Statistical Cluster Analysis for Verification of Spatial Fields

Caren Marzban, Univ. of Washington and Oklahoma
http://www.nhn.ou.edu/~marzban

Comparing gridded field with gridded field.

Each composed of features/entities/objects/events/...

Error = Shape + Size + Displacement + Magnitude

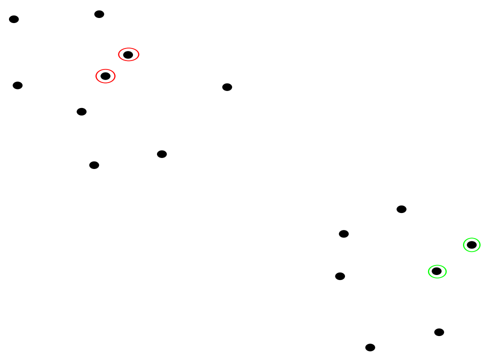
Ebert, McBride: Threshold to define CRA

Why not let cluster analysis infer the objects/events/...?

Seems obvious, but not done.

Baldwin, Lakshmivarahan, Kain:
cluster analysis → convective/nonconvective

Cluster Analysis



Agglomerative Hierarchical Techniques.

Given N data points (cases):

- Start with N clusters.
- Identify the closest clusters.
- Combine them.
- Repeat.
- End with 1 cluster.

Iterative - explores different scales

NC not fixed

D between points = Euclidean

D between clusters = average of pairwise distances between points.

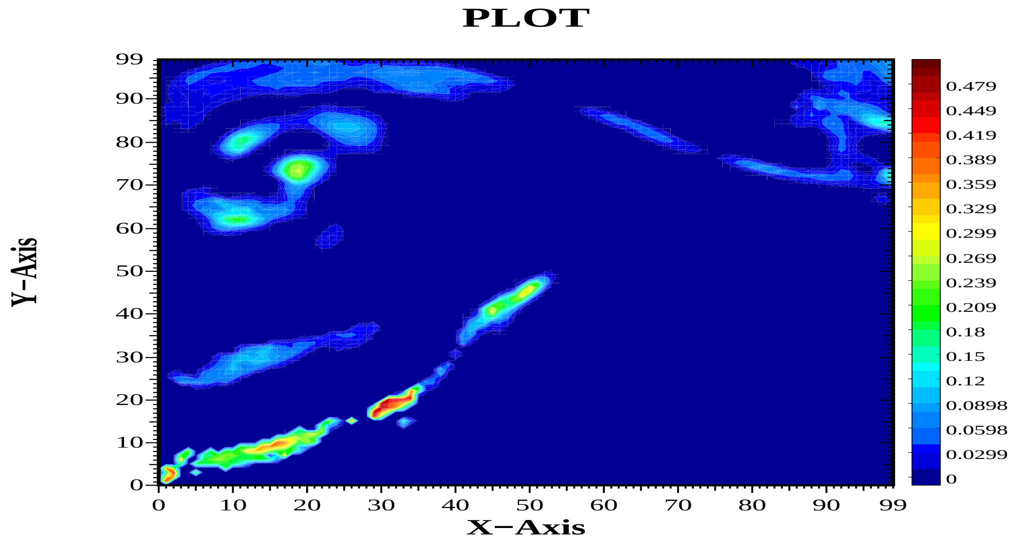
Distance in x-y or x-y-p or weighted space.

Matching clusters between fields

Distance between *fields* - overall forecast error

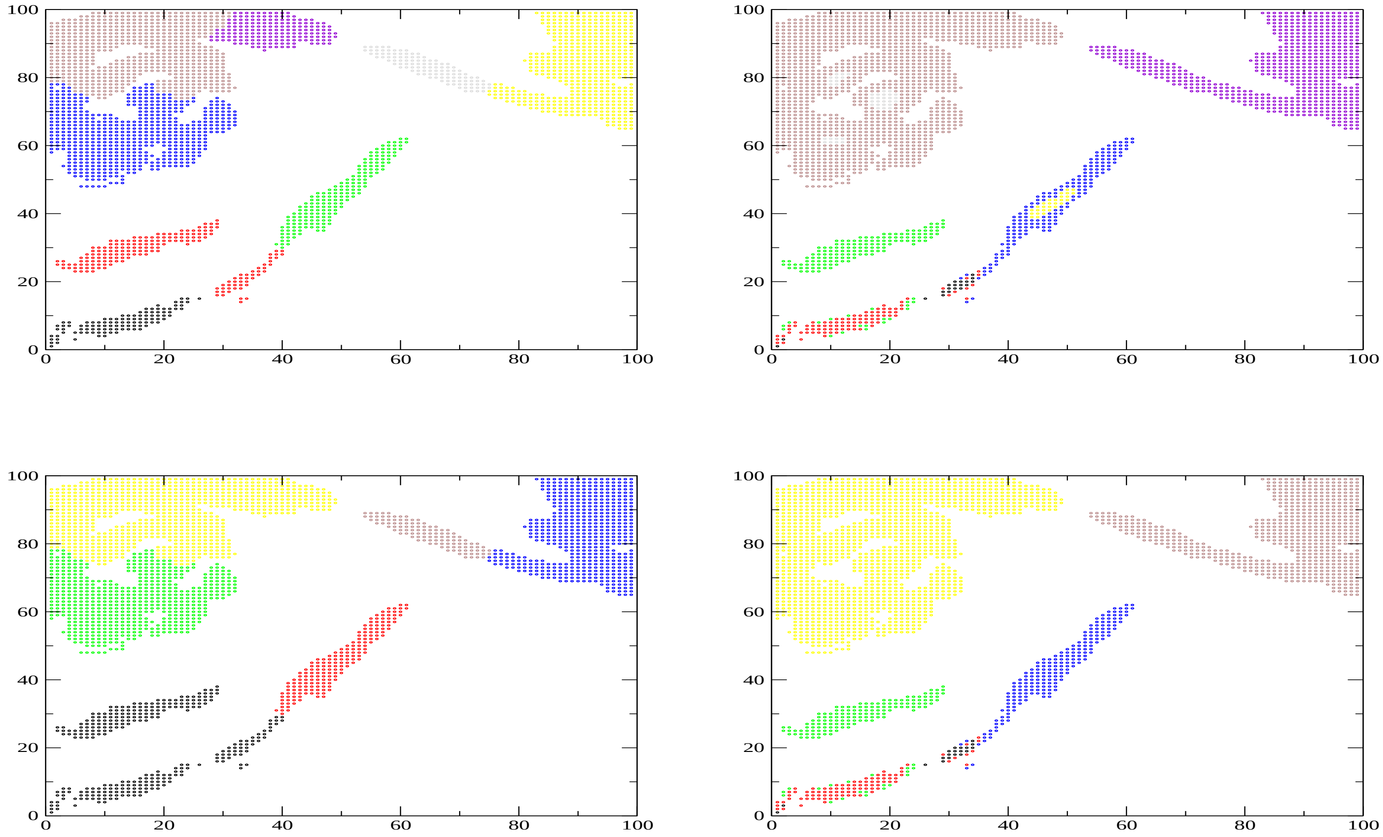
Wed Feb 25 12:27:15 2004

How well does cluster analysis agree with the human eye?

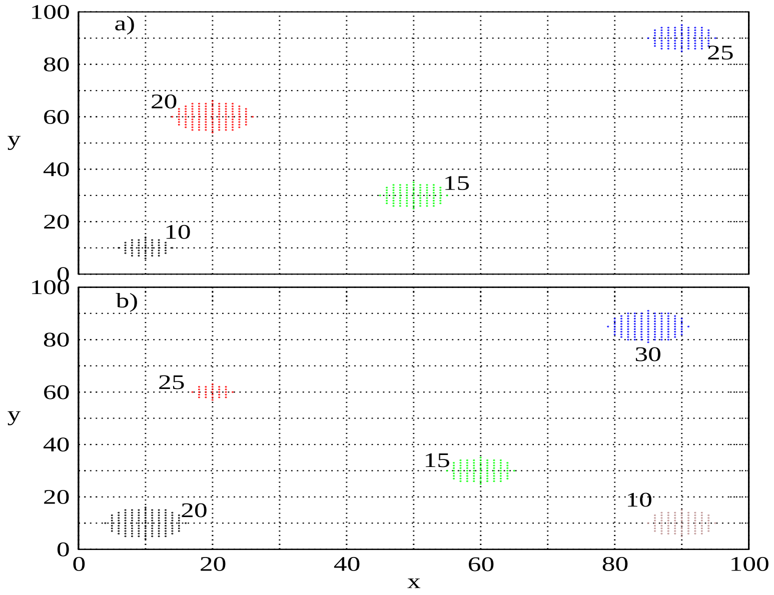


x-y

x-y-p



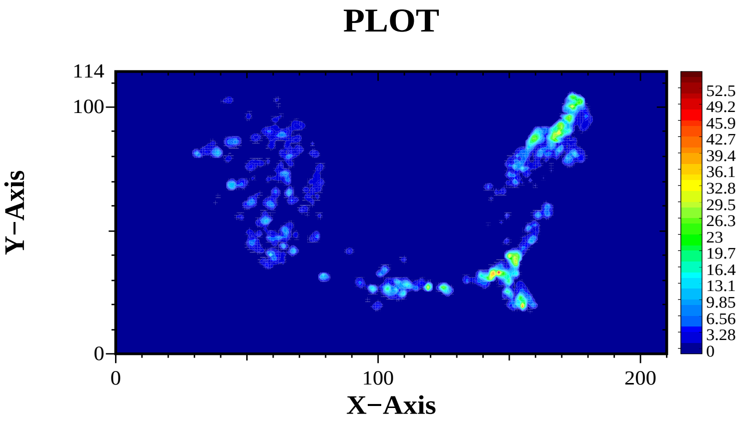
How do we match the clusters between two (fake) fields?

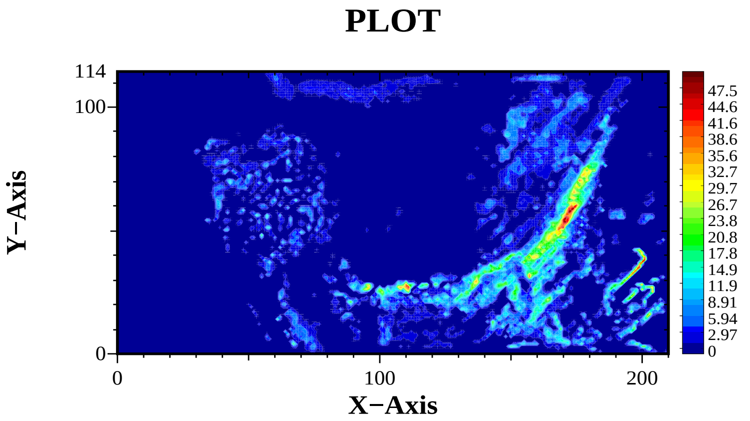


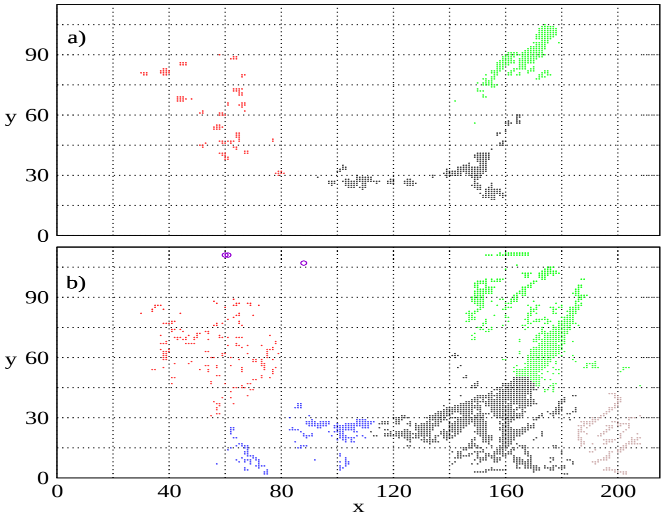
	NC in Forecast Field					
NC in Observed Field	2	3	4	5	6	7
2	0.143	0.160	0.179	0.179	0.193	0.213
3	0.155	0.378	0.163	0.158	0.167	0.172
4	0.171	0.221	0.133	**0.096**	0.103	0.106
5	0.171	0.221	0.137	0.383	0.389	0.365
6	0.171	0.221	0.142	0.387	0.701	0.619

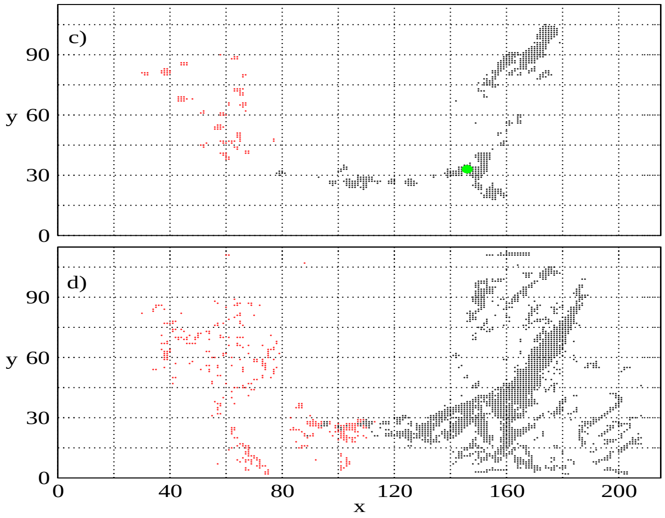
Cluster	x-y Distance	x-y-p Distance
Black	0.068	0.582
Red	0.139	0.148
Green	0.094	0.139
Blue	0.083	0.133
Average	0.096	0.250

And for real fields?





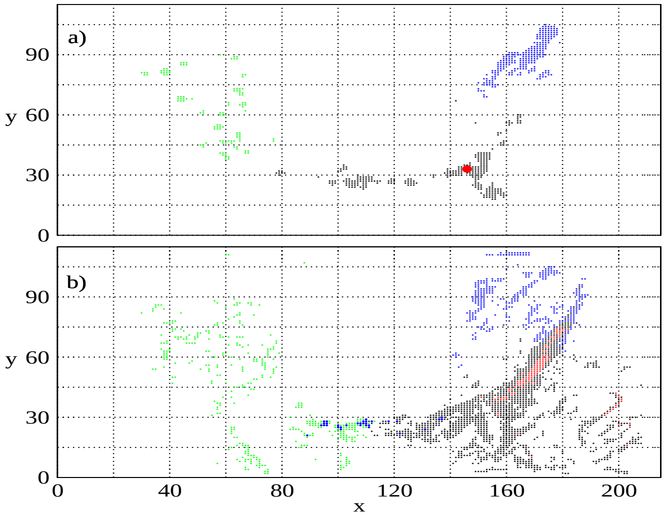


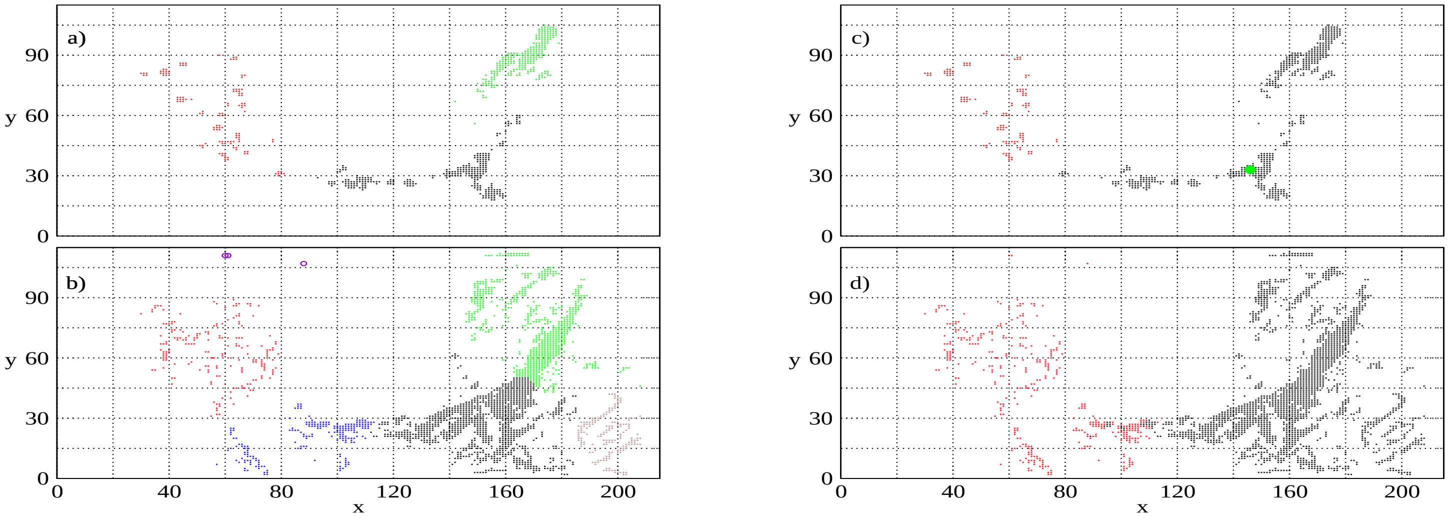


3-6

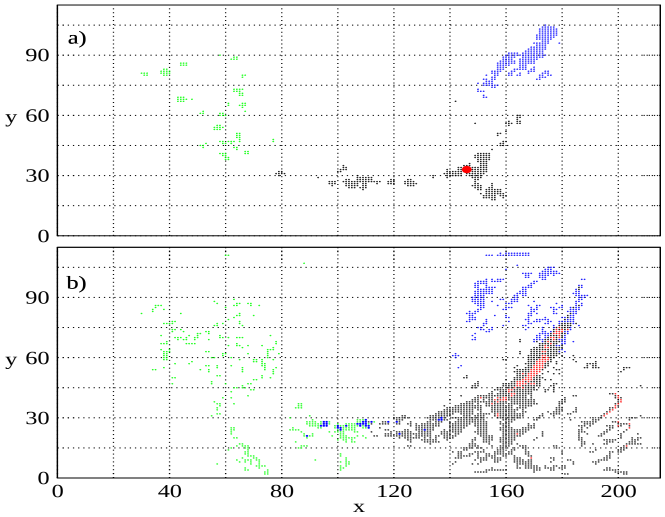
4-6

3-2





3-6 4-6 3-2



Observation	x-y (3,6)	x-y-p (3,2)	x-y-p (4,6)
Black	0.063	0.066	0.087
Red	0.181	-	5.454
Green	0.065	0.232	0.232
Blue	-	-	0.108
Average	0.103	0.149	1.470

Note: Precip in each cluster has a distribution.
So, compare clusters in terms of their means *and* variances

$$T = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1^2} + \frac{\sigma_2^2}{n_2^2}}}$$

But completely dependent, i.e. $n_i = 1$.

$$T = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}$$

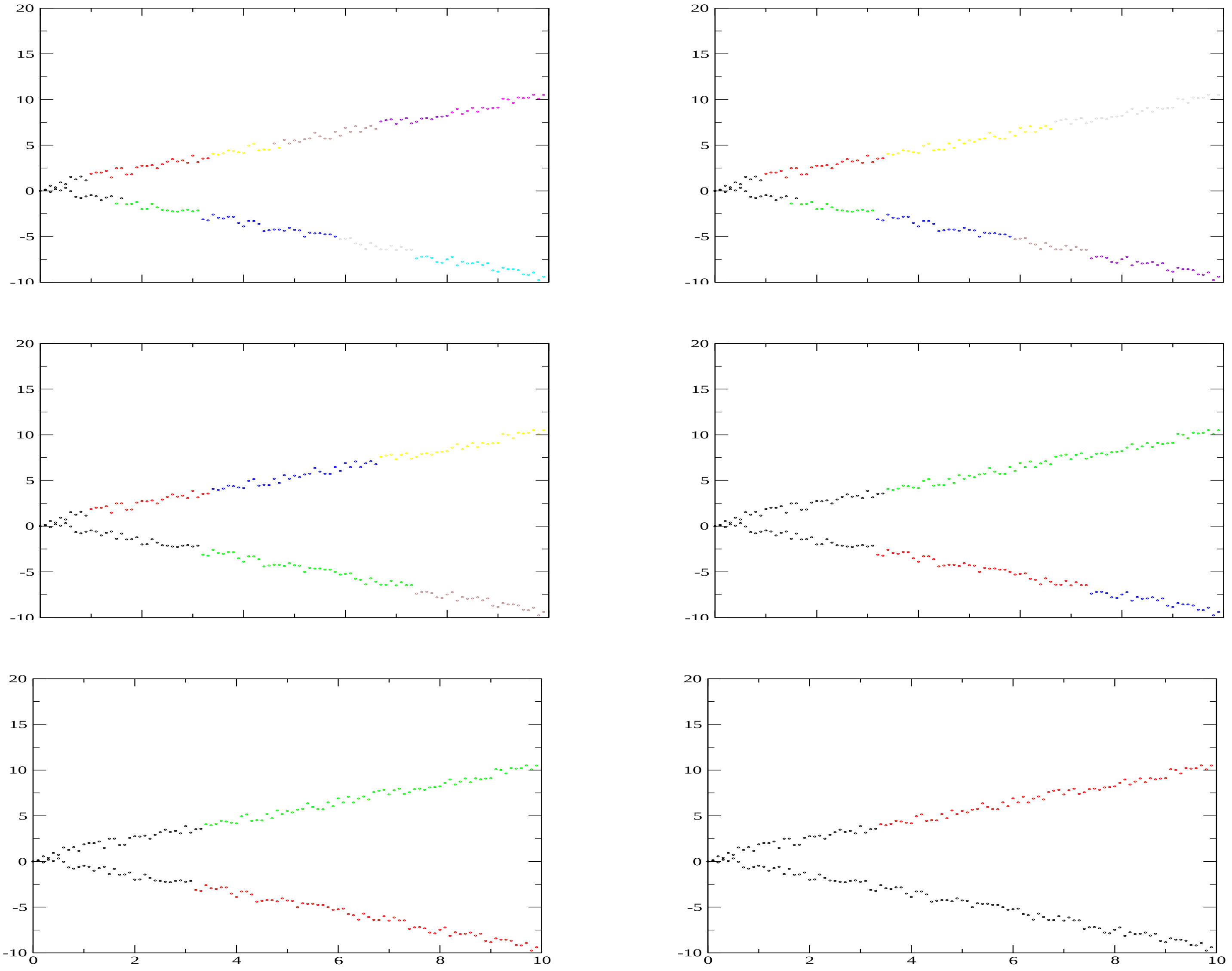
H_0: mean prcp forecast in a cluster is right.
If $|T| \geq 2$, then reject H_0, i.e. wrong forecast.
If $|T| < 2$, then no evidence for rejecting H_0, i.e. forecast=OK.

	x-y-p (3,2)			x-y-p (4,6)		
	Observed	Forecast	T	Observed	Forecast	T
Black	52.6 ± 22.0	59.1 ± 31.0	-0.2	52.9 ± 23.0	56.3 ± 20.2	-0.1
Red	-	-	-	179.00 ± 0.0	141.8 ± 28.0	1.3
Green	36.0 ± 6.3	41.8 ± 11.3	-0.4	36.0 ± 6.3	41.8 ± 11.3	-0.4
Blue	-	-	-	52.3 ± 20.7	38.26 ± 7.1	0.6

Summary

CA → objective/automated defn of object/entity.
The clusters agree with expert opinion.
CA supplemented to match clusters.
CA supplemented to compare fields.

Future Work



Model-based Cluster Analysis.