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: Hotspot 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 hot spots 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 (Hier Bruns and VIliam 2017), 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 [**?**], the standard G^*, and a modification of this statistic, the focal G^* (citate bruns and viliam). 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

- goal Here: some cluster analysis
- typical clustering apporaches -> phd symposium as examples (possible to cite =))

The goal of an analysis of temperatures in a city is to find the most interesting, significant areas: Hot spots [**?**]. This goal is similar to the hot spot analysis in the field of geo-statistics. One of the most fundamental approach is Moran´s I [**?**]. 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 [**?**] and LISA [**?**]. 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 hot spots and to measure the significance of those local areas.

The local Getis-Ord statistic [**?**] is defined as follow:

Definition 1 (Getis-Ord G_i^* statistic)**.** *Assuming a study area with n measurements, let* $X = [x_1, \ldots, x_n]$ *be all values measured in this area. Let* $w_{i,j}$ *be a spatial weight between two points i and j for all* $i, j \in \{1, \ldots, n\}$*. The Getis-Ord*

G_i^* *statistic is given as:*

$$G_i^* = \frac{\sum_{j=1}^{n} w_{i,j}x_j - \bar{X}\sum_{j=1}^{n} w_{i,j}}{S\sqrt{\frac{n\sum_{j=1}^{n} w_{i,j}^2 - (\sum_{j=1}^{n} w_{i,j})^2}{n-1}}} \quad (1)$$

where:

- $\bar{X}$ *is the mean of all measurements,*
- S *is the standard deviation of all measurements.*

As it is known, this statistic creates a z-score, which denotes the significance of an area in relation to its surrounding areas.

LISA is quite similar, as it is the local statistic for Moran's I [**?**], but the z-score has a different meaning. Apart from G_i^*, LISA does not distinguish between cold spots and hot spots as it assigns high z-score to most similar areas.

Two other well known method are the kernel density estimation [**?**] and kriging [**?**]. These do not provide significance levels. Instead, they estimate the values for each location based on the rest of the study area and a threshold value [**?**]. Therefore, results for different areas are not comparable, especially in the case of differing temperature distributions. Kriging was developed for the estimation of ore deposits [**?**], but today, applications for geo-temporal forecasts with this approach can be found, e.g. for the city of Zurich. [1]

All of the aforementioned methods use weights between pairs of points, usually based on their geographical distance. However, in real applications, the points are aggregated into rasters and the weights are represented as a weight matrix. This allows for expressing the algorithms in terms of map algebra operations, a term first coined by Dana Tomlin [**?**] and computed in a distributed fashion (e.g. using Geotrellis framework running on Apache Spark [**?**]).

In this work, we focus on the Getis-Ord statistic applied to rasters of land surface temperature. This enables us to transform the formula for the G^* statistic into a computationally more efficient form. We then propose a modification of the standard G^* statistics, to increase the stability of the hot spots found.

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 hotspot 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:

$$SoH^{\downarrow} = \frac{ParentsWithChildNodes}{Parents} = \frac{|Parents \cap Children|}{|Parents|} \quad (2)$$

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

$$SoH^{\uparrow} = \frac{ChildrenWithParent}{Children} = 1 - \frac{|Children - Parents|}{|Children|} \quad (3)$$

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

The two datasets (morning and evening flights) depicted in Fig. **??** and Fig. **??** were obtained from a thermal flight over the city of Karlsruhe on 26.09.2008 at 6:30–7:45 and 20:00–21:30. The flights were executed by the Nachbarschaftsverband Karlsruhe[2]. A single pixel in the raster represents an area of approximately $5\times5m$. The whole dataset of size $35\times25\,km$ was cropped into the inner city area of $2.4\times1.4\,km$. The temperatures in our dataset range from -1.7°C to 18.3°C. Missing values in the dataset were interpolated using a focal median function with a square matrix of 11x11 pixels, mainly for speeding up further computations and to avoid special handling of NA values.

[1] https://r-video-tutorial.blogspot.de/2015/08/spatio-temporal-kriging-in-r.html

[2] http://www.nachbarschaftsverband-karlsruhe.de/

4.2 Method

In the following text, 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 [**?**].

Definition 2 (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. 3 and 4).*

Definition 3 (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 4 (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.*

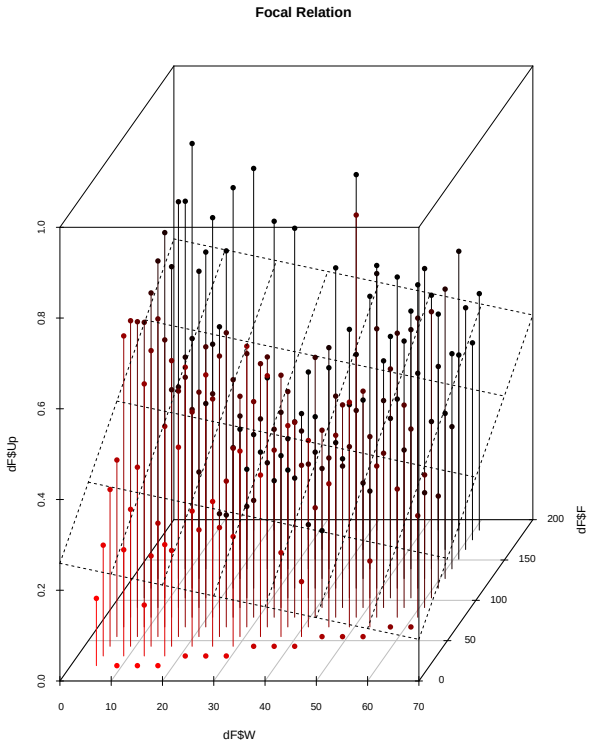
5 Results

5.1 Heatmap

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. 5. - In an ideal case, we could produce a

(a) FocalRelation



(b) WeightRealtion

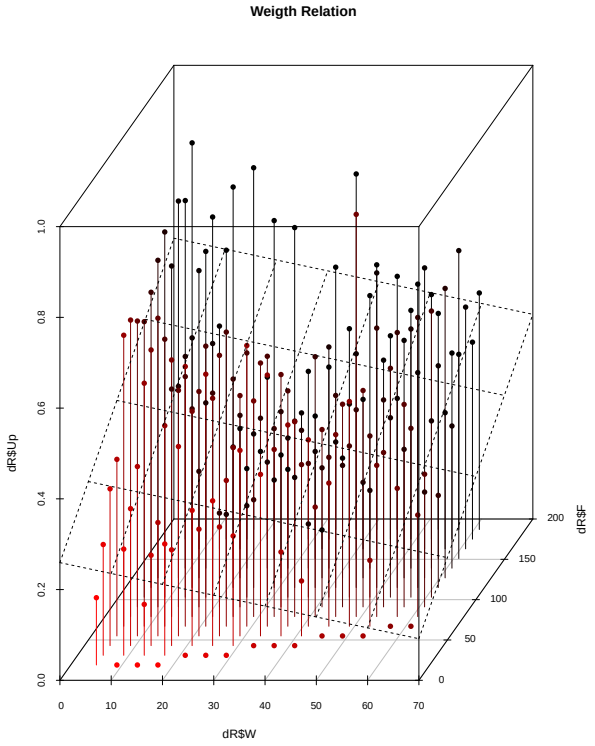


Figure 1: Heatmap of Relation between size of weight matrix W and focal matrix F

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 5 (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

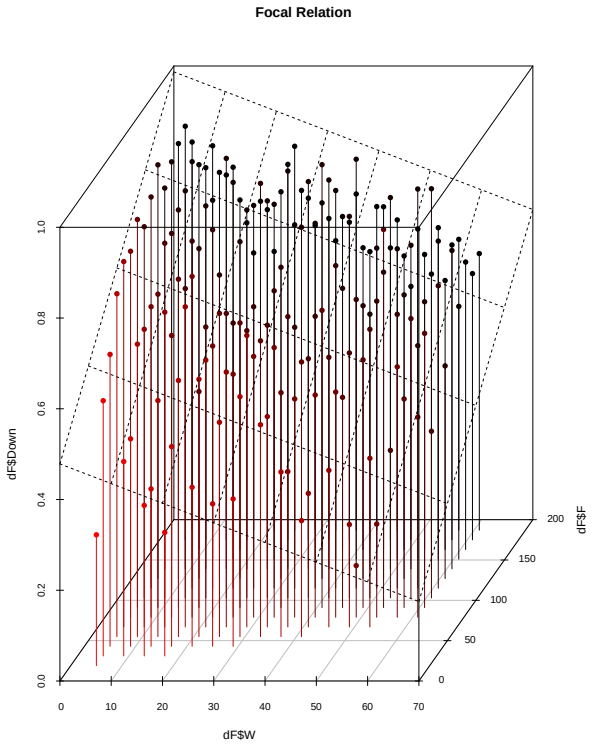
The evaluation results are plotted in Fig. **??**, each point in the graph represents the $SoH^{\uparrow}$ metric (Eq.3) between two G^* generated using weight matrices of size i and $i+2$. The focal matrix F has a fixed size of 41×41

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

6.4 Results and Discussion

The results for the hot spot analysis are found in Fig. **??** and Fig. **??** for a comparison of G^* and the Focal G^* statistic. It can easily be seen that both versions produce similar results, but the focal versions produces a more differentiated picture for larger weight matrices. Small differences on a global scale are more pronounced on a regional scale and result in smaller and finer areas for hot spots. This enables the detection of additional hot spots and interesting areas which are most easily observable for the weight matrix of size 7×7 in the evening (Fig. **??**). This enables

(a) FocalRelation



(b) WeightRealtion

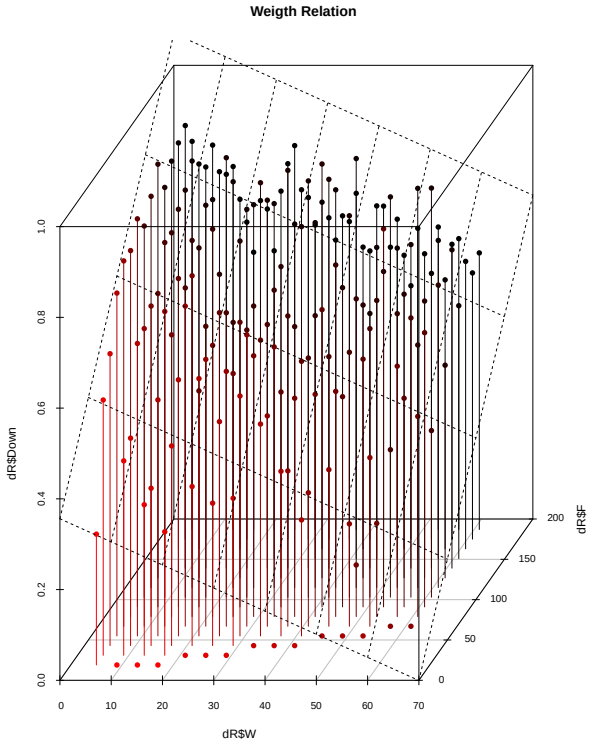
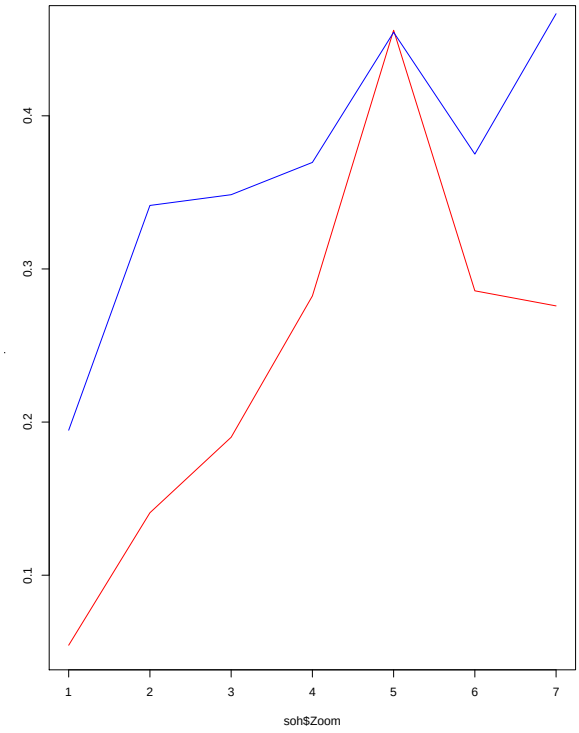


Figure 2: Heatmap of Relation between size of weight matrix W and focal matrix F

(a) Upward



(b) Downward

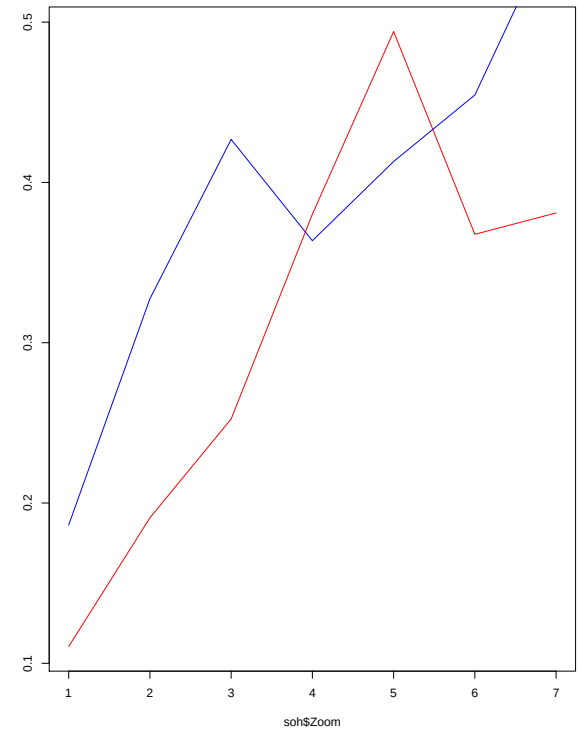
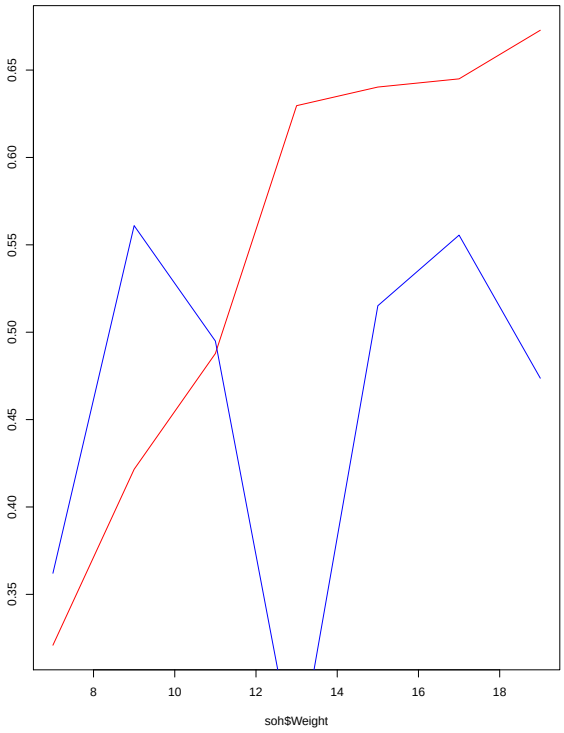


Figure 3: TODO

(a) Upward



(b) Downward

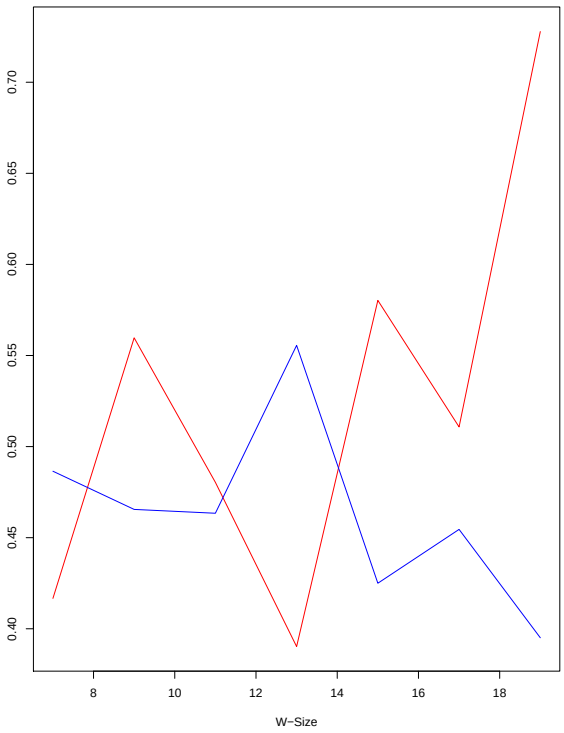


Figure 4: TODO

the detection of significant deviations from the surrounding area. In contrast the Standard G^* statistic shows larger areas as important. Therefore, depending on the need of a planner, the Focal G^* statistic is more helpful to identify individual areas of interest whereas the standard G^* statistic gives a more broad overview. For the identification of IUHI this is quite important. As a city planer wishes to detect critical areas, it is important to detect not only general hot areas, but also those points where the most extreme differences in a local context exist. Finding those areas can help identify the underlying reasons or plan individual solutions.

Based on these images one can also see that the hot spots found by the Focal G^* statistic seem to be more stable. We compare their stability using only the $SoH^{\uparrow}$ for lack of space. A plot of the results can be found in Fig. **??**. Fig. **??** compares the $SoH^{\uparrow}$ between each increasing size of the weight matrix W. At a first glance, one can see that the typical implementation with a square weight matrix is the most unstable hot spot analysis, regardless of time of day. This is to be expected as the binary weights increase the dependence on the weight matrix. The use of a decreasing weight matrix perform better. As the more outlying data points get less weight, this reduces the dependence on the weight matrix and leads therefore to more stable results. Our proposed Focal G^* statistic achieves the most stable results in almost all cases. Only data points in a restricted region around the area of interest may influence the significance result. Through this restriction high values at key points gains more weight regardless of the weight matrix and are therefore more independent of the weight matrix. This increases the stability. The decrease in stability for the largest weight matrices is most likely a result of the parametrization of the focal matrix. With increasing size of the weight matrix in relation to the focal matrix the value of each pixel is approaching to the mean of the area of the weight matrix. As can be easily seen from Def. 2 then the value for every pixel would be zero.

7 Conclusions and Future Work

In this work, we generalized the Getis-Ord statistic to deal with the problem of stability inherent in hot spot analysis. We identified possible underlying reasons for this instability: The weight matrix as well as the size of the study area. We developed a modified approach that deals with those two factors. The result is a modified G^* statistic called the *Focal G^* statistic*. It reduces the study area used for comparison into regions and achieves through this an increase in stability. To determine the effectiveness of our approach, we propose a stability metric for hot spots called *SoH*. To our knowledge no such metric existed before this work. The SoH computes the ratio of dependence of hot spots for different parametrizations of weight matrices. It enables to express the stability between each parametrization using single value restricted between zero and one. Based on this number one can decide which parametrization to use and researchers can compare the stability of their methods for unsupervised hot spot analyses. In particular, for temperature values one wishes to detect those areas which have high differences regardless of a particular parametrization. If a hot spot only appears for one parametrization, the information gained for general use is quite small and can even lead to an inefficient allocation of resources.

This research has several restrictions which have to be taken into account. First, we only tested the $SoH^{\uparrow}$ metric. While we assume, based on our graphical analysis, that the $SoH^{\downarrow}$ stability should be similar, we have no hard results. The results themselves are tested on two events in time for a fixed area of the city of Karlsruhe. We have not tested it on smaller or larger study areas, but we assume that the stability of the Focal G^* would stay the same whereas the stability of the G^* statistic would increase with a smaller study area and decrease with a larger study area. This follows the reasoning that the impact of a singular point increases with a decrease of the study area. To test this dependency is an interesting task for future work. The last restriction is the fixed size in this work of the focal matrix for the Focal G^* approach. We only tested one size in this work, but it is highly probable that the size of the focal matrix has an impact on the stability as could be seen in Fig. **??**. While an overall trend can be seen in this work when the size of the weight matrix W and the focal matrix F are almost identical, the exact ratio is beyond the scope of this work. The optimal ratio as well as when the stability suffers from a too similar size are 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". We thank the Nachbarschaftsverband Karlsruhe for the data of the thermal flight over Karlsruhe. R-packages used: `raster` [**?**], `knitr` [**?**]