Machine-Learning techniques for family demography: An application of random forests to the analysis of divorce determinants in Germany

Bruno Arpino[1], Marco Le Moglie[2], Letizia Mencarini[2]

[1.] Department of Political and Social Sciences, Universitat Pompeu Fabra, Barcelona, Spain.

[2.] Dondena Centre for Research on Social Dynamics and Public Policy, Bocconi University, Milan, Italy.

RECSM Working Paper Number 56

April 2018

Machine-Learning techniques for family demography: An application of random forests to the analysis of divorce determinants in Germany

Abstract

Demographers often analyze the determinants of life-course events with parametric regression-type approaches. Here, we present a class of nonparametric approaches, broadly defined as machine learning (ML) techniques, and discuss advantages and disadvantages of a popular type known as *random forest*. We argue that random forests can be useful either as a substitute, or a complement, to more standard parametric regression modeling. Our discussion of random forests is intuitive and we illustrate its implementation by analyzing the determinants of divorce with SOEP data for German women entered in a marriage or a cohabitation from 1984 to 2015. The algorithm is able to classify divorce determinants according to their importance, highlighting the most powerful ones, which in our data are partners' overall life satisfaction, their age, and also certain personality traits (i.e., extroversion of the partner and – though with less power – also women’s conscientiousness, agreeableness and openness). We are also able to draw partial dependence plots for the main predictors of survival of the relationship.

1. Introduction: Machine learning as a useful tool for demographic research

Family demographers are often interested in analyzing the determinants of life-course events, such as union formation and dissolution, childbearing, retirement, etc. Often, parametric regression approaches are employed. In particular, when time-to-event data are analyzed, (semi)parametric event history models are routinely estimated (Allison, 1984; Blossfeld et al. 2014; Hoem, 1993). In this paper, we discuss a class of nonparametric approaches, broadly defined as machine learning (henceforth, ML) techniques.

ML is a subfield of artificial intelligence that has attracted considerable interest in the last decade. Also in the social sciences there is a growing interest in it as proved, for example, by the increasing important role of ML techniques in Master's degrees programs specialized in data analytics, big data or data science and by the upsurge of specialized workshops, conferences and research calls on big data or data analytics within which ML has a preponderant role. Despite this growing interest, applications of ML in social sciences, and especially in demography, remain relatively scarce.

There are two main reasons behind the limited number of applications of ML in social sciences. The first one is practical and related to the complexity of these techniques. ML is a vast field, ML algorithms are not intuitive and there is a paucity of accessible resources for social scientists to learn about these techniques. Moreover, most ML algorithms are computationally demanding and were difficult to implement. This last motivation is losing relevance however, with increasing power of calculation and with the increased availability of easy to implement routines in different commercial and noncommercial software. There is also some confusion between the terms *machine learning* and *big data*. Often the two are presented together and this may give the impression that ML techniques are only useful for big data analytics and, in particular, for social media research. However, this is not the case for many applications across different fields of study, as we will show in this paper.

The second reason is more substantive. Some researchers may be skeptical about ML because results provided are often seen as "black boxes", and findings are considered difficult to interpret in substantive terms. Furthermore, differently from the typical specification of a regression model, ML algorithms are not theory-driven. Demographers are often interested in testing specific hypotheses theoretically motivated and may, consequently, dislike algorithms which are of an exploratory nature. However, although (most) ML algorithms are apparently exploratory tools that work automatically without decisive inputs from the researcher, we present here an application that serves to illustrate that they can be useful also in empirical studies motivated by solid theoretical arguments and/or previous literature.

The main goal of this paper is, therefore, to introduce demographers to the basic ideas behind a class of ML techniques, and more specifically to present a popular technique, namely the random forests, of a great potential for application in micro demographic analyses. We discuss the advantages and disadvantages of random forests, that is one of the most popular and powerful ML technique. We argue that random forests can be useful for demographers as a substitute, or a complement, to more standard parametric regression modeling. Differently to standard regression-based approaches, random forests do not impose a parametric model linking an outcome variable of interest to a set of (potentially relevant) independent variables. The key idea is to let the algorithm find the way the outcome and independent variables are linked. In this way, it is possible to automatically search for nonlinearities and interactions among independent variables. Additionally, collinearity and violations of distributional assumptions are not important concerns for random forests. To the best of our knowledge random forests and similar "ensemble" ML techniques have not been previously applied in demographic studies, though afew pioneering studies have applied relatively simpler approaches (De Rose and Pallara 1997; Billari et al. 2006).

We argue that demographers may take advantage from considering the use of ML techniques in their research and we illustrate the application of random forests to divorce.

We illustrate the implementation of random forests in the context of the analysis of the determinants of divorce using data from the SOEP (the German Socio-Economic Panel data), for women entered in a marriage from 1984 to 2015. The ML algorithm is able to classify the determinants of divorce according to their importance, highlighting the most powerful ones and to draw partial dependence plots for the main predictors of survival of the relationship.

Our discussion of the results obtained from this specific analysis of divorce determinants, apart from providing insights about predictors of divorce in Germany, is meant to inform demographers more generally about the relevance of this technique and to illuminate its potential broad usefulness for demographic research. Our application provides an example of implementation of random forests that can be easily applied to different data and different topics using the annotated code that we provide in the Appendix B to reproduce all the analysis presented in this work. We conclude by suggesting areas of research with family demography where this technique promises to be fruitful.

2. ML techniques

2.1. General features of ML

Computer intensive algorithms, belonging to the broad category of ML techniques, have their roots in artificial intelligence (McCarthy and Feigenbaum, 1990). The increased availability of both large datasets and the parallel increase in computer capabilities have opened new possibilities for data analysis and made ML techniques increasingly adopted in many fields (Athey and Imbens 2016; Raghupathy and Ragupathy, 2014).

There are two broad categories of ML techniques, the *supervised* learning and the *unsupervised* one. "Unsupervised learning" focuses on methods for finding patterns in data and for data reduction (Friedman, 1998; Trevor et al, 2009). The well-known and widely used principal components analysis can be included in the unsupervised ML class (see e.g. Bacolod and Rangel (2017) for a recent application). More recently, unsupervised learning algorithms have been developed and applied to problems like clustering or classifying images, videos and text documents into similar groups (see e.g., Vilhena et al. 2014; Athey et al. 2017).

Although unsupervised learning can also be useful for demographers, in this paper we focus on supervised ML techniques. These methods have been employed only marginally in social sciences in general and in demography, in particular. Generally speaking, supervised ML (henceforth, SML) techniques are iterative algorithms for function approximation. These methods focus primarily on prediction problems: given a "training dataset" with data on a certain outcome Y, which could be categorical, discrete or continuous, and some covariates X, the goal is to estimate a model for predicting outcomes in a new dataset (a "test" dataset) as a function of X. Despite the primary objective of these techniques is to build a predictive model, these methods can be fruitfully used to examine how a (potentially large) set of independent variables are linked to an outcome. Therefore, SML techniques can be used as a nonparametric alternative to regression-type approaches commonly employed by demographers when studying relationships between a set of independent variables and a dependent variable.

A distinctive feature of SML algorithms is that the model building is data-driven, so SML can fit complex relations in an automatic way mostly overcoming variable selection and model building efforts. More specifically, SML algorithms can detect automatically non-linearities and non-additivities. These algorithms can be useful to improve data analysis because of their flexibility, in particular when dealing with large (in terms of sample size and number of covariates) datasets.

In the following we give a brief description of some of the most popular SML algorithms that build on “classification and regression trees” (henceforth, CARTs; see Breiman, 1984).

2.2. Classification and regression trees (CARTs)

A classification tree (usually labeled regression tree in the case of a continuous outcome) uses a recursive algorithm to estimate a function describing the relationship between a multivariate set of independent variables and a single dependent variable. At each step the algorithm subsequently splits the data into subsets. These splits are defined by jointly choosing an independent variable and the value that minimizes the prediction error (defined by the sum of squared residuals). Starting with the complete dataset, the algorithm first partitions the dataset into two regions on the basis of the values of a single input variable. For example, if age and sex are the only covariates the tree might split the dataset into two partitions, one with observations with age less than 35 years and the other with observations with age greater than or equal to 35 years. Alternatively, the tree might split the dataset into males and females. Splits can occur between any pair of observed values of any of the covariates. Among all the possible splits, the algorithm selects the one that minimizes the prediction error (see e.g., McCaffrey et al 2009; Breiman et al 1984 for more details).

In the simplest version, the algorithm would stop once the reduction in the sum of squared residuals is below a given (small) threshold. A more refined approach grows the tree until the maximum number of partitions is reached and then it “prunes” the tree back by deleting subtrees that do not decrease much the prediction error. This pruning procedure has the goal of avoiding “overfitting”. Overfitting means that prediction quality is very good in the training set, but not in other samples (e.g., the test set) indicating that the obtained solution holds a strong internal validity, but it is not generalizable to other similar datasets.

A classification tree captures nonlinearities in covariates by splitting them into different intervals. This is similar to the common practices in applied work of capturing nonlinearities in a numerical variable by discretizing it, for example, by dividing it into deciles. The key difference here is that classification trees use the data to determine the appropriate points for discretization, thus potentially capturing the underlying nonlinearities with a more parsimonious and appropriate form. Going back to the previous example, it may be that the split that minimizes the prediction error in the first step is based on age: age less than 35 years versus age greater than, or equal, to 35 years. The tree detected in this way a nonlinearity at age 35 that we may have captured categorizing appropriately age. However, in a standard regression model we would have imposed a-priori a categorization of age, while the tree finds it in the data.

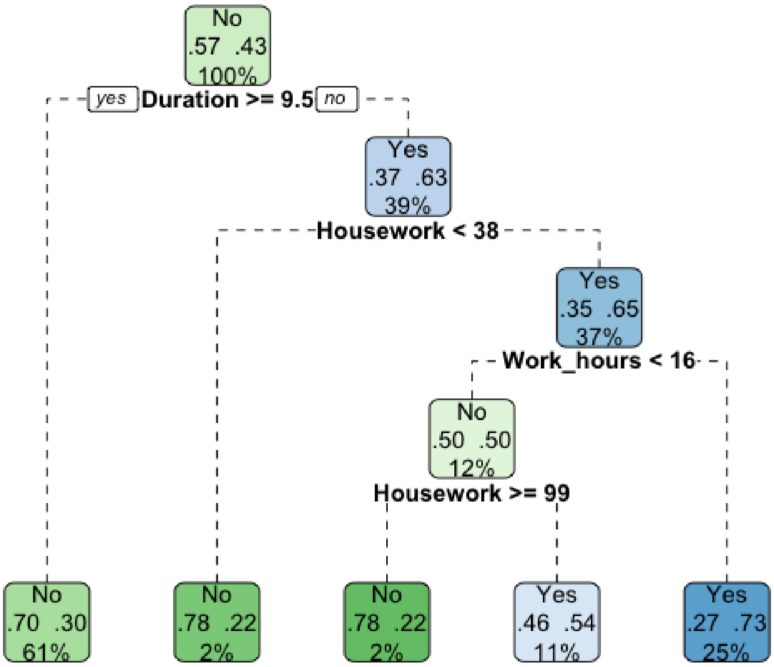
Another general feature of trees is the automatic detection of interactions. In this context, a two-way interaction between two variables X_1 and X_2 is found if a split on one variable, say X_1, in a tree makes a split on X_2 either systematically less likely or more likely. Imagine that at the second split the algorithm detects that splitting the first group (age < 35) between males and females is the split that minimizes predictions errors and suppose for simplicity that the algorithm stops here because it finds impossible to further improve prediction errors (given the prefixed tolerance threshold). The tree would consist of three branches: men younger than 35; women younger than 35; individuals above the age of 35. This means that the tree detected an interaction between age and gender. This would have been easily assessed in a regression framework by interacting the two variables. However, in the presence of many (continuous) covariates, trees may detect two-way interactions (or even more complex interactions), while in a regression framework it would be unfeasible to include all possible interactions terms.

In Figure 1 we provide an example of CART applied to a sub-sample of the dataset we employed for the main analysis in this work. It uses the same set of variables described later in the text to study the probability of separation among couples. Specifically, each node contains the information about the most likely outcome at that specific node (divorce: "No" or "Yes"), the probability of each outcome at that node (left side for "No", right side for "Yes") and the percentage of sample reaching that specific node. Also the colour patterns and their intensity reflect both the most likely outcome (green for "No", blue for "Yes") and its probability at a specific node (the more intense, the higher the probability). Below each node it is indicated the variable and its value that define the split.

According to the chart, and starting from the first split, we see that women who are in a relationship with a duration shorter than 9.5 years are more likely to separate than those having a more long-lasting relationship (63% vs 30%). On the right branch, among women who are in a relationship with a duration shorter than 9.5 years, those who carry out at least 38% of total housework are more likely to separate from their partner with respect to women doing less housework than this threshold (65% vs 22%). Among the former, if we take into consideration also the working hours, the likelihood of separating for women working more than 16 hours per week rises even further up to 73%, while for those working less than this value such probability shifts down to 50%. In particular, for this last group what becomes important for determining the likelihood of separating is a further condition on the percentage of housework, that is whether women are almost doing everything or not (i.e housework>=99%). In the first case the probability of separating is lower than in the second one, possibly signalling a complete separation of the role within the couple in this final node.

Figure 1 illustrates three important characteristics of a CART. First, it automatically finds the relationship between predictors and outcome by categorizing numerical variables. Second, it detects possibly complex nonlinearities in the effect of predictors. For example, the variable housework has been used in two splits with different values (38% and 99%) suggesting the categorization of this variable in three groups: those doing less than 38% of housework, those doing between 38% and 99% and those doing all housework. The figure indicates that women in the two extreme groups have a lower probability of divorce (22%) than in the middle group (54%). Third, the fact that the resulting tree is not symmetric indicates the presence of interactions. For example, the probability of divorce as a function of the variable working hours is modified by the share of housework.

Figure 1: Example of CART for the probability of separation among couples in Germany



Note: The figure shows a classification tree generated by applying the CART algorithm to the probability of separation among couples. The color of each node represents the most likely outcome at that node while its intensity the level of probability associated to it. The percentage at the bottom of each node provides the part of the sample reaching that node.

2.3. Ensemble algorithms and the random forest

More sophisticated algorithms have been developed in the ML literature that build on simple CARTs. These methods, called "ensemble algorithms" are based on (many) multiple trees (Berk, 2006). Ensemble algorithms use random selection features to differentiate the trees and then aggregate the results, the rationale being that averaging over several trees can improve out-of-sample predictions by reducing overfitting. Among the most important ensemble techniques we can mention *bagging*, *boosting* and *random forests*. Bagged Trees are based on a large number of trees, each tree fitting a bootstrapped sample of the data of the same size of the original dataset (Breiman, 1996). A different modification of the CART procedure is based on the idea of boosting. Boosted trees (Friedman, 2001) are based on the idea of incremental fitting: the algorithm is a linear combination of trees where each tree fits the residuals of the previous one using a different subsample of the data. In this paper we demonstrate the implementation of a particular tree-based SML algorithm called random forest (Breiman, 2001). While some rare examples of demographic studies using single decision trees exists (De Rose and Pallara 1997; Billari et al. 2006) we are not aware of previous applications of ensemble ML techniques for demographic analyses.

Random forests are one of the most popular supervised ML methods. Often, they have been found to outperform other SML techniques (Glaeser et al 2006) and are considered among the very best classifiers invented to date (Breiman, 2001). Random forests are known in the ML literature for their reliable “out-of-the-box” performance (i.e., for providing results that can be more easily generalized on different samples) that does not require excessive model tuning (Athey and Imbens 2016), i.e. choices of algorithm arguments by the researcher.

One way to think about random forests is that they are an example of “model averaging.” The prediction obtained with random forests is constructed by averaging over hundreds or thousands of distinct regression trees that differ from one another for several reasons. The name of the algorithm derives, in fact, from the characteristic of a random forest of being a multitude of trees that differ because of random selection of both the data and the variables. Random forests combine "bagging" (i.e., random selection of data) with random selection of variables, an idea introduced first by Ho (1995). For example, if there are ten predictors, before each node is split a random subset of three predictors may be chosen as candidates for defining the split. Then the best split is constructed as usual but selecting only from the three selected covariates. As in bagging, prediction is obtained by majority voting.

Random forests overcome several problems with single decision trees. First, they reduce problems of overfitting by averaging several trees. Second, by using multiple trees, they also reduces the chance of stumbling across a classifier that does not perform well because of the relationship between the train and test data. As for other tree-based approaches, random forests have several advantages over standard regression methods. Random forests can handle automatically (i.e., without need or recoding, grouping, etc.) continuous, nominal, ordinal, and missing independent variables. They can capture non-linear effects and interaction terms. Another important attribute of random forests is their ability to adaptively use a large number of covariates even if most are correlated. In other words, collinearity is not an issue for random forests.

A disadvantage of random forests (as with all ensemble methods) is that by averaging multiple trees they do not offer a single tree to interpret. As a result, there is no way to easily show how inputs are related to the output, as it is the case with single trees. However, as we shall illustrate in the following, several measures can be calculated to ease interpretation. For example, it is possible to record the decrease in the fitting measure (e.g., Gini Index) each time a given variable is used and so build a measure of how much each variable is important in predicting the outcome. In order to reveal how each predictor is related to the outcome one useful solution, is to produce “partial dependence plot” for each covariate (Friedman, 2001; Hastie et al., 2001). Partial dependence plots show the relationship between a given predictor and the response averaged over the joint values of the other predictors as they are represented in a tree structure. As such, the other predictors are being “held constant.” Partial plots generalize to quantitative responses and to responses with more than two categories. We will show their use and how to interpret them in the results section.

Random forests and similar SML techniques have been initially developed and mostly applied for cross-sectional data. However, more recent methodological developments allow using these methods also for survival analyses (Hothorn and Lausen. 2003, 2004, 2006; Ishwaran and Kogalur, 2007). In the application, we use survival random forests to examine the determinants of union dissolution implemented using the Random Survival Forest algorithm (RSF)[1]. The algorithm requires two main parameters to be set before running, that is the number of trees to grow in the forest (i.e. the number of bootstrap repetitions) and the number of variables to randomly select at each split. Specifically, we opt for using one

[1] The package is the “randomForestSRC” in the open source environment R (Ishwaran et al., 2008). Several packages for implementing ML techniques are available in the open source environments Phyton and R. More limited is the availability of packages in commercial software. In STATA, for example, there is a user written package *chaidforest* implements random forests ensemble classifier (Breiman, 2001) using the CHAID (Chi-square automated interaction detection; Kass, 1980) algorithm as the base learner.

thousand trees and five splitting variables (Ishwaran et al., 2008)[2]. To assess the performance of the algorithm in predicting the separation status we calculated the "Out-of-Bag" error rate (OOB) and the concordance index (C-index). Concerning the former, RSF does not need an independent validation dataset to get an unbiased estimate of the test set error, as it is estimated internally during the run of the algorithm. In particular, each tree of the forest is constructed by bootstrapping a sample from the original data and leaving out one-third of the cases, which represents the OOB sample (OOB). Then, each OOB case in the construction of the *k-th* tree is dropped down the tree and the algorithm estimates the percentage of times that the class assigned to each OOB case is not equal to the true class. Finally, the total OOB error is obtained as the average of this estimate across all the trees of the forest. Regarding the C-index, since related to the area under the ROC curve, it can be interpreted as the probability of correctly classifying two cases. Indeed, it estimates the probability that, in a randomly selected pair of cases, the case that fails first had a worst predicted outcome (Ishwaran et al. 2008). Differently from other measures of survival performance, the C-index does not depend on the surviving time, thus making it appealing in order to provide a general evaluation of the RSF performance. Specifically, a value of 0.5 for the C-index is not better than random guessing, whereas a value of 1 denotes full-discriminative ability. In our specific case the OOB error rate is about 35% while the C-index is 0.65, suggesting a certain capacity of the RSF algorithm to predicts the individuals' separation status, although it is not very strong.

3. An illustration of random forests: determinants of divorce in Germany

3.1. A brief literature review of the main determinants of divorce

Before going into the details of our application of random forests for the determinants of divorce, it is useful to briefly review which are these determinants according the existing literature. Naturally, there is a rather extensive literature on divorce determinants in the setting of Western countries, and also specifically in Germany. Recent reviews of the antecedents of union dissolution (e.g., Lyngstad and Jalovaara, 2010) highlight how the knowledge of the determinants of union dissolution has increased considerably over the last decade, which goes in parallel to the broader diffusion of the phenomenon itself. Most of the determinants are common to all Western countries. However, in their meta-analysis review of longitudinal

[2] Our choice of one thousand trees is made in order to minimize the OOB error rate which, as showed in the text, reaches its minimal value around such a threshold. On the contrary, the choice of five splitting variables at each node is obtained following the common practice of using the square root of the number of predictors, that in our specific case is 5.

studies on the divorce risk in Europe, Wagner and Weiss (2006) showed that the variation between contexts in the effect of typical antecedents of divorce is large. Here we refer mainly to the literature addressing the topic in European countries and – when it is possible – specifically to Germany. We have on purpose left out the rather extensive literature on divorce on the United States. One important reason for this is that patterns of divorce (i.e., its determinants) are rather different. For instance, when considering education, the way it links (negatively) with divorce in the US is the inverse of what we usually observe in Europe.

The first group of determinants is composed by the personal characteristics of the members of the couple. The most common ones are:

- *The age of the spouses and age at the marriage or start of cohabitation*. Spouses' current age and duration of the union are collinear and it seems that current age is a better predictor of divorce than age at marriage (Lutz et al. 1991). In general, the effect is that the spouse's maturity and relationship maturity are negatively correlated with divorce (Lyngstad and Jalovaara, 2010).
- *The level of educational attainment of partners.* The effect of the level of education is quite complex (Lyngstad and Jalovaara, 2010). On the one hand the education effect needs to be separated from the effects' of spouses' incomes and labor market activities. Moreover, the effect of education appears to follow a pattern linked to the so-called Goode (1962) hypothesis. That is, the effect of education depends on the prevalence of divorce in the country. More specifically, education brings about higher likelihood of divorce when divorce is not prevalent. The effect of education, however, it changes sign when divorce it is prevalent. This hypothesis is able to explain why the effect of education is always negative in the United States and Scandinavian countries, but remains positive in countries such as Italy, where the divorce prevalence is low, albeit growing (e.g., Harkonen and Dronkers 2006; Matysiak et al. 2014; Salvini and Vignoli 2011). Another element is the interaction of the education of the two partners, where homogamy tends to protect marriages from breaking up (e.g., Kalmijn 1998), although such a pattern does not appear significant in Germany (Wagner and Weiss 2006).
- *The personality of the partners.* Most of the literature on the relationship between personality and divorce relies on the "Big five" personality traits (PTs) score classification (i.e., Agreeableness, Consciousness, Extroversion, Neuroticism and Openness). Recent results show similar association between divorce and PTs in UK, Flanders and Germany. In particular, low score on consciousness and high score on

openness are found to be significant risk factors, although the latter decreasing its importance with time, as divorce becomes more common and less expensive socially and economically (Boertien et al. 2015; Boertien and Mortelmans 2017).

- *The subjective wellbeing of the partners*. Most of the literature based on cross-sectional studies (e.g., Oswald 1997) reported happiness, or overall life satisfaction, to be greater among married people than among the divorced. Other studies have tried to determine the effect of divorce on subjective well-being (e.g., Gardner and Oswald 2006). Only few studies assess the role of partners subjective well-being – and in particular their (dis-) satisfaction with relationship. They find, however, that it is an important predictor of union dissolution especially among women (Rosand et al. 2014).

The second group of determinants of divorce concerns the economic situation of the members of the couple and their (gendered) division of paid and unpaid labor.

- *Women working activities and husbands' unpaid work*: Empirical evidence, across the latter part of the past century, indicated that in many countries employed married women were more likely to divorce than those non employed, suggesting a negative association between women’s employment and marital stability (De Rose 1992; Blossfeld & Müller, 2002; Chan & Halpin, 2002; Cooke, 2006; Jalovaara, 2001; Jalovaara, 2003; Lyngstad & Jalovaara, 2010; Vignoli & Ferro, 2009; Sigle-Rushton 2010; Vignoli et al. 2016; Ozcan and Breen 2012; Poortman 2005). This gave rise to the idea of the so-called “independence effect” arising from married women working (see Cook et al. 2013). The consensus appeared to be that the independence effect dominated the income effect, hence leading to higher divorce. However, a new strand of studies has called into question standard microeconomic predictions of a positive association between women’s economic independence and marital union dissolution suggesting that women’s employment does not have a negative effect *per se*, and that women’s paid work becomes detrimental to the stability of the union only if the men’s contribution to unpaid work is limited (e.g., Mencarini and Vignoli, 2017). In light of this, an important positive determinant of union stability is gender equality within couples, i.e. manifested by the husbands' participation and sharing of domestic chores (e.g., Frisco and Williams 2003; Cooke et al. 2013; Olah and Gahler 2014; Bellani et al. 2017), but also earning equality, especially for cohabiting couples and young marriages (Kalmijn 2007; Ishizuk 2018).

The last group of determinants includes characteristics of the couple (a part from the duration of the union, that we already discussed above):

- *If the couple is married.* The rate of dissolution is generally higher for cohabitants than for married couples, independently from the presence of children (e.g. Andersson 2002, Berrington and Diamond 1999; Liefbroer and Dourleijn 2006), but the difference seems to be explained (at least to some degree) by self-selection into cohabitation or marriage (e.g. Svarer 2004). Furthermore, couples who cohabit and then marry, seem to have mixed effect on stability of the unions (Lyngstad and Jalovaara, 2010).
- *How many children they have.* Couples with children in general have a lower risk of divorce, especially after the first child. However, this lower risk seems in part to be caused by selection, whereby spouses who have little trust in the continuity of their marriage are less likely to have children (Lyngstad and Jalovaara, 2010). More recent literature seems to have repudiated also the hypothesis that sons contribute more than daughters to marital stability (e.g., Diekmann and Schmidheiny 2004).

The factors listed here summarize the more common determinants considered in studies of divorce. Specific studies, using particular surveys, have found other important characteristics linked to union dissolution, such as biological and genetic characteristics of the partners, health conditions of partners and children, migration or minority status, divorce of own parents (i.e. intergenerational transmission of divorce), values and religiosity. However, those aspects cannot be studied with the SOEP data because these information, when available, are such only on small subsamples.

3.2. Data

The sample employed in our analysis is constructed using information provided in the Socio-Economic Panel survey (SOEP), a representative ongoing longitudinal study of the German population, which started in 1984. The SOEP panel survey is well-suited for the study of divorce for two reasons. First, the length of the study allows us to follow individuals over a long period. Second, the SOEP contains all the information necessary for constructing the dependent variable, namely the separation status, and includes information on the main possible determinants of divorce.

Since our attention is focused on married or cohabiting couples, we employ a dyadic approach by selecting from the original database women whose partner is also surveyed in the SOEP. In particular, we include women less or equal than 65 years old and who started their relationship during the observation period (i.e., 1984-2015). Accordingly, those who started their relationship before entering the sample are excluded from the analysis, as those whose partner is not observed within the SOEP or who are still single when leaving the sample. The final sample consists of 18,613 observations, corresponding to 2,038 couples observed, on average, over 12.6 years.

For the dependent variable, by using the information about the identity of the partner within each couple, we construct a dummy variable, *Separation*, which is equal to 1 when we observe a change in the identifying number of the partner from year T to T+1, and 0 otherwise. After the separation we stop following both members of the couple, which means that our sample includes only individuals experiencing one separation, or any, during the observational window. The number of separating couples represents 45% of our sample (i.e. 914), while those who do not split make up 55% of the sample (i.e. 1,125).

The extreme flexibility of the ML approach in handling large set of potentially correlated independent variables allows us to study the impact on the probability of separation determined by several covariates. Specifically, we include twenty-seven explanatory variables (chosen – among those available – according to the literature on determinants of union dissolution) in order to capture several couples' characteristics and different dimensions of their everyday life. These can be classified in three groups. The first one contains the personal characteristic of both members of the couple, among which their personality as measured by the "Big five" personality traits score (Agreeableness, Consciousness, Extroversion, Neuroticism and Openness), their age, their attained level of education (Tertiary or not), their health status and their overall life satisfaction. The second group concerns with the economic situation of the members of the couple, that is the decile of labor income they belong to, their actual number of weekly working hours and their labor force status (unemployed or not). The last group includes all those variables capturing the characteristics of the relationship between the members of the couple, as the percentage of housework carried on by the woman, if it is a married couple and how many children they have. Table 1 lists all these variables and provides some summary statistics.

Table 1: Summary statistics on the independent variables

Variable	**Values**
Personal characteristics	
Age	Avg.=36 Min=17 Max=65
Age (Partner)	Avg.=38 Min=17 Max=65
Tertiary	Dummy: Avg.=0.18
Tertiary (Partner)	Dummy: Avg.=0.22
Overall life satisfaction	Avg.=7.2 Min=1 Max=10
Overall life satisfaction (Partner)	Avg.=7.1 Min=1 Max=10
Health	Avg.=2.39 Min=1 Max=5
Health (Partner)	Avg.=2..38 Min=1 Max=5
Agreeableness	Avg.=5.5 Min=1 Max=7
Consciousness	Avg.=6.1 Min=1 Max=7
Extroversion	Avg.=5 Min=1 Max=7
Neuroticism	Avg.=3.8 Min=1 Max=7
Openness	Avg.=4.6 Min=1 Max=7
Agreeableness (Partner)	Avg.=5.3 Min=1 Max=7
Consciousness (Partner)	Avg.=5.9 Min=1 Max=7
Extroversion (Partner)	Avg.=4.8 Min=1 Max=7
Neuroticism (Partner)	Avg.=4.3 Min=1 Max=7
Openness (Partner)	Avg.=4.4 Min=1 Max=7
Economic situation	
Decile of labor income	Avg.=5.6 Min=1 Max=10
Decile of labor income (Partner)	Avg.=6 Min=1 Max=10
Unemployed	Dummy: Avg.=0.06
Unemployed (Partner)	Dummy: Avg.=0.06
Weekly working hours	Avg.=24.3 Min=0 Max=80
Weekly working hours (Partner)	Avg.=40 Min=0 Max=80
Quality of relationship	
Married	Dummy: Avg.=0.72
Percentage housework	Avg.=72 Min=0 Max=100
N. Children	Avg.=1.28 Min=0 Max=7

Note: Summaries are calculated on the 18,613 observations included in the baseline sample.

3.3. Results

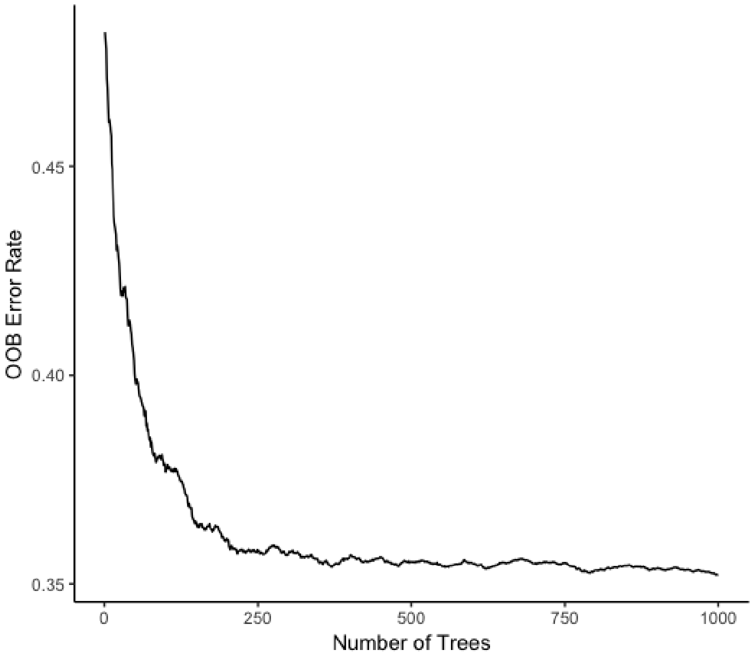
In Figure 2 we plot the value of the OOB error rate according to the number of tree within the forest, and as it can be seen, this latter almost stabilizes around the value of 35% when one thousand trees are employed in the construction of the forest.

The package we use allows us to assess the performance of the RSF (i.e. goodness of fit) at different surviving time. Specifically, we plot in Figure 3 the ROC curve at 1 year, 5 years, 15 years and 25 years. According to that, the performance of the algorithm is decreasing in the length of the time interval with the best discriminative ability recorded at 1 year and the worst one at 25 years.

Figure 4 shows the importance of each variable for the RSF classifier (VIMP). The latter measures the change in misclassification error on the test data if a specific variable is not available, given that the original forest is grown using such a variable. Large importance values are linked to variables with (some) predictive power while zero o negative values characterize those without it. With respect to our specific example, predictors can be roughly classified in four main groups according to their importance in determine the separation status. The first one includes the marriage status and the overall life satisfaction of both members of the couple, which represent the most powerful variables in term of predictive ability. The second group is made by those variables still showing a consistent predictive power, even if lower than the first group, among which there are the couple's age, the percentage of housework, woman's level of consciousness and the partner's level of extroversion. The third group includes those predictors with a limited predictive power, that is all the other personality traits for both members of the couple, their working hours, their level of education and the number of children they have and their health status. Instead, the unemployment status of both members of the couple, as well as their labor income, do not show any predictive ability.

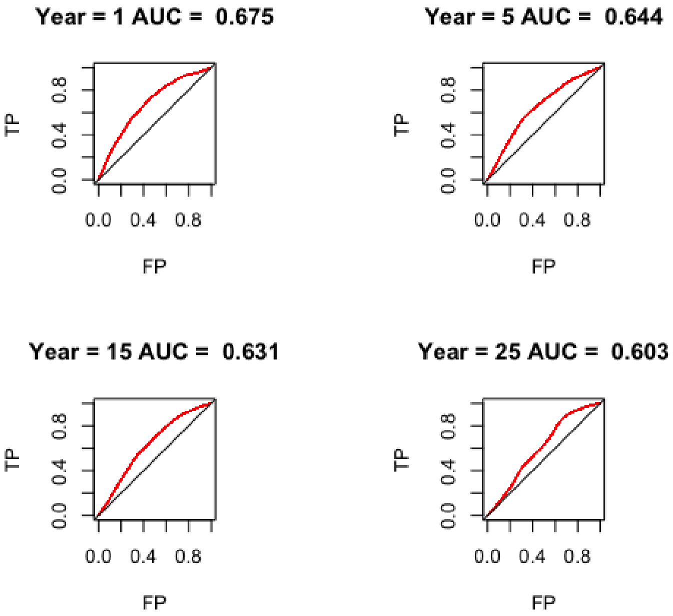
Figure 5 presents the partial dependence plot for the four most important continuous predictors in our model (i.e. overall life satisfaction of both partners, women's housework and men's age), calculated at 1 and 5 years of surviving time, while Figure 6 shows the partial dependence plot for the most important discrete variable, that is the marital status. Each point in both figures represents the average percentage of vote in favor of the "Yes trees" class across all observations, given a fixed level of the predictor. Conceptually, this type of plot is an extension of simple linear regression model parameters to complex models without any parametric specification, providing a graphical representation of the marginal effect of a given variable on the survival. According to Figure 5, the overall life satisfaction of both partner, as well as the men’s age, show a certain degree of non-linearity, especially for the prediction at 5 years, that would be difficult to properly model within a parametric specification, highlighting one possible advantage of the RSF algorithm. Regarding the sign of the effect of each predictor on the separation status, it is mostly increasing for the overall life satisfaction of both members of the couple and the woman’s percentage of housework, while an invert u-shaped effect is detected for the men’s age. Finally, in terms of the size of the effect on survival, the biggest positive ones are registered for the overall life satisfaction of both members of the couple. Concerning marital status, and by looking at Figure 6, we see that married couples are more likely to survive than unmarried couples both at 1 and 5 year, with the biggest difference in the probability between the two groups recorded at 5 year. In the Appendix A we separately provide the partial dependence plots for the entire set of continuous and discrete covariates.

Figure 2: Out of Bag Errors (OOB)



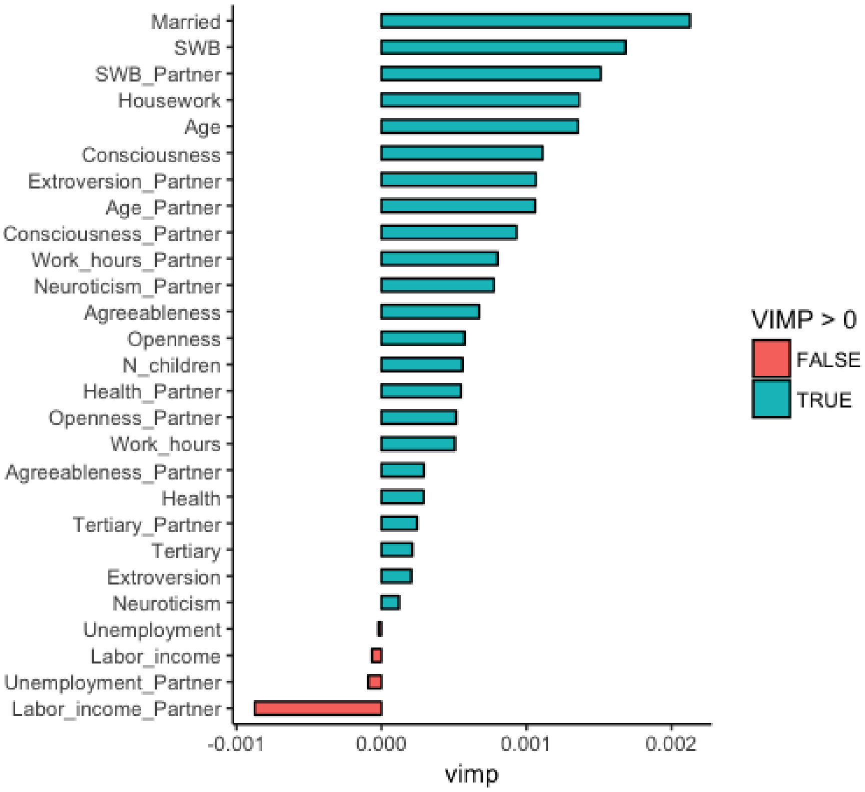
Note: The figure shows the variation of the Out of Bag error (OOB) with respect to the number of trees used in the RSF algorithm.

Figure 3: ROC curves at different surviving time



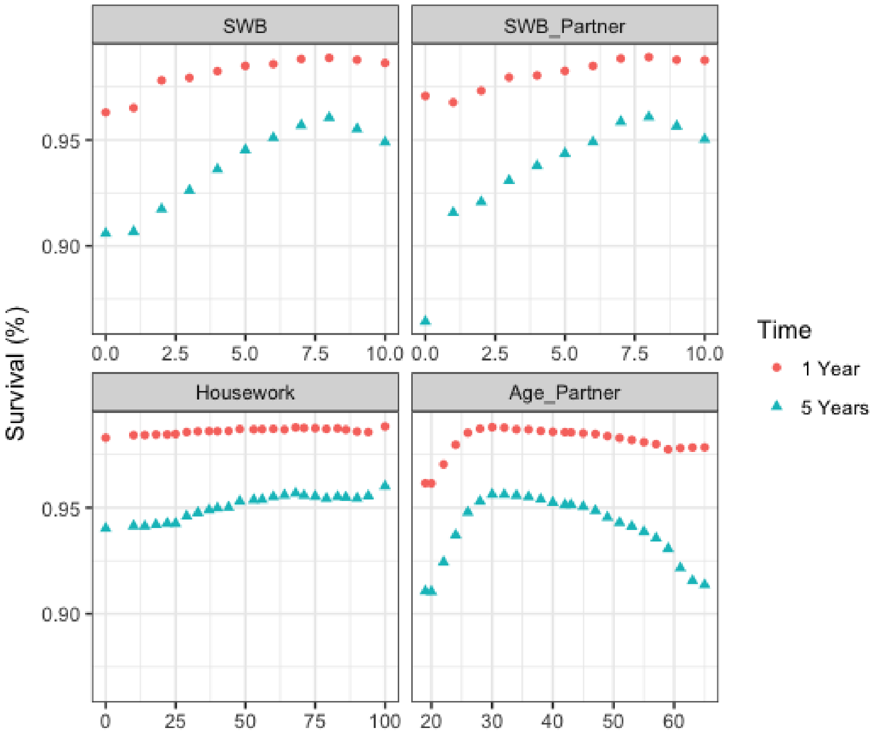
Note: The figure shows the ROC curve at different surviving time (i.e., at 1 year, 5 years, 15 years and 25 years), together with the value of the area under the curve (i.e., AUC).

Figure 4: Variable Importance



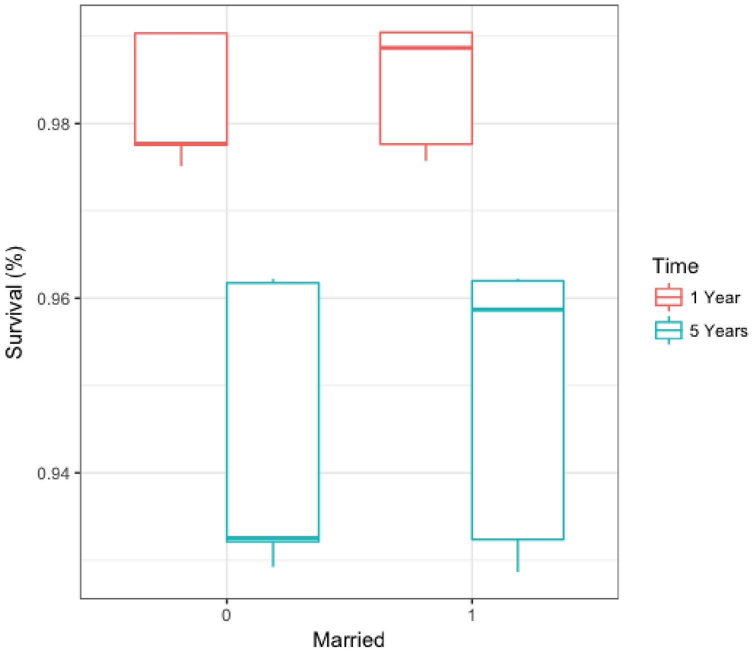
Note: The figure shows the importance measure for each of the 27 variables included as predictor in the model.

Figure 5: Partial Dependence Plots for 4 Continuous Variables



Note: The figure shows the partial dependence plot for the four most important continuous predictors according to Figure 4, at 1 and 5 years. The x-axis shows the distribution of the predictor within our sample while the y-axis provides the predicted survival associated to each value of the predictor.

Figure 6: Partial Dependence Plots for Marital Status



Note: The figure shows the partial dependence plot for marital status at 1 and 5 years. The x-axis shows the distribution of the predictor within our sample while the y-axis provides the predicted survival associated to each value of the predictor.

4. Concluding remarks

The aim of this paper was to discuss the general advantages of ML techniques over standard regression-type approaches for demographic analyses. We demonstrated the use of one of the most popular ML technique, the random forests, using the analysis of union dissolution determinants as a case study. The main advantages of random forests over simple trees is to minimize the risk of arbitrary and ad-hoc model specification, and to take into account, and control, for the risk of model overfitting.

From a substantive point of view, we have shown that random forests ware able to classify the determinants of the divorce according to their importance highlighting the most powerful ones, i.e. both partners' level of life satisfaction, their age, and also some personality traits (specifically the extroversion of the partner and – with less power – also women's conscientiousness, agreeableness and openness). This is an important strength of random forests over a typical regression-type approach because it allows identifying what predictors of the outcome under study are most strongly related to it in a non-parametric way. We also showed that, similarly to marginal effects, we were also able to draw partial dependence plots for the main predictors of survival of the relationship. These partial dependence plots highlight another strength of random forests: it automatically identifies nonlinearities in the effect of the predictors, again non-parametrically. This avoids the risk of mis-specifications and arbitrary categorizations of continuous variables of regression analysis.

Our illustration of the results obtained from the analysis of divorce determinants is of more general relevance for demographic research. Our application provides an example of implementation of random forests that can be easily applied to different data and different topics. We believe that our paper demonstrates the potentiality of random forests in particular and of ML techniques in general for demographic studies. We hope that this application will stimulate the interests of demographers in these techniques.

References

Allison, Paul D. 1984. Event History Analysis. Sage Publications.

Andersson, G. (2002). Children's experience of family disruption and family formation: Evidence from 16 FFS countries. Demographic Research 7(7): 343-364.

Athey S and Imbens GW. The state of applied Econometrics – Causality and Policy Evaluation. ArXiv 2016; 1607.00699v1.

Athey, S., Mobius, M., & Pal, J. (2017). "The Impact of Aggregators on Internet News Consumption." Stanford University Graduate School of Business Research Paper No. 17-8.

Bacolod, M., & Rangel, M. A. (2017). Economic assimilation and skill acquisition: evidence from the occupational sorting of childhood immigrants. Demography, 54(2), 571-602.

Bellani, D., Esping Andersen, G., & Pessin, L. (2017). When equity matters for marital stability: Comparing German and US couples. Journal of Social and Personal Relationships, First Published on-line June 14, 2017.

Berk R. A. (2006). An introduction to ensemble methods for data analysis. Sociological methods & research, 34(3), 263-295.).

Berrington, A. and Diamond, I. (1999). Marital dissolution among the 1958 British birth cohort: The role of cohabitation. Population Studies 53(1): 19-38.

Billari, F. C., Fürnkranz, J., & Prskawetz, A. (2006). Timing, sequencing, and quantum of life course events: A machine learning approach. *European Journal of Population*, 22(1), 37-65.

Blossfeld, H. P., Hamerle, A., & Mayer, K. U. (2014). Event history analysis: Statistical theory and application in the social sciences. Psychology Press.

Blossfeld, H.-P., & Müller, R. (2002). Union disruption in comparative perspective: The role of assortative partner choice and careers of couples. International Journal of Sociology, 32, 3-35.

Boertien, D., Mortelmans, D. (2017), Does the relationship between personality and divorce change over time? A cross-country comparison of marriage cohorts, Acta Sociologica, 1–17, on line first.

Boertien, D., von Scheve, C., Park, Mona (2015) Can Personality Explain the Educational Gradient in Divorce? Evidence From a Nationally Representative Panel Survey, Journal of Family Issues, Journal of Family Issues, Vol 38, Issue 10, pp. 1339 – 1362.

Breiman L, Friedman J, Olshen R and Stone C. (1984). Classification and Regression Trees. Wadsworth Belmont, California.

Breiman L. Bagging predictors, *Machine Learning* 1996; 24(2): 123-140.

Breiman L. Random forests, *Machine Learning* 2001; 45: 5-32.

Chan, T.W. and Halpin, B. (2002). Union Dissolution in the United Kingdom. International Journal of Sociology 32(4): 76-93.

Cooke, L. P., Erola, J., Evertsson, M., Gahler, M., Härkönen, J., Hewitt, B., . . . Trappe, H. (2013). Labor and love: Wives' employment and divorce risk in its socio-political context. Social Politics, 20, 482-509. Cooke, L. P., & Gash, V. (2010). Wives' part-time employment and marital stability in Great Britain, West Germany and the United States. Sociology, 44, 1091-1108.

Cooke, L.P. (2006). "Doing" Gender in Context: Household Bargaining and Risk of Divorce in Germany and the United States. American Journal of Sociology 112(2): 442-472.

De Rose, A. (1992). Socio-economic factors and family size as determinants of marital dissolution in Italy. European Sociological Review, 8, 71-91.

De Rose, A., & Pallara, A. (1997). Survival trees: An alternative non-parametric multivariate technique for life history analysis. *European Journal of Population*, 13(3), 223-241.

Diekmann, A. and Schmidheiny, K. (2004). Do parents of girls have a higher risk of divorce? An Eighteen-Country Study. Journal of Marriage and the Family 66(3): 651-660.

Frisco, M. L., & Williams, K. (2003). Perceived housework equity, marital happiness, and divorce in dual-earner households. Journal of Family Issues, 24, 51-73.

Gardner J, Oswald AJ. Do divorcing couples become happier by breaking up? Journal of the Royal Statistical Society (Series A) 2006; 169; 319-336.

Glaeser Edward L, Andrew Hillis, Scott Duke Kominers, and Michael Luca. Predictive cities crowdsourcing city government: Using tournaments to improve inspection accuracy. The American Economic Review, 106(5):114–118, 2016.).

Goode, W.J. (1962). Marital Satisfaction and Instability: A Cross-Cultural Class Analysis of Divorce Rates. In: Bendix, R. and Lipset, S.M. (eds.). Marital Satisfaction and Instability: A Cross-Cultural Class Analysis of Divorce Rates. New York: The Free Press.

Härkönen J., Dronkers J. (2006). Stability and change in the educational gradient of divorce: A comparison of seventeen countries. European Sociological Review, 22, 501-517.

Hoem, J. M., 1993. Classical Demographic Models of Analysis and Modern Event–history Techniques. Stockholm Research Reports in Demography 75. Stockholm University, Demography Unit, Stockholm.

Jalovaara, M. (2001). Socio-economic status and divorce in first marriages in Finland 1991-

93. Population Studies, 55, 119-133.

Jalovaara, M. (2003). The joint effects of marriage partners’ socioeconomic positions on the risk of divorce. Demography, 40, 67-81.

Kalmijn, M. (1998). Intermarriage and Homogamy: Causes, patterns, trends. Annual Review of Sociology 24(August): 395-421.

Kalmijn, M., Loeve, A., & Manting, D. (2007). Income dynamics in couples and the dissolution of marriage and cohabitation. Demography, 44(1), 159-179.

Hothorn, T. and B. Lausen. 2003. ‘Double-bagging: Combining classifiers by bootstrap aggregation,’ Pattern Recognition, 36: 6, 1303–1309.

Hothorn, T., B. Lausen, A. Benner and Ma. Radespiel-Troeger. 2004. ‘Bagging Survival Trees’. Statistics in Medicine, 23: 1, 77–91.

Hothorn, T., P. Buhlmann, S. Dudoit, A. Molinaro and M.J. van der Laan. 2006. ‘Survival Ensembles’. Biostatistics, 7: 3, 355–373.

Ho, Tin Kam (1995). Random Decision Forests (PDF). Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282.)

Imbens GM, Woolridge JM. Recent developments in the econometrics of program evaluation. National Bureau of Economic Research 2008, *NBER working paper series* no. 14251, 2008.

Ishizuka, P. (2018). The Economic Foundations of Cohabiting Couples’ Union Transitions. Demography, 1-23. First Online: 22 February 2018.

Ishwaran H and Kogalur UB. (2014). *Random Forests for Survival, Regression and Classification (RF-SRC), R package version 1.6,*

Ishwaran H, Kogalur UB, Blackstone EH and Lauer MS., (2008). *Random Survival Forests.* The Annals of Applied Statistics, 2(3), 841–860.

Ishwaran H, Kogalur UB, Chen X and Minn AJ. (2011). *Random Survival Forests for High-*Dimensional Data. Statist. Anal. Data Mining, 4, 115–132.

Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. Statistics in medicine 2010; 29(3): 337-346.

Liefbroer, A. C., & Dourleijn, E. (2006). Unmarried cohabitation and union stability: Testing the role of diffusion using data from 16 European countries. Demography, 43, 203-221.

Lutz, W., Wils, A.B., and Nieminen, M. (1991). The Demographic Dimensions of Divorce - the Case of Finland. Population Studies 45(3): 437-453.

Lyngstad, T. H., & Jalovaara, M. (2010). A review of the antecedents of union dissolution. Demographic Research, 23, 257-292.

Matysiak, A., Styrc, M., & Vignoli, D. (2014). The educational gradient in marital disruption: A meta-analysis of European research findings. Population Studies, 68, 197-215.

McCarthy, J. and Edward Feigenbaum (1990). "In Memoriam Arthur Samuel: Pioneer in Machine Learning". AI Magazine. AAAI. 11(3).

Mencarini, L., Vignoli, D. (2017), “Employed Women and Marital Union Stability: It Helps When Men Help”, Journal of Family Issues. First published online: May-17-2017.

Ng AY, Jordan, MI. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. Advances in Neural Information Processing Systems (NIPS) 2002 (14).

Olah, L. S., & Gahler, M. (2014). Gender equality perceptions, division of paid and unpaid work, and partnership dissolution in Sweden. Social Forces, 93, 571-594.

Oswald, A.J. (1997) Happiness and economic performance. Economic Journal, 107, 1815-1831.

Ozcan, B., & Breen, R. (2012). Marital instability and female labor supply. Annual Review of Sociology, 38, 463-481.

Poortman, A.-R. (2005). Women’s work and divorce: A matter of anticipation? A research note. European Sociological Review, 21, 301-309.

Røsand, G-M.B., Slinning, K., Røysamb, E., Tambs,K. (2014), Relationship dissatisfaction and other risk factors for future relationship dissolution: a population-based study of 18,523couples, Soc Psychiatry Psychiatr Epidemiol (2014) 49:109–119.

Raghupathy W and Ragupathy V. Big Data analytics in health care: promise and potential. Health Information Science and Systems 2014: 2-3.

Ripley BD. Pattern Recognition and Neural Networks. Cambridge University Press. ISBN 978-0-521-71770-0.

Salvini, S., & Vignoli, D. (2011). Things change: Women’s and men’s marital disruption dynamics in Italy during a time of social transformations, 1970–2003. Demographic Research, 24, 145-174.

Sigle-Rushton, W. (2010). Men’s unpaid work and divorce: Reassessing specialization and trade in British families. Feminist Economics, 16, 1-26.

Svarer, M. (2004). Is your love in vain? Another look at premarital cohabitation and divorce. Journal of Human Resources 39(2): 523-535.

Trevor Hastie, Robert Tibshirani, Friedman, Jerome (2009). The Elements of Statistical

Learning: Data mining, Inference, and Prediction. New York: Springer. pp. 485–586

Vignoli, D. and Ferro, I. (2009). Rising marital disruption in Italy and its correlates. Demographic Research 20(4): 11-36.

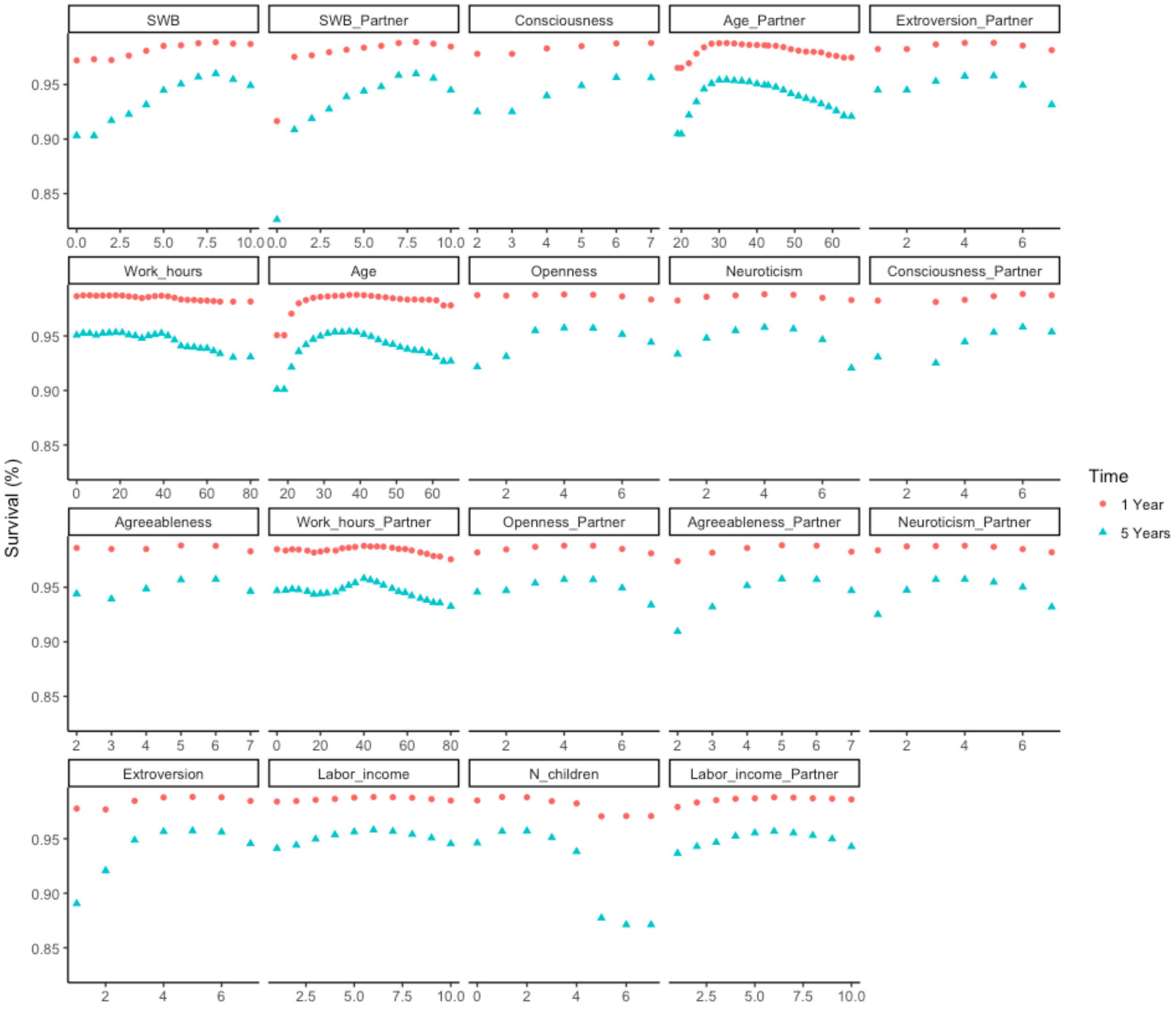
Vignoli, D., Matysiak, A., Styrc, M., & Tocchioni, V. (2016). The impact of women's employment on divorce: Real effect, selection, or anticipation? Families And Societies Working Paper Series, 59(2016).

Wagner, M., Weiß, B. (2006). On the Variation of Divorce Risks in Europe: Findings from a Meta-Analysis of European Longitudinal Studies, European Sociological Review, 22 (5),483–500 483.

Vilhena D, Foster J, Rosvall M, West J, Evans J, Bergstrom C. 2014. Finding cultural holes: how structure and culture diverge in networks of scholarly communication. Sociol. Sci. 1:221–38.

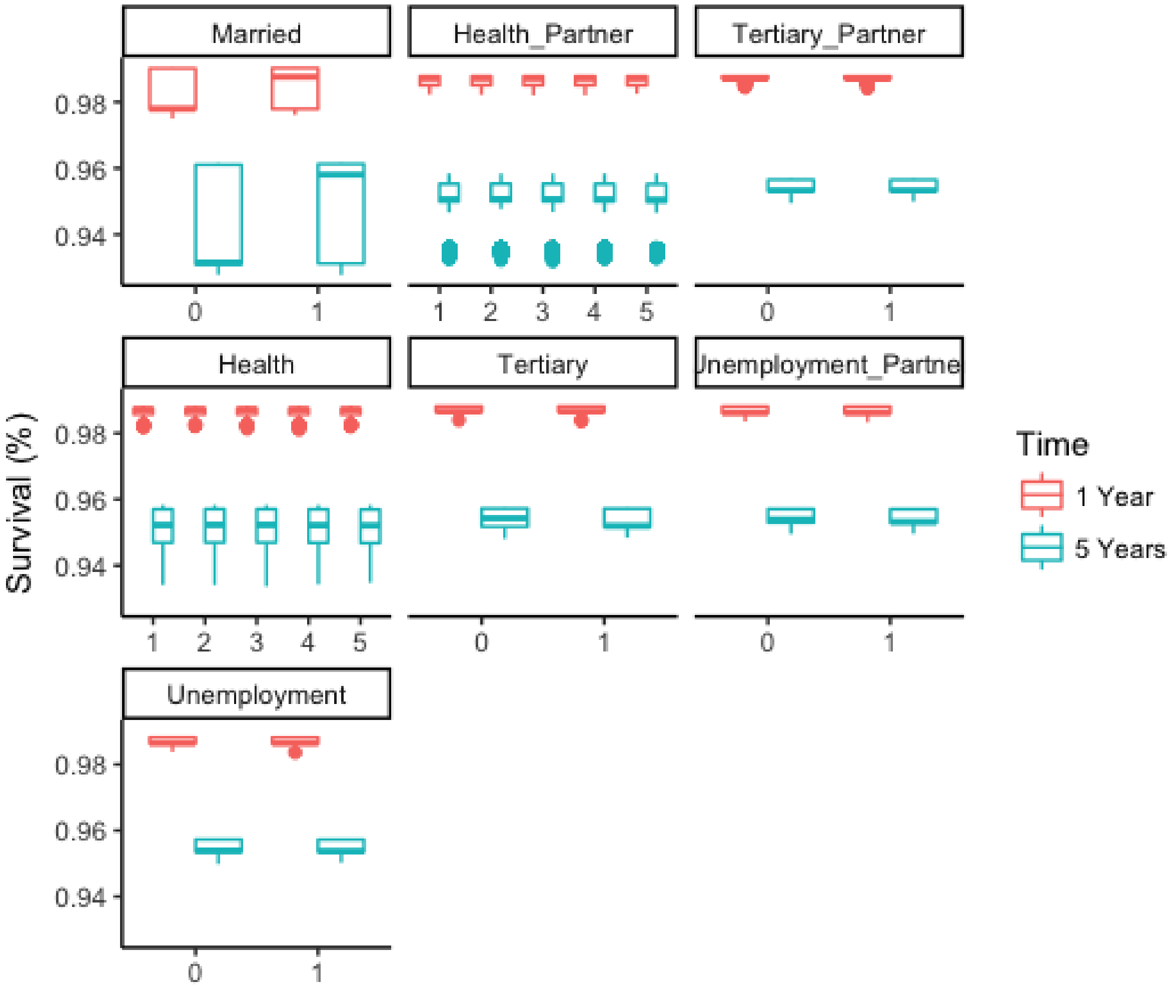
Appendix A: Partial dependence plots for all the covariates

Figure A.1: Partial Dependence Plots for Continuous Variables



Note: The figure shows the partial dependence plot for all the continuous predictors at 1 and 5 years. The x-axis shows the distribution of the predictor within our sample while the y-axis provides the predicted survival associated to each value of the predictor.

Figure A.2: Partial Dependence Plots for Discrete Variables



Note: The figure shows the partial dependence plot for all the discrete predictors at 1 and 5 years. The x-axis shows the distribution of the predictor within our sample while the y-axis provides the predicted survival associated to each value of the predictor.

Appendix B: R-code to implement all the analyses

This appendix provides all the R code we used to generate the results shown in the paper. Lines of code start with "R>".

CART

Figure 1:

```
R> library(rpart)
R> library(rpart.plot)
R> cart<-read.table("YOUR DIRECTORY/CART.csv", header=TRUE, sep=",")
R> set.seed(131)
R> frmla <- factor(cart$Separation, levels = 0:1, labels = c("No", "Yes"))
R> fit = rpart(frmla~ Agreeableness + Consciousness + Extroversion + Neuroticism +
Openness + Agreeableness_Partner + Consciousness_Partner + Extroversion_Partner +
Neuroticism_Partner + Openness_Partner + Tertiary + Tertiary_Partner + N_children + Age +
Age_Partner + Work_hours + Work_hours_Partner + Housework  + Life_satisfaction +
Life_satisfaction _Partner + Unemployment + Unemployment_Partner + Labor_income +
Labor_income_Partner + Health_Partner + Health + Married + Duration, method="class",
data= cart_F)
R> rpart.plot(fit, extra=104, box.palette="GnBu",varlen =0, tweak =0.9, compress=TRUE,
gap=0, space = 0, ycompress=TRUE,  branch.lty=2)
```

RANDOM SURVIVAL FOREST

```
R> library(randomForestSRC)
R> library(ggRandomForests)
R> rsf<-read.table("YOUR DIRECTORY/RSF.csv", header=TRUE, sep=",")
R> set.seed(131)
R> mod_F = rfsrc(Surv(Duration, Separation)~ Agreeableness + Consciousness +
Extroversion + Neuroticism + Openness + Agreeableness_Partner + Consciousness_Partner +
Extroversion_Partner + Neuroticism_Partner + Openness_Partner + Tertiary +
Tertiary_Partner + N_children + Age + Age_Partner + Work_hours + Work_hours_Partner +
Housework  + Life_satisfaction + Life_satisfaction _Partner + Unemployment +
Unemployment_Partner + Labor_income + Labor_income_Partner + Health_Partner + Health
+ Married, data=rsf, ntree= 1000, proximity=TRUE, tree.err=TRUE,  importance=TRUE,
nsplit = 10, na.action = "na.impute")
R> print(mod_F)
R> rcorr.cens(-mod_F$predicted.oob, Surv(rsf$Duration, rsf$Separation))["C Index"]
```

Figure 2:

```
R> plot(gg_error(mod_F)) + theme_classic()
```

Figure 3:

```
R> library(survivalROC)
R> library(survAUC)
R> par(mfrow = c(2, 2), pty = "s")
R> roc <- survivalROC.C(Stime=rsf$Duration, status=rsf$Separation, marker =
mod_F$predicted.oob, predict.time = 1)
R> plot(roc$FP, roc$TP, xlim=c(0,1),  type = "l",  col = "red", ylim=c(0,1), xlab=paste(
"FP"), ylab="TP", main=paste("Year = 1",  "AUC = ", round(roc$AUC,3)))
R> abline(0,1)
R> roc <- survivalROC.C(Stime=rsf$Duration, status=rsf$Separation, marker =
mod_F$predicted.oob, predict.time = 5)
R> plot(roc$FP, roc$TP, xlim=c(0,1),  type = "l",  col = "red", ylim=c(0,1), xlab=paste(
"FP"), ylab="TP", main=paste("Year = 5",  "AUC = ", round(roc$AUC,3)))
R> abline(0,1)
R> roc <- survivalROC.C(Stime=rsf$Duration, status=rsf$Separation, marker =
mod_F$predicted.oob, predict.time = 15)
R> plot(roc$FP, roc$TP, xlim=c(0,1),  type = "l",  col = "red", ylim=c(0,1), xlab=paste(
"FP"), ylab="TP", main=paste("Year = 15",  "AUC = ", round(roc$AUC,3)))
R> abline(0,1)
R> roc <- survivalROC.C(Stime=rsf$Duration, status=rsf$Separation, marker =
mod_F$predicted.oob, predict.time = 25)
R> plot(roc$FP, roc$TP, xlim=c(0,1),   type = "l", col = "red", ylim=c(0,1), xlab=paste(
"FP"), ylab="TP", main=paste("Year = 25", "AUC = ", round(roc$AUC,3)))
R> abline(0,1)
```

Figure 4:

```
R> plot(gg_vimp(mod_F)) + theme(legend.position = c(0.8, 0.2)) + labs(fill = "VIMP > 0") +
theme_classic()
```

Figures 5 and 6:

```
R> xvar <- c("Life_satisfaction", "Age_Partner", " Life_satisfaction _Partner", "Housework")
R> xvar.cat <- c("Married")
R> time_index <- c(which(mod_F$time.interest > 1)[1]-1, which(mod_F$time.interest >
5)[1]-1)
R> xvar2 <- c(xvar, xvar.cat)
R> partial_pbc <- mclapply(mod_F$time.interest[time_index],
function(tm){plot.variable(mod_F, surv.type = "surv", time = tm,  xvar.names = xvar2, partial
= TRUE, show.plots = FALSE) })
R> gg_dta <- mclapply(partial_pbc, gg_partial)
R> pbc_ggpart <- combine.gg_partial(gg_dta[[1]], gg_dta[[2]],  lbls = c("1 Year", "5 Years"))
R> ggpart <- pbc_ggpart
R> ggplot(pbc_ggpart[["Married"]], aes(y=yhat, x=Married, col=group))+
        geom_boxplot(notch = FALSE,
        outlier.shape = NA) + # panel=TRUE,
        labs(x = "Married", y = "Survival (%)", color="Time", shape="Time") +
        theme(legend.position = c(0.1, 0.2)) + theme_classic()
R> ggpart$Married <- NULL
R> plot(ggpart, panel = TRUE) +
        labs(x = "",  y = "Survival (%)", color = "Time", shape = "Time") +
        theme(legend.position = c(0.8, 0.1)) + theme_classic()
```

Figures A.1 and A.2:

```
R>  xvar3 <- c("Agreeableness", "Consciousness", "Extroversion", "Neuroticism",
"Openness", "Agreeableness_Partner", "Consciousness_Partner", "Extroversion_Partner",
"Neuroticism_Partner", "Openness_Partner",  "Age", "Age_Partner", "Work_hours",
"Work_hours_Partner", "Housework ", " Life_satisfaction ", " Life_satisfaction _Partner",
"N_children",  "Labor_income", "Labor_income_Partner")
R>  xvar3.cat <- c("Married", "Tertiary", "Tertiary_Partner", "Unemployment",
"Unemployment_Partner", "Health_Partner", "Health")
```

```
R>  time_index <- c(which(mod_F$time.interest > 1)[1]-1, which(mod_F$time.interest >
5)[1]-1)
R>  partial_pbc <- mclapply(mod_F$time.interest[time_index],
function(tm){plot.variable(mod_F, surv.type = "surv", time = tm, xvar.names = xvar3, partial
= TRUE, show.plots = FALSE) })
R>  gg_dta <- mclapply(partial_pbc, gg_partial)
R>  pbc_ggpart <- combine.gg_partial(gg_dta[[1]], gg_dta[[2]],  lbls = c("1 Year", "5
Years"))
R>  ggpart <- pbc_ggpart
R>   plot(ggpart, panel = TRUE) +
          labs(x = "", y = "Survival (%)", color = "Time", shape = "Time") +
          theme(legend.position = c(0.8, 0.1)) + theme_classic()
R>  partial_pbc <- mclapply(mod_F$time.interest[time_index],
function(tm){plot.variable(mod_F, surv.type = "surv", time = tm, xvar.names = xvar3.cat,
partial = TRUE, show.plots = FALSE) })
R>  gg_dta <- mclapply(partial_pbc, gg_partial)
R>  pbc_ggpart <- combine.gg_partial(gg_dta[[1]], gg_dta[[2]],  lbls = c("1 Year", "5
Years"))
R>  ggpart <- pbc_ggpart
R>   plot(ggpart, panel = TRUE) +
          labs(x = "", y = "Survival (%)", color = "Time", shape = "Time") +
          theme(legend.position = c(0.8, 0.1)) + theme_classic()
```