

PhD Dissertation - 2019 Joint Statistical Meeting

Monothetic Cluster Analysis with Extension to Functional Data

Tan Tran
Advisor: Dr. Mark Greenwood

?? July, 2019

Introduction

Clustering

- Unsupervised learning techniques for grouping (multivariate) responses with the goal of:
 - homogeneity — internal cohesion
 - separation — external isolation
- When to use clustering:
 - Find underlying patterns where little or no information about the data are known or to compare to known groups
 - Prediction of cluster membership based on the common characteristics of the clusters
- "A classification of a set of objects is not like a scientific theory and should perhaps be judged largely on its usefulness [...]." (Everitt, Landau, Leese, et al., 2011)

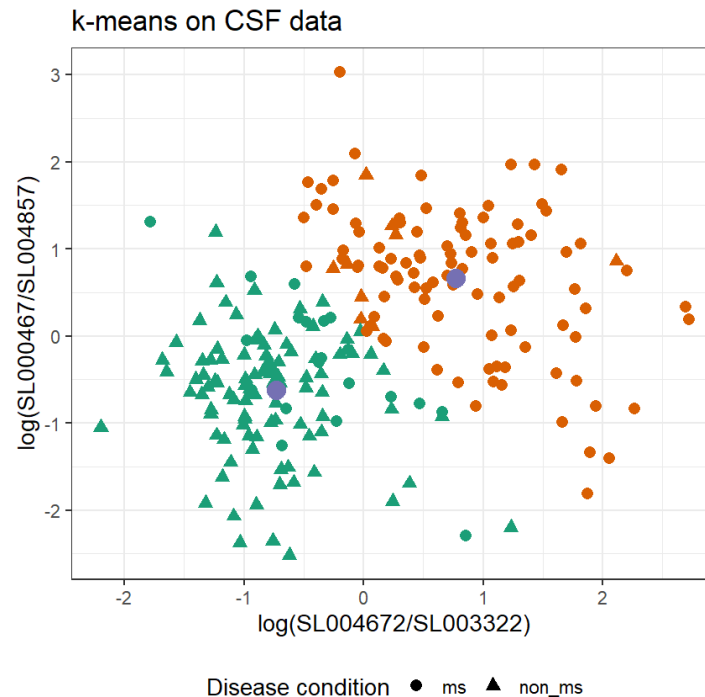
Two Clustering Techniques

- Optimization clustering techniques
 - The number of clusters, K , have to be pre-determined
 - Move the objects between clusters as long as it improves the criterion
 - k -means and partitioning around medoids (PAM, or k -medoids) are two examples of this technique
- Hierarchical clustering techniques
 - Distance measures between objects and between clusters must be defined
 - single linkage, complete linkage, Ward's method, etc.
 - Objects are fused together (agglomerative), or separated from each other (divisive) in each step based on the distance metric
 - The result is usually presented by dendrogram

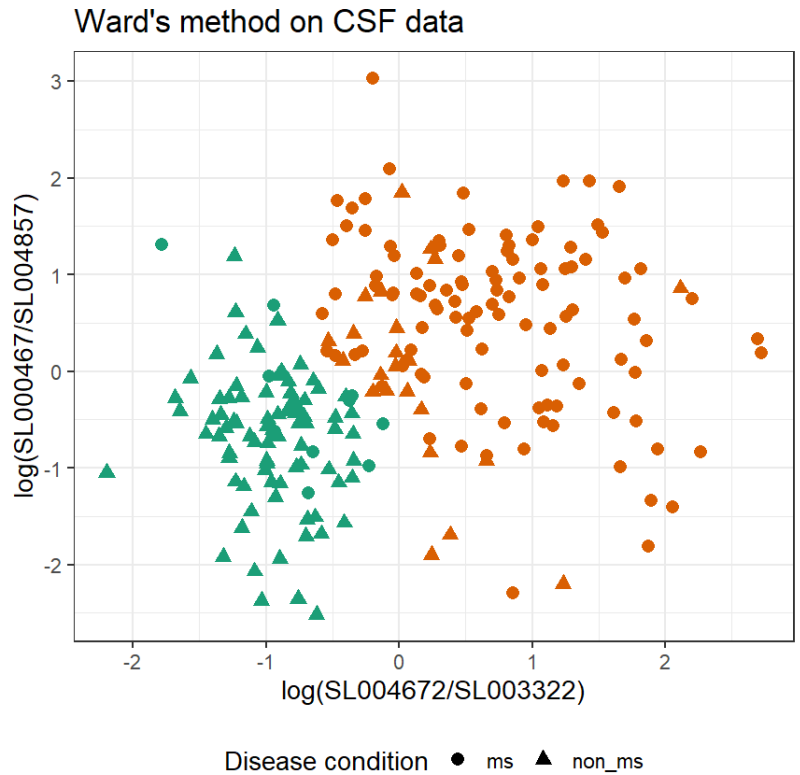
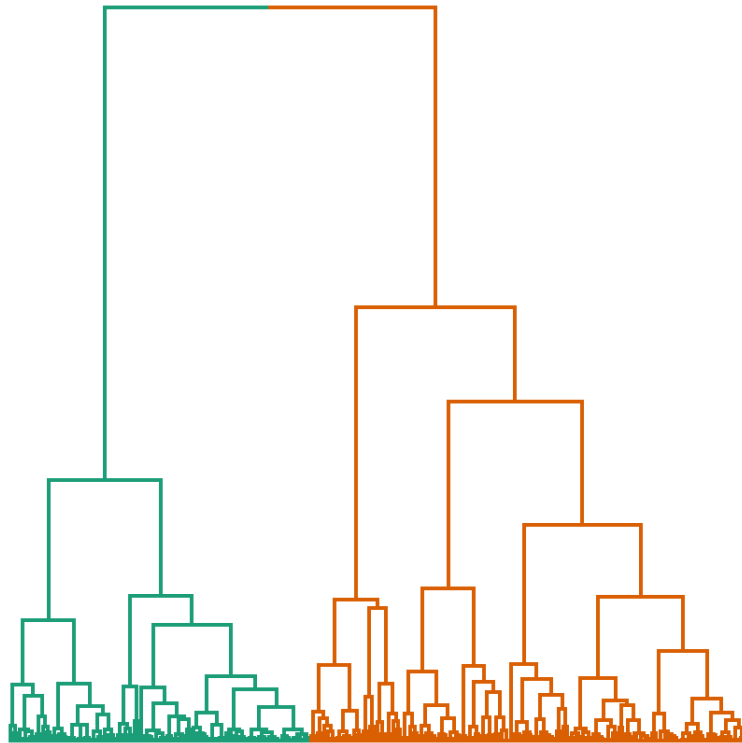
An Example: k -Means

Data from Barbour, Kosa, Komori, et al. (2017) cerebro-spinal fluid (CSF) biomarker data set with $n = 225$ subjects. The variables of interest are the standardized log ratios of 22 biomarkers (proteins).

- Multiple Sclerosis (MS) and non-MS patients are known
- $Q = 2$ proteins are displayed to visually demonstrate the method



An Example: Hierarchical with Ward's Method



Polythetic vs. Monothetic Clustering

- Popular methods like k-means and Ward's are **polythetic methods**
 - Clustered using the combined information of variables
 - Observations in a cluster are similar "on average" but may share no common characteristics
- There are also **monothetic divisive methods**
 - Data are bi-partitioned based on values of one variable at a time
 - Observations share common characteristics: in the same interval or category

Monothetic Clustering Algorithm

- Introduced in Chavent (1998) and Piccarreta and Billari (2007), inspired by classification and regression trees (Breiman, Friedman, Stone, et al., 1984)
- A global criterion called **inertia** for a cluster C_k is defined as

$$I(C_k) = \frac{1}{n_k} \sum_{(i,j) \in C_k, i > j} d^2(\mathbf{y}_i, \mathbf{y}_j)$$

where $d(\mathbf{y}_i, \mathbf{y}_j)$ is the distance between observations \mathbf{y}_i and \mathbf{y}_j and n_k is the cluster size

- Let s be a binary split dividing a cluster C_k into two clusters C_{kL} and C_{kR} . The decrease in inertia is

$$\Delta(s, C_k) = I(C_k) - I(C_{kL}) - I(C_{kR})$$

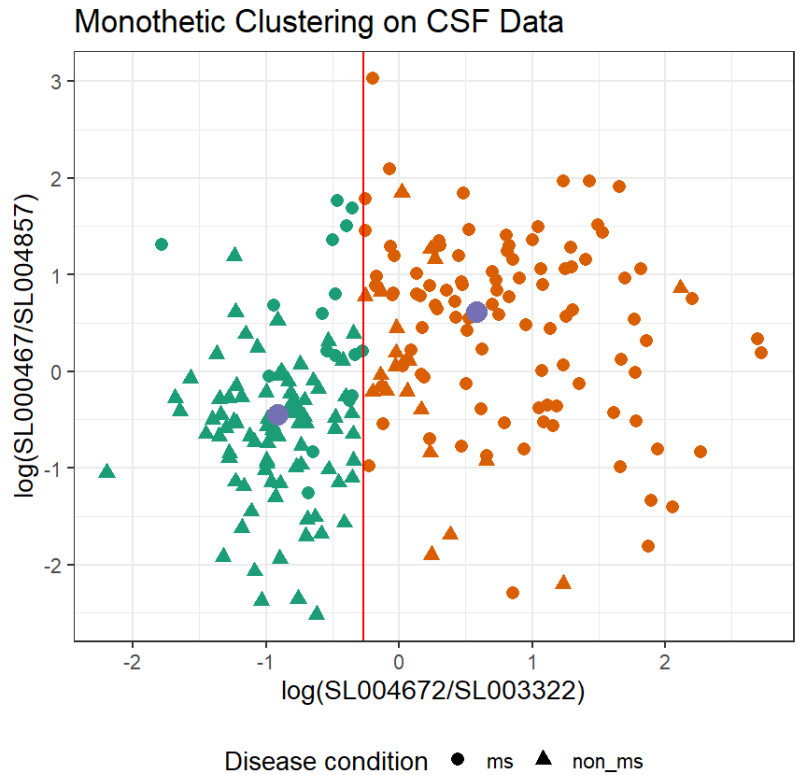
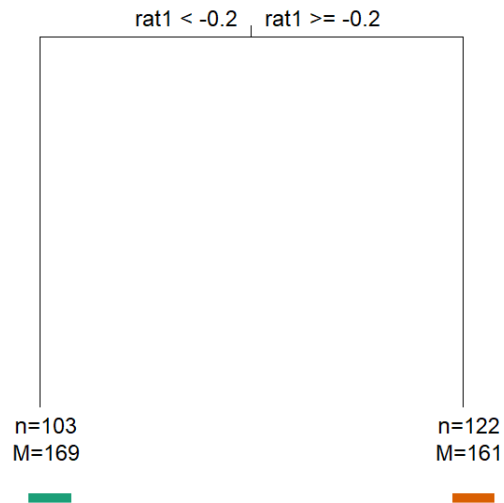
- The best split is selected as

$$s^*(C_k) = \arg \max_s \Delta I(s, C_k)$$

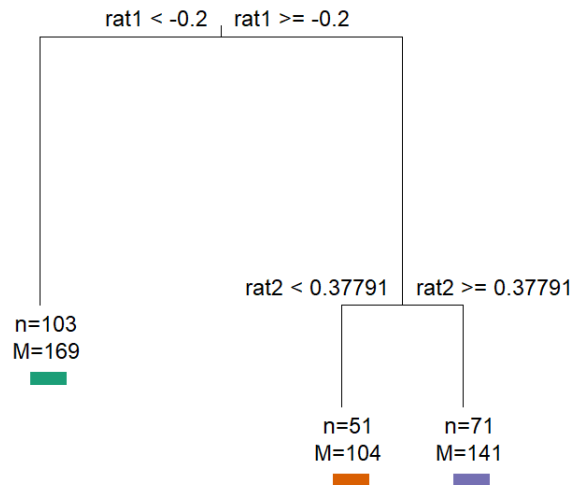
Properties of Monothetic Clustering

- Inertia is a global optimization criterion
- Bi-partition observations based on one variable at a time, making the method monothetic
- Defines rules for cluster membership
 - Easy classification of new members
- For the CSF data, monothetic clustering can be useful to allow classification into groups with shared characteristics that *might* relate to disease presence/absence

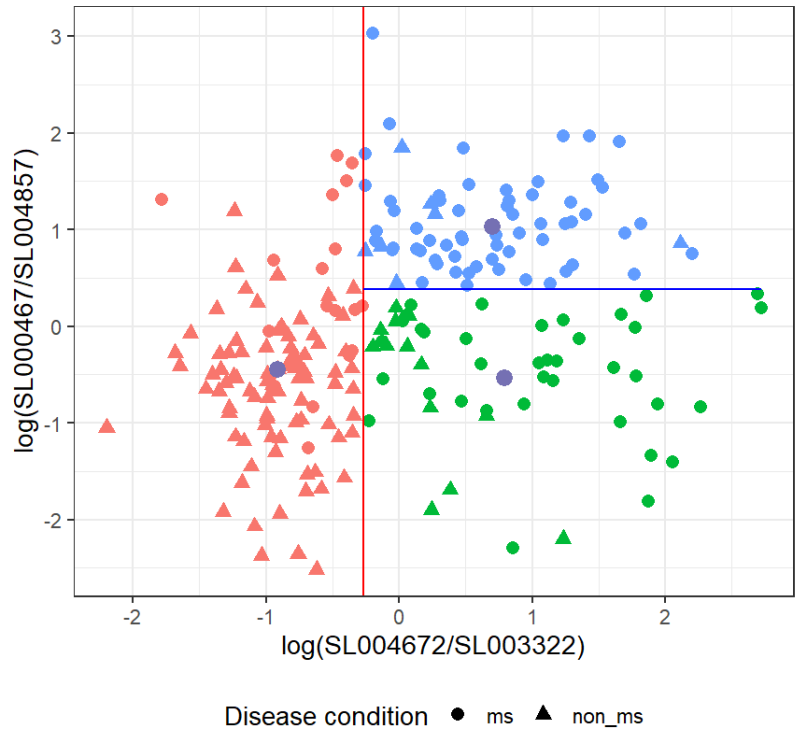
Monothetic Clustering on the CSF Data



One more split



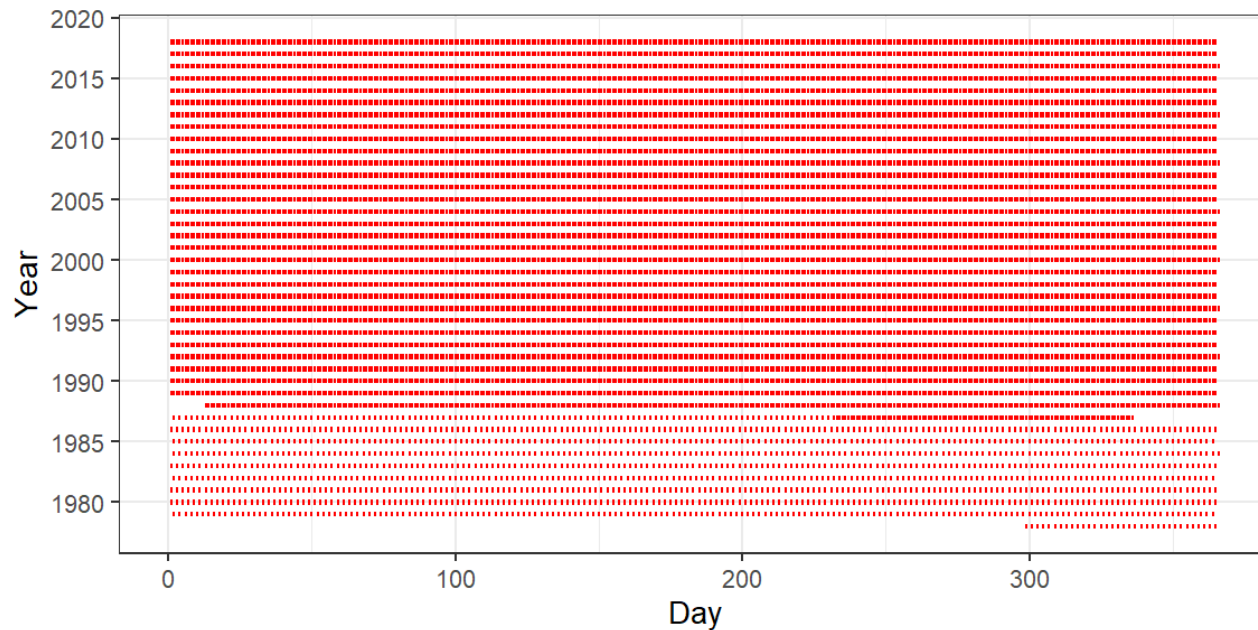
Monothetic Clustering on CSF data



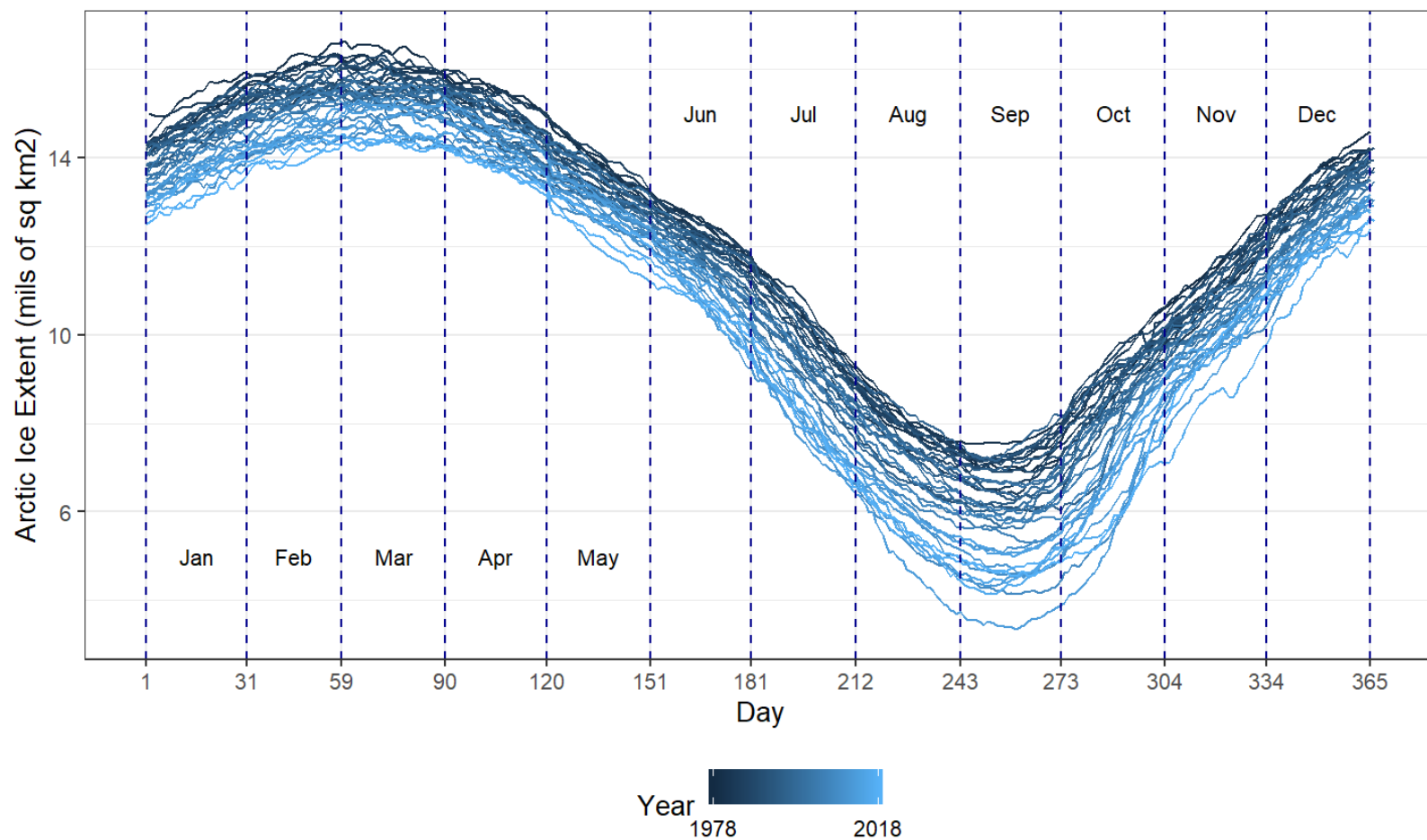
Clustering Functional Data

Arctic Sea Ice Extent Data

- Arctic Sea ice extent data set has been collected by National Snow & Ice Data Center since November 1978 (Fetterer, Knowles, Meier, et al., 2018)



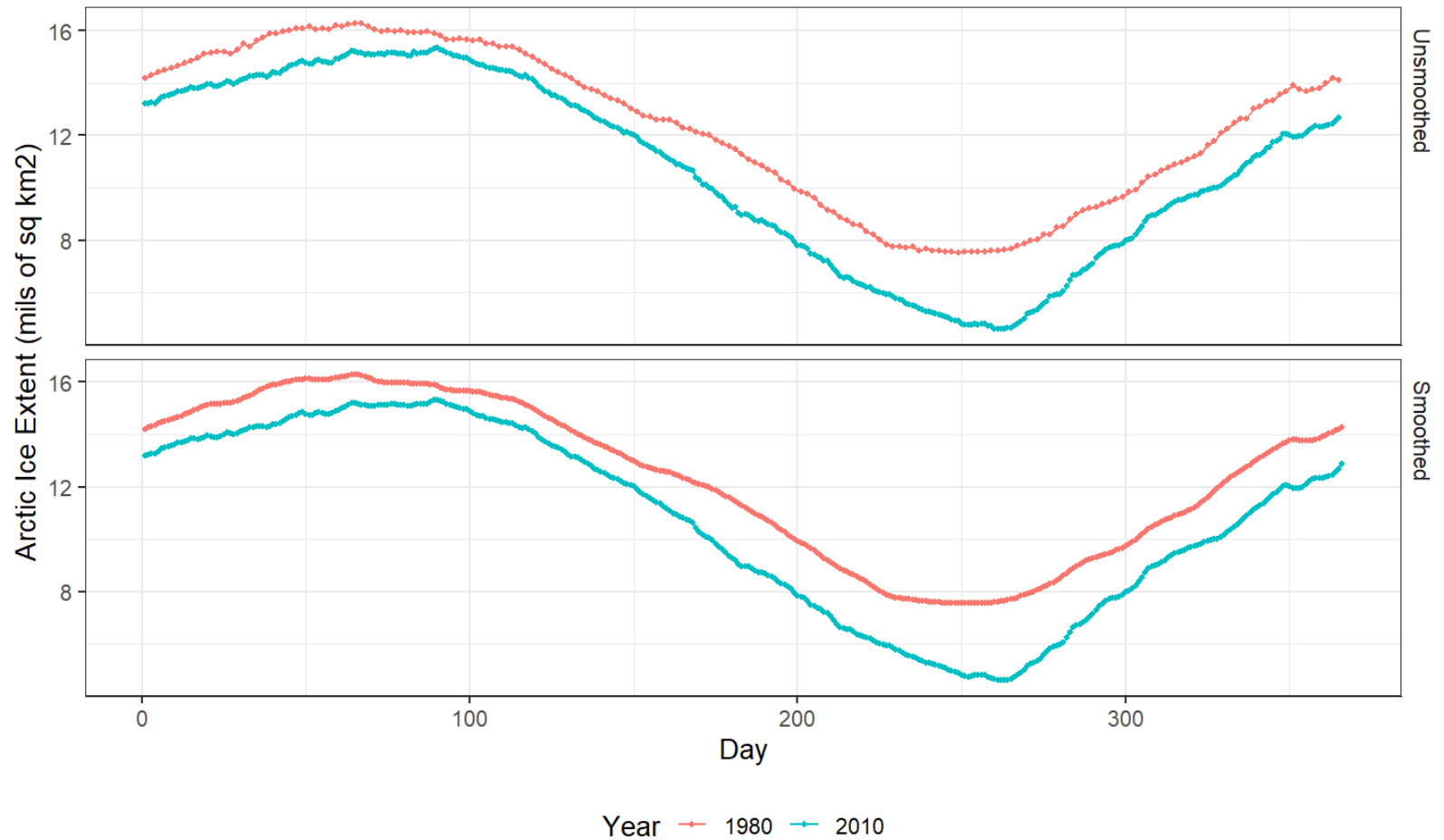
Raw Arctic Ice Extent



Functional Data

- When measurements were taken over some ordered index, such as time, frequency, or space (Ramsay and Silverman, 2005)
 - Responses are continuous as a function of the index
 - Possibly high frequency of observations and smooth underlying process or observations
- Observations are converted to functional data using basis functions:
 - **B-splines basis**, Fourier basis, Wavelets, etc.
- Penalized B-splines with knots at every day optimized with cross-validation for each curve
- Ice extent area in a year can be expressed as a function of time, $y_i(t)$
 - where i is the year, t is the day of year, and y is the ice extent at that time point

Smoothed and Interpolated Curves



Clustering Functional Data

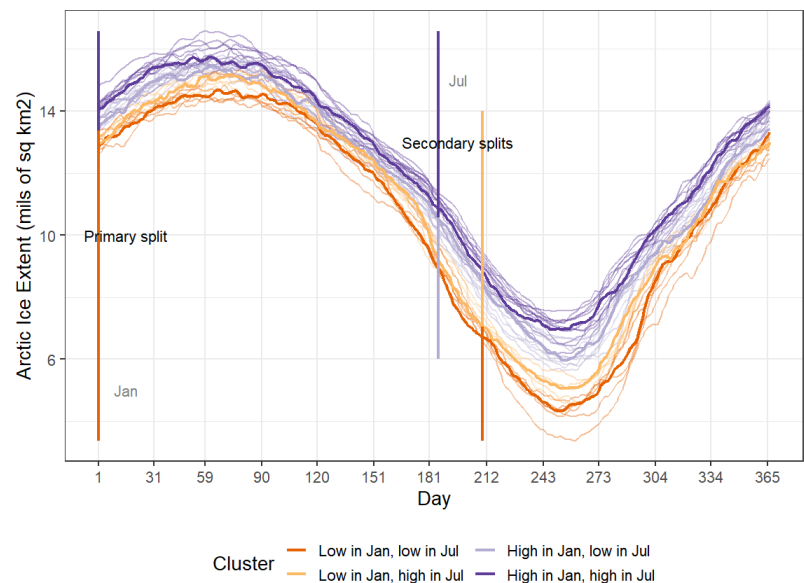
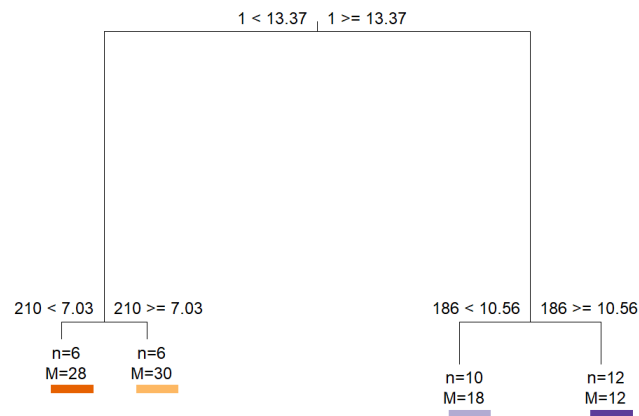
- The L_2 distance matrix can be calculated and used for non-functional clustering algorithms

$$d(y_i, y_j) = \sqrt{\int_T [y_i(t) - y_j(t)]^2 dt}$$

- Monothetic clustering uses functional data in their discretized form
 - Transform the data into functional presentations
 - Data are then estimated to a common fine grid of t , y_{it} from $y_i(t)$
 - Missing values are imputed and the data set is balanced with $t = 1, \dots, 366$ days for all years

Monothetic Clustering Result

- Evaluate functional data for each year, cluster years using the 366 "variables" (days) each year
- Splitting the data based on one variable (day) at a time
- Have to deal with many equivalent splits possible at "neighboring" variables



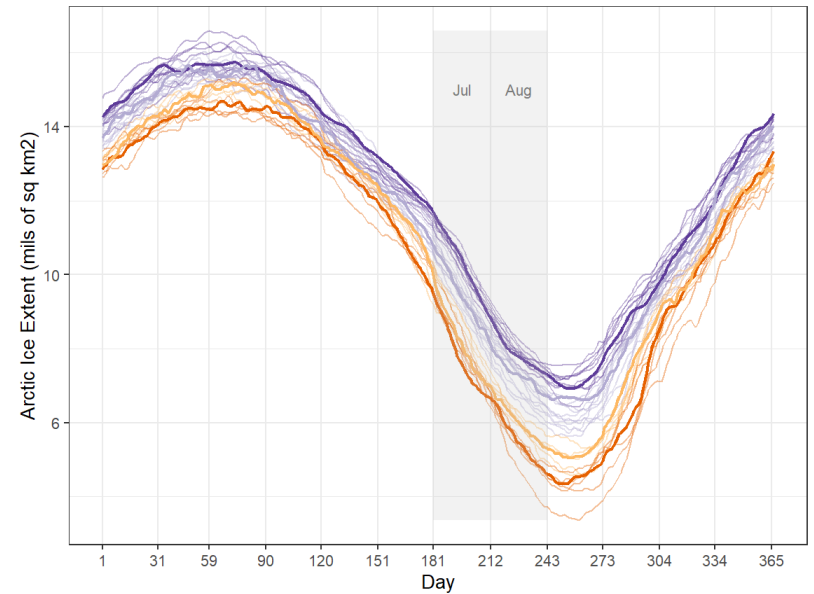
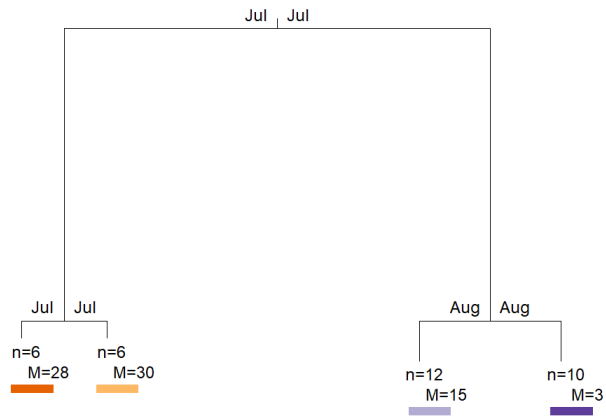
Partitioning Using Local Subregions

- By aggregating over regions of time, we develop a new clustering algorithm related to monothetic clustering that is more suited to functional data
- PULS recursively bi-partitions functional data using only groups from subregions.
- For each subregion $[a_1, b_1], \dots, [a_R, b_R]$, calculate an L_2 distance matrix

$$d_R(y_i, y_j) = \sqrt{\int_{a_r}^{b_r} [y_i(t) - y_j(t)]^2 dt},$$

- Apply a clustering algorithm (PAM, Ward's method, etc.) to create 2-group cluster solutions in R subregions
- Pick the solution that maximizes the difference in the global inertia
- Recursively apply to the newly created clusters

Partitioning Using Local Subregions Result



Comparison of Results and Cluster Prediction

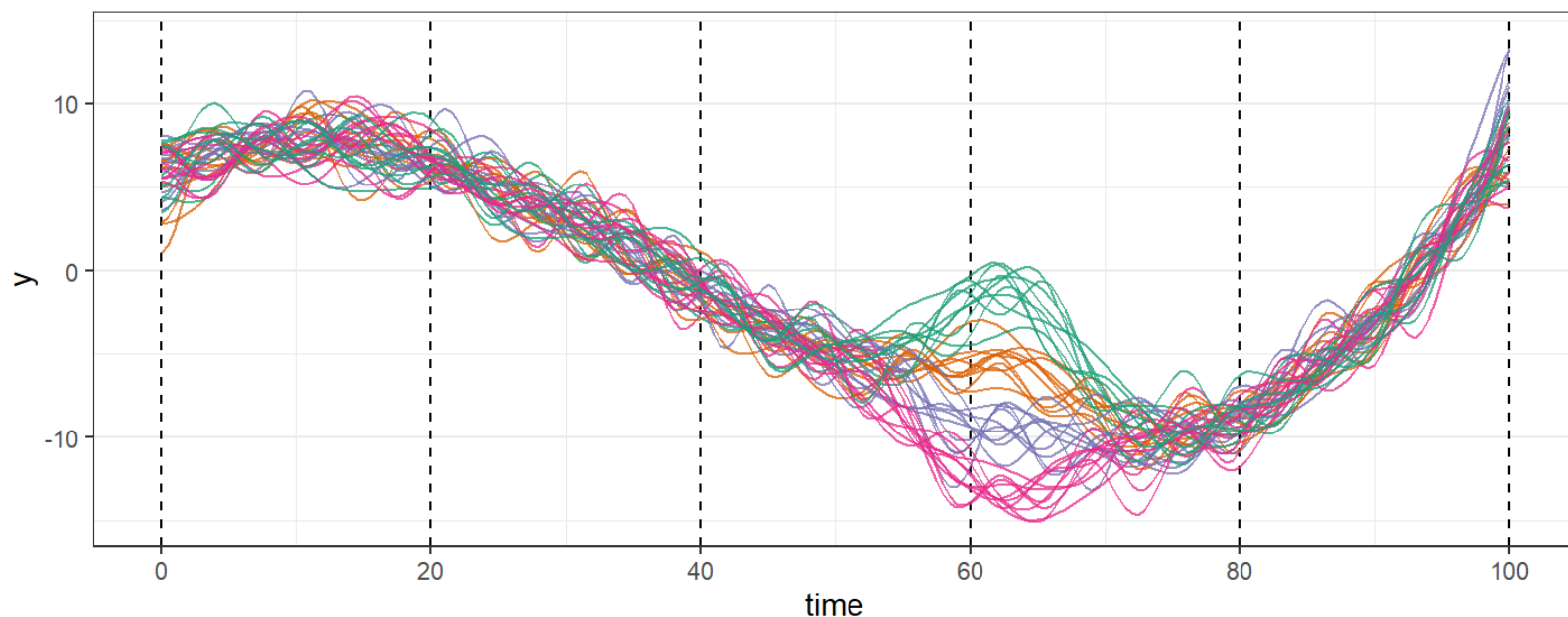
Name	PULS	MonoClust
High Jan, High Jul	1979- 1981 , 1983-1984, 1986, 1989, 1992, 1994, 1996	1979-1981, 1983-1986, 1989, 1992- 1994 , 1996
High Jan, Low Jul	1985, 1991, 1993, 1995, 1997 -2004	1991, 1995, 1997- 2000 -2004
Low Jan, High Jul	2005, 2008-2010, 2013 , 2014	2005, 2008-2010, 2013 , 2014
Low Jan, Low Jul	2007, 2011 , 2012, 2015-2017	2007, 2011 , 2012, 2015-2017

- One year in each decade was randomly withheld from the test data set:
 - 1982, 1990, 2006, 2018
- Predict the cluster from monothetic clustering's splitting rule tree

High Jan, High Jul	High Jan, Low Jul	Low Jan, High Jul	Low Jan, Low Jan
1982	1990	2006	2018

Simulation Study

- Compare the performance of various clustering techniques on functional data
 - Monothetic clustering, PAM, and Ward's method
 - PULS with PAM and PULS with Ward's method on five subregions



- Corrected Rand Index (Rand, 1971) and Pseudo-R2 are used to evaluate
- PULS and monothetic clustering both perform better than PAM and competitive with Ward's method
- Ward's method works well both within PULS and by itself

Conclusions

- Monothetic clustering bi-partitions data at the values of one variable at a time
 - Clusters share the same characteristics from split variables
 - Can interpret the resulting clusters
 - Can predict the cluster a new observation would fall into
- R packages `monoclust` and `PULS` are available on Github

Possible Extensions

- Sparse clustering (Witten and Tibshirani, 2010) can be applied to functional data to "weight" the contributions of variables (days) to the clustering process
- Monothetic Exhaustive search is slow, heuristic search algorithms may help with some trade-offs
 - Results of best split for each variable could be stored so they won't unnecessarily repeat
- Clustering the derivative of the Arctic ice extent data

References I

- Barbour, C, P. Kosa, M. Komori, M. Tanigawa, R. Masvekar, T. Wu, K. Johnson, P. Douvaras, V. Fossati, R. Herbst, Y. Wang, K. Tan, M. Greenwood, and B. Bielekova (2017). "Molecular-based diagnosis of multiple sclerosis and its progressive stage". In: *Annals of Neurology* 82.5, pp. 795-812. ISSN: 03645134. DOI: [10.1002/ana.25083](https://doi.org/10.1002/ana.25083).
- Breiman, L, J. Friedman, C. J. Stone, and R. Olshen (1984). *Classification and Regression Trees*. 1st ed. Chapman and Hall/CRC. ISBN: 0412048418.
- Chavent, M. (1998). "A monothetic clustering method". In: *Pattern Recognition Letters* 19.11, pp. 989-996. ISSN: 01678655. DOI: [10.1016/S0167-8655\(98\)00087-7](https://doi.org/10.1016/S0167-8655(98)00087-7).
- Everitt, B. S, S. Landau, M. Leese, and D. Stahl (2011). *Cluster Analysis*. 5th ed. Wiley, p. 346. ISBN: 0470749911.
- Fetterer, F., F. Knowles, W. Meier, M. Savoie, and A. K. Windnagel (2018). *Sea Ice Index, Version 3*. DOI: [10.7265/N5K072F8](https://doi.org/10.7265/N5K072F8). URL: <https://nsidc.org/data/g02135> (visited on 2018).

References II

Piccarreta, R. and F. C. Billari (2007). "Clustering work and family trajectories by using a divisive algorithm". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 170.4, pp. 1061-1078. ISSN: 0964-1998. DOI: [10.1111/j.1467-985X.2007.00495.x](https://doi.org/10.1111/j.1467-985X.2007.00495.x).

Ramsay, J. O. and B. W. Silverman (2005). *Functional data analysis*. Springer, p. 426. ISBN: 9780387400808.

Rand, W. M. (1971). "Objective Criteria for the Evaluation of Clustering Methods". In: *Journal of the American Statistical Association* 66, pp. 846-850. ISSN: 0162-1459. DOI: [10.1080/01621459.1971.10482356](https://doi.org/10.1080/01621459.1971.10482356).

Witten, D. M. and R. Tibshirani (2010). "A framework for feature selection". In: *American Statistician* 105.490, pp. 713-726. ISSN: 0162-1459. DOI: [10.1198/jasa.2010.tm09415.A](https://doi.org/10.1198/jasa.2010.tm09415.A).