Campos-Gonzalez, Jorge

Article

The impact on the skill premium of the task content of jobs: Evidence from online job ads for Chile 2009 - 2018

Estudios de Economía

Provided in Cooperation with:
Department of Economics, University of Chile

Suggested Citation: Campos-Gonzalez, Jorge (2025) : The impact on the skill premium of the task content of jobs: Evidence from online job ads for Chile 2009 - 2018, Estudios de Economía, ISSN 0718-5286, Universidad de Chile, Departamento de Economía, Santiago de Chile, Vol. 52, Iss. 1, pp. 133-189

This Version is available at:
https://hdl.handle.net/10419/315395

The impact on the skill premium of the task content of jobs: Evidence from online job ads for Chile 2009 – 2018*

El impacto del contenido de tareas de los empleos en la prima salarial de trabajadores calificados: Evidencia usando avisos de empleo online en Chile 2009 – 2018

JORGE CAMPOS-GONZÁLEZ**

Abstract

We evaluate the influence on the skill premium of the task content of jobs by exploiting the text data from online job ads covering 2009-2018 (over 189,000 ads) published by one of the leading Chilean online job portals (www.trabajando.com). Our analysis tests the expected complementarity between skilled labour, non-routine cognitive (analytical and interactive), and routine cognitive tasks. Our results show weak evidence of the influence on the skill premium of our task-related measures. Nevertheless, some implications arise from this apparent decrease in the importance of the tasks skilled workers typically perform, such as inefficient educational investment or unwanted changes in the occupational ladder.

Key words: *Task content, Skill premium, Cognitive tasks, Routinisation, Technological change, Skilled labour.*

JEL Classification: *I26, J23, J24, J31, O15, O33.*

* The author acknowledges the financial support from the National Agency for Research and Development (ANID) / Scholarship Program / DOCTORADO BECAS CHILE/2017 – 72180253. Also, the author expresses his gratitude to Trabajando.com for granting access to online job posting databases. This paper has greatly benefited from comments by the Editor of *Estudios de Economía* and two anonymous reviewers to whom the author is very grateful. The usual disclaimer applies.

** School of Agriculture, Policy and Development, University of Reading, Earley Gate, Whiteknights Road, Reading, RG6 6EU, United Kingdom. Millenium Nucleus for the Integrated Development of Territories, CEDIT, Santiago, Chile. Email: jorge.camposgonzalez@reading.ac.uk

Received: August, 2023 Accepted: November, 2024

Resumen

Este estudio evalúa la influencia del contenido de las tareas de las ocupaciones laborales en la prima salarial de trabajadores calificados mediante la explotación del texto de avisos de empleo online (más de 189.000 anuncios) publicados entre 2009-2018 por uno de los principales portales de empleo en línea de Chile (www.trabajando.com). Este análisis examina la sugerida complementariedad entre la mano de obra calificada y mediciones representando las tareas cognitivas (rutinarias y non rutinarias) de los empleos. Los resultados muestran una débil evidencia sobre la esperada influencia de nuestras medidas relacionadas con el contenido de tareas sobre la prima salarial de los trabajadores calificados. Sin embargo, a partir de esta aparente disminución de la importancia de las tareas que suelen realizar los trabajadores cualificados se derivan algunas implicancias tales como la ineficiencia de la inversión educativa o los cambios no deseados en la escala ocupacional.

Palabras clave: *Contenido de tareas, Prima salarial, Tareas cognitivas, Rutinización, Cambio tecnológico, Mano de obra calificada.*

Clasificación JEL: *I26, J23, J24, J31, O15, O33.*

1. Introducción

The impact of technological change on labour markets has been widely debated, focusing on how technology influences employment patterns and wage inequality (Autor, 2015; Autor et al., 1998). A fundamental framework used to analyse these dynamics is the Autor, Levy, and Murnane (2003) model, also known as the ALM model and task-based approach, which posits that technology complements non-routine cognitive tasks but substitutes for routine tasks, thereby reshaping the demand for different skill levels (Autor et al., 2003). While numerous studies have explored these dynamics in developed countries (Goos et al., 2014; Goos & Manning, 2007; Sebastian, 2018; Spitz-Oener, 2006), evidence from economies recently classified in this status[1], such as Chile, remains limited, particularly in the context of rapidly evolving digital technologies. Thus, computer-based technologies such as Information and Communications Technologies (ICT) and automation, among others, have fur-

[1] The World Bank classifies countries into four income groups—low, lower-middle, upp per-middle, and high-income countries— using thresholds based on Gross National Income (GNI) per capita in current USD Income. In 2012 Chile was assigned to the high-income category since its GNI per capita was higher than USD$12,615 (World Bank, 2020).

ther fuelled this debate since it has been suggested that much of the technological change in production is driven by these advancements (Acemoglu & Autor, 2011; Almeida et al., 2020).

This study seeks to address this gap by examining how changes in the task content of jobs influence the skill premium—the wage gap between skilled and unskilled workers[2] —in Chile from 2009 to 2018. Our analysis tests the ALM framework in the context of the Chilean labour market, using changes in task content as proxies for shifts in technology's impact. While we do not directly identify technology-specific shocks, the observed co-movements between task measures and the skill premium suggest that the patterns align with the expected outcomes of skill-biased technological change.

The study leverages data from online job advertisements, providing a unique opportunity to capture detailed information on job requirements. We chiefly analyse the open text data (the job title, job description, and job requirements), wages, and educational data by applying analytical processing and classification techniques. Thus, we build a monthly time series for the skill premium and task-related metrics. We use a Vector Autoregressive (VAR) framework to assess the relationship between the skill premium and the task content of jobs derived from the ALM model, explicitly focusing on non-routine analytical, non-routine interactive, and routine cognitive tasks. Our focus relies on the increase in demand for skilled labour due to computer-based and related advancements since it has been suggested that this complementarity between highly qualified workers and technology is one of the main determinants that exacerbate the skill premium (Acemoglu & Autor, 2011; Goldin & Katz, 2008).

Our main contribution is twofold. First, we comprehensively analyse how task-based shifts influence the skill premium in an economy transitioning from a developing to a developed country using high-frequency job ads data. Second, we challenge the conventional ALM model by showing that the expected complementarity between non-routine cognitive tasks and skilled labour may not hold in the Chilean context. Our findings suggest that contrary to predictions, routine cognitive tasks might play a more significant role than previously thought, potentially indicating a de-skilling trend in the Chilean labour market, as suggested by researchers (Almeida et al., 2020; Zapata-Román, 2021). This de-skilling process also aligns with the skill premium decline in recent decades (Campos-González & Balcombe, 2024; Parro & Reyes, 2017).

The results have policy implications, highlighting the need to reassess the role of education and skill development in responding to technological change.

[2] According to the conventional distinction based on workers' educational attainment, skilled and unskilled individuals are those workers with tertiary education and secondary or less schooling, respectively. An additional distinction refers to middle-skilled workers as those with only secondary schooling.

Given the ongoing expansion of tertiary education in Chile as a response to the demand for skilled labour fuelled by technological change, trade and other factors (Gallego, 2012; Murakami, 2014), understanding these dynamics is crucial for aligning educational outcomes with evolving labour market demands.

The paper will proceed by first outlining the ALM model and past evidence in Section 2. Section 3 will discuss the data before outlining our methodologies, i.e., the construction of variables and empirical strategies that we employ, in Section 4. Section 5 gives and discusses the results, and Section 6 concludes.

2. Conceptual Framework & Past Evidence

The ALM model enables us to evaluate the differentiated impact of technological advancements on different kinds of labour by positing that technology is biased towards routine or codifiable work tasks. These interactions are examined following a two-fold classification (routine/non-routine and cognitive/manual) which in turn generates five categories: 1) routine cognitive, which involves activities regarding the processing of information defined by explicit rules which can easily be programmable; 2) non-routine analytic and 3) non-routine interactive, both under the cognitive category, capture labour tasks involving reasoning skills and interactive abilities (e.g., communication and managerial skills), respectively; 4) routine manual and 5) non-routine manual refers to repetitive and non-repetitive physical work activities. This distinction tries to separate tasks that can be potentially programmable based on their degree of routineness. Thus, the ALM model assumes that routine tasks can be expressed as programmable rules or as codifiable; in that case, they could be executed by computer-related technologies. For the sake of clarity, Table 1 shows how we can allocate work activities to the five task categories discussed above following past studies (Atalay et al., 2018; Autor et al., 2003; Dengler et al., 2014; Mihaylov & Tijdens, 2019; Spitz-Oener, 2006).

TABLE 1

EXAMPLES OF THE ASSIGNMENT OF WORK ACTIVITIES TO TASK CATEGORIES

Classification and definition	**Job Tasks examples**
Non-routine cognitive analytic: non-repetitive work activities involving reasoning, critical thinking and problem-solving	Researching, analysing, evaluating and planning, making plans/constructions, designing, sketching, working out rules/prescriptions, using and interpreting rules, examining patients, and using advanced software, among others
Non-routine cognitive interactive: non-repetitive work activities involving creativity and complex communication	Negotiating, lobbying, coordinating, organising, teaching or training, selling, buying, entertaining or presenting, employing or managing personnel, pleading in courts of law, and interviewing, among others
Routine cognitive: repetitive work activities regarding the processing of information	Calculating, book-keeping, correcting texts/data, and measuring length/weight/temperature, operating systems and networks, operating laboratory and office computer equipment, inspection and quality control, among others
Routine manual: repetitive and physical work activities. (ICT and machines can automate them)	Operating, controlling or monitoring stationary machines and equipping machines; making standardised products; assembling prefabricated parts or components; sorting and storing produce, among others
Non-routine manual	Repairing or renovating houses, apartments, machines, vehicles, restoring art and monuments, serving, or accommodating, operating non-stationary and mobile equipment, driving and, guarding and protecting, among others

Source: Adapted from past studies (Atalay et al., 2018; Autor et al., 2003; Dengler et al., 2014; Mihaylov & Tijdens, 2019; Spitz-Oener, 2006).

Table 1's first column presents the task category and a general definition. In the second column, we see the most representative work activities featuring each kind of task. This allocation allows us to classify tasks according to the ALM model categories. Under the assumptions discussed above, the ALM model predicts the differentiated impact of computer-based technologies on different kinds of labour. We can describe three main potential interactions of this impact. First, technologies can become substitutes for workers doing jobs with an intensive demand for routine tasks, both cognitive and manual, such as middle-skilled labour (see footnote 2), who usually perform jobs rich in these tasks (e.g., clerical workers, assemblers). Secondly, technologies can complement non-routine cognitive activities, both analytical and interactive. These cognitive tasks mainly feature in skilled occupations (e.g., professionals, managers, associate professionals or technicians) requiring workers with specific knowledge or abilities provided by tertiary education (Autor, 2015). Thirdly, in the case of non-routine manual tasks, technologies might have a limited role as substitutes or complements for labour. We typically observe less skilled or unskilled occupations involving non-routine manual tasks (e.g., food preparation and serving, cleaning, and security services). Thus, the impact of technology will depend on the composition of the tasks of labour occupations. Implications include the possibility that computer-based technologies might harm labour outputs (e.g., demand and wages) of jobs with an intensive demand for routine tasks. On the other hand, skilled labour jobs abundant in non-routine analytical and interactive tasks might benefit from higher demand and productivity, resulting in skill premium improvements.

Some have empirically tested the differentiated technological impact predicted by the ALM model. There is evidence of the complementarity between skilled workers and jobs with an intensive demand for non-routine cognitive tasks, both analytical and interactive (Autor et al., 2003; Goos et al., 2014; Goos & Manning, 2007; Sebastian, 2018; Spitz-Oener, 2006). However, studies employing the ALM model in Chile are recent, and the evidence contradicts its main predictions. For example, Almeida et al. (2020) analysed the impact of complex software as proxies for computer-based technologies in Chilean firms between 2007 and 2013, finding that these technologies encouraged routine and manual tasks. Simultaneously, the software replaced analytical tasks, displacing skilled labour to less-skilled positions. The results of Almeida et al. (2020) are in line with recent studies showing a broader class of jobs at risk due to the potential displacement role of frontier technologies (Arntz et al., 2016; Frey & Osborne, 2017), such as robotics and artificial intelligence which can automate non-routine analytical or interactive tasks (Autor, 2015). Another study based in Chile suggests that the routine content of jobs plays an important role in earnings (Zapata-Román, 2021). Based on four waves of Chilean

household data from 1992 to 2017, Zapata-Román (2021) shows how the technological change in production, which is assumed abundant in computer-based technologies, would encourage jobs abundant in routine tasks: this contradicts ALM model predictions. Given these findings for Chile, an open question is whether the expected complementarity between skilled workers and jobs with an intensive demand for non-routine cognitive tasks can also be extended to routine cognitive tasks.

3. Online Job Ads Data

Our data cover all full-time job ads posted on the Chilean online job portal www.trabajando.com between January 1, 2009, and December 31, 2018. After some cleaning[3], our sample consists of 189,986 unique job ads. In this study, we analyse the job ads grouped by month, i.e., 120 data points. The average of job ads by month is 1,583 (standard deviation of 433), and the minimum and maximum frequencies are 657 and 2,670, respectively. For our purposes, the key advantages of this data are the detailed requirements stipulated by firms. For instance, formal qualifications (to identify job postings demanding skilled workers), offered wages (to build the skill premium), and work activities to be performed (to build our task content measures), among others.

Considered the principal internet labour market intermediary in Chile over the 2000s (Ramos et al., 2013), researchers have used data from www.trabajando.com to examine the impact on wages of job skills, job search behaviour, among other aspects of labour markets (Banfi et al., 2019, 2022; Banfi & Villena-Roldán, 2019; Campos-González, 2024; Ramos et al., 2013). Regarding the offered wages, we note that this information is required for all firms posting a job ad, although they can choose whether it is published in the job ad. Despite this feature, Banfi & Villena-Roldán (2019) show that these hidden wages are reliable measures of salaries that firms expect to pay. Table 2 shows some descriptive statistics and features of our job ads sample. These statistics are similar to Banfi & Villena-Roldán (2019), who used the same data source but in different periods (January 2008- June 2014).

[3] We exclude job ads using some criteria as past studies (Banfi & Villena-Roldán, 2019): (i) monthly wages below CLP (Chilean Pesos) 150,000 (minimum wage at the start of the period) or above CLP 5,000,000 (unfeasible), (ii) work experience above 30 years (less probable).

TABLE 2

SUMMARY OF FEATURES FOR TRABAJANDO.COM 2009-2018 JOB ADS

Required years of experience (%)	
0	16%
1	30%
2 to 3	39%
4 to 20	15%
Average years of experience (SD)	2.09 (1.89)
Required education level (%)	
Primary/secondary/technical secondary	36%
Technical tertiary	29%
College (tertiary)/graduate	34%
Other	1%
Sectors (%)	
Manufacturing	17%
Electricity/gas/water	2%
Commerce	19%
Transportation	5%
Communication	9%
Financial/business/personal service	27%
Other	21%
Offered wage (%)	
CLP <= 300,000	22%
CLP 300,001-600,000	39%
CLP 600,001-1,000,000	23%
CLP>1,000,000	16%
Average CLP offered wage (SD)	690,839 (542,024)
Observations	**189,986**

Note: CLP stand for Chilean Pesos

4. Methodologies

4.1. Construction of Variables

4.1.1. The Skill Premium

Our definition of the skill premium is based on the educational requirements specified in job advertisements and follows standard strategies used in this literature (Autor et al., 2008; Card & Lemieux, 2001; Ciccone & Peri, 2005). For Chile, see e.g., Gallego (2012), Murakami (2014) and Campos-González & Balcombe (2024), among others. Specifically, we define skilled labour as job ads requiring college or tertiary graduates and unskilled labour as job ads specifying high school education or less. Thus, our skill premium captures the difference between the offered wages for jobs categorised as skilled versus unskilled. This distinction is crucial, as our approach reflects the aggregate wage differential between groups of job postings, representing broader labour market trends rather than individual-level wage returns. To estimate the skill premium, we regress the monthly offered wage on typical wage determinants available in the data following a Mincer regression strategy using all the job ads for a given month. Then, we estimate the skill premium month by month using the differences in predicted wages between skilled and unskilled workers. We focus on job ads offering full-time positions, and we use weighted averages from education construction by experience subgroups to adjust for compositional changes. The skill premium estimation consists of the following three steps:

Step 1. Construction of education by experience sub-groups to adjust for compositional labour changes (e.g., different skill levels) within each subgroup using the educational level and experience specified in the job ads. We define four educational categories as our measure of schooling for different workers' school attainments: college graduates, some college and high school graduates and less educated (primary and high school dropouts). There are three experience subgroups: 0-2, 3-5 and 6-30 years. Combining the education and experience categories, we construct 12 education-by-experience subgroups. We use the total hours worked monthly for each sub-group as weights, assuming that full-time positions correspond to 193.5 working hours per month (45 hours per week * 4.3 weeks per month).

Step 2. Estimating the predicted wages for skilled and unskilled workers regressing a Mincer-type equation. We regress the wages for each monthly sample of job ads estimating the following standard wage equation[4]:

[4] This methodology allows controlling of the labour supply by other demographic characteristics which are not related to the education premium.

$$\log\left(W_{i,t}\right) = cons + educ\_cat'_{i,t}\,\alpha_j + \beta_1 \exp_{i,t} + \beta_2 \exp^2_{i,t} + X'_{i,t}\delta \quad (1)$$

where $W_{i,t}$ is the monthly offered (log) wage for a job ad i in month t, expressed in December 2018 Chilean pesos (CLP) using the Unidad de Fomento as a deflator[5]. $educ\_cat$ are j educational categories defined in Step 1 with "less educated" as the base category. exp is the required work experience. X is a vector containing additional determinants, such as the economic sector of the firm posting the job ad (eight industries such as agriculture, mining, and construction, among others, with manufacturing as the base category) and the firm size (big, medium, small, with micro as the base category). We use these regression results to compute the predicted wages for skilled and unskilled workers, as detailed in Step 3.

Step 3. Estimating the predicted average wage for skilled and unskilled groups and computation of the skill premium. We estimate the predicted log wages using regression results from Step 2 evaluated at the correspondent experience level (1, 4, or 10 years based on experience categories) and at base categories included in the vector X. We compute the predicted log wage difference between college and high-school graduates as our proxy for the skill premium. We use the sum of monthly hours worked for each of the education x experience sub-groups built-in Step 1 as weights. The wage differential is our monthly measure of the skill premium, which captures how firms value skilled versus unskilled labour at an aggregate level in the Chilean labour market.

4.1.2. Estimation Of Task Content Measures

Our strategy of building task content indicators from job ad data relies on the quantification and classification of tasks proposed by the ALM model. Our measures show the prevalence of each category of tasks across job postings demanding skilled labour by allocating job postings to standard occupations. These standard occupations give detailed work activity descriptions, which we can classify according to the ALM model's task categories. However, job ads do not follow standard national or international labour classifications. Besides, the task descriptions are specific to the offered jobs, resulting in a lack of information about additional general tasks. To tackle this difficulty, we developed a strategy consisting of three steps. It starts with the manual classification of work activities that feature each occupational group into the categories proposed by the ALM model, using national and international classifications of occupations as statistical tools and dictionaries from the literature. The second step categorises each job ad according to its standard occupational groups.

[5] The Unidad de Fomento (UF) is a Chilean unit of account. The exchange rate between the UF and the Chilean peso is constantly adjusted for inflation.

However, the data does not contain references to standardised classes of occupations, therefore we infer that information using the text data from the job ads by applying a classifier algorithm. The third step corresponds to constructing measures representing the task content of occupations by mapping the task analysis of occupations in step one and the classification of job ads from step two. We detail these steps as follows.

4.1.2.1 Step One: Examining The Task Content Of Standard Occupational Groups.

To evaluate the task content of standard occupational groups, we rely on the task descriptions for occupations documented in the Chilean Classification of Occupations, CIUO08-CL (INE, 2018). In turn, CIUO08-CL relies on the current International Standard Classification of Occupations, ISCO-08 (ILO, 2012). To ensure reliability, CIUO08-CL is prepared and published by the government agency in charge of national statistics for the labour sector (in Spanish, Instituto Nacional de Estadísticas, INE) (INE, 2018). Like ISCO-08, the CIUO08-CL structure is hierarchical. From the top down, ten major groups are composed of 44 sub-major groups, containing 129 minor groups. The 129 minor groups contain 444 unit groups, which the most exhaustive level of the classification. Regarding coding, 1-digit, 2-digit, 3-digit, and 4-digit codes represent the major, sub-major, minor and unit groups, respectively. To illustrate the CIUO08-CL structure, Table 3 presents an example of the hierarchy and task descriptions.

TABLE 3
EXAMPLE OF THE CHILEAN CLASSIFICATION OF OCCUPATIONS (CIUO08-CL) STRUCTURE

Groups	Codes	Occupational groups	Examples of tasks
Major Group	2	Professionals	Conducting research and analysis, developing concepts, among others
Sub-major Group	25	ICT professionals	Conducting research, planning, designing and providing among others.
Minor Group	251	Software and Applications Developers and Analysts	Evaluating, planning, and designing hardware or software for specific applications, among others
Unit groups	2511	System Analysts	Consulting with users to formulate document requirements, among others
	2512	Software Developers	Researching, designing, and developing software systems, among others
	2513	Web and Multimedia Developers	Analysing, designing, and developing Internet sites and digital animations, among others

Source: Own from (ILO, 2012; INE, 2018)

Since the analysis of task content for 4-digit and 1-digit groups can result in excessively narrow or broad descriptions of occupational duties, respectively, we analyse the task content according to the aggregation of 2-digit level groups. We exclude groups representing Armed Forces occupations and others without details about tasks or not classified. Then, our sample compounds 41 2-dig occupational groups which report 845 work activities (803 unique). We assign these work activities manually to the five ALM model categories: routine cognitive (*NR*), non-routine analytic (*NRA*), non-routine interactive (*NRI*), routine manual (*RM*) and non-routine manual (*NRM*). We support this task's classification process using translated work-task dictionaries (English to Spanish) from the literature (see Table 1). Once we have done the classification, we compute the task shares index following past studies (Autor et al., 2003; Autor & Dorn, 2013; de Vries et al., 2020; Goos et al., 2014; Mihaylov & Tijdens,

2019). Our task shares index computation aims to show the relative importance of each ALM's task category for each of the 41 occupational groups by computing the proportion of work activities for a given task category over the total of work activities as follows:

$$TS_{j,k} = \frac{number\ of\ work\ activities\ in\ task\ category\ \ j\ in\ occupation\ k}{total\ number\ of\ work\ activities\ in\ occupation\ k} \quad (2)$$

where TS is the Task Share with j referring to each of the five ALM model categories, $j = \{NRA, NRI, RC, RM, NRM\}$, as defined above and, he term k represents each of the 41 2-digit occupations. As a result, we obtain five TS measures: TS_{NRA}, TS_{NRI}, TS_{RC}, TS_{RM} and TS_{NRM}, which sum one and characterize each k occupation. These TS metrics measure the variation in intensity across the occupations. To illustrate, occupations with higher values for TS_{NRA} correspond to occupations with an intense demand for NRA tasks.

4.1.2.2 Step Two: Classification Of Job Ads Into The 41 2-Digit Occupations

This step aims to classify our job ad sample according to the 41 2-digit occupations described by CIUO08-CL. This stage consists of two sub-steps. First, we pre-process the text data (e.g., cleaning, normalisation) and construct the document-term representation, DTM, based on our job ads corpus. Secondly, we "train" and evaluate our classifier algorithm, Support Vector Machines, SVM (Cortes & Vapnik, 1995), using a training dataset. SVM, when used as a classifier algorithm, has shown reasonable accuracy in text classification using job ads. See, e.g., Guerrero & Cabezas (2019) and Boselli et al. (2018) for a Chilean and Italian application, respectively.

In our analysis, we applied SVM to classify labour data into occupational categories similar to past studies (Guerrero & Cabezas, 2019; Javed et al., 2014, 2015; Lovaglio et al., 2018; Nahoomi, 2018). Once we consider a proper SVM performance, we apply our SVM algorithm to the unlabelled observations. We evaluate the SVM prediction following the metrics *precision* and *recall* and $f1-score$ (see Appendix 2.2.3 for details on how we construct these metrics). Thus, we classify our whole data set of job ads according to the Chilean standard classification system of occupations CIUO08-CL. The techniques described are implemented using R packages like Quanteda (Benoit et al., 2018) and the Python library Scikit-learn (Pedregosa et al., 2011), among others. Appendix 2 contains a detailed description of these procedures, i.e., pre-processing and DTM representation (Appendix 2.1), the SVM theoretical overview and details of the involved algorithms and their application (Appendix 2.2.1), the training dataset construction (Appendix 2.2.2), SVM evaluation and prediction (Appendix 2.2.3).

4.1.2.3 Step Three: Construction Of Task Content Time Series Variables From Results In Step One (Section 4.1.2.1) And Step Two (Section 4.1.2.2)

This section describes the construction of a time series representing our task-related measures on a monthly basis. Similar strategies followed past studies using Chilean data from household surveys (Perez-Silva & Campos-Gonzalez, 2021). Since we are interested in the impact on the skill premium, we compute measures using only job postings requiring skilled labour, JPS, by examining the educational level required by firms (see section 3). We name our task-related measures as TM, and we compute them for all j ALM model categories (see section 4.1.2.1) over all months, t, according to our data sample, $t=\{1,2,\ldots.120\}$ (see section 3). Thus, our $TM_{j,t}$ measures stand for the share of job postings devoted to j task category relative to all job postings in t considering only JPS. For instance, the $TM_{NRA,t}$ stand for the proportion of job postings devoted to the non-routine analytical task category, NRA, relative to summing all JPS in each t.

Three sub-steps compound this stage. First, using the output from Step Two above, i.e., our labelled job ads dataset with the 2-digit occupations (see section 4.1.2.2), we obtain skilled job posting frequencies for each k occupation in each t month, i.e., $JPS_{k,t}$ where $k=\{1,2,3\ldots41\}$. Secondly, we distribute each $JPS_{k,t}$ into the five j task categories using the computed $TS_{j,k}$ metrics (see section 4.1.2.1) as weights. Notably, $TS_{j,k}$ does not depend on time since we assume that the task content of occupations is constant over time, as in past studies (Reijnders & de Vries, 2018). Thirdly, we compute the numerator of our $TM_{j,t}$ by summing the weighted quantities, i.e., the product $JPS_{k,t}*TS_{j,k}$, for a given j task category over all the k occupations and the denominator by summing all JPS over all the k in t.We represent our $TM_{j,t}$ measure as follows:

$$TM_{j,t}=\frac{\sum_k\left(JPS_{k,t}*TS_{j,k}\right)}{\sum_k JPS_t} \tag{3}$$

where TM is the task measure, as explained earlier. Therefore, we obtain our five TM measures standing for each of the ALM model categories: $TM_{NRA,t}$, $TM_{NRI,t}$, $TM_{RC,t}$, $TM_{RM,t}$ and $TM_{NRM,t}$. These metrics measure the prevalence of a given task category over t periods based on the task content of occupations. The use of TS as weights allows us to consider the variation in intensity for a given task category across the occupations. Since this research focuses on how cognitive tasks drive the relative demand for skilled labour, we evaluate the influence on the skill premium of TM_{NRA}, TM_{NRI} and TM_{RC}.

A consideration in constructing our task content measures is the potential endogeneity arising from changes in educational requirements within the same

occupation over time. Such changes may reflect shifts in the skill intensity of tasks or evolving employer expectations, leading to variations in the demand for higher education within similar job titles. For instance, an occupation like 'software developer' might require a higher degree over time as the complexity of tasks increases. While this phenomenon can introduce some degree of movement along the intensive margin of skill demand (Gu & Zhong, 2023; Harrigan & Reshef, 2015), our classification of skilled and unskilled labour is based on standard definitions in the literature (see e.g., Autor et al., 2008; Ciccone & Peri, 2005). Thus, our task content measures are primarily designed to capture broader shifts in skill demand rather than short-term fluctuations in educational requirements. Future research could address this issue by focusing on within-occupation variations to better understand the role of the intensive margin in driving changes in the skill premium.

4.2. Empirical Modelling And VAR Estimation

Following the notation developed in the last section, we examine the influence of non-routine cognitive, non-routine interactive and routine cognitive tasks as follows:

$$\omega_t = \beta_0 + \beta_1 TM_{NRA,t} + \beta_2 TM_{NRI,t} + \beta_3 TM_{RC,t} + \varepsilon_t , \tag{4}$$

where ω_t is the skill premium at t (month); TM_{NRA}, TM_{NRI} and, TM_{RC} are measures of task content related to non-routine analytical, non-routine interactive and routine cognitive tasks, respectively. ε_t is a residual term.

To test our empirical models from Eq. (4), we model our time series data interactions using the VAR framework (Sims, 1980). This modelling relies on an autoregressive model applied to a series vector, allowing us to treat each variable symmetrically (Enders, 2015). Thus, every variable is specified as endogenous and, in essence, dependent on all other lagged variables. Employing standard forms of inference within the VAR specification depends on the assumption of the stationarity of the variables, where, among other things, it is assumed that "unit roots" are not present. Testing for unit roots is required since estimation and inference in VARs become non-standard in the presence of unit roots in the data. For stationarity, we apply the Augmented Dickey-Fuller (ADF) and the Kwiatkowski, Phillips, Schmidt and Shin (KPSS) (Kwiatkowski et al., 1992). We include deterministic terms such as constant and linear trends. Also, we can add a quadratic trend, an available strategy in ADF but not in our KPSS due to software limitations[6]. We also examine seasonality by including seasonal dummies.

[6] We also found this KPSS limitation in other common statistical software like EViews, Stata, and R packages like tseries.

Also, we perform lag order testing to estimate the optimal lag order for our VAR specification. The typical approach is estimating VAR models with different lag orders beginning with higher-order lags. The selected lag order relies on inspecting minimum values of statistical information criteria, such as the Schwarz Bayesian criterion (BIC) and Hannan-Quinn criterion (HQC), that penalise overfitted models. We also examine how seasonal dummies might affect the BIC and HQC by improving the information criteria values designed to select the optimal VAR considering seasonality[7]. Since we analyse monthly data, like other labour outputs, our variables are natural candidates for seasonality. In this sense, this data evolves in 12-month rounds; then, there is a potential serial correlation at the 12th lag. Therefore, we test and control for seasonality alongside our estimation strategy testing and including relevant seasonal dummies.

4.2.1. VAR Specification And Estimation

To illustrate our VAR specification, let us suppose we are interested in capturing interactions between two economic variables, $x_{1,t}$ and $x_{2,t}$. According to Patterson (2000), in the VAR representation of this bivariate problem, $x_{1,t}$ is related to both its own lagged values and those of $x_{2,t}$, and equivalently $x_{2,t}$ is linked to its own lagged values and those of x_{1t}. Thus, two dimensions feature in a VAR model: the lag order in the autoregression, p, and the number of variables, k. In a two variables application, $k = 2$, a first-order VAR, $p = 1$, is

$$\begin{pmatrix} x_{1,t} \\ x_{2,t} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} \pi_{1,1} & \pi_{1,2} \\ \pi_{2,1} & \pi_{2,2} \end{pmatrix} \begin{pmatrix} x_{1,t-1} \\ x_{2,t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{pmatrix} \tag{5}$$

where μ are deterministic terms (e.g., a constant, a deterministic trend or both), ε_t are error terms, and t is time. A multivariate VAR generalization with order p and n variables is (Enders, 2015)

$$X_t = \mu_t + A_1 X_{t-1} + A_2 X_{t-2} + \ldots + A_p X_{t-p} + \varepsilon_t \tag{6}$$

where X_t is an $(n \times 1)$ vector containing the n variables involved in the VAR, μ_t is an $(n \times 1)$ constant vector or deterministic function of time, A_i are the $(n \times n)$ matrices of coefficients and ε_t is a $(n \times 1)$ vector of i.i.d. multivariate normal error terms. To generalise the model in Eq. (6), we may add exogenous variables as explanatory variables, and the constant term μ_t might instead represent a polynomial in time.

[7] The addition of seasonal dummy variables prevents the optimal lag being equal to the seasonal period (e.g., 12 for monthly data) at the lag selection stage, since the high additive seasonality might otherwise induce a high autocorrelation at the 12th lag.

Our empirical specification from Eq. (4), modelled under the VAR framework, assumes that our VARs are first-order. However, additional lags will be included in the lag selection phase and possibly in the optimal model selection. Thus, our VAR model with a (4×1) vector of endogenous variables and, assuming a $p=1$, is:

$$\begin{bmatrix} \omega_t \\ TM_{NRA,t} \\ TM_{NRI,t} \\ TM_{RC,t} \\ \end{bmatrix} = \begin{bmatrix} \mu_{1,t} \\ \vdots \\ \mu_{4,t} \end{bmatrix} + \begin{bmatrix} \beta_{1,1} & \cdots & \beta_{1,4} \\ \vdots & \ddots & \vdots \\ \beta_{4,1} & \cdots & \beta_{4,4} \end{bmatrix} \begin{bmatrix} \omega_{t-1} \\ TM_{NRA,t-1} \\ TM_{NRI,t-1} \\ TM_{RC,t-1} \\ \end{bmatrix} + \begin{bmatrix} \sum_{i=1}^{s-1} \rho_{1,i} D_{1,i,t} \\ \vdots \\ \sum_{i=1}^{s-1} \rho_{4,i} D_{4,i,t} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \vdots \\ \varepsilon_{4,t} \end{bmatrix} \tag{7}$$

where ω, $TM_{NRA,}$, TM_{NRI} and TM_{RC} are as in Eq. (4). μ is the deterministic trend component. $\beta_{i,j}$ stands for the elements of the matrix of coefficients of lagged variables, and $\sum_{i=1}\rho_i D_{i,t}$ stands for our $s-1$ seasonal dummies D[8] (11 for our 12-month data periodicity).

The econometric estimation of Eq. (7) can be estimated using OLS on each equation. This technique is possible because all regressions have identical right-hand side variables, and the error terms are assumed serially uncorrelated with constant variance (Enders, 2015). We focus on parameter estimation from the equation with the skill premium, ω, as the target variable. We specified the deterministic component μ as a linear time trend (constant, μ_0, and trend, μt) since ω shows a trend, as noted by past studies (Gallego, 2012; Murakami, 2014), and potentially, this trend might imply non-stationarity. By adding this trend component, we can detrend the series to obtain a stationary process (Wooldridge, 2009). Also, we apply logarithms to all variables. Thus, our equation of interest for the task content analysis is

$$\ln\omega_t = \mu_0 + \mu t + \begin{bmatrix} \beta_{1,1} \ldots \beta_{1,4} \end{bmatrix} \begin{bmatrix} \ln\omega_{t-1} \\ \ln TM_{NRA,t-1} \\ \ln TM_{NRI,t-1} \\ \ln TM_{RC,t-1} \\ \end{bmatrix} + \sum_{k=1}^{s} \rho_{1,k} D_{1,k,t} + \varepsilon_{1,t} \tag{8}$$

[8] Since our data cover the January 2009 - December 2018 period, D_1 shows we are in the first month, i.e., it takes on the value one in January and zero otherwise. D_2 applies to February and so on.

where μ_0, μt, $\beta_{i,j}$, and $\rho_{,1k}$ are our parameters of interest to be examined and interpreted.

The VAR estimation parameters in our last stage allow us to perform Granger causality testing (Granger, 1969). We evaluate the Granger causality statistics for an equation where the skill premium is the dependent variable from the VAR specification of our empirical modelling. For example, for the model represented in Eq. (7), we state null hypotheses such as 'lags of TM_{NRA} do not Granger-cause the skill premium, ω', and then they can be rejected or not based on F statistics (F statistic compared to F-value and resulting p-value). Thus, in our example we assume the Granger-causality of TM_{NRA} towards the skill premium whether the coefficients estimated on the lagged TM_{NRA} in Eq. (8) are statistically different zero as a group.

To enrich our understanding of the interaction between the variables in our VAR specification, given that the Granger causality statistics may not tell us the complete story, we apply the Impulse Response Function (IRF) analysis (Lutkepohl, 2005; Neusser, 2016). The IRF allow us to examine the response of the skill premium to an impulse in another variable specified in our VAR representation described by Eq. (7) and Eq. (8). Formally, let us assume that the error term ε_t, from our multivariate VAR generalization with p order and k variables represented by Eq. (6) can be expressed as a linear function of a vector of shocks represented by u_t (Cottrell & Lucchetti, 2021). If the elements of u_t have unit variance and are mutually uncorrelated, then $V(u_t) = I$. Assuming that the errors in the VAR can be expressed as $\varepsilon_t = Ku_t$, we can write $\sum = Vcov(\varepsilon_t) = KK'$. From this configuration, we have the following sequence of matrices C_k[9], in the following equation:

$$C_k = \frac{\partial y_t}{\partial u_{t-i}} = \Theta_k K \,. \tag{9}$$

From our VAR generalization represented by Eq. (6), we can derive the IRF of the variable i to shock j. This IRF will be the sequence of the elements in the row i and column j of the sequence of matrices C_k given by Eq. (9). Using the notation given by Cottrell & Lucchetti (2021), the IRF represented by symbols is:

$$\zeta_{i,j,k} = \frac{\partial y_{i,t}}{\partial u_{j,t-k}}. \tag{10}$$

The IRF can be plotted graphically to observe and interpret the occurrence of transmission from one specific variable to our dependent variable of interest

[9] This sequence of matrices is also called the moving average representation or VMA representation. It refers to the fact that every stationary VAR process has an infinite order vector moving average representation (Cottrell & Lucchetti, 2021).

through time. Since these results are estimations of each IRFs interaction, they are endowed with confidence intervals. We perform all the analyses described in this section using Gretl (GNU Regression, Econometrics and Time-series Library) as statistical software (Baiocchi & Distaso, 2003; Cottrell & Lucchetti, 2021). Our implementation in Gretl computes the IRF confidence intervals using bootstrap techniques, considering the construction of an artificial dataset with resampled residuals and evaluated by repetitive sampling (Cottrell & Lucchetti, 2021). In our IRF plots analysis, we set the following: the bootstrap confidence interval at $1-\alpha = 0.95$, 1,999 bootstrap iterations (by default value) and a forecast horizon of 24 months. In terms of interpretation, since our variables are in logs, we can say that a 1% unexpected shock or increase in an independent variable one, two, three, etc., periods back is an increase/decrease (expressed in percentage) in the skill premium today.

Also, during the estimation process, we compute the matrix K, which is considered a known parameter in the formula given by Eq. (9). Following standard procedures in the literature (Lutkepohl, 2005), Gretl estimates K as the Cholesky decomposition of $\sum = Vcov(\varepsilon_t) = KK'$, where K is assumed to be a lower triangular matrix (Cottrell & Lucchetti, 2021). However, the Cholesky decomposition is not unique because it depends on the ordering of the variables within the vector y_t i.e., our vector of endogenous variables. This ordering is critical since K is also the matrix of IRF at lag 0, and the assumed triangularity implies that the first variable in the vector y_t responds contemporaneously only to shock number one, the second variable only to shocks one and two, and so on. Therefore, the order of our variables is meaningful where the independent variables must be placed before our target variable, the skill premium, in the variables list. As a result, the shock in the independent variables affects our target variable instantaneously, but not vice versa. In terms of interpretation, since our variables are in logs, we can say that a 1% unexpected shock or increase in an independent variable one, two, three, etc., periods back is an increase/decrease (expressed in percentage) in the skill premium today.

It is important to mention that, given the use of the Cholesky decomposition approach, we are assuming a specific causal ordering of the variables. This ordering relies on the theoretical expectation that changes in task content measures (e.g., non-routine cognitive tasks) influence the skill premium contemporaneously. Although we hypothesise that these residuals primarily capture technological shocks (e.g., the adoption of ICT and other computer-based technologies as drivers of current technological change), it is important to recognise that they may also reflect other changes in labour demand (e.g., institutional reforms).

5. Results And Discussion

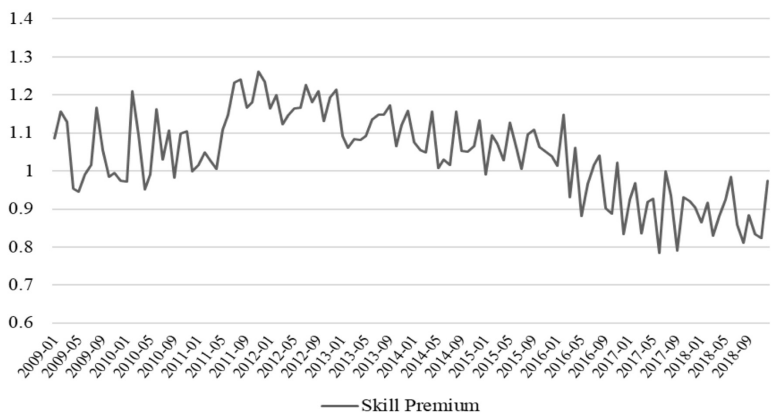
5.1. Estimation Of Variables

5.1.1.The Skill Premium

Figure 1 displays the monthly evolution of our measure for the skill premium from Jan-2009 to Dec-2018. It shows an inverted U-shaped pattern, growing to a peak of 1.26 in November 2011 and then reducing, although with fluctuations. Over two years, the skill premium average increased from 1.05 in 2009-2010 to 1.16 in 2011-2012. In turn, in 2013-2014, 2015-2016 and 2017-2018, it decreased to 1.09, 1.02 and 0.9, respectively. This pattern, composed by a reversal during most of the 2010s, has also been noted by past studies using different data sources, such as labour and household representative surveys (Campos-González & Balcombe, 2024; Murakami, 2014; Parro & Reyes, 2017).

FIGURE 1

THE SKILL PREMIUM PROXIED BY THE AGGREGATED DIFFERENCE OF OFFERED WAGES BETWEEN SKILLED AND UNSKILLED IN JOB POSTINGS: MONTHLY EVOLUTION JAN 2009- DEC 2018



Note: As stated in the section 4.1.1, the skill premium is defined as the ratio of predicted offered wages between skilled (college graduates) and unskilled (high school graduates or less) job postings, based on educational requirements in job ads.

Our estimated magnitude and pattern for the skill premium using online job ads is similar to estimations using recurrent sources like the Employment and Unemployment Survey for Greater Santiago data (in Spanish, Encuesta de Ocupación y Desempleo del Gran Santiago), EOD. The University of Chile has carried out the EOD since 1956 (University of Chile, 2020); this labour survey is generally used as source material when estimating the skill premium in Chile (Beyer et al., 1999; Campos-González & Balcombe, 2024; Gallego, 2012; Murakami, 2014; Robbins, 1994b, 1994a). Using the EOD, we estimated the average value for the skill premium for 2010-2018, around 1.06. For the same period, our estimation using job postings is 1.03. The similarity between our results and those using EOD (see a plot in Appendix 3) shows the reliability of our estimations using job posting ads data.

5.1.2. The Task Content Measures

5.1.2.1. Estimation Of The Task Content For The 41 2-Digit Occupations

Here we describe the output from the manual classification of work activities for the 41 2-digit occupational groups under analysis. On average, an occupation consists of around 20 work activities (min=5 and max= 45). Table 4 shows the global distribution of task percentage shares across occupational groups. Each row represents one of the 41 2-digit occupational groups and columns, TS_{NRA}, TS_{NRI}, TS_{RC}, TS_{RM} and TS_{NRM} depict the tasks shares of the five task types per occupational group. The score in columns TS_{NRA}, TS_{NRI}, TS_{RC}, TS_{RM} and TS_{NRM} in Table 4 ranges between zero and one. A zero score implies that a given occupational group does not contain any work activity in that task category. Alternatively, scores equal to one show that all work activities for a given occupational group belong to a unique task category. For instance, the first row of Table 4 displays the distribution of task categories for the "Chief executives, senior officials, and legislators" group. Only non-routine analytical and non-routine interactive tasks feature this occupation, given the 0.54 and 0.46 scores for the TS_{NRA} and TS_{NRI} shares, respectively. In contrast, occupations such as 2-digits codes 82, 83, 91, among others, show a score equal to zero for TS_{NRA} and TS_{NRI}.

TABLE 4

TASK CONTENT SHARES FOR THE 41 2-DIGIT OCCUPATIONAL GROUPS

2-dig Occupation Code	2-dig Occupation Name	TS_{NRA}	TS_{NRI}	TS_{RC}	TS_{RM}	TS_{NRM}
11	Chief executives, senior officials, and legislators	0.54	0.46	0.00	0.00	0.00
12	Administrative and commercial managers	0.35	0.65	0.00	0.00	0.00
13	Production and specialised services managers	0.30	0.68	0.03	0.00	0.00
14	Hospitality, retail, and related services managers	0.39	0.52	0.09	0.00	0.00
21	Science and engineering professionals	0.77	0.23	0.00	0.00	0.00
22	Health professionals	0.54	0.37	0.05	0.00	0.05
23	Teaching professionals	0.39	0.51	0.10	0.00	0.00
24	Business and administration professionals	0.50	0.44	0.06	0.00	0.00
25	ICT professionals	0.90	0.00	0.10	0.00	0.00
26	Legal, social, and cultural professionals	0.56	0.40	0.04	0.00	0.00
31	Science and engineering associate professionals (technicians)	0.39	0.21	0.27	0.06	0.06
32	Health associate professionals (technicians)	0.32	0.24	0.08	0.05	0.32

33	Business and administration associate professionals (technicians)	0.14	0.31	0.52	0.00	0.03
34	Legal, social, cultural, and related associate professionals (technicians)	0.20	0.65	0.05	0.00	0.10
35	ICT associate professionals (technicians)	0.35	0.29	0.29	0.00	0.06
36	Teaching associate professionals (technicians)	0.14	0.71	0.00	0.00	0.14
41	General and keyboard clerks	0.00	0.05	0.95	0.00	0.00
42	Customer services clerks	0.00	0.15	0.85	0.00	0.00
43	Numerical and material recording clerks	0.00	0.17	0.83	0.00	0.00
44	Other clerical support workers	0.00	0.13	0.88	0.00	0.00
51	Personal services workers	0.00	0.24	0.14	0.05	0.57
52	Sales workers	0.00	0.32	0.39	0.03	0.26
53	Personal care workers	0.00	0.67	0.00	0.00	0.33
54	Protective services workers	0.00	0.14	0.00	0.00	0.86
61	Market-oriented skilled agricultural workers and farmers	0.13	0.25	0.22	0.09	0.31
62	Market-oriented skilled forestry, fishery and hunting workers	0.05	0.20	0.05	0.05	0.65
63	Subsistence farmers, fishers, hunters, and gatherers	0.03	0.06	0.00	0.12	0.79

71	Building and related trades workers (excluding electricians)	0.00	0.11	0.00	0.00	0.89
72	Metal, machinery, and related trades workers	0.00	0.00	0.00	0.17	0.83
73	Handicraft and printing workers	0.13	0.00	0.13	0.19	0.56
74	Electrical and electronic trades workers	0.07	0.07	0.14	0.00	0.71
75	Food processing, woodworking, garment, and related trades workers	0.07	0.04	0.07	0.26	0.56
81	Stationary plant and machine operators	0.00	0.11	0.13	0.67	0.09
82	Assemblers	0.00	0.00	0.20	0.80	0.00
83	Drivers and mobile plant operators	0.00	0.00	0.00	0.06	0.94
91	Cleaners and helpers	0.00	0.00	0.00	0.00	1.00
92	Agricultural, forestry and fishery labourers	0.00	0.00	0.00	0.11	0.89
93	Labourers in mining, construction, manufacturing, and transport	0.00	0.00	0.00	0.33	0.67
94	Food preparation assistants	0.00	0.00	0.00	0.00	1.00
95	Street and related sales and services workers	0.00	0.46	0.00	0.00	0.54
96	Refuse workers and other elementary workers	0.00	0.00	0.36	0.09	0.55

Abbreviations: TS_{NRA} = non-routine analytical tasks share, TS_{NRI} = non-routine interactive tasks share, TS_{RC} = routine cognitive tasks share, TS_{RM} = routine manual tasks share and, TS_{NRM} = non-routine manual tasks share.

Also, in Table 4 we see that occupations with the two highest TS_{NRA} scores, i.e., values over 0.75 or at least with ¾ of their task content composed only of non-routine analytical tasks, are "ICT professionals" and "Science and engineering professionals". In the case of TS_{NRI}, some examples of occupations with high values are "Administrative and commercial managers" and "Production and specialized services managers" (see the second and third rows of Table 4). We see similar scores of TS_{NRI} for some occupations in the generic category of "associate professionals or technicians" (e.g., the 2-digit codes 35 and 36 in Table 4). These results are as expected since non-routine analytical and interactive work activities, such as researching, evaluating, designing, and managing, usually feature occupations performed by managers, professionals and some associate professionals or technicians. These workers are primarily highly-educated or skilled labour, given that post-secondary education provides and promotes specific knowledge and abilities, respectively. We give more insights into this relationship between non-routine analytical and interactive work activities and occupations employing skilled labour in our categorization of job ads according to the occupational classification in the next section.

5.1.2.2. Classification Of Job Ads Into The 41 2-Digit Occupational Groups

This section outlines the results of classifying our job ads sample against the 41 2-digit occupations using the SVM algorithm. The global evaluation of SVM shows that the classifier predicted as expected. The global accuracy is 0.92, and the macro and weighted averages for precision, recall and f1-score fall between 0.81 and 0.92. These results are as expected and in line with past studies, i.e., global accuracy and average precision around 0.85 (Guerrero & Cabezas, 2019). Additional details on SVM performance at occupation level can be found in Appendix 2.3.

The SVM application labelled 122,330 job ads with an occupational group. This sample plus our training dataset (67,656 job ads) represents our whole dataset (189,986 job ads, see section 3). Our labelling procedure shows six missing occupational groups in our analysis since no job ad was distributed to them[10]. Consequently, our job ads sample is distributed across 35 2-digit

[10] Unlike to the task content analysis in section 4.1.2.1 examining 41 2-digit occupational groups, our training sample is composed of only 35 occupational groups. We cannot allocate job ads to any of the following six groups (code in parentheses): (62) Market-oriented skilled forestry, fishery and hunting workers, (63) Subsistence farmers, fishers, hunters and gatherers, (63) Assemblers, (92) Agricultural, forestry and fishery labourers, (95) Street and related sales and services workers, (96) Refuse workers and other elementary workers.

occupational groups. Table 5 shows the distribution of job ads by occupational group and year, focusing on the 19 most represented groups in the dataset (these 19 groups represent 93% of the dataset). Two occupational groups, "Business and administration associate professionals (technicians)" and "Business and administration professionals", represent 34% of the sample.

Recapitulating, our tasks-related measures aim to capture the distribution of task categories across skilled labour. In this regard, Table 6 shows the composition of our sample across occupational groups. As discussed in our last section 5.1.2.1, these results align with our expectation that most occupations filled by managers, professionals and associate professionals or technicians demand skilled labour (see, e.g., the 2-digit Code Occupations 33, 21, 31 in Table 6). Regarding our sample of interest to construct measures of task content of jobs requiring skilled labour, the bottom row of Table 6 shows that our sample of job ads is 120,970.

TABLE 5

DISTRIBUTION (%) OF JOB ADS BY SELECTED 2-DIGIT OCCUPATIONS 2009-2018

2-dig Code	2- dig Name Occupation	Year										Total	
		2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	N	%
33	Business and administration associate professionals (technicians)	1,882	2,377	4,271	4,596	4,551	3,873	3,798	3,837	3,440	3,619	36,244	19.08
24	Business and administration professionals	2,098	2,838	4,315	3,759	3,157	2,737	2,761	2,410	2,413	2,792	29,280	15.41
52	Sales workers	712	1,058	2,202	2,261	2,341	1,722	1,818	1,738	1,391	1,492	16,735	8.81
21	Science and engineering professionals	854	1,422	2,370	1,966	1,522	1,283	1,599	1,467	1,201	1,442	15,126	7.96
41	General and keyboard clerks	608	941	1,570	1,575	1,511	1,192	1,100	1,121	1,001	788	11,407	6.00
42	Customer services clerks	595	593	1,517	1,398	1,234	1,868	1,602	1,217	764	510	11,298	5.95
31	Science and engineering associate professionals (technicians)	354	603	1,051	1,158	1,061	896	907	1,004	977	1,067	9,078	4.78
43	Numerical and material recording clerks	215	398	1,119	1,355	1,088	663	552	581	629	521	7,121	3.75
22	Health professionals	357	547	848	885	913	577	526	450	587	702	6,392	3.36
25	ICT professionals	604	727	1,143	790	543	447	462	550	368	540	6,174	3.25

35	ICT associate professionals (technicians)	356	401	859	811	667	534	513	649	451	478	5,719	3.01
72	Metal, machinery, and related trades workers	89	193	405	588	475	334	325	471	304	320	3,504	1.84
26	Legal, social, and cultural professionals	186	278	372	332	305	318	347	370	354	333	3,195	1.68
23	Teaching professionals	109	144	254	272	370	308	334	431	480	481	3,183	1.68
54	Protective services workers	79	233	550	485	336	218	298	321	216	188	2,924	1.54
51	Personal services workers	173	131	324	268	324	514	363	327	294	190	2,908	1.53
83	Drivers and mobile plant operators	54	109	261	327	356	275	288	364	364	283	2,681	1.41
32	Health associate professionals (technicians)	69	118	251	404	354	259	222	217	224	262	2,380	1.25
81	Stationary plant and machine operators	81	147	247	372	237	216	179	188	232	199	2,098	1.10
	Rest (16 Occupational Groups)	617	754	1,352	1,279	1,350	1,353	1,379	1,755	1,465	1,235	12,539	6.60
	Total	10,092	14,012	25,281	24,881	22,695	19,587	19,373	19,468	17,155	17,442	189,986	100

Note: 2-dig codes of the 16 occupational groups in “Rest” category: 91,74,34,12,44,36,94,53,14,75,13,71,11,73,93, and 61.

TABLE 6

JOB ADS DISTRIBUTION BY OCCUPATIONS FOR SKILLED LABOUR

2-dig Code	2-dig Name Occupation	N
33	Business and administration technicians	36,219
24	Business and administration professionals	29,280
52	Sales workers	2
21	Science and engineering professionals	15,116
41	General and keyboard clerks	-
42	Customer services clerks	4
31	Science and engineering technicians	8,902
43	Numerical and material recording clerks	1
22	Health professionals	6,346
25	ICT professionals	6,153
35	ICT associate professionals (technicians)	5,404
72	Metal, machinery, and trades workers	1
26	Legal, social, and cultural professionals	3,167
23	Teaching professionals	3,164
54	Protective services workers	54
51	Personal services workers	49
83	Drivers and mobile plant operators	31
32	Health associate professionals (technicians)	2,309
81	Stationary plant and machine operators	-
	Rest (16 Occupational Groups)	4,768
	Total	120,970

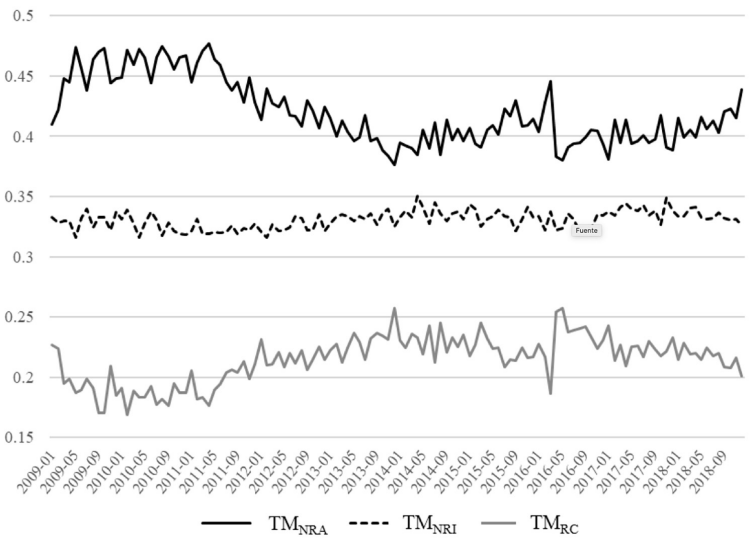
Note: 2-dig codes of the 16 occupational groups in "Rest" category: 91,74,34,12,44,36,94,53,14,75,13,71,11,73,93, and 61.

5.1.2.3. Estimation Of The Task Content Measures

Following our estimation strategy, we obtain our task measures following Eq. (3) and related statements. These metrics measure the prevalence of a given task category monthly based on the task content of occupations. The use of TS as weights allows us to consider the variation in intensity for a given task category across the occupations. Since this research focuses on how cognitive tasks drive the relative demand for skilled labour, Figure 2 plots the series representing our task content measures focusing on cognitive tasks. We focus on the task measures for the following ALM model categories: non-routine analytical (TM_{NRA} : solid black line on top of the plot), non-routine interactive (TM_{NRI} ; black dashed line at the middle of plot) and routine cognitive (TM_{RC}; solid grey line at the bottom of the plot). The TM_{NRA} series indicates that the intensity ratio of non-routine analytical tasks required by job ads for skilled labour fluctuates between 0.38 and 0.48 over the period. This measure shows an initial steady pattern, and then it decreases to grow up again, although with fluctuations.

FIGURE 2

TASK CONTENT MEASURES TIME SERIES 2009-2018



Note: TM_{NRA}, TM_{NRI} and TM_{RC} stand for non-routine analytical, non-routine interactive and routine cognitive tasks, respectively.

More generally, our results show a higher prevalence of non-routine analytical tasks in skilled occupations classified generically as managers, professionals and associate professionals or technicians in line with our expectations and previous literature (Mihaylov & Tijdens, 2019; Perez-Silva & Campos-Gonzalez, 2021; Reijnders & de Vries, 2018). Like our non-routine analytical measure, our measure stands for non-routine interactive tasks or TM_{NRI} also shows a high prevalence in these skilled occupations. Over time, TM_{NRI} fluctuated between 0.32 and 0.35, implying a more stable pattern compared to TM_{NRA}.

We also see a fluctuating pattern for the TM_{RC} measure between 0.17 and 0.2 6 but starting with an increasing trend and then a steady pattern. This measure shows the intensity ratio of routine cognitive interactive task content of jobs demanding skilled labour fluctuates. TM_{NRI} shows a narrower range, i.e. between 0.32 and 0.35, compared to the rest of the series. This pattern implies that the intensity of non-routine interactive tasks in job ads demanding skilled labour stays stable over the period. In line with expectations and previous studies (Mihaylov & Tijdens, 2019), our results show that these kinds of tasks are less prevalent in skilled jobs, with ratios fluctuating between 0.17 and 0.26. Our motivation for including this measure arose from previous studies on Chile, suggesting the relocation of skilled workers to less skilled positions due to complex software adoption as proxies for computer-based technologies (Almeida et al., 2020). Since less-skilled or middle-skilled positions are more abundant in routine cognitive tasks, as proposed by the ALM model, we might see some relationship between skilled labour and the skill premium. We return to this point in the next section on our findings on the influence of our task content measures on the skill premium.

5.2. VAR Analysis

Our results demonstrate significant co-movements between the skill premium and changes in task content measures over time, reflecting dynamic relationships in the Chilean labour market. These patterns are consistent with the predictions of the ALM framework, where technological advancements are expected to impact routine and non-routine tasks differentially. However, given the nature of our identification strategy, we refrain from making strong causal claims about the role of technology as the primary driver of these patterns. Instead, our focus is on documenting the temporal correlations that suggest a potential role of technology, leaving more definitive causal interpretations for future research.

Following formal testing described in our VAR specification and estimation techniques about stationarity (see section 4.2), our ADF and KPSS output im-

plies that our endogenous series are stationary by detrending the series using a linear and a quadratic time trend (see details in Appendix 1). Also, in our lag order testing (see section 4.2), we include the first lag order variables following our BIC and HQC results (see details in Appendix 1). Therefore, our VAR represents the interactions between our four endogenous variables, which include linear and quadratic trends and the first lag order variables. In our estimation process, we also test if the 12th variable lag (to control potential seasonality) favours the specification fitting. However, our testing cannot reject the null hypothesis that these regression parameters are zero for the 12th lag variables[11]. Therefore, we remove the 12th variable lags, which implies the estimation of a VAR with only the first lag or a VAR (1).

The results of our VAR estimation for the equation with the skill premium, ω , as the target variable are displayed in Table 7. The first three rows show that the constant, linear and quadratic trend time influence the skill premium at 10%, 1% and 1% significance levels, respectively. We do not observe influence from the lagged skill premium. Regarding our lagged task content measures, TM_{NRI} shows a positive and significant coefficient at a 5% level, and both TM_{NRA} and TM_{RC} are also positive but significant at 10%. The exposed results allow us to evaluate the dependency between variables. However, these results do not necessarily imply causality or infer how the skill premium responds to shocks in the task content variables. Hence, we apply the Granger-causality to analyse if the explanatory variables Granger-causes the skill premium and IRF analysis to examine the response of the skill premium to an impulse in another variable.

[11] We use the option given by Gretl to perform this test after the VAR estimation including the 12^{th} lag. The Wald test statistics result was Chi-square = 14.7055 and p-value>0.1 (0.538628).

TABLE 7

VAR ESTIMATION, LAG ORDER 1. OLS ESTIMATES, OBSERVATIONS 2009:02-2018:12 (T=119). RESULTS FOR EQUATION WITH THE LOGGED SKILL PREMIUM AS THE TARGET VARIABLE. SEE EQ. 8

Parameter	**Coefficient**	**Std. Error**	**t-ratio**	**p-value**	
Constant	3.9656	2.0348	1.9490	0.0538	***
Time	0.0046	0.0011	4.2140	5.11e-05	*
Time^2	–4.99e-05	0.0000	–5.698	9.96e-08	*
Skill premium, ω_{t-1}	0.1477	0.0964	1.5320	0.1282	
$TM_{NRA,t-1}$	1.4507	0.7957	1.8230	0.0709	***
$TM_{NRI,t-1}$	1.3069	0.6540	1.9980	0.0481	**
$TM_{RC,t-1}$	0.8218	0.4518	1.8190	0.0716	***
R^2	0.64				

Note: Recalling from Eq. (7) TM_{NRA}, TM_{NRI} and TM_{RC} stand for non-routine analytical, non-routine interactive and routine cognitive task content, respectively. (*), (**) and (***) denote a rejection of H_0: the regression parameter is zero at 1%, 5% and 10% significance level, respectively. All the variables, except time and time2, are in natural logs.

Table 8 outlines the Granger-causality testing results. The first column shows the stated null hypotheses and the results of F-statistic evaluation and significance level for the one lag and two lag models. We have included the results with an additional lag to show that a redundant lag still gives similar results but with some p-value changes. The TM_{NRA} variable shows significance at 10% level in both lag orders, while the TM_{NRI} is significant at 5% in both models. In the case of TM_{RC} our results show significance only in the lag one model (at 10% significance level). From these results, we assume the Granger-causality of our task content measures towards the skill premium only for the TM_{NRA} and TM_{NRI} variables at 10% and 5% of significance level, respectively.

TABLE 8
GRANGER-CAUSALITY TESTING RESULTS

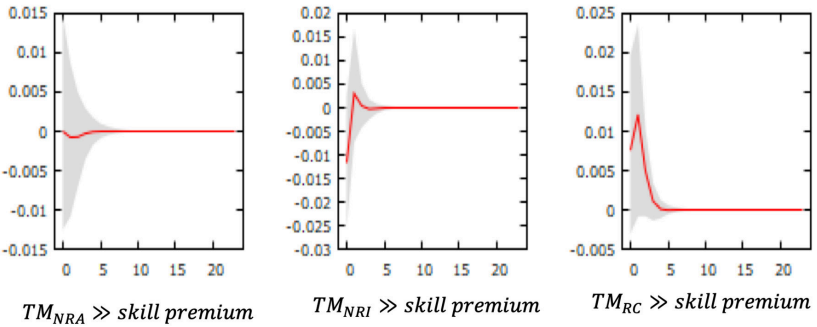
Null hypothesis	Lag Order 1 (N=112)		Lag order 2 (N=107)	
	F Statistic	p-value	F Statistic	p-value
All lags of TM_{NRA} do not Granger-cause ω	3.3243	0.0709***	2.5586	0.0821***
All lags of TM_{NRI} do not Granger-cause ω	3.9939	0.0481**	4.4575	0.0138**
All lags of TM_{RC} do not Granger-cause ω	3.3079	0.0716***	2.3304	0.1022

Note: (*), (**) and (***) denote a rejection of H_0 at 1%, 5% and 10% significance level, respectively. ω is the skill premium and, TM_{NRA}, TM_{NRI} and TM_{RC} stand for non-routine analytical, non-routine interactive and routine cognitive task content, respectively.

Related to our IRF analysis, Figure 3 displays the IRF plots where the scale refers to the sizing of the "shock" at one standard deviation of the estimated innovations in the variable stated as the origin of the impulse. Since our variables are in logs, we can say that a 1% unexpected shock or increase in a task content variable one, two, three, etc., periods back is an increase/decrease (expressed in percentage) in the skill premium today. In this regard, the left-hand plot suggests that a 1% unexpected increase in TM_{NRA} around 2-3 months back is a negligible decline in the skill premium today. In the case of TM_{NRI} and TM_{RC}, middle and right-hand plots, respectively, we see negligible increases as the response of the skill premium.

FIGURE 3

IMPULSE RESPONSE FUNCTION (IRF) PLOTS FOR THE RESPONSE OF THE SKILL PREMIUM TO A SHOCK IN TM_{NRA} (LEFT-HAND PLOT), TM_{NRI} (MIDDLE PLOT) AND TM_{RC} (RIGHT-HAND PLOT)



Note: TM_{NRA}, TM_{NRI} and TM_{RC} stand for non-routine analytical, non-routine interactive and routine cognitive task content, respectively

Although our results are weak, they support the ALM prediction about complementarity between computer-based technologies and non-routine cognitive tasks, both analytical and interactive, given the positive influence of TM_{NRA} and TM_{NRI} on the skill premium. In this regard, we show that changes in non-routine cognitive tasks may imply a greater demand for skilled labour and, consequently, a skill premium improvement. Conversely, past studies on Chile do not support this ALM prediction since they found substitution effects of computer-based technologies instead of complementarity (Almeida et al., 2020). Warnings of this substitution effect have also emerged from evidence showing a broader class of jobs at risk due to the potential ability of frontier technologies (e.g., robotics and artificial intelligence) to automate non-routine analytical or interactive tasks (Arntz et al., 2016; Autor, 2015; Frey & Osborne, 2017).

In the case of negligible influence on the skill premium of TM_{RC}, this finding is not in line with our expectations based on recent research for Chile carried out by Zapata-Román (2021), who suggested that routine tasks were important in explaining earnings variation. Relevant differences between our approaches to data and methods might explain this discrepancy[12]. However, our results agree with previous and prominent literature on other countries (Autor et al., 2003; Goos et al., 2014; Goos & Manning, 2007; Sebastian, 2018;

[12] Zapata-Román (2021) used a Chilean household survey in four waves between 1992 and 2017, with decomposition methods to observe changes in occupational structure.

Spitz-Oener, 2006). In this regard, more research is needed to understand the interactions between the skill premium and the task-related metrics analysed here, given the weakness or absence of our evidence and the potential already revealed by incipient research in this field for Chile.

We can speculate on the reasons for our lack of strong evidence on the expected role of cognitive tasks, mainly in the context of the skill premium decline observed in recent decades, as suggested by past studies (Campos-González & Balcombe, 2024; Murakami, 2014; Murakami & Nomura, 2020; Parro & Reyes, 2017). First, some suggest that the decrease in skill premium has been driven by the drop in returns to skilled labour due to the substantial expansion of Chilean tertiary education (Murakami & Nomura, 2020; Parro & Reyes, 2017). If the return to higher education, which gives knowledge and stimulates cognitive skills to perform analytical tasks, is falling, then it would be expected that this knowledge and ability has little influence on skilled labour wages. Secondly, researchers have recently reported downward movements in the occupational ladder post-2000 period, such as reassigning skilled workers to less skilled positions (Almeida et al., 2020; Zapata-Román, 2021). These downward movements could explain the declining importance of cognitive tasks and skills in explaining the wages of skilled workers.

As limitations of our analysis, we consider some characteristics of our data, potential differences between job ad salaries and survey-based wages and limitations of causal interpretation given the structure of our VAR model. Regarding data features, although we examine monthly data, the low number of observations (120 data points over 2009-2018) might not be enough to capture an adequate data variation. Additionally, categorising our global and skilled labour samples according to their occupational groups needs to be better balanced. These unbalanced data imply an over-representation of groups related to Business and administration (2-dig Code Occupations 33 and 24 in Table 5 and 6) characterized by medium or low content of non-routine cognitive analytical and interactive tasks (see 2-dig Code Occupation 33 and 24 in Table 4). Thus, observations representing non-routine cognitive task content are less represented. Future research needs to consider this potential bias towards specific occupational groups.

We also acknowledge potential differences between job ad salaries and survey-based wages[13]. Job advertisements represent the initial salary offer made by firms, which may not account for subsequent negotiations or adjustments after the hiring process. In contrast, survey-based wages reflect realised earnings, incorporating individual productivity and job-specific characteristics that emerge after matching employers and employees. Additionally, the advertised

[13] I acknowledge Estudios de Economia's reviewers for suggesting the inclusion of this discussion about job ad-salaries and survey-based wages.

wages in job postings are influenced by the strategic behaviour of firms, such as setting wage levels to attract a specific skill set or filling a particular vacancy. As a result, these job ad wages may not fully capture idiosyncratic factors affecting wage dispersion within occupations, potentially leading to an over- or under-estimation of the skill premium when compared to survey data. Despite the apparent consistency between our skill premium estimates and those derived from household surveys, as discussed in sections 3 and 5.1.1, the non-binding nature of job ad wages may introduce variability that complicates direct comparisons. Future research should consider complementing job ad data with longitudinal survey data to account for the negotiation processes and productivity-related wage differentials that occur ex-post in employer-employee relationships.

Related to limitations of causal interpretation, the interpretation of the impulse response functions (IRFs) should be approached with caution. The ordering of variables is based on theoretical considerations, but this assumption may only partially capture the contemporaneous interactions between task measures and the skill premium. Therefore, our findings primarily describe co-movements between variables rather than establish a causal relationship. Future research could employ more advanced identification methods to isolate better the effects of technological shocks from other exogenous factors.

Some policy implications, beyond our results, emerge. First, the lack of a strong relationship between the skill premium and cognitive tasks and skills might imply an unanticipated impact of technology adoption underestimated by the expected coordination between policies examining labour markets (demanding skills) and education (supplying skills). In this case, instead of the expected complementarity between technology and skilled labour, we might see a neutral or substitution effect, leading to changes in the demand for both skilled and unskilled labour. Using data from Chilean firms, Almeida et al. (2020) suggested that because of the adoption of advanced technologies like complex software, the demand for unskilled workers grew faster than for skilled labour. Here, the lower demand for cognitive tasks and skills might be a potential explanation. In this regard, the lower demand for skilled workers due to technology adoption may differ from the significant growth in supply resulting from the substantial expansion of the tertiary education system starting in the 1980s and 1990s, as discussed above.

A second policy implication arising from our lack of strong evidence about the complementarity between skilled labour and cognitive tasks could be the potential displacement of skilled labour to lesser-skilled positions. In other words, unwanted changes in the occupational ladder. For instance, displacements of skilled labour to middle-skilled positions rich in cognitive but intensive in routine tasks (Almeida et al., 2020) would push middle-skilled workers

to lower or unskilled skilled positions; in turn, these less-skilled workers can be pushed further down the occupational ladder, even affecting their chances of employment participation. These sequential downward movements represent unwanted changes in the occupational ladder for workers' educational and job prospects. Therefore, policymakers must predict these unwanted movements and mitigate their potential pervasive effects, especially among most affected employers. As pointed out above, policy efforts stimulating better coordination between higher education institutions and industry can support the development of specific skills or training systems to mitigate potential adverse effects.

6. Conclusion

The evolution of the skill premium supplies opportunities to examine how economic forces (in particular, technological change) may influence the demand for skilled labour. Research on the task content of jobs and workers' skill endowments provides material for relevant contributions to explanations of the dynamics between labour and technology, particularly the expected complementarity between cognitive tasks and more educated workers (Acemoglu & Autor, 2011; Ehrenberg & Smith, 2018; Markowitsch & Plaimauer, 2009). We examine how measures standing for cognitive work activities employing mainly skilled workers, such as reasoning, problem-solving, and persuasion, drive the skill premium. However, our analysis focuses on a period witnessing a declining trend in the skill premium when cognitive tasks might be less critical. In this regard, our results support only weakly the ALM model prediction of the complementarity between non-routine cognitive tasks and skilled labour.

Like past studies for the Chilean case, such as Almeida et al. (2020) and Zapata-Román (2021), we contribute to the recent strand of literature examining the ALM model predictions in the case of countries recently graduated from middle to high-income status. The lack of solid support for the complementarity between cognitive tasks and skilled labour is a key contribution of this study. From a policy perspective, we encourage higher levels of institutional coordination between education and labour policymakers. If the premium for analytical capability becomes less important, it might imply mismatches between skills demand and supply. Therefore, the adoption of coordinated educational and labour policies to correct these mismatches is needed. Also, our lack of solid support for the view on the complementarity between cognitive tasks, skilled labour, and technology implies a role for technological progress that would be potentially neutral or become a substitute for skilled labour. In this case, skilled workers would perform cognitive but routine tasks, which middle-skilled workers typically perform. In turn, this middle-skilled labour

would be filling lower-skilled positions. Relocating better-educated workers to less-skilled positions might imply an inefficient educational investment and produce other unwanted impacts, like deteriorating workers' prospects. Again, we highlight the importance of coordinated education and labour policies to predict and mitigate unwanted effects of technology adoption.

Referencias

Acemoglu, D., & Autor, D. (2011). Skills, Tasks and Technologies: Implications for Employment and Earnings. In O. Ashenfelter & D. Card (Eds.), *Handbook of Labor Economics* (Vol. 4B, pp. 1043–1171). Elsevier Science & Technology, Oxford, UK. https://doi.org/10.1016/S0169-7218(11)02410-5

Almeida, R. K., Fernandes, A. M., & Viollaz, M. (2020). Software Adoption, Employment Composition, and the Skill Content of Occupations in Chilean Firms. *Journal of Development Studies*, 56(1), 169–185. https://doi.org/10.1080/00220388.2018.1546847

Arntz, M., Gregory, T., & Zierahn, U. (2016). The Risk of Automation for Jobs in OECD Countries: A Comparative Analysis (*OECD Social, Employment and Migration Working Papers*). OECD. https://doi.org/10.1787/1815199X

Atalay, E., Phongthiengtham, P., Sotelo, S., & Tannenbaum, D. (2018). New technologies and the labor market. *Journal of Monetary Economics*, 97, 48–67. https://doi.org/10.1016/J.JMONECO.2018.05.008

Auria, L., & Rouslan, M. (2008). Support Vector Machines (SVM) as a technique for solvency analysis (DIW Discussion Papers). *Deutsches Institut für Wirtschaftsforschung (DIW)*. http://hdl.handle.net/10419/27334

Autor, D. (2015). Why are there still so many jobs? The history and future of workplace automation. *Journal of Economic Perspectives*, 29(3), 3–30. https://doi.org/10.1257/jep.29.3.3

Autor, D., & Dorn, D. (2013). The growth of low-skill service jobs and the polarization of the US Labor Market. *American Economic Review*, 103(5), 1553–1597. https://doi.org/10.1257/aer.103.5.1553

Autor, D., Katz, L. F., & Krueger, A. B. (1998). Computing Inequality: Have Computers Changed the Labor Market? *The Quarterly Journal of Economics*, 113(4), 1169–1213.

Autor, D., Katz, L., & Kearney, M. (2008). Trends in U. S. Wage Inequality: Revising the Revisionists. *The Review of Economics and Statistics*, 90(2), 300–323. https://www.jstor.org/stable/40043148

Autor, D., Levy, F., & Murnane, R. (2003). The Skill Content of Recent Technological Change: An Empirical Exploration. *The Quarterly Journal of Economics*, 118(4), 1279–1333.

Baiocchi, G., & Distaso, W. (2003). GRETL: Econometric software for the GNU generation. *Journal of Applied Econometrics*, 18(1), 105–110. https://doi.org/10.1002/jae.704

Banfi, S., Choi, S., & Villena-Roldán, B. (2019). Deconstructing Job Search Behavior (*MPRA Working Paper*). MPRA. https://mpra.ub.uni-muenchen.de/92482/

Banfi, S., Choi, S., & Villena-Roldán, B. (2022). Sorting On-line and On-time. *European Economic Review*, 146. https://doi.org/10.1016/j.euroecorev.2022.104128

Banfi, S., & Villena-Roldán, B. (2019). Do High-Wage Jobs Attract More Applicants? Directed Search Evidence from the Online Labor Market. *Journal of Labor Economics*, 37(3), 715–746. https://doi.org/10.1086/702627

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30), 774. https://doi.org/10.21105/joss.00774

Beyer, H., Rojas, P., & Vergara, R. (1999). Trade liberalization and wage inequality. *Journal of Development Economics*, 59(1), 103–123. https://doi.org/10.1016/S0304-3878(99)00007-3

Boselli, R., Cesarini, M., Mercorio, F., & Mezzanzanica, M. (2018). Classifying online Job Advertisements through Machine Learning. *Future Generation Computer Systems*, 86, 319–328. https://doi.org/10.1016/J.FUTURE.2018.03.035

Campos-González, J. (2025). Disasters and technological upgrading measured by changes in demand for ICT labour: Estimating the impacts with text. Nat Hazards 121, 911-957. https://doi.org/10.1007/s11069-024-06863-z

Campos-González, J., & Balcombe, K. (2024). The race between education and technology in Chile and its impact on the skill premium. *Economic Modelling*, 131, 106616. https://doi.org/10.1016/j.econmod.2023.106616

Card, D., & Lemieux, T. (2001). Can Falling Supply Explain the Rising Return to College for Younger Men? A Cohort-Based Analysis. *The Quarterly Journal of Economics*, 116(2), 705–746. https://www.jstor.org/stable/2696477

Ciccone, A., & Peri, G. (2005). Long-run substitutability between more and less educated workers: Evidence from U.S. States, 1950-1990. *The Review of Economics and Statistics*, 87(4), 652–663. https://www.jstor.org/stable/40042883

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, 20, 273–297. https://doi.org/10.1109/64.163674

Cottrell, A., & Lucchetti, R. (2021). Gretl User's Guide. https://gretl.sourceforge.net/gretl-help/gretl-guide.pdf

de Vries, G. J., Gentile, E., Miroudot, S., & Wacker, K. M. (2020). The rise of robots and the fall of routine jobs. *Labour Economics*, 66(October). https://doi.org/10.1016/j.labeco.2020.101885

Dengler, K., Matthes, B., & Paulus, W. (2014). Occupational Tasks in the German Labour Market—An alternative measurement on the basis of an expert database (*FDZ-Methodenreport*).

Ehrenberg, R. G., & Smith, R. S. (2018). Modern Labor Economics: Theory and Public Policy. Routledge, New York.

Enders, W. (2015). Applied Econometric Time Series. *John Wiley & Sons*, Ltd, New York.

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, 114, 254–280. https://doi.org/10.1016/j.techfore.2016.08.019

Gallego, F. A. (2012). Skill Premium in Chile: Studying Skill Upgrading in the South. *World Development*, 40(3), 594–609. https://doi.org/10.1016/j.worlddev.2011.07.009

Gerón, A. (2017). Hands-On Machine Learning with Scikit-Learn and TensorFlow. *O`Reilly Media, Inc.*, Sebastopol, CA.

Gil, D., & Johnson, M. (2011). Support Vector Machines in Medical Classification Tasks. In B. H. Boyle (Ed.), Support Vector Machines: Data Analysis, Machine Learning and Applications. *Nova Science Publishers, Incorporated,* New York.

Goldin, C., & Katz, L. F. (2008). The Race Between Technology & Education. *The Belknap Press of Harvard University Press*, Cambridge, MA. https://doi.org/10.2307/j.ctvjf9x5x

Goos, M., & Manning, A. (2007). Lousy and Lovely Jobs: The Rising Polarization of Work in Britain. *The Review of Economics and Statistics*, 89(1), 118–133.

Goos, M., Manning, A., & Salomons, A. (2014). Explaining Job Polarization: Routine-Biased Technological Change and Offshoring †. *American Economic Review*, 104(8), 2509–2526. https://doi.org/10.1257/aer.104.8.2509

Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, 37(3), 424–438.

Gu, R., & Zhong, L. (2023). Effects of stay-at-home orders on skill requirements in vacancy postings. *Labour Economics*, 82, 102342. https://doi.org/10.1016/j.labeco.2023.102342

Guerrero, J., & Cabezas, J. (2019). Clasificación automática de textos utilizando técnicas de text mining: Aplicación a las glosas de la Encuesta Nacional de Empleo (ENE) (Documentos de Trabajo). *Instituto Nacional de Estadísticas*.

Han, J., Pei, J., & Kamber, M. (2011). Classification: Advanced Methods. Support Vector Machines. In Data Mining: Concepts and Techniques. *Elsevier Science & Technology, Waltham*, MA.

Harrigan, J., & Reshef, A. (2015). Skill-biased heterogeneous firms, trade liberalization and the skill premium. *Canadian Journal of Economics/Revue canadienne d'économique*, 48(3), 1024–1066. https://doi.org/10.1111/caje.12167

ILO. (2012). The International Standard Classification of Occupations (ISCO-08). *International Labour Office*.

INE. (2018). Clasificador Chileno de Ocupaciones CIUO 08.CL. Instituto Nacional de Estadísticas.

Javed, F., Luo, Q., McNair, M., Jacob, F., Zhao, M., & Kang, T. S. (2015). Carotene: A job title classification system for the online recruitment domain. Proceedings - *2015 IEEE 1st International Conference on Big Data Computing Service and Applications, BigDataService 2015*, 286–293. https://doi.org/10.1109/BigDataService.2015.61

Javed, F., Mcnair, M., Jacob, F., & Zhao, M. (2014). Towards a Job Title Classification System (WSDM'14 - *Workshop on Web-Scale Classification: Classifying Big Data from the Web)*. http://hadoop.apache.org

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. How sure are we that economic time series have a unit root? *Journal of Econometrics*, 54, 159–178. https://doi.org/10.1016/0304-4076(92)90104-Y

Lovaglio, P. G., Cesarini, M., Mercorio, F., & Mezzanzanica, M. (2018). Skilg ls in demand for ICT and statistical occupations: Evidence from web-based job vacancies. Statistical Analysis and Data Mining: *The ASA Data Science Journal,* 11(2), 78–91. https://doi.org/10.1002/sam.11372

Lutkepohl, H. (2005). New Introduction to Multiple Time Series Analysis. In New Introduction to Multiple Time Series Analysis. *Springer Berlin Heidelberg*. https://doi.org/10.1007/3-540-27752-8

Markowitsch, J., & Plaimauer, C. (2009). Descriptors for competence: Towards an international standard classification for skills and competences. *Journal of European Industrial Training*, 33(8), 817–837. https://doi.org/10.1108/03090590910993652

Mihaylov, E., & Tijdens, K. G. (2019). Measuring the Routine and Non-Routine Task Content of 427 Four-Digit ISCO-08 Occupations (Tinbergen Institute Discussion Paper). *Tinbergen Institute*. https://doi.org/10.2139/ssrn.3389681

Murakami, Y. (2014). Trade liberalization and the skill premium in Chile. *México y La Cuenca Del Pacífico*, 3(6), 77–101. https://doi.org/10.32870/mycp.v3i6.418

Murakami, Y., & Nomura, T. (2020). Expanding higher education and wage inequality in Chile. *Journal of Economic Studies*, 47(4), 877–889. https://doi.org/10.1108/JES-12-2018-0445

Nahoomi, N. (2018). Automatically Coding Occupation Titles to a Standard Occupation Classification. *The University of Guelph*.

Nalepa, J., & Kawulok, M. (2019). Selecting training sets for support vector machines: A review. *Artificial Intelligence Review*, 52(2), 857–900. https://doi.org/10.1007/s10462-017-9611-1

Neusser, K. (2016). Time Series Econometrics. Springer International Publishing Switzerland.

Parro, F., & Reyes, L. (2017). The rise and fall of income inequality in Chile. *Latin American Economic Review*, 26(3), 31. https://doi.org/10.1007/s40503-017-0040-y

Patterson, K. (2000). An Introduction to Applied Econometrics, a time series approach. *Palgrave*, New York.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Perez-Silva, R., & Campos-Gonzalez, J. (2021). Agriculture 4.0? Studying the evidence for automation in Chilean agriculture. *International Journal of Agriculture and Natural Resources*, 48(3), 233–247. https://doi.org/10.7764/ijanr.v48i3.2339

Ramos, J., Coble, D., Elfernan, R., & Soto, C. (2013). The Impact of Cognitive and Noncognitive Skills on Professional Salaries in An Emerging Economy, Chile. *The Developing Economies*, 51(1), 1–33. https://doi.org/10.1111/deve.12000

Reijnders, L. S. M., & de Vries, G. J. (2018). Technology, offshoring and the rise of non-routine jobs. *Journal of Development Economics*, 135(August), 412–432. https://doi.org/10.1016/j.jdeveco.2018.08.009

Robbins, D. (1994a). Relative wage structure in Chile, 1957—1992: Changes in the structure of demand for schooling. *Estudios de Economía*, 21, 49–78. https://estudiosdeeconomia.uchile.cl/index.php/EDE/article/view/40932

Robbins, D. (1994b). Worsening relative wage dispersion in Chile during trade liberatization, and its causes: Is supply at fault? (Development Discussion Paper, p. 39). *Harvard Institute for International Development*, Harvard University.

Sebastian, R. (2018). Explaining job polarisation in Spain from a task perspective. *SERIEs*, 9(9), 215–248. https://doi.org/10.1007/s13209-018-0177-1

Sims, C. (1980). Macroeconomics and reality. *Econometrica*, 48(1), 1–48.

Spitz-Oener, A. (2006). Technical Change, Job Tasks, and Rising Educational Demands: Looking outside the Wage Structure. *Journal of Labor Economics*, 24(2), 235–270. https://doi.org/10.1086/499972

University of Chile. (2020). Encuesta de Ocupación y Desocupación en el Gran Santiago (EOD). http://www.microdatos.cl/eod

Wooldridge, J. (2009). Introductory Econometrics: A modern approach. *Cengage Learning,* Mason, OH.

World Bank. (2020). World Development Indicators. Countries Historical Classifications by Income.

Zapata-Román, G. (2021). The role of skills and tasks in changing employment trends and income inequality in Chile (WIDER Working Paper 2021/48; Wider Working Paper). *UNU WIDER: United Nations University World Institute for Development Economics Research.*

Appendix

Appendix 1. Unit-Roots And Lag Order Testing

We determine the presence of unit roots and stationarity by applying the ADF and KPSS tests, respectively, to all our endogenous variables (the skill premium, TM_{NRA}, TM_{NRI} and TM_{RC}) individually. First, we analyse the ADF results displayed in Table A.1. Columns display the variable names, modelling case (constant, "C", constant and linear trend, "C, T", and adding a quadratic trend, "C, T, TT"). Next, seasonal dummies addition, lag order (selection using criterion BIC with max order=12 given our monthly data) and test-statistics and related level of significance. ADF results show that we cannot reject the null of unit roots when modelling includes only a constant and the linear trend (the "C, T" case rows in column "Modelling case") for three of our variables (the skill premium, TM_{NRA}, and TM_{RC}). In the case of TM_{NRI}, we reject the null of unit roots for the "C, T" modelling case at 1% of significance. These results are consistent with the inclusion or not of seasonal dummies. The modelling case with a quadratic trend shows that we can reject the null of unit roots at the 1% of significance for all endogenous variables excepting TM_{NRA} variable (rejection is at 10% of significance). These ADF results with a quadratic trend are obtained with and without seasonal dummies. Overall, our ADF output implies that our endogenous series are stationary by detrending the series using a linear and a quadratic time trend.

TABLE A1
ADF RESULTS

Variable	Modelling case	Seasonal dummies	Lag order	test-statistic	p-value	
ln *skill premium*	C, T	No	2	-2.4978	0.3292	
	C, T, TT		2	-4.7600	0.0025	*
	C, T	Yes	2	-2.1447	0.5201	
	C, T, TT		2	-4.4575	0.0074	*
$\ln TM_{NRA}$	C, T	No	2	-1.9262	0.6406	
	C, T, TT		1	-5.1071	0.0006	*
	C, T	Yes	2	-1.5797	0.8012	
	C, T, TT		2	-3.6847	0.0729	***

$\ln TM_{NRI}$	C, T	No	0	-8.0841	<0.001 *
	C, T, TT		0	-8.0323	<0.001 *
	C, T	Yes	0	-7.5479	<0.001 *
	C, T, TT		0	-7.4966	<0.001 *
$\ln TM_{RC}$	C, T	No	3	-1.6561	0.7706
	C, T, TT		0	-7.9026	<0.001 *
	C, T	Yes	2	-2.055	0.5704
	C, T, TT		0	-7.0221	<0.001 *

Note: ADF H_0 = the series has a unit root. Lag order selection using criterion BIC (max was 12). (*), (**) and (***) denotes a rejection of H_0 at 1%, 5% and 10% significance level, respectively. TM_{NRA}, TM_{NRI} and TM_{RC} stand for non-routine analytical, non-routine interactive and routine cognitive tasks, respectively.

With regard to KPSS stationarity testing, we show these results in Table A2. We consider the modelling case of a constant plus a linear trend. Columns display the variable names, use of seasonal dummies, lag order (the same as in the ADF test, i.e., selection using criterion BIC with max order=12 given our monthly data) and test-statistics and related level of significance. We reject the null of stationarity at 1% of the significance level for all the endogenous variables. These results are robust to the inclusion of seasonal dummies. We confirm our ADF results for the same modelling case, i.e., constant plus linear trend or "C, T". We cannot compare the case "C, T, TT" since our KPSS implementation test the hypothesis of stationarity only around a linear trend (See section 4.2 for details). The stationarity results discussed here show that our variables are stationary around linear and quadratic trends. Therefore, we follow this modelling strategy in our VAR specification and estimation.

TABLE A2

KPSS TEST RESULTS (THE MODELLING CASE SPECIFIES A CONSTANT PLUS A LINEAR TREND)

Variable	Seasonal dummies	Lag order	test-statistic	p-value
$\ln skill\ premium$	No	2	0.7248	<0.01*
	Yes	2	0.7453	<0.01 *
$\ln TM_{NRA}$	No	2	0.5622	<0.01*
	Yes	2	0.5746	<0.01 *
$\ln TM_{NRI}$	No	0	0.2778	<0.01*
	Yes	0	0.2933	<0.01 *
$\ln TM_{RC}$	No	3	0.4548	<0.01*
	Yes	2	0.4611	<0.01 *

Note: KPSS H_0 =the series is stationary. Lag order is the same as in ADF (see Table). p-values as in Gretl output. (*), (**) and (***) denotes a rejection of H_0 at 1%, 5% and 10% significance level, respectively. TM_{NRA}, TM_{NRI} and TM_{RC} stand for non-routine analytical, non-routine interactive and routine cognitive tasks, respectively.

Regarding the optimal lag order testing, Table A3 shows the results for the tested VAR models from 1st to 12th lag order. We display the results differentiating because of the addition of seasonal dummies. The optimal number of lags to include is one, based on the minimized values of the BIC and HQC values. However, more importantly, the addition of all seasonal dummies worsened these values, i.e., a bigger penalization due to the increased number of parameters. Therefore, in our VAR estimation, we assess the addition only of the 12th lag to control the potential seasonality since the nature of our labour data can be highly seasonal alongside the first lag order following our BIC and HQC results.

TABLE A3
OPTIMAL LAG ORDER FOR THE VAR

Lags	BIC				HQC			
	Seasonal dummies				Seasonal dummies			
	No		Yes		No		Yes	
1	-17.016	*	-15.668	*	-17.429	*	-16.731	*
2	-16.644		-15.269		-17.294		-16.568	
3	-16.296		-15.040		-17.182		-16.576	
4	-15.792		-14.579		-16.914		-16.350	
5	-15.397		-14.205		-16.755		-16.213	
6	-14.905		-13.771		-16.499		-16.016	
7	-14.631		-13.537		-16.461		-16.018	
8	-14.215		-13.219		-16.282		-15.935	
9	-13.704		-12.844		-16.007		-15.797	
10	-13.382		-12.554		-15.922		-15.743	
11	-13.098		-12.250		-15.874		-15.675	
12	-12.592		-11.905		-15.604		-15.567	

Note: Results estimated from VAR systems of order 1 to max. lag order 12. The asterisks indicate the best lag order, that is, the minimized values of the respective information criteria. VAR model with constant, linear and quadratic trends and our four endogenous variables (the skill premium and the task-related measures).

Appendix 2. Pre-Processing Of Text Data And The Svm Implementation

Appendix 2.1. Pre-Processing And Dtm Representation

The pre-processing stage starts with the concatenation of the three open-text variables: job title, job description and job-specific requirements. We perform a set of standard techniques on the new concatenated variable to convert words to lowercase, remove Spanish stop words, punctuation and special symbols, tokenisation, and stemming (words being reduced to their word stem). The tokenisation allows the text content to be split into words (unigrams) and groups of two consecutive words (bigrams) denoted as tokens or features. In this study, we use unigrams and bigrams. Bigrams can be more representative for job titles composed of two words (e.g., job titles preceded by generic words like, in Spanish, "Ingeniero" ("Engineer" in English) such as "Ingeniero Informático" ("Informatics Engineer" in English).

Based on these features, we build a DTM, which shows the collection of job ads or documents represented in the vector space model. In the DTM, job ads and tokens are rows and columns, respectively. The DTM represents the corpus as a bag of words and is usually sparse; it is the primary input for SVM. We applied these pre-processing and DTM techniques to, firstly, our training sample and, secondly, our unlabelled observations to classify the job ads against the Chilean standard classification system of occupations CIUO08-CL.

Appendix 2.2. The SVM Application

The SVM, initially known as support-vector networks, is an algorithm developed by Cortes & Vapnik, (1995). SVM is a real-world-oriented application (Nalepa & Kawulok, 2019; Smola & Scholkopf, 2004) that researchers have successfully performed on classification analysis in multiple fields due to its capacity to learn from data to attain the best separation between classes or groups of data (Gil & Johnson, 2011).

Overall, the SVM algorithm predicts occupational category labels according to a subset of training data already labelled with their 2-digit occupational group or the training dataset (we detailed the training sample construction in the Appendix 2.2.2 below). SVM uses a set of functions to convert the training data into a high-dimensional space to find one or multiple optimal separating hyperplanes. An ideal hyperplane separates one class from another based on the support vectors, which refer to the critical training instances that define its margins; therefore, they give the most information about the classification (Han et al., 2011). This hyperplane should also stay as far away from the nearest training instances as possible (Gerón, 2017). We evaluate the predictive

capability of our SVM using measures developed for this purpose as detailed below (see Appendix 2.2.3).

Appendix 2.2.1. SVM Theoretical Overview And Implementation

To illustrate how, theoretically, an SVM classifier achieves its goal, we will assume a two-class problem with a dataset D linearly separable. D refers to $(X_1, y_1), \ldots, (X_{|D|}, y_{|D|})$ where X_i corresponds to the set of training instances labelled according to their class, y_i, which can take the values of +1 or -1. Since our data is linearly separable, graphically, we can draw infinite straight lines between the two classes. The SVM searches a separating hyperplane, the maximum marginal hyperplane, to discriminate between the classes in a high dimensional space. Simultaneously, the margins refer to the shortest distance between the hyperplane and the closest training instance of either class (Gerón, 2017; Gil & Johnson, 2011; Han et al., 2011). Following Han et al. (2011), we write the separating hyperplane as:

$$W \cdot X + b = 0 \tag{11}$$

where W is a row vector of weights, $W = (w_1, w_2, \ldots, w_n)$ and n is the number of attributes. In our assumption, we have two classes, denoted by a column vector $X = (x_1, x_2)$ where x_1 and x_2 are the values of attributes. b is a scalar usually associated with bias. We re-write Eq. (11) as:

$$b + w_1 x_1 + w_2 x_2 = 0 \tag{12}$$

Therefore, any data point located above or below the separating hyperplane satisfies Eq. (13) and Eq. (14), respectively

$$b + w_1 x_1 + w_2 x_2 > 0\,, \tag{13}$$

$$b + w_1 x_1 + w_2 x_2 < 0\,. \tag{14}$$

We can define the sides of the maximal margin using new hyperplanes, h_1 and h_2, based on the adjustment of weights as follow:

$$h_1 : b + w_1 x_1 + w_2 x_2 \geq 1 \text{ for } y_i = +1, \tag{15}$$

$$h_2 : b + w_1 x_1 + w_2 x_2 \leq 1 \text{ for } y_i = -1\,. \tag{16}$$

If any training data point falls on or above h_1 it will belong to class $+1$ while any training data point that falls on or below h_2 will belong to class -1. The combination of inequalities from Eq. (15) and Eq. (16) yields:

$$y_i\left(b + w_1x_1 + w_2x_2\right) \geq 1, \forall i. \tag{17}$$

The support vectors will be any training data point that falls on the sides of the maximal margin; this is the hyperplanes h_1 and h_2. Since the support vectors satisfy the Eq. (17) and are located equally near the separating hyperplane, we can use this expression to find the maximal margin between h_1 and h_2 (Gil & Johnson, 2011). By definition, the distance from a point $\left(x_0, y_0\right)$ to a line $a_x + b_x + c = 0$ is $\left|a_x + b_x + c = 0\right| / \sqrt{a^2 + b^2}$, therefore, the distance from any support vector on h_1 to the separating hyperplane is $|W \cdot X + b| / \| W \|$which is equal to $\frac{1}{\| W \|} \cdot \| W \|$ is the Euclidean distance from the origin to W, that is $\sqrt{W \cdot W}$ (recalling from Eq. (11), $W = \{w_1, w_2, \ldots, w_n\}$, then $\sqrt{W \cdot W} = \sqrt{w_1^2, w_2^2, \ldots, w_n^2}$). Since this distance is the same from any support vector on h_2 to the separating hyperplane, the maximal theoretical margin possible is $\frac{2}{W}$. Therefore, to maximize the separating hyperplane, the value of W need to be minimised with the condition given by Eq. (17) to avoid training data points falling between h_1 and h_2. We can re-write the problem as a quadratic programming problem as follows:

$$\min_{W,b} \frac{W^2}{2}, \tag{18}$$

concerning the constrain represented by Eq. (17). This formulation is usually known as the primal form (Nalepa & Kawulok, 2019), which is the problem to be solved applying SVM. SVM performs a set of mathematical functions and procedures and transformations, the so-called "fancy math tricks" (Han et al., 2011), to find the separating hyperplane and the support vectors. These mathematical approaches include Lagrangian formulations, Karush-Kuhn-Tucker conditions, and kernel functions (e.g. linear, polynomials) to handle linearly inseparable data, among others (Gil & Johnson, 2011; Nalepa & Kawulok, 2019).

To implement our SVM strategy, we use the Scikit-learn software library via an interface in Python (Pedregosa et al., 2011). We apply the linear support vector classification, LinearSVC, to solve our multi-class optimization problem. LinearSVC applies the one-vs-the rest strategy to fit one classifier per class. Thus, to obtain knowledge about a particular class, we evaluate only its computed classifier. In parameter tuning, we control the balance between max-

imising the margin and reducing misclassification using the C parameter, as explained below, and the potential imbalance between classes using weights. LinearSVC formulates Eq. (18) equivalently as:

$$\min_{W,b} \frac{1}{2} W^T W + C \sum_{i=1}^{n} \max\left(0, 1 - y_i \left(W^T \phi\left(x_i\right) + b\right)\right), \tag{19}$$

where ϕ is the identity function, and C is a real and positive constant. The LinearSVC algorithm uses the math tricks noted above with a linear kernel to optimize Eq. (19) with C as a tuning parameter. According to Gerón (2017), C controls the balance between keeping the maximal margin as wide as possible and limiting the misclassifying i.e., when instances fall in the middle of the margin or even on the wrong side. We use $C = 1$, which is the recommended value for LinearSVC (Pedregosa et al., 2011). The same parameter has been used by Guerrero & Cabezas (2019) in their study classifying occupations for Chile from national labour surveys. We also experiment with alternative values to analyse impacts on classifier performance. A lower value for *C* gives more regularization if the data contains a high number of noisy observations, and higher values for *C* (e.g., 10, 100) result in a lower generalization ability of the classifier. This impact on generalization means that SVM may classify appropriately on the training stage but its performance on new samples would be poor (Auria & Rouslan, 2008).

We also control for the expected unbalance between classes due to the natural distribution of labour. For instance, clerical workers have a significantly higher representation than managers. We apply weights to optimize the classifier performance on less represented classes. In LinearSVC, we include class weights inversely proportional to the class frequencies using the formula $weight_j = n / \left(k * n_j\right)$ where $weight_j$ is the weight to class j, n is the number of observations in the dataset, k is the number of classes and n_j is the number of observations in class j.

Appendix 2.2.2. The Training Dataset Construction

The SVM, as noted above, requires a training dataset, i.e., a job ads sample already labelled with their 2-digit occupational group code. We start by selecting the most frequent job titles. We filter 3,359 job titles whose frequencies vary between 2,000 and 20. At this point, our training sample is a subset of 67,656 job ads, i.e., 35% of our whole job ads dataset described in section 3. The rest of our whole dataset, i.e., the 122, 330 unlabelled job ads, will be labelled using SVM. The distribution of our training sample in terms of industry and educational category is similar to the whole dataset distribution as shown in Table A4. We manually label each job ad from our training sample according

to the Chilean classification CIUO08-CL, supporting this labelling by observing the educational category, economic area and job task descriptions reported by CIUO08-CL. We also support our labelling process by examining the training dataset employed by Guerrero & Cabezas (2019), which was prepared by domain experts from the Chilean National Institute of Statistics (in Spanish Instituto Nacional de Estadísticas). As observed above, since algorithms do not deal directly with text data but perform on specific text features, we apply the procedures detailed in the Appendix 2.1 to obtain our DTM representation of the training data. We use 80% of the training sample to train the SVM and the rest (20%), our testing dataset, to evaluate the SVM performance.

TABLE A4

DISTRIBUTION (IN PERCENTAGES) OF EDUCATIONAL AND INDUSTRY CATEGORIES FOR THE WHOLE AND TRAINING DATASETS

Educational and industry categories		**Whole dataset**	**Training dataset**
Educational category			
	Primary	2%	1%
	Secondary	20%	15%
	Secondary Education Technician	15%	16%
	Higher Professional Technician	29%	34%
	Graduate	34%	33%
	Postgraduate	1%	0%
	Number of observations	189,986	67,656
Industry			
	Agriculture and fishing	1%	1%
	Commerce	19%	16%
	Communication	9%	8%
	Construction	4%	5%
	Electricity, water and gas	2%	2%
	Financial services	6%	4%
	Industry	17%	17%
	Mining	2%	2%
	Other activity	5%	13%

Other services	8%	8%
Personal services	19%	18%
Public Administration	1%	0%
Restaurants and Hotels	2%	2%
Transportation	5%	5%
Number of observations	189,986	67,656

Appendix 2.2.3. SVM Evaluation And Prediction

To evaluate the SVM classification performance, we use metrics (e.g., *accuracy*, *precision*) based on four outputs by comparing the labelled categories with those predicted by SVM using our testing dataset. These outputs are true negatives, TN, when the observation is negative and predicted negative; false negatives, FN, when the observation is positive but predicted negative; true positives, TP, when the observation is positive and predicted positive; and false positives, FP, when the observation is negative but predicted positive. Typically, the classifier accuracy is calculated as the ratio of all correct predicted observations to the total number of observations, as follows:

$$accuracy = \frac{TN + TP}{TN + FN + TP + FP}.$$

We also examine *precision*, *recall*, and $F1$ *score*, which are standard metrics used to evaluate the classifier performance at the global and class level. The *precision* and *recall* measures refer to ratios to measure the ability of SVM to avoid labelling as positive an observation that is negative and to find all the positive observations, respectively and $F1$ *score* corresponds to the weighted harmonic mean of both metrics (Pedregosa et al., 2011). The formulation of these metrics is:

$$precision = \frac{TP}{TP + FP},$$

$$recall = \frac{TP}{TP + FN},$$

$$F1\,score = 2 * \frac{\left(precision * recall\right)}{\left(precision + recall\right)}.$$

Intuitively, *precision* counts for the number of observations correctly classified among that class and *recall* quantifies the number of cases for a given class found by the classifier over the total number of class cases. These metrics

can also evaluate global performance by calculating averages (*avg precision* or *avg accuracy*) which can take classes' imbalance into account. We also examine *macro avg F*1 *score*, the unweighted mean of *F*1 *score*, which results in higher penalization if the classifier does not perform appropriately with less represented classes since all classes have the same weight. We also examine *weighted avg F*1 *score*, which uses as weights the number of true positives for each class. This weighted version adjusts *macro avg F*1 *score* to account for class imbalance. The results from these metrics are presented in the form of a classification report which outlines the results from these metrics at the global and class level (see Appendix 2.3 below).

Also, some adjustments related to tokens or features (e.g., number, frequency) or balancing between classes can impact these measures.

Appendix 2.3 Results From Evaluating The SVM Algorithm

The DTM representation of our training sample (as discussed above in section 0) corresponds to a matrix of 67,656 documents and 210,689 features (unigrams and bigrams). Once we have "trained" our SVM algorithm, we evaluate the SVM prediction following the metrics *precision*, *recall* and $f1-score$) as discussed above. Table A5. displays the classification report from the results of the SVM evaluation.

The first column in Table A5 refers to 35 2-digit occupational codes (see footnote 9), and in the subsequent columns, we see the metrics *precision*, *recall* and $f1-score$ results at the occupation level. The last column shows the number of occurrences of the occupation in the training dataset. We can see that the predictive performance of SVM depends on the analysed occupational group, with better results in occupations with higher representation in the sample. Overall, by observing the global evaluation of SVM in the bottom rows of Table A.5., we see that *global accuracy* is 0.92 and the *macro* and *weighted* averages for, *recall* and $f1-score$ fall between 0.81 and 0.92.

TABLE A5

CLASSIFICATION REPORT FOR THE SVM (LINEARSVC) APPLICATION

2-dig Code	precision	recall	f1-score	N support (80% training sample)
11	0.91	0.54	0.68	143
12	0.81	0.76	0.78	571
13	0.76	0.48	0.59	219
14	0.83	0.37	0.51	94
21	0.86	0.88	0.87	3,867
22	0.93	0.97	0.95	1,418
23	0.95	0.93	0.94	533
24	0.94	0.94	0.94	8,894
25	0.89	0.88	0.88	1,661
26	0.91	0.95	0.93	1,030
31	0.9	0.86	0.88	2,589
32	0.96	0.96	0.96	805
33	0.97	0.97	0.97	12,437
34	0.86	0.8	0.83	571
35	0.86	0.92	0.88	1,653
36	0.91	0.7	0.79	211
41	0.9	0.86	0.88	3,970
42	0.84	0.87	0.85	1,420
43	0.89	0.93	0.91	2,379
44	0.93	0.86	0.89	578
51	0.92	0.86	0.89	772
52	0.91	0.96	0.93	3,748
53	0.75	0.78	0.77	99
54	0.97	0.99	0.98	902
61	0.9	0.52	0.66	86
71	0.88	0.68	0.77	117
72	0.84	0.87	0.85	885
73	0.93	0.93	0.93	75
74	0.81	0.78	0.8	497
75	0.68	0.69	0.68	106
81	0.81	0.66	0.73	479
83	0.93	0.98	0.95	634
91	0.92	0.99	0.95	530
93	0.93	0.52	0.67	50
94	0.93	0.85	0.89	102
global accuracy			0.92	54,125
macro average	0.88	0.81	0.84	54,125
weighted average	0.92	0.92	0.91	54,125

Appendix 3. Comparing The Skill Premium Estimation From EOD And Online Job Ads

FIGURE A1

COMPARING THE SKILL PREMIUM STIMATION BETWEEN JOBS ADS DATA FROM TRABAJANDO.COM AND EOD, FOR 2009-2018 (TRABAJANDO DATA GROUPED BI-ANNUALLY)