FR-glyphs for Multidimensional Categorical Data

D. C. Canlon[1], F. Paulovich[1], and M. Tennekes[2]

[1]TU Eindhoven, Eindhoven, The Netherlands
[2]Statistics Netherlands, Heerlen, The Netherlands

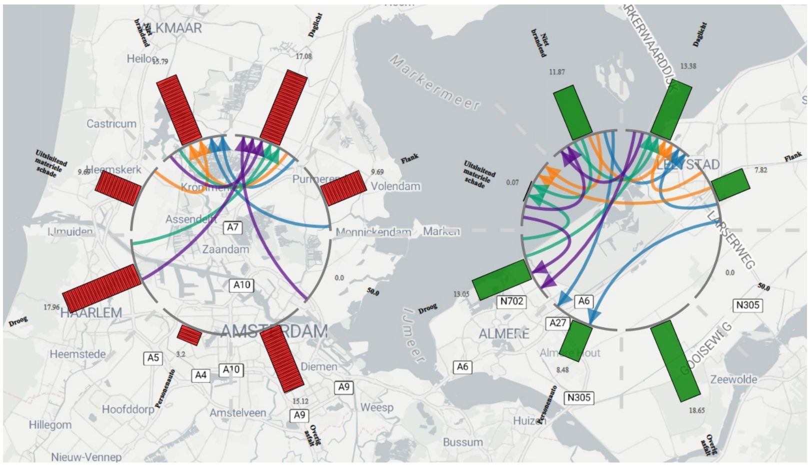

Figure 1: *Example of two frequency-relation-glyphs. The bars show the surprise for every attribute, where red indicates attributes with higher frequency than expected, and green indicates the opposite. The arrows in the middle show the relations between combinations of attributes. The arrows with the same colors link attributes belonging to the same (co-occurrence) relationship.*

Abstract

Multivariate categorical data analysis is challenging, especially when geographical information is present. Despite the widespread existence of such datasets, the current visualization solutions only typically represent frequencies of attributes, which can be misleading if uncorrelated attributes exist. We present the frequency-relation-glyphs, or FR-glyphs, as an alternative solution for these issues. FR-glyphs can (1) show deviations in the attribute's frequencies and (2) relations between combined sets of attributes. Furthermore, they can be added to geographical maps to compare multiple regions, such as provinces. We used the Bestand geRegistreerde Ongevallen in Nederland (BRON) dataset, which includes bicycle incidents, to show the usefulness of the FR-glyphs and evaluate them with stakeholders.

CCS Concepts

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

1. Introduction

One of the biggest challenges in multivariate data analysis [BB19] is finding human interpretable patterns [DKŽ*13a], especially when all data attributes (the variable values) are categorical, given that, in this case, nominal data lacks intrinsic value [PX08]. The analysis becomes even more challenging when such data is geo-located, and the analytical tasks involve comparing multiple geographical regions, such as provinces or municipalities. The challenge, then, is that the visualization for each region should present the same visual structure so visual comparison is possible.

Current methods to analyze multidimensional categorical data often use dimensionality reduction techniques to map high dimensional data onto a low dimensional representation [DKŽ*13b]. The main reason is to display the data in a format that can be visually and intuitively understood [JSLH22, DKŽ*13a]. The disadvantage is that some information is inevitably lost in this reduction, and, for geo-located data, geography is not represented in the resulting visual layouts.

Other solutions in the literature for visualizing geo-located multidimensional categorical data focus on visualizing the category frequencies of the attributes, usually through glyphs. However, only showing their frequencies can be misleading. Moreover, fre-

© 2024 The Authors.
Proceedings published by Eurographics - The European Association for Computer Graphics.
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.2312/evs.20241061

delivered by
EUROGRAPHICS
DIGITAL LIBRARY
www.eg.org diglib.eg.org

quencies do not inform about the gap between an attribute's real occurrence and the expected one. The notion of surprise, which is the difference between real and expected frequencies, offers a deeper understanding of deviations in attribute's frequencies, helping to spot anomalies quickly compared to visualizing raw frequencies [IB09, CH16].

However, neither frequencies nor surprises reflect potential relationships among multiple attributes in the data. For example, suppose that accidents during daylight and where street lights were turned on are equally frequent. Considering that streetlights are typically never turned on during daylight, visualizing only their frequencies without further information can be misleading. We may be misled to think that their combination is important while, in reality, both attributes are completely unrelated. This problem is amplified when the connection between the attributes is less straightforward.

This paper proposes *frequency-relation-glyphs* or *FR-glyphs*, a novel glyph-based visualization to represent multidimensional categorical data on geographical maps aiming to address these issues (Figure 1). Through FR-glyphs two major elements can be visualized: 1) The deviations of the attributes' frequencies to understand differences between reality and expectation, and 2) the relations between (sets of) attributes to give insight into attributes having a higher probability of occurring together. Both elements are combined in a glyph that can be generated for geographical regions. The glyphs are then plotted on a map to allow analysts to compare regions.

2. Related Work

In the literature, different visualization methods represent two or three variables on a map. One example is textured choropleth maps [War19]. The visual channels of color and texture are used to represent two variables. Color can also be used to represent two variables, as is done with relationship maps [Ber18]. The FR-glyphs are different because it is possible to visualize more than three variables. Moreover, the relations between attributes in different categories are shown explicitly in the FR-glyphs, so the issues resulting from counting frequencies of uncorrelated attributes are addressed.

Another visualization for this type of data is Chernoff faces [Che73]. Chernoff faces are glyphs where each variable corresponds to a facial feature, supporting the visualization of a dozen variables. However, it only shows the frequencies of the attributes and can suffer from resulting unnatural mappings and unintended emotive states. A star glyph [Cha18] is another type of glyph that can visualize numerical values. It does not suffer the same cognitive disadvantages as the Chernoff faces [PZS05, PAM07]. However, it also cannot represent relationships among attributes.

The literature also has visualizations that show the frequencies per combination of attributes. Mosaic plots [Fri94] and balloon plots [JW06] are methods for multidimensional categorical data based on matrix layouts. The relative frequency or ratio for each combination of attributes can be traced back to the size of each cell or balloon. Mosaic plots, however, are unsuitable for geographical region comparison because the locations of the cells change depending on the size of the neighbor cells. Another issue is that these plots scale exponentially when a variable is added. In contrast, the FR-glyphs scale linearly.

Beyond frequencies, other techniques also focus on representing the deviation from the expectation, known as Bayesian surprise [IB09]. To find the surprise, the frequencies are estimated using conditional probability, and these values are then subtracted from the real frequencies found in the data. When the real frequency of an attribute is larger than the estimated frequency, the surprise is a positive value. The surprise is negative when the real frequency is smaller than what was estimated. Larger absolute surprise values mean a large difference between the real and expected values. As proposed by M. Correll and J. Heer [CH16], Surprise Maps also use the difference between real and expected values to highlight areas with the highest surprise.

The reason for visualizing surprise instead of the real frequencies is because the frequencies alone can be misleading. Attributes could have a high-frequency count, but if this is what is to be expected, then this high number does not necessarily imply something relevant. Plotting the differences between the real and expected values can more directly show where the frequencies deviate and where more attention should be paid [CH16].

The literature mainly visualizes frequencies (or surprises) of geo-located multidimensional data, even though the relations between attributes bring useful insights. Our contribution is to create a visualization that shows deviations in the attributes' frequencies and which attributes are related or occur together.

3. Methodology

The design of the FR-glyphs consists of two major parts: the deviations in the attributes' frequencies and the relations between attributes of different categories. In this process, the data is segmented considering the different geographical regions so that relations between attributes and their deviations are calculated for every region and represented in a separate FR-glyph. Plotting the FR-glyphs on a map allows for spotting regional differences and similarities.

Before finding relations between attributes, we calculate how much their occurrence deviates from the expectation. If, for example, an attribute appears to have a high deviation, then this may explain the deviations of other attributes within the same relation. The deviations of the attributes' frequencies are calculated using the Bayesian surprise (see Section 2). The surprise is calculated in every region by subtracting the expected frequency from the real frequency. Our assumption is that an attribute's frequency depends on a variable related to every region, like the number of inhabitants or surface area.

We visualize the Bayesian surprise using star glyphs (see Section 2). A positive surprise indicates that the attribute occurred more frequently in the data for that region than is expected. A negative surprise means that it occurred less than expected. Our glyph maps negative values to green and positive to red because more frequent incidents represent a non-ideal situation. In our first study, we evaluated the glyphs using color-blind safe colors, but the involved expert participants got confused and asked for red/green. To make the

© 2024 The Authors.

glyph safe for colorblind users, we use texture. We tested this last configuration with colorblind participants, who agreed they were readable.

We use Association Rule Mining (ARM) to find the relations between (sets of) attributes. ARM is a pattern mining technique that finds relations between (sets of) attributes from different variables indicating their co-occurrence [GWBV03, Cai20]. The strength of association rule mining is that it can find relations between a set with multiple attributes to another set. With two sets of attributes, X and Y, association rules are always in the format $X \rightarrow Y$. The attributes in set X are called the antecedents, and those in set Y are the consequents. We generated association rules or relations between attributes with the standard Apriori algorithm [AS*94] for reproducibility reasons. As the data is segmented for all regions, relations are created for each geographical region in the dataset.

In our design, the association rules are visualized with a chord diagram. All attributes in the data are placed on the edge of a circle. An arrow is drawn between two attributes when the attributes are related. The arrow is always directed from an antecedent to a consequent. When multiple antecedents or consequents exist in an association rule, a chord is drawn from every antecedent to every consequent. We use categorical colors to indicate which connections belong to the same rule.

Finally, we combine the surprise with the relations to create the FR-glyph. The chord diagram is positioned in the center of the glyph. The bars of the star-glyph, indicating the deviations, are placed around the edge of the chord diagram, pointing outwards. Figure 2 illustrates an example of the FR-glyph with three attributes. The glyphs are placed on the map approximately at the location of the region they are representing.

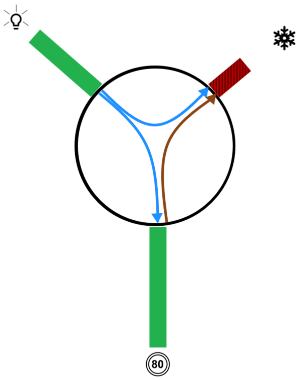

Figure 2: *Simple example of an FR-glyph with attributes "burning light" (upper left), "snowfall" (upper right), and "80 km/h" (bottom) of traffic incidents data. Accidents with "burning light" and "80 km/h" happened less frequently than was expected, while accidents with "snowfall" happened more frequently. When an accident occurs on a road with a maximum velocity of 80 km/h, there is a statistically higher likelihood that snowfall is also present (brown arrow). An accident with burning streetlights has a statistically significant probability of occurring on an 80 km/h road with snowfall (blue arrows).*

4. Results

To assess the usefulness of the FR-glyphs, we evaluated it with the data of Bestand geRegistreerde Ongevallen in Nederland, also called BRON [Rij22]. This dataset contains records of traffic incidents in the Netherlands, where each record is composed of different variables describing the incident and external factors. This dataset is a typical multidimensional dataset with many categorical values. Furthermore, the locations of the accidents are also recorded. Because the Dutch have a real biking culture, we focus only on accidents involving at least one bike.

The map with the glyphs is shown on an interactive dashboard, meaning users can zoom in and out and drag the view. We explain two use cases where fictional mobility policymakers analyze how well their province is doing regarding biking incidents. A video explanation is also available [DC].

4.1. Use Case 1: Explore

Starting from the complete map of the Netherlands showing all provinces, the policymaker moves the view to their province by dragging the view. They then zoom in such that the glyph fills the whole view (see Figure 3). They start by looking at the deviations of the attributes. It occurs to them that attributes "street lights switched off", "dry road surface", "material damage", and "daylight" are the four attributes with the highest absolute deviations. These bars are all colored green, meaning the province had fewer accidents with these attributes than expected based on population size. The attribute "50 km/h" occurs approximately as much as is expected. Three attributes occur more frequently than expected, which are "passenger car", "side", and "other asphalt". These attributes ask for further investigation.

The policymaker now looks at the relations and sees that the orange arrows indeed illustrate a relation: { passenger car, other asphalt } → { side }. They now understand that biking accidents involving a vehicle and the asphalt type "other" also often occur while the bike and car are perpendicular to each other. Lastly, they notice that the attributes "dry road surface" and "material damage" have large surprises, but there is no relation between them. This means that their combination is irrelevant, something they would not be able to conclude without the relations.

We tested exploration of the FR-glyphs with 14 real policymakers, and their main feedback was that the relations are still difficult to understand. Three indicated that it would be useful to include a textual explanation that would appear when a policymaker clicks on a relation.

4.2. Use Case 2: Compare

The policymaker now wants to know how well their province does compared to the regions around it. They zoom out such that the two glyphs are shown for Groningen and the province of Drenthe (see Figure 4). They immediately notice that Drenthe has a lot of attributes with a much lower frequency than expected. They know that Drenthe has a lower population density, which could indicate why these attributes occur less frequently for the same number of inhabitants.

© 2024 The Authors.
Proceedings published by Eurographics - The European Association for Computer Graphics.

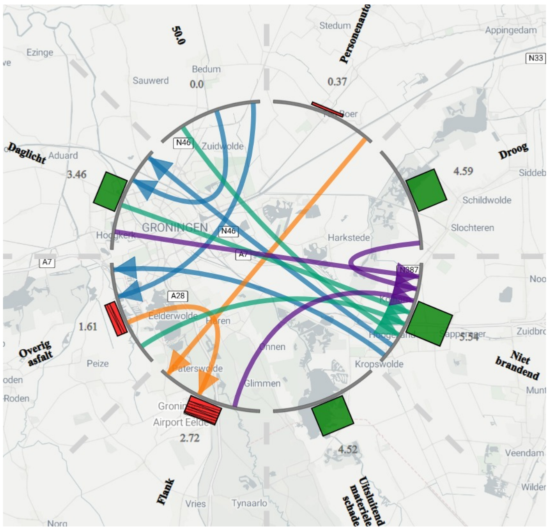

Figure 3: *FR-Glyph showing surprise (bars) and relations (arrows) for external attributes related to bicycle accidents in the Dutch province of Groningen. Attributes "Personenauto" (passenger car), "Flank" (side), and "Overig asfalt" (other asphalt) happened more often than expected, indicated by the red bars.*

The policymaker also sees that "other asphalt" is not involved with other attributes in Drenthe. The relation in Drenthe that comes closest to the { passenger car, other asphalt } → { side } is the relation with the orange arrows: { daylight, side } → { passenger car }. Perhaps they could investigate how the road surface differs in both regions. Moreover, they see that there are many more attributes present in Groningen's relations. Again, many of those are present due to the attribute "other asphalt" – a piece of potential evidence to start investigating this road surface type.

The most important feedback of the policymakers here is that they would like a feature where they can select glyphs and have them shown next to each other. Even better would be if the tool automatically shows the differences between two glyphs.

5. Discussion

While our idea is promising for visualizing multidimensional and categorical data on a geographical map, it is important also to state some limitations. First, the glyphs should be large enough to be readable but small so that different regions do not overlap. The FR-glyphs can be plotted for the 12 provinces in the Netherlands without any overlap. For the 300+ municipalities, on the other hand, there are a lot of overlaps. To circumvent this slightly, dynamic scaling has been implemented, meaning that the size of the glyphs becomes smaller when zooming in. Still, a maximum of 4 glyphs with approximately 8 attributes can be shown simultaneously before it becomes unreadable. When looking at a single glyph, approximately 30 factors can be shown, although the exact number for readability should be further investigated.

Secondly, the number of relations that can be shown in the glyph depends on the number of colors that people can easily distinguish. It is, therefore, possible to show a maximum of 7 different relations, which is even reduced for colorblind users. It is also advised to show no more than 10 arrows simultaneously in one glyph to prevent too much overlap. In the scenarios we tested, the FR-glyphs were capable of showing all relevant information (confirmed by the experts we interviewed). However, for larger datasets, other visual representations may be more suitable.

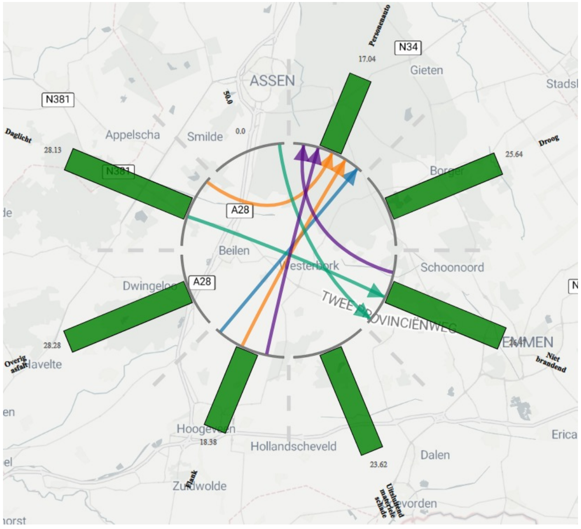

Figure 4: *FR-glyph for the Dutch province of Drenthe, showing surprise and relations for external attributes related to bicycle accidents.*

6. Conclusions and Future Work

This paper presents a novel visualization for analyzing multidimensional categorical data on a geographical map called FR-glyphs. The major advantages of FR-glyphs are that they can be compared to each other, scale linearly in terms of variables, and show both the frequencies (surprise) of different attributes and relations. User studies are in progress to evaluate the final design and see what elements should be improved. Our intermediate evaluations have shown that FR-glyphs are promising solutions for what they intend to represent.

One direction for future research is to visualize the relations using other visual metaphors, for example, using the ARMatrix [VCP22] matrix, where each row represents an attribute, and each column is an association rule. Colors could represent antecedents and consequents, and bars could represent the surprises plotted horizontally next to the matrix. Lastly, it may be interesting to experiment with symbols or icons that represent the attributes in the glyphs instead of text, such as shown in Figure 2.

References

[AS*94] AGRAWAL R., SRIKANT R., ET AL.: Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB* (1994), vol. 1215, Santiago, Chile, pp. 487–499. 3

© 2024 The Authors.

[BB19] BLACK W., BABIN B. J.: Multivariate data analysis: Its approach, evolution, and impact. In *The great facilitator: Reflections on the contributions of Joseph F. Hair, Jr. to marketing and business research*. Springer, 2019, pp. 121–130. 1

[Ber18] BERRY L.: How to make a relationship map in arcgis online. 2

[Cai20] CAI Q.: Cause analysis of traffic accidents on urban roads based on an improved association rule mining algorithm. *IEEE Access 8* (2020), 75607–75615. 3

[CH16] CORRELL M., HEER J.: Surprise! bayesian weighting for de-biasing thematic maps. *IEEE transactions on visualization and computer graphics 23*, 1 (2016), 651–660. 2

[Cha18] CHAMBERS J. M.: *Graphical methods for data analysis*. CRC Press, 2018. 2

[Che73] CHERNOFF H.: The use of faces to represent points in k-dimensional space graphically. *Journal of the American statistical Association 68*, 342 (1973), 361–368. 2

[DC] DELOREAN CANLON FERNANDO PAULOVICH M. T.: Fr glyph use cases. URL: https://youtu.be/kaq9hyyIf9w. 3

[DKŽ*13a] DZEMYDA G., KURASOVA O., ŽILINSKAS J., DZEMYDA G., KURASOVA O., ŽILINSKAS J.: Multidimensional data and the concept of visualization. *Multidimensional Data Visualization: Methods and Applications* (2013), 1–4. 1

[DKŽ*13b] DZEMYDA G., KURASOVA O., ŽILINSKAS J., DZEMYDA G., KURASOVA O., ŽILINSKAS J.: Strategies for multidimensional data visualization. *Multidimensional data visualization: Methods and applications* (2013), 15–38. 1

[Fri94] FRIENDLY M.: Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association 89*, 425 (1994), 190–200. 2

[GWBV03] GEURTS K., WETS G., BRIJS T., VANHOOF K.: Profiling of high-frequency accident locations by use of association rules. *Transportation research record 1840*, 1 (2003), 123–130. 3

[IB09] ITTI L., BALDI P.: Bayesian surprise attracts human attention. *Vision research 49*, 10 (2009), 1295–1306. 2

[JSLH22] JIA W., SUN M., LIAN J., HOU S.: Feature dimensionality reduction: a review. *Complex & Intelligent Systems 8*, 3 (2022), 2663–2693. 1

[JW06] JAIN N., WARNES G. R.: Balloon plot. *The Newsletter of the R Project Volume 6/2, May 2006 6* (2006), 35. 2

[PAM07] PROHASKA G., AIGNER W., MIKSCH S.: Glyphs and visualization of multivariate data. *Vienna University of Technology* (2007). 2

[PX08] POWERS D., XIE Y.: *Statistical methods for categorical data analysis*. Emerald Group Publishing, 2008. 1

[PZS05] PFLUGHOEFT K. A., ZAHEDI F. M., SOOFI E.: Data visualization using figural animation. *AMCIS 2005 Proceedings* (2005), 297. 2

[Rij22] RIJKSWATERSTAAT: Verkeersongevallen nederland - ongevallen 2019 - 2021 (rws). https://data.overheid.nl/en/dataset/ab73b819-42cd-4e2c-8edc-fbf640556ef9, December 2022. URL: https://data.overheid.nl/en/dataset/ab73b819-42cd-4e2c-8edc-fbf640556ef9. 3

[VCP22] VARU R., CHRISTINO L., PAULOVICH F. V.: Armatrix: An interactive item-to-rule matrix for association rules visual analytics. *Electronics 11*, 9 (2022), 1344. 4

[War19] WARE C.: *Information visualization: perception for design*. Morgan Kaufmann, 2019. 205–215. 2

© 2024 The Authors.