economies

MDPI

Article

Beyond Firm Size: Revenue Performance and Structure of Firms in the Ecuadorian Education Sector—Panel Data Analysis 2015–2025

Carlos Iván Rivera Naranjo [1], María Estefanía Sánchez Pacheco [2,3], Fernando José Zambrano Farías [2,4,*] and Gloria del Carmen García Zúñiga [5]

[1] Faculty of Social Sciences, Business Education, and Law, Universidad Estatal de Milagro, Milagro 091050, Ecuador; crivera@unemi.edu.ec
[2] Faculty of Administrative Sciences, University of Guayaquil, Guayaquil 090510, Ecuador; maria.sanchez@ug.edu.ec
[3] Faculty of Business Administration and Entrepreneurship, Life College International, Guayaquil 090511, Ecuador
[4] Business School, International University of Ecuador UIDE, Quito 170605, Ecuador
[5] Faculty of Social and Legal Sciences, University of Salamanca, 37008 Salamanca, Spain; ggarciaz5@unemi.edu.ec
* Correspondence: fezambranofa@uide.edu.ec

Abstract

Although financial performance at the firm level is understudied in Latin American educational sector, the theory predicts firm size as one of its main correlates. Using near-census level panel data (12,031 firm-year observations, 2968 firms, 2015–2025) from Superintendencia de Compañías, Valores y Seguros, Ecuador, the study employs descriptive statistics, correlation analysis, log-log regression, and k-means clustering. The regression with year and city fixed effects and clustered standard errors reveals that assets and employment are positively and significantly associated with revenue (elasticities are 0.815 and 0.353; $R^2 = 0.619$). A model with firm and year fixed effects reveals a comparable within-firm elasticity of assets (0.797) and much smaller employment elasticity (0.163). K-means clustering of the 2025 cross-section identifies a few large firms (0.44%), many medium-size firms (12.16%) and a vast majority of small firms (87.40%). Internal validity criteria show no consensus on the optimal k; the three-cluster solution is chosen as parsimonious, substantively sensible, and relatively stable, with $k = 2$, 4 and 5 reported for transparency. This study provides an empirical benchmark for firms in Ecuador's education sector, where the literature has concentrated almost exclusively on educational outcomes so far.

Keywords: business performance; education sector; firm heterogeneity; panel data; econometric modeling; K-means clustering; Ecuador

Check for updates

Academic Editor: Laura Diaconu Maxim

Received: 28 August 2026
Revised: 12 October 2026
Accepted: 19 October 2026
Published: 4 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Education is one of the rapidly developing sectors in Ecuador. It includes private schools, universities, technical colleges and other educational and training institutions, which belong to Section P of the International Standard Industrial Classification (ISIC Rev. 4; United Nations, 2008). Besides its socio-cultural meaning for the country, the education sector is economically important for Ecuador. However, few data exist on the performance, characteristics and diversification of firms in the education sector. Much of the existing literature on education in Ecuador and Latin America is oriented toward questions regarding education outcomes (Bracco et al., 2025; Elacqua et al., 2011).

The applicability of a firm-level, performance-related approach to the educational industry does not merely depend on the social function performed by this sector. Instead, it is grounded in a general principle of 'size matters' for the firm, along with certain industry-specific facts. In the context of the international practice of business, private educational firms take a special place in the theory of the firm. They are characterized by both the intensive use of human capital and low asset ratios, which is common for professional services (Von Nordenflycht, 2010), and by regulatory and reputational restrictions that do not allow implementing the same scaling strategies typical of other sectors (Crawfurd et al., 2024). The firm size, defined as the number of assets and employees, is a key determinant of the firm's revenue according to the literature on corporate performance (Nelson & Winter, 1982; Penrose, 1959; Syverson, 2011). Whether this relation also applies to a service-based sector where human capital plays a pivotal role, such as the educational industry, is an empirical issue that this research aims to answer. This issue has become particularly relevant, as 2015–2025 was not only a period of rapid growth in the number of formally registered educational firms in Ecuador, from 474 firms with data records in 2015 to 2086 firms in 2025, but also the period of an extraordinary external shock, the COVID-19 pandemic, the influence of which on education enrollment and expenditures in the region has been well-studied (Bracco et al., 2025) but never researched before in terms of the financial performance of Ecuadorian firms.

In previous researches about the education industry done in Ecuador and Latin America, the focus was more on education achievements, access, and equality (Bracco et al., 2025; Elacqua et al., 2011). The economics of educational institutions as economic actors—with respect to their size and spatial dynamics in a similar way as in other industries—had not been considered before (Syverson, 2011). As far as we know, previous researchers did not merge an econometric analysis for estimating the financial performance of firms with the use of unsupervised clustering methodology to analyze the structural diversity of the Ecuadorian education sector based on near census-type regulatory information.

The most important goal of the present research is to analyze the correlates and structural differences of revenues in firms in the educational sector in Ecuador within the period of 2015–2025 through the use of econometric models and unsupervised segmentation methods. The main goals include: (i) analyzing the behavior of financial and operational variables in firms in the educational sector; (ii) estimating the relationship between revenues, assets, and employment; (iii) developing an econometric model of revenue in firms; (iv) estimating the effect of time and territory on the revenues of firms; (v) segmenting firms by means of cluster analysis; and (vi) visualizing firm segmentation through PCA. Five hypotheses (H1–H5) are formulated at the end of each corresponding theoretical section in Section 2.

This task is accomplished by using an unbalanced panel data set constructed using the official directory and rankings of the companies held by Superintendencia de Compañías, Valores y Seguros (SCVS); this data set contains 12,031 firm-year observations of 2968 different firms from 2015 to 2025. The multiple linear regression is estimated for logarithmic values of revenue, assets, and employment, with fixed effects for years and cities along with firm-clustered standard errors and robust-regression (RLM) as a check. Then, the k-means algorithm is run for clustering on a cross-section of 2025, using the elbow method, silhouette score, Calinski–Harabasz index, Davies–Bouldin Index, and gap statistics with PCA validation and visualization.

The relationship between assets and employment is found to be positively and significantly related to firm revenue, with elasticities of 0.815 and 0.353, respectively ($R^2 = 0.619$). The city fixed effects are jointly significant (conventional joint F-test) after controlling for firm size. As explained in Section 4.6, the city effects are estimated relative to a pooled "other

cities" benchmark (lower activity cities described in Section 3.4), in which there are eight cities that have significantly higher conditional revenues than this pooled benchmark. Note that Guayaquil does not appear among them despite being a major market, therefore one should interpret these effects as conditional gaps relative to the pooled benchmark and not as a ranking of the "best city". Firms are segmented into three categories: (i) a very small number of large firms with high revenues (0.44%); (ii) a small number of medium-sized firms (12.16%), and (iii) small firms, which make up the vast majority of such businesses (87.40%). Collectively, the results imply that firm revenues in Ecuador's educational sector are strongly associated with the size of physical and human capital investment. Firm sizes are highly polarized, with the vast majority of firms operating at small scale—which fits the structural constraints faced by SMEs documented in the previous section (Section 2)—along with a secondary role played by the city where these firms are located.

In combining econometrics and unsupervised clustering using a regulated census-like data set, this paper sheds light on the heterogeneous structure of Ecuador's education sector, complementing the outcome-oriented literature about the Latin American education sector and providing the basis for further empirical evaluation of the usefulness of techniques proven valuable for analyzing SMEs and the service sector (Ibarrarán et al., 2009). The information is of value both for academics and policymakers/investors.

This paper proceeds as follows. Section 2 reviews the theoretical background of the paper and develops its hypotheses. Section 3 describes the data and method. Section 4 presents the empirical results. Section 5 discusses the results in the context of the literature. Section 6 concludes.

2. Theoretical Framework

There are two bodies of literature that serve as the basis for this study. The first one is the literature on the determinants of firm performance from the resource-based view and production function perspectives, which states that firm size, usually defined as the number of total assets and employees, is an important determinant of firm revenue and competitiveness (Nelson & Winter, 1982; Penrose, 1959; Syverson, 2011; Teece & Pisano, 1994). The second is the literature on the transaction cost and institutional perspective on the theory of the firm, where firm boundaries and size are treated as economic outcomes determined by organizational and contractual factors. The two together imply significant heterogeneity in firm size and productivity as a result of the selection and growth process (Jovanovic, 1982; Nelson, 1991) and misallocation of resources between firms of various sizes (Foster et al., 2001; Hsieh & Klenow, 2009). Another literature is concerned with the independence of firm growth from firm size according to Gibrat's law of proportional effect. The empirical evidence shows that the growth is not independent from the size of firms, because the smaller and the younger ones are growing at a higher rate than the larger and the older firms do (Evans, 1987; B. H. Hall, 1987; Sutton, 1997). Such heterogeneity in size and growth has been established for size distributions of cities and other economic entities (Eeckhout, 2004) and high levels of heterogeneity in job creation and destruction rates of firms of various sizes (Davis et al., 1996). For Latin America, such misallocation of resources between firms is a reason for lower aggregate productivity compared to more developed countries (Busso et al., 2013; R. E. Hall & Jones, 1999). It is connected to the wider discussion of the growth literature regarding the relative role of factor accumulation versus total factor productivity in explaining income differences between countries (Easterly & Levine, 2001; Klenow & Rodríguez-Clare, 1997) and importance of firm-level research for studying sectoral performance in the region. In empirical corporate finance literature, a logarithm of revenue, assets, and employees is taken in order to reduce skewness, allowing for an interpretation of the regression coefficients as elasticities. Multicollinearity is checked

via correlation analysis before regression estimation (Gujarati & Porter, 2009). When there are a strong heterogeneity and outliers in firm-level data—which is quite frequent in the case of administrative business registries—heteroskedasticity-robust standard errors (White, 1980) and robust regression (Huber, 1981) is used to overcome the possible negative impact of these characteristics.

The following hypotheses on the correlates of firm revenue are suggested on the basis of the theoretical background above:

H1. *Firm assets are positively related to the revenue of education-sector firms.*

H2. *Employment level is positively correlated with the revenue of education-sector firms.*

The second set of literature concerns firm segmentation and unsupervised clustering. K-means clustering is a popular technique for grouping heterogeneous populations into homogenous clusters characterized by similar financial and operational features (Jain, 2010). The number of clusters is usually established with the help of the elbow method and silhouette index. Principal component analysis is applied for verification and illustration purposes as a dimension reduction technique (Jolliffe & Cadima, 2016). Together, all these literature streams serve as a theoretical framework for the methodology developed herein, in which the estimation of performance determinants and unsupervised segmentation procedures are performed simultaneously to analyze heterogeneity in the sector's structure. Based on that, the following hypothesis can be stated:

H3. *There is a systematic heterogeneity of financial and operational scale in the firm population, which can be described with distinctive empirical profiles.*

Finally, there is a literature stream that touches on the issue of determinants of performance heterogeneity across firms. The resource-based view (RBV) of the firm postulates that the existence of persistent performance differences between firms positioned similarly is based on the ownership of resources that are valuable, rare, imperfectly imitable, and non-substitutable (Barney, 1991). Being applied to a human capital-intensive service sector like education, the RBV suggests viewing not only physical and human capital as the traditional factors of production, but also as the resources that the new entrant cannot reproduce rapidly due to their accumulative nature. Hence, the following testable prediction can be derived from the RBV: If physical and human capital resources are accumulated and determine performance heterogeneity in the sector, then both should be positively associated with firm revenue—a hypothesis that will be tested empirically in Section 4. Being mainly based on the provision of services via humans rather than traded goods, the employment can be viewed, despite being a rough measure, as a proxy for human capital resources, driving performance heterogeneity along with the physical capital one.

Finally, there is support for the association between human capital and firm size in the recent literature. In particular, analyzing cross-country data, Bento and Restuccia (2021) show that the average firm size increases with the share of the educated workforce in the sector. This association is not solely caused by sector composition effects, but also complementarity between skill and firm size. As both the providers and main consumers of skilled human capital, education-sector firms make the link between the classical, assets-based theory of firm size (Penrose, 1959) and the recent, human capital-based theories of firm size distribution (Bento & Restuccia, 2021). This approach provides a theoretical justification for the expected positive association between employment and firm revenue beyond the more classical channel of assets.

The extent of firm-size polarization that has emerged in the current analysis—a few large firms, the majority being small-scale firms—relates to the productivity disparity between small/medium-sized firms and large firms in Latin America. Using harmonized datasets from firm-level surveys conducted in various Latin American countries, Ibarrarán et al. (2009) reveal productivity disparities between small/medium-sized firms and large firms due to differences in their access to credit, training, and technology adoption and show that generic SME support policies do not have a significant impact on closing this productivity gap. This literature is particularly pertinent in the context of interpreting the small firms cluster obtained in this analysis (Cluster 3 in the three-cluster solution presented in Section 4.7, accounting for 87.40% of education sector firms in the 2025 cross-section). The prevalence of small-scale firms in this segment is in line with structural problems noted in the general Latin American SME literature, namely inaccessibility of credit, managerial skills, and technology adoption. However, these mechanisms are not evident in the current data set and thus cannot be tested empirically for the firms constituting the small firms cluster.

The temporal dimension of firm performance is rooted in another body of literature examining the dynamics of firms under the influence of macroeconomic shocks in Latin America. Economic shocks have not always resulted in commensurate declines in firm formation rates (Bergoeing & Repetto, 2006). A broader literature on structural transformation has related the shift of economic activity toward registered firms with improved productivity in Latin American economies (Lavopa, 2015; McMillan et al., 2014; Pagés, 2010a), consistent with earlier results linking persistent informality to low aggregate productivity (Perry et al., 2007). However, this study's firm count data cannot identify the specific channel of formalization. In the education sector, the recent survey data of Latin American households shows that school closures due to COVID-19 resulted in a disproportionate drop of private school enrollments due to household income loss (Bracco et al., 2025). Consistent with this literature on shock-driven and structural causes of temporal variation, the revenues of education sector firms in Ecuador are expected to vary systematically over the 2015–2025 period. Therefore, this leads to:

H4. *There are systematic differences in firm revenue over time.*

Finally, the territorial dimension of firm performance is rooted in agglomeration economics literature, suggesting the benefits of firms being located together due to pooled labor, knowledge spillover, and access to specialized inputs. Torres-Gutiérrez et al. (2019) reveal the significant and positive impact of urbanization on the productivity of industry and services in Ecuador over 1980–2010 based on census data for Ecuador, with the disproportionate impact coming from Ecuadorian provincial capitals due to public investments, market access, and labor market specialization in these cities. The presence of urban-agglomeration benefits for Ecuadorian firms is to be expected based on these findings in view of the presence of the two largest metropolitan areas of Ecuador—Quito and Guayaquil. However, the competition between firms in these metropolitan areas will mitigate some of those benefits for firms, meaning that territorial differences in firm revenue can persist even when the firm size is controlled for. This leads to:

H5. *The revenue of firms differs according to geographic location (city).*

As a result of this analysis of the resource-based view, human capital theory, formalization theory, SME productivity studies, and agglomeration theory literature, a shared hypothesis can be drawn concerning the sector in question: Revenue generated by firms in Ecuador's education sector will show systematic dependence on resource scale, diversity,

and geographical and temporal factors (H1–H5). This hypothesis will be tested using the method outlined in Section 3.

3. Materials and Methods

3.1. Population and Sample

This study population involves those companies operating within the Ecuadorian education industry that have been registered by the Superintendencia de Compañías, Valores y Seguros (SCVS), the official regulatory agency in the country. The following are some of the conditions used to select the sample of these firms: Inclusion in the register of companies, provision of information on their finances, and valid and realistic data regarding revenues, assets, and number of employees. In this case, there are 12,031 firm-year observations of 2968 unique firms.

For the general sample criteria, the requirement is to have valid and economically sensible values but not necessarily positive values. Revenues and employment can very well be zero for this sample (for instance, a registered company having no sales or employees in the particular year under consideration). The stronger criterion of being positive is applied only to the regression subsample considered in Section 3.4, since the regression specification in Section 3.5 applies the logarithmic transformation to revenue, assets, and employment and logarithms cannot be taken for zero values. This criterion is the reason why the regression subsample (11,774 observations) is smaller than the general sample (12,031 observations).

The period chosen for analysis was 2015–2025 instead of the initially planned 2014–2024, based on the criterion of data completeness and not on grounds of analytical reasoning. The number of observations per firm-year from the cleaned dataset gives us the following picture. There are only six observations in 2014 in contrast with 377 in 2013 and 474 in 2015 (Table 1). Since six observations will not provide us with any estimate, we decided to discard 2014 as the first year and chose 2015 instead. We performed the same test in the reverse order to see whether 2025 could be used as the last year of the panel—the results show that for years 2023 to 2025, revenue, assets, and employment are 100% non-missing for the firm-year observations in the registry. Thus, we included 2025 as the final year of our panel. The completeness check was performed on the sample of firms that are in the registry, not on the sample of firms that were active in 2025 (Table 1).

Table 1. Number of firm-year observations each year in the cleaned sample before imposing the window restriction from 2015 to 2025.

Year	Firm-Year Records	Year	Firm-Year Records
2010	1	2018	707
2011	26	2019	887
2012	348	2020	927
2013	377	2021	1075
2014	6 (excluded—gap)	2022	1263
2015	474 (period starts)	2023	1587
2016	546	2024	1854
2017	625	2025	2086 (period ends)

Source: Authors' computations based on the analytical panel obtained through the SCVS approach, before the 2015–2025 window filtering criteria presented in Section 3.1 but following the consistency filters outlined in Section 3.4. Year 2014 is not included in the analysis because of a lack of data completeness (only 6 records available vs. 377 in 2013 and 474 in 2015).

The resultant dataset of 12,031 firm-year observations on 2968 firms offers a good representation of the education companies covered in the SCVS database and is suffi-

cient for conducting the descriptive, correlational, and exploratory phases of the study (Sections 4.1–4.5). However, this dataset cannot be considered a probability or representative sample of all education providers in Ecuador, since the sample is limited to registered companies with SCVS (mostly sociedades anónimas and their equivalents) and thus does not cover possible cases of informal providers and institutions registered under different laws. Two smaller samples are taken for the regression and clustering analysis and are described in Section 3.4 below. Table 2 presents the summary of observation flow from raw data merging up to the final regression sample generation. The main decrease, from 1,342,863 firm-year observations (all ISIC sections and years from the SCVS rankings database) to 12,789 firm-year observations, is explained largely by sector scope filtering (ISIC Section P, education) along with consistency filters mentioned in Section 3.4, and not by the loss of quality data within the education sector itself. Further decreases of the sample size to 12,031 firm-year observations after 2015–2025 time window filtration and to 11,774 observations in the final regression sample are much less significant.

Table 2. Observation-flow (attrition) summary, raw merged dataset to final model estimation sample.

Stage	Firm-Year Observations	Distinct Firms
Raw merged dataset (post inner-join, all ISIC sections, all years in the ranking source)	1,342,863	n/a
After sector scope (ISIC Section P, education) + consistency filters	12,789	n/a
Study window, 2015–2025 (general cleaned dataset)	12,031	2968
Final regression estimation sample	11,774	2925

n/a: not applicable. Distinct-firm counts are not meaningfully defined for the two earliest stages, which pool all ISIC sections together.

3.2. Data Sources

The data regarding the firms was collected from two databases of SCVS for the period 2015–2025 and compiled in the form of an unbalanced panel of firms, where each firm appears for one or several periods. The directory of companies includes identification and administrative-territorial information (case-file code, company name, province, city, type of legal entity, year of registration, legal entity). The company ranking includes economic and financial information (sales, assets, number of employees, profit, equity, total debt, costs, year of observation). Two datasets were matched using the inner join on the numeric identifier present in both databases, keeping only the observations common to both databases.

3.3. Variables

Following are the major variables utilized in this study along with their types and scales of measurement.

- ingresos_ventas (revenue): total sales revenue—continuous quantitative, ratio scale.
- activos (assets): total firm assets—continuous quantitative, ratio scale.
- n_empleados (employees): number of employees—discrete quantitative, ratio scale.
- utilidad_neta (net income): net result for the period—continuous quantitative, ratio scale.
- edad_empresa (firm age): years of operation—discrete quantitative, ratio scale.
- liquidez_corriente (current liquidity): current-liquidity indicator—continuous quantitative, ratio scale.
- roe: return on equity—continuous quantitative, ratio scale.
- total_gastos (total expenses): total expenses—continuous quantitative, ratio scale.

- ciudad (city): geographic location—qualitative, nominal scale.
- anio (year): observation period—discrete quantitative, interval scale.

3.4. Data Treatment and Preparation

Several actions were performed to preserve data quality and consistency and to make sure that the data used are relevant for multivariate and econometric analysis. First of all, variable names were normalized. This was done in order to avoid inconsistencies related to spacing, accents, and special characters. Secondly, textual variables—legal status, city, and province—were normalized, and in particular white spaces were trimmed. Normalization of these variables allowed us to avoid creating duplicate categories based on different formatting.

As regards the variable of a firm's date of incorporation, the original value was kept and several date formats were tried during parsing. Dates that occurred prior to 1900 or after 2025 were considered to be implausible and were treated as missing in order to calculate firm age as the difference between the year of observation and the year of incorporation.

Finally, several consistency filters were applied in order to keep economically plausible observations, such as strictly positive assets, non-negative revenue, and employment, smaller than 50,000. The latter is a plausibility filter against implausible values of data entry in the source ranking, but not a firm-size threshold. The maximum number of employees in the final sample is only 27,340 (Section 4.1), and the limit does not bind. Furthermore, only firms belonging to ISIC Rev. 4 Section P (education) were included in the sample. In view of the typical dispersion of administrative data on firms, the outliers were filtered using consistency criteria rather than winsorization. The regression models presented in Section 4 are estimated on the whole, non-winsorized distribution of the modeling dataset. Robustness to extreme observations was checked with the help of robust-regression (RLM) estimation and residual diagnostics provided in Section 4.6.

For log-revenue, log-assets, and log-employment—constructed to mitigate distributional skewness (Section 4.1) and interpret the estimated coefficients as elasticities—the regression subsample described below is used. The purely descriptive log-scale graphs in Section 4.1 have another purpose and were plotted using the general cleaned dataset (i). As there is a handful of observations in the general dataset having the value zero for at least one of the four variables (Section 4.1), and since taking a logarithm of zero is not possible, each panel of these graphs displays the observations that are strictly positive on the particular variable. Therefore, the number of plotted observations is slightly lower than 12,031 and may vary across the panels. There are three subsets of the cleaned dataset depending on the type of analysis performed: (i) The general cleaned dataset (12,031 observations, 2968 firms, 2015–2025) is used for the descriptive and temporal and territorial analyses presented in Sections 4.1–4.5. (ii) The modeling dataset requires the positivity of the revenue, assets, and employment, groups the cities that contain less than 30 modeling dataset observations in one "other" group and therefore contains 11,774 observations of 2925 firms in 31 city categories, and is used for the regression analysis in Section 4.6. (iii) A cross-sectional subset of the dataset for 2025, which includes 2056 firms with complete values on the clustering variables, is used for the segmentation analysis in Section 4.7.

The correlation matrix reported in Section 4.5 uses the same eight variables discussed in Section 3.3, but applies a listwise (or complete-case) deletion joint for all eight, rather than calculating each pairwise correlation on its own maximum available sample. Thus, each cell in the correlation matrix is based on the same sample size, which is reported in the note to the correlation table rather than the variable-specific sample sizes reported in the descriptive statistics.

3.5. Analytical Techniques

Descriptive and inferential statistical methods were used, as well as multivariate methods. Descriptive statistics were used to explore the distributions, variability, and outliers using histograms and boxplots as well as measures of central tendency and variability (mean and median, standard deviation) (Field, 2013). A correlation matrix was used to examine the pairwise relationships among the variables. Multicollinearity was assessed using variance inflation factor (VIF) in the regression equation (Gujarati & Porter, 2009).

In the econometric analysis, the multiple linear regression model was estimated using logarithmically transformed variables in order to enhance statistical properties of the data and interpret the estimated coefficients as elasticities (Hsiao, 2014; Wooldridge, 2013). The model contains year and city fixed effects (dummy variables) to control, respectively, for the shocks and cross-sectional heterogeneity among cities. The baseline specification is:

$$\ln(\text{Revenue}_{it}) = \beta_0 + \beta_1 \ln(\text{Assets}_{it}) + \beta_2 \ln\left(\text{Employment}_{it}\right) + \sum_t \gamma_t \text{Year}_t + \sum_c \delta_c \text{City}_c + \varepsilon_{it}, \quad (1)$$

where the subscript i indicates the firm, and the subscript t indicates the year (2015–2025); Revenue_{it}, Assets_{it}, and Employment_{it} indicate sales revenue, total assets, and number of employees of firm i in year t; Year_t and City_c are year and city dummies with the corresponding coefficients γ_t and δ_c; β_1 and β_2 are the elasticities of the revenue with respect to assets and employment; and ε_{it} is the error term.

Cluster-robust, heteroskedasticity-consistent standard errors (clustered by firm) were estimated for Equation (1) according to White (1980) and further advancements in cluster-robust inference (Cameron et al., 2011; Cameron & Miller, 2015). The RLM specification was estimated to decrease the impact of outliers (Huber, 1981). For assessing multicollinearity, the Variance Inflation Factor (VIF) was used. As another robustness test, which takes into account all time-invariant, firm-specific unobservable factors, Equation (1) was estimated with firm fixed effects instead of the intercept and city dummies, identifying β_1 and β_2 solely based on the within-firm time variation.

Secondly, the K-means clustering algorithm (MacQueen, 1967), commonly used as a method to segment heterogeneous databases into clusters based on shared properties (Jain, 2010), including firms' segmentation using financial performance indicators (Herman et al., 2022), was applied to the 2025 cross-section of firms. K-means partitions n firms into k clusters as follows:

$$\text{minimize} \sum_{j=1}^{k} \sum_{x \in S_j} \left\| x - \mu_j \right\|^2, \quad (2)$$

where S_j is cluster j ($j = 1, \ldots, k$) and μ_j is the centroid of cluster j, defined over the standardized financial and operational characteristics given in Section 3.3 (revenues, assets, employment, and age). For finding the optimal number of clusters, elbow (Thorndike, 1953) and silhouette (Rousseeuw, 1987) indices—the two most commonly used methods—were considered alongside with Calinski–Harabasz index, Davies–Bouldin index, the gap statistic (Tibshirani et al., 2001), and bootstrapped cluster stability analysis (100 resamples, mean Jaccard similarity). Principal Component Analysis (PCA)—developed by Pearson (1901) and Hotelling (1933)—was considered as a tool for reducing dimensions by compressing information from the larger number of variables into a smaller number of components. Visualization of the resulting clusters was done using PCA (Jolliffe & Cadima, 2016). Another robustness test, which expands the comparison up to $k = 5$ (and, for the Calinski–Harabasz, Davies–Bouldin, and gap-statistic criteria specifically, to $k = 6$–8), was carried out using the same clustering variables, scaling, seed, and nstart parameters. Cluster membership was cross-tabulated between three- and five-cluster solutions, and the sensitivity of clustering to the variable transformation was analyzed by comparing the cluster memberships obtained based on the original standardized clustering variables with those obtained after the log1p

transformation of revenues, assets, and employment, using the Adjusted Rand Index to measure the agreement between two partitions separately for each k.

3.6. Software

All calculations were performed in R (version 4.5.1; R Core Team, 2025) via RStudio (version 2026.6.0, Build 242). The dplyr (version 1.2.0) package was used for data processing, ggplot2 (version 4.0.2) for visualization, base R and Hmisc (version 5.2.5) packages for correlations, multivariate analysis (K-means clustering and PCA)—factoextra (version 1.0.7) and cluster (version 2.1.8.1), fpc (version 2.2-15) for cluster stability bootstrapping, robust regression—MASS (version 7.3.65), sandwich (version 3.1.1) and lmtest (version 0.9.40) for heteroskedasticity testing and cluster-robust standard errors, fixest (version 0.13.2) for firm- and year-fixed effects estimation, and car (version 3.1.3) for multicollinearity diagnosis (VIF).

4. Results

4.1. Exploratory Data Analysis

First of all, the distribution of the most important financial characteristics of the sector was examined through box plots (Figure 1) and frequency histograms (Figure 2) in logarithmic scale, together with the complete descriptive statistics table given below (Table 3). Taking into account the usual dispersion of the administrative firm-level data, revenues, assets, and net income show significant variation and many outliers, whereas the number of employees is less dispersed. The logarithmic transformation of the respective variables significantly mitigates the skewness, making it appropriate for the econometric specification given in Section 4.6.

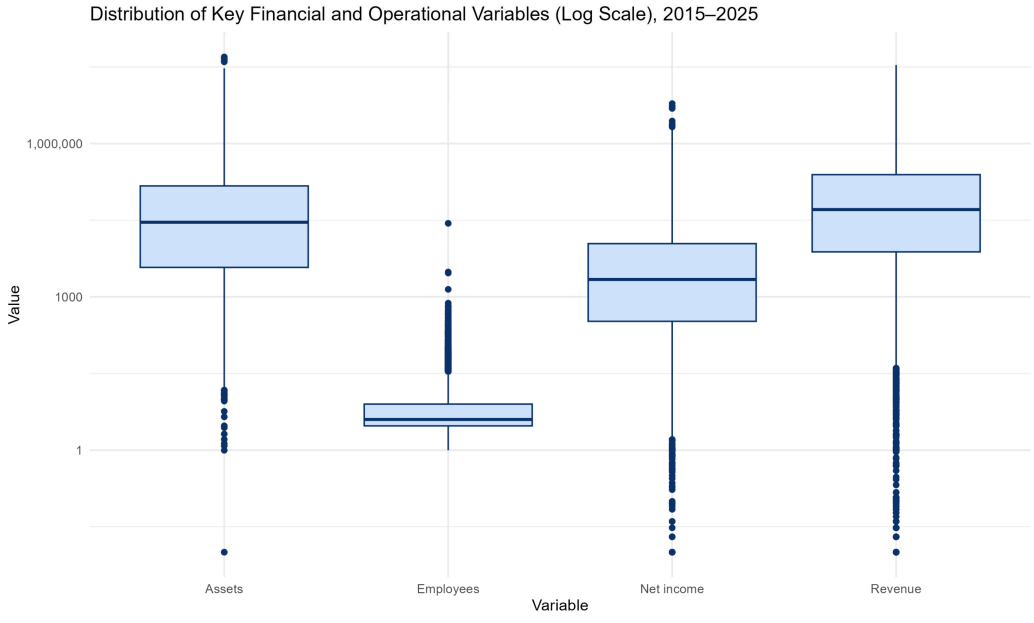


Figure 1. Distributions of revenue, assets, net income, and employees (base-10 log axis, values in original units), general dataset, 2015–2025. The base-10 logarithm used in the axis above is solely for display purposes, different from the natural-logarithm (ln) transformation of revenue, assets, and employment in the regression analysis of Section 3.5. Since the logarithm is undefined at zero, each panel graphically displays the strictly positive observations for that variable (less than the general dataset total of n = 12,031; Section 3.4); zero and negative values are included in the general dataset and are reported in Table 3. Boxes show the interquartile range (IQR); whiskers extend to 1.5 × IQR; outliers past that range are shown individually as dots.

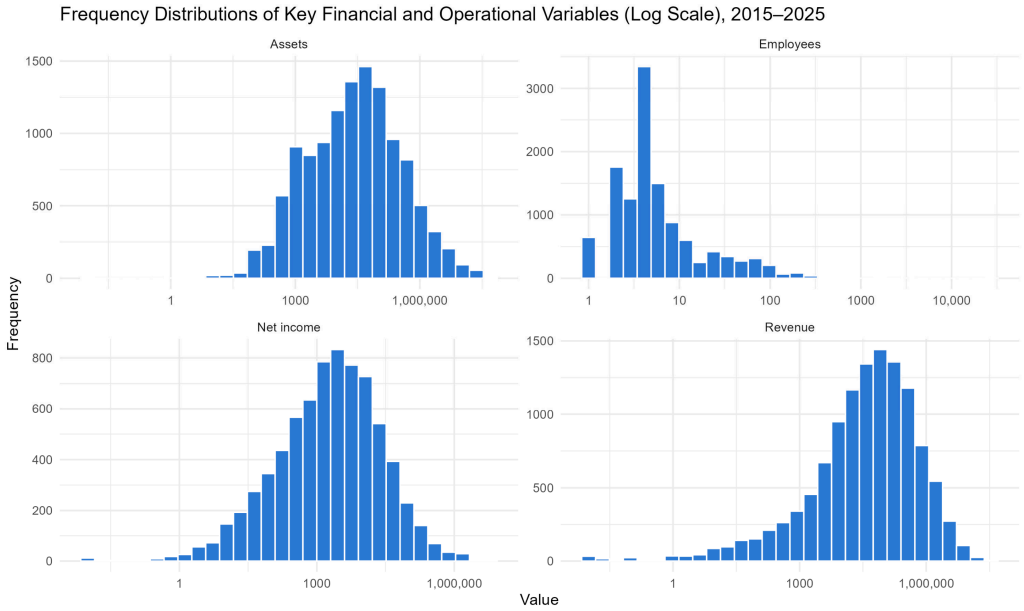


Figure 2. Frequency histograms (base-10 log axis, values in original units, 30 bins per panel) of the same four variables as in Figure 1, using the same strictly positive plot sample as in the previous figure, showing the remaining right-skewness that drives the use of logarithms in the regression specification of Section 3.5.

Table 3. Descriptive statistics for the core variables, general cleaned dataset.

Variable	*n*	Mean	Median	SD	Min	Max	Q25	Q75
Revenue (USD)	12,031	$381,882	$47,829	$1,282,727	$0	$34,665,236	$6495	$238,981
Assets (USD)	12,031	$456,492	$28,708	$2,041,087	$0.01	$49,785,722	$3761	$149,505
Employees	12,031	17.71	4	255.53	0	27,340	3	8
Net income (USD)	12,031	$9669	$167	$179,164	−$6,281,588	$6,114,830	−$0.13	$3848
Firm age (years)	12,031	7.22	4	8.76	0	64	1	10
Current liquidity	10,386	59.71	1.06	1516.24	0	100,004	0.26	2.71
ROE	10,386	1.26	0.23	15.52	0	1250.19	0.03	0.87
Total expenses (USD)	10,386	$275,820	$30,466	$1,022,427	$0	$55,171,649	$3641	$159,459

Source: Calculations by the authors. n = 12,031 firm-years for revenue, assets, employees, net income, and firm age; n = 10,386 for current liquidity, ROE, and total expenses (variables with additional missingness in the ranking source).

4.2. Evolution of the Sector over Time

The revenue and the number of firms with regulatory data were considered over the years 2015 to 2025 (Figures 3 and 4). The sector's revenue increased from $322.4 million in 2015 to $605.2 million in 2025, with a contraction to $319.0 million in 2020—from $442.1 million in 2019—due to the COVID-19 crisis, followed by growth until 2023–2025. All dollar values in the present work are nominal in USD without adjustments for inflation. The number of firms with regulatory records rose steadily over the period, from 474 in 2015 to 2086 in 2025, even in 2020, when there was a revenue contraction. As indicated below in Figure 4, this increasing number of observations might be driven by firm entry, changes in registry coverage, or variations in reporting compliance, but does not directly indicate firm entry itself.

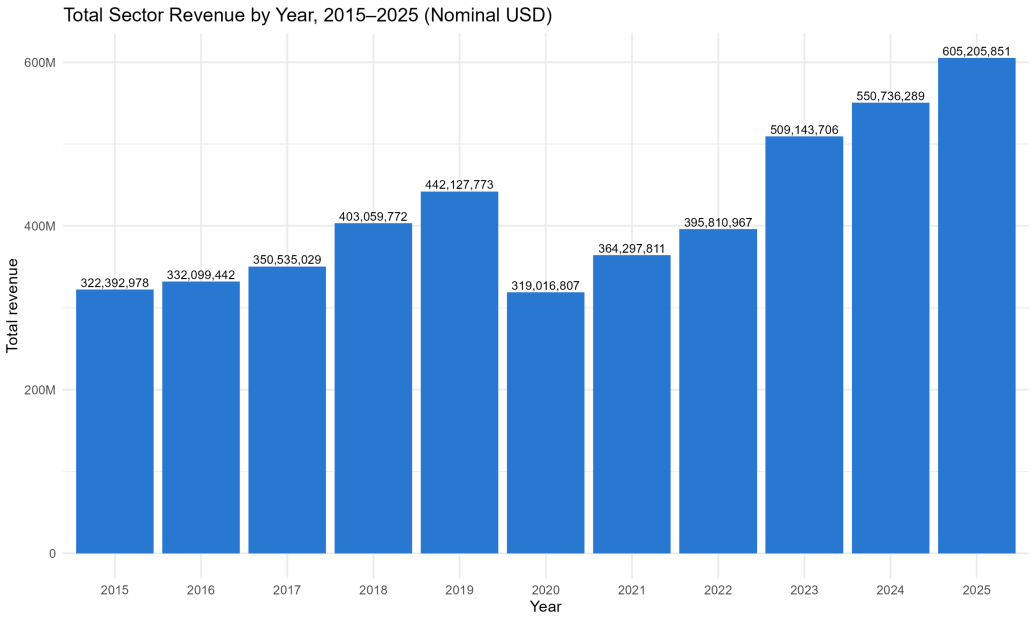


Figure 3. Annual total sector revenue (nominal USD, without inflation adjustment), pooled over all firms in the cleaned dataset, 2015–2025. The 2020 contraction and following 2021–2025 recovery are discussed in Section 4.2.

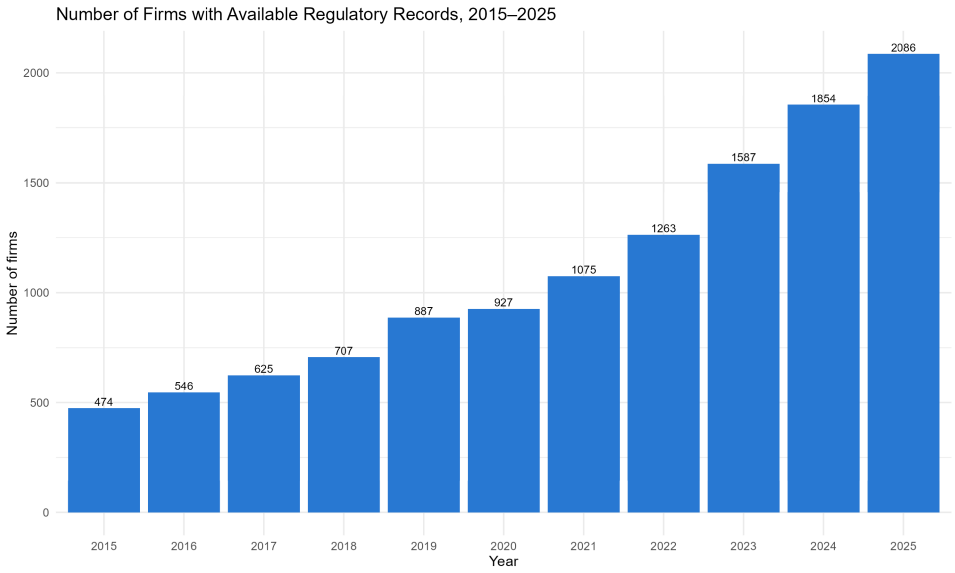


Figure 4. Number of unique education-sector firms with available regulatory data per year, 2015–2025. This graph shows how the number of firms in the dataset has evolved over time during the observation period. It could be due to new firm emergence, but could also be due to changes in coverage of the registry, or reporting compliance—do not interpret this graph as a proxy for new firm creation.

4.3. Geographical Analysis and Market Concentration

The geographical distribution of revenues in 2025 was examined at both the provincial and firm level (Figures 5 and 6). Pichincha (revenue: $247.0 million, number of firms: 830) and Guayas (revenue: $231.3 million, number of firms: 591) remain the leaders in sector revenues, followed by Azuay ($36.5 million), Manabí ($22.4 million), and El Oro ($19.7 million). The largest company in the 2025 cross-section (Colegio Menor CMSFQ S.A. from Pichincha) had a reported $33.6 million in revenue and $49.8 million in assets, which is more than twice the revenue of the second-largest firm (Academia Naval Almirante Illingworth S.A., Guayas, $15.3 million).

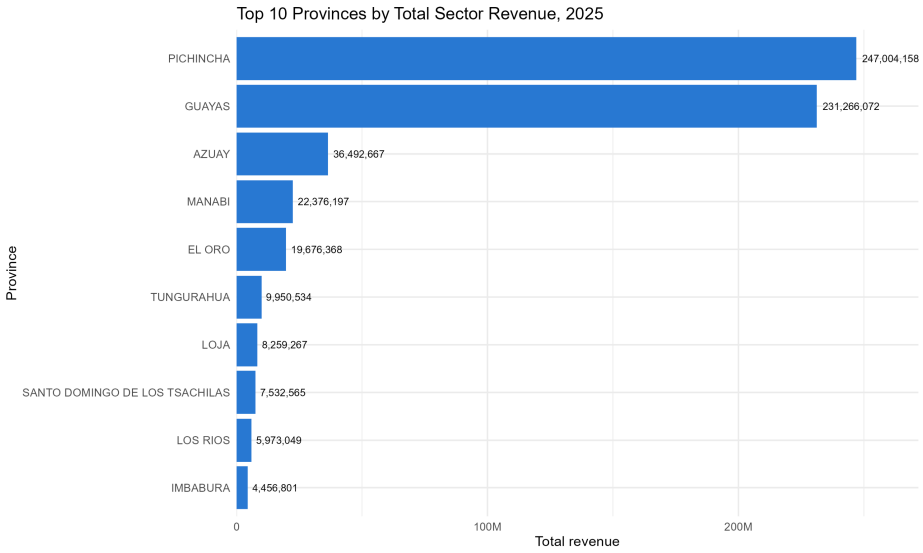


Figure 5. Top 10 provinces by total revenue of education-sector firms, Ecuador, 2025 cross-section (nominal USD). Pichincha and Guayas are leaders by the total revenue generated by firms in these provinces. However, the HHI/CR4 concentration measures presented below in Table 4 are calculated on the basis of cities' share of revenue, and not province-level.

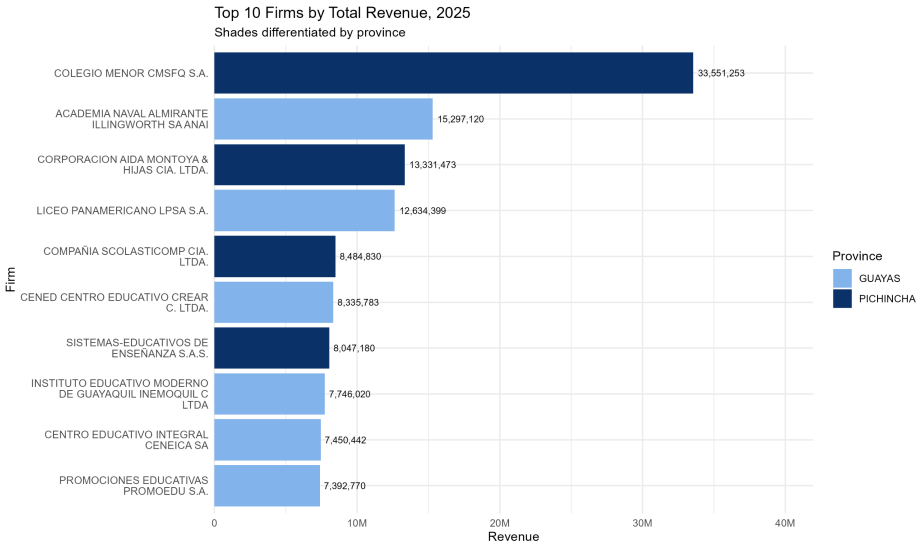


Figure 6. Top 10 firms by total revenue, 2025 cross-section (nominal USD). The shades highlight two provinces that contain firms in the top 10 (Guayas, Pichincha). The bar labels show exact revenue of each firm in 2025.

Table 4. City-level revenue concentration, 2025.

Metric	Value
HHI (city revenue shares, 2025)	1806.7
CR4 (top 4 cities, 2025)	70.8%
Top 1—Quito	30.7% of sector revenue
Top 2—Guayaquil	26.6% of sector revenue
Top 3—Samborondón	7.2% of sector revenue
Top 4—Cumbayá	6.3% of sector revenue

Source: Authors' calculations. The numbers characterize the geographic concentration of the sector's revenue and should not be considered as antitrust metrics of market power. HHI and CR4 are traditionally calculated on the basis of individual firms' share in the specified product market, rather than on the basis of cities' share in the regulatory sector.

These measures confirm the above findings at the city level (the geographic unit in the fixed effects city regression in Section 4.6). The HHI of sector revenue across cities is 1806.7, and CR4 is 70.8% (Quito: 30.7%, Guayaquil: 26.6%, Samborondón: 7.2%, Cumbayá: 6.3%). These results show a very uneven distribution of sector revenue. They are presented here as descriptive statistics and cannot be used as antitrust market power measures, since HHI and CR4 are calculated on the basis of individual firms' share of market, not cities' share of the regulatory sector.

4.4. Trend and Firm Dynamics Analysis

Trends analysis has been conducted on the provincial and firm levels (Figures 7 and 8), limited to the period between 2015 and 2025 (see Section 3.1). In the sample of the five provinces with the largest total revenue, Guayas and Pichincha demonstrate the same growth trends and decline in 2020, while Azuay, El Oro and Manabí reveal less pronounced and more stable growth. Among the ten largest companies in the 2025 cross-section, some demonstrate steady growth over several years, while others demonstrate more unstable growth trends and have recently become the market leaders.

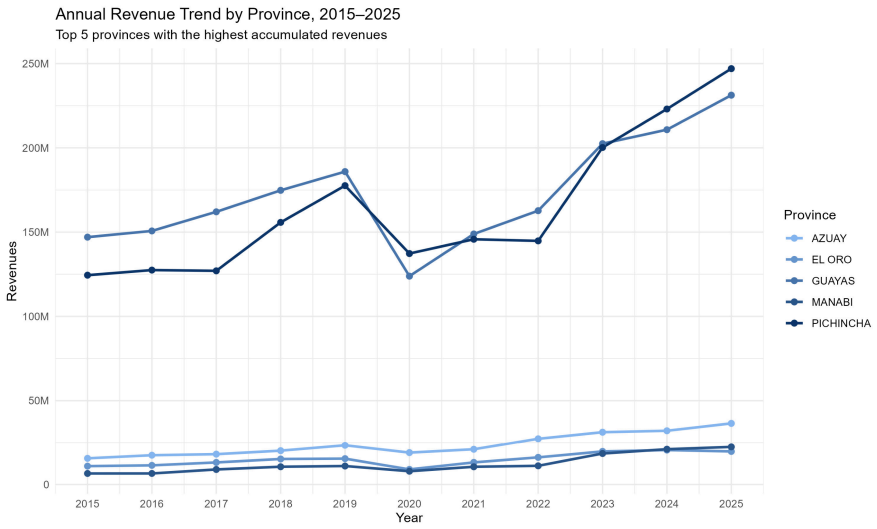


Figure 7. The annual revenue trend from 2015 to 2025 for five provinces with the largest total revenue in 2025 (nominal USD). The lines are limited to the study window from 2015 to 2025 (Section 3.1).

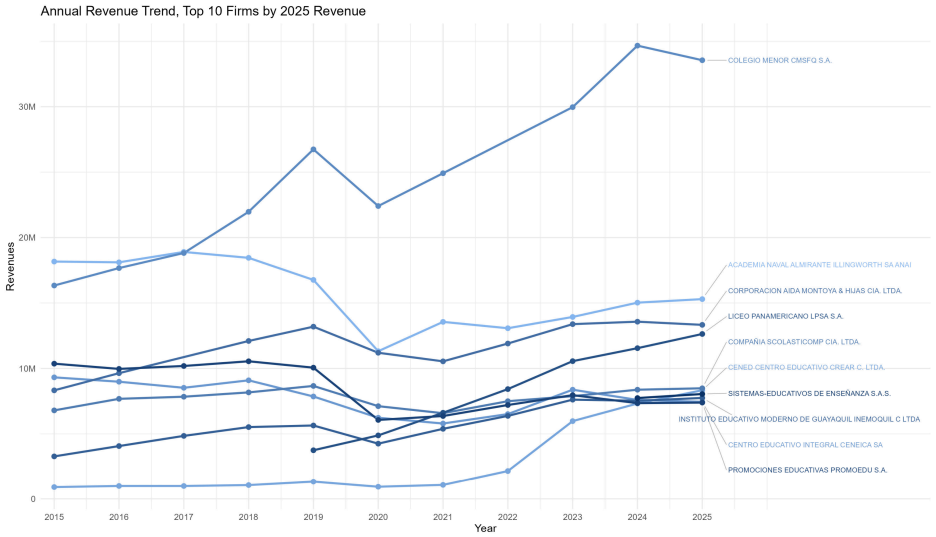


Figure 8. The annual revenue trend from 2015 to 2025 for the ten largest firms in 2025 (nominal USD). Firm names are reproduced as legally registered with the SCVS.

4.5. Correlation Analysis

The pairwise correlation coefficients have been calculated and tested for statistical significance (Table 5). The matrix is shown in Figure 9 in the form of a heatmap.

Table 5. Pearson correlation matrix with statistical significance.

Variable	Revenue	Assets	Employees	Net Income	Firm Age	Liquidity	ROE	Expenses
Revenue	1.000	0.822 ***	0.110 ***	0.503 ***	0.388 ***	−0.011	−0.011	0.724 ***
Assets	0.822 ***	1.000	0.095 ***	0.325 ***	0.369 ***	−0.008	−0.012	0.599 ***
Employees	0.110 ***	0.095 ***	1.000	0.041 ***	0.051 ***	−0.002	−0.001	0.093 ***
Net income	0.503 ***	0.325 ***	0.041 ***	1.000	0.087 ***	−0.001	0.001	0.222 ***
Firm age	0.388 ***	0.369 ***	0.051 ***	0.087 ***	1.000	−0.015	−0.008	0.368 ***
Liquidity	−0.011	−0.008	−0.002	−0.001	−0.015	1.000	−0.003	−0.010
ROE	−0.011	−0.012	−0.001	0.001	−0.008	−0.003	1.000	−0.011
Expenses	0.724 ***	0.599 ***	0.093 ***	0.222 ***	0.368 ***	−0.010	−0.011	1.000

Source: Authors' calculations. As opposed to Table 3, the matrix is calculated on a complete-case basis using all eight variables at the same time. Therefore, every cell in the table contains the same sample size: $n = 10{,}386$, which is the complete-case sample size for current liquidity, ROE, and total expenses (see Table 3), because no observation had valid values for the first three of these variables but had missing or non-finite values for the other five variables. *** $p < 0.001$ (two-sided test). Liquidity and ROE are not significantly correlated with any of the other variables in the table at the 5% significance level.

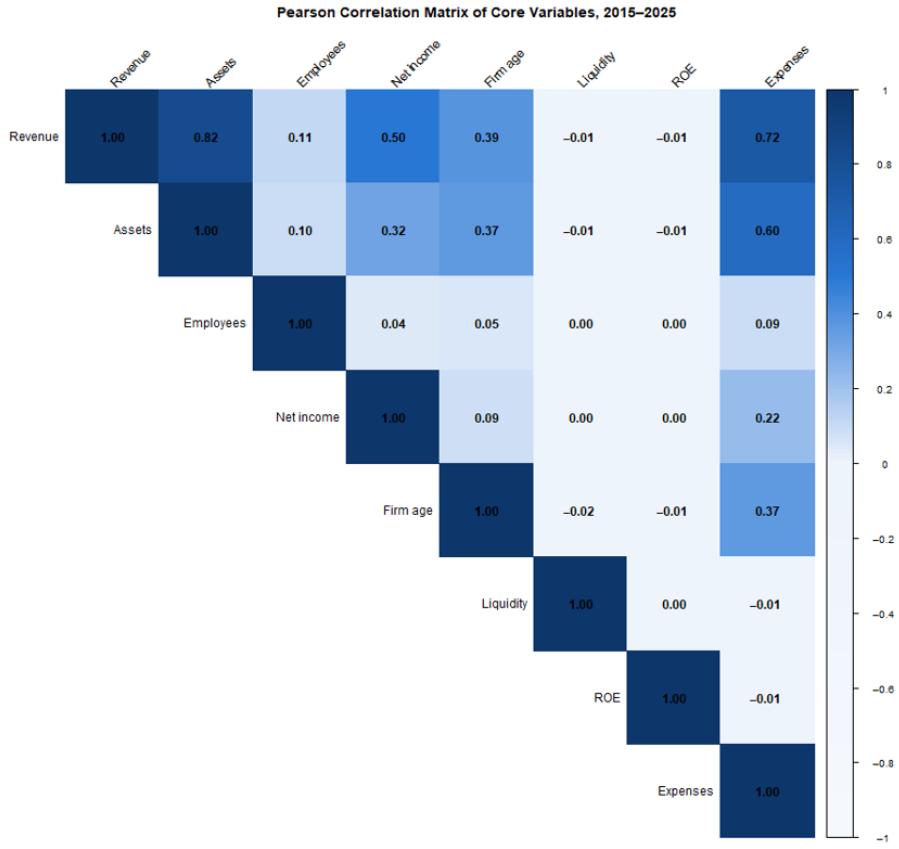


Figure 9. Correlation heatmap for the Pearson pairwise correlations between the eight core variables, complete-case subsample ($n = 10{,}386$; see Table 5 footnote). Intensity of color coding is based on the entire −1 to 1 range. The cells relating to liquidity and ROE are rather light-colored not due to the absence of negative values on the scale, but because these two variables happen to be weakly and in most cases negatively correlated with the remaining variables in the panel (Table 5), including negative correlations. Values correspond to Table 5 and are presented here only for visualization purposes, not independently obtained results.

A very strong positive correlation is present between revenue and assets ($r = 0.822$, $p < 0.001$), and an equally very strong positive correlation between revenue and total expenses ($r = 0.724$, $p < 0.001$); expenses were excluded from the regression not due to multicollinearity (which usually refers to the correlation between predictors, but not between the outcome and a predictor) but due to the structural and simultaneous determination of expenses and revenues in firm accounting. Revenue shows a rather positive correlation with the age of the firm ($r = 0.388$, $p < 0.001$) and net income ($r = 0.503$, $p < 0.001$), whereas liquidity and ROE have no significant correlation with revenue or with each other (all $|r| \leq 0.015$, $p > 0.10$).

4.6. Econometric Modeling

The multivariate log-log regression model with fixed year and city effects (Equation (1)) was estimated on 11,774 observations (Table 6).

Table 6. Baseline regression (Equation (1)): Key coefficients, cluster-robust standard errors (clustered by firm).

Term	Estimate	Std. Error	*t*	*p*-Value
Intercept	2.384	0.264	9.03	<0.001
ln(Assets)	0.815	0.016	49.45	<0.001
ln(Employees)	0.353	0.029	12.14	<0.001

Note: The model also includes 10 year dummies (2016–2025, base 2015) and 30 city dummies (base: "Other cities," a pooled category of cities with less than 30 modeling sample observations—see Sections 3.4 and 4.6). $R^2 = 0.619$; adjusted $R^2 = 0.617$; residual SE = 1.911; $n = 11{,}774$. Robust-regression (RLM) cross-check: ln(Assets) = 0.749, ln(Employees) = 0.353 ($n = 11{,}774$), roughly in line with the cluster-robust OLS estimates above, although the assets elasticity is somewhat less elastic (0.749 vs. 0.815) due to reduced outlier effect.

Both assets and employment are significantly positively related to revenue. The point estimates (with cluster-robust SEs clustered by firms) imply the following elasticities: 0.815 and 0.353 for assets and employment, respectively. That is, a 1% increase in assets would lead to roughly a 0.82% increase in revenue, and a 1% increase in employment to a 0.35% increase in revenue, ceteris paribus. As in the previous case, these elasticities should be considered merely as robust correlations, since the model is estimated using panel observational data without any identification procedure that excludes reverse causality (revenues could be used to increase assets and hiring). The firm-and-year fixed-effects regression in Section 4.6 (Robustness Check: Firm and Year Fixed Effects, below) provides an additional robustness test but not a causal identification method. The model has an excellent explanatory power ($R^2 = 0.619$; adjusted $R^2 = 0.617$), which is a great result for such a long-run and cross-firm model. A robust regression (RLM) test gives very similar estimates (0.749 for assets and 0.353 for employment; see Table 6). Variance inflation factors (Table 7) prove that there is no multicollinearity problem ($GVIF^{1/(2\cdot Df)}$ is 1.259 and 1.261 for ln(assets) and ln(employees), respectively, which is much lower than the standard 5–10. Year and city fixed-effects blocks also demonstrate low levels of $GVIF^{1/(2\cdot Df)}$: 1.004 and 1.001, respectively).

The city fixed effects are based on the reference group "Other cities". As stated in Section 3.4, this pool includes all the cities with less than 30 observations. Therefore, all the individual coefficients in Table 6 (in its full output form) represent the estimated differences in conditional revenues of these particular cities relative to the pooled group of low-activity cities. There are eight statistically significant city coefficients (at 5% significance level): Alangasí (0.856, $p = 0.001$), Cuenca (0.424, $p = 0.008$), Portoviejo (0.825, $p < 0.001$), Quevedo (0.571, $p = 0.012$), Quito (0.436, $p = 0.001$), Riobamba (0.405, $p = 0.038$), Santo Domingo de los Colorados (0.442, $p = 0.049$), and Tumbaco (0.626, $p = 0.014$). All of the firms located in

these cities show significantly higher revenues, taking into account assets, employment, and year, than firms in the pooled "Other cities" group. Despite Guayaquil being the second largest market of the sector by total revenues (see Section 4.3), it is not statistically different from the "Other cities" group (0.146, $p = 0.304$), along with 21 other cities. Again, the joint significance of the fixed effects blocks is not dependent on the choice of the reference group: the year block is jointly significant ($F(10, 11{,}731) = 9.76$, $p < 0.001$), which confirms H4; and the city block is jointly significant ($F(30, 11{,}731) = 4.10$, $p < 0.001$), which confirms H5. The two F-tests are performed on the nested OLS regressions, and thus are conventional (non-cluster-robust) tests. They are separate calculations from the firm-clustered SEs for the individual coefficients, which were already calculated, and thus are not recalculated with the firm-clustered covariance matrix. Diagnostic plots for the baseline model are shown in Figure 10.

Table 7. Variance inflation factors, baseline model.

Term	GVIF	Df	$\text{GVIF}^{1/(2\cdot\text{Df})}$
ln(Assets)	1.586	1	1.259
ln(Employees)	1.590	1	1.261
Year (10 dummies)	1.085	10	1.004
City (30 dummies)	1.080	30	1.001

Source: Authors' calculations. All the values are far below the standard concern value of 5–10 (comparison of $\text{GVIF}^{1/(2\cdot\text{Df})}$ to $\sqrt{\text{VIF}}$), hence no multicollinearity is present.

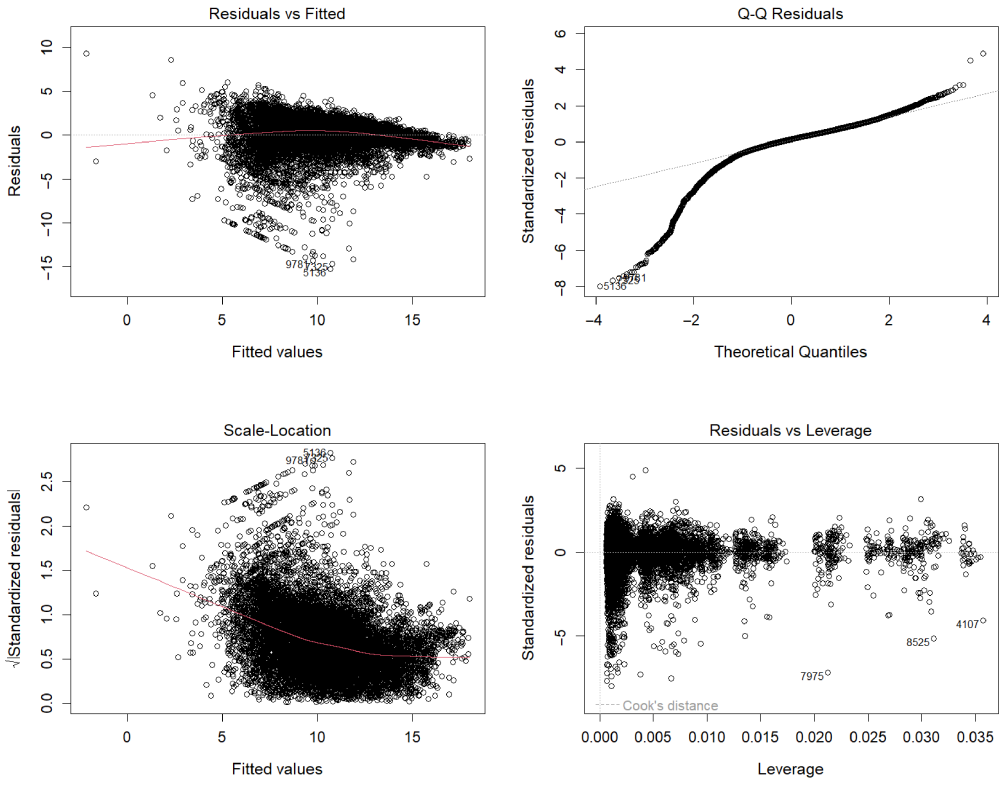


Figure 10. Diagnostic plots for the baseline model (Table 6), final estimation sample ($n = 11{,}774$): residuals vs. fitted values, normal Q-Q plot, scale-location plot, and residuals vs. leverage. Red lines are smoothed (LOWESS) trend curves; gray dotted lines are reference lines (zero residual and, in the Q-Q plot, the theoretical normal line); the gray dashed line labeled "Cook's distance" marks Cook's distance contours; numbered points identify the observations with the largest absolute residuals.

As expected for the financial data of the firms, the diagnostic plots demonstrate heteroskedasticity and a moderate departure from a normal distribution in the tails of the

residuals. There are also several high leverage points present. Taking into account the good agreement of the results obtained with OLS and RLM regressions, inference on individual coefficients in this paper is made using cluster-robust SEs and fixed effects (year and city), not classical OLS SEs. The joint F-tests for the year and city blocks (H4 and H5, above) remain conventional (non-cluster-robust) tests, as indicated above.

Robustness Check: Firm and Year Fixed Effects

Whereas cluster-robust standard errors modify the variance of the error term in order to accommodate firm-level correlation in time-series data, they do nothing to account for time-invariant firm characteristics that could correlate with both firm size and revenue (e.g., management quality, reputation, network effects). Equation (1) was therefore estimated again with firm fixed effects, in addition to year fixed effects, such that the identification of assets and employment coefficients relied exclusively on within-firm variation (Table 8). City fixed effects were not estimated alongside firm fixed effects, because city will be fixed for most firms over time and is thus contained in the firm effect.

Table 8. Firm-and-year fixed-effects model (within-firm estimates), cluster-robust standard errors by firm.

Term	Estimate	Std. Error	*t*	*p*-Value
ln(Assets)	0.797	0.034	23.28	<0.001
ln(Employees)	0.163	0.032	5.12	<0.001

Note: Model includes firm fixed effects and year fixed effects; city fixed effects are not estimated, because city is absorbed in the firm effect for most firms. n = 11,048 observations (726 fewer than the pooled model, because firms observed in only one year add no within-firm variation and are thus excluded automatically); within R^2 = 0.260. Out of the 2925 firms in the full modeling sample, 2199 (75.2%) appear in multiple years, contributing within-firm variation to the assets and employment coefficients above; the remaining 726 (24.8%) appear in only one year and are thus completely absorbed by their own firm fixed effect, adding no within-firm identifying variation.

Both coefficients remain positive and highly significant in within-firm regressions. The coefficient on assets (0.797) is quite similar to the pooled estimate from Table 6 (0.815), suggesting that there is a strong assets-revenue relationship even without accounting for any firm-time invariant differences between firms. The coefficient on employment, however, drops from 0.353 in the pooled model to 0.163 within firms—over half the value—implying that the pooled employment coefficient includes a large portion of stable firm differences (more-established firms tend to have more employees and also more revenue), rather than being driven by any within-firm effect of hiring on revenue. In other words, an increase in a firm's own asset base is associated with an increase in its revenue at the estimated elasticity of 0.797 (very similar to 0.815), whether comparing different firms or the same firms at different times. An increase in a firm's own employment is also associated with an increase in its revenue, but the within-firm elasticity of 0.163 is considerably smaller than the cross-firm elasticity of 0.353 would suggest. Assets are thus a robust correlate of revenue both between and within firms, while employment is more strongly correlated with revenue differences between firms than with revenue changes within firms over time. Table 9 compares the fit statistics of the three specifications.

4.7. Firm Segmentation

Homogeneous groups of firms in the 2025 cross-section (n = 2056 firms with complete clustering variables) were determined through k-means clustering. The elbow plot (Figure 11) indicates a rapid decrease in total within-cluster variance through k = 5 and then leveling off sharply after; the silhouette plot (Figure 12) has its highest average width

around $k = 2$ ($\approx$0.90) and then drops sharply afterwards, remaining well below its peak for all $k \geq 3$, although with some fluctuations at higher k, which does not indicate clearly any k in the $k = 2$–4 range considered below. Due to the inconsistency between criteria, further internal validity indicators were evaluated prior to determining the specification.

Table 9. Model fit comparison across specifications.

Model	*n*	R^2	Adj. R^2	Residual SE
OLS, cluster-robust SE (Table 6)	11,774	0.619	0.617	1.911
Robust regression (RLM)	11,774	—	—	1.170
Firm + Year FE, within (Table 8)	11,048	0.260 (within)	—	—

Source: Authors' calculations. There is no conventional R^2 statistic for the RLM specification; the low residual standard error reflects the outlier downweighting built into M-estimation rather than a direct indicator of fit comparable to the OLS.

Table 10. Cluster validity criteria for $k = 2$–4, 2025 cross-section ($n = 2056$ firms).

k	Silhouette	Calinski–Harabasz	Davies–Bouldin	Gap Statistic
2	0.897 (max, $k = 2$–4)	996.9	0.863 (min, $k = 2$–4)	2.85
3	0.717	1085.1 (max, $k = 2$–4)	0.889	3.00
4	0.686	1050.1	1.016	3.04

Source: Authors' calculations. Both silhouette and Davies–Bouldin favor $k = 2$ (the highest value of silhouette, the lowest value of Davies–Bouldin); Calinski–Harabasz favors $k = 3$ within the 2–4 range. There is no single value of k favored by all criteria at the same time, which is why the three-cluster solution is chosen substantively, rather than statistically (Section 4.7).

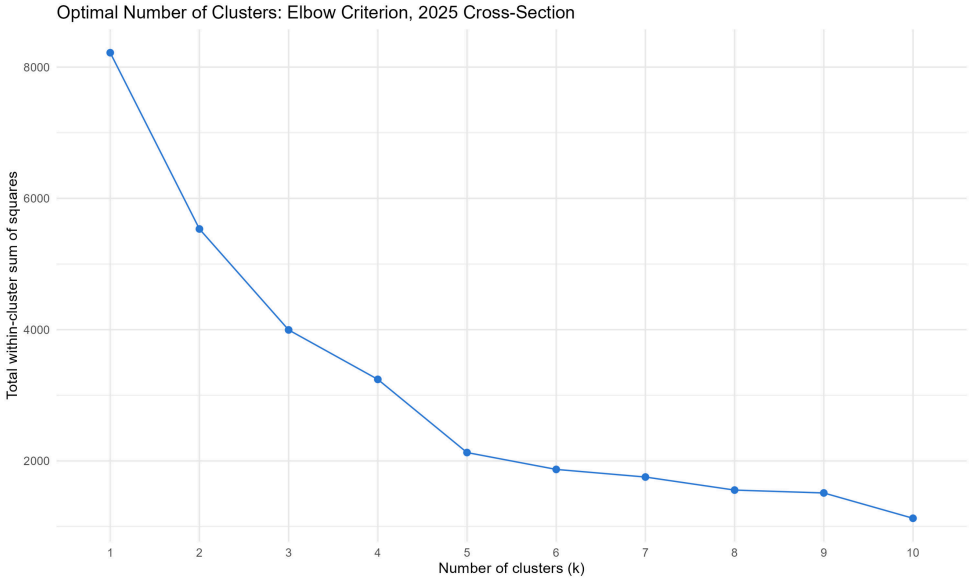


Figure 11. Elbow plot (total within-cluster sum of squares vs. $k = 1$–10), k-means clustering on the standardized 2025 cross-section ($n = 2056$ firms with complete clustering variables). Points show the total within-cluster sum of squares for each k, connected by a line. Within-cluster variance falls off rapidly to $k = 5$ and plateaus considerably from then onwards. Such behavior of the criterion does not unambiguously indicate $k = 3$, and the range $k = 2$–4 considered in Table 10 is justified more by the substantive "large/medium/small" typology outlined in this section than this graph alone.

Two additional measures of internal validity—the Calinski–Harabasz and Davies–Bouldin indices—along with the gap statistic, were calculated for $k = 2$ to $k = 4$. (Table 10. These same criteria will be extended to $k = 5$–8 at the end of this section as part of the robustness test). There is no single dominant choice within this range: the silhouette and

Davies–Bouldin indices both favor $k = 2$ (the highest value of silhouette and the lowest value of Davies–Bouldin indices), whereas the Calinski–Harabasz index is maximized at $k = 3$ within the range of $k = 2$–4 examined. The gap statistic (Figure 13) grows monotonically within this range (2.85 at $k = 2$, 3.00 at $k = 3$, 3.04 at $k = 4$) without exhibiting an elbow at $k = 2$–4, hence failing to favor one of these three choices by itself, just as expected for the highly right-skewed nature of firm size distribution rather than due to a few distinct discrete clusters.

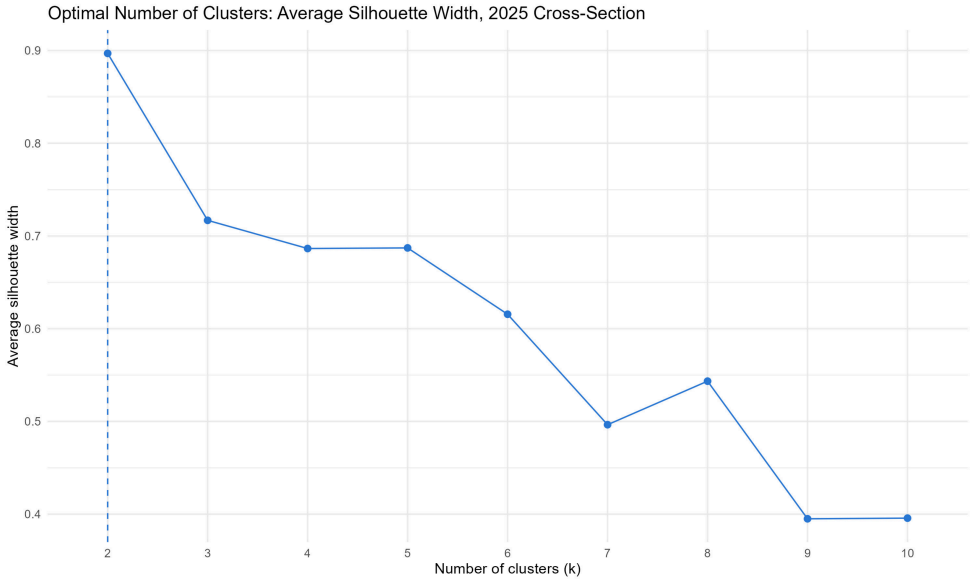


Figure 12. Average silhouette width vs. $k = 2$–10, same sample as Figure 11. Points show the average silhouette width for each k, connected by a line; the vertical dashed line marks the k with the maximum average silhouette width ($k = 2$). Maximum (average width ≈ 0.90) is achieved at $k = 2$, dropping very sharply thereafter and remaining considerably below that maximum level through $k = 10$, with some oscillation at higher values of k; refer to Table 10 for the full set of validity criteria over $k = 2$–4.

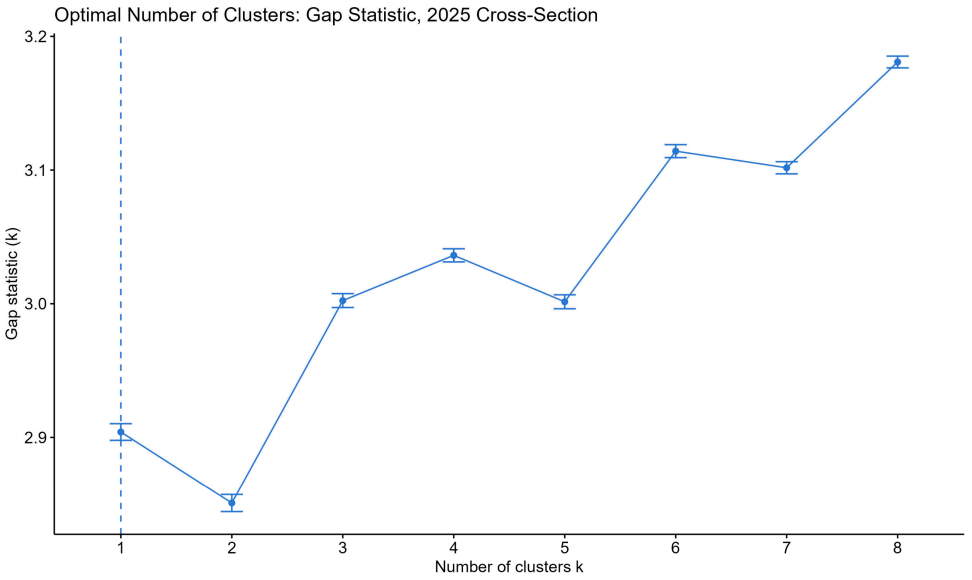


Figure 13. Gap statistic for $k = 1$–8 (1-SE band), same clustering sample as Figures 11 and 12. Points show the gap statistic for each k, with error bars of ± 1 SE, connected by a line; the vertical dashed line marks the k selected by the default "firstSEmax" rule of factoextra ($k = 1$), which lies outside the $k = 2$–4 range compared in the text. The gap statistic is strictly increasing from $k = 2$ to 4 and does not produce an "elbow" and hence does not indicate a preferred value of k in that range (Section 4.7).

The stability of each candidate solution was additionally evaluated using bootstrapping (100 resamples, average Jaccard similarity through fpc::clusterboot; Table 11). In each of the considered solutions, the largest cluster, made up of the sector's small firms, is highly stable (Jaccard ≥ 0.98, never dissolves). Meanwhile, the smallest cluster, made up of several very large firms, is relatively unstable in each k (Jaccard between 0.63 and 0.88, dissolves in 8 to 48 of 100 resamples), purely as a function of being based on very few observations (as few as 9 firms at $k = 3$) rather than being a weakness of any particular solution. This pattern—a highly stable majority cluster and an intrinsically less stable elite cluster—is true irrespective of which k is chosen.

Table 11. Bootstrap cluster stability (100 resamples, mean Jaccard similarity), $k = 2$–4.

k	Cluster	Mean Jaccard	Dissolved in n/100 Resamples
2	1 (large group)	0.988	0
2	2 (small group)	0.625	48
3	1 (medium)	0.870	2
3	2 (large)	0.744	21
3	3 (small)	0.982	0
4	1	0.672	23
4	2	0.705	40
4	3	0.878	8
4	4 (small)	0.984	0

Source: Authors' calculations (fpc::clusterboot). In each k considered, the largest, small-firm cluster is highly stable (Jaccard ≥ 0.98), while the smallest, large-firm cluster is relatively unstable and dissolves in a significant number of resamples, purely as a function of that cluster being made up of very few observations (9 firms at $k = 3$), not a weakness of that particular solution.

The three-cluster ($k = 3$) solution was selected as the main outcome, not due to it statistically outperforming $k = 2$ or $k = 4$ in all of the criteria, as it does not, but because it fits the motivationally-driven typology of large/medium/small introduced in Section 2 and preserves distinctions between medium and small firms lost in the statistically preferred $k = 2$ solution. This selection is conscious and purposeful: $k = 2$ and $k = 4$ are both valid statistical solutions and are both included, along with $k = 3$, in Tables 10 and 11 for clarity. The resulting distribution of firms across the three clusters is reported in Table 12.

Table 12. Firm distribution by cluster, three-cluster ($k = 3$) solution, 2025 cross-section.

Cluster	Firms	% of Sample
1—Medium firms	250	12.16%
2—Large firms	9	0.44%
3—Small firms	1797	87.40%

Source: Authors' calculations, $n = 2056$ firms. Cluster labels (1 = medium, 2 = large, 3 = small) follow the arbitrary numbering of the k-means algorithm in this run; they are assigned by inspecting the cluster means (Table 13).

Cluster 3 (small firms) accounts for 1797 firms (87.40%) of the 2025 clustering sample; Cluster 1 (medium firms) includes 250 firms (12.16%), while Cluster 2 (large firms) consists of 9 firms (0.44%). This labeling corresponds to the internal numbering of k-means for this particular run; since k-means numbers are arbitrary, they cannot be interpreted in isolation from the cluster means. The labels "small," "medium," "large" and the percentages given above characterize a data-driven four-variable clustering of the 2025 cross-section, but not the firm size classification (see Conclusions). Figure 14 displays the standardized profile of the three clusters, and Tables 13 and 14 report the corresponding cluster means and medians in original units.

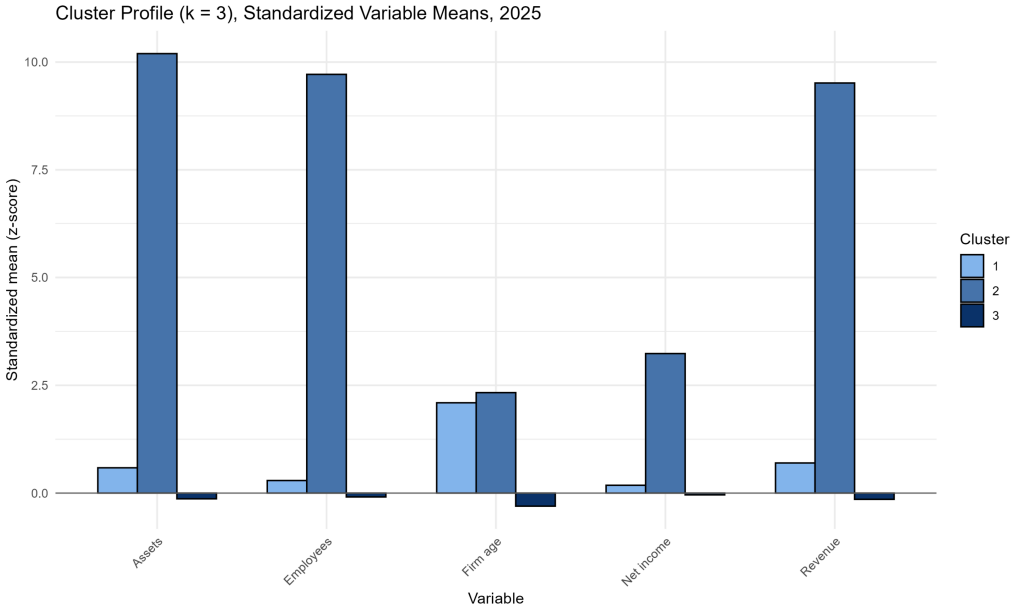


Figure 14. Financial and operational profile of the $k = 3$ solution, 2025 cross-section (n = 2056 firms). The bars are the means of the respective variables at the cluster level expressed in terms of z-scores with respect to the clustering sample. Positive and negative values show whether the cluster mean is higher or lower than the mean of the whole clustering sample. Four clustering variables (revenue, assets, employees, firm age) were used to calculate the clusters, and the net income is presented for descriptive purposes only. Large standardized values for Cluster 2 (approximately 10 SD for assets, employees, and revenue) result from the extreme right-skewness of the firm size distribution described in Table 3 concentrated in the 9 firms of that cluster, not from any numerical problems. The values in the underlying Figure 14 are presented in the directly interpretable raw form in Table 13.

Table 13. Cluster means (revenue, assets, employees, net income, firm age), $k = 3$, 2025 cross-section.

Cluster	Firms	Mean Revenue	Mean Assets	Mean Employees	Mean Net Income	Mean Age (yrs)
1—Medium	250	$1,113,895	$1,535,482	36.3	$49,384	23.2
2—Large	9	$11,454,095	$20,992,333	793.4	$721,339	25.1
3—Small	1797	$124,455	$77,772	5.7	$410	3.9

Net income is given as a separate variable column.

Table 14. Cluster medians, $k = 3$, 2025 cross-section.

Cluster	Firms	Median Revenue	Median Assets	Median Employees	Median Net Income	Median Age
1—Medium	250	$417,785	$355,890	12	$3789	20
2—Large	9	$8,484,830	$20,768,720	467	$1,179,613	21
3—Small	1797	$21,375	$10,808	4	$90	3

Medians corroborate the pattern observed in the means and indicate that Cluster 3 (small firms) is not an artifact of outliers pulling up the mean.

Similarly to the means, medians corroborate the pattern indicated by the means. The low revenues of Cluster 3 are not an artifact of the few outliers pulling the mean to higher values (median revenue is only $21,375 versus mean $124,455, the difference justified by the right skewness even within this "small firms" subgroup). The statistically preferred two-cluster solution divides 29 firms (1.41%) from 2027 firms (98.59%). The former group has a mean revenue of $6,825,779 (mean assets $12,514,304; mean employment 332.1; mean firm age 29.7 years), whereas the latter group has a mean revenue of $200,917 (mean assets $172,493; mean employment 8.3; mean firm age 6.0 years). Substantively, the solution reproduces the similar large-versus-rest pattern as the $k = 3$ solution, though the intermediate

group of firms of Cluster 1 is merged with the large group—hence, the mean revenue of the 29-firm $k = 2$ "large" group is lower than the mean revenue of the smaller 9-firm $k = 3$ Cluster 2 ($11,454,095; see Table 13). The two solutions simply place the boundary of the "large firms" category differently in response to the changing number of clusters, k. Finally, the four-cluster ($k = 4$) solution, which is reported among the $k = 2$–4 robustness checks (Tables 10 and 11), divides the sample into a 1722-firm group (83.75%), a 294-firm group (14.30%), a 34-firm group (1.65%) and a 6-firm group (0.29%). The smallest, 6-firm group reports the highest mean revenue of the four groups ($12,419,530), which supports the view that the $k = 4$ solution subdivides rather than contradicts the large-firm segment already identified at $k = 2$ and $k = 3$.

As another robustness check, the cluster-validity comparison above was additionally run beyond the initial $k = 2$–4 range: The Calinski–Harabasz and Davies–Bouldin indices were calculated for $k = 5$–8 using the same procedure as in Table 10, and the k-means solution was separately estimated at $k = 5$ using the same four standardized clustering variables, scaling, seed, and nstart as the $k = 2$–4 solutions (see Tables 15 and 16). At this wider range of k, both the Calinski–Harabasz and Davies–Bouldin indices achieve their peaks at $k = 5$ (Calinski–Harabasz = 1468.3, versus 1085.1 at $k = 3$; Davies–Bouldin = 0.783, versus 0.889 at $k = 3$; Table 16), thus reversing on the above two criteria alone the preference expressed for the narrower range of $k = 2$–4. However, three additional checks speak against the adoption of $k = 5$ on this criterion alone. First, a bootstrap-stability analysis (100 resamples, using the same procedure as in Table 11) indicates the markedly lower stability of the five-cluster solution in comparison to the three-cluster one. Four out of the five clusters dissolve in 13–43% of resamples (Jaccard 0.60–0.72), whereas one cluster is comparably stable (Jaccard = 0.95, 0% dissolve)—which contrasts with the maximum dissolve rate of 21% among the three $k = 3$ clusters (see Table 11). Second, the $k = 3$ "large firms" cluster (9 firms) does not correspond to any of the clusters at $k = 5$. It splits into three of the five $k = 5$ clusters (4, 2, and 3 firms respectively), with no dominant destination. Also, two of the five $k = 5$ clusters are extremely small (4 and 2 firms, respectively), each seemingly consisting of one or two outliers rather than of a generalized firm type (Table 15). Third, the sensitivity analysis based on the log1p-transformation of revenue, assets, and employment reveals only weak agreement between the raw-scale and transformed solutions at either k (Adjusted Rand Index of 0.181 at $k = 3$ and 0.104 at $k = 5$). Furthermore, the agreement decreases, rather than increases, at the finer partition. Overall, the above extended comparison demonstrates that $k = 5$ remains genuinely competitive, and is even superior according to Calinski–Harabasz and Davies–Bouldin criteria specifically, yet is less stable in resamples and more sensitive to the choice of variable scaling—what is consistent with the extreme right skew of firm size in this sector (Table 3), not with the clear separation into five groups. Hence, $k = 3$ is selected as the primary solution for the combination of substantive interpretation and comparative bootstrap stability, whereas the whole extended comparison is provided for clarity.

PCA projection (Figure 15) reveals the presence of a separate group of large firms along the first principal component. The first two principal components account for 76.6% of the total variance of the four standardized clustering variables (56.5% for PC1 and 20.1% for PC2).

Table 15. Robustness check: Cross-tabulation of firm membership, $k = 3$ vs. $k = 5$ solutions, and sensitivity of cluster assignment to logarithmic transformation, 2025 cross-section.

$k = 3$ Cluster	$k = 5$:1	$k = 5$:2	$k = 5$:3	$k = 5$:4	$k = 5$:5
1—Medium ($n = 250$)	0	218	1	0	31
2—Large ($n = 9$)	4	0	0	2	3
3—Small ($n = 1797$)	0	76	1721	0	0

$k = 5$ estimated on the same four standardized clustering variables, scaling, seed and nstart as the $k = 2$–4 solutions in Table 10. The $k = 3$ "large firms" cluster (9 firms) does not map into a single $k = 5$ cluster. It splits across three of the five $k = 5$ clusters (4, 2 and 3 firms respectively), with no dominant destination. Two of the five $k = 5$ clusters are themselves extremely small ($n = 4$ and $n = 2$ firms); the $n = 2$ cluster in particular has a mean of 2228 employees, versus the mean revenue below that of the $n = 4$ cluster, which suggests isolating one or two firms with an atypical employee-to-revenue ratio rather than describing a generalizable firm type. Also, a parallel bootstrap-stability analysis (100 resamples via fpc::clusterboot, using the same procedure as Table 11) indicates the marked lower stability of the five-cluster solution in comparison with the three-cluster one. Four of the five clusters dissolve in 13–43% of resamples (Jaccard 0.60–0.72), whereas one cluster is comparably stable (Jaccard = 0.95, 0% dissolve)—versus the maximum dissolve rate of 21% among the three $k = 3$ clusters (see Table 11). Finally, a complementary sensitivity analysis consists in clustering on log1p-transformed revenue, assets and employment, and comparing, separately at each k, the obtained partition to the corresponding raw-scale solution via the Adjusted Rand Index (ARI = 1: identical partitions; ARI near 0: agreement no better than random chance; this ARI does not compare the $k = 3$ and $k = 5$ partitions with each other): ARI = 0.181 at $k = 3$ and 0.104 at $k = 5$, indicating the only weak raw-vs-log agreement at both k, and worsening, rather than improving, at the finer partition.

Table 16. Extended cluster-validity criteria, $k = 5$–8, 2025 cross-section ($n = 2056$ firms).

k	Silhouette	Calinski–Harabasz	Davies–Bouldin	Gap Statistic
5	0.687	1468.3 (max, $k = 2$–8)	0.783 (min, $k = 2$–8)	3.00
6	0.616	1391.7	0.853	3.11
7	0.496	1258.5	0.916	3.10
8	0.543	1253.5	0.833	3.18

Same procedure as in Table 10 (silhouette and WSS computed via an explicit k-means loop with a fixed seed and nstart = 25 for each k, for full run-to-run reproducibility; fpc::calinhara for Calinski–Harabasz; a custom implementation for Davies–Bouldin; clusGap for the gap statistic), an extension of the $k = 2$–4 range from Table 10 to $k = 5$–8. In this extended $k = 2$–8 range, both Calinski–Harabasz and Davies–Bouldin achieve their optimum at $k = 5$ (1468.3 and 0.783 correspondingly), which are better than the $k = 2$–4 optima from Table 10 (Calinski–Harabasz = 1085.1 at $k = 3$; Davies–Bouldin = 0.863 at $k = 2$)—based on those two criteria alone, $k = 5$ is better than $k = 3$. Silhouette and the gap statistic do not prefer $k = 5$: silhouette drops sharply after reaching its maximum value at $k = 2$ (0.90) and then consistently remains low, even not monotonically decreasing (0.69 at $k = 5$, dropping to 0.50 at $k = 7$ before rising to 0.54 at $k = 8$); gap statistic drops a little at $k = 5$ (3.00, vs. 3.04 at $k = 4$) before rising at $k = 6$–8, again showing no obvious elbow in this extended range. That better single-fit compactness/separation profile for $k = 5$ is counter-balanced in Section 4.7 against the bootstrap-stability and log-scale sensitivity evidence from Table 15, which suggests not adopting $k = 5$ as the primary solution.

4.8. Integration of Results

Based on the descriptive statistics (Table 3), correlation analysis (Table 5), and $k = 3$ cluster profiles (Tables 12–14), we can draw conclusions about the presence of one consistent structure. The differentiation of firms in the education sector takes place mainly along the axis of firm size (as measured by assets and employment). Cluster 2 (0.44% of firms) represents the group with unusually high average revenue, assets, employment, and net income, compared to its relatively small number; Cluster 3 (87.40% of firms) operates on the smallest financial scale and has, significantly, the lowest average firm age (3.9 years); and Cluster 1 (12.16% of firms) operates on the intermediate financial scale despite having an average firm age (23.2 years) similar to Cluster 2 (25.1 years). Thus, as discussed in Section 5.5, age is not automatically associated with scale.

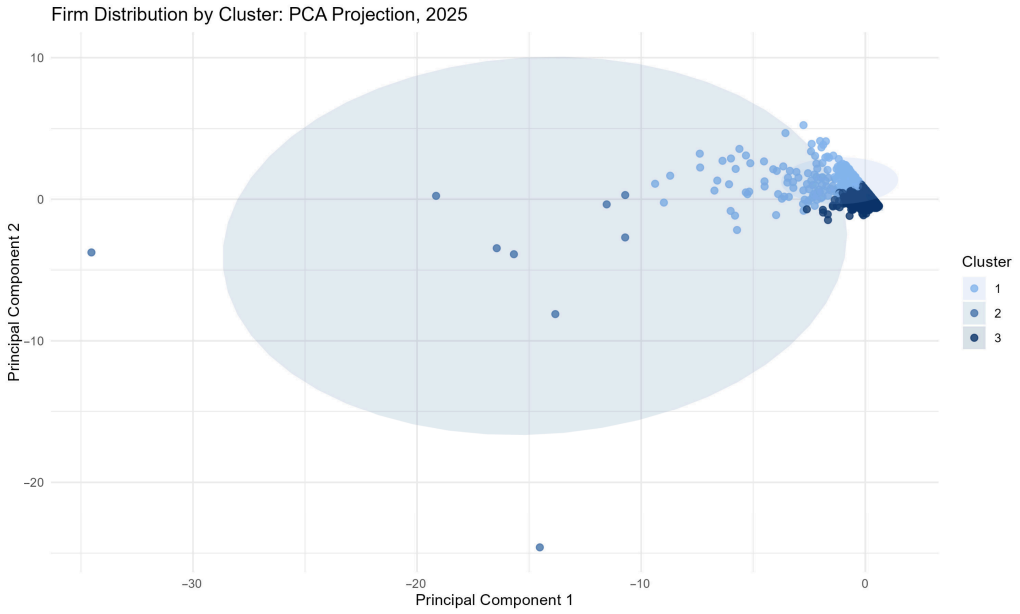


Figure 15. Distribution of firms by $k = 3$ cluster in the first two principal components of four standardized clustering variables (56.5% of variance is explained by PC1, 20.1% by PC2; together 76.6%). Shaded area shows the 95% confidence ellipse of each cluster (ggplot2's stat_ellipse, assuming the multivariate t-distribution), not the convex hull of the raw data.

5. Discussion

The combination of the cluster means and the regression results suggests that firm size is the principal axis of differentiation in Ecuador's education sector, together with time and geographic heterogeneity. This asymmetric structure, involving a highly right-skewed distribution of small firms and a few extremely large firms, fits into the wider framework of the documented right-skewed size distributions (Gabaix, 2009), although this paper did not test it for the power-law functional form. It is also consistent with the well-known pattern of younger firms being smaller and growing faster and more variably than mature incumbents, visible in the profile of Cluster 3 (Evans, 1987; B. H. Hall, 1987; Sutton, 1997). This is further confirmed by the regression results, where assets and employment are two variables most correlated with revenue generation, together with statistically significant year and city effects. Thus, while firm size and operational capacity are the main correlates of revenue generation in Ecuador's education sector, there are other factors as well.

5.1. Firm Size and Resource-Based Correlates of Revenue

Positive and statistically significant correlations of assets and employment with revenues of firms confirm and generalize the theoretical considerations presented in Section 2. At first, such correlations are in line with the theory that sees resources and production function as key factors of firm's revenues and competitive advantage (Nelson & Winter, 1982; Syverson, 2011), similarly to the traditional approach that treats firm size as an important predictor of profits (M. Hall & Weiss, 1967; Penrose, 1959). The robustness specification presented in Section 4.6 (Robustness Check: Firm and Year Fixed Effects) proves that this interpretation is justified, since the correlation remains significant when firm-fixed effects are included in the regression (elasticities of 0.797 for assets and 0.163 for employment). It is true despite the fact that observational nature of the study prevents drawing causality and establishing economies of scale from production economics point of view. In other words, even in a service-based industry where assets and employment scale are very important, the

association between revenue performance and the scale of assets and employment persists and has not been shown before for the case of the Ecuadorian education sector.

5.2. Temporal Dynamics and the COVID-19 Shock

Secondly, in support of H4's temporal component, the structure of the sector—revenue growth from $322.4 million (2015) to $605.2 million (2025), with a clear contraction of revenue in 2020 due to the pandemic and then recovery—seems to be a typical pattern of post-shock recovery. The observational nature of the data prevents a causal statement about the influence of the shock on this pattern. One of the noteworthy features of the sector structure is a stable growth in the number of firms for which there were regulatory records available during the entire study period, including the one with the revenue contraction in 2020. This pattern indicates expansion in the observed firm coverage of the sector, but, as has already been pointed out in Section 4.2, such a trend cannot by itself distinguish firm entry from the changes in the level of reporting compliance or registry coverage and should not be treated as direct evidence of firm formation or sector formalization. Broad literature on firm-level dynamics in Latin America shows that aggregate output decline did not imply a decline in firm count (Bergoeing & Repetto, 2006), while structural transformation literature associates relocation of economic activity to formal firms and thus aggregate productivity growth (Lavopa, 2015; McMillan et al., 2014; Pagés, 2010b; Perry et al., 2007). However, the regulatory records used in this paper do not provide evidence of whether the mechanisms found in this literature underlie the growth in firm counts in this sector.

This pattern of the sector is additionally confirmed by recent research on the differential effect of the coronavirus-induced lockdowns on private and public sectors of education in Latin America. Analyzing cross-nationally harmonized household survey data for thirteen Latin American countries, Bracco et al. (2025) demonstrate that COVID-19 school closures led to a disproportionate reduction in private-school enrollment relative to public education, mostly due to a decline in household incomes required for financing private schooling. This provides a plausible explanation of the observed contraction of revenue in this study from 2019 to 2020 and the subsequent recovery until 2023–2025, caused not only by the resumption of in-person teaching but also by the gradual restoration of households' ability to afford it. Another plausible mechanism may include the change in the composition of providers and demand in market segments following the shock, but the available data prevent direct testing.

5.3. Territorial Dynamics and Agglomeration

Finally, although the descriptive and correlational analysis shows high territorial concentrations of economic activity in Pichincha and Guayas, accounting for most sector revenue in 2025, the regression results demonstrate that city fixed effects cannot be considered as control variables for unobserved heterogeneity. As is stated in Section 4.6, the standard joint F-test of city effects is highly significant ($F(30, 11{,}731) = 4.10$, $p < 0.001$), which supports H5. Individual city fixed effects are estimated relative to the pooled "Other cities" baseline, see Section 4.6. Eight cities—Alangasí, Cuenca, Portoviejo, Quevedo, Quito, Riobamba, Santo Domingo de los Colorados, and Tumbaco—show higher conditional revenue than "Other cities," while despite the fact that Guayaquil has a high aggregate market size, it does not have a statistically significant fixed effect. This means that despite the fact that conditional and aggregate revenue differ from each other, the conditional revenue differences persist after controlling for firm size and year. This result is in line with the cross-country evidence indicating that firm-level performance depends on local investment climate and infrastructure even after controlling for firm characteristics (Dollar et al., 2005). This result has some implications for regional development practice, although the current cross-sectional design

of the study with city fixed effects does not allow us to determine the exact local factors behind the conditional revenue differences and the effect of specific policy interventions.

Evidence of country-specific agglomeration supports the territorial interpretation of the result in general. Analyzing economic census data for Ecuador, Torres-Gutiérrez et al. (2019) find that urbanization increases the productivity of industries and services, disproportionately concentrated in provincial capitals due to public investment, market access, and labor-market specialization. The significantly higher level of revenue for firms located in Quito, Cuenca, Portoviejo, and other cities relative to the "Other cities" baseline is consistent with this evidence of agglomeration, although the lack of a statistically significant coefficient for Guayaquil despite its large aggregate market size is a reminder that the conditional, per-firm revenue and aggregate market size are different empirical issues.

5.4. Liquidity, Profitability, and the Resource-Based View

Finally, revenue is strongly correlated with assets ($r = 0.822$) and even more strongly, concurrently, with total expenses ($r = 0.724$). It should be noted (as explained in Section 4.5) that this is not multicollinearity in a technical sense because revenue is the dependent variable in this case, but expenses were omitted from the regression due to the structural, simultaneous relation with revenue, which could cause endogeneity problems in the specification. On the other hand, the low or close to zero correlation of revenue with liquidity and ROE ($|r| \leq 0.011$, statistically insignificant) shows that, in this case, scale-related variables are significantly more strongly concurrent with revenue than these short-term financial ratios.

This pattern of correlation is consistent with the resource-based approach distinction between tangible and intangible resources (Barney, 1991; Newbert, 2008). Liquidity and ROE are short-term financial ratios which could vary significantly within one period and are not related to the resource endowment of the firm, in contrast to assets and employment, which can be considered in this case as tangible physical and human capital and thus explain differences in performance of firms within the industry. The lack of correlation between short-term financial ratios and revenue does not mean that financial management is not relevant to education sector. It only means that in this dataset, revenue is not correlated with these specific financial ratios. The descriptive statistics imply that scale-related variables are much more strongly concurrent with revenue than liquidity or profitability ratios, which confirms the greater relevance of resource-based theory for this case than short-term financial ratios.

Asymmetry of the firm-size distribution documented in Section 4.7 has implications that exceed the segmentation analysis itself. The combination of high right-skewness of the distribution of small firms and the cluster of very large firms is consistent with the right-skewed firm-size distributions documented in industry and cross-country studies (Gabaix, 2009), even though no power-law is tested here empirically. However, given the characteristics of the education sector, this distribution raises normative considerations. An industry with 9 firms (0.44% of 2025 clustering sample) occupying such a disproportional amount of assets and revenue in contrast to their number raises concerns about market power and entry barriers for the overwhelming majority of small firms in Cluster 3. In order to address these questions, additional information is required. However, the results of this segmentation analysis provide an empirical basis for further research on competition policy in Ecuador's education sector.

5.5. Firm Segmentation and Structural Heterogeneity

Fifth, k-means clustering allows us to identify three firm segments with quite different characteristics. The high domination of small-scale firms (87.40% of 2025 clustering sample, average age is about 3.9 years) and the rarity of the large-firm segment (0.44%) show high

polarization of the market structure in this sector. Such high polarization is consistent with many studies of enterprise performance in Latin America and Caribbean region, in which firm size and capacity are indicated as one of the main sources of structural heterogeneity (Grazzi et al., 2016). Also, it is consistent with the international experience of the education sector, where firm size is recognized as a major source of heterogeneity (Crawfurd et al., 2024). In general, there is extensive international evidence of lack of scaling in the education sector, where the majority of schools and educational institutions are small-scale firms (Andrews, 2016; Crawfurd et al., 2024; Elacqua et al., 2011; Hutchings & Francis, 2017; Kingdon, 2020; Smith & Baker, 2017).

Such a lack of scaling is also characteristic for countries experiencing rapid growth and formalization of education sectors, such as India, with explosive growth in the number of schools but a permanent lack of large-scale firms (Kingdon, 2020). A similar lack of performance gains associated with firm size is observed in the cases of school chains and academies in the UK and Chile (Andrews, 2016; Elacqua et al., 2011; Hutchings & Francis, 2017). Private investors' interest in the education sector usually focuses on established, large firms (Mundy & Menashy, 2014), which is consistent with this study's result of revenue and asset concentration among the large firms of Cluster 2—even though the interest of private investors in this sector is growing more and more (Assomull et al., 2015; Featherston & Ferreira, 2017; Kenny et al., 2018), and a lot of literature is devoted to the shift towards fee-based private schooling in low-income countries (Smith & Baker, 2017).

However, there is no criterion that would clearly favor the three-cluster solution. At the first $k = 2$–4 comparison stage, both the silhouette and Davies–Bouldin indices are optimal for $k = 2$, while the Calinski–Harabasz index suggests $k = 3$ (Section 4.7, Table 10). The three-cluster solution is chosen here not because of any dominance in the statistical sense but due to the alignment with the large/medium/small typology, suggested in Section 2. This solution maintains the separation between medium-sized and small firms, while the $k = 2$ solution does not (this is a substantive, not statistical choice). As a robustness check, the three-cluster solution is compared to $k = 5$–8 and finds the Calinski–Harabasz and Davies–Bouldin indices at their maxima for $k = 5$, not $k = 3$ (Section 4.7, Table 16). For these two criteria specifically, the more fine-grained solution looks more favorable. In contrast to this, the bootstrap-stability analysis finds the five-cluster solution to be less stable and the five-cluster solution contains two too-small clusters (4 and 2 firms) that do not form one cluster at $k = 3$ (Table 15). Additionally, another robustness check performed through the log-scale sensitivity analysis shows that cluster membership for $k = 3$ and $k = 5$ depends significantly on whether the variables (revenue, assets, employment) are standardized to their original or log1p-transformed scale (adjusted Rand Index is 0.181 and 0.104 respectively). The three-cluster solution is relatively, not absolutely, more robust to this change than the five-cluster solution, and both solutions cannot be treated as invariant to the transformation. All in all, the three-cluster solution is retained for its interpretation and higher stability in resampling, but not for being preferred by every criterion or for having the cluster membership immune to the variable scaling.

The stability of this structure is consistent with the empirical evidence on the productivity gap between SMEs and large firms in Latin America. Based on the evidence presented by Ibarrarán et al. (2009), this productivity gap is caused not by the size of the firms themselves, but by inequality in access to credit, training, and technology, which suggests that generic policies for supporting SMEs have low potential for closing this gap. The prevalence of small firms (87.40% of this study's cross-section of 2025 firms) is consistent with the structural constraints of the Latin American SME literature, such as constrained access to financing, management capacity, training, and technology adoption (Ibarrarán et al., 2009). Nevertheless, these mechanisms are not directly observable in the current dataset

and therefore constitute only a plausible interpretation, not an empirical explanation of the cluster structure. The comparison of cluster-level ages provides another, directly observable, piece of evidence: Cluster 1 (medium-sized) firms are almost as old on average (23.2 years) as Cluster 2 (large) firms (25.1 years), but its financial scale is much lower, which indicates that neither medium nor large size is an inevitable result of the firm's age.

Lastly, model diagnostics (Figure 10: residual plot, Q-Q plot, scale-location plot, leverage plot) show the presence of heteroskedasticity and deviations from the normality in tails of the residual distribution, consistent with the heavy-tailed nature of firm financial data. Several high-leverage observations were found, but do not seem to affect the estimation results, since there was no significant difference between the standard and robust regression estimates. Thus, these diagnostics justify using the cluster-robust standard errors as the basis of all inference in this paper.

5.6. Theoretical and Practical Implications

Apart from the direct tests of H1–H5, this research has several implications for the literature on firm performance and heterogeneity. First, this study enriches the literature on resource-based approaches and production functions (Barney, 1991; Nelson & Winter, 1982; Penrose, 1959) by applying this framework to a service-based industry where physical capital is less likely to play a role than in the manufacturing industry. Controlling for employment, the asset variable has the highest revenue elasticity estimate (0.815). This finding might be related to the complementarity of the physical and human capital in education, found by Bento and Restuccia (2021). Second, the coexistence of the optimal two-cluster solution and the theory-driven three-cluster solution illustrates the ongoing tension between statistical and substantive approaches in the firm segmentation literature (Jain, 2010; Rousseeuw, 1987). This research gives other researchers a clear example of how to report and reconcile this tension, rather than disguise it. Third, the significance of the secondary territorial effect, controlling for the size of the firm, constitutes a firm-level contribution to the Ecuadorian agglomeration economics literature, which had previously relied on the city- or canton-level aggregates of productivity (Torres-Gutiérrez et al., 2019). The firm-level approach applied here demonstrates that the aggregation and competition parts of agglomeration can coexist in the same city at the firm level.

For the managers of the firms in the study—especially the owners and administrators of the 1797 firms in the small-firm segment—these findings show that organizational size, represented by asset base and employment, is an important correlate of firm heterogeneity in the education sector. However, since the cluster analysis is conducted in a single cross-section (2025), it does not allow determining the patterns by which individual firms move between the clusters over time or the effects that the additional investment in productive assets would make a particular firm change its cluster affiliation. Such issues would have to be addressed through a longitudinal analysis of firm growth and cluster transitions, which cannot be performed using the dataset used in this study. In light of the established evidence that productivity gaps in Latin American SMEs are largely driven by credit, training, and technological access constraints (Ibarrarán et al., 2009), this cross-sectional evidence motivates future research to analyze whether those constraints could account for the established scale gap and how addressing them might lead to productive asset accumulation among the small education-sector firms. Finally, the comparison of cross-sectional characteristics of the medium-sized firms in Cluster 1 (with ages close to large firms but significantly lower scale) indicates that the firm age is not sufficient for having the resource base of large firms in line with the resource-based view of firm growth (Barney, 1991), though this, too, is a cross-sectional relationship.

For policy makers and regulators, the combination of the extremely skewed size distribution, a strong territorial effect, and the demand sensitivity to macroeconomic shocks (such as the COVID-19 contraction analyzed above; Bracco et al., 2025) suggests that the education-sector policies should refrain from "one-size-fits-all", "one instrument fits all" approach. A combination of firm-level capacity building programs (which had mixed but mostly positive effects in comparable settings (Barrera-Osorio et al., 2022; Bracco et al., 2025)) and geography-specific assistance for firms in areas outside Quito and Guayaquil could address both these aspects effectively. However, this study does not evaluate any specific intervention, and this is an implication that needs to be tested in future policy studies. For investors and development financiers, the fact that the revenue and asset base is extremely concentrated among a few mature, large firms echoes the international evidence showing that capital investment in the education sector disproportionately benefits mature and large firms over new and small ones (Mundy & Menashy, 2014). This makes a strategy of investing in the underdeveloped small and medium-firm segments viable and relatively low in competition while being fully compatible with the discussed SME productivity-gap evidence (Ibarrarán et al., 2009). From a research perspective, this study shows an approach that can be replicated for other human capital-intensive sectors in Ecuador and wider Latin American region with scarce firm-level financial data, such as health services and professional/technical services.

6. Conclusions

This paper analyzes the correlates and heterogeneity of revenue performance of firms in the Ecuadorian education sector over 2015–2025 based on econometric modeling and unsupervised segmentation of the official data from the regulatory authority Superintendencia de Compañías, Valores y Seguros (12,031 firm-year observations, 2968 firms).

H1, H2, H4, and H5 are supported by the evidence, while H3 is compatible with the evidence in that the clustering analysis identifies structural heterogeneity, although the internal validity criteria do not uniquely define the optimal number of clusters. The three-cluster solution was chosen for substantive interpretability and relative bootstrap stability despite the lack of unanimous support for the number of clusters from the internal validity criteria. Though $k = 2$ performed better on selected criteria in the initial comparison (both the silhouette and Davies–Bouldin indices suggest it over $k = 3$, while only the Calinski–Harabasz index suggests $k = 3$; Section 4.7, Table 10) and $k = 5$ has stronger Calinski–Harabasz and Davies–Bouldin scores in the extended $k = 5$–8 range, the five-cluster solution has introduced very small groups and is much more sensitive to the variable transformation (Section 4.7, Tables 15 and 16). Thus, the chosen clustering should be considered as exploratory, not uniquely determined. Accordingly, H3 has been formulated as compatible with, not uniquely confirmed by, the evidence. Firm assets and the number of employees are positively and statistically significantly correlated with firm revenue, with estimated elasticities of 0.815 and 0.353, respectively. The regression of revenue on assets, employment, year effects, and city effects explains 61.9% of the variance in log-revenue ($R^2 = 0.619$). Three firm types were identified: large firms (0.44%), medium-sized firms (12.16%), and small firms (87.40%). This is not an official classification of firm sizes but empirically derived groupings on the basis of four standardized variables (revenue, assets, employment, firm age), and the percentages should be read accordingly. The sensitivity analysis showed that cluster membership for both $k = 3$ and $k = 5$ is influenced by the transformation of highly-skewed financial and operational variables (Adjusted Rand Index of 0.181 and 0.104 between original scale and log-transformed assignments, respectively). Thus, the three identified clusters should be regarded as an interpretable, comparatively more stable empirical representation of firm heterogeneity rather than fixed, uniquely determined, or transformation-invariant categories. The analysis has also revealed some temporal patterns,

including the 2020 contraction and subsequent recovery and steady increase in firm count. There is also a secondary territorial dimension. The city fixed effects are jointly statistically significant (conventional joint F-test; Section 4.6), and eight cities (Alangasí, Cuenca, Portoviejo, Quevedo, Quito, Riobamba, Santo Domingo de los Colorados, and Tumbaco) have significantly higher conditional revenue than the "Other cities" pool serving as the baseline. Guayaquil is notably absent in the list despite its large market size.

Practical implications arise for various stakeholders. For managers, specifically for those running the 1797 small firms (average firm age 3.9 years), the finding of the study calls attention to productive assets and scale as important correlates of revenue performance in this sector. The question whether intentional increases in such resources lead to revenue growth deserves investigation. From a policy perspective, due to the extreme size concentration of the firms and a statistically significant territorial dimension, the possibility of testing whether capacity-building in firms combined with territorially varied incentives is a better tool compared to each of these instruments needs to be separately explored; this study does not test any intervention itself. Public-private partnerships and subsidizing schemes, if they are well-designed, can increase the provision of education, though the effectiveness of such policies is highly situational: from little to no effect in low-capacity outsourcing environments (Romero et al., 2020) to a more positive effect in cases of targeting underserved areas (Allende et al., 2026). Sometimes, mismatching policy tools with local capacity can generate negative results (Pritchett, 2014). This study offers firm-level benchmarks for a sector where previous literature has almost exclusively focused on educational and policy effects rather than on firm performance and points to several possible directions for future research.

There are several limitations of the current study. First, revenue is a scale-dependent measure and cannot be viewed as a sole performance measure. It captures neither profitability nor educational quality nor anything else that could be included in a wider definition of business performance. Second, the elasticities estimated above are robust association effects, not causal estimates. Though a specification with fixed effects for the firms, presented in Section 4.6 (Robustness Check: Firm and Year Fixed Effects), allows us to rule out the possibility of reverse causation among the studied variables to some degree, this problem is not fully eliminated because of the lack of data necessary for proper controls. Third, city coefficients are to be understood as conditional geographic differences, but not as territorial effects, because they might absorb unobservable local factors of price, demand, or institutions that are not identifiable in this dataset. The reference category in this specification is the pooled "Other cities" category mentioned in Section 3.4, so that the coefficients reflect gaps of each city relative to this pooled and low-activity baseline, but not to any particular city. Fourth, this study did not use dynamic panel data specification. Fifth, the clustering results are exploratory and methodologically sensitive. Although the $k = 3$ solution was chosen for its interpretability and relatively better stability by bootstrapping, Section 4.7 shows that there is no validity criterion unambiguously favoring the $k = 3$ solution over the $k = 2$ and $k = 4$ solutions. An additional comparison finds that $k = 5$ outperforms $k = 3$ according to the Calinski–Harabasz and Davies–Bouldin criteria, though it is less stable in bootstrapping (Section 4.7, Tables 15 and 16). Cluster assignments also demonstrate poor consistency between the original-scale and log1p-transformed specifications, for both $k = 3$ and $k = 5$ (Adjusted Rand Index of 0.181 and 0.104, correspondingly), and clustering is based on a single cross-section (2025) whose temporal stability has not been examined. Sixth, there were no available variables reflecting educational quality, innovations, or technology adaptation. Seventh, the SCVS registry contains only formal companies (mainly sociedades anónimas and analogues) and might miss informal firms or different structures, so the findings refer to the formal part of the sector.

Some potential extensions of the study, based on literature review, could reach in at least three directions. First, the inclusion of variables related to access to credit, education, and innovation—factors driving the SME productivity gap in Latin America (Ibarrarán et al., 2009)—would clarify whether the small size of small firms in this sector is a structural constraint or an intention. Second, addition of the canton-level variables in the spirit of Torres-Gutiérrez et al. (2019) would allow a more precise estimation of the agglomeration economies and competitive pressure discussed in Section 5. Third, the inclusion of demand-side information about the pandemic-caused drop in private-school enrollment in Latin America (Bracco et al., 2025) would allow testing the mechanisms proposed in Section 5, which was impossible because of the lack of corresponding data in this study. Possible future research could also employ dynamic panel-data methods or machine learning algorithms, and could consider the $k = 2$ and $k = 4$ clustering solutions used as robustness tests in this study as alternative solutions in their own right. Such extensions will strengthen the link this study made between the corporate performance literature and firm-level outcomes in Latin American education.

Author Contributions: Conceptualization, C.I.R.N., M.E.S.P. and F.J.Z.F.; methodology, M.E.S.P. and F.J.Z.F.; software, M.E.S.P.; validation, C.I.R.N., F.J.Z.F. and G.d.C.G.Z.; formal analysis, M.E.S.P.; investigation, C.I.R.N. and G.d.C.G.Z.; resources, F.J.Z.F.; data curation, M.E.S.P.; writing—original draft preparation, M.E.S.P.; writing—review and editing, F.J.Z.F., C.I.R.N. and G.d.C.G.Z.; supervision, F.J.Z.F.; project administration, F.J.Z.F. All authors have read and agreed to the published version of the manuscript.

Funding: This article was funded by Universidad Estatal de Milagro.

Institutional Review Board Statement: Not applicable. This study does not include any human or animal experiments; it only uses secondary firm-level data drawn from the publicly available regulatory database of the Superintendencia de Compañías, Valores y Seguros of Ecuador.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data utilized in this study were collected from the publicly available directory and rankings of the Superintendencia de Compañías, Valores y Seguros of Ecuador (https://www.supercias.gob.ec, accessed on 20 October 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CR4	four-city concentration ratio
FE	fixed effects
GVIF	generalized variance inflation factor
HHI	Herfindahl–Hirschman Index
IQR	interquartile range
ISIC	International Standard Industrial Classification of All Economic Activities
OLS	ordinary least squares
PCA	principal component analysis
RBV	resource-based view
RLM	robust linear model (robust regression)
ROE	return on equity
SCVS	Superintendencia de Compañías, Valores y Seguros (Ecuador)
SD	standard deviation
SE	standard error
SMEs	small and medium-sized enterprises
USD	U.S. dollars
VIF	variance inflation factor
WSS	within-cluster sum of squares

References

Allende, C., Gallego, F., & Neilson, C. (2026). *Approximating the equilibrium effects of informed school choice* (NBER Working Paper No. 35837). National Bureau of Economic Research (NBER). [CrossRef]

Andrews, J. (2016). *School performance in multi-academy trusts and local authorities—2015*. Education Policy Institute (EPI).

Assomull, A., Abdo, M., & Pelley, R. (2015). *Driving grades, driving growth: How private capital in education is increasing access, inspiring innovation and improving outcomes*. Parthenon-EY.

Barney, J. (1991). Firm resources and sustained competitive advantage. *Journal of Management*, *17*(1), 99–120. [CrossRef]

Barrera-Osorio, F., Blakeslee, D. S., Hoover, M., Linden, L., Raju, D., & Ryan, S. P. (2022). Delivering education to the underserved through a public-private partnership program in Pakistan. *The Review of Economics and Statistics*, *104*(3), 399–416. [CrossRef]

Bento, P., & Restuccia, D. (2021). On average establishment size across sectors and countries. *Journal of Monetary Economics*, *117*, 220–242. [CrossRef]

Bergoeing, R., & Repetto, A. (2006). Micro efficiency and aggregate growth in Chile. *Cuadernos de Economía*, *43*(127), 169–191. [CrossRef]

Bracco, J., Ciaschi, M., Gasparini, L., Marchionni, M., & Neidhöfer, G. (2025). The impact of COVID-19 on education in Latin America: Long-run implications for poverty and inequality. *Review of Income and Wealth*, *71*(1), e12687. [CrossRef]

Busso, M., Madrigal, L., & Pagés, C. (2013). Productivity and misallocation of resources in Latin America. *The B.E. Journal of Macroeconomics*, *13*(1), 903–932. [CrossRef]

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2011). Robust inference with multiway clustering. *Journal of Business & Economic Statistics*, *29*(2), 238–249. [CrossRef]

Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, *50*(2), 317–372. [CrossRef]

Crawfurd, L., Hares, S., & Todd, R. (2024). The impact of private schools, school chains and PPPs in developing countries. *The World Bank Research Observer*, *39*(1), 97–123. [CrossRef]

Davis, S. J., Haltiwanger, J. C., & Schuh, S. (1996). *Job creation and destruction*. MIT Press.

Dollar, D., Hallward-Driemeier, M., & Mengistae, T. (2005). Investment climate and firm performance in developing economies. *Economic Development and Cultural Change*, *54*(1), 1–31. [CrossRef]

Easterly, W., & Levine, R. (2001). What have we learned from a decade of empirical research on growth? It's not factor accumulation: Stylized facts and growth models. *The World Bank Economic Review*, *15*(2), 177–219. [CrossRef]

Eeckhout, J. (2004). Gibrat's law for (all) cities. *American Economic Review*, *94*(5), 1429–1451. [CrossRef]

Elacqua, G., Contreras, D., Salazar, F., & Santos, H. (2011). The effectiveness of private school franchises in Chile's national voucher program. *School Effectiveness and School Improvement*, *22*(3), 237–263. [CrossRef]

Evans, D. S. (1987). Testing alternative theories of firm growth. *Journal of Political Economy*, *95*(4), 657–674. [CrossRef]

Featherston, S., & Ferreira, D. (2017). *Education businesses in Africa*. Foundation for Studies and Research on International Development.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). Sage.

Foster, L., Haltiwanger, J., & Krizan, C. J. (2001). Aggregate productivity growth: Lessons from microeconomic evidence. In C. R. Hulten, E. R. Dean, & M. J. Harper (Eds.), *New developments in productivity analysis* (pp. 303–372). University of Chicago Press.

Gabaix, X. (2009). Power laws in economics and finance. *Annual Review of Economics*, *1*(1), 255–294. [CrossRef]

Grazzi, M., Pietrobelli, C., & Szirmai, A. (2016). Determinants of enterprise performance in Latin America and the Caribbean: What does the micro-evidence tell us? In M. Grazzi, & C. Pietrobelli (Eds.), *Firm innovation and productivity in Latin America and the Caribbean* (pp. 1–36). Palgrave Macmillan.

Gujarati, D. N., & Porter, D. C. (2009). *Basic econometrics* (5th ed.). McGraw-Hill/Irwin.

Hall, B. H. (1987). The relationship between firm size and firm growth in the US manufacturing sector. *Journal of Industrial Economics*, *35*(4), 583–606. [CrossRef]

Hall, M., & Weiss, L. (1967). Firm size and profitability. *The Review of Economics and Statistics*, *49*(3), 319–331. [CrossRef]

Hall, R. E., & Jones, C. I. (1999). Why do some countries produce so much more output per worker than others? *The Quarterly Journal of Economics*, *114*(1), 83–116. [CrossRef]

Herman, E., Zsido, K.-E., & Fenyves, V. (2022). Cluster analysis with K-mean versus K-medoid in financial performance evaluation. *Applied Sciences*, *12*(16), 7985. [CrossRef]

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*(6), 417–441. [CrossRef]

Hsiao, C. (2014). *Analysis of panel data* (3rd ed.). Cambridge University Press.

Hsieh, C.-T., & Klenow, P. J. (2009). Misallocation and manufacturing TFP in China and India. *Quarterly Journal of Economics*, *124*(4), 1403–1448. [CrossRef]

Huber, P. J. (1981). *Robust statistics*. Wiley.

Hutchings, M., & Francis, B. (2017). *Chain effects: The impact of academy chains on low-income students*. The Sutton Trust.

Ibarrarán, P., Maffioli, A., & Stucchi, R. (2009). *Policy towards SMEs and their productivity in Latin America* (IZA Discussion Paper No. 4486). Institute for the Study of Labor (IZA).

Jain, A. K. (2010). Data clustering: 50 Years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666. [CrossRef]

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202. [CrossRef] [PubMed]

Jovanovic, B. (1982). Selection and the evolution of industry. *Econometrica*, *50*(3), 649–670. [CrossRef] [PubMed]

Kenny, C., Kalow, J., Leo, B., & Ramachandran, V. (2018). *Comparing five bilateral development finance institutions and the IFC* (CGD Policy Paper No. 116). Center for Global Development. Available online: https://www.cgdev.org/publication/comparing-five-bilateral-development-finance-institutions-and-ifc (accessed on 18 October 2026).

Kingdon, G. G. (2020). The private schooling phenomenon in India: A review. *The Journal of Development Studies*, *56*(10), 1795–1817. [CrossRef]

Klenow, P., & Rodríguez-Clare, A. (1997). The neoclassical revival in growth economics: Has it gone too far? In *NBER macroeconomics annual 1997* (pp. 73–114). MIT Press.

Lavopa, A. (2015). *Structural transformation and economic development: Can development traps be avoided?* [Doctoral dissertation, Maastricht University].

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In L. M. Le Cam, & J. Neyman (Eds.), *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 281–297). University of California Press.

McMillan, M., Rodrik, D., & Verduzco-Gallo, Í. (2014). Globalization, structural change, and productivity growth, with an update on Africa. *World Development*, *63*, 11–32. [CrossRef]

Mundy, K., & Menashy, F. (2014). Investing in private education for poverty alleviation: The case of the World Bank's International Finance Corporation. *International Journal of Educational Development*, *35*, 16–24. [CrossRef]

Nelson, R. R. (1991). Why do firms differ, and how does it matter? *Strategic Management Journal*, *12*(S2), 61–74. [CrossRef]

Nelson, R. R., & Winter, S. G. (1982). *An evolutionary theory of economic change*. Harvard University Press.

Newbert, S. L. (2008). Value, rareness, competitive advantage, and performance: A conceptual-level empirical investigation of the resource-based view of the firm. *Strategic Management Journal*, *29*(7), 745–768. [CrossRef]

Pagés, C. (Ed.). (2010a). *La era de la productividad: Cómo transformar las economías desde sus cimientos*. Banco Interamericano de Desarrollo.

Pagés, C. (Ed.). (2010b). *The age of productivity: Transforming economies from the bottom up*. Palgrave Macmillan for Inter-American Development Bank.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, *2*(11), 559–572. [CrossRef]

Penrose, E. T. (1959). *The theory of the growth of the firm*. Basil Blackwell.

Perry, G. E., Maloney, W. F., Arias, O. S., Fajnzylber, P., Mason, A. D., & Saavedra-Chanduvi, J. (2007). *Informality: Exit and exclusion*. World Bank.

Pritchett, L. (2014). *The risks to education systems from design mismatch and global isomorphism: Concepts, with examples from India* (WIDER Working Paper No. 2014/039). United Nations University World Institute for Development Economics Research (UNU-WIDER).

R Core Team. (2025). *R: A language and environment for statistical computing* (Version 4.5.1). R Foundation for Statistical Computing.

Romero, M., Sandefur, J., & Sandholtz, W. A. (2020). Outsourcing education: Experimental evidence from Liberia. *American Economic Review*, *110*(2), 364–400. [CrossRef]

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. [CrossRef]

Smith, W. C., & Baker, T. (2017). From free to fee: Are for-profit, fee-charging private schools the solution for the world's poor? RESULTS Educational Fund.

Sutton, J. (1997). Gibrat's legacy. *Journal of Economic Literature*, *35*(1), 40–59. [CrossRef] [PubMed]

Syverson, C. (2011). What determines productivity? *Journal of Economic Literature*, *49*(2), 326–365. [CrossRef]

Teece, D., & Pisano, G. (1994). The dynamic capabilities of firms: An introduction. *Industrial and Corporate Change*, *3*(3), 537–556. [CrossRef]

Thorndike, R. L. (1953). Who belongs in the family? *Psychometrika*, *18*(4), 267–276. [CrossRef]

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B*, *63*(2), 411–423. [CrossRef]

Torres-Gutiérrez, T. P., Correa-Quezada, R., Álvarez-García, J., & de la Cruz del Río-Rama, M. (2019). Agglomeration Economies: An analysis of the determinants of employment in the cities of Ecuador. *Symmetry*, *11*(11), 1421. [CrossRef]

United Nations. (2008). *International Standard Industrial Classification of All Economic Activities (ISIC), Revision 4* (Statistical Papers, Series M, No. 4/Rev.4). United Nations.

Von Nordenflycht, A. (2010). What is a professional service firm? Toward a theory and taxonomy of knowledge-intensive firms. *Academy of Management Review*, *35*(1), 155–174. [CrossRef]

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*(4), 817–838. [CrossRef]
Wooldridge, J. M. (2013). *Introductory econometrics: A modern approach* (5th ed.). South-Western Cengage Learning.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.