Analysing cattle mortality in Northern Ireland with survival analysis and machine learning

Maria (Marietta) Dalamanga

***Abstract*—In this report we present the results of a large scale survival analysis of cattle mortality from across Northern Ireland exploring several risk factors including the sex, geographic area, production type, place of death, and health status. To this end, we used several parametric, nonparametric, and semiparametric methods: Kaplan-Meier estimators, the accelerated failure time and Cox methods, fitting parametric distributions—including mixture distributions—survival forests, and methods from machine learning such as random forests, extreme boosted trees and regression trees. We applied these methods to different subpopulations of the dataset. Best results were obtained using a Cox model with shared frailty term on the herd ID (C-index: 73.52%), and a survival forest on the dairy subpopulation (75.11%). We were also able to predict the life span of male animals using a random forest (RMSE 153.4 days, $R^2 = 0.65$). The sex and certain abattoirs proved to be the main risk factors for cattle survival.**

***Index Terms*—Survival analysis, Cattle mortality, Machine learning.**

I. Introduction

Monitoring and analysing the cattle mortality can have significant impact in reducing the cost and the environmental footprint of the beef and dairy industry. Moreover, understanding the factors that influence the longevity and mortality of cattle will lead to the adoption of new, more effective, management practices, which can improve both the animal welfare and public health by introducing control programs and timely responses to epidemics (Rasmussen et al. 2024, Kulkarni et al. 2021).

The key objectives of this work are (i) to explore the distributions of the life span of cattle, and (ii) identify key factors that affect them.

In our analysis we use a large dataset of mortality data of bovine animals ($N = 2,088,225$) from across Northern Ireland over four years. We use both traditional survival analysis and modern machine learning to identify possible risk factors and understand how they affect animal mortality. The results of this study can be a useful to policy-makers to enable informed decisions.

A. Mortality factors

Many factors associated with cattle mortality have been explored in the published literature. For Santman-Berends et al. (2014) the farmers' mindset and managing practices are crucial factors influencing cattle mortality. The intensification of the production (Nørgaard et al. 1999), unassisted calving (Fruscalso et al. 2020), diet (Urie et al. 2018), herd size and housing (Sarjokari et al. 2018, Reimus et al. 2020), accidents, diseases (Alvåsen et al. 2014), and the weight at birth (Urie et al. 2018) have been found to be key reasons of on-farm deaths. Among the possible environmental factors, temperature, humidity, and ventilation seem to be most important (Zablotski et al. 2024, Wisnieski et al. 2022, Morignat et al. 2015). Lastly, genetic factors, the breed and the sex of the animal have also been explored (Mõtus et al. 2018, Caraviello et al. 2003, van der Heide, Veerkamp, van Pelt, Kamphuis & Ducro 2020)

B. Taxonomy of survival analysis methods

Survival analysis is a field of statistics studying the time until an event occurs and how the distribution of such times-to-event is affected by observed and unobserved parameters (Clark et al. 2003). Following Wang et al. (2019), survival analysis methods fall into four large categories: non-parametric, semi-parametric, parametric, and machine learning.

1) Non-parametric models: Non-parametric methods seek to estimate the survival function from data without imposing assumptions on the underlying distributions. In non-parametric methods the most popular approach is that of traditional exploratory analysis (Thomsen et al. 2004) with the Kaplan-Meier (KM) estimator (Santman-Berends et al. 2018, Ciarmiello et al. 2023, Probo et al. 2018) and log-rank tests (George et al. 2014, Singh & Mukhopadhyay 2011) to compare survival curves. Another estimator is the Nelson-Aalen, which is an estimator of the cumulative hazard function (Aalen 1978, Pölsterl 2016, Colosimo et al. 2002). The KM estimates can be summarised in so-called Life Tables, which present mortality and life expectancy data (Wang et al. 2019). However, with non-parametric methods it is hard to reveal how exactly risk factors may affect the survival, especially in high-dimensional cases.

2) Semi-parametric models: Semi-parametric survival models seek to reveal the effect of risk factors on the shape of the survival function. The prime representative of this class of methods are the celebrated Cox proportional hazards (PH) models (Bradburn et al. 2003). A statistical test for verifying the PH assumption is the Schoenfeld residuals test (Borucka 2014)[1]. The Cox method has been extensively used in the literature for the survival analysis of farm animal mortality.

M. Dalamanga is a student of the MSc in Data Analytics, School of Mathematics and Physics, Queens University Belfast, Main Physics building, University Road, Belfast BT7 1NN, Student number: 40392473, Email: mdalamanga01@qub.ac.uk.

[1] In R available from the `survival` package, function `cox.zph`.

For instance, Renaud et al. (2018) identify a number of risk factors, including the barn location within the farm, season of arrival of the animal, and dehydration using Cox models, however without convincing evidence that the PH assumption is satisfied. Mõtus et al. (2018) by applying a Cox model identified metabolic/digestive disorders, traumas and accidents as key risk factors in beef youngstock, but no strong evidence is presented that the PH assumptions are satisfied. Probo et al. (2018) state that their Cox models are "highly likely to violate the PH assumption." This "inconvenience" was already highlighted by Bugnard et al. (1994). Although the PH assumption seems to be difficult to be satisfied in practice, the Cox model offers valuable insights about the effect of risk factors on the hazard function.

To account for unobserved effects, the framework of Cox models with *frailty* has been proposed. Both Alvåsen et al. (2014) and Reimus et al. (2020) use a Weibull parametric model and a gamma-distributed random effect frailty model of the herd to identify risk factors in dairy cows. Likewise, the work of Caraviello et al. (2003) uses a Cox-like model with a Weibull baseline hazard function and (time-dependent) coefficients to identify a number of risk factors.

3) Parametric models: The accelerated failure time (AFT) modelling approach consists in expressing the survival function as a *scaled* version of a (parametric) baseline survival, where the scaling factor is of the form $\exp(-\beta^\intercal x)$, where $x \in \mathbb{R}^n$ is vector of risk factors. AFT has been used in survival analysis studies extensively (Collett 2003, George et al. 2014). In a recent paper, Crowther et al. (2022) generalise the standard AFT formulation by composing the baseline survival function with a time-dependent cubic spline.

Mixture models are parametric models used to reveal the presence of sub-populations in the data (Manouchehri & Bouguila 2020). Mixture distributions have been being used in survival analysis since the 80's (Farewell 1982). In Razali & Al-Wakeel (2013) the authors propose a parameter estimation method, based on maximum likelihood, for mixtures of two and three Weibulls; these appear to be flexible enough for the type of distributions often encountered in survival analysis; see e.g., Marín et al. (2005). Likewise, log-logistic distributions (Al-Shomrani et al. 2016) have been shown to be able to model multimodal survival data effectively.

4) Machine learning: A wide range of machine learning methods have been used in survival analysis: from survival forests to deep learning methods (Wiegrebe et al. 2024). In particular, the methods that seem to be most popular are ensemble-type methods—including random forests (Rezaei et al. 2020, Ishwaran et al. 2008) and gradient boosted trees (Bai et al. 2021)— regression methods such as logistic regression (George et al. 2014) (for binary or censored data), deep neural networks (Zhao & Feng 2020), and hierarchical Bayesian methods (Bellot & van der Schaar 2019), to name a few.

As far as cattle mortality is concerned, random forests have been widely used by several authors (McBride et al. 2022, Stański et al. 2021, van der Heide, Kamphuis, Veerkamp, Athanasiadis, Azzopardi, van Pelt & Ducro 2020, Zablotski et al. 2024, Wisnieski et al. 2022) owing to their ability to reveal how the risk factors affect the survival times, especially with discrete risk factors.

Several regression methods have also been used. In the area of cattle mortality, Zablotski et al. (2024) use logistic regression to study the effect of risk factors at the level of individual animals. Notably, van der Heide, Veerkamp, van Pelt, Kamphuis & Ducro (2020) leverage both genomic and phenotypic information and use logistic regression to predict individual survival times. Likewise, logistic regression is used by Fruscalso et al. (2020) to inform policymaking based on how management practices appeared to affect animal mortality.

A general observation is that machine learning methods seem able to capture complex relationships between risk factors and survival-related targets and handle big datasets.

C. *Software packages for survival analysis*

Several software packages are available for survival analysis in R (R Core Team 2021), Python and other languages and frameworks.

In R, the `survival` library (Therneau & M. 2000) is very commonly used and allows to estimate KM, Cox, AFT and other models, plot survival curves (e.g., using `survminer` (A. Kassambra *et al.* 2021)), and compute the accompanying statistics (Therneau & M. 2000). The `caret` (Kuhn & Max 2008) and `randomForestSRC` (Ishwaran et al. 2008) are used for machine learning and survival forest models.

Python also comes with a large ecosystem of packages for survival analysis such as `scikit-survival` (Pölsterl 2020), `pysurvival` (Fotso et al. 2019), `lifelines` (Davidson-Pilon 2019), and `reliability` for mixture models (Reid 2020).

II. Methodology

A. *Introductory definitions*

Let T be the random life span of an individual. The cumulative distribution function (cdf) of T is defined as $F(t) = \mathsf{P}[T \leq t]$. The survival function, S, is defined as $S(t) = \mathsf{P}[T > t] = 1 - F(t)$. The hazard function describes the probability that a death occurs within the interval $[t, t+\delta t]$ provided the individual survives up to time t, and taking the limit $\delta t \to 0$ (Collett 2003). Concretely,

$$h(t) = \lim_{\delta t \to 0} \frac{\mathsf{P}[t \leq T < t + \delta t \mid T \geq t]}{\delta t} = -\frac{S'(t)}{S(t)}, \tag{1}$$

provided S is differentiable. The derivation can be found in Appendix A.

From Equation (1) we have $h(t) = -\frac{\mathrm{d}}{\mathrm{d}t} \log S(t)$, therefore, $S(t) = \exp\big(-\int_0^t h(\tau)\mathrm{d}\tau\big)$, which leads to the definition of the cumulative hazard function $H(t) = \int_0^t h(\tau)\mathrm{d}\tau$, thus $S(t) = \exp(-H(t))$. Next, for two random variables T and T' with survival functions S_T and $S_{T'}$ a distance between T and T' can be defined as $d_{\mathrm{KS}}(T, T') = \max_{t \geq 0} |S_T(t) - S_{T'}(t)|$; this is the Kolmogorov-Smirnov (KS) distance used in the Kolmogorov-Smirnov test of equality of probability distributions.

In survival analysis, we are typically presented with data (x_i, t_i, s_i), for $i = 1, \ldots, N$, where $x_i \in \mathbb{R}^n$ are the risk factors and t_i are life durations (or time-to-event or time-to-failure) whenever the *status* is $s_i = 0$ and lower-bounds of such times when $s_i = 1$. To put it simply, if we have observed that the animal i has died, then $s_i = 0$ and t_i is its life span. If animal i is not dead and, up to the time of the measurement, has lived t_i days, its life span will be at least t_i and $s_i = 1$. Data points with $s_i = 1$ are called censored. In this study all data are uncensored.

Harrell's concordance index—or C-index—generalises the notion of the area under the ROC curve. The C-index's definition is based on the notion of concordant pairs. Suppose $t_i < t_j$ are two observed failure times and $\hat{t}_i, \hat{t}_j$ are the corresponding predictions of a model. We say that (t_i, t_j) and $(\hat{t}_i, \hat{t}_j)$ are *concordant* if $\hat{t}_i < \hat{t}_j$; otherwise, they are called *discordant*. In the case of uncensored data, the C-index is defined as the percentage of concordant pairs among all pairs (Longato et al. 2020). According to Hartman et al. (2023), it is widely accepted that C-index values above 70% are acceptable, although in animal survival studies significantly lower concordance values are typically reported (Ellington et al. 2020).

B. The dataset

Based on mortality data from the period 2018 to 2021, we perform a traditional exploratory data analysis to gain insights about the data at hand. Details can be found in Appendix E. Our dataset consists of 2,088,225 observations with 15 possible risk factors. As outliers we consider the data that lie above the third quartile plus 3 IQR (10 years), so 82,764 animals are excluded, which corresponds approx. to 4% of the dataset.

The resulting dataset consists of 2,005,461 unique animals and counts 20,415 unique herd IDs, 17 abattoir IDs, 92 animal breeds, 3 sex types (female, male and bull), 2 places of death (abattoir or farm), 2 production types (beef or dairy), 10 divisional veterinary office (DVO) area codes, the date of birth and date of death for each animal, the number of health conditions per animal at the postmortem examination (0 to 9), the average temperature for the month when the animal died, the average humidity-temperature ratio, average herd size, and any of 72 possible health conditions discovered by AI at postmortem examination. In this work we define as *healthy* the animals that have no health conditions.

For our exploratory analysis and data manipulation we used R (R Core Team 2021) specifically the packages: `tidyverse` (Wickham et al. 2019), `dplyr` (Wickham et al. 2023), `plyr` (Wickham 2011), `tidyr` (Wickham et al. 2024), and, for visualisations, `ggplot2` (Wickham 2016).

C. Kaplan-Meier

The KM method is a non parametric estimation method for the survival function. It is widely used in medical applications and survival analysis especially for its simplicity to estimate the survival function in the presence of censored data (Kishore et al. 2010).

The simplest way to estimate the probability that the lifespan, T, of an individual from a population is more than t is by dividing the number of subjects who lived more than t by the total number of subjects. This gives rise to the KM estimator of the survival function. Given a dataset with failures at discrete times $t_1 < \ldots < t_m$ and $d_1, \ldots, d_m$ being the corresponding number of failures, the KM estimator is commonly written as (Pölsterl 2016)

$$\widehat{S}(t) = \begin{cases} 1, & \text{if } t < t_1 \\ \prod_{\{i:t_i \leq t\}} \frac{n_i - d_i}{n_i}, & \text{otherwise} \end{cases} \tag{2}$$

where n_i is the number of individuals that survived at times $t \geq t_i$. Details can be found in Appendix B.

It is often desirable to compare two estimated survival functions. This can be done using the log-rank test, which tests the alternative hypothesis that one hazard function is a multiple of the other, that is, $h_1 = ch_2$ with $c \neq 1$ (George et al. 2014, Singh & Mukhopadhyay 2011)

In R, the `survminer` (A. Kassambra *et al.* 2021) and `survival` (Therneau & M. 2000) packages were used for KM analyses and log-rank tests.

D. Fitting Parametric Distributions

A parametric distribution is a probability density function (pdf) of the form $p(t; \theta)$, where $\theta \in \mathbb{R}^n$ is a parameter vector. The objective is to estimate θ from data, and a standard way to do this is the maximum likelihood estimation method, which is based on the likelihood function. For details, see Appendix C. Once we have estimated a value of θ we can compute the KS distance between the empirical distribution and the estimated parametric distribution, $F(t; \theta)$, to assess the goodness of fit. In survival analysis, the Weibull, log-normal, exponential[2], and the gamma distributions are commonly used (Rodriguez 2010).

Often parametric distributions such as the Weibull, log-normal or gamma can prove inadequate to describe multimodal distributions that may arise from the presence of distinctly distributed subpopulations. Then, mixture models can offer more flexibility.

A mixture distribution is a parametric distribution whose pdf is a weighted average of pdfs from a standard parametric family. This way, we can have mixture of Gaussians, Weibull, etc. In other words, a mixture distribution has a pdf of the form

$$p(t) = \sum_{i=1}^{M} w_i p_i(t), \tag{3}$$

where each p_i is called a component of the mixture, and the weights, w_i, are nonnegative and sum up to 1.

E. Accelerated Failure Time Model

The idea behind the AFT model is that a risk factor affects the lifetime of the animals by "accelerating" (or decelerating) their life span. In other words, risk factors act by "stretching" the survival function.

Let us denote the survival function when risk factors $x \in \mathbb{R}^n$ are zero by S_0, that is, $S_0(t) = S(t \mid x = 0)$. Then, we assume that

$$S(t \mid x) = S_0(\exp(-\beta^\intercal x)t). \tag{4}$$

[2] The exponential distribution is a specific case of the Weibull

Typically, S_0 is assumed to follow a certain parametric distribution (e.g., log-logistic, Weibull, gamma, or other). A key requirement is that when S_0 is scaled, the resulting distribution, $S(t)$, is of the same type. (Collett 2003).

Our parametric analysis except for the mixture models is conducted in R (R Core Team 2021) using the `survival` (Therneau & M. 2000), `utils` (Bengtsson 2005) and `stats` (R Core Team 2021) packages. For mixture models we use `reliability` (Reid 2020), and `scikit-learn` (Pedregosa et al. 2011) in Python.

F. Cox proportional hazards model

1) Proportional hazards assumption: The Cox PH model is a semiparametric approach where no parametric form is assumed for the survival function (Foley 2022, Chap 9). Instead, it is assumed that the covariates enter the hazard function as follows

$$h(t) = h_0(t)\exp(\beta^\intercal x), \tag{5}$$

where $x \in \mathbb{R}^n$ is the vector of risk factors, $\beta \in \mathbb{R}^n$ is a constant vector of coefficients, and $h_0(t)$ is the so-called baseline hazard function. Breslow's estimation approach is commonly used to estimate the parameters, β, and the baseline hazard function (Lin 2007).

Note that for a δx-change of the covariates the hazard function becomes

$$\begin{aligned} h(t \mid x + \delta x) &= h_0(t)\exp(\beta^\intercal(x+\delta x)) \\ &= h_0(t)\exp(\beta^\intercal x)\exp(\beta^\intercal \delta x) \\ &= h(t \mid x)\exp(\beta^\intercal \delta x). \end{aligned} \tag{6}$$

Note that $\exp(\beta^\intercal \delta x)$ does not depend on time. Essentially, changing the covariates by δx scales the hazard function. This is why this model is referred to as the PH approach.

The *proportional hazards assumptions* can be summarised as follows (i) there is a common baseline hazard function for all individuals that corresponds to $x = 0$, (ii) the covariates, x, are independent of time, (iii) the coefficients, β, are independent of time.

To test whether the PH assumption holds the Schoenfeld test is commonly used (Therneau & Grambsch 2000). It is a test of uncorrelatedness of the residuals.

2) Models with frailty: Following Balan & Putter (2020), suppose that the proportional hazards model is valid using a vector of variables $x = (x_1, x_2)$, where x_1 is a vector of observed effects, while x_2 is unobserved. The Cox model reads

$$h(t) = h_0(t)\exp(\beta_1^\intercal x_1 + \beta_2^\intercal x_2). \tag{7}$$

We can write this equivalently as

$$h(t) = Z h_0(t)\exp(\beta_1^\intercal x_1), \tag{8}$$

where $Z = \exp(\beta_2^\intercal x_2)$ is the so-called *frailty term*, which does not depend on t and scales the baseline hazard function. The variable x_2 is known as *frailty*. Cox models with frailty introduce a random effects variable x_2 known as a *frailty* variable.

Often the data are naturally clustered into $i = 1, \ldots, K$ clusters. For example, in our case study, the animals are grouped into distinct herds, sexes, and so on. It then may make sense to identify Cox models with *shared* frailty terms within the same cluster. More precisely, the hazard function of an individual of the i-th cluster has the hazard function

$$h_i(t) = Z_i h_0(t)\exp(\beta^\intercal x), \tag{9}$$

where Z_i is the frailty term of the cluster i. For the frailty variable a parametric distribution is considered—commonly a Gamma distribution, which is the default in `coxph` in the `survival` package Therneau & M. (2000).

G. Survival forests

Survival forests are an application of the random forests methodology for the estimation of survival functions from—possibly censored— survival data, conditional on a number of variables (Ishwaran et al. 2008).

The estimation algorithm is as follows:

1) Randomly sample some data and keep aside the remaining "out-of-bag" data so we can later compute prediction errors.
2) For each sample, estimate a tree by randomly selecting for each node a subset of the variables. Split the tree according to a variable that maximises the difference between the survival functions of the respective subpopulations.
3) Keep growing the tree until it reaches a terminal node, provided that the individuals at that node are no fewer than a certain minimum number.
4) For each tree calculate the survival function (or cumulative hazard) and define the "ensemble" survival function to be the average of the (conditional) survival functions of the trees.
5) Calculate prediction errors using the out-of-bag data.

Survival forest models were trained using R's `randomForestSRC` library (Ishwaran et al. 2008).

H. Individual prediction models

We propose an approach that can make use of several regression methods from machine learning. This is facilitated by the lack of censored data in our case study. Here, instead of trying to model the survival function (or the hazard function) we aim at modelling the mortality of each individual animal as a function of known variables (e.g., sex, area, herd size, health status and more). Since the majority of variables are discrete, methods such as random forests (Cutler et al. 2012), and extreme boosted trees (Chen & Guestrin 2016) are mostly appropriate. Random forests and regression trees were trained with `caret` (Kuhn & Max 2008). For extreme boosted trees, `XGBoost` was used (Chen & Guestrin 2016).

To assess the quality of each model we use standard metric from machine learning such as the root mean squared error (RMSE) and the R^2 value of predicted-vs-actual values on a test set.

A connection can be made between the framework of individual prediction models and survival analysis; for details see Appendix D.

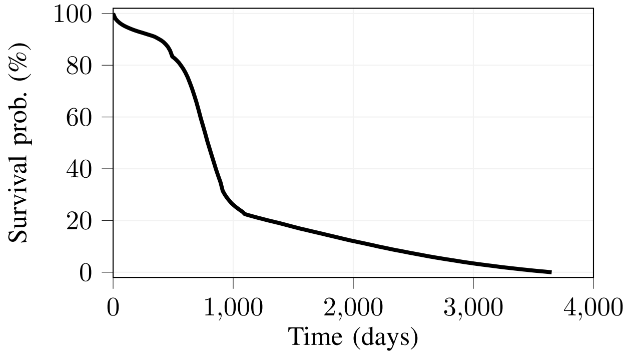


Fig. 1: Empirical (Kaplan-Meier) Survival of cattle mortality data.

Sex	all	% healthy	% unhealthy
Female	906	875	1311
Male	709	701	746
Bull	624	564	841

TABLE I: Kaplan-Meier median estimates by sex for all, healthy and unhealthy animals

I. Computing resources

All computations were performed on Dell EMC, 24 Intel® Xeon® Silver 4310 CPUs, and 96 GB RAM. The version of Python used was 3.11, and the version of R was . . . The source code is available on GitHub at https://github.com/marietta-d/cattle_mortality.

III. Results

In this section we present the results of the application of the aforementioned methodology.

A. Kaplan-Meier analysis

1) Estimated overall survival function: The overall survival curve, estimated with the KM method, is shown in Figure 1. Half of the animals live no less than 789 days (95% confidence interval: (788, 789)). We can identify three regions where the survival function has different slope, which is an indication of multimodality in the distribution of life span. We will address this in the following sections.

Next, we will look at the estimated survival curves of certain subpopulations of the data to see how the various variable affect the survival curves.

2) Sex: The KM-estimated survival curve for males, females, and bulls is presented in Figure 2. At every given time, t, female animals have the highest survival probability. The median for male animals is 624 days, for females is 906 days and for bulls 709 days. Based on the log-rank test, the survival distributions of sex are statistically significant (p-value $< 2 \cdot 10^{-16}$). All pairwise tests for the three sex categories are found to be statistically significant. The median life expectancies of healthy and unhealthy animals are shown in Table I.

3) Place of death: The impact of the place of death on the survival curve is illustrated in Figure 3. We see that the survival curves cross at 858 days: at younger ages the mortality is dominated by death in an abattoir; the opposite happens at later ages. Moreover, half of the animals who died on farm are less than 1 year old (95% confidence interval: (328, 322)), but animals who died on farm *and* survived more than 858 days have slightly longer life expectancies than animals who died on an abattoir. The place of death is statistically significant for the survival with a p-value < 0.0001 for the log-rank test. Only two animals that died on a farm were unhealthy, so it is difficult to understand the possible combined effect of both the place of death and health status.

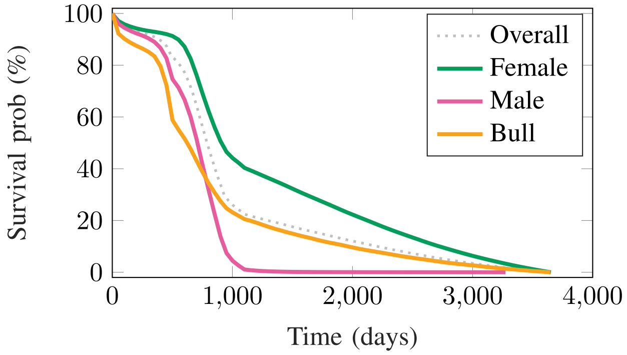


Fig. 2: Estimated survival curves for females, males, and bulls.

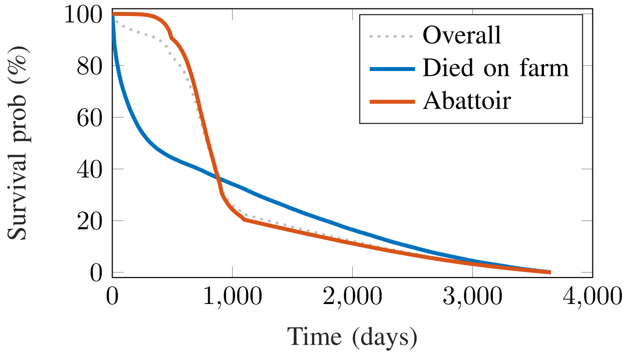


Fig. 3: Estimated survival curves for animals that died in an abattoir and ones that died on a farm.

4) Type of production: Figure 4 shows the estimated survival curves for beef and dairy animals. Young beef animals (up to 21 months of age) have a higher survival probability compared to dairy. A log-rank test shows that the type of production is significant both for healthy and unhealthy animals (< 0.0001). The median life expectancy is 759 days for beef animals and 1083 days for dairy. For healthy beef animals the median is a mere 750 days; compare with 981 days for healthy dairy, 807 days for unhealthy beef, and 1550 days for unhealthy dairy.

5) DVO area code: No significant differences are observed among the estimated survival curves conditional on DVO codes (see Figure 5). For instance, the median survival time ranges from 753 days for Derry/Londonderry to 815 days for Coleraine. In Table II we see all the median times for each DVO for all animals, healthy and unhealthy ones.

6) Breed: The survival curves of the subpopulations of animals of different breeds are shown in Figure 6. We have highlighted the beef and dairy breed and we see that they form

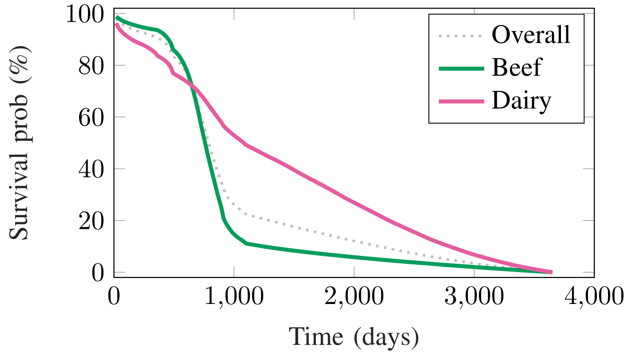


Fig. 4: Estimated survival curves for different production types (beef/dairy).

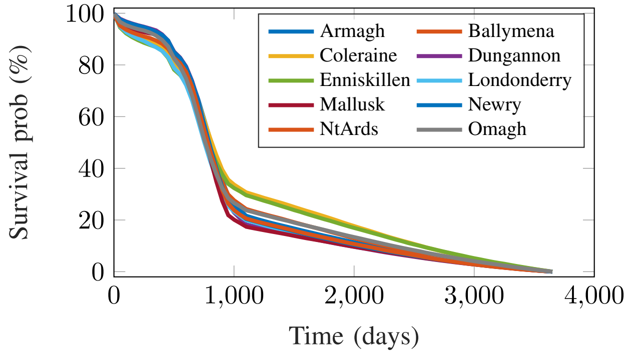


Fig. 5: Estimated survival curves for different areas. The median life spans range from 2.06 years (753 days) for Londonderry to 2.23 years (815 days) for Coleraine.

clusters.

7) Abattoir: Animals that die in different abattoirs may have significantly different survival curves as shown in Figure 7. It is notable that four clusters of survival curves can be identified. Note that Abattoirs #15, #12, #16 (blue lines) and Abattoir #10 (black line) receive—almost exclusively (>99%)—animals that died on a farm.

8) Health status: The health status (healthy/unhealthy) seems to have a significant effect on the survival function as seen in Figure 8. The survival curve of the unhealthy animals is above that of the healthy animals. The median life expectancy of healthy animals is 772 days, while unhealthy animals have a median life span of 874 days.

Next, we look at the health status combined with the sex. The resulting survival curves are shown in Figure 9. A first observation is that the health status is a significant additional feature for female animals and bulls, but offers little additional information for male animals. In particular, the survival curve of unhealthy bulls is significantly above that of healthy ones (and significantly above the survival curve of general bulls). The median life span of unhealthy bulls is 840 days, compared to 564 for healthy bulls.

DVO	all	healthy	unhealthy
Coleraine	815	791	945
Armagh	810	794	889
Newry	793	782	879
Ballymena	793	780	872
Dungannon	790	775	855
Enniskillen	786	758	889
Nt'Ards	782	769	869
Mallusk	769	757	841
Omagh	761	730	857
LondonDerry	753	735	849

TABLE II: Estimated median values of life span in different areas.

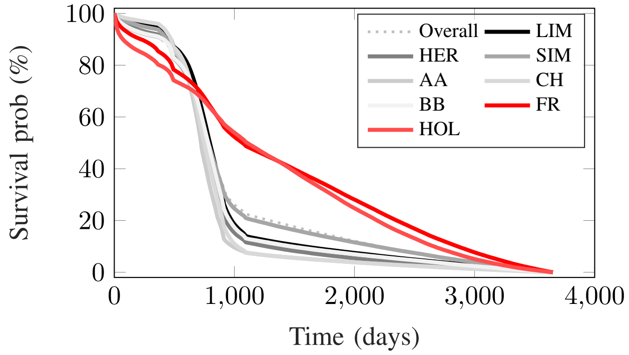


Fig. 6: Estimated survival curves for animals of different breeds. The dairy breeds, FR and HOL (shades of red), have higher survival probabilities at later ages. LIM: Limousine, HER: Hereford, SIM: Simmental, AA: Aberdeen Angus, CH: Charolais, BB: Belgian Blue, FR: Friesian, HOL: Holstein.

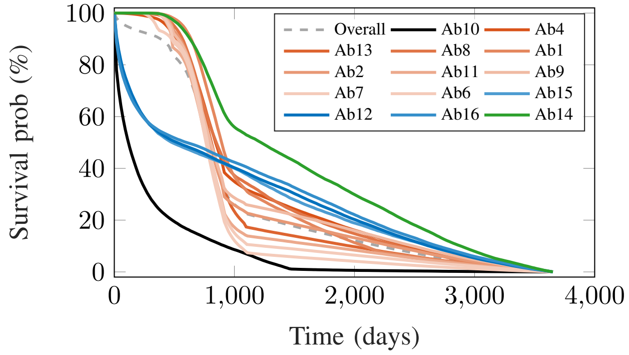


Fig. 7: Estimated survival curves for animals that died in different abattoirs.

Figure 10 shows the survival curves of those subpopulations of animals where a particular health condition is present. To better see the effect of each condition, we take those animals

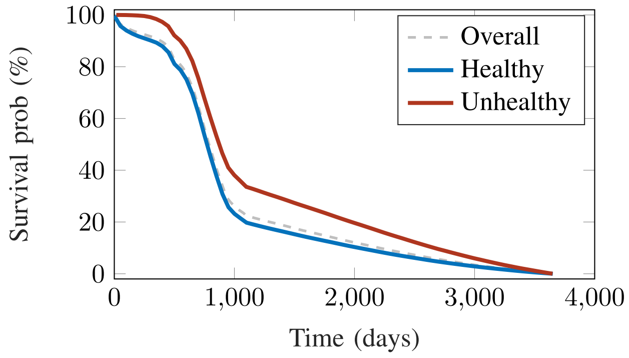


Fig. 8: Estimated survival curves for healthy and unhealthy animals.

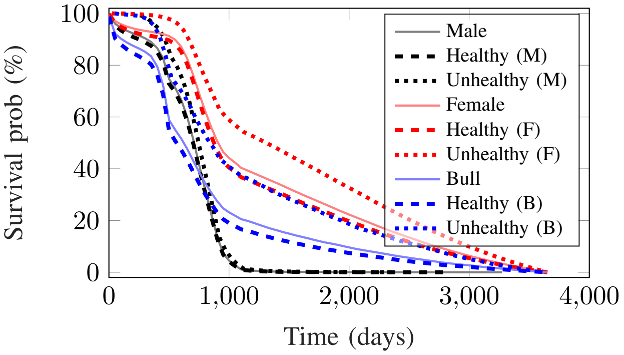


Fig. 9: Estimated survival curves for animals of different sexes and health statuses.

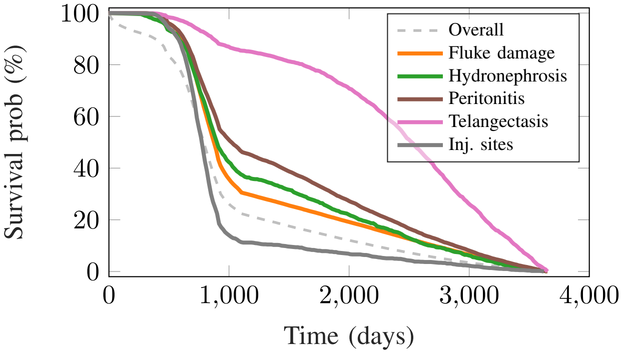


Fig. 10: Estimated survival curves for animals with particular health conditions found in a postmortem examination. Here we have selected the five health conditions that differentiate from the baseline the most.

where a single condition is present.

B. Fitting Parametric Distributions

Here we want to see whether the distribution of the data can be described by a distribution of a standard parametric family. Evidently, the survival function of Figure 1 is multimodal and can only be poorly approximated by distributions of the exponential, Weibull, or log-logistic families. Indeed, the best estimated Weibull distribution comes with a KS distance of 15.7%. This is also indicative of the presence of subpopulations that follow different distributions.

After extensive experimentation, we have been able to identify certain subpopulations that are adequately described by the Weibull distribution. Indicatively, for the data of the DVOs of Newry, Dungannon, and Newtownards, considering only male animals that died on a farm and had no conditions, the KS distance between the empirical distributions and estimated Weibulls is 5.49%, 3.32%, and 5.75%, respectively. These are the best results obtained using Weibull distributions. However, such good results are rare and the Weibull distribution seems to be an adequate approximation of only for the aforementioned subpopulations. The exponential distribution did not yield meaningful results as the KS distance was very large (>20%).

C. Mixture modelling

1) Mixtures of Gaussians: A mixture of Gaussians can point towards the presence of subpopulations. The results shown in Table III are indicative of the possible existence of at least two subpopulations in the data. However, to the best of the author's knowledge Gaussian mixtures are not popular in survival analysis, perhaps due to the facts that (i) Gaussian distributions do not take only positive values, (ii) are not right skewed as survival data often are.

Num. Gaussians	KS distance
1	24.20%
2	7.57%
3	7.70%
4	1.76%
5	1.78%

TABLE III: KS distance between estimated and empirical distributions for different numbers of components of a Gaussian mixture.

Num. Weibulls	All data	Slice A	Slice B
1	15.49%	21.77%	7.84%
2	12.49%	2.38%	1.28%

TABLE IV: KS distance between estimated and empirical distributions: single Weibull and mixture of two Weibull distributions.

2) Mixtures of Weibulls: Here we present results of fitting a mixture of two Weibull distributions. We already established in Section III-B that a single Weibull does not capture the distribution of lifespan of all animals ($d_{\mathrm{KS}} = 15.49\%$). A mixture of two Weibulls does not yield a good approximation either ($d_{\mathrm{KS}} = 12.49\%$).

Remarkably, if we further split the data into two "slices", we can obtain significantly improved results. In particular, Slice B comprises the healthy animals who died on a farm, while Slice A is the complement of B. Then, using a mixture of two Weibulls, the KS distance is just 2.38% on Slice A and 1.28% on Slice B. The results are summarised in Table IV.

The KM-estimated survival functions for Slices A and B together with the best two-Weibull mixtures are shown in Figures 11 and 12, respectively. We see that the quality of fit is very good (see also Table IV).

Lastly, mixtures of two Weibull distributions fit well upon further "slicing" of the data. For example, if we take the male/female/bull animals of Slice A, the KS distances are 0.90%/1.86%/4.11%. The corresponding KS distances of the male/female/bulls of Slice B are 5.85%/1.83%/10.28%.

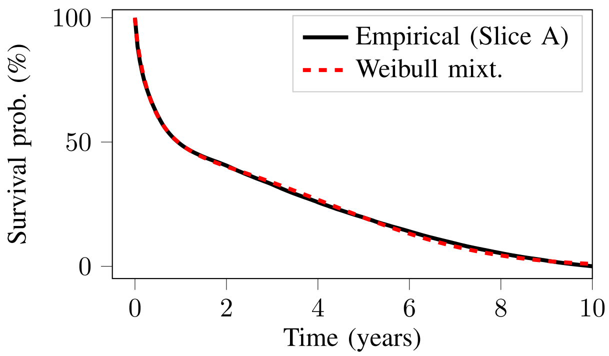


Fig. 11: Estimated mixture of two Weibull distributions on Slice A.

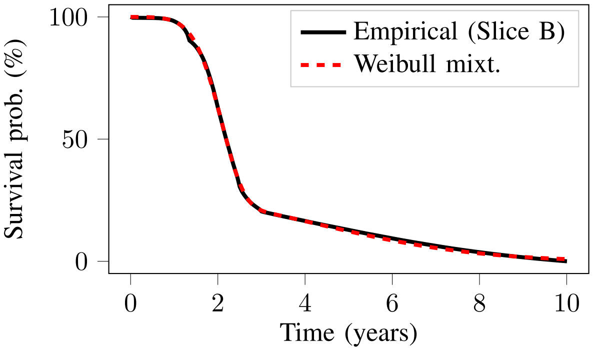


Fig. 12: Estimated mixture of two Weibull distributions on Slice B.

D. Accelerated failure time models

We constructed AFT models using the variables DVO, abattoir, sex, place of death, number of conditions, herd size, and breed, and four candidate parametric distributions: exponential, log-logistic, Weibull, and log-normal. Overall, we trained $4(2^7-1)=508$ such models.

The Weibull distribution led to the highest log-likelihood values, although these were significantly low. Two models reported very high likelihood (order of magnitude 10^8), but the estimated coefficients were null indicating, according to `survreg`'s documentation, that the estimation problem was overdetermined. The best models alongside their Akaike information criterior (AIC), log-likelihood, and C-index values are presented in Table V.

To address the overdetermination issue in the first two models, the modelling exercise was repeated with a random sample of 5% of the total data ($N=91,630$). The results are shown in Table V.

The five highest coefficients of the best model, on all data, are: Abattoir # 10 (-0.955), Abattoir # 3 (0.74), Breed Friesian (0.38), Male (-0.36), and Breed Holstein (0.31).

E. Proportional hazards models

Initially, 127 Cox models where generated and tested using all possible combinations (2^7-1) of the variables DVO, abattoir, sex, place of death, number of conditions, herd size, and breed. The obtained log-likelihood values, however, were too low—the maximum value was $-2.426\cdot 10^7$—and the PH assumption could not be verified in any of these cases using the Schoenfeld residuals test. In terms of the C-index, the best Cox model was one with variables abattoir, breed, sex, and place of death (C-index: 68.31%, see Table VI).

Next, we repeated this modelling exercise using 0.1%, 0.5%, and 1% randomly sampled data, however, the results remained equally poor. We also took the subpopulations consisting of animals without health conditions, without any significant improvement (max. log likelihood of $-1.959\cdot 10^7$).

We trained Cox models with shared frailty terms on the variables DVO, abattoir, breed, sex, place of death, number of conditions, and herd size. In all cases, the PH assumption was not satisfied and the log-likelihood was at the order of 10^7.

We further attempted to build models using a shared frailty term on the herd ID, but this proved to be computationally intractable due to the large number of data points. We, therefore, took a random sample of 5% of the data ($N=91,635$). The PH assumption was satisfied in many of the cases (see Table VI).

Overall, we trained Cox models, on all data and a 5% random sample, with and without frailty on different variables, and on subpopulations of beef/dairy and healthy/unhealthy animals (totalling 24,384 models).

To assess the predictive ability of the above Cox models—with and without frailty terms— we computed their C-indices. The C-indices for a selection of the best-performing models is given in Table VI.

The five highest coefficients of the best Cox model with frailty (with variables DVO code, abattoir, and sex) are: abattoir #10 (2.09 ± 0.033), female (-0.88 ± 0.026), abattoir #12 (0.51 ± 0.027), abattoir #15 (0.47 ± 0.053), male (0.42 ± 0.026). Recall that abattoirs #10, #15, and #12 receive animals that died on a farm.

Similarly, the five highest coefficients of the best Cox model without frailty (with variables abattoir, breed, sex, place of death, and herd size) are: abattoir #10 (1.44 ± 0.16), breed FR (-0.88 ± 0.015), death on farm (0.79 ± 0.15), breed HOL (-0.78 ± 0.019), and female (-0.71 ± 0.026).

For these two models, the forest plots are given in Figures 13 and 14, respectively.

F. Survival forests for survival analysis

The survival forests methodology was applied to the cattle mortality data using the variables DVO code, abattoir, production type (beef/dairy), sex, place of death, and number of health conditions. We experimented with different numbers of trees and different sample sizes. The modelling errors, in terms of the concordance index (C-index) are presented in Figure 15. Note that the associated computational cost and memory requirements were too high to run more models with 50% of the data (for 70 trees, the estimation takes 20.1 hours and the overall computation time was 8.2 days).

The best out-of-bag results (on training data) are obtained for the model trained on 1% of the data with 70 trees. To assess the generalisation capacity of these models, we compute the C-index values on external validation data that are not used for training. The results are presented in Figure 15.

Albeit with a small difference, the survival forest that performs the best on validation data in terms of its C-index is the one trained on 1% of the data using 190 trees (C-index: 69.41%). The library `randomForestSRC` outputs a variable importance index (Ishwaran et al. 2008, Sec. 7); for this forest the importance of each variable is shown in Table VII. The variable with the highest importance is sex, followed by the production type. For the model trained on the subpopulation of male animals, the most important variable is the place of death, followed by the herd size.

Next, we repeated the above modelling exercise on subpopulations of the data using a random sample of 10% of the dataset. In particular, we looked at animals of specific sexes and production types. In some cases, e.g., for dairy animals and bulls, better C-indices were achieved. The results are presented in Table VIII.

	All data			5% of data		
Variables	AIC	$\log L$	C-index‡	AIC	$\log L$	C-index‡
DVO, Abattoir, Breed, Sex, Death place, Num. cond., Herd size	27,802,610	-13,901,267	68.20%	139,105	-69,515	67.93%
DVO, Abattoir, Breed, Sex, Num. cond., Herd size	27,802,638	-13,901,282	68.20%	139,104	-69,515	67.93%
DVO, Abattoir, Breed, Sex, Death place, Num. cond.	27,803,492	-13,901,709	68.19%	139,111	-69,519	67.93%
DVO, Abattoir, Sex, Death place, Num. cond.*	—	—	—	139,597	-69,768	66.60%
DVO, Abattoir, Sex, Death place, Num. cond., Herd size*	—	—	—	139,599	-69,769	66.61%

TABLE V: Best AFT models estimated with the associated values of the AIC, log-likelihood value, and C-index using the entire data and a random sample of 5%. *For these models, the coefficients could not be estimated using the entire data, but estimation was successful on a random sample of 5%. ‡Validation using 20,000 randomly selected points.

	Without frailty*		Shared frailty (herd)**	
Variables	C-index	PH assumption	C-index†	PH assumption
DVO + Abattoir + Sex	**66.31%**		**73.52%**	✓
DVO + Breed + Sex	65.15%		72.21%	✓
Breed + Sex + Herd size	65.06%		72.21%	✓
Breed + Sex	64.98%		72.21%	✓
Sex + Herd size	61.80%		71.53%	✓
Sex	61.52%		71.53%	✓
DVO + Sex + Herd size	62.47%		71.53%	✓
DVO + Sex	62.41%		71.53%	✓
Breed	57.76%		69.10%	✓
Breed + Herd size	57.73%		69.10%	✓
DVO + Bree + Herd size	58.06%		69.07%	✓
Herd size	51.36%		68.42%	✓
DVO	51.55%		68.41%	✓
Abattoir + Breed + Sex + Death place + Herd size	**68.31%**		**74.34%**	
DVO + Breed + Sex + Death place + Herd size	68.31%		73.90%	
Breed + Sex + Place of death	66.93%		73.90%	
Abattoir + Sex (+ Place of death/Herd size)	65.95%		73.50%	
DVO + Sex + Place of death	64.72%		73.01%	
Sex + Place of death	63.85%		73.01%	
Abattoir + Breed	62.43%		71.49%	

TABLE VI: Concordance indices of Cox models without frailty and with a shared frailty term on the herd ID, evaluated on validation data. *Using the entire dataset, **Using 5% of the dataset, randomly sampled, for the sake of computationally feasibility. †The rows of the table are ordered by the C-index of Cox models in descending order (first those models where the PH assumption is statistically verified).

	Scaled var. importance		
Variable	All data	Male	Dairy
Sex	100	—	100
Production type	48.15	0.00	—
Abattoir	18.91	8.63	22.58
Place of death	11.27	100	20.41
Herd size	43.94	65.87	8.43
DVO	0.00	8.47	6.64

TABLE VII: Scaled variable importance for survival forest trained on 1% of data with 190 trees.

Subpopulation	Num. Trees	C-index (validation)
All data	190	69.41%
Dairy	50	75.11%
Bull	180	71.24%
Female	100	68.72%
Male	100	66.10%
Beef	110	61.54%

TABLE VIII: Survival random forests applied to subpopulations of the data. Best model reported having searched among a number of trees from 10 to 200 with a step of 10.

	Population	RMSE (train/val.)	R^2 (train/val.)
Reg. Tree	All	617.9 / 618.5	0.30 / 0.30
XGBTree	All	561.6 / 561.7	0.42 / 0.43
Random forest	All	529.9 / 513.9	0.49 / 0.52
	Beef	453.7 / 442.7	0.35 / 0.39
	Dairy	681.6 / 673.4	0.50 / 0.51
	Male	**168.9 / 153.4**	**0.58 / 0.65**
	Female	714.2 / 691.7	0.33 / 0.38
	Bull	589.7 / 563.7	0.36 / 0.42
	Healthy	519.5 / 494.4	0.46 / 0.51
	Unhealthy	599.5 / 580.8	0.47 / 0.51
	Slice A	779.2 / 761.0	0.38 / 0.40

TABLE IX: Best regression models and their RMSE and R^2 values on the training and validation datasets.

G. *Individual mortality prediction*

We applied the proposed methodology using (i) regression trees, (ii) random forests, (iii) extreme gradient boosted trees.

A selection of the best obtained models from this exercise are reported in Table IX with the RMSE and R^2 values on training and validation data. Although individual prediction models on the entire data perform poorly, with RMSE values of above 1.3 years, male animals are well within the domain of applicability of random forests with RMSE value as low as 153.4 days (0.42 years) on validation data.

IV. DISCUSSION OF RESULTS

In this section we discuss the survival analysis results of Section III. It should be noted that any relations between variables and survival do not imply a cause-and-effect relationship. For example, morbidity coincides with higher survival, but does not cause it. Instead, it is the animals that survive longer that have diseases.

A. *Kaplan-Meier models*

By estimating the survival functions of different subpopulations of the data using the KM method we see that the differences are most pronounced for animals of different production

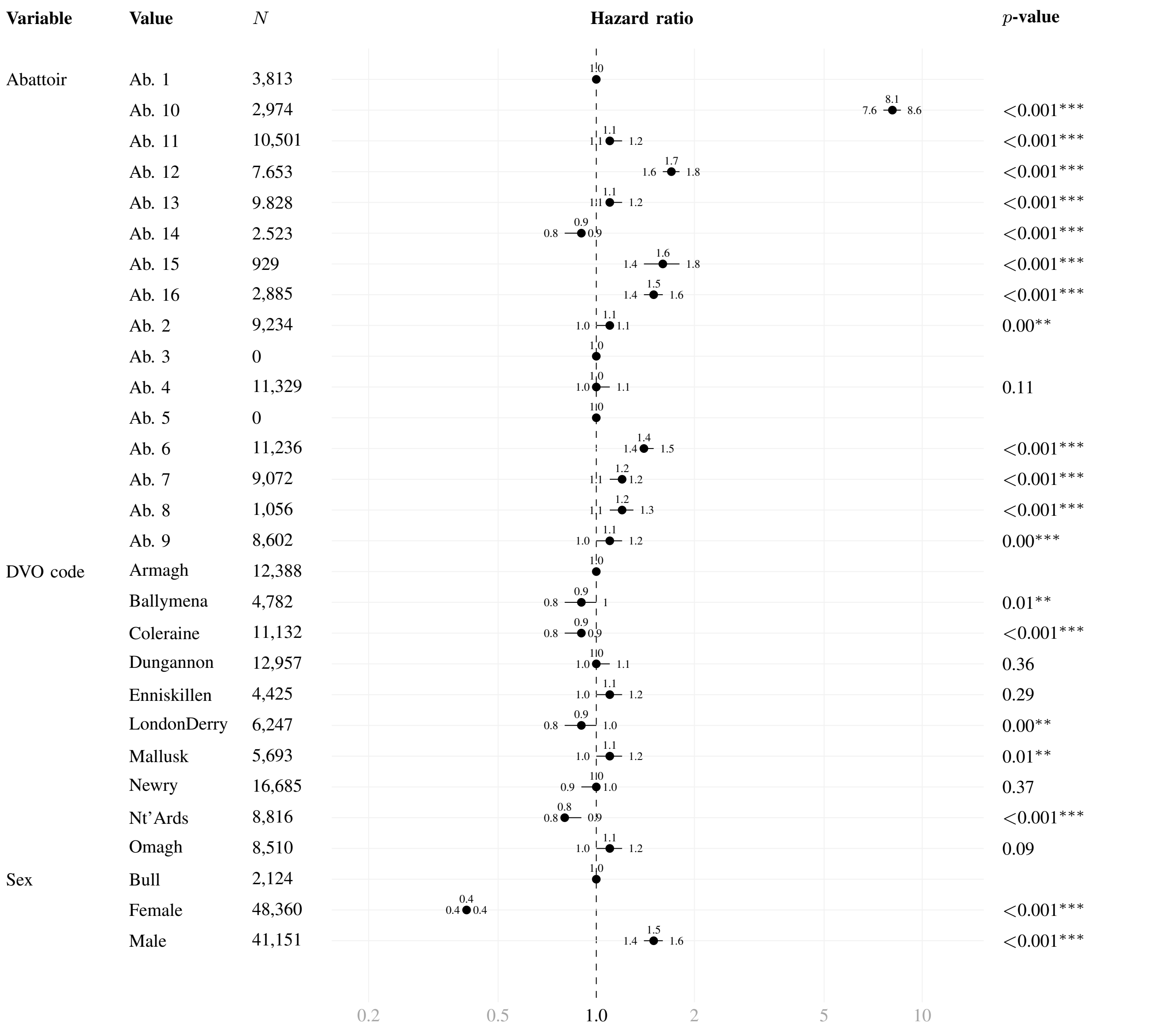


Fig. 13: Forest plot of the Cox model with variables abattoir, area code and sex, with frailty on Herd id, for which the C-index is 73.52%, The Cox PH assumption for this model can be verified (see Table VI). **Significant. ***Very significant.

types ($d_{\mathrm{KS}} = 38.32\%$), place of death ($d_{\mathrm{KS}} = 50.69\%$), and sex ($d_{KS} = 39.75\%$ between male/female, 19.48% between male/bull, and 35.53% between female/bull), as shown in Figures 2, 3, and 4.

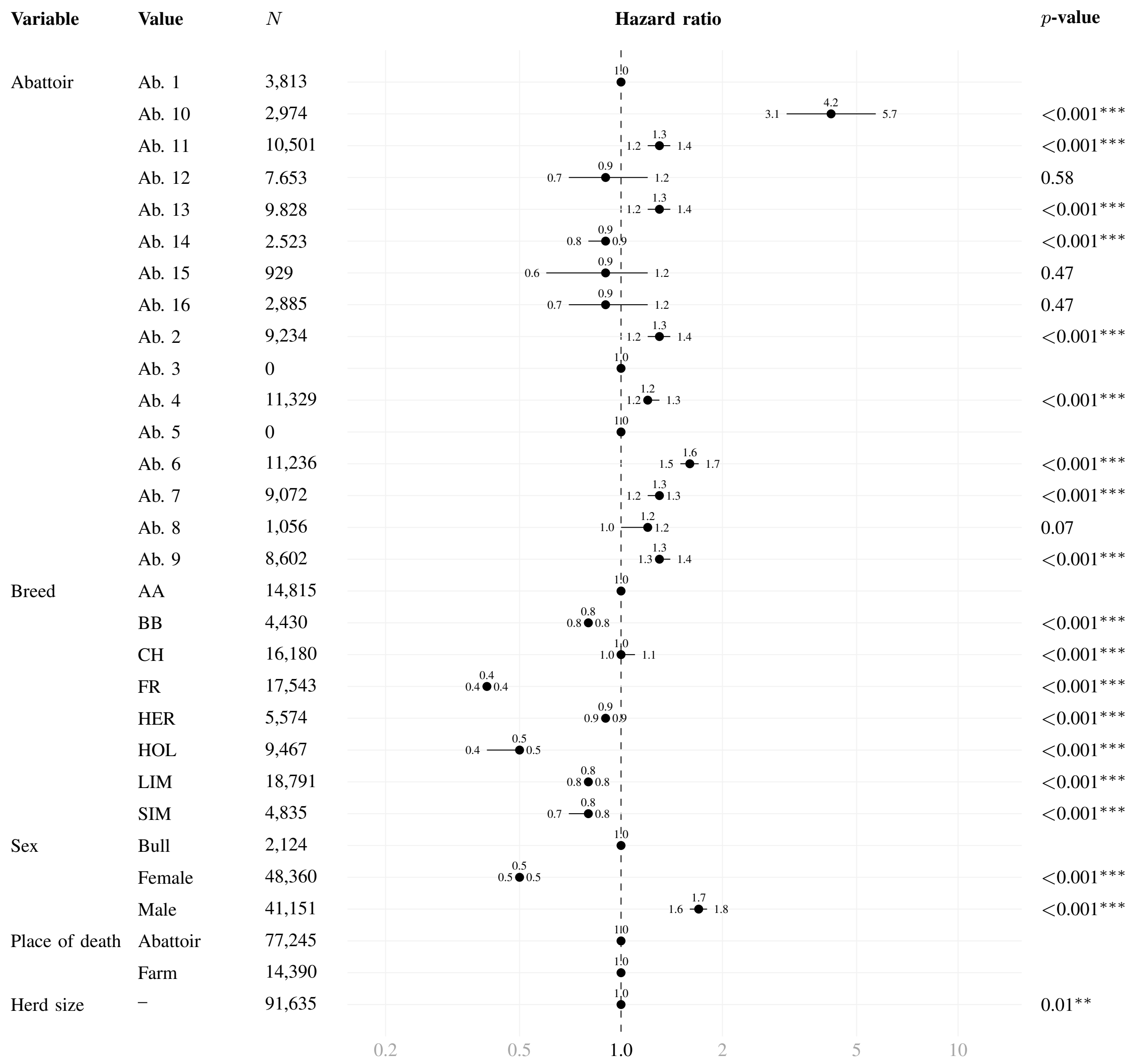
Specifically, beef animals have shorter life spans (median: 759 days) compared to dairy (median: 1082 days), and the right tail of dairy animals is thicker (95%-quantiles are 2193 and 3137 days, respectively). The opposite effect is observed on the left of these distributions: mortality in young dairy animals is higher than in young beef animals. The higher survival of dairy animals comes as no surprise, as beef animals are typically slaughtered around the age of 869 days (see Table XIII).

The survival curves based on the breed of animal can be grouped into two distinct clusters as shown in Figure 6, corresponding to beef and dairy animals. The largest median life span is 1108 days for Holstein cattle followed by Friesian with 1068 days—both dairy breeds. The smallest median is that of Aberdeen Angus at 723 days. Belgian Blue and Simmental have almost identical medians at 780 and 783 days respectively.

Likewise, as discussed in Section III-A2, female animals live longer. This is partly explained by the fact that the majority of dairy animals (64.7%) are female, but it may be due to other factors (e.g., female animals are used for reproductive purposes).

The place of death also seems an important risk factor. From Figure 3 it is evident that very few young animals die in an abattoir: 90% of the animals that die in an abattoir are over 347 days old. Young animals instead die in the farm, and, remarkably, 12.2% of them are younger than 18 days of age.

In Section III-A8 we also saw that health conditions are



Variable	Value	*N*	Hazard ratio	*p*-value
Abattoir	Ab. 1	3,813	1.0	
	Ab. 10	2,974	4.2 (3.1–5.7)	<0.001***
	Ab. 11	10,501	1.3 (1.2–1.4)	<0.001***
	Ab. 12	7.653	0.9 (0.7–1.2)	0.58
	Ab. 13	9.828	1.3 (1.2–1.4)	<0.001***
	Ab. 14	2.523	0.9 (0.8–0.9)	<0.001***
	Ab. 15	929	0.9 (0.6–1.2)	0.47
	Ab. 16	2,885	0.9 (0.7–1.2)	0.47
	Ab. 2	9,234	1.3 (1.2–1.4)	<0.001***
	Ab. 3	0	1.0	
	Ab. 4	11,329	1.2 (1.2–1.3)	<0.001***
	Ab. 5	0	1.0	
	Ab. 6	11,236	1.6 (1.5–1.7)	<0.001***
	Ab. 7	9,072	1.3 (1.2–1.3)	<0.001***
	Ab. 8	1,056	1.2 (1.0–1.2)	0.07
	Ab. 9	8,602	1.3 (1.3–1.4)	<0.001***
Breed	AA	14,815	1.0	
	BB	4,430	0.8 (0.8–0.8)	<0.001***
	CH	16,180	1.0 (1.0–1.1)	<0.001***
	FR	17,543	0.4 (0.4–0.4)	<0.001***
	HER	5,574	0.9 (0.9–0.9)	<0.001***
	HOL	9,467	0.5 (0.4–0.5)	<0.001***
	LIM	18,791	0.8 (0.8–0.8)	<0.001***
	SIM	4,835	0.8 (0.7–0.8)	<0.001***
Sex	Bull	2,124	1.0	
	Female	48,360	0.5 (0.5–0.5)	<0.001***
	Male	41,151	1.7 (1.6–1.8)	<0.001***
Place of death	Abattoir	77,245	1.0	
	Farm	14,390	1.0	
Herd size	–	91,635	1.0	0.01**

Fig. 14: Forest plot of the Cox model with variables abattoir, breed, sex, place of death, and herd size for which the C-index is 74.34%, but the Cox PH assumption cannot be verified (see Table VI). **Significant. ***Very significant.

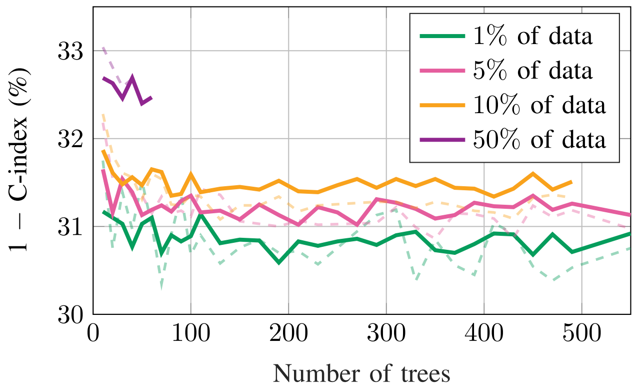


Fig. 15: (Solid lines) Modelling errors on test data in terms of 1 − C-index for survival forests against the number of trees for different sample sizes. (Dashed lines) Out-of-bag errors on the training set.

present in longer-lived animals.

In Figure 9 we looked at the joint effect of sex and health status. A first observation is that the health status of male animals has almost no effect on life expectancy—the male sex alone determines mortality. In bulls and female animals, the presence of health conditions is associated with higher life expectancies.

In Figure 7 we observe an interesting abattoir-based clustering of survival curves. At abattoir #10 we have a significantly higher mortality. This is an abattoir that receives exclusively (99.11%) animals that died on a farm. This is also the case at abattoirs #15, #16, and #12 where we have very high young mortality up to 1 year of age.

As shown in Section III-A8, most health conditions seem to appear at older ages. We see that the condition "injection sites" is associated with decreased survival after around 1,000 days of age, whereas peritonitis and telangiectasis are associated with significantly higher survival (median: 2,523 days; $N = 2,894$).

Little differentiation appears among animals of the same breed (Figure 6) conditional on their production type and among animals in different geographic areas (Figure 5).

Lastly, the KM approach does not fully allow to uncover the unique effect of the various variables since we can only look at one or two variables at a time, especially, since most variables are far from independent as discussed in Appendix E.

B. *Parametric models*

Simple parametric models—using Weibull, exponential, gamma, and log-normal—poor results were obtained. On the other hand, good results were obtained by using two-Weibull mixtures on Slices A and B of the data (see Table IV and Figures 11 and 12). This is indicative of the presence of two subpopulations within Slices A and B, which follow a Weibull distribution. However, we were not able to identify which animals exactly comprise these sub-clusters.

C. *AFT models*

Having constructed AFT models with all possible combinations of variables, we found that an AFT model with a Weibull distribution using DVO code, abattoir ID, breed, sex, place of death, number of conditions, and herd size produces the highest C-index (68.20% on all data and 67.93% on a random sample of 5%).

From the preceding discussion in Section IV-A, one would expect to see that risk factors such as the sex, and production type would be most impactful. However, from the AFT coefficients of the best model (see Section III-D) we see the emergence of a mixed effect of multiple variables such as abattoir IDs, breeds, and sex.

D. *Cox proportional hazards models*

For Cox models without frailty the PH assumption cannot be verified, and, most importantly, their C-index does not exceed 68.31%. On the other hand, some Cox models with shared frailty on the herd ID satisfy the PH assumption, and their C-index is as high as 73.52%.

Specifically, the Cox model with the highest C-index that satisfies the PH assumption uses the variables DVO code, abattoir ID, and sex. The highest negative effect on survival is from abattoirs #10 (Cox coefficient 2.09 ± 0.0327), #12 (0.508 ± 0.0265), #15 (0.469 ± 0.0534). The male and female sexes have Cox coefficients of 0.418 ± 0.0259 and -0.876 ± 0.0257, respectively. As expected, the negative sign of the female coefficient implies a reduction in hazard. Figure 7 is the forest plot of the Cox coefficients.

The Mallusk area is also associated with somewhat higher hazard (Cox coefficient 0.125 ± 0.0497). Indeed, as shown in Figure 5, the right tail of the Mallusk area is visibly below those of other areas; see also Figure 5.

Experiments on specific subpopulations of the dataset did not yield good results in terms of C-index and/or satisfaction of the PH assumption.

E. *Survival forests*

Survival forests give very good results in terms of the C-index when applied to dairy animals only (C-index: 75.11%). However, on the entire population, their C-index is comparable to that of AFT and Cox models without frailty (69.41%). In all cases, a moderate number of trees (≤ 180) is required, due to the small number of variables. The sex is the most informative variable (see Table VII). Second and third come the production type and abattoir.

F. *Machine learning methods*

In this section we present results using individual prediction models (see Section II-H). The training is performed on a training set which is a random sample of 60% of the data. The models are then validated on 20% of the dataset. In particular, we use random forests, extreme boosted trees, and regression trees.

At every node of the tree, we allowed the splitting to be conditional on five variables. As shown in Table IX, the random forest approach seems to work well only on the subpopulation of male animals in terms of RMSE. Note that the splitting rule for this model is the variance rule.

V. Conclusions and future work

In this work, we used nonparametric, semiparametric and parametric modelling approaches to estimate the survival function of cattle lifespan in Northern Ireland.

The KM method revealed that beef animals live shorter lives, but mortality in young dairy animals is higher than in young beef. Female animals live longer than male and bulls. Health conditions appear at later times of animals' lives, except injection sites, which are associated with decreased survival after around 1,000 days of age. Moreover, very few (<10%) young animals (< 1 year) are slaughtered in an abattoir. The breed does not seem to offer much information additional to the production type.

Furthermore, we were able to fit a mixture of two Weibull distributions on a partition of the dataset (Slices A and B).

Cox models yield better results in terms of the C-index than AFT. A Cox model with the DVO area code, abattoir ID, and sex as variables and shared frailty on the herd ID gives a C-index of 73.52% and the PH assumption is verified. Abattoir #10 is associated with a very high hazard ratio (8.1 times the baseline hazard). Note that abattoir #10 receives almost exclusively animals that were slaughtered on a farm (99.11%), possibly prematurely (mean age 291 days). As expected, females have decreased hazard, and males have increased hazard (see Figure 13).

Overall, the highest C-index (75.11%) was obtained by a survival forest model on dairy animals only (see Table VIII). The most informative variable was the sex. However, survival forests trained on all data do not yield as good a C-index (69.41%).

Lastly, regarding individual prediction models, decently good results on the subpopulation of male animals are obtained: a random forest gives an RMSE of 153.4 days on validation data, whereas on other subpopulations and on the entire dataset, the RMSE is over a year.

Due to the high computational complexity, we were not able to train AFT and Cox models taking into account the various health conditions. Future work will focus on the study of the effect of morbidity and comorbidity on mortality.

References

A. Kassambra *et al.* (2021), ‘survminer: Survival analysis and visualization’, https://github.com/kassambara/survminer.

Aalen, O. (1978), ‘Nonparametric inference for a family of counting processes’, *The Annals of Statistics* **6**(4), 701–726.

Al-Shomrani, A. A., Shawky, A. I., Arif, O. H. & Aslam, M. (2016), ‘Log-logistic distribution for survival data analysis using mcmc’, *SpringerPlus* **5**(1).

Alvåsen, K., Jansson Mörk, M., Dohoo, I., Sandgren, C. H., Thomsen, P. & Emanuelson, U. (2014), ‘Risk factors associated with on-farm mortality in swedish dairy cows’, *Preventive Veterinary Medicine* **117**(1), 110–120.

Bai, M., Zheng, Y. & Shen, Y. (2021), ‘Gradient boosting survival tree with applications in credit scoring’, *Journal of the Operational Research Society* **73**(1), 39–55.

Balan, T. A. & Putter, H. (2020), ‘A tutorial on frailty models’, *Statistical Methods in Medical Research* **29**(11), 3424–3454.

Bellot, A. & van der Schaar, M. (2019), ‘A hierarchical bayesian model for personalized survival predictions’, *IEEE Journal of Biomedical and Health Informatics* **23**(1), 72–80.

Bengtsson, H. (2005), ‘R.utils: Various programming utilities’.
URL: *http://dx.doi.org/10.32614/CRAN.package.R.utils*

Borucka, J. (2014), ‘Extensions of Cox model for non-proportional hazards purpose’, *Ekonometria* (3(45)).

Bradburn, M. J., Clark, T. G., Love, S. B. & Altman, D. G. (2003), ‘Survival analysis part II: Multivariate data analysis – an introduction to concepts and methods’, *British Journal of Cancer* **89**(3), 431–436.

Bugnard, F., Ducrot, C. & Calavas, D. (1994), ‘Advantages and inconveniences of the Cox model compared with the logistic model: Application to a study of risk factors of nursing cow infertility’, *Veterinary Research* (25), 134–139.

Caraviello, D., Weigel, K. & Gianola, D. (2003), ‘Analysis of the relationship between type traits, inbreeding, and functional survival in jersey cattle using a Weibull proportional hazards model’, *Journal of Dairy Science* **86**(9), 2984–2989.

Chen, T. & Guestrin, C. (2016), Xgboost: A scalable tree boosting system, *in* ‘Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining’, ACM.

Ciarmiello, A., Tutino, F., Giovannini, E., Milano, A., Barattini, M., Yosifov, N., Calvi, D., Setti, M., Sivori, M., Sani, C., Bastrieri, A., Staffiere, R., Stefanini, T., Artioli, S. & Giovacchini, G. (2023), ‘Multivariable risk modelling and survival analysis with machine learning in SARS-CoV-2 infection’, *Journal of Clinical Medicine* **12**, 7164.

Clark, T., Bradburn, M., Love, S. & Altman, D. (2003), ‘Survival analysis part I: Basic concepts and first analyses’, *British Journal of Cancer* **89**, 232–8.

Collett, D. (2003), *Modelling Survival Data in Medical Research.*, Chapman & Hall/CRC Texts in Statistical Science.

Colosimo, E., Ferreira, F., Oliveira, M. & Sousa, C. (2002), ‘Empirical comparisons between Kaplan-Meier and Nelson-Aalen survival function estimators’, *Journal of Statistical Computation and Simulation* **72**(4), 299–308.

Crowther, M. J., Royston, P. & Clements, M. (2022), ‘A flexible parametric accelerated failure time model and the extension to time-dependent acceleration factors’, *Biostatistics* **24**(3), 811–831.

Cutler, A., Cutler, D. R. & Stevens, J. R. (2012), *Random Forests*, Springer New York, pp. 157–175.

Davidson-Pilon, C. (2019), ‘lifelines: survival analysis in python’, *Journal of Open Source Software* **4**(40), 1317.

Ellington, E. H., Lewis, K. P., Koen, E. L. & Vander Wal, E. (2020), ‘Divergent estimates of herd-wide caribou calf survival: Ecological factors and methodological biases’, *Ecology and Evolution* **10**(15), 8476–8505.
URL: *http://dx.doi.org/10.1002/ece3.6553*

Farewell, V. T. (1982), ‘The use of mixture models for the analysis of survival data with long-term survivors’, *Biometrics* **38**(4), 1041.

Foley, M. (2022), ‘Supervised machine learning’, Available at https://bookdown.org/mpfoley1973/supervised-ml/.

Fotso, S. et al. (2019), ‘PySurvival: Open source package for survival analysis modeling’.
URL: *https://www.pysurvival.io/*

Fruscalso, V., Olmos, G. & Hötzel, M. (2020), ‘Dairy calves’ mortality survey and associated management practices in smallholding, pasture-based herds in southern Brazil’, *Preventive Veterinary Medicine* **175**, 104835.

George, B., Seals, S. & Aban, I. (2014), ‘Survival analysis and regression models’, *Journal of Nuclear Cardiology* **21**(4), 686–694.

Hansen, A. (2020), ‘The three extreme value distributions: An introductory review’, *Frontiers in Physics* **8**.

Hartman, N., Kim, S., He, K. & Kalbfleisch, J. D. (2023), ‘Pitfalls of the concordance index for survival outcomes’, *Statistics in Medicine* **42**(13), 2179–2190.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H. & Lauer, M. S. (2008), ‘Random survival forests’, *The Annals of Applied Statistics* **2**(3).

Kishore, J., Goel, M. & Khanna, P. (2010), ‘Understanding survival analysis: Kaplan-Meier estimate’, *International Journal of Ayurveda Research* **1**(4), 274.

Kuhn & Max (2008), ‘Building predictive models in R using the caret package’, *Journal of Statistical Software* **28**(5), 1–26.

Kulkarni, P., Mourits, M., Nielen, M., van den Broek, J. & Steeneveld, W. (2021), ‘Survival analysis of dairy cows in the netherlands under altering agricultural policy’, *Preventive Veterinary Medicine* **193**, 105398.

Lin, D. Y. (2007), ‘On the Breslow estimator’, *Lifetime Data Analysis* **13**(4), 471–480.

Longato, E., Vettoretti, M. & Di Camillo, B. (2020), ‘A practical perspective on the concordance index for the evaluation and selection of prognostic time-to-event models’, *Journal of Biomedical Informatics* **108**, 103496.
URL: *http://dx.doi.org/10.1016/j.jbi.2020.103496*

Manouchehri, N. & Bouguila, N. (2020), A frequentist inference method based on finite bivariate and multivariate beta mixture models, *in* ‘Unsupervised and Semi-Supervised Learning’, Springer International Publishing, Cham, pp. 179–208.

Marín, J. M., Rodríguez-Bernal, M. T. & Wiper, M. P. (2005), ‘Using Weibull mixture distributions to model heterogeneous survival data’, *Communications in Statistics - Simulation and Computation* **34**(3), 673–684.

McBride, K., Novakovic, A., Marshall, A. H. & Courcier, E. (2022), Knowledge discovery of bovine tuberculosis in the eurasian badger using machine learning techniques, *in* ‘2022 International Conference on Computational Science and Computational Intelligence (CSCI)’, pp. 362–367.

Morignat, E., Gay, E., Vinard, J.-L., Calavas, D. & Hénaux, V. (2015), ‘Quantifying the influence of ambient temperature on dairy and beef cattle mortality in france from a time-series analysis’, *Environmental Research* **140**, 524–534.

Mõtus, K., Viltrop, A. & Emanuelson, U. (2018), ‘Reasons and risk factors for beef calf and youngstock on-farm mortality in extensive cow-calf herds’, *Animal* **12**(9), 1958–1966.

Nørgaard, N. H., Lind, K. M. & Agger, J. F. (1999), ‘Cointegration analysis used in a study of dairy-cow mortality’, *Preventive Veterinary Medicine* **42**(2), 99–119.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011), ‘Scikit-learn: Machine learning in Python’, *Journal of Machine Learning Research* **12**, 2825–2830.

Pölsterl, S. (2020), ‘scikit-survival: A library for time-to-event analysis built on top of scikit-learn’, *Journal of Machine Learning Research* **21**(212), 1–6.
URL: *http://jmlr.org/papers/v21/20-729.html*

Probo, M., Pascottini, O. B., LeBlanc, S., Opsomer, G. & Hostens, M. (2018), ‘Association between metabolic diseases and the culling risk of high-yielding dairy cows in a transition management facility using survival and decision tree analysis’, *Journal of Dairy Science* **101**(10), 9419–9429.

Pölsterl, S. W. (2016), Algorithms for Large-scale Learning from Heterogeneous Survival Data, Phd thesis, Technischen Universität München (TUM), Munich. Available at https://mediatum.ub.tum.de/doc/1289752/68192.pdf.

R Core Team (2021), ‘R: A language and environment for statistical computing’. R Foundation for Statistical Computing, available at https://www.R-project.org/.

Rasmussen, P., Barkema, H. W., Osei, P. P., Taylor, J., Shaw, A. P., Conrady, B., Chaters, G., Muñoz, V., Hall, D. C., Apenteng, O. O., Rushton, J. & Torgerson, P. R. (2024), ‘Global losses due to dairy cattle diseases: A comorbidity-adjusted economic analysis’, *Journal of Dairy Science* **107**(9), 6945–6970.

Razali, A. M. & Al-Wakeel, A. A. (2013), ‘Mixture Weibull distributions for fitting failure times data’, *Applied Mathematics and Computation* **219**(24), 11358–11364.

Reid, M. (2020), ‘Python package matthewreid854/reliability: v0.5.1’.
URL: *https://zenodo.org/record/3938000*

Reimus, K., Alvåsen, K., Emanuelson, U., Viltrop, A. & Mõtus, K. (2020), 'Herd-level risk factors for cow and calf on-farm mortality in Estonian dairy herds', *Acta Veterinaria Scandinavica* **62**(1).

Renaud, D., Duffield, T., LeBlanc, S., Ferguson, S., Haley, D. & Kelton, D. (2018), 'Risk factors associated with mortality at a milk-fed veal calf facility: A prospective cohort study', *Journal of Dairy Science* **101**(3), 2659–2668.

Rezaei, M., Tapak, L., Alimohammadian, M., Sadjadi, A. & Yaseri, M. (2020), 'Review of random survival forest method', *Journal of Biostatistics and Epidemiology* .

Rodriguez, G. (2010), 'Parametric survival models'. Available at https://grodri.github.io/survival/ParametricSurvival.pdf.

Santman-Berends, I., Buddiger, M., Smolenaars, A., Steuten, C., Roos, C., Van Erp, A. & Van Schaik, G. (2014), 'A multidisciplinary approach to determine factors associated with calf rearing practices and calf mortality in dairy herds', *Preventive Veterinary Medicine* **117**(2), 375–387.

Santman-Berends, I., de Bont-Smolenaars, A., Roos, L., Velthuis, A. & van Schaik, G. (2018), 'Using routinely collected data to evaluate risk factors for mortality of veal calves', *Preventive Veterinary Medicine* **157**, 86–93.

Sarjokari, K., Hovinen, M., Seppä-Lassila, L., Norring, M., Hurme, T., Peltoniemi, O., Soveri, T. & Rajala-Schultz, P. (2018), 'On-farm deaths of dairy cows are associated with features of freestall barns', *Journal of dairy science* **101**(7), 6253–6261.

Singh, R. & Mukhopadhyay, K. (2011), 'Survival analysis in clinical trials: Basics and must know areas', *Perspectives in Clinical Research* **2**(4), 145.

Stański, K., Lycett, S., Porphyre, T. & Bronsvoort, B. M. d. C. (2021), 'Using machine learning improves predictions of herd-level bovine tuberculosis breakdowns in great britain', *Scientific Reports* **11**(1).

Therneau, T. M. & Grambsch, P. M. (2000), *Residuals*, Springer New York, New York, NY, pp. 79–86.

Therneau, T. M. & M., G. P. (2000), *Modeling Survival Data: Extending the Cox Model*, Springer, New York.

Thomsen, P. T., Kjeldsen, A. M., Sørensen, J. T. & Houe, H. (2004), 'Mortality (including euthanasia) among danish dairy cows (1990-2001)', *Preventive Veterinary Medicine* **62**(1), 19–33.

Urie, N., Lombard, J., Shivley, C., Kopral, C., Adams, A., Earleywine, T., Olson, J. & Garry, F. (2018), 'Preweaned heifer management on US dairy operations: Part V. factors associated with morbidity and mortality in preweaned dairy heifer calves', *Journal of Dairy Science* **101**(10), 9229–9244.

van der Heide, E., Kamphuis, C., Veerkamp, R., Athanasiadis, I., Azzopardi, G., van Pelt, M. & Ducro, B. (2020), 'Improving predictive performance on survival in dairy cattle using an ensemble learning approach', *Computers and Electronics in Agriculture* **177**, 105675.

van der Heide, E., Veerkamp, R., van Pelt, M., Kamphuis, C. & Ducro, B. (2020), 'Predicting survival in dairy cattle by combining genomic breeding values and phenotypic information', *Journal of dairy science* **103**(1), 556–571.

Wang, P., Li, Y. & Reddy, C. K. (2019), 'Machine learning for survival analysis: A survey', *ACM Comput. Surv.* **51**(6).

Wickham, H. (2011), 'The split-apply-combine strategy for data analysis', *Journal of Statistical Software* **40**(1), 1–29.
URL: *https://www.jstatsoft.org/v40/i01/*

Wickham, H. (2016), *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York.
URL: *https://ggplot2.tidyverse.org*

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K. & Yutani, H. (2019), 'Welcome to the tidyverse', *Journal of Open Source Software* **4**(43), 1686.

Wickham, H., François, R., Henry, L., Müller, K. & Vaughan, D. (2023), *dplyr: A Grammar of Data Manipulation*. R package version 1.1.4, https://github.com/tidyverse/dplyr.
URL: *https://dplyr.tidyverse.org*

Wickham, H., Vaughan, D. & Girlich, M. (2024), *tidyr: Tidy Messy Data*. R package version 1.3.1, https://github.com/tidyverse/tidyr.
URL: *https://tidyr.tidyverse.org*

Wiegrebe, S., Kopper, P., Sonabend, R., Bischl, B. & Bender, A. (2024), 'Deep learning for survival analysis: a review', *Artificial Intelligence Review* **57**(3).

Wisnieski, L., Amrine, D. E. & Renter, D. G. (2022), 'Predictive models for weekly cattle mortality after arrival at a feeding location using records, weather, and transport data at time of purchase', *Pathogens (Basel)* **11**(4), 473–.

Zablotski, Y., Voigt, K., Hoedemaker, M., Müller, K. E., Kellermann, L., Arndt, H., Volkmann, M., Dachrodt, L. & Stock, A. (2024), 'Perinatal mortality in German dairy cattle: Unveiling the importance of cow-level risk factors and their interactions using a multifaceted modelling approach', *Plos One* **19**(4), e0302004.

Zhao, L. & Feng, D. (2020), 'Deep neural networks for survival analysis using pseudo values', *IEEE Journal of Biomedical and Health Informatics* **24**(11), 3308–3314.
URL: *http://dx.doi.org/10.1109/JBHI.2020.2980204*

Appendix A
Hazard function

The hazard function is given by

$$\begin{aligned} h(t) &= \lim_{\delta t \to 0} \frac{\mathsf{P}[t \leq T < t + \delta t \mid T \geq t]}{\delta t} && \text{(10a)} \\ &= \lim_{\delta t \to 0} \frac{\mathsf{P}[t \leq T < t + \delta t \text{ and } T \geq t]}{\mathsf{P}[T \geq t]\delta t} && \text{(10b)} \\ &= \lim_{\delta t \to 0} \frac{\mathsf{P}[t \leq T < t + \delta t]}{S(t)\delta t} && \text{(10c)} \\ &= \frac{1}{S(t)} \lim_{\delta t \to 0} \frac{\mathsf{P}[t \leq T < t + \delta t]}{\delta t} && \text{(10d)} \\ &= \frac{1}{S(t)} \lim_{\delta t \to 0} \frac{F(t + \delta t) - F(t)}{\delta t} && \text{(10e)} \\ &= \frac{F'(t)}{S(t)} = -\frac{S'(t)}{S(t)}. && \text{(10f)} \end{aligned}$$

Appendix B
KM estimator

The (empirical) survival function can be simply estimated by

$$\widehat{S}(t) = \frac{\left\{ \begin{array}{l} \text{Num. subjects who} \\ \text{lived more than } t \end{array} \right\}}{N}, \tag{11}$$

where N is the total number of subjects. By the definition of conditional probability, for $t_2 > t_1 \geq 0$,

$$\mathsf{P}[T{>}t_2 \mid T{>}t_1] = \frac{\mathsf{P}[T > t_2 \wedge T > t_1]}{\mathsf{P}[T > t_1]} = \frac{\mathsf{P}[T > t_2]}{\mathsf{P}[T > t_1]}, \tag{12}$$

therefore, $\mathsf{P}[T > t_2] = \mathsf{P}[T > t_1]\mathsf{P}[T > t_2 \mid T > t_1]$. Equivalently, $S(t_2) = S(t_1)\mathsf{P}[T > t_2 \mid T > t_1]$. This gives rise to the following recursive way to estimate the survival function

$$\widehat{S}(t_2) = \widehat{S}(t_1) \cdot \frac{\left\{ \begin{array}{l} \text{Num. subjects who} \\ \text{lived more than } t_2 \end{array} \right\}}{\left\{ \begin{array}{l} \text{Num. subjects who} \\ \text{lived more than } t_1 \end{array} \right\}}. \tag{13}$$

Appendix C
Maximum likelihood estimation

Given identically independently distributed (iid) samples of survival times $T_1 = t_1, \ldots, T_N = t_N$ the likelihood function is defined as the joint pdf of $(T_1, \ldots, T_N)$ evaluated at the observations $(t_1, \ldots, t_N)$, seen as a function of θ, that is

$$L(\theta) = p(t_1, \ldots, t_N; \theta). \tag{14}$$

By virtue of the iid assumption, we can write

$$L(\theta) = \prod_{i=1}^{N} p(t_i; \theta). \tag{15}$$

Typically, it is convenient to take the logarithm of L, that is, $\ell(\theta) = \log L(\theta) = \sum_{i=1}^{N} \log p(t_i; \theta)$. The maximum likelihood method consists in determining the value of θ that maximises L, or, equivalently, ℓ. To do this, we take the gradient of ℓ with respect to θ and set it equal to zero. This can sometimes be done analytically, but often we need to do it computationally.

In survival analysis some of the parametric distributions that are commonly used are the Weibull, the gamma, and the log-normal.

The Weibull is a parametric distribution describing failure times, where the failure rate is a power of time. It involves two parameters $k, \lambda > 0$ and its pdf is given by

$$p(t; k, \lambda) = \frac{k}{\lambda} \left(\frac{t}{\lambda} \right)^{k-1} \exp\left(-\frac{t^k}{\lambda^k} \right), \tag{16}$$

for $t > 0$. The corresponding survival function is $S(t; k, \lambda) = \exp(-t^k / \lambda^k)$. When $k = 1$, the Weibull distribution becomes an exponential with parameter $1/\lambda$, it can, therefore, be thought of as a stretched exponential. Moreover, the class of Weibull distributions, alongside the Fréchet and Gumbel distributions, are the only distributions for describing extreme values, that is, maximum values drawn from a sample of fixed size (Hansen 2020). The maximum-likelihood estimates of k and λ can be determined numerically.

The gamma distribution is another generalisation of the exponential distribution. It is a two-parameter family of distributions with parameters $k, \theta > 0$ and pdf

$$p(t; k, \theta) = \frac{t^{k-1} \exp(-t/\theta)}{\Gamma(k)\theta^k}, \tag{17}$$

for $t > 0$. The maximum likelihood estimates of k and θ cannot be determined analytically.

Lastly, we say that T is log-normally distributed if $\log T$ follows the normal distribution, $\mathcal{N}(\mu, \sigma^2)$. The maximum likelihood estimates of μ and σ^2 are the sample average and sample variance of $t_1, \ldots, t_N$.

Appendix D
Individual predictions and survival analysis

The idea is that the life span, $T > 0$, or an animal is affected by variables x according to $T = \widehat{T}(x)\epsilon$, where $\widehat{T} > 0$ is a regression model and $\epsilon > 0$ is the modelling error (an error ratio). In other words, T is decomposed into two elements: a deterministic one, $\widehat{T}(x)$, and a random one, ϵ, that is independent of x. The model $\widehat{T}$ can be any regression model, e.g., random forests, regression trees, or neural networks.

Then, having obtained a model $\widehat{T}$, we see that the survival function of the random variable T can be written as

$$\begin{aligned} S_T(t) &= \mathsf{P}[T > t] && \text{(18a)} \\ &= \mathsf{P}[\widehat{T}(x)\epsilon > t] && \text{(18b)} \\ &= \mathsf{P}[\epsilon > t/\widehat{T}(x)] && \text{(18c)} \\ &= S_\epsilon(t/\widehat{T}(x)). && \text{(18d)} \end{aligned}$$

This means that the survival function of T can be obtained from the survival function of the modelling error, ϵ, denoted by S_ϵ. The proposed approach uses the following estimator of the survival function of T

$$\widehat{S}_T(t) = \widehat{S}_\epsilon(t/\widehat{T}(x)), \tag{19}$$

Year	Animals died	Mean age (days)
2018	506,702	997
2019	501,833	1004
2020	490,436	1016
2021	506,490	1015

TABLE X: Number of animals slaughtered and mean age per year (2018–2021).

	Female	male	bull
All animals	1313.00	661.60	834.60
Beef	1051.43	690.14	1011.03
Dairy	1767.44	563.68	532.82

TABLE XI: Mean age in days based on sex and production type of the animal

where $\widehat{S}_\epsilon$ is the KM-estimate of the survival function of ϵ. This is reminiscent of the AFT method, the main difference being that $\widehat{T}(x)$ can be any function.

In the special case where $\widehat{T}(x) = \exp(\beta^\intercal x)$, and $\widehat{S}_\epsilon$ is assumed to be of a parametric form (e.g., Weibull, exponential, log-logistic, or other), Equation (19) becomes exactly the AFT model.

Appendix E
Details of exploratory analysis

Each year (2018-2021) around 500,000 animals die in Northern Ireland. The mean age spans from 997 days in 2018 to 1016 days in 2020—see Table X.

The overall mean age for the years 2018 to 2021 is 1008 days, but there is a lot of variation based on the sex and the production type of animal as shown in the Table XI.

A. Sex

As shown in Table XII, in our data set 52.53% are females, 45.02% are male and 2.45% are bulls. Females dairy animals live approximately 3 times longer than dairy male and bulls, while, beef female and bulls live approximately 1 year longer, on average, than male beef animals. From the female animals 36.56% are dairy and 63.44% beef. The ratio is almost the same for bulls, with 36.89% for dairy and 63.11% for beef. For the male animals the ratio is slightly different with 22.53% for dairy and 77.47% for beef production. In all cases, more than half of the dead animals, are destined for beef production. The sex seems to be a possibly important discriminator of the survival of the animals.

B. Production type

Overall, 1,398,794 animals are destined for beef production and 606,667 for dairy. Although dairy animals live on average 1.5 times more than beef animals, beef cattle are healthier than dairy with around 83% of the animals not presenting any health conditions at the postmortem examination. Table XIII shows a summary of the mean age, the percentage of on-farm deaths and the percentage of healthy animals by production type. The production type also seems to be a possibly important discriminator of the survival of the animals.

Sex	% animals	% beef	% farm deaths	% healthy
Female	52.52	63.44	19.51	79.69
Male	45.03	77.47	11.56	84.25
Bull	2.45	63.11	20.64	81.06

TABLE XII: Percentage of animals by production type, place of death health condition and sex.

	mean age	% animals	% farm deaths	% healthy
Beef	869.84	69.75	9.407	82.87
Dairy	1326.8	30.25	31.06	79.26

TABLE XIII: Dairy animals live on average 1.5 times longer than beef, only 9% of the beef animals died on farm and over 82% have no health conditions during the post-mortem examination.

Num. Dead Animals	Num. Herds
1	2,530
<100	16,504
100–200	1,765
>200	2,146

TABLE XIV: Herd-wise distribution of the number of dead animals.

C. Herds

A herd is a group of animals housed together. Firstly, we want to see how the number of dead animals is distributed across herds of different size. This is summarised in Table XIV.

Next, we want to look at how the mortality rate distributes across different herds. In 29.5% of all herds ($N = 6,014$) we have a larger proportion of unhealthy animals compared to the general population average (18.22%). In 4.0% ($N = 815$) of the herds more than half of the dead animals died with at least one health condition.

Next, we want to gain some insight about the composition of the herds. In 58.5% ($N = 11,951$) of the herds only beef are present, while a 2.5% ($N = 550$) are dairy-only herds. The mean herd size is 218.3 animals and the maximum 2,028.3. The 2.8% ($N = 465$) of the herds has more than 375 animals. Big and medium-size herds (more than 100 animals) have mostly female animals (>54%). Lastly, small herds were mostly for beef production (80.25%).

As shown in Table XV, animals from big herds, tend to live on average 2.2 months longer.

Overall, the herd size does not seem to be too important a discriminator of animal survival.

D. Abattoirs

In total, we have 17 different abattoirs, two of which (No 3 and 17) closed after 2018 and only one animal has been slaughtered in each of them. One abattoir (No 5) was closed during the pandemic in 2020 and only 3 animals were

Herd size	mean age	% animals	% farm deaths	% healthy
<100	971.98	35.78	13.88	81.69
100–375	1022.58	48.70	17.01	82.46
>375	1045.76	15.52	17.43	79.83

TABLE XV: Big herds have on average slightly higher life expectancy than small ones. Fewer animals belonging in small herds die on a farm. Animals from big herds are in general less healthy than the general population.

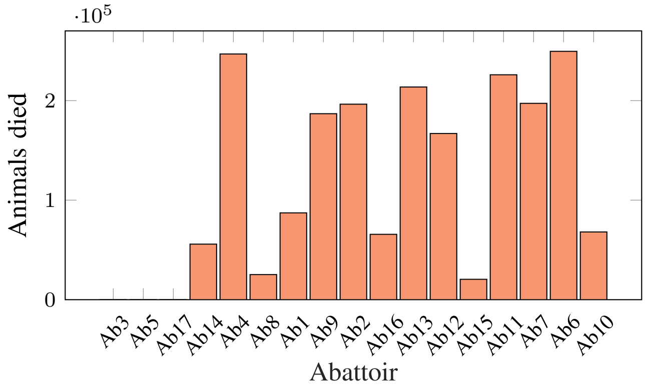


Fig. 16: Number of animals that were slaughtered in each abattoir.

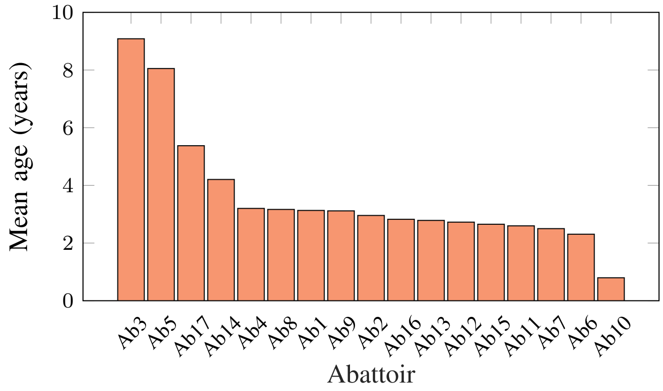


Fig. 17: Mean age per abattoir. Only one animal was slaughtered in abattoirs 3 and 17. Those abattoirs closed after 2018. Abattoir number 5 was closed during the pandemic only three animals were slaughtered there. Abattoir number 10 has the lowest mean age at 0.80 years (290 days).

slaugthered there in 2019. As shown in Figure 17, in one abattoir (number 10) the mean age is 291 days, that is, 717 days shorter than the average age of the population. 68,025 animals were slaughtered in that abattoir and the proportion of the dairy animals was around 50%. Almost all animals were healthy (3 animals out of 68,448 had one health condition each). It should be noted that abattoirs #10, #12, #15, and #16 receive animals that died on a farm.

With the exception of two abattoirs (No 14 and 10), the life expectancy in all other ones seems to be similar.

E. DVO area code

In Figure 18 we see that 18% of the total animals were from the area of Newry, while Enniskillen was the area with the fewest animals in our dataset (4.75%). As shown in Table XVI, the proportion of dairy animals per DVO code is between 24% and 37%, except for Coleraine where the proportion is 43.20%. Moreover, female animals from Coleraine live longer (1,121 days on average). Second came Omagh with 1,029 days on average. The area of Omagh has the most unhealthy animals (26% compared to 18% for the general population). Lastly, ¼ of the animals in Enniskillen died on farm instead of an abattoir. From Table XVI we see that the DVO area does not seem to have a strong effect on life expectancy.

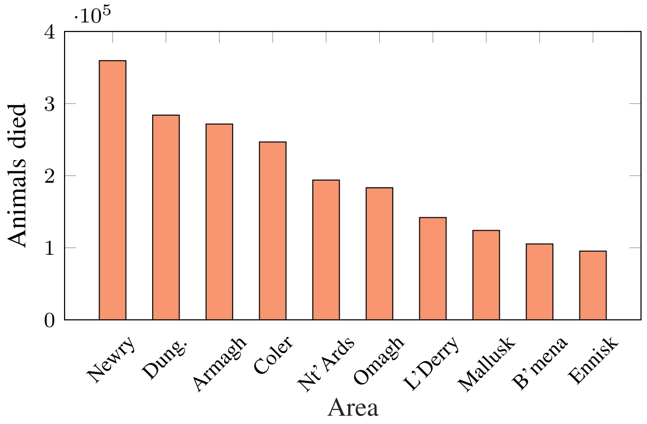


Fig. 18: Number of animals slaughtered in each area. In Newry 359,529 animals were slaughtered during the period 2018 to 2021. For the same period in Enniskillen the number of animals were 95,308.

Breed	mean age	% animals	% male	% female	% healthy
LIM	925.12	18.82	48.57	49.38	83.11
FR	1369.3	17.43	36.35	61.65	78.39
CH	816.52	16.22	50.24	47.99	83.30
AA	777.39	14.83	51.09	47.02	83.43
HOL	1288.6	9.403	25.03	70.24	80.20
HER	846.88	5.501	52.05	45.48	81.74
SIM	997.15	4.907	49.92	46.92	81.05

TABLE XVII: 7 most populous breeds in Northern Ireland. Friesian (FR) and Holstein (HOL) are dairy breeds.

F. Breed

We have 92 breeds, 33 of which are exclusively dairy breeds and 59 are exclusively beef breeds. Six breeds have more than 100,000 animals each, which accounts for the 82% of the total animals. As shown in Figure 19 the most populous breed is LIM with 19% of all animals and 27% of all beef cattle. LIM has an average mean age of 55 days above the population average for beef cattle. The proportion of sex is around 49% for male and female and 2% bulls. The most populous dairy breed is 'FR' with 17% of the total animals and 58% of dairy cattle. Mean age in days is 1369, 43 days above the average age for dairy animals. In this breed 62% are females and 36% males XVII. 29% of the animals in "FR" breed died in farm and the proportion of healthy animals is only 78.4% when in the population is 81.8%.

G. Place of death

A 40% of the animals who died on farm died before they reached 6 months of age, and 64% of the animals are female and almost 1/3 belongs to the "FR" breed. The proportion of all beef animals that died on a farm is only 9%, while for dairy is 31%. Only two out of the 320,025 animals who died on farm had some health conditions. In abattoirs 1,685,436 animals were slaughtered: 847,769 are females, 798,714 are males and 38,953 bulls. The average age of these animals is 1,038 days and 75% of the animals that were slaughtered on abattoir are for beef production. In Figure 20 we see the distribution of sexes for each place of death.

DVO	mean age	% male	% female	% dairy	% farm death	% healthy
Newry	992,41	48.37	49.21	27.68	11.83	84.55
Dungannon	972.18	45.56	52.06	23.76	11.77	79.66
Armagh	1016.7	47.27	50.80	24.63	12.88	82.06
Coleraine	1120.0	39.67	57.31	43.18	19.59	81.55
Nt'Ards	968.06	50.02	47.14	31.73	17.77	85.13
Omagh	1029.6	41.23	56.80	32.56	15.46	73.83
LondonDerry	931.96	44.71	52.32	31.11	21.95	82.85
Mallusk	943.36	46.48	51.17	26.07	15.43	85.64
Ballymena	1008.8	41.02	56.07	36.95	22.42	83.30
Enniskillen	1096.3	38.48	59.72	31.13	25.29	77.55

TABLE XVI: Mean age in days and proportions for different categories per DVO area code

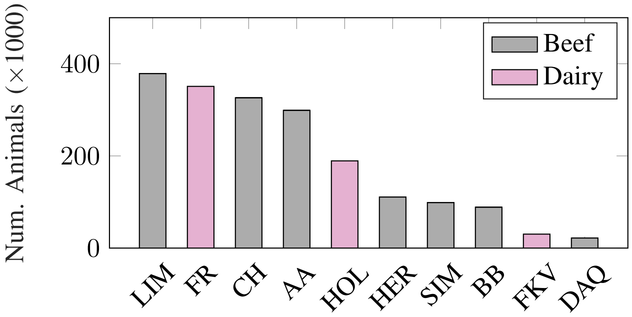


Fig. 19: The most populous breed is "LIM" with 377,332 exclusively beef cattles and second is "FR" with 349,616 exclusively dairy cattles.

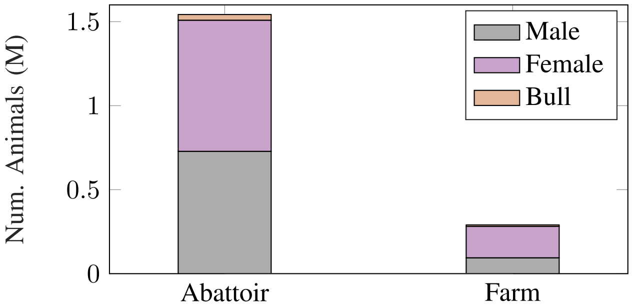


Fig. 20: A 15.96% of the animals died on farm; from these animals 64.21% are female, 32.62% male and 3.17% bulls. For the animals that died in an abattoir, 2.31% are bulls, 50.30% are female animals and 47.39% are male.

H. Health status

We consider as healthy animals the ones without any health conditions reported at the postmortem examination. All healthy animals live on average shorter than the unhealthy ones. Beef bulls with some health conditions live 1.5 times longer than beef bulls with no heath conditions. Unhealthy dairy females live almost a year longer than the healthy ones, and beef females live 297 days longer. The smallest difference in average age is observed for beef males: the unhealthy ones live 66 days longer.

In Table XVIII we see that 82% of our data have no health conditions. The mean age of healthy animals is 953 days and 71% is for beef production. The mean age increases with the number of health conditions per animal. More dairy animals have three or more health conditions and male animals are on average healthier than females.

I. Particular health conditions

In our data set there are 72 different health conditions. The unhealthy animals (animals with one or more health conditions) are 365,459. Some health conditions are extremely rare: 18 such conditions are found in only 16 cases. Moreover, 35 conditions are rare: they are found in more than 16, but less than 4,000 animals. Next, 18 conditions are found in more than 4,000 animals.

The most common health issue—as shown in Table XIX—is fluke damage which affects 100,398 animals. Only two unhealthy animals died on farm: a dairy animal with peritonitis and factory damage, and one beef animal with fluke damage.

Among the common deceases (the ones that affect at least 2000 animals) pl./pneumonia (loc), injection sites, pleurisy, and arthritis affect predominantly the male animals, whereas the remaining 19 conditions affect mostly female animals. A big split between male and female is observed for fluke damage with the condition affecting 6.34% of female animals, and 4.50% of male.

J. Temperature and humidity ratio variables

The temperature at the time of death is between 2.26°C and 16.78°C, with 50% of the observations between 6.08°C and 12.76°C. The humidity-temperature ratio is from 79.75 to 96.19 and 50% of the observations are between 86.0 and 92.0. Neither of these variables seems to have an impact on the survival of cattle as shown in the Table XX. The difference in mean life between low (<6.08°C) medium (6.08°C–12.76°C) and high (>12.76°C) temperatures is less than 10 days for both beef and dairy and less than 2 months among different sexes.

No of conditions	animals	mean age	% male	% female	% beef
0	1,640,002	953.0	46.40	51.18	70.68
1	289,974	1179.7	41.69	55.94	69.33
2	58,949	1464.0	31.05	65.71	55.71
3	12,208	1780.7	21.04	75.97	38.44
4	3,219809	1947.7	12.32	83.96	27.49
5	835	2070.0	7.425	90.30	18.56
6	222	1994.6	6.757	91.89	14.86
7	59	2122.91	5.085	93.22	13.56
8	12	1926.7	0.000	100.0	0.000
9	2	2475.0	0.000	100.0	0.000

TABLE XVIII: Summary of the number of conditions found during the postmortem examination.

	Animals			Occurrence (%)			
Condition	Num.	% of total	Mean life exp. (days)	on female	on male	on beef	on dairy
Fluke damage	100,398	5.48	1356.3	6.34	4.50	5.67	5.02
Contamination	82,797	4.52	1281.2	4.94	3.89	3.93	5.94
Fascioliasis	40,958	2.23	1217.1	2.40	2.06	2.44	1.75
Abscess, pyaemia	35,622	1.94	1246.3	2.08	1.77	1.51	2.99
TB	28,530	1.56	1180.7	2.09	0.93	1.41	1.91
Peritonitis	15,763	0.86	1634.8	1.19	0.48	0.57	1.56
Pl./Pneumonia (loc)	15,431	0.84	1142.7	0.77	0.91	0.68	1.24
Bruising	13,298	0.73	1460.9	0.95	0.47	0.45	1.38
Factory damage	10,352	0.56	1309.5	0.63	0.47	0.51	0.71
Pl./Pneumonia (gen)	8,078	0.44	1328.5	0.49	0.38	0.27	0.85
Pericarditis	7,118	0.39	1433.9	0.46	0.31	0.27	0.67
Telangectasis	6,257	0.34	2427.2	0.59	0.05	0.15	0.81
Nephritis	6,129	0.33	1645.7	0.46	0.18	0.26	0.53
Residues	5,036	0.27	1178.3	0.36	0.17	0.23	0.39
Pleurisy	4,476	0.24	1150.6	0.23	0.26	0.21	0.32
Hydronephrosis	3,806	0.21	1678.4	0.29	0.11	0.16	0.32
Oedema	3,729	0.20	2234.8	0.37	0.01	0.02	0.64

TABLE XIX: Summary of top 18 health conditions together with the number of animals having the condition, their mean life expectancy, and the percentage of occurrence of the condition on different sub-populations of the animals (female, male, beef, and dairy).

Sex/Temperature	Low	Medium	High	Population
Bulls	797.63	848.20	819.86	834.60
Females	1310.0	1314.21	1279.4	1313.0
Males	652.66	664.71	662.09	661.60

TABLE XX: Mean age in days by sex for different temperatures. Low is less than 6.08°C, medium is between 6.08°C and 12.76°C and high is above 12.76°C.