Proposal modeled after previous GSoC PySAL applications (2016, 2018)

PySAL ESDA Enhancements: Local join count and LOSH statistics

Sub-org name

Sub-organization: PySAL

Student info

- ***Name:*** Jeffery (Jeff) Sauer
- ***GitHub:*** jeffcsauer
- ***Gitter:*** jeffcsauer
- ***CV:*** recent copy of CV
- ***Email:*** jcsauer@terpmail.umd.edu
- ***Skype:*** jeffery.sauer1
- ***Twitter:*** @jcsauer_geo
- ***Time Zone:*** Eastern Standard Time (UTC - 5)

University info

- ***University Name:*** University of Maryland, College Park (USA)
- ***Research interests:*** health geography, spatial analysis, geographic data science
- ***Research project:*** Analysis of the U.S. opioid epidemic across geographic scales.
- ***Degree:*** PhD (started Fall 2019)

Code contribution

PySAL contributions:

- Fixed rendering issue for opening citation in `esda` GeoSilhouettes notebook (pull/110)
- Added citations for five example datasets (baltim, burkitt, chicago, tokyo, snow) and updated links for two example datasets in `libpysal` (pull/265)
- Raised issue and suggested fix for module import difficulties with Jupyter Book 'Interact' feature (issues/88). Later realized these would be solved by an under-review pull request from core developer (pull/86). Currently reviewing the recent update of PySAL Notebooks.

Non-pysal contributions to open source:

- arcospy: translation of the R `arcos` package into Python for cross-software functionality

Project information

Proposal Title: "PySAL ESDA Enhancements: Local join count and LOSH statistics"

Proposal Abstract:

The goal of this project is to add several recently developed spatial estimators to the exploratory spatial data analysis (`esda`) submodule of PySAL, the Python Spatial Analysis Library. This project will allow researchers to easily deploy these estimators in existing spatial workflows. Specifically, this project will contribute implementation, docstrings, tests, and example notebooks for (1) bivariate local join count statistics, (2) multivariate local join count statistics, and (3) local spatial heteroskedasticity (LOSH) statistics. In the first phase of this project, each estimator will be reviewed and pseudo-coded to identify areas of optimization. In the second phase, the estimators will be implemented with tests, with performance assessed against spatial datasets of different sizes. In the final phase, example notebooks will be drafted and polished to a quality ready for external (e.g. workshop) use.

Proposal Description:

The `esda` submodule offers nearly 20 local autocorrelation, global autocorrelation, and diagnostic spatial statistics for areal data that are routinely used in spatial analysis and the emergent field of geographic data science.(Anselin, 1996; Singleton & Arribas-Bel, 2019) To provide cutting-edge functionality to its userbase, core members of the PySAL development team have identified a list of recently developed spatial statistics for implementation (issues/61). The goal of this project is to contribute several of these priority statistics to the `esda` submodule, specifically (1) bivariate local join count statistics, (2) multivariate local join count statistics, and (3) local spatial heteroskedasticity (LOSH) statistics.

Bivariate join count statistics are a set of recently published statistics offering insight into the local autocorrelation of two or more binary variables of areal data.(Anselin & Li, 2019) These statistics are useful in examining whether areal units and surrounding neighbors do or do not have the same binary value, with potential application in a wide variety of disciplines including urban planning, spatial epidemiology, environmental sciences and more. Anselin & Li's local join count statistics offer analytical flexibility to both regular and irregular areal layouts and statistical inference through established conditional permutation tests, the latter having precedent in the `esda` submodule.(Anselin 1995) These statistics should not be confused with join counts, which are already implemented in `esda`.(Cliff & Ord, 1981)

LOSH statistics, *H*, offer information about the variance of a local spatial process.(Ord & Getis, 2012) As many local autocorrelation statistics are concerned with means, LOSH is the natural complement that focuses on variance. Users of the PySAL LOSH implementation will be able to interpret *H* in conjunction with measures of local autocorrelation already implemented in PySAL, such as the *Gi* statistic.(Getis & Ord, 1992) Additionally, users carrying out cluster analysis may

use LOSH to examine variance within or outside clusters, as well as investigate what Getis & Ord call 'transitional regions' for units that delimit clusters.(Ord & Getis, 2012)

Each of the statistics will be implemented with docstrings, tests, and example notebooks. In addition to following PySAL development guidelines, providing this robust documentation will allow users to understand how the statistics can be used in their own workflows.

Schedule and Deliverables:

Prepare for the project and interact with mentors - Community bonding period (April 27 - May 17)

During the community bonding period I will study the PySAL code base, namely the `esda` submodule, and review the PySAL developer guidelines. To build connections with the PySAL core development team I will attend the monthly developer meetings and participate in the PySAL gitter chat. Additionally, I will communicate with my mentor(s) to schedule convenient, recurring check-in meetings for the duration of the summer. If there are large time differences between myself and my mentor(s), I will adjust my work schedule to ensure portions of the workday overlap for skype or telephone calls.

Phase I: Implementation of estimators (May 18 - June 19, 5 weeks)

MILESTONE by end of week 6: Implement univariate and multivariate local join count statistics with inference.

- **Week 1 - code preparation:** I will review relevant literature on the spatial estimators proposed for implementation.(Anselin & Xi, 2019; Ord & Getis, 2012) I will pseudo-code initial attempts at the estimators with a structure that mirrors existing estimators within `esda` such as `Local_Moran`. I will discuss with my mentor(s) what type of inference should be prioritized for each estimator (i.e. both chi-square and bootstrap inference have been proposed for LOSH).(Ord & Getis 2012; Xu et al., 2014)
- **Week 2-4 - coding:** Code first attempts at each estimator. Submit draft to mentor(s) for feedback and identify priority areas for optimization. Attend PySAL developer meeting the first week of June.
- **Week 5 - code review:** Integrate mentor(s) feedback to optimize calculation of estimators and ensure code structure consistency with other `esda` estimators. I will establish expected efficiency using functions like `timeit`.

Phase II: Creating docstrings and tests (June 22 - July 17, 4 weeks)

MILESTONE by end of week 10: have docstrings and testing implemented for each of the new estimators.

- **Week 7-8 - documentation:** Build out docstrings with equations and references for each of the functions in a manner consistent with other `esda` functions. Implement unit tests that mirror existing scripts like `Local_Moran` in each of the estimators. Attend PySAL developer meeting the first week of July.
- **Week 9-10 - performance and notebook planning:** Assess performance of each test on common spatial datasets found within PySAL (e.g. Housing Prices in Berlin) and, when available, to the datasets present in original publications of each estimator. Systematically apply estimators to datasets of different sizes (e.g. 10, 100, 1000, 10000) to assess efficiency as the dataset increases in size. Review consistency and grammar of implemented estimators. Plan organization and content of demonstration notebooks.

Phase III: Writing notebooks to demonstrate estimators (July 20 - Aug 10, 3 weeks)

Phase I and II will have implemented three new estimators for the PySAL `esda` submodule. In the final two weeks I will expand documentation, writing notebooks to demonstrate examples of the estimators for interested users.

MILESTONE by end of week 12: Written and reviewed ready-for-dissemination notebooks demonstrating each of the newly implemented estimators.

- **Week 11-12 - write notebooks:** Draft example notebooks for each of the newly implemented estimators using the spatial datasets included in other PySAL notebooks. Share notebooks with mentor(s) for feedback. Attend PySAL developer meeting the first week of August.
- **Week 13 - finalize contributions:** Incorporate mentor(s) feedback into notebooks. Ensure estimators and notebooks are properly formatted according to PySAL guidelines such that they are ready for external use (e.g. PySAL workshops).

NumFOCUS Questions

Q1: Development Experience

My primary development experience comes from arcospy. While planning my dissertation research at the University of Maryland I began to review and follow media coverage of the Opioid Epidemic. In this day-to-day coverage I happened across a breaking story by *The Washington Post* about a large amount of Drug Enforcement Agency (DEA) data being made publicly available. After reading the story and exploring the data I realized its immense research potential. In the late summer of 2019, I began to familiarize myself with the data, and in the Fall 2019 semester I incorporated the data into two substantial class projects. A component of one class project was the opportunity to carry out a mini software development exercise. At the time I did not have any experience in development, but I wanted to expand my skill set and so I translated *The Washington Post's* R API wrapper into Python. This project would eventually expand into an informal partnership with the data journalists at *The Washington Post* to promote

the software as twin packages and draft a jointly authored paper for submission to open source software journals. This process provided invaluable experience in the end-to-end development of a simple software package and the practice of building a relationship with external partners.

Q2: Other Experiences

I have substantial experience working as a research assistant carrying out data-oriented tasks in geography, political science, and epidemiology. These assistantships have provided me with practical experience using a variety of technologies for scripting (R, Python, STATA), database management (SQL, PostgreSQL, PostGIS), and command-line utilities (bash). These experiences are detailed in a recent copy of my CV. In addition to work experience, I have completed several courses in statistics (non-spatial and spatial). This educational experience has provided important preliminary experience into the nuances of spatial statistics that are essential when implementing measures like local join counts and LOSH.

Q3: Why this project?

In August of 2019 I began a PhD in Geographic Information Science and Remote Sensing at the University of Maryland. I have an active interest in the spatial analysis of health outcomes, especially the examination of health inequalities using statistical products (census, data, surveys) from the U.S. government. I have come to appreciate the need for tools like PySAL in my own research and the burgeoning field of GDS. The priority statistics identified for implementation are immensely useful in spatial analysis, and incorporating the statistics into PySAL would enable the research of myself and others. Helping develop the tools I frequently use is of immense practical and personal interest, and I intend to become a regular contributor to PySAL after successfully completing GSoC 2020.

Other commitments

- N/A

Other GSoC applications

- I am not submitting applications to any other GSoC organizations or GSoC PySAL projects.

Appendix

References

Anselin L (1995) Local indicators of spatial association—LISA. Geographical Analysis 27:93–115. doi:10.1111/j.1538-4632.1995.tb00338.x

Anselin, L. (1996). Interactive Techniques and Exploratory Spatial Data Analysis. Regional Research Institute Publications and Working Papers. 200. https://researchrepository.wvu.edu/rri_pubs/200

Anselin, L., Li, X. (2019). Operational local join count statistics for cluster detection. Journal of Geographical Systems : Spatial Theory, Models, Methods, and Data, 21(2), 189-210. doi:10.1007/s10109-019-00299-x

Cliff, A., Ord, J. (1981). Spatial Processes: Models and Applications. Pion, London.

Getis, A., Ord, J. (1992). The analysis of spatial association by distance statistics. Geographical Analysis 24:189–206. doi.org/10.1111/j.1538-4632.1992.tb00261.x

Ord, J., Getis, A. (2012). Local spatial heteroscedasticity (losh). The Annals of Regional Science: An International Journal of Urban, Regional and Environmental Research and Policy, 48(2), 529-539. doi:10.1007/s00168-011-0492-y

Singleton A, Arribas-Bel D (2019) Geographic Data Science. Geographical Analysis. doi:10.1111/gean.12194

Xu, M., Mei, C. L., & Yan, N. 2014. A note on the null distribution of the local spatial heteroscedasticity (LOSH) statistic. The Annals of Regional Science, 52 (3), 697--710. doi:10.1007/s00168-014-0605-5