

Faculté des sciences

Analysis of the career length of professional basketball players by means of survival analysis

Author: Clément Tillieux

Supervisor: Ingrid Van Keilegom

Reader: Anouar El Gouch

Academic year 2025-2026

Contents

1	Introduction	5
2	Survival analysis	7
2.1	Basics	7
2.2	Non-parametric models	8
2.2.1	Kaplan-Meier estimator	8
2.3	Proportional Hazards Model (PHM)	9
2.4	Model building strategy	10
2.5	Proportional Hazards Assumption	11
2.6	Survival prediction (without time-dependent covariates)	11
2.6.1	Cox regression	12
2.6.2	Random Survival Forests	13
2.6.3	Survival Support Vector Machine	15
2.6.4	Model performance	22
2.7	Survival prediction (with time-dependent covariates)	23
3	First dataset	25
3.1	Creation of the dataset	25
3.2	Descriptive analysis	26
3.2.1	Time variables	26
3.2.2	Censoring indicator	27
3.2.3	Quantitative covariates	27
3.2.4	Qualitative covariates	29
3.3	Methods & Results	32
3.3.1	Kaplan-Meier	32
3.3.2	Proportional Hazards Model (PHM)	35
3.3.3	Proportional Hazards Assumption (PHA)	38
3.3.4	Survival prediction	41
3.4	Discussion	46
4	Second dataset	49
4.1	Creation of the dataset	49
4.1.1	Player selection	49
4.1.2	Variable selection	50
4.1.3	Final variables	51
4.2	Descriptive analysis	52
4.2.1	Time variables	52
4.2.2	Censoring indicator	53
4.2.3	Quantitative covariates	54

4.2.4	Qualitative covariates	57
4.2.5	Time-dependent covariates	60
4.3	Methods & Results	61
4.3.1	Kaplan-Meier	61
4.3.2	Proportional Hazards Model (PHM)	66
4.3.3	Proportional Hazards Assumption (PHA)	67
4.3.4	Survival prediction	70
4.4	Discussion	71
5	Global discussion	74
5.1	Outcomes	74
5.1.1	Which covariates have strong associations with career length ? .	74
5.1.2	The measures of career length	75
5.1.3	The focus on elite players	75
5.1.4	RSF and SSVM	76
5.2	Limitations	76
5.2.1	Optimization process of the SSVM	76
5.2.2	Measure of career length	76
5.2.3	Generalization	77
5.3	Future research	77
5.4	Conclusion	78
	Appendices	79
A	List of the packages	79
B	List of players who died before the end of their career	79
C	Basketball related information	80
D	Summary statistics (first dataset)	80
E	Alternate Kaplan-Meier survival curves (first dataset)	82
F	Alternate reference level comparison for covariate <i>pos</i> (first dataset)	84
G	Stratified Cox models (first dataset)	84
H	Predictive performance of stratified Cox models (first dataset)	84
I	Alternate optimization plots (first dataset)	85
J	Alternate Brier score plots	86
K	Players in the second dataset	87

1. Introduction

The career length of a professional basketball player is uncertain, as it is for many professional athletes. This uncertainty is caused, among others, by the inherent competitiveness of professional sports, by injuries and health issues, or by team decisions. Identifying the factors that influence the career length of professional basketball players would reduce this uncertainty. This thesis aims at exploring the associations between the career length and many different covariates. To do so, we use survival analysis, an area of statistics that studies time until an event happens.

Career length is an important topic in professional sports because it concerns everyone. Players want to play as long as they can. The goal of health professionals (physiotherapists, psychologists, etc.) is to help players reach that goal. Front offices (owners, presidents, general managers, etc.) want the players to play as long as possible for financial reasons. In the literature, many studies have investigated the causes of career termination in professional sports (Agresta et al., 2008; Lavalée et al., 1997; Ristolainen et al., 2011). Their objective was to determine which variables are responsible for, or at least related to, retirement among athletes. Although both subjects are related, retirement and career length are distinct concepts. However, other studies have explored career length in professional sports (Baker et al., 2012; Jones et al., 2025; Petersen et al., 2010; Steingröver et al., 2016). Some of them even focused on basketball (Martin et al., 2021; Miguel et al., 2019; Vretaros, 2021, 2022), but very few have applied survival analysis techniques to model the career length.

Survival analysis has significantly progressed during the 20th century (Fleming and Lin, 2000). The goal is to model the time until an event of interest occurs. Therefore, survival analysis has been used in many different fields: medicine (time until the death of a patient), pharmacology (time until recovery), engineering (time until a machine breaks down), etc. It has already been applied to model the career length (Bounnite, 2025; Macran et al., 1996), even in the context of professional sports. For instance, Zaman et al. (2025) have used survival analysis to study the career length of professional cricket players. However, to our knowledge, very few studies have applied survival analysis techniques to model the career length of professional basketball players. We only found Fynn and Sonnenschein (2012), who used survival analysis to explore the association between the career length of all NBA players until 2011 and four covariates: height, weight, playing position and the number of individual trophies.

In this thesis, we try to model the career length of basketball players through the use of two different datasets. The first dataset is very large and was created by Martínez et al. (2022). The second one focuses on elite players and allows for time-dependent covariates. There are very few studies that have combined the use of survival analysis and time-dependent variables to study career length of professional athletes, and as far

as we are aware, this thesis is the first one to do so for basketball players. The use of time-varying covariates was also suggested by Fynn and Sonnenschein (2012). In addition, we investigate lesser-known techniques adapting famous regression/classification techniques such as Random Forests (Breiman, 2001) and Support Vector Machines (Vapnik, 1995), to the specific area of survival analysis. Once again, to our knowledge, this thesis is the first application of Random Survival Forests (Ishwaran et al., 2008) and Survival Support Vector Machines (Shivaswamy et al., 2007; Van Belle et al., 2007, as cited in Fouodo et al., 2018; Van Belle et al., 2008, 2011) to the context of career length of basketball players.

The main objective of this thesis is to find which covariates have strong associations with the career length of basketball players. However, we also have smaller goals: to determine the best way to measure career length, to assess if the use of time-dependent covariates is pertinent, to see if separating elite players from others changes the outcomes, and to evaluate the performance of the two adapted regression techniques.

In section 2, we develop the theory. We start by explaining the basics of survival analysis, and progressively move toward the more complex regression techniques. We also present two criteria that we use to assess the predictive performance of the tested models. In sections 3 and 4, we use survival analysis to explore two datasets. The first dataset contains every basketball player who played at least one game in the NBA, while the second dataset focuses only on the best players in the history of the league. For each dataset, we start with a descriptive analysis of the covariates, then we use both non-parametric and semi-parametric models to explore the association between career length and the covariates, and finally, we assess the predictive performance of different models. Lastly, in section 5, we conclude with a discussion where we compare the outcomes of the two datasets, and we discuss the limitations of this thesis and the opportunities for future studies.

2. Survival analysis

This section is based mainly on Klein and Moeschberger (2003) and Van Keilegom (2025). Annex A contains the list of all the packages we used to implement these techniques.

2.1. Basics

The goal of survival analysis is to model time-to-event data and to explore its relation with a set of covariates.

Let X be a non-negative, continuous random variable representing the time until an event of interest occurs. In this study, we are interested in the length of a basketball player's career. For example, X could be the number of seasons in a player's career. Though the number of seasons may appear to be a discrete variable, the underlying quantity measured is time, which is continuous. It could instead have been measured in the number of games or minutes played. The number of seasons can be seen as a rounded (or discretized) measure of this underlying continuous time. Furthermore, there are many different possible values for the number of seasons. This is another reason why we can consider the number of seasons or any other discrete measure of time as a continuous variable. As X is a random variable, it has a Cumulative Distribution Function (CDF) $F(x) = P(X \leq x)$. In our example, it is the probability that an athlete plays x seasons or less during his career. In addition, survival analysis uses specific functions to characterize the data. Three of them will be particularly useful:

- The Survival Function $S(x) = 1 - F(x) = P(X > x)$. It is the probability that an athlete plays more than x seasons during his career.
- The Cumulative Hazard Function $H(x) = -\log(S(x))$.
- The Hazard Function $h(x) = \lim_{\Delta x \rightarrow 0} \frac{P(x \leq X < x + \Delta x \mid X \geq x)}{\Delta x} = \frac{\partial H(x)}{\partial x}$. It is the risk that a player who has already played x seasons retires right after the end of his x^{th} season.

Another major concept in survival analysis is censoring. When the time to event is known only to lie within an interval, rather than being observed exactly, the observation is said to be censored. There is more than one type of censoring, but in this study, we only encounter random right censoring. "Right censoring" means that we only know the lower bound of the interval that contains the true value of the survival time. For instance, right-censored observations may correspond to players who have not yet retired by the end of the observation period, when data collection is stopped. In this case, the exact end of their career is unknown, but we know that it will happen after the last season for which we collected data. Therefore, what we have is not the survival time, but the censoring time. "Random censoring" means that the censoring time and the survival

time are independent of each other. We will use the following notation from Klein and Moeschberger (2003). Given X , the survival time, and C , the censoring time, the data can be represented as a pair (T, δ) , where $T = \min(X, C)$ is the "recorded" time to event and δ is a binary variable with value 0 when the observation is censored (when $X > C$) and value 1 otherwise (when $X \leq C$).

2.2. Non-parametric models

2.2.1. Kaplan-Meier estimator

One might want to estimate the survival function based on the data. Kaplan and Meier (1958) proposed a non-parametric estimator for right-censored data. Let $t_1, \dots, t_M$ be the M distinct ordered event times, $r(t_i)$ be the number of observations at risk of experiencing the event at time t_i and $m(t_i)$ be the number of observations actually experiencing the event at time t_i (without including the censored observations). The Kaplan-Meier estimator is:

$$\hat{S}(t) = \prod_{t_i \leq t} \frac{r(t_i) - m(t_i)}{r(t_i)} \quad (1)$$

It is a decreasing step function with value 1 when $t = 0$ and, if the observation with the largest recorded time is not censored, value 0 when $t = t_M$. If the observation corresponding to the largest recorded time t_M is censored, then $\hat{S}(t)$ never reaches value 0, and it is impossible to estimate the survival function consistently for $t \geq t_M$. One way to handle this issue is to define $\hat{S}(t) = \hat{S}(t_M)$ for all $t \geq t_M$.

If one is interested in comparing the survival functions of V different groups (based on variable k), it is possible to estimate each of the V survival functions using only the observations in the corresponding group. The Kaplan-Meier estimator is computed in the same way as in (1), but only observations with value v on variable k are considered. For example, we could compare the estimated survival function of right- and left-handed basketball players. To do so, one could compute two Kaplan-Meier estimators: one using only the right-handed players in the dataset, and the other using only the left-handed players. If the variable of interest is continuous, it needs to be discretized (transformed into a finite number of groups). If variable k is time-dependent, one cannot use the Kaplan-Meier estimator to compare the survival functions of the different groups because the same person can change groups through time. For instance, a player could weigh 80 kg at the beginning of his career, then gain weight until reaching 100 kg. Consequently, the player's group membership would change during follow-up, whereas the standard Kaplan-Meier estimator assumes fixed group membership. Therefore, simply defining Kaplan-Meier curves according to the value of k at each observation would not provide a valid comparison of the survival functions between groups.

2.3. Proportional Hazards Model (PHM)

In this study, we are interested in exploring the relation between the career length of a basketball player and a set of variables, rather than a single one. For example, we could examine whether there is an association between career length and dominant hand, nationality, and/or height. One way to do so is to use the proportional hazards model (Cox, 1972) or its extension to account for time-dependent covariates.

Let $\mathbf{d}_j(t) = [d_{j1}(t) \dots d_{jp}(t)]^T$ be the vector of covariates at time t for the j^{th} observation. $d_{jk}(t)$ may be a time-dependent covariate, meaning that $d_{jk}(t_1)$ may be different from $d_{jk}(t_2)$ for $t_1 \neq t_2$. The extension of Cox's model for time-dependent covariates is:

$$h(t | \mathbf{d}(t)) = h_0(t) \exp(\boldsymbol{\beta}^T \mathbf{d}(t)) = h_0(t) \exp\left(\sum_{k=1}^p \beta_k d_k(t)\right)$$

where $\boldsymbol{\beta} = [\beta_1 \dots \beta_p]^T$ is the vector of unknown parameters and $h_0(t)$ is the baseline hazard rate. If we have two observations with covariates $\mathbf{d}_1(t)$ and $\mathbf{d}_2(t)$, the ratio of their hazard functions is:

$$\frac{h(t | \mathbf{d}_1(t))}{h(t | \mathbf{d}_2(t))} = \frac{h_0(t) \exp(\sum_{k=1}^p \beta_k d_{1k}(t))}{h_0(t) \exp(\sum_{k=1}^p \beta_k d_{2k}(t))} = \exp\left(\sum_{k=1}^p \beta_k (d_{1k}(t) - d_{2k}(t))\right)$$

If the two observations differ only in covariate d_k at time t , the hazard ratio becomes $\exp(\beta_k (d_{1k}(t) - d_{2k}(t)))$, and it leads to the following interpretations of the parameters:

- If covariate d_k is binary, $\exp(\beta_k)$ is the risk of experiencing the event of interest at time t for observation j_1 who has value 1 for covariate d_k (at time t) relative to observation j_2 who has value 0 on it (at time t). For instance, if covariate d_k is the dominant hand, $\exp(\beta_k)$ is the risk of the right-handed player j_1 retiring relative to the left-handed player j_2 retiring.
- If covariate d_k is discrete with V levels, it will be encoded as $V - 1$ binary variables and each is interpreted following the same logic as above. One level of the variable will be used as the "reference level". For instance, if we have players from Belgium, France and Germany, we create variables *France* and *Germany* and use the level *Belgium* as the reference category (value 0 on both new covariates). Now we have two binary covariates.
- If covariate d_k is continuous, $\exp(\beta_k)$ is the risk of experiencing the event of interest at time t for observation j_1 relative to observation j_2 , given that observation j_1 's value for covariate d_k (at time t) is one unit higher than observation j_2 's. For instance, if covariate d_k is height, $\exp(\beta_k)$ is the risk of player j_1 retiring at time t relative to the risk of player j_2 retiring at time t , given j_1 is 1 cm taller than j_2 at time t .

As mentioned earlier, covariates may or may not be time-dependent. For instance, dominant hand is not time-dependent. A player's dominant hand remains fixed throughout his career. However, height can be time-dependent since a player can be taller at 30 than at 20.

One can obtain an estimate of the parameter vector β by maximizing the likelihood or its corresponding log-likelihood (Efron, 1977). Given $t_1, \dots, t_M$, the M distinct ordered event times, let $\mathbb{M}(t_i)$ be the set of observations that experience the event at time t_i (without including the censored observations) and $m(t_i)$ be the number of observations in this set. Let $q(t_i) = \sum_{j \in \mathbb{M}(t_i)} \mathbf{d}_j(t_i)$ be the sum of the covariate vectors $\mathbf{d}_j(t_i)$ for all observations j who experience the event at time t_i . Let $R(t_i)$ be the set of all observations at risk of experiencing the event at time t_i . It includes all observations that have neither experienced the event nor been censored just before time t_i . Efron's partial likelihood is:

$$L(\beta) = \prod_{i=1}^M \frac{\exp(\beta^T q(t_i))}{\prod_{w=1}^{m(t_i)} \left[\sum_{j \in R(t_i)} (\exp(\beta^T \mathbf{d}_j(t_i))) - \frac{w-1}{m(t_i)} \sum_{j \in \mathbb{M}(t_i)} (\exp(\beta^T \mathbf{d}_j(t_i))) \right]} \quad (2)$$

The likelihood is called "partial" because we do not use the entire likelihood, but only the necessary part to estimate β . One solution to maximize this function is to use iterative methods on its logarithmic form $\ell(\beta) = \log(L(\beta))$. Newton-Raphson is a widely used iterative method based on the following algorithm (De Vera, n.d.):

$$\hat{\beta}_{new} = \hat{\beta}_{old} - H^{-1}(\hat{\beta}_{old}) \nabla \ell(\hat{\beta}_{old})$$

where $\nabla \ell(\beta) = \left[\frac{\partial \ell(\beta)}{\partial \beta_1} \dots \frac{\partial \ell(\beta)}{\partial \beta_p} \right]^T$ is the gradient vector and $H(\beta)$ is the Hessian matrix with elements $H_{kl} = \frac{\partial^2 \ell(\beta)}{\partial \beta_k \partial \beta_l}$ for $k, l = 1, \dots, p$. One has to determine an initial value for $\hat{\beta}$ and then iterate until convergence is reached, meaning until $\|\hat{\beta}_{new} - \hat{\beta}_{old}\| < g$, where g is a fixed, small value.

2.4. Model building strategy

One might wonder which covariates to keep in the model in order to avoid dimensionality issues (overfitting, model complexity, curse of dimensionality, among others). Many strategies exist, but we will only use AIC selection. It is a stepwise search procedure based on Akaike's Information Criterion (AIC) (Akaike, 1998). Given $L(\beta)$ the partial likelihood as computed in (2), and p the number of covariates, the AIC is:

$$AIC = -2 \log(L(\beta)) + 2p$$

There are three steps in the AIC selection:

1. Start with the full model, meaning all covariates are included.

2. Compute the AIC for the current model. Then compute the AIC for every model obtained by dropping one covariate currently in the model (if any remain), and for every model obtained by adding back a covariate that was previously dropped (if any were dropped).
3. If the current model has the smallest AIC, stop the selection. Otherwise, update the current model to the one with the smallest AIC, and return to step 2.

2.5. Proportional Hazards Assumption

In the PHM, we assume that the ratio of the hazard functions of two observations is constant over time. It is called the proportional hazards assumption (PHA). One may want to check whether this assumption holds for the data at hand. One way to do so is to use plots and tests based on Schoenfeld residuals (Schoenfeld, 1982). Given $\hat{\beta}$, the maximum partial likelihood estimator of the parameters, the Schoenfeld residual for observation j and parameter k is:

$$r_{jk} = \delta_j \left[d_{jk}(t_j) - \frac{\sum_{i \in R(t_j)} d_{ik}(t_j) \exp(\hat{\beta}^T \mathbf{d}_i(t_j))}{\sum_{i \in R(t_j)} \exp(\hat{\beta}^T \mathbf{d}_i(t_j))} \right]$$

where t_j is the time when observation j experiences the event and $R(t_j)$ is the set of all observations at risk of experiencing the event at time t_j . Once the residuals for all parameters and observations are computed, one can check the PHA in two ways: by inspecting the p plots of Schoenfeld residuals against time, each with a smoothed regression curve (one plot per covariate), or by formally testing the null hypothesis of a zero-slope regression line for each covariate. If the assumption is violated, the estimated hazard ratios should be interpreted with caution, as they may not be constant over the follow-up period.

One way to deal with the violation of the Proportional Hazards Assumption is to use a stratified model. In a model stratified on discrete covariate d_k , Cox's model becomes:

$$h(t | \mathbf{d}(t), d_k = v) = h_{0,v}(t) \exp(\beta^T \mathbf{d}(t))$$

where $\mathbf{d}(t) = [d_1(t) \dots d_{p-1}(t)]^T$ does not include the stratified covariate. In other words, we allow a different baseline hazard rate for each stratum (or level v of covariate d_k), but the regression coefficients of the $p - 1$ other covariates on the hazard function are assumed to be the same across strata (Klein and Moeschberger, 2003).

2.6. Survival prediction (without time-dependent covariates)

Once the estimator $\hat{\beta}$ is determined, one might be interested in estimating the survival of a new observation $\mathbf{z}(t) = [z_1 \dots z_p]^T$. In this part, we will only use time-invariant

covariates, meaning that $\mathbf{d}(t) = \mathbf{d}$ (or $\mathbf{z}(t) = \mathbf{z}$) for all t . This is particularly relevant in practice. For instance, a team manager may want to estimate how long a given player's career will last. There exist many methods, but we will only focus on three of them: a semi-parametric method (Cox regression), a tree-based ensemble (Random Survival Forests), and a margin-based method (Survival Support Vector Machine). The first one builds directly on the previous sections (2.3 to 2.5). The second was selected for its robustness to PHA violations, and the third for its ability to capture non-linear effects. Moreover, many studies have shown that these methods work well in general (Fouodo et al., 2018; Ishwaran et al., 2008; Kvamme et al., 2019; Shivaswamy et al., 2007).

2.6.1. Cox regression

First, one has to determine an estimate of the baseline survival function $\hat{S}_0(t)$. It is the estimated survival probability of an observation with value 0 on all covariates ($\mathbf{d} = \mathbf{0}$). Let $t_1, \dots, t_M$ be the M distinct ordered event times, $R(t_i)$ be the set of all observations at risk of experiencing the event at time t_i and $m(t_i)$ be the number of observations who actually experience the event at time t_i (without including the censored observations). Klein and Moeschberger (2003) proposed the following formula, based on Breslow's estimator (Breslow, 1972):

$$\hat{S}_0(t) = \exp \left(- \sum_{t_i \leq t} \frac{m(t_i)}{\sum_{j \in R(t_i)} \exp(\hat{\beta}^T \mathbf{d}_j)} \right)$$

Next, to estimate the survival function of a new observation with covariate vector $\mathbf{z} = [z_1 \dots z_p]^T$, one has to use the estimator:

$$\hat{S}(t | \mathbf{z}) = \hat{S}_0(t)^{\exp(\hat{\beta}^T \mathbf{z})}$$

In section 2.6.4, we will need a way to determine which of two observations j_1 and j_2 has the worst prediction outcome. In the case of Cox regression, we will use the risk score (Branders et al., 2015; Klein and Moeschberger, 2003). The risk score of observation j can be computed as:

$$RS_j = \hat{\beta}^T \mathbf{d}_j$$

Since a higher risk score corresponds to a higher hazard (and therefore, all else equal, a shorter expected survival time), observation j_1 is said to have a worse prediction outcome than observation j_2 if:

$$RS_{j_1} > RS_{j_2} \tag{3}$$

2.6.2. Random Survival Forests

This method was developed by Ishwaran et al. (2008), based on the existing Random Forests method (Breiman, 2001). The global idea of Random Survival Forests (RSF) is to build a set of decision trees based on subsamples of the original dataset.

First, draw B bootstrap samples from the data. A bootstrap sample is made by randomly drawing a specific number of observations from the original dataset, with replacement. It means that one observation can be drawn multiple times in one bootstrap sample.

Second, for each sample, grow a survival tree. Starting from a node containing the entire sample, a tree is grown by recursively splitting each node into two daughter nodes (based on a splitting criterion) until a stopping criterion is met. Thus, one needs two things to grow a survival tree: a splitting criterion and a stopping criterion. Even though there exist many criteria, in this study, we will use the log-rank splitting rule and the minimum number of observations in a terminal node, respectively.

In the log-rank splitting rule, one tests all the covariates and a finite number of possible values on each covariate to split a given node α into two daughter nodes. Given covariate d_k and split value v , the left and right daughters are $L = \{j : d_{jk} \leq v\}$ and $R = \{j : d_{jk} > v\}$, the set of observations in α who have a value respectively lower or strictly higher than v for covariate d_k .

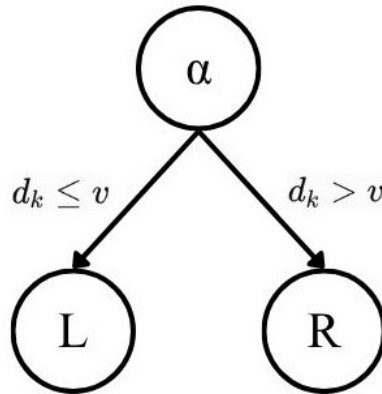


Figure 1: Illustration of a node splitting in RSF

Let $t_1, \dots, t_{M_\alpha}$ be the M_α distinct times when the event was experienced in node α (without including the censoring times), $m_L(t_i)$ and $m_R(t_i)$ be the number of observations who experience the event at time t_i (without including the censored observations) in the left and right daughter nodes respectively, $r_L(t_i)$ and $r_R(t_i)$ be the number of observations at risk of experiencing the event at time t_i in the left and right daughter nodes respectively, $m_\alpha(t_i) = m_L(t_i) + m_R(t_i)$ and $r_\alpha(t_i) = r_L(t_i) + r_R(t_i)$. The log-rank

split-statistic value for a split is:

$$LR(d_k, v) = \frac{\sum_{i=1}^{M_\alpha} \left(m_L(t_i) - r_L(t_i) \frac{m_\alpha(t_i)}{r_\alpha(t_i)} \right)}{\sqrt{\sum_{i=1}^{M_\alpha} \frac{r_L(t_i)}{r_\alpha(t_i)} \left(1 - \frac{r_L(t_i)}{r_\alpha(t_i)} \right) \left(\frac{r_\alpha(t_i) - m_\alpha(t_i)}{r_\alpha(t_i) - 1} \right) m_\alpha(t_i)}}$$

The log-rank splitting rule aims at finding covariate d_k^* and splitting value v^* such that $|LR(d_k^*, v^*)| \geq |LR(d_k, v)|$ for all d_k and v . As $|LR(d_k, v)|$ is a measure of node separation, maximizing it means getting the best possible split of the data for a given node.

One also has to choose the minimum number of observations in a terminal node n_θ . After a split, both the left and right daughter nodes must have a number of observations n_L or n_R greater than or equal to n_θ . If it is not the case, then the split is cancelled, even if it maximizes the absolute value of the log-rank split-statistic. If a node has a number of observations $n_\alpha < 2 * n_\theta$, then it cannot be split anymore and a terminal node has been reached.

Third, once a survival tree is fully grown, the Nelson-Aalen Cumulative Hazard Function (CHF) estimator (Nelson, 1972; Aalen, 1978) and the Kaplan-Meier Survival Function estimator (Kaplan and Meier, 1958) can be computed for each terminal node. Given a terminal node θ , let $t_1, \dots, t_{M_\theta}$ be the M_θ distinct times when the event was experienced in node θ (without including the censoring times), $m_\theta(t_i)$ be the number of observations in node θ who experience the event at time t_i (without including the censored observations) and $r_\theta(t_i)$ be the number of observations in node θ at risk of experiencing the event at time t_i . The Nelson-Aalen and Kaplan-Meier estimators for node θ are:

$$\hat{H}_\theta(t) = \sum_{t_i \leq t} \frac{m_\theta(t_i)}{r_\theta(t_i)} \quad \hat{S}_\theta(t) = \prod_{t_i \leq t} \frac{r_\theta(t_i) - m_\theta(t_i)}{r_\theta(t_i)}$$

For observation j with covariate vector $\mathbf{d}_j$, one needs to propagate observation j down the tree based on the values in $\mathbf{d}_j$. Because of the binary nature of the tree, $\mathbf{d}_j$ will arrive at a unique terminal node θ . The estimators of the CHF and Survival Function for j are the Nelson-Aalen and Kaplan-Meier estimators for its corresponding terminal node θ :

$$\hat{H}(t | \mathbf{d}_j) = \hat{H}_\theta(t), \text{ for } \mathbf{d}_j \in \theta \quad \hat{S}(t | \mathbf{d}_j) = \hat{S}_\theta(t), \text{ for } \mathbf{d}_j \in \theta \quad (4)$$

Fourth, the estimators in (4) are computed for a single tree. To get the ensemble estimator, one has to average over the B survival trees created from each bootstrap sample drawn in the first step. Given that $\hat{H}_b(t | \mathbf{d}_j)$ and $\hat{S}_b(t | \mathbf{d}_j)$ are, respectively, the CHF and Survival Function estimators of observation j for the b^{th} tree, the ensemble estimators are simply the averages over all the survival trees:

$$\hat{H}^E(t | \mathbf{d}_j) = \frac{1}{B} \sum_{b=1}^B \hat{H}_b(t | \mathbf{d}_j) \quad \hat{S}^E(t | \mathbf{d}_j) = \frac{1}{B} \sum_{b=1}^B \hat{S}_b(t | \mathbf{d}_j)$$

In section 2.6.4, we will need a way to determine which of two observations j_1 and j_2 has the worst prediction outcome. In RSF, the concept of mortality is used. The mortality of an observation is a summary of its CHF, obtained by summing the estimated CHF over all distinct event times (Ishwaran et al., 2008). The mortality is the predicted value used by the RSF. It measures the number of deaths expected under the hypothesis that all observations are similar to j . It can be estimated by the ensemble mortality. Given $t_1, \dots, t_M$ the M distinct event times for the entire dataset, the ensemble mortality for observation j is:

$$\bar{M}^E(\mathbf{d}_j) = \sum_{i=1}^M \hat{H}^E(t_i | \mathbf{d}_j)$$

Observation j_1 is said to have a worse prediction outcome than observation j_2 if:

$$\bar{M}^E(\mathbf{d}_{j_1}) > \bar{M}^E(\mathbf{d}_{j_2}) \quad (5)$$

2.6.3. Survival Support Vector Machine

Support Vector Machine (SVM) was introduced by Vapnik (1995) as a binary classification model. Let $\mathbf{d} = [d_1 \dots d_p]^T$ be a vector of covariates. The goal is to find the hyperplane $h(\mathbf{d}) = \beta^T \mathbf{d} + \beta_0 = 0$ (no connection with the hazard function) that best separates the two classes of the target variable $Y \in \{-1; +1\}$. If the n observations are linearly separable, then there exists a function $f(\mathbf{d}) = \beta^T \mathbf{d} + \beta_0$ such that $y f(\mathbf{d}) > 0$. The distance between a given observation $\mathbf{d}$ and the hyperplane is defined as $\Delta = \frac{y h(\mathbf{d})}{\|\beta\|}$. The margin is the minimum distance between an observation and the hyperplane:

$$\Delta^* = \min_{\mathbf{d}_j} \left\{ \frac{y_j (\beta^T \mathbf{d}_j + \beta_0)}{\|\beta\|} \right\}$$

All observations $\mathbf{d}^*$ that are located at distance Δ^* from the hyperplane are called support vectors. As there exists an infinite number of hyperplane equations ($s h(\mathbf{d}) = s(\beta^T \mathbf{d} + \beta_0)$, where s is a scalar), another condition is that the margin must be $\Delta^* = \frac{1}{\|\beta\|}$ (or $s = \frac{1}{y^* (\beta^T \mathbf{d} + \beta_0)}$). It also means that a support vector can be defined as any observation $\mathbf{d}^*$ where $\beta^T \mathbf{d} + \beta_0 = 1$ or $\beta^T \mathbf{d} + \beta_0 = -1$. Figure 2 is an illustration of SVM in two dimensions for a dataset that is linearly separable.

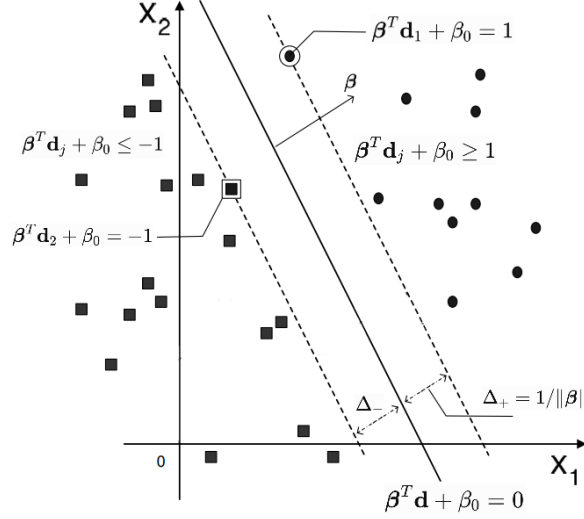


Figure 2: Illustration of the SVM for a linearly separable case.
Adapted from Hafner (2026, Slide 104).

To find the hyperplane that best separates the two classes, SVM searches for the hyperplane h^* maximizing the margin:

$$h^* = \arg \max_{\beta, \beta_0} \left\{ \frac{1}{\|\beta\|} \right\}$$

or equivalently:

$$\begin{aligned} \min_{\beta, \beta_0} \quad & \frac{1}{2} \|\beta\|^2 \\ \text{constraints:} \quad & y_j(\beta^T \mathbf{d}_j + \beta_0) \geq 1, \text{ for } j = 1, \dots, n \end{aligned} \quad (6)$$

In order to achieve optimization, we need to introduce the Lagrange multiplier α_j to turn the constrained optimization (6) into the unconstrained one (7). The new objective function is the Lagrangian:

$$La = \frac{1}{2} \|\beta\|^2 - \sum_{j=1}^n \alpha_j (y_j(\beta^T \mathbf{d}_j + \beta_0) - 1) \quad (7)$$

La needs to be minimized with respect to β and β_0 , and maximized with respect to the α_j 's.

$$\begin{aligned} \frac{\partial La}{\partial \beta} = \beta - \sum_{j=1}^n \alpha_j y_j \mathbf{d}_j = 0 & \Leftrightarrow \beta = \sum_{j=1}^n \alpha_j y_j \mathbf{d}_j \\ \frac{\partial La}{\partial \beta_0} = \sum_{j=1}^n \alpha_j y_j = 0 & \end{aligned} \quad (8)$$

Incorporating (8) into (7) gives the dual Lagrangian, which only depends on $\alpha = [\alpha_1 \dots \alpha_n]^T$.

$$\begin{aligned} \max_{\alpha} \quad & La_{dual} = \sum_{j=1}^n \alpha_j - \frac{1}{2} \sum_{j_1=1}^n \sum_{j_2=1}^n \alpha_{j_1} \alpha_{j_2} y_{j_1} y_{j_2} \mathbf{d}_{j_1}^T \mathbf{d}_{j_2} \\ \text{constraints:} \quad & \alpha_j \geq 0, \text{ for } j = 1, \dots, n \\ & \sum_{j=1}^n \alpha_j y_j = 0 \end{aligned} \quad (9)$$

It is a convex quadratic problem, which admits a unique solution α^* . Once α^* is known, one can easily compute β^* from (8) and β_0^* , from the definition of a support vector as:

$$\begin{aligned} \beta^* &= \sum_{j=1}^n \alpha_j^* y_j \mathbf{d}_j \\ \beta_0^* &= \frac{\sum_{\alpha_j^* > 0} \beta_{0j}^*}{\sum_{\alpha_j^* > 0} 1}, \text{ where } \beta_{0j}^* = y_j - (\beta^*)^T \mathbf{d}_j \end{aligned}$$

Finally, one can predict the class of a new observation $\mathbf{z} = [z_1 \dots z_p]^T$ by using hyper-plane function h^* :

$$\hat{y} = \text{sign}(h^*(\mathbf{z})) = \text{sign}((\beta^*)^T \mathbf{z} + \beta_0^*)$$

In most datasets, the data is not perfectly linearly separable. To handle it, one has to introduce the slack variable ξ_j . The ξ_j 's represent how much each observation violates the separability condition. If the observation is on the right side of the hyperplane and at least $\frac{1}{\|\beta\|}$ away from it, then $\xi_j = 0$. If the observation is on the right side of the hyperplane but inside the margin, then $0 < \xi_j < 1$. Finally, if the observation is on the wrong side of the hyperplane, then $\xi_j \geq 1$. Figure 3 is an illustration of SVM in two dimensions for a dataset that is linearly non-separable.

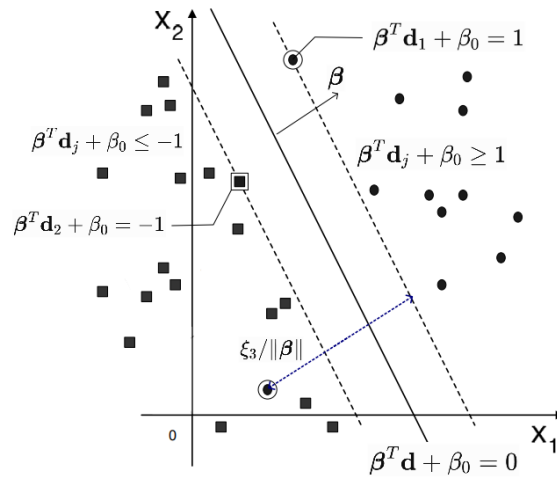


Figure 3: Illustration of the SVM for a linearly non-separable case.
Adapted from Hafner (2026, Slide 118).

Introducing the slack variables $\xi = [\xi_1 \dots \xi_n]^T$ in (6) leads to the new objective function:

$$\begin{aligned} \min_{\beta, \beta_0, \xi} \quad & \frac{1}{2} \|\beta\|^2 + \gamma \sum_{j=1}^n \xi_j \\ \text{constraints:} \quad & y_j(\beta^T \mathbf{d}_j + \beta_0) \geq 1 - \xi_j, \text{ for } j = 1, \dots, n \\ & \xi_j \geq 0, \text{ for } j = 1, \dots, n \end{aligned} \quad (10)$$

where γ is a constant quantifying the trade-off between maximizing the margin and minimizing the loss. From there, one can use the Lagrangian and dual Lagrangian to compute $\hat{y}$ as in the perfectly linearly separable case.

The last important feature of classical SVM is data transformation. The general idea is that if the data is not linearly separable in the original p -dimensional input space of observations $\mathbf{d}_j$'s, one could find a function $\phi(\mathbf{d}_j)$ that maps every observation from the input space into a new p' -dimensional feature space. The hope is that the data might be linearly separable in the feature space. All operations shown before still work thanks to the kernel trick. With the kernel trick, one never directly uses the function $\phi(\mathbf{d}_j)$ to map every point into the feature space because it might be time-consuming and computationally intensive. Instead, one uses the kernel function $K(\mathbf{d}_i, \mathbf{d}_j) = \phi(\mathbf{d}_i)\phi(\mathbf{d}_j)$. It leads to the new objective function:

$$\begin{aligned} \min_{\beta, \beta_0, \xi} \quad & \frac{1}{2} \|\beta\|^2 + \gamma \sum_{j=1}^n \xi_j \\ \text{constraints:} \quad & y_j(\beta^T \phi(\mathbf{d}_j) + \beta_0) \geq 1 - \xi_j, \text{ for } j = 1, \dots, n \\ & \xi_j \geq 0, \text{ for } j = 1, \dots, n \end{aligned} \quad (11)$$

Similarly to what was done in (7) and (9), one can obtain the following dual Lagrangian:

$$\begin{aligned} \max_{\alpha} \quad & La_{dual} = \sum_{j=1}^n \alpha_j - \frac{1}{2} \sum_{j_1=1}^n \sum_{j_2=1}^n \alpha_{j_1} \alpha_{j_2} y_{j_1} y_{j_2} \phi(\mathbf{d}_{j_1}) \phi(\mathbf{d}_{j_2}) \\ & = \sum_{j=1}^n \alpha_j - \frac{1}{2} \sum_{j_1=1}^n \sum_{j_2=1}^n \alpha_{j_1} \alpha_{j_2} y_{j_1} y_{j_2} K(\mathbf{d}_{j_1}, \mathbf{d}_{j_2}) \\ \text{constraints:} \quad & 0 \leq \alpha_j \leq \gamma, \text{ for } j = 1, \dots, n \\ & \sum_{j=1}^n \alpha_j y_j = 0 \end{aligned} \quad (12)$$

Using the derivatives from (8), one can see that $\beta^* = \sum_{j=1}^n \alpha_j^* y_j \phi(\mathbf{d}_j)$ cannot be directly computed since there is no dot product. β_0^* , however, can be computed as:

$$\beta_{0j}^* = y_j - (\beta^*)^T \phi(\mathbf{d}_j) = y_j - \sum_{j_1=1}^n \alpha_{j_1} y_{j_1} K(\mathbf{d}_{j_1}, \mathbf{d}_j)$$

And finally, one can still predict the class of a new observation $\mathbf{z} = [z_1 \dots z_p]$:

$$\hat{y} = \text{sign}((\boldsymbol{\beta}^*)^T \phi(\mathbf{z}) + \beta_0^*) = \text{sign}\left(\sum_{j=1}^n (\alpha_j y_j K(\mathbf{d}_j, \mathbf{z})) + \beta_0^*\right)$$

Drucker et al. (1996) extended the SVM to apply it to cases where the target variable is continuous: the Support Vector Regression Machine (SVR). The general idea is that SVR fits a function $f(\mathbf{d})$ through the data and tolerates a margin of error around the fitted function (Radečić, 2026). This margin is called the epsilon tube. All the observations inside the epsilon tube are considered close enough and SVR only uses the other observations to define the model (Radečić, 2026). These outside points are the support vectors in SVR. Figure 4 is an illustration of SVR in two dimensions.

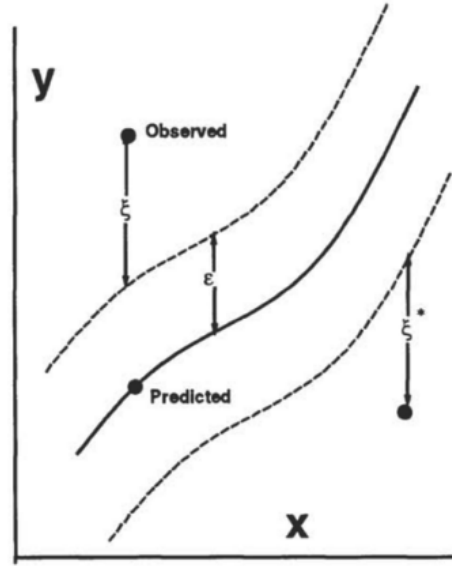


Figure 4: Illustration of the SVR. Source: Drucker et al. (1996, p. 161, Fig. 1).

Mathematically, SVR introduces two sets of slack variables ξ_j 's and ξ_j^* 's. If observation $\mathbf{d}_j$ is inside the epsilon tube, then $\xi_j = 0$ and $\xi_j^* = 0$. If the observation is above the margin, meaning that $y_j > f(\mathbf{d}_j) + \varepsilon$, then $\xi_j > 0$ and $\xi_j^* = 0$. And if the observation is below the margin, meaning that $y_j < f(\mathbf{d}_j) - \varepsilon$, then $\xi_j^* > 0$ and $\xi_j = 0$. Given $\boldsymbol{\xi} = [\xi_1 \dots \xi_n]^T$ and $\boldsymbol{\xi}^* = [\xi_1^* \dots \xi_n^*]^T$, the kernel trick still allows us to work with transformed data and the new objective function is:

$$\begin{aligned} \min_{\boldsymbol{\beta}, \beta_0, \boldsymbol{\xi}, \boldsymbol{\xi}^*} \quad & \frac{1}{2} \|\boldsymbol{\beta}\|^2 + \gamma \sum_{j=1}^n (\xi_j + \xi_j^*) \\ \text{constraints:} \quad & y_j - (\boldsymbol{\beta}^T \phi(\mathbf{d}_j) + \beta_0) \leq \varepsilon + \xi_j, \text{ for } j = 1, \dots, n \\ & (\boldsymbol{\beta}^T \phi(\mathbf{d}_j) + \beta_0) - y_j \leq \varepsilon + \xi_j^*, \text{ for } j = 1, \dots, n \\ & \xi_j \geq 0, \text{ for } j = 1, \dots, n \\ & \xi_j^* \geq 0, \text{ for } j = 1, \dots, n \end{aligned}$$

From there, one can again use the Lagrangian and dual Lagrangian to compute $\hat{y}$.

Shivaswamy et al. (2007) have extended the field of SVR to censored targets with the regression approach. Recall that the data is represented as $(\mathbf{d}_j, t_j, \delta_j)$, for $j = 1, \dots, n$, where $\mathbf{d}_j = [d_{j1} \dots d_{jp}]^T$ is the vector of covariates for observation j , $t_j = \min(x_j, c_j)$ the "recorded" time to event for observation j , and δ_j is a binary variable with value 0 if observation j is censored (when $x_j > c_j$) and value 1 otherwise (when $x_j \leq c_j$). As formulated by Fouodo et al. (2018), the new objective function is:

$$\begin{aligned} \min_{\beta, \beta_0, \xi, \xi^*} \quad & \frac{1}{2} \|\beta\|^2 + \gamma \sum_{j=1}^n (\xi_j + \xi_j^*) \\ \text{constraints:} \quad & t_j - (\beta^T \phi(\mathbf{d}_j) + \beta_0) \leq \xi_j, \text{ for } j = 1, \dots, n \\ & \delta_j ((\beta^T \phi(\mathbf{d}_j) + \beta_0) - t_j) \leq \xi_j^*, \text{ for } j = 1, \dots, n \\ & \xi_j \geq 0, \text{ for } j = 1, \dots, n \\ & \xi_j^* \geq 0, \text{ for } j = 1, \dots, n \end{aligned}$$

Notice that, due to the nature of censored data, the ε value disappears from the constraints. Recall that the epsilon tube represents a margin of tolerance between the true target value and its estimate. In censored data, the true value is unknown. Therefore, since there are no true values around which to build this tolerance interval, the epsilon tube loses its meaning. From there, one can again use the Lagrangian and dual Lagrangian to compute $\hat{y}$.

Another way to solve survival problems using the SVM is the ranking approach (Van Belle et al., 2007, as cited in Fouodo et al., 2018; Van Belle et al., 2008). Given two observations j_1 and j_2 , where at least j_1 experiences the event of interest, their classical SVM predictions are $\beta^T \phi(\mathbf{d}_{j_1}) + \beta_0$ and $\beta^T \phi(\mathbf{d}_{j_2}) + \beta_0$, respectively. The ranking approach aims at predicting the correct ranking of t_{j_1} and t_{j_2} , meaning:

$$(t_{j_2} - t_{j_1}) (\beta^T \phi(\mathbf{d}_{j_2}) - \beta^T \phi(\mathbf{d}_{j_1})) > 0$$

If we assume that $t_{j_2} > t_{j_1}$, predictions only have to verify that $(\beta^T \phi(\mathbf{d}_{j_2}) - \beta^T \phi(\mathbf{d}_{j_1})) > 0$. The new objective function is:

$$\begin{aligned} \min_{\beta, \xi^{**}} \quad & \frac{1}{2} \|\beta\|^2 + \gamma^* \sum_{j_1=1}^n \sum_{j_2=1}^n (\nu_{j_1 j_2} \xi_{j_1 j_2}^{**}) \\ \text{constraints:} \quad & (\beta^T \phi(\mathbf{d}_{j_2}) - \beta^T \phi(\mathbf{d}_{j_1})) \geq 1 - \xi_{j_1 j_2}^{**}, \text{ for } j_1, j_2 = 1, \dots, n \\ & \xi_{j_1 j_2}^{**} \geq 0, \text{ for } j_1, j_2 = 1, \dots, n \end{aligned}$$

where $\nu_{j_1 j_2}$ has value 1 if observations j_1 and j_2 are comparable, and value 0 if they are not. Two observations are comparable if the one with shorter observed time is not censored. In other words, if $t_{j_2} > t_{j_1}$ and $\delta_{j_1} = 1$ (Fouodo et al., 2018). An alternative

formulation proposed by Van Belle et al. (2011) has the comparability effect included in the inequality constraints. As formulated by Fouodo et al. (2018), the new objective function is:

$$\begin{aligned} \min_{\beta, \xi^{**}} \quad & \frac{1}{2} \|\beta\|^2 + \gamma^* \sum_{j_1=1}^n (\xi_{j_1}^{**}) \\ \text{constraints:} \quad & (\beta^T \phi(\mathbf{d}_{j_1}) - \beta^T \phi(\mathbf{d}_{j_2(j_1)})) \geq t_{j_1} - t_{j_2(j_1)} - \xi_{j_1}^{**}, \text{ for } j_1 = 1, \dots, n \\ & \xi_{j_1}^{**} \geq 0, \text{ for } j_1 = 1, \dots, n \end{aligned}$$

where $j_2(j_1)$ is the nearest comparable neighbor of observation j_1 . From there, one can again use the Lagrangian and dual Lagrangian to compute a prognostic index. Given a new observation $\mathbf{z} = [z_1 \dots z_p]^T$, the prognostic index is computed as:

$$\hat{u}(\mathbf{z}) = \sum_{j_1=1}^n \alpha_{j_1} (\phi(\mathbf{d}_{j_1}) - \phi(\mathbf{d}_{j_2(j_1)}))^T \phi(\mathbf{z}) = \sum_{j_1=1}^n \alpha_{j_1} (K(\mathbf{d}_{j_1}, \mathbf{z}) - K(\mathbf{d}_{j_2(j_1)}, \mathbf{z}))$$

where the α_{j_1} 's are the Lagrange multipliers.

Finally, Van Belle et al. (2011) have combined the regression and ranking approaches into the hybrid approach. As formulated by Fouodo et al. (2018), the new objective function is:

$$\begin{aligned} \min_{\beta, \beta_0, \xi, \xi^*, \xi^{**}} \quad & \frac{1}{2} \|\beta\|^2 + \gamma \sum_{j=1}^n (\xi_j + \xi_j^*) + \gamma^* \sum_{j_1=1}^n (\xi_{j_1}^{**}) \\ \text{constraints:} \quad & (\beta^T \phi(\mathbf{d}_{j_1}) - \beta^T \phi(\mathbf{d}_{j_2(j_1)})) \geq t_{j_1} - t_{j_2(j_1)} - \xi_{j_1}^{**}, \text{ for } j_1 = 1, \dots, n \\ & t_j - (\beta^T \phi(\mathbf{d}_j) + \beta_0) \leq \xi_j, \text{ for } j = 1, \dots, n \\ & \delta_j ((\beta^T \phi(\mathbf{d}_j) + \beta_0) - t_j) \leq \xi_j^*, \text{ for } j = 1, \dots, n \\ & \xi_j \geq 0, \text{ for } j = 1, \dots, n \\ & \xi_j^* \geq 0, \text{ for } j = 1, \dots, n \\ & \xi_{j_1}^{**} \geq 0, \text{ for } j_1 = 1, \dots, n \end{aligned}$$

From there, one can again use the Lagrangian and dual Lagrangian to compute the prognostic index (Van Belle et al., 2011).

In section 2.6.4, we will need a way to determine which of two observations j_1 and j_2 has the worst prediction outcome. In the case of survival SVM, it will simply be the observation with the lower prognostic index. Thus, observation j_1 is said to have a worse prediction outcome than observation j_2 if:

$$\hat{u}(\mathbf{d}_{j_1}) < \hat{u}(\mathbf{d}_{j_2}) \quad (13)$$

2.6.4. Model performance

To compare the performance of the models, we will use two measures: the concordance index (C-index) and the Brier score. Both are complementary given that the C-index measures discrimination while the Brier score measures prediction accuracy.

The concordance index (C-index) was created by Harrell et al. (1984). It estimates the proportion of pairs of observations where the outcomes and the predictions are concordant. It will allow us to compare the performance of the three methods: Cox regression, Random Survival Forests and Survival Support Vector Machine. There are three steps to compute the C-index:

1. Form every possible pair of observations. Drop the pairs where the observation with shorter survival time is censored. Also drop the pairs where both observations are censored and have the same survival time. Lastly, drop the pairs where both observations have the same survival time and both are uncensored. The number of pairs left is called *permissible*.
2. Determine a method to compare predicted outcomes. Recall from sections 2.6.1, 2.6.2 and 2.6.3 that we will use equations (3), (5) and (13), respectively, to determine which observation has a worse predicted outcome. Let *concordance* = 0. Then, for each pair of observations with distinct survival times, add 1 to *concordance* if the observation with shorter survival time has a worse predicted outcome than the other; and add 0.5 if both have the same predicted outcome. For each pair of observations with tied survival times, add 1 to *concordance* if both have the same predicted outcome; and add 0.5 otherwise.
3. Compute the concordance index as $C_{index} = \frac{\text{concordance}}{\text{permissible}}$.

The Brier score was created by Brier (1950) and adapted to right-censored data by Graf et al. (1999). It is close to the mean squared error. To compute the Brier score, one needs both an estimate of the Survival Function and of the Censoring Survival Function. Since survival SVM does not provide an estimate of the Survival Function, we will only use the Brier score to compare the performance of the two remaining methods: Cox regression and Random Survival Forests. Given $\hat{S}(t | \mathbf{d}_j)$ and $\hat{C}(t)$, the estimators of the Survival Function and Censoring Survival Function respectively, Graf et al. (1999) proposed the following formula to compute the Brier score:

$$BS(t) = \frac{1}{n} \sum_{j=1}^n \left(\frac{\hat{S}^2(t | \mathbf{d}_j) I(t_j \leq t) \delta_j}{\hat{C}(t^-)} + \frac{(1 - \hat{S}(t | \mathbf{d}_j))^2 I(t_j > t)}{\hat{C}(t)} \right)$$

Note that for RSF, we will use the OOB ensemble estimator $\hat{S}^{OOB}(t | \mathbf{d}_j)$. Based on the inverse probability of censoring weight, $\hat{C}(t)$ is the Kaplan-Meier estimator of the

censoring distribution (Gerds and Schumacher, 2006; Graf et al., 1999). While the Survival Function estimator is computed using (1) on (t_j, δ_j) , for $j = 1, \dots, n$, the Censoring Survival Function estimator is computed using (1) on $(t_j, 1 - \delta_j)$, for $j = 1, \dots, n$. Let $c_1, \dots, c_M$ be the M distinct ordered censoring times, $r(c_i)$ be the number of observations at risk of being censored at time c_i (it is equal to $r(t_i)$ in (1)) and $m(c_i)$ be the number of observations actually censored at time c_i (without including the observations experiencing the event). The Kaplan-Meier estimator for the Censoring Survival Function is:

$$\hat{C}(t) = \prod_{c_i \leq t} \frac{r(c_i) - m(c_i)}{r(c_i)}$$

Note that $\hat{C}(t^-)$ stops just before time t , which means that the Kaplan-Meier estimator is:

$$\hat{C}(t^-) = \prod_{c_i < t} \frac{r(c_i) - m(c_i)}{r(c_i)}$$

Finally, one might be interested in an overall score that does not depend on time. The Integrated Brier Score (IBS) provides a global assessment of the model's performance. It is defined as:

$$IBS(\tau) = \frac{1}{\tau} \int_0^\tau BS(t) dt$$

where τ denotes the largest time point for which the Brier score could be reliably estimated. The estimated Censoring Survival Function becomes unstable as it approaches zero due to the small number of observations at risk of experiencing the event.

2.7. Survival prediction (with time-dependent covariates)

In this part, time-dependent covariates are allowed. We will only focus on one method: the semi-parametric Cox regression. It is built directly on what we explained in sections 2.3 to 2.5 and is the counterpart of the model presented in 2.6.1, but allowing time-dependent covariates. We will also present an adapted version of the concordance index to assess the performance of the model. Since computing an estimation of the Survival Function is complex when there are time-dependent covariates, we will not compute the Brier score.

Once the vector of estimated parameters $\hat{\beta}$ is computed (see section 2.3), one can compute the risk score at a given time (Antolini et al., 2005; Therneau and Watson, 2017). The risk score of observation j at time t can be computed as:

$$RS_j(t) = \hat{\beta}^T \mathbf{d}_j(t)$$

where $\mathbf{d}_j(t) = [d_{j1}(t) \dots d_{jp}(t)]^T$ is the vector of covariates at time t for the j^{th} observation.

Let $RS_1(t), \dots, RS_{L(t)}(t)$ be the $L(t)$ ordered risk scores at time t . There is one risk score for each observation still at risk of experiencing the event at time t . Let $Rk_j(t)$ be

the rank of the j^{th} observation among the ordered risk scores at time t^1 and $\bar{R}k_j(t) = \frac{L(t)+1}{2}$ be the average of these ranks. To compute the adapted concordance index, one can use the formula proposed by Therneau and Watson (2017). Let t_j be the survival time of observation j and *permissible* be the number of comparable pairs as computed in section 2.6.4, the formula is:

$$C_{index} = \frac{1}{2} + \frac{\sum_{j=1}^n \delta_j (Rk_j(t_j) - \bar{R}k_j(t_j))}{permissible}$$

where each observation's risk score is only evaluated at its survival time.

¹When there are tied risk scores, $Rk_j(t)$ uses mid-ranks

3. First dataset

3.1. Creation of the dataset

This dataset already existed and was created to study the mortality risk of NBA players (Martínez et al., 2022). It contains information on the 4,374 people who played at least one game in the NBA between its inception in 1946 (National Basketball Association, 2024) and July 2019.

As target variable (time variable), we decided to use both the number of years playing in the NBA and the number of games played in the NBA. This allows us to assess whether model results are sensitive to the choice of time variable. The number of years (*years*) can be obtained in the dataset by using the columns representing a player's first season in the NBA (*from*) and his last (*fin*s). The number of games is written in the column *g*.

For the censoring indicator, we will not use the column *cens* as it was used for the study of mortality risk, and it indicates whether a player is dead or not. Instead, we will use the column *state* where *Former* will be recomputed as 1 (non-censored observation) and *Active* will be recomputed as 0 (censored observation). In addition, players who died during their career are also included in censored observations. These 20 players are listed in Annex B.

The data also includes the following variables:

- *player* is the name of the player
- *pos* is the playing position (1 = Point Guard, 2 = Shooting Guard, 3 = Small Forward, 4 = Power Forward, 5 = Center) (see Annex C for more information)
- *place* is the place of birth (*USA* if the player was born in the United States, *No USA* if he wasn't)
- *ageleft* is the age of the player at his first game in the NBA (in years)
- *kilos* is the weight of the player (in kilograms)
- *cms* is the height of the player (in centimeters)
- *lefthanded* is a binary variable specifying if the player is left-handed (*No* = right-handed, *Yes* = left-handed)

There are other variables in the dataset, but we will not use them. Finally, three observations were removed due to missing values. We thus have 4,371 observations left in the dataset.

Note that the dataset includes ABA statistics. Between 1967 and 1976, another major professional basketball league existed in the USA: the American Basketball Association (ABA). In 1976, the NBA and the ABA merged (Jozsa, 2014). Before the merger, some major NBA players played in the ABA. In order to keep these players in our dataset, we decided to take into account their ABA statistics. Simply put, if a player spent at least one season of his career in the ABA, this season is considered for the variables we selected. To simplify the reading of this study, we will only talk about the NBA but remember that it implies the presence of the ABA. Among the players who played at least one season in the ABA, there are, for example, Julius Erving (five seasons) and Moses Malone (two seasons) (Sports Reference, 2025).

Although this dataset is exhaustive in terms of observations, it does not provide a lot of variables. This is why we also use another dataset that we created with fewer observations but more variables. This is the objective of section 4.

3.2. Descriptive analysis

3.2.1. Time variables

We use two time variables as target variables: the number of games (g) and the number of years ($years$) spent in the NBA. Both distributions are shown in Figures 5 (a) and (b) while Annex D contains all the summary statistics. In the two following plots, all observations are represented jointly, censored and uncensored ones. This means that the time represented is T , which can either be the event time X or the censoring time C .

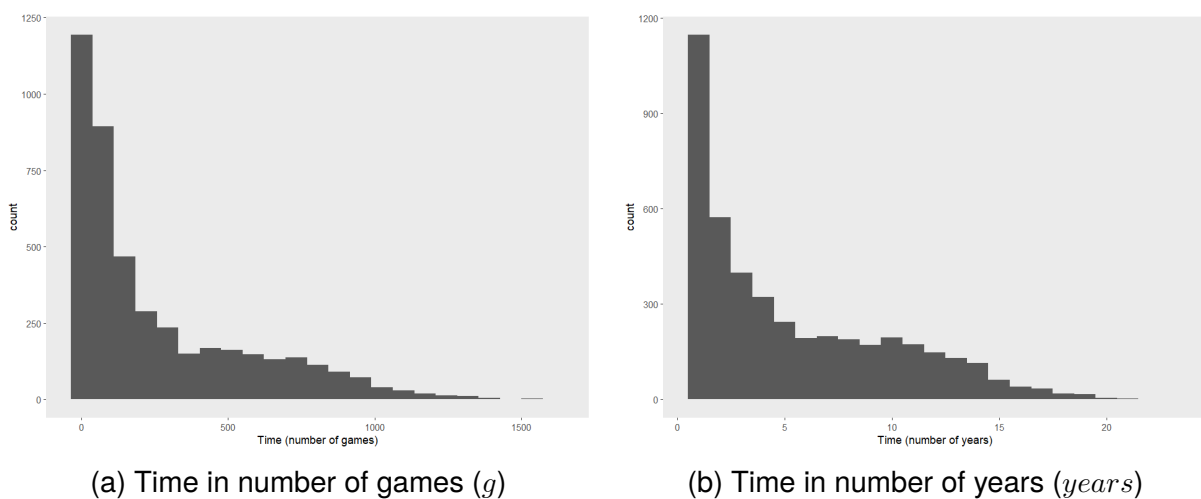


Figure 5: Distribution of time

The two distributions have a similar shape. There are a lot of observations with a small value on the time variables. It means that a lot of players played for a small amount of time in the NBA, and/or that a lot of censoring happens early in a player's career. As time increases, the number of observations goes down. For instance, less

than 25% of the players play more than 427 games (or more than 9 years) in the NBA. There is a flattening between 300 and 700 games (or between 5 and 10 years). Note that one needs to be careful when observing a right-censored variable since some values do not represent the true survival time, but only a lower bound.

3.2.2. Censoring indicator

We now turn to the distribution of the censoring indicator (*state*).

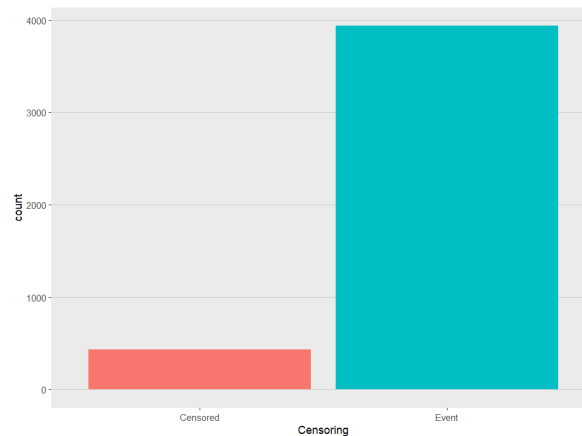


Figure 6: Distribution of *state* (censoring indicator)

Figure 6 shows that there are very few censored observations compared to non-censored ones. Of the 4,371 observations, 432 are censored. Censored observations are mainly players who have not yet retired at the end of the study (July 2019). Since the dataset spans nearly 80 years, most of the players have already retired. Therefore, we were expecting such an unbalanced distribution.

3.2.3. Quantitative covariates

There are three quantitative covariates in the dataset: the age of a player at his first game (*ageleft*), the weight of a player (*kilos*) and the height of a player (*cms*). Figures 7 to 9 represent the distributions of these variables. Annex D contains the summary statistics.

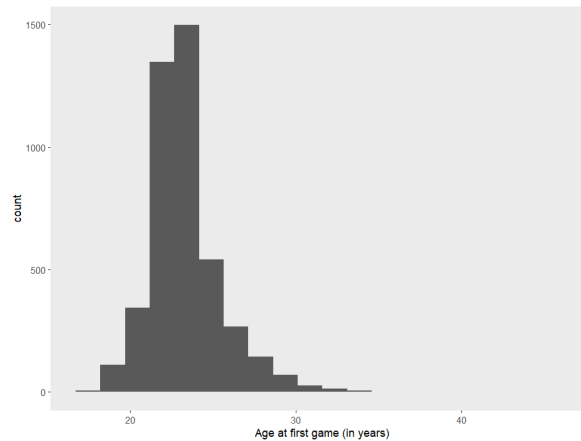


Figure 7: Distribution of *ageleft*

Figure 7 shows that a lot of observations have a low value on *ageleft*. It means that many players were young when playing their first NBA game. Three players out of four were younger than 24.10. The asymmetry of the distribution is also quantified by a skewness of 1.38. Note that only one player started his NBA career past the age of 36.

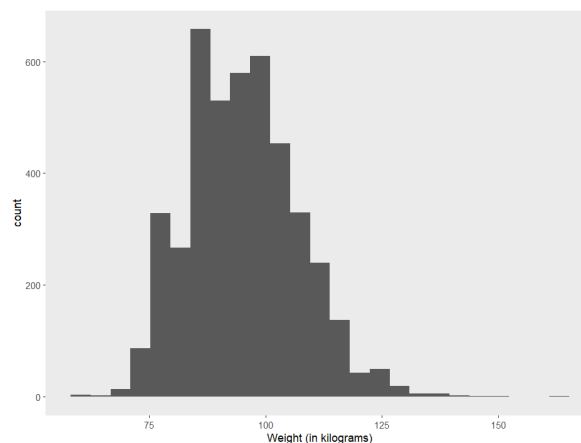


Figure 8: Distribution of *kilos*

Figure 8 shows that the weight distribution looks close to symmetric around its core, with most players clustered near 95.06 kg. The moderate skewness (0.49) is largely driven by a small number of players weighing over 130 kg, whose presence in the right tail is visible in the histogram but does not strongly affect the overall shape of the distribution.

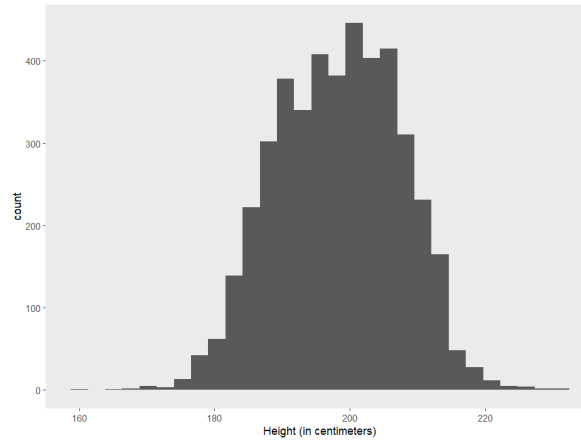


Figure 9: Distribution of *cms*

Figure 9 shows that the distribution of the height is almost normal, which is supported by a skewness of -0.08 and a kurtosis² of -0.33 . It means that around 95% of the players have a height between 180.14 cm and 216.44 cm.

Lastly, Table 1 displays the Spearman correlation between the three quantitative covariates. This reflects the well-established relationship between height and weight observed more broadly in human populations. The relations between *ageleft* and both *kilos* and *cms* are very weak.

Table 1: Spearman correlation

	<i>ageleft</i>	<i>kilos</i>	<i>cms</i>
<i>ageleft</i>	1.000	-0.129	-0.147
<i>kilos</i>	-0.129	1.000	0.845
<i>cms</i>	-0.147	0.845	1.000

3.2.4. Qualitative covariates

There are three qualitative covariates in this dataset: the playing position (*pos*), the place of birth (*place*) and the dominant hand (*lefthanded*). Figures 10, 12 and 13 represent the distribution of these variables.

²Throughout this thesis, we use the excess kurtosis.

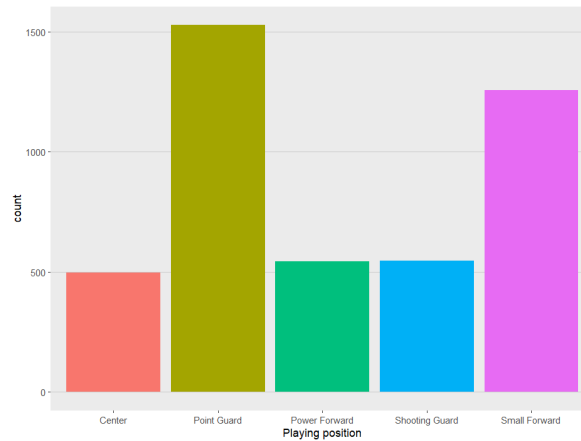


Figure 10: Distribution of pos

Figure 10 shows that there are a lot more players playing point guard and small forward than the other positions. This could mean that there is a higher turnover in players for these two positions, meaning that a point guard or small forward plays, on average, fewer games than a center, power forward or shooting guard.

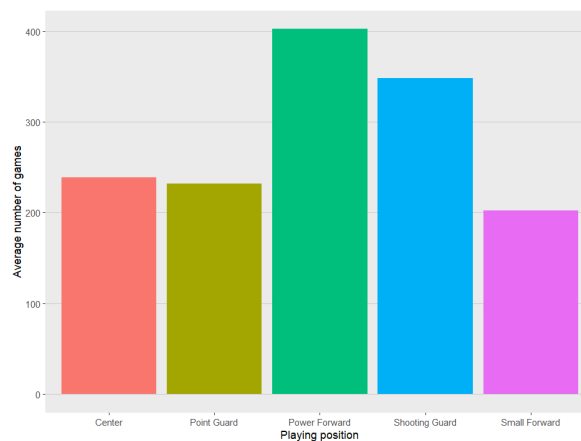


Figure 11: Average number of games per playing position

Figure 11 presents a different picture: while point guards and small forwards are the most numerous (Figure 10), they are not necessarily the positions with the lowest average number of games. Thus, our previous hypothesis is not supported by Figure 11. This unbalanced distribution is probably due to the construction of covariate pos . A player does not always play at his registered position. For instance, Patrick Beverley is registered as a point guard in the dataset, but during the 2016-17 season, he played around 67 games as a shooting guard (Basketball-Reference.com, n.d.a). Therefore, deeper investigation of the variable is needed before reaching any conclusion.

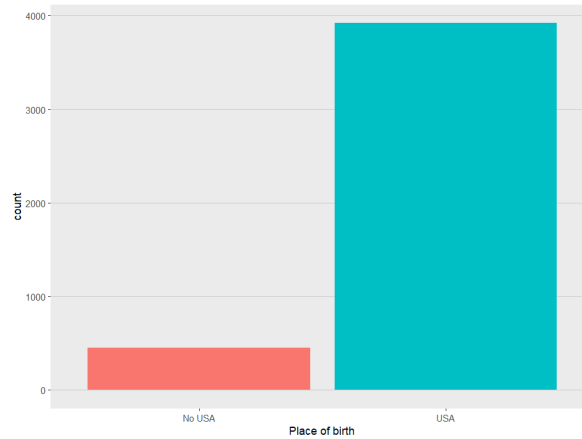


Figure 12: Distribution of *place*

Figure 12 shows that the number of players born in the USA is substantially higher than the number of players born in another country. This gap was expected as the NBA is an almost entirely American league (one team is actually located in Toronto, Canada).

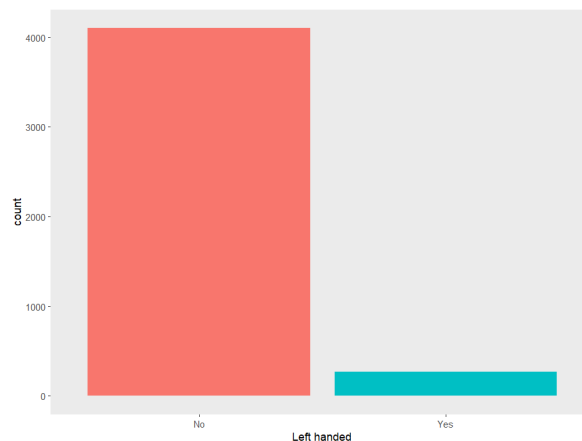


Figure 13: Distribution of *lefthanded*

Figure 13 shows that most of the players are right-handed. More specifically, around 6% of NBA players are left-handed. In the general population, the proportion of left-handed people is around 10% (Hardyck et al., 1976; Papadatou-Pastou et al., 2020; Scharoun and Bryden, 2014). This discrepancy may reflect sport-specific selection dynamics rather than a data artifact.

Lastly, Table 2 displays Cramér's V between the three qualitative covariates. Cramér's V is a measure of association between two categorical variables (Cramér, 1946) and is the counterpart of the correlation for continuous variables. The association between dominant hand and both the playing position and the place of birth is negligible. In contrast, there is a weak association between the playing position and the place of birth (Cohen, 2013).

Table 2: Cramér's V

	<i>pos</i>	<i>place</i>	<i>lefthanded</i>
<i>pos</i>	1.000	0.239	0.032
<i>place</i>	0.239	1.000	0.018
<i>lefthanded</i>	0.032	0.018	1.000

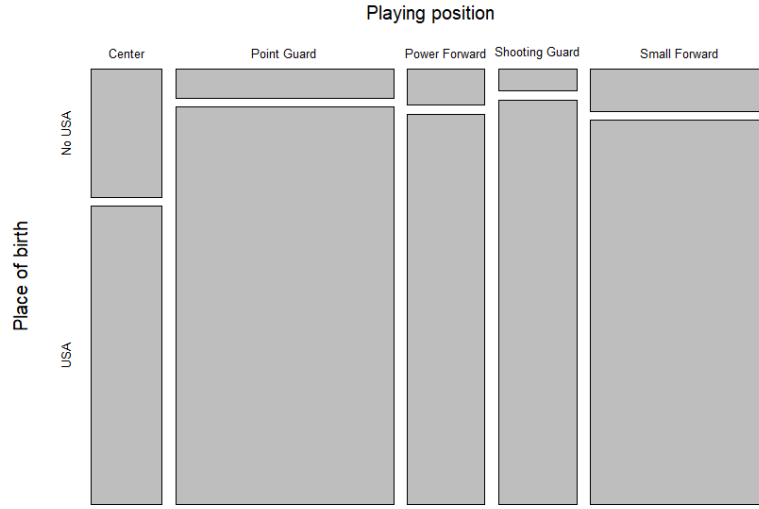
Figure 14: Relation between *pos* and *place*

Figure 14 shows that the proportion of players born outside the USA is bigger for centers than for the other playing positions.

3.3. Methods & Results

3.3.1. Kaplan-Meier

We plot the Kaplan-Meier survival curves for each variable. These curves are created by computing the Kaplan-Meier estimator for each level v of a variable k and for each event time t in each level. Then, the estimators are plotted in relation to time. In order to plot survival curves for the quantitative variables, we discretize variables *ageleft*, *kilos* and *cms* as follows:

- *ageleft* $\Rightarrow \{(\leq 21], (21 - 22], (22 - 23], (23 - 24], (24 - 25], (> 25)\}$
- *kilos* $\Rightarrow \{(\leq 80], (80 - 90], (90 - 100], (100 - 110], (> 110)\}$
- *cms* $\Rightarrow \{(\leq 190], (190 - 200], (200 - 210], (> 210)\}$

The grouping bounds were chosen using rounded values, both for interpretability and to avoid excessively unequal group sizes. Figures 15 to 20 represent the Kaplan-Meier survival curves for each variable independently. Even though differences might appear on the following plots, the significance of these differences cannot be assessed based

solely on them. In these plots, time is measured in number of games played in the NBA. Annex E displays the same plots, but time is measured using the number of seasons spent in the NBA. The overall shape of the curves is similar between the two time scales, though some differences emerge for specific variables (see the discussion of Figure 16 below).

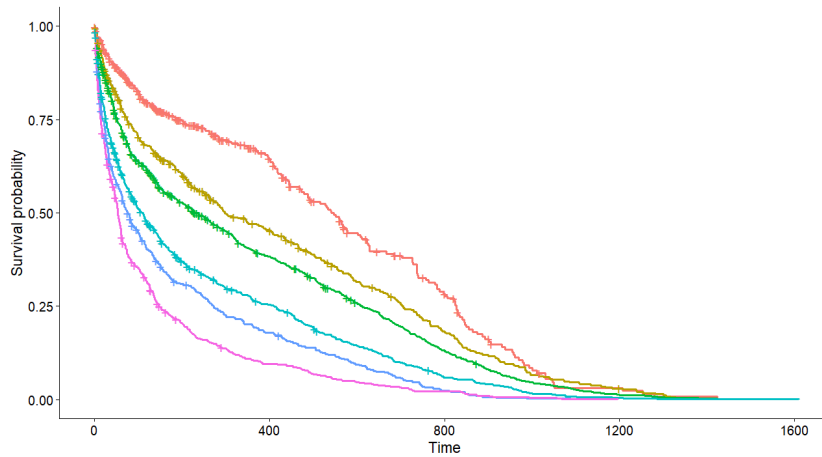


Figure 15: Survival Curves per age at first NBA game (discretized)

Based on Figure 15, players starting their NBA careers younger seem to have longer careers. Note that around the 1100th game, the curve for players starting their NBA careers before turning 21 crosses the curves for players starting between 21 and 22, and those starting between 22 and 23. This might indicate that the PHA does not hold, but we will explore that in section 3.3.3.

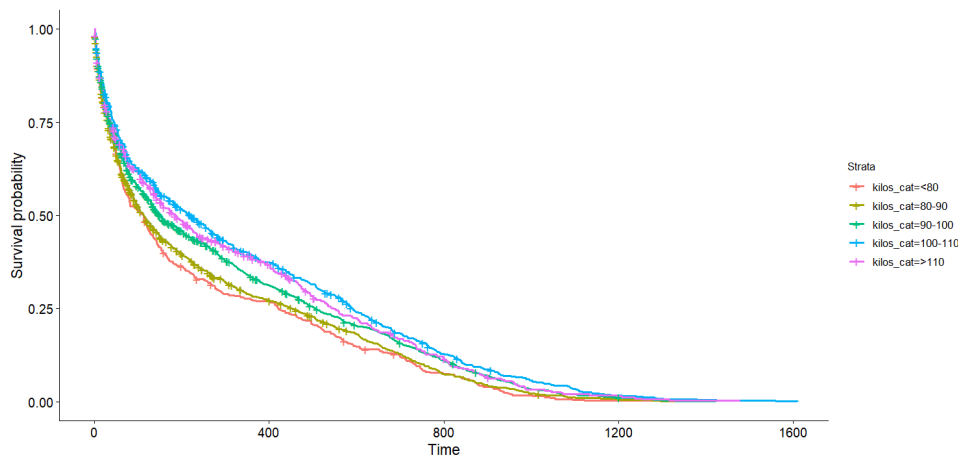


Figure 16: Survival Curves per weight (discretized)

In Figure 16, differences are less pronounced. Overall, heavier players appear to have an advantage in career length over lighter ones, except for players weighing over 110 kg, whose survival curve is lower than the survival curves of players weighing between 100 and 110 kg. Figure 51 in Annex E, on the other hand, shows that when measuring career length in number of seasons rather than in number of games, the players weighing over 110 kg seem to have longer careers than those weighing between 100

and 110 kg. However, these groups of weight contain relatively few observations, so the ordering between them should be interpreted with caution.

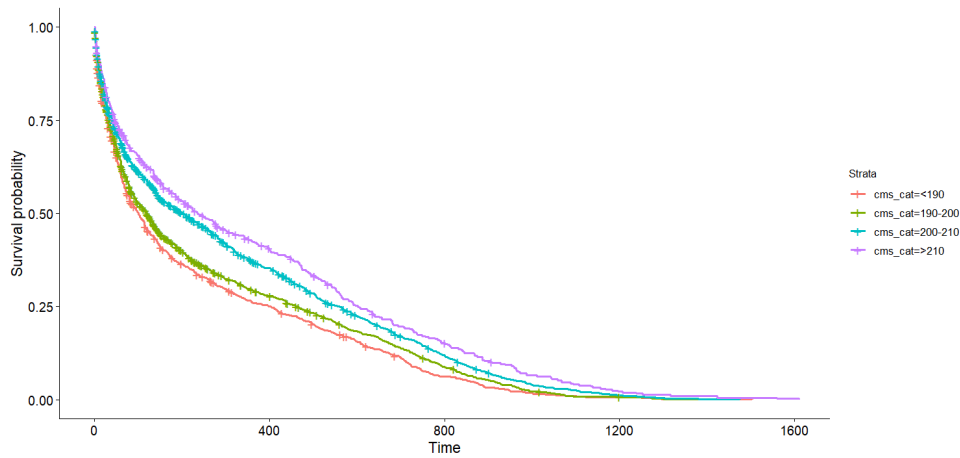


Figure 17: Survival Curves per height (discretized)

Figure 17 shows that taller players seem to have longer NBA careers than shorter ones. Moreover, there seems to be a gap between players measuring more or less than 200 cm. Note that the survival curves never intersect.

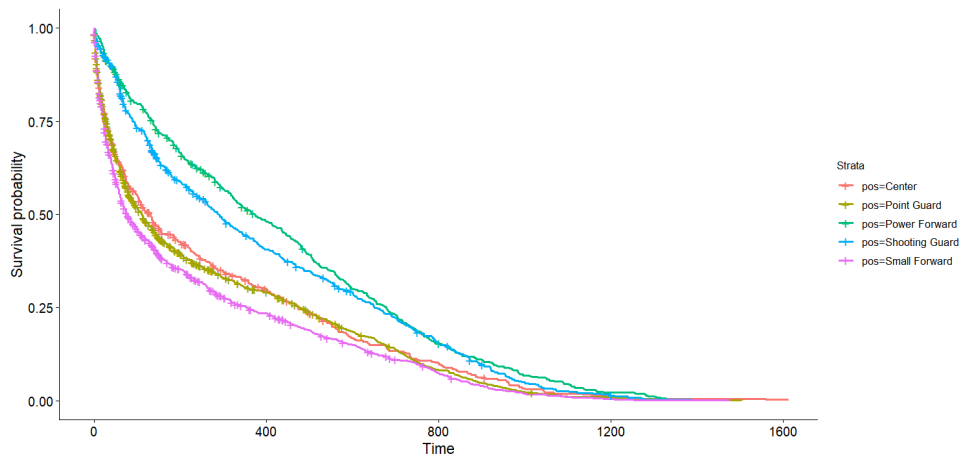


Figure 18: Survival Curves per playing position

Regarding playing positions, Figure 18 showcases a gap in survival probability between two groups. Power forwards and shooting guards seem to have longer careers than centers, point guards and small forwards. Although less pronounced, there also seem to be differences in survival probability between power forwards and shooting guards. Small forwards also seem to have lower survival probability than both centers and point guards. Lastly, note that the survival curves for centers and point guards intersect a few times.

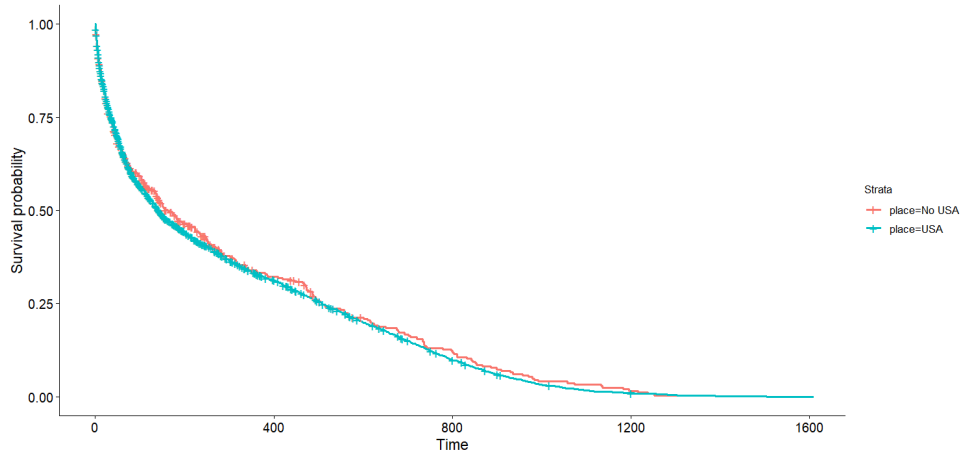


Figure 19: Survival Curves per place of birth

Figure 19 shows that the survival curves are close together. Still, non-US-born players show slightly longer expected careers than US-born players.

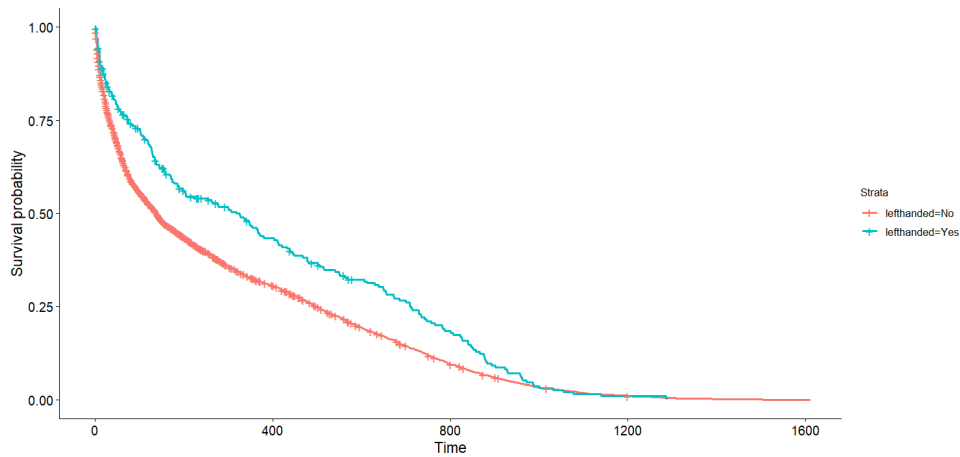


Figure 20: Survival Curves per dominant hand

Based on Figure 20, left-handed players appear to outlast right-handed ones in terms of career length. This difference disappears around the 1000th game.

3.3.2. Proportional Hazards Model (PHM)

We start by running an AIC selection to check which covariates should be included in the model when the number of games played in the NBA is used as the time variable. The final model has an AIC of 57,265 and includes four covariates: the age for the first NBA game (*ageleft*), the height (*cms*), the playing position (*pos*) and the dominant hand (*lefthanded*). We decide not to include interactions between covariates because the model becomes a lot more complex (six covariates and seven interactions after AIC selection). Two covariates were not selected: the weight (*kilos*) and the place of birth (*place*). In section 3.2.3, we found a correlation of 0.845 between the height and the weight of a player. They might provide the same knowledge. Therefore, dropping one might not lead to a major loss of information, while keeping both makes the model more

complex. Regarding the place of birth, we saw in Figure 19 (section 3.3.1) that the survival curves of players born in the USA or outside were really close to each other. Thus, the place of birth might not provide information that is substantial for modelling career length.

When we apply AIC selection on the dataset, but using the number of seasons played in the NBA as the time variable, the results are different. Indeed, with an AIC of 57,380, the final model keeps all six covariates. Thus, depending on how career length is measured, covariates *kilos* and *place* do or do not provide meaningful information. Recall that, when we looked at the Kaplan-Meier survival curves, the weight was the only variable where there was a visible difference between the two measures of time. Regarding the place of birth, no noticeable difference was spotted. Once again interactions are not included due to a major cost in complexity (six covariates and nine interactions after AIC selection).

Next, we fit two Cox models through the data using only the covariates determined by AIC selection using both measures of time respectively. Note that the reference levels for the categorical covariates are 1 (= Point Guard) for *pos*, *No USA* for *place* and *No* for *lefthanded*.

Table 3 displays the parameters for the model measuring time as the number of games. Using a type 1 error probability $\alpha = 0.05$, every covariate has a significant impact on survival time. A player starting his career one year older than another will have an 18.4% higher risk to retire after a given number of games. For every gained centimeter in height, the estimated risk to retire after a given number of games gets cut by 1.8%. A left-handed player has 21.8% less risk to retire after a given number of games than a right-handed player.

Table 3: Cox model (for time as number of games)

	$\hat{\beta}$	$\exp(\hat{\beta})$	<i>lower.95</i>	<i>upper.95</i>	<i>p.val</i>
<i>ageleft</i>	0.168	1.184	1.167	1.201	$< 2.2 \times 10^{-16}$
<i>cms</i>	-0.018	0.982	0.976	0.988	8.002×10^{-9}
<i>pos_SG</i>	-0.277	0.758	0.678	0.848	1.342×10^{-6}
<i>pos_SF</i>	0.414	1.513	1.349	1.696	1.218×10^{-12}
<i>pos_PF</i>	-0.208	0.812	0.703	0.939	4.936×10^{-3}
<i>pos_C</i>	0.341	1.407	1.175	1.684	2.011×10^{-4}
<i>lefthanded</i>	-0.246	0.782	0.683	0.895	3.674×10^{-4}

Regarding the playing position, one can check the relative retirement risks by modifying the level used as reference level. Table 4 shows the values of $\exp(\hat{\beta})$ for every possible reference level. From there, one can get a hierarchy between playing positions in terms of estimated risk to retire after a given number of games:

$$SG \leq PF < PG < C \leq SF$$

where "<" represents a significant difference, and "≤" a non-significant one. There seem to be three distinct groups: SG/PF, PG, and C/SF.

Table 4: $\exp(\hat{\beta})$ for every reference level in pos (for time as number of games) ³

Ref.level →	<i>PG</i>	<i>SG</i>	<i>SF</i>	<i>PF</i>	<i>C</i>
<i>PG</i>		1.319*	0.661*	1.231*	0.711*
<i>SG</i>	0.758*		0.501*	0.933	0.539*
<i>SF</i>	1.513*	1.995*		1.862*	1.075
<i>PF</i>	0.812*	1.071	0.537*		0.577*
<i>C</i>	1.407*	1.856*	0.930	1.732*	

Table 5 displays the parameters for the model measuring time as the number of seasons. Regarding covariates *ageleft*, *cms* and *lefthanded*, there is no major difference between the two models. All three have a significant impact on survival time. Recall that two covariates were kept in this model but not in the previous one: *kilos* and *place*. For every gained kilogram in weight, the estimated risk to retire after a given number of seasons gets significantly cut by 1.0%. A player born in the USA has a 10.8% higher risk to retire after a given number of seasons than a player born outside the USA. However, still using $\alpha = 0.05$, this difference is not significant.

Table 5: Cox model (for time as number of seasons)

	$\hat{\beta}$	$\exp(\hat{\beta})$	<i>lower.95</i>	<i>upper.95</i>	<i>p.val</i>
<i>ageleft</i>	0.189	1.209	1.192	1.225	$< 2.2 \times 10^{-16}$
<i>kilos</i>	-0.010	0.990	0.985	0.995	6.621×10^{-5}
<i>cms</i>	-0.019	0.981	0.974	0.989	3.637×10^{-7}
<i>pos_SG</i>	-0.173	0.841	0.750	0.942	2.833×10^{-3}
<i>pos_SF</i>	0.584	1.793	1.591	2.020	$< 2.2 \times 10^{-16}$
<i>pos_PF</i>	0.048	1.049	0.904	1.218	0.527
<i>pos_C</i>	0.613	1.846	1.536	2.218	6.349×10^{-11}
<i>place</i>	0.102	1.108	0.987	1.244	0.083
<i>lefthanded</i>	-0.253	0.777	0.678	0.889	2.46×10^{-4}

But the impact of the playing position on career length seems different when it is measured as the number of seasons rather than as the number of games. Using the same technique as before, one can get a hierarchy between playing positions in terms of estimated risk to retire after a given number of seasons:

$$SG < PG \leq PF < SF \leq C$$

where "<" represents a significant difference, and "≤" a non-significant one. Compared to the previous model, the hierarchy may seem very different. However, we still have

³In this table, the symbol "*" indicates a p-value lower than 0.05.

three groups: SG, PG/PF, and SF/C. Thus, there seem to be only two significant differences between the models. The difference between point guards and power forwards is no longer significant, while the difference between power forwards and shooting guards becomes significant. In other words, one could say that when measuring career length as the number of seasons, power forwards' career length gets closer to point guards and further from shooting guards than when measuring it as the number of games. Annex F contains the table used to determine the hierarchy.

3.3.3. Proportional Hazards Assumption (PHA)

We look at the Schoenfeld residuals to check if they are independent of time. Figure 21 (a) to (d) are the plots of Schoenfeld residuals through time for each covariate in the model measuring time as the number of games.

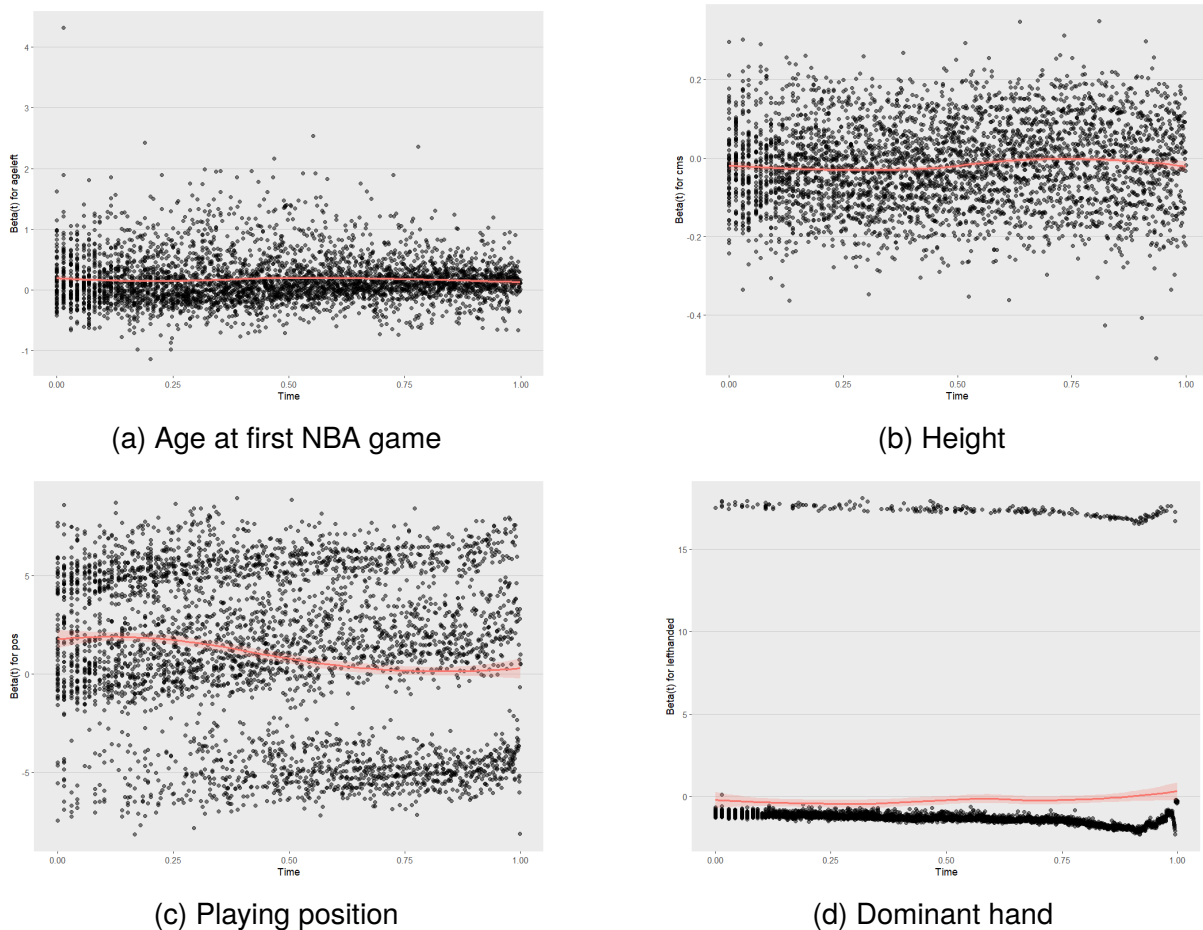


Figure 21: Schoenfeld residuals through time (measured as number of games)

The smoothed curves seem horizontal, except for *pos*. It means that the Schoenfeld residuals are independent of time for covariates *ageleft*, *cms* and *lefthanded*, but not for *pos*.

Then the hypothesis of straight regression lines with zero slope is tested both globally and locally for each covariate. Table 6 displays the p-values of each test. The results confirm what we saw in Figure 21 as the slope of the regression line is significantly

different from zero for *pos*, but not for *ageleft*, *cms* and *lefthanded* (with $\alpha = 0.05$). The p-value of the global test is also significant, which suggests that the PHA does not hold for this model, so results should be interpreted with caution, especially for the playing position. Regarding *ageleft* specifically, note that the test does not indicate a significant violation despite the curve crossing observed in Figure 15.

Table 6: Hypothesis test on Schoenfeld residuals (for time as number of games)

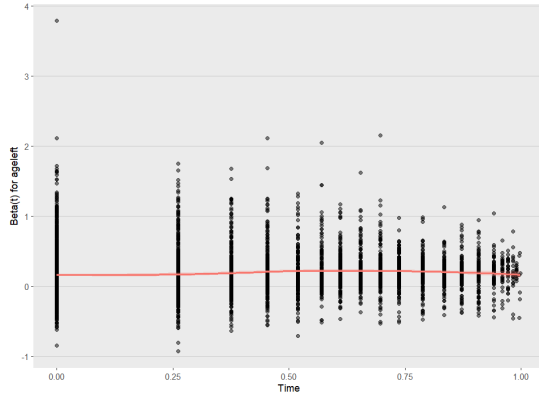
	<i>p.val</i>
<i>ageleft</i>	0.590
<i>cms</i>	0.670
<i>pos</i>	$< 2.2 \times 10^{-16}$
<i>lefthanded</i>	0.100
<i>GLOBAL</i>	$< 2.2 \times 10^{-16}$

Figure 22 (a) to (f) are the Schoenfeld residuals plots for the model measuring time as the number of seasons. Although visual inspection alone does not allow for a definitive conclusion, the smoothed curves seem horizontal for *ageleft*, *kilos*, *cms* and *place*, but not for *pos* and *lefthanded*. However, the results of the hypothesis tests displayed in Table 7 seem different. Using $\alpha = 0.05$, the slopes of the regression lines are significantly different from zero for every covariate except *kilos* and *place*. The p-value of the global test is also significant, which means that the proportional hazards assumption does not hold for this model either. Once again, we need to be cautious when interpreting the results of the Cox model. However, the discrepancy between the visual and formal tests may partly reflect the large sample size, which increases the test's power to detect even small, practically negligible departures from proportionality.

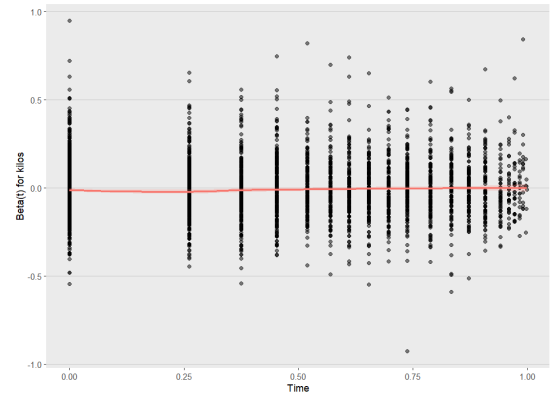
Table 7: Hypothesis test on Schoenfeld residuals (for time as number of seasons)

	<i>p.val</i>
<i>ageleft</i>	0.008
<i>kilos</i>	0.073
<i>cms</i>	0.020
<i>pos</i>	$< 2.2 \times 10^{-16}$
<i>place</i>	0.949
<i>lefthanded</i>	0.005
<i>GLOBAL</i>	$< 2.2 \times 10^{-16}$

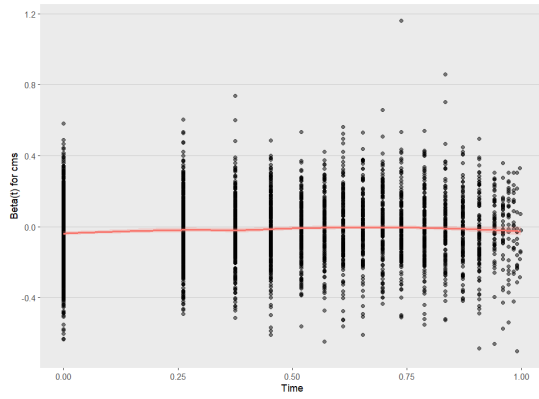
Notably, *ageleft*, *cms*, and *lefthanded* violate the PHA when time is measured in seasons but not when measured in games, suggesting that PHA diagnostics here are sensitive to the choice of time scale. In both cases, *pos* shows the strongest evidence of PHA violation among the covariates.



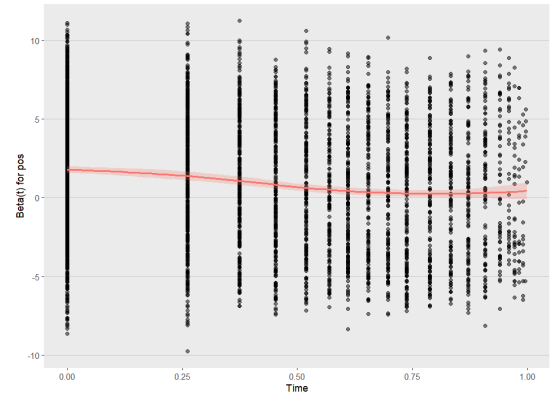
(a) Age at first NBA game



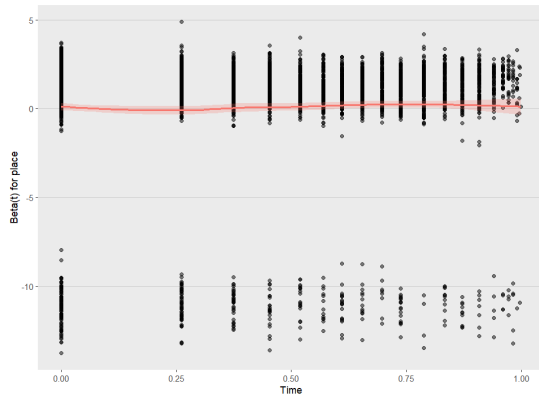
(b) Weight



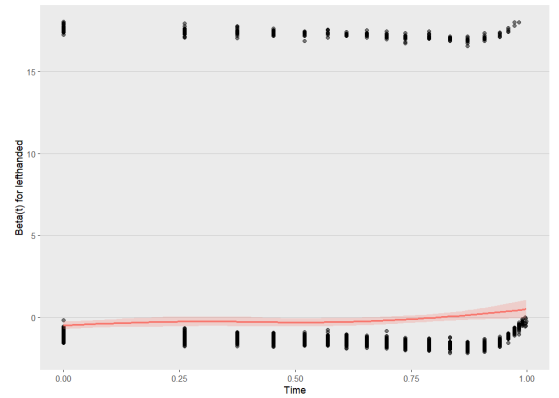
(c) Height



(d) Playing position



(e) Place of birth



(f) Dominant hand

Figure 22: Schoenfeld residuals through time (measured as number of seasons)

One way to deal with this is to use a stratified model. We now check whether it resolves the violation of the PHA by fitting a new model stratified on *pos*. Table 8 (a) and (b) displays the p-values of the hypothesis tests on Schoenfeld residuals for the model measuring time as the number of games, and for the model measuring time as the number of seasons. Based on Table 8 (a), the PHA seems to hold for the stratified model measuring time as the number of games (with $\alpha = 0.05$). However, in Table 8 (b), the p-value of the global test is still significant, and so is the p-value of the test on *lefthanded* (with $\alpha = 0.05$). It suggests that the PHA does not hold for the stratified model measuring time as the number of seasons.

Table 8: Hypothesis test on Schoenfeld residuals for stratified models

(a) Time as the number of games

	<i>p.val</i>
<i>ageleft</i>	0.675
<i>cms</i>	0.210
<i>lefthanded</i>	0.100
<i>GLOBAL</i>	0.200

(b) Time as the number of seasons

	<i>p.val</i>
<i>ageleft</i>	0.171
<i>kilos</i>	0.058
<i>cms</i>	0.067
<i>place</i>	0.697
<i>lefthanded</i>	0.005
<i>GLOBAL</i>	0.007

The coefficients of the two stratified models are almost identical to their non-stratified counterparts. Annex G contains these coefficients. There are only two differences. First, the exponentials of the coefficients now quantify the relative risks to retire for a given playing position (after a given time t). Second, we do not have coefficients for the stratified covariate anymore, meaning that its impact on career length is not measured by the stratified models.

3.3.4. Survival prediction

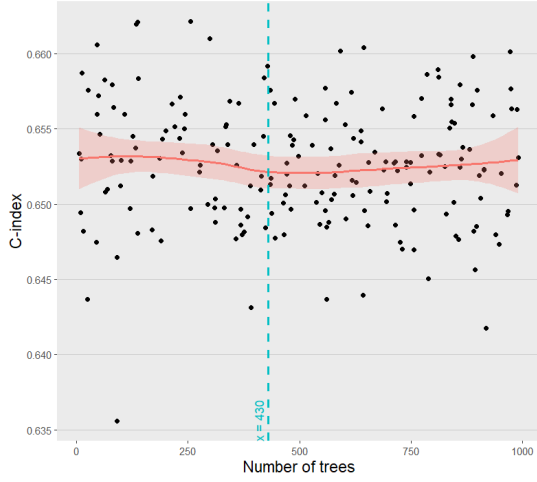
We use three different models to predict survival: Cox regression, Random Survival Forests (RSF) and Survival Support Vector Machine (SSVM). We start by optimizing the values of the hyperparameters for the RSF and the SSVM, then we evaluate the performance of the models. For Cox regression, there are no hyperparameters to optimize, but we run the performance evaluation twice: one for the stratified Cox model, and one for the non-stratified one. Since the stratified model performs worse than the other, both in terms of C-index and IBS and for both measures of time, we decide not to use stratification. We only present the results of the non-stratified model in this section, but the results of the stratified one are in Annex H. Note that we do not test stratified models for the RSF and the SSVM since they do not rely on the PHA. Our objective is to compare the performance of the different techniques used in each model. Therefore, we decide to only include the covariates selected with the AIC in section 3.3.2. This will prevent differences based on variable selection from occurring. On one hand, models measuring time as the number of games will only use four covariates: the age at first NBA game, the height, the playing position and the dominant hand. On the other hand, models measuring time as the number of seasons will use two additional covariates: the weight and the place of birth.

Following is the optimization procedure computed for the models measuring time as the number of games. Annex I contains the corresponding plots for the models measuring time as the number of seasons. The results are globally the same for both measures of time, except for one difference regarding the rank ratio of the SSVM. We will develop this point further down. During the optimization process, we often split the

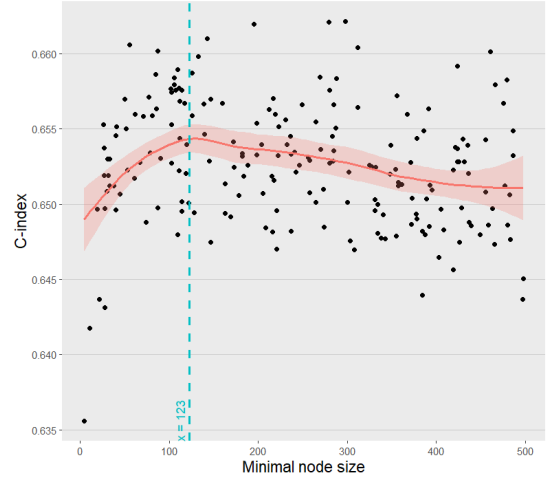
data into a training set and test set. Every time we do so, 20% of the data is selected as the test set. As the distribution of the censoring indicator is very unbalanced (see section 3.2.2), each split is stratified on it, meaning that the proportion of censored observations is the same in the original dataset, each training set and each test set.

Regarding RSF, we only optimize two hyperparameters: the number of trees in the forest and the minimum number of observations in a terminal node. The optimization is conducted manually with respect to both the C-index and the IBS. First, we select lower and upper bounds for both hyperparameters. Second, 200 values between these bounds are randomly selected. Third, for each of the 200 combinations, we randomly create 10 training and test sets from the original dataset. Fourth, for each of the 10 train/test sets created, we fit the model on the training set and compute the C-index and IBS on the test set. Fifth, we compute the mean C-index and IBS for each of the 200 combinations. We repeat the process until the optimum values of the hyperparameters can be determined. For each trial, new values of the lower and upper bounds are selected. Figure 23 (a) to (d) displays the best results of this optimization process. The red lines correspond to locally estimated regression curves (LOESS), which are included to highlight the overall trends in the data. The blue dashed lines are the values we select as optimum. In Figure 23 (a), the number of trees does not seem to have an impact on the C-index. However, it seems to have a small impact on the IBS in Figure 23 (c), with an optimum number of trees of 430. The impact of the minimal number of observations in a node is more obvious in Figures 23 (b) and (d). The optimum value for this hyperparameter is 123.

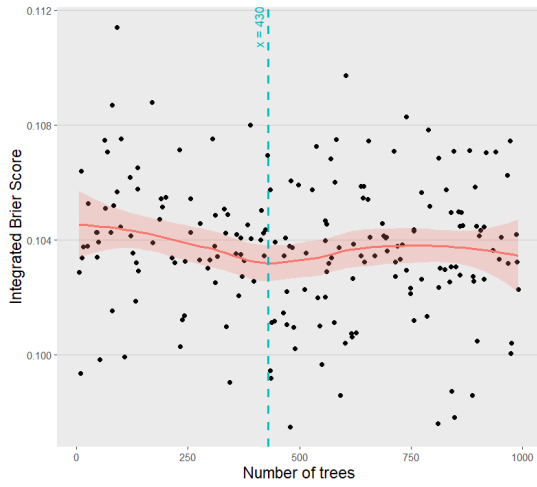
For SSVM, note that we use an algorithm slightly different from the one we presented in section 2.6.3, but the general idea does not change. The original algorithm was too time-consuming. There are three hyperparameters to optimize. First, the regularization parameter $\alpha > 0$ that determines the amount of regularization to apply (scikit-survival developers, n.d.). A smaller value of α means an increased regularization. Second, the hyperparameter controlling the trade-off between the regression and ranking approaches $r \in [0, 1]$. A value of 0 means that the model only uses the regression approach, and a value of 1, only the ranking approach. Third, the kernel function used. We only test three kernel functions: the linear kernel (linear relationship between the covariates and the prognostic index), the polynomial kernel (extends the linear kernel by modelling nonlinear interactions up to a specified polynomial degree) and the Radial Basis Function (RBF) kernel (maps the data into a high-dimensional feature space, enabling highly flexible nonlinear decision functions). In addition, the polynomial and RBF kernel functions have specific hyperparameters. We do not detail them as it goes beyond the scope of this thesis, but know that they will also be optimized.



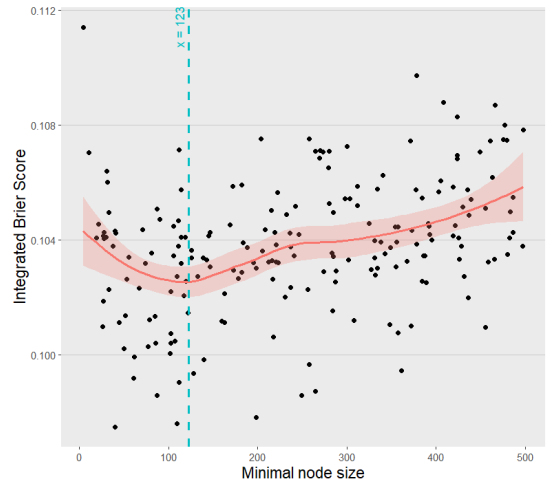
(a) C-index relative to number of trees



(b) C-index relative to minimum node size



(c) IBS relative to number of trees



(d) IBS relative to minimum node size

Figure 23: Optimization of the hyperparameters of the RSF

The optimization is conducted with respect to the C-index only since the IBS cannot be computed for SSVM. As for RSF, lower and upper bounds are selected for α and r . We split the optimization process into two steps. In step 1, we aim to determine the best kernel function. To do so, 200 values are randomly selected between the bounds and the kernel function is randomly selected. Then for each triplet $(\alpha, r, \text{kernel})$, 10 train and test sets are randomly created from the original dataset. Finally, 10 models are created, fitted on the standardized training sets, and used to compute the mean C-index over the standardized test sets. Standardization is done by removing the training set mean of each covariate and dividing by the training set standard deviation of each covariate. Figure 24 shows the results of step 1.

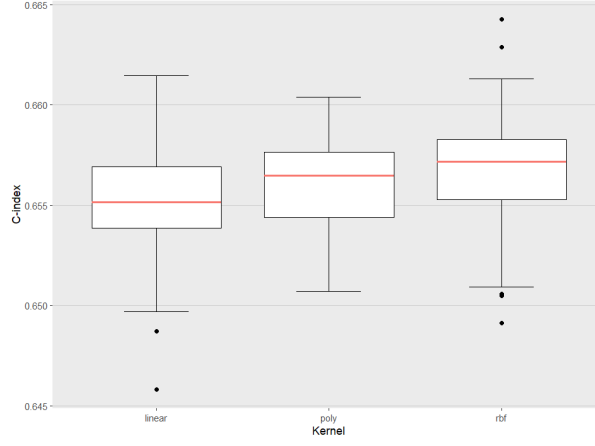
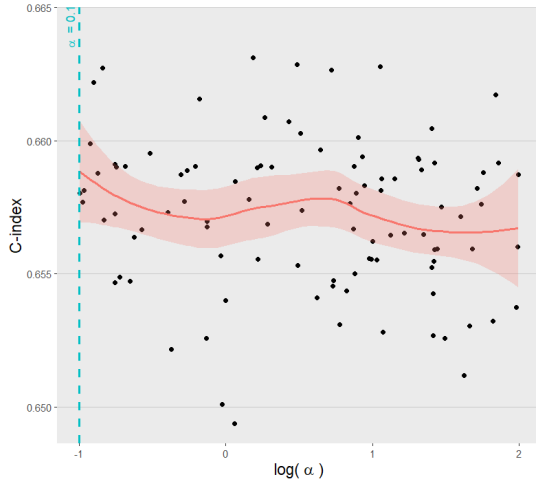


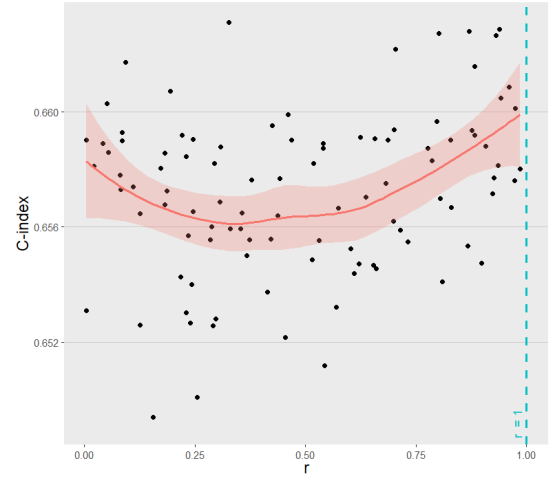
Figure 24: Optimization of the kernel function

The RBF kernel seems to perform best as it has the largest median C-index out of the three kernel functions. Therefore, we select RBF as the optimum kernel.

In step 2, we optimize α and r using the same procedure, but fixing the kernel function as the best one determined in step 1: RBF. Also, only 100 pairs (α, r) are evaluated because it takes too much time to fit an RBF-kernel. Figures 25 (a) and (b) show the relation between the two hyperparameters and the C-index. The red lines correspond to the locally estimated regression curves (LOESS), and the blue dashed lines are the values we select as optimum.



(a) C-index relative to regularization α



(b) C-index relative to rank ratio r

Figure 25: Optimization of the hyperparameters of the SSVM with RBF kernel

In Figure 25 (b), the maximum C-index seems to be achieved for $r = 1$. This result is mathematically justified since a rank ratio of 1 means that only the ranking approach is used. The ranking approach is equivalent to maximizing the C-index (Fouodo et al., 2018). However, when measuring time as the number of seasons rather than the number of games, Figure 58 (b) in Annex I shows that the rank ratio does not seem to impact the C-index. We decide to keep a rank ratio of 1 for comparability between the measures of time in the performance of the models. In Figure 25 (a), note that the C-index seems

to increase as α decreases. This is the reason we select the optimum regularization parameter $\alpha = 0.1$, with 0.1 the lower bound of the tested interval. We settle for this value because to optimize this hyperparameter further would be too time-consuming as the time to fit the model significantly increases for lower values of α . Moreover, in this thesis, we simply try to assess if SSVM could be a relevant technique to model the career length. A future study focused on finding the best SSVM model would be pertinent to explore these questions further.

Once the values of the hyperparameters are determined, we compute the performance of the three models for both time variables. To do so, we randomly split the dataset in two (the 80% training set and the 20% test set), we fit the models on the training set and compute their prediction capacity on the test set. In order not to depend on the data split, we repeat this procedure 50 times and compute the mean and the quantile-based 95% confidence interval. For Cox regression and RSF, both the C-index and the Brier score are used to assess prediction capacity, while only the C-index is used for the SSVM. Also, SSVM is fitted and tested on standardized datasets. Table 9 summarizes the results for the models where time is measured as the number of games. The three models are very close to each other in terms of C-index. It means that, based on this dataset, all three models seem to be equivalent when discriminating career length of professional basketball players. The IBS values of Cox regression and RSF are also almost equal, meaning that they seem to have equally good prediction accuracy.

Table 9: Performance of the three models (time as the number of games)

	2.5th pct	C – index	97.5th pct	2.5th pct	IBS	97.5th pct
<i>Cox</i>	0.637	0.656	0.674	0.094	0.104	0.117
<i>RSF</i>	0.641	0.657	0.669	0.094	0.103	0.115
<i>SSVM</i>	0.644	0.658	0.671	/	/	/

When looking at the predictive performance of the models where time is measured as the number of seasons, the results are different. In Table 10, we can see that the C-indexes are larger than for the other measure of time, but the IBS values are larger too. It means that, when career length is measured as the number of seasons, discrimination tasks seem to perform better, but at the cost of prediction accuracy. However, there seem to be no differences between the three models.

Table 10: Performance of the three models (time as the number of seasons)

	2.5th pct	C – index	97.5th pct	2.5th pct	IBS	97.5th pct
<i>Cox</i>	0.669	0.689	0.709	0.102	0.111	0.121
<i>RSF</i>	0.669	0.683	0.699	0.100	0.109	0.121
<i>SSVM</i>	0.669	0.684	0.698	/	/	/

Instead of looking at the Integrated Brier Score, we can also investigate the Brier scores through time. Figure 26 displays the evolution of the Brier score for increasing values of time, and for four different training/test splits (randomly created). The black curve corresponds to a reference model using none of the covariates. It uses the Kaplan-Meier estimator of the Survival Function to compute the Brier score.

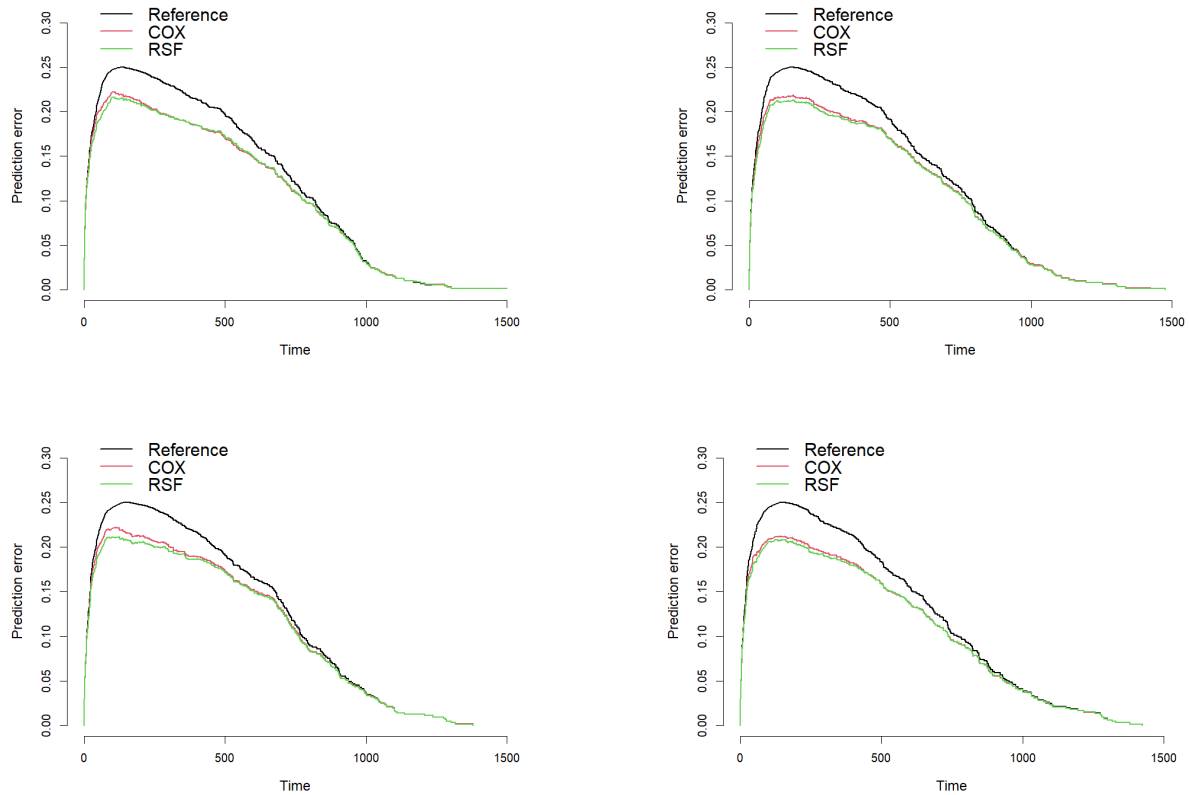


Figure 26: Brier Scores through time for Cox regression and RSF

Figure 26 shows that both Cox regression and the RSF have lower prediction error than the reference model between approximately the 100th and the 800th games. Outside this interval, there are no differences between the two models using covariates and the model using none. This means that Cox regression and the RSF perform well when predicting medium careers, but not for long (more than 800 games) and short (less than 100 games) career lengths. In the same interval, the RSF seems to perform slightly better than Cox regression, but once again, there seem to be no differences between them outside the $[100; 800]$ games interval. Annex J contains the corresponding plots for models where time is measured as the number of seasons. There seem to be no differences between the two measures of time.

3.4. Discussion

First, we have shown that some variables seem to have an impact on the career length of NBA players. The age of a player at his first NBA game is negatively corre-

lated to his career length, meaning that a player starting his career older has a shorter expected career length. An explanation could be that players tend to retire around the same age, so if they start their career earlier, they mathematically have more time to play. However, we cannot be sure that this is the reason for the association. There is a positive relation between career length and height. Indeed, taller players have longer expected careers. Left-handed players seem to have longer careers than right-handed ones. We obtained these results for all explored models: non-parametric (Kaplan-Meier in section 3.3.1) and semi-parametric (PHM in section 3.3.2) models, models measuring time as the number of games and models measuring time as the number of seasons, stratified and non-stratified models. For the Cox regression models where time is measured as the number of seasons (both stratified and non-stratified), the results concerning the dominant hand should be interpreted with caution given the PHA violation noted in section 3.3.3. However, the fact that all models achieve these results leads us to believe in their robustness. Therefore, we conclude that the career length of an NBA player is impacted by his age at his first NBA game, his height and his dominant hand.

Regarding the weight of a player, the PHM showed that it only seems to have a significant impact on career length measured as the number of seasons spent in the NBA rather than as the number of games. The same results were achieved with the stratified models. These findings echo the differences in survival curves when using different time measures. Hence, the weight of a player seems to have an impact on his career length when measured as the number of seasons spent in the NBA, but not when career length is measured as the number of games. Concerning the place of birth, the survival curves of players born in the USA and players born elsewhere were really close to each other, for both measures of time. For career length measured as the number of games, the variable was dropped during the AIC selection, and for career length measured as the number of seasons, its impact was not significant for both stratified and non-stratified models. Thus, we conclude that a player's career length is not impacted by his birthplace, but further studies may be relevant to confirm it (or not).

The impact of the playing position of a player on his career length is unclear. When looking at the survival curves, there seemed to be two different groups, but when looking at the regression coefficients of the PHM, three distinct groups appeared. Moreover, these three groups differed when using different measures of time. In addition, the playing position seems to be the main reason that the PHA does not hold for non-stratified models. When using models stratified on the playing position, the PHA seemed to hold. Also, recall that we questioned the construction of the variable *pos* in section 3.2.4. Therefore, we draw no conclusion regarding the playing position based on these results. However, further studies ought to investigate this covariate more deeply. For instance, one could consider fitting different models depending on the playing position.

Nonetheless, the impact of the covariates on the predictive performance tends to fade for both short and long careers, staying informative mainly between approximately games 100 and 800. We saw that, outside this interval, both Cox regression and RSF

seemed to perform as well as a reference model using no covariate. This suggests that the impact of the covariates is negligible for predicting short and long career lengths. However, we cannot assess this hypothesis based only on these results. Note that, in section 4, we will focus on the best players in the history of the NBA. Since they have longer careers, we will be able to explore the association between covariates and longer careers.

Second, we have compared two different measures of career length. As mentioned previously, two covariates were completely dropped when using the number of games, but not when using the number of seasons. One of them, the weight, even had a significant impact on career length when measured as the number of seasons, but not as the number of games. In addition, the predictive performance of the models were different. On one hand, when career length is measured as the number of seasons, discrimination tasks seem to perform better, but at the cost of prediction accuracy. On the other hand, when career length is measured as the number of games, prediction accuracy seems to increase, but discrimination performance decrease.

Third, we have shown that models based on survival analysis seem to be relevant to model the career length of NBA players. Although some differences were noticed, the results were generally constant through the different models used (Kaplan-Meier, Cox and stratified Cox) and for both measures of career length (number of games and number of seasons). We have also explored the predictive performance of three models through the use of two different criteria: the concordance index and the Brier score. Although it was not excellent, the models seem to perform better than random. We found no clear difference between the three models in terms of global predictive performance. This is true for both measures of time. Since the Survival Support Vector Machine is more complex and more time-consuming, we would recommend using either the classic Cox regression or the Random Survival Forests instead. However, the optimization process regarding the SSVM was not exhaustive. There exist other kernel functions that might fit the data better. Also, we settled for a suboptimal value of the hyperparameter α , since further optimization would have been too time-consuming. In addition, we only mentioned it in this thesis, but there are also kernel-related hyperparameters that one can optimize. Therefore, a new study focused on the application of the SSVM in the context of measuring basketball career length might be relevant to further assess its adequacy for this application. Finally, although no global differences between the models were noticed, RSF seemed to be slightly more accurate than Cox regression between approximately games 100 and 800. In addition, the tested Cox regression violated the PHA. It could mean that Random Survival Forests should be prioritized when trying to predict medium career lengths. Once again, further studies would be relevant to assess this hypothesis.

4. Second dataset

4.1. Creation of the dataset

4.1.1. Player selection

In order to simplify the search for the data and to be certain of its accuracy, we established a few criteria for the players included in our dataset. Moreover, it allows us to be sure that observations are comparable to each other.

First, we only focused on the American professional basketball league: the National Basketball Association (NBA). NBA player statistics are easy to find and accurate, even for people who played at the beginning of the league, in the 1940s and 1950s. One of the characteristics of the NBA is that it is a "closed" league. This means, inter alia, that teams playing in the league rarely change from one year to the next⁴ and that players can only enter the league through a "draft" (Andreff, 2019). The "draft" is a process used by some professional sports leagues to control players' entry into the league, especially young players. This system prevents teams from recruiting young players through academies or by buying their rights on the transfer market (Johnston et al., 2021). Thanks to the "closed" league system, we can consider the opposition a player faces reasonably consistent throughout his entire career in the league. Also, it allows us to control the beginning of a player's professional career. This is particularly positive for our study.

Second, we included ABA statistics. The American Basketball Association (ABA) was another major professional basketball league that existed in the USA between 1967 and 1976. Refer to section 3.1 for further explanations.

Finally, among all the remaining players, we chose to focus on the best ones. The NBA career length of the best players would hardly be comparable to the NBA career length of someone who struggled to get playing time. We based our player selection first on a list originally published in a 2011 Slam magazine special issue called "The 500 greatest NBA players of all time" and archived by Sports Reference (2011). We selected the 140 highest-ranked players who were drafted (a prerequisite for the draft-position variable used later). After that, we completed our sample based on our knowledge of the sport, mainly for people who played between 2011 and 2025. As selection criteria, we mainly used the individual and collective performances. In the end, we had a sample of 185 players. Annex K contains the list of selected players.

⁴For instance, the last time it changed was in 2004, when a new team was created (Basketball-Reference.com, n.d.c).

4.1.2. Variable selection

Many studies have already investigated the subject of variables that influence professional athletes' retirement. Based on those studies, there seem to be three categories of variables: biological, situational and sport-related.

Biological variables are the innate characteristics of the players. Even if some of them can evolve throughout the life or career of a player, they are still connected to genetics. First, a study showed that men have, on average, a longer career than women (Vretaros, 2022). In the present study, gender will not be investigated, as the NBA is an all-male sports league. Second, Fynn and Sonnenschein (2012) showed that there is a connection between the length of a player's career and his height and weight. However, this connection seems to be moderated by the player's playing position (Fynn and Sonnenschein, 2012). The present study should allow us to explore the relations between these different variables. Third, in order to check for a cohort effect, the birth year of the players will also be studied. The generational aspect of longevity has received little attention in the literature.

Situational variables are the acquired or contextual characteristics of the players. These variables cannot be qualified as "natural" and often change throughout the life or career of a player. First, there exists a connection between the length of a player's career and his nationality (Vretaros, 2022). We will investigate this variable in the present study by comparing American players with non-American players. Indeed, the path of young American athletes is really different from that of young athletes from other countries (Yan and McCullagh, 2004; Brettschneider and Brandl-Bredenbeck, 2007; Pot and Van Hilvoorde, 2013). Also, we do not have enough players from non-American countries in our sample to consider them separately. Second, one of the variables most related to the duration of a basketball player's career is injuries and health issues (Lavallee et al., 1997; Agresta et al., 2008; Ristolainen et al., 2011; Martin et al., 2021; Vretaros, 2022). We decided to study injuries and health issues through a time-varying variable, since we believe that there could be differences between players who suffered injuries early or late in their career. We only considered severe injuries and applied a heuristic criterion: a health issue is considered "severe" if a player missed 12 games or more and/or if he underwent surgery. Some variables will also vary according to the level of the injury/health variable (see the definition of *Health_Inj* in section 4.1.3).

Sport-related variables are all characteristics directly tied to athletic performance or competition. First, we have a variable specific to North America: the pick used to "draft" a player (Miguel et al., 2019; Vretaros, 2022). Each year, players wishing to enter the NBA need to register for the 'draft'. During this event, each team selects the player they want in a predefined order. Simply put, the best players are selected first. This variable also allows us to consider the hypothesis of a relation between career longevity and a player's performances prior to the NBA (Fynn and Sonnenschein, 2012). Indeed, a player with better performances before the draft is likely to be drafted before another

player with worse performances (Greer et al., 2019; Evans and Pitts, 2022). Second, some studies highlighted the relation between career length and success (Agresta et al., 2008; Fynn and Sonnenschein, 2012). We will separately consider team trophies (NBA titles) and individual awards (individual trophies). Third, some playing positions have an impact on the duration of a career (Fynn and Sonnenschein, 2012). In basketball, there are mainly five playing positions: point guard, shooting guard, small forward, power forward, and center. Note that this variable could have been considered biological since the playing position strongly depends on the physique of the player (Erčulj and Bračič, 2010; Ferioli et al., 2018). Lastly, there exists a relation between the length of a player's career and his performance (Groothuis and Hill, 2004). To measure a player's performance, we already have the collective and individual success variables. We decided to add another one: the average time a player spent on court during a game throughout his career. Indeed, it is quite easy to see that a player will play more than another if his performances are better.

4.1.3. Final variables

Here is the list of variables selected for the dataset:

- *Player* is the name of the player.
- *Retired* is a binary variable equal to 1 if the player has retired and 0 otherwise. This is the censoring indicator.
- *Nationality* is a binary variable equal to 0 if the player was born in the USA and 1 otherwise.
- *Draft_pos* is the draft pick of the player. As established in section 4.1.1, all included players were drafted.
- *Avg_min_pgm* is the average number of minutes a player spent on court during a game throughout his career. This variable also varies over time in relation to *Health_Inj*.
- *Position* is the playing position of the player (1 = Point Guard, 2 = Shooting Guard, 3 = Small Forward, 4 = Power Forward, 5 = Center) (see Annex C for more information).
- *Ind_Success* is the number of individual awards a player has won throughout his career. We will only consider the following awards: "Most Valuable Player" or "MVP" (the best player of the regular season), "NBA Finals MVP" (the best player of the finals), "NBA Conference Finals MVP" (the best players of each semi-final), "Defensive Player of the Year" (the best defensive player of the regular season), being selected in an "All-NBA Team" (the three best players of the regular season

for each position), being selected in an "All-Defensive Team" (the two best players of the regular season for each position).

- *Titles* is the number of NBA titles a player has won throughout his career.
- *Height* is the height of the player (in centimeters).
- *Weight* is the weight of the player (in kilograms).
- *Birth_year* is the birth year of the player.
- *Health_Inj* is the number of injuries and/or health problems a player has already suffered in his career. Every player starts at 0 (we do not consider injuries that have occurred before the NBA); the count increases every time a player is injured. Each new injury leads to a new entry of the player in the dataset with updated values for the following variables: *Avg_min_pgm*, *Health_Inj*, *Total_games_start* and *Total_games_end*. For example, Michael Jordan has two rows in the dataset: one covering his career up until his first severe injury, and the other covering his career until retirement.
- *Out_NBA* is a binary variable equal to 1 if the player played at least one season in another professional league after the beginning of his NBA career or if he stopped his professional career for at least one year (retired and then came back, military service, among others). It is equal to 0 otherwise.
- *Total_games_start* is the total number of games (regular- and post-season) the player has already played in the NBA at the start of the interval. This is one of the random variables that can be used to measure the time until retirement (together with *Total_games_end*). This variable also varies over time in relation to *Health_Inj*.
- *Total_games_end* is the total number of games (regular- and post-season) the player has already played in the NBA at the end of the interval. This is one of the random variables that can be used to measure the time until retirement (together with *Total_games_start*). This variable also varies over time in relation to *Health_Inj*.

4.2. Descriptive analysis

4.2.1. Time variables

We measure career length as the number of games played in the NBA. We use games as the measure of time because career length is tracked by time-dependent covariates based on injuries and/or health issues. If we were to use the number of seasons, it would be hard to split the observations for a player suffering an injury and/or

health issue in the middle of the season. Indeed, it would mean that the first part of that season would technically be before the injury and/or health issue, while the second part would be after. Measuring time as the number of games allows us to split a season in two. Since we have time-dependent covariates in the dataset, each row corresponds to a time interval and multiple rows may correspond to the same player. The time variable is split into two columns, *Total_games_start* and *Total_games_end*, representing the lower and upper bounds of the time intervals respectively. Figure 27 shows the distribution of total survival time, which is computed as the largest value of *Total_games_end* for each player. Annex L contains the summary statistics.

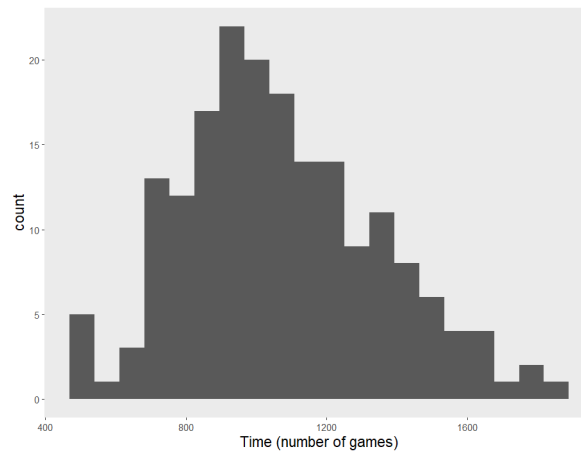


Figure 27: Distribution of the total number of games

The distribution is only slightly right-skewed (0.43), which is relatively low for a time-to-event variable. This is likely because the sample is restricted to elite players, who by construction have had long careers, pulling the bulk of the distribution away from zero and reducing the asymmetry typically seen in career-length data. Note that one needs to be careful when observing a right-censored variable since some values do not represent the true survival time, but only a lower bound.

4.2.2. Censoring indicator

We now look at the distribution of the censoring indicator (*retired*). This is a player-wise distribution, not an observation-wise one, meaning that every player is counted as one, no matter how many rows/observations correspond to that player.

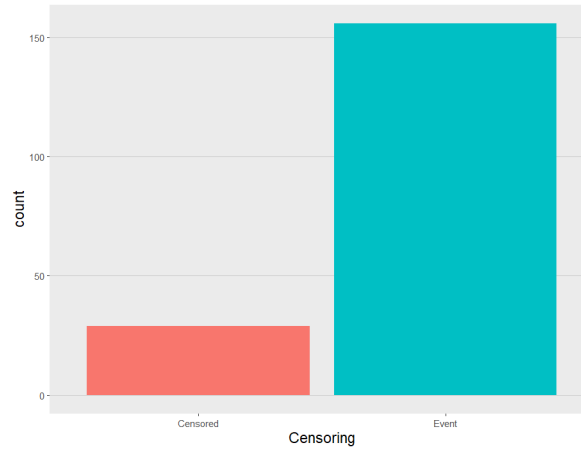


Figure 28: Distribution of the censoring indicator

Figure 28 shows that the distribution is very unbalanced. 29 out of the 185 players are censored. All of them are players that were still playing in the NBA as of July 2025 (end of the study). We focus on the best players in the history of the league, which spans nearly 80 years. Logically, most of the players considered are already retired.

4.2.3. Quantitative covariates

In the following section, we will only explore time-invariant quantitative variables. There are six of them: the position a player was selected in the draft (*Draft_pos*), the number of individual trophies a player won (*Ind_Success*), the number of NBA titles a player won (*Titles*), the height of a player (*Height*), the weight of a player (*Weight*) and the birth year of a player (*Birth_year*). Figures 29 to 34 show the distribution of these variables and Annex L contains the summary statistics.

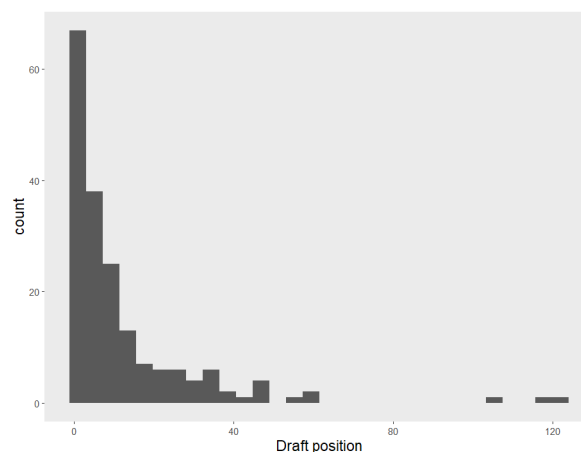


Figure 29: Distribution of *Draft_pos*

In Figure 29, we can see that the distribution is right-skewed (3.43). Recall that the draft pick is also a measure of a player's performances before entering the NBA. Since, in this dataset, we focus on the best players, we expected an asymmetric distribution. Note that the three values observed on the right side of the histogram are extreme.

Since 1989, there have never been more than 60 players selected during a single NBA draft. This is why there are only three players in this dataset that were drafted at position 106, 117 and 122 respectively. These extreme values are probably the reason for the large standard deviation of the distribution (18.18).

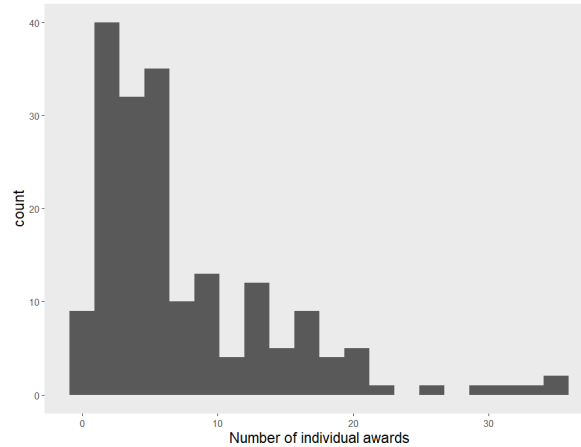


Figure 30: Distribution of *Ind_Success*

Once again, Figure 30 shows a right-skewed distribution for the individual success. However, the distribution would probably have been more skewed if we had not focused on the best players. Indeed, only nine players have never won any individual trophy and more than 25% have won at least 10.

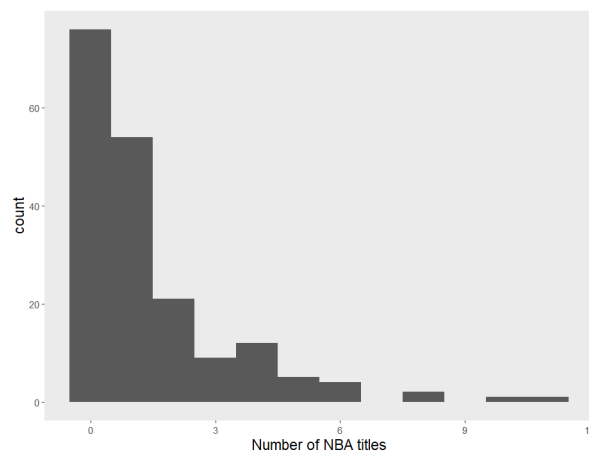


Figure 31: Distribution of *Titles*

Based on Figure 31, the distribution of the collective success is also right-skewed. As for *Ind_Success*, one could argue that the distribution would probably have been more skewed if we had selected every player in the history of the NBA.

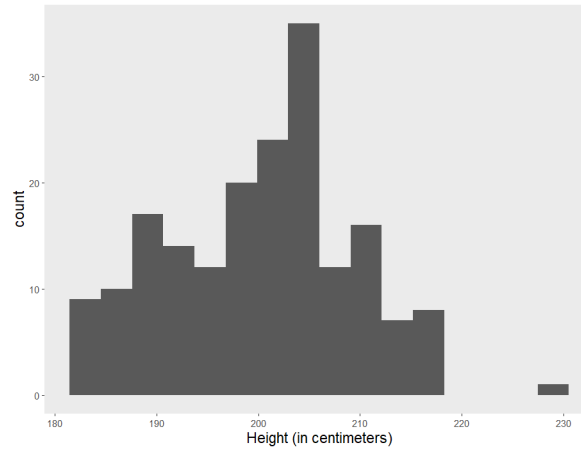


Figure 32: Distribution of *Height*

Figure 32 displays a distribution that is close to normal. This is supported by a skewness of -0.04 and a kurtosis⁵ of -0.44 . The average height of a player in the dataset is 200.21 cm. Note that we have one extremely large value: a player 229 cm tall. One important thing to note regarding the height is that it was originally measured in inches, then converted to centimeters. It means that the measures are rounded and there are only 16 different sizes.

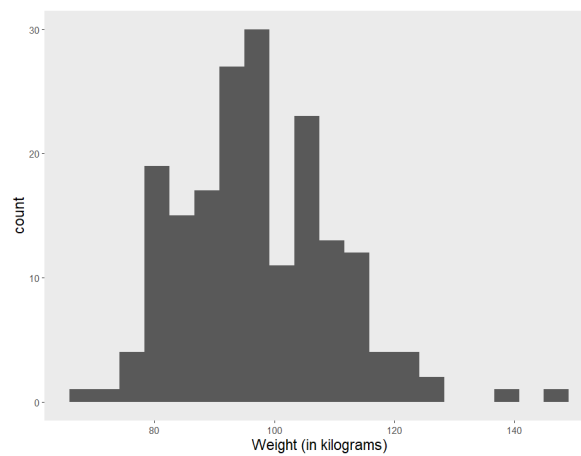


Figure 33: Distribution of *Weight*

Figure 33 shows that the distribution of weight is slightly right-skewed (0.53). Once again the measure conversion (pounds to kilograms) has rounded the values, but there are still 41 different measures of the weight.

⁵Throughout this thesis, we use the excess kurtosis.

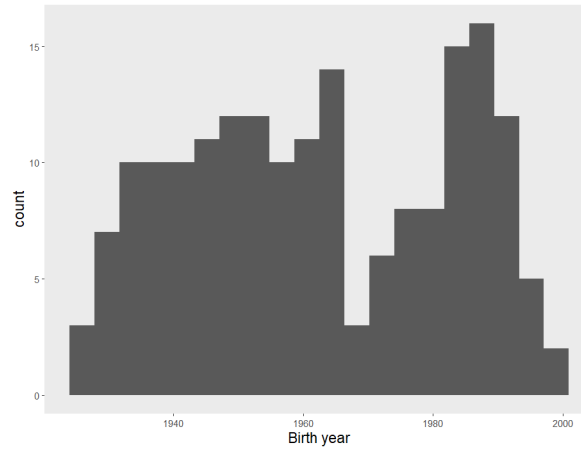


Figure 34: Distribution of *Birth_year*

In Figure 34, the distribution of the birth year is symmetrical (skewness of -0.01). However, there is a drop in density between approximately 1970 and 1980. This means that, among the best players in the history of the NBA, there are fewer players born during that decade.

Lastly, Table 11 displays the Spearman correlation between the six quantitative covariates. A correlation of 0.812 reflects the well-established relationship between height and weight observed more broadly in human populations. There also seem to be small to moderate relations between individual and collective successes (0.323) and between the weight and the birth year (0.304). While the former was expected as both variables account for the level of a player, the latter is more surprising. Indeed, the best NBA players seem to get heavier as time advances.

Table 11: Spearman correlation

	<i>Draft_pos</i>	<i>Ind_Success</i>	<i>Titles</i>	<i>Height</i>	<i>Weight</i>	<i>Birth_year</i>
<i>Draft_pos</i>	1.000	-0.113	0.087	-0.197	-0.183	-0.032
<i>Ind_Success</i>	-0.113	1.000	0.323	0.190	0.141	0.120
<i>Titles</i>	0.087	0.323	1.000	0.089	0.019	-0.130
<i>Height</i>	-0.197	0.190	0.089	1.000	0.812	0.108
<i>Weight</i>	-0.183	0.141	0.019	0.812	1.000	0.304
<i>Birth_year</i>	-0.032	0.120	-0.130	0.108	0.304	1.000

4.2.4. Qualitative covariates

In the following section, we will only explore time-invariant qualitative variables. There are three of them: if a player is from the USA or not (*Nationality*), the playing position (*Position*) and if a player spent at least one year out of the NBA between his first and last games in the NBA (*Out_NBA*). Figures 35 to 37 show the distribution of these variables.

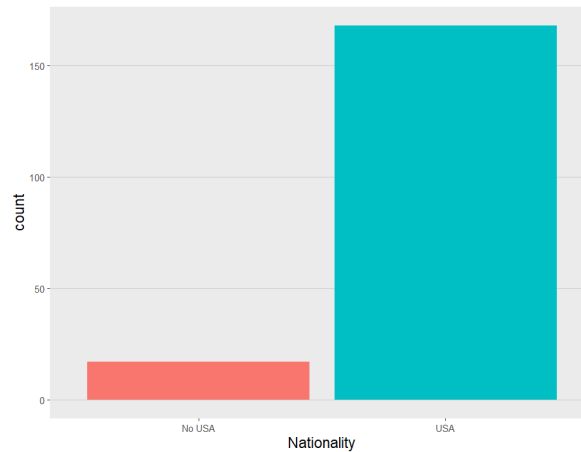


Figure 35: Distribution of *Nationality*

In Figure 35, we see that there are substantially more American players than players from other countries. Since the NBA is almost entirely American (one team in Toronto, Canada), this unbalanced distribution was expected. To be specific, 168 out of the 185 players of the dataset are Americans, that is approximately 90% of the dataset.

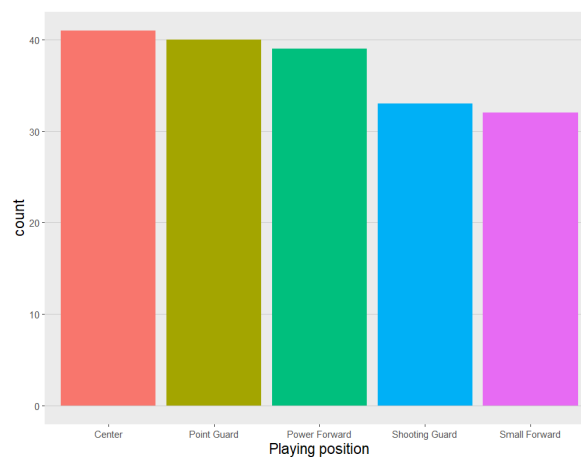


Figure 36: Distribution of *Position*

Figure 36 shows that the distribution of playing position is balanced in this dataset. Since we focus on the best players, this possibly means that there is no relation between the level of a player and his playing position. However, one cannot assess this hypothesis based only on this plot.

Based on Figure 37, the distribution between players who spent at least one year outside the NBA during their NBA career and players who played continuously in the NBA is unbalanced. 31 out of 185 players have spent at least one year outside the NBA between their first and last games in the league. Note that this does not mean that the other players did not play in another league. If they did, it happened before their first game or after their last game in the NBA. Hence, it is not taken into account when measuring the NBA career length.

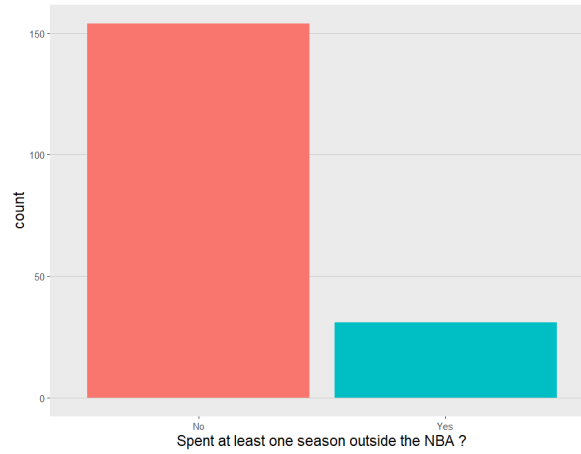
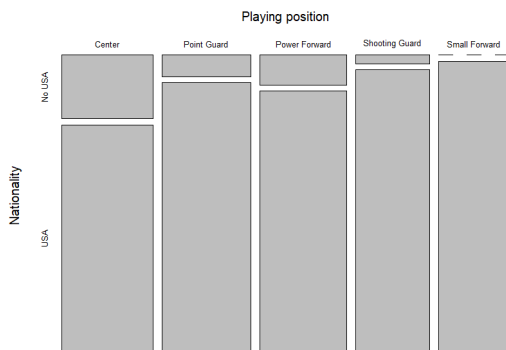


Figure 37: Distribution of *Out_NBA*

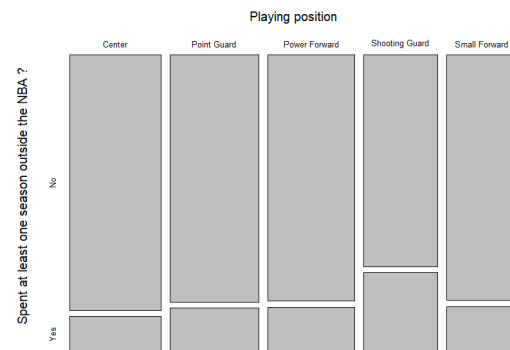
Finally, Table 12 displays Cramér's V between the three time-invariant qualitative covariates. Cramér's V is a measure of association between two categorical variables (Cramér, 1946) and is the counterpart of the correlation for continuous variables. The association between the nationality and *Out_NBA* is negligible. However, there seem to be a small association between the playing position and *Out_NBA*, and a small to moderate association between the playing position and the nationality (Cohen, 2013).

Table 12: Cramér's V

	<i>Nationality</i>	<i>Position</i>	<i>Out_NBA</i>
<i>Nationality</i>	1.000	0.264	0.058
<i>Position</i>	0.264	1.000	0.136
<i>Out_NBA</i>	0.058	0.136	1.000



(a) *Position* and *Nationality*



(b) *Position* and *out_NBA*

Figure 38: Relation between qualitative covariates

Figure 38 (a) shows that the proportion of non-American players is bigger for centers than for the other playing positions. There is not a single non-American small forward in the dataset. In Figure 38 (b), we see that the proportion of shooting guards who spent

at least one year outside the NBA during their career is larger than for any other playing position.

4.2.5. Time-dependent covariates

In this section, we explore the time-dependent covariates. There are two of them: the average number of minutes played in a game (*Avg_min_pgm*) and the number of severe injuries and/or health issues (*Health_Inj*). We will look at their distribution at the last measure for each player (end of their career or censoring) in Figures 39 (a) and 40 (a), and at the evolution of the variables through time in Figures 39 (b) and 40 (b). In the latter, each point represents one interval, plotted at *Total_games_end*, and the blue lines correspond to the locally estimated regression curves (LOESS). They are included to highlight the overall trends in the data. Annex L contains the summary statistics of *Avg_min_pgm* at the last measure for each player.

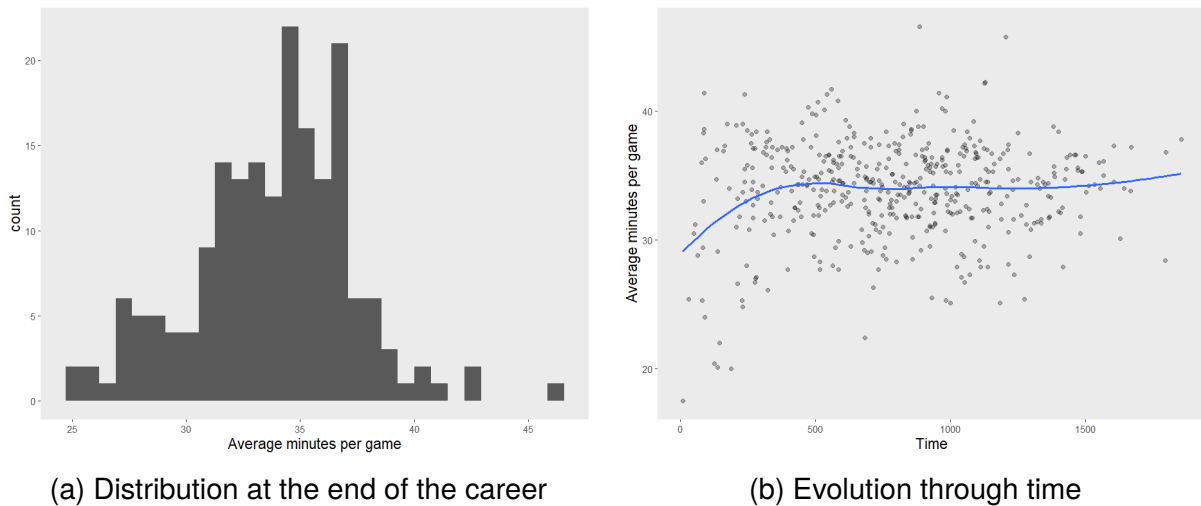


Figure 39: Time-dependent covariate *Avg_min_pgm*

Figure 39 (a) shows that the distribution of *Avg_min_pgm* is close to normal, which is supported by a skewness of -0.09 and a kurtosis of 0.52 . It means that approximately 95% of the players in the dataset have an average number of minutes played per game throughout their career between 27.01 and 40.61. In Figure 39 (b), we can see that the average time spent on court seems to increase until the 500th game. After that, it seems to stay constant. The increase at the beginning of the career was expected as a player's time on court gets progressively larger in the first part of his career, where he is progressing in terms of basketball level. One could have expected a decrease at the end of the career representing players getting older and declining in terms of basketball level. However, since we focus on the best players, one possible explanation is that they are so good that they never actually get worse. An alternate possibility is that players who are getting worse, and thus whose average time on court is decreasing, are the ones that actually stop playing. Nevertheless, these hypotheses are purely speculative and cannot be assessed based only on these plots.

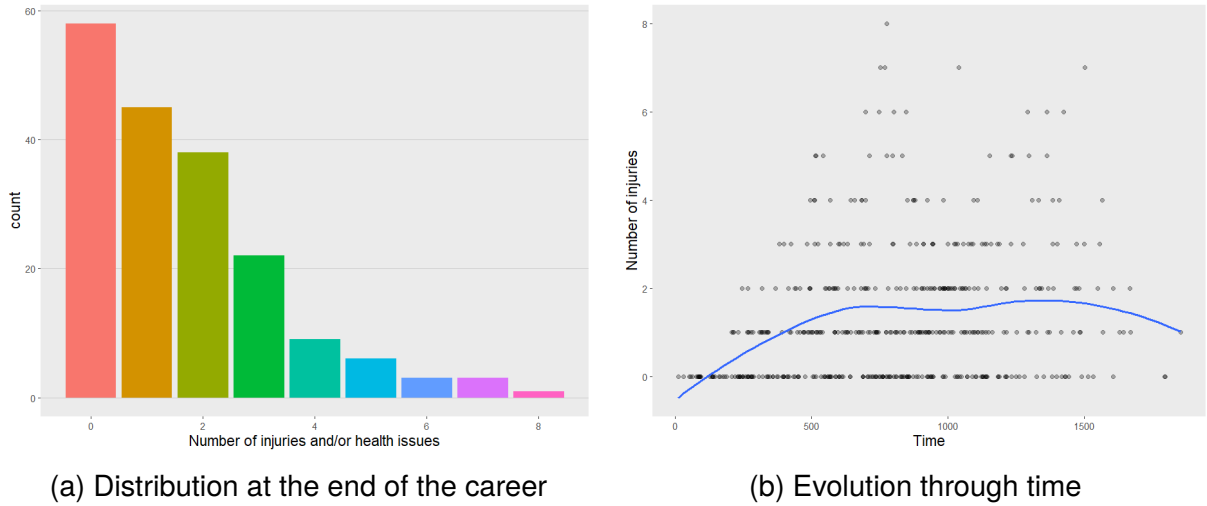


Figure 40: Time-dependent covariate *Health_Inj*

In Figure 40 (a), we can see that the distribution is right-skewed, meaning that the majority of the players suffered 1 or no injury or health issue. Figure 40 shows that the number of injuries and/or health issues tends to increase until approximately the 700th game. Then it seems to stay constant until around game 1500. After that, it tends to decrease. This pyramidal shape was expected. In the beginning of a career, the number of injuries and/or health issues can only increase by construction. And since every player in the dataset plays at least 505 games, the global tendency cannot decrease. During the constant part, the increase in injuries and/or health issues might be compensated by players retiring (or being censored). The decrease at the end probably shows that the number of players retiring (or being censored) outweighs the increase in injuries and/or health issues. Once again, these two last hypotheses cannot be assessed based only on these plots.

4.3. Methods & Results

4.3.1. Kaplan-Meier

In this section, we do not use the time-dependent covariates. As explained in section 2.2.1, the Kaplan-Meier estimator cannot be used to compare the survival functions of different groups based on a time-dependent variable. Therefore, variables *Avg_min_pgm* and *Health_Inj* are not explored using non-parametric models. For the survival time, we again use the largest value of *Total_games_end* for each player.

We plot the Kaplan-Meier survival curves for each of the 9 remaining variables. These curves are created by computing the Kaplan-Meier estimator for each level v of a variable k and for each event time t in each level. The estimators are then plotted in relation to the number of games played. In order to plot survival curves for the 6 quantitative variables, we discretize them as follows:

- *Draft_pos* $\Rightarrow \{[1 - 3], (3 - 10], (> 10)\}$

- $Ind_Success \Rightarrow \{[0 - 5], (5 - 10], (10 - 15], (> 15)\}$
- $Titles \Rightarrow \{[0], [1 - 2], (> 2)\}$
- $Height \Rightarrow \{(\leq 190], (190 - 200], (200 - 210], (> 210)\}$
- $Weight \Rightarrow \{(\leq 95], (95 - 105], (> 105)\}$
- $Birth_year \Rightarrow \{[1920's], [1930's], [1940's], [1950's], [1960's], [1970's], [1980's], [1990's]\}$

The grouping bounds were chosen both for interpretability and to avoid excessively unequal group sizes. Figures 41 to 48 show the Kaplan-Meier survival curves for each variable independently. Even if differences appear on the following plots, their significance cannot be assessed based only on these plots.

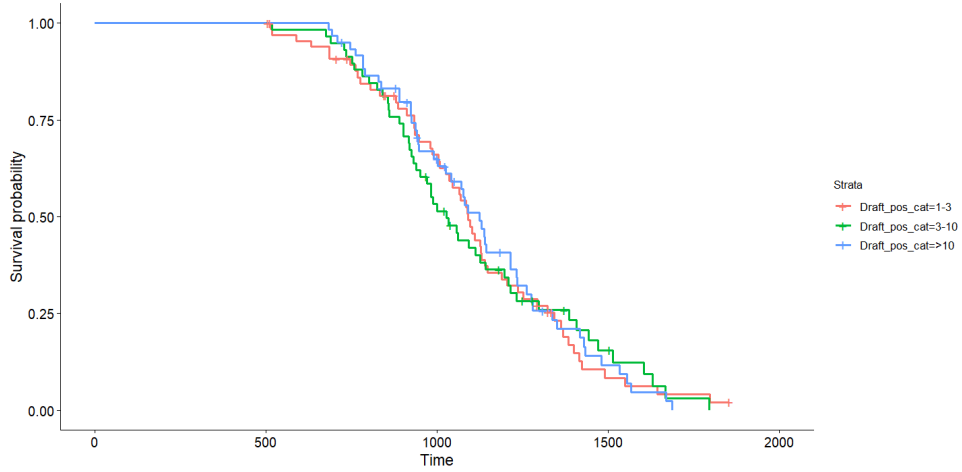
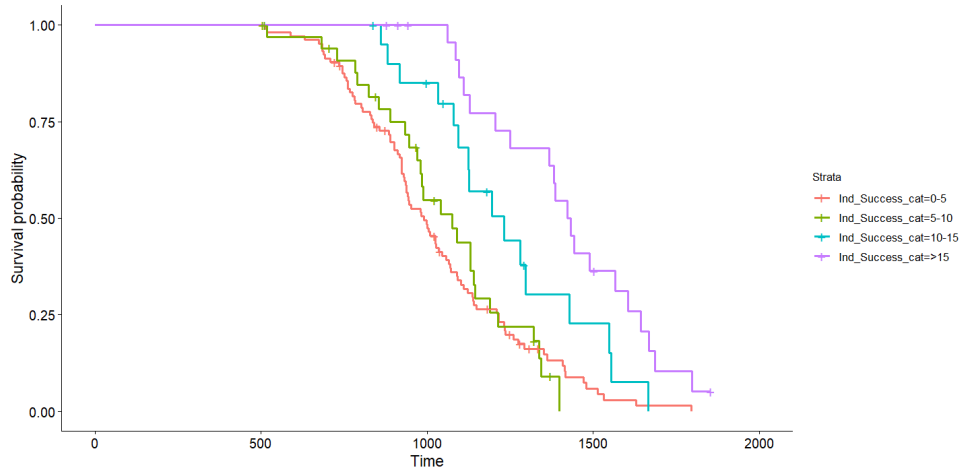


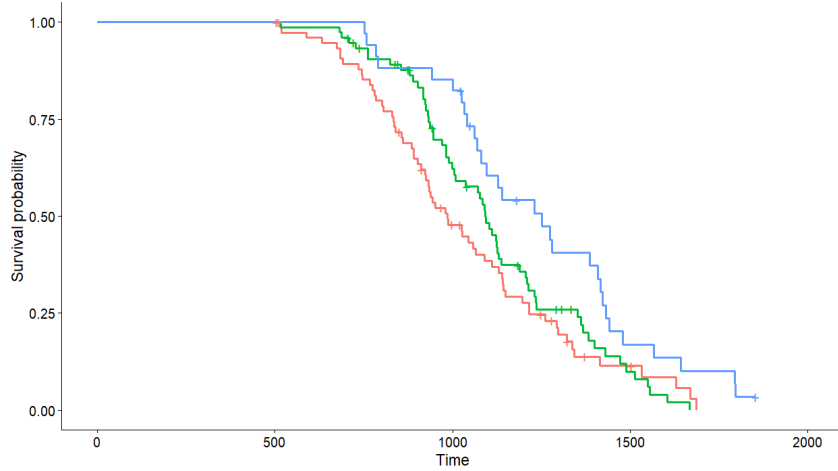
Figure 41: Survival Curves per draft position (discretized)

In Figure 41, the survival curves are close together. It means that the expected career length does not seem to be impacted by the draft position. Recall that we selected this variable to quantify the performances of a player prior to the NBA. Therefore, the survival curves suggest that the level of a player before he enters the NBA does not impact his survival probability.

Figure 42 (a) shows that players winning more individual trophies tend to have longer careers, while Figure 42 (b) shows the same effect for collective titles. We group these two variables because they were selected, *inter alia*, to quantify the level of a player. Thus, these plots show that better players seem to have longer expected careers. Another interesting finding is that, among players having won more than 15 individual trophies in their careers, none retire before approximately the 1100th game (although there are some censored players before that).



(a) Level as the number of individual trophies (discretized)



(b) Level as the number of collective trophies (discretized)

Figure 42: Survival Curves per player's level

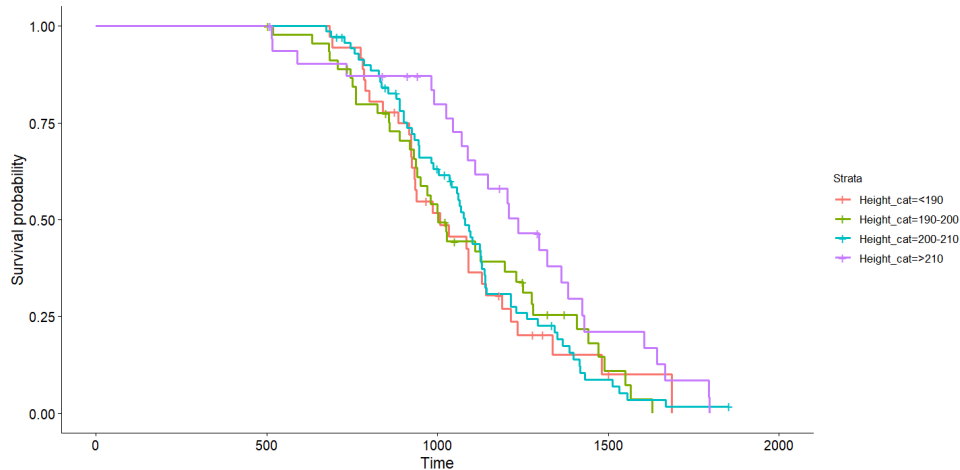


Figure 43: Survival Curves per height (discretized)

Regarding the height, Figure 43 displays a gap between players measuring more than 210 cm and those measuring less. The other survival curves are close together, meaning that the height differences for players less than 210 cm tall do not seem to have an impact on survival probability.

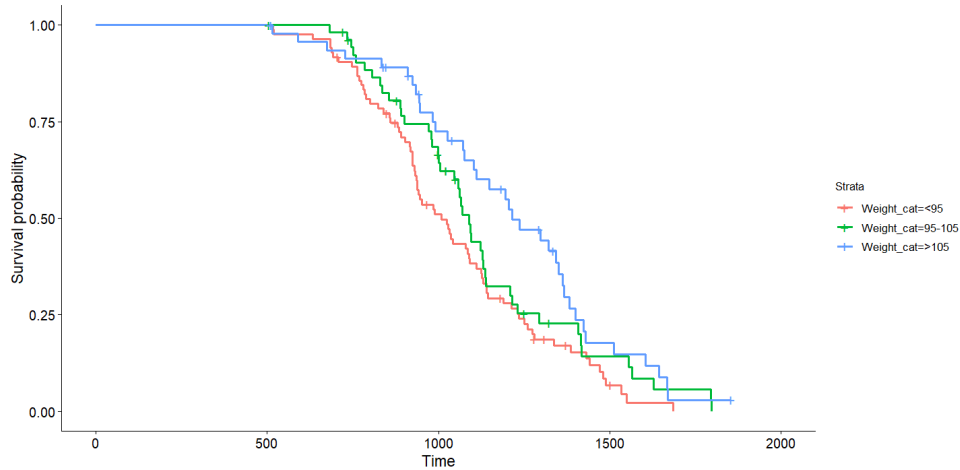


Figure 44: Survival Curves per weight (discretized)

In Figure 44, we can see that heavier players seem to have longer NBA careers than lighter ones.

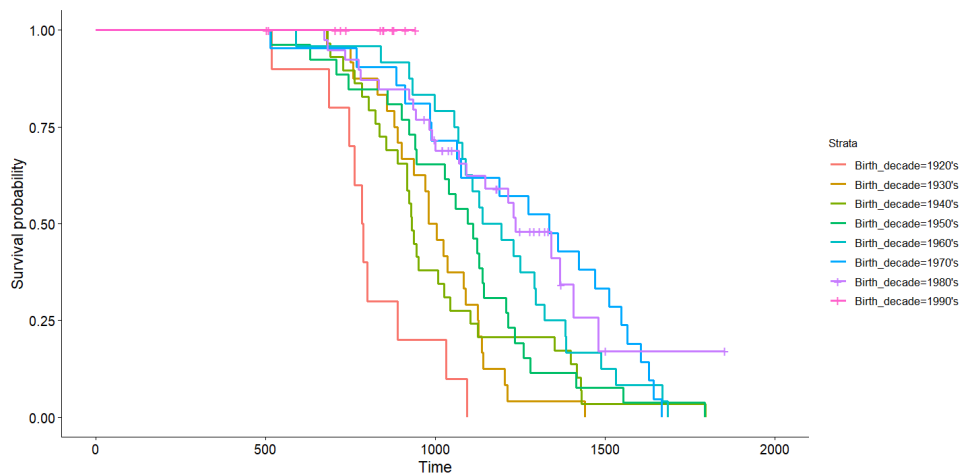


Figure 45: Survival Curves per birth year (discretized)

Figure 45 shows that the expected career length seems to increase through time. Note that all the players born in the 1990s are censored. This was expected, since by construction no player born in the 1990s can be older than 35 as of the end of the study (July 2025).

According to Figure 46, non-American players tend to outlast American ones in terms of career length. The staircase shape of the non-American players' curve highlights the small number of non-American players in the dataset.

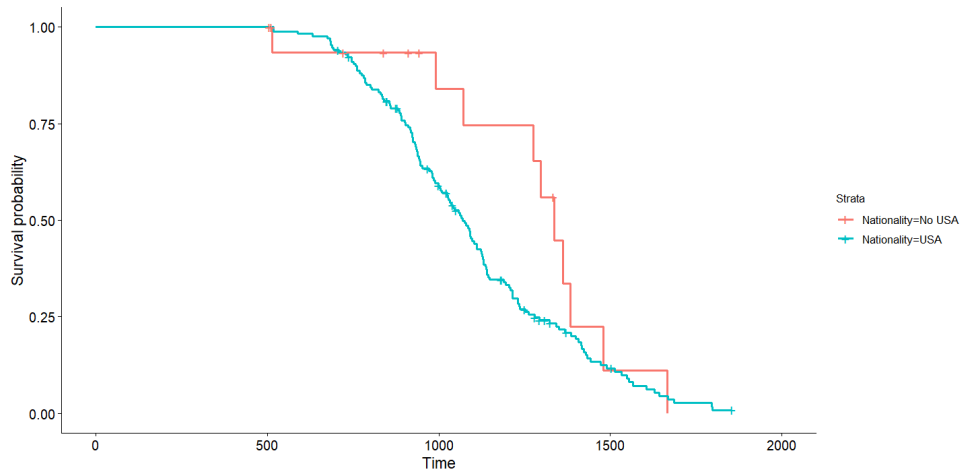


Figure 46: Survival Curves per nationality

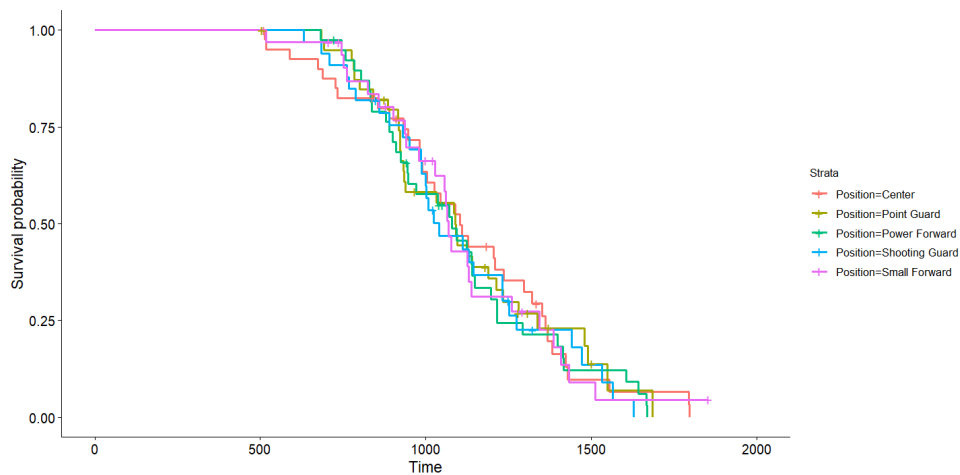


Figure 47: Survival Curves per playing position

Based on Figure 47, there does not seem to be differences in expected career length between the five playing positions.

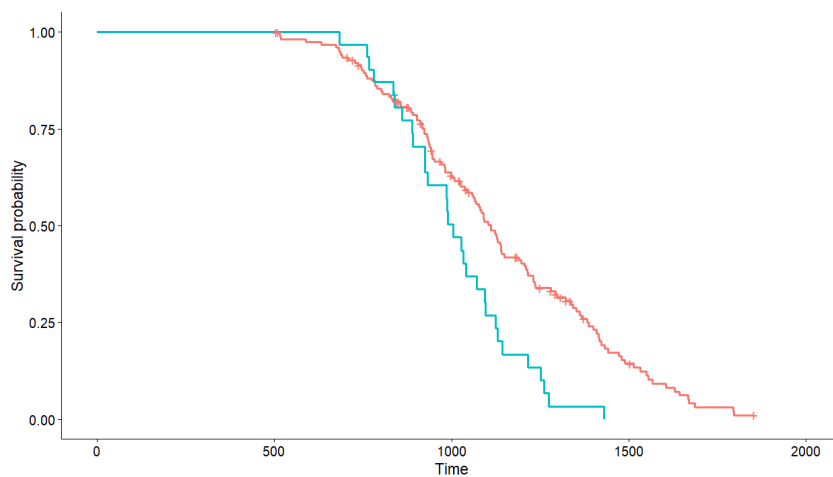


Figure 48: Survival Curves by whether the player spent a year outside the NBA

In Figure 48, players having spent at least one year outside the NBA seem to have

higher survival probability than players who have not, until around game 900. After that, the survival curves suggest the opposite.

In Figures 41 to 48, the survival curves intersect many times with each other. This might suggest that the Proportional Hazards Assumption (PHA) does not hold for the dataset. However, one cannot assess this hypothesis based only on these plots. Moreover, in the previous plots, we used a transformed dataset that does not account for time-dependent covariates. Therefore, further tests are needed before reaching the conclusion that the PHA is violated. This will be the goal of section 4.3.3.

4.3.2. Proportional Hazards Model (PHM)

In this section, we use all the covariates, including the time-dependent ones. The time-to-event variables are both *Total_games_start* and *Total_games_end*. We start by running an AIC selection to check which covariates should be included in the model. The final model has an AIC of 3477.2 and includes seven covariates: the draft pick (*Draft_pos*), the number of individual trophies (*Ind_Success*), the weight (*Weight*), the birth year (*Birth_year*), the binary variable representing if a player spent at least one year outside the NBA (*Out_NBA*), the time-dependent average number of minutes per game (*Avg_min_pgm*) and the time-dependent number of injuries and/or health issues (*Health_Inj*). Regarding interactions between the covariates, we decide not to include them because the model becomes too complex (ten covariates and eleven interactions after AIC selection).

Four covariates are not selected: the nationality (*Nationality*), the playing position (*Position*), the number of collective titles (*Titles*) and the height (*Height*). Regarding the height and the number of titles, we found in section 4.2.3 that they had respectively a strong correlation of 0.812 with the weight, and a small to moderate correlation of 0.323 with the number of individual trophies. They both might provide similar knowledge to their correlated variables. Therefore, dropping them might not lead to major information loss, while keeping all four increases the complexity of the model. Concerning the playing position, we saw in section 4.3.1 that the survival curves for each playing position were very close together and intersected many times. Dropping the nationality was not expected since the survival curves shown in section 4.3.1 seemed to show an impact on expected career length. However, the small number of non-American players in the dataset, which we noted in section 4.2.4, might be the reason.

Next, we fit a Proportional Hazards Model through the data using only the selected covariates. Note that the reference level for *Out_NBA* is 0 (= the player has not spent at least one year outside the NBA). Table 13 displays the parameters of the model. Using a type 1 error probability $\alpha = 0.05$, five out of seven covariates have a significant association with the risk of retirement. A one-unit increase in *Draft_pos* (= being drafted one position later) lowers the risk of retiring after a given number of games by 0.8%, assuming that all other covariates stay constant. At a given time, a one-minute increase

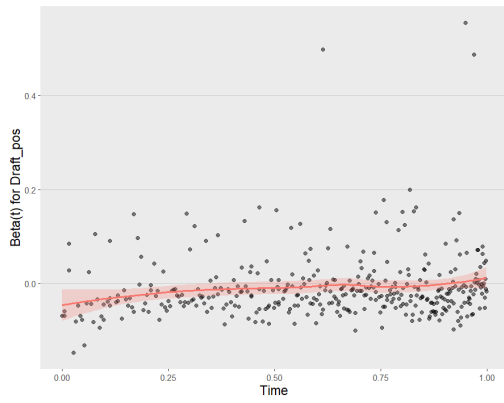
in average minutes played per game is associated with a 6.7% higher risk of retirement, conditional on the player's covariates at that time. Regarding *Ind_Success*, the effect of a precise increment should not be interpreted because the trophies summed in this variable differ substantially in prestige and are not commensurable units. We only included this variable to quantify the level of the player. Therefore, only global association should be noted: a better player (based on his individual achievements) has a lower risk to retire after a given number of games compared to another player with the same values on all other covariates. At a given time, having one more injury and/or health issue is associated with a 23.4% higher risk to retire, conditional on values of the covariates at that time. A player having spent at least one year outside the NBA has a 68.1% higher risk to retire after a given number of games. The weight and the birth year do not seem to have a significant impact on the risk of retirement.

Table 13: Cox model

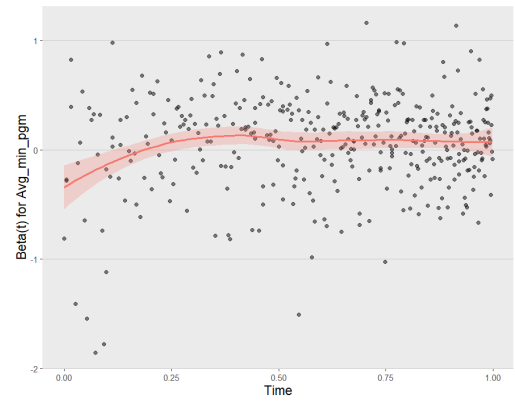
	$\hat{\beta}$	$\exp(\hat{\beta})$	<i>lower.95</i>	<i>upper.95</i>	<i>p.val</i>
<i>Draft_pos</i>	-0.008	0.992	0.986	0.998	9.892×10^{-3}
<i>Avg_min_pgm</i>	0.065	1.067	1.017	1.119	8.625×10^{-3}
<i>Ind_Success</i>	-0.061	0.941	0.920	0.962	1.174×10^{-7}
<i>Weight</i>	0.009	1.009	0.996	1.022	0.189
<i>Birth_year</i>	-0.007	0.993	0.986	1.001	0.082
<i>Health_Inj</i>	0.210	1.234	1.101	1.382	2.924×10^{-4}
<i>Out_NBA</i>	0.519	1.681	1.258	2.245	4.394×10^{-4}

4.3.3. Proportional Hazards Assumption (PHA)

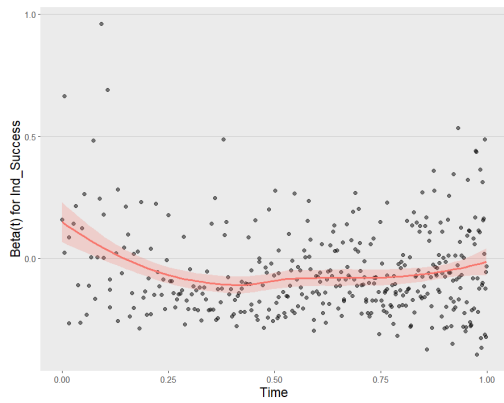
In the following, we look at the Schoenfeld residuals to assess if they are independent of time. This allows us to check if the PHA holds for the model. Figures 49 (a) to (g) are the plots of Schoenfeld residuals relative to the number of games for each covariate.



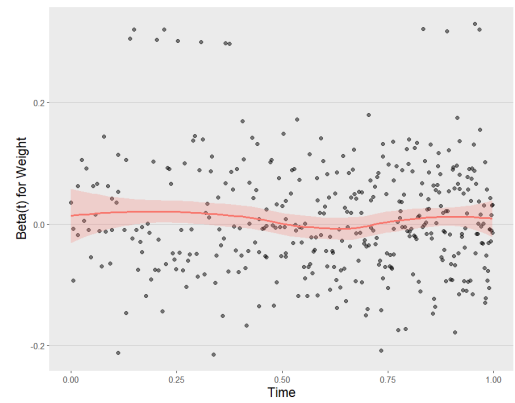
(a) Draft pick



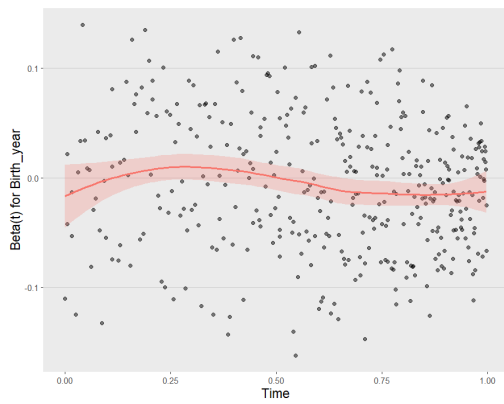
(b) Average minutes played per game



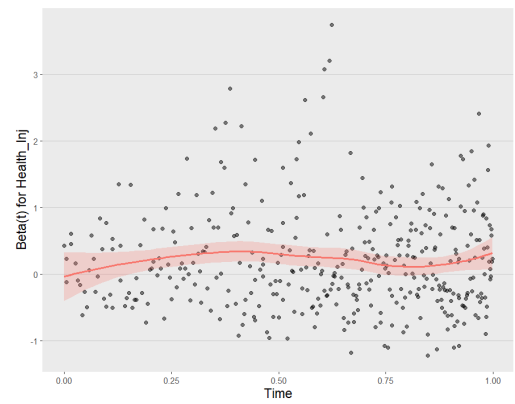
(c) Number of individual trophies



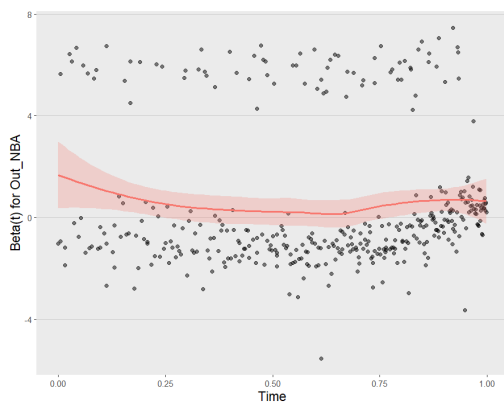
(d) Weight



(e) Birth year



(f) Injuries and/or health issues



(g) Year outside the NBA

Figure 49: Schoenfeld residuals through time

The smoothed curves seem horizontal only for *Weight* (Figure 49 (d)) and *Out_NBA* (Figure 49 (g)). It means that the Schoenfeld residuals are independent of time for these two covariates, whereas they appear non-horizontal for the other five, suggesting the residuals are not independent of time for those covariates. In order to support these findings, we test the hypothesis of straight regression lines with zero slopes, both globally and locally for each covariate. Table 14 displays the p-values of each test. Based on these p-values, three covariates seem to violate the PHA: the draft pick, the birth year and the number of injuries and/or health issues. It is supported by Figures 49 (a), (e) and (f). As expected from the plots, the p-values are non-significant (with $\alpha = 0.05$) for the weight and *Out_NBA*. This suggests that they do not violate the PHA. Regarding the average minutes per game and the number of individual trophies, it is less clear. Although their p-values are both non-significant, Figures 49 (b) and (c) seem to show, on their left side, that the Schoenfeld residuals are not constant through time. Finally, the p-value of the global test is significant, suggesting that the PHA does not hold for the model.

Table 14: Hypothesis test on Schoenfeld residuals

	<i>p.val</i>
<i>Draft_pos</i>	0.016
<i>Avg_min_pgm</i>	0.112
<i>Ind_Success</i>	0.582
<i>Weight</i>	0.156
<i>Birth_year</i>	1.6×10^{-4}
<i>Health_Inj</i>	0.007
<i>Out_NBA</i>	0.490
<i>GLOBAL</i>	6.7×10^{-4}

One way to deal with the violation of the PHA is to use a stratified model. Since *Birth_year* shows the strongest evidence of PHA violation among the covariates, it is a strong candidate for stratification. Furthermore, we already have a discretized version of the variable, the birth decade:

$$Birth_year \Rightarrow \{[1920's], [1930's], [1940's], [1950's], [1960's], [1970's], [1980's], [1990's]\}$$

Table 15 displays the p-values of the hypothesis tests on Schoenfeld residuals for the stratified model. The p-values of all local tests are non-significant (with $\alpha = 0.05$), except for *Draft_pos* which is below the threshold. Consequently, one should be cautious when interpreting its coefficient. The p-value of the global test is also non-significant for the stratified model, suggesting that the PHA holds for the stratified model.

Table 15: Hypothesis test on Schoenfeld residuals for the stratified model

	<i>p.val</i>
<i>Draft_pos</i>	0.010
<i>Avg_min_pgm</i>	0.376
<i>Ind_Success</i>	0.412
<i>Weight</i>	0.647
<i>Health_Inj</i>	0.133
<i>Out_NBA</i>	0.562
<i>GLOBAL</i>	0.052

Table 16 shows the coefficients of the new stratified model. They are broadly similar in sign and order of magnitude, though *Health_Inj* shows a larger shift in the stratified model than in the non-stratified one, and *Out_NBA* shows a smaller one. There are only two differences. First, the exponentials of the coefficients now quantify the relative risks to retire for a given decade (after a given time t). Second, we do not have coefficients for the stratified covariate anymore, meaning that its impact on career length is not measured by the stratified models. The association between the weight and the risk of retirement stays non-significant.

Table 16: Stratified Cox model

	$\hat{\beta}$	$\exp(\hat{\beta})$	<i>lower.95</i>	<i>upper.95</i>	<i>p.val</i>
<i>Draft_pos</i>	-0.008	0.992	0.986	0.999	2.014×10^{-2}
<i>Avg_min_pgm</i>	0.054	1.055	1.004	1.108	3.289×10^{-2}
<i>Ind_Success</i>	-0.059	0.942	0.922	0.963	1.604×10^{-7}
<i>Weight</i>	0.008	1.008	0.997	1.019	0.150
<i>Health_Inj</i>	0.269	1.308	1.159	1.477	1.387×10^{-5}
<i>Out_NBA</i>	0.407	1.502	1.145	1.970	3.286×10^{-3}

4.3.4. Survival prediction

We use two models: the stratified and non-stratified Cox regression adapted for time-dependent covariates. To assess the performance of the models, we compute a time-dependent generalization of the C-index. Since we only compute a concordance index, we will only assess the discrimination performance, not the prediction accuracy.

To do so, we randomly split the dataset in two. 20% of the players (not the observations) are selected as the test set, and the other 80% are selected as the training set. It leads to a very small test-set size of around 37 players. Since the distribution of the censoring indicator is very unbalanced (see section 4.2.2), each split is stratified on *Retired*, meaning that the proportion of censored players is the same in the original dataset, each training set and each test set. After the split, we fit the models on the training set and compute their prediction capacity on the test set. In order not to

depend on the data split, we repeat this procedure 500 times and compute the mean and the quantile-based 95% confidence interval. Table 17 summarizes the results for both models. The stratified and non-stratified models are very close to each other in terms of concordance index. It means that, based on this dataset, both models seem to be equivalent when discriminating career length of the best NBA players. The very large confidence intervals also show the variance of the concordance index. This is expected given the small test-set size: with so few observations, the concordance index computed on a single split can vary substantially depending on which players happen to be held out. It suggests that the discrimination performance might depend on the data split.

Table 17: Concordance index of the models

	<i>2.5th pct</i>	<i>Mean</i>	<i>97.5th pct</i>
<i>Stratified Cox</i>	0.557	0.646	0.732
<i>Non – stratified Cox</i>	0.563	0.637	0.709

4.4. Discussion

We have seen that some variables tend to be significantly associated with the career length of the best NBA players. First, according to Cox models (both stratified and unstratified), the draft position of a player is positively associated with his career length. Since the draft position was selected, *inter alia*, as a marker of a player's level before entering the NBA, it means that the better a player is before the NBA, the shorter his expected career length will be. However, the draft pick was the only covariate still violating the Proportional Hazards Assumption in the stratified model. Furthermore, the distribution of the variable was extremely skewed, with a few extreme values, and the Kaplan-Meier curves did not support this association. This suggests that the association between career length and the draft position might not be robust and should be interpreted cautiously.

Second, players with larger average time played per game seem to have a higher risk of retirement. Both Cox models have shown this association. Since one of the reasons we selected this variable was to quantify the level of a player, this outcome was not expected. It seems to show that better players tend to have shorter careers. However, if we focus on the variable itself, not what it represents, it might be more relevant. Players spending more time on court each game possibly tire out faster during their career. This might suggest that, even if a player is very good, it might be positive to limit his playing time. This hypothesis is purely speculative and cannot be assessed based only on these results.

Third, individual success and career length seem to be positively associated. As explained in section 4.3.2, the direct effect of this variable should not be interpreted. For one thing, the different trophies summed in the variable differ substantially in pres-

tige, for example, the MVP (the best player of the regular season) and a selection in an "All-NBA Team" ($\approx$ the fifteen best players of the regular season) both count as the same value in the variable. For another thing, although *Ind_Success* is treated as a fixed covariate, its value reflects the player's entire career, including trophies won after the interval being modeled. Since the dataset uses a counting-process structure with multiple rows per player, this effectively lets early intervals "see" a player's future success, which is a look-ahead bias. For instance, Michael Jordan's first row in the dataset approximately corresponds to his first season, but we already know that he will win 32 trophies during his entire career (Basketball-Reference.com, n.d.b). This look-ahead bias likely inflates the apparent protective effect of *Ind_Success*, since a player's cumulative future success is, by construction, associated with him still being active in the league long enough to keep accumulating it. Lastly, note that the dataset already focuses on the best players in the history of the NBA. Consequently, when we speak of the level of a basketball player, we speak of the level compared to the other players in the dataset. Therefore, both Cox models and the Kaplan-Meier survival curves seem to show that the very best players tend to have a lower risk of retirement than the good players.

Fourth, the career length of the best NBA players seems to be negatively associated with the number of injuries and/or health issues they suffer during their career. At a given time, a player who suffered one more injury and/or health issue than another has approximately a 23.4% higher risk to retire, assuming that they have the same values on all other covariates at that time. If we stratify the model on the birth decade, the difference rises to around 30.8%. This emphasizes the strength of the association between career length and health.

Fifth, spending at least one year outside the NBA is positively associated with the risk of retirement. This association is very strong: players having spent at least one year outside the NBA have a 68.1% higher risk to retire, conditional on values of the other covariates. However, this difference drops to 50.2% when we stratify the model on the birth decade. The Kaplan-Meier survival curves seemed to support this association for survival times beyond 900 games. As with our speculation regarding average time on court, this result might be an artifact of the dataset's construction. Most of the players with value 1 on *Out_NBA* are players who played in another league. A part of their professional career is not accounted for when we measure the NBA career. This suggests that further studies exploring the entire professional career length could be relevant.

Sixth, the associations between the risk of retirement and both the weight and the birth year are less defined. Although covariate *Weight* was selected by AIC and survival curves seemed to show an impact on survival probability, its association with career length was not significant according to both Cox models (stratified and non-stratified). Regarding the birth year, the AIC selection and the Kaplan-Meier survival curves also seemed to show an association with career length. However, the non-

stratified Cox model displayed evidence of a non-significant association. Moreover, *Birth_year* showed the strongest evidence of PHA violation, and when we explored a model stratified on the birth decade (the discretized version of *Birth_year*), the PHA seemed to hold. This suggests that the impact of the birth year might be better quantified as a stratified variable than as a covariate in the model. Further studies exploring the cohort effect on career length of NBA players are needed to confirm it.

Seventh, during the AIC selection, four covariates were dropped: the nationality, the playing position, the number of collective titles and the height. This suggests that their associations with the risk of retirement might be negligible. Regarding the nationality, note that the distribution was very unbalanced. This could be the reason the variable was not selected for the non-stratified Cox model, although the Kaplan-Meier survival curves seemed to show that non-American players tend to have a larger survival probability than American players. Concerning the number of titles and the height, we saw in section 4.2.3 that they are correlated to the number of individual trophies (0.323) and the weight (0.812) respectively. This could mean that *Titles* and *Height* do not provide additional information about career length, but we cannot confirm this hypothesis based only on these results. Both the Cox model and survival curves showed that the playing position does not seem to be associated with the career length of the best NBA players.

Lastly, we have compared the predictive performance of two models: a non-stratified Cox model and a stratified one. Both seemed to perform better than random, although the concordance indexes were not excellent. Moreover, the empirical 95% confidence intervals were both quite large, suggesting that discrimination performance depend on the data. Recall that we used a small test set (around 37 players). A larger sample would likely give smaller intervals, but since we decided to focus on elite players, including more players could undermine this objective. We found no clear difference between the two models in terms of discrimination performance. However, since the non-stratified Cox model seemed to violate the PHA while the model stratified on the birth decade did not, we recommend using the latter.

5. Global discussion

Recall that, in this thesis, we have one main objective and multiple smaller ones. First, we want to find which covariates have strong associations with the career length of basketball players. Then, we try to explore the following topics: the measure of career length, the use of time-dependent covariates, the separation of elite players from the rest, and the use of adapted regression techniques.

5.1. Outcomes

5.1.1. Which covariates have strong associations with career length ?

In the first dataset, we saw that the career lengths of NBA players are associated with their age at their first NBA game, their height and their dominant hand. In the second dataset, we found that the career length of elite NBA players seems to be associated with their relative level (measured through the number of individual trophies and the average playing time), and whether they spent at least one year outside the NBA during their career. The associations between career length and both the height and the number of trophies won were previously found by Fynn and Sonnenschein (2012). Lawler and Lawler (2011) have previously shown the association between career length and the dominant hand.

There are covariates whose associations to career length seem negligible. According to both datasets, the nationality or place of birth does not seem to impact the career length. The second dataset also showed the absence of association between the career length of elite NBA players and their relative level (measured through the number of collective titles). This suggests that, when assessing the association between career length and level, quantifying the level of a player by using the number of collective titles he won might not be relevant.

Some covariates showed associations that are less obvious. The playing position does not seem to impact the career length of elite players (second dataset), but its association with career length of all NBA players was unclear. This outcome does not support previous findings (Fynn and Sonnenschein, 2012). The birth year was not investigated in the first dataset, but it seems to have an impact on the career length of elite players. Both the playing position and the birth year led to interesting results when they were used to fit stratified models. Regarding the weight, in the first dataset, it was dropped from the model measuring time as the number of games, but kept and significant when time was measured as the number of seasons. In the second dataset, it was kept but non-significant. This suggests that the impact of the weight depends on the measure of career length. The association between draft position and career length seems to be significant. This suggests that the level of a basketball player before his professional debut impacts the length of his career. However, the use of draft pick as

single marker of the level of a player before entering the NBA is insufficient.

Lastly, in the second dataset, we introduced two time-dependent covariates: the average playing time and the number of injuries and/or health issues. Both of them seem to have a significant association with the career length of elite NBA players. Moreover, the association between a basketball player's career length and the number of injuries and/or health issues he suffered during his career seems very strong. The association between career length and health is well-supported (Lavalley et al., 1997; Agresta et al., 2008; Ristolainen et al., 2011; Martin et al., 2021; Vretaros, 2022). Regarding the average number of minutes played per game, we selected this variable to quantify the level of a player. However, since it does not suffer from the same construction problems as the number of individual trophies, we can also interpret the direct association between average playing time and career length. Surprisingly, the association seems to be negative. If we use the variable as a marker of a player's level, this outcome does not support our previous finding concerning individual success. Therefore, we reach the conclusion that, in the context of elite players' career lengths, the average number of minutes played per game might not be a relevant marker of a player's level.

5.1.2. The measures of career length

In the first dataset, we compared the use of two measure of career length: the number of games played in the NBA and the number of seasons played in the NBA. The associations between career length and the covariates were globally the same. The main difference seems to be the significance of the association with the weight when time is measured as the number of seasons, while the variable was dropped when time is measured as the number of games. The main difference appeared for survival predictions. Measuring career length as the number of games tends to slightly increase the prediction accuracy, but the discrimination performance decreases. Measuring career length as the number of seasons leads to the opposite. Based on these outcomes, we reach two conclusions. First, one must choose the unit of time used to measure career length carefully, depending on the task. Second, other units of time should be investigated. For example, the number of minutes played in the NBA could be interesting as it is an even more accurate measure of career length.

5.1.3. The focus on elite players

In the first dataset, we saw that the impact of the covariates seems negligible for predicting long (and short) NBA careers. In the second dataset, we focused on the best players in the history of the NBA whose careers are generally longer. Based on the outcomes of the second dataset, many covariates seem to be significantly associated with the career length of elite NBA players (see section 5.1.1). In terms of predictive performance, the two models evaluated on the second dataset seemed to perform moderately well. They also seemed to perform slightly worse than the models evaluated on

the first dataset, though computed via a different, time-dependent formula. However, the difference in discrimination performance might be due to the right-skewed distribution of survival time. Indeed, if the models perform better when predicting medium career lengths, the number of players with actual medium-length careers has an impact on the performance measure. Therefore, we cannot conclude that focusing only on elite players is relevant or irrelevant based only on the results of this thesis.

5.1.4. RSF and SSVM

In the first dataset, we tried to implement two other survival analysis techniques: Random Survival Forests (RSF) and Survival Support Vector Machine (SSVM). In terms of global predictive performance, we found no clear difference between these two models and Cox regression. Since SSVM is more complex and more time-consuming, we do not recommend using this model based on the results of this thesis. Nevertheless, Cox regression and RSF seem to be equally relevant survival analysis techniques.

5.2. Limitations

5.2.1. Optimization process of the SSVM

In the first dataset, during the optimization process of the SSVM, we were compelled to limit our search of the best values for the hyperparameters. The main reason was the time-consuming aspect of SSVM. Indeed, depending on the values of the hyperparameters, the time taken by our algorithms to evaluate the performance increased greatly, and we only explored three kernel functions when there exist infinite possibilities. In addition, we optimized the hyperparameters separately (except for the kernel functions), meaning that we did not account for potential interactions between them. Finally, we used a randomized optimization which is rarely the best way to proceed. We consider that all these elements go beyond the scope of this thesis. Our goal was only to assess if SSVM could be a relevant model to predict the career length of NBA players. However, since SSVM performed as well as the other models, one could argue that an SSVM optimized more thoroughly could have performed better. To a lesser extent, the same reasoning can be applied to RSF where we limited ourselves to optimizing two hyperparameters when others exist.

5.2.2. Measure of career length

As explained before, we decided to measure time using two different units: the number of games and the number of seasons. One could argue that the units "one NBA game" and "one NBA season" are not constant. In our datasets, whether a player plays 35 or 45 minutes in a game, it counts as the same increase of 1 game. On a single game, the difference is not that important, but over a career of more than 1,000 games,

this can hide major differences. In section 5.1.2, we proposed to use a new unit in future research: the number of minutes played during the entire career. This would address the issue.

5.2.3. Generalization

In both our datasets, we only focused on the NBA career of basketball players. This leads to two different limits. First, we deliberately ignored parts of some players' career. Although there were good reasons supporting this choice (easy access to reliable statistics, the "closed" league aspect of the NBA, etc.), this could prevent the results of this thesis, which are based on NBA career length, from being generalized to the more global concept of professional basketball career length. The second limit caused by our focus on the NBA is the difficulty of generalizing to other skill levels. The NBA is globally recognized as the best basketball league in the world. Therefore, some of our results might not apply to other leagues where the players are not as good. The absence of major differences in which covariates were associated with career length between our two datasets tends to show that elite players and "regular" professional players might not be too different in the context of career length. Nonetheless, this hypothesis cannot be assessed based only on the results of this thesis.

5.3. Future research

To our knowledge, this thesis was the first to combine survival analysis and time-dependent covariates to model the career length of professional basketball players. Therefore, any study aiming to reproduce or contradict our results would be relevant. The same reasoning can be applied to the use of Random Survival Forests and Survival Support Vector Machine to predict the career length of professional basketball players.

More specifically, we believe it would be worthwhile to further investigate the relations between career length and both playing position and birth year. The outcomes concerning the playing position differed between our two datasets. Based on the results of the second dataset, the association is negligible, but based on the first one, the non-stratified Cox models and Kaplan-Meier curves suggested a relation between playing position and career length, even though this covariate was also the primary source of the PHA violation. In addition, Fynn and Sonnenschein (2012) concluded that the playing position is one of the best variables to explain the career length of basketball players. Regarding the birth year, we also found a relation with career length when a stratified model with one stratum per birth decade was used. Therefore, further studies should explore the complex relations between the playing position and the career length, and between the birth year and the career length.

We also think that further studies modeling the career length of basketball players should investigate new measures of time such as the total number of minutes played

during the entire career. This new unit could address the issue of inconsistent unit increase we developed in section 5.2.2.

Finally, further studies should explore new covariates. For instance, we join Fynn and Sonnenschein (2012) in their belief that using more information from the pre-professional career of basketball players (college, high school, junior, etc.) could improve the results. In our study, we only used the draft pick as marker of the pre-professional level of a player, and it seemed to be significantly associated with career length. One solution could even be to consider that the basketball career starts with the first basketball game in college, or in high school. However, we recommend being extremely cautious with pre-professional statistics as they are not always reliable. Other possible covariates are, for example, gender or statistics such as the number of points, of rebounds, of passes, etc.

5.4. Conclusion

The goal of this thesis was to apply survival analysis to the specific case of professional basketball career length. To do so, we used two different datasets and applied multiple survival analysis techniques on them: Kaplan-Meier, Cox regression, Random Survival Forests and Survival Support Vector Machine. These techniques allowed us to explore the associations between career length of NBA players and a set of covariates, and to assess predictive performance. We ended this thesis by discussing the limitations and the opportunities for future studies.

This thesis has shown that the career length of NBA players are associated with multiple covariates such as the age of the players at their first NBA game, the height of the players, the dominant hand of the players, the relative level of the players, the number of injuries and/or health issues they suffer during their career, and whether they spent at least one year outside the NBA during their career. We have also demonstrated that the results may depend on the unit used to measure career length. In addition, we showed that Random Survival Forests and Survival Support Vector Machine can be used to predict career length, although they do not seem to perform better than the classical Cox regression. Generally, this thesis has shown that survival analysis is a good way to model the career length of basketball players, even though the predictive performance is not excellent. Future studies should further investigate the career length of professional basketball players.

Appendices

A. List of the packages

Here is the list of all the R packages we used in this thesis.

- *survival*
- *KMsurv*
- *survminer*
- *dplyr*
- *MASS*
- *summarytools*
- *ggplot2*
- *DescTools*
- *caret*
- *pec*
- *randomForestSRC*

Here is the list of all the Python packages we used in this thesis.

- *numpy*
- *pandas*
- *random*
- *sklearn.model_selection*
- *sklearn.preprocessing*
- *sksurv.util*
- *sksurv.svm*
- *sksurv.metrics*
- *matplotlib.pyplot*

B. List of players who died before the end of their career

Here is the list of all the players in the history of the NBA who died before the end of their career. In the first dataset, we consider them as censored observations.

- Ricky Berry
- Bill Buntin
- Jason Collier
- Yinka Dare
- Bryce Dejean-Jones
- Alphonso Ford
- Terry Furlow
- Eddie Griffin
- Reggie Harding
- Tyler Honeycutt
- Reggie Lewis
- Fernando Martin
- Fab Melo
- Drazen Petrovic
- Bobby Phills
- Bill Robinzine
- Malik Sealy
- Robert Traylor
- Nick Vanos
- Jackson Vroman

C. Basketball related information

Here is all the information related to basketball that one might find useful to understand this thesis.

Historically, there are five playing positions in basketball (Jr. NBA, 2016; Under Armour, n.d.):

- The point guards (PG or 1) usually run the offense. They are often the best ball handlers and the best passers on the team.
- The shooting guards (SG or 2) are the best shooters on the team. They are usually the second-best ball handlers on the team.
- The small forwards (SF or 3) are the most versatile. They need to be able to do everything on the court.
- The power forwards (PF or 4) play closer to the basket. They are often strong people who rely on their strength to score.
- The centers (C or 5) are usually the best defenders on the team. They are often the players who play the closest to the basket, both on offense and on defense.

In addition, there is often a relation between a player's position and their height. The smallest are usually the point guards (1), and they get taller as their position number increases. The tallest are usually the centers (5). However, this is a global description and not all players fit these descriptions, especially the best ones. Moreover, the role of each position is constantly evolving with the way basketball is played. Lastly, even though this is the widely accepted list of playing positions, others exist (point forward, combo guard, etc.).

D. Summary statistics (first dataset)

Here are the summary statistics for the time variable and the quantitative covariates of the first dataset.

Statistic	<i>g</i>
Mean	260.00
Std. Dev	301.70
Min	1.00
Q1	32.00
Median	122.00
Q3	427.00
Max	1611.00
Skewness	1.33
Kurtosis	0.96

Table 18: Summary statistics of *g*

Statistic	<i>years</i>
Mean	5.36
Std. Dev	4.55
Min	1.00
Q1	1.00
Median	4.00
Q3	9.00
Max	23.00
Skewness	0.93
Kurtosis	−0.12

Table 19: Summary statistics of *years*

Statistic	<i>ageleft</i>
Mean	23.32
Std. Dev	2.14
Min	17.45
Q1	22.17
Median	22.93
Q3	24.10
Max	45.75
Skewness	1.38
Kurtosis	5.35

Table 20: Summary statistics of *ageleft*

Statistic	<i>kilos</i>
Mean	95.06
Std. Dev	11.97
Min	60.33
Q1	86.18
Median	95.25
Q3	102.06
Max	163.00
Skewness	0.49
Kurtosis	0.31

Table 21: Summary statistics of *kilos*

Statistic	<i>cms</i>
Mean	198.29
Std. Dev	9.26
Min	160.02
Q1	190.50
Median	198.12
Q3	205.74
Max	231.14
Skewness	−0.08
Kurtosis	−0.33

Table 22: Summary statistics of *cms*

E. Alternate Kaplan-Meier survival curves (first dataset)

Here are the plots of Kaplan-Meier survival curves where time is measured using the number of seasons spent in the NBA rather than the number of games.

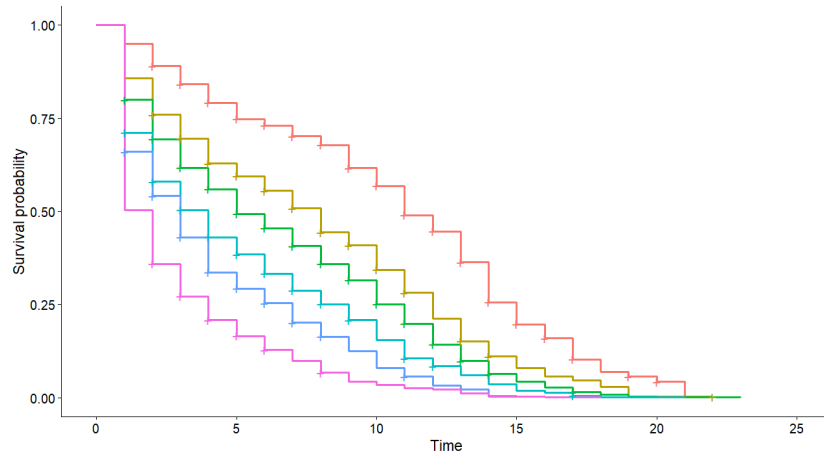


Figure 50: Alternate survival Curves per age for first NBA game (discretized)

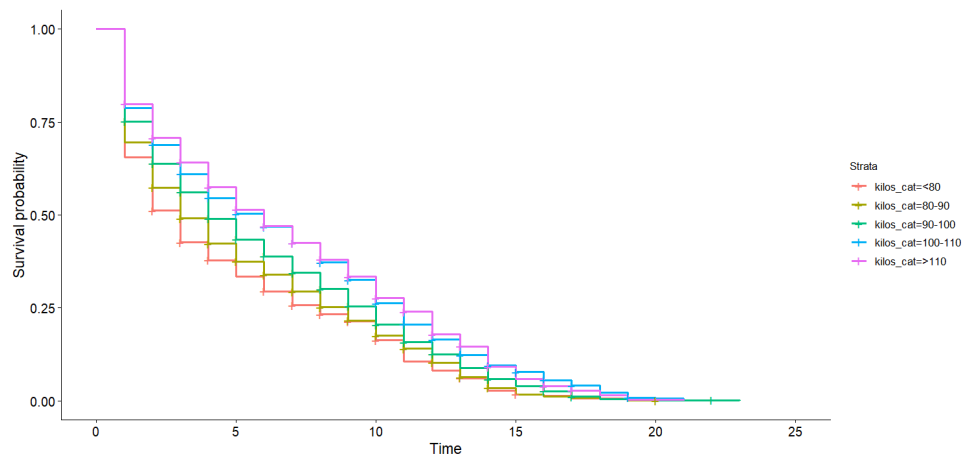


Figure 51: Alternate survival Curves per weight (discretized)

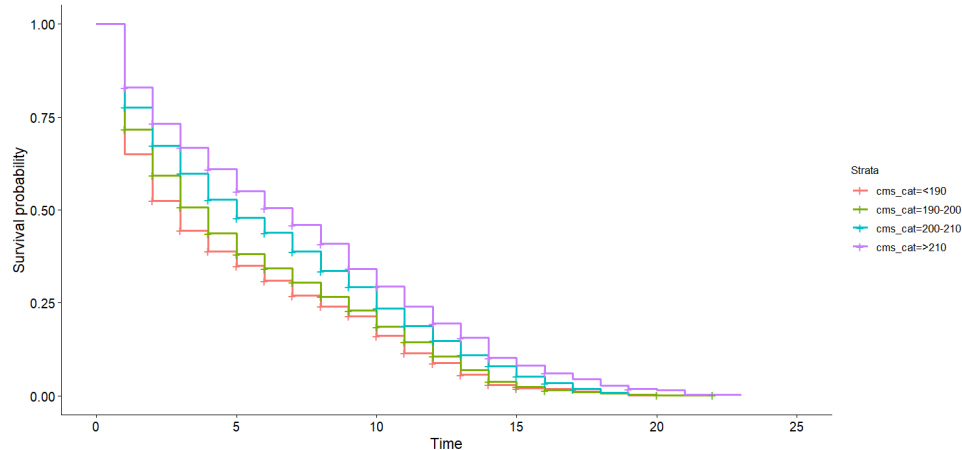


Figure 52: Alternate survival Curves per height (discretized)

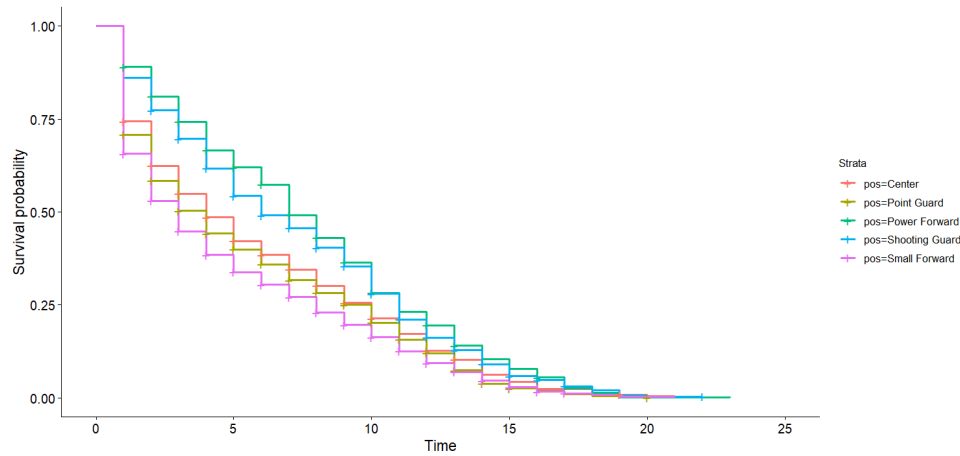


Figure 53: Alternate survival Curves per playing position

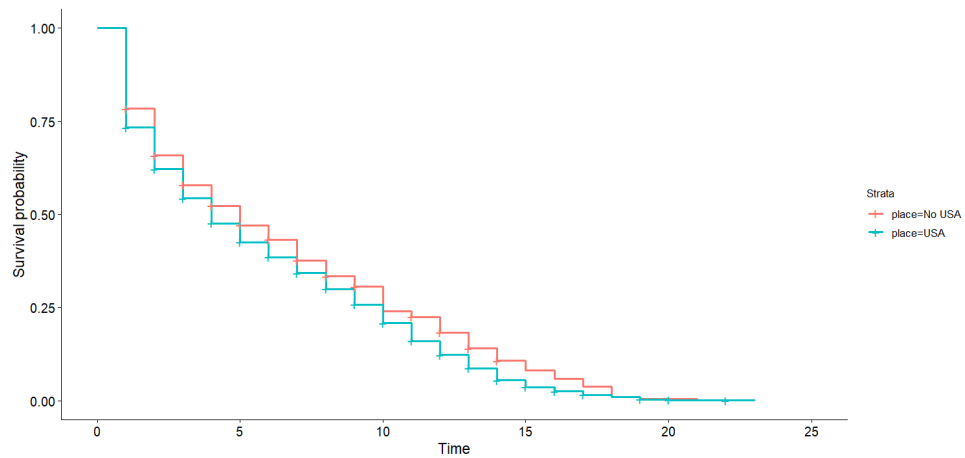


Figure 54: Alternate survival Curves per place of birth

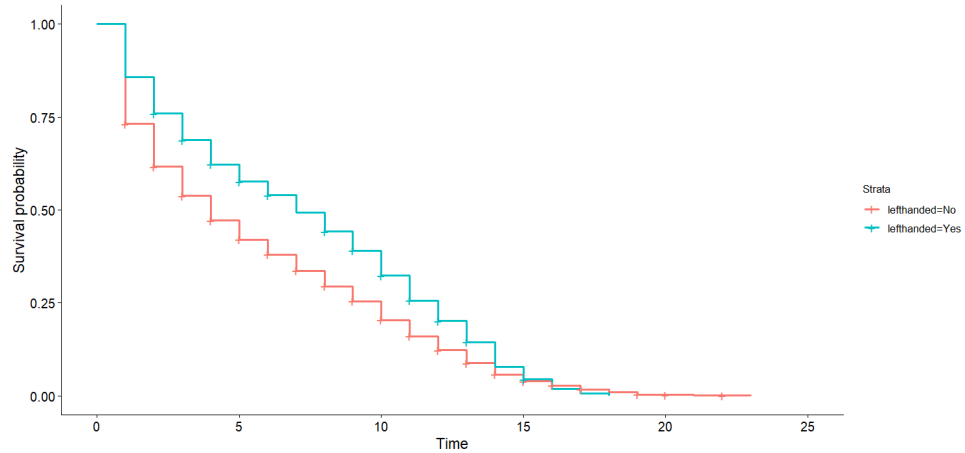


Figure 55: Alternate survival Curves per dominant hand

F. Alternate reference level comparison for covariate *pos* (first dataset)

Here is the table showing the values of $\exp(\hat{\beta})$ for every possible reference level in the model measuring time as the number of seasons spent in the NBA rather than the number of games.

Table 23: $\exp(\hat{\beta})$ for every reference level in *pos* (for time as number of seasons) ⁵

Ref.level →	<i>PG</i>	<i>SG</i>	<i>SF</i>	<i>PF</i>	<i>C</i>
<i>PG</i>		1.189*	0.558*	0.953	0.542*
<i>SG</i>	0.841*		0.469*	0.801*	0.455*
<i>SF</i>	1.793*	2.132*		1.708*	0.971
<i>PF</i>	1.049	1.248*	0.585*		0.569*
<i>C</i>	1.846*	2.196*	1.030	1.759*	

G. Stratified Cox models (first dataset)

Here are the regression coefficients for the Cox models stratified on covariate *pos*. Table 24 corresponds to the stratified model measuring time as the number of games, and Table 25 to the stratified model measuring time as the number of seasons.

Table 24: Stratified Cox model (for time as number of games)

	$\hat{\beta}$	$\exp(\hat{\beta})$	<i>lower.95</i>	<i>upper.95</i>	<i>p.val</i>
<i>ageleft</i>	0.172	1.188	1.171	1.205	$< 2.2 \times 10^{-16}$
<i>cms</i>	-0.020	0.980	0.974	0.986	6.953×10^{-11}
<i>lefthanded</i>	-0.230	0.795	0.694	0.910	8.738×10^{-4}

Table 25: Stratified Cox model (for time as number of seasons)

	$\hat{\beta}$	$\exp(\hat{\beta})$	<i>lower.95</i>	<i>upper.95</i>	<i>p.val</i>
<i>ageleft</i>	0.196	1.217	1.200	1.234	$< 2.2 \times 10^{-16}$
<i>kilos</i>	-0.010	0.991	0.986	0.995	1.717×10^{-4}
<i>cms</i>	-0.022	0.979	0.972	0.986	5.39×10^{-9}
<i>place</i>	0.113	1.120	0.997	1.258	0.056
<i>lefthanded</i>	-0.248	0.781	0.682	0.894	3.308×10^{-4}

H. Predictive performance of stratified Cox models (first dataset)

Here are the performance evaluations for the Cox models stratified on covariate *pos*. Table 26 displays the C-index and the IBS for both the stratified model measuring time

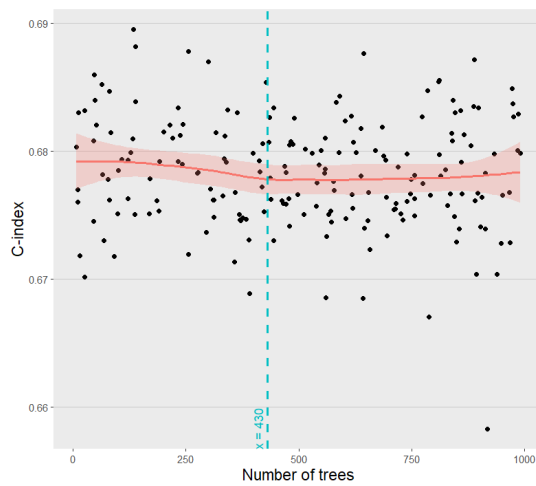
as the number of games, and the stratified model measuring time as the number of seasons.

Table 26: Performance of the stratified Cox models

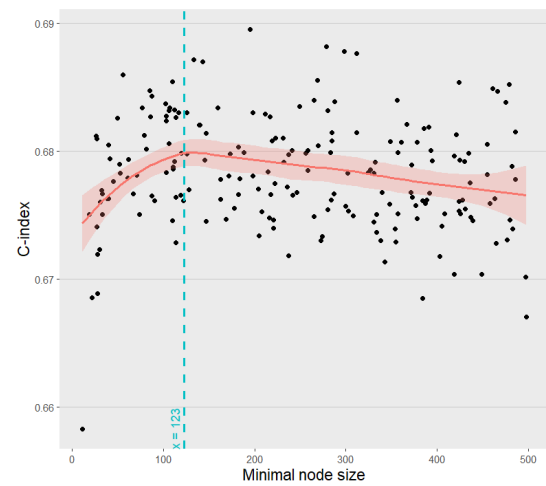
	2.5th pct	$C - index$	97.5th pct	2.5th pct	IBS	97.5th pct
Nbr of games	0.611	0.633	0.650	0.103	0.111	0.117
Nbr of seasons	0.645	0.668	0.687	0.108	0.115	0.124

I. Alternate optimization plots (first dataset)

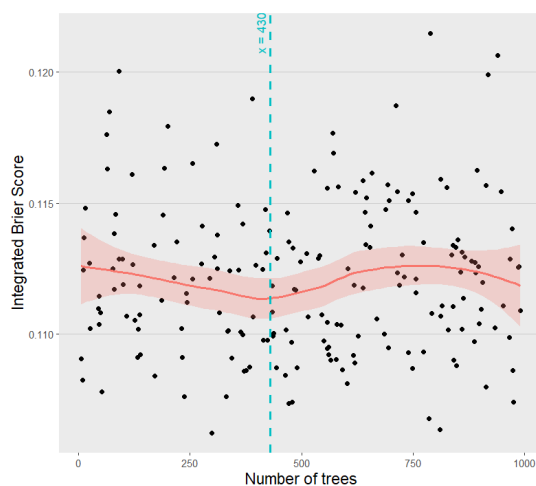
Here are the plots of hyperparameters optimization for models where time is measured using the number of seasons spent in the NBA rather than the number of games.



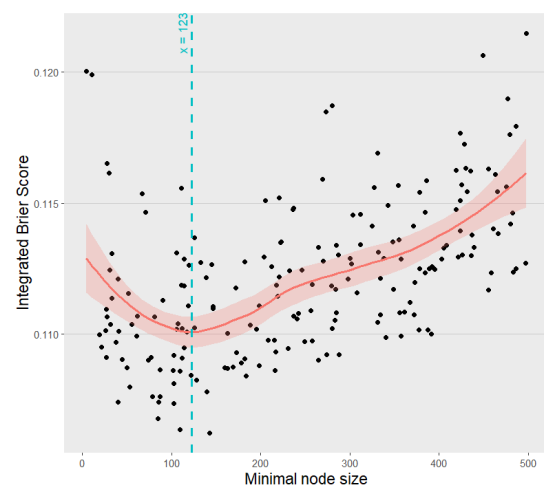
(a) C-index relative to number of trees



(b) C-index relative to minimum node size



(c) IBS relative to number of trees



(d) IBS relative to minimum node size

Figure 56: Alternate optimization of the hyperparameters of the RSF

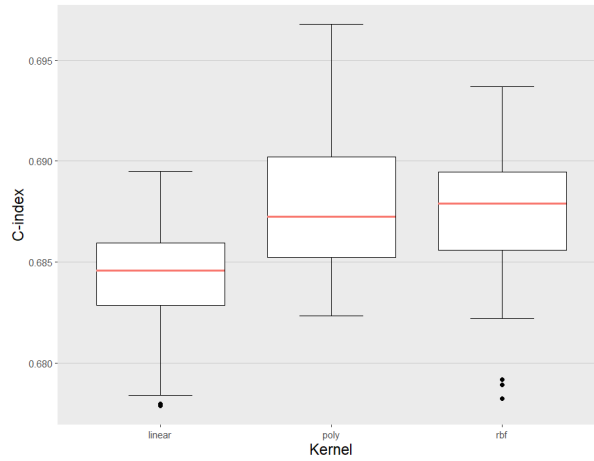
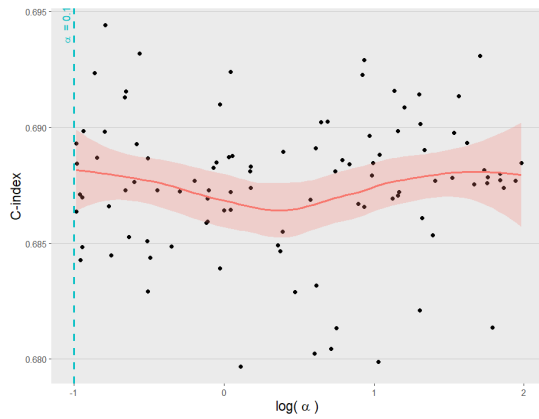
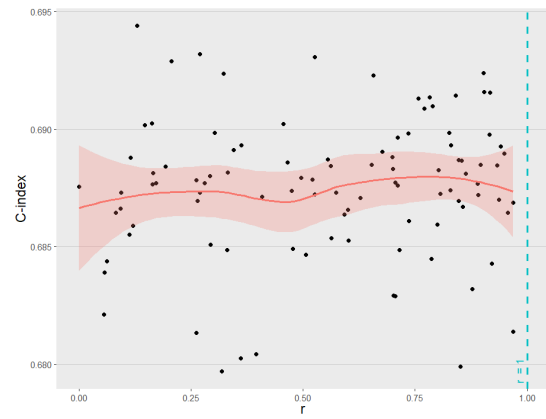


Figure 57: Alternate optimization of the kernel function



(a) C-index relative to regularization α



(b) C-index relative to rank ratio r

Figure 58: Alternate optimization of the hyperparameters of the SSVM with RBF kernel

J. Alternate Brier score plots

Here are the plots of the Brier scores through time for Cox regression and RSF where time is measured using the number of seasons spent in the NBA rather than the number of games.

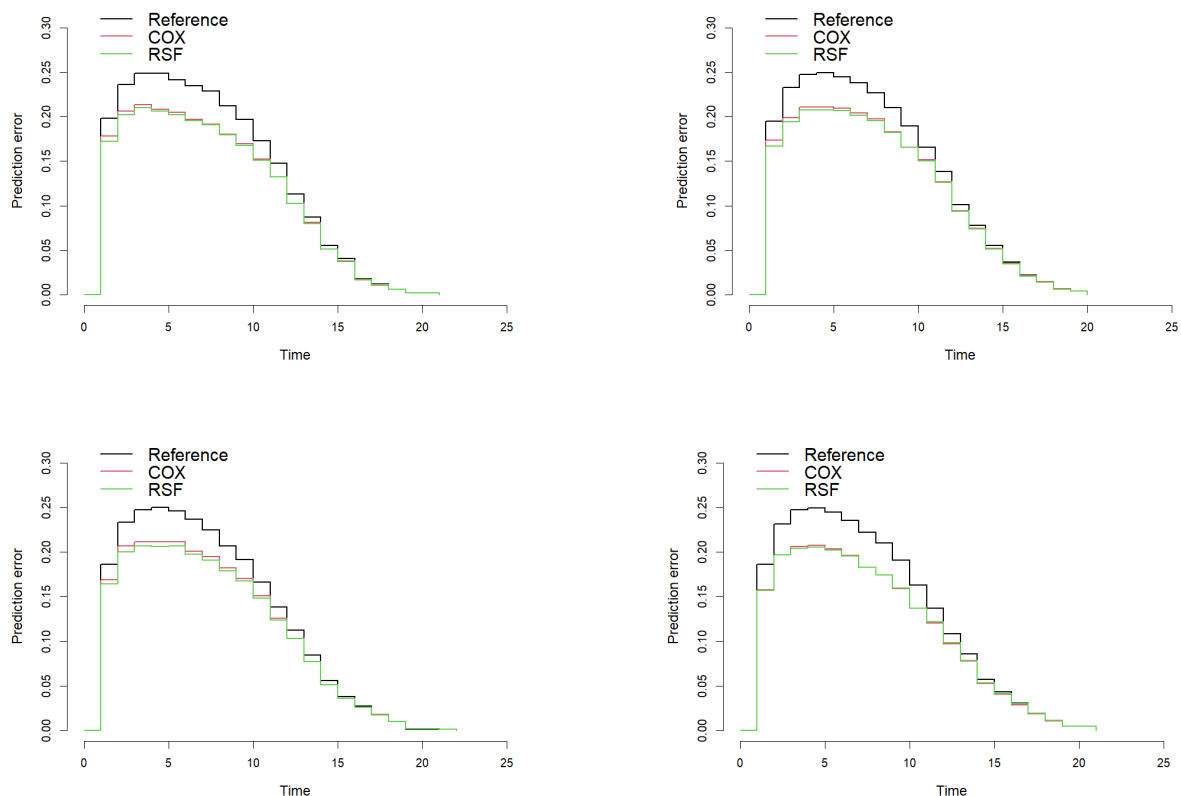


Figure 59: Alternate Brier scores through time for Cox regression and RSF

K. Players in the second dataset

Here is the list of the 185 players selected in the second dataset.

- Kareem Abdul-Jabbar
- LaMarcus Aldridge
- Ray Allen
- Carmelo Anthony
- G. Antetokounmpo
- Nate Archibald
- Paul Arizin
- Charles Barkley
- Rick Barry
- Elgin Baylor
- Bradley Beal
- Zelmo Beaty
- Walt Bellamy
- Chauncey Billups
- Dave Bing
- Larry Bird
- Carlos Boozer
- Chris Bosh
- Jaylen Brown
- Kobe Bryant
- Jimmy Butler
- Vince Carter
- Wilt Chamberlain
- Tom Chambers
- Tyson Chandler
- Maurice Cheeks
- Mike Conley
- Michael Cooper
- DeMarcus Cousins
- Bob Cousy
- Dave Cowens
- Terry Cummings
- Billy Cunningham

- Stephen Curry
- Bob Dandridge
- Adrian Dantley
- Brad Daugherty
- Anthony Davis
- Baron Davis
- Walter Davis
- Dave DeBusschere
- DeMar DeRozan
- Luka Doncic
- Clyde Drexler
- Joe Dumars
- Tim Duncan
- Kevin Durant
- Joel Embiid
- Alex English
- Julius Erving
- Patrick Ewing
- Larry Foust
- Walt Frazier
- Harry Gallatin
- Kevin Garnett
- Marc Gasol
- Pau Gasol
- Paul George
- George Gervin
- Artis Gilmore
- Manu Ginóbili
- Rudy Gobert
- Danny Green
- Draymond Green
- Hal Greer
- Blake Griffin
- Richie Guerin
- James Harden
- Anfernee Hardaway
- Tim Hardaway
- John Havlicek
- Spencer Haywood
- Elvin Hayes
- Tom Heinsohn
- Grant Hill
- Al Horford
- Dwight Howard
- Bailey Howell
- Lou Hudson
- Serge Ibaka
- Andre Iguodala
- Kyrie Irving
- Dan Issel
- Allen Iverson
- LeBron James
- Dennis Johnson
- Gus Johnson
- Kevin Johnson
- Magic Johnson
- Marques Johnson
- Nikola Jokic
- Sam Jones
- DeAndre Jordan
- Michael Jordan
- Shawn Kemp
- Red Kerr
- Jason Kidd
- Bernard King
- Bob Lanier
- Rudy LaRusso
- Kawhi Leonard
- Damian Lillard
- Bob Love
- Kevin Love
- Kyle Lowry
- Jerry Lucas
- Maurice Lucas
- Ed Macauley
- Karl Malone
- Moses Malone
- Pete Maravich
- Bob McAdoo
- George McGinnis
- Tracy McGrady

- Dick McGuire
- Kevin McHale
- Vern Mikkelsen
- Reggie Miller
- Paul Millsap
- Yao Ming
- Sidney Moncrief
- Earl Monroe
- Alonzo Mourning
- Chris Mullin
- Dikembe Mutombo
- Steve Nash
- Willie Naulls
- Joakim Noah
- Dirk Nowitzki
- Shaquille O'Neal
- Hakeem Olajuwon
- Robert Parish
- Tony Parker
- Chris Paul
- Gary Payton
- Bob Pettit
- Paul Pierce
- Scottie Pippen
- Kevin Porter
- Willis Reed
- Mitch Richmond
- Oscar Robertson
- David Robinson
- Guy Rodgers
- Dennis Rodman
- Rajon Rondo
- Derrick Rose
- Bill Russell
- Dolph Schayes
- Charlie Scott
- Bill Sharman
- Pascal Siakam
- Jack Sikma
- Paul Silas
- Phil Smith
- John Stockton
- Amar'e Stoudmire
- Jayson Tatum
- Isaiah Thomas
- David Thompson
- Klay Thompson
- Nate Thurmond
- Rudy Tomjanovich
- Jack Twyman
- Wes Unseld
- Norm Van Lier
- Dwyane Wade
- Chet Walker
- Kemba Walker
- John Wall
- Bill Walton
- Chris Webber
- Jerry West
- Russell Westbrook
- Paul Westphal
- Jo Jo White
- Lenny Wilkens
- Jamaal Wilkes
- Dominique Wilkins
- Buck Williams
- Deron Williams
- Gus Williams
- Metta World Peace
- James Worthy
- George Yardley

L. Summary statistics (second dataset)

Here are the summary statistics for the time variable and the quantitative covariates of the second dataset.

Statistic	Games
Mean	1069.68
Std. Dev	279.09
Min	505.00
Q1	880.00
Median	1036.00
Q3	1250.00
Max	1854.00
Skewness	0.43
Kurtosis	−0.18

Table 27: Summary statistics for the total number of games

Statistic	<i>Draft_pos</i>
Mean	12.45
Std. Dev	18.18
Min	1.00
Q1	3.00
Median	6.00
Q3	14.00
Max	122.00
Skewness	3.43
Kurtosis	15.03

Table 28: Summary statistics of *Draft_pos*

Statistic	<i>Ind_Success</i>
Mean	7.32
Std. Dev	7.06
Min	0.00
Q1	2.00
Median	5.00
Q3	10.00
Max	35.00
Skewness	1.72
Kurtosis	3.23

Table 29: Summary statistics of *Ind_Success*

Statistic	<i>Titles</i>
Mean	1.39
Std. Dev	1.91
Min	0.00
Q1	0.00
Median	1.00
Q3	2.00
Max	11.00
Skewness	2.15
Kurtosis	5.68

Table 30: Summary statistics of *Titles*

Statistic	<i>cms</i>
Mean	200.21
Std. Dev	9.20
Min	183.00
Q1	193.00
Median	201.00
Q3	206.00
Max	229.00
Skewness	−0.04
Kurtosis	−0.44

Table 31: Summary statistics of *Height*

Statistic	<i>kgs</i>
Mean	97.42
Std. Dev	12.70
Min	68.00
Q1	88.00
Median	97.00
Q3	106.00
Max	147.00
Skewness	0.53
Kurtosis	0.78

Table 32: Summary statistics of *Weight*

Statistic	<i>Birth_year</i>
Mean	1962.47
Std. Dev	20.54
Min	1926.00
Q1	1946.00
Median	1961.00
Q3	1982.00
Max	1999.00
Skewness	−0.01
Kurtosis	−1.26

Table 33: Summary statistics of *Birth_year*

Statistic	<i>Avg_min_pgm</i>
Mean	33.81
Std. Dev	3.47
Min	25.10
Q1	31.85
Median	34.20
Q3	36.23
Max	46.20
Skewness	−0.09
Kurtosis	0.52

Table 34: Summary statistics of *Avg_min_pgm*

References

- Aalen, O. O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics*, 6(4):701–726.
- Agresta, M. C., Brandão, M. R. F., and De Barros Neto, T. L. (2008). Causas e consequências físicas e emocionais do término de carreira esportiva. *Revista Brasileira De Medicina Do Esporte*, 14(6):504–508.
- Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In Kotz, S. and Johnson, N. L., editors, *Breakthroughs in Statistics*, pages 199–213. Springer, New York, NY.
- Andreff, W. (2019). Origins and developments of sports systems. In *The SAGE Handbook of Sports Economics*, pages 8–18. SAGE.
- Antolini, L., Boracchi, P., and Biganzoli, E. (2005). A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24(24):3927–3944.
- Baker, J., Koz, D., Kungl, A., Fraser-Thomas, J., and Schorer, J. (2012). Staying at the top: Playing position and performance affect career length in professional sport. *High Ability Studies*, 24(1):63–76.
- Basketball-Reference.com (n.d.a). 2016–17 houston rockets roster and stats. <https://www.basketball-reference.com/teams/HOU/2017.html>. Accessed: 2026-07-19.
- Basketball-Reference.com (n.d.b). Michael jordan stats, height, weight, position, draft status and more. <https://www.basketball-reference.com/players/j/jordami01.html>. Accessed: 2026-08-23.
- Basketball-Reference.com (n.d.c). Nba draft index: Latest and historical nba drafts info. <https://www.basketball-reference.com/draft/>. Accessed: 2026-08-21.
- Bounnite, M. Y. (2025). Modeling career mobility and attrition using markov chains and survival analysis. *Quality & Quantity*, 60(1):2479–2507.
- Branders, S., Frénay, B., and Dupont, P. (2015). Survival analysis with cox regression and random non-linear projections. In *Proceedings of the 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2015)*, pages 361–366.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.
- Breslow, N. E. (1972). Discussion of professor cox’s paper. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):216–217.

- Brettschneider, W. and Brandl-Bredenbeck, H. P. (2007). Youth and sport in the usa and germany: a cross-cultural comparative study. *Imprensa da Universidade de Coimbra eBooks*, pages 35–56.
- Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.
- Cohen, J. (2013). *Statistical Power Analysis for the Behavioral Sciences*. Routledge, 2nd edition.
- Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society*, 34(2):187–202.
- Cramér, H. (1946). *Mathematical Methods of Statistics*, volume 9 of *Princeton Mathematical Series*. Princeton University Press.
- De Vera, M. (n.d.). Newton–raphson algorithm notes. https://micoledevera.github.io/other_docs/NRAAlgorithm_note.pdf. Accessed: 2026-08-09.
- Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., and Vapnik, V. (1996). Support vector regression machines. In *Advances in Neural Information Processing Systems*, volume 9, pages 155–161.
- Efron, B. (1977). The efficiency of cox’s likelihood function for censored data. *Journal of the American Statistical Association*, 72(359):557–565.
- Erčulj, F. and Bračič, M. (2010). Differences between various types of elite young female basketball players in terms of their morphological characteristics. *Kinesiologia Slovenica*, 16(1):51—60.
- Evans, B. and Pitts, J. D. (2022). The determinants of draft position for nba prospects. *NEW YORK ECONOMIC REVIEW*, 52:22–37.
- Feroli, D., Rampinini, E., Bosio, A., La Torre, A., Azzolini, M., and Coutts, A. J. (2018). The physical profile of adult male basketball players: Differences between competitive levels and playing positions. *Journal of Sports Sciences*, 36(22):2567–2574.
- Fleming, T. R. and Lin, D. Y. (2000). Survival analysis in clinical trials: Past developments and future directions. *Biometrics*, 56(4):971–983.
- Fouodo, C. J. K., König, I. R., Weihs, C., Ziegler, A., and Wright, M. N. (2018). Support vector machines for survival analysis with R. *The R Journal*, 10(1):297–310.
- Fynn, K. D. and Sonnenschein, M. (2012). An analysis of the career length of professional basketball players. <https://digitalcommons.mcalester.edu/macreview/vol2/iss2/3>.

- Gerds, T. A. and Schumacher, M. (2006). Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6):1029–1040.
- Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17–18):2529–2545.
- Greer, T., Price, J. A., and Berri, D. J. (2019). Jumping in the pool: What determines which players the nba considers in the draft? *International Journal of Sport Finance*, 14(1):43–53.
- Groothuis, P. A. and Hill, J. R. (2004). Exit discrimination in the nba: A duration analysis of career length. *Economic Inquiry*, 42(2):341—349.
- Hafner, C. M. (2026). Statistical machine learning and high dimensional data analysis. Unpublished lecture slides, UCLouvain.
- Hardyck, C., Petrinovich, L. F., and Goldman, R. D. (1976). Left-handedness and cognitive deficit. *Cortex*, 12(3):266–279.
- Harrell, F. E., Lee, K. L., Califf, R. M., Pryor, D. B., and Rosati, R. A. (1984). Regression modelling strategies for improved prognostic prediction. *Statistics in Medicine*, 3(2):143–152.
- Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860.
- Johnston, K., Farah, L., Ghuman, H., and Baker, J. (2021). To draft or not to draft? a systematic review of north american sports’ entry draft. *Scandinavian Journal of Medicine and Science in Sports*, 32(1):4–17.
- Jones, M., Motesharei, A., Ball, S. V., Church, J. S., Calder, J. D. F., and Williams, A. (2025). Establishing “normal” career longevity in professional footballers allows comparison to that of players with injuries and surgery. *Knee Surgery, Sports Traumatology, Arthroscopy*, 33(9):3375–3388.
- Jozsa, F. P. (2014). National basketball association strategies. *SpringerBriefs in economics*.
- Jr. NBA (2016). Basketball positions. <https://jr.nba.com/basketball-positions/>. Published: December 19, 2016. Accessed: 2026-08-21.
- Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

- Klein, J. and Moeschberger, M. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, second edition edition.
- Kvamme, H., Borgan, Ø., and Scheel, I. (2019). Time-to-event prediction with neural networks and cox regression. *Journal of Machine Learning Research*, 20(129):1–30.
- Lavallee, D., Grove, J. R., and Gordon, S. (1997). The causes of career termination from sport and their relationship to post-retirement adjustment among elite-amateur athletes in australia. *Australian Psychologist*, 32(2):131–135.
- Lawler, T. P. and Lawler, F. H. (2011). Left-handedness in professional basketball: Prevalence, performance, and survival. *Perceptual and Motor Skills*, 113(3):815–824.
- Macran, S., Joshi, H., and Dex, S. (1996). Employment after childbearing: A survival analysis. *Work, Employment and Society*, 10(2):273–296.
- Martin, C. L., Arundale, A. J. H., Kluzek, S., Ferguson, T., Collins, G. S., and Bullock, G. S. (2021). Characterization of rookie season injury and illness and career longevity among national basketball association players. *JAMA Network Open*, 4(10).
- Martínez, J. A., Langohr, K., Felipo, J., Consuegra, L., and Casals, M. (2022). Data set on mortality of national basketball association (nba) players. *Data in Brief*, 45.
- Miguel, C. G., Milan, F. J., Soares, A. L., Quinaud, R. T., Kós, L. D., Palheta, C. E., Mendes, F. G., and Carvalho, H. M. (2019). Modelling the relationship between nba draft and the career longevity of players using generalized additive models. <http://ddd.uab.cat/record/219113>.
- National Basketball Association (2024). Our history. <https://careers.nba.com/history/>. Published: February 8, 2024. Accessed: 2026-07-19.
- Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14(4):945–966.
- Papadatou-Pastou, M., Ntolka, E., Schmitz, J., Martin, M., Munafò, M. R., Ocklenburg, S., and Paracchini, S. (2020). Human handedness: A meta-analysis. *Psychological Bulletin*, 146(6):481–524.
- Petersen, A. M., Jung, W. S., Yang, J., and Stanley, H. E. (2010). Quantitative and empirical demonstration of the matthew effect in a study of career longevity. *Proceedings of the National Academy of Sciences*, 108(1):18–23.
- Pot, N. and Van Hilvoorde, I. (2013). Generalizing the effects of school sports: comparing the cultural contexts of school sports in the netherlands and the usa. *Sport in Society*, 16(9):1164–1175.

- Radečić, D. (2026). Support vector regression (svr): How it works and when to use it. <https://www.datacamp.com/tutorial/support-vector-regression-svr>. DataCamp. Accessed: 2026-07-19.
- Ristolainen, L., Kettunen, J. A., Kujala, U. M., and Heinonen, A. (2011). Sport injuries as the main cause of sport career termination among finnish top-level athletes. *European Journal of Sport Science*, 12(3):274–282.
- Scharoun, S. M. and Bryden, P. J. (2014). Hand preference, performance abilities, and hand selection in children. *Frontiers in Psychology*, 5:82.
- Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, 69(1):239–241.
- scikit-survival developers (n.d.). Introduction to survival support vector machine. https://scikit-survival.readthedocs.io/en/stable/user_guide/survival-svm.html. scikit-survival 0.28.0. Accessed: 2026-07-19.
- Shivaswamy, P. K., Chu, W., and Jansche, M. (2007). A support vector approach to censored targets. In *Proceedings of the Seventh IEEE International Conference on Data Mining (ICDM)*, pages 655–660. IEEE.
- Sports Reference (2011). Slam 500 greatest nba players of all time — basketball-reference.com. https://www.basketball-reference.com/awards/slam_500_greatest.html. Accessed: 2026-02-16.
- Sports Reference (2025). Basketball reference. <https://www.basketball-reference.com/>. Accessed: 2025-10-15.
- Steingröver, C., Wattie, N., Baker, J., and Schorer, J. (2016). Does relative age affect career length in north american professional sports? *Sports Medicine - Open*, 2(1):18.
- Therneau, T. M. and Watson, D. A. (2017). The concordance statistic and the cox model. Technical Report 85, Mayo Clinic. Accessed: 2026-08-23.
- Under Armour (n.d.). What are the 5 basketball positions and their roles? <https://www.underarmour.com/en-us/t/playbooks/basketball/basketball-positions/>. Accessed: 2026-08-21.
- Van Belle, V., Pelckmans, K., Suykens, J. A. K., and Van Huffel, S. (2008). Survival SVM: A practical scalable algorithm. In *Proceedings of the 16th European Symposium on Artificial Neural Networks (ESANN 2008)*, pages 89–94.
- Van Belle, V., Pelckmans, K., Van Huffel, S., and Suykens, J. A. K. (2011). Support vector methods for survival analysis: A comparison between ranking and regression approaches. *Artificial Intelligence in Medicine*, 53(2):107–118.

- Van Keilegom, I. (2025). Analysis of survival and duration data (2025-2026). Power-Point, UCLouvain.
- Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer.
- Vretaros, A. (2021). Longevidade de carreira no basquetebol. *Revista Internacional de Preparacion Fisica y Ciencias Aplicadas en Basquetbol*, 2.
- Vretaros, A. (2022). Comparing the career longevity of basketball players across three continents: A preliminary exploratory study. *Adv Health Exerc*, 2(1):1–7.
- Yan, J. H. and McCullagh, P. (2004). Cultural influence on youth's motivation of participation in physical activity. *Journal of Sport Behavior*, 27:378–390.
- Zaman, Q., Khan, M. B., Sarwar, N., Irshad, M., Anwar, M., and Iqbal, S. (2025). Survival analysis of career longevity in t-20 cricket. *Competitive Research Journal Archive*, 3(3):201–209.

