CLUSTERING WITH GIS: AN ATTEMPT TO CLASSIFY TURKISH DISTRICT DATA

Ece AKSOY
TURKEY

PURPOSE OF THIS STUDY;

- comparing different non-spatial and spatial clustering techniques and algorithms (K-Means and SOM) in GIS environment;
- putting forward the facilitative usage of GIS in regional and statistical studies.

INTRODUCTION, Steps

1. Clustering techniques were researched
2. Basic software were chosen for example application
3. Turkey's statistical datum was formed and analyzed joining with geographical data in GIS environment to perform clustering experiment

INTRODUCTION

Application area:

* All districts of Turkey,which is 923 units.

Limitating factors:

* All collected geographical data were assumed as limitations for clustering.

* NUTS population thresholds were taken as a limitation for district classification and this helped us for specifying cluster number.

PRINCIPAL PROBLEMS IN SPATIAL DATA CLASSIFICATION

- Large numbers of areas
- Large numbers of variables
- Non-normal variable distributions (most geographic data usually have very complex frequency distributions)
- Non linear relationships
- Spatial dependency
- Data uncertainty
- Small number problems (It is very important that small zones and small number effects should not dominate or dictate the characteristics of the spatial classification.)
- Variable specific levels of uncertainty
- Systematic non random variations in spatial representation

KOHONEN ALGORITHM and SELF ORGANIZING MAPS

- 'Kohonen Algorithm' and his 'Self-Organizing Maps (SOM)' is the most important spatial clustering technique.
- The main applications of the SOM are:

1. The visualization of complex data in a two-dimensional display,
2. Creation of abstractions like in many clustering techniques.

ADVANTAGES OF SOM

- Very simple to implement
- "Topology-preserving" feature superior to k-means methods,
- Can be very effective for visualizing high dimensional spaces,
- Can incorporate new data quickly,
- Most successfully used on large data sets,
- Small zones and number effects can be handled with the SOM.

CLUSTERING SOFTWARE

1. ArcGIS : Districting Module
2. Crimestat II: Hotspot Analysis II- K-Means Clustering
3. SPSS 11.0 : K-Means Clustering
4. Matlab 6.2 : SomToolbox2

GEOGRAPHICAL DATA

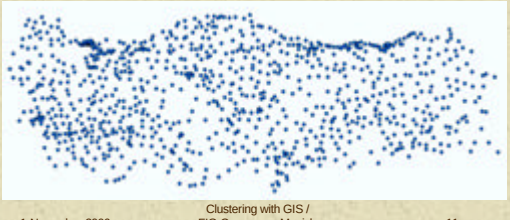
1. District Map (Polygon Data)
2. Districts Center Map (Point Data with Coordinate Data)
3. Physical Map of Turkey (Image from MRE)
4. Active Faults of Turkey Map (Image from MRE)
5. Geographical Borders Map (1st Congress of Geography, 1941, Polyline Data)
6. Basin Development Plans Maps (Polyline Data)
7. Contour Maps (Polygon Data)

DISTRICT MAP

- Contains polygon data of district's administrative borders. This map is actually base map for making thematic maps and analysis.

CENTER OF DISTRICTS MAP

- Contains point data of district's center and coordinates. This map is important for knowing where the definite settlement center of the districts and for understanding which districts are being interacted and communicate to each other.



PHYSICAL MAP OF TURKEY

- Physical map of Turkey is an image. It is useful for determining geographical borders and limits. Mountain ranges, rivers, river basins and lakes can be found out from this image.

ACTIVE FAULTS OF TURKEY MAP

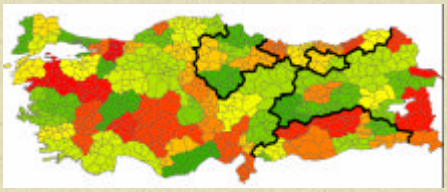
- Active faults of Turkey map is an image. It is useful for determining geographical borders and limits again.

GEOGRAPHICAL BORDERS MAP

- This is the first map for determining Turkey's geographical borders. It was made by geographers in the 'First Congress of Geography' in 1941. These divisions are not related with administrative borders, these are only related with geographical borders and limitations.

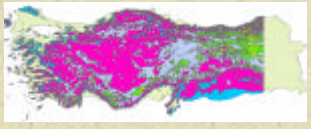
BASIN DEVELOPMENT PLANS

- This map is important for determining characteristics of under-developed province or groups while making classification of districts.



CONTOUR MAPS

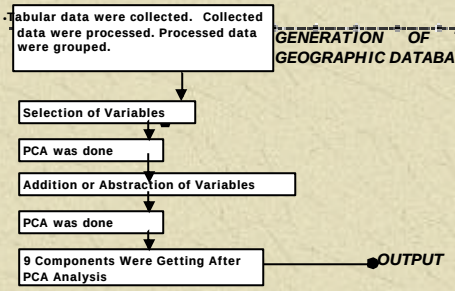
- Turkey's Border Map
- 300 Contour Map
- 600 Contour Map
- 1200 Contour Map
- 1800 Contour Map
- 2400 Contour Map
- 3000 Contour Map
- 3600 Contour Map



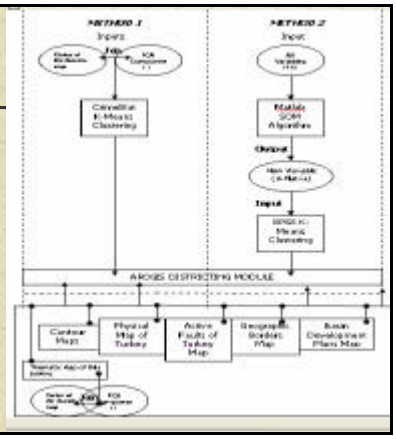
TABULAR DATA

- 142 variables were collected
- 36 variables were selected in point of view of consistency and reliability of indicators from between the whole data by means of district division.

DATABASE MANAGEMENT PROCESS



METHODS

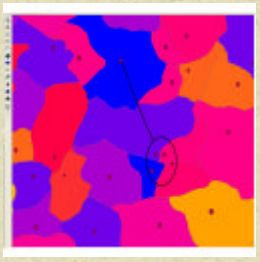


METHODS

- Each clustering techniques were applied individually and their outcomes were clustered in the Districting module by adding geographical data as layers for each method in the Districting module.
- Districts that show similar characteristic were grouped until between population thresholds.

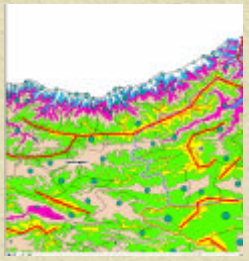
CENTER OF DISTRICTS MAP

- Some of the centers show grouping characteristic and becoming near. However, some of the districts border physically near but center of districts are in distant points.



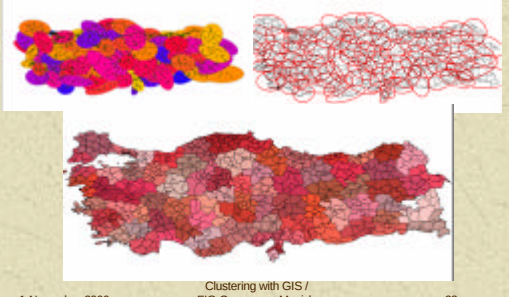
PHYSICAL LIMITATIONS

- Physical layers helped with showing geographically limits and groupings. If there is a mountain range between two districts that show same characteristics, they were not became group.



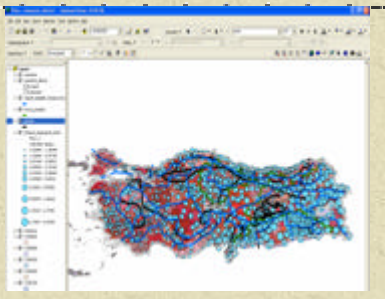
METHOD 1:

Component 1from PCA was used in K-Means Clustering analysis in CrimeStat.



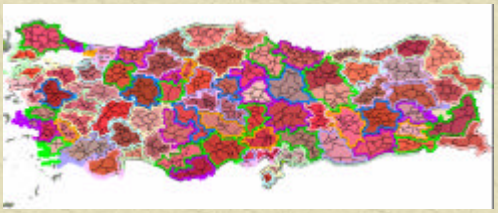
METHOD 1:

All Layers In the Same View Before Starting to Create CrimeStat Groupings



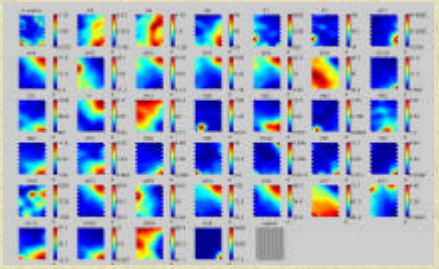
METHOD 1:

There were 84 new districts after classification

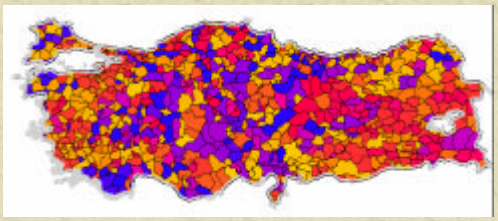


METHOD 2:

Tabular data with 36 variables was used in SOM analysis, not PCA components.

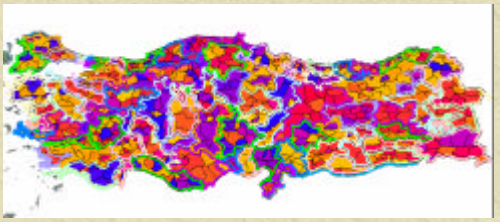


METHOD 2:



METHOD 2:

There were 87 new districts after classification



METHOD1

METHOD2

EVALUATION OF FINAL PRODUCTS

- Method 1 was the most quicker and easy method because; there have been already groupings thanks to software 'spatial' clustering routine in CrimeStat. The routine tries to find the best positioning of each centers and then assigns each point to the center that is nearest. Those groupings only were divided according to population thresholds actually.

EVALUATION OF FINAL PRODUCTS

- Method 2,had good suited and visibly groupings and best logical classification. The best clusters, which were the well fit with the geography, were obtained by the Method 2. Because the SOM algorithm is for 'spatial clustering', while it is calculating new values for total of the data it takes into consideration being neighbor. These new values are the final attribute of combination of all variables.
- But there is a problem of the step that joining with the final output of SOM algorithm and geographical data. After the making SOM algorithm procedure there is also something to need to visualize those values.

Thank You...