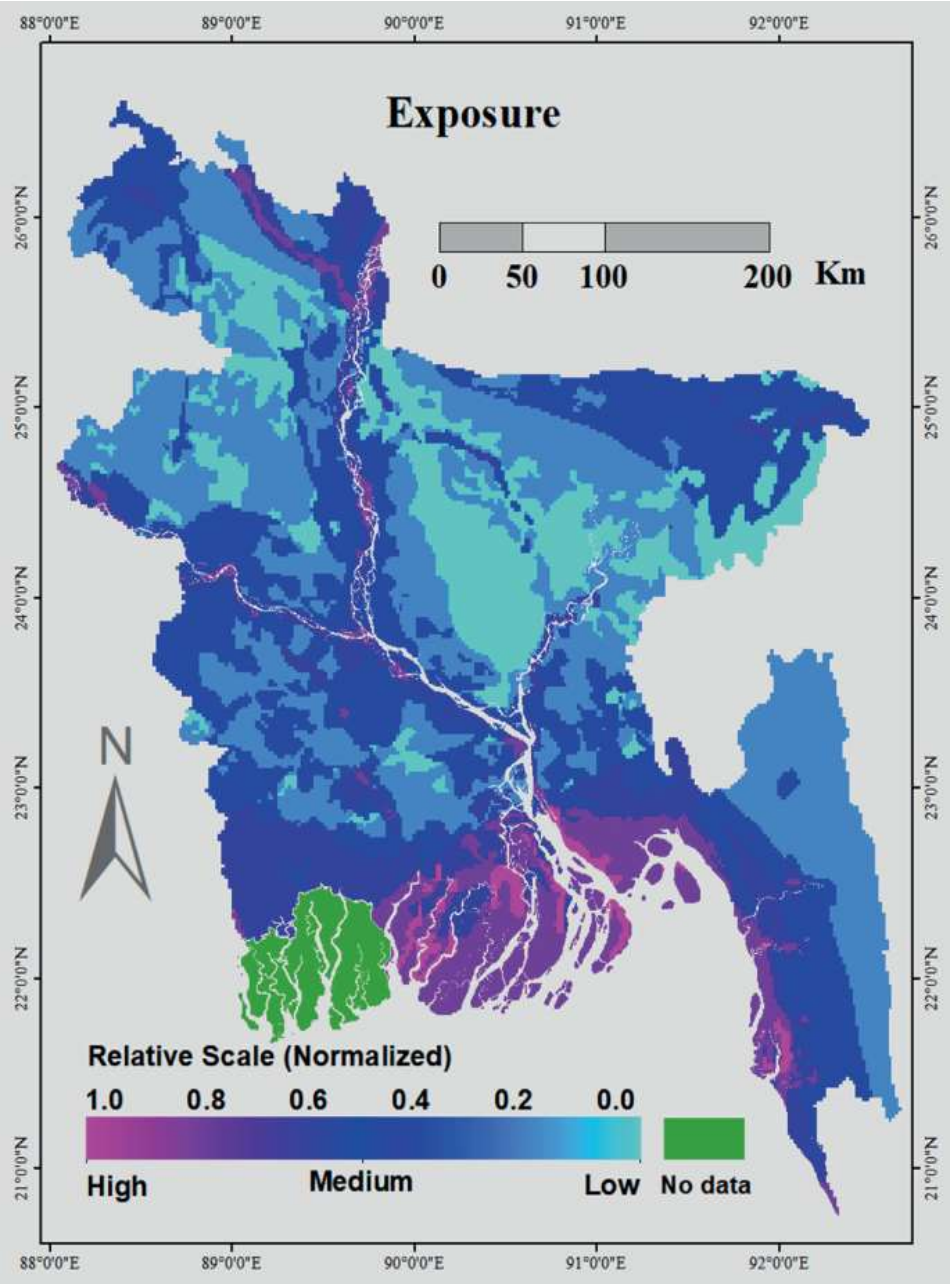

Figure 1: vulnerable zones in Bangladesh (Golam Azam et al., 2022)

Report Week 1:

Bangladesh's Transport Infrastructure

Preparing & Cleaning Data

Group 17

Ivan Temme (4955196)
Hidde Scheuer (4607325)
Philip Mueller (5809703)
Madalin Simion (5838363)
Nachiket Kondhalkar (5833884)

Supervisors: Alexander Verbraeck, Yilin Huang, Chamon Wieles, Vaibhavi Srivastava & Pascal Kampert
Course: EPA1352 Advanced Simulation
Date: 24 February 2023

Table of Contents

1. Introduction

Bangladesh, with its population of 172 million people living in just 147,000 square kilometers, faces a significant challenge due to its status as one of the most densely populated countries on the planet. The country is also highly vulnerable to the detrimental impacts of climate change, which include natural disasters that frequently ravage the nation. As a result, Bangladesh is classified as one of the world's most vulnerable countries to climate change. In light of this, there is a pressing need to enhance the country's resilience and reduce its vulnerability through effective infrastructural policies and projects.

Through the analysis of two primary indicators, criticality and vulnerability, the resilience of the infrastructure of Bangladesh can be assessed. Criticality refers to the amount of goods that can potentially be transported on a given road at a given point in time, while vulnerability refers to the likelihood of a given road or subsection being impassable (Korosh Mahmoodi, 2018). Both are significant factors in understanding traffic patterns in different hazard scenarios and identifying bottlenecks in the road network.

However, to manage these two primary criteria and to be successfully implemented, accurate, comprehensive, and understandable data on the infrastructures in question is of paramount importance. With this in mind, Bangladesh has made public data on its road network and bridges, which can be used to evaluate their resilience in the event of natural disasters. Klir (2009) suggests that a Structure-System, such as Bangladesh's transport infrastructure, is composed of various lower systems. Therefore, in order to accurately depict this system-level for the purpose of policy-making, it is essential to construct the lower level - the Data System - correctly. This report outlines the measures that were taken to prepare and refine several datasets for a data-driven simulation of Bangladesh's infrastructure. To ensure that precise simulations can be carried out for policy-making purposes, it is imperative to address inaccuracies and disparities in the datasets appropriately. In this regard, distinctive methods were utilized, each associated with the categories of data quality criteria proposed by Huang (2013).

To achieve the goal of assessing infrastructure resilience, data cleaning is a crucial step. This report will focus on data cleaning by introducing the data that will form the basis of our simulations. We will identify and prioritize data quality issues regarding roads and bridges in line with the final model purpose in chapters 3 and 4, respectively. Additionally, we will provide conceptual strategies for addressing these issues and visualization of the results in chapter 5.

By carrying out data cleaning and implementing effective infrastructural policies and projects, Bangladesh can improve its resilience and reduce its vulnerability to the detrimental impacts of climate change. This will ultimately benefit the country's population, economy, and overall well-being.

2. The Data

The Ministry of Transport has made multiple data sources available 'RMMS' contains data of all National, Regional, and 'Zila' roads in Bangladesh. 'BMMS' contains data of all bridges in Bangladesh. We will mainly work with three files that merge data from the mentioned sources: *_roads.tcv*, *Roads_InfoAboutEachLRP.xlsx*, and *BMMS_Overview.xlsx*.

__roads.tcv contains processed information from the RMMS data source about roads in Bangladesh. The roads are determined based on Location Reference Points (LRPs). This file shows all the LRPs related to each road. Each road has a starting Location Reference Point 'LRPS' and has one for the end of the road: 'LRPE'. In principle, all other LRPs are presented with a 3-digit number that indicates the distance in kilometers from the starting LRP. When multiple objects are identified within that kilometer, reference points are indicated with suffices a, b, c, etc. This also holds for LRPs that can have IDs like LRPSa (reference to an object location within the first km of the road). For instance, road 'Z1450' has starting point LRPS and since the next two LRPs are closer by than 1 km, the following points are LRPSa and LRPSb. The fourth point would then be LRP001. Each LRP has a coordinate value in the form of longitude (lon) and latitude (lat). Roads are visualized by connecting LRPs through edges on the map. The dataset contains 885 roads.

Roads_InfoAboutEachLRP.xlsx contains specific information about each LRP. This most importantly shows its specific road and its coordinates. An additional column (name) provides further information on the type of LRP (i.e., km post, bridge start, bridge end, etc.). Each LRP in the file also has a chainage value assigned to it. Chainage is used to measure the length of the road from start to the respective LRP. The file is sorted by roads and in order of increasing chainage.

BMMS_Overview.xlsx contains all data on bridges needed for our purpose. In this file, longitude (*lon*) and latitude (*lat*) give us the position of the bridge, *road* tells us on what road the bridge is located, *chainage* and *km* (identical data in both columns) specify the bridge's position on the road, and *condition* – ranging from A to D – tells us what condition the bridge is in. Every bridge also has an *LRPName*. Among other things, the year of construction, the number of lanes (*spans*), *length*, and the bridges' names are structure number are also provided in the file. In the following, we will only investigate the variables mentioned here, as all other variables do not explicitly suit the model purpose. The file contains 21407 bridges.

3. Types of Data Quality Issues

This section will discuss the different types of data quality issues using the categorization as provided by Huang (2013). Issues related to roads and bridges in Bangladesh will be identified and categorized. The results of this part will be handled in the next chapters where these data quality issues are prioritized and further assessed.

3.1 Roads

3.1.1 Data conformity (syntactic accuracy)

As mentioned in chapter 2, LRPs are the essential data points which determine the location of bridges and roads. After retrieving the data from the *_roads.tcv* file, the LRP coordinates turned out syntactic inaccurate. The data points were extracted from a file as a string separated by '\t', thus, the LRP coordinates were stored as a string. However, latitude and longitude coordinates are expected to have a float data value. Another error was observed in *Roads_InfoAboutEachLRP.xlsx*. Two strings namely, *Naogaon RHD starts* and *Palashban Gaibandha* occur in the *road's* column as names of Highways, but they are clearly not and need to be dropped.

3.1.2 Missing values (completeness)

Next, missing data was reviewed. Huang (2013) makes a distinction between semantic completeness and pragmatic completeness. Road R316 contains only one point. We assume that LRP are missing. Since the number of roads provided in *_roads.tcv* and *Roads_InfoAboutEachLRP.xlsx* are identical, we assume the roads to be semantically complete. It is important to note that we cannot assess whether the similarity of both files actually stems from one file being derived from the other.

3.1.3 Coordinate decimals (syntactic accuracy)

Decimal places for coordinates in *Roads_InfoAboutEachLRP.slxs* differ between three and seven. That corresponds to a maximum error margin of around 100 m. This minor syntactic inaccuracy is inside a neglectable error margin for the model purpose. The column 'name' in *Roads_InfoAboutEachLRP.slxs* is syntactically inaccurate due to a multitude of different typos that render the use of this column in its current state nearly impossible.

3.1.4 Road spikes (semantic accuracy)

Semantic accuracy is crucial for the intended purpose of the model, particularly when dealing with coordinate data. The primary source of inconsistency in the road data arises from erroneous LRP coordinates, causing unusual spikes in the plotted map. These errors are believed to result from data collection typos, including the inputting of multiple consecutive mistakes. It is also plausible that the database may have swapped the longitude and latitude coordinates during input.

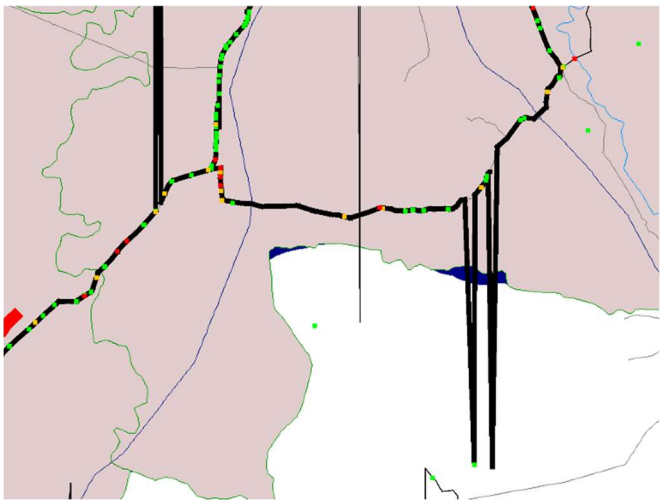

Figure 2: Spikes, stemming from false coordinate values at the corresponding LRP entry.

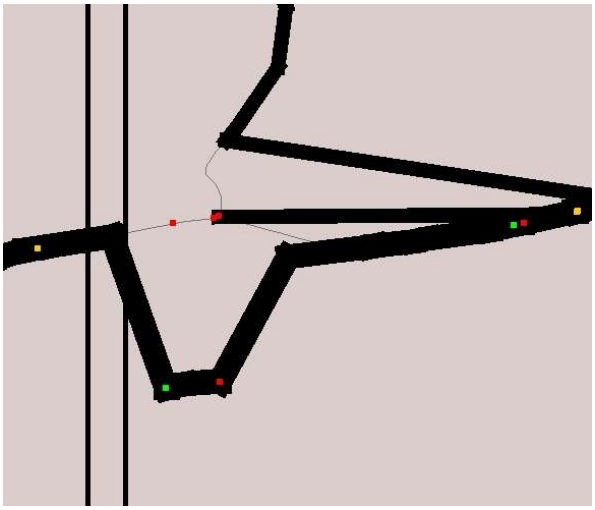

Figure 3: multiple coordinate errors in a row.

If a road is maintained, e.g. straightened or redirected around a town, the chainage changes. It is always difficult to determine whether the chainage is consistently changed for all LRPs that relate to the road. The LRP numbers (originally containing the chainage in full kilometers) do not change. Therefore, for longer roads, an LRP number of e.g., 100 can belong to a chainage of 96 or 104, or even a number that deviates more. This also resembles a semantic inaccuracy, potentially also with regards to coordinate values.

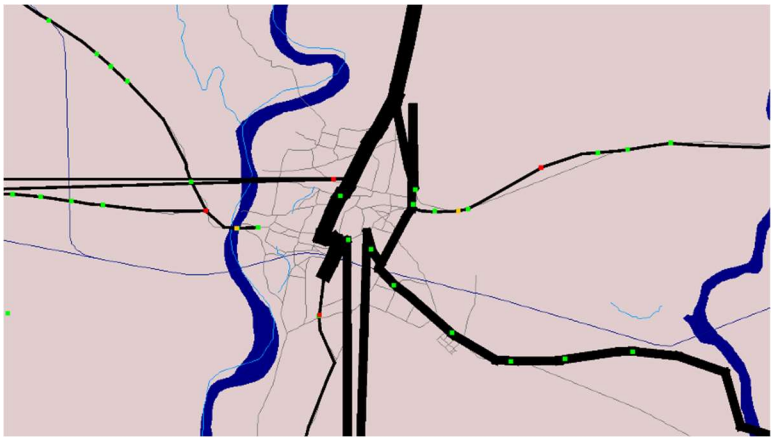

Figure 4: Potential LRP and chainage errors. When redirecting roads around cities, chainage and LRP coordinates are not always correctly updated.

3.2 Bridges

3.2.1 Mapping Inconsistency

Since information on LRP and the associated road are included in the bridge data, we can map bridges to their corresponding roads. Some bridges are referenced to roads that do not exist in the road data, however. We have detected 227 of such mapping inconsistencies.

3.2.2 Completeness

More than 14% of bridges have no data on their spans and construction year. 94 bridges have no coordinates. There are files provided, e.g. *Bridges.xlsx* whose bridges have no LRP. Using these bridges for plotting and assigning these bridges to roads is thus not possible. However, 18327 bridges are being plotted in the uncleaned data, after removing duplicates (see below). We thus render the bridge data to be semantically complete.

3.2.4 Semantic Accuracy

There are 3080 duplicates in bridges, which adds up to less than 2%. The ID-columns *StructureNr* has no duplicates, however. Duplicates were instead detected by merging the road and LRP column. It seems likely, that when an old bridge is replaced by a new one, the new bridge is added to the dataset and a new StructureNr is assigned to it, without deleting the old bridge from the dataset. The timeliness (Huang, 2013) of the bridge data hints to this possibility. Construction years for bridges date back as far as 1950.

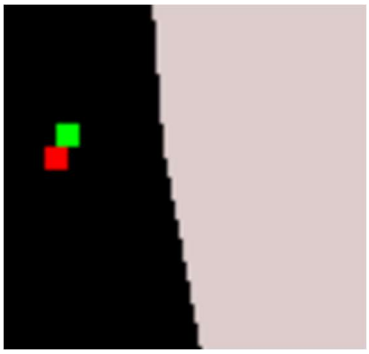

Figure 5: Two bridges on the same road and in close proximity. It can be assumed that the green bridge (quality = A) has been built to replace the red bridge (quality = D). It can furthermore be assumed that the red bridge dates further back in the data set.

Since the bridge data includes road and LRP information, bridges can be mapped to their corresponding roads and the correct LRP on that road. In some cases, the coordinated for the bridge LRP deviates from the coordinates of the road LRP. This could be due to a typo while entering the data or from merging datasets. In some cases, as can be seen in Figure 6, one bridge has multiple LRP. As explained in chapter 2, these LRP usually differ only in the suffix. Every LRP has a unique bridge name associated to it in the data, however.

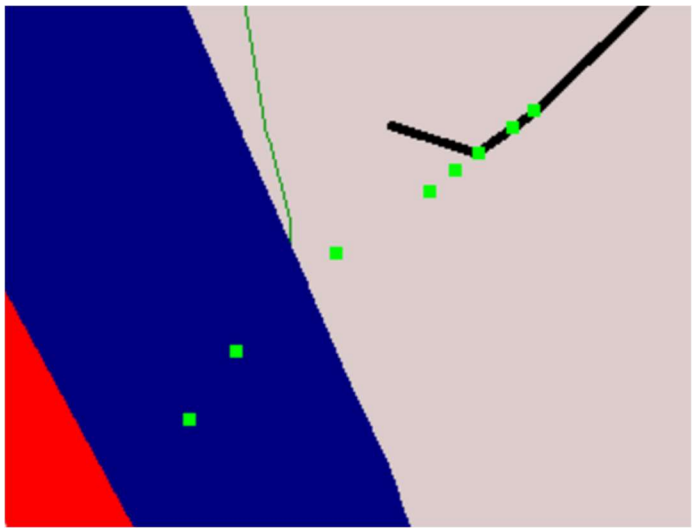

Figure 6: Multiple LRP reference to the same bridge structure.

4. Prioritizing Quality Issues

The following assignments will analyze traffic flow throughout Bangladesh, given different disaster scenarios. The model allows us to identify critical parts in the network. If these critical parts are becoming impassable, regions of Bangladesh are entirely cut off from potential disaster help. In order to produce reliable results, the model should include as many roads and bridges as possible, given they exist in reality. During data cleaning, we therefore focused mainly on completeness, semantic accuracy, and mapping consistency. All other errors solved were mostly solved to suit this purpose. Furthermore, critical parts in the network can only be assessed if the coordinates of LRP's for both roads and bridges are correct. Within the context of semantic accuracy and mapping consistency, we prioritized coordinate errors. In an attempt to control pragmatic completeness, we either corrected or kept outliers rather than deleting them.

By correcting coordinate errors for road LRPs we were able to improve accuracy in the simulated road length. Travel time and amount of goods transported per time can thus be assessed more accurately in the simulation. This helps to correctly assess the criticality for each road. Furthermore, intersections can be placed as precisely as possible. This increases accuracy in interpreting network vulnerability, as we can more precisely locate critical parts of the network. The same goes for correctly placing bridges. Since bridges are potential weak points in scenarios such as storms and floodings, they must be placed on the correct roads and coordinate.

The bridge quality is important, as it indicates a likelihood of collapse. However, there is no way to assess the validity of the provided data in a real-world context. This potential quality issue thus stays undetected here. Bridge length, starting point, end point, and number of spans were not addressed in data cleaning, as these corrections do not add to simulation accuracy. We assume here that if a bridge collapses at one point, this renders the entire bridge impassable on all spans.

5. Solution Approaches to Quality Issues

5.1 Solution to Roads

We first looked at the coordinates of road LRPs. For this purpose, we worked with *_roads.tcv*. We tried multiple approaches to detecting and correcting outliers.

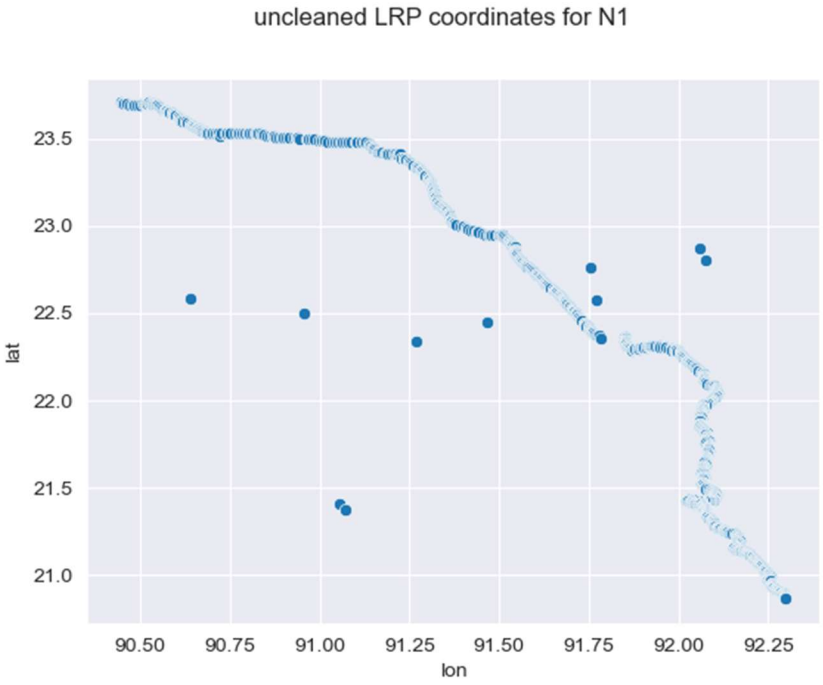

Figure 7: LRP coordinates for National Highway 1 (N1). Outliers can be clearly seen here.

We started with detecting coordinates outside of the boarders of Bangladesh. These outliers were replaced by average coordinates. We then moved on to look into each road, individually.

In the first approach, we computed a mean longitude and latitude over all LRP on a given road. All values that fall outside a tolerance margin of +/- 1% of that average were classified as outliers and overwritten with the mean coordinate. As this approach fails to capture outliers that can easily be detected visually, we decided to apply a more accurate method.

In a second approach we employed a type of moving averages. We determined outliers by computing the median over the five preceding and following points of each LRP. If an LRP lies outside a [median * 99.9%; median * 100.1%] margin it is overwritten with the median. We tested the algorithm on the first road in the data set. After two iterations, most points are sufficiently corrected for our purpose. After two iterations, the algorithm starts significantly overcorrecting coordinates. We used two iterations as a benchmark and applied the function to all roads in the data set.

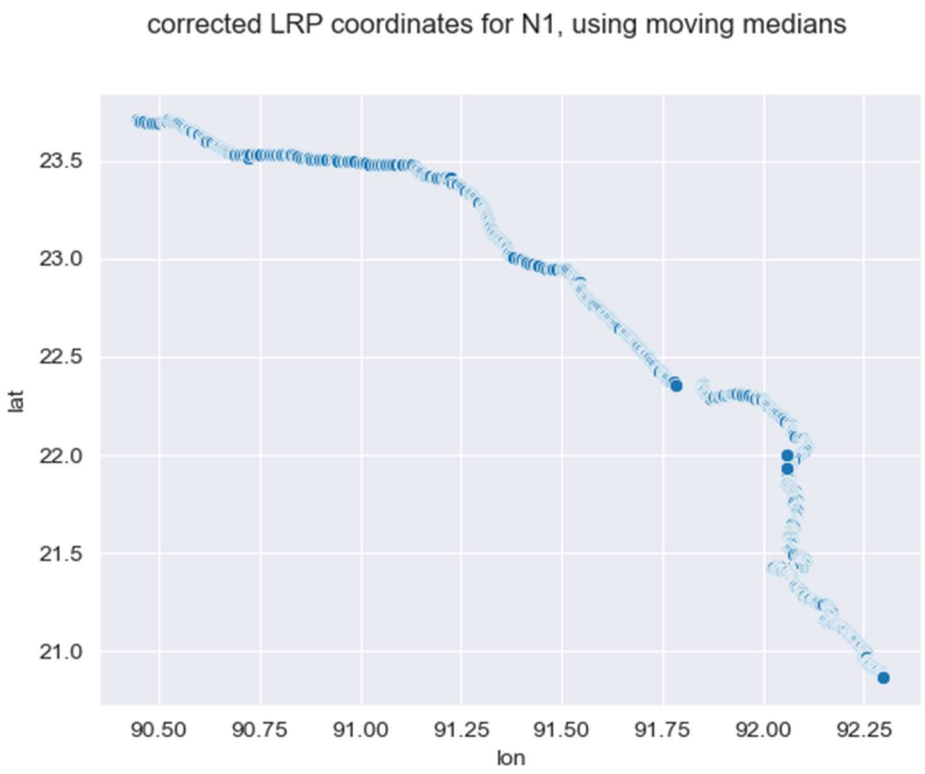

Figure 8: Corrected LRP coordinates for National Highway 1 (N1).

This approach faces three main limitations: Firstly, it struggles to recognize outliers when multiple coordinate errors were entered to the dataset successively (see Figure 3). Although the median is already robust towards outliers, if more than 5 values inside the search frame are skewed, the median gets skewed. As a result, outliers are not detected reliably and undercorrected. Thirdly, the corrected coordinates were not checked against publicly available data on street locations, such as 'OpenStreetMap'. With this approach a more precise number of iterations for the correction method could be determined for each street individually (OpenStreetMap, 2023).

Another, potentially more promising, solution would be to incorporate chainage information. By computing the Euclidean distance between coordinates of all LRPs and comparing them to the chainage, we could classify all LRPs whose corresponding coordinates are significantly further apart from each other. There are limitations to this approach too, however. As pointed above, we have no certainty on the semantic accuracy of chainage information in the provided data. We do have reason to believe that a proportion of the chainage information is not correct. The approach is also more computationally expensive.

5.2 Solution to the bridges

When analyzing the bridges data, we started with creating a UniqueID for each bridge using the road name and the LRP for the bridge. This would then allow us to check for duplicates where bridges exist on the same road and same point. This may be due to the destruction of an old bridge and a new bridge being constructed

at the same location or a new bridge being constructed right next to an old bridge on the same road. On observation of this data, we found 3080 duplicate bridges with the same LRPs. Since bridges on a particular road exist, there is little to gain by multiple bridges as we are not yet looking at any other attributes such as capacity, width, age or quality. Hence, we chose to drop the duplicates as functional connectivity was given priority.

We also created a list of these IDs and cross checked them against the road IDs from the *Roads_InfoAboutEachLRP.xlsx* file to observe if the Bridges can be placed onto roads by merging the datasets. We can see that a further 5766 bridges cannot be placed on roads directly. This does not take into account the actual coordinate placement of the bridge onto the road. Due to time limitations, we were unable to complete the processing of this data, but we have explained our tentative methodology in the markdown files in *Bridges.ipynb*. Bridges, referencing to roads that are missing in the road data can only be deleted.

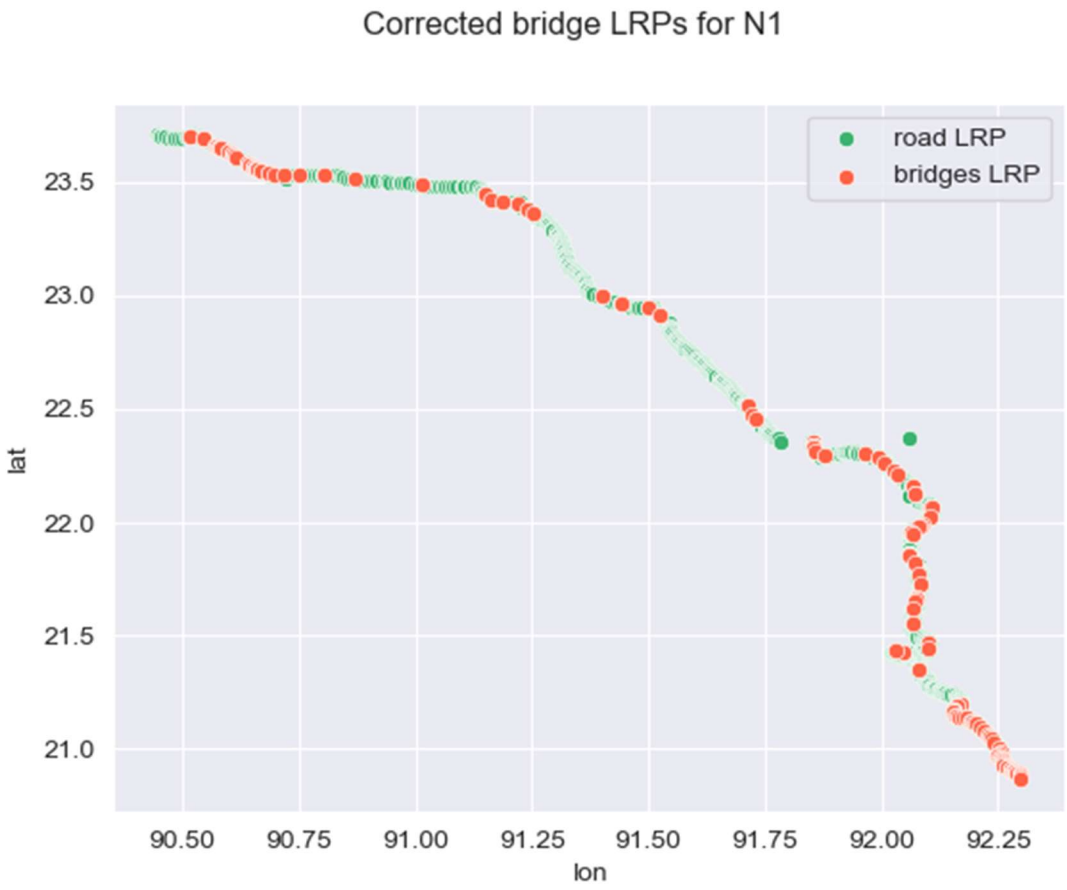

Figure 9: LRPs for the National Highway 1 (N1), depicted in green, and all bridges on N1, depicted in orange.

References

Korosh Mahmoodi, Bruce J. West, Paolo Grigolini (2018), "Self-Organized Temporal Criticality: Bottom-Up Resilience versus Top-Down Vulnerability", Complexity, vol., Article ID 8139058, 10 pages, 2018. https://doi.org/10.1155/2018/8139058

Golam Azam, M., Mujibor Rahman, M. (2022). Identification of Climate Change Vulnerable Zones in Bangladesh Through Multivariate Geospatial Analysis. In: Jana, N.C., Singh, R.B. (eds) Climate, Environment and Disaster in Developing Countries. Advances in Geographical and Environmental Sciences. Springer, Singapore. https://doi.org/10.1007/978-981-16-6966-8_5

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

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

Kulp, S.A. and Strauss, B.H. (2019) New elevation data triple estimates of global vulnerability to sea-level rise and coastal flooding. Nature Communications 10: 4844, 12pp. DOI: https://doi.org/10.1038/s41467-019-12808-z

Neha Rai, Saleemul Huq & Muhammad Jahedul Huq (2014) Climate resilient planning in Bangladesh: a review of progress and early experiences of moving from planning to implementation, Development in Practice, 24:4, 527-543, DOI: 10.1080/09614524.2014.908822

OpenStreetMap. (2023). OpenStreetMap. https://www.openstreetmap.org/#map=12/47.3775/8.4766

World Population Review.(2023) Bangladesh Population: Demographics, Maps, Graphs. Retrieved February 24, 2023, from https://worldpopulationreview.com/countries/bangladesh-population