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

Isaure Delaporte[†] Hill Kulu[‡]

Abstract

The most used techniques to analyse the multiple factors that shape people's lives outcomes are the techniques of multivariate survival data analysis. Yet, these techniques have a number of limitations. The non-parametric methods such as survival trees and tree ensembles are a useful alternative to the classical survival data analysis. This paper aims to illustrate the advantages of random survival forest (RSF) to study the fertility dynamics of immigrants and their descendants. More specifically, we examine the probability of having a first, second and third birth among immigrants and their descendants in the French population using a rich French survey named Trajectories and Origins. We first assess the performance of the algorithm in predicting the event. We then demonstrate random forest variable selection techniques using Variable Importance and Minimal Depth. This allows us to determine which variables are the most important to explain survival. We then examine how and to which extent important variables affect survival and explore potential interaction terms. Our findings justify the robust interpretability and competitive performance of the random survival forest algorithm to study the family dynamics of immigrants and their descendants.

Keywords: Machine Learning, Random Survival Forest, 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

Extended Abstract

An important aim in social sciences is to understand the multiple factors that shape people's lives outcomes. The most used techniques to model different paths of behaviour in relation to time and selected covariates are the techniques of multivariate survival data analysis. Yet, these techniques have some limitations. For instance, they become inadequate in high dimensional settings (Wang and Li 2017; Spooner et al. 2020). Indeed, as the number of covariates increases, the saturation of statistically insignificant covariates can inhibit effect size interpretation (Witten and Tibshirani 2010; Dudoit and Boldrick 2003; Whetten, Stevens and Cann 2021). Similarly, collinearity also jeopardize model interpretability. To address these concerns, conventional methods often relies on some level of subjectivity at the stage of variable selection, which in turn might limit the potential to identify the most important predictors. Besides, many parametric models require the proportional hazards assumption to hold. However, this assumption is often violated. Lastly, conventional methods of survival analysis are not best suited to detect and visualize interactions.

The non-parametric methods such as survival trees and tree ensembles are a useful alternative to the classical survival data analysis (Breiman et al. 1984; Breiman 2001; Ishwaran et al. 2008; Ishwaran and Kogalur 2007, 2008, 2014). However, despite the advantages of these methods, a limited number of studies in demography have used machine learning techniques. De Rose and Pallara (1997) rely on a tree methodology to examine the predictors of marriage formation among adult women in Italy. Billari et al. (2006) also show the usefulness of relying on decision tree learning and classification rules to detect the features that differentiate the transition to adulthood in Austria and Italy. More recently, Arpino, Le Moglie and Mencarin (2020) depart from the strategy of using single trees and apply the technique of random survival forest (RSF) to analyse the determinants of divorce for women entering in a marriage or cohabitation in Germany.[1] Overall, previous studies have stressed the usefulness of machine learning techniques mostly to identify the most important predictors of a specific behaviour. Yet, less attention has been given to the advantage of detecting interaction effects.

We contribute to the existing studies in the following ways. First, this paper uses RSF to study the fertility dynamics of immigrants and their descendants. This field has attracted increasing interest in the demographic life course literature (Kulu and González-Ferrer 2014; Kulu and Hannemann 2016). The results of existing studies indicate that immigrants exhibit higher fertility levels than natives; they start to have children earlier compared to natives while the descendants of immigrants follow more similar patterns to natives. This general finding however hides considerable heterogeneity within migrant groups and along sociodemographic characteristics. For instance, immigrants' fertility differentials differ by origin and age

[1] Apart from this study, RSF has been applied so far mostly in bio-medical settings (Breiman 2001; Fawagreh, Gaber and Elyan 2014; Ishwaran et al. 2008; Wang and Li 2017).

at arrival (Andersson 2004; Pailhé 2015; Kulu and González-Ferrer 2014; Milewski 2010; Andersson and Scott 2007; Kulu and Hannemann 2016a; Kulu et al. 2017). Besides, when we examine individuals with higher levels of education, immigrants' and natives' fertility differentials are significantly reduced (Pailhé 2017; Krapf and Wolf 2016). Therefore, there are important interaction effects. Yet, it is often difficult to detect and visualize these interaction effects 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 detect and better understand interactions between terms.

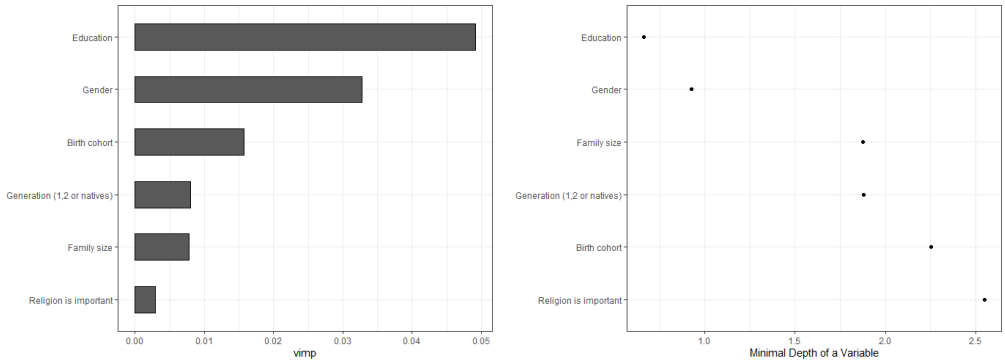
We illustrate the implementation of RSF and discuss the advantages and the limitations of the method. To carry out the analysis, this study uses a rich French survey named Trajectories and Origins which collect detailed information on immigrants, their descendants and French natives. It contains retrospective biographical data on individuals' childbearing histories. It also contains detailed information on individuals' sociodemographic characteristics. This allows us to shed light on the key differences in the fertility patterns between groups. In particular, we focus on childbearing events and examine the probability of having a first, second and third birth. After growing the random forest, we first assess the performance of the algorithm in predicting the event. We compare different models such as Cox proportional hazard regression (CPH) with RSF in order to see how each models' prediction error fluctuates. This allows us to draw some conclusions about the predictive power of the models.

We then demonstrate random forest variable selection techniques using Variable Importance (VIMP) (Breiman 2001) and Minimal Depth (Ishwaran et al. 2010; Ishwaran et al. 2011). We compare both methods to determine which variables are the most important to explain survival. Figure 1 reports the results for the most important predictors of having a first birth. VIMP measures the increase (or decrease) in prediction error for the forest ensemble when a variable is randomly "noised up". More specifically, a large VIMP value indicates that misspecification detracts from the predictive accuracy in the forest. A VIMP close to zero indicates the variable contributes nothing to predictive accuracy, and negative values indicate the predictive accurary 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 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 the birth cohort.

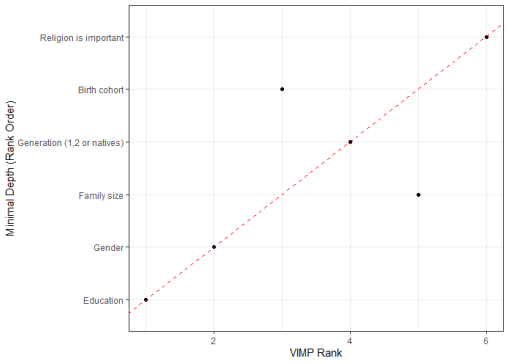
An alternative method to identify the most important predictors is to use minimal depth (Ishwaran 2007; Ishwaran et al. 2010, 2011). This method assumes that variables with high impact on the prediction are those that most frequently split nodes nearest to the root node, where they partition the largest 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. 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 in Figure 1c. The points along the red dashed line indicate where the measures are in agreement. Points above the red dashed line are ranked higher by VIMP than by minimal depth, indicating 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 more the discrepancy between measures.

Figure 1. Random Forest Variable Selection – Probability of Having a First Birth



a) VIMP b) Minimal Depth



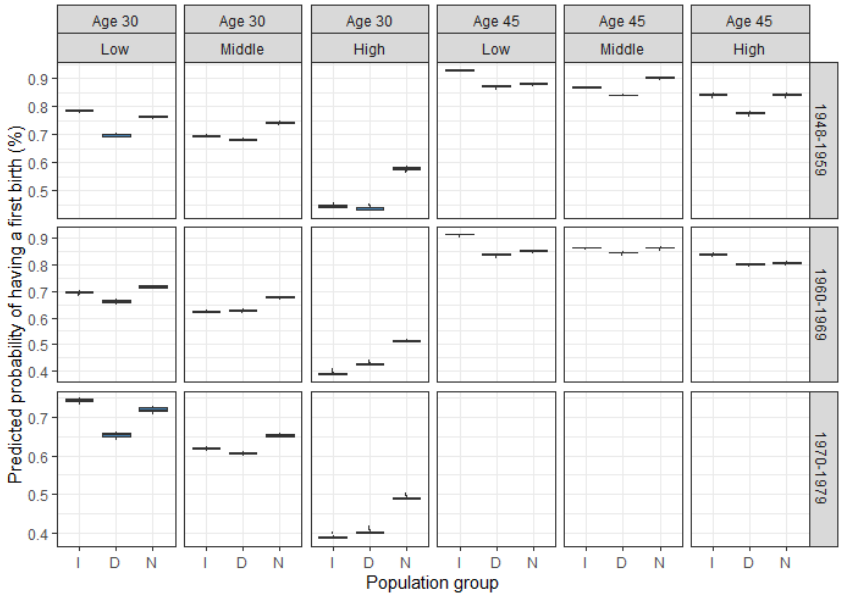
c) Comparison

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.

Once we have identified the most important predictors, we examine how and to which extent important variables affect survival and explore potential interaction terms. An example of this is reported in Figure 2. More specifically, we examine how the predicted probabilities of having a first birth differ across population groups by birth cohort and educational level. Interestingly, if we focus on the probabilities of having a first birth at the age of 30, we can see that in older cohorts, for individuals that have a low level of education, the predicted probabilities do not differ considerably between immigrants, the descendants

and natives. In the opposite, for highly educated individuals born in the 1950s, immigrants and the descendants are less likely to experience having a first birth compared to natives. The pattern differs for more recent cohorts. In particular, among low educated individuals who were born in the 1970s, immigrants are more likely to have a first birth compared to the descendants and natives. By contrast, among more educated individuals born in the 1970s, the natives have a higher probability of having a first birth compared to immigrants and the descendants. If we examine the probabilities at the age of 45, we find similar patterns: 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.

Figure 2. Predicted probability of having a first birth at age 30 and 45, by population group, 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. "I" stands for immigrants, "D" stands for the descendants of immigrants and "N" stands for natives.

Overall, our findings highlight the importance of the composition hypothesis: fertility differentials may vanish as the sociodemographic structure of an immigrant group grows to resemble that of the native population. On a methodological point of view, our results also highlight the advantages of the random survival forest technique. First, the method is ideally suited for detection of conditional information, thus helping to define group of individuals with different survival probability. Variables of both continuous and discrete type can be included since no assumptions are imposed on the covariates' probability structure (De Rose and Pallara 1997). Similarly, issues due to high dimensional data such as overfitting or

multicollinearity do not apply to RSF and this method is robust to outliers as well. When comparing the predictive power of RSF to conventional methods, we find that RSF outperforms the other models in prediction performance of survival. Therefore, our findings justify the robust interpretability and competitive performance of the random survival forest algorithm to study the family dynamics of immigrants and their descendants.