Cluster Analysis of Severe Weather Days of 2004

Jim DeArmon
MITRE/CAASD

The Environmental Working Group (EWG) of the Joint Planning and Development Office (JPDO) is charged with modeling future NAS enhancements. Modeling must consider a number of scenarios under which NAS enhancements will operate. One important scenario is severe en route weather in the CONUS. Because of the complexity of the simulation modeling and limited resources, only a few scenario days will be selected to represent the impact on the NAS of severe weather. A challenge is how to select the days for modeling. On one hand, one could argue that severe weather patterns and movements are quite different each day, making each day unique – this makes selection arbitrary and trivial.

However, there may be sufficient similarity of severe weather on certain days, and that grouping of days is feasible. If groupings are feasible, then selection of sample days could be more informed.

I was asked by the EWG to repeat an analysis I'd published regarding severe weather from 1999 and 2000 – apply cluster analysis to severe weather data, and produce groups of days of 2004. These groupings, based purely on weather data, would then be further analyzed, by Metron Aviation, with respect to NAS "responses", i.e., the characterization of TFM actions, plus flight delays, cancellations, etc. The resultant days would be selected to span the sample space (a year of severe weather in the CONUS) and become the scenario days for the simulation modeling. Results from simulation modeling could be annualized with the knowledge of how the selected days compared to the rest of the year. Other conditions, such as mostly good weather, or CONUS airport weather are considered separately from the analysis here.

This paper describes efforts in applying cluster analysis to 2004 data to find severe weather day groupings.

The data source is National Convective Weather Detection (NCWD), and is supplied by the National Center for Atmospheric Research (NCAR). The data "fuses" convective activity and lightning data and reports lat/long locations of severe weather. The analysis used the days from April 1 to October 31, 2004, since that is typically time during which severe en route weather affects air traffic in the U.S. As with most voluminous data sources, some data are missing, and not all days are represented in their entirety. If the date, however, had at least a single observation for each quarter of the subject day, then that date was deemed useable. (This rule was employed for my previous study and seemed to work well enough.) There is an obvious trade-off here between data quality and sample size. Other filtering rules than those used here are defensible. Using this filtering rule, a total of 197 dates were found usable for this analysis.

To prepare the weather data for the cluster analysis, a grid of cells sized 50 x 50 nmi was overlaid on the conterminous U.S. (CONUS). Since not all locations in the NAS are equally important with respect to air traffic, a weighting scheme was used. The top 50 origin-destination pairs for May 1, 2004 were collected from Airline Service Quality Performance (ASQP) data. Flights between these pairs were used to weight the cells which were on a great circle between the airports. (See Figure 1 for map of routes and weights.) For example, in Figure 1, the cells between Atlanta and New England are weighted higher than those from Los Angeles to Seattle, since there are more flights. These weights are applied to the NCWD weather data: for a given day, if there is weather detected in a cell, then that cell is represented with a "1", and weighted by the described scheme. Cells without weights are ignored, and are not considered in the cluster analysis. If a weighted cell has no severe weather, then a "0" is used to represent that cell. Since weather is not stationary, a sense of time was represented simply by dividing the NAS business day into "quarters" – the 17 hours from 6am to 11pm Eastern time were divided as:

Quarter 1: 6 am – 10 am
Quarter 2: 10 am – 2 pm
Quarter 3: 2 pm – 6 pm
Quarter 4: 6 pm – 11 pm

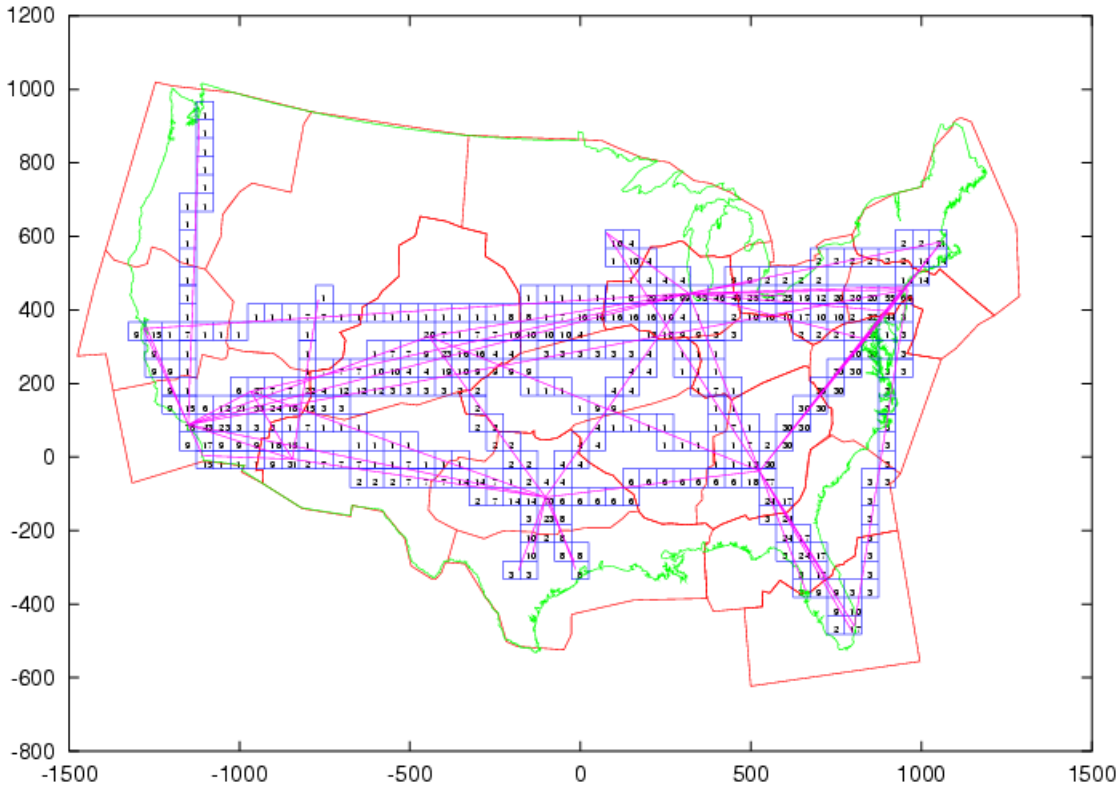


Figure 1: Cell Weights using Top 50 Origin/Destination Pairs of 5/1/2004

Creation of the data for clustering proceeds as follows. For the four "quarters" of the day, for each of the weighted cells, the presence of weather is represented as a 0 or 1. It was decided that the unit to be clustered would be a day. The resultant data structure is a rectangular array in which rows are days and columns are the many binary attributes created by weighting cells four times, one for each "quarter" of the day. From this attribute matrix, a distance matrix was created, giving the similarity of all pairs of days.

The resultant distance matrix was supplied as input to the *hclust* algorithm of Splus [Splus, 2004]. It was decided, somewhat arbitrarily, that the data would be divided into 18 clusters. But note that all possible groupings between 1 (single cluster containing all days) and 197 (197 separate clusters, one for each sample date) are defined per the clustering algorithm. In some analyses, a *pseudo-F* statistic is computed, in an attempt to find a "natural" number of clusters. The analysis here didn't do that, rather, it attempted to find a relatively small number of groups, which would be useful for summarizing the data.

Reasonableness Checking

It is important to check the results of the clustering, since several steps of data reduction and interpretation were involved in the processing. To check the clusters for reasonableness, an alternate cluster analysis of the days was undertaken. The top 50 origin/destination pairs used for the weather weighting were considered. ASQP data were used to compute, for each of the 197 days, for each of the 50 pairs, the percentage of flights which were cancelled, diverted, or delayed 30 minutes or more. This resulted in a rectangular data structure in which rows were days, and there were 50 x 3 = 150 columns of attribute data. This data structure was used as input to a cluster analysis.

At this point, two separate groupings of the 197 days had been created. The first was based solely on severe weather information. The second was based solely on what might be called "NAS response", i.e., how the FAA and airlines reacted to the environmental and other conditions of the day, as reflected in flight delay, cancellation, and diversion. How similar are these solutions? If they're similar, then one might assert that the weather day clusters were non-trivial, and have some meaning in the context of air traffic impact, and may be useful for the intended purpose here – helping to select days for simulation modeling.

The problem of testing the agreement of cluster solutions has been addressed in the open literature. One approach computes a measure called "pair classification percentage" (PCP) [Rand, 1971]. The procedure is as follows.

1. Given two cluster solutions CS1 and CS2 of some collection of items
2. Let Score = 0
3. Consider each pair of items in turn

a. If the pair are in a single cluster in CS1 and in a single cluster in CS2, then increment Score
b. If pair are in different clusters in CS1 and different clusters in CS2, then increment Score

4. PCP = Score divided by number of pairs examined

PCP values were computed for the two comparisons of interest, with the following results. Two clustering algorithms were applied to the flight data.

Weather day clusters versus Ward's method of clustering flight days: 0.783
Weather day clusters versus K-means method of clustering flight days: 0.756

In the paper by Rand, an application of PCP is shown in which the correct cluster solution is known, and various clustering algorithms are pitted in competition to find the known correct answer. In that case, the PCP is directly interpretable: the higher the PCP, then the better the clustering algorithm's accuracy.

For our application, however, there is no known correct answer, leading to the question of interpretability of the computed PCP values. A Monte-Carlo experiment of 10,000 trials was performed to construct the "null distribution", i.e., the distribution of PCP values under the assumption that items are assigned to clusters at random. This was done for both the Ward's method and the K-means method of clustering flight days. By this means, the computed PCP values of 0.783 and 0.756 shown above can be used to find p-values (aka "observed significance"). These are as follows:

Weather day clusters versus Ward's method of clustering flight days: 0.0002
Weather day clusters versus K-means method of clustering flight day : 0.006

One might interpret these values as two chances in ten thousand,and six chances in a thousand that one would see this much agreement between cluster solutions due purely to chance effects. That is, the two cluster solutions agree pretty well. There is hence some confidence that the clustering of severe weather days was not misguided, and the results have some meaning.

Appendix A presents the clustering results. Both the date, and the distance from the cluster centroid are presented.

Appendix B presents the graphical representation of the cluster centroid or center-most date, as well as a terse prose description of the displayed day.

References

Rand, W. M. (1971), “Objective Criteria for the evaluation of Clustering Methods,” Journal of the American Statistical Association, Vol. 66, pp 846-850.

Splus, 2004, Description of SPLUS Software, http://www.insightful.com/ .

Appendix A: Days Grouped into Clusters

Clusters and members are presented here. Cluster numbers are arbitrary. Dates are prefaced with a distance from the centermost date of the cluster. The units are for the abstract, high-dimensional space.

```
Cluster 1
      0 2004-06-17
   1974 2004-05-18
   3056 2004-08-20
   3265 2004-08-19
   3438 2004-07-30
   4001 2004-07-31

Cluster 2
      0 2004-09-10
   1525 2004-09-11
   1786 2004-09-09
   2243 2004-08-13
   2307 2004-07-15
   2404 2004-08-15
   2419 2004-08-14
   2604 2004-07-25
   2792 2004-08-16
   2798 2004-09-19
   2849 2004-09-18
   2929 2004-08-06
   3233 2004-09-27

Cluster 3
      0 2004-04-15
   1058 2004-05-04
   1107 2004-04-14
   1187 2004-04-04
   1210 2004-04-28
   1231 2004-04-27
   1269 2004-04-16
   1290 2004-05-03
   1293 2004-10-14
   1297 2004-04-05
   1318 2004-04-11
   1340 2004-10-08
   1341 2004-10-17
   1342 2004-09-23
   1348 2004-04-18
   1368 2004-04-26
   1374 2004-10-21
   1376 2004-10-09
   1388 2004-04-03
   1390 2004-04-29
   1408 2004-04-19
   1415 2004-09-20
   3131 2004-05-27
```

```
Cluster 3 continued
   1425 2004-10-24
   1437 2004-04-22
   1446 2004-04-09
   1462 2004-10-16
   1492 2004-10-15
   1496 2004-10-20
   1508 2004-04-17
   1553 2004-10-25
   1557 2004-05-06
   1591 2004-04-10
   1596 2004-09-22
   1617 2004-04-02
   1621 2004-09-21
   1687 2004-04-01
   1702 2004-10-30
   1709 2004-05-05
   1717 2004-10-11
   1765 2004-10-26
   1767 2004-10-12
   1776 2004-04-06
   1796 2004-04-24
   1838 2004-10-23
   1875 2004-04-12
   1876 2004-10-10
   1940 2004-04-21
   1945 2004-10-31
   1965 2004-09-29
   1979 2004-10-18
   1980 2004-05-29
   2038 2004-10-28
   2069 2004-10-13
   2075 2004-04-07
   2081 2004-04-08
   2125 2004-10-04
   2133 2004-10-06
   2173 2004-05-14
   2188 2004-05-28
   2218 2004-10-27
   2224 2004-04-25
   2414 2004-04-13
   2437 2004-09-30
   2517 2004-05-07
   2628 2004-05-15
   2659 2004-04-23
   2749 2004-05-25
   2783 2004-05-24
   2846 2004-05-26
   2866 2004-10-01
   2996 2004-10-29
```

Cluster 4

0	2004-09-15
1946	2004-04-20
2420	2004-10-22
2465	2004-05-31
2885	2004-09-14
2885	2004-10-07
3125	2004-05-30

Cluster 5

0	2004-06-07
1964	2004-09-01
2090	2004-08-31
2107	2004-06-08
2267	2004-06-19
2350	2004-06-27
2382	2004-06-30
2497	2004-10-03
2532	2004-06-26
2632	2004-06-29
2670	2004-07-08
2789	2004-06-24
2874	2004-07-29
3000	2004-08-05
3422	2004-06-28

Cluster 6

0	2004-05-17
2270	2004-06-14
2645	2004-07-04
2861	2004-07-01
3360	2004-06-16
3487	2004-07-02

Cluster 7

0	2004-07-27
1281	2004-07-12
1488	2004-06-25
1631	2004-08-30
1872	2004-07-18
2022	2004-08-01
2349	2004-08-12
2455	2004-06-22
2480	2004-08-21
2526	2004-07-17
2780	2004-08-11

Cluster 8

0	2004-08-24
1739	2004-08-23
1935	2004-08-25
2105	2004-07-06
2457	2004-07-09
2590	2004-05-22
3061	2004-06-11
3227	2004-05-12
3258	2004-06-23
3259	2004-05-20

Cluster 9

0	2004-09-13
1579	2004-09-12
2005	2004-06-06
2038	2004-06-20
2047	2004-09-02
2065	2004-06-05
2136	2004-09-26
2141	2004-09-05
2189	2004-05-01
2189	2004-06-04
2268	2004-09-04
2282	2004-09-25
2317	2004-09-06
2318	2004-08-26
2401	2004-09-24
2455	2004-08-07
2475	2004-10-05
2532	2004-09-03
2567	2004-08-08
2651	2004-08-09
2850	2004-04-30

Cluster 10

0	2004-08-27
2718	2004-05-08
3342	2004-06-09
3433	2004-05-11

Cluster 11

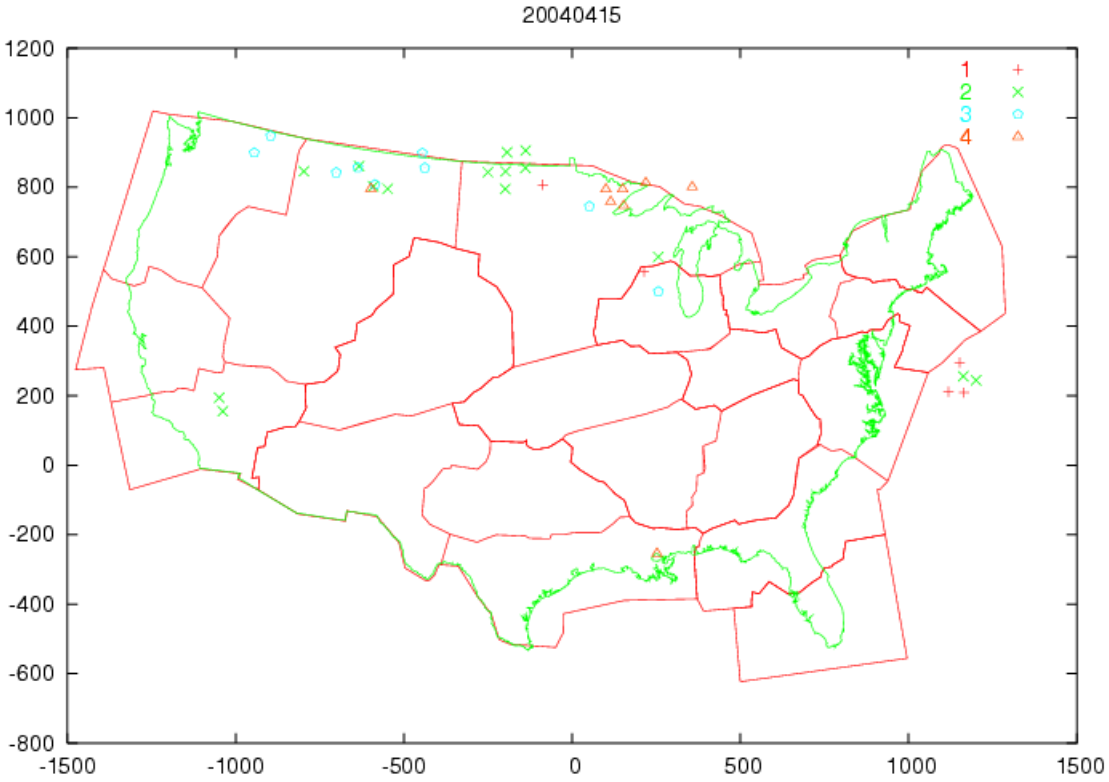
0	2004-05-09
2130	2004-05-10
2664	2004-05-23
2781	2004-07-11
3522	2004-07-22

```
Cluster 12
      0 2004-09-07
     27 2004-09-08
   2440 2004-05-02
   2791 2004-10-02
   2802 2004-07-19
   2838 2004-09-17

Cluster 13
      0 2004-10-19
   2120 2004-06-13
   2199 2004-09-16
   2289 2004-05-16
   2381 2004-05-19
   2808 2004-08-29

Cluster 14
      0 2004-07-21
   2829 2004-07-20
   3357 2004-07-16

Cluster 15
      0 2004-08-10
   2685 2004-08-04
   2974 2004-07-28

Cluster 16
      0 2004-06-10
   2111 2004-07-05
   2514 2004-08-03

Cluster 17
      0 2004-07-26
   2093 2004-08-17
   2574 2004-08-02

Cluster 18
      0 2004-09-28
   2186 2004-07-14
   2322 2004-07-23
```

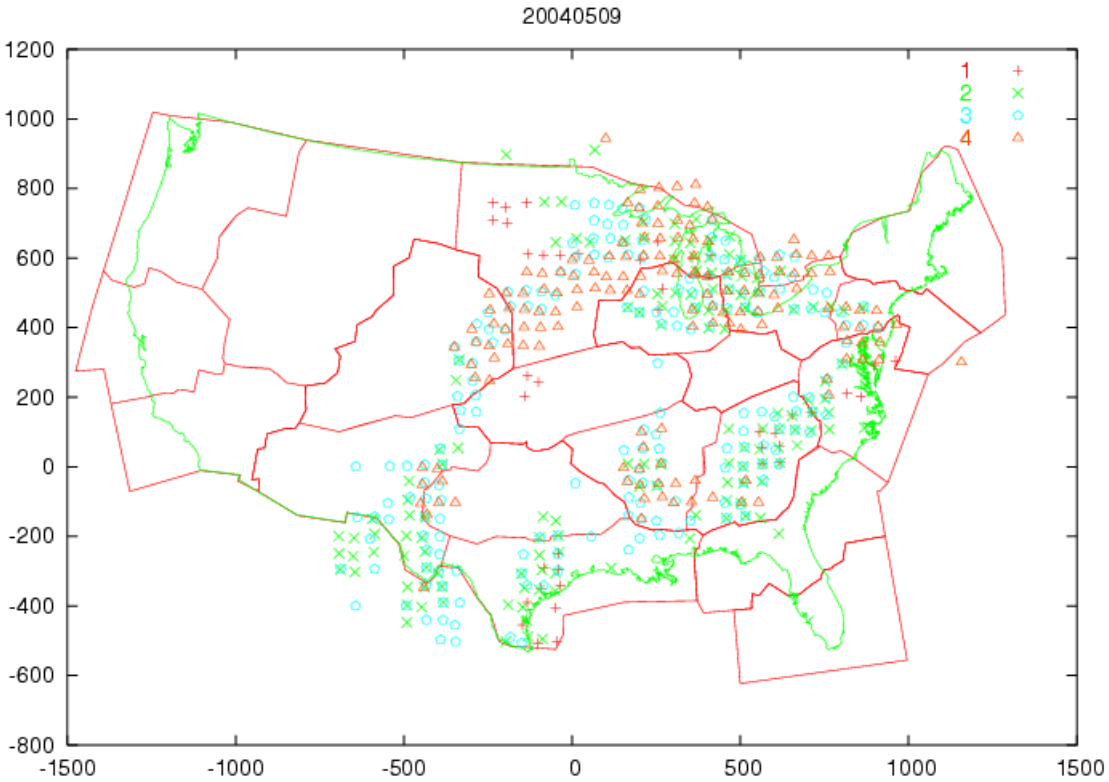
Appendix B: Graphics and Descriptions of Cluster Centroid Days

Presented below are graphical depictions of the centermost date of each cluster, and a short prose description. Note descriptions use several forms of abbreviations: airport 3-character designators, Air Route Traffic Control Center (ARTCC) 3-character designators, state 2-letter designators, and regions of the U.S.

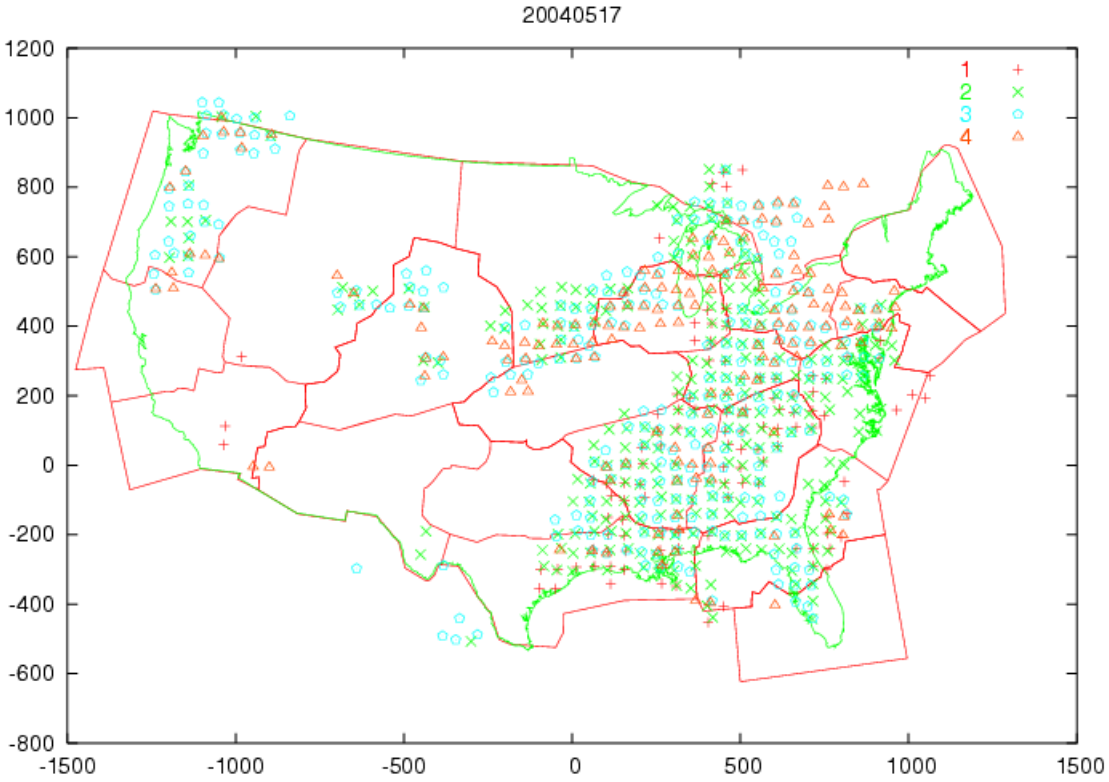
Dates are presented in chronological order, and not in cluster-number order. The legend in the upper right of each display refers to the "quarters" of the CONUS business day.



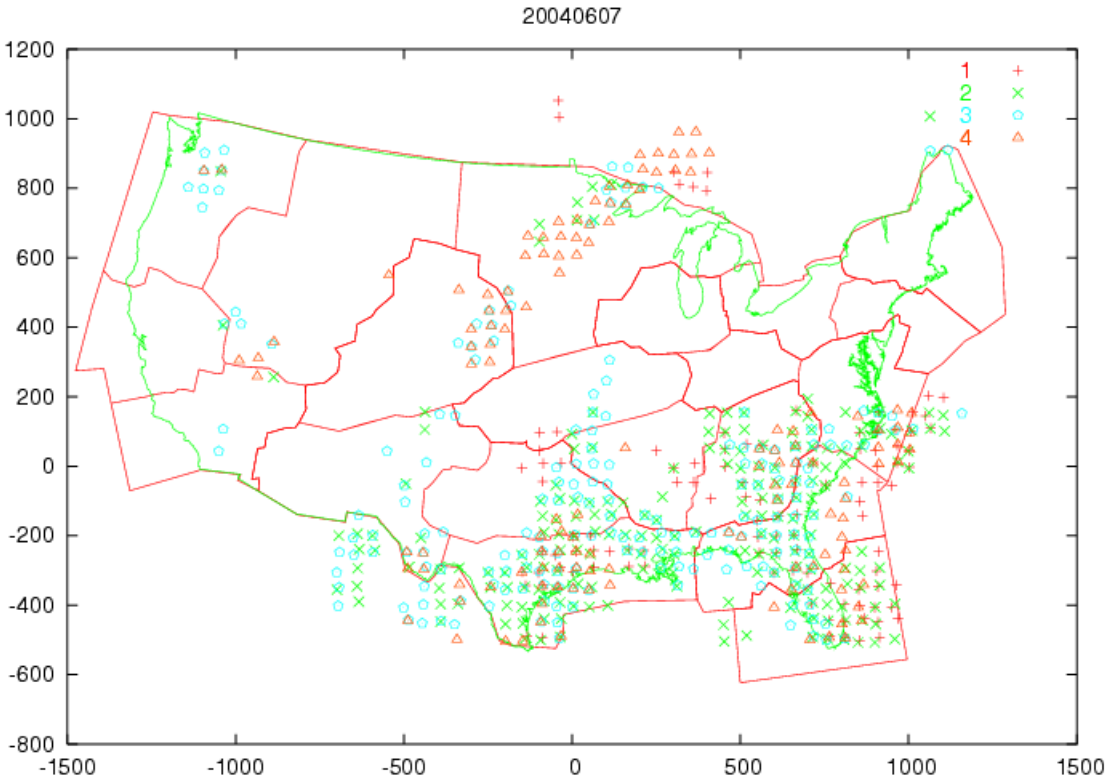
Cluster 3: Generally good weather throughout the CONUS



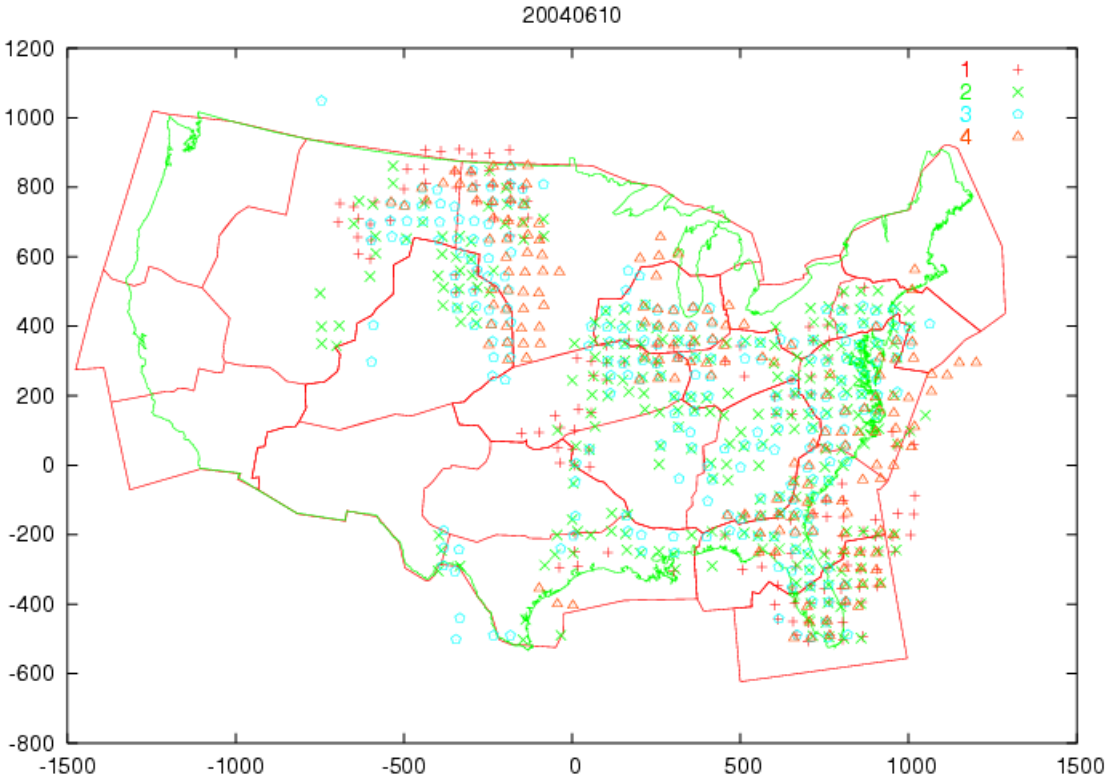
Cluster 11: Weather in northern Great Lakes, and near Atlanta late in the day



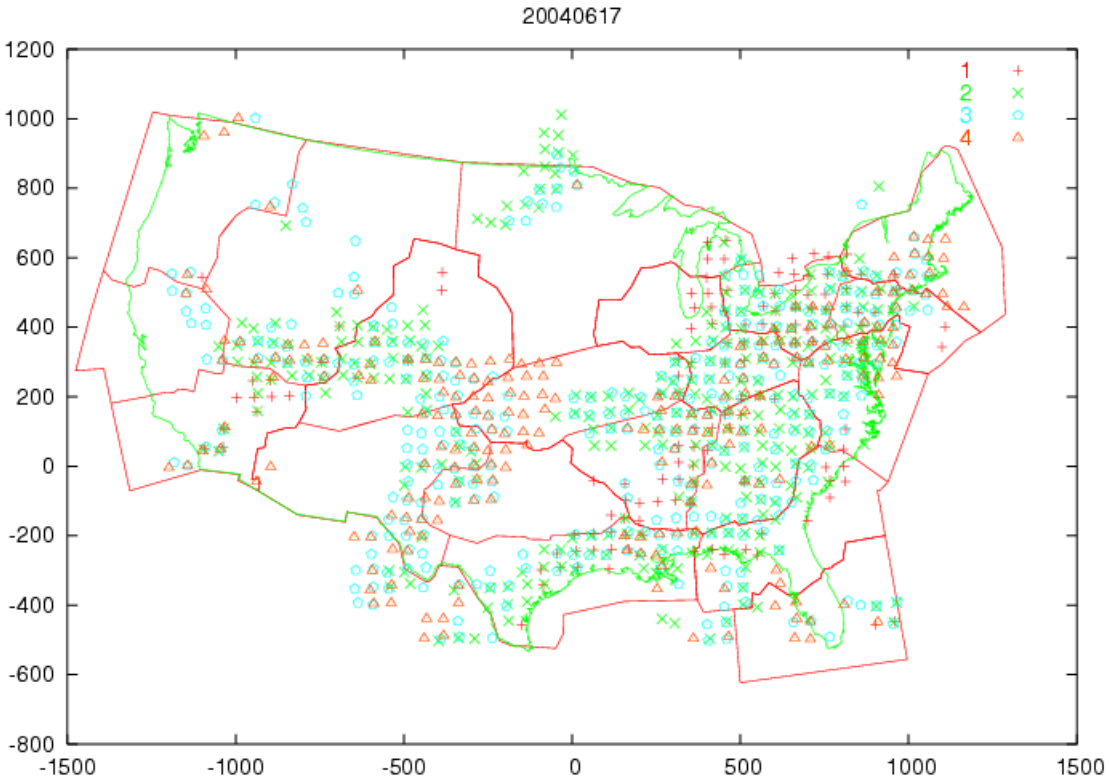
Cluster 6: Weather in a wide swath from TX to DC most of the day



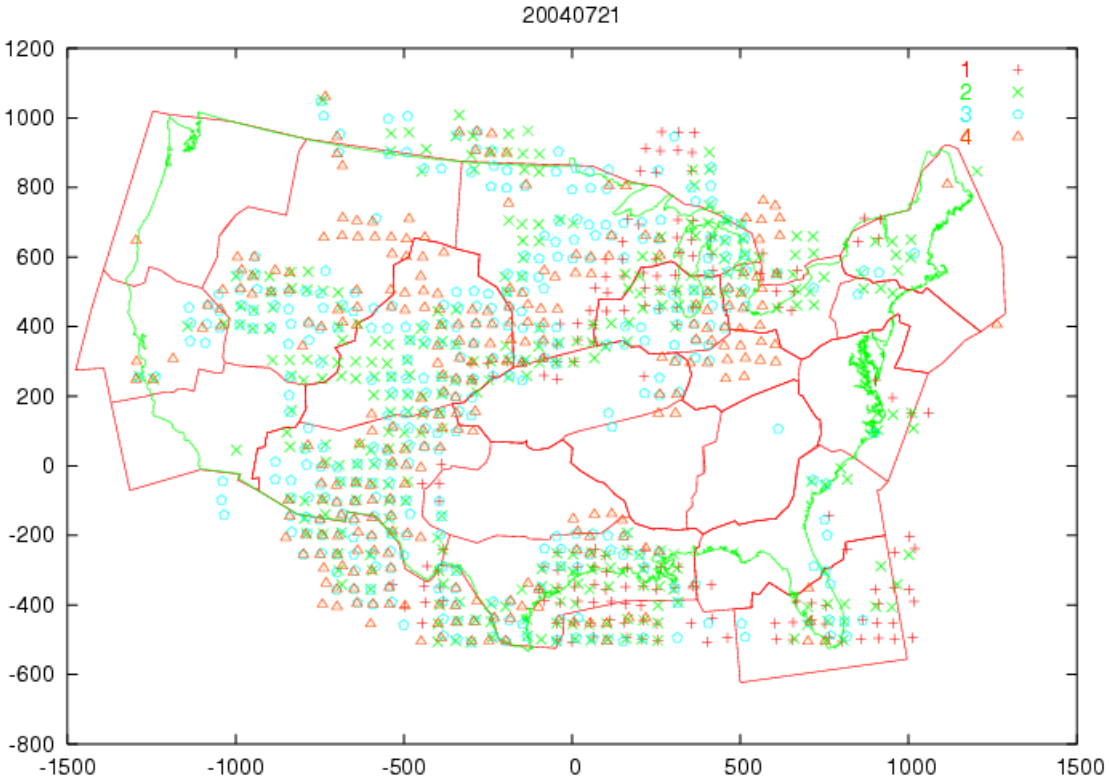
Cluster 5: Weather from TX to FL and GA most of the day



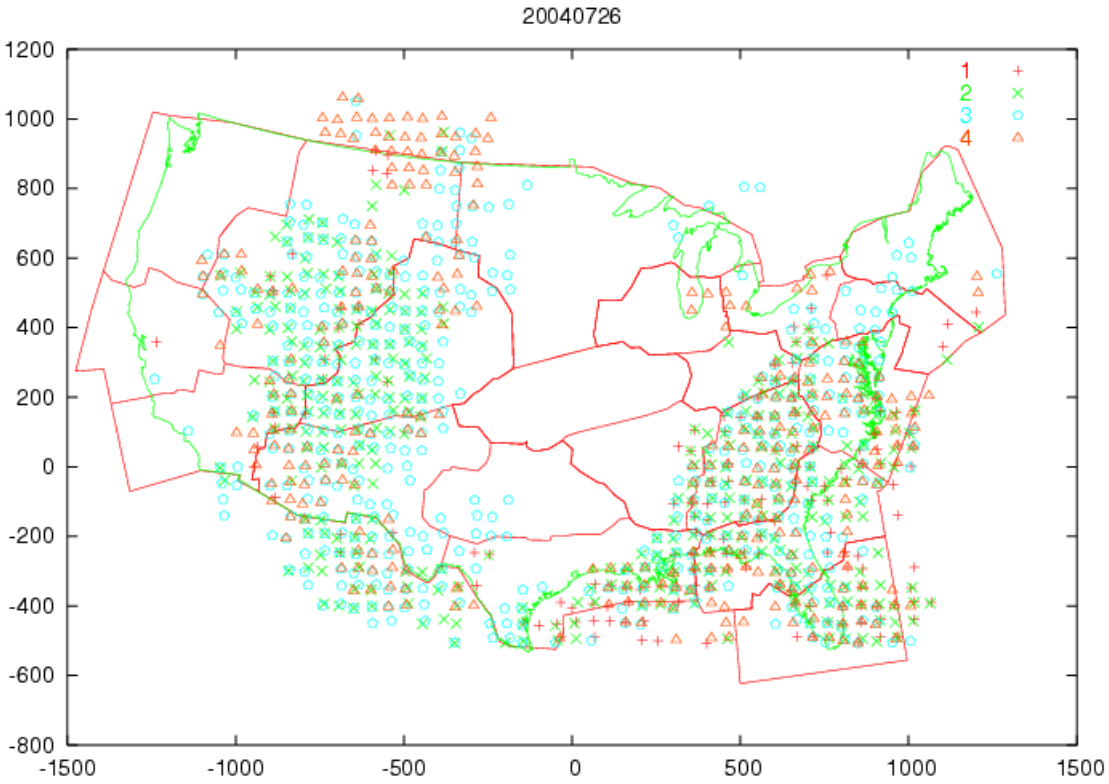
Cluster 16: Weather from FL to NY and in ZAU, after 10 AM EDT



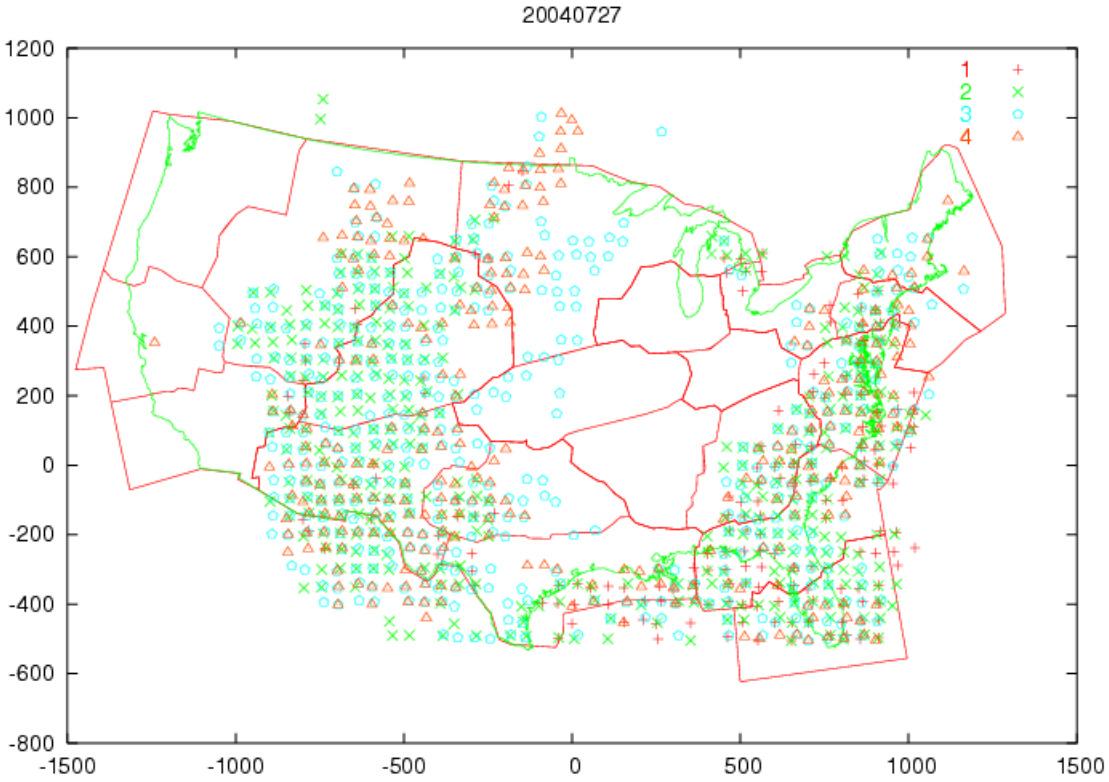
Cluster 1: Weather from NV to VA, and in NM, and ZHU, and from GA to MA



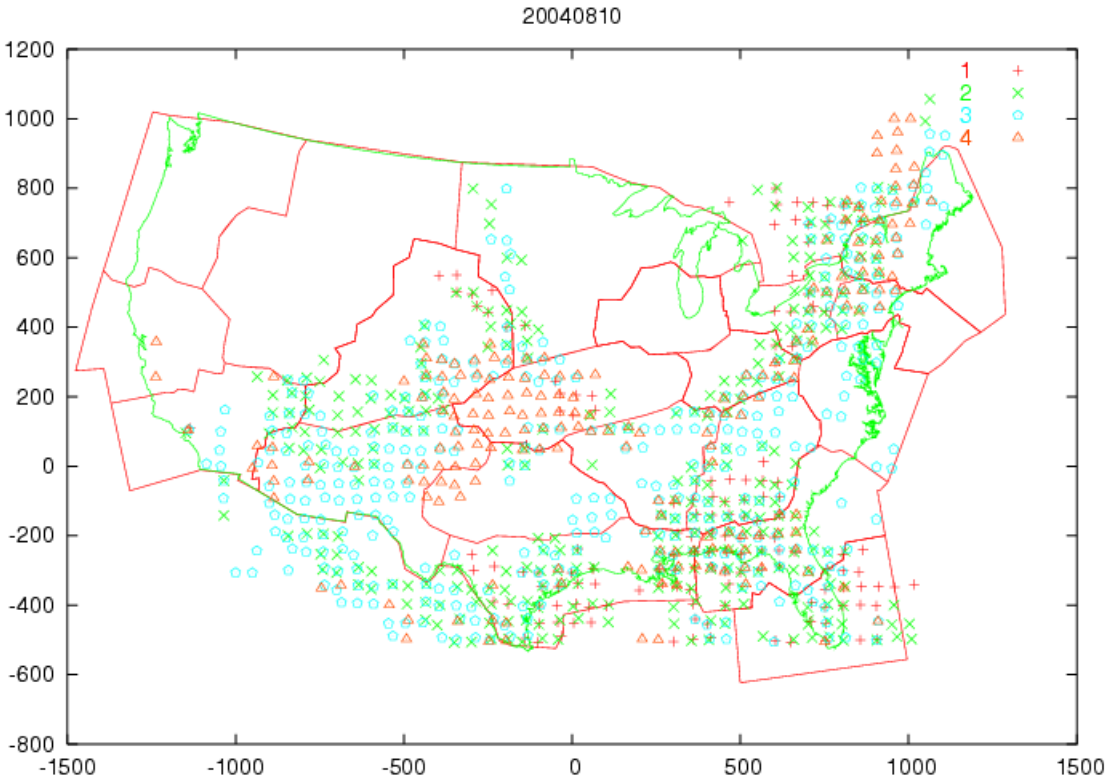
Cluster 14: Weather from PHX to MSP, and in ZHU



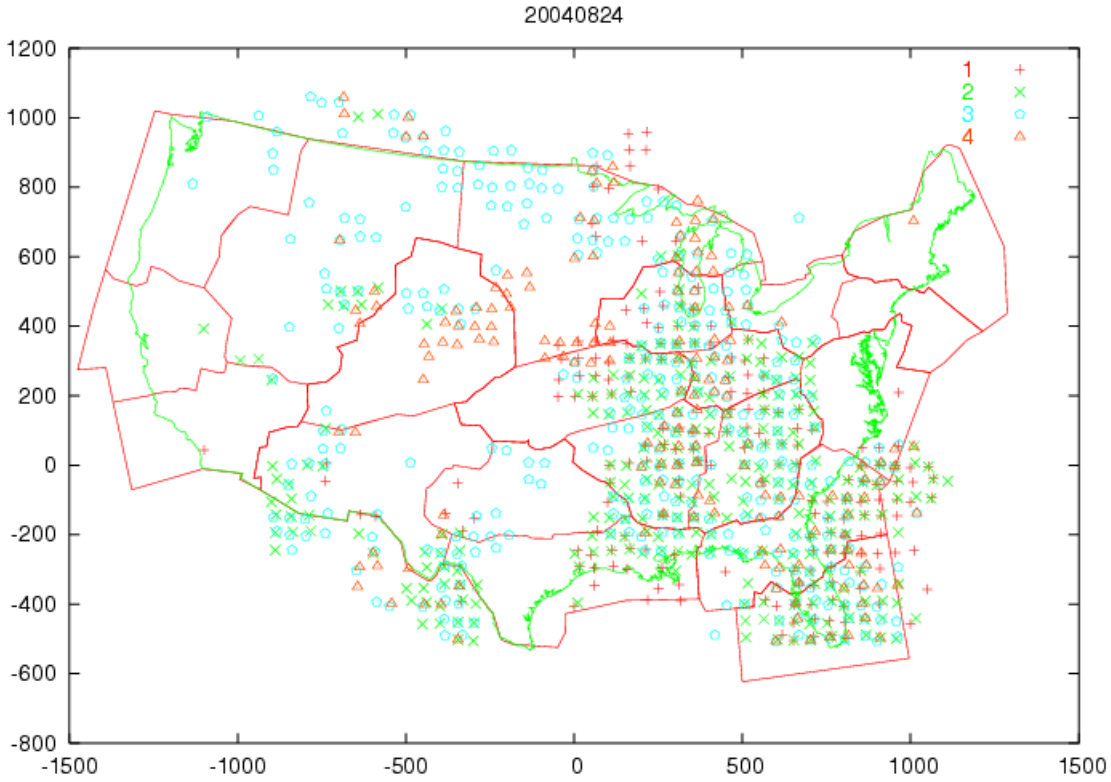
Cluster 17: Weather in ZAB, ZHU, and ZMA to ZDC



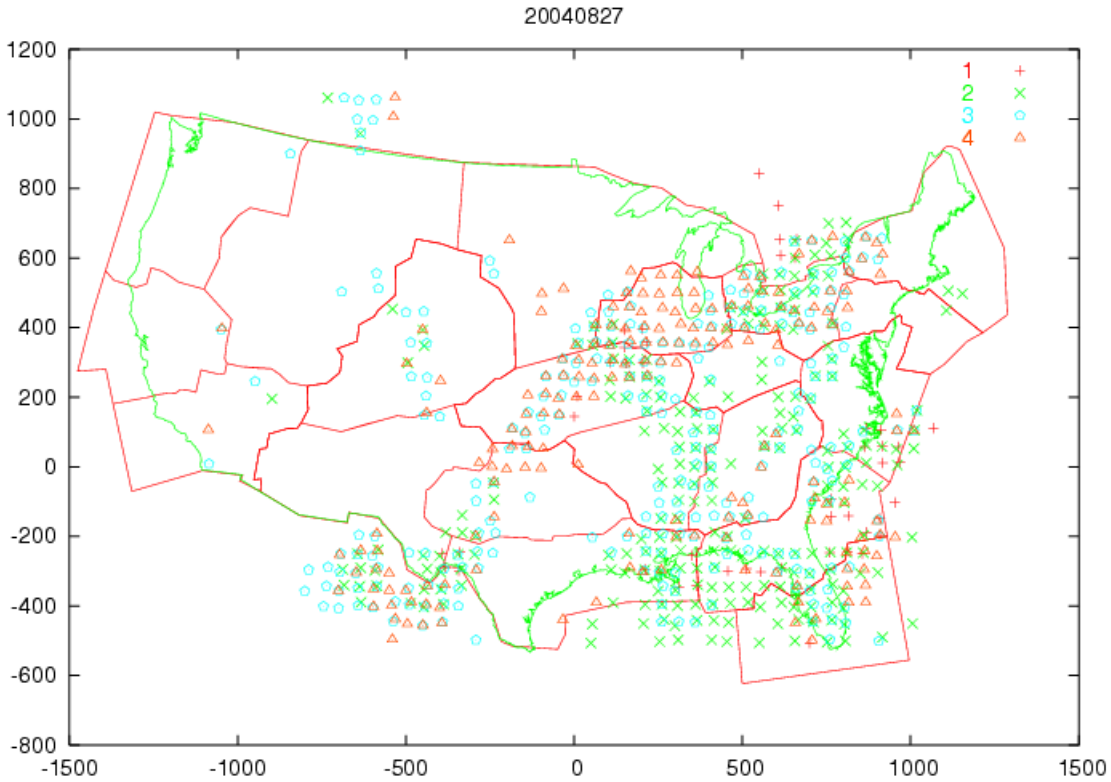
Cluster 7: Weather in ZAB, ZHU, and ZMA to ZNY



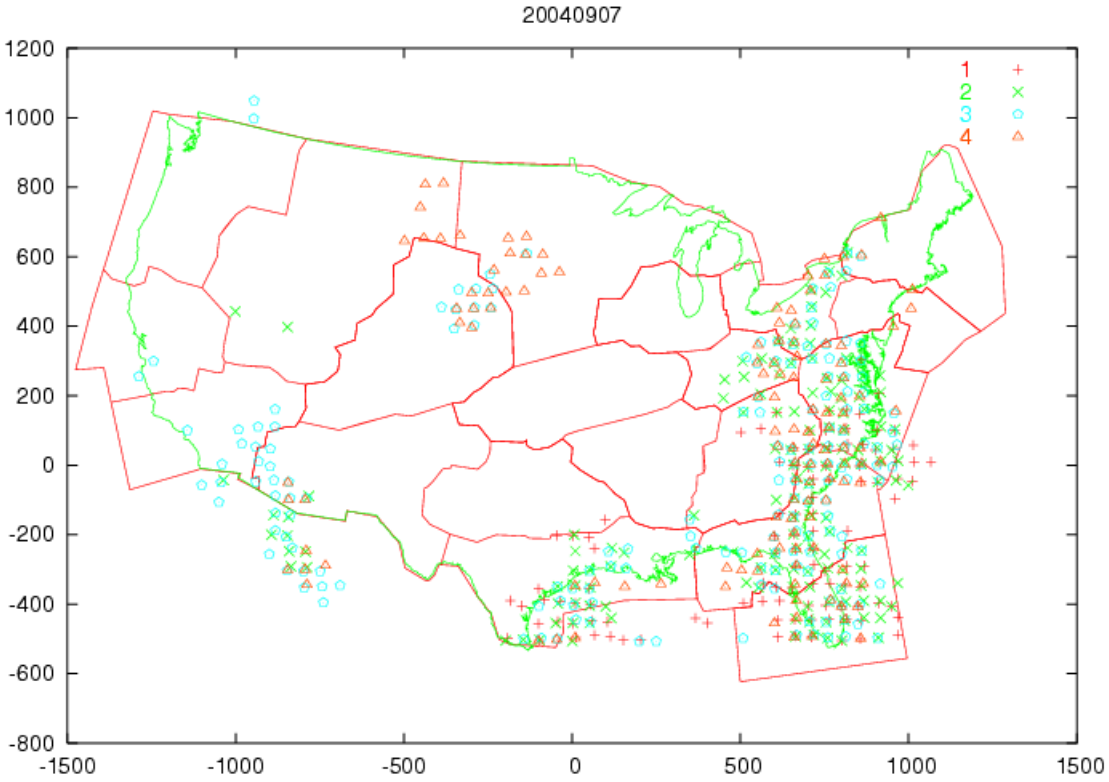
Cluster 15: Weather in ZAB moving to ZKC, also in ZHU, ZJX, ZTL, and New England



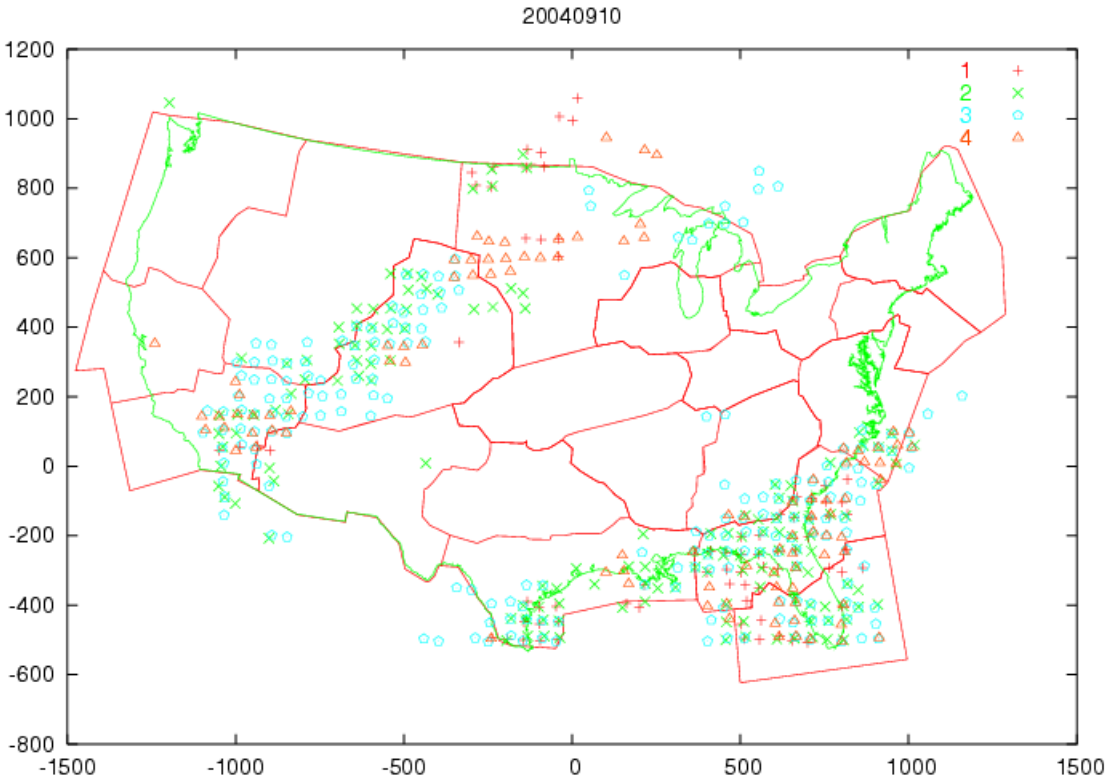
Cluster 8: Weather in a wide swath from E. ZHU north to ZAU, plus ZJX and ZMA



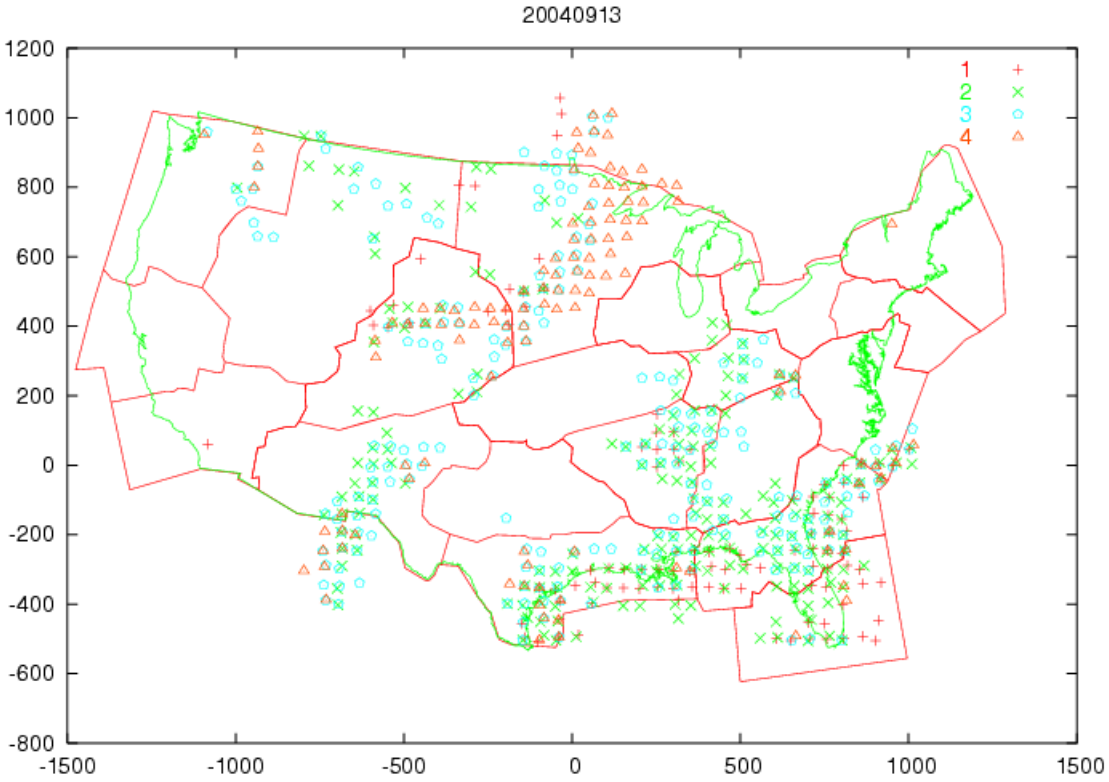
Cluster 10: Weather in ZHU and FL midday, and ZKC, ZAU, and ZOB until late



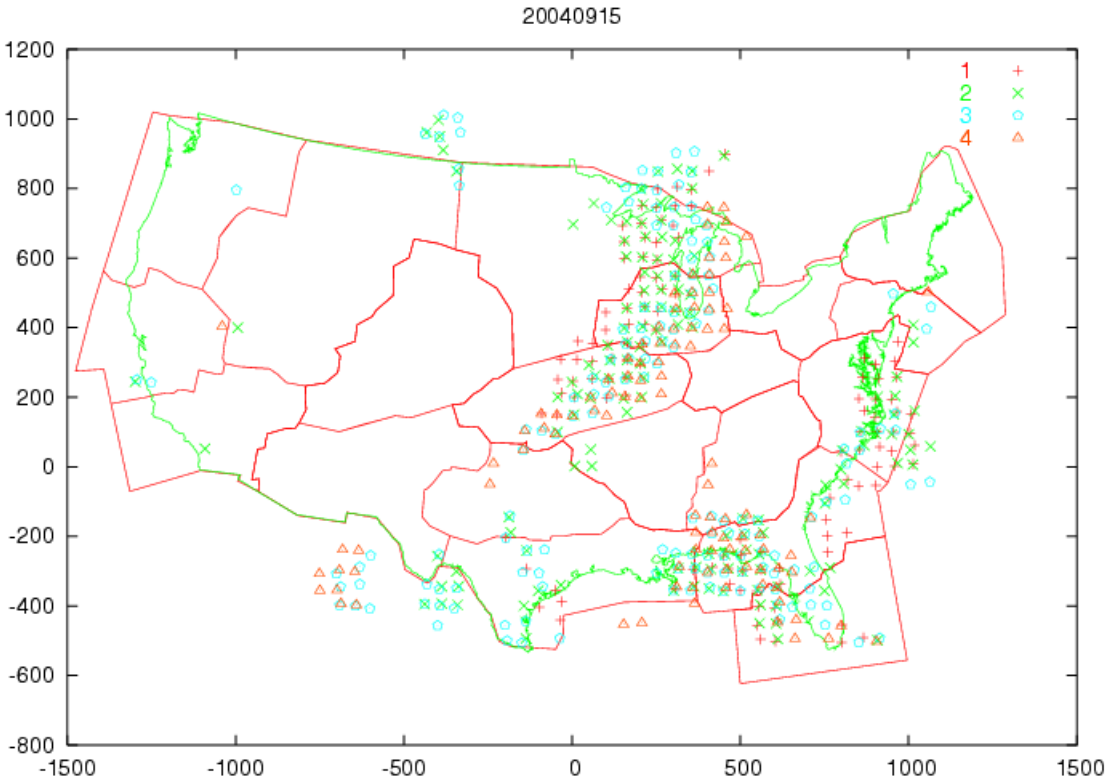
Cluster 12: Weather from S. FL to MD, and in ZHU



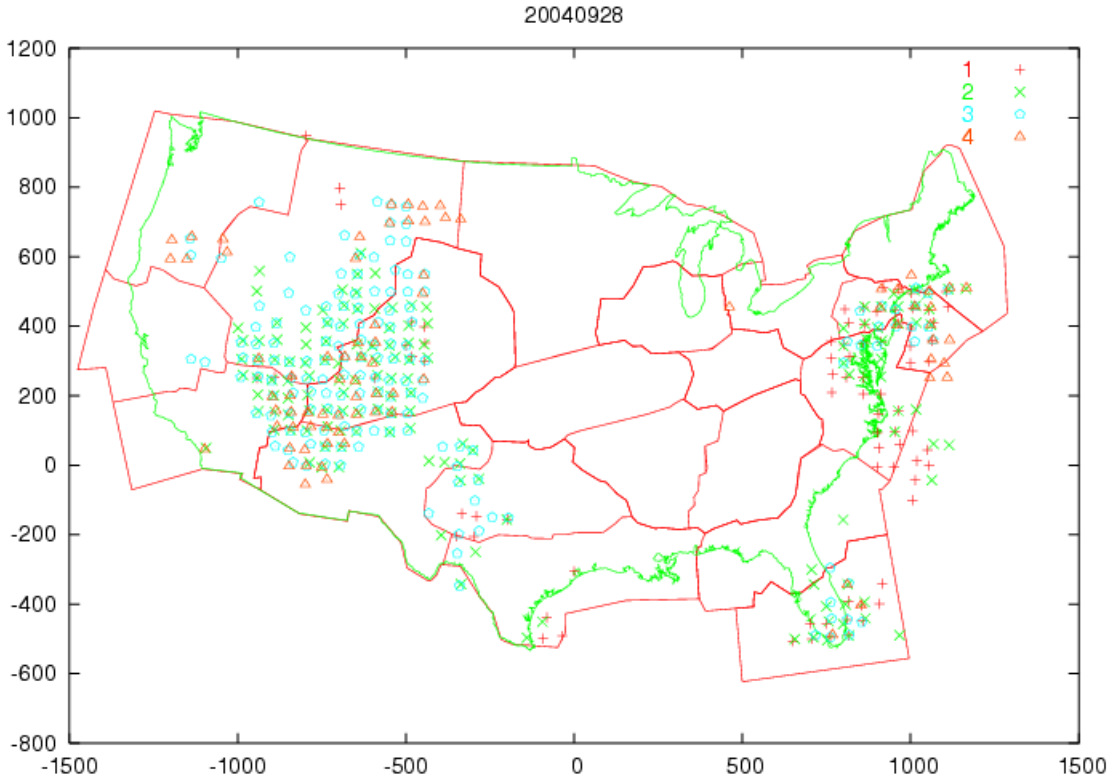
Cluster 2: Weather in ZLA and ZDV midday, also in ZHU and FL



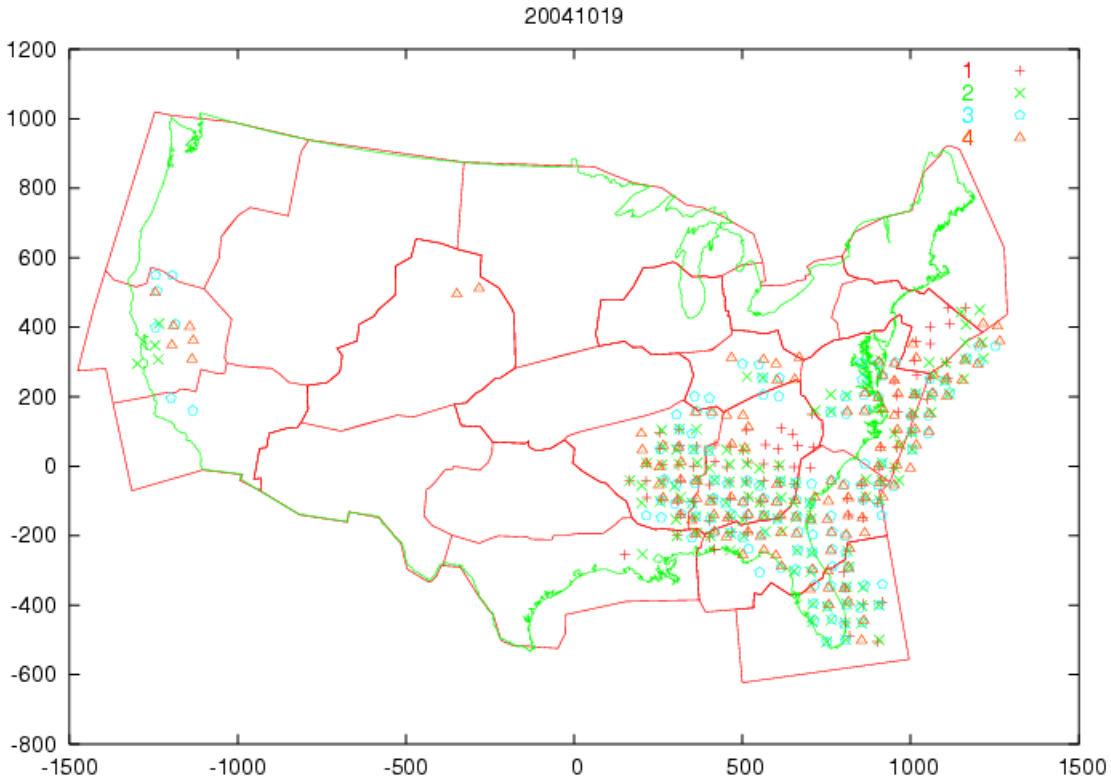
Cluster 9: Weather from ZHU to FL and ZME midday



Cluster 4: Weather in ZKC and ZAU, and in ZMA and ZJX



Cluster 18: Weather at LAS and AZ and ZDV, some weather in ZDC and ZNY



Cluster 13: Weather in ZME, ZTL and FL to off-shore MD