GIS DATA PREPARATION AND MANAGEMENT

Practical Task 3: Cleaning tabular data in preparation for use in GIS

Task description

This task will walk you through some common data errors and inconsistencies that are likely to appear in raw data files. You will look at how to identify these types of errors and think about which factors should be considered when making amendments.

Data rules covered in this task

- Not all data will have the same errors! A good understanding of how the data were collected will help you to identify potential problems.
- Do not delete any data unless the process has been approved as part of your organization's data management protocol.
- If you make any changes to the data, create new intermediate data files.
Part 1: Categorical data standardization

In Part 1 you will learn how to standardize and clean categorical data, looking specifically at correcting common errors, and prepare it for GIS use.

You may either continue working on the Excel file from Task 2 (saved in the 3_IntermediateData folder) or open the following file, which has been prepared specifically for this Practical Task 3:

...GRID3DataMng\BackupFiles\Kano_HealthFacilities_TASK3.xlsx

Steps to follow:

1. Open and visually inspect the spreadsheet
2. Check the information in the metadata tab
3. Look at the data tab to scan for obvious issues

Data tables containing a large amount of information are very difficult to inspect visually for mistakes. This task will take you through some simple processes in Excel that will help to identify potential errors in your data.

Potential errors to identify:
- Extra spaces
- Special/unusual characters
- Capitalization inconsistencies

Note: The following processes are not definitive and will not identify all potential errors in your data. These examples are designed to highlight the range of errors you should be looking for; the specific errors you find will depend heavily on the type of data you are using.
1.1 Remove extra spaces

Excel’s TRIM function removes spaces from the beginning and end of text fields and changes any multiple spaces within text fields to single spaces.

Steps to follow:
1. Sort data by uniq_id (smallest to largest)
2. Navigate to column Q (prmry_name)
3. Look at the top 15 health facility names—what do you notice?
4. Use the TRIM function—e.g. =TRIM(Q2)—in column R to remove leading and trailing spaces
5. Apply the function to all rows

6. Replace the data in column Q with the corrected data in column R

Copy > Paste special > Values Only to replace column Q with the corrected data
7. Delete the duplicated column (column R)

You can repeat this process for any other columns that are likely to have spacing errors (any text data).

1.2 Correct special characters

Typos or unusual characters may affect how the data are displayed in GIS and prevent you from running some analysis tools. These errors should always be corrected prior to any geospatial analysis. These steps will show you how to identify and correct any errors related to special characters:

Steps to follow:
1. Navigate to the **wardname** column and scroll down to cell E36
2. Note the obvious “#” typo in the cell–this would not cause any problems in a string field, but as it is indeed a typo, it needs to be resolved
3. Look at cell H36, where you will notice another error
4. The letters in this cell would cause an error in the integer field and must be replaced with zeros (see Appendix 1 for more information on date types)

Now let’s check both columns for repeats of the same errors:

Steps to follow:
1. Select the whole of column E (wardname) and press Ctrl + H to open the Find and Replace dialog box
2. Enter the special character (#) to find and replace it with the desired value (you can leave the Replace box empty to delete the special character rather than replace it with something)
3. Repeat this process for column H (lgacode) to check if there are any more cells that contain the letter “O” instead of the number “0”

*Note: When using the Find and Replace tool for this example, find and replace individual selections. Using the replace all function, while useful in some cases, may result in changing cells across the entire table and is not recommended.*
1.3 Correct place name capitalization inconsistencies

Name inconsistencies can cause issues when carrying out statistical or spatial analysis. A common error is capitalization errors. In the statename column, you can see that some entries have a capital "K" and others have a lowercase "k." Let’s standardize these to all be capitalized using the PROPER function.

Steps to follow:
1. Create a new column next to Column I (statename)
2. Use the PROPER function—e.g. =PROPER(I2)—to ensure that the first letter in each cell is capitalized
3. Apply this function to all rows
4. Copy and paste as values-only back into column I to replace the original data
5. Delete the new duplicated column

Part 2: Duplicated or missing records

In this section you’ll learn how to identify and best manage missing or duplicate records.

2.1 Duplicated records

Duplicated records can sometimes be a mistake or data entry error, but they should never be deleted without careful consideration or further investigation.

Steps to follow:
1. Sort the data in alphabetical order for primry name
2. Navigate to rows 168 and 169 and note that there are two (2) entries with the same health facility name (Jemomi Health Post)
3. Review the rest of the information for these entries:
   ■ Are the location coordinates the same?
   ■ Are they located in the same ward?
   ■ Does the attribute data match?

Note: They are located in different wards and the location coordinates are different, meaning they are different facilities—not a duplicate

You may want to flag these health facilities for later review to double-check that one name has not been entered incorrectly.

4. Navigate to rows 317 and 318, where there are another two (2) health facilities with the same name (Wasai Health Post) and investigate
5. Flag for later review or delete the duplicate (depending on the defined guidelines for your organization or team). For this training exercise, delete one of the duplicated rows.

2.2 Missing data

What you do with missing data depends heavily on the data themselves and should always be looked at on a case-by-case basis. Sort the data again by column A (uniq_id). Between rows 71 and 101 there are five (5) rows with some missing data. Assess whether you should leave each cell with missing data blank, or if you have enough information to populate a given cell with data.

Review the following examples and recommendations:

- Missing state code or name (columns I and J)
  - Because all the data are from Kano State, you can fill in any missing state codes (KN) and names (Kano)
- Missing ward name (column E)
  - Use the ward codes (column F) to identify the missing ward name and fill in the cell
- Missing health facility category (column M)
  - Information such as health facility category, type, or ownership may be used to categorize the data in spatial or statistical analysis. Rather than leaving the cell blank, mark the cell with the word "unknown" (use the Find and Replace tool as you did earlier in the task). This will now show up as a separate category in GIS, but can still be represented on the maps.

Missing coordinate data (latitude and longitude) will result in data entries not displaying in GIS. Without coordinate data, the software doesn’t know where to place the data point. It is important to flag any entries that have missing coordinate data.
Ways to identify missing data

Conditional Formatting
Select the range of data you want to analyze and then click the Conditional Formatting button on the Home tab. Then choose the type of formatting you want to apply. Any cells that do not meet the criteria will be identified. This can be a great way to quickly identify any missing values in your data.

VLOOKUP Function
Select the range of data you want to analyze and then enter the VLOOKUP formula. Specify the column you want to search and then the column you want to return a value for. Once the formula is entered, any cells that do not meet the criteria will be identified.

Pivot Tables
Select the data range you want to analyze and then click the Pivot Table button on the Insert tab. From there, you can select the columns you want to analyze and choose the type of analysis you want to view; then you can quickly scan the data to identify any missing values.

Make sure you add a comment under the metadata tab giving a brief description of the data cleaning processes you have used; be sure to include the date and editor’s name.

<table>
<thead>
<tr>
<th>Date created</th>
<th>31-Mar-24</th>
</tr>
</thead>
<tbody>
<tr>
<td>Created by</td>
<td>A.Learner, Junior data scientist, GRID3</td>
</tr>
<tr>
<td>Details</td>
<td>Health facility data for 8 LGAs in Kano state, Nigeria (Dawakin Tofa, Dawakin Kuda, Gabasawa, Gezawa, Kumbotso, Nassara, Tarauni, Ungogo). Data collected between 6 January and 10 March 2024.</td>
</tr>
<tr>
<td>Date updated</td>
<td>31-Mar-24</td>
</tr>
<tr>
<td>Edits made</td>
<td>General data cleaning and pre-processing. Duplicated row deleted (Wasai Health Post) by A.Learner on 31 March 2024</td>
</tr>
<tr>
<td>Version</td>
<td>1</td>
</tr>
<tr>
<td>Data source</td>
<td>Collected by LGA survey teams during 2024 measles campaign</td>
</tr>
<tr>
<td>CRS</td>
<td>Unknown</td>
</tr>
<tr>
<td>Terms of use</td>
<td>For internal use only</td>
</tr>
<tr>
<td>Known errors</td>
<td>Data expected from 10 LGAs but only received for 8</td>
</tr>
<tr>
<td>Additional notes</td>
<td>GPS and ODK forms were used during data collection. Metadata tab created by A.Learner on 31 March 2024</td>
</tr>
</tbody>
</table>

Part 3: Save data as CSV file

Now that you have prepared and cleaned your data, they are ready to be imported into QGIS, visualized, and checked for any further data issues. First, ensure you have saved an up-to-date version of your data as an Excel file (XLSX) in the 3_IntermediateData folder. There is already an intermediate file in this folder, so use a number system to differentiate between the two files.

...GRID3DataMng\3_IntermediateData\Kano_HealthFacilities_31Mar2024_int2.xlsx
It is important to keep a copy of the data in .xlsx format, as this ensures that the metadata are stored within the same file as the data. However, Excel files are not compatible with QGIS. To successfully import the data into GIS, you must also save a copy of the data as a CSV file.

![Image of a file format selection dialog]

Comma-separated values (CSV) is a simple text file format that uses commas to separate values, and uses new lines to separate records. A CSV file stores tabular data (numbers and text) in plain text, where each line of the file represents one data record.

Only one sheet can be stored in a CSV file, meaning that the metadata will no longer be stored with the data themselves. This is why it is important to save a copy of the data in both formats. Once you have saved the CombinedData sheet as a CSV UTF-8 file, you will see the tab name change to represent the sheet’s new name.

![Image of a file with metadata]

Summary

**Learning checklist**

You have now completed the exercises in Task 3, demonstrating good practice in data preparation and management for GIS. You can confidently say you have:

- Saved a raw data file in a readily accessible, fixed directory with a consistent naming convention
- Navigated a pre-defined data management folder structure
- Created detailed metadata by using previous knowledge of the data

In the next task, you will address anomalies with the table structure that would generate errors when you import the data into GIS, and then will create an intermediate-level data file.
Appendix 1 – Data formats and fields in QGIS

Tabular data in QGIS can take any of the following formats:

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Name</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text</td>
<td>String</td>
<td>Health facility</td>
</tr>
<tr>
<td>Whole number</td>
<td>Integer</td>
<td>321</td>
</tr>
<tr>
<td>Decimal number</td>
<td>Real</td>
<td>3.456</td>
</tr>
<tr>
<td>Date</td>
<td>Date (YYYY-MM-DD)</td>
<td>2016-07-28</td>
</tr>
<tr>
<td>Time</td>
<td>Time (HH:MM:SS+nn)</td>
<td>18:33:12+00</td>
</tr>
<tr>
<td>Date &amp; Time</td>
<td>DateTime (YYYY-MM-DD HH:MM:SS+nn)</td>
<td>2016-07-28 18:33:12+00</td>
</tr>
</tbody>
</table>