Data Readiness for BADA

BADA main study 1, FFI/Vinnova grant 2015-00677

SICS Technical Report T2017:08

Björn Bjurling Per Krueger Ian Marsh
RISE SICS AB

December 18, 2017

1 Introduction

One main motivation with the BADA program and project has been to gain understanding of requirements on data quality for big data analysis. This report aims at sharing experiences gained from dealing with automotive and traffic data in the use cases in the BADA project. We present our experiences in terms of the concept of *data readiness levels* as introduced by Niel Lawrence, Sheffield, UK [4].

In analogy with the concept of technical readiness levels [6], data readiness levels [4], were introduced with the intention to provide a tool for achieving better planning and resource allocation in data analysis projects, by characterization of various levels of data availability, quality, and suitability. As an example, a major part of the work in a data analysis project is performed in preprocessing of data. It is further not unusual that a considerable amount of time is spent waiting and asking for more relevant data than initially may have been made available. A formalization and appreciation of the data analysis process could potentially benefit the planning of data analysis projects. Lawrence's contribution is outlined in Section 3. In the area of Information logistics, one can find several related contributions dealing with the importance of getting the right data to the right data scientist at the right time [1, 5, 7].

Lawrence argues that his contribution should be seen as a first step towards formalizing the full process of data analysis. He suggests a basic order of readiness levels, and invites contributions that can be used to refine characterization of the levels of readiness. The experience from working with data in BADA can be well described in terms of Lawrence's initial levels of readiness. Further, we have encountered issues that seemingly are not covered by Lawrence's initial account. With further research, these issues may qualify for extending Lawrence's work on data readiness levels. We will make further comments on this in Section 3.

The report is structured as follows. In Section 2, we will review an analysis work flow used in the preprocessing of data in preparation for applications of analytic methods. We will also highlight some aspects of modelling process which are relevant for the concept of data readiness. Lawrence's concept of data readiness is described in Section 3. Section 4 contains an account of the data readiness levels in the use cases we have worked with in BADA and also in a project in the BADA program. These accounts should illustrate the usefulness of the proposed concept as well as some potential pitfalls in data analysis projects.

2 Data analysis work process

Data analysis and machine learning methods relies on a wide variety of methods and techniques, but share a need for usually large amounts of well structured and well understood data. Many, perhaps most, analysis projects spend a majority of their resources on collection, selection, correction, comparison, and processing of the data. This is the case regardless of the type of problem considered, whether it involves detection, diagnosis or autonomous control.

For this process to be as efficient as possible, there is a need for methods for assessing the *availability*, *quality* and *suitability* of the data, as well as a systematic workflow for processing the data and prepare it for use by analytic techniques. The three items of accessibility, quality, and availability, are also the main focus in Lawrence's acount for data readiness [4].

1. Availability
2. Data quality (e.g. correctness and completeness)
3. Suitability for a given purpose.

For each aspect, Lawrence proposes introducing a rating, which would make it easier to plan and allocate resources to a project with a given set of goals. We shall reflect these levels also in this review of the work process.

Data sources are almost universally organized as tables, normally with rows representing several aspects of a single observation, e.g. the time and place of an event or reading. For the purpose of data analysis, we refer to the observations as *data points*, or *entries*. Table columns represent several aspects or properties of the observations, and we refer to these as parameters. An observation is thus represented as one data entry with values given for some or each of the parameters. One set or sequence of data is referred to as a *data source*.

In order to prepare data for the use of analytic methods, or automatisation, we generally need to select, process and compare entries and parameters from several sources, which is an error prone and often exploratory process. In principle, this work should precede the development of models that are used by the method, but in practice it is often interleaved with at least preliminary modeling activities.

2.1 Data processing pipeline

Selection, correction, processing and collation of data sources typically occurs in repeated cycles where each step is based on preliminary assessment or analysis of earlier data versions derived from the same sources. Preparation of the data is therefore interleaved with successive interpretation and preliminary modeling. The preprocessing is incremental in the sense that a preliminary analysis step often results in a refined specification of how to interpret, correct or rearrange the data in later steps.

Challenges

In many applications, data occur in the form of logs of e.g. sensor readings, operations performed, events or transactions. Often, the data is distributed over several data bases, and entries are often only loosely coupled, e.g. by the co-occurrence of similar identifiers, numbers or text strings, or by overlapping time series with different resolution and data given in different units or formats.

Errors in data & data on errors

Values of parameters in the sources are often missing, or erroneous. Sometimes errors in the observation processes are encoded as particular values outside the normal range of the parameters. In other cases, the fact that a certain value is missing or invalid, can be valuable for the analysis, as an indication of e.g. an error in the underlying process. E.g., interpreting a missing time stamp as an indication that a process step has not been performed, or has failed. For sensor data there are also systematic errors and drift which can, when known and understood, be used to reconstruct the correct parameter values. To reduce lead times, such errors and omissions should be identified as early as possible

Deviations and anomalies

In the general case, detecting anomalous data (e.g. outliers, artefacts) is a difficult analytic problem in it's own right, but is often needed during data preprocessing and can include analysis step that are as complex as the ones used for modelling and interpretation. Failing to take this into account, can easily lead to misleading results.

2.2 Data preprocessing steps and dependencies

Figure 1 shows a selection of common preprocessing steps, their interdependencies, and possible iterations in the process of bringing the data into a shape suitable for application of analytic methods [3]. The following paragraphs elaborates and exemplifies the most crucial.

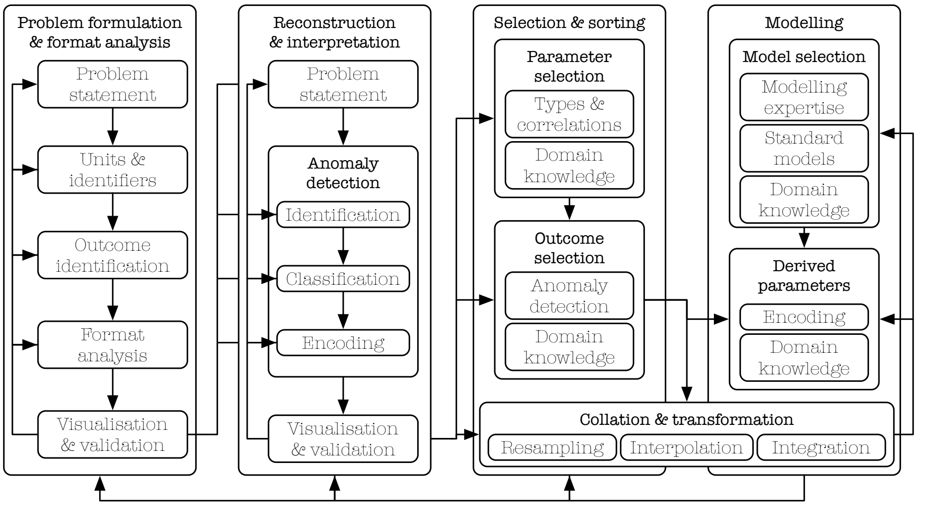


Figure 1: Preprocessing steps and dependencies

Problem formulation, unit and format analysis. The object of the analysis, e.g. anomaly detection, diagnosis, classification, or automation, clearly constrains the choice of data sources for the project[1].

It is rarely feasible to collect and prepare "all" data without at least a general idea of how it will be used. The first step in any analysis project is therefore to formulate the object of the analysis as concretely as possible. Its exact formulation is normally revised and refined as examination of the available data

[1] If the analysis is to be performed on-line, the preparation of the the data, the final version of the data preprocessing must be done continuously with further constrains the choice of sources and preprocessing complexity.

progress. Early in the process this can influence which data sources should be included in the analysis, and how to interpret their parameters.

This step in the preprocessing pipeline can include versions of at least the following steps:

1. Specification of problem formulation and selection of available data sources for the given problem.
2. Unification of units and identifiers in separate data sources
3. Determination of what constitutes an outcome relevant to the analysis object, in each data source
4. Format analysis
   (a) Determination of parameter value range (for e.g. detection of outliers and unexpected encodings)
   (b) Determination of the data type of the parameters in each data source
   (c) Rudimentary interpretation and re-encoding of free text parameters
5. Visualisation, inspection and validation of reformulated data

Reconstruction and interpretation. At least in cases where values for the parameters in a data source are generated by sensors, or humans, the data entries are often incomplete and/or inconsistent. Detecting and correcting or discounting the contribution of such entries from the analysis of each parameter is crucial to obtain reliable results. This can involve examining one or more parameters in related data entries in the same, or in other data sources, and/or reinterpretation and recoding.

1. Reconstruction of missing and obviously erroneous parameter values, where possible
2. Identification, classification and (re-)encoding of error states in the generation process that gave rise to missing and/or anomalous data
3. Visualisation, inspection and validation of reconstructed data

Selection and sorting While not all data analytic methods are sensitive to large number of parameters, many are, and selecting the ones most relevant to the object of the analysis is generally necessary and requires preliminary analysis of the correlation between parameters within and between different data sources. Some parameters may also have to be skipped if an insufficient proportion of the entries have valid values. Orthogonally, we may have to skip entries for which valid values could not be reconstructed. Finally data from several sources may have to be compared or merged into a format suitable for the modelling and interpretation of the analysis objective. This sometimes involves resampling, parameter fitting, derivation, trend analysis etc. and often involves making crucial modeling choices.

1. Selection of parameters generally depend on

   (a) the type analysis,
   (b) domain knowledge
   (c) correlation analysis between candidate parameters and against of the object model variables
   (d) identification and sub-selection of parameters representing the same or very similar information

2. Selection and sorting of entries
   (a) Filtering or discounting of entries for which values of the selected parameters are missing or invalid
   (b) Filtering or discounting of entries with uncorrectable or anomalous values

3. Sorting and merging data sources by e.g. summing, averaging, resampling, gradient detection, etc.

Modelling Depending on modelling choices related to the analysis objective, further analysis may be facilitated by the introduction of derived parameters or further specifications:

1. Deriveed parameters such as sums, differences, derivatives, or rescalings

2. Specification through development and evaluation of model for analysis objective, e.g. detection, diagnosis, prediction, or automation.

The result of this activity frequently requires that we return to an earlier data preparation step.

3 Data readiness levels

In [4] the state of the data before and during this process, is analysed as a method to improve the planing of data analysis projects. The author classifies project phases broadly in three phases, or “bands” depending on the knowledge and understanding of the available data, and it’s usefulness for a given objective. Each phase generally presupposes answers to questions to answers provided in the earlier phases, but as noted above preprocessing of the data is often an iterative process, which influences how to classify the complete set of data sources deemed necessary to achieve the analysis objective.

The first phase, “Band C”, involves assessing the availability and accessibility parameters and entries in existing data sources. Lawrence does not explicitly pose necessary and sufficient conditions for a given rating in any of the phases, but appears to propose rates running between 1 and 4 within each of them, and where rate “1” means that the project is ready to proceed with the next phase in the preparation and processing of the data.

Ideally there should be a finite number of validation criteria for each data source for the project to be able to rate it at a given level but that is not yet the case for the proposal as stated in [4]. As it stands, it proposes broad criteria for only a subset of the phases, in summary:

Band C (availability)

- **C4** Hearsay data
- **C1** Verified that data exists, can be accessed and compared, and no privacy or confidentially issues are show stoppers.

Band B (correctness and coherence)

- **B1** Missing values identified and handled, privacy issues resolved, collection process well understood, coherent use of units, data source merges verified correct, limitations and uncorrectable defects identified, preliminary modelling helps to identify problems

Band A (suitability w.r.t. a specific set or class of queries/tasks)

- **A1** Verified suitable and complete for a given task, may e.g. require manual annotation/labeling to qualify, to verify A1 status model suitable to the task must be more or less determined

There is a definite overlap between analysis of the preprocessing process as described in section 2.2 and the proposal in [4]. While [4] consider some aspects, e.g. privacy and legality concerns, it ignores the iterative character of some of the preprocessing steps, and is incomplete in terms of classification criteria. On the other hand, for managing data analysis projects, the proposed ratings, should be very useful, if only explicit and general criteria for each rating can be determined.

In the work with BADA, we have dealt with three further aspects of data readiness levels. These are seemingly are not covered by the present version of Lawrence's concept. The first issue (which is very common and which we have encountered in other projects too) is that of timing of delivery of data. Delay in delivery may obviously lead to less time for data preprocessing and less time for analysis. However, a more serious concern with delays in delivery is that the most experienced data scientist may have to leave the project for more pressing tasks. This is related to the time dimension in Information logistics, see for example [2, 5].

A second issue, related to the first one, is that the analysis may require skills that is not known in advance due to initial uncertainty abouthe quality and information content of the data. Thus, an experienced researcher may wait for data and subsequently pre-process it just to find out that another skill set would be more useful for the analysis task. Both the first and the second issues may lead to inefficiency in the project.

The third aspect concerns the role of the data scientists receiving the data to be analyzed. While academic research often require clean and reliable data, industrially motivated data analysis research has always been more tolerant to noisy and imperfect data. The success of a data analysis project lies not only in data quality. Rather, researchers with the right experience should be able to draw useful conclusions by selecting analysis methods according to the quality of the data. It could thus be relevant for the assessment of data readiness to take into account the level preparedness for dealing with imperfect data at the research organization. Further, the *suitability* of the data is relative to both the problem and the chosen analysis methods. Thus, for data analysis to succeed, it is required that the data scientist can select the models and algorithm for which he or she can make the most use of the suitability of the data. (With the wrong methods, data analysis can fail no matter the availability, quality, or suitability of the data). In BADA, we have on occasions had to reconsider the selecetion of analysis methods as quality and information contents of data have revealed.

4 Data Readiness Levels in BADA

In this section we first describe our method for assessing the data readiness levels in BADA. Then we give an account for our assessments of the data readiness levels in the two use cases Traffic Safety and Hazard Warnings. We include also an assessment of the data readiness level in a project from the BADA program.

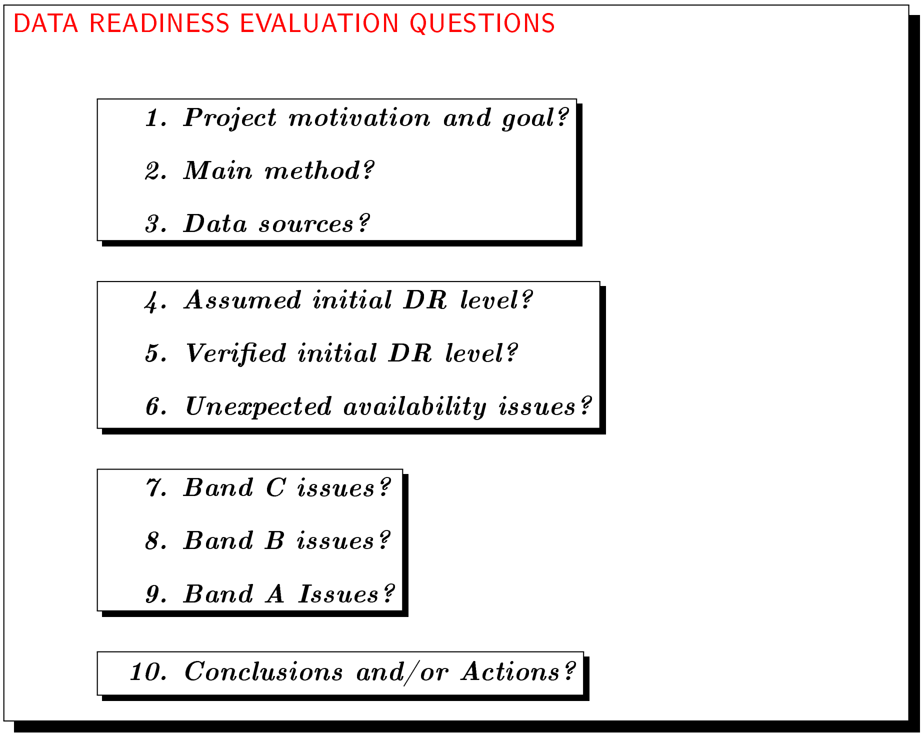
4.1 Method for assesseing data readiness

In the BADA project, we started making data readiness assessments explicitly towards the end of the project, after most of the activities described below had ended. Ad hoc assessments may have little value for a project at hand. Rather, the benefit with such ad hoc assessments lies in that encountered data readiness issues can be evaluated with respect to the impact those issues had on the project results. For example, in Anobada (see below), the high quality of the initially given data may have led the project initially to downgrade the significance of availability of ground truth data. It was not until the end of the project the non-availability of ground truth was realised as an issue.

Data readiness can be assessed before, during, and after a project. It is clear that making such assessments before a project may be beneficial by allowing better resource alloation, and setting realistic expectations of pace and outcome of the project. However, the true (or verified) data readiness level seem to be accessible only after the data has been thoroughly inspected (i.e. experimented with, wrangled, cleaned, transformed, etc.) It is only after such inspection we can know the quality of the data and potentially start to be able to evaluate its suitability for the problem. Making assessments during the project is likely a better compromise. Explicit assessment of the readiness levels of given data during the course of a project should allow for early mitigation of detected needs for more data or alternative data sources.

Our method for assessment of data readiness levels is based on posing a short list of questions to the data scientists involved in the activity at hand. The list of questions is given in the box in Figure 2. The internal boxes serve to capture the project motivation, problem formulation, data readiness expectations, detected issues, respectively conclusions based on the findings. The first question serves to ensure that the problem is well-formulated and that the expectation on the data analysis is well-concieved. The second and third questions seek to establish that the chosen methods go well with both the data sources and the expectations on the analysis. Further, the third question will also serve to set well-concieved expectations on the data quality.

Questions 4-6 caputure our assumptions about the data and contrast that to the verified state of the data. It is very common that the data readiness levels are over-estimated. For example, in one of the use cases in BADA (not included below and intended to be about air quality), it was assumed that useful



DATA READINESS EVALUATION QUESTIONS

1. Project motivation and goal?

2. Main method?

3. Data sources?

4. Assumed initial DR level?

5. Verified initial DR level?

6. Unexpected availability issues?

7. Band C issues?

8. Band B issues?

9. Band A Issues?

10. Conclusions and/or Actions?

Figure 2: Data readiness level questions

data was available and could have been classified as C1. It turned out that the data never had been stored. This was thus a clear case of hearsay data, and accordingly the verified data readiness level for the air quality use data was in fact C4.

Answers to question 7-9 should should give a comprehensive account for the status of the data readiness. Finally, based on those answers, Question 10 should lead to either a final asssessment of the data readiness or to actions for improving the data readiness levels in the project.

Note that projects with several data sources very well can have one data readines assessment for each data source.

4.2 Data readiness level assessments in BADA

4.2.1 BADA use case: Traffic Safety

1. The motivation and goal was to combining incident data (STRADA incident database) with other data sources in order to better explain root causes for traffic accidents. The intention is to enable the use of such knowledge in Real-time warnings application.

2. Main methods

   Correlation analysis of incidents with external factors such as geographic location, time of day, road type, road condition, and local weather available from public databases

3. Data sources

   STRADA: Incident database owned by Transportstyrelsen

   SMHI: Publicly available weather data incl. temperature, pressure, precipitation at time and position of incident

4. Assumed initial DR level?

   STRADA: C1

   SMHI: C1

5. Verified initial DR level?

   STRADA: C1, but privacy issues for data with high temporal resolution data,

   SMHI: B4: Low spatial resolution of weather data. Meteorological expertise required for interpolating readings to incident position

6. Unexpected availability issues

   The contact person at Transportstyrelsen went on parental leave and Transportstyrelsen did not appoint a replacement. Therefore, a request for additional data was delayed several months (in addition to the delay mentioned in item (7) below).

7. Band C issues?

   STRADA: Delay of 4-6 month for delivery of initial STRADA extract, and low temporal resolution in the first iteration

8. Band B issues?

   STRADA: "VIN no." (vehicle type) was considered very sensitive for business reasons and were never made available for all vehicles. Low temporal resolution in initial STRADA extract due to privacy issues. Higher temporal resolution was negotiated by eliminating non-essential sensitive fields in the extract for a subset of vehicles. Part of incident reports entered in natural language with an incomplete encoding of many parameters.

   Information crucial for understanding the cause and severity of the incident represented with images in pdf format.

   SMHI: Problems with the low spatial resolution of weather data so far unresolved.

9. Band A Issues?

   None. Data quality issues are being compensated for by careful selection of models and algorithms.

10. Conclusions

    Work with the use case the project is still ongoing. The low spatial resolution of weather data, the incomplete extract from STRADA, and the difficulty in interpreting bitmapped data in the STRADA data base, are factors that may limit the relevance of the outcome of the analysis. However, improving availability and the resolution of the data sources will likely improve the usefulness of the analysis.

4.2.2 BADA use case: Hazard warning

1. The motivation of this use case to *disambiguate* what an activation of hazard warning could mean. Simply put, a driver pushing a hazard in the city could mean a breakdown, parking, queue accumulation or waiting for a pickup. In a rural setting, this is much more likely to mean some form of help is needed. The motivation is to separate out these scenarios using *external* data sources, in this case the position of the vehicle.

   Name: Hazard warning

   Motivation: Essentially this is a technology showcase. As a showcase we should show how technology can be used to help in social settings. A hazard warning signal (both rear indicators blinking) is context, cultural and location-sensitive. Drivers could press the warning, often a large red triangle on the dashboard, to indicate i) parking or waiting to park ("move on, it's my place" ii) I am at the end of queue, take care and slow down iii) I have broken down iv) Even celebrations such as wedding are associated with the vehicles sending out audio and visual signals. Other uses of the hazard warnings are surely possible in other cultures.

   Goal: The goal is to clarify the use of hazard warning based on the **location of press**. By classifing and enriching the hazard warning events, correlated with with a position, helps disambiguate the reasons behind a press. As a first step separating urban or rural will give a central office some indication of what to do. Naturally, the urban presses could be disambiguated with further information, for example, from the vehicle itself (to separate breakdowns or parking). As an external information source we used a Geo-Information System, in this case, OpenStreetMaps. severe.

2. The main method we make use of is data correlation between 1 stream (presses in XML format) and 1 database lookup in OpenStreetMaps. The issue is really a press as a GPS location might not match more than 1 road, or an intersection, or a crossing, or the GPS location might not be present in OpenStreetMaps, as it does not cover all GPS pairs. Therefore, an expanding radius approach is needed, where the location is broadened to find the closest location, where a vehicle might be stationary. Recall too, that OpenStreetMaps is often a user contributed.

   In this case, we actually generated GPS data. This is not as uncommon as one imagines, as data processing pipelines need to be built, often without data sources. Occasionally privacy issues are encountered, however systems can be built and then used in company's internal systems. Open source software tends to follow this pattern, code is developed in the public domain, and then used & modified internally. We used the GPS locations to simulate that hazard warning button was pressed. Then used OpenStreetMaps to check if this location is either rural/urban and used that to classify the event into hazard/parking.

3. Data sources

   Openstreetmaps: open source map system

Hazard warning stream : (format: AMQ message)

4. Assumed readiness levels

   Openstreetmaps: A

   Hazard warning: C1

   GPS C1

5. Verified readiness levels

   Openstreetmaps: A

   Hazard warning: C1

   GPS C4

6. Unexpected availability issues:

   None

7. Band C issues. GPS data was not available and had to be simulated.

8. Band B issues. Geographical resolution of hazard warnings were different depending on car manufacturers (geolocation to likely location). Mapping to the road to the problem. Not enough metadata available (e.g. driving direction).

9. Band A issues. No evaluation of suitability done. Lack of GPS data for vehicles had negative effect on applicability of analysis.

10. Conclusions

    Data was too sparse and of too low resolution for making the analysis. However, by simulating data we were able to fulfil the main goal of the use case, which was to illustrate through implementation how a big data infrastructure would be set up for dealing with massive automotive data streams.

4.2.3 BADA program: Anobada

Anobada was a one-year project funded by FFI within the BADA program during 2016.

1. The goal of Anobada was to enable detection of deviation in vehicles operations based on operational data. The project aimed at finding interesting anomalies in existing data sets. The intended application of the results is to improve monitoring of vehicle fleets.

2. The main methods used in Anobada come from statistical data analysis. Through data analysis, the project built a number of statistical models based on the given data (including Principal Component Analysis, Guassian Mixture Models, and Markov Fields). Anomalies where defined in terms of likelihood of data points with respect to such cluster models.

3. Data came from vehicle sensors and were collected at maintenance cycles. Data was delivered as a csv file.

4. Assumed initial DR level was C1. The data was known to exist and non-disclosure agreements were signed early in the process. We expected some noise in the data as well as missing data. Unknown relevance of the data with respect to finding interesting anomalies.

5. Verified DR level was C1, alternatively B4. It turned out that data had been collected irregularly and with varying volumes from time to time. This made the analysis task harder, prompting rethinking of the choice of analysis methods. Further, the data was not too noisy and it was easy to clean and correct it. Industrial applications accept some noise in the data, therefore the B4 classification.

6. Unexpectedly the project received an additional set of data to aid in the classification of the anomalies. The data arrived towards the end of the project time, and was therefore not analysed thoroughly enough to be of use in the project.

7. The main issue in Band C was that ground truth was missing. The importance of ground truth was initially underestimated.

8. No issues in Band B. Despite some noise, irregularities, and missing data, the data analysis was performed by experienced researchers who were able to compensate and deal with unclean data.

9. In band A there were issues. The analysis helped in finding interesting statistical anomalies. However, in order to make use of the data, the anomalies needed to be categorized in terms of vehicle operations to make industrial sense. Put in other words, the project did not have access to ground truth data. The project relied on having access to experts in the the domain of vehicles for interpreting the anomalies. This turned out to be hard to access, so therefore the relevance of the found anomalies are still to be assessed.

10. Conclusions. The project was successful in analysing the initially given data, which also had a sufficiently high data readiness level. However, the missing ground truth made it hard to make practical use/sense of the otherwise successful analysis.

5 Conclusions

We have assessed the data readiness levels in BADA with respect to the classification suggested by Lawrence [4]. In that work, we have also identified some factors that potentially can be used for refining Lawrence's levels.

The goal for assessments of data readiness levels, whether they are made before, during, or after a project, should be to gain experience to make early and precise decisions in future data analysis projects. By making such assessments explicitly in writing, we believe that such experiences more easily can be shared and dissiminated with others.

References

[1] Wolfgang Deiters, Thorsten Loffeler, and Stefan Pfennigschmidt. The information logistics approach toward a user de'mand-driven information supply. *KLUWER INTERNATIONAL SERIES IN ENGINEERING AND COMPUTER SCIENCE*, pages 37–48, 2003.

[2] Andrejs Gaidukovs and Marite Kirikova. The time dimension in information logistics. In *ILOG@ BIR*, pages 35–43, 2013.

[3] Daniel Gillblad. *On practical machine learning and data analysis*. PhD thesis, 2008.

[4] Neil D. Lawrence. Data readiness levels. arXiv: 1705.02245.

[5] Magnus Lundqvist, Eva Holmquist, Kurt Sandkuhl, Ulf Seigerroth, and Jan Strandesjö. Information demand context modelling for improved information flow: Experiences and practices. In *IFIP Working Conference on The Practice of Enterprise Modeling*, pages 8–22. Springer, 2009.

[6] John C. Mankins. Technology Readiness Levels: A White Paper. Technical report, NASA, Office of Space Access and Technology, Advanced Concepts Office., 1995.

[7] Kurt Sandkuhl. Information logistics in networked organizations: selected concepts and applications. In *International Conference on Enterprise Information Systems*, pages 43–54. Springer, 2007.