

Beyond Banks and ATMs: When Institutions and Regulation Drive Bank Deposit Mobilization

Mara MAZĂLU

*Bucharest University of Economic Studies, Bucharest, Romania
mazalumara23@stud.ase.ro*

Ana-Maria-Antonia MITU

*Bucharest University of Economic Studies, Bucharest, Romania
mituana23@stud.ase.ro*

Răzvan-Auxeniu MICLESCU

*Bucharest University of Economic Studies, Bucharest, Romania
miclescurazvan23@stud.ase.ro*

Adriana AnaMaria DAVIDESCU

*Bucharest University of Economic Studies, Bucharest, Romania
adriana.alexandru@csie.ase.ro*

Ioana BÎRLAN*

*Bucharest University of Economic Studies, Bucharest, Romania
Corresponding author, ioana.birlan@csie.ase.ro

Abstract. *Why do some countries mobilize far more bank deposits than others, even when banking infrastructure appears broadly comparable? This paper examines whether deposit mobilization depends primarily on physical outreach or, more fundamentally, on the institutional environment in which financial intermediation operates. Using cross-sectional data for 2022 covering 111 countries, the study evaluates whether outstanding deposits relative to GDP are better explained by bank branches and ATMs, or by legal and regulatory quality, controlling for differences in economic development. The empirical strategy combines a benchmark OLS specification in levels, a preferred log-transformed model motivated by data asymmetry and diagnostic weaknesses of the raw-scale benchmark, and a predictive validation layer based on six machine learning algorithms. The results indicate that regulatory quality is the strongest and most robust correlate of deposit mobilization, while infrastructure variables lose statistical significance once institutional factors are introduced. In the predictive stage, regularized linear methods outperform the more flexible nonlinear alternatives. These findings suggest that policies aimed at deepening deposit-based intermediation should place greater emphasis on institutional and regulatory credibility, alongside rather than relying solely on the expansion of banking infrastructure.*

Keywords: bank deposit mobilization, institutional quality, regulatory framework, financial inclusion, banking infrastructure, cross-country analysis, machine learning validation.

Introduction

Deposits do not get much attention in the financial development literature, at least not as an outcome worth explaining in their own right. They tend to appear as a component of broader financial depth measures, absorbed into the analysis rather than featured in it. That is a mistake, we think. Deposits are the mechanism through which dispersed private savings get channeled into the

DOI: 10.2478/picbe-2026-0100

© 2026 M. Mazălu; A.-M.-A. Mitu; R.-A. Miclescu; A. A.M. Davidescu; I. Bîrlan, published by De Gruyter.
This work is licensed under the Creative Commons Attribution 4.0 License.

formal financial system at scale. Whether that happens (and how much of it happens) varies enormously across countries. And the variation is not well explained by the factors that dominate the standard financial inclusion narrative.

The starting observation is this: over the past two decades, banking infrastructure has expanded considerably in most parts of the world: more branches, more ATMs, better formal access. And yet deposit mobilization has remained stubbornly uneven. Countries with essentially similar banking footprints mobilize deposits at very different rates. The standard explanation - give people a bank account and they will save in it - does not appear to hold at the cross-country level. So what does explain it? Our conjecture, which the data support, is that the institutional environment matters far more than the physical one. Not how many branches there are, but whether the system surrounding those branches is credible enough for savers to actually use them.

There is relevant prior work to draw on, though none of it fully resolves the question. Financial development research has established - convincingly, we think - that institutional quality conditions what countries gain from financial expansion. Research on banking access has made a reasonable case for infrastructure. What has been done less often is to put both in direct empirical competition, with deposit mobilization as the specific outcome and a global cross-section as the testing ground. That is the gap we try to fill here.

The paper makes three specific contributions. First - and most basic - it takes deposit mobilization seriously as the thing to be explained, not a component of some broader financial development index. Second, it runs an explicit competitive test between infrastructure and institutional variables, rather than including both as controls without asking which dominates. Third, it uses three complementary empirical strategies - OLS in levels, log-transformed OLS, and machine learning validation on held-out data, supplemented by PCA and hierarchical clustering - not out of methodological ambition, but because each approach addresses a different potential concern about the others.

The stakes of getting this right are not trivial. Financial inclusion strategies that prioritize infrastructure expansion - on the assumption that access drives savings - may be targeting the wrong lever. The sections that follow develop the argument through research hypotheses, a review of the relevant literature, a description of the data and methods, and a presentation of results. Across all three empirical layers, the institutional signal is stronger and more consistent than the infrastructural one. We think that finding deserves to be taken seriously.

Research hypotheses

Three hypotheses guide the empirical analysis, each grounded in the relevant theoretical and empirical literature. The first is the most intuitive: more bank branches and ATMs should mean easier access to formal finance, lower participation barriers, and ultimately more deposits. Call this the infrastructure hypothesis.

The second hypothesis is more challenging: legal protection and regulatory quality are expected to outperform infrastructure in explaining cross-country deposit differences. This is not a claim pulled from thin air - Chinn and Ito (2006), among others, showed that institutional capacity shapes what countries actually get from financial expansion. The hypothesis extends that logic specifically to deposits.

Third - and most nuanced - is the heterogeneity hypothesis. Infrastructure may not be irrelevant everywhere; it may just work better in countries that already have solid legal and regulatory underpinnings. Where institutions are weak, adding branches may accomplish very

little. We explore this conditional relationship through clustering and cross-country profiling, though we stop short of a formal interaction specification.

Literature review and research gap

The question this paper asks - why do some countries mobilize bank deposits so much more effectively than others - sounds simple. It is not. Two countries can have nearly identical banking access profiles and deposit at rates that differ by a factor of three. The answer is not in the number of branches. The literature reviewed below helps explain why, and points toward the institutional environment as the more plausible driver. We organize the review around three broad themes: financial development and its determinants, the specific role of institutions, and the limited but informative set of studies that focus directly on deposit behavior.

The foundational reference is Levine (1997), who articulated why financial systems matter: they mobilize savings, allocate capital, monitor firms, and manage risk. Deposits are part of each of these functions. But Levine's framework, like most of the early financial development literature, treats financial depth as the aggregate outcome of interest. Deposits are an input to that aggregate, not a distinct outcome with its own determinants. That conflation is understandable in the context of the 1990s literature, but it has persisted in ways that are less defensible. Our paper attempts, modestly, to correct it.

A more disaggregated view of financial development comes from Svirydenka (2016), who proposes separating it into depth, access, and efficiency - and doing so independently for financial institutions and financial markets. This matters for our purposes because it makes visible a gap that aggregate measures hide: a country can score well on financial development indices while failing to mobilize deposits at any meaningful scale, if access is unequal, usage selective, or intermediation inefficient. Deposit mobilization captures something real that the indices paper over.

The evidence that institutions drive financial development is, at this point, fairly substantial. Doucouliagos et al. (2022) synthesize a large meta-regression literature and identify institutional quality as one of the most robust correlates of financial development: it shows up across samples, time periods, and model specifications in a way that most other variables do not. Lisbinski and Burnquist (2024) find this effect is especially pronounced in developing economies, which is also where deposit mobilization varies most. Chinn and Ito (2006) add a threshold dimension to the argument: financial openness generates gains in financial development only after a sufficient institutional foundation has been established. Together, these contributions suggest that institutions are not passive background conditions. They actively constrain what financial systems can deliver.

Khalid and Shafiullah (2021) document a bidirectional relationship between governance and financial development - each reinforces the other, which means causality runs in both directions and disentangling them is hard. Li et al. (2018) show that legal protection shapes financial development across market segments, not just in aggregate. What both studies imply for deposit behavior, though, is the same: a bank cannot generate depositor trust simply by opening a branch. The decision to save formally involves a judgment about whether the system surrounding the bank - legal, regulatory, governance - makes it rational to do so. Individual banks have very little control over those conditions.

The study most directly relevant to what we are doing is Nguyen (2022). She examines deposit growth in European transition economies and finds that the rule of law is a particularly important institutional driver - more so, in her analysis, than standard access variables. What makes her study methodologically useful for us is that she treats deposit mobilization as the outcome to be explained rather than as a proxy. That is the same analytical choice we make here, extended to

a much larger and more diverse global sample. If the finding holds globally - and we think it does - the policy implications are considerably broader than a European regional study can establish.

Srivastava et al. (2026) make a point worth flagging: the relationship between financial market development and bank deposits is not monotonically positive. In more advanced financial systems, banks and capital markets complement each other in attracting savings; in less developed ones, they tend to compete for the same savings pool. This non-linearity has implications for how we read cross-country regressions - linear terms may miss important structural variation. Peng et al. (2025) are useful on a different point: financial outreach can, under the right conditions, raise deposits by lowering transaction costs. The qualifier is important. "Under the right conditions" turns out to mean when access actually translates into usage - which, as we show, depends substantially on the institutional setting.

A strand of the literature that deserves more attention in this context is the work on trust. Heyert and Weill (2025) find that trust in banks has a robust positive effect on financial inclusion - an effect that holds up after controlling for income and access. The implication is that deposit behavior reflects, at least in part, a saver's assessment of whether the bank is trustworthy, not just whether it is nearby. Aracil et al. (2025) extend the argument to digital banking, showing that as financial services migrate online, the relevant form of trust shifts from interpersonal to systemic - from confidence in a banker to confidence in regulation, data security, and institutional oversight. For our purposes, the takeaway is consistent with the broader institutional argument: infrastructure, physical or digital, does not mobilize deposits unless embedded in an environment that people find credible.

What the literature, taken together, suggests is this: deposit mobilization is shaped by a combination of financial structure, institutional quality, regulation, outreach, and trust - and these forces do not operate independently. An ATM placed in a country with credible regulation and high institutional trust has different deposit effects than the same ATM in a country with weak institutions and low trust. This context-dependence is precisely why deposit mobilization cannot be treated as a mechanical consequence of infrastructure expansion. It has its own determinants, and those determinants are, we argue, primarily institutional. The empirical strategy that follows is designed to test this argument as directly as the data allow.

Methodology

Data and variables

The empirical analysis uses cross-sectional data for 2022, the most recent year with complete coverage for the variables used in the study. The full dataset includes 135 countries with different income levels and geographic profiles. For validation, the sample is split into 111 training observations and 24 test observations, close to the usual 80/20 split discussed by Gholamy, Kreinovich, and Kosheleva (2018) and Joseph (2022). The econometric models are estimated on the training sample, while the test sample is used only for out-of-sample prediction. With six predictors, the OLS model has 104 residual degrees of freedom.

The dependent variable (Y) is outstanding deposits with commercial banks as a percentage of GDP, sourced from the IMF Financial Access Survey (IMF FAS). This measure is the appropriate outcome variable for the question being asked; Nguyen (2022) uses the same measure in the closest regional precedent and finds it responds more to institutional quality than to infrastructure access. Two explanatory variables also drawn from IMF FAS proxy the banking infrastructure side of the analysis: X1, commercial bank branches per 100,000 adults, and X2, ATMs per 100,000 adults. These capture the dimensions of physical reach and transactional

accessibility that the financial inclusion literature most commonly treats as drivers of deposit behavior.

Institutional quality is proxied by two variables from the Fraser Institute's *Economic Freedom of the World* report. X3 measures Legal System and Property Rights; X4 captures Regulation. Both run on 0-10 scales, with higher values indicating stronger institutional quality. Economic development is controlled for through X5, GDP per capita at purchasing power parity in constant 2021 international dollars, sourced from the World Bank's World Development Indicators. A binary dummy variable (D) completes the specification: it takes the value 1 for countries whose Heritage Foundation economic freedom scores fall below the sample median (Low Freedom), with High Freedom as the reference category.

The distributional characteristics of the variables warrant comment before turning to estimation. The dependent variable, branch density, ATM density, and GDP per capita all display substantial positive skewness - skewness coefficients of 4.67, 1.37, 2.90, and 1.79 respectively (Table 1). The institutional indices (X3, X4) are considerably more symmetric, with skewness near zero. This divergence is not a minor descriptive footnote; it has direct implications for the choice of functional form and is the primary empirical motivation for preferring log-transformed over levels specifications in the estimation that follows.

Table 1. Descriptive statistics for the estimation sample

Variable	Description	N	Mean	Std. Dev.	Min	Max	Skewness
Y	Outstanding deposits (% of GDP)	111	65.28	63.44	6.65	548.76	4.67
X1	Bank branches per 100,000 adults	111	32.36	29.50	1.60	128.26	1.37
X2	ATMs per 100,000 adults	111	37.59	49.22	2.03	304.15	2.90
X3	Legal System & Property Rights (0-10)	111	5.55	1.57	2.53	8.93	0.45
X4	Regulation (0-10)	111	6.47	0.89	4.10	8.54	-0.24
X5	GDP per capita, PPP (\$)	111	27,196	26,549	1,477	133,572	1.79
D	Low Economic Freedom dummy	111	0.50	0.50	0	1	N.A.

Source: Authors' own calculations based on IMF Financial Access Survey, Fraser Institute, World Bank, and Heritage Foundation data.

Econometric approach

The empirical strategy unfolds in four stages. The first estimates a baseline OLS model in level form:

$$Y_i = \beta_0 + \beta_1 X1_i + \beta_2 X2_i + \beta_3 X3_i + \beta_4 X4_i + \beta_5 X5_i + \beta_6 D_i + \varepsilon_i$$

This level-form specification serves as the benchmark against which subsequent models are assessed. Standard diagnostics accompany the estimation: residual normality, heteroscedasticity, functional-form misspecification, and multicollinearity are all examined. The heteroscedasticity check follows the White (1980) procedure; influential observations are identified using Cook's distance.

The second stage estimates a log-log specification, motivated jointly by the distributional properties of the data and by the diagnostic shortcomings of the raw-scale benchmark:

$$\ln(Y_i) = \beta_0 + \beta_1 \ln(X1_i) + \beta_2 \ln(X2_i) + \beta_3 X3_i + \beta_4 X4_i + \beta_5 \ln(X5_i) + \beta_6 D_i + u_i$$

Logarithmic transformations are applied to the dependent variable, branch density, ATM density, and GDP per capita; the institutional indices and the dummy variable remain in levels. The

rationale is straightforward: the economic variables are heavily right-skewed and the relationships they enter are more plausibly multiplicative than additive. The transformation also aids interpretation - coefficients on the logged continuous variables read as elasticities, while those on X3 and X4 carry a semi-elasticity meaning.

The third stage turns to unsupervised learning to map cross-country heterogeneity in the data. Principal Component Analysis (PCA) reduces dimensionality where moderate collinearity - particularly among the institutional and development variables - would otherwise complicate interpretation (Jolliffe & Cadima, 2016). Hierarchical clustering then groups countries by their institutional and infrastructural profiles. Neither technique is meant to substitute for regression analysis; they serve as exploratory complements that help identify country typologies and probe whether the relationship between infrastructure, institutions, and deposits looks meaningfully different across structural contexts.

The fourth stage adds a machine learning validation exercise. Several algorithms are tested on raw variables, log-transformed variables, cluster-augmented data, and PCA-based inputs. Penalized linear models follow Friedman, Hastie, and Tibshirani (2010), while Elastic Net follows Zou and Hastie (2005). The ML stage uses the same train-test split and is interpreted as a predictive check, not as structural inference, following Mullainathan and Spiess (2017). Models were implemented in both R and Python; the final results are reported from the Python pipeline, with R used as a robustness check.

Results and Discussions

Exploratory data analysis

A first look at the data is instructive. Outstanding deposits range from 6.65% of GDP in the least-mobilizing country to 548.76% in the most, with a mean of 65.28% and a skewness of 4.67 - this is a heavily right-tailed distribution driven by a small number of offshore financial centers and globally integrated economies. ATM density and GDP per capita behave similarly. The institutional indices (X3 and X4) do not: they are considerably more symmetric, with skewness coefficients close to zero. This divergence in distributional behavior has direct implications for estimation strategy, and is one of the core reasons the log-transformed specification is preferred over the raw levels model rather than an arbitrary choice.

The bivariate correlation structure already tells a story consistent with what the regressions will show. Institutional quality correlates with deposit mobilization at $r = 0.31$ to 0.38 ; GDP per capita at $r = 0.37$. Branch density and ATM density produce correlations in the 0.03 to 0.15 range - present, but weak enough to raise questions about their role in any multivariate analysis. Worth flagging separately: X3, X4, and X5 are highly intercorrelated ($X3-X5 = 0.80$; $X3-X4 = 0.74$), which creates collinearity that the regression diagnostics need to account for explicitly.

Figure 1 displays the full correlation matrix, making visible the moderate-to-strong associations among the institutional and development variables that call for explicit multicollinearity diagnostics.

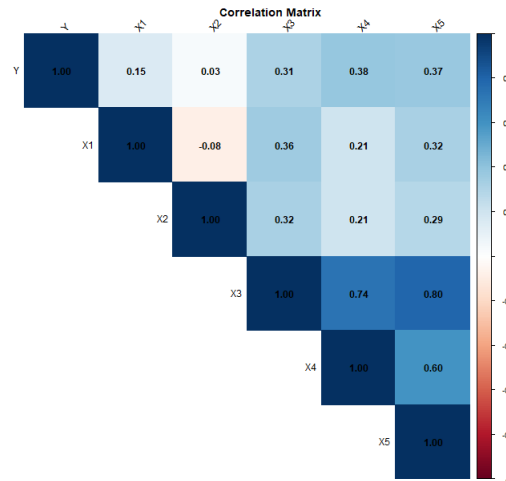


Figure 1. Correlation matrix of the explanatory variables

Source: Authors' own processing in R based on the estimation sample.

Cook's distance, using the conventional $4/n$ threshold, flags five influential cases: Hong Kong SAR, Ireland, Luxembourg, Mauritania, and Singapore. We do not treat them as data errors, because each case has an economic explanation. Hong Kong, Luxembourg and Singapore are financial hubs, Mauritania has a deposit ratio that is unusually high for its income level, and Ireland's GDP per capita is affected by the presence of multinational companies. For this reason, the observations are kept in the sample. Their influence is handled through log transformation and robustness checks, rather than by deleting them mechanically.

Figure 2 shows the distribution of outstanding deposits, illustrating the pronounced right-skewness of the dependent variable. Figure 3 confirms that GDP per capita follows a similarly asymmetric pattern.

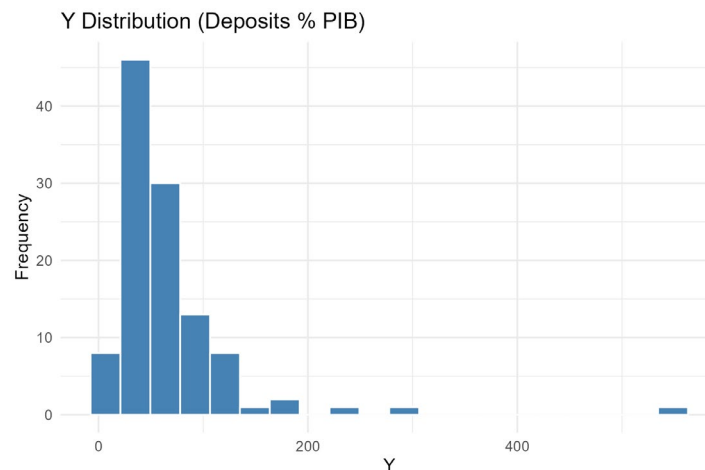


Figure 2. Distribution of outstanding deposits with commercial banks (% of GDP)

Source: Authors' own processing in R based on the estimation sample.

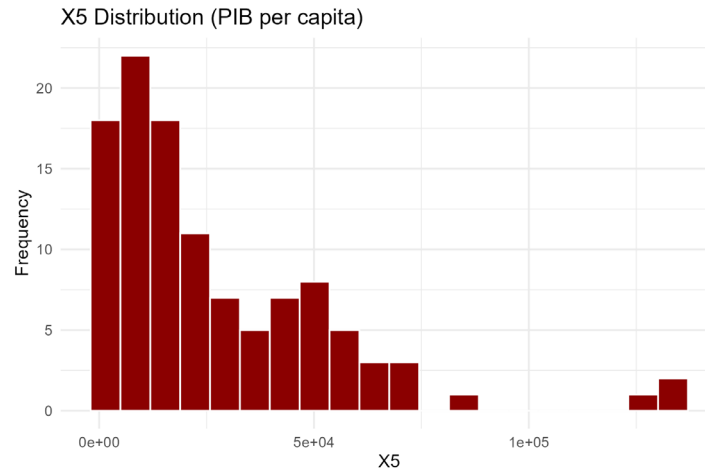


Figure 3. Distribution of GDP per capita at PPP (constant 2021 international dollars)

Source: Authors' own processing in R based on the estimation sample.

Baseline regression results

The benchmark OLS model, estimated in levels, produces limited fit: $R^2 = 0.196$, adjusted $R^2 = 0.150$. Among the six predictors, only Regulation (X4) and GDP per capita (X5) achieve statistical significance, both positively. Branch density (X1), ATM density (X2), the legal system index (X3), and the economic freedom dummy all fail to produce robust estimates. This pattern (infrastructure variables essentially absent, regulatory variable present) already emerges in the crudest specification, before any distributional corrections or transformations are applied. It is not a consequence of model refinement; it is a feature of the raw data.

Table 2. Baseline OLS regression results (level-level specification)

Variable	Coefficient	Standard Error	Significance Level
X1	0.108	0.215	-
X2	-0.074	0.126	-
X3	-7.170	7.486	-
X4	24.900	9.986	**
X5	0.001	0.0004	**
D (LowFreedom)	7.481	16.366	-
Intercept	-82.460	62.415	-

Model indicator	Value
Number of observations	111
R^2	0.196
Adjusted R^2	0.150
Residual standard error	58.487 (df = 104)
F-statistic	4.234*** (df = 6; 104)

Notes: $p < 0.1$; $p < 0.05$; $p < 0.01$.

Source: Authors' own calculations in R.

The residual diagnostics for the benchmark model are poor, though not surprising given the skewness documented in Table 1. Normality is strongly rejected - the residual histogram shows the asymmetry and long positive tail that right-skewed raw data typically generate in levels-form regressions. White's test flags heteroscedasticity. The RESET test indicates functional-form issues.

The residuals vs fitted plot confirms non-random dispersion. None of this changes the substantive result - Regulation (X4) is still the only robust predictor - but it does mean the level-level model should not be read as anything more than a benchmark reference.

Figure 4 plots residuals against fitted values for the baseline model, making the non-random structure and uneven dispersion visually apparent.

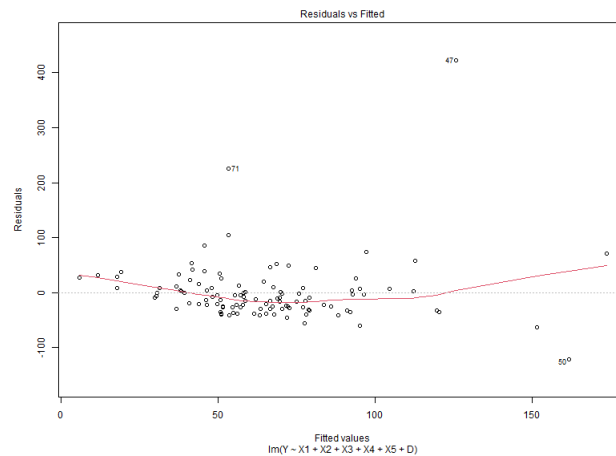


Figure 4. Residuals vs fitted values for the baseline OLS model

Source: Authors' own processing in R.

Log-transformed model estimation

The log-transformed model is the preferred specification, and the improvement over the benchmark is visible on several dimensions simultaneously. R^2 rises from 0.196 to 0.264; adjusted R^2 from 0.150 to 0.221. White's test no longer flags heteroscedasticity. The RESET test no longer indicates misspecification. The residuals vs fitted plot shows substantially more random structure than the benchmark. Box and Cox (1964) provided the theoretical justification for why logarithmic transformations restore model adequacy in cases like this - where raw-scale relationships are distorted by the multiplicative nature of the underlying economic processes and the severe skewness of the distributions.

Against the baseline, the transformed specification improves fit and stabilizes the diagnostics. Regulation (X4) remains the only clearly robust and statistically significant predictor of deposit mobilization across both specifications.

Table 3. Log-transformed regression results

Variable	Coefficient	Standard Error	Significance Level
$\ln(X1)$	0.070	0.076	-
$\ln(X2)$	-0.018	0.063	-
X3	-0.022	0.075	-
X4	0.245	0.103	**
$\ln(X5)$	0.180	0.118	-
D (Low Freedom)	0.061	0.170	-
Intercept	0.529	0.972	-

Model indicator	Value
Number of observations	111
R ²	0.264
Adjusted R ²	0.221
Residual standard error	0.604 (df = 104)
F-statistic	6.207*** (df = 6; 104)

Notes: $p < 0.1$; $p < 0.05$; $p < 0.01$.

Source: Authors' own calculations in R.

Figure 5 presents the Q-Q plot for the log-transformed model's residuals. Behavior is visibly improved relative to the baseline, though some deviation from normality persists in the tails.

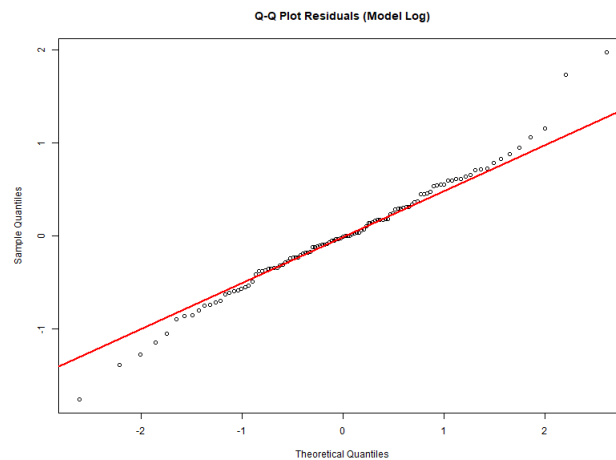


Figure 5. Q-Q plot of residuals for the log-transformed model

Source: Authors' own processing in R.

The key result of the preferred specification can be stated plainly. Regulation (X4) enters with a positive coefficient of 0.245, significant at the 5% level, and is the only variable that achieves robust statistical significance in the preferred model. Branch density (ln X1) and ATM density (ln X2) produce coefficients of 0.070 and -0.018 respectively, neither close to significance. Log GDP per capita is positive at 0.180 but also falls short. The legal system index and the economic freedom dummy contribute nothing robust. The regulatory environment, in short, is the variable doing the explanatory work - and it is doing it alone.

What should we make of this? Infrastructure is not, in our reading, irrelevant in principle - but in this cross-section it is not operating as an independent driver of deposits. Its effects, to the extent they exist, appear conditional on the institutional enabling environment - on whether the regulatory and legal conditions make it rational for savers to use a branch once it is accessible. Where institutions are credible, infrastructure may amplify deposit mobilization. Where they are not, adding branches appears to accomplish relatively little. The evidence, in our reading, supports a view in which formal banking access is necessary but not sufficient: it is one precondition among several, and not the binding one in most of the countries in this sample.

We also ran the regressions after dropping the five flagged outliers: Hong Kong, Ireland, Luxembourg, Mauritania, Singapore. The results hold. The log-transformation preference reflects the structure of the data as a whole, not the leverage of a handful of unusual economies.

Exploratory clustering and PCA

The exploratory multivariate analysis maps structural variation in the data rather than estimating conditional effects. Variables were standardized before applying PCA and clustering. The first two principal components capture about 69% of total cross-country variance, with 54.2% explained by the first component and 14.8% by the second. This suggests that much of the variation in the institutional, economic and infrastructure profile of countries can be summarized along a small number of dimensions.

Agglomerative hierarchical clustering, using Euclidean distance and Ward's D2 linkage, identifies four country groups when the dendrogram is cut at that level. The clusters follow an interpretable progression from lower-income, institutionally weaker economies, represented by countries such as India and Pakistan, to higher-income and institutionally stronger systems, represented by Hungary and Spain. This pattern is consistent with the regression results, where institutional variables are more informative than infrastructure alone.

A separate clustering of the explanatory variables, based on a $1-|r|$ correlation-distance matrix and Ward's D2 linkage, shows that branch density, legal system quality, regulation and GDP per capita group together, while ATM density separates into its own cluster. This suggests that ATMs capture a narrower transactional dimension, while the other variables are closer to the institutional-development profile of a country.

These methods do not establish causality. Their role is exploratory. Still, they show that the sample is heterogeneous in ways that are systematically related to institutional quality. This supports the idea that pooled regressions may understate some infrastructure effects within specific country groups, while the institutional signal remains more consistent across the sample.

Machine learning validation

The machine learning validation stage is introduced as a robustness complement to the econometric analysis, following the framework articulated by Mullainathan and Spiess (2017): supervised machine learning addresses predictive performance rather than structural inference, and the two should not be confused. Six algorithms were evaluated across four data representations: raw variables, log-transformed variables, PCA-based inputs, and cluster-augmented specifications - on the held-out test subsample of 24 countries. The goal was not to optimize prediction but to assess whether the patterns identified in the regression stage survive when the analysis is shifted from in-sample fit to out-of-sample accuracy.

Predictive performance varies substantially across representations. The RAW and CLUSTERS specifications produce the weakest out-of-sample results: RMSE of 26.59 and 26.35, out-of-sample R^2 of 0.065 and 0.082 respectively. The LOG representation improves on both, and the PCA-based inputs achieve the best performance: RMSE of 22.83 and R^2 of 0.281, with Ridge Regression as the selected algorithm. Two findings from this comparison are worth highlighting. First, how the input data are encoded has large effects on predictive accuracy - the gap between RAW and PCA representations is not trivial. Second, regularized linear methods (Ridge, Elastic Net) consistently outperform more flexible nonlinear alternatives, which implies the underlying relationship between the predictors and deposits has a relatively simple, approximately linear structure in the transformed space.

Table 4. Out-of-sample predictive performance across machine learning representations

Data representation	Selected model	RMSE	MAE	MAPE (%)	R ²
RAW	Elastic Net	26.5914	21.8508	55.33	0.0650
LOG (back-transformed to Y)	Ridge	23.7542	18.9742	41.04	0.2537
PCA (back-transformed to Y)	Ridge	22.8296	18.2967	39.51	0.2814
CLUSTERS	Elastic Net	26.3523	21.5762	54.71	0.0820

Source: Authors' own calculations based on the Python/R machine learning pipeline.

The machine learning results do not produce a different substantive conclusion from the regression analysis - they produce the same one, through a different method. Institutional variables, and the regulatory dimension in particular, remain more strongly associated with deposit mobilization in the predictive framework than infrastructure variables, at every level of data representation tested. What the ML stage adds beyond confirmation is a methodological finding: the encoding of input variables matters substantially for out-of-sample performance. The difference in predictive accuracy between raw and PCA-transformed inputs is large enough that it should be taken seriously by researchers applying machine learning to cross-country financial data. The models were implemented in both R and Python; final comparative results are based on the Python pipeline, with R used for robustness verification.

Discussions & conclusion

The first hypothesis receives weak support. Branch density and ATM availability have clear theoretical reasons to matter, but neither variable is statistically significant in the estimated models. Their bivariate correlations with deposit mobilization are also weak, ranging from 0.03 to 0.15. This suggests that physical access alone is not the main explanation for cross-country differences in deposits.

The second hypothesis is supported more clearly. Regulation (X4) is the only predictor that remains positive, statistically significant and substantively important across the main specifications. In the preferred log-transformed model, a one-unit increase in the Regulation index is associated with an approximate 24.5% increase in outstanding deposits as a share of GDP. This result suggests that deposit mobilization depends more on the credibility of the regulatory environment than on the simple expansion of banking infrastructure.

The third hypothesis receives only exploratory support. The PCA and clustering results show that countries form different institutional and infrastructural profiles, which is consistent with the idea that infrastructure may work differently across contexts. However, because the paper does not estimate formal interaction or threshold models, this result should be read as suggestive rather than conclusive.

Overall, the paper shows that countries do not mobilize deposits mainly by adding more branches or ATMs. Across OLS estimation, the preferred log-transformed model, PCA, clustering and machine learning validation, the strongest and most consistent signal comes from regulatory quality. Infrastructure variables remain weak across specifications.

The main policy implication is that infrastructure-led financial inclusion strategies may be incomplete. Branches and ATMs can improve access, but access is unlikely to generate deposits if the legal and regulatory environment is not credible. The study remains limited by its cross-sectional design, so the results should be interpreted as associations rather than causal effects. Even so, the consistency of the institutional result across different methods makes the finding difficult to ignore.

Acknowledgments

The research study has been elaborated within the Data Science Research Lab for Business and Economics of the Bucharest University of Economic Studies. This paper was co-financed by the Bucharest University of Economic Studies during the PhD program. During the preparation of this work, the authors used ChatGPT 4 to enhance readability and clarity of language. Following the use of this tool, the authors meticulously reviewed and revised the content as necessary and assume full responsibility for the final content of the publication.

References

- Aracil, E., Fernández-Méndez, L., Roch-Dupré, D., & Fuertes, F. J. (2025). Trust and financial inclusion: A literature review with reference to the digital transformation. *Heliyon*, 11(16), e44128.
- Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2), 211-252.
- Chinn, M. D., & Ito, H. (2006). What matters for financial development? Capital controls, institutions, and interactions. *Journal of Development Economics*, 81(1), 163-192.
- Doucouliaagos, C., de Haan, J., & Sturm, J.-E. (2022). What drives financial development? A meta-regression analysis. *Oxford Economic Papers*, 74(3), 840-863.
- Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 1-22.
- Gholamy, A., Kreinovich, V., & Kosheleva, O. (2018). *Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation* (Technical Report UTEP-CS-18-09). University of Texas at El Paso, Department of Computer Science.
- Heyert, A., & Weill, L. (2025). Trust in banks and financial inclusion: Micro-level evidence from 28 countries. *Economic Systems*, 49, 101234.
- 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.
- Joseph, V. R. (2022). Optimal ratio for data splitting. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 15(4), 531-538.
- Khalid, U., & Shafiullah, M. (2021). Financial development and governance: A panel data analysis incorporating cross-sectional dependence. *Economic Systems*, 45(2), 100819.
- Levine, R. (1997). Financial development and economic growth: Views and agenda. *Journal of Economic Literature*, 35(2), 688-726.
- Li, J., Maung, M., & Wilson, C. (2018). Governance and financial development: A cross-country analysis. *Journal of International Financial Markets, Institutions and Money*, 52, 227-239.
- Lisbinski, F. C., & Burnquist, H. L. (2024). Institutions and financial development: Comparative analysis of developed and developing economies. *Economía*, 25(2), 347-366.
- Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2), 87-106.
- Nguyen, T. A. N. (2022). The role of institutional quality in bank deposit growth in European transition economies. *Finance Research Letters*, 47(Part A), 102630.
- Peng, Y., Yang, J., Shen, J., & Gou, Q. (2025). Financial outreach, bank deposits, and economic growth. *Journal of Economic Dynamics and Control*, 171, 105036.

- Srivastava, N., Tripe, D., Haq, M., & Yuen, M. K. (2026). Financial market development and bank deposits. *Journal of International Financial Markets, Institutions and Money*, 107, 102278.
- Svirydzhenka, K. (2016). *Introducing a new broad-based index of financial development* (IMF Working Paper No. 16/5). International Monetary Fund.
- White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4), 817-838.
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301-320.