

Data Exploration and Spatial Statistics



ISTANBUL **TECHNICAL** UNIVERSITY
Sp. Anly. and Alg. in GIS
Week 8

Res. Assist. Ömer AKIN

Introduction & Aim of the Study

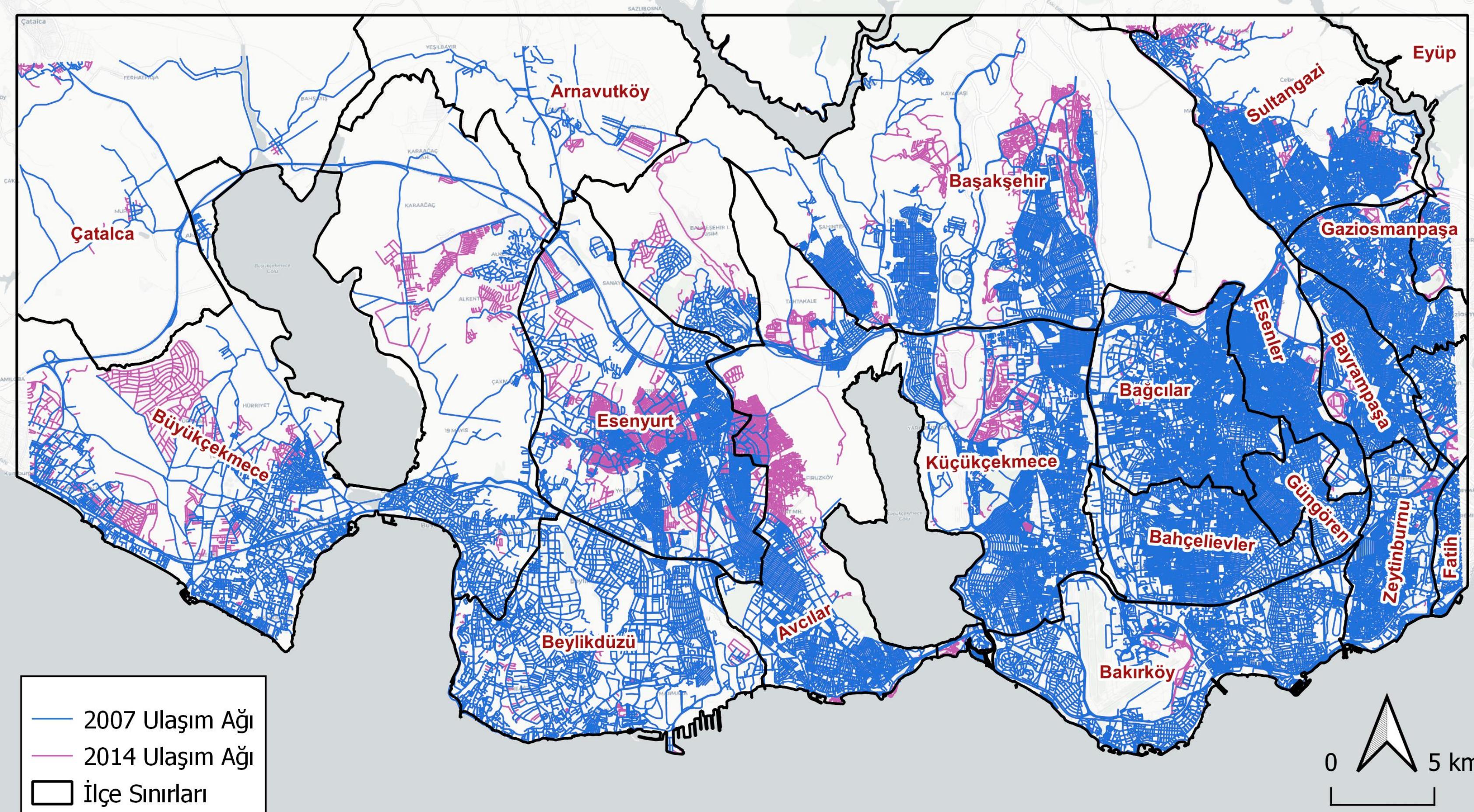
Aim of the Study:

- *Is there any relationship with population, land use and transportation?*
- *If yes, is this relationship meaningful, linear or can it be described quantitatively?*
- *Could this relationship shown spatially?*

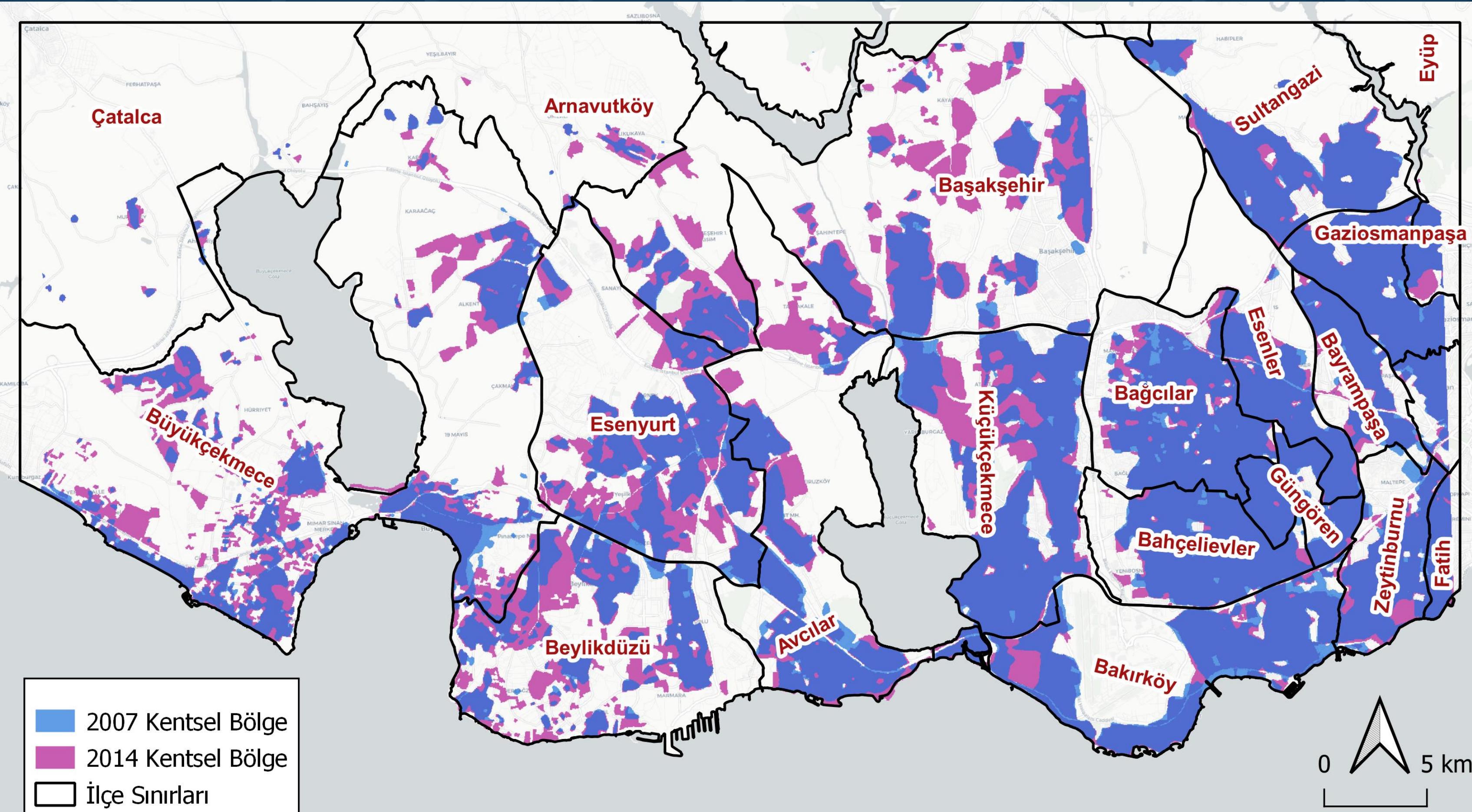
Input Data:

- *1km * 1km Grid that have (Vector-Polygon)*
 - *Population*
 - *Total Urban Areas*
 - *Total Road Length*
- *information for the years 2007 and 2014.*
- *District Boundaries (Vector-Polygon)*

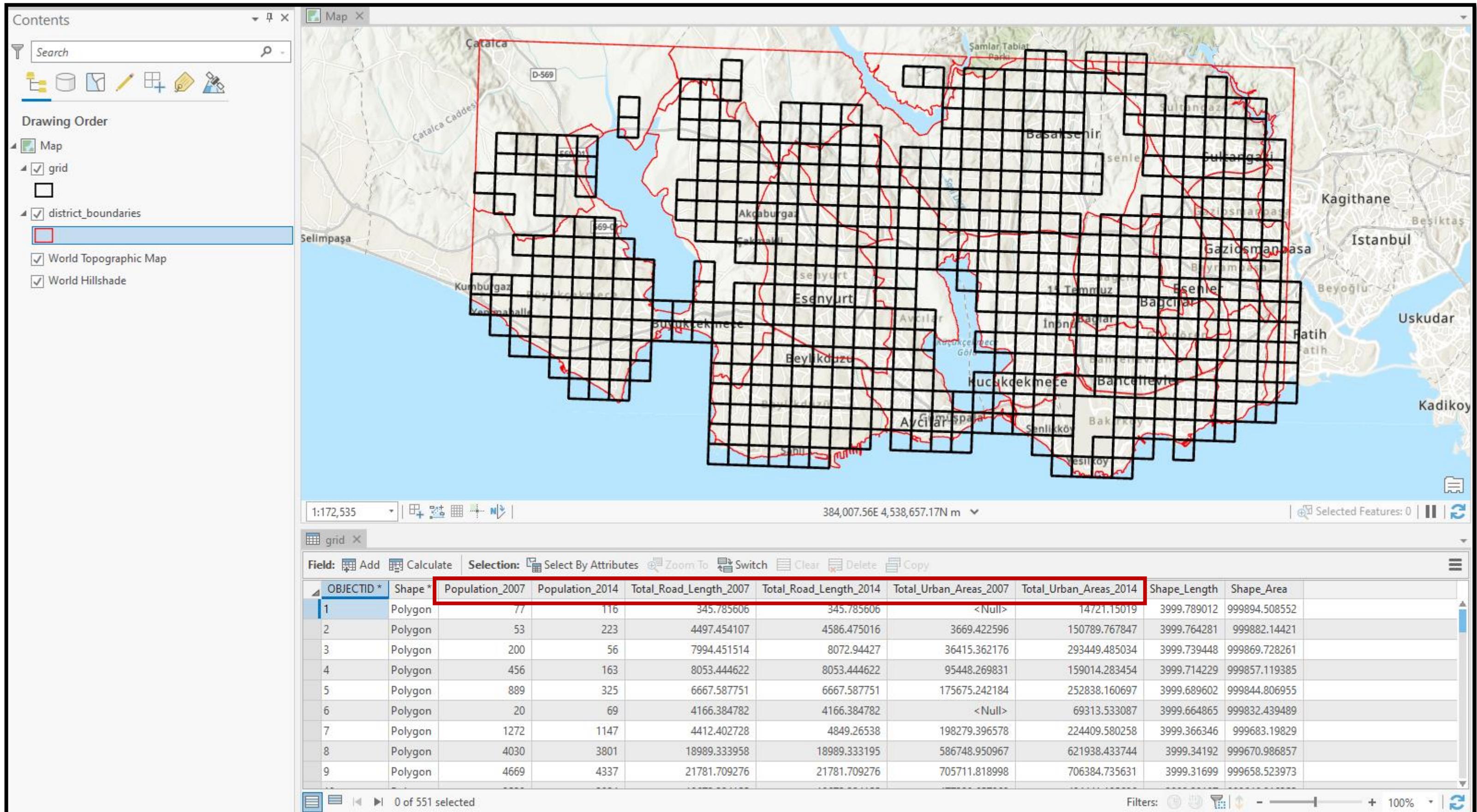
Study Area & Data



Study Area & Data



Exploring Data



Exploratory Spatial Data Analysis

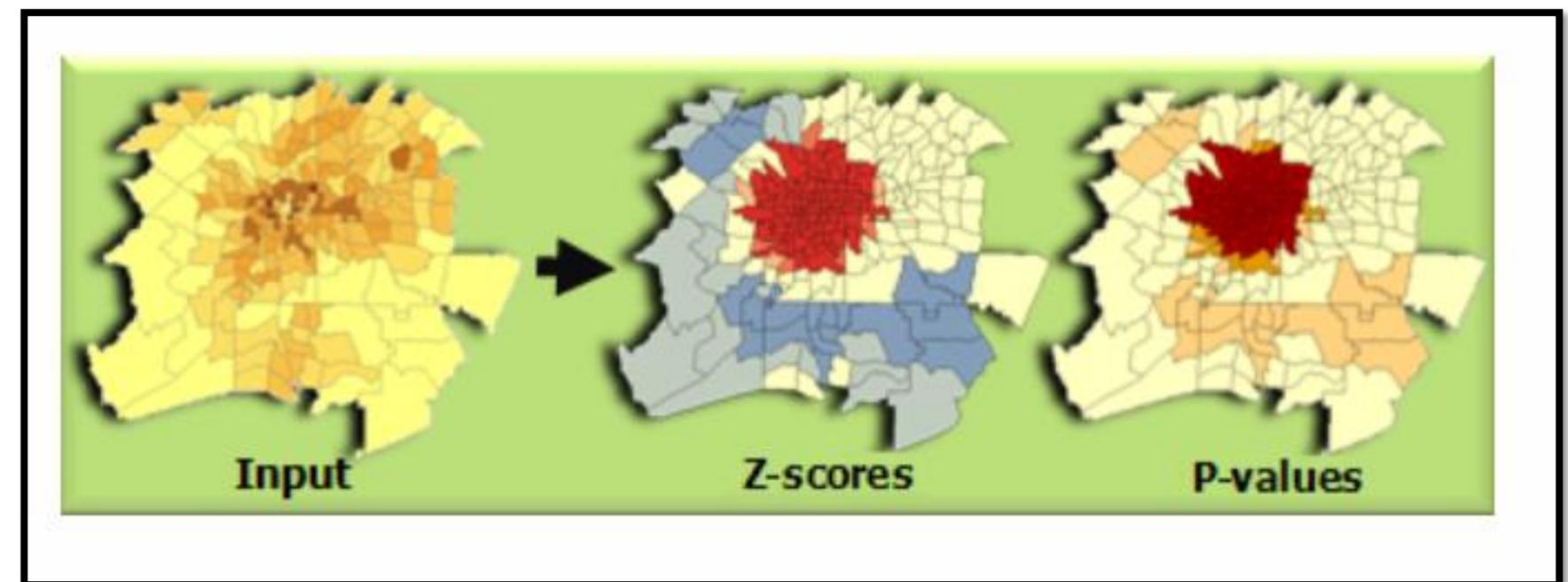
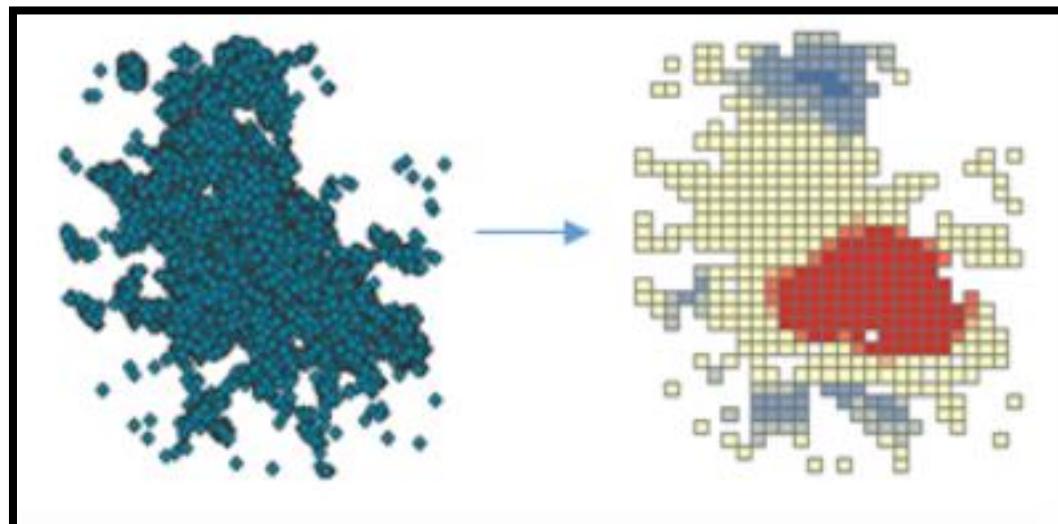
Hotspot Analysis



Examine the population value's statistical distribution on the study area by using hotspot analysis.

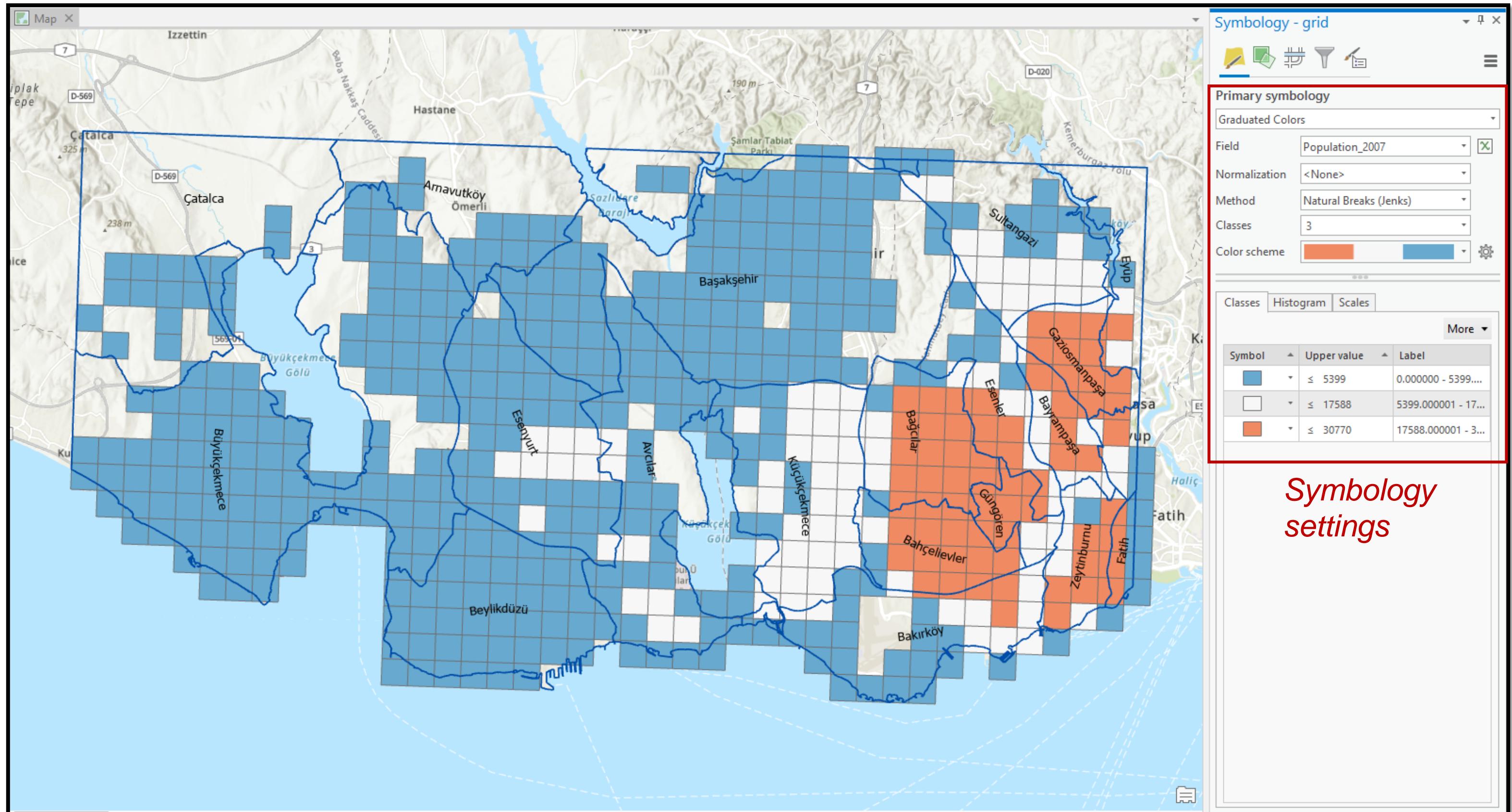
Hotspot Analysis

Given a set of weighted features, identifies statistically significant hot spots and cold spots using the Getis-Ord Gi^* statistic



Population Information on the Study Area

First, let's check the population value's distribution on the study area.



Exploratory Spatial Data Analysis

Hotspot Analysis



Geoprocessing

Hot Spot Analysis (Getis-Ord Gi*)

Parameters Environments

Input Feature Class: grid

Input Field: Population_2007

Output Feature Class: grid_HotSpots2007

Conceptualization of Spatial Relationships: Fixed distance band

Distance Method: Euclidean

Distance Band or Threshold Distance: [empty]

Self Potential Field: [empty]

Apply False Discovery Rate (FDR) Correction

Run

Conceptualization of Spatial Relationships

- Fixed distance band
- Inverse distance
- Inverse distance squared
- Fixed distance band
- Zone of indifference
- K nearest neighbors
- Contiguity edges only
- Contiguity edges corners
- Get spatial weights from file

Distance Method

- Euclidean
- Euclidean
- Manhattan

It's important to select appropriate algorithm for the case*

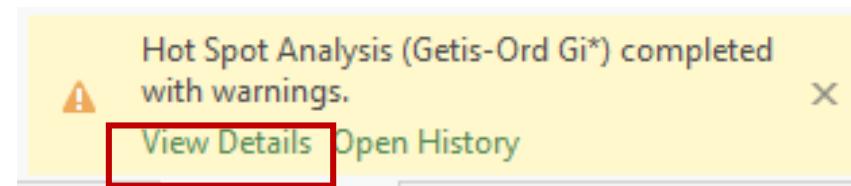
Green Road -> Euclidean Distance

Blue, Yellow and Red roads ->
Manhattan Distance

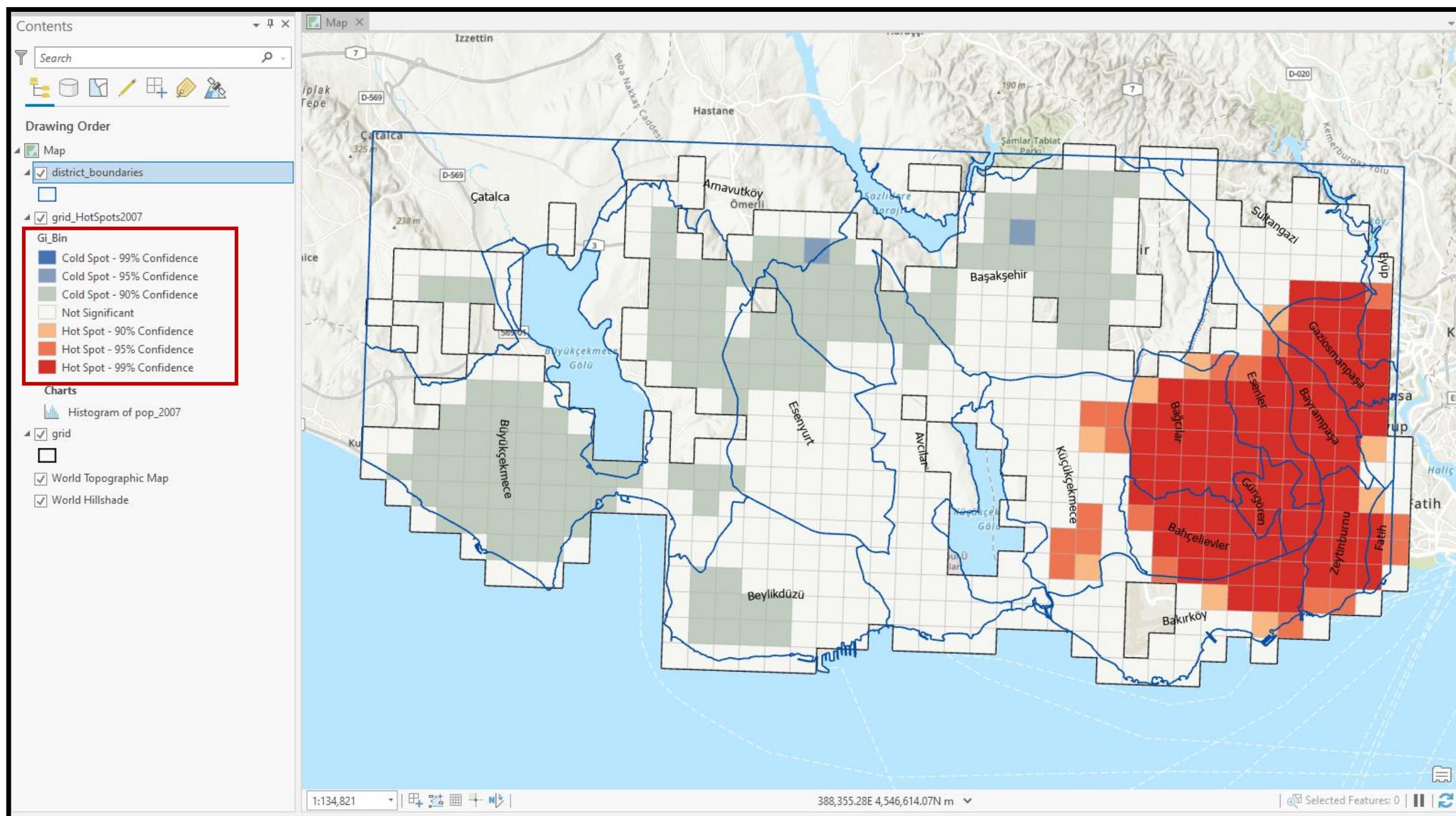
*To get more information about spatial relationship please visit:
<https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/modeling-spatial-relationships.htm>

Exploratory Spatial Data Analysis

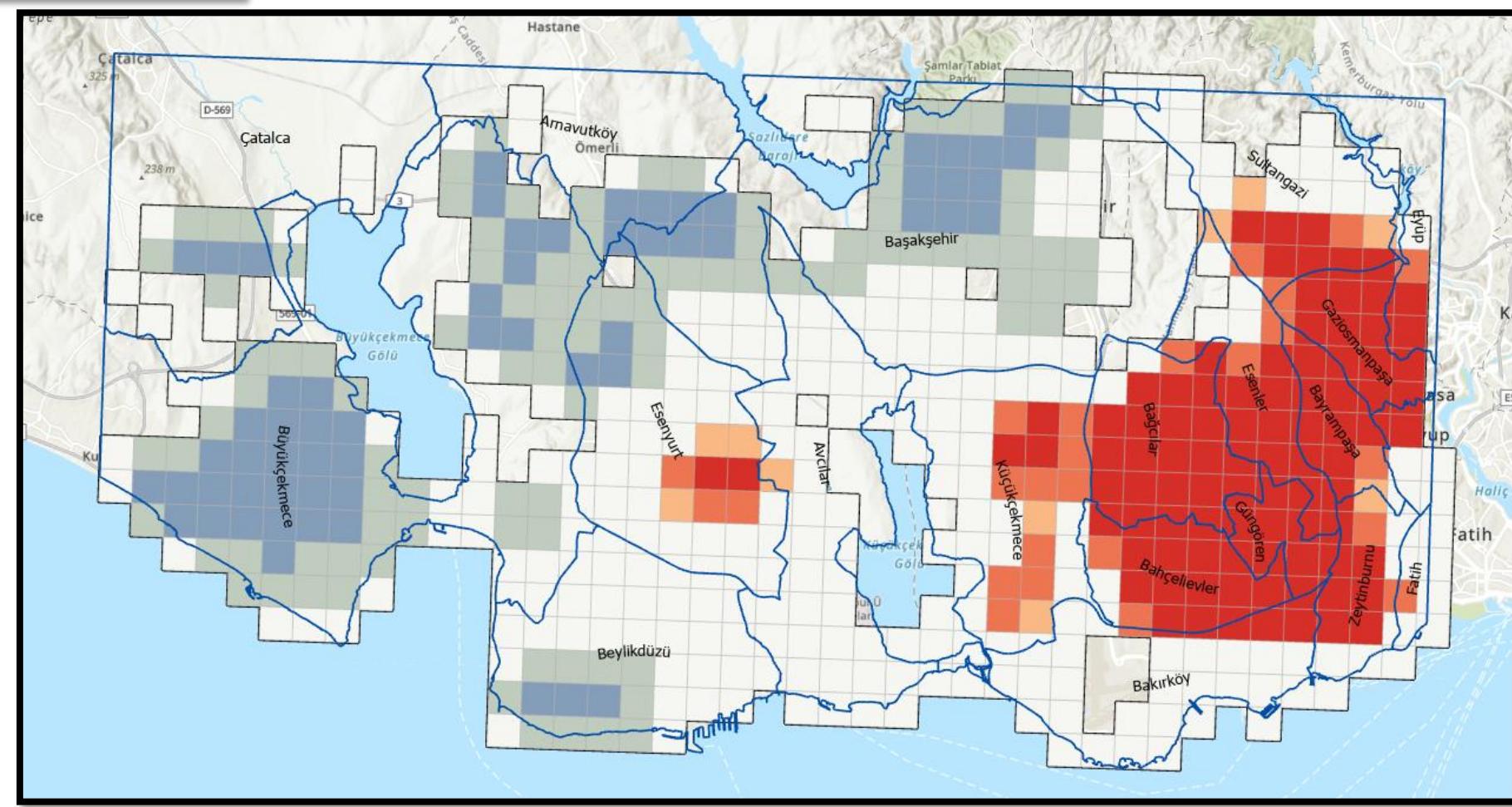
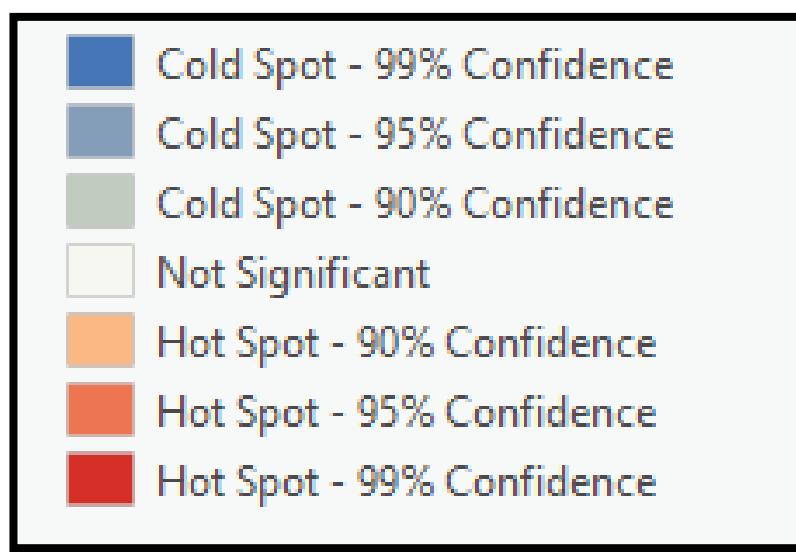
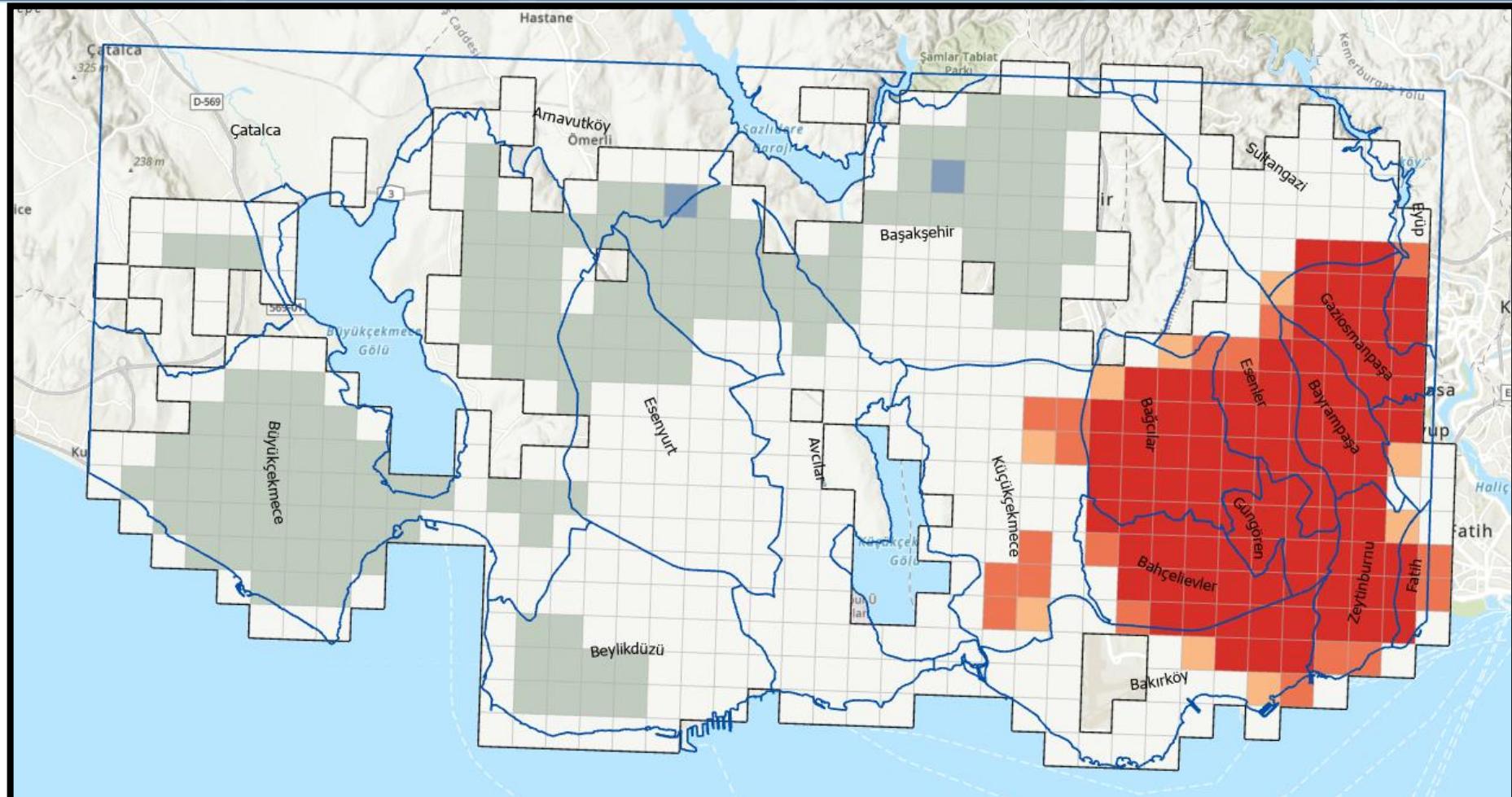
Hotspot Analysis



Since we did not fill the “Distance band or threshold distance” box, ArcGIS Pro automatically determines a default threshold distance that will ensure each feature has at least one neighbor.



Comparing the Statistical Distribution of Population for 2007 and 2014

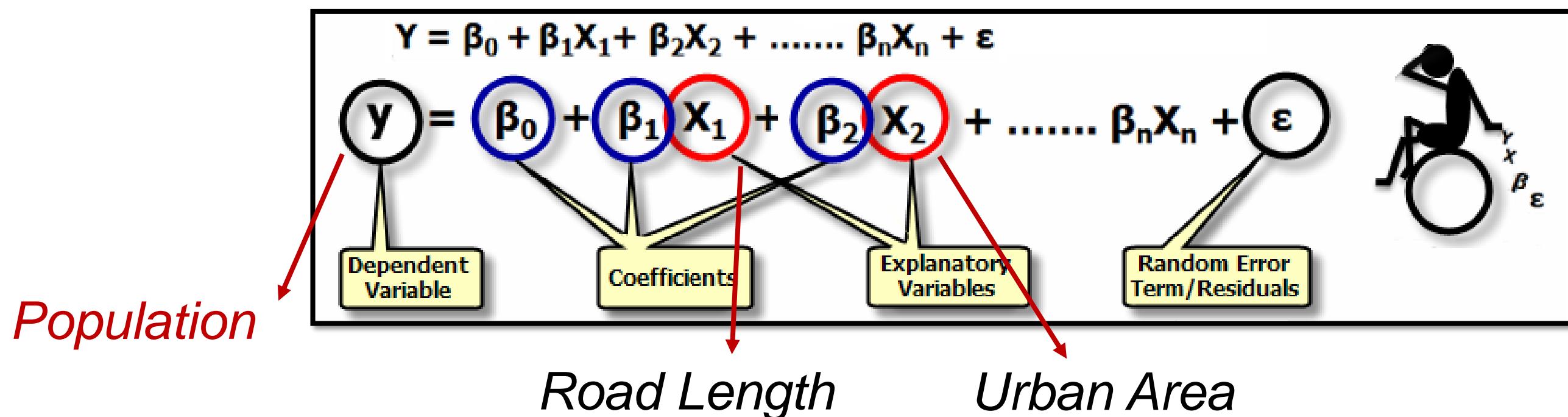


Regression Analysis

Now we are going to find the spatial/mathematical relationship between land use, transportation and population using Regression analysis

Terminology

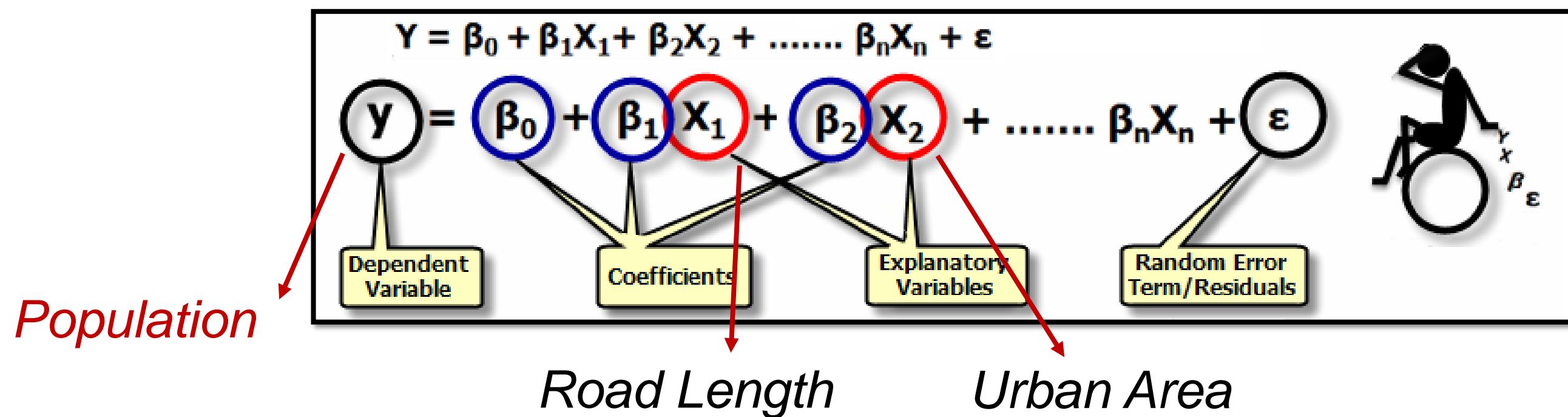
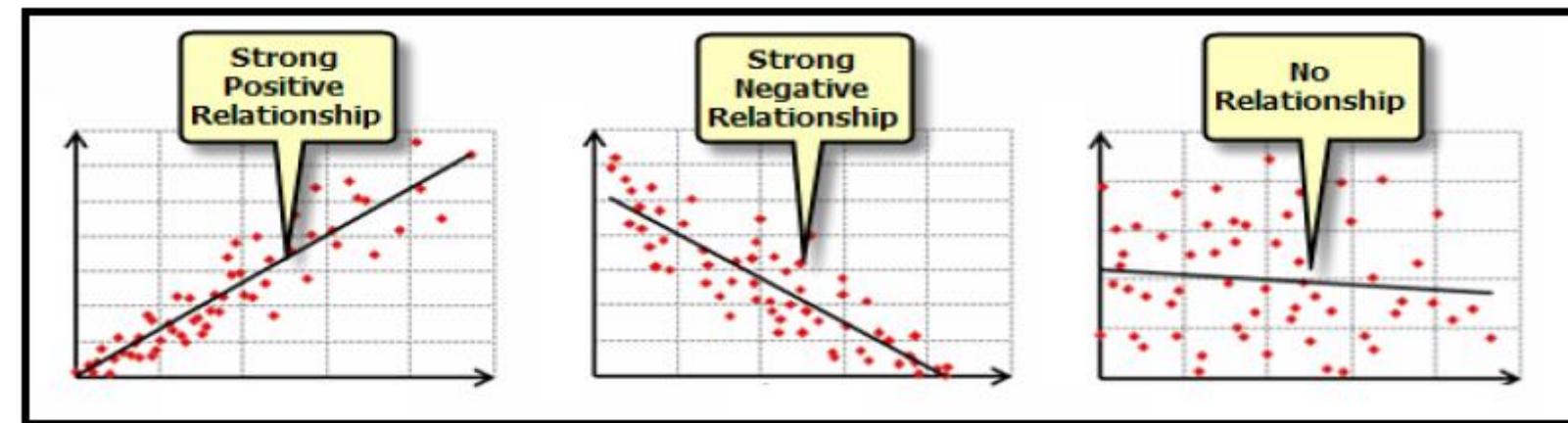
- **Dependent Variable (Y):** What are you trying to model or predict
- **Explanatory Variables (X):** Variables you believe influence or help explain the model
- **Coefficients (β):** Values computed by the regression analysis to reflect the relationship and strength of explanatory variables
- **Residuals (ϵ):** Portion of the dependent variable that isn't explained by the model



Regression Analysis

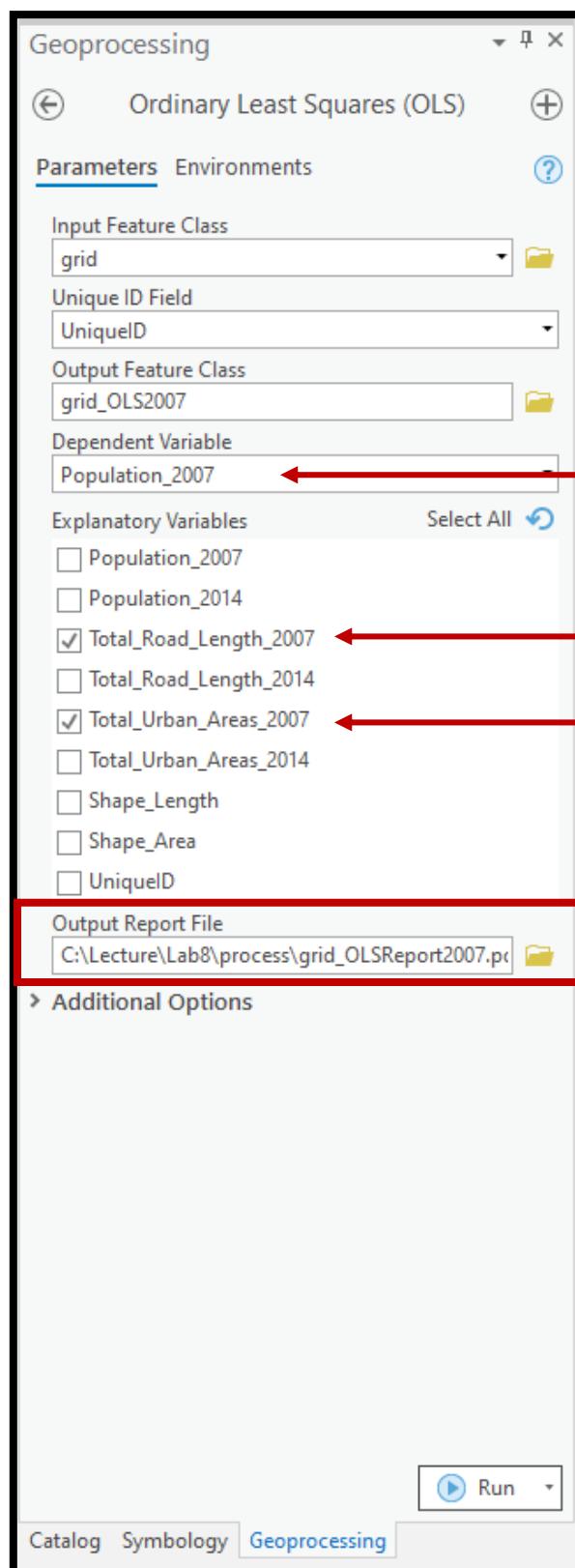
Terminology

- Sign of β coefficients tells whether the relationship is positive or negative.

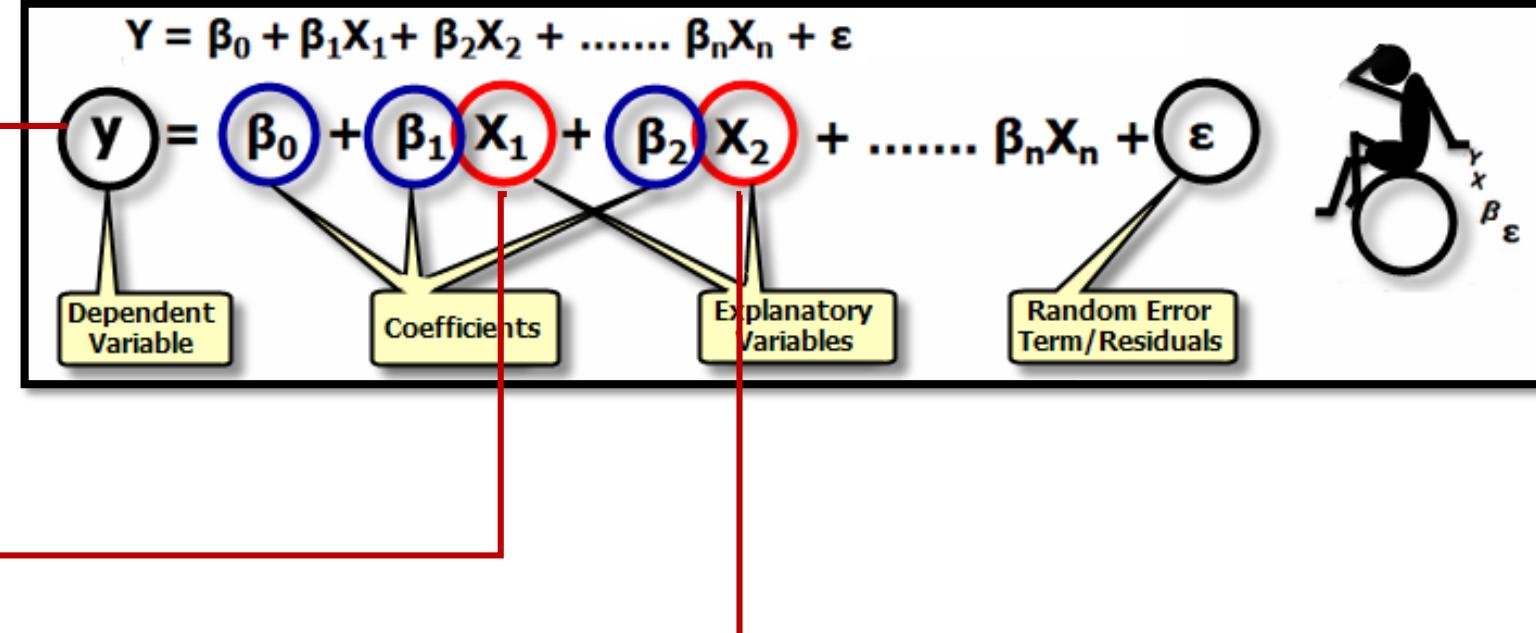


Ordinary Least Squares (OLS)

- Ordinary Least Squares (OLS) is the best known of the regression techniques.
- It provides a global model of the variable or process you are trying to understand or predict by creating a single regression equation to represent that process.



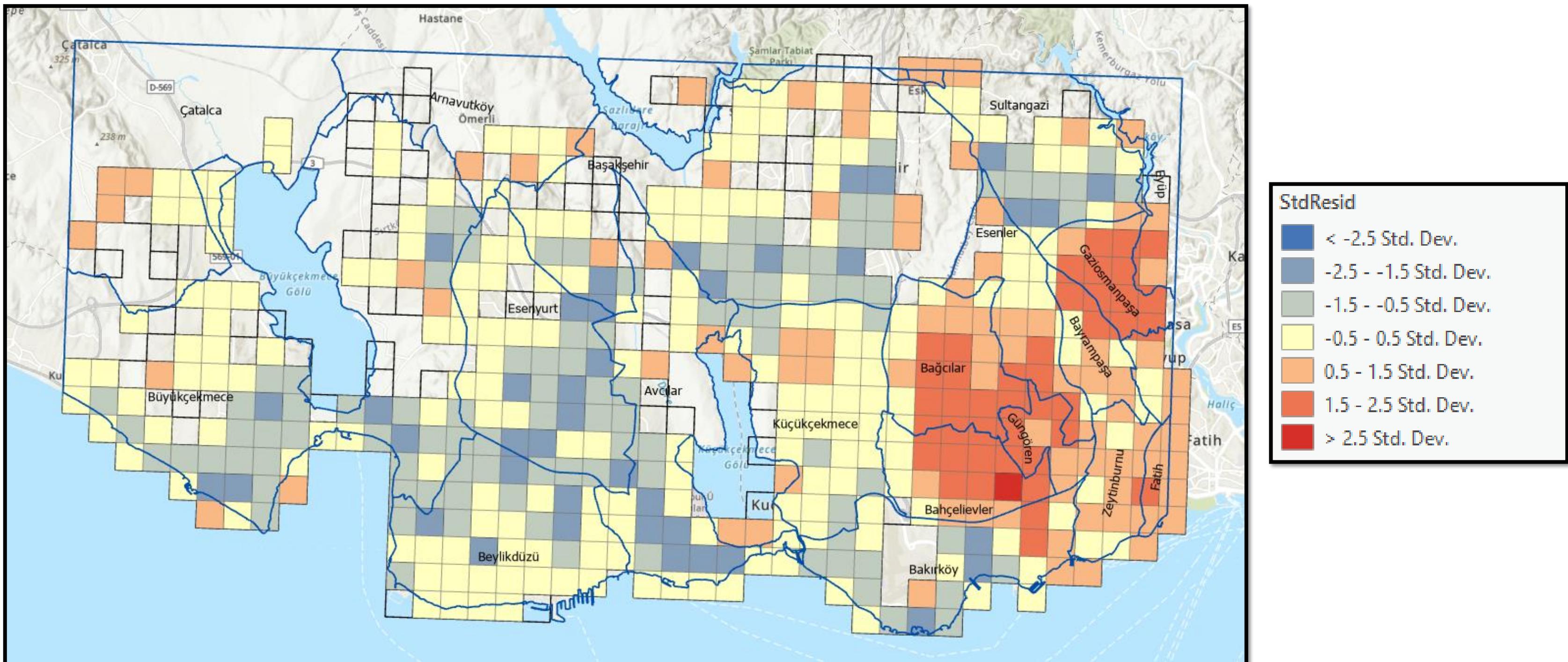
The screenshot shows the 'Geoprocessing' interface with the 'Ordinary Least Squares (OLS)' tool selected. The 'Parameters' tab is active. The 'Input Feature Class' is set to 'grid' and 'Unique ID Field' is 'UniqueID'. The 'Output Feature Class' is 'grid_OLS2007' and the 'Dependent Variable' is 'Population_2007'. In the 'Explanatory Variables' section, 'Total_Road_Length_2007' and 'Total_Urban_Areas_2007' are checked. The 'Output Report File' is set to 'C:\Lecture\Lab8\process\grid_OLSReport2007.pdf'. The 'Run' button is at the bottom.



The diagram illustrates the OLS equation $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \epsilon$. It identifies the components: Y as the Dependent Variable, $\beta_0, \beta_1, \beta_2, \dots, \beta_n$ as Coefficients, X_1, X_2, \dots, X_n as Explanatory Variables, and ϵ as the Random Error Term/Residuals. A red box highlights the text 'Variable that will be explained' pointing to Y .

We should check the statistical report of the analysis from the report file that we created

Exploring OLS Results



Interpreting the OLS Results

Check the Signs of β coefficients

Variable	Coefficient [a]	StdError	t-Statistic	Probability [b]	Robust_SE	Robust_t	Robust_Pr [b]	VIF [c]
Intercept	-3332.791479	407.626521	-8.176091	0.000000*	279.008038	-11.945145	0.000000*	-----
SUM_SHAPE_LE	0.311387	0.038190	8.153693	0.000000*	0.038293	8.131691	0.000000*	2.948502
SUM_SHAPE_AR	0.013119	0.001248	10.510569	0.000000*	0.001388	9.451260	0.000000*	2.948502

Should be less than 7.5. If no, two variables are telling the same story

Model Performance

OLS Diagnostics		
Input Features:	grid	Dependent Variable:
Number of Observations:	474	Akaike's Information Criterion (AICc) [d]:
Multiple R-Squared [d]:	0.664430	Adjusted R-Squared [d]:
Joint F-Statistic [e]:	466.290457	Prob(>F), (2,471) degrees of freedom:
Joint Wald Statistic [e]:	622.829490	Prob(>chi-squared), (2) degrees of freedom:
Koenker (BP) Statistic [f]:	240.977030	Prob(>chi-squared), (2) degrees of freedom:
Jarque-Bera Statistic [g]:	1.166495	Prob(>chi-squared), (2) degrees of freedom:
		0.558083

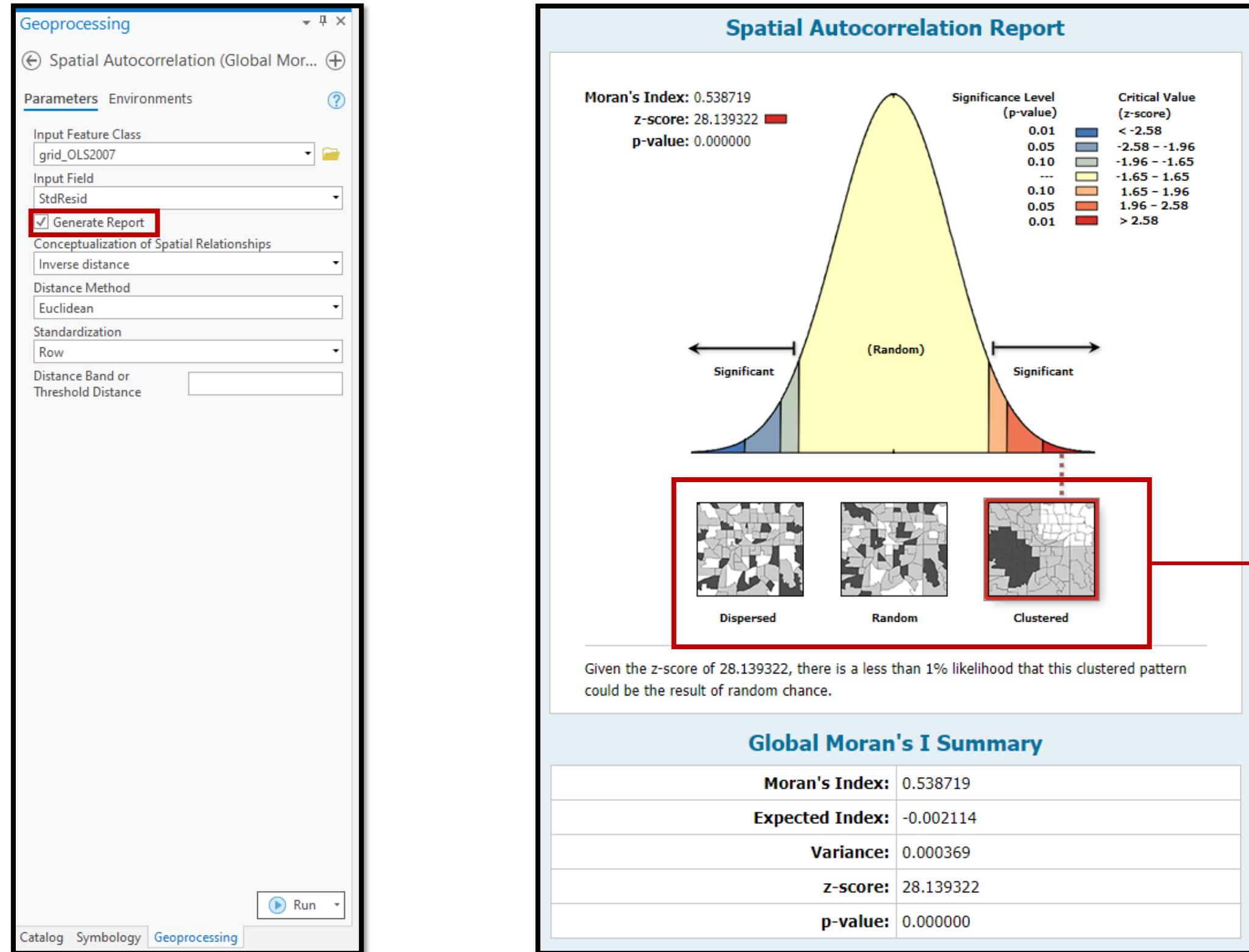
*Should have * sign. Means that the variables are significant.*

- After OLS, always run the Spatial Autocorrelation (Moran's I) tool on the regression residuals to ensure that they are spatially random.
- Statistically significant clustering of high and low residuals (model under- and overpredictions) indicates a key variable is missing from the model (misspecification). OLS results cannot be trusted when the model is misspecified

*It shouldn't have * sign. It means model is biased*

Spatial Autocorrelation

Measures spatial autocorrelation based on feature locations and attribute values using the Global Moran's I statistic.

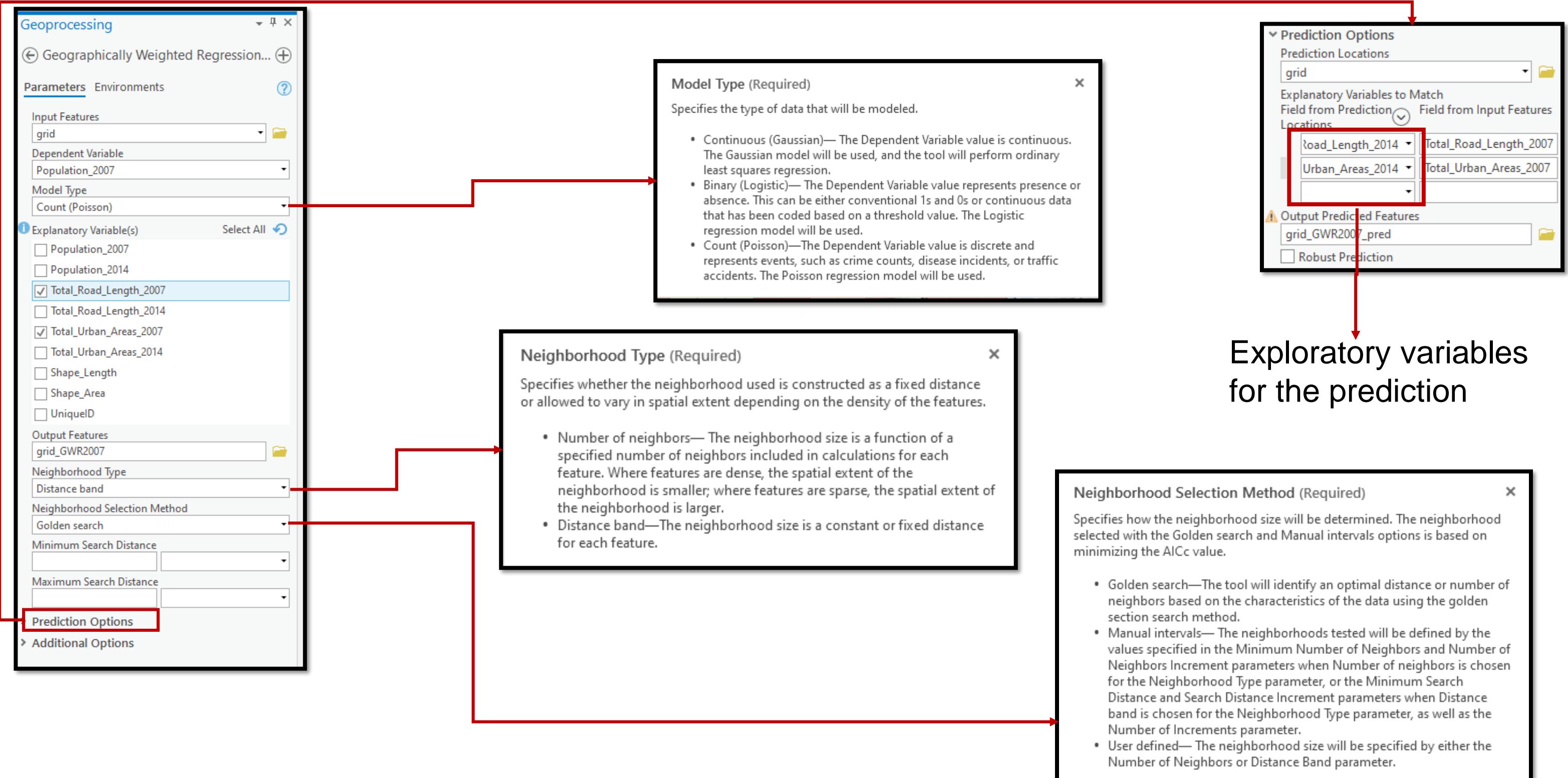


When there is spatial clustering of the under/overpredictions coming out of the model, it introduces an overcounting type of bias and renders the model unreliable.

*To get more information about spatial autocorrelation please visit:
<https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/h-how-spatial-autocorrelation-moran-s-i-spatial-st.htm>

Geographically Weighted Regression (GWR)

Geographically Weighted Regression (GWR) is the local form of regression used to model spatially varying relationships.



Geoprocessing

Geographically Weighted Regression...

Parameters

Input Features: grid
Dependent Variable: Population_2007
Model Type: Count (Poisson)

Explanatory Variable(s): Select All

- Population_2007
- Population_2014
- Total_Road_Length_2007
- Total_Road_Length_2014
- Total_Urban_Areas_2007
- Total_Urban_Areas_2014
- Shape_Length
- Shape_Area
- UniqueID

Output Features: grid_GWR2007

Neighborhood Type: Distance band

Neighborhood Selection Method: Golden search

Minimum Search Distance: []

Maximum Search Distance: []

Prediction Options

Model Type (Required)

Specifies the type of data that will be modeled.

- Continuous (Gaussian)—The Dependent Variable value is continuous. The Gaussian model will be used, and the tool will perform ordinary least squares regression.
- Binary (Logistic)—The Dependent Variable value represents presence or absence. This can be either conventional 1s and 0s or continuous data that has been coded based on a threshold value. The Logistic regression model will be used.
- Count (Poisson)—The Dependent Variable value is discrete and represents events, such as crime counts, disease incidents, or traffic accidents. The Poisson regression model will be used.

Neighborhood Type (Required)

Specifies whether the neighborhood used is constructed as a fixed distance or allowed to vary in spatial extent depending on the density of the features.

- Number of neighbors—The neighborhood size is a function of a specified number of neighbors included in calculations for each feature. Where features are dense, the spatial extent of the neighborhood is smaller; where features are sparse, the spatial extent of the neighborhood is larger.
- Distance band—The neighborhood size is a constant or fixed distance for each feature.

Neighborhood Selection Method (Required)

Specifies how the neighborhood size will be determined. The neighborhood selected with the Golden search and Manual intervals options is based on minimizing the AICc value.

- Golden search—The tool will identify an optimal distance or number of neighbors based on the characteristics of the data using the golden section search method.
- Manual intervals—The neighborhoods tested will be defined by the values specified in the Minimum Number of Neighbors and Number of Neighbors Increment parameters when Number of neighbors is chosen for the Neighborhood Type parameter, or the Minimum Search Distance and Search Distance Increment parameters when Distance band is chosen for the Neighborhood Type parameter, as well as the Number of Increments parameter.
- User defined—The neighborhood size will be specified by either the Number of Neighbors or Distance Band parameter.

Prediction Options

Prediction Locations: grid

Explanatory Variables to Match: Field from Prediction
Field from Input Features: Locations

Total_Road_Length_2014	Total_Road_Length_2007
Urban_Areas_2014	Total_Urban_Areas_2007

Output Predicted Features: grid_GWR2007_pred

Robust Prediction

Exploratory variables for the prediction

*To get more information about GWR please visit:
<https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/geographicallyweightedregression.htm>

Results of GWR

Geographically Weighted Regression (GWR) (Spatial Statistics Tools)

⚠ Completed with warnings.

Started: Today at 13:56:32
 Completed: Today at 13:56:38
 Elapsed Time: 6 Seconds

- Errors and warnings
- Parameters
- Environments
- ▼ Messages

8865.6941	8852.5586
8765.0923	8847.2833
8702.9170	8843.9646
8664.4906	8841.8931
8640.7417	8840.6056
8626.0641	8839.8072
8616.9929	8839.3128
8611.3865	8839.0069
8607.9216	8838.8176
8605.7802	8838.7007
8604.4567	8838.6283
-
- ⚠ **WARNING 110306:** The final model didn't have the lowest AICc encountered in the Golden Search Results.
- ⚠ **WARNING 110259:** At least one local regression had very limited variation after applying the weights. Use caution when interpreting the results.
- Analysis Details -----
 Number of Features: 474
 Dependent Variable: POP_2007
 Explanatory Variables: SUM_SHAPE_LENGTH
 SUM_SHAPE_AREA
 Distance Band (Meters): 8604.4567
-
- Model Diagnostics -----

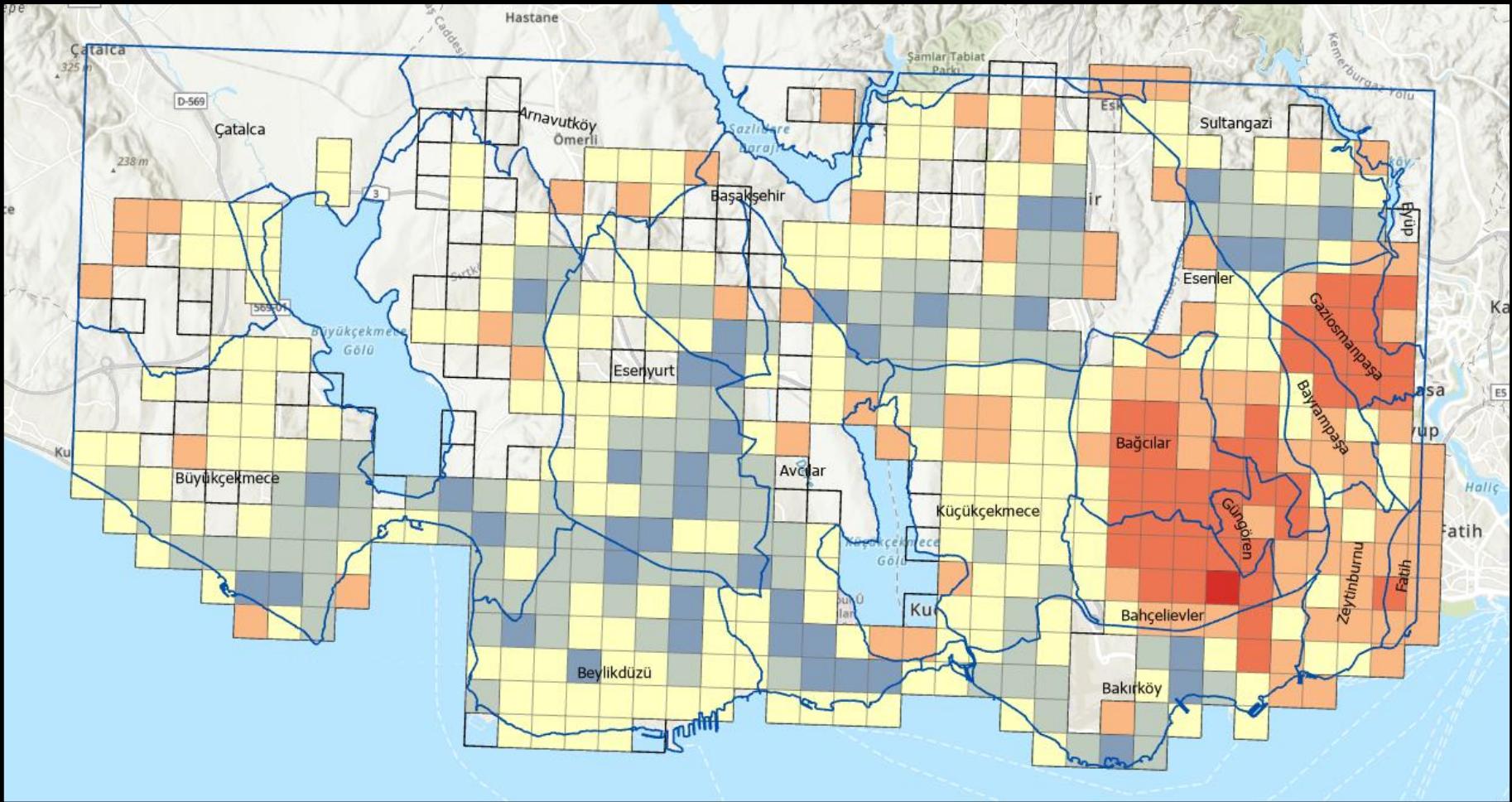
R2:	0.9188
AdjR2:	0.9116
AICc:	8838.6283
Sigma-Squared:	6978444.6636
Sigma-Squared MLE:	6409219.5428
Effective Degrees of Freedom:	435.3363
-
- ⚠ **WARNING 000642:** Problems reading 13 of 551 total records.
- ⚠ **WARNING 000848:** Features with bad records (only includes first 30): OBJECTID = 353, 354, 385, 426, 455, 498, 518, 519, 535, 542, 543, 546, 548.

Succeeded at Tuesday, May 4, 2021 13:56:37 (Elapsed Time: 5.32 seconds)

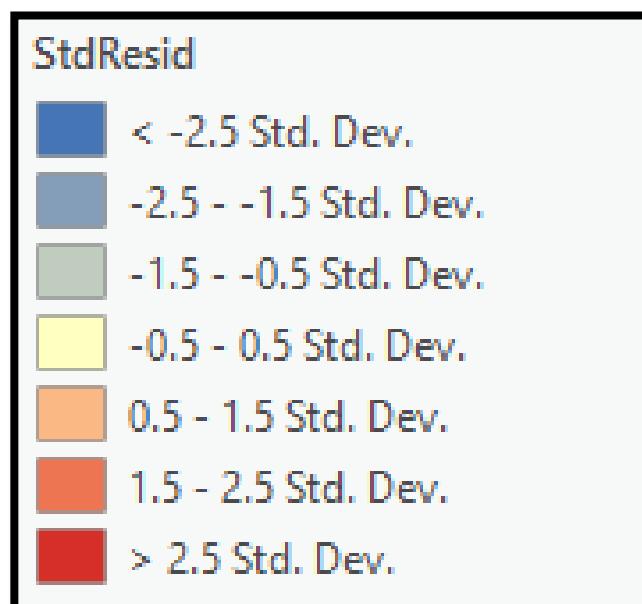
Population_2007	Predicted (POP_2007)	Residual	Std Residual
53	-3.593756	56.593756	0.022986
200	223.833516	-23.833516	-0.009329
456	481.827402	-25.827402	-0.010051
889	870.578121	18.421879	0.007277
1272	327.151358	944.848642	0.375451
4030	9354.395339	-5324.395339	-2.051056
4669	12156.993244	-7487.993244	-2.894616
3520	4973.018571	-1453.018571	-0.571428
1940	1602.8479	337.1521	0.132514
1102	893.543898	208.456102	0.080242
475	601.567749	-126.567749	-0.049151
621	829.471454	-208.471454	-0.082681
786	993.745458	-207.745458	-0.079887
543	810.513082	-267.513082	-0.103242
483	190.318741	292.681259	0.113702
359	181.170986	177.829014	0.069901
1179	1281.074563	-102.074563	-0.040871
2415	3305.41099	-890.41099	-0.345147
1785	2313.577373	-528.577373	-0.204564
593	564.795516	28.204484	0.011112
2740	3361.789573	-621.789573	-0.243573

Comparing the Results of OLS and GWR

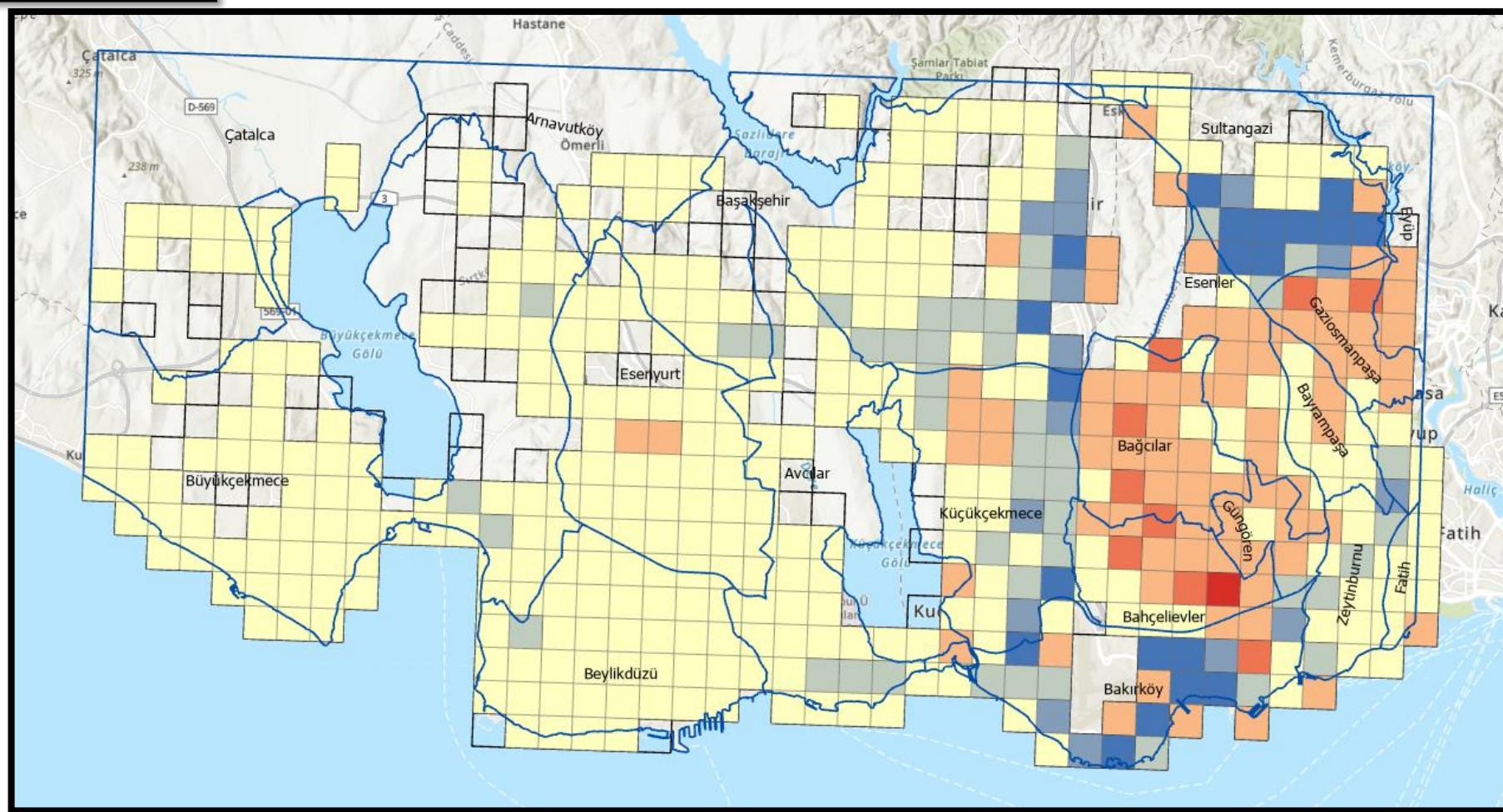
OLS



Local variations are interpreted better with GWR

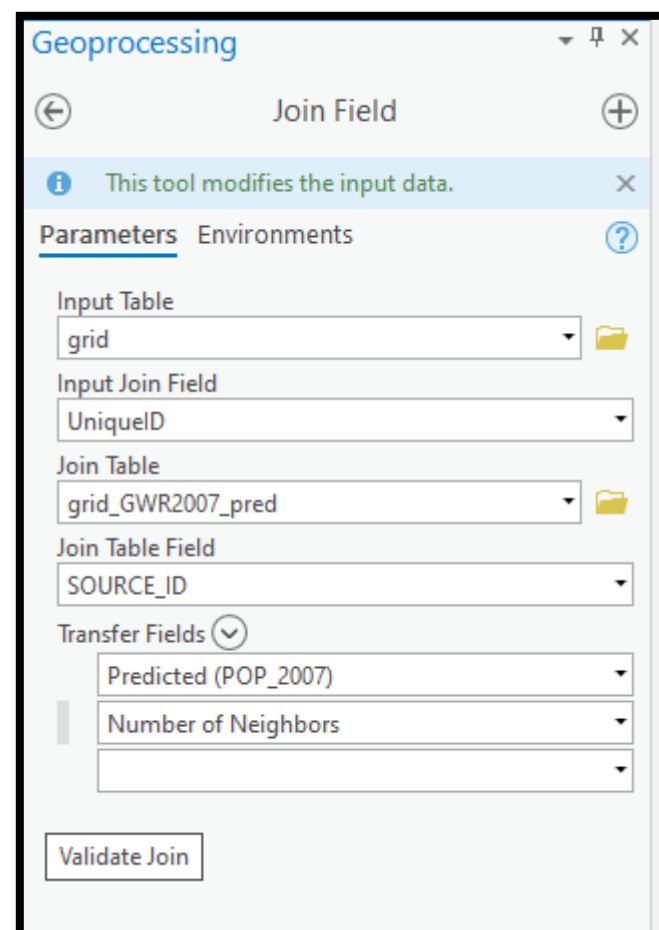


GWR



GWR Prediction

Join the GWR predictions with input grid data to visually interpret the 2014 results.



Population_2014	Predicted (POP_2007)
163	839.969139
325	1378.157783
69	347.926451
1147	908.154103
3801	11153.387124
4337	13866.854096
3024	6907.725697
1398	2486.661314
1130	1685.38405
740	1818.396898
635	1909.614668
609	1787.676183
858	879.784951
1095	687.079283
1288	414.985094
3326	1177.808986
5803	4016.621891
4571	2795.603423
2277	1437.14302
3125	3928.674038
4924	8536.807223

With GWR model, better statistical predictions were made. However, the predictions are still far from the real values. For a complex attribute such population, more exploratory variables are needed to calibrate the model.

Interpreting the Results

- There is a positive relationship between population, land use and transport network density, but this relationship cannot be explained using only linear models.
- The model created using only neighboring grids gives more accurate results in terms of spatial distribution of variables, but global comments or predictions cannot be made.
- The model was able to estimate above average density in Bağcılar, Bahçelievler, Esenler, Güngören and Gaziosmanpaşa regions, which are among the most populated areas of Istanbul.
- In the model that tries to explain the population by only using land and roads, the fact that these regions yield above-average results may be the result of vertical growth / dense urbanization in these regions.
- In the solution of such nonlinear problems; methods such as logarithmic models or artificial neural networks can be used, or analysis can be repeated by adding more independent variables to linear models.

Results & Take Home

Aim of the Study:

- *Is there any relationship with population, land use and transportation?*
- *If yes, is this relationship meaningful, linear or can it be described quantitatively?*
- *Could this relationship shown spatially?*

Output Data:

- *Hotspot analysis results for 2007 and 2014 (Vector-Polygon)*
- *OLS results for 2007 (Vector-Polygon)*
- *GWR results for 2007 (Vector-Polygon)*
- *GWR predictions for 2014 (Vector-Polygon)*

Take Home:

- *Perform the same analysis techniques with different parameters for the year 2014 and interpret the statistical and spatial results.*



Contact:

akinom@itu.edu.tr