11 –Spatial Autoregressive Models with GeoDaSpace

Ness Sandoval
Sociology
Saint Louis University

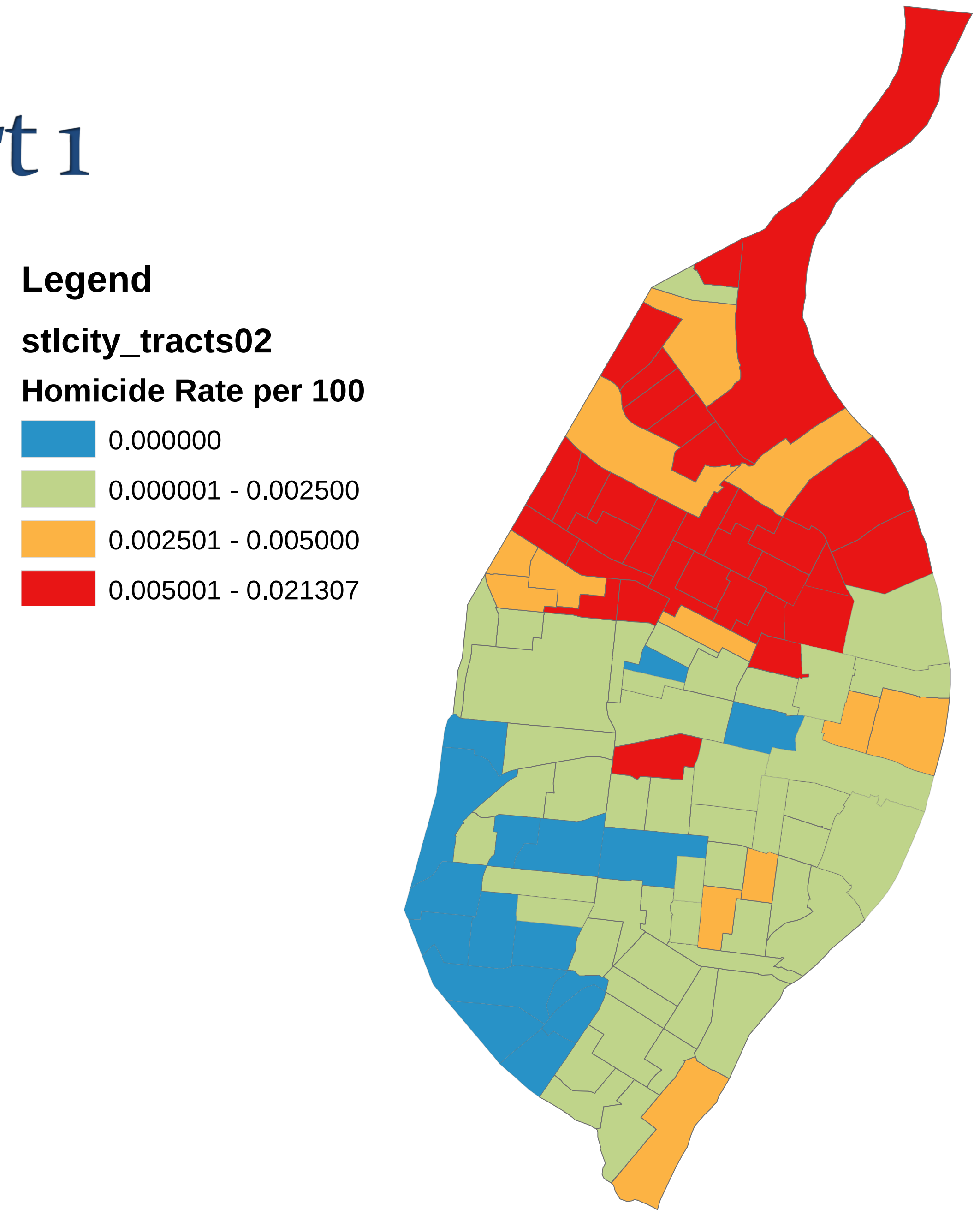
Outline

- Spatial Heterogeneity vs. Spatial Dependency
- Higher Ordered Spatial Regression
- GeoDa vs. GeoDaSpace
- Overview of GeoDaSpace
- Lab

Spatial Heterogeneity vs. Spatial Dependency

A nuanced discussion – Part 1

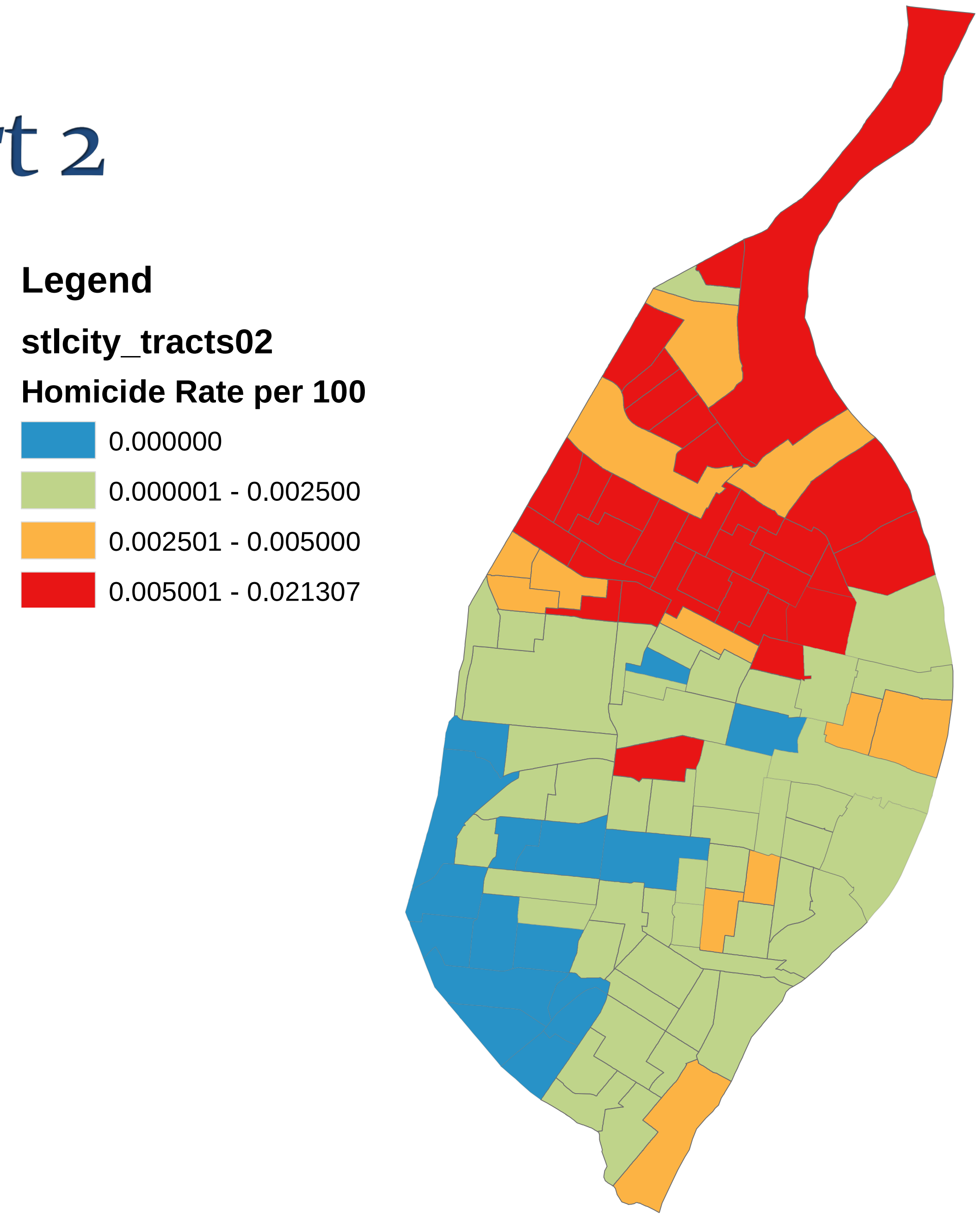
- One way to think about *spatial heterogeneity* and *spatial dependency* is to ask yourself:
  - (a) Is the intensity of occurrence of an event equally distributed across the landscape?
  - (b) Does the intensity at one location influence the intensity at neighboring locations?



A nuanced discussion – Part 2

If the answered is “yes” to the first question we are dealing with spatial heterogeneity

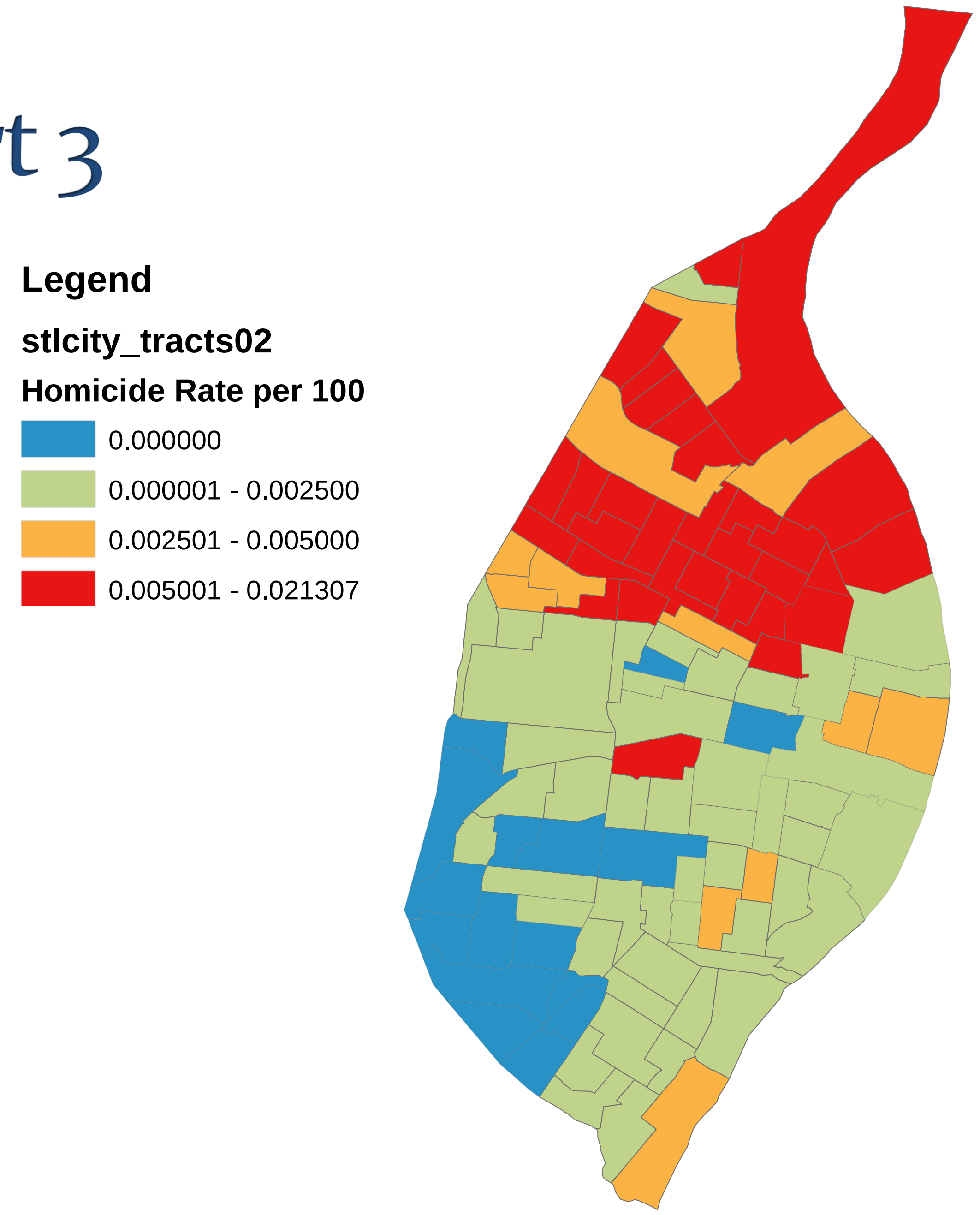
If the answered is “yes” to the second question we are dealing with spatial dependency.

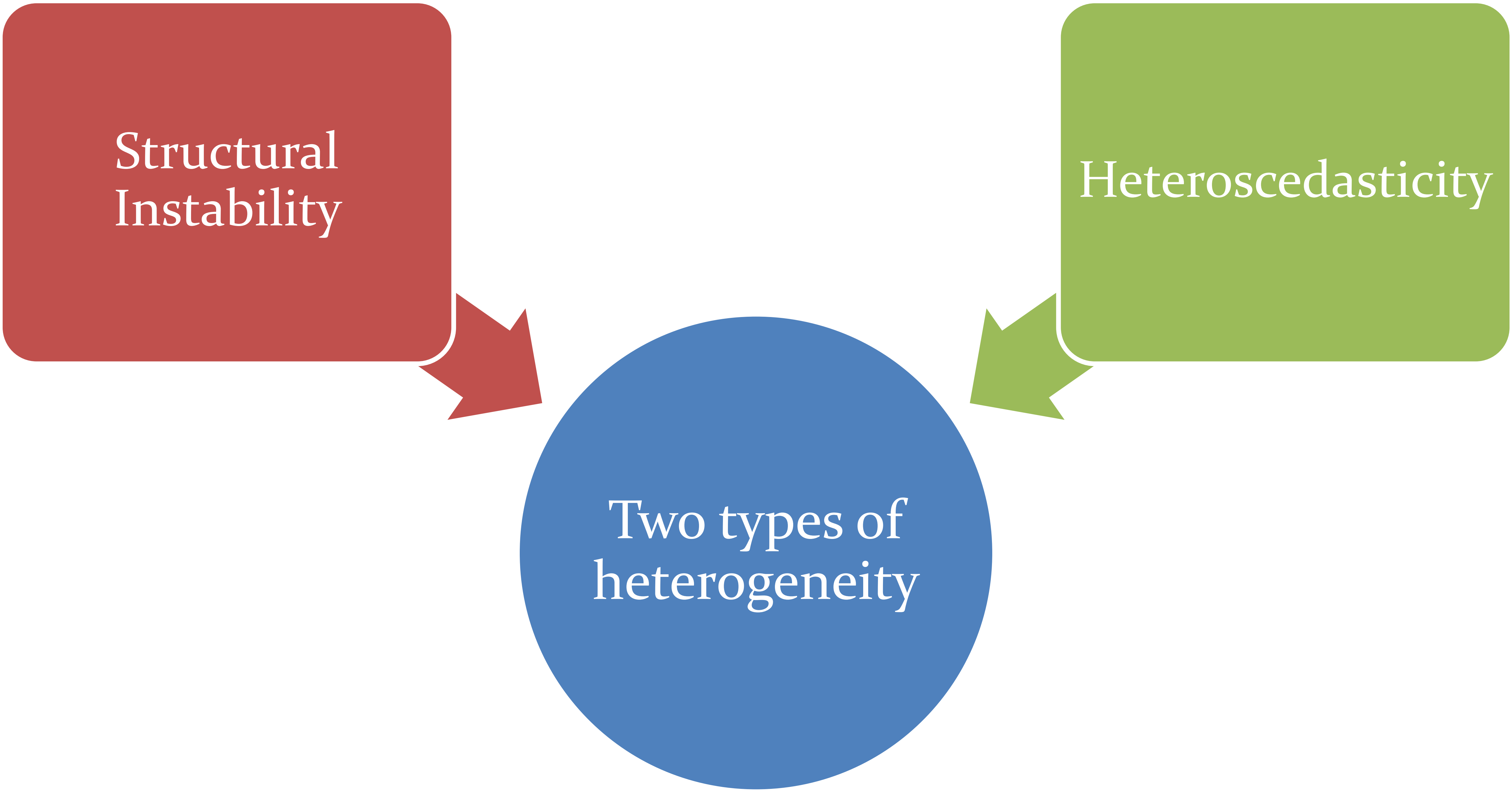


A nuanced discussion – Part 3

- Heterogeneity can be related to spatial structure or the spatial process generating data

- We need to challenge the assumption that regression coefficients are fixed throughout the sample




Structural Instability
Heteroscedasticity
Two types of heterogeneity

Spatial Regime Models

Spatial fixed effects models

- The spatial regime model is suitable in certain instances in which the assumption of a fixed relation between the explanatory variables and the dependent variable across the study is not tenable.

- Spatial Heterogeneity may be present, in the form of different intercepts and/or slopes in the regression equation for subsets of the data.

- This is often referred to as structural instability or structural change in the academic literature.

- If this is the case, we may want to include additional variables in the SAR models. When the different subsets in the data correspond to regions or spatial clusters, it is called spatial regimes model.

Higher Ordered Spatial Regression

Review of SAR Models – Baseline Models

Spatial Lag

$$y = B_0 + \rho Wy + X\beta + \varepsilon$$

Spatial Error

$$y = B_0 + X\beta + \lambda \mathrm{W}\varepsilon + \xi$$

Higher Ordered Models

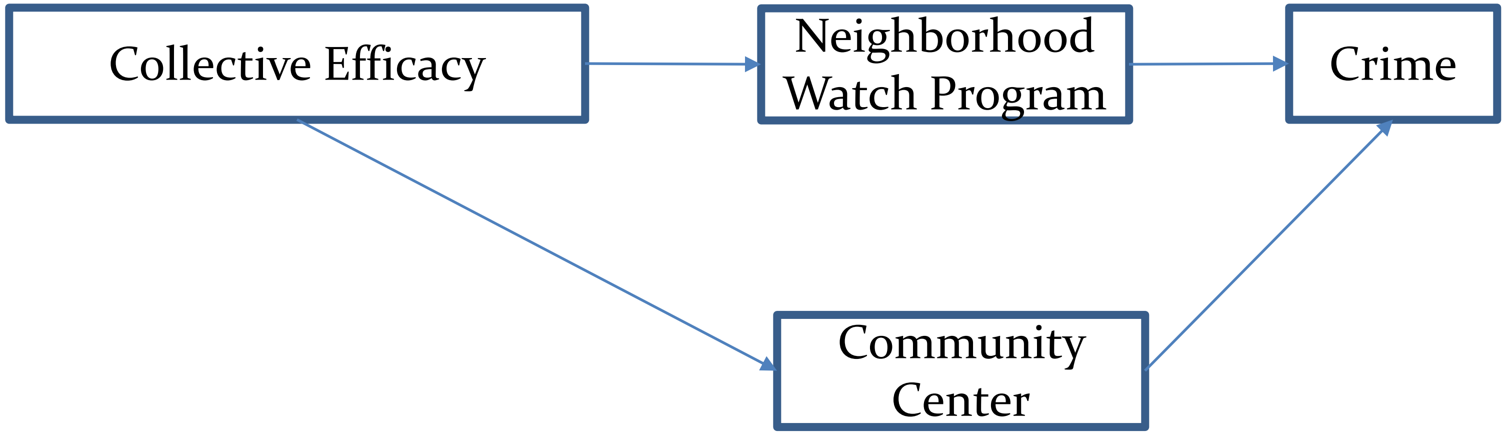
Spatial Lag

$$y = B_0 + \rho Wy + X\beta + Y\gamma + \varepsilon$$

Where $Y\gamma$ is an endogenous variable

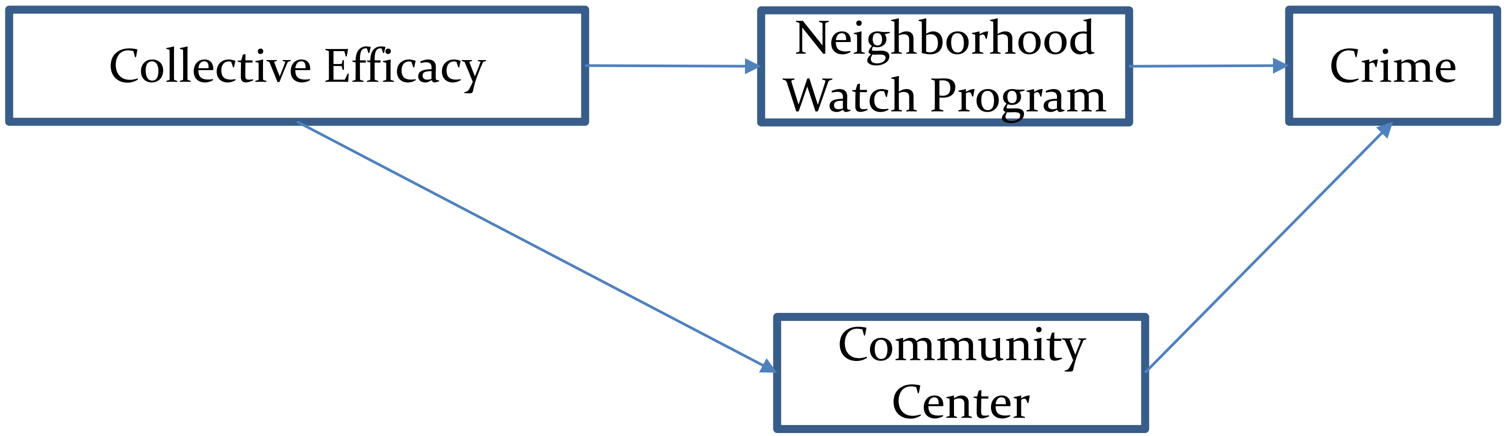
Let’s define some terms

- What is an endogenous variable?
  - One can think of an endogenous variables as been determined or influenced by “other” variables. We typically call the these “other” variables are called exogenous variables



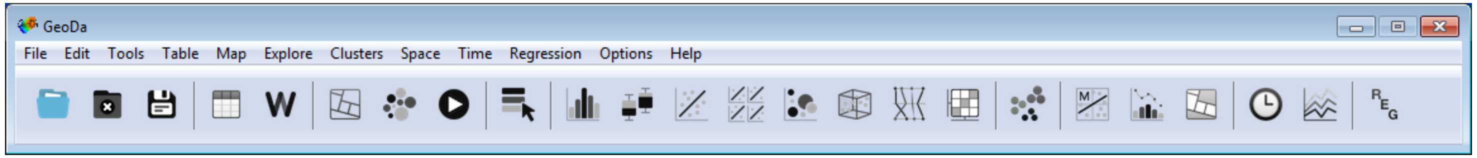
Let's define some terms

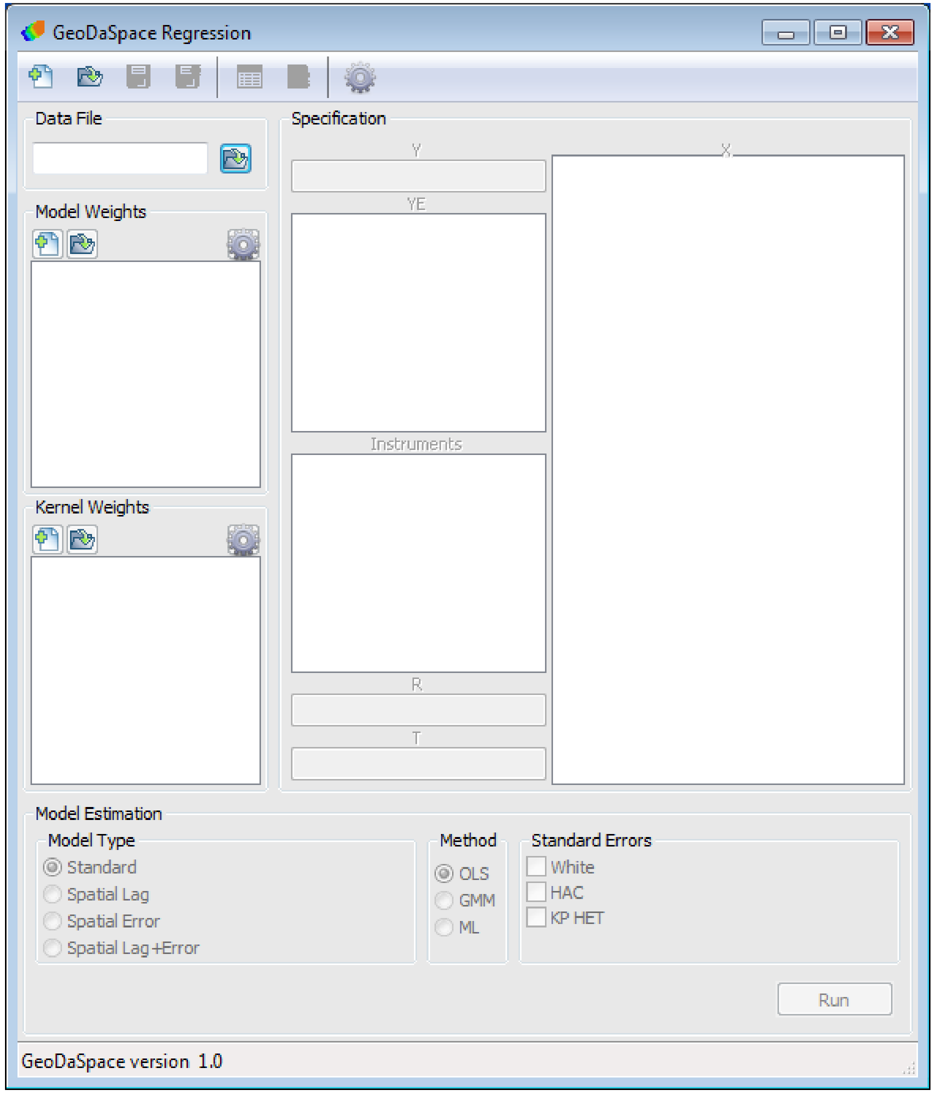
- What is an instrumental variable?
  - An instrumental variable does not have a direct correlation with the dependent variable
  - However, an instrumental variable does have a direct correlation with an independent variable which has a correlation with the dependent variable.



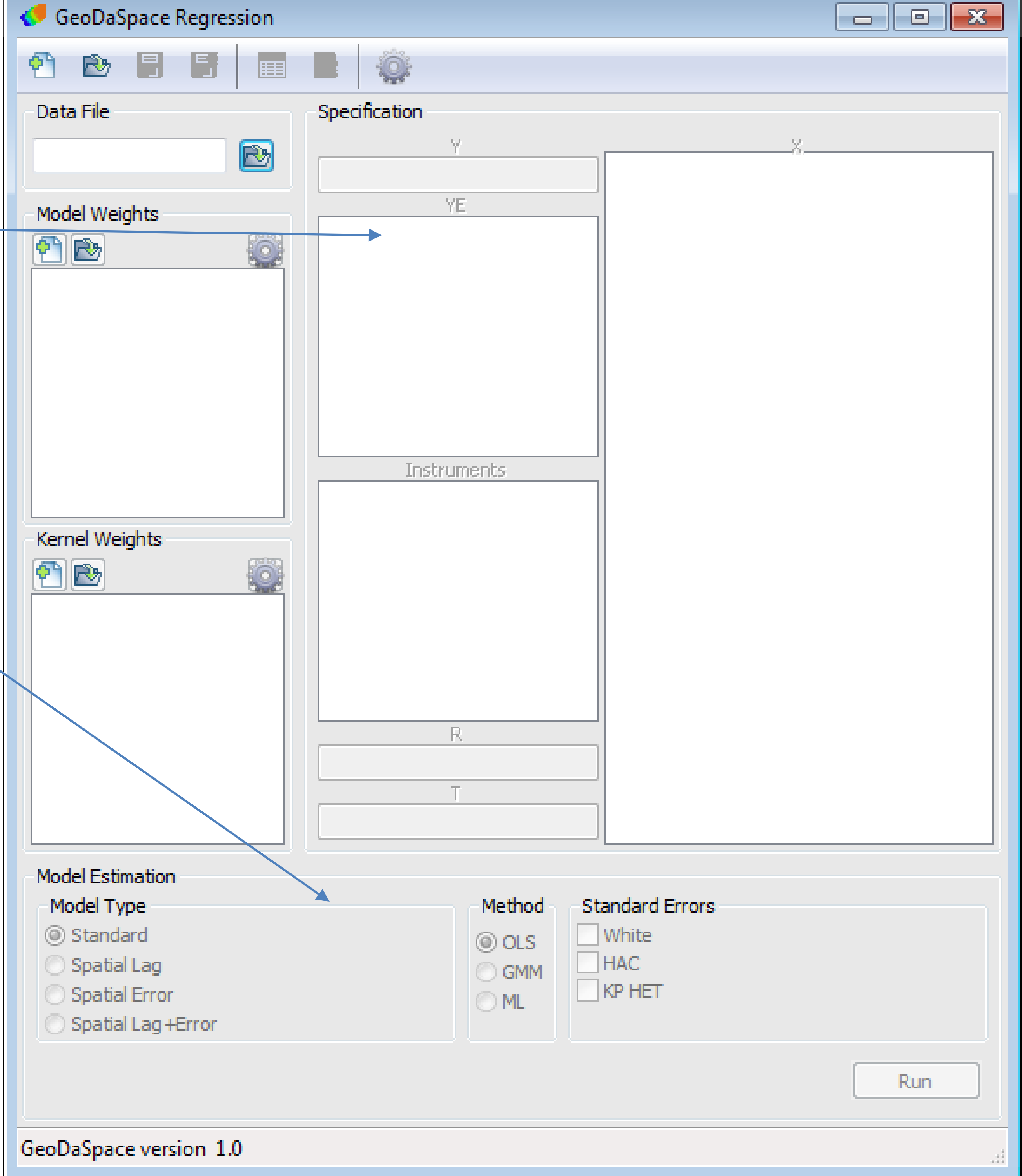
GeoDa vs GeoDaSpace

GeoDa vs GeoDaSpace



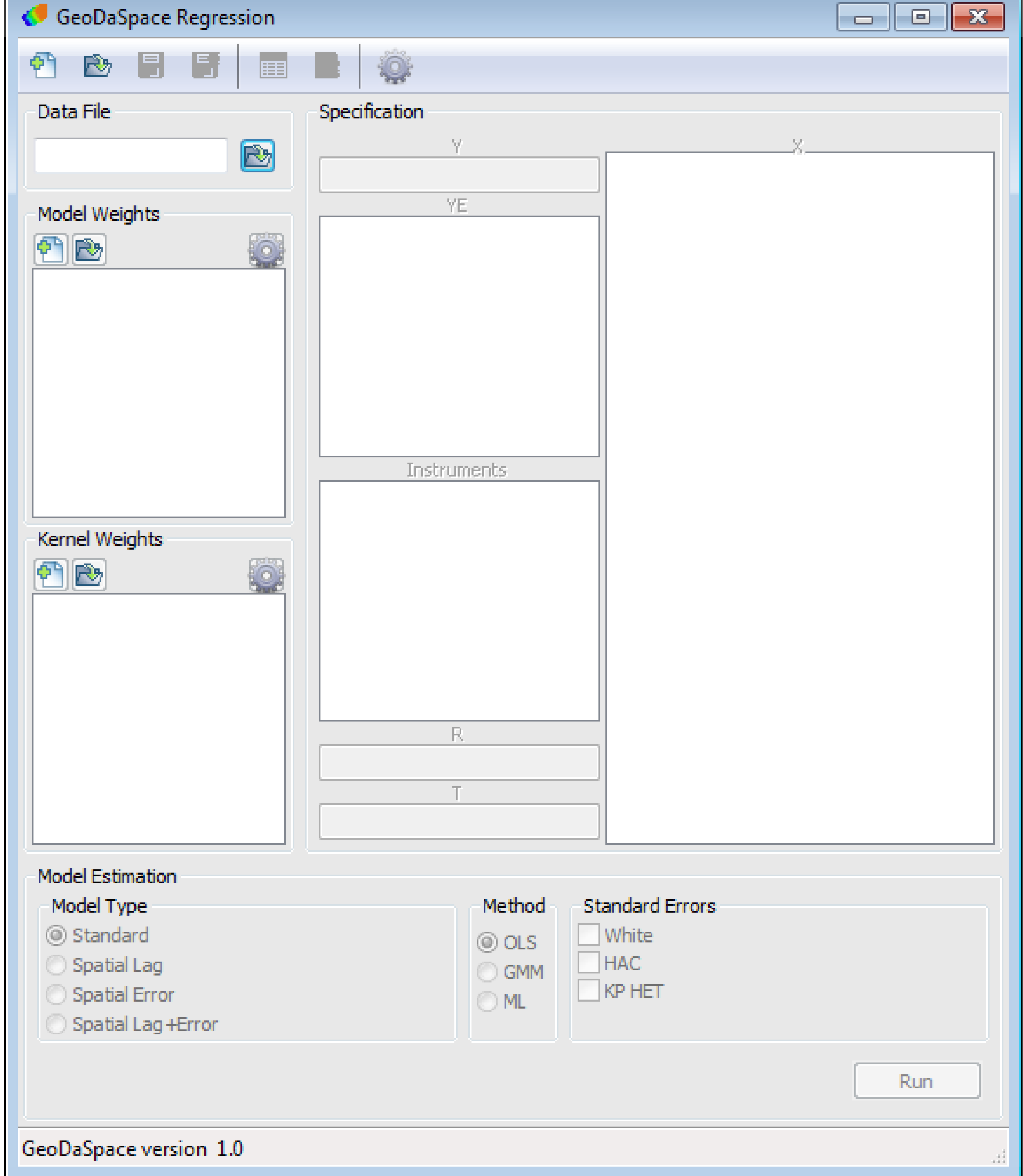


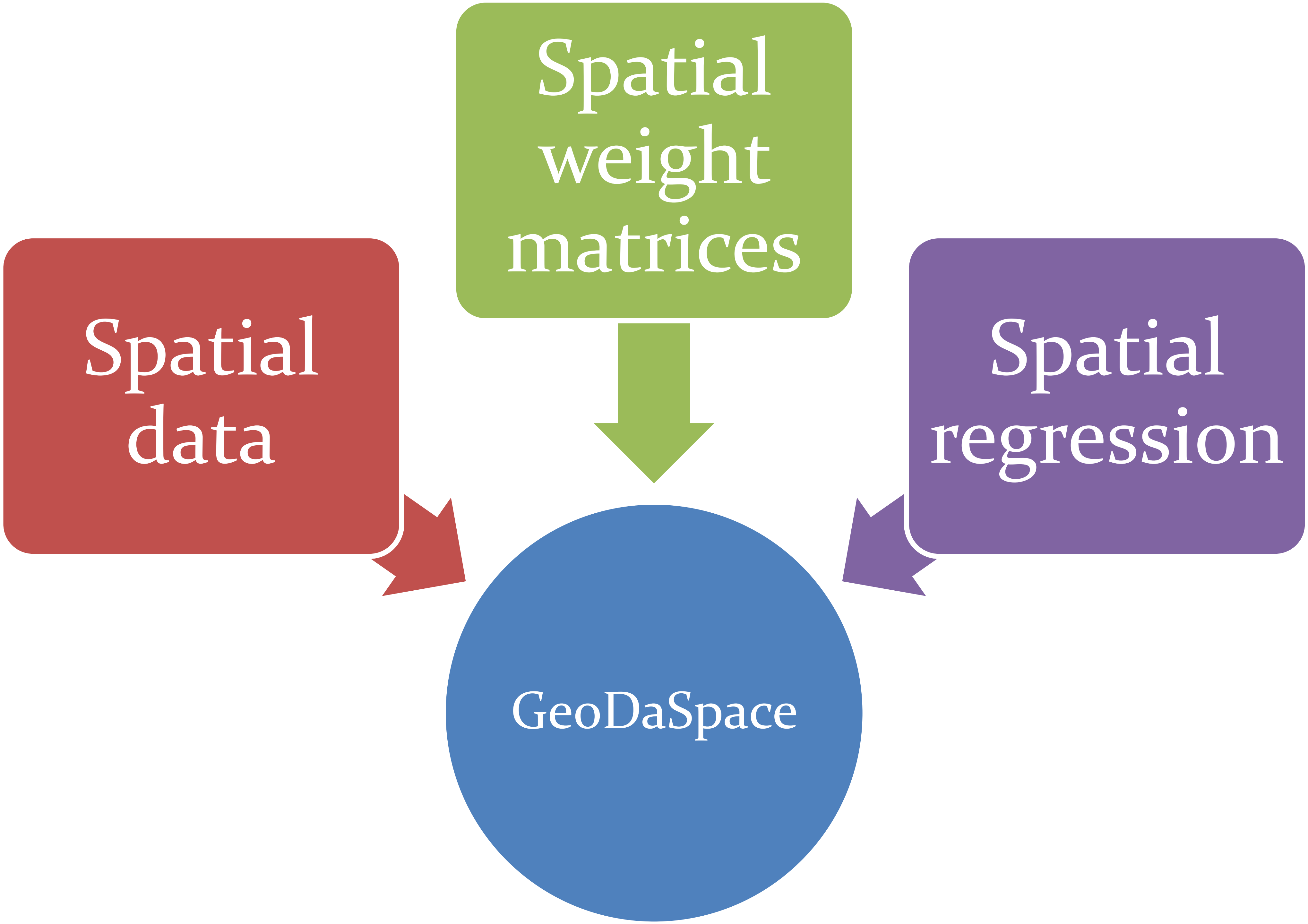
- *GeoDaSpace* includes two menus for spatial regression:
  - specification
  - model estimation



- The design of *GeoDaSpace* ***does not*** consist of an interactive environment combining maps with statistical graphs, using a technology of dynamically linked windows.

- *GeoDaSpace* is geared to the higher order analysis of discrete geospatial data
    - characterized by their location in space either as points (point coordinates) or polygon (polygon boundary coordinates).



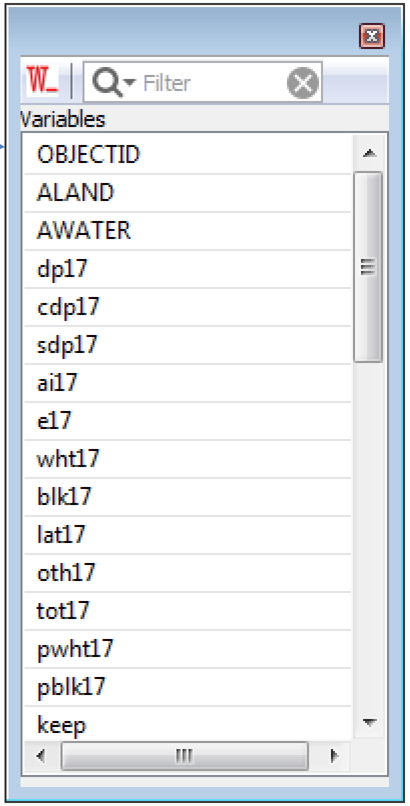

Spatial weight matrices
Spatial data
Spatial regression
GeoDaSpace

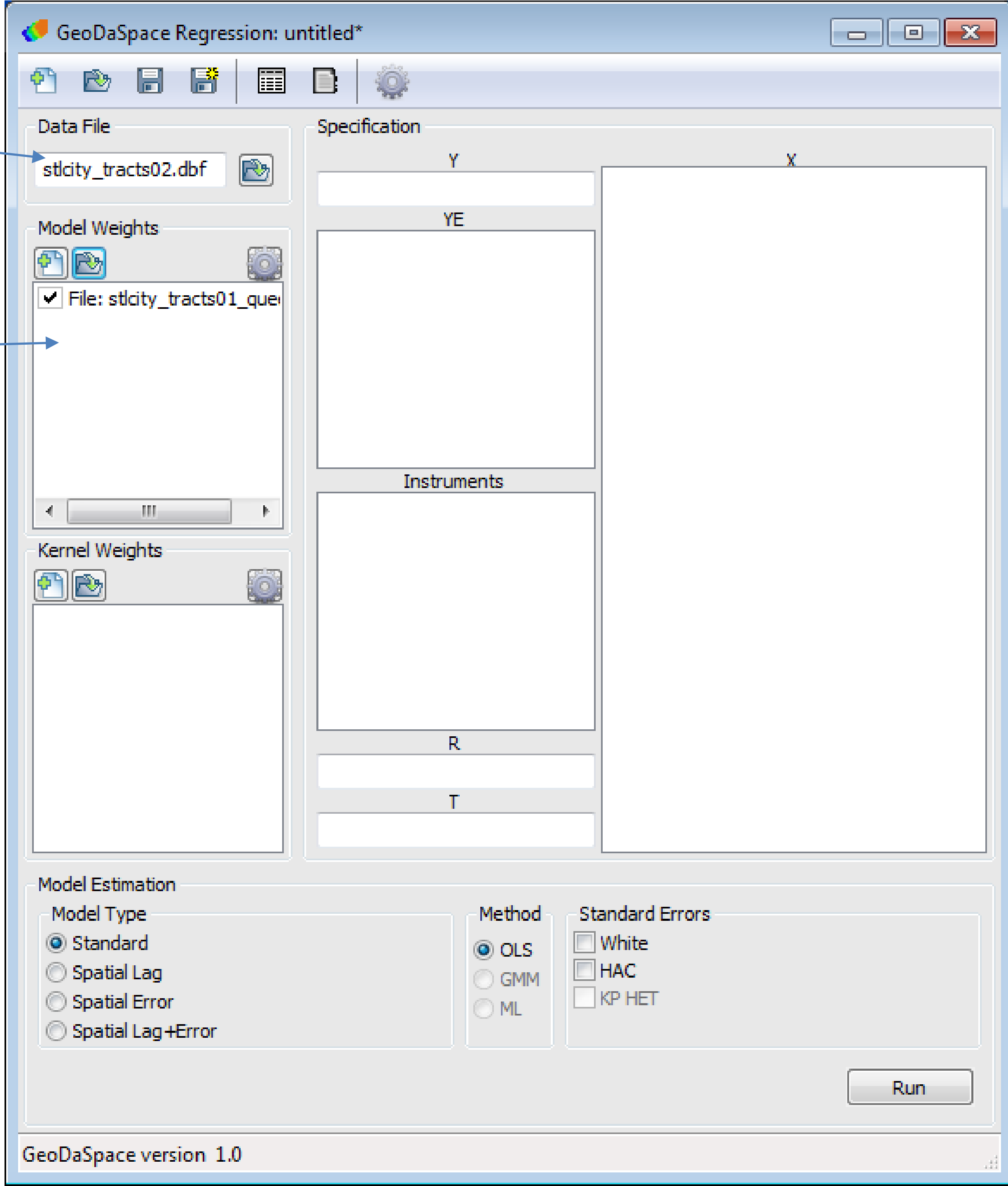
Working with GeoDaSpace

1. Load your dbf file from the shapefile

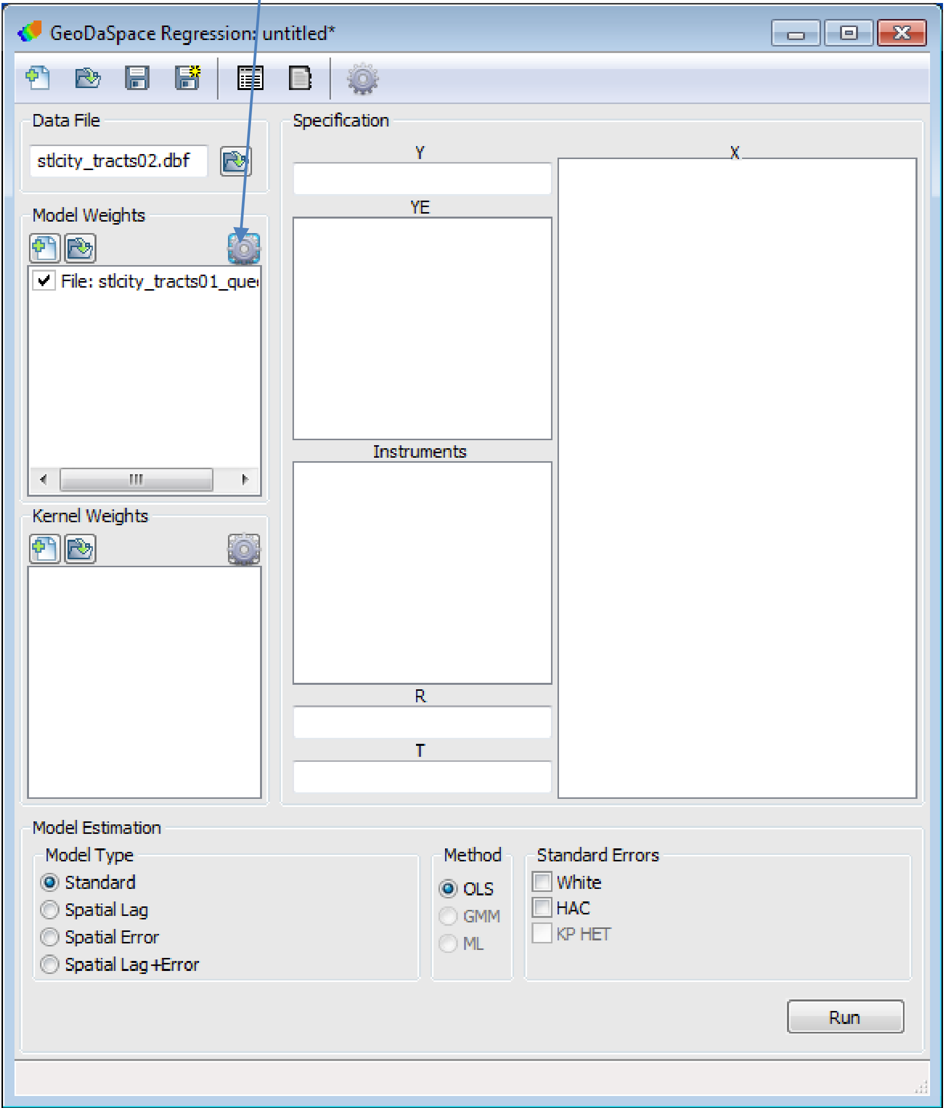
2. Load your weight file

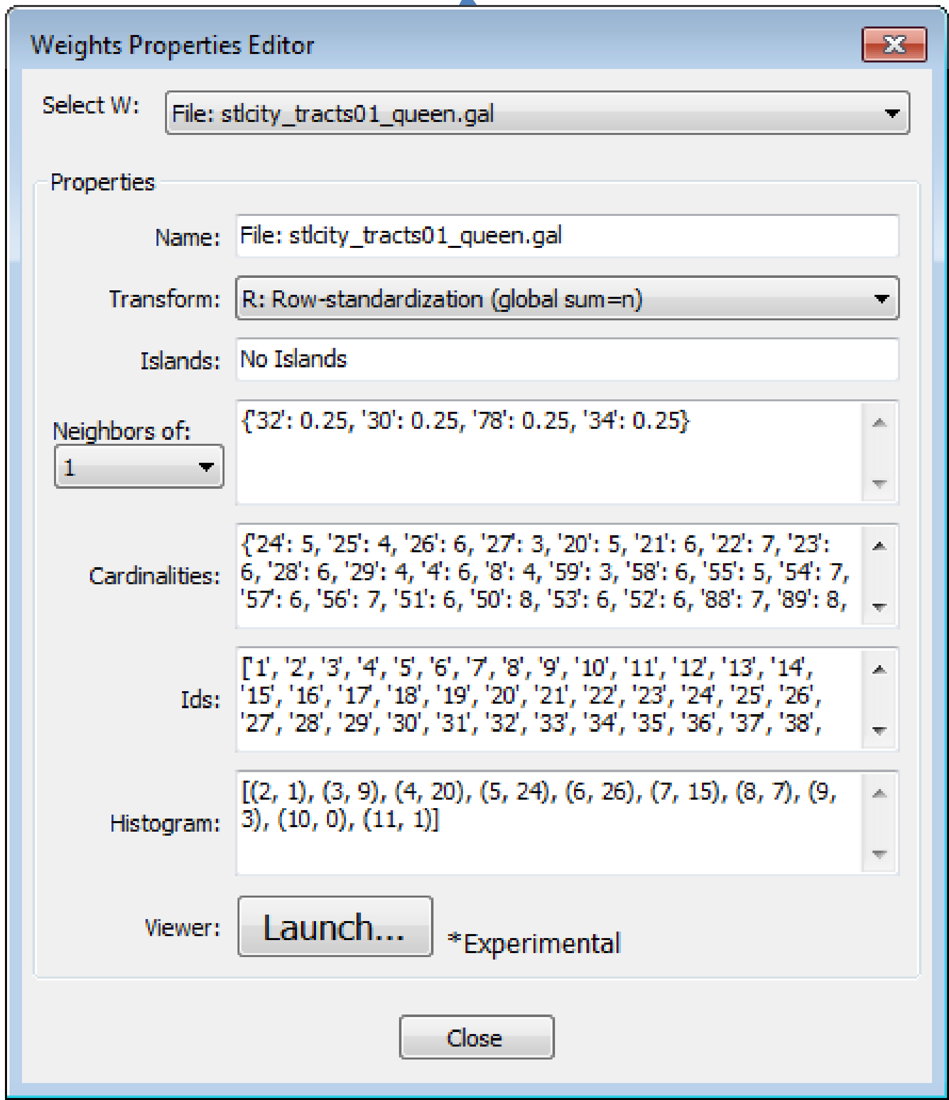
3. You should have box with your variables



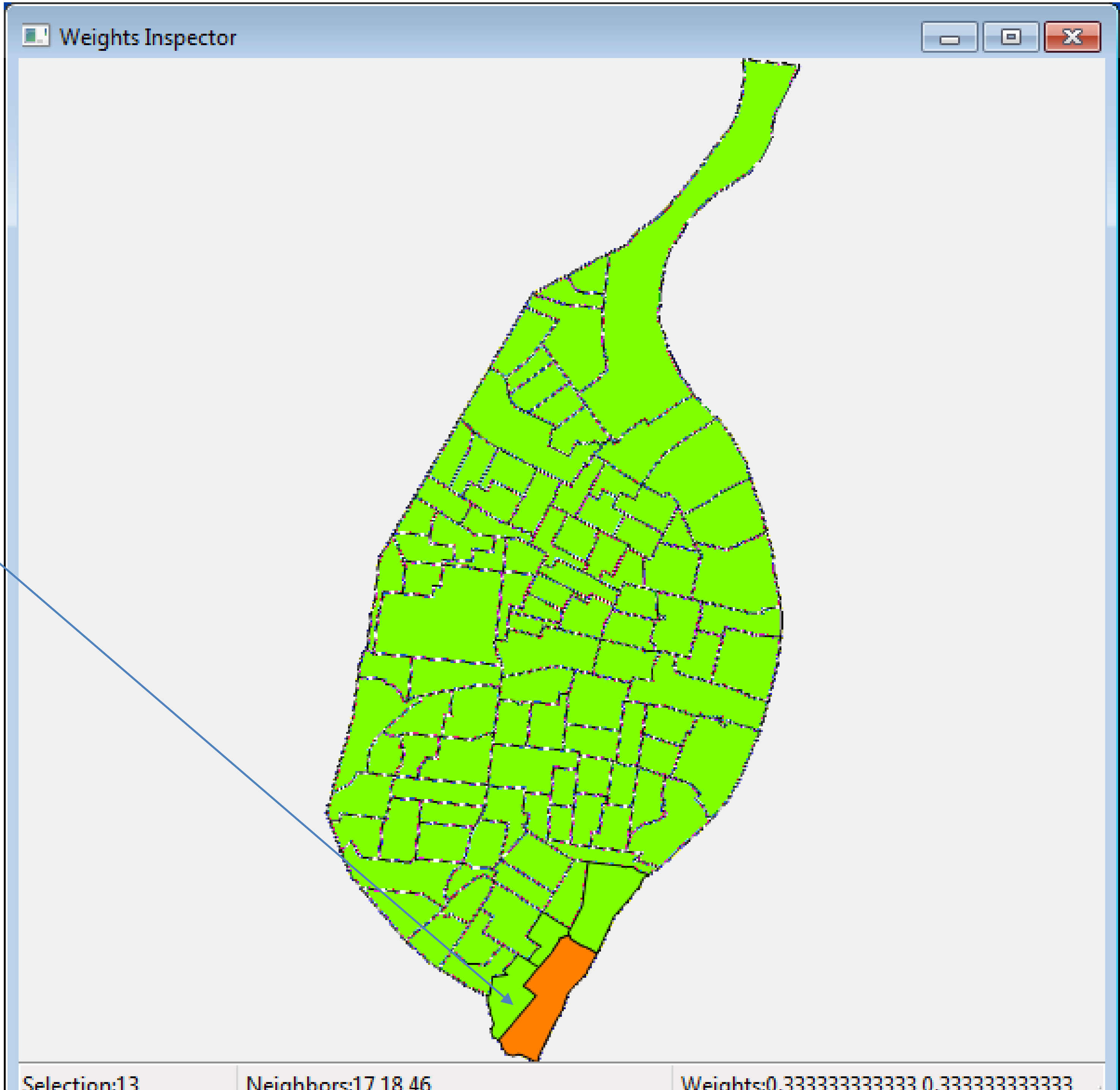


4. Select the “setting” buttons in the Model Weights box, and you should get this box

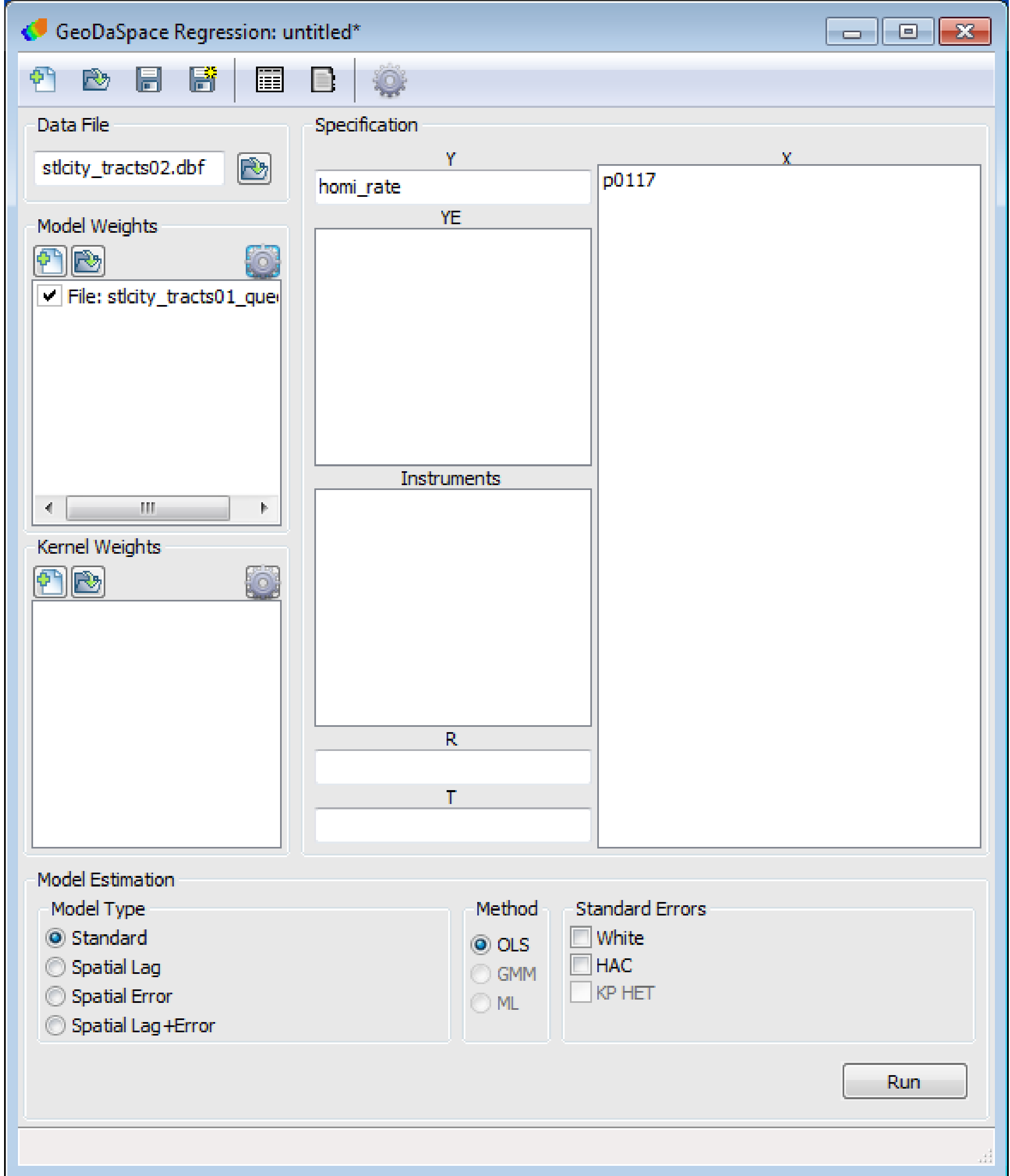




5. Link the dbf to the shapefile and you view the weight file and how it constructed the weight matrix.



6. Let's replicate the results from GeoDa



OLS

Results from GeoDa

Regression Report

```
>>03/27/19 08:39:21
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set            :  stlcity_tracts02
Dependent Variable  :   homi_rate  Number of Observations:  106
Mean dependent var  :  0.00340081  Number of Variables   :    2
S.D. dependent var  :  0.00407594  Degrees of Freedom    :  104

R-squared           :    0.240954  F-statistic           :     33.0141
Adjusted R-squared  :    0.233655  Prob(F-statistic)     :9.18912e-008
Sum squared residual:  0.00133669  Log likelihood        :     447.486
Sigma-square        :1.28528e-005  Akaike info criterion :    -890.971
S.E. of regression  :  0.00358507  Schwarz criterion     :    -885.644
Sigma-square ML     :1.26103e-005
S.E of regression ML:  0.00355109

-----------------------------------------------------------------------------
       Variable      Coefficient       Std.Error     t-Statistic   Probability
-----------------------------------------------------------------------------
          CONSTANT  -0.000400913    0.000747688       -0.536204      0.59296
             p0117     0.0140372     0.00244304         5.74579      0.00000
-----------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER   4.047352
TEST ON NORMALITY OF ERRORS
TEST                  DF           VALUE             PROB
Jarque-Bera            2           224.9842          0.00000

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                  DF           VALUE             PROB
Breusch-Pagan test     1             8.4759          0.00360
Koenker-Bassett test   1             2.0447          0.15274

DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : stlcity_tracts01_queen
   (row-standardized weights)
TEST                           MI/DF        VALUE           PROB
Moran's I (error)              0.4312        7.7260         0.00000
Lagrange Multiplier (lag)         1         69.3286         0.00000
Robust LM (lag)                   1         17.9192         0.00002
Lagrange Multiplier (error)       1         51.6868         0.00000
Robust LM (error)                 1          0.2774         0.59844
Lagrange Multiplier (SARMA)       2         69.6059         0.00000
============================== END OF REPORT ================================
```

Results from GeoDaSpace

Untitled*

```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES
-----------------------------------------
Data set            :stlcity_tracts02.dbf
Weights matrix      :File: stlcity_tracts01_queen.gal
Dependent Variable  :   homi_rate                Number of Observations:         106
Mean dependent var  :      0.0034                Number of Variables   :           2
S.D. dependent var  :      0.0041                Degrees of Freedom    :         104
R-squared           :      0.2410
Adjusted R-squared  :      0.2337
Sum squared residual:       0.001                F-statistic           :     33.0141
Sigma-square        :       0.000                Prob(F-statistic)     :   9.189e-08
S.E. of regression  :       0.004                Log likelihood        :     447.486
Sigma-square ML     :       0.000                Akaike info criterion :    -890.971
S.E of regression ML:      0.0036                Schwarz criterion     :    -885.644

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     t-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT      -0.0004009       0.0007477      -0.5362041       0.5929625
               p0117       0.0140372       0.0024430       5.7457892       0.0000001
------------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER          4.047

TEST ON NORMALITY OF ERRORS
TEST                             DF        VALUE           PROB
Jarque-Bera                       2       224.984          0.0000

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                             DF        VALUE           PROB
Breusch-Pagan test                1         8.476          0.0036
Koenker-Bassett test              1         2.045          0.1527

DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST                           MI/DF       VALUE           PROB
Lagrange Multiplier (lag)         1        69.329          0.0000
Robust LM (lag)                   1        17.919          0.0000
Lagrange Multiplier (error)       1        51.687          0.0000
Robust LM (error)                 1         0.277          0.5984
Lagrange Multiplier (SARMA)       2        69.606          0.0000

================================ END OF REPORT =====================================
```

Spatial Lag

Results from GeoDa

Regression Report

```
>>03/27/19 08:42:20
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set             : stlcity_tracts02
Spatial Weight       : stlcity_tracts01_queen
Dependent Variable   :    homi_rate  Number of Observations:  106
Mean dependent var   :   0.00340081  Number of Variables   :    3
S.D. dependent var   :   0.00407594  Degrees of Freedom    :  103
Lag coeff.   (Rho)   :     0.720932

R-squared            :     0.614184  Log likelihood        :     476.368
Sq. Correlation      : -             Akaike info criterion :    -946.736
Sigma-square         :6.40967e-006   Schwarz criterion     :    -938.745
S.E of regression    :   0.00253173

-----------------------------------------------------------------------------
       Variable       Coefficient      Std.Error       z-value    Probability
-----------------------------------------------------------------------------
       W_homi_rate      0.720932       0.0733364       9.83048     0.00000
          CONSTANT  -0.000578612     0.000536325      -1.07885     0.28066
             p0117    0.00579562      0.00183897       3.15156     0.00162
-----------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                         DF       VALUE        PROB
Breusch-Pagan test                           1          2.0104     0.15623

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL LAG DEPENDENCE FOR WEIGHT MATRIX : stlcity_tracts01_queen
TEST                                         DF       VALUE        PROB
Likelihood Ratio Test                        1         57.7647     0.00000
============================== END OF REPORT ================================
```

Results from GeoDaSpace

```
REGRESSION
----------
SUMMARY OF OUTPUT: MAXIMUM LIKELIHOOD SPATIAL LAG (METHOD = FULL)
-----------------------------------------------------------------

Data set            :stlcity_tracts02.dbf
Weights matrix      :File: stlcity_tracts01_queen.gal
Dependent Variable  :   homi_rate                Number of Observations:       106
Mean dependent var  :      0.0034                Number of Variables   :         3
S.D. dependent var  :      0.0041                Degrees of Freedom    :       103
Pseudo R-squared    :      0.6235
Spatial Pseudo R-squared:  0.4008
Sigma-square ML     :       0.000                Log likelihood        :   476.368
S.E of regression   :       0.003                Akaike info criterion :  -946.736
                                                 Schwarz criterion     :  -938.745

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     z-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT      -0.0005786       0.0005363      -1.0788828       0.2806400
         W_homi_rate       0.7209334       0.0733390       9.8301534       0.0000000
               p0117       0.0057956       0.0018389       3.1516165       0.0016237
------------------------------------------------------------------------------------
================================ END OF REPORT =====================================
```

Spatial Error

Results from GeoDa

Results from GeoDaSpace

Regression Report

```
>>03/27/19 08:45:08
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL ERROR MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : stlcity_tracts02
Spatial Weight      : stlcity_tracts01_queen
Dependent Variable  :   homi_rate  Number of Observations:  106
Mean dependent var  :    0.003401  Number of Variables   :    2
S.D. dependent var  :    0.004076  Degrees of Freedom    :  104
Lag coeff. (Lambda) :    0.769312

R-squared           :    0.602553  R-squared (BUSE)      : -
Sq. Correlation     : -            Log likelihood        :  473.454778
Sigma-square        :6.60291e-006  Akaike info criterion :     -942.91
S.E of regression   :  0.00256961  Schwarz criterion     :    -937.583

-----------------------------------------------------------------------------
       Variable       Coefficient      Std.Error       z-value    Probability
-----------------------------------------------------------------------------
          CONSTANT    0.00245817     0.00123074       1.99731      0.04579
             p0117    0.00438642     0.00226379       1.93764      0.05267
            LAMBDA      0.769312      0.0696768       11.0411      0.00000
-----------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                        DF       VALUE         PROB
Breusch-Pagan test                          1          1.6470      0.19936

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL ERROR DEPENDENCE FOR WEIGHT MATRIX : stlcity_tracts01_queen
TEST                                        DF       VALUE         PROB
Likelihood Ratio Test                       1         51.9385      0.00000
============================== END OF REPORT ================================
```

```
REGRESSION
----------
SUMMARY OF OUTPUT: MAXIMUM LIKELIHOOD SPATIAL ERROR (METHOD = FULL)
-------------------------------------------------------------------
Data set            :stlcity_tracts02.dbf
Weights matrix      :File: stlcity_tracts01_queen.gal
Dependent Variable  :   homi_rate                Number of Observations:         106
Mean dependent var  :      0.0034                Number of Variables   :           2
S.D. dependent var  :      0.0041                Degrees of Freedom    :         104
Pseudo R-squared    :      0.2410
Sigma-square ML     :       0.000                Log likelihood        :     473.455
S.E of regression   :       0.003                Akaike info criterion :    -942.910
                                                 Schwarz criterion     :    -937.583

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     z-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT       0.0024582       0.0012307       1.9973160       0.0457909
              lambda       0.7693111       0.0696769      11.0411133       0.0000000
               p0117       0.0043864       0.0022638       1.9376453       0.0526665
------------------------------------------------------------------------------------
================================ END OF REPORT =====================================
```

An Higher Ordered Model

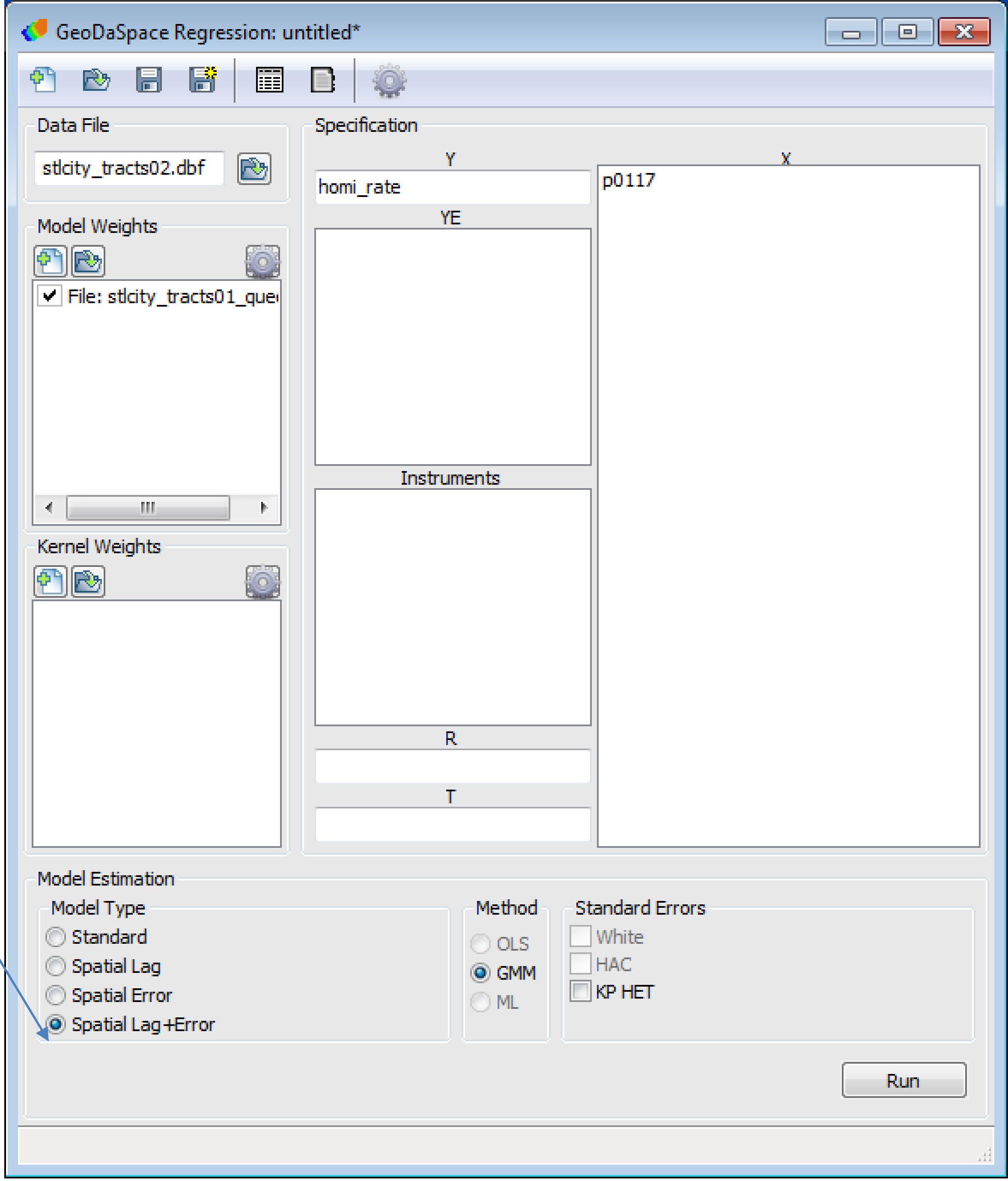
Spatial Lag and Spatial Error Model

$$y = B_0 + \rho W y + X\beta + \varepsilon$$

Where

$$\varepsilon =$$

$$\lambda W \varepsilon + \varepsilon$$



```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIALLY WEIGHTED TWO STAGE LEAST SQUARES (HOM)
-------------------------------------------------------------------
Data set            :stlcity_tracts02.dbf
Weights matrix      :File: stlcity_tracts01_queen.gal
Dependent Variable  :   homi_rate                Number of Observations:         106
Mean dependent var  :      0.0034                Number of Variables   :           3
S.D. dependent var  :      0.0041                Degrees of Freedom    :         103
Pseudo R-squared    :      0.6329
Spatial Pseudo R-squared:  0.4684
N. of iterations    :           1

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     z-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT      -0.0006033       0.0003979      -1.5159112       0.1295418
         W_homi_rate       0.9106472       0.1314021       6.9302327       0.0000000
               p0117       0.0033861       0.0024243       1.3966910       0.1625065
              lambda      -0.4582897       0.3759493      -1.2190201       0.2228366
------------------------------------------------------------------------------------
Instrumented: W_homi_rate
Instruments: W_p0117
================================ END OF REPORT =====================================
```

Variable	OLS	SLM	SEM	SLM and SEM	SLM and SEM With KP HET
Poverty	.0140372 (.00244)***	0.00579562 (0.00183897) **	0.00438642 (0.00226379)*	0.0033861 (0.0024243)	0.0033853 (0.4240851)
Constant	-0.0004 (0.0007)	-0.000578612 0.000536325	0.00245817 (0.00123074)	-0.0006033 (0.0003979	-0.0006025 (0.0003500)
ρ		0.720932 (0.0733365) ***		0.9106472 (0.1314021)***	0.9101785 (0.1708339)***
λ			0.769312 (0.0696768)***	-0.4582897 (-0.4582897)	-0.1473622 (0.4240851)
r-square	0.240954	0.614184	0.602553	0.6329	0.6329
Log likelihood	447.486	476.368	473.454	NA	NA
AIC	-890.971	-946.736	-942.91	NA	NA
Moran's I Residual	.431166***	-0.0570314	-0.0455698	NA	NA
N	106	106	106	106	106

$* \le .05, ** \le .01, *** \le .001$

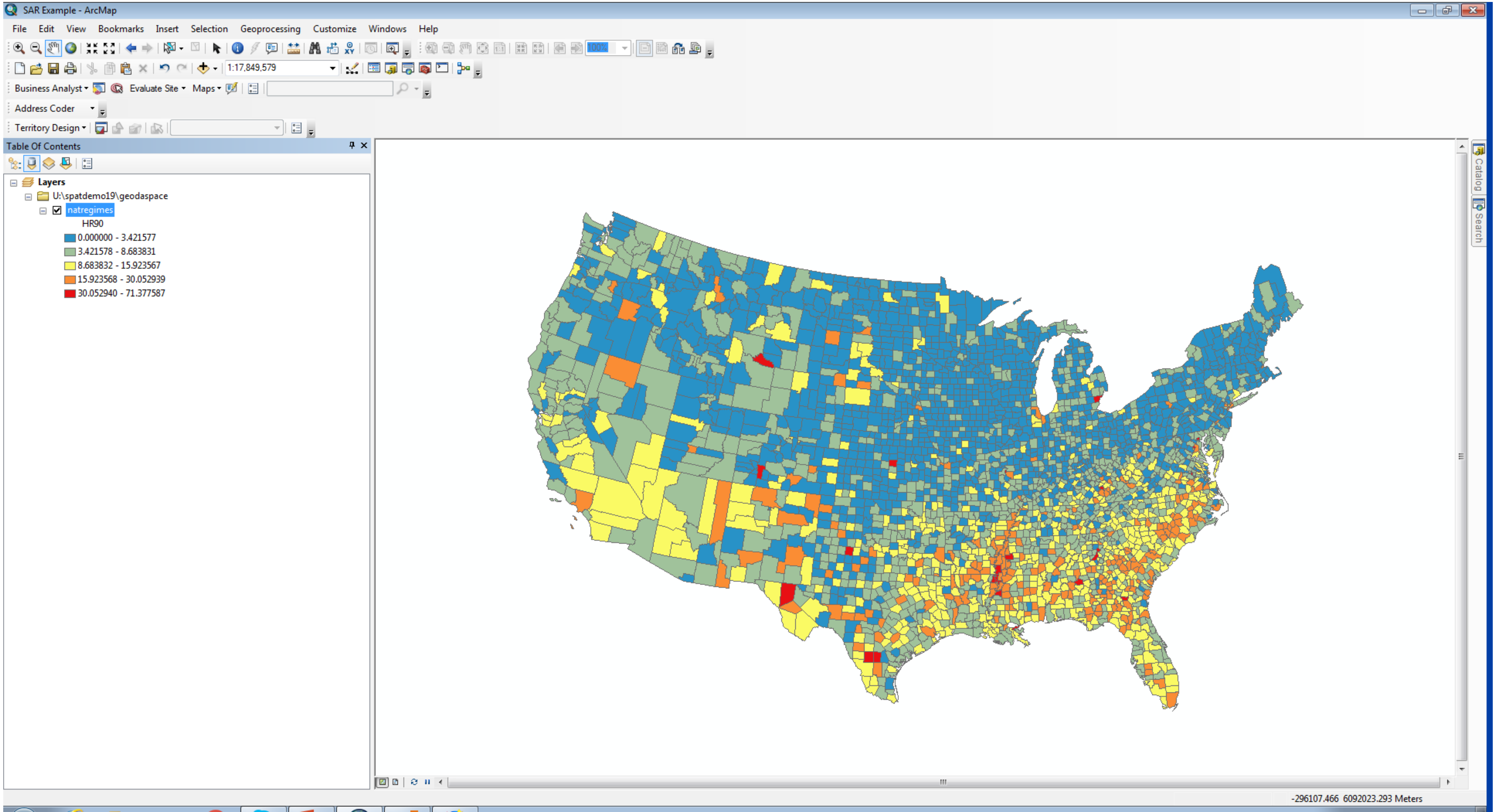
Standard Errors in Parentheses

Reflections on SLM and SEM

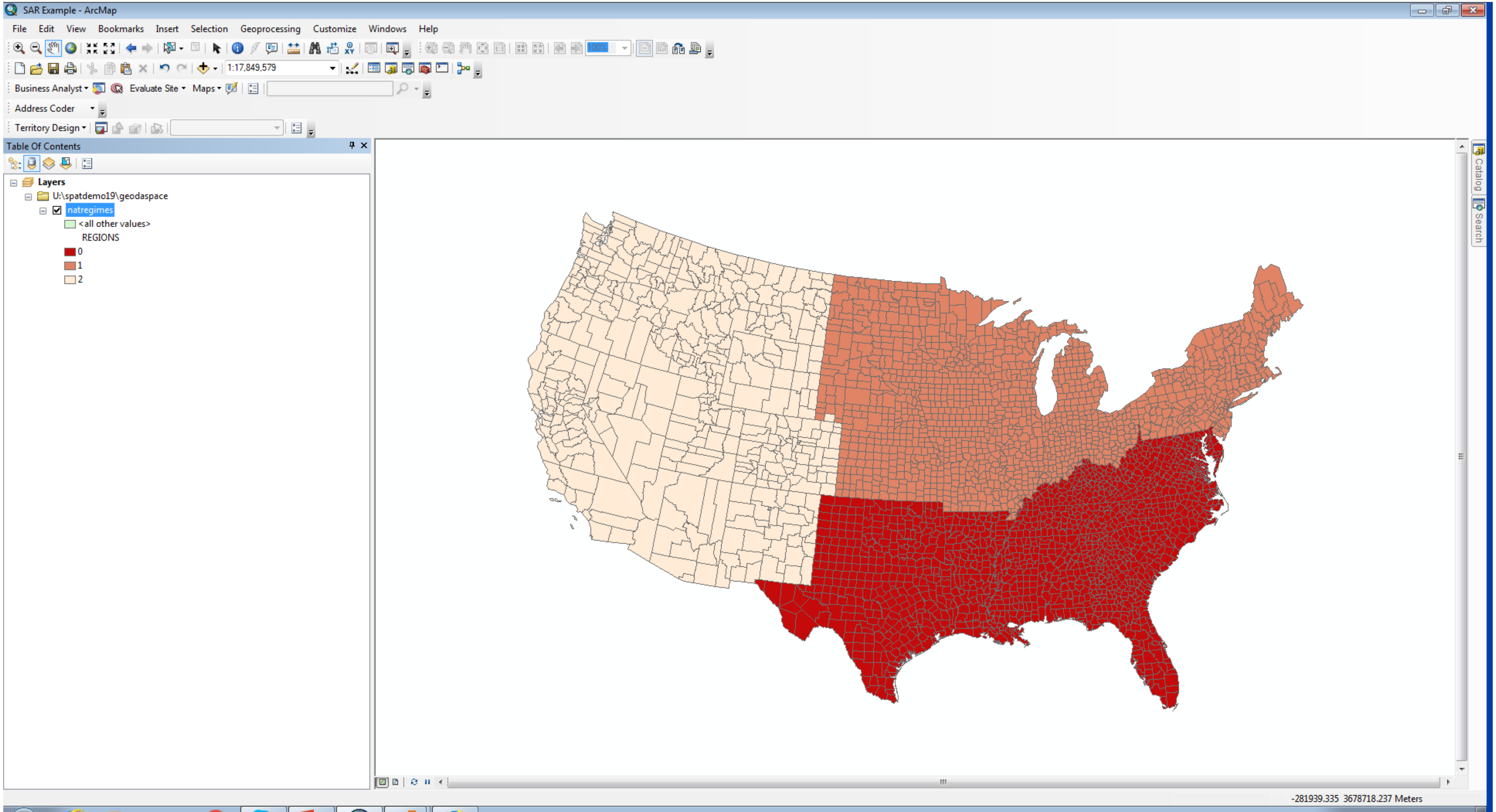
- The joint model is a special case…
  - Spatial Heteroskedasticity is a problem, especially in the error model
  - Spatial Autocorrelation is still a problem, especially in the error model
  - There is no clear preference for the SLM or SEM
- It is possible that both ρ and λ can be significant

Spatial Regime Models

This is a map of the Homicide Rate for the U.S. Counties for 1990



This is a map of the three regions for the U.S. Counties for 1990



First we want to build a basic OLS model

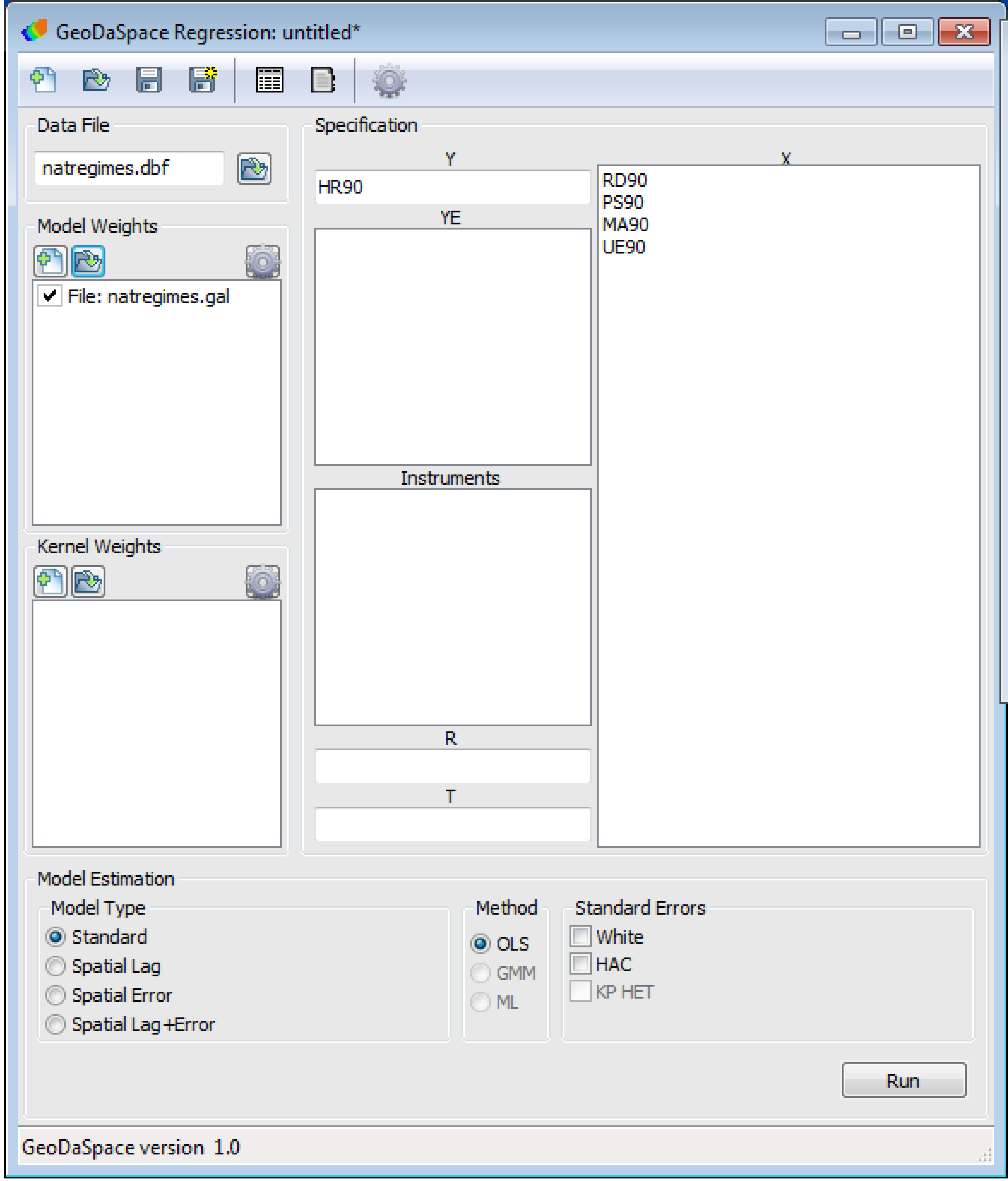
Where:
HR90=Homicide Rate

RD90=Resource Deprivation

PS90=Population Structure Component

MA90 = Median Age

UE90=Unemployment Rate

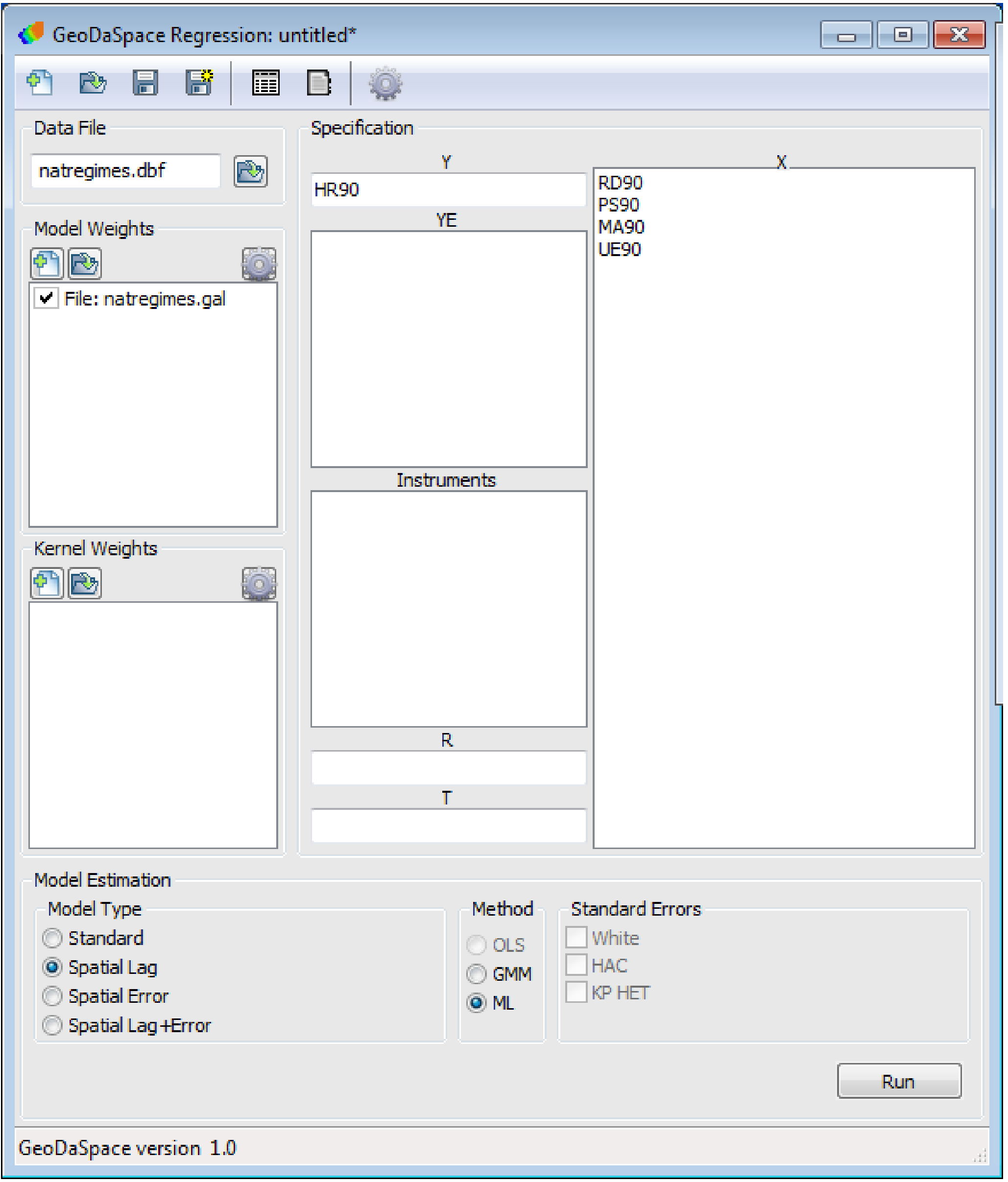


Review the output on the right from GeoDaSpace

Untitled*

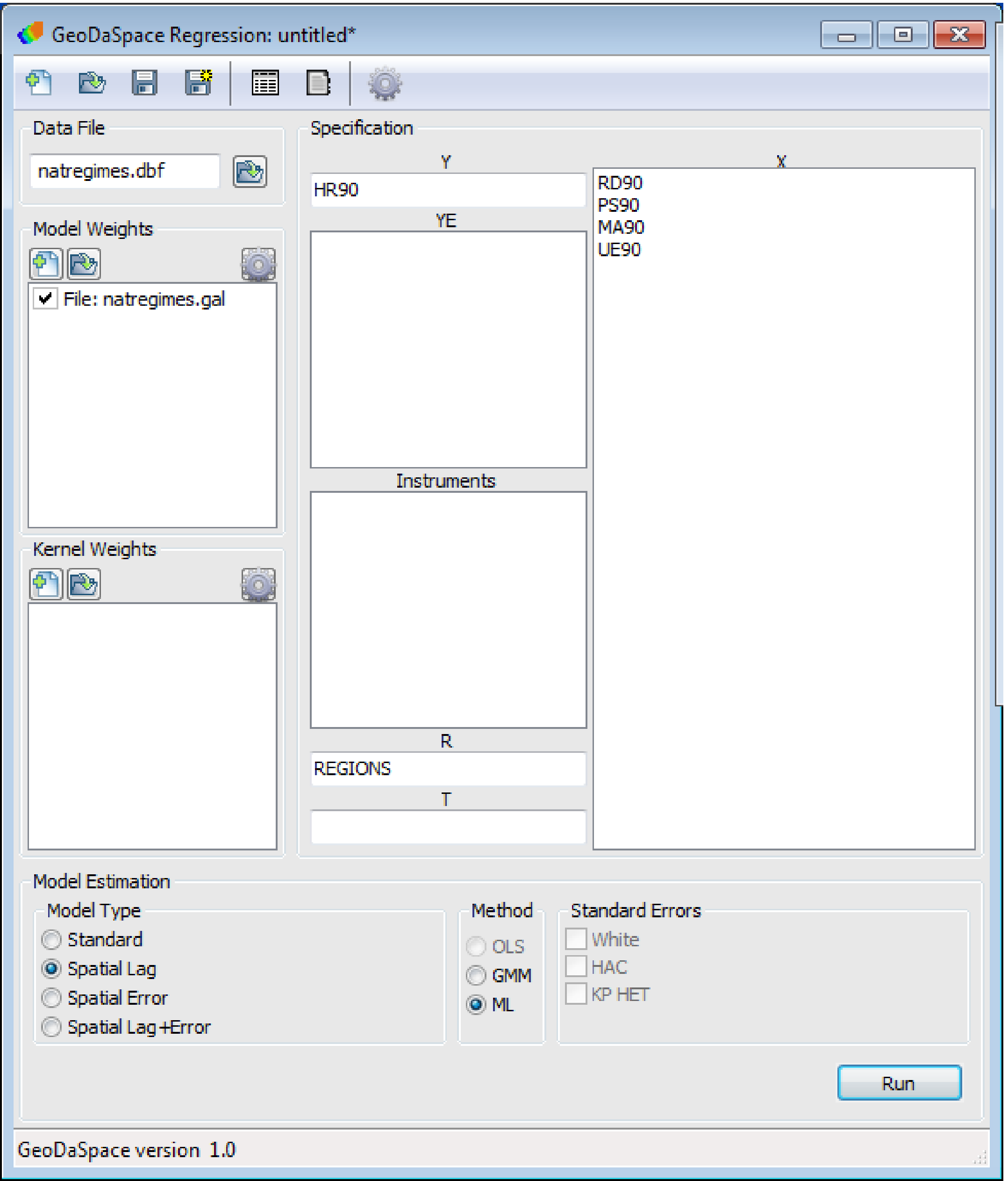
```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES
-----------------------------------------
Data set              :natregimes.dbf
Weights matrix        :File: natregimes.gal
Dependent Variable    :        HR90              Number of Observations:     3085
Mean dependent var    :      6.1829              Number of Variables   :        5
S.D. dependent var    :      6.6414              Degrees of Freedom    :     3080
R-squared             :      0.3951
Adjusted R-squared    :      0.3943
Sum squared residual:     82288.591              F-statistic           :  502.8748
Sigma-square          :      26.717              Prob(F-statistic)     :        0
S.E. of regression    :       5.169              Log likelihood        : -9442.503
Sigma-square ML       :      26.674              Akaike info criterion : 18895.006
S.E of regression ML:        5.1647              Schwarz criterion     : 18925.178

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     t-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT       9.9009950       1.0334618       9.5804168       0.0000000
                MA90      -0.0517906       0.0281109      -1.8423692       0.0655172
                PS90       1.7109846       0.1005362      17.0185857       0.0000000
                RD90       4.5446821       0.1229004      36.9785919       0.0000000
                UE90      -0.2912908       0.0397067      -7.3360647       0.0000000
------------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER           25.660

TEST ON NORMALITY OF ERRORS
TEST                             DF        VALUE           PROB
Jarque-Bera                       2      54094.592         0.0000

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                             DF        VALUE           PROB
Breusch-Pagan test                4       1276.655         0.0000
Koenker-Bassett test              4        116.736         0.0000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST                           MI/DF       VALUE           PROB
Lagrange Multiplier (lag)         1        198.183         0.0000
Robust LM (lag)                   1         18.580         0.0000
Lagrange Multiplier (error)       1        199.733         0.0000
Robust LM (error)                 1         20.129         0.0000
Lagrange Multiplier (SARMA)       2        218.313         0.0000

================================ END OF REPORT =====================================
```

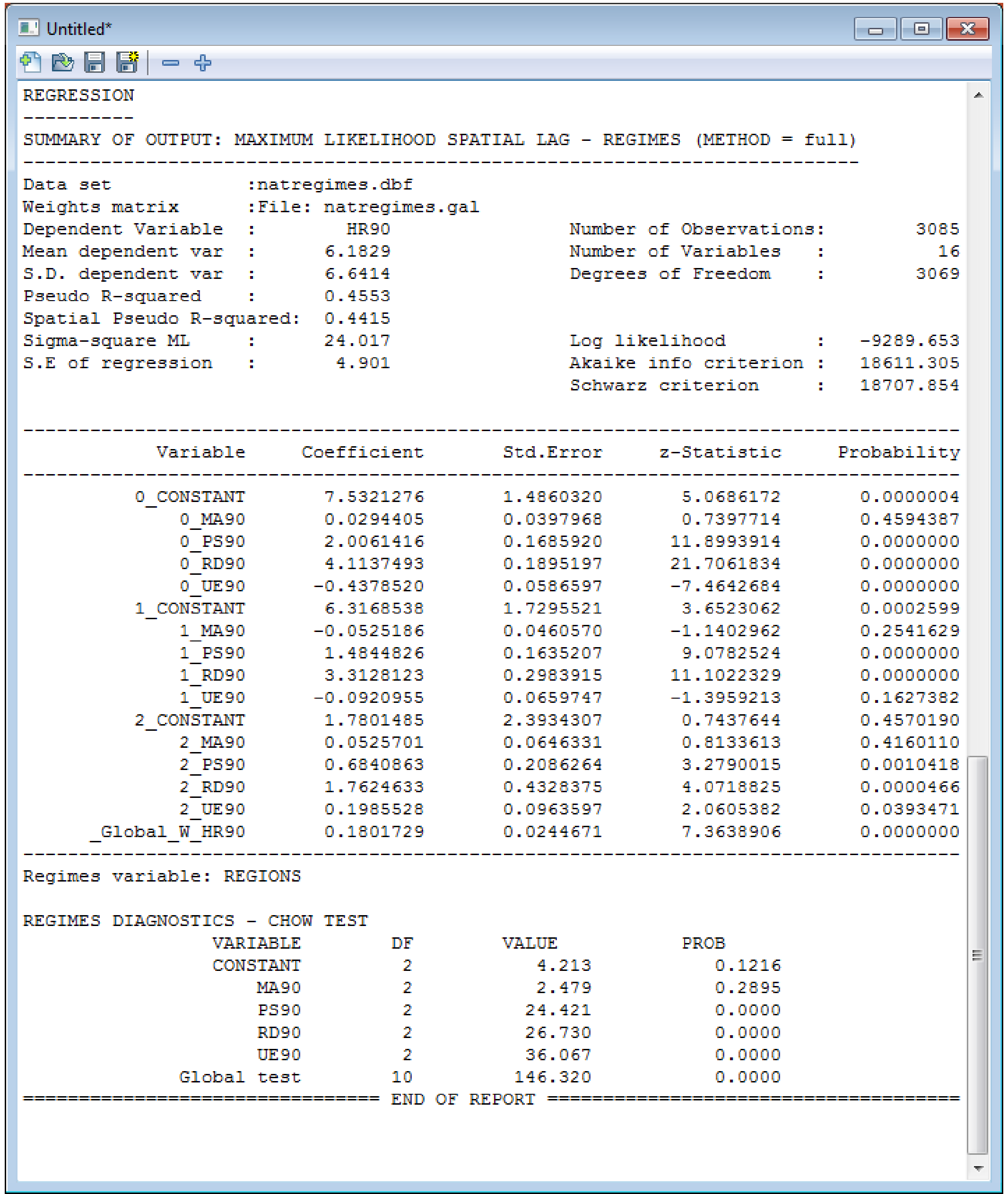
Now I will build a Spatial Lag Model – MA90 is no longer significant




```
REGRESSION
----------
SUMMARY OF OUTPUT: MAXIMUM LIKELIHOOD SPATIAL LAG (METHOD = FULL)
-----------------------------------------------------------------
Data set            :natregimes.dbf
Weights matrix      :File: natregimes.gal
Dependent Variable  :        HR90                Number of Observations:        3085
Mean dependent var  :      6.1829                Number of Variables   :           6
S.D. dependent var  :      6.6414                Degrees of Freedom    :        3079
Pseudo R-squared    :      0.4349
Spatial Pseudo R-squared:  0.3982
Sigma-square ML     :      24.927                Log likelihood        :   -9360.236
S.E of regression   :       4.993                Akaike info criterion :   18732.473
                                                 Schwarz criterion     :   18768.679

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     z-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT       6.4739021       1.0145427       6.3811034       0.0000000
                MA90      -0.0203630       0.0271588      -0.7497753       0.4533901
                PS90       1.4365892       0.0994735      14.4419316       0.0000000
                RD90       3.6242501       0.1389813      26.0772478       0.0000000
                UE90      -0.1986542       0.0385328      -5.1554564       0.0000003
              W_HR90       0.2788219       0.0220929      12.6204172       0.0000000
------------------------------------------------------------------------------------
================================ END OF REPORT =====================================
```

Now I will build a spatial regime model





```
REGRESSION
----------
SUMMARY OF OUTPUT: MAXIMUM LIKELIHOOD SPATIAL LAG - REGIMES (METHOD = full)
----------------------------------------------------------------------------
Data set            :natregimes.dbf
Weights matrix      :File: natregimes.gal
Dependent Variable  :        HR90                Number of Observations:      3085
Mean dependent var  :      6.1829                Number of Variables   :        16
S.D. dependent var  :      6.6414                Degrees of Freedom    :      3069
Pseudo R-squared    :      0.4553
Spatial Pseudo R-squared:  0.4415
Sigma-square ML     :      24.017                Log likelihood        :   -9289.653
S.E of regression   :       4.901                Akaike info criterion :   18611.305
                                                 Schwarz criterion     :   18707.854

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     z-Statistic     Probability
------------------------------------------------------------------------------------
          0_CONSTANT       7.5321276       1.4860320       5.0686172       0.0000004
              0_MA90       0.0294405       0.0397968       0.7397714       0.4594387
              0_PS90       2.0061416       0.1685920      11.8993914       0.0000000
              0_RD90       4.1137493       0.1895197      21.7061834       0.0000000
              0_UE90      -0.4378520       0.0586597      -7.4642684       0.0000000
          1_CONSTANT       6.3168538       1.7295521       3.6523062       0.0002599
              1_MA90      -0.0525186       0.0460570      -1.1402962       0.2541629
              1_PS90       1.4844826       0.1635207       9.0782524       0.0000000
              1_RD90       3.3128123       0.2983915      11.1022329       0.0000000
              1_UE90      -0.0920955       0.0659747      -1.3959213       0.1627382
          2_CONSTANT       1.7801485       2.3934307       0.7437644       0.4570190
              2_MA90       0.0525701       0.0646331       0.8133613       0.4160110
              2_PS90       0.6840863       0.2086264       3.2790015       0.0010418
              2_RD90       1.7624633       0.4328375       4.0718825       0.0000466
              2_UE90       0.1985528       0.0963597       2.0605382       0.0393471
      _Global_W_HR90       0.1801729       0.0244671       7.3638906       0.0000000
------------------------------------------------------------------------------------
Regimes variable: REGIONS

REGIMES DIAGNOSTICS - CHOW TEST
                 VARIABLE        DF        VALUE           PROB
                 CONSTANT         2        4.213           0.1216
                     MA90         2        2.479           0.2895
                     PS90         2       24.421           0.0000
                     RD90         2       26.730           0.0000
                     UE90         2       36.067           0.0000
              Global test        10      146.320           0.0000
================================ END OF REPORT =====================================
```