An User-friendly Python Application for Exploratory and Structural Spatial Dependence Analysis for Sample Points of Spatial Attributes

Carlos A. Felgueiras[1], Jussara O. Ortiz[1], Eduardo C. G. Camargo[1]

[1]Divisão de Processamento de Imagens (DPI) – Instituto Nacional de Pesquisas Espaciais (INPE)
[1]Caixa Postal 515 – São José dos Campos – SP – Brazil

{carlos.felgueiras,jussara.ortiz,eduardo.camargo}@inpe.br

Abstract. *This article describes the functionalities and implementation details of the PyESSDA, an easy to use Python application, that allows for performing exploratory and structural spatial dependence analysis on a set of sample points representing geographic attributes. Exploratory analysis makes it possible to view the sample set in 2D and 3D projections, to report its univariate statistics and to generate its histogram. A semivariogram map can be generated to evaluate the isotropic or anisotropic spatial behavior of the investigated attribute. The analyzes of spatial dependencies, for determining the attribute spatial correlation structures, comprise the interactive creation of experimental and mathematical semivariograms. The functionalities of the developed application are illustrated with a set of real elevation data sampled in a region of Jacareí city of São Paulo state, Brazil.*

Resumo. *Este artigo descreve as funcionalidades e detalhes de implementação do PyESSDA, um aplicativo Python, de fácil uso, que permite realizar análises exploratória e estrutural de dependência espacial sobre um conjunto de pontos amostrais representando atributos geográficos. A análise exploratória possibilita visualizar o conjunto amostral em projeções 2D e 3D, relatar estatísticas básicas e visualizar o histograma dos dados de entrada. Um mapa de semivariograma pode ser gerado para se avaliar o comportamento espacial isotrópico ou anisotrópico do atributo investigado. As análises de dependências espaciais, para se determinar as estruturas de correlação espacial do atributo, são realizadas pela criação interativa de semivariogramas experimentais e matemáticos. A fim de ilustrar as funcionalidades do aplicativo, utilizam-se um conjunto de pontos de elevação amostrados em uma região da cidade de Jacareí do estado de São Paulo, Brasil.*

1. Introduction

Spatial analysis is a research paradigm that provides a unique set of techniques and methods for analyzing events — events in a very general sense — that are distributed in geographical space [Bailey 1995 and Fischer 2006]. As subset of general spatial analysis, the Exploratory Spatial Data Analysis (ESDA) and the Structural Spatial Dependence Analysis (SSDA) of spatial attributes, frequently sampled as a set of points, spatial locations, are important issues for modeling the behavior of spatial attributes inside a geographical region in Geographical Information System (GIS) applications [Anselin et al. 2006, Burrough 1998].

Python is an interpreted, high-level, easy to learn, general-purpose and powerful programming language. Created by Guido van Rossum and first released in 1991, Python's design philosophy emphasizes code readability with its notable use of significant whitespace. A very helpful tutorial of the Python language can be found in [Rossum 2018]. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects [Kuhlman 2012]. The python language has been widely used in applications involving manipulation and analysis of spatial data [Rey and Anselin 2007].

In this context, this article describes the functionalities and implementation details of a Python application, named PyESSDA (Python application for Exploratory and Structural Spatial Dependence Analysis), for accomplishing the ESDA and the SSDA on a set of sampled points of spatial attributes. The ESDA application interface contains methods for plotting the 2D and 3D sample sets, reporting their univariate statistics, and visualizing their histogram. A semivariogram map, also known as surface or anisotropy map [Robertson 2008], can be plotted in order to determine the attribute spatial anisotropy. The SSDA interface contains tools for interactively creating experimental and mathematical semivariograms that model the attribute spatial correlations. The application implementation aims to enable users easily creating semivariograms that better represent the attribute variability mainly for short distances, smaller than the semivariogram range.

The semivariograms, obtained with the python application, are used mainly as input for geostatistical procedures of estimations and simulations of spatial attributes [Isaaks and Shrivastava 1989, Deutsch and Journel 1992]. Even in deterministic estimation approaches, as the Inverse Distance Weighted (IDW) for example, the range of the resulting semivariograms is an important information to define the search radius to find the nearest neighbors to be used in this prediction method.

In order to illustrate the PyESSDA functionalities, a case study is presented using a set of real elevation information sampled in a region of Jacareí municipality of São Paulo state, Brazil.

2. Concepts

2.1. Exploratory Spatial Data Analysis

ESDA generally comprises a series of techniques which are used to statistically analyze spatial data and mine necessary knowledge of features' spatial structure and correlation (Haining and Wise, 1997, Symanzik, 2013).

For spatial attributes, sampled as a set of points, the most common tools for ESDA are: i) visualizing the data spatial distribution in 2D and 3D, presentations that help the analyst to better understand the spatial attribute sampling geometry, such as the occurrence of clusters, for example; ii) reporting univariate and multivariate basic statistics, such as minimum and maximum, mean, variance, median, skewness, kurtosis and quantile values, that summarize and describe the distribution of the investigated attribute; iii) plotting histograms, normal graphics, and others, make possible a faster perception of the variable distribution.

For geostatistics analysis purposes the input sample information must be stationary with constant mean and variance depending only on the spatial distance vector. This requirement can be achieved by using residuals information obtained from taking off the tendency and analyzing regional variabilities of the original sample set [Deutsch and Journel 1998, Goovaerts 1997].

2.2. Structural Spatial Dependence Analysis

The SSDA comprises two steps: i) identify the spatial directional continuity of the investigated attribute which is attained through a semivariogram map; ii) detect the spatial dependence structure presented in the attribute, accomplished by building experimental semivariograms and fitting them with empirical or mathematical models. A semivariogram is a graphic that represents the variability of the semivariogram values related to the spatial separation vector **h**.

2.2.1. Semivariogram Maps

Semivariogram map is employed to identify if the spatial continuity of the phenomenon occurs in some privileged directions, the anisotropic case, or equally in all directions, the isotropic case. When the range (geometric anisotropy) or sill (zonal anisotropy) or both (combined anisotropy) vary according to the angular direction considered, there is anisotropic behavior of the attribute. In the case of invariant spatial continuity, isotropy occurs (Bettini, 2007).

The anisotropy map is an image, or a raster representation, that contains the experimental semivariogram value for each point of the raster grid. Each semivariogram value is evaluated, from the sample set, considering the distance and the angle of the grid point to its origin (location 0,0). The isotropy or anisotropy can be easily detected by visual inspection. Figure 1(a) shows an isotropy and 1(b) an anisotropy case.

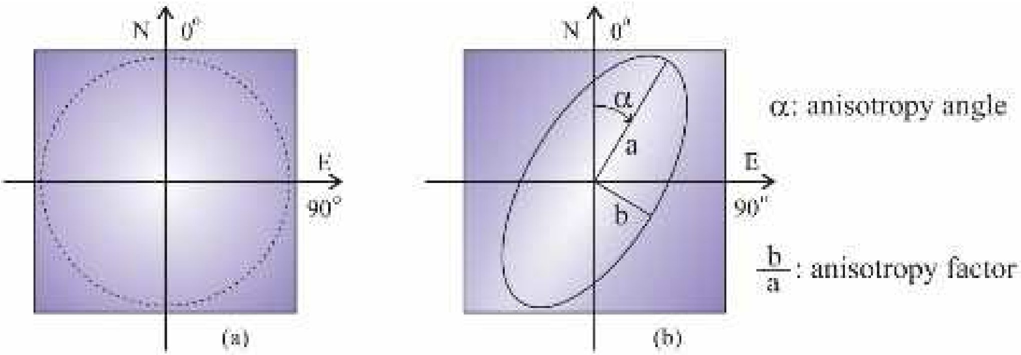


Figure 1. Illustration of spatial attribute continuities: (a) isotropic and (b) anisotropic

2.2.2. Experimental Semivariograms

Experimental semivariograms are built directly from the set of sample points. A semivariogram is a graphic that represents the variability of the semivariogram values related to the spatial separation vector **h**, as can be seen as black dots in Figure 2. Thus, the semivariogram describes and models the structural spatial dependence of geographic attributes. The experimental semivariances are estimated using the Equation 1.

$$\hat{\gamma}_{(\mathbf{h})} = \frac{1}{2N(\mathbf{h})}\sum_{j=1}^{N(\mathbf{h})}\left[z(\mathbf{u}_j) - z(\mathbf{u}_j + \mathbf{h})\right]^2 \qquad (1)$$

where $z(\mathbf{u}_j)$ and $z(\mathbf{u}_j+\mathbf{h})$ are the j-th values of the attribute Z of the samples separated by the direction and distance of vector **h**, and N(**h**) is the number of the sample pairs of **h.**

The experimental curve points are obtained first defining a directional angle, along with an angle tolerance, and, as distance parameters, the number of lags, the lag increment and a tolerance around the increment. For each lag h, the experimental semivariogram value $\hat{\gamma}_{(\mathbf{h})}$ is assessed from the set of the pair of points suitable for the angular and directional parameter values by means of the Equation 1.

2.2.3. Empirical Semivariograms

A mathematical model is used to fit the graphic points of the experimental semivariogram. This model is considered the mathematical or empirical semivariogram and will be used to obtain spatial correlation values for geostatistical procedures, for example. The spherical, exponential and gaussian models, illustrated in Figure 2, are the most widely used models in practice. The mathematical equations of these models are [Deutsch and Journel 1992]:

$$\text{Spherical: } \gamma(h) = c.\,\mathrm{Sph}\left(\frac{h}{a}\right) = \begin{cases} c.\left[1.5\frac{h}{a} - 0.5\left(\frac{h}{a}\right)^3\right], \text{if } h \le a \\ c, \qquad\qquad\qquad\qquad if\ h > a \end{cases} \qquad (2)$$

$$\text{Exponential: } \gamma(h) = c.\mathrm{Exp}\left(\frac{h}{a}\right) = c.\left[1 - \exp\left(-3\frac{h}{a}\right)\right] \qquad (3)$$

$$\text{Gaussian: } \gamma(h) = c.\mathrm{Exp}\left(\frac{h}{a}\right) = c.\left[1 - \exp\left(-3\left(\frac{h}{a}\right)^2\right)\right] \qquad (4)$$

Where c is the contribution and a is the range of the experimental semivariogram parameters. Also, the semivariogram can present a nugget effect, explained by measure or low scale errors, and in this case the sill is the contribution added to the nugget effect.

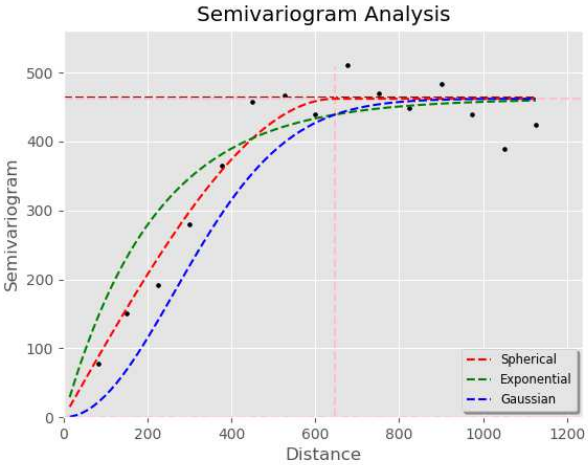


Figure 2. Examples of different mathematical models used to fit an experimental semivariogram (black dot marks)

3. Application Implementation

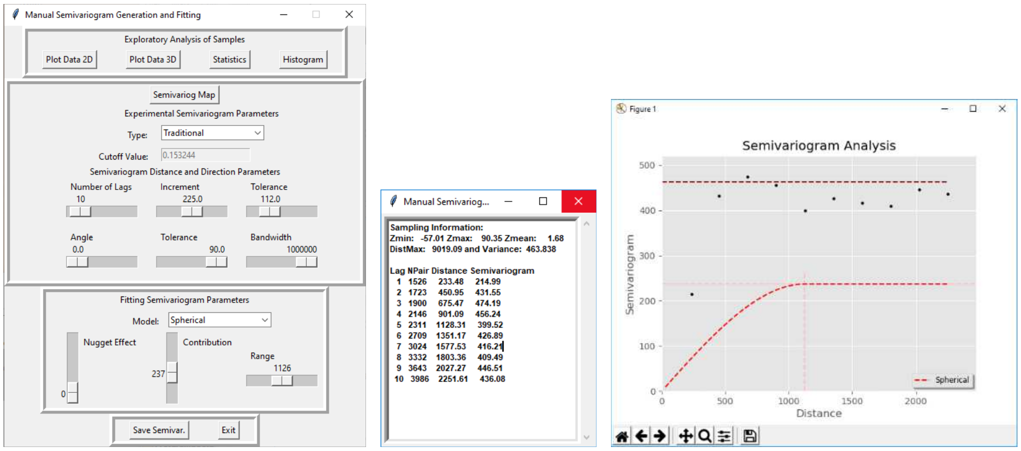
The following steps were done in order to implement the PyESSDA application:

1. Specification of the tools that compose the ESDA as 2D and 3D data plotting, data statistics reporting and histogram visualization.
2. Python codification of the widgets to be used in the ESDA submodule
3. Python codification and tests of the functionalities presented in the ESDA submodule.
4. Specification, codification and tests of the semivariogram map that is called as a button inside the experimental semivariogram submodule
5. Specification of the tools that compose the experimental semivariogram parameters, distances and angle directions along with tolerances, submodule
6. Python codification and test of the functionalities available in the experimental semivariogram parameters submodule
7. Specification of the tools that compose the fitting semivariogram parameters (mathematical model, nugget effect, sill and range) submodule
8. Inclusion of the Save Semivariogram and Exit buttons at the end of the window application

4. Results and Analysis

4.1. Activating the Application

On activating the application, the user has to choose an input csv, comma delimited, file containing a header with the x, y, z1 and z2 (optional) names followed by the respective sample numerical data values, each x, y and z values in a new line. Figure 3 depicts the first 3 windows that are opened just after the application has been activated and has been read the csv file data with a sample set.



Experimental Semivariogram Parameters

Type: Traditional

Cutoff Value: 0.153244

Semivariogram Distance and Direction Parameters

Number of Lags: 10

Increment: 225.0

Tolerance: 112.0

Angle: 0.0

Tolerance: 90.0

Bandwidth: 1000000

Fitting Semivariogram Parameters

Model: Spherical

Nugget Effect: 0

Contribution: 237

Range: 1126

```
Sampling Information:
Zmin:  -57.01 Zmax:   90.35 Zmean:    1.68
DistMax:  9019.09 and Variance:  463.838

Lag NPair Distance  Semivariogram
  1   1526    233.48    214.99
  2   1723    450.95    431.55
  3   1900    675.47    474.19
  4   2146    901.09    456.24
  5   2311   1128.31    399.52
  6   2709   1351.17    426.89
  7   3024   1577.53    416.21
  8   3332   1803.36    409.49
  9   3643   2027.27    446.51
 10   3986   2251.61    436.08
```

(a) (b) (c)

Figure 3. First windows output: (a) the main window; (b) the report of the experimental semivariogram data and (c) the graphic of semivariograms

Figure 3(a) shows the main window of the application, 3(b) depicts a report window with the numerical information of an experimental semivariogram and 3(c) presents the graphic of semivariograms. The Tkinter Python standard GUI (Graphical User Interface) package is used to design an create the widgets presented in figure 3(a). In the application, the following commands were used to import tkinter functionalities: *from tkinter package from tkinter import *; from tkinter import ttk; from tkinter import scrolledtext; from tkinter import filedialog; from tkinter import messagebox*.

The input csv file was read in a df data frame using the command *filename=filedialog.askopenfilename*, to ask for the name of the input file, and the command *df = pandas.read_csv("filename")*, to read the file in a df pandas data frame structure that requires the command *import pandas*.

4.2. Exploratory Analysis of Samples

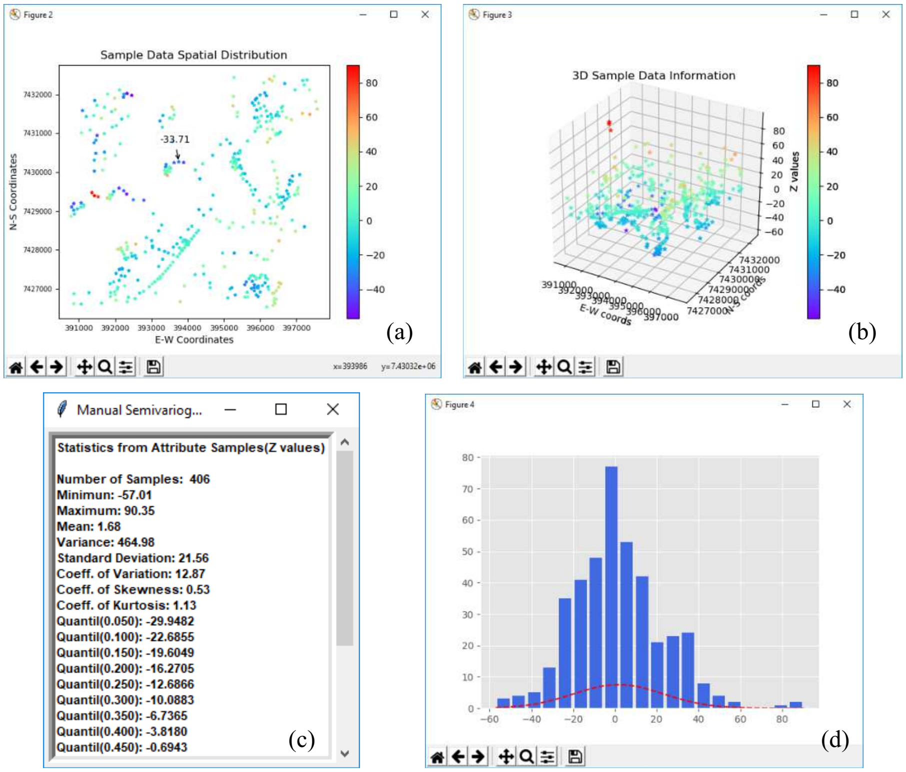
Exploratory analysis of the samples can be performed using the buttons offered at the top of the main window of the PyESSDA. The available analysis options are: 2D plot, 3D plot, statistics and histogram of the input data. Figure 4 illustrates these options. In this work it was used as input data a set of 406 samples of altimetry information from a geographic region in the municipality of Jacareí, São Paulo state, Brazil. The limits of the region are: W 46^o 4' 4.98'' to W 46^o 0' 2.82'' and S 23^o 16' 2.91'' to S 23^o 12' 47.23''.

Besides plotting the distribution, as shown in Figure 4(a), the user can read in this graphic the x, y and z values of each sample. The 3D plotting of Figure 4(b) allows 3 axis graphic rotations. In Figures 4(a) and 4(b), each sample is plotted in a colored mark according its z value following the legend on the right side of the graphic.

Univariate statistics are reported in the scrolled text widget of Figure 4(c) including the percentiles 0.05 to 0.95 of the z values. The histogram is plotted in blue in Figure 4(d) along with the respective mean and variance gaussian curve. All the 2D data visualizations, figures and styles, of this application were done by the Python plotting library Matplotlib setting the initial commands: *import matplotlib.pyplot as plt* and *from matplotlib import style*. For example, for plotting the samples in Figure 2(a) the command *plt.scatter* was used, for plotting the samples in Figure 2(b) the commands *ax3 = fig3.add_subplot(111, projection='3d')* and *ax3.scatter* were used. The legends were included in Python figures with the command *plt.colorbar()* after defining a color map by the command *cmap= plt.cm.rainbow*, for example, inside the plotting commands. The histograms of Figure 2(d) were visualized using the commands *plt.bar*, for plotting the blue bars, and *plt.plot*, for the showing the red dashed line.

4.3. Structural Analysis of Samples

As pointed out in section 2.2, structural analysis of the samples comprises the building of the experimental semivariograms and the fitting them with a empirical, or mathematical, models. The Structural Analysis of the PyESSDA application allows to create Traditional, Indicator Continuous, Indicator Categorical and Traditional Crossed Univariate Directional and Omnidirectional Experimental Semivariograms. The experimental semivariogram is then fitted with a Conceptual Semivariogram by means of a Spherical, Exponential or Gaussian mathematical model.



```
Statistics from Attribute Samples(Z values)

Number of Samples:  406
Minimun: -57.01
Maximum: 90.35
Mean: 1.68
Variance: 464.98
Standard Deviation: 21.56
Coeff. of Variation: 12.87
Coeff. of Skewness: 0.53
Coeff. of Kurtosis: 1.13
Quantil(0.050): -29.9482
Quantil(0.100): -22.6855
Quantil(0.150): -19.6049
Quantil(0.200): -16.2705
Quantil(0.250): -12.6866
Quantil(0.300): -10.0883
Quantil(0.350): -6.7365
Quantil(0.400): -3.8180
Quantil(0.450): -0.6943
```

Figure 4. Exploratory Spatial Data Analysis options of the PyESSDA applicaton

On activating the PyESSDA, the application automatically fills for the second and third fields of the sub windows of the main window with default parameters as seen in Figure 3. These parameter values, along with the sample values and by means of the Equation 1, are used for assessment of the experimental and empirical semivariograms that are firstly presented in the report window showed in Figure 3(b) and in the graphic of Figure 3(c).

Information of the report window contains minimum, maximum, maximum distance and variance values evaluated from the sample set. Besides it presents a table that informs, for each lag number, its number of pairs, its mean distance and its semivariogram value.

Before working on creating unidirectional semivariograms, the user can select the button Semivariog Map, of Figure 3(a), to create an image representing experimental semivariogram values for different distance and directions. For this functionality the user must set the following parameters of the window of Figure 3(a): the angle tolerance value, using the slide labeled as (Direction) Tolerance and the maximum distance value, using the slide labeled as Range. The maximum distance value defines the x and y resolutions of the semivariogram map preset as 51 rows by 51 columns. The angular tolerance is considered for create an angular interval related to the angular direction defined by each point of the image location and the center of the map.

Figure 5 depicts the semivariogram map obtained for the Jacarei´s sample set using 677 as maximum distance and 30^o as the angle tolerance. In this figure the user can observe an anisotropy with greater continuity in the direction of ~0^o, considering 0^o

degrees in the north direction and angles increasing positively on a clockwise direction. For plotting the image of Figure 5 it was set the classic plotting style with the Python command *plt.style.use('classic')*. The image was plotted considering the number of lines and columns, nlins and ncols, and the distance tolerance increment, incrtol, as the parameters of the Python command *plt.imshow(semivar, extent=(xmin, xmax, ymin, ymax)), cmap=plt.cm.Blues)* where xmin=ncols*incrtol, xmax=ncols*incrtol, ymin=-nlins* incrtol, ymax=nlins*incrtol

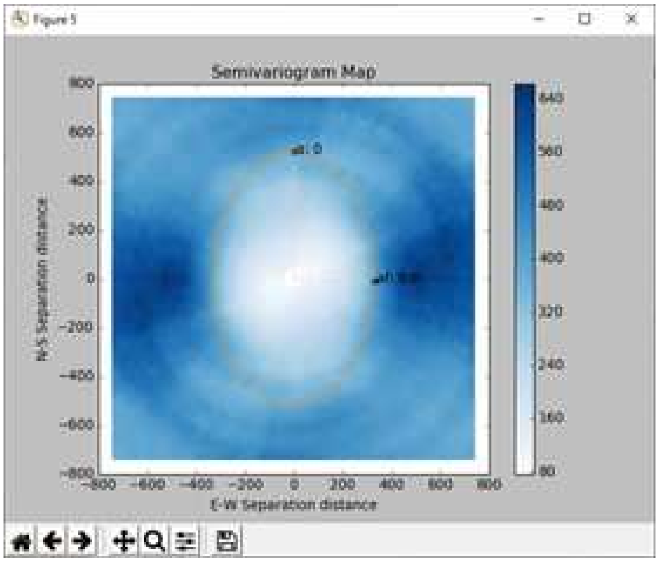
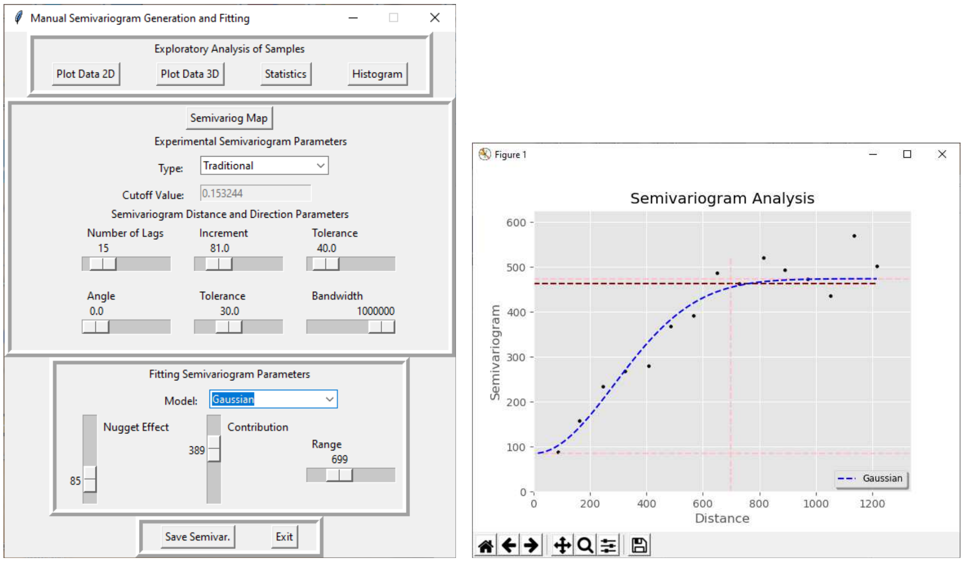


Figure 5. Semivariogram map using 677m as the maximum distance and 30° as angular tolerance.

The graphic of semivariogram of Figure 3(c) presents the experimental semivariogram in dark dots and semivariogram mathematical model as a curve in red dashed line. The parameters of the second and third sub windows of the main windows can be interactively changed by the user to obtain better semivariogram representations. The user can also set the parameters of the mathematical semivariogram selecting and moving, with mouse click and pan, the horizontal and vertical pink lines presented in the graphic of Figure 3(c). The implementation of this functionality is facilitated by use of mouse events provided by Tkinter package. It's possible to bind Python functions and methods to an event going on in a widget. When the event occurs, the "handler" function is called with an event object. For reference the graphic of semivariograms presents also the variance of the samples in a red horizontal dashed line.

Figure 6 shows the fitting of a Gaussian model, curve of blue dashed line in Figure 6(b), for the direction of greatest continuity, 0°, of the altimetry of the Jacarei's region. The parameters of the experimental and empirical semivariograms appear in Figure 6(a).

Figure 7 shows the fitting of a Gaussian model, curve of blue dashed line in Figure 7(b), for the direction of lowest continuity, 90°, of the altimetry of the Jacarei's region. The parameters of the experimental and conceptual semivariograms appear in Figure 7(a).



Manual Semivariogram Generation and Fitting

Exploratory Analysis of Samples

Experimental Semivariogram Parameters

Type: Traditional

Cutoff Value: 0.153244

Semivariogram Distance and Direction Parameters

Number of Lags 15 | Increment 81.0 | Tolerance 40.0

Angle 0.0 | Tolerance 30.0 | Bandwidth 1000000

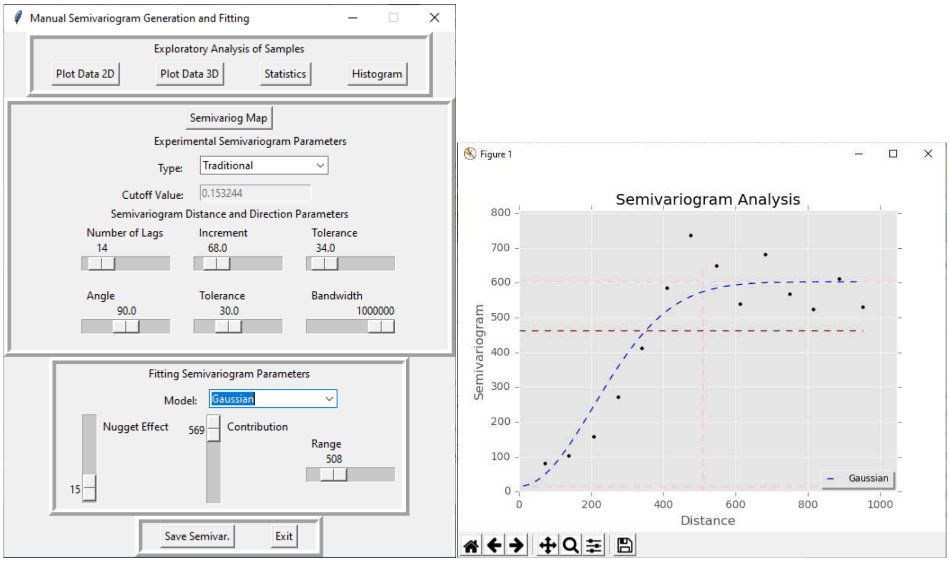
Fitting Semivariogram Parameters

Model: Gaussian

Nugget Effect 85 | Contribution 389 | Range 699

(a) (b)

Figure 6. Gaussian model, blue dashed line in window (b), fitted for the greatest spatial continuity, 0°, with the parameters defined in window (a)



Manual Semivariogram Generation and Fitting

Exploratory Analysis of Samples

Experimental Semivariogram Parameters

Type: Traditional

Cutoff Value: 0.153244

Semivariogram Distance and Direction Parameters

Number of Lags 14 | Increment 68.0 | Tolerance 34.0

Angle 90.0 | Tolerance 30.0 | Bandwidth 1000000

Fitting Semivariogram Parameters

Model: Gaussian

Nugget Effect 15 | Contribution 569 | Range 508

(a) (b)

Figure 7. Gaussian model, blue dashed line in window (b), fitted for the lowest spatial continuity, 90°, with the parameters defined in window (a)

To take into account the anisotropic behavior of the attribute altimetry of the presented case study, the two above empirical variogram models must be considered by the user for geostatistical prediction and simulation procedures. Table 1 presents a summary of the parameters of these empirical variograms. For spatial attributes with isotropic behavior the angular tolerance must be set to 90º in the main window of the application.

Table 1. Parameters of the empirical semivariograms for anisotropic modeling of the altimetry of the Jacareis' region

	Model	Nugget Effect	Contribution	Range
0º	Gaussian	85	389	699
90º	Gaussian	15	569	508

4.4. Saving Semivariogram and Exit the Application

Selecting the buttons of last sub window, in the bottom of the main window, the user can save the current semivariograms and, or, exit the application.

The semivariogram parameters are saved in a text file containing the information of the experimental and conceptual semivariograms. This is important as documentation, as information to further reproductions of the semivariogram modeling and as input for geostatistical procedures.

On pressing the exit button of the application, the user is warned about saving the semivariogram before exiting and after this all the opened application window is closed.

5. Conclusions

This article presented the functionalities and implementation details of an easy to use python application to perform exploratory and structural spatial dependence analysis of a set of sample points. The implementation shows that Python is a very suitable, scriptable program language to develop windows applications. The language offers a lot of basic packages for GUI implementation and widget generation for visualizing data in general, texts, graphics, images, etc. Although this article uses the Windows, Python scripts can be developed for different operational system environments Linux, Mac, Android.

For users that work on modelling attributes of geographical data, the presented spatial analysis tool, when compared with other free of charge options, enables users easily creating semivariograms that better represent the attribute variability mainly for short distances. Also, it is very important as basic investigation of spatial data to be used in multivariate spatial analysis on Geographical Information Systems (GIS) environments. The application is available for free use in the web location: http://www.dpi.inpe.br/spring/portugues/manuais.html.

The application implementation aims to enable users easily creating semivariograms that better represent the attribute variability mainly for short distances.

For the future, it is intended to include in the presented application functionalities related to geostatistical estimation and simulation procedures that make use of the modeled empirical semivariograms.

6. References

Anselin L, Syabri I, Kho Y (2006) GeoDa: An introduction to spatial data analysis. Geographical Analysis:38(1):5–22

Bailey T.C. and Gatrell A.C. (1995): *Interactive Spatial Data Analysis*, Longman, Essex.

Bettini, C. (2007). Conceitos básicos de geoestatística. In: MEIRELLES, M. S. P.; CÂMARA, G.; ALMEIDA, C. M. (Ed.). Geomática: modelos e aplicações ambientais. cap. 4. Brasília: Embrapa.

Burrough, P.A.; McDonell, R. (1998). Principles of Geographical Information Systems. Oxford, Oxford University Press.

Deutsch, C. e A. Journel (1992). GSLIB: Geostatistical Software Library and user's guide. New York, Oxford University Press.

Fischer, M. M. (2006). In: *Spatial Analysis and GeoComputation*. Springer, Chapter2, Berlin, Heidelberg

Goovaerts, P. (1997). "Geostatistics for Natural Resources Evaluation". Oxford Univ. Press, New-York, 483 p.

Haining, R. and Wise, S. (1997) Exploratory spatial data analysis. NCGIA core curriculum in GIScience. http://www.ncgia.ucsb.edu/giscc/units/u128/u128.html

Issaks, M. and Srivastava E. (1989). *An Introduction to Applied Geostatistics*. New York,

Oxford University Press.

Kuhlman, D. (2012) "A Python Book: Beginning Python, Advanced Python, and Python Exercises".

Rey, S. J. and Anselin, L. (2007) Pysal, a python library of spatial analytical methods. The Review of Regional Studies, vol. 37, n. 1, pp. 5-27

Robertson, G.P. 2008. GS+ : "Geostatistics for the Environmental Sciences", Gamma Design Software, Plainwell, Michigan USA. Pdf document available for free at: https://geostatistics.com/files/GSPlusUserGuide.pdf

Rossum, G. van. and Python development team. (2018) "Python Tutorial release 3.7.0". Python Software Foundation. Pdf document available for free at: https://bugs.python.org/file47781/Tutorial_EDIT.pdf

Symanzik, J. (2013) *Exploratory spatial data analysis*, handbook of regional science. Springer, pp. 1295-1310