Data Mining and Knowledge Discovery manuscript No.
(will be inserted by the editor)

Supplementary Materials to

Discrete-Time Survival Forests with Hellinger Distance Decision Trees

Matthias Schmid · Thomas Welchowski ·
Marvin N. Wright · Moritz Berger

A Results of the Simulation Study – Effect of the Minimum Node Size on Prediction Accuracy

Figures 1 and 2 show the prediction accuracy of the *ICcforest*, *HD* and *GI* methods for various choices of the minimum node size and various values of k.

Matthias Schmid, Thomas Welchowski, Moritz Berger
Department of Medical Biometry, Informatics and Epidemiology, University of Bonn
Venusberg-Campus 1, D-53127 Bonn, Germany
Tel.: +49-228-28715400, e-mail: matthias.c.schmid@uni-bonn.de
Marvin N. Wright
Department of Biometry and Data Management, Leibniz Institute for Prevention Research and Epidemiology – BIPS GmbH
Achterstraße 30, D-28359 Bremen, Germany

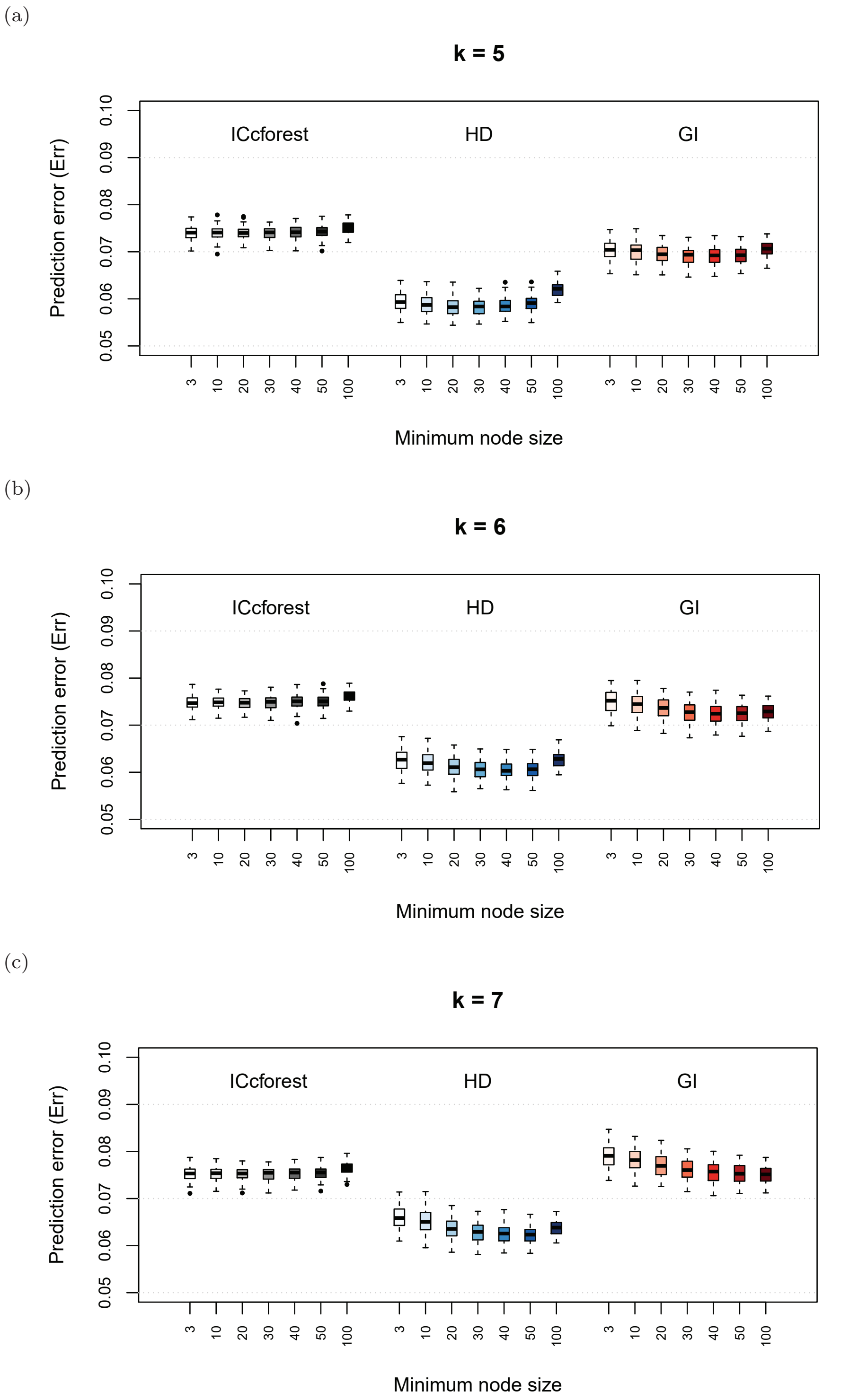

Fig. 1: Results of the simulation study. The boxplots visualize the prediction accuracy (as evaluated by the summary measure *Err*) of the *ICcforest*, *HD* and *GI* methods for various choices of the minimum node size and various values of k.

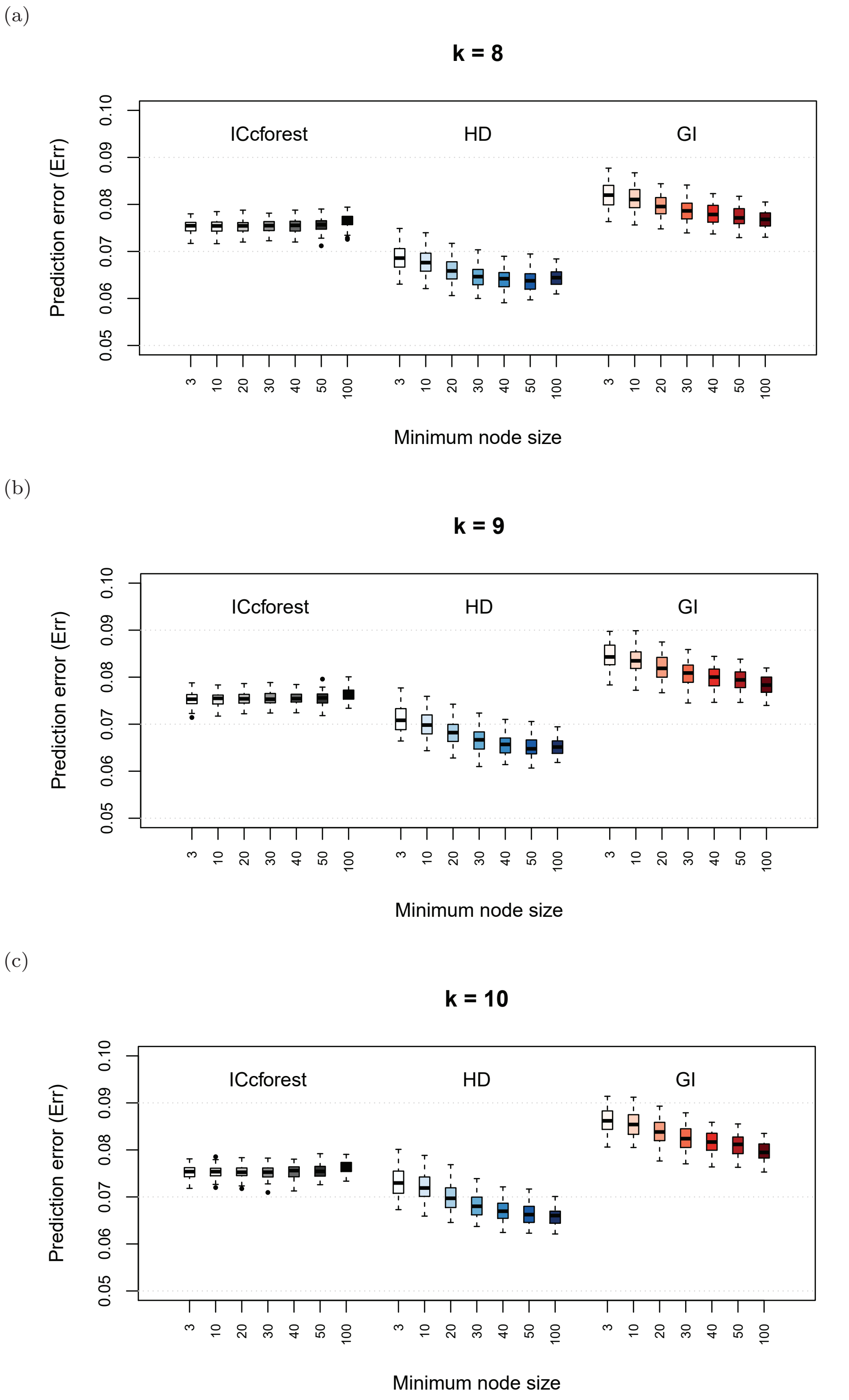

Fig. 2: Results of the simulation study. The boxplots visualize the prediction accuracy (as evaluated by the summary measure *Err*) of the *ICcforest*, *HD* and *GI* methods for various choices of the minimum node size and various values of k.

B Application: Time to First Childbirth

Here we present the results of another real-world application in which we analyzed the time to first childbirth in German women. More specifically, we evaluated data from the first nine waves of the German Family Panel ("pairfam"), which provides data on family dynamics and relationships in Germany (Brüderl et al., 2018). The first wave of the survey in 2008 collected data from a nationwide random sample comprising more than 12,000 respondents of the birth cohorts 1971-1973, 1981-1983, and 1991-1993 and their families. The main focus was on so-called *anchor persons* of a certain birth cohort who were annually interviewed to get detailed information on topics like the development of partnership, family plans and formation as well as attitudes regarding parenting in general. Further details on the study are described in Huinink et al. (2011).

As all information was collected in one-year intervals, the observed duration times of the pairfam study are discrete ($t = 1, \ldots, 9$). The event of interest of our analysis was defined by whether an anchor woman gave birth to her first child or not. Following Groll and Tutz (2017), we restricted our consideration to women of the birth cohorts 1971-1973 and 1981-1983. The resulting analysis data set comprised 4,077 observations of 861 anchor women who stated to have no children in the initial wave. The censoring rate was 68% (corresponding to 273 observed childbirths).

As covariates, we included the educational level of the anchor women measured in years (*yeduc*), the educational levels of the parents of the anchor women measured in years (*myeduc* and *fyeduc*), the degree of life satisfaction of the anchor women (*sat6*, with higher values indicating a higher life satisfaction), the status of the relationship of the anchor women (*relstat*, 0: single, 1: married and/or cohabitation), the employment status of the anchor women (*casprim*, 0: not employed, 1: employed), the number of siblings of the anchor women (*siblings*), and the amount of leisure time spent for going to bars/cafés/restaurants, doing sport, meeting with friends and/or going to a discotheque (*leisure*, 1: daily, 2: at least once a week, 3: at least once a month, 4: less often). A descriptive overview of the covariates in the first wave 2008 is given in Table 1.

To investigate the performance of the discrete-time RSF approach, we conducted a benchmark experiment that was based on 100 random partitions of the pairfam data. Each partition consisted of a learning data set of size $n = 688$ and a test data set of size $n_{\text{test}} = 173$. For model comparison we considered the same approaches as in Section 4 of the paper. To evaluate the predictive performance of the RSF fits, we applied the estimator of the discrete concordance index (C-index, Schmid et al. 2018). Furthermore, we computed estimates of the integrated squared prediction error in each test data set, as described in Section 4 of the paper.

The estimates of the integrated squared prediction error are presented in panel (a) of Figure 3. It is seen that, in contrast to the analysis of the unemployment data in Section 4 of the paper (and in line with the results of the

Table 1: Summary statistics of the covariates that were used to model time to first childbirth in the pairfam data (first wave 2008, $n = 861$).

Variable	Categories / unit	Sample proportion / median (range)
yeduc	years	13 (8.0−20.0)
myeduc	years	11.5 (8.0−20.0)
fyeduc	years	11.5 (8.0−20.0)
sat6		8 (0−10)
relstat	0 / 1	53.4% / 46.6%
casprim	0 / 1	32.9% / 67.1%
siblings		1 (0−16)
leisure	1 / 2	16.1% / 73.5%
	3 / 4	8.2% / 2.1%

simulation study), the *ICcforest* method did not perform better than the *HD* method. Apart from this finding, the numerical results of the paper were again confirmed: In particular, the discrete-time RSF approaches with splitting by Hellinger's distance (*HD* and *HD_BA*) performed better than the respective approaches with splitting by the Gini impurity (*GI_BA* and *GI_BA*). The median values of the integrated squared prediction error (as estimated from the 100 test data sets) were 0.203 (*RSF_cont*), 0.201 (*ICcforest*), 0.195 (*HD*), 0.201 (*HD_BA*), 0.200 (*GI*), 0.206 (*GI_BA*), 0.216 (*E_net*), and 0.245 (*GI_SMOTE*). Similar results were obtained from the estimates of the C-index presented in panel (b) of Figure 3. The median values of the C-index (as estimated from the 100 test data sets) were 0.674 (*RSF_cont*), 0.671 (*ICcforest*), 0.673 (*HD*), 0.673 (*HD_BA*), 0.664 (*GI*), 0.663 (*GI_BA*), 0.648 (*E_net*), and 0.659 (*GI_SMOTE*).

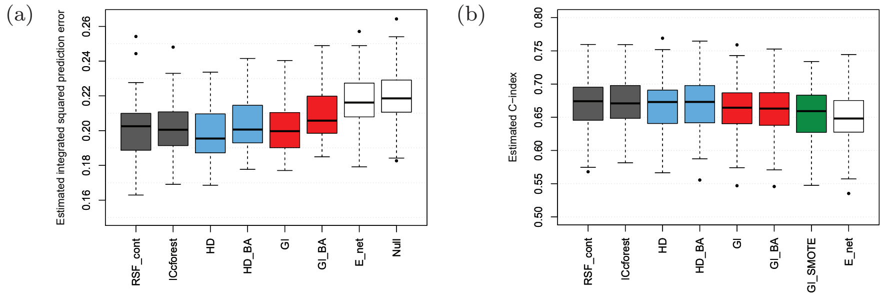

Fig. 3: Analysis of the time to first childbirth. The boxplots in panel (a) visualize the estimates of the integrated squared prediction error, as obtained from fitting various RSF models to 100 pairs of learning and test samples generated from the pairfam data. The rightmost boxplot in panel (a) refers to a discrete-time RSF model that included the time intervals $1, \ldots, \tilde{T}_i$ as only covariate. This model served as a "null" model that would have been used in the absence of any covariate information. Note that the boxplot referring to the *GI_SMOTE* method was excluded from panel (a), as the respective estimates of the integrated squared prediction error (median $= 0.245$, range $= [0.204, 0.308]$) were far higher than the values of the null model. The boxplots in panel (b) visualize the estimated values of the C-index. A reference value for the C-index is given by the value 0.5 (not depicted in the right panel), which corresponds to the C-index of the covariate-free null model.

References

Brüderl J, Drobnic S, Hank K, Huinink J, Nauck B, Neyer F, Walper S, Alt P, Borschel E, Bozoyan C, Buhr P, Finn C, Garrett M, Greischel H, Hajek K, Herzig M, Huyer-May B, Lenke R, Müller B, Peter T, Schmiedeberg C, Schütze P, Schumann N, Thönnissen C, Wetzel M, Wilhelm B (2018) The German Family Panel (pairfam) GESIS Data Archive, Cologne. ZA5678 data file version 9.1.0, doi: 10.4232/pairfam.5678.9.1.0

Groll A, Tutz G (2017) Variable selection in discrete survival models including heterogeneity. Lifetime Data Analysis 23:305–338

Huinink J, Brüderl J, Nauck B, Walper S, Castiglioni L, Feldhaus M (2011) Panel analysis of intimate relationships and family dynamics (pairfam): Conceptual framework and design. Journal of Family Research 23:77–101

Schmid M, Tutz G, Welchowski T (2018) Discrimination measures for discrete time-to-event predictions. Econometrics and Statistics 7:153–164