



A Data-Driven Approach to Job Skills and Occupations: Evidence from Latin America

Charles River Economics Labs

Utkarsh Dandanayak (Project Lead)

Kiran Duggirala, Ishaan Goel, Giyoung Kwon, Katherine Papen, Prakhar Saxena
(Analysts)*

Spring 2024

*The authors are from the University of Chicago. They are grateful to Harry Moroz at the World Bank for his invaluable support and guidance. We also thank Mauro Pelucchi at Lightcast for his insights and advice on our project.

Contents

1	Team and Project Overview	3
1.1	Charles River Economics Labs	3
1.2	Partner Organization	3
1.3	Results in Brief	3
2	Introduction	3
3	Data	4
4	A Data-Driven Skills Taxonomy	5
4.1	Literature	5
4.2	Methodology	6
4.2.1	Continued Pre-training of Language Model	6
4.2.2	Skill Embedding	7
4.2.3	Skill Clustering	7
4.2.4	Taxonomy Generation	8
4.3	Results	9
4.3.1	Skill Clusters	9
4.3.2	Skill Hierarchy and Taxonomy	9
4.4	Evaluation	10
5	A Skills-Based Approach to Occupational Similarity	12
5.1	Literature	12
5.2	RCA	12
5.2.1	Methodology	12
5.2.2	Results	13
5.3	OpenAI Embeddings	13
5.3.1	Methodology	14
5.3.2	Results	14
5.4	Evaluation	15
6	Conclusion and Further Research	19
7	Appendix	21

1 Team and Project Overview

1.1 Charles River Economics Labs

Charles River Economics Labs is a group of undergraduate researchers at the University of Chicago, Harvard, and Stanford that completes pro-bono economics research for governments and other social enterprises. The organization is advised by graduate researchers and faculty members at Harvard and the University of Chicago. Its past and current research partners include the United Nations Development Programme, the Department of Transportation, the Environmental Protection Agency, the Brookings Institution, and the US Department of State, among others.

1.2 Partner Organization

The Social Protection and Labor Global Practice at the World Bank develops policy solutions to alleviate poverty and improve employment-related outcomes for predominantly low-income households. They provide financing for social protection and job programs to serve the poor, and offer advisory and analytical services pertaining to these programs for governments. The World Bank is currently “providing close to \$29 billion in financing” globally through its social protection and jobs programs¹. A team of its economists are currently working on an initiative to strengthen labor market information systems in Argentina and Uruguay.

1.3 Results in Brief

The primary contribution of our project is methodological in nature. First, we develop a new, data-driven taxonomy to hierarchically categorize job skills based on a large dataset of online job postings in Latin America. We take advantage of recent developments in Natural Language Processing (NLP) to generate context-specific representations of these skills, and compare our results to that of a “manually” generated taxonomy (developed by labor economists at the World Bank). Second, we analyze similarities between occupations using two different approaches - one that is grounded in the economic literature and is based on the idea of “revealed comparative advantage”, and another that is grounded in the data and uses embeddings-based techniques. Our analysis suggests that there are notable differences between traditional and data-driven approaches to understanding job skills and occupations. We discuss the nuances of these differences in further detail below.

2 Introduction

Our work forms an intermediate component of an ongoing World Bank project on strengthening labor market information systems (LMIS) in Argentina and Uruguay. In particular, the initiative is specifically focused on **labor market intelligence**, seeking to encourage governments to take

¹Accurate as of February 2024. See [here](#) for more information.

advantage of novel sources of granular labor market data in order to facilitate the effective development of job support programs. Labor market information systems are important for the efficient functioning of labor markets - among other things, they expedite job transitions during periods of disruption, and enable human capital reallocation in response to changes in the nature of labor demand. This is particularly relevant today, given the fact that labor markets in many countries are currently experiencing significant structural changes (e.g. the introduction of artificial intelligence tools in the workplace, the transition away from fossil fuels towards green energy etc).

Advanced LMISs facilitate a detailed analysis of labor market trends, and offer opportunities to respond to those trends. If individuals lose their job as a result of labor market disruptions, LMIS mechanisms allow those individuals to rapidly transition to alternative roles which align with their previous skills and experiences. Therefore in order to fulfil their role in human capital reallocation, LMISs should be supplemented with a granular understanding of the relationships between different job skills and occupations. In our work we sought to contribute to this understanding using both classical and modern techniques for labor market data analysis. In particular, we:

1. create a novel job skills taxonomy designed specifically for Latin America using NLP techniques.
2. map similarities between occupations across entire network of ISCO occupations using skills-based and embeddings-based methods.

Our analysis proceeds as follows. Section 3 describes the job postings data that is used in our subsequent analysis, Section 4 explains the procedure for generating our new job skills taxonomy using NLP, and Section 5 discusses two methods for calculating occupational similarity (RCA and embeddings-based techniques) and evaluates the differences between them. We conclude in Section 6 and discuss opportunities for further research in this field.

3 Data

As discussed above, the ongoing World Bank project aims to take advantage of newly developed sources of detailed labor market data in Argentina and Uruguay. Our study specifically is grounded in a dataset of online job postings collected by Lightcast, a leading provider of labor market data. The dataset contains several million job postings across Argentina, Uruguay, and Chile in the period 2020-2023. We have access to each job posting’s ISCO classification (an internationally recognized standard for occupational classification), the date it was posted, a list of associated skill requirements for the job, and the raw posting text. A smaller proportion of postings also include information on salary, although we do not use this in our analysis.

We encountered a number of challenges in processing and analyzing this data, particularly when working with the raw job posting text. Firstly, almost all of the job postings were in Spanish, so we began by implementing an automated procedure for translating these postings at scale. Many postings also contained irrelevant HTML elements, so we also cleaned each posting of these

elements. Due to computational constraints, we only translate and process a (representative) sample of 500,000 job postings for use in most of our subsequent analysis. We discuss the caveats associated with our sampling procedure in further detail below.

4 A Data-Driven Skills Taxonomy

4.1 Literature

Djumalieva et al. (2018) and Gallagher et al. (2022) serve as primary benchmarks for our study, both employing a data-driven approach to develop a new skills taxonomy in the UK. Similar to this project, Gallagher et al. (2022) aimed to extract skill names from online job postings and generate hierarchical skill clusters based on embedded skill sentences. It utilized a transformer-based Large Language Model for embedding and the DBSCAN algorithm for clustering, methodologies shared with our research. However, our study advances these previous projects in two significant ways. First, we employ a continually pre-trained version of the BERT model, rather than the default version, to enhance domain-specific accuracy in skill embeddings and clusters. Second, rather than manually transforming skill names post-clustering, we propose using the Generative Pre-Trained Transformer (GPT) to automatically generate interpretable labels for each cluster. Another difference between this study and previous studies, unrelated to methodology, is that skill extraction from raw job postings was not part of our data generation process. The primary reason is that the extraction was already completed and provided to us. But this setting also implies a methodological advantage compared to previous studies, as it allows us to achieve more accurate skill embeddings by eliminating uncertainties associated with skill extraction, which is based on clusters of skill sentences.

Gu et al. (2021) and Ji et al. (2023) emphasize the advantages of domain-specific pre-training of Large Language Models, particularly for tasks requiring precise handling of expert jargon (e.g., medical terms) not commonly included in the default corpus on which Large Language Models are pre-trained. Since the skills involved in our clustering task encompass not only generic skills (e.g., Python) but also expert skills (e.g., cardiovascular imaging), we pre-trained the basic BERT model with an additional corpus to provide the model with extra context about the skills relevant to our study and clustering efforts.

Grootendorst (2019) provides a comprehensive overview of BERTopic, a topic modeling technique that integrates BERT and TF-IDF as its core methodological components. The study compares the performance of BERTopic with conventional topic modeling models such as Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF). The highlighted advantages of this framework are its ability to parallel the state-of-the-art development of NLP models, the separation of procedures from document embedding to topic representation, and the dynamic topic representation using TF-IDF. Ao et al. (2023) and Shehu and Gjika (2024) illustrate the application of BERTopic within the context of the job market. Although these studies do not prominently highlight BERTopic’s comparative advantage over conventional models, this is attributable to the

fact that the effectiveness of topic modeling is highly task-specific. In our study, for instance, the multilingual nature of our data and the generative nature of our final output render BERTopic the most suitable framework due to its structural flexibility.

4.2 Methodology

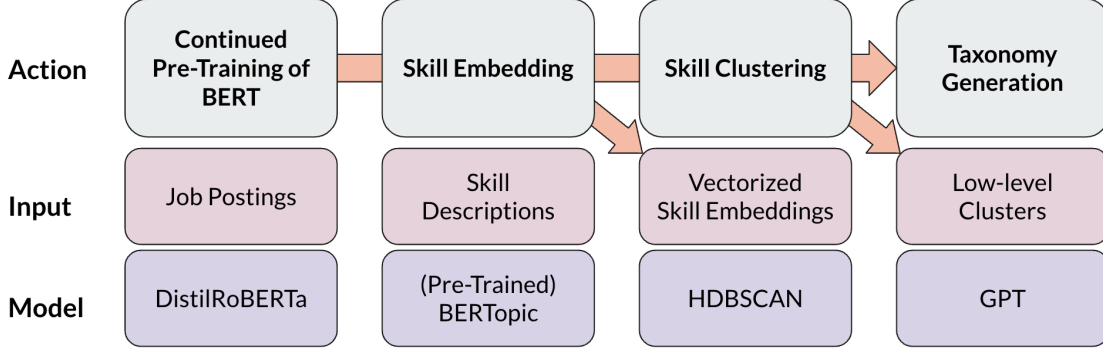


Figure 1: Pipeline Overview

As depicted in Figure 1, our methodology for generating the new skills taxonomy from job postings and skill descriptions data encompasses four distinct steps. Initially, we extended the pre-training of a Bidirectional Encoder Representations from Transformers (BERT) variant using the textual content of job postings. Subsequently, we pre-processed the skill descriptions text and transformed them into embedding vectors utilizing the model derived from the first step. It is important to note that the skill descriptions were excluded from the continued pre-training of BERT, as these descriptions, sourced from Lightcast, are predominantly based on Wikipedia pages, a corpus already extensively covered during BERT’s original pre-training. In the third step, we employed the HDBSCAN algorithm to generate hierarchical clusters of the embedded skill descriptions. Finally, we utilized the ChatGPT API to rename the clusters, producing more interpretable labels. The computations for the second through fourth steps were conducted using methods provided by the BERTopic package. In the following sections, we delineate the details of each methodology.

4.2.1 Continued Pre-training of Language Model

BERT (Bidirectional Encoder Representations from Transformers), introduced by Google in 2018, is a pre-trained language model designed to understand the context of words in a sentence bidirectionally by looking at the surrounding text. Leveraging a transformer architecture, BERT can capture complex dependencies in language, making it highly suitable for the job skills embedding task where understanding nuanced relationships between terms is crucial. DistilBERT, introduced by Hugging Face in 2019, is a smaller and more efficient version of BERT, making programs that use it faster and more reproducible due to its lighter model size with only insignificant loss of performance for computationally lighter tasks.

Specifically, we continued the Masked Language Model pre-training process of DistilBERT to expand its baseline corpus by feeding it tokenized job postings. We utilized Hugging Face’s Trainer API for hyperparameter tuning and implementation. Since the continually pre-trained DistilBERT model itself is not used to achieve our final task (clustering), there was less emphasis on minimizing training loss by spending extensive time and running over many epochs.

Given the vast number of job postings available, to ensure the model’s scalability (both in terms of computational efficiency and external validity of the results), it must be robust to finite sample sizes up to a certain level. To address this, instead of using the entire dataset of nearly 2 million job postings, we randomly sampled 500,000 postings that share the same distribution of manual skill descriptions as the entire population and used them as our final input.

4.2.2 Skill Embedding

We pre-processed and embedded the job skill descriptions prior to inputting them into the clustering algorithm. Instead of utilizing the skill names directly, we employed the first sentences from each skill’s detailed descriptions as our model input. This approach was adopted because skill names alone often provide limited information for generating embeddings, particularly when the names do not clearly reflect the actual content of the skill or when they consist of acronyms (e.g., Python, AWS)

From skill embedding to taxonomy generation, the theoretical methodologies employed in this study were implemented using multiple classes and methods provided by the BERTopic package in Python. The modularity of BERTopic allowed us to select the most suitable methods for our dataset at each step and integrate them through a single function call. Utilizing an open-source library like this significantly simplifies the clustering generation pipeline compared to previous attempts, which indeed aligns with one of our practical objectives despite its decreased degree of freedom to choose the optimal algorithm.

Following the extraction of the initial sentences from each skill description, we used the basic NLTK package in Python to eliminate stop words (e.g., articles) and excessively common but uninformative keywords (e.g., professional, certificate). This step was crucial for preventing incorrect cluster representations. For instance, without removing stop words, BERTopic could generate clusters dominated by words like “the, an, is,” which are not meaningful for our purposes.

4.2.3 Skill Clustering

In the previous section we described how we obtained an embedding model which could represent the (cleaned) skill descriptions as vectors. The next step was to cluster these description vectors to obtain meaningful skill categories. The approach we used was HDBSCAN (Hierarchical Density Based Spatial Clustering of Applications with Noise), a non-parametric clustering algorithm which identifies clusters in noisy data by isolating groups of points which are mutually close to one another. On a high level, the algorithm works as follows: First, the algorithm defines a mutual reachability distance between vectors which is larger for either pairs of far away vectors or

for vectors in non-dense areas. Based on this mutual reachability distance, it then constructs a minimum-distance spanning tree of the dataset to identify the smallest distances needed to connect vectors to find points which naturally group together. Next, based on a pre-determined “minimum cluster size,” HDBSCAN traverses down the tree and sequentially splits the data into smaller and smaller clusters until the minimum size is reached. The algorithm then determines which of the resulting split clusters are robust enough to be extracted as stand-alone clusters, and which ought to be re-combined into larger clusters. Points which are too far away from a robust cluster to be assigned cluster membership in this step are determined to be outliers.² We employed a specific form of HDBSCAN called soft clustering which, in addition to producing clusters, assigns a vector of probabilities to every point where each component denotes the probability that the point belongs to the component’s corresponding cluster. We used this to deal with outlier points, as outlined in the next section.

We preferred the non-parametric HDBSCAN method over parameteric methods like k-means clustering as it performs better in noisy settings with many outliers, which we determined the skills dataset exemplified. Additionally, we preferred HDBSCAN’s ability to organically identify clusters in the data without being forced to create an arbitrary number of clusters (as k-means does). Finally, the implementation of HDBSCAN is particularly useful for hierarchical clustering, which is explored in the next section.

BERTopic allowed us to integrate HDBSCAN clustering with our trained embedding model to produce topics. In particular, we defined an HDBSCAN object with a minimum cluster size of 50, which we found worked best for producing sensibly sized clusters with the appropriate specificity. Then, for visualization purposes, we added a UMAP dimensionality-reduction layer which could represent the high-dimensional feature space in a two-dimensional plot. In total, these layers transformed the cleaned skills descriptions into vectors, perform HDBSCAN clustering to obtain clusters, and visualize the resulting clusters in a 2D-plot.

4.2.4 Taxonomy Generation

Generating the skills taxonomy consisted of two main steps. First, we needed to construct sensible names for the clustered skills to form our taxonomy’s lowest-level skill categories. This was achieved through Topic Modeling in BERTopic together with OpenAI word representations. In particular, as part of the clustering process, BERTopic produces a collection of words which define each cluster based on the common semantic meaning of the words contained in the cluster. Then, by feeding these topic words into ChatGPT 3.5, we were able to obtain interpretable names for each skill cluster obtained from HDBSCAN. Second, we needed to organize these base-level skill categories into a hierarchical organization of higher-level skill groupings, which is where the Hierarchical component of HDBSCAN was useful. In particular, HDBSCAN is able to automatically generate this hierarchy by virtue of the clustering method: the minimum spanning tree HDBSCAN generates already connects the entire dataset, so by reading off the spanning tree we can recover

²For more information see [here](#).

which clusters are closest to each other and iteratively build the hierarchy. BERTopic supports hierarchical topic modeling for these hierarchical clusters, which allowed us to build a complete taxonomy with human-interpretable terms for the resultant higher-level clusters.

4.3 Results

4.3.1 Skill Clusters

The HDBSCAN clustering model organizes the entire sample of skills into 50 distinct lowest-level clusters based on topic similarity. Figure 2 illustrates the colored two-dimensional embedding of these clusters following dimensionality reduction from the initial high-dimensional embedding. As detailed in the methodology section, two versions of labels are provided for each cluster: the preliminary output generated by BERTopic’s default representation model, and the final labels refined by GPT. During the implementation phase, users have the capability to zoom in and interactively explore the clusters, allowing them to view the specific skills associated with each cluster.

In Figure 2, each data point is distinctly allocated to a cluster without the presence of duplicates or outliers. These observations suggest the next steps. Firstly, it is notable that HDBSCAN does not accommodate the assignment of a skill to multiple clusters. This restriction poses limitations to our model as certain skills may have attributes that warrant classification into more than one cluster. Given the inherent transferability of job skills across domains, the imposition of a one-to-one mapping constraint may not always align with the nature of the data. Secondly, the outcome of HDBSCAN typically involves the outliers. In the context of our study, outliers signify skills that defy conventional categorization within established fields. However, our observation revealed that many of these outliers are, in fact, the skills that intuitively belong within certain clusters but are erroneously classified as outliers. Consequently, we opted to employ the `reduce_outliers` option within the BERTopic framework, which mitigates the presence of outliers by eliminating probability thresholds and enforcing a unique assignment of each skill to a cluster without exceptions. This decision was based on our observation that fine-tuning probability thresholds failed to discern true outliers from false ones, leading to another proliferation of inaccurate outliers. Hence, we suggest further research aimed at devising strategies to effectively reduce the number of outliers without inadvertently generating false negatives.

4.3.2 Skill Hierarchy and Taxonomy

The resulting hierarchical taxonomy from running HDBSCAN on the clusters is shown in Figure 7 of the Appendix. Since the resultant hierarchy was highly complex and uneven with a large number of intermediate layers, we manually reorganized the hierarchy into a 4-layer categorization, depicted in Figure 8. The reorganization maintains the hierarchical relationships established by our clustering while flattening and assigning more interpretable names to the levels. The first result of note is the highest-level division between Digital and Non-Digital skills, indicating that

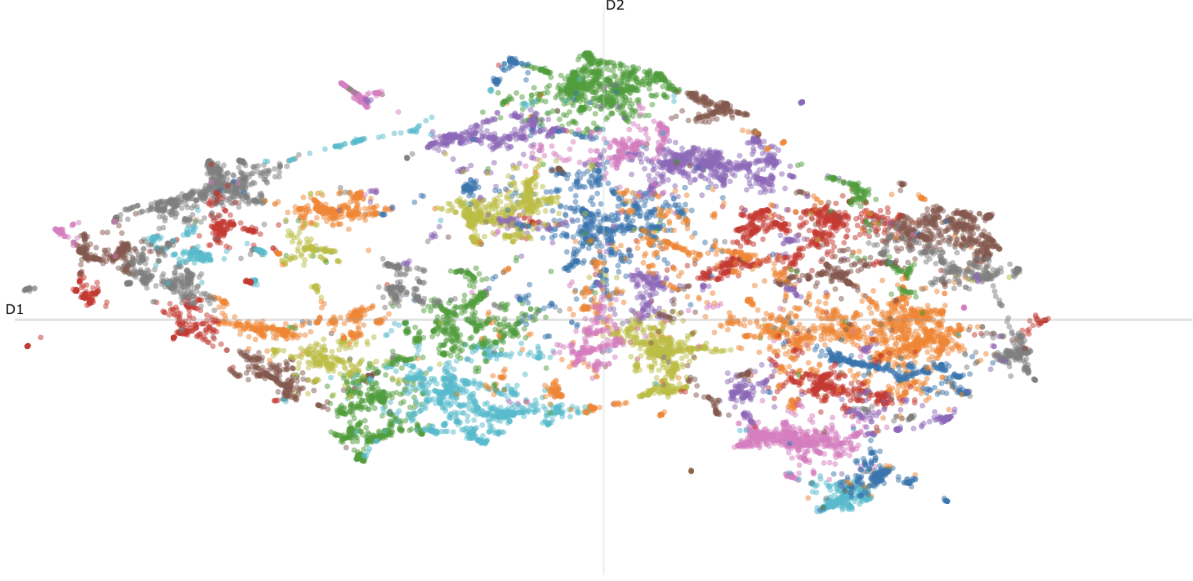


Figure 2: Cluster Visualization Sample

digital skills represent a highly divergent category of skills which cannot be closely approximated by any non-digital skill group. This would seem to indicate that digital skills are the hardest skill groups to adopt for non-digital laborers. Next, it’s noteworthy that our taxonomy recovers a distinction between human-oriented and technical skills, which concords with common understandings of “STEM” skills as being different from “Humanities” skillsets. Finally, the fact that Data Science skills, which also encompasses statistics and its mathematical applications, are categorized substantially differently from other Digital skills indicates the recent emergence of Data Science as a unique skillset from other software roles.

4.4 Evaluation

On a high level, our method succeeds in producing a sensible clustering of skills into a skills taxonomy. The clustering seen in Figure 2 produces many well-formed clusters which group together similar skills, and the higher-level clusters such as “Mechanical,” “Language” and “Data Science” skills seem to accord with intuitive groupings of skills. This is encouraging for our overall approach, since it suggests that there are substantial trends in the skills data which an unsupervised and non-parametric method can successfully model, indicating that this methodology can serve as the basis for future research. Furthermore, the purely data-driven approach means it may be more flexibly extended to new datasets. This can be used to both extend the taxonomy through historical data, to see how the skills categories have changed historically, and to future skills data, to see how new skills may integrate into the taxonomy.

Nevertheless, our approach has some drawbacks that future researchers will have to address. In particular, the large number of outliers in the initial clustering step demands more thorough investigation, since they might indicate skills that are either caught between skill categories or skills

that are highly divergent and belong to no category. Our current methodology is insufficient at identifying and differentiating between these cases, which may be a task for future research.

Our final evaluation task was to compare the taxonomy we generated to the manual skills taxonomy provided by our project partners at the World Bank. To get a basic sense for the similarity between our taxonomies, we compared the lowest-level clusters in our data to the 33 skills categories provided to us by the World Bank. The main metric we employed was the mutual information score, which for two distinct clusterings $U = \{U_1, \dots, U_n\}$ and $V = \{V_1, \dots, V_m\}$ of the same N datapoints is defined as

$$MI(U, V) = \sum_{i=1}^n \sum_{j=1}^m \frac{|U_i \cap V_j|}{N} \log \left(\frac{N|U_i \cap V_j|}{|U_i||V_j|} \right)$$

This metric essentially measures the extent to which the clusterings U and V may be modeled as arising from independent random variables, where if the clusterings are completely independent, $MI(U, V)$ will be near 0. However, when there are a large number of clusters (as there are in this case), this metric reports clusterings as more similar than they actually are due to random chance alignments. Therefore, we consider the *adjusted* mutual information score

$$AMI(U, V) = \frac{MI(U, V)}{\frac{H(U) + H(V)}{2} - \mathbf{E}[MI(U, V)]}$$

where $H(U)$ denotes the entropy (i.e. total information content) of U .

For the comparison between our taxonomy and the World Bank taxonomy, the AMI is 0.00225, indicating that the taxonomies are very different at the low level. There may be a few reasons for this discrepancy. For one, we have a larger number of clusters at this level, meaning that there is bound to be misalignment between the clusterings. However, examining the clusterings more closely reveals some more fundamental differences in the taxonomies. The World Bank categories aim for generality and breadth of coverage, with the broad categories having little overlap and aiming to cover as many different conceivable skills as possible (eg. categories such as “Business” and “Education and Training”). However, our taxonomy emerge from the skills that are represented in the dataset, leading it to focus on finer distinctions between highly-represented skills and ignore skill categories with little representation. For example, the World Bank categorization has one category to represent “Information Technology” whereas our taxonomy has over 10, which emerges because software skills were highly represented in our dataset. Conversely, our taxonomy either subsumes or omits entirely several of the World Bank categories such as “Agriculture, Horticulture, and Landscaping,” or “Property and Real Estate” for which there were few skills in our skills dataset.

This may be seen both as an advantage and a disadvantage for our approach, depending on the use-case. As an advantage, we can view our approach as grounding the taxonomy in currently relevant skills, which carries information that may be both more pertinent to researchers who want to understand the current labor market and which may carry more time-specific information that can make the model valuable in studying time-varying skills dynamics. As a disadvantage, this may

indicate that our approach is highly sensitive to sampling bias in the provided list of skills, meaning extra attention needs to be paid to the data used to build any unsupervised skills taxonomy.

5 A Skills-Based Approach to Occupational Similarity

5.1 Literature

Building a measure of occupational similarity based on the underlying skills in each occupation allows us to better understand what constitutes an occupation, how occupations change over time, and how workers might transition between occupations. Previous research on classifying occupations based on skills has overwhelmingly used revealed comparative advantage (Alabdulkareem et al., 2018; Dawson et al., 2021; José-García et al., 2023). Revealed comparative advantage has also been used in several other contexts such as identifying the key exports of nations (Hidalgo et al., 2007) and the key industries in cities (Glaeser et al., 1992). In the context of occupational similarity, revealed comparative advantage allows us to uniquely identify the key skills for a given occupation, thus allowing us to distinguish occupations by these core skills.

5.2 RCA

Our first skills-based analytic approach is to apply the RCA method used by Alabdulkareem et al. (2018) and others to our data.

5.2.1 Methodology

We follow Alabdulkareem et al. (2018), Dawson et al. (2021), and José-García et al. (2023) in using revealed comparative advantage to measure the relevance of a skill s in occupation j , computed as:

$$\text{RCA}(j, s) = \frac{x(j, s) / \sum_{s' \in S} x(j, s')}{\sum_{j' \in J} x(j, s') / \sum_{s' \in S, j' \in J} x(j, s')}$$

where $x(j, s) = 1$ if skills s is required for occupation j and is 0 otherwise, S is the set of all distinct skills, and J is the set of occupations in our data set.

An RCA value greater than one indicates that skill s is more important to occupation j than average. Following Alabdulkareem et al. (2018) and Dawson et al. (2021), we normalize our RCA values using "effective use of skills," where $e(j, s) = 1$ if the RCA value is greater than one and $e(j, s) = 0$ otherwise. Although we explored other normalization methods, we ultimately found that the effective use of skills provided the most useful measures of occupational similarity (i.e., occupational similarity scores in similar orders of magnitude).

For each country and year, we calculate the effective use of skills for each occupation and skill combination. We use these values to create a vector of skills to represent each occupation, where each element in the vector represents a particular skill and contains the corresponding effective



Figure 3: Note: For clarity, only occupations with cosines scores greater than 0.3 are connected.

use of skills value. We use these vectors to calculate our measures of occupational similarity using cosine similarity, which we found to perform better than the other similarity measures we tested (e.g. Jaccard).

5.2.2 Results

Using the vectors of effective skills and the cosine similarity scores allows us to visualize changes and continuities in occupational similarity in a single country over time (see Figure 9), while also allowing us to observe differences between countries (see Figure 10). Note that in both figures, some of the increases in connections can be explained by increases in the amount of data available, which increases from the left-most figure to the right-most figure.

These visualization methods also allow us to track changes in occupational similarity as well as demand for occupations. Figure 3 displays two network maps showing the relative distances between all occupations in Chile in 2020 and 2023. While many occupations that were close in 2020 remained close in 2023, we see that economists (denoted by the red dots) and policy and planning managers (denoted by the blue dots) grew relatively far apart between 2020 and 2023. This corresponds to a 0.07 decrease in cosine similarity. Figure 4 shows occupations scaled by occupational demand, as measured by the number of job postings for a given occupation. Fewer occupations in 2023 made up more than 2.5% of the listed occupations, which are denoted by the red dots. However, in general, occupational demand remained fairly similar between 2022 and 2023 for Argentina.

5.3 OpenAI Embeddings

We also used an empirical approach for analyzing occupational similarity through OpenAI’s API which gave us access to their ‘text-embedding-3-small’ vector embedding model. We made use of this model on job descriptions in postings, as we assume that professions which are similar also have postings that are semantically similar.

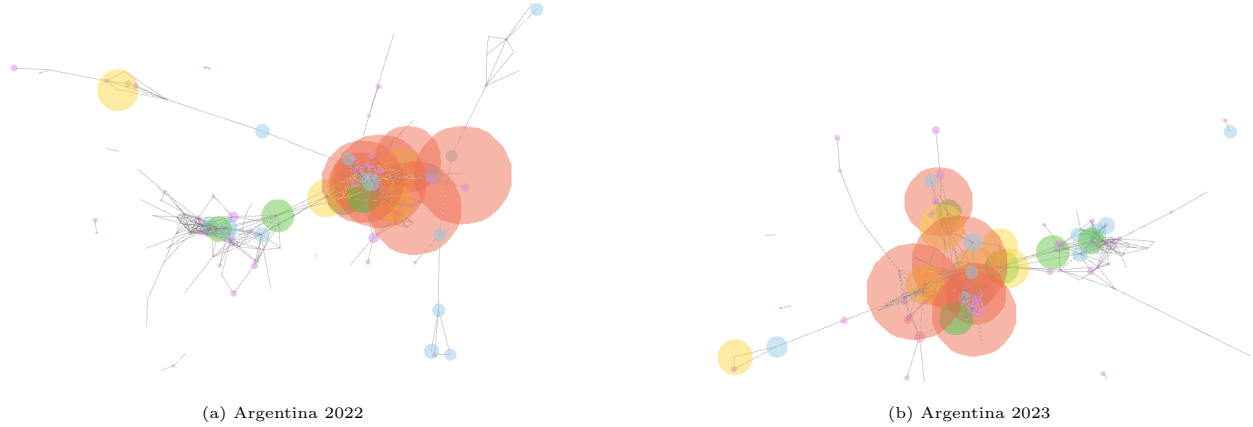


Figure 4: Occupational Similarity by Demand

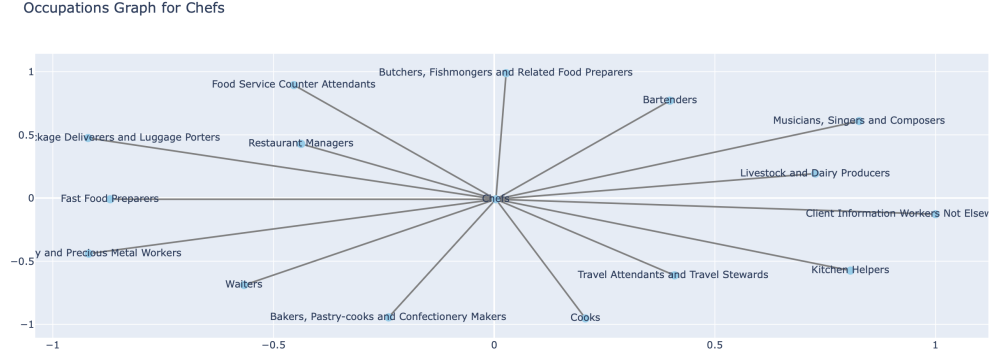
5.3.1 Methodology

We translated a representative sample of all the postings we had from Spanish into English. This sample contained 500,000 postings from Argentina, Uruguay and Chile from 2020-2023. Approximately 90,000 of these postings did not have an ISCO occupation classification so we dropped them for this task. From the remaining subset we chose roughly half of the postings at random and used OpenAI’s ‘text-embedding-3-small’ model to generate vector embeddings for each of the chosen postings’ job descriptions. We then aggregated these postings according to their ISCO occupation name to obtain an ‘average’ embedding score for each occupation’s postings’ descriptions. Using cosine similarity on these aggregated vectors allowed us to compare how similar one occupation’s average posting description is to another which should provide us with an indication of how similar two occupations are.

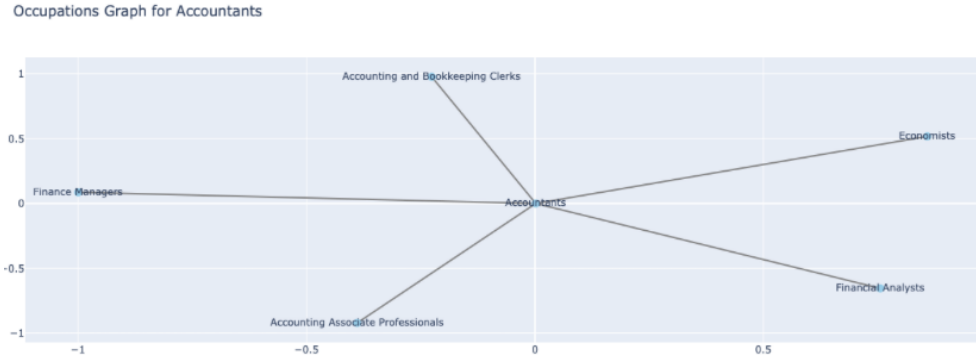
5.3.2 Results

As this methodology also makes use of vectors and cosine similarity we can produce similar plots to those obtained from the RCA method, including plots which show how occupational similarities vary over countries and times. Taking a look at some of the results may show where the strengths and weaknesses of this approach lie. For example, Figure 5b shows a network graph with the 5 occupations which most similar to ‘Accountants’. All 5 of these occupations seem intuitively similar to ‘Accountants’ as we get results like ‘Accounting Clerks’ and ‘Finance Managers’. Similarly, Figure 5a shows the 15 closest occupations to ‘Chef’ over the whole aggregated sample. A lot of these results are intuitively similar to ‘Chef’, as we have occupations like ‘Butchers’, ‘Cooks’ and ‘Kitchen Helpers’. However some other occupations like ‘Musicians’ and ‘Precious Metal Workers’ seem unrelated to ‘Chef’. This could show that the OpenAI classification is more useful when searching for a smaller number of similar occupations (all the intuitively similar occupations were in the top 5 closest occupations to ‘Chef’).

Furthermore, the classifications seem to be more effective for certain industries than others.



(a) 15 closest occupations to Chef



(b) Network graph between every occupation and its 5 closest occupations

For example, finance and STEM occupations, including medical ones, usually have other similar occupations that intuitively ‘make sense’. For other service sector occupations, like ‘Librarians’ (top 5 similar occupations include ‘Actors’ and ‘Executive secretaries’) and ‘Painters’ (top 5 similar occupations include ‘Actors’ and ‘Executive secretaries’) the results do not seem very useful. So, the OpenAI embeddings method may work particularly well for only a few industries.

5.4 Evaluation

We produced two distinct methodologies to produce occupational similarity scores (that I henceforth call scoring systems), so it was necessary to evaluate their performance. We have no absolute “truth”, beyond general human intuition, for occupational similarity. This makes an empirical evaluation of the performance of an individual scoring system infeasible. From a qualitative view, we assessed both systems by examining the 5 most similar occupations to a selected group of professions using similarity scores produced by each systems. The OpenAI scores appeared to produce somewhat more sensible results than the RCA scores in certain cases, in that the closest occupations it recommended seemed to co-incide with our general understanding of the industry. However, the difference in performance along this metric was not appreciably large.

A more interesting question, then, was to study the level of agreement between the two systems. That is, given an occupation, did both systems agree on what they considered to be the closest

N occupations to the given one? This is important in the larger context of this project because if our systems diverge significantly in providing closest occupations, then we do not have a clear understanding of which one to use in the process of generating a LMIS recommendation system.

What do we define as strong agreement? Given an occupation O , and a space of other occupations S , we make both systems rank the occupations in S in descending order of similarity score to O . Then, if the top N ranked occupations generated by RCA and OpenAI show a strong overlap, we have grounds to believe that the two systems exhibit an agreement on what they consider to be the closest relatives of O . The relative ranks of the top N professions also carry value, but only if there is a high degree of agreement as otherwise we do not have a like comparison. To quantify agreement, we use the Jaccard similarity coefficient, which is given by

$$\frac{|A \cap B|}{|A \cup B|}$$

where A and B are the sets of top N occupations generated by the RCA and OpenAI methods.

In implementation, we compared similarity matrices from both systems separately for each country-year combination, to ensure like comparisons for specific periods and geographies. We reduced the similarity matrices of both systems to the space of common occupations for which data was present in both matrices, to ensure that we had the same space S for each occupation O between both methods. To calculate Jaccard scores, it was important to pick a suitable N . We consider $N = 10$ to be a reasonable pick, since one cannot imagine a worker caring about more than the 10th closest occupation to their current one. However, we also ran the Jaccard scores with $N = 15$ to gain a broader comparison for theoretical purposes.

Then, for each country-year combination, we generated a Jaccard score for each common occupation. We present the histogram of Jaccard scores for each combination below for $N = 10$.

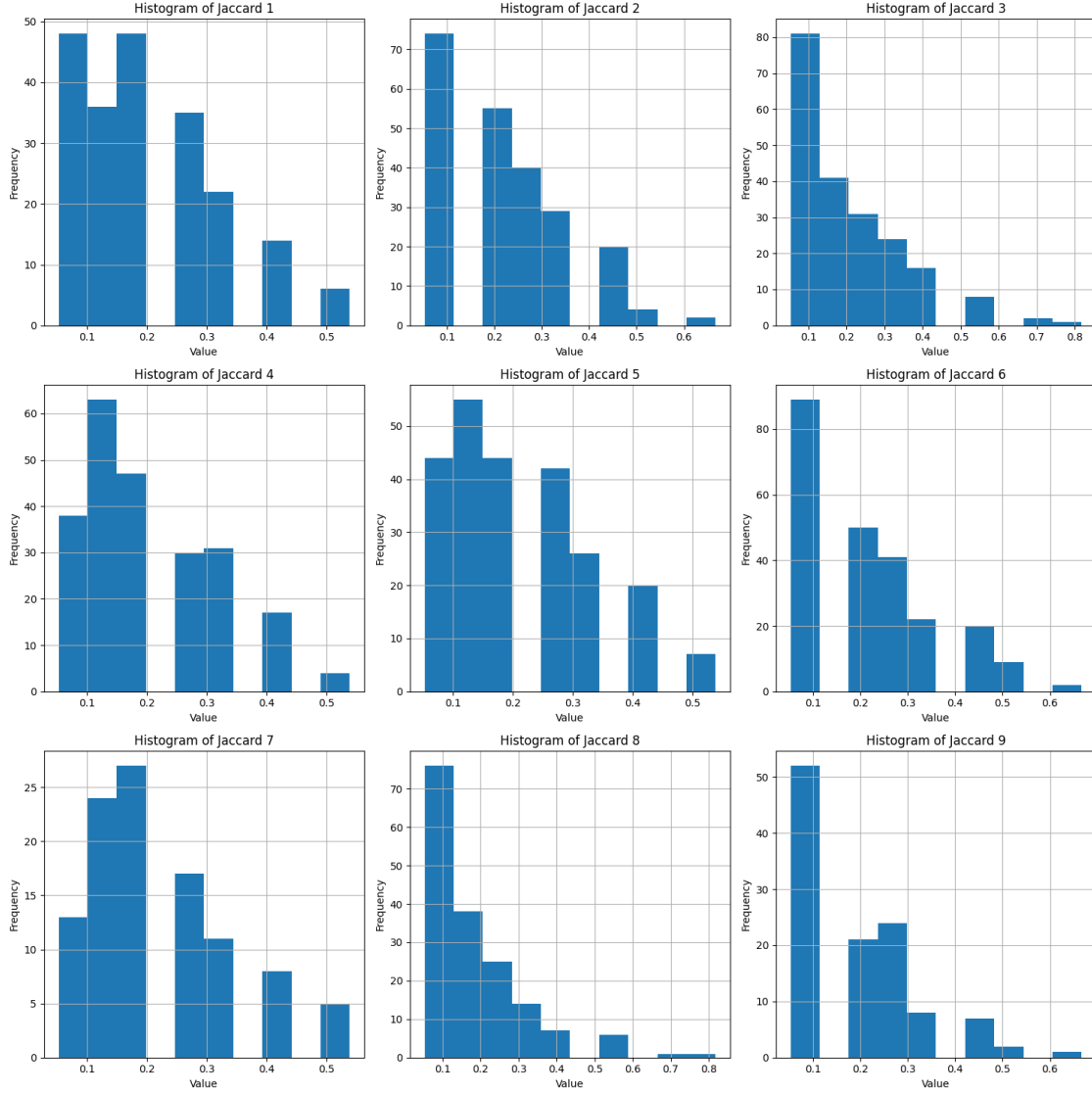


Figure 6: Histograms of Jaccard scores for all occupations in each country-year combination ($N = 10$). The rows in the grid represent Argentina, Chile and Uruguay, and the columns in the grid represent the years 2020, 2021, 2022, in that order.

The vast majority of occupations exhibited Jaccard scores less than 0.3, which means that the number of occupations the systems had in common within their top 10 ranks was less than 30% of the total number across both lists. In fact, the mode seems to be around 0.1 for most combinations, which is very low. This indicates that agreement between the systems is generally low. As a robustness check, we drew the same histograms for $N = 15$ (see Appendix), and the outcome remained the same - an overwhelming concentration of occupations had scores below 0.3.

A low level of agreement between the two systems is surprising, since on a larger level it indicates a divergence between the theoretical and empirical approaches we took to this problem. We do not have a clear reason as of now as to why this level of disagreement occurs, which would be an interesting area of future research. The disagreement forces us to confront our original problem of evaluating the objective accuracy of a system. If our two systems do not agree, then we cannot

resolve a conflict between their recommendations without having an objective basis to believe in one over the other. Therefore, further research is also required on which one of the two systems is more reflective of intuitive human understanding of occupational similarity.

On the bright side, a sizable number of occupations *did* exhibit reasonably high Jaccard scores - a score of 0.4 appeared to form a natural cutoff for the highest-scoring occupations. We then tabulated the occupations with Jaccard scores greater than 0.4 within each combination, and created an ordered list of occupations based on their frequency of appearance on the high-scoring list. A frequency of 4 was considered to be the minimum for an occupation to be deemed to have consistently high Jaccard scores across country-year combinations. Such occupations totaled 17 in number, less than 5% of all occupations mentioned in the data. These are presented in Table 1.

Occupation	Score
graphic and multimedia designers	9
accounting associate professionals	8
accountants	8
finance managers	8
sales and marketing managers	8
financial analysts	7
advertising and marketing professionals	6
systems administrators	6
computer network and systems technicians	6
public relations professionals	5
software developers	5
applications programmers	5
chefs	4
accounting and bookkeeping clerks	4
software and applications developers and analysts not elsewhere classified	4
web and multimedia developers	4
commercial sales representatives	4

Table 1: List of occupations with Jaccard scores greater than 0.4 in 4 or more country-year combinations

We observe that the occupations tend to largely belong to the finance, business administration and software/IT sectors. Our hypothesis for this is that these sectors tend to have niche, well-defined skills, which would mean that the semantic similarity found by the OpenAI model and the skill words picked up in the RCA would be similar to a greater degree than for other professions. For this subspace of occupations, it is possible to create a recommendation of closest occupations based on the intersection of the top-N ranked lists between the two systems, since we are guaranteed of a large intersection for around half the country-year combinations. In fact, a small test following this methodology for “graphic and multimedia designers” recommended occupations like advertising, marketing, public relations, theater and journalism, which are very sensible and therefore promising.

6 Conclusion and Further Research

Our study explores new, cutting-edge methods for labor market data analysis in the context of strengthening labor market information systems in Latin America.

First, we leverage advancements in topic modelling techniques and language models to produce a novel job skills taxonomy. Our taxonomy generation pipeline accounts for the context within which skills appear in job postings, by using a BERT-based transformer model that is continually pre-trained on several hundred thousand job postings from Latin America. Although the creation of our taxonomy required **some** manual intervention, the vast majority of the pipeline is designed to “let the data speak for itself”. Our model can be rerun at a relatively low computational cost in order to incorporate new skills into the taxonomy as they appear in the data.

Next, we compare two approaches for calculating similarity between (ISCO) occupations. Broadly speaking, the RCA method is computed using the **frequency of co-appearance** between occupations and **job skills**, while the embeddings-based method relies on **semantic similarity** between the underlying **posting text** from each occupation. Surprisingly we find that in many cases these methods produce very different results, and we see this as an opportunity for further research.

While our work extends previous research on the subject in multiple ways, there are several avenues for further development especially with regards to skills categorization. First, as detailed in Section 4.4, outliers in each cluster should be appropriately removed to enhance the practical utility of the output. Although domain experts can still perform the final cleaning (as done in previous studies), we hope that future research can develop a data-driven optimization for handling outliers to make the entire process fully scalable. Second, researchers may explore the use of different evaluation metrics to compare the model output against ground truths. In our case, as detailed in preceding sections, we utilized the Mutual Information Score for this purpose. Lastly, similar to our utilization of job postings from Argentina and Uruguay for the continual pre-training of DistilROBERTa aiming at enhancing skill clustering performance within these countries, researchers could consider employing diverse document sets as inputs for additional corpus to augment default transformer models as well as different combinations of the default NLP models.

References

- Alabdulkareem, A., M. R. Frank, L. Sun, B. AlShebli, C. Hidalgo, and I. Rahwan (2018). Unpacking the polarization of workplace skills. *Science advances* 4(7), eaao6030.
- Ao, Z., G. Horváth, C. Sheng, Y. Song, and Y. Sun (2023). Skill requirements in job advertisements: A comparison of skill-categorization methods based on wage regressions. *Information Processing Management* 60(2), 103185.
- Dawson, N., M.-A. Williams, and M.-A. Rizoiu (2021). Skill-driven recommendations for job transition pathways. *Plos one* 16(8), e0254722.
- Djumaieva, J., A. Lima, and C. Sleeman (2018). Classifying occupations according to their skill requirements in job advertisements. (ESCoE DP-2018-04).
- Gallagher, E., I. Kerle, C. Sleeman, and G. Richardson (2022). A new approach to building a skills taxonomy. (ESCoE-TR-16).
- Glaeser, E. L., H. D. Kallal, J. A. Scheinkman, and A. Shleifer (1992). Growth in cities. *Journal of political economy* 100(6), 1126–1152.
- Grootendorst, M. (2019). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.
- Gu, Y., R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon (2021, oct). Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare* 3(1).
- Hidalgo, C. A., B. Klinger, A.-L. Barabási, and R. Hausmann (2007). The product space conditions the development of nations. *Science* 317(5837), 482–487.
- Ji, S., T. Zhang, K. Yang, S. Ananiadou, E. Cambria, and J. Tiedemann (2023). Domain-specific continued pretraining of language models for capturing long context in mental health. *arXiv preprint arXiv:2304.10447*.
- José-García, A., A. Sneyd, A. Melro, A. Ollagnier, G. Tarling, H. Zhang, M. Stevenson, R. Everson, and R. Arthur (2023). C3-ioc: A career guidance system for assessing student skills using machine learning and network visualisation. *International Journal of Artificial Intelligence in Education* 33(4), 1092–1119.
- Shehu, M. and E. Gjika (2024). A comprehensive review of the three main topic modeling algorithms and challenges in albanian employability skills. *European Scientific Journal, ESJ* 20(12), 31.

7 Appendix

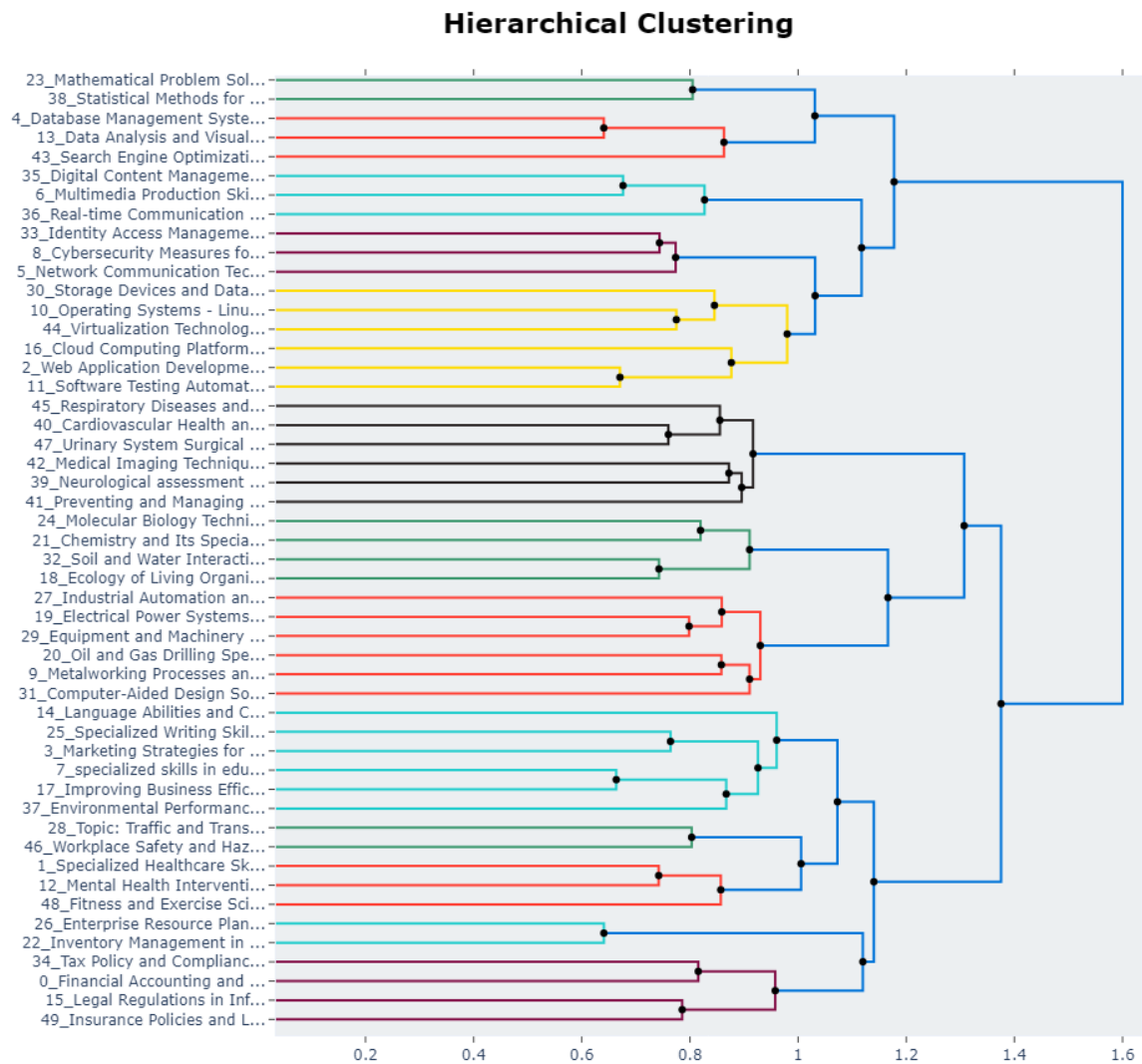


Figure 7: Hierarchical Clustering Result

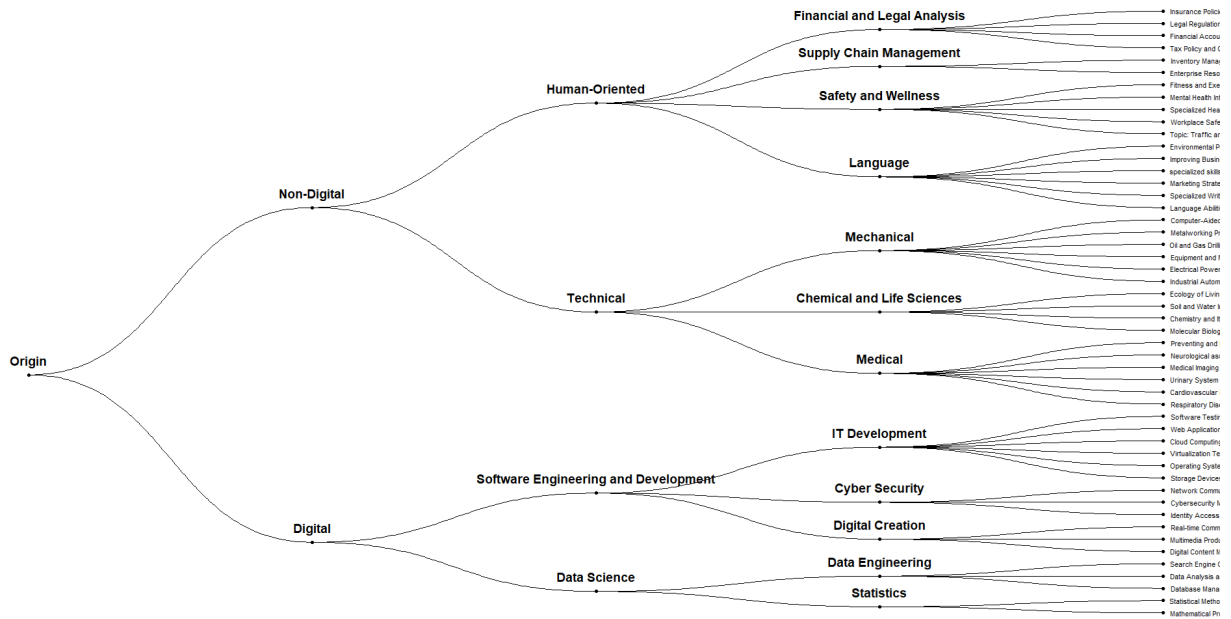


Figure 8: Simplified Taxonomy (Based on Hierarchical Clustering)

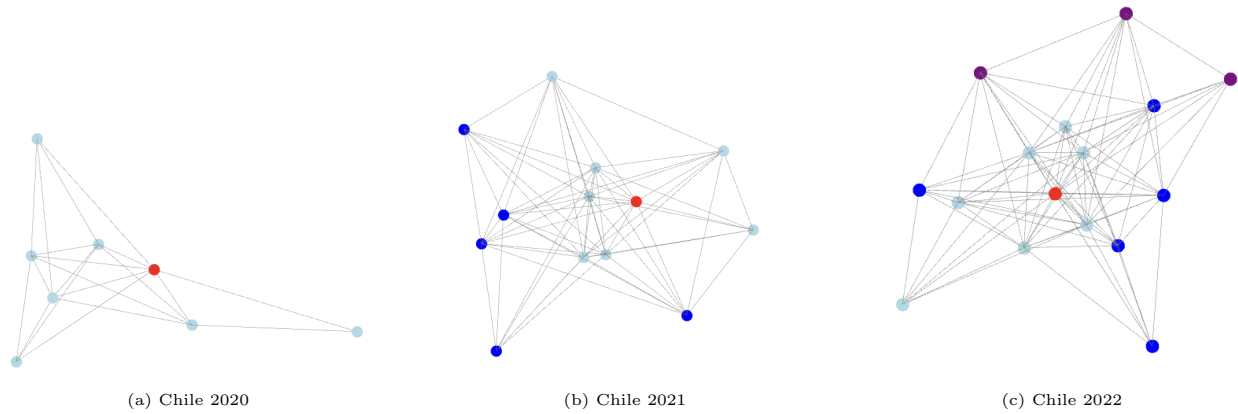


Figure 9: Occupational Similarity for Software Developers by Year

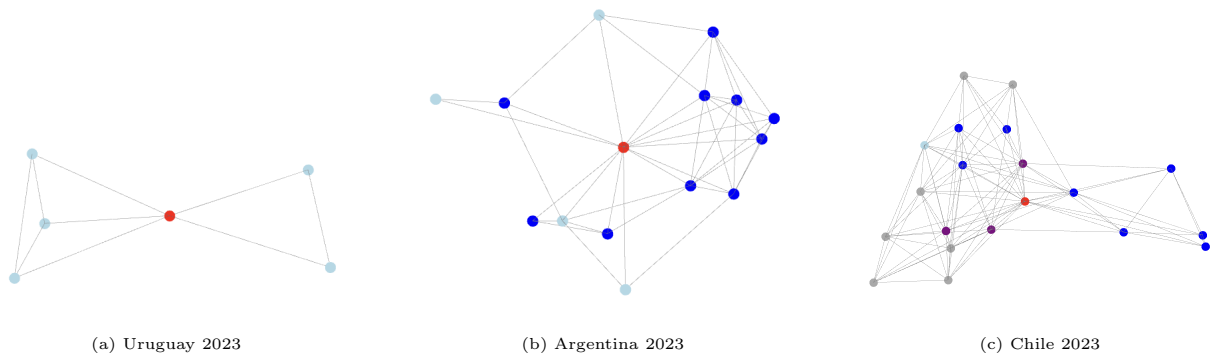


Figure 10: Occupational Similarity for Mechanical Engineers by Country (2023)

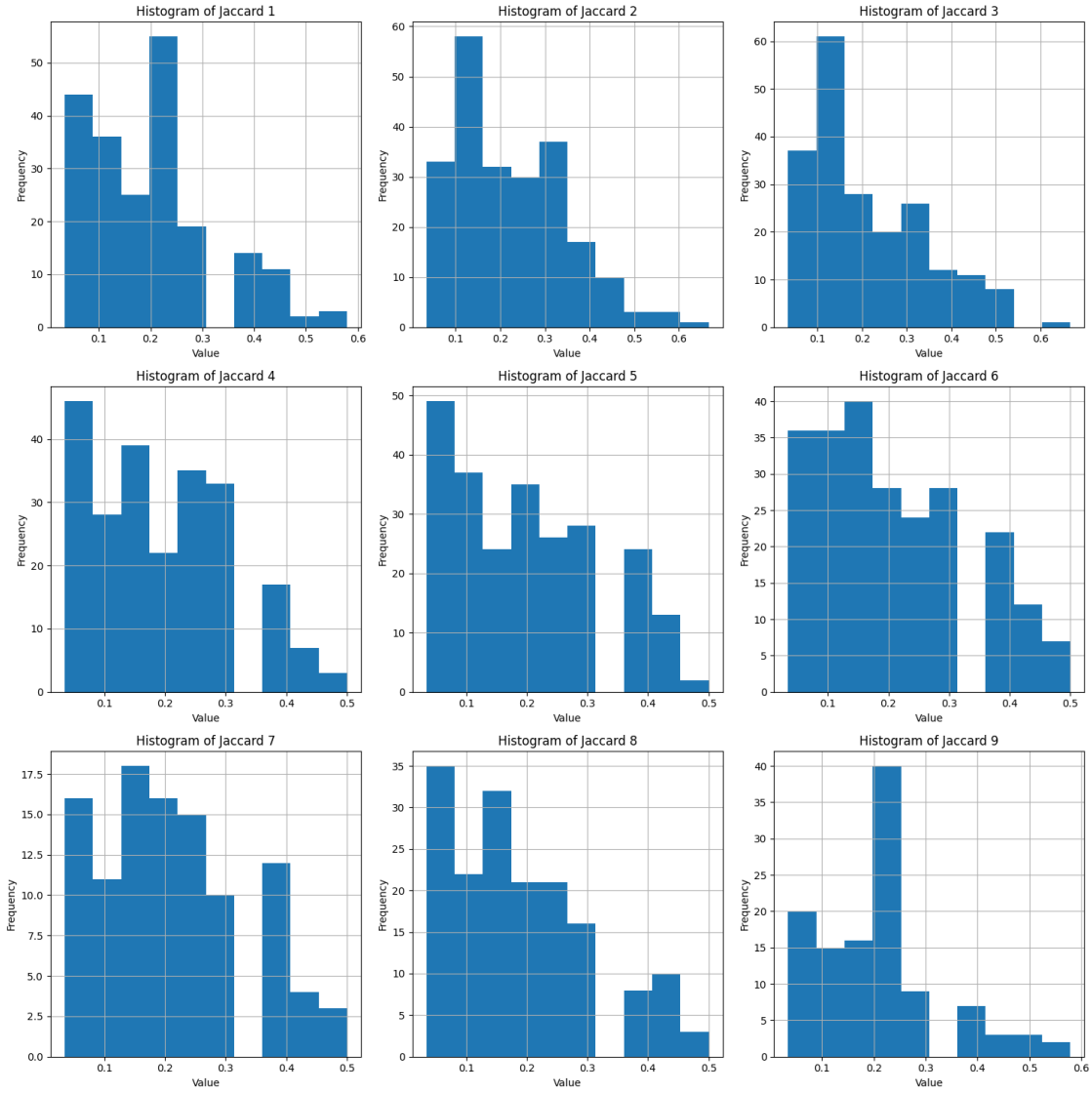


Figure 11: Histograms of Jaccard scores for all occupations in each country-year combination ($N = 15$). The rows in the grid represent Argentina, Chile and Uruguay, and the columns in the grid represent the years 2020, 2021, 2022, in that order.