Stable Hot Spot Analysis (Draft)

Marc Gassenschmidt, Viliam Simko, and Julian Bruns
todo@fzi.de, simko@fzi.de, bruns@fzi.de

FZI Forschungszentrum Informatik
am Karlsruher Institut für Technologie
76131, Haid-und-Neu-Str. 10-14
Karlsruhe, Germany

Abstract: Hot spot analysis is essential for geo-statistics. It supports decision making by detecting points as well as areas of interest in comparison to their neighbourhood. However, these methods are dependent on different parameters, ranging from the resolution of the study area to the size of their neighbourhood. This dependence can lead to instabilities of the detected hotspots, where the results can highly vary between different parameters. A decision maker can therefore ask how valid the analysis actually is. In this study, we examine the impact of key parameters on the stability of the hotspots, namely the size of the neighbourhood, the resolution and the size of the study area, as well as the influence of the ratio between those parameters. We compute the hotspots with the well known Getis-Ord (G^*) statistic as well as its modification, the *Focal* G^* statistic. We measure the stability of the hotspot analysis using a recently introduced *stability of hotspots* metric (SoH) and compare the results to intuitive visual analysis. We evaluate the results on real world data with the well-known yellow cab taxi data set from New York, Manhattan. Our results indicate a negative impact on the stability with an increase of the size of the neighbourhood as well as a reduction of the size of the study area, regardless of the resolution.

1 Introduction

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.

But, similar to a cluster analysis, there does exist a high dependency of the identified hotspots on the detection method and in particular the parametrization of this method. The identified areas as well as their shape can vary highly. This volatility can lead to a decrease in trust in the result or in suboptimal allocations of scarce resources. Therefore it is necessary to measure and evaluate the stability of a hotspot analysis as well as the different parametrizations. In our initial work [3], we introduced a method to measure the stability of hotspots, the *stability of hotspots* metric (SoH) and showed its use on the basis of temperature data . Here, we build upon that work and examine in more detail the impact of the different instantiations of the most typical parameter. We use the well known Getis-Ord statistic [7], the standard G^*, and a modification of this statistic, the focal G^* [3]. Those parameter are the size of the study area (focal matrix), the detail of the resolution (zoom) and the size of the neighbourhood (weight matrix). By varying over these parameter, we can compare the stability for all possible combinations and isolate the effect of single parameter by aggregation over the other parameter. We evaluate on the well-known taxi data set to show the applicability on real world use cases as well as to enable a simple replication.

2 Related Work

2.1 Quality of Clustering

The problem of assessing the quality in unsupervised learning is well known. In the case of the k-mean algorithm, the quality of the clustering is mostly dependent on the value of the k and a miss-specification can lead to highly irregular clusters. In a simple 2D clustering, they can be easily recognized by visual analysis, but in higher dimensionality, this is impossible. One method, to measure the quality of such a clustering is the compactness of the clusters, see e.g. [9]. This enables the comparison between different clusters. Another possibility is the Silhouette Coefficient by Kaufman et Rousseeuw 1990. This metric measures the similarity of objects in a cluster in comparison to other clusters. For density based clustering, e.g. for DBSCAN [4], OPTICS [1] gives a simple method to tune the essential parameter for this clustering. This is only a small overview of methods to influence and measure the quality of different clustering methods. But it shows that this problem is not easily solved and dependent on the chosen algorithm. To our knowledge, there does not exist a method to overall measure the stability of a clustering.

2.2 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 [6]. 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 [7] and LISA [2]. 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. A more in depth overview of methods to identify and visualize spatial patterns and areas of interest can be found in [8].

3 Stability of Hot Spot Analysis

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 (Fig. **??**a) whereas decreasing the size can be seen as a form of "sharpening" (Fig. **??**b).

For a data analyst, when exploring the data interactively by choosing different filter sizes (weight matrices) or point aggregation strategies (pixel sizes), it is important that the position and size of a hot spot changes in a predictable manner. We formalize the intuition in our stability metric.

We define a hot spot found in comparably more coarse resolutions as *parent* (larger weight matrix or larger pixel size) and in finer resolutions as *child* (smaller weight matrix or smaller pixel size).

That is

To be stable, one assumes that every parent has at least one child and that each child has one parent. For a perfectly stable interaction, it can be easily seen that the connection between parent and child is a injective function and between child and parent a surjective function. To measure the closeness of connection, we propose a metric called the *Stability of Hot spot* (SoH). It measures the deviation from a perfectly stable transformation of resolutions.

In its downward property (from parent to child, injective) it is defined as:

$$SoH^{\downarrow} = \frac{ParentsWithChildNodes}{Parents} = \frac{|Parents \cap Children|}{|Parents|} \tag{1}$$

And for its upward property (from child to parent, surjective):

$$SoH^{\uparrow} = \frac{ChildrenWithParent}{Children} = 1 - \frac{|Children - Parents|}{|Children|} \tag{2}$$

where $ParentsWithChildNodes$ is the number of parents that have at least one *child*, $Parents$ is the total number of *parent*, $ChildrenWithParent$ is the number of children and $Children$ as the total number of children. The SoH is defined for a range between 0 and 1, where 1 represents a perfectly stable transformation while 0 would be a transformation with no stability at all.

4 Focal Getis-Ord

4.1 Dataset

Our results are based on the cab taxi data set from New York, Manhattan [**?**]. We used only the dataset of January 2016. From that dataset we used the *pickup_longitude* and *pickup_latitude* column. The borders of the raster are between (40.699607, -74.020265) and (40.769239, -73.948286). Zoom size 1 means we aggregate every point, which was in range of 100x100*0.000001 into one pixel. In zoom size N we aggregated NxN pixel from zoom size 1 into a new pixel. For our measurements we specified the following data series:

- The weight matrix W is from 3 to 43 with a stepsize of 4.
- The focal matrix F is from 17 to 125 with a stepsize of 12.
- The zoom level is from 1 to 5 with a stepsize of 1.

Therefore we calculate around 10*10*5=500 results for each fixed x. The following conditions must hold F>W, F<Raster.cols und F<Raster.rows.

4.2 Method

In the following text, we use the notation $R \overset{\texttt{op}}{\circ} M$ to denote a focal operation op applied on a raster R with a focal window determined by a matrix M. This is rougly eqivalent to a command `focal(x=R, w=M, fun=op)` from package *raster* in the R programming language [5].

Definition 1 (G^* function on rasters)**.** *The function G^* can be expressed as a raster operation:*

$$G^*(R,W,st) = \frac{R \overset{\texttt{sum}}{\circ} W - M * \sum_{w \in W} w}{S\sqrt{\frac{N*\sum_{w \in W} w^2 - (\sum_{w \in W} w)^2}{N-1}}}$$

where:

- *R is the input raster.*
- *W is a weight matrix of values between 0 and 1.*
- *$st = (N,M,S)$ is a parametrization specific to a particular version of the G^* function. (Def. 2 and 3).*

Definition 2 (Standard G^* parametrization)**.** *Computes the parametrization st as global statistics for all pixels in the raster R:*

- *N represents the number of all pixels in R.*
- *M represents the global mean of R.*
- *S represents the global standard deviation of all pixels in R.*

Definition 3 (Focal G^* parametrization)**.** *Let F be a boolean matrix such that: $all(dim(F) \geq dim(W))$. This version uses focal operations to compute per-pixel statistics given by the focal neighbourhood F as follows:*

- *N is a raster computed as a focal operation $R \overset{\texttt{sum}}{\circ} F$. Each pixel represents the number of pixels from R convoluted with the matrix F.*
- *M is a raster computed as a focal mean $R \overset{\texttt{mean}}{\circ} F$, thus each pixel represents a mean value of its F-neighbourhood.*
- *S is a raster computed as a focal standard deviation $R \overset{\texttt{sd}}{\circ} F$, thus each pixel represents a standard deviation of its F-neighbourhood.*

5 Results

6 Evaluation

- To evaluate, we compare G* with FocalG* on the same dataset. - We use NY taxi dropoffs - a single evaluation run is defined in Def. 4. - In an ideal case, we could produce a 3D-plot: - x-axis would be growing pixel sizes 100..1000 by 100 (aka zoom-out) - y-axis would be growing weight matrix sizes (e.g. 5..41 by 2) - z-axis would be performance of G* vs FocalG* x and y coordinates - we could also repeat the computation for multiple datasets to obtain multiple samples and to compute error bars. - due to paper size limit, we choose projection in which matrix size is set to constant 7x7 pixels and the only variable is the pixel size. This way, we produce a 2D plot. We also generate just a single sample.

Definition 4 (Evaluation Run)**.** *We define a single evaluation run as a tuple:*

$$E = (V, m, p, w)$$

where:

- *V is the input dataset of points, representing the taxi dropoffs in our case.*
- *m is the metric used, in our case either $SoH^{\uparrow}$ or $SoH^{\downarrow}$.*
- *p represents the pixel size for aggregating points from V, e.g. 100×100 meters.*
- *w represents the size of a weight matrix. In our case, we chose a weight matrix depicted in Figure **??**(a) for both the G* and FocalG* cases.*

6.1 Clumping

6.2 Zoom

Blueline is Focal G^*. Redline is G^*

6.3 Blur

The evaluation results are plotted in Fig. **??**, each point in the graph represents the $SoH^{\uparrow}$ metric (Eq.2) between two G^* generated using weight matrices of size i and $i+2$. The focal matrix F has a fixed size of 41×41

$$SoH^{\uparrow}(G^*(R, W_i, st), G^*(R, W_{i+2}, st))$$

6.4 Results and Discussion

In figure 3a you can see that an increase of the focal size leads to better results for $SoH^{\uparrow}$. The standard deviation decreases also with a higher focal size. Only at focal size F of 17x17 there is no standard deviation. At focal size F 17x17 the standard deviation is missing, because the weight size W must be smaller than F. When F and W have similar size, there are no clusters. (TODO assumtion for SOH, if no clusters we asume Up/Down = 0) $SoH^{\downarrow}$ in figure 3b reaches its maximum for focal size F 65x65 in the regarded area.

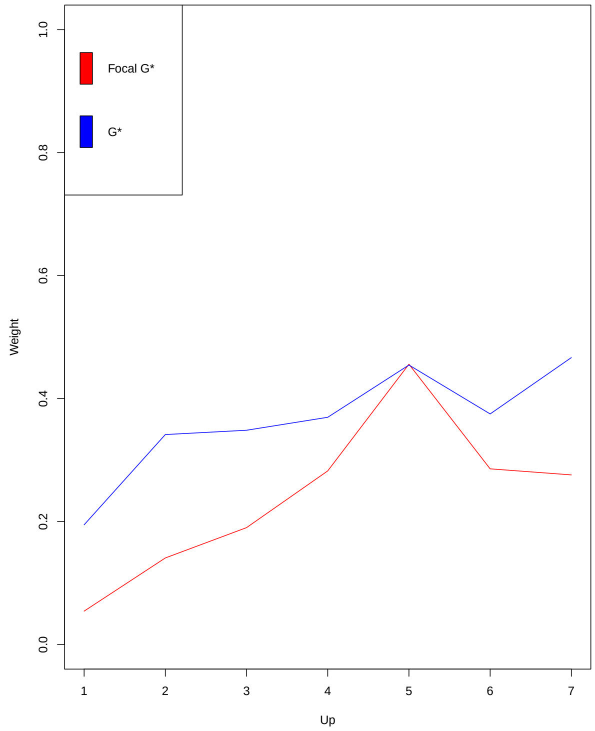
In figure 4b one can see that an increase of the weight size leads to better results. Focal G^* is not allways better than G^* which would be a perfect result we hoped for. But it seems that $G*$ has at 23 and 39 some parametrisation which leads to small break downs. The standard deviation is in this area higher than in the other points. In figure 4a the break downs are also visible. The high standard deviation indicates that it is possible to use this parametrisation for better results The $SoH^{\uparrow}$ in 4a is mostly better than G^* which was our conclusion in **??**.

The last variation we examined was the zoom level. The results can seen in figure 5. One can see that G^* is always better than Focal G^*. The target area of Focal G^* increases with every zoom step.

7 Conclusions and Future Work

In this work, we compared the Getis-Ord statistic with Focal Getis-Ord statisitic on a different dataset. We examined further parameters and evaluated that G^* and Focal G^* performe better if the weight is higher. Our results indicate a high Focal range for the upward property and a maximum of 65x65 for the downward property. Therefore F should be not less than 65x65. If we calculate the SoH for the zoom level, Getis-Ord leads to better results. In general the zoom level has the least influence if G^* or Focal G^* performe better. We assume that if Focal G^* has a different Focal size for a different zoom level, it performes better, but this is beyond the scope of this work.

(a) Upward



(b) Downward

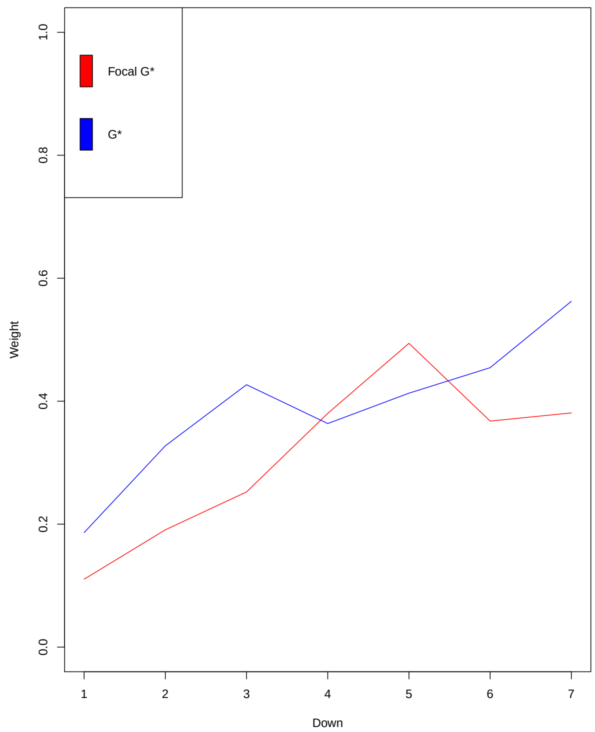
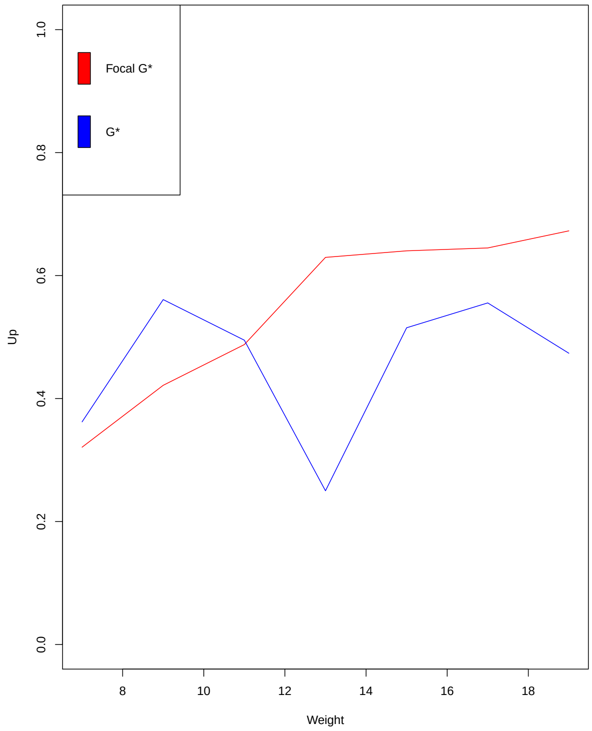


Figure 1: TODO

(a) Upward



(b) Downward

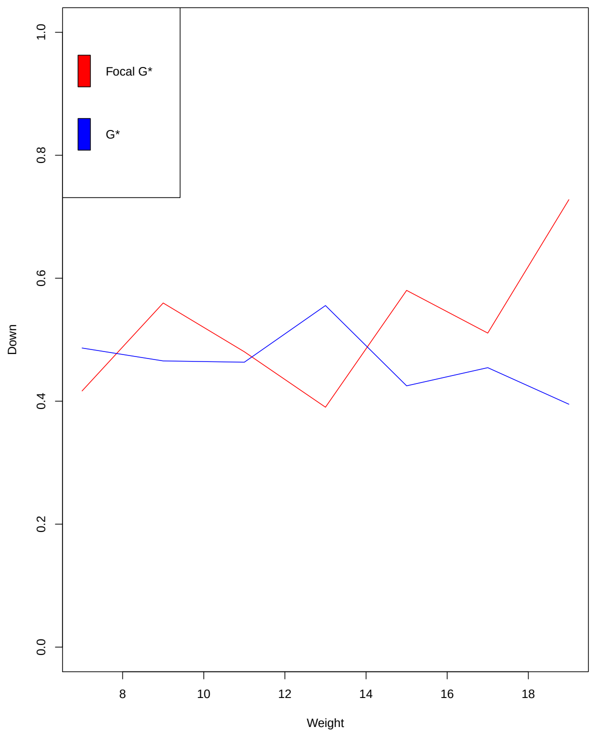
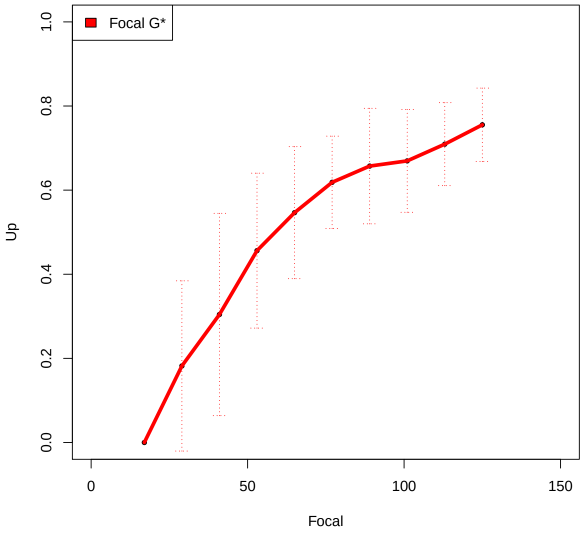


Figure 2: TODO

(a) SoH up for focal size



(b) SoH down for focal size

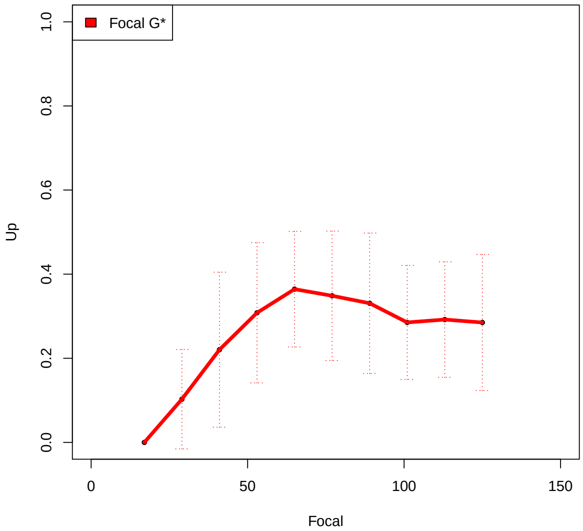
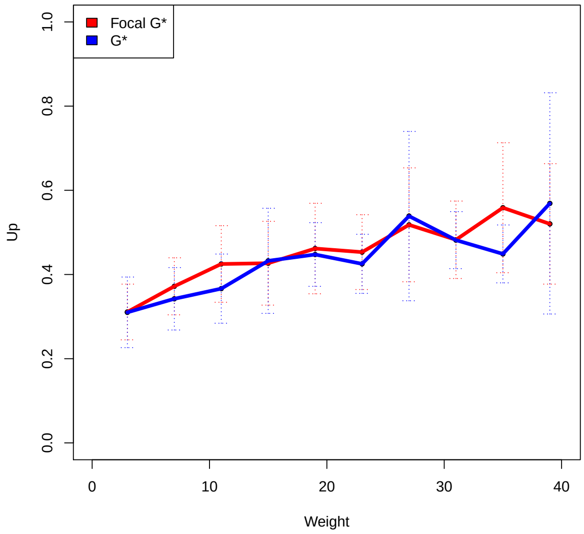


Figure 3: SoH for focal size

(a) SoH up for weight size



(b) SoH down for weight size

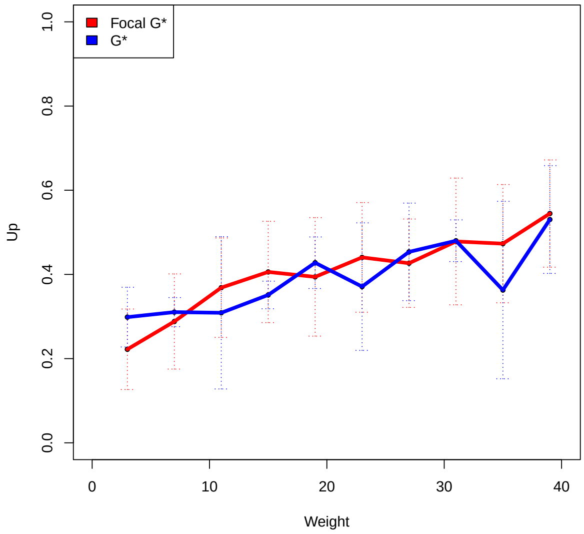


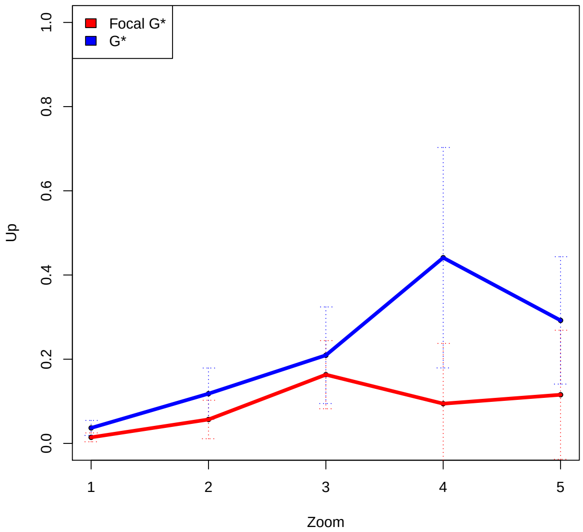
Figure 4: SoH for weight size

With this work we made a step further to the optimal parametrisation for G^* and Focal G^*. All our results are based on the SoH. Further work should validate or improve this metric and compare if Focal G^* can have better results than G^* in all dimension with fixed weight and focal size.

8 Acknowledgements

This work is part of the research project BigGIS (reference number: 01IS14012) funded by the Federal Ministry of Education and Research (BMBF) within the frame of the programme "Management and Analysis of Big Data" in "ICT 2020 – Research for Innovations". We thank the Nachbarschaftsverband Karlsruhe for the data of the thermal flight over Karlsruhe. R-packages used: `raster` [5], `knitr` [10]

(a) SoH up for focal size



(b) SoH down for focal size

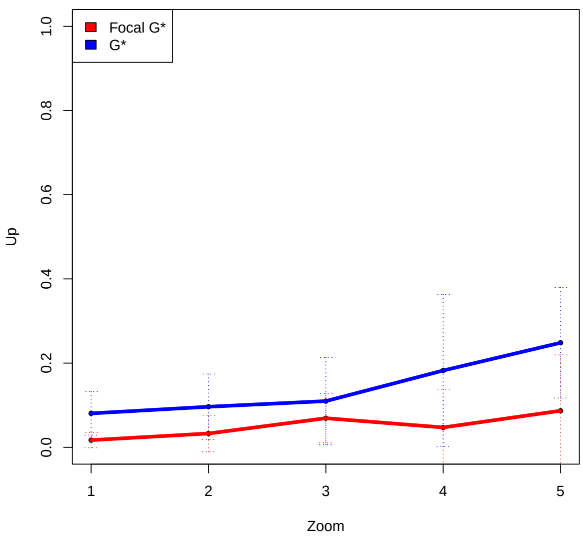


Figure 5: SoH for focal size

References

[1] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, and Jörg Sander. Optics: Ordering points to identify the clustering structure. *SIGMOD Rec.*, 28(2):49–60, June 1999.

[2] Luc Anselin. Local indicators of spatial association - lisa. *Geographical Analysis*, 27(2):93–115, 1995.

[3] Julian Bruns and Viliam Simko. Stable hot spot analysis for intra urban heat islands, forthcoming. *GI FORUM*, (1), 2017.

[4] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. pages 226–231. AAAI Press, 1996.

[5] Robert J. Hijmans. *raster: Geographic Data Analysis and Modeling*, 2016. R package version 2.5-8.

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

[7] J. K. Ord and Arthur Getis. Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis*, 27, 1995.

[8] Shashi Shekhar, Michael R Evans, James M Kang, and Pradeep Mohan. Identifying patterns in spatial information: A survey of methods. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3):193–214, 2011.

[9] Y. Song. Class compactness for data clustering. In *2010 IEEE International Conference on Information Reuse Integration*, pages 86–91, Aug 2010.

[10] Yihui Xie. *knitr: A Comprehensive Tool for Reproducible Research in R*. Chapman and Hall/CRC, 2014. ISBN 978-1466561595.