Journal of
Social and Economic Statistics

Vol. 14, No. 2, Winter 2025
DOI: 10.2478/jses-2025-0007

UNCOVERING PATTERNS OF INFORMALITY AND SKILLS IN THE EU THROUGH CLUSTER ANALYSIS

Maria-Bianca BOLBOAȘĂ[a], Marina-Diana AGAFIȚEI[b1], Adriana AnaMaria DAVIDESCU[c]

[a, b, c] *The Bucharest University of Economic Studies, Romania;*
National Scientific Research Institute for Labour and Social Protection, Bucharest, Romania

Abstract

The informal economy remains a persistent structural challenge across European countries, closely linked to governance quality, socio-economic development and skill disparities. This study examines how informality relates to skills by combining four alternative estimates of the informal economy with a wide set of educational, digital and institutional indicators. Using Elastic Net regularisation, we identify the most robust determinants of economic performance and subsequently apply k-means clustering to classify 26 European countries according to their profiles of informality and skills. Tertiary education, regulatory quality, digitalisation and life expectancy emerge as recurrent structural factors. The results consistently reveal two clusters: one comprising high-capacity, high-skill, low-informality countries in Western and Northern Europe, and another including lower-skill, higher-informality countries in Southern and Eastern Europe. The study contributes new evidence to a relatively limited European literature on informality and skills, highlighting the need for more granular regional analyses to capture internal heterogeneity and better inform targeted policy interventions.

Keywords: Skills, Competencies, Informal Economy, Elastic Net, Cluster Analysis

JEL Classification: O17; J24; C38

Acknowledgments: The research study was elaborated within the Business Economics Data Science Lab of the Bucharest University of Economic Studies, within the project 'Analysis of the Economic Recovery and Resilience Process in Romania in the Context of Sustainable Development', EconST2025.

1. Introduction

Across the European Union, the persistence of informal economic activity continues to challenge policymakers, researchers, and labour-market institutions. Despite decades of regulatory harmonisation, digital transformation, and rising educational attainment, informality remains deeply embedded in the economic and social fabric of many Member States. This resilience raises an important question: why do substantial differences in informality endure within one of the world's most integrated economic blocs? Understanding the answer requires moving beyond single-factor explanations to explore the complex interplay between skills, governance, socio-economic structures, and digital capacity.

Existing research has provided valuable insights into the nature and determinants of informality, yet important gaps remain. A large body of work emphasises its conceptual ambiguity and the challenges of measurement, noting that informal, shadow, undeclared, or hidden economic activities capture related but distinct phenomena (International Monetary Fund, 2021). At the same time, recent contributions by Medina and Schneider (2024) shift the debate toward structural drivers of informality, such as regulatory

[1]Corresponding author,Marina-Diana Agafiței – diana.agafitei@csie.ase.ro

burden, institutional trust, and labour-market conditions, while also pointing out that education, though consistently correlated with informality, remains underrepresented in cross-country analyses.

Parallel evidence highlights the role of digitalisation, governance quality, and social vulnerability in shaping informal employment patterns. Recent studies have shown that digital tools, e-governance systems, and online financial infrastructures may either reduce informality by increasing transparency and formal participation or, paradoxically, facilitate new forms of informal work. Moreover, Agafiței, Bolboașă and Davidescu (2024) provide robust empirical support for this duality, demonstrating, through a LASSO-based approach applied to Central and Eastern European countries, that governance factors such as the rule of law exert a far stronger influence on informality than digitalisation indicators themselves. Their results underscore that digitalisation alone is insufficient: its effectiveness depends critically on institutional quality and the broader socio-economic environment.

However, these strands of literature often evolve separately. Very few studies systematically examine how skills ecosystems, including tertiary education, digital competencies, and training opportunities, interact with informal economic structures across EU Member States. This omission is increasingly consequential as Europe undergoes simultaneous transitions in the digital, green, and demographic landscape.

Moreover, while the EU is often treated as a relatively homogeneous institutional space, the reality is far more variegated. Deep structural divides persist between Western and Northern European countries, which typically exhibit stronger governance and lower levels of informality, and several Southern and Eastern European economies, where informal work remains a widespread livelihood strategy. Capturing the heterogeneity of these national profiles requires multidimensional, data-driven approaches that can disentangle how skills, institutions, and informality co-evolve.

This paper addresses these gaps by offering one of the first integrated empirical assessments of informality and skills across 26 European countries using a combination of Elastic Net regularisation and k-means clustering. Building on recent advances in informality measurement (Asllani, Schneider &Dell'Anno, 2024), we incorporate four alternative estimates of the informal economy alongside a rich set of educational, digital, socio-economic, and governance indicators. This methodological strategy allows us not only to identify the most robust structural correlates of informality but also to uncover distinct clusters of European countries sharing similar profiles of skills, governance capacity, and informal-economy dynamics.

Our contribution is threefold. First, we integrate skills, particularly tertiary education and digital competencies, into the analysis of informality, responding to the call in recent literature to better understand the role of human capital in shaping undeclared work. Second, we leverage advanced regularisation techniques to navigate multicollinearity and small-sample constraints, extracting the most informative predictors from a wide conceptual space. Third, by clustering countries according to their structural characteristics, we provide a new comparative typology that reveals persistent divides, unexpected country positions, and policy-relevant patterns.

In doing so, the study demonstrates that informality is neither an isolated labour-market niche nor merely a symptom of weak enforcement, but rather a reflection of broader ecosystems of skills, institutions, and socio-economic development.

2. Literature review

The informal economy has emerged as one of the most persistent structural features of contemporary labour markets, cutting across levels of development, institutional systems, and socio-economic contexts. Scholarly work has long emphasised its conceptual ambiguity, with multiple partially overlapping definitions complicating empirical analysis. The IMF's analytical framework on measuring informality demonstrates that terms such as informal, shadow, underground and hidden economy refer to distinct yet interrelated phenomena, each capturing different facets of non-regulated production and employment (International Monetary Fund, 2021). Complementing this definitional perspective, Medina and Schneider (2024) place stronger emphasis on the determinants of informal economic activity, highlighting how institutional quality, tax morale, regulatory burden and labour-market conditions interact in shaping informality. Their work also underscores that education, though consistently correlated with informality, is frequently underrepresented in macro-comparative models, suggesting that human capital remains an undervalued dimension in explaining cross-country variation in informal employment. The International Monetary Fund, therefore, advocates an SNA-aligned conceptualisation that incorporates unincorporated or unregistered production units, informal wage employment, self-employment without formal registration, and own-use production.

A central empirical anchor in the global debate is the Women and Men in the Informal Economy: A Statistical Brief, produced jointly by the ILO and WIEGO. The report shows that 61% of global workers, amounting to over two billion people, are engaged informally, with higher prevalence in developing and emerging economies, but with substantial variation even within high-income countries. The publication provides disaggregated insights into gender, age, poverty and occupational patterns, illustrating that informality disproportionately affects women, young workers, older workers, migrants and individuals in low-income households. Informal work is also shown to vary sharply across sectors, with agriculture dominating rural informality while urban informality is concentrated in services, micro-enterprise activity, home-based production and domestic work. These descriptive regularities are consistent with findings from the ILO's broader analyses of informal employment dynamics, which outline global patterns of vulnerability, limited access to social protection and regulatory exclusion (Bonnet, Vanek and Chen, 2019).

Within Europe, the phenomenon of undeclared and unregistered employment has been studied extensively. Williams (2021) argues that informality is less a failure of enforcement and more a reflection of institutional asymmetry, a misalignment between formal rules and citizens' normative expectations. This perspective aligns with earlier multi-level analyses showing that both individual-level characteristics (education, hardship, tax morale) and country-level institutional factors (governance quality, administrative efficiency, corruption) shape the propensity to engage in undeclared work (Williams &Horodnic, 2018; Krasniqi& Williams, 2017; Williams &Horodnic, 2017). Complementary evidence from European service industries refines this understanding: informal employment is particularly pronounced in hospitality, cleaning, care work and small-scale retail sectors, characterised by seasonal demand, labour intensity and structural cost pressures, which often make informal hiring more attractive than formal compliance (Williams &Horodnic, 2018). These studies converge toward a view in which informality is embedded not only in economic incentives but also in socio-institutional contexts marked by low trust, weak governance and perceived unfairness of the tax-benefit system.

A parallel strand of literature addresses the economic and social determinants of informality through macro- and microeconometric frameworks. Several works, including those drawing on human capital theory as formulated by Becker (1962), along with more recent analyses by the International Monetary Fund (2021), examine how individuals weigh risks, expected earnings and perceived institutional legitimacy when deciding between formal and informal engagement. While Becker's contribution lies in conceptualising how education and skills shape labour-market behaviour more broadly, contemporary empirical studies apply and extend these foundations to understand participation in informal activities. This body of research also highlights that the informal economy often functions as a buffer during adverse shocks, providing countercyclical employment in contexts where formal labour markets fail to absorb displaced workers, yet simultaneously perpetuating low productivity and persistent vulnerability.

Human capital theory provides an important conceptual foundation for understanding the labour-market dynamics that underpin informal employment. Yet this theoretical framework has been increasingly scrutinised. For instance, Auerbach and Green (2024) argue that mainstream models overstate the causal link between education and productivity while neglecting how structural inequalities shape access to skills, occupational opportunities and earnings trajectories. Complementarily, (Marginson, 2019) emphasises that skill acquisition is embedded in broader socio-political and institutional contexts that cannot be reduced to an investment decision by individuals. These critiques are salient in the context of informality: skill deficits are not merely individual shortcomings but outcomes of systemic exclusion, and returns to education are highly uneven across formal and informal segments of the labour market.

Empirical research on skills development in the informal sector offers further nuance. The World Bank's analyses of Sub-Saharan Africa demonstrate that workers in informal employment often acquire skills through traditional apprenticeships, informal training networks and learning-by-doing, which lack certification, quality assurance and labour mobility potential (Adams, Johansson de Silva and Razmara, 2013a; 2013b). Without structured pathways to upgrade skills, workers remain trapped in low-productivity activities with limited prospects for formalisation. These dynamics resonate with global policy analyses, such as Comyn's review of skills, employability and lifelong learning in the SDGs and the 2030 labour market, which argues that achieving SDG 4 (inclusive and equitable quality education) and SDG 8 (decent work and economic growth) requires systemic reforms expanding access to lifelong learning, digital skills and labour-market-relevant competencies (Comyn, 2018).

More recent scholarship emphasises evolving forms of informality generated by transformations in the nature of work. Matcu-Zaharia et al. (2024) show that although self-employment and platform-mediated work are formally classified within the standard labour framework, they often mirror informal employment in terms of income instability, exclusion from social protection, inadequate representation and exposure to economic shocks. These hybrid forms of informality complicate the traditional binary distinction between formal and informal work, suggesting that indicators of precarious self-employment, flexibilised labour arrangements and digital platform dependence must be integrated into contemporary analyses.

Digitalisation itself plays a dual role in shaping informality. On one hand, digital public services, automated tax systems, e-government platforms and electronic payment infrastructures reduce transaction costs and increase the traceability of economic activity, creating opportunities for formalisation. On the other hand, digital platforms facilitate new forms of informal or semi-formal work that escape conventional regulatory frameworks, as noted in the IMF's analysis of measurement challenges (International Monetary Fund, 2021). This duality implies that digitalisation does not automatically lead to greater formalisation or increased informality; instead, its effects depend on institutional readiness, regulatory adaptation, and the acquisition of digital skills.

3. Data, models and methods

3.1. *Data*

This study relies on a diverse set of indicators compiled from multiple sources, including Eurostat, the European Innovation Scoreboard, Eurobarometer, Cedefop, and the Worldwide Governance Indicators. To estimate the size of the informal economy, we use the four estimates from Asllani et al. (2024), each calibrated with a different benchmark. The analysis also incorporates a wide range of socioeconomic variables, measures of digital and green skills, and governance indicators, in accordance with the literature. A detailed description of all variables, units of measurement, and data sources is provided in the Appendix. All variables are used for the latest available year and cover 26 European countries: Belgium, Bulgaria, Czechia, Denmark, Germany, Estonia, Ireland, Greece, Spain, France, Croatia, Italy, Cyprus, Latvia, Lithuania, Luxembourg, Hungary, the Netherlands, Austria, Poland, Portugal, Romania, Slovenia, Slovakia, Finland, and Sweden.

3.2. *Methodology*

To examine how European countries compare in terms of informal economic activity and skill-related characteristics, and to identify groups of nations with similar structural profiles, we follow a two-step methodological approach: (1) correlation-based variable reduction and Elastic Net variable selection, and (2) cluster analysis. This strategy ensures that the final set of indicators is both statistically robust and conceptually meaningful for characterising national profiles of informality and skills. The empirical analysis is conducted at the country level and includes all European Union member states for which consistent and comparable data are available. Malta is excluded from the analysis due to the absence of reliable estimates of the informal economy across all four measurement approaches considered, resulting in a final sample that covers the full EU with the exception of this single case.

We begin by constructing the full correlation matrix for all indicators related to education, digitalisation, governance, inclusion, and the four estimates of the informal economy. Because several variables are conceptually close and empirically overlapping, we remove those that display very high pairwise correlations (above 80%). This screening step reduces multicollinearity and retains only the most informative indicators. After this process, the working dataset includes 17 core variables that cover all structural dimensions relevant to the study. Given the country-level nature of the analysis and the inherently limited number of observations at this level of aggregation, this initial reduction step is essential for ensuring model stability and interpretability.

Even after correlation filtering, strong associations persist among some indicators. To further refine the set of explanatory variables and understand which factors are most strongly connected to economic development, we apply Elastic Net regression. Elastic Net offers several advantages, including its ability to handle multicollinearity, perform simultaneous variable selection and regularisation, stabilise estimates in the presence of highly correlated predictors, and outperform LASSO or Ridge when neither penalty alone is sufficient (Zou and Hastie, 2005), as can be seen in equation 1:

$$\hat{\beta} = arg\ arg\ \left\{\frac{1}{2n}\sum_{i=1}^{n}(y_i - x_i^{\top}\beta)^2 + \lambda[\alpha||\beta||_1 + (1-\alpha)||\beta||_2^2]\right\} \quad (1)$$

Where: y is the dependent variable, x_i is the vector of predictors, $\lambda>0$ controls the degree of regularisation, and $\alpha\in[0,1]$ determines the balance between LASSO (l_1) and Ridge (l_2) shrinkage. Importantly, Elastic Net is particularly well suited to empirical settings characterised by a relatively small number of observations combined with a comparatively large set of potentially correlated predictors, as is typical in cross-country analyses. Its regularisation properties mitigate overfitting risks and allow for reliable variable selection even when traditional regression techniques would be inappropriate.

Table no. 1: Variables included in Elastic Net regression

Predictors	Estimates of the IE using Dell'Anno (2024); NA projections and ILO (2019) for calibration
	Estimates of the IE using Elgin et al. (2021) - DGE for calibration
	Estimates of the IE using Elgin et al. (2021) - MIMIC for calibration
	Estimates of the IE using Medina and Schneider (2021) – MIMIC for calibration
	Population with tertiary education
	Employment of young persons not in education and training
	At-risk-of-poverty rate
	Material and social deprivation rate
	Percentage of Enterprises that employ ICT specialists
	Persons in long-term unemployment (% of population in the labour force)
	Early leavers from education and training
	Participation rate in education and training
	Life expectancy
	Individuals with above basic overall digital skills
	Proportion of individuals with digital skills for job activities
	Lack of training opportunities as the main barrier to improving digital skills
	Lack of clarity about which digital skills need improvement as the main barrier for improving digital skills
	Digitalness
	Digital pervasiveness
	Greenness
	Regulatory Quality
Independent variable	GDP per capita in PPS

Source: Authors' own contribution

We use GDP per capita as the dependent variable because it anchors the selected indicators in a broad macroeconomic context (Table no. 1). Importantly, we repeat the Elastic Net procedure four times, each time substituting one of the alternative measurements of the informal economy available in Asllani et al. (2024). This allows us to account for uncertainty regarding the true level of informality and to identify variables that consistently emerge as significant regardless of the chosen estimate. This repeated-estimation strategy further enhances robustness by ensuring that results are not driven by a single specification or by idiosyncrasies of one particular informality measure, which is especially relevant in small-sample cross-country settings. Moreover, it is important to note that GDP per capita is employed in the Elastic Net framework as an anchoring outcome variable to identify the set of structural indicators most robustly associated with overall economic performance, rather than to estimate causal effects of informality or skills. Consequently, references to persistence and structural features should be interpreted as descriptive patterns emerging from the combined regression–clustering framework, not as causal inferences derived from a growth regression.

Elastic Net hyperparameters were selected using 10-fold cross-validation (CV). For each candidate α, λ was chosen by the CV-min criterion; the selected (α, λ) and the number of non-zero coefficients are calculated by Stata 17. Model performance is documented using MSE and R^2 for the penalised fit, and we additionally calculate postselection fit after refitting on the selected predictors as a robustness check. Finally, we assessed robustness by re-estimating the Elastic Net under alternative specifications (e.g., forcing key variables to remain, swapping correlated predictors), and by varying the estimates/proxies of the informal economy to identify those macroeconomic factors that remain consistently relevant when the informality dimension is measured in different ways, comparing the stability of the selected predictors across these specifications.

In the final stage, we classify European countries into groups based on the variables identified through the Elastic Net. All indicators are standardised to ensure comparability and avoid scale effects. We use the k-means clustering algorithm, which partitions countries into internally homogeneous groups based on similarity in their profiles of informality, education, digital skills, governance, and inclusion. We apply the k-means clustering algorithm following the approach described by AlboukadelKassambara (2017). The method partitions observations into k groups by iteratively assigning each country to the nearest cluster centroid and then updating the centroids based on the current group memberships. This process of cluster assignment and centroid recalculation continues until convergence, that is, until no further changes occur in the cluster structure.

To determine the appropriate number of clusters, we rely on two complementary diagnostic tools: the elbow plot and the silhouette score. The elbow plot identifies when additional clusters no longer significantly reduce within-group variability, while the silhouette score assesses how well each country fits within its assigned cluster relative to neighbouring clusters. The corresponding elbow and silhouette diagnostics are reported in the Appendix.

This combined approach enables us to identify coherent clusters of countries that share similar structural characteristics and to analyse patterns linking informal economic activity with education, digitalisation, and governance outcomes. In other words, Elastic Net serves as a supervised feature-selection step, while k-means provides an unsupervised typology based on the retained informality-related macro-structural features. Overall, the methodological design explicitly addresses the constraints imposed by the small number of observations inherent to country-level analyses by combining dimensionality reduction, regularised regression, and clustering techniques that are robust in high-dimensional, low-sample contexts.

4. Results and discussions

This section presents the empirical results of the analysis in two steps. First, we present the results of the Elastic Net estimations, focusing on the variables that emerge as robustly associated with GDP per capita across the four alternative measures of the informal economy. Varying the informal economy estimates is explicitly used as a robustness strategy to determine which macroeconomic, governance, education, and digitalisation factors remain consistently selected across quantifications of informality. These results allow us to identify the key dimensions linking informality, skills, and macroeconomic environment performance, despite multicollinearity and the structure of the dataset.

The results presented in Table 2 reveal some differences in the explanatory power of the four alternative measures of the informal economy. Because Elastic Net estimates are penalised and primarily used for variable selection in a small-sample, multi-predictor setting, we do not interpret coefficient magnitudes as classical marginal effects. Instead, we interpret the results in terms of (i) which variables enter the active set (non-zero coefficients), (ii) the sign of these coefficients, and (iii) the stability of selection across alternative informal economy measures and specification checks. In this sense, Table 2 is used to identify a smaller, robust subset of informality-related macro-structural indicators to be carried forward to the clustering step.

Table no. 2: Elastic Net results for the four models

	Model 1: Estimates of the IE using Dell'Anno (2024); NA projections and ILO (2019) for calibration		**Model 2: Estimates of the IE using Elgin et al. (2021) - DGE for calibration**		**Model 3: Estimates of the IE using Elgin et al. (2021) - MIMIC for calibration**		**Model 4: Estimates of the IE using Medina and Schneider (2021) – MIMIC for calibration**	
Selected variables	Regulatory Quality	0.2854	**Estimates of the IE using Elgin et al. (2021) - DGE for calibration**	**-0.339**	**Estimates of the IE using Elgin et al. (2021) - MIMIC for calibration**	**-0.225**	Population with tertiary education	0.2244
	Population with tertiary education	0.2284	Population with tertiary education	0.2538	Population with tertiary education	0.2120	Regulatory Quality	0.2057
	Digital pervasiveness	0.1475	Regulatory Quality	0.1451	Regulatory Quality	0.1815	**Estimates of the IE using Medina and Schneider (2021) – MIMIC for calibration**	-0.1288
	Life expectancy	0.1127	Digital pervasiveness	0.0024	Life expectancy	0.0514	Digital pervasiveness	0.0760
	Percentage of Enterprises that employ ICT specialists	0.0538	Life expectancy	0.0023	Digital pervasiveness	0.0389	Life expectancy	0.0758
	Persons in long-term unemployment	-0.007					Percentage of Enterprises that employ ICT specialists	0.0084
N	26		26		26		26	
No. of predictors	18		18		18		18	
Selected α	0.500		1.000		0.500		0.500	
Selected λ	.2422		.1546		.3201		.3201	
R²	0.6105		0.6279		0.6002		0.5783	
MSE	.3745256		.357762		.3844675		.4055259	

Source: Authors' own contribution

In Model 1, which relies on the Dell'Anno (2024); NA projections and ILO (2019) for calibration, the informal economy estimate does not appear among the selected variables, indicating that this measure does not provide additional explanatory value for GDP per capita once other structural factors are taken into account. Because the IE estimate is not retained as a relevant predictor in this specification, Model 1 does not offer a consistent basis for identifying country profiles of informality.

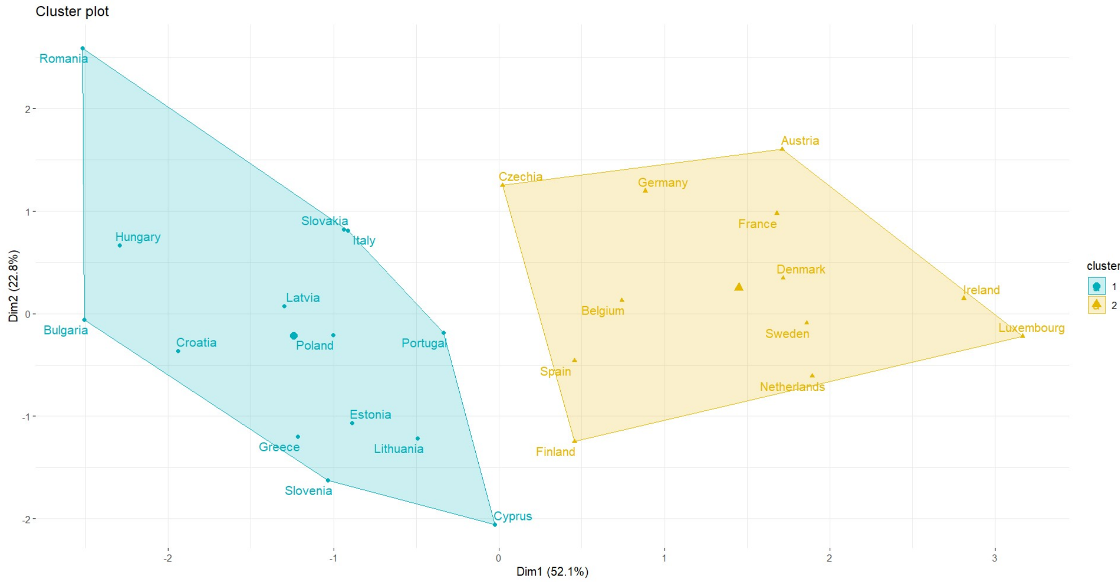
By contrast, the IE estimates used in Models 2 and 3, based on Elgin et al. (2021) DGE and MIMIC calibrations, emerge as the strongest predictors in their respective regressions. In both cases, the informal economy is selected with the largest coefficient magnitude, confirming that these estimates capture substantial macro-structural variation across European countries. Their negative signs are consistent with theoretical expectations: higher levels of informality are associated with lower economic outcomes, reflecting reduced tax revenues, productivity constraints and weaker institutional environments. These two models, therefore, position informality as a central structural driver of cross-country differences in GDP per capita.

In Model 4, the IE estimate calibrated using Medina and Schneider's (2021) MIMIC is also retained, although its magnitude is smaller than in Models 2 and 3. This suggests that while the measure still helps explain economic performance, its explanatory power is less dominant. Nevertheless, the fact that IE is selected in three out of four models strengthens the evidence that informality is systematically linked to macroeconomic outcomes in Europe when robust calibration methods are applied.

Because the IE estimate is not retained as a relevant predictor in this specification, Model 1 provides limited leverage for deriving informality-centred country profiles. Therefore, we focus the clustering exercise on Models 2–4, where informality is explicitly retained among the selected features.

Across Models 2, 3 and 4, a set of structural determinants emerges, underscoring the interplay between human capital, governance, digitalisation and broader socio-economic development. Importantly, the same core variables are preserved across all three models, indicating a stable selection pattern despite differences in the underlying informal economy measure. The population with tertiary education is consistently selected, with sizeable positive coefficients, highlighting the pivotal role of advanced human capital in shaping economic performance. Regulatory Quality also appears systematically, underscoring the fundamental importance of effective governance, rule enforcement and institutional capacity in influencing both formal economic activity and the extent of informality. Although with a smaller impact, digital pervasiveness is retained across all models, signalling the growing relevance of digital skills and digitally intensive labour demand in differentiating countries' economic structures. Finally, life expectancy repeatedly enters the model as a proxy for long-term welfare and social development, further reinforcing the idea that economic prosperity, institutional quality and human wellbeing are tightly interlinked. Taken together, these results illustrate that informality is not an isolated phenomenon but part of a broader structural configuration that connects skills, governance, digital transformation and socio-economic conditions.

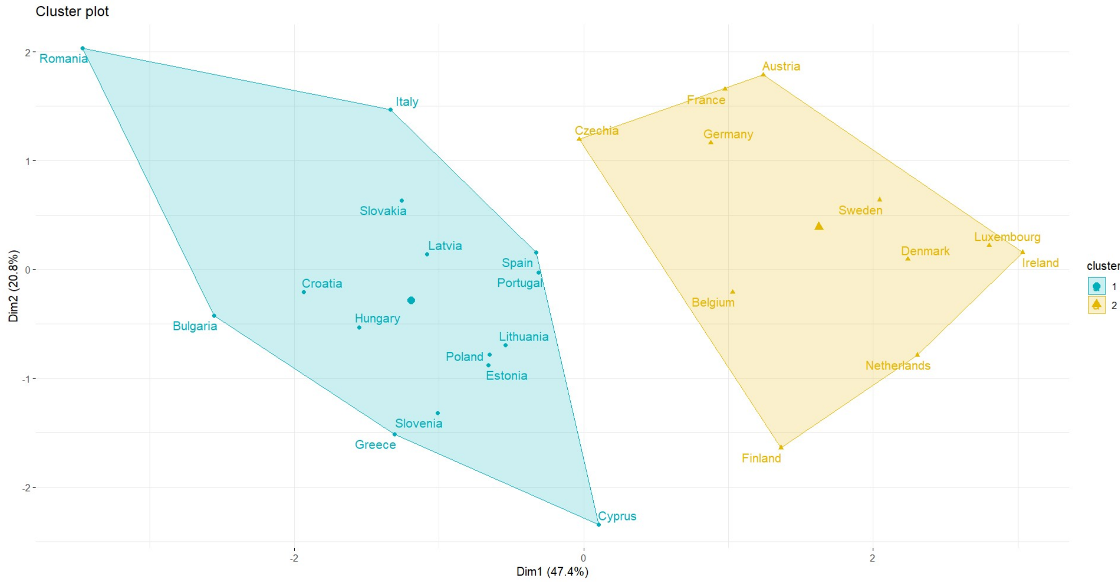
To ensure methodological coherence between the two steps, the variables retained by Elastic Net in each specification are used as inputs to k-means; we then compare cluster membership across Models 2-4 to evaluate whether the typology is robust to alternative informal economy measures and the associated selected feature sets.


Cluster plot
Romania
Hungary
Slovakia
Italy
Latvia
Bulgaria
Croatia
Poland
Portugal
Estonia
Greece
Lithuania
Slovenia
Cyprus
Czechia
Germany
Austria
France
Denmark
Belgium
Ireland
Sweden
Luxembourg
Spain
Netherlands
Finland
cluster
1
2
Dim2 (22.8%)
Dim1 (52.1%)

(a)


Cluster plot
Cyprus
Slovenia
Greece
Estonia
Lithuania
Finland
Netherlands
Spain
Croatia
Luxembourg
Ireland
Sweden
Poland
Latvia
Bulgaria
Portugal
Denmark
Belgium
Italy
Hungary
Slovakia
France
Germany
Czechia
Austria
Romania
cluster
1
2
Dim2 (22.6%)
Dim1 (51.7%)

(b)



(c)

Figure no. 1: K-means clustering for models 2 (a), 3 (b) and 4 (c)

Source: Authors' own contribution

Across the three models, the clustering results display a stable pattern, suggesting that the country groupings are robust to the choice of informal economy measure and to the inclusion of additional covariates. In all cases, k-means yields two clearly differentiated clusters that broadly align with a high-capacity, high-skill, low-informality group, comprising mainly Western and Northern European economies such as Germany, France, Denmark, Sweden, Ireland, the Netherlands, Luxembourg and Austria, and a lower-capacity, lower-skill, higher-informality group, dominated by Southern and Eastern European countries including Bulgaria, Romania, Greece, Croatia, Portugal, Poland, Latvia and Lithuania. While the exact position of a few borderline countries (like Spain for the third model, when percentage of Enterprises that employ ICT specialists included) shows some sensitivity to the underlying informality estimate and the variable set, the overall configuration remains consistent: higher regulatory quality, tertiary education, digital pervasiveness and life expectancy systematically co-occur with lower informal economy levels, whereas weaker governance and skills are associated with structurally higher informality.

An interesting case is Estonia, which, despite being one of the most digitally advanced countries in Europe (Ministry of Economic Affairs and Communication Estonia, 2021) and consistently achieving strong educational outcomes (OECD, 2024a; OECD, 2024b; OECD, 2024c), falls within the cluster of Southern and Eastern European economies. This counterintuitive placement reflects the multidimensional nature of the variables used: although Estonia scores exceptionally well on digital infrastructure, e-government, and ICT uptake, it continues to exhibit higher levels of informality than Western and Nordic countries. Estonia's position, therefore, illustrates that even highly digitalised economies can remain closer to Eastern and Southern European patterns when informality, governance capacity, and demographic conditions are jointly taken into account.

A noteworthy case in the clustering results is Czechia, a Central and Eastern European country that is consistently grouped with Western economies rather than with its regional peers. This positioning reflects Czechia's comparatively strong governance performance and lower levels of informality, which pull it closer to the institutional profile of high-capacity EU members. Although Czechia records one of the lowest shares of tertiary-educated adults (around 30%), this limitation is outweighed in the clustering algorithm by its relatively robust regulatory quality, institutional effectiveness, and moderate informality levels. These characteristics distinguish Czechia from other CEE countries and allow it to converge toward the Western European cluster, illustrating how governance strength and low informal activity can compensate for weaker educational attainment when structural profiles are assessed multidimensionally.

Romania appears as one of the most distant members of the higher-informality cluster. Its placement is driven by persistently high informal economy estimates, lower regulatory quality, and gaps in human capital and demographic indicators such as life expectancy.

5. Conclusions

This study examined how European countries differ in their levels of informality and skills, using Elastic Net regression and k-means clustering to capture the multidimensional nature of informal economic activity. The results show that informality remains a persistent structural feature of European labour markets, strongly shaped by governance quality, socio-economic development, and human capital. Although education is not emphasised as a primary determinant in studies such as Medina and Schneider (2021), our findings reveal that tertiary education systematically contributes to lower informality and higher economic performance, underscoring its underappreciated role in comparative studies of undeclared work.

The analysis also highlights a notable gap in the European literature: very few studies explore the link between informality and skills, despite clear evidence that digitalisation, educational attainment and labour-market competencies are closely intertwined with informal employment patterns. The consistent emergence of two country clusters, one high-skill and low-informality, the other low-skill and high-informality, further illustrates how skill ecosystems and institutional conditions jointly shape informal work. These clusters capture stable configurations of co-occurring characteristics across countries, rather than direct determinants of informality, and should therefore be read as typological profiles that summarise structural similarities within the EU.

From a broader analytical perspective, the results underscore the value of combining regularised regression with clustering techniques when investigating complex, multidimensional phenomena such as informality. By using GDP per capita as an anchoring variable for feature selection rather than as a target of causal interpretation, the analysis is able to identify a smaller set of structural indicators that meaningfully differentiate country profiles. The subsequent clustering exercise shows that these indicators tend to co-occur in stable configurations across countries, reinforcing the view that informality should be analysed as part of an interconnected structural system rather than through isolated explanatory factors. This integrated approach offers a replicable framework for future comparative studies seeking to map structural heterogeneity under data and sample-size constraints.

The study has several limitations. First, the reliance on national-level data prevents capturing important within-country heterogeneity, as regional disparities in informality, skills, and institutional capacity remain unobserved. Increasing the number of observations through a regional-level analysis would substantially enrich the empirical framework and allow for a more fine-grained assessment of structural differences. However, at present, such an extension is constrained by the absence of consistent and comparable estimates of the informal economy at the regional level across European countries. Second, the analysis is based on 26 national observations, which necessarily limits the complexity of the modelling strategy. Future research could address these constraints as new regional datasets and longitudinal measures of informality become available.

Overall, the findings reaffirm that informality is multifaceted and deeply embedded in broader structures of governance, education and digital capacity - areas that must be addressed together to design effective policy responses. Rather than establishing causal mechanisms, the study provides an empirically grounded mapping of how informality, skills, and institutional features are jointly configured across European countries, offering a foundation for future hypothesis-driven and causally oriented research.

References

Adams, A.V., Johansson de Silva, S., &Razmara, S. (2013b). Strategies for improving skills in the informal sector. In Improving skills development in the informal sector: Strategies for Sub-Saharan Africa. Washington, D.C.: World Bank, 91–116. doi: 10.1596/978-0-8213-9968-2.

Adams, A.V., Johansson de Silva, S.,&Razmara, S. (2013a). The role of skills in the informal sector.In Improving skills development in the informal sector: Strategies for Sub-Saharan Africa. Washington, D.C.: World Bank, 61–90. doi: 10.1596/978-0-8213-9968-2.

Agafiței, M. D., Bolboașă, M. B., &Davidescu, A. A. (2025, May). Shadow Economies in the Digital Age: A LASSO Approach to Measuring the Impact of Digitalization on Informality in Central and Eastern European Countries. In International Conference on Digital Economy (250-273). Cham: Springer Nature Switzerland.

Alban Asllani, Schneider, F.G., &Dell'Anno, R. (2024). Mapping the Informal Economy Around the World with an Enhanced MIMIC Approach: New Estimates for 110 Countries from 1997-2022. doi:https://doi.org/10.2139/ssrn.5041391.

AlboukadelKassambara (2017). Practical guide to Cluster Analysis in R : unsupervised machine learning. S.L.: Createspace Independent Publishing Platform.

Auerbach, P.,& Green, F. (2024). Reformulating the critique of human capital theory. Journal of Economic Surveys, 39(5), 1839–1851. doi: 10.1111/joes.12675.

Becker, G.S. (1962). Investment in human capital: a theoretical analysis.In Universities-National Bureau Committee for Economic Research (ed.) Investment in human beings. Chicago: University of Chicago Press, 9–49.

Bonnet, F., Vanek, J.,& Chen, M. (2019). Women and men in the informal economy: a statistical brief. Manchester: WIEGO.

Comyn, P.J. (2018). Skills, employability and lifelong learning in the Sustainable Development Goals and the 2030 labour market. International Journal of Training Research, 16(3), 200–217. doi: 10.1080/14480220.2018.1576311.

International Monetary Fund (2021). Measuring the informal economy. Washington, D.C.: IMF Policy Paper.

Krasniqi, B.A.,& Williams, C.C. (2017). Explaining individual- and country-level variations in unregistered employment using a multi-level model: evidence from 35 Eurasian countries. South East European Journal of Economics and Business, 12(2), 61–71. doi: 10.1515/jeb-2017-0017.

Marginson, S. (2019). Limitations of human capital theory. Studies in Higher Education, 44(2), 287–301. doi: 10.1080/03075079.2017.1359823.

Matcu-Zaharia, M., Horodnic, I.A., Williams, C.C.,&Nistor, G.C. (2024). Self-employed workers and the achievement of the 2030 Agenda for Sustainable Development: an overview of their social benefit entitlements across 31 European countries. Sustainability, 16(7), p. 2625. doi: 10.3390/su16072625.

Medina, L.,& Schneider, F. (2024). The evolution of shadow economies through the 21st century. Washington, D.C.: International Monetary Fund.

Ministry of Economic Affairs and Communication Estonia (2021). Estonia's Digital Agenda 2030. [online] Ministry of Economic Affairs and Communication Estonia. Available at: https://www.mkm.ee/en/e-state-and-connectivity/digital-agenda-2030.

OECD (2024a). Education at a Glance 2024 Estonia. [online] OECD. Available at: https://www.oecd.org/en/publications/education-at-a-glance-2024-country-notes_fab77ef0-en/estonia_4bad887a-en.html.

OECD (2024b). New PISA results on Creative Thinking: can students think outside the box? [online] OECD. Available at: https://www.oecd.org/en/publications/new-pisa-results-on-creative-thinking_b3a46696-en.html.

OECD (2024c). Survey of Adult Skills 2023. Country note: Estonia. [online] OECD. Available at: https://www.oecd.org/en/publications/survey-of-adults-skills-2023-country-notes_ab4f6b8c-en/estonia_60a7b919-en.html.

Williams, C.C. (2021). Tackling unregistered employment in European regions: towards a variegated policy approach. Brussels: European Commission.

Williams, C.C.,&Horodnic, I.A. (2017). Evaluating the relationship between social exclusion and participation in the informal sector in the European Union. International Journal of Manpower, 38(3), 489–503. https://doi.org/10.1108/IJM-10-2015-0179

Williams, C.C.,&Horodnic, I.A. (2018). Extent and distribution of unregistered employment in the service industries in Europe. The Service Industries Journal, 38(11–12), 856–874. doi: 10.1080/02642069.2018.1481209.

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. doi:https://doi.org/10.1111/j.1467-9868.2005.00503.x.

Appendix

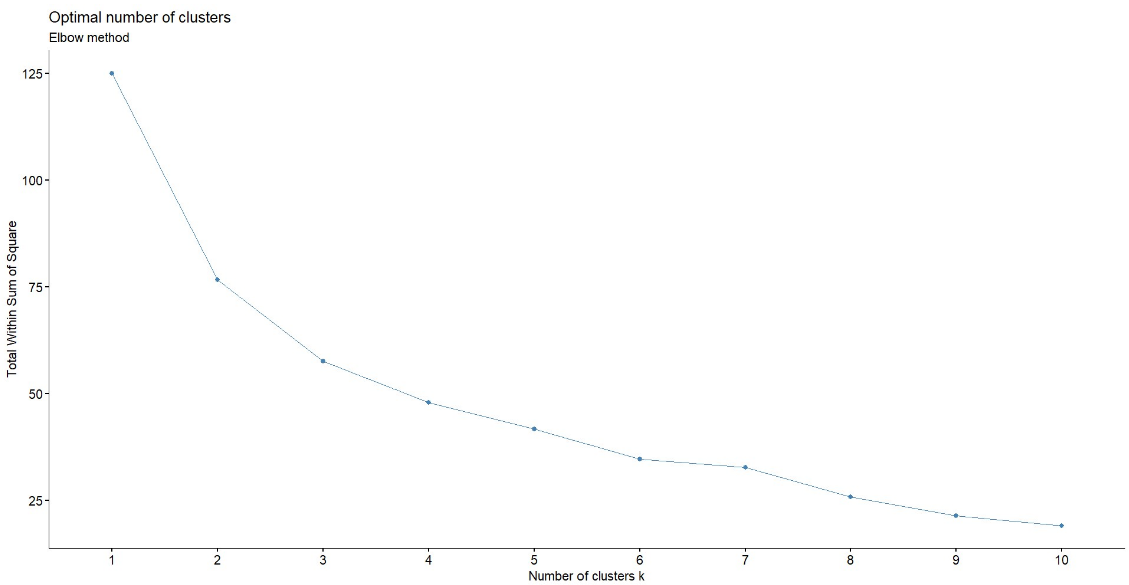
Table no. A1: Variables description

Variables	Description	Units of measure	Source
Estimates of the IE using Dell'Anno (2024); NA projections and ILO (2019) for calibration	Estimates of the size of the informal economy using the MIMIC (Multiple Indicators Multiple Causes) model, calibrated with various benchmark sources: NA projections, ILO (2019), Elgin et al. (2021) using both DGE and MIMIC approaches, as well as Medina and Schneider (2021). The model relies on a broad set of economic and governance indicators covering fiscal conditions, labour market flexibility, employment, human development, and educational characteristics, social vulnerability, or institutional quality.	%	Asllani, Schneider &Dell'Anno (2024)
Estimates of the IE using Elgin et al. (2021) - DGE for calibration		%	
Estimates of the IE using Elgin et al. (2021) - MIMIC for calibration		%	
Estimates of the IE using Medina and Schneider (2021) – MIMIC for calibration		%	
Population with tertiary education	The percentage of individuals aged 25 to 34 who have completed a form of tertiary education.	%	Eurostat
Young persons neither in employment nor in education and training	The share of young people aged 15–29 who, at the time of the EU-LFS survey, are neither in employment nor participating in formal or non-formal education or training. This indicator reflects the degree of youth integration into the labour market and the education system.	%	Eurostat
Employment of young persons not in education and training	The percentage of individuals aged 15–34 who are not in education or training but are employed. Reported annually as a share of the total population in this age group, the indicator reflects the labour-market integration of youth outside the education system.	%	Eurostat
Human resources in science and technology (% population in the labour force)	The share of individuals with higher education and/or employed in science and technology fields in the total active population.	%	Eurostat
At-risk-of-poverty rate	The share of individuals with equivalised disposable income after social transfers below 60% of the national median.	%	Eurostat
Material and social deprivation rate	The share of people unable to afford at least five of the thirteen items deemed essential for a decent standard of living.	%	Eurostat
Percentage of Enterprises that employ ICT specialists	The percentage of enterprises (by NACE Rev. 2 activity) that employ ICT specialists (employees whose primary tasks involve developing, maintaining, or supporting information and communication technologies).	%	Eurostat
Persons in long-term unemployment (% of population in the labour force)	The share of the labour force aged 15–64 who have been unemployed for at least 12 months and are still actively seeking employment.	%	Eurostat
Early leavers from education and training	The share of 18–24-year-olds with no more than lower secondary education who are not pursuing any further education or training.	%	Eurostat
GDP per capita in PPS	Per-capita value of all final goods and services	PPS	Eurostat

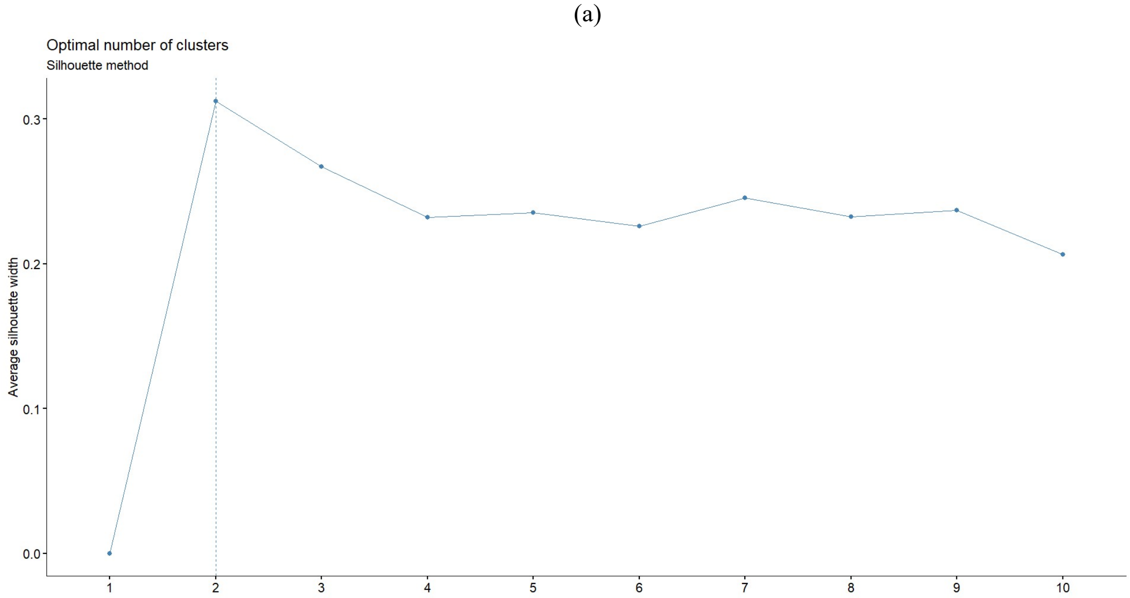
Variables	Description	Units of measure	Source
	produced in a country, adjusted for purchasing power (PPS).		
Participation rate in education and training (last 4 weeks)	The share of adults aged 25–64 who participated in formal or non-formal education or training during the four weeks preceding the EU-LFS survey, used as a measure of lifelong learning.	%	Eurostat
Life expectancy	The average number of years a newborn is expected to live if current age-specific mortality rates remain constant.	years	Eurostat
Individuals with above basic overall digital skills	The percentage of individuals aged 16–74 who possess above-basic overall digital skills. This highest skill level is determined based on a composite indicator reflecting specific online activities performed in the areas of information, communication, problem-solving, and content creation within the previous three months of the EU survey on the ICT usage in households and by individuals.	%	European Innovation Scoreboard
Proportion of individuals with digital skills for daily activities	The share of respondents in a country who state that they are sufficiently skilled in using digital technologies for daily tasks, relative to the total number of respondents in that country.	%	Eurobarometer 92.4
Proportion of individuals with digital skills for job activities	The share of respondents in a country who state that they are sufficiently skilled in using digital technologies to perform their work tasks, relative to the total number of respondents in that country.	%	Eurobarometer 92.4
Lack of training opportunities as the main barrier to improving digital skills	The share of respondents in a country who identified the lack of training opportunities as the main obstacle to developing digital skills, relative to the total number of respondents in that country.	%	Eurobarometer 92.4
Lack of clarity about which digital skills need improvement as the main barrier for improving digital skills	The share of respondents in a country who stated that they do not know which digital skills they should improve, relative to the total number of respondents in that country.	%	Eurobarometer 92.4
Digitalness	The share of digital skills out of the total number of unique skills observed in a given European region. In this study, the indicator is used at the country level, calculated as the average of regional values.	%	Cedefop
Digital pervasiveness	The proportion of online job advertisements (OJAs) in a given European region that include at least one type of digital skill, relative to all job advertisements in that region. In this study, the indicator is used at the country level, calculated as the average of regional values.	%	Cedefop
Greenness	The share of green skills out of the total number of unique skills observed in a given European region. In this study, the indicator is used at country level, calculated as the average of regional values.	%	Cedefop
Green pervasiveness	The proportion of online job advertisements (OJAs) in a given European region that include at least one green skill, relative to all job advertisements in that region. In this study, the indicator is used at the country level, calculated as the average of regional values.	%	Cedefop

Variables	Description	Units of measure	Source
Regulatory Quality Rule of Law Government Effectiveness Control of Corruption	These governance indicators assess the quality of a country's institutional environment by capturing how effectively public authorities design and implement regulations (Regulatory Quality), uphold laws and property rights (Rule of Law), deliver public services and credible policies (Government Effectiveness), and prevent the misuse of public power through corruption (Control of Corruption). Each indicator ranges from -2.5 to +2.5, where higher values reflect better governance performance and lower values indicate weaker governance.	standardized scores	World Bank

Source: Authors' own contribution



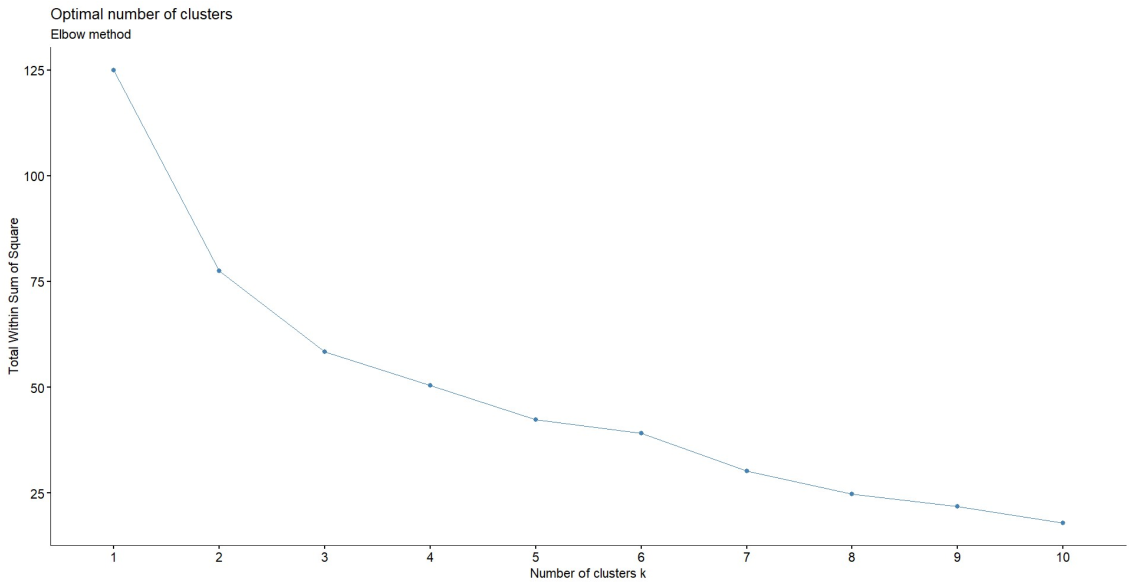
(a)



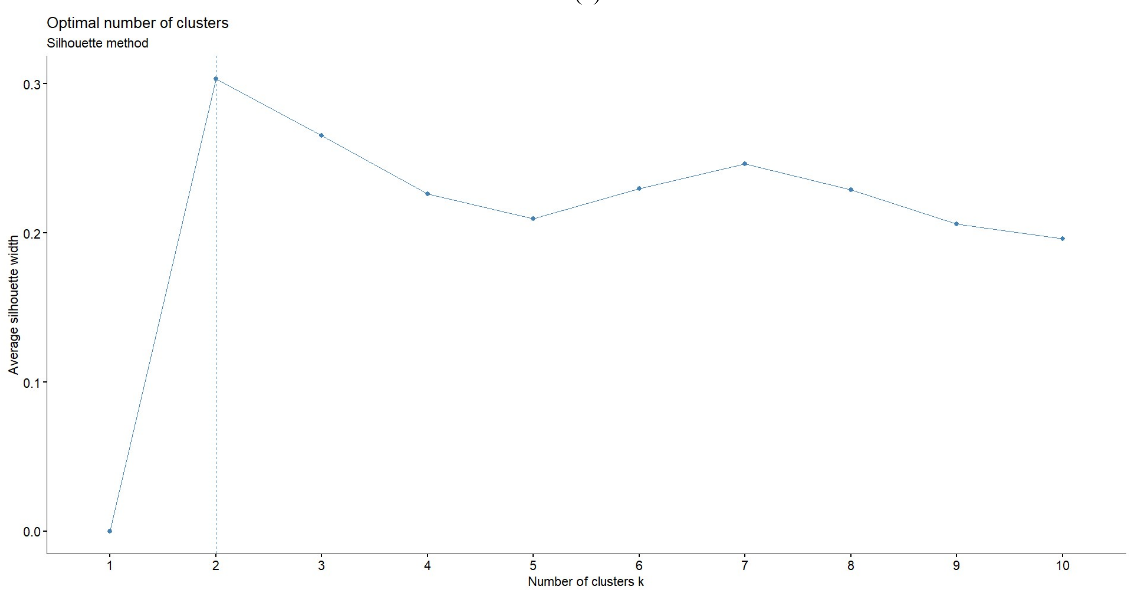
(b)

Figure no. A1: The optimal number of clusters based on the Elbow (a) and the Silhouette (b) methods for Model 2

Source: Authors' own contribution



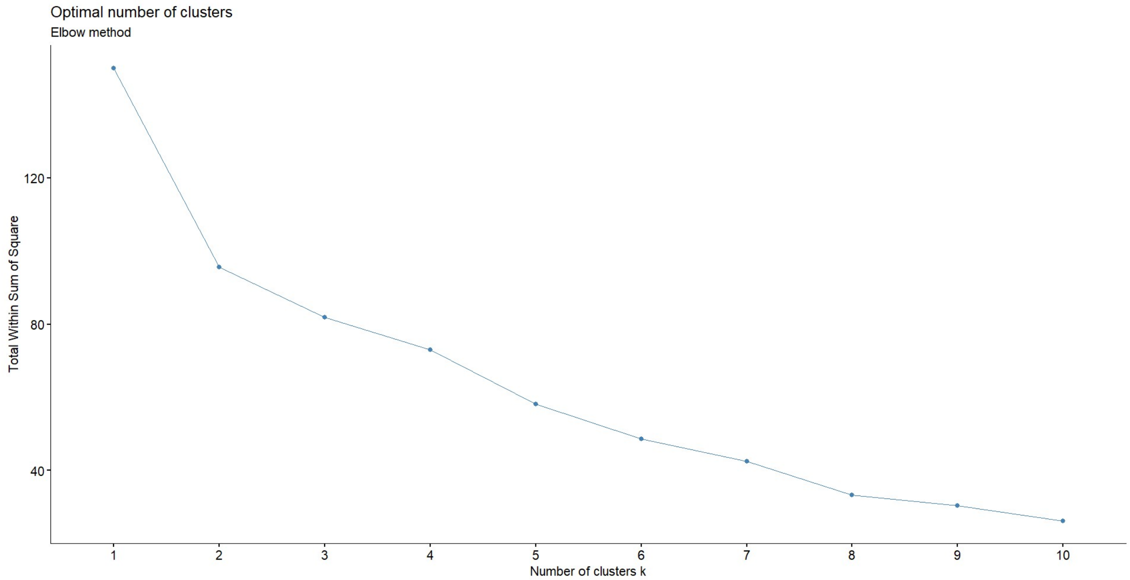
(a)



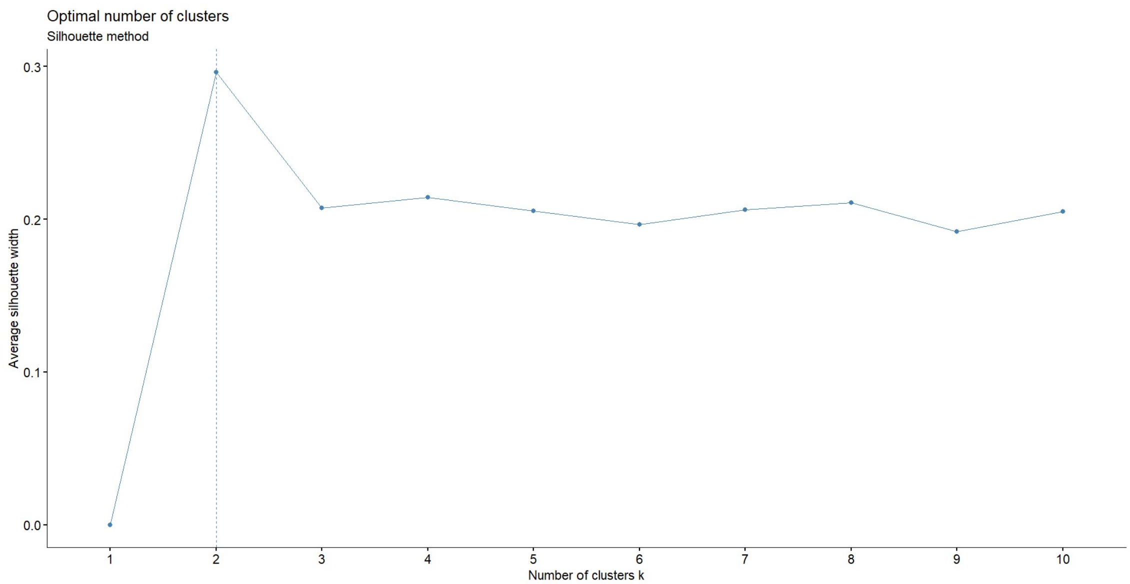
(b)

Figure no. A2: The optimal number of clusters based on the Elbow (a) and the Silhouette (b) methods for Model 3

Source: Authors' own contribution



(a)



(b)

Figure no. A3: The optimal number of clusters based on the Elbow (a) and the Silhouette (b) methods for Model 4

Source: Authors' own contribution