Module 8
Geographically Weighted Regression in GeoDa

Regression is the statistical dependence of the mean value of a random variable on the values of another random variable (or several random variables).

Geographically weighted regression (GWR) helps explain certain variables using other variables with due account of the spatial effect. In other words, GWR tells us the degree to which the spatial effect strengthens or weakens the regression. The goal of GWR is to find a model that reliably explains the dependent variable while taking the neighbourhood effect into account.

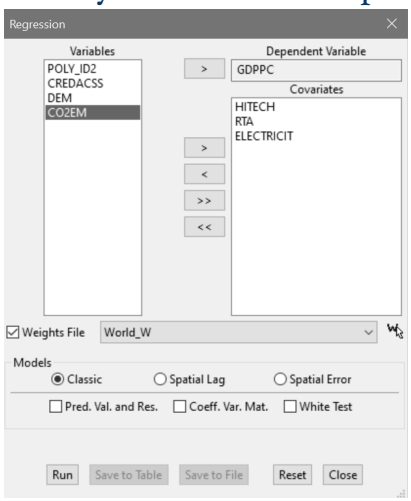
The spatial lag model (SAR) and the spatial error model (SEM) are two main GWR models.

For this task we will use the "WorldR" archive available on our Google Drive (https://drive.google.com/file/d/1iY2ruAQH_41qPjdwZihXUiyy3k9LP8gF/view?usp=drive_link) and our website (https://mgimo.ru/upload/2023/05/worldr-geographically-weighted-regression.rar).

Open the WorldR .shp file in GeoDa, and load the World_W weights. The attribute table contains a set of seven variables: GDPPC – GDP per capita; CREDACCS – number of credit cards per 1000 people; CO2EM – CO2 emissions (metric tons per capita); DEM – Democracy Index; ELECTRICIT – energy supply; RTA – participation in regional trade agreements; HITECH – the share of medium- and hi-tech industries in industrial production.

Let's find out whether the GDPPC variable can be explained by a model consisting of three explanatory variables – HITECH, ELECTRICIT and RTA. This will allow us to determine the extent to which a country's GDP per capita determines this particular set of variables.

1) Select Regression in the toolbar and set the following parameters in the dialog box:
Dependent Variable: GDPPC
Covariates (independent variables): HITECH, RTA, ELECTRICIT
Weights File: World_W
Models: Classic
Once you have set the parameters, click Run to get the regression report.



When you have the regression report, run a few tests to determine whether the model is statistically correct.
1. p-value (probability) represents the level of statistical significance of the given value: the greater the p, the lower the level of significance. The maximum allowable p-value in social sciences is 0.05.
All p-values other than CONSTANT must satisfy this condition. The p-value in our case is below this threshold, so Test 1 passed.
The p-value of the diagnostics for heteroskedasticity (BP test or KB test) should also be under 0.05, while the p-value for the test on normality of errors (Jarque–Bera) should be higher than 0.1 (i.e. there should be no such normality), although this test is only valid for big datasets and can thus be disregarded for now.

```
РЕГРЕССИЯ
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set            :  v22_2608
Dependent Variable  :       GDPPC  Number of Observations:  144
Mean dependent var  :     23593,6  Number of Variables   :    4
S.D. dependent var  :     22068,2  Degrees of Freedom    :  140

R-squared           :    0,654078  F-statistic           :     88,2385
Adjusted R-squared  :    0,646665  Prob(F-statistic)     :4,12296e-032
Sum squared residual:2,42592e+010  Log likelihood        :    -1568,17
Sigma-square        : 1,7328e+008  Akaike info criterion :     3144,34
S.E. of regression  :     13163,6  Schwarz criterion     :     3156,22
Sigma-square ML     :1,68467e+008
S.E of regression ML:     12979,5

--------------------------------------------------------------------------------
       Variable      Coefficient      Std.Error    t-Statistic   Probability
--------------------------------------------------------------------------------
          CONSTANT        1045,6        1919,74       0,544658     0,58685
            HITECH       425,992        81,4701        5,22881     0,00000
               RTA        339,21         86,935        3,90188     0,00015
        ELECTRICIT       1,54489        0,18615        8,29918     0,00000
--------------------------------------------------------------------------------
```

2. R squared (the coefficient of determination) is the proportion of the variation explained by the model. It is generally accepted that a model with an R squared (R^2) of less than 0.5 cannot be used, and that above 0.8 is best. In this case, the R^2 is 0.65, which makes the model usable. This means that the set of variables explains the value of GDP per capita by 65%.

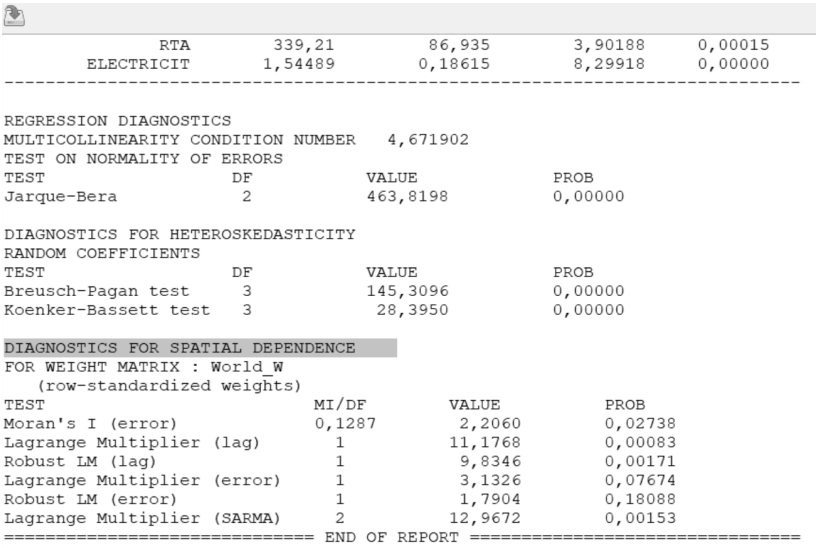

```
РЕГРЕССИЯ
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set            :  v22_2608
Dependent Variable  :       GDPPC  Number of Observations:  144
Mean dependent var  :     23593,6  Number of Variables   :    4
S.D. dependent var  :     22068,2  Degrees of Freedom    :  140

R-squared           :    0,654078  F-statistic           :     88,2385
Adjusted R-squared  :    0,646665  Prob(F-statistic)     :4,12296e-032
Sum squared residual:2,42592e+010  Log likelihood        :    -1568,17
Sigma-square        : 1,7328e+008  Akaike info criterion :     3144,34
S.E. of regression  :     13163,6  Schwarz criterion     :     3156,22
Sigma-square ML     :1,68467e+008
S.E of regression ML:     12979,5

--------------------------------------------------------------------------------
       Variable      Coefficient      Std.Error    t-Statistic   Probability
--------------------------------------------------------------------------------
          CONSTANT        1045,6        1919,74       0,544658     0,58685
            HITECH       425,992        81,4701        5,22881     0,00000
               RTA        339,21         86,935        3,90188     0,00015
        ELECTRICIT       1,54489        0,18615        8,29918     0,00000
--------------------------------------------------------------------------------
```


3. You should also pay attention to the Akaike information criterion (AIC): the lower the AIC, the more preferable the model. While there is no upper limit to the AIC, it can nevertheless be used when comparing two linear regressions.

The SAR and SEM models are not always equally appropriate, and spatial tests are carried out to determine the acceptable model using the Lagrange Multiplier (LM), which estimates the level of spatial autocorrelation of the

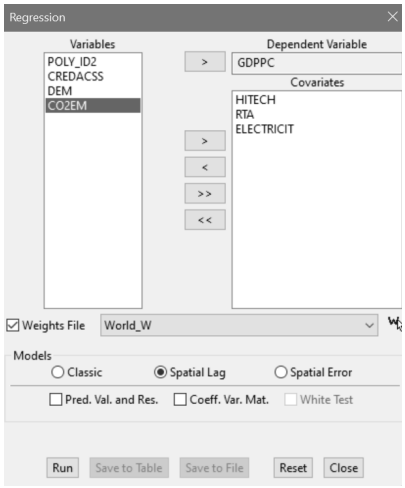
dependent variable (for the spatial lag model) or the ***residuals*** (for the spatial error model).

Take a look at the last column in the section "Diagnostics for Spatial Dependence," titled PROB (p-value), and choose the model – Lagrange Multiplier (lag) or Lagrange Multiplier (error) – for which the p-value is less than or equal to 0.05. If both models pass this test, then do the same with Robust LM (lag) and Robust LM (error). If the models pass this test as well, then choose the model with the lowest p-value. In this case, the Lagrange Multiplier (error) is above 0.05, meaning that the Lagrange Multiplier (lag) is the only usable model here.



```
             RTA          339,21          86,935         3,90188      0,00015
      ELECTRICIT        1,54489         0,18615         8,29918      0,00000
------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER    4,671902
TEST ON NORMALITY OF ERRORS
TEST                    DF           VALUE             PROB
Jarque-Bera              2           463,8198          0,00000

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                    DF           VALUE             PROB
Breusch-Pagan test       3           145,3096          0,00000
Koenker-Bassett test     3            28,3950          0,00000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : World_W
   (row-standardized weights)
TEST                           MI/DF        VALUE          PROB
Moran's I (error)              0,1287        2,2060        0,02738
Lagrange Multiplier (lag)           1       11,1768        0,00083
Robust LM (lag)                     1        9,8346        0,00171
Lagrange Multiplier (error)         1        3,1326        0,07674
Robust LM (error)                   1        1,7904        0,18088
Lagrange Multiplier (SARMA)         2       12,9672        0,00153
============================== END OF REPORT ================================
```

To calculate GWR follow the same steps as with linear regression, but choose the Spatial Lag model, as it is the only valid model for this particular set of variables.



Once you have your regression report, look at the p-values of the variables, the Diagnostics for Spatial Dependence, and the Diagnostics for Heteroskedasticity (BP test).

Note that a new variable has appeared (compared to linear regression) – W_GDPPC, which is the spatial lag of the dependent variable we included in the regression model (in the spatial error model, this will be the LAMBDA variable, reflecting the spatial autocorrelation of the ***residuals***).

In each of these cases, the p-value should not be greater than 0.05. The next thing you need to do is look at the R^2 and the AIC, which we have already covered, as well as at the Schwarz information criterion.

```
РЕГРЕССИЯ
----------
SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : v22_2608
Spatial Weight      : World_W
Dependent Variable  :      GDPPC  Number of Observations:  144
Mean dependent var  :    23593,6  Number of Variables   :    5
S.D. dependent var  :    22068,2  Degrees of Freedom    :  139
Lag coeff.   (Rho)  :   0,315804

R-squared           :   0,699586  Log likelihood        :   -1560,29
Sq. Correlation     : -           Akaike info criterion :    3130,57
Sigma-square        :1,46304e+008 Schwarz criterion     :    3145,42
S.E of regression   :    12095,6

------------------------------------------------------------------------------
      Variable       Coefficient     Std.Error       z-value    Probability
------------------------------------------------------------------------------
          W_GDPPC      0,315804      0,0707651       4,46271     0,00001
         CONSTANT       -1354,9        1880,41     -0,720538     0,47119
           HITECH       343,969        76,5332       4,49438     0,00001
              RTA       242,665        85,2539       2,84638     0,00442
       ELECTRICIT       1,19979       0,179967       6,66675     0,00000
------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                DF      VALUE        PROB
Breusch-Pagan test                  3       131,2247     0,00000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL LAG DEPENDENCE FOR WEIGHT MATRIX : World_W
TEST                                DF      VALUE        PROB
Likelihood Ratio Test               1        15,7634     0,00007
============================== END OF REPORT ================================
```

Just like with the AIC, there is no upper limit for Schwarz, and the optimal model is the one that shows the lowest values for this criterion. As the GWR report shows, the p-value tests have been passed, and the R^2 is sufficiently high, meaning that the model can be used as an explanatory model.

There are a number of ways to compare regression models: the Akaike information criterion can be used to compare all kinds of models – for example, non-spatial and spatial models, non-spatial models with other non-spatial models, and spatial models with other spatial models (and, in turn, lags and errors, lags with other lags, and errors with other errors).

The Schwarz information criterion only allows you to compare spatial models with other spatial models. This makes it more accurate for comparing spatial models than the Akaike information criterion. The regression report for the lag model produces an R^2 of 0.69, which also allows this set of variables to be used as an explanatory variable.

To compare these two models, we need to use the Akaike information criterion, which shows that the GWR turned out to be slightly more accurate than the linear regression model, with an acceptable R^2 value in both.

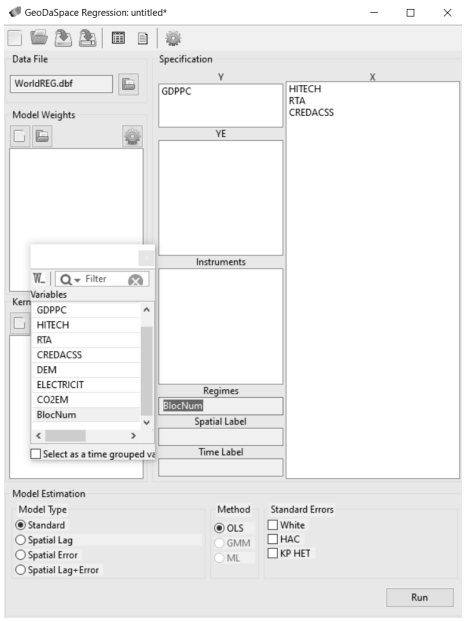
GeoDa Space (available for download at https://geodacenter.github.io/GeoDaSpace/download.html) allows you to perform regression analysis within individual groups of objects, for example, regions of the world.
1) To analyse an entire set of objects, from the start window, select a .dbf file in the Data File field, set the dependent variable (in this case, GDPPC) in the Y field, and explanatory variables (HITECH, RTA, ELECTRICT) in the X field. Select which regression model you need in the Model Type window.



The resulting report will be identical to those discussed earlier.

```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES
-----------------------------------------
Data set            :        1.dbf
Weights matrix      : File: 1.gal
Dependent Variable  :      Clergy                Number of Observations:          85
Mean dependent var  :     43.9294                Number of Variables   :           3
S.D. dependent var  :     24.7130                Degrees of Freedom    :          82
R-squared           :      0.1067
Adjusted R-squared  :      0.0849
Sum squared residual:   45829.901                F-statistic           :      4.8950
Sigma-square        :     558.901                Prob(F-statistic)     :    0.009812
S.E. of regression  :      23.641                Log likelihood        :    -387.937
Sigma-square ML     :     539.175                Akaike info criterion :     781.873
S.E of regression ML:     23.2201                Schwarz criterion     :     789.201

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     t-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT      50.8084117       6.2691545       8.1045078       0.0000000
             Infants       0.0001476       0.0003044       0.4848826       0.6290525
             Suicids      -0.0002651       0.0000855      -3.0995030       0.0026560
------------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER            5.325

TEST ON NORMALITY OF ERRORS
TEST                             DF        VALUE           PROB
Jarque-Bera                       2          3.215          0.2004

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                             DF        VALUE           PROB
Breusch-Pagan test                2          0.507          0.7760
Koenker-Bassett test              2          0.968          0.6164

DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST                           MI/DF       VALUE           PROB
Lagrange Multiplier (lag)         1         26.676          0.0000
Robust LM (lag)                   1          2.885          0.0894
Lagrange Multiplier (error)       1         23.878          0.0000
Robust LM (error)                 1          0.088          0.7673
Lagrange Multiplier (SARMA)       2         26.763          0.0000

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

2) The spatial regimes function allows you to analyse individual groups of objects. To do that, you need to select a variable in the Regime field that marks an observation as belonging to a group (in this case, select BlocNum – a numerical value that corresponds to one of the geopolitical blocs of the world).



The resulting report will be presented both for the entire set of objects and for each group (Regime) individually.

```
REGRESSION
----------

SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION - REGIME 1
-----------------------------------------------------------------
Data set            :       1.dbf
Weights matrix      : File: 1.gal
Dependent Variable  :    1_Clergy                Number of Observations:          17
Mean dependent var  :     34.0588                Number of Variables   :           3
S.D. dependent var  :     19.0475                Degrees of Freedom    :          14
R-squared           :      0.1560
Adjusted R-squared  :      0.0354
Sum squared residual:    4899.649                F-statistic           :      1.2934
Sigma-square        :     349.975                Prob(F-statistic)     :      0.3052
S.E. of regression  :      18.708                Log likelihood        :     -72.263
Sigma-square ML     :     288.215                Akaike info criterion :     150.527
S.E of regression ML:     16.9769                Schwarz criterion     :     153.027

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     t-Statistic     Probability
------------------------------------------------------------------------------------
          1_CONSTANT      54.0802479      17.5090850       3.0886964       0.0080102
           1_Infants      -0.0011902       0.0007401      -1.6082290       0.1300962
           1_Suicids      -0.0000024       0.0004748      -0.0051597       0.9959560
------------------------------------------------------------------------------------
Regimes variable: CATEGORIES

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER           7.985

TEST ON NORMALITY OF ERRORS
TEST                             DF        VALUE           PROB
Jarque-Bera                       2          1.723          0.4225

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                             DF        VALUE           PROB
Breusch-Pagan test                2          0.644          0.7248
Koenker-Bassett test              2          1.556          0.4593

DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST                           MI/DF       VALUE           PROB
Lagrange Multiplier (lag)         1          0.949          0.3300
```