Ontwikkeling van een datasysteem voor toerisme

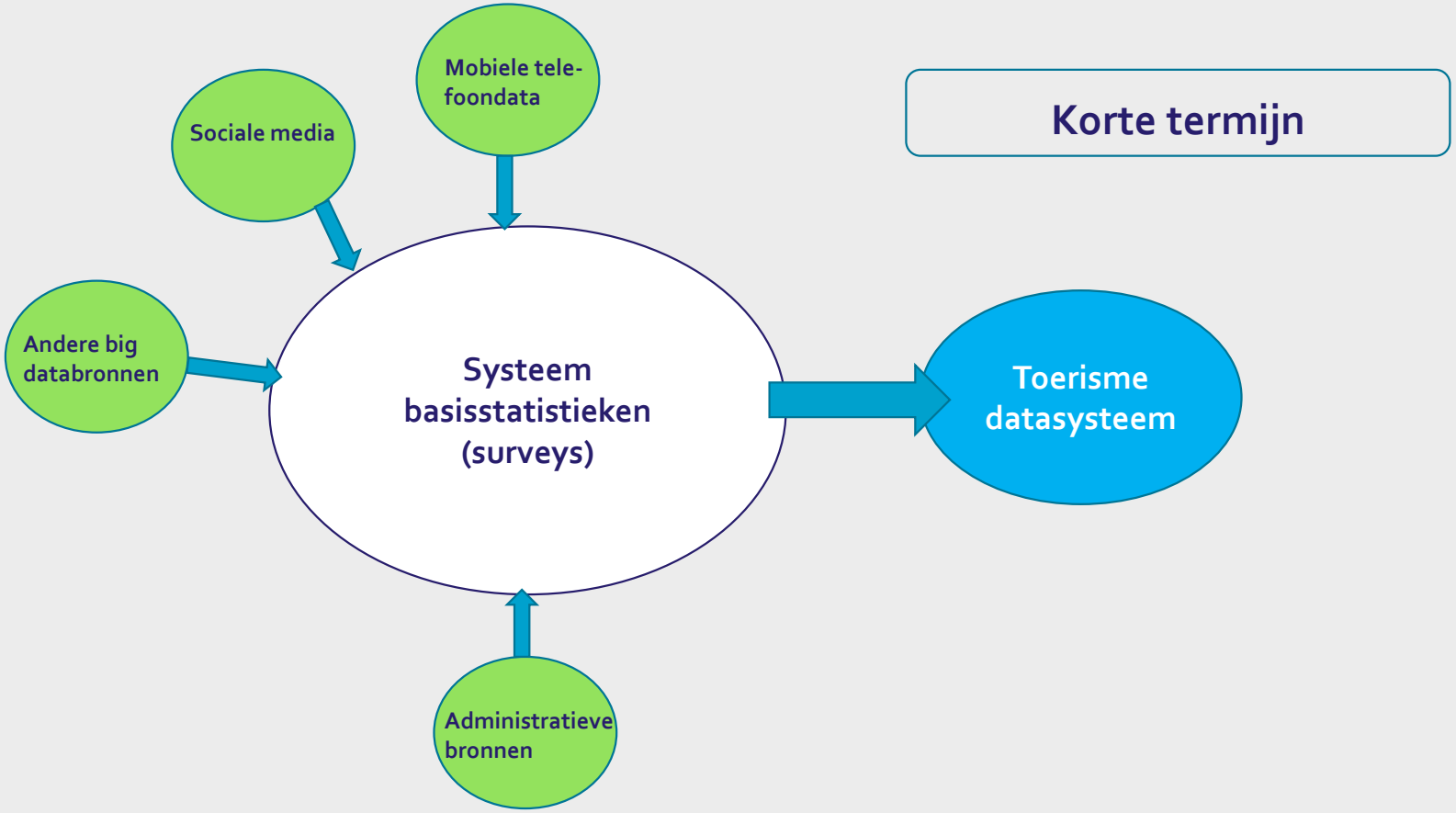
Nico Heerschap, november 2017, CBS

Centraal Bureau voor de Statistiek

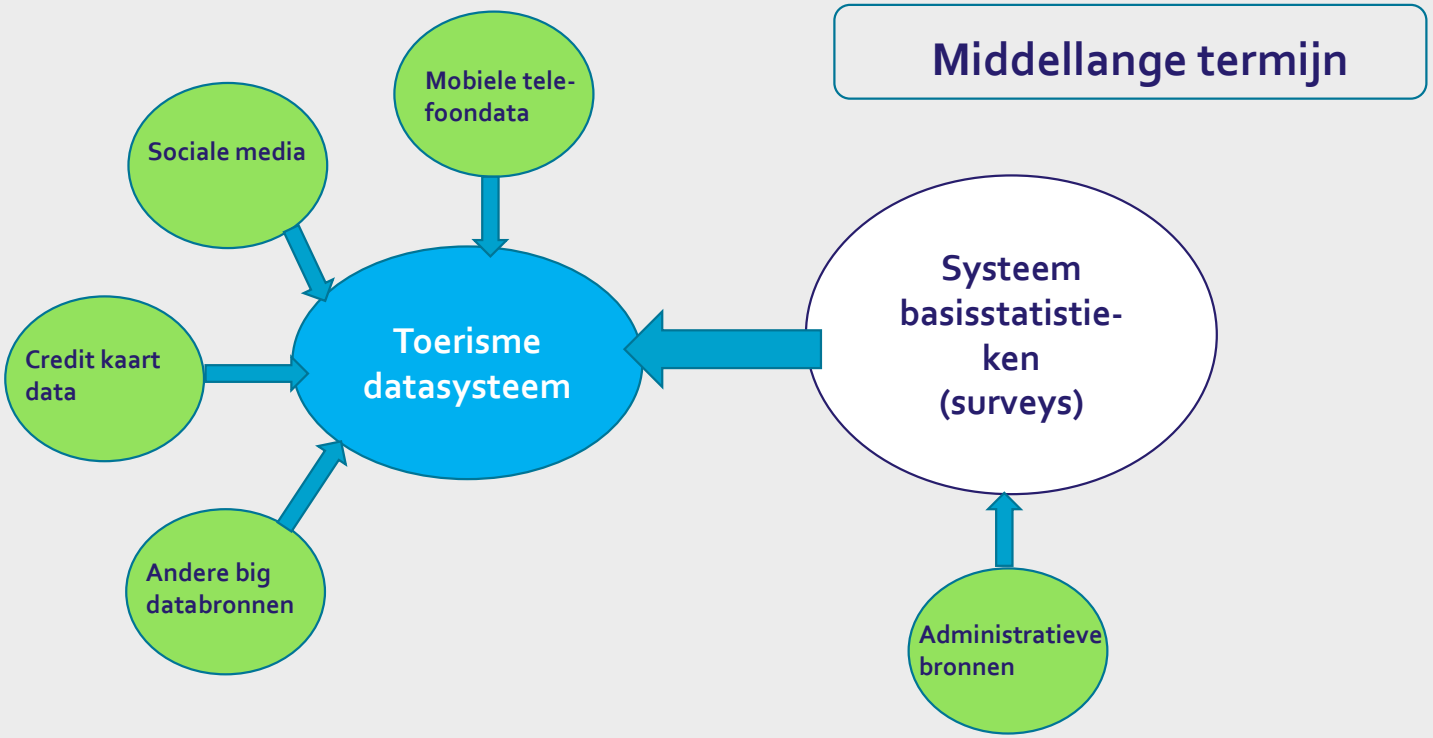
Airbnb

Bron: Dwarshuis

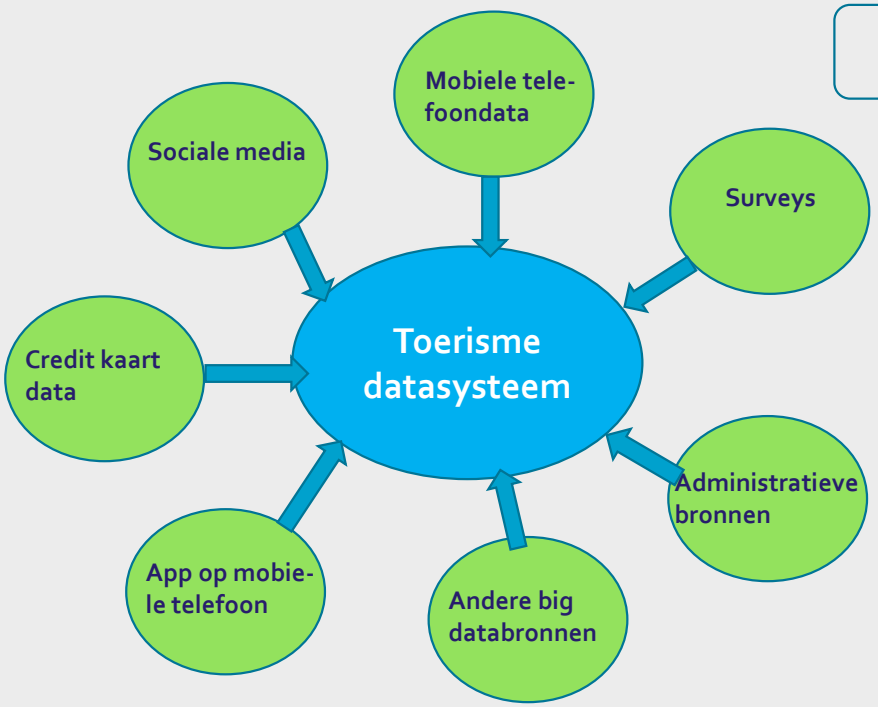
Ontwikkeling datasysteem toerisme



Ontwikkeling datasysteem toerisme



Ontwikkeling datasysteem toerisme



Lange termijn

Combinatie van bronnen
Populatieframes

Echter:
Welk systeem?
Eén systeem?
Wie beheert?
Eéncijfer gedachte?

Mogelijke nieuwe data bronnen

Bron: o.a . Eurostat

Voordelen nieuwe data bronnen, maar…

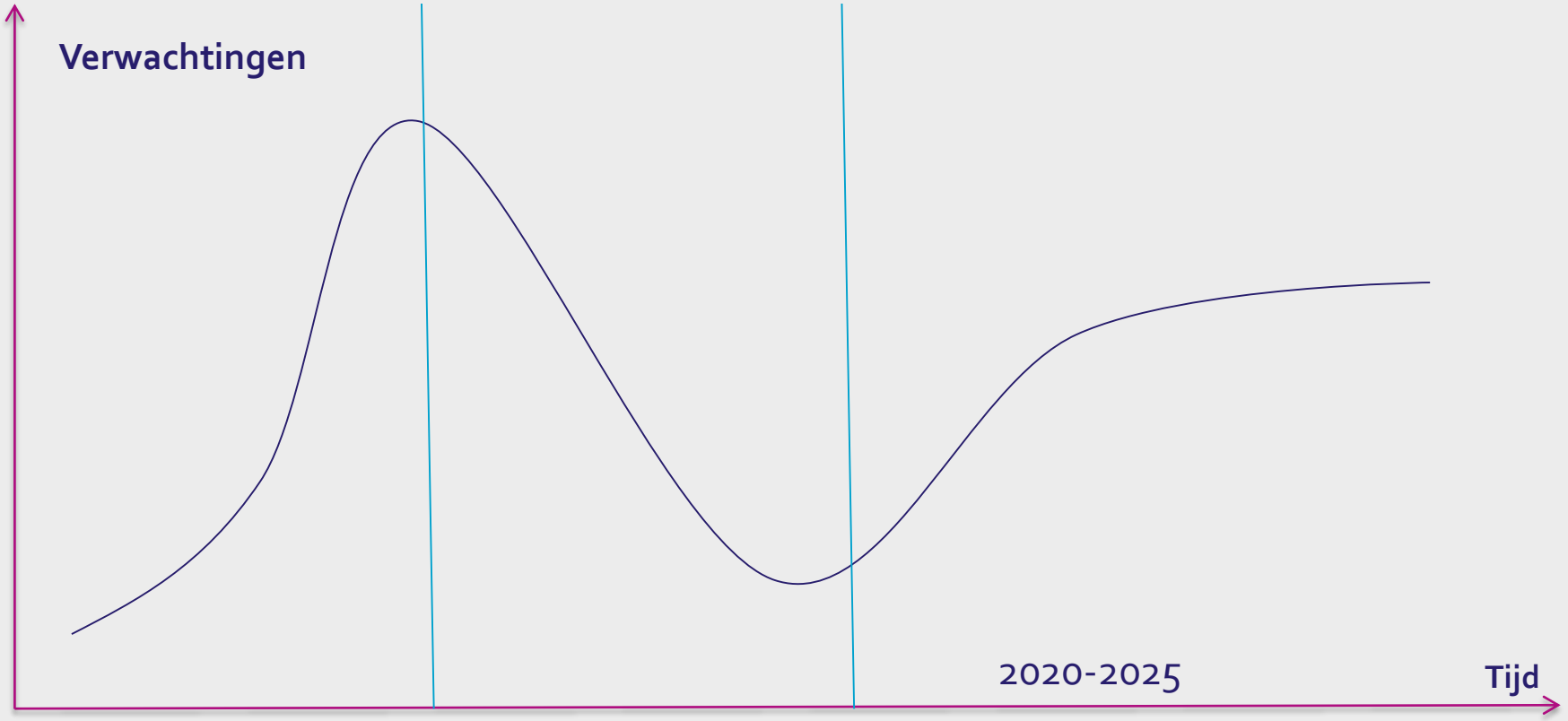
Voordelen, o.a.:

- Meer detail naar ruimte en tijd (vaak als aanvulling op bestaande statistiek)
- Snellere statistiek (voor beleid real-time?) en voorspellingen (predictive analysis)
- Nieuwe statistiek (bijvoorbeeld stromen van toeristen, gevoelens)
- (Mogelijk) efficiënter en dus goedkoper?

Maar…..

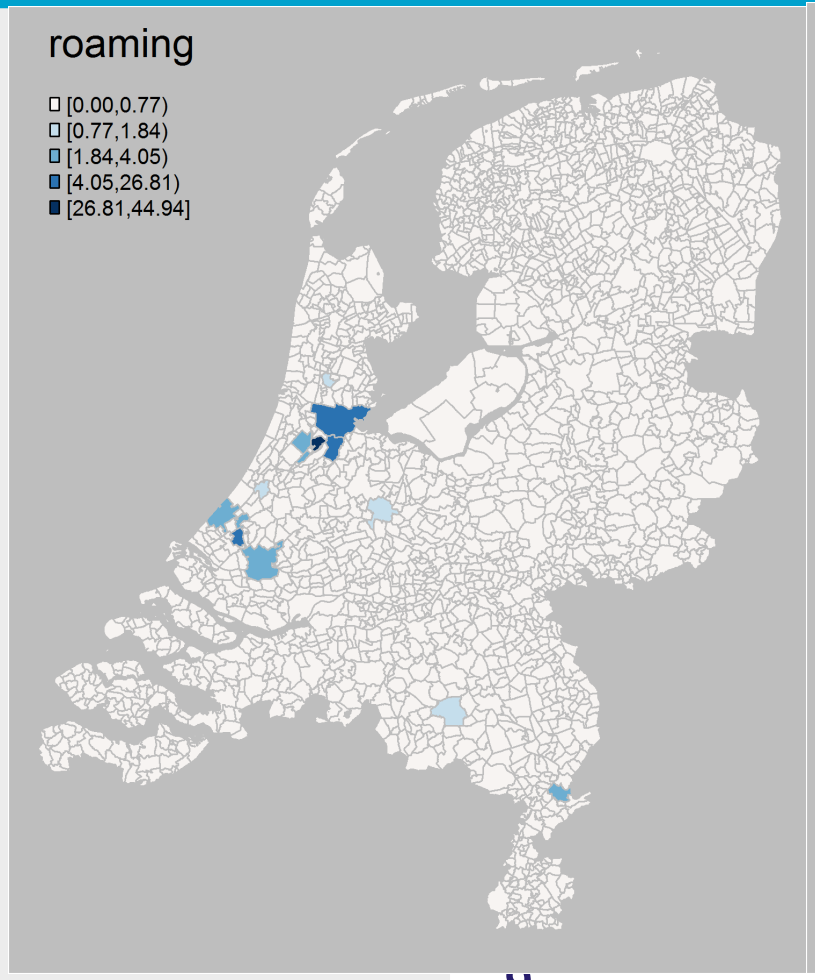
- Nog weinig daadwerkelijke structurele productie van statistieken: experimentele fase.
- Toegang tot data (mobiel telefoon, creditkaartdata). In de CBS-wet opnemen.
- Juridische problemen: privacy (vertrouwen), maar wetten (databankwet; iedereen scrapt bijvoorbeeld Airbnb, Booking e.d.)
- Representativiteit van de data. Hieraan wordt veel gewerkt.
- Aansluiting bij definities en concepten (werken met proxy's)
- Samenwerking (nu vaak concurrentie; verdienmodel)
- Big data overmoed (zijn de cijfers en conclusies wel juist?)

Gartner's hype curve

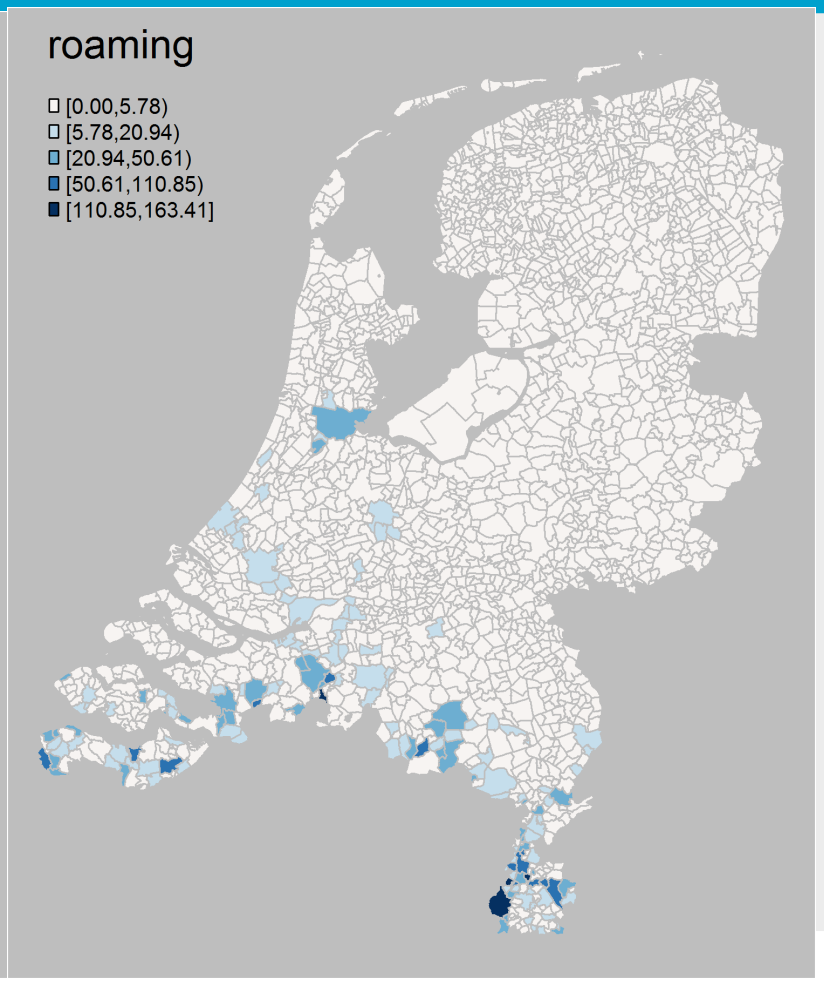


1. Mobiele telefonie data

Aziatische toeristen

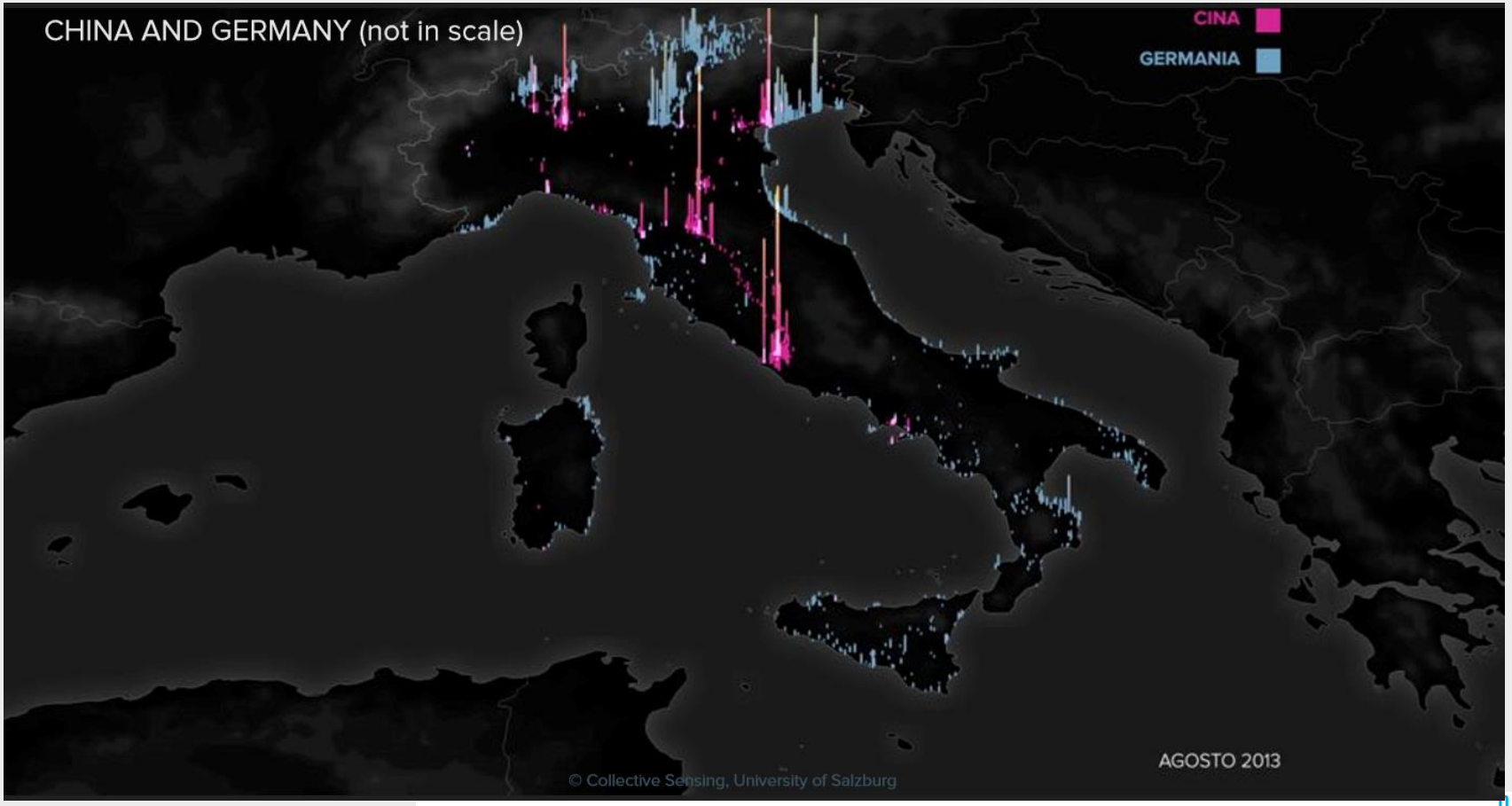


Belgische toeristen

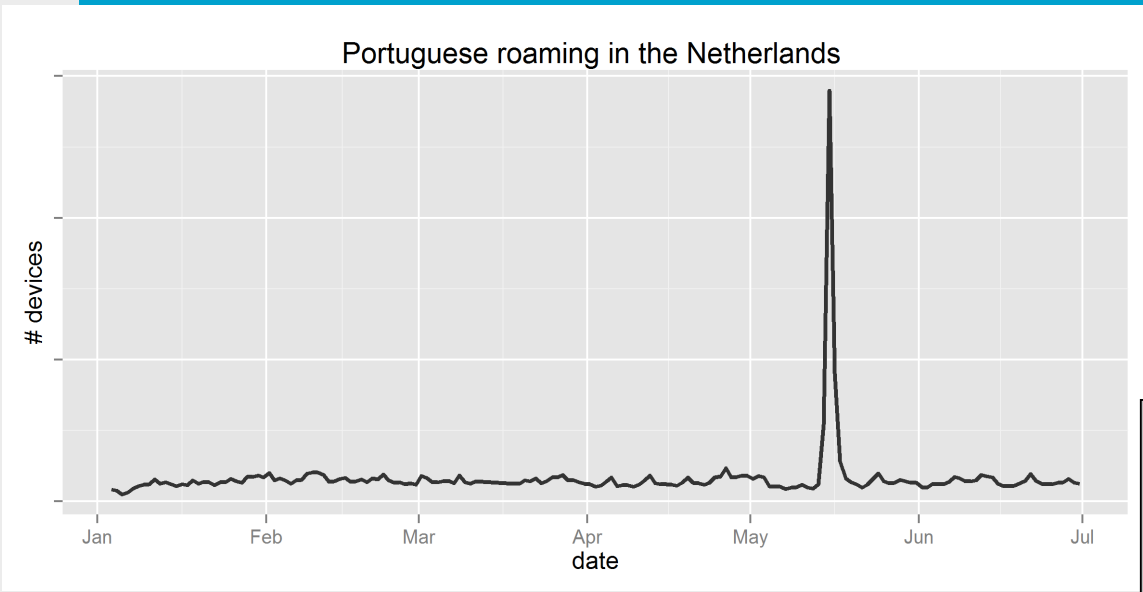


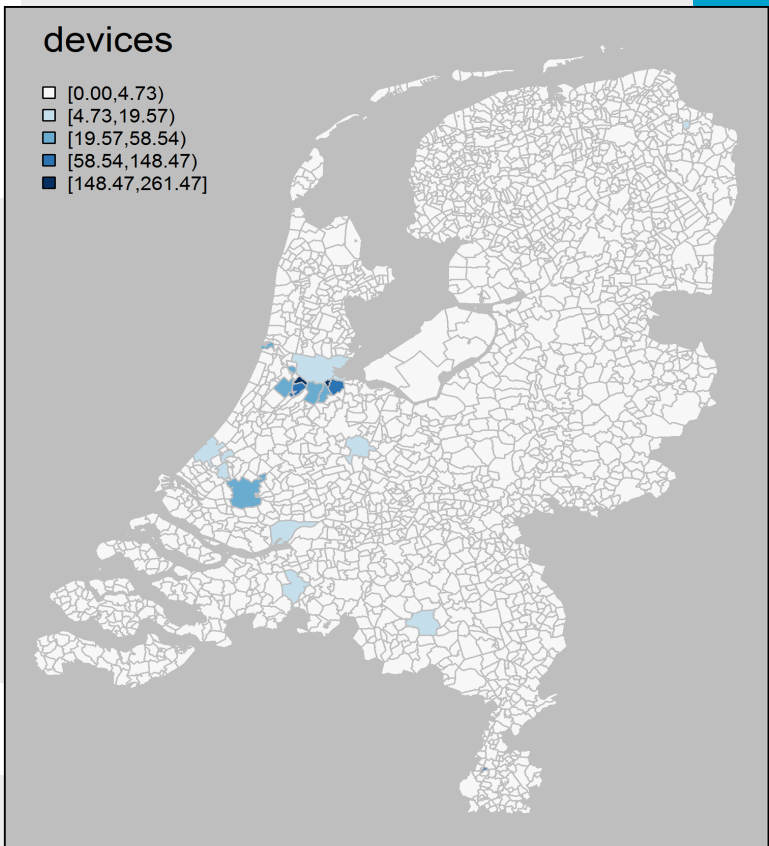
Bron: Vodafone/Mezuro, samengesteld CBS

Source: Vodafone/Mezuro, compiled by SN


CHINA AND GERMANY (not in scale)
CINA
GERMANIA
AGOSTO 2013
© Collective Sensing, University of Salzburg

Portugese toeristen

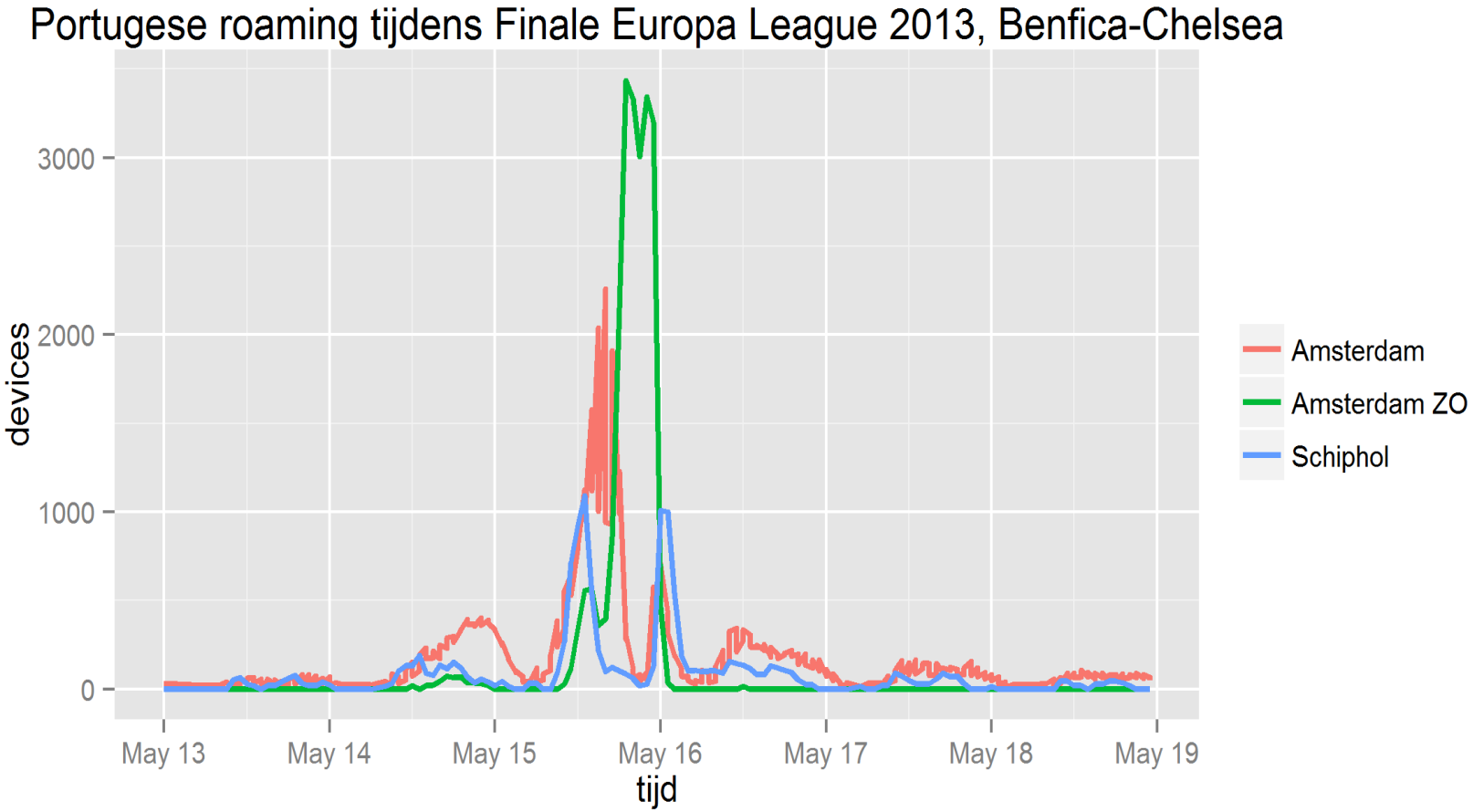




Portugese roaming data tijdens 2013 UEFA Cup
Benfica (Portugal) - Chelsea (England)

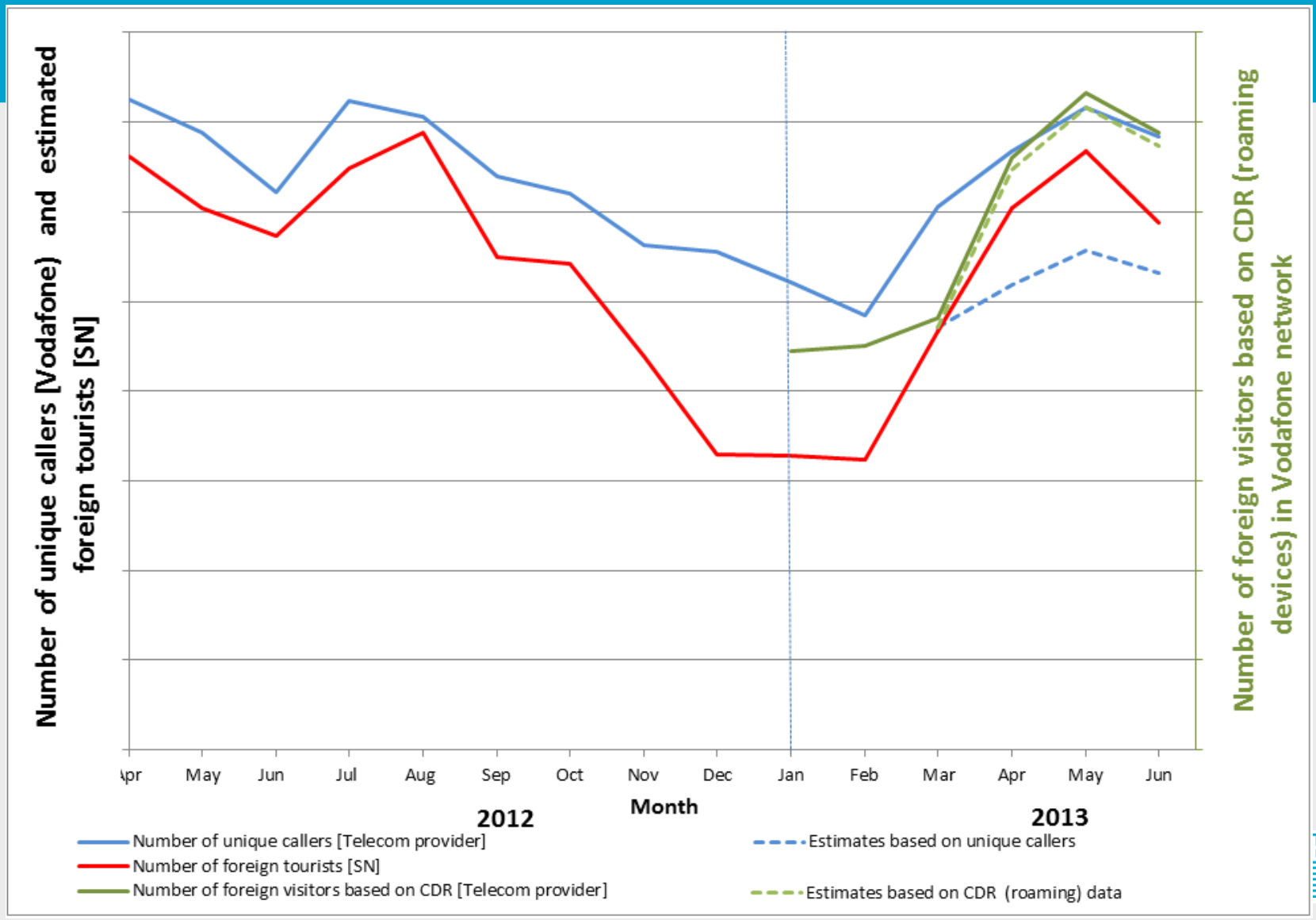
Bron: Vodafone/Mezuro, samengesteld CBS

Portugese toeristen



Bron: Vodafone/Mezuro, samengesteld CBS

Vergelijking met Logiesaccommodaties



Source: Vodafone/Mezuro/SN

Stand van zaken (bij CBS)

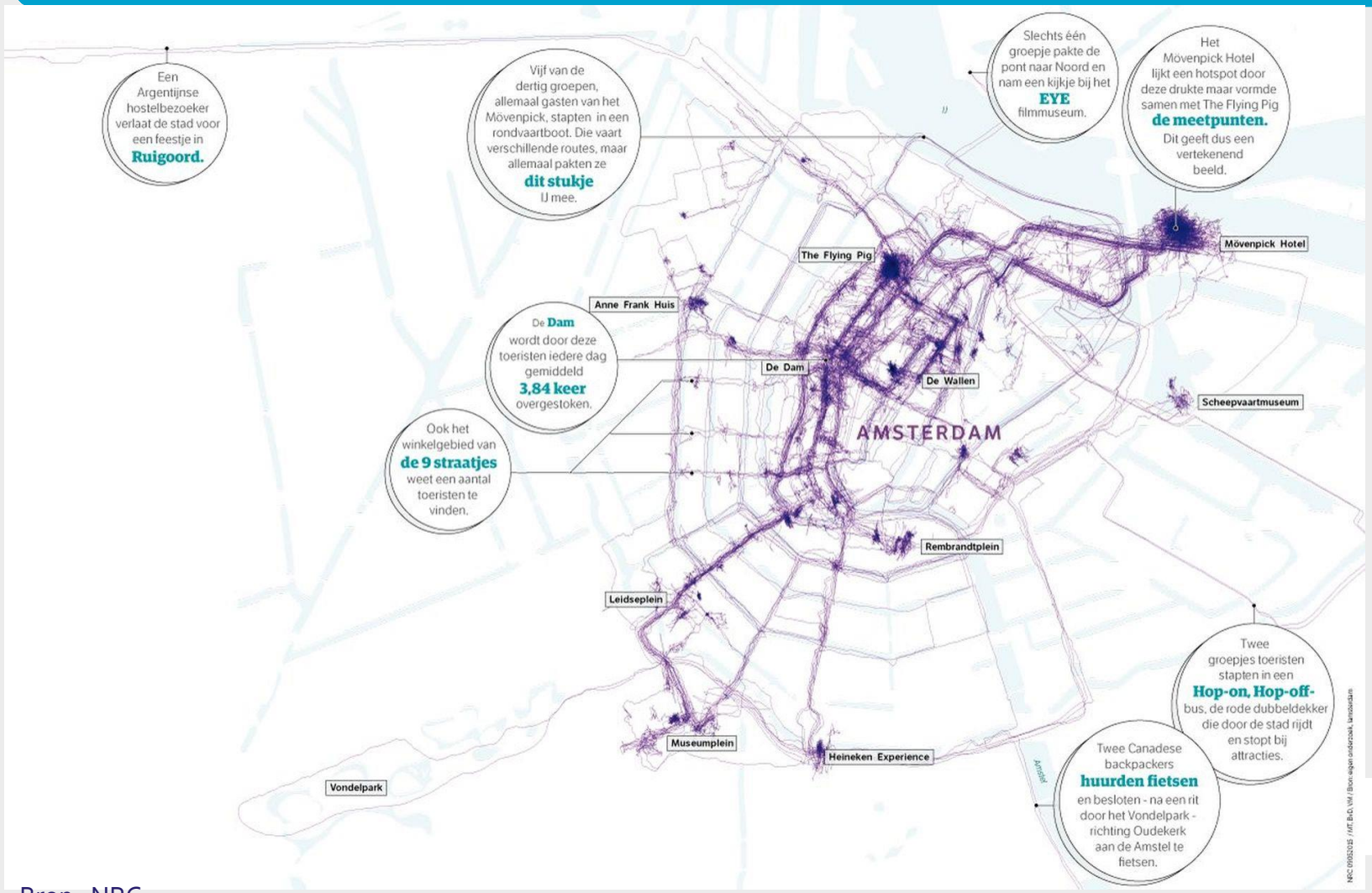
Onderzoek op basis van Vodafone-data (2012-2016)

- Zinvolle resultaten: vooral trends en meer detail in tijd en ruimte (bestaande statistiek).
- Zowel inkomend als uitgaand toerisme, minder als het gaat om binnenlands toerisme (usual environment)

Ontwikkelingen:

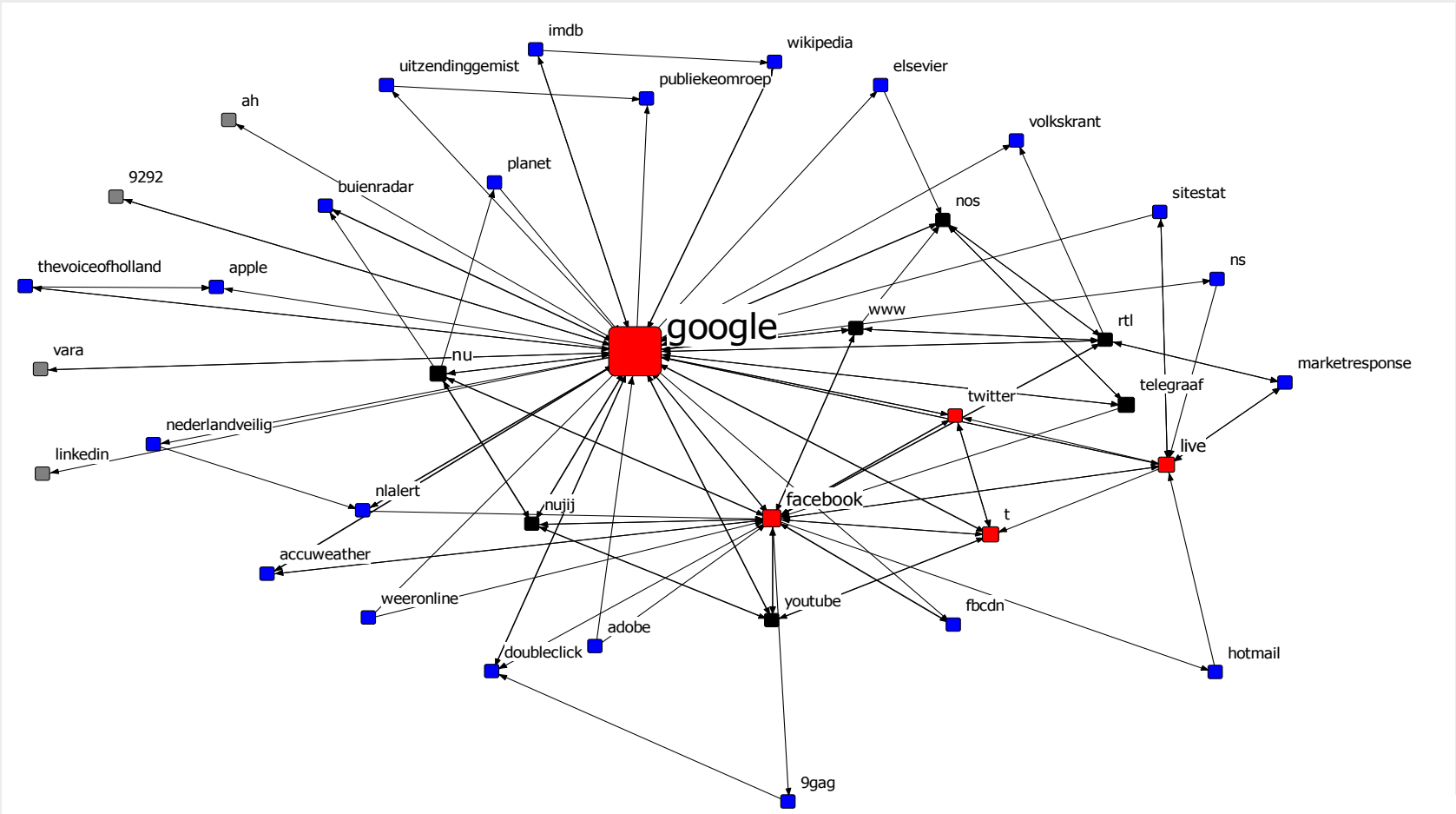
- Er wordt nu gewerkt met T-Mobile. Gesprekken lopen met KPN en Vodafone
- Datatoegang opnemen in CBS-wet (vergelijk Finland; ook Eurostat)
- Niet alleen CDR's, maar ook contactevents
- Representativiteit van de data. Hieraan wordt veel gewerkt.
- Eventueel ook administratieve data van provider (achtergrondkenmerken)
- Eerder trends dan volumes/aantallen

2. Apps op mobiele telefoon /GPS-tracker/pushberichten



Bron: NRC

Surfgedrag → stromen toerisme

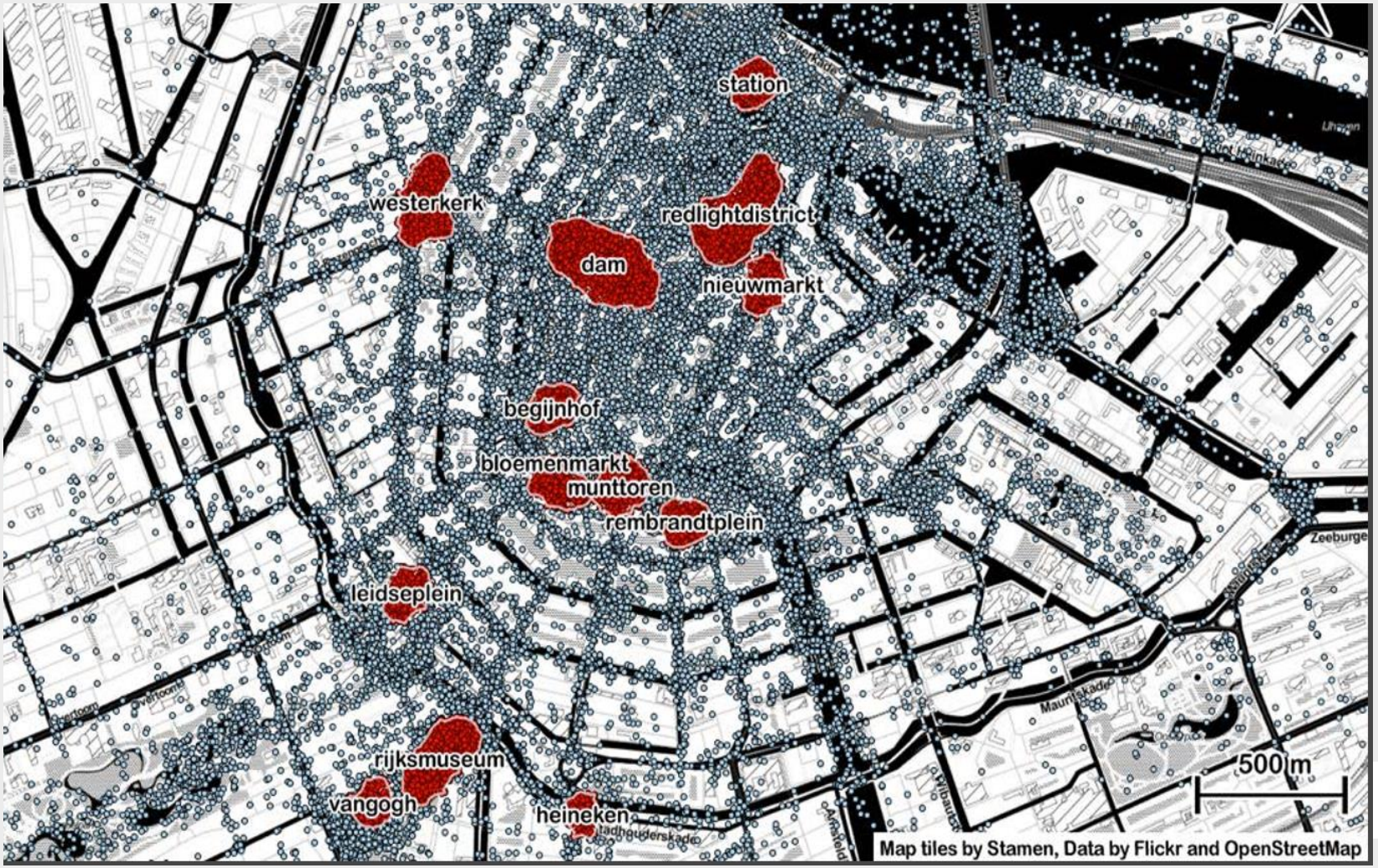


Bron: CBS / TU-Delft

Stand van zaken (bij CBS)

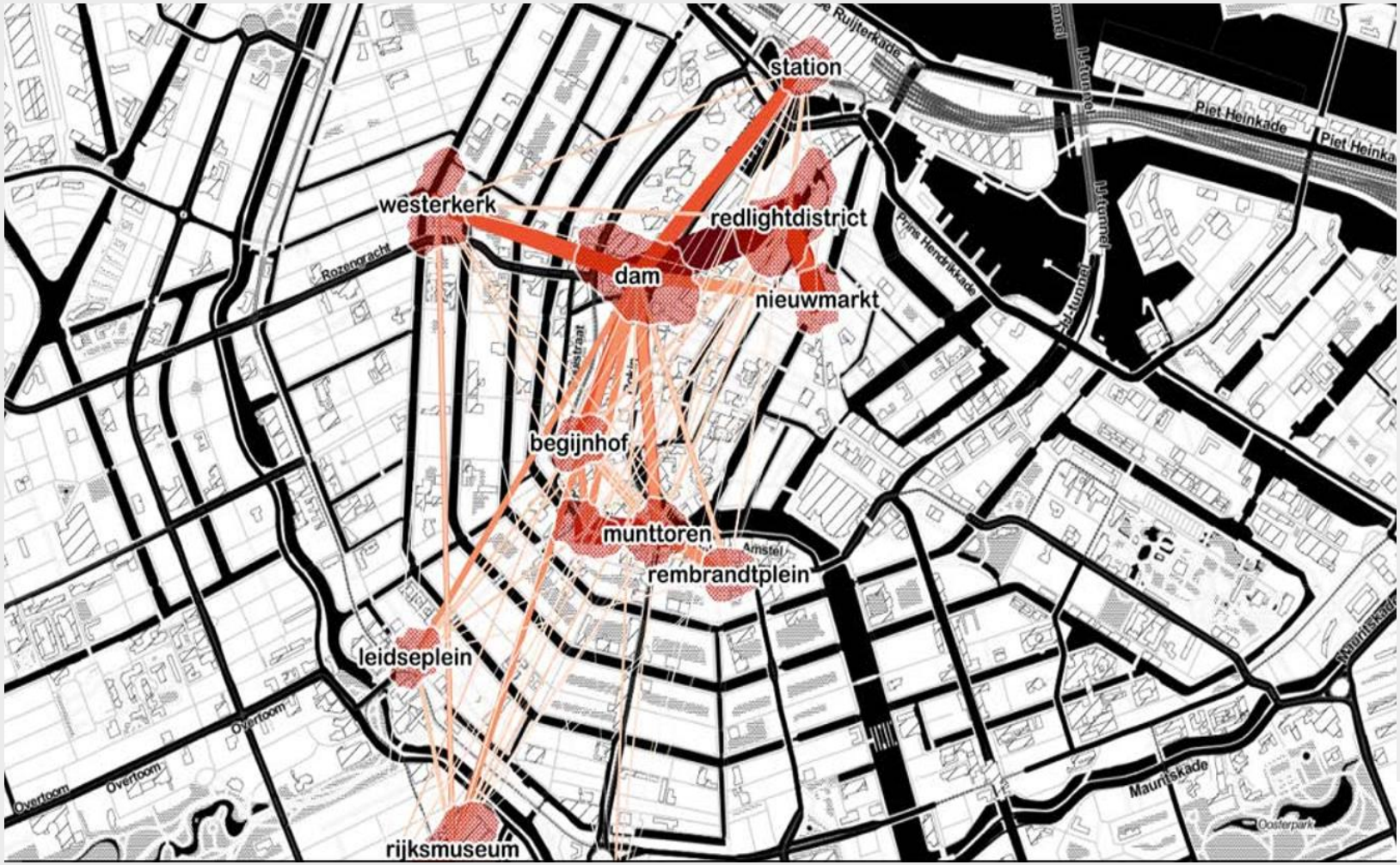
- Voordelen:
  - Achtergrondkenmerken beschikbaar en mogelijkheid om te combineren met vragenlijst (in feite traditionele statistiek; representativiteit)
  - Sprake van GPS-data (locatie nauwkeuriger dan mobiele telefoondata)
  - Consent van deelnemers
- Voor gebruik in OViN. Ook kijken naar toerisme.
- CBS heeft al twee jaar geëxperimenteerd (voor ICT-onderzoek)
- Software beschikbaar (vaak niet alle platformen)
- Zie ook: elektronische kortingskaart/app (korting, informatie; vergelijk Taiwan)
- Issues
  - Popup vragen (Kanthar)
  - Batterij snel leeg
  - Over de streep trekken van deelnemers (voor Inkomend toerisme 2018?)

3. Sociale media



Bron: Sander van der Drift

Flickr



Bron: Sander van der Drift

Toeristen stromen

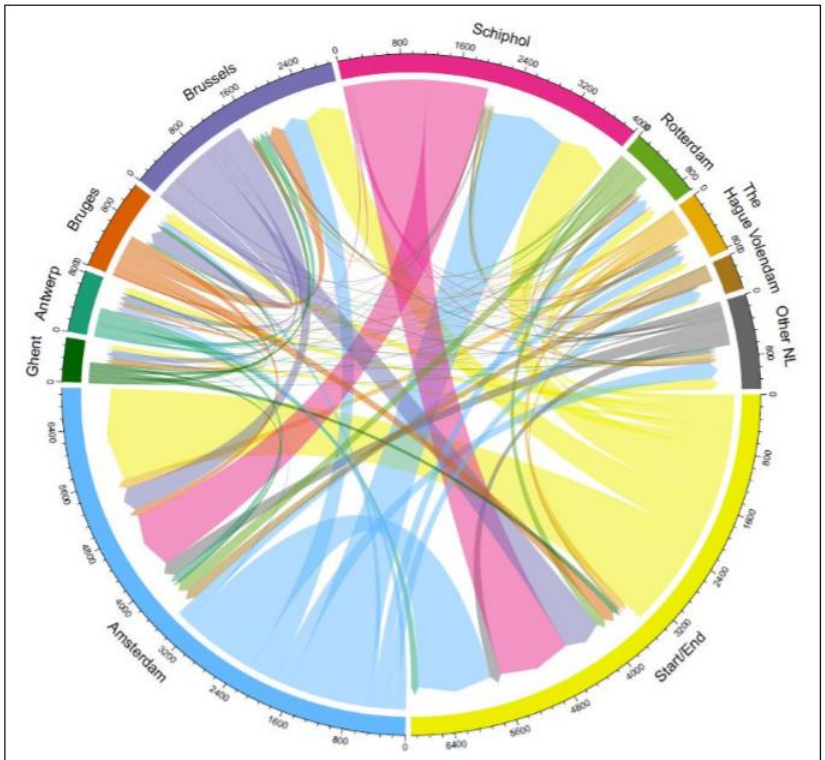
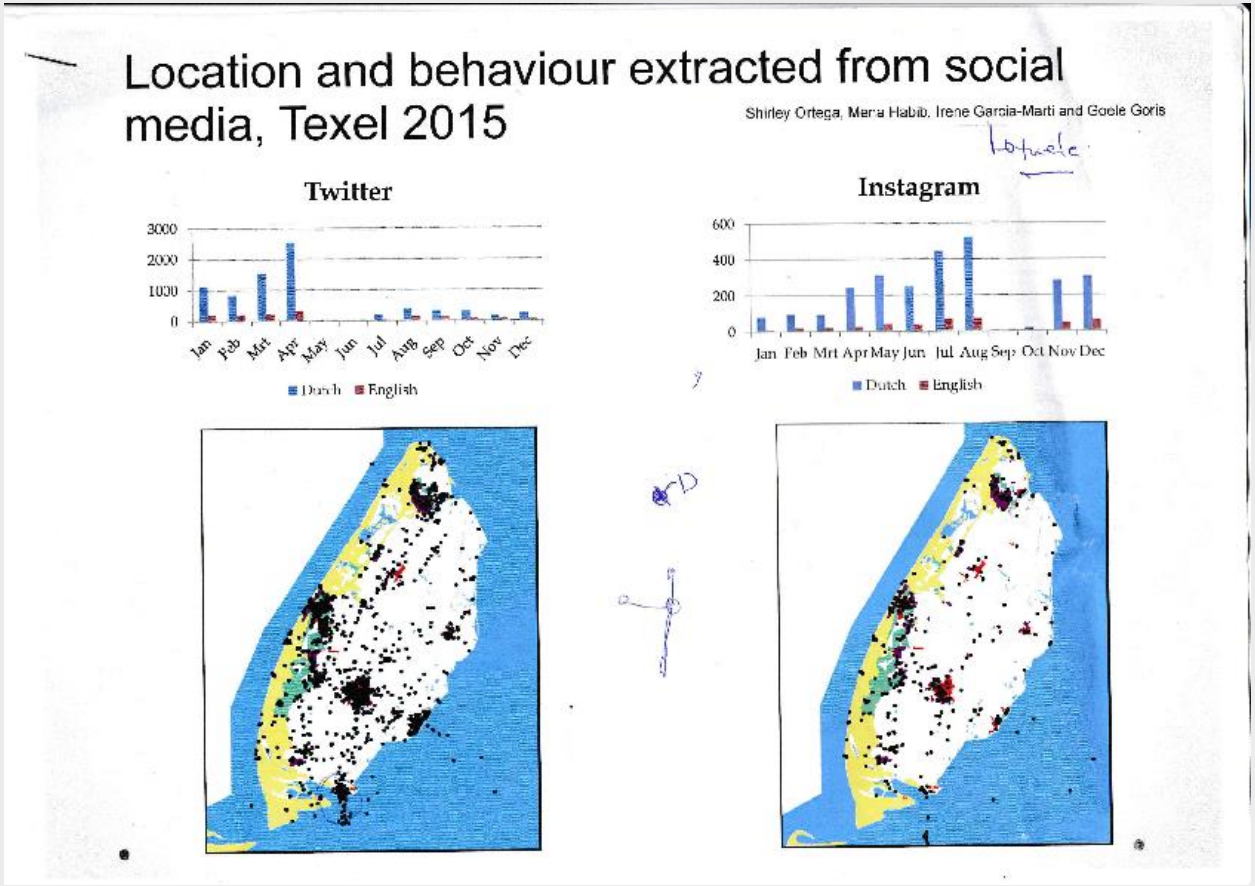


Figure 8. Footprint patterns across cities. Arrow size represents the number of tourists moving from one location to the next. Graph created with R package "circlize" (Gu et al, 2014).

Bron: M. Habib / N. Krol

Twitter posts op Texel



Locatiedata naar activiteit op basis van woordenwolken in posts

(Wie is de toerist?)

Bron: Shirley Ortega

Stand van zaken (bij CBS)

- Vergelijkbaar met mobiele telefoondata: locatie en tijd (klein deel van posts heeft deze kenmerken)
- Locatie gebaseerd op GPS (nauwkeuriger dan mobiele telefoon data).
- Data relatief gemakkelijk te verkrijgen (geen directe privacy issues).
- Kan meer informatie halen uit posts, o.a. positieve en negatieve sentimenten (echter vaak nog primitieve algoritmes) of woorden(wolken).
- CBS is gestart met een experiment. Mogelijk samenwerking met universiteiten Maastricht en Utrecht. Probleem: resources
- Issues
  - Bepalen van toerist moeilijk
  - Taal niet altijd goede ingang voor onderscheid naar land van herkomst.
  - Toeristen uit verschillende landen gebruiken verschillende sociale media
  - Representativiteit

4. Wikipedia

Evaluate the use of these data as a source of information for the identification of factors that drive tourism to an area and whether it is possible to predict tourism flows using these data.

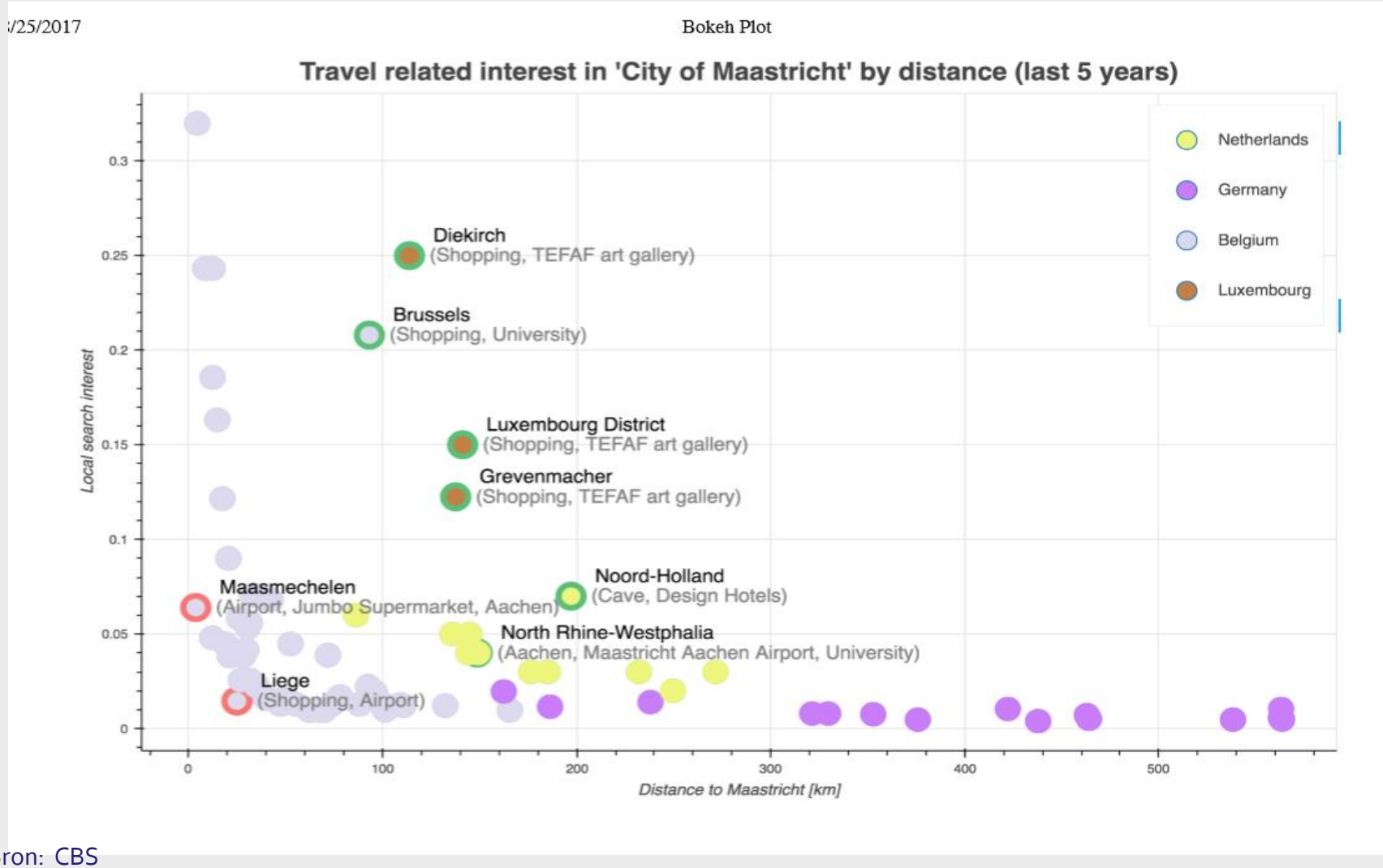
Bron: Eurostat

Stand van zaken

- Aardige manier om te kijken naar de factoren achter de aantrekkingskracht van steden of regio's.
- Points of interest (locatie).
- Eurostat.
- Issue:
  - Representativiteit (wie gebruikt Wikipedia en wie is actief op Wikipedia), misschien beter Googletrends (op zoektermen).

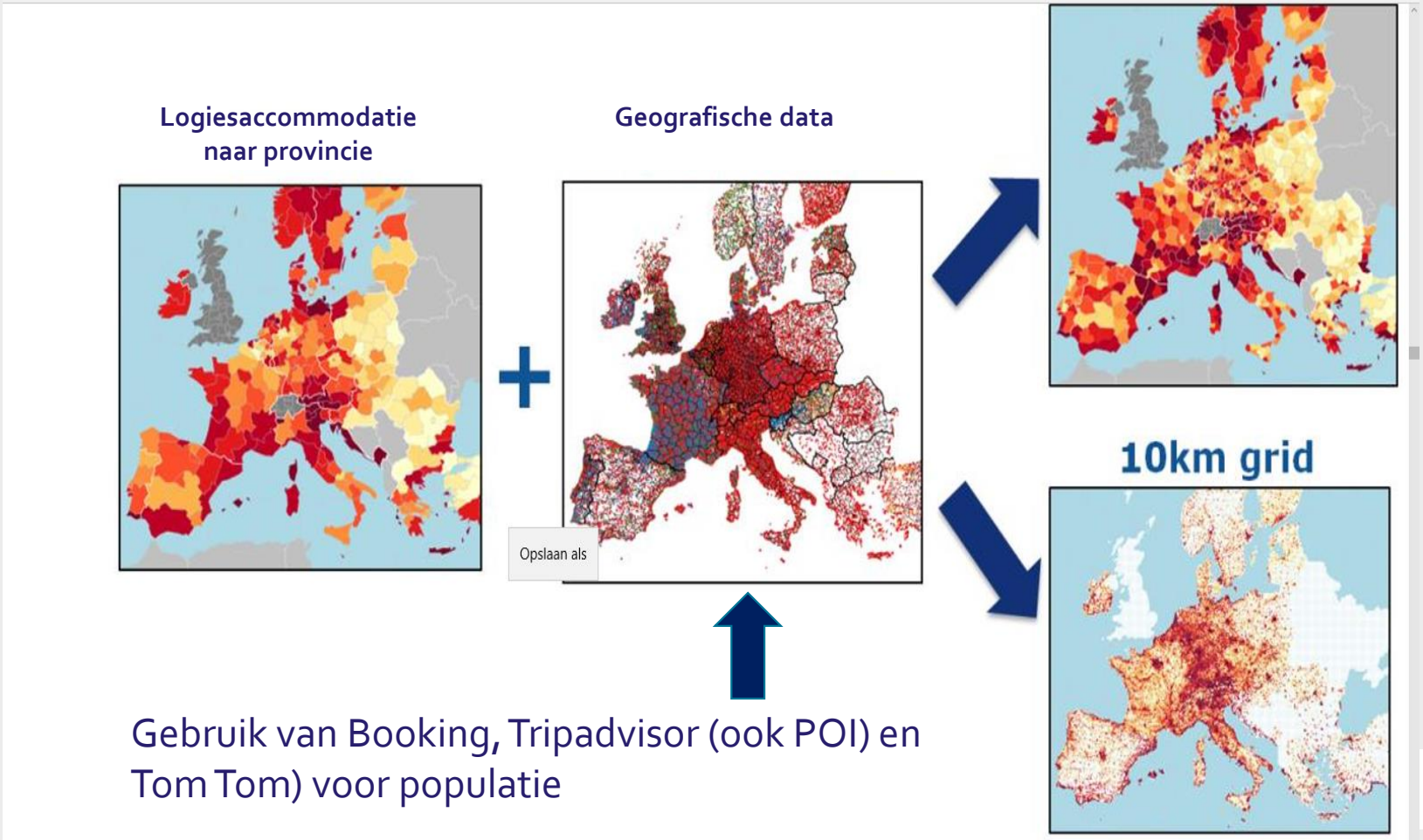
Bron: NRC

5. Googletrends



Bron: CBS

6. Web scraping



Gebruik van Booking, Tripadvisor (ook POI) en Tom Tom) voor populatie

Bron: Eurostat

Stand van zaken (bij het CBS)

- Gebruik van scraping voor het bepalen van de populatie (en kenmerken) Logiesaccommodaties, o.a. Booking, boerencampings, Airbnb.
- CBS heeft veel ervaring met het scrapen van o.a. bedrijvensites (bedrijfskenmerken), CPI, woningen, vacatures e.d. Ook in Europees verband.
- Verder detailleren resultaten Logiesaccommodaties. Populatie wordt bepaald aan de hand van Booking e.d. Vervolgens wordt met verschillende algoritmen gasten en overnachtingen uit de statistiek Logiesaccommodaties verdeeld naar kleinere regio's
- Issue: betrouwbaarheid voor kleine regio's.
- Meer op het aanbod van toerisme

- Activiteiten op het terrein van de platformeconomie (met name ook logies en vervoer), zowel bij EZ als Eurostat. Ook CBS.

Bron: NRC

Verder

- CBS onderhandelt over het verkrijgen van **creditkaart- en pinpasdata.** Misschien is volgend jaar iets mogelijk. Scannerdata van kassa's al beschikbaar.
- **Nachtregister**: afspraken lopen met Amsterdam, maar als mogelijk uitbreiden Nederland.
- CBS maakt gebruikt van **douane-gegevens** bij Caribisch Nederland, gebaseerd op **vluchtgegevens.** Onduidelijk of dat ook kan voor Nederland?
- **Openbaarvervoer data**? Er wordt naar gekeken. Beginfase.
- CBS heeft beschikking over satellietdata
- CBS heeft beschikking over **belastingdienst- en werkgelegenheidsdata** (Polis, ook big data).
- CBS maakt statistieken op basis van **verkeerslussen**. Mogelijk ook te koppelen aan het domein toerisme.

Samenvattend:

- Erg gebeurt erg veel op het terrein van nieuwe databronnen en het gebeurt overal.
- Toerisme favoriet onderwerp
- Vaak nog in experimentele fase (Gartner's hype curve). Nog weinig echte statistiekproductie
- Veel aandacht voor vraagkant (gedrag van toeristen), minder voor aanbodkant (bedrijven, POI's)
- Moet je als organisatie focussen: wat is belangrijk en kansrijk? Ook methodologie, toegang tot data, juridisch, dataknooppunt en één cijfergedachte (wat is nog waar?)
- Vaak onvoldoende resources / middelen. Niet concurreren, maar samenwerken!
- Waar hangt het laaghangend fruit?
- Hoe gaan beleidsmakers om met deze nieuwe werkelijkheid?