3D LISA: A Flexible Program for Calculating the Local Moran's I in 1, 2 or 3D Illustrated with Examples from Soil Science

R. Kerry*[1], B. R. Ingram[2], P. Goovaerts[3], F. Meza[2] and D. Gimenez[4]

[1]Department of Geography, Brigham Young University, Provo, UT, USA
[2]Facultad de Ingeniería, Universidad de Talca, Curicó, Chile.
[3]BioMedware Inc., Ann Arbor, MI, USA.
[4]Department of Environmental Science, Rutgers University, New Brunswick, NJ, USA

*Email: ruth_kerry@byu.edu

Abstract

This paper presents an efficient and flexible program for computing in 1, 2 or 3D, the most common Local Indicator of Spatial Association (LISA) statistic, the Local Moran's I. The program exploits the parallel nature of the calculation of the LISA statistic and is implemented using efficient data structures for fast spatial indexing. The performance is illustrated with examples from Soil Science, a discipline where the prevalent 2D local Moran's I has only been rarely used. The application of this method to soil science resulted in the need to expand upon traditional implementations by developing it to perform in 3D. Also included in the program is the ability to compute a multi-variate LISA and to compute indicator variograms of the output to determine the average size of significant clusters. The program is illustrated with examples from Soil Science in 1, 2 and 3D.

Keywords: Local indicator of Spatial Association, Soil Science, Parallel processing, Spatial indexing algorithms

1. Introduction

Local Indicator of Spatial Association (LISA) statistics and particularly, the most commonly used of them, the Local Moran's I (Anselin, 1995), have been commonly computed for 2D datasets in Geography to identify clusters and spatial outliers in phenomena such as air pollution (Zhao et al. 2017, Lu et al 2017 and O'Leary et al. 2016), plant distributions (Chance et al. 2016), disease (Caswell 2016, Goovaerts and Jacquez, 2004 and 2005, Mahara et al. 2016, Tsai and Teng, 2016), socio-economic phenomena (Hirsch et al. 2016) and crime (Kerry et al. 2010). Such analysis is often used along with various regression methods to try to explain the causes of clusters/outliers (Mahara et al. 2016, Kerry et al. 2010). In contrast, LISA techniques have seldom been used in Soil Science or Geology even though they give different insights compared to geostatistical methods which are more commonly used in these disciplines. Indeed, a 20-year literature search of article titles from 1995-2015 only revealed one soil study that used the Local Moran's I (Zhang et al., 2008). While considering the promotion of the wider use of the Local Moran's I in Soil Science its expansion to 3D was prompted as well as its use in 1D to examine patterns of temporal autocorrelation.

Here we present an efficient and flexible program for computing the LISA statistic for 1, 2 or 3D datasets with data in various file formats. The utility of the program is illustrated with examples from Soil Science in 1, 2 and 3D. The 3D LISA program includes three major innovations which improve on the 2D LISA software that is currently available: 1) a multivariate LISA statistic was developed to test for clusters in several images (or variables) simultaneously, 2) the extension of the LISA statistic to 3D data and 3) indicator variograms of the LISA statistic output can be computed to give the average size of significant clusters. The implementation of the software is fast and uses memory efficiently. It was developed to exploit multi-core processor architectures to improve the calculation speed. The determination of neighbourhoods in the data is implemented using various user-selected data-structures that allow a trade-off between computational speed and memory usage allowing flexibility depending on the dataset size.

2. Methods

The local Moran's I, is computed for each point/pixel/geographic unit *X* as:

$$\text{LISA}(X) = \left[\frac{\hat{z}(X) - m}{s}\right]\left[\frac{1}{J}\sum_{i=1}^{J}\frac{\hat{z}(X_i) - m}{s}\right]$$

where *m* and *s* are the mean and standard deviation of the data $\hat{z}(X)$. In the LISA the standardized value at *X* (called the kernel value) is compared with the average of the *J* neighboring standardized values. For gridded data, the first or second order etc. queen or rook neighborhoods can be used. Positive values indicate positive spatial autocorrelation or clusters (HH-high values surrounded by high values or LL-low values surrounded by low values) whereas negative values indicate negative spatial autocorrelation and spatial outliers (HL-high values surrounded by low values and LH-low values surrounded by high values). Determining whether the LISA value is significantly different from zero requires knowledge of the distribution of values under the null hypothesis of spatial randomness, which is built using Monte Carlo simulation (Goovaerts and Jacquez, 2005). The False Discovery Rate (FDR) method (Castro and Singer, 2006) can then be used to correct for multiple testing and reduce the proportion of false positives.

2.1. 3D LISA Program Details

The program is written in C with a choice of file and data formats and allows data in 1, 2 or 3D to be analysed. The program can automatically detect the number of cores in the computer processor. As the LISA statistic calculation is essentially a parallel algorithm, the processor cores can be exploited to increase the calculation speed. The program also grants the user to select the number of threads to use during execution.

The determination of neighbourhoods for clustering of the data is performed using spatial indexing algorithms. Currently two such algorithms, Quad-/Oct- Trees and KD-Trees (Samet, 1990), are included in the program. More spatial indexing algorithms can be easily added in the future. The Quad-/Oct- Trees have a larger memory overhead than KD-Trees, but are faster for searching. For a baseline comparison, a sequential searching algorithm is included but its use is not recommended for large datasets as it is prohibitively slow. For example, when using the sequential algorithm to determine neighbourhoods the performance of the program scales quadratically with the size of the data compared to logarithmic scaling when using the Quadtree or KD-Tree search methods.A choice of methods for correcting for multiple testing is available as well as options for defining neighborhood size in terms of the number of nearest neighbours or 1st, 2nd order queen or rook neighbours for grid data. Indicator variograms of the LISA output can be computed to determine average cluster size**.**

General Program Outline:

1. Load data set into memory using pre-determined spatial indexing algorithm
2. Calculate neighbourhoods based on desired configuration
3. Calculate LISA statistic for each neighbourhood
4. Calculate P-values and Significance levels (1000 Monte Carlo simulations used by default)
5. Correct for multiple testing
6. Save calculations into file

3. Example Applications in Soil Science

3.1. 1D: Detection of Clusters in Soil Moisture Time Series Data

In time series analysis of soil volumetric water content (VWC) identification of periods with values above (HH) or below (LL) the average for a site/depth and also periods of transition between the two states (NS) is important. Over a five-year period, hourly measurements of VWC were made with five CS615 water content reflectometers installed at particular depths at Cream Ridge Fruit farm, NJ, USA. LITA (Local Indicator of Temporal Autocorrelation) analysis of this data is shown in Figure 1.

The soil surface (Figure 1, depth 1: 10 cm) shows periods of soil water depletion (LL) that generally coincide with the growing season and periods of replenishment (HH) when the vegetation is dormant. These features are transmitted downward (Figure 1, depths 2-5) while being filtered by the soil, shown by the reduced magnitude of the variations around the means (NS grey areas, Figure 1) with depth. Periods of HH and LL VWC occur at about the same time at different depths with the only exception being at depth 2: 46 cm, which coincided with a soil layer enriched in clay and thus a greater capacity for buffering VWC. The size of the HH period at 2.7 m enclosed by the circle seems unrelated to surface changes and could be caused by a rising groundwater table. The ability of the 3D LISA program to do multivariate LITA analysis is promising for future studies as links in temporal patterns of VWC with soil temperature and respiration patterns at the site could facilitate interpretations of the response of the system to climatic inputs of energy and water.

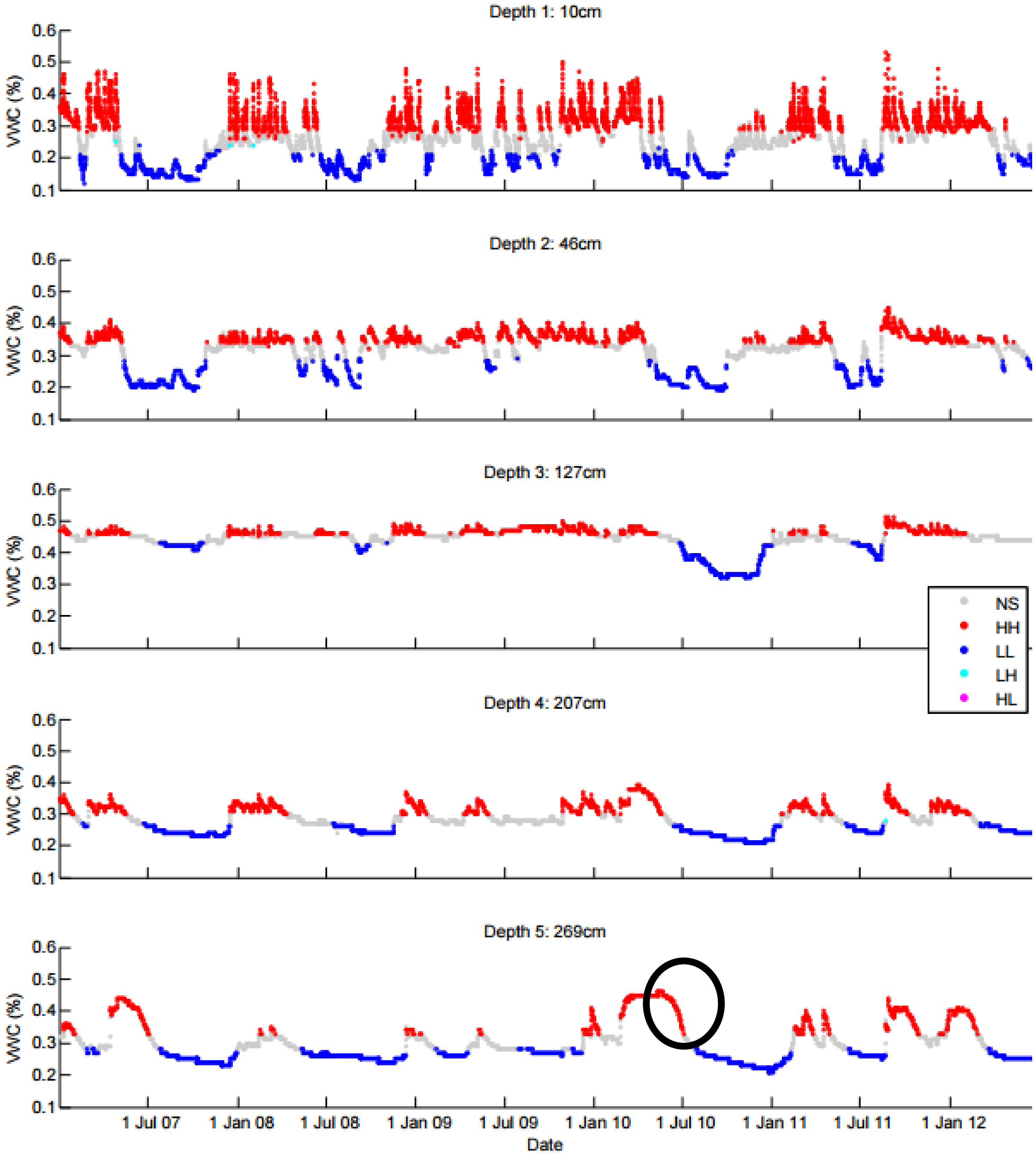

Figure 1: LITA analysis of soil VWC time series at Cream Ridge, NJ

3.2. 2D: Detection of Soil Pollution Hotspots

A possible 2D use of the multivariate LISA in soil science could involve detection of soil pollution hotspots (HH) and whether these coincide with clay or OM hotspots (HH) that could bind pollutants and make them biologically unavailable that would mean that the area needing excavating for remediation may be kept to a minimum. A soil pollution application is illustrated using the Swiss Jura data (Atteia et al. 1994). Figure 2 shows that LL clusters of Ni tend to occur on the Argovian rock type and the HH clusters on the Kimmeridgian. The HH clusters show the most continuous areas in need of potential expensive remediation.

Soil pollution hotspots with the largest values are often very isolated and small, LISA analysis could also potentially identify these points as spatial outliers (HL) meaning that the area to be remediated would be even smaller.

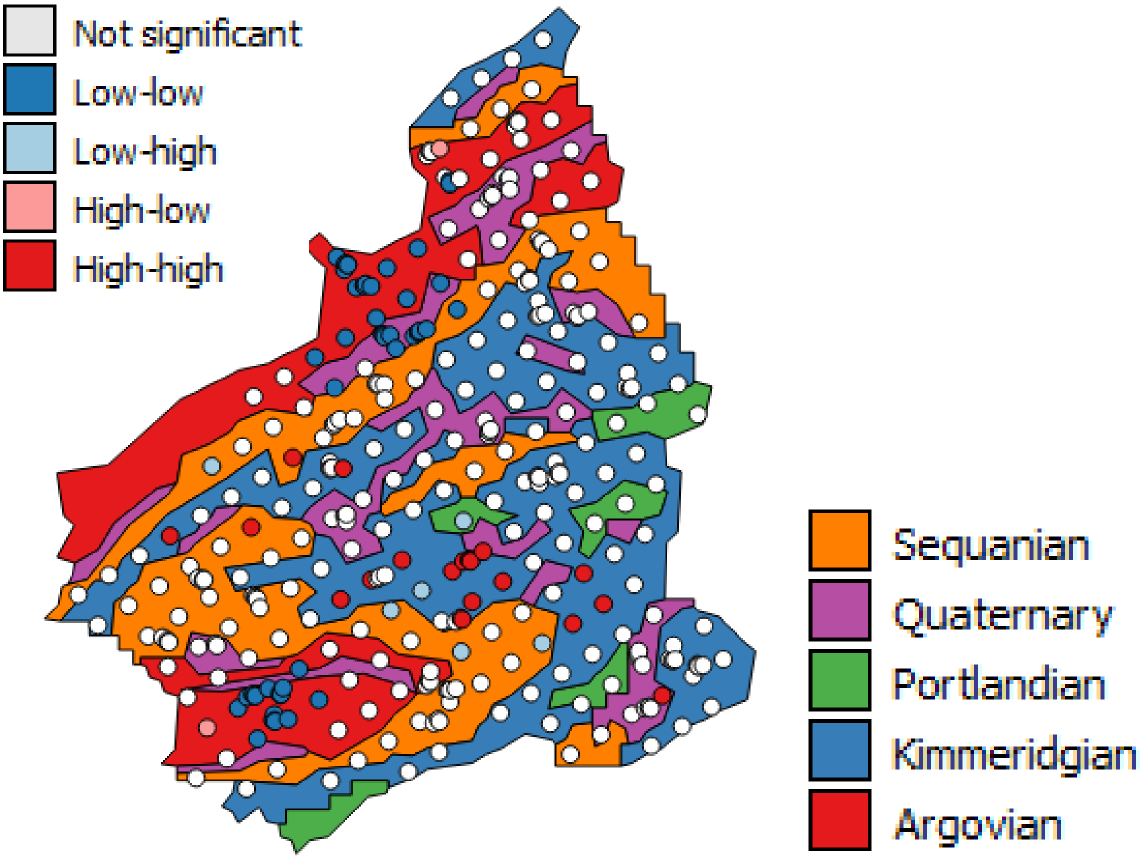

Figure 2: LISA map for soil Ni for the Swiss Jura in relation to rock type

3.4. 3D: Cluster Detection in Soil Cores

Soil morphology and micromorphology studies need to distinguish between soil structures and pore spaces from 3D images of soil cores such as CAT scans or electrical resistivity data. Usually this is done with some sort of image segmentation technique (Hapca et al. 2013, Houston et al. 2013). Here we suggest using the 3D LISA as an image segmentation technique for this purpose. 3D electrical resistivity (ER) data from a 50 cm depth and 30 cm diameter soil core. With 144 electrodes arranged in 8 planes resistivity data at 62208 points were generated (Figure 3a). The resistivity values of the core were used to analyze the distribution of water within the core at different flow rates. LISA analysis (Figure 3b) helped identify regions within the soil retaining relatively less (HH) or more (LL) water. Physical analysis of the core revealed heavier textured clay materials coinciding with the LL clusters. Computation of indicator variograms of the LISA output showed that the LL clusters were larger but showed less spatial structure than the HH clusters. The spatial location and shape of the LISA clusters can be used to inform 3D models of transport processes through the soil. Future developments of the algorithm should allow the ability to define different sized neighbourhoods in different directions. For example, in soil different processes act in the vertical and horizontal directions, so it should be possible to define the neighbourhood size differently in the x and y directions compared to the z. The efficient way in which the 3D LISA program has been designed allows the analysis of far larger data sets such as CAT scans of soil cores.

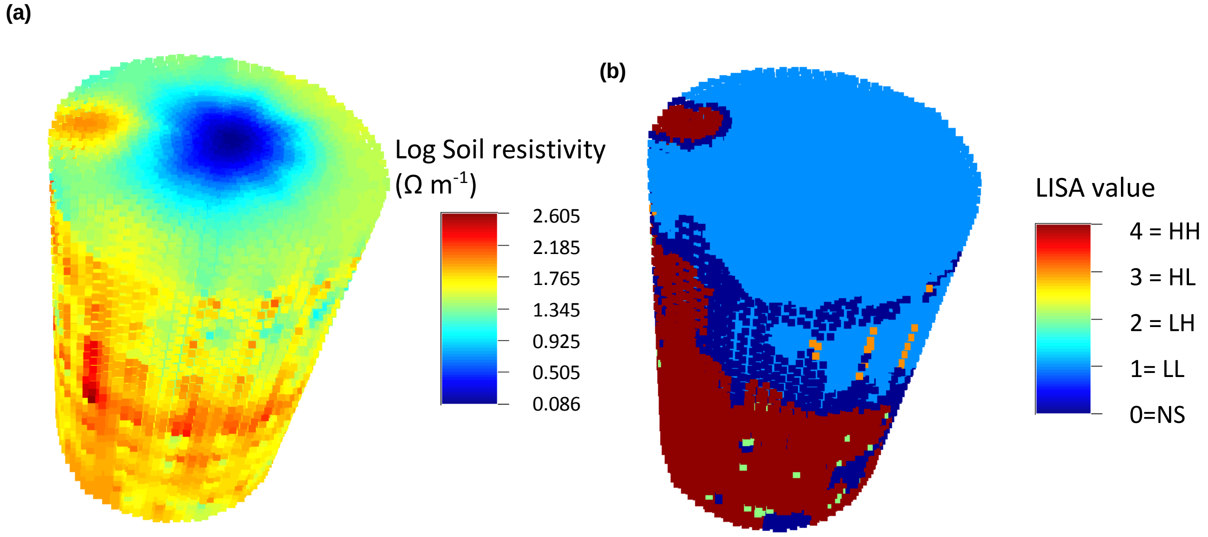

Figure 4: (a) Electrical resistivity data and (b) LISA results for a soil core

4. Conclusions

The local Moran's I is a useful tool for soil scientists and the 3D LISA program allows fast and flexible analysis with several data types in 1, 2 or 3D. The inclusion of multi-threaded algorithms enables the user to exploit multi-core processor architectures. Flexibility in the selection of spatial indexing algorithms gives the user the freedom to trade-off between speed and memory usage. The most efficient spatial indexing algorithms scale logarithmically rather than quadratically as in sequential searching so can be much faster. The program currently requires that all data fit into available computer memory but future versions will allow partial loading for large datasets such as CAT scans of soil cores and will include more spatial indexing algorithms. The software can be used on Windows, Linux or Mac systems. Future versions will also incorporate the ability to define anisotropic neighbourhoods, particularly for 3D data in soil studies where this would be beneficial.

5. Acknowledgements

M. Qin (Rutgers, NJ) collected Cream Ridge VWC data, A. Binley (Lancaster Uni.), L. Slater (Rutgers, NJ) helped with ER data collection

6. References

Anselin, L. 1995. Local Indicators of Spatial Association—LISA. Geographical Analysis, 27, 93–115.

Atteia O., Dubois J. P. and Webster R. 1994. Geostatistical analysis of soil contamination in the Swiss Jura. Environmental Pollution 86, 315–327.

Castro, M. C. and Singer, B. H. 2006. Controlling the False Discovery Rate: A New Application to Account for Multiple and Dependent Tests in Local Statistics of Spatial Association. Geographical Analysis 38, 180–208.

Caswell, J. M. 2016. Exploring spatial trends in Canadian incidence of hospitalization due to myocardial infarction with additional determinants of health. Public Health 140, 136-143.

Chance, C. M., Coops, N. C. and Plowright, A. A. 2016. Invasive Shrub Mapping in an Urban Environment from Hyperspectral and LiDAR-Derived Attributes. Frontiers in Plant Science 7, 1528.

Goovaerts, P. and Jacquez, G. M. 2004. Accounting for regional background and population size in the detection of spatial clusters and outliers using geostatistical filtering and spatial neutral models: the case of lung cancer in Long Island, New York. International Journal of Health Geographics, 3. 14.

Goovaerts, P. and Jacquez, G. M. 2005. Detection of Temporal Changes in the Spatial Distribution of Cancer Rates Using Local Moran’s I and Geostatistically Simulated Spatial Neutral Models. Journal of Geographical Systems 7, 137–59.

Hapca, S., Houston, A., Otten, W., and Baveye, Ph., 2013. New local segmentation method for 3D images of natural porous media, based on minimizing the intraclass grayscale variance, Vadose Zone Journal, 12 (3)

Hirsch, J. A., Grengs, J. and Schulz, A. 2016. How much are built environments changing, and where? Patterns of change by neighborhood sociodemographic characteristics across seven US metropolitan areas. Social Science & Medicine 169, 97-105.

Houston, A, Schmitt, S., Otten, W., Baveye, Ph. and Hapca, S. 2013. Effect of scanning and image reconstruction settings in X-ray computed tomography on soil image quality and segmentation performance, Geoderma, 207-208, 154-165.

Kerry, R., Goovaerts, P., Haining, R. P. and Ceccato, V. 2010. Geostatistical Analysis of Car Theft and Robbery in the Baltic States. Geographical Analysis. 42, 53-77.

Mahara, G., Wang, C. and Yang, K. 2016. The Association between Environmental Factors and Scarlet Fever Incidence in Beijing Region: Using GIS and Spatial Regression Models. International Journal of Environmental Research and Public Health. 13, 1083.

O'Leary, B, Reiners, J. J. and Xu, X. 2016. Identification and influence of spatio-temporal outliers in urban air quality measurements. Science of the Total Environment 573, 55-65.

Samet, H. 1990. *The Design and Analysis of Spatial Data Structures*, Addison-Wesley, MA.

Tsai, P. and Teng, H. 2016. Role of *Aedes aegypti* (Linnaeus) and *Aedes albopictus* (Skuse) in local dengue epidemics in Taiwan. BMC Infectious Diseases 16, 662.

Zhang, C., Luo, L. and Xu, W., 2008. Use of local Moran's I and GIS to identify pollution hotspots of Pb in urban soils of Galway, Ireland. Science of the Total Environment, 398, 212-221.

Zhao, X., Deng, C. and Huang, X. 2017. Driving forces and the spatial patterns of industrial sulfur dioxide discharge in China. Science of the Total Environment 577, 279-288.