Identifying Storms in Noisy Radio Frequency Data via Satellite: an Application of Density Estimation and Cluster Analysis

Tom Burr, Angela Mielke
Safeguards Systems Group
Los Alamos National Laboratory

Abram Jacobson
Space and Atmospheric Science Group
Los Alamos National Laboratory

Presented at:
U.S. Army Conference on Applied Statistics, Santa Fe, New Mexico
October 24-26, 2001
LA-UR-01-4874.

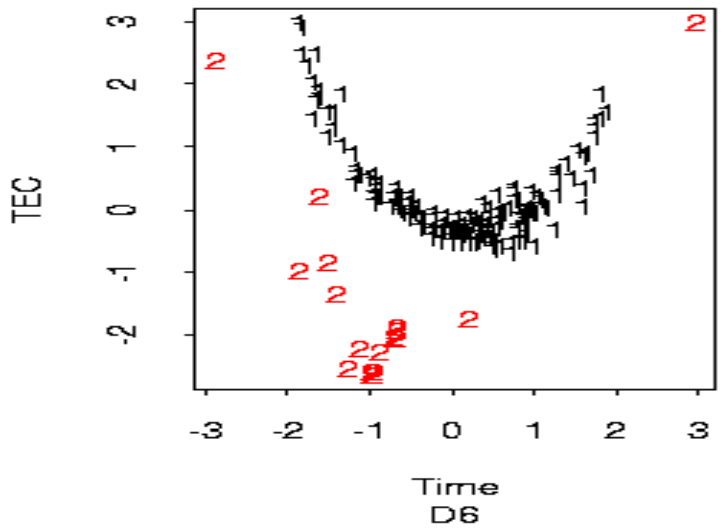
Abstract

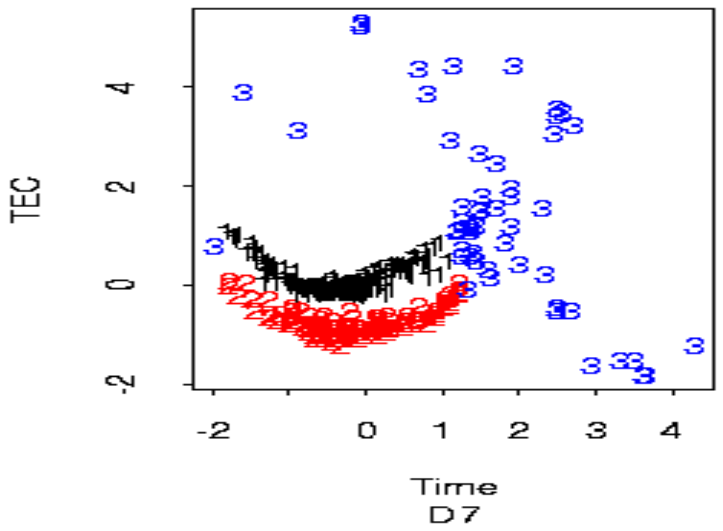
Abstract

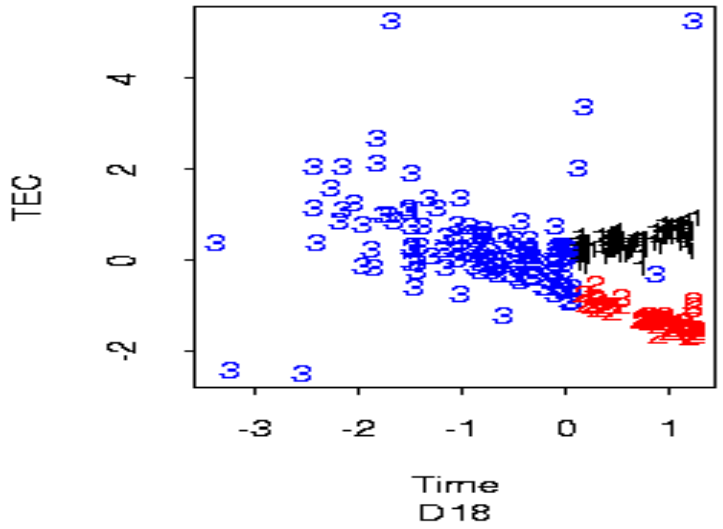
The FORTE (Fast On-Orbit Recording of Transient Events) satellite collects records of radio frequency events that exceed a threshold. Here we consider processed (dechirped) data from storm-like micro events. Each data point is the total electron count (TEC) accumulated over 400 microseconds. Each data record contains approximately 100 to 400 micro events. Some data records contain well-defined storm events which consist of many data points (micro events) in a specialized cluster. We present a method involving noise rejection and cluster analysis to identify well-defined storms from the data records. We first remove noise using density estimation and then apply hierarchical clustering to the higher-density micro events. For each identified cluster of micro events, we fit TEC as a quadratic function of time (a quadratic shape is anticipated from atmospheric physics), and find more micro events that belong to the cluster using a careful extrapolation. The overall performance of finding each storm and identifying which micro events belong to which storm is assessed by comparing our results to test data produced by a human analyst.

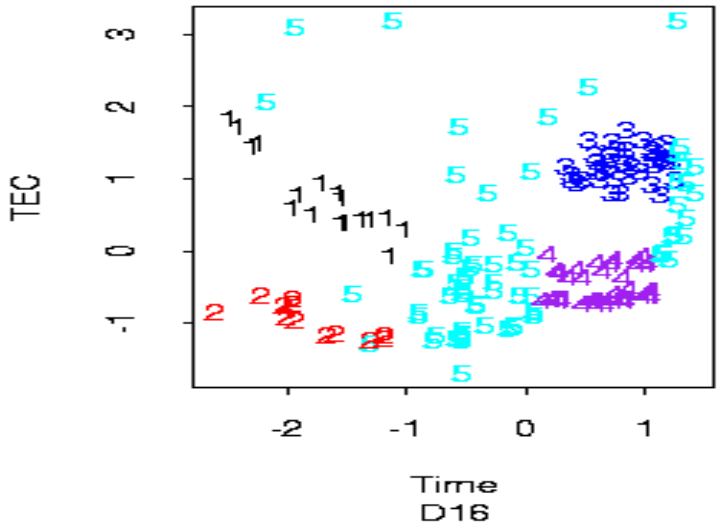
Examples - data sets 6, 7, 18, 16

TEC and time are transformed to unit variance, 0 mean









Outline

Problem Statement: find smiles in TEC vs time

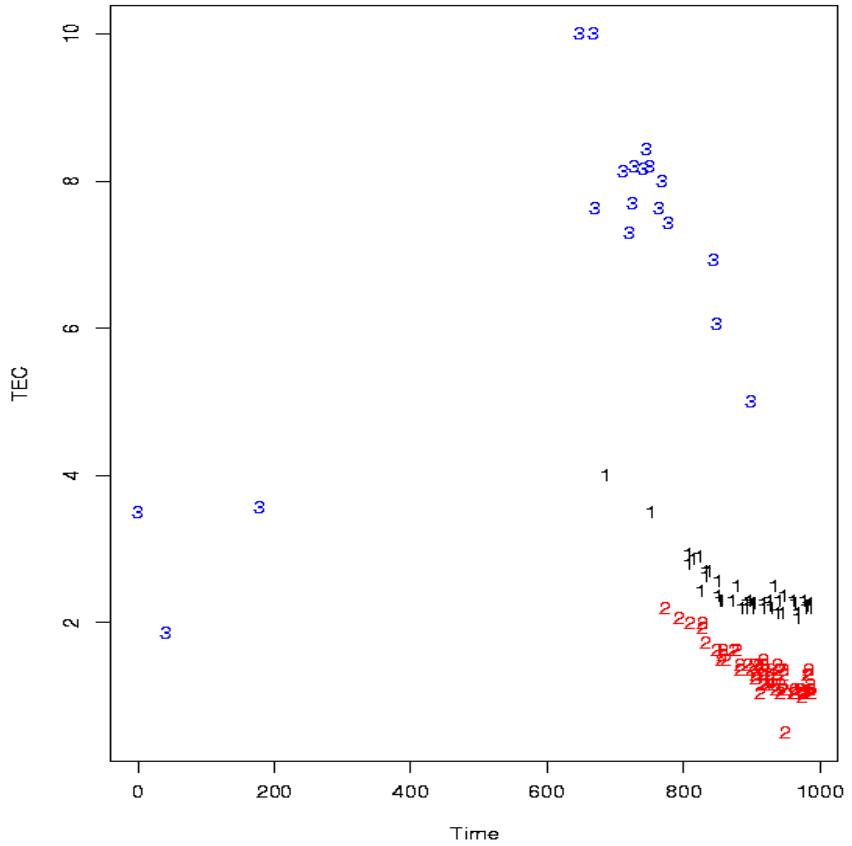
want small false positive rate

Methods

Performance Results

Simulation Studies

Summary/Future



D1

Problem Statment

Find smiles in TEC vs time

-each data point is the total electron count over 400 μs

-each data record contains 100 to 400 micro events

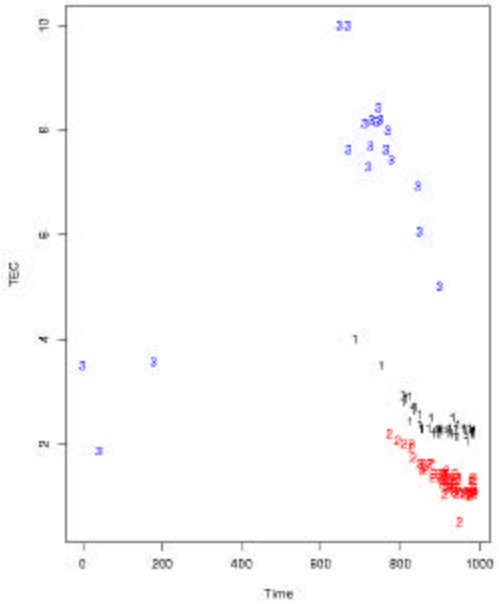
-most data records contain storms which consist of many data points (micro events) in a specialized cluster

Anticipated cluster shape: bowl (“smile” or region of it)

-total record time approx 800 seconds, or 1 pass of satellite

Methods (in Splus now, PERL next)

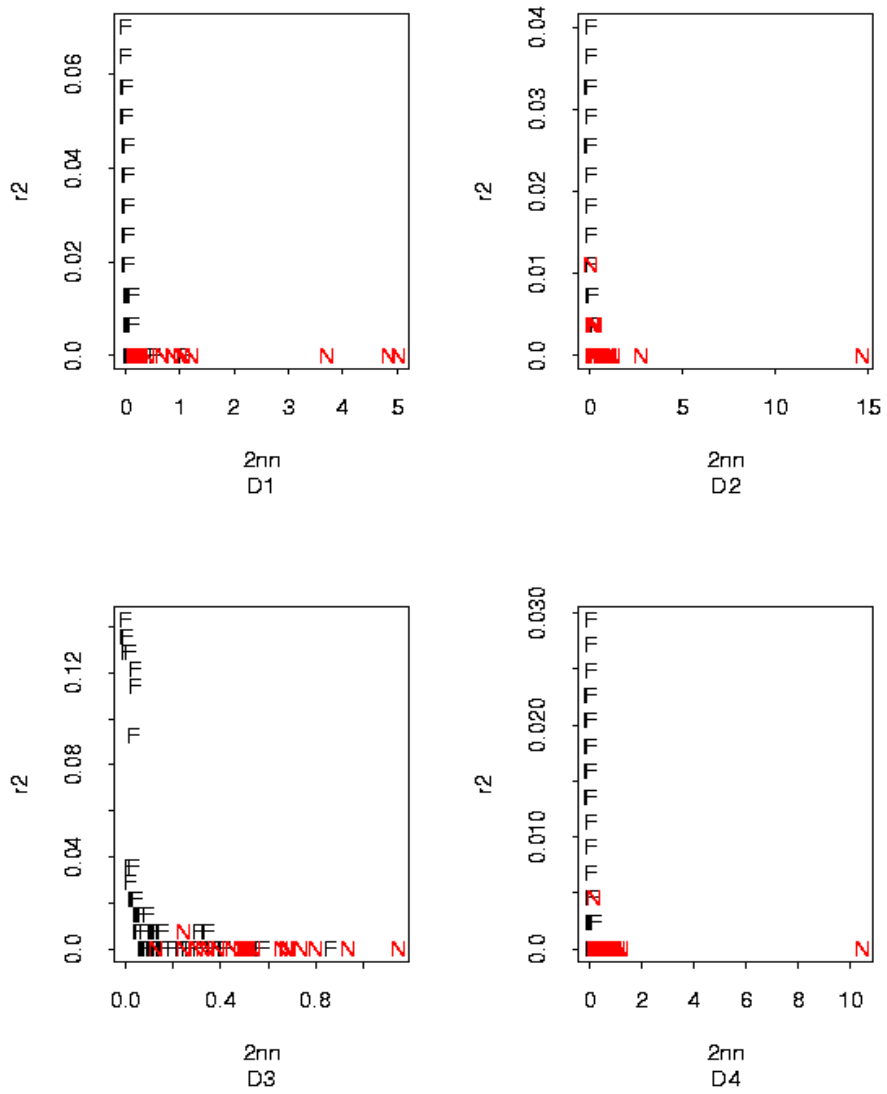
- Noise -- remove via simple density estimation
- Hierarchical clustering and experiment with:
  - metric
  - cutpoint using quantiles of distances
  - clustering method: long, thin clusters
  - rule for rejecting small clusters



- Compare to model-based clustering (Raftery and others)

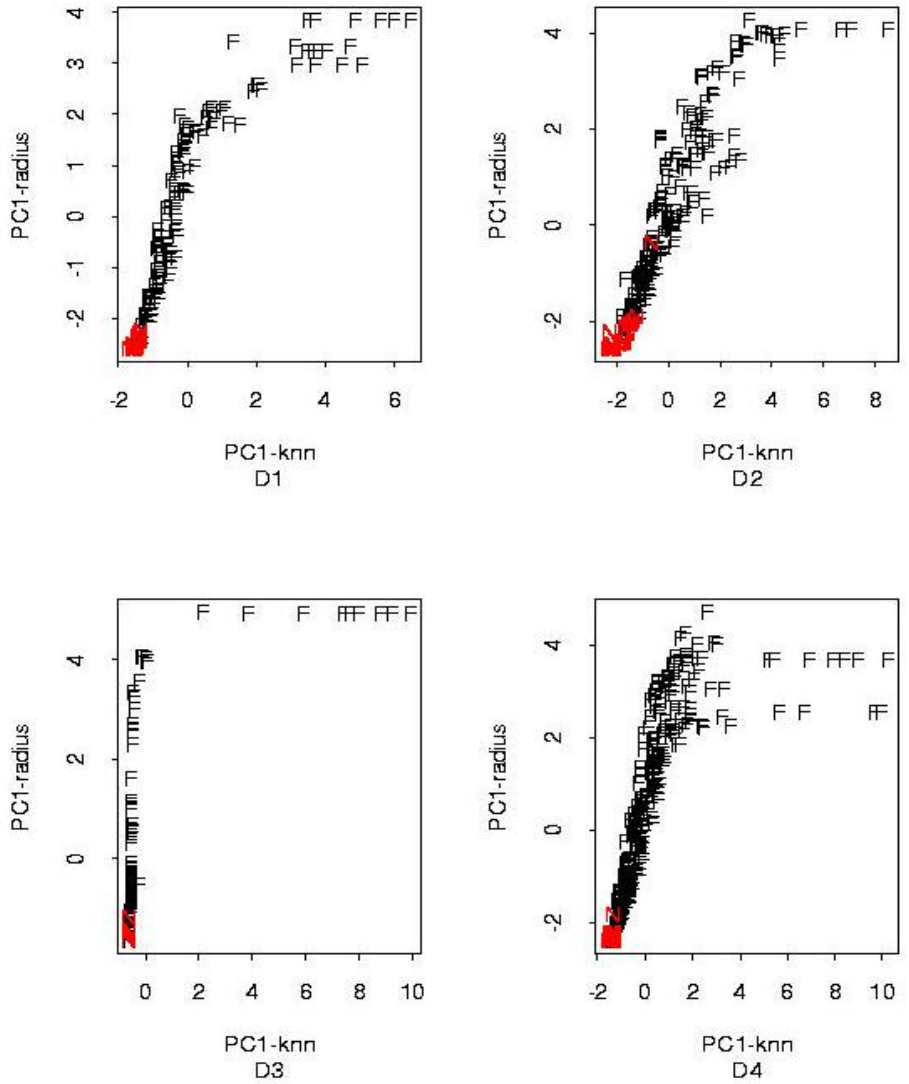
Ref: Stanford and Raftery 2000 IEEE Transactions on Pattern Analysis involved principal curve clustering with noise. CEM - PCC with BIC and model for likelihood using hierarchical clustering as key steps

Noise rejection performance: false positive and false negative rates



FP: 0.03 0.08 …. 0.33

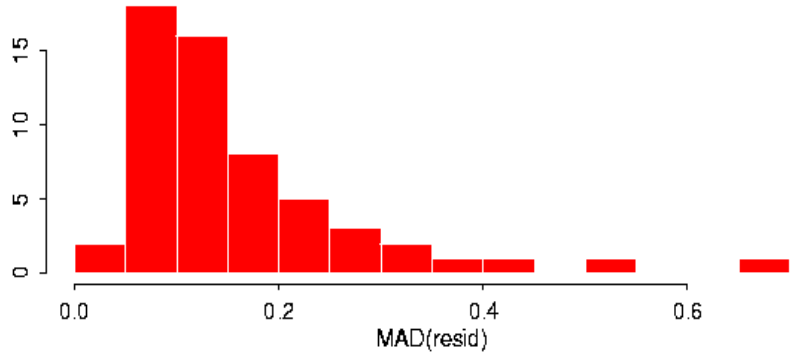
FN: 0.79 0.62 …. 0.36

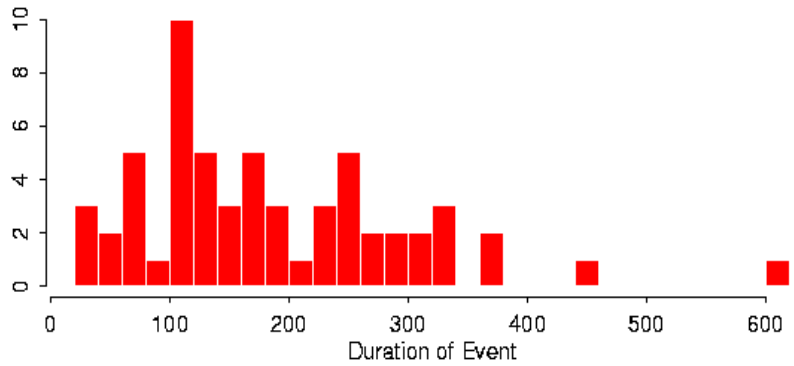


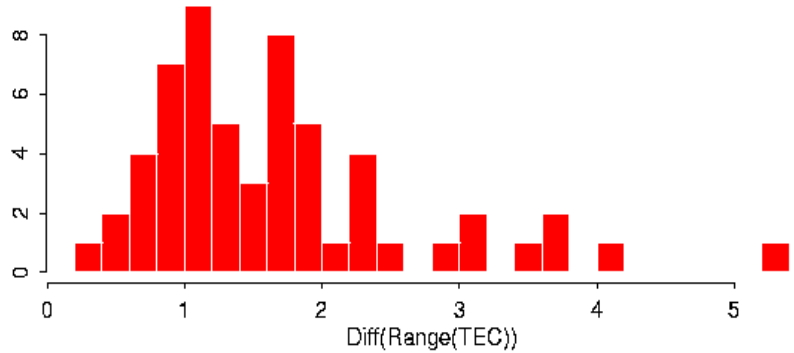
FP: 0.02 0.08 …. 0.32

FN: 0.79 0.61 …. 0.35

Cluster Features







Final QC check on any purported storm can require "reasonable values" for any subset of:

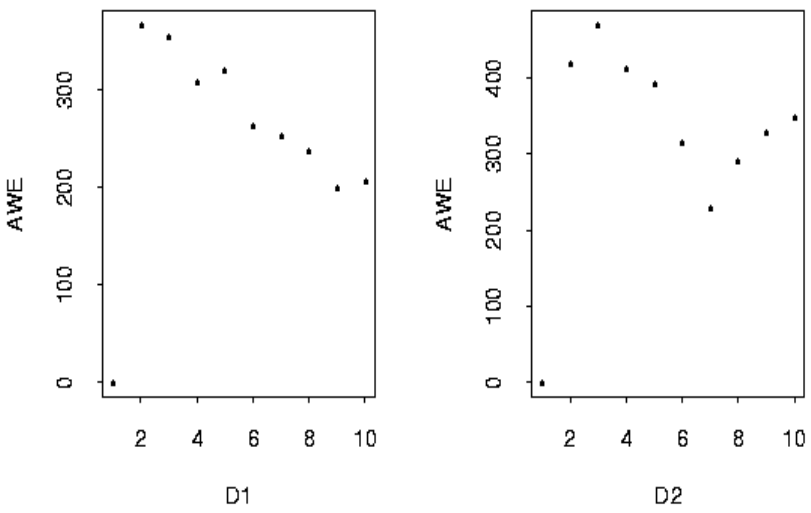
MAD(resid)

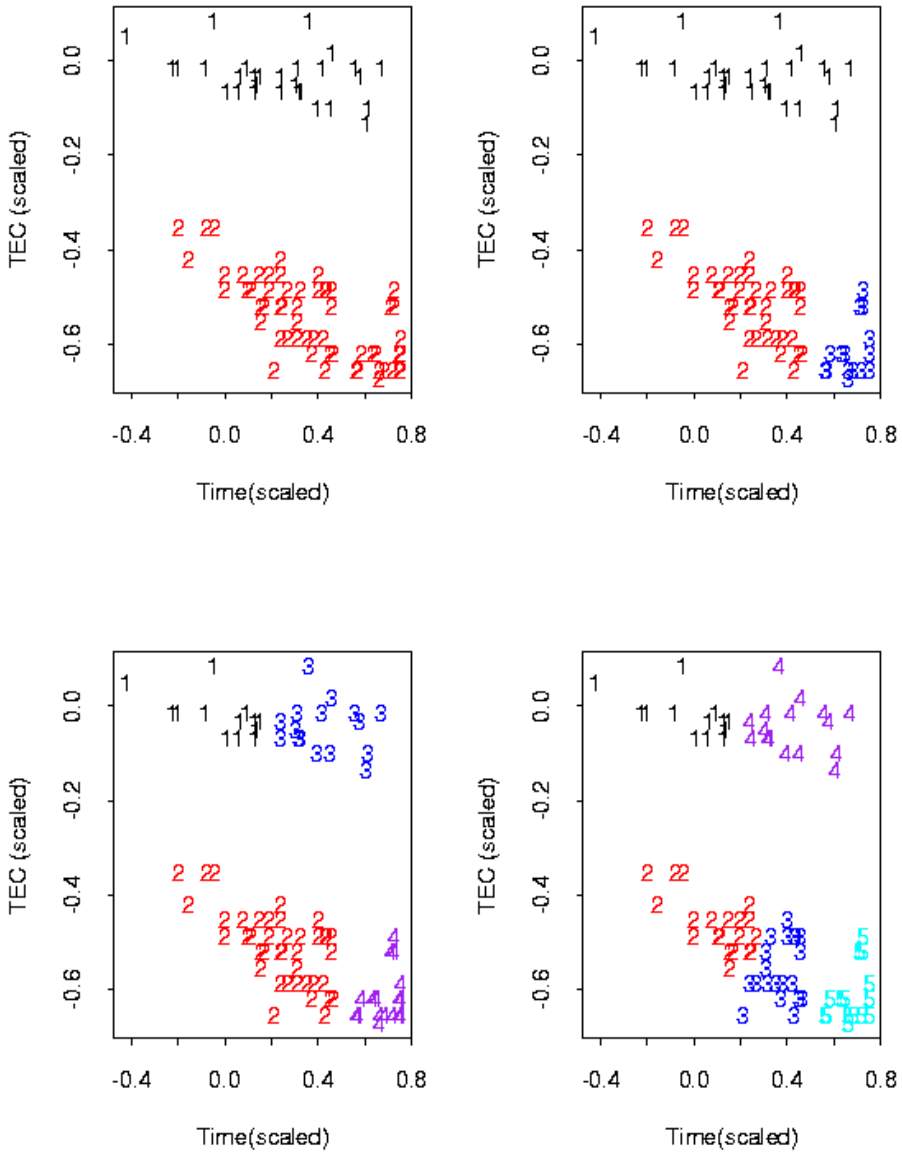
event duration

range(TEC)

And, concave up, not down

Model-based clustering: approximate (bayes) weight of evidence for candidate cluster numbers





Closest existing method: HPCC/CEM

Stanford and Raftery 2000 IEEE PAMI:

principal curve clustering with noise

Issues here:

Noise model, gaps, quadratic anticipated and OK

BIC: $2 \log(L(X|\theta)) - M \log(N)$

$L(X|\theta)$ depends on noise model, feature model

$M = k(DF + 2)$, DF= deg freedom in curve fit,

N points, k = no. of clusters

BIC for “true” versus current best, near best guesses:

selects true 10 times, best 9 times and near best 11 times

Future: experiment with likelihood for noise and feature

Methods

6 factors considered in a search:

2 noise rejection thresholds - relative (f1) and absolute (f2)

3 factors related to hierarchical clustering -

cut tree at some high percentile f3

reject clusters with small relative no of observations f4

reject clusters with small absolute no of observations f5

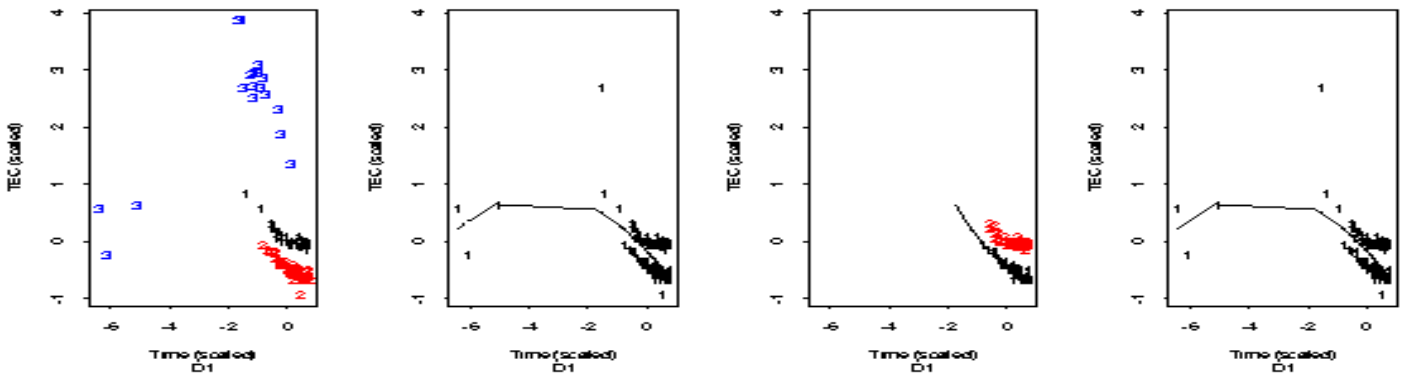
1 factor related to extrapolation from original cluster f6

what fraction of range of original cluster to allow extrapolation

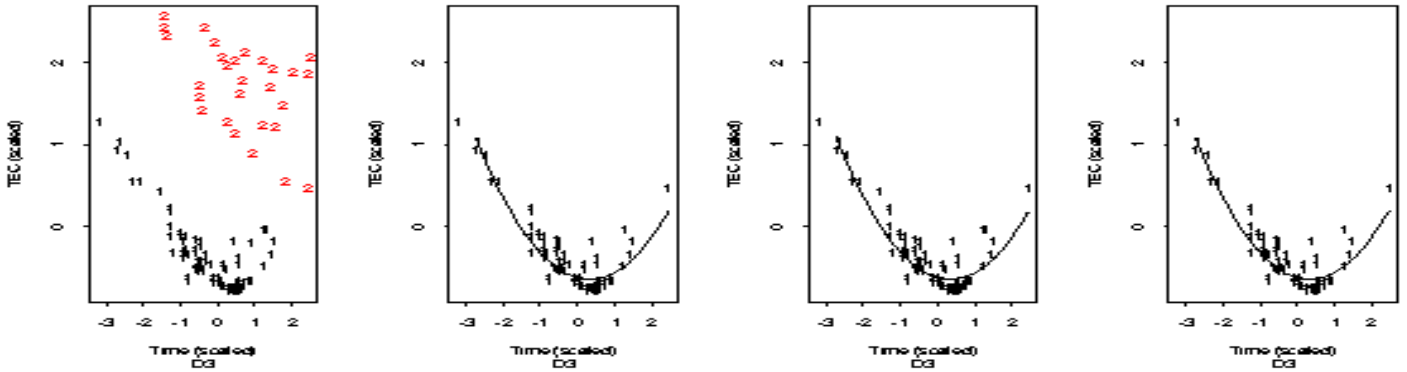
Search over 3^6 runs to find good values. Result: optimal values over 30 data sets approx same as those chosen from D1.

Examples - with noise - slide 16 has method description

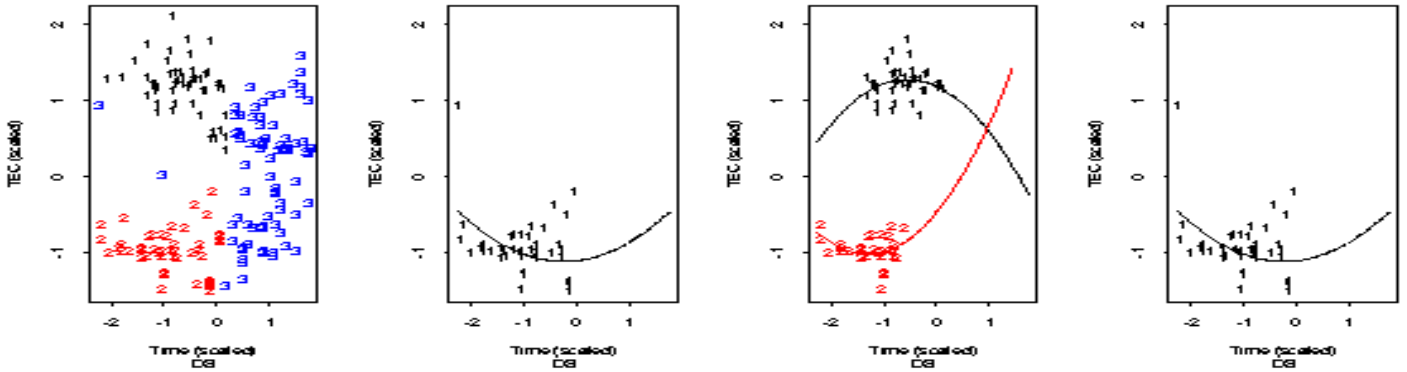
Data Clusters found: A and B parameters Final cluster



Find 1 of 2



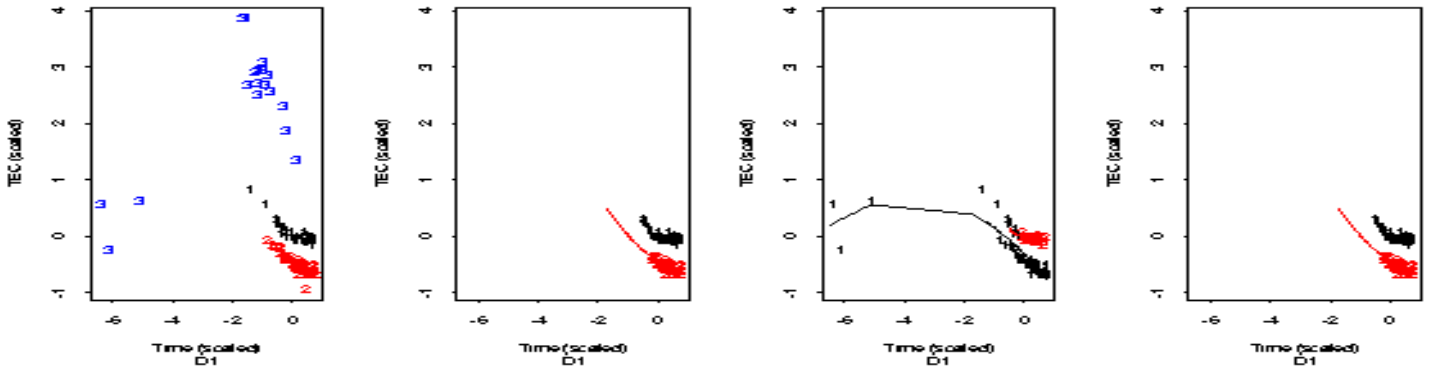
Find 1 of 2



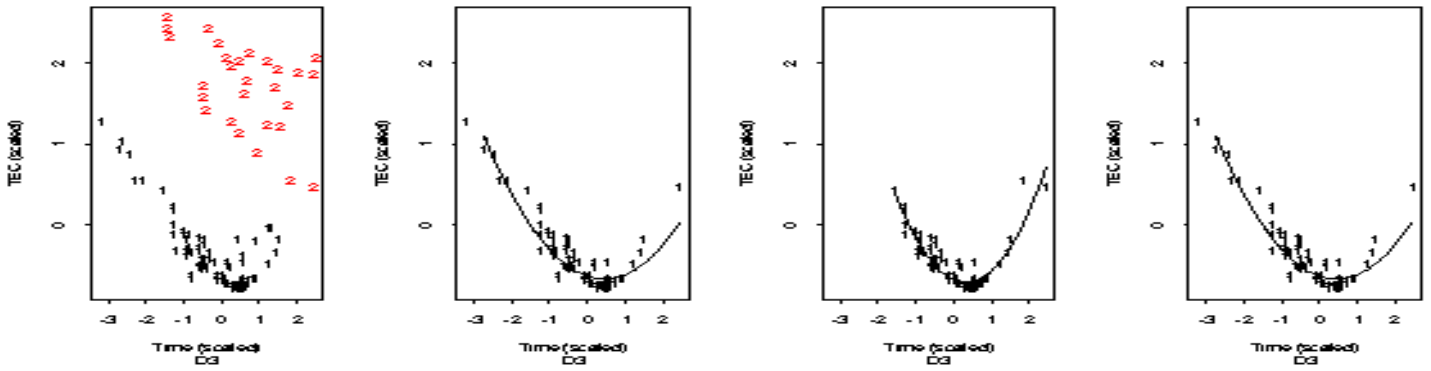
Find 1 of 2

Examples - some noise removal

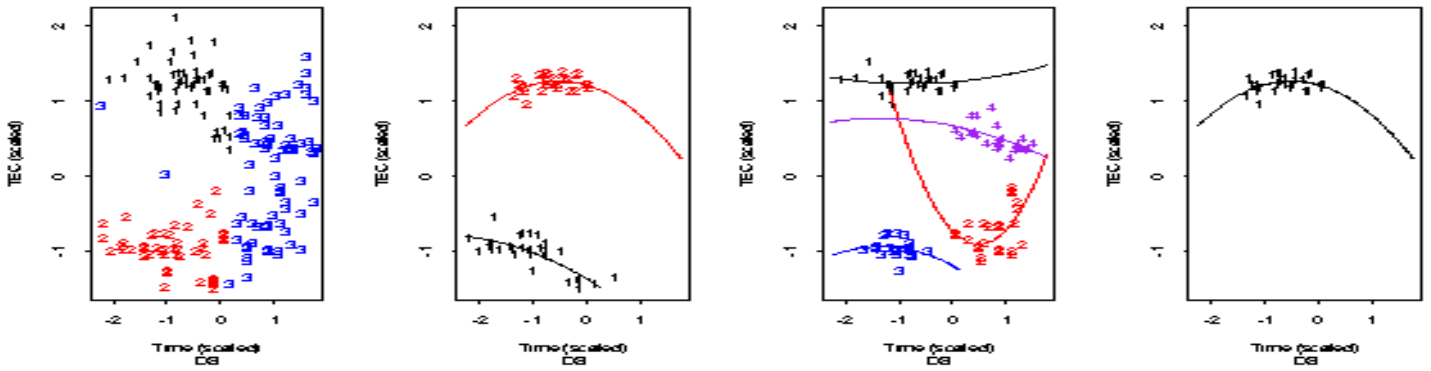
Data Clusters found: A and B parameters Final cluster



Find 2 of 2



Find 1 of 1

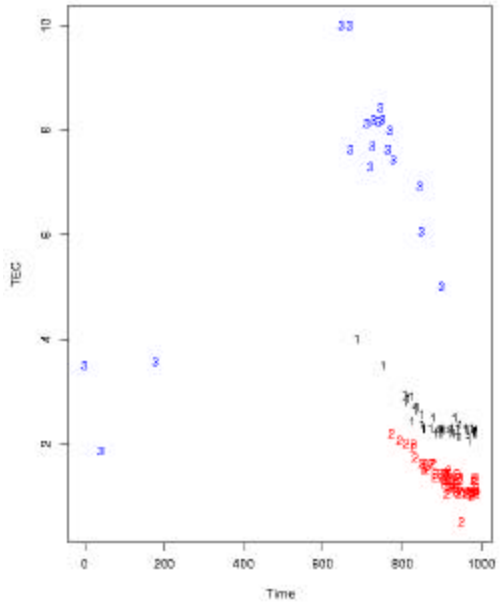


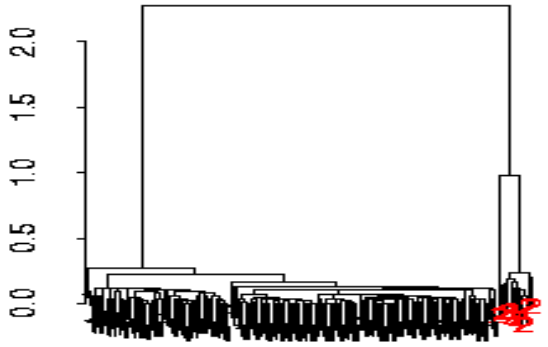
Find 1 of 2

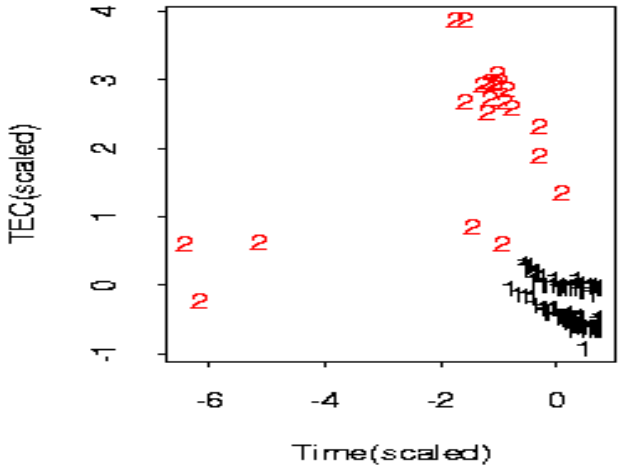
Example - with/without noise removal

Top:

no noise removal

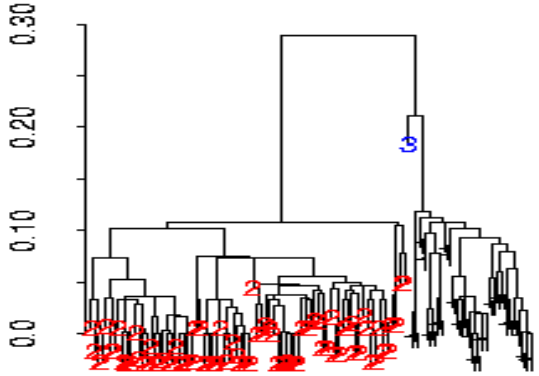


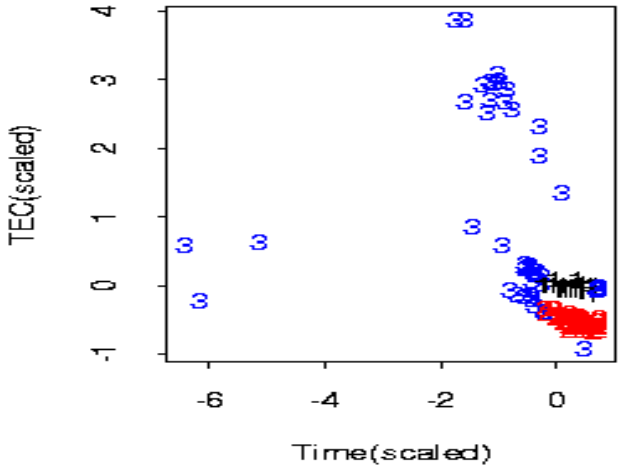


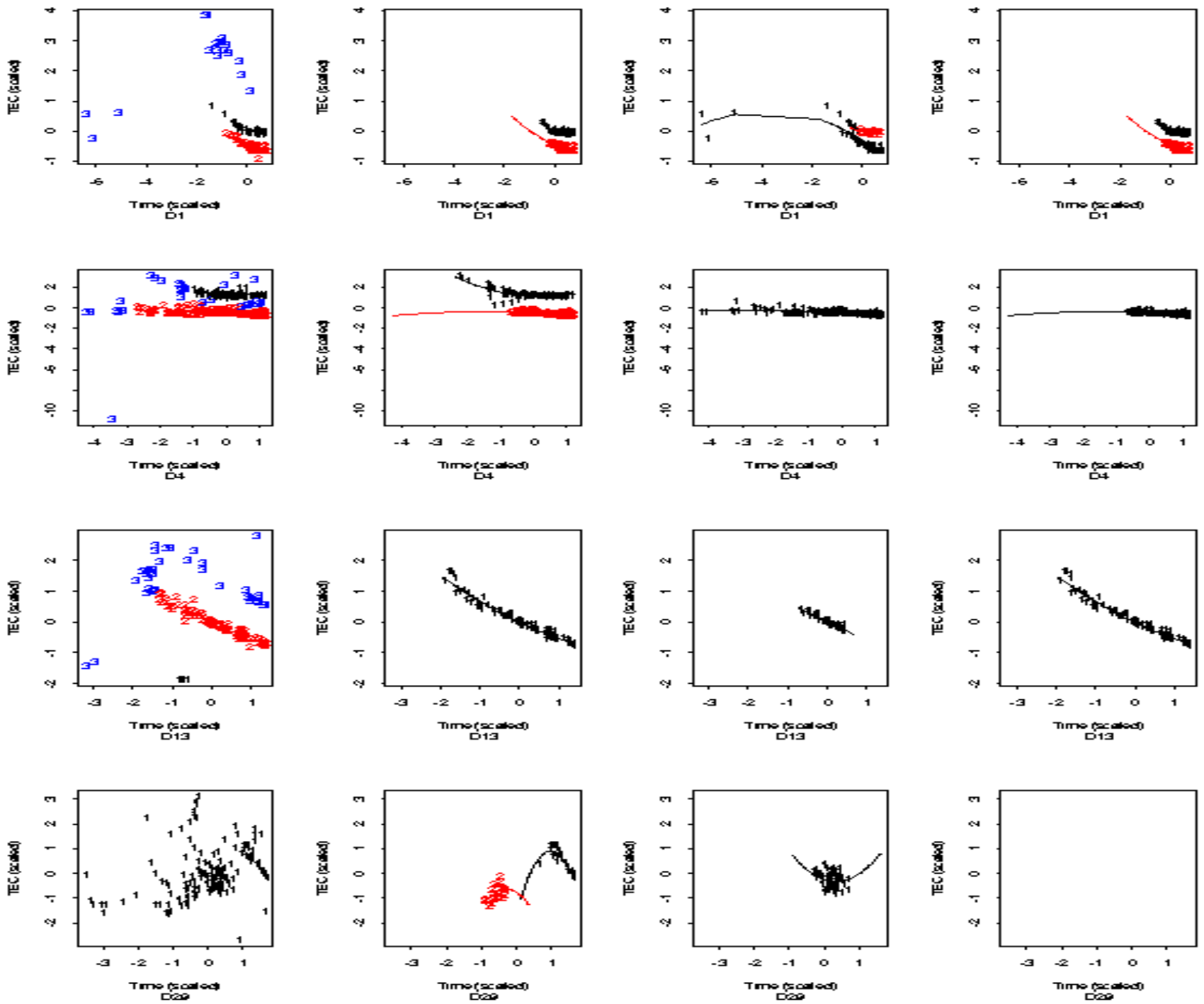


Bottom:

some noise removal







Find 2 of 2

Find 1 of 2

Find 1 of 2

Find 0 of 0

Method 1

Using values chosen from D1, and one set of nearby values:

1. reject noise
2. cluster result A with optimal values, B with near optimal
3. For each cluster, extrapolate using quadratic fit and “zone of ownership” to avoid ambiguous points.
4. Compare A and B results. For each cluster in A that is confirmed in B, accept cluster as a storm.

Other methods* informally evaluated, but results for 1 are:

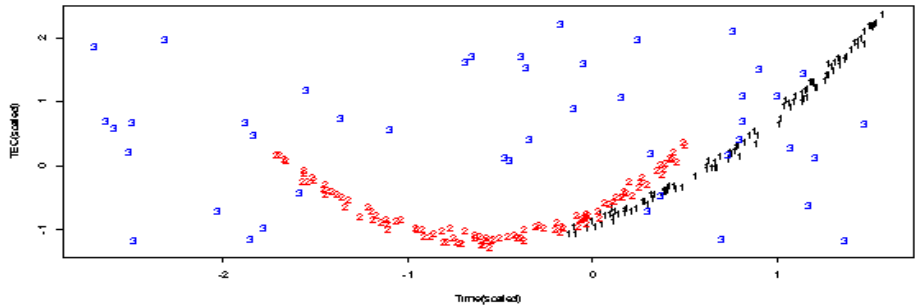
false positives: 0 (found 24 of 59) false negatives: 35/59

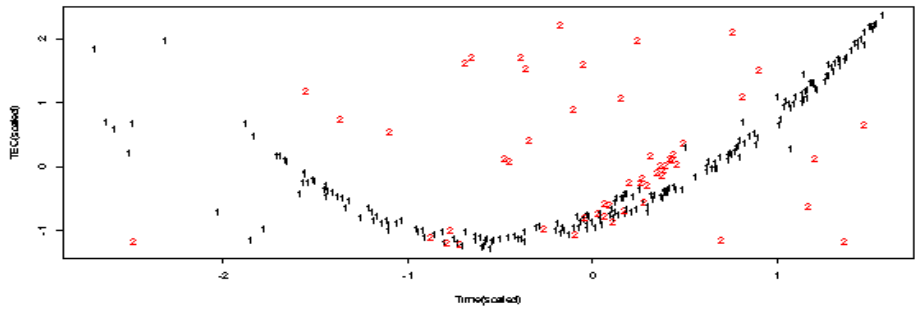
false positive rate: 0.21 false negative rate: 0.09

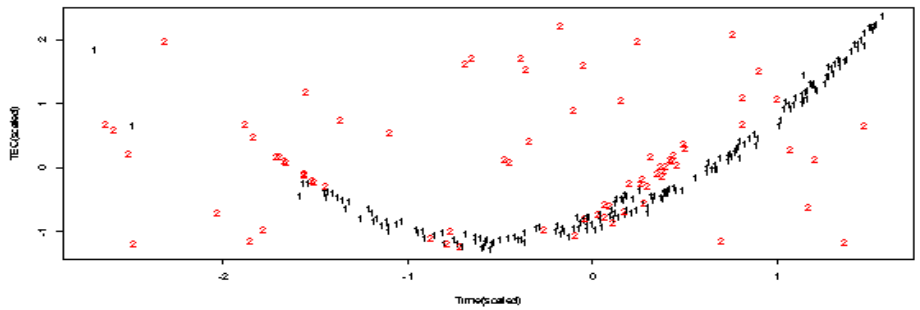
* Example: BIC as in Stanford and Raftery 2000 evaluated for true and estimates not yet working well, issue is likelihood.

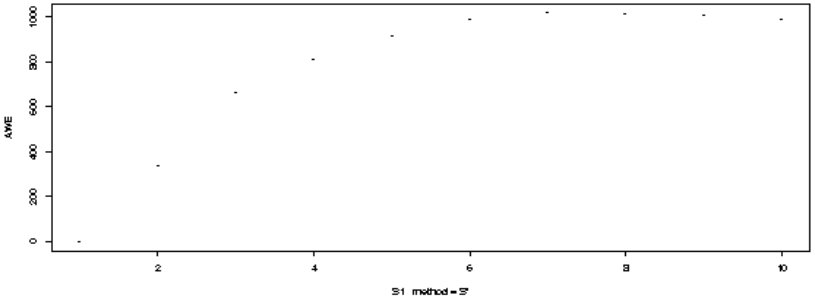
Simulation Study: LHS:Method1, RHS:mclust

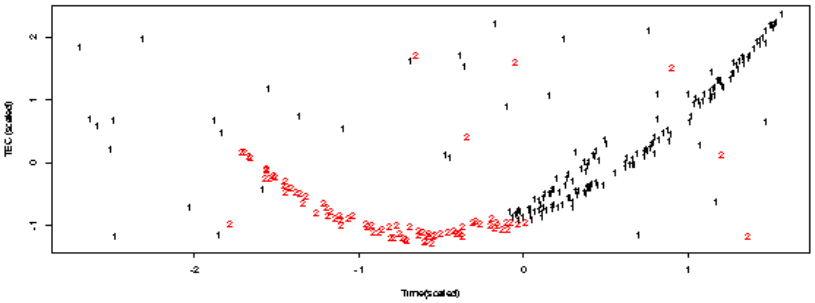
Goals: Estimate performance, **quantify difficulty**, effective number of clusters, identify other methods

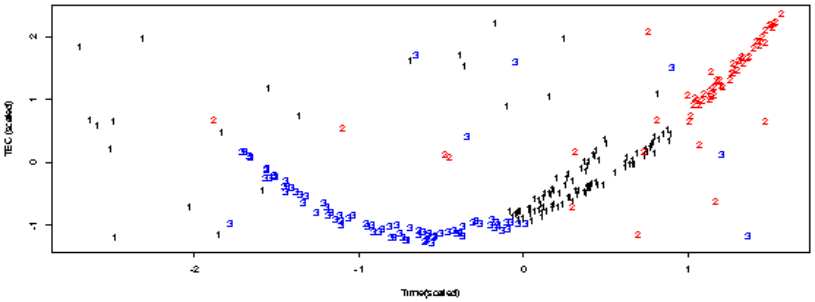












Summary/Future

- SUMMARY

Method1: Combination of noise rejection, hierarchical clustering, and extrapolation with zone of ownership

Metaparameters chosen by hand working with D1, validated in 3^6 search over all 30 data sets.

Method 1 to be implemented in PERL and results compared to manual results in large testing set.

- FUTURE

Storms in TEC vs time plots to be evaluated using ground-based observation data.

More analytical/simulation work on effective no. of clusters and quantifying difficulty of each case