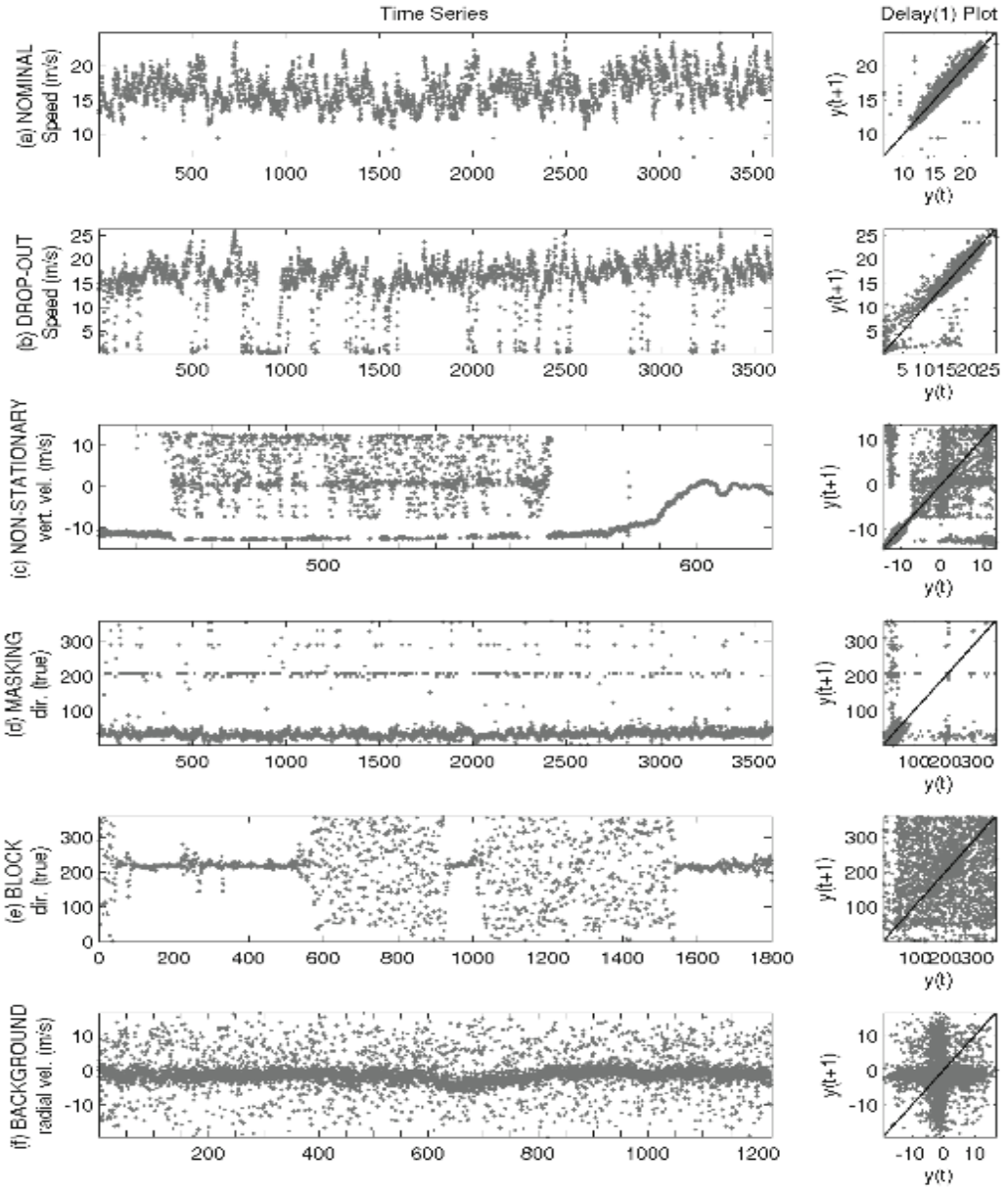
Intelligent Outlier Detection Algorithm

A harsh environment

Cases studied



Cases studied

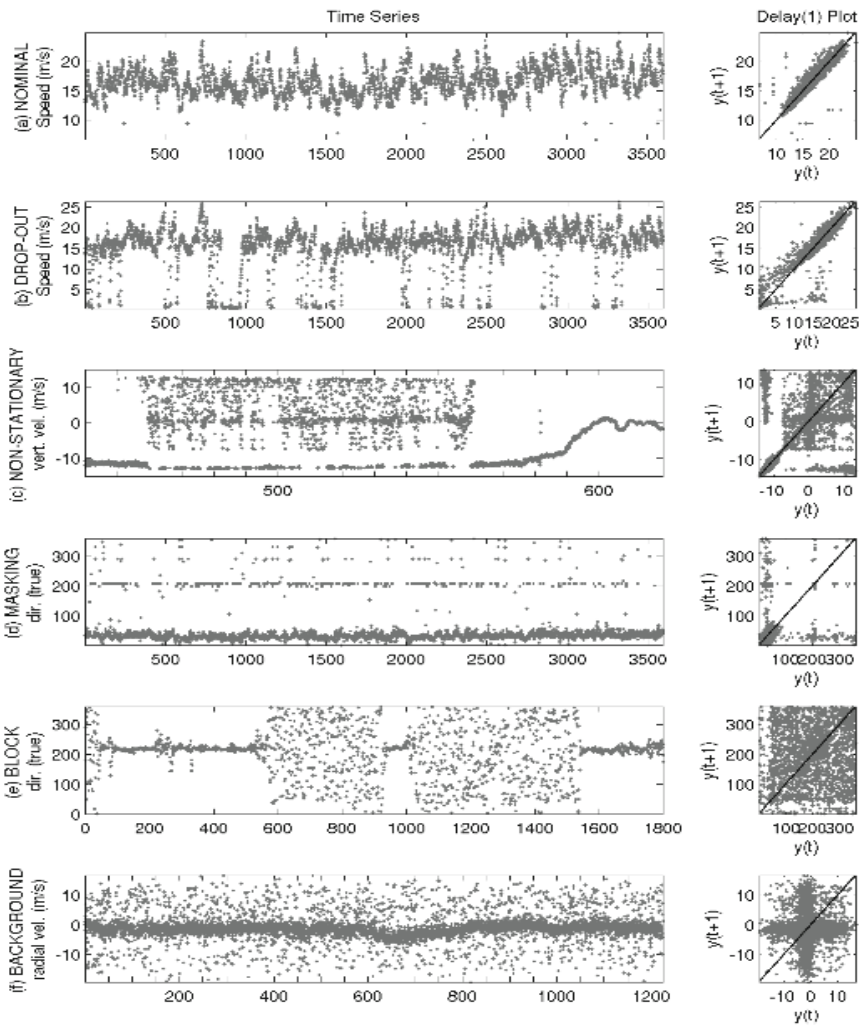
- Nominal (anemometer wind speed)
- Drop-out (anemometer wind speed)
- Non-stationary (aircraft vertical velocity)
- Masking (anemometer wind direction)
- Block (anemometer wind direction)
- Uniform noise (LIDAR Radial Velocity)

IODA

- Motivated by time-series data collected in Juneau Alaska
- Numerous cases of instrument failures
- Build a time-series quality control algorithm that mimics the humans ability to identify bad data
- Essentially an optimization problem i.e. find the largest subset set of points to maximize the autocorrelation.

Motivation

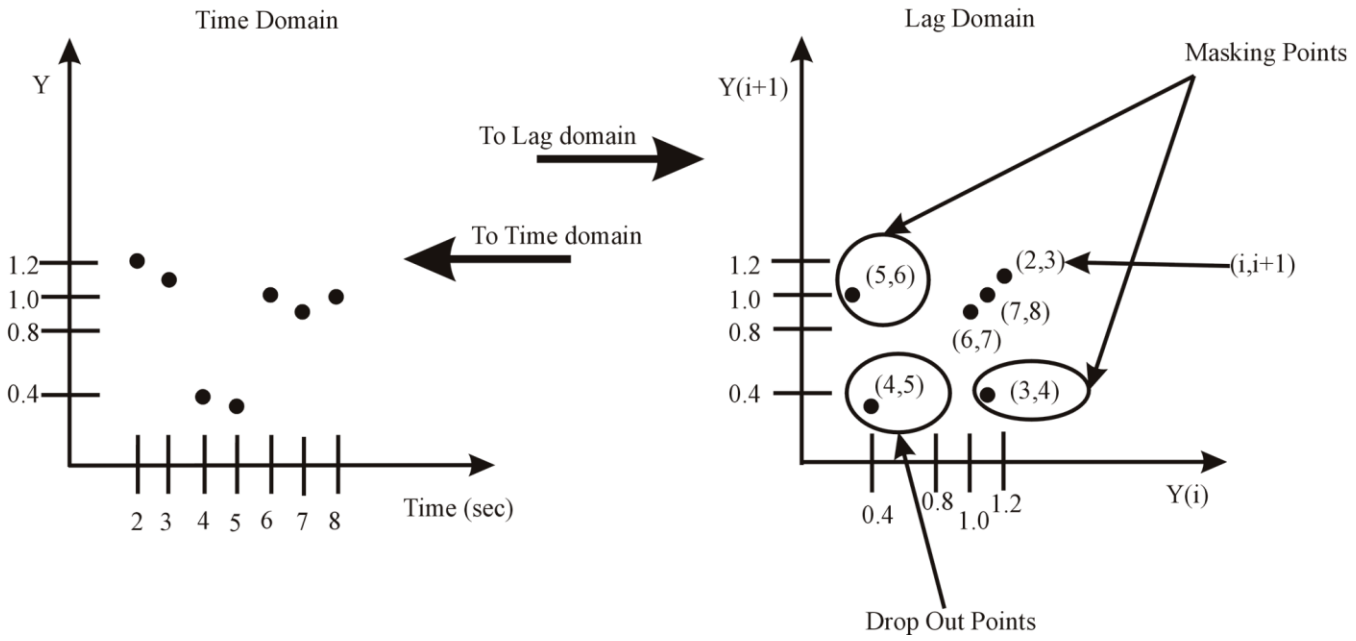
- A typical approach to QC the data might use a Z-statistic and over a window of data
- The results depend on the size of the window used.
- If most of the data in the window is bad then the test will fail.
- A single method will not work on all failures
- A multi-resolution technique is needed
- Need a framework to easily add new failure modes as they are discovered



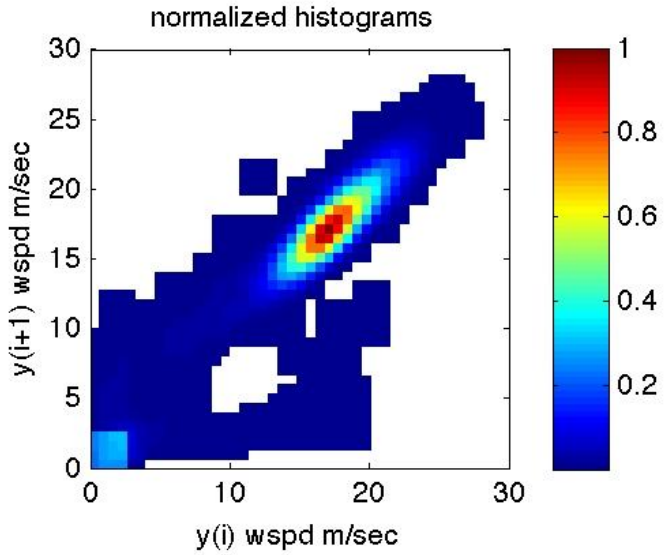
IODA

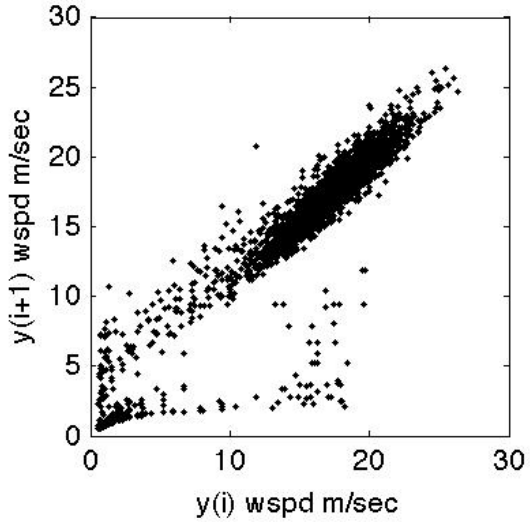
- Image processing applied to time series data to detect changes in auto-correlation
- Cluster in time and delay space (time-delay embedding)
- Use a decision tree to identify the failure mode
- Score the data depending on the failure mode and "type" of data point
- R. A. Weekley, R. K. Goodrich, and L. B. Cornman, "An Algorithm for Classification and Outlier Detection of Time-Series Data," *Journal of Atmospheric and Oceanic Technology*, vol. 27, no. 1, pp. 94–107, 2010.
- U.S. Patent Number 6735550 issued May 11, 2004

Time-delay Embedding



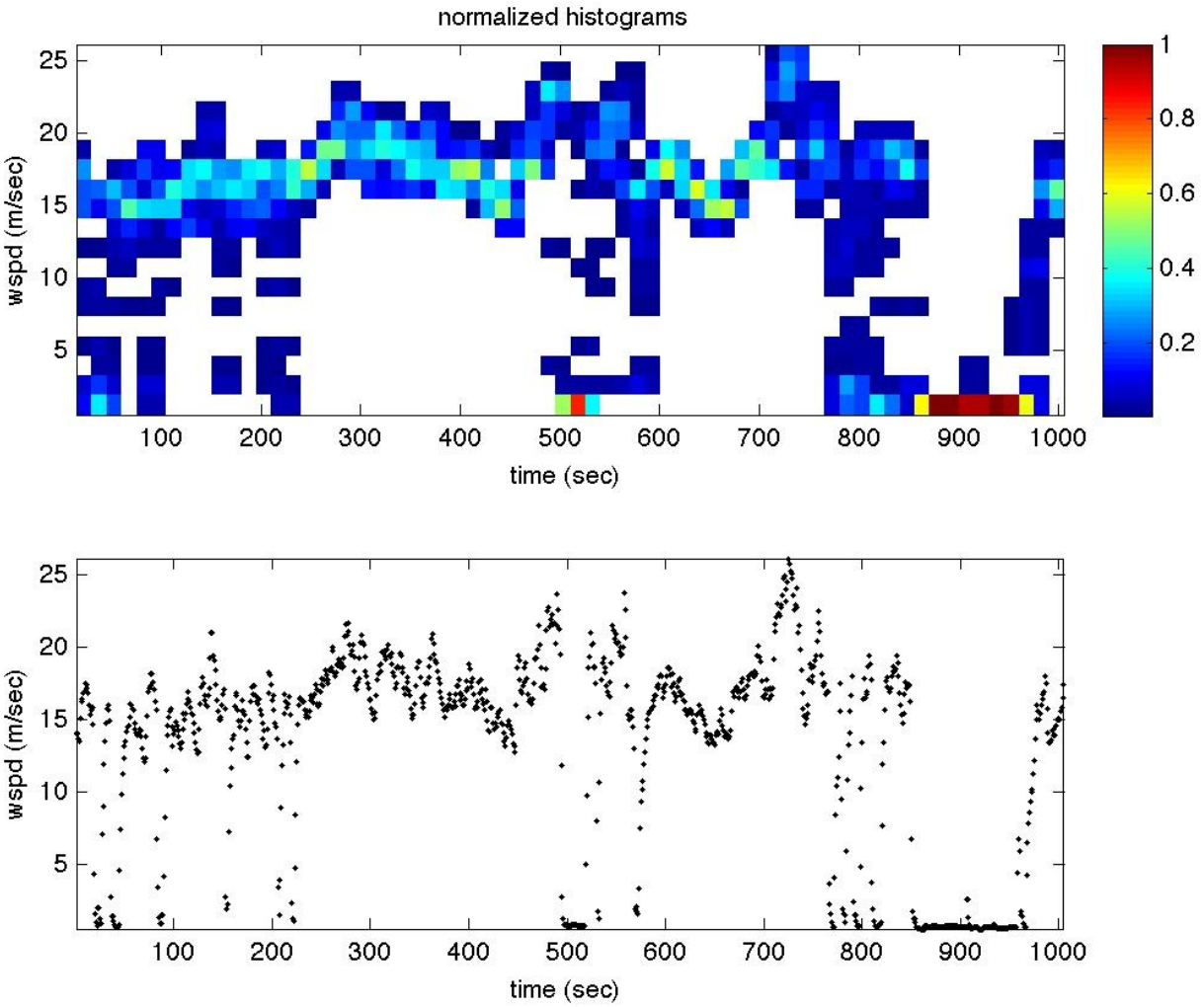
Density Map (overlapping tiles)





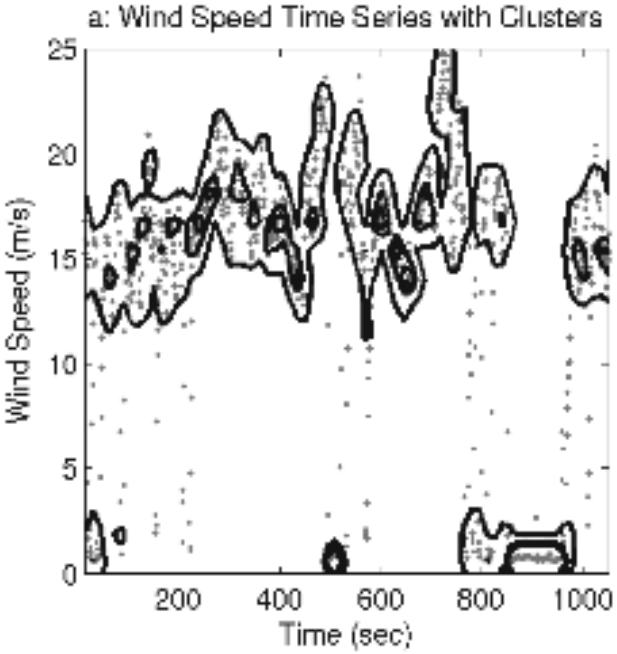
- Calculate density in lag domain with overlapping tiles
- normalize by total number of points

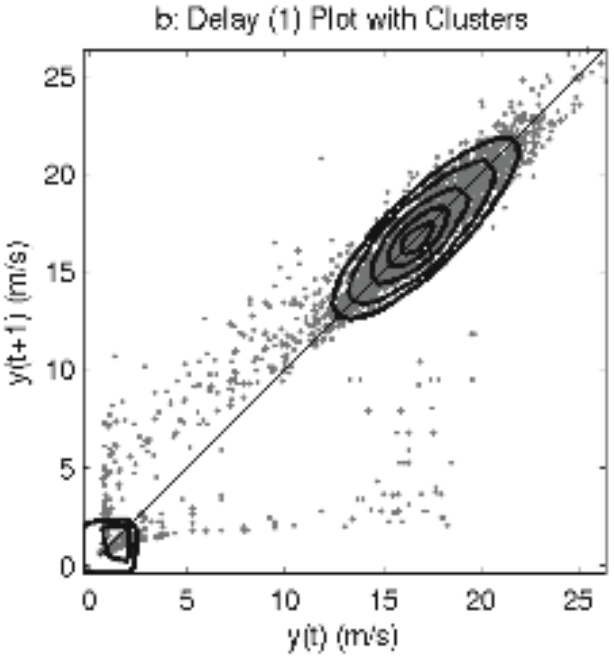
Density Map (stacked histograms)



- Calculate histogram of n data points
- Overlap moving window
- normalize histogram by total number of points

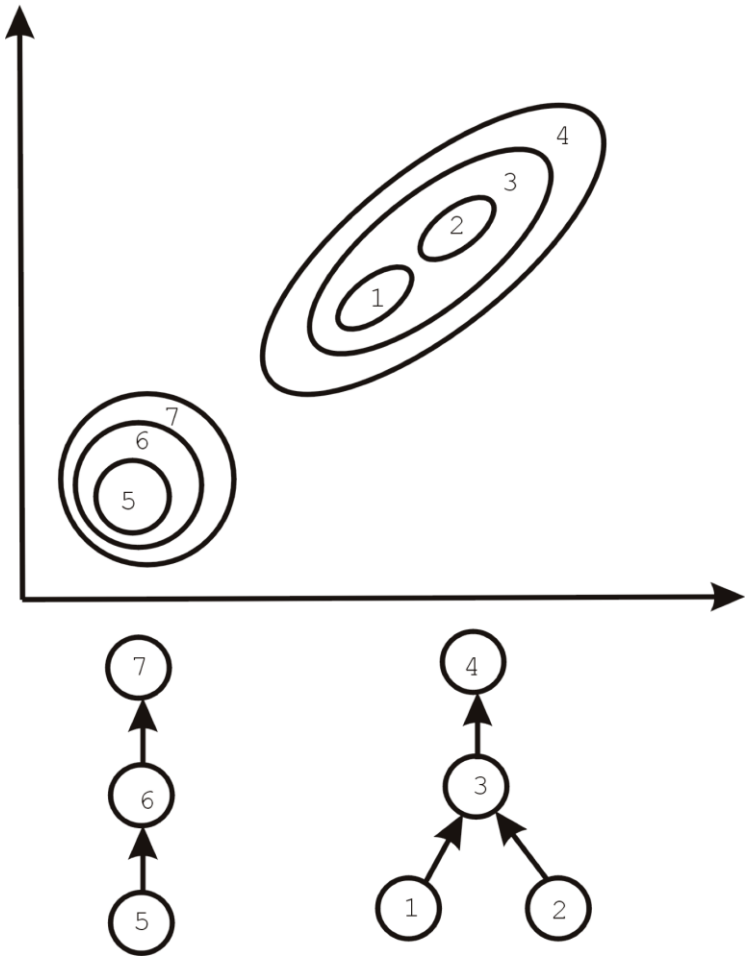
Density Cluster from multiple thresholds





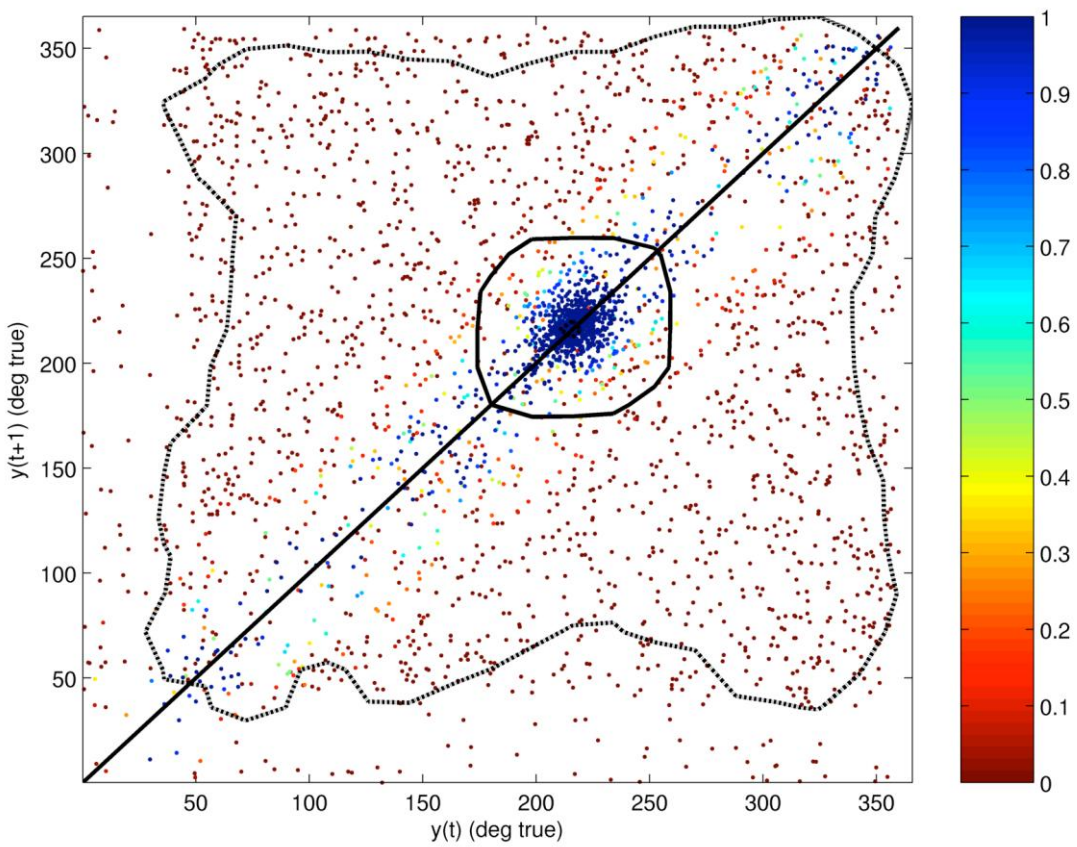
- find clusters in density maps using multiple thresholds
- sequence of nested coincident clusters

Cluster Graphs (delay space)



- Construct a graph that represents coincident clusters (not necessarily binary trees)
- Calculate the convexity of the clusters
- Select largest cluster in each tree that has a convexity above a threshold
- can build a graph of the clusters in the time domain

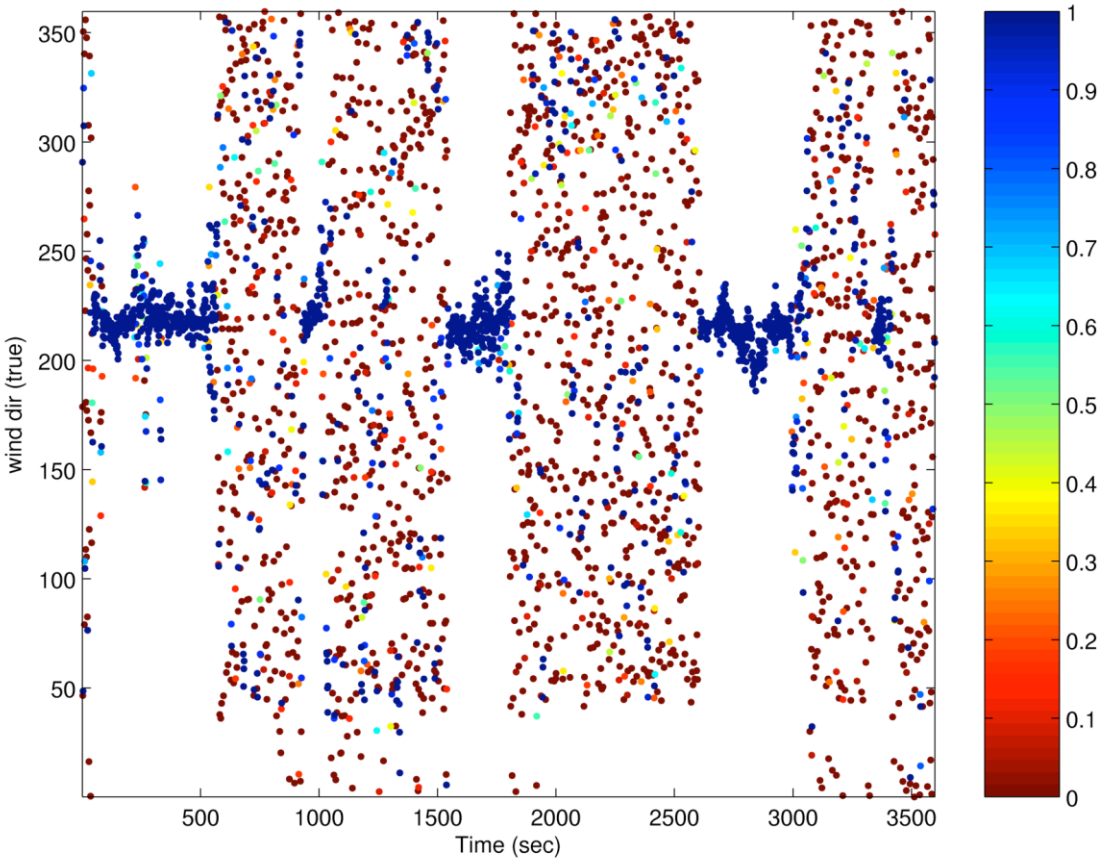
Distance Score (delay space)



- Calculate sample deviation of data inside optimal lag cluster

- calculate score based on distance from the line y=x normalized by the sample deviation

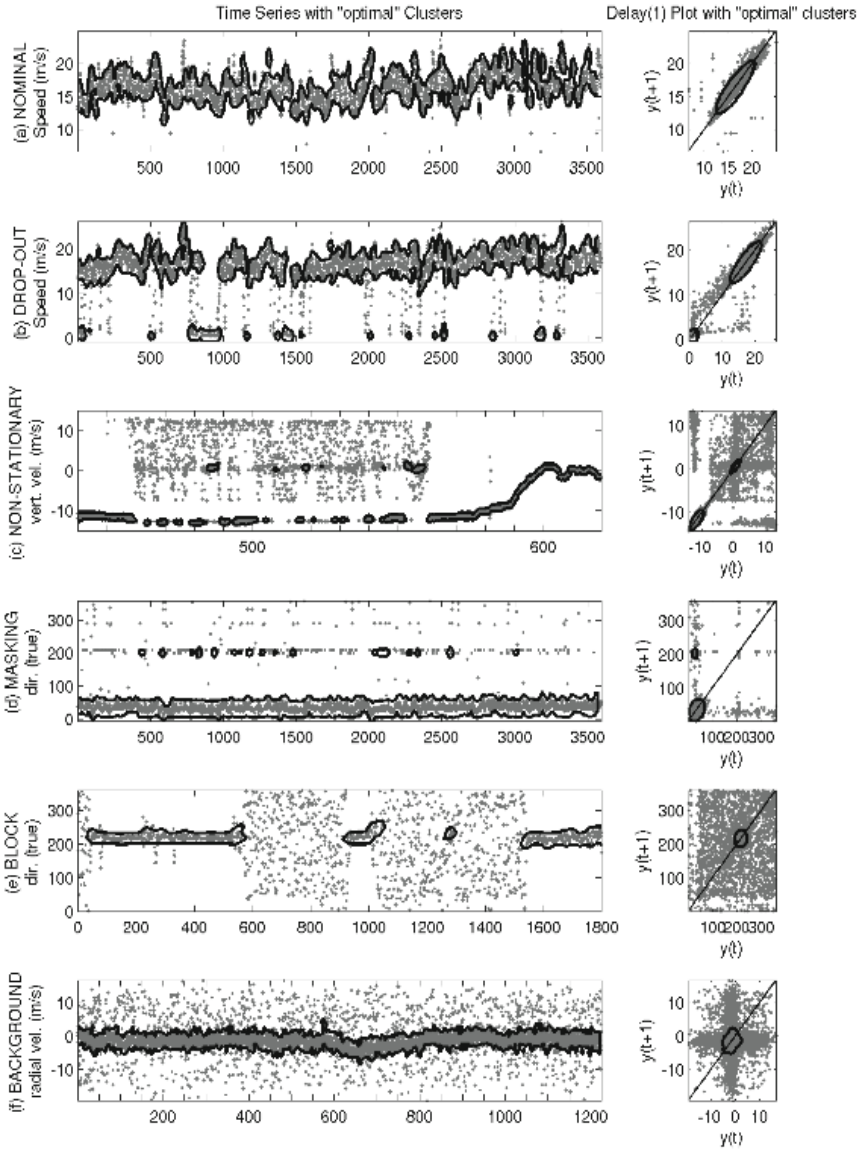
Distance score in time domain



- The distance score in the time domain is the geometric mean of the distance score for (i-1,i) and (i,i+1)

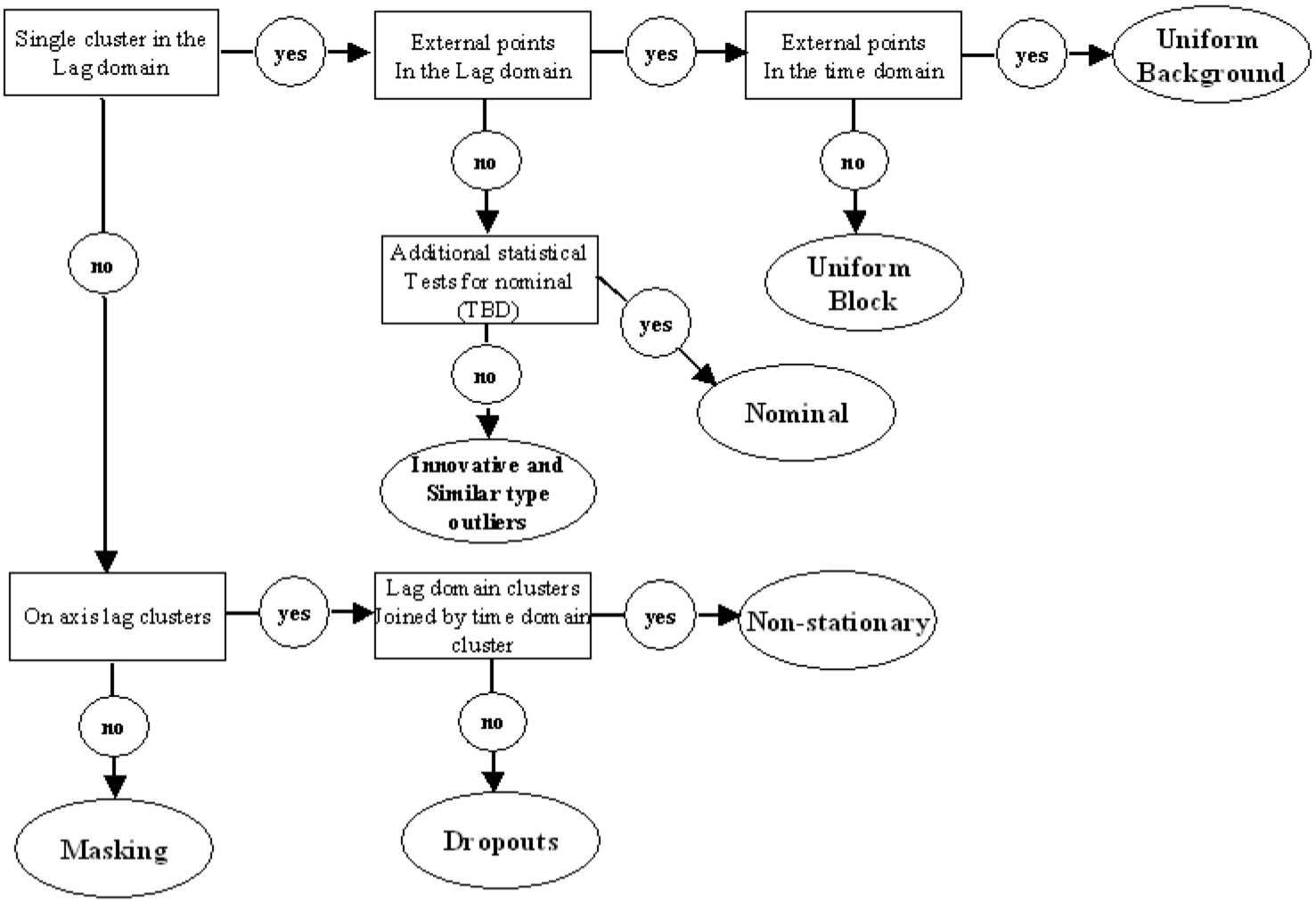
- Optimal clusters are found in the time domain by finding the lowest water level cluster with a predominate number of point with a high distance score.

Cases studied with optimal clusters

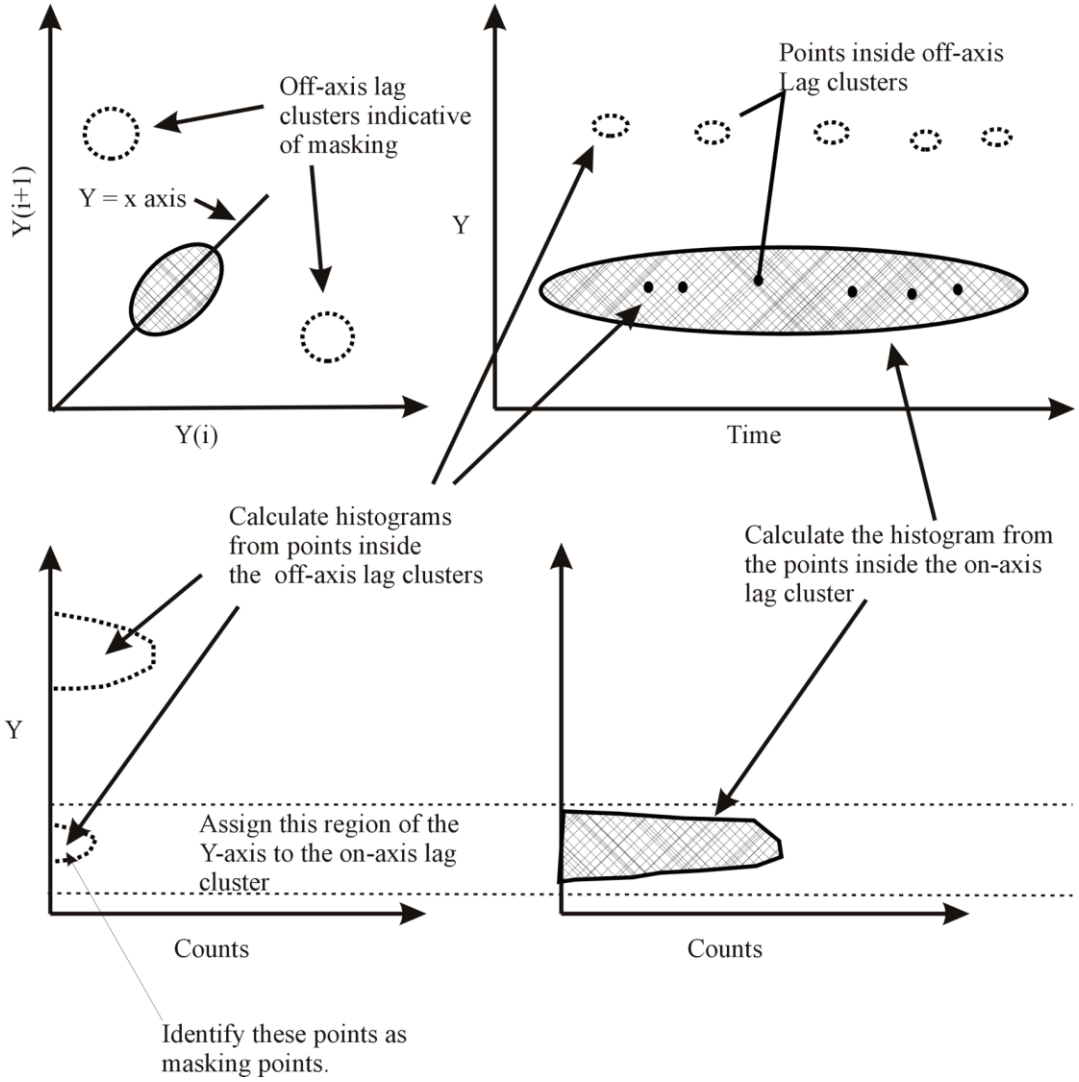


- optimal clusters in time domain and delay space
- build a “feature” in the time domain from optimal clusters in the time domain
- drop-out and non-stationary cases have two optimal clusters in delay space, but distinct representations in the time domain.
- nominal, block and uniform cases have single clusters in delay space but distinct representations in the time domain

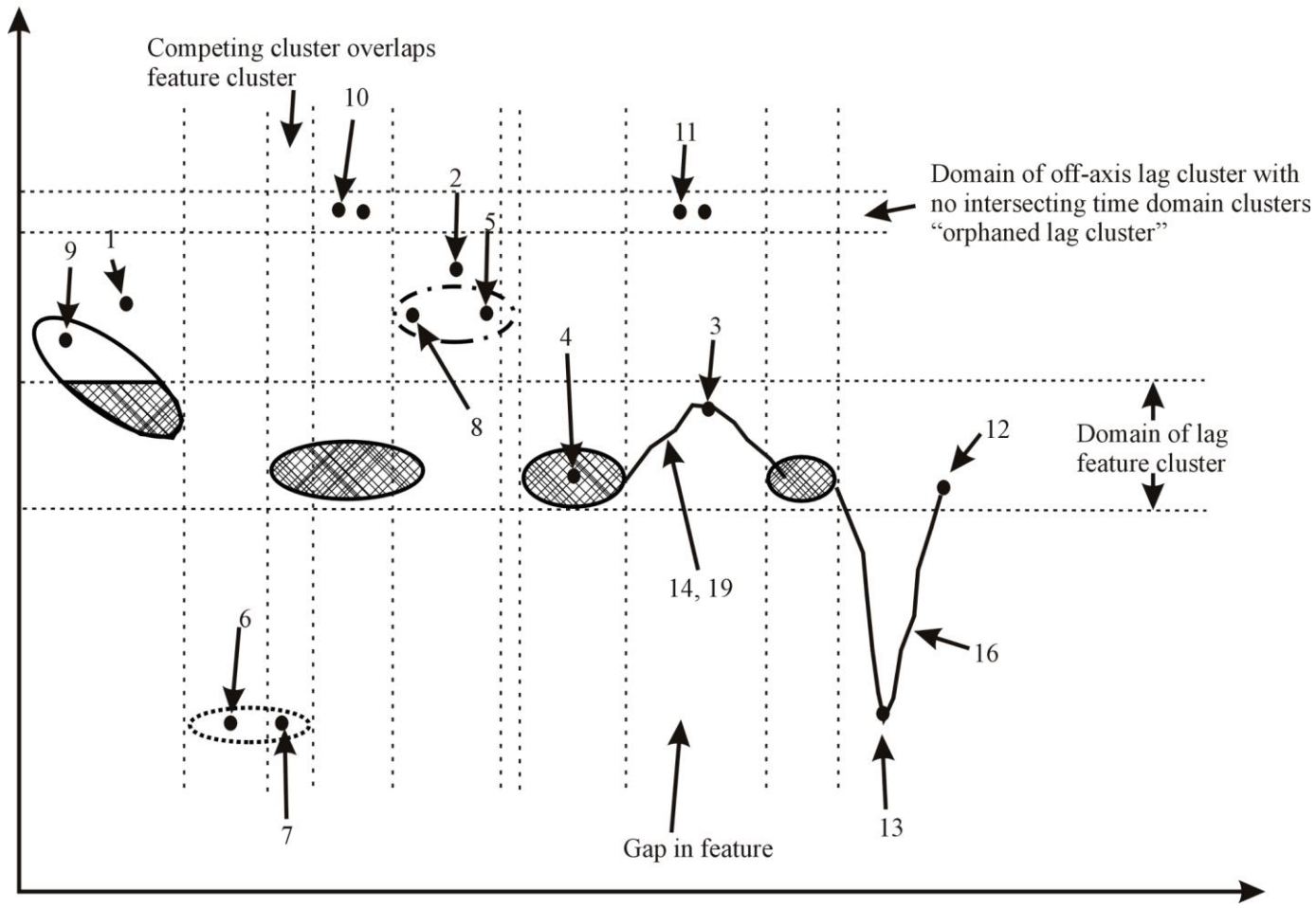
Failure mode decision tree



Resolving masking

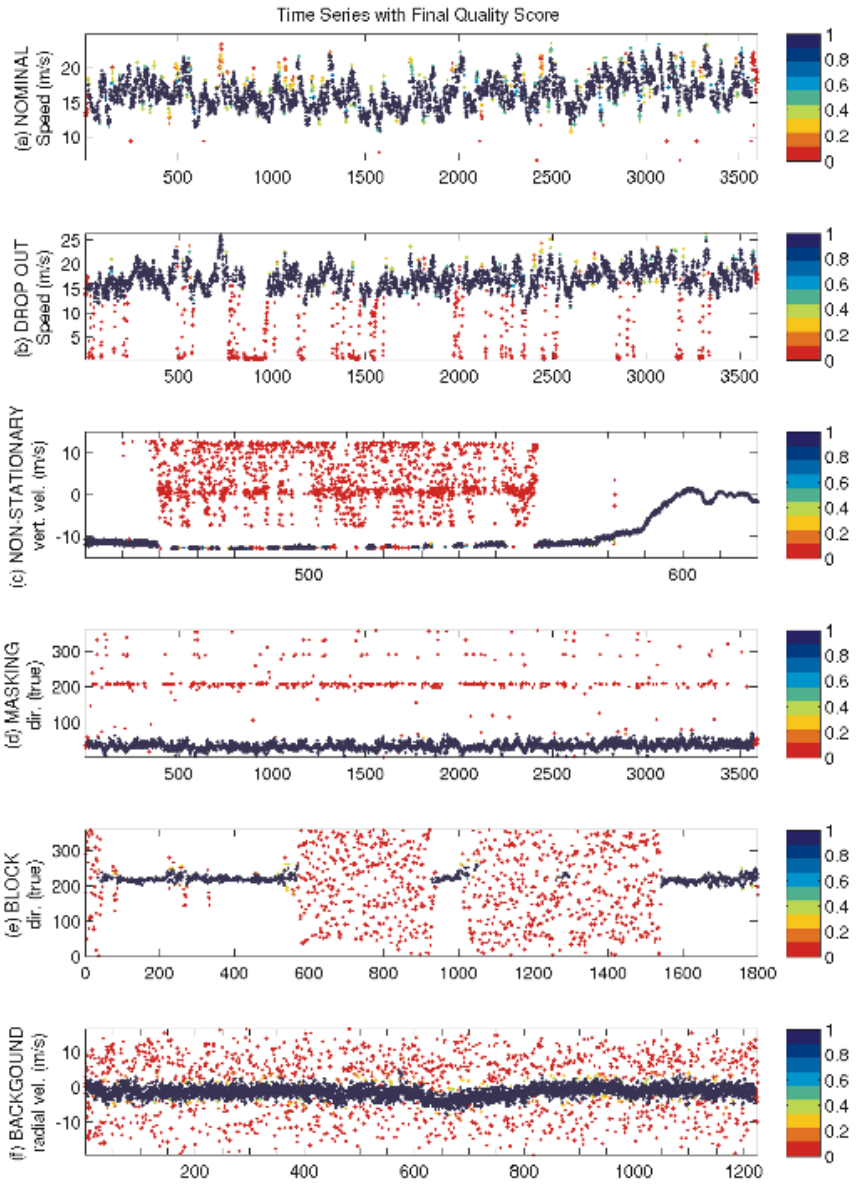


Types of points



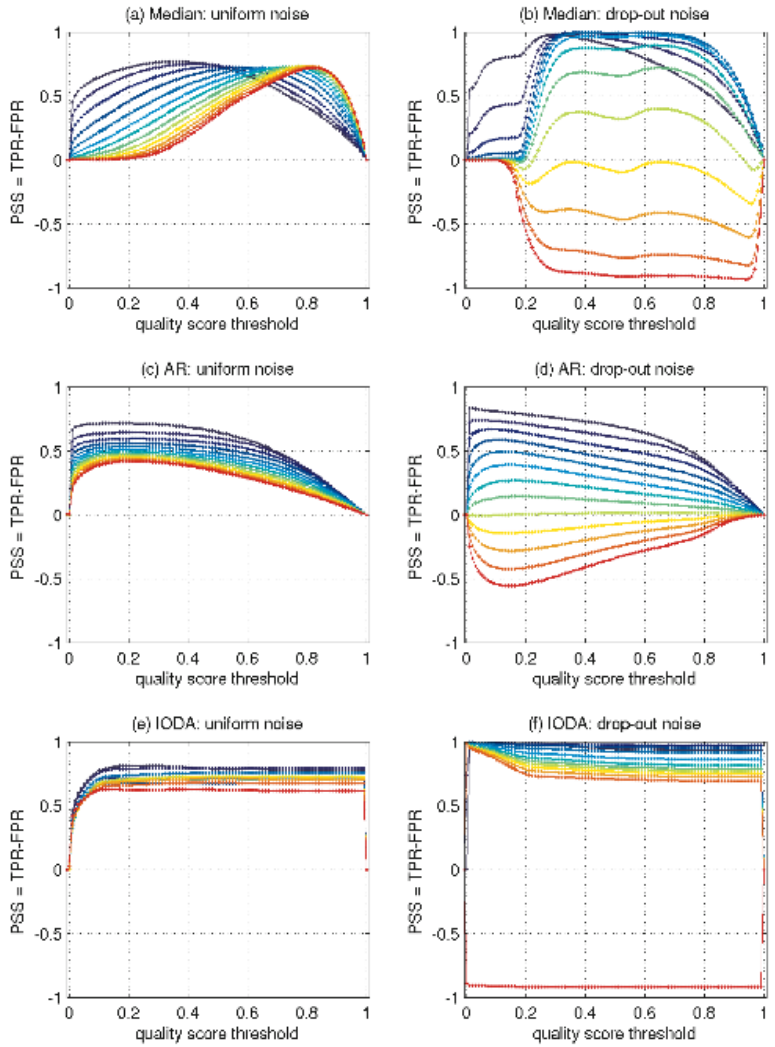
- points can be assigned a type that is a function of their location relative to the feature and optimal clusters

Final Confidence



- assign a confidence to a point given the failure mode and the point type
- Heuristically, the algorithm matches how a human might score the data
- In the non-stationary case the algorithm correctly finds the auto-correlated data but in reality this data is probably bad
- auto-correlation is not always enough to correctly classify the data

Performance



- two simulation scenarios, uniform noise and drop-out
- skill score as a function of confidence threshold and percent bad data
- There exists a single threshold such that IODA performs well for all scenarios.

References

- Ban, A.I. and S.G. Gal, 2002: *Defects of Properties in Mathematics; Quantitative Characterizations*. World Scientific, 364 pp.
- Barnett, V. and T. Lewis, 1977: *Outliers in Statistical Data, 3rd ed*. John Wiley and Sons, 604 pp.
- Bohm, C., K Kailing, P. Kroger, and A Zimek, 2004: Computing Clusters of Correlation connected Objects. Proceedings, *Int. Conf. on Management of Data*, Pairs, France, SIGMOD, 455-466.
- Box, G.E.P. and G.M. Jenkins, 1970: *Time series Analysis: Forecasting and Control* Holden-Day, 784 pp.
- Chen, J. R. 2007: Useful Clustering Outcomes from Meaningful Time Series Clustering, Proceedings, *Sixth Australasian data Mining Conference*, Gold coast, Australia, CRPIT, 101-109
- Chi, Z.,Y. Hong, and P. Tuan, 1996: *Fuzzy Algorithms: with Applications to Image Processing and Pattern Recognition.* Wold Scientific, 225 pp.
- Cornman, L.B.,R.K. Goodrich,C.S. Morse, and W.L. Ecklund,1998: A Fuzzy Logic Method for Improved Moment Estimation From Doppler Spectra, *Journal of Atmospheric And Oceanic Technology*, **15 No 6**, American Meteorological Society, Boston, , 1287-1305
- Frehlich, Rod, S. Hannon, and S. Henderson, 1994: "Performance of a 2-mu m coherent Doppler Lidar for wind measurements," *Journal of Atmospheric And Oceanic Technology*, 11, 1517-1528.
- Hartigan, J.A., 1975: Clustering Algorithms, Wiley, 366 pp.
- Jolliffe, I.T., 2002: Principal *Component Analysis, 2nd ed*. Springer, 502 pp.
- Luenberger, D.G., 1984: *Linear and Nonlinear Programming 2nd ed*. Addison-Wesley, 546 pp.
- Priestley, M.B., 1981: *Spectral Analysis and Time series*, Academic Press, 890 pp.
- Rosenstein M.T. and P.R. Cohen, 1998: Concepts for Time series, Proceedings, *Fifteenth National Conference on Artificial Intelligence*, Madison, WI, AAAI, 739-745
- Weekley, R.A., R.K. Goodrich, and L.B. Cornman, 2003: Fuzzy Image Processing Applied to Time series Analysis, Preprint*, 3rd Conf. Artificial Intelligence Applications to the Environmental Sci*., Long Beach, CA, Amer. Metero. Soc., CD-ROM, 4.3
- Wilks D. J., 2006: *Statistical Methods in the Atmospheric Sciences 2nd ed*., Academic Press, 617 pp.
- Wishart D., 1969: Mode Analysis: A Generalization of Nearest Neighbor Which reduces chaining effects. *Numerical Taxonomy*. (A. J. Cole, Ed.), Academic Press, 328 pp.
- R. A. Weekley, R. K. Goodrich, and L. B. Cornman, “An Algorithm for Classification and Outlier Detection of Time-Series Data,” *Journal of Atmospheric and Oceanic Technology*, vol. 27, no. 1, pp. 94–107, 2010.