Bayesian Vehicle Fleet Survival Analysis based on Workshop-Service Data

Simon Steinberg[1], Wolf Baumann[1], Rene Gegusch[1], Philipp Schmiechen[1], and Dominik Gütermann[1]

[1] *IAV GmbH, Berlin, Germany*
simon.steinberg@iav.de
wolf.baumann@iav.de
rene.gegusch@iav.de
philipp.schmiechen@iav.de
dominik.guetermann@iav.de

ABSTRACT

This paper presents a fully Bayesian approach for the survival analysis of an automotive vehicle fleet using real workshop-service data as input. It explores a problem instance containing more than 170 000 individual vehicles driving in 100 different countries exhibiting a certain failure, pre-selected for this study.

The suggested fleet surival analysis consists of a combination of two probabilistic models. The first model predicts the *mileage* of each individual of the fleet for a given point in time in the future. The second model attempts to forecast the *total number of failures* that will arise for the entire fleet.

Both probabilistic models are fully Bayesian, i.e., all parameters of the models are implemented as probability distributions and computations are solely performed on distributions rather than on summarizing statistics. As a consequence, uncertainty of the predictions is made accessible in a very natural way and can be taken into account in the decision-making process explicitly.

1. INTRODUCTION

In automotive industry, every car manufacturer is repeatedly facing the problem of in-advance spare-part stock production. With every new model series, the question of how many spare parts to produce and store arises. A precise estimation of the necessary number of spare parts is required in order to process the upcoming component failures adequately. In case of underestimation, there is a risk that a car cannot be repaired at all, which amounts to a total loss. At the same time, any overestimation leads to significant additional costs for production and storage. Comparing both scenarios, it becomes obvious that the underlying risk function shows an asymmetric distribution. As a consequence for decision-making, not only the point estimate of total failures but also its expected distribution is of great importance.

In this paper we present a novel approach for the survival analysis of a vehicle fleet by combining two Bayesian models, in order to capture and process all inherent uncertainties of the data. The first model, $\mathcal{M}$, aims at predicting the total mileages for all vehicles of the fleet for a desired future point in time. Based on this estimation, the second model, $\mathcal{S}$, estimates the total number of losses or survived individuals, respectively. One advantage of a fully Bayesian approach is, that the uncertainties of the mileage estimation model can be directly taken into account by the model for lifetime estimation. A higher uncertainty in mileage estimation directly leads to a higher uncertainty in predicting the probability of survival.

The outline of the paper is as follows: Section 2 introduces the vehicle fleet dataset used for the analysis. The section also describes both probabilistic models, $\mathcal{M}$ and $\mathcal{S}$, and explains the underlying assumptions. Based on these models, Section 3 shows the outcome of the analysis of the investigated component failure and provides an interpretation of the findings. Finally, we draw a conclusion of the investigation in Section 4 and discuss advantages and limitations of the method.

2. METHODOLOGY

Bayesian methods have become a valuable tool in the field of survival analysis (Ibrahim, Chen, & Sinha, 2004), especially with the availability of probabilistic programming sampling methods (Kelter, 2020). Recent publications confirm their suitability, e.g. in medical science (Zhou, Hanson, & Zhang, 2020; Brilleman, Elci, Novik, & Wolfe, 2020) or engineering applications (Feng, Shengyi, & Dai, 2019).

The presented survival analysis makes use of these ideas with an adaptation to the automotive use case. It is based on two probabilistic models in series. The first model is a mileage estimator, $\mathcal{M}$, which is responsible for estimating the total mileage distribution of each individual in the vehicle fleet for any point in time in the future. The output distributions of $\mathcal{M}$ act as inputs to the second probabilistic model, the lifetime estimator $\mathcal{S}$. The lifetime estimator performs the actual survival analysis of the fleet based on the uncertain future mileage of each individual vehicle.

The parameter distributions of the models were fit to the data using Markov-Chain Monte-Carlo (MCMC), utilizing the Stan-implementation (Stan Development Team, 2021) of a No-U-Turn Sampler (NUTS) (Hoffman & Gelman, 2014).

2.1. Input Data

The input to the survival analysis is a data table $\mathcal{D}$ containing time-series data of vehicle states. Each row in $\mathcal{D}$ corresponds to a single data reading event. $\mathcal{D}$ comprises 4 columns: vehicle id i, time-stamp t, total mileage m, and a binary variable δ signifying the repair state of the vehicle. $\delta = 1$ indicates a repairing event and $\delta = 0$ healthy vehicles. Furthermore, let N be the number of vehicles in the fleet, i.e. $i \in [1 \dots N]$.

Note that $\mathcal{D}$ constitutes a censored data set, since only a fraction of vehicles suffered a failure during the study. In order to make predictions for the future, the survival analysis also requires a query time t_{future} which is provided by the user.

Also note that all data originate solely from workshop-service, i.e., no regular reporting from the vehicle fleet via cloud service was available. As a consequence, for the majority of vehicles in the fleet, the time-intervals between two subsequent data reading events are irregular, ranging from several days to even several years.

2.2. Mileage Estimator

In a first step, the Mileage Estimator, $\mathcal{M}$, learns the distribution of distance driven per day, $p(d_i)$, for each vehicle i. We assume that the total mileage of every vehicle in $\mathcal{D}$ is strictly increasing, i.e., $d_i > 0,\ \forall\, i \in \mathcal{D}[\text{id}]$, where $\mathcal{D}[\text{id}]$ denotes the vehicle-id column of $\mathcal{D}$. Accordingly, d_i should be generated by a non-negative distribution. Furthermore, we assume that, for each individual, the mean and variance of $p(d_i)$ is time-invariant and finite. Under these constraints, the maximum entropy distribution is the Log-Normal distribution, which we chose as likelihood function to keep our assumptions to a minimum,

$$d_i \quad \sim \quad \text{Log-Normal}(\mu_i, \sigma_i)\,. \tag{1}$$

The Log-Normal distribution has two parameters, its mean μ_i and variance σ_i^2.

We chose the prior distribution for μ_i to be weakly informative, where we assume that the great majority of all drivers on average drive between $1\ \frac{\text{km}}{\text{day}}$ and $500\ \frac{\text{km}}{\text{day}}$ and a value of approximately $25\ \frac{\text{km}}{\text{day}}$ has the highest probability. These assumptions can be expressed by a Normal prior distribution

$$\mu_i \quad \sim \quad \text{Normal}(\mu_\alpha, \sigma_\alpha) \tag{2}$$

with corresponding parameters μ_α and σ_α.

It is natural to expect a certain amount of variability σ_α in a vehicle's daily mileage. On the one hand, variability may be caused by minor fluctuations in traffic management or differences in personal driving behavior. On the other hand, there exist also events that cause great variability such as changing drivers (car sharing, car selling, family car, etc.) or varying major routes (avoiding traffic jams caused by larger construction sites, new job location, new partner, etc.). We conclude that a variability of zero is not very plausible in this setting, which excludes the Exponential distribution from our candidate list of maximum entropy prior distributions for standard deviation parameters. However, the standard deviation parameter σ_i must be positive to avoid multi-modality of its posterior distribution[1]. Remember that we assume the variability for a single vehicle is fixed and finite. Again, the Log-Normal distribution represents the maximum entropy distribution under these constraints. However, this time we utilize it as prior distribution of the variability parameter σ_i

$$\sigma_i \quad \sim \quad \text{Log-Normal}(\mu_\beta, \sigma_\beta)\,. \tag{3}$$

We chose the parameters of the prior distribution for σ_i such that it becomes weakly informative, where we assumed that an individual's variability in daily mileage performance, on average, ranges from $0.1\ \frac{\text{km}}{\text{day}}$ to $100\ \frac{\text{km}}{\text{day}}$ and a value of approximately $3\ \frac{\text{km}}{\text{day}}$ has the highest prior probability. The values of μ_β and σ_β were chosen accordingly.

As a final step, the model has to convert the daily mileage distribution to a distribution of total mileage driven by the vehicle at some point t in the future. This is accomplished by integration over time. Let t_i be the most recent timestamp available in $\mathcal{D}$ for the i-th vehicle. Now, if we take a look at the variables themselves, we get the relationship between total mileage and distance driven per day is

$$m_i(t) - m_i(t_i) \;=\; \int_{\tau=t_i}^{t} d_i \,\mathrm{d}\tau \;=\; d_i \cdot (t - t_i)\,. \tag{4}$$

From this follows that the integration over time simply corresponds to a coordinate transformation of the argument of the

[1] Both σ_i and $-\sigma_i$ would lead to the same variance of σ_i^2, and thus have the same compatibility to a given data set $\mathcal{D}$.

probability distribution

$$P\left(m_i(t) = x\right) \quad = \quad P\left(d_i = \frac{x - m_i(t_i)}{(t - t_i)}\right). \qquad (5)$$

Accordingly, in order to compute the probability distribution $p(m_i(t))$, it is sufficient to know $p(d_i)$ and the constants t_i, $m_i(t_i)$.

One of the most important use-cases of the mileage estimator is the estimation of the current fleet mileage given past data. Suppose the current time is t_0. Let t_i be the most recent timestamp for vehicle i available in $\mathcal{D}$. The fact that no failure was reported within the time interval $[t_i,\, t_0]$ bears valuable information for the survival analysis. Suppose, for vehicle i a model $\mathcal{M}_i$ was fit to the data $\mathcal{D}_i$, where $\mathcal{D}_i$ corresponds to all data points in $\mathcal{D}$ with $\mathcal{D}[\text{id}] = i$. Each of these models outputs samples from the posterior predictive distribution $p(d_i|\mathcal{D})$. Furthermore, by looking up the values of t_i and $m(t_i)$, we can compute an estimate for the total mileage of vehicle i at time t_0 for each sample drawn from $p(d_i|\mathcal{D})$. This process is equivalent to drawing samples from the distribution $p(m_i(t_0)|\mathcal{D})$. By inserting the tuple (i, t_0, one sample from $p(m_i(t_0)|\mathcal{D})$, $\delta = 0$) into $\mathcal{D}$ we have generated a sample from the distribution $p(\mathcal{D}_i(t_0)|\mathcal{D}_i(t_i))$.

2.3. Lifetime Estimator

In order to perform the lifetime estimation, the censored data set $\mathcal{D}$ has to be transformed using the product-limit estimator[2] (Cox & Oakes, 1998) first. The result of this transformation is a non-parametric estimate of the survival function, $\hat{S}(m)$, as a function of the total mileage m. Recall that the value of the survival function at a specific value of m is a probability, i.e., $0 \leq \hat{S}(m) \leq 1,\, \forall\, m$.

In the next step we need to fit the parameter distributions of a parametric model $S(m)$ to the non-parametric estimate $\hat{S}(m)$. For this regression problem, we chose a normal distribution as likelihood function

$$\hat{S}(m) \quad \sim \quad \text{Normal}(S(m), \sigma_e) \qquad (6)$$

where σ_e can be interpreted as a measure describing the residual model error between $S(m)$ and $\hat{S}(m)$. Due to the constraint $\sigma_e \geq 0$, the prior distribution over σ_e is chosen to be an Exponential distribution, the maximum entropy distribution under this constraint,

$$\sigma_e \quad \sim \quad \text{Exponential}(\lambda)\,, \qquad (7)$$

with weakly informative rate parameter $\lambda = 100$.

In our model we assume that the moment a specimen is produced, it is also assigned a certain failure mechanism $k = 1, \ldots, K$ with probability θ_k. This failure mechanism will cause the specimen to fail eventually. The lifetime distribution $p(l_k)$ for each of the K failure classes is modelled using a Weibull distribution

$$l_k \quad \sim \quad \text{Weibull}(\alpha_k, \sigma_k)\,. \qquad (8)$$

The prior distributions over the α_k's and σ_k's were chosen to accommodate specimen half-lifes ranging from 100 km to 1 000 000 km.

$$\alpha_k \quad = \quad \alpha_{\min} + \tilde{\alpha}_k \qquad (9)$$
$$\sigma_k \quad = \quad \sigma_{\min} + \tilde{\sigma}_k \qquad (10)$$
$$\tilde{\alpha}_k \quad \sim \quad \text{Exponential}(\lambda_\alpha) \qquad (11)$$
$$\tilde{\sigma}_k \quad \sim \quad \text{Exponential}(\lambda_\sigma) \qquad (12)$$

The complementary cumulative distribution function of the Weibull distribution is the corresponding survival function

$$S_k(m) \quad = \quad \exp\left(-\left(\frac{m}{\sigma_k}\right)^{\alpha_k}\right). \qquad (13)$$

As we need to consider all possible failure mechanisms simultaneously, the K different survival processes are combined

$$\mu(m) \quad = \quad \sum_{k=1}^{K} \theta_k \cdot \mu_k(m)\,. \qquad (14)$$

The θ_k's are the hidden states of the Weibull mixture model and are drawn from a Dirichlet distribution

$$[\theta_1, \ldots, \theta_k] \quad = \quad \texttt{sort}\{\theta\}\,. \qquad (15)$$
$$\theta \quad \sim \quad \text{Dirichlet}(\gamma) \qquad (16)$$

In order to mitigate the danger of a multi-modal posterior distribution, the mixing parameters must be sorted, i.e., $\theta_1 \leq \theta_2 \leq \ldots \leq \theta_K$. Since no prior knowledge on the composition of Weibull distributions was available, we chose γ to be a K-dimensional vector of ones. This prior assigns an equal probability to every K-simplex.

2.4. Failure Ratio

The survival function $S(m)$, which is part of the model $\mathcal{S}$, and the individual total mileage estimates m_i, which are obtained from the N models $\mathcal{M}_i$, can now be used to compute the failure ratio of the vehicle fleet.

Given that the i-th vehicle survived until time t, its probability of survival until some point in the future $t_{\text{future}} > t$ is

$$\pi_i(t_{\text{future}}|t) \quad = \quad 1 - S(m_i(t)) + S(m_i(t_{\text{future}}))\,. \qquad (17)$$

Accordingly, a predicted failure ratio ϕ for the entire vehicle fleet can be computed by summing up the probabilities of

[2] Also known as Kaplan-Meier estimator.

survival for each vehicle

$$\phi(t_{\text{future}}|t) \;=\; \frac{\hat{N}_{\text{failures}}(t_{\text{future}}|t)}{N} \;=\; \sum_{i=1}^{N} \pi_i(t_{\text{future}}|t)\,, \quad (18)$$

where $\hat{N}_{\text{failures}}(t_{\text{future}}|t)$ is the estimated number of failures at time t_{future} given the health state of the fleet at time t.

3. RESULTS

The results presented in this section were obtained from analyzing a data set $\mathcal{D}$ containing more than $745\,000$ data points obtained from more than $170\,000$ different vehicles from 100 different countries. All vehicles were of identical brand and type. Furthermore, the vehicles were produced in 2014/Q1, i.e., from January, 1st 2014 to March, 31st 2014. It took, however, until the beginning of 2015/Q1 for all produced vehicles to be sold and used in the field. The survival analysis takes into account failures of a particular engine component. If that particular component did not break during the study, the vehicle is labeled as *survived*. On the other hand, if a vehicle experienced a failure of the component of interest, we assume that it was repaired immediately. Thus, starting its new lifecycle from $m = 0\,\text{km}$.

To validate our method, we simulated the successive accumulation of data over time. To be more precise, let $\mathcal{D}(t_0) \subseteq \mathcal{D}$ be the data that was available at timestamp t_0, i.e., the timestamp of each data point contained in $\mathcal{D}(t_0)$ is less or equal to t_0. In this paper we show results for training data available at timestamps $t_0 \in \{2015/\text{Q2}, 2016/\text{Q2}, 2017/\text{Q1}, 2017/\text{Q2}\}$, each instance predicting the *future* at $t_{\text{future}} = 2020/\text{Q4}$. I.e., while the ground truth is available for 2020/Q4 and is used as validation, for model training only data readings up to one of the above-mentioned time-stamps have been used. This reflects a typical real-world scenario, where such models are usually fed with the latest gathered information.

3.1. Mileage Estimator Results

3.1.1. Estimation of distance driven per day, d

The mileage estimation models aim at the prediction of each vehicle's mileage for a specific point in time. To gain an exemplary insight into the data, Fig. 1 shows four different data readings $\mathcal{D}_i$ and their corresponding posterior predictive distribution (PPD). Whereas some vehicles, like $i = 141417$, exhibit a very repeatable pattern of daily driving, others, e.g., $i = 000055$, show high fluctuation in daily driving. As explained in Section 2.2 this may originate from the individual use cases, e.g., private car vs. car sharing. Note that the prior assumptions of (2) lead to a higher probability around the value of $25\ \frac{\text{km}}{\text{day}}$, also for vehicle $i = 000055$. However, in case of strong evidence of deviation from prior assumptions, significantly different posterior predictive distributions, like for vehicle $i = 092381$, are nevertheless compatible with the model.

3.1.2. Prediction of total mileage, m

Fig. 2 shows the distribution of total mileage of the fleet (left) and the distribution of number of years to forecast (right). The left side of Fig. 2 depicts three different distributions.

- The **gray shaded area** (■) corresponds to the available data readings so far.
- The **black solid line** $(-)$ shows the prediction of the propagated future fleet mileage distribution.
- The **black dashed line** (- -) marks the ground truth.

In order to demonstrate the increase of model quality with new data, we predict the final data readings from 2020/Q4 in every iteration. Starting from the available data in 2015/Q1 in Fig. 2 (top), we loop over new data readings up to 2017/Q2 in Fig. 2 (4th row). This means whereas in 2015/Q1 (top) we predict over a horizon of 5 years with a very limited amount of data, in 2017/Q2 (4th row) we predict the fleet mileage over 3.5 years into the future.

We can see that for $t_0 = 2015/\text{Q2}$, Fig. 2 (top), there are no vehicles present with a higher mileage than $50\,000\,\text{km}$. This is no surprise, as the total time span of driving for all considered vehicles is only up to 12 months in this plot. At the same time, the predicted fleet mileage is already surprisingly accurate. Although there is still a difference in the mode of the posterior predictive distribution and the distribution of the future final readings, the overall shapes of both distributions already looks very similar. Note that this result is obtained from a forecast five years into the future.

One year later, for $t_0 = 2016/\text{Q2}$ (2nd row), the prediction of the total mileage distribution has improved greatly (left), while a larger fraction of the fleet did not report any new data (right), i.e., a large portion of probability mass is still around the value of 6 years to forecast.

For $t_0 = 2017/\text{Q2}$ (3rd row) and $t_0 = 2017/\text{Q2}$ (4th row), the predicted mileage distribution and ground truth are almost identical (left). The wide distribution of years to forecast illustrate the strong irregularity of the time intervals between data reporting events (right). Note that there is a distinct peak of the years-to-forecast distribution, Fig. 2 (right), hovering over the 6-years mark. This means that some car owners did not carry their vehicle to a service inspection until at least 2017/Q2.

In what follows, we will assume that each vehicle will be used by its driver in the future according to distribution of distance driven per day seen so far. Thus, if a driver changes his or her behavior over time, the distance driven per day distribution will become wider, thereby expressing the greater variability in this user's behavior.

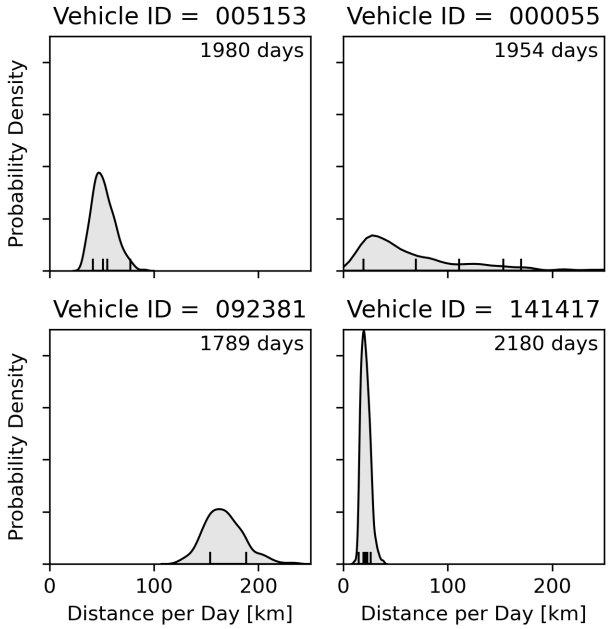

Figure 1. Posterior predictive distributions $p(d_i|\mathcal{D})$ (gray, filled) and actual data readings (black ticks) for four different vehicles. The time interval the data readings correspond to are shown in the upper right of each figure.

3.2. Lifetime Estimator Results

3.2.1. Description

The result obtained from the lifetime estimator are shown in Fig. 3.

Each row of Fig. 3 corresponds to a different training data set $\mathcal{D}(t_0)$ with $t_0 \in \{2015/\text{Q2}, 2016/\text{Q2}, 2017/\text{Q2}, 2017/\text{Q2}\}$, thus simulating the accumulation of more and more training data over time. The model is validated via comparison of the results obtained from the complete dataset $\mathcal{D} = \mathcal{D}(t_0 = 2020/\text{Q4})$

The **left column** shows the development of the posterior prediction of the quantity $\log(F(m)) = \log(1 - S(m))$, which is simply the logarithm of the cumulative lifetime probability distribution.

The **solid thick black line** $(-)$ is the median of the posterior predictive distribution

$$p(\log(F(m))|\mathcal{D}(t_0))\,, \tag{19}$$

where only data up to timestamp t_0 has been taken into account to train the model. The gray shaded area (■) between the thin solid black lines correspond to the 99%-compatibility interval[3] of (19).

The **black dots** $(\bullet)$ correspond to the values obtained from the non-parametric product limit estimator

$$\log(\hat{F}(m)) = \log(1 - \hat{S}(m)) \tag{20}$$

evaluated for different samples drawn from the posterior predictive distribution $p(m|\mathcal{D}(t_0))$.

The **black dashed line** (- -) corresponds to the median of

$$p(\log(F(m))|\mathcal{D}(t_{\text{future}})) \tag{21}$$

at $t_{\text{future}} = 2020/\text{Q4}$, i.e., when all available data was taken into account. Again, the gray shaded area (■) between the black dashed lines corresponds to the 99%-compatibility interval of (21).

The **right column** compares the actual and predicted values of the fleet failure ratio ϕ.

The **black dashed line** (- -) corresponds to the actual failure ratio $\phi_{\text{true}} \approx 0.18$ of the data set $\mathcal{D} = \mathcal{D}(t_0 = 2020/\text{Q4})$.

The **gray shaded area** (■) shows the posterior predictive distribution of the failure ratio

$$p(\phi(t_{\text{future}})|\mathcal{D}(t_0))\,. \tag{22}$$

3.2.2. Discussion

At $t_0 = 2015/\text{Q2}$ (top), the posterior predictive distribution of the failure ratio (right) is very flat. This corresponds to a great amount of uncertainty within the prediction. The cause of this uncertainty can easily be derived. Note that at $t_0 = 2015/\text{Q2}$ the majority of the fleet drove less than $40\,000$ km. The mileage estimator, however, expects the majority of the fleet to have driven more than $40\,000$ km at $t_{\text{future}} = 2020/\text{Q4}$. Accordingly, there is no data available that describes how well the fleet is doing for total mileages above $40\,000$ km. In consequence, the lifetime estimator $\mathcal{S}$ produced an extremely flat PPD of the survival function (left). The uncertainty in the survival function is then transported to the failure ratio (right).

One year later, at $t_0 = 2016/\text{Q2}$ (2^{nd} row), the situation looks very different already. Some vehicles drove for more than $75\,000$ km and the PPD of the failure ratio (right) becomes more concentrated around the (unknown) true value of $\phi(t_{\text{future}}) \approx 0.18$. However, the PPD of the failure ratio is bimodal which suggest that two different settings of parameter distributions within $\mathcal{S}$ are *fighting* for prevalence.

Nine month later, at $t_0 = 2017/\text{Q1}$ (3^{rd} row), the most active drivers have reached approximately $100\,000$ km. Note that the models still have to forecast four to six years of driving behavior for the majority of the fleet. Both PPDs have narrowed substantially, and we get our first reliable estimate of the survival function and prediction of the fleet failure rate. The lifetime estimator $\mathcal{S}$ now assigns great probability to a Weibull mixture model corresponding to $K = 2$ different failure mechanisms present in the fleet. Also, the PPD of the fail-

[3] We use the term *compatibility interval*, which was introduced by Richard McElreath (McElreath, 2020), to describe the compatibility of the model and the data.

ure ratio has become uni-modal, developing a distinct peak (around the true value!). If one had to decide how many spare parts were required within the upcoming 45 months in a probable worst-case scenario, the survival analysis presented here suggests the following answer: $\approx 0.22 \cdot N$ or about 37 000 parts necessary until 2017/Q4 for the entire fleet, compare to Fig. 3 (right, 3rd row).

The results show that a reliable estimate of the fleet survival function S and thus a reliable prediction of the fleet failure ratio ϕ is only possible after having gathered a sufficient amount of data. For the case regarded in this paper we required to collect data from January 2014 to March 2017 (27 months) in order to make reliable predictions for December 2020, i.e., forecasting a total of 45 months. However, at each moment in time the amount of uncertainty within the posterior predictive distributions correctly reflected the level of confidence one could have in the predictions.

However, there is still some residual uncertainty left. There are two mechanisms that make the PPD more narrow, i.e., more confident, with increasing timestamp of the snapshots shown in Figures 2 and 3. *First*, short prediction horizons lead to more confident predictions. As time goes on, the number of days until the fixed future date 2020/Q4 is reached gets smaller. Now, when simulating the future behavior of the fleet, the uncertainty about its state can only increase with increasing simulation time. Thus, for small prediction horizons there are only few opportunities for the uncertainty to inflate. *Second*, rich data sets, especially ones containing many high-mileage vehicles, cf. Fig.2 (left), lead to more confident predictions. If there is no historical data available for a certain (high-value) mileage interval, the model must extrapolate when making predictions about such a region. In this case, the model can only rely on the expert knowledge that was incorporated into it via the prior-distributions. If the relevant priors are not informative, the extrapolation is accompanied by an excessive growth of uncertainty. This behavior is typical for Bayesian model, and it is a desirable behavior, since the user is always informed on in how far the model's predictions can be trusted.

4. Conclusion

Survival analysis of mechanical or electrical components is a common tasks in engineering and it bears a huge potential for cost reduction as well as a high risk for cost increase in the decision-making process. To overcome the impacts of asymmetric cost functions, the use of Bayesian models has been suggested as a fundamental step to process the inherent uncertainties. The advantages of this approach have been demonstrated for an automotive use case with a vehicle fleet of over 170 000 cars and a specific component failure of interest.

In the presented scenario, a set of two combined Bayesian models has been used. The first model estimates the mileage of each individual car for a desired time point in terms of its posterior density. This estimate is used for modeling the predictive posterior distribution of the survival rate. It has been shown, that with each new information the posterior distribution becomes more narrow and finally converges to the real value (provided by the data of 2020/Q4.)

By analyzing the PPD of the survival rate, the decision-making process can be supported in an optimal way, as all the possible values along with their probabilities are available. The knowledge about the uncertainty of one's predictions is an indispensable attributre of the presented approach. Compared to a point estimate it gives way more insight into the underlying data generation process and helps to predict the expected failure rates more realistically.

5. Appendix

5.1. The Log-Normal Distribution

The Log-Normal distribution is defined over $y \in \mathbb{R}^+$ as

$$\text{Log-Normal}(y|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\,\sigma}\frac{1}{y}\exp\left(-\left(\frac{\log(y)-\mu}{\sqrt{2}\sigma}\right)^2\right), \quad (23)$$

with $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$.

5.2. The Dirichlet Distribution

For $K \in \mathbb{N}$, the Dirichlet distribution is defined as

$$\text{Dirichlet}(\theta|\alpha) = \frac{\Gamma\left(\sum_{k=1}^{K}\gamma_k\right)}{\prod_{k=1}^{K}\Gamma(\gamma_k)}\prod_{k=1}^{K}\theta^{\gamma_k-1}, \quad (24)$$

where $\gamma \in (\mathbb{R}^+)^K$ is a vector of shape parameters and $\theta \in K$-simplex is a vector of probabilities. The Γ-function is defined as

$$\Gamma(z) = \int_0^\infty x^{z-1}\exp(-x)\,\mathrm{d}x\,. \quad (25)$$

5.3. The Weibull Distribution

The Weibull distribution is defined over $y \in [0,\infty)$ as

$$\text{Weibull}(y|\alpha,\sigma) = \frac{\alpha}{\sigma}\left(\frac{y}{\sigma}\right)^{\alpha-1}\exp\left(-\left(\frac{y}{\sigma}\right)^{\alpha}\right), \quad (26)$$

where $\alpha \in \mathbb{R}^+$ is a shape parameter and $\sigma \in \mathbb{R}^+$ is a scale parameter.

5.4. The Exponential Distribution

The Exponential distribution is defined over $y \in \mathbb{R}^+$ as

$$\text{Exponential}(y|\lambda) \quad = \quad \lambda \exp\left(-\lambda \cdot y\right) , \qquad (27)$$

where $\lambda \in \mathbb{R}^+$ is the rate parameter.

REFERENCES

Brilleman, S. L., Elci, E. M., Novik, J. B., & Wolfe, R. (2020). *Bayesian Survival Analysis Using the rstanarm R Package.*

Cox, D., & Oakes, D. (1998). *Analysis of survival data*. CRC Press.

Feng, Q., Shengyi, S., & Dai, L. (2019). Bayesian survival analysis model for girth weld failure prediction. *Applied Sciences*, *9*, 1150.

Hoffman, M. D., & Gelman, A. (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, *15*, 1593-1623.

Ibrahim, J., Chen, M., & Sinha, D. (2004). Bayesian Survival Analysis..

Kelter, R. (2020). Bayesian Survival Analysis in STAN for Improved Measuring of Uncertainty in Parameter Estimates. *Measurement: Interdisciplinary Research and Perspectives*, *18*, 101 - 109.

McElreath, R. (2020). *Statistical Rethinking, 2nd edition*. CRC Press.

Stan Development Team. (2021). *Stan reference manual, version 2.26.*

Zhou, H., Hanson, T., & Zhang, J. (2020). spBayesSurv: Fitting Bayesian Spatial Survival Models Using R. *Journal of Statistical Software, Articles*, *92*(9), 1–33. Retrieved from `https://www.jstatsoft.org/v092/i09` doi: 10.18637/jss.v092.i09

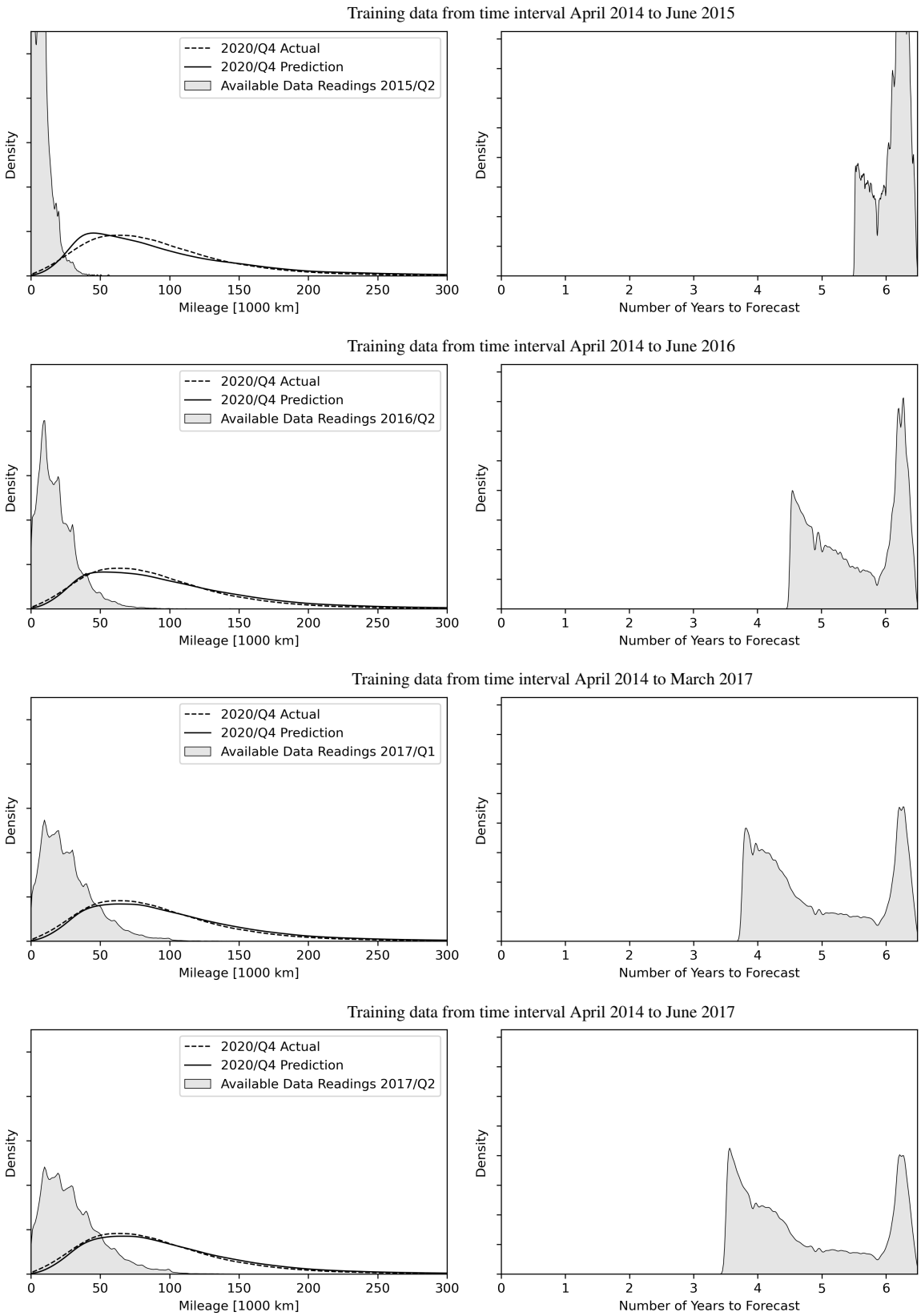

Figure 2. Development of mileage estimation result over time. Left: distribution of total mileage data readings (gray, filled) in $\mathcal{D}$ as well as the predicted (black, solid) and actual (black, dashed) total mileage distribution of the vehicle fleet over total mileage m for $t_{\text{future}} = 2020/\text{Q4}$ and different $t_0 = [2015/\text{Q2}, 2016/\text{Q2}, 2017/\text{Q1}, 2017/\text{Q2}]$ (rows). Right: distribution of years to forecast, giving an impression of the irregularity of data reporting events.

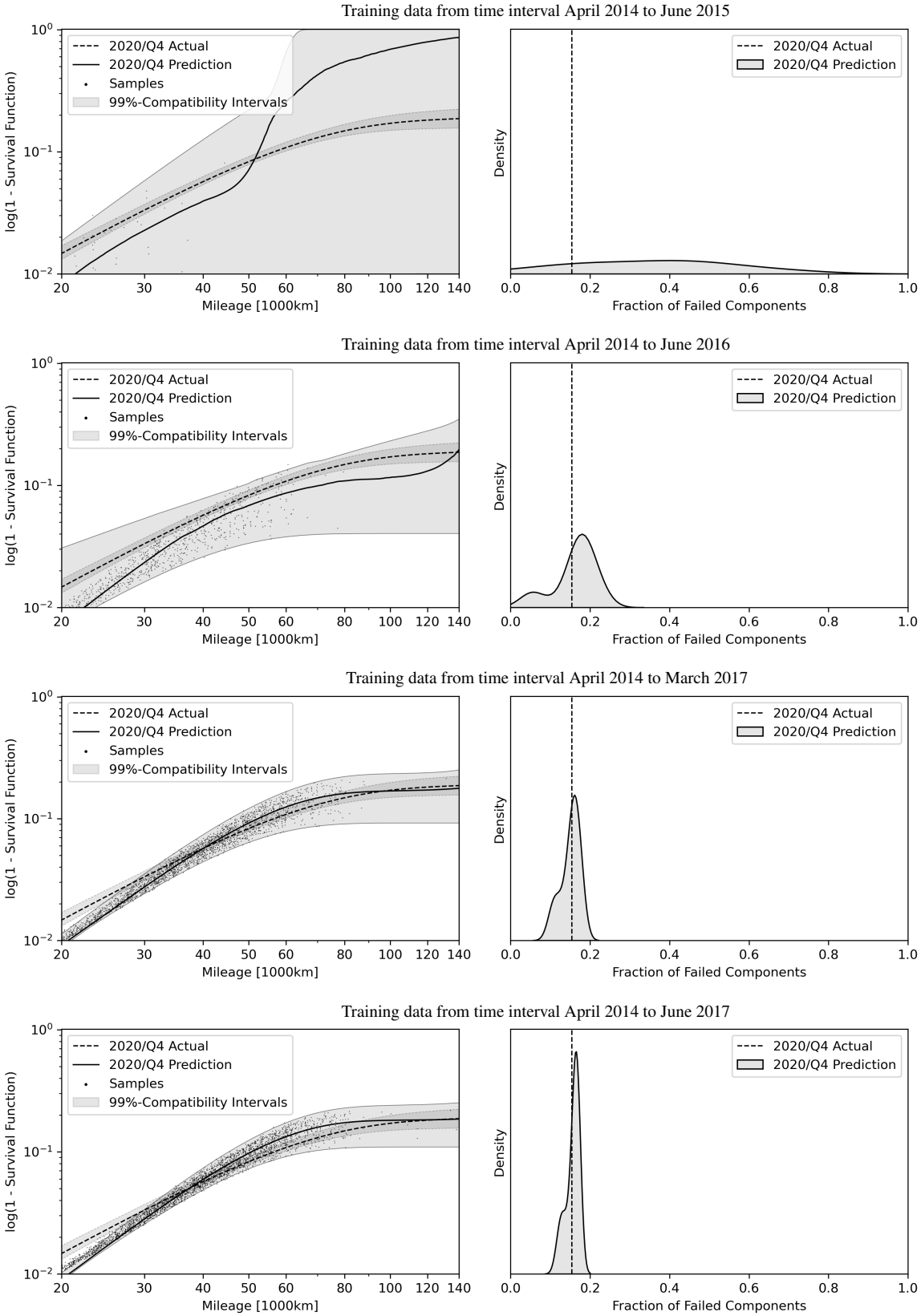

Figure 3. Development of survival analysis result over time. Left: logarithm of predicted (black, solid) and actual (black, dashed) cumulative lifetime probability distribution $\log(1 - \mu(m))$ over total mileage m for $t_{\text{future}} = 2020/\text{Q4}$ and $t_0 = [2015/\text{Q2}, 2016/\text{Q2}, 2017/\text{Q1}, 2017/\text{Q2}]$. Right: posterior predictive distribution of fraction of failures (gray, filled) and the actual value (black, dashed) at $t_{\text{future}} = 2020/\text{Q4}$.