Data Science Training Webinars
Spatiotemporal Simulation

The Workbench for Spatial Data Science

Hanchen Yu

Center for Geographic Analysis
Harvard University

June 25, 2022

Center for Geographic Analysis
Harvard University

Geo-computation Center for Social Sciences
Wuhan University

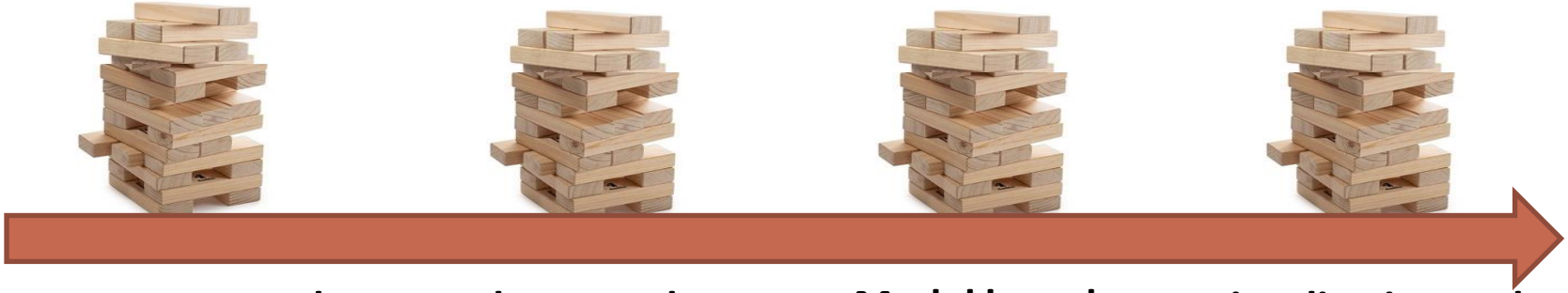
China Data Institute

Future Data Lab

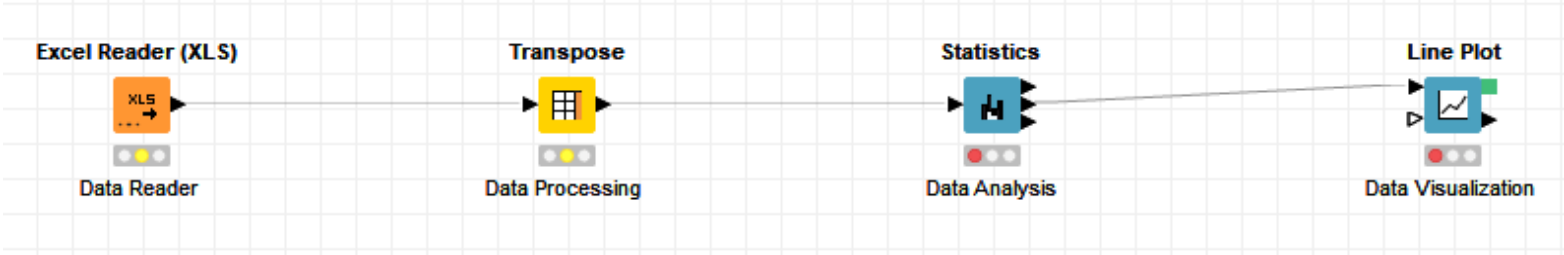
What is the Workbench for Spatial Data Science (WSDS)?

A **workbench** is a collection of nodes and workflows, which can be used to build different workflow projects for teaching, research and business applications.

Features： Assemblable, Reproducible, Replicable, Expandable, Shareable



Data access and processing | **Exploratory data analysis** | **Model based analysis** | **Visualization and diagnostic**



Goals and Objectives

This project will explore methodologies and establish protocols for developing workbenches for spatial data science in research, teaching and business applications.

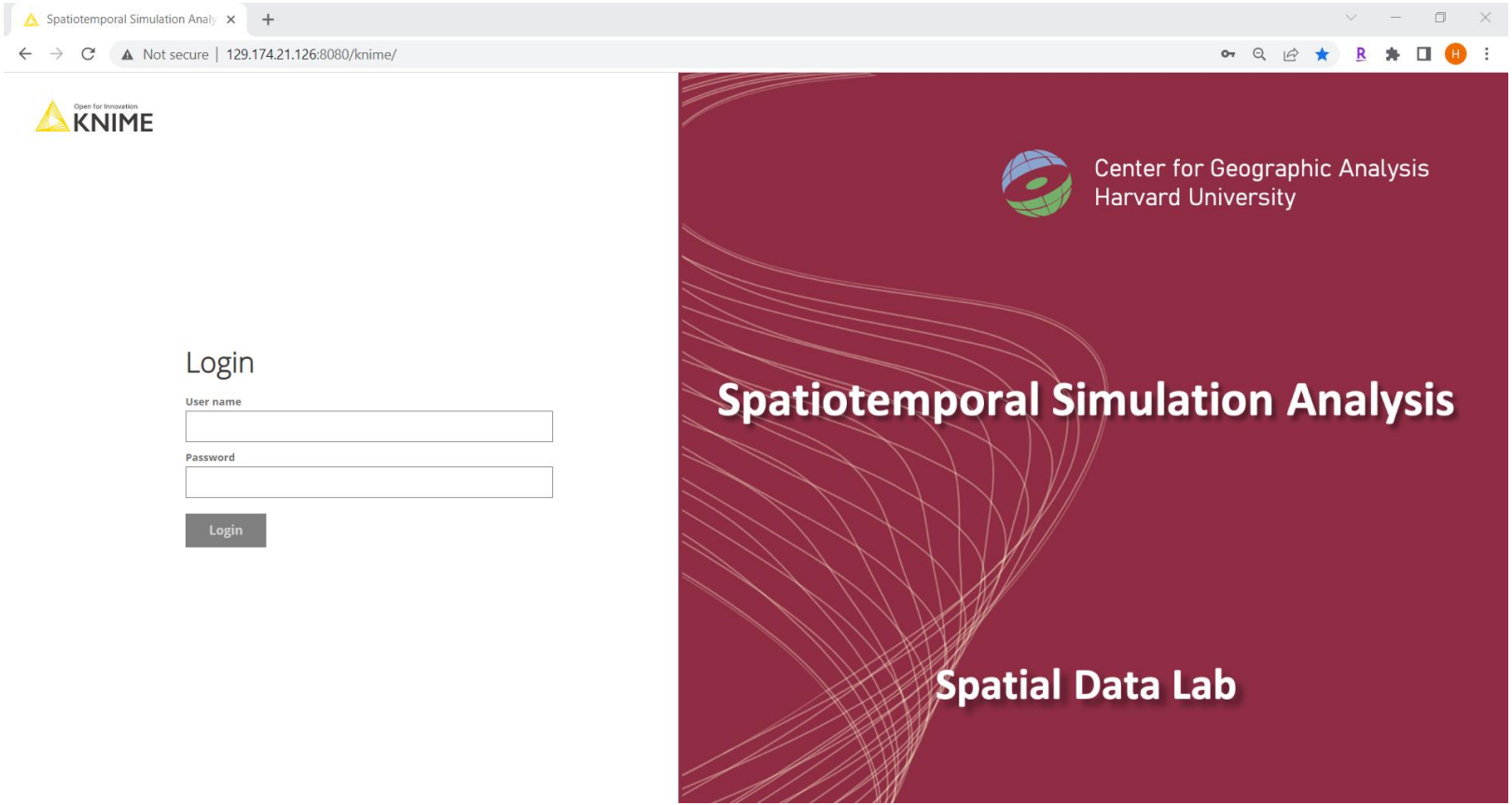
The objectives of this project are to provide:

1. An easy, efficient and customizable toolkit for spatial data analysis with new added nodes.
2. An integration of data, methodology and applications for spatial data science.
3. Workflow-based case studies for teaching and research in spatial social science.
4. A training base for users with no skills in GIS and advanced methodology.

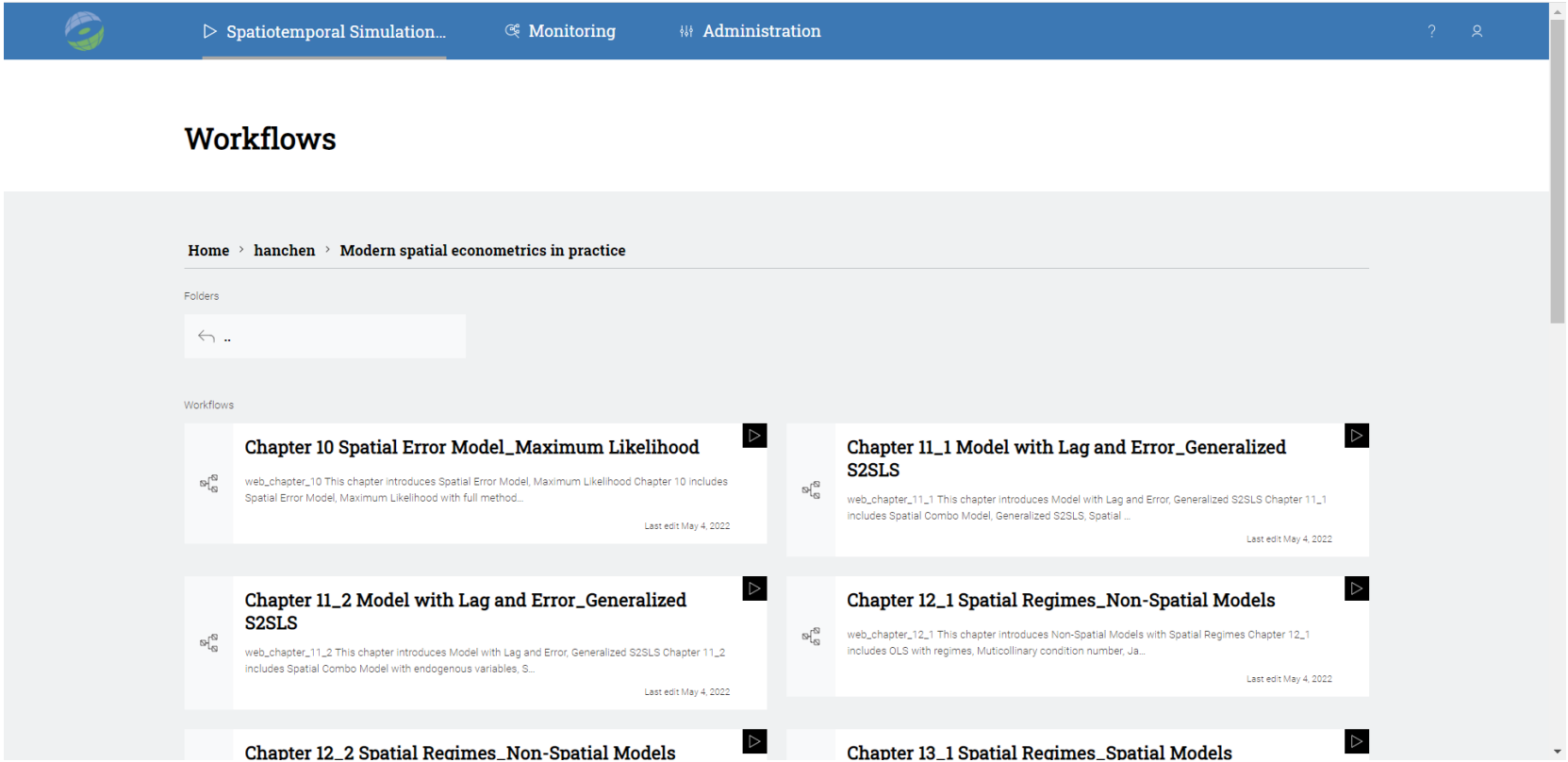
Comparison

	ArcMap	Pysal	GeoDa	GeoDaspace	MGWR	Matlab	Workbench
SLX						YES	YES
SLM		YES	YES	YES		YES	YES
SEM		YES	YES	YES		YES	YES
SDM						YES	YES
SDEM						YES	YES
SAC		YES		YES		YES	YES
GNS						YES	YES
Spatial regimes		YES		YES			YES
Endogeneity		YES		YES			YES
GWR	YES	YES			YES	YES	YES
MGWR		YES			YES		YES

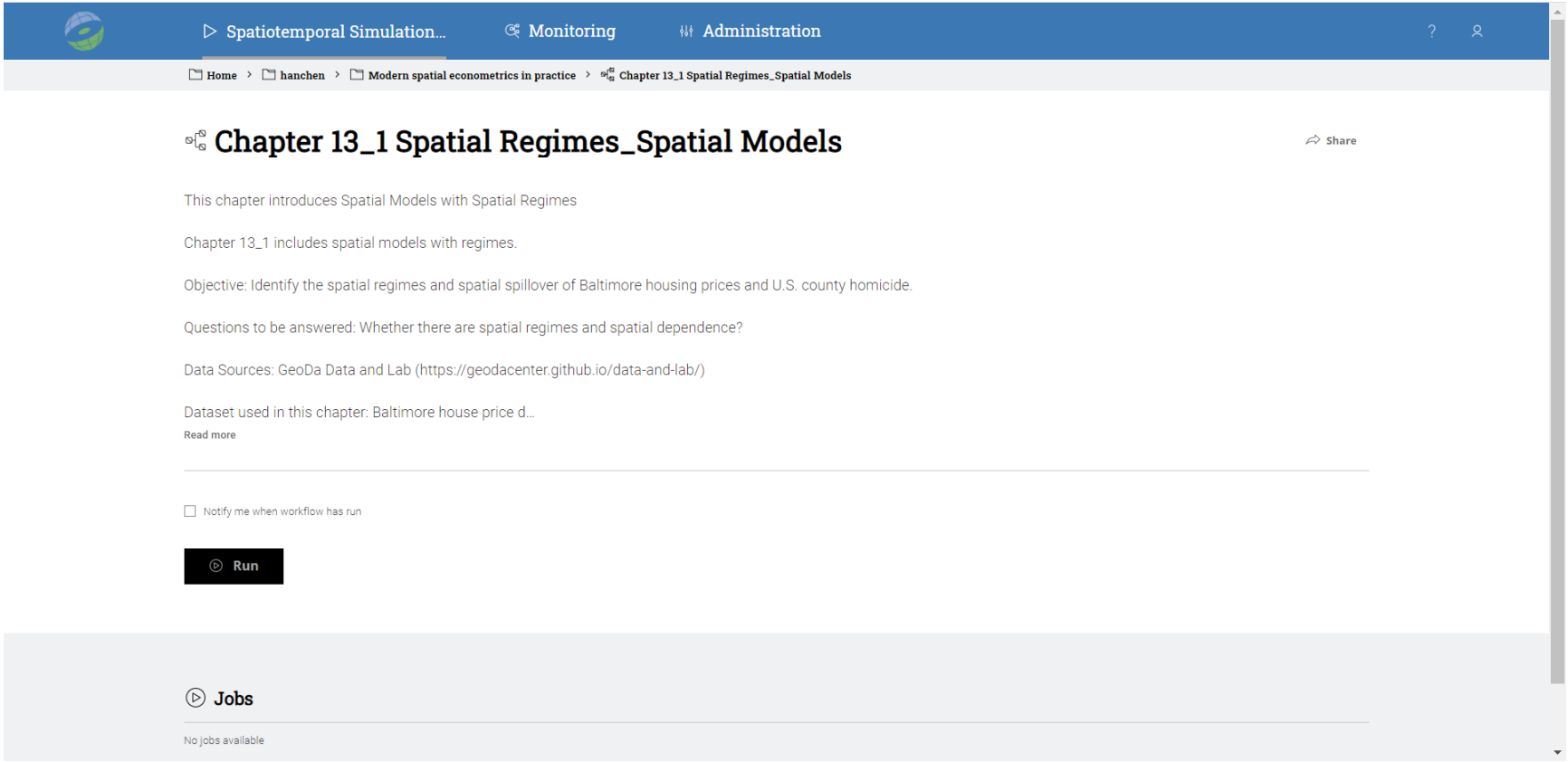
Introduction



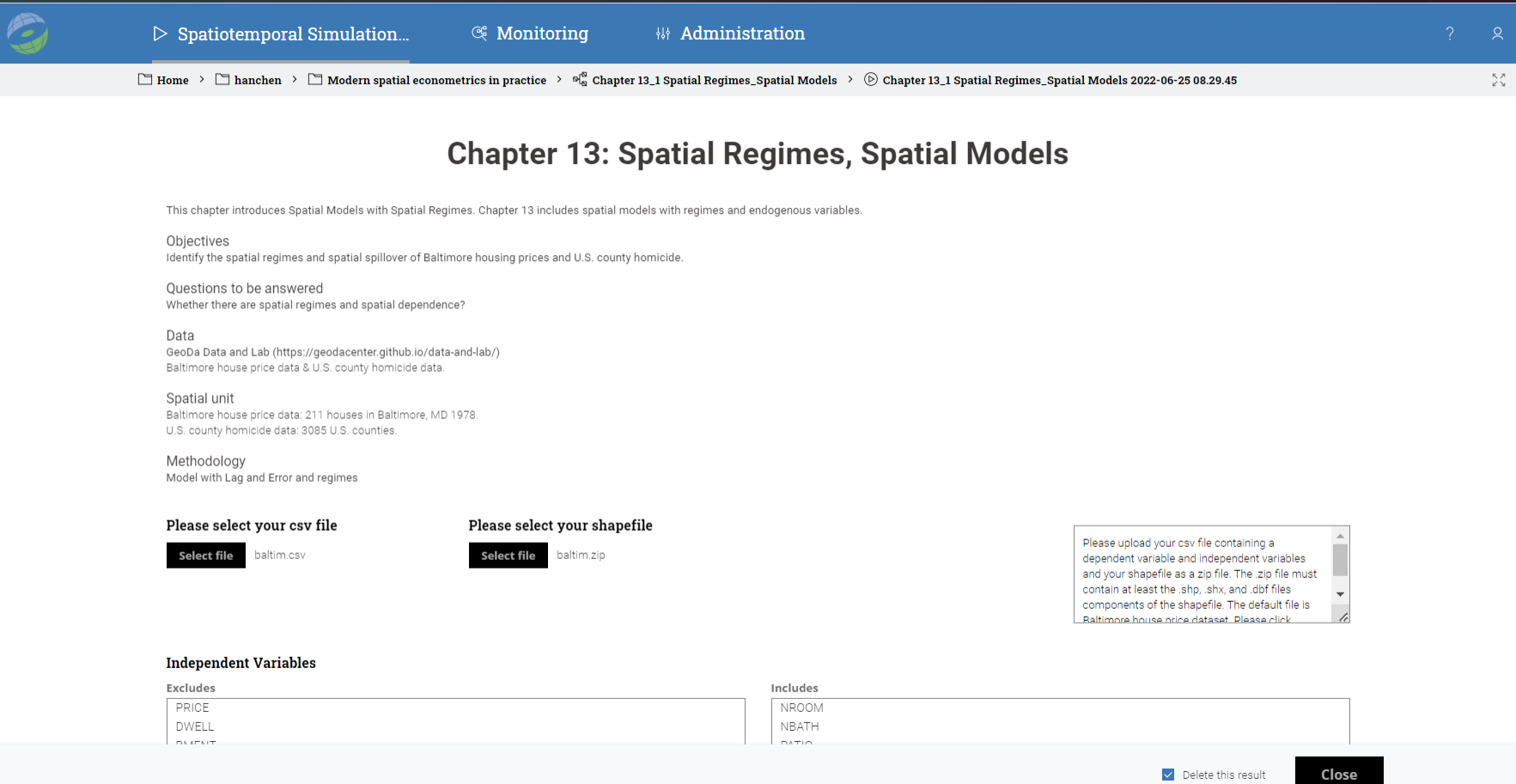
Introduction



Introduction

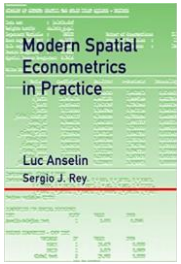


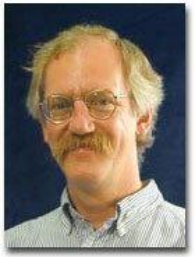
Introduction



Content

Modern Spatial Econometrics in Practice, A Guide to GeoDa, GeoDaSpace and PySAL



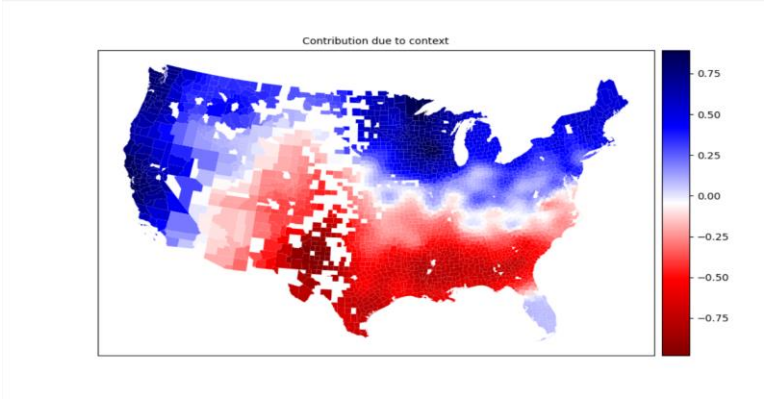


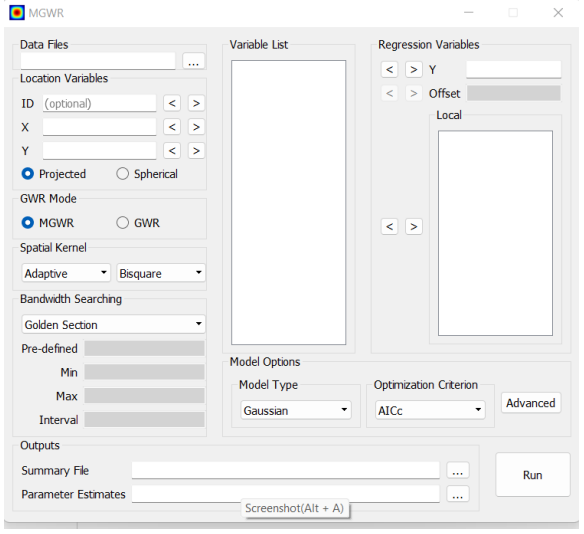
Chapter	Status
Chapter 1: Introduction	Done
Chapter 2: Getting Started	Done
Chapter 3: Spatial Weights: Contiguity	Done
Chapter 4: Spatial Weights: Distance	Done
Chapter 5: Ordinary Least Squares	Done
Chapter 6: Two Stage Least Squares (2SLS)	Done
Chapter 7: Spatial Lag Model, Spatial Two Stage Least Squares (S2SLS)	Done
Chapter 8: Spatial Lag Model, Maximum Likelihood	Done

Chapter	Status
Chapter 9: Spatial Error Model, General Method of Moments	Done
Chapter 10: Spatial Error Model, Maximum Likelihood	Done
Chapter 11: Model with Lag and Error, Generalized S2SLS	Done
Chapter 12: Spatial Regimes, Non-Spatial Models	Done
Chapter 13: Spatial Regimes, Spatial Models	Done

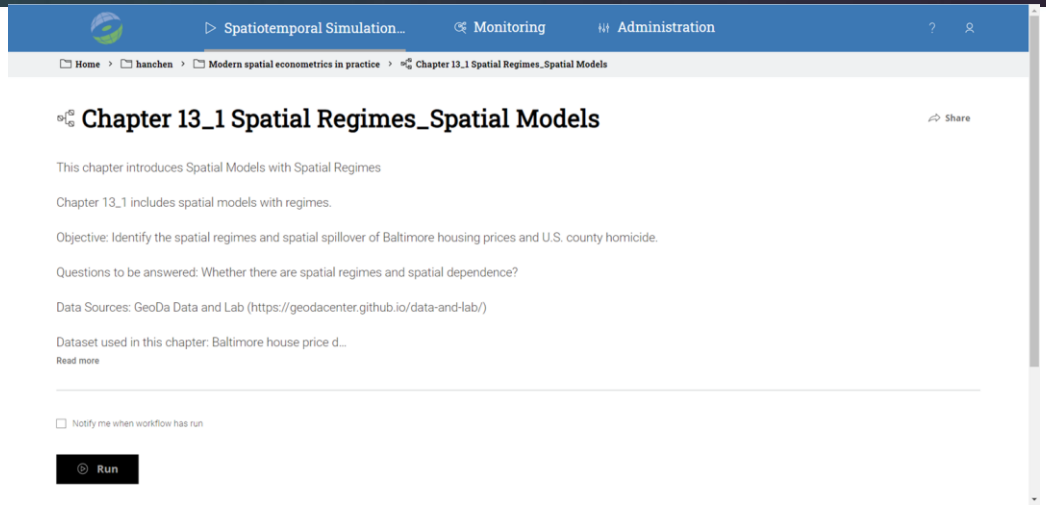
Content

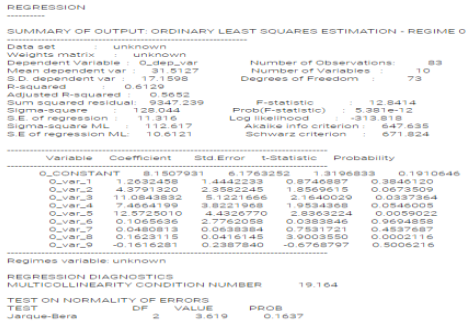
Multi-scale Geographically Weighted Regression(MGWR)



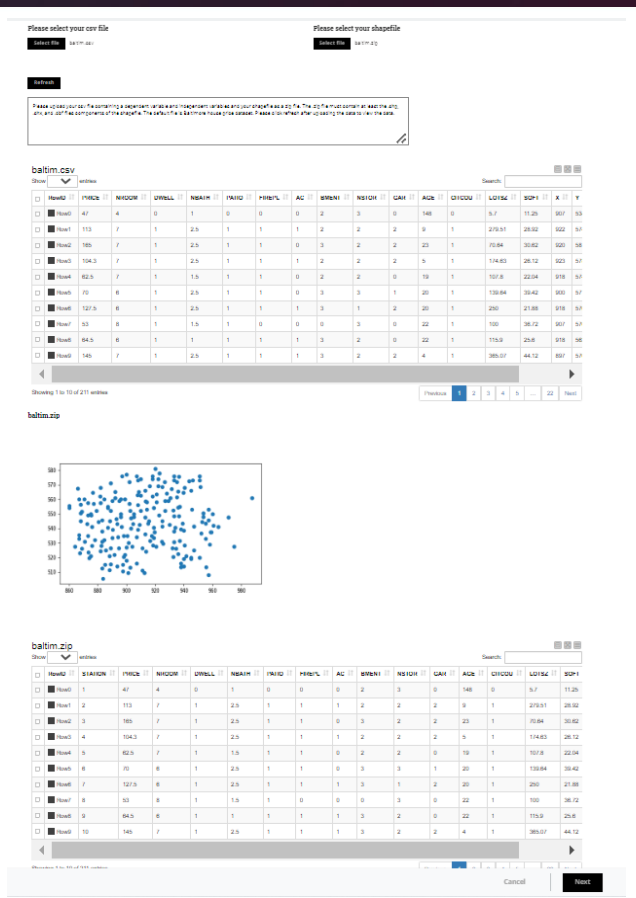


Introduction





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

SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION - REGIME 0
-----------------------------------------------------------
Data set            :    unknown
Weights matrix      :    unknown
Dependent Variable  :  0_dep_var                Number of Observations:          83
Mean dependent var  :     31.5127               Number of Variables   :          10
S.D. dependent var  :     17.1598               Degrees of Freedom    :          73
R-squared           :      0.6129
Adjusted R-squared  :      0.5652
Sum squared residual:    9347.239               F-statistic           :      12.8414
Sigma-square        :     128.044               Prob(F-statistic)     :   5.381e-12
S.E. of regression  :      11.316               Log likelihood        :    -313.819
Sigma-square ML     :     112.617               Akaike info criterion :     647.635
S.E of regression ML:      10.612               Schwarz criterion     :     671.824

-----------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     t-Statistic     Probability
-----------------------------------------------------------------------------
          0_CONSTANT      8.1507931      6.1763252      1.3196833      0.1910646
           0_var_1      1.2632458      1.4442233      0.8746887      0.3846120
           0_var_2      4.3791320      2.3582245      1.8566615      0.0673509
           0_var_3     11.0843832      5.1221666      2.1640029      0.0337364
           0_var_4      7.4664199      3.8221968      1.9534368      0.0546005
           0_var_5     12.5725010      4.4326770      2.8363224      0.0059022
           0_var_6      0.1065636      2.7762058      0.0383846      0.9694858
           0_var_7      0.0480813      0.0638384      0.7531721      0.4537687
           0_var_8      0.1623115      0.04161145     3.9003550      0.0002116
           0_var_9     -0.1616281      0.2387840     -0.6768797      0.5006216
-----------------------------------------------------------------------------
Regimes variable: unknown

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER        19.164

TEST ON NORMALITY OF ERRORS
TEST                     DF        VALUE           PROB
Jarque-Bera               2        3.619         0.1637
```



Introduction

Please select your csv file

Select file baltim.csv

Please select your shapefile

Select file baltim.zip

Refresh

Please upload your csv file containing a dependent variable and independent variables and your shapefile as a zip file. The .zip file must contain at least the .shp, .shx, and .dbf files components of the shapefile. The default file is Baltimore house price dataset. Please click refresh after uploading the data to view the data.

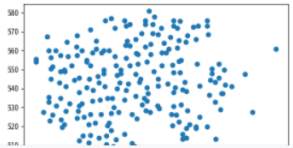
baltim.csv

Show 10 entries Search:

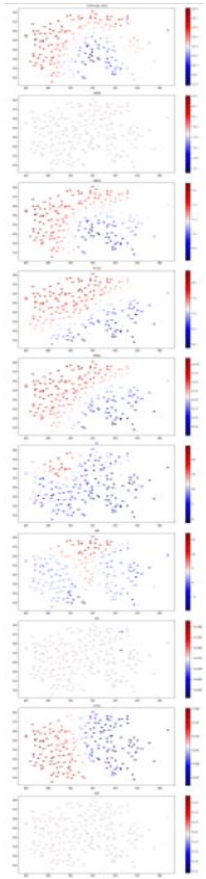
	RowID	STATION	PRICE	NROOM	DWELL	NBATH	PATIO	FIREPL	AC	BMENT	NSTOR	GAR	AGE	CITCOU	LOTSZ	SQFT
☐	Row0	1	47	4	0	1	0	0	0	2	3	0	148	0	5.7	11.25
☐	Row1	2	113	7	1	2.5	1	1	1	2	2	2	9	1	279.51	28.92
☐	Row2	3	165	7	1	2.5	1	1	0	3	2	2	23	1	70.64	30.62
☐	Row3	4	104.3	7	1	2.5	1	1	1	2	2	2	5	1	174.63	28.12
☐	Row4	5	62.5	7	1	1.5	1	1	0	2	2	0	19	1	107.8	22.04
☐	Row5	6	70	6	1	2.5	1	1	0	3	3	1	20	1	139.64	39.42
☐	Row6	7	127.5	6	1	2.5	1	1	1	3	1	2	20	1	250	21.88
☐	Row7	8	53	8	1	1.5	1	0	0	0	3	0	22	1	100	36.72
☐	Row8	9	64.5	6	1	1	1	1	1	3	2	0	22	1	115.9	25.6
☐	Row9	10	145	7	1	2.5	1	1	1	3	2	2	4	1	365.07	44.12

Showing 1 to 10 of 211 entries Previous 1 2 3 4 5 … 22 Next

baltim.zip



Cancel Next



```
==========================================================================
Model type                                  Gaussian
Number of observations:                          211
Number of covariates:                             10

Global Regression Results
--------------------------------------------------------------------------
Residual sum of squares:                   40960.463
Log-likelihood:                             -855.223
AIC:                                        1730.446
AICc:                                       1733.773
BIC:                                        39884.740
R2:                                             0.650
Adj. R2:                                        0.634

Variable                 Est.         SE  t(Est/SE)    p-value
------------------------------------------------------------------------
X0                     23.270      5.224      4.454      0.000
X1                      0.222      1.228      0.181      0.856
X2                      5.648      2.018      2.799      0.005
X3                     10.336      3.120      3.312      0.001
X4                     11.173      2.732      4.089      0.000
X5                      7.854      2.702      2.906      0.004
X6                      5.402      1.964      2.751      0.006
X7                     -0.213      0.057     -3.722      0.000
X8                      0.095      0.017      5.740      0.000
X9                      0.188      0.189      0.993      0.321

Multi-Scale Geographically Weighted Regression (MGWR) Results
--------------------------------------------------------------------------
Spatial kernel:                     Adaptive bisquare
Criterion for optimal bandwidth:                AICc
Score of Change (SOC) type:                Smoothing f
Termination criterion for MGWR:                 1e-05

MGWR bandwidths
--------------------------------------------------------------------------
Variable     Bandwidth      ENP_j  Adj t-val(95%)  Adj alpha(95%)
X0         210.000      1.127          2.023          0.044
X1          46.000     10.797          2.862          0.005
X2         200.000      1.218          2.056          0.041
```

Case 1: Columbus Crime

- ❑ Objective: To identify the determinants of neighborhood crime.
- ❑ Questions to be answered:
  - ❑ Whether there is spatial dependence?
- ❑ Data Sources:
- GeoDa Data and Lab: https://geodacenter.github.io/data-and-lab/
- Dependent variable: CRIME(residential burglaries and vehicle thefts per 1000 households)
- Independent variable: INC(household income (in $1,000)), HOVAL(housing value (in $1,000)), OPEN(open space (area))
- ❑ Spatial Unit: 49 contiguous Planning Neighborhoods in Columbus, OH, 1980

Reference：

- Anselin, Luc (1988). Spatial Econometrics. Boston, Kluwer Academic, Table 12.1, p. 189.

Outputs

SLM

```
----------
SUMMARY OF OUTPUT: SPATIAL TWO STAGE LEAST SQUARES
----------
Data set            :     unknown
Weights matrix      :     unknown
Dependent Variable  :     dep_var                Number of Observations:          49
Mean dependent var  :     35.1288                Number of Variables   :           5
S.D. dependent var  :     16.7321                Degrees of Freedom    :          44
Pseudo R-squared    :      0.6424
Spatial Pseudo R-squared:  0.6172

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     z-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT      57.8537536       8.3837142       6.9007306       0.0000000
               var_1      -0.2783675       0.0921089      -3.0221577       0.0025098
               var_2      -1.3534671       0.3376643      -4.0083222       0.0000612
               var_3       0.3345079       0.3173690       1.0540030       0.2918815
           W_dep_var       0.0431994       0.0294060       1.4690682       0.1418143
------------------------------------------------------------------------------------
Instrumented: W_dep_var
Instruments: W_var_1, W_var_2, W_var_3

DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST                           MI/DF       VALUE           PROB
Anselin-Kelejian Test              1           0.632          0.4267
```

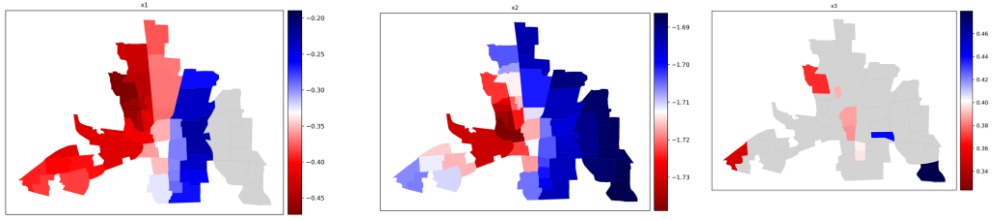
SEM

```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIALLY WEIGHTED LEAST SQUARES (HET)
----------
Data set            :     unknown
Weights matrix      :     unknown
Dependent Variable  :     dep_var                Number of Observations:          49
Mean dependent var  :     35.1288                Number of Variables   :           4
S.D. dependent var  :     16.7321                Degrees of Freedom    :          45
Pseudo R-squared    :      0.5531
N. of iterations    :           1                Step1c computed       :          No

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     z-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT      62.5429047       4.6203059      13.5365291       0.0000000
               var_1      -0.3214740       0.1696411      -1.8950240       0.0580892
               var_2      -1.1242548       0.4285375      -2.6234691       0.0087039
               var_3       0.3376159       0.2149787       1.5704623       0.1163076
              lambda       0.5436331       0.1428935       3.8044628       0.0001421
```

LM tests

```
REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER           6.972

TEST ON NORMALITY OF ERRORS
TEST                             DF         VALUE           PROB
Jarque-Bera                       2         1.187           0.5525

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                             DF         VALUE           PROB
Breusch-Pagan test                3         8.714           0.0333
Koenker-Bassett test              3         6.605           0.0856

SPECIFICATION ROBUST TEST
TEST                             DF         VALUE           PROB
White                             9        22.030           0.0088

DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST                           MI/DF        VALUE           PROB
Moran's I (error)             0.2557        3.107           0.0019
Lagrange Multiplier (lag)          1       10.450           0.0012
Robust LM (lag)                    1        5.720           0.0168
Lagrange Multiplier (error)        1        6.540           0.0105
Robust LM (error)                  1        1.810           0.1785
Lagrange Multiplier (SARMA)        2       12.260           0.0022
```

MGWR



Case 2: Election

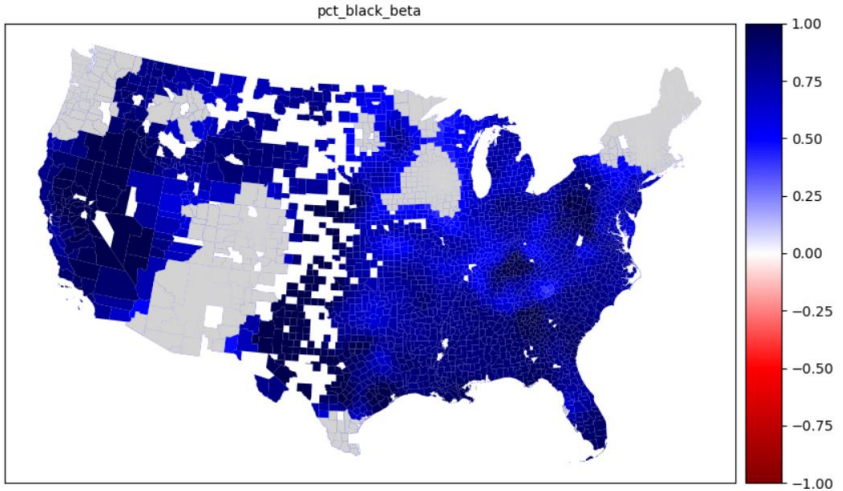
- ❑ Objective: To identify the impact of location(geography) other than socioeconomic factors on voting. To identify the spatial heterogeneity of the impact of socioeconomic factors on voting.
- ❑ Questions to be answered:
  - ❑ Is there spatial heterogeneity in the impact of socioeconomic factors on voting?
  - ❑ Whether location other than socioeconomic factors have an impact on voting?
- ❑ Data Sources:
- The MIT Election Lab: https://electionlab.mit.edu/data
- US Census Bureau: https://www.census.gov/programs-surveys/acs/technical-documentation/table-and-geography-changes/2016/5-year.html

Reference：

- Stewart Fotheringham, A., Li, Z. and Wolf, L.J., 2021. Scale, Context, and Heterogeneity: A Spatial Analytical Perspective on the 2016 US Presidential Election. *Annals of the American Association of Geographers*, pp.1-20.

The impact of percentage black

- The impact of percentage black:The tendency for African American voters to favor the Democratic Party is well known and is borne out by both the global result and the local parameter estimates displayed in the figure. Most counties across the United States have extremely strong relationships between the percentage of African Americans in each county and the Democratic vote percentage. The only exceptions are swathes of counties in the Northwest, in the Southwest, the upper Midwest, New England, and the southern tip of Texas. What unites all of these counties is a very low proportion of African American residents, so there is very little variation locally in the percentage of African Americans.



Case 3: Baltimore house price

- ❏ Objective: To identify the influencing factors of housing prices, Baltimore, MD 1978.

- ❏ Questions to be answered:
  - ❏ Whether there is spatial dependence?
  - ❏ Whether there is spatial heterogeneity?

- ❏ Data Sources:
- GeoDa Data and Lab: https://geodacenter.github.io/data-and-lab/

- ❏ Spatial Unit: 211 houses in Baltimore, MD 1978.

Reference：

- Dubin, Robin A. (1992). Spatial autocorrelation and neighborhood quality. Regional Science and Urban Economics 22(3), 433-452.

SLM

- Questions:

  Whether there is a spatial lag of Y?

  There is a spatial lag of Y, the spatial coefficient is 0.58 and significant which indicates the spatial dependency in the model.

```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL TWO STAGE LEAST SQUARES
--------------------------------------------------
Data set            :     unknown
Weights matrix      :     unknown
Dependent Variable  :     dep_var               Number of Observations:         211
Mean dependent var  :     44.3072               Number of Variables   :          11
S.D. dependent var  :     23.6061               Degrees of Freedom    :         200
Pseudo R-squared    :      0.7278
Spatial Pseudo R-squared:  0.6928

White Standard Errors
------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     z-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT      -2.5762742       7.0147591      -0.3672648       0.7134215
               var_1       0.9440746       1.4002856       0.6742015       0.5001832
               var_2       5.5981348       2.1605285       2.5910951       0.0095671
               var_3       5.8424768       2.9445656       1.9841558       0.0472385
               var_4       6.4579185       2.4500195       2.6358641       0.0083923
               var_5       5.4871926       2.6021469       2.1087175       0.0349690
               var_6       4.3565951       2.2070747       1.9739228       0.0483905
               var_7      -0.0730060       0.0976079      -0.7479516       0.4544894
               var_8       0.0579765       0.0237454       2.4415887       0.0146228
               var_9       0.0395330       0.2355809       0.1678105       0.8667323
           W_dep_var       0.5823313       0.1325884       4.3920220       0.0000112
------------------------------------------------------------------------------------
Instrumented: W_dep_var
Instruments: W_var_1, W_var_2, W_var_3, W_var_4, W_var_5, W_var_6, W_var_7,
             W_var_8, W_var_9
```

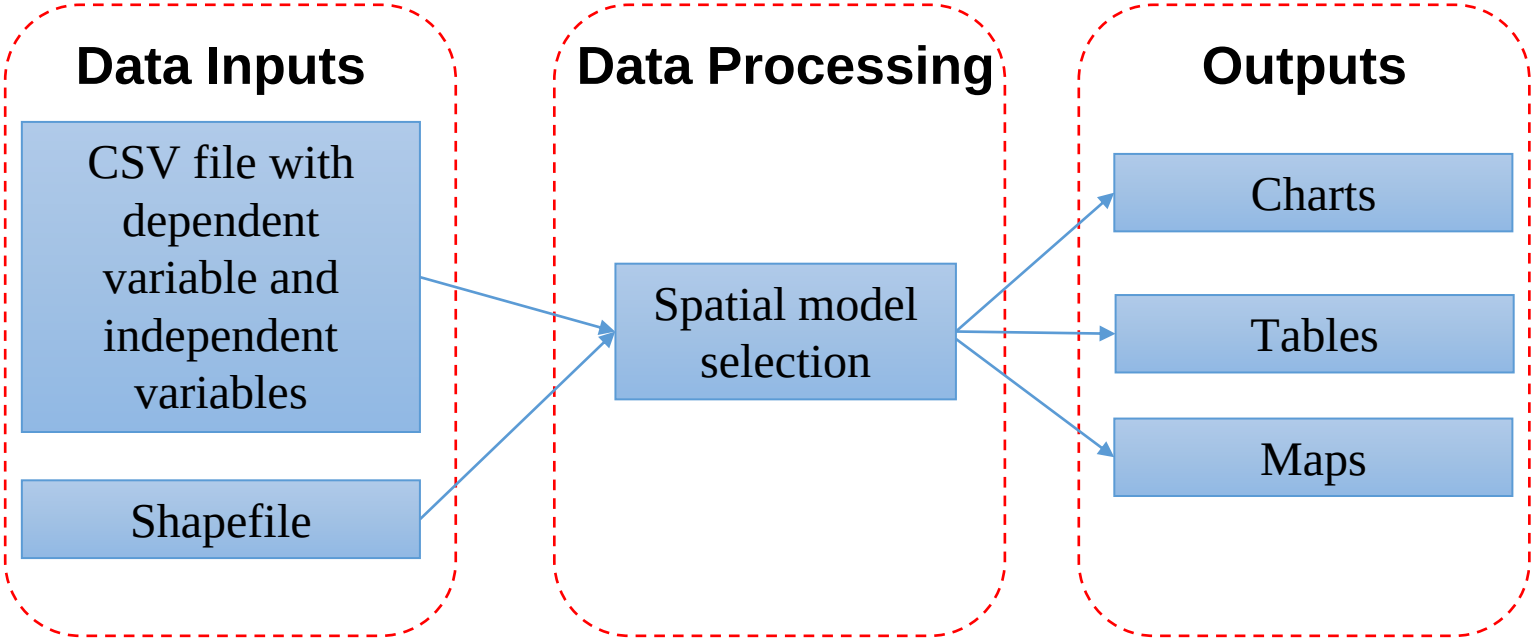
Workflow for Spatial Model Selection

- Objective: Selection of Spatial Regression Model Based on LM Tests.
- Questions to be answered: Which spatial model should be used?
- Data Sources: GeoDa Data and Lab (https://geodacenter.github.io/data-and-lab/)
- Dataset: Baltimore house price data
- Spatial unit: 211 houses in Baltimore, MD 1978.

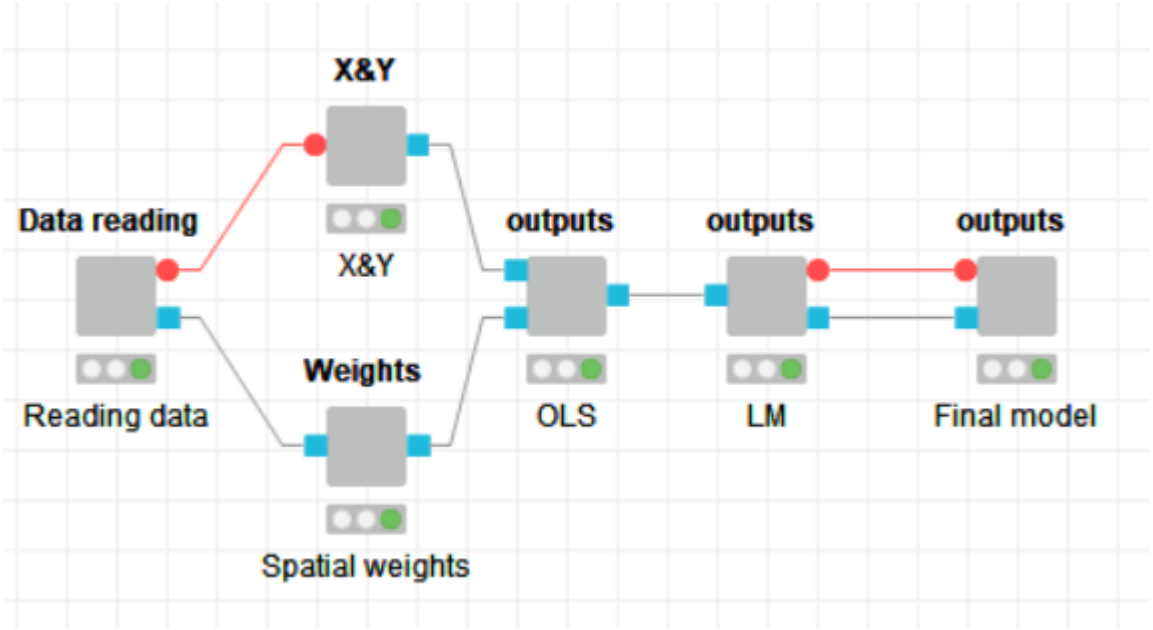
Methodology

- Spatial Lag Model
- Spatial Error Model
- Lagrange Multiplier Test (for SLM and SEM)
- Robust Lagrange Multiplier Test (for SLM and SEM)

Flowchart

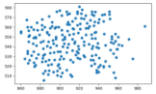

Data Inputs
CSV file with dependent variable and independent variables
Shapefile
Data Processing
Spatial model selection
Outputs
Charts
Tables
Maps

Workflow Implementation



Results

baltim.zip



baltim.zip

Show 10 entries Search:

RowID	STATION	PRICE	NROOM	DWELL	NBATH	PATIO	FIREPL	AC	BMENT	NSTOR	GAR	AGE	CITCOU	LOTSZ	SQFT
Row0	1	47	4	0	1	0	0	0	2	3	0	148	0	5.7	11.25
Row1	2	113	7	1	2.5	1	1	1	2	2	2	9	1	279.51	28.92
Row2	3	165	7	1	2.5	1	1	0	3	2	2	23	1	70.64	30.62
Row3	4	104.3	7	1	2.5	1	1	1	2	2	2	5	1	174.63	26.12
Row4	5	62.5	7	1	1.5	1	1	0	2	2	0	19	1	107.8	22.04
Row5	6	70	6	1	2.5	1	1	0	3	3	1	20	1	139.64	39.42
Row6	7	127.5	6	1	2.5	1	1	1	3	1	2	20	1	250	21.88
Row7	8	53	8	1	1.5	1	0	0	0	3	0	22	1	100	36.72
Row8	9	64.5	6	1	1	1	1	1	3	2	0	22	1	115.9	25.6

LM Test

Lagrange Multiplier tests:

LM error test=7.8844p-value=0.005
LM lag test=44.2041p-value=0.0

Both statistics are significant, run robust LM test

Robust LM error test:3.2567p-value=0.0711
Robust LM lag test=39.5764p-value=0.0

Only robust LM-Lag is significant. Spatial Lag Model is Recommended

Please select the model you want

Both

Model Regression Results

```
-----------------------------------------------------------------------------
Instrumented: W_PRICE
Instruments: W_AC, W_AGE, W_FIREPL, W_GAR, W_LOTSZ, W_NBATH, W_NROOM,
        W_PATIO, W_SQFT

DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST                    MI/DF     VALUE        PROB
Anselin-Kelejian Test        1       4.873       0.0273
================================ END OF REPORT ====================================



Spatial Error Model

REGRESSION
----------
SUMMARY OF OUTPUT: SPATIALLY WEIGHTED LEAST SQUARES
-----------------------------------------------------------------------------
Data set          :    baltim
Weights matrix    :     Queen
Dependent Variable :     PRICE          Number of Observations:      211
Mean dependent var :    44.3072          Number of Variables   :       10
S.D. dependent var :    23.6061          Degrees of Freedom    :      201
Pseudo R-squared   :     0.6470

-----------------------------------------------------------------------------
        Variable    Coefficient     Std.Error    z-Statistic    Probability
-----------------------------------------------------------------------------
        CONSTANT     20.2706731     5.0850205     3.9863503     0.0000671
          NROOM      1.0343307     1.1452993     0.9031095     0.3664678
          NBATH      6.3658124     1.9294575     3.2992759     0.0009693
          PATIO      8.3841811     3.0158917     2.7800008     0.0054359
         FIREPL      9.2742263     2.5854857     3.5870344     0.0003345
             AC      6.7924620     2.5589015     2.6544445     0.0079439
            GAR      4.5596828     1.8924946     2.4093504     0.0159809
            AGE     -0.1820243     0.0601426    -3.0265460     0.0024737
          LOTSZ      0.0860646     0.0162254     5.3043259     0.0000001
           SQFT      0.1075179     0.1796858     0.5983658     0.5495959
         lambda      0.2604128
-----------------------------------------------------------------------------
================================ END OF REPORT ====================================
```

Website and Contact

Project website:

http://spatialdatalab.org

Contact:

spatialdatalab@lists.fas.harvard.edu