Bangladesh's Transport Infrastructure:

Preparing & Cleaning Data

Anna Noteboom 4564979
Auriane Tecourt 5397243
Floris Boendermaker 4655605
Job Onkenhout 4595769
Zara-Vé van Tetterode 4577701

Group 22

February 17, 2021

Figure 1: Dhaka on the Buriganga river, 2018 (foto: Auriane Tecourt)

1 | Introduction

With 166 million people living on 147.000 square kilometers, Bangladesh is one of the most densely populated countries in the world (World Population Review, 2021). The country sadly also suffers frequently from the negative effects of climate change, such as natural disasters. These factors, among others, result in the fact that Bangladesh can be seen as one of the most vulnerable countries to the effects of climate change (Kulp & Strauss, 2019). Infrastructural policies and projects can play a big part in increasing the country's resilience, making it less vulnerable (Ministry of Environment and Forests Government of the People's Republic of Bangladesh, 2008). However, in order to implement adequate policies, the data on these infrastructures should be correct, complete and interpretable.

According to Klir (2009), a Structure-System, such as Bangladesh's transport infrastructure, is defined by a set of lower systems. Therefore, to properly describe this system-level for policy-making purposes, the lower level - the Data System - needs to be correctly constructed. Therefore, this report describes the steps that were taken in order to prepare and clean multiple datasets for an eventual data-driven simulation, regarding Bangladesh's infrastructure. Before accurate simulations can be run for purposes of policy-making, errors and inconsistencies in the datasets need to be handled properly. Five different methods were used to realize this, each in relationship to the classes of data quality criteria, as proposed by Huang (2013).

The cleaning process focuses on the road-data and bridge-data, using the 'BMMS.xlsx' and the '_roads.tcv' files. Section 2 discusses the data and the initial modifications to the data. Section 3 explains the cleaning of the data related to the roads in Bangladesh. Thereafter, the fourth section presents the amendments made for the bridges. The fifth and last section discusses and explains the limitations of the data cleaning process.

2 | Data

This section firstly discusses the right categorization of variables, after which it digs deeper into the missing data and duplicates.

2.1 | Floats as strings (syntactic accuracy)

We first ensure that the variables are categorized correctly. In this case this means identifying numbers as floats, not as strings. In figure 2 numbers identified as strings. This issue is fixed in figure 3.

```
['N1',
 'LRPS',
 '23.7060278',
 '90.443333',
 'LRPSa',
 '23.7029167',
 '90.4504167',
 'LRPSb',
 '23.7027778',
 '90.4504722']
```

Figure 3: Floats as strings

```
['N1',
 'LRPS',
 23.7060278,
 90.443333,
 'LRPSa',
 23.7029167,
 90.4504167,
 'LRPSb',
 23.7027778,
 90.4504722]
```

Figure 2: Floats as floats

2.2 | Missing data and duplicates (pragmatic completeness and timeliness)

An exploration of the BMMS data file revealed the presence of duplicate entries with more (or less) missing information than the other corresponding entries. Huang (2013) categorizes this 'missing information for a given use' as pragmatic (in-)completeness. We therefore keep the row with the most information to maximize the pragmatic completeness of the dataset, and delete the duplicate.

A 'count_NaN' column was added showing the amount of missing (relevant) information for each entry. The relevance of information was hypothesised through reflection on the intended purpose of the data. The dataset is then sorted by road, LRPName, construction year and count_NaN. The construction year is used to ensure the timeliness of the dataset as defined by Huang (2013). The duplicates are then dropped from the sorted dataset based on the higher count_NaN value, resulting in less missing information and the preservation of relevant information. A lack of knowledge on the necessary variables for a future model causes a preservation of most columns.

3 | Roads

This section first discusses the roads that consist of only one data point, then outliers.

3.1 | One point roads

Some roads only have one data point, but a line needs at least two points to be drawn. There are two different explanations for this issue, thus classifying it as either a semantic accuracy or semantic completeness issue (Huang, 2013). If the road only has one data point due to an error in the road name, it is a semantic accuracy problem. If the road only has one data point because the other data points of the road are not in the dataset, it is a semantic completeness problem. Due to time constraints and lack of data, it is not possible to know which of these problems is occurring. These points are thus deleted.

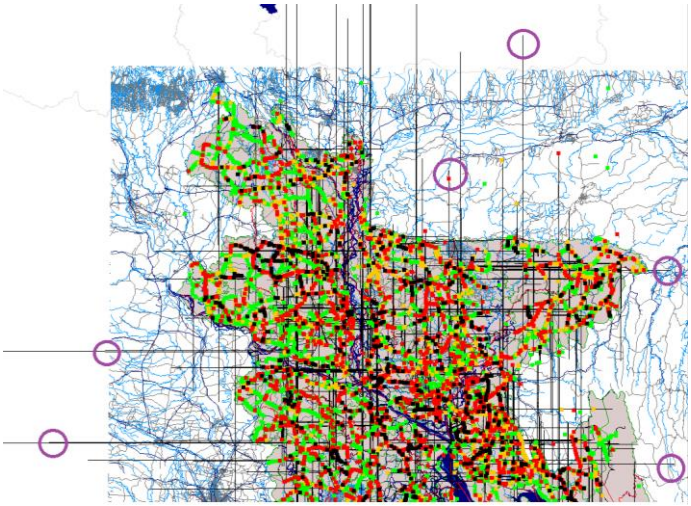

Figure 4: Before cleaning, outlier roads

3.2 | Outliers

A visible issue in the data is the outliers in the road, as can be seen highlighted with purple circles in figure 4. Since the roads don't look like that in reality, these outliers show a semantic accuracy issue of the data (Huang, 2013). After the cleaning method described below, the data looks like figure 5

Two types of outliers are tackled in the cleaning process: outliers that have two neighbouring data points that are correct, and outliers for which one of the neighbours is also an outlier. A data point was considered an outlier if it was over 0.1 in either longitude or latitude apart from its neighbour. This translates to a distance of 11 kilometres.

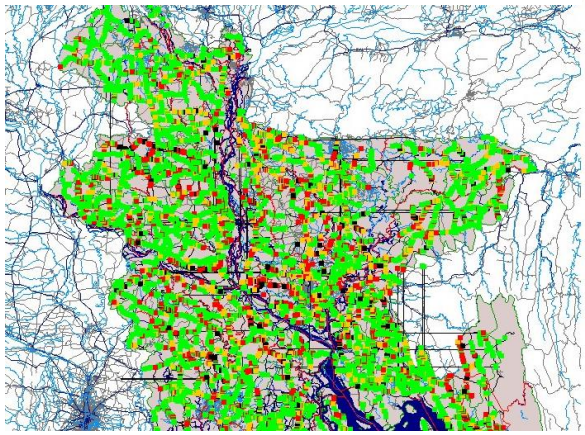

Figure 5: After cleaning, outlier roads

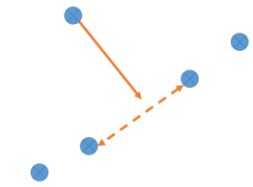

Figure 6: Outlying coordinate: mean

Outliers with two neighbouring correct data points: the outlying coordinate is changed to the mean value of the corresponding coordinate of the neighbouring data points. Figure 6 shows this case schematically.

Subsequent outliers: In this case, as depicted in figure 7, the mean value of the two neighbouring data points would not result in an improvement in the data. Therefore, the solution implemented here is to extrapolate the trend of the two previous data points to generate a more feasible location for the outlier.

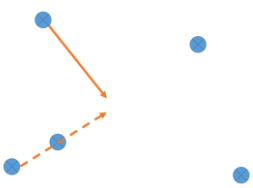

Figure 7: Outlying coordinate: extrapolate

To prevent missing a curve in the road due to the linearisation used in this method, the method is only applied if it does not bring the data point further away from its non-outlier successor. Figure 8 below compares a case where the method should be applied (left) to a case where it should not be applied (right)

The cleaning method for road outliers is still prone to errors as it may generate data that is not necessarily accurate, although it is usually still an improvement in accuracy. It does not address cases where there are too many consecutive (more than 3) outliers. This issue could be addressed through an expansion of the cleaning method.

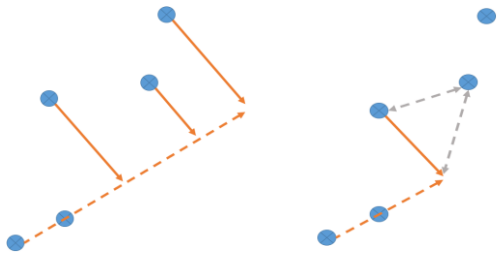

Figure 8: Compare cases

4 | Displaced bridges

As shown in figure 9, some bridges are on semantically inaccurate places, either because they are not on a road or because they are not on a river, or both. The bridges' dataset identifies bridges through the road on which they're on as well as their individual road point (LRP). Since the road dataset also identifies points with the road name and the LRP, our cleaning process replaces the coordinates of the bridges in the bridges dataset with those of the same LRP on the same road in the roads file. Bridges that could not be matched to a road are deleted. While this cleaning process increases semantic accuracy, it does so at the cost of semantic completeness: about 30% of the bridges are deleted. Since these bridges do not have corresponding road data, opportunities to use this data are limited, therefore the pragmatic completeness of the dataset remains unaffected (Huang, 2013). Figure 10 shows the results after this cleaning step.

To preserve as much data as possible, the method is refined to address two exception when the bridge should not be deleted although it does not match a road point:

- Some roads are missing in the roads dataset while bridges on that same road are included in the bridges data.
- Some LRPNames differ between the roads and the bridges data at the start of the road

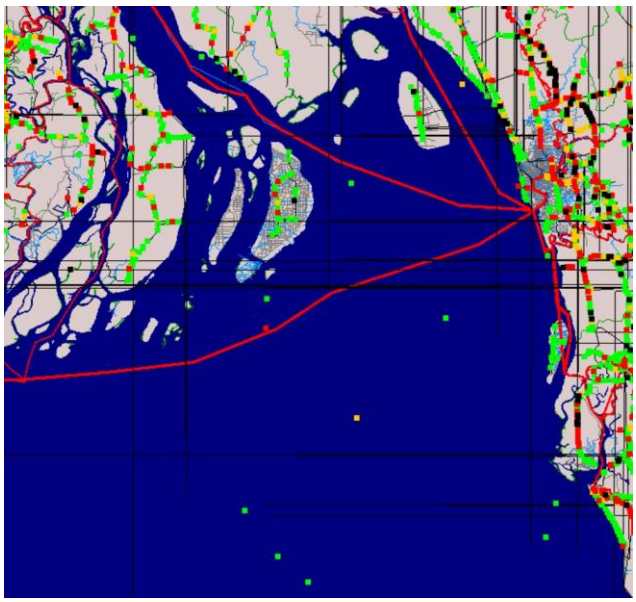

Figure 10: Bridges in the sea

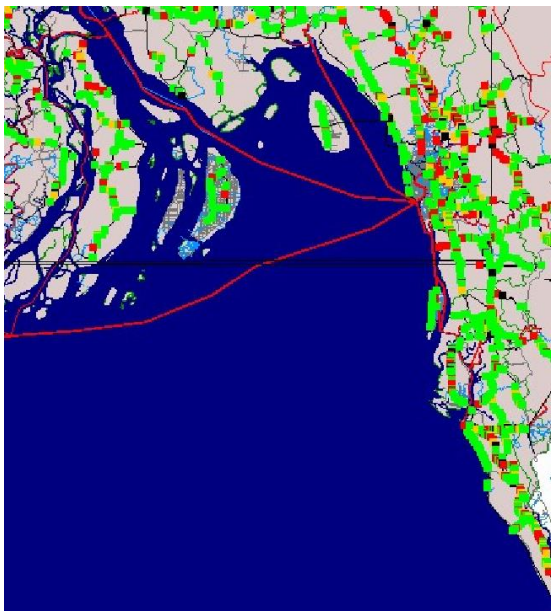

Figure 9: No bridges in the sea

4.1 | Missing roads: bridge does exist, but the road it is on is not documented

If the bridge file indicates that a bridge is on a road that is not included in the roads dataset, a road is created, connecting all bridges that state to be on that same road. The new road is very coarse, as it is only straight lines between bridges. If there is only one bridge on that new road, it is assumed to be an error in the bridge file instead of missing data in the roads file and this bridge is deleted.

4.2 | Missing connection: Different names, same coordinates

Sometimes, the LRPName of the bridge does not match the corresponding LRPName of the road even though they refer to the same point. One example here is that the LRP at the start of a road may be called LRPS, LRPSf, of LRPSg. This is a semantic inaccuracy (Huang, 2013) that is addressed in the cleaning process: when matching the LRPs of the bridges to the roads, these three names are considered as referring to the same point in a given road.

5 | Limitations & Discussion

Due to the limited time available for this assignment, not all observed errors and inconsistencies could be addressed in the cleaning process. However, we provide a short overview of other mistakes that could be addressed in future work:

- Bridgenames can have non-corresponding LRP names in the two datafiles. Currently we delete these cases, but with more time available these could be checked and preserved.
- Some outliers for the roads and bridges still exist. Further cleaning could portray why this is the case.
- Some roads are still missing. Remote sensing could be used to match the roads on a map with the roads on the road file
- Some bridges are missing. The road and bridges files could be matched with a river file to see where there is a crossing between a road and a river that is not accompanied by a bridge.

Furthermore, our cleaning process has limitations. Firstly, it results in the loss of 30% of all bridges in the dataset, which is quite significant. Some bridges are being deleted when they shouldn't be. This would be the case when a bridge can be matched to an existing road, while the LRP name that the bridge should be on cannot be found on that road. This can have unwanted implications for the future simulation, but the impact this has is yet unclear at this stage. In some cases, roads that connect bridges are coarse and do not have a realistic shape. This is due to them being drawn as connections between bridges without looking at other spatial variables.

We also encountered small errors in the dataset such as swapped longitudes and latitudes, and wrongly inserted decimals (e.g. 2.21 vs. 22.1). These 'easy' mistakes could be solved more accurately with a dedicated algorithm for each type of error instead of a broad one-size-fits-all algorithm such as here. The accuracy of the algorithm was deemed less necessary than the efficiency and consistency.

References

Huang, Y. (2013). Automated Simulation Model Generation, *Delft University of Technology*

Kulp, S. A., & Strauss, B. H. (2019). New elevation data triple estimates of global vulnerability to sea-level rise and coastal flooding. *Nature Communications, 10(1)*. https://doi.org/10.1038/s41467-019-12808-z

Klir, G. J. (2009). W. Ross Ashby: a pioneer of systems science. *International journal of general systems, 38(2)*, 175-188.

Ministry of Environment and Forests Government of the People's Republic of Bangladesh. (2008). Bangladesh Climate Change Strategy and Action Plan. ISBN 978-984-8574-25-6.

World Population Review.(2021). Bangladesh Population: Demographics, Maps, Graphs. Retrieved February 15, 2021, from https://worldpopulationreview.com/countries/bangladesh-population