A Note on Kendall's Tau Coefficient for Gap Times in Presence of Right Censoring *

Cecilia Castro[0000−0001−9897−8186] and Ana Paula Amorim[0000−0003−3957−1129]

University of Minho, Centre of Mathematics, Braga, Portugal
{cecilia,apamorim}@math.uminho.pt

Abstract. In several clinical and epidemiology studies, data from events that occur successively in time in the same individual, are frequently reported. Among these, the most common are recurrent events where each subject may experience a number of failures over the course of follow-up. Examples include repeated hospitalization of patients, recurrences of tumor, recurrent infections, among others. In this work, the interest is to study the correlation between successive recurrent events, gap times, in the presence of right censoring. To measure the association between two gap times we use the Kendall's τ correlation coefficient, by incorporating suitable bivariate estimators of the joint distribution function of the gap times and of the marginal distribution function of the second gap time, into the integrals that define the probability of concordant pairs and the probability of discordant pairs. Two of the estimators of the joint distribution function of the gap times considered in this work are already known, but we consider also estimators with Kaplan-Meier weights defined by using decision trees and random forests methodology. We conclude that all the estimators perform better in a scenario of negative association. When the association is moderately negative, the performance of the estimator with smoothed weights using random forests is superior. In the case of strong positive association, the best estimator is the presmoothed nonparametric but, in the case of moderate positive association, this estimator has identical performance as the estimator with presmoothed weights using random forests.

Keywords: Right Censoring · Decision Trees · Kendall's Tau · Random Forests.

1 Introduction

Recurrent events that occur in the same subject successively in time, are frequently reported in clinical studies. Examples include repeated hospitalizations of patients, tumor recurrences, recurrent infections, among others.
In this work, the interest is to study the correlation between gap times, that is, times between two successive recurrent events.

* Supported by Portuguese funds through Centre of Mathematics, University of Minho.

Correlations between gap times are of interest in themselves, when investigating whether the first gap time is predictive of the occurrence of the second event. To measure the possible association between two gap times, it is usual to use the nonparametric estimator of the correlation coefficient Kendall's τ, because it has good properties, like invariance to monotone transformations and robustness in the presence of outliers [12].
Measuring correlation can be challenging in the presence of right censoring where some data values are not observed due to an upper detection limit, dropout or due to the end of the study.
Right censoring is present in a wide range of survival data, so it is natural that one or both of the gap times may not be observed.
Several different methods have been proposed to measure and test the correlation between two right-censored time-to-event variables [5, 8, 10, 11].
In this paper we define estimators of τ that accounts for joint information of the random pair of gap times in the presence of right censoring. In fact, a natural way to estimate τ is to incorporate a suitable bivariate estimator of the joint distribution function into the integrals that define the probability of concordant pairs and the probability of discordant pairs in the context of Kendall's tau definition. This is not the usual approach in the papers that have been published on this topic. In general, the authors use a compact formula for the estimator, considering the complementarity of concordant and discordant events.

When looking to a single subject, the censoring time distribution may be the same for time to the first failure and for time to the second failure. However, the censoring time for the second gap time depends on the time to the first failure and on the censoring time for the total time.

Two of the estimators of the joint distribution function of the gap times used in this work are already known (see [1, 2]), but we propose another estimators for the Kaplan-Meier weights, defined with the same reasoning used in the definition of the known semiparametric estimator of the bivariate distribution function (see [2]), but using decision trees and random forests to define the weights used in the Kaplan-Meier estimator for the joint distribution function of the gap times. The study presented is supported by simulations.

This paper is divided into 6 sections. In the first section we present an introduction to the topic, a brief bibliographical review and we establish the notation. In the second section we present estimators of the joint distribution function of gap times and justifications for using decision trees and random forests methodologies to define the Kaplan-Meier weights. In the third section we define Kendall's tau estimators based on the probability of concordance and probability of discordance of pairs of gap times, and we justify this approach based on theoretical results. This section also presents a detailed description of the numerical procedure for obtaining estimates of the probabilities of concordance and probabilities of discordance. The fourth section is dedicated to the simulation procedure and the main results. In section 5 is presented an application example of the proposed methodology with real data. Finally, in the last section, are the main conclusions of the work.

1.1 Notations and Definitions

Let T_1 be the time from the begining of the study to the first occurrence of the event of interest or *failure* (first gap time) and T_2 be the time between the first *failure* and the second *failure* (second gap time). The random times T_1 and T_2 are possibly correlated. Let (T_{1i}, T_{2i}) and (T_{1j}, T_{2j}), $i \neq j$, be independent realizations from (T_1, T_2).

Definition 1. *The pair (i, j) is said to be concordant if $(T_{1i}-T_{1j})(T_{2i}-T_{2j}) > 0$ and discordant if $(T_{1i} - T_{1j})(T_{2i} - T_{2j}) < 0$. If T_1 and T_2 are continuous, the Kendall's correlation between T_1 and T_2 is given by*

$$\tau = P\left((T_{1i} - T_{1j})(T_{2i} - T_{2j}) > 0\right) - P\left((T_{1i} - T_{1j})(T_{2i} - T_{2j}) < 0\right) \quad (1)$$

The correlation coefficient, τ, is such that $-1 \leq \tau \leq 1$ and $\tau = 0$ if (T_1, T_2) are independent.

Denoting marginal and joint cumulative distribution functions of T_1 and T_2 as $F_1(x) = P(T_1 \leq x)$, $F_2(y) = P(T_2 \leq y)$ and $F_{12}(x, y) = P(T_1 \leq x, T_2 \leq y)$, respectively, and defining $F_.(x^-) = \lim_{t \uparrow x} F_.(t)$, we have

$$p_c = P\left((T_{1i} - T_{1j})(T_{2i} - T_{2j}) > 0\right) = 2 \int_0^{+\infty} \int_0^{+\infty} F_{12}(x^-, y^-) F_{12}(dx, dy) \quad (2)$$

and

$$p_d = P\left((T_{1i} - T_{1j})(T_{2i} - T_{2j}) < 0\right) = 2 \int_0^{+\infty} \int_0^{+\infty} U(x^-, y^-) F_{12}(dx, dy) \quad (3)$$

where

$$U(x, y) = P(T_1 > x, T_2 < y) = F_{12}(\infty, y^-) - F_{12}(x, y^-)$$

Now the tau Kendall's coefficient is given by:

$$\tau = p_c - p_d \quad (4)$$

2 Estimators of the Bivariate Distribution Function for Censored Gap Times

Let C be the right censoring time. This censoring time is the minimum between the time from the start of study to the end of the study, and the time from the start of study to dropout. So, the support of C is bounded.
We made the standard assumption that the first gap time, T_1, and the total time, $Y = T_1 + T_2$, are subject to independent right censoring. As T_1 and Y are observed in a single subject, the distribution function of the censoring time C, say $G(.)$, may be the same for both T_1 and Y. So, the marginal distribution of the first gap time F_1, can be consistently estimated by the Kaplan and Meier estimator, based on the observable pair $(\widetilde{T}_1, \delta_1)$ where $\widetilde{T}_1 = \min\{T_1, C\}$ and the distribution of the total time, Y, say F, can also be estimated by the Kaplan and

Meier estimator based on $(\widetilde{Y}, \delta_2)$ where $\widetilde{Y} = \min\{Y, C\}$ [9]. The indicator variables $\delta_j, j = 1,2$ are defined by $\delta_j = 1, \quad \text{if} \quad T_i \leq C,$ and equal to $0,$ otherwise. However, the second gap time, T_2, and the censoring time, $C_2 = (C - T_1)\delta_1$, are in general dependent. Let $\widetilde{T}_2 = \min\{T_2, C_2\}$ and the marginal distributions of $\widetilde{T}_1$ and $\widetilde{T}_2$ are $H_1(x) = P(\widetilde{T}_1 \leq x)$ and $H_2(y) = P(\widetilde{T}_2 \leq y)$ and the joint distribution of $\left(\widetilde{T}_1, \widetilde{T}_2\right)$ is $H(x,y) = P(\widetilde{T}_1 \leq x, \widetilde{T}_2 \leq y)$.
The estimators for the bivariate distribution function of gap times (T_1, T_2), $F_{12}(x,y)$, are weighted Kaplan-Meier estimators with the same weights used in the definition of the estimator of the total time distribution function $\widehat{F}(y)$ (see [1,2,8]), based on the ranks of $\widetilde{Y}_i$, $R_i = Rank(\widetilde{Y}_i)$, where, in the case of ties, the ranks of the censored observations $\widetilde{Y}_i$'s are higher than the ranks of the uncensored observations.

$$\widehat{F}_{12}(x,y) = \sum_{i=1}^{n} W_i \mathrm{I}\left(\widetilde{T}_{1i} \leq x, \widetilde{T}_{2i} \leq y\right) \tag{5}$$

with $\mathrm{I}(A)$ the usual indicator function of the event A.
The second gap time distribution function estimator is easily obtained from equation (5). In fact, we have

$$\widehat{F}_2(y) = \widehat{F}_{12}(\infty, y) = \sum_{i=1}^{n} W_i \mathrm{I}\left(\widetilde{T}_{2i} \leq y\right) \tag{6}$$

The weights W_i in equation (5) presented in the expressions (7) and (9) are already known and the corresponding estimators have already been studied (see [1,2]).

$$W_i = \frac{\delta_{2i}}{n - R_i + 1} \prod_{j=1}^{i-1} \left(1 - \frac{\delta_{2j}}{n - R_j + 1}\right) \tag{7}$$

With the weights defined in (7), the estimator (5) only assigns positive mass to pairs of gap times with both components uncensored.

In order to assign positive mass to pairs of gap times in which only the second gap time T_2 is censored, while the weight assigned to pairs with the first gap time censored remains zero, a binary classification model $m(x,y)$ can be used, which, based on the observed values of the first gap time and the total time, assigns a non-zero probability to the event $\delta_2 = 1$.

$$m(x,y) = P\left(\delta_2 = 1 | \widetilde{T}_1 = x, \widetilde{Y} = y\right), \; x \leq y \tag{8}$$

$$W_i = W_i(m) = \frac{m(\widetilde{T}_{1i}, \widetilde{Y}_i)}{n - R_i + 1} \prod_{j=1}^{i-1} \left(1 - \frac{m(\widetilde{T}_{1i}, \widetilde{Y}_i)}{n - R_j + 1}\right) \tag{9}$$

When the model m is parametric, like the logistic model, we must estimate the model parameters, typically computed by maximizing the conditional likelihood of the δ_2's given $(\widetilde{T}_1, \widetilde{T}_2)$ for those cases with $\delta_1 = 1$ (see [6,7]).

An alternative way for the definition of the Kaplan-Meier weigths, is to consider the probability $m(x, y)$ in equation(8) given by decision trees or random forests methodologies. The incorporation of smoothed Kaplan-Meier weigths in the estimation of the bivariate function aims to reduce the bias imposed by right censoring. When estimating the probabilities of the second gap time observations being censored, knowing the values of the first gap time and the total time, the objective is not to explain but to predict. So it might make sense to use decisions trees or random forests to get these probabilities. In fact, in general terms, if the focus is mainly on explanation, logistic regression tends to perform better than random forests, but this in not completely true if the focus is on prediction rather than explanation [4]. On the other hand, logistic regression requires there to be little or no multicollinearity among the independent variables. This means that the independent variables should not be too highly correlated with each other which, in the case under analysis, is not verified since the first gap time x and the total time y can be strongly associated.

3 Estimators of the Kendall's Tau Coefficient for Censored Gap Times

With the definition and notations of subsection 1.1, to estimate the correlation τ between two gap times, T_1 and T_2, we use the definition for Kendall's tau coefficient as the difference between the concordance probability, p_c, and discordance probability of T_1 and T_2, p_d, given by expressions (2) and (3), respectively. These probabilities depend only on the joint distribution function of the interval times, F_{12}, since the marginal distribution of the second interval time, T_2, can be obtained from the joint distribution function F_{12}. Under right censoring, the estimator of p_c, $\widehat{p}_c$, obtained from the distribution function estimator $\widehat{F}_{12}$, only converges to p_c in a restricted domain, and the same goes for the estimator $\widehat{p}_d$ of p_d. In general we have $\widehat{p}_c + \widehat{p}_d \leq 1$, therefore we will calculate these estimates separately. In fact, denoting by τ_H the upper bound of the support of the distribution function of $\widetilde{Y}$, say H_y, variable assumed to be continuous, and defining

$$F_{12}^0 = P(T_1 \leq x, T_2 \leq y, T_1 + T_2 \leq \tau_H) \tag{10}$$

it was proved that the estimators of F_{12}, defined on section 2, converges to F_{12}^0, as $n \to \infty$, and not to F_{12} (see [2] for detailed explanation). The same situation occurs for the estimator of the marginal distribution function of T_2, $\widehat{F}_2$, which is given by

$$\widehat{F}_2(y) = \widehat{F}_{1,2}(\infty, y) = \sum_{i=1}^{n} W_i(m) I(\widetilde{T}_{2i} \leq y) \tag{11}$$

In fact,

$$\lim_{n\to\infty} \widehat{F}_2(y) = P(T_2 \leq y, T_1 + T_2 \leq \tau_H) \equiv F_2^0(y) \neq F_2(y) \tag{12}$$

3.1 Procedure for obtaining Kendall's Tau Estimates

In this subsection we present the numerical procedure for obtaining estimates of Kendall's tau coefficient. A data matrix with 4 columns and n rows is given, $M \equiv M[i,j]$, $i = 1, \dots, n$; $j = 1, \dots, 4$. Each line i corresponds to one case.

- $M[,1]$ – time until the first event occurs;
- $M[,2]$ – total time until the second event occurs;
- $M[,3]$ – boolean variable: 1 if the time until the occurrence of the first event is observed, 0 if it is censored;
- $M[,4]$ – boolean variable: 1 if the time until the occurrence of the second event is observed, 0 if it is censored.

Step 1 The values of the columns of M are sorted in such a way that the uncensored observations relative to the first time and relative to the total time appear first.

Step 2 Assign a weight to each observation in such a way that the observations with the highest rank have a greater weight. In the case of the estimator proposed by J. de Unã-Álvarez and L. Meira-Machado [1], the censored observations both in the first time and in the total time have a weight of 0. In the remaining estimators, the weight assigned to observation i is a function of the probability of this observation being censored in the second time, knowing that it was not censored in the first time.

Example of the R code for this procedure:

```
R <- rank(M[, 2], ties.method="first")
n <- nrow(M)

  Pkm <- rep(1,n)
  for (i in 1:n){
    for (j in 1:n){
     if (R[j] < R[i])
      Pkm[i] <- Pkm[i]*(1 - M2[j,4]/(n-R[j]+1))
     }
    Wkm[i,1] <- Pkm[i]*M2[i,4]/(n-R[i]+1)
  }

   n1 <- sum(M[,3])

  glm.fitted <- fitted (glm(M[1:n1, 4] ~ M[1:n1,1] + M2[1:n1, 2],
   family=binomial))

  Mlogit <- c(glm.fitted, rep(0, n-n1))

  P1 <- rep(1,n)
  for (i in 1:n){
```

```
    for (j in 1:n){
     if (R[j]<R[i])
      P1[i] <- P1[i]*(1-Mlogit[j]/(n-R[j]+1))
       }
    W1[i,1]<-P1[i]*Mlogit[i]/(n-R[i]+1)
    }
}
```

Step 3 :: Define two indicator matrices I1 and I2 to indicate, respectively, the concordant and discordant pairs in the data set.

R code for this procedure:

```
 t2 <- M [ , 2] - M [ , 1]
 for (i in 1:n) {
  for (j in 1:n) {
    if((M[j,1]<M[i,1]  &  t2[j]<t2[i])|(M[j,1]>M[i,1]
          & t2[j] > t2[i])) I1[i j] <- 1 elseI I1[i,j] <- 0
    if((M[j,1] > M[i,1] & t2[j] < t2[i])|(M[j,1]<M[i,1]
          & t2[j] > t2[i])) I2[i,j] <- 1 elseI I2[i,j] <- 0
    }
    }
```

Step 4 :: Calculate the estimates of probability of concordance, probability of discordance and the estimate of Kendall's tau coefficient.

R code for this procedure:

```
hatpc1 <- t(as.matrix(W1))%*%I1%*%(as.matrix(W1))
hatpckm <- t(as.matrix(Wkm))%*%I1%*%(as.matrix(Wkm))
hatpd1 <- t(as.matrix(W1))%*%I2%*%(as.matrix(W1))
hatpdkm <- t(as.matrix(Wkm))%*%I2%*%(as.matrix(Wkm))
tau1 <- hatpc1-hatpd1
tau_km <- hatpckm-hatpdkm
```

4 Simulation Study

We can simulate correlated gap times by using copulas. In this work, we simulated gap times with unitary exponential marginal distribution, obtained from the Frank copula. The motivation for using the Frank copula is justified because it allows to obtain positive or negative, strong or moderate, correlations. Furthermore, this copula does not have any tail dependence, so the dependences are relatively similar for all values of the marginals.

The Frank copula is an archimedean copula, with associaton parameter $\alpha \in \mathbb{R} - \{0\}$, with generator ϕ given by

$$\phi(t) = -\log\left(\frac{\mathrm{e}^{-\alpha t} - 1)}{\mathrm{e}^{-\alpha} - 1}\right),\ t \in [0, 1]$$

For this copula, the Kendall's tau coefficient is given by

$$\tau = 1 + \frac{4(D(\alpha) - 1)}{\alpha}, \quad \text{with} \quad D(\alpha) = \frac{1}{\alpha}\int_0^\alpha \frac{t}{\mathrm{e}^t - 1}\mathrm{d}t$$

Samples of dimensions 50, 100, 150, 200 and 250 were considered. To implement random censoring, for both the first gap time and the total time, we independently generated uniform times on the interval $[0, N]$, where N was selected to achieve a given proportion of censoring. Assuming independence between the random variables $C \sim U[0, N]$ with distribution function G (density g) and $Y \sim Exp(1)$ with density f_y, we have

$$P(C < Y) = \int_0^\infty \int_0^y f_y(y) g(c) \mathrm{d}c\mathrm{d}y = \int_0^N \frac{1 - \mathrm{e}^{-y}}{N}\mathrm{d}y = \frac{N - \sinh(b) + \cosh(b) - 1}{N}$$

In the present work we take $N = 4$ to reach about 25% censoring for the first gap time and a little more than 48% censoring for the total time.

```
#t1 first gap time; y total time
  cens[,1] = runif(n,0,N)
  for (i in 1:n){
    ytilde[i,1] = min(y[i,1],cens[i,1])
    d[i,1]=1
    d1[i,1]=1
  }
  for (i in 1:n){
    if (ytilde[i,1] < y[i,1]) d[i,1]=0
  }
  for (i in 1:n){
    t1tilde[i,1] = min(t1[i,1],cens[i,1])
  }
   for (i in 1:n){
    if (t1tilde[i,1] < t1[i,1]) d1[i,1]=0
  }
```

We considered 10,000 repetitions of each procedure for generate the estimates of Kendall's tau coefficient. The final estimate was the mean of the estimates produced in the simulation process. The standard deviation, the bias and the Mean Squared Error of the estimate were also calculated.

4.1 Simulation Results

The tables presented in this section contain the results of the simulations, namely the estimates for τ and the corresponding bias, standard deviation (SD) and the Mean Squared Error (MSE) of the estimator. The different methods are identified as `WKM` for the weights given by expression (7), `WSP` for the weights in the semiparametric estimator (9), `WTree` and `WRF` have the same expression for the weights as the latter, but the estimated probabilities for the weights based on the model m (see (8)) are calculated with decision trees and random forests methodologies, respectively.

In all cases, the values of standard deviation and MSE decrease as the sample size increases, implying the consistency of the estimates. In what concerns to bias it gets smaller and smaller with increasing sample size in all cases except for low negative association in WSP estimator (Table 2).

Low Positive Association The Table 1 show the performance for all estimators for gap times with low positive association. In this case, WSP performs better with lower SD and MSE for all sample sizes considered.

Table 1. True tau 0.1100

n	*method*	$\widehat{\tau}$	$\widehat{\tau}-\tau$	$SD(\widehat{\tau})$	$MSE(\widehat{\tau})$
50	WSP	0.0270	−0.0830	**0.0983**	**0.0165**
	WKM	0.0241	−0.0860	0.1269	0.0235
	WTree	0.0223	−0.0878	0.1065	0.0190
	WRF	0.0215	−0.0885	0.1087	0.0197
100	WSP	0.0348	−0.0753	**0.0737**	**0.0111**
	WKM	0.0272	−0.0828	0.0952	0.0159
	WTree	0.0293	−0.0807	0.0818	0.0132
	WRF	0.0272	−0.0828	0.0822	0.0136
150	WSP	0.0367	−0.0733	**0.0623**	**0.0093**
	WKM	0.0277	−0.0823	0.0805	0.0132
	WTree	0.0291	−0.0810	0.0698	0.0114
	WRF	0.0285	−0.0816	0.0702	0.0116
200	WSP	0.0381	−0.0719	**0.0557**	**0.0083**
	WKM	0.0277	−0.0823	0.0719	0.0119
	WTree	0.0292	−0.0809	0.0625	0.0104
	WRF	0.0289	−0.0812	0.0629	0.0105
250	WSP	0.0402	−0.0698	**0.0513**	**0.0075**
	WKM	0.0280	−0.0820	0.0659	0.0111
	WTree	0.0292	−0.0808	0.0581	0.0099
	WRF	0.0297	−0.0803	0.0579	0.0098

Low Negative Association In what concerns to data with low negative association, the results present in Table 2, reveal that WSP estimator is again the best estimator, but as the sample size grows, there is a change in the sign of the bias of this estimator. The smallest values of bias are achived for moderately sized samples ($n = 100, 150$).

Table 2. True tau -0.1100

n	*method*	$\widehat{\tau}$	$\widehat{\tau}-\tau$	$SD(\widehat{\tau})$	$MSE(\widehat{\tau})$
50	WSP	**−0.1052**	**0.0048**	**0.1097**	**0.0121**
	WKM	−0.1211	−0.0111	0.1286	0.0166
	WTree	−0.1163	−0.0063	0.1148	0.0132
	WRF	−0.1209	−0.0109	0.1136	0.0130
100	WSP	**−0.1099**	**0.0002**	**0.0841**	**0.0071**
	WKM	−0.1260	−0.0161	0.0986	0.0100
	WTree	−0.1192	−0.0092	0.0872	0.0077
	WRF	−0.1246	−0.0146	0.0874	0.0078
150	WSP	**−0.1109**	**−0.0009**	**0.0704**	**0.0050**
	WKM	−0.1272	−0.0171	0.0827	0.0071
	WTree	−0.1214	−0.0114	0.0730	0.0054
	WRF	−0.1254	−0.0154	0.0737	0.0057
200	WSP	**−0.1127**	**−0.0027**	**0.0631**	**0.0040**
	WKM	−0.1286	−0.0186	0.0728	0.0056
	WTree	−0.1231	−0.0130	0.0650	0.0044
	WRF	−0.1270	−0.0170	0.0650	0.0045
250	WSP	**−0.1131**	**−0.0031**	**0.0586**	**0.0034**
	WKM	−0.1294	−0.0194	0.0664	0.0048
	WTree	−0.1248	−0.0147	0.0595	0.0038
	WRF	−0.1276	−0.0176	0.0597	0.0039

Moderate Positive Association In case of moderate positive association, the results presented in Table 3 show that although the estimative of τ is closer to WKM than the other estimates, this one has greater variability. In terms of consistency, both WSP and WRF perform better with a lower MSE, the latter being slightly better because of its lower bias.

Table 3. True tau 0.3881

n	*method*	$\widehat{\tau}$	$\widehat{\tau}-\tau$	$SD(\widehat{\tau})$	$MSE(\widehat{\tau})$
50	`WSP`	0.1922	−0.1959	0.0966	0.0477
	`WKM`	0.2070	−0.1811	0.1249	0.0484
	`WTree`	0.1890	−0.1992	0.0981	0.0492
	`WRF`	0.1958	−0.1923	0.1028	**0.0476**
100	`WSP`	0.2093	−0.1788	0.0722	**0.0372**
	`WKM`	0.2188	−0.1694	0.0944	0.0376
	`WTree`	0.2048	−0.1833	0.0765	0.0395
	`WRF`	0.2122	−0.1759	0.0788	**0.0372**
150	`WSP`	0.2153	−0.1729	0.0615	0.0337
	`WKM`	0.2222	−0.1659	0.0816	0.0342
	`WTree`	0.2124	−0.1758	0.0679	0.0355
	`WRF`	0.2184	−0.1697	0.0689	**0.0335**
200	`WSP`	0.2208	−0.1673	0.0553	**0.0311**
	`WKM`	0.2256	−0.1626	0.0732	0.0318
	`WTree`	0.2180	−0.1701	0.0627	0.0329
	`WRF`	0.2231	−0.1650	0.0621	**0.0311**
250	`WSP`	0.2242	−0.1639	0.0506	**0.0294**
	`WKM`	0.2269	−0.1613	0.0676	0.0306
	`WTree`	0.2209	−0.1673	0.0587	0.0314
	`WRF`	0.2256	−0.1625	0.0573	0.0297

Moderate Negative Association In the case of moderate negative association, the results presented in the Table 4 show a better performance for all estimators, with the bias being considerably reduced compared to the corresponding values in the case of moderate positive association. However, as in the previous case, the estimator `WKM` exhibits greater variability and a higher value for the MSE than any of the estimators `WSP` and `WRF`. In this case `WRF` performs better.

Table 4. True tau -0.3881

n	*method*	$\widehat{\tau}$	$\widehat{\tau}-\tau$	$SD(\widehat{\tau})$	$MSE(\widehat{\tau})$
50	`WSP`	-0.3056	0.0826	0.1152	0.0201
	`WKM`	-0.3167	0.0714	0.1261	0.0210
	`WTree`	-0.3130	0.0751	0.1137	0.0186
	`WRF`	-0.3154	0.0727	0.1128	**0.0180**
100	`WSP`	-0.3234	0.0648	0.0855	0.0115
	`WKM`	-0.3309	0.0572	0.0939	0.0121
	`WTree`	-0.3307	0.0575	0.0863	0.0107
	`WRF`	-0.3298	0.0583	0.0843	**0.0105**
150	`WSP`	-0.3311	0.0570	0.0720	0.0084
	`WKM`	-0.3380	0.0502	0.0796	0.0088
	`WTree`	-0.3374	0.0508	0.0730	0.0079
	`WRF`	-0.3372	0.0510	0.0715	**0.0077**
200	`WSP`	-0.3364	0.0517	0.0634	0.0067
	`WKM`	-0.3408	0.0474	0.0709	0.0073
	`WTree`	-0.3411	0.0470	0.0652	0.0065
	`WRF`	-0.3405	0.0476	0.0638	**0.0063**
250	`WSP`	-0.3392	0.0488	0.0582	0.0057
	`WKM`	-0.3435	0.0446	0.0644	0.0061
	`WTree`	-0.3435	0.0446	0.0596	0.0055
	`WRF`	-0.3433	0.0448	0.0582	**0.0054**

High Positive Association In relation to high positive association, according to the results shown in the Table 5, all estimators present worse performance when compared to the corresponding ones in the case of moderate positive association. Both the bias, the standard deviation and the MSE present higher values in all estimators. For strong positive association, the estimator `WSP` presents a smaller MSE and the estimator `WTree` has lower SD.

Table 5. True tau 0.7626

n	*method*	$\widehat{\tau}$	$\widehat{\tau}-\tau$	$SD(\widehat{\tau})$	$MSE(\widehat{\tau})$
50	WSP	0.4849	−0.2778	0.1226	**0.0921**
	WKM	0.4856	−0.2770	0.1368	0.0954
	WTree	0.4136	−0.3489	**0.1058**	0.1329
	WRF	0.4465	−0.3160	0.1153	0.1132
100	WSP	0.5212	−0.2414	0.0995	**0.0681**
	WKM	0.5079	−0.2547	0.1088	0.0767
	WTree	0.4410	−0.3216	**0.0790**	0.1096
	WRF	0.4776	−0.2850	0.0922	0.0897
150	WSP	0.5363	−0.2263	0.0889	**0.0591**
	WKM	0.5128	−0.2498	0.0937	0.0712
	WTree	0.4593	−0.3033	**0.0705**	0.0970
	WRF	0.4888	−0.2738	0.0799	0.0813
200	WSP	0.5485	−0.2140	0.0814	**0.0524**
	WKM	0.5205	−0.2421	0.0847	0.0658
	WTree	0.4748	−0.2878	**0.0663**	0.0872
	WRF	0.5006	−0.2619	0.0730	0.0739
250	WSP	0.5538	−0.2088	0.0786	**0.0498**
	WKM	0.5219	−0.2407	0.0809	0.0645
	WTree	0.4840	−0.2785	**0.0638**	0.0816
	WRF	0.5053	−0.2573	0.0694	0.0710

High Negative Association The results in the Table 6 are in line with what happens in moderate negative association. There is a considerable reduction in the bias, standard deviation and MSE of the estimators in a scenario of strong negative association. There is also a better performance of the estimator `WSP` in relation to the others.

Table 6. True tau -0.7626

n	*method*	$\widehat{\tau}$	$\widehat{\tau}-\tau$	$SD(\widehat{\tau})$	$MSE(\widehat{\tau})$
50	WSP	−0.6310	0.1315	0.0862	**0.0247**
	WKM	−0.6268	0.1358	0.1094	0.0304
	WTree	−0.6047	0.1578	0.1021	0.0353
	WRF	−0.6171	0.1455	0.0970	0.0306
100	WSP	−0.6609	0.1017	0.0569	**0.0136**
	WKM	−0.6530	0.1096	0.0819	0.0187
	WTree	−0.6507	0.1119	0.0716	0.0176
	WRF	−0.6494	0.1132	0.0717	0.0179
150	WSP	−0.6729	0.0897	0.0457	**0.0101**
	WKM	−0.6656	0.0970	0.0700	0.0143
	WTree	−0.6654	0.0972	0.0592	0.0130
	WRF	−0.6638	0.0988	0.0615	0.0135
200	WSP	−0.6796	0.0830	0.0386	**0.0084**
	WKM	−0.6717	0.0908	0.0628	0.0122
	WTree	−0.6741	0.0885	0.0522	0.0106
	WRF	−0.6717	0.0909	0.0548	0.0113
250	WSP	−0.6834	0.0792	0.0343	**0.0074**
	WKM	−0.6753	0.0873	0.0573	0.0109
	WTree	−0.6786	0.0839	0.0471	0.0093
	WRF	−0.6760	0.0866	0.0494	0.0099

5 Example of Application With Real Data

In this section, the methods described in Section 3 are applied to data from a bladder cancer study in which patients had superficial bladder tumors that were removed. Some patients had multiple recurrences of tumors during the study [3]. The R survival package contains data from 85 subjects in the placebo and thiotepa treatment groups. Considering the first two recurrences times (in months) and the corresponding gap times, T1 and T2 we have, of the total of 85 patients, 47 relapsed at least once and 29 of these had a new recurrence. This data contain a high percentage of total censored time. In fact 66% of total observations is censored and 44.7% of the observations on first gap time are censored (see Figure 1).

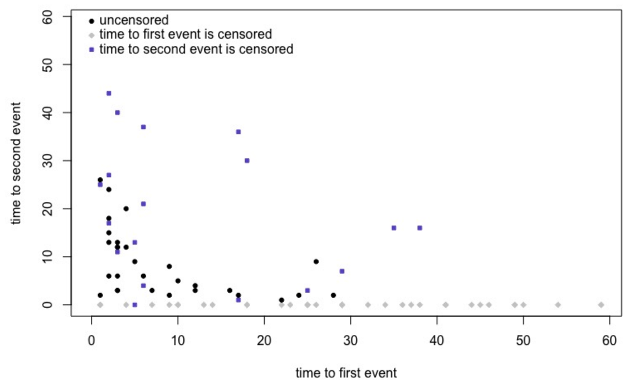

Fig. 1. Bladder data with censored observations

In this example, the estimates obtained with the presmoothing estimator, `WSP`, and with the estimator with weights obtained using the random forests methodology, `WRF`, show less variability, according with the results of a nonparametric bootstrap approach to calculate the confidence intervals of Kendall's τ (see Table 7). Table 7 also contains the value of the point estimate of τ, using the various estimators, as well as the mean and standard deviation calculated from 200 bootstrap samples. If we consider the results of simulation presented in this study, `WSP` is the best estimator in this case, so there is no evidence of association of the two gap times in this study (bootstrap quantiles in Table 7).

Table 7. Tau Estimatives and Bootstrap CI: Bladder Data

Method	$\widehat{\tau}$	Mean	SD	$\chi_{.025}$	$\chi_{.975}$
`WSP`	**−0.0614**	−0.0619	**0.0345**	**−0.1253**	**0.0047**
`WKM`	−0.0915	−0.0915	0.0366	−0.1701	−0.0268
`WTree`	−0.0979	−0.0979	0.0449	−0.1542	0.0186
`WRF`	−0.0857	−0.0857	**0.0350**	−0.1513	−0.0175

6 Conclusions

In this paper we estimate Kendall's tau coefficient from the estimates of the probability of concordance and discordance of right-censored gap times pairs. This is not the usual approach in the papers that have been published on this topic. In general, the authors use a compact formula for the estimator, considering the complementarity of concordant and discordant events. Both for estimating the probability of concordance and for estimating the probability of discordance, we used estimators of the joint distribution function of the gap times under right censoring, as well as estimators for the marginal distribution function of the second gap time. Recall that for the second gap time, the distributions of T_2 and censorship are not independent. The approach followed to accommodate this dependency is not new. However, this paper presents two new alternatives for smoothing the weights of the estimator. These alternatives consist of considering decision trees and random forests methodologies, to calculate the probabilities associated with the occurrence of censoring in the second gap time, given the total time and the values of the first gap time.

The results of the simulations are compatible with the behavior already known of the estimators of the joint distribution function of pairs of gap times previously studied. In fact, the lower variability of the presmoothed semiparametric estimator (see [2]) in relation to the weighted Kaplan-Meier is also presented in the corresponding Kendall's tau estimators. Regarding the estimator with smoothed weights using random forests, the simulation results are compatible with the best performance of this one in relation to both the weighted Kaplan-Meier and the smoothed weight estimator using decision trees.

Estimators generally perform better in a scenario of negative association of gap times. Furthermore, if the association is strongly negative, there is a very marked reduction in bias, standard deviation and MSE of all estimators, even in a context of small or moderate samples. When the association is moderately negative, the performance of estimator with smoothed weights using random forests is superior to the other estimators.

In the case of positive association, the best performing estimator is the presmoothed nonparametric, if the association is strong and, in the case of moderate positive association, this estimador and the estimator with smoothed weights using random forests perform identically.

Acknowledgements This work was supported by Portuguese funds through the CMAT - Research Centre of Mathematics of University of Minho - within projects UIDB/00013/2020 and UIDP/00013/2020.
Authors thank to referees for their careful reading and for their constructive suggestions.

References

1. Unã-Álvarez, J., Meira-Machado, L.: A simple estimator of the bivariate distribution function for censored gap times. Statistics and Probability Letters **78**, 2440–2445 (2008). https://doi.org/10.1002/bimj.201000063
2. Unã-Álvarez, J., Amorim, A.: A simple semiparametric estimator of the bivariate distribution function for censored gap times. Biometrical Jornal **53**(1), 113–127 (2011). https://doi.org/10.1002/bimj.201000063
3. Byar, D.P. : The Veterans Administration Study of Chemoprophylaxis for Recurrent Stage I Bladder Tumours: Comparisons of Placebo, Pyridoxine and Topical Thiotepa. In: Pavone-Macaluso, M., Smith, P.H., Edsmyr, F. (eds) Bladder Tumors and other Topics in Urological Oncology. Ettore Majorana International Science Series, vol 1. Springer, Boston, MA. (1980) https://doi.org/10.1007/978-1-4613-3030-1_74

4. Couronné, R., Probst, P., Boulesteix, AL.: Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinformatics **19**, 270 (2018). https://doi.org/10.1186/s12859-018-2264-5
5. Dabrowska, D. M.: Kaplan–Meier estimate on the plane: weak convergence, LIL, and the bootstrap. Journal of Multivariate Analysis **29**(2), 308–325 (1989). https://doi.org/10.1016/0047-259X(89)90030-4
6. Dikta, D.: On semiparametric random censorship models. Journal of Statistical Planning and Inference **66**, 253–279 (1998). https://doi.org/10.1016/S0378-3758(97)00091-8
7. Dikta, D.: The strong law under semiparametric random censorship models. Journal of Statistical Planning and Inference **83**, 1–10 (2000). https://doi.org/10.1016/S0378-3758(99)00086-5
8. Fan, J., Hsu, L., Prentice, R. L.: Dependence estimation over a finite bivariate failure time region. Lifetime Data Analysis **6**(4), 343–355 (2000). https://doi.org/10.1023/a:1026557315306
9. Kaplan, E. L., Meier, P.: Nonparametric estimation from incomplete observations. Journal of the American Statistical Association **53** (282), 457–481 (1958).
10. Oakes, D.: A model for association in bivariate survival data. Journal of the Royal Statistical Society: Series B (Methodological) **44**(3), 414–422 (1982).
11. Oakes, D.: Bivariate survival models induced by frailties. Journal of the American Statistical Association **84**(406), 487– 493 (1989). https://doi.org/10.2307/2289934
12. Oakes, D.: On consistency of Kendall's tau under censoring. Technical report. Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY (2006).