Data Exploration and Spatial Statistics

ISTANBULTECHNICALUNIVERSITY

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?*

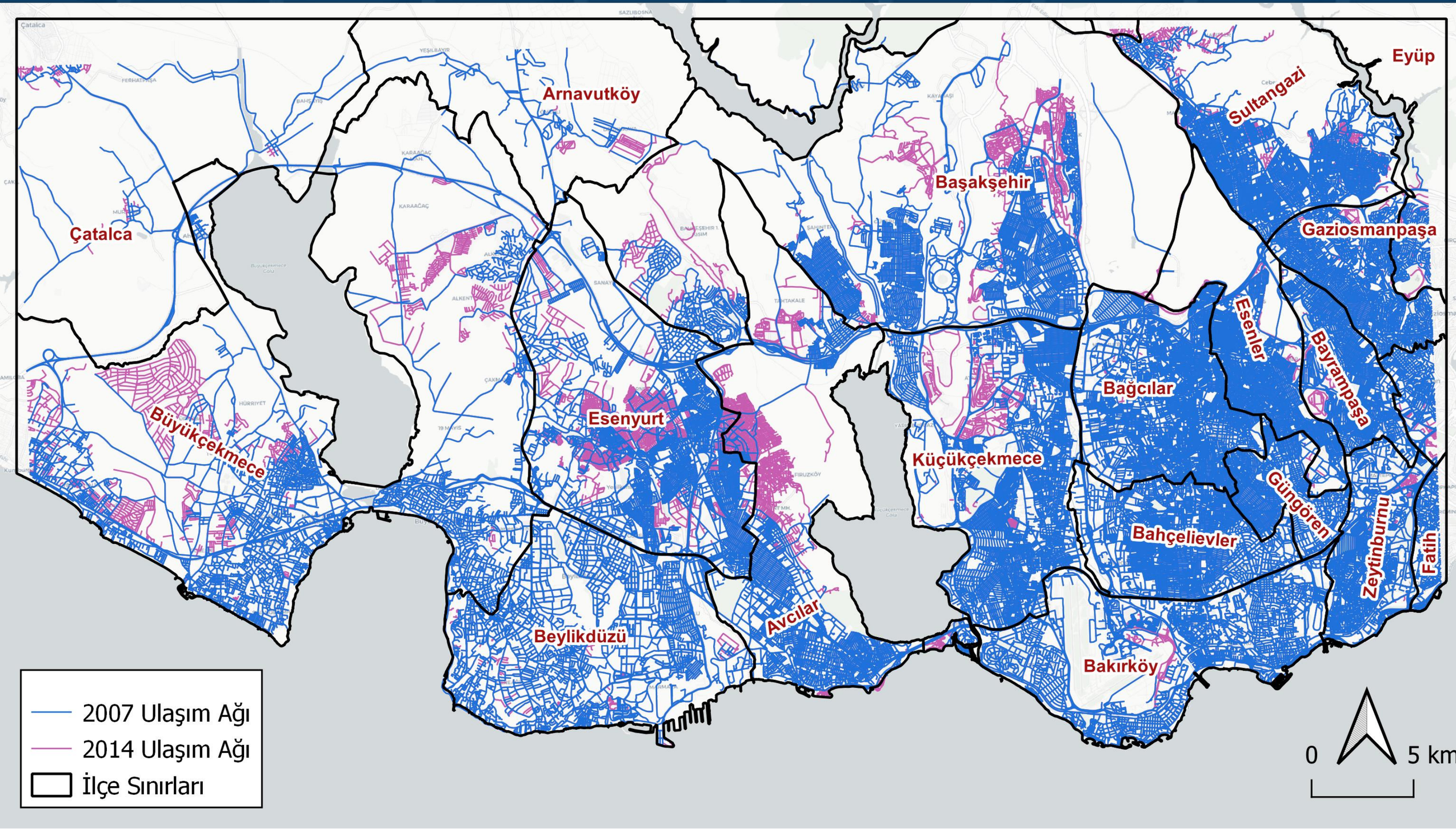
Input Data:

- *1km * 1km Grid that have (Vector-Polygon)*
  - *Population*
  - *Total Urban Areas*
  - *Total Road Length*

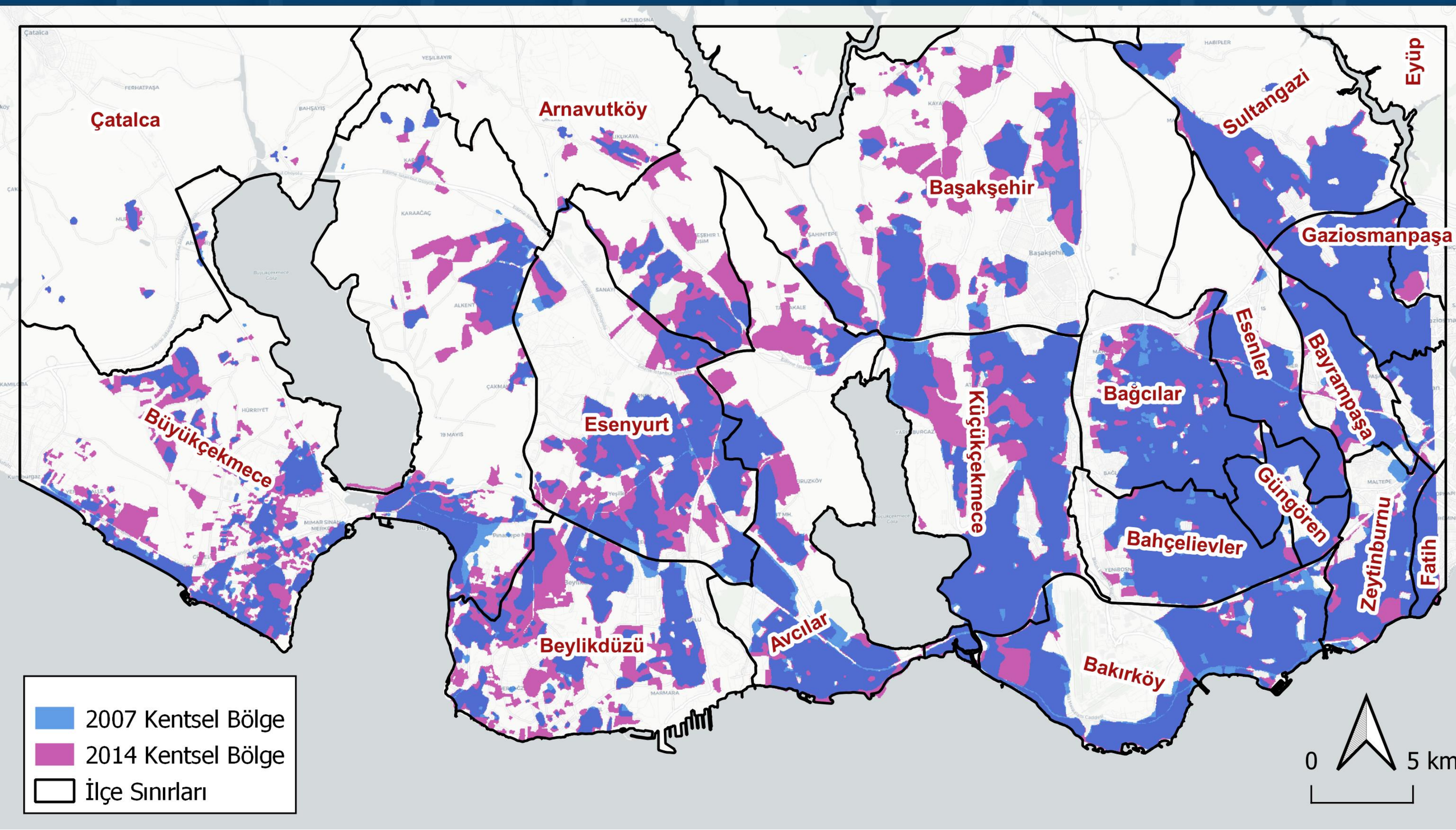
information for the years 2007 and 2014.

- *District Boundaries (Vector-Polygon)*

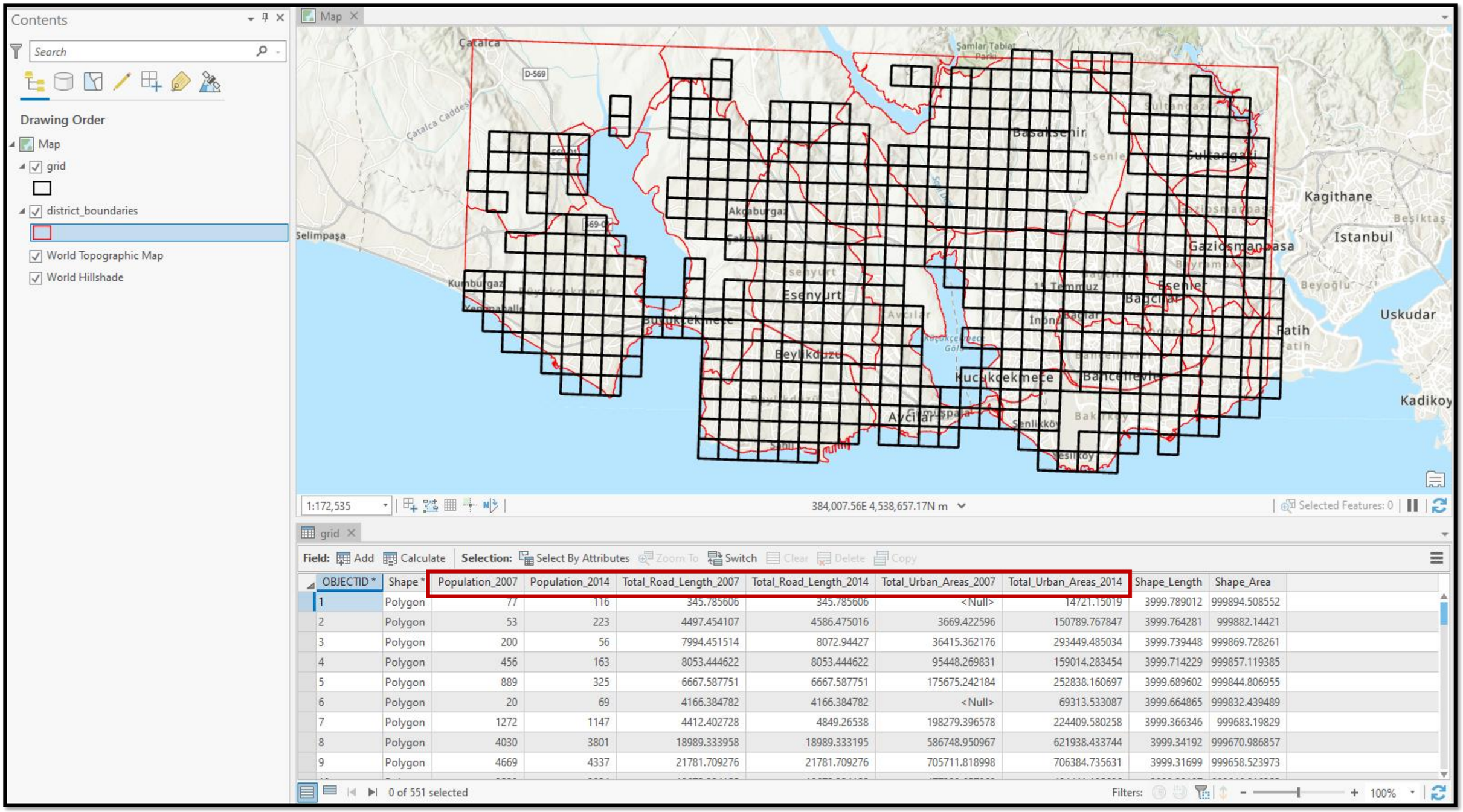
Study Area & Data



Study Area & Data



Exploring Data



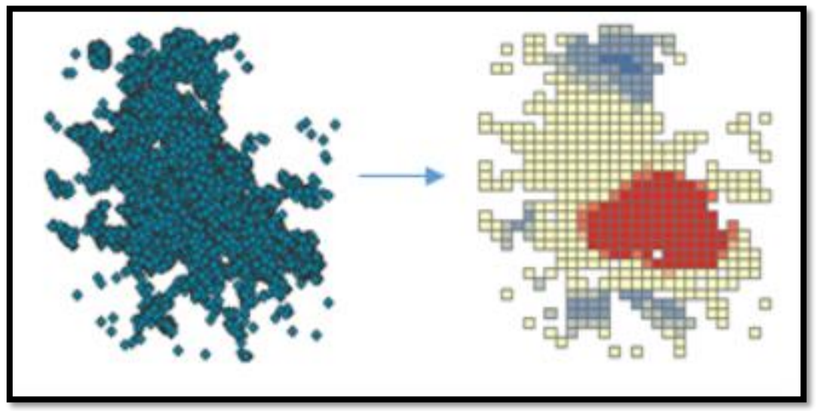
OBJECTID *	Shape *	Population_2007	Population_2014	Total_Road_Length_2007	Total_Road_Length_2014	Total_Urban_Areas_2007	Total_Urban_Areas_2014	Shape_Length	Shape_Area
1	Polygon	77	116	345.785606	345.785606	<Null>	14721.15019	3999.789012	999894.508552
2	Polygon	53	223	4497.454107	4586.475016	3669.422596	150789.767847	3999.764281	999882.14421
3	Polygon	200	56	7994.451514	8072.94427	36415.362176	293449.485034	3999.739448	999869.728261
4	Polygon	456	163	8053.444622	8053.444622	95448.269831	159014.283454	3999.714229	999857.119385
5	Polygon	889	325	6667.587751	6667.587751	175675.242184	252838.160697	3999.689602	999844.806955
6	Polygon	20	69	4166.384782	4166.384782	<Null>	69313.533087	3999.664865	999832.439489
7	Polygon	1272	1147	4412.402728	4849.26538	198279.396578	224409.580258	3999.366346	999683.19829
8	Polygon	4030	3801	18989.333958	18989.333195	586748.950967	621938.433744	3999.34192	999670.986857
9	Polygon	4669	4337	21781.709276	21781.709276	705711.818998	706384.735631	3999.31699	999658.523973

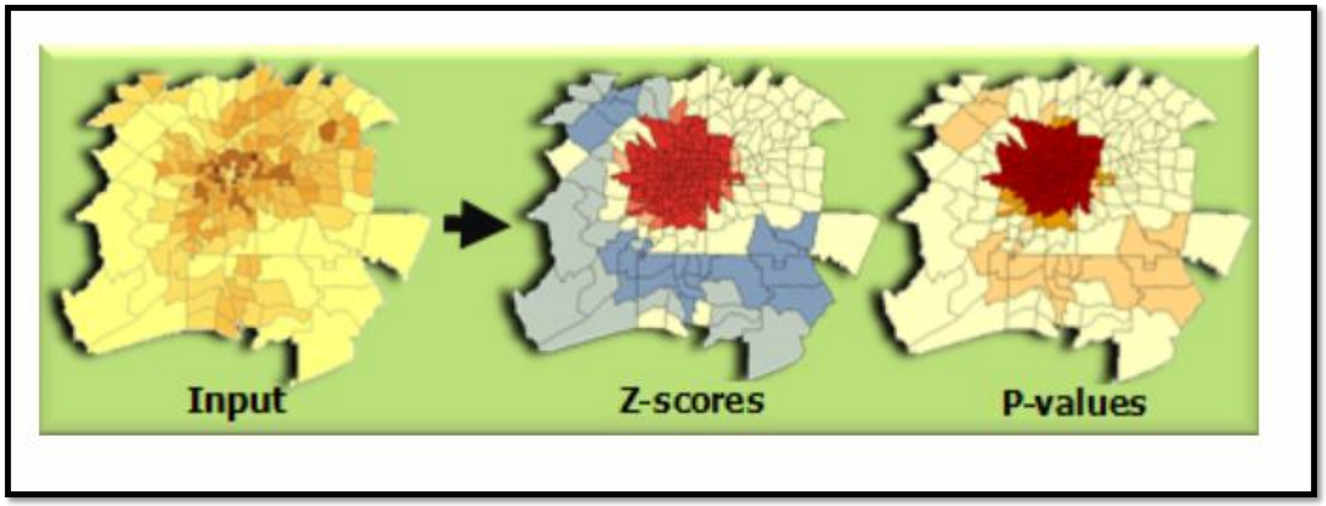
Exploratory Spatial Data Analysis
Hotspot Analysis

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

Hotspot Analysis

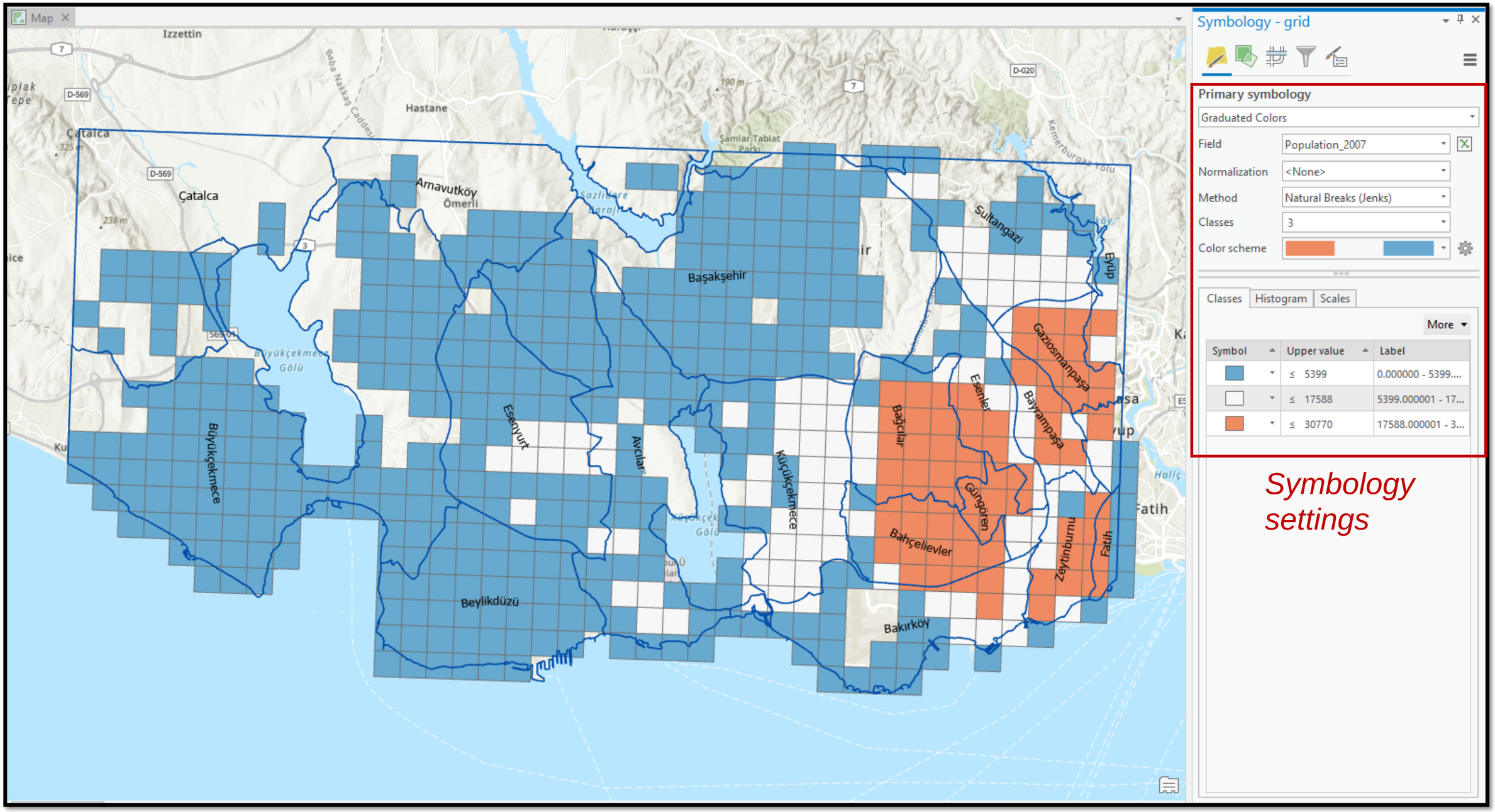
Given a set of weighted features, identifies statistically significant hot spots and cold spots using the Getis-Ord Gi* statistic





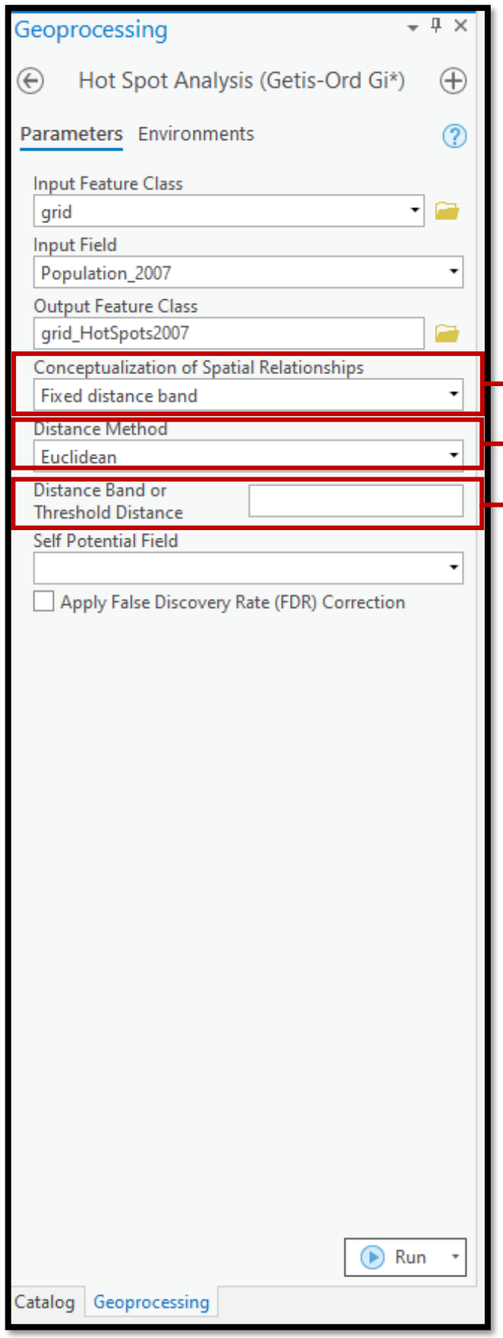
Population Information on the Study Area

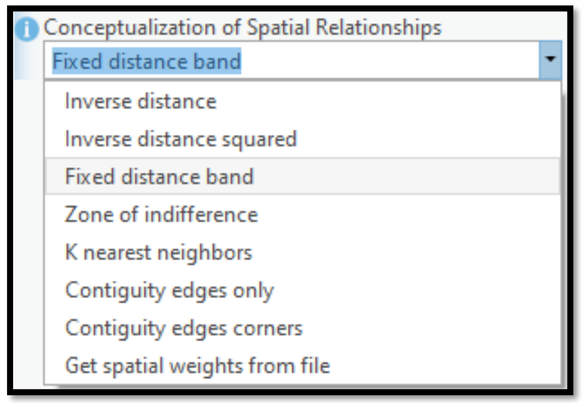
First, let’s check the population value’s distribution on the study area.



Symbology settings

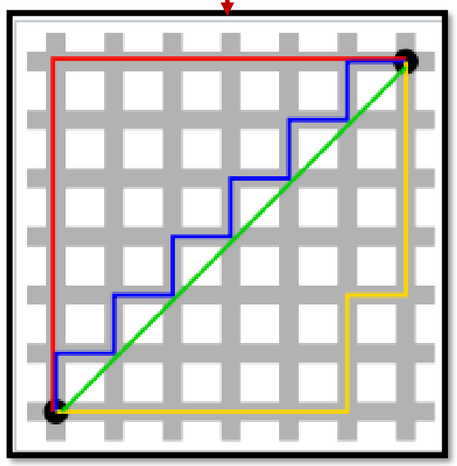
Exploratory Spatial Data Analysis
Hotspot Analysis





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

Do not fill the threshold distance for this example



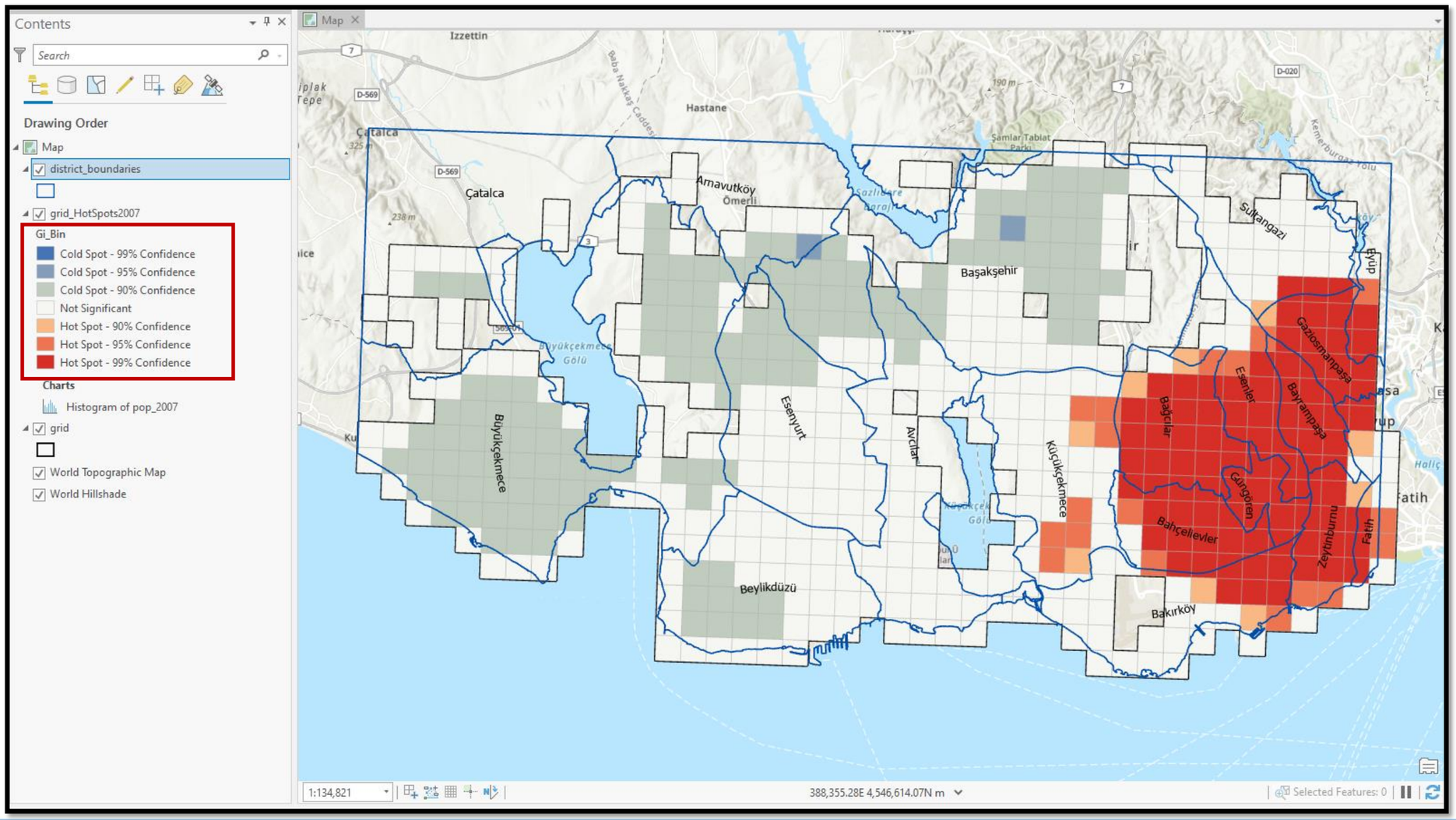
Green Road -> Euclidean Distance

Blue, Yellow and Red roads -> Manhattan Distance

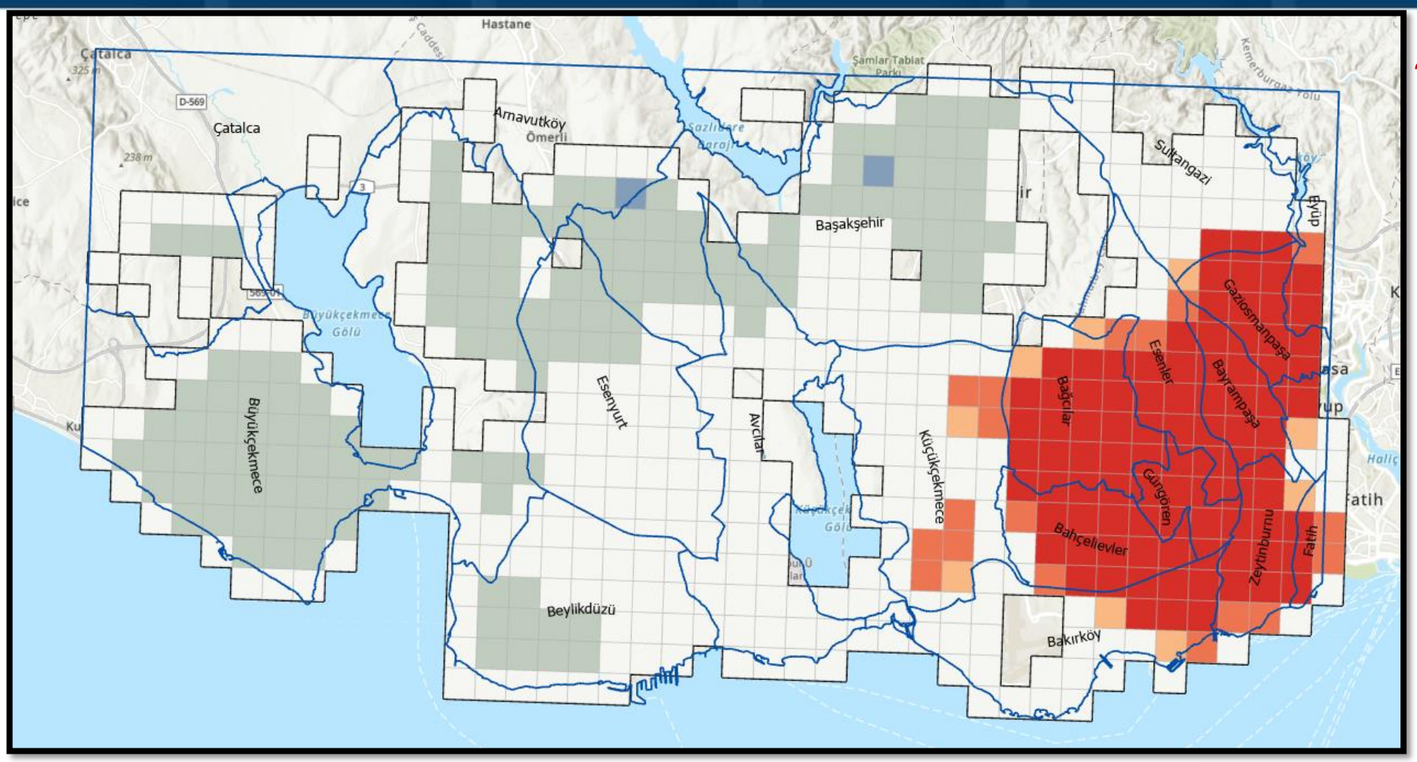
*To get more information about spatial relationship please visit:
https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/modeling-spatial-relationships.htm

Exploratory Spatial Data Analysis
Hotspot Analysis

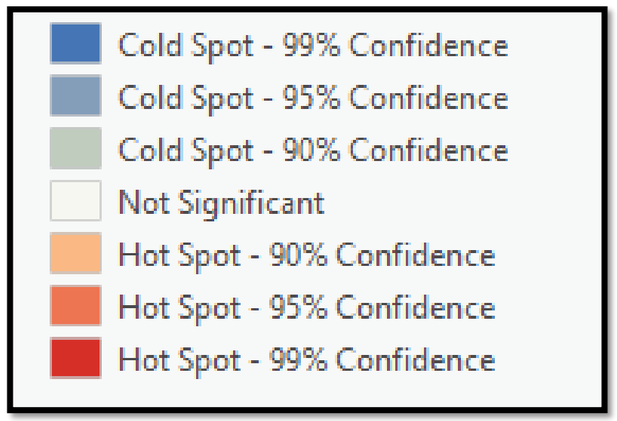
Since we did not fill the “Distance band or threshold distance” box, ArcGIS Pro automatically determines a default threshold distance that will ensure each feature has at least one neighbor.

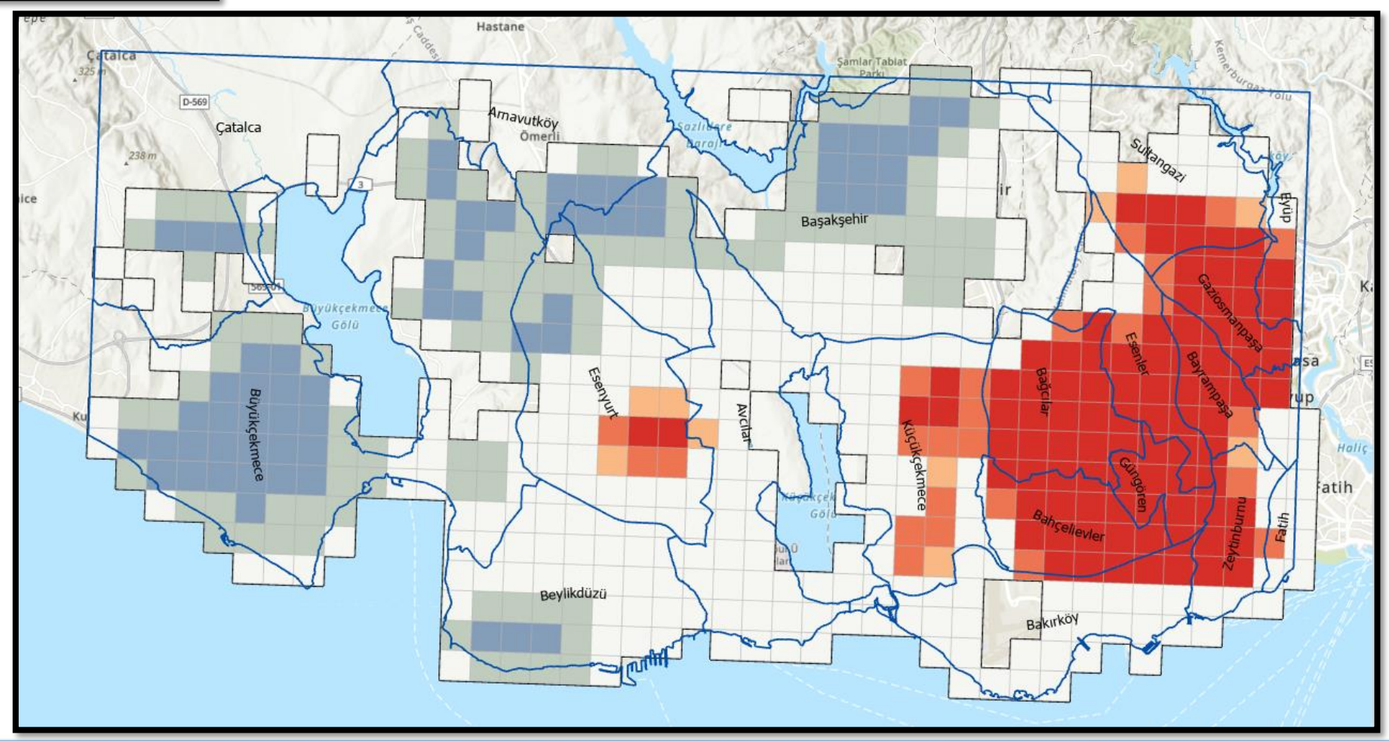


Comparing the Statistical Distribution of Population for 2007 and 2014



2007





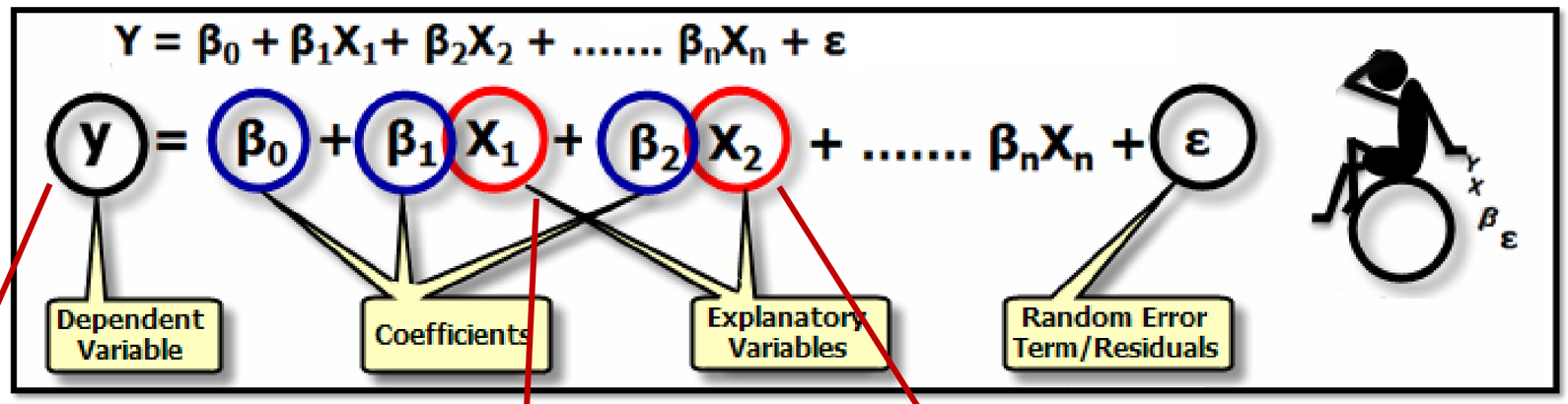
2014

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 (ß):** 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



Population

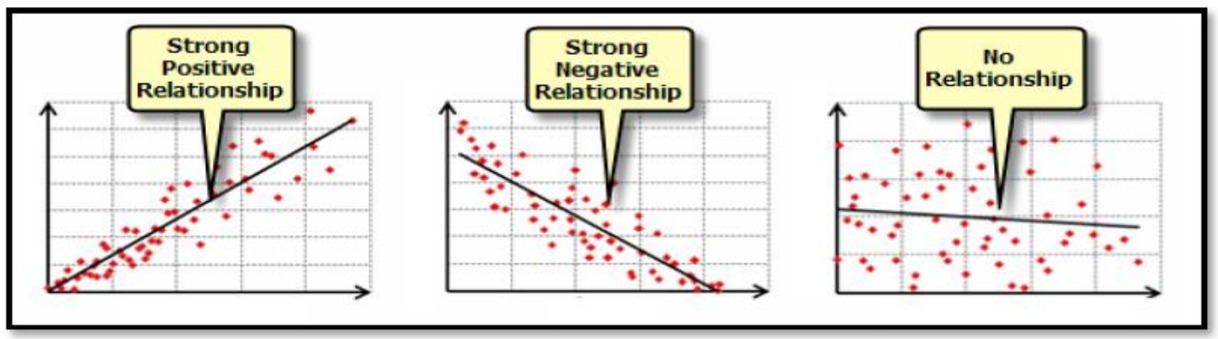
Road Length

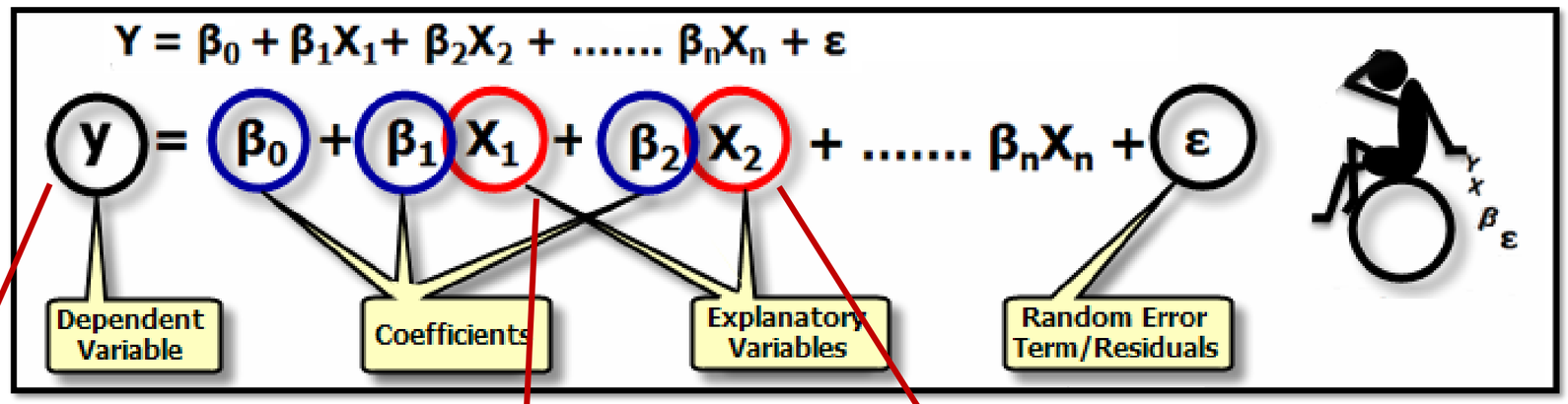
Urban Area

Regression Analysis

Terminology

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





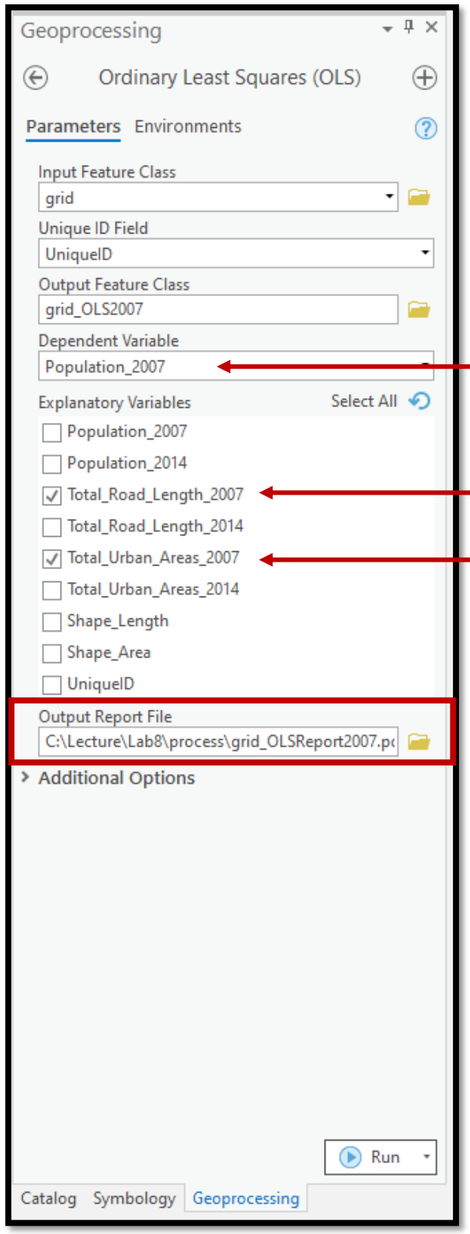
Population

Road Length

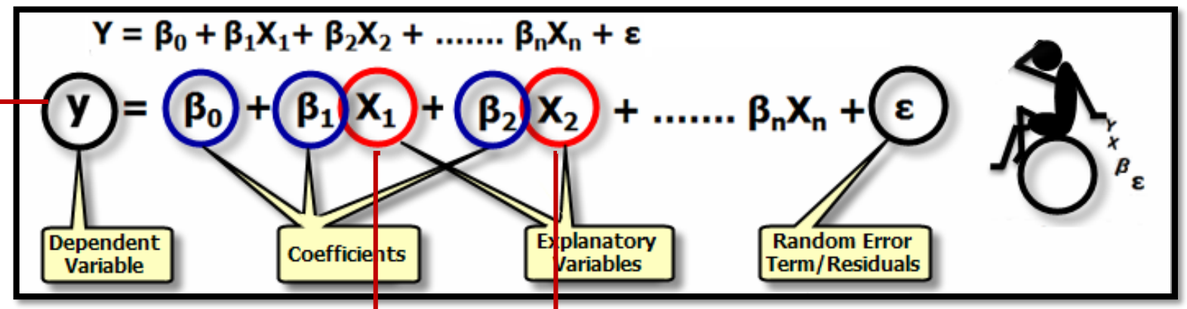
Urban Area

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.

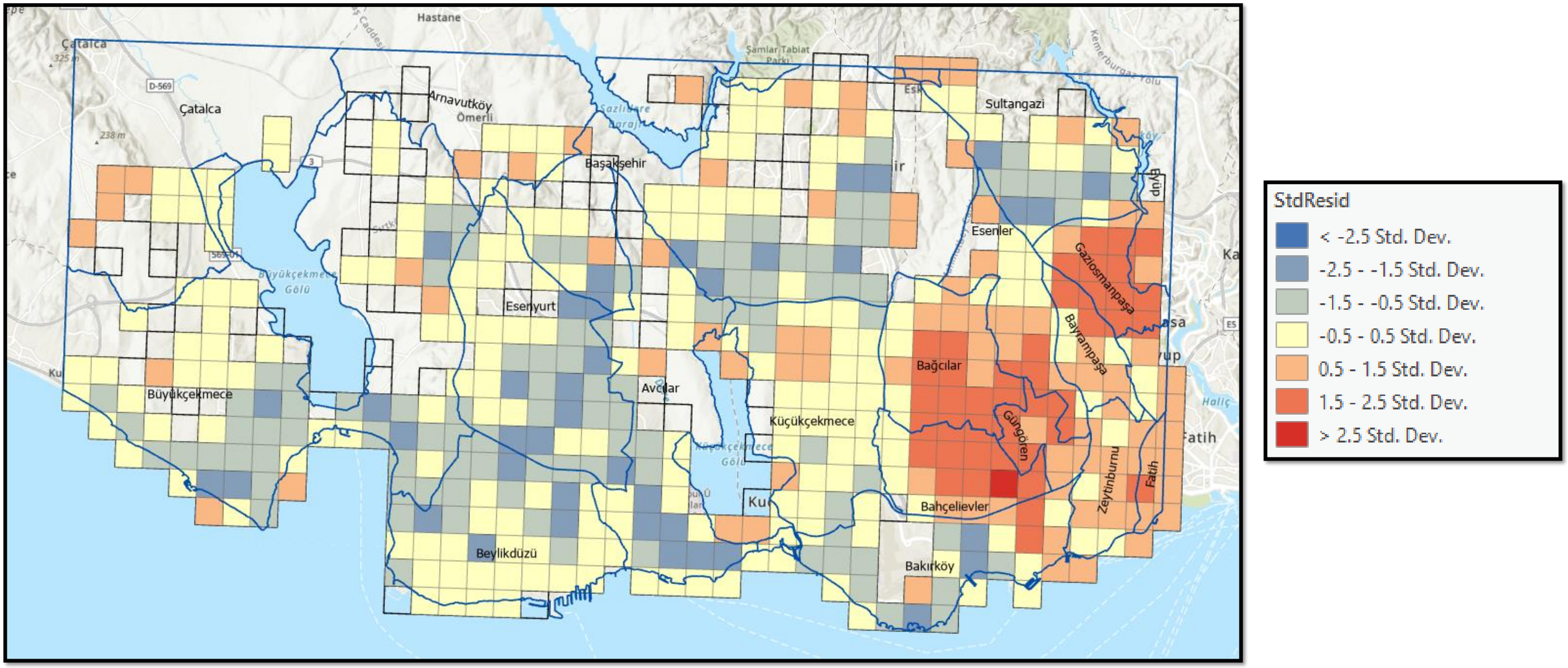


Variable that will be explained



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 ß 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

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

OLS Diagnostics

Input Features:	grid	Dependent Variable:	POP_2007
Number of Observations:	474	Akaike's Information Criterion (AICc) [d]:	9454.986320
Multiple R-Squared [d]:	0.664430	Adjusted R-Squared [d]:	0.663005
Joint F-Statistic [e]:	466.290457	Prob(>F), (2,471) degrees of freedom:	0.000000*
Joint Wald Statistic [e]:	622.829490	Prob(>chi-squared), (2) degrees of freedom:	0.000000*
Koenker (BP) Statistic [f]:	240.977030	Prob(>chi-squared), (2) degrees of freedom:	0.000000*
Jarque-Bera Statistic [g]:	1.166495	Prob(>chi-squared), (2) degrees of freedom:	0.558083

Model Performance

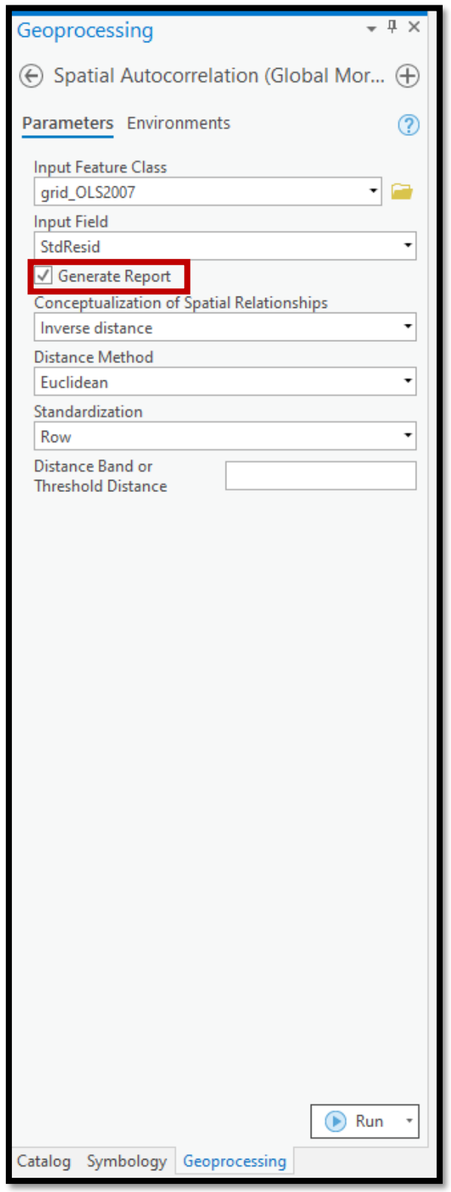
*It shouldn't have * sign. It means model is biased*

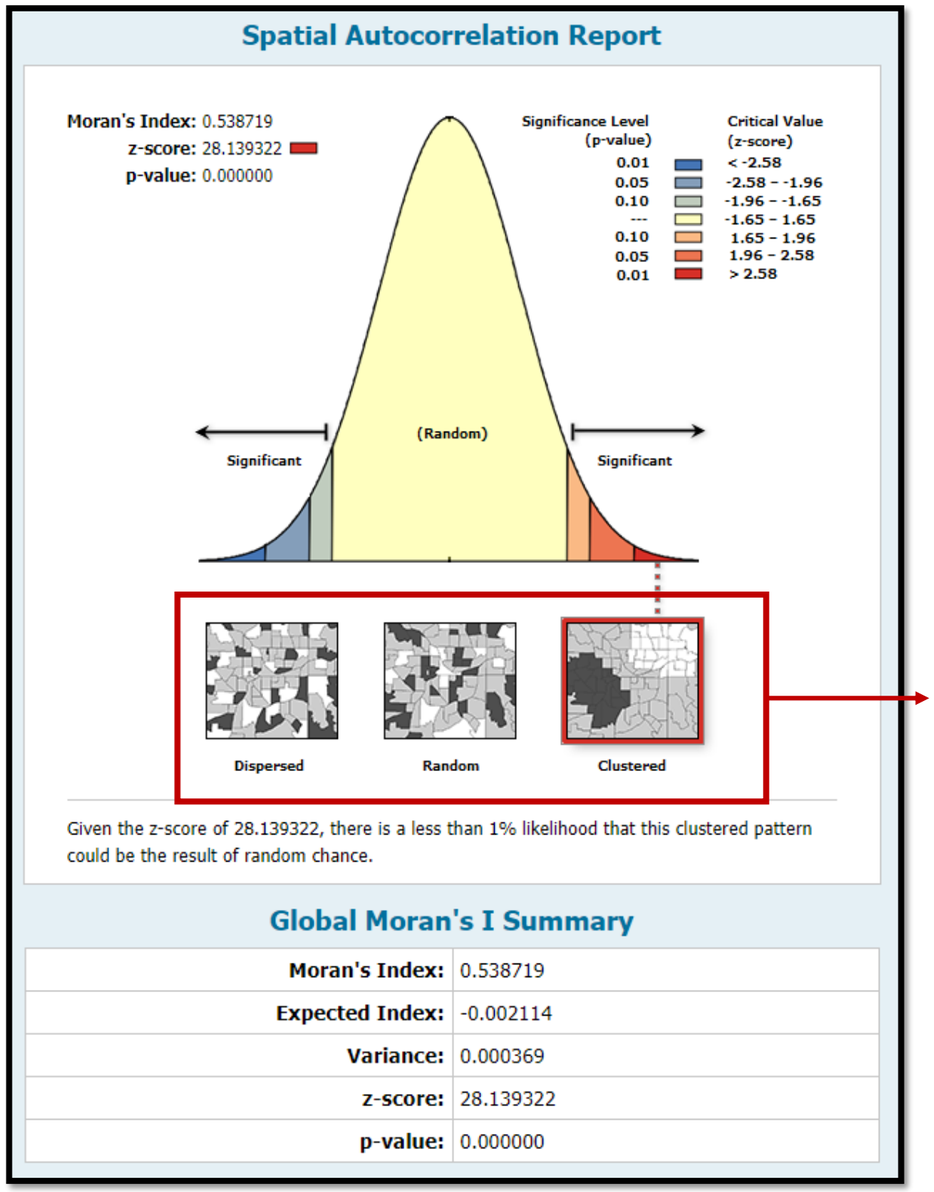
- After OLS, always run the Spatial Autocorrelation (Moran's I) tool on the regression residuals to ensure that they are spatially random.

- 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

Spatial Autocorrelation

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



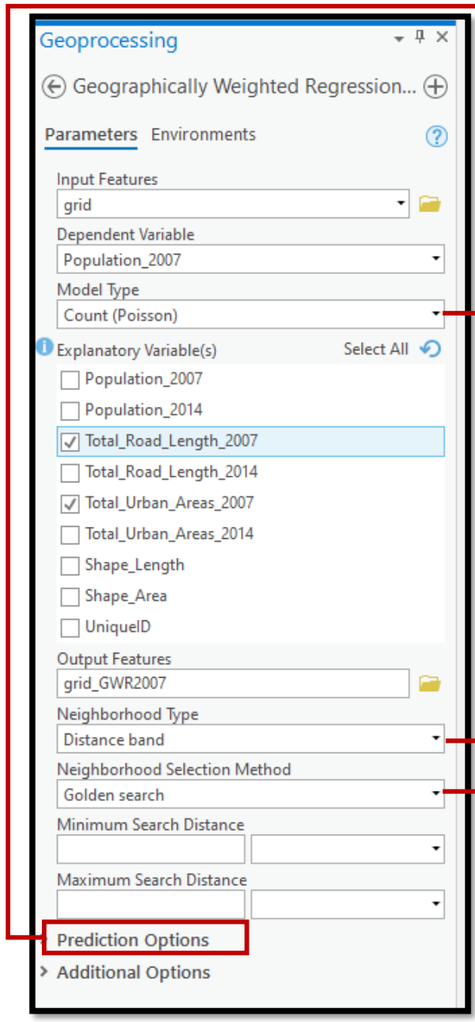


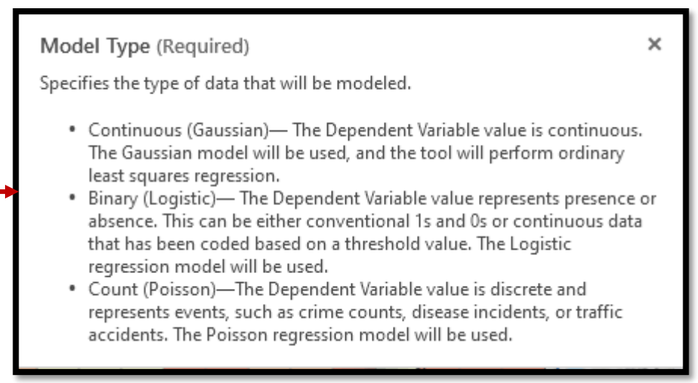
When there is spatial clustering of the under/overpredictions coming out of the model, it introduces an overcounting type of bias and renders the model unreliable.

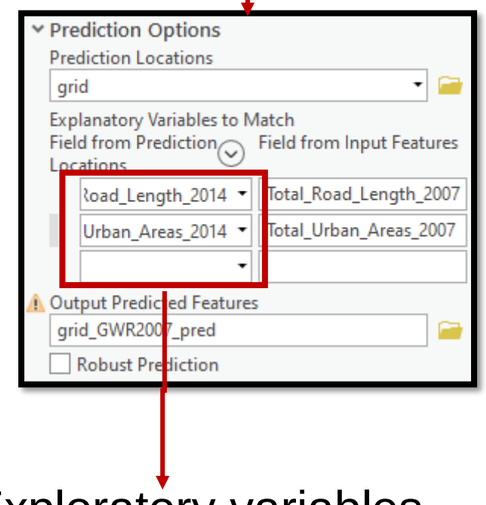
*To get more information about spatial autocorrelation please visit:
https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/h-how-spatial-autocorrelation-moran-s-i-spatial-st.htm

Geographically Weighted Regression (GWR)

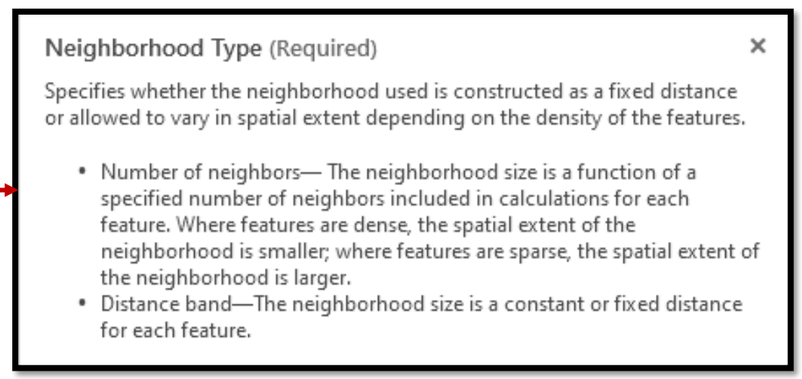
Geographically Weighted Regression (GWR) is the local form of regression used to model spatially varying relationships.

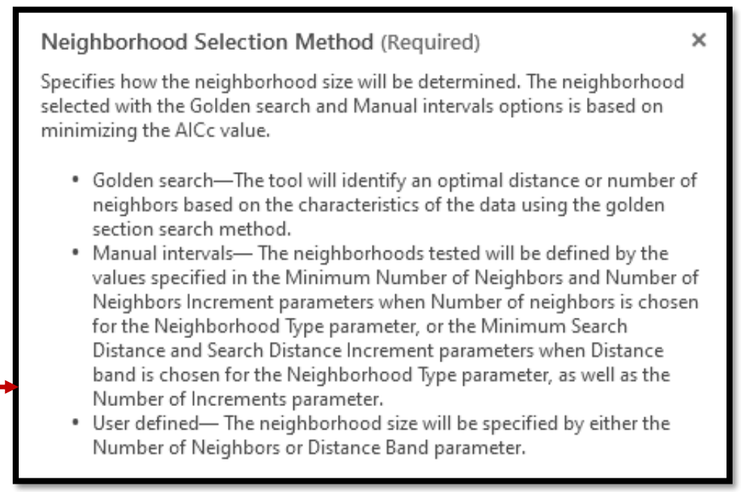






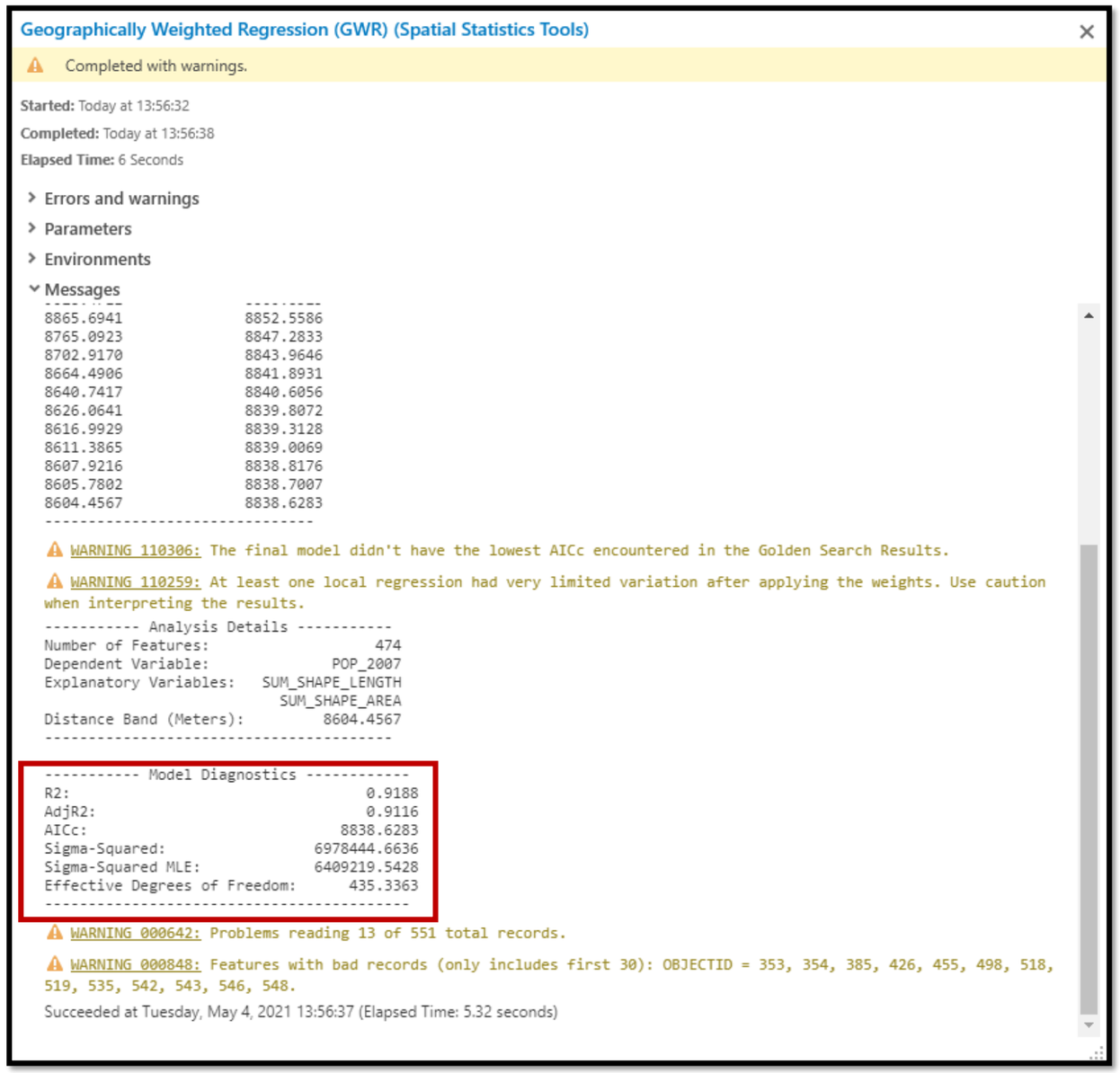
Exploratory variables for the prediction





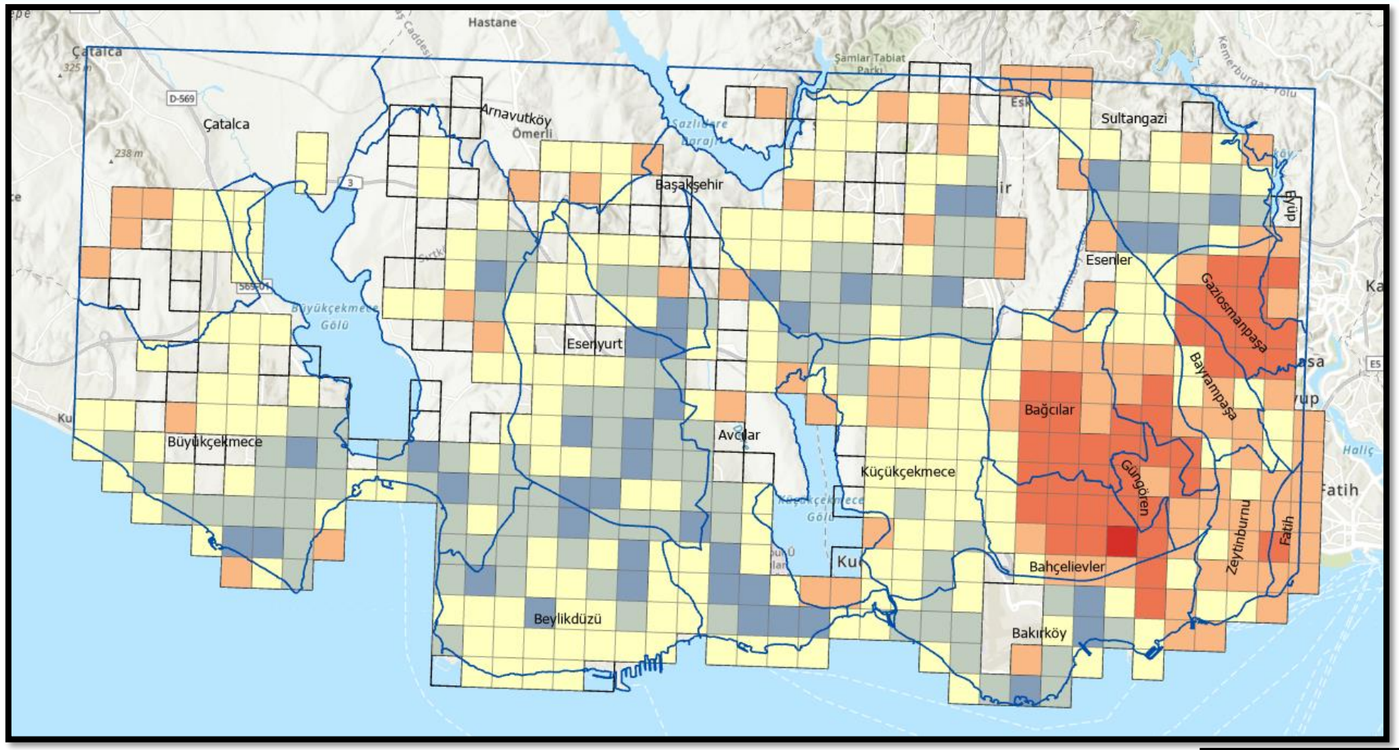
*To get more information about GWR please visit:
https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/geographicallyweightedregression.htm

Results of GWR



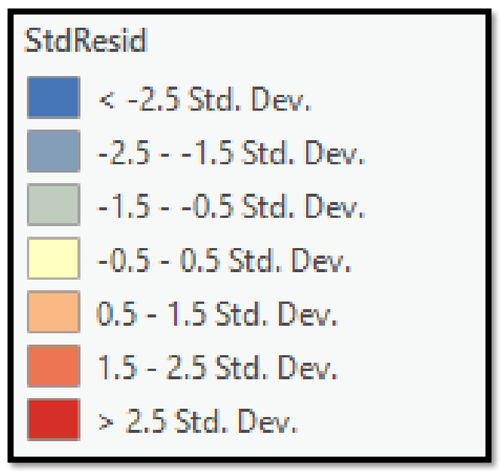
Population_2007	Predicted (POP_2007)	Residual	Std Residual
53	-3.593756	56.593756	0.022986
200	223.833516	-23.833516	-0.009329
456	481.827402	-25.827402	-0.010051
889	870.578121	18.421879	0.007277
1272	327.151358	944.848642	0.375451
4030	9354.395339	-5324.395339	-2.051056
4669	12156.993244	-7487.993244	-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

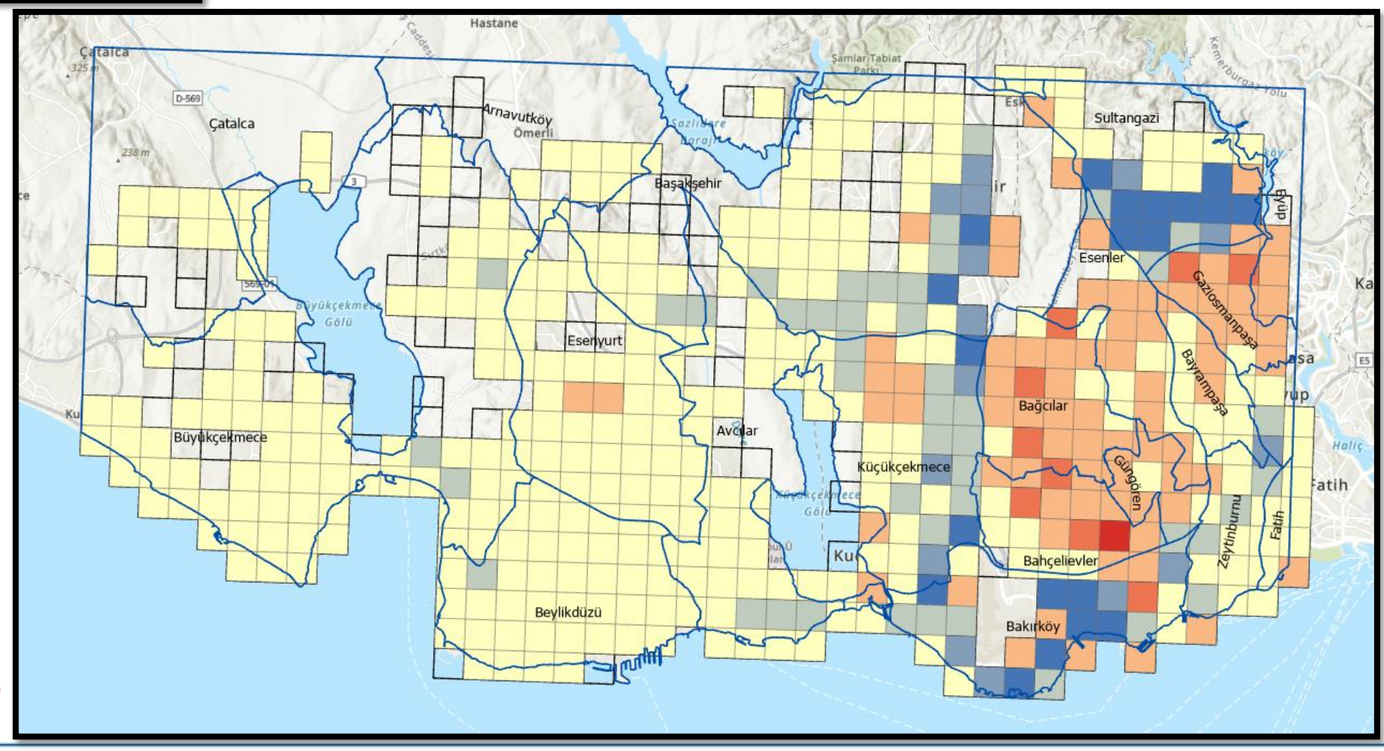
Comparing the Results of OLS and GWR



OLS

Local variations are interpreted better with GWR

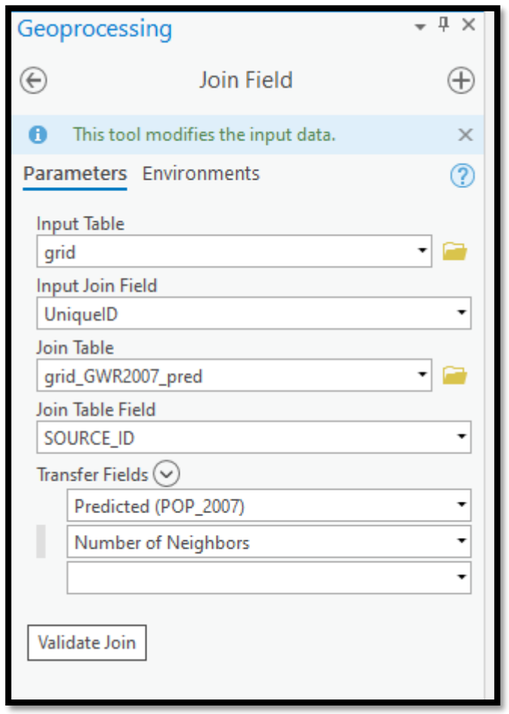




GWR

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