Aalto-yliopisto
Aalto-universitetet
Aalto University

CS-E4840
Information Visualization

Assignment 2

Aitor Urruticoechea Puig
aitor.urruticoecheapuig@aalto.fi
Student N°101444219
April 2024

Contents

List of Figures

Acronyms

EAFRD European Agricultural Fund for Rural Development. 2, 4–6

EC European Comission. 2, 7

ERDF European Regional Development Fund. 1, 2, 4–6

ESF European Social Fund. 2

EU European Union. 1–7

1 Exercise 1: EU Open Data

Take a look at the European Union (EU) *Open Data Portal database at* https://data.europa.eu/en/. *From there, download one of the open data sets and visualize some interesting phenomena in the data. Write a short (2–3 pages) report that describes your approach, obtained results, and the methods you used. Explore at least three of the topics discussed in the course (see below) and explain in your report how you have used them.*
The purpose is not to do an analysis that covers the data from all possible angles. Focus on one or two aspects and make a clear visualisation of them. Try to relate your work to the topics discussed in the course, i.e., Tufte's principles, pre-attentive features, Gestalt laws, etc., when preparing the visualisations. Also, make sure your images convey at least some of their information when printed on a black and white printer.
The use of available visualisation tools—e.g., R, Python (with libraries like Matplotlib), Matlab, Illustrator, (Open)Office, etc.—is encouraged. However, implementing your own scripts and small programs for processing the data will probably be necessary.

1.1 Introduction

The EU Cohesion Policy "aims to correct imbalances between countries and regions" in order to "strengthen economic, social and territorial cohesions" throughout the union [1]. To do so, it uses certain strategies of fund allocation to regional entities across the union with different purposes, requisites, or objectives. Overall, however, the policy has the following high-level goals [2]:

1. A more competitive and smarter Europe.
2. A greener, low carbon transitioning towards a net-zero carbon economy.
3. A more connected Europe by enhancing mobility.
4. A more social and inclusive Europe.
5. A Europe closer to citizens by fostering the sustainable and integrated development of all types of territories.

To do so, the European Comission (EC) has historically employed different funding schemes. Most notably, the European Regional Development Fund (ERDF), aiming at creating economic structures in underdeveloped regions (not countries) to bring them up to speed, create jobs of their own. Other than that, the European Agricultural Fund for Rural Development (EAFRD) helps in developing sustainable agriculture, and aims to balance out the inequalities between rural and urban.

While these are the main funding plans, the EU Cohesion Policy does have other mechanisms, like the European Social Fund (ESF) among others, but they are quite reduced in scope when compared to the ERDF and the EAFRD. In this exercise, the author aims to illustrate which EU regions have benefitted the most from these funds, and how the decisions on the data visualization can shape the discourse and the perception the population has of these funding schemes. To do so, the data regarding the all-time sum of funds received by regions [3] is to be visualized, and later on a normalization of that data by present populations [4] is to be performed for better analysis (discussed in Section 1.4). Note that data from the United Kingdom of Great Britain and Northern Ireland is included, as they did receive European funds up to their exit from the community in 2019.

1.2 Continental Europe and overseas territories

When visualizing data on a map, as it is of interest in this exercise, the EU runs into the question on how to represent overseas territories, which in many cases (though notably, not all) are part of the EU due to them being an integral part of their countries; despite the geographical distance in respect to the continent itself. Often times, the solution chosen is to represent these regions in little boxes around the map (see Figure 1). While this solution serves some of Tufte's principles for combining information in a clear and concise way, it does over-represent these regions. With the notable exception of French Guayana, all of them take a larger area than what they actually take on Earth's surface. Population and economic-wise, this is even worse, for they represent a very minor part of both the European population and economy; so their zoomed-in version in boxes places them in a higher-relevance spot despite their actual position in many rankings.

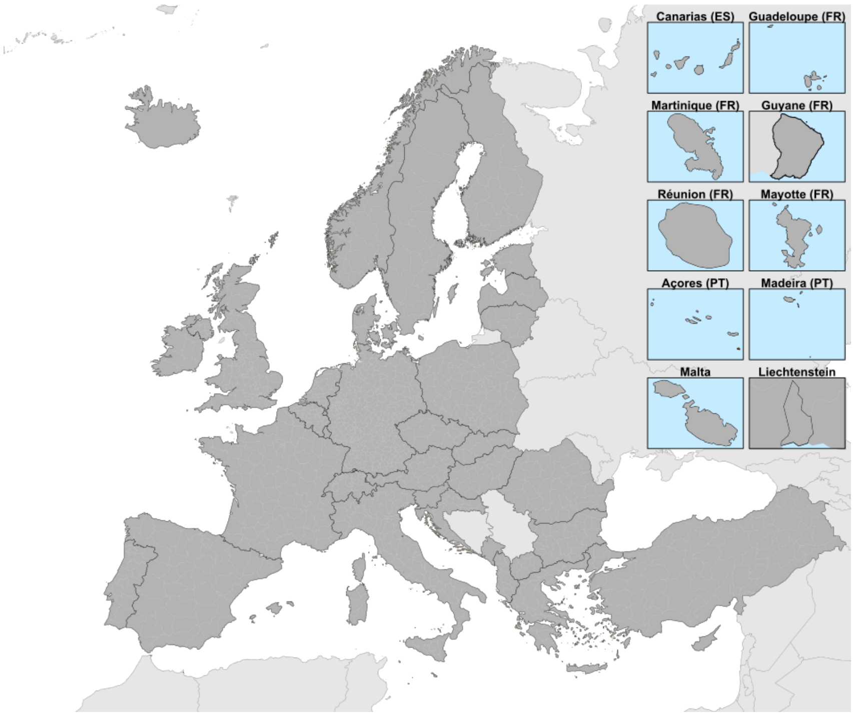


Figure 1: Eurostat Level-3 region map, with overseas territories in boxes.

What is even worse, arguably, from an Information visualization perspective, is the fact that Gestalt laws play against them when choosing to represent them this way. By grouping them all in a corner in little boxes, alongside whole countries that are not even overseas regions like Malta and Liechtenstein, one presumably groups them as a whole and mentally associates them to a single pattern, rather than recognise their own idiosyncrasies. It is hard to argue that the realities of the Canary Islands (the Spanish archipelago off the coast of the Western Sahara) are even remotely comparable to those of La Réunion (the French overseas territory in the Indian Ocean) or the French Guayana in South America. In the same way that it would clearly be misleading to represent the data from Finish Lapland next to the data of the Greek Mediterranean Islands, this representation invites an association of terms detached from actual realities.

On the flip side, if one is to represent the overseas regions and territories in their right place on the world map, the sheer distances and sizes make for a difficult to read map that can hardly follow Tufte's principles (see Figure 2); while adding little to no useful extra information. For that reason, the representation of this Cohesion Policy funds data has been chosen to be only representative of the European continent. While this does mean that overseas regions will not be represented; it also means that focus can be put on the actual effect these policies have had on the continent as a whole without outlier and disproportionate data points from sparsely populated and economically complex overseas regions compromising the usability of the visualizations of the rest of the regions.

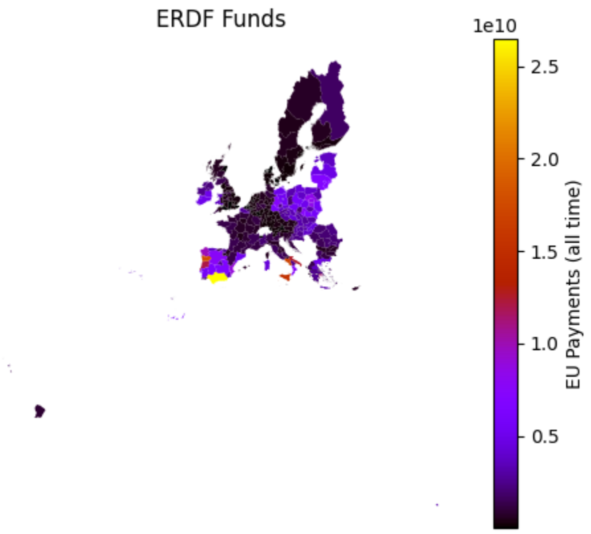


Figure 2: All-time ERDF distribution per regions, including overseas.

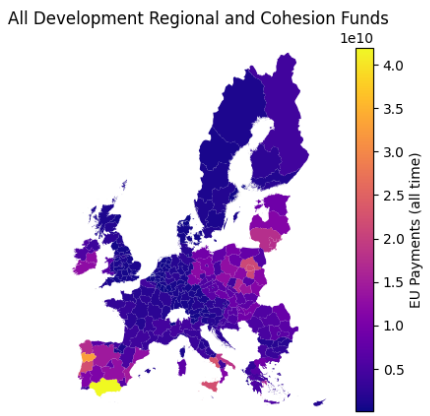
1.3 Colour gradients

When representing the absolute all-time amounts of certain funding programs, deciding on the right colour scale is key in identifying patterns without producing biased views of the results. The distributions observed tend to showcase a lot of regions being in the low-to-mid part of the spectrum when talking about fund allocation; which means the colour scale chosen needs to be able to show the richness and difference in that zone; while displaying outliers in the higher end of the spectrum distinctly enough. Figure 3 showcases some different colour scale options; with a direct comparison to a gnuplot colour scale visualization, the final one chosen.

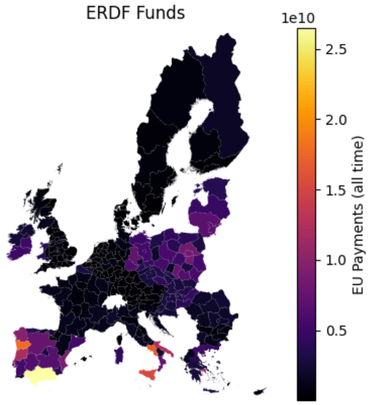
Importantly, gnuplot is able to easily display the nuances of the aforementioned low-to-mid parts of the spectrum, unlike plasma and inferno colour scales shown in Figures 3b and 3a respectively. Added to that, no part of the gnuplot spectrum is light enough to be easily confused with the sea or non-EU countries, like YlGnBu (Figure 3c); while it does not sacrifice colour variation and ease of translation to black-and-white like viridis (Figure 3d).

1.4 Population normalization

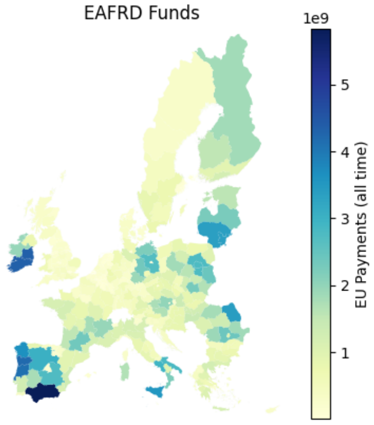
An easy trick one could easily pull off here is to finish the discussion with the colour scales and the like; but that would be disingenuous at best and with malicious intentions at worse. The EU is quite the diverse place; and no two regions are the same. When looking at funding allocation, this can easily show some regions as being receptive of disproportionate amounts of funds; yet one must put that into context. An easy and direct way to do so, to some extent, is to normalize the data by population (per capita received funds, as shown in Figure 4). This is an incomplete solution, as funding schemes like the EAFRD aim precisely at combating the inequalities faced by the dwindling populations of the rural regions; and will naturally not be favouring more densely populated regions (see Figure 4c). Yet, this action can easily help in showcasing which regions might not actually be huge outliers when one takes into consideration how many people are to actually feel the effects of such schemes.



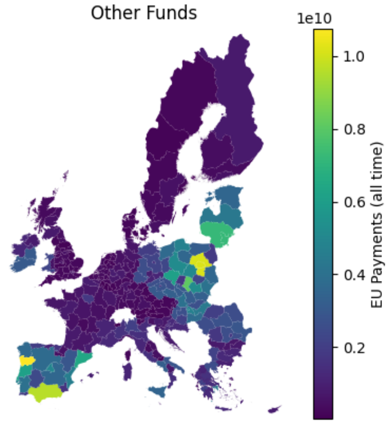
(a) All funds, in plasma colour scale.



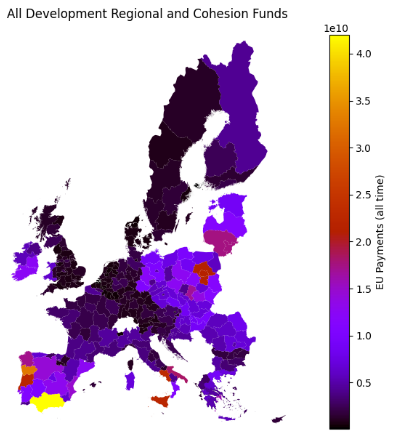
(b) ERDF, in inferno colour scale.



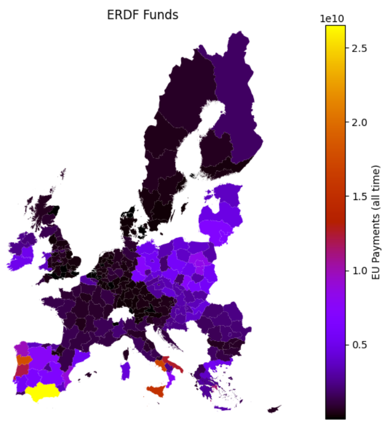
(c) EAFRD, in YlGnBu colour scale.



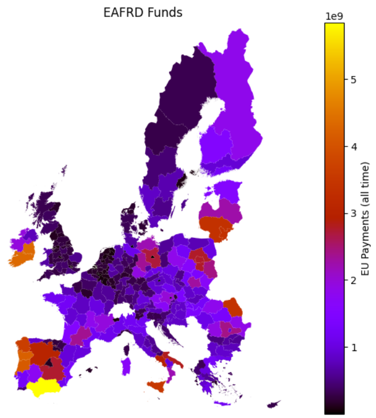
(d) Other funds, in viridis colour scale.



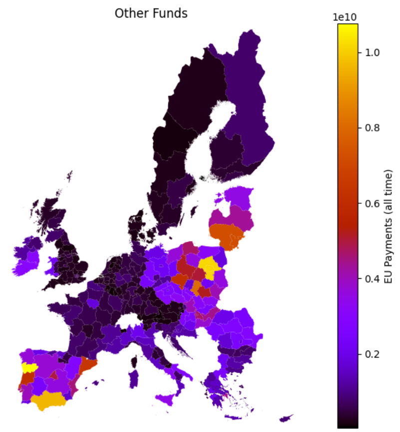
(e) All funds, in gnuplot colour scale.



(f) ERDF, in gnuplot colour scale.

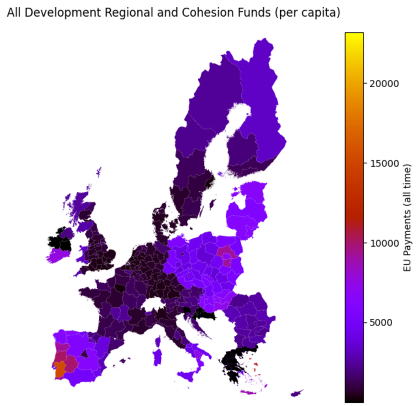


(g) EAFRD, in gnuplot colour scale.

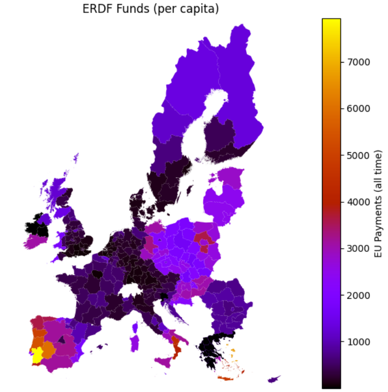


(h) Other funds,in gnuplot colour scale.

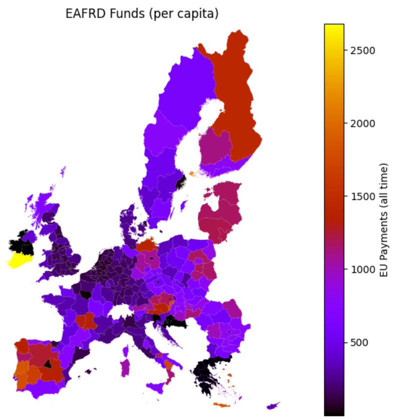
Figure 3: Colour scale comparison, using various EU fund distributions.



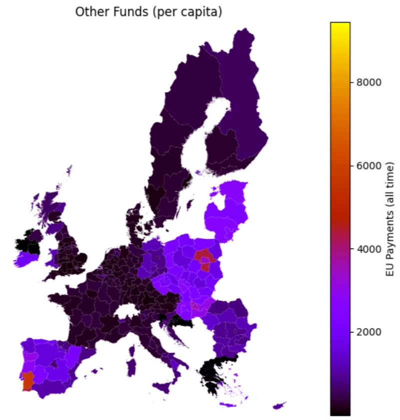
(a) All funds received (per capita).



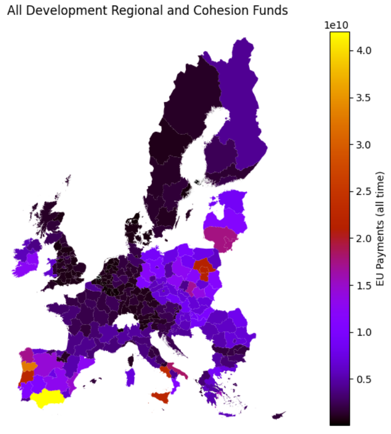
(b) ERDF payments (per capita).



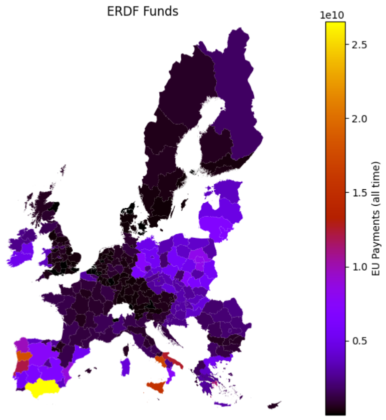
(c) EAFRD payments (per capita).



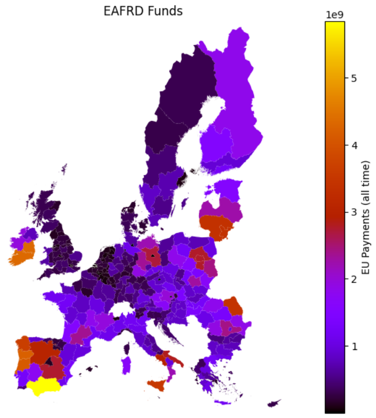
(d) Other funds payments (per capita).



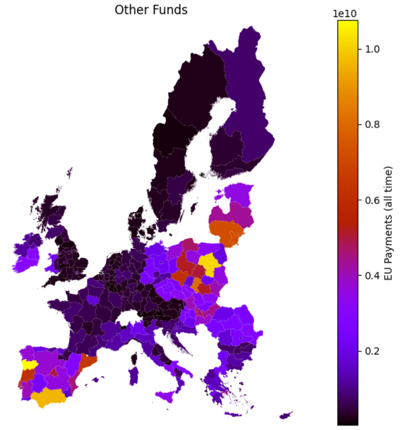
(e) All funds, in absolute terms.



(f) ERDF, in absolute terms.



(g) EAFRD, in absolute terms.



(h) Other funds, in absolute terms.

Figure 4: Region-by-region all-time EU regional and cohesion funds payments, in absolute and per-capita terms.

1.5 Conclusions

Taxpayer money allocation is a difficult discussion and can easily become controversial. This exercise shows how easily the conversation can be changed and stirred to widely different directions if visualizations are not chosen properly and carefully enough. Different maps, colour gradients, and data treatments can get the best of human cognitive abilities and be used to push different agendas if one is not aware of tricks like these. It is interesting to see how, nonetheless, where the divides exist in the EU once the data is properly analysed, to understand the motivations behind the fund allocations done by the EC.

The most clear divide seems to still be the East/West divide; and, most surprisingly, places typically thought of as worse like continental Greece and rural Ireland seem to actually be receiving fewer funds per capita than what one may think initially. Other than that, there are less surprising but still consistent trends: the North/South divide in Italy, the wide variety of realities of the Iberian Peninsula (with Catalonia and the Basque Country receiving notably less aid than Central Portugal or Andalusia), the East/West divide of Finland (with Northern Finland being a notable beneficiary from many of these funds), or the well-known German East/West not being so unified yet.

2 Exercise 2: Simultaneous brightness and contrast effect

Discuss the simultaneous brightness and contrast effect shown in Figure 5 and explain it using the Difference of Gaussians model. Your answer does not have to contain any mathematical calculations. Instead, try to give a clear explanation why the rectangles appear to be of a different gray tone while they in fact are all of the same "color." What are the implications of this phenomena for the design of color scales?.

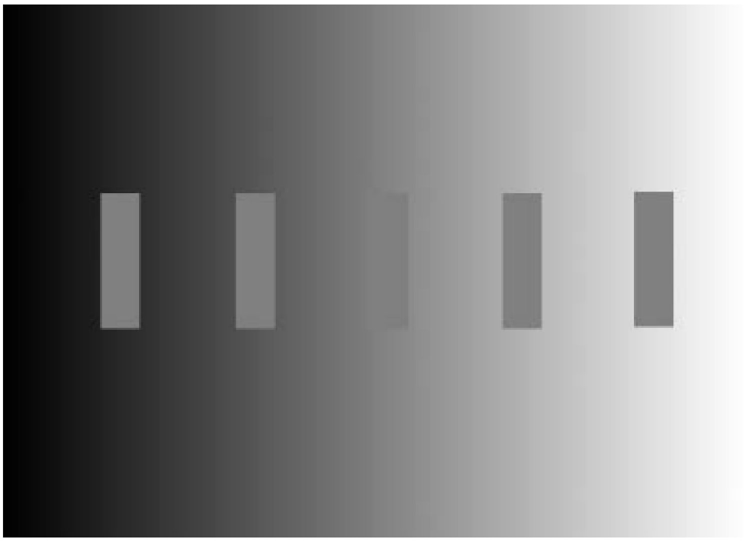


Figure 5: Simultaneous brightness and contrast effect.

The simultaneous brightness and contrast effect is a visual phenomenon. It describes how the perceived brightness or darkness of a region is inevitably influenced by its surroundings. This is quite clearly seen in Figure 5 where, despite all the grey rectangles being of the same gray colour, they are seen as different by a human eye.

The Difference of Gaussians model suggests that the visual system processes images through the comparison of blurred versions of the image at different scales. The actual model is a mathematical tool that acts similar to a band-pass filter, creating contours in an image where the contrast between colours is the highest. The steps to get that result involve taking the difference between two Gaussian-blurred versions of the image to accentuate edges and contrasts; which ultimately results in a black-and-white image where the perception of edges is enhanced while suppressing noise.

If one tries to understand human vision as an apparatus that is constantly performing a difference of Gaussians to everything the eye sees, the effect seen in Figure 5 is quite easier to understand. Due to the effect of perceiving the contours of highly-contrasted objects, regions with higher contrast edges against their surrounding context will appear brighter, while regions with lower contrast edges will appear darker. This has, of course, inevitable consequences in the design of colour scales. If one is to design one, they must be mindful of how the surrounding context influences the perception of colours. For example, when creating colour scales for data visualization, it's essential to consider how adjacent colours interact and how they might influence each other's perceived brightness or darkness. Otherwise, areas with the same value and thus associated with the same colour can be perceived differently if their surroundings are not exactly the same (something that rarely happens when visualizing data).

3 Exercise 3: Colour Scales

Figure 6 is a map of Europe. If you were to show statistical data for comparing the countries, which scale (Greys, Spectral or YlGnBu) would you use? Justify your answer with at least two reasons based on the functioning of the human eye.

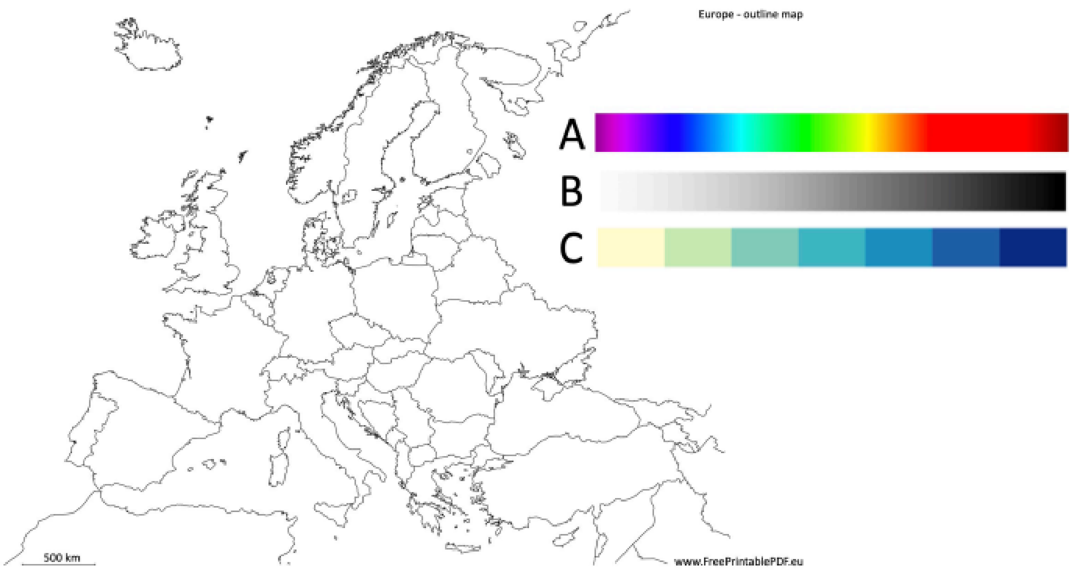


Figure 6: Three colour scales (A: Spectral, B: Greys, C: YlGnBu) and map of Europe with country borders.

In the case of map painting, and while this answer will always be context-dependant according to where and how the data is to be displayed, and what the data is actually talking about; it seems to be generally better to use YlGnBu. It avoids bias and offers a perceived uniformity in a way that the other two options do not.

A major point favouring the YlGnBu scale is designed to be perceptually uniform, meaning that the perceptual difference between adjacent colours is consistent. Since the human eye is more sensitive to differences in brightness than differences in hue or saturation; the YlGnBu scale is an obvious choice, as it takes this as a strength without sacrificing variability and the ability to easily discern variations. This is something that the spectral scale clearly fails at; as it completely lacks uniformity. Yes, it is visually appealing, but since the human perception of colour is non-linear, colours may appear more dominant or visually distinct than others; thus misleading viewers. On the flip side, the Greys Scale does have consistency, but does so by sacrificing hue variation. While great for black-and-white printings, it lacks the diversity provided by colours across the spectrum. Thus, its usage may result in a less informative visualization, as it can be more challenging for viewers to differentiate between data points.

The YlGnBu scale also provides with a good balance between warm and cool colours. Unlike the Spectral scale, which contains both warm and cool colours, YlGnBu avoids the potential confusion that can arise from using diverging scales. Added to that, the spectral scale plays the human cognition in another way: since the human eye tends to associate warmer colours with positive values and cooler colours with negative values. However, when comparing countries on various metrics, using a scale that does not bias perceptions is crucial in allowing a nuanced analysis. YlGnBu's progression from yellow to green to blue maintains a neutral stance, reducing the risk of misinterpretation based on colour biases.

4 Exercise 4: Multivariable glyph

Design a glyph that enables the pre-attentive perception of at least three variables (discrete or continuous). How many variables can you represent with your glyph, and what can you say about how the different variables are perceived?

The proposed glyph aims to solve some of the issues found in Assignment 1 regarding the penguins dataset [5]. In this dataset, each point can be associated with one penguin, that has three discrete variables associated to it:

- **Sex**: either female or male
- **Island**: Torgersen, Biscoe, or Dream
- **Species**: Adélie, Gentoo, or Chinstrap

The proposed glyph, as seen in Figure 7, aims to offer an adaptative and intuitive solution to the representation of this multivariable reality. The outline colour of the circle is used to represent the sex. Initially, the proposal is to use cyan for female and yellow for male; but this could be expanded using the colours between these two if one wanted to represent sex as a continuous variable for the whole sex spectrum.For the island, the left side of the circle is used; with different colours and different sixths of the circle associated to the different islands. The same is used for the species on the right side of the circle. See Figure 8 for an example on how to use the proposed glyph.

By using two different factors to identify the characteristics of Island and Species, the precognition is maximised, as not only colour but also location are used as a key to reading the glyph. Interestingly, this also allows for easy conversion to a black-and-white printing, as well as easy support for color-blinded people. It is also worth noting that a conversion to a continuous scale is also theoretically possible for the Island and Species variables, by partially filling the circular sections associated with each one. It is also possible to add more variables and/or more island and species variability into the glyph; by simply splitting the circle into more sections and/or subsections. This has, of course, practical limitations. For human cognition reasons, the circle can only be split so much before it becomes a useless tool for displaying information. 12, maybe 16 subdivisions, would probably be the highest one could reasonably get before the colours and sectors mix too much with one another and render the glyph unreadable.

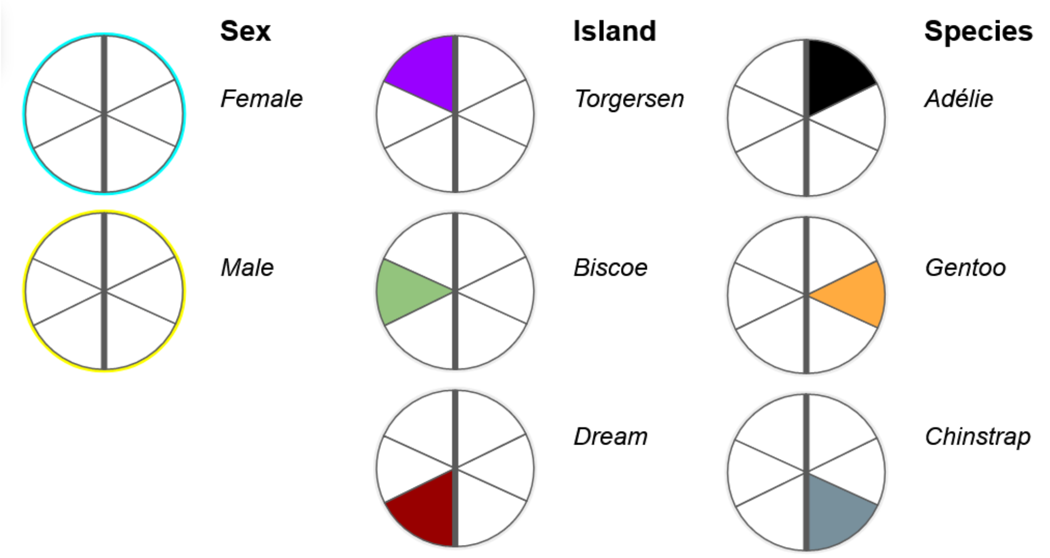


Figure 7: Proposed glyph for the penguins database.

Figure 8: Example of usage for the proposed glyphs

References

1. *New Cohesion Policy* [online]. European Comission, 2021 [visited on 2024-03-25]. Available from: https://ec.europa.eu/regional_policy/2021-2027_en.
2. *Priorities for 2021-2027* [online]. European Comission, 2021 [visited on 2024-03-25]. Available from: https://ec.europa.eu/regional_policy/policy/how/priorities_en.
3. *EU Cohesion Policy : Historic EU payments - regionalised and modelled* [online]. Open EU Data, 2022 [visited on 2024-03-25]. Available from: https://data.europa.eu/data/datasets/eu-cohesion-policy-historic-eu-payments-regionalised-and-modelled?locale=en.
4. *Population statistics at a regional level* [online]. Eurostat, 2023 [visited on 2024-03-25]. Available from: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Population_statistics_at_regional_level.
5. HORST, Allison Marie; HILL, Alison Presmanes; GORMAN, Kristen B. *palmerpenguins: Palmer Archipelago (Antarctica) penguin data*. 2020. Available from DOI: 10.5281/zenodo.3960218. R package version 0.1.0.