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 a reduction 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 [8], 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. [10]. 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 [7]. 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 [8] 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 [9].

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. If $|Children| = 0$ or $|Parents| = 0$ SoH is 0.

4 Focal Getis-Ord

Section is analog to [3] and incorporates the most important definitions for the computation. We use the notation $R \overset{op}{\circ} M$ 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 [6].

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

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

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

5 Evaluation

5.1 Dataset

We evaluate our data on the New York city yellow cab data set [1]. 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, -74.020265) and (40.769239, -73.948286). By using this data set, we reduce the computational effort while still being able to show the applicability on real world data and problems.

5.2 Treatments

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.*

Aggregation size 1 means we aggregate every point, which was in range of $100x100 \cdot 0.000001$ into one pixel. For aggregation size Z we aggregated ZxZ pixel from aggregated size 1 into a new pixel. For our measurements we specified the following data series:

- W_i: The weight matrix W is from 3 to 47 with a stepsize of 4.
- F_j: The focal matrix F is from 17 to 137 with a stepsize of 12. F is ignored for G^* but not for Focal G^*
- R_z: The aggregation level is from 1 to 6 with a stepsize of 1. R represents the raster

Definition 5. *The SoH for a singled evalution run in the weight dimension is defined as:*

$$SoH^{\uparrow}(G^*(R_z, W_{3+i\cdot 4}, st, F_{17+j\cdot 12}), G^*(R_z, W_{3+(i+1)\cdot 4}, st, F_{17+j\cdot 12}))$$

$$SoH^{\downarrow}(G^*(R_z, W_{3+i\cdot 4}, st, F_{17+j\cdot 12}), G^*(R_z, W_{3+(i+1)\cdot 4}, st, F_{17+j\cdot 12}))$$

$$z \in [1,6], i, j \in [0,10], G^* \in [Standard, Focal]$$

The SoH for a singled evalution run in the focal dimension is defined as:

$$SoH^{\uparrow}(G^*(R_z, W_{3+i\cdot 4}, st, F_{17+j\cdot 12}), G^*(R_z, W_{3+i\cdot 4}, st, F_{17+(j+1)\cdot 12}))$$

$$SoH^{\downarrow}(G^*(R_z, W_{3+i\cdot 4}, st, F_{17+j\cdot 12}), G^*(R_z, W_{3+i\cdot 4}, st, F_{17+(j+1)\cdot 12}))$$

$$z \in [1,6], i, j \in [0,10], G^* \in [Standard, Focal]$$

The SoH for a singled evalution run in the aggregation dimension is defined as:

$$SoH^{\uparrow}(G^*(R_z, W_{3+i\cdot 4}, st, F_{17+j\cdot 12}), G^*(R_{z+1}, W_{3+i\cdot 4}, st, F_{17+j\cdot 12}))$$

$$SoH^{\downarrow}(G^*(R_z, W_{3+i\cdot 4}, st, F_{17+j\cdot 12}), G^*(R_{z+1}, W_{3+i\cdot 4}, st, F_{17+j\cdot 12}))$$

$$z \in [1,6], i, j \in [0,10], G^* \in [Standard, Focal]$$

Therefore we would calculate $10 \cdot 10 \cdot 5 + 10 \cdot 5 = 550$ results for each dimension. The following conditions must hold dim(F)>dim(W) and dim(F)<dim(R) which leads e.g. to 460 results in the z dimension. We vary over the weight matrix W, the focal matrix F and the aggregation level, as motivated in the introduction. To compare the impact each of these parameters has we compute all variations over two parameter and holding one parameter fix. We then calculate the mean as well as the standard deviation for the SoH for the fixed value based on the two other parameters. This allows us to isolate the impact of the variation in the single parameter. The results are then plotted in Fig. 1-6, one graphic for each fixed parameter and the direction of the SoH. The hotspots to use the SoH on are computed by the focal G^* and standard G^* method. The focal G^* allows us, to isolate the impact that the size of the study area has on the stability of the results without recomputing the total dataset.

6 Results and Discussion

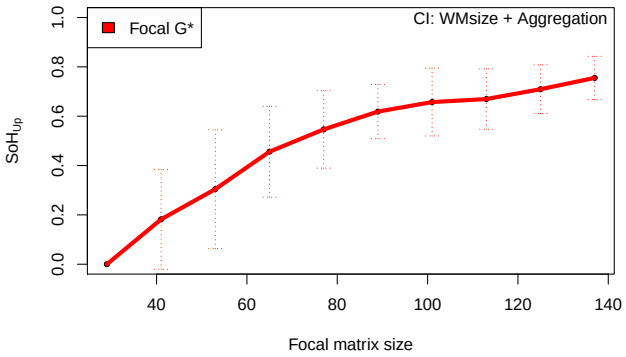
6.1 Impact of study area

The first parameter we want to isolate is the size of the study area. This enables us to discuss the optimal size of the study area for the given research/decision question. To do so, we have to fix the focal size F and use it as the x-axis and plot the mean value as well as the mean +- the standard deviations of the SoH for all other parameter combinations as the y-axis. In this scenario, only the Focal G^* can produce varied results as the standard G^* has by definition a fixed study area by always using the total size of the study area. In figure 1 the $SoH^{\uparrow}$ and $SoH^{\downarrow}$ is shown for G^* and Focal G^*. Higher values indicate that the found hot spots are more stable. The $SoH^{\uparrow}$ in figure 1a growth with an increase of the focal size. It can see that an increase of the focal size leads to better results for $SoH^{\uparrow}$. With a higher focal size you can see that the dotted line, which represent the standard deviation, is getting smaller. This indicates that an increase in the size of the study area does not only lead to an overall increase of stability but also that the it

[1] $http://www.nyc.gov/html/tlc/html/about/trip\_record\_data.shtml$

reduces the volatility of the variations in other parameters. When F and W have similar size, there are no clusters. Because the values have no much variation. $SoH^{\downarrow}$ (figure 1b)shows a different picture then the $SoH^{\uparrow}$. It reaches its maximum SoH for a focal size F of 65x65 in our evaluation set. The standard deviation is not decreasing with an increase of the focal size.

(a) SoH up for focal size



(b) SoH down for focal size

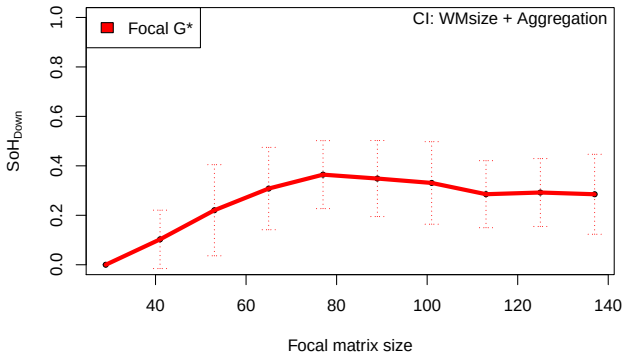
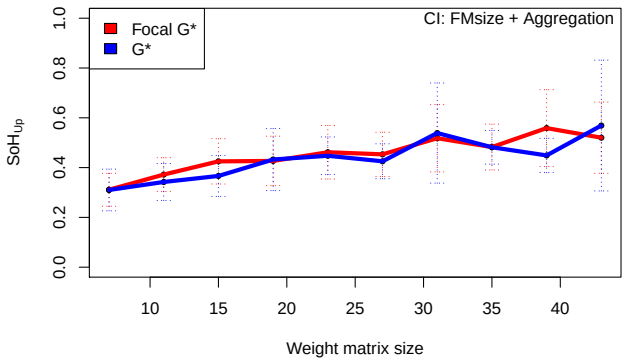


Figure 1: SoH for focal matrix size

increasing. We assume that the reason behind this effect is the inclusion of the water near Manhattan, which leads to a stronger differentiation between areas near water and areas more in the inner city. by increasing the weight matrix, the number of points which are influenced increase in number. The other parameter can regulate this impact and therefore the variance is stronger.

(a) SoH up for weight size



(b) SoH down for weight size

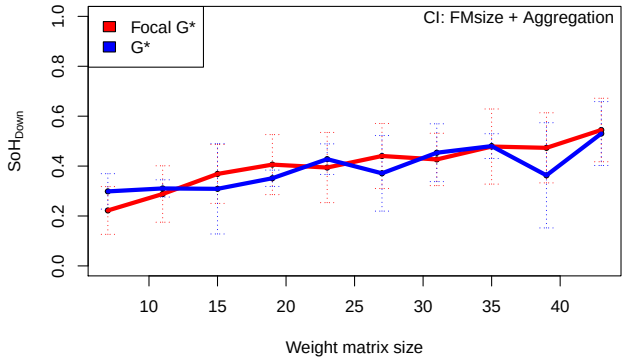


Figure 2: SoH for weight size

6.2 Impact of Neighbourhood size

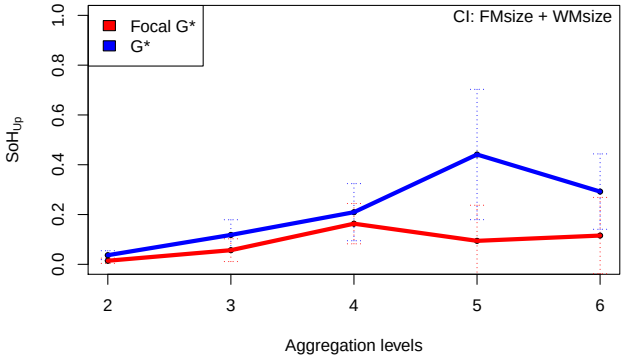
Next, we want to isolate the effect of the neighbourhood size by fixing the weight matrix size W and using it as x-axis. With the impact of the weight matrix W we gain information on how far the interconnectness is between different points as indicated in [5]. The results for the weight matrix can be seen in figure 2a for $SoH^{\uparrow}$ and figure 2b for $SoH^{\downarrow}$. Here it can be seen that in general for both evaluations metrics as well as the different hot spot methods an increase in the size of the weight matrix W leads to more stable results as well as that the Focal G^* matrix is slightly more stable in most cases for the mean values. An interesting phenomena is the dip in stability for the standard G^* for the value of 39. Despite our effort we could not determine the reason behind this result, as no overall trend could be extracted. The standard deviation shows similar no particular trend, but for the values above 37 is

6.3 Impact of Aggregation level

Finally, we isolate the impact of the different aggregation levels on the stability of the hot spot analysis. We fix the aggregation and use it as x-axis. This allows us to examine the impact the resolution of a data set has on the results and allow us indirectly to reduce the computational effort on future computations by using the maximal aggregation as useful. The results can seen in figure 3. First, we can see that for the zoom level, in contrast to previous results, the standard G^* seems to be more stable.

For the $SoH^{\uparrow}$ there is a huge increase between aggregation level 4 and 5. This increase can't be seen for $SoH^{\downarrow}$. One can see that G^* is always better than Focal G^*. This could be beacause the target area of Focal G^* increases with every aggregation step.

(a) SoH up for focal size



(b) SoH down for focal size

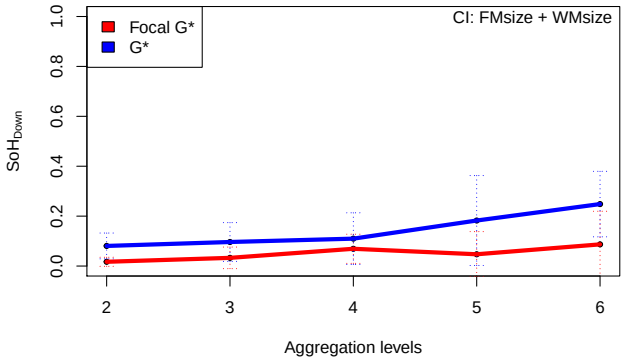


Figure 3: SoH for focal size

7 Conclusions and Future Work

In this work we examined the influence of different parameter on the stability of hot spot analysis on the basis of the Getis-Ord and Focal Getis-Ord statistic. We validated our results with the use of the SoH metric on a well known real world data set to ensure external validity and easy replication. Based on the result we can show several insights, given the restrictions of this work, for future use of hot spot analysis. First, the greater the area size, the more stable the results seem to be. The same relation seem to be given for the size of the weight matrix. The greater the comparable area, the more stable are the results. Given the study are, the focal range should be at least be a size of 65x65 tiles, to have a good trade-off between $SoH^{\uparrow}$ and $SoH^{\downarrow}$. In the case of the aggregation level, the stability does not seem to be impacted in the case of the Focal G^*, but for the standard G^* statistic we see an increase in stability for higher aggregation levels. We assume that the high aggregation of data and the inherent increase of examined area reduce the impact of outliers, but reduce the potential to differentiate. The worse results for the Focal G^* statistic compared to the G^* statistic are most likely a result of the increased focus on a smaller region. A more in depth look at the interaction between focal size and zoom level would be an interesting question for the future, but this is beyond the scope of this work.

With this work we made a step further to the optimal parametrisation for G^* and Focal G^*. Our results indicate that the parametrisation for G^* and Focal G^* besides what is defined as parent and child has a huge impact on the stability of hotspots. We can see the increase of up to 0.6 $SoH^{\uparrow}$ for e.g. the focal matrix size. This emphasizes the importance of a metric for the stability of hotspots. But this work also shows the shortcomings of the existing metric, as the trade-off between $SoH^{\uparrow}$ and $SoH^{\downarrow}$ is not yet fully explored and should be the focus of future work. Another interesting field for future research is the stability of spatio-temporal hot spots and their different parametrizations. As G^* is often applied with regard to temporal impacts, the efficient computation of the focal G^* for spatio-temporal and the impact on stability is the second main path for future research. We evaluated our data on the aggregate of a singular month, but one could assume that at lunchtime there will be hotspots at restaurants and they are not consistent with the hotspots over one month. Therefore the impact of time on the results is important and could also influence the metric for stability in profound ways. As in the example the question is, how fine grained should the temporal clustering be and when is the result unstable for useful reasons? This leads to the final future work: To which hot spots should the comparison be made for the metric? In the current state only the next parametrization is compared for computational time reasons. How the comparison hot spots should be chosen and how many different comparisons have to be used is another quite interesting question for future work.

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". R-packages used: `raster` [6], `knitr` [11]

9 Appendix

9.1 Clumping

The SoH needs a parent and a child. Clusters can be calculated different. For G^* a sigmoid can be used. Values above 2.58 and under -2.58 means in less then 1% of the cases you are wrong. The values of Focal G^* depend on the focal size F. Therefore, we decide all values in the top quantile belong to a clusters. For clustering we grouped the neighbourhood with values in the top quantile to one cluster. Neigbours are are pixels with touching edges which leads for every pixel to 8 neigbours. The clustering can be expressed in R with clumping, TODO

which package and parametrisations. In Fig. 4 we visualized the clusters. The clusters are marked in different colours yellow (child) and violet (parent). It can seen that the results from G^* and Focal G^* differ. Focal G^* has more smaller clusters which are distributed over the complete raster. Not overlapping clusters should be not stable because if we changed the parametrisation we want that the clusters do not disappear or appear. The SoH defines in more detail, which changes for the clusters are stable and which changes indicate that the clusters are instable.

9.2 Aggregate

The aggregation level is a slice plane through our three dimensional space. This is an example how the results can look with visualize representations of the calculated rasters and the raw data. We evalute only the aggregation with fixed W and fixed F. In contrast we included the negative clusters. They where exluded for our 3 dimensional space because the water had to much influence.

Definition 6. *Aggregation run:*

$$SoH^{\uparrow}(G^*(R_z,W,st,F),G^*(R_{z+1},W,st,F))$$

$$SoH^{\downarrow}(G^*(R_z,W,st,F),G^*(R_{z+1},W,st,F))$$

$$z \in [1,6], dim(W) = 11, dim(F) = 41, G^* \in [Standard, Focal]$$

The results can be found in figure 5. It can be seen that with an increase of the aggregation level the values increase. TODO if clusters increase/decrease For the upward property, G^* is always better than Focal G^*, compared to the results from figure 3. Focal G^* reaches the same result as G^* at zoom level 5 . For the downward property, Focal G^* leads to better results at zoom level 5.

9.3 Blur

We also made an example for the variation of the weight size W. Blur is a slice plane through our three dimensional space. W is changed from 7 to 29 with a stepsize of 2, F and the aggregation level is fixed. For blur we also included the negative clusters which resulted in a surprising result.

Definition 7. *Blur run:*

$$SoH^{\uparrow}(G^*(R_z,W_{7+i},st,F),G^*(R_z,W_{7+i+2},st,F))$$

$$SoH^{\downarrow}(G^*(R_z,W_{7+i},st,F),G^*(R_z,W_{7+i+2},st,F))$$

$$z = 3, i \in [1,6], dim(F) = 41, G^* \in [Standard, Focal]$$

The blur results are plotted in Fig. 6. The weight matrix is shown in the first row, with increasing size. The scond and thrid row sho how G^* and Focal G^* changes based on W. This example is a slice plane of the weight dimension ploted in figure 2. It can be seen that increase of the weight size leads in general to better results for Focal G^*. G^* compared to the mean of our whiskerplot does not fit in the picture because it leads to better results than G^*.

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. *Journal GI_Forum 2017*, (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] Arthur Getis and J. K. Ord. The analysis of spatial association by use of distance statistics. *Geographical Analysis*, 24(3):189–206, 1992.

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

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

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

[9] 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.

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

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

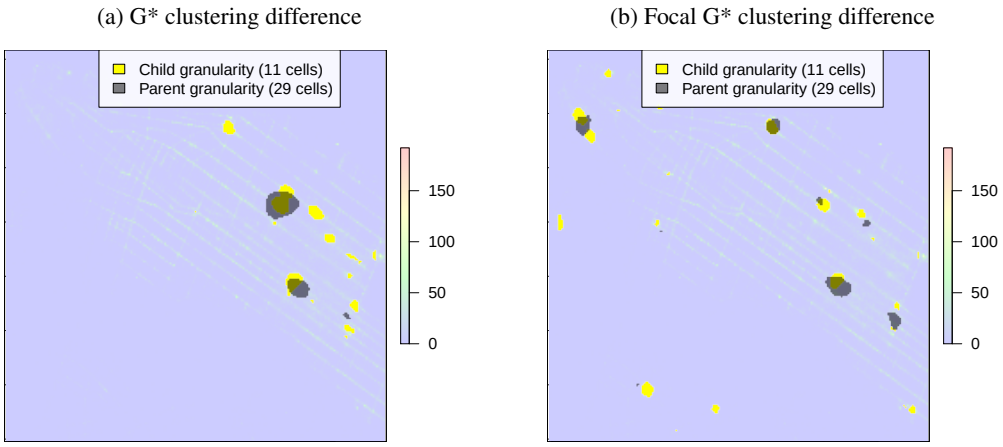


Figure 4: Example clustering for G* and Focal G* used as a input for SoH computation.

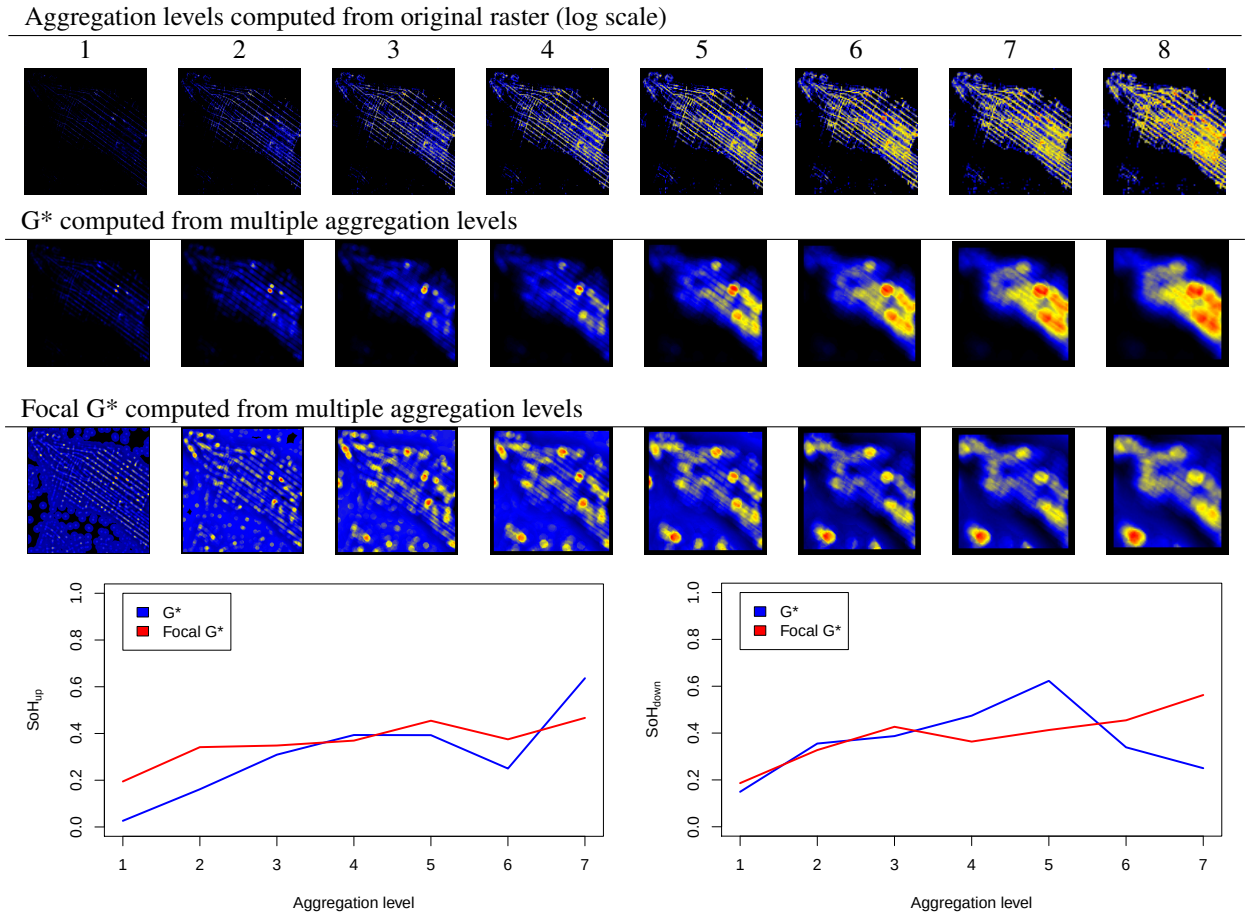


Figure 5: This image shows all aggregation levels for G* and Focal G* together with two metrics – $SoH^{\uparrow}$ and $SoH^{\downarrow}$.

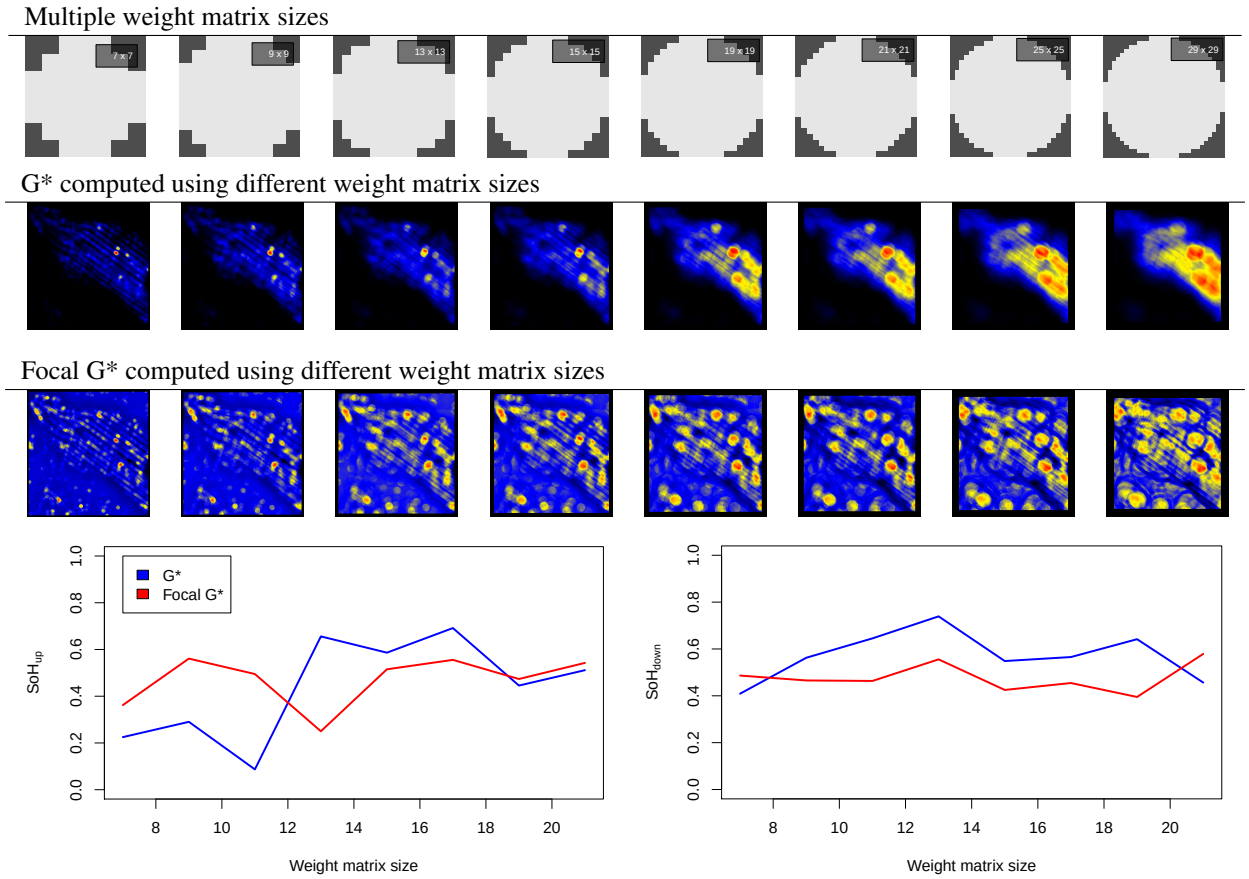


Figure 6: This image shows different weight matrix sizes for G* and Focal G* together with two metrics – $SoH^{\uparrow}$ and $SoH^{\downarrow}$.