# Focus en context voor afstandsafhankelijke data: de circle chart visualisatie

**Nick Michiels**
Universiteit Hasselt
Master student
Nick.Michiels@student.uhasselt.be

**ABSTRACT**
Deze paper beschrijft een methode om afstandsafhankelijke data op een betrouwbare manier weer te geven op een circle chart visualisatie. Met afstandsafhankelijk wordt bedoeld dat de data gecorreleerd is met de afstand tot een bepaalde plaats. Door middel van een distortie functie wordt er getracht dit probleem op te lossen. De algemene taxonomy over focus en context van Y.K. Leung en M.D. Apperley [2] zijn hiervoor zeer bruikbaar. Bijkomende dimensies van de informatie worden gegeven op de manier waarop de data op de circle chart visualisatie wordt geplaatst. In de visualisatie wordt de data ontkoppeld van de geografische ligging.

**Author Keywords**
Informatievisualisatie, circle chart, afstandsafhankelijk visualisatie, datatransformatie

**ACM Classification Keywords**
H.5.m Information Interfaces and Presentation: Miscellaneous—*Information visualization*

## INTRODUCTIE

Veel visualisaties maken gebruik van geografische data. Een voor de hand liggende manier om deze data te tonen, is op een kaart. Hierdoor krijgt de gebruiker direct een zicht op de geografische spreiding. In alternatieve aanpakken wordt de geografische data ontkoppeld van de geografische ligging. A. Becker, S. Eick en A. Wilks beschrijven hiervoor een matrix display [4]. Een probleem stelt zich echter als geografisch afhankelijke data gebruikt wordt in zulke alternatieve visualisatie. Dit is het gevolg van de zogenaamde ¨Tobler's First Law of Geography´´: Alles is gerelateerd aan alles, maar dichtere dingen zijn meer gerelateerd dan verdere dingen [6]. Dit probleem wordt ook wel spatial-autocorrelation genoemd [3]. In zulke data kan er een vervorming in functie van de afstand aanwezig zijn. Doordat de notie van geografische ligging wordt weggehaald, kunnen er verkeerde interpretaties worden gedaan. Een belangrijke opmerking die hierbij moet worden gemaakt, is dat afhankelijk van de context van de data dit probleem zich al dan niet stelt. Dit wil zeggen dat als de gegevens van de data in verhouding staan met de afstand tot de geografische plaats, de data slecht geïnterpreteerd kan worden. Als daarentegen bijvoorbeeld het aantal griepzieken per gemeente wordt gevisualizeerd, zal er geen probleem zijn, omdat deze niet afhankelijk is van de plaats waar de gemeente zich bevindt (in de veronderstelling dat het niet gaat over de verspreiding ervan en de griephaard). In de volgende sectie wordt het probleem besproken dat deze paper tracht te visualiseren.

### Probleemstelling

De universiteit van Hasselt wil een marketingcampagne uitvoeren. Hiervoor heeft ze gegevens nodig om te kunnen zien welke scholen het goed en minder goed doen. Ze beschikt over het aantal studenten per school en per jaar die komen studeren aan hun universiteit. Deze studenten per school zijn gerelaterd aan de afstand tot Diepenbeek. Scholen verder gelegen sturen standaard minder studenten dan scholen dichterbij gelegen. Indien de data gevisualiseerd zou worden in een visualisatie dat onkoppeld is van een geografische focus, zou dit kunnen resulteren in slechte interpretaties. Hoe kan de data gevisualiseerd worden zodat verkeerde interpretaties worden uitgesloten? In deze paper zullen de scholen voor de gemakkelijkheid per gemeente worden gebundeld. Verder zullen er ook meerdere dimensies van data worden aangeboden. Volgens Tufte [7] is hypervariate data [5] de meest interessante data om te begrijpen .

## GERELATEERD WERK

Het gerelateerd werk is gebaseerd op de klassieke paper van Y.K. Leung en M.D. Apperley [2].

### Focus en context

Vele visualisaties moeten een groot aantal data kunnen tonen. Een veel voorkomend probleem is dat de oppervlakte van een visualisatie te klein is om al deze data te kunnen tonen. Hierdoor zal de data slecht te interpreteren zijn en kunnen er maar moeilijk verbanden worden gezocht. Y.K. Leung en M.D. Apperley [2] maken gebruik van extra informatie afkomstig van de gebruiker of van de data om dit probleem op te lossen. Dikwijls wil de gebruiker zich focussen op een bepaald deelaspect van een visualisatie of kan er gebruik worden gemaakt van de context van de data. Op basis van deze twee aspecten kan er een bepaalde ordening, transformatie, distortie of andere techniek worden gebruikt om de data duidelijker voor te stellen. Deze acties vallen in

Paper voor het vak informatievisualisatie, geschreven door Nick Michiels.

twee categorieën: non-distortie technieken en distortie technieken.

*Non-distortie technieken*

Een non-distortie techniek zal de data niet veranderen, maar wel inwerken op de ordening of zichtbaarheid van de data. Enkele voorbeelden van non-distortie technieken zijn: het weglaten van uitschieters, alfabetische ordening, zoomen, panning (viewpoint veranderen), . . .

*Distortie technieken*

In tegenstelling tot non-distortie technieken, zal een distortie techniek de presentatie of de data wel vervormen. Afhankelijk van de context of de focus kan er een vervorming op de presentatie van de data worden toegepast. Het hoeft niet altijd een vervorming op de presentatie te zijn. Soms kan ook de presentatie behouden blijven, maar wordt er voordat de data wordt uitgetekend eerst nog een vervorming op de data gedaan. Enkele voorbeelden van distortie technieken zijn de bifocal display, perspective walls en fisheye views.
Om deze distortie voor te stellen wordt er typisch gebruik gemaakt van een transformatiefunctie en magnificatiefunctie. Een voorbeeld voor de perspective wall vindt u in Figuur 1.

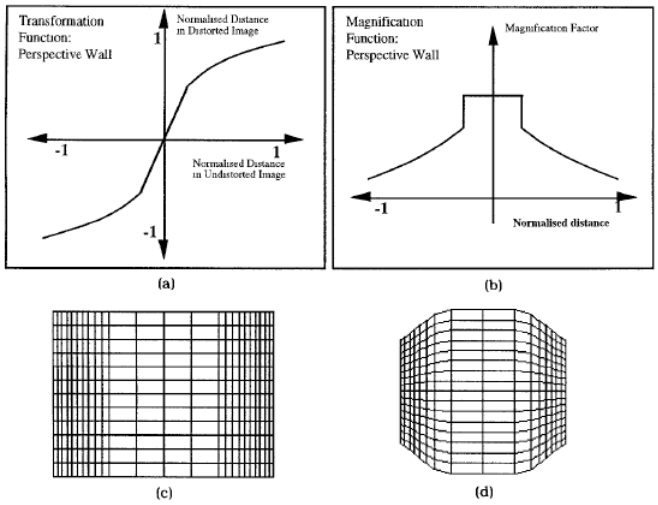


**Figure 1. De perspective wall: (a) transformatie functie; (b) magnification functie; (c) de visualisatie in 2D; (d) de visualisatie in 3D. Figuur van [2].**

## CIRCLE CHART

### Overzicht

Herinner, deze paper probeert het aantal studenten per gemeente afhankelijk van de afstand te visualiseren. Dit gebeurt in de circle chart. We zullen ons niet enkel beperken tot de inwerking van de afstand op de data, maar ook trachten meerdere dimensies van de data toe te voegen (afstand, goedheid, positie, kleur). Het is een visualisatie waar de geografische ligging van de gemeentes wordt weggelaten. Ter verduidelijking zal er een voorbeeld van het resultaat gegeven worden, te zien in Figuur 2.

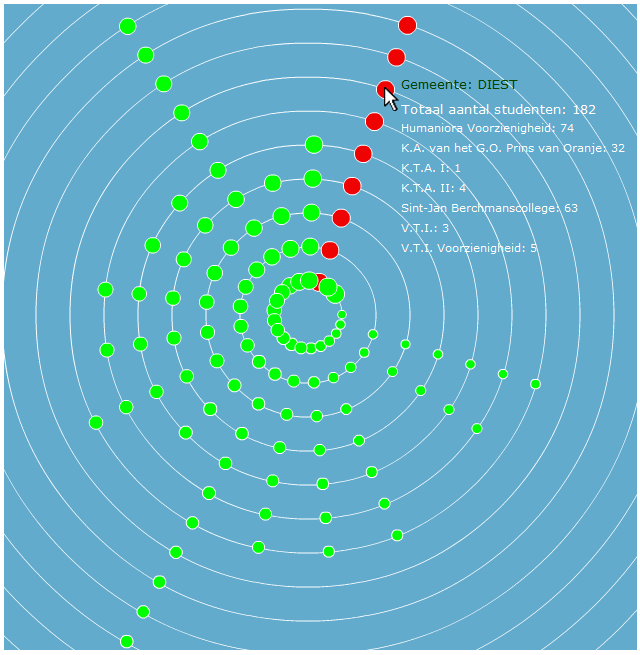


**Figure 2. Voorbeeld van een circle chart.**

Er zijn een aantal vragen die zich stellen bij de aanmaak van deze visualisatie. Wat wordt aangegeven met de kleur en grootte van een bolletje? Welke transformatiefunctie wordt gebruikt om de data in functie van de afstand te vervormen? Welke gemeentes hebben een gelijke hoek en liggen dus op dezelfde lijn? Hoe worden de gemeentes op die lijn geordend? De laatste twee vragen zijn niet noodzakelijk voor de betrouwbaarheid, maar geven wel een extra dimensie aan de data.

*Waarvoor staat de grootte van een bolletje?*

Het aantal studenten voor een gemeente weerspiegelt zich in de grootte van een bolletje. Let op, dit is wel het getransformeerde aantal. De grootte hangt dus rechstreeks af van de keuze van de transformatie functie.

*Welke transformatie functie gebruiken?*

Over het algemeen is het computer intensief om een auto correlation probleem op te lossen [1]. In deze paper wordt er gebruik gemaakt van een eigen alternatieve manier die relatief goede resultaten oplevert. In deze visualisatie gaat er geen distortie gebeuren op de presentatie van de data, maar wel op de data zelf. De probleemstelling levert ons een context van de data. De context is namelijk de volgende: studenten gegroepeerd per gemeente die komen studeren aan de universiteit van Hasselt. Deze context impliceert ook direct een focus van de data: de universiteit van Hasselt is gelegen in Diepenbeek. Centraal in onze visualisatie bevindt zich dus Diepenbeek. Uit de focus en context kan afgeleid worden dat het aantal studenten afhankelijk is van de afstand tot de universiteit. De afstandsfactor kan gebruikt worden in de transformatie functie. In feite kan de huidige data reeds als getransformeerd worden beschouwd. De grootte van de data is omgekeerd gecorreleerd met de afstand er-

van. De functie die het best mapt met deze correlatie is dus in feite de huidige magnificatiefunctie. De functie is van de vorm $y = -x + maxY$ (hierbij is maxY de maximale waarde voor het aantal studenten). Het beste is om de waardes te normaliseren zodat de maximale waarde 1 is. Dit komt dan overeen met geen vervorming voor Diepenbeek en hoe groter de afstand wordt, hoe meer vervorming er op aanwezig is. Deze functie wordt bekomen door de beste lijn te fitten in de gecorreleerde punten. De bedoeling is om deze correlatie weg te werken. Er moet dus een transformatiefunctie worden gekozen om de magnificatiefunctie terug vlak te trekken. Dit is geïllustreerd in Figuur 3. Links ziet u de de magnificatiefunctie van de huidige data. Rechts de magnificatiefunctie die we trachten te bekomen na transformatie. Het is wel belangrijk in te zien dat er nu geen transformatie gebeurt op de presentatie, maar wel op de data zelf. De magnificatiefunctie geeft in dit geval dus de vervorming van de data aan. Omdat het vinden van de beste gefitte lijn ons te ver naar de statistiek leidt, wordt er hier gebruik gemaakt van een eigen methode. De afstand vermenigvuldigen met het aantal studenten geeft een goede benadering:

**Transformatie.** $\forall$ *gemeente G, met afstand afstand* $X_G$ *en aantal studenten* $S_G$*, hebben we de getransformeerde data* $T_G$*:*

$$T_G = \log(X_G * S_G) \tag{1}$$

Er wordt een log aan de transformatie toegevoegd om er voor te zorgen dat grote uitschieters niet té grote bolletjes krijgen in het resultaat.

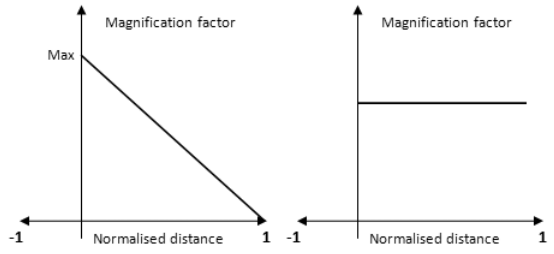


**Figure 3. Links op de afbeelding ziet u de huidige magnification functie van de data. Rechts de magnification functie hoe ze er uit zou moeten zien na transformatie.**

*Welke gemeentes liggen onder een gelijke hoek?*
De keuze om de gemeentes onder te verdelen in hoeken is niet noodzakelijk. Men zou deze ook gewoon op 1 lijn kunnen plaatsen van klein naar groot. Toch kan het voordelen hebben om ze te ordenen per hoek. Zo is er een grotere groepering van bijeenhorende data. Er zijn veel mogelijkheden om een hoek toe te kennen aan een gemeente. Ook dit is weer sterk afhankelijk van de context. In ons resultaat is er voor gekozen om gemeentes die het min of meer even goed doen op dezelfde lijn te leggen. Min of meer even goed uit zich in bollen met dezelfde grootte. Het gebruik van deze ordening geeft de gebruiker direct alle gemeentes terug die het even goed doen. Andere ordeningensmethodes om de gemeentes te groeperen zijn o.a. per afstands interval, per geografishe hoek, per provincie, ... Het groeperen per afstandsinterval geeft per lijn een overzicht van de goede en slechte gemeentes voor dezelfde afstand.

*Hoe de gemeentes op een lijn ordenen?*
Ook deze ordening is enkel nodig om meer dimensie te geven aan de visualisatie. Men kan er voor kiezen om de afstanden van een gemeente te weerspiegelen aan de afstand tot het middelpunt. Dit geeft in één oogopslag de gemeentes die het slecht doen en dichtbij liggen. Het zijn ook net de gemeentes die dichterbij liggen die belangrijker zijn voor de universiteit. Toch geeft dit niet altijd de gewenste resultaten. We komen hier in een volgende sectie op terug. Een tweede manier is i.p.v. de geografische afstand tot een gemeente, een vaste afstand tussen de gemeentes te nemen. Bijkomend hierbij worden de gemeentes van dichtbij naar ver geordend. Op die manier is er nog wel een notie van de echte afstand, maar wordt de structuur van de visualisatie niet verstoord.

*Kleurkeuze van de gemeente*
De kleurkeuze van de gemeente kan nóg een extra dimensie geven aan de data. In onze implementatie is deze niet erg uitgebuit. Hij wordt enkel gebruikt voor te hoveren. Standaard worden alle gemeentes op één lijn ingekleurd als de muis over één van die gemeentes komt. Ook hier bestaan weer andere variaties waarvan er één nog in de resultaten aan bod komt.

### Variaties

Voor de vier keuzes die in de vorige sectie werden aangehaald, kunnen er nog heel wat variaties worden gevonden. Deze sectie gaat er een aantal aanhalen en bespreekt zijn voor- en nadelen. Sommige conclusies baseren zich op de gebruikerstest die in de discussie nog aan bod komt.

Er is reeds gezegd dat de geografische afstand spiegelen aan de afstand tot het middelpunt niet zo'n goede keuze is. Uit een gebruikerstest blijkt dat dit niet altijd een overzichtelijk resultaat geeft. Het probleem is dat ten eerste gemeentes kunnen overlappen, zodat sommige gemeentes niet meer zichtbaar zijn. Ten tweede zijn er bepaalde aspecten die voor chaos kunnen zorgen. Er kunnen een aantal uitschieters zijn die ver liggen waardoor er moet worden uitgezoomd om alles op het beeld te kunnen krijgen. Het uitzoomen zorgt er voor dat in het centrum veel gemeentes overlappen. Ook de gemeentes die op 1 lijn liggen zijn niet meer zo duidelijk, omdat er nu veel spatie tussen is. De combinatie van dit alles zorgt ervoor dat de algemene structuur en duidelijkheid van de visualisatie verloren gaat.

Een tweede variatie is dat de algemene grootte van de gemeentes gemapped worden van 1 tot 10. Dit vermijdt grote uitschieters en kan dus overweg met alle soorten data. Het voordeel impliceert ook direct het nadeel. De grote uitschieters worden gemapped op tien, maar dit wil ook zeggen dat alle andere kleinere gaan gedownsized worden waardoor ze allemaal klein worden en min of meer dezelfde visuele grootte hebben. Het doel is juist om de variaties in grootte goed te kunnen opmerken.

De twee keuzes van hoekindeling die deze paper geeft, zijn wel sterk aan elkaar gewaagd. De keuze tussen de twee is sterk afhankelijk van de context van de vraag die men wenst op te lossen. Zo kan een hoekindeling volgens gemeentes van dezelfde grootte goed gebruikt worden om in te schatten welke het minder goed doen en welke het ongeveer even goed doen. Als men eerder wil weten welke gemeentes binnen eenzelfde afstand het slecht doen is de hoekindeling volgens afstand weer beter. Zo zijn alle gemeentes binnen een bepaalde afstand gebundeld. De eerste kan ook in de tweede worden verwerkt door kleuren te gebruiken voor de gemeentes. De implementatie kan bijvoorbeeld alle gemeentes inkleuren die het slechter of even goed doen dan de geselecteerde. Een voorbeeld van deze implementatie vindt u in Figuur 4.

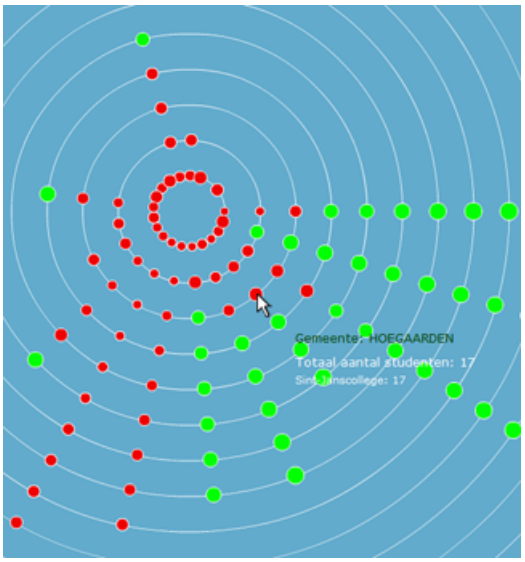


**Figure 4. Voorbeeld van een circle chart waar gemeentes rood worden ingekleurd indien ze relatief slechter zijn dan de geselecteerde gemeente.**

## DISCUSSIE/CONSLUSIE

Deze visualisatie is onderdeel geweest van een user test. Hierbij waren er twee categorieën van gebruikersgroepen: vier personen van de universiteit Hasselt die de input data gebruiken, kortom de kenners, en vier personen die de data nog nooit hebben gezien. Uit de groep van kenners kan er vooral worden afgeleid dat zo'n transformatie noodzakelijk is om de data correct te kunnen interpreteren. Ze zijn dan ook zeer positief over de uitkomst die uit de visualisatie komt. Een groter probleem is het begrijpen van de visualisatie zelf. Het is moeilijk om de visualisatie door te hebben zonder er eerst een opleiding over gekregen te hebben. Omdat het zo moeilijk begrijpbaar is, is het geen goede visualisatie om op grote schaal, zoals het internet, te gebruiken. Bij de groep van leken was de reactie ongeveer hetzelfde. Men moet al wat achtergrond hebben om de circle chart te kunnen interpreteren en daarenboven vergt het ook veel tijd om te begrijpen hoe de visualisatie gestructureerd is. Toch zijn ze positief als ze het éénmaal doorhebben. Het gebruik van extra labels en extra interactief opvraagbare informatie zal volgens hen de bruikbaarheid ook verhogen. Enkele gebruikers vonden het (veelvuldig) flikkeren van groene naar rode bolletjes storend voor de ogen. Kort gezegd is het dus een zeer nuttige visualisatie, maar kunnen er nog heel wat stappen worden ondernomen om de duidelijkheid en begrijpbaarheid te verhogen.

## REFERENCES

1. L. Anselin. Local indicators of spatial association-lisa. *Geographical Analysis*, 27:93–115, 1995.

2. Y. K. Leung and M. D. Apperley. A review and taxonomy of distortion-oriented presentation techniques. *ACM Trans. Comput.-Hum. Interact.*, 1(2):126–160, 1994.

3. H. J. Miller and J. Han. Geographic data mining and knowledge discovery. pages 139–149, 2001. Chapter 6, ISBN: 0415233690.

4. A. R. W. Richard A. Becker, Stephen G. Eick. Visualizing network data. *IEEE Transactions on Visualizationand Computer Graphics*, 1(1), 1995.

5. R. Spence. Information visualization. pages 33–51, 2006. Chapter 3, ISBN: 0132065509.

6. W. Tobler. A computer movie simulating urban growth in the detroit region. *Economic Geography*, 2(46):34–40, 1970.

7. E. Tufte. *The Visual Display of Quantitative Information.* Graphics Press, 1983.