Tutorial for STADIA Classification Stage using the "Diagnostic Mode"

The diagnostic mode built into STADIA is an opportunity for the user to contribute experience and domain knowledge into the classification process as the segmented DI length history data is separated into cluster groups. This document serves as a tutorial for operating STADIA in the diagnostic mode, and how to interpret the output in selecting the number of clusters to use in the classification stage to run the full, automated mode of STADIA. Here, the gap statistic is used to help identify the optimal number of clusters, the *k*-value, that should be used to separate the data considered in each scenario (Tibshirani, Walther, and Hastie 2001). Note that for this tutorial, the user is expected to be familiar with the STADIA stages, and to understand that the segmentation stage is completed at the point where the diagnostic mode analysis takes over, such that the clustering is performed on 3-D points each representing the three line segment characteristics identified from the length history data.

Gap statistic analysis

In all of the variations of the diagnostic mode, the *k*-means algorithm is repeated 100 times for each of the *k*-values in the range 1 to 12. For each run, the gap statistic is measured, and the results are displayed as the gap statistic plot, indicating the mean gap value and standard deviation whiskers for each *k*-value. The gap statistic is a comparison of the within-cluster dispersions measured for the results applied to the data in question and a "null" reference data set drawn from a random uniform distribution. Larger gap statistic values imply better clustering results. In their paper, Tibshirani et al. describe the gap statistic criterion, which suggest that the optimal number of clusters is the first *k*-value such that the following inequality is satisfied:

$$Gap(k) \geq Gap(k+1) - s_{k+1}$$

where $Gap(k+1)$ and s_{k+1} are mean the standard deviation of the gap statistic respectively when using (*k*+1)-many clusters (Tibshirani, Walther, and Hastie 2001)**.** Choosing the first *k*-value that satisfies this condition promotes choosing a lower number of clusters, which also prevents overfitting. Also, stricter interpretations can require choosing the first local maximum, which stems from the fact that choosing the optimal number of clusters is a difficult task considering an unsupervised machine learning approach, where the ground truth is unknown. For this reason, the user must use discretion and apply scientific domain expertise when choosing the final *k*-value.

Running the three diagnostic mode versions in STADIA

The diagnostic mode in STADIA can run the gap statistic analysis for *k*-means clustering in three ways: for **all segments**, for **positive sloped segments**, and for **negative slopes segments**. The corresponding user-defined parameters for running the STADIA's classification stage in the different diagnostic mode are as follows:

- KMEANS_DIAGNOSTIC_ALLFLAG: an indicator for conducting gap statistics analysis using all of the segment points. Acceptable values are 1 (on) or 0 (off).
- KMEANS_DIAGNOSTIC_PosSlopeFLAG: an indicator for conducting gap statistics analysis using only the points from positive sloped segment. Acceptable values are 1 (on) or 0 (off).
- KMEANS_DIAGNOSTIC_NegSlopeFLAG: an indicator for conducting gap statistics analysis using only the points from negative sloped segment. Acceptable values are 1 (on) or 0 (off).

For each variation of the diagnostic mode, after conducting the gap statistic analysis and saving the corresponding output files, STADIA halts the classification stage by stopping the current run and displaying a warning message indicating the version of diagnostic mode that is turned on. This way only one diagnostic mode variant is run at a time.

In addition to the output files plot, STADIA outputs the suggested *k*-value in the dialogue printed to the terminal for each of the diagnostic mode options. This *k*-value is also indicated in the gap statistics result plot, where the data is clustered using *k*-value, and it is also printed in the plot title. This automatically determined *k*-value corresponds to the gap statistic criterion. However, the user should use prior knowledge of their data and its characteristics when making a decision in the event that this automated suggestion is contradictory to case scenario being explored. For example, if the user feels that the suggested *k*-value suggests too few clusters, the next local maximum may be the user's choice to yield the desired results.

After a diagnostic run is completed, the user is expected to study the gap statistic results, make selections for STADIA variables accordingly, and move to the next diagnostic mode option. Once all the diagnostic mode options have been exhausted, the user should have all the information needed from the diagnostic mode, specifically the two *k*-values for the positive and negative slope segments. The user can then turn off the diagnostic modes by setting each variable = 0, and execute the code again to run STADIA in the fully automated mode.

Diagnostics Mode Part 1: All Segments

The purpose of attempting to cluster **all segments** together is primarily for demonstration purposes. When setting KMEANS_DIAGNOSTIC_ALLFLAG=1, or turning this mode on, gap statistics analysis is performed on the data comprised of all the segment points. STADIA developers regularly observed this approach to yield a gap statistic plot that is monotonically increasing, where a clear local maximum or a point satisfying the gap criterion is not evident, or appearing at a *k*-value larger than what is reasonably expected. This is an indicator that the data may have substructures not yet considered, which would benefit from considering subsets of the data separately. Thus, this variation in the diagnostic mode is merely for the user to verify that this indeed is not the best way to conduct the clustering procedure, and to justify the need to break up the data into smaller subsets prior to clustering.

Furthermore, there is no *k*-value associated to conduct clustering data for all the segments together. For this reason, the STADIA classification stage and the remaining part of the diagnostic mode moves forward by clustering positive slope segments and negative slope segments separately. At this time, the user should switch off the "all segments" variation of the diagnostic mode by setting KMEANS_DIAGNOSTIC_ALLFLAG=0, and re-running STADIA to conduct the next step in the process, which is to analyze the positive sloped segments.

Diagnostics Mode Part 2: Positive Sloped Segments

Setting KMEANS_DIAGNOSTIC_PosSlopeFLAG=1 turns on the diagnostic mode for conducting the gap statistic analysis on the positive sloped segments data. Typical results yield the optimal *k*-value to be

between 1 through 3. Values greater than or equal to 4 are typically not interpretable and difficult to explain, though they are included as part of the gap statistic analysis in STADIA for completeness. At this time, the user should decide on the *k*-value that represents the number of clusters that best separates the positive sloped segments. The user should assign that value to the KMEANS_NumClust_PosSlope variable in the "Input_and_Run.m" file. After completing this step, the user should switch off the "positive sloped segments" variation of the diagnostic mode by setting KMEANS_DIAGNOSTIC_PosSlopeFLAG =0, and re-running STADIA to conduct the next step in the process, which is to analyze the negative sloped segments.

Diagnostics Mode Part 3: Negative Sloped Segments

Setting KMEANS_DIAGNOSTIC_NegSlopeFLAG=1 turns on the diagnostic mode for conducting the gap statistic analysis on the positive sloped segments data. Typical results yield the optimal *k*-value to be between 1 through 3. Values greater than or equal to 4 are typically not interpretable and difficult to explain, though they are included as part of the gap statistic analysis in STADIA for completeness. At this time, the user should decide on the *k*-value that represents the number of clusters that best separates the positive sloped segments. The user should assign that value to the KMEANS_NumClust_NegSlope variable in the "Input_and_Run.m" file. After completing this step, the user should switch off the "positive sloped segments" variation of the diagnostic mode by setting KMEANS_DIAGNOSTIC_NegSlopeFLAG =0, and re-running STADIA to conduct the next step in the process, which is to run STADIA in the fully automated mode

Special Considerations:

As mentioned in the positive and negative sloped segment analysis sections, expected selections for the optimal *k*-values are the integers 1, 2, or 3. The data sets that yielded that identified the largest number segments from length history data was generated from 10 hour long simulations, which was helpful in filling up the feature space as much as possible. The analysis of even this data never yielded a *k*-value greater than 3 (Mahserejian et al., n.d.). For this reason, the automated mode only accepts *k*-values no more than 3 for the KMEANS_NumClust_PosSlope and KMEANS_NumClust_PosSlope variables. Choosing anything larger than 3 will terminate STADIA early and return an error explaining the circumstance.

Another special circumstances exists when selected a *k*-value of 2 for either the positive or negative sloped segment data subsets. Choosing *k*=1 is obvious, since all of the segment points would be categorized as growth or shortening, and the stutter label would be ignored. Choosing *k*=3 would assign the three sub-phase labels to each cluster: up stutter, brief growth, and sustained growth for the positive sloped segments, and down stutter, brief shortening, and sustained shortening for the negative sloped segments. However, it is unclear how to assign two of the possible three labels when a user selects *k*=2. To aid with the matter in this case, the user needs provide one of the following additional options which determines how the relevant data subset will be labeled.

So, if the user assigns KMEANS_NumClust_PosSlope = 2, then the STADIA variable KMEANS_Pos2_Option needs to be assigned one of the following values:

- KMEANS_Pos2_Option = 'A' : two growth phases, such that the two clusters will be assigned brief growth or sustained growth class labels
- KMEANS_Pos2_Option = 'B' : one growth and one stutter phase, such that the two clusters will be assigned sustained growth or up stutter class labels

If the user assigns KMEANS_NumClust_NegSlope = 2, then the STADIA variable KMEANS_Pos2_Option needs to be assigned one of the following values:

- KMEANS_Neg2_Option = 'A' : two shortening phases, such that the two clusters will be assigned either brief shortening or sustained shortening class labels
- KMEANS_Neg2_Option = 'B' : one shortening and one stutter phase, such that the two clusters will be assigned either sustained shortening or down stutter class labels

Output content:

Resulting output files from the STADIA diagnostic mode are stored into a subdirectory contained within the same output directory that other STADIA output files are stored. Depending on the diagnostic mode option that is turned on, files will be generated into one of the corresponding directories:

- AllSegments_Diagnostics
- PositiveSlopeSegments_Diagnostics
- NegativeSlopeSegments_Diagnostics

For each diagnostic mode variant, the following output files are produced:

- GapStat_Table_[DATE]_[TIME].txt: a comma delimited table with columns holding the *k*-value, gap statistic values, and standard deviations, which collectively represent the gap statistic analysis results
- 3DRawData_Figure_[DATE]_[TIME].png (or .fig): a 3D plot of the raw segment points data that are analyzed in this run of the diagnostic mode
- 3DStandardizedData_Figure_[DATE]_[TIME].png (or .fig): a 3D plot of the log transformed and standardized segment points data that are passed in directly to the *k*-means algorithm in this run of the diagnostic mode
- GapStatClustering_Figure_[DATE]_[TIME].png (or .fig): the gap statistic analysis results shown in two plots. The gap statistic plot (left) used to determine the optimal *k*-value to cluster the input data, and the clustering results (right) when using the optimal *k*-value, which is also reported in the plot title.

Note that the [DATE] and [TIME] placeholders in the filenames represent the date and time at the time that STADIA was run to produce that content. Additionally, note that three figures are saved as both PNG images, as well as matlab figure files.

BIBLIOGRAPHY

Mahserejian, Shant M., Jared P. Scripture, Ava J. Mauro, Elizabeth J. Lawrence, Erin M. Jonassen, Kris S. Murray, Jun Li, et al. n.d. “Stutter: A Transient Dynamic Instability Phase That Is Strongly Associated with Catastrophe.”

Tibshirani, Robert, Guenther Walther, and Trevor Hastie. 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): 411–23. https://doi.org/10.1111/1467-9868.00293.