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 [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. 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 [5].

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

Our results are based on the cab taxi dataset 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).

- To evaluate, we compare G* with FocalG* on the same dataset. - We use NY taxi dropoffs

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:*

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

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

$$z \in [1,6], i, j \in [1,11], G^* \in [Standard, Focal]$$

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

$$SoH^{\uparrow}(G^*(R_z, W_i, st, F_j), G^*(R_z, W_i, st, F_{j+1}))$$

$$SoH^{\downarrow}(G^*(R_z, W_i, st, F_j), G^*(R_z, W_i, st, F_{j+1}))$$

$$z \in [1,6], i, j \in [1,11], G^* \in [Standard, Focal]$$

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

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

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

$$z \in [1,6], i, j \in [1,11], G^* \in [Standard, Focal]$$

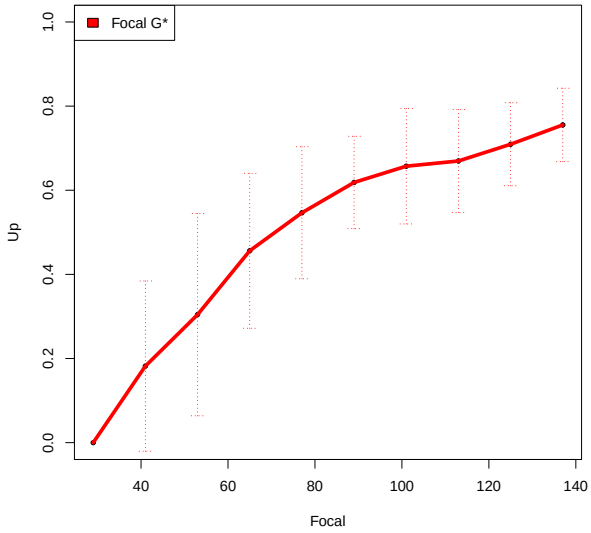
Therefore we would calculate $10 \cdot 10 \cdot 5 + 10 \cdot 5 = 550$ results for each dimension. The following conditions must hold dim(F)>dim(W) and dim(F)<dim(R) which leads e.g. to 460 results in the z dimension. We vary over the weight matrix W, the focal matrix F and the aggregation level, as motivated 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.

6 Results and Discussion

First we fixed the focal size F as our x-axis. Therefore we have no G^* to compare against Focal G^* because G^* has a fixed study area. To evaluate this dimension is essential because if it is not known how big the study area needs to be. Based on these results we want to fix the focal size F so that in most cases focal G^* is better than G^*. In figure 1 the $SoH^{\uparrow}$ and $SoH^{\downarrow}$ is shwon for G^* and Focal G^*. Higher values mean 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. Therefore the standard deviation decreases with a higher focal size. Only at focal size F of 17x17 there is no standard deviation. Which is presented with the absent of the dotted line. At focal size F 17x17 the standard deviation is missing, because the weight size W must be smaller than F. And for focal size F the values for W which are caluclated are only 3,7,11. When F and W have similar size, there are no clusters. Because the values have no much variation. $SoH^{\downarrow}$ can also be calculated with the same images. $SoH^{\downarrow}$ in figure 1b reaches its maximum for focal size F 65x65 in the regarded area. The standard deviation is not decreasing with an increase of the focal size. Also the standard deviation is missing for the first point.

Next we fixed the weight dimension as our x-axis. The values are increasing with an increase of the weight size

(a) SoH up for focal size



(b) SoH down for focal size

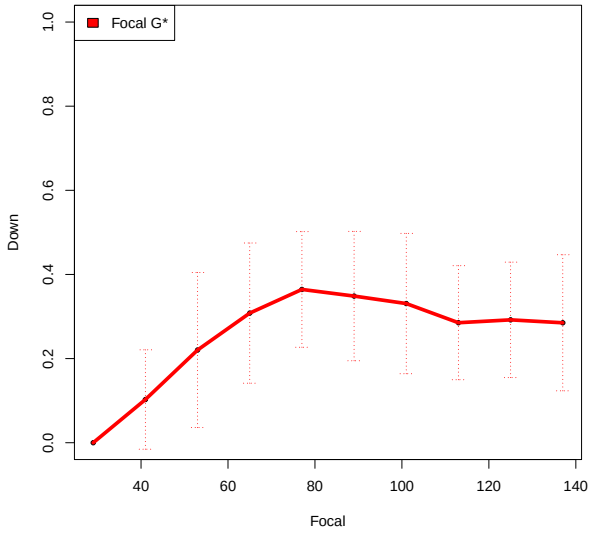
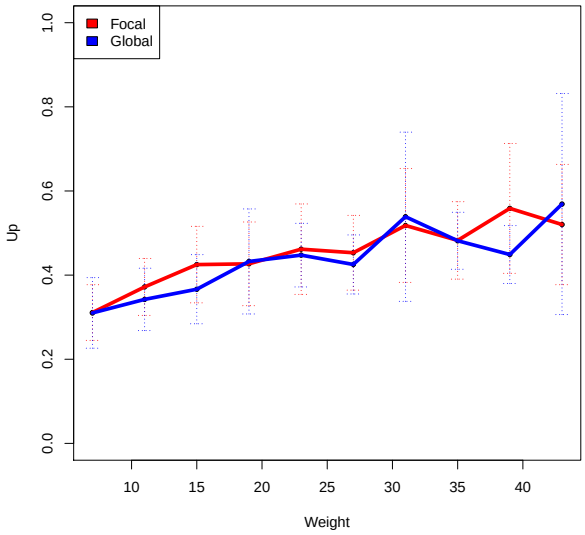


Figure 1: SoH for focal size

W. Higher values means the results are more stable. In figure 2a one can see that an increase of the weight size leads to better results forh the $SoH^{\uparrow}$. We compare G^* wiht Focal G^* . The Focal G^* is not always above the G^*. Therefore Focal G^* is not allways better than G^* which would be a perfect result we hoped for. It seems that G^* has at 23 and 39 some small break downs. At first glance one would say that some parametrisation leads to that result but the standard deviation is decreasing which indicate the opposite. The standard deviation is in this area higher than in the other points. In figure 2b the break downs are also visible which uses the same images as before but calculate the $SoH^{\downarrow}$. The high standard deviation in general indicates that it is possible to use better parametrisation for better results. Which would be the zoom size. The break downs at 23 and 39 are also visible but this time the standard deviation is increasing. The $SoH^{\downarrow}$ and $SoH^{\uparrow}$ is mostly better than G^* with fixed x-axis for the weight. Which was the conclusion in **??**.

(a) SoH up for weight size



(b) SoH down for weight size

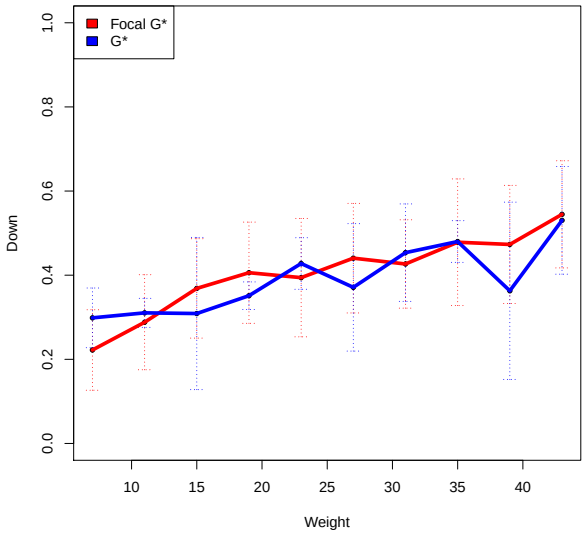
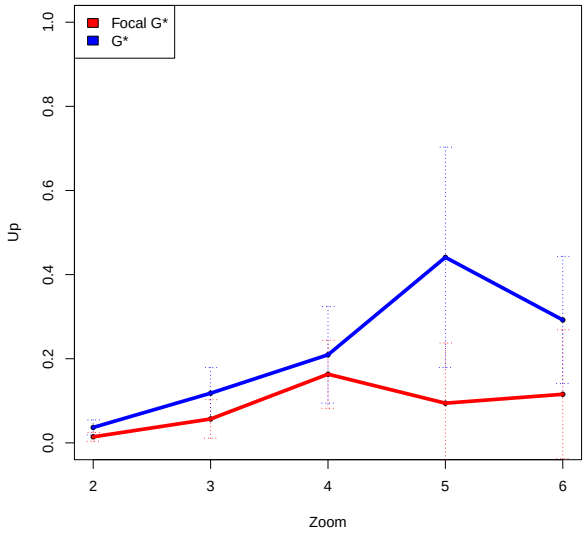


Figure 2: SoH for weight size

The last variation we examined was the aggregation

level. There we fixed the x-axis as our aggregation level. And the focal size and weight size variation lead to a sample set. The results can seen in figure 3. 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

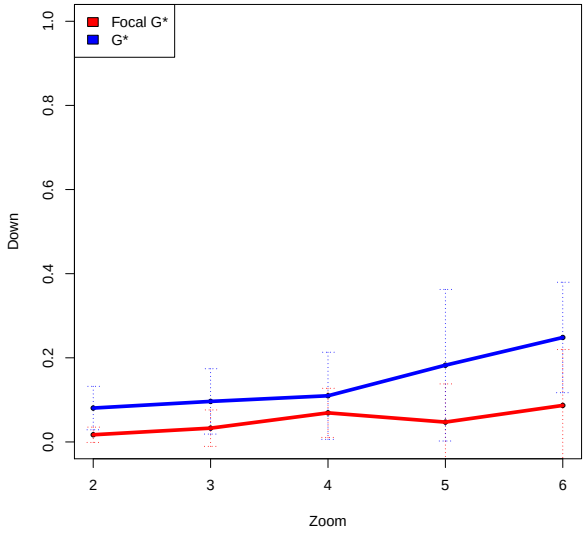


Figure 3: SoH for focal size

6.1 Clumping

6.2 Aggregate

The aggregation level is a slice plane through our three dimensional space. We evalute only the aggregation with fixed W and fixed F.

Definition 6. *Aggregation run:*

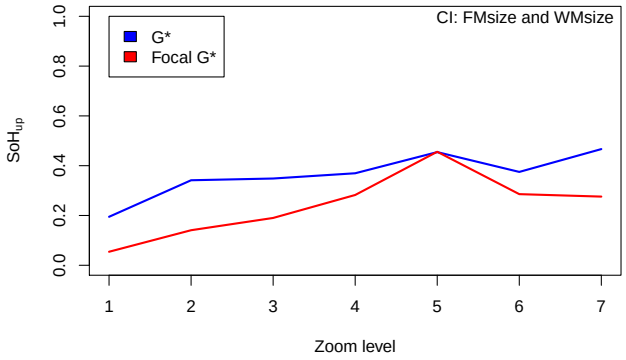
$$SoH^{\uparrow}(G^*(R_z,W,st,F),G^*(R_{z+1},W,st,F))$$

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

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

The results can be found in figure 4. 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.

(a) Upward for zoom



(b) Downward for zoom

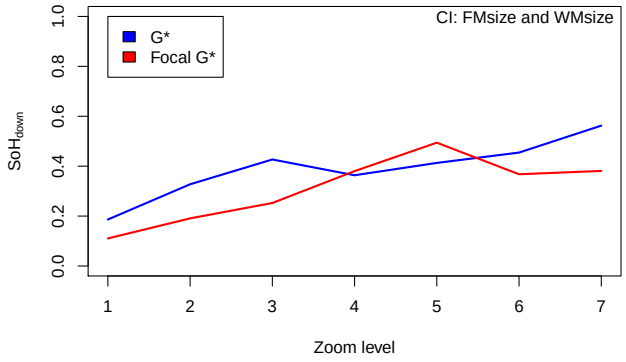


Figure 4: SoH zoom

6.3 Blur

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.

Definition 7. *Blur run:*

$$SoH^{\uparrow}(G^*(R_z,W_i,st,F),G^*(R_z,W_{i+1},st,F))$$

$$SoH^{\downarrow}(G^*(R_z,W_i,st,F),G^*(R_z,W_{i+1},st,F))$$

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

The blur results are plotted in Fig. 5. The slice plane of the weight dimension is consistent with figure 2. Higher weight leads in general to better results for Focal G^*. G^* doesn't fit to the mean of our whiskerplot.

7 Conclusions and Future Work

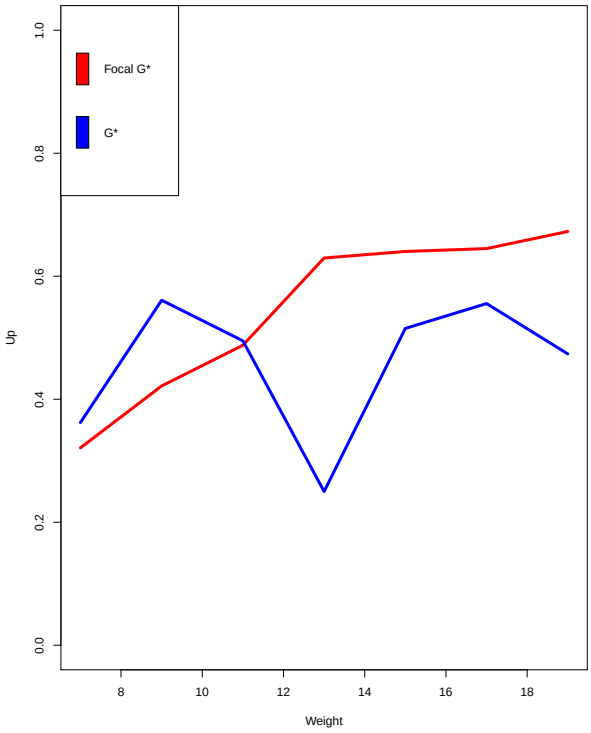
In this work, we compared the Getis-Ord statistic with Focal Getis-Ord statisitic. We validate the result from **??** with a real world dataset an with a larger dataset. The dataset was only one month projected in one raster. We examined further parameters and evaluated that G^* and Focal G^* performe 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.

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. Wrong parametrisation can lead to good results, the mean reaches for $SoH^{\uparrow}$ in the weight dimension around 0.6. It can also lead to bad result, $SoH^{\uparrow}$ in the zoom dimension reaches nearly zero. Therefore we shown that it is important to have a metric for the stability of hotspots. Also we presented indication which parametrisation should be avoided and which parametrisations are promising. G^* can also be applied to spatio-temporal data. Therefore Focal G^* should be also applied to the spatio-temporal data. We now have a good indication which weight and focal size is usefull. We didn't evaluate the results on other month because we assume that they are similar But it is interesting to find clusters for different times in the day. One could asume that at lunchtime there will be hotspots at restaurants and they are not consistent with the hotspots over one month. All our results are based on the SoH. Further work should validate or improve this metric also for the temporal dimension.

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

(a) Upward for blur



(b) Downward for blur

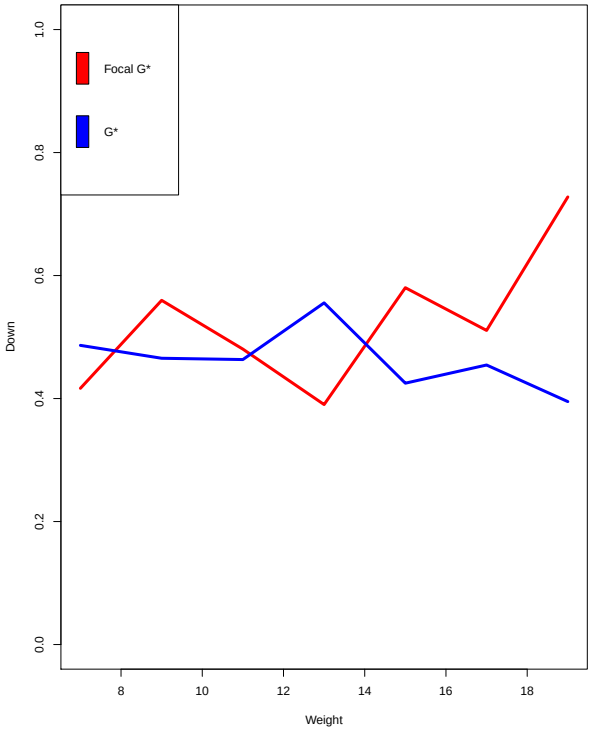
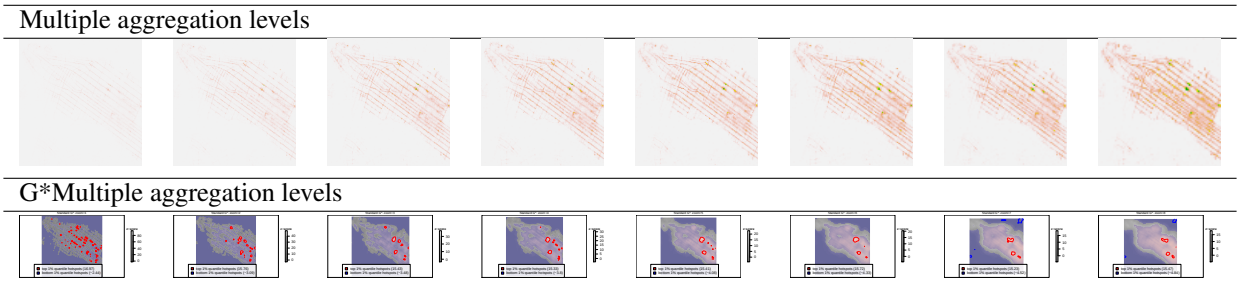


Figure 5: SoH blur



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]

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 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] 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.