Novel innovation indicator for Peruvian universities

Marcelo Gallardo (†) & Juan León (§)

February 13, 2025

(†) **Pontificia Universidad Católica del Perú**
(§) **Grupo de Análisis para el Desarrollo**

marcelo.gallardo@pucp.edu.pe
leon.jjm@pucp.edu.pe

Abstract

This paper proposes an innovation indicator for Peruvian universities, focusing on scientific innovation in fields such as engineering and pure sciences. The indicator is constructed using a selected dataset and confirmatory factor analysis (CFA) to ensure robust measurement, with Tucker-Lewis Index (TLI) and Comparative Fit Index (CFI) used to validate the model fit. K-means clustering is applied to identify innovation clusters among universities. Its validity is examined through standard correlation with university rankings and econometric analysis linking the indicator with wage per hour and overeducation. To address potential sample selection bias, we implement a Heckman two-step correction, incorporating the inverse Mills ratio (IMR) into the wage equation. Additionally, we correct for heteroscedasticity by employing heteroscedasticity-robust standard errors (HC3) and assess model reliability through diagnostic tests such as the Breusch-Pagan and White tests.

JEL Classification: O31, O33, I23, C38, C31, C36, C51

The authors wish to express their gratitude to Luis Randy Loayza Arroyo, Gabriela Alejandra Benites Camacho and Nicolás Alberto Velarde Freundt for their assistance with gathering data references and processing, and also for providing help with the STATA code for the treatment of the ENAHO database. Finally, our appreciation is also directed to Juan Jose Tapia Montenegro for his code allowing web scraping Scopus.

1 Introduction

In the context of higher education in developing countries such as Peru, universities play a crucial role in driving scientific progress and economic growth. The increasing global emphasis on STEM (science, technology, engineering, and mathematics) fields highlights the need for specific indicators to evaluate innovation, as universities excelling in these areas are vital for advancing technology, scientific understanding, and societal development. This paper aims to develop an innovation indicator to provide insights with academic and policy implications for Peru.

In Peru, there is a lack of specific tools to measure university performance in terms of scientific innovation. Existing studies, such as SUNEDU's third biennial report in 2021 SUNEDU (2021), provide general metrics on university quality but fail to focus on innovation in STEM fields. Additionally, Millones-Gómez and et al. (2021) reveals that while research policies in Peruvian universities do not significantly drive scientific production in major databases, factors such as management style and the number of researchers have a notable impact. This underscores the necessity of an innovation indicator tailored to Peruvian universities, focusing on scientific advancement and technological development. To address this gap, we propose a methodology that includes: (1) selecting relevant variables based on theoretical frameworks, previous studies, and the theory of human capital; (2) applying K-means clustering A. Jain and Flynn (1999); Trevor Hastie and Friedman (2009) to identify groups of universities with similar characteristics; and (3) using Confirmatory Factor Analysis (CFA) (Brown, 2015; Mulaik, 2010; Hoyle, 2012; Kline, 2015) to construct the innovation indicator. The robustness of the indicator is evaluated by correlating it with existing rankings, analyzing its relationship with graduates' earnings through regression and performing a logit where the dependent variable is the overeducation.

Our research is limited by data unavailability, such as detailed metrics on the proportion of top-tier PhD faculty members at each university. Future studies should incorporate more comprehensive datasets to enhance the understanding of innovation dynamics in Peruvian higher education. This research contributes to a nuanced understanding of innovation in Peruvian universities, providing a foundation for improved educational policies and practices that can foster national progress in science and technology. While our primary goal is to develop an innovation indicator, we consciously opt not to restrict our analysis to purely scientific publications. This decision reflects the recognition that interdisciplinary collaboration is crucial in the current era of technological advancement, driven by artificial intelligence and machine learning. By incorporating a broader range of publications, we aim to capture the multifaceted nature of innovation that transcends traditional academic boundaries.

The paper begins by detailing the data selection process for the innovation indicator and explains the application of K-means clustering to categorize Peruvian universities. K-means clustering is widely used for partitioning observations into distinct groups based on feature similarity (Trevor Hastie and Friedman, 2009). This methodology allows for an objective classification of universities based on innovation-related characteristics.

Subsequently, we apply Confirmatory Factor Analysis (CFA) to validate the innovation indicator, ensuring that the constructed measure accurately captures the underlying latent construct. CFA is a well-established technique in psychometrics and applied statistics, often used to confirm the factor structure of an indicator by testing the relationships between observed variables and their latent dimensions (Sarmento and Costa, 2019).

Finally, we correlate our findings with existing university rankings and proceed with regression analyses to assess the relationship between innovation and labor market outcomes. Specifically, we estimate the effect of innovation on log hourly wages, controlling for a relevant set of covariates using an econometric specification inspired by standard human capital models. To correct for potential sample selection bias, we employ the Heckman selection model (Heckman, 1979), a two-step procedure that accounts for non-random selection into employment. This methodology is widely used in labor economics and econometric applications where omitted selection mechanisms can lead to biased estimates (Vella, 1998).

For the overeducation analysis, we estimate a logit model (McFadden, 1974), a standard binary choice framework in discrete choice modeling. This approach allows us to analyze the probability of an individual being overeducated, incorporating relevant individual and job-related characteristics. The logit model is appropriate due to its ability to handle dichotomous dependent variables and has been extensively used in labor market research.

To ensure the robustness of our results, we conduct diagnostic tests on our regression models. We assess normality using the Jarque-Bera test (Jarque and Bera, 1987), check for serial correlation with the Durbin-Watson statistic (Durbin and Watson, 1950), and evaluate the presence of misspecification errors through the Wald, Likelihood Ratio, and Lagrange Multiplier tests (Engle, 1984).

2 Data

We considered 90 universities in Peru, which provided the data relevant to our study. The list of these universities and their respective information is available at GitHub. We recorded the publications of each university up to 2020 using Scopus. Additionally, the student ratios in courses across various fields - Agriculture, Architecture, Biology, Earth Sciences, Civil Engineering, Environmental Studies, Electrical and Energy Engineering, Electronics and Automation, Statistics, Pharmacy, Physics, Food Industries, Mathematics, Sanitary Engineering, Mechanical and Metallurgical Engineering, Medicine, Obstetrics, Dentistry, Fisheries Engineering, Textile Production, Mining, Chemistry, Forestry, Telecommunications, Motorized Vehicles, Ships and Aircraft, and Veterinary Medicine—were obtained through the Tuni portal. Furthermore, the number of patent applications per university in the list was sourced from the WIPO PatentScope. Finally, through manual record-keeping, we gathered data on the number of publications in Q1 and Q2 journals and a variable assessing the existence of graduate studies at each university. For this variable, master's studies were assigned a value of 0.5,

doctoral studies a value of 1, and the variable was set to 0 otherwise.

Hereafter, we present key statistics to provide a thorough understanding of the characteristics of our dataset. These include measures of central tendency (mean) and dispersion (standard deviation), which reveal general patterns and variability in the data. Such statistical measures are foundational for understanding the underlying characteristics of the variables. Additionally, normality tests using the Shapiro-Wilk method were conducted to assess the distributional properties of each variable.

Let us define the variables as follows[1]:

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \\ X_7 \end{pmatrix} = \begin{pmatrix} \text{Publications} \\ \text{Proportion of students in science fields} \\ \text{Application patents} \\ \text{Q1} \\ \text{Q2} \\ \text{Graduate indicator} \\ \text{Accreditation} \end{pmatrix}. \tag{1}$$

Note that X_7 is a binary variable, i.e., $X_7 \in \{0, 1\}$, and $X_6 \in \{0, 0.5, 1\}$, where 0 represents the absence of graduate programs, 0.5 indicates the presence of master's programs only, and 1 corresponds to the presence of doctoral programs.

In the context of a CFAnalysis aimed at determining an innovation indicator, each variable included plays a crucial role in capturing different dimensions of innovation capacity. *Publications* measure the academic and scientific output, serving as a proxy for the institution's research productivity and its contribution to knowledge creation. The *proportion of students in science fields* emphasizes the focus on STEM disciplines, which are essential for developing technical skills and fostering innovative thinking. *Application patents* provide a direct indicator of the translation of research into practical and marketable technologies, reflecting the institution's ability to produce tangible outcomes. The variables *Q1* and *Q2* measure the quality of the research produced, with *Q1* capturing publications in top-tier journals and *Q2* highlighting significant contributions in second-tier journals, ensuring a broad perspective on research impact. The *graduate program indicator* evaluates the institution's ability to offer advanced education (e.g., master's and doctoral programs), which is key for training future leaders in innovation. Finally, *accreditation* serves as a measure of institutional quality, ensuring that the university adheres to high academic and research standards. Together, these variables create a robust framework to assess and quantify the innovation potential of an institution comprehensively.

All variables in (1) were normalized:

$$x \to \frac{x - x_{\min}}{x_{\max} - x_{\min}} \in [0, 1].$$

The descriptive statistics for these variables are presented below:

[1] For accreditation, we considered the information delivered here.

	X_1	X_2	X_3	X_4	X_5	X_6	X_7
count	90.000000	90.000000	90.000000	90.000000	90.000000	90.000000	90.000000
mean	0.058286	0.428851	0.045029	0.036826	0.054739	0.633333	0.066667
std	0.159127	0.217739	0.137704	0.127766	0.162435	0.415973	0.250841
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.003287	0.308982	0.000000	0.000836	0.001205	0.500000	0.000000
50%	0.011191	0.439022	0.000000	0.003135	0.006627	1.000000	0.000000
75%	0.045625	0.555160	0.015038	0.015468	0.032831	1.000000	0.000000
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Normality tests are essential for determining whether data are distributed in a manner consistent with a normal distribution, which is a critical assumption for many statistical analyses. The Shapiro-Wilk test, introduced in the seminal paper by Shapiro and Wilk (1965), is a widely used method for assessing normality. The null hypothesis of the test posits that the data follow a normal distribution. If is too small and the -value falls below the significance level (commonly), the null hypothesis is rejected, indicating non-normality.

The Shapiro-Wilk test results for our dataset are summarized in Table 2. Based on the Shapiro-Wilk test results and considering a 95% confidence level, we conclude that only the second variable, namely the proportion of students in science fields, follows a normal distribution. This is evident from its p-value of 0.167, which is greater than the significance level of 0.05. For all other variables, the p-values are significantly below 0.05, leading to the rejection of the null hypothesis of normality. These findings highlight the importance of using non-parametric or distribution-agnostic methods for subsequent analyses involving the non-normal variables. Figures 1-6) complete the anaklysis of normality form the graphical approach.

Variable	Statistic	p-value
Publications	3.670970e-01	7.179654e-18
Science students ratio	9.795203e-01	1.672049e-01
Patent applications	3.652966e-01	6.800968e-18
Q1	2.949021e-01	8.898616e-19
Q2	3.502554e-01	4.343951e-18
Graduate program indicator	7.420062e-01	2.819779e-11
Acreditacion	2.695735e-01	4.444075e-19

3 Clustering analysis

Normalization is especially critical in unsupervised learning methods like K-means, as it prevents variables with larger ranges from dominating the clustering process. By limiting the analysis to 2 or 3 variables at a time, we also achieve clearer graphical representations, which are essential for understanding the underlying data structure. We perform K-means clustering on the dataset $\mathbf{X}$ and visualize the results for selected subsets. These subsets, denoted as $S_j \subset \mathbf{X}$ for $j = 1, 2, 3$, were chosen to explore specific combinations of the data dimensions. The subsets are defined as follows:

1. $S_1 = \{\text{Publications}, \text{Science students ratio}\}$,
2. $S_2 = \{\text{Publications}, \text{Science students ratio}, \text{Patent applications}\}$,
3. $S_3 = \{\text{Publications}, \text{Science students ratio}, \text{Graduate studies indicator}\}$.

The K-means algorithm, a foundational method in unsupervised learning, partitions the data into K clusters by minimizing intra-cluster variance. Figures 8, 9, 10, and 11 illustrate the results of this clustering for the selected subsets. These visualizations reveal the intrinsic structure of the data and highlight relationships between the variables. The elbow method was used to determine an optimal number of clusters, balancing the trade-off between model simplicity and explanatory power. This analysis provides valuable insights for evaluating the impact of innovation on university choice using a model with a non-continuous dependent variable. Moreover, it enables the identification of distinct groups within the dataset. Specifically, we observe that three universities consistently stand out in the subset analysis due to their high number of publications. However, raw publication counts must be interpreted carefully, as they do not account for the size and structure of the university. Moreover, it is important to emphasize once again that the indicator measures innovation in the context of sciences and engineering, and does not specifically account for the quality of education in areas such as the social sciences. For instance, Universidad del Pacífico is a university that only offers professional programs (administration, law, economics, finance...) and is recognized for its academic prestige (e.g., QS Ranking 2024).

4 Confirmatory Factor Analysis

Let us recall that Confirmatory Factor Analysis (CFA) is a statistical technique used to verify the factor structure of a set of observed variables. The CFA model is based on the assumption that each observed variable is directly influenced by certain factors, which are unobservable and represented as latent variables. In the context of constructing an innovation indicator, CFA can be employed to evaluate the extent to which a set of observed variables—such as the number of publications, patent applications, the number of publications in $Q1$ and $Q2$ journals, and our «graduate program indicator»—captures the underlying construct of «scientific innovation» within universities.

Formally, the CFA model is expressed as:

$$\mathbf{Y} = \Lambda\xi + \delta,$$

where:

1. $\mathbf{Y}$ represents the vector of observed variables,
2. ξ denotes the latent factor (e.g., scientific innovation),
3. Λ is the matrix (or vector) of loadings, which quantifies the strength of the relationship between the observed variables and the latent factor,
4. δ captures the errors or uniquenesses associated with the observed variables, representing the variance not explained by the latent factor.

This formulation provides a robust framework for modeling and validating the relationships between measurable indicators and the latent construct of scientific innovation, allowing for both empirical validation and theoretical interpretation.

The factor loadings in Λ are critical as they show how strongly each observed variable is associated with the latent factor. The goal of CFA is to estimate these loadings and assess how well the hypothesized factor structure fits the data. Using these loadings, we compute our innovation indicator, which is a weighted sum of the observable variables.

Before exposing our results, we shall present some key statistic concerning the validity of employing CFA. Barlett test (Bartlett, 1950) is used to test the null hypothesis that variables in a dataset are uncorrelated and therefore not suitable for factor analysis (Sarmento and Costa, 2019). Mathematically, the test checks whether the observed correlation matrix significantly differs from an identity matrix (a matrix where all correlations are zero). For this, it is used the statistic

$$\chi^2 = -\left[(n-1) - \frac{2p+5}{6}\right] \ln|(\det\mathbf{R})|$$

where n is the number of observations, p the number of variables and $\mathbf{R}$ the correlation matrix. A high chi-square value suggests that the correlations among variables are strong enough and not a random even. A very low p-value leads to rejecting this null hypothesis, indicating that the variables are, in fact, correlated. We obtained

$$\text{Chi-square value} = 655.8708438531937,$$
$$p\text{-value} = 2.740532093846338e-125.$$

We complement our analysis of the Bartlett test with the Kaiser-Meyer-Olkin (KMO) test (Kaiser, 1974): The KMO test measures the suitability of data for factor analysis. It evaluates the magnitude of partial correlations between variables, with higher values (closer to 1) being indicative of greater

suitability for analysis. The KMO statistic is given by:

$$\text{KMO} = \frac{\sum\sum_{j\neq k} r_{jk}^2}{\sum\sum_{j\neq k} r_{jk}^2 + \sum\sum_{j\neq k} \psi_{jk}^2}$$

where r_{jk} are the coefficients of correlation between variables, and ψ_{jk} are the partial correlation coefficients. A KMO value greater than 0.6 generally confirms the suitability for factor analysis. More than 0.8 is very good and near 0.7 is good. In our case,

$$\text{KMO Model} = 0.6853506639636759.$$

Hence, combing this with our $p-$value of the Barlett test, we support out methodology (though not marvelous in the words of Kaiser, good enough). Hereafter, our results of the CFA using semopy package in Python, using as base variable X_4.[2]

lval	op	rval	Estimate	Std. Err	z-value	p-value
X4		innovacion	1.000000	-	-	-
X1		innovacion	1.303167	0.050400	25.856750	0.000000
X2		innovacion	0.189018	0.188812	1.001093	0.316782
X3		innovacion	0.535539	0.107611	4.976630	0.000001
X5		innovacion	1.342833	0.048552	27.657448	0.000000
X6		innovacion	0.848092	0.352512	2.405853	0.016135
X7		innovacion	0.983800	0.195648	5.028425	0.000000
innovacion		innovacion	0.014469	0.002394	6.042964	0.000000
X1		X1	0.000466	0.000136	3.440766	0.000580
X2		X2	0.046363	0.006911	6.708204	0.000000
X3		X3	0.014600	0.002176	6.707931	0.000000
X4		X4	0.001673	0.000259	6.468706	0.000000
X5		X5	0.000000	0.000124	0.000000	1.000000
X6		X6	0.160614	0.023943	6.708190	0.000000
X7		X7	0.048226	0.007189	6.707919	0.000000

In the context of Structural Equation Modeling (SEM) and Confirmatory Factor Analysis (CFA), the Maximum Likelihood Estimation (MLE) method is commonly used to estimate the model parameters. The likelihood function $L(\Lambda, \Phi, \Theta)$ in this context is defined as in Joreskog (1969) or Kaplan (2000),

$$L(\Lambda, \Phi, \Theta) = -\frac{n}{2}\left[\log|\Sigma(\Lambda, \Phi, \Theta)| + \text{tr}(S\Sigma^{-1}(\Lambda, \Phi, \Theta)) - \log|S| - p\right] \tag{2}$$

where Λ represents the matrix of factor loadings, Φ is the covariance matrix of the latent factors, Θ denotes the unique variances of the observed variables, $\Sigma(\Lambda, \Phi, \Theta)$ is the model-implied covariance

[2] Note that only X_2 has low significance.

matrix, S is the observed covariance matrix, n is the number of observations, p is the number of observed variables, log denotes the natural logarithm and $|\cdot|$ denotes the determinant of a matrix.

We now focus on analyzing key indices in Structural Equation Modeling (SEM) statistics: the Tucker-Lewis Index (TLI), the Comparative Fit Index (CFI), and the Root Mean Square Error of Approximation (RMSEA). The TLI is defined as

$$TLI = 1 - \frac{\chi^2_{\text{model}} - \text{df}_{\text{model}}}{\chi^2_{\text{baseline}} - \text{df}_{\text{baseline}}},$$

where χ^2_{model} and df_{model} represent the chi-square value and degrees of freedom for the model, while χ^2_{baseline} and $\text{df}_{\text{baseline}}$ are those for a baseline (usually null) model. Similarly, the CFI assesses the relative improvement in fit of the model compared to the baseline. The RMSEA is defined as

$$\text{RMSEA} = \sqrt{\frac{\chi^2_{\text{model}} - \text{df}_{\text{model}}}{\text{df}_{\text{model}}(n-1)}},$$

where n is the sample size, and it evaluates the model's error of approximation in the population.

In interpreting these indices for our model, a TLI of 0.8671428197413257 and a CFI of 0.911428546 4942171 indicate a good fit, as values above 0.9 are generally considered acceptable and 0.85 good enough following for Digital Research and Education (nd). However, the RMSEA of 0.21769883488039693, exceeding the typical threshold of 0.1, highlights potential challenges with the model fit. This discrepancy may result from factors such as model complexity or sample size. Nonetheless, the combination of indices supports the overall robustness of the model, particularly given the primary objective of constructing an innovation indicator rather than deriving conclusions solely from the model itself.

With the groundwork established, we proceed to construct the innovation indicator using the loadings from the CFA as weights. For each university j, the innovation indicator ζ_j is defined as:

$$\zeta_j = \Lambda_{\text{significant}} \mathbf{X}^j_{\text{significant and min-max normalized}}.$$

The results are presented in Appendix C. Finally, we compute the correlation between the innovation indicator and an existing quality ranking SUNEDU (2021). The calculated correlation is 0.8495538574066844, demonstrating a strong positive relationship. Detailed results can be found in Appendix C.1.

5 Effects over wages

We use data from the *Encuesta Nacional de Hogares* (ENAHO) to analyze individual characteristics and their relationship with wages. The key variables of interest include `estado_civil` (marital status), `lengua_materna` (native language), `modalidad` (study modality), `ocupinf` (informal occupation), and `ocupacion` (occupation), as well as annual and hourly wages. Each individual is linked to their respective university, from which we link the innovation indicator.

Our primary objective is to evaluate whether the innovation indicator impacts wages, specifically the logarithm of hourly wages. To address potential sample selection bias, we employ the Heckman selection model Heckman (1979). The selection equation is specified as:

$$\mathbb{P}(\text{works}_i = 1|\Upsilon_i) = \Phi(\beta \cdot \Upsilon_i + \varepsilon_i). \tag{3}$$

where,

$$\Upsilon_i = \begin{pmatrix} \text{age}_i & \text{sex}_i & \text{has\_children}_i & \text{education}_i & \text{native\_language}_i & \text{education}_i^2 \end{pmatrix}$$

and

- age_i: Continuous variable representing the age of individual i.
- sex_i: Binary variable equal to 1 if the individual is female and 0 otherwise.
- has_children_i: Binary variable equal to 1 if the individual has children and 0 otherwise.
- education_i: Categorical variable taking the value 5 if the highest level attained is undergraduate and 8 if the individual has completed a postgraduate degree (Master's or PhD).
- native_language_i: Binary variable equal to 1 if the individual's native language is Spanish and 0 otherwise.
- education_i^2: The squared term of the education variable to capture potential nonlinear effects.

In (3), the dependent variable works_i is a binary indicator, equal to 1 if the individual is employed and 0 otherwise. The function $\Phi(\cdot)$ represents the cumulative distribution function of the standard normal distribution. Then, the wage equation, corrected for selection bias, is given by:

$$\begin{aligned} \log(\text{wage}_i) = \gamma_0 &+ \gamma_1 \text{innovation}_i + \gamma_2 \text{sex}_i + \gamma_3 \text{marital\_status}_i \\ &+ \gamma_4 \text{education}_i + \gamma_5 \text{informal\_occupation}_i + \gamma_6 \text{age}_i + \gamma_7 \text{native\_language}_i + \gamma_8 \lambda_i + \eta_i, \end{aligned}$$

where:

- innovation_i: A continuous variable representing the innovation indicator assigned to the individual's university.
- marital_status_i: Binary variable equal to 1 if the individual is married and 0 otherwise.
- $\text{informal\_occupation}_i$: Binary variable equal to 1 if the individual holds an informal job and 0 otherwise.
- λ_i is the inverse Mills ratio (IMR) obtained from the first stage probit regression, correcting for selection bias.

- $\gamma_8 = \rho\sigma_u$, where ρ represents the correlation between the error terms of the selection and wage equations, and σ_u is the standard deviation of the error term in the wage equation.

We verify the presence of multicollinearity in the wage regression by analyzing the Variance Inflation Factor (VIF). Furthermore, we assess heteroskedasticity using the Breusch-Pagan test. To obtain robust standard errors, we employ the HC3 estimator:

$$\hat{V}(\hat{\beta}) = (X^T X)^{-1} X^T \hat{H} X (X^T X)^{-1}, \tag{4}$$

where $\hat{H}$ is a diagonal matrix with elements $h_{ii} = X_i (X^T X)^{-1} X_i^T = e_i^2/(1-h_{ii})^2$, accounting for leverage effects.

The first-stage probit regression (Table 1) shows that age, sex, marital status, and native language are statistically significant in predicting employment. Notably, being a woman (sex = 1) reduces the probability of employment by approximately 0.24, and speaking Spanish (native language = 1) increases employment likelihood.

Table 1: Probit Regression Results

Variable	Coefficient	Std. Error	z-value	p-value
Constant	0.6052	0.062	9.792	0.000
Age	-0.0229	0.001	-40.641	0.000
Sex (1=Female)	-0.2399	0.016	-14.642	0.000
Has Children	0.0100	0.017	0.573	0.566
Education	0.2299	0.008	29.261	0.000
Marital Status	0.2029	0.019	10.738	0.000
Native Language (1=Spanish)	-0.1520	0.037	-4.120	0.000

In the second-stage wage regression (Table 2), we find that the innovation indicator has a positive and significant effect on log hourly wages. Specifically, an increase in the innovation index is associated with a 8.67% increase in wages. The inverse Mills ratio (IMR) is negative and statistically significant, indicating selection bias in the original sample.

The Durbin-Watson statistic (1.698) suggests weak autocorrelation. The Jarque-Bera test Jarque and Bera (1987) rejects normality of residuals (p-value = 0.00), and HC3 standard errors suggest some degree of heteroskedasticity, which the robust estimation corrects. However, HC3 does not fully eliminate heteroskedasticity, possibly due to misspecification or influential observations.

Overall, our results indicate that innovation positively impacts wages, with additional controls confirming robustness. Future work could refine the model by incorporating instrumental variables Angrist and Krueger (2001) or alternative selection models Vella (1998).

In our statistical analysis, we found a Pearson correlation coefficient of 0.1572 between innovation and log hourly wages, with a p-value of 0.0000. This indicates a modest but statistically significant

Table 2: OLS Regression Results for Log Hourly Wage

Variable	Coefficient	Std. Error	t-value	p-value
Constant	1.4967	0.157	9.538	0.000
Innovation	0.0867	0.005	18.372	0.000
Sex (1=Female)	-0.0599	0.015	-4.001	0.000
Marital Status	0.1049	0.014	7.276	0.000
Education years	0.0805	0.011	7.100	0.000
Informal Occupation	0.4407	0.012	37.879	0.000
Age	0.0114	0.001	8.585	0.000
Native Language (1=Spanish)	0.0526	0.018	2.856	0.004
IMR	-0.8286	0.405	-2.044	0.041

positive relationship. The significance suggests that higher innovation levels are associated with higher wages, but the modest magnitude implies that innovation alone is not a strong determinant of earnings. Other factors such as industry, experience, and job type likely play a more crucial role in wage formation.

While innovation may be linked to *higher education and cognitive ability*, its direct impact on wages is not immediate. This suggests that:

- Innovation is a long-term investment that does not always translate into short-term earnings.
- Labor market frictions may prevent innovative individuals from capturing their full earning potential immediately.
- Higher education might create a pathway to higher wages, but the effect of innovation alone is limited.

6 Overeducation Analysis

In this section, we estimate a logit model to assess the impact of innovation on overeducation. The dependent variable is defined as follows:

$$\text{overeducation}_i = \begin{cases} 1, & \text{if years of studies}_i > \overline{\text{years of studies required for } i\text{'s job}} \\ 0, & \text{otherwise.} \end{cases}$$

The probability of being overeducated follows the logistic function:

$$\mathbb{P}(\text{overeducation}_i = 1) = \frac{e^{X_i\beta}}{1 + e^{X_i\beta}},$$

where X_i represents the vector of explanatory variables and β is the vector of coefficients. The estimation results for the logit model are presented in Table 3.

Variable	Coefficient	Std. Error	p-value
Constant	-2.6602	0.095	0.000
Age	0.0320	0.001	0.000
Female (Sex = 1)	0.1007	0.033	0.002
Has children	-0.1999	0.034	0.000
Marital status	0.4091	0.036	0.000
Native language	-0.0884	0.068	0.195
Innovation indicator	0.0403	0.018	0.026

Table 3: Logit Model: Predicting Overeducation

The results indicate that age, being female, and marital status are positively associated with a higher probability of being overeducated, while having children is negatively associated. The coefficient for the innovation variable is positive and statistically significant at the 5% level, suggesting that individuals in more innovative environments are slightly more likely to be overeducated. However, the magnitude of this effect is small. To interpret the results in probability terms, we compute the marginal effects:

$$ME_j = \frac{\partial \mathbb{P}(\text{overeducation}_i = 1)}{\partial X_{ij}} = \beta_j \mathbb{P}(\text{overeducation}_i = 1)(1 - \mathbb{P}(\text{overeducation}_i = 1)).$$

The estimated marginal effects are presented in Table 4. The results confirm that age, gender, and

Variable	Marginal Effect	Std. Error	p-value
Age	0.0053	0.000	0.000
Female	0.0167	0.006	0.002
Has children	-0.0332	0.006	0.000
Marital status	0.0679	0.006	0.000
Native language	-0.0147	0.011	0.195
Innovation indicator	0.0067	0.003	0.026

Table 4: Marginal Effects on Overeducation.

marital status have non-negligible effects on the probability of being overeducated. The marginal effect of innovation is 0.0067, meaning that a one-unit increase in the innovation variable raises the probability of being overeducated by approximately 0.67 percentage points, holding other factors constant.

To ensure robustness, we check for multicollinearity using the Variance Inflation Factor (VIF).

Table 5 shows that all VIF values are well below the common threshold of 10, indicating no serious multicollinearity issues.

Variable	VIF
Age	1.26
Female	1.02
Has children	1.01
Marital status	1.25
Native language	1.02
Innovation	1.02

Table 5: Variance Inflation Factor (VIF).

To validate our findings, we also estimate a probit model. The results, shown in Table 6, are consistent with those of the logit model. The coefficient for innovation remains positive and statistically significant.

Variable	Coefficient	Std. Error	p-value
Constant	-1.6119	0.056	0.000
Age	0.0194	0.001	0.000
Female	0.0620	0.019	0.001
Has children	-0.1192	0.020	0.000
Marital status	0.2376	0.021	0.000
Native language	-0.0510	0.040	0.205
Innovation	0.0240	0.011	0.025

Table 6: Probit Model: Predicting Overeducation.

The findings suggest that higher innovation levels are associated with an increased probability of overeducation. However, the effect size remains relatively small, implying that other labor market frictions and individual characteristics play a more significant role in determining overeducation. The stability of the results across different model specifications further supports their robustness.

7 Conclusion

In this paper, we constructed an innovation indicator for Peruvian universities using Confirmatory Factor Analysis (CFA). Our proposed indicator strongly correlates with an existing university quality metric, validating its relevance in assessing scientific and technological innovation. The construction

of this index is based on key variables, particularly the number of indexed publications in high-impact journals, reflecting the research productivity of institutions.

We also find that our innovation index positively correlates with wages and remains statistically significant in a wage regression that accounts for selection bias. While the estimated impact is not large, it is still meaningful. The relatively modest effect size may be attributed to labor market frictions or stronger selection bias in higher education outcomes. It is important to emphasize that our analysis does not establish causality, and results should be interpreted with this limitation in mind.

Regarding overeducation, we employed a standard yet simple measure and found that higher innovation levels are associated with greater overeducation. This result may stem from the fact that more innovative universities—likely of higher quality—produce a larger number of postgraduates (Master's and PhDs). Given the characteristics of the Peruvian labor market, which features heterogeneity and structural constraints, these highly educated individuals may end up in positions where they are overqualified.

A more robust analysis would require refining the dataset and improving the innovation index. Future research should consider segmented regressions by population groups, a more thorough endogeneity analysis, and enhanced robustness checks (e.g., Newey-West standard errors, M-estimation, see Newey and West (1987), Huber (1964)). Further work could also incorporate alternative identification strategies to strengthen causal inference.

A Data set variable distribution

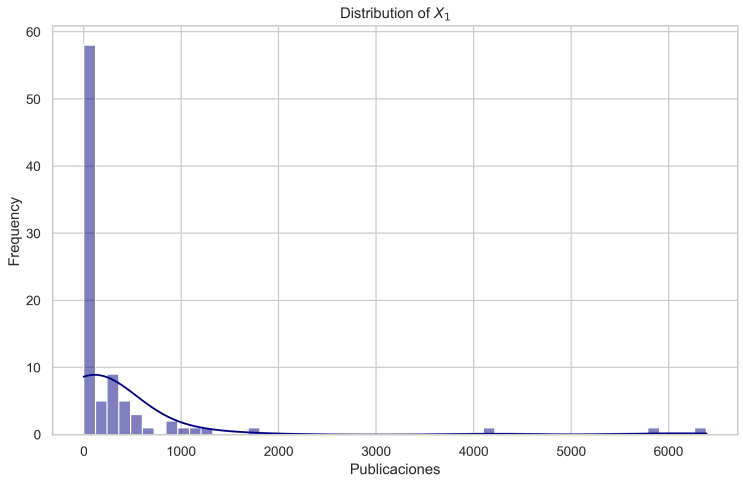

Figure 1: Distribution of X_1.

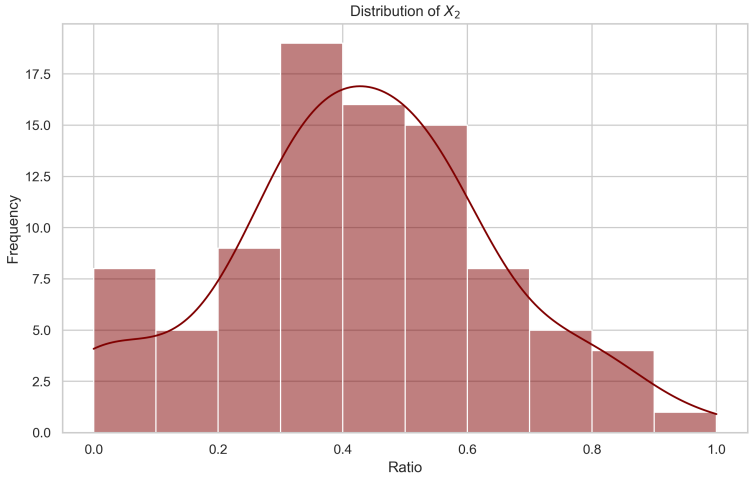

Figure 2: Distribution of X_2.

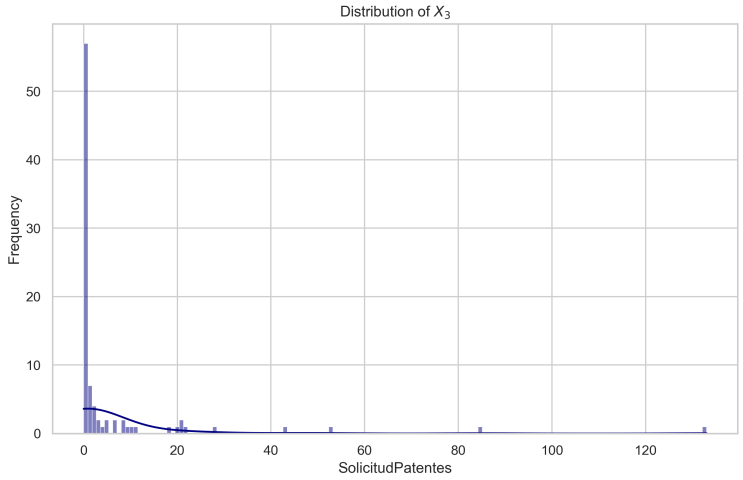

Figure 3: Distribution of X_3.

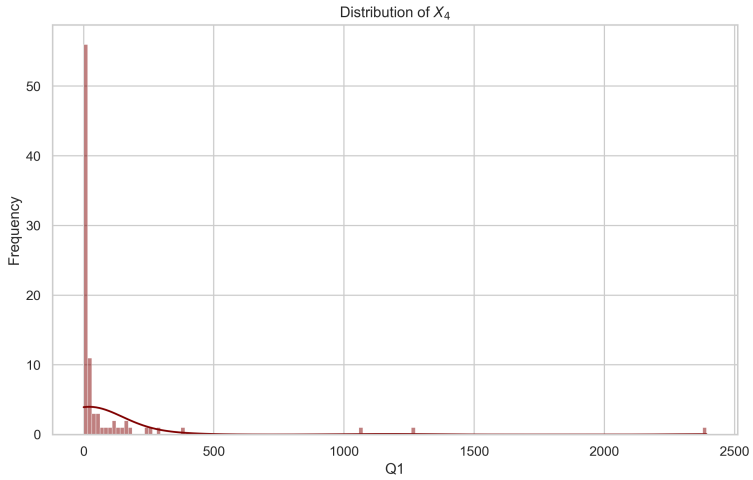

Figure 4: Distribution of X_4.

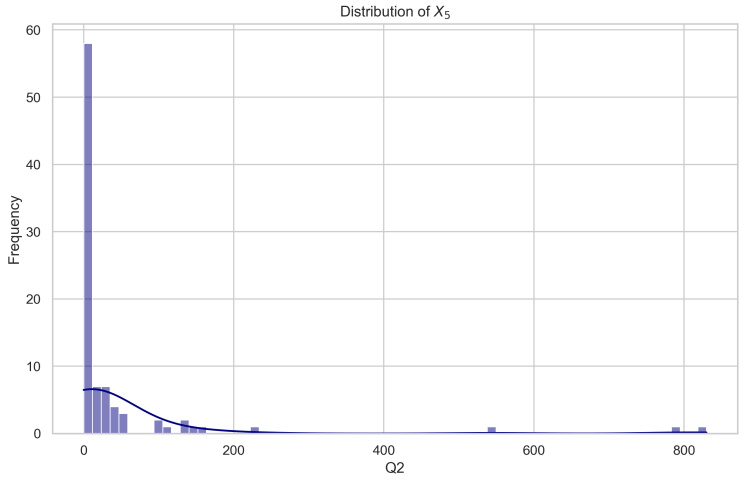

Figure 5: Distribution of X_5.

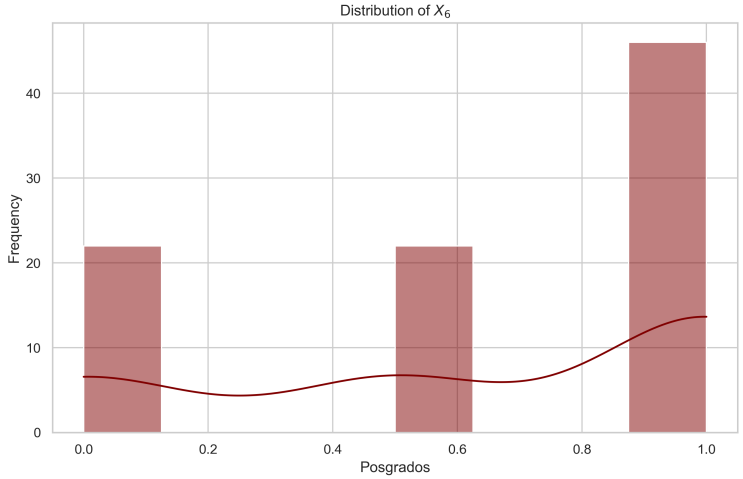

Figure 6: Distribution of X_6.

B K-means

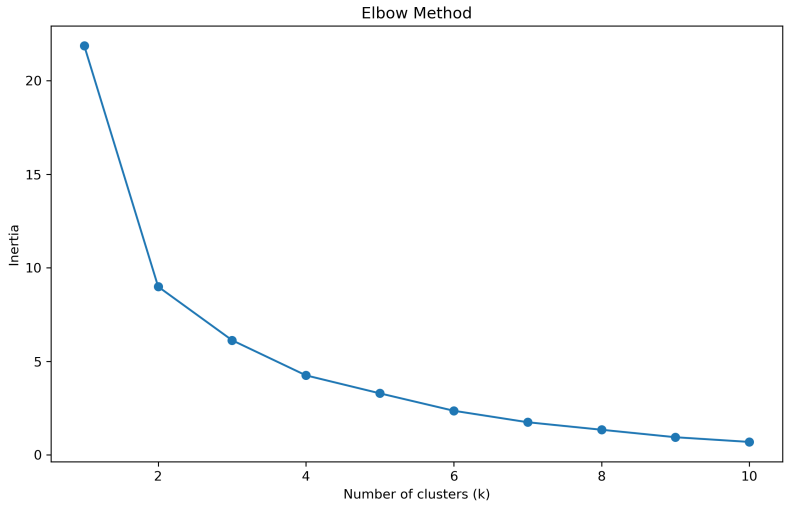

Figure 7: Elbow method.

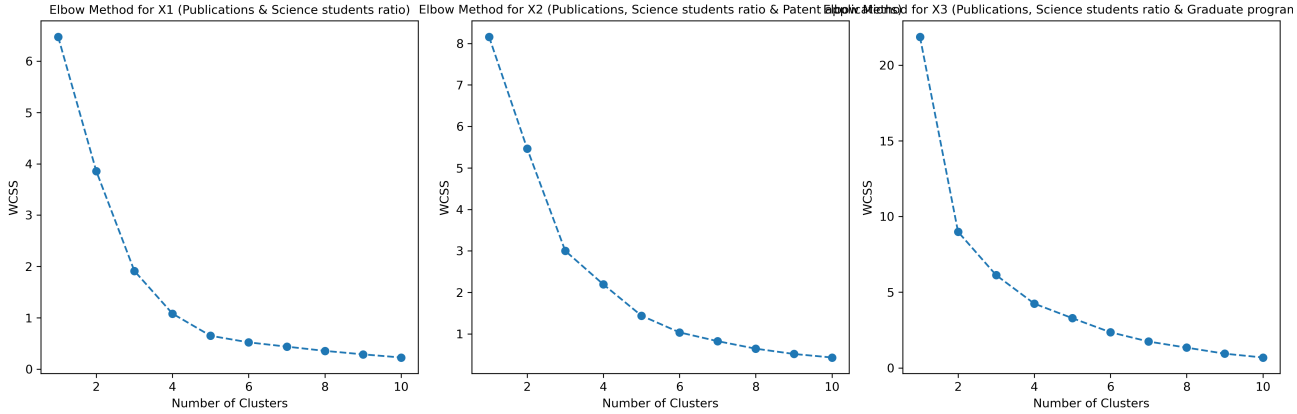

Figure 8: Elbow method for the subsets S_j.

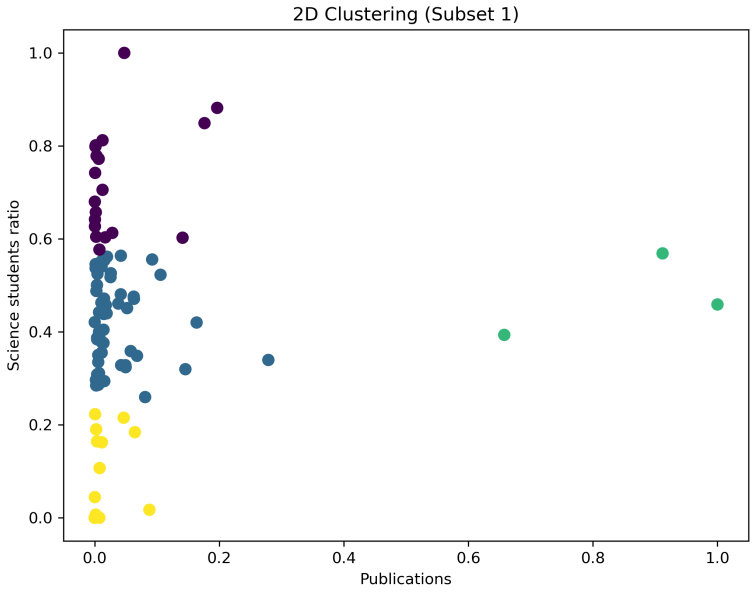

Figure 9: Cluster for S_1.

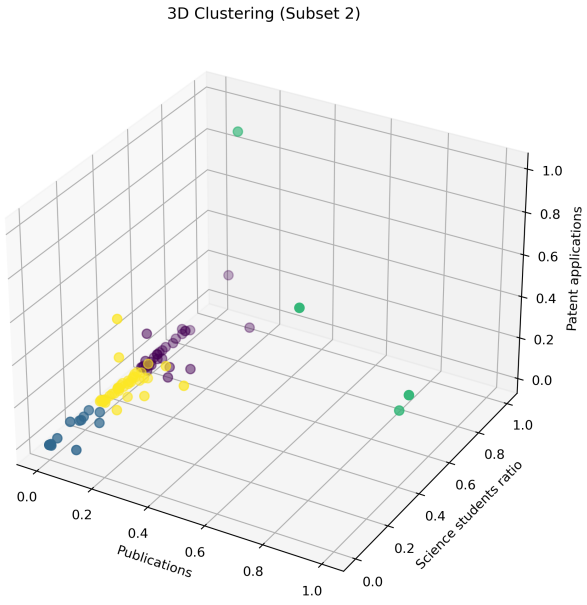

Figure 10: Cluster for S_2.

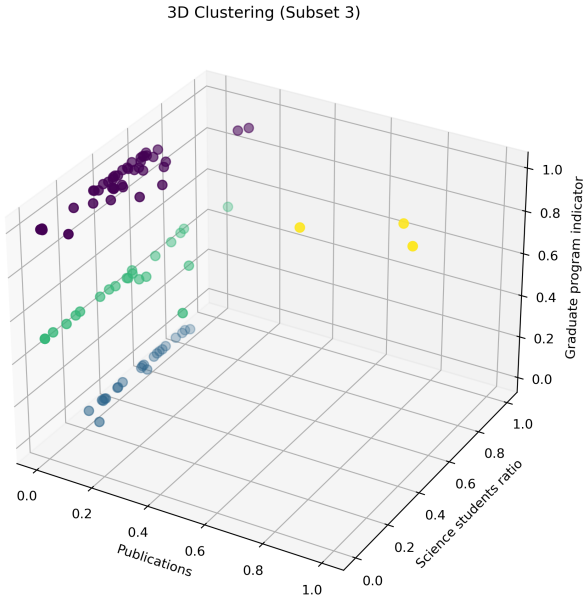

Figure 11: Cluster for S_3.

B.1 Cluster and university

G	X_1	X_2	X_3	X_4	X_5	X_6
0	0.856524	0.473623	0.365915	0.658166	0.870281	1.000000e+00
1	0.042187	0.390074	0.055779	0.023868	0.040347	1.000000e+00
2	0.032307	0.414475	0.024607	0.012561	0.023987	5.000000e-01
3	0.006880	0.512916	0.000684	0.001691	0.002410	2.220446e-16

Table 7: Centroids per cluster $G \in \{0, 1, 2, 3\}$.

University	Cluster
Asociación Civil Universidad de Ciencias y Huma...	3
Facultad de Teología Pontificia y Civil de Lima	1
Pontificia Universidad Católica del Perú	0
Universidad Andina del Cusco	1
Universidad Antonio Ruiz de Montoya	2
Universidad Autónoma de Ica S.A.C.	3
Universidad Autónoma del Perú S.A.C.	1
Universidad Católica San Pablo	2
Universidad Católica Santo Toribio de Mogrovejo	1
Universidad Católica Sedes Sapientiae	2
Universidad Católica de Santa María	1
Universidad Católica de Trujillo Benedicto XVI	2
Universidad Científica del Sur S.A.C.	2
Universidad Continental S.A.C.	2
Universidad César Vallejo S.A.C.	1
Universidad ESAN	1
Universidad Femenina del Sagrado Corazón	1
Universidad Jaime Bausate y Meza	2
Universidad La Salle	2
Universidad Le Cordon Bleu S.A.C.	2
Universidad Marcelino Champagnat	1
Universidad María Auxiliadora S.A.C.	2
Universidad Nacional Agraria La Molina	1

Continued on next page

University	Cluster
Universidad Nacional Agraria de la Selva	2
Universidad Nacional Amazónica de Madre de Dios	3
Universidad Nacional Autónoma Altoandina de Tarma	3
Universidad Nacional Autónoma de Alto Amazonas	3
Universidad Nacional Autónoma de Chota	3
Universidad Nacional Autónoma de Huanta	3
Universidad Nacional Autónoma de Tayacaja Danie...	3
Universidad Nacional Daniel Alcides Carrión	1
Universidad Nacional Federico Villarreal	1
Universidad Nacional Hermilio Valdizán de Huánuco	1
Universidad Nacional Intercultural Fabiola Sala...	3
Universidad Nacional Intercultural de Quillabamba	3
Universidad Nacional Intercultural de la Amazonía	3
Universidad Nacional Intercultural de la Selva ...	3
Universidad Nacional Jorge Basadre Grohmann	1
Universidad Nacional José Faustino Sánchez Carrión	1
Universidad Nacional José María Arguedas	3
Universidad Nacional Mayor de San Marcos	0
Universidad Nacional Micaela Bastidas de Apurímac	2
Universidad Nacional San Luis Gonzaga	3
Universidad Nacional Santiago Antúnez de Mayolo	1
Universidad Nacional Tecnológica de Lima Sur	3
Universidad Nacional Toribio Rodríguez de Mendo...	1
Universidad Nacional de Barranca	3
Universidad Nacional de Cajamarca	1
Universidad Nacional de Cañete	3
Universidad Nacional de Educación Enrique Guzmá...	1
Universidad Nacional de Frontera	3
Universidad Nacional de Huancavelica	1
Universidad Nacional de Ingeniería	1
Universidad Nacional de Jaén	2
Universidad Nacional de Juliaca	3
Universidad Nacional de Moquegua	2
Universidad Nacional de Piura	1

Continued on next page

University	Cluster
Universidad Nacional de San Agustín de Arequipa	1
Universidad Nacional de San Antonio Abad del Cusco	1
Universidad Nacional de San Martín	1
Universidad Nacional de Trujillo	1
Universidad Nacional de Tumbes	1
Universidad Nacional de Ucayali	1
Universidad Nacional del Altiplano	1
Universidad Nacional del Callao	1
Universidad Nacional del Centro del Perú	1
Universidad Nacional del Santa	1
Universidad Peruana Cayetano Heredia	0
Universidad Peruana Los Andes	3
Universidad Peruana Unión	1
Universidad Peruana de Ciencias Aplicadas S.A.C.	2
Universidad Privada Antenor Orrego	1
Universidad Privada Norbert Wiener S.A.	2
Universidad Privada San Juan Bautista S.A.C.	2
Universidad Privada de Huancayo Franklin Roosev...	3
Universidad Privada de Tacna	1
Universidad Privada del Norte S.A.C.	1
Universidad Ricardo Palma	1
Universidad San Ignacio de Loyola S.R.L.	1
Universidad Señor de Sipán S.A.C.	2
Universidad Tecnológica de los Andes	2
Universidad Tecnológica del Perú S.A.C.	2
Universidad de Ciencias y Artes de América Lati...	2
Universidad de Huánuco	1
Universidad de Ingeniería y Tecnología	2
Universidad de Lima	1
Universidad de Piura	1
Universidad de San Martín de Porres	1
Universidad del Pacífico	1
Universidad para el Desarrollo Andino	3

C The innovation indicator

Table 9: Universities and their Innovation Indicators

University	Indicator
Asociación Civil Universidad de Ciencias y Huma...	0.093946
Facultad de Teología Pontificia y Civil de Lima	0.850634
Pontificia Universidad Católica del Perú	3.911445
Universidad Andina del Cusco	0.876946
Universidad Antonio Ruiz de Montoya	0.431524
Universidad Autónoma de Ica S.A.C.	0.005046
Universidad Autónoma del Perú S.A.C.	0.873995
Universidad Católica San Pablo	0.572887
Universidad Católica Santo Toribio de Mogrovejo	0.876672
Universidad Católica Sedes Sapientiae	0.446112
Universidad Católica de Santa María	0.975928
Universidad Católica de Trujillo Benedicto XVI	0.429892
Universidad Científica del Sur S.A.C.	0.848429
Universidad Continental S.A.C.	0.591538
Universidad César Vallejo S.A.C.	0.950655
Universidad ESAN	0.971418
Universidad Femenina del Sagrado Corazón	0.852469
Universidad Jaime Bausate y Meza	0.425419
Universidad La Salle	0.455545
Universidad Le Cordon Bleu S.A.C.	0.425419
Universidad Marcelino Champagnat	0.850838
Universidad María Auxiliadora S.A.C.	0.433536
Universidad Nacional Agraria La Molina	1.495406
Universidad Nacional Agraria de la Selva	0.460119
Universidad Nacional Amazónica de Madre de Dios	0.025765
Universidad Nacional Autónoma Altoandina de Tarma	0.002652
Universidad Nacional Autónoma de Alto Amazonas	0.002040
Universidad Nacional Autónoma de Chota	0.003263
Universidad Nacional Autónoma de Huanta	0.002448
Universidad Nacional Autónoma de Tayacaja Danie...	0.001632

Continued on next page

Table 9: Universities and their Innovation Indicators

University	Indicator
Universidad Nacional Daniel Alcides Carrión	0.854917
Universidad Nacional Federico Villarreal	1.014470
Universidad Nacional Hermilio Valdizán de Huánuco	0.875054
Universidad Nacional Intercultural Fabiola Sala...	0.000612
Universidad Nacional Intercultural de Quillabamba	0.001020
Universidad Nacional Intercultural de la Amazonía	0.008539
Universidad Nacional Intercultural de la Selva ...	0.000612
Universidad Nacional Jorge Basadre Grohmann	0.878304
Universidad Nacional José Faustino Sánchez Carrión	0.864054
Universidad Nacional José María Arguedas	0.009941
Universidad Nacional Mayor de San Marcos	3.610864
Universidad Nacional Micaela Bastidas de Apurímac	0.448657
Universidad Nacional San Luis Gonzaga	0.058126
Universidad Nacional Santiago Antúnez de Mayolo	0.881892
Universidad Nacional Tecnológica de Lima Sur	0.022993
Universidad Nacional Toribio Rodríguez de Mendo...	0.888002
Universidad Nacional de Barranca	0.003658
Universidad Nacional de Cajamarca	0.911518
Universidad Nacional de Cañete	0.008539
Universidad Nacional de Educación Enrique Guzmá...	0.869519
Universidad Nacional de Frontera	0.010157
Universidad Nacional de Huancavelica	0.890389
Universidad Nacional de Ingeniería	2.846555
Universidad Nacional de Jaén	0.430490
Universidad Nacional de Juliaca	0.005697
Universidad Nacional de Moquegua	0.451301
Universidad Nacional de Piura	0.924150
Universidad Nacional de San Agustín de Arequipa	1.328332
Universidad Nacional de San Antonio Abad del Cusco	1.199840
Universidad Nacional de San Martín	0.874373
Universidad Nacional de Trujillo	1.074658
Universidad Nacional de Tumbes	0.880886

Continued on next page

Table 9: Universities and their Innovation Indicators

University	Indicator
Universidad Nacional de Ucayali	0.879405
Universidad Nacional del Altiplano	0.948121
Universidad Nacional del Callao	0.952293
Universidad Nacional del Centro del Perú	0.934103
Universidad Nacional del Santa	0.868105
Universidad Peruana Cayetano Heredia	4.437239
Universidad Peruana Los Andes	0.009179
Universidad Peruana Unión	0.891246
Universidad Peruana de Ciencias Aplicadas S.A.C.	1.089825
Universidad Privada Antenor Orrego	0.962927
Universidad Privada Norbert Wiener S.A.	0.457534
Universidad Privada San Juan Bautista S.A.C.	0.472749
Universidad Privada de Huancayo Franklin Roosev...	0.000408
Universidad Privada de Tacna	0.860804
Universidad Privada del Norte S.A.C.	1.173580
Universidad Ricardo Palma	1.982142
Universidad San Ignacio de Loyola S.R.L.	1.055946
Universidad Señor de Sipán S.A.C.	0.452220
Universidad Tecnológica de los Andes	0.434175
Universidad Tecnológica del Perú S.A.C.	0.461980
Universidad de Ciencias y Artes de América Lati...	0.425011
Universidad de Huánuco	0.857312
Universidad de Ingeniería y Tecnología	0.630604
Universidad de Lima	2.075553
Universidad de Piura	1.050196
Universidad de San Martín de Porres	2.218327
Universidad del Pacífico	1.121581
Universidad para el Desarrollo Andino	0.000204

C.1 Ranking Indicator versus Innovation Indicator

University	Ranking Indicador	Innovation Indicator
Universidad Peruana Cayetano Heredia	100.00	4.437239
Pontificia Universidad Católica del Perú	78.25	3.911445
Universidad Nacional Mayor de San Marcos	54.72	3.610864
Universidad Nacional de Ingeniería	25.30	2.846555
Universidad Nacional Agraria La Molina	24.90	1.495406
Universidad Nacional de San Antonio Abad del Cusco	23.10	1.199840
Universidad Peruana de Ciencias Aplicadas S.A.C.	15.65	1.089825
Universidad Nacional de San Agustín de Arequipa	12.04	1.328332
Universidad Científica del Sur S.A.C.	12.03	0.848429
Universidad Nacional de Trujillo	11.49	1.074658
Universidad de San Martín de Porres	10.21	2.218327
Universidad de Lima	10.21	2.075553
Universidad de Piura	9.20	1.050196
Universidad del Pacífico	9.16	1.121581
Universidad de Ingeniería y Tecnología	8.05	0.630604
Universidad Nacional del Altiplano	7.16	0.948121
Universidad San Ignacio de Loyola S.R.L.	6.15	1.055946
Universidad Ricardo Palma	5.86	1.982142
Universidad ESAN	5.74	0.971418
Universidad Privada del Norte S.A.C.	5.36	1.173580
Universidad Nacional Federico Villarreal	4.77	1.014470
Universidad Católica de Santa María	4.29	0.975928
Universidad Nacional de Cajamarca	3.86	0.911518
Universidad Católica San Pablo	3.84	0.572887
Universidad Nacional de Piura	3.71	0.924150
Universidad Continental S.A.C.	3.61	0.591538
Universidad Nacional Agraria de la Selva	2.89	0.460119
Asociación Civil Universidad de Ciencias y Huma...	2.85	0.093946
Universidad Privada Antenor Orrego	2.79	0.962927
Universidad Nacional Jorge Basadre Grohmann	2.74	0.878304
Universidad Nacional del Callao	2.64	0.952293
Universidad Nacional Toribio Rodríguez de Mendo...	2.50	0.888002

Continued on next page

University	Ranking Indicador	Innovation Indicator
Universidad Nacional Santiago Antúnez de Mayolo	2.40	0.881892
Universidad Nacional San Luis Gonzaga	2.19	0.058126
Universidad Nacional Amazónica de Madre de Dios	2.19	0.025765
Universidad Peruana Unión	1.86	0.891246
Universidad Nacional de Huancavelica	1.77	0.890389
Universidad Nacional de Tumbes	1.54	0.880886
Universidad Privada San Juan Bautista S.A.C.	1.43	0.472749
Universidad Nacional de Educación Enrique Guzmá...	1.33	0.869519
Universidad César Vallejo S.A.C.	1.32	0.950655
Universidad Nacional Micaela Bastidas de Apurímac	1.19	0.448657
Universidad Católica Sedes Sapientiae	1.07	0.446112
Universidad Andina del Cusco	0.94	0.876946
Universidad Nacional Hermilio Valdizán de Huánuco	0.80	0.875054
Universidad Nacional de Ucayali	0.75	0.879405
Universidad Privada Norbert Wiener S.A.	0.70	0.457534
Universidad Señor de Sipán S.A.C.	0.61	0.452220
Universidad Privada de Tacna	0.60	0.860804
Universidad Tecnológica del Perú S.A.C.	0.55	0.461980
Universidad Nacional Tecnológica de Lima Sur	0.46	0.022993
Universidad La Salle	0.39	0.455545
Universidad Autónoma del Perú S.A.C.	0.37	0.873995
Universidad Católica Santo Toribio de Mogrovejo	0.16	0.876672
Universidad Nacional del Centro del Perú	0.00	0.934103
Universidad para el Desarrollo Andino	0.00	0.000204
Universidad Nacional de San Martín	0.00	0.874373
Universidad Nacional Autónoma de Chota	0.00	0.003263
Universidad Peruana Los Andes	0.00	0.009179
Universidad Nacional Intercultural de la Amazonía	0.00	0.008539
Universidad Nacional de Cañete	0.00	0.008539
Universidad Nacional de Juliaca	0.00	0.005697
Universidad Autónoma de Ica S.A.C.	0.00	0.005046
Universidad Nacional de Barranca	0.00	0.003658
Universidad Nacional Autónoma Altoandina de Tarma	0.00	0.002652
Universidad Nacional del Santa	0.00	0.868105

Continued on next page

University	Ranking Indicador	Innovation Indicator
Universidad Nacional Autónoma de Huanta	0.00	0.002448
Universidad Nacional Autónoma de Alto Amazonas	0.00	0.002040
Universidad Nacional Autónoma de Tayacaja Danie...	0.00	0.001632
Universidad Nacional Intercultural de Quillabamba	0.00	0.001020
Universidad Nacional Intercultural de la Selva ...	0.00	0.000612
Universidad Nacional Intercultural Fabiola Sala...	0.00	0.000612
Universidad Nacional José María Arguedas	0.00	0.009941
Universidad Nacional de Frontera	0.00	0.010157
Universidad de Ciencias y Artes de América Lati...	0.00	0.425011
Universidad Jaime Bausate y Meza	0.00	0.425419
Universidad Le Cordon Bleu S.A.C.	0.00	0.425419
Universidad Católica de Trujillo Benedicto XVI	0.00	0.429892
Universidad Nacional de Jaén	0.00	0.430490
Universidad Antonio Ruiz de Montoya	0.00	0.431524
Universidad María Auxiliadora S.A.C.	0.00	0.433536
Universidad Tecnológica de los Andes	0.00	0.434175
Universidad Nacional de Moquegua	0.00	0.451301
Facultad de Teología Pontificia y Civil de Lima	0.00	0.850634
Universidad Marcelino Champagnat	0.00	0.850838
Universidad Privada de Huancayo Franklin Roosev...	0.00	0.000408
Universidad Nacional Daniel Alcides Carrión	0.00	0.854917
Universidad de Huánuco	0.00	0.857312
Universidad Nacional José Faustino Sánchez Carrión	0.00	0.864054
Universidad Femenina del Sagrado Corazón	0.00	0.852469

References

A. Jain, M. M. and Flynn, P. (1999). Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323.

Angrist, J. D. and Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, 15(4).

Bartlett, M. (1950). Test of significance in factor analysis. *British Journal of Statistical Psychology*, 3(2):77–85.

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research*. Guilford Publications.

Durbin, J. and Watson, G. S. (1950). Testing for serial correlation in least squares regression: I. *Biometrika*, 37(3/4):409–428.

Engle, R. F. (1984). Wald, likelihood ratio, and lagrange multiplier tests in econometrics. *Handbook of Econometrics*, 2:775–826.

for Digital Research, I. and Education, U. (n.d.). A practical introduction to factor analysis and confirmatory factor analysis. Online Resource. Accessed: 2025-01-24.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1):153–161.

Hoyle, R. H. (2012). *Handbook of Structural Equation Modeling*. The Guilford Press.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35:73–101.

Jarque, C. M. and Bera, A. K. (1987). A test for normality of observations and regression residuals. *International Statistical Review*, 55(2):163–172.

Joreskog, K. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34(2).

Kaiser, H. (1974). An index of factorial simplicity. *Psychometrika*, (30):31–36.

Kaplan, D. (2000). *Structural Equation Modeling: Foundations and Extensions Advanced Quantitaitve Techniques in the Social Sciences Series*. Sage Publications, 1 edition.

Kline, R. B. (2015). *Principles and Practice of Structural Equation Modeling*. The Guilford Press.

McFadden, D. (1974). The measurement of urban travel demand. *Journal of Public Economics*, 3:303–328.

Millones-Gómez and et al., Y.-V. (2021). Research policies and scientific production: A study of 94 peruvian universities. *PLoS ONE*, 16(5).

Mulaik, S. A. (2010). *Foundations of Factor Analysis*. CRC Press.

Newey, W. K. and West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708.

Sarmento, R. P. and Costa, V. (2019). Confirmatory factor analysis: A case study. *arXiv preprint arXiv:1905.05598*.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3–4):591–611.

SUNEDU (2021). *III Informe bienal sobre la realidad universitaria en el Perú*. Superintendencia Nacional de Educación Superior Universitaria, 1 edition.

Trevor Hastie, R. T. and Friedman, J. (2009). *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. Springer Verlag, 2 edition.

Vella, F. (1998). Estimating models with sample selection bias: A survey. *The Journal of Human Resources*, 33(1):127–169.