Effect of color palettes in heatmaps perception: a study

E. Molina and C. Middel and P. Vázquez

ViRVIG Group
Universitat Politècnica de Catalunya

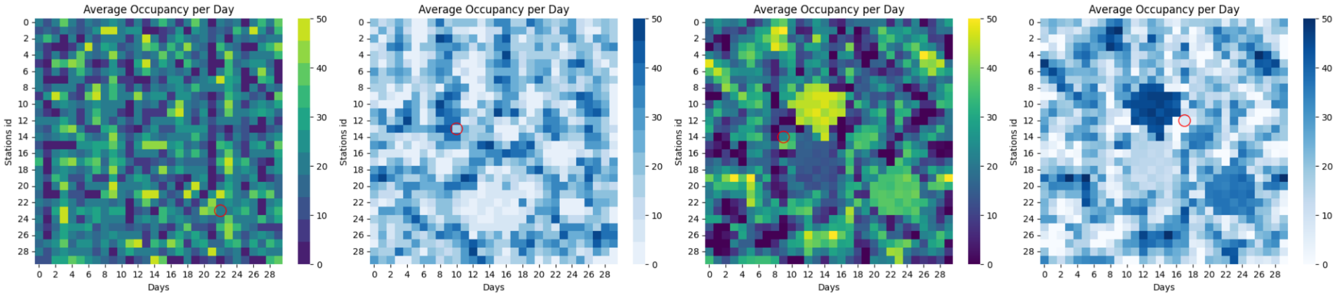

Figure 1: *Example of heatmaps used in our study. In order: Viridis and Blues discrete, and Viridis and Blues continuous.*

Abstract

Heatmaps are a widely used technique in visualization. Unfortunately, they have not been investigated in depth and little is known about the best parameterizations so that they are properly interpreted. The effect of different palettes on our ability to read values is still unknown. To address this issue, we conducted a user study, in which we analyzed the effect of two commonly used color palettes, Blues and Viridis, on value estimation and value search. As a result, we provide some suggestions for what to expect from the heatmap configurations analyzed.

CCS Concepts

• ***Human-centered computing** → **Heat maps; Visualization design and evaluation methods;***

1. Introduction

Heatmaps are a popular visualization technique that can be used to detect clusters, outliers, and summarize data [Mun14]. They are designed as a two-dimensional matrix of cells where two categorical (or sometimes ordered) variables are used as indices, and a quantitative variable is encoded as a color. Though very commonly used, to the best of our knowledge, there are no guidelines that advise on how to design them. Existing studies on how good people's estimation of values is when reading heatmaps depend on different visual cues are scarce. With our study, we present insights into value estimation with different palettes, and provide a more in-depth understanding of how the ranges of values in the palettes affect, which can be a starting point for new investigations.

2. Related Work

Heatmaps are very common in visualization nowadays [KAB*20]. They can be traced back to 1957, when Sneath already talked about reordering rows and columns of matrices that represented similarity values as shades [Sne57] to help users find patterns. Bertin describes them under the name of reorderable matrices [Ber73]. In the approach depicted by Wilkinson [Wil12], oftentimes the two keys can be reordered, as with clustered heatmaps [Mun14, DTT*15]. However, heatmaps also encode time data [CSL*15, KIM*16] that cannot be reordered. Calendar heatmaps are a variant where the cells represent days in a calendar (e.g., [LWW*20]) and the order is fixed. Under the name of heatmaps, we find many other designs: using hexagonal cells [CLNL87] (though commonly called hexplots [TSSB22]), or maps where the axes are linked to spatial positions [PSSC16, WSWW13].

Perception studies are needed to understand how humans interpret charts. There is a large body of research [FPS*21], but there are still unknowns, such as the effect of the palettes [LH18]. Furthermore, perceptual studies are often limited due to the enormity of the space of parameters, so we must constrain the conditions to effectively obtain statistically valid results. Besides, there are

DOI: 10.2312/evs.20231039

some practices that are widespread (such as the use of red-green palettes) despite not being the most suitable according to science [BGP*11]. Unfortunately, there are also contradictory results, rejecting a palette [RT98] and recommending it [RSGP21]. Regarding heatmaps, some focus on comparing them with other types of visual encoding techniques [GFC05], but just a few papers have analyzed how they are perceived, as Słomska-Przech et al. do with respect to the generalization level [SPPP21] or as Rostislav et al. do with geographical information [NPS18]. However, the rest mostly deal with comparative visualization tasks. Krakov and Feitelson analyze how to better encode differences between hexplots [KF13], but no perceptual analysis is performed. Kraus et al. also concentrate on visual comparison tasks for heatmaps and height maps in 3D environments [KAB*20]. But the tasks they address don't involve estimating values or changing parameters like the palette. Our work has some analogies to the one by Tory et al., where the authors compare how points, encoded in a sequential palette of greens, are read by users under different distributions in scatterplots [TSD09]. Trautner et al. have analyzed the perception of different configurations of honeycomb plots, which resemble heatmaps [TSSB22].

3. Experiment Design

Our objective is to investigate the relationship between heatmap configurations and the perception of values. To identify the most common configurations used by researchers and practitioners, we first analyzed a set of sources.

3.1. Common heatmap parameters

The initial goal was to find whether there were some configuration parameters that we could use as a basis for the design of our tests. We first downloaded the EuroVis and IEEE Vis papers from 2019 and 2020 (263). 75 had some kind of heatmap or heatmap-like chart with color coding that showed confusion matrices. The tasks the designers were trying to help solve were mostly related to the exploration of correlation matrices or the search for patterns, shapes, and trends. We also checked the more common usage in non-scientific publications and searched "heatmap chart" in Google Images. We checked the primary results (122 images). 114 were heatmaps or heatmap-like charts. With this analysis, summarized in Table 1, we saw that there are no common parameters to start with. In fact, there are good practices that need to become more widespread, like avoiding red-green palettes, to make them accessible to color-blind people. The lack of consensus on the best design led us to fix a set of parameters (cell size and dimension) and vary the palette. The proposed tasks were:

- **Task 1:** Estimation of the value of the indicated cell. Its goal is to check whether the values are estimated correctly and which parameters may have an impact.
- **Task 2:** Select a cell with the indicated value. The objective here is to study if values are properly found.

Due to the massive space of possible configurations, we leave out of this work other elements such as clusters, outliers, and trends detection/interpretation.

	Papers	**Google**
Dimensions	From 4x4 to hundreds	12-35 up to 50
Cell size	1px, majority >10 px	Big, rectangular
Palette	Sequential: blue/green. Multi-hue: viridis/magma (red-green)	Sequential: blue/orange Multi-hue: red-green.

Table 1: *More common parameters found in the EuroVis and IEEE Vis 2019/20 papers and in Google Images.*

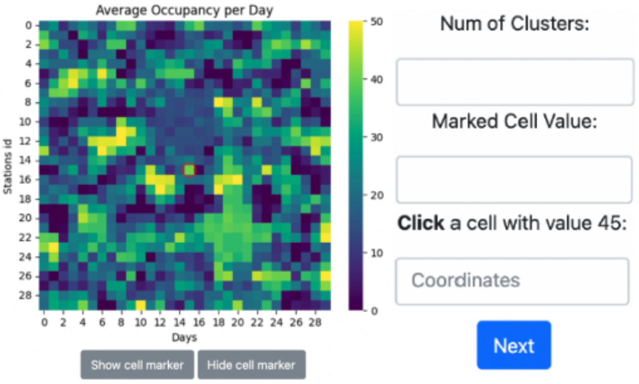

Figure 2: *Example of a heatmap with the different tasks to be solved. Participants had the fields to be filled under the chart.*

3.2. Hypotheses

We restricted ourselves to changes in palettes: single-hue vs multi-hue ones, and discrete vs continuous ones, since both are found in the literature. Since continuous palettes encode more values and multi-hue palettes use a combination of color hue and luminance, we hypothesized that:

- Continuous palettes are better than discrete ones to: *a)* **H1:** estimate values and *b)* **H3:** search for a given value.
- Multi-hue palettes are better than single hue ones to: *c)* **H2:** estimate values and *d)* **H4:** search for a given value.

Hypotheses 1 and 2 are related to Task 1 and the others to Task 2. We also analyze the ranges of the palettes used to encode values and the time to answer.

3.3. Charts Design

This paper is part of a larger study, where we analyze other variables such as cluster counting (not included here due to the lack of space). Therefore, the data was generated synthetically to ensure a certain number of clusters (0 to 3) in each heatmap. To make the data relatable, we informed the participants that the heatmaps were encoding the number of bicycles per station (Y-axis) per day (X-axis). The dimensions selected were 30 × 30. Despite this, we later saw that participants tended to ignore the meaning of the axes.

Concerning ***resolution***, to ensure that all fit on any screen, we analyzed several cell sizes and layouts, keeping a total size of around

500px. We finally opted for a square design as shown in Figure 2, but with the widgets on the bottom. Charts were created with each cell having a size large enough to be easily seen (13×13 px).

Regarding ***palettes***, since blue colors predominated, we opted to use the Blues as a sequential palette and Viridis for the multi-hue. Note that these two palettes also stand out in previous experiments [LH18]. The final combinations selected were: **BluesCont:** Blues continuous (BC), **BluesDisc:** Blues discrete (BD), **ViridisCont:** Viridis continuous (VC), and **ViridisDisc:** Viridis discrete (VD). We then used the `https://colorbrewer2.org/` tool to decide which color (red) would be the most suitable to highlight a cell with a ring, needed for Task 1. Finally, we created 34 charts, as the ones in Figure 1.

3.4. Structure

The study was designed as a web application, delivered through Prolific [pro] with the following steps: First, the objectives were presented and the heatmaps and tasks were described. Then, a demographic survey (age, country, gender, education, eyesight, and screen size) was conducted. The training tasks, which required the participant to solve three heatmaps, were followed by the actual tasks, which were randomized. Lastly, a questionnaire to evaluate understanding and satisfaction. The demographic survey was later used to identify non-suitable participants (e.g., poor eyesight).

3.5. Participants

We gathered 50 participants (worldwide, English speaking and 50% gender balance) through Prolific and 53 extra participants by announcing the study through social networks. The duration was estimated to be about 20 minutes, with two pilot participants. The amount considered a fair wage by Prolific was £7.5 per hour. To ensure the ecological validity of the data, participants who displayed no understanding or carelessness (average errors larger than 20, from a 1-50 range) during the training phase were not allowed to continue. We also included duplicated tests for double-checking click-through strategies. For the actual tasks, participants exhibiting too large errors were interpreted as click-through and cleaned. This led us to keep 85 valid participants (47 and 38, respectively).

These 85 participants (40 female) were from 8 different countries. Age ranges: 69 in the range 18-35, 10 in 35-50, and 6 above 50. Concerning education, 15 had a high school degree and 53 had higher education. None of them were color-blind. 79 participants declared they understood the experiment, 61 were satisfied with the performance, and 15 were *undecided*. 24 found the tasks not easy.

4. Results

The data analysis was performed by comparing the errors in each task using Repeated Measures ANOVA [Gir92], followed by the Bonferroni posthoc test [Bon36]. Next, we discuss the significant results, also illustrated in Figure 3. Liu and Heer found differences when comparing values in Blues in low ranges [LH18]. Participants had to check which of two colors was more similar to a reference one. In our case, we wondered whether the Blues palettes exhibited larger errors at lower ranks. Thus, we partitioned the data in ***ranges***: small (S) 0-16, medium (M) 17-33, and large (L) 34-50. Since the values were selected randomly for the experiment, there is a different number of samples per range. For each combination, we use the maximum number of common answers, indicated in parentheses. The results are summarized in Table 2 and also illustrated in 3 with the p-values displayed.

4.1. Task 1: Estimating cell values

Participants had to write the correct value of a circled cell. This allows us to evaluate H1 and H2.

For **H1: Continuous vs discrete** and **H2: Multi-hue vs single hue** significant differences were found. Best ones: Discrete and Multi-hue. Thus, **we reject H1** and **accept H2**. Hence, between individual palettes, we could expect the best to be ViridisDisc. However, ViridisCont is the best one. This result is in line with our prediction but needs more investigation. Top left in Figure 3.

Same palette, different *ranges* There are only differences in BluesDisc (79 samples) between S and L, being S better. Middle left in Figure 3.

Same range, different *palette* S (79 samples): differences between both Viridis and both Blues. The best ones, in order, ViridisCont and BluesDisc. M (67): Same as in S. L (68): Both Continuous are different, and ViridisDisc is different from the other Discrete and the other Viridis. ViridisCont is also the best one. In general, with Blues, participants tend to underestimate, and with Viridis, to overestimate. Middle in Figure 3.

4.2. Task 2: Searching given values

For **H3: Continuous vs discrete** we found significant differences (p-value 0.045). Both underestimate, but continuous palettes are more accurate. Therefore, we **accept H3**. No differences were found for **H4: Multi-hue vs single hue**, so it **cannot be accepted**.

Same palette, different *ranges* BlueCont (81): Differences between L and the others, being the worst. BlueDisc (73): Differences between all, being M the best. ViridisCont (83): Differences between all, S is the best. ViridisDisc (83): Same as in BlueDisc. It seems that the almost white colors in Blues are difficult to distinguish and that the variation between greenish and yellow tones in Viridis favors the distinction. Right in Figure 3.

Same range, different *palette* There are only differences in M (83) between BluesCont and the others, being the best one. Bottom left in Figure 3.

5. Conclusions and Future work

We have analyzed the effect of the palette on estimating values in heatmaps. The significant results are summarized in Table 2.

The main **takeaways** are: With the task of estimating values, when we analyze palettes 2 vs 2, our intuition about the possible superiority of continuous palettes for having more values is rejected. However, when we analyze the results of the 4 possibilities, the combination of Multi-Hue and Continuous confirms our intuitions by showing the best individual results in VC. Individually, the

Hypothesis testing				
	Task 1		Task 2	
palettes	C vs D	V vs B	C vs D	V vs B
C		NA	✓	NA
D	✓	NA		NA
V	NA	✓	NA	-
B	NA		NA	-

	Task 1: Estimate value of cell							Task 2: Select cell with value						
		same palette			same range				same palette			same range		
palettes	All	S	M	L	S	M	L	All	S	M	L	S	M	L
VC	✓	-	-	-	✓	✓	✓	-	✓			-		-
VD		-	-	-				-		✓		-		-
BC		-	-	-				-	✓	✓		-	✓	-
BD		✓	-		✓	✓	✓	-		✓		-		-

Table 2: *Summary of the study results. C= Continuous, D= Discrete, V= Viridis, and B= Blues. When testing Task 1, D and V palettes perform better (left table), so we would expect VD to be the best palette, but VC is the one that performs best when treated as 4 different palettes, right table. In Task 2, C palettes perform better, as expected, but there is no difference between individual palettes. Within the same range, right table, in Task 1 VC and BD are better at estimating values in all ranges, with VC exhibiting a slightly better behavior. In Task 2, within the same palette, we see that the ranges with the best results are the M/S, which makes sense, being the ones where there is a greater variation between the two ends of the palette. A ✓ means the best result, a – means no significant differences and NA means not applicable*

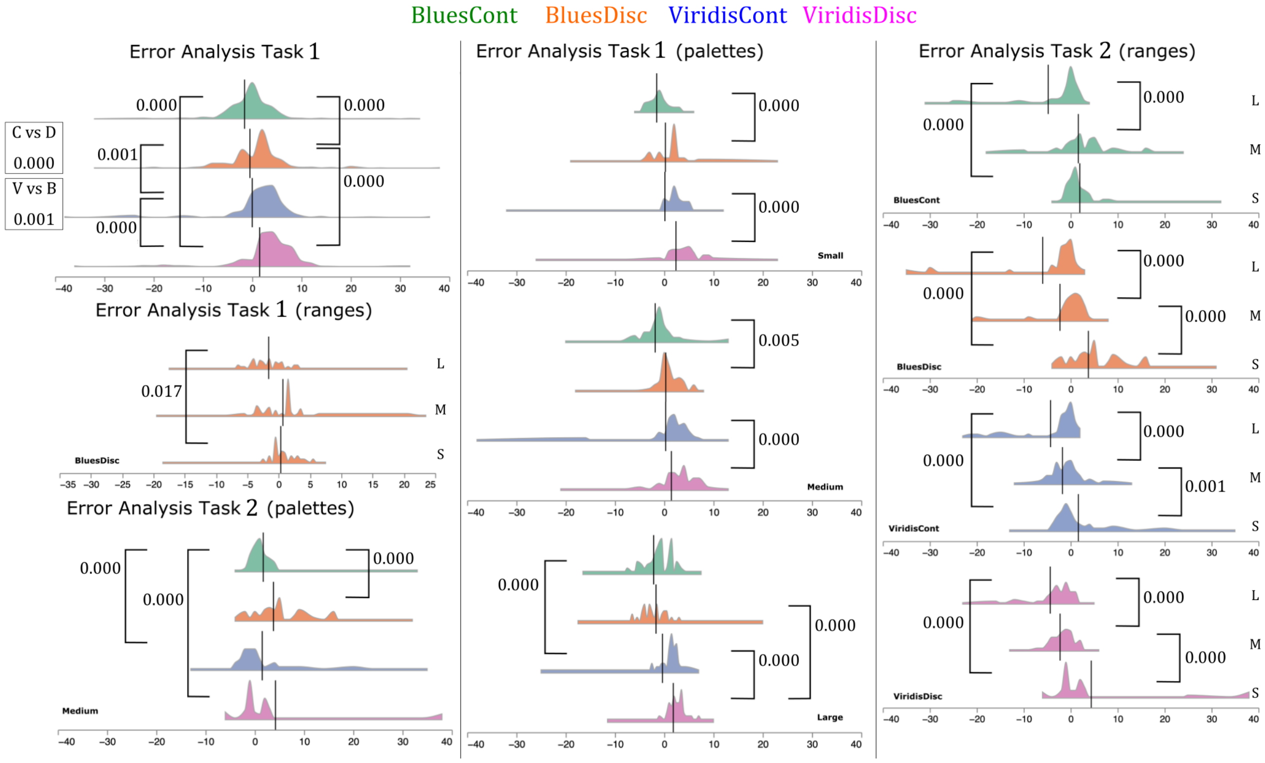

Figure 3: *Error distribution, X axis = answer given - correct value, Y axis = number of answers. Vertical lines = averages. C= Continuous, D= Discrete, V= Viridis, and B= Blues. P-values are shown next to each comparison. Left: Top: analysis H1 & H2 between individual palettes. Middle: Task 1 BluesDisc per range Small(S), Medium(M) & Large(L). Bottom: Task 2 Medium per palette. Middle: Task 1 the 3 ranges per palette. Right: Task 2 the 4 palettes per range.*

palettes do not show differences, but taking into account the ranges the BC palette works best, so it seems the most indicated. Although it must be taken into account that reading values in high ranges gives worse results than in the middle and low ranges. However, it is true that in this task the participants had more freedom by not having the cell to be evaluated restricted. We did not obtain any significant difference between time error per palette. We would suggest using VC since it gives the best results for Task 1 and in Task 2 there is none that dominates.

In the future, besides cluster counting, we want to analyze the effect of other parameters, such as cell size. Additionally, we want to study if there is a correlation between time and error per task.

Acknowledgments

This paper has been supported by PID2021-122136OB-C21 from the Ministerio de Ciencia e Innovación, by 839 FEDER (EU) funds. Elena Molina has been supported by FI-SDUR doctoral grant from Generalitat de Catalunya, and FPU grant from the Ministerio de Ciencia e Innovación.

References

[Ber73] Bertin J.: *Sémiologie graphique: Les diagrammes-Les réseaux-Les cartes*. Tech. rep., Gauthier-VillarsMouton & Cie, 1973. 1

[BGP*11] Borkin M., Gajos K., Peters A., Mitsouras D., Melchionna S., Rybicki F., Feldman C., Pfister H.: Evaluation of artery visualizations for heart disease diagnosis. *IEEE transactions on visualization and computer graphics 17*, 12 (2011), 2479–2488. 2

[Bon36] Bonferroni C.: *Teoria statistica delle classi e calcolo delle probabilità*. Pubblicazioni del R. Istituto superiore di scienze economiche e commerciali di Firenze. Seeber, 1936. URL: https://books.google.de/books?id=3CY-HQAACAAJ. 3

[CLNL87] Carr D. B., Littlefield R. J., Nicholson W., Littlefield J.: Scatterplot matrix techniques for large n. *Journal of the American Statistical Association 82*, 398 (1987), 424–436. 1

[CSL*15] Cao N., Shi C., Lin S., Lu J., Lin Y.-R., Lin C.-Y.: Targetvue: Visual analysis of anomalous user behaviors in online communication systems. *IEEE transactions on visualization and computer graphics 22*, 1 (2015), 280–289. 1

[DTT*15] Dehingia M., Talukdar N. C., Talukdar R., Reddy N., Mande S. S., Deka M., Khan M. R., et al.: Gut bacterial diversity of the tribes of india and comparison with the worldwide data. *Scientific reports 5*, 1 (2015), 1–12. 1

[FPS*21] Franconeri S. L., Padilla L. M., Shah P., Zacks J. M., Hullman J.: The science of visual data communication: What works. *Psychological Science in the Public Interest 22*, 3 (2021), 110–161. 1

[GFC05] Ghoniem M., Fekete J.-D., Castagliola P.: On the readability of graphs using node-link and matrix-based representations: A controlled experiment and statistical analysis. *Information Visualization 4*, 2 (2005), 114–135. URL: https://doi.org/10.1057/palgrave.ivs.9500092, arXiv:https://doi.org/10.1057/palgrave.ivs.9500092, doi:10.1057/palgrave.ivs.9500092. 2

[Gir92] Girden E. R.: *ANOVA: Repeated measures*. No. 84. sage, 1992. 3

[KAB*20] Kraus M., Angerbauer K., Buchmüller J., Schweitzer D., Keim D. A., Sedlmair M., Fuchs J.: Assessing 2d and 3d heatmaps for comparative analysis: An empirical study. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020), pp. 1–14. 1, 2

[KF13] Krakov D., Feitelson D. G.: Comparing performance heatmaps. In *Workshop on Job Scheduling Strategies for Parallel Processing* (2013), Springer, pp. 42–61. 2

[KIM*16] Kumatani S., Itoh T., Motohashi Y., Umezu K., Takatsuka M.: Time-varying data visualization using clustered heatmap and dual scatterplots. In *2016 20th International Conference Information Visualisation (IV)* (2016), IEEE, pp. 63–68. 1

[LH18] Liu Y., Heer J.: Somewhere over the rainbow: An empirical assessment of quantitative colormaps. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (2018), pp. 1–12. 1, 3

[LWW*20] Lin Y., Wong K., Wang Y., Zhang R., Dong B., Qu H., Zheng Q.: Taxthemis: Interactive mining and exploration of suspicious tax evasion groups. *IEEE Transactions on Visualization and Computer Graphics* (2020). 1

[Mun14] Munzner T.: *Visualization analysis and design*. CRC press, 2014. 1

[NPS18] Netek R., Pour T., Slezakova R.: Implementation of heat maps in geographical information system – exploratory study on traffic accident data. *Open Geosciences 10*, 1 (2018), 367–384. URL: https://doi.org/10.1515/geo-2018-0029 [cited 2023-03-01], doi:doi:10.1515/geo-2018-0029. 2

[pro] Prolific · quickly find research participants you can trust. https://www.prolific.co/. URL: https://www.prolific.co/. 3

[PSSC16] Pahins C. A., Stephens S. A., Scheidegger C., Comba J. L.: Hashedcubes: Simple, low memory, real-time visual exploration of big data. *IEEE transactions on visualization and computer graphics 23*, 1 (2016), 671–680. 1

[RSGP21] Reda K., Salvi A. A., Gray J., Papka M.: Color nameability predicts inference accuracy in spatial visualizations. 2

[RT98] Rogowitz B. E., Treinish L. A.: Data visualization: the end of the rainbow. *IEEE spectrum 35*, 12 (1998), 52–59. 2

[Sne57] Sneath P. H.: The application of computers to taxonomy. *Microbiology 17*, 1 (1957), 201–226. 1

[SPPP21] Słomska-Przech K., Panecki T., Pokojski W.: Heat maps: Perfect maps for quick reading? comparing usability of heat maps with different levels of generalization. *ISPRS International Journal of Geo-Information 10*, 8 (2021). URL: https://www.mdpi.com/2220-9964/10/8/562, doi:10.3390/ijgi10080562. 2

[TSD09] Tory M., Swindells C., Dreezer R.: Comparing dot and landscape spatializations for visual memory differences. *IEEE transactions on visualization and computer graphics 15*, 6 (2009), 1033–1040. 2

[TSSB22] Trautner T., Sbardellati M., Stoppel S., Bruckner S.: Honeycomb Plots: Visual Enhancements for Hexagonal Maps. In *Vision, Modeling, and Visualization* (2022), Bender J., Botsch M., Keim D. A., (Eds.), The Eurographics Association. doi:10.2312/vmv.20221205. 1, 2

[Wil12] Wilkinson L.: The grammar of graphics. In *Handbook of computational statistics*. Springer, 2012, pp. 375–414. 1

[WSWW13] Willems N., Scheepens R., Wetering H. v. d., Wijk J. J. v.: Visualization of vessel traffic. In *Situation Awareness with Systems of Systems*. Springer, 2013, pp. 73–87. 1

Proceedings published by Eurographics - The European Association for Computer Graphics.