Robotics Exposure in European Labour Markets

Bintang Satyahutama , Joohyeon Lee, Leonie Gray, Siyao Zhang

Université Paris 1 Panthéon Sorbonne

UNIVERSITÉ PARIS 1 PANTHÉON SORBONNE

Introduction & Motivation

The rapid advances of industrial robotics present a challenge to labour markets, because they directly threaten the "routine-manual" jobs that sustain the middle and lower tiers of the wage distribution. For measuring which occupations are exposed to risk the most, it is crucial to have reliable exposure indices. In our analysis, we build on Webb (2019) and use Latent Semantic Analysis (LSA), that capture semantic similarity beyond exact word matches. More specifically, we ask the following research question: To what extent can LSA improve the identification of robotics-prone tasks in comparison to simple keyword matching, and how are European wages reshaped by robotics exposure?

Literature Review

We are building on two strands of literature in our analysis. First, the task-based framework of Acemoglu and Autor (2011) positing, that technological change affects labour markets by the automation of specific tasks rather than entire occupations. This in turn, drives wage polarisation, because middle-skill routine jobs are displaced. Second, Webb (2019) builds on that by constructing occupational exposure scores from the textual overlap between job task descriptions (O*NET) and patent titles. He finds that robotics exposure predicts wage and employment declines in the United States.

We contribute to this literature by extending the Webb (2019) method to the European context. Besides that, we replace exact keyword matching with LSA. Therefore, we can identify occupations tied to "precision manual labour" even when they do not share lexical matches with the patent vocabulary.

Data

We use three data sources. Occupational text comes from ESCO v1.2.1, provided by the European Commission and covering 3,039 occupations and 13,939 skills. For comparison purposes, we use data from O*NET Task Statements (US Department of Labor), covering 923 occupations with 18,796 core task descriptions. The robotics keyword vocabulary is derived from Webb (2019) Table 2, and consists of verb-noun pairs, which were extracted from USPTO robotics patent titles. Wage data comes from the Eurostat Structure of Earnings Survey 2022, giving information about the mean hourly earnings in EUR by ISCO 1-digit occupation group across 34 European countries.

Methodology

Our methodology proceeds in four stages. All occupational texts are first preprocessed through lowercasing, removal of punctuation and digits, lemmatization, and domain-specific stopword filtering.

Stage 1 Keyword baseline: Following Webb (2019), we compute a baseline robotics exposure score as the fraction of tokens in each occupation's text that match a robotics patent keyword vocabulary derived from USPTO patent titles. This serves as our benchmark.

Stage 2 LSA exposure: We construct a joint TF-IDF matrix from O*NET and ESCO texts (20,000 features, unigrams and bigrams) and apply Truncated SVD to obtain 100-dimensional LSA vectors. A robotics centroid is defined as the mean LSA vector of the top-30 keyword-scoring O*NET occupations; the LSA exposure score for each ESCO occupation is then its cosine similarity to this centroid.

Stage 3 Prediction Model: Random Forest and LightGBM regressors are trained on the 100 LSA dimensions to predict the keyword baseline score, evaluated via 5-fold cross-validation. SHAP values decompose which latent dimensions drive exposure predictions.

Stage 4 Wage regression: We estimate wage effects using OLS with country fixed effects across 34 European countries and 9 ISCO 1-digit groups, with standard errors clustered by country.

Results

The results from the baseline analysis suggest that the highest-exposed ISCO groups are machine operators and assemblers (OC8, LSA score = +1.32), as well as craft and related trades workers (OC7, LSA score = +1.30). In contrast, managers (OC1) and service workers (OC5) are the least exposed. The emerging pattern is also consistent with Webb (2019)'s finding that physical, repetitive, and precision manual tasks are mostly targeted by robotics. These findings are presented in the right graph of Figure 1.

Figure 1. Exposure Overview
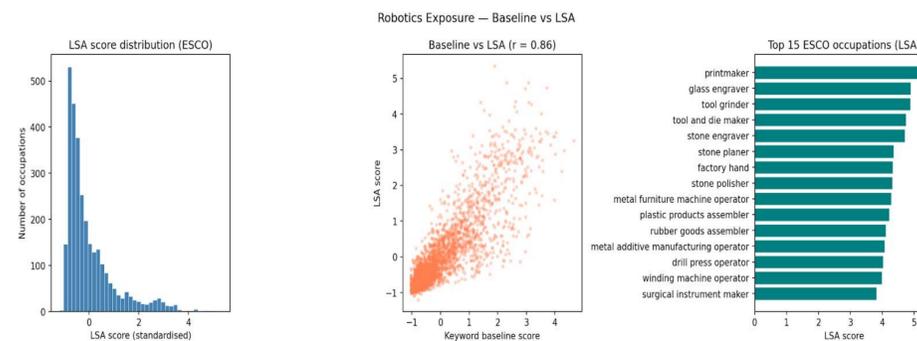
The LSA scores correlate strongly with the keyword baseline (Pearson r = 0.86), which confirms the success of LSA replicating Webb's ranking (Figure 1). In subsequent analysis, LightGBM is trained on 100 LSA dimensions, which it was not trained on before, and achieves a cross-validated R^2 of 0.915. This result demonstrate that the semantic structure captured by LSA is predicting the keyword-based measure with high precision. To better understand the drivers of prediction, we apply SHAP analysis, which identifies the most important LSA dimensions that contribute to the model's output. Here, lsa_001 is identified as the dominant dimension in predicting robotics exposure (Figure 2). Importantly, higher values of this dimension are correlated with lower predicted robotics exposure. This pattern suggests that the dimension lsa_001 represents a latent semantic axis related to manual versus cognitive tasks. Occupations that are more associated with manual and physical activities tend to load more on this dimension. They are more linked to lower robotics exposure in our model.
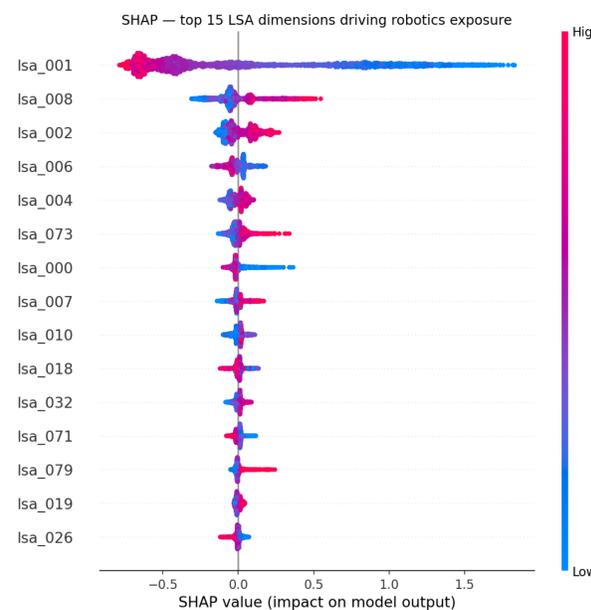

Figure 2. SHAP Ranking
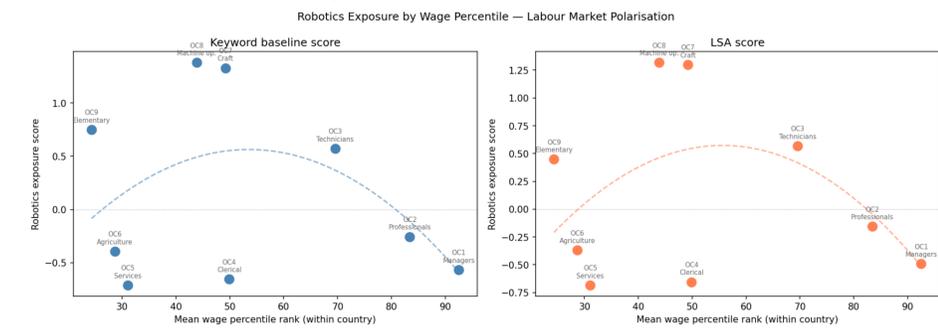

Figure 3. Labour Market Polarisation.

Figure 3 plots mean robotics exposure against mean wage percentile rank for each of the 9 ISCO 1-digit groups. Both the keyword baseline and the LSA score suggest an inverted-U pattern. Exposure peaks in middle-wage occupations, particularly OC8 (Machine operators) and OC7 (Craft workers), and declines at both the lower and upper ends of the wage distribution. This pattern closely matches Acemoglu and Autor's (2011) polarisation hypothesis and supports the view that routine, middle-skill occupations are the most exposed to automation risk.

Table 1. Wage regression

Variable	Coef.
LSA exposure score	−0.072***
R^2	0.750
N	601
Country fixed effects	Yes
Clustered SE	
Countries	34
ISCO groups	9

Notes: Dependent variable is log mean hourly wage (EUR). 95% CI [−0.086, −0.059]. *** p < 0.001. A one standard deviation increase in LSA robotics exposure is associated with a 7.2% decrease in mean hourly wages.

Table 1 reports the OLS coefficient of LSA exposure on mean hourly wages, estimated with country fixed effects across 34 countries and 9 ISCO groups. A one standard deviation increase in robotics exposure is associated with a 7.2% decrease in mean hourly wages (β = −0.072, p < 0.001). This result holds after absorbing country-level wage differences via fixed effects, suggesting that the wage penalty associated with robotics exposure reflects occupational task content rather than cross-country variation in wage levels.

Discussion & Conclusion

This analysis show that LSA can extend Webb (2019)'s patent-task overlap method to European occupations without a manual crosswalk between classification systems. The polarisation pattern we document is consistent with the task-based theory of technological change. It is middle-wage, routine-task occupations that bear the greatest automation risk, not the bottom of the wage distribution.

A key limitaion is that Eurostat publishes earnings data at the ISCO 1-digit level only, requiring us to aggregate occupation-level LSA scores across broad groups. This coarse aggregation masks within-group heterogeneity and may attenuate the estimated wage effects. Furthermore, our LSA model operates on English text only, leaving non-English ESCO descriptions unexamined. Future work could apply the LSA exposure index to individual-level EU-LFS microdata, enabling a more precise test of labour market polarisation, and extend the semantic model to multilingual ESCO descriptions.

References

Acemoglu, D. & Autor, D. (2011). Skills, tasks and technologies: Implications for employment and earnings. Handbook of Labor Economics, 4, 1043–1171.

Webb, M. (2019). The impact of artificial intelligence on the labor market. SSRN Working Paper 3482150.