

Advanced Simulation

Assignment 1- Report

AS Group 10: Angelica Saglimbeni (6556671), Elena Deckert (6580300), Finn Maingay (5415586), Tadas Lukavičius (5525047), Viola Clerici (6542034)



Table of Contents

Table of Contents	2
1. Introduction	3
2. Identification of Data Quality Issues	3
3. Prioritisation of Data Quality Issues	5
4. Conceptual Strategies for Data Cleaning	7
5. Implementation and Results	9
5.1 Roads correction algorithm	9
5.2 Bridges correction algorithm	9
5.3 Visualisation	10
6. Discussions	12
7. Acknowledgement and Contribution	13
References	14

1. Introduction

Bangladesh often faces climate-related disasters, such as floods and cyclones, which threaten economic activity, transportation and the livelihood of the population (Ministry of Environment and Forests, n.d.) To address these risks, the World Bank aims to assess Bangladesh's potential transport infrastructure investments by identifying which infrastructure sections are most vulnerable and critical.

A key challenge in this project is data quality. Marsden and Pingry argue that Information Systems (IS) research often overlooks numerical data quality, despite its importance in ensuring sound research and analysis (2018). Because this dataset could positively impact the development of Bangladesh and benefit its citizens, it is important that the research based on it is sound. This means that the data behind it must be of high quality. Therefore, in this assignment, we will focus on identifying and tackling types of data quality issues within the World Bank dataset for the Bangladesh transport network.

To do this, we follow a multi-layered approach. The work begins with data processing, working from the dataset provided. First, we identify potential issues by visually inspecting the network on the map and detecting structural irregularities and outliers. A key challenge in this phase is determining which data errors to address first. Section 3 presents a prioritised overview of the main data quality issues, covering roads, bridges, ferries, and traffic, identifying which issues are crucial to investigate. Section 4 then proposes possible cleaning techniques, while Section 5 provides a detailed explanation of how selected strategies were implemented and the solutions obtained.

2. Identification of Data Quality Issues

In this section, we present the main data quality issues identified through visual inspection of the map and data. These errors were noticeable without processing, as they clearly deviated from geographic patterns. The issues are summarized in Table 1 and categorized based on the object type, such as roads, bridges, and waterways. For each issue, the table reports both a qualitative description and the corresponding quality type based on the definition from (Huang, 2013)

We distinguish between three semiotic levels to define the corresponding data mistake categories:

- **Syntactic mistakes** arise from errors in data format or structure, such as swapped latitude and longitude values or invalid coordinate representations.
- **Semantic mistakes** occur when the data are syntactically correct but do not accurately represent real-world conditions.
- **Pragmatic mistakes** refer to cases where the data may be present and semantically correct, but are insufficient or unreliable for the intended use, such as missing or inaccurate quality attributes.

This classification helps in understanding what is wrong in the dataset, and also why the error matters and how it affects the usability of the data for modeling and analysis. Moreover, considering the type of error can help develop efficient solutions to fix the data issues.

Table 1: Types of data quality issues

	Issue	Type	Description
Road	Missing road	semantic completeness	The dataset is incomplete compared to reality. There can be missing LRPs or entire roads that are not documented.
	Wrong coordinates	syntactic /semantic accuracy	A "swap" (lat and lon) is a syntax error. An outlier located kilometers away from the actual road is an error in meaning (semantics).
	LRPs off-road	semantic accuracy	An LRP should be located directly on the corresponding road segment. If it is positioned beside the road, the recorded data does not accurately reflect the actual geographic alignment. Moving the LRPs back onto the road is a way of restoring semantic truth.
	Inconsistency in LRP data	Mapping consistency	When two different records refer to the same physical point, they must have identical coordinates. If they do not, the mapping is inconsistent.
	Traffic	Semantic completeness	Missing data in the traffic data set is considered a semantic completeness mistake. This could be solved by using neighboring values and size of roads. There could also be false documentation of traffic intensity.
Bridges	Inconsistency in bridges data	Mapping consistency	This represents a conflict between two different sources. The degree to which different data collections agree on the same object determines the consistency of your overall system.
	Completeness and accuracy values for quality	Pragmatic completeness	The data may be semantically present (the bridge is in the list), but we require the quality rating. Without this value, the data is not usable for the purpose. Filling these in based on neighboring bridges is a pragmatic solution.
	Road and bridge points are shifted together	Mapping consistency	A bridge is relationally dependent on a road. If the road is shifted but the bridge remains in its old position, the logical relationship between these two objects in the model is broken.
	Missing bridges	Semantic completeness	Missing bridge data can be identified in cases where a road crosses water but no bridge is documented.
	Bridge length	Semantic completeness	If the length is missing entirely, it is a matter of Semantic Completeness. If the length is present but insufficiently accurate for the calculations, it is a lack of Precision.
Waterways	Ferries	Semantic	The absence of a ferry connection even though the corresponding road includes "ferry" in its description.

3. Prioritisation of Data Quality Issues

To prioritize the data quality issues, the MoSCoW analysis technique has been used. This technique ranks issues in the categories Must have, Should have, Could have, and Will or will not have (Kravchenko et al., 2022). Must-haves are non-negotiable needs that are mandatory for further analysis. Should-haves are important initiatives that are not vital, but add significant value. Could-haves are nice to have initiatives that will have a small impact if left out. And lastly, 'will not haves' are initiatives that are not a priority for this specific time frame. To know what to focus on, we considered the requirements for the next assignments, which will focus on bridges and their quality, main roads and traffic, resulting in an economic analysis. Table 2 below gives an overview of all issues, put into the correct category of the MoSCoW frame, thereby showing our priority order.

Table 2: prioritisation of data quality issues

	Issue	Why
Must have	Wrong coordinates	Having wrong coordinates makes analysis futile because solutions don't fit reality. Without the correction of these coordinates, distance calculations and spatial mapping won't be possible for economic analysis. As Chojka (2020) mentions, syntactic interoperability is a necessary precondition for all other analyses. This being a syntactic error therefore makes it a must have to fix, because semantic accuracy and economic modelling cannot exist without a syntactically sound foundation.
	LRP's off-roads & Inconsistency in LRP data	Crucial for semantic accuracy, LRP's are the anchor points for all other data. An off-road or shifted LRP creates a rift from reality, where bridges cannot be correctly associated with the road they are on. As Newson et al. (2009) point out, map matching is important, but mapping every noisy point to the road can create very unlikely road paths.
	Inconsistencies in bridges data	Fixing the bridges' data resolves mapping consistency issues. When merging the RMSS and BMSS data, duplicates or conflicting locations must be resolved to make sure the bridge count is accurate. Otherwise, the simulation might consist of 'phantom' maintenance costs. As Xin et al. (2023) state, creating a unified bridge data archive is a needed prerequisite for sound evaluation.
Should have	Road and bridge points are shifted together	Bridges are dependent on roads. Ensuring that they move as 'one' prevents errors in the network and ensures that the graph shows real-world connections.
	Missing bridges	This is not considered a must-have because the simulation can run without every single bridge. However, missing major crossings, especially the ones over big rivers, would lead to an underestimation of infrastructure value and risks. Therefore, the model should include fixes for this issue.

		Missing bridges are often also new bridges, meaning that they are of good quality and do not need to be prioritised for policy interventions.
	Missing roads	Just like missing bridges, a semantic completeness issue. The model can work without every single road, but to get accurate route finding and total distance-based economic metrics, it should include as many as possible. Using interpolation between LRPs to fill gaps creates this continuous road network. If roads are completely missing they are most likely small and less relevant and therefore not crucial for our analysis.
Could have	Traffic	Linked to pragmatic quality, traffic data is not strictly necessary for geometric cleaning. However, traffic volume will be essential in the following assignments. It is, therefore, nice to have, but not vital for this phase
	Quality of bridges	Fixing the quality of bridges enhances pragmatic completeness. Having correct ratings for the quality makes for a more nuanced simulation of maintenance costs. But, if left out, the model will still be able to run using average assumptions; however, with less granular detail.
Will not have	Bridge length	Not needed, because in this macro-level assessment, the sole existence and quality of a bridge are way more important than the precise length. By excluding, the model gets less cluttered, and scope creep can be avoided.
	Ferries	Leaving ferries out is a pragmatic exclusion. The ferries involve information that is outside of the scope of road-based infrastructure. Leaving it out puts more focus on what really matters.

Thanks to the MoSCoW framework, we were able to prioritise all the issues in order to provide a solution for each of them in a structured and transparent way. In the following section, a conceptual solution for all the data cleaning issues is proposed.

4. Conceptual Strategies for Data Cleaning

In this section, we propose conceptual strategies on how to solve the previously identified quality issues. The proposed solutions are developed based on detailed data inspection and aim to resolve inconsistencies in a structured and transparent manner.

Table 3: Conceptual strategies for each data cleaning issue

Issue	Conceptual Strategies for Data Cleaning
Wrong coordinates, LRP's off-roads and inconsistencies	Wrong coordinates can be caused by latitude-longitude reversals, degree-minute formatting errors, or simple typographical mistakes. Because it is difficult to identify the exact cause of each error separately, we adopt a general algorithm. We detect suspicious LRPs by looking for unusually large jumps in distance between consecutive points and for points that strongly deviate from the general alignment of their neighbouring LRPs. Once such inconsistencies are found, we correct them by reconstructing their positions based on surrounding points. If the erroneous LRPs lie between two reliable points, their coordinates are interpolated; if they occur at the beginning or end of a road, their positions are extrapolated using the direction and spacing of nearby valid points. Our primary objective is to restore a coherent sequence and plausible spatial structure of each road segment for subsequent economic analysis, rather than to achieve perfectly accurate geographic coordinates.
Inconsistencies in bridges data	To correct the bridges, we align each bridge with its corresponding road and reference point and then verify its spatial plausibility. First, we match bridges to valid roads and, where possible, replace their coordinates with LRP coordinates to ensure consistency with the road reference system. Next, we check whether each bridge is located reasonably close to its assigned road geometry. Bridges that are slightly offset are projected onto the nearest point along the road to correct minor positional inaccuracies, while bridges that are implausibly far from their road are removed.
Road and bridge points are shifted together	When adjusting an LRP's coordinates as part of the road correction process, we inherently shift all bridges associated with that LRP to the updated location. In this way, the spatial consistency between roads and bridges is preserved, and no additional adjustment step is required for bridges that are anchored to corrected reference points.
Missing bridges	For missing bridges, we assume that the most common reason is recent construction that is not reflected in the dataset. In cases where a road clearly crosses a river but no bridge is recorded, we would classify the crossing as a Category A bridge and treat it as a newly constructed structure. This ensures that major crossings are not omitted from the analysis simply due to data gaps.
Missing roads	For missing road segments, we reconstruct the network wherever sufficient reference information is available. When gaps occur between

	known LRPs, we use interpolation to estimate the missing geometry, thereby restoring a continuous road alignment. If an entire road is missing from the available data, there is no reliable way to reconstruct its geometry. In such cases, we assume these are most likely minor roads with limited relevance to the overall analysis and acknowledge them as residual data limitations.
Traffic	For missing or incomplete traffic data, we could use values from neighboring road segments under the assumption that adjacent sections of the same corridor tend to have similar traffic patterns. Structural features such as the road's functional classification (main road versus secondary road), width, or overall size could also be used to inform estimates. Another factor that could be considered is the proximity to urban areas, since roads near cities typically carry more traffic than those in remote locations.
Quality of bridges	The bridge quality data is complete, so the main task would be to verify its accuracy. Where evaluation is needed, the condition or reliability of a bridge could be inferred from nearby bridges and features of the surrounding road network. Traffic intensity could also be taken into account, since higher volumes generally lead to greater wear and affect its lifetime.
Bridge length	The length of a bridge should correspond to the width of the feature it crosses (such as a river, road, railway, or valley), plus some additional margin for structural supports like abutments and safety space. For example, if a bridge crosses another road, its length should be comparable to the width of that road corridor. If it crosses water, it should reflect the mapped width of it at that location. Values that are significantly shorter than the crossed feature are likely to be incorrect. If the value is incorrect, it can be approximated by taking the width of the body being crossed plus some additional relative margin.
Ferries	For ferries, we could identify crossings by examining road segments on either side of a body of water. If the road names on both sides contain the word "ferry," we could infer the presence of a ferry connection and represent it with a dedicated ferry line (yellow).

Weighing the implementation complexity of potential corrections against their relevance for the subsequent analysis, we developed and implemented an algorithmic approach to address inconsistencies in both bridge and road data. Rather than attempting to correct each issue manually, we applied a set of procedures to detect, evaluate, and correct irregularities. The primary objective is to ensure an analytically reliable representation of the transport network while maintaining a pragmatic level of accuracy appropriate for the intended analysis.

5. Implementation and Results

5.1 Roads correction algorithm

This algorithm automatically detects and corrects spatial inconsistencies in Location Reference Points (LRPs) along road centerlines using geometric continuity constraints. LRP are sequentially analysed per road segment after standardising naming conventions across road and bridge datasets. For each road, the spatial distances between consecutive LRP are computed to estimate a representative median spacing that reflects the expected geometric continuity of the road alignment.

Two complementary error-detection mechanisms are then applied. First, LRP that introduce abnormal spacing relative to neighbouring points are identified as potential positional outliers based on a distance-jump threshold. Second, a local geometric consistency test is performed by computing the perpendicular deviation of each LRP from the line formed by its adjacent neighbours. Points that significantly deviate from the expected linear trajectory of the road are marked as spatial anomalies.

Flagged LRP are grouped into contiguous sequences and corrected using either linear interpolation (for internal outlier segments) or directional extrapolation (for boundary segments). This ensures that reconstructed LRP preserve both spacing regularity and directional continuity along the road geometry. The corrected coordinates are subsequently propagated to associated infrastructure datasets (e.g., bridges) using LRP identifiers, maintaining spatial alignment between linear road references and point-based assets.

This outlier detection method is applied iteratively because spatial errors in LRP affect not only the incorrect points themselves but also the local geometric context used to detect neighbouring inconsistencies. Since the deviation and spacing checks rely on adjacent LRP to define an expected road trajectory, the presence of corrupted points may initially mask secondary outliers by distorting local alignment and spacing estimates. This approach is inspired by trajectory reconstruction methods such as the Improved Kinematic Interpolation (IKI) framework proposed by Shaoqing Guo et al. (2021).

The general inspiration for this algorithm is taken from statistical outlier detection techniques, particularly those based on median distance measures, as discussed by Irad Ben-Gal (2005), as well as from trajectory urban traffic anomaly detection methods that incorporate road angle (Djenouri et al., 2019).

5.2 Bridges correction algorithm

To improve the positional accuracy of the bridges in the dataset, we implemented a three-step correction algorithm that integrates road matching with geometric validation and projection onto the road network. The objective was to harmonize bridge records with the road data, correct spatial inaccuracies, and remove clearly deviating entries.

In the first step, bridges were matched to the official road reference data using the road name and the associated LRP. The bridge dataset was merged with the road data containing

LRP information, together with its coordinates. Where a bridge matched a known LRP, its coordinates were replaced with the corresponding LRP coordinates. This ensured that bridges located at reference points inherited consistent position data and were placed on the matching position on the road accordingly.

The second step focused on ensuring that every bridge was associated with a valid road geometry. Bridges whose road names did not exist in the road dataset were excluded from further processing.

In the third step, we checked whether each bridge was plausibly located on its assigned road. For every bridge, we measured the distance to the nearest segment of the corresponding road. If a bridge was located more than 10 kilometres away from its road, it was considered implausible and removed from the dataset. If the bridge was reasonably close to the road but slightly offset (more than 25 meters away), its coordinates were adjusted by projecting it onto the nearest point along the road. Bridges that were already within 25 meters of the road were left unchanged. This approach allowed us to correct small positional inaccuracies while filtering out clearly incorrect entries.

Overall, the procedure ensures that bridges are either aligned with official reference coordinates, adjusted to match the road geometry more accurately, or removed if they fail consistency checks. The result is a cleaner and spatially coherent bridge dataset that is fully aligned with the underlying road network.

5.3 Visualisation

In Figures 1 and 4, we displayed the initial data of roads N1 and N2, as these are considered central roads in Bangladesh and are relevant for our further analysis. We plotted the roads and associated bridges based on their latitude and longitude coordinates. The visualisation reveals several incorrectly positioned LRPs, which create visible spikes and distortions in the road geometry. In addition, we plotted the bridges as red "X" markers along their corresponding roads. This makes it apparent that some bridges are misaligned with the road network.

In Figures 2 and 5, we displayed the results after applying the road correction algorithm. The yellow points indicate the LRPs that were identified as outliers and consequently interpolated according to the neighbours. You can also observe that misplaced bridges at the same position as misplaced LRPs were shifted onto the road together (Figure 5).

In Figures 3 and 6, you can observe the results of the bridge correction algorithm. Depending on their distance from the assigned road, bridges were either projected onto the corresponding road segment to correct minor misalignments or removed from the dataset if the deviation was considered too large to be plausible.

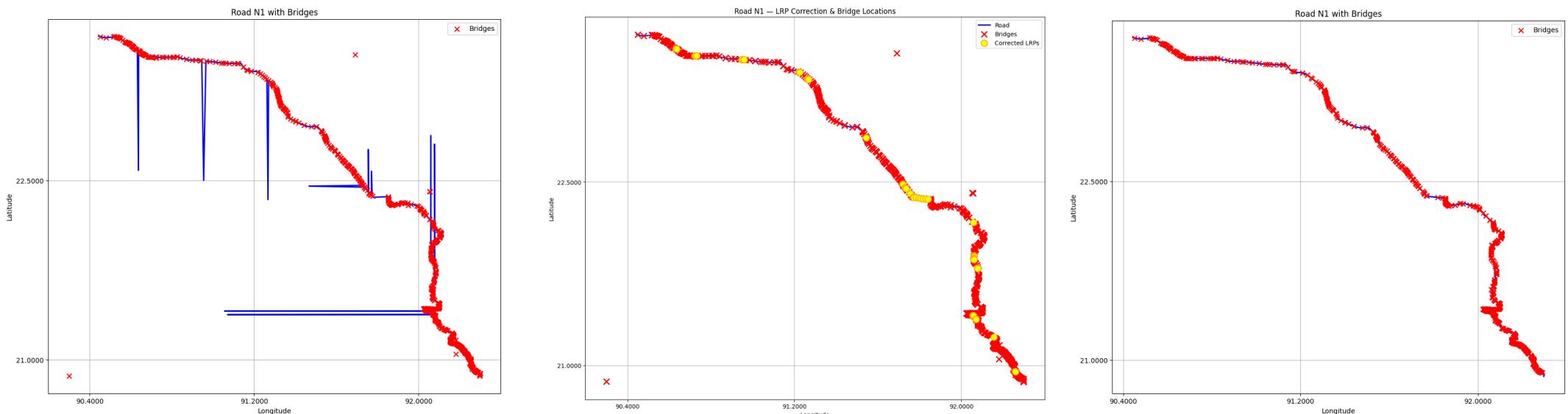


Figure 1. Road N1 before the algorithms.

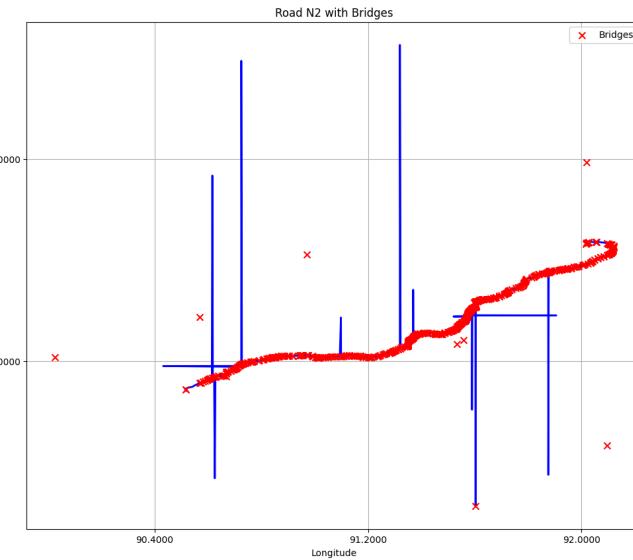


Figure 4. Road N2 before the algorithms.

Figure 2. Road N1 after applying road-fixing algorithm.

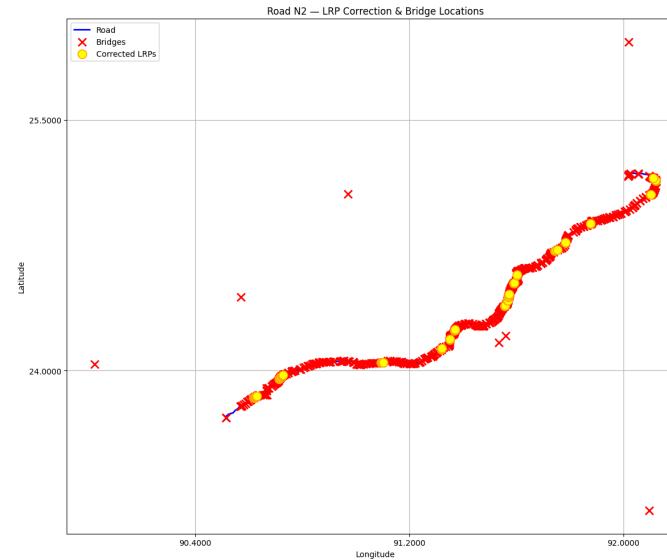


Figure 5. Road N2 after applying road algorithm.

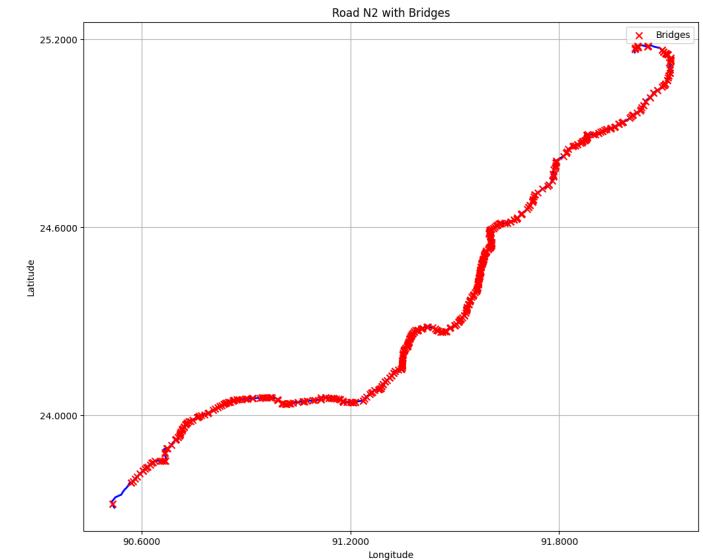


Figure 6. Road N2 after applying road and bridge algorithms.

6. Discussions

The results of this assignment are faced with certain limitations regarding data quality and algorithm implementation. Because of time constraints and the prioritisation scheme explained, some of the data quality issues were not tackled at all in this assignment. Some of the issues that were included from our analysis were aspects of infrastructure, e.g. bridges, that did not exist in the dataset. Locating and incorporating these missing features would have required substantial additional work with limited immediate benefit, so this issue is suggested for future analysis.

In addition to dataset issues, there are limitations in the implementation of data corrections. First, although most of the described issues were tackled, some errors might remain. For example, there are roads that still have gaps between them even after the implementation of the fixing algorithms. The effort required to resolve these smaller problems was considered disproportionate to the expected benefit, so this issue has also been left for possible future work. Another limitation of the implemented code is that we did not consider the source of the error, especially in syntactic errors like the switch of latitude and longitude. A more thorough investigation into the source of these errors could have further informed the implementation of the code. This lack of reflection could have guided the building of the algorithm differently. Finally, the code implementation focused roughly on approximating the simulated points, not achieving true values. This was considered sufficient given the economic nature of the planned intervention derived from the results, but it should be taken into consideration.

The implementation of a prioritisation scheme based on the necessity of the fix, the importance for next assignments and the ease of it, also raises ethical and methodological considerations. Based on the intended goal of the World Bank's database, it could be problematic to only focus on certain roads, as this could further amplify existing structural and economic inequalities of Bangladesh's population. While addressing this issue might fall out of the intent of this assignment, it remains an important concern, especially when dealing with data quality, and should be taken into account in future work.

7. Acknowledgement and Contribution

We acknowledge the use of AI in this assignment. Grammarly was used in the report for text-editing and proofreading, and ChatGPT was used to assist in coding the algorithms with editing and debugging. AI suggestions were considered critically and modified if needed.

All members took part in the conceptualisation of data quality issues, and editing of the report but the rest of the tasks were divided as follows:

Group member	Contribution
Angelica Saglimbeni	<ul style="list-style-type: none">• Identification of data quality categories of the issues• Report: Introduction, Discussion, Acknowledgements
Elena Deckert	<ul style="list-style-type: none">• Bridge correction algorithm• Report: Conceptual strategy, Implementation of results
Finn Maingay	<ul style="list-style-type: none">• Identification of data quality categories of the issues• Report: Identification and prioritisation of data quality issues
Tadas Lukavičius	<ul style="list-style-type: none">• Data preprocessing, Road correction algorithm, Visualisations• Report: Conceptual strategy, Implementation of results
Viola Clerici	<ul style="list-style-type: none">• Identification of data quality categories of the issues• Report: Identification and prioritisation of data quality issues

References

- Ben-Gal, I. (2005). Outlier detection. In O. Maimon & L. Rokach (Eds.), Data mining and knowledge discovery handbook (pp. 131–146). Springer.
https://doi.org/10.1007/0-387-25465-X_7
- Chojka, A. (2020). Semantic and syntactic interoperability issues in the context of SDI. *Geomatics and Environmental Engineering*, 14(3), 5–20.
<https://doi.org/10.7494/geom.2020.14.3.5>
- Djenouri, Y., Belhadi, A., Lin, J. C.-W., Djenouri, D., & Cano, A. (2019). A survey on urban traffic anomalies detection algorithms. *IEEE Access*, 7, 12192–12205.
<https://doi.org/10.1109/ACCESS.2019.2893124>
- Guo, S., Mou, J., Chen, L., & Chen, P. (2021). Improved kinematic interpolation for AIS trajectory reconstruction. *Ocean Engineering*, 234, 109256.
<https://doi.org/10.1016/j.oceaneng.2021.109256>
- Huang, Y. (2013). Automated Simulation Model generation. *Research Repository (Delft University of Technology)*.
<https://doi.org/10.4233/uuid:dab2b000-eba3-42ee-8eab-b4840f711e37>
- Kravchenko, T., Bogdanova, T., & Shevgunov, T. (2022). Ranking requirements using MOSCOW methodology in practice. In *Lecture notes in networks and systems* (pp. 188–199). https://doi.org/10.1007/978-3-031-09073-8_18
- Kuruman. (n.d.). *Rural road in Bangladesh* [Cover Photograph]. Retrieved from
<https://www.sasec.asia/index.php?page=news&nid=274&url=dhaka-sylhet-highway-upgrade>
- 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/10.1016/j.dss.2018.10.007>

Ministry of Environment and Forests. (n.d.). Climate change and infrastructure in Bangladesh: Information brief. Government of the People's Republic of Bangladesh.
<https://iucn.org/sites/default/files/import/downloads/infrastructure.pdf>

Newson, P., Krumm, J., Microsoft Research, & Microsoft Corporation. (2009). Hidden Markov map matching through noise and sparseness. In *Microsoft Research* [Journal-article].
<https://www.microsoft.com/en-us/research/wp-content/uploads/2016/12/map-matching-ACM-GIS-camera-ready.pdf>

Xin, G., Galant, F. L., Zhang, W., Xia, Y., & Zhang, G. (2023). A practical data extraction, cleaning, and integration method for structural condition assessment of highway bridges. *Infrastructures*, 8(12), 183. <https://doi.org/10.3390/infrastructures8120183>