Spatial Analysis in CyberGIS

towards a spatial econometrics workbench

Luc Anselin, Sergio Rey and Myunghwa Hwang
GeoDa Center
School of Geographical Sciences and Urban Planning
Arizona State University
• Acknowledgment

• NSF Award OCI-1047916: CyberGIS Software Integration for Sustained Geospatial Innovation
• Team

• Luc Anselin
• Sergio J. Rey
• Robert Pahle
• Myunghwa Hwang
• Philip Stephens
• Xing Kang
Goals
towards cyberinfrastructure in support of spatial analysis
(Goodchild, 2010)
• Some Specific Project Goals

• Integrate and sustain a core set of composable, interoperable, manageable, and reusable CyberGIS software elements based on community-driven and open source strategies
Some Specific Project Goals (2)

- Empower high-performance and scalable CyberGIS by exploiting spatial characteristics of data and analytical operations for achieving unprecedented capabilities for geospatial knowledge discovery
• Challenge

• most current spatial analysis/spatial econometrics software written for single CPU

• rethink and rewrite analytical, algorithmic and processing facilities to integrate into a cyberinfrastructure

• lack of interoperability
• Spatial Econometrics Workbench
  • framework for supporting spatial econometric research in a cyberscience era (Anselin and Rey, IJGIS 2012)
  • support for scientific workflow
  • leverages PySAL
Leveraging PySAL
• PySAL

• open source library of Python routines for spatial analysis: geocomputation, spatial weights, spatial autocorrelation, spatial econometrics

• http://pysal.org

• hosted on google code
Welcome

PySAL is an open source cross-platform library of spatial analysis functions written in Python. It is intended to support the development of high level applications for spatial analysis.

Documentation

PySAL 1.4, released 2012 07 31

All Versions

- Unreleased development version
- PySAL 1.3, released 2012 01 31
- PySAL 1.2, released 2011 07 31
- PySAL 1.1, released 2011 01 31
- PySAL 1.0, released 2010 08 01

News

2012-07-31 PySAL 1.4 Stable (Downloads)
2012-01-31 PySAL 1.3 released
2012-01-19 PySAL 1.3 release code sprint on January 23
• PySAL Functions for CyberGIS (selected examples)

• spatial weights

• global/local spatial autocorrelation

• simulation estimators
• Migrating to CyberGIS

  • provide functionality as web services

  • performance = need for parallelization + refined algorithms

  • interoperability = need for metadata and provenance tracking
Web Services
• PySAL Web Services
  • wrapping PySAL code as services
  • providing access through portal
  • infrastructure = link between UIUC and ASU Decision Theater Cluster
  • prototype: spatial weights operations
Service Implementation

Client Tier

- Web Application
- Command Line Shell
- Desktop GIS Extensions
- Other Clients

Server Tier

- Apache2 modwsgi
- Python django
- PySAL
- DT Cluster
Integration with Token Service

1. Get a token
2. Token
3. Request a service with token, domain & remote_addr
4. Verify a token
5. Token Verification
6. Process the request
7. Results

User

Token Service
(UIUC Cluster)

PySAL Web Services
(DT Cluster)
Web 2.0 Interface to PySAL Services
• **Command-Line Access**

• open to cyberGIS gateway users

• authentication: gateway token

• **Examples:**

```bash
./kernelW.py -g ohio.zip -d 2 -f false -k 5 -t quadratic -i FIPSNO
./knnW.py -g ohio.zip -d 2 -k 5 -i FIPSNO
./lag.py -d ohio.dbf -y POPMW68 -w queen.gal
./moran.py -d ohio.dbf -y POPMW68 -w queen.gal -m global -s 99
```
Parallelization
• Why?

  • real-time analysis in a decision support environment

  • 8 second rule

  • spatial econometric methods: spatial weights manipulations, simulation estimators
• Approach

• focused on PySAL and use of Python parallelization modules

• multicore and GPU, clusters (ongoing)

• available Python libraries: pyopencl, multiprocessing, parallel python

• not all spatial analysis methods suitable for parallelization
• Initial Results

• embarrassingly parallel: permutation inference (simulations), local statistics, simulation estimators

• not so straightforward: sequential analysis combined with parallel

• experimentation needed: results not always as expected

no obvious general results
• Some Examples

• implementing parallel Lisa (Local Moran)

• Fisher-Jencks map classification

• regionalization (contiguity-constrained clustering)
Example: Fisher-Jencks Map Classification

Comparative Speedups k=7

Ts/TP vs n

- PyOpenCL
- Multiprocessing
- PP
Provenance and Standards
• Scientific Workflow

• data provenance

• model provenance

• lineage of data sources, manipulation and analytical operations (DAG)

• enable replication
• Specialized Needs

• spatial operations combine data with processing

• requires metadata for attribute data

• requires metadata for operations
Example: Spatial Weights

- provenance = lineage

- includes spatial data source, type of weights (e.g., contiguity, distance), any standardization or manipulation (e.g., higher order)
Example: PySAL spgreg

what do we know about south_k6.gwt and south_ep_k20.kwt
• Early Attempts

• header line in gal and gwt files

• “standard” initially in SpaceStat and later adopted by R, GeoDa, PySAL, etc.

• insufficient
• Taxonomy of Weights Formats in PySAL

```python
>>> pysal.open.check()
PySAL File I/O understands the following file extensions:
Ext: '.shp', Modes: ['r', 'wb', 'w', 'rb']
Ext: '.mtx', Modes: ['r', 'w']
Ext: '.swm', Modes: ['r', 'w']
Ext: '.mat', Modes: ['r', 'w']
Ext: '.shx', Modes: ['r', 'wb', 'w', 'rb']
Ext: '.stata_text', Modes: ['r', 'w']
Ext: '.geoda_txt', Modes: ['r']
Ext: '.dbf', Modes: ['r', 'w']
Ext: '.dat', Modes: ['r', 'w']
Ext: '.gwt', Modes: ['r', 'w']
Ext: '.gal', Modes: ['r', 'w']
Ext: '.arcgis_text', Modes: ['r', 'w']
Ext: '.wk1', Modes: ['r', 'w']
Ext: '.arcgis_dbf', Modes: ['r', 'w']
Ext: '.geobugs_text', Modes: ['r', 'w']
Ext: '.csv', Modes: ['r']
Ext: '.wkt', Modes: ['r']
```
• Conceptual Framework

• separate data source from operations

• data source: polygon or coordinate files with standard metadata (projection, origin, etc.)

• operations: no metadata yet (initial work in OGC WPS etc. but not for analysis)
• Operations

  • creating the network structure
    0-1 connectivity graph

  • computing the weights values
• Prototype Weights Metadata

• content model = vocabularies to describe spatial weights provenance and operations

• representation = XML Schema
Core Operations
<xs:complexType name="CreationOperationType" abstract="true">
  <xs:complexContent>
    <xs:restriction base="wgt:WeightsOperationType">
      <xs:element ref="wgt:DataSource" />
    </xs:restriction>
  </xs:complexContent>
</xs:complexType>

<xs:complexType name="ContiguityOperationType">
  <xs:complexContent>
    <xs:restriction base="wgt:CreationOperationType">
      <xs:sequence>
        <xs:element name="Type" type="xs:string" />
        <xs:element name="Order" type="xs:int" />
        <xs:element name="LowerOrderNeighbors" type="xs:boolean" />
      </xs:sequence>
    </xs:restriction>
  </xs:complexContent>
</xs:complexType>
<?xml version="1.0" ?>
    xmlns:weights="http://www.swm.org/2011/weights"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
    <weights:WeightsOperation xsi:type="operations:ContiguityOperationType">
        <weights:TimeStamp>2012-01-18T00:46:11</weights:TimeStamp>
        <weights:Software>
            <weights:Name>GeoDa Center Weights Service</weights:Name>
            <weights:Version>0.1</weights:Version>
        </weights:Software>
        <weights:DataSource>
            <weights:TemporalCoverage>
                <weights:StartTime>2010-01-01T12:00:00</weights:StartTime>
                <weights:EndTime>2010-07-01T12:00:00</weights:EndTime>
            </weights:TemporalCoverage>
            <weights:SpatialCoverage>
                <weights:PlaceName>United States, U.S., State or Equivalent Entity, Arizona, AZ, 04</weights:PlaceName>
                <weights:GeographicExtent>
                    <weights:MinX>-114.816591</weights:MinX>
                    <weights:MinY>31.332177</weights:MinY>
                    <weights:MaxX>-109.045223</weights:MaxX>
                    <weights:MaxY>37.00426</weights:MaxY>
                </weights:GeographicExtent>
                <weights:GeometryType>Polygon</weights:GeometryType>
                <weights:GeographicCoordinateSystem>
                    ...
                </weights:GeographicCoordinateSystem>
                <weights:Scale>Census Tract, Tract</weights:Scale>
            </weights:SpatialCoverage>
            <weights:DataLocation>/[Path]/arizona_tracts_2010/t1_2010_04_tract10.shp</weights:DataLocation>
            <weights:DataType>ESRI Shape File</weights:DataType>
            <weights:IdVariable>GEOID10</weights:IdVariable>
            <weights:NumberOfObservations>1526</weights:NumberOfObservations>
        </weights:DataSource>
        <operations:Type>Rook</operations:Type>
        <operations:Order>1</operations:Order>
        <operations:LowerOrderNeighbors>true</operations:LowerOrderNeighbors>
    </weights:WeightsOperation>
</weights:WeightsMetadata>
Moving Forward

- automatic generation of provenance information
- automatic retrieval
- semantic interoperability
Roadmap
• Methodological Refinements

• parallelizing spatial methods and algorithm improvement

• spatial probit/tobit

• space-time methods (space-time scan statistic)
• Operationalization

• web services / portal

• weights metadata

• automatic provenance generation

• spatial econometrics workbench
• Community Building

• demonstrate usefulness

• demonstrate added value

• training and documentation

• facilitate participation