Visualization and communication of statistics

Keywords: Data visualization, perception, cognition

Introduction

Statistics Netherlands (CBS) is responsible for compiling and publishing reliable statistics of the Netherlands serving as a quantitative backbone of the Dutch society. Policy makers, journalists, and teachers are among the groups that actively use output produced by CBS. As intended by the European Code of Practice [1] CBS puts much effort in producing high quality statistics and developed expertise in assessing the quality of surveys, registers, data editing, estimation and aggregation. The production of high quality statistics is essential, but not sufficient to accomplish the mission of a statistical office: statistics have to be published and communicated to be useful for society.

Current approaches of visualizing and communicating official statistics are often based on convenience, user experience and common journalistic principles and practices. Quality aspects with respect to the understanding of the message of statistical communication are less known, and have not received much attention yet. It is unclear how information that CBS publishes comes across and how it effects the understanding and interpretation of the published statistic. Naïve statistical communication may introduce the risk of visual bias in a chart or prime a reader to draw premature conclusions. Furthermore effectively communicating (in)accuracy of statistics is important: it indicates the value for society.

This paper suggests several research directions that can improve the effectiveness of (visual) communication. We will focus on visualizations, since those are effective means to summarize data, but the research can be extended to textual communication. We plan to conduct in-depth user studies go obtain insights about the effectiveness of our communication methods, and propose alternatives when needed. Furthermore, our aim is to make checklists and guidelines that help visual designers.

The central research question that we propose is: *How can statistical information be effectively visualized so that it is understandable and usable by the intended audience?* Although it may seem a trivial question, surprisingly little research has been done to tackle it. Research has been done in several academic disciplines, such as information visualization, statistics, computer science (information theory), and psychology, but a holistic, interdisciplinary, approach is missing. We propose a model for data visualization that can be used as a guide to develop relevant research questions as well as user experiments to meet the needs of official statistics.

A data visualization model

Rensink [2] views visualization as a process in which graphical representation of data is converted via a visual percept to a conceptual representation. Chen and Golan [3] have an information theoretic perspective on data visualization, where the amount of information (Shannon entropy) is purposely reduced in order improve the efficiency of the visualization. Questionnaire design is similar to data visualization in many respects as the way a respondent perceives the response task together with the visual design of a survey question and his answer options. Tourangeau [4,5,6] breaks this cognitive question-answer process down to four stages: comprehension, retrieval, judgement and response. In Figure 1, a model of data visualization is proposed where these perspectives are united.

Data visualization has emerged as part of statistics and computer science [7], and less as part of visual perception and cognition. Therefore, it is common that large organizations only recruit data visualization designers who have a background in statistics, computer science, of graphic arts. However, little is known how users perceive and interpret visualizations. Therefore, extra knowledge and research is needed from visual perception and cognition.

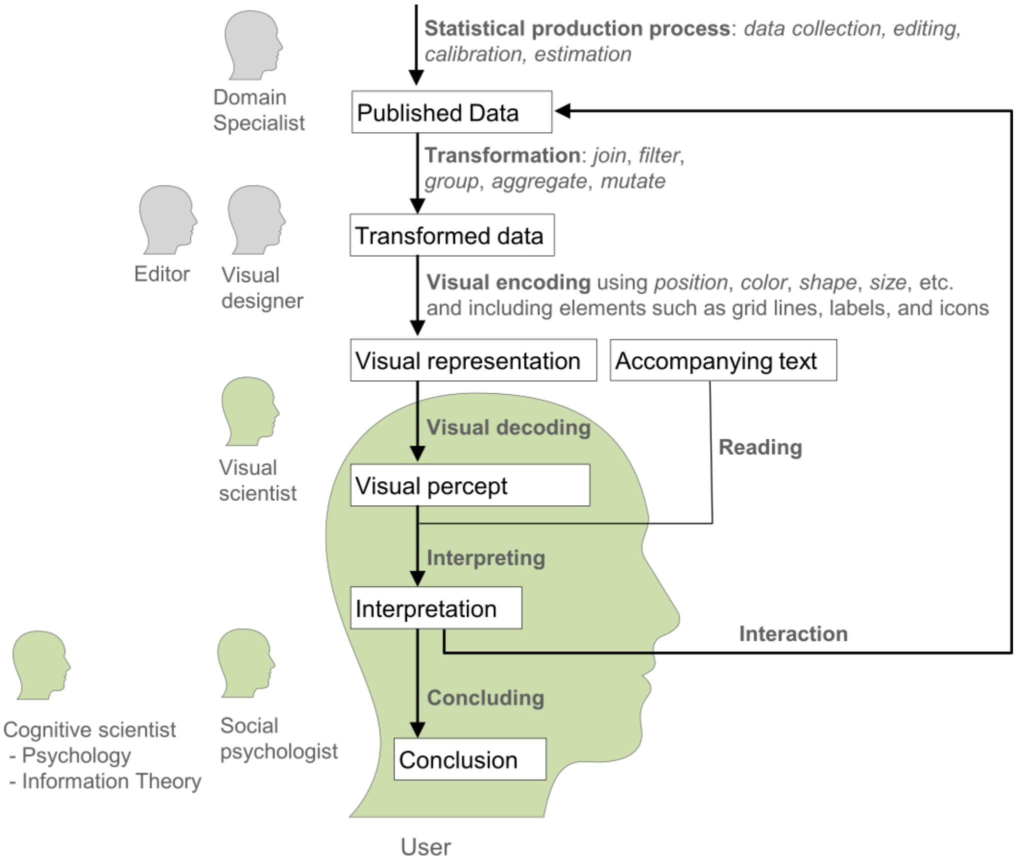

Figure 33. A data visualization model. The process is depicted on the right hand side. The required roles for designing data visualizations are shown on the left hand side.

Quality Guidelines

The research that we propose should gradually and eventually result in guidelines that can be used in practice to improve overall quality of official statistics. Table 1 illustrates the relationship between quality dimensions for Official Statistics [8] on the one hand (rows) and communication (text publications) and visualizations on the other hand (columns).

Table 1. Relationship between quality indicators and communication / visualization

	communication	**visualization**
relevance		
accuracy	perception biases based on opinions and culture, leading to ambiguities.	non-linear behavior of the visual system (see next paragraph)
timeliness		
accessibility	e.g. readability of the text	easiness vs. attractiveness

interpretability	e.g. definitions	a good visualization should guide the user in interpreting the statistics
coherence		tool to integrate different statistics and gives the ability to compare.

Relevance refers to the degree to which the statistical output is able to meet the real needs of the client. When it comes to accuracy, it refers to how well the phenomena are described. This dimension includes different types of errors. When it comes to timeliness, it's about how long it takes for a phenomenon to reach its final publication. Accessibility refers to the ease of obtaining information from the NSO. Interpretability refers to the availability of metadata and additional information to assist in the interpretation of the data. Lastly, coherence describes the extent to which the statistics are able to be integrated consistently.

Visual decoding

Research should be performed that focusses on bringing the theoretical knowledge on visual decoding into practice, and thereby providing information on how reliability and validity of visualization can be improved. As an example, Cleveland and McGill [9] studied the effect of using different visual variables on the accuracy of the perceived data. The findings of this empirical study, which are verified [10] are summarized in Figure 7.

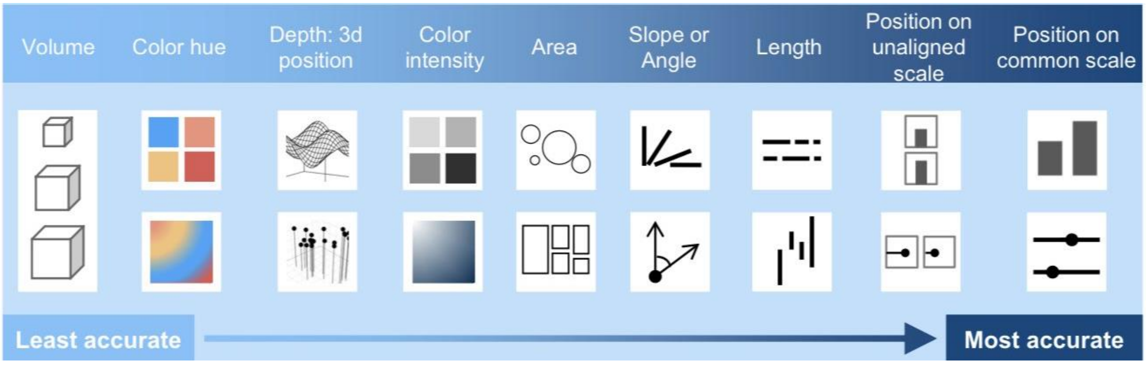

Figure 2. Accuracy of visual variables according to Cleveland and McGill (1984)

It is important to realize is that the human visual system evolved over millions of years in a natural environment in which 3D objects were present. Consequently, Gregory [11] and Marr [12] suggest that every image, including a 2D picture, appears as a certain view in a 3D environment. Visual elements, such as angles orientation, and colors are used by the human visual system as depth cues, leading to the creation of visual illusions. The presence of these factors could introduce biases into the perception of charts, even the simplest of them.

Another source of bias is shown in Figure 3. A five color blue scale is used in this choropleth, but in some cases is it hard to compare non-neighboring regions, for instance the two regions marked in red. Although the two marked areas have exactly the same value, they are perceived differently due to contrast effects. This is due to the fact that the perception of colour is often relative to its surroundings.

Interpretation

Visualizations could definitely be improved so that they truly ease the understanding and adequate use of statistics. But visualizations could also misguide the user. Most statistical output consist of summary statistics such as means and totals and bar charts are often used to visualize these. Kerns and Wilmer [13] suggested that proper mean bar charts can nevertheless be misinterpreted, because users may not realize that the underlying data distribution may be less smooth than it may seem. Just using a visualization may not be sufficient, and it is unclear whether such side-effects occur with other visualizations of standard statistics.

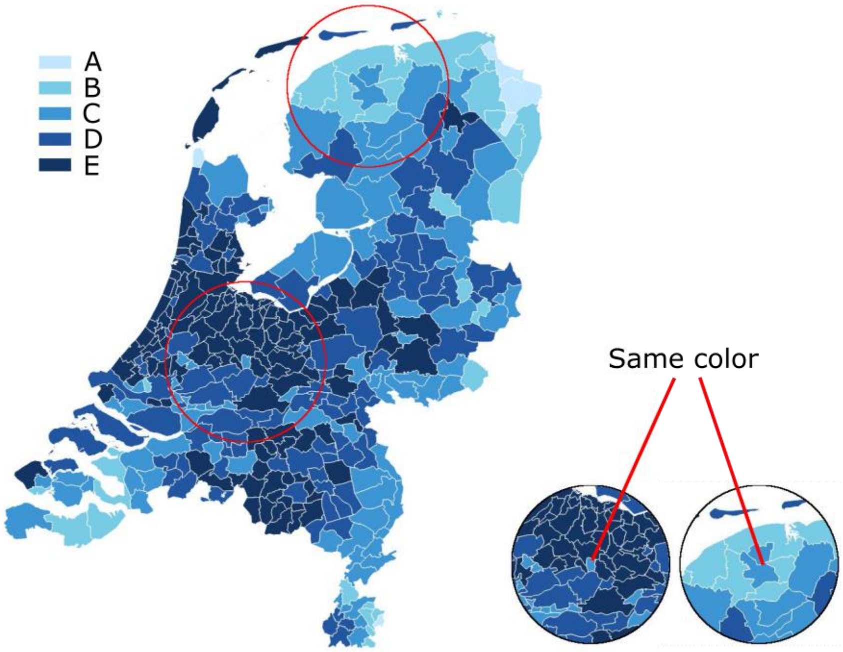

Figure 3. Choropleth showing that color perception depends on the surroundings

While visualizing standard statistics may already cause problems, things may even get more complex with more advanced statistical output like uncertainty. Official statistics has the aim to produce accurate, precise and valid estimates, so each statistical estimate should include an indication of its accuracy, precision and validity. Knowledge with respect to communicating uncertainty is scarce and proper guidelines are lacking [14, 15]. A review of different visualization methods for communication uncertainty in Official Statistics is given in [16].

Conclusions

Effective communication of Official Statistics is an essential part of the mission of NSI’s. However, while the quality of official statistics is highly regarded, the quality of visualization and communication of those statistics is often overlooked. Communicating statistics has be done for many years, but perceptual, cognitive and other presentation errors are abound. Visualizations may suffer from all kinds of perception biases and misinterpretations. Surprisingly, the practice of communication of statistics lacks scientific rigor. Furthermore, statistics have limited precision, but accuracy of statistics is seldom communicated.

Therefore, we encourage the official statistics community to investigate how users perceive, interpret, and use our statistical output and in particular visualizations. We propose to develop guidelines based on latest insights from several scientific fields and based on user studies to elevate the quality of communication above “best practices”.

As a start, we plan to conduct user studies at CBS, both in a controlled environment (for studies on visual perception), and in a free environment (for qualitative studies). In this way, different visual designs and components, such as colour schemes, can be compared systematically. The results should eventually lead to checklists and guidelines for visual designers.

References

[55] Eurostat (2017). European Statistics Code of Practice (https://ec.europa.eu/eurostat/web/quality/european-quality-standards/european-statistics-code-of-practice). Luxembourg: Eurostat.

[56] Rensink, Ronald. (2014). On the Prospects for a Science of Visualization. 10.1007/978-1-4614-7485-2_6.

[57] Chen, M. & Golan, A (2016) What may visualization processes optimize? IEEE Transactions on Visualization and Computer Graphics 22:12, 2619–2632.

[58] Tourangeau, R. (1984). Cognitive science and survey methods. *Cognitive Aspects of Survey Methodology: Building a Bridge between Disciplines*, (Eds., T.B. Jabine, M.L. Straf, J.M. Tanur and R. Tourangeau). Washington, D.C.: National Academy Press, 73-100.

[59] Tourangeau, R., Conrad, F. and Couper, M. (2013). *The Science of Web Surveys.* Oxford University Press, New York.

[60] Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). The psychology of survey response. New York: Cambridge University Press.

[61] Tufte, E.R. (1983) The Visual Display of Quantitative Information

[62] Brackstone, G. (1999), Statistics Canada, Survey Methodology, Catalogue No. 12-001-XPB, 1 Vol. 25 No. 2, December 1999

[63] Cleveland, W. S., & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. Journal of the American statistical association, 79(387), 531-554.

[64] Heer, J., & Bostock, M. (2010). Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 203-212). ACM.

[65] Gregory, R.L. (1968). Perceptual Illusions and Brain Models. Proceedings of the Royal Society of London, Series B, Biological Sciences, Vol. 171, No. 1024. A Discussion on the logical Analysis of Cerebral Functions (Dec. 13, 1968), pp. 279-296.

[66] Marr, D. (1982), Vision: A Computational Approach, San Francisco, Freeman & Co.

[67] Kerns, S.H. & Wilmer, J.B. (2021) Two graphs walk into a bar: Readout-based measurement reveals the Bar-Tip Limit error, a common, categorical misinterpretation of mean bar graphs. Journal of Vision, 21: 17, 1-36.

[68] Petersen, A.C., Jansen, P.H.M., Van der Sluijs, J.P., Risbey, J.S., Ravetz, J.R., Wardekker, JA., Martinson Hughes, H. (2013) Guidance for Uncertainty Assessment and Communication 2nd Edition, PBL

[69] Mastrandrea, M.D., Field, C.B., Stocker, T.F., Edenhofer, O., Ebi, K.L., Frame, D.J., Held, H., Kriegler, E., Mach, K.J., Matschoss, P.R. et al. Guidance note for lead authors of the IPCC fifth assessment report on consistent treatment of uncertainties. 2010.

[70] De Jonge, E. (2020), Communicating uncertainties in official statistics — A review of communication methods, Eurostat, Communicating uncertainties in official statistics — A Review of communication methods — 2020 edition - Products Statistical working papers - Eurostat (europa.eu)