03_model

February 4, 2019

1 *Model* spatial data

In [1]:
```
%matplotlib inline

        import pandas as pd
        import geopandas as gpd
        from pysal.lib import weights
        from pysal.model import spreg
        from pysal.model.mgwr.gwr import GWR
        from pysal.model.mgwr.sel_bw import Sel_BW

        db = gpd.read_file('../data/demo_data.gpkg')
        w = weights.Queen.from_dataframe(db)
        w.transform = 'R'
```

```
/opt/conda/lib/python3.6/site-packages/pysal/model/spvcm/abstracts.py:10: UserWarning: The `dill
  from .sqlite import head_to_sql, start_sql
```

1.1 Non-spatial regression (OLS)

$$y_i = X_i\beta + \epsilon_i$$

In [2]:
```
y = 'arturo_score'
        x = ['land_use_mix', 'closeness_small_parks',
             'residence_ratio', 'age_diversity']

        ols = spreg.OLS(db[[y]].values,
                        db[x].values,
                        w=w,
                        name_y=y,
                        name_x=x,
                        spat_diag=True)

        ols
```

Out[2]:
```
<pysal.model.spreg.ols.OLS at 0x7fcc6f7503c8>
```

```
In [3]: print(ols.summary)

REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES
-----------------------------------------
Data set            :     unknown
Weights matrix      :     unknown
Dependent Variable  :arturo_score                Number of Observations:         128
Mean dependent var  :     24.8409                Number of Variables   :           5
S.D. dependent var  :      2.4354                Degrees of Freedom    :         123
R-squared           :      0.3087
Adjusted R-squared  :      0.2862
Sum squared residual:     520.758                F-statistic           :     13.7286
Sigma-square        :       4.234                Prob(F-statistic)     :   2.766e-09
S.E. of regression  :       2.058                Log likelihood        :    -271.432
Sigma-square ML     :       4.068                Akaike info criterion :     552.865
S.E of regression ML:      2.0170                Schwarz criterion     :     567.125

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     t-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT      21.2746322       1.0339182      20.5767082       0.0000000
        land_use_mix   -6289.8511551    1995.5131280      -3.1519969       0.0020372
closeness_small_parks      -0.0008897       0.0018640      -0.4772935       0.6340002
     residence_ratio       4.1686706       1.7333574       2.4049688       0.0176641
       age_diversity    1678.1725955     332.9280823       5.0406460       0.0000016
------------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER           16.367

TEST ON NORMALITY OF ERRORS
TEST                             DF        VALUE           PROB
Jarque-Bera                       2           4.649           0.0978

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                             DF        VALUE           PROB
Breusch-Pagan test                4           6.934           0.1394
Koenker-Bassett test              4           5.420           0.2468

DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST                           MI/DF       VALUE           PROB
Lagrange Multiplier (lag)         1          27.890           0.0000
Robust LM (lag)                   1          11.875           0.0006
Lagrange Multiplier (error)       1          17.274           0.0000
Robust LM (error)                 1           1.259           0.2619
```

```
Lagrange Multiplier (SARMA)       2          29.149           0.0000

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

1.2 Spatial Lag

$$y_i = \sum_j w_{ij} y_j + X_i \beta + \epsilon_i$$

```
In [4]: slm = spreg.ML_Lag(db[[y]].values,
                           db[x].values,
                           w=w,
                           name_y=y,
                           name_x=x)

        print(slm.summary)
```

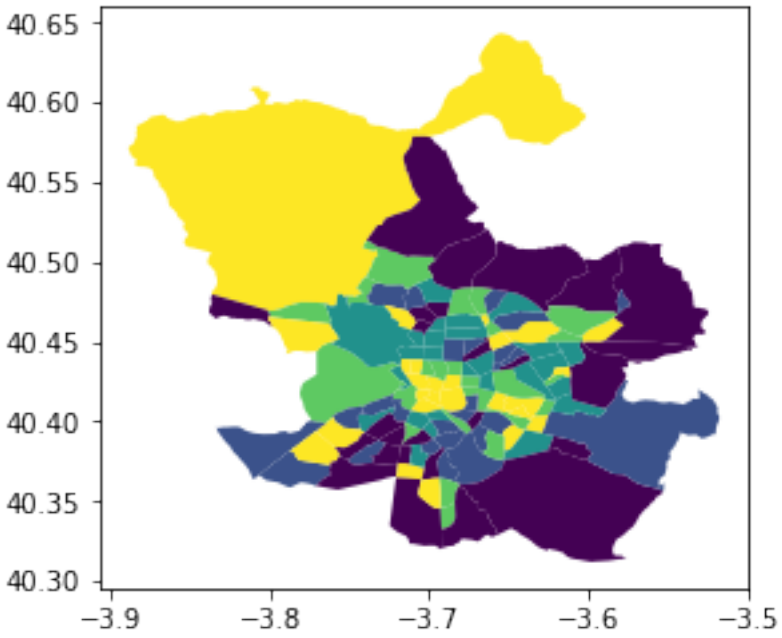
```
REGRESSION
----------
SUMMARY OF OUTPUT: MAXIMUM LIKELIHOOD SPATIAL LAG (METHOD = FULL)
-----------------------------------------------------------------
Data set            :     unknown
Weights matrix      :     unknown
Dependent Variable  :arturo_score                Number of Observations:         128
Mean dependent var  :     24.8409                Number of Variables   :           6
S.D. dependent var  :      2.4354                Degrees of Freedom    :         122
Pseudo R-squared    :      0.4608
Spatial Pseudo R-squared:  0.3602
Sigma-square ML     :       3.193                Log likelihood        :    -259.557
S.E of regression   :       1.787                Akaike info criterion :     531.114
                                                 Schwarz criterion     :     548.226

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     z-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT       8.6975172       2.4024646       3.6202478       0.0002943
        land_use_mix    -4443.8788366    1753.7192835      -2.5339739       0.0112777
closeness_small_parks      -0.0003619       0.0016207      -0.2233116       0.8232930
     residence_ratio       3.9356923       1.5139087       2.5996893       0.0093308
       age_diversity     865.2543040     307.1346045       2.8171827       0.0048447
      W_arturo_score       0.5416816       0.0971072       5.5781846       0.0000000
------------------------------------------------------------------------------------
================================ END OF REPORT =====================================


/opt/conda/lib/python3.6/site-packages/scipy/optimize/_minimize.py:761: RuntimeWarning: Method '
  "defaulting to absolute tolerance.", RuntimeWarning)
```

1.3 Spatial regimes

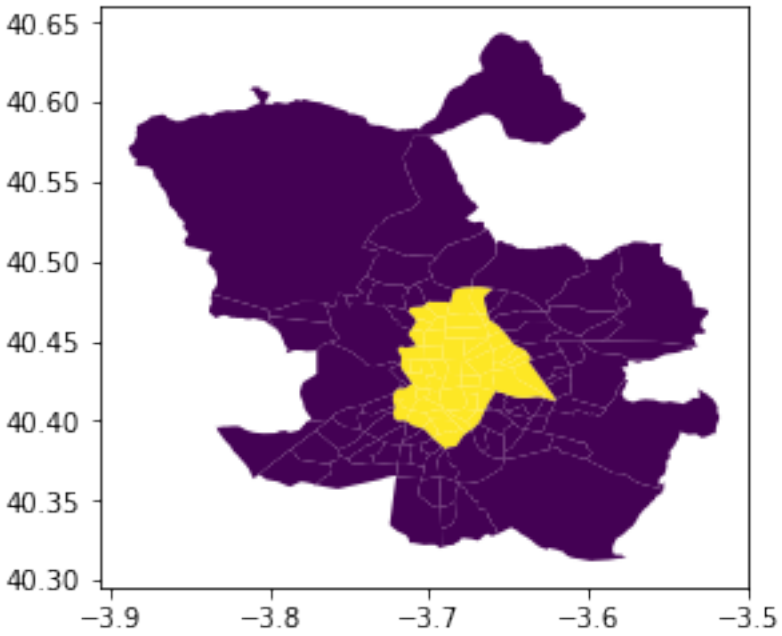
$$y_i = X_i \beta_r + \epsilon_i$$

```
In [5]: db.assign(resid=ols.u)\
          .plot(column='resid', scheme='quantiles')
```

```
Out[5]: <matplotlib.axes._subplots.AxesSubplot at 0x7fcc448ebd68>
```




```
In [6]: centro = ['Tetuán', 'Chamartín', 'Ciudad Lineal',
                 'Salamanca', 'Chamberí', 'Centro', 'Arganzuela',
                 'Retiro']
        db['centro'] = 0
        db.loc[db['neighbourhood_group'].isin(centro), 'centro'] = 1

        db.plot('centro');
```




```
In [7]: ols_r = spreg.OLS_Regimes(db[[y]].values,
                                  db[x].values,
                                  db['centro'].values,
                                  w=w,
                                  name_y=y,
                                  name_x=x)

        print(ols_r.summary)
```

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

SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION - REGIME 0
-----------------------------------------------------------------
Data set            :     unknown
Weights matrix      :     unknown
Dependent Variable  :0_arturo_score                Number of Observations:          76
Mean dependent var  :     23.8474                Number of Variables   :           5
S.D. dependent var  :      2.0900                Degrees of Freedom    :          71
R-squared           :      0.2446
Adjusted R-squared  :      0.2020
Sum squared residual:     247.464                F-statistic           :      5.7476
Sigma-square        :       3.485                Prob(F-statistic)     :   0.0004584
S.E. of regression  :       1.867                Log likelihood        :    -152.699
Sigma-square ML     :       3.256                Akaike info criterion :     315.399
S.E of regression ML:      1.8045                Schwarz criterion     :     327.053
```

```
------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     t-Statistic     Probability
------------------------------------------------------------------------------
          0_CONSTANT      21.0656224       1.3127447      16.0470062       0.0000000
      0_land_use_mix    -8820.2629643    2958.1366539      -2.9816956       0.0039250
0_closeness_small_parks      -0.0028580       0.0024126      -1.1846180       0.2401189
   0_residence_ratio       6.9625674       2.1490206       3.2398793       0.0018209
     0_age_diversity    1050.6023704     459.9046757       2.2843916       0.0253407
------------------------------------------------------------------------------
Regimes variable: unknown

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER           16.676

TEST ON NORMALITY OF ERRORS
TEST                             DF        VALUE           PROB
Jarque-Bera                       2           0.507           0.7761

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                             DF        VALUE           PROB
Breusch-Pagan test                4           7.585           0.1080
Koenker-Bassett test              4           7.604           0.1072
----------

SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION - REGIME 1
-----------------------------------------------------------
Data set            :     unknown
Weights matrix      :     unknown
Dependent Variable  :1_arturo_score                Number of Observations:          52
Mean dependent var  :     26.2930               Number of Variables   :           5
S.D. dependent var  :      2.1738               Degrees of Freedom    :          47
R-squared           :      0.2255
Adjusted R-squared  :      0.1596
Sum squared residual:     186.662               F-statistic           :      3.4206
Sigma-square        :       3.972               Prob(F-statistic)     :     0.01554
S.E. of regression  :       1.993               Log likelihood        :    -107.014
Sigma-square ML     :       3.590               Akaike info criterion :     224.029
S.E of regression ML:      1.8946               Schwarz criterion     :     233.785

------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     t-Statistic     Probability
------------------------------------------------------------------------------
          1_CONSTANT      26.2497879       2.0419239      12.8554196       0.0000000
      1_land_use_mix    -3749.2772633    2529.1418521      -1.4824306       0.1449003
1_closeness_small_parks       0.0044967       0.0028221       1.5933780       0.1177792
   1_residence_ratio      -4.7230823       3.0240452      -1.5618425       0.1250341
```

```
     1_age_diversity    1491.8201792     562.6179389       2.6515688       0.0108874
--------------------------------------------------------------------------------
Regimes variable: unknown

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER           21.766

TEST ON NORMALITY OF ERRORS
TEST                             DF        VALUE           PROB
Jarque-Bera                       2            7.624          0.0221

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                             DF        VALUE           PROB
Breusch-Pagan test                4           20.819          0.0003
Koenker-Bassett test              4           11.749          0.0193

REGIMES DIAGNOSTICS - CHOW TEST
                 VARIABLE        DF        VALUE           PROB
                 CONSTANT         1            4.561          0.0327
            age_diversity         1            0.369          0.5437
    closeness_small_parks         1            3.924          0.0476
             land_use_mix         1            1.698          0.1926
          residence_ratio         1            9.922          0.0016
              Global test         5           22.923          0.0003
================================= END OF REPORT =====================================
```

1.4 Geographically Weighted Regression

$$y_i = X_i \beta_i + \epsilon_i$$

- Set up

```
In [8]: ptX = db.centroid.x
        ptY = db.centroid.y
        coords = pd.DataFrame({'X': ptX,
                               'Y': ptY}).values
```

- Bandwith selection

```
In [9]: selector = Sel_BW(coords,
                         db[[y]].values,
                         db[x].values)
        bw = selector.search()
```

- Fit

```
In [10]: gwr = GWR(coords,
                   db[[y]].values,
                   db[x].values,
                   bw)\
                .fit()
```

- Visualise results for the intercept

```
In [11]: ax = db.assign(c=gwr.params[:, 0])\
                .plot('c', scheme='quantiles', k=12,
                      figsize=(9, 9))
         ax.set_title('Spatial heterogeneity of Intercept');
```