Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *http://www.jstatsoft.org/*

The PySAL ecosystem: philosophy and implementation

PySAL Developers

Abstract

PySAL is a library for geocomputation and spatial/geographic data science. Written in Python, the library has a long history of supporting novel science and broadening methodological impacts far afield of academic work. Recently, many new techniques, methods of analyses, and development modes have been implemented, making the library much larger and more encompassing than that previously discussed in the literature (Rey and Anselin 2007, e.g.). As such, we provide an introduction to the library as it stands now, as well as the scientific and conceptual underpinnings of its core set of authors. Finally, we provide a prospective look at the library's future evolution.

Keywords: spatial analysis, open-source computation, spatial econometrics, statistics, data science, spatial data science.

Introduction

In recent years, it has become increasingly important for scientists to adopt open science practices (Piwowar, Day, and Fridsma 2007; Piwowar and Vision 2013), especially for junior or early-career researchers (Allen and Mehler 2019). New tools and platforms have also lowered the technical barriers to entry for contributing to open science projects. Open approachs enhance reproducibility, transparency and speed of scientific workflows and discovery. One critical part of open science practices is the development, improvement, maintenance, and use of open science tooling (FOSTER 2014). Alongside the broader trends in quantitative research towards computation-driven inquiry (Efron and Hastie 2016), geography has provided a fertile ground for open science (Rey, Anselin, Li, Pahle, Laura, Li, and Koschinsky 2015; Singleton, Spielman, and Brunsdon 2016). Large-scale collaborations on technical and scientific infrastructure have long been a requirement in Geography, owing to distinctive spatial data representations, statistical concerns, and computational requirements. But, in the past, many of these large-scale, open collaborations have been outpaced in functionality and

computational performance by closed source, proprietary platforms. This led to widespread awareness of the challenges of "disabling technologies" in the field (Gahegan 1999), where the implementation of a specific suite of analytical capabilities limited the conceptual and practical reach of spatial science. During the past decade, however, the situation has begun to change, as progress in methodology of spatial analysis has been aided by the availability of open source (and thus verifiable) software, in contrast to the closed source black box implementations of proprietary software, where the underlying assumptions were often not made explicit.

As a result, the dependence on proprietary software has been waning, as there is now a strong case to be made for open science in geography (Rey 2009). Treating scientific code as text, enmeshed and integral to the scientific work, has pedagogic, scientific, and societal benefits. As part of this process, packages such as **spdep** in R (Bivand, Pebesma, and Gomez-Rubio 2013) or **PySAL** in Python (Rey and Anselin 2007) serve as open libraries in two senses. First, in terms of computation, they are open libraries that support scientists *doing spatial science* through the analyses they make possible & reproducible. Second, in terms of literature, they are open libraries that support students *learning spatial science* through the algorithms they make explicit. Thus, it is important to ensure long-term contributions, development, and maintenance to open scientific libraries.(Gahegan 2018)

PySAL is one of these long-term projects that has proven successful due to its handling of both the learning and doing of spatial science. Since its initial public release in 2010 (Rey 2019), PySAL has demonstrated the benefits of an open source geographic science library and seen widespread adoption across a diverse set of applications. As a software library, PySAL is relied upon by a number of other analysis packages to develop specialized tools for spatial analysis, prominent examples include **geopandas**, **geoplot**, **momepy**, and **geosnap**. PySAL is also used by researchers in the analysis of a wide array of topics across many disciplines including political science (Ingram and Harbers 2019), criminology (Jendryke and McClure 2019), economics (Felkner and Townsend 2011), planning (Nourian, Ohori, and Martinez-Ortiz 2018), public health (Joo 2017), engineering (Fan, Zhu, She, Guo, and Guo 2018), environmental science (Heilmayr and Lambin 2016), chemistry (Spiridon and Minh 2017), physics (Jakubska-Busse, Janowicz, Ochnio, and Ashbourn 2018), religion (Ferguson and Tamburello 2015), biology (Noorbakhsh, Farahmand, Soltanieh-ha, Namburi, Zarringhalam, and Chuang 2019), neuroscience (Burt, Demirtaş, Eckner, Navejar, Ji, Martin, Bernacchia, Anticevic, and Murray 2018), epidemiology (Hughes, Naik, Sengupta, and Saxena 2014), technology forecasting (Kwakkel, Carley, Chase, and Cunningham 2014), climate change (Ozturk, Chaudhary, Votava, and Kotfila 2016), organizational dynamics (Vaz, Miki, de Noronha, and Cusimano 2017), information visualization (Cottam and Lumsdaine 2012), ecology (Theodoridis, Nogués-Bravo, and Conti 2019), and sociology (Manduca and Sampson 2019), among others.

PySAL has evolved significantly since its original inception, both technologically and as a collaborative research endeavour. This paper frames recent changes against the backdrop of the project's history, and presents the ecosystem model that was recently adopted as a solution to some of the challenges posed by its own success. The remainder of the paper is organised as follows: Section 1 reviews the process of growth and change experienced by the project since its early years to the recent move to a federated model; Section 2 introduces the new structure of the package and, in doing so, reviews the current set of functionality available in PySAL; Section 3 considers non-technical aspects of the project, including governance

practices, the approach to community-building, and pedagogy; and Section 4 concludes with some reflections on the future challenges and next steps.

1. Understanding PySAL 2.0: Original Design, Evolution, and Current Model

Original design principles

To understand the recently adopted model, we first need to frame it under the evolution the library has experienced since its inception, now over ten years ago. In the early days of PySAL, the Python scientific ecosystem was largely devoid of any packages covering geospatial analysis. PySAL was conceived as an initial attempt to fill this void. Our target audience was data scientists who wanted to engage with spatial analysis using Python, as well as developers who could leverage the library to build new applications across the growing number of delivery platforms including desktop, plugins to standard geographic information systems, and web-based applications. To support those users and fuel the dissemination of the library, we wanted to ensure that installation of PySAL was streamlined.[1] We also stressed the importance of interoperability with the wider geospatial stack outside of Python, such as the proliferation of spatial analysis packages in R, Matlab, and stata. These other ecosystems brought with them community-specific implementations of spatial data formats and interfaces that were beginning to limit collaboration between and across user and developer communities. Python had already been widely recognized as an excellent "scientific glue" (van Rossum 1989) that could be used to leverage disparate scientific code. In designing PySAL, we wanted to leverage this feature of Python.

The original model to develop PySAL was community-driven and centralised. The code was structured as a single package with several closely interrelated submodules. Since there was little existing Python code within the domain, the first versions were focused on covering the minimum functionality required to start a typical spatial analysis. This included functionality such as file readers and writers, and foundational data structures on which several techniques relied, e.g. spatial weights matrices. Once these building blocks were created, development was mainly driven by volunteered time or funded time from grants on related projects. So, subsequent functionality focused on areas related to ongoing grants or research interests of developers. In this period, the development team was five to ten people who already had a history of collaboration through other research projects, all based within the same academic department. These circumstances allowed for direct communication, rapid iteration of ideas, and rapid progress. The library was reliable, stable, and reflected a research group's consensus about how to do cutting edge spatial science. However, this also meant that efforts to establish more formal channels of communication and develop materials to help integrate new contributors external to the PySAL project remained aspirational.

Growth in a time of (technical) debt

While this first model of growth and community-driven development was innovative for the field of spatial sciences and spatial econometrics it also resulted in "technical debt" (Kruchten, Nord, and Ozkaya 2012). Though it succeeded in getting the project off the ground, choices

[1] At the time, there were no dedicated package managers such as Anaconda, Inc.'s **conda** or container technologies such as Docker, so installing certain dependencies was substantially more difficult.

about the structure of the software affected the software's subsequent growth, maintenance, and future stability in the following ways.

First, designing the package as a monolithic distribution of spatial analysis functions meant that the most commonly-used parts of the library were very tightly coupled together (Perrow 2011). This meant that changes introduced into these components of the library immediately and substantially affected other parts. In some cases, this resulted in "cascades" of faults: changes in the "stack", the numerical and computational libraries upon which PySAL relied (e.g. **numpy**, van der Walt, Colbert, and Varoquaux 2011; **scipy**, Jones, Oliphant, Peterson *et al.* 2001–) could introduce software faults in one part of PySAL and the entire package would thus be affected. Errors within this tight integration, however, would only be discovered when a new contribution made them apparent. Often this new contribution was not even to the subsystem in which the fault was introduced. Additionally, attempts to resolve these kinds of faults would sometimes introduce new issues in other parts of the library. Triage of these problems and their fixes was difficult, especially while accepting new contributions and continuing the development of the package as new scientific advances were made. This situation made the library confusing to use for those not directly involved in the development process and complicated the process of accepting outside contributions.

Second, the history of the package significantly guided how the package was distributed and discussed. Since many in the original development group understood most of the functionality across the library, all functionality was distributed together and exposed at the same Application Program Interface (API) endpoint. For example, users interested in conducting spatial regression analyses would also be exposed to a set of statistics and tools for the analysis of Markov Chains in the program's interface. This API design also meant that explanations of what PySAL could do were difficult to focus, further harming user experience. Beyond PySAL being perceived as a large toolbox filled with many tools used for very different purposes, it remained unclear to a new user how the pieces fit together.

Third, the maintenance required by tight coupling muddied the efforts of new contributors. Because parts of the package were tightly coupled, new contributions often required large amounts of editorial work by a team of maintainers, before novel functionality or enhancements could be integrated. This caused a situation akin to the Matthew effect (Merton 1968): a new contributor may make a significant and novel addition, but this addition would be credited to experienced contributors. Significant contributions from new community members would require integration work elsewhere in the package. This integration work would normally be done by senior maintainers. The whole contribution would get "credited" to the senior maintainer that made it possible to include the new functionality instead of to the new contributor. While this is primarily a social problem, the tight coupling of the software exacerbated the integration effort required to include new contributions.

Finally, the tight coupling between library components seriously limited the library's ability to grow, refactor, and integrate with new *dependencies* as the Python ecosystem grew.[2] As a design principle, the Python language adopts a loose collection of statements set forth in Peters (2010). The thirteenth statement, "There should be one—and preferably only one—obvious way to do [a task]," became particularly challenging to obey due to the tight internal coupling in PySAL. Some parts of PySAL had been written before Python was mainstream in

[2] *Dependencies* are other Python packages that provide algorithms or computational objects that form the foundations upon which PySAL's analytical functionality was built.

science. Critically, Python's widespread adoption in spatial sciences meant that new packages often replicated and improved the infrastructure that PySAL built in pursuit of its main goal: scientifically novel spatial analytics. This new community infrastructure usually did not provide new analytical methods, but instead provided simpler data processing methods or more efficient data structures. And, these new packages grew up outside of PySAL; they were not "new contributions" to the package, but were new contributions to the wider ecosystem that PySAL would need to integrate from scratch. But, it was difficult to justify changing PySAL to rely on these newcomers: since the packages of interest only provided computational infrastructure, the work needed to re-tool PySAL for a newer and simpler infrastructure was seen as less important than work to support novel spatial science. This meant that new packages in the community were ignored, even when they significantly improved existing functionality or simplified user experience. As the number of new packages replicating or improving basic functionality increased, PySAL's internal consistncy also lead users to think that PySAL could not integrate with these other packages. As the tide of geospatial packages in Python continued to rise, PySAL needed to cut this tight-coupling tether; it needed to rely on the growing geospatial infrastructure in Python to hone the library's comparative advantage in spatial analytics.

A "federated" solution

To do this, the tight coupling needed to loosen. The solution to the growing challenge of maintaining and expanding PySAL was to move from a tightly integrated to a *federated* model. Rather than contributing all code to a monolithic package that holds all functionality, the project moved to a model where functionality split into several, smaller packages with a clearly delimited area of focus. Each of these packages are now independent Python packages in their own right. As such, they may have different maintainers, release cycles, and sets of dependencies. In this context, the library **PySAL** becomes an "aggregator", or a meta-package, that brings together all of these packages under a common brand and interface with a single install: every six months, PySAL collects the latest release of each federated package, wraps them under a common API, and releases it in a bundle.

This model brings together benefits from our prior monolithic approach with that of a fully distributed software community. Because the functionality is split across independent, self-contained packages, development is faster and more agile. Developers can rely on official versions of other packages to develop their own, and can focus on expanding functionality rather than ensuring their changes do not affect other ends of the library. Equally, testing any single package and catching faults is faster since each package's tests are now isolated from other packages. Furthermore, since packages are independent, releases of each sub-package can take place as soon as the developers agree to, without having to coordinate with a larger team. Users interested in only the functionality contained in one package can install only that package, bringing a smaller footprint and a more limited set of dependencies. Finally, it is easier to explain the purpose and functionality of each smaller package, as they focus on and contain only related functionality. These independent packages with more focused functionality also make the pathway and credit for new contributions much more clear. As discussed above, the monolithic approach lends itself more to focus attention on a smaller set of developers and maintainers, even though a larger group might be contributing functionality. A federated approach opens the option to include more developers in lead roles as package maintainers, and provides more opportunities to disseminate the functionality in independent

papers (e.g. Lumnitz, Arribas-Bel, Cortes, Gaboardi, Greiss, Oshan, Wolf, and Rey) or other venues, such as citable software releases through `zenodo` (Nielsen 2019). This ensures the community is healthy, broad, well-integrated and provides incentives to grow in diversity and functionality (Wolf, Rey, and Oshan 2019b). At the same time, the meta-package retains the stability and regularity of the monolithic approach. Users with more general needs can rely on the six-month release to provide a stable, one-install version that requires all the dependencies and provides the entire set of functionality in the federation. This "aggregator" also acts as a platform with higher visibility that makes it easier to discover functionality.

2. Current Analytical Capabilities

The new federated approach discussed above means **PySAL** is a meta-package that re-distributes several independent smaller packages. The new meta-package groups several loosely-connected packages together by common themes in order to bundle, organise, and assure consistent quality as a platform for spatial analytics. This de-couples the software structure of the library from the final "analysis platform" that is easy to access, learn, and deploy. Distribution and development issues are now resolved within each federated package, while the concerns about consistency and pedagogical clarity are addressed in the meta-package. Since the number of packages that **PySAL** encompasses is relatively large[3], and is expected to grow over time, the team decided to re-organise functionality in more general thematic categories, such as visualization or data exploration. The result is PySAL 2.0, released first in 2019.

The **PySAL** 2.X series organizes functionality around four main areas or domains: `lib` -core data structures and foundational algorithms-, `explore` - spatial data exploration -, `model` - explicitly-spatial modelling -, and `viz` - tools for visualization of spatial statistical analysis. Each of these domains is broadly aligned with different components of a spatial analysis workflow, and accordingly houses packages providing related functionality. To reflect this feature, each federated package is imported from within its own domain. The remainder of this section briefly describes the packages present in each domain for the original 2.0 release.

Foundational Algorithms: libpysal

Underpinning the three domains, **libpysal** provides foundational algorithms and data structures that support the rest of the library. This currently includes the following modules: input/output (`io`), which provides readers and writers for common geospatial file formats[4]; weights (`weights`), which provides the main class to store spatial weights matrices, as well as several utilities to manipulate and operate on them; computational geometry (`cg`), with several algorithms, such as Voronoi tessellations or alpha shapes (Edelsbrunner and Mücke 1994) that efficiently process geometric shapes; and an additional module with example data sets (`examples`). This domain is also a single stand-alone package due to its core importance to other domains.

[3]The first meta-package version of PySAL (2.0) consisted of 14 packages.

[4]Much of these are provided in a legacy mode to avoid breaking backwards compatibility. However, the consensus among the development team is to offload much of this area to related packages such as **geopandas** or **rasterio**.

Exploratory Spatial Data Analysis: explore

The `explore` layer of PySAL includes modules to conduct exploratory analysis of spatial and spatio-temporal data. At a high level, packages in `explore` are focused on enabling the user to better understand patterns in the data and suggest new interesting questions rather than answer existing ones. They include methods to characterize the structure of spatial distributions (either on networks, in continuous space, or on polygonal lattices). In addition, this domain offers methods to examine the *dynamics* of these distributions, such as how their composition or spatial extent changes over time.

esda

Exploratory spatial data analysis (ESDA) involves the interrogation of patterns in spatial data. Common topics in ESDA include the analysis of *spatial dependence*, where realizations from a random spatial process depend on other nearby realizations and *spatial heterogeneity* where a process may exhibit different behavior in different areas. In exploratory spatial data analysis, *spatial autocorrelation*, statistical dependence of a given variable with other nearby measurements of that same variable, is often critical to identify and understand. The **esda** package implements methods for the analysis of both global (map-wide) and local (focal) spatial autocorrelation (Anselin 1995), for both continuous and binary data. In addition, the package increasingly offers cutting-edge statistics about boundary strength (Wolf, Knaap, and Rey 2019a) and measures of aggregation error in statistical analyses (Duque, Laniado, and Polo 2018).

giddy

Geospatial Distribution Dynamics (**giddy**) is an extension of **esda** to spatio-temporal data. The package hosts state-of-the-art methods that explicitly consider the role of space in the dynamics of distributions over time (Kang, Rey, Stephens, Malizia, Wolf, Lumnitz, Gaboardi, Laura, Schmidt, Knaap, and Eschbacher 2019). A full set of spatially-extended discrete Markov chain models, including Spatial Markov, LISA Markov, Full Rank Markov, and Geographic Rank Markov models (Rey 2001, 2014) are available for users who are interested in the underlying transitional dynamics of a process as well as how the spatial structure shapes such dynamics. Global and Local Indicators of Mobility Association (GIMA and LIMA) —see Rey (2016)—are also provided in **giddy**. These indicators assess the degree to which changes in the positions in an (income) distribution over two time periods displays a global or local spatial pattern.

inequality

Indices for measuring inequality over space and time are included in the **inequality** package. These comprise classic measures such as the Theil T information index and the Gini index in mean deviation form; but also spatially-explicit measures that incorporate the location and spatial configuration of observations in the calculation of inequality measures. For example, the Theil inequality index can be decomposed into between and within inequality contributions, the so-called inter- and intra-regional inequality (Rey 2004). Complementing this partition-based approach, the package also provides a Spatial Gini decomposition (Rey and Smith 2013) that can be used to test if inequality is distinct between observations that are spatial neighbors and those that are not. Complementing the implementation of measures of

inequality, several statistics also include inference methods that use a variety of permutation-based and analytical approaches.

pointpats

The statistical analysis of point data is supported by the **pointpats** package (Rey, Kang, Shao, Wolf, Seth, Gaboardi, and Arribas-Bel 2019). This package provides methods to characterise the spatial structure of an observed point pattern: a collection of locations where some phenomena of interest have been recorded. Measures of centrography provide overall geometric summaries of the point pattern, including central tendency, dispersion, intensity, and extent. In addition, **pointpats** supports a flexible window, or geometric frame, that is used in the calculation of these descriptive measures and in visualizations. This window is also used to implement formal tests for clustering or co-location, including quadrat-based methods and distance-based methods (van Lieshout and Baddeley 1996).

segregation

The **segregation** package (Cortes, Rey, Knaap, and Wolf 2019) calculates over 40 different segregation indices and provides a suite of additional features for measurement, visualization, and hypothesis testing that together represent the state-of-the-art in quantitative segregation analysis. These methods are exposed through a streamlined interface that allows users to calculate common and advanced measures of segregation, including aspatial, spatial, two-group, multi-group, and localized indices. In addition, the spatial structure of a dataset can be represented using spatial weights from the **libpysal** package, or street network distances that can depict a more detailed picture of urban accessibility. Users of **inequality** can also perform simulation-based hypothesis testing for single values (e.g. when testing for the presence or absence of segregation) or value pairs (e.g. when testing whether a given city is more segregated than another), as well as decompose comparisons into spatial and demographic structures.

spaghetti

Many spatial processes are constrained to networks, and hence, studying them in a euclidean-based framework may lead to results that are less representative of reality (Barthélemy 2011; Ducruet and Beauguitte 2014). Therefore, **Spa**tial **G**rap**h**s: N**et**works, **T**opology, & **I**nference (**spaghetti**) was developed to provide data structures and analytical methods to study networks and statistical processes on networks (Gaboardi, Laura, Rey, Wolf, Folch, Kang, Stephens, and Schmidt 2018). For instance, the Network K Function allows for the statistical testing of clusters on networks (Okabe and Sugihara 2012, Ch. 6). In order to make these kinds of statistics efficient, **spaghetti** provides a robust all-to-all Dijkstra shortest path algorithm with multiprocessing functionality. Other current functionality includes high-performance geometric and spatial computations using **geopandas** that are necessary for high-resolution interpolation along networks, and the ability to connect near-network observations onto the network (Gaboardi, Folch, and Horner 2019).

Explicitly-Spatial Statistical Modelling: model

In contrast to `explore`, the `model` layer focuses on confirmatory analysis. In particular, its

packages focus on the estimation of spatial relationships in data with a variety of linear, generalized-linear, generalized-additive, nonlinear, multi-level, and local regression models.

mgwr

Geographically-weighted regression (GWR) is a central tool in geographical analysis (Fotheringham, Brunsdon, and Charlton 2002). At its core, geographically-weighted regression models are a *local regression* technique (Cleveland and Devlin 1988) that borrows data from nearby locations to estimate place-specific coefficients. The method recognizes that parameters may vary across the spatial domain when the same stimulus elicits a different response depending upon geographical context. Recent innovations in the GWR methodology remove the limitation that only one scale is considered; typically a single "bandwidth" controls how far sites are allowed to borrow data for all of relationships in the model. Multiscale GWR is a new approach based on generalized additive models (Wood 2006) that allows for bandwidths that vary uniquely for each predictor (Fotheringham, Yang, and Kang 2017). This means that data borrowing might be more local for some covariates than others, suggesting more nuanced patterns in the relationships between a set of covariates and a response. Altogether, the **mgwr** package provides scalable algorithms for estimation, inference, and prediction using single- and multi-scale geographically-weighted regression models in a variety of generalized linear model frameworks, as well model diagnostics tools (Oshan, Li, Kang, Wolf, and Fotheringham 2019).

spglm

In order to solve geographical modelling problems efficiently, it is useful to employ sparse matrix operations where possible (Bivand and Piras 2015). Existing generalized linear modelling frameworks in Python, such as **statsmodels** (Seabold and Perktold 2010), did not fully incorporate sparse methods in its generalized linear modelling frameworks. To address this gap, **spglm** implements a set of generalized linear regression techniques, including Gaussian, Poisson, and Logistic regression, that allow for sparse matrix operations in their computation and estimation to lower memory overhead and decreased computation time.

spint

Spatial interaction models are a class of geographical models for studying the interaction between places (Fotheringham and O'Kelly 1989; Roy and Thill 2003; Batty 2013). **spint** seeks to provide a collection of tools to study spatial interaction processes and analyze spatial interaction data (Oshan 2016). A primary functionality of **spint** is to facilitate the calibration and interpretation of a family of gravity-type spatial interaction models, including those with *production* constraints (where total outgoing flows predicted by the model must be unbiased), *attraction* constraints (where total incoming flows predicted by the model must be unbiased), or a combination of the two constraints (Wilson 1971). Given the unique structure of calibrating models with constraints, **spint** provides scalable algorithms by leveraging sparse matrix operations in **spglm**.

spreg

The package **spreg** supports the estimation of classic and spatial econometric models. Currently it contains methods for estimating standard Ordinary Least Squares (OLS), Two Stage

Least Squares (2SLS) and Seemingly Unrelated Regressions (SUR), in addition to various tests of homokestadicity, normality, spatial randomness, and different types of spatial autocorrelation. There is also a suite of tests for spatial dependence in models with binary dependent variables (Amaral, Anselin, and Arribas-Bel 2013). The package enables the incorporation of both spatial dependence and spatial heterogeneity into traditional econometric models. To deal with spatial dependence, the package contains methods for estimating spatial lag and/or error models. Different flavours of these methods are available according with the characteristics of the specification: with/without heteroskedasticity or with/without endogenous predictors. Most of these models can then be fit via Generalised Method of Moments—GMM—or Maximum Likelihood—ML. To incorporate spatial heterogeneity, **spreg** allows the specification of spatial regimes in all of its methods, and provides tests for coefficient stability. For spatial panel estimation, **spreg** contains classic Spatial Seemingly-Unrelated Regression (SUR), Spatial Three Stage Least Squares, Lag SUR and Error SUR, in addition to Likelihood Ratio, Lagrange Multipliers and Chow tests to assess model specification or evaluate parameters. Additional details on these methods, as well as its implementation in the package, can be found in Anselin and Rey (2014).

spvcm

Variance components models are a kind of multilevel model used extensively in social science (Gelman and Hill 2006; Hox, Moerbeek, and van de Schoot 2010). They are most useful in situations where the differences between groups are of interest, but groups are of varying sizes or have differing levels of variation. Variance components methods partition variation into "within" group and "between" group variation, allowing for separate group-level and individual-level error terms. These models can be estimated using a variety of Bayesian and Maximum Likelihood methods (Browne and Draper 2006). In **spvcm**, a general framework for estimating spatially-correlated variance components models is provided. This class of models allows for spatial dependence in the variance components, so that nearby groups may affect one another (Lacombe and McIntyre 2016). The **spvcm** package also provides a general-purpose framework for estimating models using Gibbs sampling in Python, accelerated by the **numba** package (Lam, Pitrou, and Seibert 2015).

Visualisation Layer: `viz`

The `viz` layer provides functionality to support the creation of geovisualisations and visual representations of outputs from a variety of spatial analyses. Visualization plays a central role in modern spatial/geographic data science. Current packages provide classification methods for choropleth mapping and a common API for linking PySAL outputs to visualization toolkits in the Python ecosystem.

mapclassify

Choropleth maps are thematic maps that rely on shading, color, or patterning to represent the measurement of a statistical attribute across polygonal areas. The effective design of a choropleth map requires careful consideration of the symbolization as well as the choice of classification scheme that assigns observations to different map classes. The **mapclassify** package in PySAL addresses the second design imperative. Currently, fifteen different classification schemes are available in **mapclassify**, including a highly-optimized implementation

of Fisher-Jenks optimal classification (Rey, Stephens, and Laura 2017). Each scheme inherits a common structure that ensures computations are scalable and supports applications in streaming contexts. The popular geoprocessing and visualization packages **geopandas** and **geoplot** use **mapclassify**.

splot

The **splot** package provides statistical visualizations for spatial analysis (Lumnitz *et al.*). The package offers, i.e. methods for visualizing global and local spatial autocorrelation (through Moran scatterplots and cluster maps), temporal analysis of cluster dynamics (through heatmaps and rose diagrams), and multivariate choropleth mapping (through value-by-alpha maps; Roth, Woodruff, and Johnson 2010). A high level API supports the creation of publication-ready visualizations. Functionality that provides multiple views (i.e. scatterplots combined with cluster maps) and small multiples (i.e. facet plots) help to guide users in their visual analytics workflow and parameter choices through a "grammar of graphics." **splot**'s functionality is implemented across different graphical engines available in Python (including matplotlib and bokeh) to allow for static and interactive visualizations.

3. Pedagogy and Community

Since its inception in Rey and Anselin (2007), **PySAL** has aimed to satisfy two distinctive goals. The first goal is a scientific one: to serve as a platform that makes cutting-edge spatial analytic techniques available and accessible to a wide range of users. The second goal is more pedagogical: to employ computer programming as a medium to communicate advanced statistical concepts. At the same time, the project has been built following standard approaches in the world of open-source development; this has now reached beyond pure software development and into community building, which is structured through a transparent governance model. Over the years, the role of these two aspects —-pedagogy and community—- has grown in both relevance and the amount of effort devoted. This section unpacks some of the approaches adopted and provides further detail on the processes established.

The pedagogical ethos of the project comes across in a few broad areas: scientific documentation, open teaching, participatory governance, and community service. First, an exhaustive, clear, and updated documentation is complemented through direct access to the source code. From the very beginning, a compulsory requirement for any functionality added to **PySAL** has been to include a "docstring" together with new code. These are human-language explanations of what the method, class, or package does, along with a list of what is required to pass as input, and what the user can expect to receive as output, as well as a small example demonstrating its use. This close integration between computer code and human explanation, although by no means new or unique to **PySAL** (Knuth 1984), has been a distinctive feature of the library enhancing the understanding of functionality with wide coverage and consistency. The rationale behind this approach to developing community code is the belief that, by making the code easily accessible and complementing it with concise explanations, the user is more likely to use "code as text," as Rey (2009) argues. This supports and facilitates the transition from users of the package into developers and computational scholars. Well-documented code is easier to inspect and understand, so these users can get involved in the library's inner workings, and obtain a deeper insight into the computation and methodological details. This

approach supports the library, in that it trains new developers and contributors, but it also supports the broader academic discipline, because it makes the procedures involved in new science explicit.

Second, much of the effort of the team has been directed not only to detailed software documentation but also at creating broader materials on teaching spatial analysis. These study resources integrate **PySAL** with other community software to support broad instruction in geographic scince. For example, Arribas-Bel (2019) and Rey, Arribas-Bel, and Wolf (2021, *under contract*) introduce students to the nascent field of Geographic Data Science. To do so, they feature **PySAL** extensively. This material serves the purpose of extended, narrative documentation for the software; at the same time, the pedagogical approach to theoretical concepts is enriched by being able to take an explicitly-computational perspective, illustrating statistics with code snippets. The value of these materials is augmented by an additional effort to promote **PySAL** in a wide range of workshops and short courses.[5]

In addition to pedagogy, **PySAL** has paid special attention to governance. Its first ten years of existence saw the project grow from a small team localised in the same department, to a larger collective distributed across the world. To make this transition successful, several activities that used to take place in an informal setting in the early days were taken forward more proactively. First, collaboration around code was from the early days coordinated through an open, version control-based platform (Google Code first, Github currently). These platforms offer a detailed log of changes and, through "commit messages", "issues" and "pull requests", allow to reconstruct the evolution of the project as well as the technical discussions that surrounded it. Given the geographical distribution of developers, the team uses an open monthly call to cover aspects of the development for which written discussion was not practical. Topics such as the transition to Python 3 or the reorganisation of the library in subpackages were fleshed out in these calls, but also coordination around conference attendance or workshop proposals. Even though development is technically possible with the practices just described, the team has been purposeful about maintaining a regular schedule of face-to-face meetings. Usually held in the form of "code sprints" alongside academic conferences (such as the American Association of Geographers, the North American Regional Science, or the Scientific Python Conferences), these events serve a double purpose: first, they focus attention to particular areas of the project (maintenance, documentation, code development) that the group has identified as a priority; second, they act as a "social glue", keeping team members involved and engaged.

As the project has grown, it has become important to formalise how to integrate and foster external contributions. We have developed a code of conduct[6] that provides guidelines for interaction and collaboration around **PySAL** to any individual interested in contributing. As described above, part of the rationale behind moving to a federated model is to foster external contributions and to have a more flexible framework to incorporate cognate packages. To make this process easier, **PySAL** also has a package template[7] that details expected requirements from any package that wishes to join the federation.

Besides setting forth paticipation guidelines and governance within the **PySAL** community, the team has also embraced contributing to the larger Python community for data science.

[5] For an an illustration of materials developed with this outlets in mind, the reader is referred to: http://pysal.org/notebooks

[6] A copy is available at: https://github.com/pysal/governance/blob/master/conduct/code_of_conduct.rst

[7] A copy is available at: https://github.com/pysal/submodule_template

Rather than "reinventing the wheel", our goal is to provide the spatial analytic layer that makes cutting edge geospatial techniques available and integrates seamlessly within the larger ecosystem of Python packages and tools. This means we now fully aim to integrate our functionality with other large Python packages. An example of this strategy is the integration of choropleth classification schemes from **mapclassify** into the **geopandas** plotting API, or the interoperability between most of **PySAL** and `GeoDataFrames`, the foundational tabular data structure for geospatial data in Python. These technical integrations have been possible thanks to (but have also contributed to) closer collaboration with the development teams of other components of the ecosystem such as **geopandas** or **matplotlib**. Thanks to architecture and governance changes, **PySAL** is now much more embedded into the ecosystem at large, and stands to increase its integration going forward.

Finally, a note on funding. Although much of the work devoted to **PySAL** has come out of traditional "research time", the team has begun to explore alternative and complementary funding models to support development. Many of the features currently available were developed as part of larger research projects and grants that required a computational implementation of a method that was not available. For example, the **giddy** package emerged out of substantive research on income inequality dynamics carried out by members of the team (e.g. Rey and Montouri 1999; Rey 2016; Kang and Rey 2019a,b). A more dedicated funding stream has been in the form of the Google Summer of Code[8], a program run by Google that funds students to work on implementing new features on open-source projects. **PySAL** has used this model to rewrite internal core data structures, to add new functionality to already existing packages, or to develop brand new packages such as **spint** or **splot**. More recently, an additional funding approach includes joining the membership of NumFocus.

4. Future Plans & Next Steps

The first ten years of **PySAL** have seen the project evolve from a small, single package into a synchronised federation of packages that collectively enhance spatial analytics in Python. In this process, the technical and human infrastructure that support it has experienced profound changes, evolving to meet the demands of the given time. With all this ground covered, the logical question is: *What's next?* In this concluding section, we explore what lies ahead; what we consider as the main opportunities for the project to continue growing, but also the main challenges. The section is split first into several specific plans for the short- and mid-term, and a second set of longer-run reflections.

Specific plans

A keen interest of several contributors to **PySAL** has been to build a first-class module for spatial optimisation and regionalization. Compared to other functionality in the project, optimisation problems require a significantly larger set of underlying computational tools to solve. Spatial optimization algorithms usually rely heavily on general purpose optimisation or linear programming libraries once the spatial information has been expressed as a standard optimisation problem. These general-purpose optimization libraries must operate at peak performance given the difficulty of solving many spatial optimization problems, and so usually are written in C or Fortran. Since early releases, **PySAL** included a **region** module with a few

[8] https://summerofcode.withgoogle.com

algorithms implemented separately in Python. However, it soon became clear that a more unified approach that offloads heavy computations to a general linear programming library would be more efficient. This led to a Google Summer of Code to re-write **region** with a unified approach to its API. Recently however, as the ambitions of other packages such as **spaghetti** have expanded into domains that also require optimisation routines, the team has decided to move development to a new **spopt** package that unifies the approach, and provides underlying spatial optimisation routines in a more flexible and general way. In this context, there is an ample agenda to implement core algorithms, expose them through general interfaces, and then use them to build applications related to regionalization problems (e.g. spectral clustering, the SKATER or REDCAP algorithms), spatial optimisation along networks (e.g. optimal facility location modeling), or other domains where it might be useful.

A second area of interest aims to provide better integration with related libraries from the Python data science ecosystem. As mentioned above, the first efforts in **PySAL** had to be spent in building a set of utilities that, even though were not planned as a core part of the library, allowed the user to interface with spatial data (e.g. shapefile readers and writers). As the Python community evolved, these tasks were taken up by more comprehensive projects and the main priority of **PySAL** in this respect became to appropriately interface with these projects. A good example of this is **geopandas**, a package that extends functionality of **pandas** datatypes to spatial data. Once the project matured, it allowed **PySAL** to drop support for file I/O and focus on analytics. But **geopandas** also became a direct user of the **PySAL** ecosystem by using choropleth classification algorithms from **mapclassify** instead of re-implementing them. As the ecosystem matures and foundational libraries become more established and stable, a priority of **PySAL** is to further integrate with this functionality, making it not only possible but pleasant to write code that seamlessly knits different projects into a unified workflow that favors developer productivity and computational performance.

Finally, we increasingly see Python as one of many environments with which scholars and industry researchers conduct their work. So-called "polyglot" environments that seamlessly allow scholars to use packages from different computer languages in a single analysis are becoming increasingly common. This suggests that users of the library and developers building on top of the library may actually be coming from entirely different computing platforms. Further, documenting the interactions between software ecosystems becomes important when considering actual analytical workflows, where it may be easier to conduct some parts of an analysis in some environments and not others (Arribas-Bel, de Graaff, and Rey 2017). Thus, it becomes important for the library to document and build upon its integrations with other packages, including desktop GIS software (QGIS, ESRI ArcGIS), and other computing languages (such as Julia or R), in order to ensure that **PySAL** is usable in whatever environment the user actually resides.

General reflections

Beyond specific ideas, a series of guiding principles and ambitions are likely to be at the heart of the next "big decisions." The first one is the sense that the ground work required to build a platform of spatial analytics, and its place in the broader data science Python ecosystem, is largely completed. Maintenance (not a light task) aside, this makes it possible to focus entirely on ensuring the cutting edge methods are implemented shortly after they are invented. Our plan is that each federated package stays at the frontier of the domain

whose functionality it represents. A key ingredient of this idea is to reach out to scientists beyond the core development team and work with them to integrate their methods in **PySAL** code. Much of this process is enhanced with the move to a federated model discussed in the second section and, to some extent, it is already at work. For example, the authors of the "S-MAUP" statistic proposed in Duque *et al.* (2018) contributed their code to begin its implementation as part of **esda**.[9] Given recent changes in the library, we can effectively integrate contributions directly from the original authors rather than having to shoulder the burden of re-implementing cutting edge algorithms ourselves. Going forward, we will continue to integrate cutting-edge spatial science into **PySAL** given its new governance and technical structures.

To end this paper, we would also like to reflect on what we believe has been the most successful lesson learned over this period: the ability to maintain a flexible approach to adapt as the environment changes. It is important to be willing to change your own mindset to accommodate paradigm shifts in order to remain useful. This flexibility may slow achievement of short-term goals, but is the only way we have found to stay relevant. Our original intention was not to write file readers and writers, but there was no other way to make functionality in **PySAL** available to a wider audience. Neither were we enthusiastic about the work required to become compatible with Python 3. But, the rest of the ecosystem was moving in that direction, and ignoring it would have relegated the project to obsolescence; even the move to a federated model required a lot of additional developer time that could have been spent implementing new features. Flexibility can be expensive to attain, but it is a valuable investment for the future. We do not know what the scientific computing world will look like in ten years. But, as long as Python is playing a key role, we would like PySAL to continue contributing the spatial analytic layer to its larger ecosystem. We are sure that ensuring this contribution continues will take time, effort, and adaptation.

References

Allen C, Mehler DMA (2019). "Open Science Challenges, Benefits and Tips in Early Career and Beyond." *PLOS Biology*, **17**(5), e3000246. ISSN 1545-7885. doi:10.1371/journal.pbio.3000246.

Amaral PV, Anselin L, Arribas-Bel D (2013). "Testing for spatial error dependence in probit models." *Letters in Spatial and Resource Sciences*, **6**(2), 91–101. doi:10.1007/s12076-012-0089-9.

Anselin L (1995). "Local indicators of spatial association-LISA." *Geographical Analysis*, **27**(2), 93–115.

Anselin L, Rey S (2014). *Modern Spatial Econometrics in Practice: A Guide to GeoDa, GeoDaSpace and PySAL*. GeoDa Press, Chicago.

Arribas-Bel D (2019). "A course on Geographic Data Science." *The Journal of Open Source Education*, **2**(14). doi:https://doi.org/10.21105/jose.00042.

[9]The original pull request, with discussion and progress made for the contribution, is available at: https://github.com/pysal/esda/pull/58

Arribas-Bel D, de Graaff T, Rey SJ (2017). *Looking at John Snow's Cholera Map from the Twenty First Century: A Practical Primer on Reproducibility and Open Science*, pp. 283–306. Springer International Publishing, Cham. ISBN 978-3-319-50590-9. doi:10.1007/978-3-319-50590-9_17. URL https://doi.org/10.1007/978-3-319-50590-9_17.

Barthélemy M (2011). "Spatial networks." *Physics Reports*, **499**(1-3), 1–101. ISSN 03701573. 1010.0302.

Batty M (2013). *The New Science of Cities*. The MIT Press. ISBN 978-0-262-01952-1.

Bivand R, Piras G (2015). "Comparing Implementations of Estimation Methods for Spatial Econometrics." *Journal of Statistical Software*, **63**(18), 1–36.

Bivand RS, Pebesma E, Gomez-Rubio V (2013). *Applied spatial data analysis with R, Second edition*. Springer, NY. URL http://www.asdar-book.org/.

Browne W, Draper D (2006). "A Comparison of Bayesian and Likelihood-Based Methods for Fitting Multilevel Models." *Bayesian Analysis*, **1**(3), 473–514.

Burt JB, Demirtaş M, Eckner WJ, Navejar NM, Ji JL, Martin WJ, Bernacchia A, Anticevic A, Murray JD (2018). "Hierarchy of transcriptomic specialization across human cortex captured by structural neuroimaging topography." *Nature neuroscience*, **21**(9), 1251.

Cleveland WS, Devlin SJ (1988). "Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting." *J. Am. Stat. Assoc.*, **83**(403), 596–610.

Cortes RX, Rey S, Knaap E, Wolf LJ (2019). "An open-source framework for non-spatial and spatial segregation measures: the PySAL segregation module." *Journal of Computational Social Science*, pp. 1–32.

Cottam JA, Lumsdaine A (2012). "Spatial Autocorrelation-based Information Visualization Evaluation." In *Proceedings of the 2012 BELIV Workshop: Beyond Time and Errors - Novel Evaluation Methods for Visualization*, BELIV '12, pp. 8:1–8:8. ACM, New York, NY, USA. ISBN 978-1-4503-1791-7. doi:10.1145/2442576.2442584. URL http://doi.acm.org/10.1145/2442576.2442584.

Ducruet C, Beauguitte L (2014). "Spatial Science and Network Science: Review and Outcomes of a Complex Relationship." *Networks and Spatial Economics*, **14**(3-4), 297–316. ISSN 15729427. doi:10.1007/s11067-013-9222-6.

Duque JC, Laniado H, Polo A (2018). "S-Maup: Statistical Test to Measure the Sensitivity to the Modifiable Areal Unit Problem." *PLOS ONE*, **13**(11), e0207377. ISSN 1932-6203. doi:10.1371/journal.pone.0207377.

Edelsbrunner H, Mücke EP (1994). "Three-Dimensional Alpha Shapes." *ACM Trans. Graph.*, **13**(1), 43–72. ISSN 0730-0301. doi:10.1145/174462.156635.

Efron B, Hastie T (2016). *Computer Age Statistical Inference*, volume 5. Cambridge University Press.

Fan Y, Zhu X, She B, Guo W, Guo T (2018). "Network-constrained spatio-temporal clustering analysis of traffic collisions in Jianghan District of Wuhan, China." *PLoS one*, **13**(4), e0195093.

Felkner JS, Townsend RM (2011). "The Geographic Concentration of Enterprise in Developing Countries." *The Quarterly Journal of Economics*, **126**(4), 2005–2061. ISSN 0033-5533. doi:10.1093/qje/qjr046. http://oup.prod.sis.lan/qje/article-pdf/126/4/2005/5426002/qjr046.pdf, URL https://doi.org/10.1093/qje/qjr046.

Ferguson TW, Tamburello JA (2015). "The natural environment as a spiritual resource: A theory of regional variation in religious adherence." *Sociology of Religion*, **76**(3), 295–314.

FOSTER (2014). "Open Science Taxonomy." *Technical report*, FOSTER Open Science. URL fosteropenscience.eu/resources.

Fotheringham AS, Brunsdon C, Charlton M (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Wiley.

Fotheringham AS, O'Kelly ME (1989). *Spatial Interaction Models:Formulations and Applications*. Kluwer Academic Publishers. URL http://www.springer.com/earth+sciences+and+geography/geography/book/978-0-7923-0021-2.

Fotheringham AS, Yang W, Kang W (2017). "Multiscale Geographically Weighted Regression (MGWR)." *Annals of the American Association of Geographers*, (6), 1247–1265. doi:10.1080/24694452.2017.1352480.

Gaboardi JD, Folch DC, Horner MW (2019). "Connecting Points to Spatial Networks: Effects on Discrete Optimization Models." *Geographical Analysis*, **0**, 1–24. doi:10.1111/gean.12211.

Gaboardi JD, Laura J, Rey S, Wolf LJ, Folch DC, Kang W, Stephens P, Schmidt C (2018). "pysal/spaghetti." doi:10.5281/zenodo.1343650. URL https://github.com/pysal/spaghetti.

Gahegan M (1999). "Guest Editorial: What Is Geocomputation?" *Transactions in GIS*, **3**, 203–206.

Gahegan M (2018). "Our GIS Is Too Small." *The Canadian Geographer / Le Géographe canadien*, **62**(1), 15–26. ISSN 1541-0064. doi:10.1111/cag.12434.

Gelman A, Hill J (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Heilmayr R, Lambin EF (2016). "Impacts of nonstate, market-driven governance on Chilean forests." *Proceedings of the National Academy of Sciences*. ISSN 0027-8424. doi:10.1073/pnas.1600394113. https://www.pnas.org/content/early/2016/02/23/1600394113.full.pdf, URL https://www.pnas.org/content/early/2016/02/23/1600394113.

Hox JJ, Moerbeek M, van de Schoot R (2010). *Multilevel Analysis: Techniques and Applications, Second Edition*. 2 edition edition. Routledge, New York. ISBN 978-1-84872-846-2.

Hughes C, Naik VS, Sengupta R, Saxena D (2014). "Geovisualization for cluster detection of Hepatitis A & E outbreaks in Ahmedabad, Gujarat, India." In *Proceedings of the Third ACM SIGSPATIAL International Workshop on the Use of GIS in Public Health*, pp. 39–44. ACM.

Ingram MC, Harbers I (2019). "Spatial Tools for Case Selection: Using LISA Statistics to Design Mixed-Methods Research." *Political Science Research and Methods*, pp. 1–17.

Jakubska-Busse A, Janowicz M, Ochnio L, Ashbourn J (2018). "Pickover biomorphs and non-standard complex numbers." *Chaos, Solitons & Fractals*, **113**, 46–52.

Jendryke M, McClure SC (2019). "Mapping crime - Hate crimes and hate groups in the USA: A spatial analysis with gridded data." *Applied Geography*, **111**, 102072. ISSN 0143-6228. doi:https://doi.org/10.1016/j.apgeog.2019.102072. URL http://www.sciencedirect.com/science/article/pii/S014362281831004X.

Jones E, Oliphant T, Peterson P, *et al.* (2001–). "SciPy: Open source scientific tools for Python." [Online; accessed <today>], URL http://www.scipy.org/.

Joo Y (2017). "Spatiotemporal study of elderly suicide in Korea by age cohort." *Public Health*, **142**, 144 – 151. ISSN 0033-3506. doi:https://doi.org/10.1016/j.puhe.2016.07.016. URL http://www.sciencedirect.com/science/article/pii/S0033350616301871.

Kang W, Rey S, Stephens P, Malizia N, Wolf LJ, Lumnitz S, Gaboardi JD, Laura J, Schmidt C, Knaap E, Eschbacher A (2019). "pysal/giddy: giddy 2.2.2." doi:10.5281/zenodo.3401736. URL https://doi.org/10.5281/zenodo.3401736.

Kang W, Rey SJ (2019a). "Inference for Income Mobility Measures in the Presence of Spatial Dependence." *International Regional Science Review*, pp. 1–30.

Kang W, Rey SJ (2019b). "Smoothed Estimators for Markov Chains with Sparse Spatial Observations." *Geographical Analysis*.

Knuth DE (1984). "Literate programming." *The Computer Journal*, **27**(2), 97–111.

Kruchten P, Nord RL, Ozkaya I (2012). "Technical debt: From metaphor to theory and practice." *Ieee software*, **29**(6), 18–21.

Kwakkel JH, Carley S, Chase J, Cunningham SW (2014). "Visualizing geo-spatial data in science, technology and innovation." *Technological Forecasting and Social Change*, **81**, 67 – 81. ISSN 0040-1625. doi:https://doi.org/10.1016/j.techfore.2012.09.007. URL http://www.sciencedirect.com/science/article/pii/S0040162512002193.

Lacombe DJ, McIntyre SG (2016). "Local and Global Spatial Effects in Hierarchical Models." *Applied Economics Letters*, **23**(16), 1168–1172. doi:10.1080/13504851.2016.1142645.

Lam SK, Pitrou A, Seibert S (2015). "Numba: A LLVM-Based Python JIT Compiler." In *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, LLVM '15, pp. 7:1–7:6. ACM, New York, NY, USA. ISBN 978-1-4503-4005-2. doi:10.1145/2833157.2833162.

Lumnitz S, Arribas-Bel D, Cortes RX, Gaboardi JD, Greiss V, Oshan TM, Wolf L, Rey S (????). *The Journal of Open Source Software*. doi:10.21105/joss.01882.

Manduca R, Sampson RJ (2019). "Punishing and toxic neighborhood environments independently predict the intergenerational social mobility of black and white children." *Proceedings of the National Academy of Sciences*, **116**(16), 7772–7777.

Merton RK (1968). "The Matthew effect in science: The reward and communication systems of science are considered." *Science*, **159**(3810), 56–63.

Nielsen LH (2019). "Software citations now available in Zenodo." URL https://blog.zenodo.org/2019/01/10/2019-01-10-asclepias/.

Noorbakhsh J, Farahmand S, Soltanieh-ha M, Namburi S, Zarringhalam K, Chuang J (2019). "Pan-cancer classifications of tumor histological images using deep learning." *bioRxiv*. doi:10.1101/715656. https://www.biorxiv.org/content/early/2019/07/26/715656.full.pdf, URL https://www.biorxiv.org/content/early/2019/07/26/715656.

Nourian P, Ohori KA, Martinez-Ortiz C (2018). "Essential Means for Urban Computing: Specification of Web-Based Computing Platforms for Urban Planning, a Hitchhiker's Guide." *Urban Planning*, **3**(1), 47–57.

Okabe A, Sugihara K (2012). *Spatial Analysis along Networks*. John Wiley & Sons, Ltd. ISBN 9781119967101. doi:10.1002/9781119967101.

Oshan T, Li Z, Kang W, Wolf L, Fotheringham A (2019). "mgwr: A Python Implementation of Multiscale Geographically Weighted Regression for Investigating Process Spatial Heterogeneity and Scale." **8**(6), 269. ISSN 2220-9964. doi:10.3390/ijgi8060269. URL https://www.mdpi.com/2220-9964/8/6/269.

Oshan TM (2016). "A Primer for Working with the Spatial Interaction Modeling (SpInt) Module in the Python Spatial Analysis Library (PySAL)." *REGION*, **3**(2), R11–R23. ISSN 2409-5370. doi:10.18335/region.v3i2.175.

Ozturk D, Chaudhary A, Votava P, Kotfila C (2016). "GeoNotebook: Browser based Interactive analysis and visualization workflow for very large climate and geospatial datasets." In *AGU Fall Meeting Abstracts*.

Perrow C (2011). *Normal accidents: Living with high risk technologies-Updated edition*. Princeton university press.

Peters T (2010). "The zen of python." In *Pro Python*, pp. 301–302. Springer.

Piwowar HA, Day RS, Fridsma DB (2007). "Sharing Detailed Research Data Is Associated with Increased Citation Rate." *PLOS ONE*, **2**(3), e308. ISSN 1932-6203. doi:10.1371/journal.pone.0000308.

Piwowar HA, Vision TJ (2013). "Data Reuse and the Open Data Citation Advantage." *PeerJ*, **1**, e175. ISSN 2167-8359. doi:10.7717/peerj.175.

Rey S, Kang W, Shao H, Wolf LJ, Seth M, Gaboardi JD, Arribas-Bel D (2019). "pysal/pointpats: pointpats 2.1.0." doi:10.5281/zenodo.3265637. URL https://doi.org/10.5281/zenodo.3265637.

Rey SJ (2001). "Spatial empirics for economic growth and convergence." *Geographical Analysis*, **33**(3), 195–214.

Rey SJ (2004). "Spatial analysis of regional income inequality." In M Goodchild, D Janelle (eds.), *Spatially Integrated Social Science: Examples in Best Practice*, pp. 280–299. Oxford University Press, Oxford.

Rey SJ (2009). "Show me the code: spatial analysis and open source." *Journal of Geographical Systems*, **11**(2), 191–207.

Rey SJ (2014). "Rank-based Markov chains for regional income distribution dynamics." *Journal of Geographical Systems*, **16**(2), 115–137.

Rey SJ (2016). "Space–time patterns of rank concordance: Local Indicators of Mobility Association with application to spatial income inequality dynamics." *Annals of the American Association of Geographers*, **106**(4), 788–803. doi:10.1080/24694452.2016.1151336. http://dx.doi.org/10.1080/24694452.2016.1151336, URL http://dx.doi.org/10.1080/24694452.2016.1151336.

Rey SJ (2019). "PySAL: the first 10 years." *Spatial Economic Analysis*, **0**(0), 1–10. doi:10.1080/17421772.2019.1593495. https://doi.org/10.1080/17421772.2019.1593495, URL https://doi.org/10.1080/17421772.2019.1593495.

Rey SJ, Anselin L (2007). "PySAL: A Python library of spatial analytical methods." *The Review of Regional Studies*, **37**(1), 5–27.

Rey SJ, Anselin L, Li X, Pahle R, Laura J, Li W, Koschinsky J (2015). "Open Geospatial Analytics with PySAL." *ISPRS International Journal of Geo-Information*, **4**(2), 815–836.

Rey SJ, Arribas-Bel D, Wolf LJ (2021, *under contract*). *Geographic Data Science with Python and the PyData Stack*. CRC Press, Boca Raton, FL.

Rey SJ, Montouri BD (1999). "US regional income convergence: a spatial econometric perspective." *Regional studies*, **33**(2), 143–156.

Rey SJ, Smith RJ (2013). "A spatial decomposition of the Gini coefficient." *Letters in Spatial and Resource Sciences*, **6**, 55–70.

Rey SJ, Stephens P, Laura J (2017). "An evaluation of sampling and full enumeration strategies for Fisher Jenks classification in big data settings." *Transactions in GIS*, **21**(4), 796–810.

Roth RE, Woodruff AW, Johnson ZF (2010). "Value-by-Alpha Maps: An Alternative Technique to the Cartogram." *The Cartographic Journal*, **47**(2), 130–140. ISSN 0008-7041. doi:10.1179/000870409X12488753453372.

Roy JR, Thill JC (2003). "Spatial interaction modelling." **83**(1), 339–361. ISSN 1056-8190, 1435-5957. doi:10.1007/s10110-003-0189-4. URL http://doi.wiley.com/10.1007/s10110-003-0189-4.

Seabold S, Perktold J (2010). "Statsmodels: Econometric and statistical modeling with python." In *9th Python in Science Conference*.

Singleton AD, Spielman S, Brunsdon C (2016). "Establishing a framework for Open Geographic Information science." *International Journal of Geographical Information Science*, **30**(8), 1507–1521.

Spiridon L, Minh DD (2017). "Hamiltonian Monte Carlo with Constrained Molecular Dynamics as Gibbs Sampling." *Journal of Chemical Theory and Computation*, **13**(10), 4649–4659.

Theodoridis S, Nogués-Bravo D, Conti E (2019). "The role of cryptic diversity and its environmental correlates in global conservation status assessments: Insights from the threatened bird's-eye primrose (Primula farinosa L.)." *Diversity and Distributions*, **25**(9), 1457–1471.

van der Walt S, Colbert SC, Varoquaux G (2011). "The NumPy Array: A Structure for Efficient Numerical Computation." *Computing in Science Engineering*, **13**(2), 22–30. doi:10.1109/MCSE.2011.37.

van Lieshout MNM, Baddeley AJ (1996). "A nonparametric measure of spatial interaction in point patterns." *Statistica Neerlandica*, **50**(3), 344–361. doi:10.1111/j.1467-9574.1996.tb01501.x.

van Rossum G (1989). "Glue it all together with Python." In *OMG-DARPA-MCC Workshop on Compositional Software Architecture*. CNRI.

Vaz E, Miki J, de Noronha T, Cusimano M (2017). "New methods for resilient societies: The geographical analysis of injury data." *Journal of Spatial and Organizational Dynamics*, **5**(1), 12–26.

Wilson AG (1971). "A family of spatial interaction models, and associated developments." **3**, 1–32. URL https://illiad.lib.asu.edu/illiad/illiad.dll?Action=10&Form=75&Value=1221819.

Wolf LJ, Knaap E, Rey S (2019a). "Geosilhouettes: Geographical Measures of Cluster Fit." *Environment and Planning B: Urban Analytics and City Science*, p. 2399808319875752. ISSN 2399-8083. doi:10.1177/2399808319875752.

Wolf LJ, Rey SJ, Oshan TM (2019b). "Open code is not enough: towards a replicable future for geographic data science." doi:10.31235/osf.io/3hbnt. Position paper for the Third Geospatial Software Institute Workshop on Strategic Planning and Governance.

Wood SN (2006). *Generalized Additive Models: An Introduction with R*. CRC Press, Boca Raton.

Affiliation:

Sergio J. Rey
Center for Geospatial Sciences
University of California Riverside
900 University Ave
Riverside CA, 92521, USA
E-mail: sergio.rey@ucr.edu
URL: https://sergerey.org

Luc Anselin
Center for Spatial Data Science
University of Chicago
Searle Lab
5735 S Ellis Ave, Room 230
Chicago, Illinois 60637
E-mail: anselin@uchicago.edu

Dani Arribas-Bel
Geographic Data Science Lab
Department of Geography and Planning
University of Liverpool
Roxby Building, 74 Bedford St S,
Liverpool, L69 7ZT, United Kingdom
E-mail: D.Arribas-Bel@liverpool.ac.uk
URL: https://darribas.org

Levi John Wolf
Center for Multilevel Modelling
School of Geographical Sciences
University of Bristol
University Road, Clifton,
Bristol, BS8 1SS, United Kingdom
E-mail: levi.john.wolf@bristol.ac.uk
URL: https://ljwolf.org
ORCID: 0000-0003-0274-599X

Taylor M. Oshan
Center for Geospatial Information Science
Department of Geographical Sciences
University of Maryland, College Park
Lefrak Hall, 7251 Preinkert Drive
College Park, MD 20742, United States
E-mail: toshan@umd.edu

James David Gaboardi
Department of Geography
The Pennsylvania State University
302 Walker Building
University Park, PA 16802, USA
E-mail: jgaboardi@psu.edu
URL: https://github.com/jGaboardi

Elijah Knaap
Center for Geospatial Sciences
University of California-Riverside
900 University Ave
Riverside, CA 92521, USA
E-mail: knaap@ucr.edu
URL: https://knaaptime.com
ORCID: 0000-0001-7520-2238

Wei Kang
Center for Geospatial Sciences
University of California Riverside
900 University Ave
Riverside CA, 92521, USA
E-mail: weikang@ucr.edu
URL: https://weikang9009.github.io
ORCID: 0000-0002-1073-7781

Hu Shao
Environmental Systems Research Institute (Esri)
380 New York Street
Redland, CA, 92373, USA
E-mail: HShao@esri.com
ORCID: 0000-0003-3852-3176

Renan Xavier Cortes
Center for Geospatial Sciences
University of California Riverside
900 University Ave
Riverside CA, 92521, USA
E-mail: renanxcortes@gmail.com
URL: https://renanxcortes.github.io

Stefanie Lumnitz
Department of Forest Resource Management
University of British Columbia
2045 - 2424 Main Mall
Vancouver, BC V6T 1Z4, Canada
E-mail: stefanie.lumnitz@gmail.com
ORCID: 0000-0002-7007-5812

Pedro Amaral
Department of Economics
Centre for Development and Regional Planning (Cedeplar)
Universidade Federal de Minas Gerais (UFMG)
Av. Antonio Carlos 6627
Belo Horizonte, MG, 31270-901, Brazil
E-mail: pedroamaral@cedeplar.ufmg.br

Ziqi Li
School of Geographical Sciences and Urban Planning
Arizona State University
975 S Myrtle Ave
Tempe, AZ 85281, USA
E-mail: liziqi1992@gmail.com
ORCID: 0000-0002-6345-4347

Ran Wei
Center for Geospatial Sciences
University of California Riverside
900 University Ave
Riverside CA, 92521, USA
E-mail: ran.wei@ucr.edu

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume VV, Issue II *Submitted:* yyyy-mm-dd
MMMMMM YYYY *Accepted:* yyyy-mm-dd