Before we continue on with the analysis of the month's worth of trajectories using cluster methods, I would like you to go back to the frequency analysis section, and save the CONTROL file that we had used, that we are created to do the month's worth of trajectories. So go to the trajectory tab, setup run, and you'll see if this is the correct example, it started on September 1st, so go ahead and do a save as, to traj_freq_control.txt. You may need this file later on.

Now another way of analyzing a lot of trajectories is to use cluster analysis. In the previous section, the trajectories from multiple simulations were counted as they passed over an arbitrary grid. Another approach is to merge trajectories that are near each other and represent those groups, called clusters, by their mean trajectory. Differences between trajectories within a cluster are minimized while differences between clusters are maximized. Computationally the trajectories are combined until the total variance of the individual trajectories about their cluster mean starts to increase. This occurs when disparate clusters are combined. The clustering computation is described in more detail in the next section. Here we will do the calculation first.

So go ahead and open up the graphical user interface. Now we are assuming that you have already computed one month's worth of trajectories in the previous section. If you have not, then go back and do the month's worth of calculations using the run daily menu tab.

Click on trajectory, special runs, clustering, standard. This menu contains several options that we need to change. The first is the default in the menu is 36, we know our trajectories are 48 hours long, so we want to look at the entire length of the trajectory, when we compare trajectories. Remember I said that the clustering looks at the difference between two trajectories.

You can go from the origin point out to as many hours as you would like. In this case, we're going all the way to look at every hour, and we're not going to skip any trajectories. So we would look at every trajectory. The endpoints folder, the folder where the trajectory endpoints are to be found is in the hysplit4/working directory. We do not need to change the working folder or the archive folder at this point. The base name, the wild card name for looking at these files, if as you recall, they started with fdump, followed by a starting time, so the wild card base name will be fdump. And the first step, of course, is to make the INFILE. Now this is identical to the method we used in the previous section, so you would open up the windows explorer to the INFILE and make sure that there are no file names that do not match. But in this case, we set the working directory in cluster, and there is the INFILE, and you can see the same issue here is that the fdump file without a identifying start time is not wanted, so we delete that and do a save. Right, so if we had changed the working folder to hysplit4/working here, then the intermediate results would've been written to the hysplit4/working, but instead they are written to hysplit4/cluster/working.

The next step is to run the cluster analysis. So it looked at 112 files but found 111 of them usable, and the clustering is complete. Now at this point we can display the total spatial variance.

When the calculation starts, the number of clusters would be equal to 112 or 111, and a total spatial variance would be very small because each trajectory is its own cluster. But as the search procedure during clustering, that is the calculation will look at the spatial variance between one trajectory and all other trajectories, and find the pair of trajectories that have the minimum spatial variance between those two trajectories and merge those two trajectories together to give you a mean trajectory, reducing the total number of clusters from say 111 to 110. So one of the clusters at that point has two trajectories and all the other clusters still have one. And this process continues, that trajectories are merged until the total spatial variance starts to increase. And this is the area that we're interested in, so here we have 1, 2, 3, this cluster here, is the fourth, so we have 4 clusters. After 4 clusters we really start, you can say four or five, we start a rapid increase in the total spatial variance.

So if I were to select four clusters as the number that we're interested in, and now we do run, the code assigns each trajectory to a particular cluster number. So these first trajectories are assigned to cluster number one and then this group of trajectories is assigned to cluster number two. So cluster number two has a mean trajectory

associated with it and the spatial variance for cluster number two is the variance of each one of these trajectories that are contained within cluster number two from the mean trajectory. So you can imagine if you have …, and then the total spatial variance is that you add the variances of all the clusters. So you can imagine if you had a trajectory … let's go on and look at the picture and then I'll go back and make that explanation.

So let's display the cluster means. So these are now the four trajectories, the four cluster mean trajectories. You can see that cluster #1 the trajectories go to the southwest, and that cluster #2 to the north, and #3 to the east, and #4 to the northeast.

And the point here, from the standpoint of the CAPTEX experiment, 38% of the trajectories went over Pennsylvania, which is the downwind domain, and 22% went to the north, northeast. So if you look at these two trajectories, that really represents over 50% of the flow regimes, somewhere going over, across this CAPTEX domain during September, it makes it a very favorable period for doing this experiment, if we selected Dayton as the starting location.

The point I was making about the cluster means, so for instance, if I were to combine cluster one with cluster two, the mean trajectory, you know, might be some trajectory that really goes nowhere, just hangs around the source point, which means that when you compute the total spatial variance, the spatial variance of the individual

trajectories that makeup that particular merged cluster, it's going to go up. It will be much higher because these individual trajectories going to the northeast and the southwest are so much different from the cluster mean. And you can look at those individual trajectories as well. So these are the individual trajectories that compose cluster one, cluster two, the ones going to the north, cluster three, and cluster four.

So you can redo this with different number of clusters to see how it changes, but in the next action we will review the computational method in a little more detail.