Validation of Data for Use in Civil Infrastructure Big Data Applications

CONNOR O'HIGGINS*, CONNOR KENT, DAVID HESTER and SU TAYLOR

ABSTRACT

Big data applications are becoming more popular across many different fields and civil engineering is no different. The benefits of big data lie in its potential to provide valuable insights into various large datasets. Big data applications can identify patterns and trends that were previously unknown, which can help them make informed decisions and develop effective strategies. In the case of civil engineering, this could be taking large datasets that have been produced in relation to pieces of infrastructure and using them to create more efficient management strategies. One of the issues with using big data is that if the input dataset is flawed then the output resulting from any big data analysis will be compromised. Therefore, data validation is used, which is the process of ensuring that data is accurate, complete, and consistent. The consequence of not undertaking data validation may be Inaccurate or inconsistent data which can lead to incorrect insights and decisions in big data applications. This paper explores the necessity of validating data before it is used in a big data application. It outlines some of the different methods used for validating data and provides an overview of the potential issues that may arise from data validation errors. A case study is then presented showing the process of validation on data collected from four bridges and provides recommendations for implementing data validation as part of a larger big data workflow. The results of the case study show that validation of data is an important step in the big data process both for confidence in the outputs and to make big data applications more useful and more common in the civil engineering field. The discussion and presented case study in this paper highlight the necessity of validating data. It has shown some of the potential issues that may arise from not undertaking data validation.

INTRODUCTION

Big data applications have become increasingly popular in various fields, including civil engineering. [1] shows the increased popularity in big data applications in 2016 and this trend does not seem to have decreased in the recent years. There are a range of potential benefits of big data, providing valuable insights into large datasets that were previously unknown and predicting future observations. By identifying patterns and trends, big data applications can inform decision-making processes and develop effective and efficient strategies. However, the accuracy of the input data is crucial, in [2] it is stated that inaccurate or non-valid data cannot serve as a basis for extracting insight. Data validation is a process used to ensure that data is accurate, complete, and consistent. This paper explores the necessity of validating data before it is used in a big data application. Various methods for validating data are presented as well as an overview of the potential issues that may arise from data validation errors. The paper also presents a case study of data validation on data collected from four bridges and provides recommendations for implementing data validation as part of a larger big data workflow. The case study results demonstrate the importance of data validation in the big data process, both for confidence in the outputs and for making big data applications more useful in the civil engineering field.

BACKGROUND ON BIG DATA APPLICATIONS IN CIVIL ENGINEERING

Civil engineering deals with a massive amount of data be that in the design, construction, and maintenance of infrastructure such as roads, bridges, buildings, or other infrastructure such as water systems. Between these infrastructures a vast amount of data is produced, including sensor data, traffic data, weather data, and maintenance records. In recent years, big data analytics has emerged as a powerful tool in civil engineering for analysing and making sense of this vast amount of data. This is evidenced by the various studies that make use of big data applications such as; [3] in which data from a population of structures is used to make insights into induvial structures; and [4]where a data-driven method for detecting damage to structures is presented. A review of some of the opportunities presented by big data applications and some current research is presented in [5]. By applying big data analytics to these datasets, civil engineers can gain valuable insights into infrastructure performance, identify patterns and trends, and develop effective management strategies. This, in turn, can lead to more efficient maintenance and repair, improved safety, and cost savings.

BENEFITS OF BIG DATA IN PROVIDING INSIGHTS AND PREDICTIONS INTO LARGE DATASETS

Utilising big data analytics can provide numerous advantages when it comes to obtaining insights or making predictions from large datasets. Below are a few of the benefits and examples of how that may be applied to engineering.

Identifying patterns and trends: Big data analytics is capable of identifying patterns and trends within large datasets that may not be easily noticeable through traditional analysis methods. For instance, civil engineers can analyse traffic data over an extended period to identify traffic patterns like peak hours and congestion hotspots.

Predictive analytics: Using big data applications, it's possible to create predictive models that can forecast future events or trends based on historical data. An example of this is when engineers analyse maintenance records and sensor data from a bridge to predict when it will require repairs or maintenance.

Improved decision-making: An additional benefit to identifying patterns and trends is that it can also aid the decision-making process. For instance, engineers can make informed choices about which materials to use in upcoming projects by examining data on the properties of various materials utilised in previous construction projects.

DATA VALIDATION IN BIG DATA APPLICATIONS

Data validation is the process of ensuring that data is accurate, complete, and consistent. Generally, data validation aims to ensure that the data is credible and error-free so that any outputs resulting from using the data are accurate and can be reliably used. [6] presents an overview of data validation across different industries and some of the overarching concepts. The reasons why data validation is important vary from field to field but there are some reasons that are common to all types of data and applications. Accuracy is the most obvious and possibly the most important. While the accuracy of data is vital in any data analysis in big data even small errors in input data can cause significant errors in the results. This problem is worsened because it is common for big data applications to use black box methods or at least methods that are difficult to understand how outputs are processed from the inputs. Thus, making it harder to detect errors when reviewing

results. Other reasons to undertake data validation is to ensure the completeness and consistency of the dataset. The completeness of a dataset refers to if there is any missing data. Missing data may be significant and may mean that certain insights may be overlooked or the predictive power of the model may be reduced. Consistency of data comes into focus when using multiple datasets as part of your inputs. Consistency refers to aspects like the units of your data; the format of your dates; the capitalisation of letters. All these properties if not consistent will most likely invalidate any comparisons that are made between the datasets. The last reason for data validation, which is sometimes overlooked, is trust in the applications/method. This can be a large factor if the big data application is used for something such as decision-making. Errors in the input data that result in poor decision can quickly erode stakeholder's trust and make it far less likely to be used in the future.

METHODS FOR DATA VALIDATION

This section will look at some of the methods that can be used for data validation. While there is no definitive set of methods that can be used in all situations, the broader categories of data validation will be discussed here and as well as some common methods that can be used.

SYNTAX VALIDATION

Syntax validation is the process of checking data to ensure that it follows the correct format. Generally, there will be a set of predefined rules or schema that the data will be compared to ensuring it is in the correct format. Syntax validation will ensure that there is consistency in the dataset. This type of validation is very important if a large number of datasets are going to be used in the analysis. Examples of syntax validation would include checking the format of a date or checking that there are only numbers contained within a certain measurement field.

SEMANTIC VALIDATION

Semantic validation focuses more on the actual content and context of the data. This type of validation checks for such issues as logical errors or internal inconsistencies. Again, like the syntax validation a predefined set of rules can be used to check the data against but this time the checks are to ensure that the data is meaningful for the intended use. Examples of semantic validation would be checking if an air temperature reading falls within a given range for the location and time of year. An air temperature reading of 100°C would pass a syntax validation check but may not pass a semantic validation check.

FIELD-LEVEL VALIDATIONS

Field-level validation describes the method of checking data on a field-by-field basis. This type of validation is used to check common errors such as missing values and incorrect syntax. These errors are not dependent on any other aspects of the data and so can be checked on a field-by-field basis.

CROSS-FIELD VALIDATION

The cross-field validation method involves checking the relationships between fields within a dataset to ensure that they are behaving as expected. This type of validation can discover errors in the data that may not be apparent when using a field-level validation. Examples of cross-field validation would be ensuring that the time values are in a particular order or that the temperatures that are recorded during the night and less than those recorded during the day. The use of cross-field validation can give confidence and credibility to the dataset before use in a big data application.

STATISTICAL VALIDATIONS

Statistical validation encompasses a wide range of methods but in general, it is the process of using statistical methods to find any issues/problems with a dataset. One of the most common methods that fall under this category is outlier detection, here statistics from the data will be compared to individual observations and any that deviate too far from a central value would be considered an outlier. Other forms of statistical validation include comparisons of data distributions this method gives a quick way to determine if a dataset conforms to an expected/predicted distribution and allows for a range of issues to be detected such as repeated values and rounding errors.

CASE STUDY: VALIDATION OF DATA FROM FOUR BRIDGES

This section presents a case study in which data was collected for 21 days across four bridges. The presented case study was part of a larger piece of work that involves ongoing data collecting, this case study gives a template for how to validate the data that has been collected as well as how to validate future data.

COLLECTION OF DATA

The bridges used in this case study are a mixture of different span lengths, span numbers and construction types. The four bridges chosen have span lengths ranging from 8.9 m – 98 m with span numbers ranging from 1 – 3. The construction type/material of these bridges also varied as the sample includes a steel bridge, reinforced concrete bridges and a steel-concrete composite bridge.

The SHM system used to collect the data consisted of one MEMS accelerometer and one environmental. The accelerometer used was a Multifunction Extended Life (MEL) accelerometer. This accelerometer measures acceleration in 3 axes within a range of ± 2 g and has a real-time clock to timestamp every acceleration measurement. The sensor data is stored locally on an SD card at a sample rate of 128 Hz and is powered from an internal battery.

The environmental sensors used were the ‘OM-EL-USB Series’ from Omega Engineering. The environmental sensors measure both air temperature and humidity and, like the accelerometers, store the data locally with a corresponding timestamp.

VALIDATION OF DATA

In this section, the steps that were undertaken to validate the acceleration and environmental data will be presented. As discussed in the previous section, there

is no all-purpose validation procedure and the steps presented in this section are tailored to the collected dataset. The use of data is an important factor in the data validation process. For this case study, it is assumed that the data will be used to predict future data for each of the bridges and make comparisons between the bridges. Because of this data consistency becomes a factor between each of the datasets.

ACCELERATION DATA

Figure 1 shows a sample of the acceleration data that is representative of the data that was collected across all four bridges across the monitoring period. Plotting a sample of the data in this way allows for basic checks on the data such as if the acceleration data looks credible and the data being centered around 9.8 m/s^2 due to the effects of gravity.

Step 1: Check for missing data and syntax of data.

The first and most basic check was for missing data and to ensure the correct syntax of the data. Here a simple script was written to undertake a field-level validation check. The script first checked for any NaN or null values and then checked that the acceleration, reading consisted of only floating numbers and that the date format was consistently in the format of DD/MM/YY HH:MM:SS:sss.

Step 2: Verify the sampling rate of the data.

The sampling rate of each of the four monitoring systems was set at 128 Hz. It is important for this to be consistent as processing the data to extract features such as the natural frequency will be altered by varying sample rates. A field-level validation check was undertaken on the timestamps of the data to ensure the difference between consecutive steps was 7.8125 e-3 seconds.

Step 3: Check for and clipping of the acceleration data.

The accelerometers being used have a measurement range of ± 2 g. In this step, the values of the acceleration data are checked to ensure that the magnitude of any values does not exceed 2 g. While during normal operation the accelerations would never come close to their measurement limit if a value of 2 g was recorded it may indicate that the accelerometer was struck by something or that the sensor is not functioning as expected.

Step 4: Statistical validation of the acceleration data.

There are some basic statistical methods that are valid for all most all continuous datasets. TABLE I shows the mean and interquartile ranges of the four bridge datasets. These basic statistical properties allow domain knowledge to be used to check if the data behaves as expected. In the presented dataset the accelerometer should be measuring the gravity constant so the acceleration experienced by the bridge should be centred around 9.81 m/s^2.

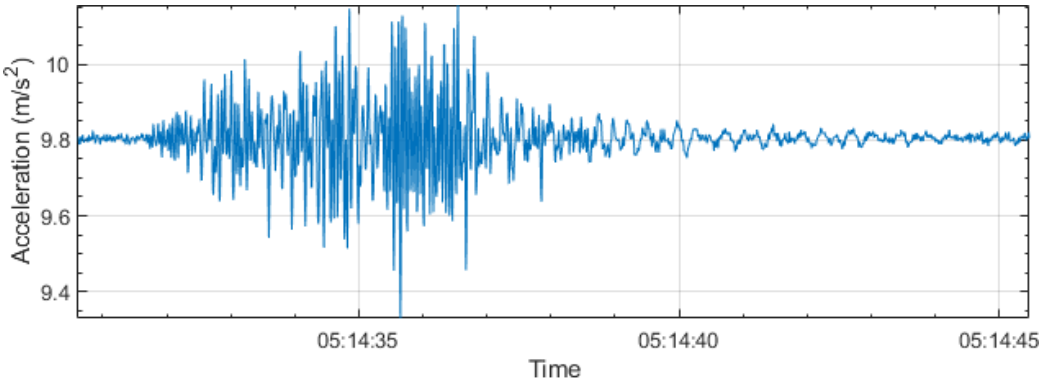

Figure 1. Sample of the collected acceleration data

TABLE I. Basic statical properties of the four bridge datasets

Bridge Ref.	*Mean (g)*	*Interquartile range (g)*
Bridge 1	9.8049	0.0132
Bridge 2	9.7557	0.0119
Bridge 3	9.8068	0.0119
Bridge 4	9.8418	0.0117

From TABLE I it can be seen that the actual mean values range from 9.7557 m/s^2 to 9.8418 m/s^2, this is caused by discrepancies in the calibration of the sensors. For the analysis of the data, the data can either be detrened so that the mean is 0 for all data sets or the calibration factor changed so that the data mean is the same for each dataset. The interquartile ranges of the datasets vary between 0.117 m/s^2 and 0.132 m/s^2 from experience this is the range that would be expected as the bridges vary in both their construction and span length.

The distributions of the acceleration can also be studied. Figure 2 shows the distribution of the accelerations across the monitoring period for bridge 1. From this figure, we can see that the distribution is approximately normal, centred around 9.81 g and showing no significant skew. These properties are what is expected and give credibility to the datasets.

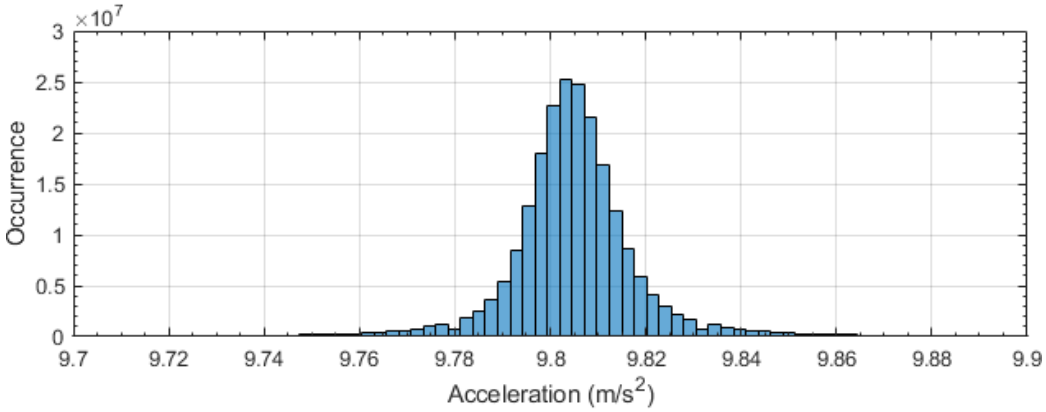

Figure 2. Histogram showing the distribution of acceleration data for bridge 1

ENVIRONMENTAL DATA

The environmental data consists of air temperature and air humidity readings. For the validation of this data steps 1 and 2 described in the last section are largely the same. During the validation of the timestamps, an issue was found regarding the adjusted daylight saving time which was not present in the acceleration data. If left unchecked this would have caused the temperature data and acceleration data to be out of sync by 1 hour.

For the statistical validation the same process was followed, first, check to ensure that the range of temperature readings was realistic for the time of year. The distribution of the temperature data again showed the expected normal distribution. The distribution for the humidity readings was skewed towards the 100% range. After some research into this, it was determined that that was not uncommon for the location of the bridges.

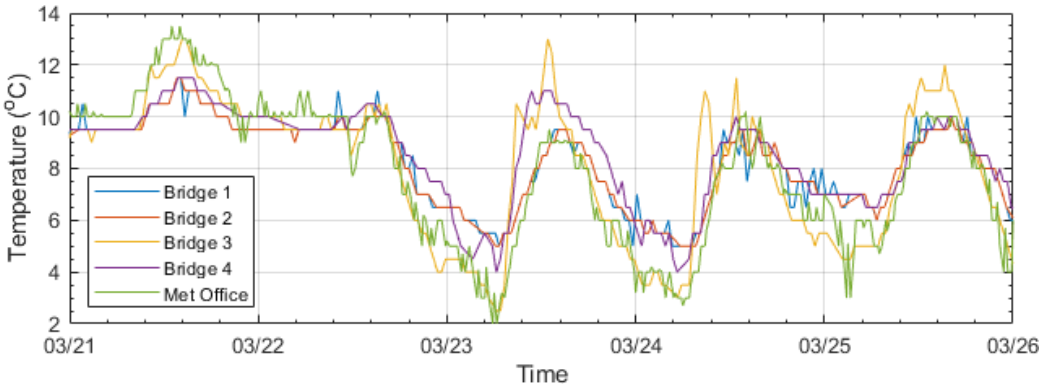

Figure 3. Temperature data over 7 days from each of the monitored bridges as well as met office data for the same period

As a final check for the temperature data to ensure that it was credible, it was compared to data from a nearby weather station. Figure 3 shows seven days of temperature data with the blue, orange, purple and green lines representing bridges 1, 2, 3 and 4, respectively. Temperature data between each of the bridges is fairly consistent, albeit with slight differences depending on the location of the bridge. The corresponding air temperature from the weather station is also plotted (green line). This Met Office weather station ranges from 11 to 14.5 miles from the four monitored bridges. The broad temperature trends at all 4 of the bridges match well with the met office air temperature. There are some small differences, but this is likely due to the met office temperature being taken in a weather station whereas the temperature at the bridges is taken close to a structure which could account for the slight differences.

The case study presented here was a small proportion of a much larger dataset. The steps described give a basis for a workflow to be developed for the validation of this dataset. This workflow can in turn be used as a template to develop automatic processes that can be applied to the larger dataset and to any newly collected data.

RECOMMENDATIONS FOR IMPLEMENTING DATA VALIDATION IN A BIG DATA WORKFLOW

As stated throughout this paper there is no one process or method that can be used on all datasets. The correct data validation methodology will be based on the data and the intended use of that data. However, there are some strategies that can be applied to all data validation problems. Some of these strategies are described below:

Define data quality standards: The big data application that will be used to process the data will inform what data quality standards will be needed. The quality standards would include such criterial as: the required accuracy to the data, the required completeness of the data and which parts of the data do and do not require validation.

Define a validation methodology: It is always advisable to create a plan for the validation of data. This both allows for the consistency on how the data is validated and gives a method to allow new data to be validated in the same way. The validation methodology could include such information as: what methods are going to be used, any schema of predefined lists that the data will be compared to, if automation will be used.

Monitor the data: Depending on the size of the dataset and the amount of new data being added to the dataset you may be required to automate the data validation process. If this is the case having a process in place to sample and test your data to ensure that the validation is processing the data as expected/planned may be beneficial.

CONCLUSION

The use of big data applications is becoming more popular in all field and civil engineering is no different. The benefits of these applications are clear and there is a significant potential for improving decision-making process regarding infrastructure and gaining insights into datasets that may otherwise be missed. However, the fact still remains that the output from any analysis, including that of big data is only as good as the inputs. Data validation of a dataset ensures that the input data meets the requirements of the application. The specific requirement will change from project to project but every application will have an underlying assumption of quality data. In this paper, the broad categories of validation have been discussed and examples are given of how they may be applied to different datasets. The case study presented shows how a validation methodology may be applied to data and the reason why certain techniques may be used.

Data validation has always been an important step in the analysis of data, however, this process becomes critical in big data applications. When the amount of data grows to the point where manual checking of inputs is not feasible or the analysis methods used are black box techniques, then data validation adds repeatability, credibility and accuracy to the outputs of the application.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the support of the UK Engineering and Physical Sciences Research Council (EPSRC) through the ROSEHIPS project (Grant EP/W005816/1).

REFERENCES

1. Kapliński, O.; Košeleva, N.; Ropaitė, G. BIG DATA IN CIVIL ENGINEERING: A STATE-OF-THE-ART SURVEY. *Engineering Structures and Technologies* **2016**, *8*, 165–175, doi:10.3846/2029882X.2016.1257373.
2. Gahi, Y.; Guennoun, M.; Mouftah, H.T. Big Data Analytics: Security and Privacy Challenges. In Proceedings of the 2016 IEEE Symposium on Computers and Communication (ISCC); IEEE, June 2016; pp. 952–957.
3. Bull, L.A.; Gardner, P.A.; Gosliga, J.; Rogers, T.J.; Dervilis, N.; Cross, E.J.; Papatheou, E.; Maguire, A.E.; Campos, C.; Worden, K. Foundations of Population-Based SHM, Part I: Homogeneous Populations and Forms. *Mech Syst Signal Process* **2021**, *148*, 107141, doi:10.1016/j.ymssp.2020.107141.
4. Gui, G.; Pan, H.; Lin, Z.; Li, Y.; Yuan, Z. Data-Driven Support Vector Machine with Optimization Techniques for Structural Health Monitoring and Damage Detection. *KSCE Journal of Civil Engineering* **2017**, *21*, 523–534, doi:10.1007/s12205-017-1518-5.
5. Bilal, M.; Oyedele, L.O.; Qadir, J.; Munir, K.; Ajayi, S.O.; Akinade, O.O.; Owolabi, H.A.; Alaka, H.A.; Pasha, M. Big Data in the Construction Industry: A Review of Present Status, Opportunities, and Future Trends. *Advanced Engineering Informatics* **2016**, *30*, 500–521, doi:10.1016/j.aei.2016.07.001.
6. Gao, J.; Xie, C.; Tao, C. Big Data Validation and Quality Assurance -- Issuses, Challenges, and Needs. In Proceedings of the 2016 IEEE Symposium on Service-Oriented System Engineering (SOSE); IEEE, March 2016; pp. 433–441.