Spatial Social Science

Workflow-Based Practices Series II

The Workbench for Spatial Data Science

Presenter: Hanchen Yu, CGA, Harvard University
Chair: Wendy Guan, CGA, Harvard University

Spatial Data Lab
CGA @ Harvard University
GCSS@Wuhan University

Background: What is the WSDS?

- Workbench is the place where you will be building your workflows. Workbench is a collection of workflows?
- WSDS integrates a large number of spatial analysis workflows. It contains a large number of nodes for building a spatial analysis workflow.
- We will build workflows which use the most commonly used and the latest methods and add them to the workbench.
- After the release, we will establish a user group to update and feedback on the workbench in time.
- Workbench can serve different types of research. You could build your own workflow by using the nodes in the workbench or you could directly use the workbench to run all possible models and analyses.

Background: Why WSDS?

- **Demands**
  - Repetitive data analysis
  - Replicable data analysis
  - Customized data analysis
  - High efficient data analysis
  - Data analysis for everyone without professional skill

- **Challenges**
  - Most tools offer interactive and step by step data analysis with low efficiency
  - Most tools offer fixed functions for data analysis or limited space for expanded functions
  - Most data analysis (procedure and code) are difficult to replicate
  - Most data analysis need professional skills while data analysis become more and more complicated

- **Solutions** (Combinable, replicable, expandable, suitable for users of any background)
  - Tools with assemblable functions for data analysis
  - Tools with replicable process
  - Tools with expandable functions
  - Tools with easy operation for everyone

Objectives

This project aims to build a workbench for spatial data science based on workflow tools

The objectives are to provide:

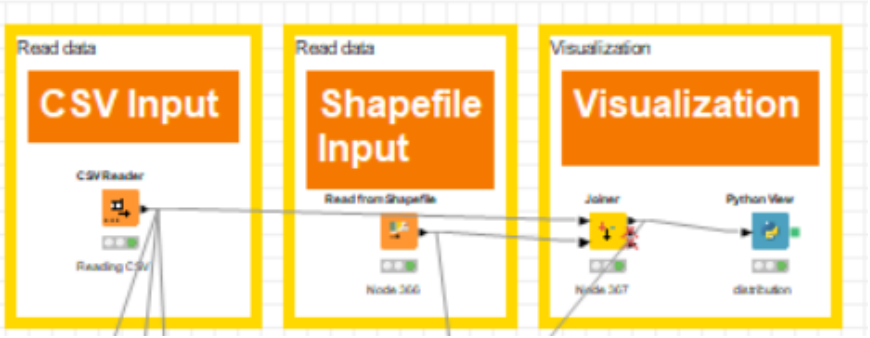
1. Easy, efficient and customizable toolkit for spatial data analysis with new added nodes
2. Workflow based case studies for research in spatial social sciences
3. Integration of data, methodology and applications for spatial data science
4. Training base for users with no skills in GIS and advanced methodology

Tasks

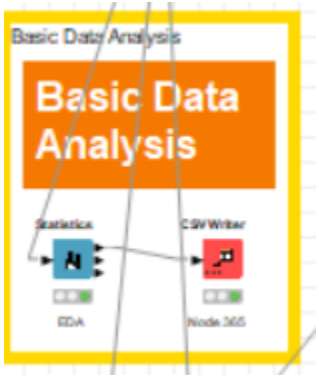
- ❑ Workbench development
  - Data access
  - Data processing and integration
  - Data analysis
    - Exploratory data analysis
    - Data modeling
    - Data visualization
  - Data applications
- ❑ New node development
- ❑ Research applications
  - Case studies
  - Publications
- ❑ Teaching applications
  - Curriculum
  - Teaching case studies
- ❑ Publications
  - User Guides
  - Journal papers and Books

The Workbench for Spatial Statistics built on KNIME

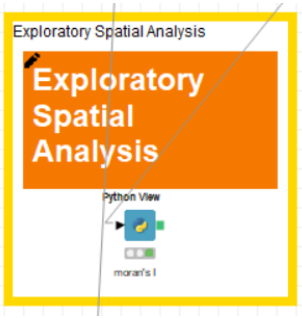
1 data input **2 basemap input** **3 visualization**



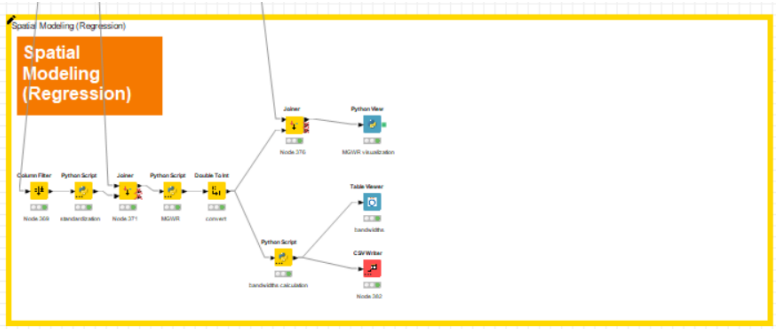
4 EDA



5 ESDA



6 OLS, MGWR, spatial regression, spatial filtering

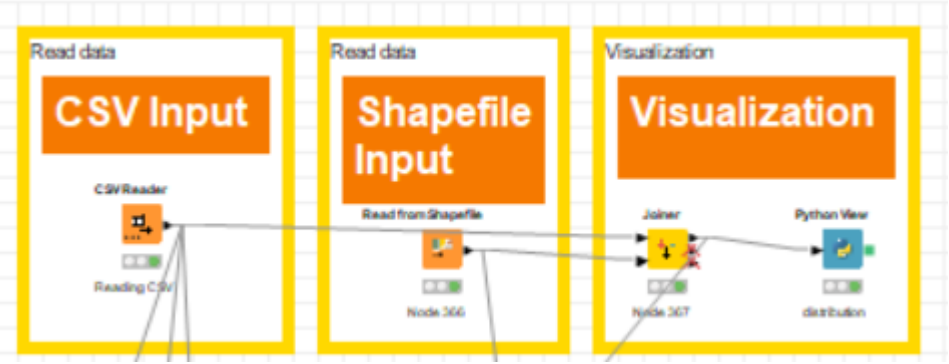


Plans

1. Research data collection
2. Workflow & node development
3. Case development
4. Curriculum development
5. User Guide

Plans 1. Research data collection

- Data collections on Harvard Dataverse
- Research data
- Base maps(Global, China, US).



Plan 2. Workflow & node development

Data Access

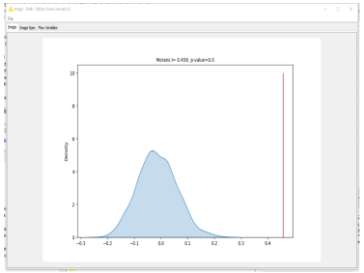
- Dataverse
- GiHub
- CGE
- ….

EDA

- Descriptive reports
- Box plot
- Histogram
- Multi-vari chart
- Run chart
- Pareto chart
- Scatter plot
- Stem-and-leaf plot
- Parallel coordinates
- Odds ratio
- Multidimensional scaling
- Targeted projection pursuit
- Principal component analysis
- Multilinear PCA

ESDA

- Variogram
- Moran's I
- LISA
- Geit's G
- Geary's C
- Outlier detection
- Trend detection



Modeling

- OLS regression
- Spatial regression
- nonlinear regression
- panel regression
- Spatial filtering
- MGWR/MGTWR

Advanced modeling

- Machine learning
- Cellular automata
- Agent based model

Spatial data visualization

- Static maps
- Dynamic maps

Plans 3 &4: Case and Cruuiculum Development

3. Case development

- Case studies for selected papers
- Workbook for selected books

4. Curriculum development

- Spatial econometrics and spatiotemporal innovation workshop
- Spatial social science course
- Spatial data analysis course

Data Sources and Code Souirces

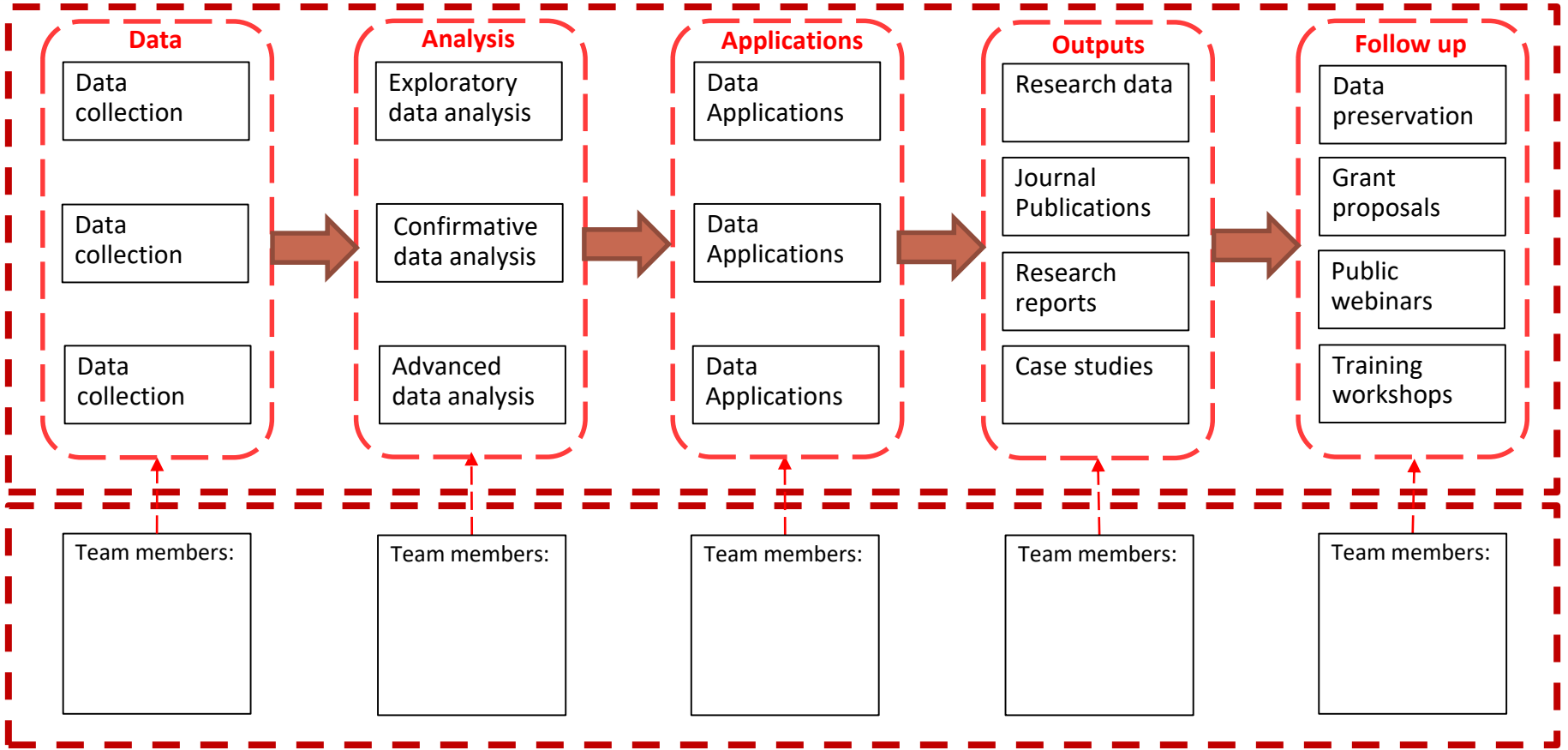
Data Sources:

- Dataverse
- GiHub
- China Data Online
- ….

Programs and codes:

- Pysal/GeoDa
- Matlab Spatial econometrics toolbox
- MGWR
- Spatial Econometric Functions in R
- Matlab Spatial filtering

Project Flowchart



Deployment and Sharing

- Dataverse at Harvard
- SDL at Harvard
- SDL at ECUST
- SDL at Wuhan Uni
- SDL at PKU

Sample Works

Spatial Statistics workbench

Sample Work 1: Spatial regression

- ❑ Objective: To identify the factors of spatial difference in neighborhood crime.
- ❑ 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 (2003) “An Introduction to Spatial Regression Analysis in R”, http://labs.bio.unc.edu/Buckley/documents/AnselinIntroSpatRegres.pdf

Methodology

- Ordinary least squares (OLS):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \boldsymbol{\varepsilon}$$

- Spatial regression(SLX, SEM, SLM):

- SLX:

$$y = X\beta + WX\theta + \varepsilon$$

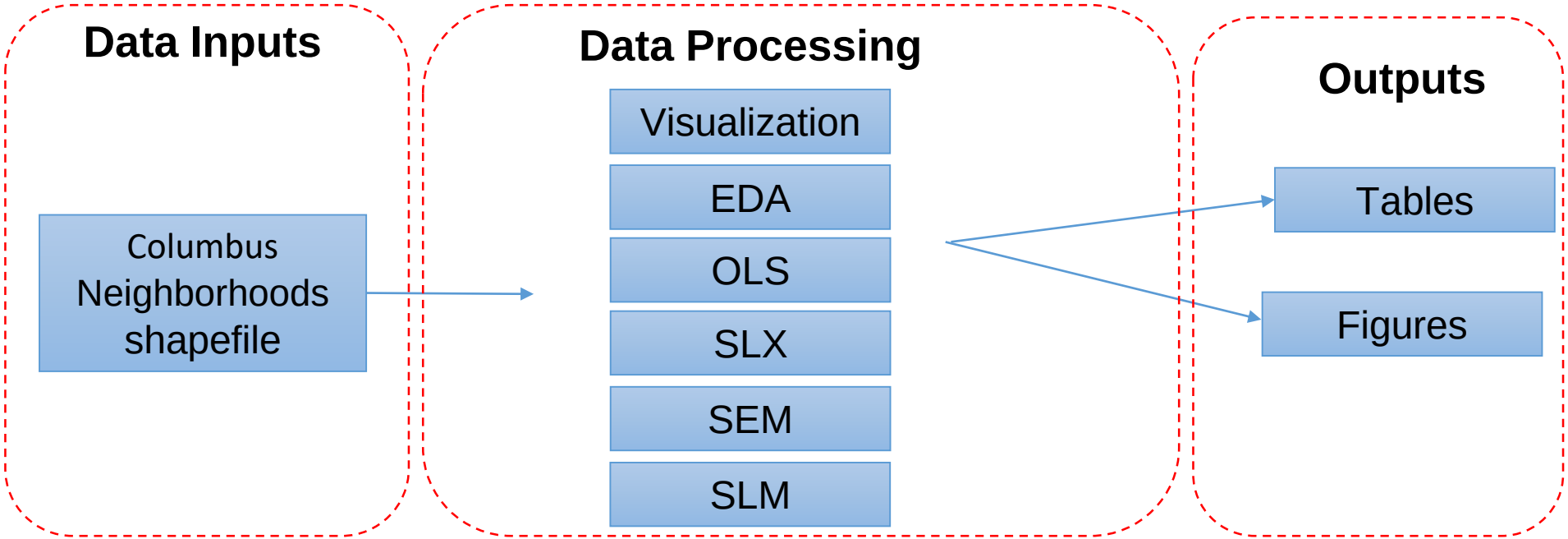
SEM:

$$\mathrm{y} = \mathrm{X}\beta + (\mathrm{I_n} - \lambda \mathrm{W})^{-1}\varepsilon$$

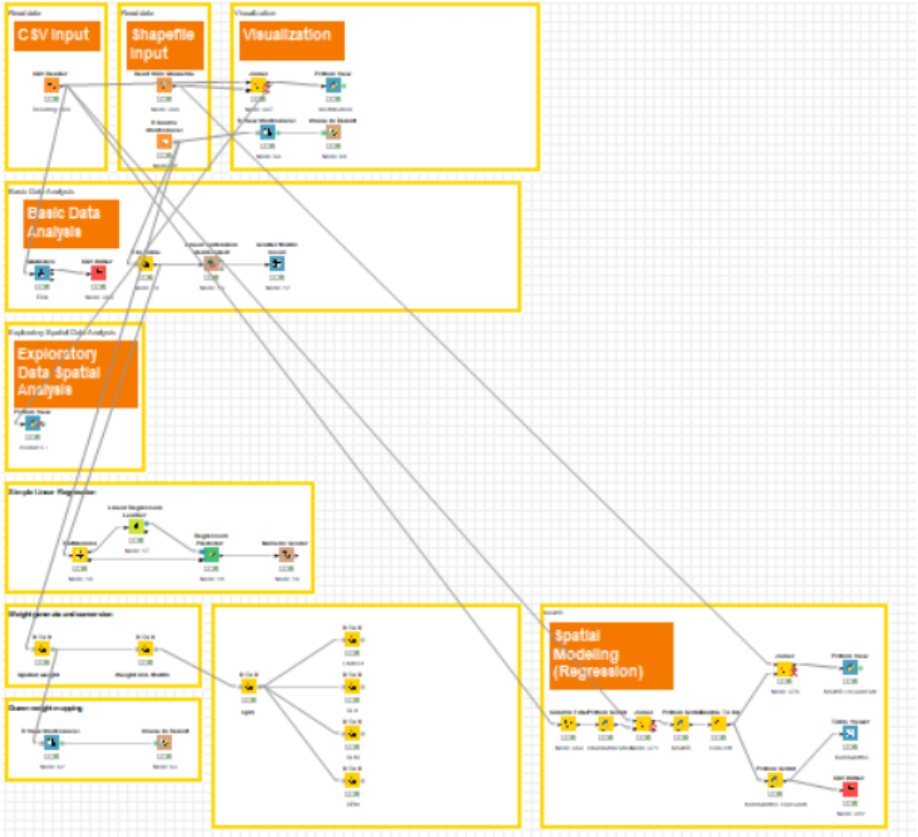
SLM:

$$\mathrm{y} = \rho \mathrm{W}\mathrm{y} + \mathrm{X}\beta + \varepsilon$$

Flowchart of Data Analysis



Workflow Implementation



Required R packages:

- library("rgdal")
- library("maptools")
- library("spdep")
- library("tripack")
- library("RANN")
- library("plm")
- library("splm")

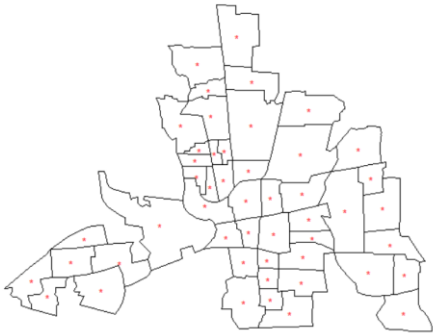
Data Input

- Data Input:

Configure Shapefile Input node: Change the path to where the data file (columbus.shp) is located.

```
library("rgdal")
library("maptools")
library("spdep")
library("tripack")
library("RANN")
library("plm")
library("splm")


f<-CRIME~ HOVAL+INC+OPEN
mdata <- readOGR("C:/Users/12074/Desktop/Spatial statistics workbench/columbus/columbus.shp")
df<- mdata@data
map_crd <- coordinates(mdata)
```

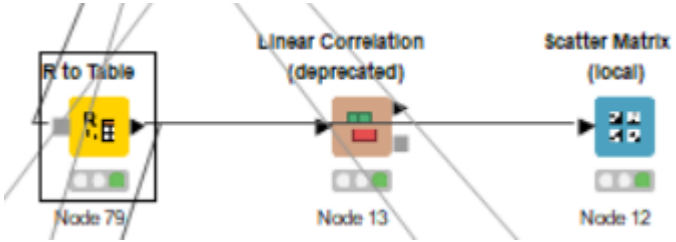


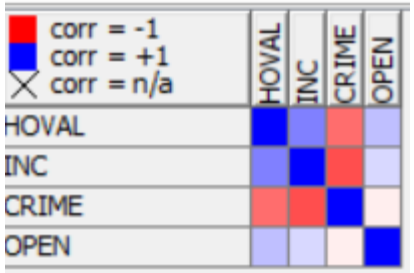
Exploratory Data Analysis

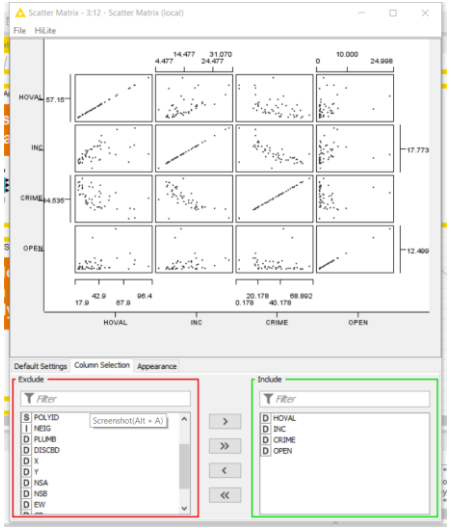
❑ Questions:

- Whether there are correlations between variables?
- If there is any outliers?

Run the EDA node, right click on the node, then click Linear Correlation & Scatter Matrix.







Results from OLS

- Question: What factors affect the spatial distribution of crime?

Statistics on Linear Regression

Variable	Coeff.	Std. Err.	t-value	P>|t|
HOVAL	-0.1134	0.2054	-0.5521	0.585
INC	-1.7991	0.4357	-4.1294	0.0003
OPEN	0.2519	0.3798	0.6633	0.5122
Intercept	65.5304	5.4041	12.126	4.30E-13

Multiple R-Squared: 0.5954
Adjusted R-Squared: 0.5549

Diagnostic Tests for Spatial Dependence

❑ Questions: :

- Whether there is spatial autocorrelation in OLS residuals?
- Which model should be applied: SLM or SEM?

```
        Lagrange multiplier diagnostics for spatial dependence

data:
model: lm(formula = f, data = mdata)
weights: w_cn_mat

RLMerr = 0.10028, df = 1, p-value = 0.7515


        Lagrange multiplier diagnostics for spatial dependence

data:
model: lm(formula = f, data = mdata)
weights: w_cn_mat

RLMlag = 3.4878, df = 1, p-value = 0.06182


        Lagrange multiplier diagnostics for spatial dependence

data:
model: lm(formula = f, data = mdata)
weights: w_cn_mat

SARMA = 8.9576, df = 2, p-value = 0.01135
```

```
        Global Moran I for regression residuals

data:
model: lm(formula = f, data = mdata)
weights: w_cn_mat

Moran I statistic standard deviate = 2.8765, p-value = 0.00201
alternative hypothesis: greater
sample estimates:
Observed Moran I      Expectation         Variance
     0.227662705     -0.032487594      0.008179232


        Lagrange multiplier diagnostics for spatial dependence

data:
model: lm(formula = f, data = mdata)
weights: w_cn_mat

LMerr = 5.4698, df = 1, p-value = 0.01935


        Lagrange multiplier diagnostics for spatial dependence

data:
model: lm(formula = f, data = mdata)
weights: w_cn_mat

LMlag = 8.8573, df = 1, p-value = 0.002919


        Lagrange multiplier diagnostics for spatial dependence
```

Results from SLM

❑ Questions:

- Which factors have significant impact on the spatial distribution of crime?
- If the spatial dependence is significant in the spatial lag model?

```
Call:lagsarlm(formula = f, data = mdata, listw = w_cn_mat)

Residuals:
     Min       1Q   Median       3Q      Max
-36.1366  -4.6581  -0.2158   6.9110  23.4083

Type: lag
Coefficients: (asymptotic standard errors)
             Estimate Std. Error z value  Pr(>|z|)
(Intercept) 45.796780   7.205228  6.3560  2.07e-10
HOVAL       -0.284854   0.090211 -3.1576 0.0015906
INC         -1.063214   0.304671 -3.4897 0.0004835
OPEN         0.312493   0.311979  1.0016 0.3165136

Rho: 0.41935, LR test value: 9.3503, p-value: 0.0022294
Asymptotic standard error: 0.11922
    z-value: 3.5176, p-value: 0.00043553
Wald statistic: 12.373, p-value: 0.00043553

Log likelihood: -182.1766 for lag model
ML residual variance (sigma squared): 94.999, (sigma: 9.7467)
Number of observations: 49
Number of parameters estimated: 6
AIC: 376.35, (AIC for lm: 383.7)
LM test for residual autocorrelation
test value: 0.31597, p-value: 0.57404
```

Summary and Discussion

❑ **Summary:**

- There is significant spatial dependence in the crime distribution over the space.
- OLS model should not be applied directly due to the spatial dependence of the error term.
- Spatial lag model (SLM) is a better choice.

❑ **Discussion on possible expansions:**

- Allow tests of more other variables
- Allow tests of more other models
- Expand to other spatial studies

Instructions for running the workflow

Step 1: **Download data files** from \Spatial statistics\workbench\01_data\columbus_data\columbus.shp
Step 2: **Download workflow** from \Spatial statistics workbench\02_workflow\workbench.knwf
Step 3: **Open KNIME** (version 4.4.1) from local PC
Step 4: Install the following R packages if not yet:

"rgdal" "plm" "splm" "spdep"
"maptools" "RANN" "tripack"
"spdep" "tripack"

Step 5: **Import KNIME workflow** file (workbench.knwf)
Step 6: **Configure "CSV Input" & "Shapefile Input"**
Step 7: Click **Run** button on the top menu
Step 8: **Display the outputs** (see outputs folder)

Sample Work 2: MGWR

- ❑ Objectives:
- 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 has any 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.

Data

- Dependent variable: Votes for the Democratic Party in the 2016 U.S. presidential election

Independent variables:

1. gender ratio
2. percentage of young voters
3. percentage of older voters
4. income disparity
5. percentage employed in manufacturing
6. income
7. black population
8. people with at least a bachelor's degree
9. Hispanic population
10. foreign-born population
11. people with health insurance
12. voter turnout
13. the log of population density
14. votes for a third-party candidate.

Spatial unit： US counties

Methodology

- Spatial data visualization with GIS maps: County-level percentage vote for the Democratic Party, equal interval Graduated color(red to blue).

- Exploratory Data Analysis(EDA): Min, Mean, Median, Max, Std. Dev., Skewness, Kurtosis, Missing, Infinity, Distribution.

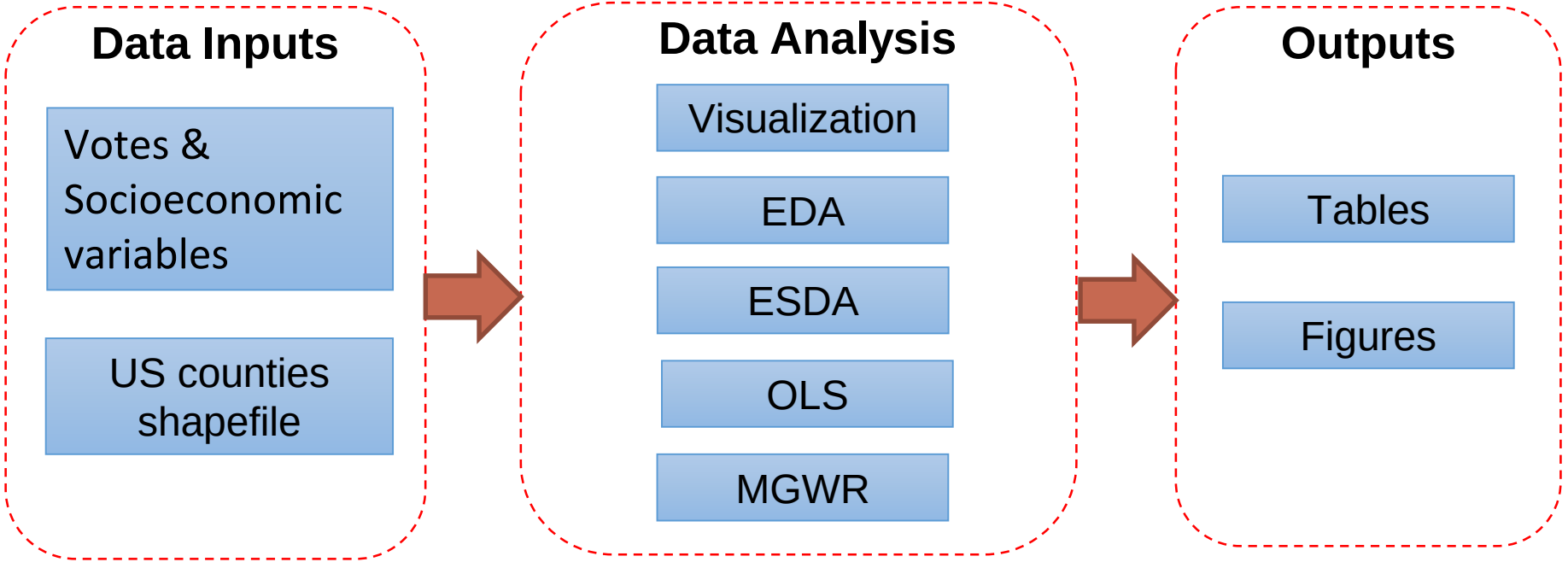
- Moran's I: $$I = \frac{N}{W} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

- Ordinary least squares (OLS): $$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\epsilon}$$

- Multiscale geographically weighted regression (MGWR):

$$y_i^* = \alpha_i + \sum_j \beta_{ij} x_{ij}^* + \varepsilon_i$$

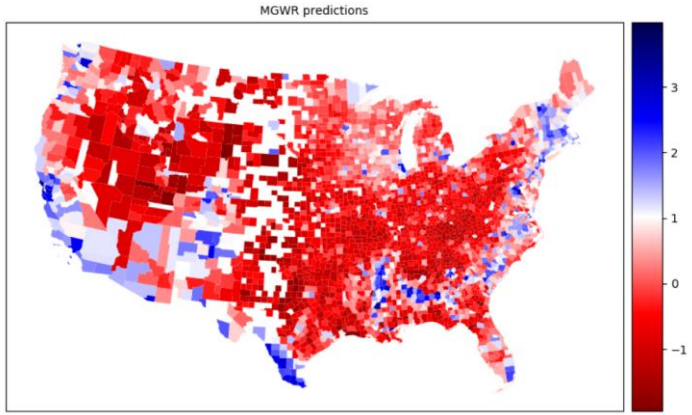
Flowchart of Data Analysis



MGWR

❑ Questions: :

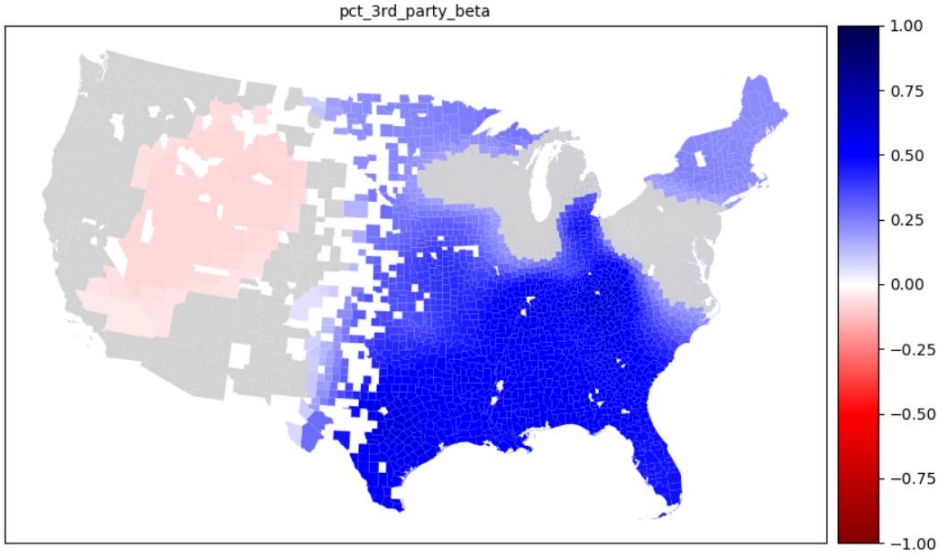
1. Is there spatial heterogeneity in the impact of socioeconomic factors on voting? How each variable affects voting locally?
2. Whether location other than socioeconomic factors have an impact on voting? How?



S variables	D bandwi...	D bandwi...	D bandwi...
Intercept_beta	43	43	44
sex_ratio_beta	603	600	696
pct_black_beta	43	43	45
pct_hisp_beta	543	446	543
pct_bach_beta	208	196	210
income_beta	2,659	2,158	2,659
pct_65_over...	656	656	696
pct_age_18_...	58	56	65
gini_beta	763	696	850
pct_manuf_b...	2,809	2,158	2,809
log_pop_den...	387	387	446
pct_3rd_part...	160	137	160
turn_out_beta	117	115	137
pct_FB_beta	1,424	1,350	1,754
pct_isured_b...	43	43	45

The impact of percentage third-party votes

- The impact of percentage third-party votes: Third-party candidates are often varied and have a regional, rather than national, attraction to voters. This is clearly seen in Figure, where across much of the eastern part of the country, and particularly in the Southeast, third-party candidates drew more heavily from Republican voters, increasing the Democratic vote share in a straight fight between Democrats and Republicans. The opposite is true in western states centered on Utah, where the third-party candidate drew more heavily from traditional Democratic voters and therefore increased the share of the Republican votes at the expense of Democratic votes.



Summary and discussion

❑ **Summary:**

- There is spatial heterogeneity in the impact of socioeconomic factors on voting. Variables have different effects on voting in different regions.
- Location factor other than socioeconomic factors are important for voting.

❑ **Discussion on further questions and expansions:**

- Tests on possible spatial autocorrelation in the results from MGWR models
- Explore potential impacts of geographical context
- Apply non-linear MGWR models

Instructions for running the workflow

Step 1: **Download data files** from \Spatial statistics workbench\01_data\election_data\election.csv, US_counites.shp

Step 2: **Download workflow** from \Spatial statistics workbench\02_workflow\MGWR.knwf

Step 3: **Open KNIME** (version 4.4.1) from local PC

Step 4: Install the following python packages if not yet:

- libpysal 4.4.0
- esda 2.3.6
- seaborn 0.11.2
- mgwr 2.1.2
- pandas 1.1.3
- matplotlib 3.2.2
- geopandas 0.9.0
- descartes 1.1.0
- pysal 1.14.4

Step 5: **Import KNIME workflow** file (MGWR.knwf)

Step 6: **Configure “CSV Input” & “Shapefile Input”**

Step 7: Click **Run** button on the top menu

Step 8: **Display the outputs** (see outputs folder)

List of Documents for the Sample Work

Documents	Description	Files
Knime workflows	1 workflow file	MGWR.knwf
Presentation	1 PPTX	MGWR.pptx
Data	1 data table in CSV	election.csv
	1 map in Shapefile	US_counties.shp
Output files	24 Output files	statistics.csv, accuracy.csv, bandwidth intervals.csv mgwr_summary.txt distribution.png, Moran's I.png, x1.png-x14.png, Contribution due to xs.png, Intercept.png, MGWR predictions.png, MGWR residuals.png

How to Join the Project?

- **Requirements:**
  - Self-motivation on spatial data science and spatial social sciences
  - Commitment of 3-6 hours per week on the project
  - Participation in regular online meetings on the project
  - Timing reports on the project by following the schedule
- **Tasks:**
  - Reading the references for the project
  - Identifying methodology for spatial data analysis
  - Preparing the data for case studies
  - Testing (or developing) nodes for the workbench
  - Writing the user guide
  - Drafting the PPT for project presentation
- **To Apply:**
  - Send a copy of CV with photo to spatialdatalab@lists.fas.harvard.edu
  - Join an online interview

What You Can Benefit From the Project?

- **Improvement:**
  - Advanced methodology
  - Cutting-edge Technology
  - Frontier applications
  - Enhanced visibility
- **Recognition:**
  - Certificate of completion
  - Project team member on the Lab website (http://spatialdatalab.org)
  - Academic recognition
  - Public recognition
- **Capacity:**
  - Infrastructure support
  - Networking support
  - Leadership support

Website and Contact

Project website:

http://spatialdatalab.org

Contact:

spatialdatalab@lists.fas.harvard.edu