Leading Firms and the Future of Work

Max Brès Fabrizio Colella David Koll Philipp Kircher
Vinzenz Ziesemer

June 26, 2024

This is a preliminary version.
Please do not circulate without the permission of the authors

1 Introduction

The landscape of employment is undergoing rapid transformation, driven by advancements in technology and shifts in economic structures. According to the influential paper by Frey and Osborne (2017), a significant portion of the workforce faces potential displacement due to automation. This risk is particularly high in administrative, sales, and service jobs, while managerial and STEM fields face comparatively lower risks.

When technological changes disproportionately benefit workers with particular skill sets, it is referred to as *Skill-Biased Technological Change* (SBTC). SBTC has far-reaching effects on the labor market, influencing job availability, wage distribution, and often fueling wage inequality. Autor and Dorn (2013) demonstrate how technology has also favored the related phenomenon of job and wage polarization, characterized by the simultaneous growth of high-skill and low-skill jobs at the expense of middle-skill jobs and their wages. Similarly, Acemoglu and Restrepo (2020) provide evidence on the impact of robots on U.S. labor markets, suggesting that automation not only substitutes for routine tasks but also complements more complex tasks, leading to a nuanced view of technological impacts.

The rise of "superstar firms" (Autor et al., 2017) further complicates labor market dynamics. These firms, benefiting disproportionately from technological and market advantages, contribute to the de-

The authors gratefully acknowledge funding through the European Research Council Grant No. 818859. Access to confidential data, on which this work is based, has been made possible within a secure environment offered by CASD – Centre d'accès sécurisé aux données (Ref. 10.34724/CASD). The views expressed in this paper are solely the responsibility of the authors and should not be interpreted as reflecting the views of Sveriges Riksbank.
Brès: Sveriges Riksbank. max.bres@riksbank.se; Colella: USI Lugano. fabrizio.colella@usi.ch; Koll: University of Mannheim. koll@uni-mannheim.de; Kircher: Cornell University. pk532@cornell.edu; Ziesemer: Instituut voor Publieke Economie. v.ziesemer@instituut-pe.nl.

clining labor share in the economy. Caselli and Manning (2019) challenge the pessimistic view of technological change, arguing through a rigorous theoretical framework that while new technologies can disrupt labor markets, they do not uniformly harm all workers. Instead, they highlight the conditions under which technological advancements can lead to overall wage growth and improved labor market outcomes. Greenwood et al. (1997) provide a long-term perspective on technological change, emphasizing the importance of investment-specific technological advancements in shaping economic growth and labor market structures.

Further exploring the implications of technological change, Acemoglu and Autor (2012) analyze the race between education and technology, arguing that while technological advancements can drive wage inequality, investments in human capital can mitigate some adverse effects. To equip workers with the necessary skills, it is crucial to predict which jobs will be in demand in the future. Understanding future labor demand will also provide valuable insights for policymakers, enabling them to navigate the evolving employment landscape and design effective retraining programs for the unemployed. Despite the clear benefits of predicting future labor market trends, the literature has yet to establish a comprehensive and general method for forecasting labor market changes.

This paper introduces an algorithm to forecast the occupational structure of the labor market. We employ a machine learning approach to identify the characteristics of *leading firms*, which are those whose occupational structures are ahead of broader economic trends. Using this information, we estimate a simple forecast equation to predict future job trends based on the current and past occupational structures of the economy and of the leading firms. We estimate our model using comprehensive data on firms and workers in France[1] for the period 1997–2001 and test its performance using the same data for 2002–2006. We benchmark our model against a baseline forecast model that does not differentiate leading firms from others and a model that uses an alternative productivity-based methodology to identify leaders. Our comparisons, based on root mean square errors, show that the proposed procedure significantly outperforms the other two models.

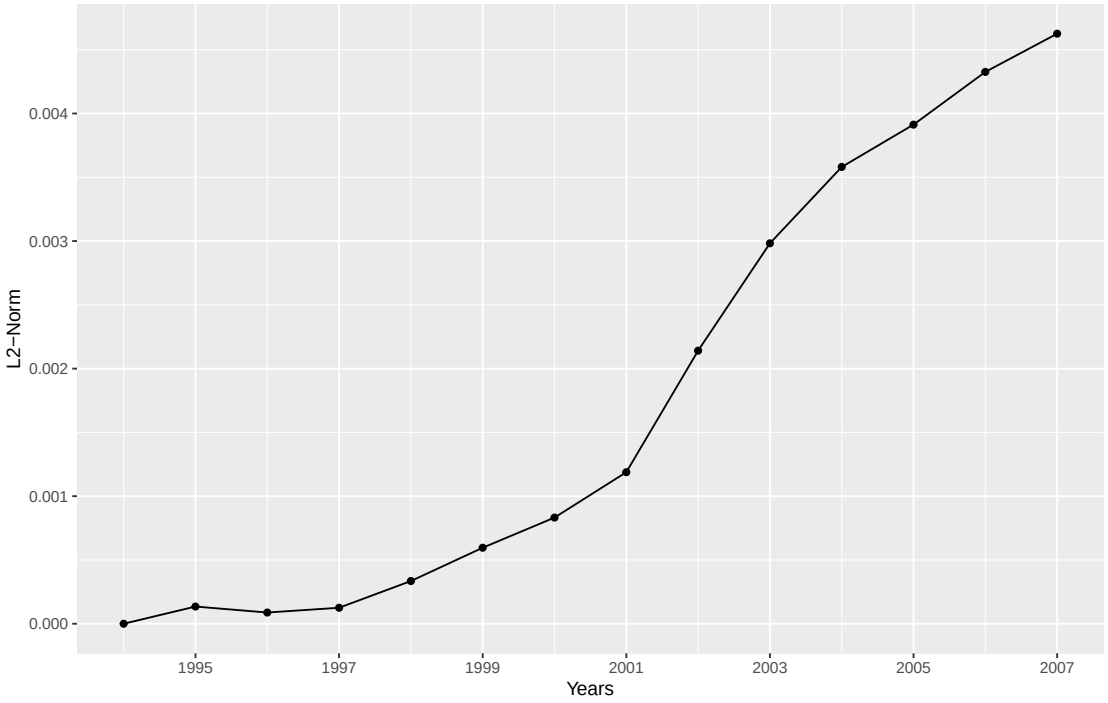
2 Measuring Occupational Differences

We begin by defining the occupational structure of an economy as the distribution of working hours across different occupations. The share of hours worked in occupation o at time t is represented as S_t^o, and the occupational structure as the vector $S_t = [S_t^1, \ldots, S_t^O]$. To measure the difference between two occupational structures at times t and s, we use the L-2 norm, which calculates the root of the sum of squared differences in these shares across all occupations: $\|S_t - S_s\|_2 = \sqrt{\sum_{o \in O}(S_t^o - S_s^o)^2}$.

[1]We use two main datasets. The first contains firm-level balance sheet data for all private sector firms in France (FARE/FICUS), provided by the *Institut national de la statistique et des études économiques* (INSEE) and the *Direction générale des Finances publiques*, which includes information about firms' balance sheets, e.g., assets, ... The second consists of job-level social declarations of all private sector employees (DADS), provided by the INSEE, which includes occupation, wage payments, hours worked, and socio-economic characteristics of workers. We combine them into a panel of firms covering 1997–2006, with plans to use data from 2007–2018 for future tests of our model.

Figure 1 shows the occupational distance between S_{1994} and S_t in France for the years 1995 to 2007, illustrating how the distribution of work hours among various occupations has evolved over time. The steadily increasing measure indicates that the occupational structure has been diverging from its 1994 baseline, with an average difference per occupation of 2.08% between 1994 and 2007.[2]

Figure 1: Occupational Distance between year t and year 1994



The figure reports the occupational distance between year t and year 1994 for the French workforce. 2-digits occupational codes are used for a total of 27 occupations. *Source:* FICUS/FARE and DADS Databases.

3 Predicting Leading Firms

We then turn to the identification of "leading firms", which in this context are defined as those whose occupational structure at time t closely mirrors the industry-wide structure at time $t+h$. Our interest lies not in the leading firms *per se*, but in the characteristics that make these firms leaders. To this extent, we employ a firm-level machine learning (ML) algorithm[3] and estimate the firm's characteristics that predict the following distance:

$$\left\| S_t^f - S_{t+h}^{(i_f,-f)} \right\|_2 = \sqrt{\sum_{o \in O} \left(S_t^{o,f} - S_{t+h}^{o,(i_f,-f)} \right)^2} \tag{1}$$

where S_t^f is the occupational structure of firm f at time t and $S_{t+h}^{(i_f,-f)}$ is the occupational structure of all other firms operating in the same industry as firm f at time $t+h$. This procedure provides a predicted

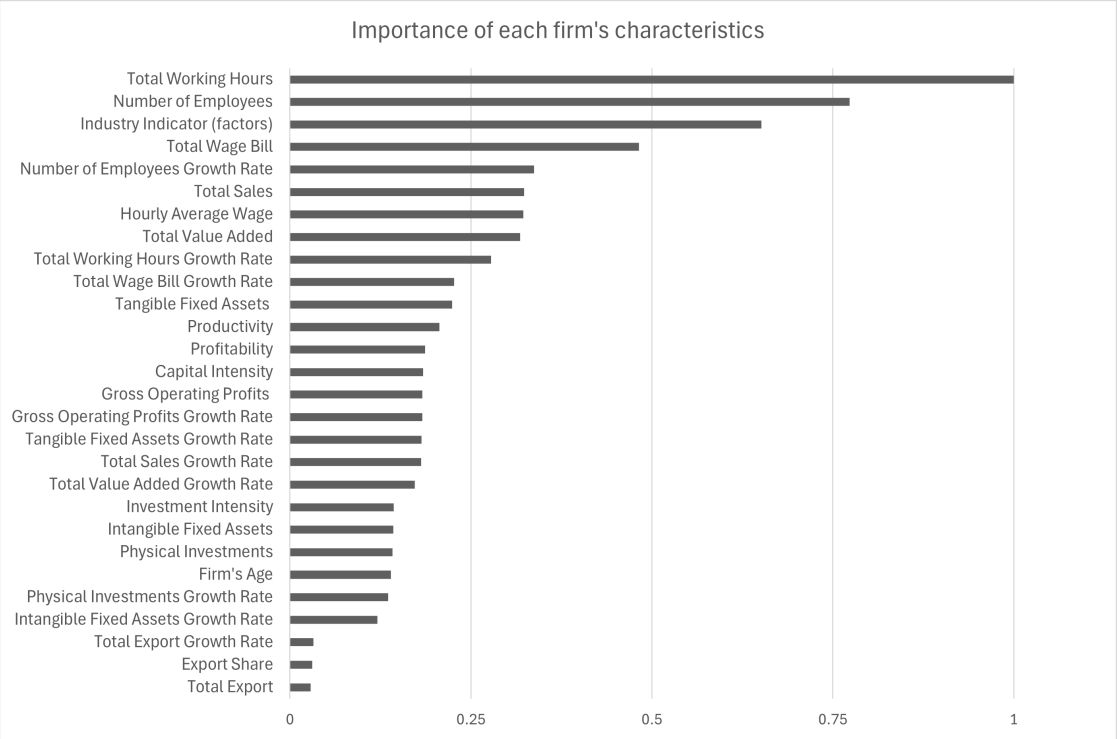
[2]The average difference per each of the 27 occupations is $\frac{0.004}{\sqrt{27}} \approx 0.00077$. Given the average occupation share is $\frac{1}{27} \approx 0.037$, the relative change per occupation is $\frac{0.00077}{0.037} \approx 0.0208$, or approximately 2.08%.

[3]We rely on a Random Forest Algorithm with 3000 trees and seven randomly chosen features for each tree.

distance between each firm's current occupational structure and the future occupational structure of the economy. We then identify as leaders the 20% of firms with the lowest predicted distance by industry, weighted by firm size.[4]

We estimate the ML model for the quasi-universe of French firms between 1999 and 2001, using as predictors the averages and growth rates of several firm characteristics between 1997 and 1999. Figure 2 shows all the variables used in the estimation, ranked by their importance (predictive power) in predicting the occupational distance in equation 1. The figure reveals two notable facts. First, firm size indicators play a crucial role in predicting occupational distances, whereas measures like productivity and profitability are less relevant. Second, stock variables are more informative than flow variables.

Figure 2: Variable Importance in the Random Forest Estimation of Leading Firms



The figure reports all the variables used in the Leading Firms Prediction Algorithm. Importance refers to the power of each characteristic to predict the occupational distance between the firm in t and the industry in $t+h$. Values are normalized such that the most important characteristic gets the value 1. Variables at t are average values between 1997 and 1999, growth rates are computed between 1997 and 1999. $t+h$ is 2001. Industry Indicators are 187 Industry Dummies at the 3 Digits level, while Firm's Age refers to the Number of Years Since Creation. Profitability = Gross Operating Profits / Total Value Added, Investment Intensity = Physical Investment / Tangible Fixed Assets, Capital Intensity = Tangible Fixed Assets / Total Sales, Export Share = Total Export / Total Sales, Productivity = Total Value Added / Total Working Hours. *Source:* FICUS/FARE and DADS Databases.

4 Forecasting

The procedure described in the previous section can be used to predict the leaders in an economy at any point in time, based on their characteristics. The final step is to use these leaders to forecast

[4]This exercise can be performed for any length of h; we do not make any specific recommendation on the time horizon at this stage.

the occupational structure of the future. This is achieved by estimating the following model at the occupation-industry level:

$$S_{t+h}^{o/i} = \alpha + \beta_{L_1} L_{t,t}^{o/i} + \beta_{L_2} L_{t,t-h}^{o/i} + \beta_{S_1} S_t^{o/i} + \beta_{S_2} S_{t-h}^{o/i} + \epsilon, \tag{2}$$

Where $S_t^{o/i}$ is a vector representing the occupational structure of industry i at time t, and $L_{t,s}^{o/i}$ is a vector representing the occupational structure at time s of the predicted leading firms at time t. This model predicts the future $(t+h)$ occupational structure of the economy using the contemporaneous (t) and lagged $(t-h)$ occupational structure of the whole economy and the current predicted leaders.

We estimate the coefficients of this model using the same sample of French firms employed for the ML algorithm. Column (3) of Table 1 reports the beta coefficients and the Mean Squared Error of this estimation. To compare our model with alternative specifications, we also estimate the same model without the lagged occupational structure of the leaders (column 2) and without including any information about the leaders (column 1) as specified in equation 3.

$$S_{t+h}^{o/i} = \tilde{\alpha} + \tilde{\beta}_{S_1} S_t^{o/i} + \tilde{\beta}_{S_2} S_{t-h}^{o/i} + \varepsilon. \tag{3}$$

The sum of the two coefficients in the first column is approximately 1. Including the leaders' information decreases this sum to 0.67 in column (2) and 0.74 in column (3), indicating that this additional information significantly contributes to the forecast. The consistent reduction in the Root Mean Squared Error (RMSE) across the three columns further confirms that incorporating firm leaders in the estimation improves the prediction of the occupational structure.

5 Assessment

The estimation results above show a strong in-sample performance of the proposed forecasting methodology. However, to properly assess our method's performance, we perform an out-of-sample prediction and forecast. First, we use the procedure from Section 3 to predict the leaders in the French economy in 2004. We then use the occupational structure of this group of leaders in 2002 ($L_{t,t-h}^{o/i}$) and 2004 ($L_{t,t}^{o/i}$), along with the β coefficients estimated in Section 4, to forecast the occupational structure of each industry in 2006 ($\widehat{S}_{t+h}^{o/i}$). This is done using the full model (equation 2), the baseline model without firm leaders (equation 3), and a partial model using only the contemporaneous occupational structure of the leaders.

Columns (1-3) of Table 2 report the results of this forecasting exercise. The RMSE is 0.0190 in the baseline model and 0.0178 and 0.0180 in the two models with the firm leaders. The corresponding Theil Ratios[5] for these two models compared to the baseline model are 0.9334 and 0.9402, respectively.

[5]The Theil Ratios are computed as the ratios between the RMSE of the models with the firm leaders and the RMSE of the baseline model.

Table 1: Forecast Regression Results

		ML Leaders	
	Baseline Model	Partial Model	Full Model
	(1)	(2)	(3)
Dependent Variable: Future Occupational Structure, Industry Level, $S^{o/i}_{t+h}$			
Regressors:			
Industry: Current Occ. Structure, $S^{o/i}_{t}$	0.7510***	0.4752***	0.3887***
	(0.0122)	(0.0125)	(0.0128)
Industry: Lagged Occ. Structure, $S^{o/i}_{t-h}$	0.2545***	0.1939***	0.3528***
	(0.0122)	(0.0107)	(0.0129)
Leaders: Current Occ. Structure, $L^{o/i}_{t,t}$		0.3153***	0.5232***
		(0.0076)	(0.0126)
Leaders: Lagged Occ. Structure, $L^{o/i}_{t,t-h}$			-0.2786***
			(0.0138)
Intercept	-0.0002	0.0006***	0.0005***
	(0.0002)	(0.0001)	(0.0001)
RMSE	0.0103	0.0089	0.0085
Adj. R^2	0.9854	0.9891	0.9899
Num. obs.	5049	5022	5022

Notes: Regression Results of three forecasting models for the occupational structure. The model in column (1) uses only the current and lagged occupational structure of the industry. The model in column (2) adds the current occupational structure of the leaders, while the model in column (3) adds both the current and the lagged occupational structure of the leaders. *** p<0.01, ** p<0.05, * p<0.1. Standard errors are in parentheses. The number of observations is the product between the 27 occupations and the 187 industry categories in the sample. *Source:* FICUS/FARE and DADS Databases.

Including the firm leaders in the forecasting reduces the RMSE by 6-7% compared to the baseline model, indicating improved predictive accuracy. Notably, the full model does not outperform the partial one, suggesting that the improvement in prediction stems from the contemporaneous occupational structure of the leaders rather than their past structures.

Table 2: Forecast, Out of Sample Assesment

		ML Leaders		Productivity L.	
	Baseline Model	Partial Model	Full Model	Partial Model	Full Model
	(1)	(2)	(3)	(4)	(5)
Predicted Variable: Future Occupational Structure, Industry Level, $S^{o/i}_{t+h}$					
Predictors:					
Industry, Current Occ. Structure, $S^{o/i}_{t}$	✓	✓	✓	✓	✓
Industry, Lagged Occ. Structure, $S^{o/i}_{t-h}$	✓	✓	✓	✓	✓
Leaders, Current Occ. Structure, $L^{o/i}_{t,t}$		✓	✓	✓	✓
Leaders, Lagged Occ. Structure, $L^{o/i}_{t,t-h}$			✓		✓
RMSE	0.0191	0.0178	0.0180	0.0191	0.0189
Theil Ratio (vs the Baseline Model)	—	0.9334	0.9402	1.0000	0.9900
Num. obs.	5049	5049	5049	5049	5049

Notes: Out-of-sample forecast assessment results of five forecasting models for the occupational structure of the future. The model in column (1) uses only the current and lagged occupational structure of the industry. The models in columns (2) and (4) add the current occupational structure of the leaders, while the models in columns (3) and (5) add both the current and the lagged occupational structure of the leaders. Leaders in columns (2-3) are predicted minimizing the using the proposed ML algorithm; Leaders in columns (4-5) are predicted using the most productive firms. *** p<0.01, ** p<0.05, * p<0.1. Standard errors are in parentheses. The number of observations is the product between the 27 occupations and the 187 industry categories in the sample. *Source:* FICUS/FARE and DADS Databases.

To gain additional insights into the source of the forecast improvement, we perform another forecast of the occupational structure using both the partial and full models, but with an alternative definition of leaders: we identify the 20% most productive firms by industry in 2004 as leaders. The results of this alternative forecast exercise are reported in Columns (4-5) of Table 2. The Theil Ratios for the productivity-based leaders models are 1 and 0.99, indicating that this alternative model does not improve the forecast compared with the baseline model. From this result, we conclude that the majority of the improvement in the forecast is attributable to the accurate prediction of leaders using our original method based on the ML algorithm.

Conclusion

This paper presents an innovative machine learning approach to forecast the future occupational structure by identifying leading firms whose current occupational practices predict broader industry trends. Our findings demonstrate that incorporating the occupational structure of these leaders significantly improves forecast accuracy compared to traditional models. This approach offers valuable insights for creating effective training and retraining programs, helping to prepare workers for future job demands. We plan to further refine the methodology and apply it to longer time horizons.

References

Acemoglu, D. and D. Autor (2012): "Skills, tasks and technologies: Implications for employment and earnings," *Handbook of Labor Economics*, 4, 1043–1171.

Acemoglu, D. and P. Restrepo (2020): "Robots and jobs: Evidence from US labor markets," *Journal of Political Economy*, 128, 2188–2244.

Autor, D., D. Dorn, L. F. Katz, C. Patterson, and J. Van Reenen (2017): "Concentrating on the Fall of the Labor Share," *American Economic Review*, 107, 180–185.

Autor, D. H. and D. Dorn (2013): "The growth of low-skill service jobs and the polarization of the US labor market," *American Economic Review*, 103, 1553–1597.

Caselli, F. and A. Manning (2019): "Robot Arithmetic: New Technology and Wages," *American Economic Review: Insights*, 1, 1–12.

Frey, C. B. and M. A. Osborne (2017): "The future of employment: How susceptible are jobs to computerization?" *Technological Forecasting and Social Change*, 114, 254–280.

Greenwood, J., Z. Hercowitz, and P. Krusell (1997): "Long-run implications of investment-specific technological change," *American Economic Review*, 87, 342–362.