



Utrecht University

Applied Data Science Master's degree programme

Spatial Data Analysis and Simulation Modelling

Instruction manual for Laboratory 3.1: **Overlay and aggregation**

Document version: 0.1

Document modified: 2020.11.22

Dr. Simon Scheider, Eric Top, Haiqi Xu

Department of Human Geography and Spatial Planning

Faculty of Geosciences

s.scheider@uu.nl, h.xu1@uu.nl

Table of Contents

Vector overlay and Map algebra	3
Set up QGIS analysis environment	4
Coordinate Reference System	4
Load data	5
Task 1 Simple area interpolation for (re-)aggregating statistics	8
Generate new model	9
Overlay: Intersection	10
Calculate the product of area and the statistics attribute	11
Aggregate region statistics	12
Task 2: Map algebra for aggregating landuse areas	15
Convert landuse polygons to raster and select park areas	16
Aggregate park areas into PC4 by zonal map algebra	18

Vector overlay and Map algebra

In Lab 3.1, you will learn how to apply diverse raster-based and vector-based GIS methods in order to summarize data into statistical areas. This will allow you to assess the spatial quality of an urban area. All techniques are implemented as executable workflows using QGIS Model Designer.

Suppose you are asked by the municipality of Amsterdam to analyze the suitability of the city for older people on the spatial level of postcode areas (level 4). You are given a set of urban geodata sources which is openly available at the open geodata portal of the Amsterdam data lab:

https://maps.amsterdam.nl/open_geodata

To save time for the practical, these data sources are already packaged in terms of a set of shapefiles in the appendix of this manual (*dataAmsterdam*).

Your task in this assignment is to compute the following characteristics on the level of postcode areas of level 4:

1. People want to meet peers of their same age in their neighbourhood. Where in Amsterdam are the PC4 areas with a *high percentage* of old people?
2. People like to have extended parks in their neighbourhood that allows them to take a walk. Where in Amsterdam are the PC4 areas with a *large density* of parks and green?

The tasks are done by *vector overlay (area interpolation)* and by *map algebra analysis* using rasters. Through these practicals, you will acquire the following *GIS competences*:

- Generate *workflows* in Model Designer and execute them
- Learn how to make use of *Vector overlay techniques*
- Learn how to make use of Map Algebra for raster processing
- Learn *basic analytical workflows* that can be reused for many similar tasks (compare lecture):
 1. Areal interpolation of regional statistics (re-aggregate region data into overlapping regions)
 2. Aggregate continuous ("field") data into spatial regions

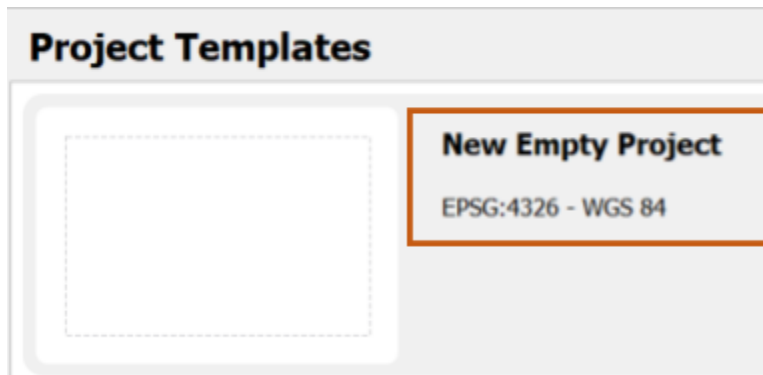
In particular, the following QGIS tools are used and practiced:

- Setting the geoprocessing environment
- General Vector tools
 - "Graduated Color" (Choropleth mapping)
 - "Add Field"
 - "Field calculator"
- Vector Overlay tools
 - "Intersection"
 - "Aggregate"
- Map algebra (raster) tools
 - "Rasterize (vector to raster)"
 - "Zonal Statistics"

Set up QGIS analysis environment

Start with downloading and unzipping “dataAmsterdam.zip” to some folder on a local drive, e.g. C:\Temp\dataAmsterdam\.

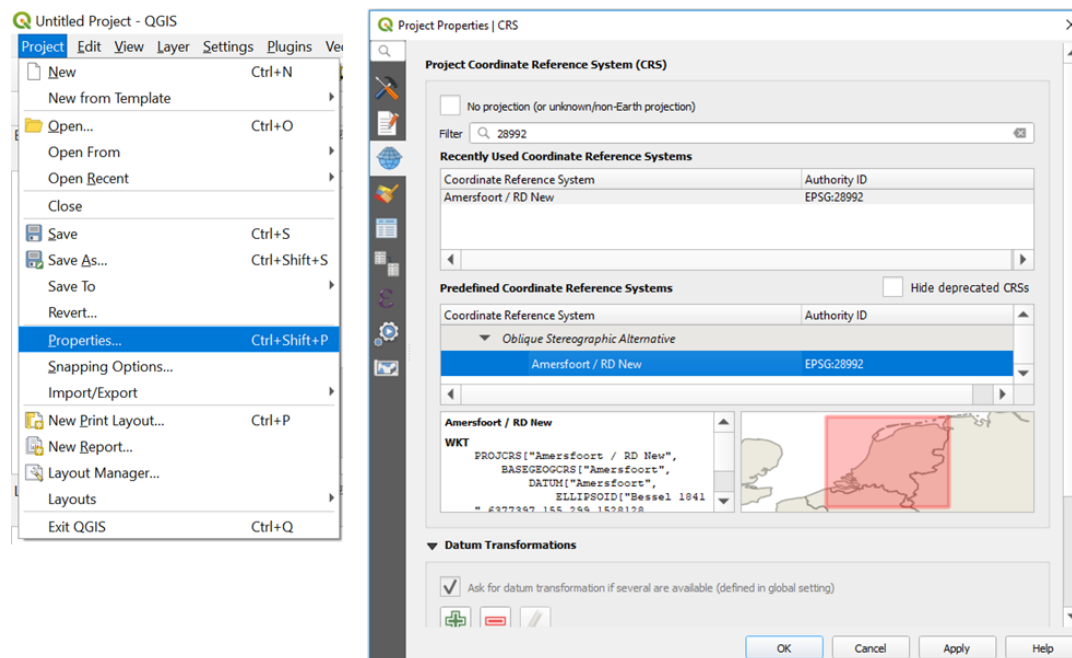
Open QGIS Desktop. Then create a QGIS project by clicking **New Empty Project**:



Coordinate Reference System

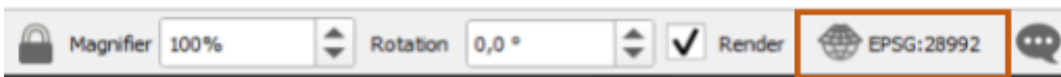
QGIS projects use a global Geographic Coordinate System (GCS) – “WGS 84” (EPSG: 4326) by default. However, because of geometric distortion, “WGS 84” does not reflect the relative size and shapes of reality in the Netherlands. We need to change it into the Dutch standard Coordinate Reference System (CRS) - “Amersfoort / RD New” (EPSG:28992). “RD New” is the projection based on the GCS Amersfoort Ellipsoid.


Under **Project** click **Properties....** Then choose **CRS** tab and search the EPSG code of “Amersfoort/ RD New” in **Filter** and apply it.



Note that all datasets that you want to analyze together need to be in the same projection. However, other data sources differ. QGIS supports automatic CRS transformation for both vector and raster data. After you define the CRS for a project, QGIS will project all layers that you load into **Layers** and which are not in “RD New” into this CRS *on the fly*.

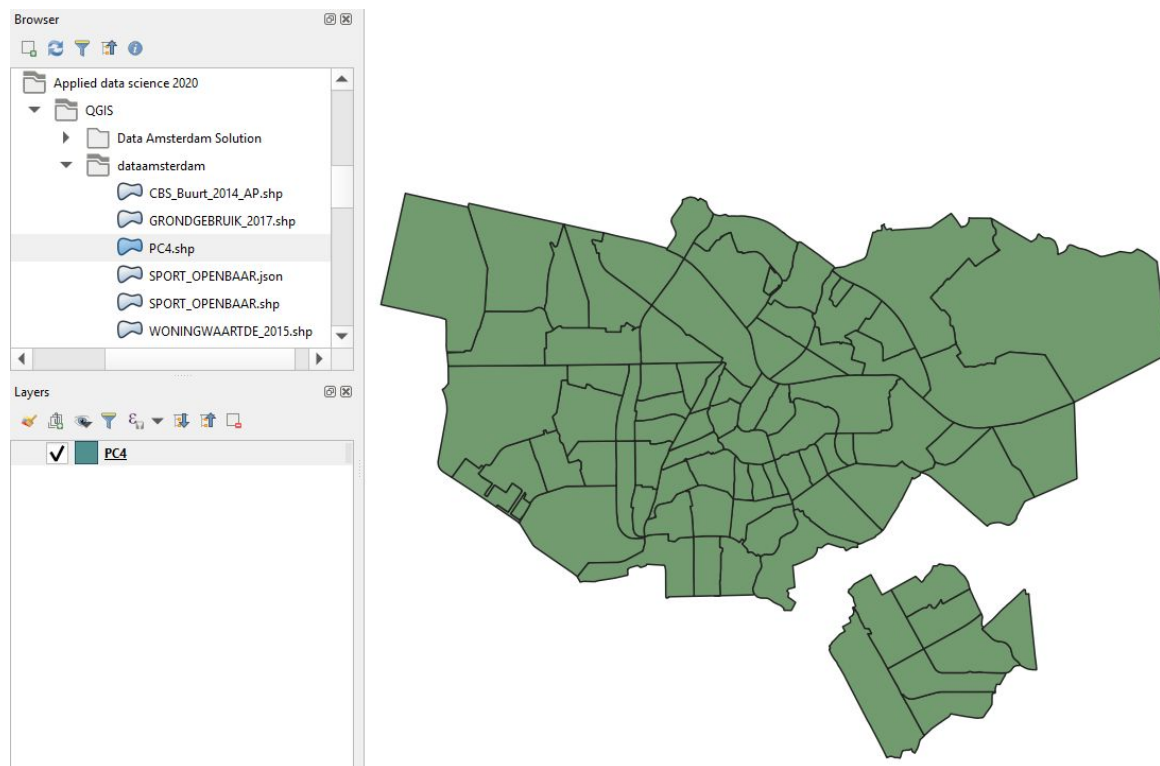
You can check the used CRS for the current project at the right-bottom of the interface.



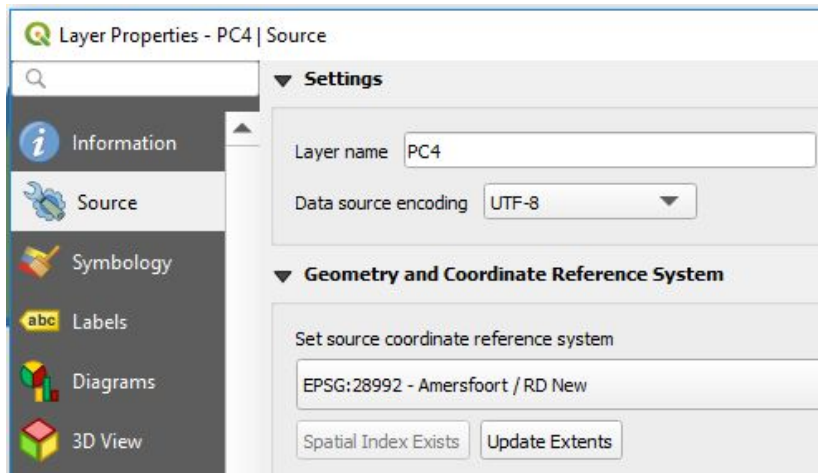
Click **Save** button  in the toolbar, create a new folder “Lab 3.1 solution” and save the QGIS project as “Lab 3.1” in it. Next time QGIS can directly open all layers and maps in the stored layout.

Load data

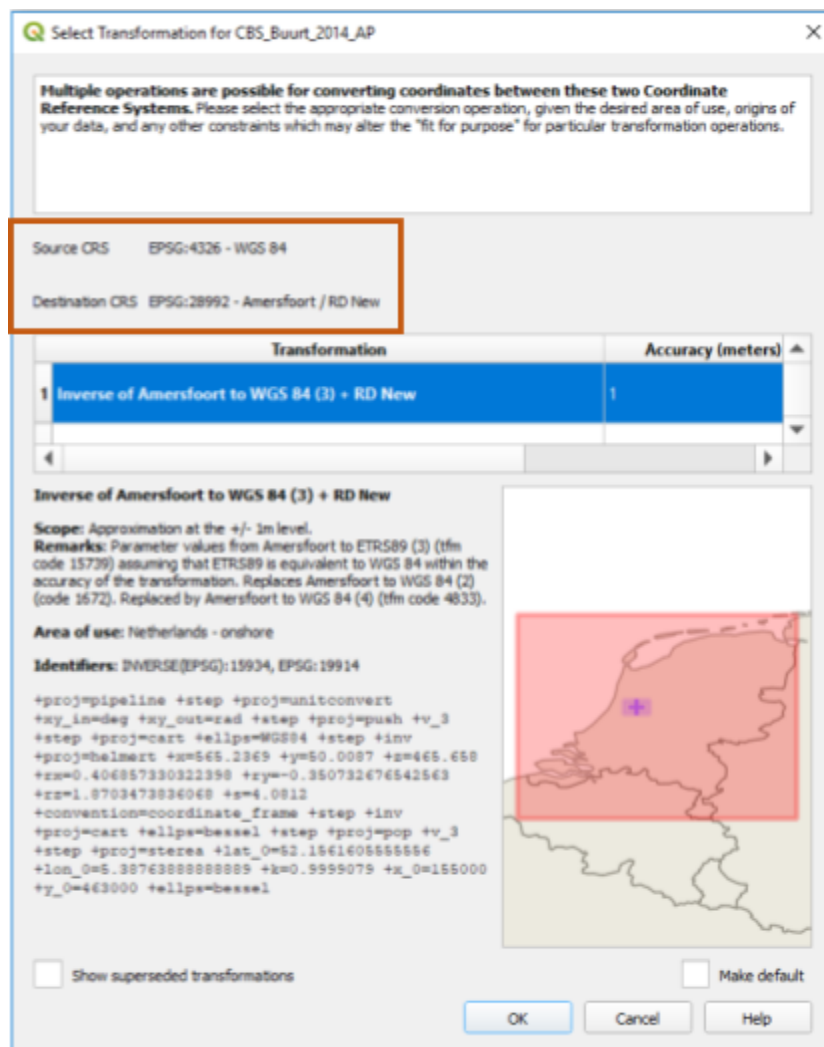
In the **Browser** panel, locate the “Lab3.1_dataAmst” folder in your directory. Then select “PC4.shp” and drag it onto the **Layers** panel. A map should appear which looks like this:



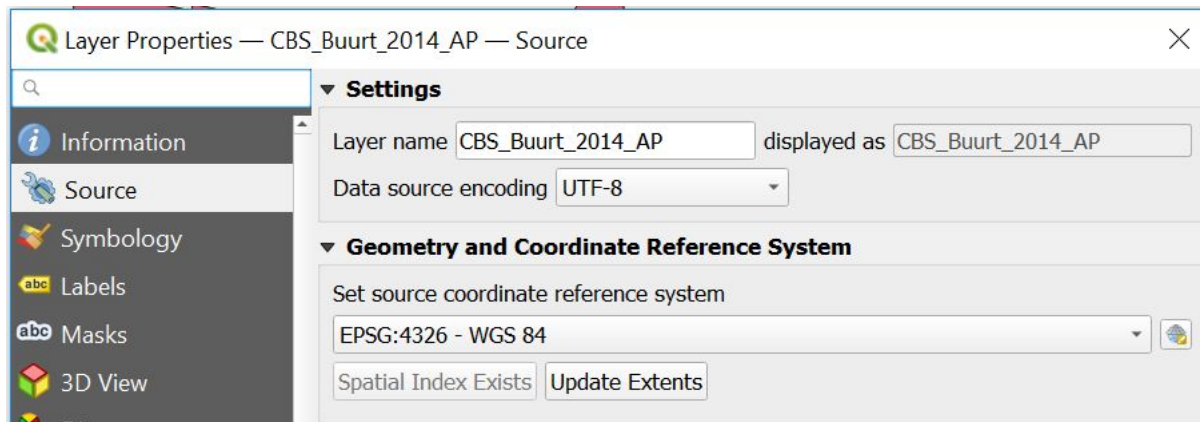
Right click or double click on the “PC4” layer in the **Layers** and click **Properties**, then go to **Source** tab. This shows you that the data itself is projected in “RD New” based on the GCS Amersfoort Ellipsoid.



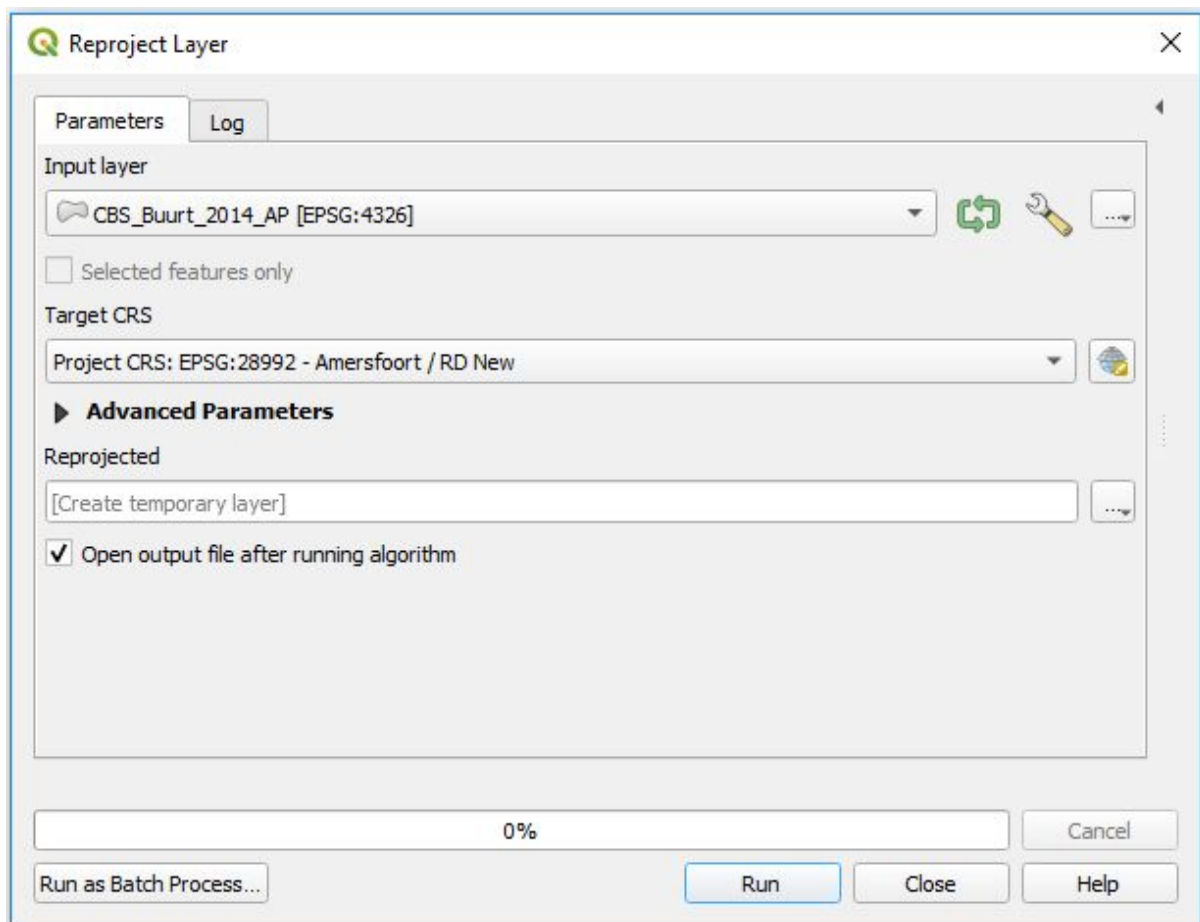
Drag and drop the file “*CBS_Buurt_2014_AP.shp*” at the top of the **Layers**, a dialog will prompt you to transform coordinate on the fly, because the CRS of CBS data is WGS 84. Click OK and the CBS is displayed exactly on top of the “*PC4*” layer.



However, this transformation does not change the original CRS, but only display CBS in RD New in the current project. Open **Layer Properties** and check the CRS of CBS. It should look like this:



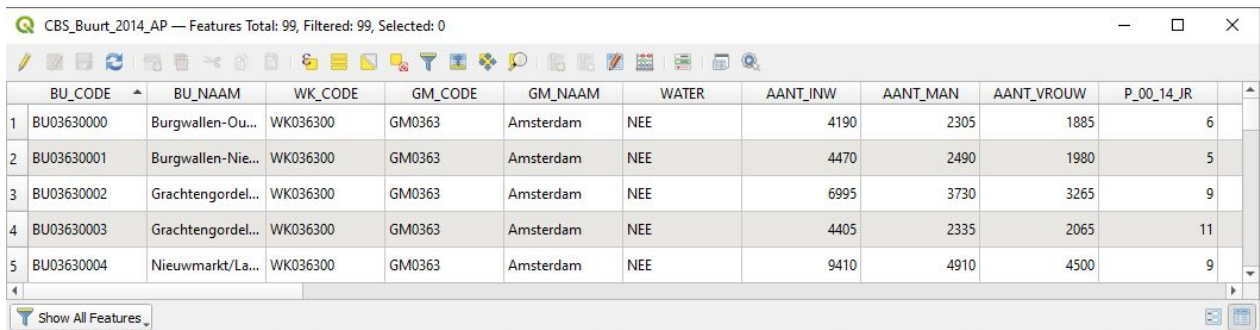
Using data with different CRS in analysis may bring some errors. It is better to use a QGIS tool **Reproject layer** which creates a new layer with the same feature as the input but with geometries projected to a new CRS. You can try to run this tool for CBS data. To save time, the data for Lab 3.1 and 3.2 are already projected to RD New.



Task 1 Simple area interpolation for (re-)aggregating statistics

In the first analysis step, your task will be to determine the percentage of inhabitants which are older than 65 years, in order to assess whether a given PC4 area hosts older people. The Centraal Bureau voor de Statistiek (CBS) openly publishes such data, however only on the level of Wijken en Buurten (Kerncijfers Wijken en buurten¹). The geodata file “CBS_Buurt_2014_AP_RDNew.shp” contains this information for the year 2014

Open the attribute table by right-clicking on this layer in the *Layers* menu and opening “*Open Attribute Table*”. Then search for the right attribute which denotes percentage of people older than 65, also by checking the data description online².



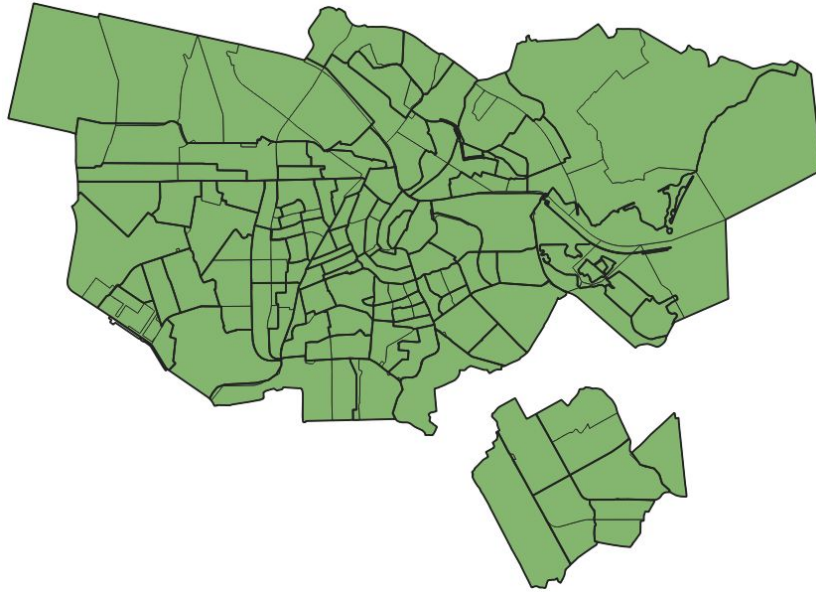
	BU_CODE	BU_NAAM	WK_CODE	GM_CODE	GM_NAAM	WATER	AANT_INW	AANT_MAN	AANT_VROUW	P_00_14_JR
1	BU03630000	Burgwallen-Ou...	WK036300	GM0363	Amsterdam	NEE	4190	2305	1885	6
2	BU03630001	Burgwallen-Nie...	WK036300	GM0363	Amsterdam	NEE	4470	2490	1980	5
3	BU03630002	Grachtengordel...	WK036300	GM0363	Amsterdam	NEE	6995	3730	3265	9
4	BU03630003	Grachtengordel...	WK036300	GM0363	Amsterdam	NEE	4405	2335	2065	11
5	BU03630004	Nieuwmarkt/La...	WK036300	GM0363	Amsterdam	NEE	9410	4910	4500	9

Now double-click the CBS layer and navigate to *Properties* > *Symbology*. In the dropdown menu at the top of the symbology pane you can select a symbology type. You can leave this on *Single symbol*. Click on *Simple fill* and set the *Fill style* to *No Brush*. Make sure the *Stroke color* is black and set the *Stroke width* to 0.5 mm. The result should be similar to the following map:

¹ <https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/wijk-en-buurtstatistieken>

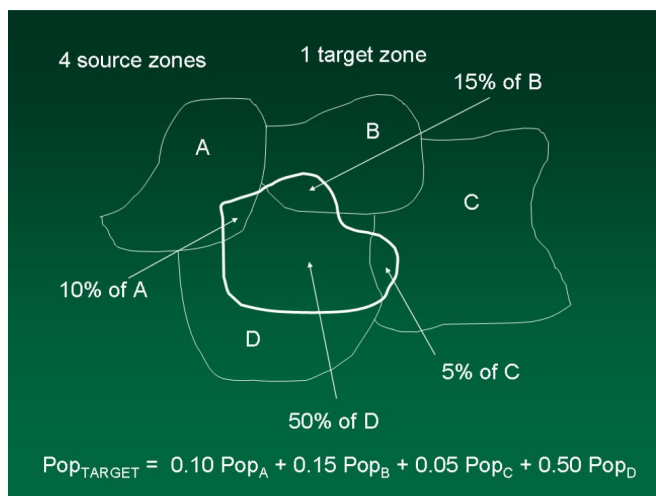
²

https://www.cbs.nl/-/media/cbs/dossiers/nederland-regionaal/wijk-en-buurtstatistieken/_pdf/toelichting-variabelen-kwb-2014.pdf



As you can see, the PC4 area boundaries do not coincide with the CBS buurt boundaries. In order to assess the statistics for the PC4 areas on the basis of the given statistics in the CBS data, you need to *interpolate* the CBS value.

This problem is called *area interpolation* in GIS. There are sophisticated tools for this purpose in QGIS³. However, for the sake of simplicity, you will implement the following interpolation technique in terms of Overlay tools, which computes the new statistics as an *area-weighted* average:

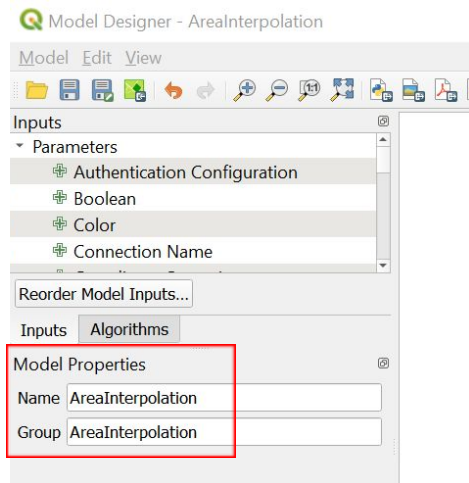


Generate new model

We can now start making a workflow for the area interpolation. For this purpose, we will use the *Graphical Modeler*. Go to *Processing > Graphical Modeler*. This opens up a new window named


³https://docs.qgis.org/3.10/en/docs/gentle_gis_introduction/spatial_analysis_interpolation.html

Model Designer. In *Model Properties*, enter a name for the model and a name for its group and save the model.



When you look in the top left corner of the window, you can find a list of input types. Drag and drop the *Vector Layer* type into the workspace on the right side of the window. You will be prompted to give a description of the input. Name it “Target Layer”. Then add another *Vector Layer* input and name it “Source Layer”. These will be the input parameters for the workflow.

Overlay: Intersection

Now we can start adding operations to the workflow. In the *Algorithms* tab, find the *Intersection* operation and drop it into the workspace. A new window pops up for the configuration of the algorithm. The inputs are automatically configured to take *Value* as input. While this setting allows us to integrate specific layers into the workflow, we want to be able to specify new layers each time the workflow is run. To enable this, click the  icon and select *Model Input* for both the *Input Layer* and the *Overlay Layer*. Assign *Target Layer* to the *Input Layer* setting and *Source Layer* to the *Overlay Layer* setting and click OK. The input parameters and the *Intersection* operation should now be linked in the workspace.

Intersection

Properties Comments

Description Intersection

Show advanced parameters

Input layer

Using model input Target Layer

Overlay layer

Using model input Source Layer

Input fields to keep (leave empty to keep all fields) [optional]

123

Overlay fields to keep (leave empty to keep all fields) [optional]

123


Intersection

[Enter name if this is a final result]

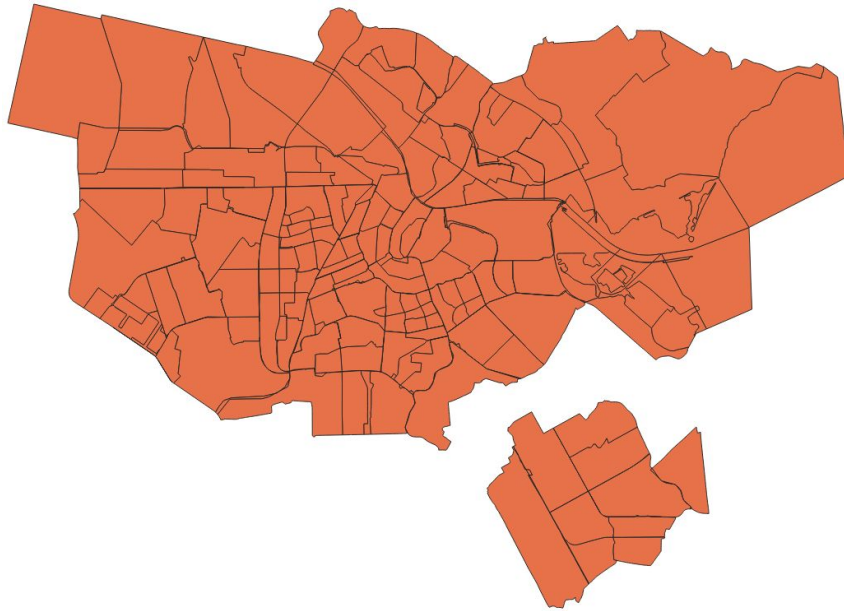
Dependencies

0 dependencies selected


OK Cancel Help

Get the result of the operation by entering the name “clipped” for the output in the last box and running the model (click the  icon in the *Model Designer* window). If the operation output is given a name, it will output the result in the QGIS project, even if it is not the last operation in the model.


Enter CBS_Buurt_2014_AP as the *Source Layer* and PC4 as the *Target Layer* and click run. By default, layers are generated as temporary scratch layers, which will be lost after QGIS gets closed. To avoid this, specify a file location for the output. The output should look like this:



Calculate the product of area and the statistics attribute

Right-click the intersection output layer from the previous step and select *Open Attribute Table*. Here you can see the values that are attributed to each spatial feature. As you can see, the intersection output layer contains no information about the size of the area of the newly generated polygons (Column “Opp_m²” contains the sizes of the corresponding postal code areas). In order to compute an area-weighted average, we first need to measure the area of the newly generated polygons. To do so, Add the *Add geometry attributes* operation to the model and set the output of the intersection operation as the *Input layer*. You can do this by clicking the  icon and selecting *Algorithm Output*. Generate the result by naming the output “add_geom” and running the model. The result of the operation contains two new columns called “area” and “perimeter”.

Next up, we need to generate a new attribute field which should contain the product of the area with the statistics attribute (in our case “P_65_EO_JR”). To do this, search and add the *Field Calculator* tool to the workflow. Select the output of the previous operation as the *Input layer*, enter “product” as the *Result field name* and set the *Field type* to *Float*. Name the output so it gets generated in the QGIS project.

Then, go to formula and click the  icon on the right of the input field. A new window pops up where you can perform advanced operations on attribute data of layer files. The *Expression* tab lets you specify operations with a custom QGIS language, which is similar to SQL, while the *Function Editor* lets you specify them using Python. We will be using the *Expression* tab this time.

We can navigate to specific attribute values by using the syntax `attribute(feature, 'column name')`. Because the *Field Calculator* tool uses values from a layer which is



generated earlier in the workflow, we cannot call the features of this layer directly. Instead, we call `$currentfeature`, which returns the feature that is currently being evaluated. Enter the following text in the text box:

```
CASE
WHEN attribute($currentfeature, 'BU_CODE') IS NULL
OR attribute($currentfeature, 'P_65_EO_JR') <= 0
THEN NULL
ELSE attribute($currentfeature, 'area') * attribute($currentfeature,
'P_65_EO_JR')
END
```

The expression fills the “product” attribute with the product of the newly generated “area” and the P_65_EO_JR values. However, this is done only whenever the attribute is non-negative and filled with any values, otherwise NULL is entered into the field⁴. Take some time to understand the what is happening in the expression and click OK. Name the output “calc” and run the model to generate the layer with the new “product” column.

Aggregate region statistics

Drag “calc.shp” to the *Layers*, so that we can use it in the *Aggregate* tool.

Add the *Aggregate* tool to the model. This tool unifies polygon regions of an input layer into aggregated regions based on some attribute or expression (*Group by expression*). We use it here to aggregate the products of CBS Buurt attributes and areas into PC4 regions. Again, select the output of the previous operation as the *Input layer*. In the *group by expression* field, write Postcode4. The *Aggregates* field is currently empty, but should contain the attributes we want to include in the output layer. Specifying the relevant attributes manually would work, but since we generated the input layer in the previous step, we can load the attributes from a template layer. Select “calc” in the *Load fields from template layer* field and click *Load Fields*. This generates the specifications of the attribute fields of that layer. Change the *Aggregate Function* of “Postcode4” to *first_value* and delete all attributes except “Postcode4”, “area”  and “product” with the icon  (Use shift-click to select multiple attributes at once).

⁴ This is due to the fact that the original CBS data has missing values denoted by negative values.

Aggregate

Properties Comments

Description: Aggregate

Input layer: Using algorithm output "Calculated" from algorithm "Field calculator"

Group by expression (NULL to group all features): Postcode4

	Source Expression	Aggregate Function	Delimiter	Name	Type	Length	Precision
0	"Postcode4"	first_value	,	Postcode4	Whole number (integ	4	0
1	"area"	sum	,	area	Decimal number (doi	23	15
2	"product"	sum	,	product	Decimal number (doi	10	3

Load fields from template layer: calc Load Fields

Aggregated: C:/UU/3_Courses/Applied data science 2020/Lab3.1/Lab3.1 solution/area_interpolation_data/aggr.shp

Dependencies: 0 dependencies selected

OK Cancel Help

Among these attributes are “area” and “product”, which will be the two terms for building a weighted sum⁵: the sum of products and the sum of weights (in our case a spatial weighting with “area”). Note that the aggregations of the original CBS Buurt attributes are not spatially weighted. Name the output layer “aggr”. Remove “calc” from *Layers* and run the model. The new layer should look exactly like the old PC4, with the difference that now you have two new attributes in there.

Finally, you have to divide one aggregated sum by the other. For this purpose, use the *Field Calculator* tool to add a field (of type *Float*) “WVALUE” and to fill it with the ratios of the sums of products divided by the sums of areas by entering the custom expression:

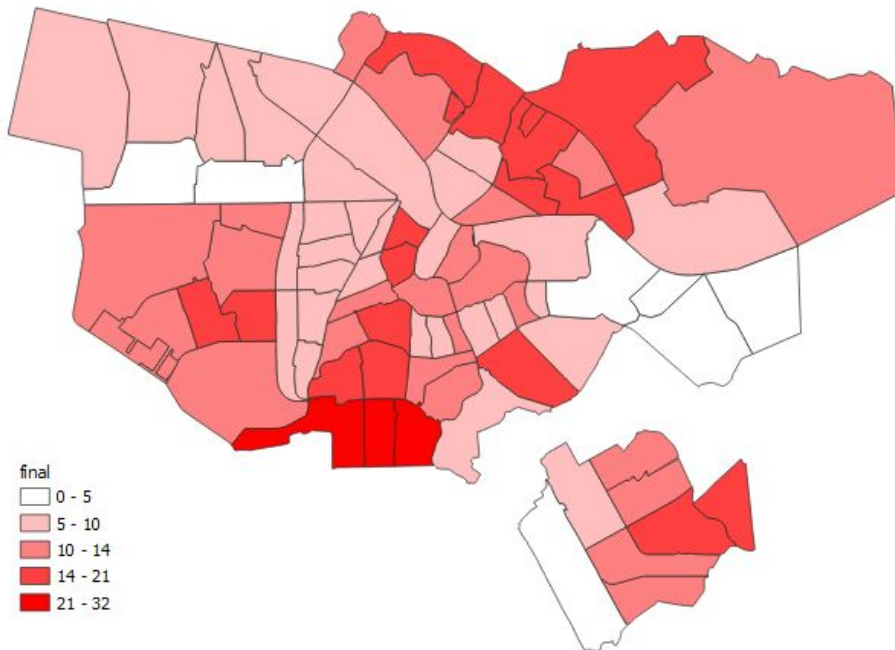
```
attribute($currentfeature, 'product') / attribute($currentfeature, 'area')
```

Make sure you properly integrate the *Field Calculator* tool in the model. Name the output as “final”.

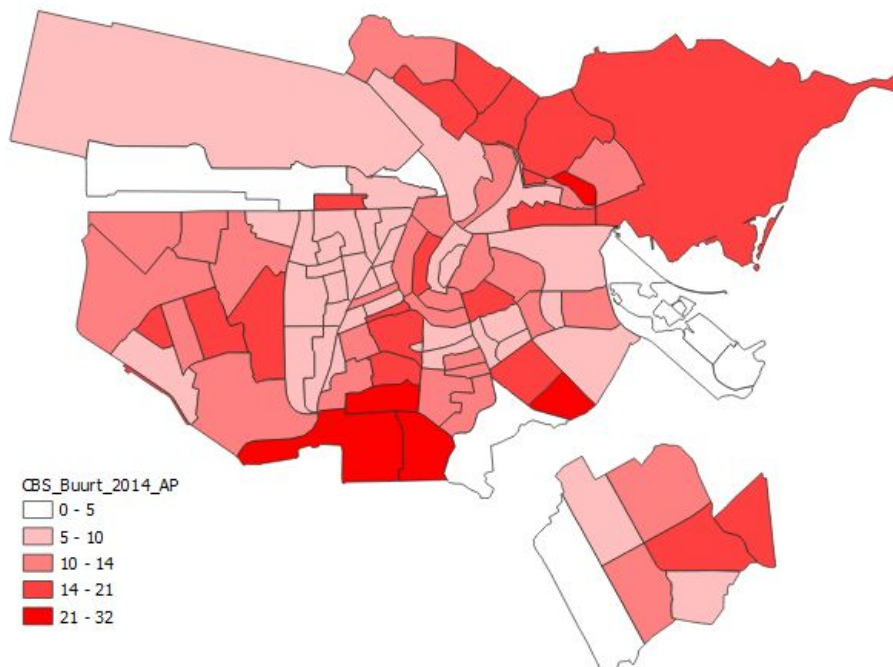
The final layer displays PC4 regions and should contain the newly computed attribute “WVALUE”. Map this attribute in a choropleth map by opening the layer’s *Properties* > *Symbology*, choose *Graduated* and select the attribute “WVALUE” as *Value*. Click on *Classify*. After the classes are generated, set the upper values to 5, 10, 14, 21 and 32. Click OK.

Then go to *Project* > *New Print Layout*, give a name to create a new layout for “final”. Under *Add Item*, click *add map* and *add legend*, then you will have a map shown below. You can export the map as image by clicking *Export as Image* under *Layout*.

⁵ Remember that a weighted sum is defined as: wv , w = weights and v = values



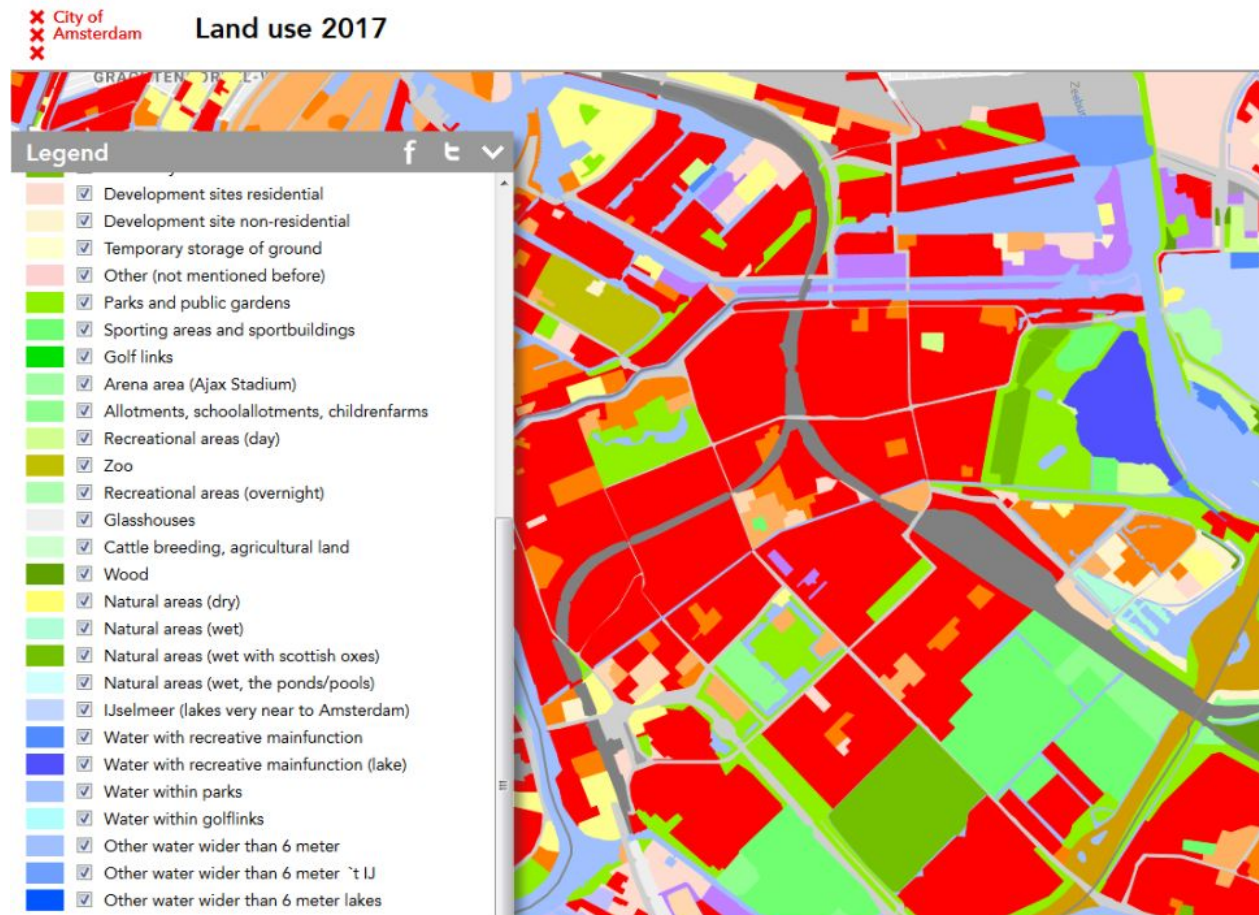
Finally, classify and map the CBS layer and its statistical attribute “P_65_EO_JR” in the same way (also with the upper values of 5, 10, 14, 21 and 32). Set the lower bound of the lowest category to 0 to exclude missing values. The result looks like this. Compare it with the last map. It gives a good impression of the quality of our re-aggregation method:



In *Model Designer*, under *Model* click *Export*, try to export the model you created as Image and Python scripts. You will submit the images and python scripts of your models, as well as the maps with legends for Lab 3.1 and Lab 3.2.

Task 2: Map algebra for aggregating landuse areas

In the second analysis step, your task will be to aggregate landuse data into PC4 areas in order to determine the coverage of these areas with parks and green and thus whether they are suitable for older people who would like to take a walk in the park. The Amsterdam Data Lab publishes high resolution data on landuse in terms of vector polygons with an associated landuse attribute⁶.



Your task is to compute the area of parks and green areas within each PC4 region. This can be done by applying several raster analysis tools from map algebra to the data as explained in the following.

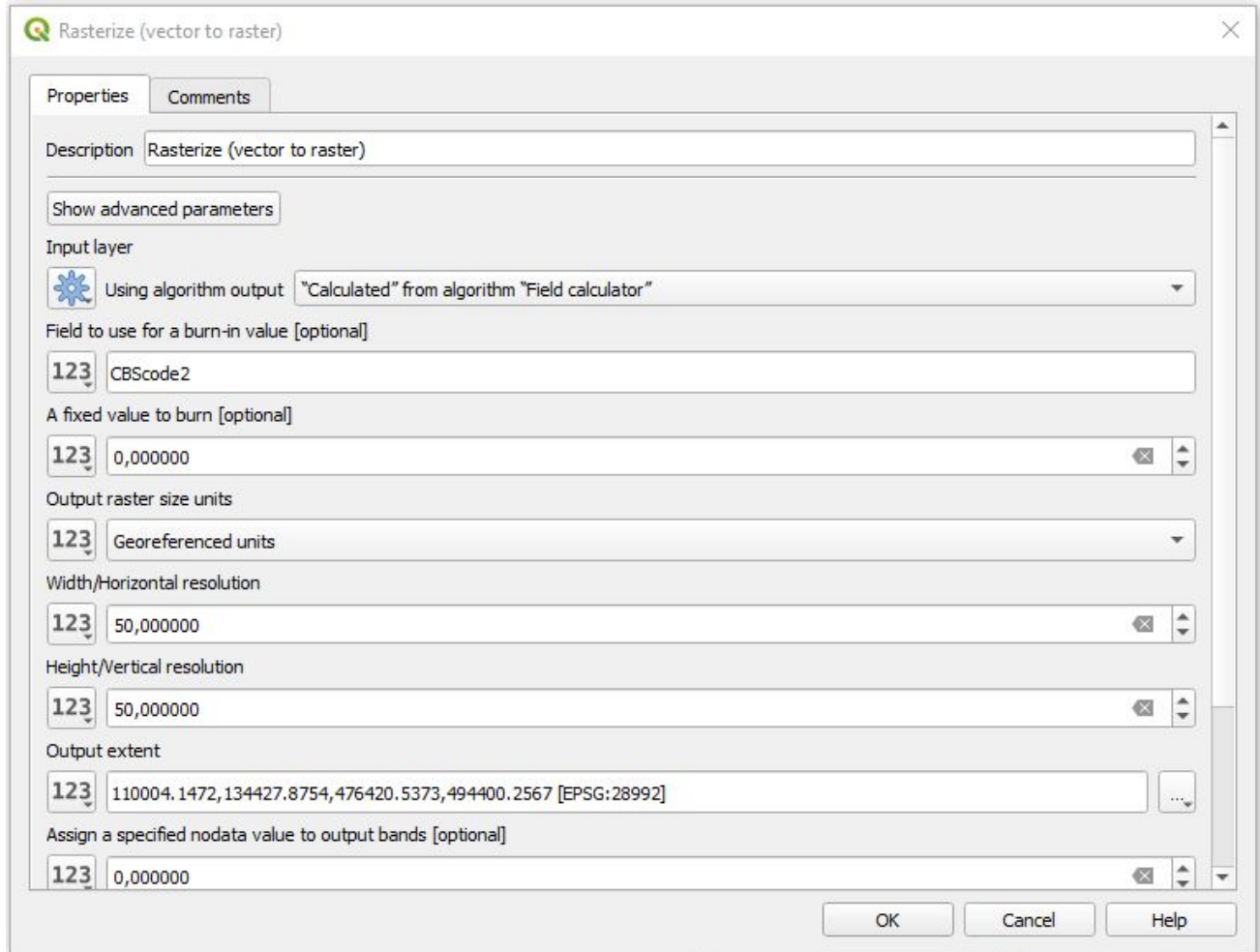
Open the landuse file and add it as a layer to your map ("GRONDGEBRUIK_2017_RDNew.shp"). Also open its attribute table and study the different landuse classes given in attribute "CBScode2" and the explanations in "CBScode2_O". Note that code 40 denotes parks.

Finally, open a new workflow model file. Since we are aggregating discrete field data (landuse), we can call this model "FieldAgg".

⁶ <https://maps.amsterdam.nl/grondgebruik/?LANG=en>

Convert landuse polygons to raster and select park areas

Search and drag the tool “Rasterize” into the *Model Designer* window to convert the land use polygon to raster. Set the *Input Layer* to *Land use* and type *CBScode2* in the *burn-in value* field (This refers to the attribute we will use to generate raster values). Set the *Output raster size units* to *Georeferenced units*, so it will use the project’s distance metric (This should be **meters**). Set the *Width* and *Height* to 50 meters each. This operation requires an extent to define the area to be processed. In this case, we can select PC4 data as extent. Save the rasterized output as “Landuse_raster”.




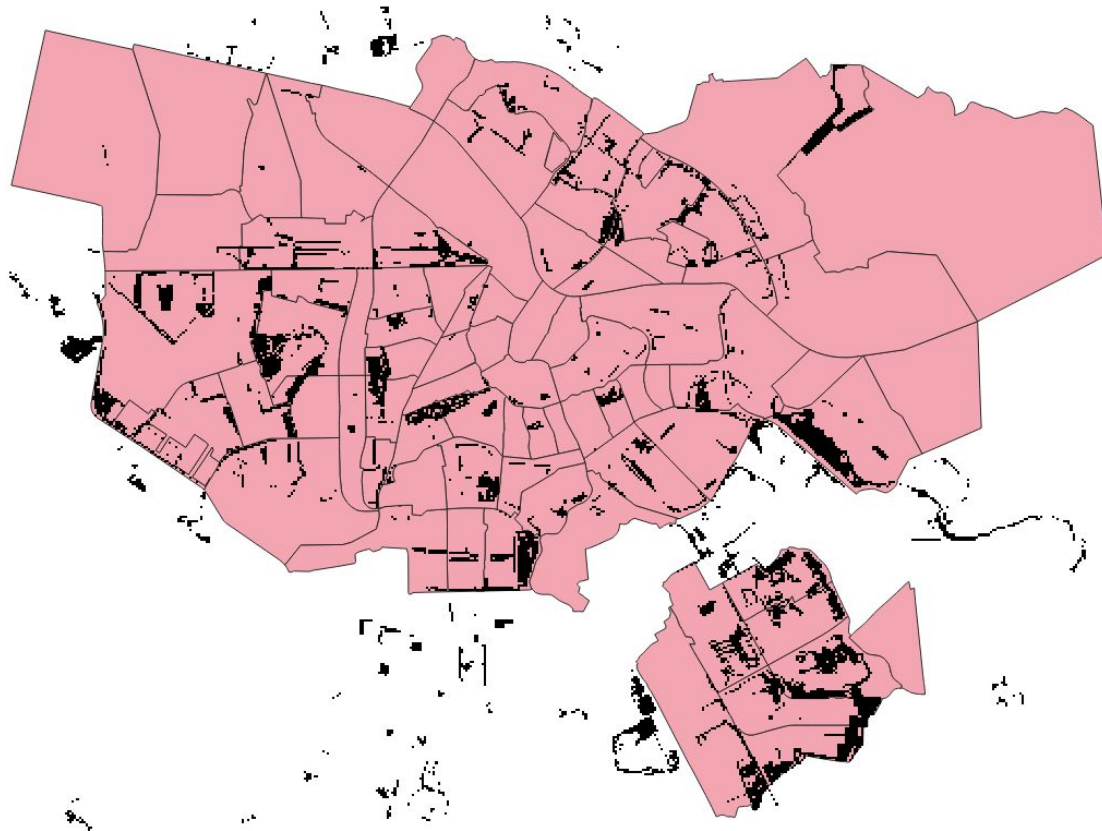
The screenshot shows the "Rasterize (vector to raster)" dialog box in QGIS. The "Properties" tab is active. The "Description" field contains "Rasterize (vector to raster)". Below it is a "Show advanced parameters" button. The "Input layer" section shows a gear icon and the text "Using algorithm output" followed by a dropdown menu showing "Calculated" from algorithm "Field calculator". The "Field to use for a burn-in value [optional]" section has a dropdown menu showing "123" and a text field containing "CBScode2". The "A fixed value to burn [optional]" section has a dropdown menu showing "123" and a text field containing "0,000000". The "Output raster size units" section has a dropdown menu showing "123" and a text field containing "Georeferenced units". The "Width/Horizontal resolution" section has a dropdown menu showing "123" and a text field containing "50,000000". The "Height/Vertical resolution" section has a dropdown menu showing "123" and a text field containing "50,000000". The "Output extent" section has a dropdown menu showing "123" and a text field containing "110004.1472,134427.8754,476420.5373,494400.2567 [EPSG:28992]". The "Assign a specified nodata value to output bands [optional]" section has a dropdown menu showing "123" and a text field containing "0,000000". At the bottom right are "OK", "Cancel", and "Help" buttons.

Note that an unprojected vector data (e.g., “CBS_Buurt_2014_AP.shp”) only has “degrees” as its Unit (check under *Layer Properties > Information*). By projecting the vector data to a certain CRS, the unit will become “meters” for further area calculations.

If you have an error in this step, this might be because a string field can not be used as a burn-in field in some QGIS versions. Then you can use the “Field calculator” tool to create a new field that has the same value as CBScode2 but is an integer field instead of a string field. Enter the following text into *Expression* in Field calculator to convert field type :

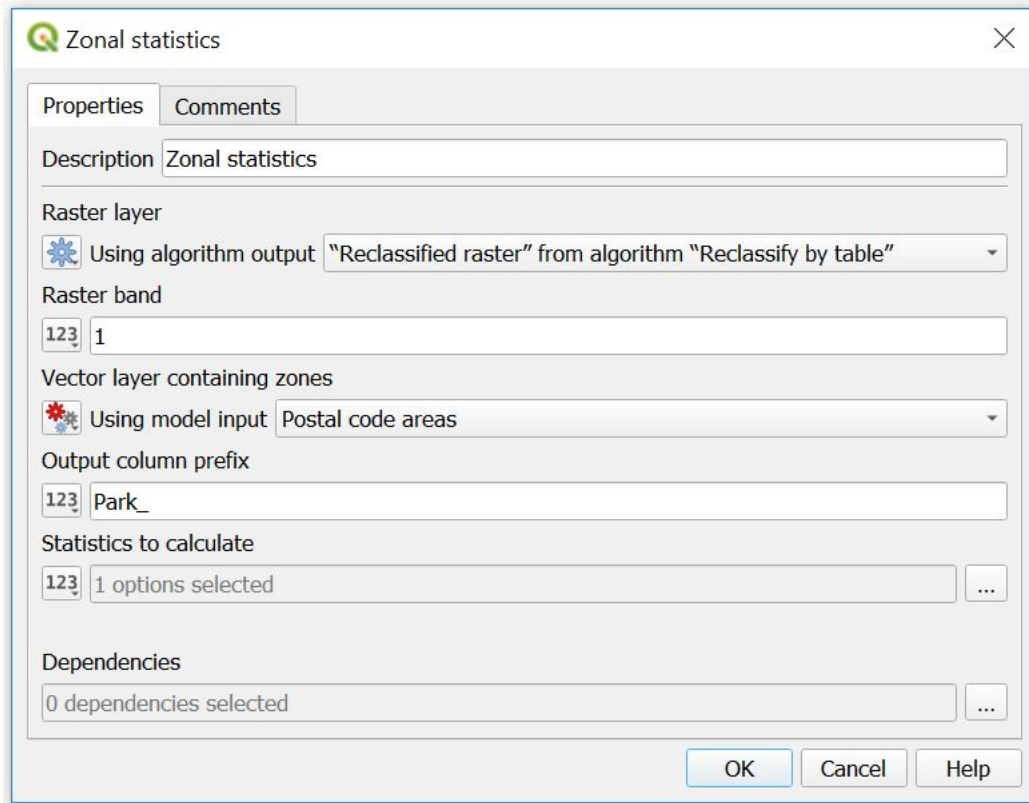
```
to_int(attribute($currentfeature,'CBScode2'))
```

Next, we need to select those raster cells that denote park areas. For this purpose, use the tool *Reclassify by table* in order to generate a new raster which has “1” values for all cells inside of green areas and “NoData” values for all cells that lie outside park areas. Figure out on your own how to specify a reclassification mapping of the attribute “CBScode2” using the *Reclassification table* (Click the -icon on the right of the input field). In the advanced parameters, set *Use no data when no range matches value* to Yes. Output the result as “parks” and execute the model with “GRONDGEBRUIK_2017_RDNew.shp” as *Land use*. The resulting raster layer should look like this:



Aggregate park areas into PC4 by zonal map algebra

In the second step, we need to aggregate the park raster into PC4 areas. First drag a Vector Layer and name it as “*Postal code areas*”. Then we use the *Zonal Statistics* tool which computes a zonal map algebra operation using “*Postal code areas*” as zone definition. The map algebra can make use of several algebraic functions.



In our case we are only interested in the function “count” which is generated by default⁷. Search and add this tool. Use the output of the previous operation as the *Raster layer* and use *Postal code areas* as *Vector layer containing zones*. Write “Park_” in Output column prefix. The results from this operation will not be stored in a new layer. Instead, when you run the workflow with the same inputs as before, three new attributes will be added to the “PC4” layer. The attribute we need is “Park_count”. This attribute denotes the number of non-Null cells within each postcode 4 area of the raster layer, and thus captures the number of cells that lie within a park.

In order to compute the coverage of park areas in each postal code area, we need to divide the area of these park cells by the polygon area. The areal coverage of park cells can be estimated by:

$$\frac{\text{number of cells in polygon} \times \text{area of a cell in m}^2}{\text{area of polygon in m}^2}$$

The area of each cell can be based on the cellsize of your raster. The “PC4.shp” layer already contained an attribute measuring the polygon area, namely “Opp_m2”. Add a *Field Calculator* operation to the model and calculate, using “Park_count” and “Opp_m2”, a new attribute “parkarea”. Select the output of the previous operation as the *Input layer* and run the operation. Finally, change the symbology to display a graduated colors map with “parkarea” as attribute (You can use the classes that are automatically generated after clicking *Classify*) and place the “parks” layer above it. The result should look like this:

⁷ In this exercise, we use the number of cells in order to measure landuse area.

