

Wel binnen de lijntjes blijven! Over kleur in visualisaties

Robin Marx
Hasselt University
marx.robin@gmail.com

ABSTRACT

In deze paper worden drie verschillende visualisaties voor items op een landkaart besproken. Elk van deze visualisaties maakt veel gebruik van kleur om het mogelijk te maken verschillende items met elkaar te vergelijken en van elkaar te onderscheiden. De concrete kleurkeuzes per visualisatie worden besproken en gemotiveerd. Er wordt geconcludeerd dat kleur een krachtige tool is om gebruikers te helpen bij het interpreteren van visualisaties, maar dat er veel tijd en aandacht moet worden besteed aan de juiste kleurkeuze, wat niet altijd even makkelijk is.

Author Keywords

infovis, kleur, kleurschema's, landkaart, vergelijken

ACM Classification Keywords

H.5.2 Information Interfaces and Presentation: Miscellaneous

INLEIDING

Kleur is een krachtige maar ook gevaarlijke tool bij het weer-geven van data in visualisaties. Er moet rekening gehouden worden met contrast, betekenis, zichtbaarheid etc. Als ze goed gebruikt worden kunnen ze echter helpen een visua-lisatie duidelijker te maken.

Het doel van deze paper is om kleuren te gebruiken in vi-sualisaties waarbij het vergelijken van verschillende items belangrijk is. In deze visualisaties gebruiken we kleur om items op een landkaart te plotten. Elk item staat voor één gemeente en we proberen deze verschillende gemeenten met elkaar te vergelijken.

De vraag is hoe kleur kan helpen om deze vergelijkingen te vergemakkelijken. Dit wordt gedaan in drie verschillende contexten waarbij de kleuren telkens een andere betekenis krijgen toegekend. Na een overzicht van enkele bronnen be-spreek ik eerst alle visualisaties, waarna ik de kleurkeuzes voor elke visualisatie apart onder de loep neem.

GERELATEERD WERK

Het idee om kleur te gebruiken in data visualisaties is natu-urlijk niets nieuws. Zo wordt kleur vaak gebruikt om data onder te verdelen in categorieën of verzamelingen, om ver-schillende datasets tegelijk te laten zien of om bepaalde items te highlighten. Anderzijds zijn er ook heel wat visualisaties die toelaten items met elkaar te vergelijken. We denken hier-bij bijvoorbeeld aan een parallel coordinate plot, starplot, piechart, landkaarten, etc.. Een mooi voorbeeld dat kleur en vergelijking koppelt is figuur 1.

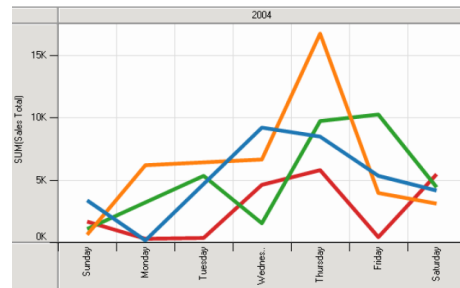


Figure 1. Elke lijn stelt een andere dataset voor, de kleuren helpen bij het onderscheiden. Afkomstig uit [4].

Brewer beschrijft enkele kleurschema's die in verschillende situaties kunnen ingezet worden [2] en haar website Color-Brewer [3] wordt vaak vermeld in verband met het kiezen van goede kleurschema's. Stone [4] geeft enkele inter-essante voorbeelden van verschillende kleurschema's. Eén ding waar alle schrijvers het over eens zijn, is het advies van Tufte: "avoiding catastrophe becomes the first principle in bringing color to information: Above all, do no harm." [5], waarmee hij wilt zeggen dat er moet opgepast worden met het gebruik van kleur omdat er snel dingen mis kunnen gaan.

VISUALISATIE 1 : NUMBERRENDERER

Een eerste visualisatie moet voor elk item een getal weer-geven. Hoge getallen stellen voor dat de locatie het beter doet dan locaties met lagere getallen.

Zoals te zien in figuur 2 zijn er een aantal verschillende mo-gelijkheden om dit probleem te visualiseren. Ik heb ervoor gekozen ze allemaal tegelijk te gebruiken zoals te zien in d. Interessant is hier dat er eigenlijk univariate data (één getal per item) wordt weergegeven op drie verschillende manieren, terwijl we meestal juist andersom werken: verschillende di-mensies in de data allemaal op een andere manier weergeven. Een interessante vraag is dan ook of deze combinatie extra

meerwaarde met zich meebrengt en of we eventueel één onderdeel kunnen weglaten.

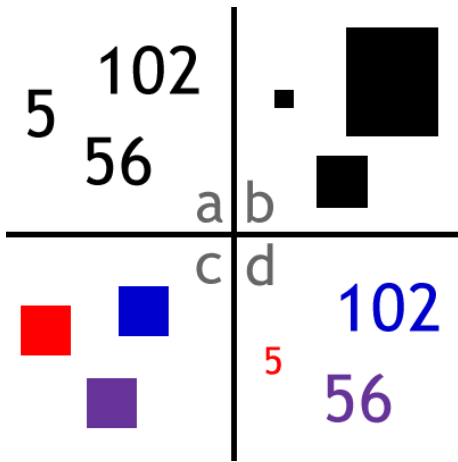


Figure 2. a. getallen in zelfde grootte, b. getallen als oppervlakten, c. getallen als kleur, d. NumberRenderer

Weergegeven in grote aantallen op een volledige landkaart ziet de NumberRenderer er beter en overzichtelijker uit dan de andere drie alternatieven. De belangrijke en minder belangrijke items zijn in één oogopslag duidelijk en outliers kunnen snel gevonden worden. Het weglaten van kleur uit de NumberRenderer maakt het geheel minder aantrekkelijk. Het weglaten van het getal en werken met symbolen lijkt onlogisch omdat we extra concrete informatie verliezen en er maar weinig voor terugkrijgen.

Een belangrijk probleem is hier de opbouw van de data. Indien er veel uitschieters zijn gaan de heel grote en heel kleine items duidelijk zijn, maar alle items rond het gemiddelde gaan ongeveer even groot zijn en dezelfde kleur hebben. Dit komt omdat de kleur en grootte van de items relatief is t.o.v. de ganse beschouwde dataset. We verliezen dus heel wat detail als de data niet mooi uniform is. Een oplossing hiervoor is de PlotSlider. Deze gaat alle waarden plotten op een verticale schaal. Via een slider kunnen de gebruikers bepalen welke gevallen moeten worden gefilterd. Hierdoor kan er een meer uniforme verdeling worden bekomen waardoor de visualisatie beter weergeeft wat er rond het gemiddelde gebeurt en waarbij de kleuren en groottes anders worden verdeeld, zoals te zien is in figuur 3.

Tijdens usertests werd de NumberRenderer vergeleken met een traditionele heatmap. De proefpersonen vonden de heatmap interessanter voor een totaaloverzicht maar hadden toch liever de NumberRenderer als ze echt dieper moesten ingaan op de data. De PlotSlider werd ook goed onthaald en gebruikt om nadruk te leggen op bepaalde segmenten van de data. De plotslider plot alle waardes echter in één kleur. Een interessante aanpassing zou zijn om de kleuren van de items op de PlotSlider gelijk te maken aan de kleuren van de overeenkomstige items op de kaart, voor een betere conceptuele mapping tussen de twee.

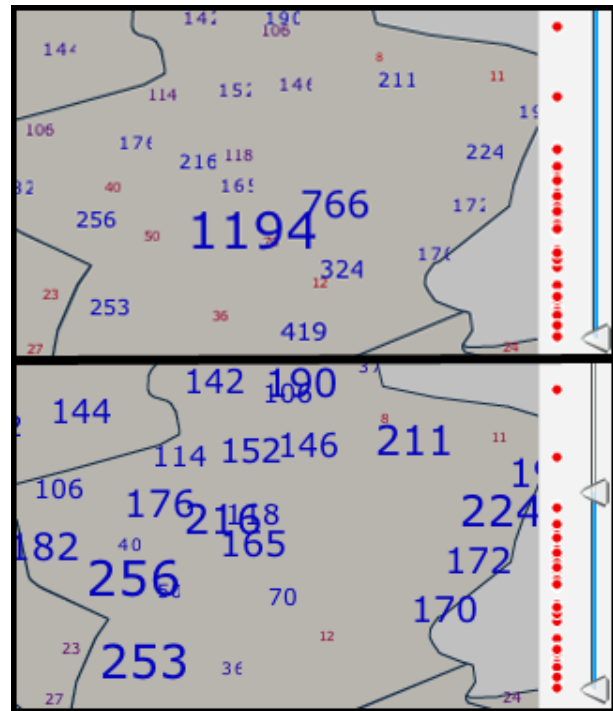


Figure 3. Boven: alle data weergegeven. Onder: hoge waarden gefilterd met plotslider.

VISUALISATIE 2 : ONE COLOR CATEGORY

Deze tweede visualisatie had als doel voor elke gemeente weer te geven welke studierichting daar het meest populair was. Eén mogelijkheid is om voor elke gemeente een piechart te tonen met daarin alle richtingen. Dit zal echter niet altijd even duidelijk aantonen welke nu de grootste is. De gekozen mogelijkheid is om elke studierichting een andere kleur te geven en voor elke gemeente enkel de kleur van de populairste richting te tonen. Dit geeft iets minder informatie maar vermindert de clutter en vergemakkelijkt de interpretatie. De kleuren werden tevens zo gekozen dat studierichtingen uit dezelfde faculteit gelijkaardige kleuren kregen. Dit om hun gelijkenissen te benadrukken en het verschil met andere faculteiten in de verf te zetten.

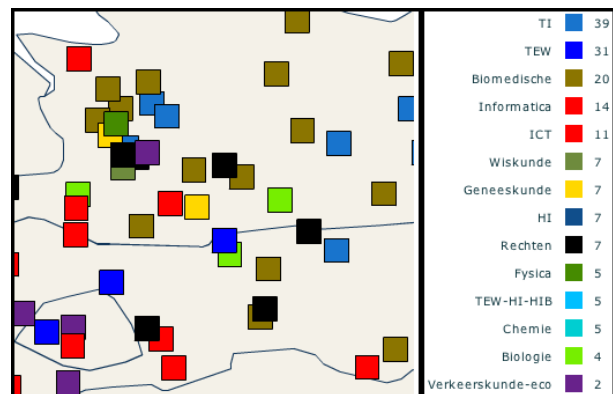


Figure 4. Links: de verschillende items, Rechts: detailweergave voor één gemeente. De kleuren worden herhaald om de koppeling met de landkaart makkelijker te maken.

Deze aanpak heeft het voordeel dat ze snel outliers toont en tendenzen in bepaalde gebieden makkelijk duidelijk maakt. Zo komen er in figuur 4 opvallend meer studenten biomedische uit Antwerpen dan uit de regio rond Brussel. Een nadeel is dat sommige items onderdrukt worden als er veel scholen in hetzelfde gebied liggen.

Ik heb geprobeerd om ook hier de grootte van de items aan te passen naarmate er meer studenten uit die gemeente kwamen. Dit gaf echter mogelijk een fout beeld omdat de aandacht daar dan misschien meer uitgaat naar het aantal studenten en minder naar de tendenzen qua studiekeuze per gemeente en per regio. In dat opzicht is de visualisatie echter ook niet altijd optimaal. Gemeentes die twee studenten naar een richting sturen staan naast gemeentes die er 50 sturen waardoor er misschien een vertekend beeld wordt gegeven over het pure aantal studenten dat een richting begint.

De usertests waren zeer positief over deze visualisatie. Het was meteen duidelijk wat werd weergegeven en de detailweergave was volledig wat er verwacht werd. Het feit dat een vertekend beeld mogelijk is werd echter ook opgemerkt.

VISUALISATIE 3 : ONE COLOR EVOLUTION EN GRADIENTSPARKLINE

Het doel van deze visualisaties was om de evoluties in studentenaantallen per gemeente weer te geven. De evolutie per gemeente was op zich niet zo'n probleem want daarvoor bestaan veel mogelijkheden zoals columncharts, sparklines, line charts etc. De vraag was hoe we de evoluties konden weergeven op de landkaart zodat het overzichtelijk bleef en zodat de verschillende evoluties met elkaar konden worden vergeleken.

Gedetailleerde charts plotten per gemeente geeft veel clutter en mogelijk ook teveel detail. Zelfs wanneer we enkel een sparkline per gemeente zouden plotten is het moeilijk de items te vergelijken. Een oplossing is om het idee van de One Color Category over te zetten naar evoluties, zie figuur 5. Als er een algemeen positieve evolutie heeft plaatsgevonden is de kleur groen, als er een algemeen negatieve evolutie was, is het rood. Deze aanpak geeft weinig concrete details maar laat wel heel goed de tendenzen zien en wel op zo een manier dat we ze goed met elkaar kunnen vergelijken. Zo zijn rode items in Limburg waarschijnlijk belangrijk om eens van dichtbij te bekijken omdat we daar toch vooral groene items verwachten.

Het is belangrijk hoe we precies bepalen of de algemene evolutie positief of negatief is geweest. We kunnen bijvoorbeeld kijken naar de eerste en de laatste waarde, het aantal stijgingen en dalingen over de ganse periode, of er de laatste paar jaar een goede stijging is geweest etc. Deze keuze bepaalt of een item rood of groen wordt en dus ook of het door de gebruiker aanzien wordt als belangrijk of niet. De specifieke keuze is dus data afhankelijk en afhankelijk van waarop we precies de nadruk willen leggen. Voor mijn implementatie bekijk ik het aantal dalingen en stijgingen van jaar tot jaar over de gehele periode van de data. Meer dalingen geeft een rode kleur, stijgingen geeft groen.

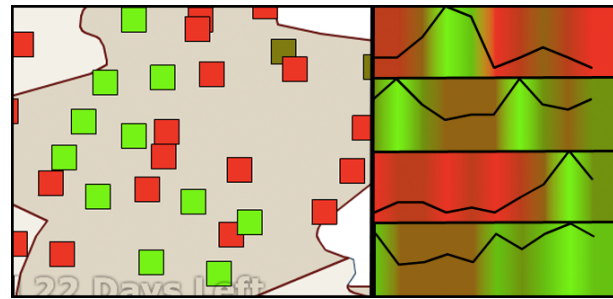


Figure 5. Links: éénkleur per item duidt de algemene evolutie aan, Rechts : GradientSparklines over 4 verschillende datasets.

Het naderbij bekijken van items gebeurt op twee manieren. Ten eerste is er een detailweergave per item waarbij de evolutie wordt geplott in een columnchart. Deze visualisatie geeft zeer concrete detailinformatie voor het item waardoor het nadeel van de One Color renderers een beetje wordt tegengegaan. Een tweede detailweergave zien we als we inzoomen op de landkaart. Als we dit ver genoeg doen veranderen de items in zogenaamde Gradient Sparklines. Deze manier van visualisatie is geïnspireerd door de gewone sparkline [6] maar stelt deze voor in combinatie met een gradient, een overvloeiing van kleuren. Hierbij heeft elk jaar (de x-as) een kleur die voorstelt hoe het studentenaantal dat jaar was t.o.v. het vorige jaar.

De GradientSparkline was de meest experimentele van de visualisaties. Hun doel is om de gebruiker toe te laten snel belangrijke items te identificeren (bijvoorbeeld de items die veel rood bevatten) en het daarna makkelijk te maken om ze concreet met elkaar te vergelijken. Het identificeren werkt relatief goed omdat de kleuren visueel zo sterk staan dat gebruikers makkelijk grote vlakken rood of groen kunnen onderscheiden. Het vergelijken werd iets bemoeilijkt omdat de kleuren niet genoeg contrasterend zijn om kleinere details te bekijken, zeker in de gebieden die niet duidelijk groen of rood zijn en overgangen. Een oplossing hiervoor is om de gewone sparkline over de gradient te tekenen, zie figuur 6. Dit geeft de gebruikers de mogelijkheid om snel een globaal idee te krijgen van de evolutie door de kleuren en daarna meer in detail te kijken door de sparkline zelf. Een andere oplossing zou zijn om de kleuren niet in elkaar te laten overvloeien en zo eerder een discreet in plaats van continu overzicht te geven van de opeenvolgende jaren.

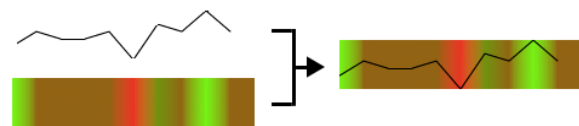


Figure 6. Linksboven: een gewone sparkline, Linksonder: een gradient, Rechts : combinatie van sparkline met gradient.

Hoewel de GradientSparklines dus redelijk goed werken om evoluties van verschillende items met elkaar te vergelijken bleek het toch moeilijk om een mooie visualisatie te creëren met enkel de GradientSparklines op de landkaart. Er is dan

gewoon te veel informatie. Daarom werd gekozen om de One Color Renderers te gebruiken als totaaloverzicht, waarbij men kan inzoomen op de GradientSparklines. Hierdoor worden er minder sparklines tegelijk getoond wat de vergelijking makkelijker en duidelijker maakt. Gecombineerd met de detailweergave geeft dit drie lagen van abstractie waarmee veel informatie over de evoluties kan verzameld worden.

De gebruikers die de usertests deden waren duidelijk gecharmeerd door de One Color renderers, net zoals door de One Color renderers voor richtingen. De GradientSparklines waren iets minder interessant en er werd geopperd dat de gewone sparkline echt wel nodig was om veel concreet nut te hebben. Hiernaast ondersteunden ze echter wel de theorie dat de kleuren van de gradients het makkelijk maken om snel belangrijke delen te identificeren en dat dit moeilijker zou zijn zonder de kleuren.

KLEURKEUZES

De kleurkeuze voor de evoluties was redelijk makkelijk en duidelijk. Rood en groen zijn niet enkel twee kleuren die hard met elkaar contrasteren en dus duidelijk van elkaar te onderscheiden zijn, ze hebben ook een sociale conventie die de visualisatie helpt. Rood staat vaak voor slecht, terwijl de tegenhanger groen iets positief aanduidt. Gebruik maken van deze conventie helpt de gebruikers om snel te snappen dat negatieve evoluties in het rood worden weergegeven. Dit gebeurt automatisch als de enige andere gebruikte kleur groen is en de gebruiker weet over wat voor data het gaat. Dit is een voorbeeld van een diverging color scheme [2]. Hoewel Brewer deze beschrijft met een neutrale kleur in het midden, past het bruinige van het gemengde rood/groen wel bij deze visualisatie en contrasteert het genoeg met het rood en groen om duidelijk te maken dat het ergens tussenin ligt.

Een totaal andere aanpak werd ook geprobeerd in het project, waarbij enkel zwart en wit werden gebruikt in de sparklines. Hierdoor bekomen we gradients van grijswaarden. Deze zijn minder prominent aanwezig op het scherm (de rood/groene sparklines komen soms zeer fel over) en geven volgens mij toch nog altijd een mooie aanduiding van de hoge en lage waardes. Het gebruik van grijswaarden wordt afgeraden omdat het soms ongewenste artefacten geeft, maar aangezien het bij de gradients toch niet de bedoeling is dat er gedetailleerde informatie uit kan worden afgeleid speelt dit minder een rol dan de intuïtieve aanvoeling dat de witte gebieden positief zijn versus de "lagere" zwarte gebieden. Het was een interessant experiment en variatie op het rood/groene kleurschema.

De keuze van rood en blauw voor de NumberRenderer was moeilijker. Gewoon rood en groen nemen zou hier juist ongewenste informatie kunnen verspreiden. Het is niet omdat een gemeente weinig leerlingen stuurt dat ze daarom slecht bezig is. Het aantal leerlingen is immers gekoppeld aan de afstand tot de universiteit (een feit waar een andere visualisatie, de CircleChart, de nadruk op legt). Rood en blauw zijn daarom interessantere keuzes omdat ze eveneens goed contrasteren met elkaar, maar er minder sociale conventies aan gekoppeld zijn die fouten in de interpretatie zouden kunnen opleveren. Het gekende rainbow-kleurschema

[1], waarbij lage waardes blauw zijn, gemiddelde geel en hoge rood, is een andere mogelijkheid. Dit werkt volgens mij echter beter op continue data(plots).

De keuzes voor kleuren voor de verschillende richtingen was de minst exacte. Het is duidelijk dat bijna alle mogelijke verschillende basiskleuren moesten gebruikt worden om het grote aantal richtingen te kunnen weergeven en goed onderscheidbaar te maken. De vraag is welke kleur we kiezen voor welke richting. De faculteiten zijn nu onderverdeeld in de duidelijkst contrasterende kleuren, waarbij de richtingen per faculteit een variatie op de faculteit-hoofdkleur krijgen. Zo heeft de faculteit wetenschappen groentinten, de faculteit economie blauw en informatica wordt aangeduid met rood. Op deze manier behouden we genoeg contrast om uitschieters tussen verschillende faculteiten te bekijken, maar zijn de richtingen van dezelfde faculteit gelijkend genoeg om te identificeren dat ze iets met elkaar te maken hebben. Het geeft de gebruiker ook minder moeilijkheden om alle verschillende kleuren te onthouden, waardoor hij minder moet teruggrijpen naar de legende en dus meer met de data zelf bezig kan zijn. Een mogelijke verbetering hier is dat volgens Brewer [2] de kleuren binnen de faculteit best enkel van de basiskleur verschillen in lightness, terwijl er nu soms ook verschillen in hue en saturation bijzitten. Ook zijn er reeds bestaande kleuren voor de verschillende faculteiten die verschillen van de gebruikte kleuren in de visualisatie. Het vervangen van de faculteitkleuren door de intern gebruikte kleuren zou interessant zijn voor gebruikers die reeds gewend zijn om met de kleur-faculteit associatie te werken.

Een interessante anekdote is dat er elders in het project de mogelijkheid werd geboden aan de gebruikers om de gebruikte kleuren aan te passen en waarbij enkele standaardkleuren als startwaarden werden gegeven. Tijdens de usertests bleek dat de gebruikers geen nood hadden om de kleuren zelf aan kunnen te passen en dat ze ook zeker geen vereiste feature voor het programma vonden. Dit maakt een goede kleurkeuze van de applicatie-ontwerper natuurlijk nog belangrijker, omdat men moeilijk van de gemiddelde gebruiker kan verwachten dat ze zelf een gepaste kleurschaal opstellen.

CONCLUSIE

Kleuren kunnen heel krachtig zijn in visualisaties en de gebruiker goed helpen met het trekken van de benodigde conclusies. Vooral voor het onderscheiden van verschillende groepen, schakeringen of categorieën van data zijn kleuren handig. Zo kunnen ook ze goed ingezet worden in visualisaties waarbij verschillende datasets met elkaar moeten vergeleken worden.

Tijdens het project gebruikte ik drie verschillende visualisaties die elk kleur proberen te gebruiken om de gebruiker in één oogopslag een overzicht te kunnen bieden en het daarna mogelijk te maken de verschillende items goed met elkaar te vergelijken. Elk van deze visualisaties had zijn plus- en minpunten maar telkens was toch duidelijk dat kleuren een goed middel waren om de beoogde data voor te stellen op een bruikbare manier. Hierbij werd vaak gebruik gemaakt van

een abstractie waardoor een deel van de data werd samengebundeld tot één kleur. Dit leverde duidelijkere en minder drukke visualisaties op en hielp de gebruikers focussen op de interessante delen, waarna via detailweergaves nog kon worden overgegaan tot het bekijken van de gedetailleerdere data.

Belangrijk is hierbij de kleurkeuze en ook welke kleuren samen met elkaar worden gebruikt. Zo kan men handig gebruik maken van sociale conventies om bepaalde conclusies te vergemakkelijken. Maar zelfs als er kleuren gekozen worden is het soms nog moeilijk om deze correct toe te passen binnen de visualisatie, zoals bij de One Color Renderer voor evoluties waarbij de keuze voor rood of groen voor één item zeer belangrijk is. Dit volgt uit het feit dat er door slechts een beperkt aantal kleuren te gebruiken heel wat informatie verloren gaat. Het is dus altijd een uitdaging om de data te presenteren met een beperkt aantal kleuren en er toch voor te zorgen dat de gebruiker geen foute conclusies gaat trekken.

REFERENCES

1. D. Borland and R. M. Taylor. Rainbow color map (still) considered harmful. In *IEEE Computer Graphics and Applications*, 2007.
2. C. Brewer. Color use guidelines for data representation. In *Proceedings of the Section on Statistical Graphics*, 1999.
3. C. Brewer. Colorbrewer. In www.colorbrewer.org, 2011.
4. M. Stone. Choosing colors for data visualization. In www.perceptualedge.com/articles/b-eye/choosing_colors.pdf, 2006.
5. E. Tufte. *Envisioning Information*. 2000.
6. E. Tufte. *Beautiful Evidence*. 2006.