




Aggregation Bias: A Proposal to Raise Awareness Regarding Inclusion in Visual Analytics

Andrea Vázquez-Ingelmo¹ , Francisco J. García-Peñalvo¹ ,
and Roberto Therón^{1,2} 

¹ GRIAL Research Group, Computer Sciences Department,
Research Institute for Educational Sciences, University of Salamanca,
Salamanca, Spain

{andreavazquez, fgarcia, theron}@usal.es

² VisUSAL Research Group, University of Salamanca, Salamanca, Spain

Abstract. Data is a powerful tool to make informed decisions. They can be used to design products, to segment the market, and to design policies. However, trusting so much in data can have its drawbacks. Sometimes a set of indicators can conceal the reality behind them, leading to biased decisions that could be very harmful to underrepresented individuals, for example. It is challenging to ensure unbiased decision-making processes because people have their own beliefs and characteristics and be unaware of them. However, visual tools can assist decision-making processes and raise awareness regarding potential data issues. This work describes a proposal to fight biases related to aggregated data by detecting issues during visual analysis and highlighting them, trying to avoid drawing inaccurate conclusions.

Keywords: Data bias · Information visualization · Data visualization · Inclusion awareness

1 Introduction

Information has grown in size and relevance over the last years; technology has not only increased the generation of data but also their accessibility. People with an Internet connection can consult a wide range of datasets about almost any topic: crime data, healthcare data, weather data, financial data, etc.

These data can be employed to make informed decisions regarding different domains. For example, businesses can employ demographic data to create personalized advertisements or to segment the market. Governments can employ their data to design new policies. Any person regularly uses data to make informed decisions. A simple question like “should I get a coat to go out today?” can be answered through data (made available by weather services) to make an informed decision that, in the end, seeks some kind of benefit (in this case, the benefit of avoiding hypothermia).

However, delegating decisions solely in data might turn out to be a two-edged sword. Data not only can be wrong or false, but it can also be incomplete, and making

decisions using wrong data leads to wrong decisions. There are several cases in which relying on the wrong data has provoked undesired results, mostly because of data bias or even algorithmic bias [1–3].

So it seems clear that if the data that you are using to make decisions is not the best for your problem, you could end up with decisions that are also not the best for your problem. But how can people avoid such inconveniences with data? Bias is generally introduced unconsciously, and it can be hard to detect our own biases and be aware of them while collecting data. For these reasons, data should be thoroughly examined to identify gaps or inconsistencies before using them in decision-making processes.

One of the most used methods to ease the analysis and exploration of datasets is visual analytics [4, 5]; using information visualizations, users can interact and explore datasets through visual marks that encode certain information [6]. However, visualizations could hide data issues by lifting the attention from the analysis process carried out on the raw data to the discovered patterns. Patterns can be seen as shortcuts that tell us properties about the data, for example, if there are correlations among the visualized variables [7]. But visual analysis shouldn't be reduced to just the identification of patterns and to trust them blindly, because patterns can likewise lead to wrong conclusions [8].

This work describes a proposal for raising awareness during visual analysis, helping users to make informed decisions taking into account the flaws or potential issues of their datasets. Specifically, issues related to data aggregation, which can be very harmful in data-driven decision-making processes. The main goal is not only to improve decision-making, but to address inclusion problems when dealing with data, as data biases can lead to decisions that (involuntarily, or not) discriminate individuals.

The rest of this paper is organized as follows. Section 2 introduces some issues related to data analysis and data aggregation. Section 3 describes the methodology followed to design the proposal. Section 4 presents a proposal to raise awareness during visual data analysis. Section 5 discusses the proposal, following by Sect. 6, in which the conclusions derived from this work are outlined.

2 Background

The outcomes of decision-making processes are actions that affect the context in which decisions are being made. When deciding which action to take, the decision-maker will have an assumption on how the action's effect will affect the context, looking for a benefit or a pursued result. However, the critical fact is that assumptions can be very personal and could vary depending on the person's beliefs, background, domain knowledge, etc.

Even when the decision-maker support its decisions on data (embracing data-driven decision-making [9]), there are still problems. As introduced before, data is not the holy grail of decision-making, because as well as personal traits can influence the decision-maker, the collected data and performed analyses can be influenced by other harmful factors like data biases [10] or poor analysis.

There are specific fields of study, like uncertainty visualization, that try to find methods to visualize uncertain data, thus warning users regarding the uncertain nature of the results they are consuming through their displays [11, 12]. However, uncertainty

visualization is complex, and several concepts could be difficult to understand by non-technical or non-statistical audiences, such as probabilities or densities, resulting in users ignoring or misinterpreting uncertainty [13].

On the other hand, the data that is being visualized can present issues that could be concealed and not considered through information visualizations, like excessive (or not appropriate) aggregation levels, which could result in wrong conclusions.

Summary statistics summarize a set of observations through a collection of values that simplify the comprehension of the datasets. But this simplification comes with a price; while performing these summaries, a lot of information can be lost. One of the most famous examples of this drawback is Anscombe's quartet [14], in which different datasets that tell very different stories have the same mean and variance. Anscombe highlighted the usefulness of graphics [14] to avoid these issues (Fig. 1).

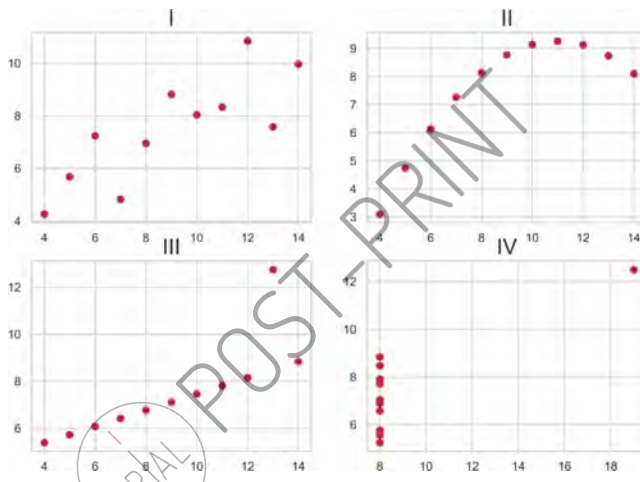


Fig. 1. The Anscombe's Quartet. The four datasets have the same mean and variance values on both variables represented on the X and Y axes.

So aggregated data ease the analysis process, but they can lead to a loss of information. Aggregated data can also be vulnerable to phenomena like the ecological fallacy and the Simpson's paradox [15].

Inferring individual behavior by using aggregated data is a common extrapolation mistake, where analysts might conclude that the behavior of a group is also accurate to explain the behavior of the individuals within that group [16, 17].

Simpson's paradox is also related to the data aggregation-level. In this case, there might exist lurking variables that could entirely "change" the conclusions derived from aggregated data [18, 19].

These aggregation-related issues can be very harmful if not taken into account [20], especially if the audience is biased or not statistically-trained (or both).

Some works have tried to address these aggregation drawbacks through detection algorithms [21, 22], but a few tried to address them during visual exploration [23].

3 Methodology

The proposal focuses on how to draw attention to potential aggregation biases and fallacies during visual analysis. A simple workflow has been considered to automatically seek for aggregation issues regarding the data being presented to the user. Specifically, issues involving the Simpson's paradox and underrepresentation of categories.

Each categorical variable is considered as a potentially influencing variable. Of course, as it will be discussed, this methodology is limited to the available variables within the dataset. If the whole dataset has a small set of categories, the results would not be as useful as it could be with a richer dataset.

The workflow follows a naïve approach to detect Simpson's paradoxes [23]:

1. Every possible grouping at any possible level is computed on categorical to obtain a set of potential disaggregation variables.
2. When the user visualizes data, the current aggregation level is retrieved (i.e., the categorical columns used to group the data)
3. These data are then grouped by the variables identified in the first step.
4. The results of the performed disaggregation are sorted and compared with the original scenario (i.e., the aggregated data values) trend.
5. If the disaggregation results differ from the originally aggregated results (a threshold can be defined to specify which proportion of values need differ from the original trend to consider the paradox), the Simpson's paradox is considered for the disaggregated attributes

However, even visualizing the disaggregated data by the identified attributes in the fifth step, there could still be aggregation issues if data are in turn aggregated by a function such as the mean, mode, ratios, etc. These functions can, in turn, distort the reality of data.

To avoid relying on aggregation functions, when the detected Simpson's paradoxes are inspected, a sunburst diagram complements the display to give information about the raw data sample sizes regarding the disaggregated values.

Sunburst diagrams are usually employed to represent hierarchies; in this context, they are useful to display how the number of observations of the variable being inspected varies its size among the different disaggregation levels.

The primary purpose is to have another perspective of data, drawing attention over potential underrepresentation or overrepresentation in datasets.

4 Proposal

A simple proof-of-concept has been developed to illustrate the proposal. The employed test data is from one of the most famous cases involving Simpson's paradox: the student admission at UC Berkeley in 1975 [24]. This dataset holds the following information about each student: gender, the department in which the application was issued, and the result of the application (admitted or rejected).

If the gender variable aggregates this data, the results yield a significant gender bias against women: only 35% of women were admitted, in contrast with the 44% of admitted males. This data could help the decision-makers to design new policies trying to address the discovered gender bias.

However, this high-level aggregation hides some parts of the picture. If data is, in turn, disaggregated using the department in which the application was issued, we see a different scenario: the majority of the departments shown higher admission rates for women than men. What was happening is that women applied to more competitive departments than men, who issued the majority of applications to departments with a high rate of admissions (resulting in higher admissions rates among male students).

This case is a famous example of Simpson’s paradox, but misleading conclusions can be present in any context if these potential issues in data analysis are not accounted for. For this reason, the interface presented in Fig. 2 is proposed.

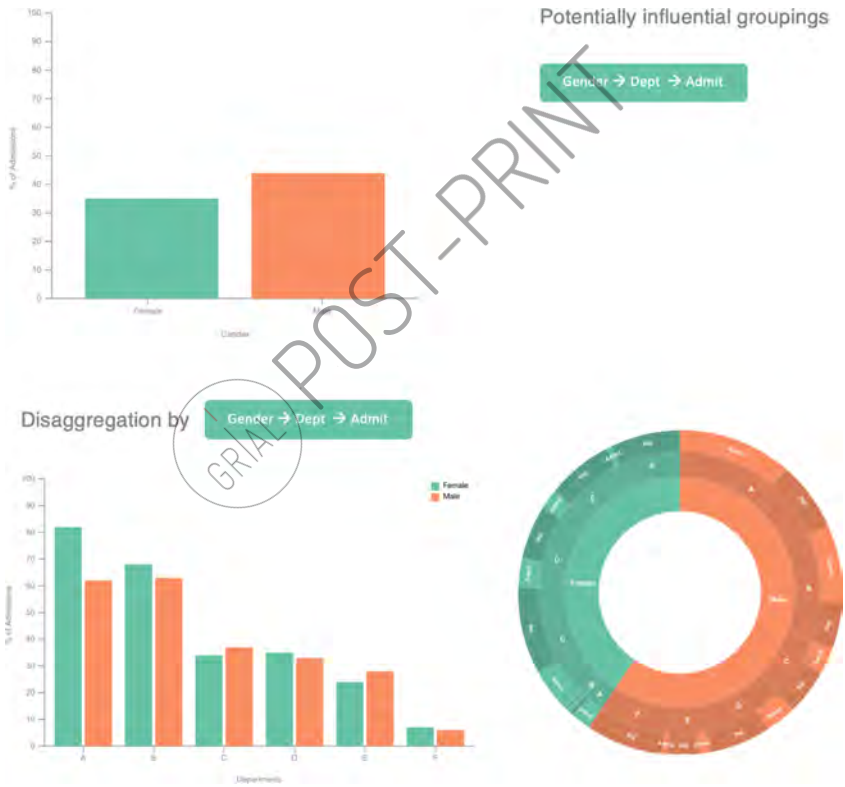


Fig. 2. Interface proposal for detecting aggregation issues.

When the user is exploring her dataset, Simpson’s paradox detector starts searching for potentially influential groupings that change the trend of the currently displayed

variables. If any grouping changes the trend, the categorical variables identified are displayed (top section of Fig. 2).

The user then can click on each detected grouping to explore how the disaggregation affects the value that she was examining, in addition to a sunburst diagram that shows the distribution of occurrences of each observation under the selected grouping (bottom section of the Fig. 2). In this specific example, the user can observe how women apply less to departments with high admission rates (like department A, for example) and issue more applications to more competitive departments, obtaining a complete view of the examined data.

5 Discussion

Aggregating data is useful to summarize observations, but it can overlook crucial aspects of data, like, for example, underrepresentation or overrepresentation of the samples. Raising attention over this matter is essential, especially when studying behavioral data or data that involve human beings.

How can this approach benefit decision-makers regarding inclusion-related issues? Our biases could blind ourselves and make us not prone to ask skeptical questions about the analyzed data. If data confirms something we believe, we might trust the results without carrying out further analyses [25].

This approach forces analysts (or any kind of audience) to have a more in-depth look at aggregated data, which sometimes can conceal underlying patterns or trends. Having a deeper look is crucial when analyzing data for inclusion-related research contexts because it is possible to visualize if aggregated results are due to the overrepresentation of certain categories and to identify if any category is missing or not represented at all.

Not considering aggregation issues can strengthen the belief that “one size fits all”, which can lead to (involuntary) discrimination. If you design a product (referring to an object, an algorithm, a policy, a treatment, etc.) for “people” and you use data that only represent a particular portion of people or don’t bring attention to their differing characteristics, you end up with a personalized product for a segment. There is nothing wrong with personalized products; what is wrong is to think that this unconsciously personalized product is universal and should fit every individual.

The underrepresentation of certain categories depends, of course, on the data context. For example, in the Berkeley dataset, the underrepresentation of women’s applications to some departments is due to the preference of the students to apply to specific departments. But there are other cases in which the underrepresentation is due to selection bias or a not representative sampling of the population. It is essential to take this into account to avoid data bias against minorities (or even against non-minorities, like women [20]).

A proposal for visually identifying aggregation issues (especially those related to the Simpson’s paradox) has been developed. Of course, this proposal does not try at all to replace statistical methods but to deliver a visual tool to understand better our datasets.

The proposal has been focused on raising awareness regarding how disaggregating data could change the patterns identified during the analysis of aggregated data. It also could be used as an informative tool to educate people through a friendly interface regarding the underlying issues of data aggregation and their dangerous effects on decision-making processes.

Educating people in data skepticism and regarding potential biases is important because data visualizations can be very persuasive and could influence people's beliefs. Relying on data visualizations tools to raise awareness can be powerful due to the possibility of presenting information in understandable manners and also to the possibility of enabling individuals to freely interact with data [26, 27].

The methodology seeks for sub-groups that "change" the original scenario (i.e., the trends identified on aggregated data). It is important to mention that, in this case, statistical significance has not been considered because the main goal was to draw attention to changes in visual patterns, no matter how small. However, complementing this methodology with the computation of statistical significance could be more powerful in some contexts [23].

Statistically-trained audiences might be aware of these issues. However, other audiences could reach wrong insights about data if attention is not raised regarding potential issues, thus distorting the decision-making process without even notice.

For example, when dealing with policies that affect individuals, it is crucial to rely on disaggregated data to avoid ignoring the necessities of minorities [28–30].

But when talking about disaggregated data, there are some limitations to take into account. Demographic variables are meaningful for inclusion-related research contexts, but also sensitive. Some of these variables can be difficult to collect because of privacy policies or privacy concerns.

In fact, for some activities as for example, hiring people, having such data available could introduce the risk of biasing the decisions made during some phases of the process [31, 32]. So analysts and decision-makers must understand the level of analysis and goals to anonymize or omit these attributes accordingly.

To sum up, it is important to foster critical thinking and some skepticism toward data. When dealing with information about individuals, accounting for data gaps is a responsibility, because the decisions made could have a high impact in the context of application, and sometimes, this impact is not beneficial for everyone.

6 Conclusions

This work presents a proposal for raising awareness in decision-making processes through visual analysis. Relying on inappropriate data could lead to wrong decisions. But identifying flaws in data is not a trivial task; bias, beliefs, and uncertainty can show up both at data collection time and analysis time, resulting in distorted insights.

Through the detection of existing Simpson's Paradox and the disaggregation of the displayed data, the presented proposal tries to draw attention to issues like excessive or inappropriate aggregation levels and potential overrepresentation or underrepresentation of data attributes or categories.

Future work will involve the evaluation and refinement of the proposal to improve its effectiveness to obtain a tool to raise awareness about inclusion in different fields.

Acknowledgments. This research work has been supported by the Spanish *Ministry of Education and Vocational Training* under an FPU fellowship (FPU17/03276). This work has been partially funded by the Spanish Government Ministry of Economy and Competitiveness throughout the DEFINES project (Ref. TIN2016-80172-R) and the Ministry of Education of the Junta de Castilla y León (Spain) throughout the T-CUIDA project (Ref. SA061P17).

References

1. Sweeney, L.: Discrimination in online ad delivery. arXiv preprint [arXiv:1301.6822](https://arxiv.org/abs/1301.6822) (2013)
2. Garcia, M.: Racist in the machine: the disturbing implications of algorithmic bias. *World Policy J.* **33**, 111–117 (2016)
3. Hajian, S., Bonchi, F., Castillo, C.: Algorithmic bias: from discrimination discovery to fairness-aware data mining. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2125–2126. ACM (2016)
4. Keim, D.A., Andrienko, G., Fekete, J., Görg, C., Kohlhammer, J., Melançon, G.: Visual analytics: definition, process, and challenges. In: Kerren, A., Stasko, J., Fekete, J., North, C. (eds.) *Information Visualization*, pp. 154–175. Springer, Heidelberg (2008)
5. Thomas, J.J., Cook, K.A.: *Illuminating the path: the research and development agenda for visual analytics*. National Visualization and Analytics Center, USA (2005)
6. Munzner, T.: *Visualization Analysis and Design*. AK Peters/CRC Press, Boca Raton (2014)
7. Harrison, L., Yang, F., Franconeri, S., Chang, R.: Ranking visualizations of correlation using weber’s law. *IEEE Trans. Visual Comput. Graph.* **20**, 1943–1952 (2014)
8. O’Neil, C.: *On Being a Data Skeptic*. O’Reilly Media, Inc., Newton (2013)
9. Patil, D., Mason, H.: *Data Driven*. O’Reilly Media Inc, Newton (2015)
10. Shah, S., Horne, A., Capellá, J.: Good data won’t guarantee good decisions. *Harvard Bus. Rev.* **90**, 23–25 (2012)
11. Bonneau, G.-P., Hege, H.-C., Johnson, C.R., Oliveira, M.M., Potter, K., Rheingans, P., Schultz, T.: Overview and state-of-the-art of uncertainty visualization. In: *Scientific Visualization*, pp. 3–27. Springer, Heidelberg (2014)
12. Brodlie, K., Osorio, R.A., Lopes, A.: A review of uncertainty in data visualization. In: *Expanding the Frontiers of Visual Analytics and Visualization*, pp. 81–109. Springer, Heidelberg (2012)
13. <https://medium.com/multiple-views-visualization-research-explained/uncertainty-visualization-explained-67e7a73f031b>
14. Anscombe, F.J.: graphs in statistical analysis. *Am. Stat.* **27**, 17–21 (1973)
15. Pollet, T.V., Stulp, G., Henzi, S.P., Barrett, L.: Taking the aggravation out of data aggregation: a conceptual guide to dealing with statistical issues related to the pooling of individual-level observational data. *Am. J. Primatol.* **77**, 727–740 (2015)
16. Kramer, G.H.: The ecological fallacy revisited: aggregate-versus individual-level findings on economics and elections, and sociotropic voting. *Am. Polit. Sci. Rev.* **77**, 92–111 (1983)
17. Piantadosi, S., Byar, D.P., Green, S.B.: The ecological fallacy. *Am. J. Epidemiol.* **127**, 893–904 (1988)
18. Blyth, C.R.: On Simpson’s paradox and the sure-thing principle. *J. Am. Stat. Assoc.* **67**, 364–366 (1972)
19. Wagner, C.H.: Simpson’s paradox in real life. *Am. Stat.* **36**, 46–48 (1982)

20. Perez, C.C.: *Invisible Women: Exposing Data Bias in a World Designed for Men*. Random House, New York (2019)
21. Alipourfard, N., Fennell, P.G., Lerman, K.: Can you trust the trend?: discovering Simpson's paradoxes in social data. In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pp. 19–27. ACM (2018)
22. Xu, C., Brown, S.M., Grant, C.: Detecting Simpson's paradox. In: *The Thirty-First International Flairs Conference* (2018)
23. Guo, Y., Binnig, C., Kraska, T.: What you see is not what you get!: detecting Simpson's paradoxes during data exploration. In: *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics*, p. 2. ACM (2017)
24. Bickel, P.J., Hammel, E.A., O'Connell, J.W.: Sex bias in graduate admissions: data from Berkeley. *Science* **187**, 398–404 (1975)
25. Nickerson, R.S.: Confirmation bias: a ubiquitous phenomenon in many guises. *Review of general psychology* **2**, 175–220 (1998)
26. Hullman, J., Adar, E., Shah, P.: The impact of social information on visual judgments. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1461–1470. ACM (2011)
27. Kim, Y.-S., Reinecke, K., Hullman, J.: Data through others' eyes: the impact of visualizing others' expectations on visualization interpretation. *IEEE Trans. Visual Comput. Graph.* **24**, 760–769 (2018)
28. Mills, E.: 'Leave No One Behind': Gender, Sexuality and the Sustainable Development Goals. *IDS* (2015)
29. Stuart, E., Samman, E.: Defining "leave no one behind". ODI Briefing Note. London: ODI (www.odi.org/sites/odi.org.uk/files/resource-documents/11809.pdf) (2017)
30. Abualghaib, O., Groce, N., Simeu, N., Carew, M.T., Mont, D.: Making visible the invisible: why disability-disaggregated data is vital to "leave no-one behind". *Sustainability* **11**, 3091 (2019)
31. Rice, L., Barth, J.M.: Hiring decisions: the effect of evaluator gender and gender stereotype characteristics on the evaluation of job applicants. *Gend. Issues* **33**, 1–21 (2016)
32. Alford, H.L.: *Gender bias in IT hiring practices: an ethical analysis* (2016)