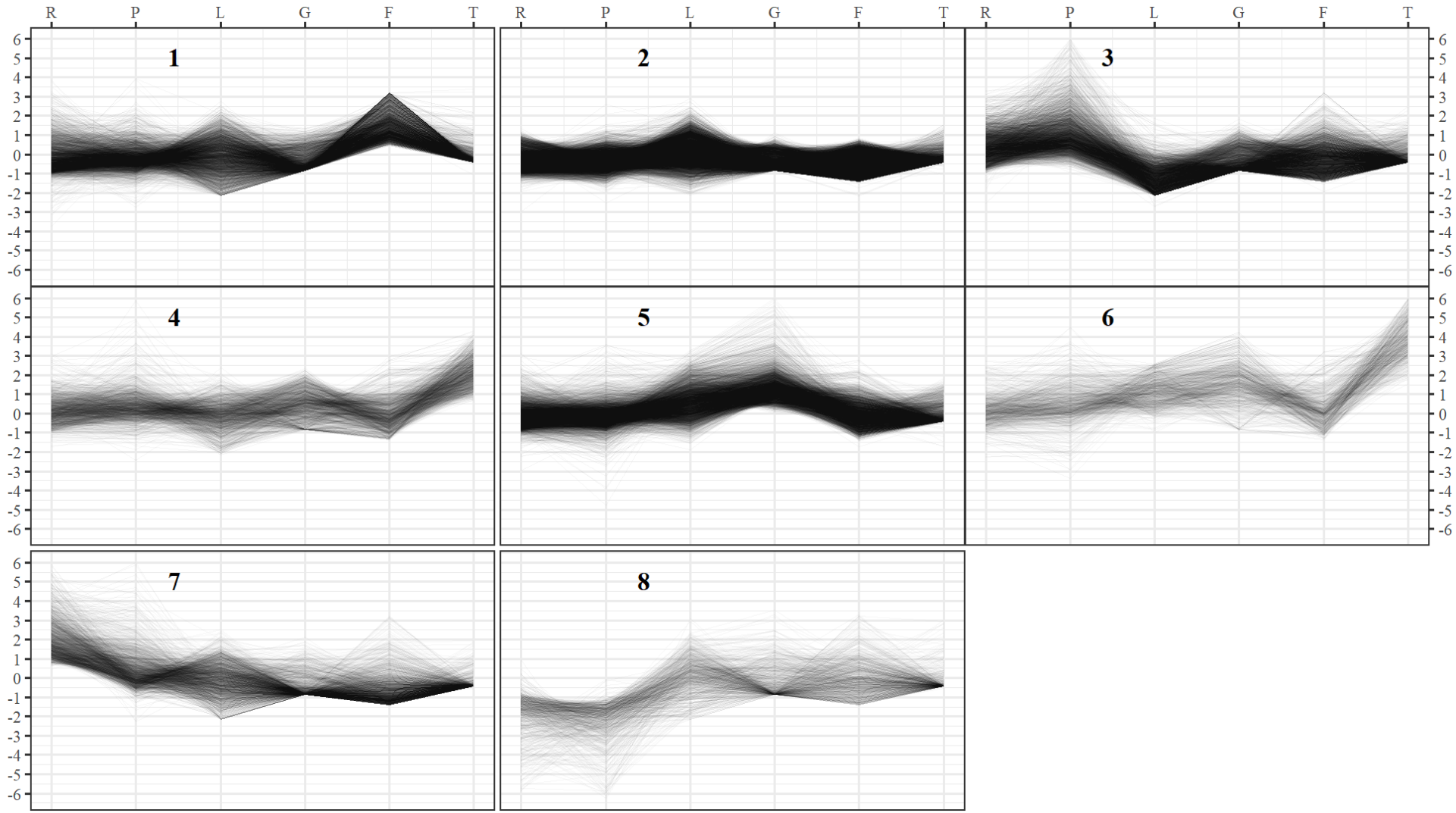
Determination of the best k, demonstration of the effect of outliers removal using clustering methods on hk-means results



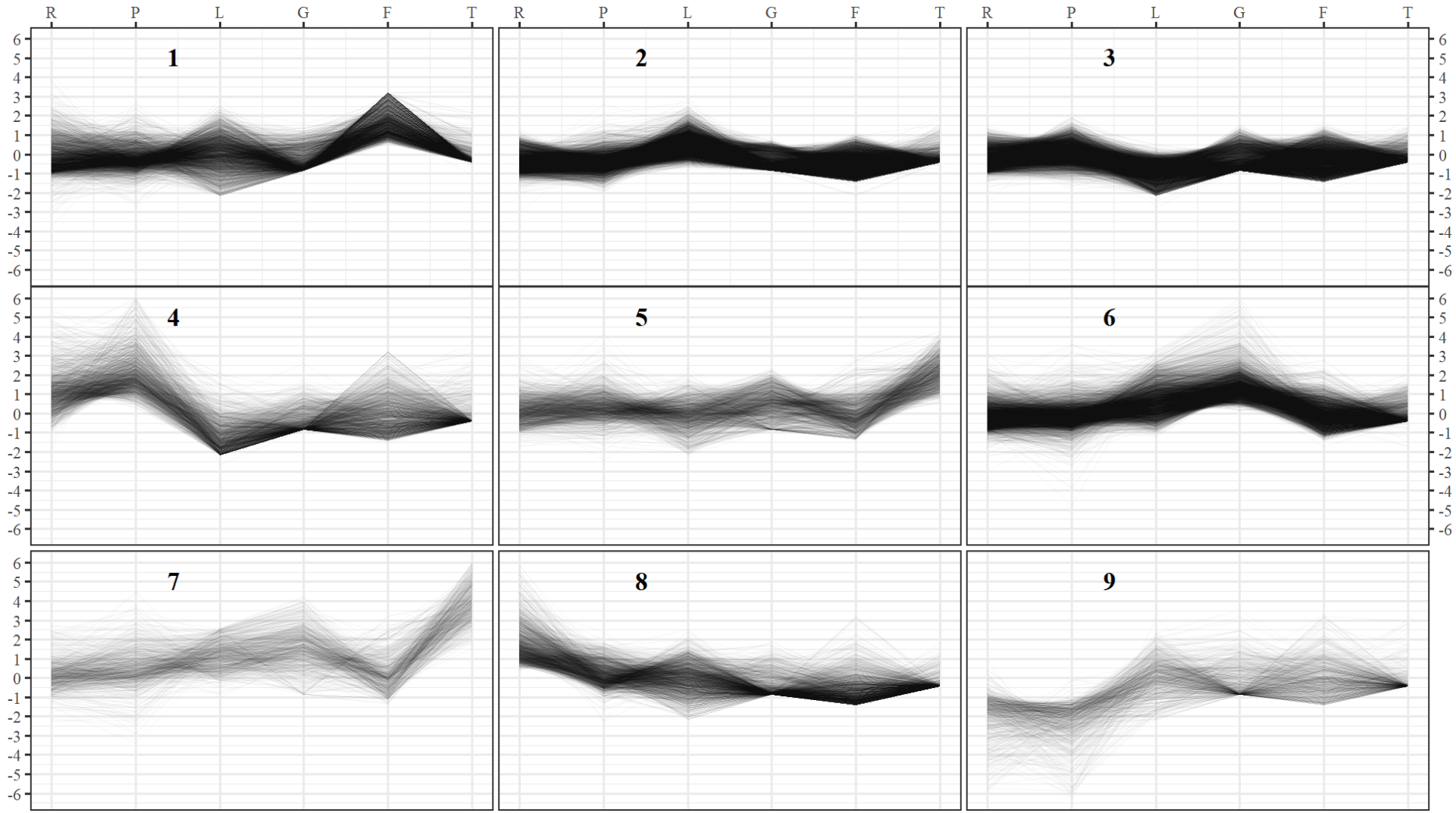
Finding The Best k, k = 7



Finding The Best k, k = 8



Finding The Best k, k = 9

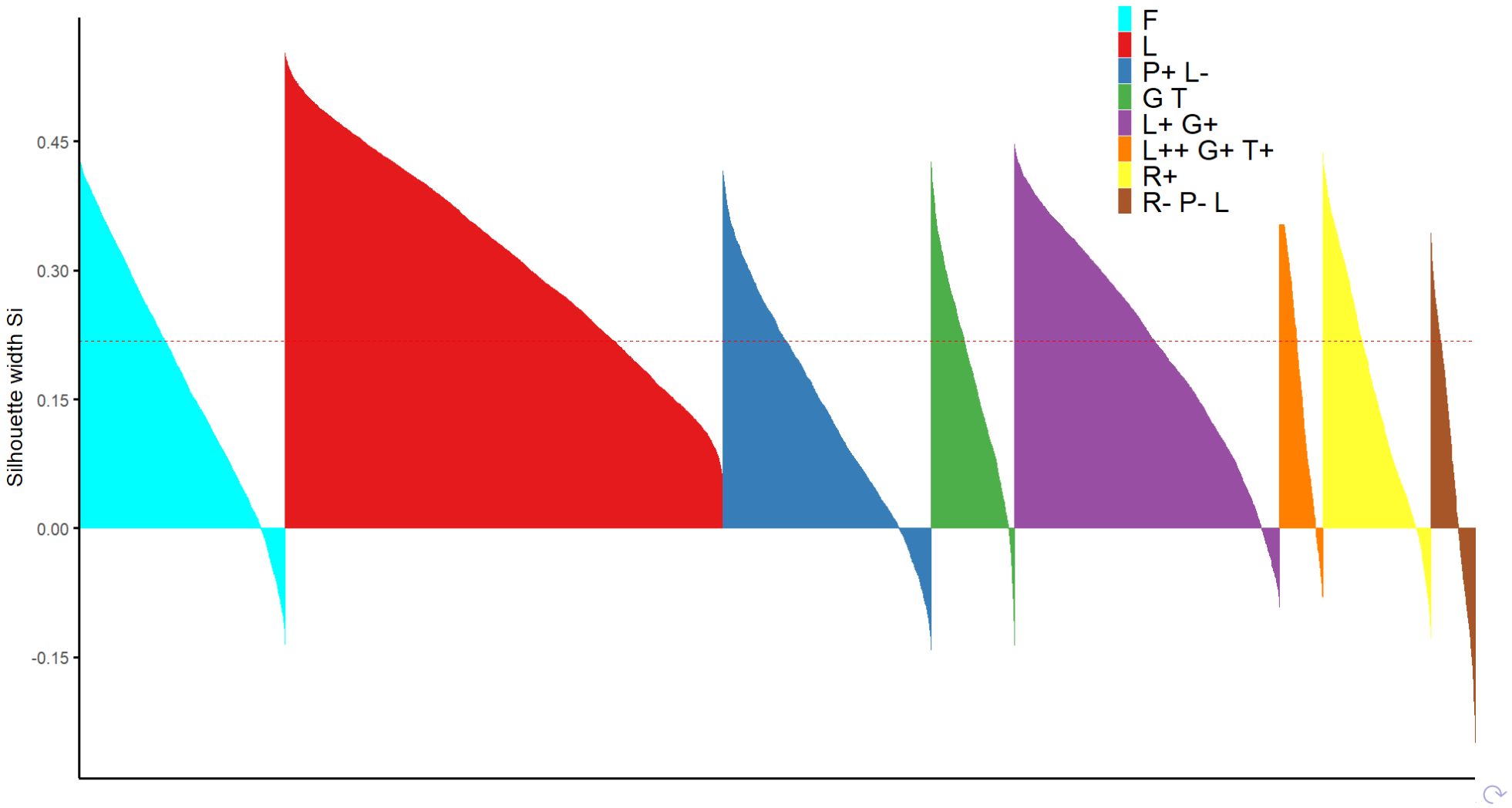


Outline

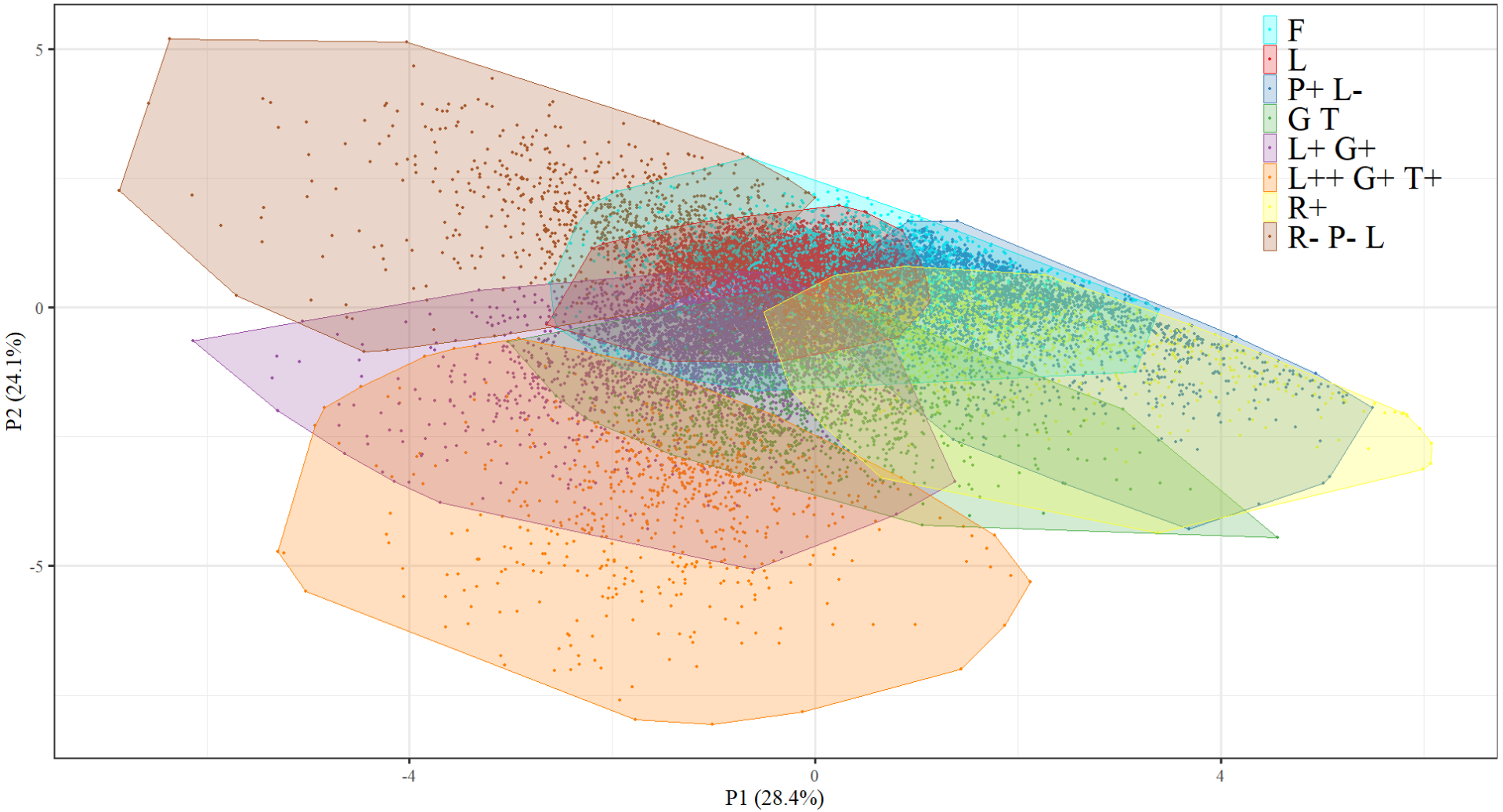
1. Clusterization Preparation
   - Short History Behind The Problem
   - Data Preparation
2. Clusterization
   - Choosing The Right k (number of clusters)
   - Clusterization Results ($k = 8$)

Silhouette Width Plot

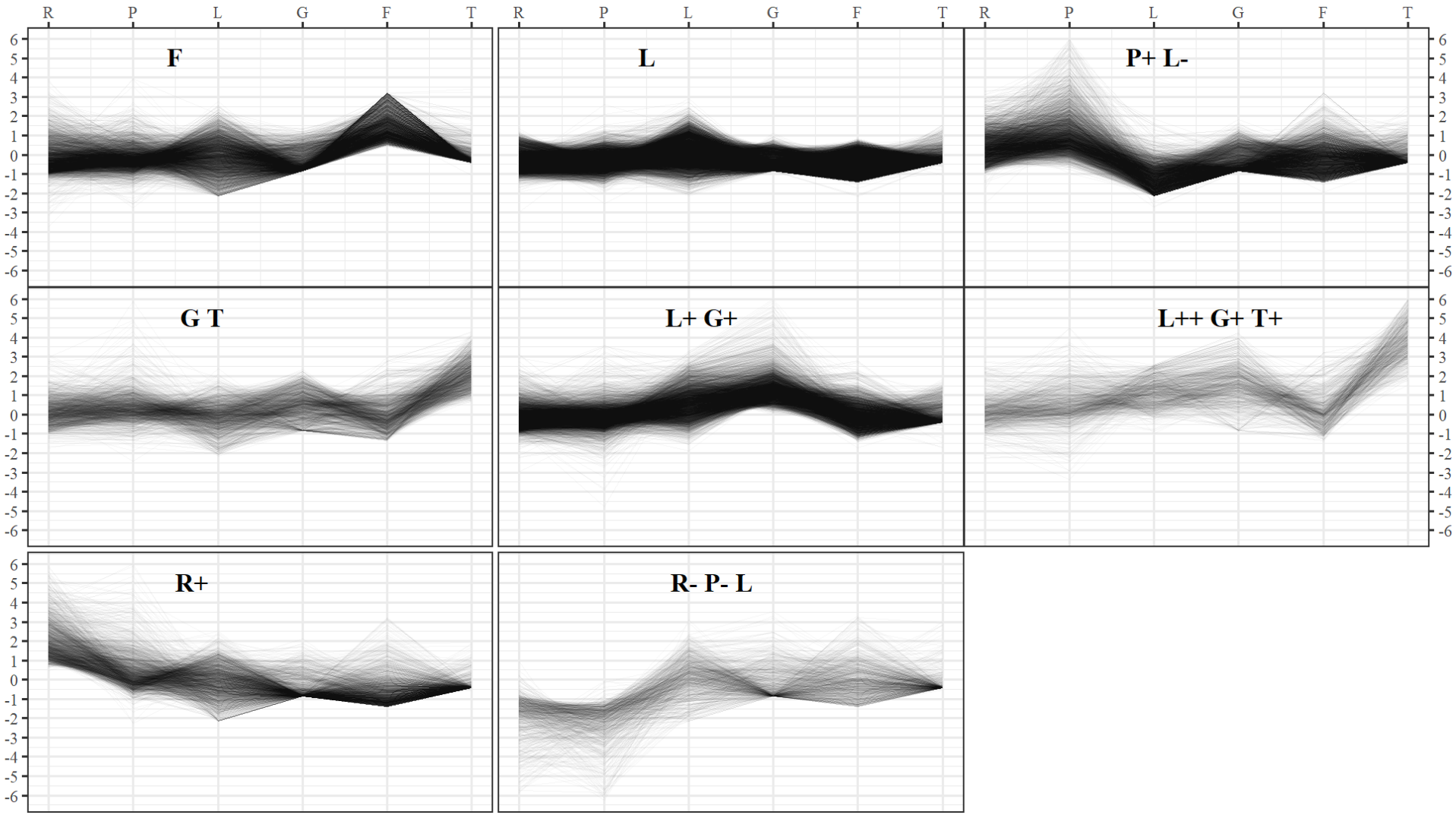
Average Silhouette width = 0.218



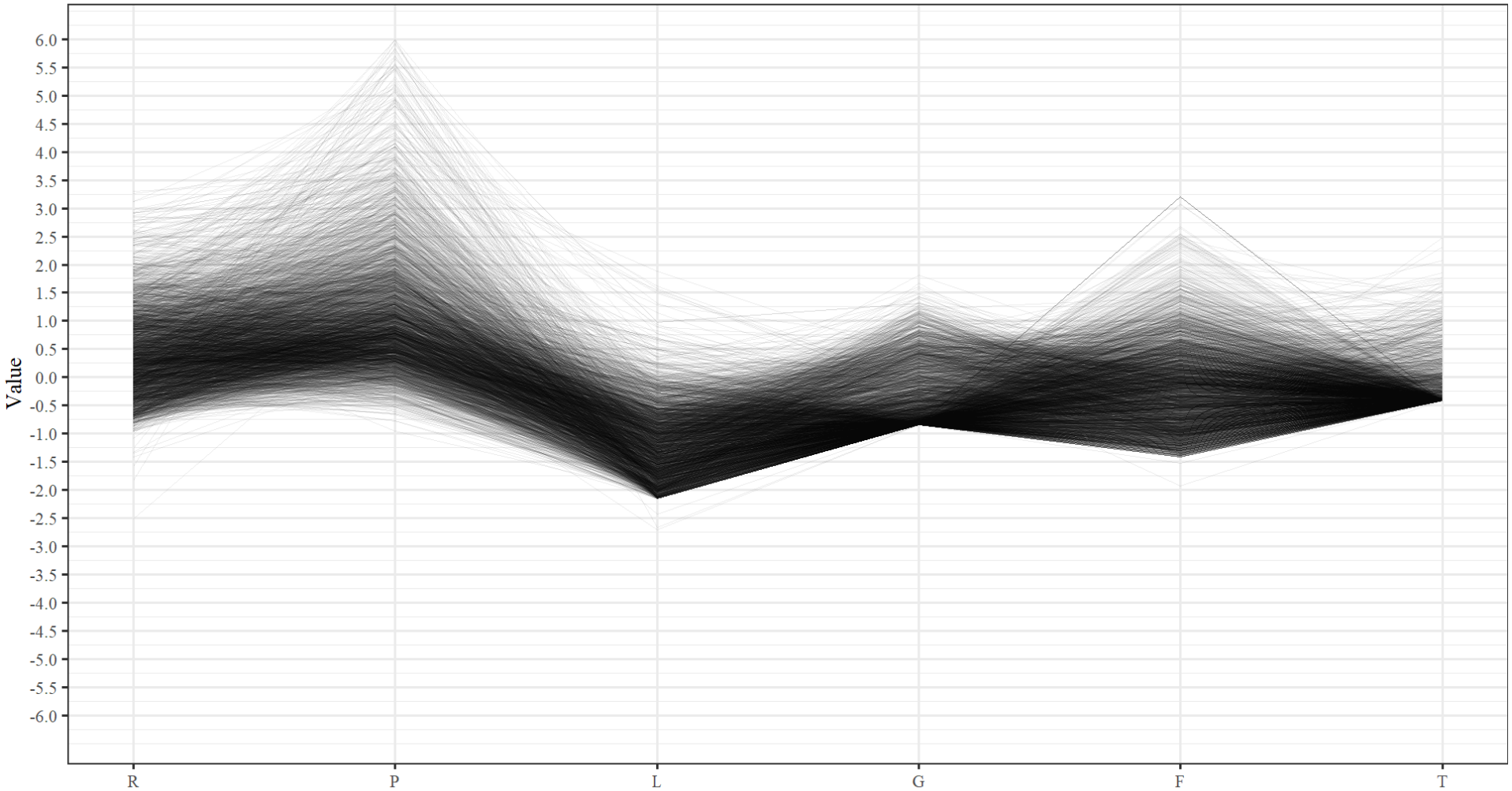
Principal Components Plot



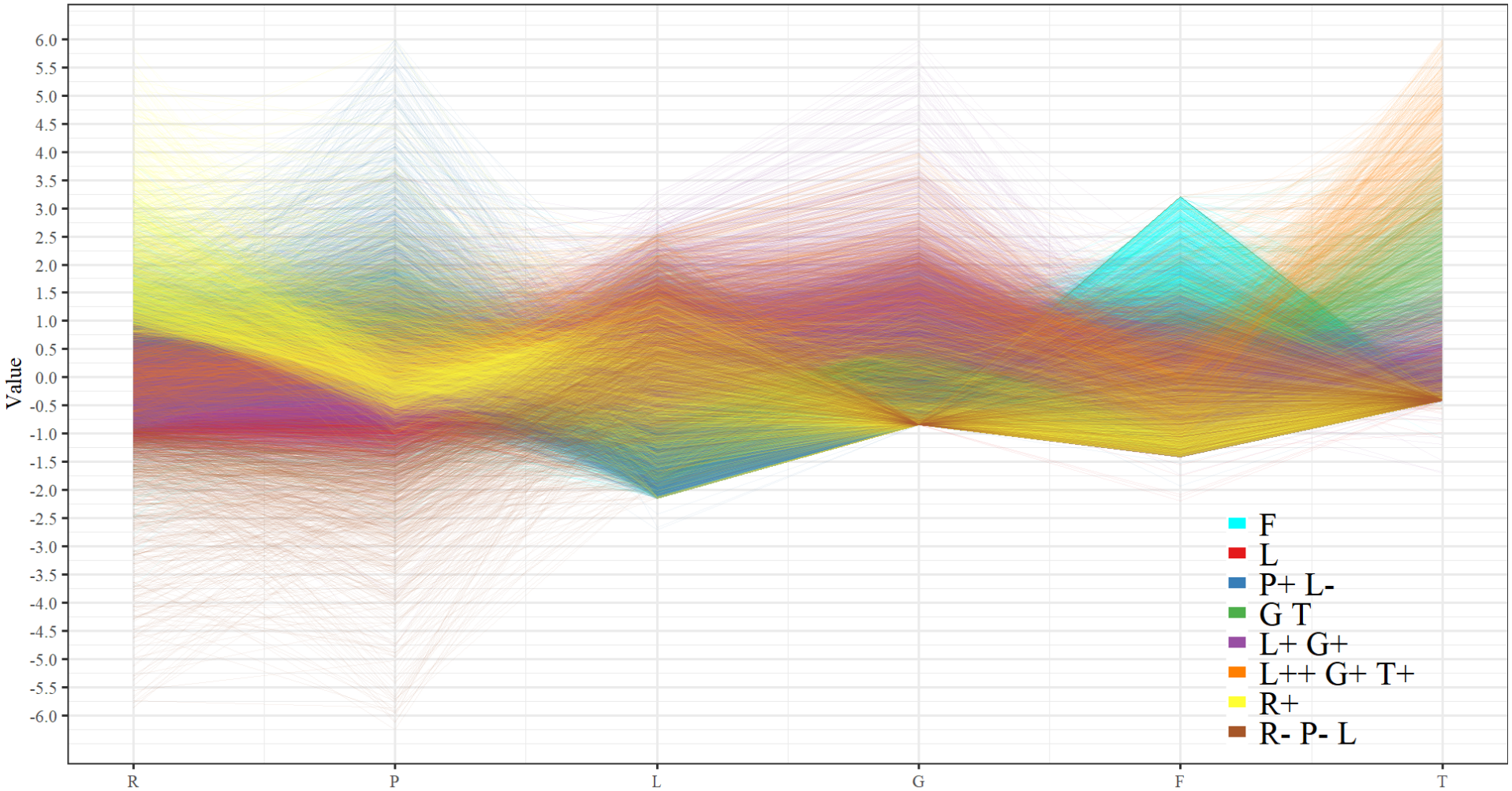
Parallel Coordinates Plot



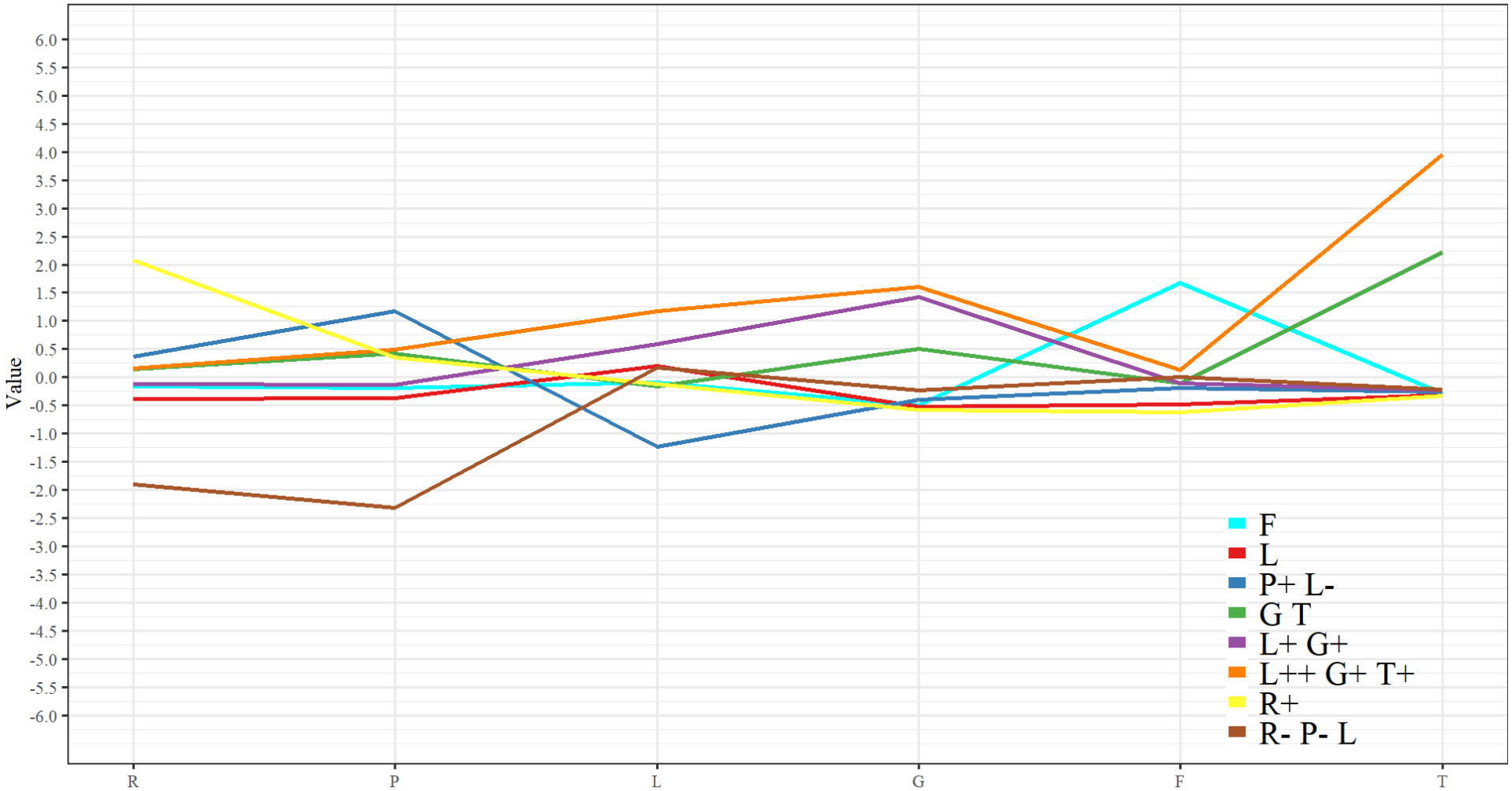
We Can Take A Closer Look At Any Pattern



We Can Also Look At Them All At Once



And We Can Analyze Their Average Values



Key Findings

1. We presented an modified method for identifying structural outliers using overclusterization, which is a mix of supervised and unsupervised clustering
2. The pattern analysis helps to make sure that the final value of k is the best by visually comparing the clustering results at k with $k-1$ and $k+1$
3. If in $k-1$ clustering the structurally unique pattern disappears from the results, and at $k+1$ a new structurally unique pattern does not appear, then the previously determined value of k is most likely the best one

References


Antonio Alfredo Ferreira Loureiro, Luís Torgo, and Carlos Soares.
Outlier detection using clustering methods: a data cleaning application.
In *Proceedings of KDNet Symposium on Knowledge-based systems for the Public Sector*, 2004.


Peter J. Rousseeuw.
Silhouettes: A graphical aid to the interpretation and validation of cluster analysis.
Journal of Computational and Applied Mathematics, 20:53 – 65, 1987.