

How To: Prepare for, carry out and display a cluster analysis in ArcMap 9.3 using the Ward's cluster analysis technique

This paper aims at providing a useful step-by-step guide on how to display potential geographic clusters of political subculture utilizing the ESRI software ArcMap version 9.3 on the basis of a cluster analysis. As such, the paper will provide a practical, hands-on introduction to the process of displaying the results of a cluster analysis in ArcMap. It will, however, do so without introducing the formal theory behind cluster analysis itself¹. To be able to display any kind of cluster analysis in ArcMap 9.3 two elements must necessarily come first; 1) preparing for and 2) carrying out the cluster analysis. This paper thus consists of three parts: first part concerns the preparation for carrying out a cluster analysis, second part concerns carrying out the cluster analysis using a standard statistical software package and third part concerns the issue of how to display the cluster analysis in ArcMap. This means that only the third part will actually be carried out in ArcMap. As a practical illustration of this procedure the paper will make use of an explicit research example, the research question of which is: *“looking at every European Parliament election in Denmark between 1979 and 2004 is there any indication of neighboring nomination districts clustering together on the basis of regionally equal party swings from one election to the next?”* In other words, is it possible to find somewhat *geographically* stable political subcultures within a Danish context in the light of a so-called cultural component in the voting behavior as some theory in the field of voting behavior has suggested².

¹ For a great introduction to the theory behind various kinds of cluster analyses you may consult Everitt, Brian S., Sabine Landau & Morven Leese (2001):“Cluster Analysis”, Institute of Psychiatry, Kings College London.

² See for example Thomsen, Søren R. (1987). ”Danish elections 1920-79. A logit approach to ecological analysis and inference”. Århus: Politica, Thomsen, Søren R., & Jørgen Elklit (2007). *Hvad betyder de personlige stemmer for partiernes tilslutning?* In J. Goul Andersen, & J. Andersen (Eds.), *Det nye politiske landskab. folketingsvalget 2005 i perspektiv* (pp. 307-334). Århus: Academica and Thomsen, Søren R. i Jørgen Elklit og Jens Blom-Hansens *Perspektiver på Politik*, pp. 57-63.

1) Preparing for the cluster analysis

The first step in carrying out a cluster analysis must necessarily include the preparation of data for use in the cluster analysis. In this example we will make use of the Danish election data from Den Danske Valgdatabase³. Specifically we want voting data for all European Parliament elections held in Denmark from 1979 to 2004 with the Danish nomination districts serving as the geographic reference unit on which the number of votes for each party is aggregated⁴. When downloading data from Valgdatabasen, data will be delivered in an Excel-file in long format with party specific voting data from each election placed after each other in the same columns. For one to be able to use each election year as an independent case in the following cluster analysis, data must be transformed into wide format such that variables from each election year have their own columns. Having done this, data should be converted into SPSS, Stata or any other file format that is capable of doing data manipulation and further statistical analysis. Using the StatTransfer software will make this conversion quite easy.

2) The cluster analysis

Next, the cluster analysis can be carried out using one of various cluster analysis techniques available in many statistical software packages. In this example, we will utilize one of SPSS's built-in techniques, namely the Ward's cluster technique⁵. It is found by clicking the "Analysis"-option in the main menu bar of the SPSS-window and navigating to the "Classify"-option and from here selecting the "Hierarchical Cluster Analysis"-option. In this case we are interested in finding clusters of potential *geographically coherent* political subcultures on the basis of changes in logit-transformed vote percentages for all parties within each nomination district from one election to the next for more election periods. Therefore, we first have to make percentages of all the parties' vote share within each nomi-

³ Den Danske Valgdatabase is part of the PEDA project. It is a scientific database containing Danish election data for each party running for election in each type of election, be it a National Parliament election, a European Parliament election or a national referendum, held in Denmark since 1979. It is regularly updated with the most recent election data.

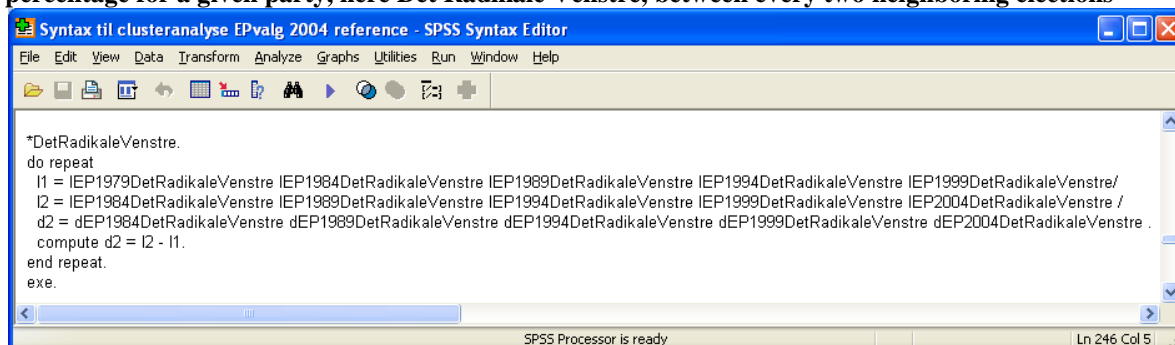
⁴ The geographic extent of these nomination districts was maintained up until January 1st 2007 when a major structural reform changed the administrative landscape of Denmark. The 103 former nomination districts which are used in this example were replaced by 92 new nomination districts.

⁵ For a detailed description of this specific cluster technique consult for example Everitt, Brian S., Sabine Landau & Morven Leese (2001): "Cluster Analysis", Institute of Psychiatry, Kings College London.

nation district for each election year and next to logit-transform these percentages. This can straightforwardly be done by the use of “do repeat”-“end repeat” commands in the SPSS-syntax. Having run such commands, the necessary logit-transformations (i.e. changes in logit-transformed vote percentage for each party within each nomination district between each two on each other following elections) are now available and the Ward’s cluster analysis can subsequently be carried out in SPSS.

An alternative way of making the same Ward’s cluster analysis exists. This involves constructing the logit-transformed vote percentages within SPSS only and then do the Ward-making within Stata instead. Is this the procedure chosen, then after having created the logit-transformed vote percentages, it is necessary in SPSS to compute a set of new variables (d_1 - d_n) for each party that has run for a European Parliamentary election at least twice between 1979 and 2004, where each new party specific set of variables (d_1 - d_n) contains the differences in logit-transformed vote percentage for that party between every two neighboring elections. For an example of this “do repeat” - “end repeat” command, see the following piece of SPSS syntax in figure 1 for the Danish party Det Radikale Venstre, which produces five new variables (dEP1984DetRadikaleVenstre, dEP1989DetRadikaleVenstre, dEP1994DetRadikale Venstre, dEP1999DetRadikaleVenstre, dEP2004DetRadikale-Venstre) because Det Radikale Venstre has run at all six European Parliamentary elections between 1979 and 2004, which again implies that there are five changes in the vote percentage for Det Radikale Venstre to be estimated.

Figure 1: SPSS-syntax creating party specific sets of variables as differences in logit-transformed vote percentage for a given party, here Det Radikale Venstre, between every two neighboring elections

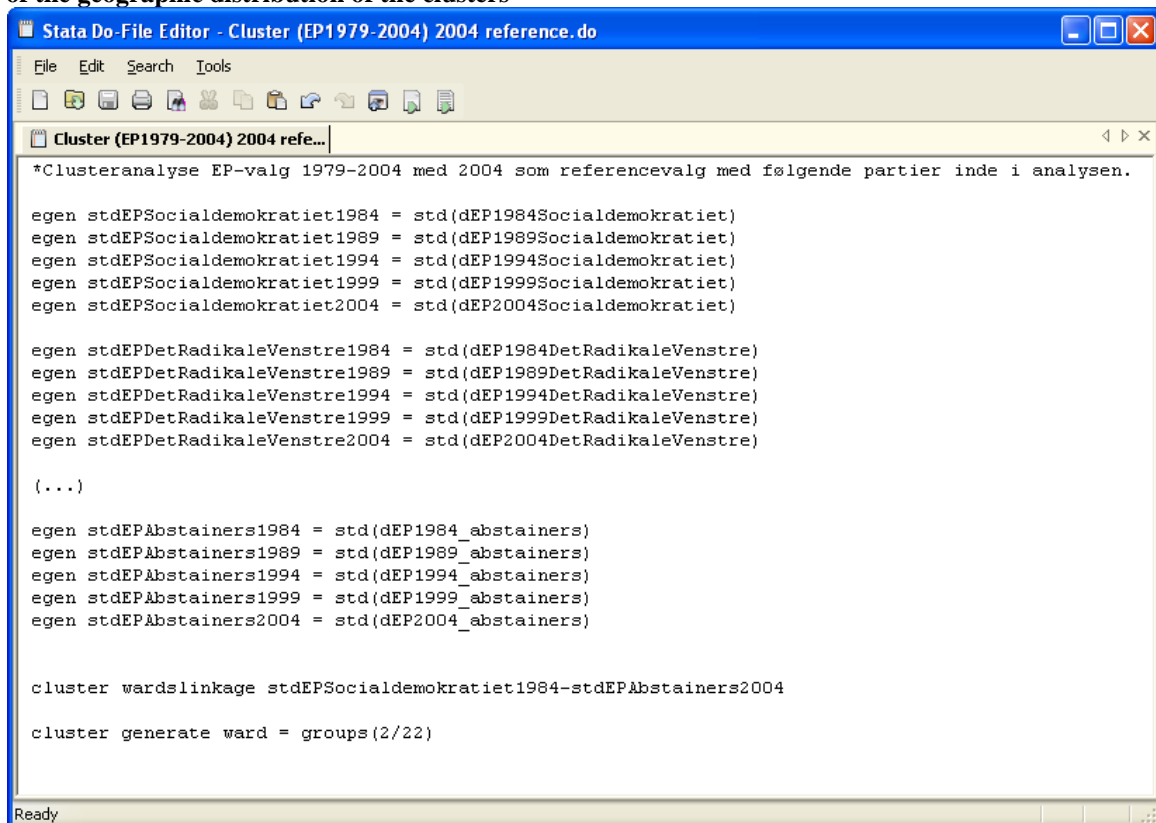


```
*DetRadikaleVenstre.
do repeat
  I1 = IEP1979DetRadikaleVenstre IEP1984DetRadikaleVenstre IEP1989DetRadikaleVenstre IEP1994DetRadikaleVenstre IEP1999DetRadikaleVenstre /
  I2 = IEP1984DetRadikaleVenstre IEP1989DetRadikaleVenstre IEP1994DetRadikaleVenstre IEP1999DetRadikaleVenstre IEP2004DetRadikaleVenstre /
  d2 = dEP1984DetRadikaleVenstre dEP1989DetRadikaleVenstre dEP1994DetRadikaleVenstre dEP1999DetRadikaleVenstre dEP2004DetRadikaleVenstre .
compute d2 = I2 - I1.
end repeat.
exe.
```

When having run the SPSS-syntax that creates these party specific sets of variables, you should save the newly created SPSS data in Stata-file format and then open it from within Stata. In Stata, the same Ward-variables (containing potential geographic clusters of political

subculture) that are made using the Ward's cluster technique in SPSS can now be made by typing a few commands in the do-file, see the do-file example in figure 2 below. The number of Wards to be created is a choice which must be made by the researcher using a certain command in the do-file, see the last line within figure 2. In this example we have chosen to create twenty-one Ward-variables beginning with a variable which is restricted to contain two Wards (Ward2), i.e. containing two potential geographic clusters, and ending with a variable (Ward22) which is restricted to contain twenty-two potential geographic clusters.

Figure 2: Stata Do-file generating the Ward-variables to be used subsequently in ArcMap for a display of the geographic distribution of the clusters



```
Stata Do-File Editor - Cluster (EP1979-2004) 2004 reference.do
File Edit Search Tools
Cluster (EP1979-2004) 2004 refe...
*Clusteranalyse EP-valg 1979-2004 med 2004 som referencevalg med følgende partier inde i analysen.

egen stdEPSocialdemokratiet1984 = std(dEP1984Socialdemokratiet)
egen stdEPSocialdemokratiet1989 = std(dEP1989Socialdemokratiet)
egen stdEPSocialdemokratiet1994 = std(dEP1994Socialdemokratiet)
egen stdEPSocialdemokratiet1999 = std(dEP1999Socialdemokratiet)
egen stdEPSocialdemokratiet2004 = std(dEP2004Socialdemokratiet)

egen stdEPDetRadikaleVenstre1984 = std(dEP1984DetRadikaleVenstre)
egen stdEPDetRadikaleVenstre1989 = std(dEP1989DetRadikaleVenstre)
egen stdEPDetRadikaleVenstre1994 = std(dEP1994DetRadikaleVenstre)
egen stdEPDetRadikaleVenstre1999 = std(dEP1999DetRadikaleVenstre)
egen stdEPDetRadikaleVenstre2004 = std(dEP2004DetRadikaleVenstre)

(...)

egen stdEPAbstainers1984 = std(dEP1984_abstainers)
egen stdEPAbstainers1989 = std(dEP1989_abstainers)
egen stdEPAbstainers1994 = std(dEP1994_abstainers)
egen stdEPAbstainers1999 = std(dEP1999_abstainers)
egen stdEPAbstainers2004 = std(dEP2004_abstainers)

cluster wardslinkage stdEPSocialdemokratiet1984-stdEPAbstainers2004

cluster generate ward = groups(2/22)

Ready
```

3) Displaying potential geographic clusters in ArcMap

At this point both procedures for creating clusters using the Ward's cluster technique have been described - the first one using SPSS only and the second one using both SPSS and Stata. We are now finally ready to use ArcMap 9.3 as the software to create our final geographic display of the cluster analysis. Before we can make our final display, however, two things

must be done: 1) Any attribute data contained in a standalone table (i.e. a table without a direct geographic representation attached as opposed to a shapefile), as is the case with our Ward-variables containing a value for each nomination district in the SPSS- or Stata-file, must be converted into dBase-file format before it can be added to and thereby used in ArcMap. This can easily be done using StatTransfer (remember, however, not to save too many variables, since dBase can only handle 255 at a time!). And 2) when having opened ArcMap 9.3 and added both the shapefile of interest, i.e. the shapefile of the 103 former nomination districts, as well as the newly made dBase-file containing the Ward-variables, a so-called “Join”-operation must be carried out. The “Join”-operation links one or more table attributes in a standalone table to a geographic boundary file (a layer-file/shapefile), that is, by using the “Join”-operation our potential geographic cluster data can now be linked to the respective nomination districts in the shapefile of the 103 former Danish nomination districts. For this “Join”-operation to succeed, however, a unique key variable linking the voting data in the dBase-file to the nomination district features in the shapefile must exist. In this case the unique key variable will be the former nomination district code found in both the dBase-file⁶ and the shapefile. Having successfully made the join, we are finally ready to find out if our voting data reveals *geographically bounded* clusters of political subculture through time.

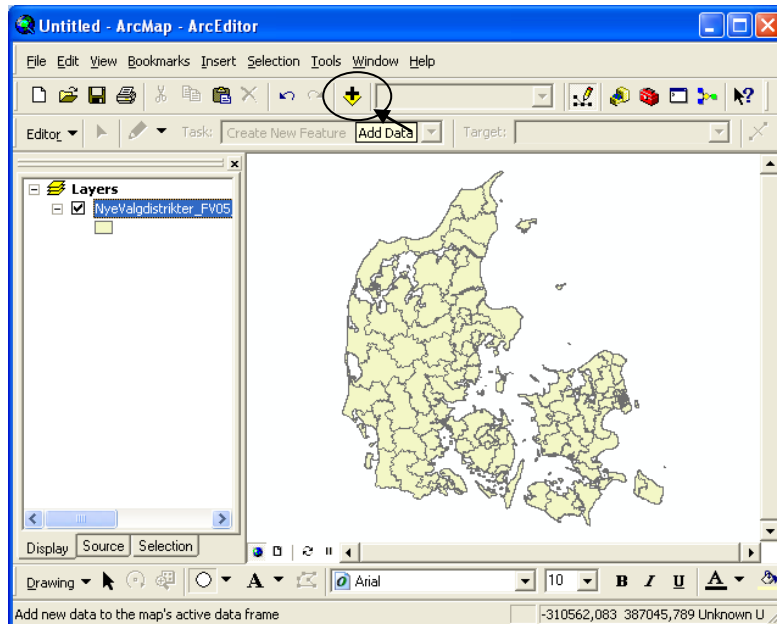
3.1) How to proceed in ArcMap

Open ArcMap 9.3 (or an earlier version) and choose to start with a new empty map. Then add to the empty data frame the shapefile containing the 103 former nomination districts of Denmark⁷ as well as the dBase-file containing the Ward-variables. To add this data to your map you should select either the “Add Data”-icon (depicting a black addition sign on top of a yellow tilted square) positioned right above the main panel (see figure 3 below) or use the “File”-drop bar in the menu bar at the top of the screen and from there choose the “Add Data”-option.

⁶ A nomination district code was originally attached when downloading the voting data from Den Danske Valgdatabase.

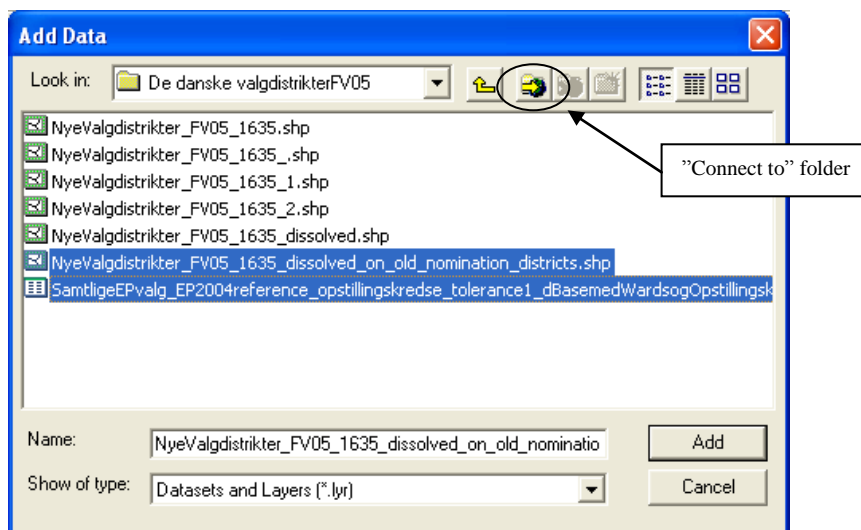
⁷ If your shapefile at the outset displays less aggregated features, e.g. the Danish polling districts, rather than the old nomination districts, on which the cluster analysis will be based, a “Dissolve”-operation should first be carried out. For a guide on how to use the “Dissolve”-function, consult the paper “How To: Make a map displaying comparable voting districts through time using 0-tolerance data from Den Danske Valgdatabase and ArcMap 9.3” by Malene Rode Larsen (O:\PEDA\GIS\Introductory papers to ArcMap 9.3).

Figure 3: One way to add data for analysis in ArcMap is to use the “Add Data”-icon



Choosing either option, an “Add Data”-window pops up, see figure 4 below. You must now navigate to the relevant folder(s) where the layer-file(s)/shapefile(s) and dBase-file(s) of interest are located.

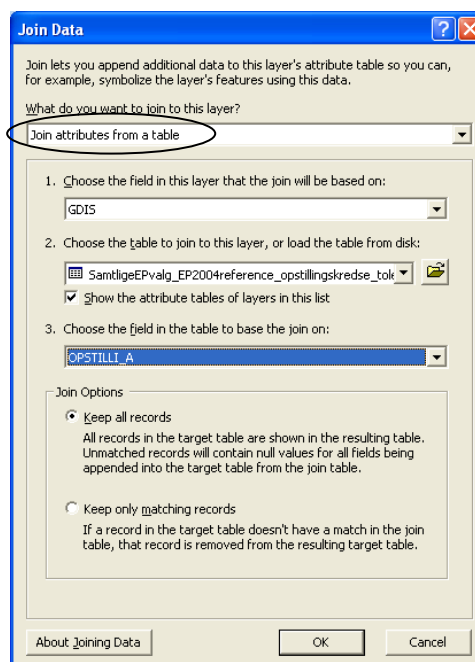
Figure 4: The “Add Data”-window



In case the relevant drive (e.g. the O-drive) from where you need to pick your data is not readily accessible, you will have to establish a connection to this drive in the program's file ordering system. This is done by first clicking the "Connect To"-folder icon inside the "Add Data" pop-up-window (depicting a yellow arrow pointing to a globe), see figure 4 above, and next by navigating to and choosing the appropriate drive and folder (in this case we will choose O:\PEDA\ Clusteranalyse Danmark EPvalg 1979-2004\Shape files and layer files of Denmark). Now add the relevant data, i.e. the shapefile named "NyeValgdistrikter_FV05_1635_dissolved_on_old_nomination_districts.shx".shp as well as the dBase-file named "SamtligeEPvalg_EP2004reference_opstillingskredse_tolerance1_dBasemedWardsog OpstillingskredsNr.".dbf to the empty data frame and ignore for the present purpose any potential warnings of a lack of spatial reference information.

When all necessary data have been added to the data frame, right-click the relevant layer-file/shapefile which you want to join data *to* (in this instance the shapefile of the 103 former nomination districts) and navigate to the "Join"-option via the "Joins and Relates"-option in the drop bar. The "Join Data"-window pops up as shown below. Fill out the window as shown (GDIS is the name of the variable containing the former nomination district codes in the shapefile's attribute data, whereas Opstilli_A is the former nomination district code-variable in the dBase-file).

Figure 5: The "Join Data"-window



Clicking the “OK”-button at the bottom of the “Join Data”-window, the Ward-variables containing the results of the cluster analysis are now part of the shapefile’s attribute table and are as such ready to be represented graphically. An example is given in figure 6 below for the variable WARD14 which contains up to fourteen potential geographic clusters. It is worth mentioning here that a “join” is not a permanent operation in ArcMap and that the join can easily be removed again by right-clicking the layer-file/shapefile upon which the join was made and then selecting the “Remove Join(s)”-option from the “Joins and Relates”-drop bar. The join, however, can also be made permanent. This is done by right-clicking the layer-file/shapefile and from there selecting “Export Data” from the “Data”-drop bar and then saving the layer-file/shapefile under a new file name.

Joining the Ward-variables to the shapefile does nothing to the graphic representation of the former nomination districts right away. All geographic features (the nomination districts) are still displayed with a single random color as in the default setting, see figure 3 above. However, since the Ward-variables containing the desired attribute, that is, from two to twenty-two potential geographic clusters, are now part of the shapefile’s attribute table, the shapefile can be coloured on these variables.

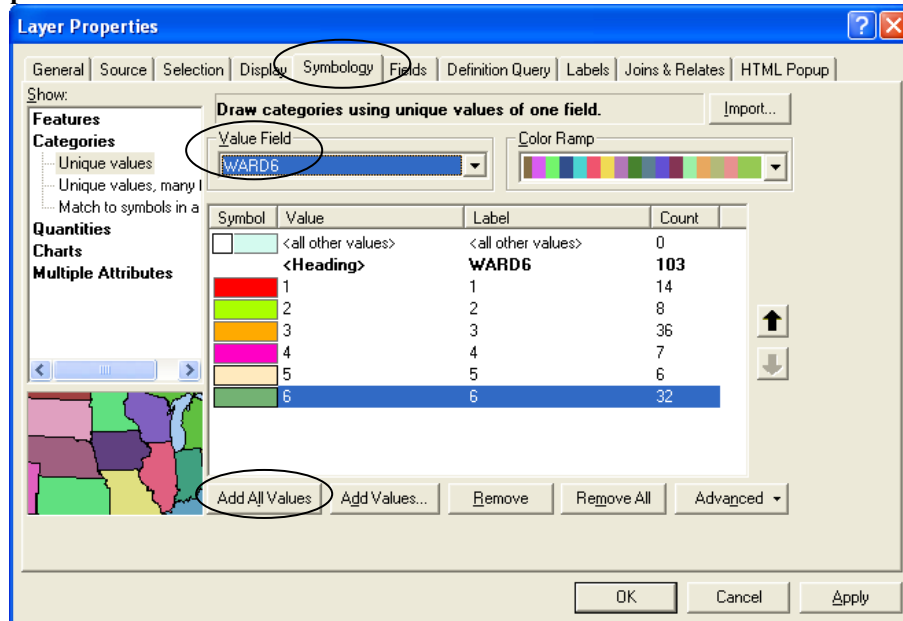
Figure 6: The shapefile’s “Attribute Table” showing the joined Ward-variables each containing a different numbers of potential geographic clusters

The screenshot shows the 'Attributes of NyeValgdistrikter_FV05_1635_dissolved_on_old_nomination_districts' window. The table contains 103 records, each representing a nomination district. The columns include WARD12 through WARD21, OPSTILLING, and OPSTILLI A. The data shows the number of potential geographic clusters for each ward variable across different nomination districts.

WARD12	WARD13	WARD14	WARD15	WARD16	WARD17	WARD18	WARD19	WARD21	OPSTILLING	OPSTILLI A
12	12	13	14	15	15	16	16	18	Varde	66
12	13	14	15	16	17	18	19	21	Esbjerg	67
12	12	13	14	15	15	16	16	18	Ribe	68
12	12	13	14	15	15	16	16	18	Grindsted	69
7	7	8	9	10	10	11	11	12	Fredericia	70
7	7	8	9	10	10	11	11	12	Kolding	71
7	7	8	9	10	10	11	11	12	Vejle	72
7	7	8	9	10	10	11	11	12	Givø	73
7	7	8	9	10	10	11	11	12	Juelsminde	74
7	7	8	9	10	10	11	11	12	Horsens	75
12	12	13	14	15	16	17	17	19	Ringkøbing	76
12	12	13	14	15	16	17	17	19	Holstebro	77
12	12	13	14	15	16	17	17	19	Herning	78
12	12	13	14	15	16	17	17	19	Skjern	79
5	5	6	6	7	7	8	8	9	Ørhus Øst	80
5	5	6	6	7	7	8	8	9	Ørhus Nord	81
5	5	6	6	7	7	8	8	9	Ørhus Syd	82
8	8	9	10	11	11	12	12	13	Ørhus Vest	83
12	12	13	14	15	15	16	16	18	Mariager	84
12	13	14	15	16	17	18	19	21	Randers	85
4	4	5	5	6	6	7	7	8	Hammel	86
12	12	13	14	15	15	16	16	18	Grenaa	87

The final task thus consists in colouring our shapefile on one of more of our Ward-variables. To do this right-click the shapefile and select the “Properties”-option at the bottom of the drop bar.

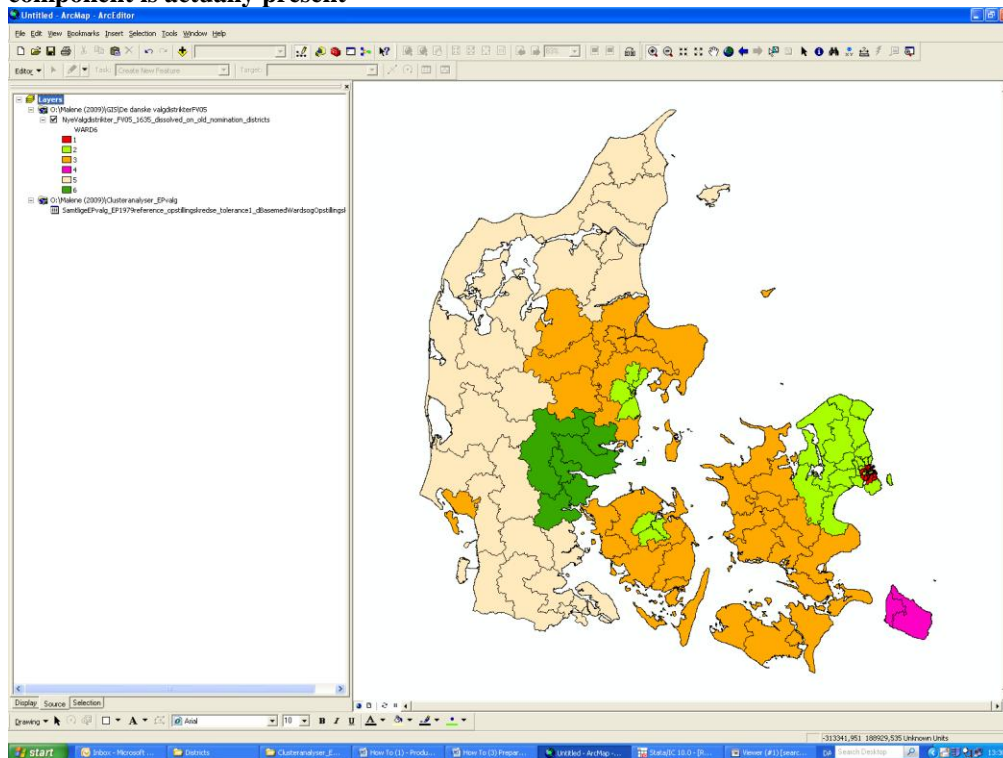
Figure 7: Using the “Layer Properties”-window to colour the potential geographic clusters



When in the “Properties”-window (see figure 7 above), choose the “Symbology”-tab. Inside this tab numerous different possibilities exist for colouring a layer-file/shapefile (for a thorough guide to the many possibilities for colouring features in ArcMap 9.3, consult the paper “How To: Colour geographic features and different types of attributes using ArcMap 9.3”). In this case we will choose the “Categories”-option with unique values found in the right panel of the “Symbology”-tab. We do this because the Ward-variables contain categories and because we will choose to colour on the WARD6-variable which contains “only” six different categories and that is not too many to distinguish from each other. From the “Value Field”-drop bar choose the variable WARD6 and click the “Add All Values”-button. Now, six distinct random colours will have been chosen for you and by clicking the “OK”-button you will return to the “Data View”. The shapefile will now be coloured in these six distinct random colours, see figure 8 below. Colours can also be chosen manually inside the “Symbology”-tab by double-clicking each colour icon or alternatively by double-clicking each of the colour icons displayed in the “Table of Content” to the right in the “Data View”.

This same procedure can of course be carried out for each of the other Ward-variables in the shapefile's "Attribute Table"⁸. For instance, colouring the shapefile on the WARD14-variable which we looked at in figure 6 means that fourteen different clusters will emerge on our map displaying politically stable subregions over time, and if the theory holds these politically stable subregions should also appear to be geographically bounded. Figure 8 below depict the cluster analysis using the WARD6-variable, that is, when restraining the number of potential geographic clusters to six. The map of the six clusters shows that there *is* actually a great number of geographically connected nomination districts. This cluster analysis example thus seems to be in support of the theory claiming that a geographically locally bounded cultural component is present in the Danish voting behavior over time.

Figure 8: The final display of the six politically stable subregions shows that a geographic component is actually present



Malene Rode Larsen, January 2010

⁸On how to insert a title, legend, north arrow, scale bar etc. on the final map and how to export the map, consult the paper "How To: Make a map displaying party support using ArcMap 9.3" by Malene Rode Larsen (O:\PEDA\GIS\Introductory papers to ArcMap 9.3)

References:

- Larsen, Malene Rode: “How To: Colour geographic features and different types of attributes using ArcMap 9.3”.doc (O:\PEDA\GIS\Introductory papers to ArcMap 9.3)
- Larsen, Malene Rode: “How To: Make a map displaying party support using ArcMap 9.3” (O:\PEDA\GIS\Introductory papers to ArcMap 9.3)