Data driven case base construction for prediction of success of marine operations

Bjørn Magnus Mathisen, Agnar Aamodt, Helge Langseth

Norwegian University of Science and Technology
bjornmm@ntnu.no, agnar@ntnu.no, helge.langseth@ntnu.no

Abstract. It is a common situation to have lots of recorded data that you want to use for improving a process in your organization or make use of this data to provide new services or products. Starting with one primary data set we describe a system that enhances this data set to a level such that it can be used by a deep learning system. This deep learning system then creates a model based on this data set, trying to predict operational windows for marine operations. Using this model the system extracts cases for use in a CBR-system aimed at providing operational support. This paper describes the partial implementation and results of this system.

Keywords: Data Science, Deep Neural Networks, Data Analytics, Case-based Reasoning

1 Introduction

Critical operations are often meticulously planned and subject to many parameters that decide if and how these operations are performed. Some of these parameters are called operational time windows, which in marine environments often are connected to external factors such as weather.

This paper uses machine learning to predict favorable operational time windows or warn of unfavorable operational windows, so that critical operations can be planned with better accuracy, e.g. when the operation should ideally take place. One way of doing this is to look at historical data of previously executed operations. By combining data on successful and unsuccessful operations with the relevant context of that operation, we create a data set that can be used to find indicators for success or failure in advance. Which context that is relevant is dependent on the nature of operational window; wind and fog are important contexts for aviation, while waves and current are important for marine operations but not aviation.

This paper focuses on marine operations, and we analyze event data captured from boats moving in and out of zones connected to aquaculture installations. Next, we calculate the duration of these events and connect them to the relevant context and the associated success or failure classification.

The data used in this analysis is gathered as part of the EXPOSED project[1]. This project aims to develop enabling- and applied technologies for exposed

[1] http://exposedaquaculture.no/en/

aquaculture operations. The work we describe aims to improve planning of operations on aquaculture installations on exposed locations.

The data is a subset of boats moving across geofences attached to aquaculture installations. This system consists of two zones around every aquaculture installation in Norway: One outer zone 400 meters from the outer points of the structures holding the fish themselves (not including the control building/fishfeed silos). The inner zone is 20 meters from the structure. These limits are in adherence to government regulations that no boat should fish within the outer zone and no boat should move within the inner zone unless the boat is there to operate on the installation.

An example of geofencing zones are shown in Fig. 1 below.

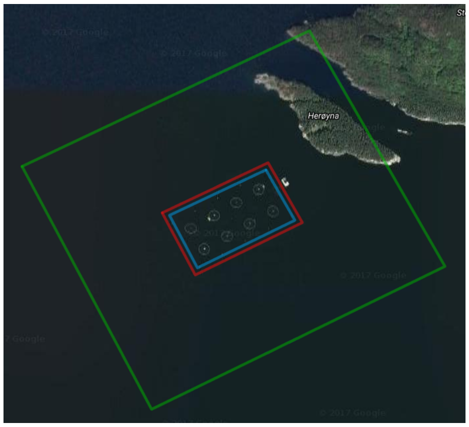


Fig. 1: The Green line show the outer geofence zone, the red line shows the inner geofence zone.

An event is created each time a boat crosses any of the geofence zones, marking the time. Table 1 below shows an example of a typical event.

Event ID	Location-ID	Vessel Name	Time	LocationZone	EventType
81766	12966	Vessel A	2014-09-02 21:39:32	1	1
81767	12966	Vessel A	2014-09-02 21:40:11	1	2

Table 1: This table shows an example of two events of a vessel entering (EventType=1) and leaving (EventType=2) the outer zone (LocationZone=1) of location 12966.

In data gathered in the EXPOSED project, the aquaculture industry reports on several possible problems with fish feed carriers interacting with aquaculture installations: Approaching the feed barges, often placed in shallow waters; Knowing which barge container to fill with what feed; Planning according to weather and route to enable the installation crew to attend the operation; And the fact that impact and currents from the boat can damage the installation.

As our data only gives us the time spent in two different proximities to the aquaculture installation there will be limits to which types of operational problems we can detect, and it will be very hard to discern between different causes (other than bad weather which is very general) of any detected problem.

The architecture of the full decision support system for EXPOSED is illustrated in Fig. 2. In this paper we only present results from parts of the system. Future work will integrate these results with the other modules (e.g. knowledge models) to complete the system to a state where it can be verified in the field.

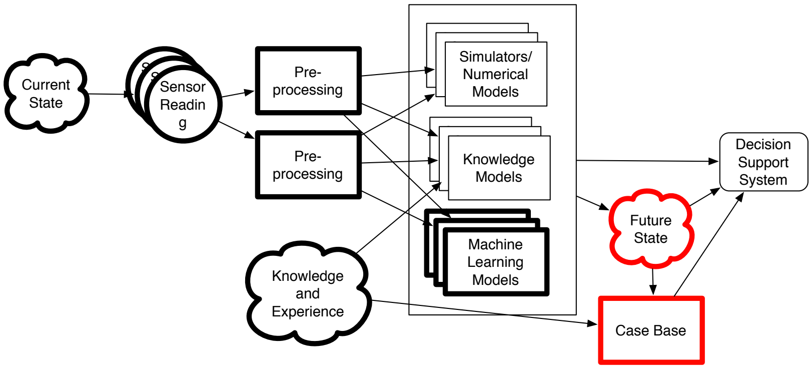


Fig. 2: The architecture of the planned systems. The parts implemented are highlighted, the case base and the future state is highlighted in red as being the current target for development.

Our main hypothesis is that given enough contextual weather data a deep neural network should be able to predict the length of a maritime operation at a aquaculture installation, enabling us to predict favorable operational windows. The main contribution of this paper is to show the reader the process of gathering, collating, filtering of data and subjecting this data to an analysis.

This paper is structured as follows; Section 2 introduces related work and our work in the light of this previous work. Section 3 describes the methods used in our work as well as the data sources used. Section 4 shows the result of our experiments, while section 5 presents the conclusion along with a discussion of the results.

2 Related work

In this work we aim to extract cases from a time series of events, CBR research has been done on several aspects of automatic case-authoring.

In CBR there has been a lot of focus on how to measure competence and utility of a case-base [1,2]. In [3], they do this via reversing deletion policies constructed in [4] that try to improve case base utility without degrading competence.

Several works [5,6,7] use NLP to extract cases from structured and unstructured ([8,9,10]) text.

More specifically connected to the task of extracting cases from time series is the work done by Bach et.al. [11] where they employ clustering of time-series events in time and space, in combination with other detection methods. Funk et. al [12] uses different models of how predictive (or discriminatory) different time-series patterns are to different medical diagnosis of stress. For more insight into work done in time-series analysis connected to CBR research we suggest chapter 3.3 in [13]

The work presented in this paper shares the approach of Bach et al. [11] in that we try to extract the useful data points from the time series via clustering and filtering. Our work differs from the previous work in that we have very few verified cases apriori or during learning. In other words, the time-series is in all practical sense unlabeled for our use. We will try to apply common knowledge about how long an operation usually takes to perform. Then we can extract failed operations from the even time series to create cases that exemplify failed operations.

3 Method

To enable the deep learning system to correctly model and predict the time spent at an installation, we need to provide it with as much context data as possible for each of the event data points. In addition, we need the data to be as noise free as possible, thus we want to filter away operations that naturally have a high degree of variation in time spent at the location. We address these two requirements by combining the primary data set with other data sets, to enable us to provide filtering and context. An illustration of this process can be seen in Fig. 3 . Below we describe each of the data sets.

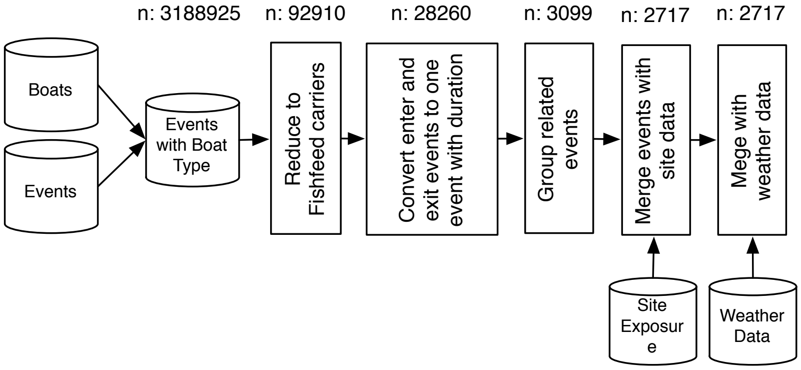


Fig. 3: This figure illustrates how the different data sources are combined and filtered to provide the deep learning system as much context as possible.

Boat data set As mentioned in the introduction we do not want to analyze all the traffic data of all of the boats. To verify that our method is usable in at least one instance, we want to look at a specific type of boat that has stable characteristics when it comes to the parameters (e.g. time and stability of time) of the operations it executes on the installation. We chose fishfeed boats in this case, as they only do one type of operation. That way we do not need to deduce the type of operation from the event data (one less hidden variable). In addition, this operation should be stable in the time it takes to execute it. To filter the data accordingly we need to combine the event data set with a data source that describes the boats. We can then easily extract the fishfeed boats.

NORA10 data set NORA10 [14,15] is a data set that describes output of a precise weather model (hind-cast), that is validated by measurements. It has a higher resolution (10km) than most other models (e.g. the much used ERA[2] model with 80km resolution) as it is re-sampled for this specific region around Norway. We sample this model for each of the installations and at each time of each event (in the case of long events we use the median time of the event). We sample every datatype that we think will have an impact on the time spent on an operation: wind speed, wave direction, wind direction, significant swell wave height and significant wave height.

Exposure data set SINTEF EXPOSED has produced a data set [16] that describes the degree of exposure for a large number of the installations that are used in the event data set. This data set provides a level of exposure for 360 degrees around the installation (from 0 to max, where max is no land in sight). We combine our weather data with this (described above), thus we combine the wind direction of the wind with how exposed the location is in the direction of the wind using a filter that combines exposure level from +/- 10 degrees around the direction of the wind.

3.1 Extracting time spent in zones.

The data set needs to contain the time spent in the zones around the aquaculture installations. The raw data only contains events of entering and exiting the zones. To extract this we sequentially find each exit from a zone then search backwards for the entry to that zone by the same boat, then compute the time spent in that zone.

3.2 Grouping events close in time

After converting all discrete events into events with a duration, we still ended up with a lot of extremely short events. This is most probably caused by boats trying to stay close to the installation but the dynamic positioning system moves them in and out of the inner or outer zones. To counter this fact we grouped all events with the same boat at the same location within 1 hour into one event. However, after this grouping there is still 63% (or 244) of the events within the first 10 minute window. These are events within a zone that is less than ten minutes in duration and without another event in the same location within one hour of the original event. There are three possible explanations for these strange events: 1. The boat is passing through the location, and not returning for at least one hour. Or otherwise briefly enters and exists the zone, without this fact having any effect on the operation. 2. The boat tries to perform an operation at the location but has to abort and leaves within ten minutes. 3. The event was not registered correctly when the data was gathered. The most probable cause for most of these events are boats that travel through the zone heading for another location. This hypothesis can be tested by removing outer zone events from the distribution. As the inner zone is small, very few of these big fishfeed carrier boats would drive through the inner zone of an aquaculture installation when

[2] `http://www.ecmwf.int/en/research/climate-reanalysis/era-interim`

heading somewhere else. We can still see 244 events that are of duration 10 minutes or less within the inner zone of an aquaculture installation. Figure 4 looks at the 1 minute distribution within the first 10 minutes to try to find the causes for the high number of short stay events. And once again we can see that many of the events are very short, with very few events lasting more than 3 minutes. This further supports our first hypothesis.

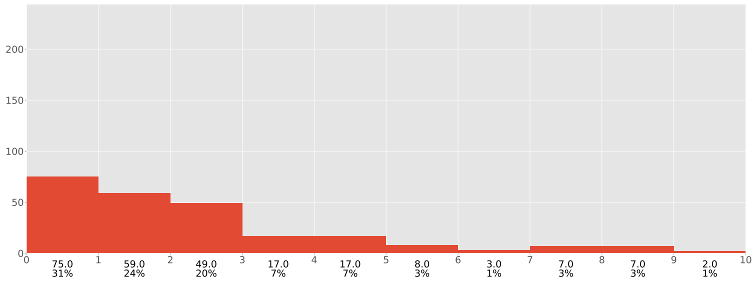


Fig. 4: Distribution of events over length of stays in all inner zone after grouping all events within a 1 hour time window. Zoomed into the first 10 minutes.

One problem with our approach so far is that some events are very far apart in time as well as having different zone types. One example being one boat having a 0 second stay in the inner zone of location 31437 at 18:23 the 28th of November, however the boat entered the outer zone of the same site at 17:04 the same day, and exited zone 1 of that location at 18:24. We can then conclude that the boat spent approximately 1 hour and 20 minutes at the location in the outer zone, then very briefly entered the inner zone before leaving the location. Again supporting the first hypothesis. From this we can see that including inner zone in analyzing fishfeed carrier operations adds very little information to our analysis as the fishfeed carriers do not enter the inner zone when transferring fishfeed. As a consequence we discard the inner zone data. We are still left with 2401 events with a duration shorter than 10 minutes. Fig 5 shows the distribution of these events length in stay. We can see that most of these are shorter than 5 minutes, and most probably does not represent actual maritime operations (or failed tries), but rather traveling through the zone. Thus we discard events shorter than 10 minutes, giving us the final distribution shown in Fig. 6.

3.3 Predicting the operational time using Deep Learning

To extract cases that exemplify instances where the weather conditions stops a fishfeed operation from being successfull, we are currently building a deep learning model aimed at predicting the time spent at the installation, with the given weather and level of exposure at the time and location. The input to the model is: draft and length of the boat, wind speed[3], distance between the model grid point and actual site coordinate, wave direction[3], wind direction[3], maximum level of exposure at location, significant swell wave height[3], month, hour, wind effect (wind speed combined with

[3] Measured at the closest grid point in NORA10

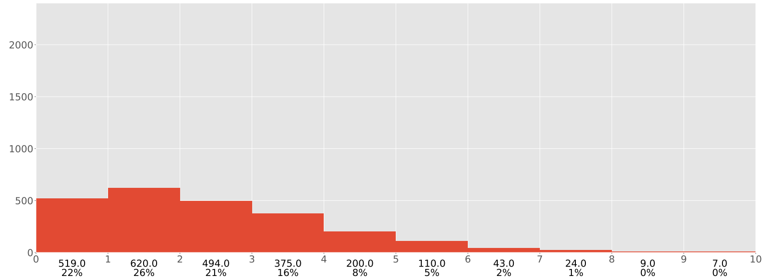


Fig. 5: Distribution of events over length of stays in all outer zone after grouping all events within a 1 hour time window. Zoomed into the first 10 minutes.

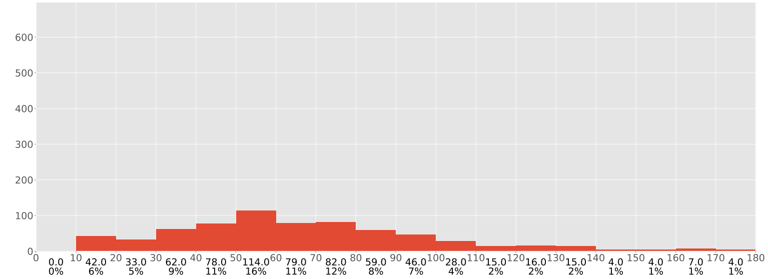


Fig. 6: Distribution of events over length of stays in all outer zone after grouping all events within a 1 hour time window. With all stays smaller than 10 minutes removed.

exposure levels in the wind direction +/- 10 degrees) and significant wave height. The output of the model is the amount of time spent on the installation.

The regression was implemented using python. We used sklearn for preprocessing and scaling (MinMax scaling) of input data (including regression target). The Keras library for deep learning was used for the regression itself, with a input layer of $inputcolumns + 1 = 14$ nodes. We used 3 hidden layers with 13 nodes each and a output layer of 1 node. All nodes used the ReLU activation function.

4 Results

The current results show that there is little information in the gathered data (through the NORA10 model and exposure levels) that account for the variance shown in the time spent at the locations. The neural network models presented in the previous Section 3.3 gets very low accuracy (0.11%, which means the predictor is very slightly better than just outputting the average) in terms of predicting how long a fish feed boat stays at a aquaculture installation. Figure 7 shows the length of all of the events in the chronologically in blue and the predicted length in orange. The "Time Spent" axis is normalized values of the time spent in near a installation where $y = 1.0$ represents the longest stay recorded in the training data. There are obvious differences between predicted and true values; predicted values consistently returns too high values, and fails to predict short stays. A cross validated ($cv = 5$) hyper parameter grid search was performed and showed no better performance at 10 hidden layers with 56 nodes in each hidden layer.

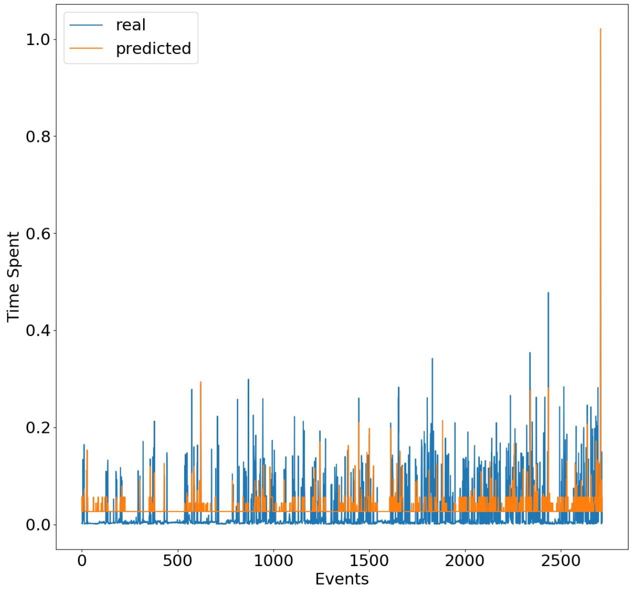
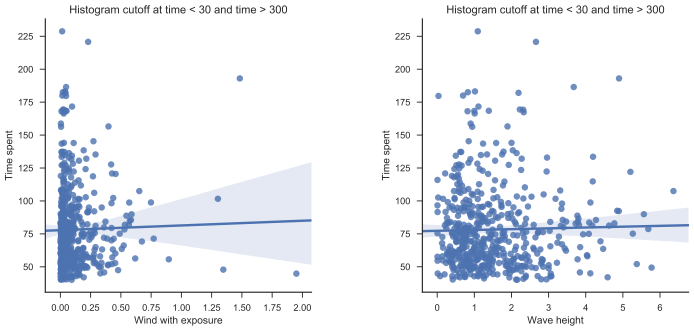


Fig. 7: This shows the DNN model try to predict the amount of time spent at a installation in orange, and the actual time spent in blue. The X-axis is simply the record number, where the record are ordered along the time axis.

After we received the disappointing results we created scatter plots of two weather variables in relation to the length of stay at the installations. Typically most would assume there would be a pattern of some correlation between the weather and the length of stay. However Figure 8 shows that neither wind (8a) or waves (8b) reveals any obvious correlation patterns against time spent at installations.

In addition we did a principal component analysis of the data, to discover if there where any clear principal components that could contain the variance in the data. The components returned: $C = (0.127, 0.117, 0.109, 0.099, 0.091, 0.039, 0.034, 0.028, 0.020, 0.011, 0.008, 0.004, 0.002, 0.000)$ Where the sum of components $sum(C) =$ 0.6967 indicating that the total of the components could account for little of the variance. Finally we tried a standard method for non-linear regression as a base-line result to measure the DNN against. We tried Epsilon-Support Vector Regression (SVR) which scored with a coefficient of determination $R^2 = -0.83$ which is worse than constantly predicting the



(a) Wind correlation

(b) Wave correlation

Fig. 8: Scatter plot illustrating correlation between the weather and the time spent at the installation

mean of the target (which would give $R^2 = 0.0$). This final result shows in the context of the other results listed above us that the data set may not contain the features needed to predict the length of the stay at a installation.

5 Conclusions and future work

We started the work with a hypothesis that whether or not a fishfeed boat operation (loading of fishfeed from boat to barge) succeeded depended on the weather, and that such a failure could be detected from the length of time the fishfeed boat stayed at the aquaculture installation. Our analysis did not find any deterministic correlation between the weather and location data and the length of the stay at the installation. There can be many reasons for this, we will try to list some of the reasons we think are probable;

The first possibility is that despite our efforts to remove noise from the data, the data still contains noise. This includes the three factors listed in the introduction section and other possibilities we have not considered.

Second, given the size of the boats and their stability, they can operate during harsh conditions. In addition these boats are expensive in operation, and even more expensive if they fail to deliver feed at the appointed time, possibly starving the fish at the installation. Thus these boats are already subject to careful operational planning. It may therefore be that there is none to very few failed fishfeed operations in the data captured. An additional consequence is that the time spent during operations has very low variance.

Extending this work would start with confirming these possible explanations for the lack of correlation found in our data. We would also like to gather further data, extending the number of events beyond the current 2700. This would enable us to train and test our models with more rigor and less uncertainty.

6 Acknowledements

None of the work done in this paper would have been possible without the support of the EXPOSED project. Special thanks to ANTEO (`http://anteo.no/`) for providing data to this experiment and working with us to make use of this data.

References

1. Barry Smyth and Elizabeth McKenna. *Modelling the competence of case-bases*, pages 208–220. Lecture Notes in Computer Science. Springer Nature, 1998.
2. Barry Smyth and Elizabeth McKenna. *Building Compact Competent Case-Bases*, pages 329–342. Case-Based Reasoning Research and Development. Springer Nature, 1999.
3. Jun Zhu and Qiang Yang. Remembering to add: competence-preserving case-addition policies for case-base maintenance. In *IJCAI*, volume 99, pages 234–241, 1999.
4. B Smyth and M Keane. Remembering to forget: A competence-preserving deletion policy for cbr. In *Proceedings IJCAI-95*, 1995.
5. Chunsheng Yang, Benoit Farley, and Bob Orchard. Automated case creation and management for diagnostic cbr systems. *Applied Intelligence*, 28(1):17–28, Feb 2007.
6. Qiang Yang and Hong Cheng. Case mining from large databases. *Lecture Notes in Computer Science*, page 691–702.
7. Marvin Zaluski, Nathalie Japkowicz, and Stan Matwin. Case authoring from text and historical experiences. *Lecture Notes in Computer Science*, page 222–236, 2003.
8. Kerstin Bach, Klaus-Dieter Althoff, Régis Newo, and Armin Stahl. *A Case-Based Reasoning Approach for Providing Machine Diagnosis from Service Reports*, pages 363–377. Case-Based Reasoning Research and Development. Springer Nature, 2011.
9. Valmi Dufour-Lussier, Florence Le Ber, Jean Lieber, and Emmanuel Nauer. Automatic case acquisition from texts for process-oriented case-based reasoning. *Information Systems*, 40(nil):153–167, 2014.
10. Benoit Farley. From free-text repair action messages to automated case generation. In *Proceedings of AAAI 1999 Spring Symposium: AI in Equipment Maintenance Service & Support, Technical Reprot SS-99-02, Menlo Park, CA, AAAI Press*, pages 109–118, 1999.
11. Kerstin Bach, Odd Erik Gundersen, Christian Knappskog, and Pinar Öztürk. Automatic case capturing for problematic drilling situations. In *International Conference on Case-Based Reasoning*, pages 48–62. Springer, 2014.
12. Peter Funk and Ning Xiong. Case-based reasoning and knowledge discovery in medical applications with time series. *Computational Intelligence*, 22(3-4):238–253, Aug 2006.
13. Odd Erik Gundersen. *Enhancing the Situation Awareness of Decision Makers by Applying Case-Based Reasoning on Streaming Data*. PhD thesis, NTNU, 2014.
14. Øyvind Breivik, Magnar Reistad, and Hilde Haakenstad. A high-resolution hindcast study for the north sea, the norwegian sea and the barents sea. In *10th International Workshop on Wave Hindcasting and Forecasting*, 2007.
15. Magnar Reistad, Øyvind Breivik, Hilde Haakenstad, Ole Johan Aarnes, Birgitte R Furevik, and Jean-Raymond Bidlot. A high-resolution hindcast of wind and waves for the north sea, the norwegian sea, and the barents sea. *Journal of Geophysical Research: Oceans*, 116(C5), 2011.
16. Pål Lader, David Kristiansen, Morten Alver, Hans. V Bjelland, and Dag Myrhaug. Classification of aquaculture locations in norway with respect to wind wave exposure. In *Proceedings of the ASME 2017 36th International Conference on Ocean, Offshore and Arctic Engineering OMAE2017*, 2017.