P20: Clustering methods for selecting representative samples in chemical databases

Felipe V. Calderan[1], Johnatan Mucelini[2], Juarez L. F. da Silva[2], & Marcos G. Quiles[1]

[1]Federal University of São Paulo (Unifesp)- emails: {fvcalderan, quiles}@gmail.com

[2]University of São Paulo (USP) - emails: johnatan.mucelini@gmail.com, juarez_dasilva@iqsc.usp.br

Objectives of this project

We investigate some Machine Learning clustering methods [1] for selecting representative samples from a given dataset of molecules. Here, we show the effects of minimizing the maximum variance of molecules amount per cluster in a clustering process.

Methodology

The program consists of 2 algorithms working together: the clustering algorithm K-means and the optimization algorithm Simulated Annealing. In a simplified manner, the program can be represented by the diagram on the right.

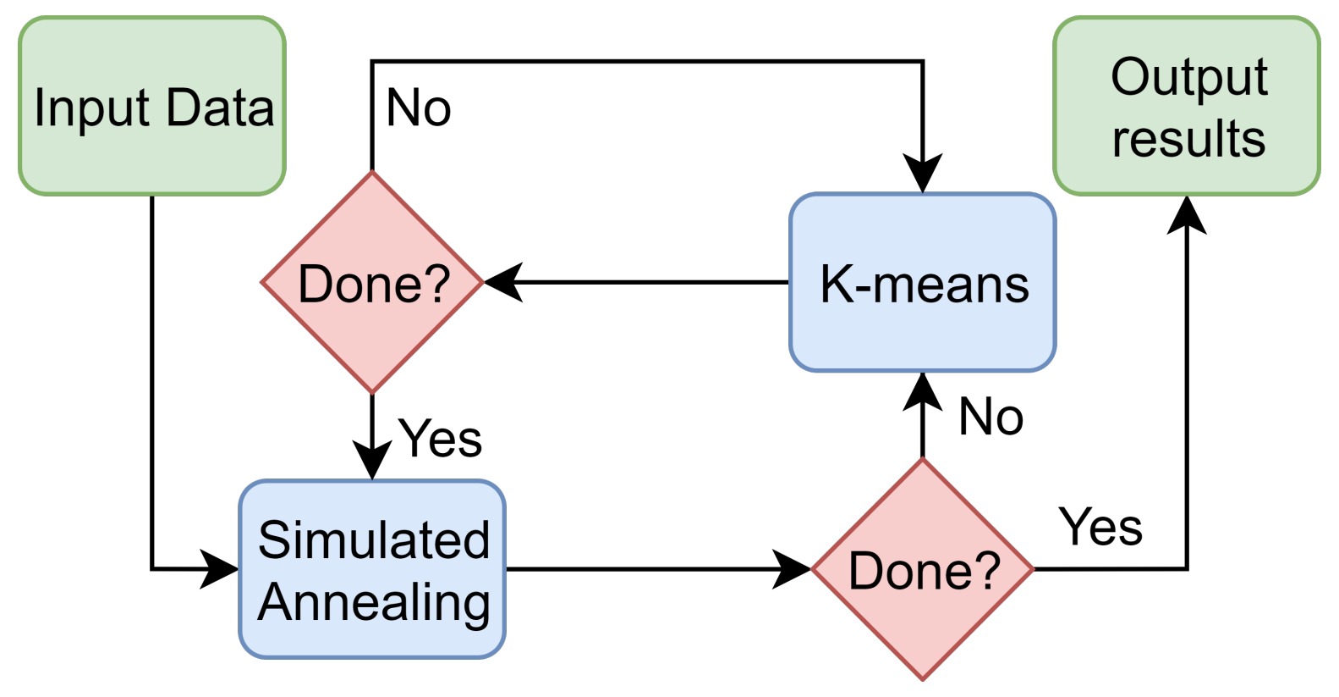

Figure: Program diagram

Clustering and K-means:

Clustering is a technique that, given a dataset, tries to group elements based on similarity [2], that is, elements within the same cluster are more similar with each other than those in different clusters. For this study, in particular, an algorithm called K-means was used.

Basically, it starts with a random initial partition and iteratively reassigns the patterns to clusters based on the similarity between the pattern and the cluster center, named centroid, until it converges.

Optimization and Simulated Annealing:

Mathematical optimization tries to find the best available values for a function, given certain constraints and objectives. One algorithm that accomplishes this task is Simulated Annealing, which is based on the analogy between the simulation of the annealing of solids and the problem of solving large combinatorial optimization problems [3]. One important feature of this algorithm is the fact that it avoids local minima/maxima and tries to find global ones.

Working together:

In the program, a dataset of features from various molecules is received. Simulated Annealing runs K-means as many times as it deems necessary (with a hard-coded limit), applying, each time, different weights to the features, in an attempt to minimize the maximum variance of molecules amount per cluster.

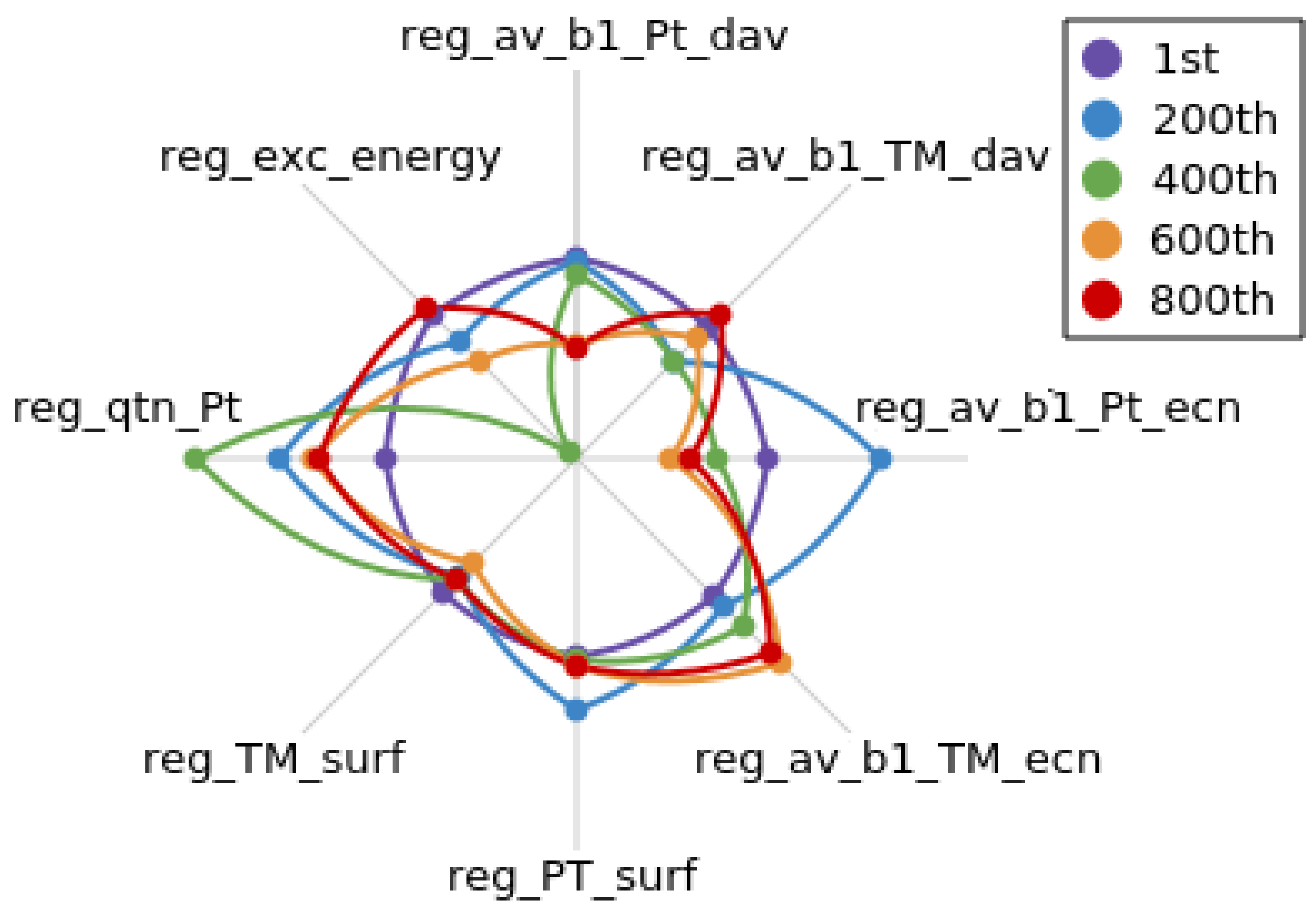

Figure: Convergence of the weights as the variance decreases / iterations increase

Results

2 Clusters:

Observing the graph on the right, we can see a disparity in *reg_qtn_Pt* (quantity of platinum), *reg_TM_surf* (amount of transition metals on the surface) and *reg_Pt_surf* (amount of platinum on the surface). It's easy to notice that the molecules have been clustered considering the amount of platinum and transition metals.

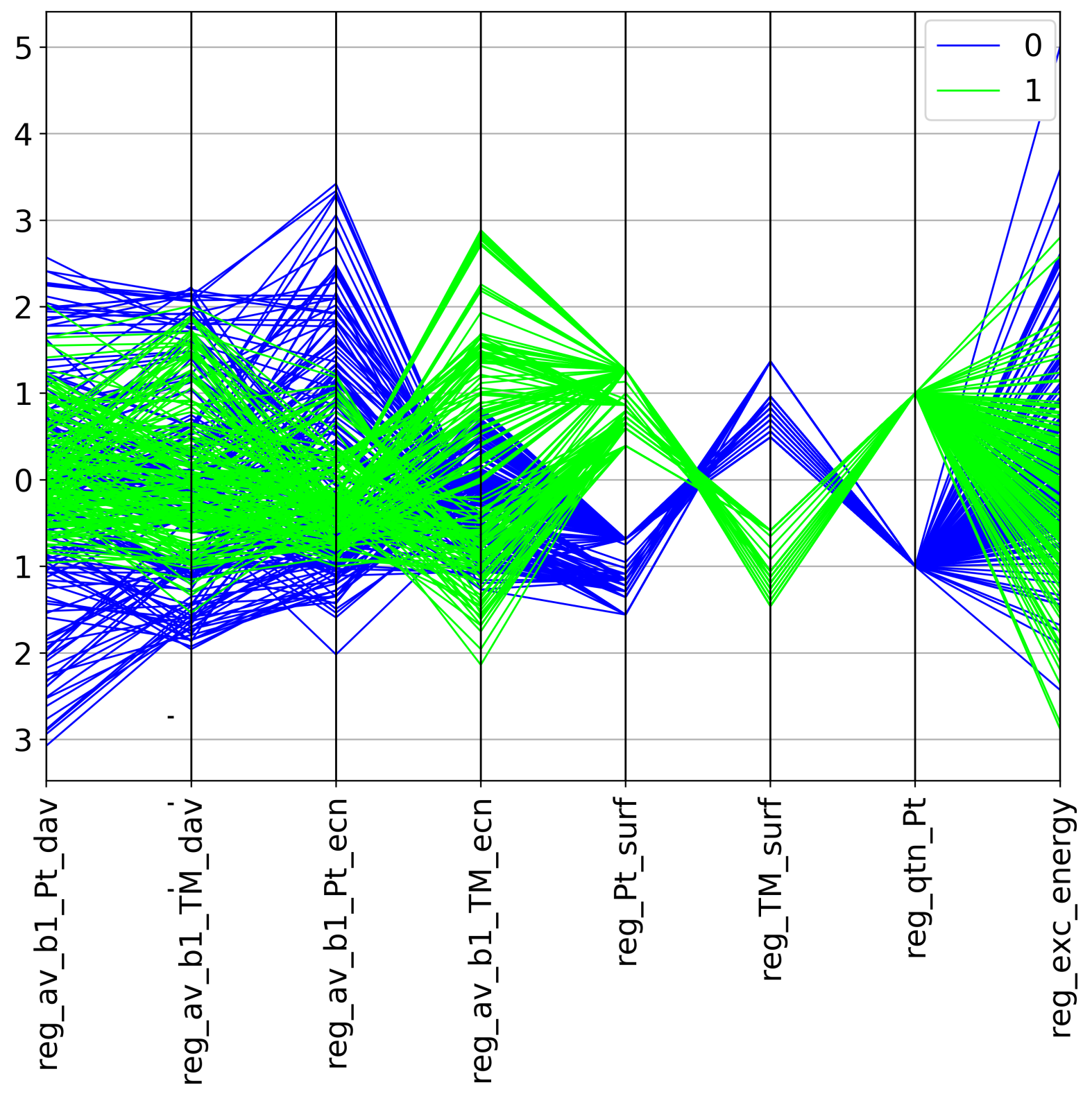

Figure: Parallel Coords for 2 clusters

4 Clusters:

With 4 clusters, we have the same platinum and transition metals division, but now each side has subdivided into 2 new clusters. This new division is given by *reg_av_b1_Pt_ecn* and *reg_av_b1_TM_ecn* (average quantity of neighbors of Pt or TM in the molecule). Since it's not clear just from the image beside, here are some histograms:

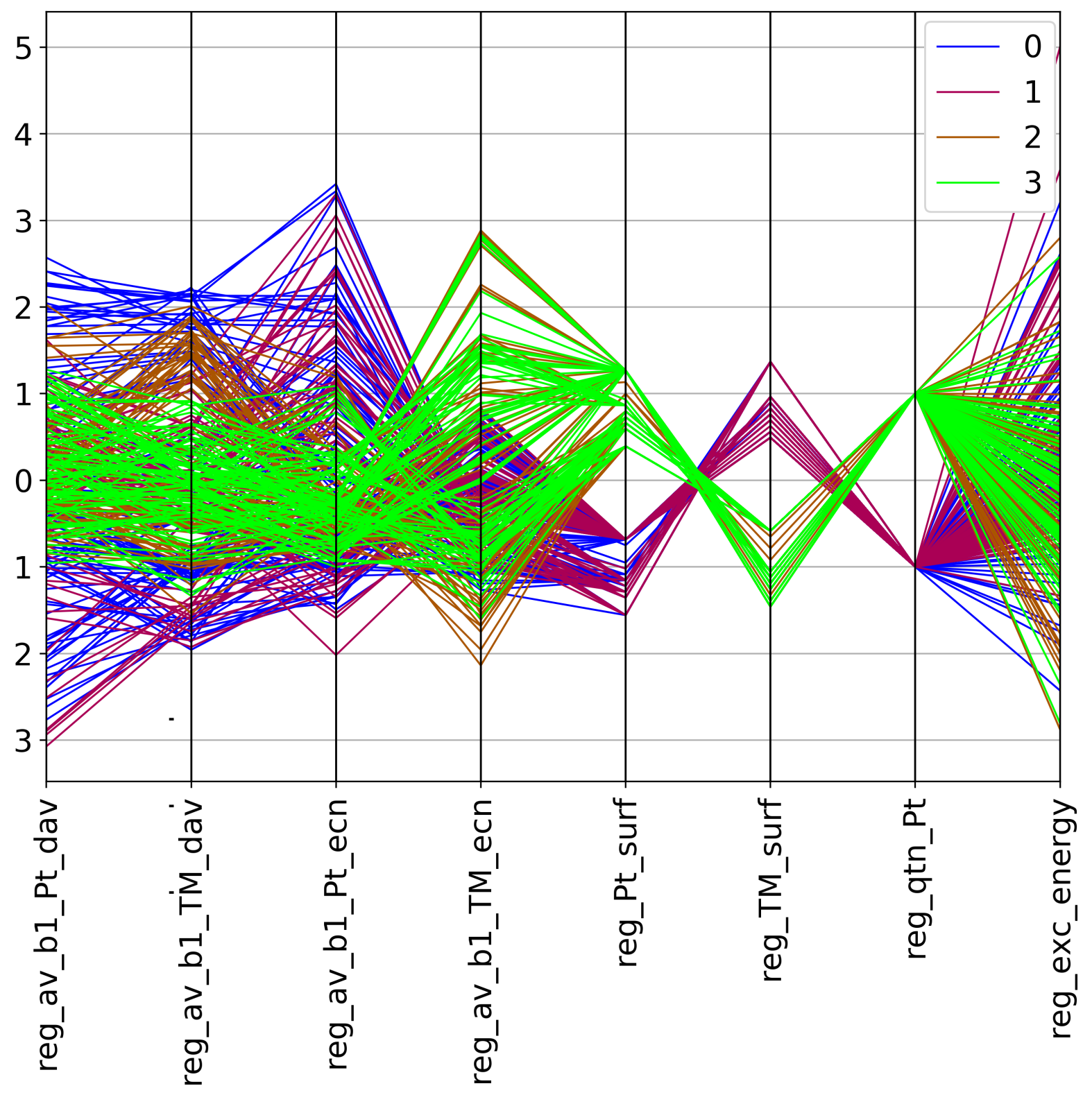

Figure: Parallel Coords for 4 clusters

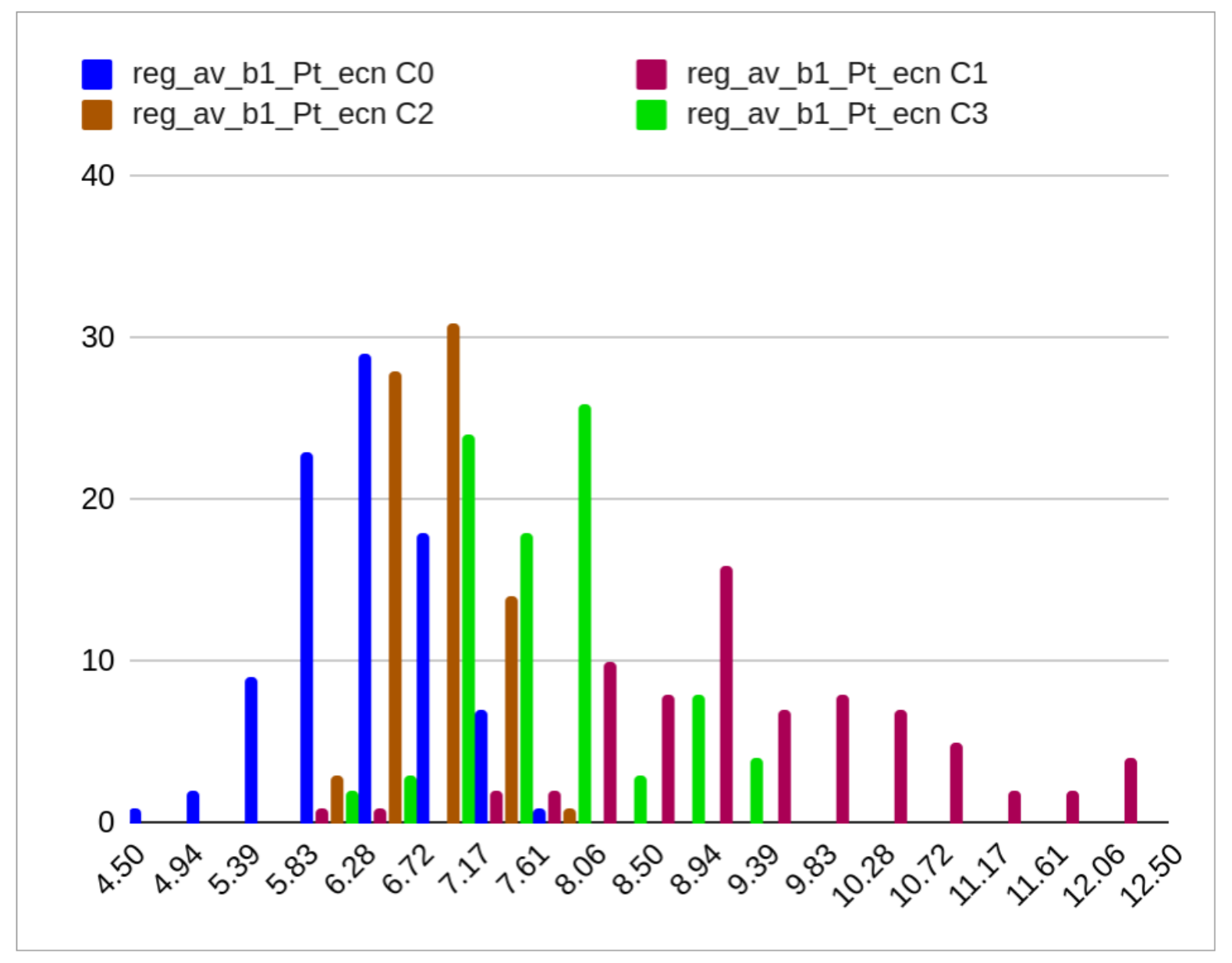

Figure: Histogram for Pt_ecn

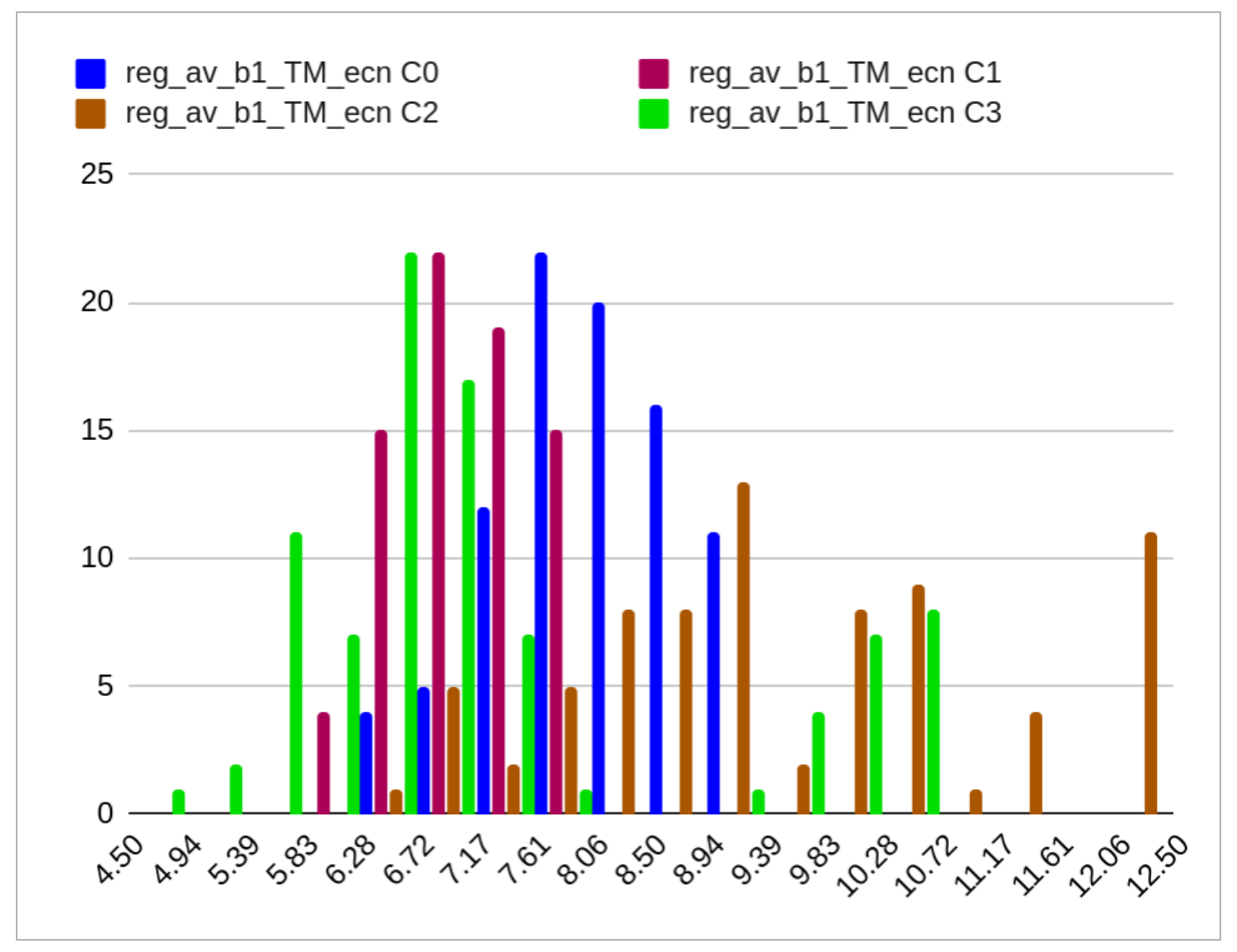

Figure: Histogram for TM_ecn

References

[1] A. K Jain.
Data clustering: 50 years beyond k-means.
Pattern Recognition Letters, 31:651–666, 2010.

[2] A. K. Jain, M. N. Murty, and P. J. Flynn.
Data Clustering: A Review.
ACM Computing Surveys, 31(3):265–323, September 1999.

[3] Aarts E.H.L. van Laarhoven P.J.M.
Simulated annealing.
In *Simulated Annealing: Theory and Applications.*, pages 7–15. Springer, Dordrecht, 1987.

Acknowledgments

CNPq, FAPESP, GPAM, ICT, UNIFESP, QTNANO, IQSC, USP