Supplementary material : The impact of churn on client value in health insurance, evaluation using a random forest under various censoring mechanisms

1 Choice of the parameters *minleaf* and *maxdepth*

The values taken by the parameters $minleaf$ and $maxdepth$ in the numerical applications of the article are optimized so that the different random forest algorithms achieve good performances. In this section, we first provide a sensitivity analysis performed on the simulated data presented in Section 3.1.1 which shows that the parameters $minleaf = 50$ and $maxdepth = 4$ are optimal for the majority of the random forest algorithms considered. We then discuss these optimal parameter values. Finally, we give the results of the sensitivity analysis performed on real data, along with some comments.

1.1 Study of the model's sensitivity to the parameters *minleaf* and *maxdepth* on simulated data

1.1.1 Setting and results

This study is based on the simulated datasets used in Section 3 which corresponds to a Weibull distribution (Case 1), a function $\phi(t) = \log(t + 1)$, and a censoring rate $q \in \{0.1, 0.3, 0.5\}$. The same models as those represented on Fig. 1 are used, except the Cox model (Cr) which is not relevant in this analysis, and $swRF2$ that we added to the compared models. Each model is evaluated under the chosen parameters $maxdepth = 4$ & $minleaf = 50$ and the other settings given in Tab. 4. The means of the MSE over the 100 i.i.d. replicates of the simulation process are given in Fig. 5. For the sake of clarity, the models are divided in two subsets, each represented on the left side and right side of the figure.

$maxdepth$	$minleaf$
4	**50**
10	10
10	20
10	50
10	100
10	200

Tab. 4: Parameters used for the sensitivity analysis performed on simulated data.
In bold, the parameter values used for the applications in the article

For the cases $q = 0.1$ and $q = 0.3$, we can observe that every model, except $RRTr$ (when $q = 0.1$) and $swRF11$ (when $q = 0.3$), achieves its best performance with the setting $maxdepth \in \{4, 10\}$ & $minleaf = 50$. The results are more contrasted when the rate of censoring is higher ($q = 0.5$): while some models still achieve their best MSE with $maxdepth \in \{4, 10\}$ & $minleaf = 50$ ($swRF32$, $RSFr$, $RRRr$, $swRF34$), other models such as $swRF11$, $swRF13$ and $swRF22$ reach their best performances for $minleaf \in \{100, 200\}$. Thus, there is here a clear distinction between the models which use the Kaplan-Meier estimator in each terminal leaf and the models which employ in terminal leaves the IPCW used to grow the trees, these latter requiring more observations in each leaf to give good results.

1.1.2 Comments about the choice of *minleaf*= 50

The value $minleaf = 50$ is bigger than the $minleaf$ values usually reported in the literature. As an example, Zhu & Kosorok (2012) limit the number of observed failures in terminal nodes to six when measuring the performances of RSF and *recursively imputed survival trees* (RIST). Moreover, a common assertion found in the literature is that random forest algorithms perform well if the individual trees are grown to full size or nearly full size, which corresponds to small $minleaf$ values (e.g. $minleaf \leq 5$). For example, we refer to Sun (2010) or Biau & Scornet (2016). We would like to challenge this common belief in view of the results of our sensitivity analysis. Our results clearly demonstrate that, no matter which random forest model is used, a $minleaf$ of 10 is not big enough to reach the optimal range of MSE for our application. In fact, as supported by Segal (2004), the choice of the parameters which control the size of the trees involves a bias-variance trade-off. Scornet (2017) endorses this point of view, arguing that a high signal/noise ratio in the data, for a given classification of regression problem, causes large trees to perform well, while a lower signal/noise ratio leads to small trees performing better. The article analyzes the optimization of the parameters in the random forest algorithm and finds no

theoretical reason to use the default values proposed by Breiman ($minsplit = 5$ for regression), concluding that optimizing the parameters which control the size of the terminal leafs and the size of the bootstrap samples improve the performance. From a practical point of view, many recent works insist on the importance of the tree size optimization for random forests : e.g. Boulesteix et al. (2012), Huang & Boutros (2016), or Probst et al. (2018).

1.2 Sensitivity analysis on real data

The results of the sensitivity analysis performed on real data are given on Fig. 6. The chosen parameters $maxdepth = 5$ & $minleaf = 100$ are compared with the settings given in Tab. 5. The results justify our parameter choices and are coherent with the analysis made for simulated data (with $q = 0.5$). Indeed, our real data application is an example of a situation where the signal/noise ratio is low. The C-index values given in Fig. 3 are around 0.56, which is low compared to the results obtained with the various datasets (except the *transplant* dataset) in Ishwaran et al. (2008). The percentage of explained variance $R2$ is about 0.03, and thus quite low. Therefore, it is necessary to use small trees, with $maxdepth = 5$ & $minleaf = 100$ for a training set composed of 5000 observations, to achieve optimal performances with random forest algorithms.

$maxdepth$	$minleaf$
5	**100**
10	50
10	100
10	200
10	500

Tab. 5: Parameters used for the sensitivity analysis performed on real data. In bold, the parameter values used for the applications in the article

2 Other results on simulated data

The Fig. 7 shows the complete results obtained for the sixteen models considered in the simulated data experiment of Section 3, completing the information presented in Fig. 1. Of course, the results presented on Fig. 7 are consistent with the analysis made in Section 3.2.1.

3 Other results on real data and further comments

The Fig. 8 & 9 complete the results obtained on real data presented in Section 4.3. It is worthwhile noting that the model $RLTr$ performs slightly better than the model $nRLTr$. It is not surprising to have such a small difference between the two models since the data is not high-dimensional, with only six covariates. The fact that $swRF32$ is the best model in our real data application suggests that conditional IPCW are effective at selecting the optimal splits in the early steps of the tree growing (with a split criteria being unbiased under the conditional independence assumption), but less reliable to estimate individual tree predictions in terminal leaves especially when the censoring rate is high and the leaves contain few non-censored observations.

Even if the explained variance of our model is only about three percent, it is still very useful to optimize a model for churn prediction, because at the aggregated level of an insurance portfolio made of 200 000 policies, small improvements in the prediction of individual risks result in large impacts on the cash flows of the company. In fact, this situation of low signal/noise ratio is common in the domain of survival analysis. Sometimes, even when a model manages to rank the relative risks of the observations with a good accuracy, it can not predict the target duration with a low uncertainty. For instance in reliability analysis, it is usually hard to predict precisely the failure time of a component. This is illustrated in the work of Hong et al. (2009), who found in their study that the prediction intervals for the remaining lifetimes of power transformers are very large.

References

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, *25*(2), 197–227.

Boulesteix, A.-L., Janitza, S., Kruppa, J., & König, I. R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *2*(6), 493–507.

Hong, Y., Meeker, W. Q., & McCalley, J. D. (2009). Prediction of remaining life of power transformers based on left truncated and right censored lifetime data. *The Annals of Applied Statistics*, *3*(2), 857–879.

Huang, B. F., & Boutros, P. C. (2016). The parameter sensitivity of random forests. *BMC bioinformatics*, *17*(1), 331.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. *The annals of applied statistics*, 841–860.

Probst, P., Wright, M. N., & Boulesteix, A.-L. (2018). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, e1301.

Scornet, E. (2017). Tuning parameters in random forests. *ESAIM: Proceedings and Surveys*, *60*, 144–162.

Segal, M. R. (2004). *Machine learning benchmarks and random forest regression* (Tech. Rep.). UCSF: Center for Bioinformatics and Molecular Biostatistics.

Sun, Y. V. (2010). Multigenic modeling of complex disease by random forests. In *Advances in genetics* (Vol. 72, pp. 73–99). Elsevier.

Zhu, R., & Kosorok, M. R. (2012). Recursively imputed survival trees. *Journal of the American Statistical Association*, *107*(497), 331–340.

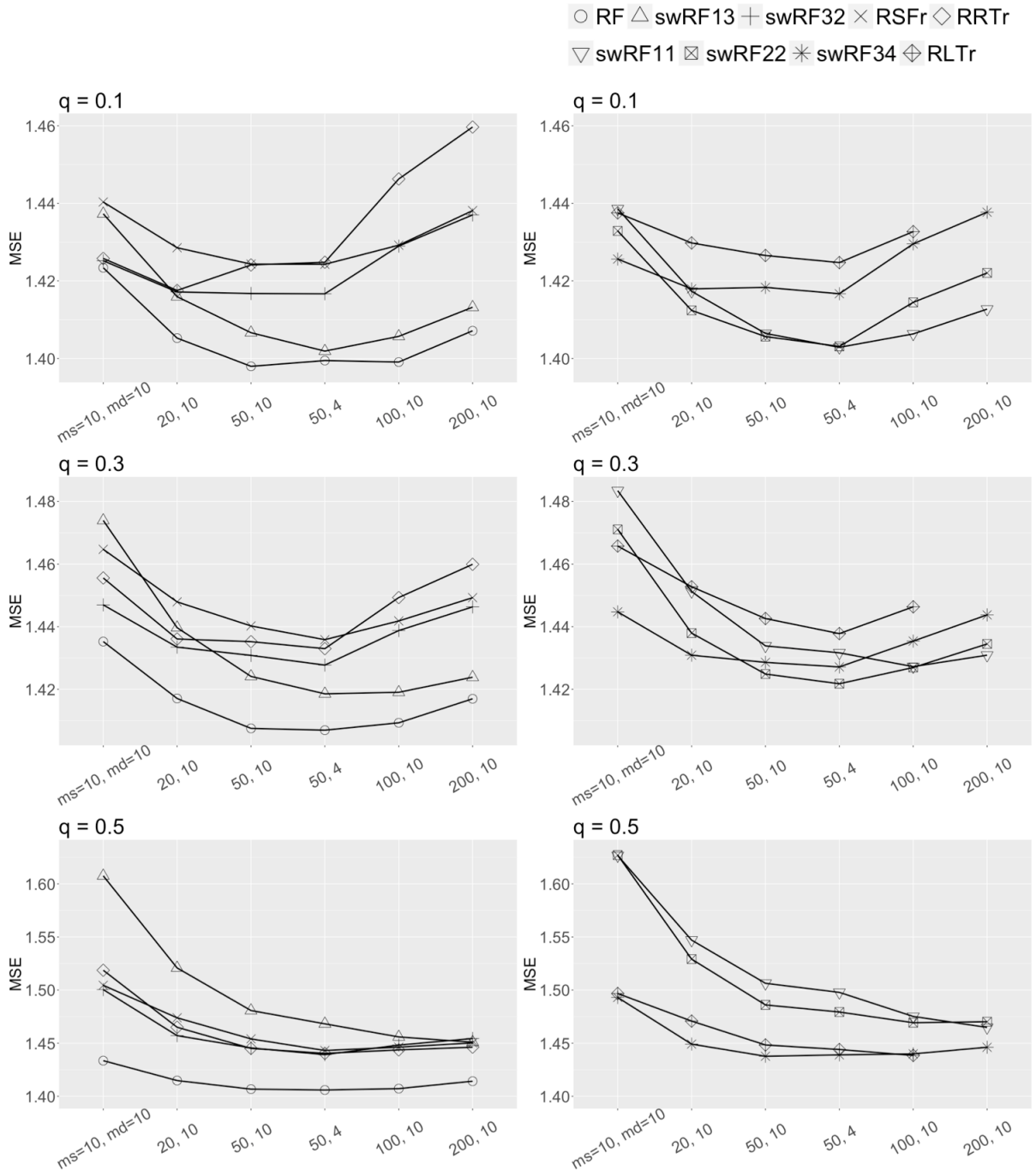

Fig. 5: Results of the sensitivity analysis on simulated data.

The mean of the MSE over the 100 i.i.d. replicates of the simulation process, for each random forest model and each pair of parameters *maxdepth* (*md*) & *minleaf* (*ms*). For *RLTr*, *embed.ntrees* is set to 10. For each random forest, *ntree* = 100 and *mtry* = p = 6.

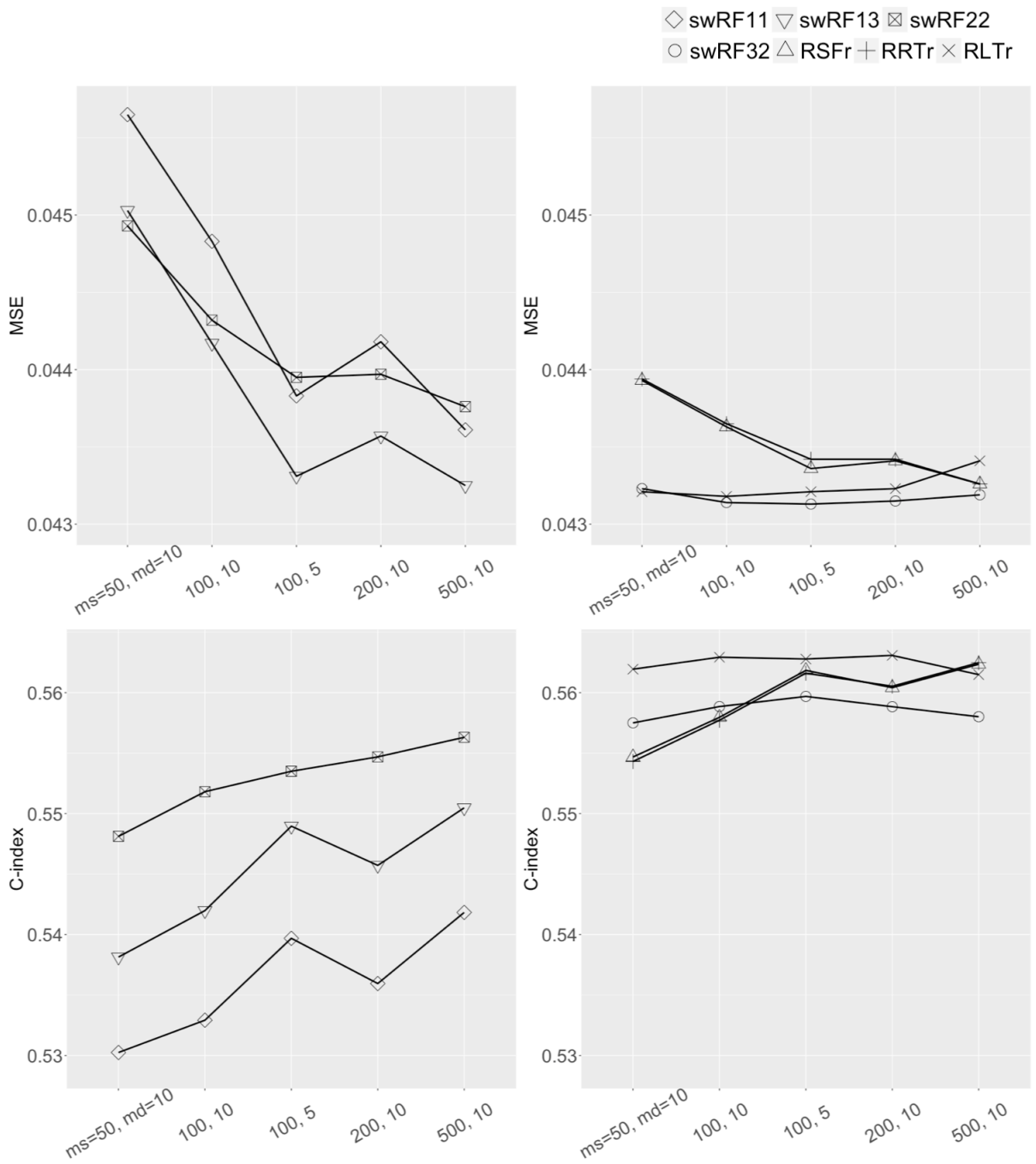

Fig. 6: Results of the sensitivity study on real data.

The mean of the MSE (estimated with RSF weights) and C-index over the 100 i.i.d. replicates of the simulation process, for each random forest model and each pair of parameters *maxdepth* (*md*) & *minleaf* (*ms*). For *RLTr*, *embed.ntrees* is set to 10. For each random forest, *ntree* = 100 and *mtry* = p = 6.

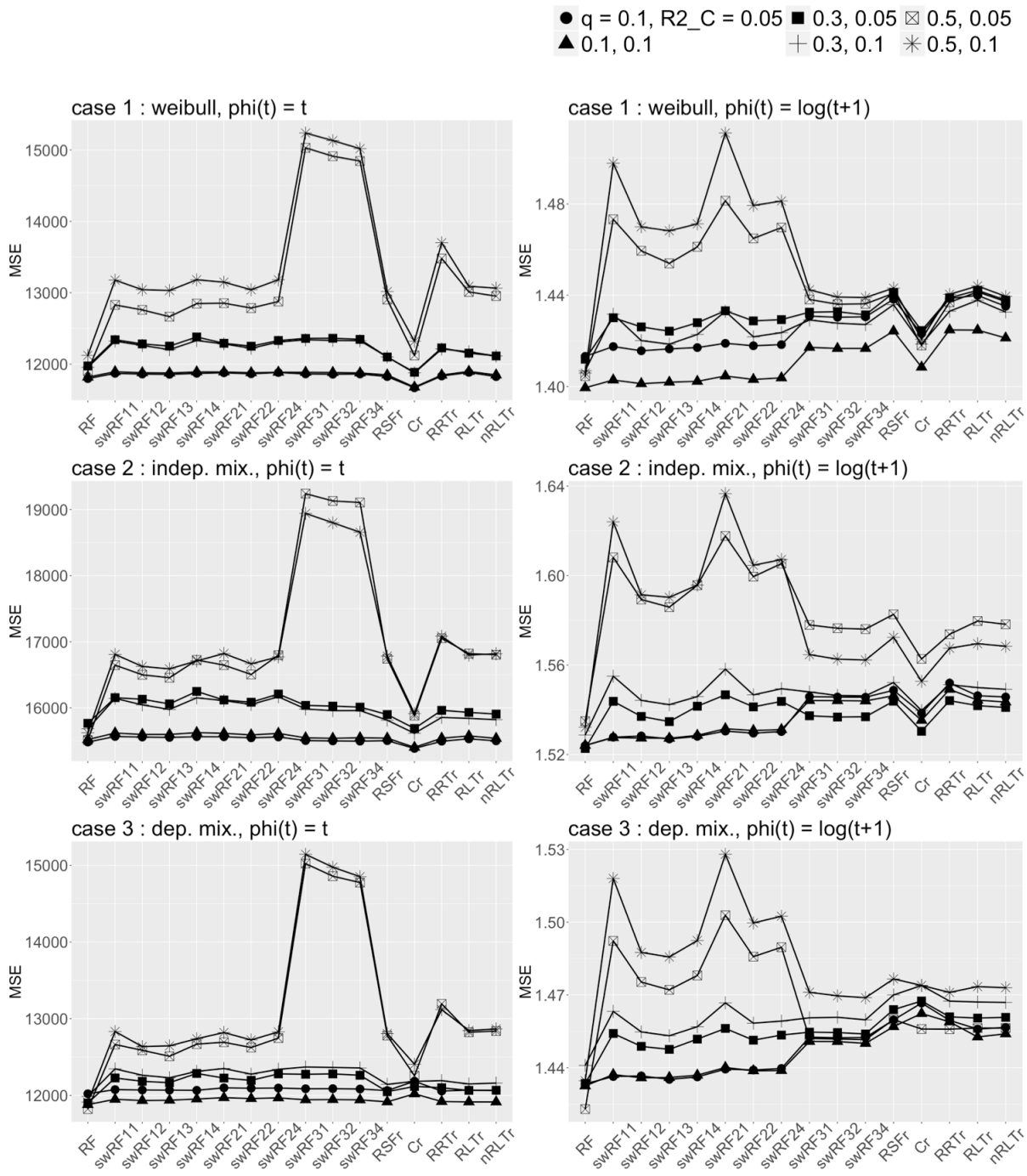

Fig. 7: Results on simulated data - all models

For each setting, the mean of the MSE values over 100 i.i.d. replicates of the simulation process is shown. The censoring rate q is equal to 0.1, 0.3, or 0.5, while the percentage of explained variance of C given X: $R2_C$, is set to 0.05 or 0.1.

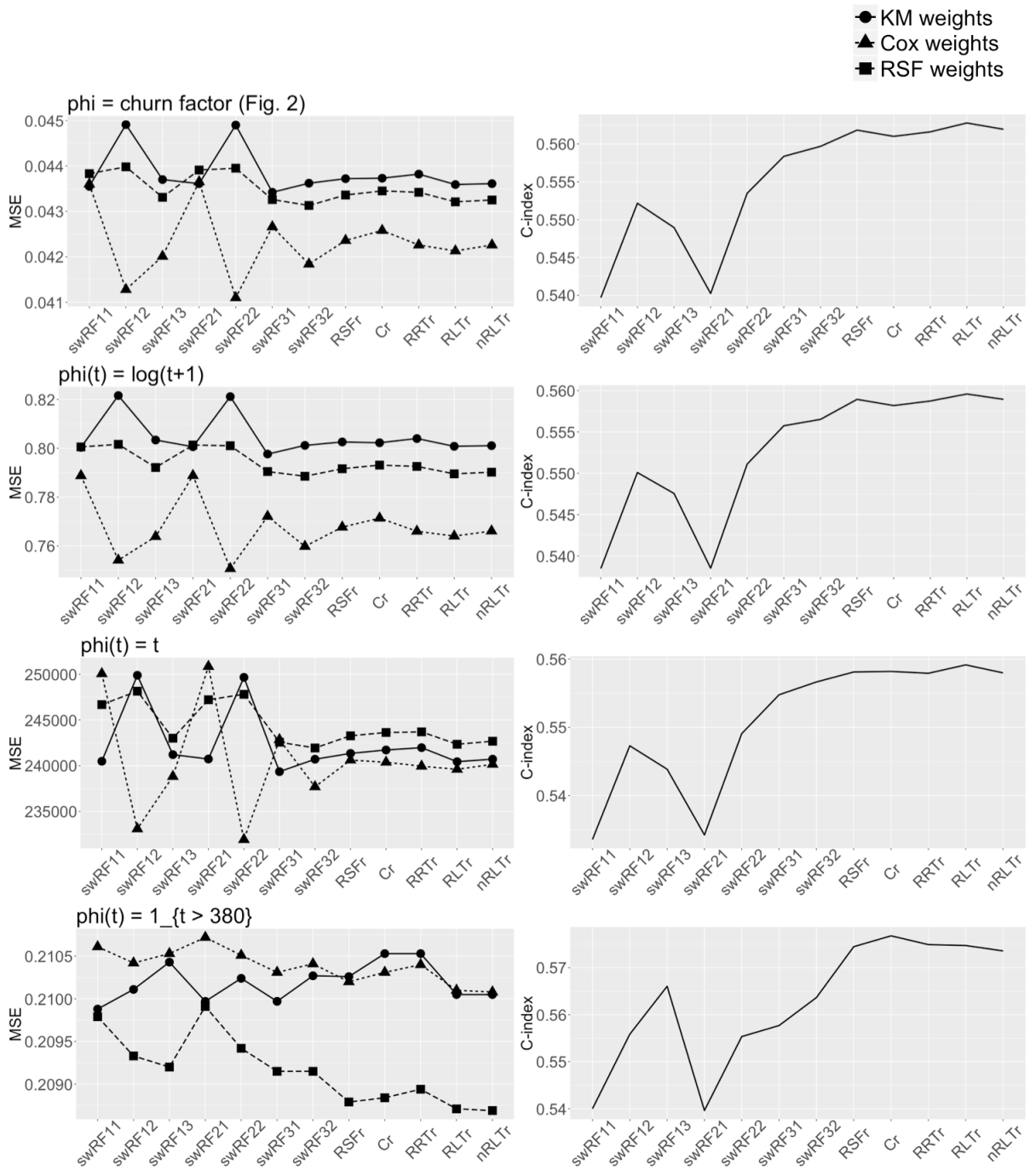

Fig. 8: Results on real data - all models

Left : the mean of the MSE values over 100 i.i.d. replicates of the simulation process, computed with KM, Cox and RSF weights. Right : the mean of the C-index over 100 i.i.d. replicates.

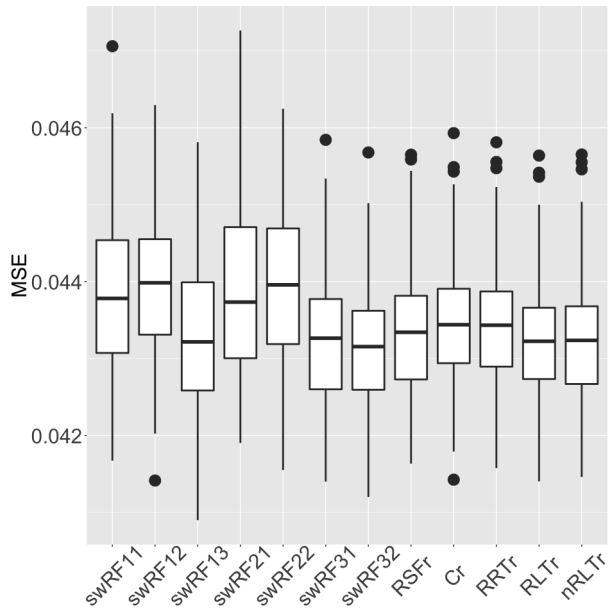

Fig. 9: Boxplot of the performances (MSE) of the models; $\phi = \phi_{ch}$, with RSF weights.

model	mean rank
$swRF11$	9.0
$swRF12$	9.5
$swRF13$	4.8
$swRF21$	9.6
$swRF22$	9.6
$swRF31$	4.7
$swRF32$	2.8
$RSFr$	5.9
Cr	6.9
$RRTr$	6.9
$RLTr$	3.7
$nRLTr$	4.4

Tab. 6: Mean ranks of the models; $\phi = \phi_{ch}$, with RSF weights.