



## Practical 5

### Data management for mapping

## Aim of practical

---

This practical aims to introduce the main validation steps during data cleaning to ensure a high quality of survey data. The practical focuses on specific procedures for the validation of mapping data. Moreover, you will learn how to transform individual level data to group level data that can be presented on maps and how to merge datasets.

## Key learning skills

---

In this practical, you will learn the following skills:

- Performing structural checks on the database
- Verifying the plausibility of data
- Performing logical checks
- Verifying the coordinates of study locations
- Transforming individual level data into group level data (aggregation)
- Merging datasets



This work is licensed under a [Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License](#). This means that users are free to copy and share this material with others. Requests for creating new derivatives should be sent to the primary author ([www.thiswormyworld.org](http://www.thiswormyworld.org)).

## Outline

---

|  |          |
|--|----------|
| <b>How to ensure good quality of data .....</b>        | <b>4</b> |
| <b>Data set description .....</b>                      | <b>5</b> |
| <b>Practical .....</b>                                 | <b>6</b> |
| 1. Developing a data cleaning plan .....               | 6        |
| 2. Structural data checks .....                        | 6        |
| 2.1. Format of the database .....                      | 6        |
| 2.2. Removing duplicates.....                          | 9        |
| 2.3. Coding of variables.....                          | 10       |
| 3. Verifying the plausibility of data.....             | 12       |
| 3.1. Coding of variables.....                          | 12       |
| 3.2. Using a filter to detect outliers .....           | 12       |
| 4. Logical data checks.....                            | 14       |
| 4.1. Cross-tabulations .....                           | 14       |
| 4.2. Formulas.....                                     | 16       |
| 5. Verifying coordinates of mapping data.....          | 16       |
| 6. Preparing data for mapping .....                    | 21       |
| 6.1. Collapse numerical values by study location ..... | 22       |
| 6.2. Creating new variables using formulas .....       | 25       |
| 6.3. Merged variables to collapse data.....            | 27       |

## How to ensure good quality of data

---

To obtain reliable results from a study, a good quality of underlying data needs to be guaranteed. For this purpose, in addition to a good questionnaire design, checks need to be performed at several steps during data processing:

1. Checks during data collection - using data collection tools (e.g. tablets)

2. Checks during data entry - using data entry surfaces (e.g. EpiData, Access, etc.)

3. Checks during data cleaning - using Excel or other statistical software

consistent processing of data



This practical will mainly focus on the validation steps that should be performed during data cleaning. These include three types of checks:

- Structural checks: e.g. unique ID, duplicates, format
- Validation of plausibility - possible range of values e.g. min/max expected age
- Logical checks - inconsistencies in answers e.g. occupation/age

The practical will be conducted using Microsoft Excel and QGIS software. The introduced concepts also apply when cleaning data with other statistical software, even though the exact procedures will differ.

The practical was developed based on Excel 2007. The exact layout and location of functions will slightly differ in other Excel versions.

## Data set description

---

| File name                      | Description  | Variables/Source   |
|--------------------------------|--|--|
| Kenya_epidemiological_data.xls | Excel file containing data from an epidemiological survey in Kenya | child_id<br>school_name<br>school_id<br>district_id<br>age<br>sex<br>stool<br>hookworm_epg<br>ascaris_epg<br>trichuris_epg<br>smansoni_epg<br>hookworm_inf<br>ascaris_inf<br>trichuris_inf<br>anysth_inf<br>smansoni_inf<br>dewormed |
| Kenya_epidemiological_dict.xls | Excel file containing description of variables and coding          |  |
| Kenya_school_location.csv      | Text file containing the location information of each school       | school_id<br>school_name<br>latitude<br>longitude<br>district_name   |
| Kenya_school_dict.xls          | Excel file containing description of variables and coding          |  |
| Kenya_admin_shp                | District boundaries for Kenya, ArcGIS shapefile                    | Digital Chart of the World   |

In this training, you will work with data from surveys for soil-transmitted helminths (STH) and schistosomiasis in Western, Nyanza, and Rift Valley Province in Kenya in 2008-2010. The surveys were conducted in 41 schools and included 3,877 school children. Each child was asked to provide a stool sample which was examined for the presence and number of eggs using the Kato-Katz method. The slides were double-read and the mean egg count was used to calculate the eggs/g faeces (epg).

## Practical

---

### 1. Developing a data cleaning plan

---

Generally, data should be processed consistently. Therefore, based on the characteristics of the variables in the dataset, a list of useful checks should be created before manipulating the data. These can be adapted if additional errors appear at a later stage that were not considered when developing the list. First, you will need to make yourself familiar with the dataset, and ensure that you understood the meaning of all variables.

- Open the “Kenya\_epidemiological\_data.xls” in Microsoft Excel. Additionally open the corresponding dictionary “Kenya\_epidemiological\_dict.xls”, which contains a description of all variables and information about their coding. Ensure that you understand the meaning of the variables and their format. You should make note of the number of entries in your database, as you will need to keep track of any changes e.g. when you remove duplicates at a later stage.



*Note: Generally, you should keep track of any modifications that you are doing on a dataset (deleting records, recoding etc.). Some statistical software will provide options such as do-files or log-files. You can also do this simply by creating a text document (e.g. in Word) listing the changes and dates when the changes were made. You should also note the new number of entries and the reason whenever you remove records to ensure that no data gets just “lost”.*

- To develop a data cleaning plan you should consider the following questions:
  - ✓ Which variables are string variables, which are numerical?
  - ✓ What is the expected range of continuous variables?
  - ✓ Which values are possible for categorical variables?



Based on the introduced types of checks (structural, plausibility, logical), write down a short outline of checks that you consider as useful to detect errors in this dataset.

### 2. Structural data checks

---

#### 2.1. Format of the database

- Before manipulating your dataset you should always save a copy so you can come back to a previous version if anything goes wrong. It is useful to include

numbers or dates to identify the different versions. Therefore save the version you are now working on as *Kenya\_epidemiological\_data\_2.xls*.

Remember the required format of the database:

- ✓ The database should contain one head row with informative variable names followed by the data
- ✓ No empty rows (rows without any data, or rows to subdivide the dataset) should be included
- Ensure that these criteria are met and if necessary remove all rows that do not contain any data or the variable names. You can freeze the first row of your dataset; this will allow you to see the variable names when scrolling down the data. Go to the **View** tab, click **Freeze Panes** and choose **Freeze Top Row**.

| child_id          | school_nai | school_id | district_na | district_id | age | sex | stool | hookworm | ascaris_eg_trich | th.inf | smansonii | dewormed |
|-------------------|------------|-----------|-------------|-------------|-----|-----|-------|----------|------------------|--------|-----------|----------|
| 70107028 TEREBEN  | 1190       | BARINGO   | 701         | 9           | F   | 1   | 0     | 0        |                  | 0      | 0         | 0        |
| 70105018 NGINYANI | 1203       | BARINGO   | 701         | 12          | F   | 1   | 0     | 0        |                  | 0      | 0         | 0        |
| 70105055 NGINYANI | 1203       | BARINGO   | 701         | 11          | F   | 1   | 0     | 0        |                  | 0      | 0         | 0        |
| 70107038 TEREBEN  | 1190       | BARINGO   | 701         | 10          | F   | 1   | 0     | 0        |                  | 0      | 0         | 1        |
| 70105029 NGINYANI | 1203       | BARINGO   | 701         | 7           | F   | 1   | 0     | 0        |                  | 0      | 0         | 0        |
| 70105085 NGINYANI | 1203       | BARINGO   | 701         | 14          | F   | 1   | 0     | 0        |                  | 0      | 0         | 0        |
| 70405040 NGINYANI | 1203       | BARINGO   | 701         | 7           | M   | 1   | 0     | 0        |                  | 0      | 0         | 0        |

When working with anonymised data we need to be able to identify each record by a unique identifier. In the Kenya dataset, a unique identifier has been created previously (*child\_id*).

### Creating unique identifiers

An easy and useful way to create a unique identifier is to combine several identifier numbers to obtain a unique number.

For example when combining the *district\_id*: 702, *school\_id*: 654 and *child\_id*: 032 the unique\_id would be: 702654032.

This is convenient as the unique identifier directly contains all information about where the participant was coming from.

- Verify that *child\_id* identifies the participants uniquely by using the conditional formatting function in Excel.

## GIS training for Neglected Tropical Diseases mapping Data management for mapping using QGIS 2.0.1-Dufour

- First select the ‘child\_id’ column, then click **Conditional Formatting** in the **Home** tab. You can use the preset rule for “Duplicate values” in “Highlight cells rules”.

The screenshot shows a Microsoft Excel spreadsheet titled "Kenya\_epidemiological\_data [Compatibility Mode] - Microsoft Excel". The data consists of 17 rows of epidemiological information. The columns include child\_id, school\_name, school\_id, district\_name, district\_id, age, sex, stool, hookworm, ascaris\_ept, trichuris\_e, smansoni\_e, ascaris\_int, trichuris\_ir, anysth\_inf, smansoni\_deworm. The "Conditional Formatting" ribbon tab is selected, and the "Highlight Cells Rules" dropdown is open, with "Duplicate Values..." highlighted in yellow.

- Chose a highlight format from the dropdown, for example *Light red fill with Dark red text*. By clicking **OK** the rule is applied to the selected cells.
- To filter the highlighted cells you can choose ‘Filter by colour’ and then the highlighted cell colour. First, you need to enable the option **Filter** under **Sort & Filter** tools.

The screenshot shows the same Microsoft Excel spreadsheet. The "Sort & Filter" dropdown menu is open, and the "Filter" option is circled in red.

The screenshot shows the "Filter by Color" dropdown menu open, displaying options for "No Fill", "Red", and "Automatic".

## GIS training for Neglected Tropical Diseases mapping Data management for mapping using QGIS 2.0.1-Dufour

- Inactivate the filter (by clicking on *Filter*, under **Sort & Filter**) and then formatting (by clicking on **Clear rules > Clear Rules From Entire Sheet**).

The screenshot shows a Microsoft Excel spreadsheet titled "Kenya\_epidemiological\_data [Compatibility Mode] - Microsoft Excel". The ribbon at the top has the "Conditional Formatting" tab selected. A context menu is open over a range of data starting from cell A1. The menu path "Conditional Formatting > Clear Rules" is highlighted, and the option "Clear Rules from Entire Sheet" is also highlighted with a red box.

### 2.2. Removing duplicates

Sometimes ID duplicates are caused by errors in the ID number, however, it is also possible that an entire record is entered twice. To ensure that each study participant is included only once, we need to search for any possible duplicates.

- Delete duplicates using the **Remove Duplicates** function in the **Data** tab. Select all columns to delete duplicate values (all columns are selected already by default).

The screenshot shows the Microsoft Excel ribbon with the "Data" tab selected. In the "Sort & Filter" group, the "Remove Duplicates" button is circled in red. A "Remove Duplicates" dialog box is open over the data range, showing the "Columns" section where all columns (child\_id, school\_name, school\_id, district\_name, district\_id, age) are selected for removal. The "OK" button is visible at the bottom right of the dialog.

 Note: In older Excel versions the **Remove Duplicates** function does not exist and duplicates can be removed using the advanced filter function in the Data tab. Here you have to select the list range (which is your entire dataset) and your criteria range (which

are all variables). Then you have to tick **Unique records only** and OK. You can copy the filtered data then to a new tab or spreadsheet. You can also use the conditional formatting to highlight the duplicates and then remove them by hand.



Compare the number of previous entries to the number of entries after removal of duplicates. Were any duplicates removed?



**Note:** In cases where only the ID number is duplicated but the rest of the data differs, you will have to go back to the original files to find out where the error occurred.

### 2.3. Coding of variables

Categorical variables can be coded as string (e.g. M/F) or numbers (e.g. 1/2). For computation purposes, it is often more convenient to work with numbers than strings. Therefore, you might want to consider recoding some of the variables. In this dataset, sex was coded as M for male and F for female. However, we would like to code the categories using numerical values (1 for male and 2 for female).

- In Excel, you can recode variables using the **Replace** function. Select the ‘sex’ column: in the **Home** tab go to **Find & Select** (or click **Ctrl+F**, simultaneously), then to **Replace**. In the field **Find what** type ‘M’ and in **Replace with** type ‘1’. Then click on **Replace All**. Excel will give a message stating how many values have been replaced. Repeat the steps to replace F with 2.

| district | id  | age | sex | stool | hookworm | ascaris_eg | trichuris_eg | mansoni | hookworm | ascaris_in | trichuris_in | anysth_in | mansoni_in | dewormed |
|----------|-----|-----|-----|-------|----------|------------|--------------|---------|----------|------------|--------------|-----------|------------|----------|
| IGO      | 701 | 9   | F   |       | 1        | 0          | 0            | 0       | 0        | 0          | 0            | 0         | 0          | 0        |
| IGO      | 701 | 12  | F   |       | 1        | 0          | 0            | 0       | 0        | 0          | 0            | 0         | 0          | 0        |
| IGO      | 701 | 11  | F   |       | 1        | 0          | 0            | 0       | 0        | 0          | 0            | 0         | 0          | 0        |
| IGO      | 701 | 10  | F   |       | 1        | 0          | 0            | 0       | 0        | 0          | 0            | 0         | 0          | 1        |
| IGO      | 701 | 7   | F   |       | 0        | 0          | 0            | 0       | 0        | 0          | 0            | 0         | 0          | 0        |
| IGO      | 701 | 14  | F   |       | 0        | 0          | 0            | 0       | 0        | 0          | 0            | 0         | 0          | 0        |
| IGO      | 701 | 7   | M   |       | 0        | 0          | 0            | 0       | 0        | 0          | 0            | 0         | 0          | 0        |
| IGO      | 701 | 10  | F   |       | 0        | 0          | 0            | 0       | 0        | 0          | 0            | 0         | 0          | 0        |
| IGO      | 701 | 8   | F   |       | 0        | 0          | 0            | 0       | 0        | 0          | 0            | 0         | 0          | 0        |
| IGO      | 701 | 12  | M   |       | 0        | 0          | 0            | 0       | 0        | 0          | 0            | 0         | 0          | 0        |
| IGO      | 701 | 10  | M   |       | 0        | 0          | 0            | 0       | 0        | 0          | 0            | 0         | 0          | 0        |
| IGO      | 701 | 10  | F   |       | 0        | 0          | 0            | 0       | 0        | 0          | 0            | 0         | 0          | 0        |
| IGO      | 701 | 10  | M   |       | 0        | 0          | 0            | 0       | 0        | 0          | 0            | 0         | 0          | 0        |
| IGO      | 701 | 9   | M   |       | 1        | 0          | 0            | 0       | 0        | 0          | 0            | 0         | 0          | 0        |
| IGO      | 701 | 6   | F   |       | 1        | 0          | 0            | 0       | 0        | 0          | 0            | 0         | 0          | 0        |

- To check whether all values have been replaced accordingly you can use the filter function. The drop down menu will list all possible values that are now included in the column.
- Don’t forget to update the coding in the data dictionary.



*Note: Ensure that coding is applied consistently throughout the dataset and that a dictionary with all values and labels is created.*

There are several possibilities to encode missing values in a dataset e.g. using numerical values as 999, string as N/A, or fields are just left blank. The coding of missing values can also depend on the statistical software used for analysis. STATA for example encodes missing values as “.”

- In Excel, probably the easiest way to find out about missing values is to use the filter option drop down menu.



Are there any missing values in your dataset? How are missing values encoded?

#### Coding of variables

- **Decide which answer options to provide**

You should test the answers previously to ensure that all relevant options are included. You might also allow to additionally writing in text other answer options that were not included.

- **Be consistent with your coding**

If you used 0/1 for No/Yes you should apply the coding to all questions with a Yes/No answer option and not change to for example 1/2.

- **Be clear about your coding**

Answer options with their codes should be provided in the questionnaire; additionally you should create a dictionary so that this information does not get lost at later stages of data processing and analysis.

- **Decide how to code missing values**

There might be different reasons for a missing value. The question might not be applicable for the participant, the answer might be unknown, or recording the answer was forgotten. Ensure not to use a missing code such as 999 for continuous data as this could be a possible value e.g. for epg, and it could not be detected as missing

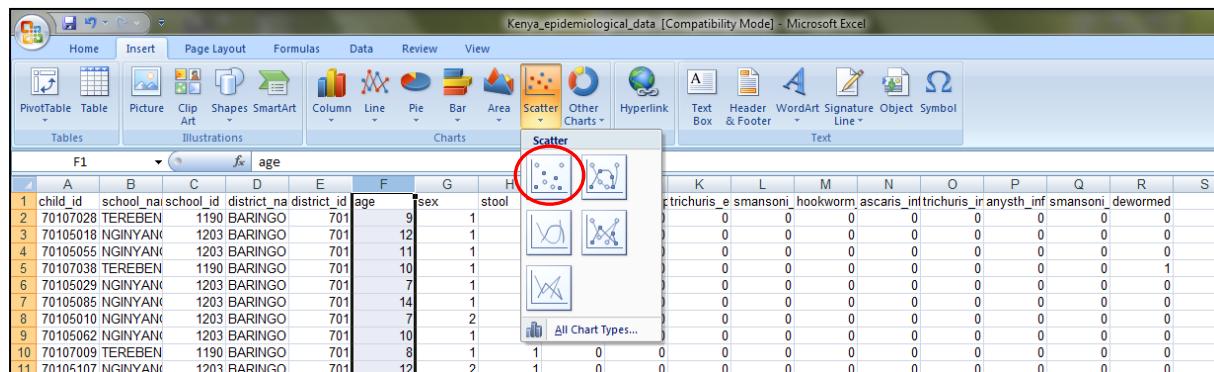
### 3. Verifying the plausibility of data

#### 3.1. Coding of variables

By verifying the plausibility of data, you ensure that all entered values are in the possible range. For example, if you are working on data from school children a certain age range can be expected. If a school child has an age of 75, it is very likely to be an error. Also, for categorical data where certain answer options were provided in the questionnaire (such as values of 1, 2 & 3), a value different from these (such as 4) is an error.

You can first explore the data using the scatter plot function to visualise the data. This is just to get a first impression on how erroneous each variable is. We will correct the errors at a later stage.

- Create a scatterplot for age. Therefore, select the ‘age’ column, then choose **Scatter** from the **Insert** tab. Choose a scatterplot displaying the data points only.



 Do you observe any age outlier in the data?

- You can now also check the other variables
- Delete the graphs from your spreadsheet

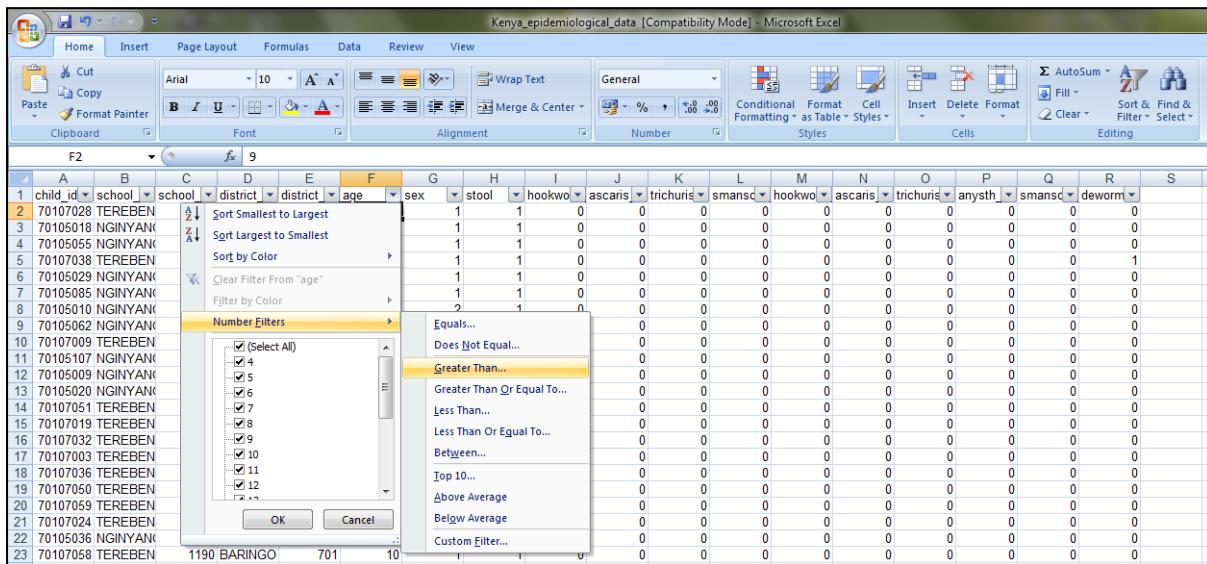
#### 3.2. Using a filter to detect outliers

Scatterplots are a fast way to become aware of any data outliers. However, to correct these errors we will need to localise them exactly. To identify the outliers

GIS training for Neglected Tropical Diseases mapping  
*Data management for mapping using QGIS 2.0.1-Dufour*

in the spreadsheet, we will now use the filter function that was already introduced to detect duplicates.

- In the Home tab click on Sort & Filter then choose Filter. Open the drop down menu in the ‘age’ column and click on Number Filters, select the Greater than option and type ‘20’ in the “is greater than” field.



| child_id | school   | school   | district | district | age | sex | stool | hookwo | ascaris | trichuris | smansc | hookwo | ascaris | trichuris | anysth | smansc | deworm | S |
|----------|----------|----------|----------|----------|-----|-----|-------|--------|---------|-----------|--------|--------|---------|-----------|--------|--------|--------|---|
| 1        | 70107028 | TEREBEN  |          |          | 1   | 1   | 1     | 0      | 0       | 0         | 0      | 0      | 0       | 0         | 0      | 0      | 0      | 0 |
| 2        | 70105018 | NGINYANK |          |          | 1   | 1   | 1     | 0      | 0       | 0         | 0      | 0      | 0       | 0         | 0      | 0      | 0      | 0 |
| 3        | 70105055 | NGINYANK |          |          | 1   | 1   | 1     | 0      | 0       | 0         | 0      | 0      | 0       | 0         | 0      | 0      | 0      | 0 |
| 4        | 70107038 | TEREBEN  |          |          | 1   | 1   | 1     | 0      | 0       | 0         | 0      | 0      | 0       | 0         | 0      | 0      | 0      | 0 |
| 5        | 70105029 | NGINYANK |          |          | 1   | 1   | 1     | 0      | 0       | 0         | 0      | 0      | 0       | 0         | 0      | 0      | 0      | 1 |
| 6        | 70105085 | NGINYANK |          |          | 1   | 1   | 1     | 0      | 0       | 0         | 0      | 0      | 0       | 0         | 0      | 0      | 0      | 0 |
| 7        | 70105010 | NGINYANK |          |          | 1   | 1   | 1     | 0      | 0       | 0         | 0      | 0      | 0       | 0         | 0      | 0      | 0      | 0 |
| 8        | 70105062 | NGINYANK |          |          | 1   | 1   | 1     | 0      | 0       | 0         | 0      | 0      | 0       | 0         | 0      | 0      | 0      | 0 |
| 9        | 70105009 | NGINYANK |          |          | 2   | 1   | 0     | 0      | 0       | 0         | 0      | 0      | 0       | 0         | 0      | 0      | 0      | 0 |
| 10       | 70107009 | TEREBEN  |          |          |     |     |       |        |         |           |        |        |         |           |        |        |        |   |
| 11       | 70105107 | NGINYANK |          |          |     |     |       |        |         |           |        |        |         |           |        |        |        |   |
| 12       | 70105020 | NGINYANK |          |          |     |     |       |        |         |           |        |        |         |           |        |        |        |   |
| 13       | 70107051 | TEREBEN  |          |          |     |     |       |        |         |           |        |        |         |           |        |        |        |   |
| 14       | 70107019 | TEREBEN  |          |          |     |     |       |        |         |           |        |        |         |           |        |        |        |   |
| 15       | 70107032 | TEREBEN  |          |          |     |     |       |        |         |           |        |        |         |           |        |        |        |   |
| 16       | 70107003 | TEREBEN  |          |          |     |     |       |        |         |           |        |        |         |           |        |        |        |   |
| 17       | 70107036 | TEREBEN  |          |          |     |     |       |        |         |           |        |        |         |           |        |        |        |   |
| 18       | 70107050 | TEREBEN  |          |          |     |     |       |        |         |           |        |        |         |           |        |        |        |   |
| 19       | 70107059 | TEREBEN  |          |          |     |     |       |        |         |           |        |        |         |           |        |        |        |   |
| 20       | 70107059 | TEREBEN  |          |          |     |     |       |        |         |           |        |        |         |           |        |        |        |   |
| 21       | 70107024 | TEREBEN  |          |          |     |     |       |        |         |           |        |        |         |           |        |        |        |   |
| 22       | 70105036 | NGINYANK |          |          |     |     |       |        |         |           |        |        |         |           |        |        |        |   |
| 23       | 70107058 | TEREBEN  |          |          |     |     |       |        |         |           |        |        |         |           |        |        |        |   |

- Now delete the values of the outliers (replace them as missing values). In practice, you might want to go back to the original records (cross-checking against field forms) to verify these entries.
- Remove the filter from the dataset.
- Remember to make a note about the changes in your log-file



This can be expanded to perform more sophisticated checks if necessary (e.g. the age range in a certain class) by creating new variables that takes the value 1 if the entry has an error or 0 if the values are correct and then calculate their values with a formula. More details will be provided in the Logical Checks section.

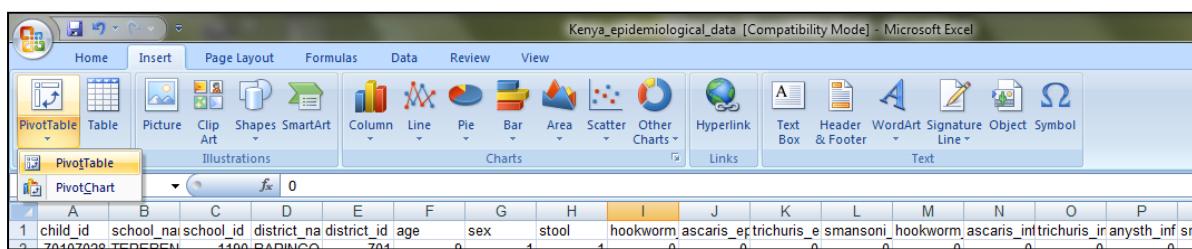
Another option is to use a data validation function in Excel. Here you can set up a range of possible values (even before entering the data) and define an error message. To avoid entry errors you could also include a drop-down list or an allowed range of values during data entry.

## 4. Logical data checks

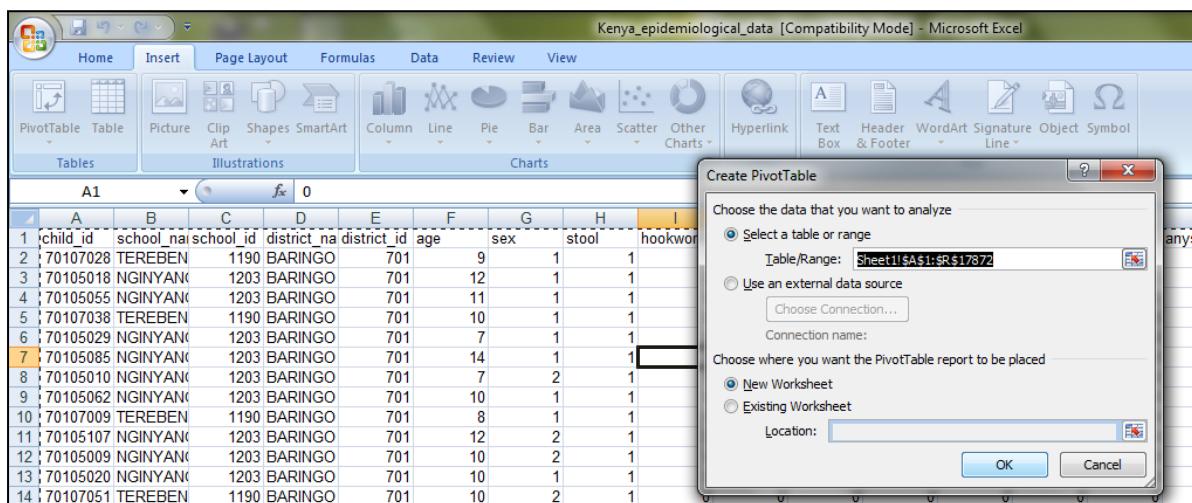
### 4.1. Cross-tabulations

To detect any inconsistencies in the data, logical checks need to be performed. In this dataset, for example, infection status can only be assessed for children who have provided a stool sample. These inconsistencies can be detected using cross-tabulations.

- In Excel, such tables can be generated using the **PivotTable** function. You can find the function in the **Insert** tab.



- Ensure that the entire dataset is selected (this should be done by default)
- Tick **New Worksheet** option to tell Excel that you want to place the table in a new sheet.



- An empty table is inserted in a new sheet and a window will open on the right hand side named PIVOT TABLE FIELD LIST.

## GIS training for Neglected Tropical Diseases mapping Data management for mapping using QGIS 2.0.1-Dufour

The screenshot shows the Microsoft Excel interface with the PivotTable Tools ribbon selected. The PivotTable Field List pane on the right lists various fields: child\_id, school\_name, school\_id, district\_name, district\_id, age, sex, stool, hookworm\_egg, ascari\_egg, trichuris\_egg, and manson\_egg. A red arrow points from the 'Row Labels' section of the PivotTable Field List to the 'stool' field in the list.

- Now click on the variable ‘stool’ in the Field List and draw it into the “Row labels” field.
- Click on *anysth\_inf* and draw it into the “Column labels” and “Σ Values” field.
- To include the count of observations in the table you might need to change the value field settings to count.

The screenshot shows the Microsoft Excel interface with the PivotTable Tools ribbon selected. The PivotTable Field List pane on the right shows the 'stool' field selected. A red box highlights the 'Value Field Settings...' button at the bottom of the list.

|   | B                 | C | D | E | F    | G | H           | I | J | K | L | M | N | O |   |
|---|-------------------|---|---|---|------|---|-------------|---|---|---|---|---|---|---|---|
| 1 | A                 | B | C | D | E    | F | G           | H | I | J | K | L | M | N | O |
| 2 |                   |   |   |   |      |   |             |   |   |   |   |   |   |   |   |
| 3 | Sum of anysth_inf |   |   |   |      |   |             |   |   |   |   |   |   |   |   |
| 4 | stool             |   |   | 0 |      | 1 | Grand Total |   |   |   |   |   |   |   |   |
| 5 | 0                 |   | 0 |   |      | 0 |             |   |   |   |   |   |   |   |   |
| 6 | 1                 |   | 0 |   | 3525 |   | 3525        |   |   |   |   |   |   |   |   |
| 7 | Grand Total       |   | 0 |   | 3525 |   | 3525        |   |   |   |   |   |   |   |   |

Do you observe any inconsistencies? Are there any children with infection results that haven't provided stool samples according to the data?

## 4.2. Formulas

To exactly identify the study participant with the inconsistency in the data, you can create a new variable which will take a specific value if the inconsistency occurs (e.g.1). This will allow you to correct the error in the specific entry. You will need to apply a formula to the cells.

- Create a new variable called ‘check’
- We want the new variable to have the value 1 if a STH infection was assessed but no stool sample was provided and 0 if the infection was assessed and a stool sample was provided.
- In the first cell type the formula =IF(AND(H2=0, NOT(P2="")),1,0)

| C            | D           | E           | F   | G   | H     | I        | J          | K           | L        | M        | N          | O             | P            | Q         | R     | S                              | T | U | V |
|--------------|-------------|-------------|-----|-----|-------|----------|------------|-------------|----------|----------|------------|---------------|--------------|-----------|-------|--------------------------------|---|---|---|
| school_id    | district_na | district_id | age | sex | stool | hookworm | ascaris_eg | trichuris_e | smansoni | hookworm | ascaris_in | trichuris_inf | smansoni_inf | deworming | check |                                |   |   |   |
| 1190 BARINGO | 701         | 9           | 1   | 1   | 0     | 0        | 0          | 0           | 0        | 0        | 0          | 0             | 0            | 0         | 0     | =IF(AND(H2=0, NOT(P2="")),1,0) |   |   |   |

- Then copy the formula to all other cells (ensure that the formula is copied to all rows in your dataset)
- Now use the filter to show only entries with a check value of 1
- In reality you might want to go back to the original records to investigate the error. For the purpose of the practical, just replace the provided stool samples to yes (1) and take note of the changes in your log file.

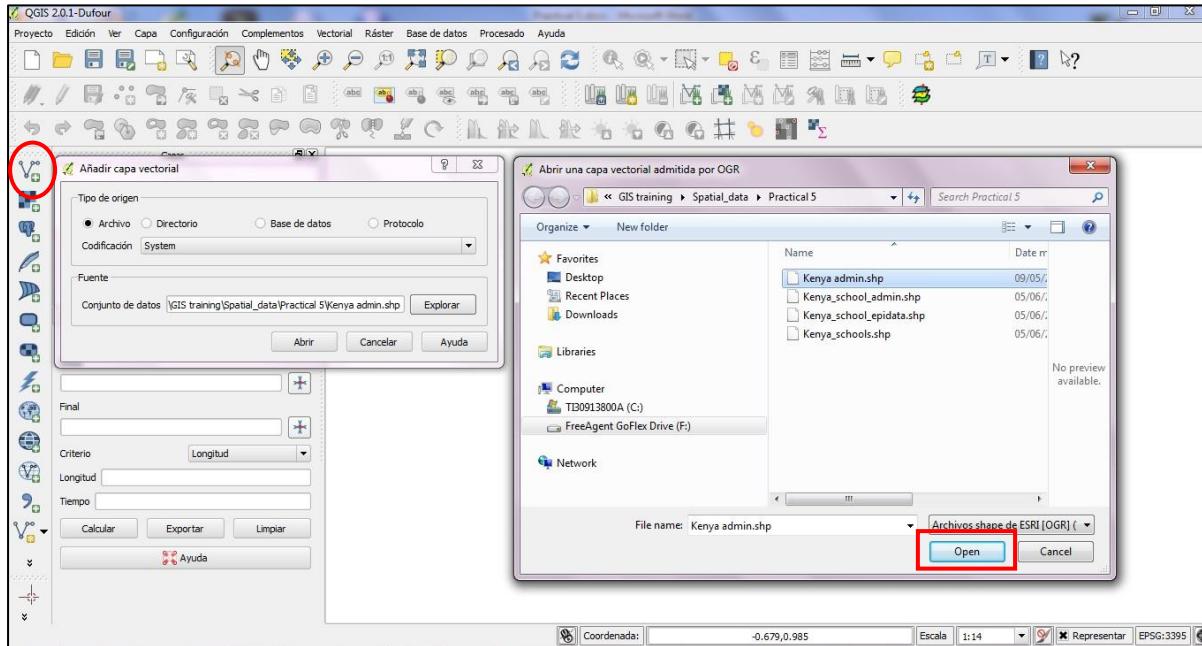
## 5. Verifying coordinates of mapping data

We also need to verify if the coordinates of study locations are correct. Therefore, we will need to display the school locations on a map of Kenya and compare the mapped location with the supposed location. We can use the included district information (variable district\_name) to verify the location.

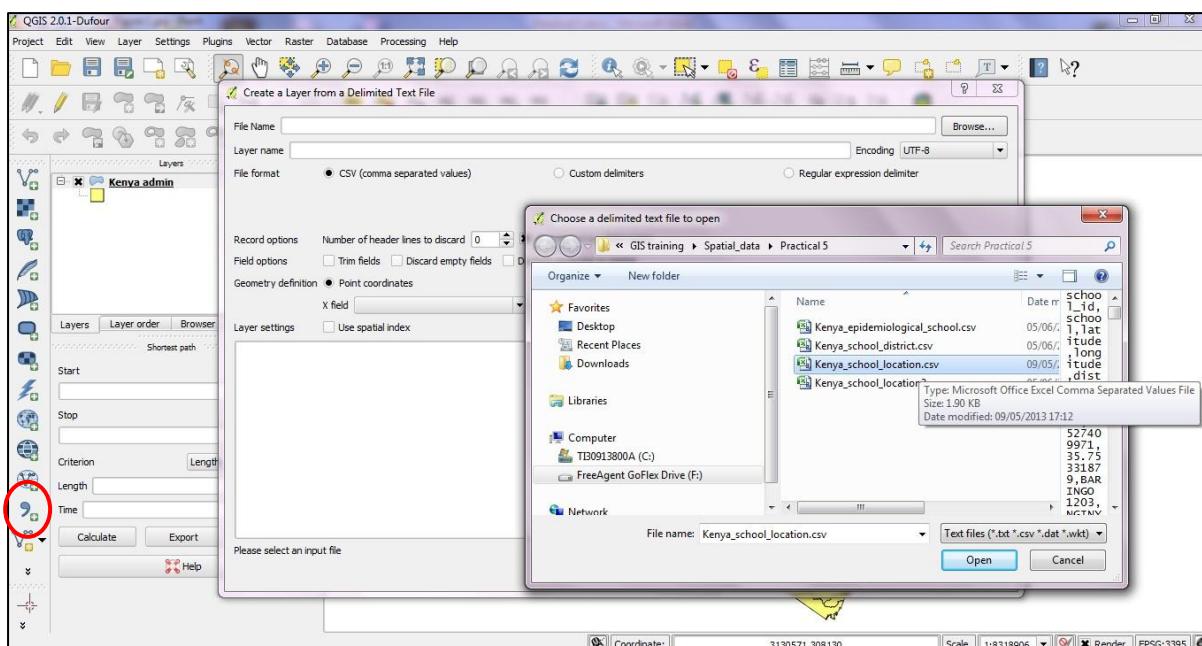
## GIS training for Neglected Tropical Diseases mapping Data management for mapping using QGIS 2.0.1-Dufour

We will use the file *Kenya\_school\_location.csv*. This is a list of the 41 schools including the school\_id, longitude, latitude, and district\_name. The file *Kenya\_school\_dict.xls* contains information about the variables.

- Open a QGIS project and add the Kenyan administrative boundaries *Kenya\_admin.shp* to your project (using the **Add Vector Layer** function). You do not need to worry about the visualisation now.

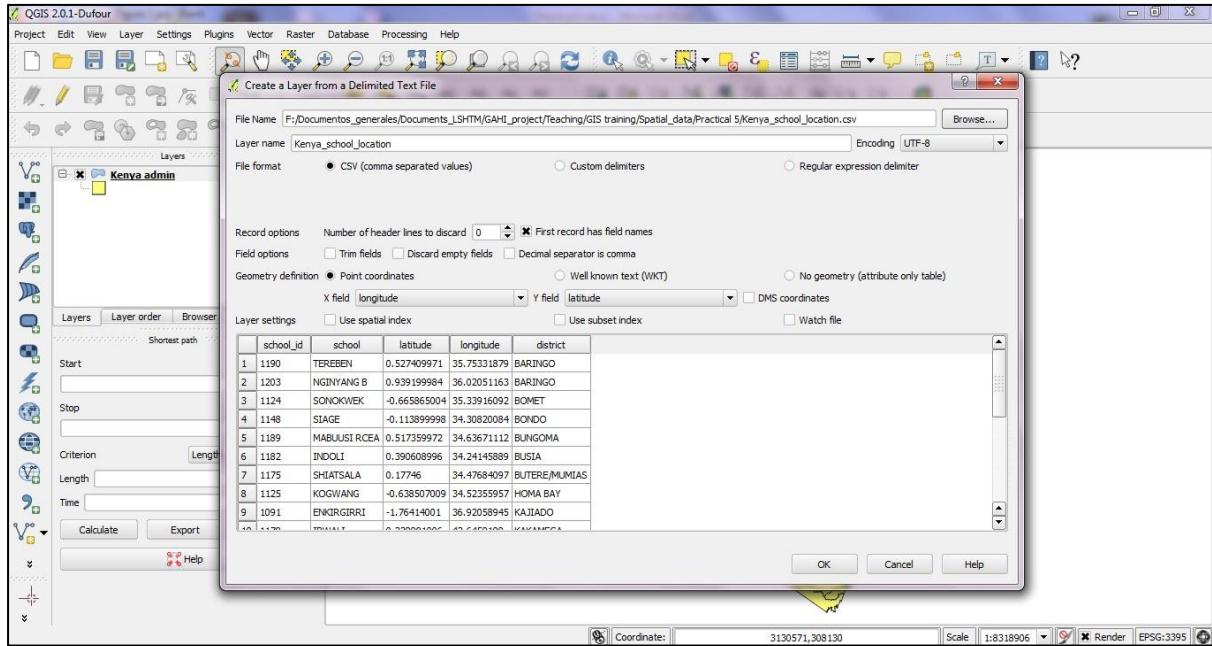


- Add *Kenya\_school\_location.csv* by displaying the schools using their coordinates on the map. Therefore, click on **Add delimited text layer** in the menu, and browse to select *Kenya\_school\_location.csv*.

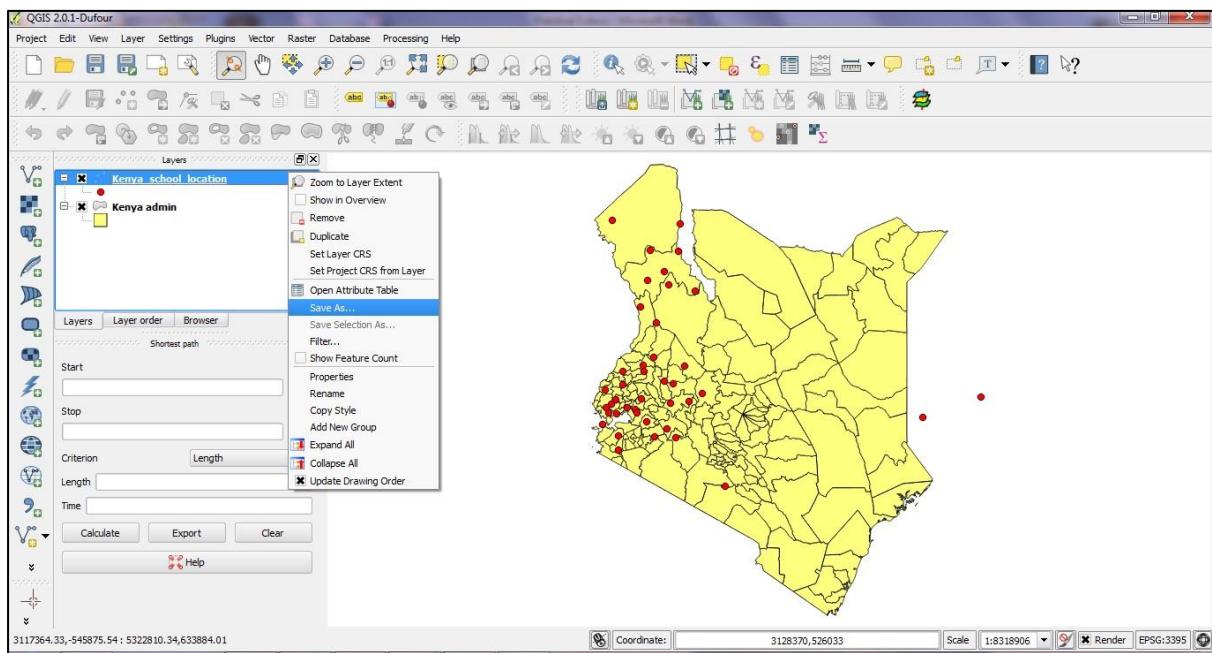


## GIS training for Neglected Tropical Diseases mapping Data management for mapping using QGIS 2.0.1-Dufour

- Select Comma as delimiter and longitude in the x field and latitude in the y field. Press OK and select WGS84 as coordinate system. The exact steps are also described in Practical 2, Section 3.1 (3.1. Geo-positioning a set of locations from a data table).



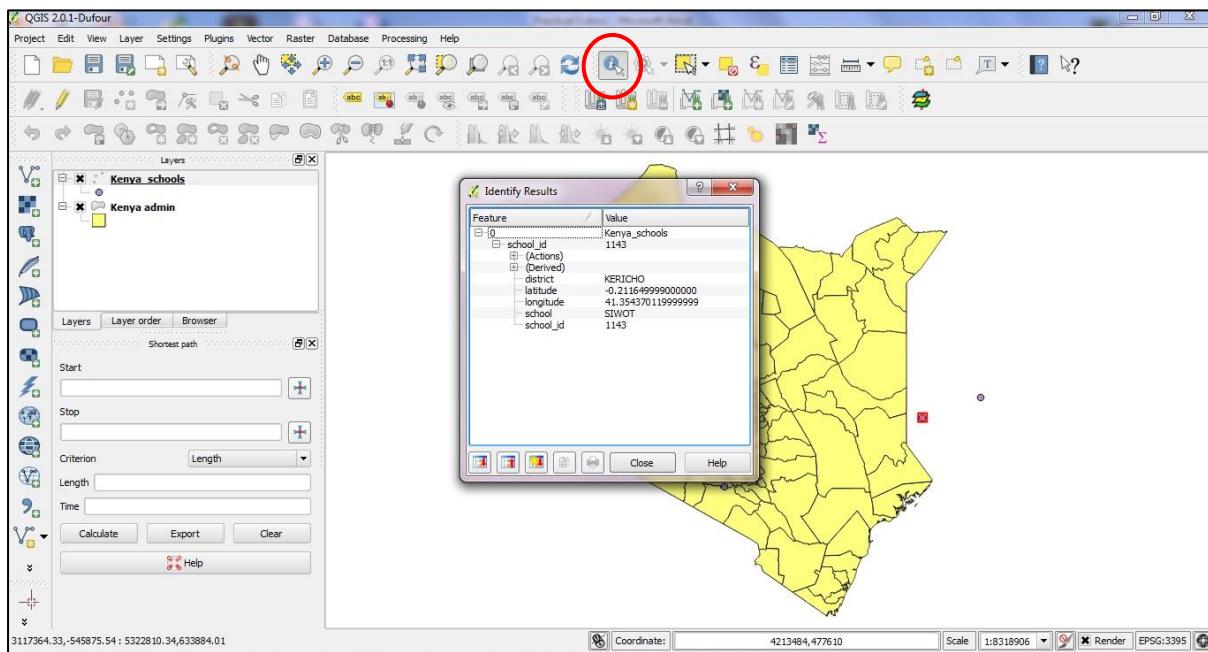
- Save the layer as a shapefile (*Kenya\_schools.shp*) and add the shapefile to your map (you can remove the *Kenya\_school\_location* events layer).



Are there any schools that map outside of Kenya?

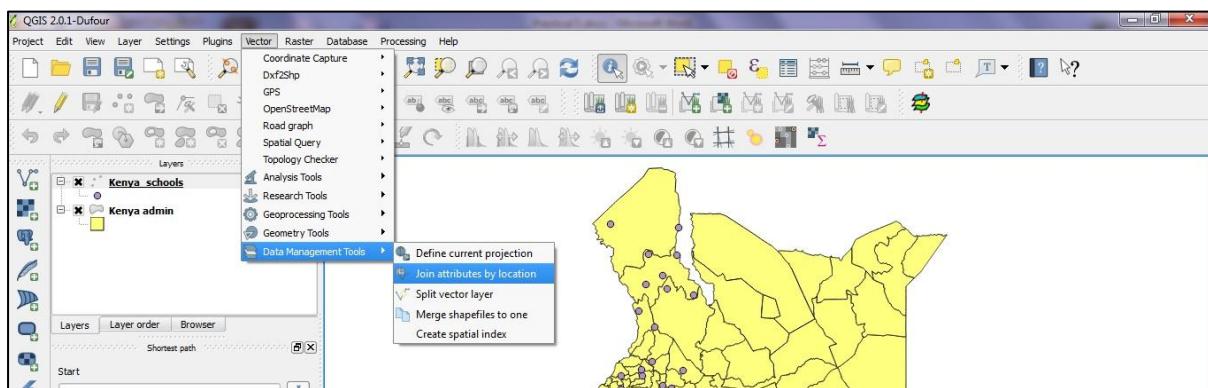
GIS training for Neglected Tropical Diseases mapping  
Data management for mapping using QGIS 2.0.1-Dufour

- After selecting the *Kenya\_schools* layer in the TOC, activate the **Identify Features** tool and click on the points outside of Kenya to obtain names of schools that do not have the correct coordinates. (The tool was introduced in Practical 1, Section 4.1). Note the names of the schools; you will need them to change the coordinates in the original file.



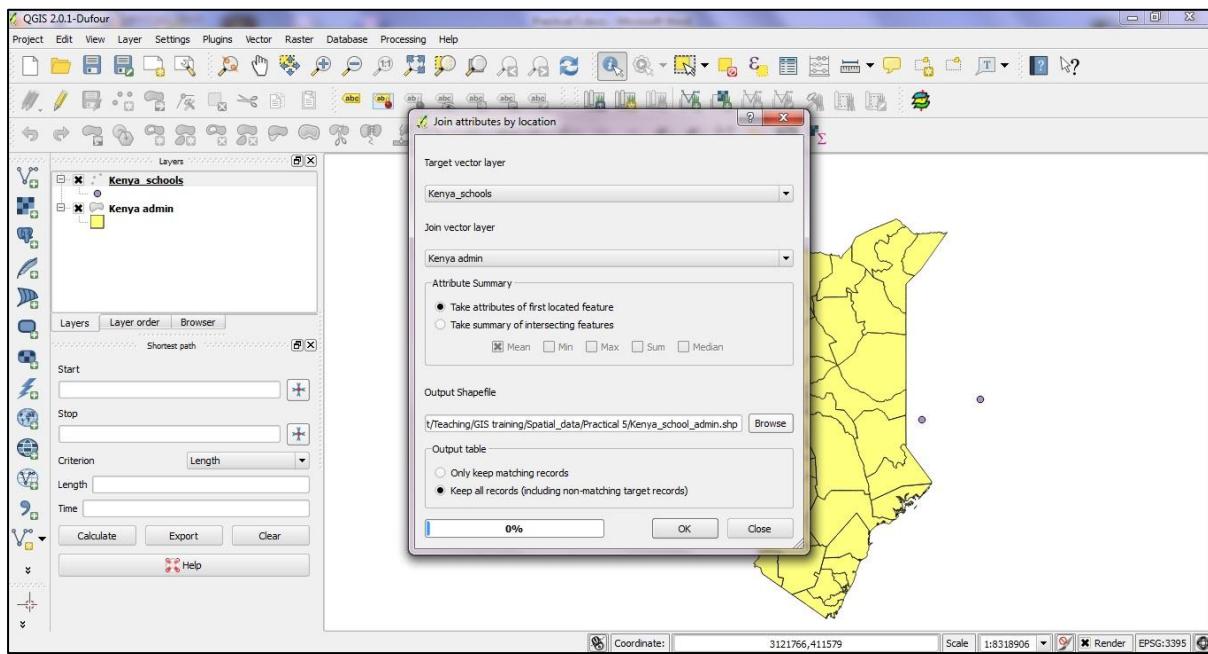
Now we want to add the district information to the school data. Therefore we will join information based on the geographical localisation.

- In the menu click on **Vector**, then **Data Management Tools**, then select **Join attributes by location**.

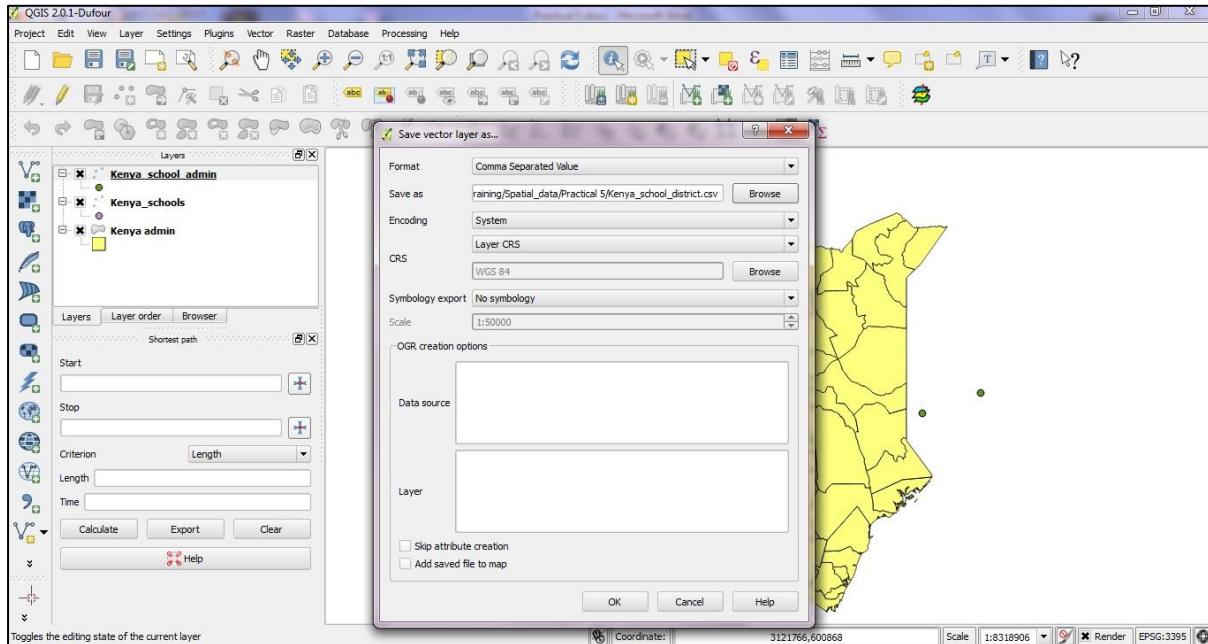


- Then select *Kenya\_schools* as target layer and *Kenya\_admin* as join layer and save the new shapefile as *Kenya\_school\_admin*. Select **Keep all records**.

## GIS training for Neglected Tropical Diseases mapping Data management for mapping using QGIS 2.0.1-Dufour



- Display the new layer on the map
- Save the joined data as .csv file. You find the **Save as** function in the dropdown menu when you right-click on the new layer in the TOC. Choose Comma separated value and save as *Kenya\_school\_district*.



- You can now open the data in Excel. Scroll down comparing ‘district’ with ‘NAME’ added by the spatial join. You can hide the other columns to make the comparison easier.

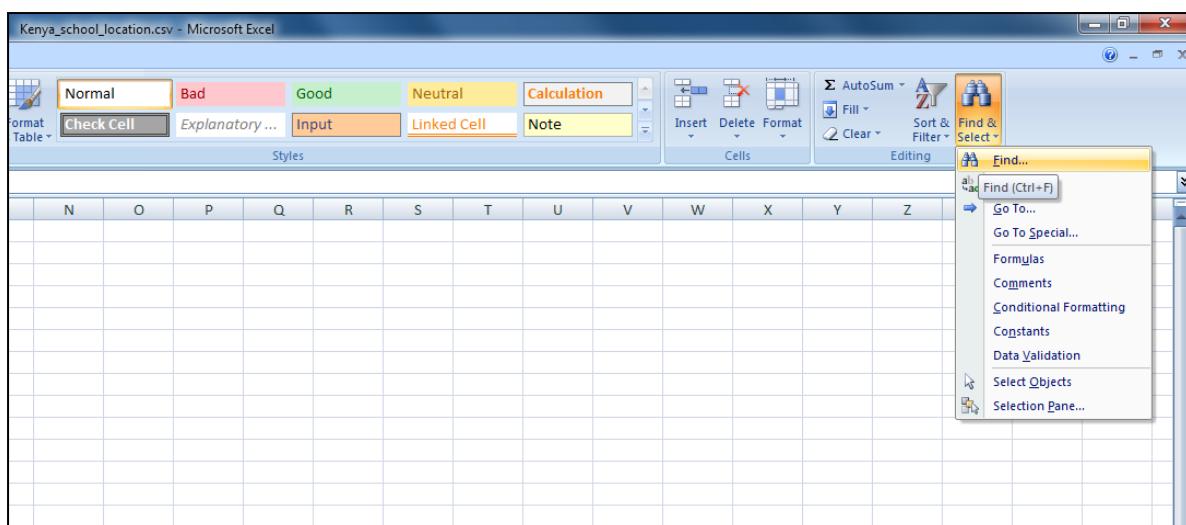


Do you observe any inconsistencies in district names?



*Note: The district boundaries of the shapefile serve only as a guideline to detect mapping errors and you will need to consider several points during the verification. The provided administrative region names might be administrative districts or health districts and these might not be the same. Moreover, the names of districts and their boundaries might have changed over time (as it is the case for Kenya!). Therefore, if the information does not match to your shapefile, it does not necessarily mean that your coordinates are wrong, and you should investigate these inconsistencies further.*

- For the purpose of the practical, ignore the district name inconsistencies (we assume now that the discrepancies originate from changes in administrative boundaries and district names). However, replace the coordinates of schools that map outside of Kenya in the *Kenya\_school\_location.csv* file as follows:  
IBWALI: Longitude 34.6459198  
SIWOT: Longitude 35.35437012
- You can use the Find function to find the relevant schools.



- Save the file as *Kenya\_school\_location2.csv*.

## 6. Preparing data for mapping

---

The Kenya dataset contains information individually for each study participant. You could display each individual on the map, however all children in one school would map exactly to the same location and a meaningful presentation of the

results would therefore not be possible. Therefore, to visualise the survey results on a map, you will need to summarise your data for each study location. In this case we want to display summaries for each school, but it might also be communities or districts etc. There are different types of statistics to summarise data such as counts, sums, mean values, min, max etc.



Think about which types of statistics (count, sum, mean...) you could use present a meaningful summary estimate of each variable per school?

### 6.1. Collapse numerical values by study location

- Go back to the individual level data (*Kenya\_epidemiological\_data\_2.xls*)
- We will now collapse the data by school using the **PivotTable** function in Excel (which was already used in section 4). You can find the function in the **Insert** tab.
- Ensure that the entire dataset is selected (this should be done by default)
- Tick **New Worksheet** to tell Excel that you want to place the table in a new sheet.
- An empty table is inserted in a new sheet and a window will open on the right hand side named **PIVOT TABLE FIELD LIST**.

A screenshot of Microsoft Excel showing the PivotTable Field List dialog box. The dialog box is circled in red and contains a list of fields: child\_id, school\_name, school\_id, district\_name, district\_id, age, sex, stool, hookworm\_egg, ascariasis\_egg, trichuris\_egg, smanson\_egg, and haemonchus\_egg. The main Excel worksheet shows a PivotTable structure with columns for Row Labels and Values. The PivotTable Tools ribbon tab is selected, showing options like Active Field, Group Selection, Ungroup, Group Field, Sort, Refresh, Change Data Source, Clear, Select, Move, PivotTable Actions, PivotChart, Formulas, OLAP tools, Tools, Field List, Buttons, Headers, and Show/Hide.

GIS training for Neglected Tropical Diseases mapping  
*Data management for mapping using QGIS 2.0.1-Dufour*

- Now click on school\_id and draw it in the “Row labels” field at the bottom.
- Click on the variables that you wish to summarize by school and draw them one by one into the “Σ Values” field (this time we do not need any column labels). Let’s start with district\_id.
- By clicking on the variable name in the “Σ Values” field a drop down list will open, click on **Value Field Settings**.... Now you can choose the type of statistic that you would like to use to summarize the variable. As district\_id should be identical for all individuals in the same schools choose **Average** as type of calculation.

| school_id | Total |
|-----------|-------|
| 1091      | 704   |
| 1119      | 613   |
| 1123      | 712   |
| 1124      | 702   |
| 1125      | 610   |
| 1133      | 710   |
| 1139      | 616   |
| 1141      | 706   |
| 1143      | 706   |
| 1145      | 611   |
| 1146      | 614   |
| 1148      | 609   |
| 1150      | 611   |
| 1155      | 808   |
| 1156      | 615   |
| 1163      | 615   |
| 1164      | 707   |
| 1169      | 707   |
| 1175      | 803   |
| 1178      | 711   |
| 1179      | 804   |

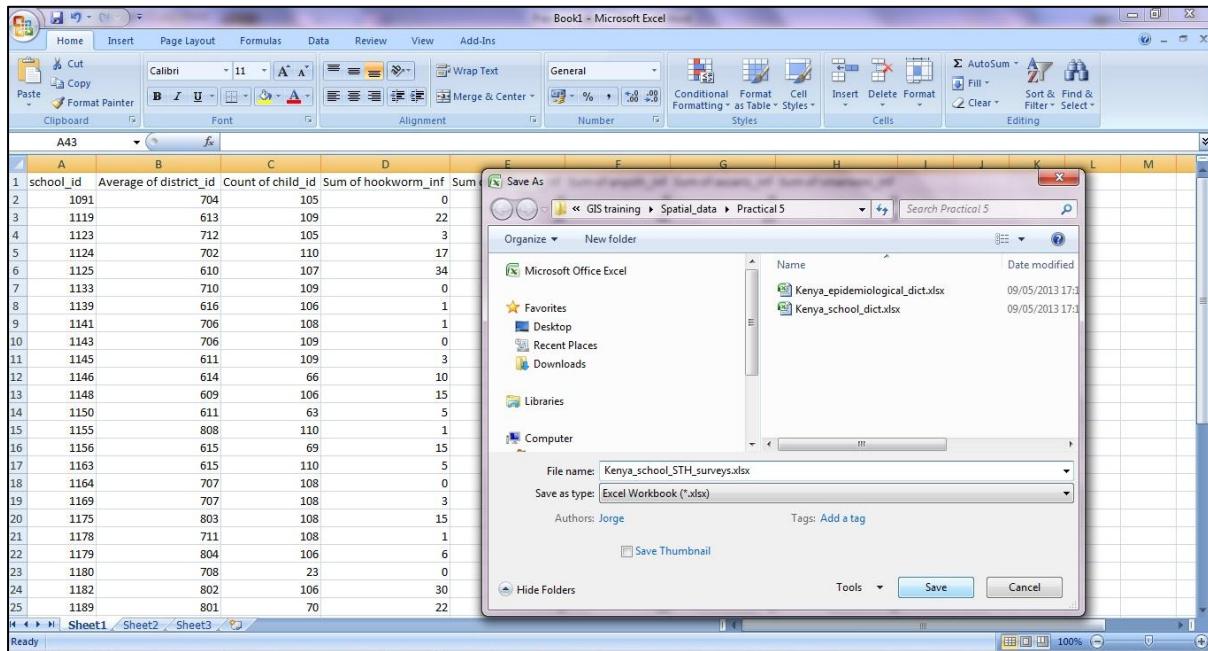
- Now draw child\_id into the “Σ Values” field and choose **Count** to summarize the results. This will give you the number of children per school.
- We would also like to know the number of children infected with any STH per school. Draw anysth\_inf into the field and choose **Sum**. As infected children have a value of one and not-infected children have a value of 0 the sum will represent the number of infected children.
- You can also summarize other variables as you wish.

## GIS training for Neglected Tropical Diseases mapping Data management for mapping using QGIS 2.0.1-Dufour

- Now copy the table, open a new Excel file and paste the values into the new spreadsheet. Therefore, click **Paste** in the **Home** tab, then choose **Paste special** and paste only the values. You can now modify the headings of the table to have a head row with variable names as you wish. You can also remove the last row which represents total values.

- Save the new data table as *Kenya\_school\_STH\_surveys.xlsx*.

## GIS training for Neglected Tropical Diseases mapping Data management for mapping using QGIS 2.0.1-Dufour



### 6.2. Creating new variables using formulas

You might also want to introduce new variables to the school data. For example, you could present STH prevalence per school instead of the total count of infected children.

- Therefore, create a new variable called sth\_prev by dividing the number of infected children by the number of children and multiplying by hundred to obtain a value in percent (%). In the first cell under the heading type =F2/C2\*100 (this assumes that your number of children is in row C and the number of infected children in row F; if you had summarized additional variables you will need to adjust the column label accordingly); then copy the formula to the other cells in the column.

| A         | B                      | C                 | D                   | E                    | F                 | G                  | H                  | I        | J | K | L | M |
|-----------|------------------------|-------------------|---------------------|----------------------|-------------------|--------------------|--------------------|----------|---|---|---|---|
| school_id | Average of district_id | Count of child_id | Sum of hookworm_inf | Sum of trichuris_inf | Sum of anysth_inf | Sum of ascaris_inf | Sum of smanson_inf | sth_prev |   |   |   |   |
| 1091      | 704                    | 105               | 0                   | 0                    | 0                 | 0                  | 0                  | 0/C2*100 |   |   |   |   |
| 1119      | 613                    | 109               | 22                  | 18                   | 45                | 21                 | 8                  |          |   |   |   |   |
| 1123      | 712                    | 105               | 3                   | 0                    | 3                 | 2                  | 0                  |          |   |   |   |   |
| 1124      | 702                    | 110               | 17                  | 9                    | 34                | 27                 | 0                  |          |   |   |   |   |
| 1125      | 610                    | 107               | 34                  | 8                    | 39                | 11                 | 1                  |          |   |   |   |   |
| 1133      | 710                    | 109               | 0                   | 1                    | 1                 | 0                  | 0                  |          |   |   |   |   |

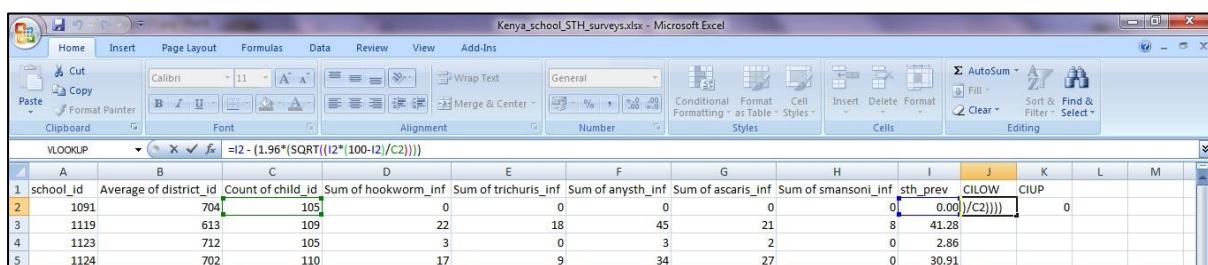
Additionally, you could calculate the 95% confidence interval (CI) of your prevalence. The statistical formula for calculating the 95% CI is as follows

## GIS training for Neglected Tropical Diseases mapping Data management for mapping using QGIS 2.0.1-Dufour

$C_{low} = sth\_prev - 1.96 * (\text{SQRT}((sth\_prev * (100 - sth\_prev)) / nr\_children)))$

$C_{up} = sth\_prev + 1.96 * (\text{SQRT}((sth\_prev * (100 - sth\_prev)) / nr\_children)))$

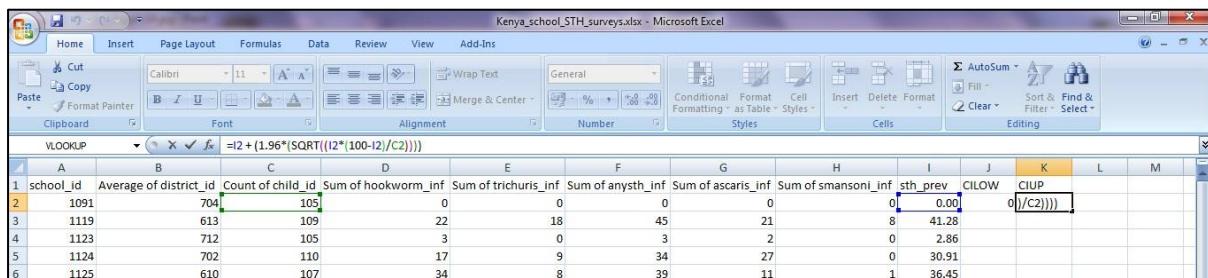
- You will have to create the following variables first: “ $C_{low}$ ” (lower confidence limit) and “ $C_{up}$ ”(upper confidence limit)
  - You can calculate the limits by typing the formulas as follows and then copying to the other cells:
- $C_{low}:$   
 $=I2 - (1.96 * (\text{SQRT}((I2 * (100 - I2)) / C2))))$



| A         | B                      | C                 | D                   | E                    | F                 | G                  | H                   | I        | J     | K    | L | M |
|-----------|------------------------|-------------------|---------------------|----------------------|-------------------|--------------------|---------------------|----------|-------|------|---|---|
| school_id | Average of district_id | Count of child_id | Sum of hookworm_inf | Sum of trichuris_inf | Sum of anysth_inf | Sum of ascaris_inf | Sum of smansoni_inf | sth_prev | CLOW  | CIUP |   |   |
| 2         | 1091                   | 704               | 105                 | 0                    | 0                 | 0                  | 0                   | 0        | 0.00  |      |   |   |
| 3         | 1119                   | 613               | 109                 | 22                   | 18                | 45                 | 21                  | 8        | 41.28 |      |   |   |
| 4         | 1123                   | 712               | 105                 | 3                    | 0                 | 3                  | 2                   | 0        | 2.86  |      |   |   |
| 5         | 1124                   | 702               | 110                 | 17                   | 9                 | 34                 | 27                  | 0        | 30.91 |      |   |   |

$C_{up}:$

$=I2 + (1.96 * (\text{SQRT}((I2 * (100 - I2)) / C2))))$



| A         | B                      | C                 | D                   | E                    | F                 | G                  | H                   | I        | J     | K     | L | M |
|-----------|------------------------|-------------------|---------------------|----------------------|-------------------|--------------------|---------------------|----------|-------|-------|---|---|
| school_id | Average of district_id | Count of child_id | Sum of hookworm_inf | Sum of trichuris_inf | Sum of anysth_inf | Sum of ascaris_inf | Sum of smansoni_inf | sth_prev | CLOW  | CIUP  |   |   |
| 2         | 1091                   | 704               | 105                 | 0                    | 0                 | 0                  | 0                   | 0        | 0.00  | 0.00  |   |   |
| 3         | 1119                   | 613               | 109                 | 22                   | 18                | 45                 | 21                  | 8        | 41.28 | 41.28 |   |   |
| 4         | 1123                   | 712               | 105                 | 3                    | 0                 | 3                  | 2                   | 0        | 2.86  | 2.86  |   |   |
| 5         | 1124                   | 702               | 110                 | 17                   | 9                 | 34                 | 27                  | 0        | 30.91 | 30.91 |   |   |
| 6         | 1125                   | 610               | 107                 | 34                   | 8                 | 39                 | 11                  | 1        | 36.45 | 36.45 |   |   |

 Note: When making conclusions from a sample about the prevalence in the entire population we will have to take into account uncertainty that comes from randomly drawing a subset of the people. The 95% confidence interval gives you the range of prevalence within which your real population prevalence will lie with 95% confidence. This means that if you would draw 100 random samples, in 95 of these samples the prevalence would be within the range.

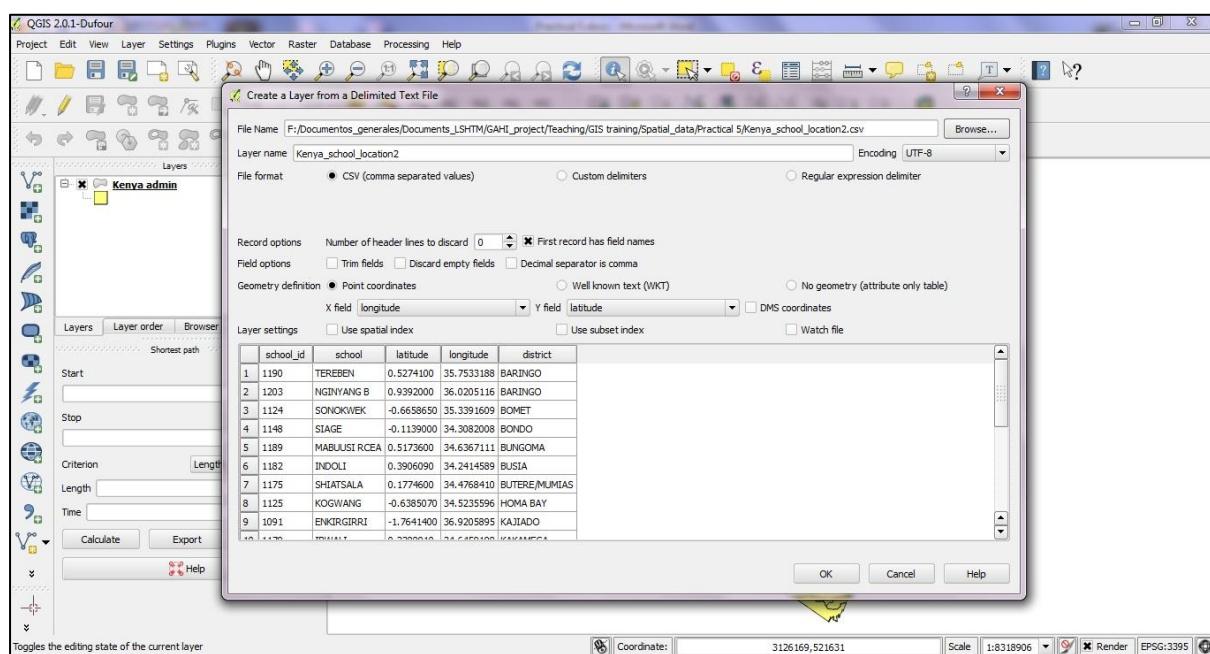
- When you are happy with all numerical variables save the file as *Kenya\_epidemiological\_school.csv* file.

### **6.3. Merged variables to collapse data**

Some information will be collected only at school level e.g. school coordinates, school characteristics. Moreover, it is not possible to collapse string variables in the same way as numerical variables e.g. the school or district names. Therefore, these need to be processed separately and then merged with the rest of the dataset.

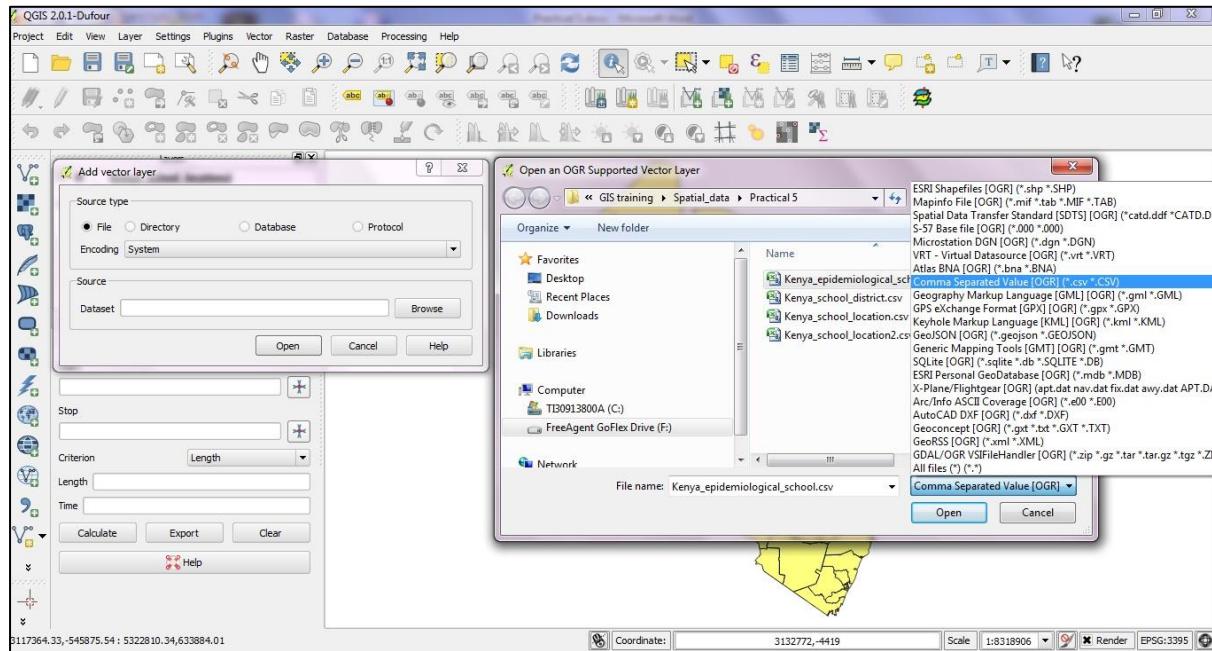
We will now merge the epidemiological data with the school information.

- Open a new QGIS project.
- First, add *Kenya\_school\_location2.csv* to the project displaying their longitude and latitude (this should be the file with the correct coordinates).. Click on **Add delimited text layer** in the menu, and browse to select *Kenya\_school\_location.csv*, and follow the same steps described in the section 5 to geolocate the schools.

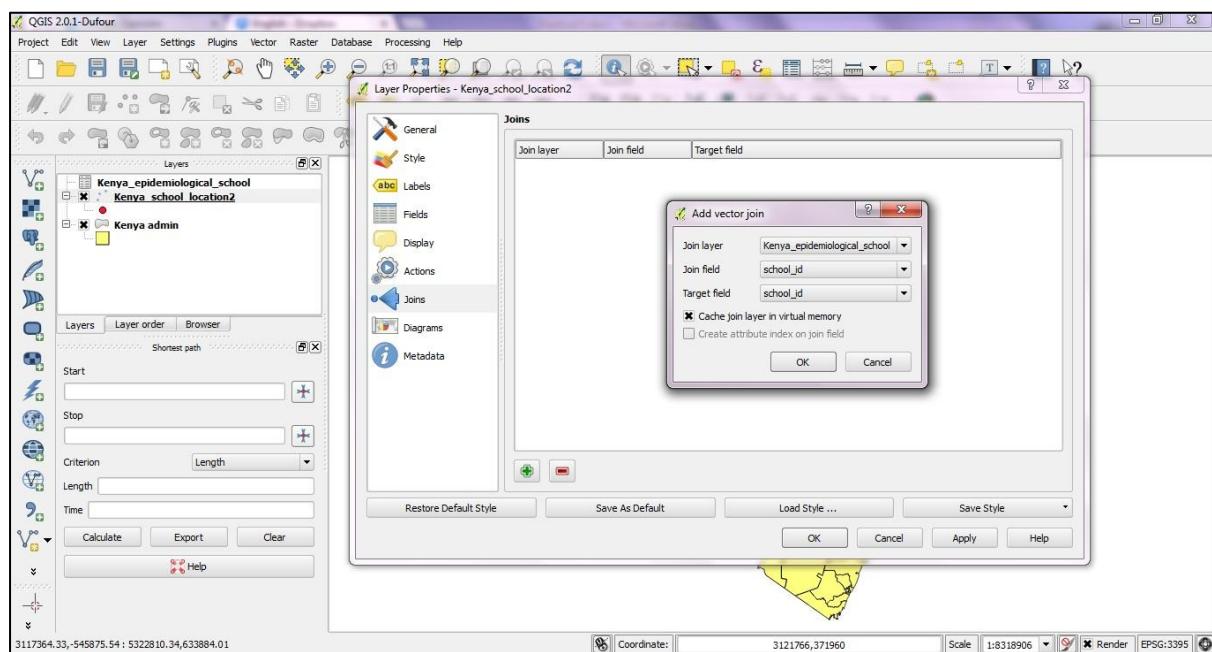


- Click on **Add vector data** and add *Kenya\_epidemiological\_school.csv*.

## GIS training for Neglected Tropical Diseases mapping Data management for mapping using QGIS 2.0.1-Dufour

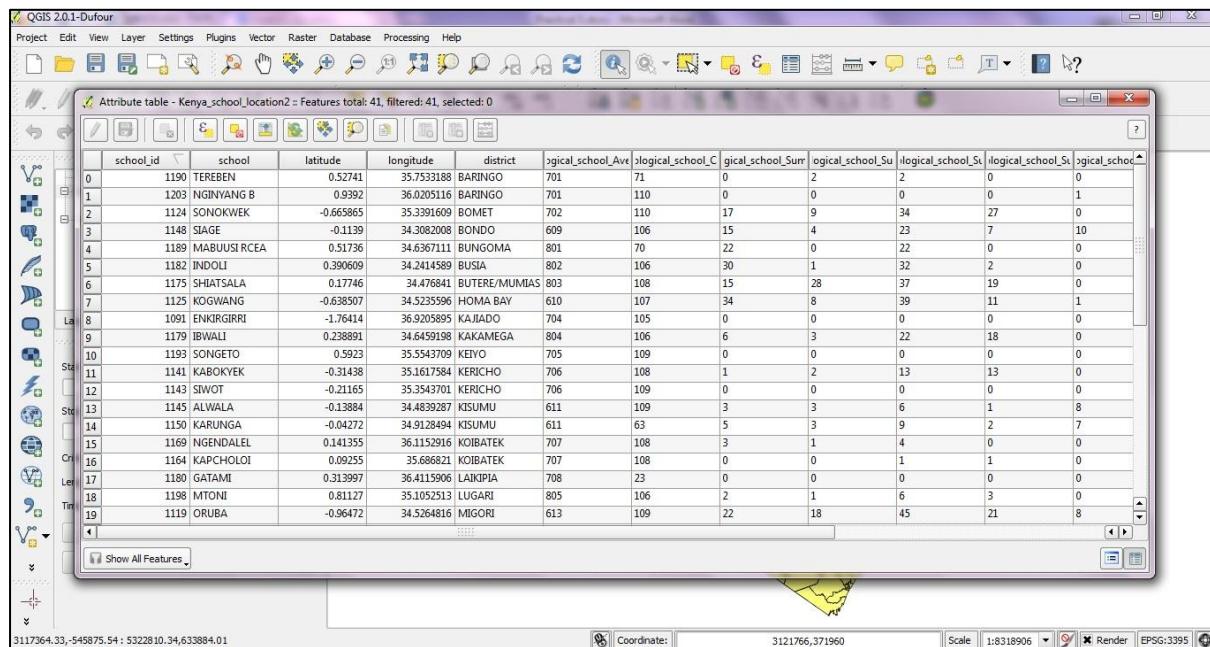


- Now open the properties of the Kenya \_school\_location2 layer and select join from the menu.
- Add a vector join (by clicking on the plus).
- Specify Kenya\_epidemiological\_school as ‘join layer’ and school\_id as ‘Join field’ and ‘Target field,’ then click OK.



## GIS training for Neglected Tropical Diseases mapping Data management for mapping using QGIS 2.0.1-Dufour

- If you open now the attribute table of the Kenya\_school\_location2 layer you will see that the epidemiological variables are included.
- Save *Kenya\_school\_epidata.shp* as a new shapefile. You can use it now to visualise your data.
- Open the attribute table of the new shapefile to check that the join has been done correctly and our layer contains the fields needed.



- You can additionally save the new dataset as .csv file, which you can open in Excel.



**Note:** The same principles apply when you want to summarize your data by district. First summarize your data by district using the PivotTable. Using the district\_id, you can join the data to a shapefile with administrative regions using the same district\_id. If you don't have a district\_id included in your data, you can add these by a spatial join as was shown when verifying the coordinates.

**Main steps from data collection to presenting your data on maps**

- Data entry (ideally double entry) using an interface that allows for structural checks (necessary fields), plausibility (range of values), and logical checks (inconsistencies)
- Export of data into a format that can be further processed (e.g. xls for Excel)
- Data cleaning by verifying the structure of your database, the range of values, and the consistency of data
- Verify the coordinates of study locations by displaying them on a map
- Transform individual level data to group level data (e.g. schools, districts)
- Create new variables of interest (e.g. prevalence)
- Merge datasets from different sources (e.g. school level data)
- Create your maps and think about how to display the attributes of your data to present the results of interest