



An Analysis of Argentina's *Fomentar* Employment Program

Project Lead: Darshan Kommanapalli

Analysts: Adam Cheema, Ece Ella Tumer, Koren Gila, Arko Sinha, Ana Cury *

Fall/Spring 2025

*All of the listed authors study at the University of Chicago. Special thanks to Harry Moroz and Evelyn Vezza at the World Bank for the support and patience throughout this project, and to Shivanth Krishnan and Utkarsh Dandanayak for the opportunity with CREL.

Contents

1	CREL and Project Overview	3
1.1	Charles River Economics Labs	3
1.2	Project Background	3
2	Introduction and Motivations	4
2.1	Motivation	4
2.2	Summary of Methodologies	4
3	Literature Review	4
3.1	NLP Matching	4
4	Data and Methodology	5
4.1	<i>Fomentar</i> Datasets	5
4.1.1	Course Database	6
4.1.2	Participant Database	6
4.1.3	ESCO and ISCO Occupation Hierarchy	6
4.1.4	Course Mappings	7
4.2	Unique Course Filtering	7
4.3	Geography and Course Offering Accessibility	9
4.4	Previous Employment and Course Alignment	12
4.4.1	Participant Mappings	12
4.4.2	Course Mappings	15
4.4.3	Creating Course Alignment	16
4.5	Course Participation and Alignment Regressions	16
5	Results	17
5.1	Course Distances and Participation Rates	17
5.2	Course Alignment Statistics	21
5.3	Course Participation Regression	23
5.4	Course Alignment Regression	24
5.4.1	Discussion	24
5.4.2	Issues	26
6	Conclusion	27
6.1	Conclusion	27
6.2	Next Steps	27
7	Appendix	27

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.

This project falls under the *Social Protection & Labor Global Practice* which delivers evidence-backed strategies, programs and structures to give people access to better livelihoods through social safety assistance and labor market programs.

1.2 Project Background

In September 2021, the Argentinian government created the *Fomentar* program to increase the employment of unemployed and informally employed workers. It falls under the Secretariat of Labor, Employment and Social Security (*Secretario de Trabajo, Empleo y Seguridad Social*, STESS) in the Ministry of Human Capital (*Ministerio de Capital Humano*). As well as receiving a stipend, program participants undergo labor market orientation (online and in person), vocational training, internships and labor market intermediation (for example, job matching). The program targets young people (men and women) aged 18 and 24, women aged 25 to 59, men aged 45 to 64 and several other smaller disadvantaged groups.

As of 2024, participants of another program, *Volver a Trabajo (VAT)* (*return to work*) were also eligible for Fomentar's services. VAT participants are required to take a training course to receive a stipend, Fomentar participants are not.

This project falls under the World Bank's *Social Protection & Labor Global Practice*, which delivers evidence-backed strategies, programs and structures to give people access to better livelihoods through social safety assistance and labor market programs.

2 Introduction and Motivations

2.1 Motivation

This project focuses on the vocational training provided as part of Fomentar’s services. Training courses are published on the government of Argentina’s employment portal. Courses are sponsored by counterparties across a wide variety of industries. There is ongoing research on the effectiveness of these courses for improving employment outcomes, but the composition of these courses, and in particular how representative they are of labor market demands, is not fully understood.

Our goal is to investigate if participants are using the program to move vertically or horizontally within the labor force. That is to say: are individuals looking to further specialize in fields that they are already established in, or are they looking to pivot into different fields? We use Natural Language Processing and several probit regression models to investigate the program and its efficacy. More general descriptive statistics demonstrate notable trends across participant activity relevant to course modality and geographic location.

2.2 Summary of Methodologies

3 Literature Review

3.1 NLP Matching

The rapid development of Artificial Intelligence has dramatically increased the efficiency with which researchers can classify large quantities of data. Machine Learning assisted taxonomy serves as a promising method to organize large unstructured qualitative datasets. One specific approach being adopted by labor economists is Natural Language Processing (NLP). The NLP approach allows researchers to universally compare large amounts of labor market data by matching qualitative data based on the language used in it’s descriptions. This process is repeatedly ran at more granular levels, creating a tiered classification system of job markets. The leading paper employing this method, Djumalieva 2018, used clustering based on skills and salaries to create a classification system of Britain’s labor market using 37 million online job postings. Furthermore, Dandanayak 2024 uses Latin American labor market data to create a network of occupations grouped by skills.

Majority of the literature involving NLP matching performs two main steps in their methodology. First, descriptive data such as skills and salaries, are embedded from their original text to vectors. The literature does not agree upon a standardized process for this, with Djumalieva using the GloVe model, and Dandanayak using BERT. BERT specializes in recognizing words in context, creating embeddings using both the words as well as their context. GloVe generates a single vector for each word, trading contextual embedding for the benefit of reduced noise. For further context, the embedding process is repeated on labor dictionaries, such as the Dictionary of Skills and Competences (DISCO).

The next step of this process is the repeated clustering of the vector embeddings. Clustering matches vectors based on semantic understanding, grouping occupations that have similar skill sets or salaries. For example, a job that requires "coding experience" may be matched to a job that requires "python experience". Similar to the embedding process, there is no standardizes process that the literature agrees on. In fact, most papers use several methods as the taxonomy becomes more granular. Djumalieva uses Gaussian Mixture Models, as well as K-means clustering. Dandanayak employs HDBSCAN, a model similar to K-means clustering, but without the predetermined number of clusters, allowing for more versatile pairings. This process is then repeated for several iterations on the smaller groups.

The final result is a pyramid shaped taxonomy, grouping jobs based on qualitative data. Without NLP, this process would have to be organized by an individual. Although this task is mundane at a small level, it allows for previously impossible in-depth analysis at a large scale. Djumalieva used this method to categorize 37 million job postings in the United Kingdom.

4 Data and Methodology

4.1 *Fomentar* Datasets

The majority of data analysis utilized four datasets provided by the World Bank across *Fomentar* and VAT participants, courses offered by *Fomentar* and its partnering institutions, as well as occupation classifications.

4.1.1 Course Database

The World Bank provided a dataset includes a list of 13,633 training courses offered between April 2022 and September 2024. Each item includes a unique identifier, the course name, information on the institution sponsoring the course, a topical occupation category (“typology”), the modality of the course (e.g. in-person, remote), and an address of the building where sessions are held for in-person and hybrid courses. Some courses are offered multiple times over the 2.5-year period, though with different modalities, locations, or names. A subset of 6,408 courses contain additional information, including a brief description of course content, prerequisite skills, and workload. Much of this information is text-based, and hence slightly less structured with more parsing errors. We address these concerns in (link unique course filtering section).

4.1.2 Participant Database

The World Bank provided a list of 1,242,416 participants registered with the *Fomentar* or VAT programs. Each participant has basic demographic information, including age, sex, education level, and municipality in Argentina at time of registration. Other information of interest provided by this dataset include the unique identifiers of courses taken by participants, as well as a brief description of a participant’s previous labor experience should they have any. The “Participant List” recognizes more courses than the “Course List”, as many participants enroll in courses with unique course identifiers that do not exist in “Course List”. It’s possible that these courses were offered prior to April 2022 or lack information in other categories (e.g. modality) that excluded them from being added to “Course List”.

4.1.3 ESCO and ISCO Occupation Hierarchy

The European Skills, Competences, Qualifications and Occupations (ESCO) system recognizes 3,039 occupations and 13,939 related skills within the EU labor market. Each ESCO occupation is a child node of a four-level hierarchy of occupation groups of varying specificity recognized by the International Standard Classification of Occupation (ISCO). For example, a “branch” of the ISCO-ESCO scheme is “Professionals ¿ Health professionals ¿ Medical doctors ¿ Specialist medical practitioners ¿ Specialized doctor,” in order of the first four ISCO groups and final ESCO

occupation. Despite ESCO being oriented toward the EU, the classification scheme is widely used and valuable for assessing the similarity between occupations and outcomes.

4.1.4 Course Mappings

For each course in a collection of 110 courses recently offered by *Fomentar* in February and March 2025, the World Bank provided a set of skills and occupations related to the course content. Unlike the “typology” for the courses in “Course List”, each occupation and skill in this dataset belongs to the set of ESCO occupations and skills. Standardizing related occupations by the ESCO scheme enables us to use the ESCO/ISCO dataset as a bridge between participant-specific occupations and course-specific occupations (link alignment section). The ESCO/ISCO classification can also be downloaded natively in both Spanish and English rather than manually translating the English dataset.

4.2 Unique Course Filtering

The 6,408 courses in the “Course List” with textual information on course content are especially valuable because they can be parsed for skills or compared to other courses in terms of content similarity. Attempting a similar procedure on only the basis of the name of a course, which is sometimes only 1-3 words, may be uninformative. A “duplicate” in this set of courses is a course that has been offered multiple times between April 2022 and September 2024. To avoid comparing a course against itself, duplicates must be removed. This is challenging because 1) every course offered at each time has a unique identifier, including duplicate courses, and 2) the name or content of a course may slightly change, even though the course is fundamentally the same.

We take a systematic approach to course filtering by using sorting and string-similarity methods. We will use string-similarity methods multiple times throughout our analysis. 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 claim that the majority of true “duplicate” courses should not have substantial changes to course content or name, and hence the Levenshtein distance between duplicate courses with different names or content descriptions should be reasonably small (and the similarity score reasonably close to 1). One benefit of the Levenshtein metric is that it is language-agnostic and independent of context; the course list is in Spanish, and using machine learning methods to accomplish a similar goal through the contextual meaning of name or content would depend on the quality of the Spanish data used to train the model.

A course group is defined as a set of courses that are duplicates. To identify duplicates, we first sort courses alphabetically by name, then by offering institution, and then by course content. Note that course duplicates often differ by variations that come at the end of the course name, and thus sorting likely groups duplicate courses together. Fixing the first course as a “reference course,” we compute the similarity score between each sequential course and the reference course for the name, offering institution, and course content. Taking the product of these three similarity scores gives an aggregate similarity score. We designate a similarity threshold of 0.9 for each category;

that is, courses with aggregate similarity score less than $0.9^3 = 0.729$ are considered members of a new course group. Given that we sorted alphabetically across three categories, sequential course similarity scores for a fixed reference course are generally decreasing, as alphabetical distance correlates to similarity distance. We recognize slight variations in this assumption in the following paragraph. Then the first sequential course that scores below 0.729 is considered the end of the previous course group and the start of a new course group; the process is repeated with this course as the new reference course. Formally, for any fixed reference course A and sequential courses $B_1, B_2, \dots, B_m$, the course group is given by $G = \{B_i \mid s(A_{\text{name}}, B_{i,\text{name}}) \cdot s(A_{\text{content}}, B_{i,\text{content}}) \cdot s(A_{\text{inst.}}, B_{i,\text{inst.}}) \geq 0.729\}$. Thus, we obtain a set of course groups $\{G_1, G_2, \dots, G_n\}$. Picking the reference course of each group, we obtain a list of unique courses $\{A_1, A_2, \dots, A_n\}$.

Finally, to ensure that courses are truly unique, we compute similarity scores for each pair of courses and drop an individual course when $s(A_i, A_j) \geq 0.729$. In a handful of cases, special characters or text parsing errors caused a highly similar course to be located several positions from its reference course in the original sorting. This course should be considered a duplicate, but instead it forms a new group as its own reference course. The second round of similarity scores filter these courses. We conclude with 816 unique courses and additional information on the number of times each unique course was offered by modality since April 2022.

We excluded a number of other fields provided for these 6,408 courses, such as the address of in-person course sessions or the average workload. It is possible that these variables could produce a more precise course filtering. However, as we consider more variables, we need to lower the similarity threshold to be more tolerant of courses that score highly in a number of fields but fail in one (e.g. all fields match, except the course is offered in a different location). This tolerance may enable fundamentally different courses to be considered duplicates. In general, for courses from the same institution, we can expect course information to not deviate significantly between offerings, i.e. we can be relatively confident in our filtering.

4.3 Geography and Course Offering Accessibility

In order to investigate the geographic accessibility of course offerings, we make use of the address associated with each course provided in “Course List”. We attempt to create a dataset which shows how far away each course is from each municipality in Argentina.

Specifically, we generate driving distance isochrones for each municipality in Argentina. In this context, an isochrone represents a set of locations that are within a certain driving distance away from a point. A “10 minute” isochrone would represent all the locations that one could drive to in 10 minutes. We generate six different isochrones for each municipality, and then group courses into their corresponding isochrones: Within 10 minutes driving distance, between 10 and 20 minutes driving distance, within 20 and 30 minutes driving distance, within 30 and 40 minutes driving distance, within 40 and 50 minutes driving distance, within 50 and 60 minutes driving distance, and greater than 60 minutes driving distance.

To create this dataset, we first need to quantify each course address and each municipality in terms of a location coordinate. To do this, we make use of the Google Maps API. We find that the Google Maps API is most robust in finding accurate coordinates, especially in cases where addresses have slight typos or variations in spelling. To that end, we approximate the location of a municipality as the specific location coordinate that the Google Maps API associates with it.

We now proceed with creating our desired dataset. To generate isochrones, we make use of the OpenRouteService API. An isochrone, in this context, is a polygonal region containing all points reachable within a specified travel time from a municipality, given existing road networks and transportation constraints. Conceptually, this involves modeling the road system as a weighted directed graph, where each node represents a junction or road segment, and each edge has an associated cost, which is calculated based on average speed.

Mathematically, this is a variant of the shortest-path problem. Let the road network be represented by a graph $G = (V, E)$, where:

- V is the set of vertices (e.g., intersections, road segment endpoints),
- E is the set of edges (roads)
- and each edge $e \in E$ has a travel time cost $c(e)$.

Given a source node $s \in V$ (a municipality), the API computes the shortest travel time $d(v)$ from s to every other node $v \in V$, using a modified Dijkstra’s algorithm adapted for travel-time constraints. For a specified threshold T , the isochrone is defined as:

$$\text{Isochrone}(s, T) = \{\text{location}(v) \in \mathbb{R}^2 : d(v) \leq T\}$$

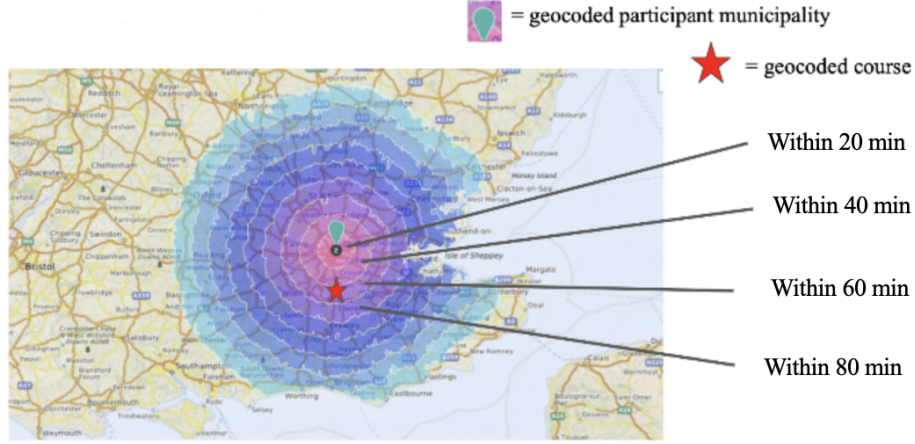


Figure 1: Visualized isochrones for London

This set is then approximated to form a continuous polygon in geographic space.

This is a much more genuine measurement of accessibility than air distance, as it tracks the time opportunity cost of participants who would decide to take a certain course.

We then create a database of all the courses that are within our 6 buckets of driving distance for each municipality. To do this, we first create a list of all potential course candidates that could be within 60 minutes driving distance by filtering for all courses that are within 100 air miles from the center of each municipality. We then check which bucket they fall into, if any. We place all the courses that do not fall into any bucket, including those that were not in our original 100 air mile filter, in the 7th bucket, of greater than 60 minute driving distance. Doing this generates a dataset with all the courses that are within a reasonable driving distance and a list of all the courses which are not reasonable to travel to from each municipality. Note that this analysis is only done for in-person or hybrid courses, i.e. courses that participants would have to attend in person at least once a week.

Because we do not have information about participant addresses, and only have access to their municipality as a proxy for location, when making use of this dataset, we approximate each participant as living at the municipality coordinate generated by the Google Maps API. Although this may lead to slight imprecisions with regards to the amount of time it would take a participant who lives in a certain municipality to drive to a certain course, most municipalities in Argentina are less than 10 minutes of driving from edge to edge, this approximation does not have a significant affect on which driving time bucket a course theoretically should fall into for a given participant.

Another limitation of applying the municipality data to participants is that it is possible that participants move locations after registering for the Fomentar program. Given this, when analyzing this data, we expect to see a spike of people who we report as taking courses greater than 60 minutes away. Assuming that the data from the first buckets is generalizable, we can estimate the true proportion of people we expect who take courses greater than 60 minutes away. We can compare this to the data we measure to analyze the extent to which our greater than 60 minute data is inaccurate.

4.4 Previous Employment and Course Alignment

One area of interest is to identify how participation in a specific course is related to the occupation history of a program participant. A potential indicator of the Fomentar and VAT programs being used for vertical mobility is data suggesting that participants take courses very similar to their previous occupation. However, if participants enroll overwhelmingly in courses unrelated to their previous jobs, course participation may reflect lateral mobility. We employ natural-language processing methods to complete multiple “mappings” of participant occupations and course typologies to the ISCO occupation hierarchy. The World Bank provided a correlation matrix comparing each of the 437 ISCO Level 4 occupation groups through NLP methods on the skills associated with them. Specifically, for any two ISCO occupations, the matrix provides a quantitative score $s \in [0, 1]$, where s equals 1 if and only if the two ISCO occupations are exactly the same. Thus, once both participant occupations and course typologies are mapped to the *same* ISCO hierarchy, we can readily compute similarity scores using the correlation matrix.

4.4.1 Participant Mappings

We first map participants’ previous occupations onto the ISCO hierarchy. Data provided by the World Bank includes the name of the previous occupation, in Spanish, for about 804,000 participants enrolled in the Fomentar program. These occupation names are recorded at the time of program enrollment by the participant, and hence they are non-standard. Many occupations are significant textual variations of each other but are fundamentally similar. These differences prevent standard text-based filtering with Levenshtein scores as we previously used in 4.2, as we must consider the contextual meaning of the words within an occupation name. Natural-language-

processing allows for the comparison of two strings based on *contextual* similarity, which produces a similarity score between 0 and 1 that is distinct from Levenshtein scores. The standard logic of NLP mapping is to create a “domain” list of previous occupation names, a “co-domain” list of ISCO occupation names and descriptions in Spanish, and then use a language model to compute a many-to-one function that maps each element of the domain to the most closely related element in the co-domain. If two occupation names are fundamentally similar but textually distinct, they likely have similar contextual meanings, and NLP methods will theoretically map them to the same ISCO group as an element in the co-domain.

We use a multilingual model of BERT (Bidirectional Encoder Representations from Transformers) as the language model to map participant occupation name to ISCO group name. We forgo an in-depth explanation of BERT and provide only its basic procedure. For two strings, BERT will tokenize each string into substrings (“tokens”), create vector embeddings for each token, and finally compute cosine similarity scores between vectors. The highest cosine similarity score determines the mapping between domain and co-domain.

One challenge is to determine which ISCO level to use as the co-domain. Although the correlation matrix provided by the World Bank is in terms of ISCO Level 4, we need to assess any opportunities to produce more accurate mappings that may provide better comparisons against course typologies. The intuition is often to map to the co-domain with greatest specificity, as this co-domain likely has the most meaningful textual information that can improve the accuracy of the mapping function. However, the most specific ISCO Level—the set of ESCO occupations—suffers from a similar problem as the participant data, as the 3,000 ESCO occupations are also highly contextually similar but textually distinct. Likewise, when using the ESCO occupations as the co-domain, many contextually similar participant occupations are mapped to distinct ESCO occupations. This approach produces a set of mappings with very few many-to-one matches (where multiple participant occupations map to the same ESCO occupation) despite mappings being far more similar than their separation would suggest.

For each participant occupation, BERT can also compute a vector of the top 3 highest cosine similarity scores with the co-domain, or the three most closely related ISCO occupations ordered by rank. For a participant occupation x , we have $x \mapsto ((o_1, s_1), (o_2, s_2), (o_3, s_3))$ where o_i is an ISCO occupation and $f(x, o_1) = s_1 > s_2 > s_3$ is the BERT-generated similarity score. We are interested in

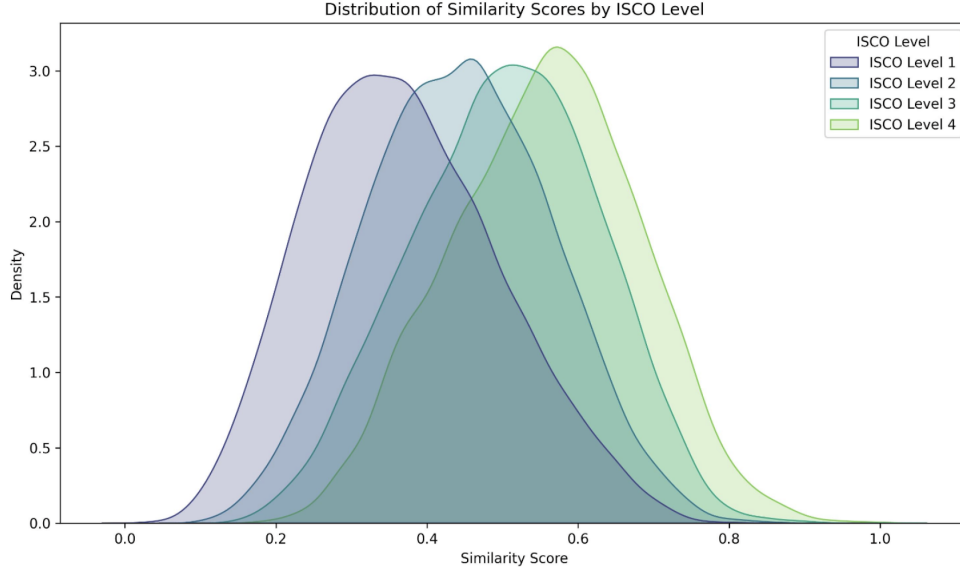


Figure 2: Similarity score distributions with different ISCO co-domains

mappings where $(s_1 - s_2) \gg 0$, or where there is a significant “drop-off” in similarity score between the most closely related ISCO occupation and the second most closely related ISCO occupation for a given participant. A large similarity score difference suggests that there is a single ISCO occupation that is a best match, rather than a collection of ISCO occupations that are all arbitrarily similar to one another and the participant occupation in question. Indeed, when completing a map with ESCO occupations as the co-domain, we generally find that $s_1 - s_2 \approx s_2 - s_3 > 0$.

To identify whether to use ISCO 1, 2, 3, or 4 as the co-domain for the calculation of the alignment variable, we run 4 separate models with the participant occupation as the domain and each of ISCO 1-4 groups as the co-domain. For each participant occupation within each model, we record both the highest similarity score with an ISCO group and the difference in similarity score between the closest matching and second-closest matching ISCO group. Figure 3 plots these “drop-off” percentages. We recognize that ISCO Level 4 has generally lower drop-off percentages than ISCO 2-3, but not by a significant amount. Similarly, ISCO Level 4 also demonstrates higher similarity scores on average than ISCO 1-3. We select ISCO 4 as the final co-domain, as there is a compromise between increased specificity and accurate matches for the purpose of future regressions. We conclude by mapping every participant’s previous labor experience to its most related ISCO Level 4 group.

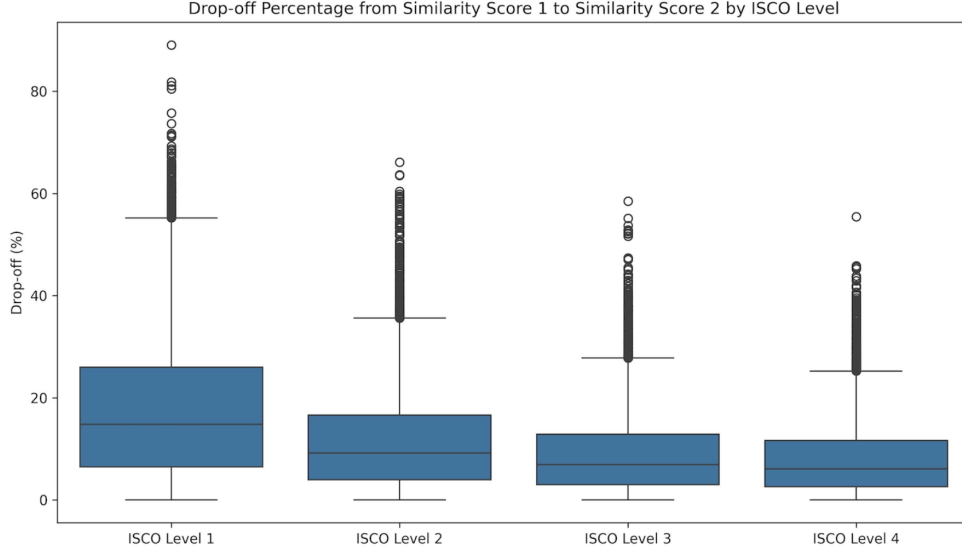


Figure 3: Similarity score differences between ISCO closest matches

4.4.2 Course Mappings

We next map courses to ISCO occupations. We make use of a provided World Bank dataset which, for a of 119 recent courses, outputs a set of approximately five ESCO occupations which the courses are associated with. This dataset was generated through a third party company called Lightcast.

Our first goal is to map courses in Course Database to one of the 119 recent courses, for which we have associated ESCO occupations. Of course, not every course in Course Database has a representative course in one of the 119 recent courses, and we only include those that do in our analysis. To match courses, we use a standard text-based filtering with Levenshtein scores as previously discussed in section 4.2, except we only match based on course name, as we don't have course content, and we don't expect institution to be the same for these courses. Similar to section 4.2, we use a similarity threshold of 0.9. Doing this, we are able to match 9203 of the 13633 courses in Course Database.

Once we map courses to courses in Course Database to a set of ESCO occupations, it is trivial to map them to ISCO occupations, as each ESCO occupation maps directly to an ISCO occupation which it is a part of. Performing this mapping, for 9203 courses in Course Database, we have a set of associated ISCO occupations.

4.4.3 Creating Course Alignment

Once we have participants' previous occupation mapped to an ISCO group, and a set of courses mapped to ISCO groups, we can proceed with creating the alignment variable using the provided correlation matrix. We do this by first forming a dataset of all participant course pairs where we have data for both the participant and the course they took. We end up with $\approx 150,000$ such pairs. Note that if a participant took multiple courses, they will appear in multiple participant course pairs. Next, we use the correlation matrix to calculate alignment between the participant and the course they took. Because we have multiple ISCO occupations for each course, we compute a similarity score between the participant ISCO group and each associated course ISCO group, and then choose the max to be the course alignment score.

One drawback of this dataset is that a subset of the courses provided in Course Database are general education courses: that is, they do not teach any specific skill-set, and instead teach general principles for how to prepare for work. Because any course alignment score with these courses would be meaningless, we choose to ignore these courses both when generating descriptive statistics with this dataset and when using this dataset for our regression. After removing this courses, our set of courses participant pairs decreases to $\approx 90,000$ pairs.

4.5 Course Participation and Alignment Regressions

In order to investigate co-linearities and better understand the data before running our final regression, we chose to run a preliminary regression answering the question: *What factors determine whether or not someone takes a course?* . Given the binary nature of our outcome variable, and the standard distribution of our independent variables, we chose to run a probit model. Our equation took the form

$$Pr(y_i = 1|x_i) = \Phi(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i}) \quad (1)$$

With x_1 = work experience, x_2 = employment status, x_3 = sex, x_4 = population of municipality, and x_5 = age. Work experience was defined as a dummy variable with a base value of no work experience. Employment status is a binary variable indicating whether or not an individual was working at the time of joining the program. Note that the fomentar program targets unemployed individuals, so

current employment refers to informal employment. Age has been split up into categories which are presented in the regression as dummy variables. Note, that the education variables refer to that level of education having been completed. We chose these variables after extensive research into the descriptive statistics and make-up of participants. This preliminary regression confirmed a statistically significant relationship between our selected independent variables and whether or not a participant took a course.

Our second regression focused on what specific factors may influence an individual to pivot into a new industry through the program. As a result, we tried to determine what factors influence whether or not someone takes a course aligned with their past work experience? Our output was a binary variable that measured whether or not an individual's work history aligned with the content of their course. It's important to note that in order to run this regression, we filtered out any participants taking general studies courses, as these courses do not align with any specific work history. Given the binary nature of the outcome variable, we also used a probit model for this regression. This equation took the form:

$$Pr(y_i = 1|x_i) = \Phi(\beta_0 + \beta_1x_{1i} + \beta_2x_{2i} + \beta_3x_{3i} + \beta_4x_{4i} + \beta_5x_{5i}) \quad (2)$$

5 Results

5.1 Course Distances and Participation Rates

Before we use apply our course distance database to our regression, we first analyze the data directly. We are interested in two questions:

Before applying our distance database and running our regression, we performed a preliminary data analysis. Within the analysis, we were interested in two questions. How much time are participants spending on getting to their courses and how are courses distributed across Argentina's provinces?

A satisfying answer to our first question comes from Figure 4. It reveals that over 70% of participants who take in-person courses live within 10 minutes of driving distance from the participant's residence. It also shows the expected spike of people who we calculate that take courses over 60 minutes away from their reported residence, which comes from relocation. We aim

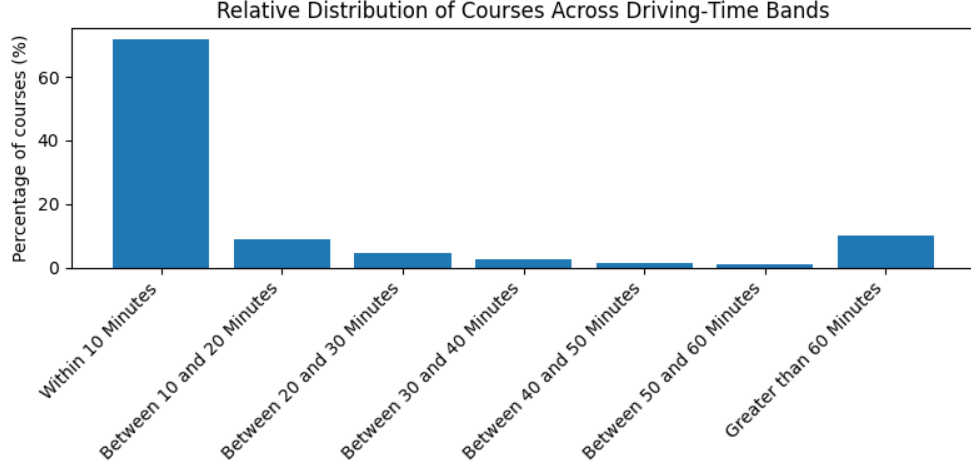


Figure 4

to investigate how much of this spike is expected, and how much likely comes from participants who have moved from their original reported address.

Figure 5 shows various equations of fit for our data on courses within 60 minutes. We find that a Weibull fit best models our data, and produces an R^2 of 0.998. Quantitatively, we express this fit as:

$$F(t) = 1 - e^{\left[-\left(\frac{t}{4.8483}\right)^{0.6189}\right]} = P(T \leq t)$$

Where $F(t)$ represents the percentage of people who take courses within t driving minutes.

Extrapolating from this fit, we find that the expected percentage of participants who take in-person courses and drive further than 60 minutes away is 0.9%, much less than the 10% reported in our data. Thus, we conclude that $\approx 90\%$ of the data for courses which are reported to be taken by those who live more than 60 minutes away actually comes from people who have moved, and thus have unreliable location data. Because we cannot differentiate between these courses to figure out the $\approx 0.9\%$ of reliable data, we simply remove all data points where we report that a participant traveled more than 60 minutes to take a course when performing our regression analysis. Descriptively, we find that our Weibull fit predicts that in reality, almost 80% of participants who take in-person courses live within 10 minutes of the course address.

We next investigate how courses are distributed across Argentina. We find the following data for the municipalities with the most courses within 60 minute driving distance. The municipalities with the most courses within 60 minutes driving distance all reside in Buenos Aires, which is

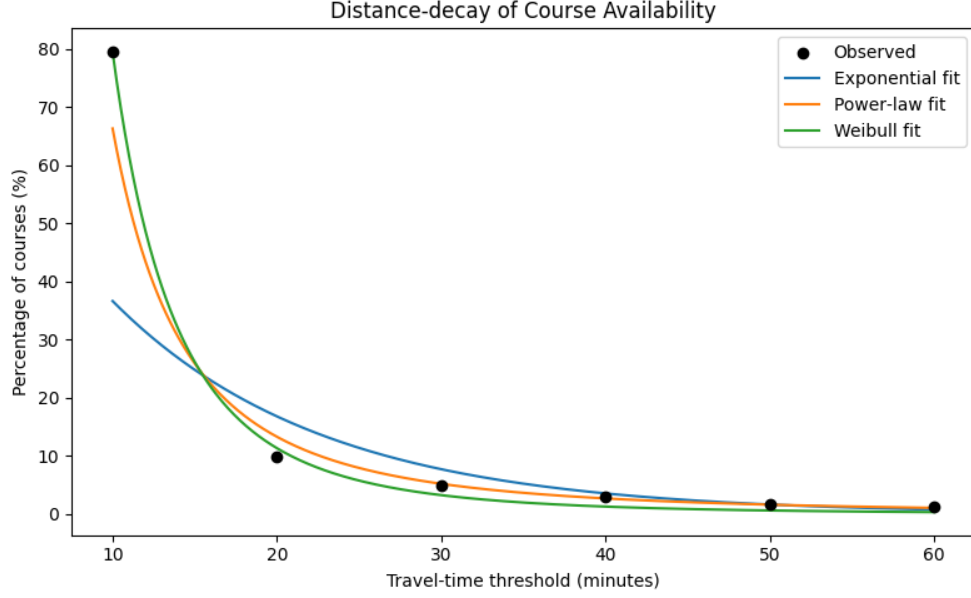


Figure 5

in correspondence to Buenos Aires being the most populated province. Note that all of these municipalities share the exact same courses within 60 minute driving distance, which speaks to the density of Buenos Aires as a province, and the fact that all the courses here reside inside the urban area.

The correspondence to population falters slightly when we look at Figure 6, which shows the top ten municipalities by number of courses, but filters for only one municipality from each province. Here, we see that the number of courses near Buenos Aires is roughly four times that of the next most province, which is less than what we would expect if the number of accessible courses were strictly proportional to population. Moreover, the province with the second most courses is Tucumán, which is only the fifth most populated province in population. The rest of the provinces in the top 7 of population, along with provinces 11-13th are all found in the top 10, but the order is not in accordance to population. The second most populated province, Córdoba, has only the 10th highest number of courses. Thus, if we define equity as having accessible courses proportional to population, the distribution of courses throughout Argentina could be significantly improved.

However, as Figure 7 shows, participants are not distributed throughout Argentina strictly proportional to population. For example, Tucumán, which is the fifth most populated province in Argentina, contains the second most participants. Recall that it also contains the second most courses within 60 minutes driving distances. When we examine how the courses are distributed

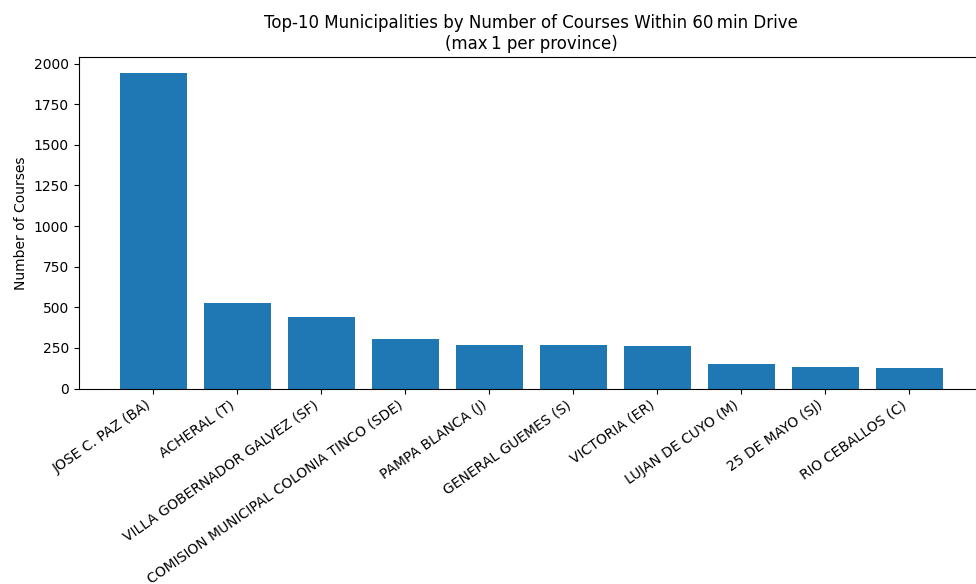


Figure 6

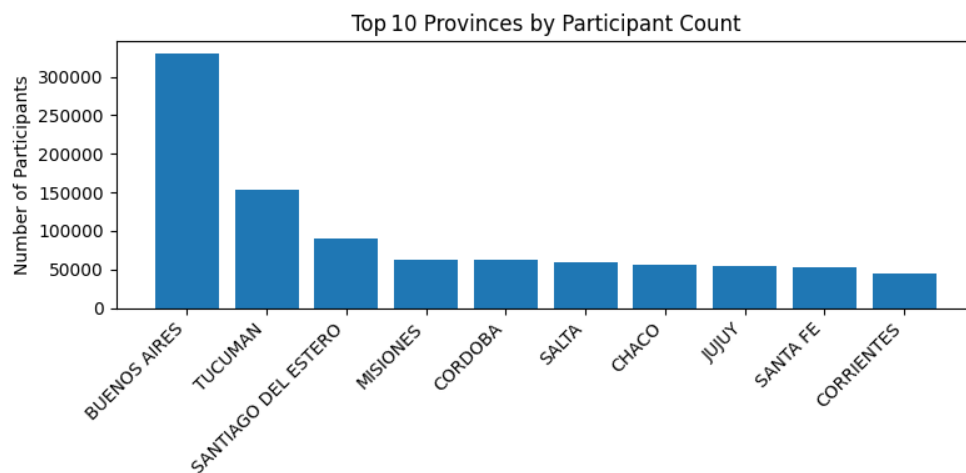


Figure 7

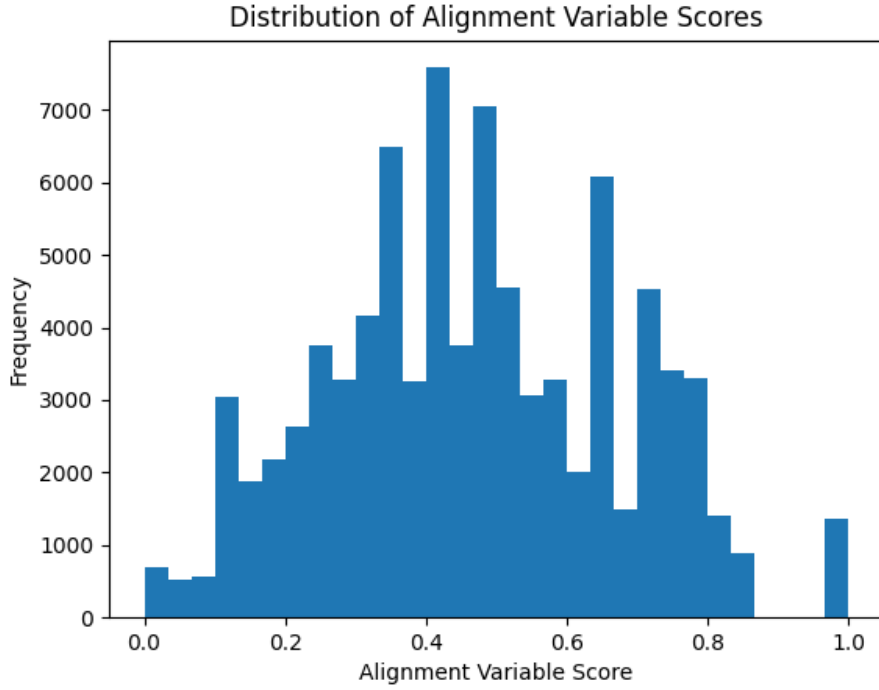


Figure 8: Excluding General Courses

proportional to the participants enrolled in the program, we see that they are indeed very equitable. One interesting facet of Figure 7 is that although Buenos Aires is the province with the most participants, it only has approximately twice as many as Tucumán, even though Tucumán has $\approx 10\%$ the population. Córdoba and Santa Fe are also underrepresented in terms of participants, as they have the second and third largest populations but only the fifth and eighth most participants. Overall while participants are not distributed proportionally to Argentina’s population, the courses established by the Fomentar program do capture the skewed distribution of participants well. Note, however, that perhaps one reason why there exists this skewed distribution of participants is because of this distribution of courses. This affect is likely secondary, however, because a vast majority of participants take courses entirely virtually, and thus do not care about the availability of in person courses near them.

5.2 Course Alignment Statistics

Before we use the alignment variable in our regression, we extract information through general descriptive statistics. We first look at the distribution of alignment variable scores in Figure 8.

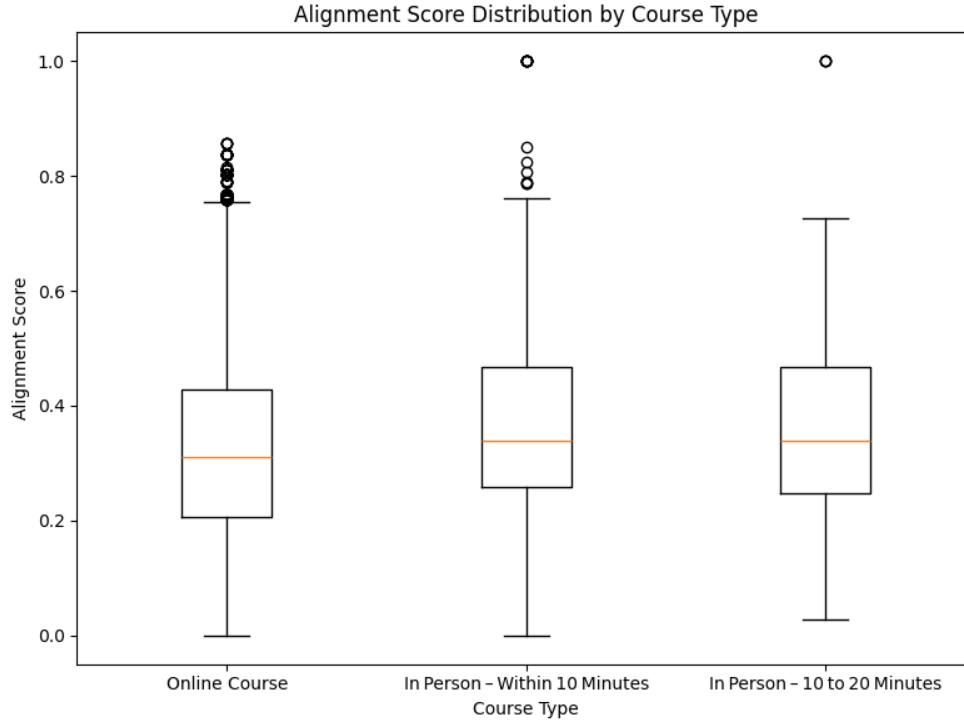


Figure 9: Excluding General Courses

Although the magnitude of the alignment variable itself does not contain much information, we note that the scores are well distributed. Furthermore, we note that there is a low frequency of scores near one. This suggests that not many participants are taking courses that exactly align with their previous labor experience.

We also note how the distribution of the alignment variable changes by course type in Figure 9. We note that the alignment scores shift upward between online courses and in-person courses. This suggests that when participants take a course in-person, they are more likely to take a course that is aligned with their previous occupation. Alternatively, this can be interpreted as, people are only willing to take a course in person if it is aligned with their previous occupation.

5.3 Course Participation Regression

Main factors associated with increased probability of taking a course	
Fomentar	VAT
Lower levels of education	Higer levels of education
Being younger	Being older
Being female	Being female
Not being currently working	Not being currently working
Having no work experience	Having work experience

For Fomentar participants, the strongest predictors of taking an aligned course are having completed tertiary education, being older (45+) and being female. All these variables are associated with a decrease in the probability that someone takes a course.

For VAT, the strongest predictors of taking a course are having completed tertiary education, having completed secondary education and being older (45+). Notably, for VAT participants, all education factors are associated with a decrease in the probability of taking a course. The opposite is true for Fomentar participants. The impact of sex is also smaller. For both Fomentar and VAT participants, having a job currently decreases the probability that they take a course. VAT participant's work experience increases the probability of taking a course. The opposite is true for Fomentar. For both Fomentar and VAT, the impact of SSN contributions, municipality population, and experience (age^2) is relatively small.

Table 1: Course Participation Regression

	Fomentar	VAT
<i>tertiary_completed</i>	-0.614*** (0.010)	0.605*** (0.018)
<i>age > 45</i>	-0.344*** (0.012)	-0.280*** (0.011)
<i>sex</i>	-0.311*** (0.003)	-0.043*** (0.006)

Continued on next page

(Continued from previous page)

	Fomentar	VAT
$25 < age \leq 45$	-0.257*** (0.004)	-0.188*** (0.006)
<i>primary_completed</i>	-0.206 (0.005)	0.149*** (0.012)
<i>currently_working</i>	-0.208*** (0.004)	-0.225*** (0.005)
<i>work_experience</i>	-0.177*** (0.003)	0.251*** (0.006)
<i>secondary_completed</i>	-0.043*** (0.005)	0.452*** (0.012)
<i>social_security_contributions</i>	-0.020*** (0.000)	-0.002** (0.001)
<i>municipality_population</i>	-0.001*** (0.000)	-0.001*** (0.000)
<i>experience (age²)</i>	0.000*** (0.000)	0.000 (0.000)
<i>constant</i>	0.382*** (0.006)	-1.376*** (0.013)
<i>Observations</i>	803 854	438 490
<i>Pseudo R²</i>	0.0484	0.042

5.4 Course Alignment Regression

5.4.1 Discussion

In this regression, the coefficient β_j represents the expected change in alignment score for a 1-unit increase in the explanatory variable, holding all other variables constant. For example, say the alignment score for a person with no work experience is 0.5. If they were to get work experience,

we would expect their alignment score to increase to 0.56.

Main factors associated with increased alignment	
Fomentar	VAT
Having work experience	Having work experience
Being female	Being female
Higher levels of education	Lower levels of education
Not being currently working	Being currently working
Being older	Being older
Taking a course >10 mins away	Taking a course >10 mins away

For Fomentar participants, the factors associated with the largest change in expected alignment are work experience, sex and secondary education. Having work experience, being male, and completing secondary education are associated with an increase in alignment score.

For VAT participants, the factors associated with the largest change in expected alignment are work experience, currently working and secondary education. They are all associated with an increase in alignment score. Notably, our regression shows that the effect of taking a remote course, hybrid course, or online self paced course as opposed to taking an in-person course has no strong significant impact alignment score.

Similar to the regression on course taking, the impact of social security contributions, municipality population and experience are small.

Table 2: Course Alignment Regression

	Fomentar	VAT
<i>work_experience</i>	0.058***	0.057***
<i>sex</i>	-0.045***	-0.034***
<i>secondary_completed</i>	0.036***	0.051***
<i>tertiary_completed</i>	0.033***	0.041***
<i>course_dist</i> > 10 min	0.036**	-0.197***
<i>remote</i>	-0.024	0.057

Continued on next page

(Continued from previous page)

	Fomentar	VAT
<i>primary_completed</i>	−0.021***	0.027**
<i>currently_working</i>	−0.013***	0.057***
<i>hybrid</i>	−0.015	0.012
<i>age > 45</i>	0.015***	0.043***
<i>25 < age ≤ 45</i>	0.014***	0.020***
<i>online_self_paced</i>	−0.006	0.107*
<i>social_security_contributions</i>	0.002***	0.001*
<i>municipality_population</i>	0.000***	0.000**
<i>experience (age²)</i>	−0.000***	−0.000***
<i>constant</i>	0.412	0.337***
<i>Observations</i>	66 670	10 882
<i>Adjusted R²</i>	0.049	0.051

5.4.2 Issues

The results of this regression can be used to make tentative conclusions. They are vulnerable to errors in the NLP methods used to create the alignment variable due to low string length, as well as errors arising from the process of creating isochrones.

Note, the regression setup includes multiple observations corresponding to the same person taking different courses. These observations are assumed to be independent, however, in reality it is possible that the decision to take one course will influence other courses they take. Across Fomentar and VAT, 18.8% of people take multiple courses but 17.3% take multiple courses in the same typology. This means that their alignment scores across multiple courses will be similar and correlated. To account for this, standard errors have been clustered by person ID.

6 Conclusion

6.1 Conclusion

The results from the regression demonstrate that for the sampled populations, the further along an individual is in their education, the more likely they are to remain in the same field. This can be demonstrated in both the Fomentar and VAT groups, as completing tertiary education was a strong predictor of alignment in both cases. This makes sense, as individuals who have invested more time in a field would be less likely to try to pivot to a new one. Another strong predictor of alignment was age, specifically being older than 45. Individuals in this age group are likely of a similar mindset to participants with higher degrees of education. That is to say, the more time an individual has invested in a field, the less likely they are to transition to a new one. The significance of these variables suggests that the Fomentar program is being used for different purposes by each participant. Younger and less educated participants are less likely to take courses, and in the cases that they do, they are more likely to pursue a new field. This information can help structure future initiatives in the program, targeting younger participants for horizontal development, and older participants for vertical development. Lastly, the statistical significance of our regression output lends credibility to Natural Language Processing, and its role in the future of structuring large datasets.

6.2 Next Steps

It should be noted that our group was largely limited by the computing power available to us. For future reference, a larger scale computing center could provide greater accuracy in computing alignment score. Additionally, the results of this study could be used to design a field experiment involving participants. An experimental approach would add further credibility to any causal claims made in this paper.

7 Appendix