Data Processing

Julia Jansen

Shannon Bakker

Tom Schoufour

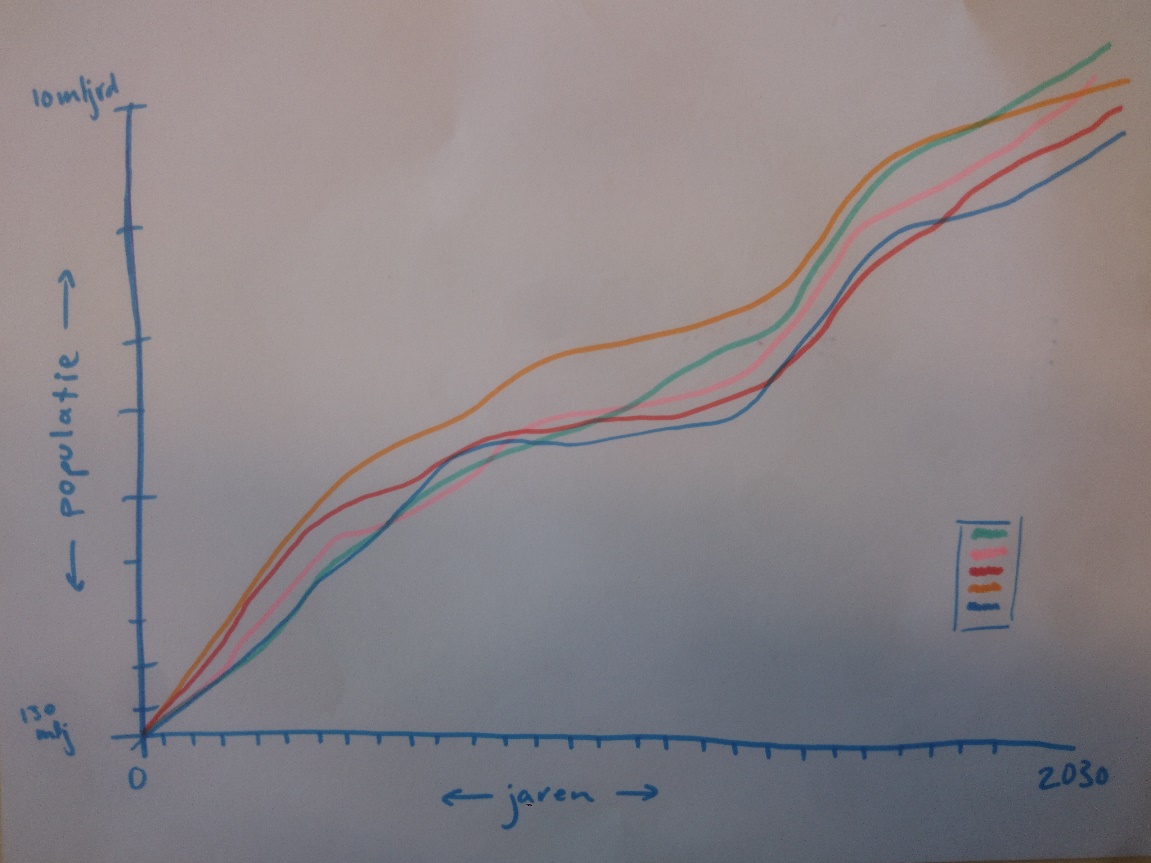
Thom Mekelenkamp

# Analyse

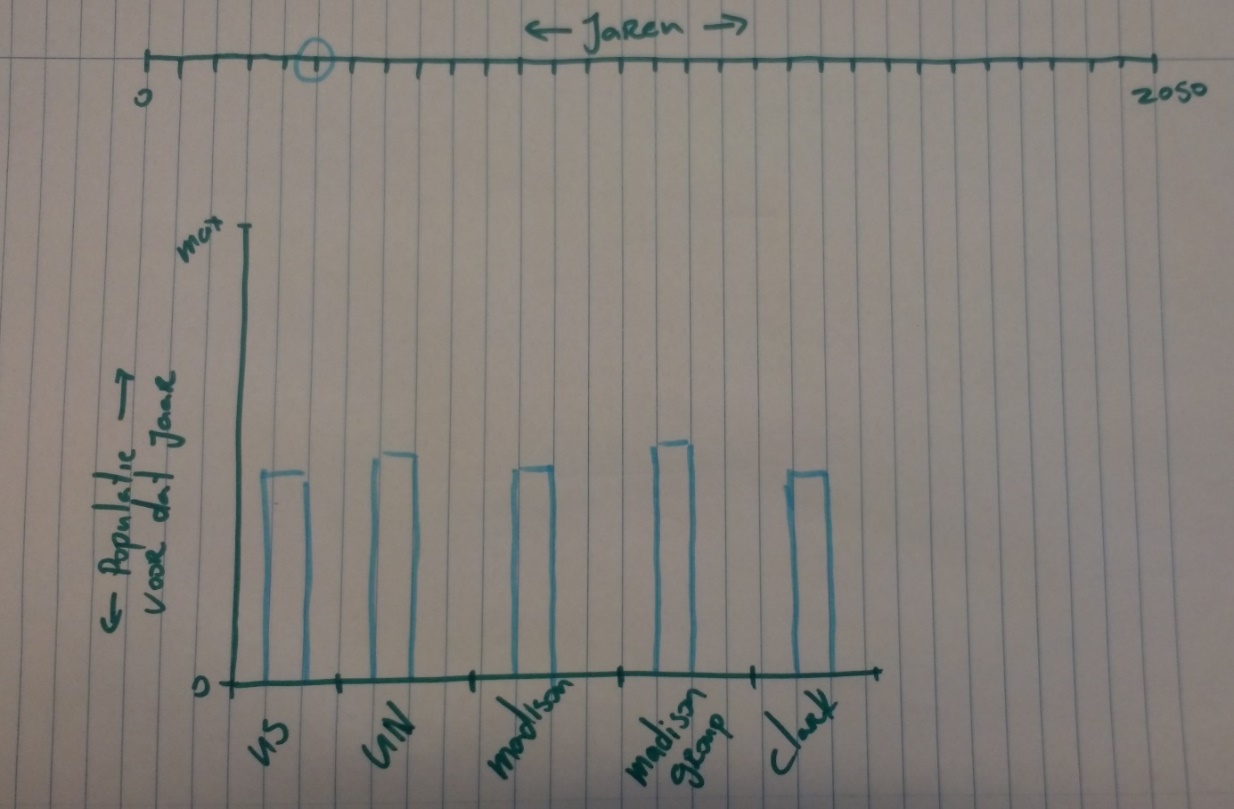
* Trends in de data
  + In de bijna gehele tabel stijgt het aantal mensen.
  + Tm 1500 stijgt de populatie gematigd, na 1500 neemt de groei per jaar toe.
  + Een aantal keer voor 1950 neemt de bevolking af
  + Vanaf 1900 begint de bevolking explosief te groeien
  + Vanaf 1950 groeit de bevolking met ongeveer 100 miljoen per jaar.
  + Vanaf 8 miljard mensen lijkt de groei te stabiliseren.
* Verschillen in de data
  + Voor 1850 zijn er grote verschillen tussen de verschillende partijen.
  + Sommige partijen geven slecht eens in de x jaar een schatting, waar sommige partijen ieder jaar een schatting uitbrengen.
  + US, UN en Madison lopen vrijwel gelijk op vanaf 1950. En Madison group ligt van 1950 tot 2003 onder de schattingen van de rest en daarna ver boven..
* Verschillen in schattingen (absolut en relatief)
  + Verschillen tussen de schattingen worden groter, maar die zijn relatief tot de populatie niet zo groot. Naarmate de populatei groeit woden de relatieve verschillen kleiner.
  + Absolut:
    - Het verschil in de schatting in 2010 het tussen Madison en de USCB groot met een verschil van ongeveer 900 miljoen.
  + Relatief:
    - Het verschil tussen HYDE en McEvedy & Jones voor -4000, daar is het verschill relatief het grootst met 400%
* Data representatie
  + Ja, het kan worden weergeven door de schaal van de as aan de totale populatie aan te passen. Als je dit doet wordt het relatieve verschil in de schattingen waarheidsgetrouw weergegeven. Het nadeel van deze vorm is wel dat dermate lastig te lezen wordt dat het niet aan te raden is om deze vorm te gebruiken.
* Lineaire interpolatie
  + Alleen als de data op de zelfde jaren is gebaseerd en de intervallen niet te groot zijn kan je lineaire interpolatie gebruiken. Als dit niet het geval is, is er sprake van een soort schijnzekerheid. Het zou kunnen voor Madison Group, Madison en de UN vanaf 1950 tot en met 2009.

# Grafieken

1. In deze grafiek is de y-as is een geschaalde as op basis van de populatiegrootte. Op de x-as staan de jaartallen van 0 t/m 2050. We hebben voor een geschaalde as gekozen, omdat je op deze manier de verschillen tussen de schattingen eerlijker kan zien. Als je dit niet doet dan lijken de verschillen in het begin heel klein, terwijl ze relatief heel groot zijn. Iedere lijn in de grafiek geeft de data van één instituut weer.



1. De tweede grafiek is een staafdiagram. Op de tijdlijn bovenaan de grafiek kan je kiezen van welk jaar je de gegevens wilt zien. Wanneer er op een jaar geklikt wordt krijg je alle data van dat jaar met op de x-as de verschillende instituten en op de y-as een schaal passend voor dat jaar. ‘Passend’ betekent dat de hoogste waarde van de schaal gelijk is aan de hoogste schatting van de populatie voor dat jaar. Op deze manieren zijn de relatieve verschillen goed zichtbaar.

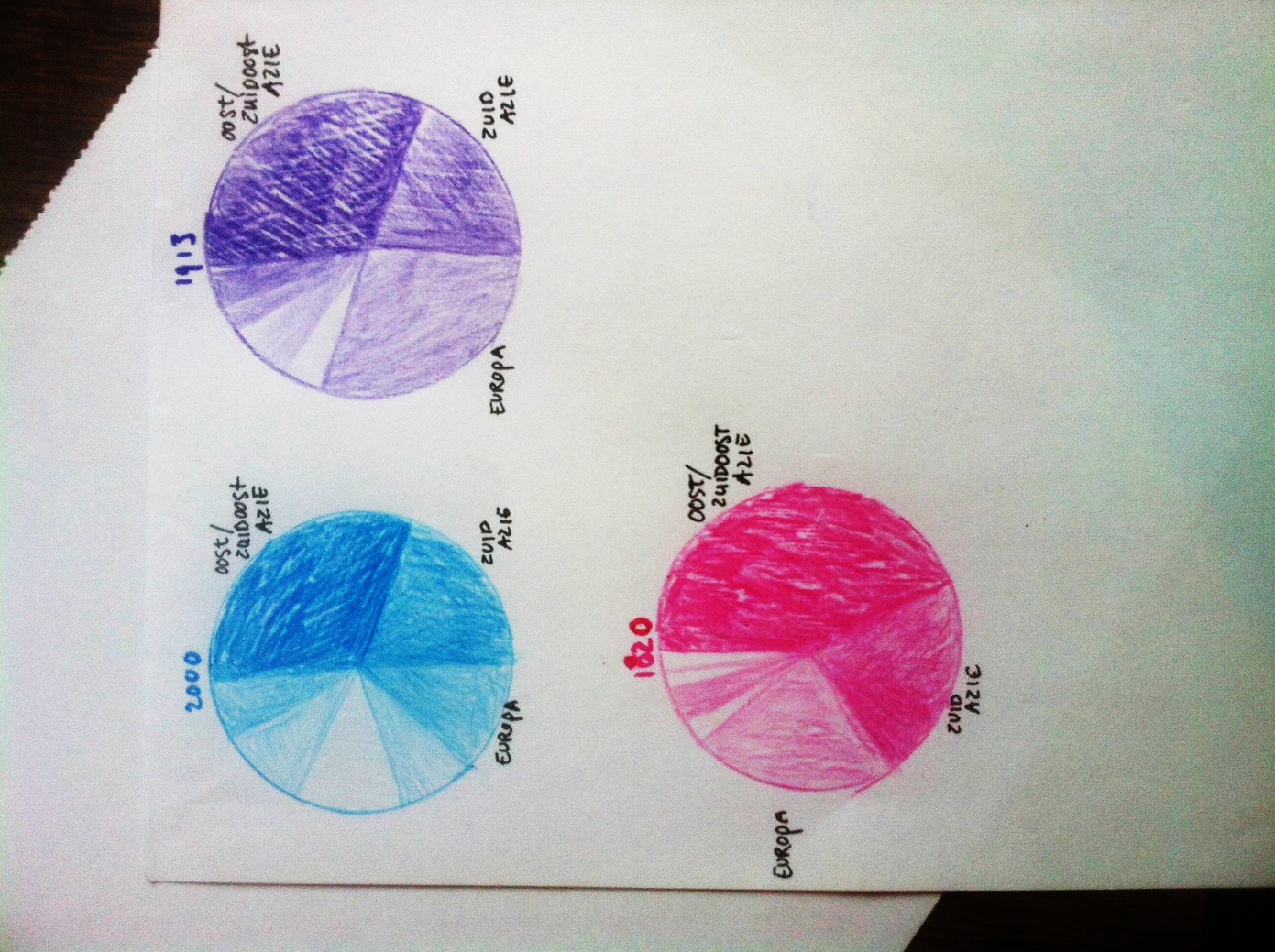


3. De derde grafiek heeft een lineair geschaalde y- as, met daarop de populatiegrootte en een x-as waarop alle jaren staan. Het voordeel van een lineair geschaalde y-as is dat de groei van de bevolking duidelijk is weer te geven. Daarnaast is de y-as makkelijk af te lezen. De data wordt in deze grafiek weergeven met stippen en ieder onderzoek heeft een eigen kleur. We hebben gekozen voor stippen omdat je zo duidelijk kan weergeven hoeveel datapunten ieder onderzoek heeft.

1

4. Grafiek vier laat met behulp van bollen de populatiegrootte zien. Wanneer men op een bol klikt ziet men de populatieverdeling over de verschillende continenten. Deze grafiek is visueel aantrekkelijker dan een lijn- of staafdiagraam. Je ziet direct dat de populatie groter is geworden, omdat de bol ook groter is geworden.





# Discussie

We bespreken eerst waarom we sommige grafieken niet gaan gebruiken:

Ten eerste gebruiken we grafiek 1 niet. Het belangrijk nadeel van de eerste grafiek is dat je de groei niet goed kan zien, dit komt door de geschaalde as. Daarnaast trek je een lijn tussen verschillende datapunten terwijl het aantal datapunten per onderzoek verschilt.

De derde grafiek gebruiken we ook niet. Dit komt doordat de stippen erg moeilijk leesbaar zijn. De oorzaak hiervoor is dat er erg veel datapunten zijn. Daarnaast zijn door de lineair geschaalde y-as de relatieve verschillen tussen de schattingen niet goed te zien aan het begin van de grafiek.

De vierde grafiek gebruiken we niet, omdat we hier niet goed de verschillen tussen de partijen in kunnen laten zien.

We kiezen ervoor om de tweede grafiek wel te gebruiken. Het voordeel van deze grafiek is dat de verschillende voorspellingen van de partijen duidelijk zichtbaar zijn. Door de schaal steeds aan te passen is zowel het relatieve als het absolute verschil tussen de partijen duidelijk zichtbaar. Het nadeel van deze grafiek is wel dat het bijna niet te zien is hoe de populatie zich over meerdere jaren ontwikkeld. Om dit op te lossen kiezen we ervoor om de lezer eerst een lijngrafiek te laten zien. Op deze lijngrafiek staat er in het zwart een trendlijn voor de bevolkingsgroei en in het lichtgrijs staan de stippellijnen van de individuele onderzoeken weergegeven. Wanneer een lezer met zijn muis over één van deze stippellijnen gaat wordt de stippellijn uitgelicht en wordt weergegeven wie de bron is voor de schattingen in deze stippellijn.

Wanneer de lezer met zijn muis over de trendlijn gaat ziet men de staafgrafiek voor dat jaar met alle verschillende onderzoeken. Deze staafdiagram ziet eruit als de tweede grafiek, alleen dan zonder lijn om een jaartal te kiezen.

