AN APPLICATION-ORIENTED IMPLEMENTATION OF HEXAGONAL ON-THE-FLY BINNING METRICS FOR CITY-SCALE GEOREFERENCED SOCIAL MEDIA DATA

D. Weckmüller [1] *, A. Dunkel [1]

[1] Institute of Cartography, Technische Universität Dresden, Helmholtzstr. 10, 01062 Dresden, Germany –
dominik.weckmueller@mailbox.tu-dresden.de, alexander.dunkel@tu-dresden.de

KEY WORDS: geo-social media, hexagonal binning, signed chi, urban planning, spatial dashboards

ABSTRACT:

The use of georeferenced social media data (GSMD) for informing municipal policy-making has significant potential, particularly in addressing pressing socio-environmental challenges. Geospatial dashboards have emerged as a powerful tool for knowledge communication and supporting urban sustainability. However, there has been little emphasis on how to display and make GSMD more accessible, partly due to their complex nature. Existing visualization tools lack sophisticated methods, especially for complex urban contexts, and the methodological choice can significantly impact the interpretation of results. In this study, we propose the use of hexagonal binning as an interactive visualization method and assess three different on-the-fly binning metrics for mapping GSMD. We expand the use of the signed chi metric for spatial purposes and apply it in a case study in Bonn, Germany. We evaluate the advantages and disadvantages of the proposed metrics as well as visualizations and highlight the challenges of visualizing GSMD particularly in the context of Instagram. Our findings highlight the importance of using appropriate context-dependent visualization methods when analyzing data at the municipal level.

1. INTRODUCTION

The analysis of georeferenced social media data (GSMD) holds broad potential for informing municipal policy-making. Local adaptation to climate change and disaster resilience, transforming city centers, gentrification, and demographic change are significant challenges for municipalities.

In light of these pressing topics, a growing awareness for data-driven decision making has fostered geospatial interfaces that allow practitioners to interactively explore data source (Kitchin, 2016; Jing et al., 2019). As Jing et al. (2019) put it, geospatial dashboards are a potent tool for knowledge communication and supporting urban sustainability. Social media data are "increasingly location-based" (Sui and Goodchild, 2011) and offer the potential of a live feed and continuous reflection of events at scale (Ilieva and McPhearson, 2018).

Despite the pressing demand for a purposeful and tailored representation of spatial data, there has been a lack of attention given to the display of this data, as well as to the facilitation of access to GSMD (Weckmüller and Dunkel, 2022) which is to be attributed partially due to the already intricate nature of social media data (Ilieva and McPhearson, 2018).

In fact, besides the big data character, Kitchen (2016) summarizes the main challenge to be the translation of insight "into new urban theory (fundamental knowledge) and actionable outcomes (applied knowledge)". For this purpose, many studies on map-based visualization of GSMD use traditional cartographic methods, such as pins or choropleth maps based on rasterization methods or statistical areas (such as the European NUTS 2021 classification or US census tract or county level, e.g., Li et al. (2021)) with varying color scales or heatmaps to represent absolute or relative values (Heikinheimo, 2020).

Existing web-GIS solutions are often unsuited to more complex visualizations (Kulawiak, 2022). However more complex visualization methods weighing carefully between appropriate shapes, binning algorithms, colors and color classification methods can seldomly be found.

We argue for a use-case specific choice of visualization methods based on hexagonal bins and show that depending on the user question, neglecting the right means can lead to a false – and in the worst case even diametral – impression which is critical for laypeople.

We assess three different binning metrics, absolute value aggregation, mean calculation and the signed chi metric, building on the original adaptation of the signed chi metric to spatial purposes of Wood et al. (2007) and the recent adaptation of Dunkel et al. (2023).

The metrics are tested in a case study for Bonn using an on-the-fly hexagonal binning method for frontend applications. We then evaluate the advantages and disadvantages of the various proposed metrics and visualizations in terms of their practical applications.

As the overview by Teles da Mota and Pickering (2020) has shown, research involving GSMD from different platforms has become increasingly popular but bears specific problems inherent to the characteristics of big data, often referred to as multiple V's: volume, veracity, velocity, variety and value (Laney, 2001; Abkenar et al., 2021). Access to SM databases, such as Instagram or Twitter, is nowadays limited to capital intensive partner companies (Toivonen et al., 2019) partially in consequence to the Cambridge Analytica scandal (Bruns, 2019). Instagram's public-facing API for example is undocumented and opaque to end-users, causing uncertainty about data selection criteria (Dunkel, 2023). It is highly restrictive and limits the amount of data (volume) one can download, negatively affecting the possibility of up-to-date analysis (velocity). The lack of knowledge about data context and possible biases (veracity) can affect the representativeness of the data subset (value) just like "super users" sharing repeated content which may create noise and skew analysis outcomes if absolute values are solely considered (value).

Teles da Mota and Pickering (2020) point out that research has been conducted mainly for large areas ranging from national parks (e.g. Heikinheimo et al., 2017; Barros et al., 2020; Sinclair et al., 2020; Zhang et al., 2021) to entire countries (often the US,

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-4-W7-2023-253-2023 | © Author(s) 2023. CC BY 4.0 License.

e.g. Li et al., 2021) or seldomly even the whole world (e.g. Dunkel et al., 2023).

In the more recent years instead, a trend towards a more local approach can be observed, growing with the dominating paradigm of the smart city but creating new challenges like an increasing need for user privacy especially for high spatial resolution.

Studies working with data on the municipal level where individual locations and differences of only a few meters play a significant role, are usually not focusing on methodological cartographic issues like appropriate binning methods, color classes or values. The trend of treating the right visualization method as something granted using default methods generally intensifies, the more interdisciplinary a research topic and hence the study complexity becomes.

Place, as practiced by Instagram, poses a unique problem for researchers as well as the platform itself. Users are allowed to create public "Instagram Locations" and tag their posts with a coordinate of their choice, which can then be referenced by other users as well. However, the user is not obliged to provide a clear definition of what geospatial extent is meant by the coordinate they choose, creating ambiguity. For instance, the "Bonn" location's coordinates (50.7333, 7.1, instagram.com/explore/locations/107481562) are situated right in the city's center. What it actually refers to is entirely subject to the user. It could refer to differently interpreted extents of the city center, the official administrative boundaries of Bonn or anything loosely associated with Bonn, including cultural references, events or personal opinions. This ambiguity that Meta is aware of (Dalvi et al., 2014) can be unanimously observed on different zoom levels such as city districts, cities, countries or continents throughout Instagram data and poses an enormous challenge to researchers working with city-scale areas of interest. It is a great example of current challenges of an increasingly spatial society (Sui and Goodchild, 2011) where consumers become prosumers. Oyana and Scott (2008) identified the crucial “expanding subfield of geospatial data management” to be “relatively new”, justifying the lack of more sophisticated methods with a still growing interest. In the spirit of this critique, we aim to lay a solid foundation for both guided application and a starting point for further research in this field.

2. METHODOLOGY

We propose a flexible on-the-fly hexagonal binning approach in JavaScript to tackle the various challenges georeferenced social media data pose. The functional approach presented here is implemented based on the popular Leaflet package (github.com/Leaflet/Leaflet) for interactive web maps and the leaflet-d3 plugin (github.com/bluehalo/leaflet-d3) porting d3 visualizations to Leaflet.

2.1 Hexagonal Binning

Hexagonal binning – also referred to as hexbinning or simply hexbins – “has a long history in spatial analysis” (Khan et al., 2023) and recently become increasingly popular for big data (Poorthuis and Zook, 2015) and GSMD.

As a form of regular zoning, it refers to the division of a geographical area into a tessellation of hexagons by which the underlying data points are aggregated. Hexagons are one of only three regular polygons along with squares and equilateral triangles that can tessellate a plane (Carr et al., 1992).

Based on a binning function, a value is calculated for each hexbin and used in a classification function for assigning color values. These functions can each be fine-tuned in our app, as described below.

Due to their advantages over conventional binning methods (Poorthuis and Zook, 2015), hexbins have been applied to a wide range of fields such as big remote sensing data (e.g. Yao et al., 2023), flood mapping (e.g. Li et al., 2022), ecosystem management (e.g. Levine and Feinholz, 2015), crime mapping (e.g. Rickson, 2023) or risk mapping (e.g. Abante, 2020).

In the context of GSMD, hexbins were already used e.g. in the context of crisis mapping (e.g. Cresci et al., 2015), earthquakes (e.g. Avvenuti et al., 2017), but also fashion (e.g. Poorthuis et al., 2020; Power et al., 2015), regional identity (e.g. Rock and Taber, 2020), urban socio-spatial inequality and spatial justice (e.g. Shelton et al., 2015; Weckmüller and Dunkel, 2022).

Their advantages are tightly linked to their geometrical properties. Firstly, hexagons can reduce sampling bias. Grid-based binning methods with rectangles or triangles are subject to very uneven bin sizes with acute angles, resulting in so-called edge-effects. Hexagons also “yield better approximations than square partitions” and hence have a higher “representational accuracy” (Carr et al., 1992). Particularly, connectivity or movement paths are represented more accurately and create less visual bias in comparison to rectangular shapes (Lewin-Koh, 2021). Hexagons are suited better to represent curves and might have a higher “visual appeal” for hexagons (Carr et al., 1992).

Even though they need more vertices (six) to be represented than squares (four) or triangles (three), they are the closest shape to a circle that can tesselate. Hexagons can easily be represented in common data exchange formats such format such as GeoJSON whereas circles are not supported according to the RFC 7946 specification (Internet Engineering Task Force, 2016).

The similarity to circles yields another advantage for visualization of two variables, as not only the color but also the radius can be used. Different diameters of squares or triangles instead are difficult for humans to differentiate, while the radius of a circle or similar shapes are easier to interpret.

On larger scales, the earths curvature should be respected by choosing an appropriate equal area projection. For the municipal level however, this effect can be neglected.

2.2 Binning Functions

While the above-mentioned studies prove a wide adaptation of hexbins in the context of georeferenced social media, notwithstanding the plethora of practical challenges (Dunkel et al., 2023) little emphasis has been put on how to appropriately bin the data. We observed a general tendency in research visualizations to use easy-to-interpret metrics like absolute or relative values. However, two studies need to be pointed out that try to use more sophisticated metrics by comparing a subset of data to entirety of the data to overcome the shortcoming of absolute and relative values. Such functions are needed in order to identify over- or underrepresented spatial patterns (Visvalingam, 1978).

Poorthuis and Zook (2015) propose a so-called odds ratio for hexbins as a simple method to normalize raw counts in large spatial point pattern datasets, such as social media data. It allows for the normalization by any chosen variable, providing easy-to-interpret ratios that capture the distribution of a phenomenon within the social media use context rather than its popularity within the overall population. It is defined as

$$odds\ ratio = \frac{{}^{obs}/_{exp}}{{}^{\Sigma_{obs}}/_{\Sigma_{exp}}} \qquad (1)$$

where obs is the number of posts in a hexagon respecting a certain filter condition such as a topic classification and exp is an expected value as a baseline such as the number of all posts i.e.

the overall size of the random sample in the hexbin (Poorthuis and Zook, 2015). When e.g. multiple Instagram locations are found in a hexbin, each location has a certain sample size exp and number of matching posts obs. What is considered as the observed value for normalization is up to the analyst (Wood et al., 2007) as any kind of filtered subset could be used here. Poorthuis and Zook (2015) point out a small numbers problem on a large scale where low posts numbers distort the overall representation. They propose a function adding confidence intervals in order to exclude posts with highly varying variance. Wood et al. (2007) propose the use of so-called chi expectation surfaces instead, adapted by Dunkel et al. (2023). We build on these efforts and add a fine-grained approach to tackle particular challenges occurring on the municipal scale. In our demo app, we use three binning functions for different purposes that serve as input for the color classification function. The functions all imply random samples of data which is a challenge and discussed later. The terminology follows Wood et al. (2007) and Dunkel et al. (2023).

Absolute values provide a direct representation of the total count of posts within each hexbin, offering insights into the overall intensity or concentration of social media activity in different geographic areas. This metric can be defined as

$$abs = \Sigma_{obs} \tag{2}$$

Mean values provide a normalized perspective by calculating a quotient based on the number of query-related posts divided by all posts of the sample in a hexbin. The metric reflects the query-related "purity" and is defined as

$$mean = \frac{\Sigma_{obs}}{\Sigma_{exp}} \tag{3}$$

Following Dunkel et al. (2023) the signed chi metric quantifies the significance of a hexbin within a certain confidence interval, providing insights into both the over- and underrepresentation of query-related posts in that area. It is defined as

$$chi = \frac{((obs*norm)-exp)}{\sqrt{exp}} \;\; ; \;\; norm = \frac{\Sigma_{exp}}{\Sigma_{obs}} \tag{4}$$

where $norm$ is a normalizing factor.

2.3 Color Functions

Choosing the right colors for communicating research results is critical (Coalter, 2020).

The color function maps a value from the binning functions – here abs, mean or chi – to a color on a color range. Brewer (2016) presents a valuable guide to best practices and a popular tool, ColorBrewer, which is also utilized in this context, for creating color ranges that can reduce common mistakes in interpretability for those with color blindness (github.com/axismaps/colorbrewer).

Apart from choosing the right colors, the actual color classification plays a crucial role in cartography (Jiang, 2015). We use four common algorithms, each of which is suited for different purposes:

1. equal breaks
2. quantiles
3. natural breaks (Jenks)
4. head/tail breaks

While the sum (Equation (2)) and $mean$ (Equation (3)) calculation are using equal breaks ranging between the minimum and maximum value as default in our app, signed chi requires slightly more overhead. We used a critical value of chi of 3.84 (1 degree of freedom, $p < 0.05$) and hence used red for values below -3.84, full transparency for values from -3.84 to 3.84 as they are non-significant and five different shades of blue for values greater than 3.84.

2.4 Case Study Data

This case study uses a sample of 946.955 posts from Instagram between 2011 (founding year) and October 2022, mined from the publicly available API endpoint.

For data mining, we used the location-based social media architecture by Dunkel, Löchner and Burghardt (2020) in combination with their respective privacy-aware HyperLogLog database (gitlab.vgiscience.de/lbsn/databases/hlldb/). Using predefined thematic HyperLogLog sets allowed us to create a deduplicated database of the social media data without needing to save the original raw data. The algorithm introduces an error rate of 2-4% (Dunkel et a., 2020) which is neglectable for the purpose of this study. The case study of Weckmüller (2022) presents a possible implementation of a dashboard based on this data structure and provides more insight why its practical usage is favorable for user privacy in live-applications. The thematic data excerpts on our GitHub repository (github.com/do-me/hexbins) are derived from such a HyperLogLog database.

As mentioned earlier, the functions always imply a random sample of data. Unfortunately, there is no method to verify whether this is actually the case or whether the data are skewed towards most popular posts or similar (Dunkel et al., 2023). The complete Instagram database is scarcely accessible, thereby rendering it nearly impossible to compare the mined data and conduct a test for random distribution. Dunkel et al. (2023) propose a simple a method to mitigate this limitation. By operating on a global scale and examining two topics, one can assess the ratio between the random mined sample and the global figures comprising all posts on the platform obtainable via Instagram's API. Although a similar ratio is necessary, it is still not sufficient for the assumption of a random distribution. In practice, this approach is only viable for a restricted number of use cases, specifically for comparing topics on a considerably large scale (e.g. global), but not at the municipal level.

Even though social media have significant usage biases, such a positivity bias (cf. Schreurs and Vandenbosch, 2021), studies show, that in some contexts they can serve as a good quantitative estimate (Wilkins et al., 2021). Still, one must be careful as not to perpetuate predominant spatial hegemonies and reproduce spatial inequality (Metcalf and Crawford, 2016). By misinterpreting data, further spatial exclusion of certain population cohorts would be a potential threat (boyd and Crawford, 2012), simply because excluded cohorts might not use the platform in the exclusive locations and hence have no locatable voice. Accessibility serves as a good example: if an urban park is only accessible via stairs, there might be no posts locatable within the park about missing accessibility.

As for the purpose of this case study, the focus lays on the methodological approach and less on the actual results. Still, the authors were able to verify some of the results due to their local knowledge indicating that the sample might have been random.

The methodology presented here is suitable to all georeferenced social media networks and could also mitigate above biases by combining data from more than one origin.

3. RESULTS

All results provided here can be reproduced with our public demo app (geo.rocks/hexbins) but might slightly differ due to initial random seeding of the hexagons.
The objective of this section is to provide a summary of our experience with different metrics and offer guidance on when to use each metric, where possible.

3.1 Absolute Values

By summing up all the posts of interest for each hexbin, the absolute values reveal the areas with the highest post density. The overview in Figure 1 shows the absolute distribution of all case study posts according to Equation (2), with a very clear center of gravity located in and around the city center of Bonn and Bad Godesberg as a lesser frequented district (DE: Stadtbezirk) of Bonn in the south.

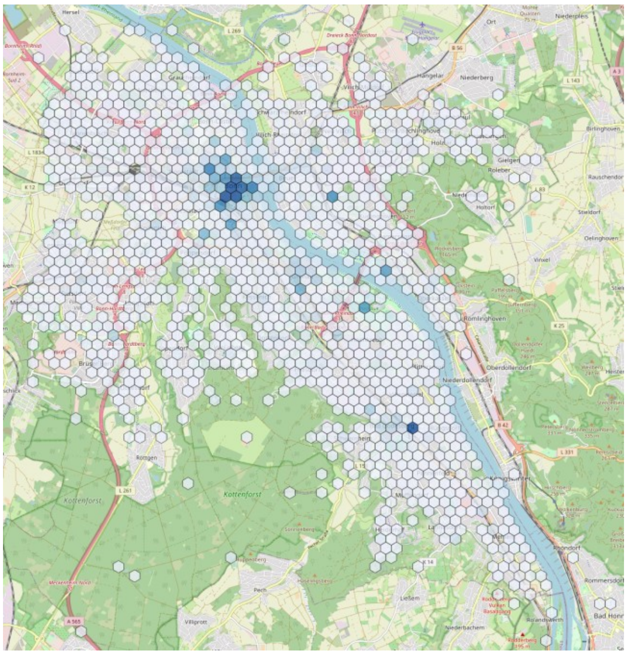

Figure 1. Study area with aggregated absolute values of all posts per hexbin, equal breaks. Darker blues equal more posts. Map data: © OpenStreetMap contributors.

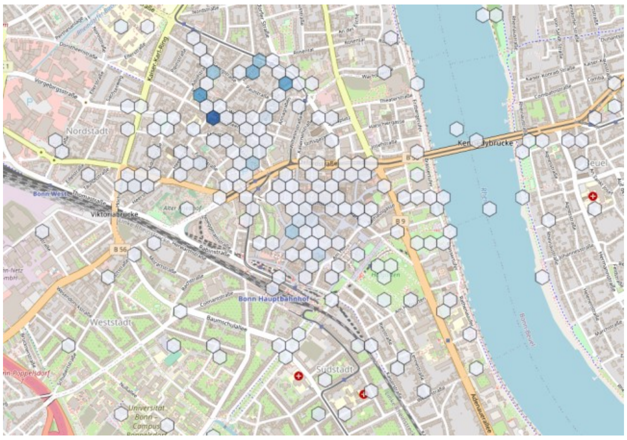

Figure 2. Aggregated absolute values of all posts per hexbin for “cherry blossom”, equal breaks. Map data: © OpenStreetMap contributors.

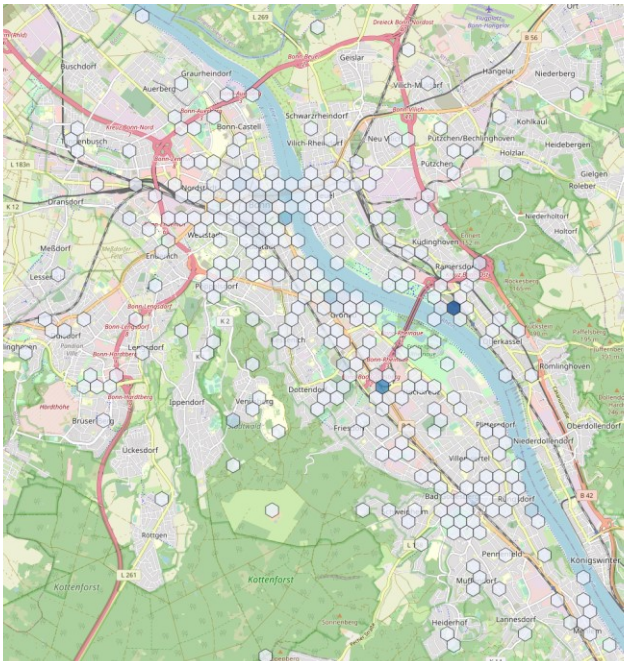

Figure 3. Aggregated absolute values of all posts per hexbin for “hotels”, equal breaks. Map data: © OpenStreetMap contributors.

Absolute metrics may be suitable for certain purposes where overall popularity is relevant. Seasonal occurrences, such as the cherry blossom in Bonn (Figure 2) can be identified through absolute values since people rarely post about them from other locations during the rest of the year. Moreover, business studies could be conducted as hotels (Figure 3) or restaurants may want to determine their social media ranking in comparison to their competitors. However, this is not universally applicable to other use cases or events. In general, absolute values are primarily suitable for comparing overall popularity. Beyond this specific application, absolute values should be used with caution.

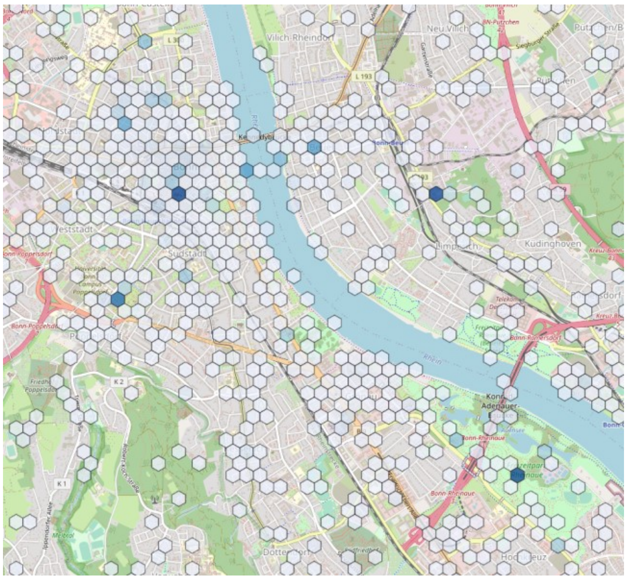

Figure 4. Aggregated absolute values of all posts per hexbin for “nature”, equal breaks. Map data: © OpenStreetMap contributors.

A significant limitation of the absolute metric becomes apparent when employing broader, more general terms. The term "nature" for instance is vague and not solely used for categorizing elements such as vegetation; it is frequently used in marketing and other contexts as well. In these cases, thematic posts tend to align with the overall trend of posts. As a result, the center of Bonn may have a high number of nature posts, even if there may be no apparent direct connection to nature.

The phenomenon can partially be explained with the ambiguous character of Instagram locations explained earlier. Users who post something tagging the "Bonn" location in the center of Bonn do not necessarily mean the center coordinates but rather the whole area of Bonn in general, referring to the cultural entirety of the city.

This effect occurs on any spatial level on Instagram but increases in severity on the local level. If for example, someone would create a "Germany" location with the coordinates of the Brandenburg Gate in Berlin and foreign tourists consistently use this location for geotagging posts throughout Germany, these posts would be inaccurately attributed to Berlin or, worse, the "Mitte" district. The larger the area these locations actually refer to, the greater the volume of posts they attract and the more significant the adverse impact on the absolute representation of posts in hexbins.

3.2 Mean Values

To mitigate the effect, a simple mean function (Equation (3)) can be used. Mean values represent the purity of a hexbin which is well-suited for cases where the question does not depend on the most popular locations overall.

As the previously mentioned meta-locations attract a wide range of diverse posts the mean values of thematic subsets of posts tend to behave like the overall average mean value. However, there is a rare case where a meta-location might still influence the results. For instance, if a hypothetical Germany location has a high percentage of posts about "beer" and a query might try to identify all breweries in Berlin Mitte, this would pose a problem.

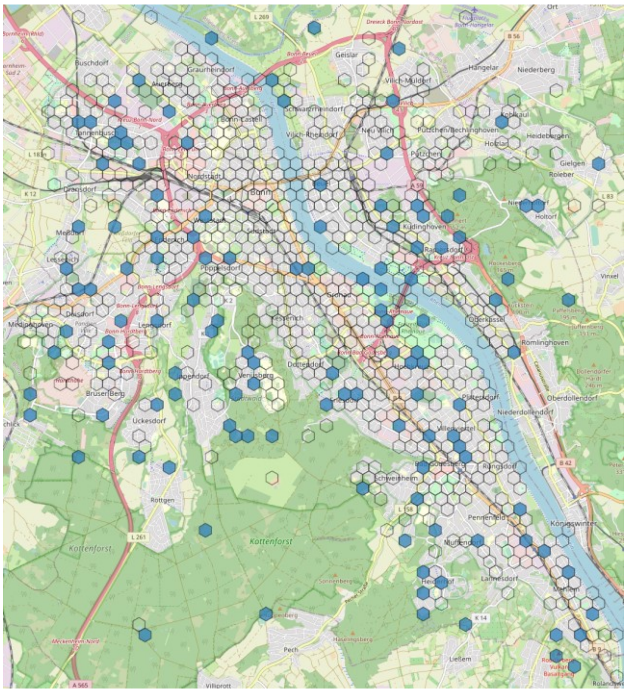

Figure 5. Hexbins with a mean greater than the average mean plus a margin of 0.1 for "nature". Map data: © OpenStreetMap contributors.

In contrast to Figure 4, Figure 5 shows an entirely different image for nature-related posts. As the center of Bonn does not yield a mean greater than the average plus a margin of 0.1, it is not displayed anymore.

The mean value is suited when trying to identify highly related locations to the query and the absolute number of posts is not of interest. For instance, when the subject in question is uniformly distributed throughout the area of interest, but does not yield many absolute results, like for the topic "ping-pong" in Bonn, mean values are suited as one would like to include locations with only one post too.

The function could be improved by allowing a user-defined minimum number of posts per location, so artifact locations with e.g. one post only could be excluded as they might not necessarily be significant and relevant to all queries

The basic $odds\ ratio$ (Equation (1)) behaves in a similar way and is suited to the same use cases.

3.3 Signed Chi

If neither the absolute frequency nor the purity is of sole interest, but one might need a metric that covers both, the signed chi (Equation (4)) function is suitable. It indicates not only the significancy of a location or hexbin within a certain confidence interval but also the over- and underrepresentation of thematic posts (positive/negative chi), compared to the average ratio for the whole area under investigation. In practice, there might be more use cases for the positive values as they indicate where something is significantly overrepresented. Dealing with nonexistence is typically methodologically challenging, but it may have its applications, such as the absence of party-related terms in posts suggesting potential areas of calm.

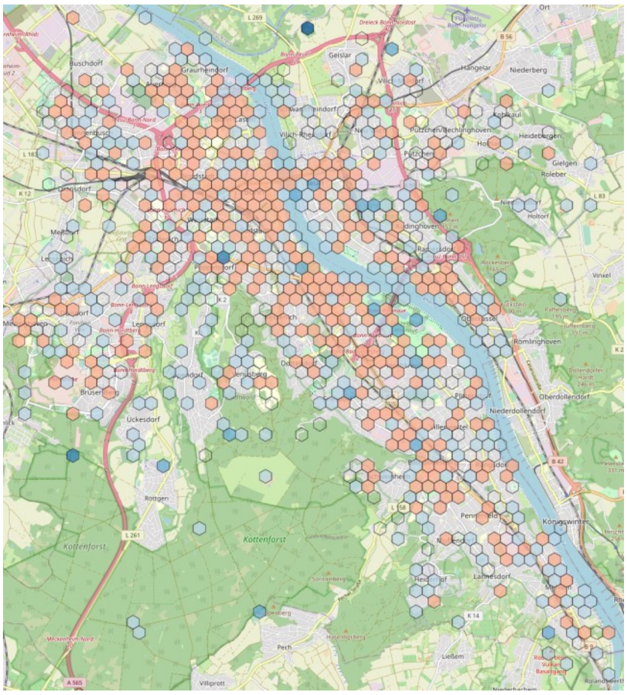

Figure 6. Hexbins based on signed chi values for "nature", equal breaks. Map data: © OpenStreetMap contributors.

Figure 6 shows that the three hexbins with the highest positive signed chi value lay indeed in areas with plenty of vegetation: a forest in the south-west (Kottenforst), the botanical garden (Poppelsdorf) close to the center and a natural reserve in the north (Siegaue). In this instance, not only is Bonn's center excluded from the hexbins with positive scores, but it also receives a negative score, indicating a potential lack of nature compared to the other hexbins.

Through the utilization of diverse color classification techniques, including the implementation of natural breaks instead of equal breaks, the identification of regions exhibiting elevated chi scores is facilitated, while also mitigating the influence of outlier locations characterized by exceptionally high chi values. Additionally, adjusting the radius proportionally to the number of topic posts enables the identification of hexbins with high scores of significance, thereby preserving a measure of general popularity (Figure 7).

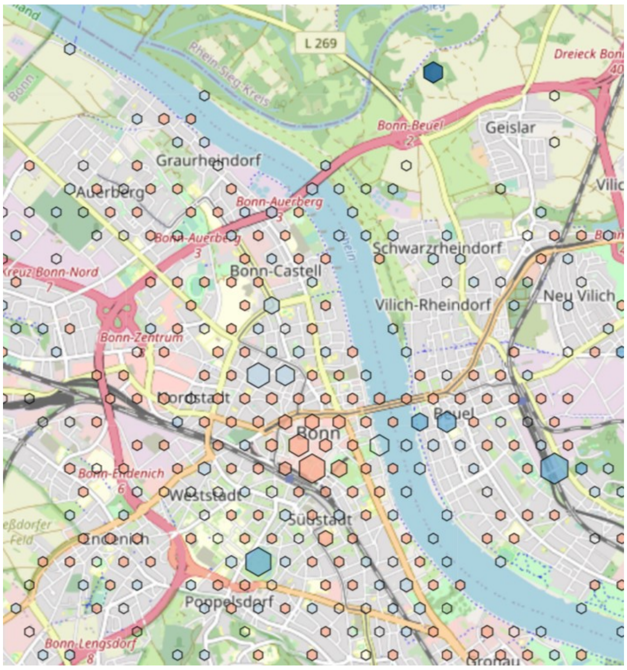

Figure 7. Hexbins based on signed chi values for “nature”, equal breaks, proportional radius, Bonn center. Map data: © OpenStreetMap contributors

4. DISCUSSION AND CONCLUSION

We propose a flexible approach consisting of different binning and color classification functions each suited to different use cases. We evaluate these metrics through a hexagonal on-the-fly binning approach with different color scaling and propose easily customizable scripts for the leaflet-d3 plugin.

Our findings suggest that all of the investigated metrics can offer insight into GSMD, but their appropriate use is highly dependent on the research question at hand, as demonstrated by the example of nature-related posts. The absolute metric should be used sparingly and is only appropriate when overall popularity is relevant, such as for competition analysis in specific sectors like hotels, restaurants, or similar businesses.

The relative metric is suitable for analysis when overall popularity is not a factor, and the focus is on the purity of the underlying binned locations. This is useful for identifying less frequented but highly specialized and evenly distributed locations, such as ping-pong tables or niche shops. However, locations with few posts may distort the overall result and lead to misinterpretation, as the number of posts is not reflected in the metric.

The signed chi metric addresses these shortcomings by reflecting significant values and is generally the least error-prone. It yields the best overall results for our queries so that we recommend the metric as a default for visualizations.

However, we acknowledge the various challenges of effectively binning GSMD that still require significant awareness.

Firstly, it is important to note that all of the discussed binning functions require either the entirety of data or random samples. However, due to the policies of some of the biggest commercial social networks, transparent and random sampling or querying may not be allowed or possible, which could introduce bias. Unless platform policies change, little can be done to address this issue.

Choosing the right color palette for data visualization has become easier with the availability of convenient online tools that account for color-blindness and perceptual accuracy. It is crucial for app users to be aware of different color classifications functions and their influence on the representation of data. Additionally, users should have the option to toggle between different color classifications to best understand the crucial differences. By comparing the different representations on a map, app users can develop a better understanding of the data. Researchers should provide guidelines on which metric and classification to use in different cases, as they have different purposes. To ensure that data visualizations are accurate and meaningful, it is important to prevent possible misinterpretation of data.

While testing our demo app, we encountered yet another visualization challenge for mean and chi values. Due to the relative nature of both functions, hexbins vary greatly in their value, depending on the variance of the underlying locations. If for example, on a lower zoom-level a hexbin contains three locations, two with very low mean values in the west and one with a high mean value in the east, the hexbin would likely have a mean below average and hence be indicated as red. When zooming in, the hexbin would then be split up in two separate hexbins, one with a very low mean in the west, and the eastern one with a very high mean, now becoming green. This phenomenon of color bouncing is rooted in the nature of all binning approaches and independent from the used geometry. However, it can be visually confusing for app users.

To avoid this effect, only the best location within a hexbin can be used for color classification, ignoring all others. This approach creates a consistent experience across all zoom-levels, which can be useful for finding locations with low post numbers in busy areas with plenty of posts. For instance, it is possible to identify specific restaurants in the city center with many irrelevant posts even at a low zoom level because the "noise" locations around them do not even out the high score of the matching location but instead are merely ignored.

We propose to investigate hexbins further by trying new metrics and color classifications (e.g. standard deviation and z-scores) depending on the user’s needs and by analyzing the capabilities of three-dimensional representation with a z-height as additional variable besides the radius and color.

In conclusion, our study lays the foundation for better suited GSMD hexagonal binning metrics and provides practical usage recommendations, as well as an open-source implementation.

5. REFERENCES

Abante, A. M. R. 2020. Risk hotspot conceptual space characterized by hexagonal data binning technique: an application. *International Journal of Computing Sciences Research*, 5(1), 550-567.

Abkenar, S. B., Kashani, M. H., Mahdipour, E., Jameii, S. M., 2021. Big data analytics meets social media: A systematic review of techniques, open issues, and future directions. *Telematics and Informatics*, 57, 101517.

Antonelli, F., Azzi, M., Balduini, M., Ciuccarelli, P., Valle, E. D., Larcher, R., 2014. City sensing: visualising mobile and social data about a city scale event. In *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces* (pp. 337-338).

Avvenuti, M., Cresci, S., La Polla, M. N., Meletti, C., Tesconi, M., 2017. Nowcasting of earthquake consequences using big social data. *IEEE Internet Computing*, 21(6), 37-45.

Barros, C., Moya-Gómez, B., Gutiérrez, J., 2020. Using geotagged photographs and GPS tracks from social networks to analyse visitor behaviour in national parks. *Current Issues in Tourism*, 23(10), 1291-1310.

boyd, D., Crawford, K., 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5), 662-679.

Brewer, C. A., 2016. *Designing Better Maps: A Guide for GIS Users*, 2nd ed. Esri Press.

Bruns, A., 2019. After the ‘APIcalypse’: Social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, 22(11), 1544-1566.

Carr, D. B., Olsen, A. R., White, D., 1992. Hexagon Mosaic Maps for Display of Univariate and Bivariate Geographical Data. *Cartography and Geographic Information Systems*, 19(4), 228–236.

Coalter, J. 2020. ColorBrewer 2.0 and the Rainbow: Using Color Tools to Choose Appropriate Color Schema for your Data Visualization. *Issues in Science and Technology Librarianship*, (94).

Cresci, S., Cimino, A., Dell’Orletta, F., Tesconi, M., 2015. Crisis mapping during natural disasters via text analysis of social media messages. In *Web Information Systems Engineering–WISE 16th International Conference* (pp. 250-258). Springer International Publishing.

Dalvi, N., Olteanu, M., Raghavan, M., Bohannon, P., 2014. Deduplicating a places database. In *Proceedings of the 23rd international conference on world wide web* (pp. 409-418).

Dunkel, A., Hartmann, M. C., Hauthal, E., Burghardt, D., Purves, R. S., 2023. From sunrise to sunset: Exploring landscape preference through global reactions to ephemeral events captured in georeferenced social media. *Plos one*, 18(2), e0280423.

Dunkel, A., Löchner, M., Burghardt, D., 2020. Privacy-Aware Visualization of Volunteered Geographic Information (VGI) to Analyze Spatial Activity: A Benchmark Implementation. *ISPRS International Journal of Geo-Information*, 9(10), 1-21.

Giglio, S., Bertacchini, F., Bilotta, E., Pantano, P., 2019. Using social media to identify tourism attractiveness in six Italian cities. *Tourism management*, 72, 306-312.

Heikinheimo, V., Di Minin, E., Tenkanen, H., Hausmann, A., Erkkonen, J., Toivonen, T., 2017. User-generated geographic information for visitor monitoring in a national park: A comparison of social media data and visitor survey. *ISPRS International Journal of Geo-Information*, 6(3), 85.

Heikinheimo, V., Tenkanen, H., Bergroth, C., Järv, O., Hiippala, T., Toivonen, T., 2020. Understanding the use of urban green spaces from user-generated geographic information. *Landscape and Urban Planning*, 201, 103845.

Hu, J., Wang, Y., Li, P., 2017. Online city-scale hyper-local event detection via analysis of social media and human mobility. In IEEE International Conference on Big Data (pp. 626-635). IEEE.

Ilieva, R. T., McPhearson, T., 2018. Social-media data for urban sustainability. *Nature Sustainability*, 1(10), 553-565.

Internet Engineering Task Force, 2016. The GeoJSON Format (RFC 7946). IETF.

Jiang, B., 2013. Head/Tail Breaks: A New Classification Scheme for Data with a Heavy-Tailed Distribution, *The Professional Geographer*, 65:3, 482-494.

Jing, C., Du, M., Li, S., Liu, S., 2019. Geospatial dashboards for monitoring smart city performance. *Sustainability*, 11(20), 5648.

Kankanamge, N., Yigitcanlar, T., Goonetilleke, A., Kamruzzaman, M., 2020. Determining disaster severity through social media analysis: Testing the methodology with South East Queensland Flood tweets. *International journal of disaster risk reduction*, 42, 101360.

Kitchin, R., 2016. The ethics of smart cities and urban science. Philosophical Transactions of the Royal Society. *Mathematical, Physical and Engineering Sciences*, A374(2083), 20160115.

Kulawiak, M., Kulawiak, N., Sulima, M., Sikorska, K., 2022. A novel architecture of Web-GIS for mapping and analysis of echinococcosis in Poland. *Applied Geomatics*, 14(2), 181-198.

Levine, A. S., Feinholz, C. L., 2015. Participatory GIS to inform coral reef ecosystem management: Mapping human coastal and ocean uses in Hawaii. *Applied Geography*, 59, 60-69.

Lewin-Koh, N., 2021. Hexagon binning: An overview. cran.r-project.org/web/packages/hexbin/vignettes/hexagon_binning.pdf (23 April 2023).

Li, M., McGrath, H., Stefanakis, E., 2022. Multi-Scale Flood Mapping under Climate Change Scenarios in Hexagonal Discrete Global Grids. *ISPRS International Journal of Geo-Information*, 11(12), 627.

Li, Z., Huang, X., Ye, X., Jiang, Y., Martin, Y., Ning, H., Hodgson, E., Li, X., 2021. Measuring global multi-scale place connectivity using geotagged social media data. *Scientific Reports*, 11(1), 1-19.

Metcalf, J., Crawford, K., 2016. Where are human subjects in Big Data research? The emerging ethics divide. *Big Data & Society*, 3(1), 1-14.

Oyana, T. J., Scott, K. E., 2008. A geospatial implementation of a novel delineation clustering algorithm employing the k-means. *The European Information Society: Taking Geoinformation Science One Step Further*, 135-157.

Poorthuis, A., Zook, M., 2015. Small stories in big data: Gaining insights from large spatial point pattern datasets. *Cityscape*, 17(1), 151-160.

Poorthuis, A., Power, D., Zook, M., 2020. Attentional social media: Mapping the spaces and networks of the fashion industry. *Annals of the American Association of Geographers*, 110(4), 941-966.

Power, D., Zook, M., Poorthuis, A., 2015. *The world tweets Norway: The Norwegian music and fashion industry in global social media*. Knowledge Works: the Norwegian National Centre for Cultural Industries.

Rickson, C., 2023. *Investigating the Influence of Tessellate Shapes on Crime Hot Spot Mapping Results: Do Hexagonal Grid Thematic Maps Outperform Conventional Thematic Maps?*. The University of Tampa.

Rock, A., Taber, J., 2020. Home Tweet Home: Can social media define a community?. *Journal of Appalachian Studies*, 26(1), 87-105.

Schreurs, L., Vandenbosch, L., 2021. Introducing the Social Media Literacy (SMILE) model with the case of the positivity bias on social media. *Journal of Children and Media*, 15(3), 320-337.

Shelton, T., Poorthuis, A., Zook, M., 2015. Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information. *Landscape and urban planning*, 142, 198-211.

Sinclair, M., Mayer, M., Woltering, M., Ghermandi, A., 2020. Using social media to estimate visitor provenance and patterns of recreation in Germany's national parks. *Journal of Environmental Management*, 263, 110418.

Sui, D., Goodchild, M., 2011. The convergence of GIS and social media: challenges for GIScience. *International journal of geographical information science*, 25(11), 1737-1748.

Teles Da Mota, V., Pickering, C., 2020. Using social media to assess nature-based tourism: Current research and future trends. *Journal of Outdoor Recreation and Tourism*, 30, 100295.

Toivonen, T., Heikinheimo, V., Fink, C., Hausmann, A., Hiippala, T., Järv, O., Tenkanen, H., Di Minin, E., 2019. Social media data for conservation science: A methodological overview. *Biological Conservation*, 233, 298-315.

Visvalingam, M., 1978. The signed chi-square measure for mapping. *The Cartographic Journal*, 15(2), 93-98.

Weckmüller, D., Dunkel, A., 2022. Developing a Privacy-Aware Map-Based Cross-Platform Social Media Dashboard for Municipal Decision-Making. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48.

Wilkins, E. J., Howe, P. D., Smith, J. W., 2021. Social media reveal ecoregional variation in how weather influences visitor behavior in US National Park Service units. *Scientific Reports*, 11(1), 2403.

Wood, J., Dykes, J., Slingsby, A., Clarke, K., 2007. Interactive visual exploration of a large spatio-temporal dataset: Reflections on a geovisualization mashup. *IEEE transactions on visualization and computer graphics*, 13(6), 1176-1183.

Yao, F., Wang, Y., 2020. Towards resilient and smart cities: A real-time urban analytical and geo-visual system for social media streaming data. *Sustainable Cities and Society*, 63, 102448.

Yao, X., Yu, G., Li, G., Yan, S., Zhao, L., Zhu, D., 2023. HexTile: A Hexagonal DGGS-Based Map Tile Algorithm for Visualizing Big Remote Sensing Data in Spark. *ISPRS International Journal of Geo-Information*, 12(3), 89.

Zhang, H., van Berkel, D., Howe, P. D., Miller, Z. D., Smith, J. W., 2021. Using social media to measure and map visitation to public lands in Utah. *Applied Geography*, 128, 102389.