DEMOGRAPHIC RESEARCH
A peer-reviewed, open-access journal of population sciences

DEMOGRAPHIC RESEARCH

VOLUME 53, ARTICLE 21, PAGES 629–660
PUBLISHED 14 OCTOBER 2025
https://www.demographic-research.org/Volumes/Vol53/21
DOI: 10.4054/DemRes.2025.53.21

Research Article

Analysing migrant fertility using machine learning techniques: An application of random survival forest to longitudinal data from France

Isaure Delaporte

Hill Kulu

Andrew Ibbetson

© 2025 Isaure Delaporte, Hill Kulu & Andrew Ibbetson.

This open-access work is published under the terms of the Creative Commons Attribution 3.0 Germany (CC BY 3.0 DE), which permits use, reproduction, and distribution in any medium, provided the original author(s) and source are given credit.
See https://creativecommons.org/licenses/by/3.0/de/legalcode.

Contents

Demographic Research: Volume 53, Article 21
Research Article

Analysing migrant fertility using machine learning techniques: An application of random survival forest to longitudinal data from France

Isaure Delaporte[1]

Hill Kulu[2]

Andrew Ibbetson[3]

Abstract

BACKGROUND
The fertility of immigrants and their descendants is shaped by many factors. Survival and event history techniques are methods commonly used to study the determinants of individuals' childbearing behaviour. Yet, machine learning techniques such as survival trees and tree ensembles are a useful alternative to classical methods.

OBJECTIVE
This paper analyses the predictors of having a first, second, and third birth among immigrants and their descendants in France.

METHODS
This study applies random survival forest (RSF) to longitudinal data from the Trajectories and Origins survey.

RESULTS
Our findings illustrate the potential of machine learning techniques in two ways. First, RSF allows us to identify the most important predictors of a life event. Our results show that predictors differ by parity: Educational level is the most important predictor of having a first child, whereas parents' family size is the most important predictor of having a second and third child. Second, RSF allows us to easily detect and visualize interactions. For instance, our results of a four-way interaction show that highly educated migrants are closer to the native population in their childbearing behaviour than migrants with low education.

[1] International Labour Organization, Geneva, Switzerland. Email: delaporte@ilo.org.
[2] University of St Andrews, UK. E-mail: hill.kulu@st-andrews.ac.uk.
[3] University College London, UK. E-mail: andrew.ibbetson.13@ucl.ac.uk.

CONTRIBUTION
Our application of RSF to the analysis of immigrant fertility behaviour shows that the method can easily be applied in life course research and that research on migrant fertility should pay more attention to how education shapes childbearing patterns among minority populations.

1. Introduction

The fertility behaviour of immigrants and their descendants is subject to multiple influences (Kulu and González-Ferrer 2014; Kulu and Hannemann 2016a). Previous studies have shown that immigrants exhibit higher fertility levels than natives; they also have children earlier compared to the native population (defined here as native-born individuals with two native-born parents). By contrast, the descendants of immigrants often exhibit family patterns that are more similar to those of the native population. 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 2022). Regarding the fertility patterns of immigrants' descendants, the sociocultural distance between the parents' country of birth and the host country as well as structural determinants play an important role (Pailhé 2017; Krapf and Wolf 2016).

Survival and event history techniques are commonly used methods to identify the factors that shape individuals' fertility behaviour. Yet, the technique of survival analysis is not without limitations. For instance, survival analysis cannot be easily applied in high-dimensional settings (Wang and Li 2017; Spooner et al. 2020): With a large number of covariates in the model, many statistically insignificant parameters may complicate the interpretation of the results (Witten and Tibshirani 2010; Dudoit, Shaffer, and Boldrick 2003; Whetten, Stevens, and Cann 2021). When many covariates are simultaneously included in the model, collinearity also jeopardizes interpretation of the results. Furthermore, it is difficult to detect and visualize interactions between two or more variables. Lastly, parametric models require the proportional hazards assumption to hold. Since migrant fertility behaviour is a complex process, non-parametric methods may be useful to identify important predictors and address nonlinearities.

Survival trees and tree ensembles can be a useful alternative to classical survival analysis (Breiman et al. 1984; Breiman 2001; Ishwaran et al. 2008; Ishwaran and Kogalur 2008, 2014). However, to date, only a limited number of studies in demography have

used machine learning techniques (see Kashyap et al. 2022). 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, Fürnkranz, and Prskawetz (2006) also apply decision tree learning and classification rules to detect the predictors 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 predictors of divorce among married and cohabiting women in Germany. Apart from these studies, so far RSF has been applied 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, Abbasgholizadeh, and Abazari 2017; Hanson et al. 2019; Cafri et al. 2018).

In this study, we apply random survival forest to investigate the fertility behaviour of immigrants and their descendants in France. We use a rich survey named Trajectories and Origins which contains detailed information on immigrants, immigrants' descendants, and French natives. It contains retrospective biographical data on individuals' childbearing histories as well as information on their sociodemographic characteristics. This allows us to predict the likelihood of having a first, second, and third birth. We first examine the predictive performance of the algorithm. We then analyse which predictors are important to explain a first or subsequent birth. Finally, we examine possible interaction effects. To the best of our knowledge, this is the first study to apply RSF to the topic of immigrant and ethnic minority fertility. Furthermore, previous studies have stressed the usefulness of machine learning techniques mostly to identify the most important predictors of a specific behaviour. This study also demonstrates their ability to detect and visualise (complex) interaction effects.

Our results show, first, that the most important predictors differ for a first, second and third birth: Educational level is the most important predictor of having a first child, whereas parents' family size is the most important predictor of having a second and third child. Second, our results show that highly educated migrants are closer to natives in their childbearing patterns than migrants with low education.

2. Understanding immigrant fertility behaviour

Competing hypotheses have been proposed to understand differences in fertility behaviour between immigrants and the majority population. According to the socialisation hypothesis, the social environment in which individuals grow up has an important impact on individuals' family preferences (Andersson 2004; Kulu and Milewski 2007). Therefore, immigrants' family behaviour is largely shaped by the family norms of their country of origin. By contrast, the adaptation hypothesis argues that

immigrants adapt to the host country's social and economic environment and their fertility behaviour gradually converges to that of the native population (Andersson 2004; Kulu et al. 2019). The selection hypothesis states that immigrants are a select group of people with life expectations, aspirations, and values similar to those in the destination country.

If the descendants of immigrants grow up under the influence of a minority subculture they will exhibit family patterns that closely resemble those of their parents. Equally, the second generation may grow up under the influence of the mainstream society and thus show family patterns similar to those of natives (Kulu et al. 2019; Delaporte and Kulu 2022).

Previous studies have shown that immigrants start childbearing at a younger age and have higher fertility levels than the native population (Kulu and González-Ferrer 2014; Kulu et al. 2019; Rojas, Bernardi, and Schmid 2018). These differences in fertility levels are especially pronounced for specific groups (Kulu and Hannemann 2016b; Pailhé 2017; Mussino and Strozza 2012; Andersson and Scott 2007; Delaporte and Kulu 2022). For instance, immigrants in France from Turkey and Southern Europe have higher first birth risks compared to natives. In France, the risk of having a second or a third child is also significantly higher among immigrant women from Turkey and North Africa (Kulu et al. 2017).

Similar findings have been found in other European countries. For instance, in the United Kingdom, immigrants from Pakistan and Bangladesh have higher first-birth risks than natives (Kulu and Hannemann 2016b). In Sweden, Andersson and Scott (2007) report that immigrants from high fertility countries have significantly higher second- and third-birth levels than Swedish-born women. In West Germany, Milewski (2010) shows that second- and third-birth levels are relatively high for immigrant women from Turkey. Mussino and Strozza (2012) find that in Italy, immigrants from North Africa have significantly higher fertility levels.

Immigrants' and natives' fertility also differ at different ages (Wilson 2020) and across birth cohorts (Erman 2022). There are also some differences across generations (Kulu and Hannemann 2016b; Pailhé 2017; Mussino and Strozza 2012; Andersson and Scott 2007; Delaporte and Kulu 2022). 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). The reason for migration (e.g., work vs. family reunification) is also important in explaining fertility behaviour (Mussino and Cantalini 2022). Briefly, origin group, age at arrival, migrant generation, and reason for migration are all potentially important predictors of having a first or subsequent birth among migrant populations.

Besides personal characteristics, structural determinants also play an important role. Fertility differentials between migrants and natives may vanish when the

sociodemographic structure of an immigrant group changes to resemble that of the native population (Milewski 2007, 2010). Access to higher education is a crucial factor in reducing the differences between groups (Krapf and Wolf 2016; Pailhé 2017). High educational aspirations among ethnic minority women may lead to a significant postponement of family formation and smaller family size (Kulu and Hannemann 2016b). By contrast, poor employment prospects among some ethnic minority groups due to low levels of education and/or discrimination in the labour market may promote high completed fertility (Kulu and Hannemann 2016b).

Finally, there are other factors that influence individuals' childbearing behaviour. For instance, the role of family background in fertility outcomes has been extensively studied (Hays and Guzzo 2022; Baudin 2015; Berghammer, 2009). Existing studies suggest that family size as well as family complexity can be transmitted across generations. Family values such as the importance of religion also influence individuals' childbearing behaviour (Baudin 2015; Berghammer 2009).

Overall, previous studies highlight a number of important predictors of fertility. However, it may be difficult to include all potentially important variables when using conventional methods and often the researcher has to pre-select variables. Using RSF allows us to test the importance of a large number of potential predictors for different outcomes of interest. Furthermore, the existing literature suggests the presence of subgroups with specific family behaviour. Conventional survival analysis is not always best suited to detect interaction effects, especially if more than two variables are involved, whereas RSF allows us to easily detect and visualise interactions.

3. Methods

3.1 Data

To carry out the analysis we use Trajectories and Origins, a rich French survey collected in 2008. Information was collected on immigrants, immigrants' descendants, and French natives. The survey provides retrospective childbearing histories for all individuals on a monthly time scale. We also have detailed information on individuals' personal characteristics. For the purpose of this study, we analyse three outcomes: the event of having a first, second, or third birth.

When predicting the event of having a first birth, all individuals are examined, although we exclude from our analysis individuals who were born in the 1990s since they were too young to have had a first birth at the time of interview. The sample consists of 20,346 individuals including 8,233 immigrants, 8,609 descendants of immigrants, and 3,504 natives. When predicting the event of having a second birth, only individuals that

have had a first child are included. The sample consists of 12,600 individuals including 6,492 immigrants, 3,894 descendants of immigrants, and 2,214 natives. Lastly, when predicting the event of having a third birth, only individuals that have had two children are examined. The sample includes 9,336 individuals with 5,122 immigrants, 2,610 descendants of immigrants, and 1,604 natives. For all samples, the largest migrant group was North Africans and the smallest was Turkish, representing 22% and 7% of individuals, respectively (Table A-1).

For each birth, we construct a dummy variable that is equal to 1 when the individual experiences the birth, and 0 otherwise. On the basis of the literature review, we decided to include a series of predictors of childbearing which are time-constant, such as gender, birth cohort, size of the family of origin, education, religiosity, and a number of variables reflecting family background, and the reason for migration (for migrants only). We distinguish immigrants, their descendants, and natives. The birth cohorts are 1948–1959, 1960–1969, 1970–1979 and 1980–1989. 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). Religiosity is a dummy variable equal to 1 if the respondent reported that religion was important in their upbringing and 0 otherwise. The variables measuring family background are coded 0 or 1; the reason for migration (for migrants only) has 8 categories.

3.2 Random survival forest

To predict births, this study uses RSF, an extension of random forests to right-censored survival or time-to-event data (Ishwaran 2007, Ishwaran et al. 2008).[4] To run the algorithm, we use the randomForestSRC package in R.[5] Over the last two decades, many applications of random forest have been developed across different disciplines such as biostatistics, medicine, bioinformatics and computational biology, and economics and finance (Fawagreh, Gaber, and Elyan 2014; Best et al. 2021; 2022). Yet, its application in the field of demography remains limited (Kashyap et al. 2022).

[4] 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 (Ishwaran et al. 2014; Keramati et al. 2020; Hamidi 2017; Wang, Li, and Reddy 2019).

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

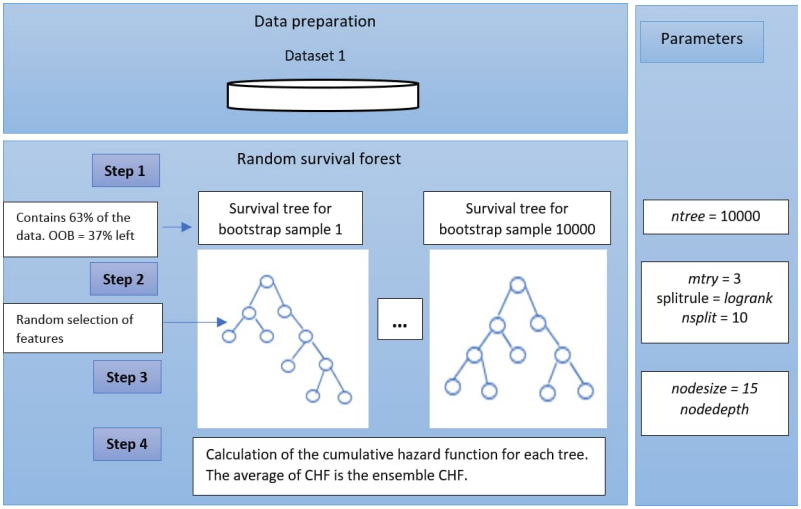
Developed by Breiman (2001), random forest combines Breiman's bagging sampling approach[6] and the random selection of features[7] introduced by Ho (1995, 1998) and Amit and German (1997) to construct a collection of decision trees. In practice, random forest runs the analysis over many sub-datasets made by randomly selecting features (Breiman 2001). The prediction is obtained by averaging over hundreds or thousands of distinct regression trees, which 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 is a useful method for predicting births because it can be used with a large number of correlated variables, thereby allowing us to identify important interactions and non-linear associations between variables that would typically be excluded when using conventional methods.

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 1, 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 (Wang, Li, and Reddy 2019). Each step of the algorithm involves defining specific parameters (see Figure 1).

[6] The bagging sampling approach is an important characteristic that allows the reduction of overfitting issues. As each split 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.

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

Figure 1: Steps of the random survival forest algorithm



3.2.1 The cumulative hazard estimate

To formalise our approach, 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) is calculated using the Nelson-Aalen estimator (Ishwaran et al. 2008, 2009):

$$H(t_r) = \sum_{t_{k,r} \leq t} \frac{E_{k,r}}{N_{k,r}} \tag{1}$$

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. To better understand this notion of cumulative hazard estimate, we briefly present the basic functions of survival analysis.

3.2.2 Basic functions of survival analysis

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:

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t, T \geq t)}{\Delta t} \tag{2}$$

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:

$$S(t) = \Pr(T \geq t) \tag{3}$$

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, we can then calculate the value of the survivor function at t by integrating the hazard function from 0 to t:

$$S(t) = exp\left\{- \int_0^t h(\tau) d\tau \right\} \tag{4}$$

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.

$$H(t) = \int_0^t h(\tau) d\tau \tag{5}$$

In non-parametric analysis, the value of the cumulative hazard 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 varies in empirical applications. Therefore, the values of the hazard function are erratic, and identifying a specific pattern is difficult. It follows from Equations 4 and 5 that the survivor function can easily be calculated using the cumulative hazard function:

$$S(t) = exp\{-H(t)\} \tag{6}$$

Therefore, when performing an RSF analysis, after calculating the cumulative hazard function (CHF) from (1), the survivor function is estimated using the Kaplan-Meier estimator:

$$S(t_r) = \prod_{t_{k,r} \leq t} \left(1 - \frac{E_{k,r}}{N_{k,r}}\right) \tag{7}$$

All episodes (or individuals) within r have the same CHF and 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):

$$H(t|X) = H_r, \qquad \mathrm{S}(t|X) = S_r, \qquad \text{if } X \in r \tag{8}$$

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

$$\bar{H}(t|X) = \frac{1}{N}\sum_{n=1}^{N} H_n(t|X), \qquad \bar{S}(t|X) = \frac{1}{N}\sum_{n=1}^{N} S_n(t|X) \tag{9}$$

where H_n is the *n*th survival tree with N trees.

3.3 Model parameters

To predict the event of having a first, second, and third birth we grow separate forests for each of these outcomes. For all outcomes, we opt for the default number of trees: *ntree* = 1000.[8] Similarly, to specify the number of candidate variables, *mtry*, we used the default setting where *mtry* is equal to the square root of the total number of features. The number of split points considered for each variable is also given by *nsplit*; we use *nsplit* = 10. We use a log-rank splitting rule,[9] i.e., the random splitting rule, which is the default option.

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

[9] A splitting rule needs to be defined, which can be *logrank*, *logrankscore,* or *random*. The first two rely on the log-rank-score statistic, and both 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

3.4 Assessment of predictive performance

To assess the performance of the RSFs in predicting births, we calculate the 'Out-of-Bag' (OOB) error rate 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 the 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.

We also assess the performance of the RSF (i.e., goodness of fit) at different survival times. Specifically, we plot the ROC curve at four points in time: at the individuals' ages of 20, 30, 40, and 50. The Area Under the Curve (AUC) tells us how well we can classify individuals into two groups: those who experience the outcome of interest and those who do not. AUC ranges 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.

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 prediction's concordance index (c-index) over time. The c-index gives the probability of concordance between the predicted and the observed survival.

3.5 Assessing the importance of variables in predicting outcomes

To assess the importance of variables in predicting births, we use both Variable Importance (VIMP) (Breiman 2001) and Minimal Depth (Ishwaran et al. 2010; Ishwaran et al. 2011) methods. VIMP for a variable is the difference between the prediction error when the variable is randomly permuted and the prediction error under the observed values. Therefore, a large VIMP value indicates that misspecification detracts from the

single split point is randomly chosen in each variable, and the largest log-rank statistic decides the choice of split.

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 mis-specified. Therefore, we ignore variables with negative and near-zero values of VIMP and rely on the variables with large positive values. Minimal Depth 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 predictors have smaller minimal depth values.

3.6 Assessing response dependency

Once we have identified the most important predictors, we examine how these variables are related to the outcome. This is done using partial dependence plots, which are generated by integrating out the effects of variables other than the covariate of interest. We choose to report partial dependence plots where the models are adjusted for gender, family size, and religiosity.

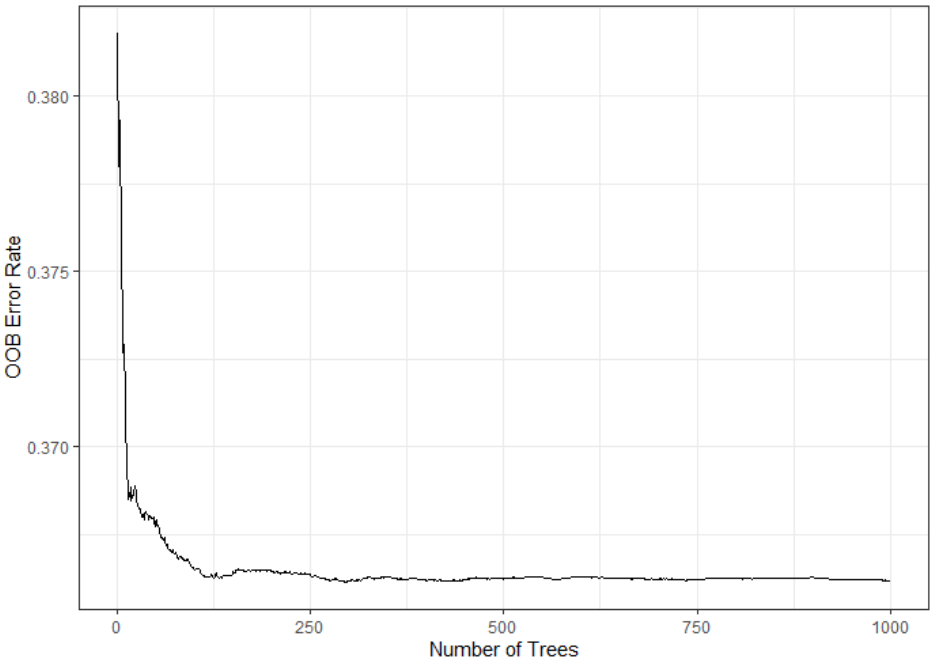
4. Results

4.1 First birth

4.1.1 Assessing predictive performance

Our RSF predicted first birth with an OOB error rate of 36%, while the c-index was 0.65, suggesting that it does a good job in predicting individuals' parenthood status. When plotting the value of the OOB error rate according to the number of trees in the forest (Figure 2), we can see that the OOB error rate stabilizes at around 125 trees to a value of around 36%.

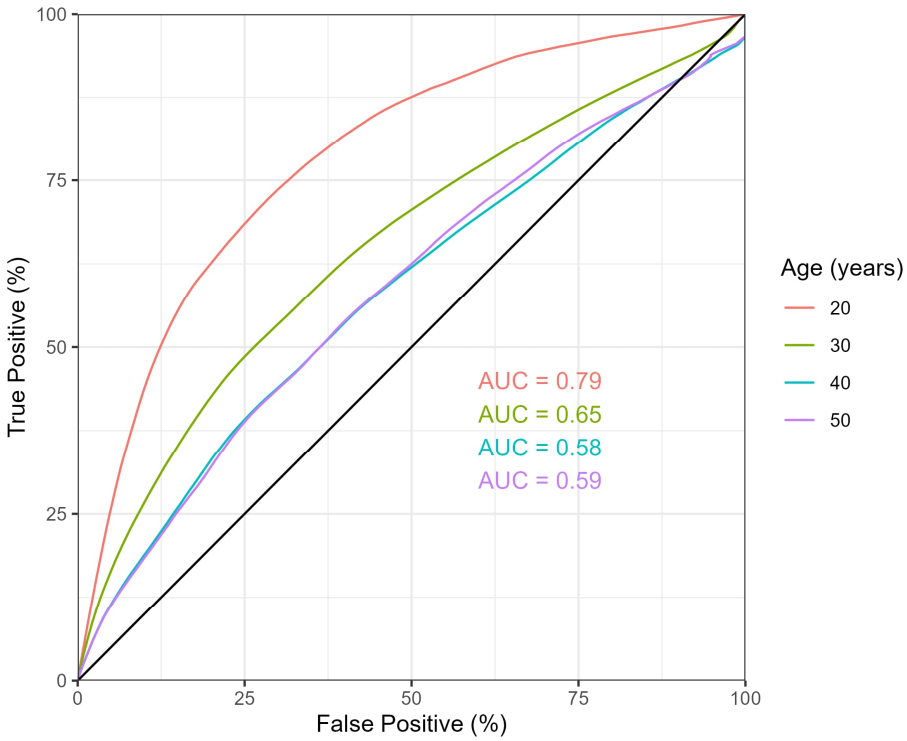
Figure 2: Out of Bag Errors (OOB)



Source: Trajectories and Origins, authors' own calculations.

Figure 3 displays the ROC curves at different surviving times. They demonstrate that the algorithm has a moderate-to-good discriminative ability over the life course, with AUCs ranging from around 0.8 to 0.6 for ages 20 to 50.

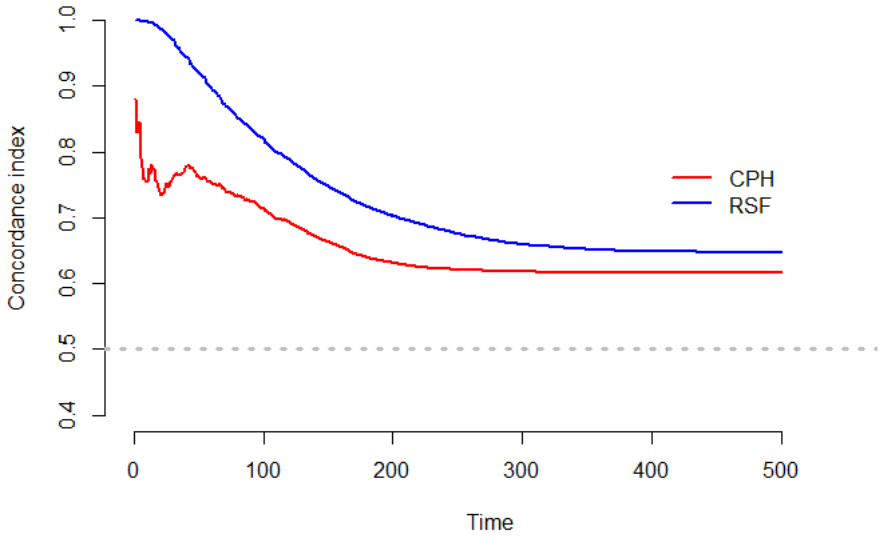
Figure 3: ROC curves at different surviving times



Source: Trajectories and Origins, authors' own calculations.

Figure 4 shows that RSF consistently outperforms the Cox proportional hazards model (CPH) in predicting first births across the life course. RSF achieves substantially higher concordance early on (up to 0.2 above CPH), though the advantage decreases over time, and both models converge to similar performance at later ages.

Figure 4: Comparison of c-indexes



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

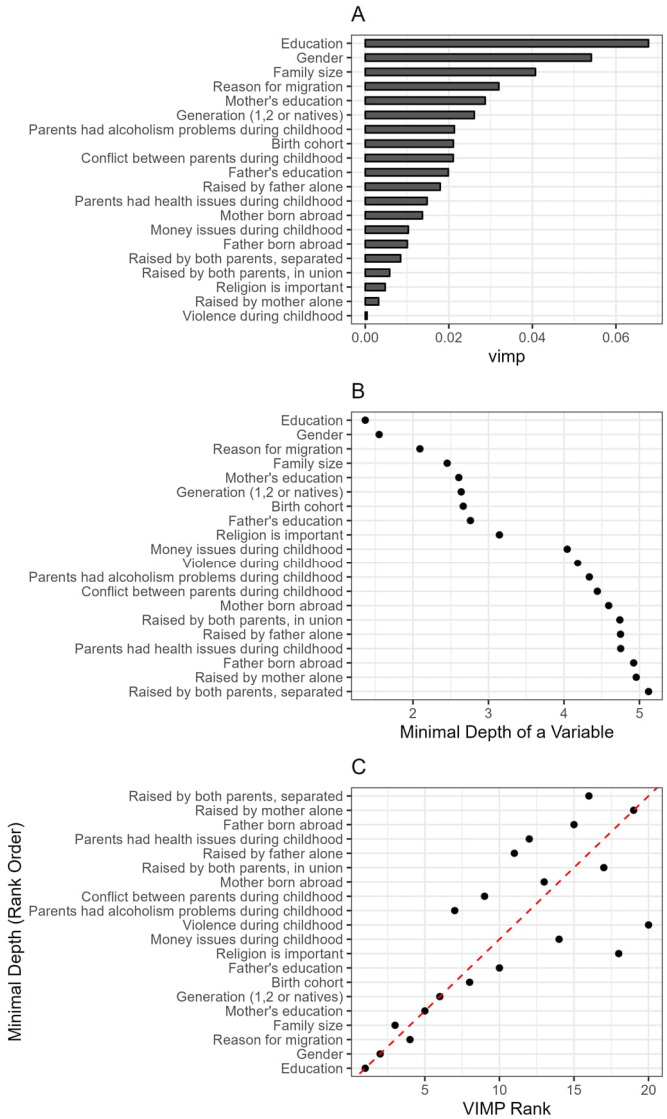
4.1.2 Variable selection

Using the VIMP method, we find that the most important predictor of having a first birth is the level of education, followed by gender and family size, with VIMPs of around 0.07, 0.05, and 0.04, respectively (Figure 5a). The other predictors had VIMPs ranging from around 0 to 0.03.

Using minimal depth, the results show that the most important features are now education, gender, and reason for migration (Figure 5b). Since the VIMP and Minimal Depth measures use different criteria, it is not surprising that the variable ranking tends to be somewhat different.

A comparison of the rankings of Minimal Depth and VIMP (Figure 5c) indicates that both measures are largely in agreement regarding the predictors with higher VIMP; however, the level of agreement declines with decreasing VIMP.

Figure 5: Random forest variable selection – first birth

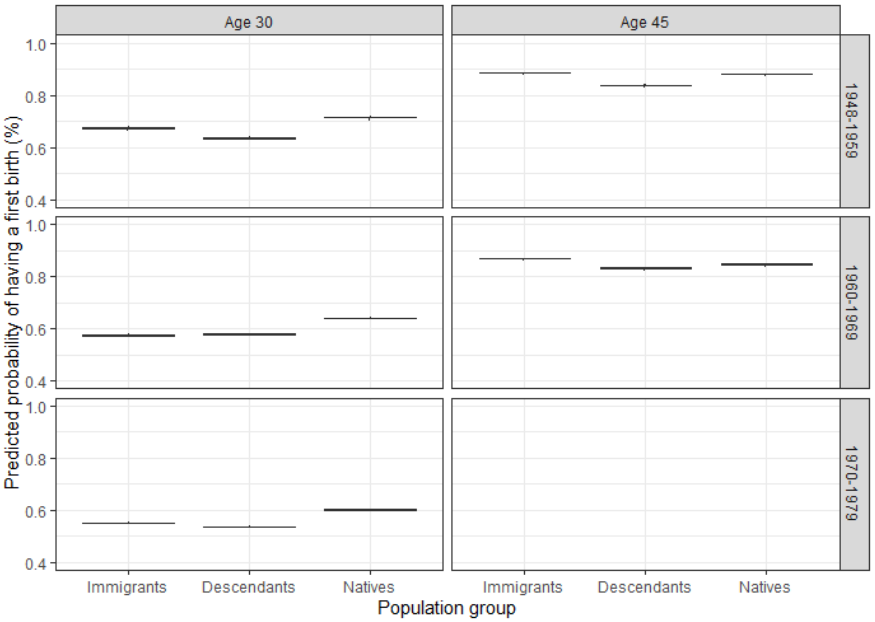


Source: Trajectories and Origins, authors' own calculations.
Notes: In (a) we present the results of Variable Importance (VIMP). Importance is relative to length of bars. In (b) we present the results using Minimal Depth. Low minimal depth indicates important variables. Lastly, in (c) we compare the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

4.1.3 Assessing response dependency

Next, we focus on birth cohort, education, and migrant generation, which all are important variables. We start with a three-way interaction and examine the differences in fertility behaviour between immigrants, their descendants, and natives at two points in time: by the ages of 30 and 45 and across birth cohorts. This also allows us to check if the proportionality assumption holds. The results (Figure 6) show only small differences between migrants and natives in the predicted probabilities of having a child. The differences are also stable across birth cohorts. Overall, natives have a slightly higher probability of having a first birth by the age of 30 compared to immigrants and their descendants, whereas immigrants are more likely, although only marginally, to have a first birth by the age of 45 compared to their descendants and natives. Furthermore, individuals in more recent cohorts have a lower probability of having a first birth, especially by age 30, suggesting the postponement of childbearing.

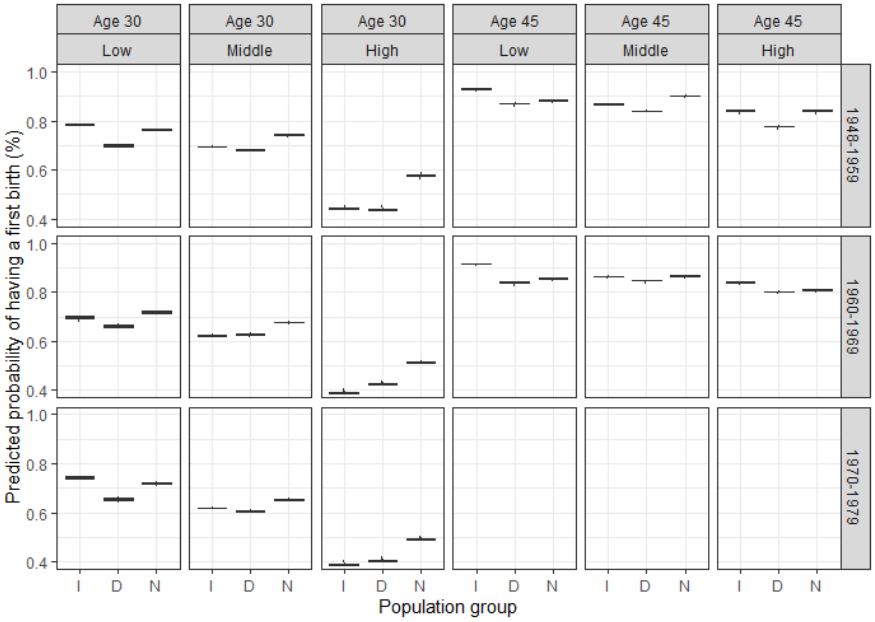
Figure 6: Predicted probabilities of a first birth by age 30 and age 45, by migrant generation and birth cohort



Source: Trajectories and Origins, authors' own calculations.
Note: The figure displays a partial dependence plot where the model has been adjusted for gender, family size, religiosity, and education. 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 as they have not yet reached this age.

Next, we examine a four-way interaction by also including education (Figure 7). Interestingly, if we focus on the likelihood of having a first birth by the age of 30, among low-educated individuals the patterns do not differ considerably between immigrants and natives: Only immigrants' descendants have slightly lower first-birth levels. By contrast, among highly educated individuals, both immigrants and their 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 parenthood by the age of 45, we see that among low-educated individuals, immigrants are more likely to have a first birth compared to their descendants and natives. By contrast, among highly educated individuals, immigrants' fertility differentials are reduced. The likelihood of highly educated immigrants having a child by age 45 is much more similar to that of natives than that of low-educated immigrants.

Figure 7: Predicted probabilities 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.
Note: The figure displays a partial dependence plot where the model has been adjusted for gender, family size, and religiosity. 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 yet reached this age.

4.2 Second birth

4.2.1 Variable selection

Next, we predict the event of having a second birth among parents of one child and examine the importance of a number of predictors. Family size, mother's education, reason for migration, conflict between parents during childhood, and whether the mother was born abroad are found to be the most important features to predict whether individuals have a second birth, with VIMPs ranging from 0.017 to 0.022. Reason for migration, family size, and religion are important and education has the lowest minimal depth, with values ranging from 1.6 to 2.3. The correlation of VIMP with minimal depth is not as good as for first births, and displays no clear pattern by VIMP.

4.2.2 Assessing response dependency

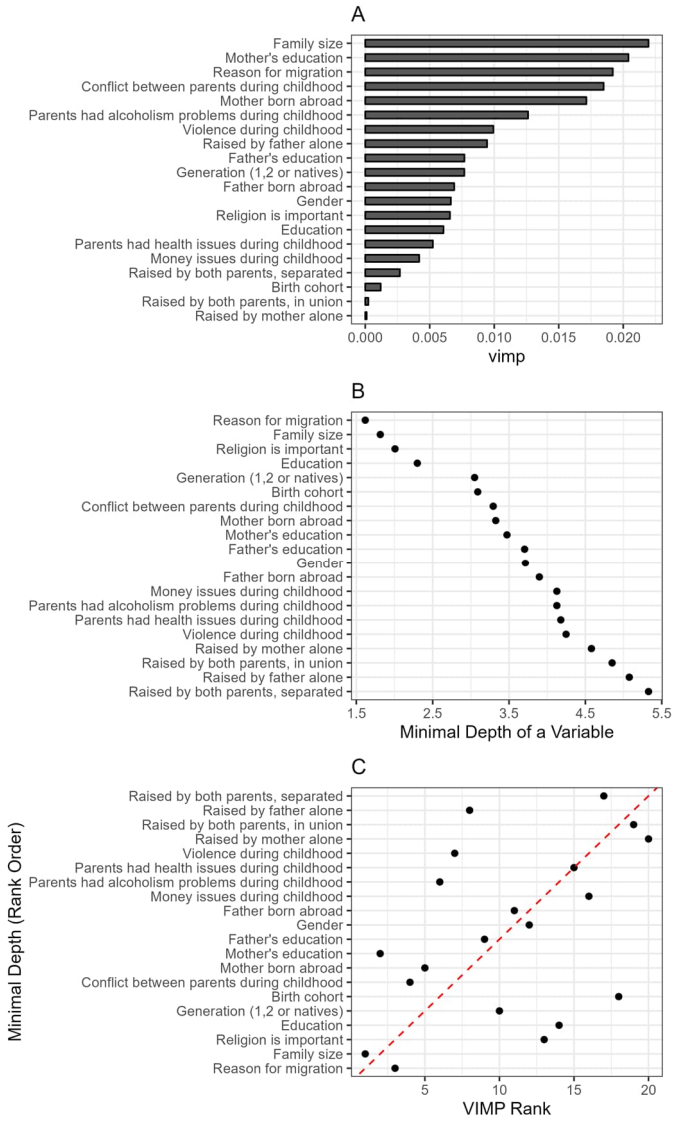
We examine the probability of having a second birth by migrant generation, birth cohort, and educational level at 5 and 10 years after the first birth (Figure 9). Among individuals with low levels of education, immigrants are more likely to have a second birth compared to descendants of immigrants and 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 their descendants. Most importantly, the group differences are reduced for highly educated individuals.

4.3 Third birth

4.3.1 Variable selection

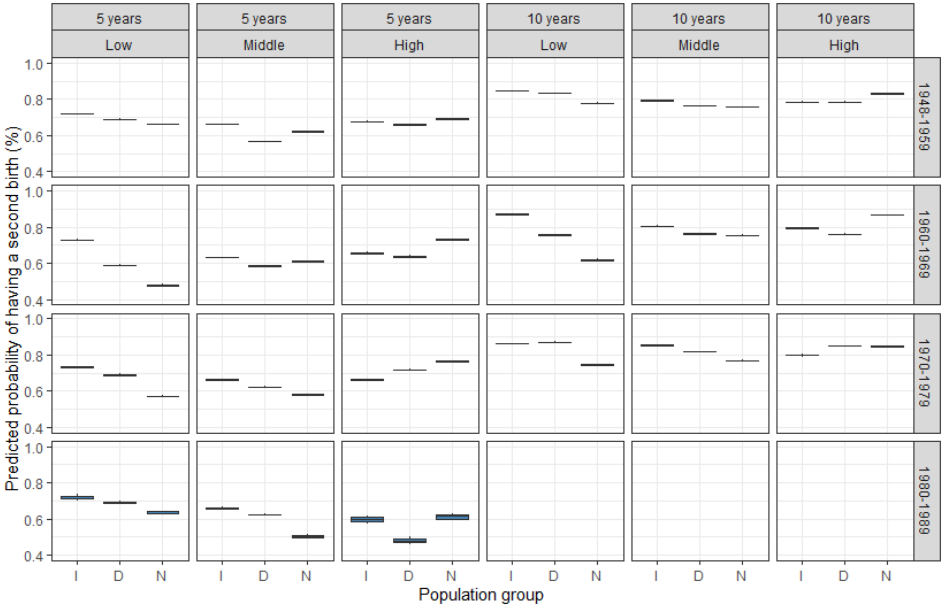
Finally, we predict the event of having a third child among individuals who have two children and identify the most important predictors of having a third child (Figure 10). Family size is found to be the most important predictor of having a third birth and it is also the predictor with the lowest minimal depth. As was the case for second births, the correlation of VIMP with minimal depth is weak.

Figure 8: Random forest variable selection – second birth



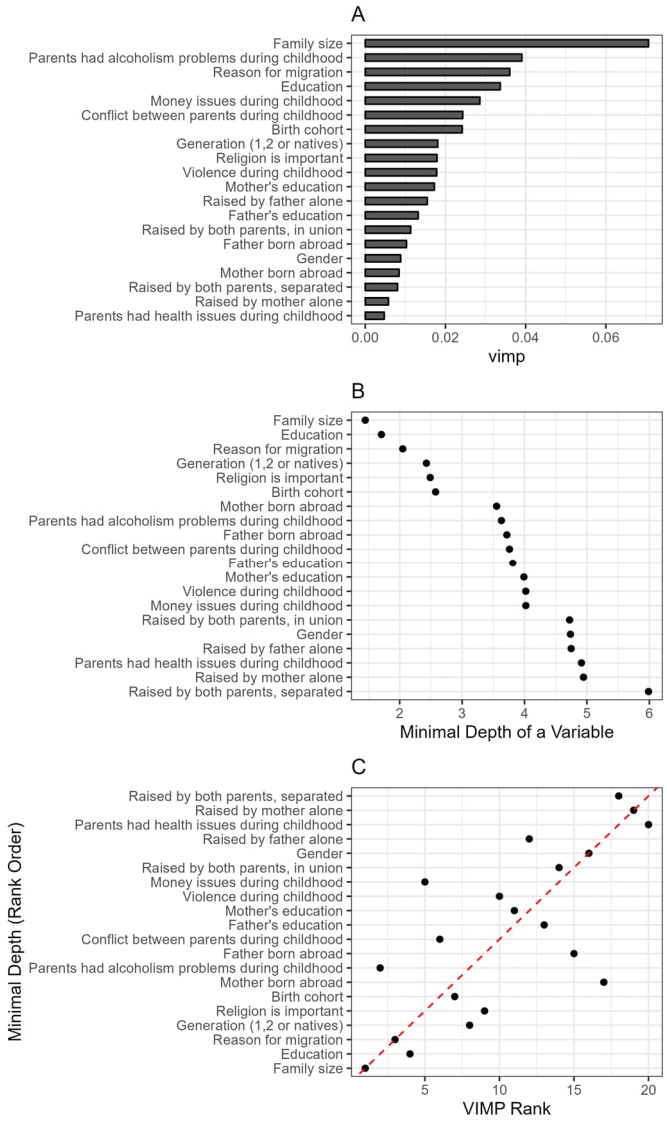
Source: Trajectories and Origins, authors' own calculations.
Note: We use the VIMP method to identify the most important predictors of having a second birth. Importance is relative to the length of the bars (with positive VIMP values).

Figure 9: Predicted probabilities 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.
Note: The figure shows a partial dependence plot where the model has been adjusted for gender, family size, and religiosity. 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 1980–1989 10 years after the first birth since they have not yet reached this stage.

Figure 10: Random forest variable selection – third birth

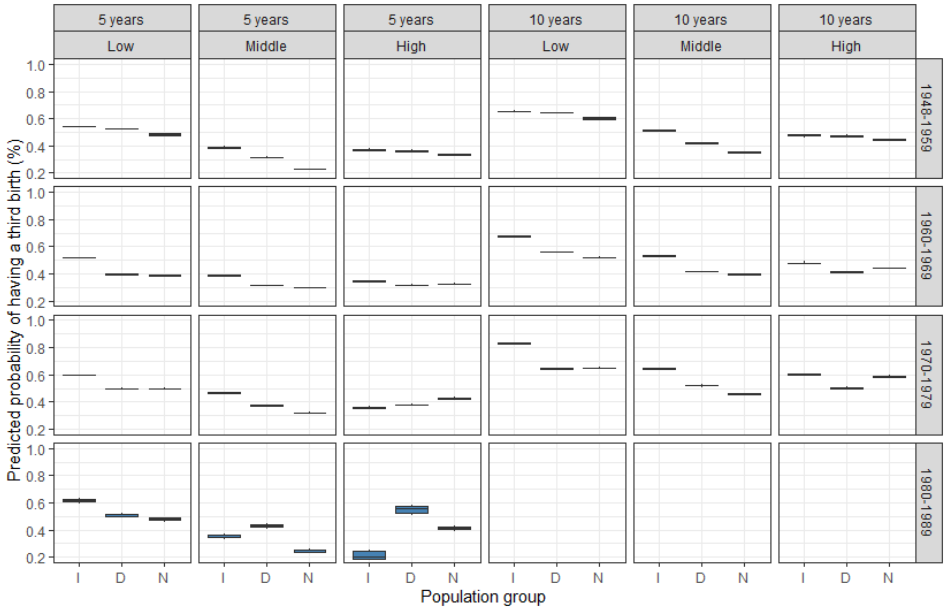


Source: Trajectories and Origins, authors' own calculations.
Note: We use the VIMP method to identify the most important predictors of having a third birth. Importance is relative to the length of the bars.

4.3.2 Assessing respondence dependency

We examine the probability of having a third birth by migrant generation, birth cohort, and educational level at 5 and 10 years after the second birth (Figure 11). Immigrants are more likely to have a third birth, irrespective of the birth cohort and level of education. Again, the differences between immigrants, descendants, and natives in third-birth probabilities are reduced among highly educated individuals.

Figure 11: Predicted probabilities 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.
Note: The figure shows a partial dependence plot where the model has been adjusted for gender, family size, and religiosity. 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 1980–1989 for the period 10 years after the second birth since they have not yet reached this stage.

5. Discussion

This paper applies RSF to predict 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 important findings in relation to the fertility behaviour of immigrants. We find that fertility differences between immigrants and natives are smaller among highly educated individuals compared to those with low education. The study shows that highly educated migrants are similar to natives in their childbearing behaviour. This is a novel finding: While previous research has shown similarity in the first birth rates of descendants of immigrants and natives (Krapf and Wolf 2016), we demonstrate that this similarity is already observed for immigrants and for all three parities. Although our findings apply to immigrants as a group and there is still some variation across subgroups, this variation is significantly reduced among highly educated immigrants.

Our results show that RSF can be a useful method to analyse individuals' fertility behaviour. First, the technique allows us to assess the predictive importance of many covariates. Although it is possible to identify the determinants of fertility behaviour using conventional methods, only a limited number of potential determinants can be included. By contrast, RSF allows us to assess the importance of a high number of predictors. Our results show that education is the most important predictor of a first birth. Family size is the most important predictor of a second and third birth. Second, the method is ideally suited to detect interaction effects. We are able to analyse interactions of more than 2 variables, which can become complicated when relying on conventional methods.

However, although RSF allows us to overcome some of the issues that conventional methods of survival analysis face, the RSF technique also suffers from shortcomings (Salganik et al. 2020; Garip 2020). The most notable drawback of RSF and machine learning techniques in general is that the models are 'black boxes' that can be hard to understand (Best et al. 2022; Salganik et al. 2020; Garip 2020). Still, RSF may represent a suitable tool for exploratory analysis of survival or time-to-event data where previous knowledge is limited. Our application of RSF to the analysis of immigrant fertility behaviour shows that the method can easily be applied in life course research and that research on migrant fertility should pay more attention to how education shapes childbearing patterns among minority populations.

6. Acknowledgments

This paper is part of the MigrantLife Project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement number 834103). We express our gratitude to

three reviewers and the Associate Editor of Demographic Research for their valuable comments and suggestions.

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* 18(1): 129. doi: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. doi: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. doi: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(30): 897–938. doi:10.4054/DemRes.2007.17.30.

Arpino, B., Le Moglie, M., and Mencarini, L. (2021). What tears couples apart: A machine learning analysis of union dissolution in Germany. *Demography* 59(1): 161–186. doi:10.1215/00703370-9648346.

Baudin, T. (2015). Religion and fertility: The French connection. *Demographic Research* 32(13): 397–420. doi:10.4054/DemRes.2015.32.13.

Berghammer, C. (2009). Religious socialisation and fertility: Transition to third birth in the Netherlands/Socialisation religieuse et fécondité: L'arrivée du troisième enfant aux Pays-Bas. *European Journal of Population/Revue européenne de Démographie* 25: 297–324. doi:10.1007/s10680-009-9185-y.

Best, K.B., Gilligan, J.M., Baroud, H., Carrico, A.R., Donato, K.M., Ackerly, B.A., and Mallick, B. (2021). Random forest analysis of two household surveys can identify important predictors of migration in Bangladesh. *Journal of Computational Social Science* 4(1): 77–100. doi:10.1007/s42001-020-00066-9.

Best, K., Gilligan, J., Baroud, H., Carrico, A., Donato, K., and Mallick, B. (2022). Applying machine learning to social datasets: A study of migration in southwestern Bangladesh using random forests. *Regional Environmental Change* 22(52): 1–12. doi:10.1007/s10113-022-01915-1.

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. doi:10.1007/s10680-005-5549-0.

Breiman, L. (2001). Random forests. *Machine Learning* 45: 5–32. doi:10.1023/A:1010933404324.

Breiman, L., Friedman, J., Olshen, R.A., and Stone, C.J. (1984). *Classification and regression trees*. Belmont, CA: Thomson Wadsworth.

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. doi:10.1080/02664763.2017.1414166.

Cleves, M., Gutierrez, M., Gould, W., and Marchenko, Y. (2010). *An introduction to survival analysis using Stata*. Third Edition. College Station: Stata Press.

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. doi:10.1023/a:1005844818027.

Delaporte, I. and Kulu, H. (2022). Interaction between childbearing and partnership trajectories among immigrants and their descendants in France: An application of multichannel sequence analysis. *Population Studies* 77(1): 55–70. doi:10.1080/00324728.2022.2049856.

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

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

Erman, J. (2022). Cohort, policy, and process: The implications for migrant fertility in West Germany. *Demography* 59(1): 221–246. doi:10.1215/00703370-9629146.

Fawagreh, K., Gaber, M.M., and Elyan, E. (2014). Random forests: from early developments to recent advancements. *Systems Science & Control Engineering* 2(1): 602–609. doi:10.1080/21642583.2014.956265.

Garip, F. (2020). What failure to predict life outcomes can teach us. *Proceedings of the National Academy of Sciences* 117(15): 8234–8235. doi:10.1073/pnas.2003390117.

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). doi:10.2427/12663.

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. doi:10.1097/JU.0000000000000416.

Hays, J.J. and Guzzo, K.B. (2022). Does sibling composition in childhood contribute to adult fertility behaviors? *Journal of Marriage and Family* 84(1): 53–79. doi:10.1111/jomf.12788.

Ho, T.K. (1995). Random decision forests. In: *Proceedings of 3rd international conference on document analysis and recognition, Volume 1*. Montreal, QC: IEEE: 278–282. doi:10.1109/ICDAR.1995.598994.

Ho, T.K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20(8): 832–844. doi:10.1109/34.709601.

Hsich, E., Gorodeski, E.Z., Blackstone, E.H., Ishwaran, H., and 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. doi:10.1161/CIRCOUTCOMES.110.939371.

Ishwaran, H. (2007). Variable importance in binary regression trees and forests. *Electronic Journal of Statistics* 1: 519–537. doi:10.1214/07-EJS039.

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. doi:10.1093/biostatistics/kxu010.

Ishwaran, H., Kogalur, U.B., Blackstone, E.H., and Lauer, M.S. (2008). Random survival forests. *Annals of Applied Statistics* 2(3): 841–860. doi:10.1214/08-AOAS169.

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. doi:10.1198/jasa.2009.tm08622.

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. doi:10.1002/sam.10103.

Ishwaran, H. and Kogalur, U.B. (2008). RandomSurvivalForest 3.2. 2. R package. doi:10.1214/08-AOAS169.

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

Jiang, S. (2019). Prediction based on Random Survival Forest. *American Journal of Biomedical Science and Research* 6(2). doi:10.34297/AJBSR.2019.06.001005.

Kashyap, R., Rinderknecht, R.G., Akbaritabar, A., Alburez-Gutierrez, D., Gil-Clavel, S., Grow, A., Kim, J., Leasure, D.R., Lohmann, S., Negraia, D.V., Perrotta, D., Rampazzo, F., Tsai, C.-J., Verhagen, M.D., Zagheni, E., and Zhao, X. (2022). Digital and computational demography. SocArXiv. doi:10.31235/osf.io/7bvpt.

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 and Prevention* 144: 105683. doi: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: Hank, K. and Kreyenfeld, M. (eds.). *Social Demography – Forschung an der Schnittstelle von Soziologie und Demographie*. Wiesbaden: Springer VS: 137–164. doi:10.1007/978-3-658-11490-9_7.

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. doi:10.1007/s10680-014-9322-0.

Kulu, H. and Hannemann. T. (2016a). Introduction to research on immigrant and ethnic minority families in Europe. *Demographic Research* 35(2): 31–46. doi: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(49): 1441–1488. doi:10.4054/DemRes.2016.35.49.

Kulu, H., Hannemann, T., Pailhé, A., Neels, K., Krapf, S., González-Ferrer, A., and Andersson, G. (2017). Fertility by birth order among the descendants of immigrants in selected European countries. *Population and Development Review* 43(1): 31–60. doi:10.1111/padr.12037.

Kulu, H. and Milewski, N. (2007). Family change and migration in the life course: An introduction, *Demographic Research* 17(19): 567–590. doi:10.4054/DemRes.2007.17.19.

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(46): 1345–1374. doi:10.4054/DemRes.2019.40.46.

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: Lacković, I. and Vasic, D. (eds.). *6TH European conference of the international federation for medical and biological engineering*. Cham: Springer: 740–743. doi:10.1007/978-3-319-11128-5_184.

Milewski, N. (2007). First child of immigrant workers and their descendants in West Germany: Interrelation of events, disruption, or adaptation? *Demographic Research* 17(29): 859–896. doi:10.4054/DemRes.2007.17.29.

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. doi:10.1007/s10680-010-9211-0.

Mussino, E. and Cantalini, S. (2022). Influences of origin and destination on migrant fertility in Europe. *Population, Space and Place* 28(7): e2567. doi:10.1002/psp.2567.

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. doi:10.1007/s10680-012-9261-6.

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(45): 1361–1398. doi: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. doi:10.18502/jbe.v6i1.4760.

Rojas, E.A.G., Bernardi, L., and Schmid, F. (2018). First and second births among immigrants and their descendants in Switzerland. *Demographic Research* 38(11): 247–286. doi:10.4054/DemRes.2018.38.11.

Salganik, M.J., Lundberg, I., Kindel, A.T., Ahearn, C E., Al-Ghoneim, K., Almaatouq, A., ... and McLanahan, S. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences* 117(15): 8398–8403. doi:10.1073/pnas.1915006117.

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. doi: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): 20410. doi:10.1038/s41598-020-77220-w.

Taylor, J.M.G. (2011). Random survival forests. *Journal of Thoracic Oncology* 6(12): 1974–1975. doi:10.1097/JTO.0b013e318233d835.

Wang, H. and Li, G. (2017). A selective review on random survival forests for high dimensional data. *Quantitative Bio-Science* 36(2): 85. doi: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* 51(6): 1–36. doi: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. doi: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. doi: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. doi: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. doi:10.1002/widm.1114.

Appendix

Table A-1: Migrants and their descendants, by origin

	Sample (%)		
Origin	1st birth	2nd birth	3rd birth
Native	17	18	17
North Africa	22	22	22
Other Europe	9	10	9
South East Asia	9	8	8
Southern Europe	19	21	21
Sub-Saharan Africa	13	12	12
Turkey	6	7	7
Missing	5	4	4

Note: Percentages may not total 100% due to rounding."

Example of R-code is available at https://github.com/aibbetson/rsf_migrant_fertility.