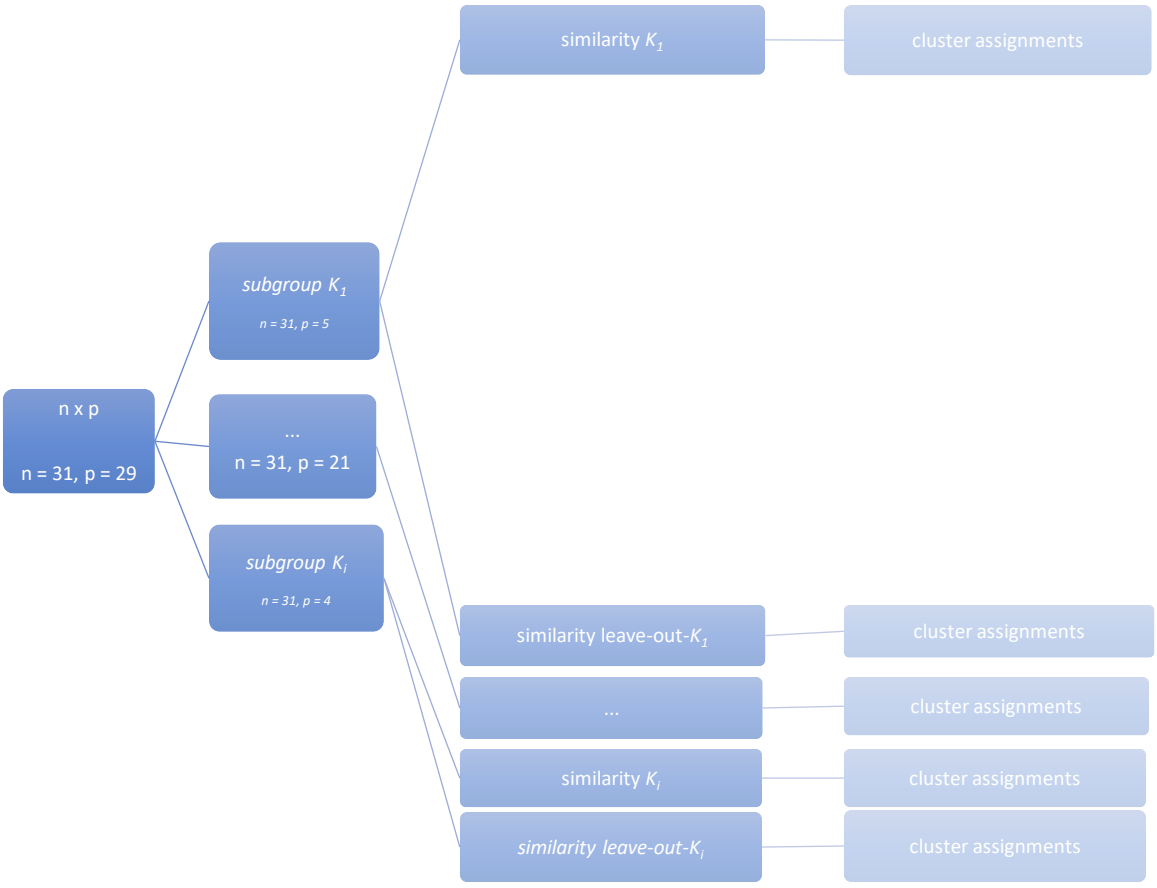
TSEClustering = Threshold Smoothing Ensemble Clustering – is an unsupervised clustering algorithm that allows smoothing over noisey data with low observational size in high dimensional space.

Smooth Ensemble Workflow, Steps:

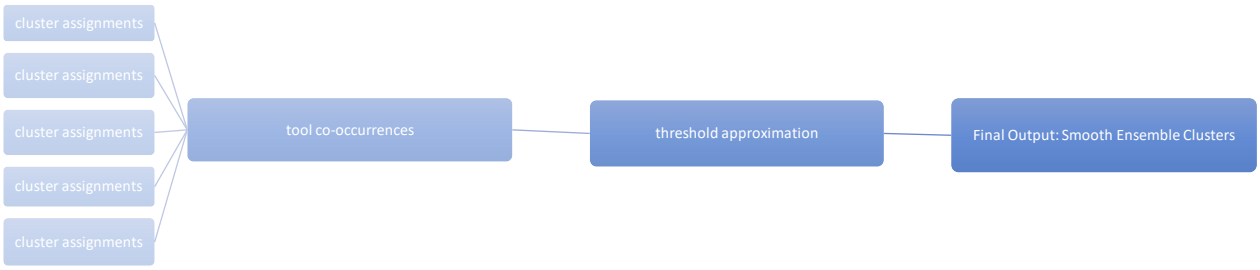
1. Input data of *n x p* (observations x predictors)
2. Intra-observational similarity matrices are generated through Gap statistic in various subsets of predictor space
3. Subset predictor space (*i.e.* subsets of the features) is calculated manually by selecting groups of features by use / type
4. Analysis without each predictor subset is performed (*i.e. leave-one-out*)
5. Resulting similarity matrices are used to independently cluster observations through the optimal number of clusters in Kmeans
6. Resulting cluster assignments are used to obtain co-occurrence of observations in relation to total number of analyses
7. A threshold approximation is applied to perform observational similarity dropout
8. Smooth Ensemble of observational correlations for Kmeans analysis

Smooth Ensemble Workflow, Part 1:



Smooth Ensemble Workflow, Part 2:

Kmeans -> pairwise co-occurrence = raw co-occurrence matrix -> smoothing function (math function) -> smooth ensemble (corr matrix) = probabilistic co-occurrence matrix



Steps for feature importance:

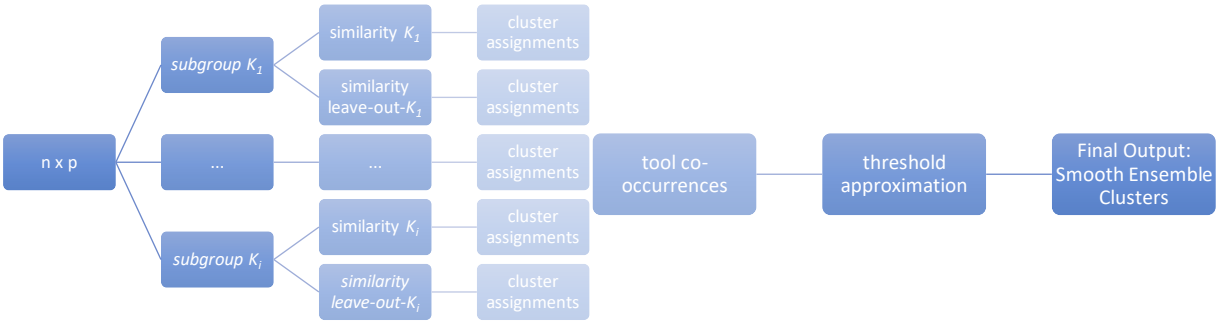
1. Each subgroup of features has
    a. within- vs. between-cluster variance
        i. measure of cluster goodness
    b. actual cluster assignment & prob. Of assignment
        i. measure of park goodness / sureness
2. Compare:
    a. 1.a.i across subgroups to get influence of features on cluster assignments
    b. 1.a.ii across subgroups to get influence of features on park assignment
3. Take high probs of assignments / importance from 1. and 2. to find what factors determined cluster assignment
    a. To get equation (e.g. linear combination) to get to cluster assignment
    b. E.g. only use probs > 0.9

References:

Tibshirani, R., Walther, G. and Hastie, T., 2001. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), pp.411-423.

Ciss, S., 2015. Random Uniform Forests for Classification, Regression and Unsupervised Learning.

Alternative Figure Full:



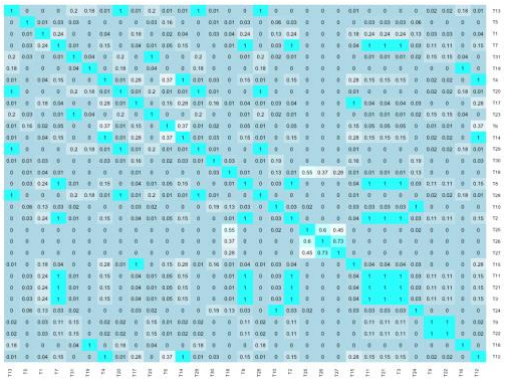
A: Data

a: Data

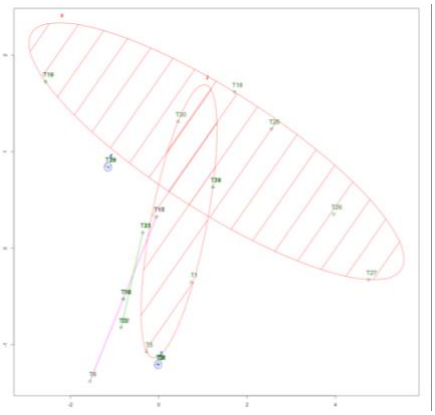
n x p
n = 31 (Tools)
p = 29
(All features)

p = {g_1, g_2, g_3, g_4, g_5, g_6}
g_1 = p_1 to p_6
…
g_6 = p_{26} to p_{29}

b: Similarity Matrices



c: K-means



for each g_i:

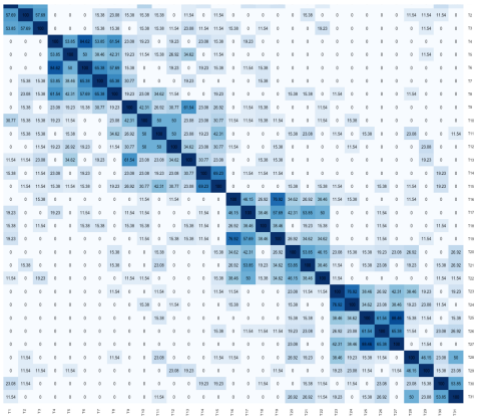
i) generate similarity matrix (*b*) & Kmeans (*c*) for n x g_i
ii) generate similarity matrix (*b*) & Kmeans (*c*) for n x (for all g_x not g_i)
iii) generate similarity matrix (*b*) & Kmeans (*c*) for all data

d: Raw Pairwise Co-Occurrences

Aggregate cluster assignments from K-means for scenarios in b

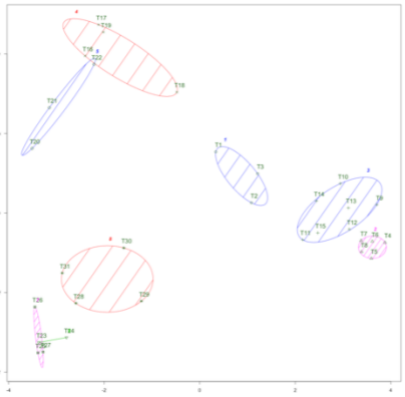
(i.e. 13 K-means = 6 groups + without each group + all data)

e: Probabilistic Pairwise Co-Occurrences



From aggregated cluster assignments (d) with threshold smoothing applied

f: Ensemble K-means



Final K-means Clusters from Probabilistic Co-Occur Matrix (e)