



An Analysis of Argentina's Mining Sector

Project Lead: Koren Gila

Analysts: Gavin An, Adam Cheema, Ludovico Graber, Diego Williams *

Fall 2025

*All of the listed authors study at the University of Chicago. Special thanks to Harry Moroz at the World Bank for the support and patience throughout this project, and to Aleksandr Popov and Darshan Kommanapalli for the opportunity with CREL.

Contents

1	CREL and Project Overview	4
1.1	Charles River Economics Labs	4
1.2	Partner Organization	4
1.3	Project Background	4
1.4	Motivation	5
2	Data and Methodology	5
2.1	Datasets	5
2.1.1	Job Postings Database	6
2.1.2	Skills Database	6
2.2	Datasets Qualitative Findings	6
2.3	Random Forest	7
2.3.1	Random Forest Theory	9
2.3.2	Application	11
2.4	Country Grading	12
2.4.1	Completeness	13
2.4.2	Consistency	13
2.4.3	Quantity	14
2.4.4	Composite Ranking	15
2.5	Labor and Skills Demand Analysis	16
2.5.1	Data Refinement and Outlier Exclusion	16
2.5.2	Geographic and Corporate Market Share	17
2.5.3	Relative Demand and Disproportionality	17
2.5.4	Longitudinal Growth Measurement	18
3	Case Study: Quantifying Job Demand Growth in Mexico	18
3.1	Data Collection and Cleaning	19
3.2	Occupational Taxonomy and Classification	19
3.3	Measurement of Job Growth	20
3.4	Observed Growth Patterns in Mexico	20
3.5	Limitations	21
4	Results	21
4.1	Random Forest Results	21
4.2	Country Grading	23
4.3	Labor and Skills Demand in the Argentinian Mining Sector	26
4.3.1	Aggregate Vacancy Trends	26
4.3.2	Geographic Distribution and Evolution	27
4.3.3	Corporate Landscape	28

4.3.4	Occupational and Skills Demand	30
4.3.5	Longitudinal Growth and Digitization (2023–2024)	31
5	Conclusion	32
5.1	Conclusion	32
5.2	Limitations	33

1 CREL 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, World Bank, the US Department of State, and the Census Bureau, among others.

1.2 Partner Organization

The World Bank is an international development organization that aims to reduce poverty and improve living standards. In addition to lending money to the governments of its poorer members, it conducts research in development.

1.3 Project Background

Argentina’s mining sector has experienced a sharp expansion in recent years, driven by large-scale investments in lithium, copper, gold, and silver projects. This growth has been concentrated primarily in the northwest (NOA) and Cuyo regions and is expected to continue as several projects transition from feasibility to construction and production. Despite the scale of announced investment, mining companies report that the availability of skilled labor is a binding constraint on project execution and long-term sector growth.

Between 2023 and 2025, the World Bank Group (WBG), through joint work by the International Finance Corporation (IFC) and the International Bank for Reconstruction and Development (IBRD), conducted surveys and workshops with mining companies in collaboration with CAEM. These consultations revealed a persistent mismatch between labor demand in the mining sector and the existing supply of trained workers. Firms consistently cited shortages in technical and professional occupations, limited alignment between educational curricula and operational needs, and difficulties attracting and retaining workers in remote mining regions.

In response, the WBG is supporting the creation of a public-private mining talent platform.

The platform is designed as a coordination and financing mechanism rather than a training provider. Its role is to align mining companies, governments, and selected educational institutions around a common strategy for skills development. In practice, the platform aims to:

1. Identify occupational and skill demand under different mining investment scenarios
2. Support the redesign and expansion of education and training programs to better match industry needs
3. Reduce regulatory and institutional frictions that limit the scale and speed of training delivery

1.4 Motivation

This project contributes to the broader World Bank initiative aimed at addressing skill shortages in Argentina’s mining sector. Mining companies have identified labor constraints as a binding limitation on project execution, yet the composition of labor demand in the sector—and how it evolves across regions and over time—is not well documented. While firm-level surveys provide qualitative evidence of skill gaps, there is limited empirical work that systematically measures demand using labor market data.

Our objective is to use job postings and skills data from across Latin America to quantify labor demand within Argentina’s mining sector. In doing so, we confront several methodological challenges, including limitations in the coverage and quality of job postings data and ambiguity in defining the mining sector in a consistent and operational manner. We address these challenges by first investigating the data more broadly. Using the knowledge from those investigations, we combine text-based classification methods with a random forest model to identify the mining sector and the corresponding occupations and skills in demand.

2 Data and Methodology

2.1 Datasets

The majority of data analysis utilized two datasets provided by the World Bank across all countries in Latin America. The data sets center on job postings posted in each country, starting from January of 2020.

2.1.1 Job Postings Database

The World Bank provided a dataset of 39,738,185 jobs listings posted between 2020 and 2025 across Latin America, of which 4,918,460 were from Argentina. Each listing consists of information including a unique identifier, the title, the company name, the salary, ESCO (European Skills, Competences, Qualifications and Occupations) data, the modality of the job (e.g. in-person, remote), and location.

2.1.2 Skills Database

The World Bank also provided a dataset of 504,312,398 skills that were in demand over the same time period of 2020 to 2025 across Latin America, of which 42,269,485 were from Argentina. Each skill consists of a unique skill ID, the nation in which the skill was desired for, the name of the skill (e.g. Project Management, Marketing), the type of skill (e.g. Specialized Skill, Certification), a skill category name and ID (e.g. Business), a subcategory name and ID, and whether it was a software related skill. These two databases are related in that each entry in the database is information on a skill that a particular job listing in the same country and time period recommends candidates have. As a result, skills in the dataset are often duplicates.

2.2 Datasets Qualitative Findings

Before analyzing the Argentinian mining sector, we first ran some descriptive statistics on the dataset itself. We found that the skills dataset was highly concentrated in certain countries, with the five countries most represented in the skills dataset (Mexico, Brazil, Colombia, Argentina, and Peru) making up 71% of total entries. This was even more pronounced in the jobs listings dataset, with Mexico, Brazil, Chile, Peru, and Colombia making up roughly 90% of total entries. We noticed dataset size varied across time as well, with around 11.9% million jobs listings in the year 2023 compared to 1.4% million in 2020 and 4% million in 2025.

When assessing data quality, we used three metrics: completeness (for each row, how many columns contain values?), consistency (how does completeness vary with time?), and quantity (how

much data is there?). In particular, we let

$$Q_i = \sum_{j \in \text{Months}} \omega_{ij} q_{ij}$$

be completeness for country i , where

$$\omega_{ij} = \frac{t_{ij}}{\sum_{l \in \text{Months}} t_{il}}$$

and

$$q_{ij} = \sum_{k \in \text{Months}} \left(\frac{m_{ijk}}{t_{ij}} \right),$$

with t_{ij} being the number of postings in country i in month j and m_{ijk} is the number of postings with missing entries in column k for country i in month j . We let consistency for country i be

$$C_i = \frac{Q_i}{1 + \sigma(q_{ij})},$$

and we let quantity of country i be

$$A_i = \frac{1}{\# \text{ of months}} \sum_{j \in \text{months}} a_{ij}$$

where $a_{ij} \in \{1, \dots, 27\}$ denotes the ranking of country i 's postings in month j . We found that Argentina's data was below average in relation to its peers, with a completeness score of about 68% and a consistency score of 66%, while ranking 5th in quantity. Averaging the rankings across the three metrics, we found it ranked 19th out of the 27 countries in the dataset for quality. Specifically, listings location, occupation, employer, and sector data was significantly underpopulated compared to other countries, while salary information was almost entirely missing.

2.3 Random Forest

A central objective of this research is to accurately capture labor and skills demand within the mining sector of Argentina. One question of interest is how we should define the mining sector. Following ([Abowd](#)), we argue that the economic classification of the worker is tied to their

employer’s sector. Thus, we define jobs in the mining sector as all jobs under an employer in the mining sector. We define an employer in the mining sector broadly as a firm which contributes to the extraction or processing of natural minerals.

One benefit of this definition of the mining sector is the presence of the “Employer” column for each job posting. We then proceed by extracting mining sector job postings by matching them to firms via this column. We construct an initial set of Argentinian mining sector firms by combining online sources such as ([Minería](#)) with LLMs which both checked the sources for accuracy and found additional firms in the mining sector.

However, we run into roadblocks when attempting to extract a set of job postings to define the mining sector. Specifically

1. The Employer field is particularly badly populated for Argentina. Specifically, only 65% of postings in Argentina have a non-missing Employer field.
2. The Employer field is not standardized. That is, because the data was scraped from online job-postings, there is no set format or spelling which the company names come in.
3. Our initial set of mining sector firms is likely not exhaustive.

To combat the non-standardization of employer names, we use text-similarity methods to match our list of mining companies to names in the Employer column. Specifically, a match consists of two strings which have a sufficiently low Levenshtein distance. To understand Levenshtein distance, suppose we express two strings of length n and m as an ordered sequence $x = (x_1, x_2, \dots, x_n)$ and $y = (y_1, y_2, \dots, y_m)$, where each element x_i is the i th letter of x . Let $|x|$ denote the length of the word. Furthermore, suppose $\text{head}(x) = x_1$, the first letter, and $\text{tail}(x) = (x_2, x_3, \dots, x_n)$, where we form a new word by removing the first letter. The Levenshtein distance between two strings x and y is given by:

$$\text{lev}(x, y) = \begin{cases} |x| & \text{if } |y| = 0 \\ |y| & \text{if } |x| = 0 \\ \text{lev}(\text{tail}(a), \text{tail}(b)) & \text{if } \text{head}(a) = \text{head}(b) \\ \text{lev}_2(x, y) + 1 & \text{otherwise} \end{cases} \quad \text{lev}_2(x, y) = \min \begin{cases} \text{lev}(\text{tail}(a), b) \\ \text{lev}(a, \text{tail}(b)) \\ \text{lev}(\text{tail}(a), \text{tail}(b)) \end{cases}$$

In non-mathematical terms, the Levenshtein distance is the number of letter substitutions, additions, or deletions that we need to transform string x into string y . For example, we can get from “fog” to “dog” by substituting “f” for “d,” and hence $\text{lev}(\text{“fog”}, \text{“dog”}) = 1$. When computing similarity between pairs within a set of strings, we need to normalize Levenshtein scores along the length of the strings so that the scores can be meaningfully compared. The similarity score between strings x and y is defined as the ratio:

$$s(x, y) = 1 - \frac{\text{lev}(x, y)}{|x| + |y|}$$

We define a match as any Employer cell which has a similarity score of greater than 90%, which is a conservative threshold.

To combat the incompleteness of the Employer field in Argentina and the non-exhaustive nature of our list of mining firms, we make use of random forest methods to capture additional job postings in the mining sector which we miss on our initial matching of mining firms to the employer field.

2.3.1 Random Forest Theory

We choose a random forest architecture for its robustness against overfitting and its ability to handle the high-dimensional sparsity inherent in natural language data. Formally, a random forest is an ensemble estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

Suppose we have a training set $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$ where $x_i \in \mathbb{R}^p$ represents the vector of features (derived from tokenized job descriptions and skill embeddings) and $y_i \in \{0, 1\}$ represents the class label, where 1 denotes a mining-sector vacancy. We generate B bootstrap samples $Z_1, \dots, Z_B$ from the original data D . For each sample Z_b , we construct a decision tree T_b .

Unlike standard decision trees, which search the entire feature space for the optimal split, the random forest algorithm introduces randomness by selecting a random subset of features $m < p$ at each candidate split. This de-correlation of trees is critical for our application; without it, strong predictors (such as the word “mining” itself) would dominate the upper nodes of every tree, preventing the model from learning subtler, skill-based patterns required to identify unlabeled

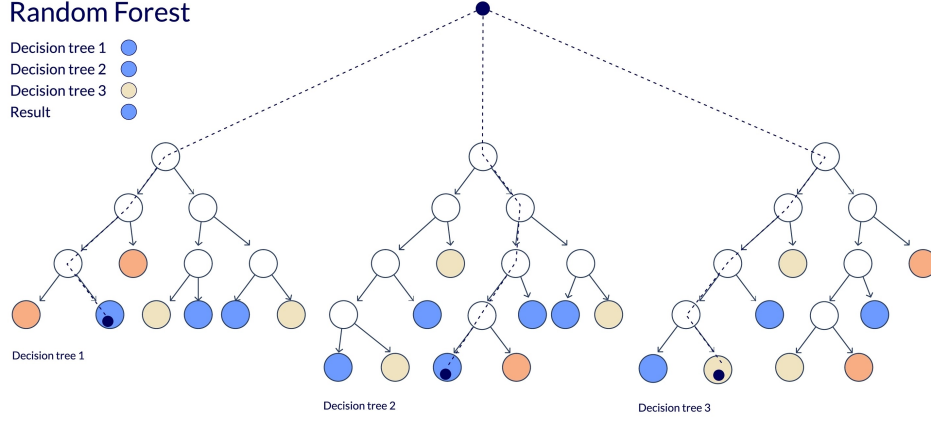


Figure 1: Architecture of the Random Forest Ensemble

vacancies.

The splitting criterion at any node t is determined by minimizing the Gini Impurity. Let p_{mk} be the proportion of training observations in the m -th region that are from class k . The Gini Impurity $H(Q_m)$ for node m is defined as:

$$H(Q_m) = \sum_{k=1}^K p_{mk}(1 - p_{mk})$$

In our binary case ($K = 2$), this simplifies. The algorithm selects the split $\theta = (j, t_m)$ consisting of a feature j and a threshold t_m that maximizes the Information Gain, or equivalently, minimizes the weighted impurity of the two child nodes Q_{left} and Q_{right} :

$$G(Q, \theta) = \frac{n_{left}}{N_m} H(Q_{left}) + \frac{n_{right}}{N_m} H(Q_{right})$$

where N_m is the total number of samples in the parent node.

Finally, for a new unlabelled job posting x' , the predicted class $\hat{y}$ is obtained by a majority vote of the ensemble:

$$\hat{y} = \operatorname{argmax}_k \sum_{b=1}^B I(\hat{C}_b(x') = k)$$

where $\hat{C}_b(x')$ is the class prediction of the b -th tree and $I(\cdot)$ is the indicator function.

2.3.2 Application

In practice, we first created a training dataset of job postings. The positive class ($y = 1$) comprised the set of vacancies successfully identified via the Levenshtein employer matching process described in Section 2.3. We classify exactly 13,721 job postings as part of the mining sector in this way. Because mining sector job postings consist of less than one percent of total postings, training a model on the raw distribution would lead to heavy bias toward the majority class. To address this class imbalance, the negative class ($y = 0$) was generated by randomly sampling non-mining job postings to achieve a strict 1:3 ratio of mining to non-mining observations in the training set.

The random forest classifier was trained on a comprehensive feature set derived from 14 distinct data points in the job postings data. These features were processed through a column transformation pipeline consisting of:

1. **Quantitative Metrics:** Minimum years of experience, standardized and imputed.
2. **Categorical Identifiers:** Nine features including geographic market names, NACE industry codes, and standardized ESCO/IDESCO occupation hierarchies, processed via one-hot encoding.
3. **Text Features:** Natural language text from job titles and skill taxonomies (names, categories, and subcategories). These were transformed using Term Frequency-Inverse Document Frequency (TF-IDF) vectorization, filtering for English stop words and capping the maximum feature space to prevent high-dimensional noise.

Importantly, we excluded the employer column from our feature set, both because our goal was to classify job postings which had an empty or non-matching employer field, and to avoid overfitting the model on the column from which the set of job postings was initially formed.

We trained the random forest, configured with an ensemble of 100 decision trees, using an 80-20 train-test split. That is, we trained the model on 80 percent of the balanced dataset and tested its precision on the remaining 20 percent. As discussed in Section 2.3.1, the output of the random forest model is the aggregated consensus of the ensemble, expressed as a probability score $\hat{p} \in [0, 1]$. We minimized the risk of false positives by imposing a strict confidence threshold. A vacancy was classified as mining only if the random forest predicted a probability $\hat{p} > 0.70$, exceeding the

standard majority vote threshold of 0.5.

The final output of the random forest model is a robust set of job postings which we classify as part of the Argentinian mining sector. This includes both jobs which explicitly have an employer identified as a mining corporation, and those extrapolated via our machine learning pipeline.

2.4 Country Grading

Before proceeding with granular labor demand analysis, we systematically evaluate the reliability and coverage of the job posting data across our set of 27 countries. We explore this data quality exercise both to better understand the data, and to motivate exploring labor demand changes in a country other than Argentina, as a means of comparison.

To establish a robust, comparative measure of data viability, we evaluated each country’s data along three primary dimensions: Completeness, Consistency, and Quantity. We then compare each country’s respective order in each of these three metrics to construct a quasi-ordinal ranking scheme to assign a final data quality rank to each country.

Countries with large populations have more complete data in terms of magnitude alone. That is, their data sets are larger, and thus will in general contain a large number of quality entries compared to a country with a smaller population. However, countries with small populations may still exhibit high quality data, even if the number of high quality quality entries is small. A major challenge in establishing these rankings was trying to ensure that small countries would not be punished for having less data, provided that their data is of high quality. As a result, in simply counting the number of complete entries, countries with large populations will almost always beat out countries with smaller populations. At the same time, if a country does have a large amount of good, workable data, it should be rewarded. As we noted, this is often the case with large countries.

Another goal of these rankings is to provide some sense of how the data quality varies with time. While we provide an overall score to each countries data, it is also important to evaluate data quality at a more granular level. Indeed, it possible that there are time periods in which data quality is strongest, which an overall score would fail to capture. Thus, our methodology calculates the metrics of Completeness, Consistency, and Quantity at a month over month level, and then uses these figures to provide an overall ranking.

2.4.1 Completeness

The completeness metric, denoted as Q_i , measures the density of the data; specifically, for each row, it evaluates how many columns contain valid values. We first calculate a monthly completeness score q_{ij} for country i in month j , defined formally as:

$$q_{ij} = \sum_{k \in \text{Columns}} \mu_k \left(\frac{m_{ijk}}{t_{ij}} \right)$$

where t_{ij} represents the total number of job postings in country i during month j , m_{ijk} denotes the number of postings with missing entries in column k for country i in month j , and μ_k is a predefined weight capturing the relative importance of column k . In practice, we proceed with an equal weighting for each column.

To aggregate this into a single, country-level completeness score Q_i , we calculate a weighted average across all months, where the weight ω_{ij} is proportional to the monthly posting volume:

$$Q_i = \sum_{j \in \text{Months}} \omega_{ij} q_{ij}$$

$$\omega_{ij} = \frac{t_{ij}}{\sum_{l \in \text{months}} t_{il}}.$$

Weighing by posting volume is to ensure that months with the largest amounts of data still dominate. The “true” completeness of data would be given by simply counting the number of complete entries across all months, rather than calculating a q_{ij} for each month j . If there were months with very small amounts of data that were, for example, very complete, our overall metric would be biased significantly upwards compared to this “true” completeness metric. However, in computing such a metric, one misses out on the granularity of the month-over-month approach. Weighing by posting volume ensures that our overall completeness score Q_i is not biased upwards (downwards) by months of low posting volume, but high (low) quality data.

2.4.2 Consistency

We desire that the data quality be stable with respect to time, and so to account for temporal volatility, we compute a consistency metric, C_i , which evaluates how the completeness of the data

varies with time. This metric adjusts the baseline completeness score by penalizing countries with high month-to-month variance in their reporting quality:

$$C_i = \frac{Q_i}{1 + \sigma(q_{ij})}$$

where the standard deviation (or “volatility” of the data), σ , is taken with respect to $j \in months$. One issue with computing volatility alone is that countries with consistently bad data would have a low volatility. However, this should not be rewarded. For this reason we multiply by Q_i . Effectively ensures that countries with low completeness volatility are only rewarded if the data is already complete.

Finally, this metric is also another means to bias our rankings against large countries, giving another metric through which smaller countries may penetrate the rankings.

2.4.3 Quantity

Finally, we are motivated by the utility of a large net amount of data to analytical models reward countries that have more data. We measure the quantity of data, A_i , using a monthly ranking system. For each month j , we rank all 27 countries by their total volume of job postings, assigning a rank $a_{ij} \in \{1, \dots, 27\}$, where 1 indicates the highest volume. The overall quantity score A_i is effectively calculated as the average of these monthly ranks (see the actual algorithm for more detail):

$$A_i = \frac{1}{\#months} \sum_{j \in months} a_{ij}$$

This initially appears to be a strange means of scoring posting volume. A perhaps more natural approach would be to simply look at the total number of postings across time and simply give countries with the most postings the highest scores. There are two reasons we do not take this approach. First, is that we would not obtain any monthly figures by basing our score on total volume. Second, is that it by definition provides no means for countries with smaller data sets to penetrate the rankings.

The main appeal of this approach is that it allows for the possibility of smaller countries to be rewarded for months of relatively high posting volume. Indeed, it is possible that a large

country may only have a few months of high posting volume, with the remaining months rest being average or below average in posting volume. During these months of lower posting volume, it is entirely possible that a country of smaller overall posting volume may actually have more data than the aforementioned large country. Aggregating across all months, this fact would disappear, and the large country always wins. However, by calculating rankings within each month, the smaller country may reach a higher ranking than it would by looking at postings in absolute terms. One major draw back of this approach, however, is that it gives no sense of how much better posting the volume of one country is over another country.

2.4.4 Composite Ranking

To arrive at a final determination of data viability, we calculated the rank of each country across the three isolated metrics: Completeness, Consistency, and Quantity. By averaging the rankings across these three metrics, we achieved a final, composite quality ranking. We choose to average the rankings instead of averaging a standardized score in order to not skew the results toward any one of the three metrics. Indeed, all three metrics provide scores on vastly different attributes of the data, any one of which may be more or less useful for any given task. Additionally, given the unique way figures in each category were calculated, the numeric values of our criteria vary from each other significantly. For example, most completeness scores however between 0.6 and 0.8, whereas the quantity scores are between 1 and 27. There is thus no sensible way to combine the scores for any given country other than to simply state how well a country ranks across the categories, on average. This is exactly what we opt to do.

Being an average of scores given ordinally, we dub the final score as being “quasi-ordinal.” The final score is not itself a strict ranking. Being an average of rankings, it communicates exactly how a country ranks on average across the three categories. In fact, across the 27 countries, there are only 18 unique scores, which themselves range from 1 to 23. One could of course rank countries ordinally based on these scores, but it would remove yet another layer of interpretability, which the rankings within each category already obscure. Again, by ranking we lose the ability to see “how much” better one countries is over another. The reported averages give a sense of how well a country is performing across the three specified dimensions, without giving any extra favor towards a particular metric.

2.5 Labor and Skills Demand Analysis

Following the extraction and classification of the Argentinian mining sector dataset via both direct employer matching and the random forest ensemble detailed in Section 2.3, our objective shifts to characterizing the labor and skills demand within this sector. We explore this demand across four primary dimensions: temporal volume, geographic and corporate consolidation, relative occupational and skill prevalence, and longitudinal growth.

To establish a robust analytical framework, we apply several targeted filtering and comparative methodologies to the raw, classified postings. A major challenge in this stage of the analysis is isolating genuine industry signals from both programmatic noise and the overwhelming volume of generic, non-specialized labor data.

2.5.1 Data Refinement and Outlier Exclusion

Before conducting comparative analyses on human capital requirements, it was necessary to address structural noise within the skills data. The text data of job postings is inherently unstructured, and algorithms used by data providers like Lightcast to extract and tag discrete skills are susceptible to parsing failures.

To identify these failures, we analyzed the distribution of the number of skills requested per job posting.

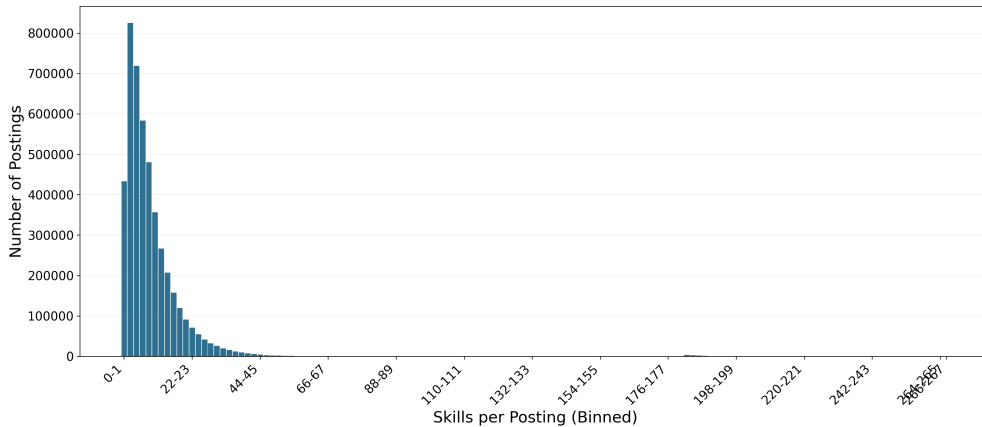


Figure 2: Distribution of Skills Extracted per Job Posting in Argentina

As illustrated in Figure 2, the distribution of skills per job posting broadly follows a log-normal curve, with 95% of postings identified as having less than 25 unique skills. However, a distinct,

anomalous long tail exists where a subset of postings were identified as having as having over 180 skills. As these matchings were clearly a result of a failure in Lightcast’s algorithm of extracting skills, we drop in posting with 180 skills from our analysis of the mining sector.

2.5.2 Geographic and Corporate Market Share

When evaluating the geographic spread and corporate concentration of the mining sector, a significant methodological hurdle is the presence of missing data. A substantial fraction of job postings lacks a defined region or explicitly named employer.

If we were to calculate the market share of a specific province (e.g., Salta) out of the total universe of mining postings, the resulting percentage would be artificially deflated by the volume of postings with “Unknown” geography. Therefore, to accurately capture the structural concentration of the sector, we calculate geographic and corporate shares exclusively relative to the subset of postings with known entities.

Let P_{total} be the total number of mining job postings, and $P_{unknown}$ be the number of postings lacking a geographic label. For a given region r , with p_r postings, its true structural market share S_r is calculated as:

$$S_r = \frac{p_r}{P_{total} - P_{unknown}}$$

This methodology ensures that our evaluation of regional dominance (e.g., the concentration of labor in the Lithium Triangle) and corporate monopolization accurately reflects the observable market landscape. We apply the same exclusion principle when calculating the hiring share of top employers.

2.5.3 Relative Demand and Disproportionality

A central question of interest is what makes the mining sector unique compared to the broader Argentinian economy. Simply ranking the most frequently occurring occupations or skills within the mining sector often yields generic roles (e.g., general administrators) or ubiquitous soft skills (e.g., basic communication) that are prevalent in almost every industry.

To isolate the specific human capital requirements that characterize the mining sector, we compute a metric of relative disproportionality. We calculate the net difference in the concentration

of a specific occupation or skill within the mining sector versus its concentration in the rest of the national economy.

Let N_m be the total number of mining postings, and $n_m(k)$ be the number of mining postings requiring an attribute k (where k is either an ESCO Level 4 occupation or a specific skill). Similarly, let N_{nm} be the total number of non-mining postings, and $n_{nm}(k)$ be the number of non-mining postings requiring attribute k . The relative proportion of attribute k in each sector is given by $P_m(k) = \frac{n_m(k)}{N_m}$ and $P_{nm}(k) = \frac{n_{nm}(k)}{N_{nm}}$. The net difference in percentage points, $\Delta_{net}(k)$, is defined as:

$$\Delta_{net}(k) = (P_m(k) - P_{nm}(k)) \times 100$$

By sorting occupations and skills by $\Delta_{net}(k)$ in descending order, we effectively filter out macroeconomic noise. This allows us to surface the highly specialized roles and distinct skill vectors that employers disproportionately demand when operating specifically within the mining sector.

2.5.4 Longitudinal Growth Measurement

Finally, we measure the evolution of the mining sector by analyzing occupational and skill shifts between 2023 and 2024, which are the last two calendar years where we have complete data.

We evaluate growth along two dimensions: absolute increase and percentage increase. Let $C_{2023}(k)$ and $C_{2024}(k)$ denote the raw count of postings requesting attribute k in the years 2023 and 2024, respectively.

$$\text{Absolute Increase} = C_{2024}(k) - C_{2023}(k)$$

$$\text{Percentage Increase} = \frac{C_{2024}(k) - C_{2023}(k)}{C_{2023}(k)} \times 100$$

While percentage increase is often informative, it is highly sensitive to small denominators. Therefore, we rank the fastest-growing occupations and skills by their absolute increase, and display the percentage increase to give readers a sense of the magnitude of these increases.

3 Case Study: Quantifying Job Demand Growth in Mexico

To quantify job demand dynamics in Latin America, we adopt a job-posting-based methodology and illustrate its application through a focused case study of Mexico using complete data from

2023 and 2024. This period represents the most recent interval for which full annual coverage was available across all relevant data sources.

3.1 Data Collection and Cleaning

The underlying dataset consists of online job postings purchased by the World Bank from Lightcast, covering employment platforms operating across Latin America. Prior to analysis, the data underwent a multi-stage cleaning and standardization process conducted by the data provider, including deduplication of postings, normalization of text fields, and harmonization of occupational labels to improve temporal consistency. Despite these procedures, residual classification and labeling errors remain; these limitations are discussed in detail in a subsequent section. Notably, among countries in the region, Mexico exhibits some of the highest overall data coverage and consistency, making it a suitable case study for methodological illustration.

Following provider-level cleaning, job postings were programmatically processed and counted using an automated counter. Each posting was recorded once per unique occurrence, with counts aggregated by job type and year. This aggregation strategy enables consistent longitudinal comparisons of job demand across occupational categories.

3.2 Occupational Taxonomy and Classification

Each job posting was categorized using a hierarchical occupational taxonomy provided by the data source. The taxonomy included the following fields:

- CAREER_AREA_NAME
- OCCUPATION_GROUP_NAME
- OCCUPATION_NAME
- SPECIALIZED_OCCUPATION_NAME

While more granular classifications (e.g., SPECIALIZED_OCCUPATION_NAME) offer higher specificity, they often fragment postings into narrowly defined categories, reducing interpretability at scale. Conversely, broader groupings (e.g., CAREER_AREA_NAME) obscure meaningful occupational trends.

After exploratory analysis, we selected `OCCUPATION_NAME` as the primary unit of analysis. This choice reflects a deliberate balance between occupational specificity and the ability to observe macro-level labor market trends. The selection was made manually following inspection of classification stability and semantic consistency across years.

3.3 Measurement of Job Growth

For each occupation, total job postings were aggregated separately for 2023 and 2024. Job growth was then quantified using two complementary metrics:

- **Absolute change:** the difference in the number of postings between 2023 and 2024.
- **Percentage change:** the relative change in postings year-over-year, calculated as the difference divided by the 2023 baseline.

To mitigate volatility arising from low-frequency postings, we excluded occupations with fewer than 10,000 total postings across the two-year period. This threshold was imposed to avoid misleading growth rates, such as extreme percentage changes resulting from very small absolute counts (e.g., an increase from one to two postings).

3.4 Observed Growth Patterns in Mexico

Applying this methodology to the Mexican labor market revealed distinct patterns of occupational expansion and contraction between 2023 and 2024. The fastest-growing occupations were predominantly service-oriented and operational roles, including building and general maintenance technicians, travel agents, cable technicians/installers, patient service representatives, loss prevention specialists, and bakers. These occupations exhibited double-digit percentage growth, indicating sustained demand in sectors tied to infrastructure, tourism, retail services, and healthcare administration.

In contrast, several technology-related occupations exhibited notable declines in job postings over the same period. These included computer programmers, IT project managers, IT managers/directors, software QA engineers/testers, and web developers. Importantly, this observed contraction is not corroborated by official government labor statistics for Mexico, which do not indicate a commensurate decline in technology employment.

We therefore interpret these declines with caution. They likely reflect artifacts of the job-posting ecosystem rather than true labor market contraction. Possible explanations include changes in occupational naming conventions, reclassification or regrouping of technology roles by the data provider, or shifts in how employers advertise technical positions (e.g., consolidation under broader titles or internal hiring channels).

Building and General Maintenance Technician	9122	12021	21143	31.78%	2899
Travel Agent	4795	5780	10575	20.54%	985
Cable Technician / Installer	9319	10505	19824	12.73%	1186
Registrar / Patient Service Representative	5316	5984	11300	12.57%	668
Loss Prevention / Asset Protection Specialist	7053	7824	14877	10.93%	771
Baker	7197	7949	15146	10.45%	752
Computer Programmer	7836	4448	12284	-43.24%	-3388
IT Project Manager	16013	9466	25479	-40.89%	-6547
IT Manager / Director	15616	9525	25141	-39.00%	-6091
Software QA Engineer / Tester	15475	9754	25229	-36.97%	-5721
Web Developer	29747	19107	48854	-35.77%	-10640

Figure 3: Top 5 highest and top 5 lowest occupations by growth.

3.5 Limitations

While job postings provide a timely and granular proxy for labor demand, they are subject to platform-specific and classification-related biases. The Mexico case study underscores the importance of interpreting posting-based trends in conjunction with institutional labor statistics and of carefully monitoring taxonomy changes over time.

4 Results

4.1 Random Forest Results

The primary objective of implementing the random forest ensemble was to capture additional job postings beyond our initial extraction of 13,721 mining sector vacancies identified via the Levenshtein distance matching of known mining companies to the employer column of the job postings data outlined in Section 2.3.

The predictive performance of the model was evaluated on the held-out 20% test set. Table 1 details the performance metrics of the random forest classifier on this validation set.

The model achieved an overall accuracy of 98%, which indicates that across the entire held-out test set, the model correctly categorized 98% of all job postings into their true class. Crucially

Class	Precision	Recall	F1-score	Support
Non-mining (0)	0.97	1.00	0.99	8,244
Mining (1)	0.99	0.92	0.95	2,732
Accuracy			0.98	10,976
Macro Avg	0.98	0.96	0.97	10,976
Weighted Avg	0.98	0.98	0.98	10,976

Table 1: Random Forest Classification Report

for our methodology, the classifier achieved a precision of 0.99 on the positive class (mining sector jobs). This indicates that of the times that the model concluded that a job was from the mining sector, 99% of the time the job was indeed from the mining sector. The recall for the positive class was lower, at 0.92, which indicates that the model identified 92% of mining job postings as indeed mining. Thus, our model is conservative. Even further, when applying this model to identify additional jobs in the mining sector, we use an even stricter probability threshold $\hat{p} = 0.70$, as opposed to the $\hat{p} = 0.50$ used for the classification report in 1. Applying this trained model to the unclassified pool of job postings successfully identified an additional 1,334 mining sector jobs, representing an approximately 10% expansion of our working dataset.

We can begin to both verify that our model learned economically meaningful patterns rather than statistical artifacts and learn about the mining sector by extracting the Gini importance of the features used in the decision splits. The top features driving the model’s classifications align strongly with the established geographic and operational realities of the Argentinian mining sector.

As illustrated in Figure 4, the most highly predictive features include one-hot encoded geographic markers and specific TF-IDF skill vectors. Notably, the locations *Salta* and *Jujuy* emerge as the top geographic predictors. This is consistent with expectations, as these northwestern provinces constitute the Argentinian portion of the “Lithium Triangle,” a region characterized by intense mineral extraction and foreign investment. A job posting located in these specific provinces is inherently much more likely to be associated with mining operations.

Furthermore, the most predictive skills include operations, quality control (QC), compliance, law, and engineering. These reflect the structural composition of the mining industry, which is heavily reliant on complex physical engineering and field operations. Simultaneously, the prominence of compliance and legal skills reveal that there are integral aspects of the mining sector beyond just physical mineral extraction. The heavily regulated nature of the sector requires specialized

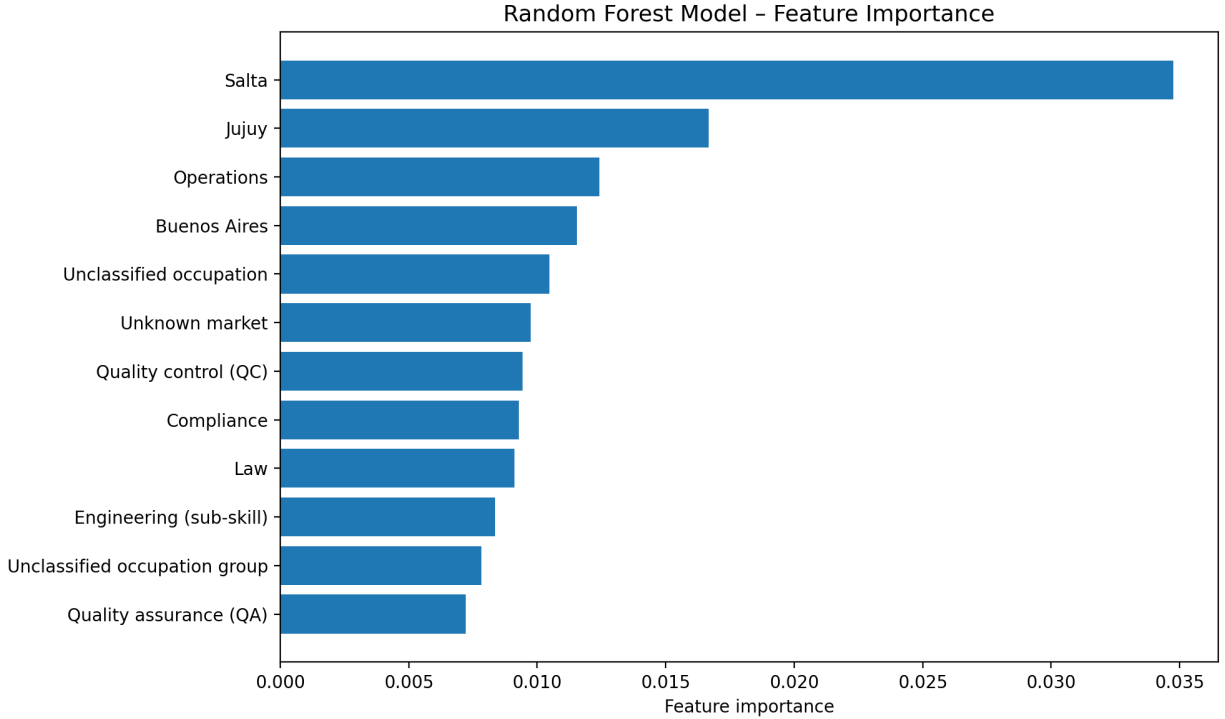


Figure 4: Top Feature Importances in the Random Forest Classifier

administrative labor to navigate local and federal environmental frameworks.

Interestingly, three unclassified fields appear as important features. We find that these are indeed more likely among jobs that we initially classified as the mining sector. While not directly insightful towards the mechanics of the industry, this indicates that work in the mining sector is less standardized than other sectors.

4.2 Country Grading

Applying the the country grading methodology in Section 2.4, we achieve a ranking of the 27 counties investigated, displayed in Figure 5.

Mexico emerged as the undisputed leader in data reliability, achieving a perfect rank of 1.0 across all three isolated metrics. It is followed in the composite rankings by Puerto Rico and Costa Rica, respectively.

A longitudinal analysis of the monthly data measurement quality—visualized in Figure 6 across the 60-month observation window—reveals several distinct temporal trends.

First, there is a macro-level trend of gradual improvement; aggregate completeness scores

country	completeness	consistency	volatility	quantity_rank	completeness_rank	consistency_rank	final_quality
Mexico	0.791356	0.772832	0.023969	1.0	1.0	1.0	1.0
Puerto Rico	0.780250	0.761180	0.025053	8.0	2.0	2.0	4.0
Costa Rica	0.762941	0.745958	0.022767	7.0	5.0	3.0	5.0
Colombia	0.749169	0.724078	0.034652	2.0	9.0	10.0	7.0
Brazil	0.729991	0.712421	0.024662	2.0	14.0	12.0	9.0
Paraguay	0.763092	0.735328	0.037757	20.0	4.0	4.0	9.0
Trinidad and Tobago	0.764839	0.732671	0.043905	21.0	3.0	5.0	10.0
Guatemala	0.745922	0.726060	0.027357	11.0	10.0	8.0	10.0
Bolivia	0.751214	0.726230	0.034403	17.0	8.0	7.0	11.0
Uruguay	0.733420	0.718673	0.020519	12.0	12.0	11.0	12.0
Cuba	0.762132	0.730320	0.043559	26.0	6.0	6.0	13.0
Guyana	0.752659	0.725365	0.037628	24.0	7.0	9.0	13.0
Ecuador	0.709692	0.694207	0.022307	10.0	17.0	16.0	14.0
Nicaragua	0.730425	0.701232	0.041631	19.0	13.0	14.0	15.0
Dominican Republic	0.721804	0.695650	0.037598	15.0	16.0	15.0	15.0
Barbados	0.741099	0.710784	0.042651	22.0	11.0	13.0	15.0
Peru	0.680048	0.660790	0.029144	5.0	21.0	21.0	16.0
Argentina	0.680045	0.661395	0.028198	5.0	22.0	20.0	16.0
Jamaica	0.722253	0.682051	0.058944	15.0	15.0	17.0	16.0
Venezuela	0.676649	0.665806	0.016285	11.0	23.0	19.0	17.0
Chile	0.659622	0.630423	0.046316	5.0	25.0	25.0	18.0
Bahamas	0.703066	0.673450	0.043977	24.0	18.0	18.0	20.0
Panama	0.642997	0.630303	0.020139	11.0	27.0	26.0	21.0
Honduras	0.661971	0.630800	0.049416	18.0	24.0	24.0	22.0
Belize	0.697011	0.657704	0.059765	26.0	19.0	22.0	22.0
Suriname	0.685136	0.634028	0.080609	23.0	20.0	23.0	22.0
Haiti	0.657760	0.628578	0.046426	17.0	26.0	27.0	23.0

Figure 5: Composite Data Quality Rankings by Country

broadly shift from lower baselines in late 2020 toward higher, more robust densities by early 2025. However, this stability is heavily stratified by country. Top-performing nations exhibit consistent, high-density reporting (with quality scores reliably exceeding 0.75) and minimal month-to-month variance. Conversely, countries in the lower quartile, such as Suriname, Belize, and Haiti, demonstrate severe temporal volatility. Their data pipelines are characterized by erratic, isolated spikes in measurement quality interspersed with prolonged periods of profound data missingness.

Examining the monthly volume of postings, as illustrated in Figure 7, illuminates a critical disconnect between data quantity and data quality.

As expected, log-volume is predictably dominated by the largest regional economies; Mexico, Brazil, Colombia, and Argentina consistently maintained the highest posting frequencies throughout the five-year window, exhibiting horizontal bands of high density with minimal seasonal interruption. Crucially, however, our results indicate that high posting volume does not insulate a country from systemic data sparsity. For instance, while Argentina and Chile both rank 5th overall in absolute posting quantity, they suffer from significant field-level incompleteness, ranking 22nd and 25th in measurement quality, respectively.

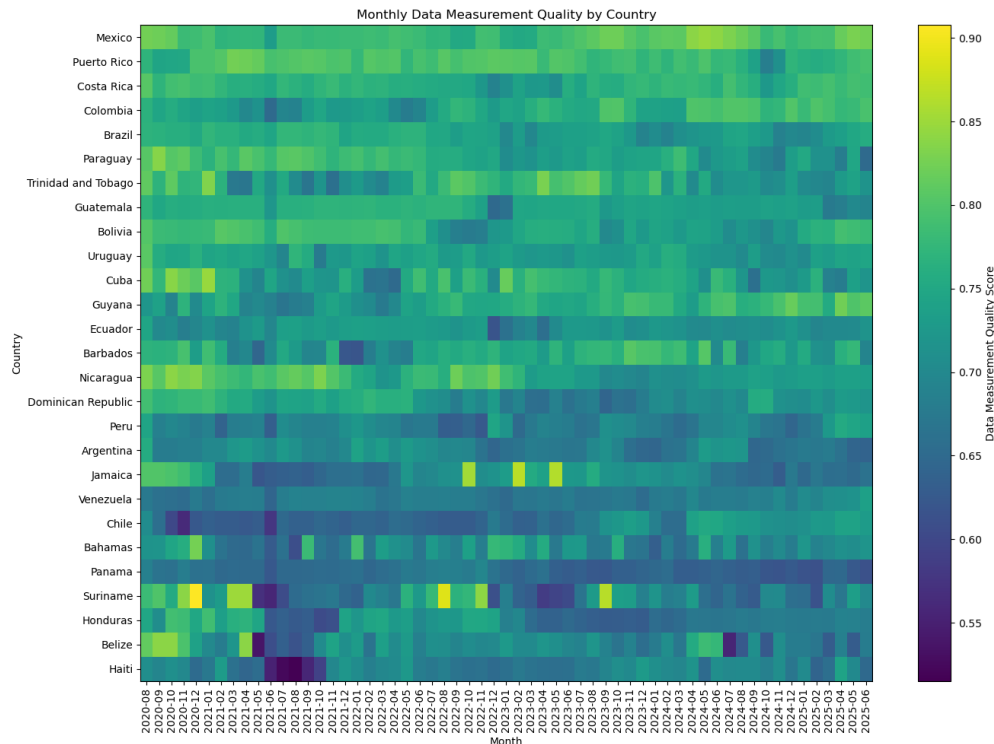


Figure 6: Monthly Data Measurement Quality by Country

These findings directly inform the structural approach of our subsequent empirical analyses. The pristine nature of the Mexican dataset—ranking first in volume, completeness, and consistency—justifies its use for direct, un-imputed occupational growth tracking. In contrast, the severe missingness observed in the Argentinian data (ranking 16th overall) necessitates the deployment of the predictive machine learning techniques—specifically the random forest ensemble detailed in Section 2.3—to reliably extract and reconstruct sector-specific labor demand.

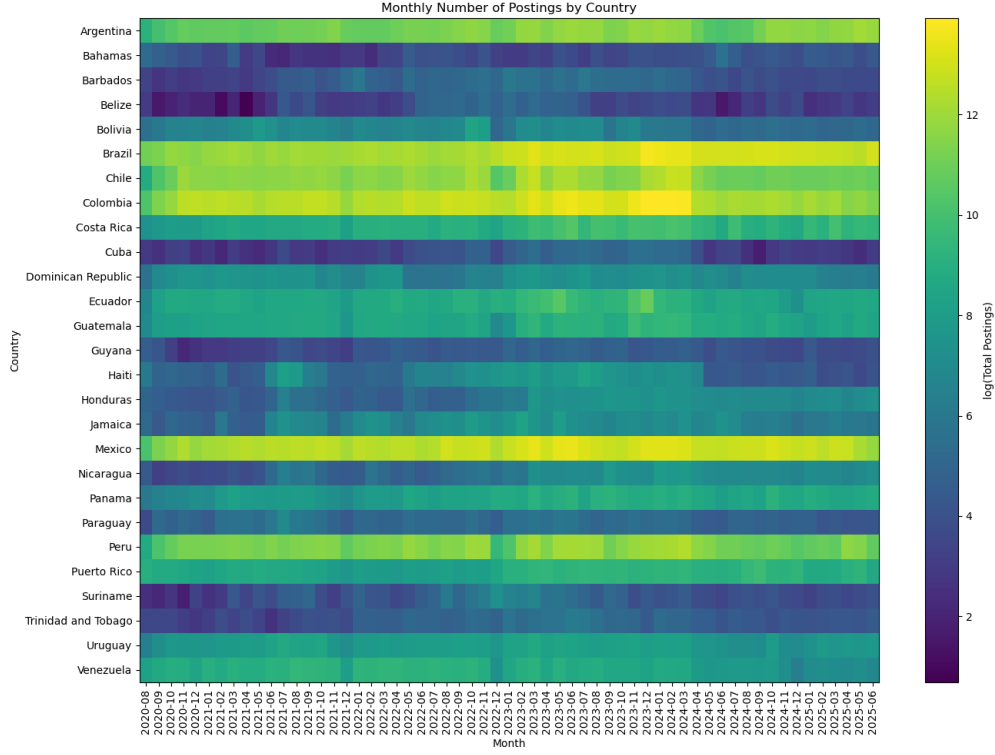


Figure 7: Monthly Number of Postings by Country (Log Scale)

4.3 Labor and Skills Demand in the Argentinian Mining Sector

Having established a robust dataset using both exact employer matching and the random forest extrapolation described in Section 2.3, we turn to a granular analysis of labor and skills demand. The following sections explore the aggregate volume, geographic spread, corporate composition, and specific human capital requirements of the Argentinian mining sector.

4.3.1 Aggregate Vacancy Trends

We begin with examining the overarching volume of mining vacancies over time.

As shown in Figure 8, the absolute volume and relative share of mining vacancies increased rapidly in 2022, and then maintained a fraction of around 0.4% of total national job postings from 2023 to 2025-06, which is the last month where we have job postings data. Note that while the increase in mining job postings from 2020 to 2023 is significant, the volatility in the percentage of total national job postings between 2023 to 2025-06 is likely due to noise in the data.

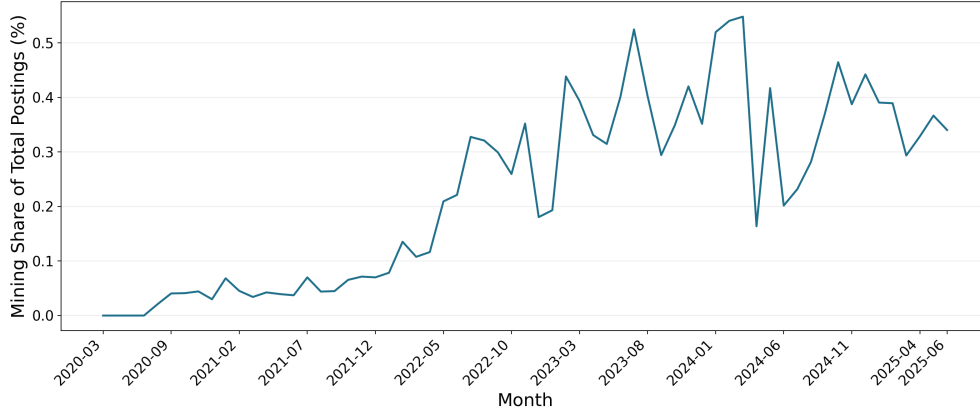


Figure 8: Mining Sector Share of Total Job Postings Over Time

4.3.2 Geographic Distribution and Evolution

The demand for mining labor in Argentina is highly concentrated geographically. Table 2 outlines the primary geographic hubs driving the sector’s hiring volume as a share of total localized postings.

Top Regions (Cities)	Share of Known Postings
Salta (ARG)	52.8%
Jujuy (ARG)	14.1%
Buenos Aires (ARG)	13.6%
San Juan (ARG)	4.2%
Catamarca (ARG)	4.0%

Table 2: Concentration of Mining Job Postings by Region (Excluding Unknowns)

We see that over 80% of mining job postings come from the three cities of Salta, Jujuy, and Buenos Aires. This reinforces the importance of geography to the mining sector—specifically, the provinces of Salta and Jujuy, which are the Argentinian anchor of the “Lithium Triangle”—dominate the sheer volume of regional demand, accounting for over two-thirds (66.9%) of all localized hiring.

We further show that this spatial concentration is not static. Figure 9 shows the distribution of postings across regions over time. Note that the table shows that the concentration of mining job postings has increased in time. Specifically, we see that in recent years the proportion of mining jobs in Salta has increased dramatically.

We further look into what occupations are present by region. Table 3 provides a three-way contrast of the most prevalent classified roles across Salta, Jujuy, and Buenos Aires, as defined by the ESCO level four occupation Lightcast’s algorithm associates with each job posting.

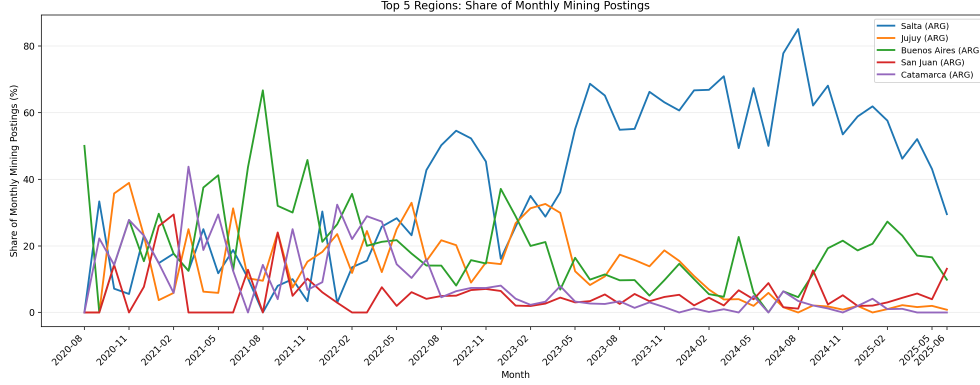


Figure 9: Share of Mining Postings Across Top 5 Regions Over Time

Salta		Jujuy		Buenos Aires	
Top Occupations	Share	Top Occupations	Share	Top Occupations	Share
Environmental health prof.	5.8%	Business services/admin	9.7%	Business services/admin	5.0%
Professional services mgmt	4.8%	Accountants	3.4%	Admin/exec secretaries	4.6%
Business services/admin	4.0%	Environmental health prof.	3.3%	Electronics engineers	3.1%
Policy admin professionals	3.6%	Manufacturing managers	3.1%	Accountants	2.4%
Geologists and geophysicists	3.0%	Freight handlers	2.6%	Civil engineering techs	2.4%

Table 3: Comparative Occupational Demand in Key Mining Regions (ESCO Lvl 4)

These results show minimal differences in what occupations are demanded in the mining sector by region. Although Buenos Aires does see the majority of occupations in corporate support, both Jujuy and Salta similarly include administrators, with administrators and accountants even being the two highest demanded roles in Jujuy. And while Salta does see the highest concentration of mechanical occupations, with their most demanded occupation being environmental health professionals, the differences in demand by region are not of statistical significance given the size and accuracy of our dataset. We thus conclude no meaningful difference in the occupational demand by region.

4.3.3 Corporate Landscape

The corporate landscape is similarly consolidated among a few key multinational players. Remarkably, over half (51.2%) of all postings tied to a known employer belong to just three mega-corporations. Table 4 details this corporate market share.

We investigate the hiring volume of these top employers over time in Figure 10. We note that the concentration of the mining sector in a few key employers is decreasing in time.

We next investigate occupational data, separating by the top three employers. We again

Top Employers	Share of Known Postings
Rio Tinto Group	25.6%
Allkem	16.1%
Ganfeng Lithium Co., Ltd.	9.5%
Posco Argentina	9.2%
Eramet	6.9%

Table 4: Concentration of Mining Job Postings by Top Employers (Excluding Unknowns)

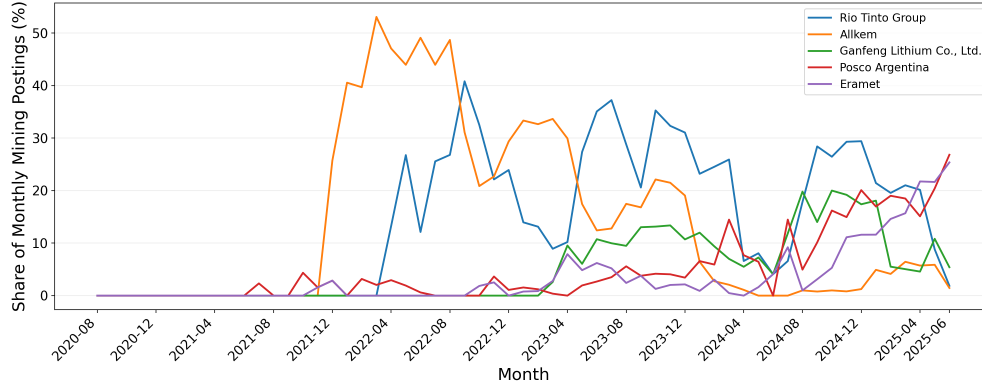


Figure 10: Share of Mining Postings by Top Employers Over Time

measure occupations by ESCO level four and display our results in Table 5.

Ganfeng Lithium		Allkem		Rio Tinto Group	
Top Occupations	Share	Top Occupations	Share	Top Occupations	Share
Construction supervisors	9.0%	Business services/admin	9.0%	Environmental health prof.	7.8%
Electronics engineers	5.7%	Admin/exec secretaries	4.8%	Professional services mgmt	6.3%
Business services/admin	4.9%	Civil engineering techs	3.8%	Policy admin professionals	5.5%
Metal working operators	4.5%	Environmental health prof.	3.5%	Business services/admin	4.3%
Engineering professionals	4.2%	Supply/distribution mgrs	3.5%	Admin/exec secretaries	3.5%

Table 5: Comparative Occupational Demand Across Top 3 Mega-Corporations (ESCO Lvl 4)

Overall, we see that similar roles are demanded across the three firms, but with different magnitudes. For example, Ganfeng Lithium’s heavy involvement in lithium processing is reflected in a bias toward more mechanical roles such as construction supervisors, metal working operators, and engineering professionals. The Rio Tinto Group, on the other hand, as a large, diversified, and highly regulated multinational, exhibits relatively greater demand for environmental health professionals, policy administrators, and management roles, consistent with the governance, compliance, and coordination requirements of a global mining conglomerate.

4.3.4 Occupational and Skills Demand

To understand the unique human capital requirements of the mining sector, we evaluate occupational and skills demand relative to the broader Argentinian economy. Table 6 highlights the top occupations and skills that are disproportionately concentrated in the mining sector compared to all other industries.

Most Disproportionate Occupations		Most Disproportionate Skills	
Occupation (ESCO Lvl 4)	Net Diff (pp)	Skill Name	Net Diff (pp)
Business services/admin managers	+3.14	Operations	+33.03
Geologists and geophysicists	+2.94	English language	+22.93
Environmental/occupational health	+2.43	Management	+15.42
Professional services managers	+2.24	SAP applications	+14.15
Electronics engineers	+1.93	Companionship	+12.64

Table 6: Occupations and Skills with the Highest Net Difference in Mining vs. Non-Mining Postings

The occupational data confirms that modern natural resource extraction is a highly complex administrative endeavor. A mix of both highly technical roles such as geologists and engineers, and administrative roles such as business managers and environmental health professionals are in high demand in the mining sector. This mirrors how the sector both requires technical expertise to perform the labor and managerial oversight to navigate complex regulatory frameworks.

The skills taxonomy mirrors this complexity. The most relatively desired skill is operations, which reinforces the technical nature of the mining sector. The massive premium placed on English language proficiency highlights the heavily foreign-invested nature of Argentinian lithium extraction. Finally, soft skills such as Companionship also appear, which is suggestive of the remote camp-based mining operations where extreme isolation makes teamwork and interpersonal cohesion critical to operational safety.

Furthermore, when we isolate just the top 10 most common mining occupations and extract their associated skills, the necessity for robust administrative and interpersonal capabilities becomes even clearer.

As shown in Table 7, nearly half of the most critical roles in the sector require administrative skills such as management capabilities and technical skills such as operations. Furthermore, over 20% of mining roles require the English language, Management, and Operations skills, reinforcing the interdisciplinary nature of the mining sector.

Rank	Top Skills for Most Common Mining Jobs	Share of Top Jobs Postings (%)
1	English language	47.5%
2	Management	44.9%
3	Operations	44.4%
4	Communication	27.6%
5	Companionship	24.2%

Table 7: Skills Most Frequently Associated with the Top 10 Mining Occupations

4.3.5 Longitudinal Growth and Digitization (2023–2024)

Finally, analyzing year-over-year changes between 2023 and 2024 suggest modernization of the Argentinian mining sector.

Rank	Fastest Growing Job Titles	Abs. Increase	% Increase
1	Web Developer	+49	2450.0%
2	Municipal / License / Court Clerk	+32	355.6%
3	Civil Engineer	+25	96.2%
4	Computer Programmer	+24	480.0%
5	Software Developer / Engineer	+23	230.0%

Table 8: Top 5 Fastest Growing Occupations in Argentinian Mining (2023–2024)

As outlined in Table 8, the occupations exhibiting the highest absolute growth are concentrated in information technology. Roles such as Web Developers, Computer Programmers, and Software Engineers are among the five with the highest absolute growth between 2023 and 2024.

This occupational shift is directly validated by the fastest-growing skills over the same period, detailed in Table 9.

Rank	Fastest Growing Skills	Abs. Increase	% Increase
1	Timelines	+329	51.2%
2	Localization	+223	3185.7%
3	Branding	+216	378.9%
4	Poco (C++ library)	+215	132.7%
5	Authorization (computing)	+202	297.1%

Table 9: Top 5 Fastest Growing Skills in Argentinian Mining (2023–2024)

The surge in demand for highly specific software engineering tools such as the Poco C++ library and Authorization computing aligns with the influx of software roles. Crucially, the explosion of software and IT roles in the Argentinian mining sector stands in direct contrast to the broader contraction of technology roles observed in the overall Mexican national economy (detailed

in Section 3.4). This divergence indicates a sub-sector undergoing rapid, localized digitization and technological upgrading, likely driven by the modernization of extraction techniques and the integration of automated monitoring systems within the Lithium Triangle.

5 Conclusion

5.1 Conclusion

This project set out to empirically quantify labor and skill demand within Argentina’s rapidly expanding mining sector, providing a data-driven foundation for the World Bank Group’s talent initiatives. By applying a random forest classifier to a massive database of online job postings, we were able to navigate a sparse data environment and reconstruct a detailed, longitudinal profile of the sector’s hiring needs.

Our findings indicate that labor demand in the “Lithium Triangle”—particularly in Salta and Jujuy—extends far beyond physical extraction. Today’s mining operations require a highly complex mix of human capital. We observe a disproportionate need for administrative, regulatory, and managerial professionals who can navigate localized environmental compliance frameworks. The consistent demand for English proficiency highlights the multinational nature of these investments, while the prevalence of soft skills like companionship speaks to the unique, isolated conditions of camp-based fieldwork.

The data from 2023 to 2024 reveals a sharp increase in demand for IT professionals, software engineers, and specific programming skills. This localized wave of digitization—occurring even as broader technology hiring contracted elsewhere—suggests that multinational mining operators are aggressively modernizing their extraction and monitoring infrastructure.

For the World Bank Group and its public-private talent platform, these insights offer clear operational direction. Addressing the sector’s binding talent bottleneck will require more than expanding traditional vocational training. To meet industry demand, local educational institutions must also adapt to supply the sector with environmental compliance experts, localized civil engineers, and IT professionals, while fostering the interpersonal and language skills necessary to operate within global mining conglomerates.

5.2 Limitations

While this methodology provides a granular look at the mining sector, our observations are inherently limited by the nature of our data source. Most importantly, online job postings reflect digital recruitment behavior, which does not capture the full reality of the labor market. The strong demand we observe for managerial, technical, and IT roles is likely amplified by the fact that white-collar jobs are heavily advertised on digital platforms. Conversely, blue-collar roles, site operators, and foundational trades may be recruited through offline channels—such as local labor unions, community centers, or internal referrals—leading to their structural underrepresentation in our dataset.

Additionally, as established in our country grading analysis, Argentinian job posting data suffers from significant field-level incompleteness. While our random forest model proved highly precise in identifying missing mining jobs, any machine learning imputation carries probabilistic uncertainty and relies on the assumption that unclassified postings follow the same feature distribution as our training set.

Finally, our analysis is bound by the constraints of the data provider’s text-parsing algorithms. As demonstrated by the necessity of filtering out postings with over 180 tagged skills, the natural language processing models used to extract data from unstructured job descriptions are susceptible to misclassification, parsing errors, and keyword stuffing. Therefore, while our findings provide a robust, macro-level proxy for evolving industry needs, they are best utilized alongside on-the-ground institutional labor statistics and direct firm-level surveys to inform final policy and curriculum design.

References

- John Abowd. The analysis of labor markets using matched employer-employee data. In *Handbook of Labor Economics*, volume 3, pages 2629–2710. Elsevier. ISBN 978-0-444-50188-2. doi: 10.1016/S1573-4463(99)30026-2. URL <https://linkinghub.elsevier.com/retrieve/pii/S1573446399300262>.
- Argentina Mining-Oportunidades de Negocios en Exploración Minería, Geología y.

Mining companies in argentina.
[empresas-mineras-en-argentina/](https://www.argentinamining.com/en/empresas-mineras-en-argentina/).

URL <https://www.argentinamining.com/en/>