Mapping human population

A data science approach

Martijn Tennekes
London, autumn 2019

Official Statistics

- Statistics published by government agencies or other public bodies (e.g. international organizations) as a public good.
- National Statistical Institutions (NSIs) describe economic and social phenomena in a country and respond to national developments and events.

Statistics Netherlands

Office for National Statistics

eurostat

Netherlands

United Kingdom

European level

Data sources for Official Statistics

- **Survey data**

- **Administrative data**

Examples: tax data, health data, insurance data, vehicle registration data.

- **Big data**

Examples: mobile phone road sensor social media internet

Data science in official statistics

Traditional statistics	Modern statistics
Need for data	Abundance of data
Hypothesis-driven	Data-driven (machine learning)
Frequentist approach	Bayesian approach
Tables	Data visualization
Fixed (scheduled) topics	Hot topics (e.g. climate change)

Mapping human population

Many statistics are about humans:

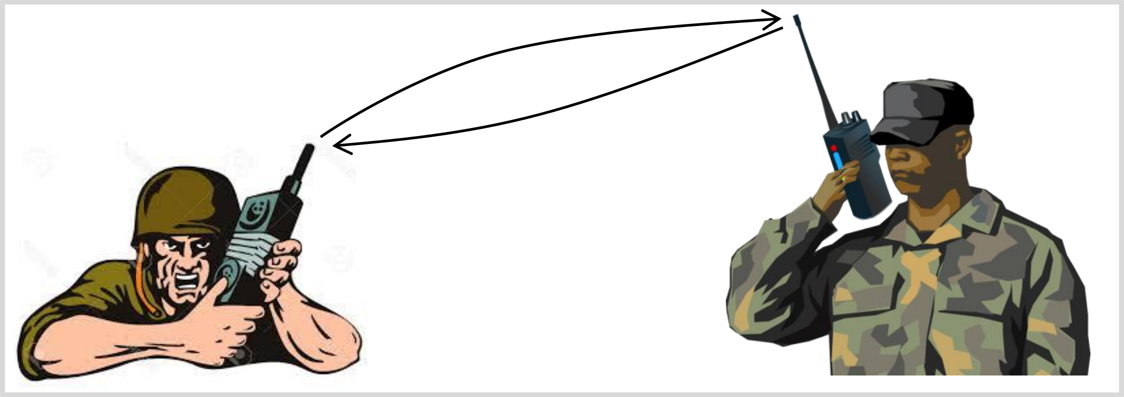
- What are the demographics of certain regions?
- Where are people during daytime?
- How do people commute?
- Where do tourists go to?
- …

Two projects are presented in this presentation:

- **Mobile phone data**; a new source with huge potential
- **Dot maps**; a visualization method for spatial data.

Mobile phone data

Predecessors of Mobile Phones

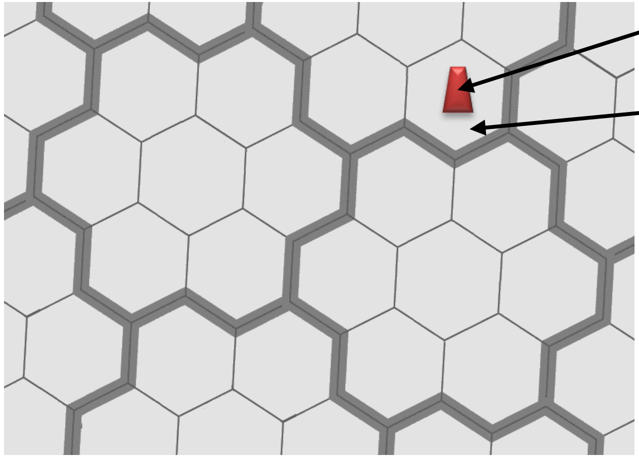


Walkie-talkie

Car telephone system

Why are mobile phones called ‘cell-phones’?

The target area is chopped into small cells such that each cell is covered by a cell tower.



Hexagon cell-plan

Advantages:

- Close proximity to antennas -> small batteries
- Communication frequencies can be reused without disturbance from other antennas

Type of antennas

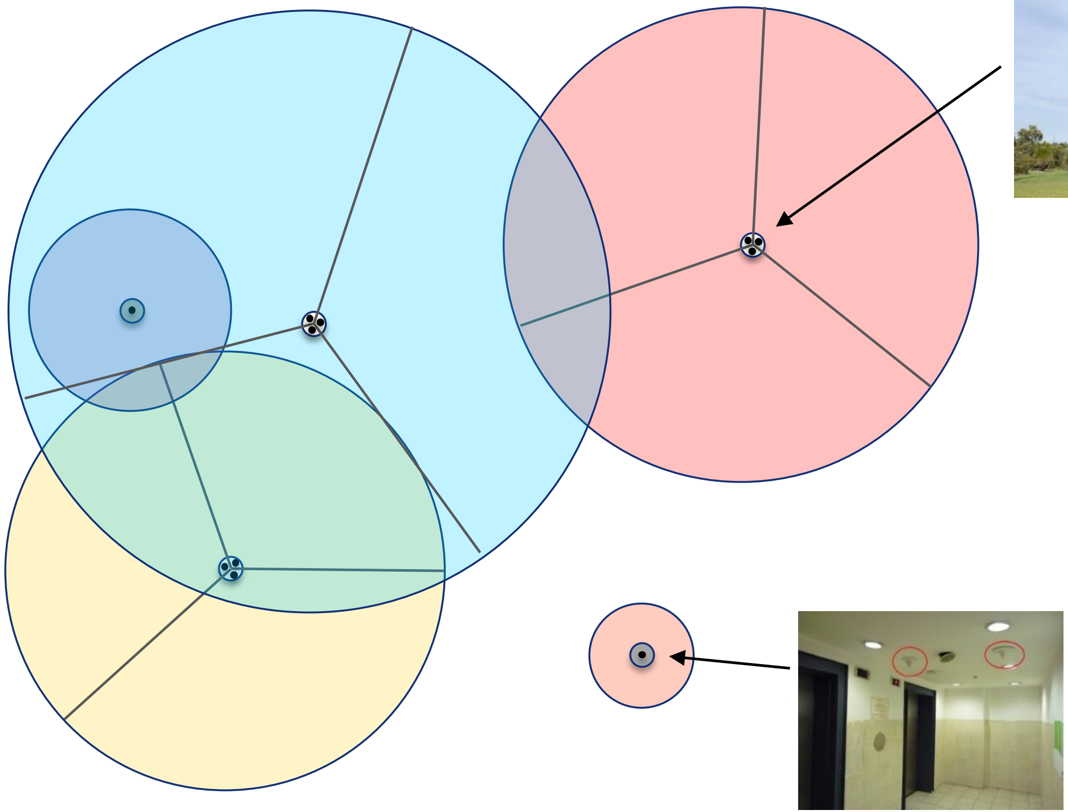
Cell tower
- 3 antennas, each covering 120º
- Coverage up to 40 km

Rooftop cell site
- Coverage up to 40 km

Indoor cell
- Coverage 200 m

Small cell
- Coverage up to 2 km

Simplified cell-plan…



- Cell site (BTS)
- Cell antenna
- Cell coverage area

Mobile phone generations

	Generation / description	Year of introduction
0G	Mobile radio telephone, used in car telephones.	1940's - 1970's
1G	Mobile analog telecommunications.	1981
2G	Global System for Mobile Telecommunications (GSM) standard. Digital encryption used. Introduction of SMS and MMS messages.	1991
3G	Universal Mobile Telecommunications Service (UMTS) and CDMA2000 standards. Introduction of mobile internet. 10 Mb/s	2001
4G	Mobile broadband data, including voice over data. Enabling video conferencing and cloud computing. Download rates: - 100 Mb/s at high mobility (cars/trains) - 1 Gb/s at low mobility (pedestrians)	2008
5G	High speed mobile internet. Probably around 10Gb/s.	2020

Signaling data / Call Detail Records

Signaling data
- 100 variables, e.g.
  - Antenna id (geolocation)
  - Time/date
  - Country
  - Provider
  - Type of event
- Hundreds of records per device per day (4G)

Mobile phone usage

- **Call** (incl. being called)
- **SMS** (send and receive)
- **Data** (continuous logging)

Events trigged by movements, e.g. handovers from one area to another.

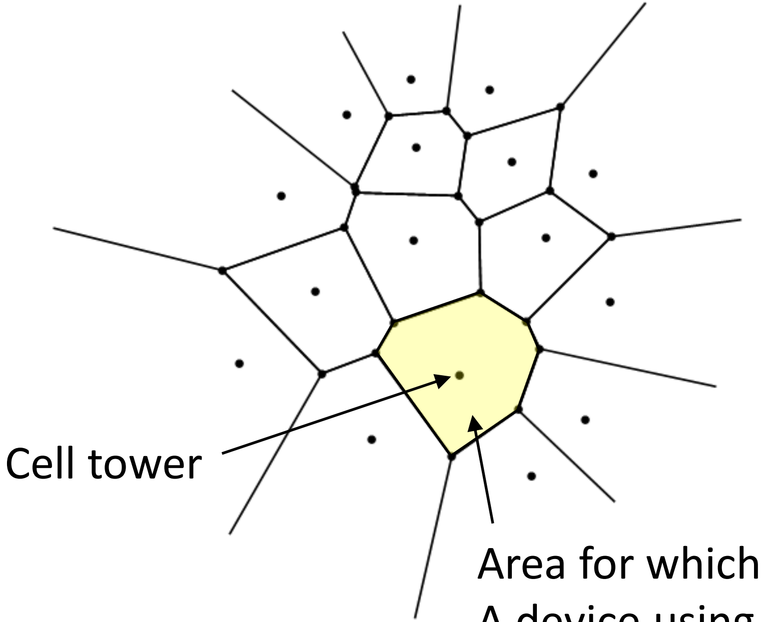
Call Detail Records (CDR)
- Used for billing
- Every provider should have them

Applications for Official Statistics

1. **Day Time Population**: the number of people in a certain region at a certain time. Useful for visitor counts during events, infrastructure planning, emergency management.
2. **Tourism statistics**: what places do they visit, where do they overnight, where do they come from?
3. **Commuting patterns**: where do people live and work? How and when do they commute?
4. **Urban planning / smart city**: what trips do people make in urban areas? By what mode of transport?
5. **Social networking**: who is connected to whom?
6. **Natural disasters**: what are the migration flows over time?

How to determine geolocation?

Voronoi tessellation



Area for which it is the nearest cell tower
A device using the cell tower is supposed to be somewhere in this polygon (uniform distribution)

Assumptions:

- All antennas are omnidirectional
- Areas do not overlap

Taking overlap into account

Bayesian approach

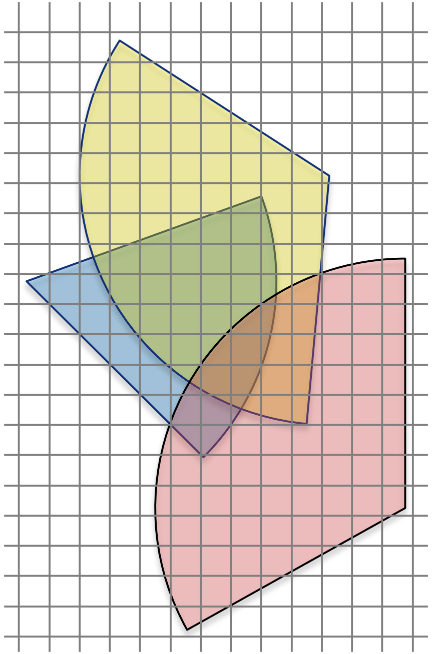
$$\mathrm{P}(g|a) = \frac{\mathrm{P}(a|g)\,\mathrm{P}(g)}{\mathrm{P}(a)}$$

where g is a grid cell and a an antenna

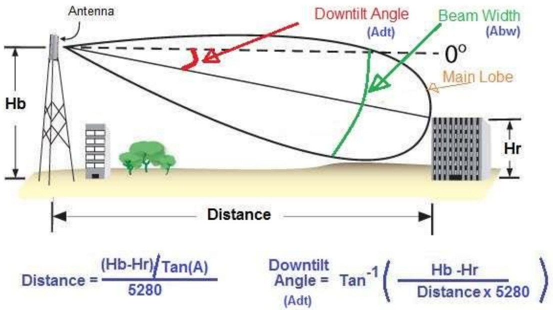
- $\mathrm{P}(g)$ specifies a prior probability that a device is in grid cell g
- $\mathrm{P}(a)$ serves as a normalization constant
- $\mathrm{P}(a|g)$ is the likelihood, which can be defined as:

$$\mathrm{P}(a|g) = \begin{cases} 0 & \text{if grid cell } g \text{ is not covered by } a \\ \dfrac{s(g,a)}{\sum_{a' : g \in B(a')} s(g,a')} & \text{if grid cell } g \text{ is covered by } a \end{cases}$$

where $s(g, a)$ the (relative) signal strength of antenna a in grid cell g and $B(a')$ is the set of grid cells covered by a'



Signal strength is complex in reality…

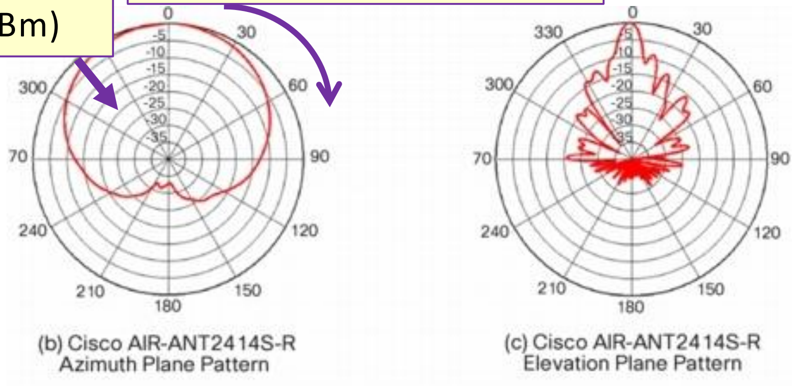


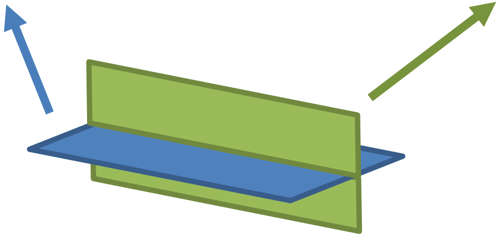
Beam (simplified) for which signal strength is good

Radiation plots for a specific antenna:

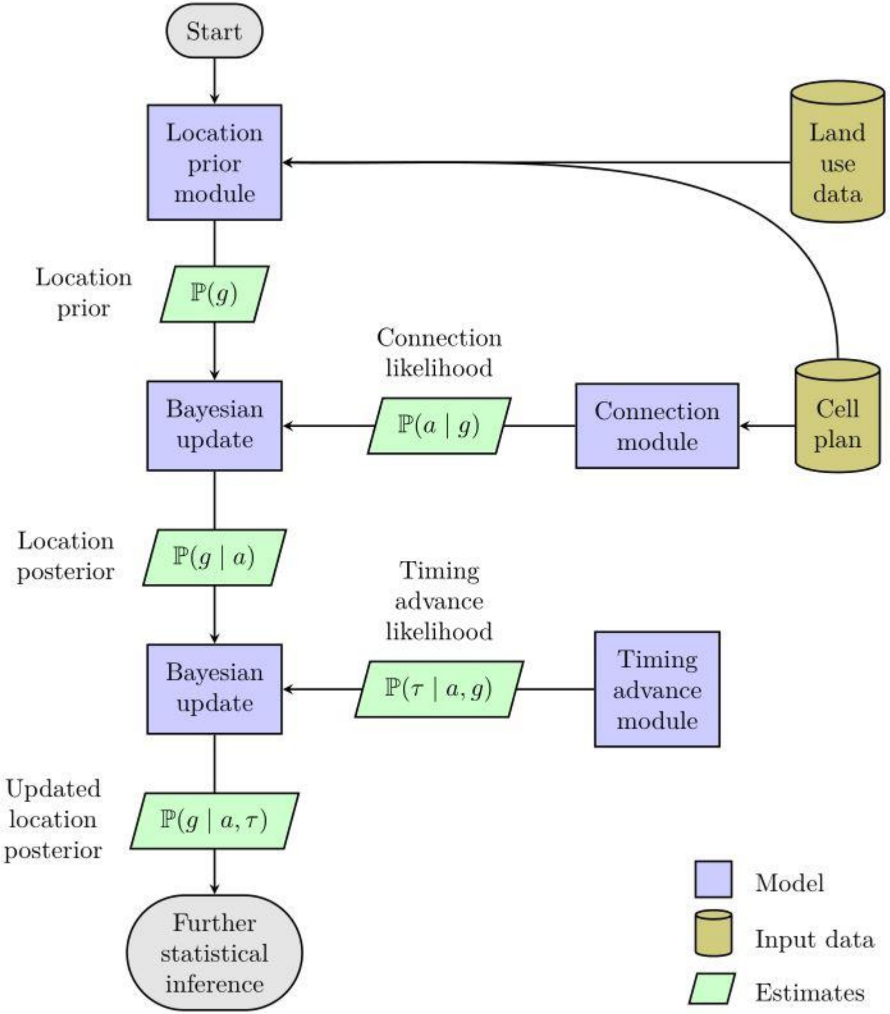
Angle w.r.t. main direction

Signal delta (dBm)

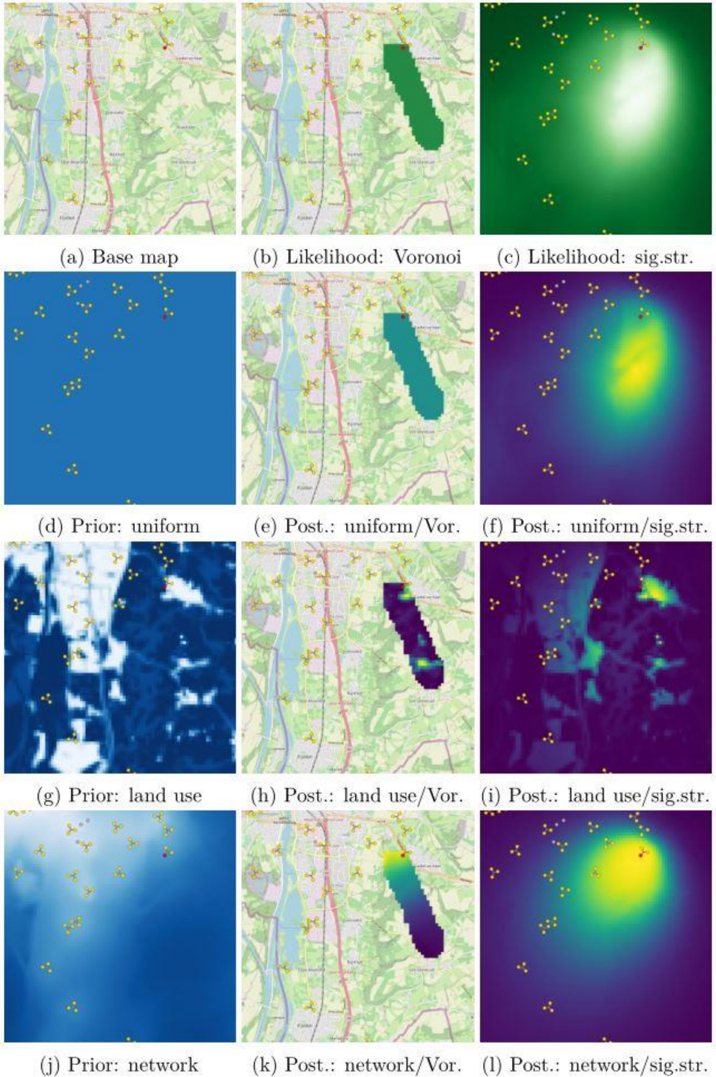




Location estimation process



Example



(a) Base map (b) Likelihood: Voronoi (c) Likelihood: sig.str.

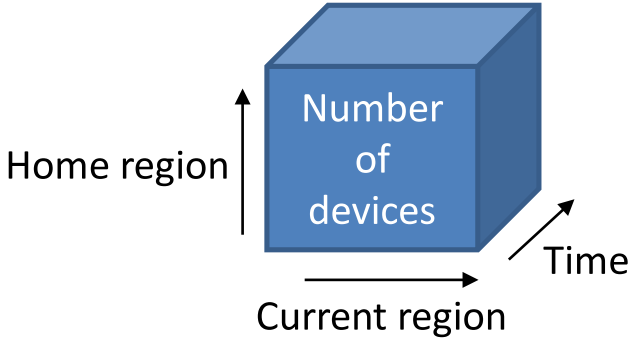
(d) Prior: uniform (e) Post.: uniform/Vor. (f) Post.: uniform/sig.str.

(g) Prior: land use (h) Post.: land use/Vor. (i) Post.: land use/sig.str.

(j) Prior: network (k) Post.: network/Vor. (l) Post.: network/sig.str.

From location to estimates

1. Deriving home location (needed because signaling data / CDR does not contain customer data. Method: find the 'home' antennas of a device, and map the probabilities to the administrative region of interest (e.g. municipality)
2. Aggregate likelihood values per time frame (e.g. one hour) per device
3. Data cube:



4. Calibrate with population registers and education registers.

Further research

- Validate the results with GPS data
- Use particle filter to estimate the route and mode of transport
- Use clustering methods to group devices based on their routes, for instance in order to detect tourists (domestic and foreign).

Dot maps

Classic dot map

Cholara outbreak in London (1854) by John Snow

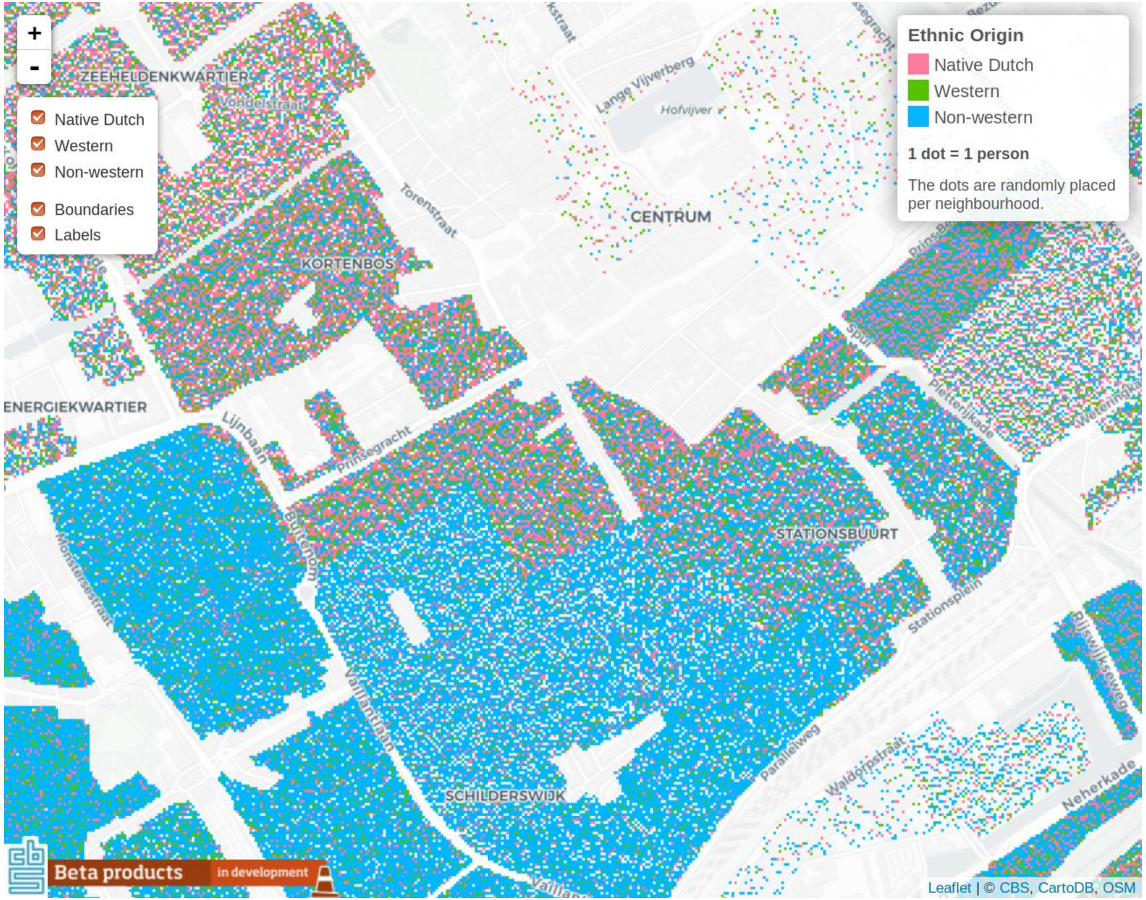
Dots instead of bars

Let there be… COLOR

Position of the dots: **density**

Colors of the dots: **composition**

What happens when you zoom out?



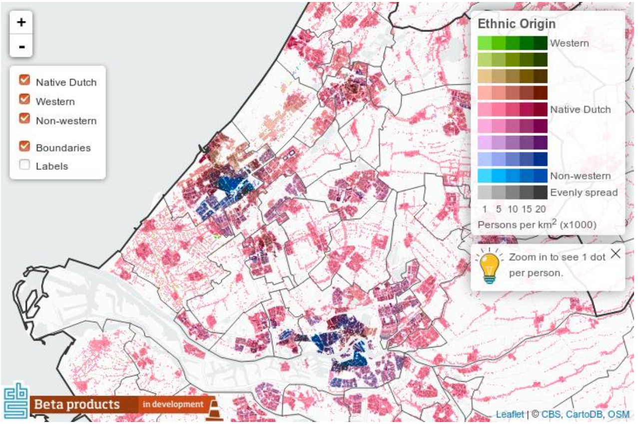
Position of the dots:
density

Colors of the dots:
composition

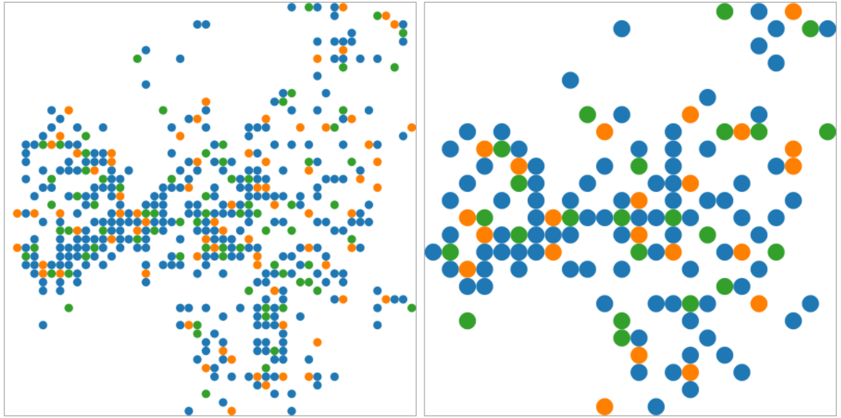
Out of pixels

How to aggregate the dots?

We propose two approaches:



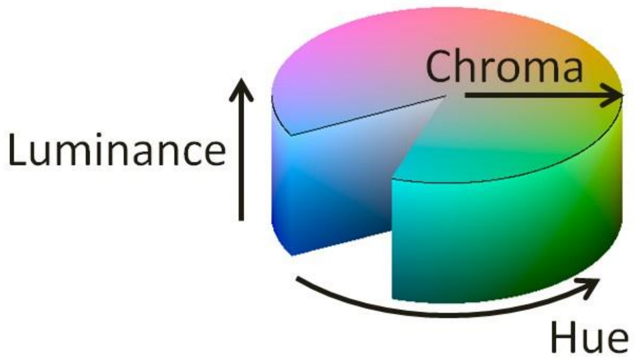
1. Blended colours

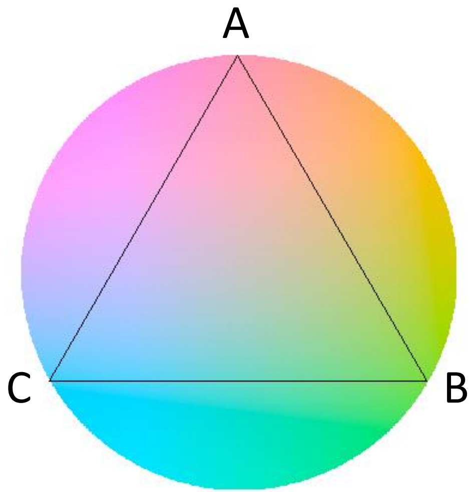


2. Super dots

Blended colours

Pixel colours are selected from the HCL colour space:



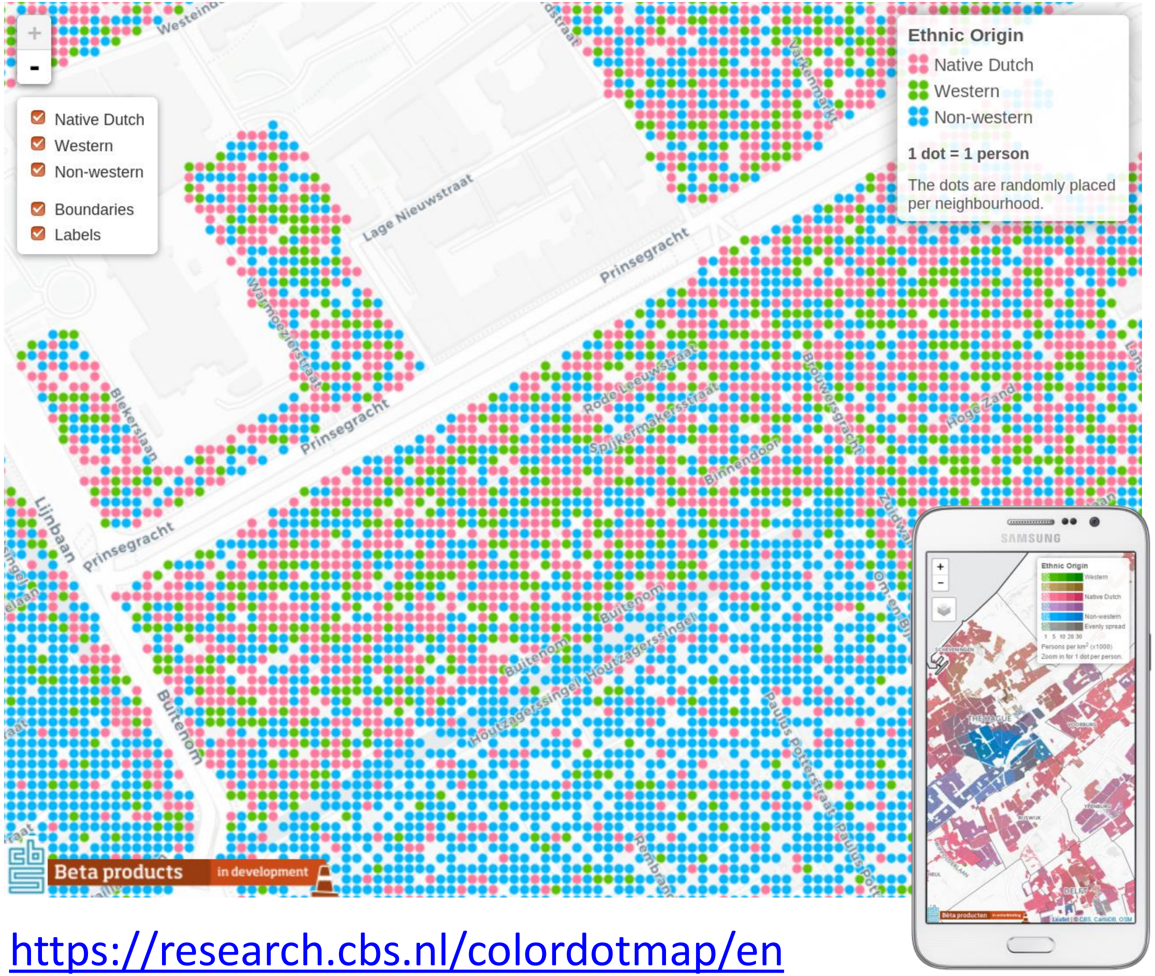


- **Luminance** for **density**
- **Hue** and **Chroma** for **composition**

Application

Migration background of the Dutch population

Dots are distributed uniformly per neighbourhood and placed in the land use category “residential”



Published a CBDS beta product: https://research.cbs.nl/colordotmap/en

Application

Migration background of the Dutch population

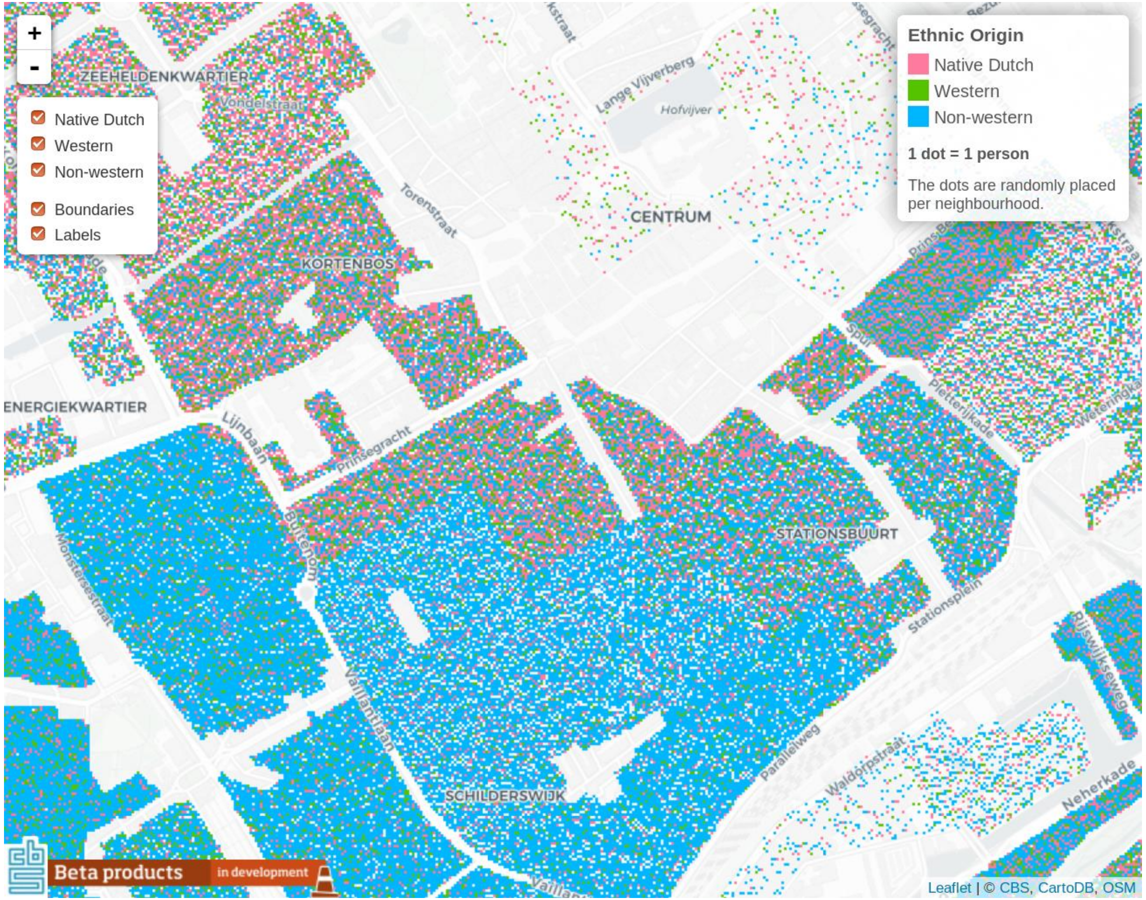
Dots are distributed uniformly per neighbourhood and placed in the land use category “residential”

Published a CBDS beta product: https://research.cbs.nl/colordotmap/en

Application

Migration background of the Dutch population

Dots are distributed uniformly per neighbourhood and placed in the land use category “residential”

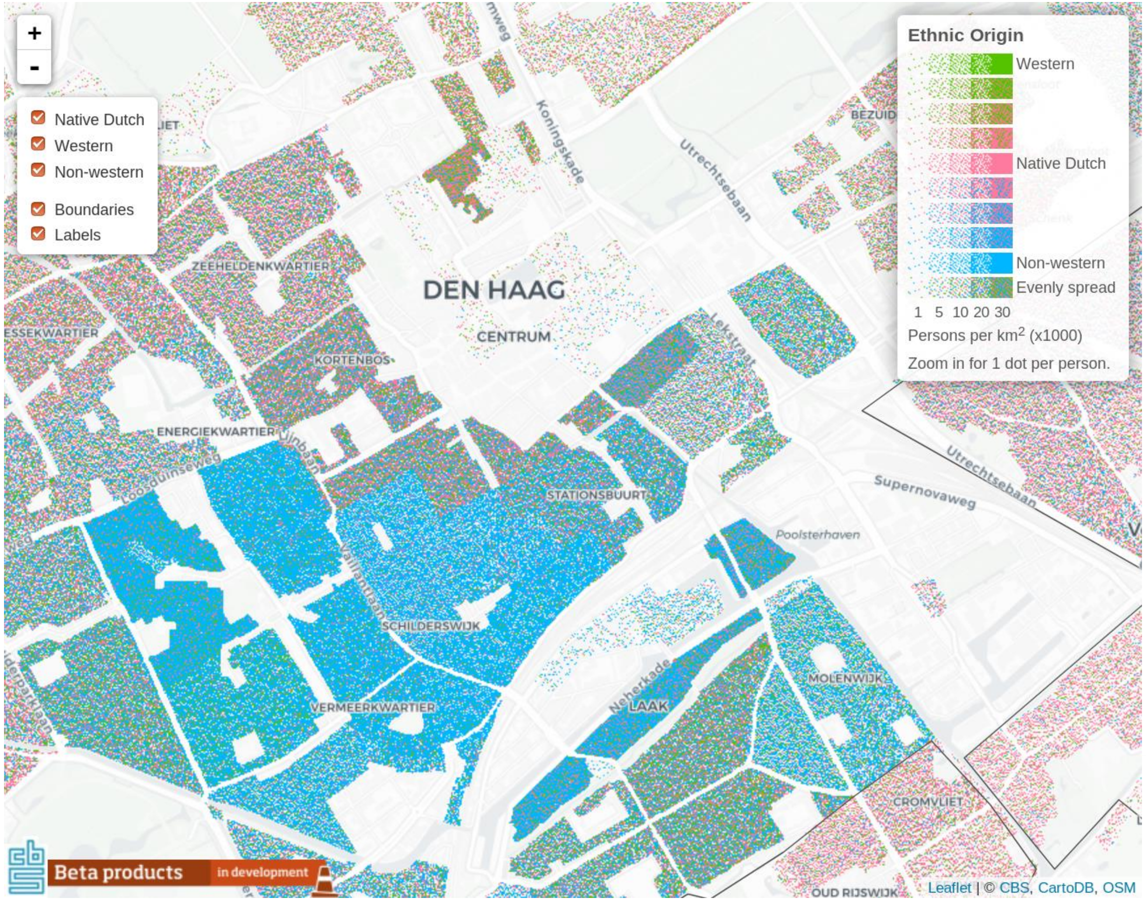


Published a CBDS beta product: https://research.cbs.nl/colordotmap/en

Application

Migration background of the Dutch population

Dots are distributed uniformly per neighbourhood and placed in the land use category “residential”

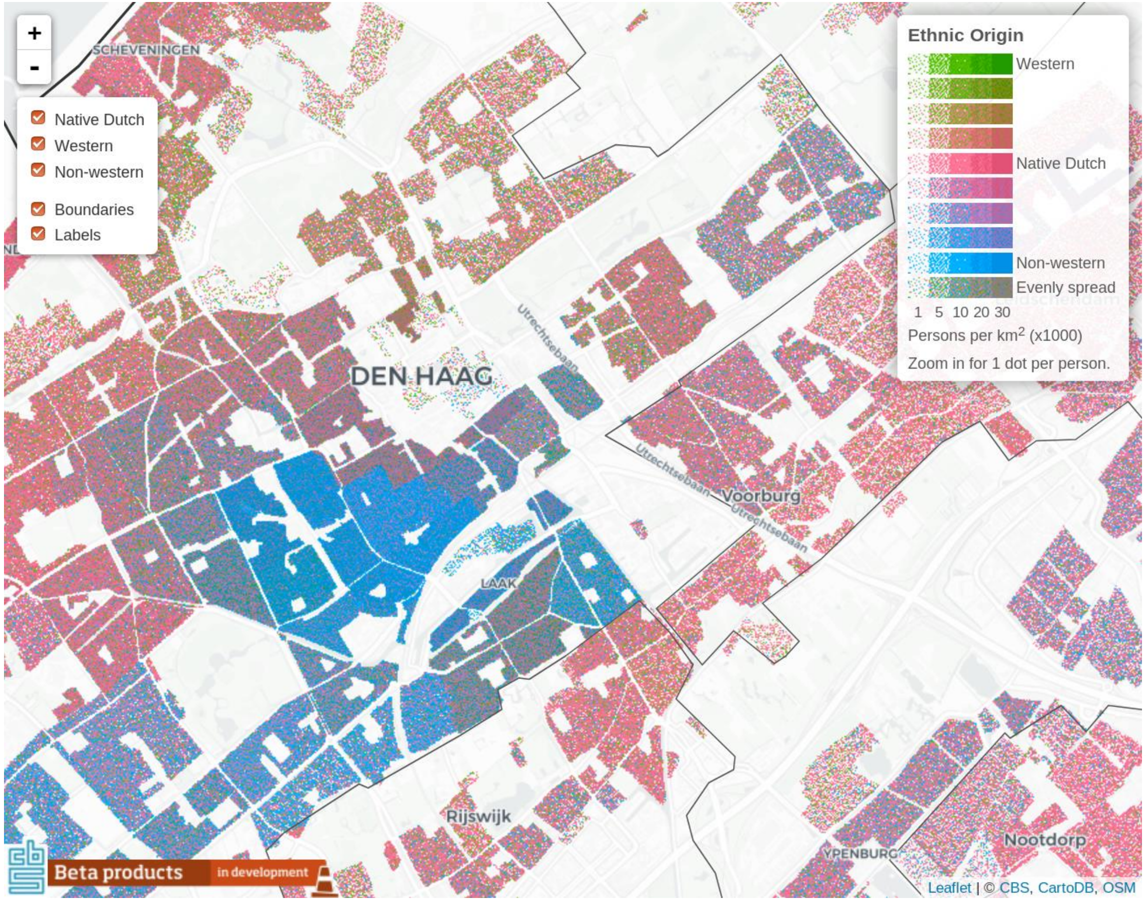


Published a CBDS beta product: https://research.cbs.nl/colordotmap/en

Application

Migration background of the Dutch population

Dots are distributed uniformly per neighbourhood and placed in the land use category “residential”

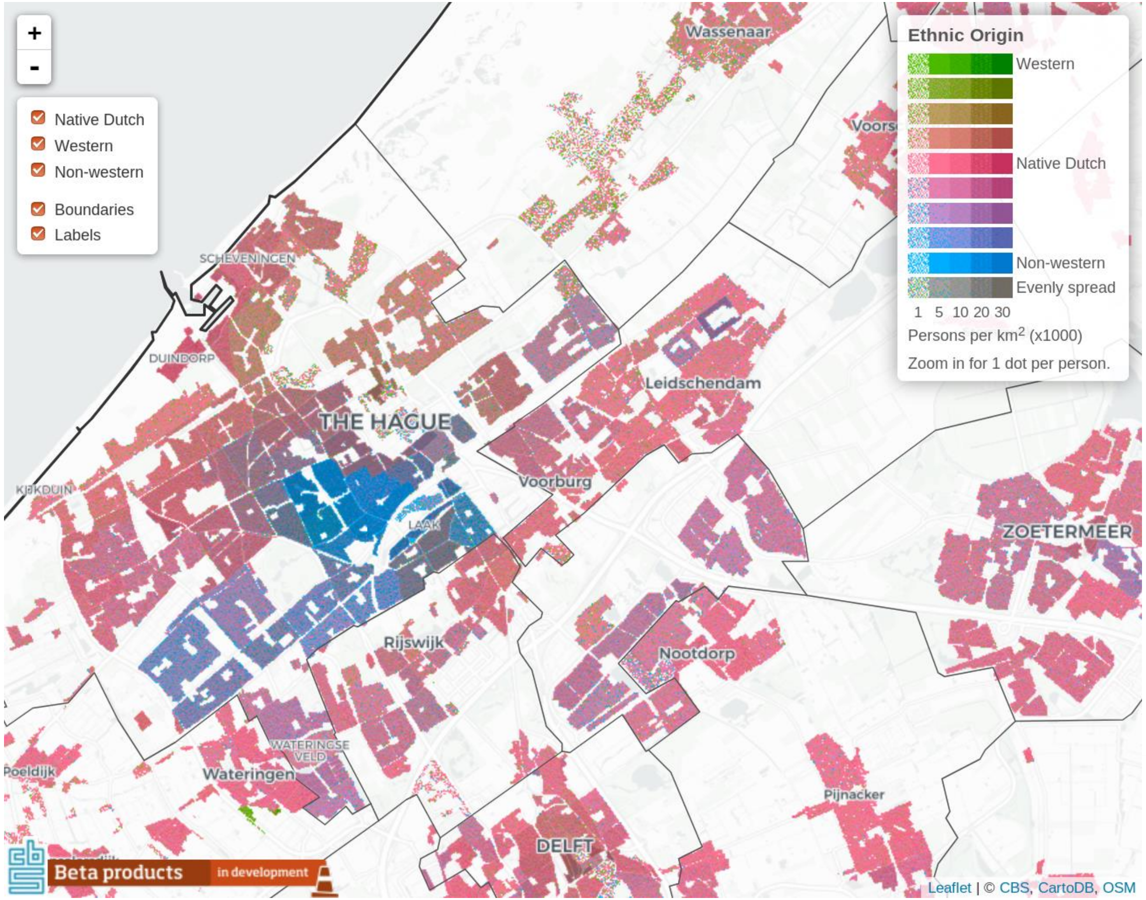


Published a CBDS beta product: https://research.cbs.nl/colordotmap/en

Application

Migration background of the Dutch population

Dots are distributed uniformly per neighbourhood and placed in the land use category “residential”

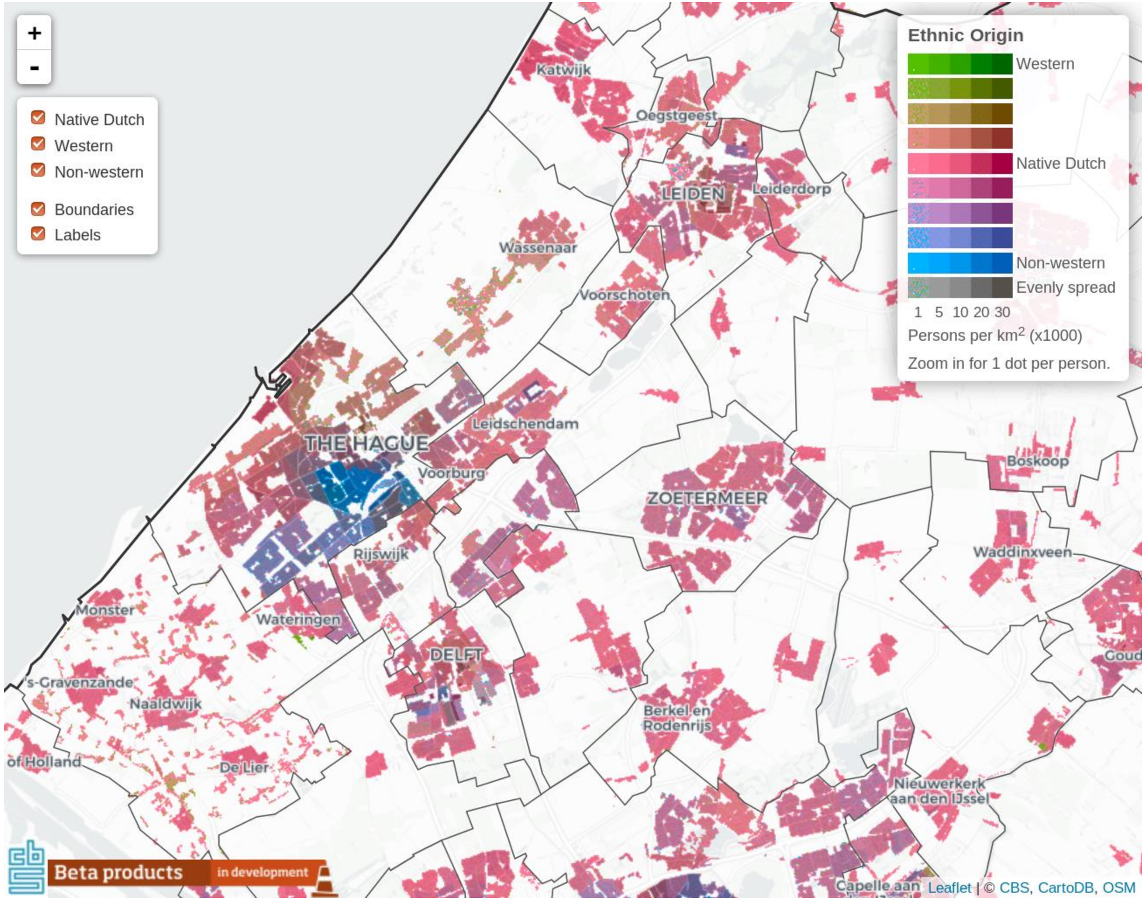


Published a CBDS beta product: https://research.cbs.nl/colordotmap/en

Application

Migration background of the Dutch population

Dots are distributed uniformly per neighbourhood and placed in the land use category “residential”

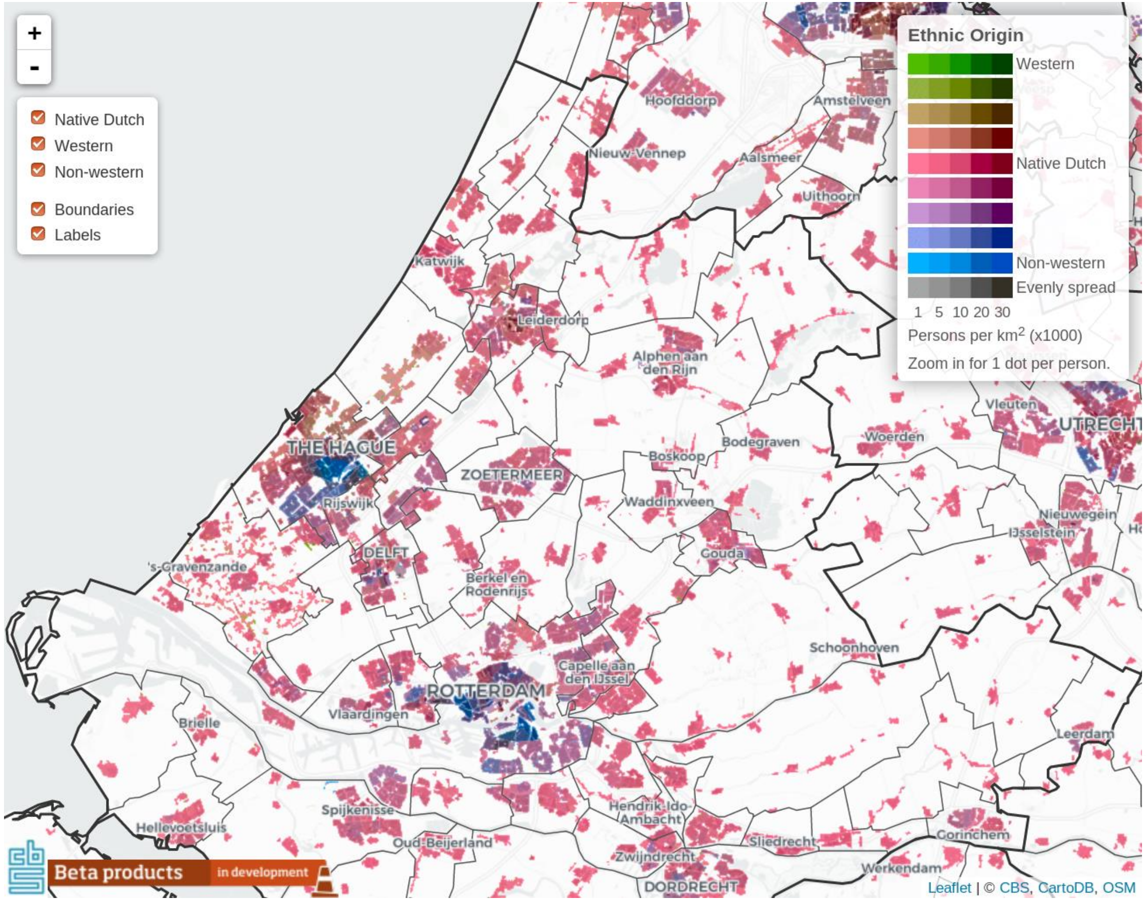


Published a CBDS beta product: https://research.cbs.nl/colordotmap/en

Application

Migration background of the Dutch population

Dots are distributed uniformly per neighbourhood and placed in the land use category "residential"

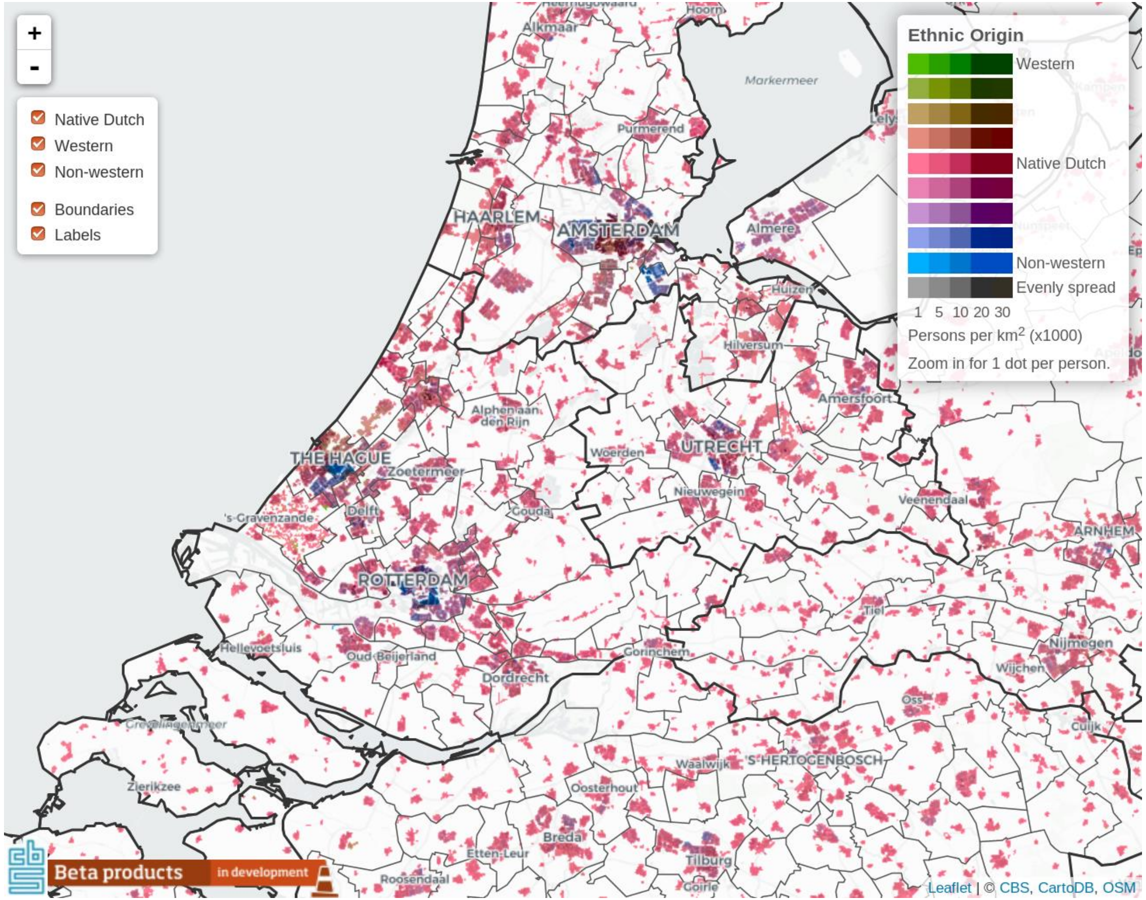


Published a CBDS beta product: https://research.cbs.nl/colordotmap/en

Application

Migration background of the Dutch population

Dots are distributed uniformly per neighbourhood and placed in the land use category “residential”

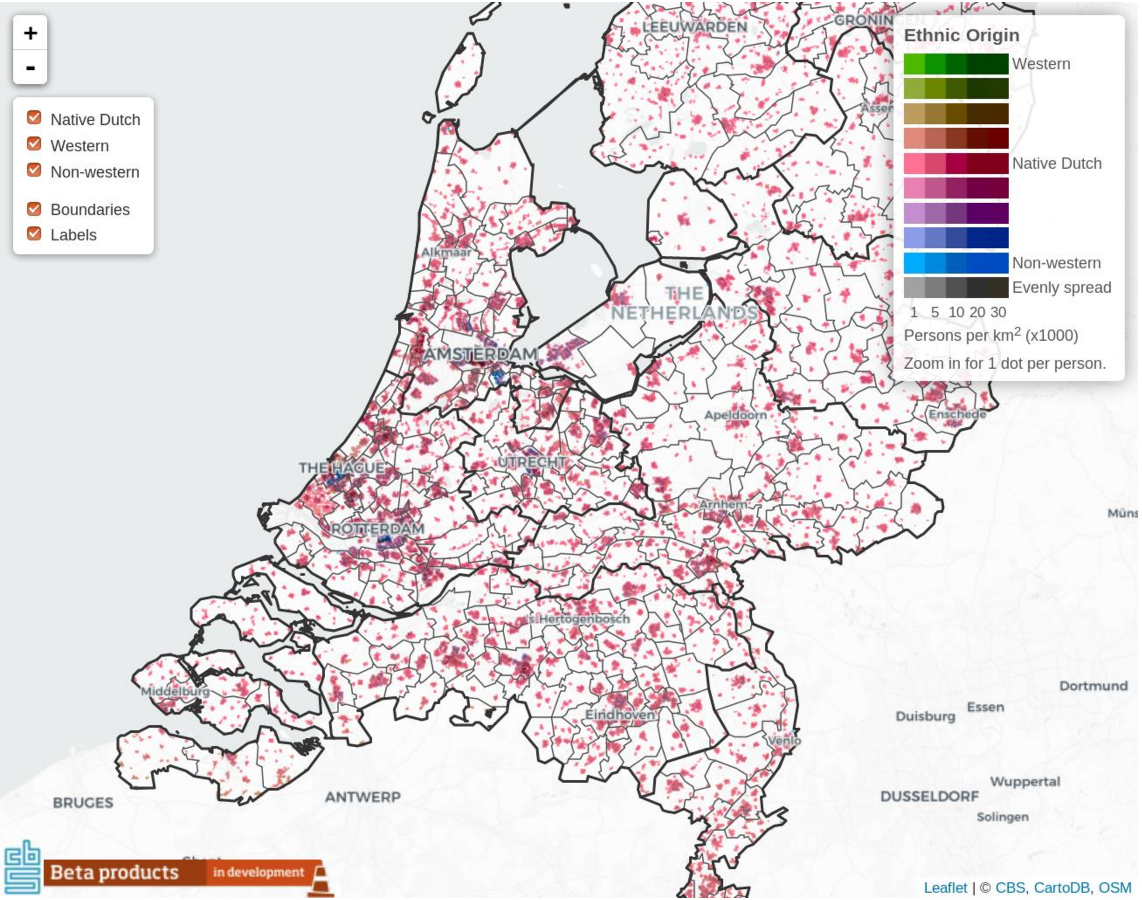


Published a CBDS beta product: https://research.cbs.nl/colordotmap/en

Application

Migration background of the Dutch population

Dots are distributed uniformly per neighbourhood and placed in the land use category “residential”

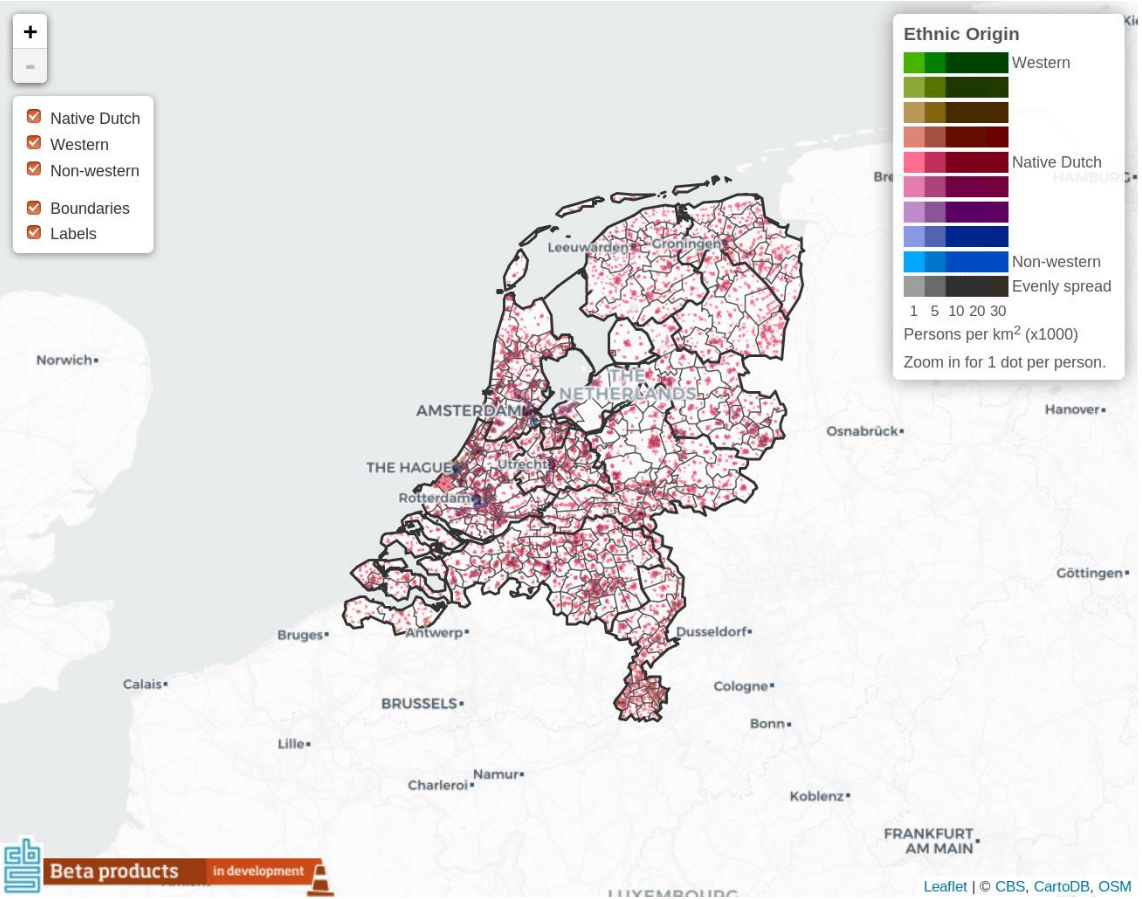


Published a CBDS beta product: https://research.cbs.nl/colordotmap/en

Application

Migration background of the Dutch population

Dots are distributed uniformly per neighbourhood and placed in the land use category “residential”



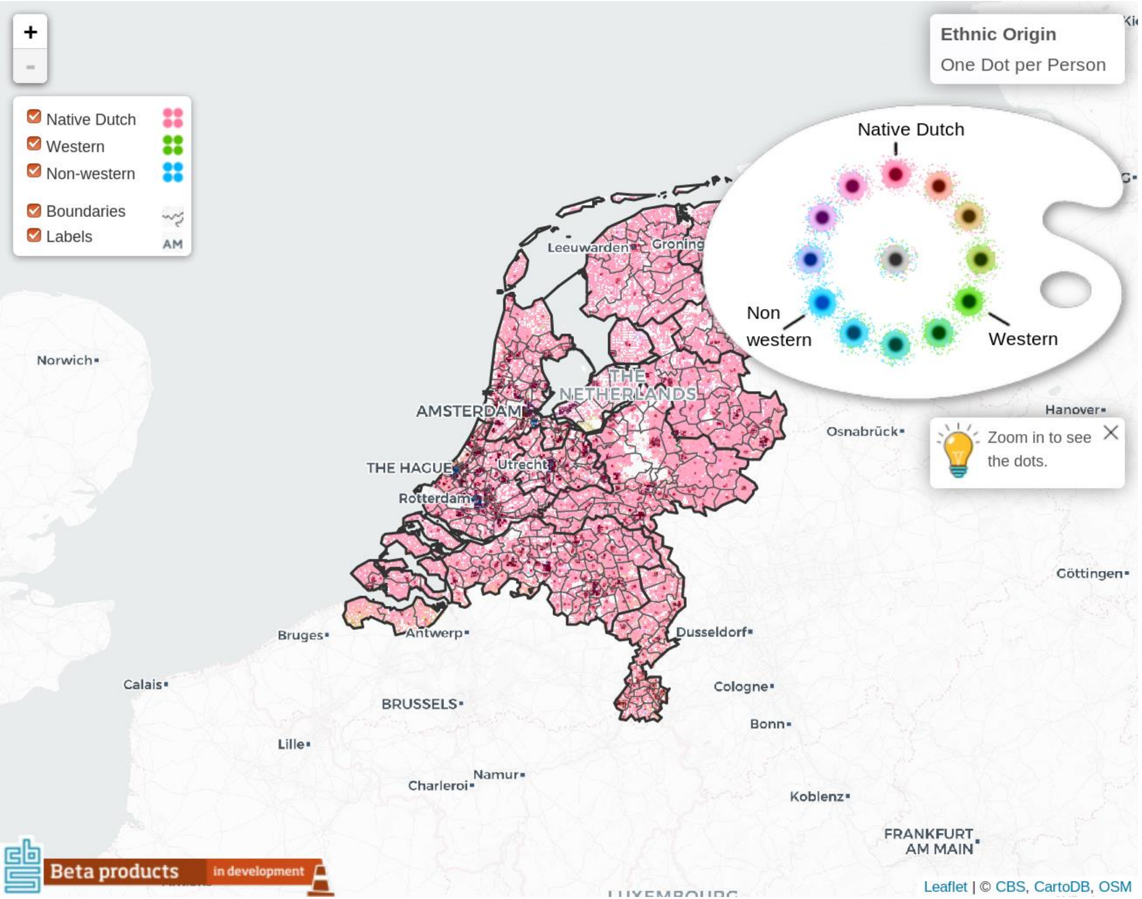
Published a CBDS beta product: https://research.cbs.nl/colordotmap/en

Application

Migration background of the Dutch population

Experimental version:

- Dots are placed in building areas (using the BAG register)
- “Artistic” legend

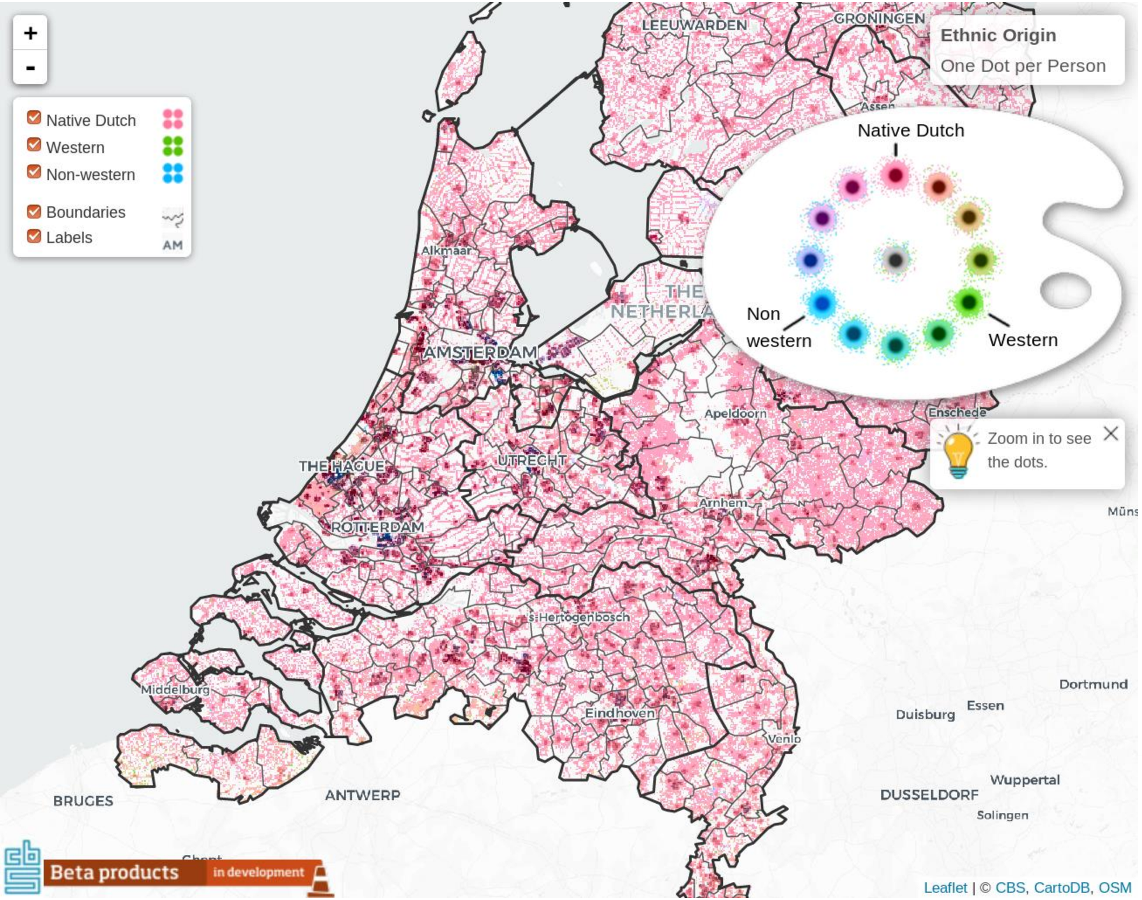


Application

Migration background of the Dutch population

Experimental version:

- Dots are placed in building areas (using the BAG register)
- “Artistic” legend

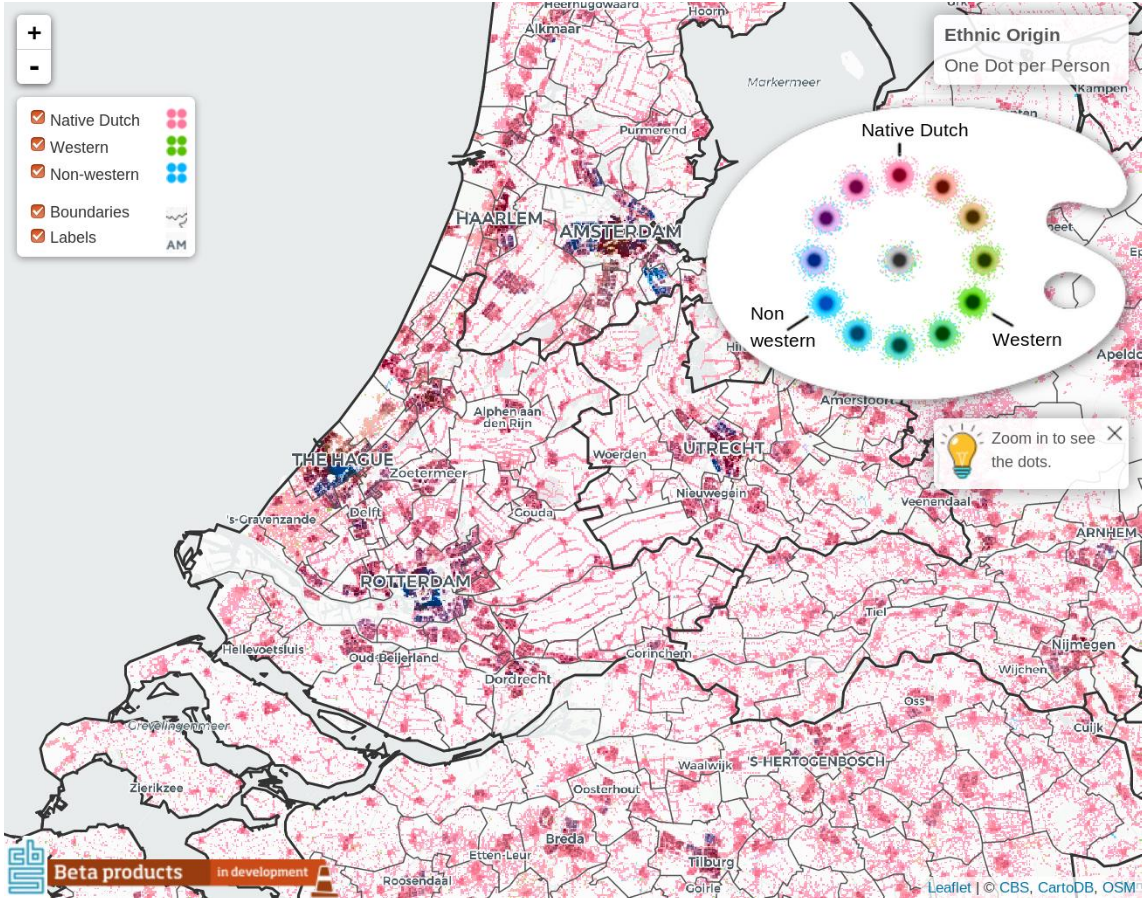


Application

Migration background of the Dutch population

Experimental version:

- Dots are placed in building areas (using the BAG register)
- “Artistic” legend

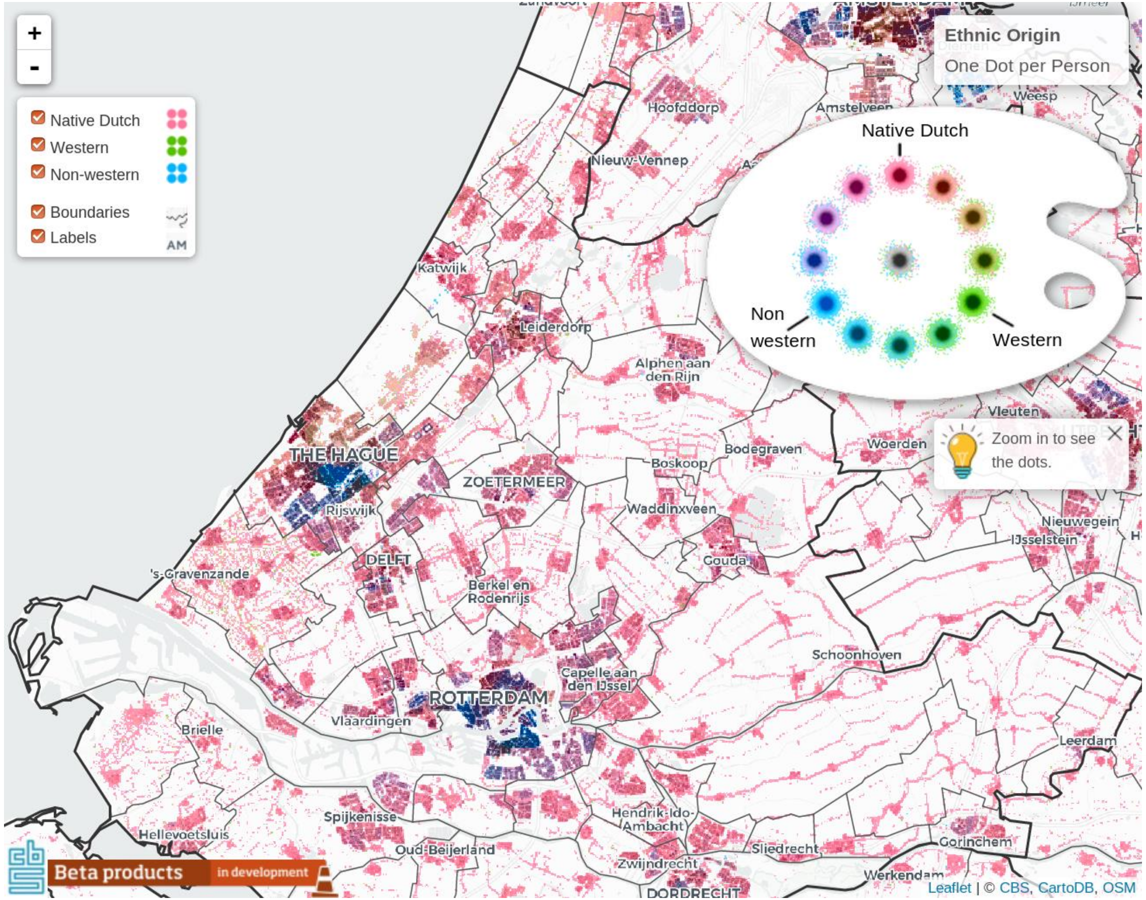


Application

Migration background of the Dutch population

Experimental version:

- Dots are placed in building areas (using the BAG register)
- “Artistic” legend

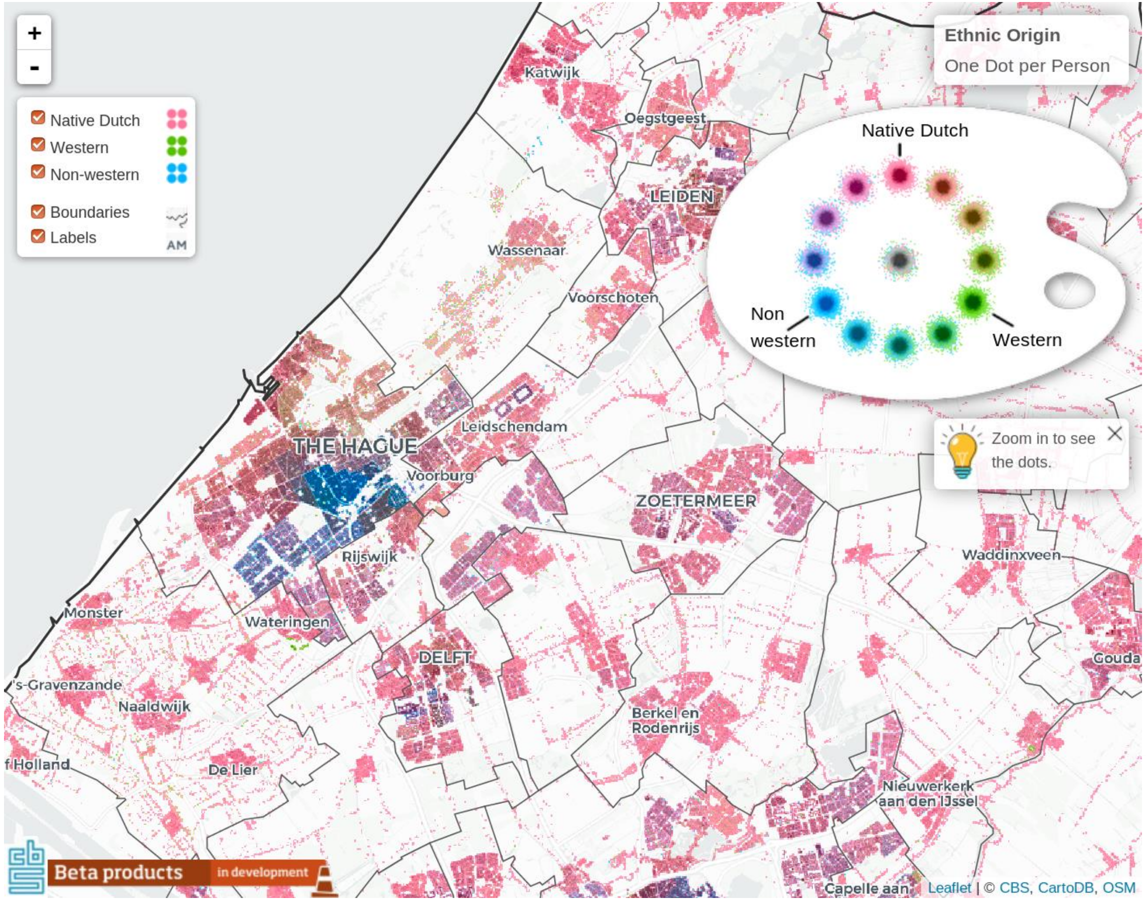


Application

Migration background of the Dutch population

Experimental version:

- Dots are placed in building areas (using the BAG register)
- “Artistic” legend

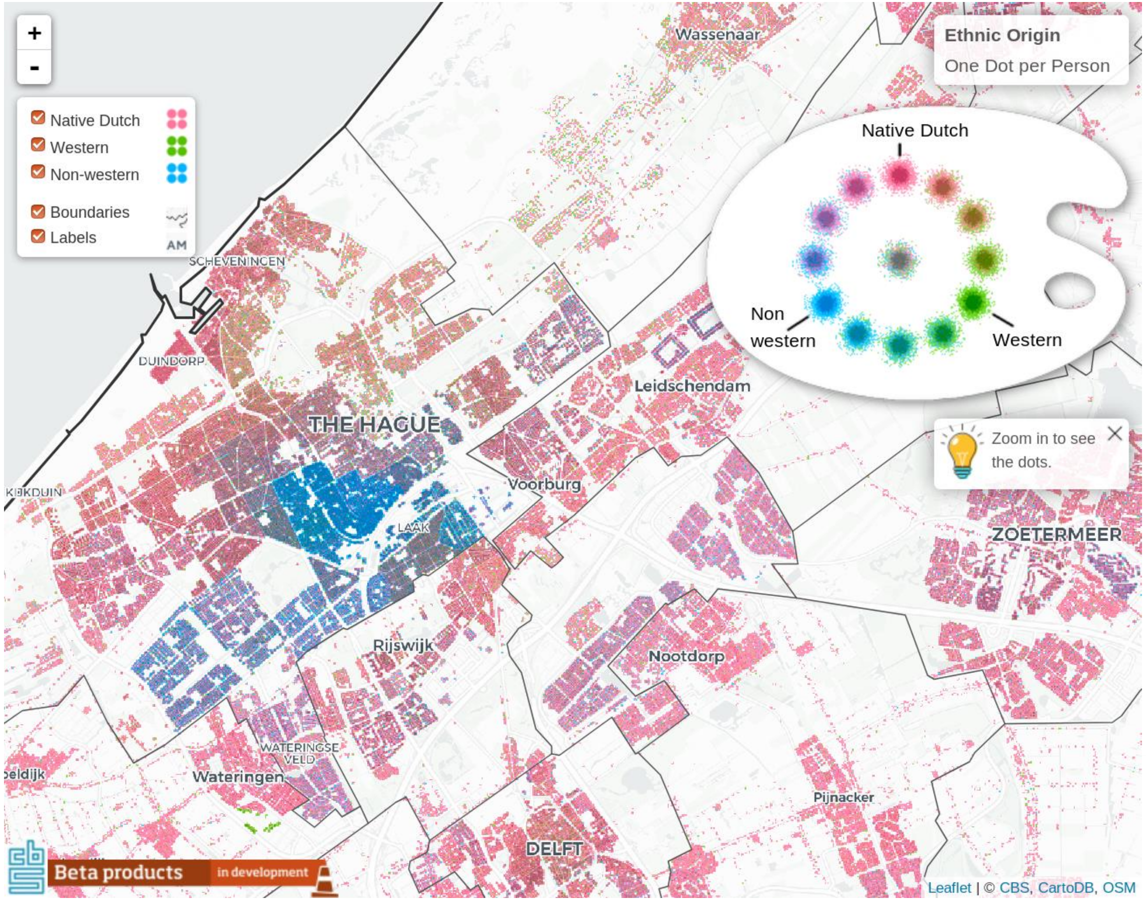


Application

Migration background of the Dutch population

Experimental version:

- Dots are placed in building areas (using the BAG register)
- “Artistic” legend

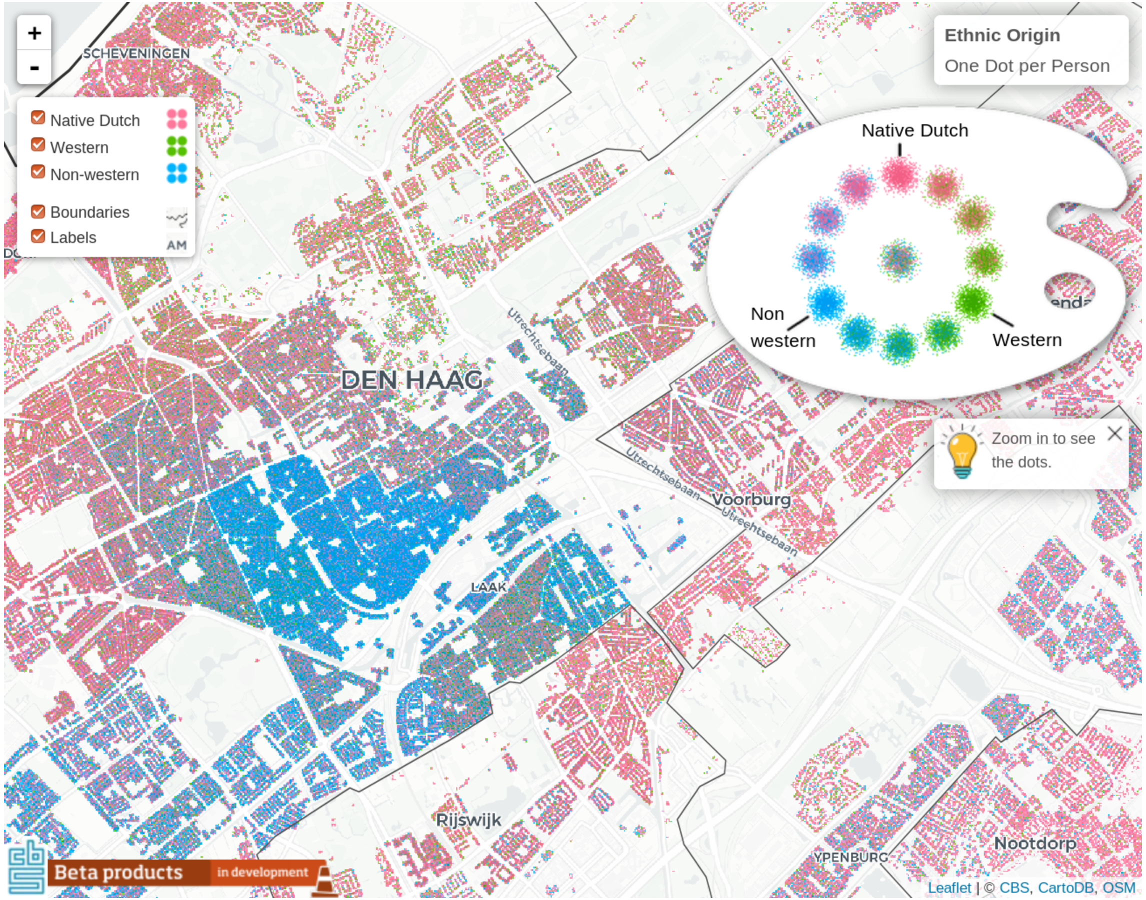


Application

Migration background of the Dutch population

Experimental version:

- Dots are placed in building areas (using the BAG register)
- “Artistic” legend

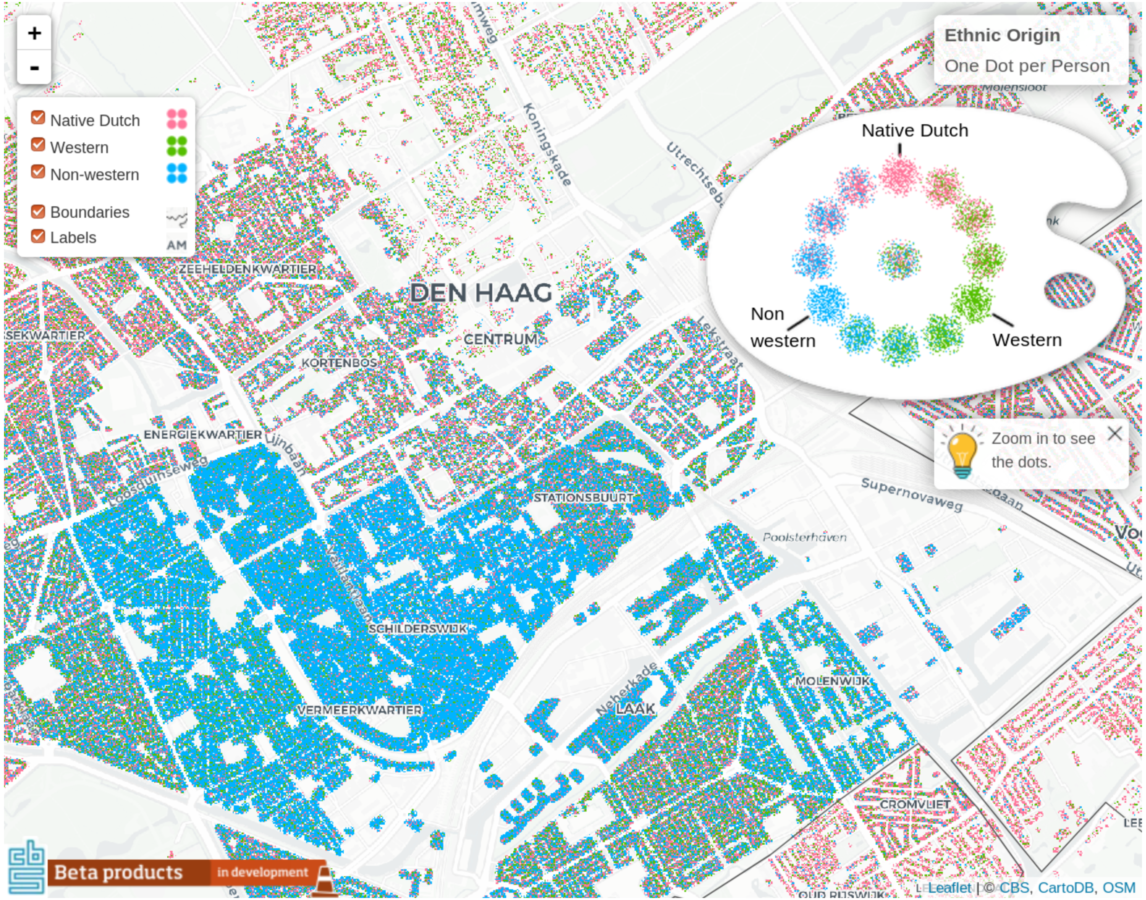


Application

Migration background of the Dutch population

Experimental version:

- Dots are placed in building areas (using the BAG register)
- “Artistic” legend

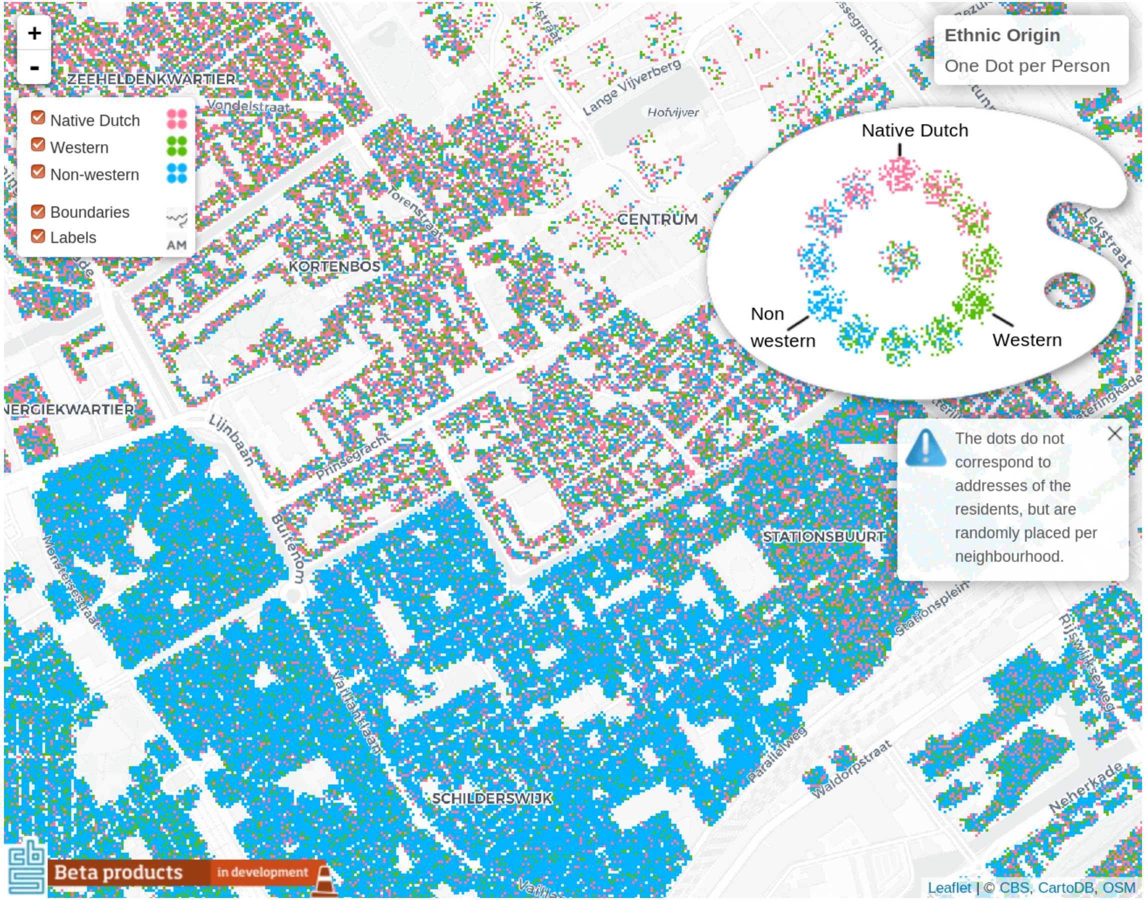


Application

Migration background of the Dutch population

Experimental version:

- Dots are placed in building areas (using the BAG register)
- "Artistic" legend



Application

Migration background of the Dutch population

Experimental version:

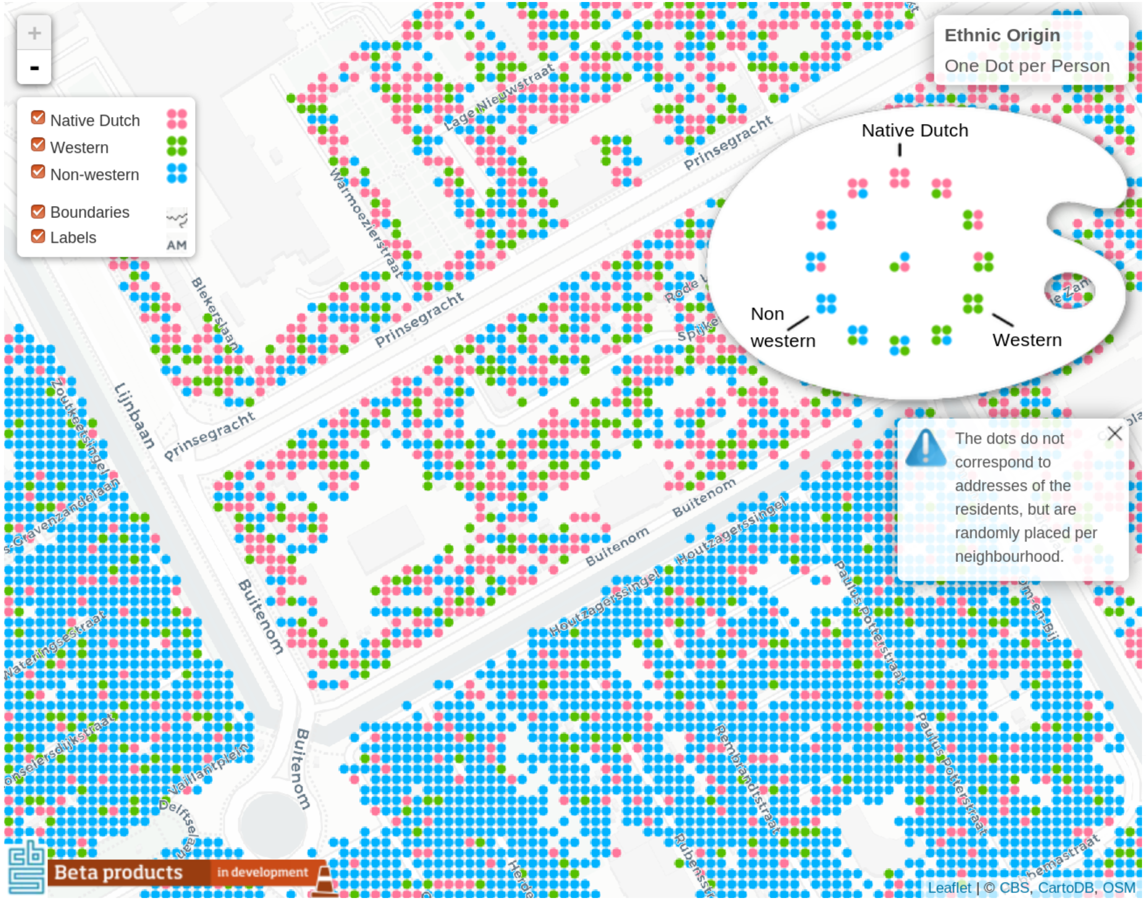
- Dots are placed in building areas (using the BAG register)
- “Artistic” legend

Application

Migration background of the Dutch population

Experimental version:

- Dots are placed in building areas (using the BAG register)
- “Artistic” legend



User study

Comparison between original and experimental version with eye-tracking.

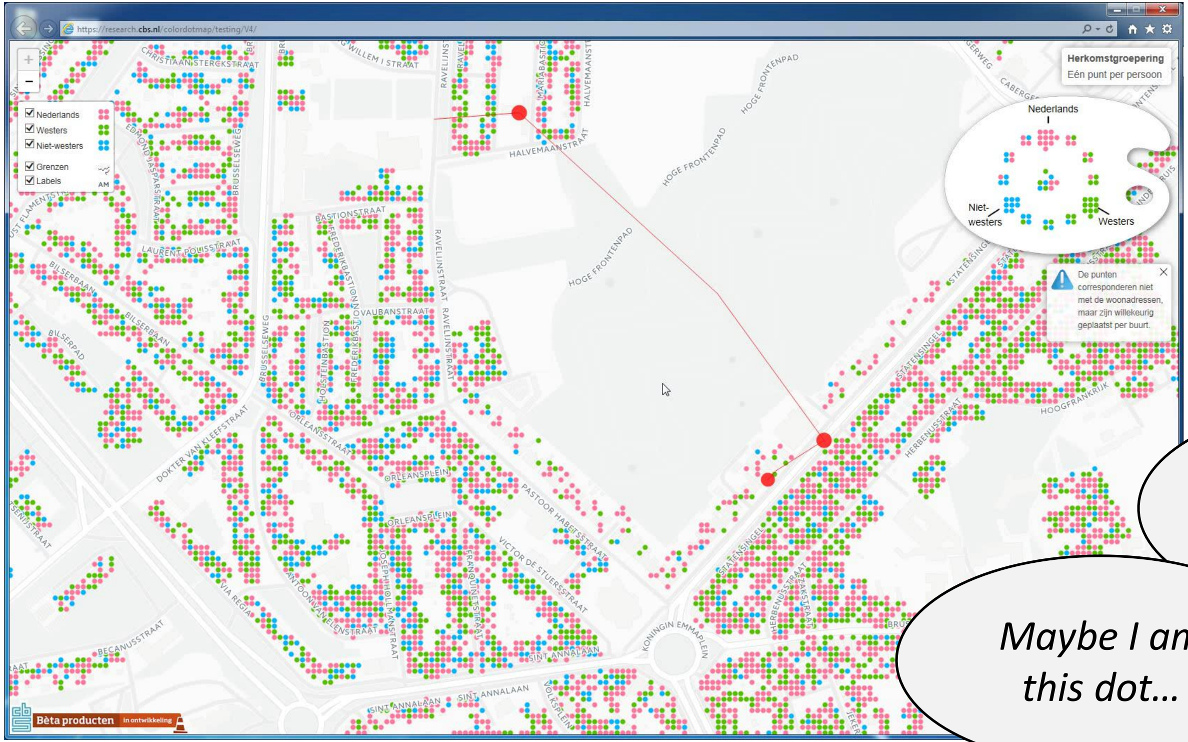
User study

Comparison between **original** and **experimental version** with **eye-tracking**.

Pink is dominant, and therefore it's hard to distinguish between green and blue.

User study

Comparison between **original** and **experimental version** with **eye-tracking**.



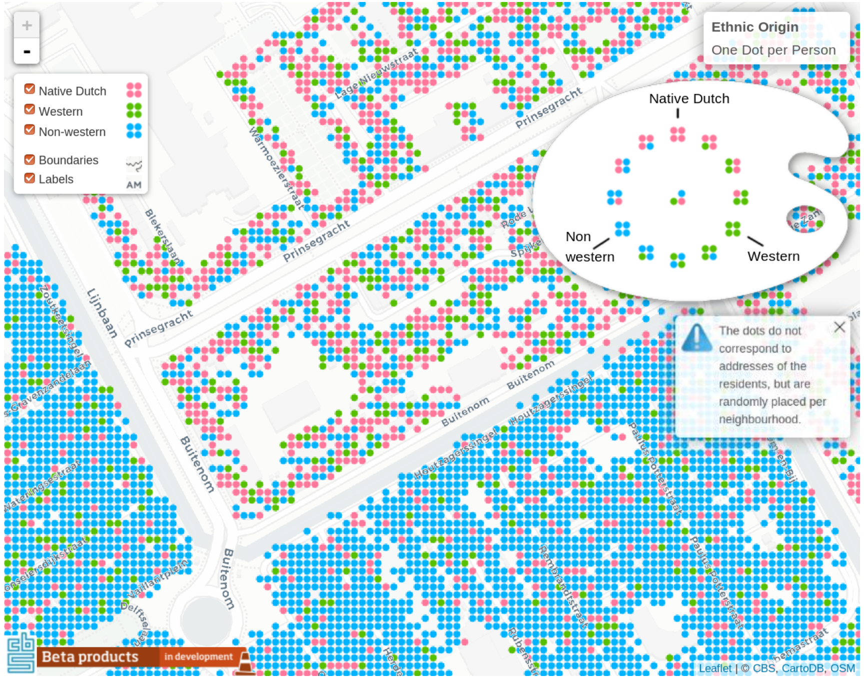
This property is indeed empty.

Maybe I am this dot…

User study

Conclusion:

- Discrepancy between nearby and distant views, although users were able to read and interpret composition and density correctly.
- Legend was difficult to interpret (both versions).
- Most users thought that the dots where placed on actual addresses.



How to deal with privacy?

Some ideas / guidelines:

- Areas should not be too detailed (global land use is better than detailed building areas)
- Draw neighbourhood borders
- Limit the zoom level (not to close)

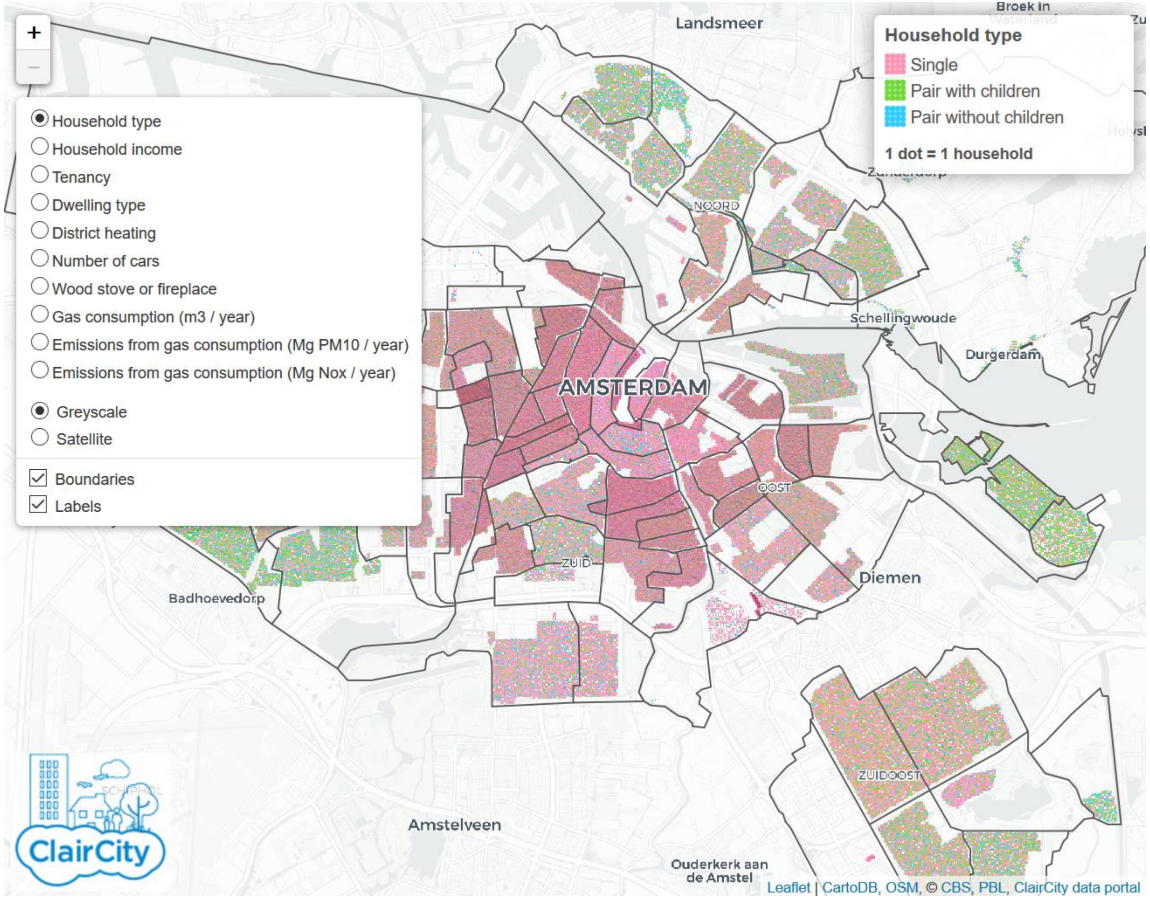
Application

- Simulated data on neighbourhood level for Amsterdam
- Each dot represents a household
- Dots are placed in residential areas (OpenStreetMap) per neighbourhood

http://www.claircity.eu/

Application

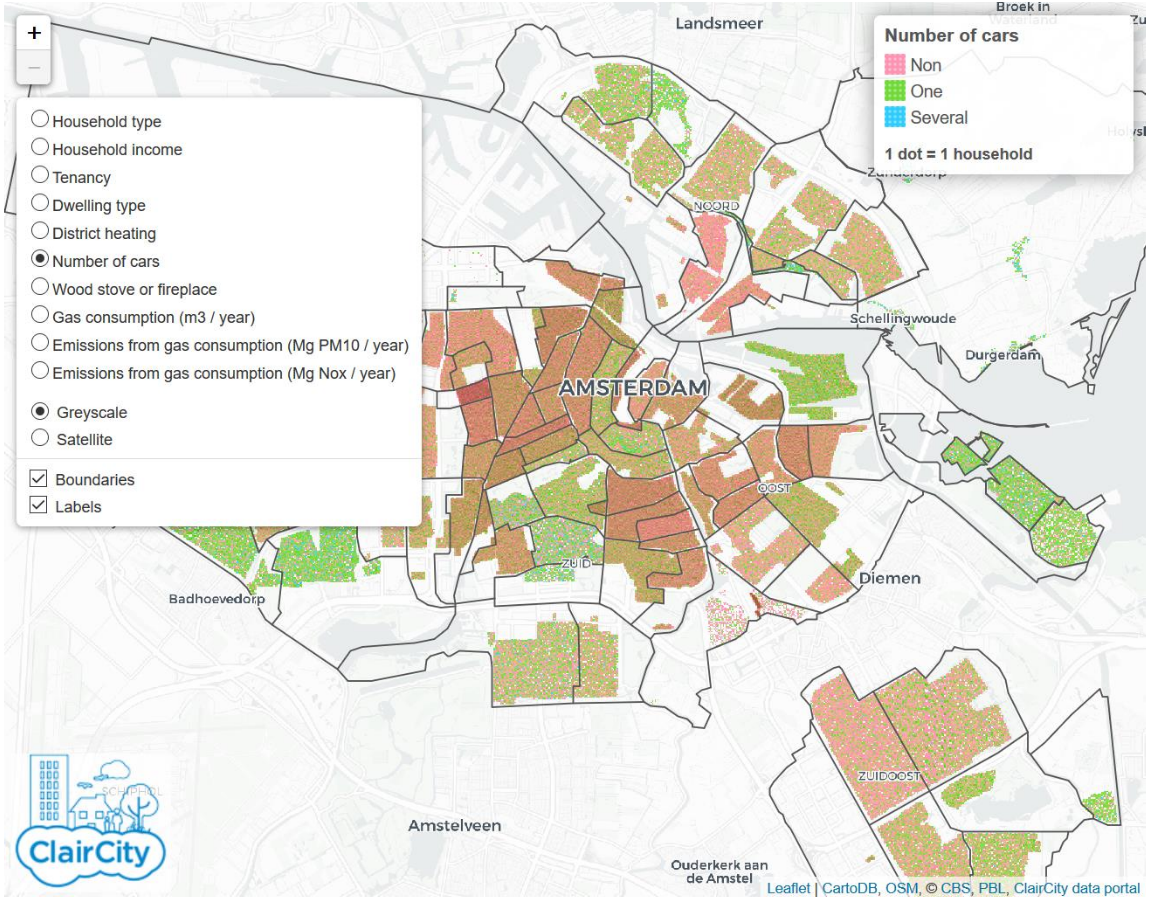
- Simulated data on neighbourhood level for Amsterdam
- Each dot represents a household
- Dots are placed in residential areas (OpenStreetMap) per neighbourhood



https://claircitydata.cbs.nl/pages/dotmaps

Application

- Simulated data on neighbourhood level for Amsterdam
- Each dot represents a household
- Dots are placed in residential areas (OpenStreetMap) per neighbourhood



https://claircitydata.cbs.nl/pages/dotmaps

Application

- Simulated data on neighbourhood level for Amsterdam
- Each dot represents a household
- Dots are placed in residential areas (OpenStreetMap) per neighbourhood

https://claircitydata.cbs.nl/pages/dotmaps

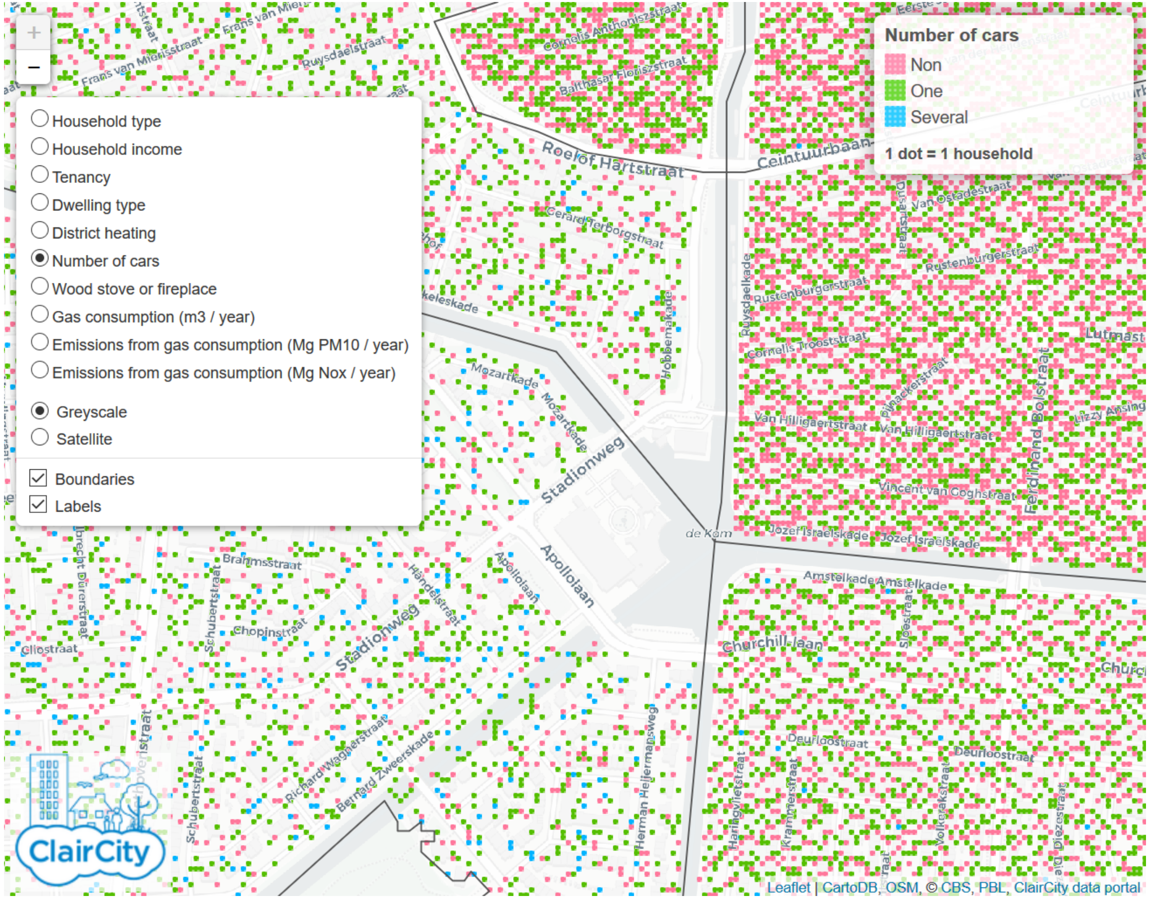
Application

- Simulated data on neighbourhood level for Amsterdam
- Each dot represents a household
- Dots are placed in residential areas (OpenStreetMap) per neighbourhood

https://claircitydata.cbs.nl/pages/dotmaps

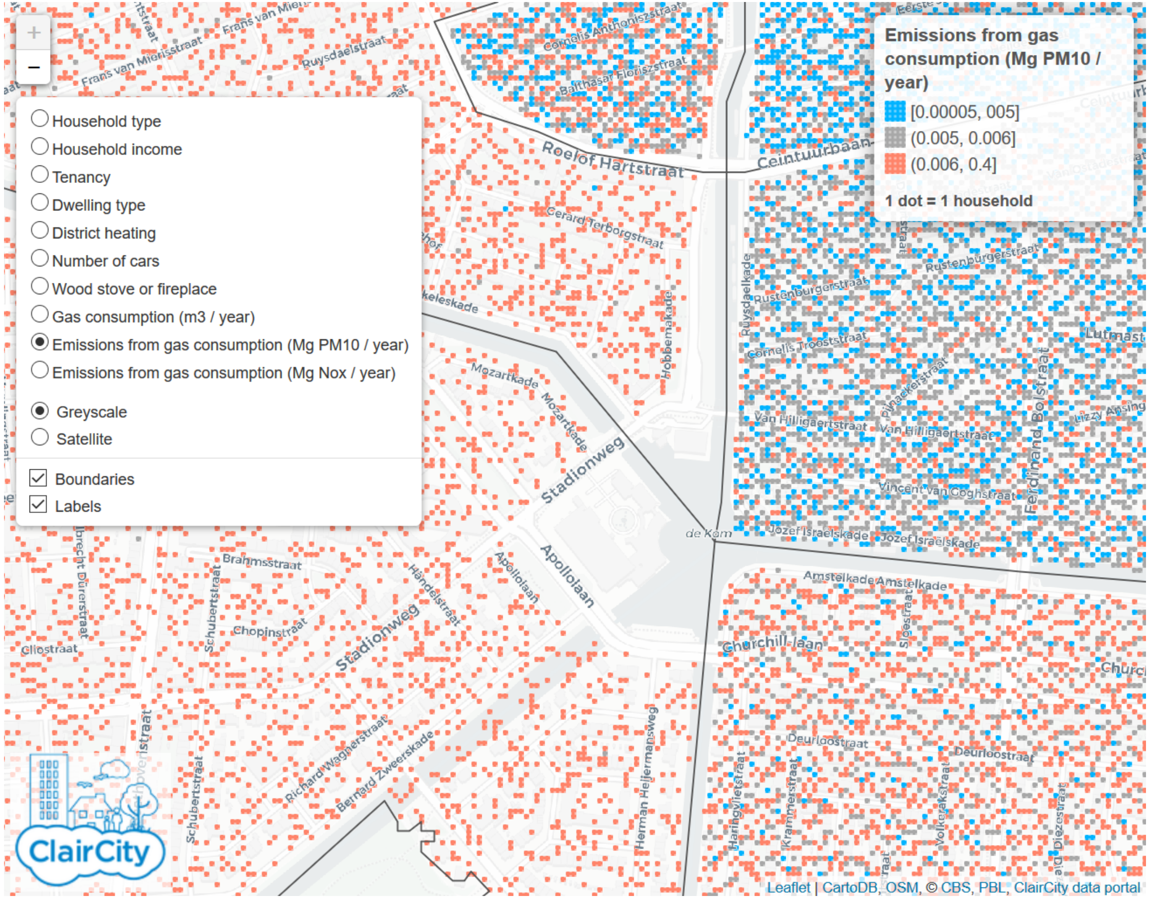
Application

- Simulated data on neighbourhood level for Amsterdam
- Each dot represents a household
- Dots are placed in residential areas (OpenStreetMap) per neighbourhood



https://claircitydata.cbs.nl/pages/dotmaps

Application

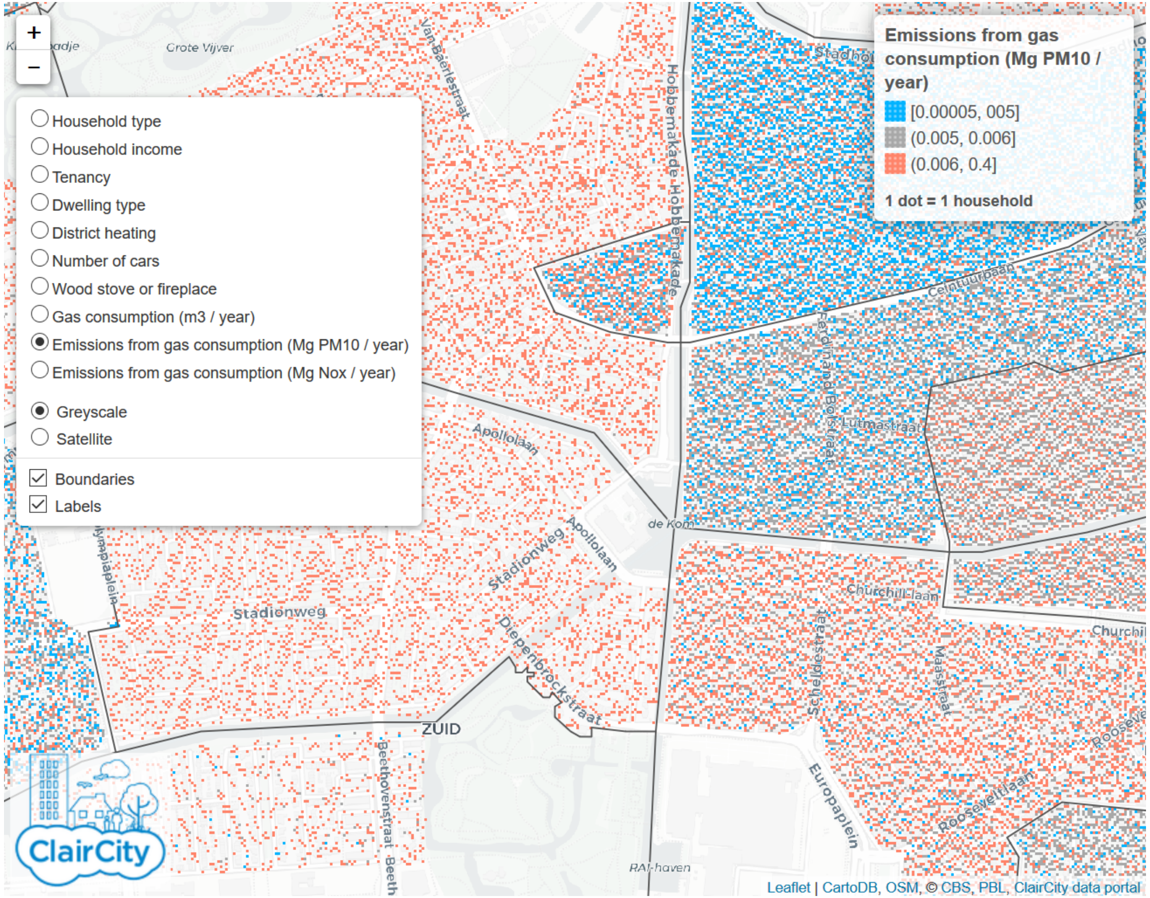


- Simulated data on neighbourhood level for Amsterdam
- Each dot represents a household
- Dots are placed in residential areas (OpenStreetMap) per neighbourhood

https://claircitydata.cbs.nl/pages/dotmaps

Application

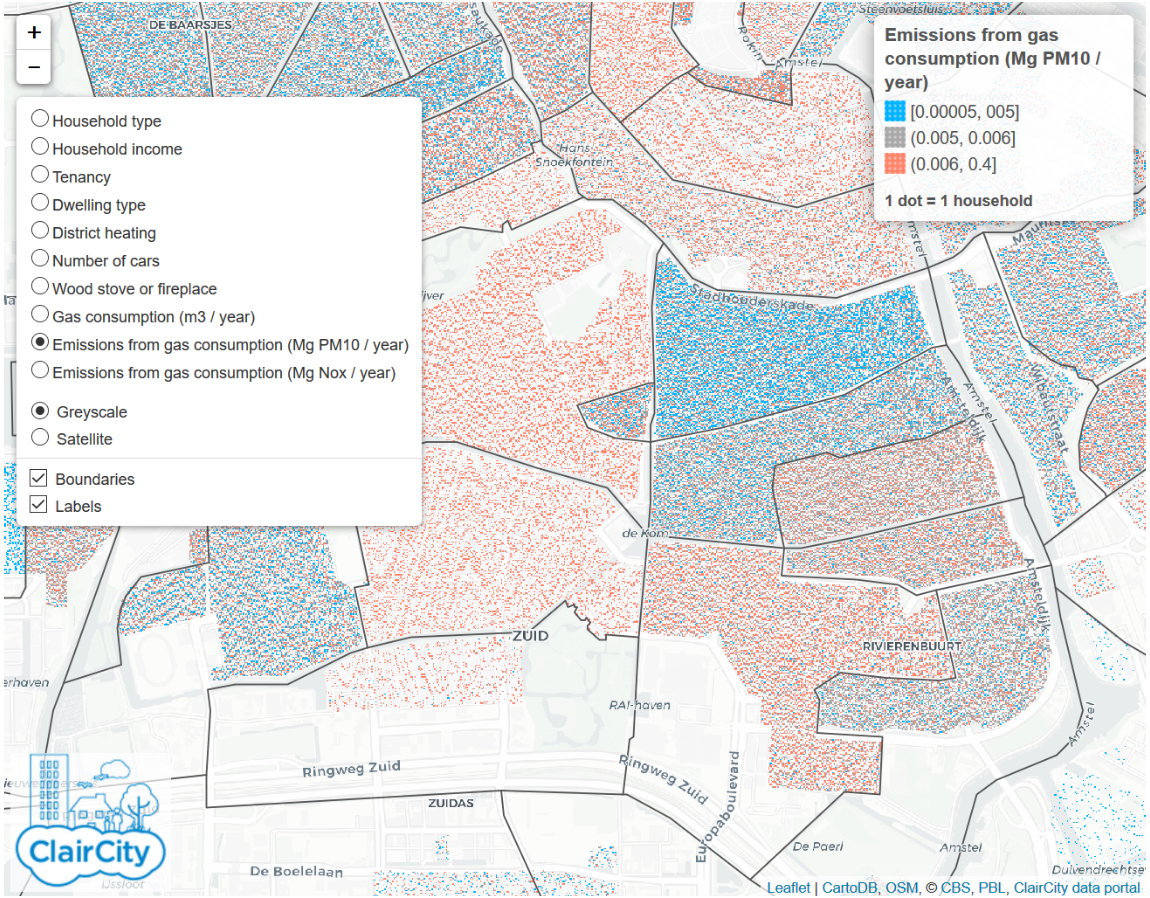
- Simulated data on neighbourhood level for Amsterdam
- Each dot represents a household
- Dots are placed in residential areas (OpenStreetMap) per neighbourhood



https://claircitydata.cbs.nl/pages/dotmaps

Application

- Simulated data on neighbourhood level for Amsterdam
- Each dot represents a household
- Dots are placed in residential areas (OpenStreetMap) per neighbourhood



https://claircitydata.cbs.nl/pages/dotmaps

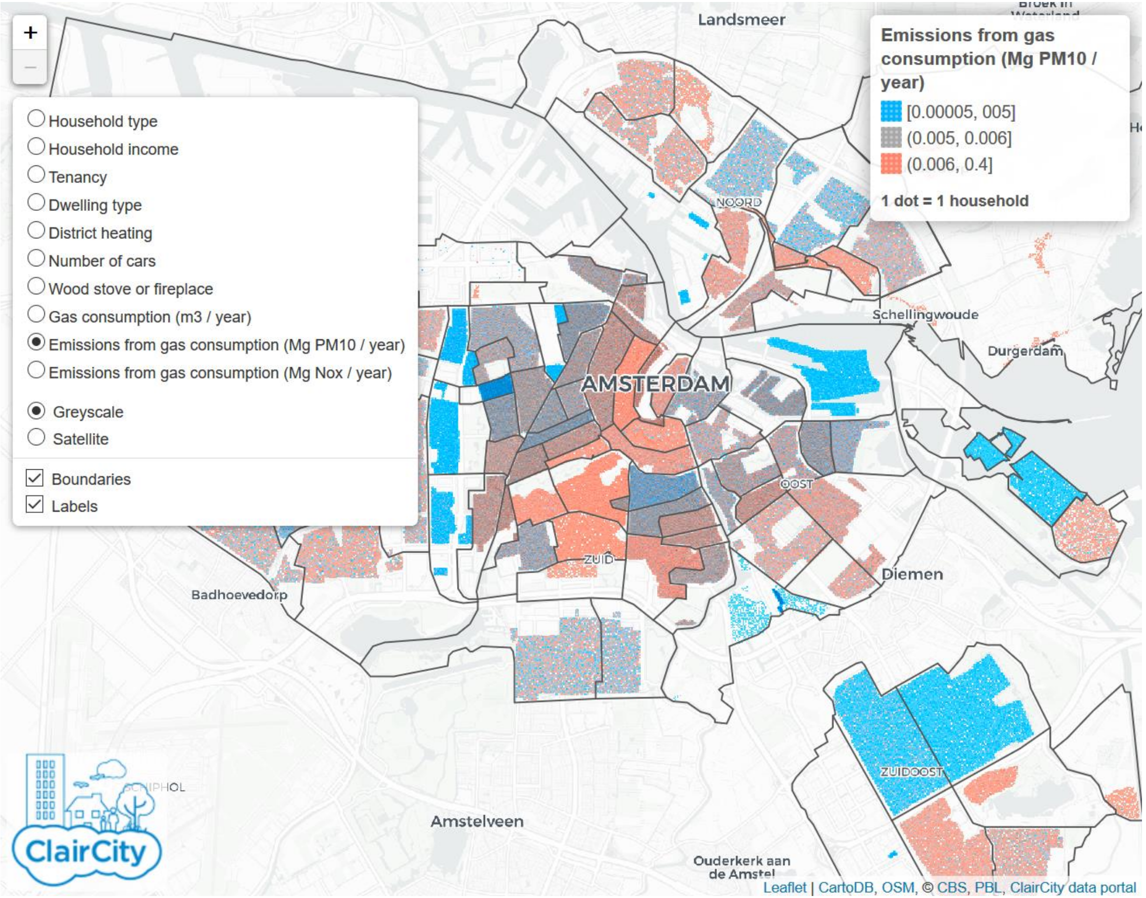
Application

- Simulated data on neighbourhood level for Amsterdam
- Each dot represents a household
- Dots are placed in residential areas (OpenStreetMap) per neighbourhood

https://claircitydata.cbs.nl/pages/dotmaps

Application

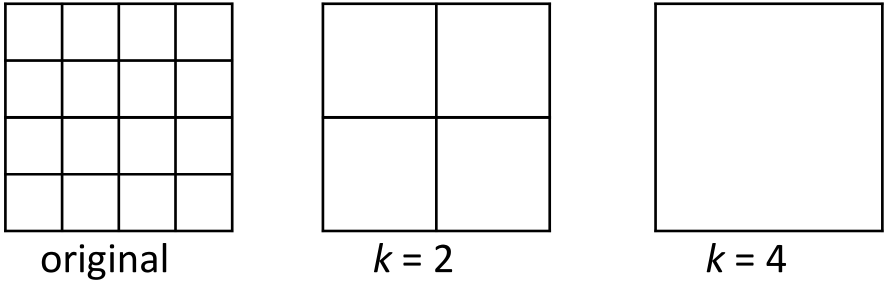
- Simulated data on neighbourhood level for 6 European cities (including Amsterdam and Bristol)
- Each dot represents a household
- Dots are placed in residential areas (OpenStreetMap) per neighbourhood



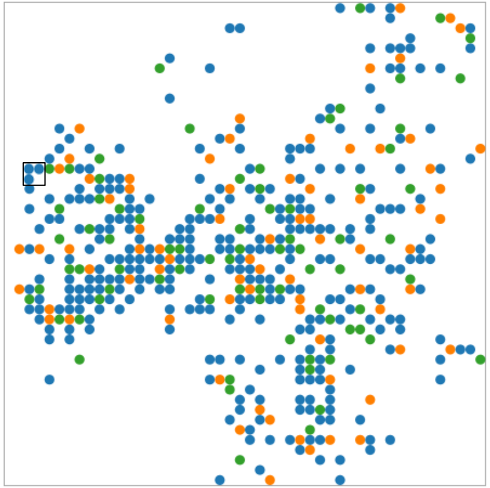
https://claircitydata.cbs.nl/pages/dotmaps

Super Dots

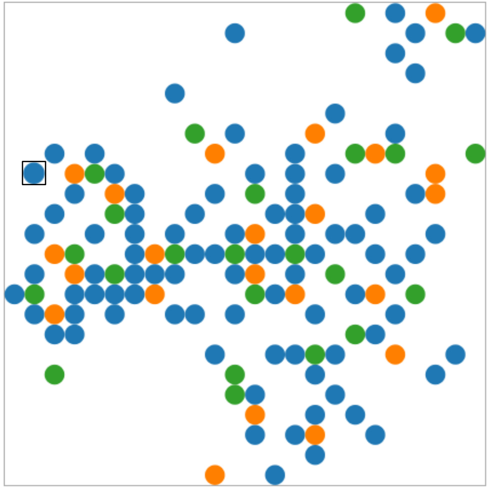
k by *k* grid cells in **original matrix** = 1 grid cell in **aggregated matrix**



Example:

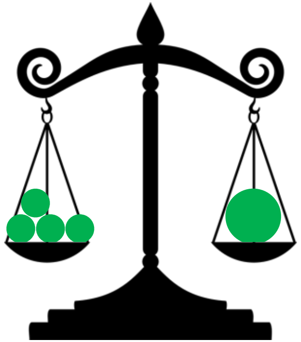


original dot map



aggregated dot map (k = 2)

What is a good aggregation?

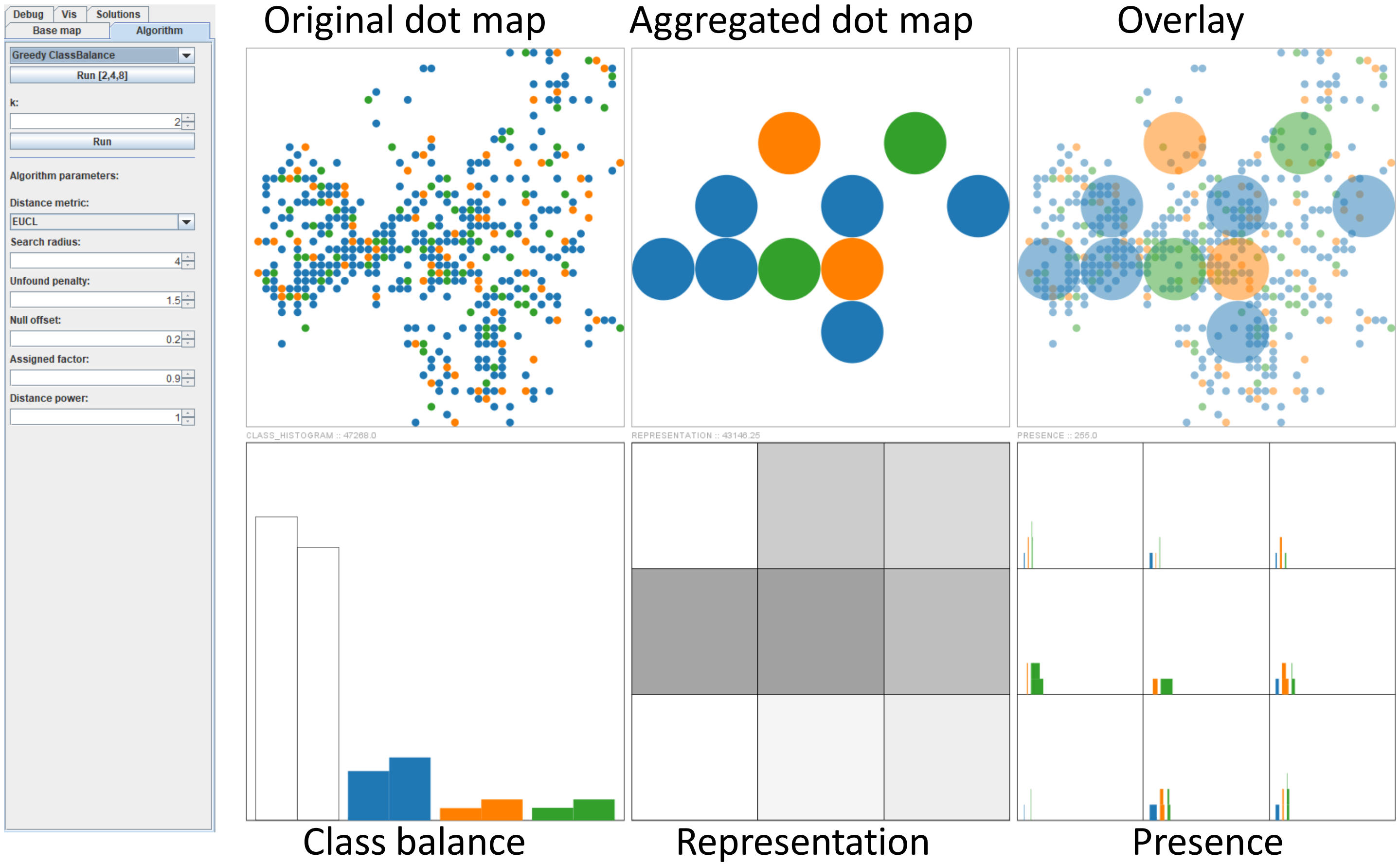


- **Class Balance** Total number of super dots per class should represent the total number of small dots per class

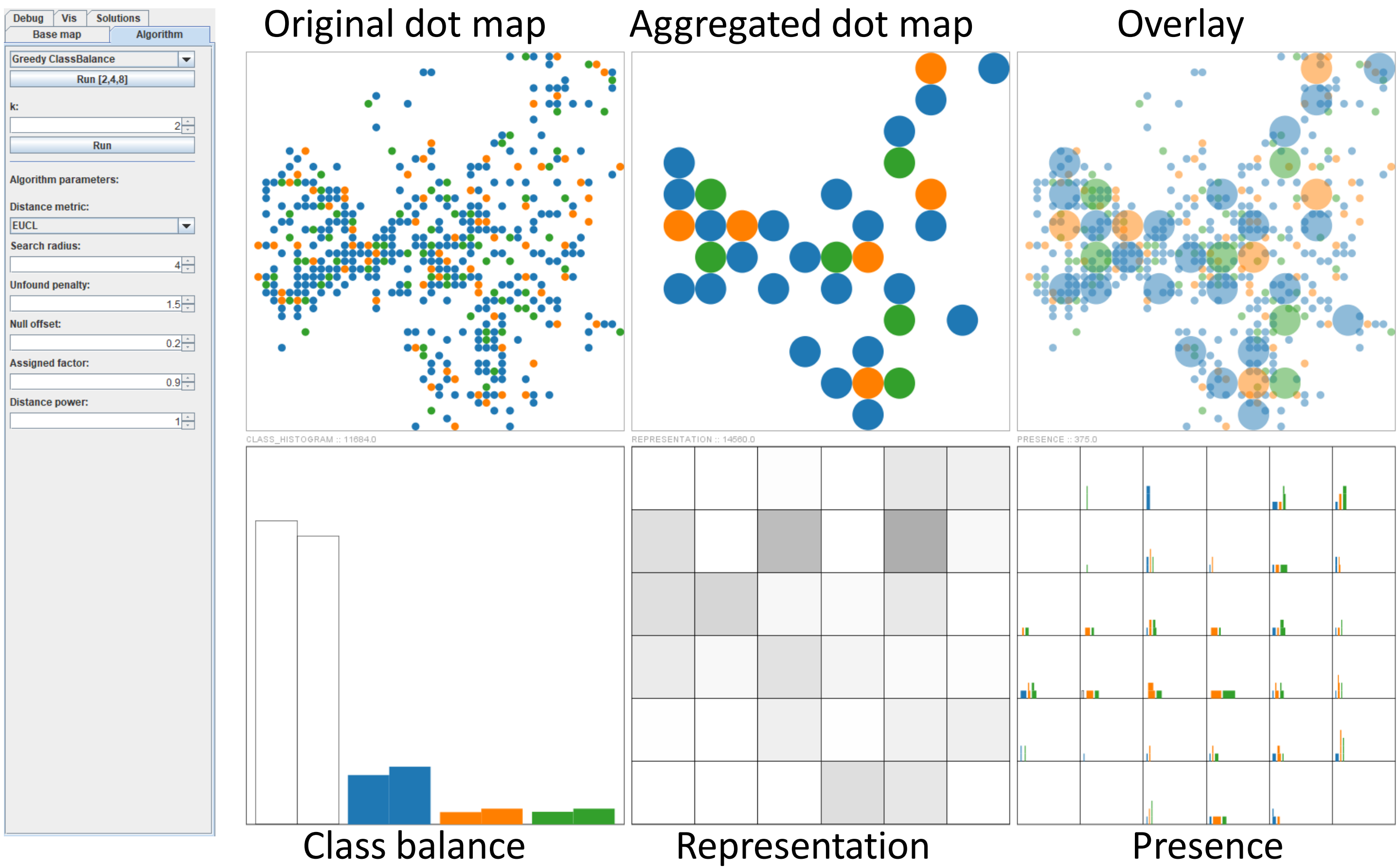
- **Representation** How well do the super dots represent the small dots? Each small dot is represented at most once, and each super dot can represent at most k^2 small dots.

- **Presence** How well are the small dots represented by the super dots? For each small dot, the distance to the nearest super dot is measured.

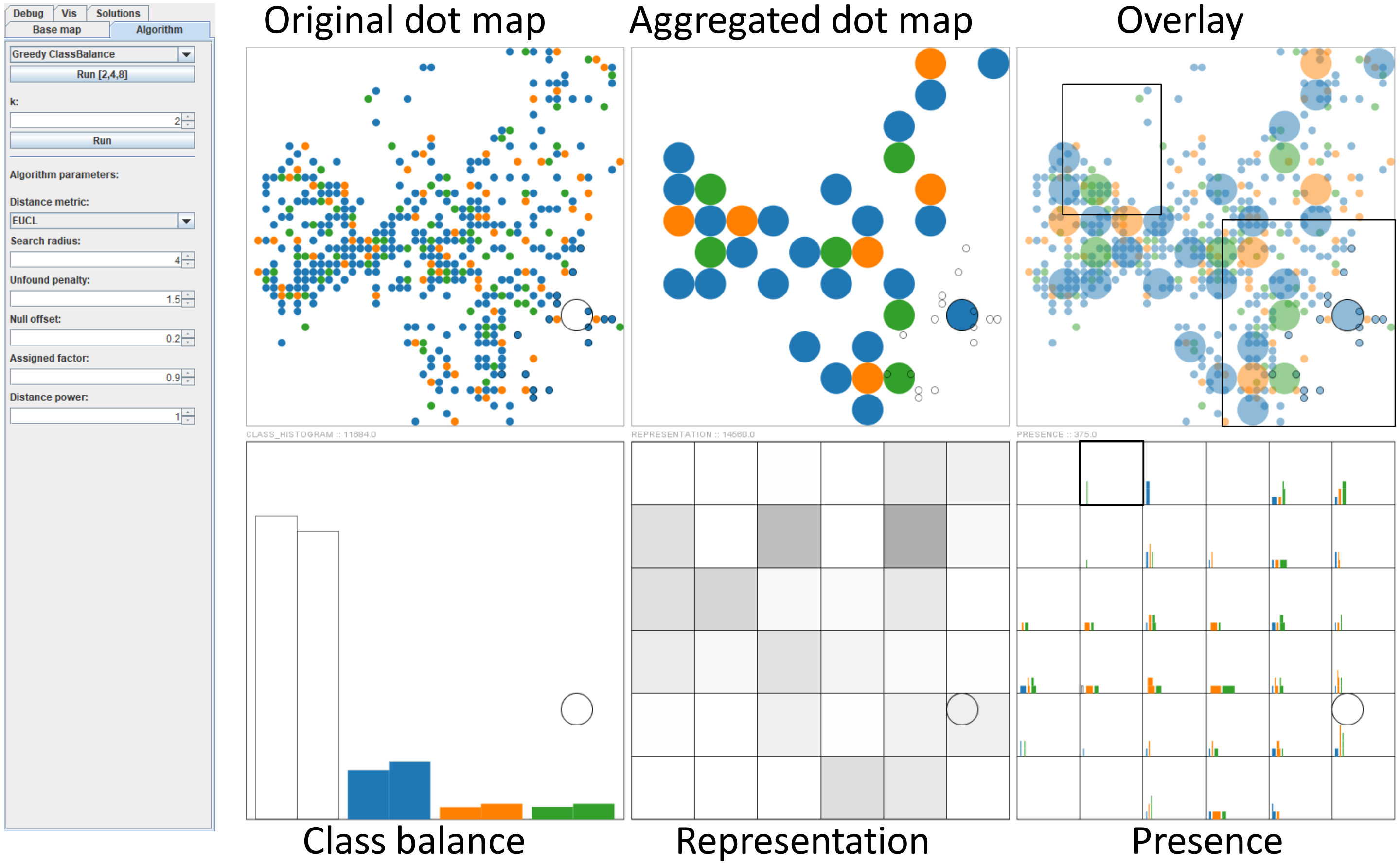
Aggregation analyses tool



Aggregation analyses tool



Aggregation analyses tool



Algorithms (sketches)

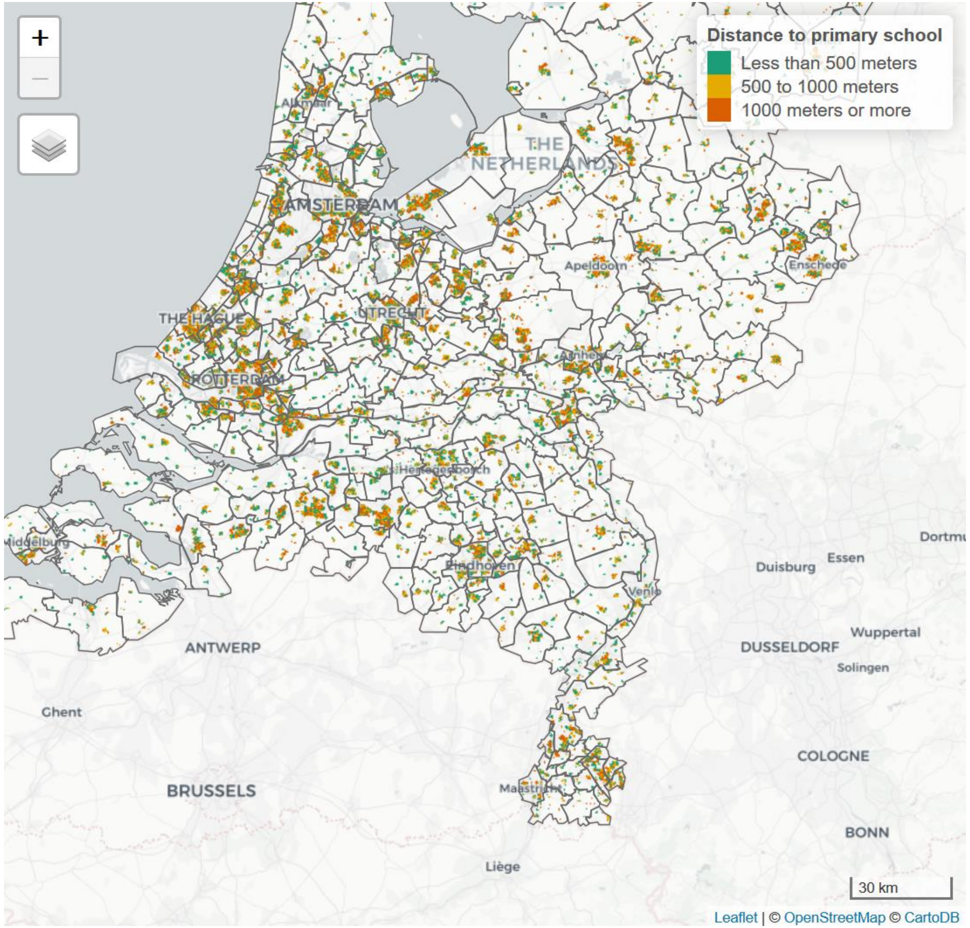
Greedy Class Balance Algorithm

1. Start with an empty map.
2. Pick the class with the largest imbalance and place a super dot of this class on the spot with the best representation.
3. Repeat step 2 until all super dots are placed.

Kernel Density Sampling Algorithm

1. For each class, estimate 2D kernel density.
2. Place super dots where total density is above a certain threshold.
3. Per super dot, sample its class using the density values as probabilities.

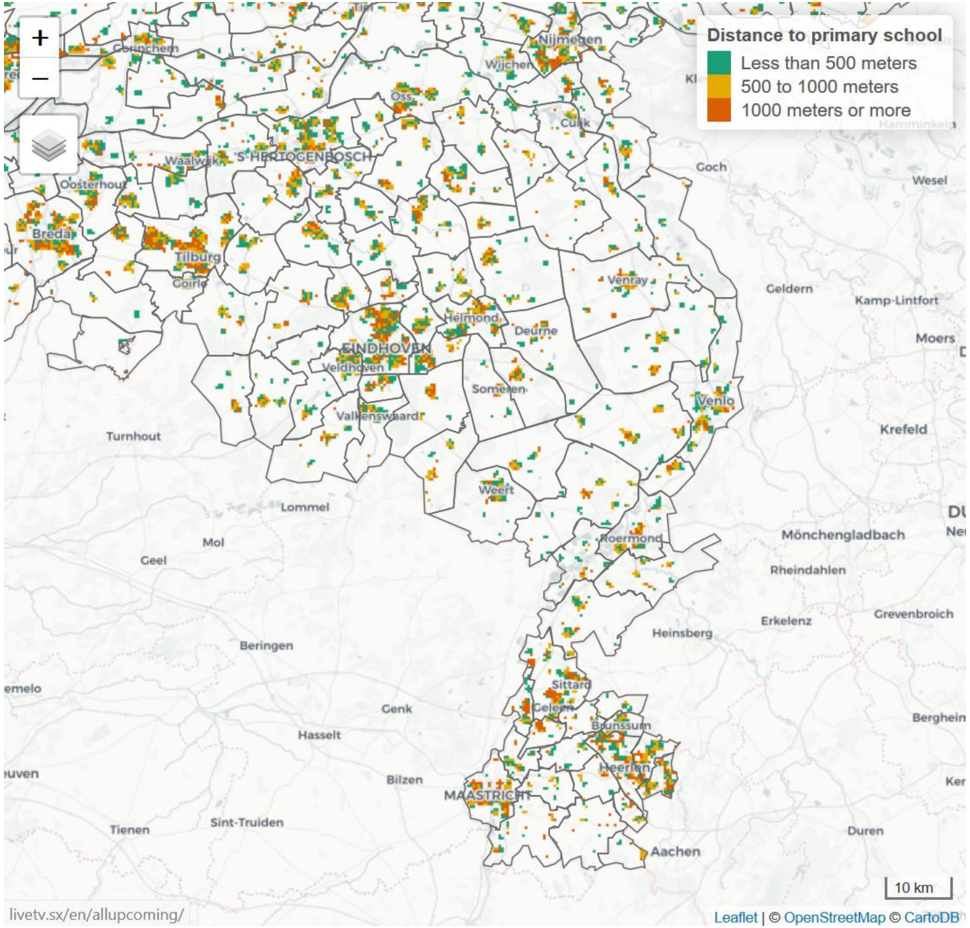
Application



Distance to school

- Dots represent children who go to primary schools
- Colour indicates distance to their primary school (not necessarily the nearest one)
- Used data: education registers
- Draft version (not published yet)
- Dots aggregated using the Kernel Density Sampling Algorithm (only one aggregation)

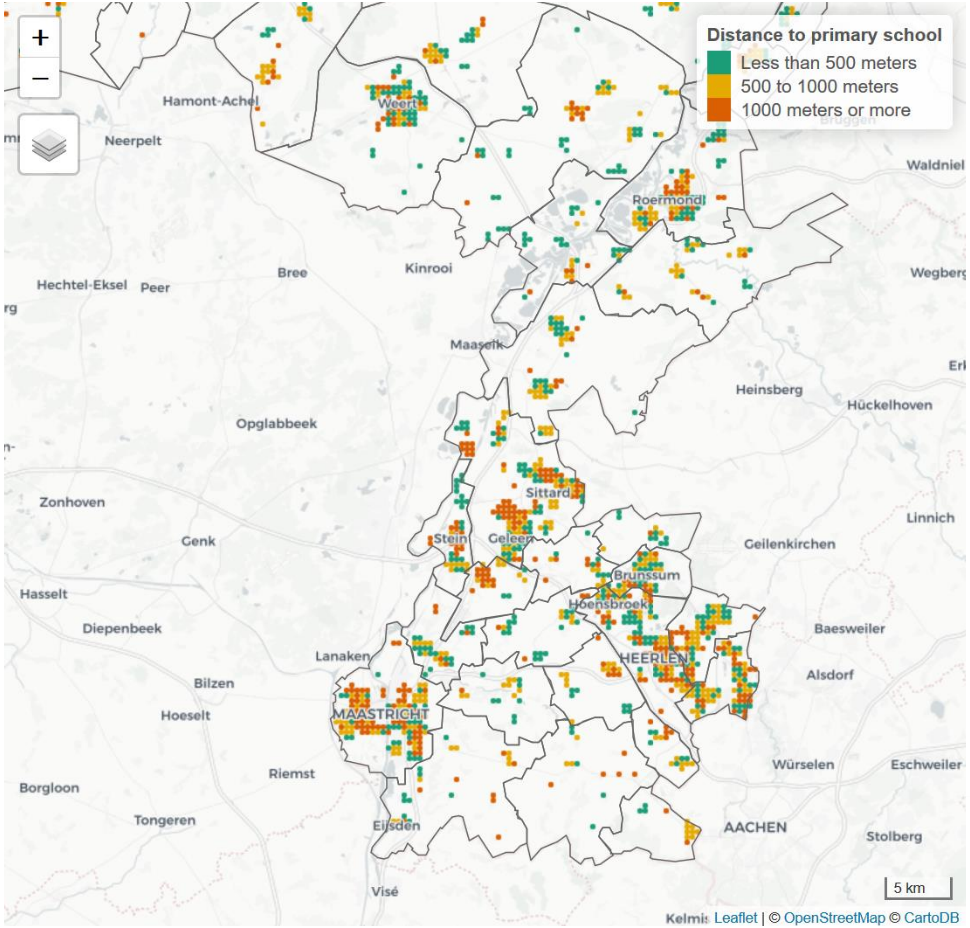
Application



Distance to school

- Dots represent children who go to primary schools
- Colour indicates distance to their primary school (not necessarily the nearest one)
- Used data: education registers
- Draft version (not published yet)
- Dots aggregated using the Kernel Density Sampling Algorithm (only one aggregation)

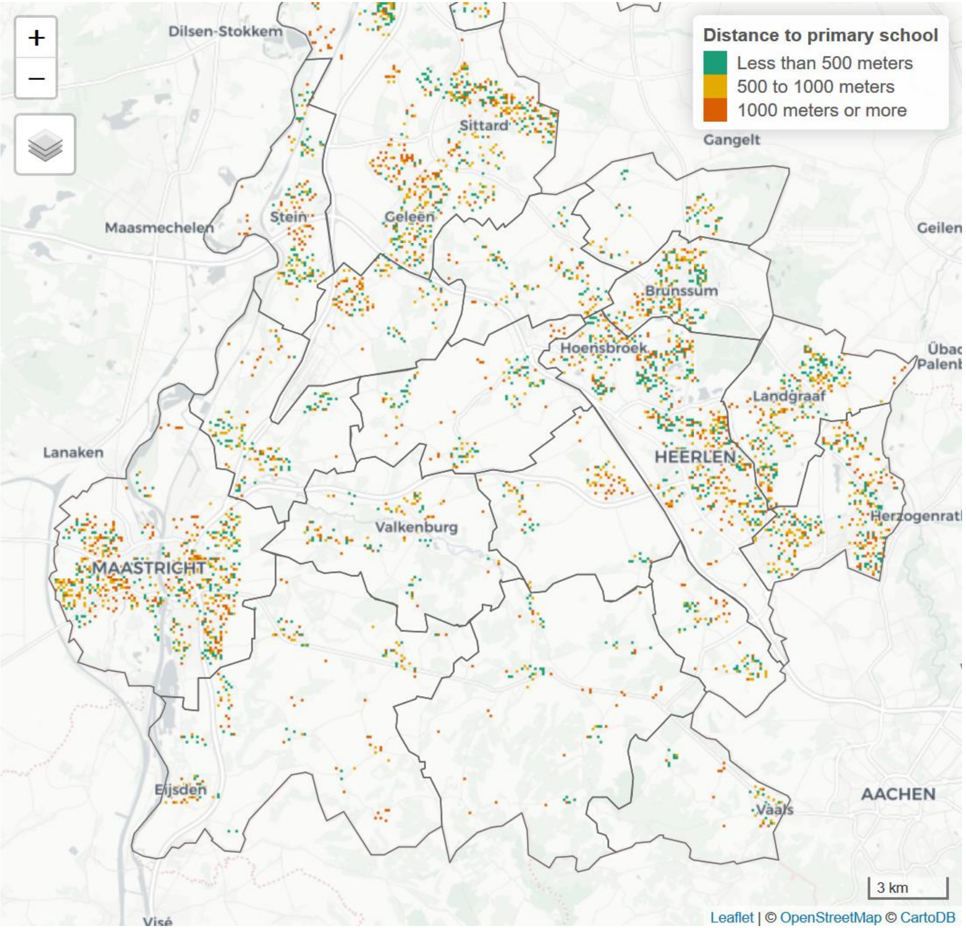
Application



Distance to school

- Dots represent children who go to primary schools
- Colour indicates distance to their primary school (not necessarily the nearest one)
- Used data: education registers
- Draft version (not published yet)
- Dots aggregated using the Kernel Density Sampling Algorithm (only one aggregation)

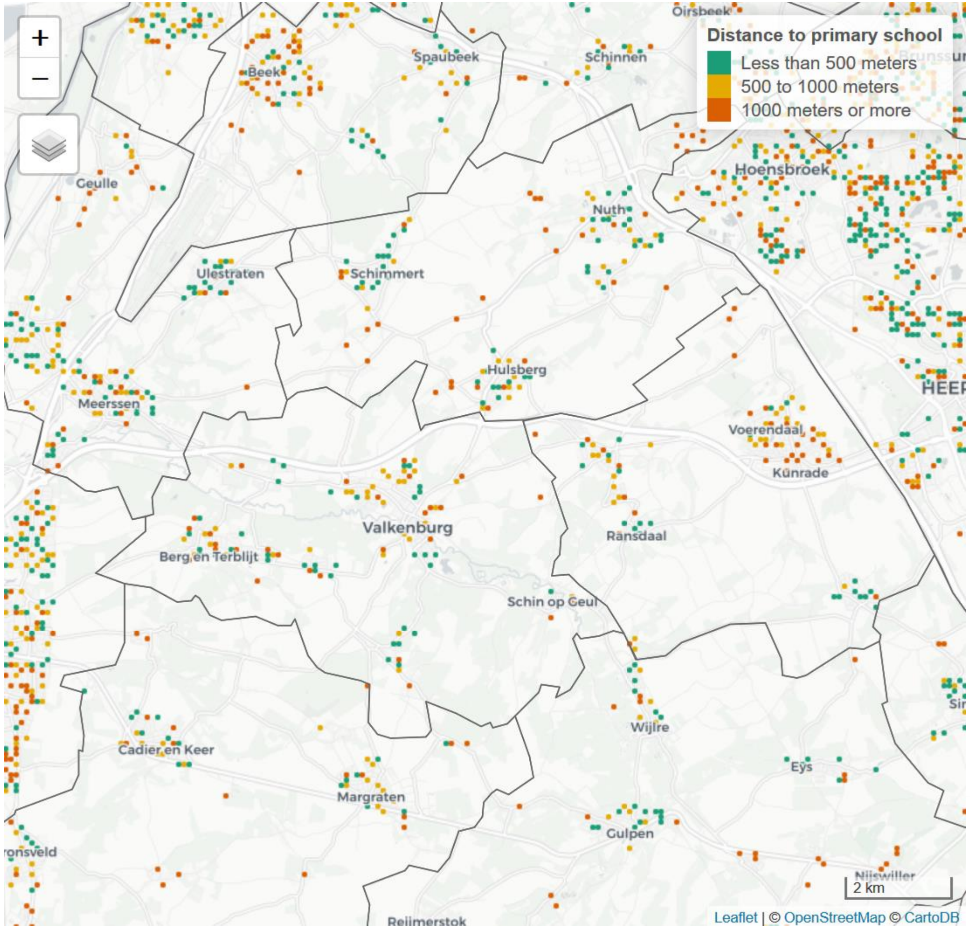
Application



Distance to school

- Dots represent children who go to primary schools
- Colour indicates distance to their primary school (not necessarily the nearest one)
- Used data: education registers
- Draft version (not published yet)
- Dots aggregated using the Kernel Density Sampling Algorithm (only one aggregation)

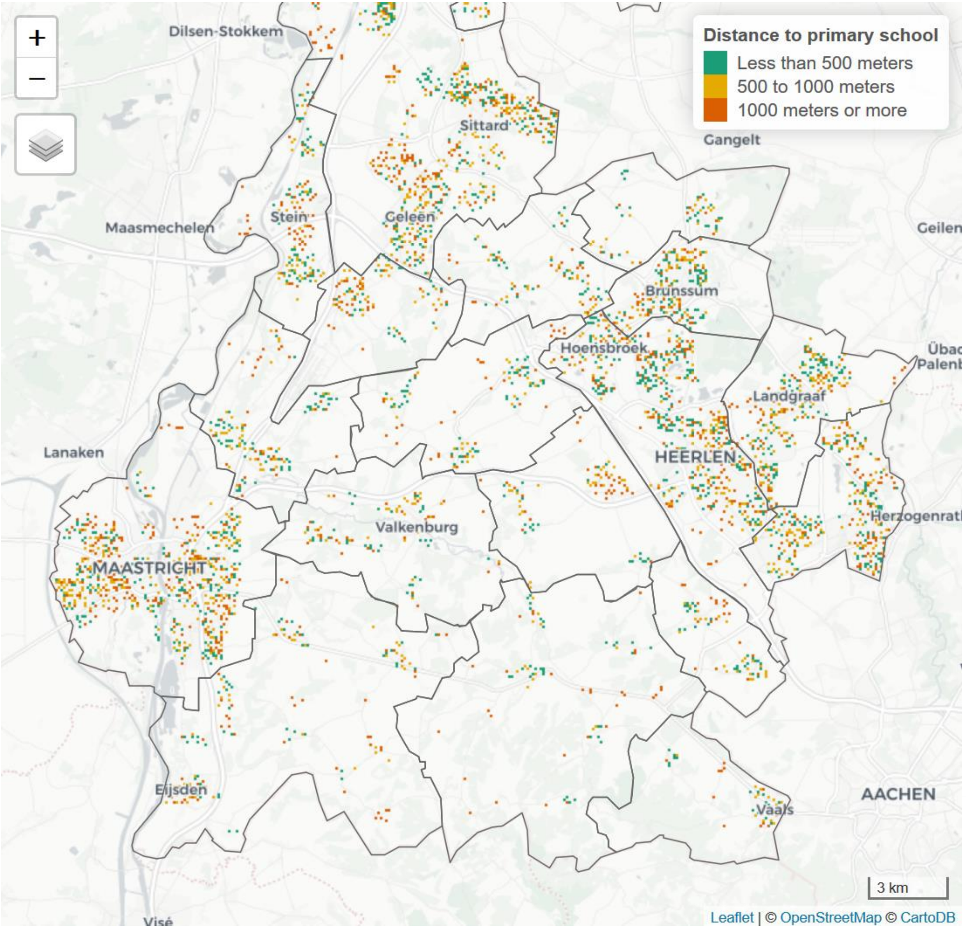
Application



Distance to school

- Dots represent children who go to primary schools
- Colour indicates distance to their primary school (not necessarily the nearest one)
- Used data: education registers
- Draft version (not published yet)
- Dots aggregated using the Kernel Density Sampling Algorithm (only one aggregation)

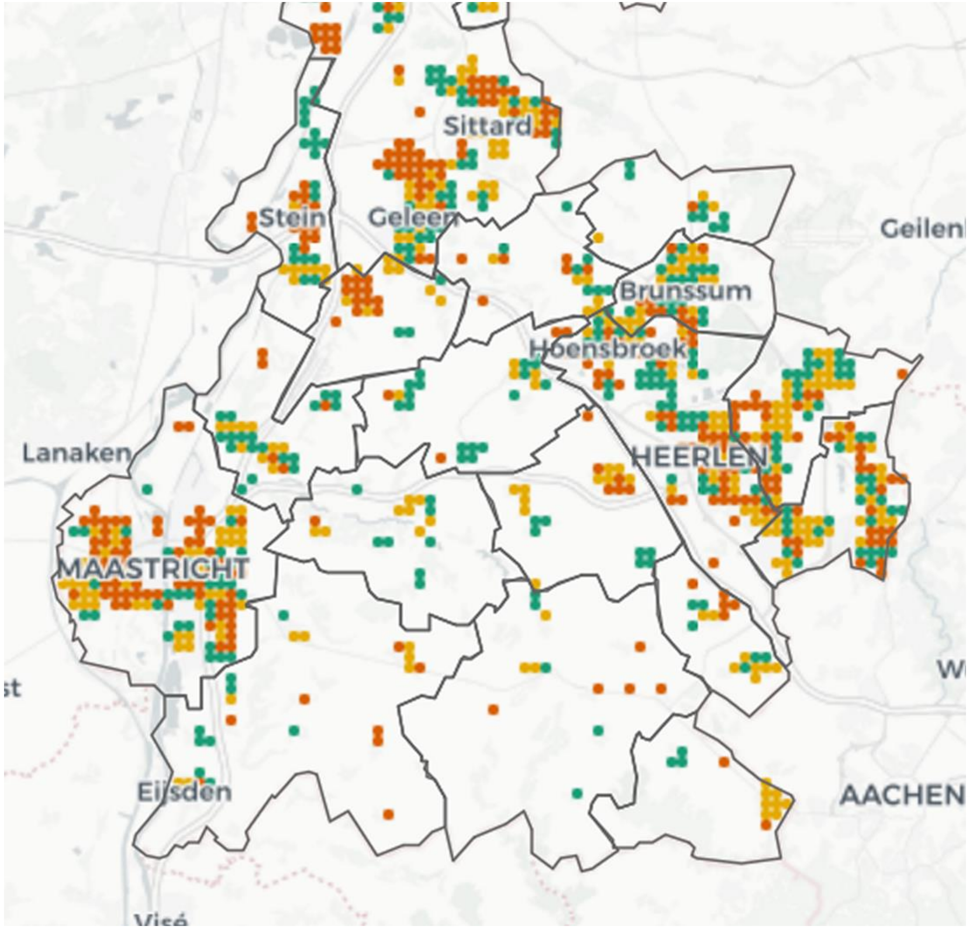
Application



Distance to school

- Dots represent children who go to primary schools
- Colour indicates distance to their primary school (not necessarily the nearest one)
- Used data: education registers
- Draft version (not published yet)
- Dots aggregated using the Kernel Density Sampling Algorithm (only one aggregation)

Application



Distance to school

- Dots represent children who go to primary schools
- Colour indicates distance to their primary school (not necessarily the nearest one)
- Used data: education registers
- Draft version (not published yet)
- Dots aggregated using the Kernel Density Sampling Algorithm (only one aggregation)

Comparison

Blended colours

\+ Sense of immensity of the data

– Dots hard to distinguish and categorize

– Difficult to create simple legend

– Tricky to pick suitable colours (visual perception is complex)

Super dots

\+ Simple and clear representation

\+ Keeps the overall distribution and composition

– Loss of local detail

Software implementation

Super dots analysis tool

- Java application (available upon request)

Creating tiles

- Tiles are 512x512 sized png images (also used by Google Maps, Bing Maps, OSM)
- R package **dotmap**
  - In development: https://github.com/mtennekes/dotmap
  - Both methods (blended colours and super dots) are implemented
  - Working, but no documentation yet

Visualization

- R package **tmap** or Javascript library **leaflet**
- Dynamic legend: Javascript

Acknowledgements

Dot maps:
- Edwin de Jonge and Chantal Melser (CBS)
- Wouter Meulemans (TU Eindhoven)
- François Engelen (Hogeschool Zuyd)

Mobile phone data:
- Yvonne Gootzen, Shan Shah, May Offermans, Marco Puts, Sander Scholtus, Harm Jan Boonstra, Ralph Meijers, Sigrid van Hoek, Edwin de Jonge, Jan van der Laan, and Marc Ponsen (CBS)
- Fabio Ricciato, David Salgado, Benjamin Sakarovitch, Roberta Radini, Tiziana Tuoto, and Sandra Hadam (ESSnet Big Data)
- Arne Jol, Guido Diepen, Arjan Schoe, and Rob Hormes (T-Mobile NL)