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

Bruno Arpino
Pompeu Fabra University
bruno.arpino@upf.edu

Marco Le Moglie
Bocconi University
marco.lemoglie@unibocconi.it

Letizia Mencarini
Bocconi University
letizia.mencarini@unibocconi.it

BigSurv 2018, Barcelona

Machine learning (ML) in social sciences

- Algorithms that give computers the ability to learn without being explicitly programmed (Samuel 1959)
- Growing interest but limited use in social sciences, in general, and demography, in particular
  - Computational burden and absence of software are loosing importance
  - “Black box” / exploratory approach
- Demographers have used ML for (semi)automatic coding of big data (e.g., Mencarini et al. 2017)
- Our aim is to demonstrate the use of ML as an alternative or complement to standard regression approaches

Application: divorce determinants in Germany (1/2)

- Many and different types of determinants have been examined
  - Personal: age, personality, health and wellbeing, education
  - Economic: income, paid work and housework
  - Couple: n. of children, married vs cohabiting, union duration
- Nonlinearities: e.g., the effect of wife's income (Rogers 2004)
- Interactions: Husband does > 50% housework * wife earns < 50% (Cooke 2006)

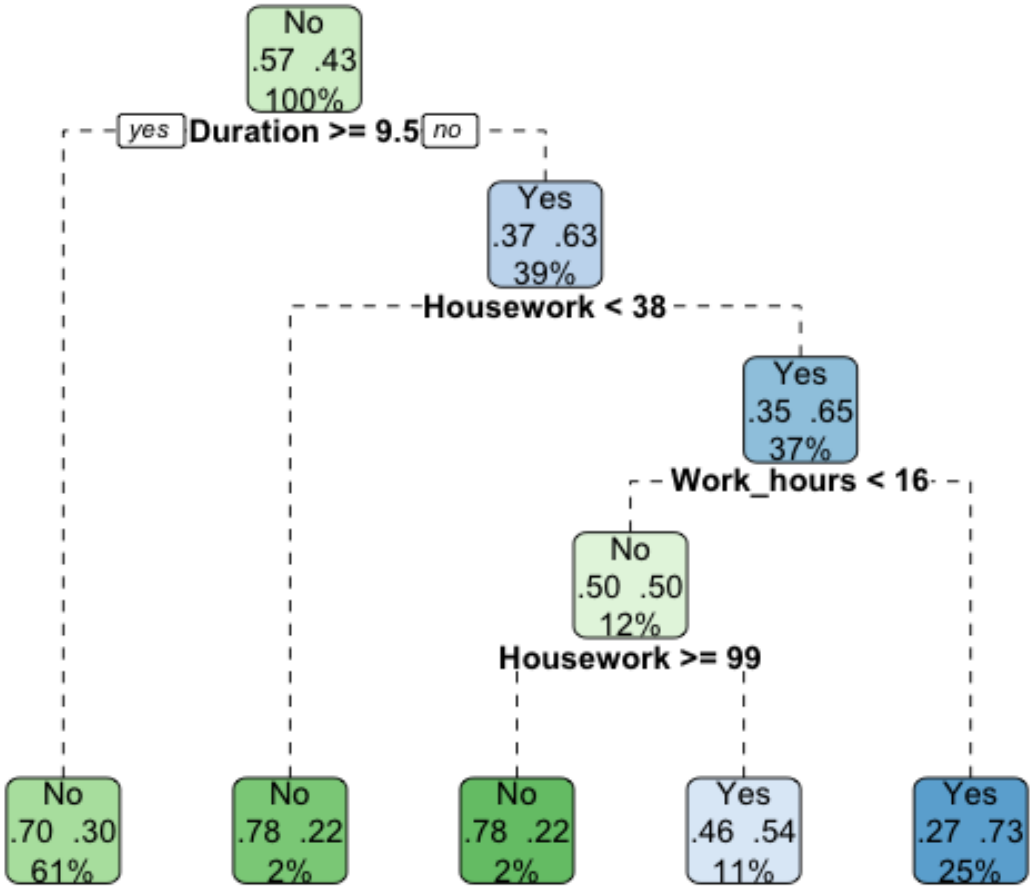
Application: divorce determinants in Germany (2/2)

- We use longitudinal data from the German Socio-Economic Panel survey (SOEP)
- We selected individuals aged up to 65 years and who started their relationship during the observation period (i.e., 1984-2015)
  - Followed until the union ends or until the last available observation
- The final sample consists of 18,613 observations (2,038 couples observed, on average, over 12.6 years).
- 45% of couples split up during the observation period

Classification and Regression trees (CARTs)

- At each step the algorithm iteratively splits the data into subsets
- Splits defined by jointly choosing an independent variable, X, and the value of X that minimizes the prediction error (defined by the sum of squared residuals)
- CARTs capture automatically nonlinearities and interactions
- In Demography, CARTs-based approaches have been used by De Rose and Pallara (1997); Billari et al. (2006)

An example of CART for divorce



Note: Colors represent the most likely outcome at each node while its intensity the level of probability associated to it. The percentage at the bottom of each node provides the part of the sample reaching that node.

Random forests

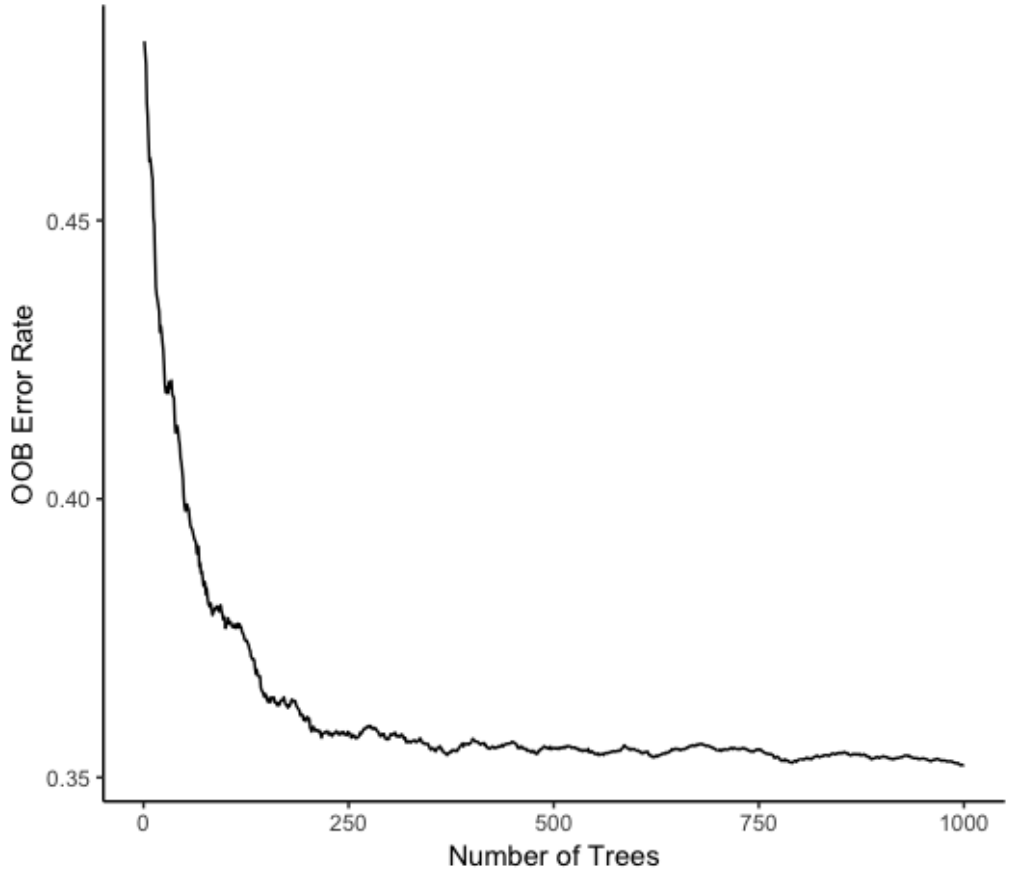
- Belong to the class of "ensemble algorithms" based on (many) multiple trees (Berk, 2006).
  - Random selection features to differentiate the trees
  - Aggregation of results to improve out-of-sample predictions
- A random forest is a multitude of trees that differ because of random selection of both the data (bootstrap sample) and the variables (random selection of few Xs)
- Random forests often outperform other ML algorithms (Breiman, 2001; Glaeser et al 2006)

Random survival forests

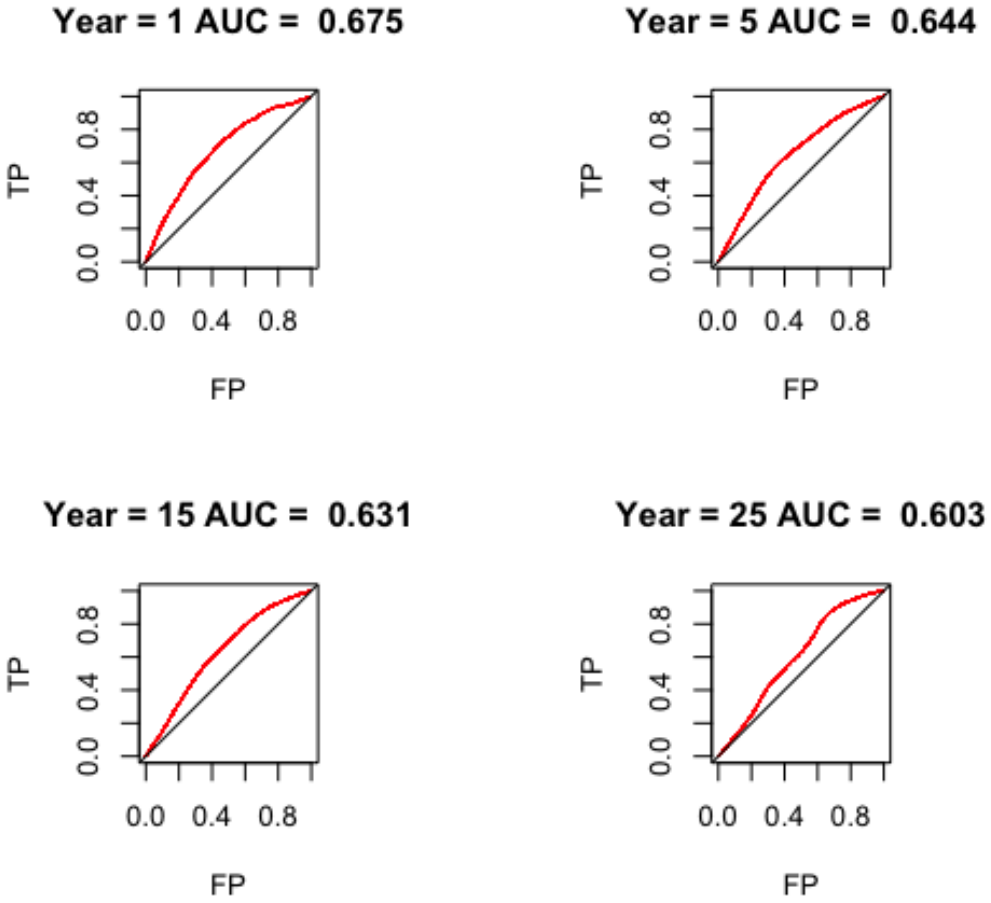
- Standard CARTs and ensemble techniques can be applied to cross-sectional analyses
- More recent developments also for survival data
- Similar to survival models, survival random forests can be used to predict the timing of events
- A good split for a node maximizes survival difference between daughters
- Survival random forests can be implemented using the R package *randomForestSRC* (Ishwaran et al., 2008). We provide replication code

RESULTS

Determining the number of trees

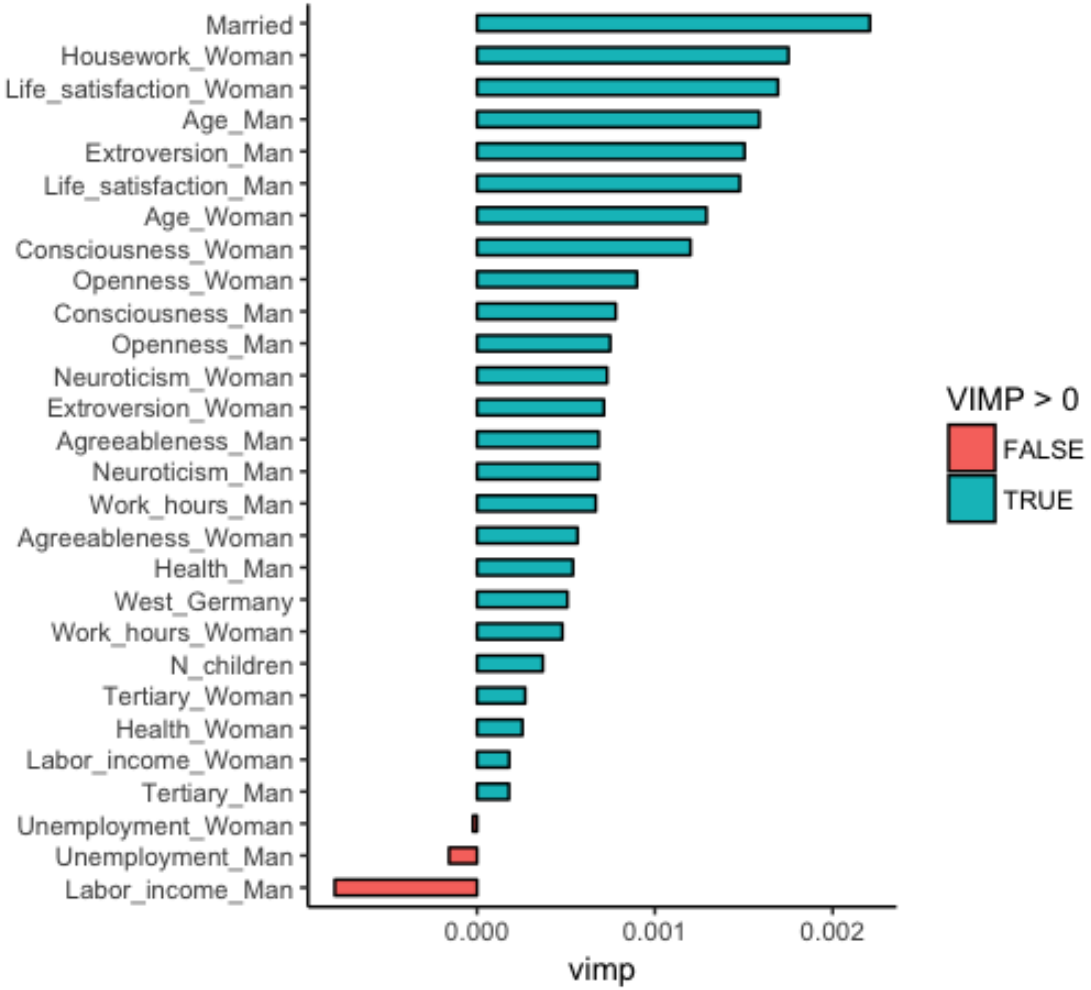


Predictive power



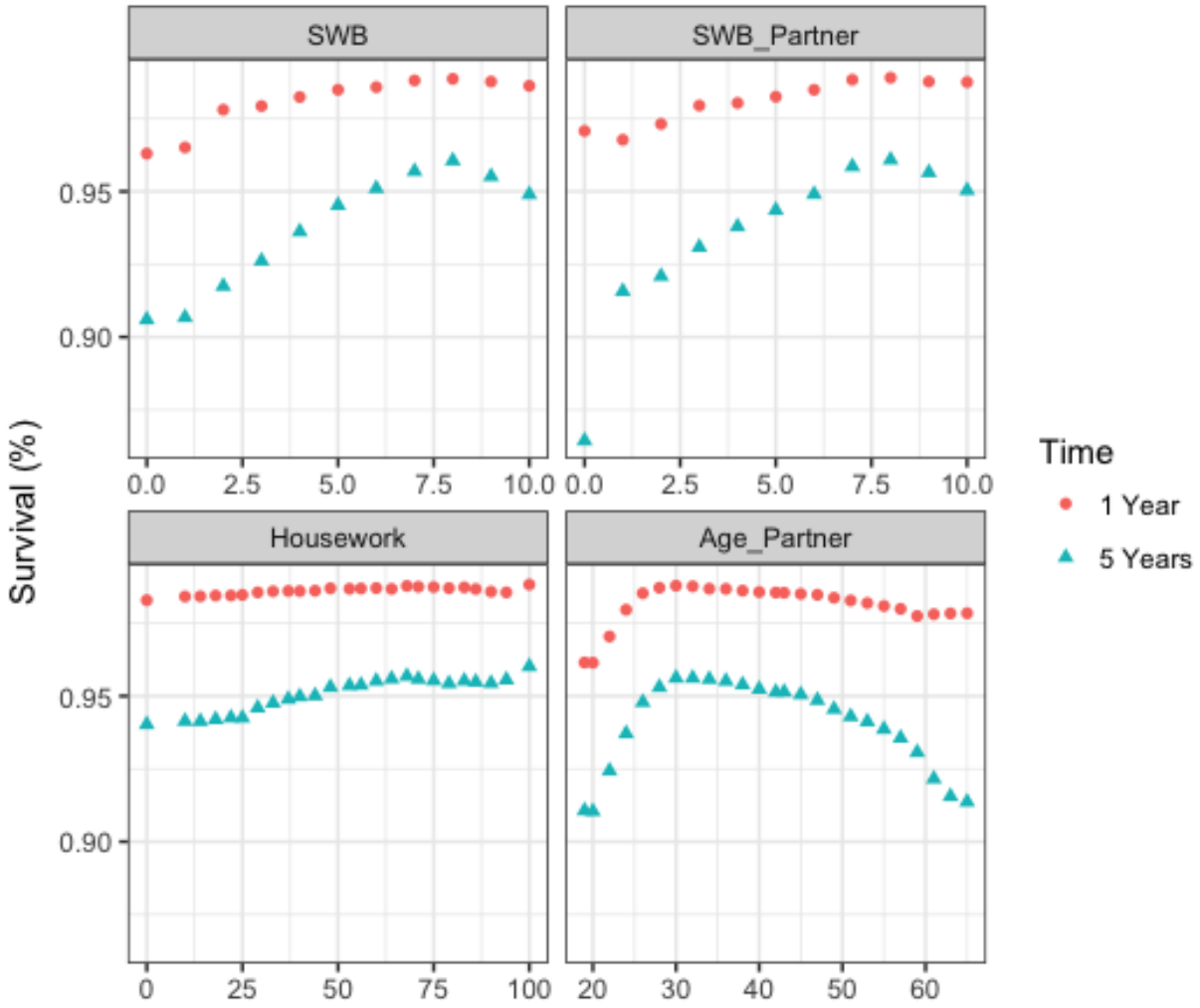
OOB error rate is about 35% (C-index = 0.65)

Variables importance

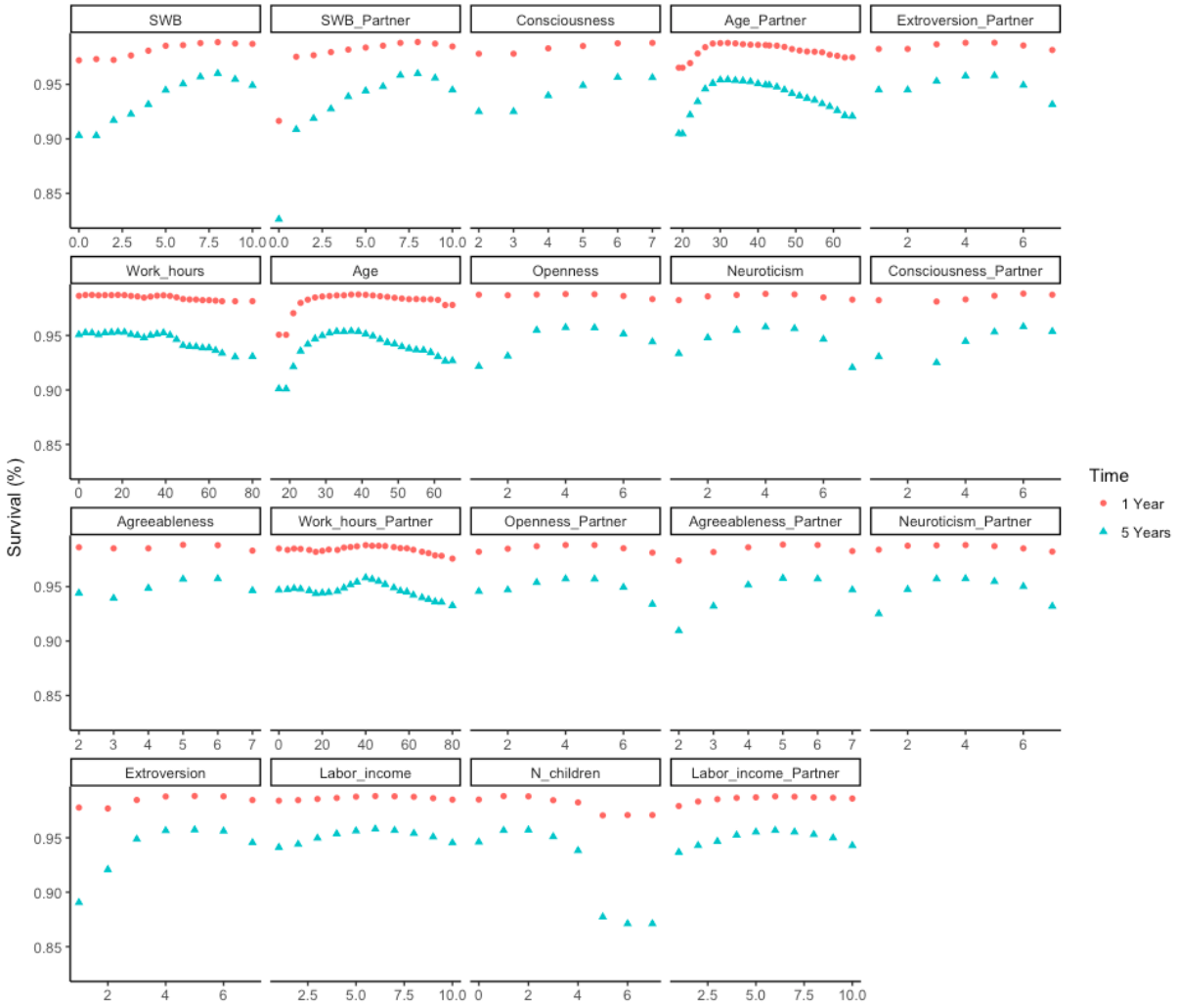


VIMP (X) is the change in prediction error if X were not available

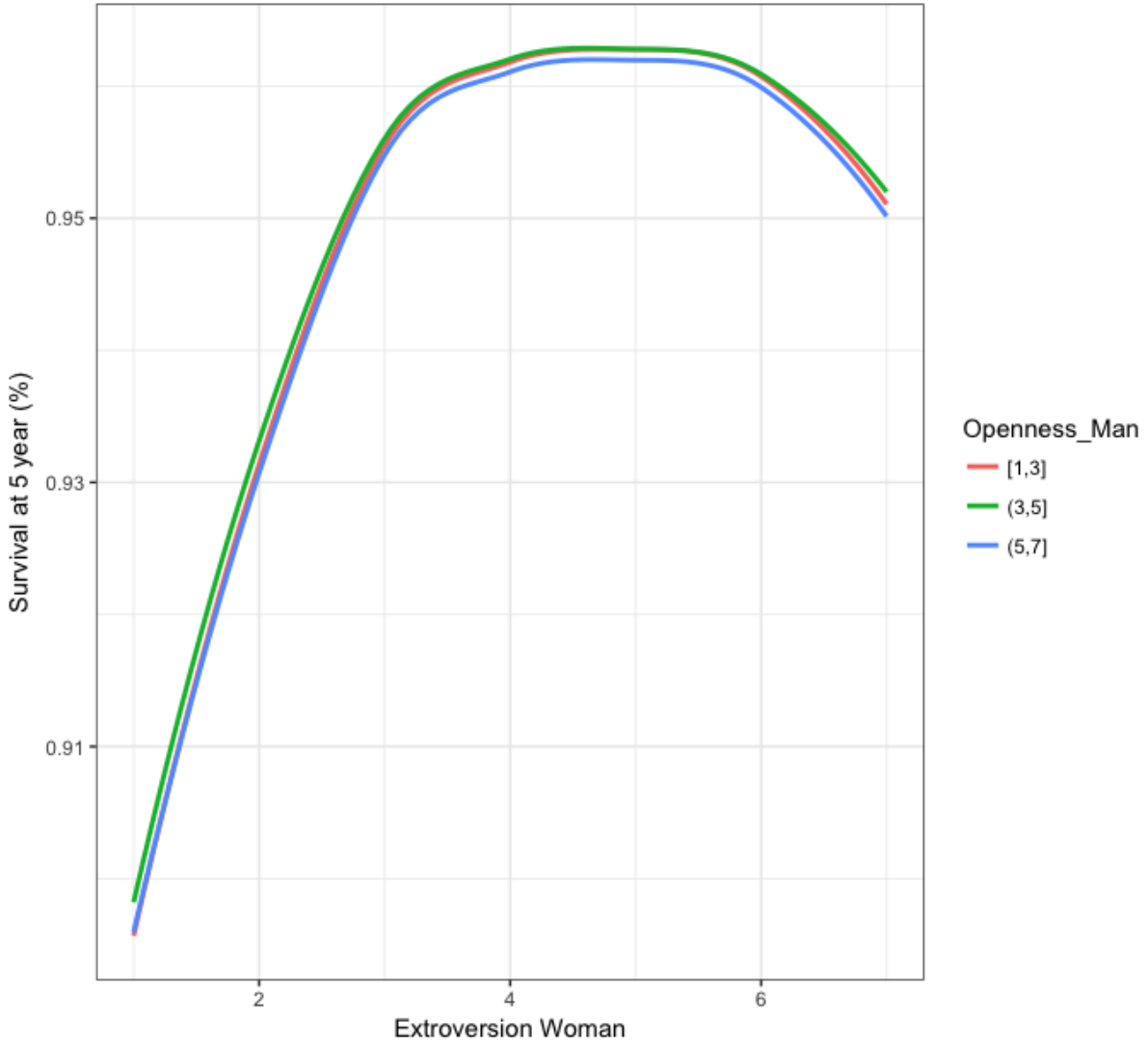
Partial dependence plots (1/2)



Partial dependence plots (2/2)



Partial dependence co-plots



Summary and concluding remarks (1/2)

- Advantages of RFs over parametric regression:
  - Can manage a considerable amount of data and variables
  - Automatic detection of interactions and nonlinearities
  - Collinearity is not an issue
  - Variable importance
  - Partial dependence plots similar to marginal effects plots
  - They can easily handle missing data

Summary and concluding remarks (2/2)

- RFs and similar techniques can be used in substitution or in alternative to parametric regression. Eg.:
  - Compare RF and parametric alternative as a sort of robustness check
  - Use RF for variable selection
  - Use RF to improve model specification (e.g., categorical variables with many levels; numerical variables)
- Machine learning techniques are growing in importance and scope (Athey and Imbens 2016) and demographers may find them useful

Bruno Arpino
Pompeu Fabra University
bruno.arpino@upf.edu

Marco Le Moglie
Bocconi University
marco.lemoglie@unibocconi.it

Letizia Mencarini
Bocconi University
letizia.mencarini@unibocconi.it

BigSurv 2018, Barcelona