Parallel PySAL

Autoregression and Complex System Framework Integration

Jason Laura, Robert Pahle, Sergio Rey, Luc Anselin

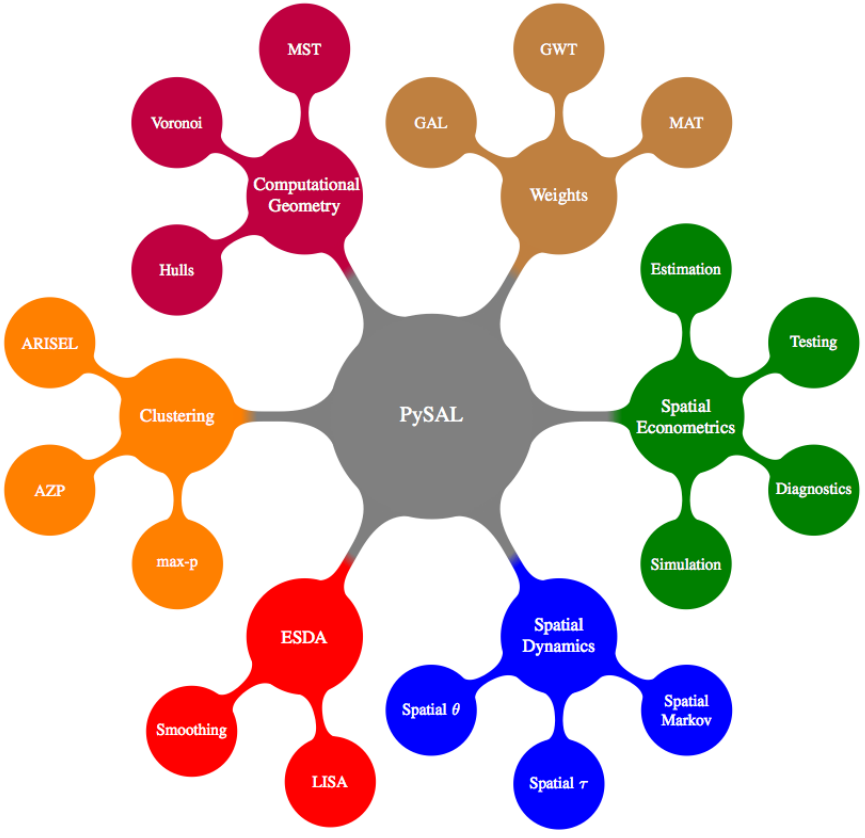
GeoDa Center for Geospatial Analysis and Computation
Arizona State University

August 17, 2014

Outline

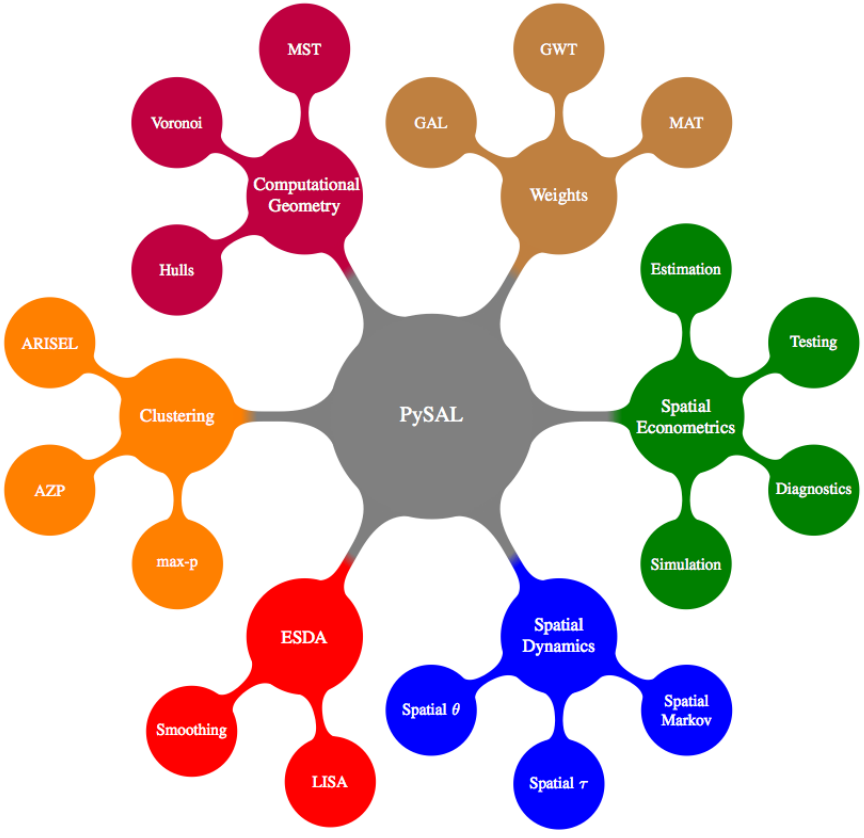
PySAL

- 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

Spatial Econometrics

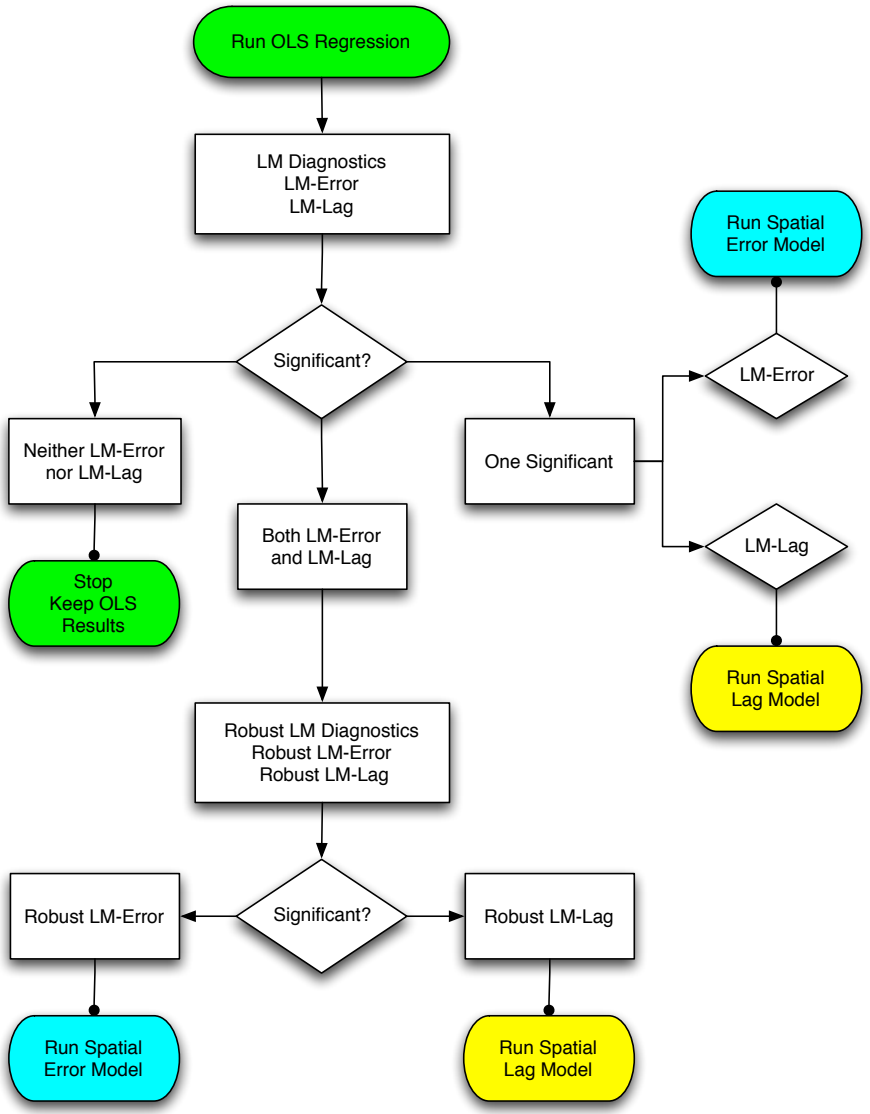
GeoDaSpace

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

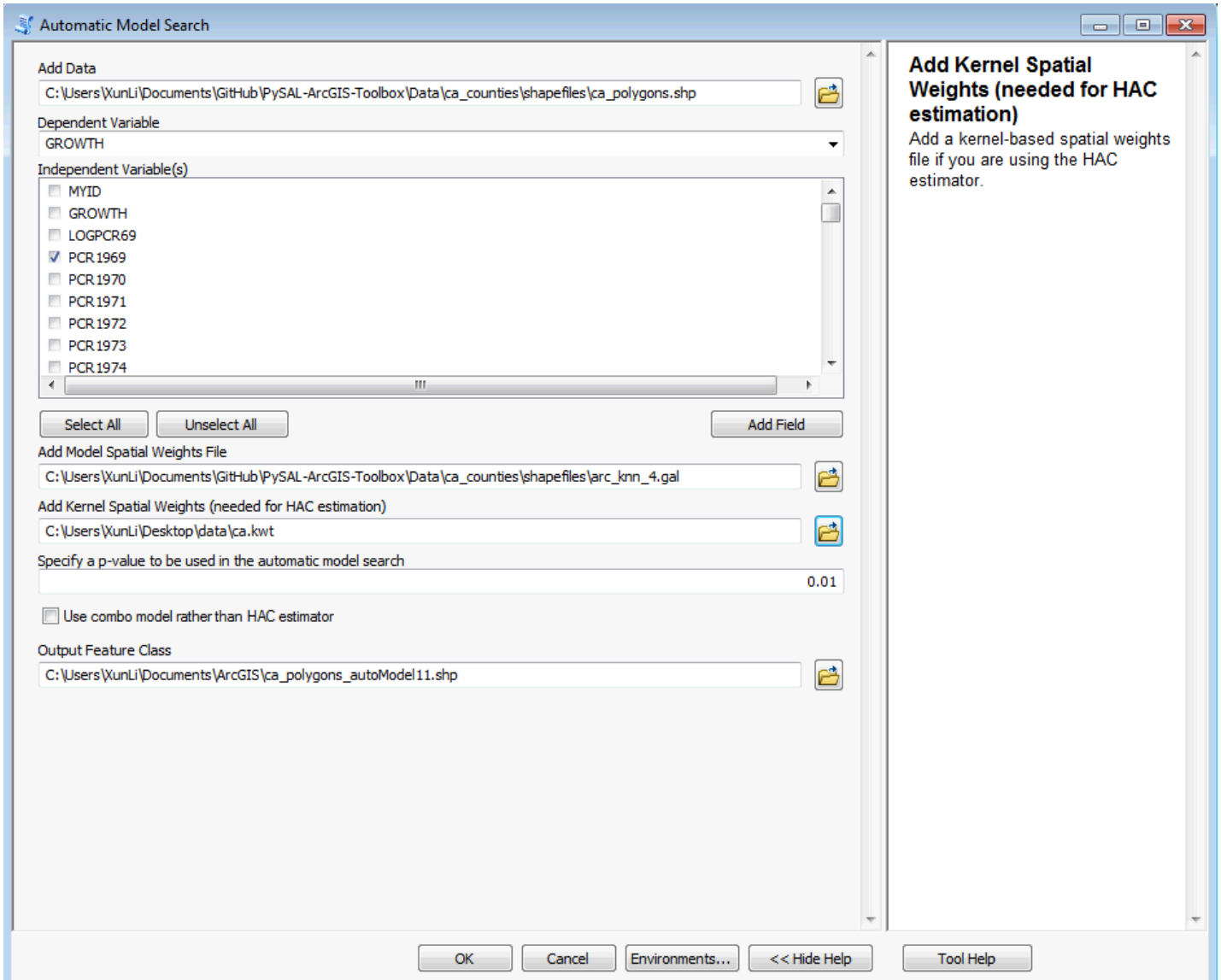
Specification Searches

- 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

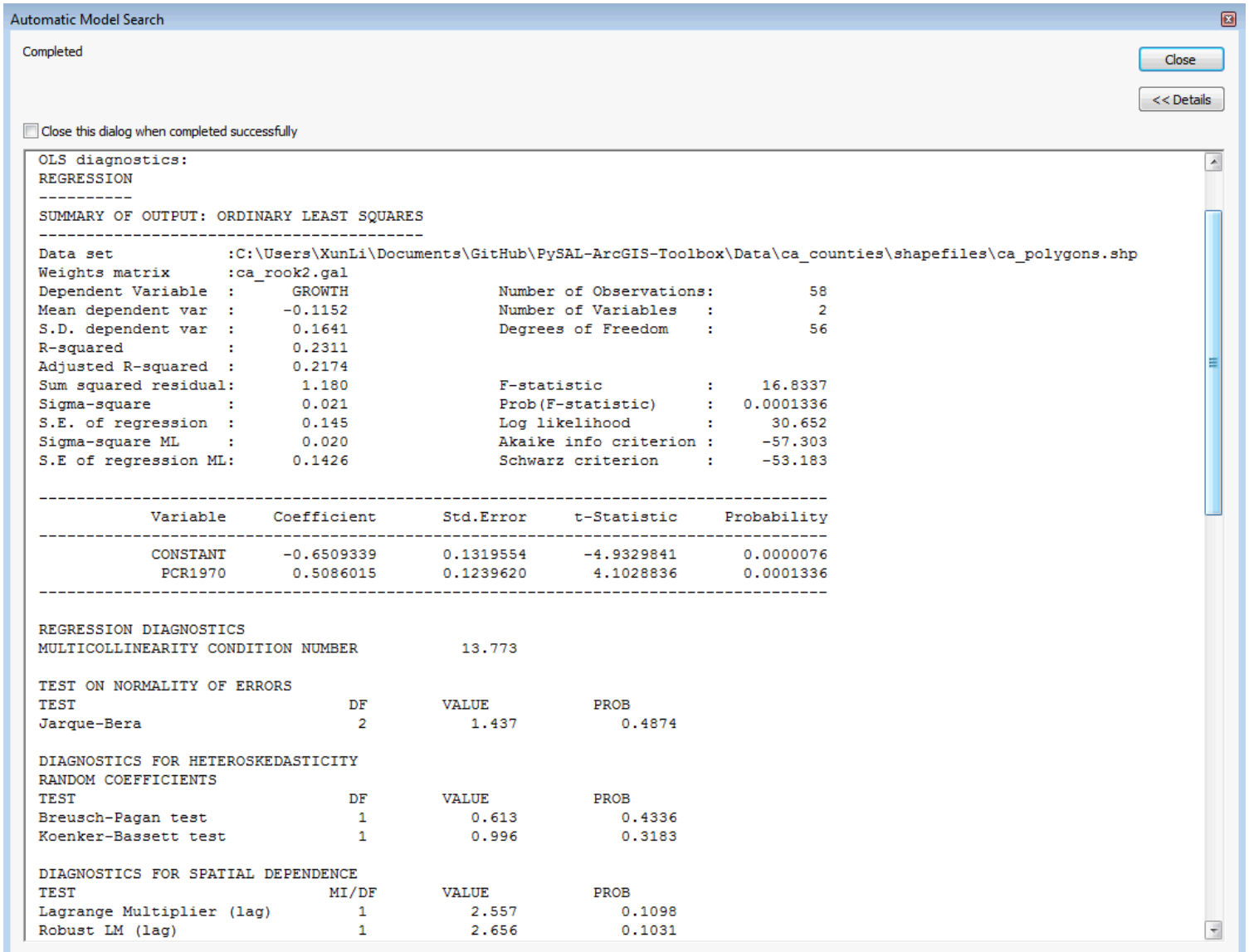
LM Based Specification



ArcGIS Toolbox



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"

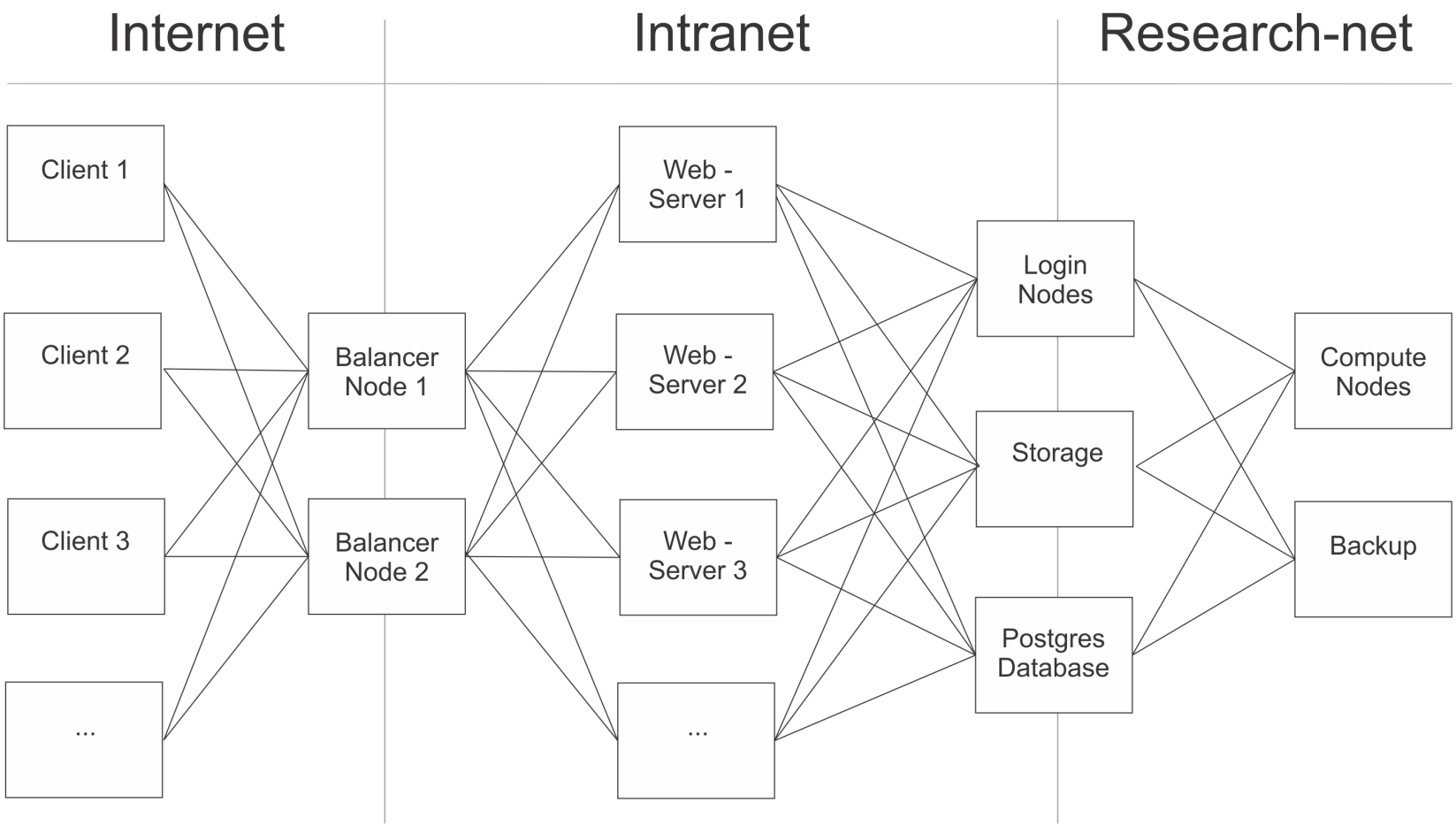
D. Else Lagrange Multiplier Test for Spatial Error Model > p-value AND Lagrange Multiplier Test for Spatial Lag > p-value

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"

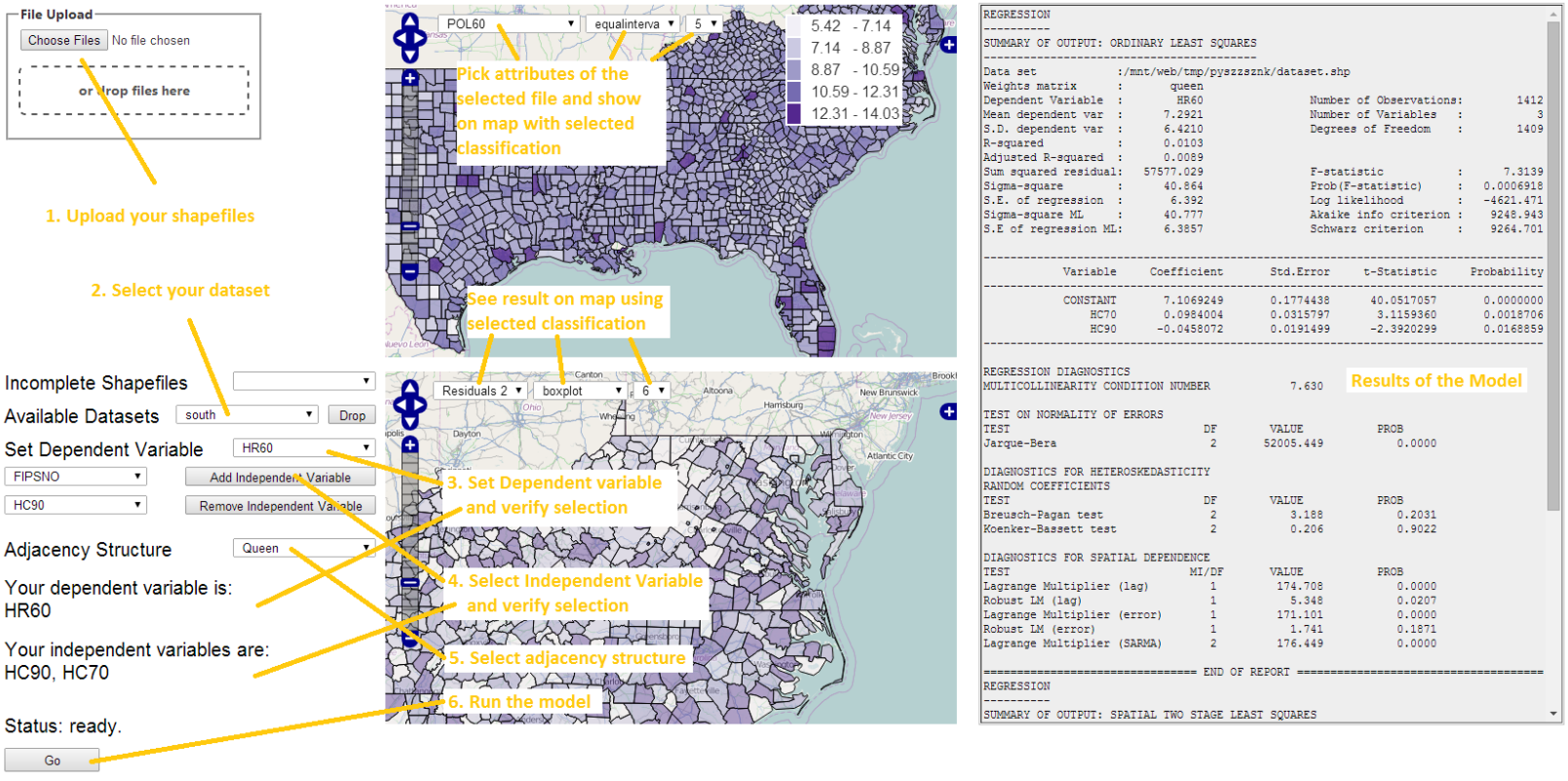
Parallel Strategy

- 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)

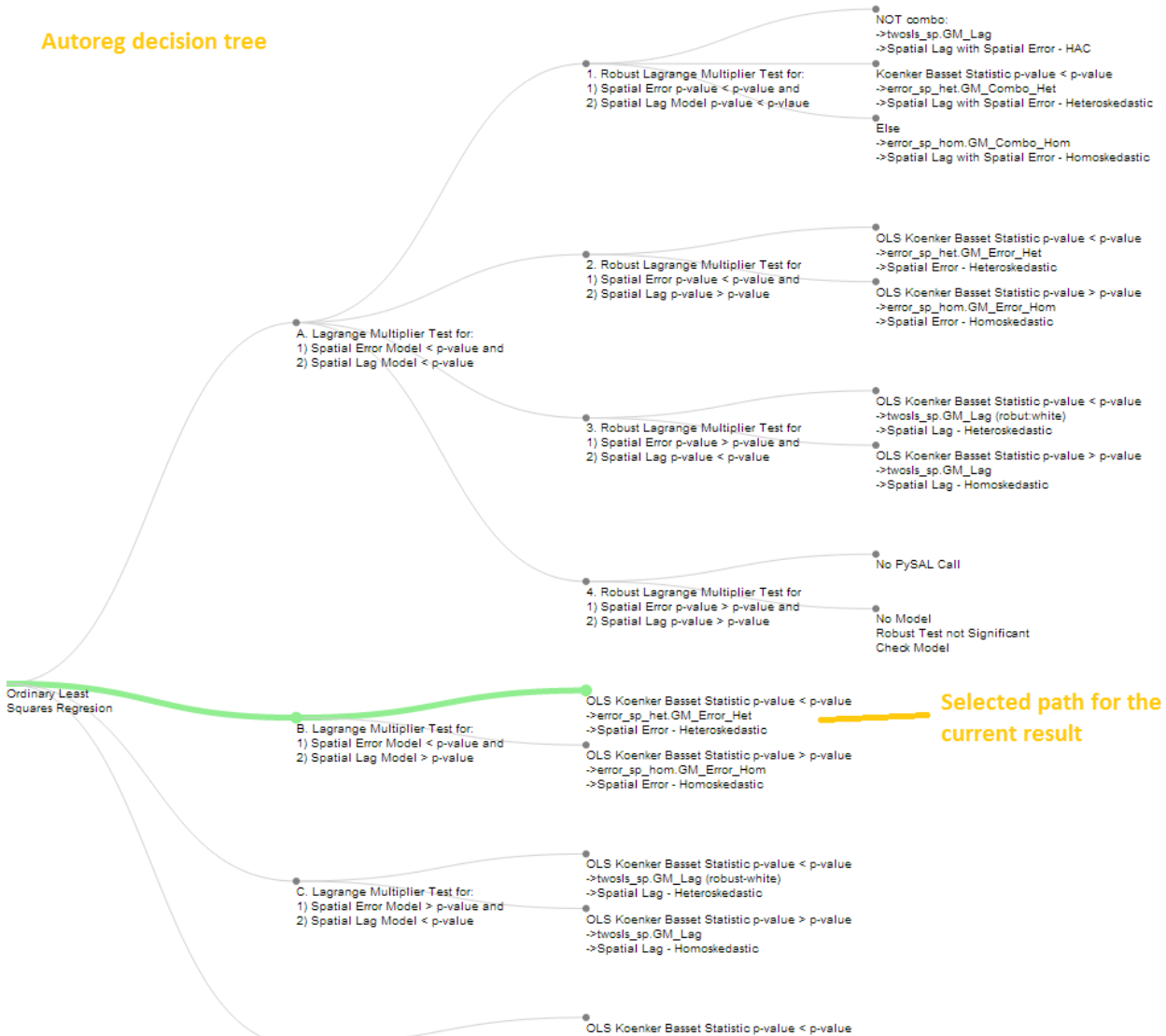
Complex Systems Framework



Autoreg in CSF



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
OLS Koenker Basset Statistic p-value > p-value
->error_sp_hom.GM_Error_Hom
->Spatial Error - Homoskedastic
Selected path for the current result
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!