Data Quality Issues for Data-Driven Simulation

ADVANCED SIMULATION

EPA133A

Authors:

Bettin Lorenzo (6132928)
Dell'Orto Alessandro (6129161)
Le Grand Tangui (6172075)
Precup Ada (9876543)
van Engelen Ralph (4964748)

February 21, 2025

TU Delft Delft University of Technology

Contents

Introduction

Today, Bangladesh faces the threat of natural disasters, including earthquakes, floods, and cyclones, which could worsen with the advent of climate change. In this context, the World Bank wishes to assess and improve the country's existing infrastructure. To do so, data was collected on Bangladesh's roads and bridges, which were mapped using Java.

In the world of data science, verifying that the data being analyzed is of sufficient quality and fit for purpose is imperative. Publications such as Marsden and Pingry (Marsden & Pingry, 2018), Dong (Dong, 2019), and The Economist (2013) deplore the lack of attention paid to data quality in research, as some scientists warn of a study reproducibility crisis (Baker, 2016).

To address this, Marsden and Pingry (2018) call on readers to consider whether articles they come across address data quality, whether their data-gathering methodology is sufficiently detailed, and whether the authors answer seven key questions behind their data selection process: "What?", "When?", "Where?", "How?", "Who?" and "Which?".

Figure 1: Victims of flooding in Bangladesh, August 26, 2024. From "Bangladesh floods leave 71 dead, fears of waterborne disease rise", by R. Paul, M.P. Hossain. 2024 September 4, REUTERS

The objective of this report is thus to scrutinize the data, solve errors that would have a significant, immediate impact on the analysis, and guide future research using this data. To this end, the categorization of data quality, as laid out by Huang (HUANG, 2013), will be used. It makes use of three semiotic levels (Price & Shanks, 2005) to define three corresponding data quality categories: syntactics (the degree to which stored data conforms to database rules), semantics (the degree to which stored data corresponds to the represented referents that are relevant to the purposes for which the data is stored) and pragmatics (the degree to which stored data is suitable for a given use). These are further divided by criterion, which will be detailed in the paper where relevant.

The following data quality assessment and data cleaning will not be exhaustive. In particular:

- This document will only focus on **roads** and **bridges**, as indicated in the Assignment description.
- The data was analyzed and cleaned through the **derived files** for a matter of time constraint.
- Original datasets were only observed as a beginning step to have in mind their content and structure.

Therefore, what follows is a systematic overview of the issues in the roads and bridges data and the outline of a set of strategies to solve all these issues. Moreover, algorithmic solutions are proposed after assessing which issues hold the highest priority for simulation purposes.

1 Data Quality Issues

Issues in the raw data were assessed by directly investigating the tables or using basic exploratory data analysis (EDA) using Python. The results of this assessment are associated with one of the *data quality criteria* outlined by Huang (Huang, 2013).

1.1 Bridges

1.1.1 Naming issues

Identical names: Some bridges have identical names but different location reference point (LRP) names or properties. For example, there are two bridges named "HAR BAG BOX CULVERT", which have different LRPs (LRP319a and LRP319b), length (4.5 and 3), slightly different chainage, width, and lat/lon (both from bsc1). Most likely, one of the bridges has a different name, or they both lack information to distinguish them. *(Semantic Accuracy)*

Different names: Other bridges appear to be the same because they have almost identical data but different (yet often similar) names. For example, bridges "Sarkarpara Bridge" and "SARKAR PARA BRIDGE" have identical entries for road, Km, and LRP name but differ slightly in other attributes, possibly due to missing data. Since they differ in location estimation (bsc1 and road_precise), the redundancy could be due to TU Delft position processing. Most likely, these entries refer to the same real-world bridge but are inconsistently named. *(Mapping Consistency)*

Left/Right bridges: Many bridges are divided into their left and right sides, but the notation is inconsistent: some use numbers (1 and 2), others use "L" and "R"; some use a mix of different fonts and styles. Additionally, slight errors in width, length, or positioning can cause spatial overlap. This inconsistency makes it harder to identify and correctly categorize bridges. *(Syntactic Consistency)*

Formatting: Some bridge names include the bridge type (e.g., "RAIPUR BOX CULVERT"), while others use different formats, such as adding a dot at the end. The lack of a uniform naming convention makes duplicate detection more difficult. *(Syntactic Consistency)*

1.1.2 Missing Data

Unnamed bridges: Some bridges lack names and only display a dot character. This missing information makes classification and referencing difficult. *(Semantic Completeness)*

Width, construction year, and span: Around 3000 bridges are missing these three values, particularly those derived from TU Delft processing rather than bcs1. The absence of these attributes reduces data completeness and usability. *(Semantic Completeness)*

Bridges with no latitude/longitude: Around 100 bridges are missing location data, which makes spatial analysis unreliable and impacts data usability. *(Semantic Completeness)*

Missing entries: A simple online query reveals that the longest bridge in the country, *Padma Bridge*, is missing from the dataset because construction year data stops at 2013. This suggests that bridges built after 2013 are not included, leading to an outdated dataset. *(Timeliness)*

1.1.3 Structural Data Issues

Too short bridges: Some bridges are recorded as being less than 1m long, which is highly unlikely and suggests incorrect data entry. *(Semantic Accuracy)*

Too wide bridges: Two bridges are recorded with widths of 702m and 100m, which are unrealistic values likely caused by data entry errors. *(Semantic Accuracy)*

1.1.4 Spatial and Chainage Issues

Bridges not on a road: Some bridges do not align with any road in the dataset, indicating possible errors in coordinate placement or missing road data. Most likely, either the coordinates are incorrect or the road data is incomplete. *(Semantic Accuracy)*

Precise and extrapolated bridges: Differences between measured and interpolated bridge locations raise concerns about spatial accuracy. Most likely, errors in interpolation or measurement cause these discrepancies. *(Semantic Accuracy)*

Misplacement: Some bridges are placed far from roads, in the sea, or outside Bangladesh's borders, likely due to incorrect coordinate entries. *(Semantic Accuracy)*

1.2 Roads

1.2.1 Missing Data

No data: Some roads lack any LRP information and only display names, making it difficult to establish their full structure. Most likely, key attributes were not recorded or extracted properly. *(Pragmatic Completeness)*

No starting LRP: Road N045 starts with LRP036c instead of LRPC, which deviates from the expected road structuring. Most likely, this is a data input or processing error. *(Syntactic Accuracy)*

No ending LRP: Roads should correctly terminate with an LRPE, but some do not, requiring an algorithm to ensure completeness. Most likely, road termination points were not systematically enforced. *(Syntactic Accuracy)*

1.2.2 Spatial and Chainage Issues

Incorrect Chainage: Java visualization reveals that many roads have highly misplaced LRP lat/lon values, causing abrupt and unrealistic deviations. Most likely, incorrect geolocation data or chainage calculations are responsible. *(Semantic Accuracy)*

2 Prioritization of issues

The role of simulation in research and policy is still being debated today , especially its suitability for predictive purposes; however, it is clear that simulations are useful to get insights on the problem structure and unfolding, even if the simulation model does not fully mirror the real world (Baker, 2016).

The main requirement for a simulation model, therefore, is not its accuracy but being **fit for purpose**. Therefore, the priority for this model is to function properly and to be insightful on the problem unfolding, even if its results are not accurate, complete, or up-to-date (Oreskes, 1998).

Intuitively, the most important thing for this model is that *roads work as roads* and *bridges work as bridges*. On that line, the encountered issues were given the following priority, from less to more important:

- **Pragmatic Completeness:** Missing LRPC at the beginning of the road doesn't prevent the visualization from running, and chainage is not inevitably affected. However, it could lead to unexpected simulation errors or inaccuracies further on with the suimulation and, therefore, should be fixed.
- **Syntactic Consistency:** The simulation can run even if the naming is not consistent. Nonetheless, the purpose of the simulation is to identify vulnerable infrastructures, so systematically aligning all bridge naming can help in communicating results. Moreover, it is a useful tool to deal with duplicates.
- **Timeliness:** As with syntactic consistency, the simulation can run and be insightful on the unfolding of Bengali road mobility and its link to infrastructural vulnerability even if the data comes from before 2013. Also, if data becomes available, the model can be easily updated.

- **Semantic Completeness:** Missing width data for bridges could lead to simulation problems, and missing construction year and span impacts the quality of vulnerability assessment.
- **Syntactic Accuracy:** Ensuring the data conforms to its domain is important for pragmatic suitability. A numeric value in a string column would, for example, not be able to be processed in the analysis and would result in the model not running or missing values.
- **Mapping Consistency:** Ideally, an object such as a bridge would be mapped into our software with a single key. Issues arise when the mapping is inconsistent, for example, if a bridge is assigned two keys, meaning it is present twice in the simulation instead of once, or if the wrong key is assigned to a bridge, possibly leading to overlapping.
- **Semantic Accuracy:** The conformity of a data value to its real-world value is very important in this case. If the location of a data point representing a bridge or a road is offset, the whole simulation is incorrect as distances traveled are significantly altered. This is by far the most important criterion here.

That being clarified, the implementation will follow different approaches for the two datasets. For the bridges dataset, the cleaning process will begin by ensuring consistency in bridge names. This is essential for accurately assessing vulnerable bridges later on and, most importantly, for removing duplicates. After that, the focus will be on correcting as many bridge misplacements as possible. On the other hand, the roads dataset does not require formatting adjustments for road names, as they are already consistent, and duplicates are not an issue. Instead, road cleaning will prioritize major roads that carry the most traffic, focusing on fixing significant misalignments that could impact simulation outcomes.

3 Conceptualization of Data Quality Solutions

3.1 Strategy for Addressing Data Quality Issues

A structured and systematic approach has been adopted to address data quality issues and obtain a robust, high-quality dataset. This strategy covers the following points:

- **Standardization and Consistency Enforcement:** Creating uniform, global naming conventions that include formatting rules and structured data entries in order to improve consistency across the whole dataset.
- **Validation and Error Detection:** Implementing threshold-based validation, outlier detection, and logical checks to flag and rectify erroneous values.
- **Data Imputation and External Verification:** Leveraging historical data, geospatial datasets, and external sources to fill in missing values and confirm accuracy.
- **Geospatial Corrections and Alignment:** Applying geospatial interpolation, proximity analysis, and network-based corrections to ensure accurate road and bridge placements.
- **Visualization and Mapping:** Geospatial visualization is used to detect misalignments and identify missing segments.
- **Distance and Outliers:** Implementing distance calculations and spatial clustering to flag potential data inconsistencies.
- **Automated and Manual Review Processes:** Combining algorithmic techniques with manual validation for ambiguous cases that require human judgment.

3.2 Conceptual Solutions for Bridges

When it comes to addressing bridge-related data issues, the approach begins by refining how names are managed. Naming conventions and fuzzy matching are needed to detect duplicate or similar names, while standardized formatting minimizes variations from differences in capitalization or punctuation. For bridges that are missing critical details, default names, and estimated values are derived using geospatial inference and historical data, ensuring that every bridge is identifiable. In cases where geolocation data is absent location inference techniques are applied to estimate accurate coordinates. Any bridges with unrealistic dimensions are scrutinized using range constraints and statistical analysis to pinpoint outliers. Lastly, when bridges are misaligned with the road network, geospatial road network data is used to realign them correctly, ensuring consistency with the overall infrastructure layout.

3.3 Conceptual Solutions for Roads

For road data, incorrect elements are addressed through interpolation techniques, filling in the gaps with informed approximations. A rule-based validation process is implemented to tackle roads lacking clearly defined start or end points, ensuring every road segment begins and ends at the appropriate LRPs. Misaligned or incorrect chainage is then corrected using spatial smoothing algorithms alongside verification of the coordinate system to adjust the segments most accurately.

4 Implemented Solutions

4.1 Data Quality Solutions for the Bridges

Six issues were resolved by making the algorithm to resolve the prioritized data quality issues in the bridges file. These are as follows:

- Bridges not inside of Bangladesh
- Bridges not connected to land
- Bridge name formatting errors
- Duplicate bridges
- Swapped longitude and latitude
- Bridges that are too small
- Bridges not connected to roads

In the following paragraphs, the solutions to these problems are described.

The data contained several bridges that were either outside of Bangladesh or not within its onshore boundaries. To remove these bridges, a bounding box containing the approximate geographical coordinates of the bounds of Bangladesh was established. Next to this, an approximate land polygon for Bangladesh's landmass was made. Then, the bridges that were not in this bounding box were filtered out to ensure that all bridges were inside Bangladesh. Bridges that were not located on land were filtered out by checking whether their coordinates fell within a predefined land polygon. Using the `contains()` method, a Boolean column was created, and only the bridges where this condition was true were kept, ensuring that incorrectly placed bridges in water bodies or coastal areas were removed.

A function, `clean_name`, was defined to standardize bridge names by converting them to uppercase, removing extra spaces, and eliminating special characters. It also removed the word "BRIDGE" to ensure uniformity across entries.

The function `extract_lr`, was made to extract Left/Right indicators from bridge names and standardize them as 'L' or 'R'. It converted the name to uppercase and checked for patterns matching "LEFT", "RIGHT", "L", "R", "1", or "2", assigning the corresponding standard notation. The function was applied to the `name` column, creating a new column called `LR_indicator`.

Fuzzy duplicates (duplicates with nearly identical character values) and spatial duplicates (bridges within a small lat/lon distance) were removed from the dataset. A function `find_duplicates` was made to compare each bridge's name within the same road group using a similarity threshold, latitude, longitude, and chainage differences to detect near-duplicates. If the pairs the function iterates through exceeds the threshold or falls within the small spatial differences, then it is marked as a potential duplicate. In the `merge_duplicates` function, the duplicate rows are then filled in using other row data, and the duplicate entries are removed.

Then, bridges with swapped latitude and longitude values were corrected by creating a Boolean mask that contained bridges where the latitude was larger than the longitude. Next, the latitude and longitude values were swapped for the affected rows using `.loc[]`.

Moreover, bridges with a length of under 1 meter and a width under 50 meters are filtered out by creating thresholds and then filtering the length and width column based on these thresholds.

Finally, the bridges that were not connected to roads were filtered out. A **KDTree-based spatial search** was used to make sure all bridges within a **100 meters** of a road are considered connected. This prevented

excessive removal of bridges due to small coordinate differences. Only the bridges that were near roads were retained.

This was done by performing an inner join between the dataset of the bridges and the dataset containing the information on the roads on the latitude and longitude columns. This ensured that only bridges containing coordinates that were also present in the roads dataset were retained, and loose bridges were filtered out. The results were plotted in figure 2. This figure clearly shows that the algorithm has removed outliers and ensures that the dataset contains only relevant infrastructure data. The cleaned dataset contains 13905 entries as opposed to 21259 of the original dataset.

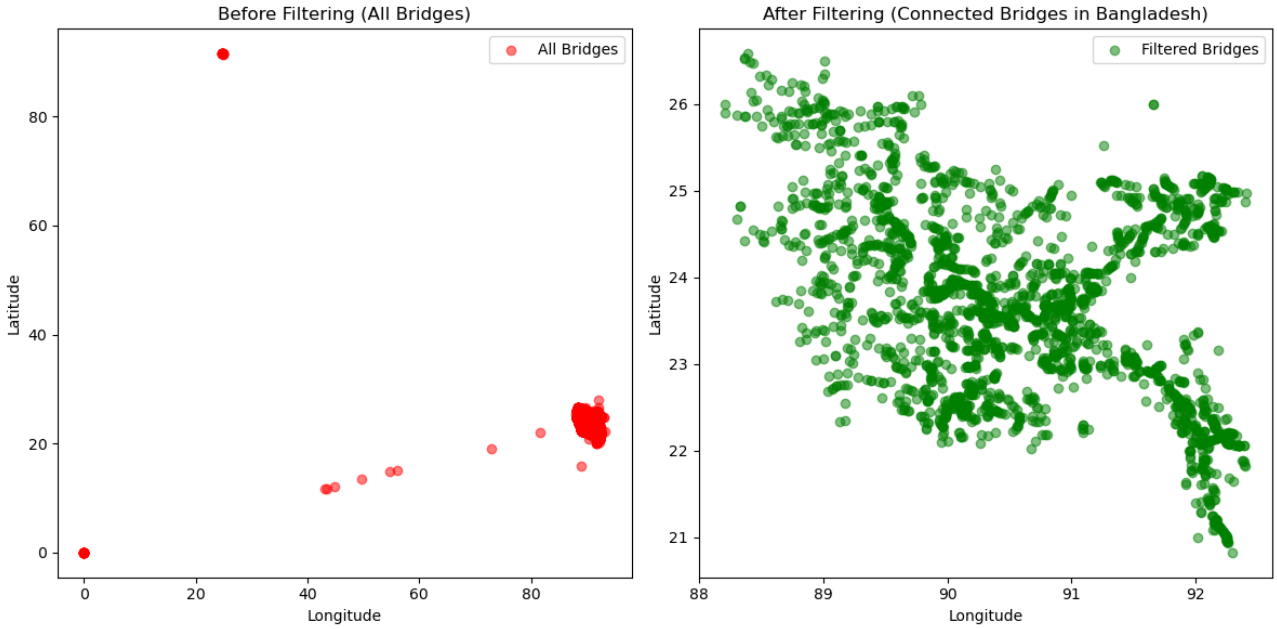


Figure 2: Bridge Locations Before and After Filtering

4.2 Data Quality Solutions for the Roads

In tackling the quality of the roads file, the most striking issue of the data was the non-linearity of LRPs, as seen from Figure 3. This plot shows that the chainage issues prioritized in section 2 significantly impact the length of a road, therefore negatively impacting ulterior criticality assessments (Jin et al., 2022), vulnerability studies (Ansari Esfeh et al., 2022), or financial estimations for maintenance or renovations.
To address the jumps in the roads map (Figure 3), we design three functions in our code that can be easily applied to the entire dataset and increase the validity and quality of the inputs for our model. These functions represent incremental steps in identifying and eliminating outliers from the LRP set of each road, ensuring the processed file is a continuous representation of reality. These steps help resolve the semantic accuracy of the roads file. The group of functions created estimates the average distance found between LRPs across a single road, then isolates the LRPs outside of an arbitrary range as outliers and modifies them.

The first function, named `find_distance`, is the first step in improving the chainage of the roads. It focuses on calculating the distance between two LRPs in a row of the roads data file, by extracting the columns containing the longitude and latitude coordinates of each road. With this information, an imaginary right-angled triangle can be formed using the coordinates of every 2 LRPs in the row, in which the hypotenuse of the triangle represents the distance between the LRPs. We then employ Pythagoras' Theorem to calculate this hypotenuse and store the evaluated distances for further processing.

The second function, denominated `find_outlier`, uses the distances array as input to estimate which coordinates result from incorrectly introduced LRP spatial data. It fulfills this purpose by identifying the distances outside the n^{th} percentile of distances, where n designates an arbitrarily decided value, and correlating the outlier distance with the relevant LRP's. The isolated outlying LRPs are then stored and can be later analyzed.

The final function, titled `interpolate_outliers`, applies corrective measures to the data deemed erroneous in the previous step. It modifies an outlier by averaging the coordinates two steps previous to and subsequent to

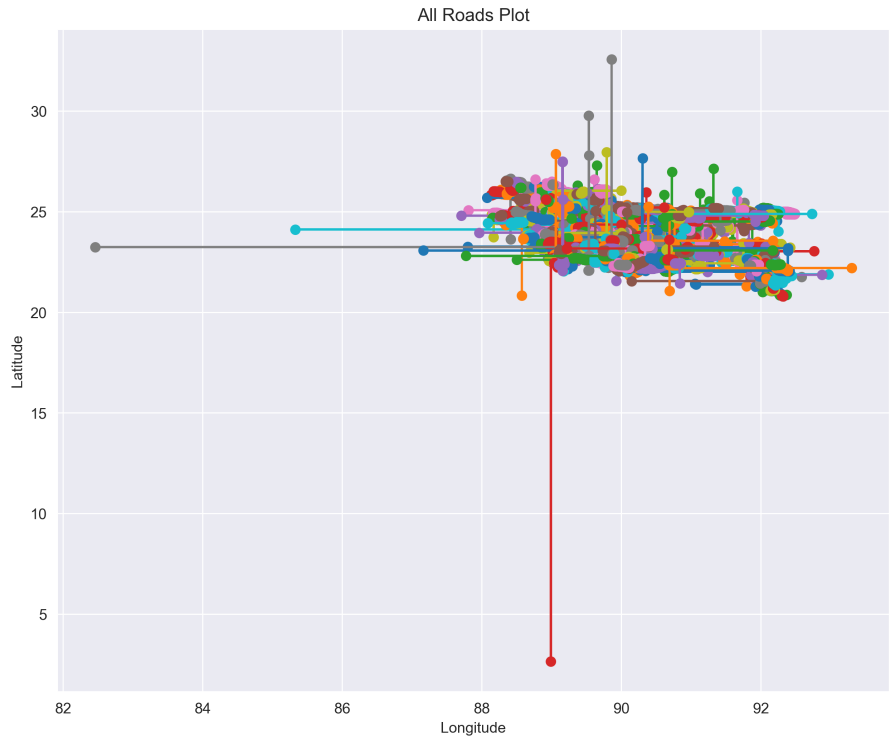


Figure 3: Overview of the raw data describing the Bangladeshi road network.

it, ensuring that even consecutive faulty LRPs are resolved. Once the changes are implemented to all outliers in the selected row, they are stored in a new array.

The created algorithm is tested and applied to Bangladesh's two main national roads, N1 and N2, as seen in Figure 4.

5 Discussion and Conclusions

The current assignment was a rewarding challenge for our team, allowing us to resolve our team dynamics and better define our roles for future tasks. The collaboration in the team was healthy, and no conflicts arose.

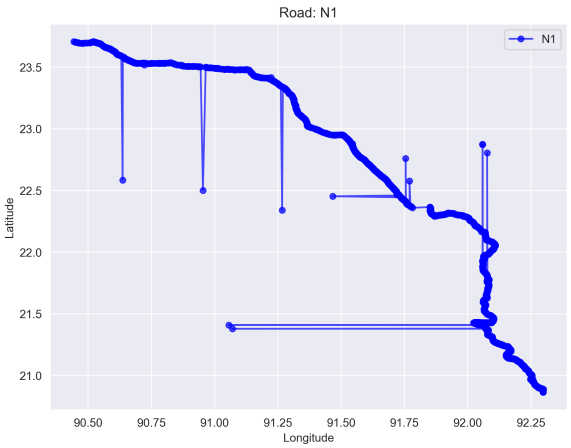
The theoretical exploration of the assignment was completed with guidance from lectures held within the course EPA133a, and it was a journey of further understanding the influencing factors and relevance of data quality. Thanks to the efforts of linking theoretical knowledge with our data set, we discovered the shortcomings and strengths of our files and gained confidence in the raw material of our future simulation.

The application aspects of this task were problematic, mainly due to code-sharing struggles. GitHub remains a challenge for our team to master in the future. However, the data clean-up solutions designed and coded by the team turned out, at least partially, successful.

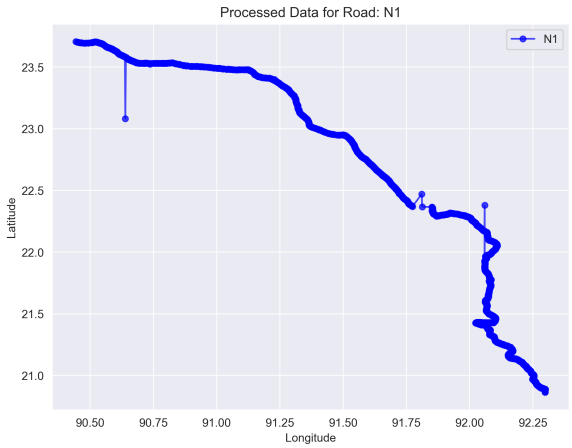
In the case of the bridges file, many issues were tackled in an isolated fashion. This process rendered a highly performing raw data set for bridges and was implemented significantly faster than that of roads. An improvement to the code could be implementing an automatization to identify outliers in the widths of bridges, therefore avoiding manually imputing minimum and maximum values of what can be accepted.

In the case of the roads data file, the approach to design inter-dependent functions resulted in a time-consuming procedure, where the team overcame challenges related to correlated indexing across arrays and mathematical approximation tools. However, the road processing results were only tested on two roads from the data file. Other improvements to the road clean-up process include creating a boundary box for the territory of Bangladesh to ensure chainage is in the territory of interest.

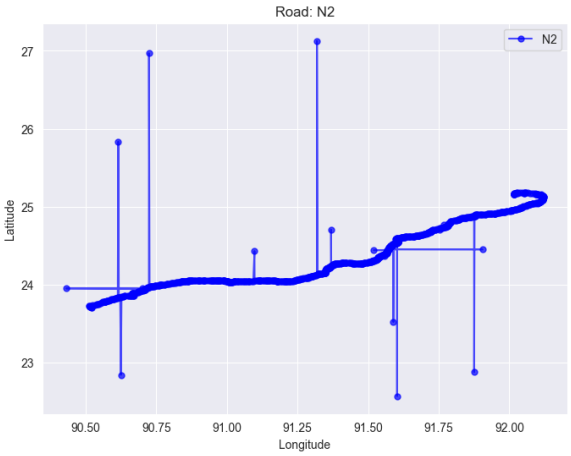
It is important to note that the solutions implemented for bridges and roads can be extended to the entire raw dataset and are not limited to individual instances. This is due to function-oriented coding, which ensures



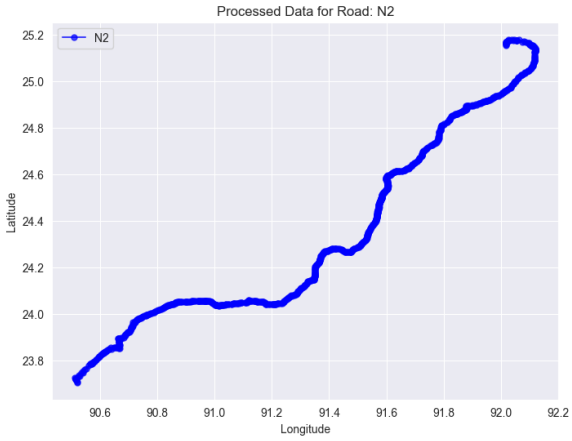
(a) National road 1 in Bangladesh before processing.



(b) National road 1 in Bangladesh after processing.



(c) National road 2 in Bangladesh before processing.



(d) National road 2 in Bangladesh after processing.

Figure 4: Visual reconstruction of a national road in Bangladesh before and after the designed algorithm was applied. The reconstruction was obtained using LRP information from the raw and processed data files.

flexibility in application. For future attempts at this task, we recommend a more extensive literature review of data quality issues before conceptualizing algorithmic solutions and performing more extensive testing and integration of found solutions. For example, attempting to use the newly processed data as input for the JAVA model.

6 Aknowledgment

6.1 Use of AI

AI tools (mainly ChatGPT) were used for assistance in writing the report, mainly grammar and spelling, coherence with the assignment guidelines, and translating the text to LaTex document. Another use of AI tools was summarization of all information of the case, assignment, and rubric for easy accessibility. Moreover, ChatGPT was used for guidance on setting up GitHub and Pycharm, as all group members were new to the two coding tools. Lastly, ChatGPT was also used for debugging during the coding process.

6.2 Contributions

The group divided the workload during the first two lab sessions, trying to give equal space to both coding and reporting efforts:

- **Bettin Lorenzo:** conceptual strategy for solving data quality issues (reporting).
- **Dell'Orto Alessandro:** conceptual analysis of data quality issues and prioritization of solutions (reporting)

- **Le Grand Tangui:** introduction, data exploration, reflection and conclusion (reporting/coding)
- **Precup Ada:** solving spatial issues of roads (coding) and the written description of the implementation (reporting)
- **van Engelen Ralph:** solving spatial issues of bridges (coding) and the written solutions (reporting)

References

Ansari Esfeh, M., Kattan, L., Lam, W. H., Salari, M., & Ansari Esfe, R. (2022). Road network vulnerability analysis considering the probability and consequence of disruptive events: A spatiotemporal incident impact approach. *Transportation Research Part C: Emerging Technologies*, *136*, 103549. https://doi.org/https://doi.org/10.1016/j.trc.2021.103549

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*(7604), 452–454. https://doi.org/10.1038/533452a

Dong, J. Q. (2019). Numerical data quality in simulation research: A reflection and epistemic implications [Perspectives on Numerical Data Quality in IS Research]. *Decision Support Systems*, *126*, 113134. https://doi.org/https://doi.org/10.1016/j.dss.2019.113134

Huang, Y. (2013). Data quality issues in automated simulation model generation. *Delft University of Technology*, 44–52.

HUANG, Y. (2013). *Automated simulation model generation* [Doctoral dissertation, Technische Universiteit Delft].

Jin, K., Wang, W., Li, X., Hua, X., Chen, S., & Qin, S. (2022). Identifying the critical road combination in urban roads network under multiple disruption scenarios. *Physica A: Statistical Mechanics and its Applications*, *607*, 128192. https://doi.org/https://doi.org/10.1016/j.physa.2022.128192

Marsden, J. R., & Pingry, D. E. (2018). Numerical data quality in is research and the implications for replication. *Decision Support Systems*, *115*, A1–A7. https://doi.org/https://doi.org/10.1016/j.dss.2018.10.007

Oreskes, N. (1998). Evaluation (not validation) of quantitative models. *Environmental Health Perspectives*, *106*(6), 1453–1460.

Price, R., & Shanks, G. (2005). A semiotic information quality framework: Development and comparative analysis. *Journal of Information Technology*, *20*(2), 88–102. https://doi.org/10.1057/palgrave.jit.2000038