Received: 30 August 2020 | Revised: 4 March 2021 | Accepted: 13 March 2021

DOI: 10.1002/ail2.22

LETTER

WILEY

Hierarchical spline for time series prediction: An application to naval ship engine failure rate

Hyunji Moon[1] | **Jinwoo Choi**[2]

[1]Department of Industrial Engineering, Seoul National University, Seoul, Republic of Korea

[2]Defense Management, Korea National Defense University, Nonsan, Republic of Korea

Correspondence
Jinwoo Choi, Defense Management, Korea National Defense University, Hwangsanbeol-ro 1098, Nonsan, Republic of Korea.
Email: chlwlsdn8570@gmail.com

Abstract

Predicting equipment failure is important because it could improve availability and cut down the operating budget. Previous literature has attempted to model failure rate with bathtub-formed function, Weibull distribution, Bayesian network, or analytic hierarchy process. But these models perform well with a sufficient amount of data and could not incorporate the two salient characteristics: imbalanced category and sharing structure. Hierarchical model has the advantage of partial pooling. The proposed model is based on Bayesian hierarchical B-spline. Time series of the failure rate of 99 Republic of Korea Naval ships are modeled hierarchically, where each layer corresponds to ship engine, engine type, and engine archetype. As a result of the analysis, the suggested model predicted the failure rate of an entire lifetime accurately in multiple situational conditions, such as prior knowledge of the engine.

KEYWORDS

B-spline, failure rate, hierarchical model, naval ships data, Stan, time series forecasting

1 | INTRODUCTION

Predicting failure rate is important as it serves as a standard for preventive measures and inventory management. Both over and underestimation of failure are detrimental to the system. Underestimation can lead to mission failure due to maintenance, and overestimation can lead to wasted budget and reduced operational efficiency due to excessive spare part purchases. Therefore, taking account of the features of failure data into the model is important. Two characteristics of failure rate data, imbalanced category and sharing structure, are the main motivation for this paper and we propose a hierarchical spline (HS) model for improvement. First, an imbalanced category refers to the fact that the amount of data corresponding to each age or product type has a high variance. The second is sharing structure. In our case of predicting the failure rate of an engine of each ship, as engines are shared among ships, ships with the same type of engine display similar failure rate patterns. The underlying process also supports the empirical results, as the same engine types share design patterns and are made from the same factory.

Hierarchical model provides a systemic structure to address both imbalanced and sharing nature of data. In our setting, even the failure rate of certain periods without historical data could be predicted through pooling. For this purpose, we have constructed the three-layer model as the following: a root layer that accounts for the core characteristics of an engine, that is, engine archetype, a second layer that corresponds to each type of an engine, and lastly, the third layer that explains the specific characteristics of each ship.

The proposed model has additional advantages in terms of predicting the failure of new engine types. Republic of Korea (ROK) Navy battleships evolve continuously; for example, FF (Fate Frigate) class has been replaced by FFG (Fast Frigate Guided-missile). Predicting the failure rates of a new battleship is hard but necessary. Most existing time series

models such as ARIMA or ETS (exponential smoothing) model struggles in a situation where no quantitative data exist. However, a hierarchical model can construct the outline of the failure function based on the prior qualitative information. For instance, as we will elaborate in section 5, engines constructed in a similar era show similar patterns. Therefore, information on which era the unforeseen engine was made could be utilized to predict its failure rates.

The main contribution of this paper lies in applying HS model to failure data from ROK Navy. Compared to the previous models, the proposed model not only improves overall prediction accuracy but also is capable of predicting failure rates for categories with scarce data robustly. Moreover, the hypothetical similarity between each category can be tested and proved using our model; this enables users to utilize the qualitative knowledge on the unforeseen, ships with new engines for example, for predicting. These results, when used as a reference for maintenance policy and budget allocation, could contribute greatly to the Navy's operating system. However, this model is not limited to the naval domain. When it comes to predicting failure rates, the circumstances where data are hierarchical, imbalanced, or insufficient are common and therefore, our model is widely applicable. For example, mechanical equipment consists of several parts. The generator, which is a part of the wind turbine, is composed of parts such as a motor and a transformer[1] in a hierarchical structure. Using the HS model, it is also possible to predict the failure of equipment components in a hierarchical structure.

The remainder of this paper consists of five sections. Section 2 introduces the background for failure predicting and HS model. The advantage of the chosen model is explained especially in terms of data characteristics for our setting. In section 3, details of ROK Navy data are introduced. Also, HS model is compared with two existing and one newly created model. Section 4 contains an analysis of the experimental models, and lastly, conclusions are presented in section 5.

2 | LITERATURE REVIEW

In this section, imbalanced and missing nature of naval ship engine data is analyzed. Previous prediction methodologies suitable for constructing the failure function are review and the reasons for applying the Bayesian hierarchical model are described in this section. Three models are selected for comparison: ARIMA, Prophet, and Globalized-Prophet. ARIMA is a canonical time series model while Prophet is a relatively recent model. However, Prophet is adapted widely due to its high accuracy and scalability. Our proposed HS model has the advantage over these two models in that it has a global framework. In global model, shared parameters are jointly learned from heterogeneous time series. For a fair comparison between global models, we have additionally developed a new model by adding a global structure to Prophet, namely Globalized-Prophet (G-Prophet).

2.1 | Failure rate in naval ship setting

Before the mission, each naval ship is loaded with a predicted amount of spare parts. An underestimated prediction has a risk of mission failure as spare parts cannot be resupplied during mission times. An overestimated prediction may lead to reduced operating efficiency due to a load of unnecessary spare parts. Moreover, from a system point of view, overestimation induces unnecessary use of budget and even leads to inventory shortage for other ships. So, defining the optimal set of spare parts is crucial for mission success.[2]

For accurate prediction, several special features resulting from the navy's system should be noted. First of all, imbalances are observed in two categories of the data: age period and engine types. A major cause of the missing data is ROK Navy information system's inability to record the operation. From our data, an early age has less data than the rest of the age period; this might be problematic as the failure rate of young ships is needed for operation. Also, the distribution of ships for each engine type category is not balanced. In our dataset with 99 ships, there are 6, 27, 43, 19, and 4 ships for each engine type category. In this case, while a satisfactory model could be obtained from an engine type with a large amount of data, other models might suffer a lack of data problems.

Moreover, the similarity between ships and engines should also be noted as they undergo the same maintenance process; planned maintenance is performed by ROK Navy regardless of the engine type.[3] Based on these circumstances, where engines of different ships share certain qualities, the model with layered parameter structure is needed; it should be able to learn the specific structure between and within each layer from the data.

2.2 | Failure forecasting models

Several models exist such as ARIMA, exponential smoothing, and seasonal trend decomposition using loess[4] that could model time series characteristics of failure rate. Among the existing time series models, Prophet, which adopts Bayesian generalized additive model (GAM), shows high accuracy. Moreover, it decomposes time series into trend, seasonal, other regressor factors that enhance both its application and interpretability.[5] Along with the hierarchical model framework (section 2.3), GAM is known for its ease of information sharing; when two concepts are combined by adding a hyperparameter to GAM, it becomes a hierarchical GAM[6,7] which has been applied in much research.[8]

More specific models concentrating on the characteristics of failure have been suggested. A bathtub is a typical shape pattern observed in the failure rate. Also, Weibull or Poisson distribution is often used as a distribution of failure rate. Wang and Yin[9] performed failure rate predicting with the stochastic ARIMA model and Weibull distribution. Time series data have been decomposed into bathtub-shape assumed trend and stochastic factors. Parameters of the Weibull distribution were separately learned for the increase, decrease, and flat period of the bathtub. The stochastic element was obtained using ARIMA, and the time series failure rate was calculated as the sum of the trend and stochastic elements. Sherbrooke[10] proposed Pareto-optimal algorithms, named constructive algorithms, based on Poisson distribution. However, it had limits in determining the parameter. Zammori et al.[2] tried to solve the problem of parameter estimation of Sherbrooke's[10] model by applying time series Weibull distribution. Other attempts such as Pareto-optimal, Monte-Carlo,[10] ARMA, and least-squares logarithm[9] have been made to add the effect of stochastic factors to this distribution.

Attempts have been made to integrate time series models with information about system architecture. In the risk analysis of deep waters drilling riser fracture,[11] Bayesian network was used to predict the fracture failure rate. Bayesian network could also be used to analyze and prevent the cause of a ship's potential accidents.[12] Time series predicting based on Bayesian network[13] and analytic hierarchy process[3] illustrate these approaches. They are based on the assumption that equipment, engines for example, within the same group follow similar failure patterns.

2.3 | Hierarchical model

The hierarchical model has an edge in representing the features of navy data introduced in section 2.1; imbalanced category and sharing structure, by information pooling. Gelman et al.[14] explained that hierarchical models are highly predictive because of pooling. When a hierarchical model is used, there is almost always an improvement, but to different degrees that depends on the heterogeneity of the observed data.[16,33] When updating the model parameters, such as prior parameters, the relationship between the part of the data being used and the whole population should always be considered. Pooled effects between subclusters are partial as they are implemented through shared hyperparameters, not parameters.

By properly setting the hyperprior structure, we can find a reasonable balance between overfitting and underfitting, as hyperpriors are known to serve as a regularizing factor. Many examples of applying hierarchical structure in cross-sectional data exist in diverse domains, such as ecology, education, business, and epidemiology.[16] The structure of cross-sectional data where the whole population is divided into multiple and nested subcategories provides an excellent environment for a hierarchical model. Previous literature on comparing the education effects of multiple schools has shown that incorporating the nested structure of the state, school, and class in the model had substantial improvement in terms of accuracy and interpretability.[17]

Januschowski et al.[18] have classified methods in the predicting domain into two: global and local. Global methods jointly learn parameters using all available time series while the local methods learn independently from each time series. HS model is global as its hierarchical structure provides the framework to predict new types of engine that has next to no quantitative information via pooling. Note that predictions for new groups are difficult in a model without pooling. The concept of pooling is not restricted to hierarchical model; Trapero et al.[19] achieved pooling by replacing a regression coefficient of stock-keeping units with a limited amount of data with the coefficient calculated from multiple SKUs. Other examples include recurrent neural network models with globally calculated weights.[20] Models that balance global and local information in the context of pooling have also been suggested, an example being pooling within each cluster[21] and twofold spatial attention mechanism in recurrent neural network.[20] For a fair comparison, we have constructed another global model, G-Prophet, by using the fact that Prophet model could be extended to hierarchical GAM as introduced in section 2.2. A detailed description of G-Prophet is in section 3.4.

2.4 | Model evaluation measures

Time series cross-validation and k-fold cross-validation, along with the expanding prediction method, can be used to measure prediction accuracy in time series.[4] Several sets of training and test data are created in a walk-forward mode, and prediction accuracy is computed by averaging over the test sets. Various measures of prediction error exist, including the mean absolute, root mean squared and mean absolute percentage error. To compare the results on different datasets, scale-independent errors including SMAPE (Symmetric Mean Absolute Percentage Error), MAPE (Mean Absolute Percentage Error) are preferred.[22] However, the presence of the predicted or real data in the denominator makes the measure unstable when the values take near-zero values.[4] Also, based on the case where SMAPE takes negative values, Hyndman and Koehler[22] recommended not to use SMAPE. Based on this recommendation and as our comparison experiments are based on one set of data, we chose root mean square error (RMSE) as our measure.

Specific to Bayesian models, measures that could diagnose model fit are provided in Stan, a Bayesian computation software. Bayesian fraction of missing information (BMFI) and effective sample size (ESS) are two examples. BMFI quantifies the efficacy of the momentum resampling between Hamiltonian trajectories and ESS quantifies the accuracy of the Markov chain Monte Carlo estimator of a given function.[23] Garby et al.[24] show graphical summaries based on these measures. Information criteria used to measure the fit of a model in Bayesian models include widely applicable information criterion and the leave-one-out cross-validation (LOOCV); they are preferred to other criteria such as Akaike information criterion and deviance information criterion.[25] Due to computational problems, approximate LOOCV methods exist, including Pareto smoothed importance sampling, which is implemented in a package called loo.[26] The package provides means for model comparison with approximated expected log pointwise predictive density (ELPD). Choosing the validation set to address the sequential characteristic of time series has also been proposed.[27] However, as our proposed and compared models are curve fitting that does not directly address the sequential trait of time series, except for ARIMA, RMSE, and ELPD measures are mainly used in this paper.

Note that several diagnostics on the model and its computation inference tool should be checked before comparing the model. Model checking methods such as predictive checks compare the assumption or result of the model with the given data.[15] Moreover, for accurate parameter estimation, convergence of a Markov chain should be checked; trace plots of parameter samples and numerical summaries such as the potential scale reduction factor, R-hat[28] are used. R-hat lower than 1.05, for each parameter, is recommended.

3 | DATA AND MODEL

In this section, the aforementioned two characteristics, imbalanced and missing, are confirmed on a real dataset. Details of model construction including how basis functions and coefficients for B-spline were designed hierarchically are described. Also, we describe the details of G-Prophet.

3.1 | Data

Data consist of 99 ship engines that are categorized into five types of engines. Therefore, our hierarchical model has a 1-5-99 structure; 1 engine archetype, 5 engine types, and 99 ship engines. The numbers of ships in the five categories are also different as in section 2.1.

Figure 1 shows the annual failure count of each ship by its age. Due to military security concerns, it is impossible to disclose specific values of the data used in this study (data subject to third party restrictions). Engine archetype in the figure highlights the existence of shared characteristics between different types of engines. Engine types are categorized into five types. As can be seen from the figure, there are a lot of missing data. Most missingness is due to the absence of ROK record system. In other words, there was no information system before 2000s when type 2 and 3 engines were in their early ages. Note the gaps between data, for example age 9-11 and 13-14 for fifth ship engine in type 1, are missing mainly because of information system manager's mistake. Also, the amount of data for each category is highly imbalanced. Moreover, the similarity between data under the same category could be inferred; for example, data with the same type of engine display a similar age period. Annual failure counts of 99 ship engines are rearranged according to their types (from type 1 to 5) and ages (from year 1 to 31) resulting in Figure 1. Note that only the records from direct maintenance workshop are included; data for warranty repair which take place at shipyard were unavailable. Due to this lowered failure count data, the early period could have different pattern from the other period which could be an obstacle to partial pooling (section 4.2.1).

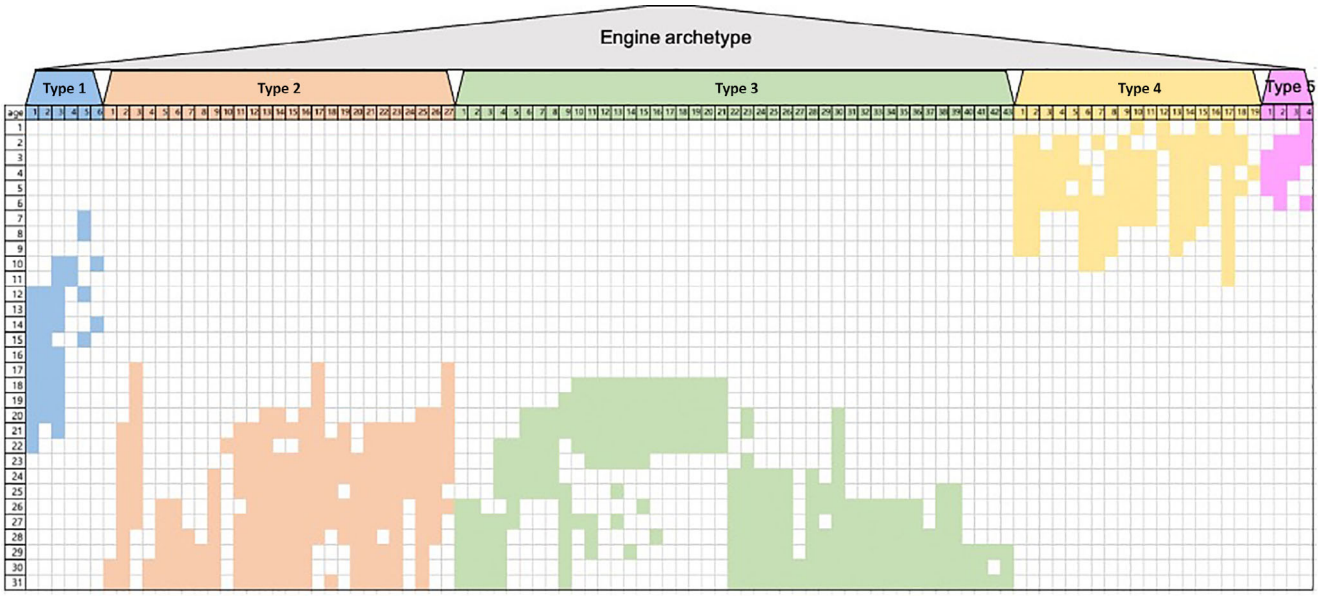

FIGURE 1 Existing data by age and engine type

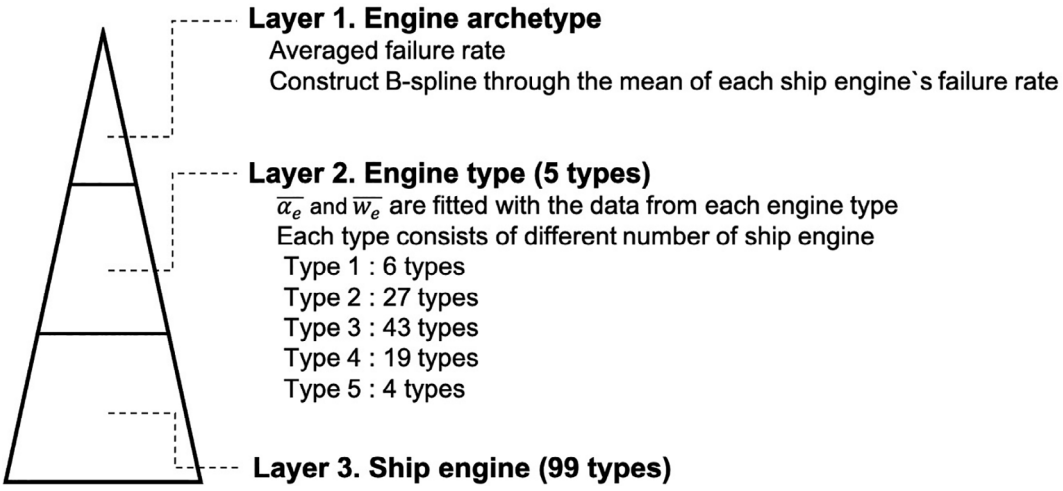

FIGURE 2 Hierarchical structure of ship engine failure

3.2 | HS model

Naval ship engines applied in the proposed model are classified as Figure 2. The ship engine (layer 3) of each ship belongs to the same engine type (layer 2), and five types belong to the engine archetype (layer 1).

Since the naval ship data are nonlinear time series data, polynomial and spline regression are considered. In polynomial regression, to achieve flexibility, a degree should be increased; however, the risk of overfitting becomes higher with its degree. To prevent this and to endow the model a form of locality, the B-Spline model is suggested: the overall life cycle, or age period, is first divided into several sections. Then, low-dimensional polynomial is fitted for each section to form a piecewise polynomial spline.

The third layer of ROK naval ship hierarchy, representing the ship engine, is modeled with B-Spline. As can be seen from Equation (1), parameters for each layer share hyperparameter which enables pooling. To be more specific, basis functions for a degree 3, cubic, spline is constructed. Note that weight, $\mathbf{w}$, is a vector whose length is determined by the number of knots and B_k is the basis function for B-Spline. The value of B_k depends on the number of knots. By pre-fitting these basis to the averaged time series, we get the layer 1 parameters $\overline{\alpha_0}$, $\overline{w_0}$. With these pre-fitted engine archetype (layer 1) parameters, engine type (layer 2) parameters, $\overline{\alpha_e}$ and $\overline{w_e}$ are learned. Ship engine (layer 3) parameters, μ_s, are calculated as α_s, w_s and B_k. α_s

and w_s are learned from $\overline{\alpha_e}$, $\overline{w_e}$. The standard deviation parameters, $\sigma_y, \sigma_w, \sigma_\alpha, \sigma_{\overline{\alpha}}, \sigma_{\overline{w}}$ are calibrated with prior predictive checks. Also, we chose power transformation (Yeo-Johnson) to scale our data to match our prior distributions.

$$
\begin{aligned}
Y_s &\sim \text{Normal}\left(\mu_s, \sigma_y\right) \\
\mu_s &= \alpha_s + \Sigma_{k=1}^{K} w_{k,s} B_k \\
\alpha_s &\sim \text{Normal}(\overline{\alpha_e}, \sigma_\alpha) \\
w_s &\sim \text{Normal}(\overline{w_e}, \sigma_w) \\
\overline{\alpha_e} &\sim \text{Normal}(\overline{\alpha_0}, \sigma_{\overline{\alpha}}) \\
\overline{w_e} &\sim \text{Normal}(\overline{w_0}, \sigma_{\overline{w}}) \\
\sigma_y &\sim \text{Exponential}(1) \\
\sigma_\alpha &\sim \text{Gamma}(10, 10) \\
\sigma_w &\sim \text{Gamma}(10, 10) \\
\sigma_{\overline{\alpha}} &\sim \text{Exponential}(1) \\
\sigma_{\overline{w}} &\sim \text{Exponential}(1)
\end{aligned} \tag{1}
$$

Stan, a probabilistic programming language, is used to implement Equation (1) and the codes could be found in the Appendix. Stan uses HMC (Hamiltonian Monte Carlo) algorithm for sampling, which has the advantage of fast and robust convergence, compared to other algorithms such as Gibbs. Moreover, a detailed diagnosis of the sampling process is provided.[29,30] Diagnosis includes divergence, BFMI, ESS, and R-hat. Figure 3 is the summary of diagnostics provided by the software and shows that HS has satisfied the above criteria overall. Detailed analysis on each of the above could be achieved with the help of posterior package.[31] For example, Figure 4 uses kernel density estimate and trace plot to examine convergence for each chain. In the left of Figure 4, likelihood is plotted against values of $\overline{\alpha_e}$ for each chain. The right of Figure 4 shows that all chains converged well. Figures 5 and 6 present BFMI, ESS of $\overline{\alpha_e}$. BFMI diagnoses the accuracy of the HMC sampler, especially its momentum resampling and BFMI values over 0.2 are recommended for each chain. Also, ESS greater than 100 times the number of chains (which is four) is recommended. ESS surpasses the minimum level (Figure 6). Therefore, plots indicate that the samples from our constructed model are safe to use for further inference. For more information on each diagnosis method, refer to Betancourt.[32] Lastly, R-hat measures convergence by comparing the variance within and between different chains and values. All R-hat values from different parameters were within the recommend bounds, from 1.0 to 1.05.[23]

Posterior predictive check was used to validate the model. Posterior predictive check simulates data from a fitted model and compares it with the real data.[33] Systematic discrepancies between real and simulated data[15] could be detected with this test. The result of our model is shown in Figure 7.

3.3 | Process

Workflow is organized as shown in Figure 8. From failure rate data, a rough trend of the failure rate over a lifetime is deduced by averaging existing ship engine failure data from 99 ships. This averaged time series is used to determine the values of the layer 1 hyperparameters. With initial values and model, we fit the model to existing data. Note that this learning process is iterative as it includes model calibration based on predictive check and parameter plots as suggested in section 3.2. Lastly, with the inferred parameters, our quantity of interest, annual failure count, is predicted.

```
Checking sampler transitions for divergences.
No divergent transitions found.

Checking E-BFMI - sampler transitions HMC potential energy.
E-BFMI satisfactory for all transitions.

Effective sample size satisfactory.

Split R-hat values satisfactory all parameters.

Processing complete, no problems detected.
```

FIGURE 3 Diagnosis results of Stan

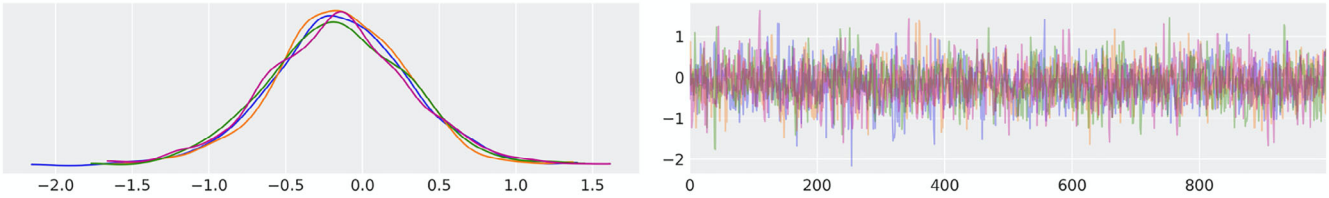

FIGURE 4 Kernel density estimate and trace plot for \bar{\alpha_e}

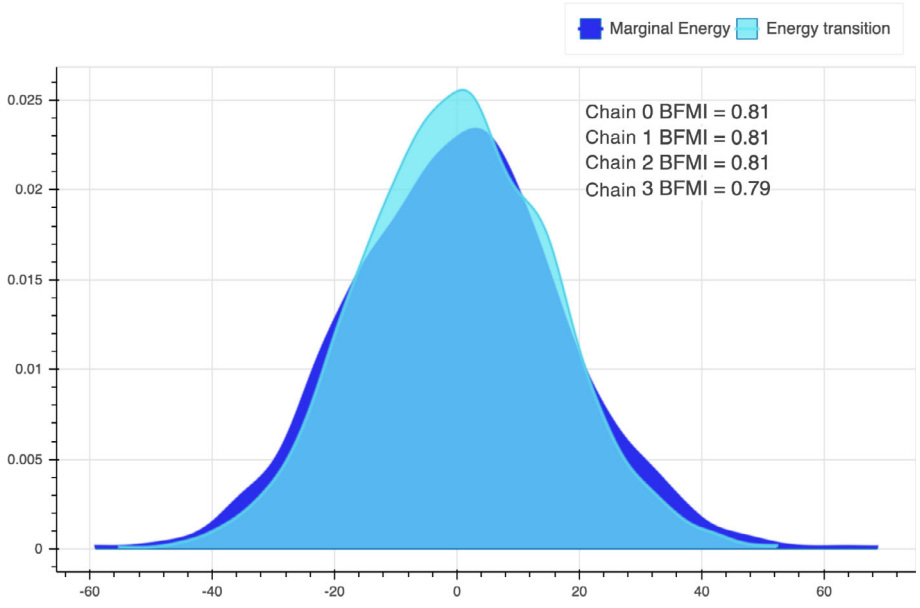

FIGURE 5 BFMI of \bar{\alpha_e}

3.4 | G-Prophet model

As G-Prophet is newly created for comparison purposes, it needs further description. Compared to Prophet which constructs a separate model for each ship engine, G-Prophet pools information from other ship engines before predicting

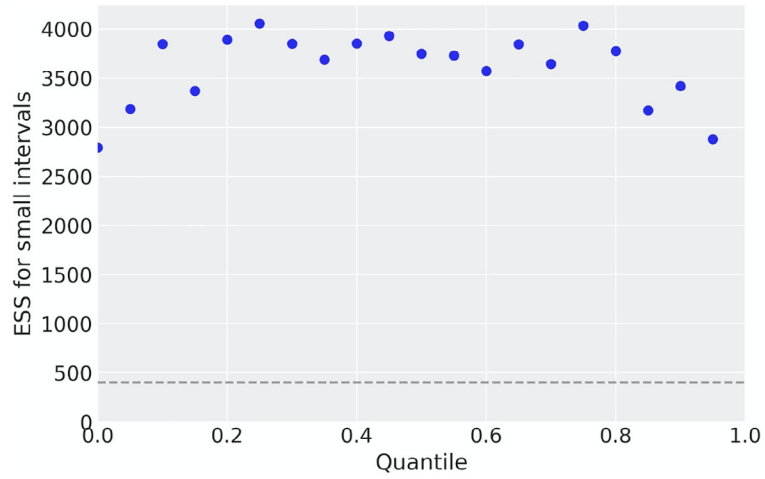

FIGURE 6 Effective sample size (ESS) of \bar{\alpha_e}

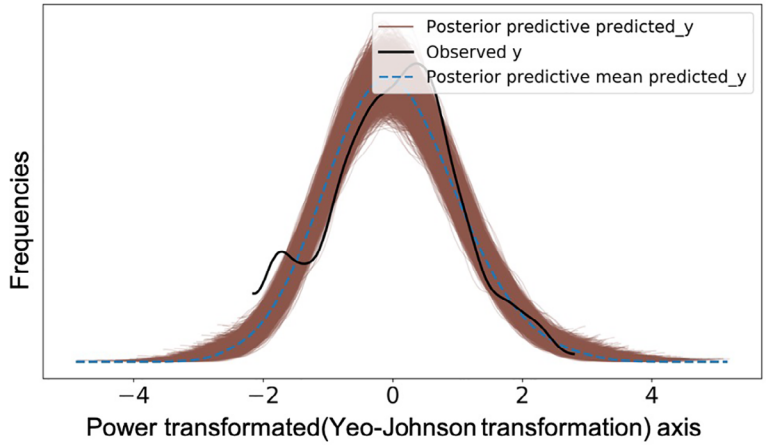

FIGURE 7 Posterior predictive check

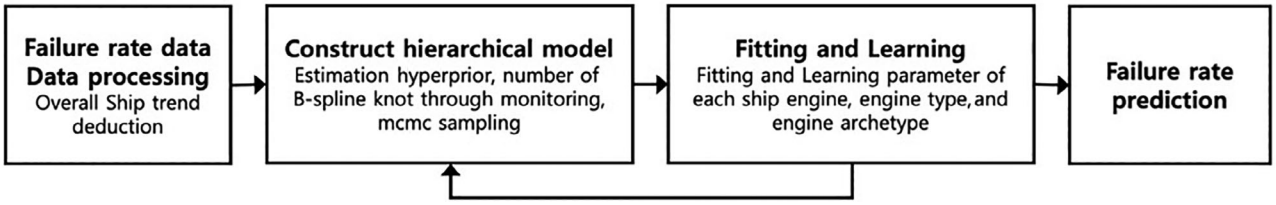

FIGURE 8 Workflow

the failure count of a specific ship engine. This is through weight averaging parameters from each layer. The process is as follows. First, coefficients of time series components (recall that Prophet is GAM) are separately learned with three time series from each layer: averaged time series of all 99 ship engines (layer 1), averaged time series of a certain type of engine (layer 2), and each ship engine (layer 3). Then, parameters from each layer are averaged with a weight that is inversely proportional to the errors of the previous fit. Averaged parameter values are plugged in to the GAM model for the final prediction of each ship engine. Overfitting is prevented by G-prophet which is important as the amount of data is very small for layer 3 (section 3.1). Differences caused by transforming the local Prophet model to global are shown in Figures 9 and 10. Predicted failure counts of layer 3 are much stable for G-Prophet.

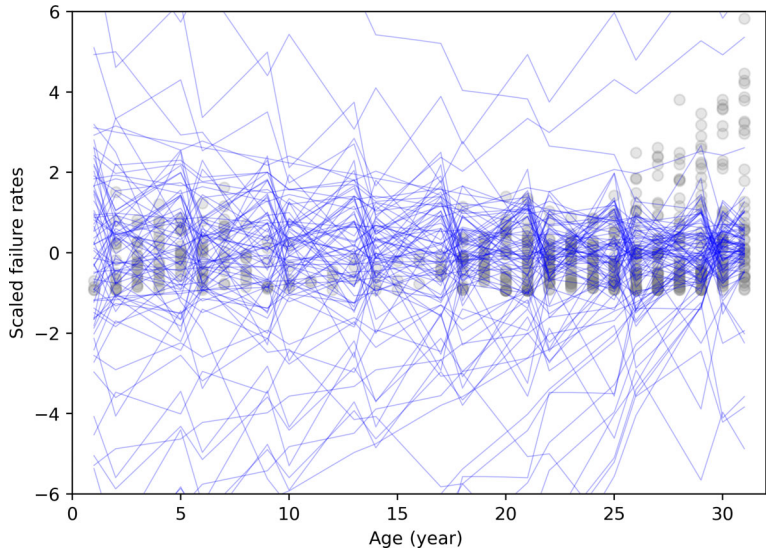

FIGURE 9 Prophet fitting (layer 3)

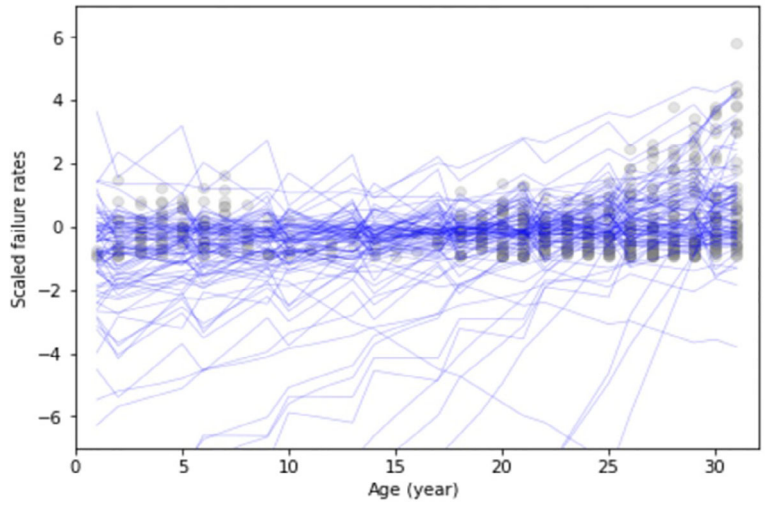

FIGURE 10 G-Prophet fitting (layer 3)

4 | RESULTS AND DISCUSSION

Predictions from four models are plotted and compared. Considering the ROK Navy system, four types of prediction scenarios exist. The first two are based on train sets and the last two are on test sets. We first compare the overall RMSE and ELPD of 99 ship engines and secondly of each subcategory: engine type. The purpose is to observe the effect of pooling. Test set predictions are divided according to whether the engine types of test data are included in the train set or not. Different results between the model and contexts are analyzed. Also, the similarities between the failure functions of each engine type are measured and analyzed.

4.1 | Accuracy comparison

Figures 11, 12, 13, and 14 show the data and four models' prediction results for the total lifetime failure rate of 99 ship engines. Data are plotted with dots while the model predictions are plotted with lines. For two global models, HS and G-

Prophet, we have used different colors for predictions from each layer: engine archetype (blue), engine type (purple), and ship engine (green). The blue line from Figures 13 and 14 is Prophet and ARIMA's prediction for averaged time series. For local models, models are separately constructed with data from each ship engine; as it is impossible to present 99 plots averaged time series was used for comparison in Figures 13 and 14. We would like to note that when ARIMA model returned 0 value due to lack of data (unable to fit), we have adjusted the model to use only moving average without the seasonality.

Accuracy comparison was performed in a way that considers the model's application. In general, when it is necessary to introduce a new ship engine or predict the ship engine in use, refer to the data of the same engine type. The prediction accuracy of the ship engine is used as a reference to predict the spare parts of the ship engine in use, and the prediction accuracy between the engine type and the ship engine is used as a reference when introducing a new ship engine. Therefore, Table 1, the accuracy of the ship engine prediction (predicted layer 3) and observed ship engine data (actual layer 3) values were compared, and second (Table 2), the accuracy of engine type prediction (predicted layer 2) and observed ship engine data (actual layer 3) values were compared. RMSE and ELPD (except ARIMA) were used for the error (accuracy) measure. Lower RMSE and higher ELPD indicate a better model. The values in Tables 1-4) are the average values of RMSE and ELPD calculated from each data.

The first was to compare the average by obtaining the prediction accuracy of each of the 99 ship engines. To model the three-layer dataset, only one option exists for the hierarchical model and G-Prophet. This is because the hierarchical

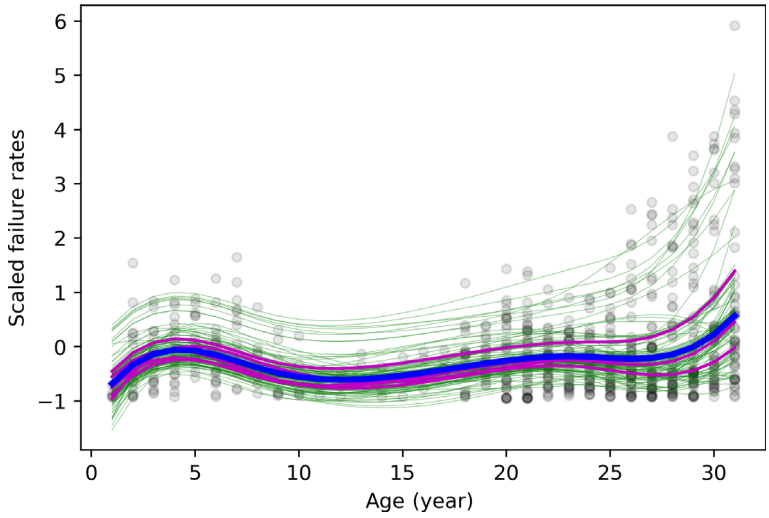

FIGURE 11 Hierarchical spline

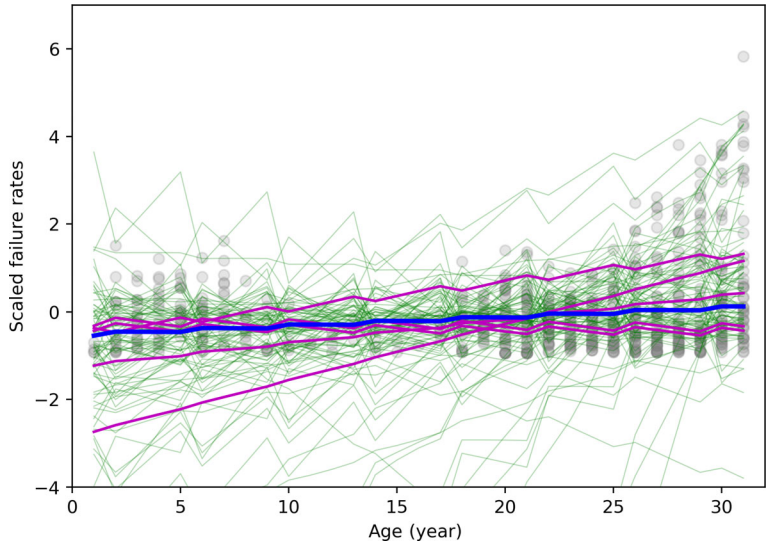

FIGURE 12 G-Prophet

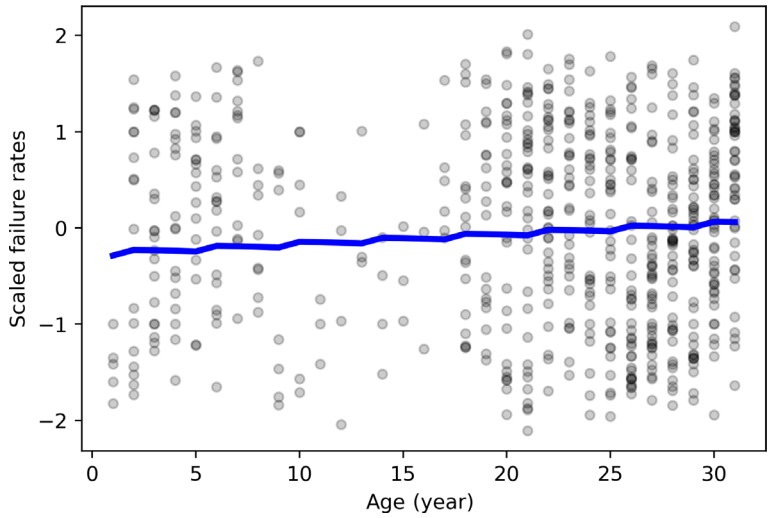

FIGURE 13 Prophet

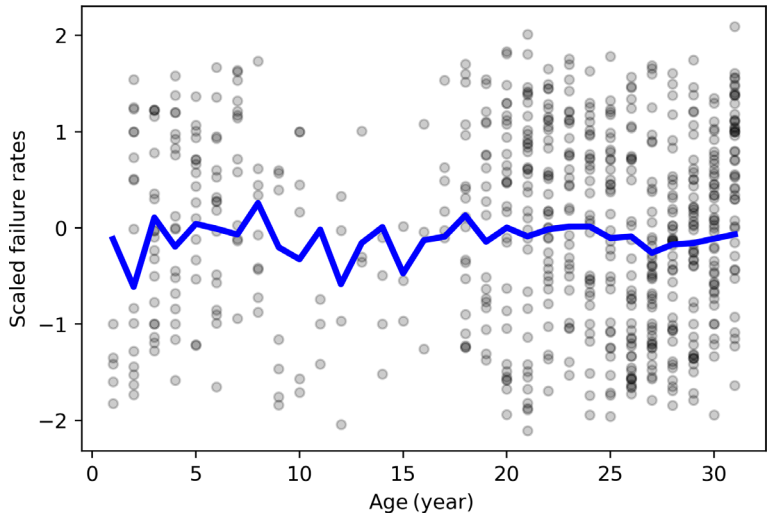

FIGURE 14 ARIMA

model predicts using the information from all layers. G-Prophet predicts using the weighted trend of all layers and seasonality of layer 3. For Prophet and ARIMA, which are unable to represent the hierarchical structure, input data should be preprocessed, by averaging (as in section 3.2), to learn the parameters.

As can be seen from Tables 1 and 2, HS model has the best results with the lowest RMSE and highest ELPD. Note that the difference in RMSE from Table 1 is large once unscaled to original values which unfortunately could not be presented due to security concerns. However, based on the current budget amount, at least several million dollars of budget savings are expected upon application of this model. The effect would be even greater if this model is applied to other armed force departments, Airforce and Army, which have a similar structure with Navy. HS model best describes the data (Tables 1 and 2).

4.2 | Predicting new ship engines or engine types

When we fit the hierarchical model with failure rates of 99 ships, the learned results are stored in the model in the form of each parameter's distribution, that is, posterior. For example, whose prior had exponential form would evolve into a posterior distribution. Bayes formula explains this mechanism. As discussed in the introduction, engine failure rate of a

	HS	G-Prophet	Prophet	ARIMA
RMSE	0.9475	1.0068	1.0395	0.9553
ELPD	−0.6307	−1.2185	−1.2506	—

Abbreviations: ELPD, expected log pointwise predictive density; HS, hierarchical spline; RMSE, root mean square error.

TABLE 1 Average RMSE and ELPD of existing ship engine (layer 3) prediction

	RMSE (ELPD)			
Engine type	HS	G-Prophet	Prophet	ARIMA
Type 1	0.9761 (−0.8347)	1.0080 (−1.6198)	1.0188 (−1.4417)	0.9773
Type 2	0.9567 (−0.8878)	0.9956 (−1.8647)	0.9950 (−1.0791)	0.9709
Type 3	0.9697 (−1.2031)	0.9911 (−3.6487)	0.9981 (−3.8671)	0.9883
Type 4	0.9426 (−0.9662)	1.0462 (−1.1958)	1.0103 (−1.2017)	0.9432
Type 5	0.9593 (−0.9119)	1.0082 (−1.3414)	1.0123 (−1.3511)	0.9802

Abbreviations: ELPD, expected log pointwise predictive density; HS, hierarchical spline; RMSE, root mean square error.

TABLE 2 Average RMSE and ELPD of existing ship engine (layer 3) from existing engine type (layer 2)

new type of engine or ship is frequently needed. Depending on its engine type, the way by which the hierarchical model should be applied differs. If its engine type is present among the data, the posterior of parameters corresponding to layer 2 could be used for the prediction (section 4.2.1). On the other hand, if the engine type is new as well, the only information we could borrow from the previous data are posteriors of layer 1 parameters (section 4.2.2).

Test set data is shown in Figure 15. Type 1 to 5 are the same as the five engine types included in train set data. One ship engine data was obtained for each engine type and prepared as a test set. Type 6 to 10 are new engine types not included in train set data. Ship engine data corresponding to five new engine types were prepared as a test set for each type.

4.2.1 | New ship engine with engine type included in the training set

Posterior of $\overline{\alpha_e}$ and $\overline{w_e}$ could be directly used to predict engine failure of a new ship engine whose engine type is among the five trained engine types. As in section 4.1, G-Prophet, Prophet, and ARIMA were used as comparative models. The results are shown in Figure 16. RMSE of type 3 and type 5 for HS were higher than the other models. As mentioned in section 3.1, early period data show a different pattern compared to the rest of the period (exclusion of warranty repair). Therefore, pooling might have made the prediction less accurate as different patterns are mixed. The fact that type 5 is a relatively minor category also contributes to this analysis; type 4 also corresponds to the early period, but as it has a larger amount of data (five times larger than type 5) the advantageous and disadvantageous effects of pooling could have been offset.

On the other hand, type 3 corresponds to the last age. As shown in Figures 11 to 14, failure rates have high variation at the last age. Accumulated differences of usage environment could be the cause; some operators manage the engine poorly while others with great care. Due to this great variance, we believed test samples that only include six instances (type 3 for example) were not representative enough which is why we used cross-validation for all types for the accuracy measure. Figure 17 is the summary of the process. In general, cross-validation divides the data into train and test set. Unlike this, Figure 17 performed by replacing the test set with the train set. Table 3 shows the results.

The HS model had the highest mean of ELPD and the lowest mean of RMSE (except for type 5). For type 5, the difference in RMSE from other models is reduced compared to Figure 16. In section 4.1, HS model showed lower RMSE than G-Prophet, Prophet and ARIMA too. Compared to section 4.1, the increase in RMSE means of engine types were the smallest in HS. The effect of hierarchical information pooling of HS model was significant when applying new data that was not learned. ROK Navy continues to introduce the same engine type over the long term. Table 3 is useful in many cases, for example, securing a maintenance plan or purchase budget for spare parts during the total life cycle.

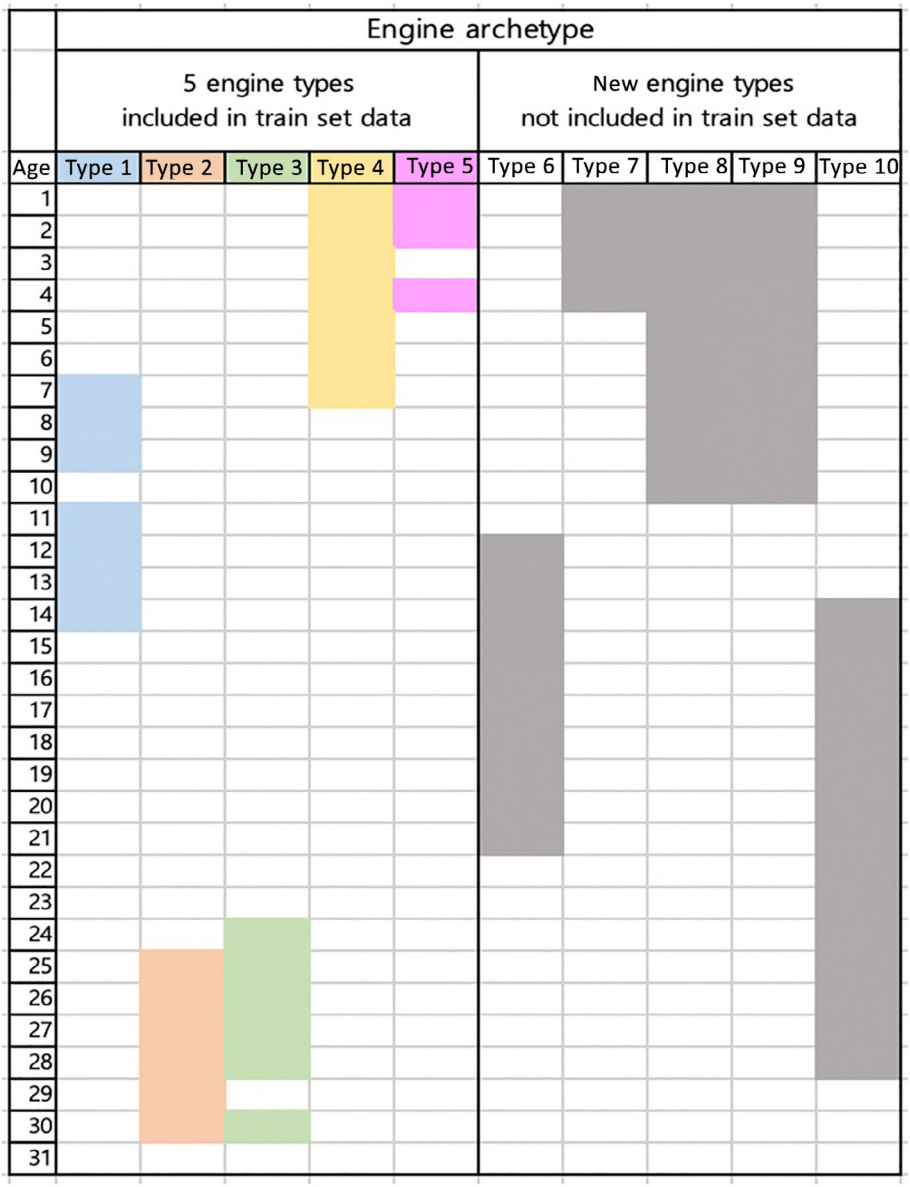

FIGURE 15 Test set data

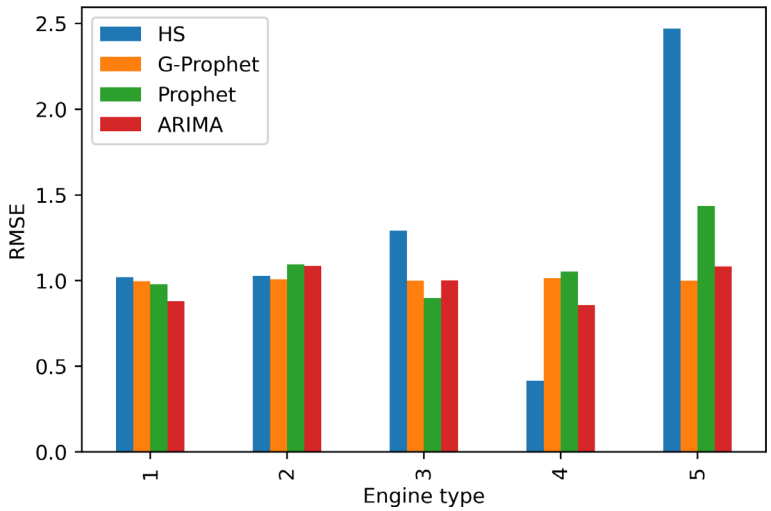

FIGURE 16 Root mean square error (RMSE) of type 1 to 5 test set

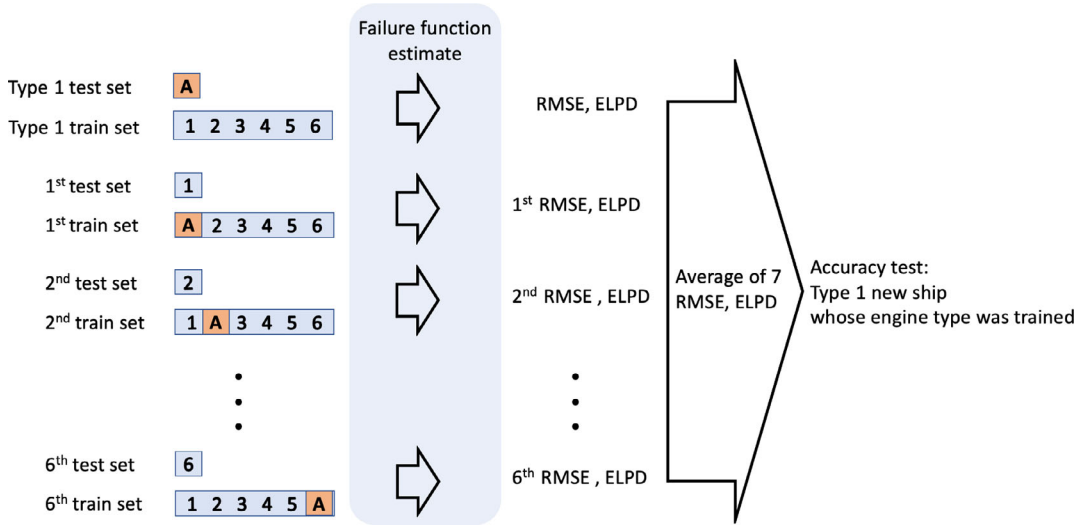

FIGURE 17 Cross-validation example of type 1

Engine type	RMSE (ELPD)			
	HS	G-Prophet	Prophet	ARIMA
Type 1	0.9374 (−0.8431)	0.9959 (−2.5303)	0.9622 (−2.4390)	1.1886
Type 2	0.9884 (−0.8054)	1.0073 (−3.1071)	0.9887 (−2.4391)	1.1151
Type 3	0.9731 (−1.0025)	0.9988 (−2.4545)	0.9996 (−2.5149)	1.0471
Type 4	0.9658 (−1.0218)	1.0133 (−3.5159)	1.0594 (−3.0400)	1.0342
Type 5	1.0131 (−0.8532)	0.9984 (−1.7807)	1.2092 (−2.8526)	1.0962

Abbreviations: ELPD, expected log pointwise predictive density; HS, hierarchical spline; RMSE, root mean square error.

TABLE 3 Average RMSE and ELPD of new ship engine (layer 3) from existing engine type (layer 2)

4.2.2 | New ship with its engine type not included in train set

Using HS, estimation for unforeseen engine type could be performed in a robust manner; information on engine archetype is stored in hyperparameters of layer 1 with which prediction can be made. In other words, the pre-fitted values of $\overline{\alpha_0}$, $\overline{w_0}$, and B-spline on averaged time series data are used for $\overline{\alpha_s}$ and $\overline{w_s}$ (layer 3 parameters) from Equation (1).

For new engine type (layer 2) prediction, information can be pooled from engine archetype (layer 1). These situations are very frequent since there is a constant need to replace or upgrade the engine following the technology development. We confirmed that HS performed well in this case (Table 4).

HS model showed lower RMSE compared to the other three models for most engine types, except type 8. G-Prophet and Prophet have the same prediction results. As explained in section 3.4, G-Prophet is constructed with weight averaged parameters of layers corresponding to data present. For new ship engine (layer 3) from existing engine type (layer 2), layers 1 and 2 are used for layer 3 prediction. However, for new ship engine (layer 3) with new engine type (layer 2), only layer 1 parameters are available which is why Prophet and G-Prophet results are the same. Averaged value is for which could be used as expected error levels for new type of ship engine. HS shows the best result on average followed by ARIMA and Prophet. Note that even though hyperparameters are from the second layer established with previous data (type 1-5), the resulting predictions for new engine types (type 7-10) are different because their existing data are added to the model.

4.3 | Reflecting the qualitative knowledge

Prediction can be improved in the presence of qualitative knowledge, construction era of the new engine type, for example. This act of translating qualitative into quantitative knowledge could be justified by analyzing their

TABLE 4 Average RMSE and ELPD of new ship engine type (layer 3) from new engine type (layer 2)

	RMSE (ELPD)			
New engine type	**HS**	**G-Prophet**	**Prophet**	**ARIMA**
Type 6	0.9620 (−2.9535)	1.0155 (−4.8511)	1.0155 (−4.8511)	1.1078
Type 7	0.7782 (−2.3136)	1.0629 (−5.1454)	1.0629 (−5.1454)	0.7961
Type 8	1.1069 (−3.1876)	1.0563 (−5.1652)	1.0563 (−5.1652)	1.0263
Type 9	0.8546 (−2.5311)	1.0352 (−5.0259)	1.0352 (−5.0259)	1.0031
Type 10	0.9128 (−2.2613)	1.0064 (−4.8539)	1.0064 (−4.8539)	0.9991
Average	0.9229 (−2.6494)	1.0353 (−5.0083)	1.0353 (−5.0083)	0.9865

Abbreviations: ELPD, expected log pointwise predictive density; HS, hierarchical spline; RMSE, root mean square error.

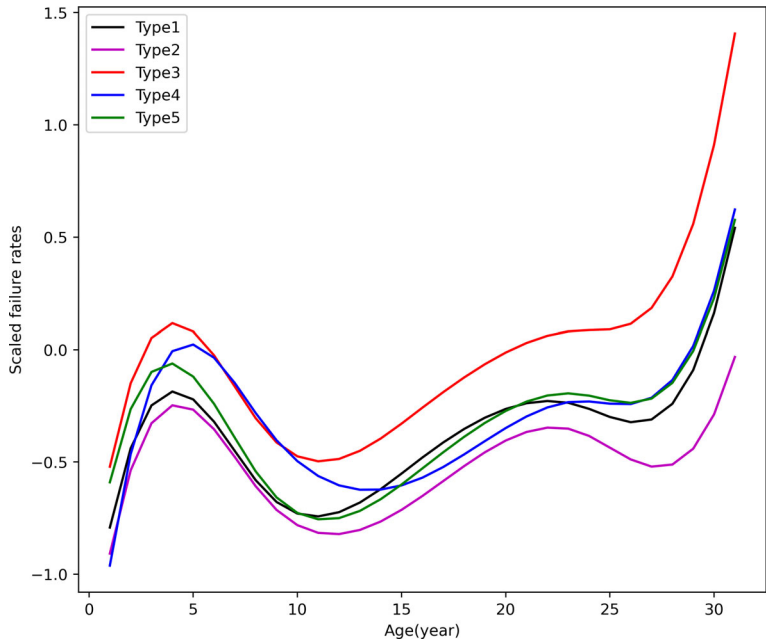

FIGURE 18 Distance between engine types

relationship with the existing failure functions of five engine types. Figure 18 and Table 5 give interpretable results. Engine failure function is largely influenced by the construction era. Based on the historical data, we have classified type 4 and 5 as Early, type 1 as Middle, and type 2 and 3 as Last. As shown in Table 5, in general, the Euclidean distance between Early and Middle is small compared to Early and Last. This could be understood in terms of the development of engine technology and supports the result of our model. To be more specific, a similar construction era resulted in a similar trend in failure function except type 2 and 3. The distance between these old engines is large. We think this could be the result of accumulated differences in the ship usage environment. Early aged engines would show similar failure patterns between types compared to older engines. Other factors including environmental (East or West sea) and purpose (shipping or guarding in the frontal line) factors would affect failures in old aged engines greatly. Second is technology development. It can be said that the latest engines have similar failure functions. Type 4 and 5 engines were constructed after 2010 while type 2 and 3 engines were constructed in the 1990s.

Based on the qualitative knowledge on the closeness of new engine types with the existing engine types, posterior of $\overline{\alpha_e}$ and $\overline{w_e}$ of previous engine types could be used as a hyperprior for new ship's $\overline{\alpha_s}$ and $\overline{w_s}$. Compared to using the original prior Normal$(\overline{\alpha_0}, \sigma_{\overline{\alpha}})$ and Normal$(\overline{w_0}, \sigma_{\overline{w}})$ from Equation (1), this would give more accurate results as more prior knowledge could be reflected for the prediction.

TABLE 5 Euclidean distance

Engine type		Euclidean distance
Early vs Middle	1 vs 5	0.0425
Early vs Early	4 vs 5	0.0461
Early vs Middle	1 vs 4	0.0466
Middle vs Last	1 vs 3	0.1057
Early vs Last	2 vs 5	0.1076
Middle vs Last	1 vs 2	0.1090
Early vs Last	3 vs 5	0.1237
Early vs Last	2 vs 4	0.1244
Early vs Last	3 vs 4	0.1278
Last vs Last	2 vs 3	0.1836

5 | CONCLUSIONS

We have proposed using HS to develop a hierarchical model for predicting failure rates. This approach shines especially when the data are imbalanced and hierarchically structured. We demonstrated the applicability of the model using a real-world dataset of failure rate data from Naval ship engines and compared it with previous methods. Through these comparisons, we confirmed that the prediction performance of our novel model in the given dataset was greatly improved. Moreover, we have shown how qualitative knowledge, such as belonging to the same series or construction era, could be incorporated into the model; this approach was justified by further analyzing the relationship between each parameter. These techniques could greatly improve naval ship management efficiency.

Some improvements could be noted for further studies. First, prevention repair which may affect the failure pattern could be considered. A more advanced model that incorporates the probability of failure after the prevention repair is needed to design a model. Second, due to substantial operational differences between combat and noncombat ships, only combat ships are used in this paper. However, if the differences could be incorporated in the further models, by using categorical variables, a more accurate model could be possible based on a larger amount of data.

HS can contribute greatly to the following areas. First, failure rate prediction could be used as a quantitative reference when establishing a maintenance policy. Proper maintenance not only improves the availability and mission completion rates but also reduces the budget by reducing unnecessary maintenance. Second, from a broader perspective, the predicted failure trend can be a qualitative reference for designing the optimal life cycle of a ship. For instance, based on our results, the failure rate increases dramatically as the ship becomes senile. Therefore, an optimal retirement period could be decided by balancing the maintenance and construction costs.

DATA AVAILABILITY STATEMENT

Data subject to third party restrictions.

ORCID

Jinwoo Choi https://orcid.org/0000-0002-0509-5964

REFERENCES

1. Scheu MN, Tremps L, Smolka U, Kolios A, Brennan F. A systematic failure mode effects and criticality analysis for offshore wind turbine systems towards integrated condition based maintenance strategies. *Ocean Eng*. 2019;176:118-133.
2. Zammori F, Bertolini M, Mezzogori D. A constructive algorithm to maximize the useful life of a mechanical system subjected to aging, with non-resuppliable spares parts. *Int J Ind Eng Comput*. 2020;11(1):17-34.
3. Yoo JM, Yoon SW, Lee SH. SNA-based trend analysis of naval ship maintenance. *J Korea Soc Comput Inf*. 2019;24(6):165-174.
4. Hyndman RJ, Athanasopoulos G. *Forecasting: Principles and Practice*. Australia: OTexts; 2018.
5. Taylor SJ, Letham B. Forecasting at scale. *Am Statist*. 2018;72(1):37-45.
6. Smith AC, Edwards BP. Improved status and trend estimates from the North American Breeding Bird Survey using a hierarchical Bayesian generalized additive model. *bioRxiv*; 2020.

7. Wood SN. *Generalized Additive Models: An Introduction with R*. 2nd ed. Portland, OR: CRC Press; 2017.
8. Pedersen EJ, Miller DL, Simpson GL, Ross N. Hierarchical generalized additive models in ecology: an introduction with mgcv. *PeerJ*. 2019;7:e6876.
9. Wang J, Yin H. Failure rate prediction model of substation equipment based on Weibull distribution and time series analysis. *IEEE Access*. 2019;7:85298-85309.
10. Sherbrooke CC. *Optimal Inventory Modeling of Systems: Multi-Echelon Techniques*. Vol 72. Stanford University: Springer Science & Business Media; 2006.
11. Chang Y, Zhang C, Wu X, et al. A Bayesian network model for risk analysis of deepwater drilling riser fracture failure. *Ocean Eng*. 2019; 181:1-12.
12. Afenyo M, Khan F, Veitch B, Yang M. Arctic shipping accident scenario analysis using Bayesian network approach. *Ocean Eng*. 2017; 133:224-230.
13. Dikis K, Lazakis I. Dynamic predictive reliability assessment of ship systems. *Int J Nav Archit Ocean Eng*. 2019;1-44. https://doi.org/10.1016/j.ijnaoe.2019.01.002.
14. Gelman, A. (2006). Multilevel (hierarchical) modeling: what it can and cannot do. *Technometrics*, 48(3):432-435.
15. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. *Bayesian Data Analysis*. Florida: CRC Press; 2013.
16. McElreath R. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Florida: CRC Press; 2020.
17. Rubin, D. B. (1981). The bayesian bootstrap. *The annals of statistics*, 9(1):130-134.
18. Januschowski T, Gasthaus J, Wang Y, et al. Criteria for classifying forecasting methods. *Int J Forecast*. 2020;36(1):167-177.
19. Trapero, J. R., Kourentzes, N., & Fildes, R. (2015). On the identification of sales forecasting models in the presence of promotions. *Journal of the operational Research Society*, 66(2):299-307.
20. Hewamalage H, Bergmeir C, Bandara K. Recurrent neural networks for time series forecasting: current status and future directions. *Int J Forecast*. 2020;37(1):388-427.
21. Moon H, Song B. Time unit clustering model for pallet movement amount. *Korean J Logist*. 2019;27(4):1-10.
22. Hyndman RJ, Koehler AB. Another look at measures of forecast accuracy. *Int J Forecast*. 2006;22:679-688.
23. Betancourt M. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*; 2017.
24. Gabry J, Simpson D, Vehtari A, Betancourt M, Gelman A. Visualization in Bayesian workflow. *J R Stat Soc A*. 2018;182:389-402.arXiv preprint: http://arxiv.org/abs/1709.01449.
25. Vehtari A, Lampinen J. Bayesian model assessment and comparison using cross-validation predictive densities. *Neural Comput*. 2002;14 (10):2439-2468.
26. Vehtari A, Gelman A, Gabry J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat Comput*. 2017; 27(5):1413-1432.
27. Bürkner PC, Gabry J, Vehtari A. Approximate leave-future-out cross-validation for Bayesian time series models. *J Stat Comput Simul*. 2020;90:2499-2523.
28. Stan Development Team. 2017.
29. Betancourt M, Girolami M. Hamiltonian Monte Carlo for hierarchical models. *Curr Trends Bayesian Methodol Appl*. 2015;79(30):2-4.
30. Carpenter B, Gelman A, Hoffman M, et al. Stan: a probabilistic programming language. *J Stat. Softw*. 2017;76(1):1-32.
31. Vehtari A, Gelman A, Simpson D, Carpenter B, Bürkner PC. Rank-Normalization, Folding, and Localization: An Improved R for Assessing Convergence of MCMC. *Bayesian Analysis*. 2021;1(1):1-28. https://doi.org/10.1214/20-BA1221.
32. Betancourt M. Diagnosing suboptimal cotangent disintegrations in Hamiltonian Monte Carlo. *arXiv preprint arXiv:1604.00695*; 2016.
33. Gelman A, Hill J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press; 2006.

How to cite this article: Moon H, Choi J. Hierarchical spline for time series prediction: An application to naval ship engine failure rate. *Applied AI Letters*. 2021;2:e22. https://doi.org/10.1002/ail2.22

APPENDIX A

Stan code hierarchical spline (HS) model

```
data {
    int <lower = 1> K;
    int <lower = 1> N;
    int <lower = 1> T;
    int <lower = 1> S;
    int <lower = 1> E;
```

```
    int <lower = 1> Age[N];
    int <lower = 1> Ship[N];
    int <lower = 1> S2E[S];
    matrix[T,K] B;
    real mu_a_bar;
    real mu_w_bar[K];
    vector [N] Y;
}

parameters {
    vector[S] a;
    real a_bar[E];
    vector[K] w[S];
    vector[K] w_bar[E];
    real<lower=0> s_a;
    real<lower=0> s_w;
    real<lower=0> s_a_bar;
    real<lower=0> s_w_bar;
    real<lower=0> s_Y;
}

transformed parameters {
    vector [N] mu;
    for (n in 1: N){
        mu[n] = a[Ship[n]] + B[Age[n]] * w[Ship[n]];
    }
}

model {
    s_a ~ gamma(10,10);
    s_w ~ gamma(10,10);
    s_a_bar ~ exponential(1);
    s_w_bar ~ exponential(1);
    s_Y ~ exponential(1);

    for (s in 1:S){
        a[s] ~ normal(a_bar[S2E[s]], s_a);
        w[s] ~ normal(w_bar[S2E[s]], s_w);
    }

    for (e in 1:E){
        a_bar[e] ~ normal(mu_a_bar, s_a_bar);
        w_bar[e] ~ normal(mu_w_bar,s_w_bar);
    }

    Y ~ normal(mu, s_Y);
}
generated quantities{
    vector[N] log_likelihood;
    for (i in 1:N) {
        log_likelihood[i] = normal_lpdf(Y[i]|mu[i], s_Y);
    }
}
```