PROCESSING OF COMBINE HARVESTER YIELD MONITOR DATA IN QGIS

Andrei Girz and Tuomas J. Mattila,

Finnish Environment Institute SYKE

Combine harvester yield monitoring equipment is becoming more common. Most hardware manufacturers also supply software for interpreting the data, but the data can also be analyzed in QGIS an open access software. Analyzing the data yourself allows combination with external datasets (topography, vegetation NDVI index, etc.). Several years can be combined to identify management zones and statistical analyses can be run to determine if experiment strips resulted in a statistical difference. This short guide describes a workflow for cleaning, processing and analyzing data from yield monitors. We will use a few Finland specific examples, but the workflow is applicable across the globe. The Finland specific parts are in grey boxes.

1. DOWNLOAD AND INSTALL QGIS

QGIS can be freely downloaded from https://www.qgis.org/en/site/forusers/download.html Follow the instructions on the website for installing. We recommend creating a QGIS project folder also on your harddrive (QGISprojects/ etc.), where you can store separate project files and common resources used across many projects.

2. OPEN AND SET UP A QGIS PROJECT AND ADD A BACKGROUND MAP

Select **"Project → "New"**. Then **"Project → Save"** and save to a project folder. You now have a blank sheet to work on. Next we add a background map.

Internationally you can use the Google Earth XYZ tile server (**"Layer → Add Layer → Add XYZ layer → New"**. Then copy the following address: https://mt1.google.com/vt/lyrs=s&x={x}&y={y}&z={z}). You should now have a world map in WGS 84 Pseudo Mercator coordinate system. In the next step we will navigate to the field.

Some regions have WMS services for more detailed maps, for example in Finland there is the local Maanmittauslaitos MML Avoin.

- *Sign up for a username from MML https://omatili.maanmittauslaitos.fi/ -> Create an API key and copy it*
- *In QGIS: Layer > Add Layer > Add WMS…*
- *Choose New, add this to the URL https://avoinkarttakuva.maanmittauslaitos.fi/avoin/wmts?SERVICE=WMTS&REQUEST=GetCapabilities and your API key to the username.*
- *Connect to the server and choose for example "Ortokuva" aerial photo.*

3. IMPORT THE YIELD POINTS FROM THE YIELD MONITOR

Yield monitoring software should have the possibility of exporting data in a Shapefile (*.shp, coming with three files .shp,.shx and .dbf). Store the files in your project folder. Add the layer through **"Layer → Add layer … → Add vector layer"** and select the .shp file). In the QGIS Layer panel right click the added layer and choose **"Zoom to layer"**. You should now have the yield points on top of the aerial background photo (If the layer ends up in the wrong place in the world, it is probably in the wrong coordinate system and needs fixing during the Add Layer menu). The points are all of the same colour, making interpretation difficult. Next we will add some colour.

Add colors for the values. Right click on the yield point layer. While you are there, rename the layer to something easy to use (**"Rename layer"** for example FARM_Field_crop_year). Then choose from the right click menu "**Properties**… **→ Symbology →** change **Single Symbol** to **Graduated → Value →** your yield option (i.e. "yldkg_ha")". Now you have QGIS drawing different colors to points based on the recorded yield. The default color ramp is not very useful, so that could be changed.

In the same **Symbology** menu choose **Color ramp** → **Spectral.** That color scale is neutral for average yield, blue for good and red for bad yields. Clicking **Classify** will show the classification to colours based on equal counts of points in each group. So in theory each color has the same field area. You can check the **Histogram** tab (**Load values** – button) to see how the points are classified and which are the yield ranges in the raw, unfiltered data. Click **Apply** to use the colours in the map. The result should look like Figure 1 below.

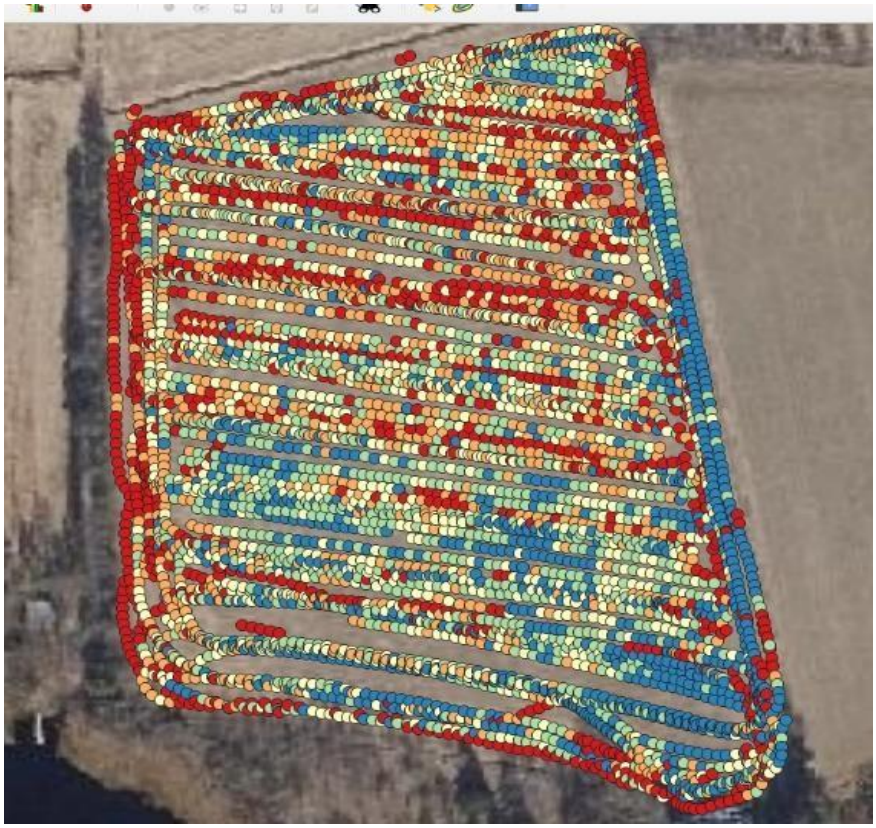

Figure 1: Color-coded yield monitor data with Spectral color ramp.

NOTE! Remember to save your work! Preferably after each major step.

4. DATA CLEANING

The yield monitor produces some unrealistically high and low yield results, especially when the combine slows down, turns or cuts a partial swath. Yield monitor data cleaning is necessary for making a reliable analysis and drawing realistic conclusions. At this point, most of the data cleaning is directed towards removing the outliers.

There are two steps involved in data cleaning. First, the outliers must be identified and second, they must be removed from the attribute table.

As a first step, make a copy of the original data. Right click then **Export** → **Save features as…** You can save it as a .shp file with a file name something like FARM_Field_Crop_Yield__Year_Cleaned .

IDENTIFYING THE OUTLIERS

The outlier identification can be easily and statistically sound done by using the standard deviation as a yardstick. A standard deviation describes how spread out the data is from it's average value. At two standard deviations distance, 95% of the datapoints that are closest to the average yield are retained and 5% of the data is removed. The removed data represents very high and very low values.

The first step is to remove all the points which have an abnormal driving speed.
Choose **Symbology** → **Value** → "Speedkmh" (or similar). Apply it to the map and see if you can identify points of too high or too low speeds. Usually these could be around < 3 km/h and > 8 km/h, but it depends on the combine operator and crop.
Right click the Layer and select **Open attribute table.** This lets you see all the datapoints.
Choose **Select by expression** (Figure 3 blue arrow). Put the expression "***speedkmh <= 3 OR speedkmh >= 7.5***" to the Expression box and choose **Select features.** You now have the points which fall outside the desired speed range marked as blue. Click the yellow pen edit toggle button and then click the red trash bin to delete all those observations. Click the **Save edits** disk and close the attribute table. You can compare the removed points with the original by clicking the saved copy layer and the original layer on and off in the layer menu.

Now let's repeat the cleaning for yield, using a bit more detailed cut-off points. Figure 2 shows an example of this from the **Symbology** → **Histogram** tab. The yield data points in this histogram are not equally distributed to the right and left of the yield distribution peak. On the left of the peak, very low values are counted, and on the right, the yield goes up to 11 400 kg/ha. The mean (μ) is at about 8200 kg/ha and the yield between one standard deviation (σ) to the right and left of the mean (marked with green) ranges between ~7000 kg/ha and 9200 kg/ha. At two standard deviations to the right and left of the mean (area marked with red), 95% of the data points are included and the yield ranges between 6000 and 10300 kg/ha. This is a more realistic depiction of the yield.

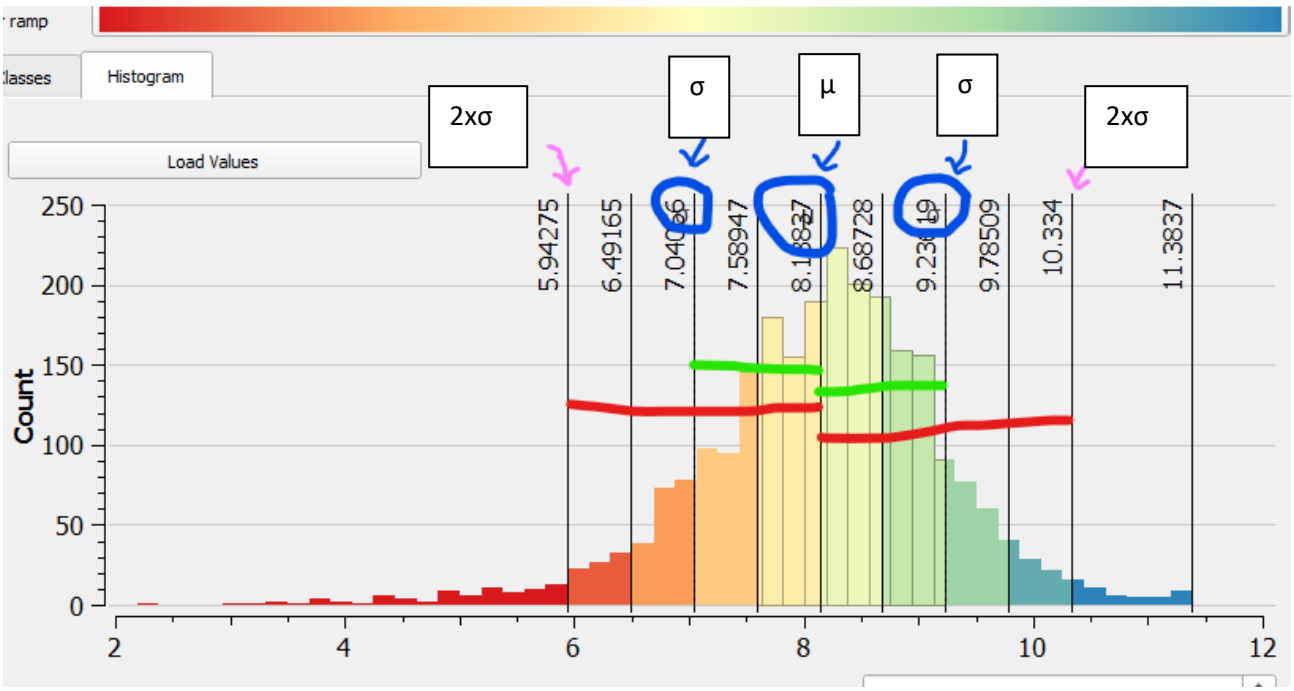

Figure 2: Histogram showing the yield on x-axis and yield frequency on the y-axis. The standard deviation (σ) is covered by the yield values but then is emphasized in blue on both sides of the mean yield (μ). The 2xσ yield limits are pointed with pink arrows.

Choose **Symbology** → **Mode** → **Standard Deviation**" (you can add more classes in the right side of the window) → **Classify.** Now, to visualize the data. In the same window, click on **Histogram → Load values.** Write down the limit values of the 2 x the standard deviation (2σ), in this case 5.94 t/ha and 10.33 t/ha.

REMOVING THE YIELD OUTLIERS

This follows the same approach as for the speed. Right-click on the Yield layer: **Open Attribute Table** → Activate editing mode (pen icon upper left corner). **Select by Expression** icon (Figure 3 blue arrow) and write the expression "*yldkg_ha <= 5940 OR yldkg_ha >= 10330*" (replacing the 2 Std Dev values for the numbers). Then **Select Features** and delete the abnormal entries with the red rubbish bin icon. Save edits by the disk and you are done cleaning the data. You can again compare with the original data to see which points were removed.

Once you understand the process, you can replace the writing down of stddev-values by using the following expression in selecting the outliers:

"*yldkg_ha <= (mean(yldkg_ha) – 2* stdev(yldkg_ha)) OR yldkg_ha >= (mean(yldkg_ha) + 2* stdev(yldkg_ha))*"

If you are uncertain of deleting data, you can also create a new cleaned data column in the layer. That process would be to Open Attribute Table → Open Field Calculator. Output field name = “Clean_Yld” (See Figure 3.) Copy/paste the expression formula below in the expression field -> replace the name of the yield column to have the same name as in your attribute table (if necessary), and introduce the yield limits you have noted down.
if(("yldkg_ha" > 2500 and "yldkg_ha" <6700), "yldkg_ha" , NULL)

Now, in the attribute table, the last column should be the cleaned yield data points. Use this column for all further calculations and visualizations.

To visualize the cleaned yield data points, in the Symbology panel, change the Value field entry with the clean yield one. This approach allows you to change filtering and does not lose data.

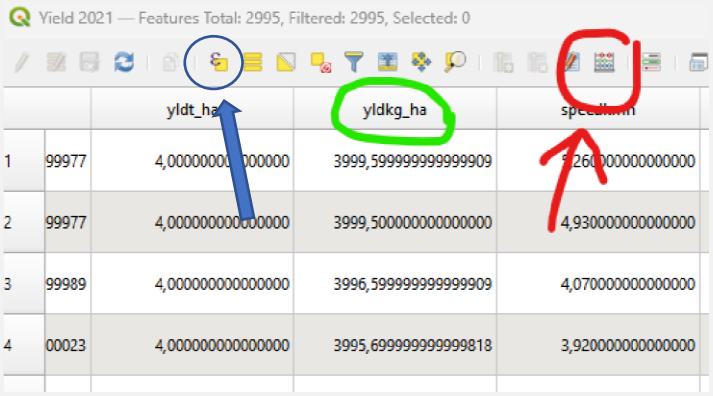

Figure 3: Attribute table. Red circles the field calculator button , and green is the name of the yield column.

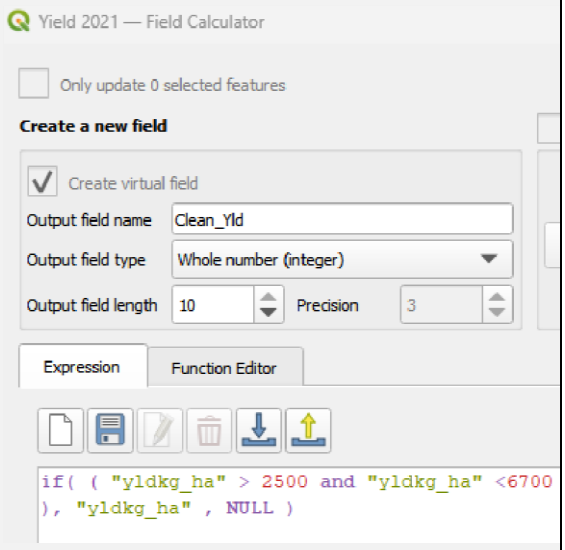

Figure 4: Field calculator example snapshot.

Remove headland area

The headland is usually receiving high traffic along the season and the yield is usually negatively impacted. The headland areas serve as turning and operative space for the combine driver. Usually, 2-3 passes are done along the headlands, depending on width of the cutterbar, before the rest of the field is harvested. This part can also be removed from the dataset before analysis. It can be removed by eye, selecting the headland harvest passes or by a certain distance from the field edge. Here it will be showed the first option.

- If you do not have the editing mode activated, do it by pressing the pen icon (red circle in Figure 5)
  - “Select Features by Polygon” (green circle and arrow)
  - Select by hand the datapoints you want to keep (close the selection by right-click)-> “Invert Feature Selection” (blue circle)
  - if happy with the selection, then press “Delete Selected” (pink circle).

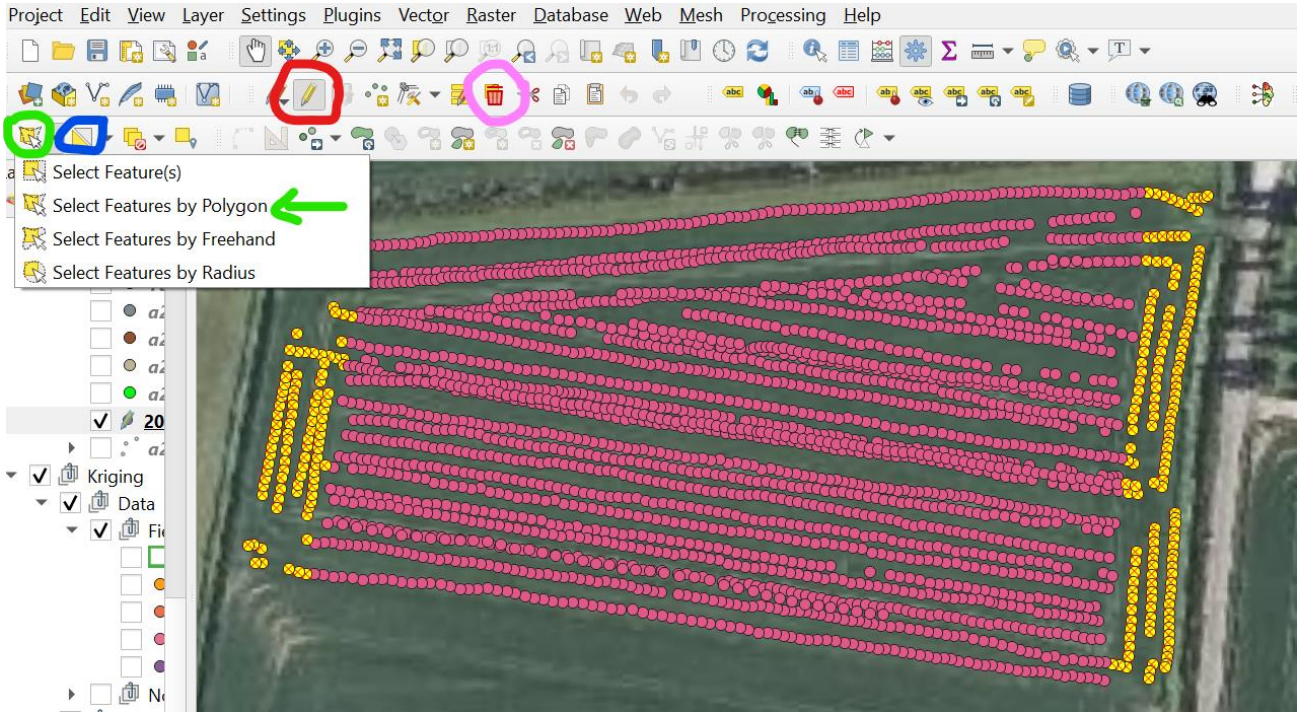

FIGURE 5: VISUAL AID FOR HOW TO REMOVE HEADLAND YIELD POINTS.

Of course, the headlands can be selected one-by-one but for uniform fields this way is faster.

Now the data is clean!

For the comparison, Figure 6 shows how many points (in blue) were removed in the cleaning process.

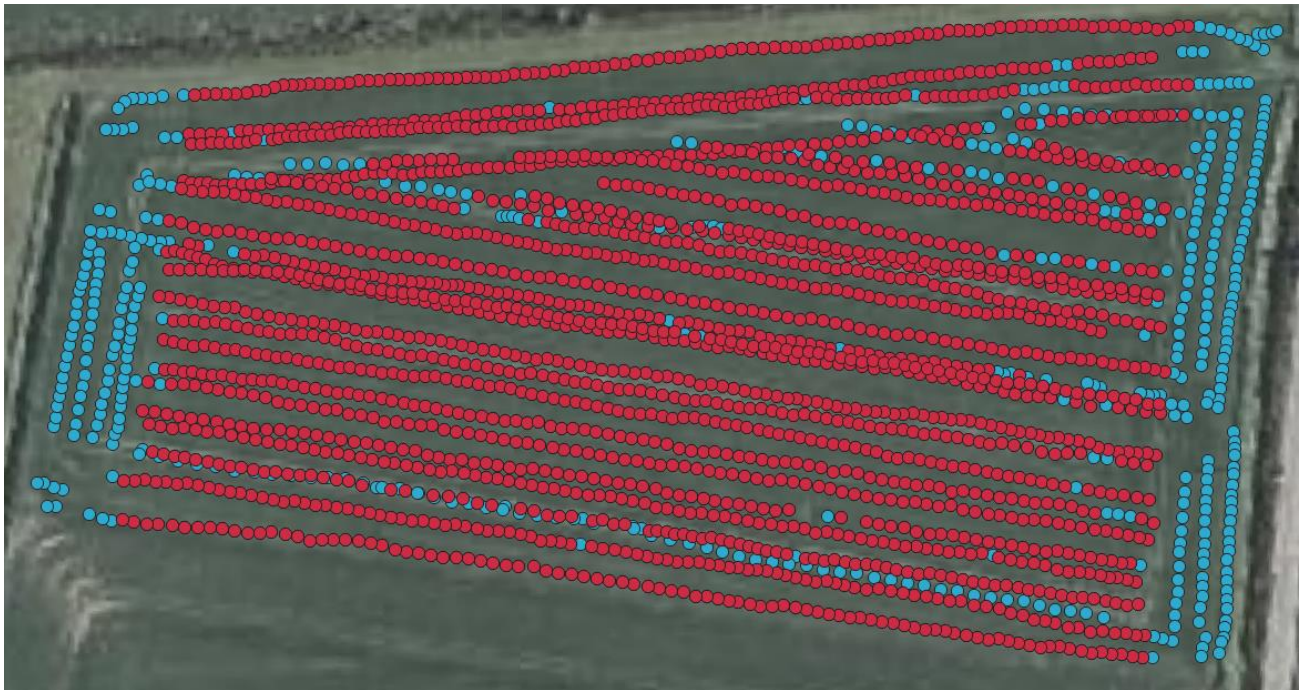

FIGURE 6: COMPARISON OF BEFORE AND AFTER CLEANING THE YIELD MONITOR DATA. BLUE IS BEFORE AND RED IS AFTER.

5. CHECK AVERAGE YIELD FOR FIELD PARTS

Sometimes you would like to know the yield from different parts of the field. Maybe from an experimental strip or from former field subdivisions. The first step is to represent the field parts as polygons. You can either draw those elsewhere and import them (**Add vector layer** as for points) or **New temporary scratch layer** with **Geometry type: Polygon** features, then click the new layer as editable and draw the polygons by left clicking the corner points and finishing the polygon by right clicking.

The average yield is the calculated by Processing toolbox **Join attributes by location (summary)** (Figure 7). **Input layer** = boundaries of the field parts, **Join layer** = Yield map points, **Geometric predicate** = contains, **Fields to summarise** = yieldkg_ha, **Summaries to calculate** = mean, stdev (or another statistic of interest).

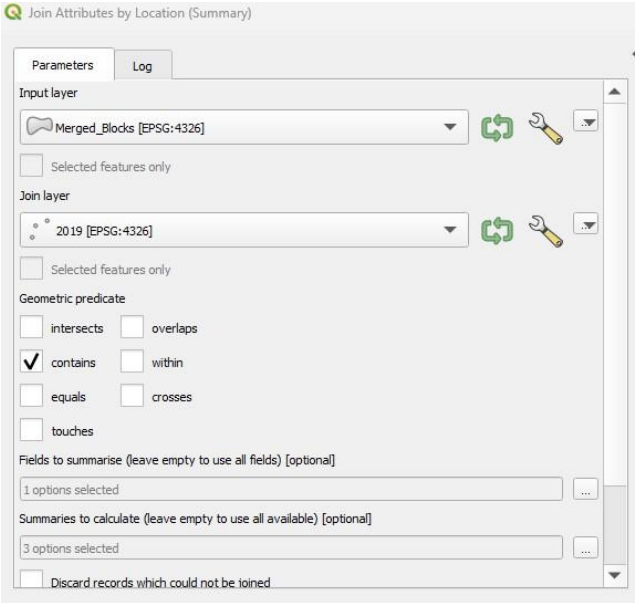

FIGURE 7: JOIN ATTRIBUTES BY LOCATION (SUMMARY) EXAMPLE SNAPSHOT

Visualise the means per exp. unit: double-click the joined layer **Labels**: **Single labels.** click on the ε and insert the following equation *round(yldkg_ha_mean,-1) || '±' || round(yldkg_ha_stddev,-1)* . This presents the mean and standard deviation on top of each polygon.

6. DRAW A RASTERIZED VERSION OF THE YIELD MAP.

The colorized yield points can be valuable for interpretation, but usually we are interested in field regions. Therefore a rasterized map, which averages the points over an area can be helpful. We present two approaches to this, one is based on the QGIS base package interpolation and the other is based on a package called SmartMap.

For basic rasterization, you can use the interpolation tools found under: **Processing** > **Toolbox … Interpolation** to the search bar > **TIN Interpolation** (Figure 8).

Choose the cleaned yield points for the vector layer, choose the yield for the attribute and press the green plus sign. Choose the extent from the yield map layer. output raster size rows – depends on the field size. For a 2-3 ha, then 20 - 40 is a good number. If the field is larger, then higher values must be used. The aim is to get about 10-20 m pixel size. Finally choose a file name in the **Interpolated** menu and save to file and click **Run**.

Figure 7 and 8 shows an example of the TIN interpolation. The colors can be changed in **Symbology**: **Render type > Singleband pseudocolor**. Then **Color ramp > Spectral**. And **Mode > Quantile** and **Classify** and **Apply**.

Create contour lines to highlight yield areas.

In “Processing Tool” search field type “contour”. Open the **Contour tool** under the **GDAL. Input Layer** is the created raster layer. **Interval between contour lines** is in kg/ha and depends how much detail is needed (in Figure 7 below is at 1000 kg/ha).

Add labels: Double click on the contour layer select **Labels** → **Single labels**. Value = ELEV. Click OK.

In Figure 7 below, can be seen the result of the two steps in this section. The yield distribution map with yield contour lines at 1000kg/ha (red is low yield, green is high yield).

For most basic analysis, this would be sufficient, in the following sections we will look at two advanced analyses: management zones based on multi year yield maps and combining yield data with background datasets.

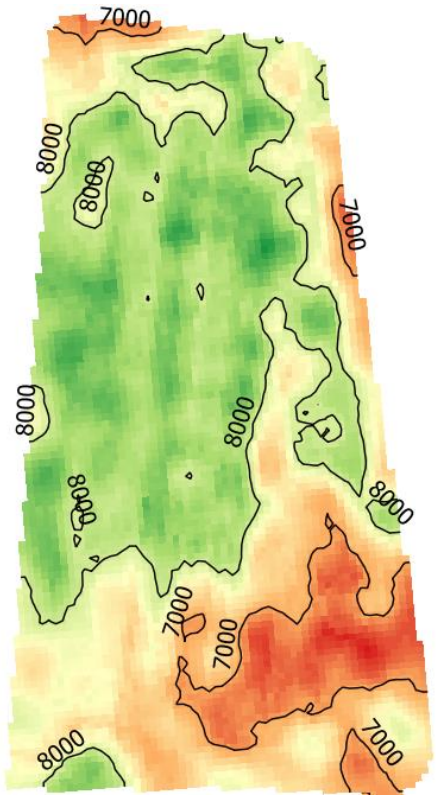

FIGURE 7: RESULT OF THE INTERPOLATION FUNCTION WITH YIELD CONTOUR LINES AT 1000 KG/HA.

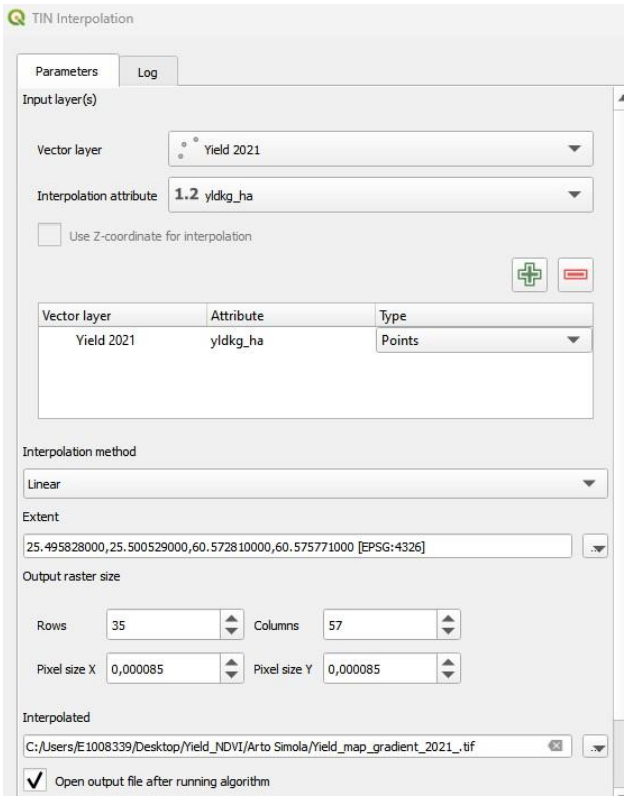

FIGURE 8: TIN INTERPOLATION SNAPSHOT

7. MULTIYEAR YIELD AVERAGE FROM RASTER MAPS AND MANAGEMENT ZONES

One might be interested in the performance of a field over several years and ask “How does the yield look like in the field over x years?”. This might present patterns in the field, which can be further investigated and addressed with a different management.

For this example we calculated the raster map with Smart-Map plugin. Another requirement is to have the layers to be used in this operation in a UTM format and to have a boundary around the yield points. Once the raster layers of each year’s yield have been created, then the average will be calculated.

Save point layers (yield layers) in UTM format.

Choose **Layer** → **Save as**, then choose a file name (FARM_Field_Crop_year_UTM) and the appropriate UTM coordinate system. You can find a map of UTM zones in https://en.wikipedia.org/wiki/Universal_Transverse_Mercator_coordinate_system .

Create a field boundary around the cleaned data.

This is the same approach as for the polygons in section 5. After a polygon has been created, go to polygon settings -> symbology, and choose to have only the outline of the polygon, without the filling. **IMPORTANT! ->** Make sure to choose the same UTM format.

For Finland, for example, there are two UTM zones, UTM zone 34 and 35 (EPSG:32634 and 32635).

The result of this step should look like in Figure 9.

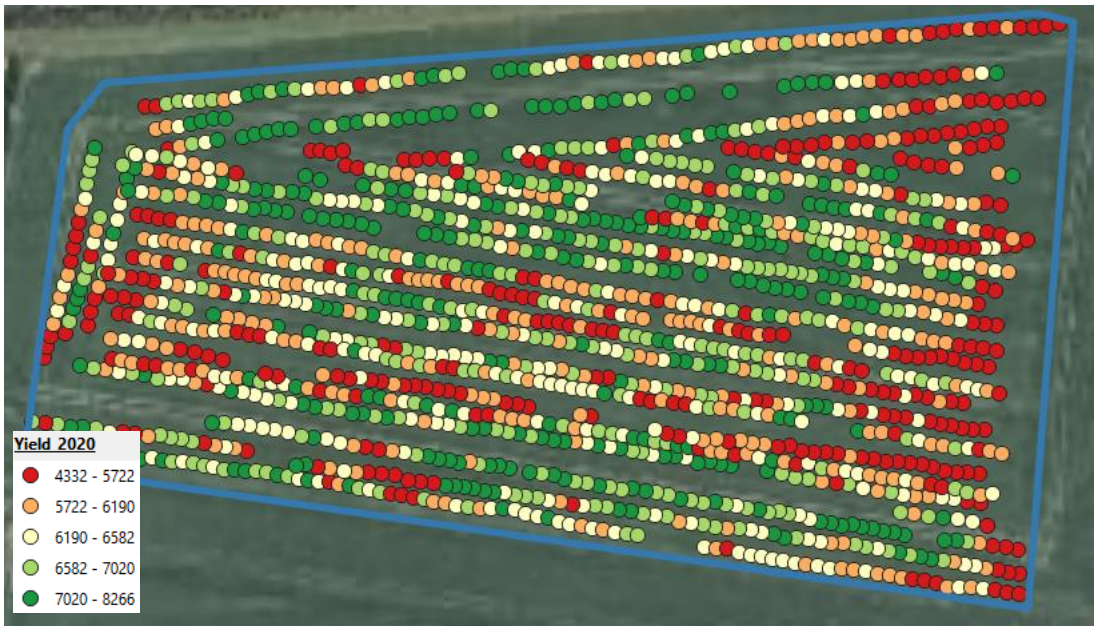

FIGURE 9: CLEANED YIELD POINT LAYER SURROUNDED BY A BOUNDARY.

Install Smart-Map plugin

Go to **Plugins** → **Manage and Install Plugins**. Type in the search “Smart-Map” and then install.

Create a raster map from yield points by kriged interpolation with Smart-Map.

Start the Smart-Map plugin , **Input Layer** is UTM formatted yield points, **Z:**yldkg_ha **Import…** it . Move to **Grid** tab, set **Pixel size** to approximately 2 meters. Select the **Outline polygon** to be the field boundary. Move to **Interpolation** tab. Set the model as a linear model and then under **Ordinary kriging** adjust the **Maximum distance** and click **Calculate..** until the variogram looks linear. (You don’t want to include the leveling off part of the curve to the estimation. Usually something around 20-50 m is appropriate. Adjust the lag between 2-3 to get more points on the line. Click **Interpolate…** to make the map.

Smart-Map creates only one file per project and overwrites that each time. So rename the .kri file when you are satisfied with the result.

Using the data in Figure 9 to create Figure 10, map A was created with the settings: “Maximum Distance”= 30 and the “Lag”= 2, Model= Exponential, R^2=0,916. Map B was created with the settings: “Maximum Distance”= 20, the “Lag”= 2,5, Model= Linear, R^2=0,645.

Map B looks like a better representation of the yield over the field. Therefore, we recommend using a linear model and perhaps a few settings tryouts before deciding which map to use.

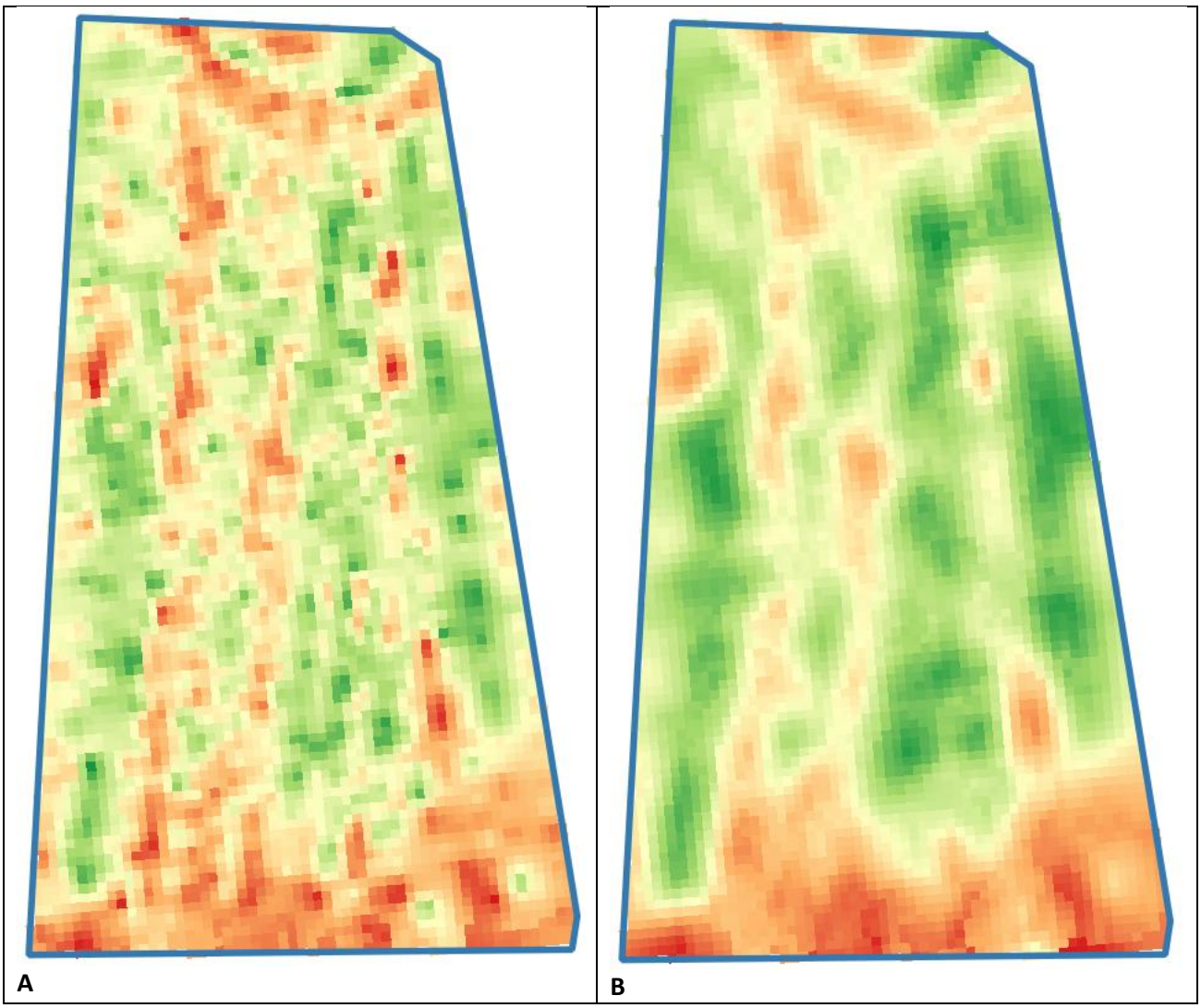

FIGURE 10. EXAMPLE OF 1TWO DIFFERENT SETTING USED IN YIELD MAP CREATION BY KRIGING INTERPOLATION WITH SMART-MAP PLUGIN.

Management zones

Smart map can also classify the points to management zones, representing similar areas. The yield zones are based on the kriged rasters created with the Smart-Map. So, at this point will be needed the .kri files which you saved for the yield map of each year.

Open Smart-Map plugin and go to "management Zones" tab. A list with the interpolated files should be available. If the .kri files you are looking for are not there, click on "Add External Variables" and find the files and add them one by one. They will appear in the "Variables to generate ZM" tab.
Next go to "FPI/NCE Chart" and click "Calculate". Now, go to the last tab "Map Management Zones" and chose the number of zones you are looking for (3-4 zones is usually enough) -> "Generate". A zone map is generated for you.

Calculate multiyear average of the field

To calculate the multiyear average of the yield over the field, we followed the work of Keller et al., (2012) (*Keller, T., Sutter, J. A., Nissen, K., & Rydberg, T. (2012). Using field measurement of saturated soil hydraulic conductivity to detect low-yielding zones in three Swedish fields. Soil and Tillage Research, 124, 68–77. https://doi.org/10.1016/j.still.2012.05.002*)

All these calculations within and between the generated rasters from the kriging, are done with the raster calculator tool (in the Processing Toolbox panel -> type in the search field **Raster Calculator**). In order for this to work, the rasters need to have the same dimensions and location. Therefore generating them with the Smart-Map tool is recommended.

For example, the average yield of a field over 4 years is given by the formula:

$$RY4y = (RYy1 + RYy2 + RYy3 + RYy3)/4$$

Where the RY = relative yield of each yield pixel (or point if there is a point layer map) to the mean yield of the whole field. This will give how much % of the yield of each data pixel differs from the field mean and is calculated with the following formula:

$$RY = \frac{\text{Yield point}}{\text{field mean}} * 100$$

So calculate first the average yield for each yield year, then divide each year raster map with that year, add them together and divide by the number of years.

To calculate the coefficient of variation (CV) which measures the consistency of yields at a specific location over time, expressed as a percentage, the following formula is used:

$$CV = \sqrt{\frac{((RYy1 - RY4y)^2 + (RYy2 - RY4y)^2 + (RYy3 - RY4y)^2 + (RYy4 - RY4y)^2)}{4}} : RY4y$$

Note! A lower CV indicates more stable yields over time, while a higher CV suggests greater variability.

8. COMPARE YIELD WITH NDVI VEGETATION INDEX

NDVI – Normalized Difference Vegetation Index is used to determine the green coverage of a plot of land. This is helpful to see crop emergence and growth in the spring and identify management zones. You can get an NDVI index from https://apps.sentinel-hub.com/eo-browser/ for free, but you need to be a registered user to be able to download it.

To get an NDVI raster that can be used in QGIS, it has to be calculated based on two band rasters that can be downloaded from the sentinel hub. Below is described the download process and how to calculate the NDVI raster, assuming you have an account.

Go to sentinel EO-browser-> Discover->Sentinel-2->Visualize

- Navigate to the desired field
- chose the suitable day in spring based on the least cloud coverage (clear sky) and the presence of difference in crop establishment over the field. One can investigate any period, including autumn, to check cover crops establishment or when the crop starts to mature.
- on the right-hand-side panel -> download image
- set the options as in Figure 16, and tick RAW B04 and B08. ->Download

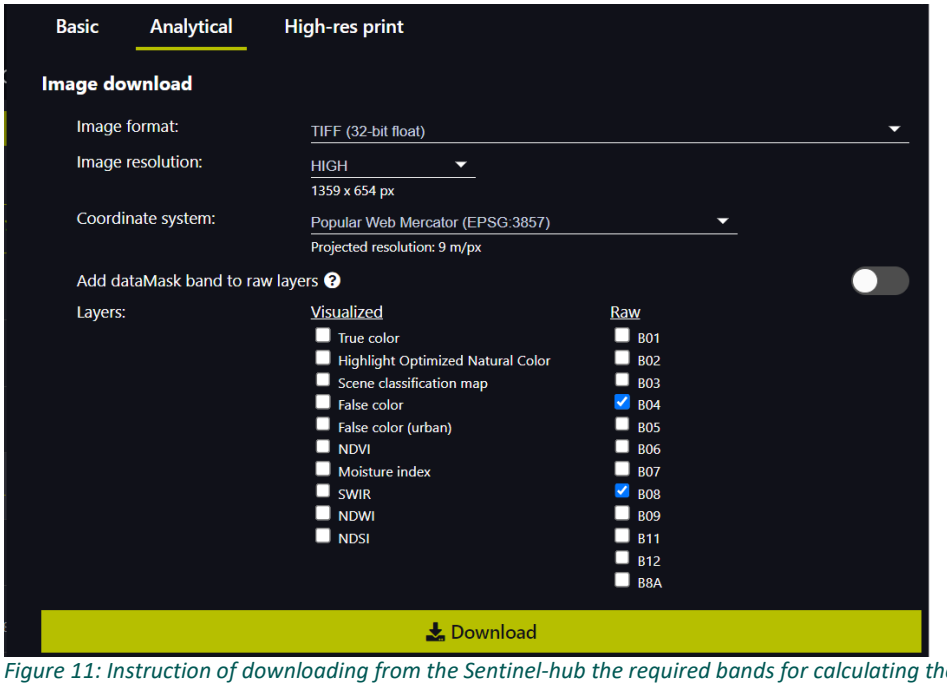

Figure 11: Instruction of downloading from the Sentinel-hub the required bands for calculating the NDVI.

In QGIS

- Open the downloaded file in QGIS
- In Processing Toolbox search field type: “Raster Calculator” -> open the first raster calculator

- Input layer A = band 8 raster (….B08_(Raw) -> Raster band = Band 1
- Input layer B = band 4 raster (….B04_(Raw) -> Raster band = Band 1

- Calculate NDVI from the two raster bands
  - Calculation in gdalnumeric syntax….. = (A-B)/(A+B) this is based on the band formula (B8-B4)/(B8+B4)
  - Calculated -> click dropdown arrow ->Save to file (save it as .tif format) ->Run
- To investigate the raster, use "Greens" for the "Colour ramp". Usually the range between 0.4-0.8 NDVI is the most interesting, so you can select cut-off min-max values for the greenness.

A topographic wetness index (DTW, depth to water). Instructions for Finland.

Figure 12 shows what the DTW of an area looks like.

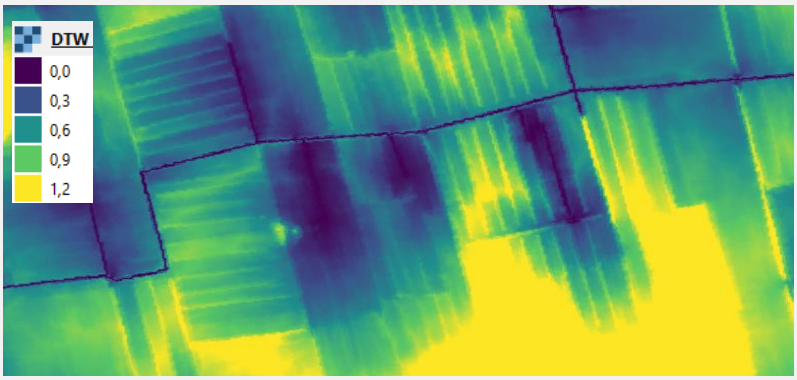

FIGURE 12: DTW IMAGE OF AN AREA. PURPLE = HIGH WATER TABLE (METERS), YELLOW = DEEP WATER TABLE (METERS.

- In Finland DTW maps are publicly available, interna tionally they are in creasi ngly being made available.
- Go to https://paituli.csc.fi/do wnload.htm l
- Select dataset as in Figure 15

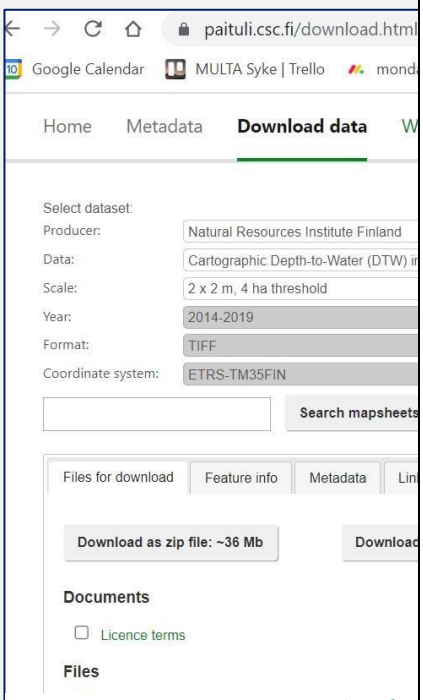

Figure 15: Downloading DTW data from www.paituli.csc.fi example snapshot

- **Important!** click on "Metadata" and read the info. The "4 ha threshold" is relevant to dry summer conditions, while the "1 ha threshold" is to wet spring conditions.
- Select the map area, it is available as map grids.
- Download

Insert the raster file in QGIS

- Investigate the raster: Double click on file->Symbology->
  - render type = Singleband pseudocolor
  - Color ramp = Viridis
  - Adjust the Min at 0 and play with the Max until you start to see differences of color in the area of interest. **Note**: every time you change the min or max values press the "Classify" button and then "Apply" to visualize the change on the map.

- If you want to create correlations between Yield and DTW or NDVI follow the steps in section 9.

9. CHECK CORRELATIONS

Through correlations, it is possible to understand whether the field conditions such as NDVI, DTW (or other) could have an influence on the yield.

The steps in this section are the same for NDVI as for DTW. DTW is used here as an example.

This action will create a new raster with DTW values that are determined by the yield registered in that specific place in the field. In this way, it is possible to correlate the yield with the DTW. Doing this operation will also tell more precisely what the highest and lowest water table distances from the soil's surface across the field are.

IN QGIS : SAMPLE THE RASTER LAYER UNDER THE YIELD MAP POINTS

Sample the raster by Yield point locations. Choose **Sample raster values** from the processing toolbox. **Input layer** = yield points , **Raster layer** = NDVI or DTW, **Output column prefix** = name for new columns. **Sampled** = save to file. You can visualize the outcome from the **Open Attribute table** .

If the sampling is successful, right click the new layer and choose to **Export** it as .csv file for opening in Excel.

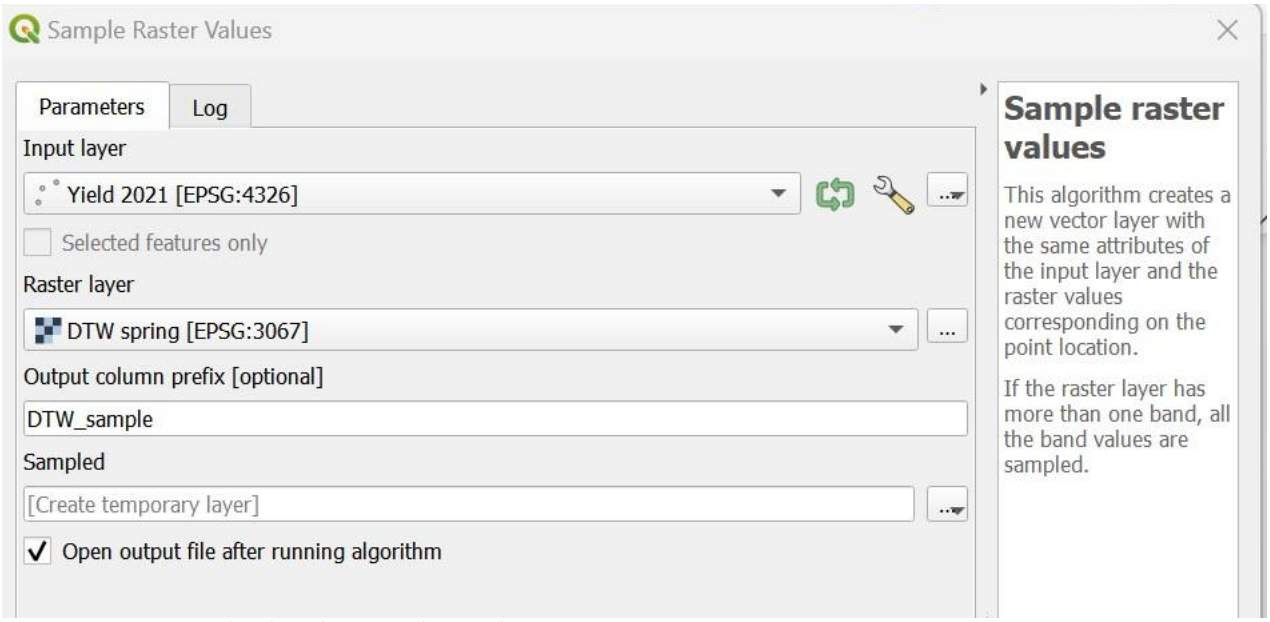

Figure 13: Raster sampling by value example snapshot.

IN EXCEL : CREATE A CORRELATION BASED ON THE SAMPLED POINTS

Create a correlation matrix between yield and and the raster layer DTW. It should look something like in the Figure 14

- Open the .csv file in Excel
- Insert -> Scatter plot
- Insert trendline and correlation factor R^2

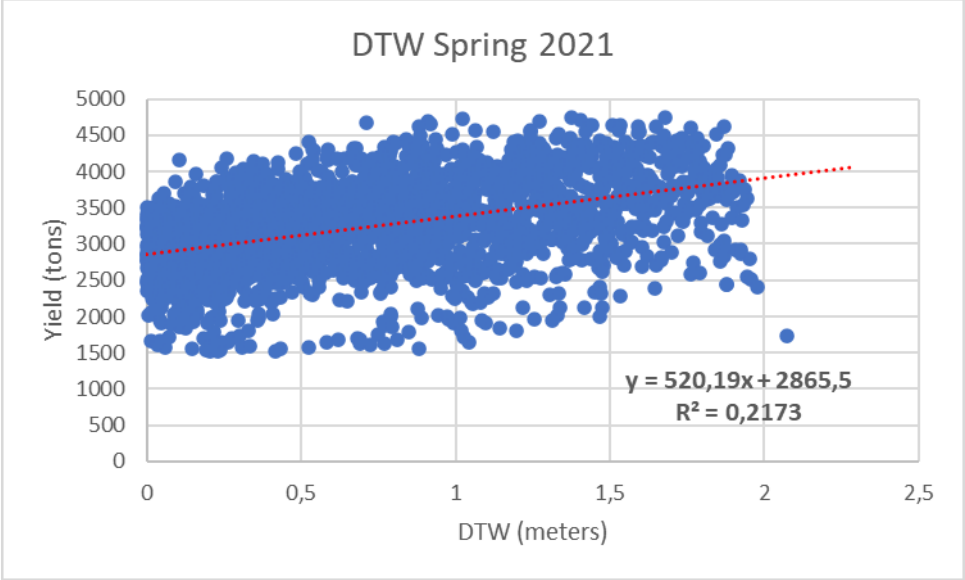

Figure 14: Correlation plot between yield and DTW. The red dotted line is the trendline. The R2 value tells that about 20% of the yield could be influenced by the water depth in spring. The deeper the depth of the water table in the spring, the higher the yield.

IN CONCLUSION

Yield mapping is a powerful tool for precision agriculture. The access to open source GIS software, freely available datasets (Sentinel NDVI etc.) and packages such as Smart-Map make it applicable to farmers, extension personnel and anybody interested in improving agriculture. At the same time, the amount of data and the various data sources can make it a difficult topic to approach. We hope that this workflow and the examples presented here will make the learning path to mastering QGIS analysis a bit easier.