Utrecht University

Applied Data Science Master's degree programme

Spatial Data Analysis and Simulation Modelling course

Instruction manual for Lab 2.1: **Core concepts, data models and measurement levels**

Document version: 1.0

Document modified: 2020.10.14

Dr. Enkhbold Nyamsuren, Dr. Simon Scheider

Department of Human Geography and Spatial Planning

Faculty of Geosciences

s.scheider@uu.nl

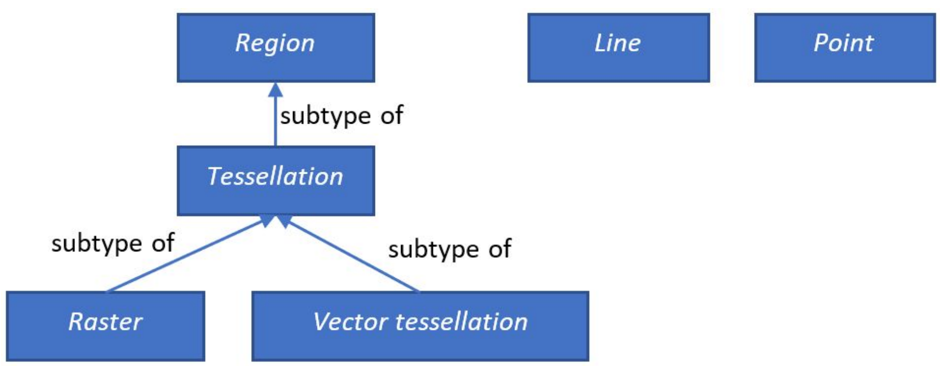
Table of Contents

Introduction

In this lab, you will explore core concepts and measurement levels based on concrete datasets. You should try loading and visualizing each example dataset in QGIS. While visualizing the datasets related to data entities, you should try to use a proper visual style similar to ones used in the example maps shown in this manual. You should include the screenshots of your visualizations in the "Submission template for Lab 2.1" document *provided on blackboard*. At the end of this lab, you should submit this document.

Data geometry

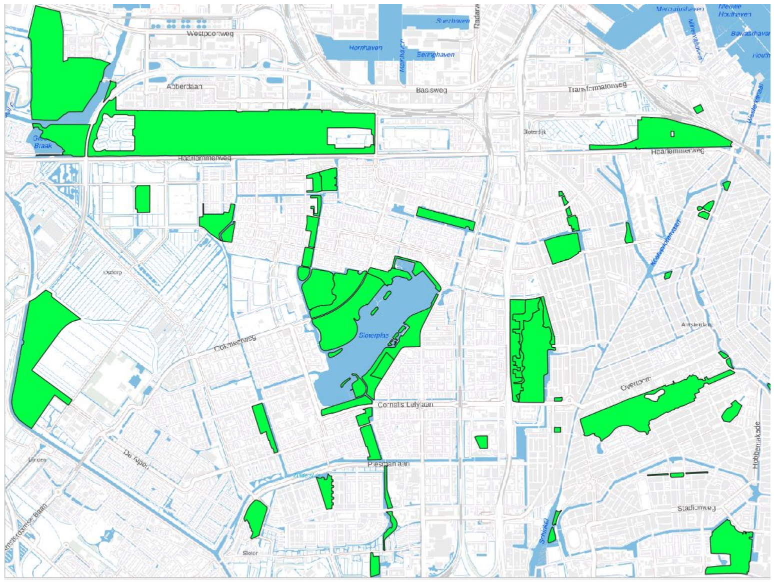
Any geodataset can be annotated with respect to the type of geometry used in the dataset and the types of entities these geometries represent (their semantics). Let's start with geometry.
We distinguish six types of geometries shown in picture below: *Region*, *Tessellation*, *Raster*, *Vector tessellation*, *Line*, and *Point*. Some of these types are organized in a supertype/subtype hierarchy. For example, *Region* is a supertype of *Tessellation*. This means that any *Tessellation* is also *Region*, but not every *Region* is *Tessellation*. There are also mutually exclusive types. For example, *Region* cannot also be *Line* or *Point* and vice versa.



Let's discuss the geometry types with some more details based on actual datasets that can be found in the Amsterdam open geodata portal (https://maps.amsterdam.nl/open_geodata/?LANG=en) and PDOK (https://www.pdok.nl/datasets). We provide download links for each dataset mentioned in this manual. You should try loading and visualizing each dataset in QGIS. It is recommended to use GeoJSON distributions in the Amsterdam portal and WFS/WMS servers in PDOK to download the relevant datasets. Alternatively, *all datasets are also accessible together with this lab manual on blackboard*.

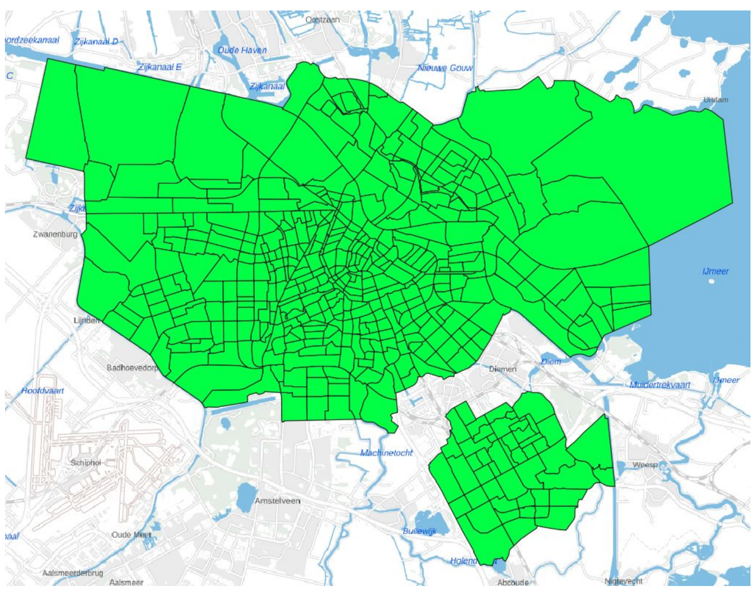
Region.

A dataset where the geometric primitives are regions (polygons or cells). Below is example of *Region* geometry that is not *Tessellation*. The map depicts parks and green areas in Amsterdam (https://maps.amsterdam.nl/open_geodata/?k=99).



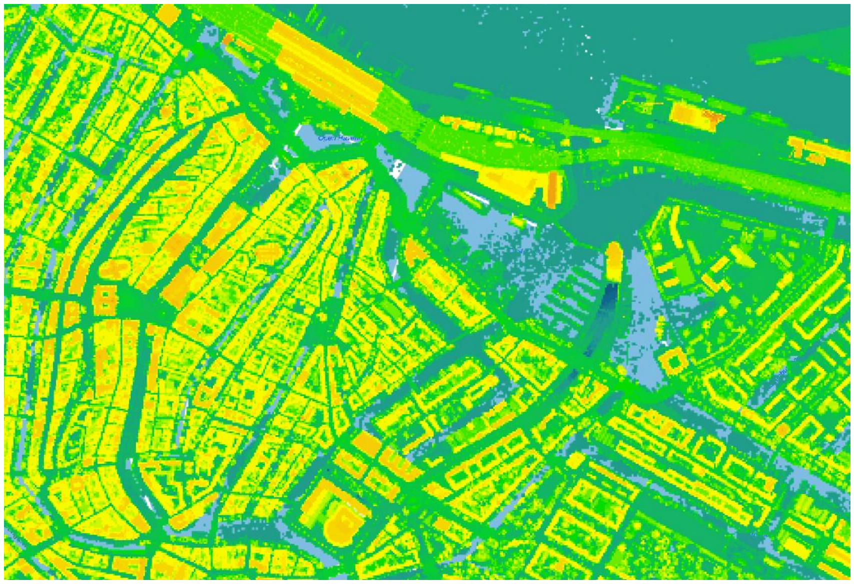
Tessellation.

A specific form of *Region*. Regions are non-overlapping, and fill the entire extent of the dataset without gaps (for example, counties of England). A good example of tessellation is the dataset of Dutch provinces you worked with in Lab 1.1. Dataset of Amsterdam neighborhoods/buurten (https://maps.amsterdam.nl/open_geodata/?k=198) is also a tessellation.



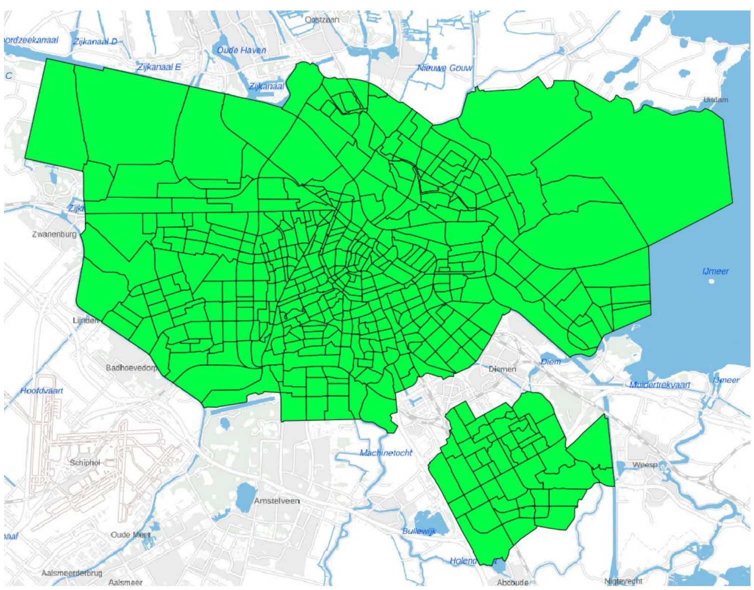
Raster.

A special kind of *Tessellation*, where the regions are all squares (cells). A typical example is satellite imagery. Another example is the digital elevation map of Amsterdam shown below. The dataset can be downloaded from PDOK (https://www.pdok.nl/geo-services/-/article/actueel-hoogtebestand-nederland-ahn1-). You need to use the *WMS* URI (instructions in Lab 1.1) and import the *ahn1_5m* layer to get the same map as below.



Vector Tessellation.

Any kind of *Tessellation* that is not a *Raster*. Below is an example of *Vector Tessellation* geometry. The map depicts Amsterdam neighborhoods/buurten (https://maps.amsterdam.nl/open_geodata/?k=198).

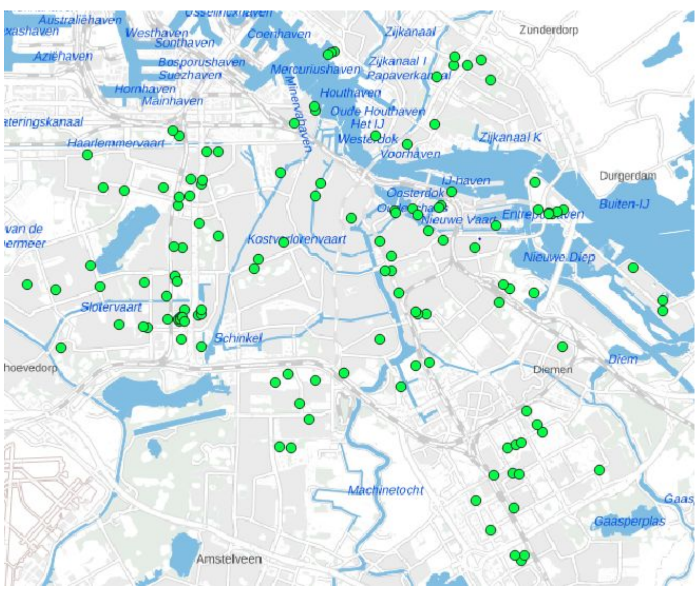


Line.

A dataset where the geometric primitives are lines. The map below depicts Amsterdam streets accessible to pedestrians (https://maps.amsterdam.nl/open_geodata/?k=298). Each street section is encoded as a line.

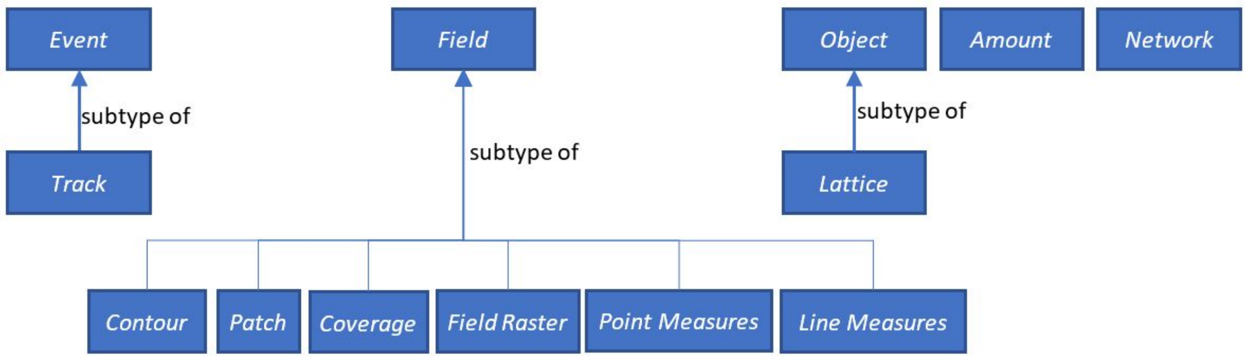
Point.

A dataset where the geometric primitives are points. For example, in Amsterdam geodata portal, youth and student housing in Amsterdam (https://maps.amsterdam.nl/open_geodata/?k=115) are depicted with points as shown below.



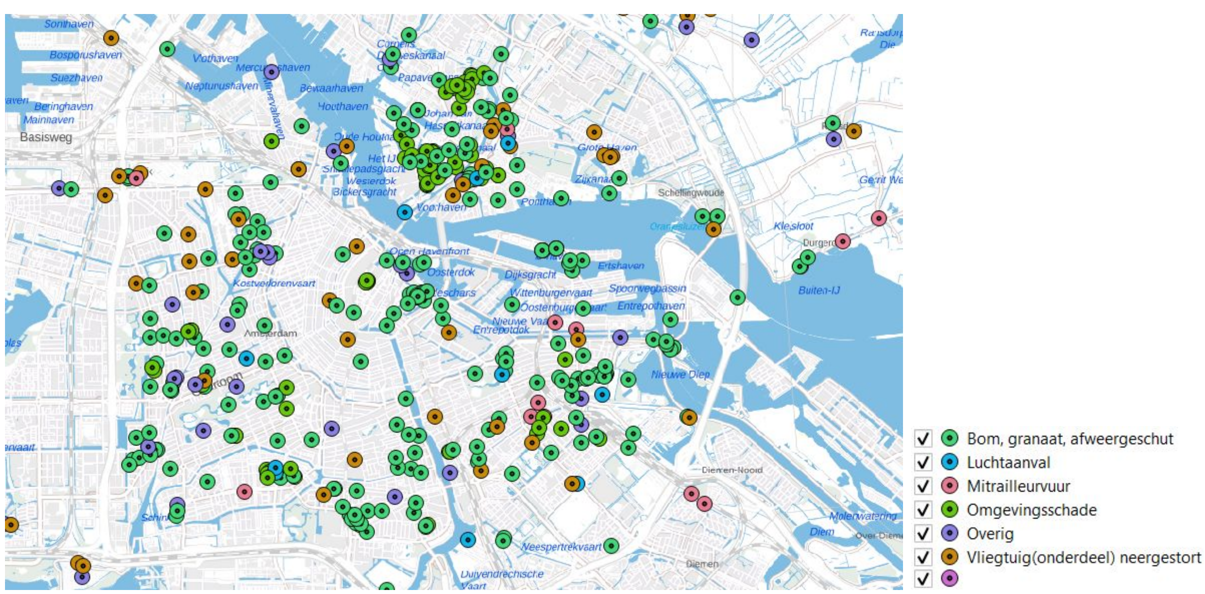
Data entity

It is important to note that a dataset can (and often should!) be annotated with multiple types. In particular, at least one semantic type and one geometry type are needed for each dataset. For example, *Point* and *Point Measures* are used in the case of a point-measured temperature field dataset. **It can also be the case that more than one entity type applies**. For example, road network dataset may be regarded both as *Network* and *Object*. Similar to the geometry types, we review each entity datatype based on actual datasets that can be found in the Amsterdam open geodata portal (https://maps.amsterdam.nl/open_geodata/?LANG=en), PDOK (https://www.pdok.nl/datasets), and Movebank portal (https://www.movebank.org/cms/webapp?gwt_fragment=page=search_map). We provide download links for each dataset mentioned in this manual. You should try loading and visualizing each dataset in QGIS. While visualizing the datasets, you should try to use a proper visual style similar to ones used in the example maps shown in this manual. It is recommended to use GeoJSON distributions in the Amsterdam portal and WFS/WMS servers in PDOK to download the relevant datasets. Alternatively, all datasets are also accessible *on blackboard*.



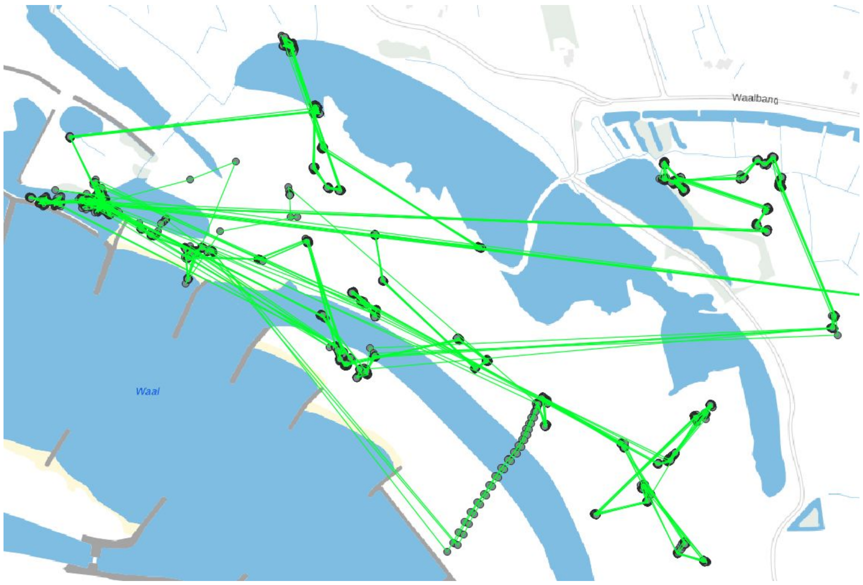
Event.

An event is a thing which happens in time. Events therefore have a start and end time as well as a duration, even if this is not explicitly represented in a dataset. *Event* denotes a dataset **where each data item can be regarded as an event** in this sense. For example: a map layer depicting occurrences of rhinos in South Africa, or house sales in London, or earthquakes in Iceland, or building construction phases in Amsterdam. The map below depicts the categorized events of attacks on Amsterdam during World War 2 (*Soort* attribute, https://maps.amsterdam.nl/open_geodata/?k=100).



Track.

It is a subtype of *Event*. In a *Track* dataset, data items are snapshots of a trajectory of some moving object. For example: a dataset of tracks of birds, where each point is a position of a bird. The map below shows the flight paths of a goose family during a migration (https://www.movebank.org/cms/webapp?gwt_fragment=page=studies,path=study1049685237). Note that the map below actually depicts two datasets: a dataset of track points and a dataset of lines connecting the track points.

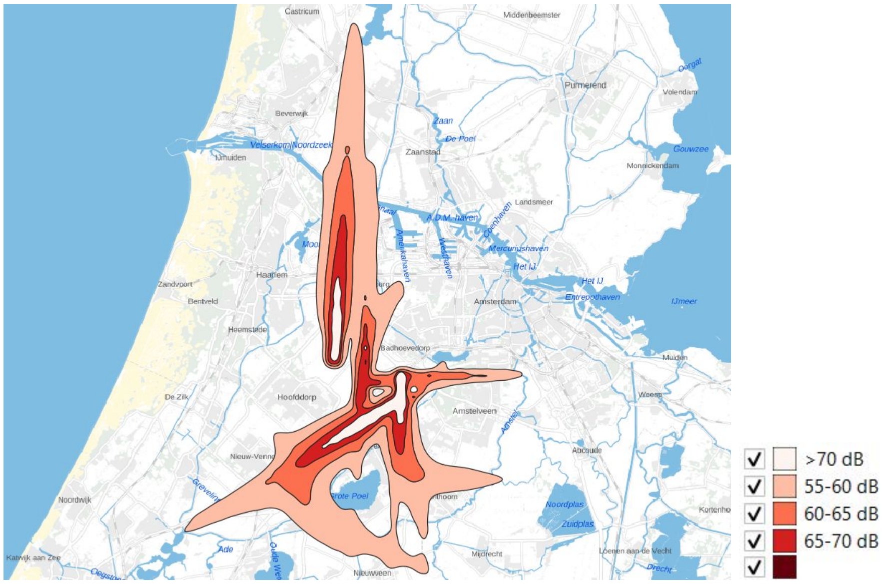


Field.

A field is a continuous surface of values in space. For example: Air pressure. Air temperature. Noise or acoustic pressure. Reflection of light from the earth surface. *Field* is a dataset **that (as a whole) represents an excerpt of a field as a layer**. We will explore specific subtypes of *Field* with concrete example datasets.

Contour.

Contour is a subtype of *Field*. A dataset that represents *Field* by contours. Each contour encloses field values that lie inside a given interval. For example, a terrain contour map or an air pressure map in meteorology. In the image below, there is a contour map of noise levels caused by air traffic around Amsterdam (*KLASSE* attribute, https://maps.amsterdam.nl/open_geodata/?k=252).

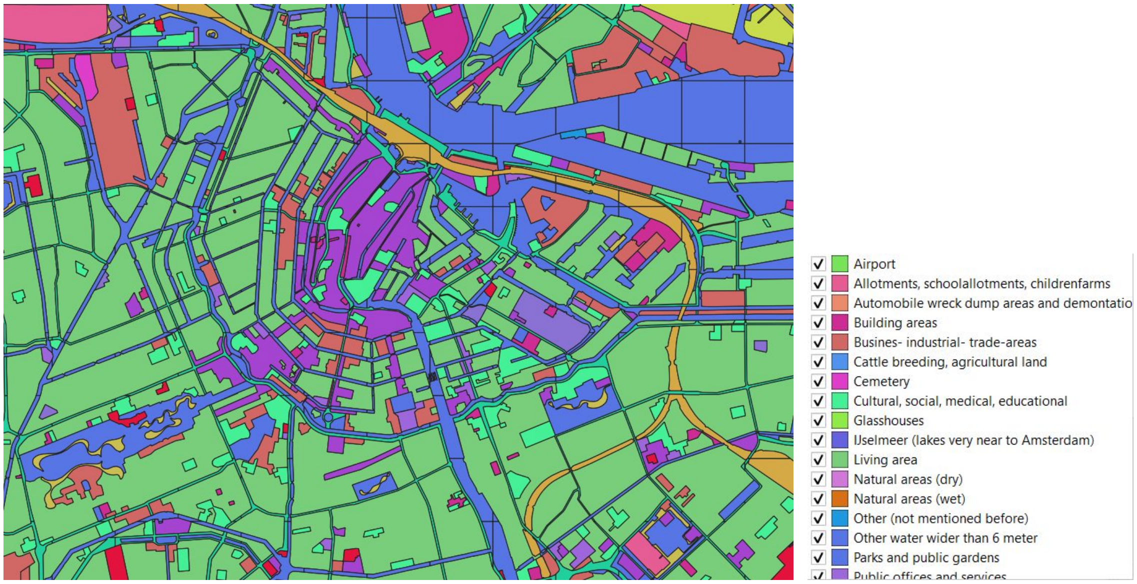


Patch.

Patch is a subtype of *Field*. *Patch* is a region dataset that represents irregular patches of a field containing homogeneous field values. For example, the map below visualizes the areas of Amsterdam parks with corresponding levels of restrictions (Off-leash, restricted, prohibited) for dogs (*Soort* attribute, https://maps.amsterdam.nl/open_geodata/?k=245).

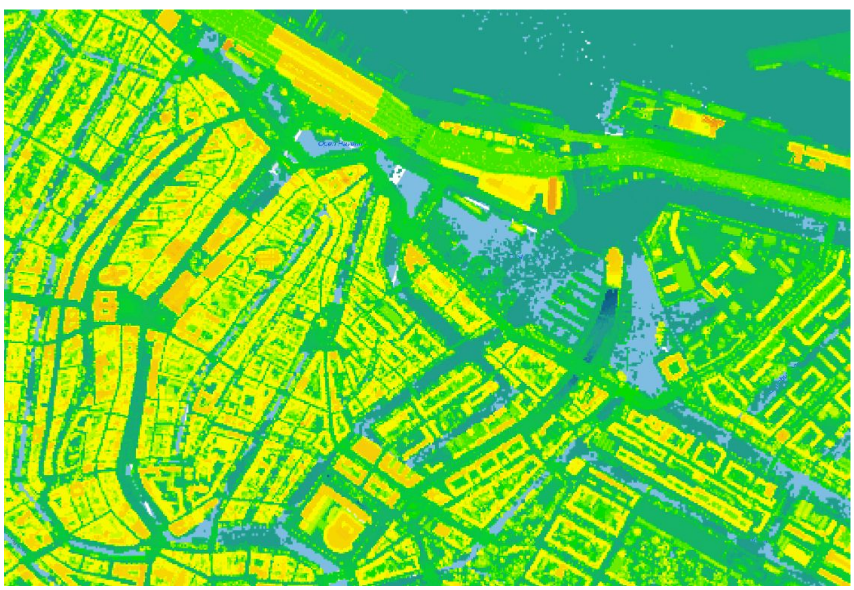
Coverage.

Coverage is a subtype of *Field*. Conceptually, *Coverage* is similar to *Patch*. The distinction from *Patch* is that *Coverage* is also a form of *Tessellation*. For example, a dataset describing soil types is *Coverage*. Another example of *Coverage* is the map shown below depicting the landuse types in Amsterdam (*CBScode2_OmsEng* attribute, https://maps.amsterdam.nl/open_geodata/?k=152).



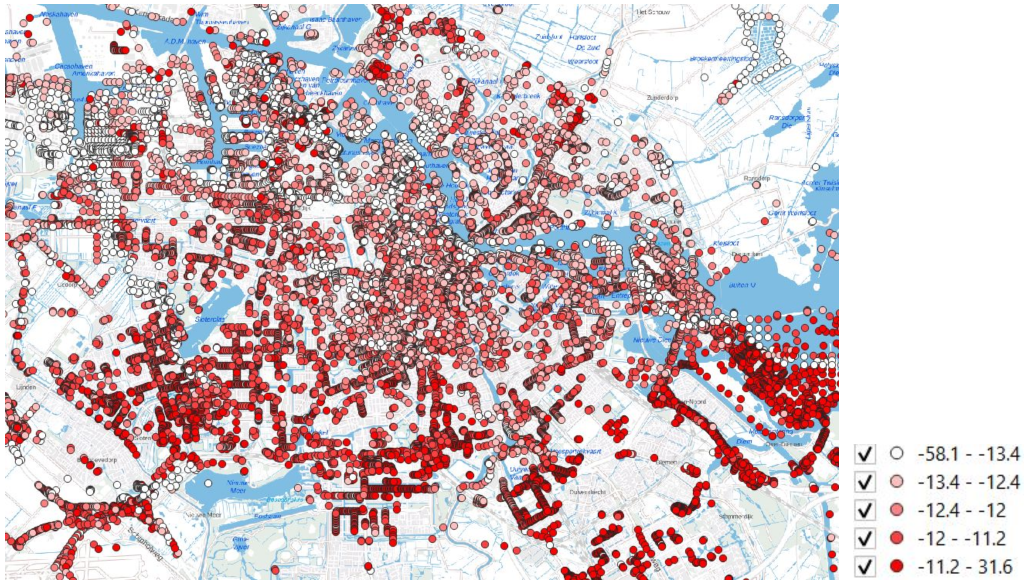
Field Raster.

Field Raster is a subtype of *Field*. A dataset representing a field in terms of raster cells. For example, a remote sensing image, where the represented field is light intensity, or a digital terrain model. The previously shown digital elevation map of Amsterdam is also an example of *Field Raster*.



Point Measures.

Point Measures is a subtype of *Field*. A dataset representing a field in terms of a sample of point-like measurements. For example, a record of temperature sensor measurements in a city. Another example in the map below are CPT measurements (Geology) to identify soil's load-bearing capacity in Amsterdam (*Top_Pleistoceen* attribute, https://maps.amsterdam.nl/open_geodata/?k=243).



Line Measures.

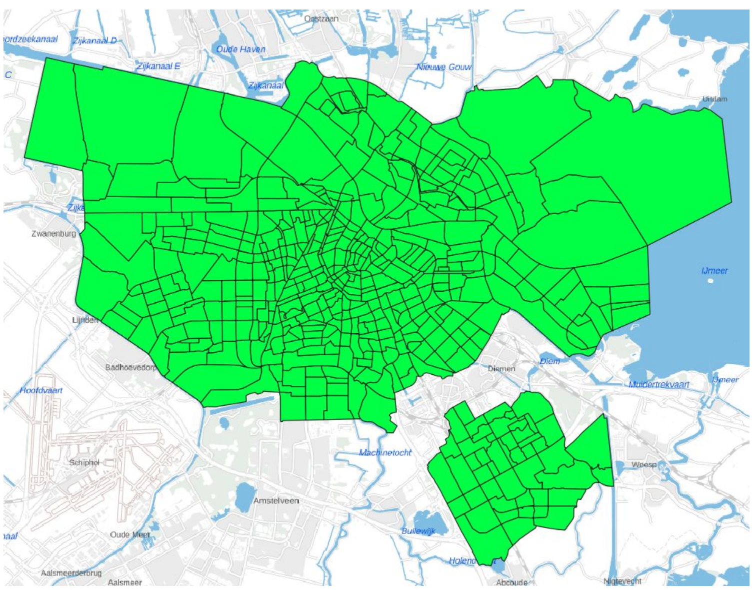
Line Measures is a subtype of *Field*. A dataset representing a field in terms of lines of homogeneous values. For example, *Line Measures* is a coastline, which is a line depicting sea level height as shown in the map below (dark blue line). This dataset is from PDOK (https://www.pdok.nl/geo-services/-/article/zeegebieden), and can be downloaded via the WFS server by choosing the *SR.Coastline* layer.

Object.

Objects are things that have a name and are spatially bounded. Their locations are represented for a temporal snapshot. In contrast to events they don't have a start or end. An *Object* is a dataset **where each data item is a spatial object in this sense**. Compared to *Field*, *Object* is a whole and indivisible entity of its own. Attributes report about qualities of these objects, for example, statues in Groningen, or streets of London. The map below depicts as points the individual trees in Amsterdam (https://maps.amsterdam.nl/open_geodata/?k=254).

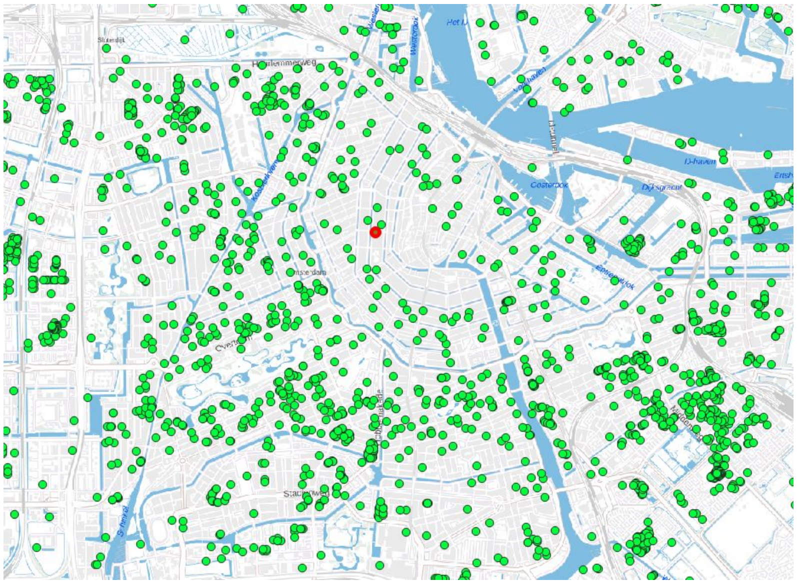
Lattice.

An *Object* which represents statistical areas forming a *Tessellation,* and attributes are spatial summaries. For example, official statistics of the municipalities of the Netherlands. The previously shown map of Amsterdam neighborhoods (https://maps.amsterdam.nl/open_geodata/?k=198) is an example of *Lattice*.



Amount.

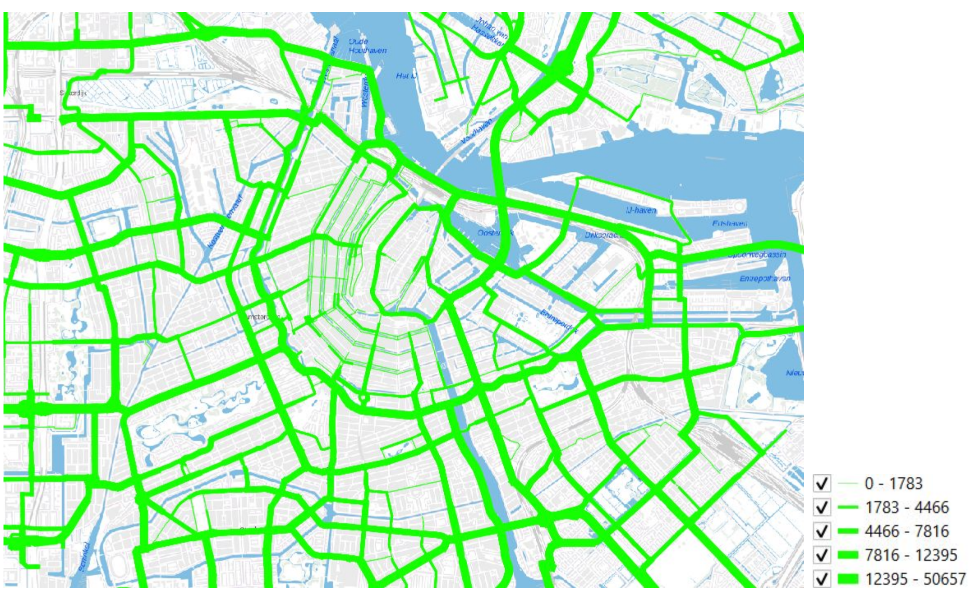
A layer where each data item represents an amount of objects or matter, but is itself **not** an object (and therefore the layer is not classified as *Object*). If this is a raster dataset, e.g., then each raster cell represents an amount of objects. For example, the amount of residents living in a cell, or the percentage of elderly. In the example below, each point depicts the number of solar panels (*Gedetecteerde_panelen* attribute) detected at the location and the amount of energy (*Vermogen* attribute) produced by these panels (https://maps.amsterdam.nl/open_geodata/?k=251).



Functie	NietWonen
Gedetecteerde_panelen	4
Vermogen	1000

Network.

Network is a quantified relation between pairs of spatial objects, for example, traffic flow between two road intersections during rush hour, or commuter flow between municipalities of the Netherlands, or migration flow between cities. *Network* is a dataset that represents the measures of these pairs, either in terms of a linked list or a matrix or an object quality. The map below depicts traffic flows on Amsterdam roads during a 24-hour weekday (*etmaal* attribute, https://maps.amsterdam.nl/open_geodata/?k=253).



Measurement scales

We differentiate between four measurement scales: *Nominal*, *Ordinal*, *Interval*, and *Ratio*. Additionally, we can consider *Count* as a special case of *Ratio* scale.

Scales of measurement with examples			
Measurement scale	**Value type**	**Example attribute**	**Example unit of measurement**
Nominal	String	Ski resort name	Text
Ordinal	String, number	Resort popularity ranking	Rank number
Interval	Number	Average winter temperature	Celsius
Ratio	Unsigned number	Size of ski area	m^2
Count	Unsigned integer	Number of ski slopes	Count number

Nominal.

On a nominal scale, numbers are used to establish identity. For example, numbers on a nominal scale include telephone numbers or ski passcodes. These numbers cannot be processed in a mathematical sense since they do not represent order or relative value. Adding, subtracting or dividing numbers on a nominal scale will not produce a useful result. Adding together two phone numbers is possible, but the answer is meaningless.

Ordinal.

The numbers in an ordinal scale establish order. Location in a ski lift queue is an example. For example, the ordinal scale is used to publish the top 10 cafés and ski runs based on the number of people using them each week. Using an ordinal scale, you can obtain an impression of the order of numbers, but no information about relative sizes. The most popular ski run (ranked 1) is not necessarily twice as popular as the ski run which is ranked second. Arithmetic operations, whilst possible on ordinal data, will again give meaningless results. It is also possible to have string values in an ordinal scale. As an example, let’s consider the 5-point Likert scale with values between [1, 5] (note that the Likert scale is always ordinal). However, it is often the case that such Likert scale can have string values instead, for example, *Fully Disagree*, *Disagree*, *Neutral*, *Agree*, and *Fully Agree*.

Interval.

On an interval scale, the difference between numbers is meaningful but the scale does not have a real origin. Temperatures, in degrees Celsius, are a good example. On a temperature scale it is possible to say that there is a 10-degree difference between a thermometer that records a value of 10 degrees and one that records a value of 20 degrees. Thus, differences can be calculated. However, it would be incorrect to say that 20 degrees are twice as warm as 10 degrees, because zero degrees on the Celsius scale is not a true zero. There is still a temperature when the thermometer reads zero! Negative numbers are also possible on an interval scale.

Ratio.

On a ratio scale, measurements can have an absolute/ real zero, and the difference between the numbers is significant. Snow depth is an example. It is impossible to have a negative value for snow depth. Something is also known about relationships between data, for example, a snowpack that is 3 m deep is twice as deep as one that is 1.5 m deep.

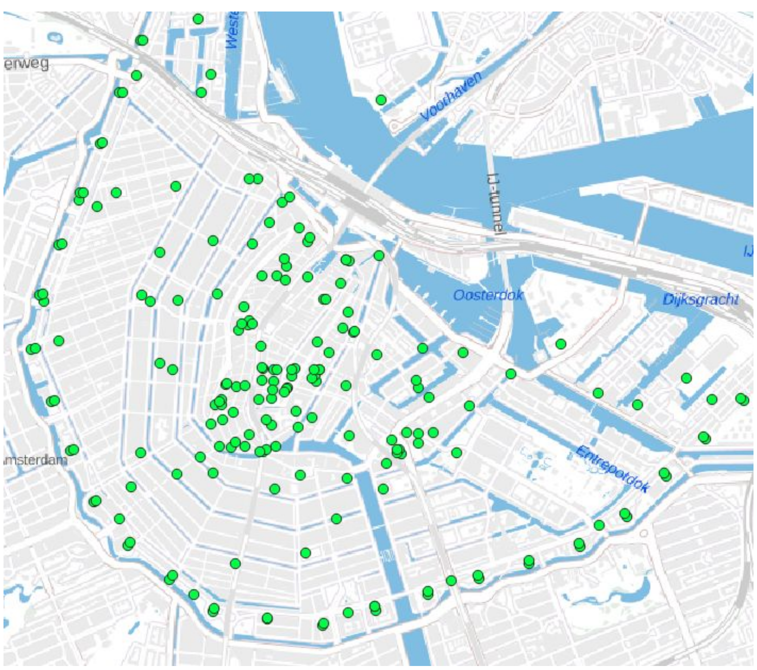
Count.

Finally, we can consider count as a special case of the ratio scale, where values are integer numbers representing count of something. For example, the number of ski slopes in a ski resort will be count. Since we are counting individual indivisible objects, it does not make sense to have a fractional number. Hence, we assume that the scale contains only integer numbers equal to or above zero.

Example annotation measurement scales of real datasets

In this section, we will make an example annotation of measurement scales of attributes in a dataset about historical buildings of Amsterdam city. The dataset can be found on the open geodata portal of the municipality of Amsterdam: https://maps.amsterdam.nl/open_geodata/?k=118.

1. Download the JSON distribution of the dataset.
2. Open the downloaded JSON file with QGIS as a map layer. The resulting map should represent a point dataset similar to one below:



3. Open the attribute table for the dataset by right-clicking on the layer and selecting the *Open Attribute Table* option from the popup menu. The attribute table should look like below:

HISTORISCHE_BEBOUWING :: Features Total: 200, Filtered: 200, Selected: 0

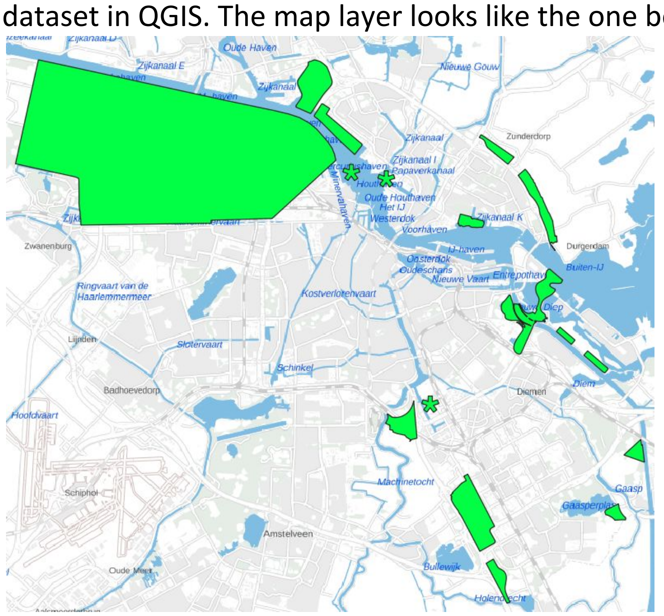
	Naam	Hoofdgroep	Functie	Foto	Website	Opmerking	Begin	Eind	SELECTIE
1	Armeense Kerk 1714	Religie	Kerk	Persiaanse kerk.jpg		Bron: Gerrit de ...	1714	0	RELIGIE
2	Oude Walenkerk 1586	Religie	Kerk	Walenkerk.jpg		Bron: Balthasar ...	1586	0	RELIGIE
3	Leidsepoort 1663-1862	Militair	Poort	Leidsepoort.jpg		Bron: Gerrit de ...	1663	1862	MILITAIR
4	Magazijn Admiraliteit 1579-...	Openbaar	Admiraliteit				1579	1632	OPENBAAR
5	Olofspoort gevangenis	Openbaar	Strafinrichting	SintOlofspoort.jpg		Bron: Cornelis A...	1425	1618	OPENBAAR

Show All Features

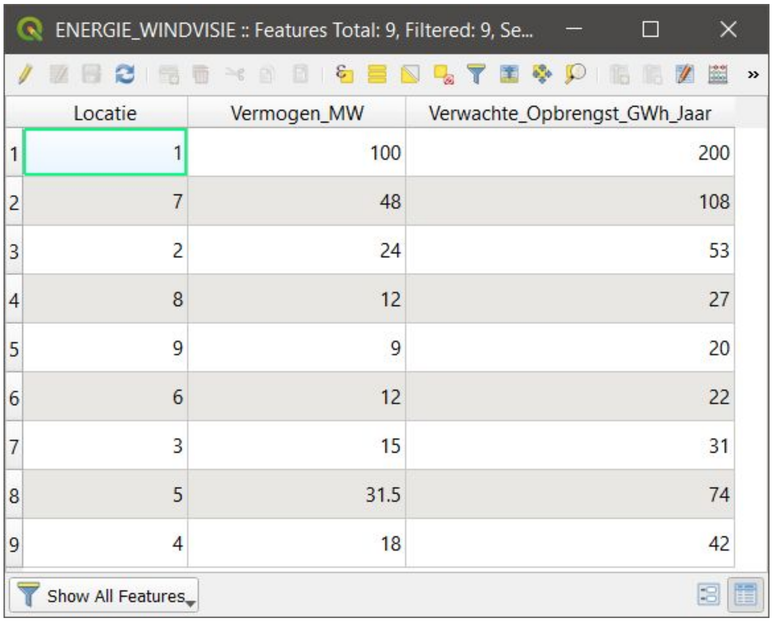
Now, we will annotate individual attributes in the dataset. Since we are mainly interested in annotating thematic data, we select the following columns for annotation: Naam, Hoofdgroep, Functie, Begin, and Eind. These annotations are shown in the table below. Naam, Hoofdgroep, and Functie are of nominal scale since they define dataset in terms of named categories and individuals. On the other hand, Begin and Eind are dates and, therefore, interval scales.

Column name	**Measurement scale**
Naam (Place name)	nominal
Hoofdgroep (Place main category)	nominal
Functie (Place function)	nominal
Begin (Construction start year)	interval
Eind (Construction end year)	Interval

Let's annotate the measurement scale of attributes in another dataset depicting potential areas for deploying wind turbines around Amsterdam city (https://maps.amsterdam.nl/open_geodata/?k=96). Similarly, download and visualize the dataset in QGIS. The map layer looks like the one below:



And the corresponding attribute table is as follows



	Locatie	Vermogen_MW	Verwachte_Opbrengst_GWh_Jaar
1	1	100	200
2	7	48	108
3	2	24	53
4	8	12	27
5	9	9	20
6	6	12	22
7	3	15	31
8	5	31.5	74
9	4	18	42

There are three attributes. While, based on its values, *Locatie* looks like belonging to ordinal scale, the scale is actually nominal. *Locatie* is the location id unique to each area. *Vermogen_MW* is power output measured in megawatt. Since power output cannot be below zero, the attribute has a ratio scale. *Verwachte_Opbrengst_GWh_Jaar* is expected yield measured as Gigawatt per year. It is technically possible for yield to be negative (if energy consumed is more than energy produced), therefore the attribute has an interval scale.

Column name	**Measurement scale**
Locatie (Location)	nominal
Vermogen_MW (Power, MW)	ratio
Verwachte_Opbrengst_GWh_Jaar (Expected yield, GWh/year)	interval

Assignment: Entity datatype and applicable GIS operations

Understanding what a dataset conceptually represents allows us to also understand what GIS operations can be meaningfully applied on the dataset. In this exercise, you are given four GIS operations below:

- Intersection: https://docs.qgis.org/3.10/en/docs/user_manual/processing_algs/qgis/vectoroverlay.html?highlight=#intersection
- Raster layer zonal statistics: https://docs.qgis.org/3.10/en/docs/user_manual/processing_algs/qgis/rasteranalysis.html?highlight=zonal%20statistics#raster-layer-zonal-statistics
- Zonal statistics: https://docs.qgis.org/3.10/en/docs/user_manual/processing_algs/qgis/rasteranalysis.html?highlight=zonal%20statistics#zonal-statistics

- Interpolation: https://docs.qgis.org/3.10/en/docs/gentle_gis_introduction/spatial_analysis_interpolation.html?highlight=interpolation

First study these operations in detail by reading the web descriptions. In the "Submission template for Lab 2.1" document, you can find a matrix of entities and operations similar to one shown below. In that matrix, put 'X' in a cell if the GIS operation can be meaningfully applied on the corresponding entity type.

	Intersection	Raster layer zonal statistics	Zonal statistics	Interpolation
Event				
Track				
Contour				
Patch				
Coverage				
Field Raster				
Point Measures				
Line Measures				
Object				
Lattice				
Amount				
Network				

For each entity type, please write some text arguing why you think an operation is or is not applicable. Submit this document on blackboard.

Optional exercises

These exercises are optional, and you don't have to submit anything for these exercises.

Exercise 1: Visualizing attributes of different measurements scales

QGIS support multiple visualization styles for attributes in a map layer. Among these styles are *Categorized* and *Graduated* (note that gradation can be done based on either color or symbol size, see https://qgis.nl/2012/11/24/proportionele-symbolen-2/?lang=en). For point features, QGIS additionally supports *Point Cluster* and *Heatmap* styles.

1. For each measurement scale, discuss by providing concrete arguments whether or not the two visualization styles can used on the scale.
2. Apply *Categorized* and *Graduated* visualization styles on two datasets (one dataset per style) of your choice downloaded from either Amsterdam open geodata portal or PDOK.

Exercise 2: Annotating a dataset

Here is a dataset of properties of a housing association in Amsterdam: https://maps.amsterdam.nl/open_geodata/?k=208. You should annotate the data geometry and the data entities for this dataset. For each attribute in the dataset, annotate its measurement level.