The Journal of Open Source Software

Pyinterpolate: Spatial Interpolation in Python for point measurements and aggregated datasets

Szymon Moliński[1]

1 Data Lions company, Poland, https://datalions.eu

DOI: 10.21105/joss.02869

Software

- Review
- Repository
- Archive

Editor: Hugo Ledoux

Reviewers:

- @chrisbrunsdon
- @kenohori
- @sdesabbata

Submitted: 26 October 2020
Published: 28 June 2021

Summary

Spatial Interpolation techniques are used to interpolate values at unknown locations and/or filter and smooth existing data sources. Those methods work for point observations and areal aggregates. The basic idea under this family of algorithms is that every point in a space can be described as a function of its neighbors values weighted by the relative distance from the analyzed point. It is known as the Tobler's First Law of Geography, which states: *everything is related to everything else, but near things are more related than distant things* (Tobler, 1970).

Kriging technique designed for mining applications exploits this statement formally and nowadays it has gained a lot of attention outside the initial area of interest. Today *kriging* is a set of methods which can be applied to problems from multiple fields: environmental science, hydrogeology, natural resources monitoring, remote sensing, epidemiology and ecology and even computer science (Chilès & Desassis, 2018). Most commonly Kriging interpolates values from point measurements or regular block units but many real-world datasets are different. Especially challenging are measurements of rates over areas of irregular shapes and sizes, as example administrative units in every country (Goovaerts, 2007).

Pyinterpolate is designed to tackle a problem of areas of irregular shapes and sizes with Area-to-Area and Area-to-Point Poisson Kriging functions. With those algorithms Pyinterpolate became an interpolation and filtering tool which is useful for social, environmental and public health sciences. Moreover, the package offers basic Kriging and Inverse Distance Weighting techniques and can be utilized in every field of research where geostatistical (distance) analysis gives meaningful results. Pyinterpolate merges basic Kriging techniques with more sophisticated Area-to-Area and Area-to-Point Poisson Kriging methods.

Statement of need

Pyinterpolate is a Python package for spatial interpolation and it is designed to perform predictions from point measurements and areal aggregates of different sizes and shapes. Pyinterpolate automates tasks performed by spatial statisticians, it helps with data exploration, semivariogram estimation and kriging predictions. Thing that makes Pyinterpolate different from other spatial interpolation packages is the ability to perform Kriging of areas of different shapes and sizes and this type of operation is extremely important in the context of social, medical and ecological sciences.

Importance of areal (block) Kriging

Areas of irregular shapes and sizes are especially challenging for analysis and modeling. The ability to transform areal aggregates into point support maps is desired by many applications.

As an example in public health studies data is aggregated over large areas due to the protection of citizens' privacy but this process introduces bias to modeling and makes policy-making more complex. The main three reasons behind transformation of choropleth maps with aggregated counts into point support models are:

1. The presence of extreme unreliable rates that typically occur for sparsely populated areas and rare events.
2. The visual bias resulting from aggregation of data over administrative units with various shapes and sizes.
3. The mismatch of spatial supports for aggregated data and explanatory variables. This prevents their direct use in models based on the correlation (Goovaerts, 2006).

In this context Area-to-Area Poisson Kriging serves as the noise-filtering algorithm or areal interpolation model and Area-to-Point Poisson Kriging is designed to interpolate and transform values and to preserve coherence of the prediction so the sum of average of disaggregated estimates is equal to the baseline area value (Goovaerts & Gebreab, 2008). Area-to-Point Poisson Kriging can be useful in the chained-models systems where change of support is required to perform a study.

Researchers may use centroids of areas and perform point kriging over a prepared regular point grid. However this method has pitfalls. Different sizes and shapes of units may lead to imbalanced variogram point pairs per lag. Centroid-based approach does not catch spatial variability of the linked variable, for example population density over area in the context of infection rates.

To disaggregate areal data into point support one must know point support covariance and/or semivariance of a regionalized variable. Then the semivariogram deconvolution is performed. In this iterative process experimental semivariogram of areal data is transformed to fit the semivariogram model of a linked point support variable. Example of it is the use of spatial distribution of population to transform a semivariogram of disease rates which are the number of cases divided by population. Semivariogram deconvolution is the core step of the Area-to-Area and Area-to-Point Poisson Kriging operations. Poisson Kriging is widely used in the social sciences, epidemiology and spatial statistics (Goovaerts, 2007; Goovaerts & Gebreab, 2008; Kerry et al., 2013).

Interpolation methods within Pyinterpolate

Package performs six types of spatial interpolation at the time of paper writing; five types of Kriging and inverse distance weighting:

1. **Ordinary Kriging** which is a universal method for point interpolation.
2. **Simple Kriging** which is useful when the mean of the spatial process is known and it is used for the point interpolation.
3. **Centroid-based Poisson Kriging**. This method of Kriging is based on the assumption that each block can be collapsed into its centroid. It is much faster than Area-to-Area and Area-to-Point Poisson Kriging but introduces bias related to the transformation of areas into single points. It is used for areal interpolation and filtering.
4. **Area-to-Area Poisson Kriging**. Point support is included in the analysis and if it varies over area. Model is able to catch this variation. It is used for areal interpolation and filtering.
5. **Area-to-Point Poisson Kriging**. Areal support is deconvoluted in regards to the point support. Output map has spatial resolution of the point support while coherence of analysis is preserved (sum of rates is equal to the output of Area-to-Area Poisson Kriging). It is used for point-support interpolation and data filtering.

User starts with semivariogram exploration and modeling. Next researcher or algorithm chooses the theoretical model which best fits the semivariogram. This model is used to predict values at unknown locations. Areal data interpolation, especially transformation from areal aggregates into point support maps, requires deconvolution of areal semivariogram. This is an automatic process which can be performed without prior knowledge of kriging and spatial statistics. The last step is Kriging itself. Poisson Kriging is especially useful for counts over areas. On the other spectrum is Ordinary Kriging which is an universal technique which works well with multiple point data sources. Predicted data is stored as a DataFrame known from the *Pandas* and *GeoPandas* Python packages. Pyinterpolate allows users to transform given point data into a regular numpy array grid for visualization purposes and to perform large-scale comparison of different kriging techniques prediction output. Use case with the whole scenario is available in the paper package repository.

Package performs many steps automatically. User has the option to control prediction flow with Python optional parameters in the function call. Package was initially developed for epidemiological study, where areal aggregates of infections were transformed to point support population-at-risk maps and multiple potential applications follow this algorithm. Initial field of study (epidemiology) was the reason behind automation of many tasks related to data modeling. It is assumed that users without a wide geostatistical background may use Pyinterpolate for spatial data modeling and analysis, especially users which are observing processes related to the human population.

The example of a process where deconvolution of areal counts and semivariogram regularization occurs is presented in the Figure 1.

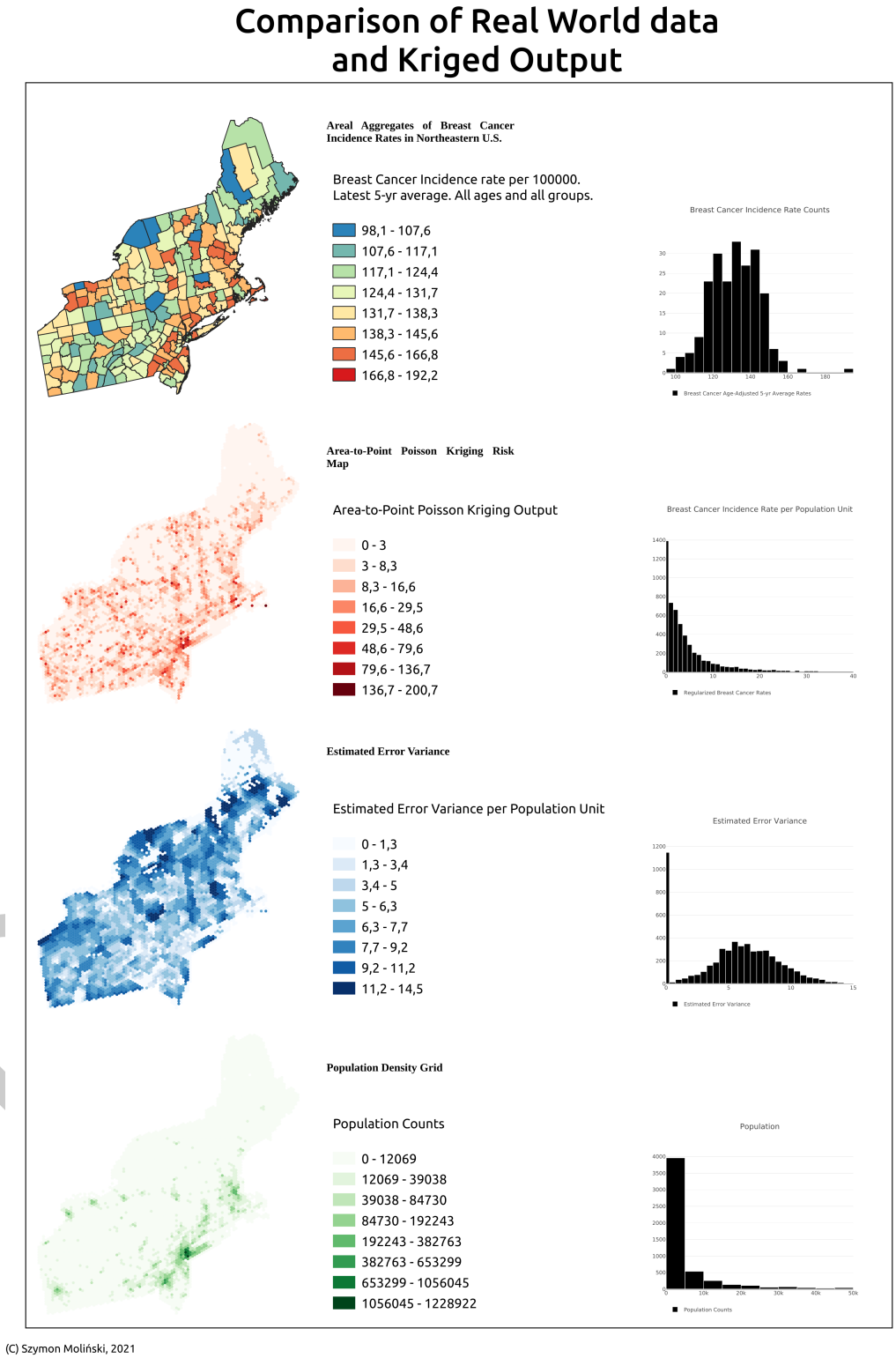

Figure 1: Structure of Pyinterpolate package.

Methodology

Chapter presents general methodology of calculations within package. Concrete use case is presented in document here. Comparison of algorithm with the **gstat** package is available here.

Spatial Interpolation with Kriging

Kriging, which is the baseline of the Pyinterpolate package, is an estimation method that gives the best unbiased linear estimates of point values or block averages (Armstrong, 1998). Kriging minimizes variance of a dataset with missing values. Baseline technique is the **Ordinary Kriging** where value at unknown location $\hat{z}$ is estimated as a linear combination of K neighbors with value z and weights λ assigned to those neighbors (1).

(1)

$$\hat{z} = \sum_{i=1}^{K} \lambda_i z_i$$

Weights λ are a solution of following system of linear equations (2):

(2)

$$\sum_{j=1} \lambda_j C(x_i, x_j) - \mu = \bar{C}(x_i, V); i = 1, 2, ..., K$$

$$\sum_{i} \lambda_i = 1$$

where $C(x_i, x_j)$ is a covariance between points x_i and x_j, $\bar{C}(x_i, V)$ is an average covariance between point x_i and all other points in a group (K points) and μ is a process mean. The same system may be solved with semivariance instead of covariance (3):

(3)

$$\sum_{i=1} \lambda_j \gamma(x_i, x_j) + \mu = \bar{\gamma}(x_i, V); i = 1, 2, ..., K$$

$$\sum_{i} \lambda_i = 1$$

where $\gamma(x_i, x_j)$ is a semivariance between points x_i and x_j, $\bar{\gamma}(x_i, V)$ is an average semivariance between point x_i and all other points. Semivariance is a key concept of spatial interpolation. It is a measure of a dissimilarity between observations in a function of distance. Equation (4) is a experimental semivariogram estimation formula.

(4)

$$\frac{1}{2N} \sum_{i}^{N} (z_{(x_i+h)} - z_{x_i})^2$$

where z_{x_i} is a value at location x_i and $z_{(x_i+h)}$ is a value at translated location in a distance h from x_i.

In the next step theoretical models are fitted to the experimental curve. Pyinterpolate package implements linear, spherical, exponential and gaussian models but many others are applied for specific cases (Armstrong, 1998). Model with the lowest error is used in (3) to estimate γ parameter.

Ordinary Kriging is one of the classic Kriging types within the package. **Simple Kriging** is another available method for point interpolation. Simple Kriging may be used when the process mean is known over the whole sampling area. This situation rarely occurs in real world. It can be observed in places where sampling density is high (Armstrong, 1998). Simple Kriging system is defined as:

(5)
$$\hat{z} = R + \mu$$

where μ is a process mean and R is a residual at a specific location. Residual value is derived as the first element (denoted as $\mathbf{1}$) from:

(6)
$$R = ((Z - \mu) \times \lambda)\mathbf{1}$$

Number of values depends on the number of neighbours in a search radius, similar to equation (1) for Ordinary Kriging. λ weights are the solution of following function:

(7)
$$\lambda = K^{-1}(\hat{k})$$

where K is a semivariance matrix between each neighbour of size NxN and k is a semivariance between unknown point and known points of size $Nx1$.

Package allows use of the three main types of Poisson Kriging: Centroid-based Poisson Kriging, Area-to-Area Poisson Kriging and Area-to-Point Poisson Kriging. Risk over areas (or points) for each type of Poisson Kriging is defined similarly to the equation (1) but weights associated with the λ parameter are estimated with additional constraints related to the population weighting. The spatial support of each unit needs to be accounted for in both the semivariogram inference and kriging. Full process of areal data Poisson Kriging is presented in (Goovaerts, 2006) and semivariogram deconvolution which is an intermediate step in Poisson Kriging is described in (Goovaerts, 2007).

Modules

Pyinterpolate is designed from seven modules and they cover all operations needed to perform spatial interpolation: from input/output operations, data processing and transformation, semivariogram fit to kriging interpolation. Figure 2 shows package structure.

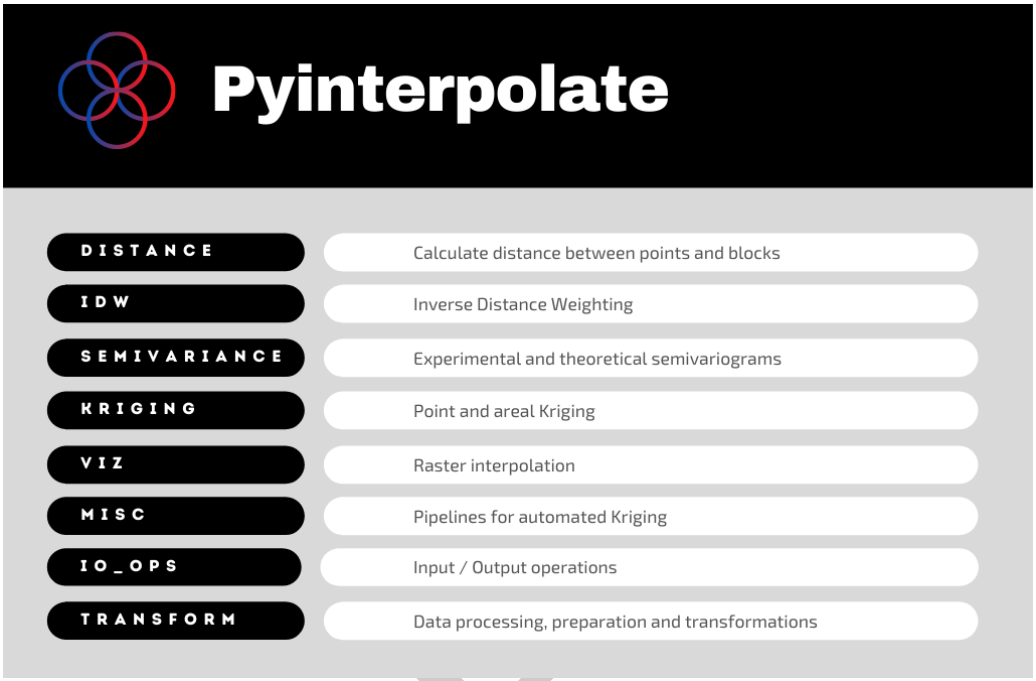

Figure 2: Structure of Pyinterpolate package.

Modules follow typical data processing and modeling steps. The first module is **io_ops** which reads point data from text files and areal or point data from shapefiles, then changes data structure for further processing. **Transform** module is responsible for all tasks related to changes in data structure during program execution. Sample tasks are:

- finding centroids of areal data,
- building masks of points within lag.

Functions for distance calculation between points and between areas (blocks) are grouped within **distance** module. **Semivariance** is most complex part of Pyinterpolate package. It has three special classes for calculation and storage of different types of semivariograms (experimental, theoretical, areal and point types). **Semivariance** module has other functions important for spatial analysis:

- function for experimental semivariance / covariance calculation,
- weighted semivariance estimation,
- variogram cloud preparation,
- outliers removal.

Kriging module contains three main types of models Ordinary and Simple Kriging models as well Poisson Kriging of areal counts models. Areal models are derived from (Goovaerts & Gebreab, 2008), simple Kriging and ordinary Kriging models are based on (Armstrong, 1998).

It is possible to show output as numpy array with **viz** module and to compare multiple kriging models on the same dataset with **misc** module. Evaluation metric for comparison is an average root mean squared error over multiple random divisions of a passed dataset.

Comparison to Existing Software

Pyinterpolate is one package from a large ecosystem of spatial modeling and spatial interpolation packages written in Python. The main difference between Pyinterpolate and other packages is focus on areal deconvolution methods and Poisson Kriging techniques useful for ecology, social science and public health studies in the presented package. Potential users may choose other packages if their study is limited to point data interpolation.

The most similar and most important package from Python environment is **PyKrige** (Murphy et al., 2020). PyKrige is designed especially for point kriging. PyKrige supports 2D and 3D ordinary and universal Kriging. User is able to incorporate own semivariogram models and/or use external functions (as example from **scikit-learn** package (Pedregosa et al., 2011)) to model drift in universal Kriging. Package is well designed, and it is actively maintained.

GRASS GIS (GRASS Development Team, 2020) is well-established software for vector and raster data processing and analysis. GRASS contains multiple modules and GRASS functionalities can be accessed from multiple interfaces: GUI, command line, C API, Python APU, Jupyter Notebooks, web, QGIS and R. GRASS has two functions for spatial interpolation: `r.surf.idw` and `v.surf.idw`. Both use Inverse Distance Weighting technique, first interpolated raster files and second vectors (points).

PySAL is next GIS / geospatial package which can be used to interpolate missing values – but this time at areal scale. Package's **tobler** module can be used to interpolate areal values of specific variable at different scales and sizes of support (knaap et al., 2020). Moreover, package has functions for multisource regression, where raster data is used as auxiliary information to enhance interpolation results. Conceptually tobler package is close to the Pyinterpolate, where main algorithm transforms areal data into point support derived from auxiliary variable.

R programming language offers **gstat** package for spatial interpolation and spatial modeling (Pebesma, 2004). Package is designed for variogram modelling, simple, ordinary and universal point or block kriging (with drift), spatio-temporal kriging and sequential Gaussian (co)simulation. Gstat is a solid package for Kriging and spatial interpolation and has the largest number of methods to perform spatial modelling. The main difference between gstat and Pyinterpolate is availability of area-to-point Poisson Kriging based on the algorithm proposed by Goovaerts (Goovaerts, 2007) in Pyinterpolate package. Comparison to **gstat** is available in the paper repository.

Appendix

1. **Paper repository** with additional materials
2. **Package repository**

References

Armstrong, M. (1998). *Basic linear geostatistics*. Springer. https://doi.org/10.1007/978-3-642-58727-6

Chilès, J.-P., & Desassis, N. (2018). Fifty years of kriging. In B. S. Daya Sagar, Q. Cheng, & F. Agterberg (Eds.), *Handbook of mathematical geosciences: Fifty years of IAMG* (pp. 589–612). Springer International Publishing. https://doi.org/10.1007/978-3-319-78999-6_29

Goovaerts, P. (2006). Geostatistical analysis of disease data: Accounting for spatial support and population density in the isopleth mapping of cancer mortality risk using area-to-point

poisson kriging. *International Journal of Health Geographics*, *5*. https://doi.org/10.1186/1476-072X-5-52

Goovaerts, P. (2007). Kriging and semivariogram deconvolution in the presence of irregular geographical units. *Mathematical Geosciences*, *40*, 101–128. https://doi.org/10.1007/s11004-007-9129-1

Goovaerts, P., & Gebreab, S. (2008). How does poisson kriging compare to the popular BYM model for mapping disease risks? *International Journal of Health Geographics*, *7*, 6. https://doi.org/10.1186/1476-072x-7-6

GRASS Development Team. (2020). *Geographic resources analysis support system (GRASS GIS) software*. Open Source Geospatial Foundation. https://grass.osgeo.org

Kerry, R., Goovaerts, P., Smit, I. P. J., & Ingram, B. R. (2013). A comparison of multiple indicator kriging and area-to-point poisson kriging for mapping patterns of herbivore species abundance in kruger national park, south africa. *International Journal of Geographical Information Science*, *27*, 47–67. https://doi.org/10.1080/13658816.2012.663917

knaap, eli, Cortes, R. X., Rey, S., Gaboardi, J., & Frontiera, P. (2020). *Pysal/tobler: Release v0.5.4* (Version v0.5.4) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.4385980

Murphy, B., Müller, S., & Yurchak, R. (2020). *GeoStat-framework/PyKrige v1.5.1* (Version v1.5.1) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.3991907

Pebesma, E. J. (2004). Multivariable geostatistics in s: The gstat package. *Computers & Geosciences*, *30*(7), 683–691. https://doi.org/https://doi.org/10.1016/j.cageo.2004.03.012

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography*, *46*, 234. https://doi.org/10.2307/143141