Vision Paper: Efficient Spatio-Temporal Geo-Statistics for Exploratory Data Analysis

Julian Bruns[1] **and Joan R. Access**[1]

1 **GIScience Research Group, Institute of Geography, Heidelberg University, Heidelberg, Germany**
firstname.lastname@uni-heidelberg.de

Abstract

Hot spot analysis is essential for geo-statistics.

1998 ACM Subject Classification Dummy classification – please refer to http://www.acm.org/about/class/ccs98-html

Keywords and phrases Dummy keyword – please provide 1–5 keywords

Digital Object Identifier 10.4230/LIPIcs.CVIT.2016.23

1 Introduction and Related Work

The goal of hotspot analysis is the detection and identification of interesting areas. It achieves this goal by computing statistically significant deviations from the mean value of a given study area. This allows a decision maker to easily identify those areas of interest and allows further focus in sub-sequential data analysis or the decision focus. Typical applications range from crime detection over identification of disease outbreaks to urban heat islands. In such applications, scarce resources are then often applied in only those identified hotspots or used as the basis for the allocation. The general approach is an unsupervised learning method similar to a cluster analysis.

1.1 Examples

Spatio-temporal sentiment hotspot detection using geotagged photos [25]

A framework for evacuation hotspot detection after large scale disasters using location data from smartphones: case study of Kumamoto earthquake [23]

[22] BigGIS Vision Paper

[7] The era of big spatial data

[14] Spatio-Temporal Hotspot Computation on Apache Spark (GIS Cup)

Giscup 2016 [1]

2 State of the Art

2.1 Geostatistic

While the standard hot spot analysis approaches allow for a fast and automated exploratory analysis of spatial data sets, they are highly dependent on their parametrizations and underlying data. This is quite similar to the challenges in clustering, in particular for

[1] http://sigspatial2016.sigspatial.org/giscup2016/home

the well-known k-means, first coined in [13], or DBScan ([8]) algorithm. This leads to an instability of the analysis results, which are then difficult to use with high certainty. Here, a selection of approaches and discussions of the last 15 years are presented how to tackle this instability.

The most general approach is to pre-determine and calculate spatial dependencies, based on the a-priori knowledge of the analyst, the empirical data set or best-practices from previous, similar analyses. A good example for this approach can be found in [18]. In their work, the authors examine the effects of scale on temperature and in particular urban heat island modeling. The pre-determined range of spatial influences is called *buffer zones*; these buffer zones indicate the range of impact in an inner-city temperature measurement scenario. They found that for their empirical data set a buffer zone of 1000m provides the best results. This is familiar to the well-known approach to solve the inherent problem of DBScan, the determination of its distance: OPTICS ([2]). The disadvantage of this approach is that it has to be manually pre-determined, which results in similar problems as the semi-supervised methods to detect points of interest. It can not be done automatically without introducing an element of uncertainty, which is opposed to our goal to create a more stable approach.

An automated approach is presented with the *A Multidirectional Optimal Ecotope-Based Algorithm* (AMOEBA) in [1]. The idea behind this approach is to automatically create the optimal, scale-invariant weight matrix and then use this weight matrix in conjunction with a clustering approach to create a graphical overview map of areas of interest. The term ecotope is used for this areas, which is the technical term from the field of biology for the habitat of species. The result is a consistent identification of spatial clusters on a map. In their work they use the G^* statistic as the underlying statistic. The clustering approach is quite similar to DBScan in its approach of creating ecotopes.

A true modification of the G^* statistic is presented in a later work ([10]) of the same authors called the LSM (local statistics model). They base their modification on the Kriging approach and its ability to model the spatial autocorrelation as a function dependent on the distance. The idea is to model the weight matrix W as a function of the spatial autocorrelation, where each entry of the matrix is a value derived from the empirical (semi-)variogram. This leads to continuous values up to the so called *critical distance*, which is "defined as the distance beyond which no discernible increase in clustering exist" ([10]). They compare their configuration to other, well-known spatial configuration approaches for the weight matrix W. These are taken from [11] and in the words of [10]: "Research on W has been reviewed by Griffith (1996, p. 80), who concludes that five rules of thumb aid in the specification of weights matrices:

1. "It is better to posit some reasonable geographic weights matrix than to assume independence." This implies that one should search for or theorize about an appropriate W and that better results are obtained when distance is taken into account.
2. "It is best to use surface partitioning that falls somewhere between a regular square and a regular hexagonal tessellation." Griffith suggests that for planar data, a specification between four and six neighbors is better than something either above six or below four. Of course, the configuration of the planar tessellations will play a role here ([4]).
3. "A relatively large number of spatial units should be employed, n > 60." Following from the law of large numbers, most spatial research, especially due to unequal size spatial units, would require fairly large samples.
4. "Low-order spatial models should be given preference over higher-order ones." Following from the scientific principle of parsimony, it is always wise to choose less complicated models when the opportunity presents itself.

5. "In general, it is better to apply a somewhat under-specified (fewer neighbors) rather than an over-specified (extra neighbors) weights matrix." [9] found this result by identifying the power of tests. Overspecification reduces power. They recognize that "Uncertainty with respect to proper specification has long been recognized as a fundamental problem in applied spatial econometric modeling" (p. 132).

"

[17] discuss the question in how to formulate the G^* statistic to focus more on local pattern, while still accounting for the global autocorrelation. They propose the O statistic which uses the (semi-)variogram to subdivide the data set into several " 'relatively homogeneous' subregions" ([17]). This allows the identification of smaller, more local hot spots, which can be overshadowed in bigger data sets. Finally, they restrict the general applicability in that the version presented in their work requires spatial stationarity.

[20] present a scale-sensitive version of the local G^* statistic, which they call the *GS statistic*. The motivation for this modification is to account for the differences in the scale (the impact of the area under investigation) of the data set, i.e. whether a data set includes only the inner city or also its surrounding area. The problem lies in the detection and use of the local context of the gathered data. A fixed weight matrix W does not include the difference in context, e.g. in Twitter feeds. Their approach is to redefine the neighborhood of a data point with upper and lower distance thresholds, which are then used in pairwise comparisons. Only sufficiently connected data points within their thresholds are considered to be viable as a hot spot and only those points are used for the global mean and global deviation values. They evaluated their approach on Twitter data of the city of San Francisco, USA and show that this leads to reduction or even negation of cross-scale interference. The main restrictions of this approach lies in its increased computational costs as well as its reliance on continuous distance functions.

For futher literature regarding the creation of optimal weight matrices for spatial associations we refer to the work of [1], where they provide an exhaustive overview of the state of the art.

[5] for focal G^* and efficient parametrization of G^*

2.2 Frameworks and Methods

An Effective High-Performance Multiway Spatial Join Algorithm with Spark [6]

GeoSpark: a cluster computing framework for processing large-scale spatial data [24]

Spatio-Temporal Join on Apache Spark [21]

Algorithmically-Guided User Interaction [19]

GeoMesa: a distributed architecture for spatio-temporal fusion [12]

geotrellis[2]

[2] https://geotrellis.io/

3 Vision

4 Challenges

5 Conclusion

5.1 Hot Spot Analysis

The goal of hotspot analysis is the detection of interesting areas as well as patterns in spatial information. One of the most fundamental approach is Moran´s I [15]. There it is tested whether or not a spatial dependency exists. This gives the information on global dependencies in a data set. Upon this hypothesis test several geo-statistical tests are based. The most well known are the Getis-Ord statistic [16] and LISA [3]. In both cases the general, the global statistic of Moran´s I is applied in a local context. The goal is to detect not only global values, but instead to focus on local hotspots and to measure the significance of those local areas.

Existing methods for determining hot spots are dependent on the parametrization of the weight matrix as well as on the size of the study area. Intuitively, increasing the size of a weight matrix has a "blurring" effect on the raster (

6 Evaluation

6.1 Dataset

We evaluate our data on the New York city yellow cab data set [3]. This data set includes all taxi drives from the yellow cabs in New York City, from location, to passengers and many more informations. In this study, we compare the total pickups over January from 2016 in the Manhattan area. The borders of the rasters are (40.699607 °(N), -74.020265 °(E)) and (40.769239 °(N), -73.948286 °(E)) after WGS84. By using this data set, we reduce the computational effort while still being able to show the applicability on real world data and problems.

6.2 Treatments

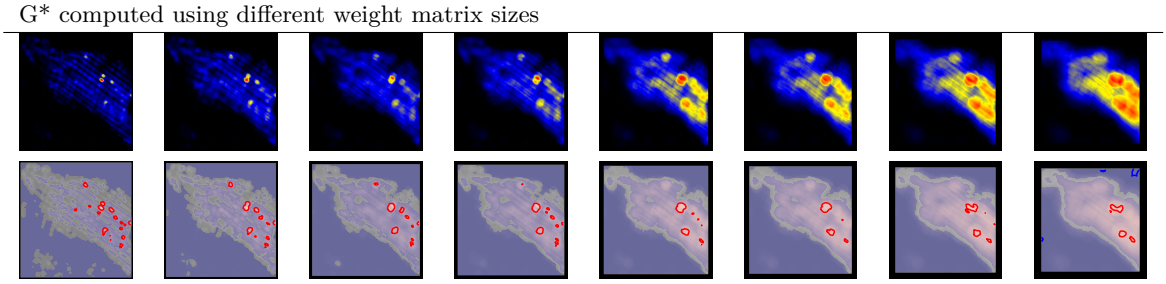

Figure 1 This image shows different weight matrix sizes for G* and Focal G* together with two metrics – $SoH^{\uparrow}$ and $SoH^{\downarrow}$.

[3] *http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml*

References

1 Jared Aldstadt and Arthur Getis. Using amoeba to create a spatial weights matrix and identify spatial clusters. *Geographical Analysis*, 38(4):327–343, 2006.

2 Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander. Optics: ordering points to identify the clustering structure. In *ACM Sigmod record*, volume 28, pages 49–60. ACM, 1999.

3 Luc Anselin. Local indicators of spatial association - lisa. *Geographical Analysis*, 27(2):93–115, 1995. doi:10.1111/j.1538-4632.1995.tb00338.x.

4 Barry Boots and Michael Tiefelsdorf. Global and local spatial autocorrelation in bounded regular tessellations. *Journal of Geographical Systems*, 2(4):319–348, 2000.

5 Julian Bruns and Viliam Simko. Stable hotspot analysis for intra-urban heat islands. *GI_Forum*, 1:79–92, 2017. URL: https://doi.org/10.1553/giscience2017_01_s79, doi:10.1553/giscience2017_01_s79.

6 Zhenhong Du, Xianwei Zhao, Xinyue Ye, Jingwei Zhou, Feng Zhang, and Renyi Liu. An effective high-performance multiway spatial join algorithm with spark. *ISPRS International Journal of Geo-Information*, 6(4), 2017. URL: http://www.mdpi.com/2220-9964/6/4/96, doi:10.3390/ijgi6040096.

7 A. Eldawy and M. F. Mokbel. The era of big spatial data. In *2015 31st IEEE International Conference on Data Engineering Workshops*, pages 42–49, April 2015. doi:10.1109/ICDEW.2015.7129542.

8 Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

9 Raymond JGM Florax and Serge Rey. The impacts of misspecified spatial interaction in linear regression models. In *New directions in spatial econometrics*, pages 111–135. Springer, 1995.

10 Arthur Getis and Jared Aldstadt. Constructing the spatial weights matrix using a local statistic. In *Perspectives on spatial data analysis*, pages 147–163. Springer, 2010.

11 Daniel A Griffith. Some guideline for specifying the geographic weights matrix contained in spatial statistical models. *Practical handbook of spatial statistics*, pages 65–82, 1996.

12 James N Hughes, Andrew Annex, Christopher N Eichelberger, Anthony Fox, Andrew Hulbert, and Michael Ronquest. Geomesa: a distributed architecture for spatio-temporal fusion. In *Geospatial Informatics, Fusion, and Motion Video Analytics V*, volume 9473, page 94730F. International Society for Optics and Photonics, 2015.

13 James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

14 Paras Mehta, Christian Windolf, and Agnès Voisard. Spatio-temporal hotspot computation on apache spark (gis cup). In *24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2016.

15 Patrick AP Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23, 1950.

16 J. K. Ord and Arthur Getis. Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis*, 27, 1995. doi:10.1111/j.1538-4632.1995.tb00912.x.

17 J Keith Ord and Arthur Getis. Testing for local spatial autocorrelation in the presence of global autocorrelation. *Journal of Regional Science*, 41(3):411–432, 2001.

18 Juuso Suomi, Jan Hjort, and Jukka Käyhkö. Effects of scale on modelling the urban heat island in turku, sw finland. *Climate Research*, 55(2):105–118, 2012.

19 Thomas C. van Dijk and Alexander Wolff. Algorithmically-guided user interaction. In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL'17, pages 11:1–11:4, New York, NY, USA, 2017. ACM. URL: http://doi.acm.org/10.1145/3139958.3140032, doi:10.1145/3139958.3140032.

20 Rene Westerholt, Bernd Resch, and Alexander Zipf. A local scale-sensitive indicator of spatial autocorrelation for assessing high-and low-value clusters in multiscale datasets. *International Journal of Geographical Information Science*, 29(5):868–887, 2015.

21 Randall T. Whitman, Michael B. Park, Bryan G. Marsh, and Erik G. Hoel. Spatio-temporal join on apache spark. In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL'17, pages 20:1–20:10, New York, NY, USA, 2017. ACM. URL: http://doi.acm.org/10.1145/3139958.3139963, doi:10.1145/3139958.3139963.

22 Patrick Wiener, Manuel Stein, Daniel Seebacher, Julian Bruns, Matthias Frank, Viliam Simko, Stefan Zander, and Jens Nimis. Biggis: A continuous refinement approach to master heterogeneity and uncertainty in spatio-temporal big data (vision paper). In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '16, pages 8:1–8:4, New York, NY, USA, 2016. ACM. URL: http://doi.acm.org/10.1145/2996913.2996931, doi:10.1145/2996913.2996931.

23 Takahiro Yabe, Kota Tsubouchi, Akihito Sudo, and Yoshihide Sekimoto. A framework for evacuation hotspot detection after large scale disasters using location data from smartphones: Case study of kumamoto earthquake. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '16, pages 44:1–44:10, New York, NY, USA, 2016. ACM. URL: http://doi.acm.org/10.1145/2996913.2997014, doi:10.1145/2996913.2997014.

24 Jia Yu, Jinxuan Wu, and Mohamed Sarwat. Geospark: A cluster computing framework for processing large-scale spatial data. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '15, pages 70:1–70:4, New York, NY, USA, 2015. ACM. URL: http://doi.acm.org/10.1145/2820783.2820860, doi:10.1145/2820783.2820860.

25 Yi Zhu and Shawn Newsam. Spatio-temporal sentiment hotspot detection using geotagged photos. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '16, pages 76:1–76:4, New York, NY, USA, 2016. ACM. URL: http://doi.acm.org/10.1145/2996913.2996978, doi:10.1145/2996913.2996978.