Outliers and missing data in datasets

In the previous blog post on data verification, see: [link] we mentioned the need to identify outliers in the data. This blog post will look at some of the humanitarian aid datasets/formats and the types of techniques we apply to identify outliers and how to effectively deal with them. The remainder of this blog post will focus on the occurrence of missing data and possible reasons for their occurrence.

Outliers

Outliers are data points whose values are much lower, or much higher than the rest of the data points. We need to identify them, as they may impact the predictive accuracy and model fit when applying simple or multivariate regression analyses. If there are many outliers in the higher values, it is likely that the model will underestimate these values. Similarly, if there are many outliers in the lower values, the model can overestimate them.

Outliers can be classified as either “valid” or “invalid”, depending on the underlying cause(s). For example, in the case of a survey, an observation may have been wrongly entered, or numerical values may have been inserted where descriptive text was expected. Other causes may be the inconsistent use of zero, or “non-applicable”. If a CSV file (“comma separated values”), or a TSV file (“tab separated values”) does not have the proper encoding format (e.g. UTF-8) then rendering their contents on a UTF-8 encoded website may result in specific characters being replaced by black squares or question marks. As a side effect, due to the encoding errors, table entries may be shifted and so end up in the wrong columns. If the table is “scraped” from a website using scripting, these encoding errors will also appear in the downloaded data.

Identifying outliers

In case of large structured datasets such as surveys, manually identifying outliers remains cumbersome.

In order to accelerate the identification we use a number of methods:

- Visual inspection methods such as histogram (Figure 1) or box-and-whisker plots (Figure 2) help us to ‘see’ and interpret the distribution of the data
- Non-visual inspection methods using spreadsheet functions or scripts help us to identify any empty entries, inconsistencies between numerical values and their textual entries, etc.

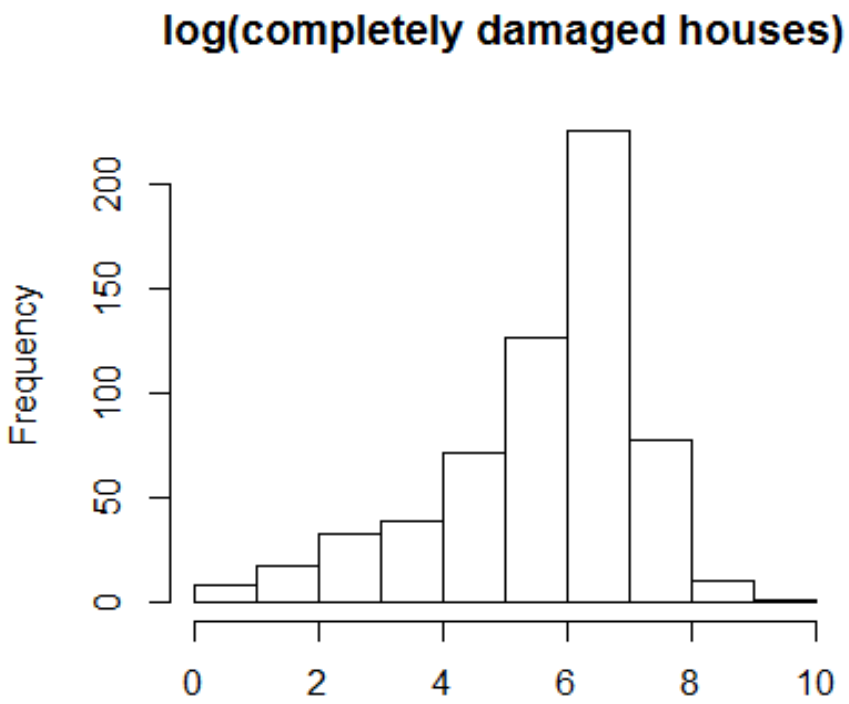

Figure 1: An example of a histogram, showing the number of completely damaged houses due to the Gorka earthquake in Nepal.

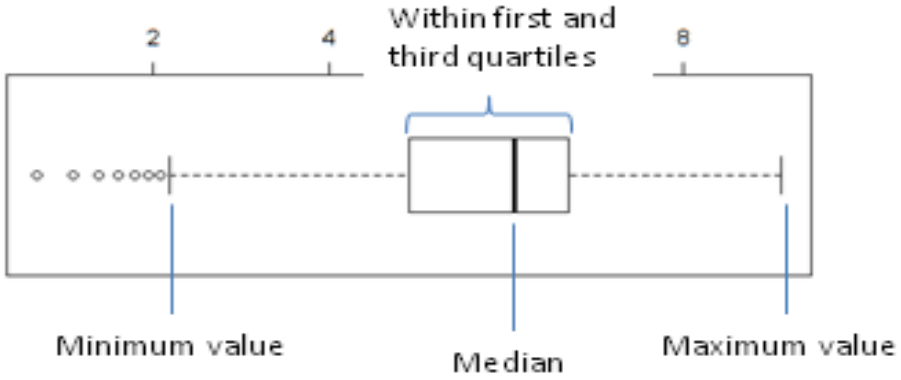

Figure 2: An example of a box-and-whisker plots showing outliers below the “minimum value mark”.

Handling outliers

Once the outliers have been identified, the next step is to determine what to do with them:

- Retaining outliers that appear to be valid data.
- Replacing outliers with a known (or derived) entry from related datasets.
- Deleting outliers from the dataset.

In early stages after a natural disaster, when detailed data is still scarce,

humanitarian aid organizations often publish high-level data which then gets updated over time. Also, as more and more data becomes available, we can use triangulation, where outliers are retained if different sources report similar/equal values and are removed if values differ in two or more of the datasets.

Missing data

In other instances values may be missing from the dataset, either intentionally (subjects refusing to provide answers to survey questions), or unintentionally (data got corrupted or subjects were no longer available to complete a survey). The extent to which the missing data impacts further analyses starts with determining the type of missing data:

- Missing Completely At Random (**MCAR**): data are missing independently of both observed and unobserved data. An example of this would be: entire surveys that, at random, were not submitted, leading to missing values.
- Missing At Random (**MAR**): given the observed data, data are missing independently of unobserved data. An example of this would be: collecting data about a subject's profession where it is known that certain professions are more likely not to share their income. Within subgroups of the profession, missing incomes will be random.
- Missing Not at Random (**MNAR**): missing observations related to values of unobserved data. An example of this would be: people with a low income are less likely to report their income on a data collection form.

We can ignore missing data (= omit missing observations) if we have MAR or MCAR.

In a recently held survey in 2016 on competitiveness index for municipalities in the Philippines, only 1245 municipalities out of a total of more than 1600 municipalities were ranked and no reason was given for the missing 400 municipalities. In such instances it is advised to reach out to the researchers to understand why not all data was published.

After the Gorkan earthquake in Nepal, the number of damaged houses was reported at the lowest administrative level (level 4, Village Development Committee). Each VDC was associated with an identification label. The p-coding of the document was done automatically by means of an algorithm searching for matching letters in the label. Unfortunately, as the administrative borders in Nepal are rather dynamic, several VDCs did not have an associated p-code.

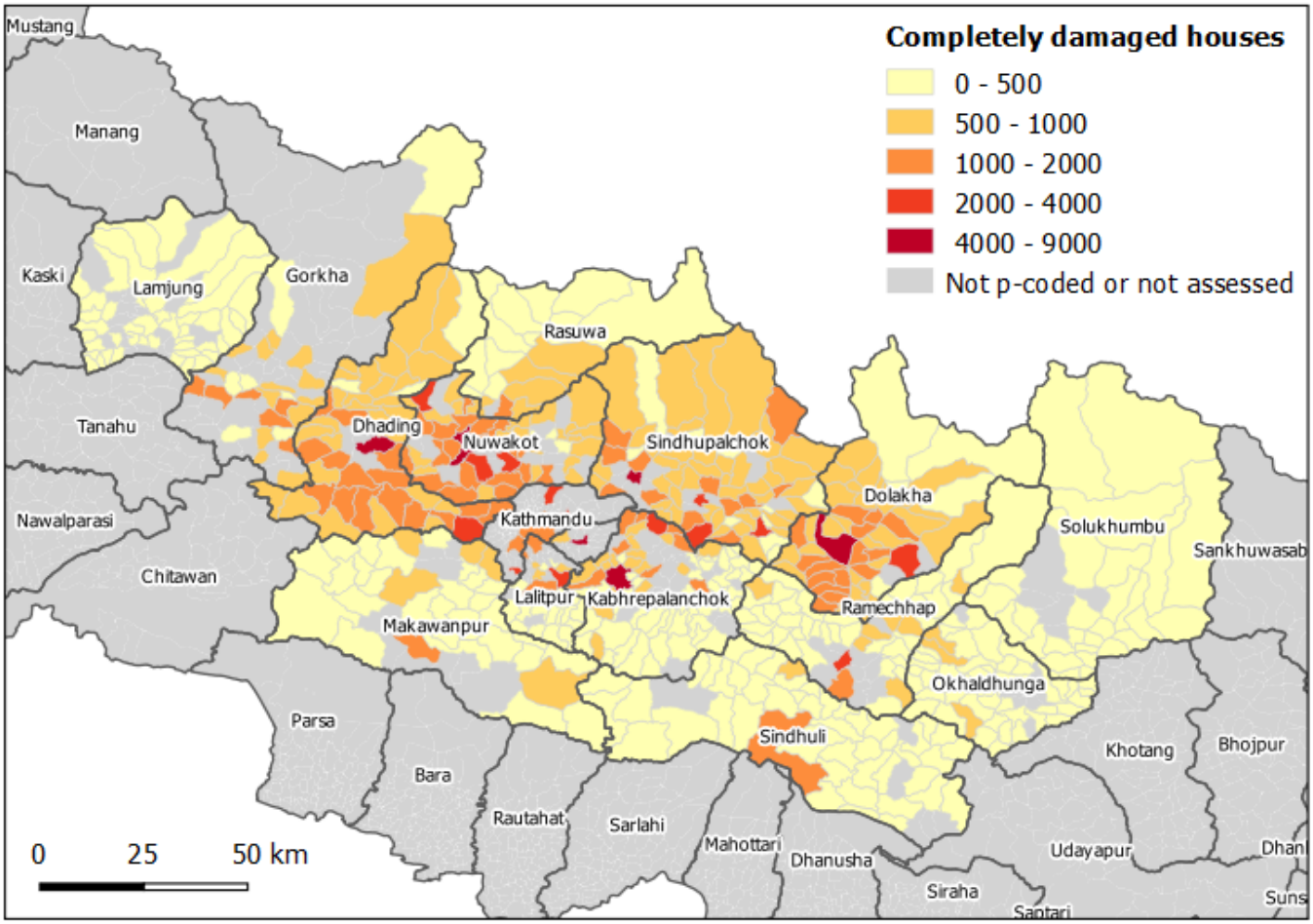

Figure 3: A visualisation of the number of completely damaged houses in Nepal due to earthquake Gorka.

Our champions