Parallel PySAL

Autoregression and Complex Systems Framework Integration

Jason Laura, Robert Pahle, Sergio Rey, Luc Anselin

GeoDa Center for Geospatial Analysis and Computation
Arizona State University

August 18, 2014

Outline

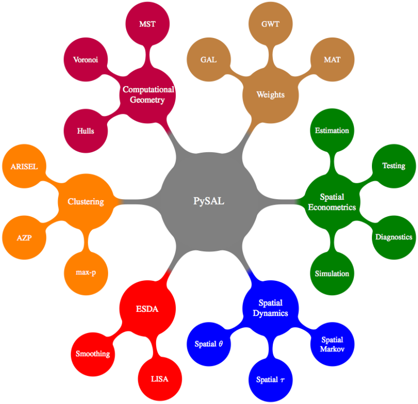
PySAL

Substantive Application: Spatial Econometrics

Implementation

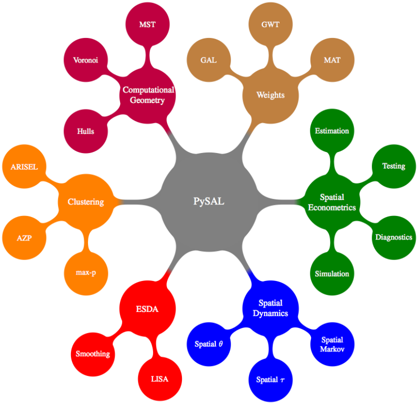
PySAL

- Spatial analysis library
- Big data world
- v 1.8 July 2014



pPySAL

- contiguity builder
- max-p region
- p-lisa
- fisher jenks
- spatial regimes



Lessons Learned

- Hardware dependence
- No holy grail of automatic parallelization
- Need a roadmap = Taxonomy
    - Guidance on "best practice"
    - Identify dead ends

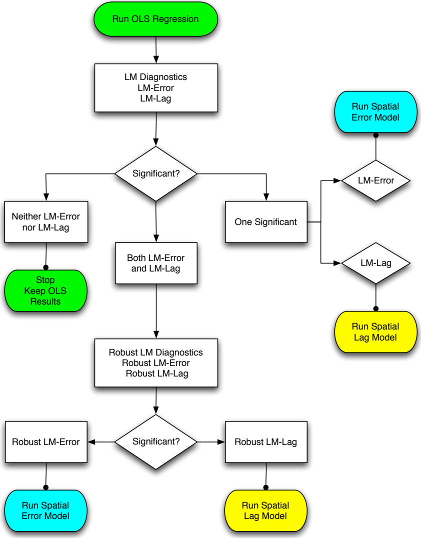
GeoDaSpace: Spatial Econometrics

- GUI ontop of spreg
- Subset of spreg functionality
- Cross-platform

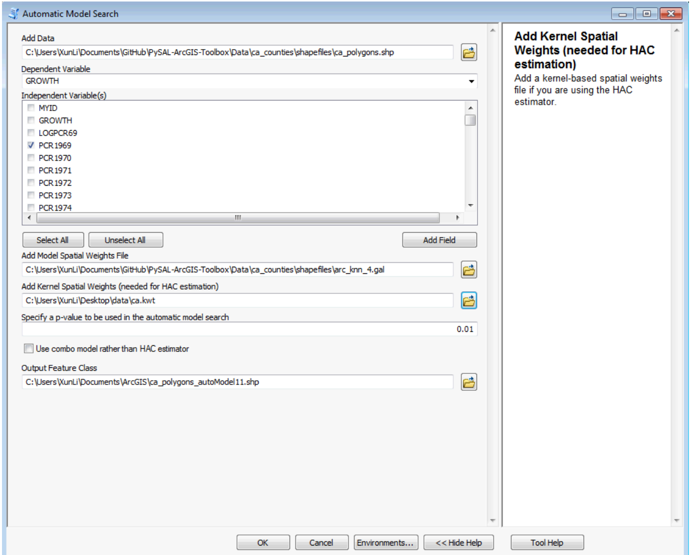
Specification Searches

- Specific to General
  - $y = X\beta + \epsilon$
  - OLS + Lagrange Multiplier Tests
- General to Specific
  - $y = \rho Wy + X\beta + (I - \lambda W)^{-1}\nu$
  - ML + Restrictions

LM Based Specification



ArcGIS Toolbox



Automatic Model Search

Add Data
C:\Users\XunLi\Documents\GitHub\PySAL-ArcGIS-Toolbox\Data\ca_counties\shapefiles\ca_polygons.shp

Dependent Variable
GROWTH

Independent Variable(s)
- □ MYID
- □ GROWTH
- □ LOGPCR69
- ☑ PCR1969
- □ PCR1970
- □ PCR1971
- □ PCR1972
- □ PCR1973
- □ PCR1974

Add Model Spatial Weights File
C:\Users\XunLi\Documents\GitHub\PySAL-ArcGIS-Toolbox\Data\ca_counties\shapefiles\arc_knn_4.gal

Add Kernel Spatial Weights (needed for HAC estimation)
C:\Users\XunLi\Desktop\data\ca.kwt

Specify a p-value to be used in the automatic model search
0.01

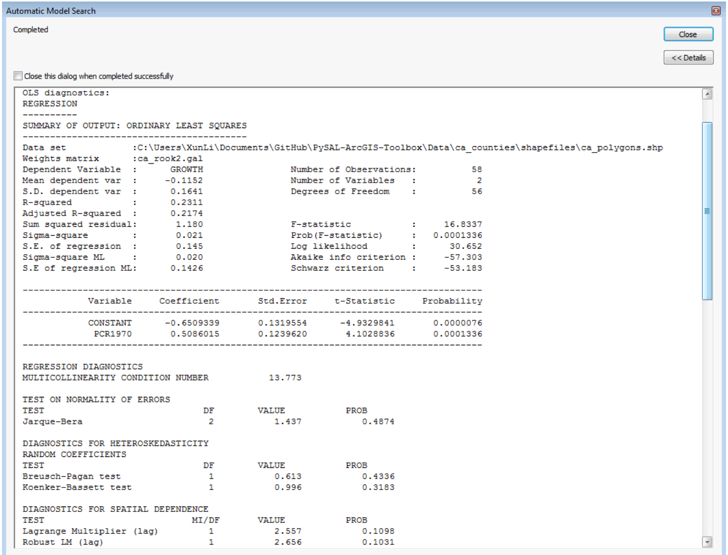
□ Use combo model rather than HAC estimator

Output Feature Class
C:\Users\XunLi\Documents\ArcGIS\ca_polygons_autoModel11.shp

Add Kernel Spatial Weights (needed for HAC estimation)

Add a kernel-based spatial weights file if you are using the HAC estimator.

ArcGIS Toolbox



```
OLS diagnostics:
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES
-----------------------------------------
Data set            :C:\Users\XunLi\Documents\GitHub\PySAL-ArcGIS-Toolbox\Data\ca_counties\shapefiles\ca_polygons.shp
Weights matrix      :ca_rook2.gal
Dependent Variable  :      GROWTH                Number of Observations:          58
Mean dependent var  :     -0.1152                Number of Variables   :           2
S.D. dependent var  :      0.1641                Degrees of Freedom    :          56
R-squared           :      0.2311
Adjusted R-squared  :      0.2174
Sum squared residual:       1.180                F-statistic           :     16.8337
Sigma-square        :       0.021                Prob(F-statistic)     :   0.0001336
S.E. of regression  :       0.145                Log likelihood        :      30.652
Sigma-square ML     :       0.020                Akaike info criterion :     -57.303
S.E of regression ML:      0.1426                Schwarz criterion     :     -53.183

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     t-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT      -0.6509339       0.1319554      -4.9329841       0.0000076
             PCR1970       0.5086015       0.1239620       4.1028836       0.0001336
------------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER           13.773

TEST ON NORMALITY OF ERRORS
TEST                             DF        VALUE           PROB
Jarque-Bera                       2           1.437           0.4874

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                             DF        VALUE           PROB
Breusch-Pagan test                1           0.613           0.4336
Koenker-Bassett test              1           0.996           0.3183

DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST                           MI/DF       VALUE           PROB
Lagrange Multiplier (lag)         1           2.557           0.1098
Robust LM (lag)                   1           2.656           0.1031
```

```
Root Node: Ordinary Least Squares Regresion

Then

A. If Lagrange Multiplier Test for Spatial Error Model < p-value AND Lagrange Multiplier Test for Spatial Lag Model < p-value

    1. If Robust Lagrange Multiplier Test for Spatial Error p-value < p-value and Robrust Lagrange Multiplier Test for Spatial Lag Model p-value < p-vlaue:
        a. If NOT combo
            i. twosls_sp.GM_Lag
            ii. "Spatial Lag with Spatial Error - HAC"
        b. Elif Koenker Basset Statistic p-value < p-value
            i. error_sp_het.GM_Combo_Het
            ii. "Spatial Lag with Spatial Error - Heteroskedastic"
        c. Else
            i. error_sp_hom.GM_Combo_Hom
            ii. "Spatial Lag with Spatial Error - Homoskedastic"

    2. Else If Robust Lagrange Multiplier Test for Spatial Error p-value < p-value and RLM for Spatial Lag p-value > p-value:
        a. If OLS Koenker Basset Statistic p-value < p-value
            i. error_sp_het.GM_Error_Het
            ii. "Spatial Error - Heteroskedastic"
        b. Else If OLS Koenker Basset Statistic p-value > p-value
            i. error_sp_hom.GM_Error_Hom
            ii. "Spatial Error - Homoskedastic"

    3. Else If RLM for Spatial Error > p-value and RLM for Spatial Lag < p-value
        a. If OLS Koenker Basset Statistic p-value < p-value
            i. twosls_sp.GM_Lag (robut:white)
            ii. "Spatial Lag - Heteroskedastic"
        b. Else If OLS Koenker Basset Statistic p-value > p-value
            i. twosls_sp.GM_Lag
            ii. "Spatial Lag - Homoskedastic"
    4. Else If RLM for Spatial Error > p-value and RLM for Spatial Lag > p-value
        a. No PySAL Call
        b. No Model - Robust Test not Significant - Check Model.

B. Else If Lagrange Mutiplier Test for Spatial Error Model < p-value AND Lagrange Mutiplier Test for Spatial Lag > p-value
    1. If OLS Koenker Basset Statistic p-value < p-value
        i. error_sp_het.GM_Error_Het
        ii. "Spatial Error - Heteroskedastic"
    2. Else If OLS Koenker Basset Statistic p-value > p-value
        i. error_sp_hom.GM_Error_Hom
        ii. "Spatial Error - Homoskedastic"

C. Else If Lagrange Multiplier Test for Spatial Error Model > p-value AND Lagrange Mutliplier Test for Spatial Lag < p-value
    1. If OLS Koenker Basset Statistic p-value < p-value
        i. twosls_sp.GM_Lag (robust-white)
        ii. "Spatial Lag - Heteroskedastic"
    2. Else If OLS Koenker Basset Statistic p-value > p-value
        i. twosls_sp.GM_Lag
        ii. "Spatial Lag - Homoskedastic"

D. Else Lagrange Multiplier Test for Spatial Error Model > p-value AND Lagrange Multiplier Test for Spatial Lag > p-value

    1. If OLS Koenker Basset Statistic p-value < p-value
        i. ols.OLS (robust-white)
        ii. "No Space - Heteroskedastic"
    2. Else If OLS Koenker Basset Statistic p-value > p-value
        i. ols.OLS
        ii. "No Space - Homoskedastic"
```

Parallel Strategy

- Speculative Parallelism
  - Solve' all branches of a search tree
  - Leverage an excess computation model
  - No dependency in execution order
  - Synchronization at the completion of all computation
- Implementation
  - Utilize a processing queue
  - One manager, and n workers
  - Workers draw a regression model from the queue, process, and return the result
  - Scales to where n = number of models to compute
  - Potential to extend to variable parameter specification (larger tree)

Tensions

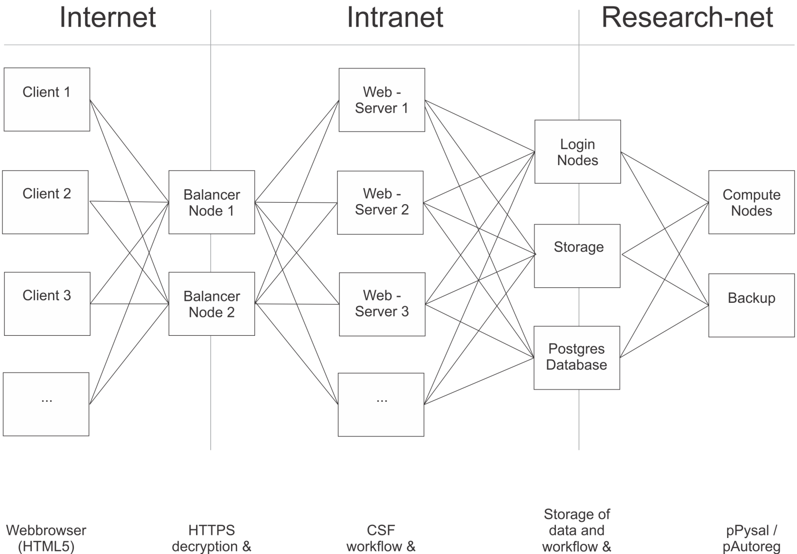
Trade-off

- Trading elegant econometric theory for data mining
- Gain speed and coverage of model space over the sequential approach

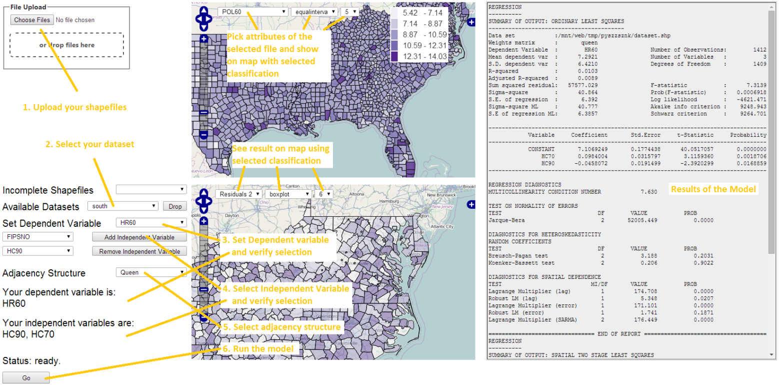
Issues

- Distributional properties of big data approach unknown
- Purists take a dim view of “data mining”

Complex Systems Framework



Autoreg in CSF



File Upload
Choose Files No file chosen
or drop files here

1. Upload your shapefiles

2. Select your dataset

Incomplete Shapefiles
Available Datasets: south ▾ Drop
Set Dependent Variable: HR60 ▾
FIPSNO ▾ Add Independent Variable
HC90 ▾ Remove Independent Variable
Adjacency Structure: Queen ▾

Your dependent variable is:
HR60

Your independent variables are:
HC90, HC70

Status: ready.

Go

Pick attributes of the selected file and show on map with selected classification

See result on map using selected classification

3. Set Dependent variable and verify selection

4. Select Independent Variable and verify selection

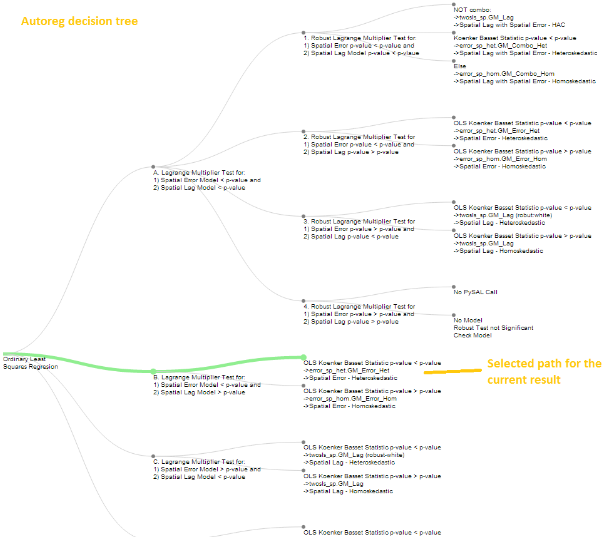
5. Select adjacency structure

6. Run the model

Results of the Model

```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES
-----------------------------------------
Data set            :/mnt/web/tmp/pyszzsznk/dataset.shp
Weights matrix      :       queen
Dependent Variable  :        HR60                Number of Observations:        1412
Mean dependent var  :      7.2921                Number of Variables   :           3
S.D. dependent var  :      6.4210                Degrees of Freedom    :        1409
R-squared           :      0.0103
Adjusted R-squared  :      0.0089
Sum squared residual:   57577.029                F-statistic           :      7.3139
Sigma-square        :      40.864                Prob(F-statistic)     :   0.0006918
S.E. of regression  :       6.392                Log likelihood        :   -4621.471
Sigma-square ML     :      40.777                Akaike info criterion :    9248.943
S.E of regression ML:      6.3857                Schwarz criterion     :    9264.701

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     t-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT       7.1069249       0.1774438      40.0517057       0.0000000
                HC70       0.0984004       0.0315797       3.1159360       0.0018706
                HC90      -0.0458072       0.0191499      -2.3920299       0.0168859
------------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER           7.630

TEST ON NORMALITY OF ERRORS
TEST                             DF        VALUE           PROB
Jarque-Bera                       2       52005.449           0.0000

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                             DF        VALUE           PROB
Breusch-Pagan test                2           3.188           0.2031
Koenker-Bassett test              2           0.206           0.9022

DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST                           MI/DF       VALUE           PROB
Lagrange Multiplier (lag)         1         174.708           0.0000
Robust LM (lag)                   1           5.348           0.0207
Lagrange Multiplier (error)       1         171.101           0.0000
Robust LM (error)                 1           1.741           0.1871
Lagrange Multiplier (SARMA)       2         176.449           0.0000

================================ END OF REPORT =====================================
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL TWO STAGE LEAST SQUARES
```

Model Path



Autoreg decision tree

Ordinary Least
Squares Regresion

A. Lagrange Multiplier Test for:
1) Spatial Error Model < p-value and
2) Spatial Lag Model < p-value

1. Robust Lagrange Multiplier Test for:
1) Spatial Error p-value < p-value and
2) Spatial Lag Model p-value < p-vlaue

NOT combo:
->twosls_sp.GM_Lag
->Spatial Lag with Spatial Error - HAC

Koenker Basset Statistic p-value < p-value
->error_sp_het.GM_Combo_Het
->Spatial Lag with Spatial Error - Heteroskedastic

Else
->error_sp_hom.GM_Combo_Hom
->Spatial Lag with Spatial Error - Homoskedastic

2. Robust Lagrange Multiplier Test for
1) Spatial Error p-value < p-value and
2) Spatial Lag p-value > p-value

OLS Koenker Basset Statistic p-value < p-value
->error_sp_het.GM_Error_Het
->Spatial Error - Heteroskedastic

OLS Koenker Basset Statistic p-value > p-value
->error_sp_hom.GM_Error_Hom
->Spatial Error - Homoskedastic

3. Robust Lagrange Multiplier Test for
1) Spatial Error p-value > p-value and
2) Spatial Lag p-value < p-value

OLS Koenker Basset Statistic p-value < p-value
->twosls_sp.GM_Lag (robut:white)
->Spatial Lag - Heteroskedastic

OLS Koenker Basset Statistic p-value > p-value
->twosls_sp.GM_Lag
->Spatial Lag - Homoskedastic

4. Robust Lagrange Multiplier Test for
1) Spatial Error p-value > p-value and
2) Spatial Lag p-value > p-value

No PySAL Call

No Model
Robust Test not Significant
Check Model

B. Lagrange Multiplier Test for:
1) Spatial Error Model < p-value and
2) Spatial Lag Model > p-value

OLS Koenker Basset Statistic p-value < p-value
->error_sp_het.GM_Error_Het
->Spatial Error - Heteroskedastic

Selected path for the
current result

OLS Koenker Basset Statistic p-value > p-value
->error_sp_hom.GM_Error_Hom
->Spatial Error - Homoskedastic

C. Lagrange Multiplier Test for:
1) Spatial Error Model > p-value and
2) Spatial Lag Model < p-value

OLS Koenker Basset Statistic p-value < p-value
->twosls_sp.GM_Lag (robust-white)
->Spatial Lag - Heteroskedastic

OLS Koenker Basset Statistic p-value > p-value
->twosls_sp.GM_Lag
->Spatial Lag - Homoskedastic

OLS Koenker Basset Statistic p-value < p-value

Next Steps

Parallel Autoreg

- Ensemble of search strategies
  - short
  - full
  - hybrid
- Candidate Variables
- Candidate Ws

Integration

- CyberGIS Gateway
- Strategies

Come see the demo!