University of St Andrews

MigrantLife
Working Paper 7 (2022)

Analysing Migrants' Fertility Behaviour Using Machine Learning Techniques: An Application of Random Survival Forest to French Data

Isaure Delaporte and Hill Kulu

This paper is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 834103).

MigrantLife

Understanding Life Trajectories of Immigrants and Their Descendants in Europe and Projecting Future Trends.

Website | Twitter

Analysing Migrants' Fertility Behaviour Using Machine Learning Techniques: An Application of Random Survival Forest to French Data*

Isaure Delaporte† Hill Kulu‡

Abstract

Survival and event history analyses have become widely used techniques in life-course and longitudinal research. Machine learning methods such as survival trees and tree ensembles are a useful alternative to classical methods. This paper aims to illustrate the advantages of random survival forest (RSF). We apply the method to analyse migrant fertility: the probability of having a first, second and third birth among immigrants and their descendants in France. The results of the RSF indicate that even though immigrants have a higher probability of having a birth than natives, highly educated immigrants are much closer to natives in their childbearing patterns than low educated migrants. Our findings illustrate the usefulness of machine leaning techniques in two ways. First, RSF allows us to easily identify the most important predictors of a life event. Second, it allows us to detect and visualize interactions and therefore to identify groups of individuals with different survival probability.

Keywords: Machine Learning, Random Survival Forest, Survival Analysis, Immigrants, Fertility.

* This paper has been prepared within the framework of the MigrantLife project which aims at: "Understanding the Life Trajectories of Immigrants and their Descendants in Europe and Projecting Future Trends". This project is led by Hill Kulu and funded by the European Research Council.
† University of St Andrews, UK. E-mail: icmd1@st-andrews.ac.uk
‡ University of St Andrews, UK. E-mail: hill.kulu@st-andrews.ac.uk

1. Introduction

An important aim in life-course and longitudinal research is to understand the multiple factors that shape people's life outcomes. The most common method to study life events or transitions in the life course of individuals is the technique of survival analysis. Yet, this technique is not without limitations. For instance, survival analysis method cannot be applied directly in high-dimensional settings (Wang and Li 2017; Spooner et al. 2020): as the number of covariates increases, the saturation of statistically insignificant covariates can inhibit effect interpretation (Witten and Tibshirani 2010; Dudoit, Shaffer and Boldrick 2003; Whetten, Stevens and Cann 2021). Collinearity also jeopardizes interpretation of the results. Furthermore, it is difficult to detect and visualize interactions between two or more variables. Lastly, many parametric models require the proportional hazards assumption to hold.

Non-parametric methods such as survival trees and tree ensembles are a useful alternative to classical survival analysis (Breiman et al. 1984; Breiman 2001; Ishwaran et al. 2008; Ishwaran and Kogalur 2008, 2014). To date, only a limited number of studies in demography have used machine learning techniques. De Rose and Pallara (1997) show the usefulness of using a tree methodology to examine the predictors of marriage formation among women in Italy. Billari et al. (2006) also apply decision tree learning and classification rules to detect determinants of the transition to adulthood in Austria and Italy. More recently, Arpino, Le Moglie and Mencarin (2021) depart from the strategy of using single trees and apply random survival forest (RSF) to analyse the determinants of divorce among married and cohabiting women in Germany. Apart from these studies, RSF has been applied so far mostly in bio-medical research (Breiman 2001; Fawagreh, Gaber and Elyan 2014; Ishwaran et al. 2008; Wang and Li 2017; Rezaei et al. 2020; Hsich et al. 2011; Miao et al. 2015; Scheffner et al. 2020; Adham et al. 2017; Hanson et al. 2019; Cafri et al. 2018). Overall, previous studies have stressed the usefulness of machine learning techniques mostly to identify the most important predictors of a specific behaviour; much less attention has been paid to their ability to detect interaction effects.

This paper applies RSF to study the fertility behaviour of immigrants and their descendants. This field has witnessed increasing interest in the demographic life course literature (Kulu and González-Ferrer 2014; Kulu and Hannemann 2016a). The results of previous studies indicate that immigrants exhibit higher fertility levels than natives; they also have children earlier compared to natives; the descendants of immigrants often follow similar patterns to the natives. However, there is considerable heterogeneity within migrant groups and along sociodemographic characteristics. For instance, immigrants' fertility patterns differ by origin and age at arrival (Andersson 2004; Pailhé 2017; Kulu and González-Ferrer 2014; Milewski 2010; Andersson and Scott 2007; Kulu and Hannemann 2016a; Kulu et al. 2017; Delaporte and Kulu 2021). Fertility differences between immigrants and natives are significantly reduced once we focus on highly educated individuals (Pailhé 2017; Krapf and Wolf 2016). Technically, this implies that there are important

interaction effects, which are often difficult to detect using conventional methods. We contribute to the existing literature on the fertility dynamics of immigrants and their descendants by showing how RSF can be used to better detect and understand patterns specific to population subgroups.

To illustrate the application of RSF, we use a rich survey from France named Trajectories and Origins. This survey collects detailed information on immigrants, immigrants' descendants, and French natives. It contains retrospective biographical data on individuals' childbearing histories as well as detailed information on their sociodemographic characteristics. We examine the likelihood of having a first, second, and third birth. We will first compare the predictive performance of the Cox proportional hazards model with RSF. We will then examine which variables are the most important determinants of a first or subsequent birth. Lastly, we will examine possible interaction effects to detect subgroup-specific patterns. Our findings illustrate the competitive performance of the RSF algorithm in two ways. First, RSF allows us to identify the most important predictors of a life event. Second, it allows us to detect and visualize interactions and therefore identify groups of individuals with different survival probability.

The rest of the paper proceeds as follows. Section 2 presents the main functions of survival analysis, whereas Section 3 introduces the ensemble learning methods with a focus on random survival forest. Section 4 presents the application of RSF to the analysis of migrant fertility behaviour. Lastly, Section 5 concludes on the advantages and potential pitfalls of the method.

2. Survival Analysis

Basic functions

Let T be a continuous random variable to represent the duration of an episode, the waiting time, until an event occurs (Cleves et al. 2010). The hazard function, $h(t)$, is defined as follows:

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t+\Delta t, T \geq t)}{\Delta t}, k = 1,2, \dots, K \quad (1)$$

The numerator of the formula is the probability that an event occurs for a randomly selected individual in the time interval from t to $t + \Delta t$ given that they have not experienced an event before. The denominator includes the length of the interval. The survivor function, $S(t)$, is defined as follows:

$$S(t) = \Pr(T \geq t) \quad (2)$$

The survivor function of T represents the probability that the episode's duration is at least t. The survivor function thus measures the likelihood of 'surviving', i.e., not experiencing an event up to the time point t. If we know the hazard function, then we can calculate the value of the survivor function at t by integrating the hazard function from 0 to t:

$$S(t) = exp\left\{- \int_0^t h(\tau) d\tau\right\} \quad (3)$$

The cumulative hazard function, $H(t)$, is another function often used in survival analysis. It measures the total amount of hazard that has been accumulated up to time t.

$$H(t) = \int_0^t h(\tau)d\tau \tag{4}$$

In non-parametric analysis, the value of the cumulative distribution function rather than the hazard function is often calculated at duration t. This is because the length of the intervals used to estimate the hazard at various durations vary in empirical applications. Therefore, the values of the hazard function are erratic, and the pattern identification is difficult. It follows from 3 and 4 that the survivor function can easily be calculated using the cumulative hazard function:

$$S(t) = exp\{-H(t)\} \tag{5}$$

Maximum likelihood estimation

Parameters of a survival model are estimated using the maximum likelihood method (Hosmer and Lemeshow 1999). The first step is to construct a likelihood function. For episode i ending with an event, E, we calculate the probability of having an event at the observed time point, t_i:

$$f(t_i) = \Pr(T_i = t_i) \tag{6}$$

where $f(t)$ denotes the probability density function. For episode i ending with right-censoring, Z, we calculate the probability of surviving until time point, t_i:

$$S(t_i) = \Pr(T_i \geq t_i) \tag{7}$$

The likelihood function for all episodes, N, is the joint probability:

$$L(t_i, \dots, t_n) = \prod_{i \in E} f(t_i) \prod_{i \in Z} S(t_i) \tag{8}$$

Given that $f(t) = S(t)h(t)$, the likelihood function may be written:

$$L(t_i, \dots, t_n) = \prod_{i \in E} h(t_i) \prod_{i \in N} G(t_i) \tag{9}$$

We now take the natural logarithm of both sides to obtain the log-likelihood function for censored survival data:

$$l(t_i, \dots, t_n) = \sum_{i \in E} \log h\,(t_i) + \sum_{i \in N} \log S(t_i) \tag{10}$$

The final step is to maximise the log-likelihood function. This involves taking the derivative of the log-likelihood function with respect to unknown parameters (e.g., the hazard rate for a model without covariates), setting the derivate of the function equal to zero and solving the resulting equation, and conduct the second derivative test. Standard software uses a numerical iteration procedure to maximise the log-likelihood function (Hosmer and Lemeshow 1999).

3. Ensemble Learning Methods

The main objective of supervised machine learning techniques is to estimate a model for predicting an outcome as a function of covariates. The outcome can be categorical, discrete, or continuous (Loh 2011). There are many different supervised machine learning models; among them, we have the ensemble learning methods. These methods aim to obtain so called "trees" by recursively partitioning a data set into subsets called nodes. The final regions created on termination of the growth of a tree are known as terminal nodes, or leafs, while the initial node is commonly referred to as the root node. These trees aim to classify individuals into classes based on the values taken by their covariates. Each class has a specific outcome. Among these ensemble learning methods, one technique used for classification and regression is random forest (Liaw and Wiener 2002; Fawagreh, Gaber and Elyan 2014).

Random forest

Developed by Breiman (2001), random forest combines Breiman's bagging sampling approach[1] and the random selection of features[2], introduced by Ho (1995, 1998) and Amit and German (1997) to construct a collection of decision trees. In practice, random forest runs the analysis over many sub-datasets made by randomly selecting features (Breiman 2001). The prediction is obtained by averaging over hundreds or thousands of distinct regression trees that differ from one another in the sense that the correlation between trees is low (Taylor 2011). This allows the reduction of overfitting issues and the mitigation of the instability of regression trees (Jiang 2019). Random forest can be used with a large number of variables, even if they are highly correlated with each other. Besides, both numeric and categorical variables can be included. Lastly, the method allows us to identify interactions and non-linear associations. Random forests are one of the most popular supervised machine learning techniques. Over the last two decades, many applications of random forest were developed across different disciplines (Fawagreh, Gaber and Elyan 2014).

[1] The bagging sampling approach is an important characteristic that allows the reduction of overfitting issues. As each split of a node is dependent on previous partitioning, a single tree can be unstable. Therefore, the sensitivity of a single tree to minor training data variations is likely to result in poor generalization to new data. Introducing bootstrapping consists of having individual trees that are grown for multiple bootstrap samples. These trees are subsequently aggregated instead of producing a single ensemble tree.

[2] The term "features" refer to variables. The random selection of variables allows for the selection of less strong predictors as splitting variable. This could lead to the inclusion of relevant interaction effects that would be missed otherwise in standard bagging procedure. The random selection of features also ensures that the individual trees in the forest differ from each other.

Random survival forest

Random forests can be extended to right-censored survival or time-to-event data with RSF (Ishwaran 2007, Ishwaran et al. 2008).[3] In RSF, the outcome is an ensemble cumulative hazard estimate which is calculated over all trees in the forest (Ziegler and König 2014). As illustrated in Figure A.1 in the Appendix, the application of RSF involves the following principles: a) survival trees are grown using bootstrapped data; b) random feature (or variable) selection is used when splitting tree nodes; c) trees are generally grown deeply; and d) the survival forest ensemble is calculated by averaging tree survival predictors (Ishwaran et al. 2019; Wang, Li and Reddy 2019). Each step of the algorithm involves defining specific parameters (see Figure 1) which we discuss in the results section.

Survival trees are binary trees grown by recursive splitting of tree nodes. A tree is grown starting at the root node, which is the top of the tree comprising all the data. Using a predetermined survival criterion, the root node 1 is split into two nodes: 2 and 3. Each new node, in turn, is split into two further nodes, 2 into 21 and 22 and 3 into 31 and 32. The process is repeated for each subsequent node. An optimal split for a node maximises survival difference between the two new nodes. There are several methods to determine the difference between the nodes; the log-rank test, often used in survival analysis to compare survival of two groups, is a commonly used method (Ishwaran et al. 2008). Eventually, the survival tree reaches a saturation point when no new nodes can be formed. The ends of the tree are called the terminal nodes.

Let r be a terminal node of the tree. There are k points in time where at least one of the episodes ends with an event:

$$0 < t_{1,r} < t_{2,r} < t_{3,r} < \cdots < t_{k,r}$$

For each node, the cumulative hazard function (CHF, see formula 4) is calculated using the Nelson-Aalen estimator (Ishwaran et al. 2008; 2009):

$$H(t_r) = \sum_{t_{k,r} \leq t} \frac{E_{k,r}}{N_{k,r}} \tag{11}$$

Where $E_{k,r}$ denotes the number of events in node r at t_k and N is the number of episodes (or individuals) in the risk set in r at t_k. The survivor function is estimated using the Kaplan-Meier estimator:

$$S(t_r) = \prod_{t_{k,r} \leq t} \left(1 - \frac{E_{k,r}}{N_{k,r}}\right) \tag{12}$$

[3] Within the framework of random survival forest, a number of extensions have also been developed (Wang and Li 2017) such as random survival forest to competing risks (Frydman and Matuszyk 2020; Ishwaran et al. 2014; Keramati et al. 2020; Hamidi 2017; Wang, Li and Reddy 2019).

All episodes (or individuals) within r have the same CHF and the survivor function. This is because the survival tree has partitioned the data into homogeneous groups (i.e., terminal nodes) of individuals with similar survival behaviour. If we wish to estimate $H(t|X)$ and $S(t|X)$ for a given feature (or variable) X, we drop X down the tree. Because of the binary nature of a tree, X will fall into a unique terminal node r. The CHF and survival estimator for X's terminal node are then (see Ishwaran et al. 2019):

$$H(t|X) = H_r, \qquad \mathrm{S}(t|X) = S_r, \qquad \text{if } X \in r, \tag{13}$$

The ensemble CHF and survivor function are calculated by averaging the tree estimator (Ishwaran et al. 2008; 2019):

$$\bar{H}(t|X) = \frac{1}{N}\sum_{n=1}^{N} H_n(t|X), \qquad \bar{S}(t|X) = \frac{1}{N}\sum_{n=1}^{N} S_n(t|X) \tag{14}$$

where H_n is the nth survival tree with N trees.

4. Application: Immigrant Fertility Behaviour in France

We use the RSF technique to analyse the fertility behaviour of immigrants and their descendants in France.

4.1. Data

This analysis uses Trajectories and Origins, a rich retrospective French survey collected in 2008. Information is collected on immigrants, immigrants' descendants, and French natives. The sample consists of 20,380 individuals including 8,259 immigrants, 8,614 descendants of immigrants and 3,507 French natives. The survey provides retrospective childbearing histories for all individuals on a monthly time scale. For the purpose of this study, we examine three outcomes: the event of having a first birth, a second birth and a third birth. If individuals have not experienced an event, they are right-censored.

We include in our models the following time-constant variables: gender, origin group, birth cohort, size of the family of origin, education, and religiosity. The origin group distinguishes immigrants and their descendants from North Africa, Sub-Saharan Africa, South East Asia, Turkey, Southern Europe and other Europe. In some models, we include a variable which only distinguishes immigrants, their descendants, and natives. The birth cohorts are: 1948-1959, 1960-1969, 1970-1979 and 1980-1989. We exclude individuals who were born in 1990-1999 since they were too young at the time of interview. The family size refers to the respondent's number of siblings. The educational levels are low (no qualification or primary education), middle (lower- and higher-secondary education) and high (two years or more in higher education). Lastly, religiosity is a dummy variable equal to 1 if the respondent reports that religion was important in his/her upbringing, 0 if otherwise. We use the RSF algorithm as implemented in "randomForestSRC" package in

R.[4] An analysis of the fertility behaviour of immigrants and their descendants in France provides an ideal case study to illustrate the advantages of using RSF. We discuss the existing literature to explain further how the use of RSF can improve our understanding of determinants of migrant fertility.

4.2. Understanding Immigrant Fertility Behaviour

Previous studies show that immigrants tend to have higher fertility levels than the natives (Kulu and González-Ferrer 2014; Kulu et al. 2019) and they start childbearing at a younger age compared to natives (Rojas, Bernardi and Schmid 2018); the descendants of immigrants are more similar to natives in their fertility behaviour. This pattern has been observed in many European countries. For instance, Andersson and Scott (2007) report that immigrants from high fertility countries in Sweden have significantly higher second- and third-birth levels than Swedish-born women. Similarly, Milewski (2010) show that second and third-birth levels were relatively high for immigrant women from Turkey in West Germany. Mussino and Strozza (2012) draw similar conclusions from the analysis of second-birth rates among immigrants in Italy.

However, there is considerable heterogeneity among migrant populations. For instance, immigrants' and natives' fertility differ at different ages (Wilson 2020). Fertility levels are also different by origin groups and generations (Kulu and Hannemann 2016b; Pailhé 2017; Mussino and Strozza 2012; Andersson and Scott 2007; Delaporte and Kulu 2021). For instance, in France, women of Southeast Asian origin deviate from the fertility pattern of their parents, while those of Turkish descent exhibit fertility patterns similar to those of their parents (Pailhé 2017). Similarly, in the UK, immigrants from Pakistan and Bangladesh have higher first-birth risks than the natives, whereas European and other immigrants have lower first-birth levels (Kulu and Hannemann 2016b). In Italy, immigrants from North Africa had significantly higher fertility levels than those who came from Eastern European countries (Mussino and Strozza 2012).

Fertility differentials between immigrants and natives may also vanish as the sociodemographic structure of an immigrant group grows to resemble that of the native population (Milewski 2007, 2010). Indeed, there are significant differences between immigrants from lower and higher socio-economic backgrounds (Krapf and Wolf 2016). Besides, access to a higher level of education is a crucial factor in reducing the differences between the groups (Pailhé 2017). Further, it is likely that education and employment-related factors play a key role in shaping the fertility behaviour of the descendants of immigrants. Successful structural integration suggests that high educational aspirations may lead to a significant postponement of family formation and smaller family size among ethnic minority women. In

[4] A useful document to get to know the package on R is Ehrlinger (2016).

contrast, poor employment prospects among some ethnic minority groups due to lower levels of education and/or discrimination in the labour market may promote high completed fertility.

These differences within migrant and descendant groups suggest the presence of so-called interaction effects and the need to explore the role of different sociodemographic characteristics in reducing or deepening fertility differences between immigrants and natives. Yet, conventional survival analysis methods are not best suited to detect interaction effects, especially if more than two variables are involved. The next section highlights the advantages of using RSF to determine the most important predictors of childbearing events and to identify important interaction effects.

4.3. Results

Growing the forest in R

We grow the forest for each outcome of interest – the event of having a first, second and third birth. The RSF algorithm consists of four main steps. First, a fixed number of bootstrap samples are drawn randomly from the dataset.[5] This number of bootstrap samples leads to an equal number of trees. In our case, we opt for the default number of trees: *ntree* = 1000.[6] Several options are also available for the bootstrap. By default, about 63% of the original observations will occur one or more times in the bootstrap sample and 37% of the original data will not occur in the bootstrap sample. These observations are said to be out-of-bag (OOB) with respect to the bootstrap sample.

The second step of the algorithm involves the splitting of the nodes. For each bootstrap sample, a survival tree is built by randomly selecting features. The number of candidate variables is specified through *mtry*. We used the default setting where *mtry* is equal to the square root of the total number of features. This gives us *mtry* = 3. The number of split points considered for each variable is also given by *nsplit*. We use *nsplit* = 10. The node is split by using the candidate feature that maximizes the difference in survival between child nodes. The difference in survival is evaluated by a log-rank splitting rule.[7] We use the default option: the random splitting rule. The third step of the algorithm consists in growing the tree to the full size under the constraint that each leaf node contains no less than a single unique event. Additional constraints can be imposed by increasing the minimal size of the terminal nodes (*nodesize*) and by limiting the total

[5] We do not split our dataset into a training and a test set. Indeed, 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.

[6] We demonstrate later on that our results are robust to a different number of trees.

[7] A splitting rule needs to be defined, which can be *logrank*, *logrankscore* or *random*. The former two rely on the log-rank-score statistic, respectively, both of which quantify the difference in survival curves between two groups in this context between the two daughter nodes for a potential split point. When *random* is specified as the splitting rule, a single split point is randomly chosen on each variable, and the largest log-rank statistic decides the choice of split.

number of splits (*nodedepth*). By default, in survival, *nodesize* = 15. Lastly, using the non-parametric Nelson-Aalen estimator, the ensemble cumulative hazard function (CHF) of OOB data is calculated by taking the average of the cumulative hazard function of each tree (Wang, Li and Reddy 2019). We can then calculate the prediction error for the ensemble CHF using only the OOB data.

4.3.1. First Birth

Our first outcome of interest is the likelihood of having a first birth among the French population. We first examine the predictive performance of the model.

Predictive performance

To assess the performance of the algorithm in predicting the birth event, we calculate the "Out-of-Bag" error rate (OOB) and the concordance index (c-index). The OOB error rate is obtained as follows. First, 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. The algorithm then estimates the percentage of times that the outcome assigned to each OOB case is not equal to the true outcome. Finally, the total OOB error rate is obtained as the average of this estimate across all the trees of the forest.

The c-index is another measure that allows us to assess the performance of the algorithm. It can be interpreted as the probability of correctly classifying two cases as it is related to the area under the receiver operating characteristic (ROC) curve. More specifically, it estimates the probability that, in a randomly selected pair of cases, the case that fails first had a worst predicted outcome. The c-index differs from other measures of survival performance as this measure does not depend on the survival time. Therefore, the c-index provides a general evaluation of the performance: a value of 0.5 is not better than random guessing, whereas a value of 1 denotes full-discriminative ability.[8]

We obtain an OOB error rate of 36% while the c-index is 0.65, suggesting that the RSF algorithm does a good job in predicting individuals' parenthood status. Figure A.2 in the Appendix plots the value of the OOB error rate according to the number of trees within the forest. We can see that the OOB error rate stabilizes around the value of 36%. The figure also demonstrates that a similar OOB error rate is reached when specifying a smaller number of trees than 1000. We also assess the performance of the RSF (i.e., goodness of fit) at different survival times. Specifically, we plot in Figure 1 the ROC curve at four points in time: at the individuals' ages of 20, 30, 40 and 50. FP on the x-axis refers to "False Positive" and TP on the y-axis refers to "True Positive". The Area Under the Curve (AUC) tells us how well we can classify

[8] The literature so far does not provide more guidance on whether these numbers can be interpreted as rather high or low.

individuals in two groups: those who experience the outcome of interest and those who do not. AUC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0 while one whose predictions are 100% correct has an AUC of 1. According to the results displayed in the figure, the algorithm has a good discriminative ability over the life course. The performance of the algorithm is the best at the age of 20.

< Figure 1 here >

It is also possible to compare the predictive performance of the RSF with conventional survival analysis methods such as the Cox proportional hazards regression model (CPH). We examine each model's concordance index (c-index) over time. The c-index gives the probability of concordance between the predicted and the observed survival. Figure A.3 in the Appendix shows that the RSF outperforms CPH when it comes to predicting the event of having a first birth throughout the life course.

Variable selection

Next, we examine the importance of the variables to predict the likelihood of having a first child. We use both Variable Importance (VIMP) (Breiman 2001) and Minimal Depth (Ishwaran et al. 2010; Ishwaran et al. 2011) methods and compare the results (Figure 2).[9] VIMP for a variable is the difference between the prediction error when the variable is randomly permuted compared to the prediction error under the observed values. Therefore, a large VIMP value indicates that misspecification detracts from the predictive accuracy in the forest. A VIMP close to zero indicates that the variable contributes nothing to predictive accuracy, and negative values indicate the predictive accuracy improves when the variable is misspecified. Therefore, we ignore variables with negative and near zero values of VIMP and rely on the variables with large positive values. Using the VIMP method, we find that the most important feature to predict the event of having a first birth is the level of education of the individual, followed by gender and birth cohort (Figure 2 panel a).

An alternative method to identify the most important predictors is to use minimal depth (Ishwaran 2007; Ishwaran et al. 2010, 2011). This method assumes that variables with high impact on the prediction are those that most frequently split nodes nearest to the root node, where they partition the largest samples of the population. These variables have smaller minimal depth values. Using minimal depth, the results show that the most important features are now the educational level of the individual, gender, and the family size (Figure 2 panel b). Since the VIMP and Minimal Depth measures use different criteria, it is not surprising that the variable ranking tends to be somewhat different. We compare the rankings between minimal depth and VIMP (Figure 2 panel c). The points along the red dashed line indicate where the

[9] We report the results obtained when including migrant generation rather than migrant origin. Indeed, we do not include both variables in the same model since the two variables refer to similar characteristics.

measures are in agreement. Points above the red dashed line are ranked higher by VIMP than by minimal depth, indicating that the variables are more sensitive to misspecification. Those below the line have a higher minimal depth ranking, indicating they are better at dividing large portions of the population. The further the points are from the line, the larger the discrepancy between measures. Our results indicate that both measures are in agreement about the importance of the education level of individuals to predict the event of having a first birth.

< Figure 2 here >

Response dependency

Once we have identified the most important predictors, we can examine how these variables are related to the likelihood of having an event. We focus on the birth cohort, the educational level and origin group. We can either examine variable dependence plots or partial dependence plots. While variable dependence plots show the predicted response relative to a covariate of interest, partial dependence plots give us an adjusted variable dependence. In other words, partial dependence plots are generated by integrating out the effects of variables besides the covariate of interest. We choose to report partial dependence plots. We start with a two-way interaction and examine the differences in fertility behaviour between immigrants, the descendants, and natives at two points in time: by the ages of 30 and 45. This allows us to check if the proportionality assumption holds.

The results (not reported but available upon request) show that natives have a slightly higher probability of having a first birth by the age of 30 compared to immigrants and the descendants, whereas immigrants are more likely, although only marginally, to have a first birth by the age of 45 compared to the descendants and natives. As expected, Turkish and Southern European immigrants are the most likely to have a first birth by the age of 30 compared to other groups. Similarly, among the descendants of immigrants, the children of Turkish and Southern European immigrants have the highest probability of a first birth. The differences in the predicted probabilities between groups are reduced when we examine the probabilities by the age of 45. The results thus show some group differences in the share of childlessness and also indicate differences in the timing of childbearing.

Interactions

We now examine whether and how the relationship between migrant generation and fertility changes by birth cohort (Figure 3). The results show that the differences in the likelihood of having a child between migrants and natives are stable across birth cohorts. Overall, individuals in more recent cohorts have a lower probability of having a first birth, especially by age 30 suggesting the postponement of parenthood. The analysis supports that the French natives are slightly more likely to have a first birth compared to

immigrants and their descendants by the age of 30. By the age of 45, however, immigrants are slightly more likely to have a first birth.

< Figure 3 here >

Next, we examine a four-way interaction by also including the educational level of individuals (Figure 4).[10][11] Interestingly, if we focus on the likelihood of having a first birth by the age of 30, the patterns do not differ considerably between immigrants and natives among low educated individuals; only their descendants have slightly lower first-birth levels. By contrast, for highly educated individuals, both immigrants and the descendants are less likely to have a first birth compared to natives. This pattern remains similar across birth cohorts, although the predicted probabilities are lower for all population groups among more recent cohorts. If we examine the likelihood of having a child by the age of 45, we see that among low educated individuals, immigrants are more likely to have a first birth compared to the descendants and natives. By contrast, among highly educated individuals, immigrants' fertility differentials are reduced. Highly educated immigrants are much more similar to natives in their likelihood of having a child by age 45 than low educated immigrants.

< Figure 4 here >

4.3.2. Second Birth

Next, we examine the probability of having a second birth among parents. As previously, we examine the effect of a number of time-constant variables on the probability that the event occurs. First, we explore variable importance using VIMP and minimal depth (See Figure A.5 for the results using VIMP and minimal depth). Figure 5 shows that migrant generation is the most important variable to predict whether individuals have a second birth; education was the most important variable for having a first birth.

< Figure 5 here >

We then examine how the likelihood of having a second child differs across migrant generations and birth cohorts 5 and 10 years after the first birth (Figure A.6 in Appendix). We find no significant differences across birth cohorts: immigrants are more likely to experience a second birth than the descendants or natives and for all birth cohorts. Furthermore, the difference in the probabilities between immigrants and natives is stable over time since the first birth. This result supports that birth cohort is not important to predict the event of a second birth.

[10] Additionally, Figure A.4 in the Appendix also presents the predicted probabilities of having a first birth by age 30 and age 45, by migrant origin, birth cohort and educational level.

[11] We also examine whether these patterns are similar for men and women. The results (not reported but available upon request) show that the patterns are overall similar for men and women, except that men overall have lower probabilities of a first birth than women, especially by the age of 30. Second, among low educated individuals, immigrants' and descendants' fertility differentials are larger among men than among women.

Lastly, we examine the probability of having a second birth by migrant generation, birth cohort and educational level (Figure 6).[12][13] Among individuals who have low levels of education, immigrants are more likely to have a second birth compared to the descendants and natives, at both durations and across all birth cohorts. By contrast, among highly educated individuals, the natives are slightly more likely to have a second birth compared to immigrants and the descendants. Most importantly, the group differences are reduced for highly educated individuals.

< Figure 6 here >

4.3.3. Third Birth

Finally, we examine the probability of having a third child among individuals who already have two children. We identify the most important variables to predict the likelihood of having a third child (Figure 7, see Figure A.8 in Appendix for the results using VIMP and minimal depth methods separately). The ranking is different from the previous ones: family size is now the most important variable to predict the event of having a third birth.

< Figure 7 here >

We first examine how the probability of having a third birth differs by migrant generation and birth cohort (Figure A.9 in Appendix). There are no significant differences across birth cohorts. Immigrants are more likely to have a third birth compared to the descendants and natives. This is the case both at 5 years and at 10 years after the second birth. Next, we examine the probability of having a third birth by migrant generation, birth cohort and educational level (Figure 8).[14][15] Immigrants are more likely to experience an event, irrespective of the birth cohort and level of education. Again, the differences in the third-birth probabilities between immigrants, descendants and natives are reduced for highly educated individuals.

< Figure 8 here >

---

[12] Figure A.7 in the Appendix also presents the predicted probabilities of having a second birth 5 and 10 years after the first birth, by origin group, birth cohort and educational level.

[13] We examine whether the results differ for men and women. The results (available upon request) show that overall, the patterns are similar. Among low educated individuals, immigrants have the highest probability while among highly educated individuals, the natives have the highest probability. However, one difference to note is that the patterns differ among men and women for the most recent cohort. Among men born in the 1980s, immigrants have the highest probability of having a second birth no matter the level of education, while among women born in the 1980s, the level of education matters: while among low educated individuals, immigrants have the highest probability, among highly educated individuals, the natives have the highest probability.

[14] Figure A.10 in the Appendix also presents the predicted probabilities of having a third birth 5 and 10 years after a second birth, by origin group, birth cohort and educational level.

[15] When exploring potential differences between men and women, we find significant differences among the most recent cohorts. For individuals born in the 1950s and in the 1960s, among both men and women, immigrants have the highest probability. Among individuals born in the 1970s, for men, immigrants have the highest probability no matter the level of education whereas for women, immigrants have the highest probability only among low educated individuals. For highly educated women, natives have the highest probability. Lastly, among individuals born in the 1980s, the fertility differentials are larger among women than men.

5. Discussion

This paper demonstrates that random survival forest can be useful as a substitute or a complement to standard survival and event history analysis. We applied the RSF technique to examine the probability of having a first, second and third birth among immigrants, their descendants, and natives using rich longitudinal data from France. Our analysis shows some interesting findings in relation to the fertility behaviour of immigrants. On average immigrants have a higher probability of having a first and third birth compared to natives. However, and importantly, fertility differences between immigrants and natives are much smaller among highly than low educated individuals. In other words, the study shows that highly educated immigrants are very similar to natives in their childbearing behaviour.

Our results highlight some advantages of RSF over conventional survival analysis. First, the method requires no pre-selection of variables. Second, the technique allows us to assess the predictive importance of the covariates. In particular, our results show that the educational level matters the most to explain the probability of a first birth. Migrant generation matters the most for the probability of a second birth, whereas family size is the most important predictor of a third birth. Third, the method is ideally suited for detection of complex interaction effects, thus helping to define groups of individuals with specific survival probability.

Although RSF allows us to address some issues that conventional methods of survival analysis face, the technique of RSF also suffers from shortcomings. The most notable drawback of RSF and machine learning techniques in general is that the models are "black boxes" that can be hard to interpret. This is due to the fact that RSF is completely data driven and thus independent of any hypothesis testing. It simply seeks a model that best explains the data. RSF may nevertheless represent a suitable tool for exploratory analysis of survival or time-to-event data where previous knowledge is limited. Our application of the method to the analysis of immigrant fertility behaviour shows that random forest can easily be applied to different settings in life-course research and that research on migrant fertility should pay more attention to how education shapes childbearing patterns among minority populations. We hope that our paper demonstrates the potential of RSF and other machine learning techniques for demographic research.

References

Adham, D., Abbasgholizadeh, N., and Abazari, M. (2017). Prognostic factors for survival in patients with gastric cancer using a random survival forest. Asian Pacific journal of cancer prevention: APJCP, 18(1), 129. https://doi.org/10.22034/APJCP.2017.18.1.129

Amit, Y., and Geman, D. (1997). Shape quantization and recognition with randomized trees. Neural computation, 9(7), 1545-1588. https://doi.org/10.1162/neco.1997.9.7.1545

Andersson, G. (2004). Childbearing after migration: Fertility patterns of foreign-born women in Sweden. International migration review, 38(2), 747-774. https://doi.org/10.1111/j.1747-7379.2004.tb00216.x

Andersson, G., and Scott, K. (2007). Childbearing dynamics of couples in a universalistic welfare state: The role of labor-market status, country of origin, and gender. Demographic research, 17, 897-938. https://doi.org/10.4054/DemRes.2007.17.30

Andersson, G., Obućina, O., and Scott, K. (2015). Marriage and divorce of immigrants and descendants of immigrants in Sweden. Demographic Research, 33, 31-64. https://doi.org/10.4054/DemRes.2015.33.2

Arpino, B., Le Moglie, M., and Mencarini, L. (2021). What Tears Couples Apart: A Machine Learning Analysis of Union Dissolution in Germany. Demography; 9648346. https://doi.org/10.1215/00703370-9648346

Billari, F. C., Fürnkranz, J., and Prskawetz, A. (2006). Timing, sequencing, and quantum of life course events: A machine learning approach. European Journal of Population/Revue Européenne de Démographie, 22(1), 37-65. https://doi.org/10.1007/s10680-005-5549-0

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

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

Cafri, G., Li, L., Paxton, E. W., and Fan, J. (2018). Predicting risk for adverse health events using random forest. Journal of Applied Statistics, 45(12), 2279-2294. https://doi.org/10.1080/02664763.2017.1414166

De Rose, A., and Pallara, A. (1997). Survival trees: An alternative non-parametric multivariate technique for life history analysis. European Journal of Population/Revue européenne de Démographie, 13(3), 223-241. https://doi.org/10.1023/a:1005844818027

Delaporte, I., and Kulu, H. (2021). Interaction between Childbearing and Partnership Changes among Immigrants and their Descendants: An Application of Multichannel Sequence Analysis to Longitudinal Data from France.

Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. Statistical Science, 18(1), 71-103. https://doi.org/10.1214/ss/1056397487

Ehrlinger, J. (2016). ggRandomForests: Exploring random forest survival. arXiv preprint arXiv:1612.08974.

Fawagreh, K., Gaber, M. M., and Elyan, E. (2014). Random forests: from early developments to recent advancements. Systems Science & Control Engineering: An Open Access Journal, 2(1), 602-609. https://doi.org/10.1080/21642583.2014.956265

Hamidi, O., Tapak, M., Poorolajal, J., Amini, P., and Tapak, L. (2017). Application of random survival forest for competing risks in prediction of cumulative incidence function for progression to AIDS. Epidemiology, Biostatistics and Public Health, 14(4).

Hanson, H. A., Martin, C., O'Neil, B., Leiser, C. L., Mayer, E. N., Smith, K. R., and Lowrance, W. T. (2019). The relative importance of race compared to health care and social factors in predicting prostate cancer mortality: a random forest approach. The Journal of urology, 202(6), 1209-1216.

Ho, T. K. (1995, August). Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition (Vol. 1, pp. 278-282). IEEE.

Ho, T. K. (1998). The random subspace method for constructing decision forests. IEEE transactions on pattern analysis and machine intelligence, 20(8), 832-844. https://doi.org/10.1109/34.709601

Hsich, E., Gorodeski, E. Z., Blackstone, E. H., Ishwaran, H., & Lauer, M. S. (2011). Identifying important risk factors for survival in patient with systolic heart failure using random survival forests. Circulation: Cardiovascular Quality and Outcomes, 4(1), 39-45. https://doi.org/10.1161/CIRCOUTCOMES.110.939371

Ishwaran, H. (2007). Variable importance in binary regression trees and forests. Electronic Journal of Statistics, 1, 519-537.

Ishwaran, H., Gerds, T. A., Kogalur, U. B., Moore, R. D., Gange, S. J., and Lau, B. M. (2014). Random survival forests for competing risks. Biostatistics, 15(4), 757-773.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. Annals of Applied Statistics, 2(3), 841-860.

Ishwaran, H., Kogalur, U. B., Gorodeski, E. Z., Minn, A. J., and Lauer, M. S. (2010). High-dimensional variable selection for survival data. Journal of the American Statistical Association, 105(489), 205-217.

Ishwaran, H., Kogalur, U. B., Chen, X., and Minn, A. J. (2011). Random survival forests for high-dimensional data. Statistical Analysis and Data Mining: The ASA Data Science Journal, 4(1), 115-132.

Ishwaran, H., and Kogalur, U. B. (2008). RandomSurvivalForest 3.2. 2. R package.

Ishwaran, H., and Kogalur, U. B. (2014). RandomForestSRC: Random forests for survival, regression and classification (RF-SRC). R package version, 1(0).

Keramati, A., Lu, P., Iranitalab, A., Pan, D., and Huang, Y. (2020). A crash severity analysis at highway-rail grade crossings: The random survival forest method. Accident Analysis & Prevention, 144, 105683. https://doi.org/10.1016/j.aap.2020.105683

Krapf, S., and Wolf, K. (2016). Persisting differences or adaptation to German fertility patterns? First and second birth behavior of the 1.5 and second generation Turkish migrants in Germany. In Social

Demography Forschung an der Schnittstelle von Soziologie und Demografie (pp. 137-164). Springer VS, Wiesbaden. https://doi.org/10.1007/s11577-015-0331-8

Kulu, H. (2005). Migration and fertility: Competing hypotheses re-examined. European Journal of Population/Revue européenne de Démographie, 21(1), 51-87. https://doi.org/10.1007/s10680-005-3581-8

Kulu, H., and González-Ferrer, A. (2014). Family dynamics among immigrants and their descendants in Europe: Current research and opportunities. European Journal of Population, 30(4), 411-435. https://doi.org/10.1007/s 10680-014-9322-0

Kulu, H., and Hannemann. T. (2016a). Introduction to research on immigrant and ethnic minority families in Europe. Demographic Research, 35, 31-46. https://doi.org/10.4054/DemRes.2016.35.2

Kulu, H., and Hannemann, T. (2016b). Why does fertility remain high among certain UK-born ethnic minority women?. Demographic Research, 35, 1441-1488. https://doi.org/10.4054/DemRes.2016.35.49

Kulu, H. et al. (2017). Fertility by birth order among the descendants of immigrants in selected European countries. Population and Development Review, 31-60.

Kulu, H., Milewski, N., Hannemann, T., and Mikolai, J. (2019). A decade of life-course research on fertility of immigrants and their descendants in Europe. Demographic Research, 40, 1345-1374. https://doi.org/10.4054/DemRes.2019.40.46

Liaw, A., and Wiener. M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.

Loh, W. Y. (2011). Classification and regression trees. Wiley interdisciplinary reviews: data mining and knowledge discovery, 1(1), 14-23.

Miao, F., Cai, Y. P., Zhang, Y. T., and Li, C. Y. (2015). Is random survival forest an alternative to Cox proportional model on predicting cardiovascular disease?. In 6TH European conference of the international federation for medical and biological engineering (pp. 740-743). Springer, Cham.

Milewski, N. (2007). First child of immigrant workers and their descendants in West Germany: Interrelation of events, disruption, or adaptation?. Demographic Research, 17, 859-896.

Milewski, N. (2010). Immigrant fertility in West Germany: Is there a socialization effect in transitions to second and third births?. European Journal of Population/Revue européenne de Démographie, 26(3), 297-323. https://doi.org/10.1007/s10680-010-9211-0

Mussino, E., and Strozza, S. (2012). Does citizenship still matter? Second birth risks of migrants from Albania, Morocco, and Romania in Italy. European Journal of Population/Revue européenne de Démographie, 28(3), 269-302. https://doi.org/10.1007/s 10680-012-9261-6

Pailhé, A. (2017). The convergence of second-generation immigrants' fertility patterns in France: The role of sociocultural distance between parents' and host country. Demographic Research, 36, 1361-1398. https://doi.org/10.4054/DemRes.2017.36.45

Rezaei, M., Tapak, L., Alimohammadian, M., Sadjadi, A., and Yaseri, M. (2020). Review of Random Survival Forest method. Journal of Biostatistics and Epidemiology, 6(1), 59-68.

Rojas, E. A. G., Bernardi, L., and Schmid, F. (2018). First and second births among immigrants and their descendants in Switzerland. Demographic Research, 38, 247-286. https://doi.org/10.4054/DemRes.2018.38.11

Scheffner, I., Gietzelt, M., Abeling, T., Marschollek, M., and Gwinner, W. (2020). Patient survival after kidney transplantation: important role of graft-sustaining factors as determined by predictive modeling using random survival forest analysis. Transplantation, 104(5), 1095-1107. https://doi.org/10.1097/TP.0000000000002922

Spooner, A., Chen, E., Sowmya, A., Sachdev, P., Kochan, N. A., Trollor, J., and Brodaty, H. (2020). A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. Scientific reports, 10(1), 1-10. https://doi.org/10.1038/s41598-020-77220-w

Wang, H., and Li, G. (2017). A selective review on random survival forests for high dimensional data. Quantitative bio-science, 36(2), 85. https://doi.org/10.22283/qbs.2017.36.2.85

Wang, P., Li, Y., and Reddy, C. K. (2019). Machine learning for survival analysis: A survey. ACM Computing Surveys (CSUR), 51(6), 1-36. https://doi.org/10.1145/3214306

Whetten, A. B., Stevens, J. R., and Cann, D. (2021). The implementation of random survival forests in conflict management data: An examination of power sharing and third party mediation in post-conflict countries. Plos one, 16(5), e0250963. https://doi.org/10.1371/journal.pone.0250963

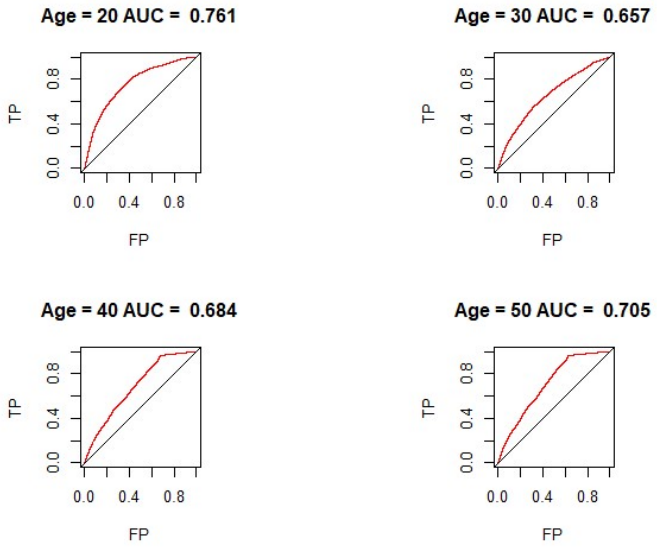
Wilson, B. (2020). Understanding how immigrant fertility differentials vary over the reproductive life course. European Journal of Population, 36(3), 465-498. https://doi.org/10.1007/s10680-019-09536-x

Witten, D. M., and Tibshirani, R. (2010). Survival analysis with high-dimensional covariates. Statistical methods in medical research, 19(1), 29-51. https://doi.org/10.1177/0962280209105024

Ziegler, A., and König, I. R. (2014). Mining data with random forests: current options for real-world applications. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 4(1), 55-63.

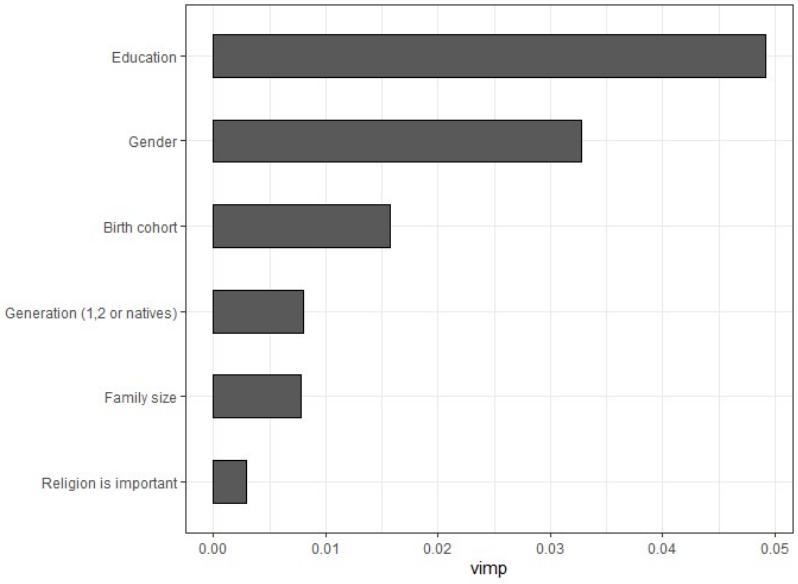
Tables and Figures

Figure 1. ROC Curves at Different Surviving Time

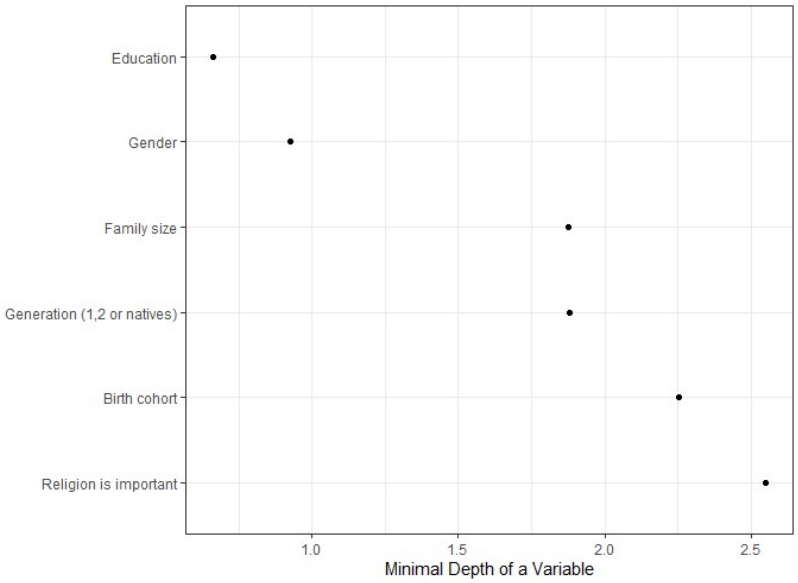


Source: Trajectories and Origins, authors' own calculations.

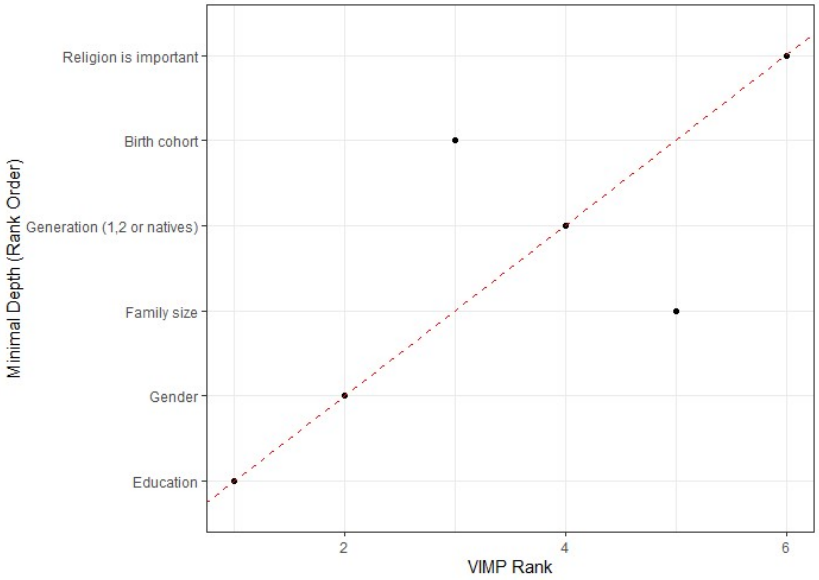
Figure 2. Random Forest Variable Selection – Probability of a First Birth



a) VIMP



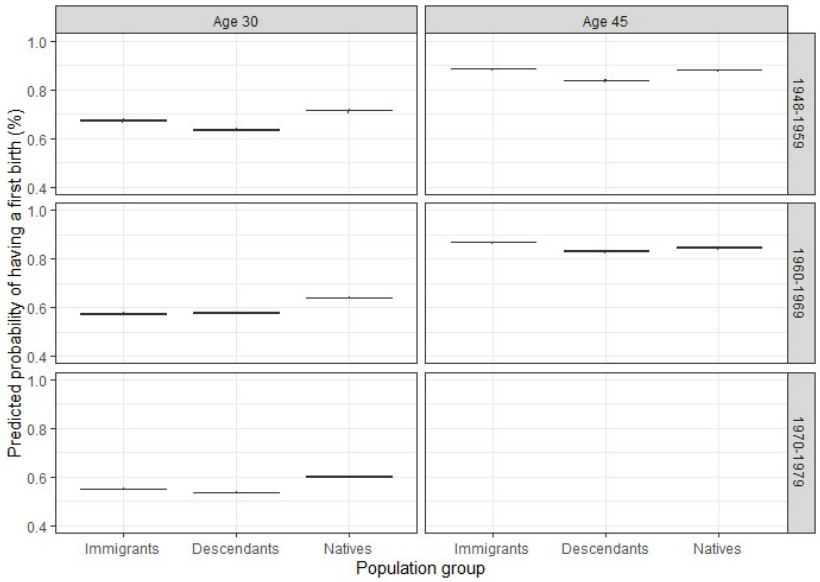
b) Minimal Depth



c) Comparison

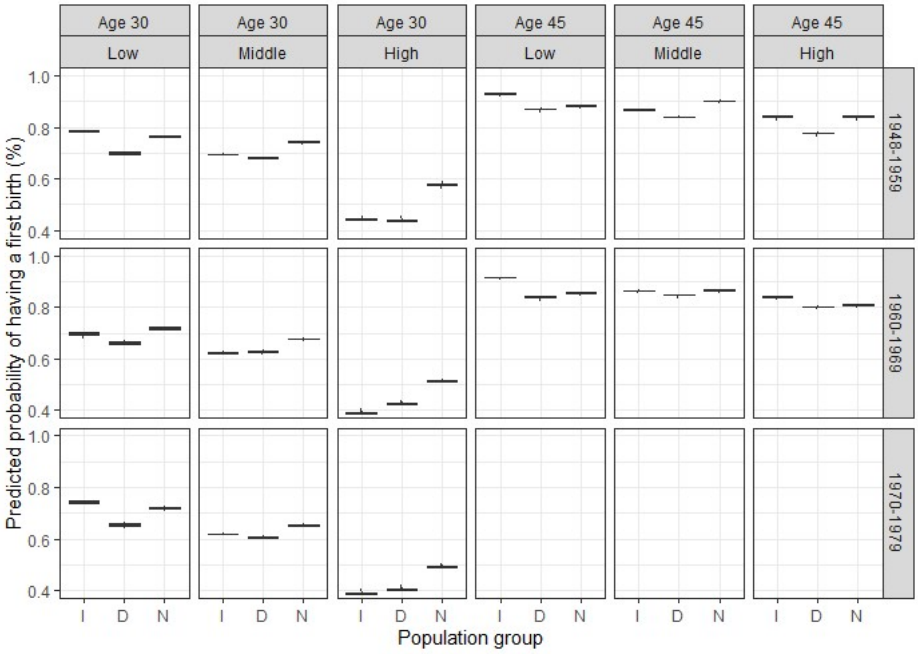
Source: Trajectories and Origins, authors' own calculations. Notes: In (a), we present the results of Variable Importance (VIMP). Importance is relative to length of bars. In (b), we present the results using Minimal Depth. Low minimal depth indicates important variables. All variables are above the threshold of maximum value for variable selection. Lastly, in (c), we compare the two variable rankings.

Figure 3. Predicted Probability of a First Birth by Age 30 and Age 45, by Migrant Generation and Birth Cohort



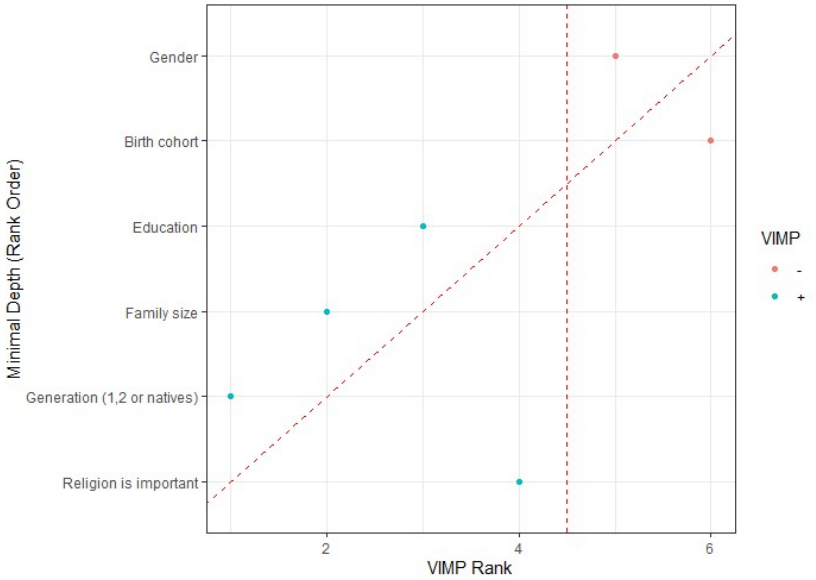
Source: Trajectories and Origins, authors' own calculations. Notes: The black lines are the median values for each group. There are no predicted probabilities for the cohort 1970-1979 by the age of 45 since they have not reached this age yet.

Figure 4. Predicted Probability of a First Birth by Age 30 and Age 45, by Migrant Generation, Birth Cohort and Educational Level



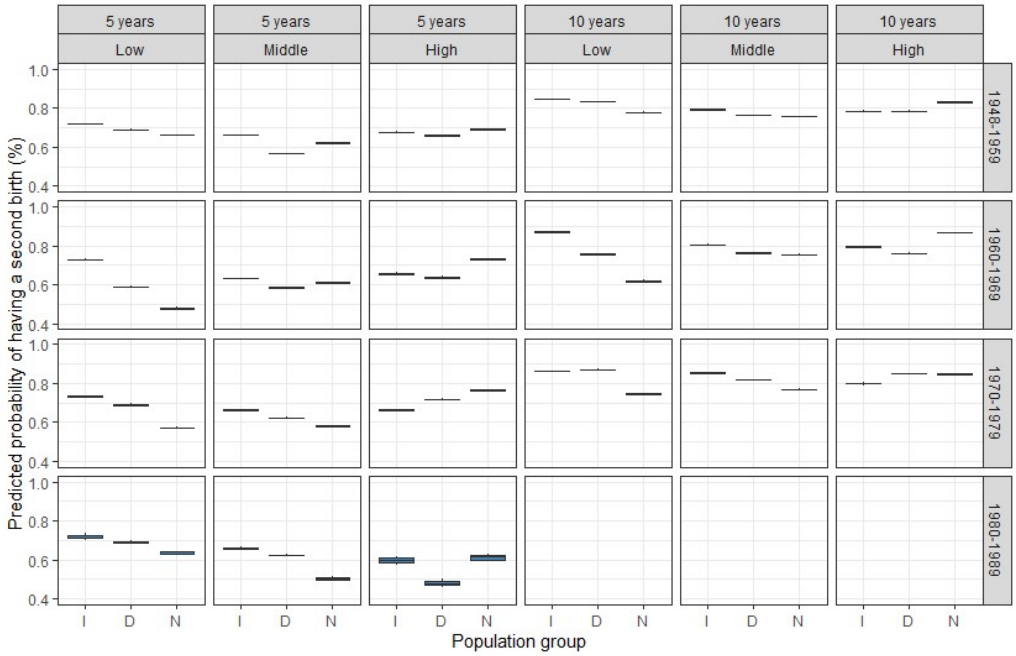
Source: Trajectories and Origins, authors' own calculations. Notes: The black lines are the median values for each group. "I" stands for immigrants, "D" stands for the descendants of immigrants and "N" stands for natives. There are no predicted probabilities for the cohort 1970-1979 by the age of 45 since they have not reached this age yet.

Figure 5. Random Forest Variable Selection – Probability of a Second Birth



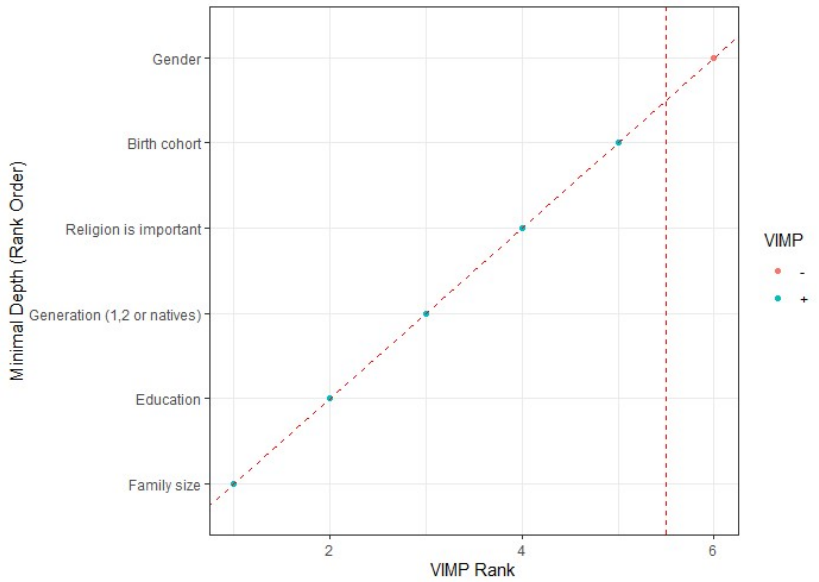
Source: Trajectories and Origins, authors' own calculations. Notes: we compare the two variable rankings: i) using VIMP and ii) using minimal depth methods.

Figure 6. Predicted Probability of a Second Birth at 5 and 10 Years Since First Birth, by Migrant Generation, Birth Cohort and Educational Level



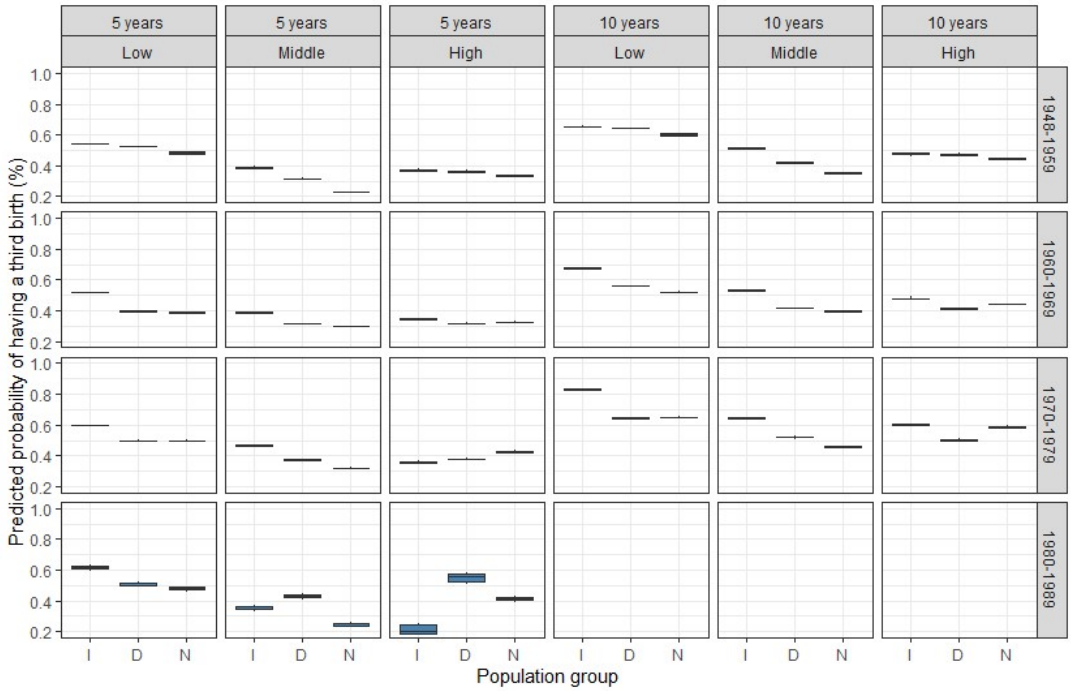
Source: Trajectories and Origins, authors' own calculations. Notes: The black lines are the median values for each group. The vertical size of the boxes represents the interquantile range. "I" stands for immigrants, "D" stands for the descendants of immigrants and "N" stands for natives. There are no predicted probabilities for the cohort 1980-1989 for the period 10 years after the first birth since they have not reached this stage yet.

Figure 7. Random Forest Variable Selection – Probability of a Third Birth



Source: Trajectories and Origins, authors' own calculations. Notes: we compare the two variable rankings: i) using VIMP and ii) using minimal depth methods.

Figure 8. Predicted Probability of a Third Birth at 5 and 10 Years Since Second Birth, by Migrant Generation, Birth Cohort and Educational Level



Source: Trajectories and Origins, authors' own calculations. Notes: The black lines in the middle of the boxes are the median values for each group. The vertical size of the boxes represents the interquantile range. Lastly, the flattened arrows extending out of the box are the minimum and maximum values. There are no predicted probabilities for the cohort 1980-1989 for the period 10 years after the second birth since they have not reached this stage yet.

Appendix

Figure A.1. Steps of the Random Survival Forest Algorithm

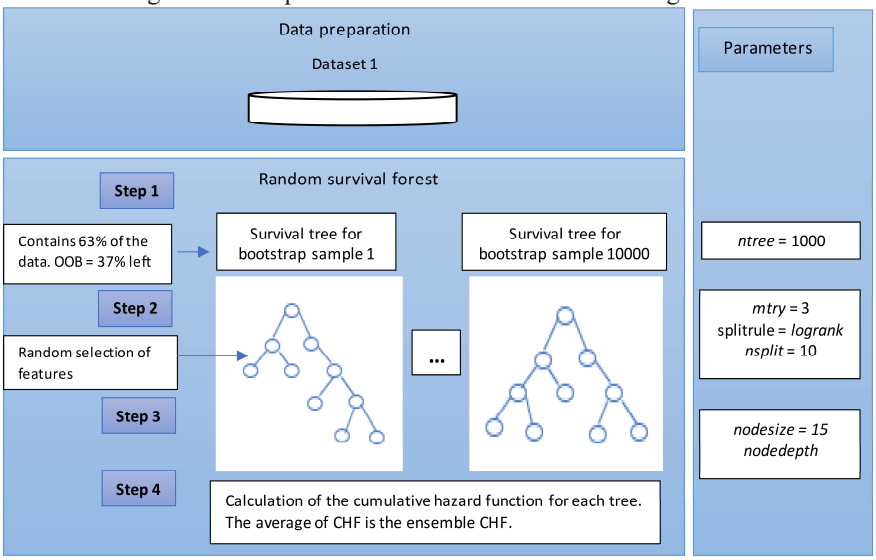
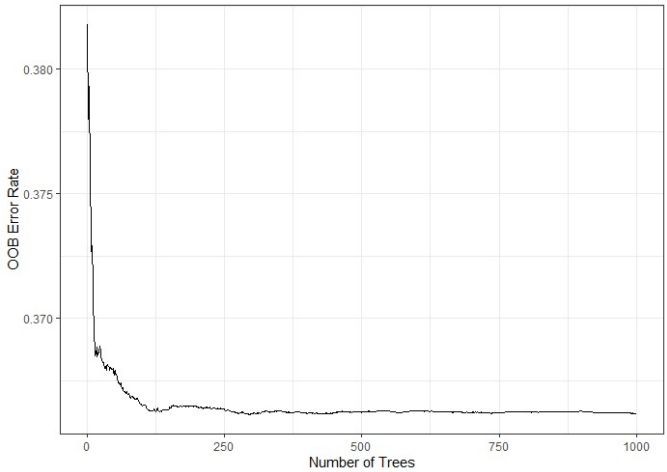
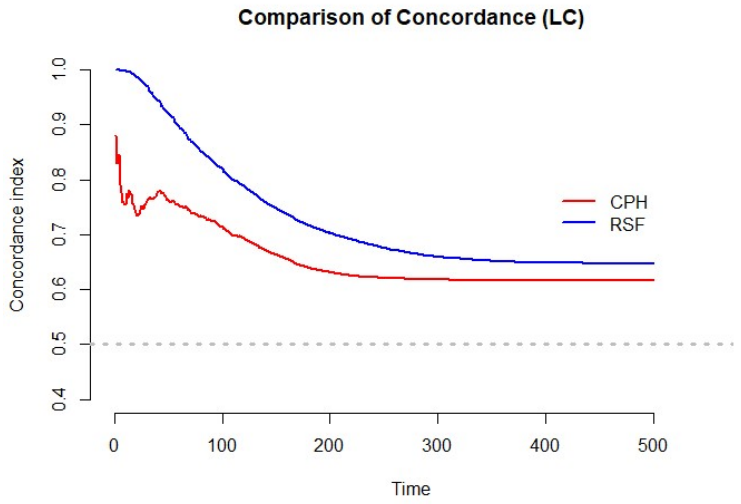


Figure A.2. Out of Bag Errors (OOB)



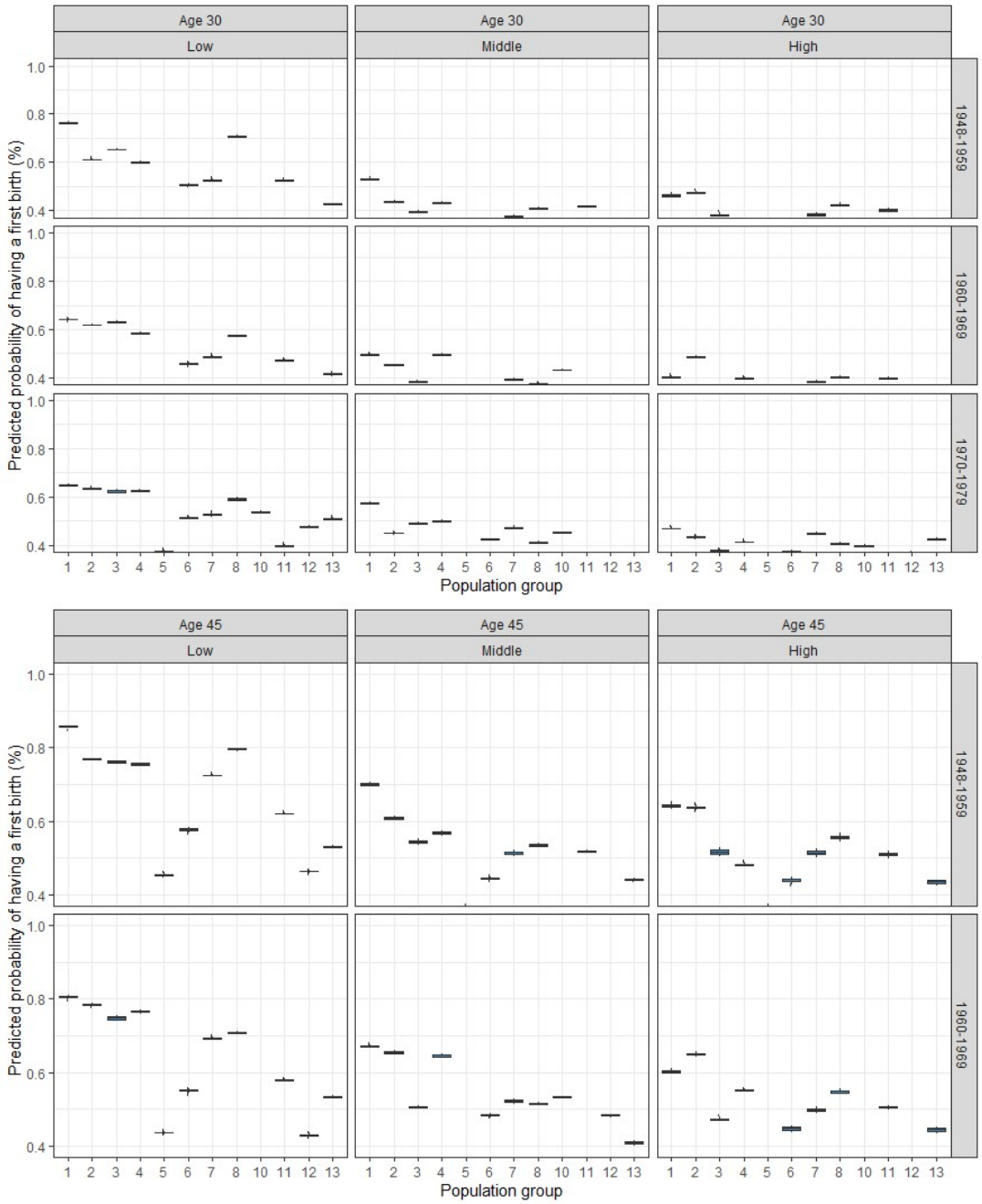
Source: Trajectories and Origins, authors' own calculations.

Figure A.3. Comparison of c-indexes



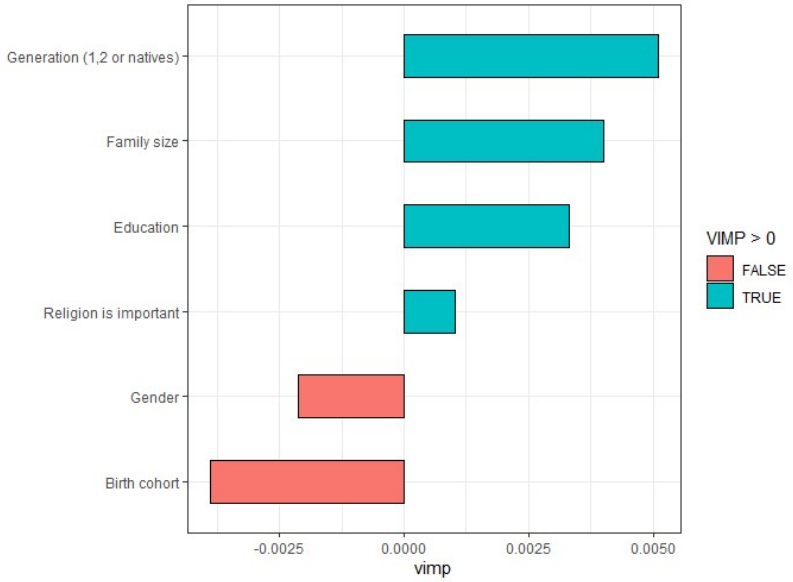
Source: Trajectories and Origins, authors' own calculations. Notes: the time on the x-axis goes from the ages of 15 to 40 approximately.

Figure A.4. Predicted Probability of a First Birth by Age 30 and Age 45, by Migrant Generation, Birth Cohort and Educational Level

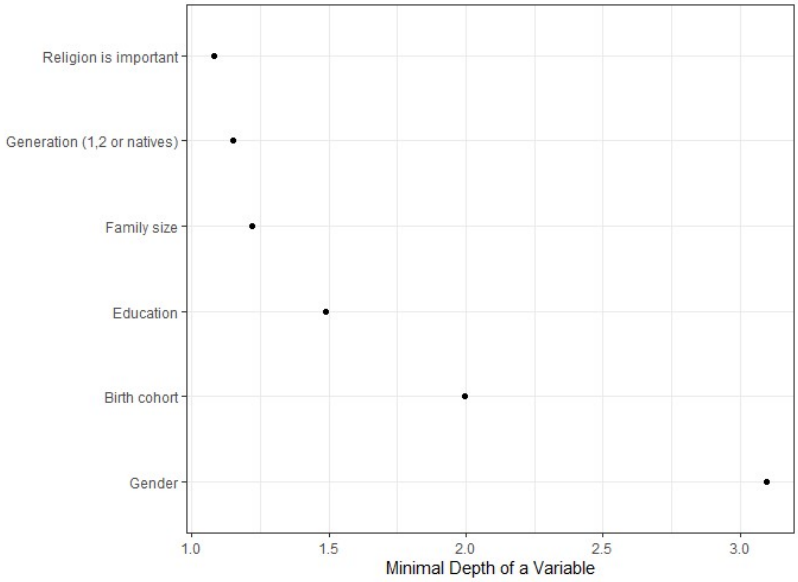


Source: Trajectories and Origins, authors' own calculations. Notes: The black lines are the median values for each group. "Low", "Middle" and "High" refer to the level of education. The population groups from 1 to 13 stands for: "1G North Africa", "1G Sub-Saharan Africa", "1G South East Asia", "1G Turkey", "1G Southern Europe", "1G other Europe", "2G North Africa", "2G Sub-Saharan Africa", "2G South East Asia", "2G Turkey", "2G Southern Europe", "2G other Europe" and lastly "Natives".

Figure A.5. Random Forest Variable Selection – Probability of a Second Birth



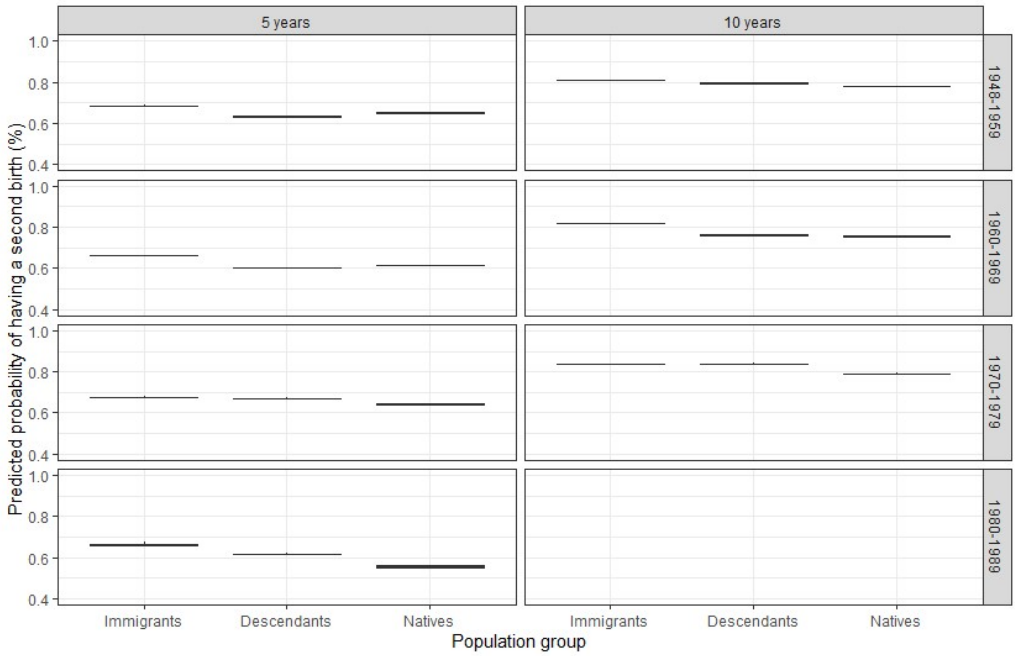
a) VIMP



b) Minimal Depth

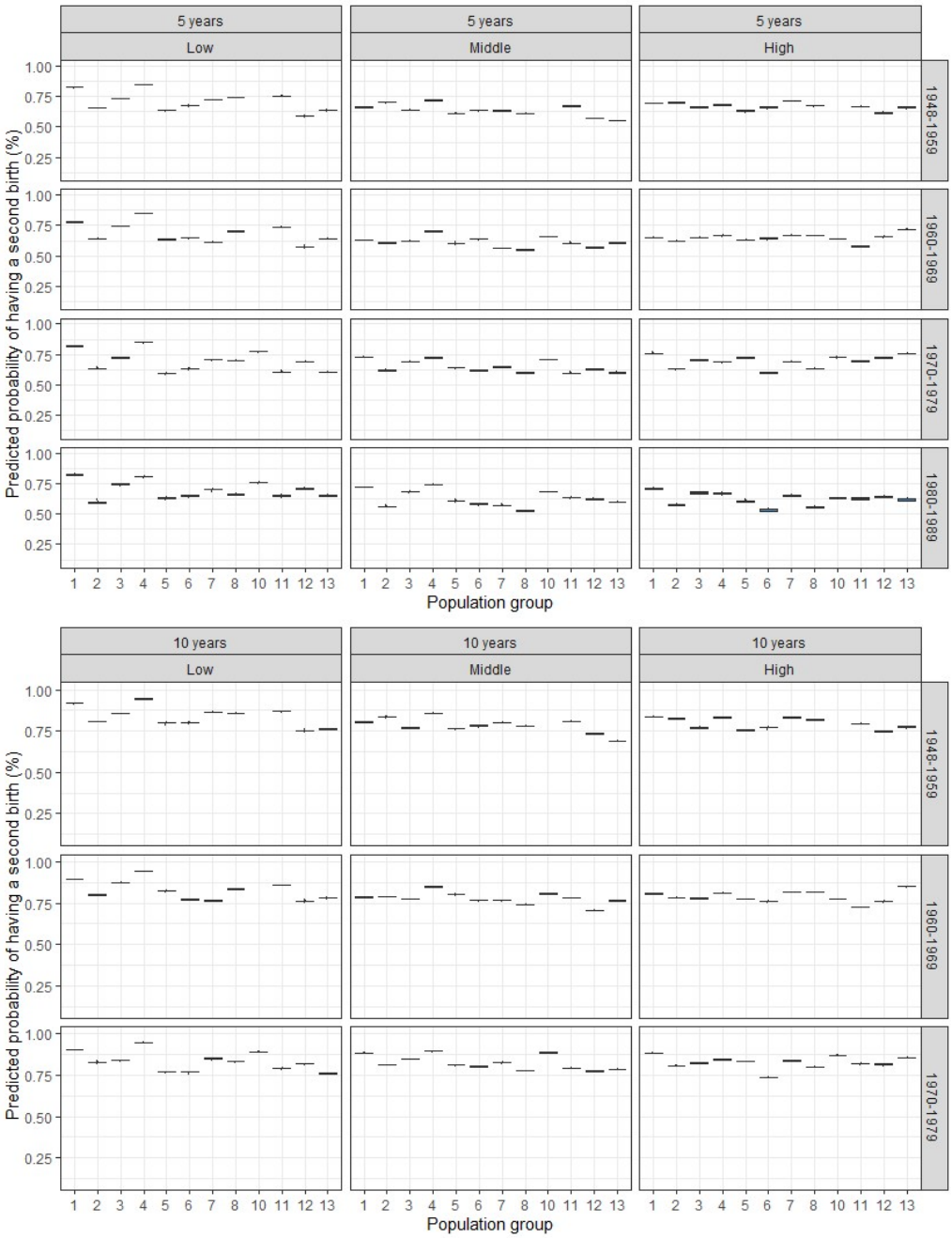
Source: Trajectories and Origins, authors' own calculations. Notes: In (a), we present the results of Variable Importance (VIMP). Importance is relative to length of bars. In (b), we present the results using Minimal Depth. Low minimal depth indicates important variables. All variables are above the threshold of maximum value for variable selection.

Figure A.6. Predicted Probability of a Second Birth at 5 and 10 Years Since First Birth by Migrant Generation and Birth Cohort



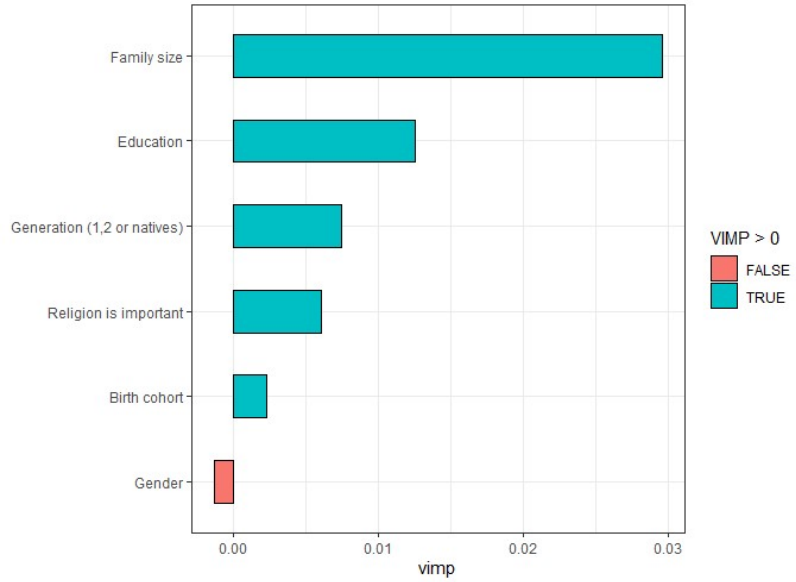
Source: Trajectories and Origins, authors' own calculations. Notes: The black lines are the median values for each group. There are no predicted probabilities for the cohort 1980-1989 for the period 10 years after the first birth since they have not reached this stage yet.

Figure A.7. Predicted Probability of a Second Birth at 5 and 10 Years After the First Birth, by Migrant Generation, Birth Cohort and Educational Level

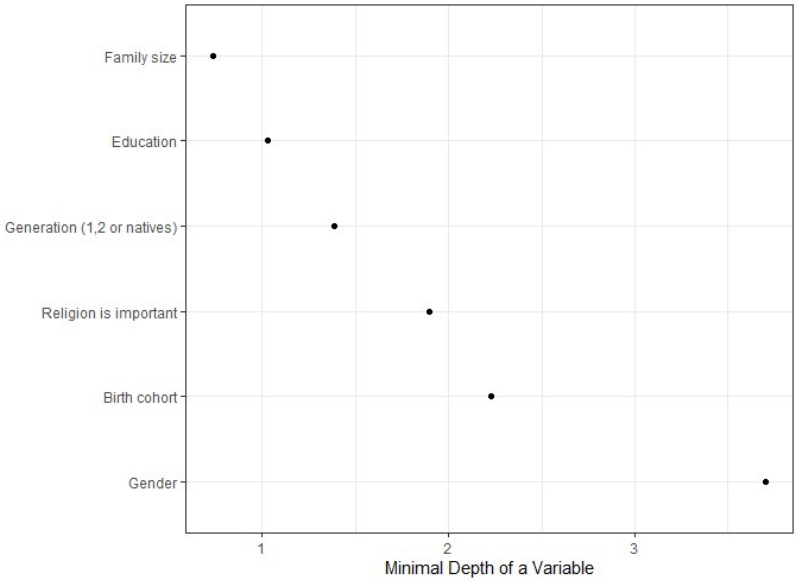


Source: Trajectories and Origins, authors' own calculations. Notes: The black lines are the median values for each group. "Low", "Middle" and "High" refer to the level of education. The population groups from 1 to 13 stands for: "1G North Africa", "1G Sub-Saharan Africa", "1G South East Asia", "1G Turkey", "1G Southern Europe", "1G other Europe", "2G North Africa", "2G Sub-Saharan Africa", "2G South East Asia", "2G Turkey", "2G Southern Europe", "2G other Europe" and lastly "Natives".

Figure A.8. Random Forest Variable Selection – Probability of a Third Birth



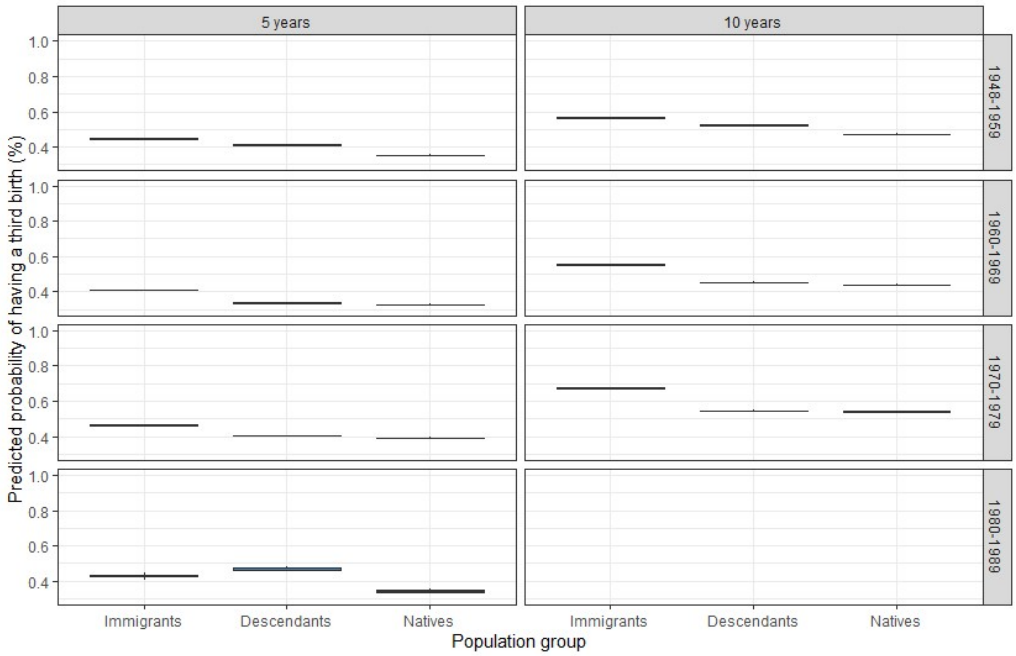
a) VIMP



b) Minimal Depth

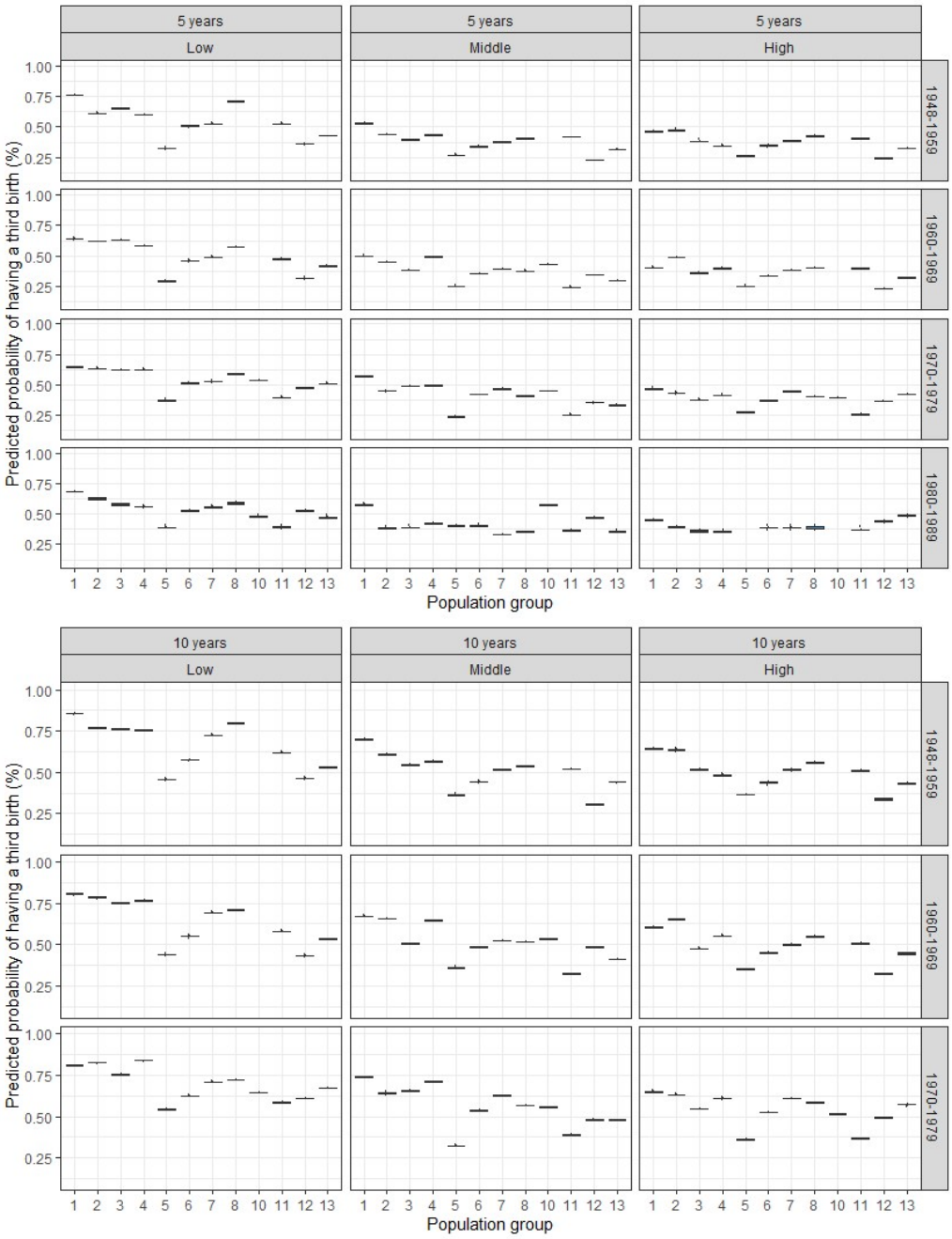
Source: Trajectories and Origins, authors' own calculations. Notes: In (a), we present the results of Variable Importance (VIMP). Importance is relative to length of bars. In (b), we present the results using Minimal Depth. Low minimal depth indicates important variables. All variables are above the threshold of maximum value for variable selection.

Figure A.9. Predicted Probability of a Third Birth at 5 and 10 Years Since Second Birth, by Migrant Generation and Birth Cohort



Source: Trajectories and Origins, authors' own calculations. Notes: The black lines are the median values for each group. There are no predicted probabilities for the cohort 1980-1989 for the period 10 years after the second birth since they have not reached this stage yet.

Figure A.10. Predicted Probability of a Third Birth at 5 and 10 Years After the Second Birth, by Migrant Generation, Birth cohort and Educational Level



Source: Trajectories and Origins, authors' own calculations. Notes: The black lines are the median values for each group. "Low", "Middle" and "High" refer to the level of education. The population groups from 1 to 13 stands for: "1G North Africa", "1G Sub-Saharan Africa", "1G South East Asia", "1G Turkey", "1G Southern Europe", "1G other Europe", "2G North Africa", "2G Sub-Saharan Africa", "2G South East Asia", "2G Turkey", "2G Southern Europe", "2G other Europe" and lastly "Natives".