Random Forest in Economic Panel Data Analysis: Fundamentals and Methodology

Martín Gabriel Cargnel

May, 2025

Table of contents

Summary

Panel data analysis has historically been approached through classical econometric techniques, such as linear fixed and random effects models. While these methods are widely used for causal inference and control of unobserved heterogeneity, they impose strong functional restrictions and require a priori specification of relationships between variables. In contexts where the data structure is complex and non-linear, these limitations can affect the validity or accuracy of the results. In this framework, machine learning models (particularly tree ensembles like Random Forest) emerge as complementary tools that allow modeling complex relationships without the need to specify rigid functional forms. Although traditionally considered "black box" methods, in the last decade interpretative techniques have emerged that allow extracting useful and comprehensible information for applied researchers. This work aims to contribute to the economic analysis of panel data by introducing a theoretical and practical framework for the use of Random Forest, offering applied economists a flexible and complementary alternative to conventional econometric approaches.

Keywords: Ensemble of trees, Bagging, Random Forest, Panel Data Analysis, Econometrics.

1 Introduction

Panel data, which consists of observations on multiple individuals (such as people, firms, or countries) over multiple time periods, is widely used in econometrics. The inclusion of a temporal dimension in cross-sectional data offers several advantages, such as greater variability than a purely cross-sectional sample, the ability to control for unobservable heterogeneity between individuals, and the opportunity to analyze dynamic effects, conduct event studies, and evaluate policy impacts (Cameron & Trivedi, 2005). However, despite their widespread use in econometrics, panel data methods share two key limitations with the classic linear model: the treatment of interactions and the assumption of linearity.

1.1 Limitations of Classical Panel Data Models

The reason of the limitations rely on the underlying model behind the estimation. If we think about the specification of a linear fixed or random effects models, we can see the following

The **linear fixed effects model** can be written as:

$$y_{it} = \alpha_i + \mathbf{x}'_{it}\beta + \varepsilon_{it}$$

where y_{it} is the outcome variable for individual i at time t, α_i the individual-specific intercept (captures unobserved heterogeneity that is constant over time for each individual), $\mathbf{x}_{it}$ the vector of observed explanatory variables for individual i at time t, β the vector of coefficients associated with $\mathbf{x}_{it}$ and ε_{it} the idiosyncratic error term.

The **linear random effects model** is specified as:

$$y_{it} = \alpha + \mathbf{x}'_{it}\beta + u_i + \varepsilon_{it}$$

where: α: overall intercept (common to all individuals), u_i: individual-specific random effect (assumed to be uncorrelated with $\mathbf{x}_{it}$) and the other terms are as defined above.

In both models, the key distinction is how the individual-specific effect is treated: as a fixed parameter to be estimated for each individual (α_i) in the fixed effects model, or as a random variable (u_i) in the random effects model.

Both specifications assume linearity and independence between variables, assumptions that are often unrealistic in practice. A common workaround, as in classical linear regression, is to manually add transformations (to address non-linearity) and interaction terms. Addressing the non-linearity manually can be problematic because it might rely on subjective choices and can lead to model misspecification and lack of interpretability.

Addressing interactions manually quickly becomes impractical as the number of variables increases, because the number of possible interaction terms grows rapidly. For example, with p variables, the number of possible k-way interaction terms is given by the binomial coefficient:

$$\text{Number of } k\text{-way interactions} = \binom{p}{k}$$

Thus, the total number of possible terms (main effects and all possible interactions, not just two way) in a fully specified linear model is:

$$\sum_{k=1}^{p} \binom{p}{k} = 2^p - 1$$

For instance, with $p = 5$ variables: Main effects ($k = 1$): $\binom{5}{1} = 5$, Two-way interactions ($k = 2$): $\binom{5}{2} = 10$, Three-way interactions ($k = 3$): $\binom{5}{3} = 10$, Four-way interactions ($k = 4$): $\binom{5}{4} = 5$ and Five-way interaction ($k = 5$): $\binom{5}{5} = 1$. Leading to a total numbers 31 terms. As p increases, the number of terms grows combinatorially, making it infeasible to specify and interpret all possible interactions in a linear model. A more visual way of this can be seeing this phenomenon can be found in Figure 1.1 where the number of interaction terms grows as a function of the number of variables.

1.2 Machine Learning as an Alternative

To the best of the author's knowledge, there are currently no practical solutions within the traditional linear modeling framework to fully overcome the challenges of specifying and interpreting all possible non-linearities and interaction effects as the number of variables increases. This motivates the need for a different methodological approach, which is the focus of this document.

In this work, a methodology is proposed that combines three key elements: (1) the use of a flexible, non-parametric algorithm: Random Forests; (2) the application of interpretable machine learning techniques; and (3) the adaptation of these tools to account for the time-dependent structure of panel data. Random Forest is an ensemble learning method that aggregates the predictions of multiple decision trees to improve predictive accuracy and robustness. Unlike linear models, Random Forests do not require the analyst to specify the functional form of

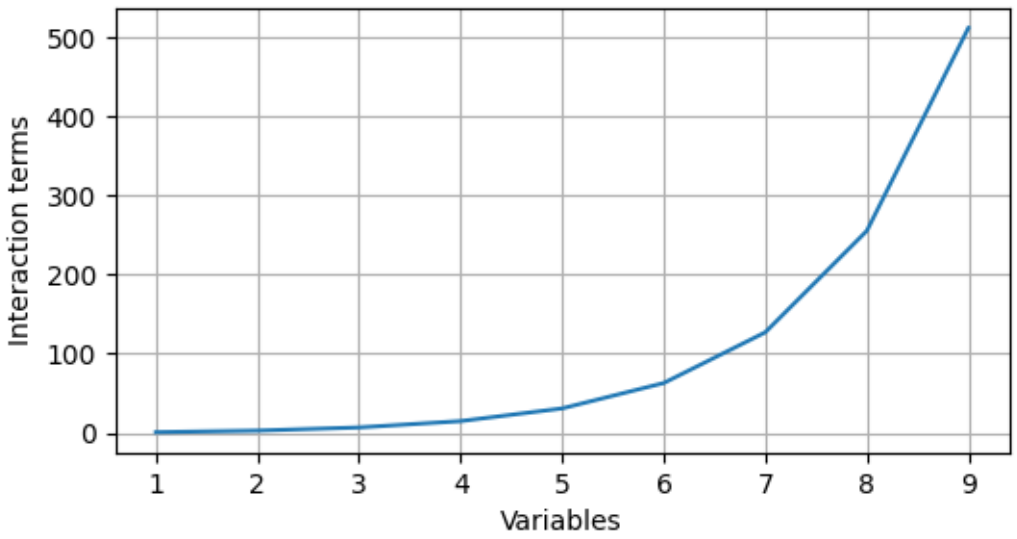

Figure 1.1: Number of interaction terms, as a function of the number of variables. It can be seen that the number of interaction terms grows exponentially with the number of variables.

relationships between variables or to manually enumerate interaction terms. The algorithm naturally captures complex, nonlinear relationships and high-order interactions among features, thereby alleviating the risk of model misspecification due to omitted nonlinearities or interactions.

However, adopting Random Forests in the context of panel data analysis introduces two main challenges. First, the standard Random Forest algorithm assumes that all observations are independent, which is often violated in panel data where repeated measurements are taken from the same individuals over time. Second, machine learning models, including Random Forests, are frequently criticized for their lack of interpretability, as their internal workings are less transparent than those of traditional statistical models. This "black-box" nature can hinder substantive understanding and limit their adoption in applied research.

The first challenge, handling time dependence, will be addressed in later sections, where we discuss modifications to the modeling approach that account for the temporal structure of the data.

To address the second challenge, interpretability, this document introduces and demonstrates three interpretable machine learning techniques: Permutation Feature Importance (PFI), Individual Conditional Expectation (ICE) plots, and Partial Dependence Plots (PDP). PFI provides a global ranking of the importance of each independent variable by measuring the increase in prediction error when the variable's values are randomly permuted. ICE plots visualize how the predicted outcome for individual observations changes as a single feature

varies, offering insight into heterogeneous effects. PDPs, on the other hand, show the average effect of one or more features on the predicted outcome, helping to reveal general patterns and marginal relationships. Together, these tools enable researchers to move beyond "black-box" predictions and gain a deeper understanding of the relationships captured by the Random Forest model, both at the global and individual level.

In summary, this document presents for a modern, flexible, and interpretable approach to panel data analysis that leverages the strengths of machine learning while addressing its traditional limitations in the context of social science research. In this context, it's important to notice that, although PDPs can have a causal interpretation as we will mention later, the goal of this document is not to replace the current methodologies but to provide an additional tool for applied social scientists.

1.3 Document Structure

The remainder of this document is organized as follows. Chapter 2 reviews the existing literature on panel data analysis and the application of machine learning methods, with a particular emphasis on Random Forests and their use in economics and the social sciences. Sections 3, 4, and 5 will be mostly a summary of the methodology and will assume some familiarity with the concepts, Section 3 introduces the structure and key characteristics of panel data, discusses traditional modeling approaches, and highlights the challenges associated with high-dimensional and time-dependent data. Section 4 examines the limitations of conventional linear models in capturing complex relationships and interactions, thereby motivating the need for more flexible methodologies. Section 5 provides an overview of the Random Forest algorithm, outlining its advantages and potential for modeling panel data, as well as discussing its strengths and weaknesses in this context. Section 6 details methodological adaptations and strategies for applying Random Forests to panel data, addressing issues such as time dependence and repeated measurements. Section 7 presents practical examples and empirical applications to illustrate the proposed methodology, including the use of interpretable machine learning techniques. Finally, Section 8 concludes with a summary of key findings, implications for applied research, and suggestions for future work.

2 Related Work

This section overviews the current work in the topics of Panel Data Analysis and Classical Econometrics Methods and Machine Learning in Economics and Panel Data. We will also include the main sources for this study and the contributions that this document offers.

2.1 Panel Data Analysis in Economics

The analysis of panel data has been a widely used in empirical economics research and there are many papers and textbooks documenting the methodology. Classical approaches, as documented by (Cameron & Trivedi, 2005) and (Wooldridge, 2010), have primarily relied on linear fixed and random effects models. These methods have proven valuable for controlling unobserved heterogeneity and conducting causal inference. However, these traditional approaches often struggle with complex, non-linear relationships and high-dimensional interactions.

2.2 Machine Learning in Economics

Comprehensive overviews of classic machine learning algorithms—including regularization techniques, tree-based ensembles, and neural networks—are provided in (Hastie, Tibshirani, & Friedman, 2009), while (James, Witten, Hastie, & Tibshirani, 2023) offers a more introductory perspective. The foundational details of Classification and Regression Trees (CART) are discussed in (Breiman, Friedman, Olshen, & Stone, 1984), and the Random Forest algorithm was specifically introduced by (Breiman, 2001a).

In recent years, (Mullainathan & Spiess, 2017) have explored the practical applications of machine learning in econometrics, particularly emphasizing prediction and cautioning against drawing causal conclusions about the effects of independent variables without careful consideration. (Varian, 2014) highlights how big data and machine learning techniques, especially tree-based models, can complement traditional econometric methods, excelling in settings with non-linearities and complex interactions. Another good review of how machine learning is being integrated into economic analysis can be found in (Desai, 2023), specifically showing how these algorithms are gaining popularity among the years.

The integration of machine learning methods into economics has accelerated, with increasing attention on adapting these tools for econometric analysis. For example, (Athey & Imbens,

2015) provide a framework for using machine learning algorithms in causal inference; (Wager & Athey, 2017) develop methods for estimating heterogeneous treatment effects using Random Forests; (Grimmer, Messing, & Westwood, 2017) extend these approaches to ensemble methods; and (Athey, Tibshirani, & Wager, 2018) introduce Generalized Random Forests for causal inference.

2.3 Tree-Based Methods for Panel Data

Several approaches have been proposed to extend random forests for longitudinal data analysis, (Hu & Szymczak, 2023) presented a comprehensive review of extensions of this algorithm to longitudinal data. However, in this document out focus will be on combining general linear models with Random Forest (or other flexible model). Hoping to capture the non-linearities with a flexible algorithm.

In particular, this document will be focused on MERF (Mixed Effects Random Forest) as refers to the approach where the mean behavior function f in a mixed effects model is estimated using a Random Forest. This method was proposed by (Hajjem, Bellavance, & Larocque, 2014) and can be seen as a generalization of their earlier method, MERT (Mixed Effects Random Trees), which used a single CART tree to estimate f (Hajjem, Bellavance, & Larocque, 2011). Both MERT and MERF employ a variant of the EM algorithm (Laird & Ware, 1982), iterating between estimating the fixed effects (via tree/forest on a modified outcome), predicting random effects, and updating variance parameters. The method assumes that errors are conditionally independent given the random effects. A further generalization of MERF, which includes a stochastic process component, is called SMERF (Capitaine, Genuer, & Thiébaut, 2021).

2.4 Interpretability in Machine Learning

Interpretability in machine learning has been discussed by many authors in recent years. One of the most influential works is (Breiman, 2001b), which highlights the differences between the objectives of machine learning and statistics. Another important philosophical article is (Shmueli, 2010), which clearly distinguishes between the goals of prediction and explanation.

There are now several techniques that aim to combine the predictive power of complex, "black box" machine learning models with methods to interpret them. A comprehensive overview of these techniques can be found in (Molnar et al., 2023). However, it is important to note that many of these methods focus on describing the behavior of the model itself, rather than uncovering the underlying data generating process (DGP) as is common in classical statistics. Despite this, recent research, such as (Molnar et al., 2023), has started to address these limitations. Further discussion of these and other challenges to adopt supervised machine learning in science can be found in (Freiesleben & Molnar, 2024).

2.5 Main Contributions

Most of the works cited above provide a theoretical background for applying machine learning to various fields. However, to the best of the author's knowledge, there is not a comprehensive guide on how to apply these models specifically to economics. Therefore, the goal of this document is to fill the gap between theory and application by providing a comprehensive guide on how to use Random Forest in economic research.

References

Athey, S., & Imbens, G. W. (2015). *Machine Learning for Estimating Heterogeneous Causal Effects* (Research Papers No. 3350). Stanford University, Graduate School of Business. Retrieved from Stanford University, Graduate School of Business website: https://ideas.repec.org/p/ecl/stabus/3350.html

Athey, S., Tibshirani, J., & Wager, S. (2018). *Generalized random forests*. Retrieved from https://arxiv.org/abs/1610.01271

Breiman, L. (2001a). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L. (2001b). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231. https://doi.org/10.1214/ss/1009213726

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Routledge. https://doi.org/10.1201/9781315139470

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and applications*. Cambridge University Press.

Capitaine, L., Genuer, R., & Thiébaut, R. (2021). Random forests for high-dimensional longitudinal data. *Statistical Methods in Medical Research*, *30*(1), 166–184. https://doi.org/10.1177/0962280220946080

Desai, A. (2023). *Machine Learning for Economics Research: When What and How?* (Papers No. 2304.00086). arXiv.org. Retrieved from arXiv.org website: https://ideas.repec.org/p/arx/papers/2304.00086.html

Freiesleben, T., & Molnar, C. (2024). *Supervised machine learning for science: How to stop worrying and love your black box*. Retrieved from https://ml-science-book.com/

Grimmer, J., Messing, S., & Westwood, S. J. (2017). Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. *Political Analysis*, *25*(4), 413–434. https://doi.org/10.1017/pan.2017.15

Hajjem, A., Bellavance, F., & Larocque, D. (2011). Mixed effects regression trees for clustered data. *Statistics & Probability Letters*, *81*(4), 451–459. https://doi.org/https://doi.org/10.1016/j.spl.2010.12.003

Hajjem, A., Bellavance, F., & Larocque, D. (2014). Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation*, *84*, 1313–1328. https://doi.org/10.1080/00949655.2012.741599

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer. https://doi.org/10.1007/978-0-387-84858-7

Hu, J., & Szymczak, S. (2023). A review on longitudinal data analysis with random forest. *Briefings in Bioinformatics*, *24*(2), bbad002. https://doi.org/10.1093/bib/bbad002

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). *An introduction to statistical learning: With applications in r*. Springer. https://doi.org/10.1007/978-3-031-38747-0

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, *38*(4), 963–974.

Molnar, C., Freiesleben, T., König, G., Herbinger, J., Reisinger, T., Casalicchio, G., … Bischl, B. (2023). Relating the partial dependence plot and permutation feature importance to the data generating process. In *Explainable artificial intelligence* (pp. 456–479). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-44064-9_24

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106. https://doi.org/10.1257/jep.31.2.87

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, *25*(3), 289–310. https://doi.org/10.1214/10-STS330

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, *28*(2), 3–28. https://doi.org/10.1257/jep.28.2.3

Wager, S., & Athey, S. (2017). *Estimation and inference of heterogeneous treatment effects using random forests*. Retrieved from https://arxiv.org/abs/1510.04342

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data* (Vol. 1). The MIT Press. Retrieved from https://ideas.repec.org/b/mtp/titles/0262232588.html