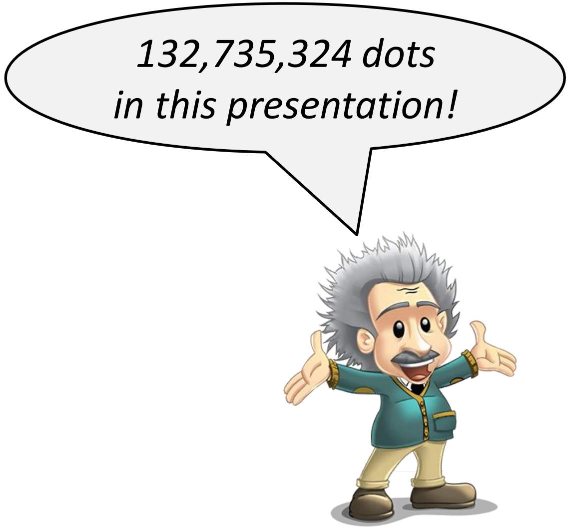
Aggregation of Dots

Methods for Big Data in Official Statistics

Martijn Tennekes

Heerlen, October 5, 2018

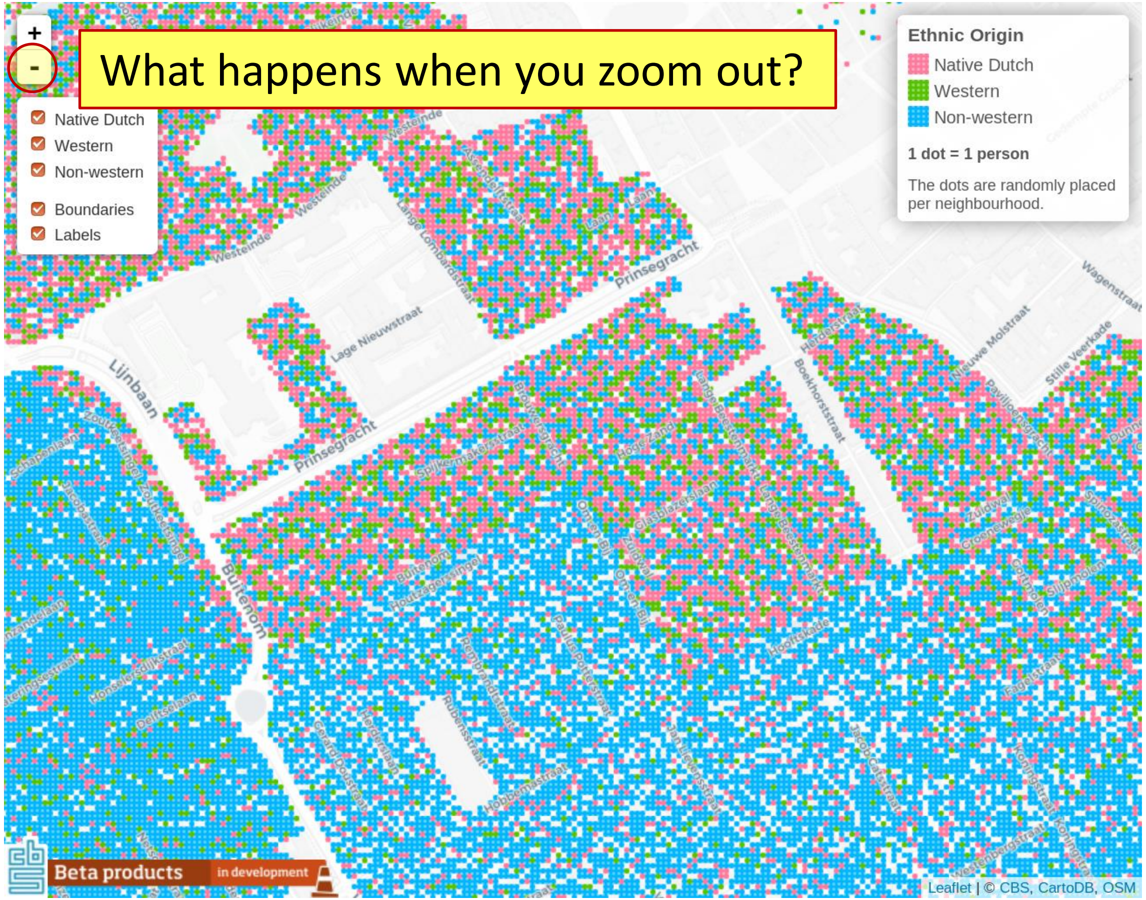


Classic dot map

Cholara outbreak in London (1854) by John Snow

Dots instead of bars

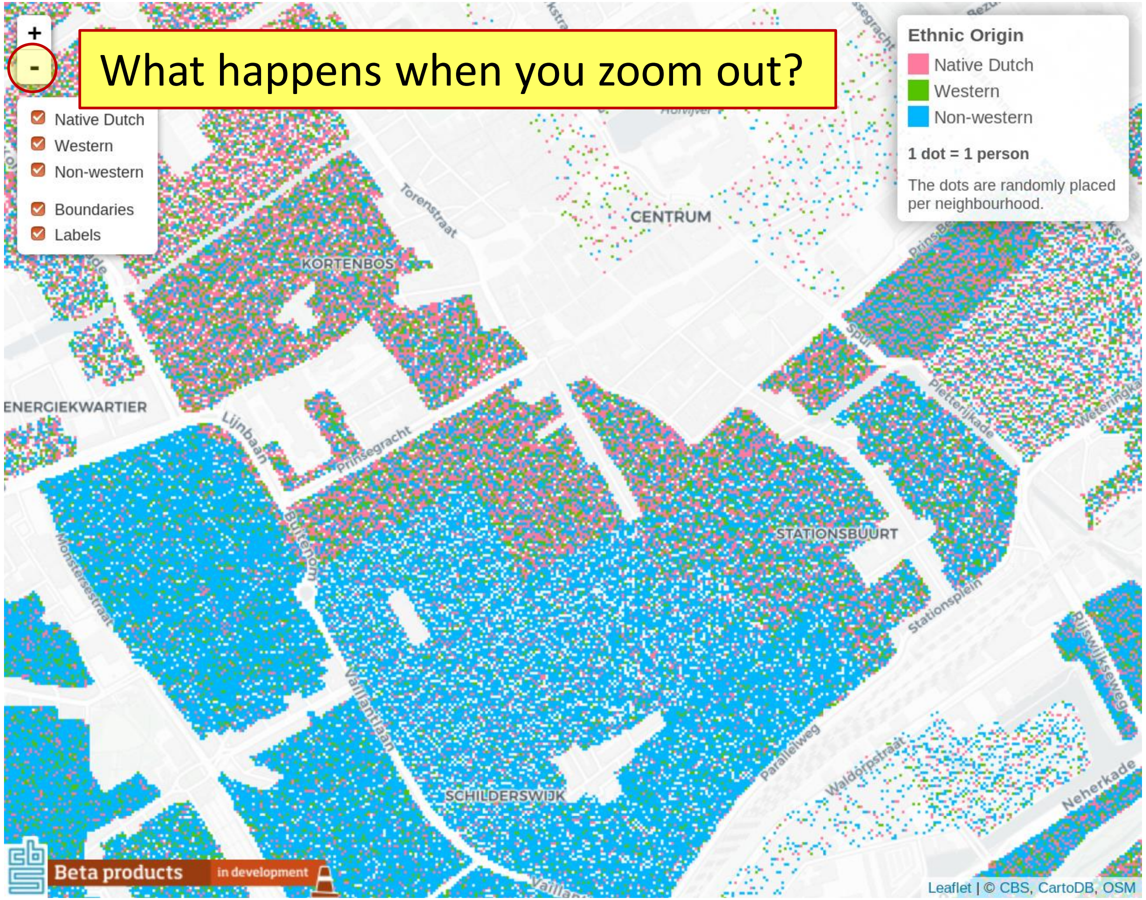
Let there be… COLOR



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

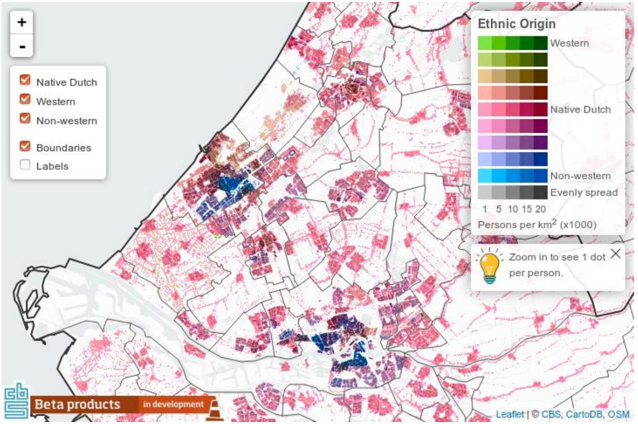
Out of pixels ☹

Hmm, still can't see the dots...

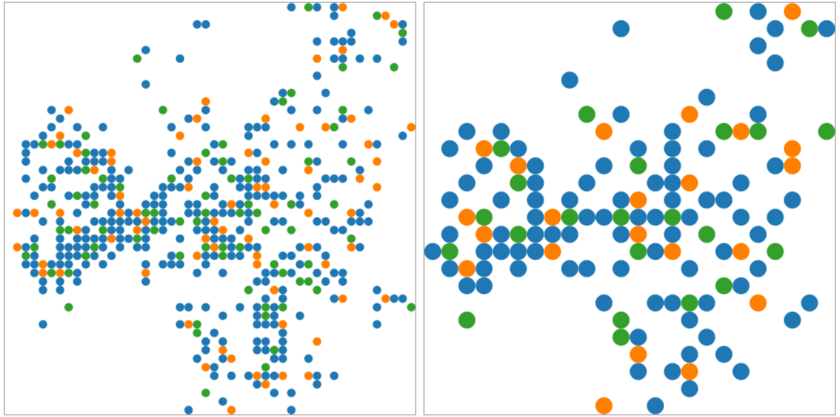
Wait until 8K UHD becomes the standard? Nope.

How to aggregate the dots?

We propose two approaches:



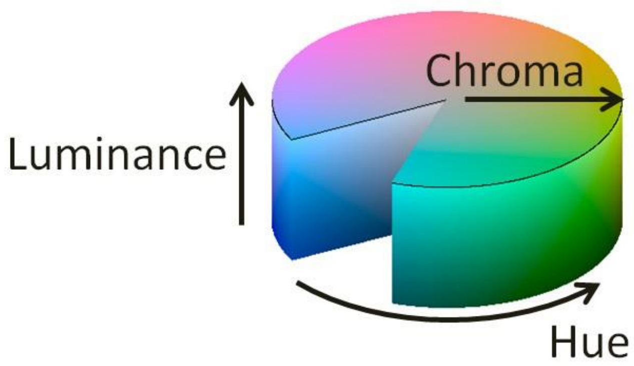
1. Blended colours

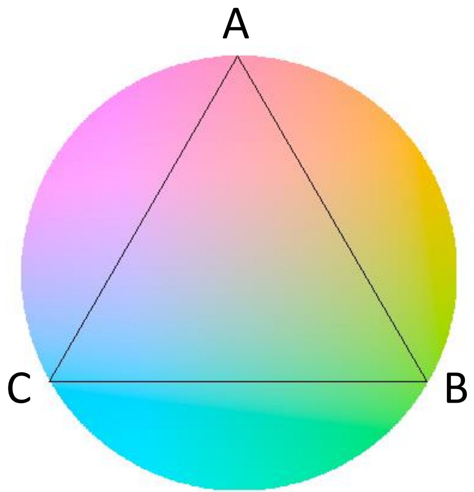


2. Super dots

Blended colours

Pixel colours are selected from the HCL colour space:



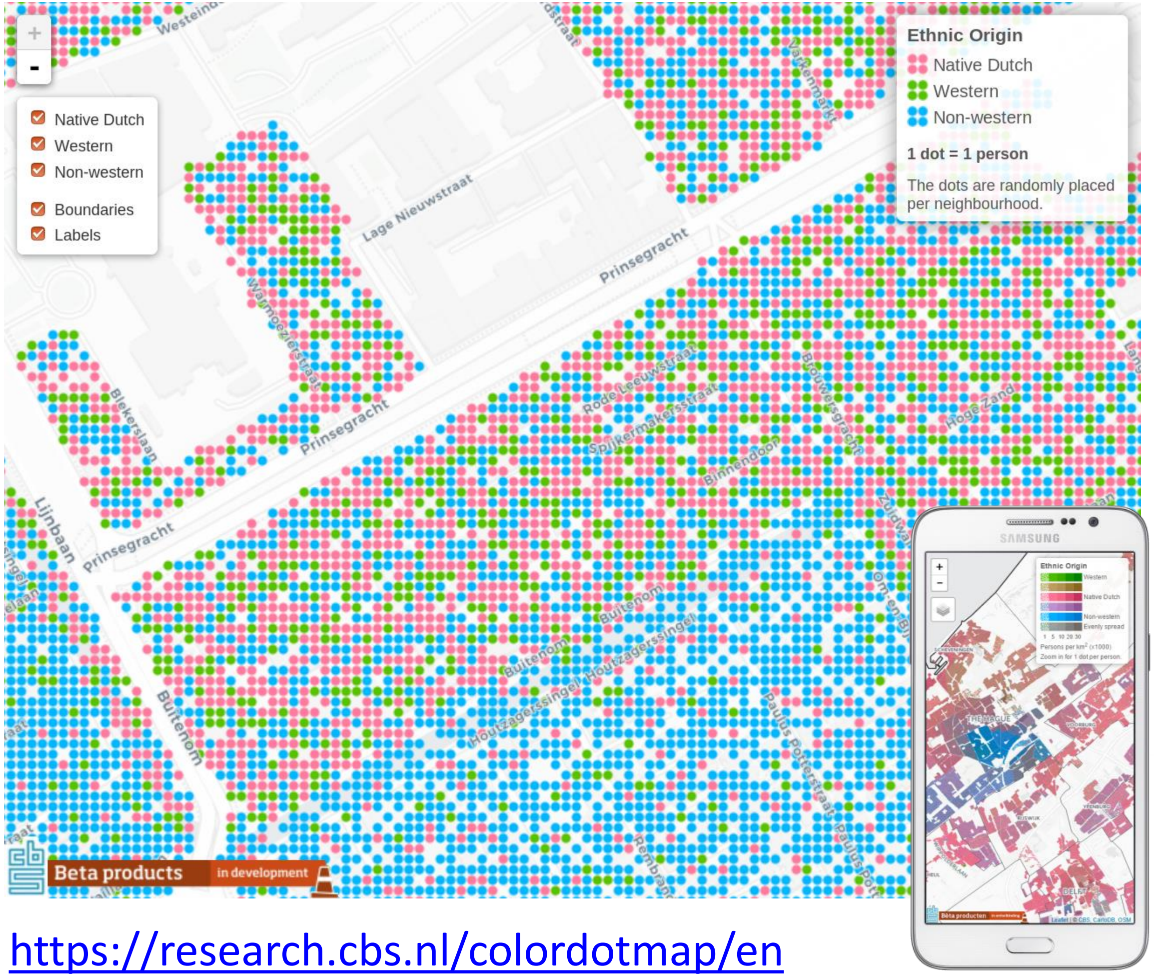


– **Luminance** for **density**
– **Hue** and **Chroma** for **composition**

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

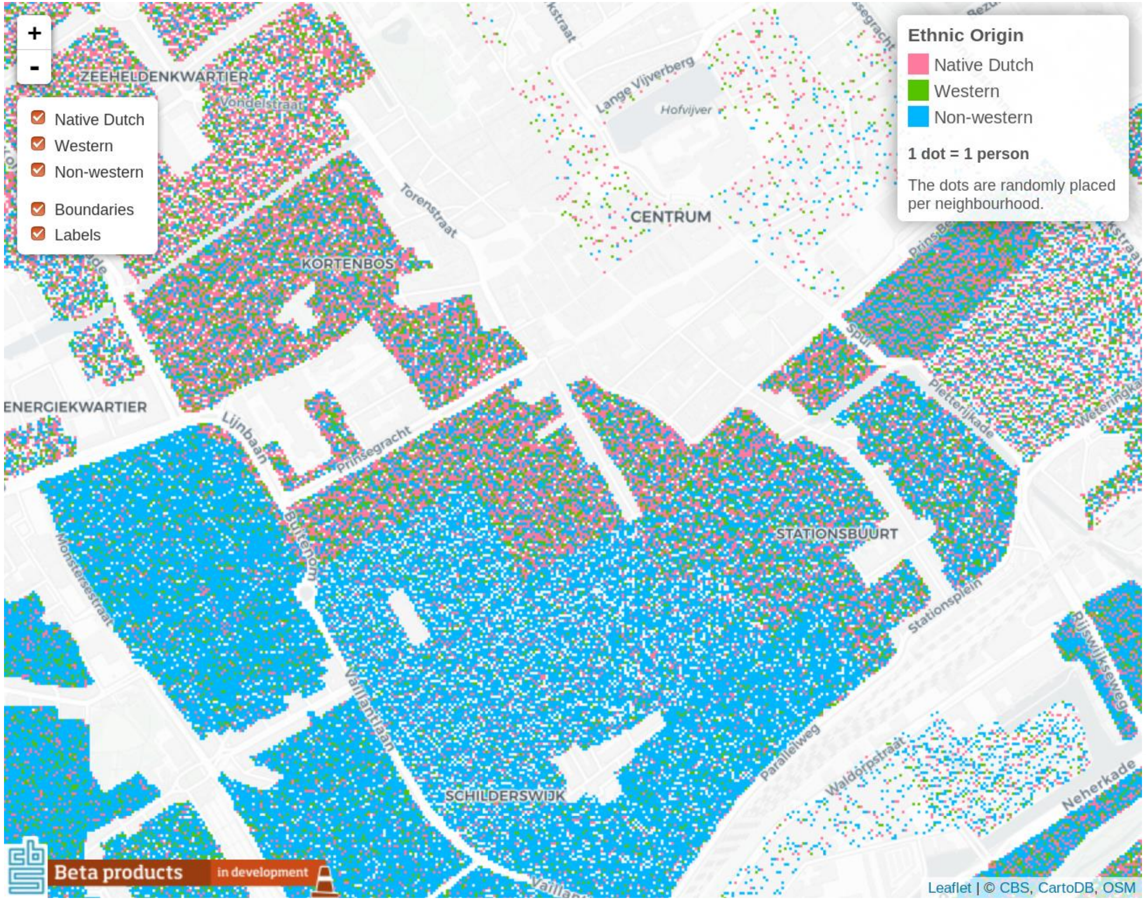
Dots are distributed uniformly per neighbourhood and placed in the land use category “residential”

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”

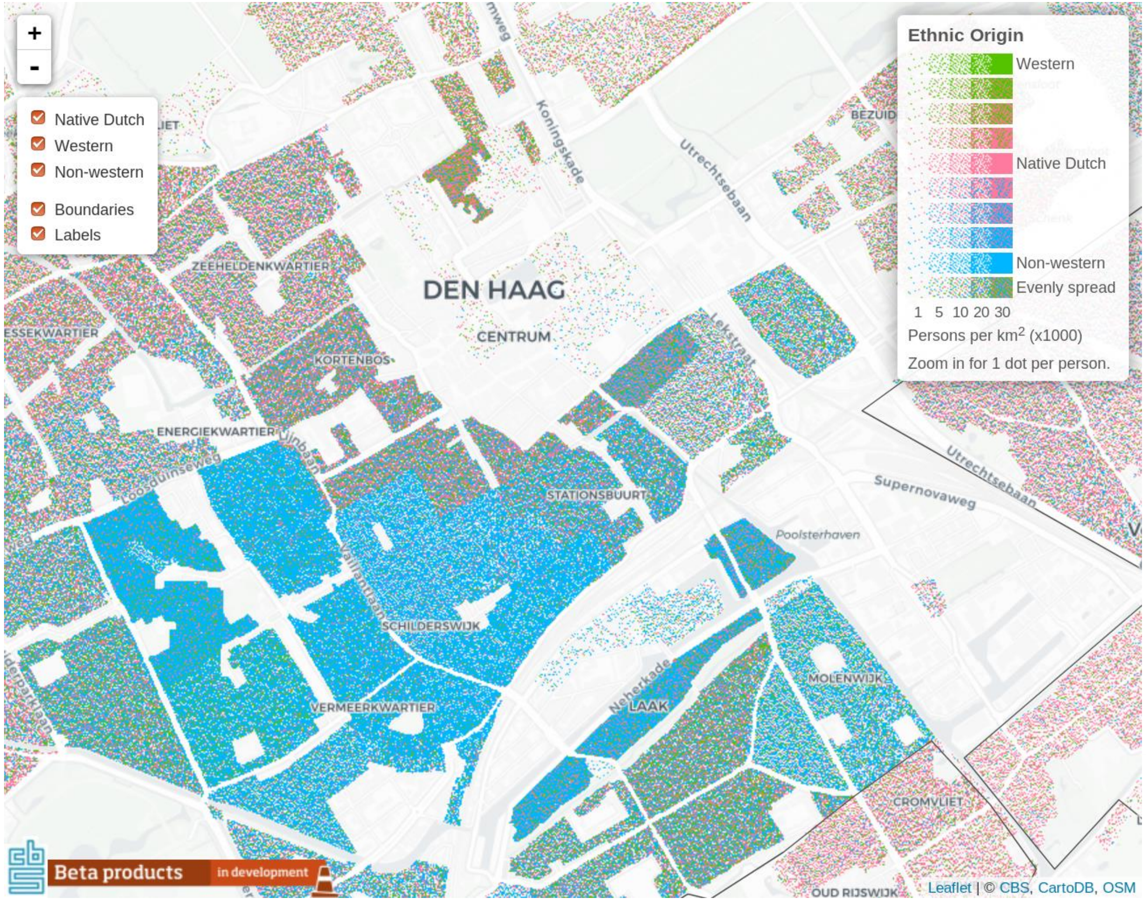


Published a CBDS beta product: https://research.cbs.nl/colordotmap/en

Application

Migration background of the Dutch population

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

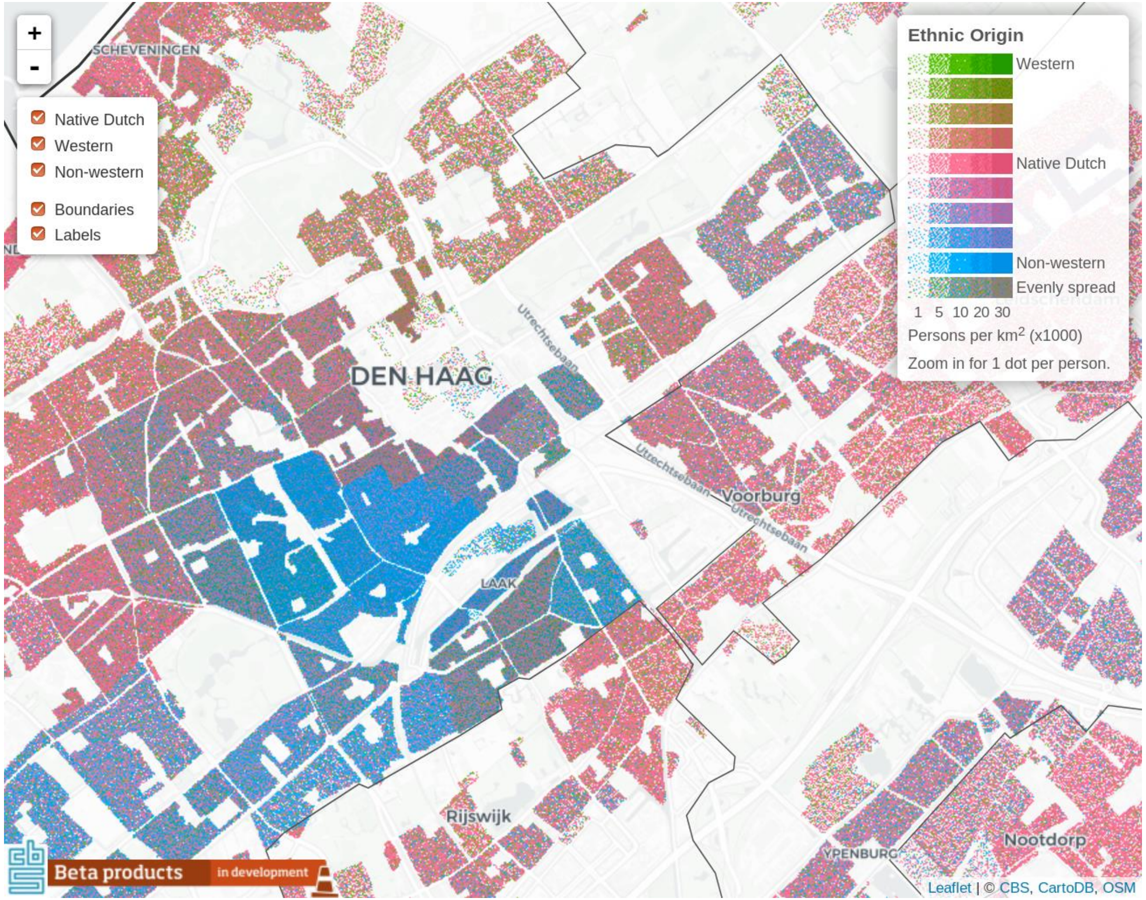


Published a CBDS beta product: https://research.cbs.nl/colordotmap/en

Application

Migration background of the Dutch population

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

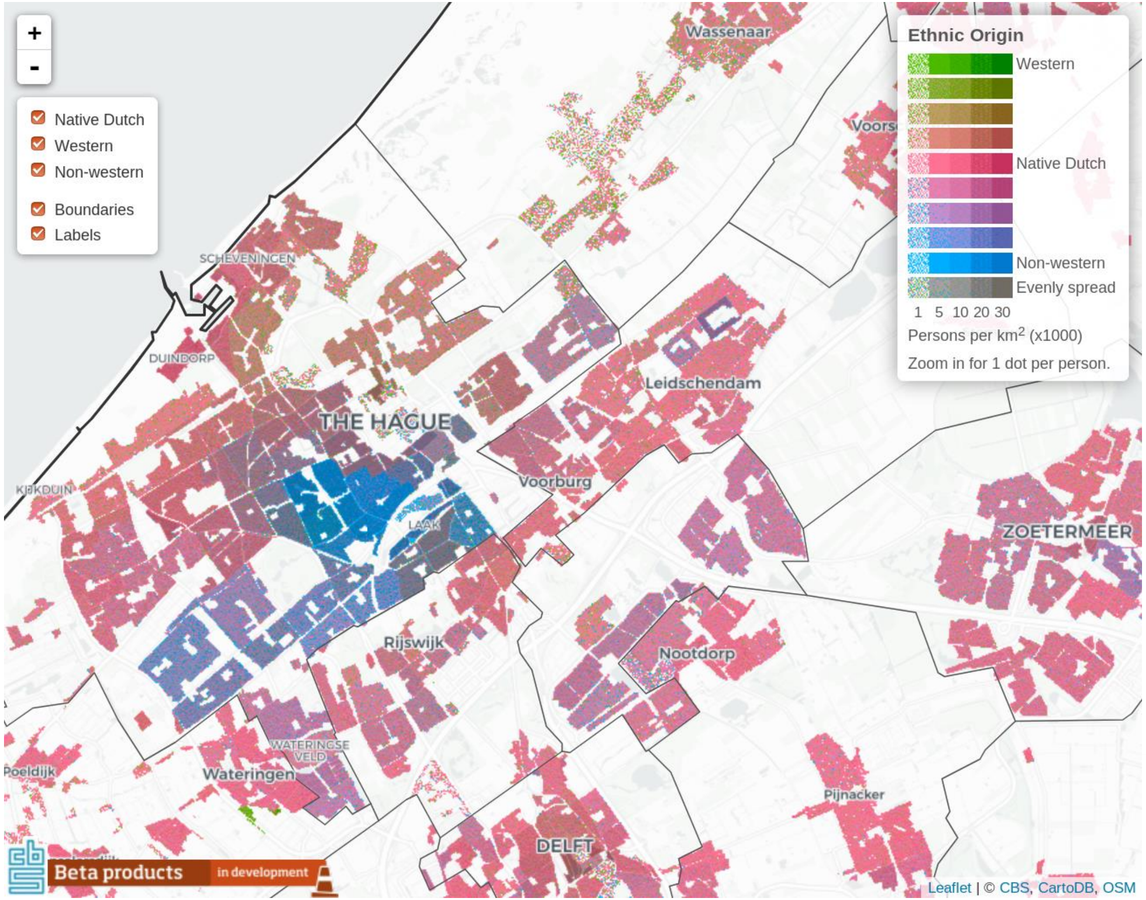


Published a CBDS beta product: https://research.cbs.nl/colordotmap/en

Application

Migration background of the Dutch population

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

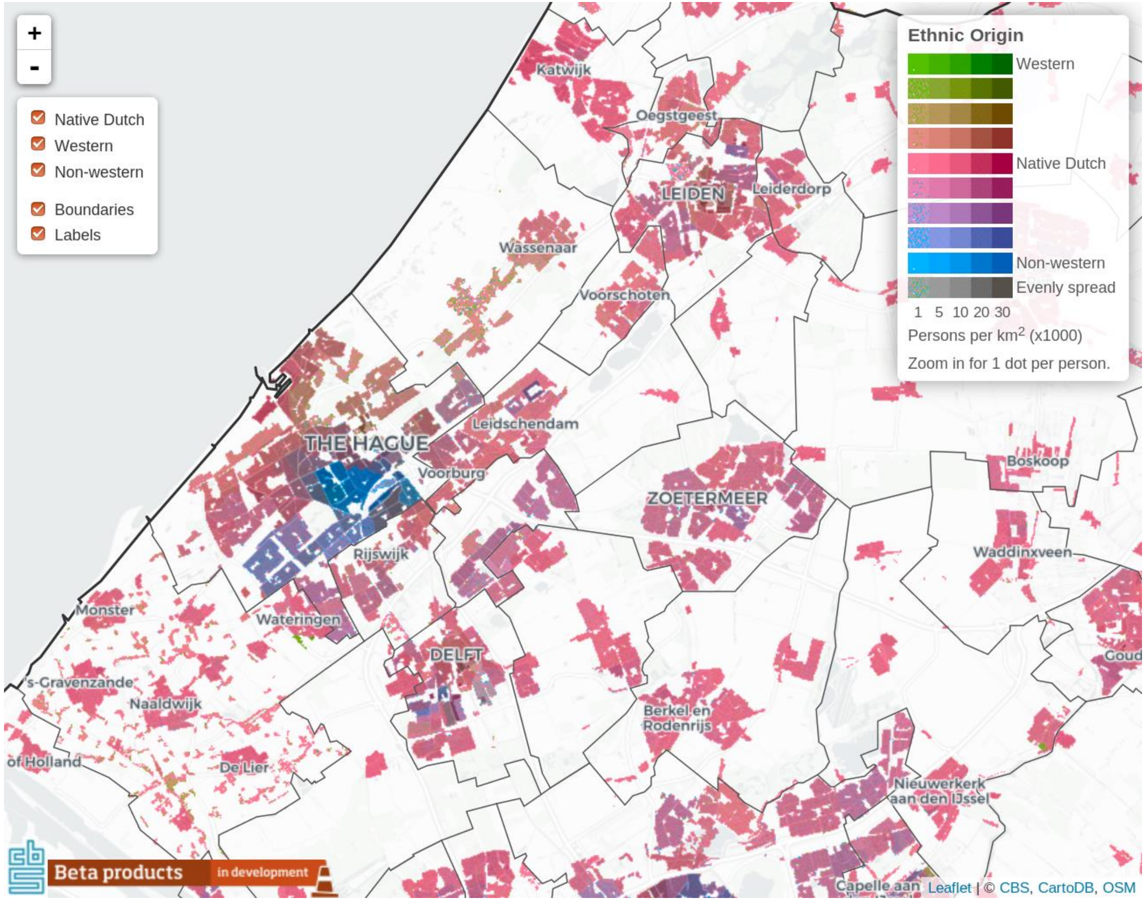


Published a CBDS beta product: https://research.cbs.nl/colordotmap/en

Application

Migration background of the Dutch population

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

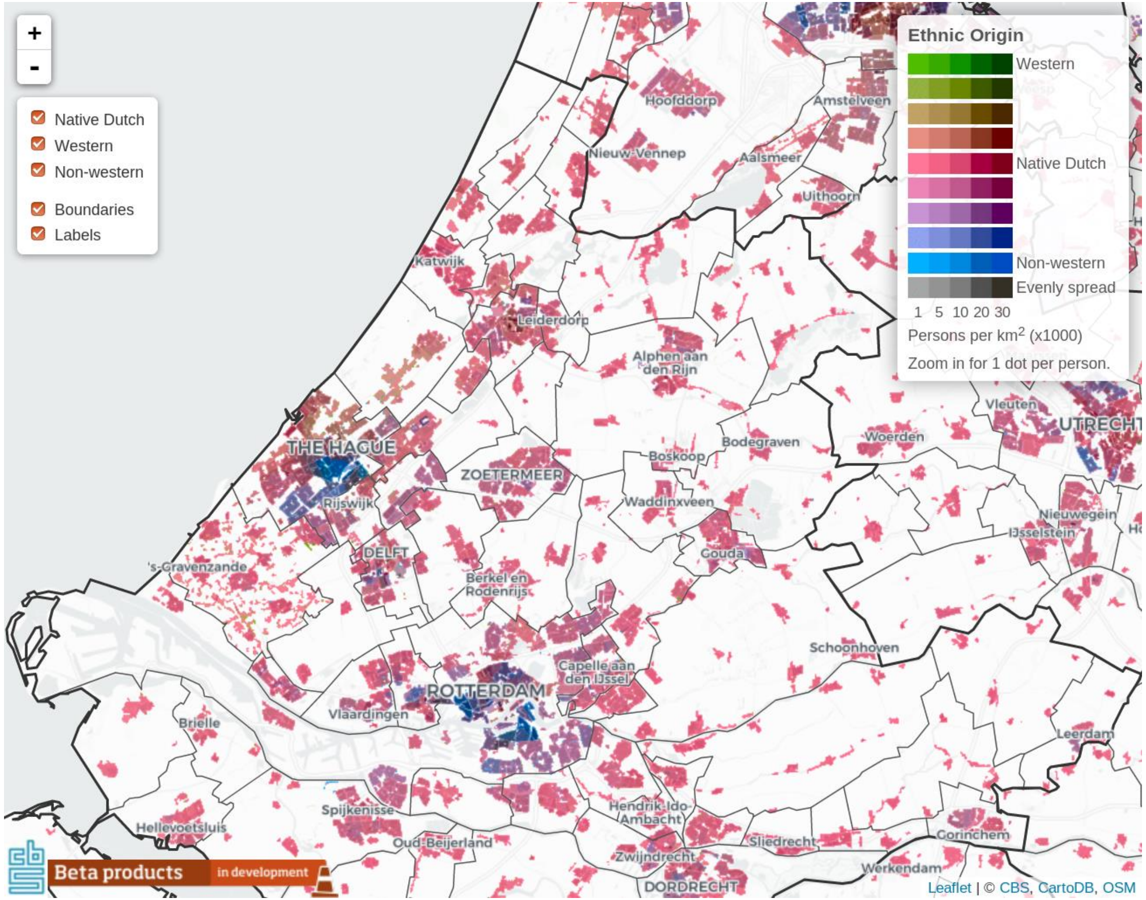


Published a CBDS beta product: https://research.cbs.nl/colordotmap/en

Application

Migration background of the Dutch population

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

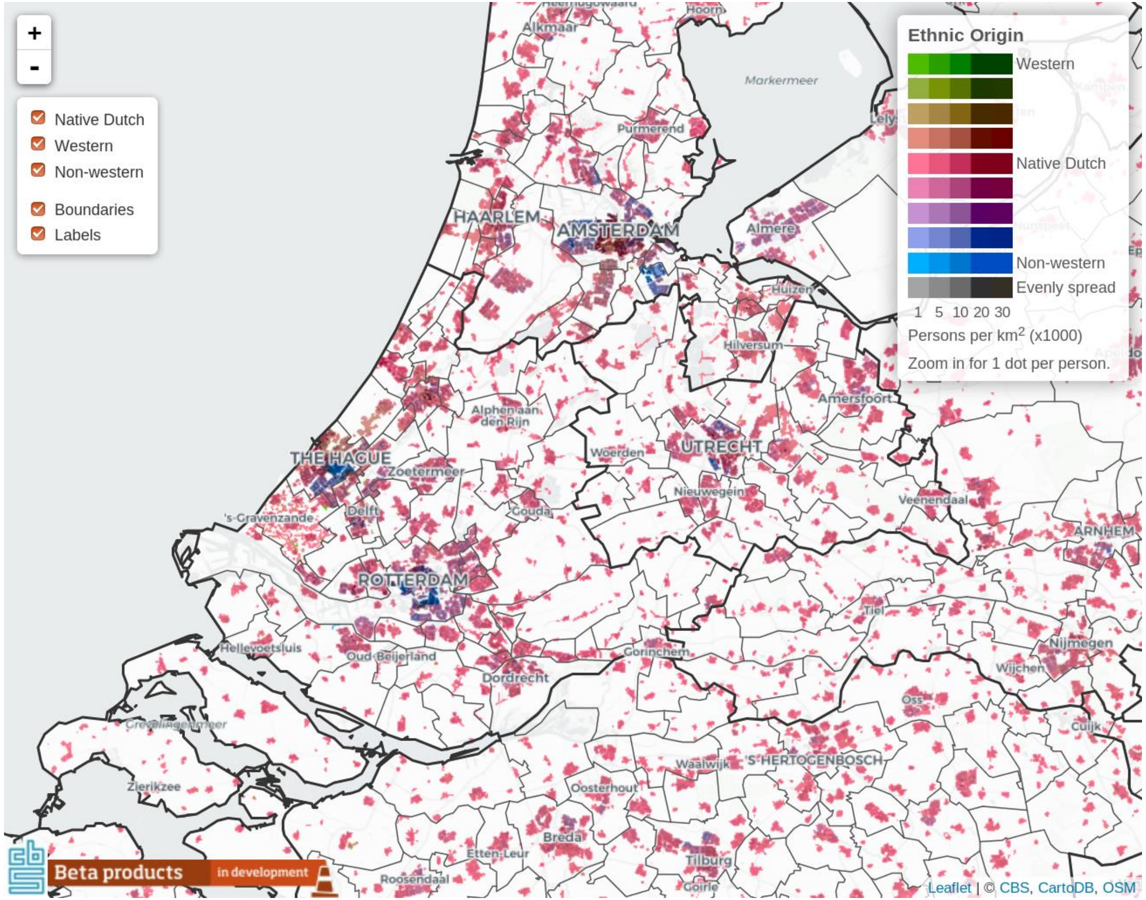


Published a CBDS beta product: https://research.cbs.nl/colordotmap/en

Application

Migration background of the Dutch population

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

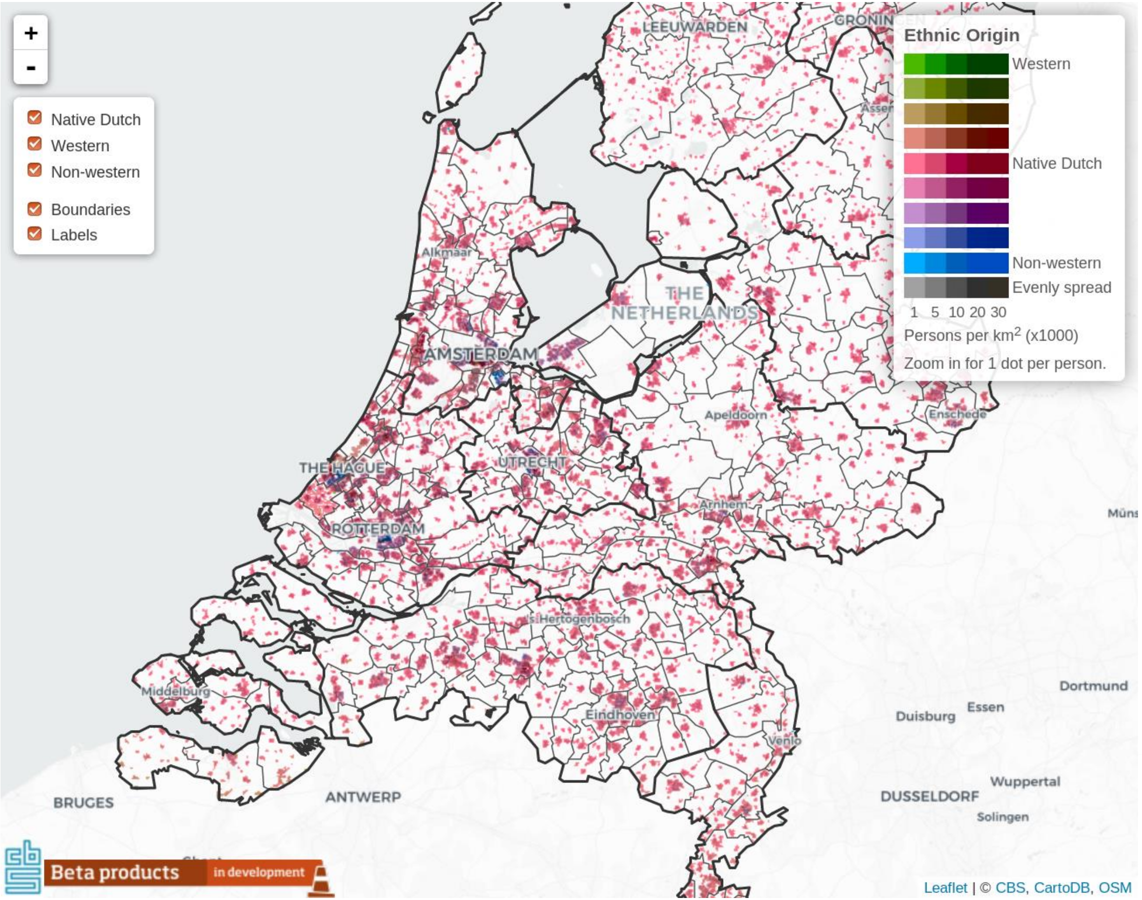


Published a CBDS beta product: https://research.cbs.nl/colordotmap/en

Application

Migration background of the Dutch population

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

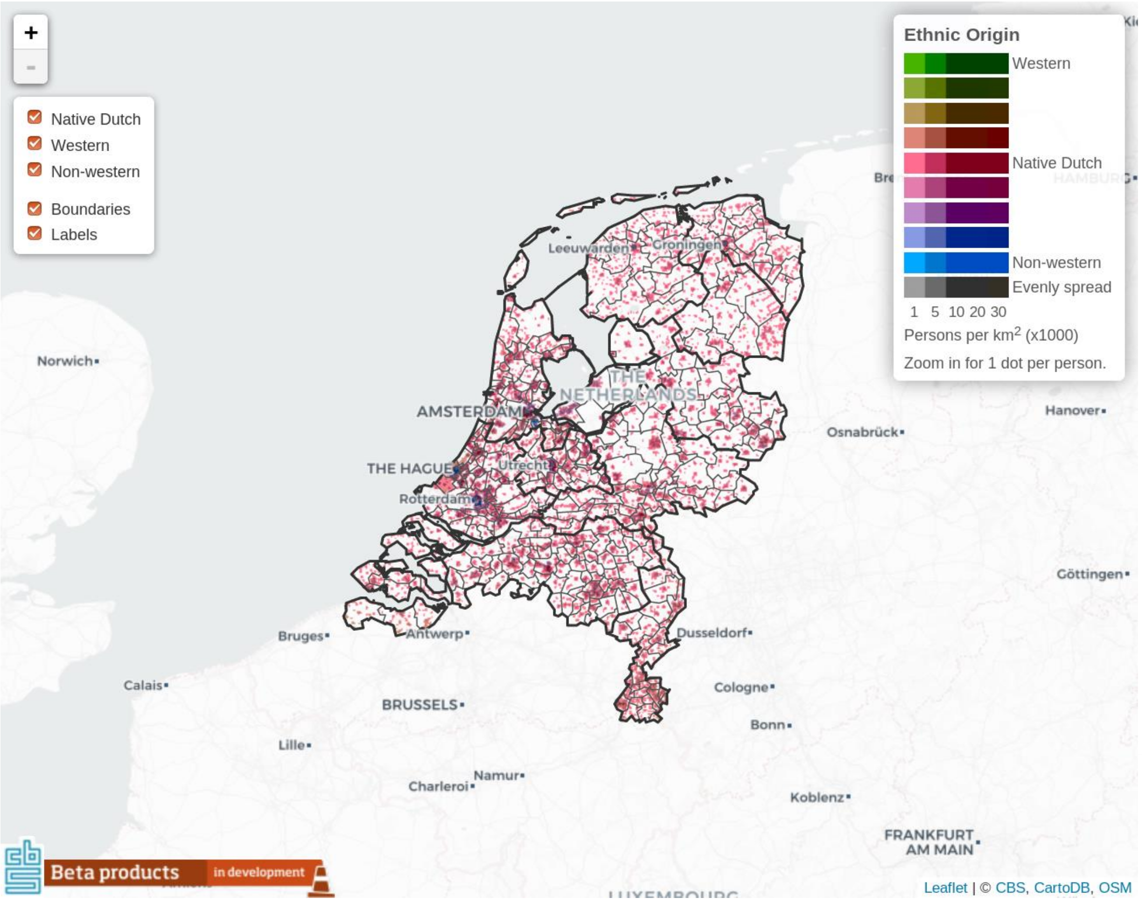


Published a CBDS beta product: https://research.cbs.nl/colordotmap/en

Application

Migration background of the Dutch population

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



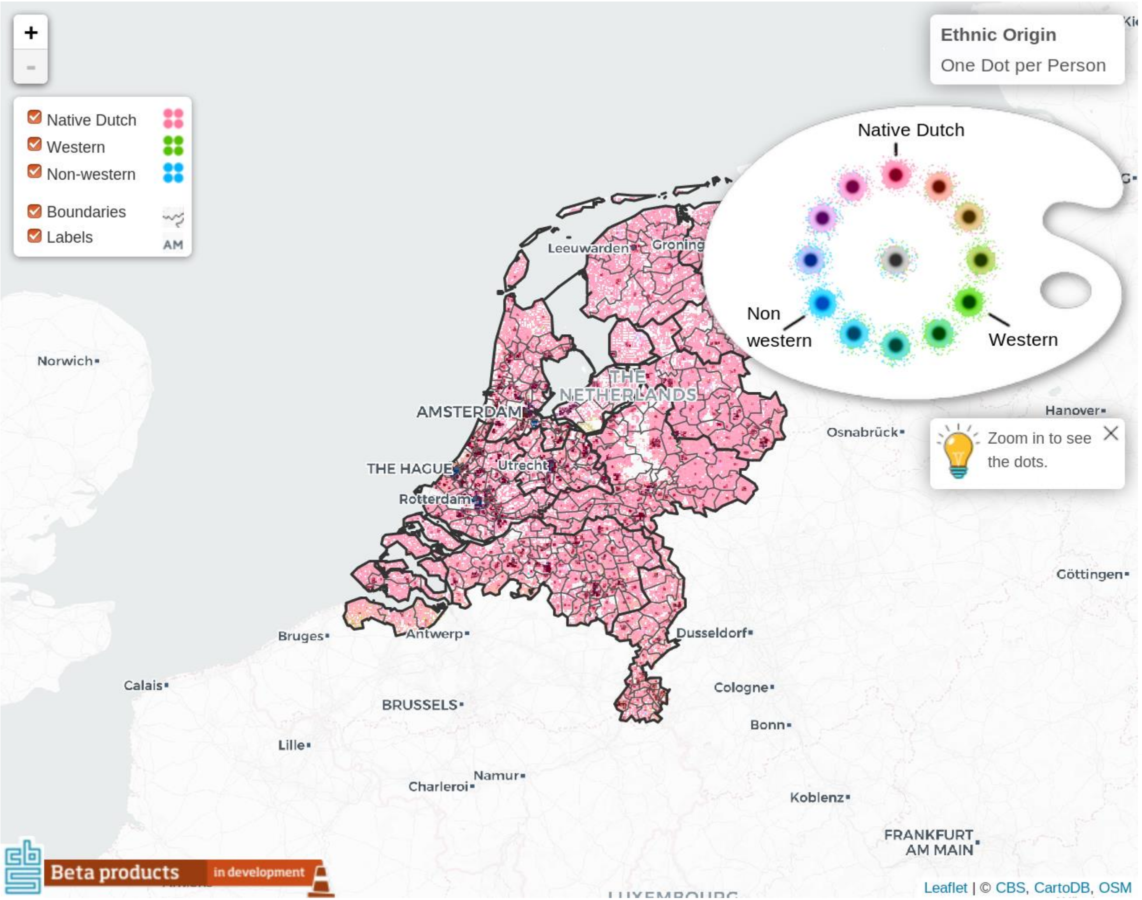
Published a CBDS beta product: https://research.cbs.nl/colordotmap/en

Application

Migration background of the Dutch population

Experimental version:

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

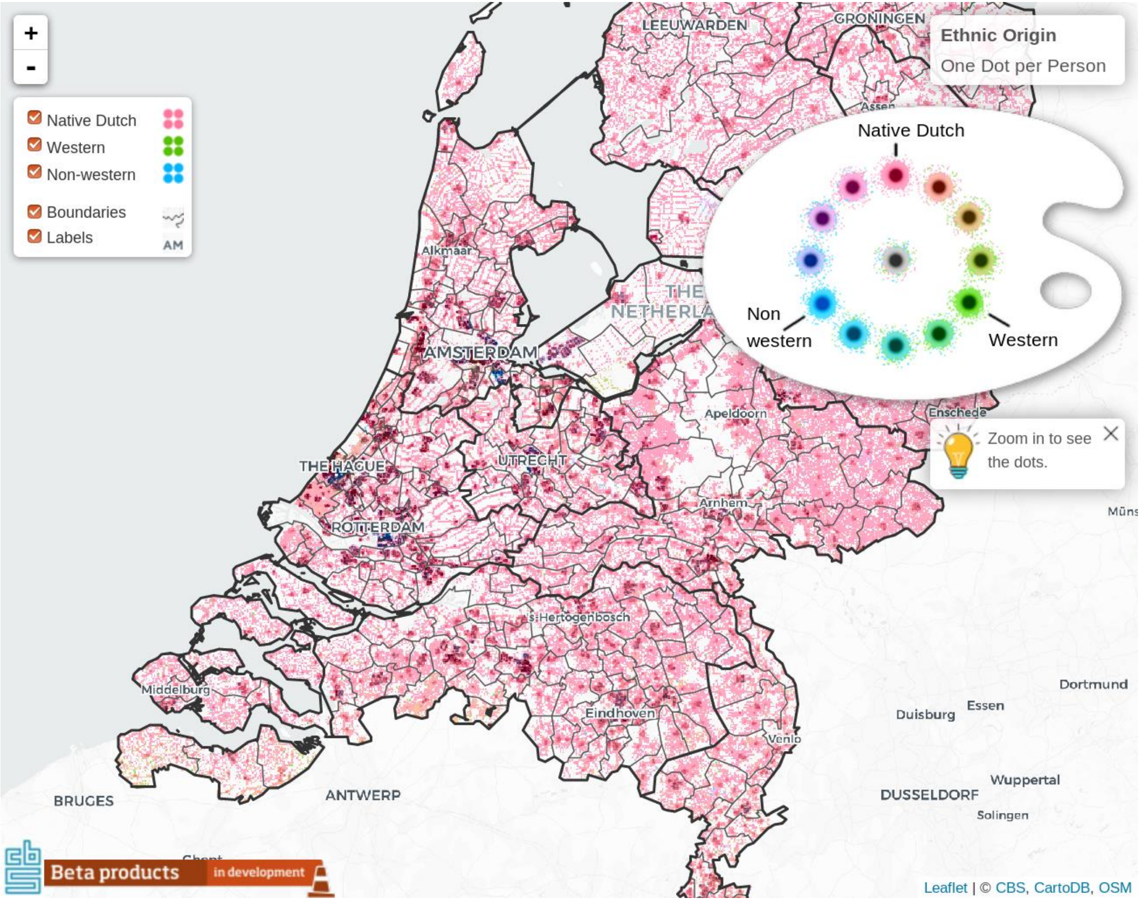


Application

Migration background of the Dutch population

Experimental version:

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

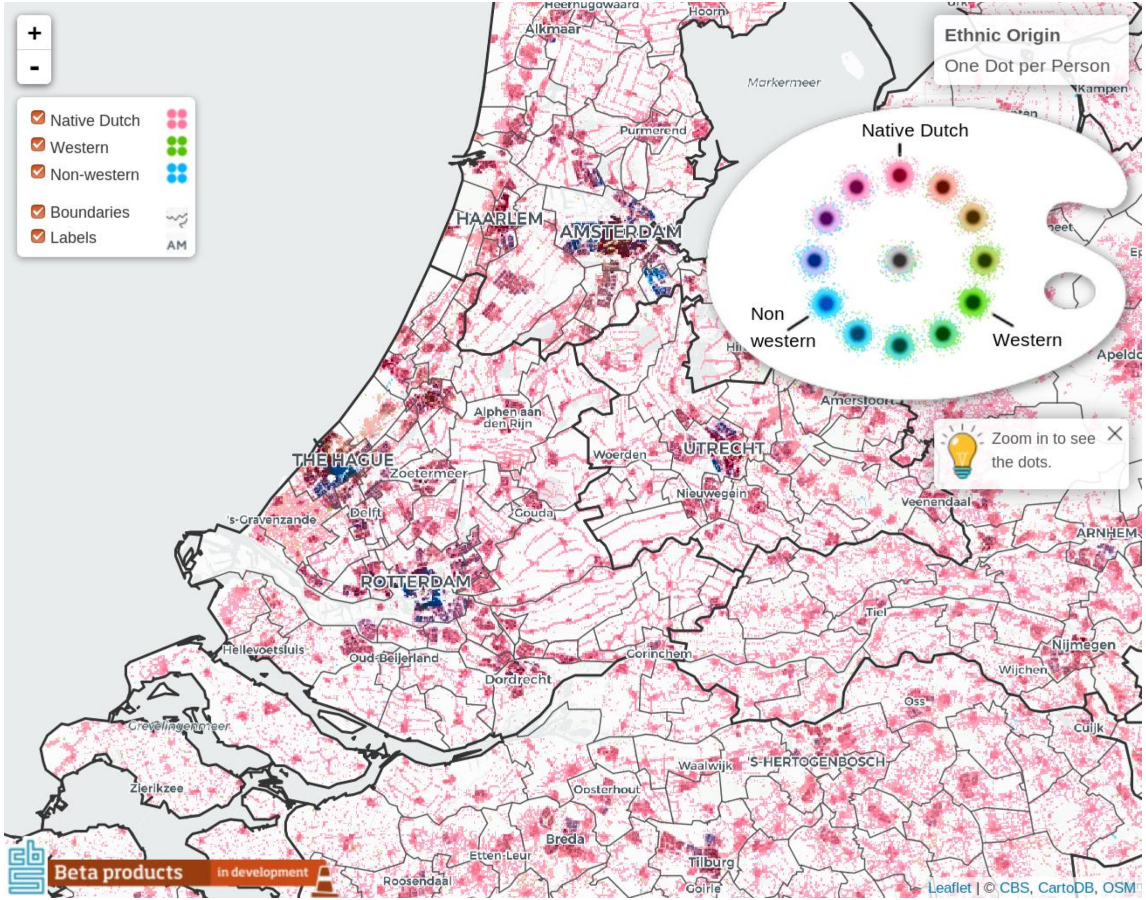


Application

Migration background of the Dutch population

Experimental version:

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

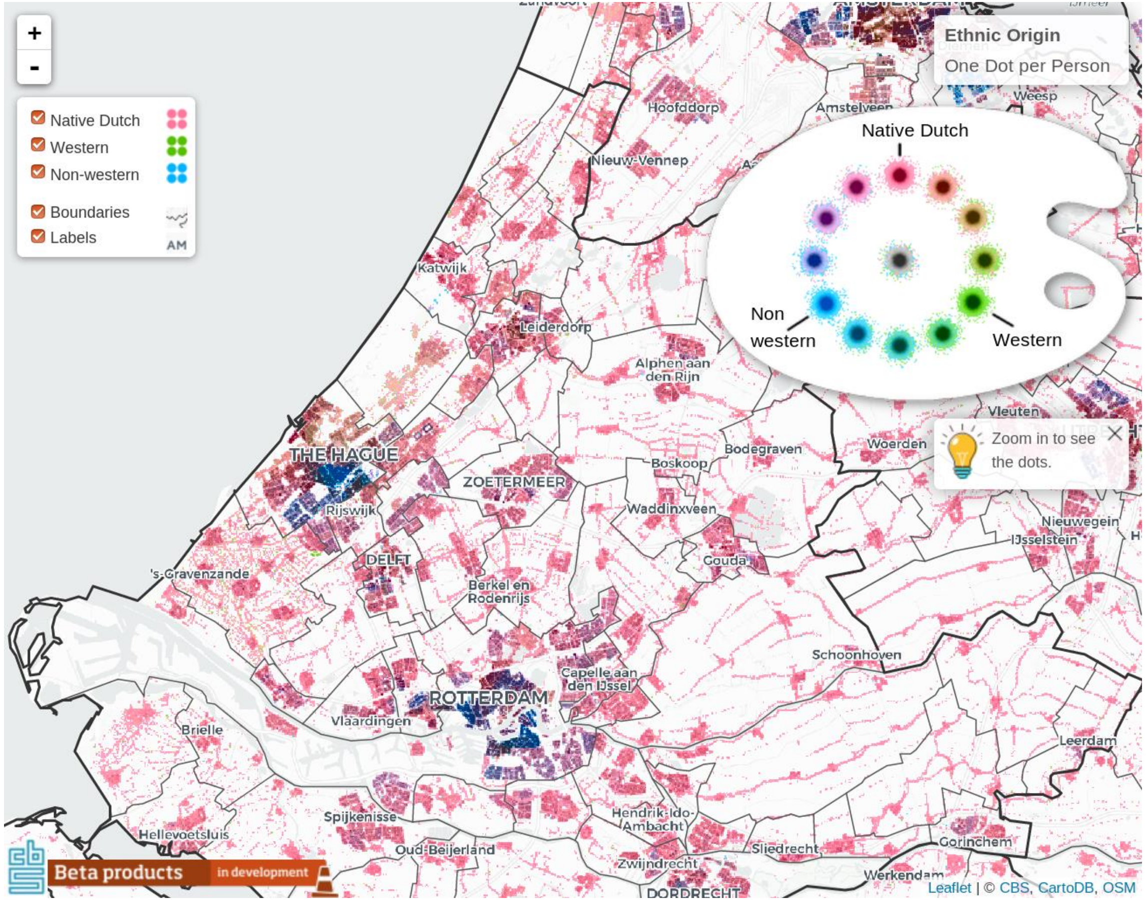


Application

Migration background of the Dutch population

Experimental version:

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

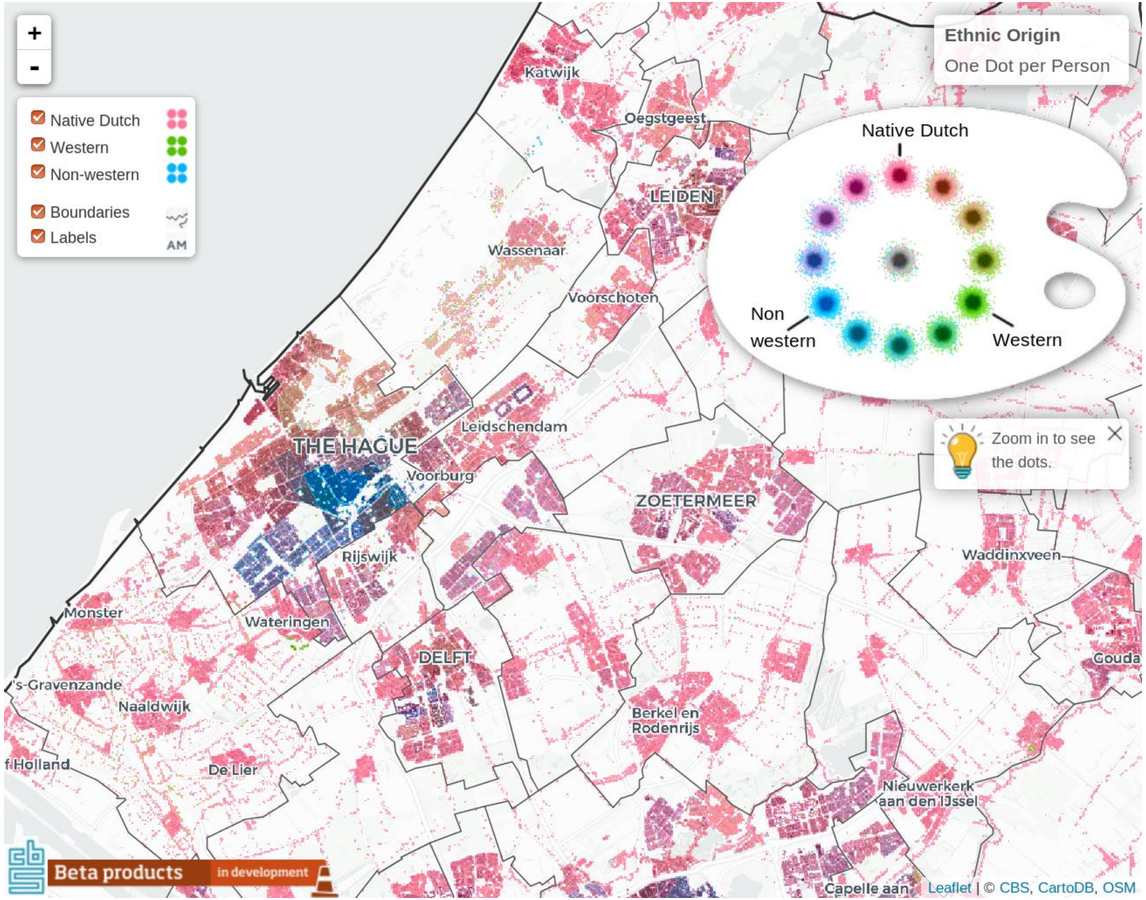


Application

Migration background of the Dutch population

Experimental version:

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

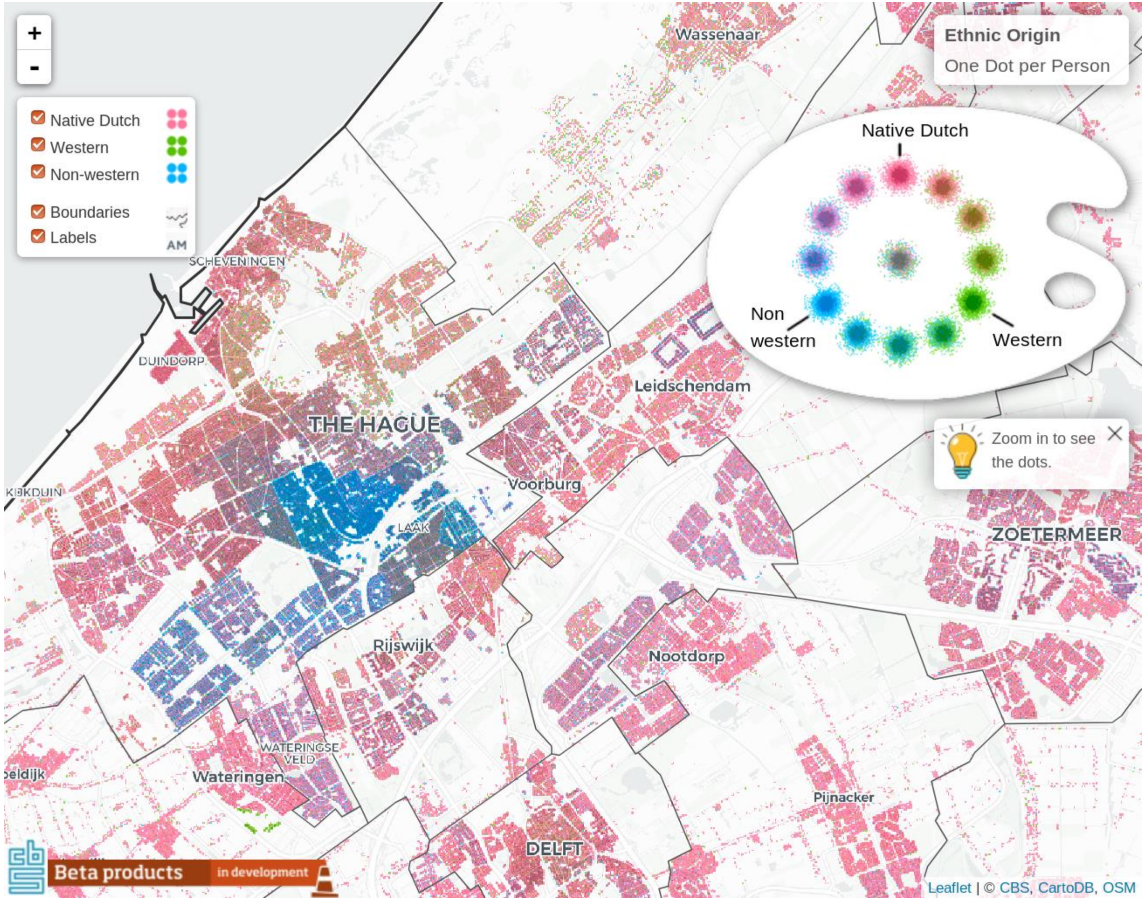


Application

Migration background of the Dutch population

Experimental version:

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

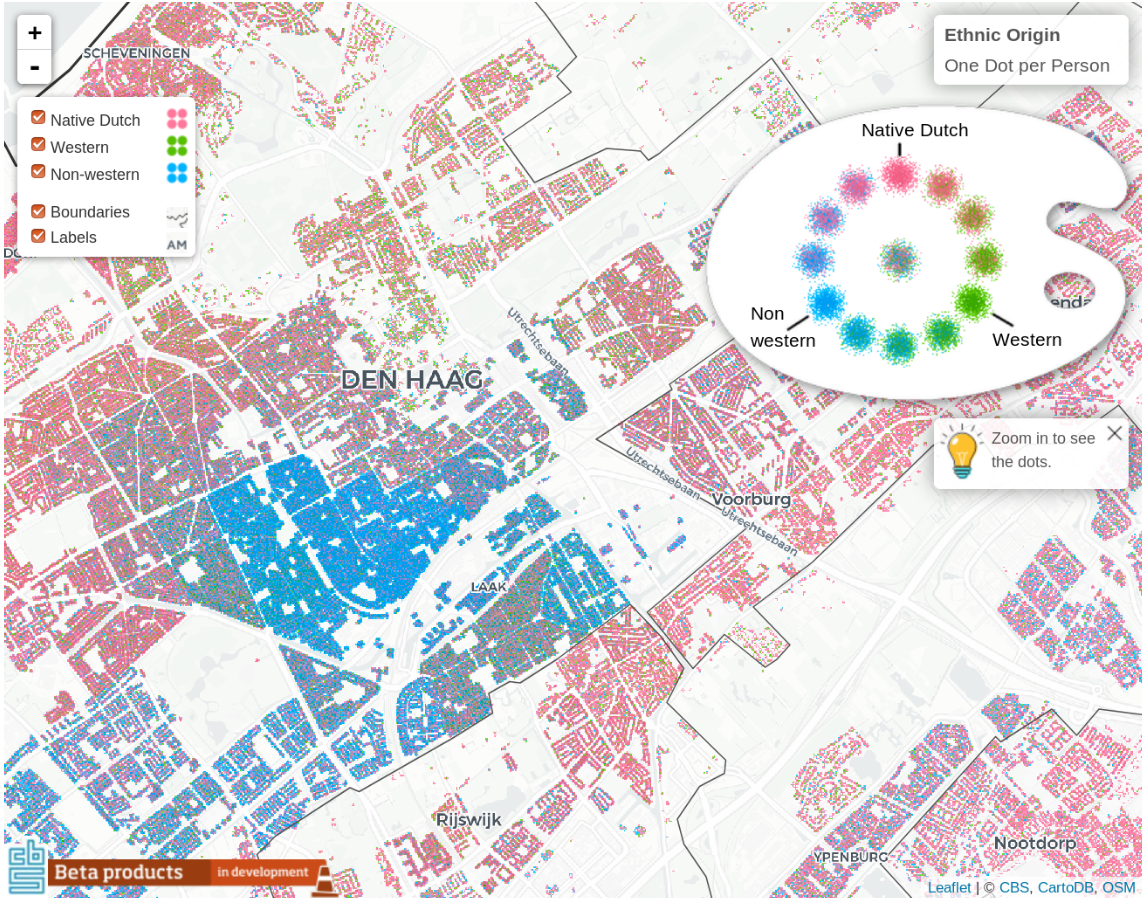


Application

Migration background of the Dutch population

Experimental version:

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

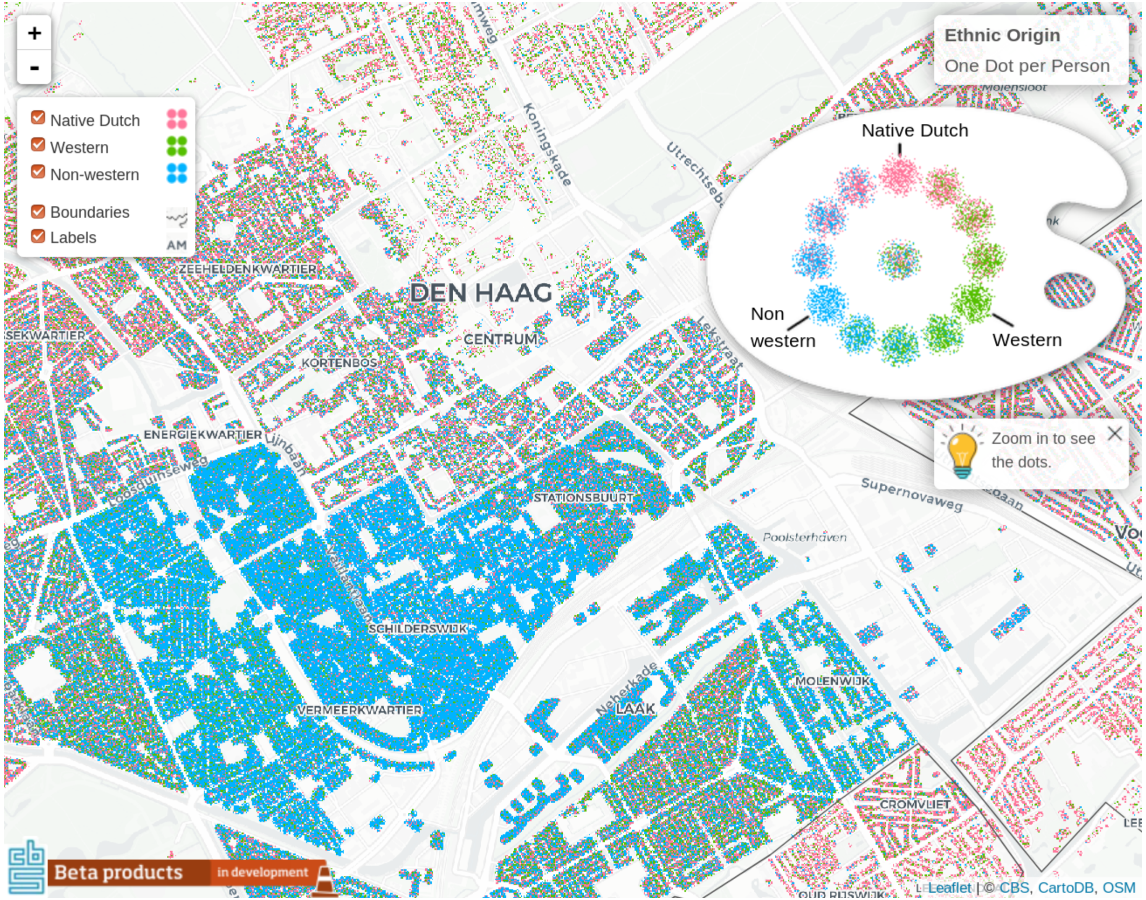


Application

Migration background of the Dutch population

Experimental version:

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

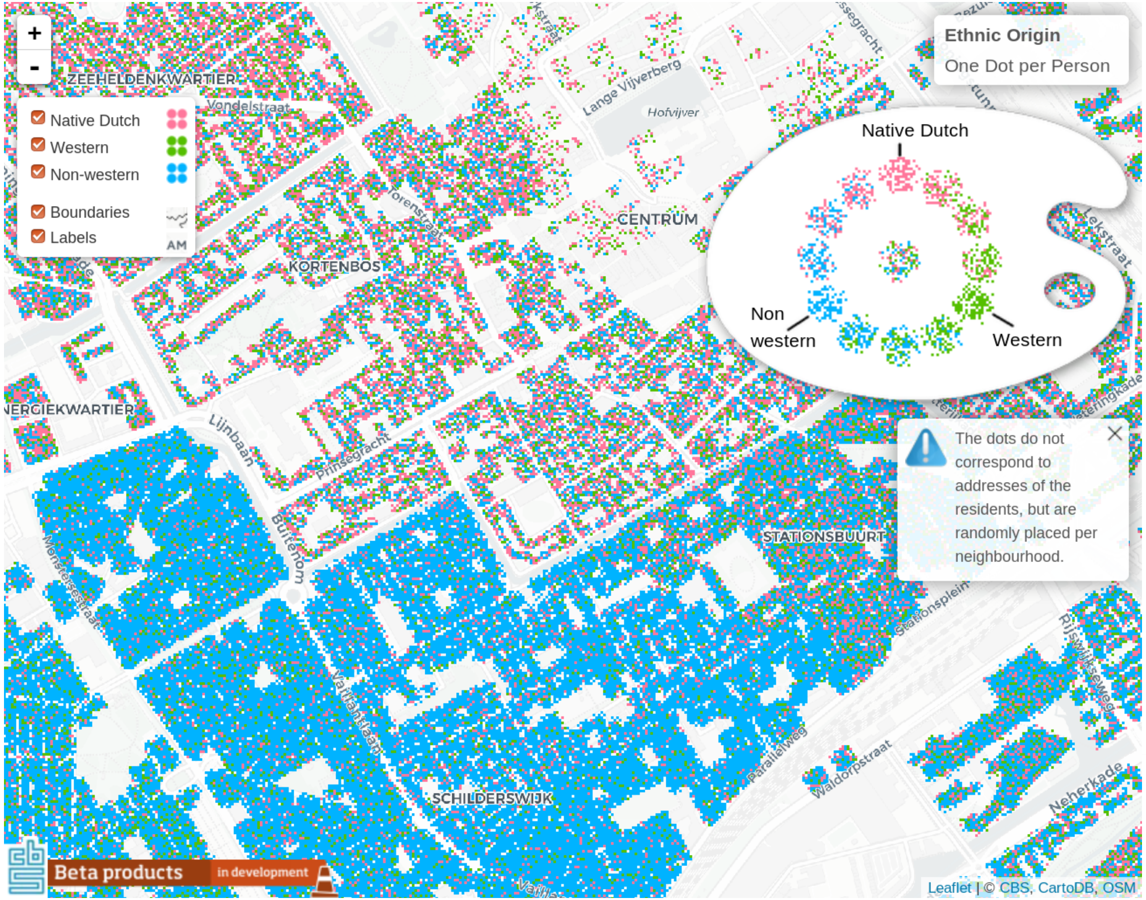


Application

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:

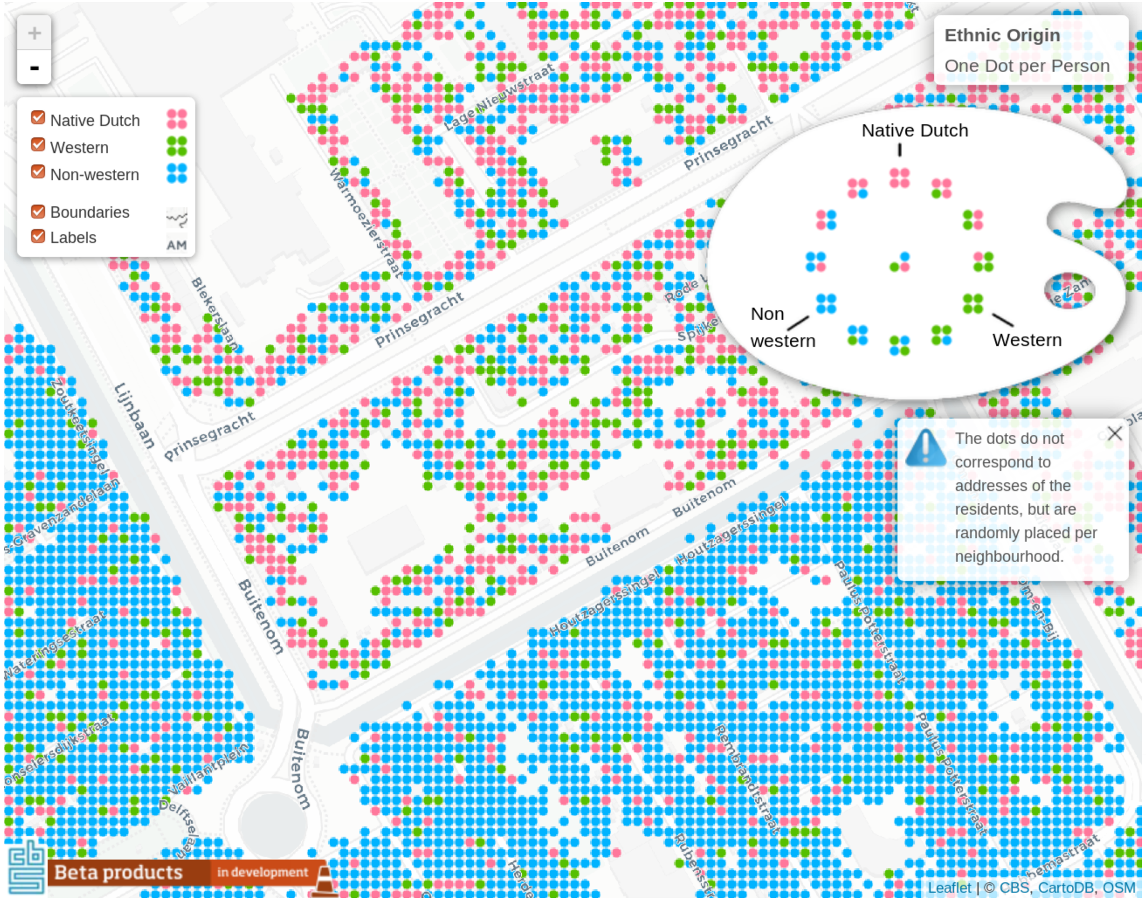
- Dots are placed in building areas (using the BAG register)
- "Artistic" legend

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.

Strange… Neighbourhoods appear pink from a distance, but from nearby, you clearly see the mix.

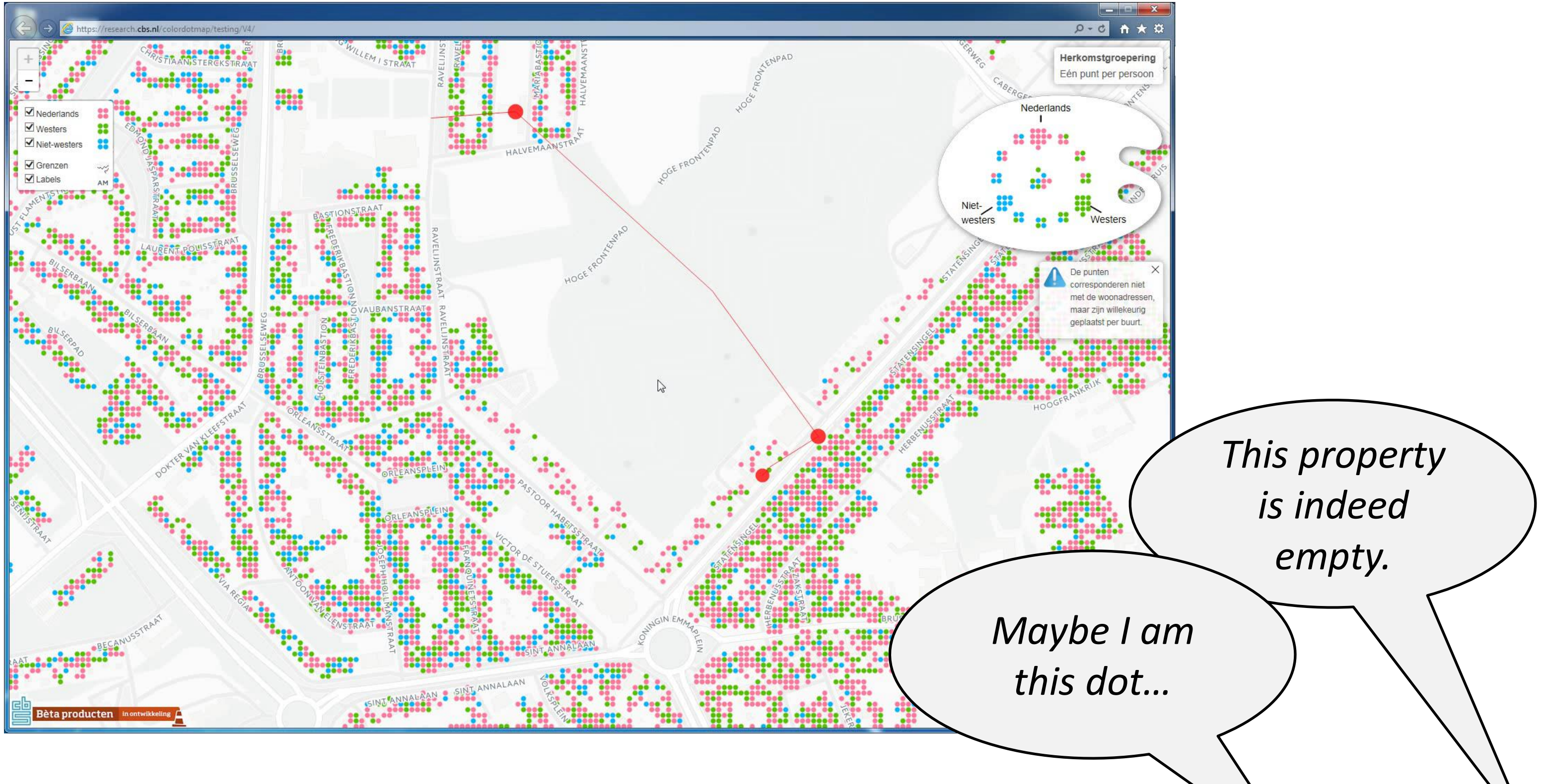
User study

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

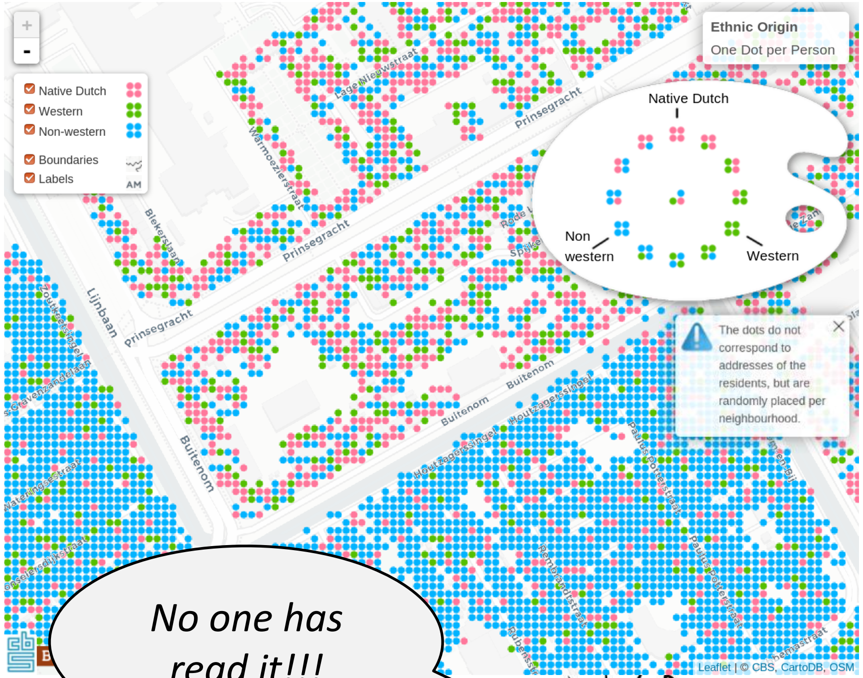
Comparison between **original** and **experimental version** with **eye-tracking**.



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.



No one has read it!!!

A good visualization should be self-explanatory.

How to deal with privacy?

Some ideas / guidelines:

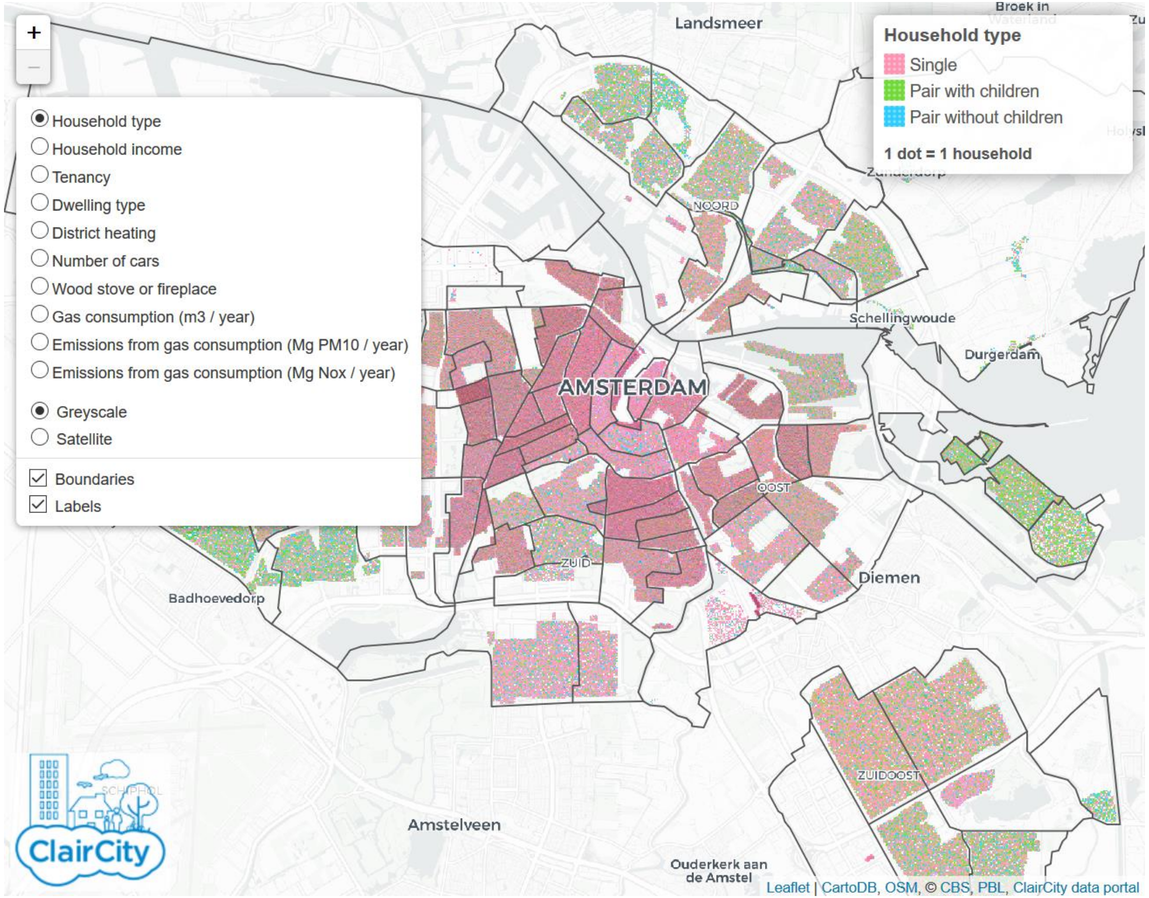
- Areas should not be too detailed (global land use is better than detailed building areas)
- Draw neighbourhood borders
- Limit the zoom level (not to close)

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

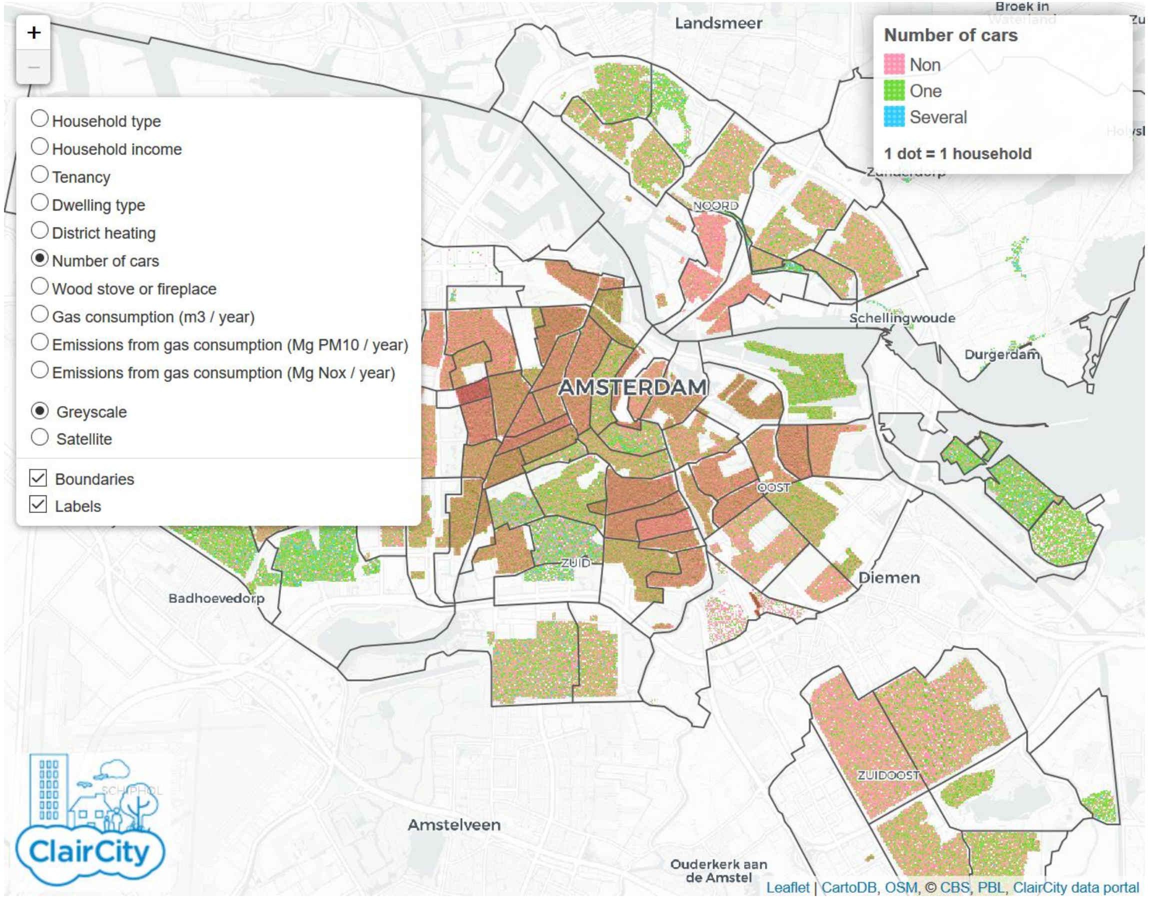


- Simulated data on neighbourhood level for Amsterdam
- Each dot represents a household
- Dots are placed in residential areas (OpenStreetMap) per neighbourhood

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



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

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

http://www.claircity.eu/

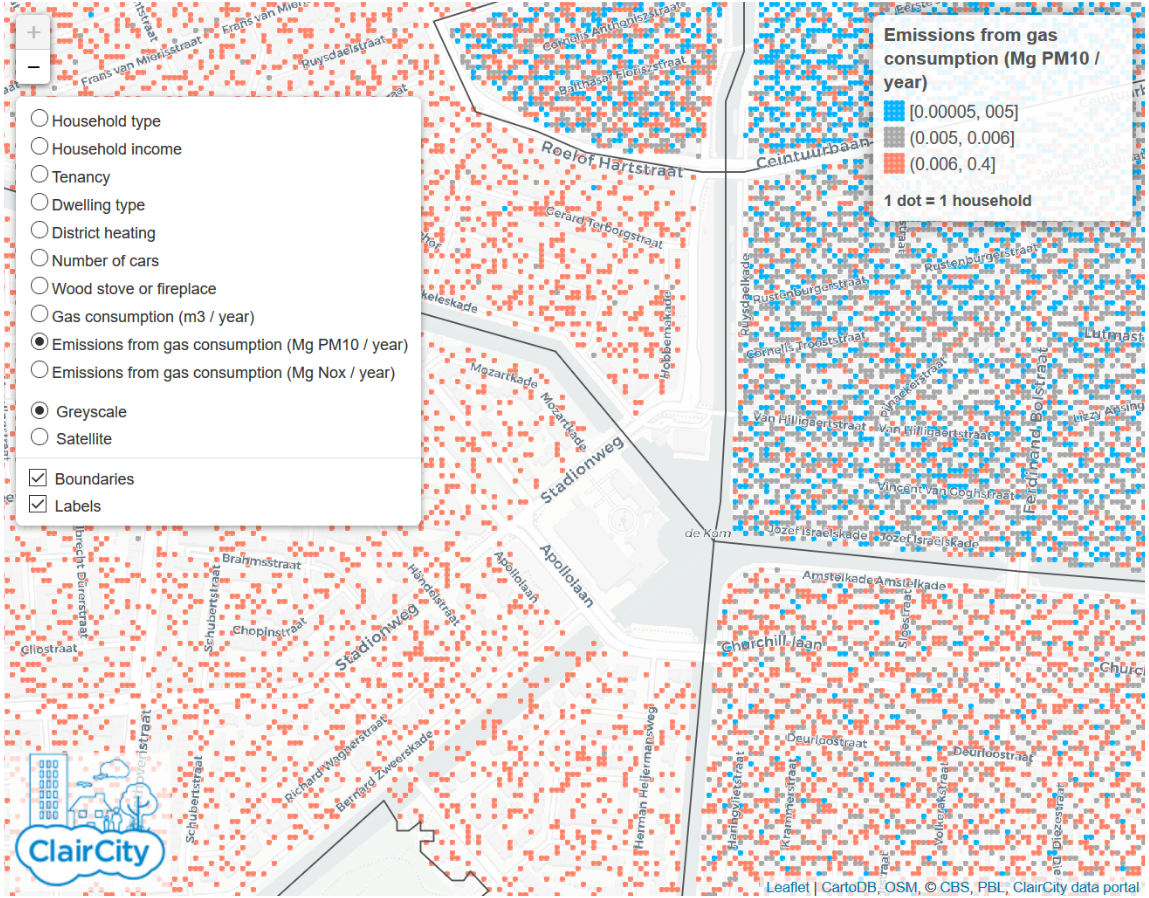
Application

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

http://www.claircity.eu/

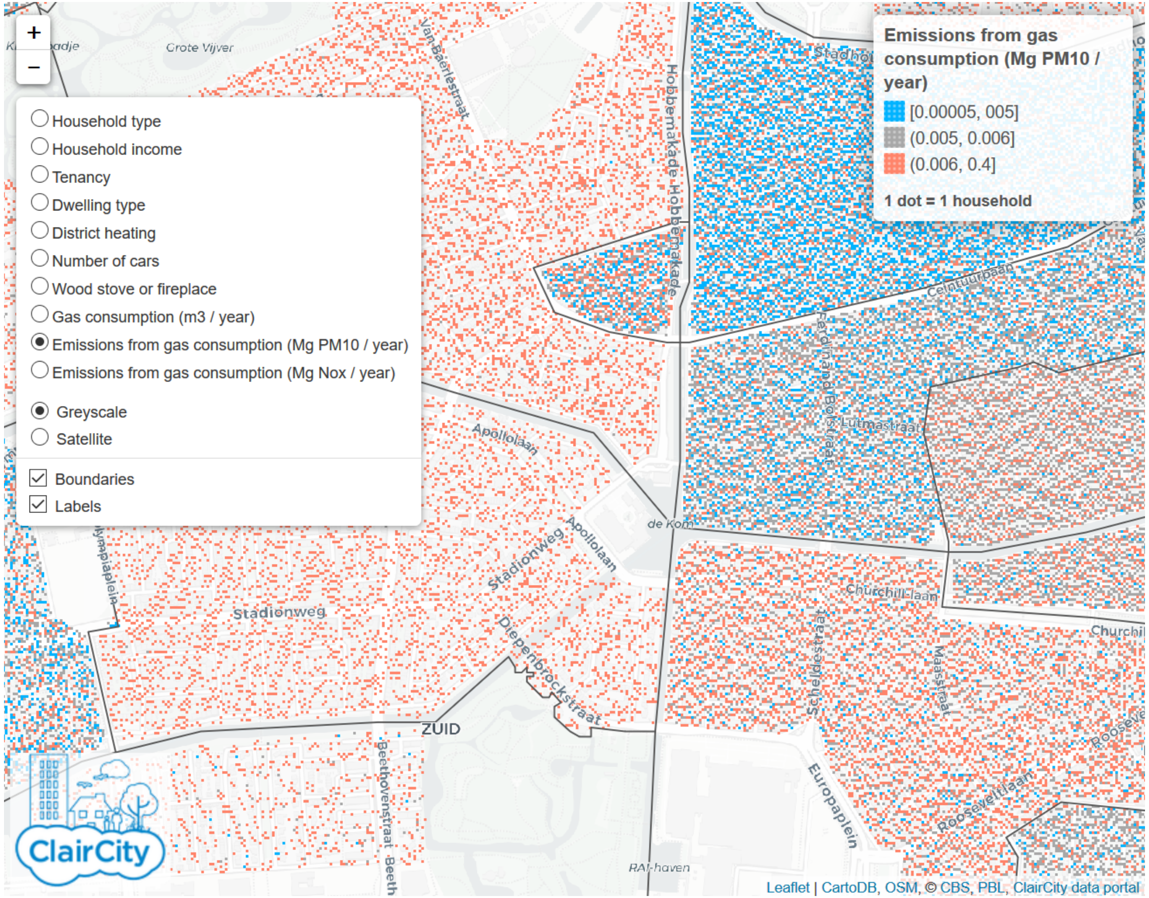
Application

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



http://www.claircity.eu/

Application

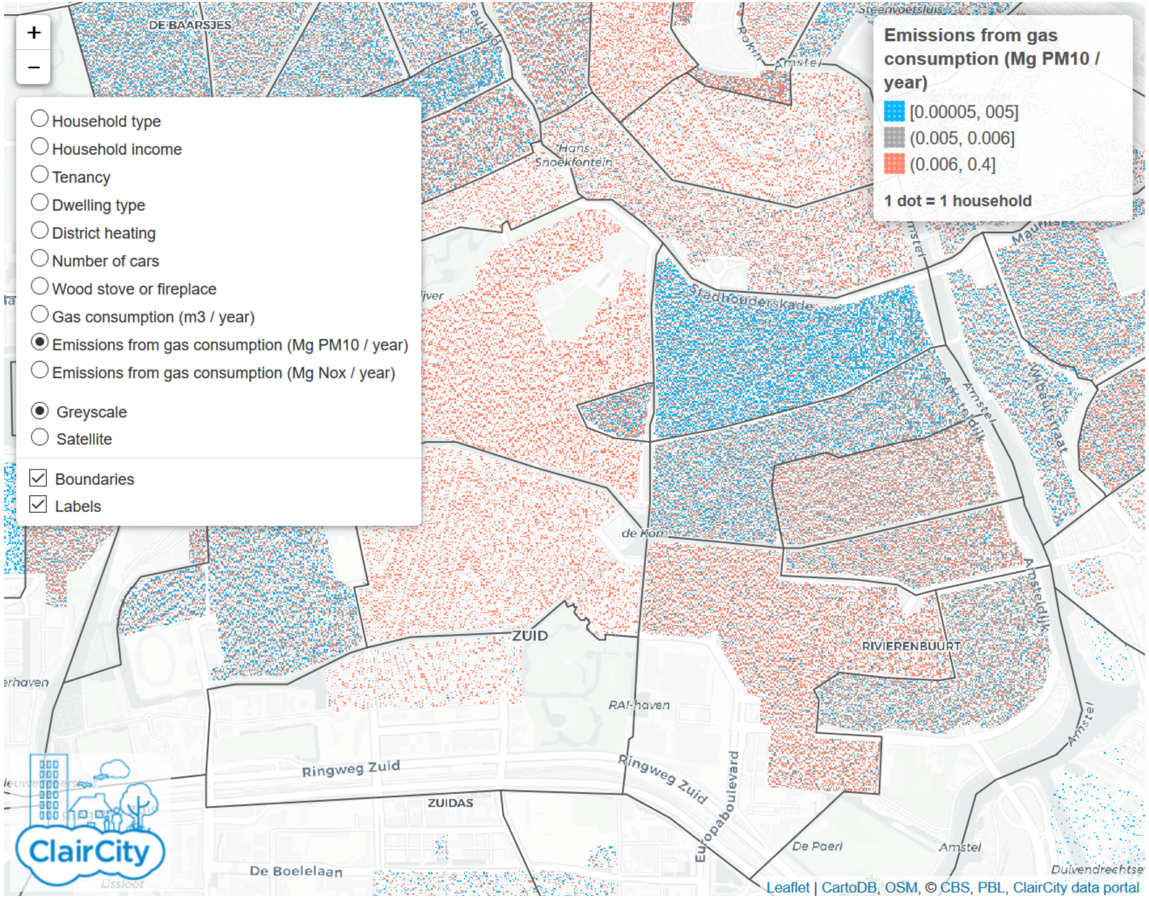


- Simulated data on neighbourhood level for Amsterdam
- Each dot represents a household
- Dots are placed in residential areas (OpenStreetMap) per neighbourhood

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



http://www.claircity.eu/

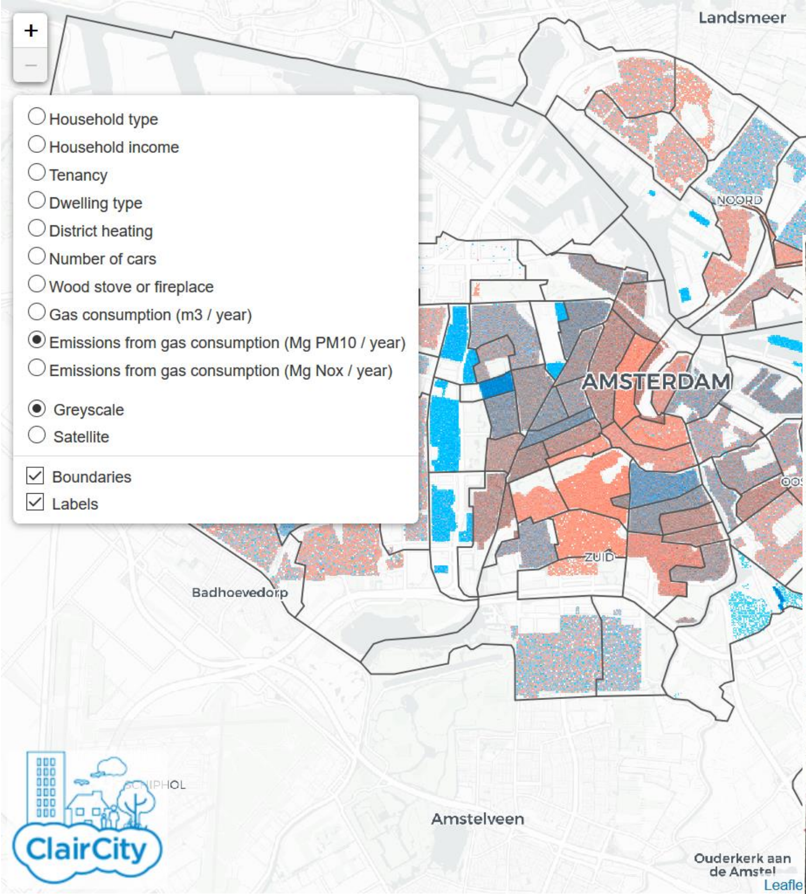
Application

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

http://www.claircity.eu/

Application

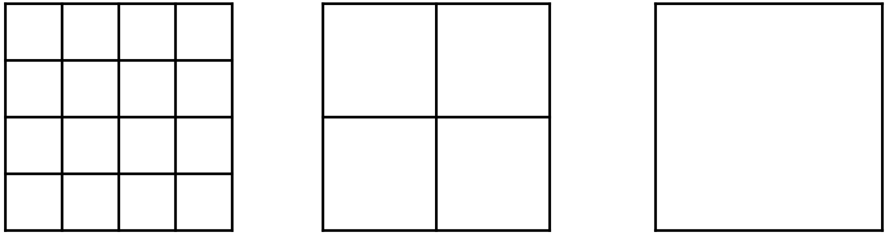
- Simulated data on neighbourhood level for Amsterdam
- Each dot represents a household
- Dots are placed in residential areas (OpenStreetMap) per neighbourhood



http://www.claircity.eu/

Super Dots

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

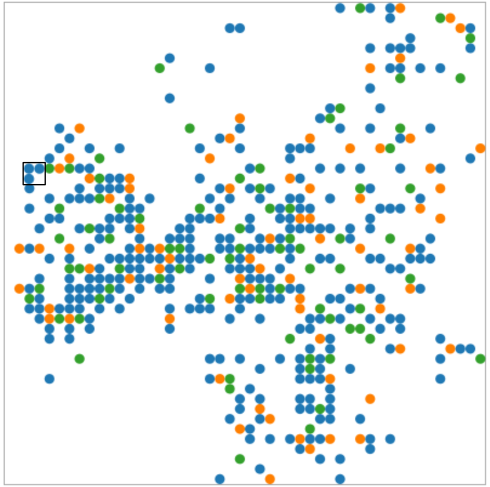


original

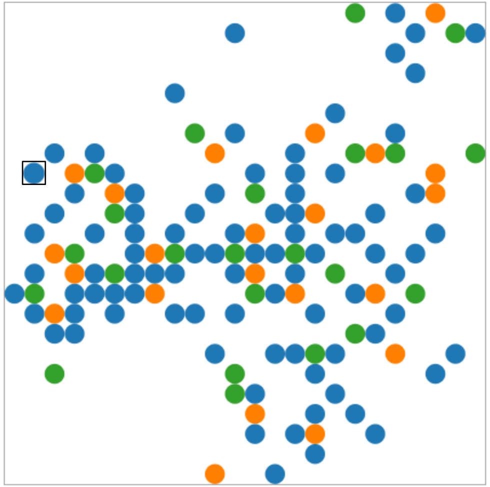
k = 2

k = 4

Example:

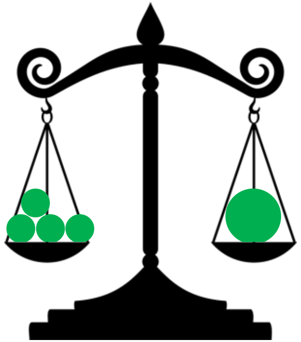


original dot map



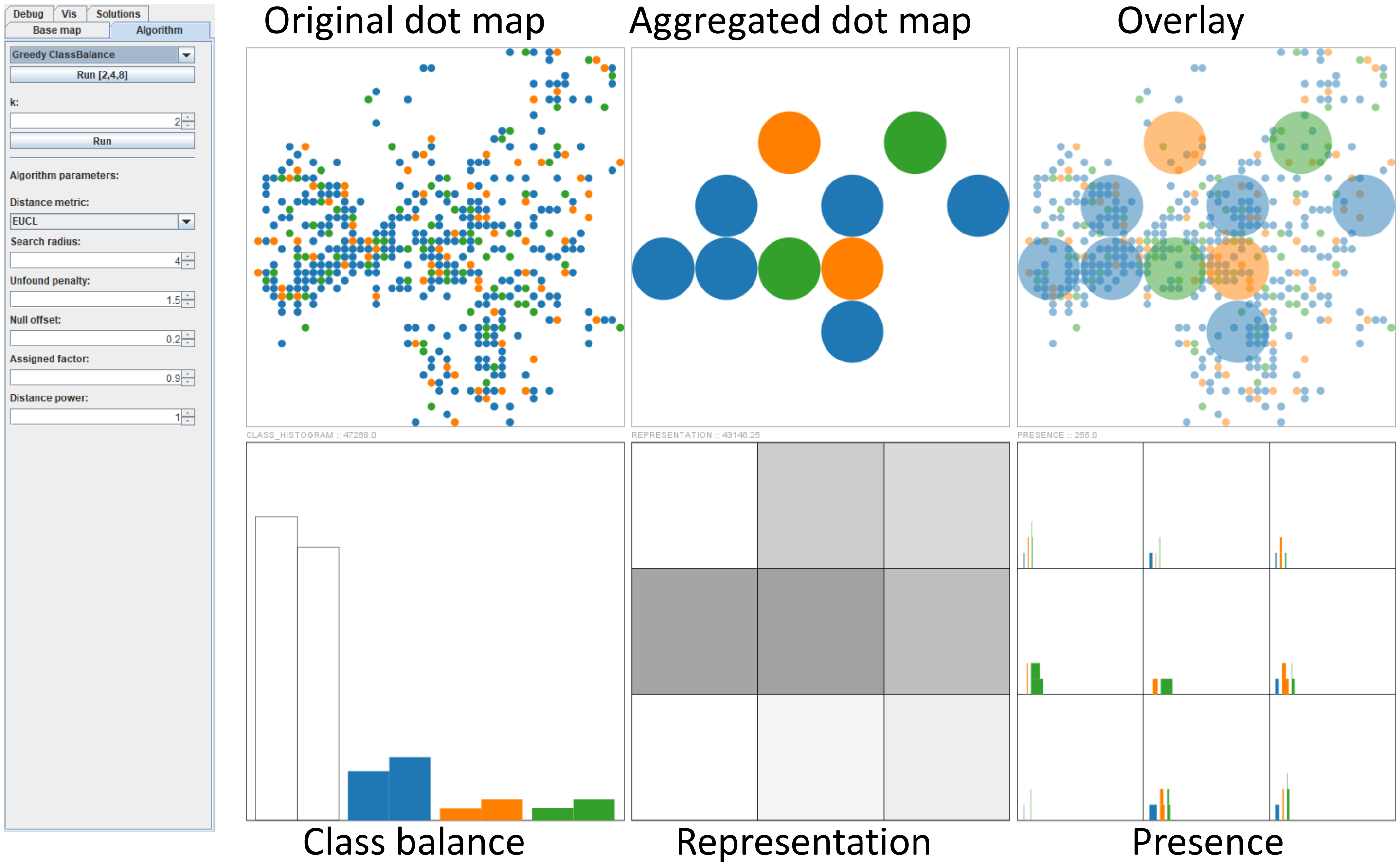
aggregated dot map (k = 2)

What is a good aggregation?

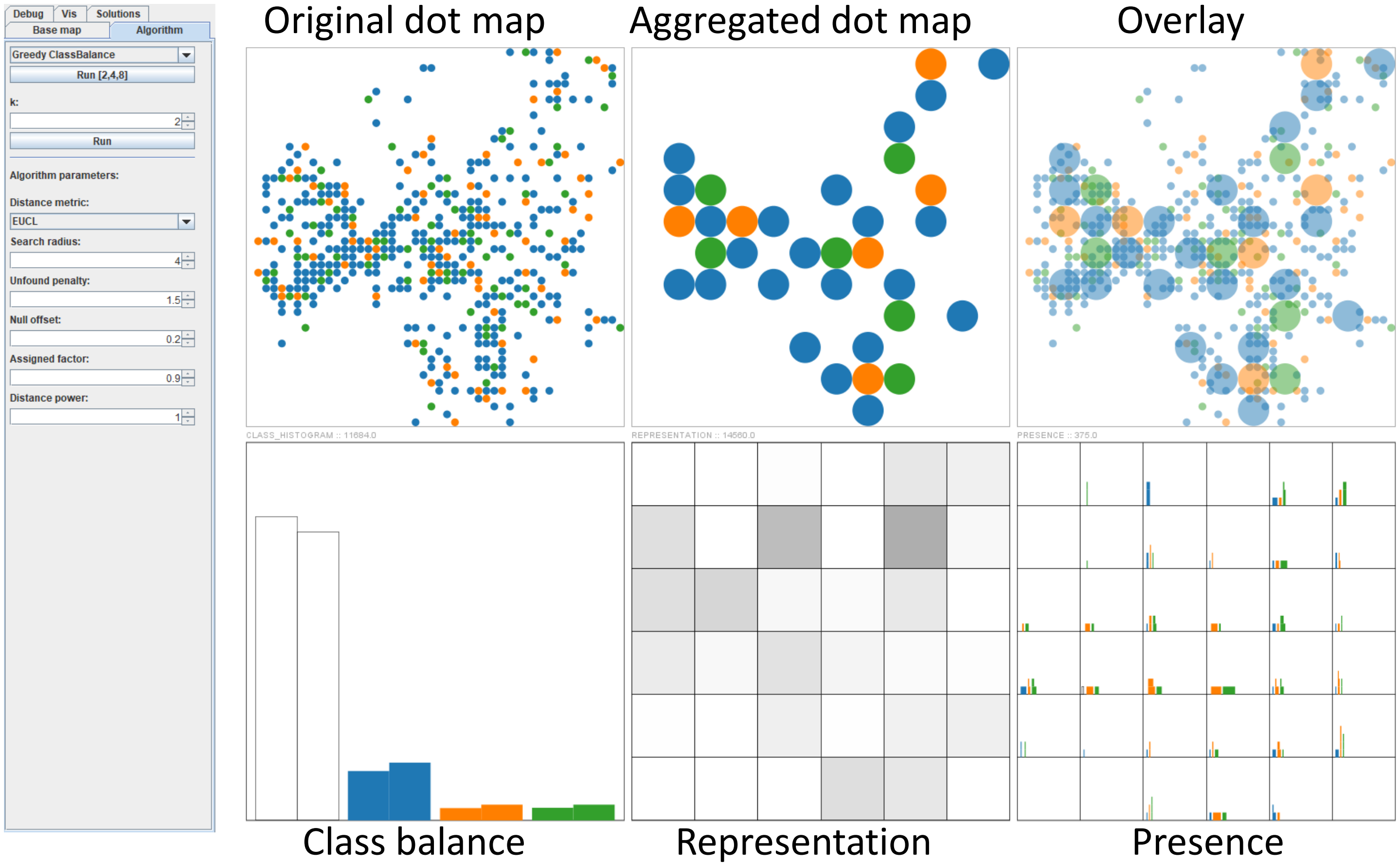


- **Class Balance** Total number of super dots per class should represent the total number of small dots per class
- **Representation** How well do the super dots represent the small dots? Each small dot is represented at most once, and each super dot can represent at most k^2 small dots.
- **Presence** How well are the small dots represented by the super dots? For each small dot, the distance to the nearest super dot is measured.

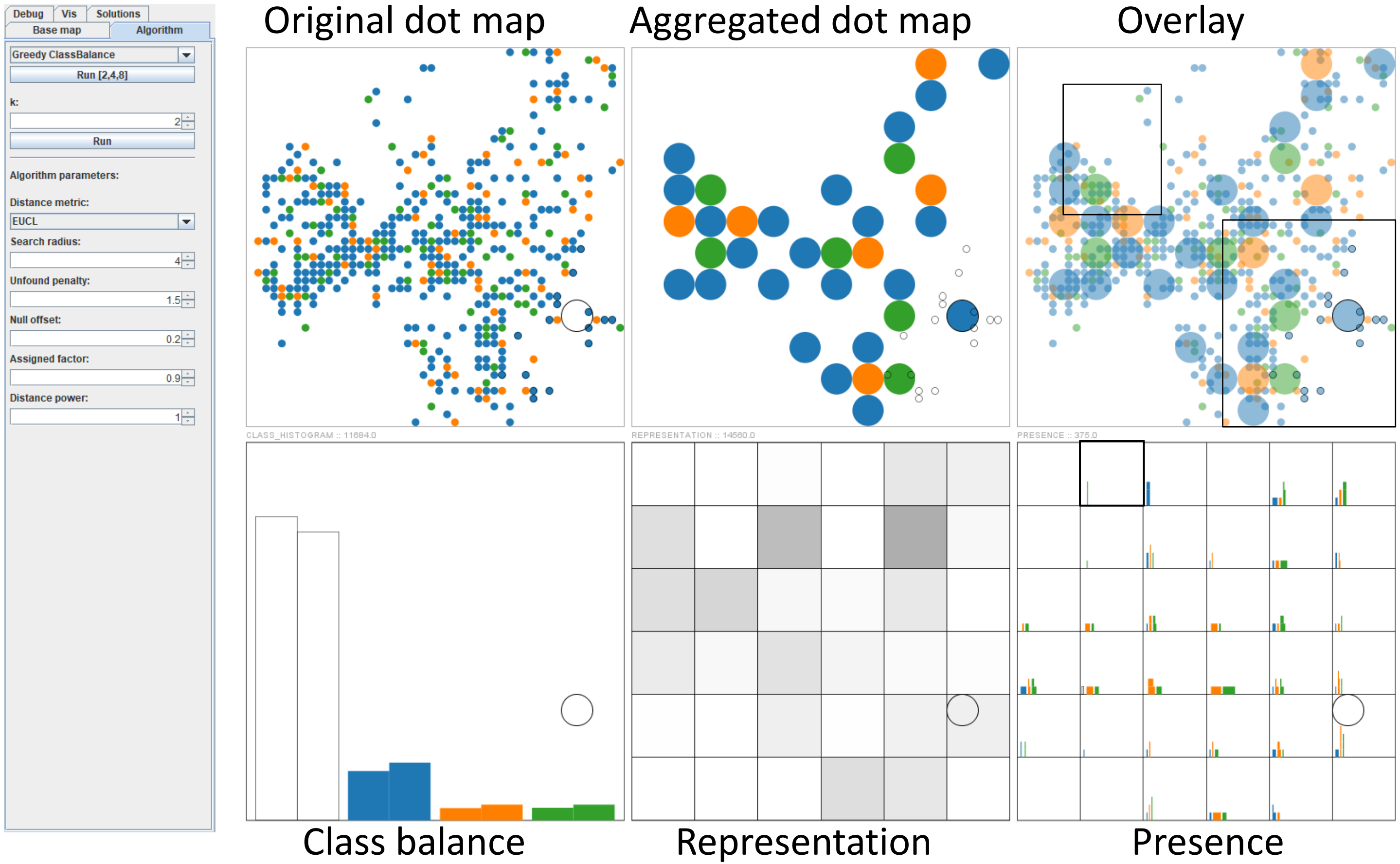
Aggregation analyses tool



Aggregation analyses tool



Aggregation analyses tool



Algorithms (sketches)

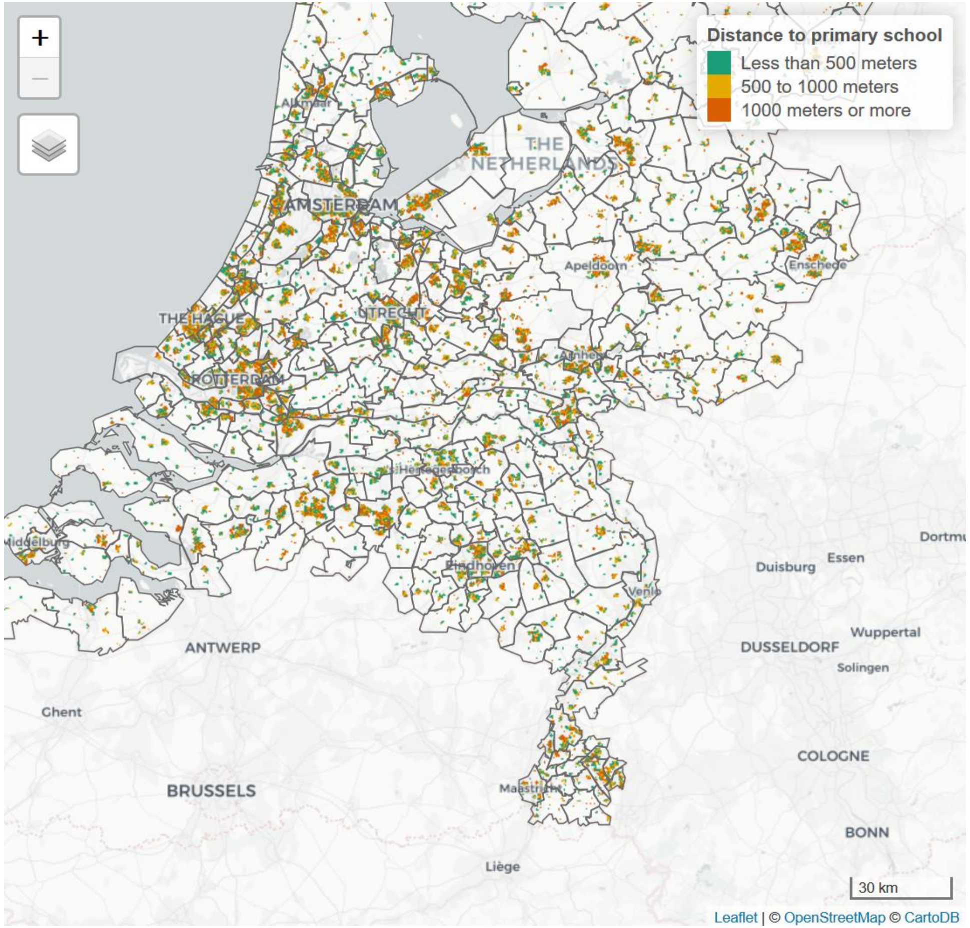
Greedy Class Balance Algorithm

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.

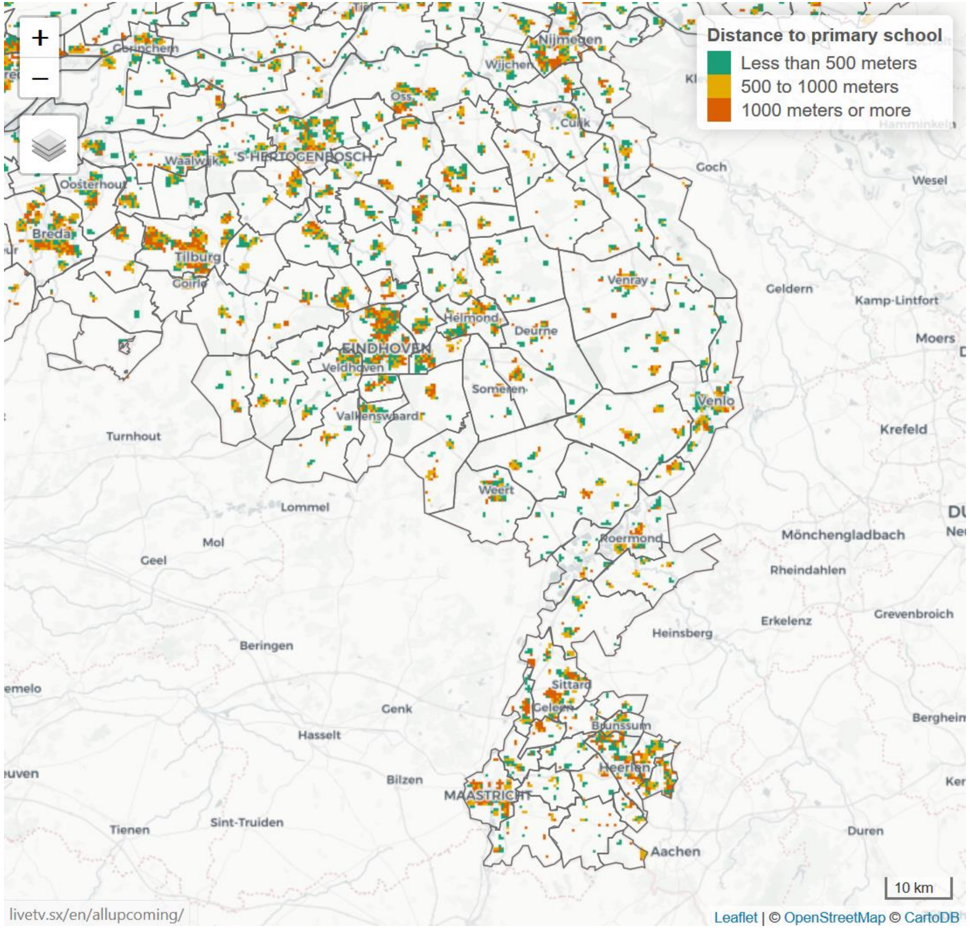
Application



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)

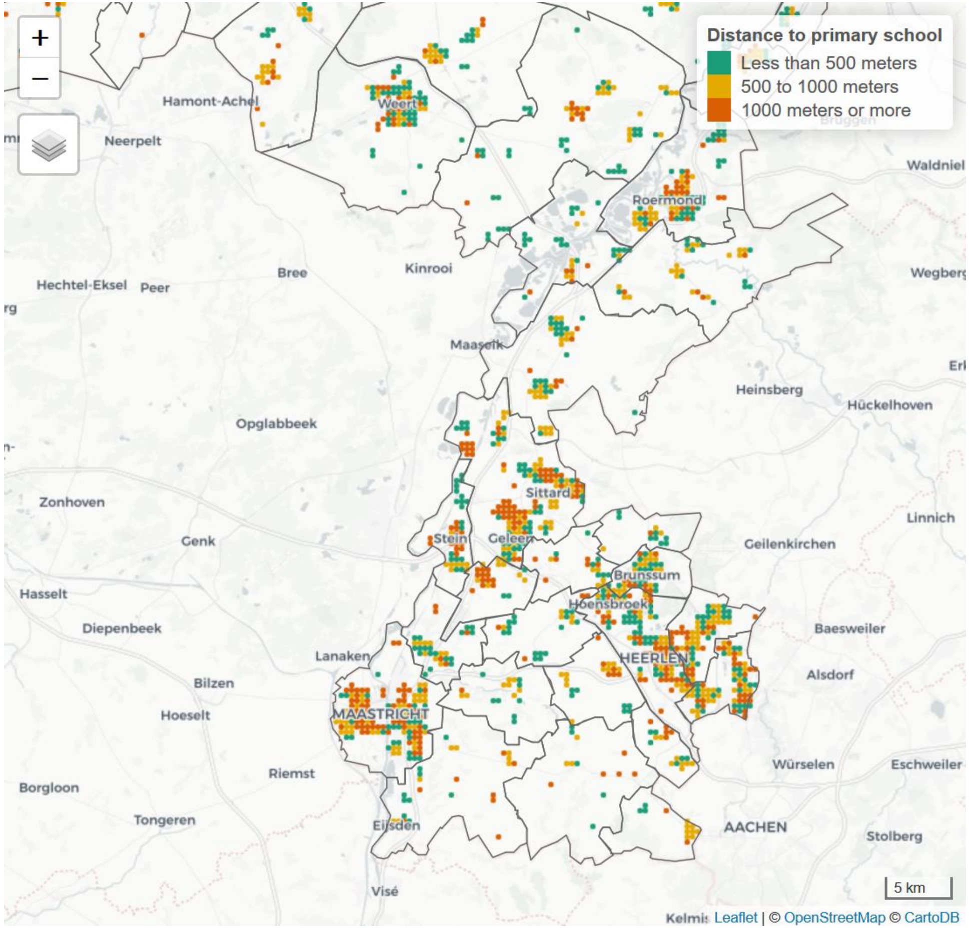
Application



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)

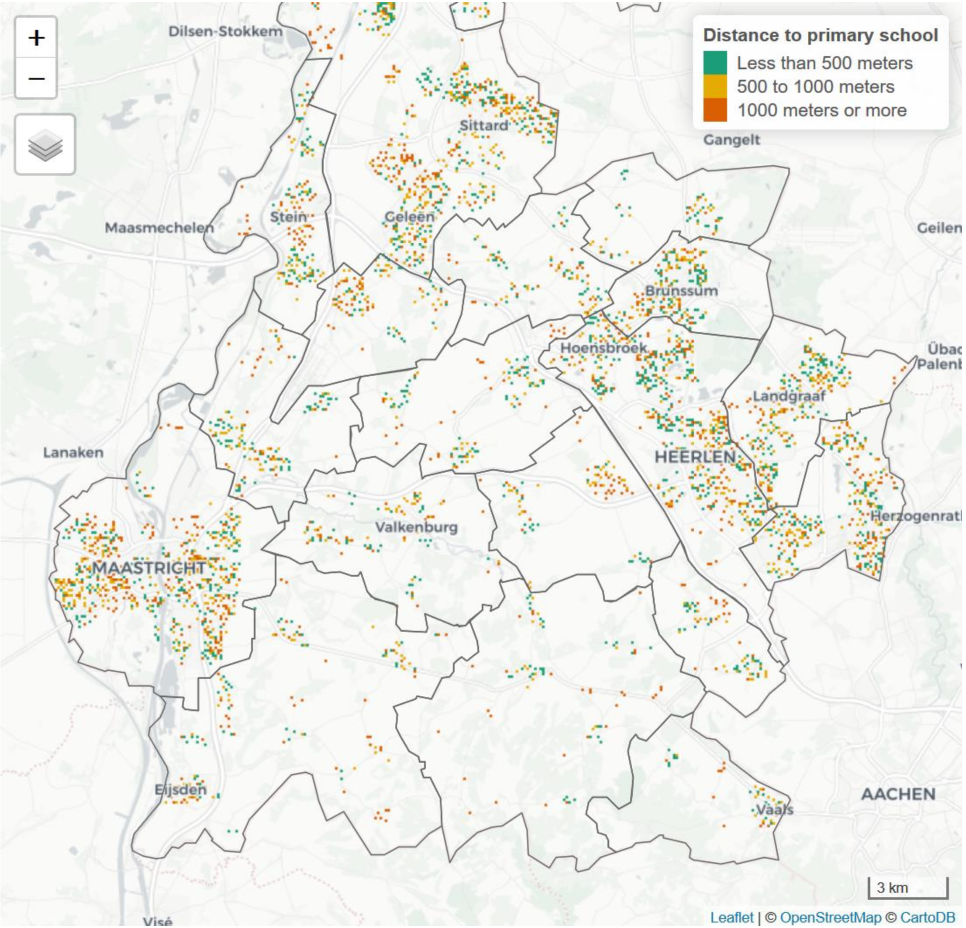
Application



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



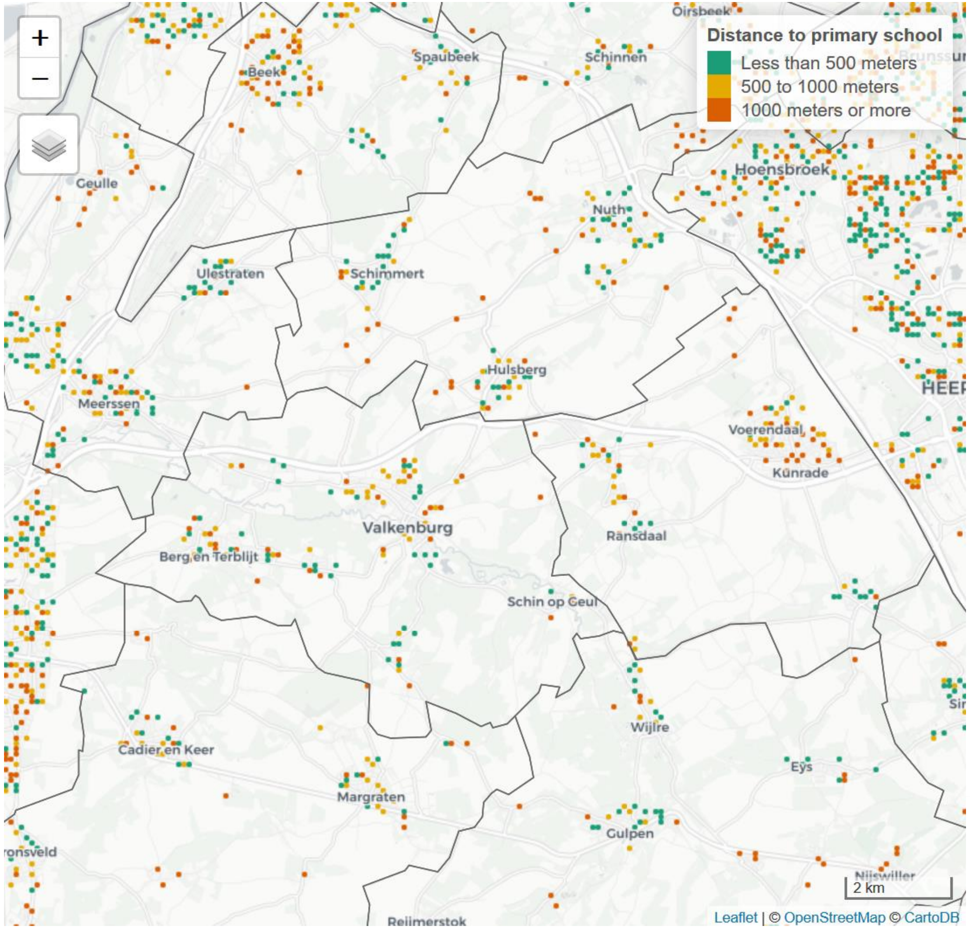
Distance to school

- 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

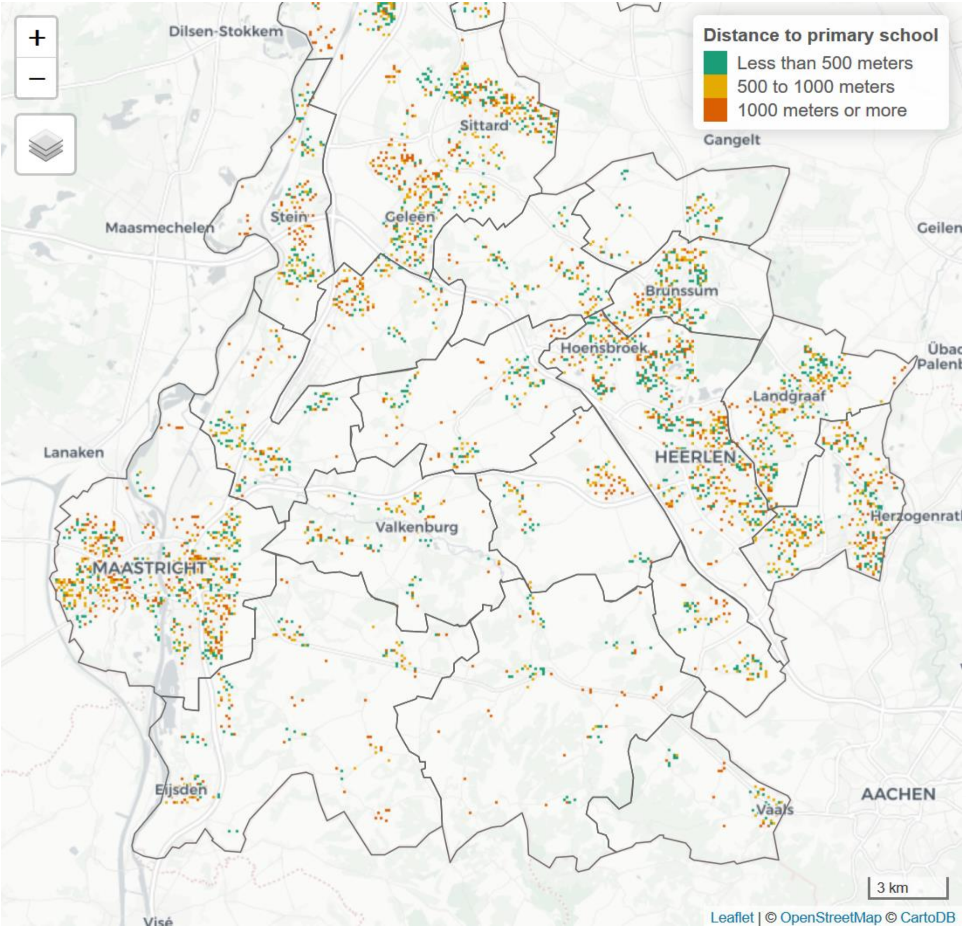
Distance to school

- 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)

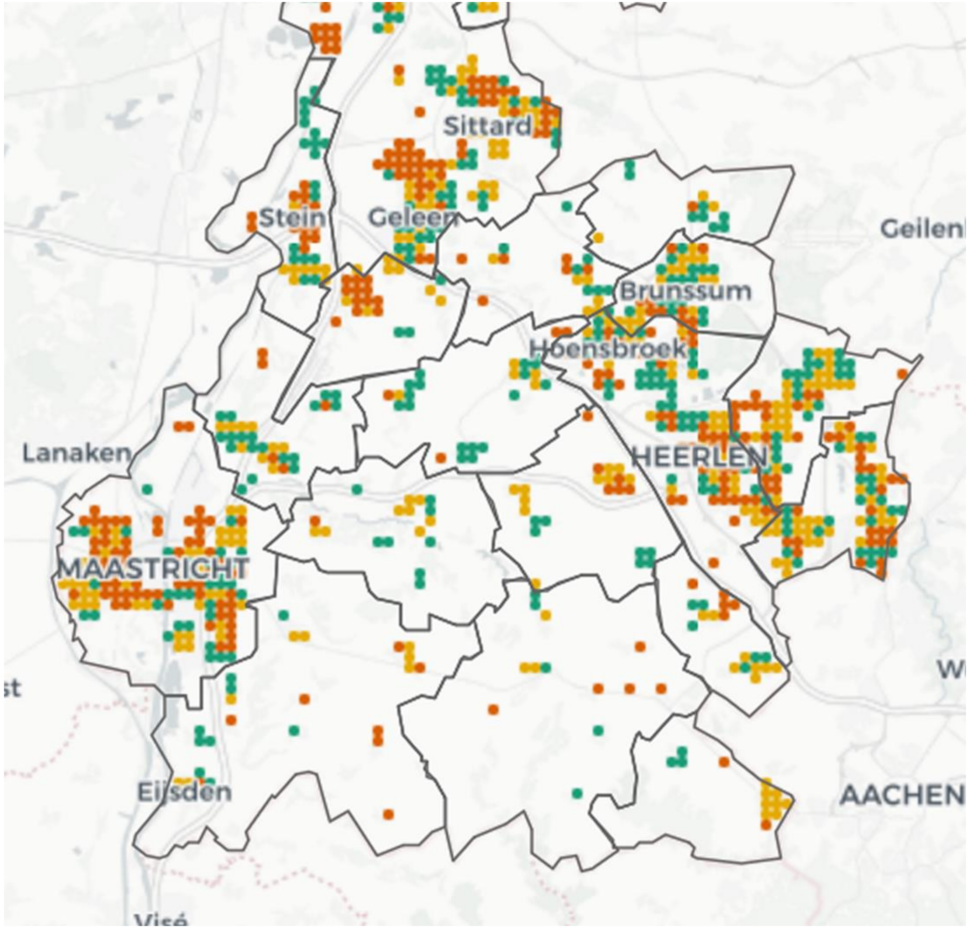
Application



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

Edwin de Jonge (CBS)

Wouter Meulemans
(TU Eindhoven)

Chantal Melser (CBS)

François Engelen
(Hogeschool Zuyd)