POVERTY RISK AND TERRITORIAL DISPARITIES IN EUROPE: A CLUSTERWISE LASSO REGRESSION APPROACH

Simona Cafieri, Gianmarco Borrata

Abstract. In 2024, more than 93 million people in the EU were at risk of poverty or social exclusion, highlighting persistent territorial disparities and the urgent need for accurate and robust measurement. This article proposes an innovative two-stage methodology to improve the estimation of poverty risk across European territories by leveraging high-dimensional socio-economic data. The approach combines a clustering phase, which captures territorial heterogeneity, with Lasso regression for variable selection and model simplification, ensuring parsimony and interpretability. The application to different degrees of urbanisation reveals that the determinants of poverty vary spatially, with significant differences both in the composition of clusters and in the relevance of explanatory variables. The results provide new insights for the design of effective, territorially targeted anti-poverty policies and contribute to the European debate on regional inequalities through a comparative and statistically sound approach.

1. Introduction

Over the past decades, poverty risk has remained a core topic in socio-economic research, involving diverse disciplines such as economics, sociology, and statistics. A widely used metric is the at-risk-of-poverty rate, defined as the share of individuals with equivalised disposable income below 60% of the national median (Eurostat, 2024). This relative measure captures income distribution within countries and serves as a proxy for broader dimensions such as inequality and social exclusion.

Another measure widely used in the European policy context is the At Risk of Poverty or Social Exclusion (AROPE) indicator, which corresponds to the sum of persons who are either at risk of poverty, or severely materially and socially deprived or living in a household with a very low work intensity. It is one of the main indicators to monitor the EU 2030 target on poverty and social exclusion and was the headline indicator to monitor the EU 2020 Strategy poverty target. As one of the main indicators of the EU 2030 Agenda, AROPE provides insight into socioeconomic vulnerability and highlights the multidimensional nature of poverty and exclusion in Europe (Eurostat, 2021).

Poverty risk is shaped by a complex interplay of socio-economic determinants, including welfare policies (Chircop *et al.*, 2022), labour market conditions (Reluga *et al.*, 2025), and access to quality education (Eurostat, 2025) and healthcare services (Renahy, 2018). Empirical evidence shows that regions with higher educational attainment and stronger social infrastructure tend to report lower poverty risks (Copus *et al.*, 2015). Conversely, structural unemployment and limited public services increase vulnerability (Serban, 2021).

Poverty, inequality, and exclusion often overlap conceptually and empirically, this overlap makes measurement and policy formulation more complex. Moreover, shared drivers such as long-term unemployment affect these dimensions simultaneously, further blurring analytical boundaries.

A spatial perspective adds another layer of complexity. Urban, suburban, and rural areas differ significantly in terms of opportunities, costs of living, and access to services. While urban centres offer broader access to education and employment, they are often associated with higher housing costs and income disparities. Rural areas, by contrast, may face chronic limitations in infrastructure and labour market integration, leading to persistently higher poverty rates. Suburban contexts show intermediate characteristics but can also experience uneven development.

This study investigates how the relationship between socio-economic variables and poverty risk varies across degrees of urbanisation in Europe. By applying an innovative two-stage methodology—clustering followed by sparse regression—we aim to uncover latent spatial structures and identify the most relevant predictors in each context. The goal is to provide actionable insights for developing differentiated, evidence-based policies tailored to the specific challenges of cities, towns, and rural areas. The next section explores how poverty risk differs across degrees of urbanisation and identifies the multi-dimensional vulnerabilities specific to each context.

2. Poverty Risk and Urbanisation in Europe

Poverty risk in Europe displays marked territorial disparities, closely linked to the degree of urbanisation—typically classified into cities, towns and suburbs, and rural areas. According to Eurostat data, in 2024, about 93.3 million people in the EU were at risk of poverty or social exclusion. Of these, 5.6 million lived in households experiencing all three dimensions of severe disadvantage: monetary poverty, severe material and social deprivation, and very low work intensity.

Additional combinations also reveal multi-dimensional vulnerability: 11.4 million experienced both income poverty and low work intensity, while 8.4 million combined poverty with severe deprivation. These patterns are shaped by structural features specific to different territorial settings.

Urban centres often provide broader access to employment and education but face challenges such as high living costs and greater income inequality, which may offset these advantages for lower-income groups. Rural areas, by contrast, are often characterised by limited access to services, lower educational attainment, and fewer job opportunities—conditions that contribute to persistently high poverty risks. Towns and suburbs tend to fall between these two extremes, showing more heterogeneous profiles depending on national and regional contexts.

These spatial dynamics highlight the importance of incorporating territorial dimensions into anti-poverty strategies. Effective policies must be adapted to the unique constraints and opportunities of each area type, with a focus on balancing accessibility, affordability, and service provision across the urban–rural spectrum. In light of these spatial disparities, we apply a novel statistical framework to disentangle the role of socio-economic factors in explaining poverty risk across urban, suburban, and rural areas.

3. Methodology

Clusterwise Regression Model (CRM) (Spath, 1979; 1982) is an effective framework to model the relationship between a dependent variable and a set of explanatory variables in the presence of unobserved heterogeneity among observations. CRM assumes an underlying clustering structure within the data, which enables a more nuanced interpretation of regression results, especially in high-dimensional and heterogeneous socio-economic data; CRM has been widely applied in domains such as market segmentation, manufacturing, process modelling (Ganjigatti *et al.*, 2007), infrastructure management (Khadka *et al*., 2017), spatial analysis (Openshaw, 1977), meteorological forecasting (Bagirov *et al*., 2017), air quality prediction (Poggi *et al*., 2011), and machine learning (Karmitsa *et al*., 2020). Its flexibility in capturing data- driven clusters makes it a valuable tool for exploring complex, multi-faceted relationships.

In this study, we extend the CRM approach by integrating it with the Least Absolute Shrinkage and Selection Operator (Lasso) regression. The resulting method—Clusterwise Regression with Lasso—combines unsupervised clustering with regularised regression, allowing for both identification of latent subgroups and simultaneous variable selection within each cluster. Specifically, the dataset is partitioned into clusters based on an optimal number determined empirically, and a separate Lasso regression model is estimated for each subgroup. This combined approach handles heterogeneity and multicollinearity, while improving model clarity and prediction accuracy.

The general formulation of Clusterwise Lasso regression can be described as follows:

$$\min_{\beta} \sum_{k=1}^{K} \left[\sum_{l \in C_k} \left(y_l - \beta_{k0} - \sum_{j=1}^{J} \beta_{kj} x_{lj} \right)^2 + \lambda \sum_{j=1}^{J} \left| \beta_{kj} \right| \right] \quad (1)$$

where:

K is the number of clusters,

C_k represents the set of indices of data points in cluster k,

$\beta_{k0}$and β_{kj} are the intercept and coefficients for the k-th cluster,

λ is the regularization parameter controlling the sparsity of the solution.

Cross-validation is a widely used method for selecting the optimal regularization parameter λ in Lasso regression (Roberts, 2014).

This method proceeds by splitting the data into “folds”. Then over a sequence of tuning parameter values, penalized models are fitted to all but one of the folds and the predictive performance of each model is gauged over the “left-out” fold. This process is repeated until each fold has been left out. The value of the tuning parameter is then chosen to be the value in the sequence that has the smallest aggregated prediction error (or cross-validation error). By selecting the optimal lambda, cross-validation helps identify the most relevant predictors, simplifying the model while retaining essential information (Hastie, 2009).

The CRM algorithm is executed alternating between the following two steps until convergence to a stationary value:

Step1 – Representation phase (best fitting):

The local Lasso regression models are estimated by minimizing the objective function (Equation 1).

Step 2 – Assignment phase (partition P_k):

The optimal clusters that minimize the objective function are obtained according to the following assignment rule:

$$C_k = \left\{e_i \in E \colon |\widehat{e_{\iota(k)}}|^2 < |\widehat{e_{\iota(h)}}|^2\right\}, \mathit{for\,all}\, h \neq k$$

where:

- $\hat{e}_i$ is the *i*-th residual from the regression model.

Therefore, observation $\hat{e}_i$ is assigned to cluster C_k if the sum of squared errors is minimal for that cluster's regression model[1].

4. Evaluating Poverty Risk in Europe through Clusterwise Lasso Regression

The application of Clusterwise Lasso Regression allows for a nuanced assessment of poverty risk determinants across different levels of urbanisation in Europe. By combining clustering with penalised regression, the approach captures latent heterogeneity and enhances interpretability through context-specific variable selection.

[1] The statistical analysis was carried out by using the software R-studio and package used for Lasso regression is glmnet in R programming environment available on the CRAN (https://CRAN.R-project.org/package=glmnet).

The analysis is based on Eurostat data[2], drawing on harmonised indicators from the Sustainable Development Indicators (SDI) and the Degree of Urbanisation classification. A set of 20 socio-economic variables (see Table 1), covering education, employment, income, and housing, was selected based on theoretical relevance and data availability for 2023.

This integrated method identifies homogeneous territorial clusters and estimates separate regression models for each group. The resulting framework accommodates multicollinearity and structural differences between urban, suburban, and rural areas, offering insight into the most influential poverty risk predictors across spatial contexts.

Figure 1 illustrates the spatial distribution of poverty risk by urbanisation level across European countries, while Figure 2 reports the optimal number of clusters identified via the elbow method. Three clusters were selected in each context to balance model fit and parsimony. Figure 3 displays the clustering results obtained through Clusterwise Lasso Regression algorithm, illustrating the classification of European countries according to their respective degrees of urbanisation. The analysis was conducted separately for cities, towns and suburbs, and rural areas.

The spatial distribution of the at-risk-of-poverty rate across European countries shows that, in Western and Northern Europe, urban areas generally display lower poverty rates, reflecting higher levels of economic development and more comprehensive social protection systems. In contrast, urban centres in certain Southern and Eastern European countries exhibit comparatively higher poverty risks, likely due to structural economic disparities and varying levels of service accessibility.

Rural areas reveal the most pronounced geographical inequalities, with consistently elevated poverty rates observed in Central and Eastern Europe. These regions are often characterised by limited labour market opportunities, infrastructural deficits, and reduced access to public services. Towns and suburban areas show more heterogeneous patterns: while some align closely with national averages, others diverge significantly, particularly in Southern European countries where regional imbalances remain substantial.

Table 2 reports the estimated coefficients from Clusterwise Lasso Regression model applied to cities, with the dataset segmented into three clusters, as determined by the elbow method. Each cluster includes countries sharing similar socio-economic profiles that influence poverty risk in urban settings.

[2] https://ec.europa.eu/eurostat/web/database

Table 1 - *Socio-economic Indicators used in the analysis.*

Variable	Indicator	Definition
Y	At- Poverty Risk Percentage	Percentage of people at risk of poverty on the total population
X1	Low Education	People aged 18-24 who have completed at lower secondary education and have not attended further education or training courses
X2	Education Rate	Rate of participation in education or training - Age 18-64
X3	Low-Level Education Population	Population by education level: "Less than primary, primary and lower secondary education (Level 0-2) - Age 15-64"
X4	Medium-Level Education Population	Population by education level: "Upper secondary and post-secondary non-tertiary education (Level 3 and 4) - Age 15-64"
X5	High-Level Education Population	Population by education level: "Tertiary education (Level 5-8) - Age 15-64"
X6	Employment Rate	Employment rates - Age 15-64
X7	NEETs 15-24	Young people not in employment, education, or training - Age 15-24
X8	NEETs 25-34	Young people not in employment, education, or training - Age 25-34
X9	Youth Employment (Level 3-4)	Employment rate of young people not in education or training: "Upper secondary and post-secondary non-tertiary education (Level 3 and 4)"
X10	Youth Employment (Level 5-8)	Employment rate of young people not in education or training: "Tertiary education (Level 5-8)"
X11	Housing Cost Overburden	Housing cost overburden rate
X12	Median Income	Median income
X13	Severe Material Deprivation	Severe material deprivation rate
X14	High Income Share	Share of people with income at or above 130% of median income
X15	Under-occupied Housing	Percentage of people living in under-occupied dwellings
X16	Part-Time Employment (Level 0-2)	Part-time employment rate by job type: "Less than primary, primary and lower secondary education (Level 0-2)"
X17	Part-Time Employment (Level 3-4)	Part-time employment rate by job type: "Upper secondary and post-secondary non-tertiary education (Level 3 and 4)"
X18	Part-Time Employment (Level 5-8)	Part-time employment rate by job type: "Tertiary education (Level 5-8)"
X19	Self-Employed Percentage	Percentage of self-employed workers
X20	Unemployment Rate 15-64	Unemployment rate of people aged 15-64

In Cluster 1, participation in education and youth employment are positive variables for reducing the risk of poverty, while the presence of NEET young adults significantly increases the risk of poverty. In Cluster 2, low education and material deprivation are negative determinants, while a higher share of high incomes helps to reduce poverty. In Cluster 3, widespread secondary education has a positive effect in reducing poverty, but the risk remains high in the presence of material deprivation and disadvantaged housing conditions.

While urban areas show some commonalities in educational and employment-related determinants of poverty, suburban regions reveal more heterogeneous and context-dependent patterns.

Figure 1 *Maps of Poverty Risk Rate in Europe for different degree of urbanisation.*

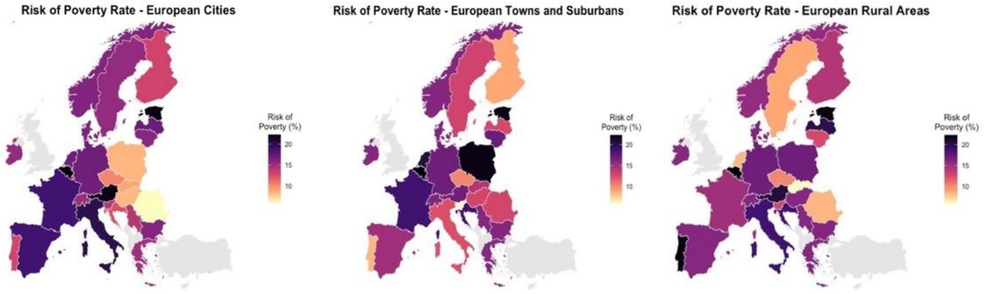
Risk of Poverty Rate - European Cities
Risk of Poverty Rate - European Towns and Suburbans
Risk of Poverty Rate - European Rural Areas
Risk of Poverty (%)
20
15
10

Figure 2 - *Elbow plot.*

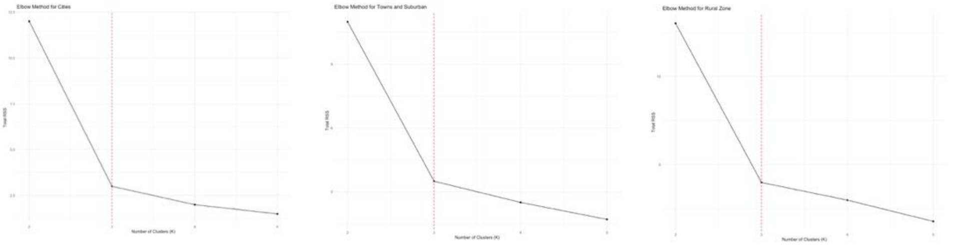

Figure 3 - *Classification of European Regions for different urbanisation degree*.

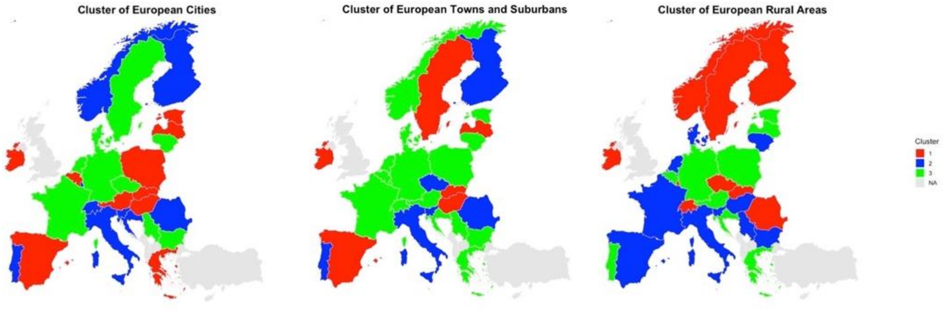
Cluster of European Cities
Cluster of European Towns and Suburbans
Cluster of European Rural Areas
Cluster
1
2
3
NA

Table 2 – *Coefficients for European Cities.*

Variable	Cluster 1	Cluster 2	Cluster 3
Low Education	0.11	0.50	0
Education Rate	-0.23	0	-0.16
Low-Level Education Population	0	-0.11	-0.31
Medium-Level Education Population	0	0	-0.79
High-Level Education Population	0	0	0
Employment Rate	0	0	0
NEETs 15-24	0	0	0
NEETs 25-34	0.44	0	0.01
Youth Employment (Level 3-4)	-0.67	0	-0.05
Youth Employment (Level 5-8)	0	-0.06	0
Housing Cost Overburden	0	0	0
Median Income	0	0	0
Severe Material Deprivation	0	0.28	0.29
High Income Share	0	-0.05	0
Under-occupied Housing	0	0	0.05
Part-Time Employment (Level 0-2)	0	0	0
Part-Time Employment (Level 3-4)	0	0	0
Part-Time Employment (Level 5-8)	-0.39	-0.05	0
Self-Employed Percentage	-0.47	-0.19	0
Unemployment Rate 15-64	0.08	0	0

Table 3 shows the coefficients from the Clusterwise Lasso model for European towns and suburbs, grouped into three clusters by socio-economic traits.

In Cluster 1, the risk of poverty increases in the presence of young NEETs, high housing costs, and under-occupation of the home, while youth employment acts as a positive factor in reducing poverty. In Cluster 2, vulnerability is associated with low education, NEETs, and unemployment, while higher levels of education and higher employment contribute to reducing the risk. In Cluster 3, there is a high level of vulnerability linked to NEETs, housing costs, and labor market fragility, which is only partially offset by positive factors such as tertiary education, self-employment, and a higher proportion of high incomes.

Compared to urban and suburban areas, rural regions tend to exhibit more structural constraints, including lower service availability and more persistent unemployment. Table 4 reports coefficients from the Clusterwise Lasso model for rural areas, divided into three clusters.

Table 3 – *Coefficients for European Towns and Suburbs.*

Variable	Cluster 1	Cluster 2	Cluster 3
Low Education	0	0.22	0
Education Rate	-0.18	0	0
Low-Level Education Population	0	0	0
Medium-Level Education Population	0	0	0
High-Level Education Population	0	-0.28	-0.10
Employment Rate	0	-0.10	-0.07
NEETs 15-24	0.21	0	0.35
NEETs 25-34	0	0.15	0.28
Youth Employment (Level 3-4)	0	0	0
Youth Employment (Level 5-8)	-0.06	0	-0.15
Housing Cost Overburden	0.23	0	0.35
Median Income	0	0	0
Severe Material Deprivation	0	0	0
High Income Share	0	0	-0.09
Under-occupied Housing	0.20	0	0
Part-Time Employment (Level 0-2)	0	0	0
Part-Time Employment (Level 3-4)	0	0	0
Part-Time Employment (Level 5-8)	0	0	0
Self-Employed Percentage	0	-0.15	-0.22
Unemployment Rate 15-64	0	0.22	0

In Cluster 1, the presence of young NEETs and under-occupation of housing increase the risk of poverty, while higher average incomes, education, and youth employment play a positive role in reducing poverty. In Cluster 2, the risk increases in the presence of low education and adult NEETs, but is mitigated by youth employment, higher education, and a higher share of high incomes. In Cluster 3, the main negative determinants are material deprivation, housing costs, and unemployment, while education and overall employment remain positive variables.

In some clusters, the coefficient associated with the variable “Under-occupied Housing” shows a positive sign, which may appear counterintuitive. Rather than indicating improved living standards, this association can reflect structural inefficiencies in the housing stock or demographic imbalances.

As highlighted by Eurofound, in several post-communist and Southern European countries many households occupy relatively large dwellings that are energy-inefficient and costly to maintain, which can generate financial strain despite the apparent availability of space (Eurofound, 2023). Other studies further suggest that housing deprivation contributes to poverty even among households that appear well-off in terms of space but lack adequate quality or affordability (Hick *et al.,* 2022). These findings indicate that the positive sign of under-occupation in our results should be interpreted as a signal of latent housing-related vulnerabilities.

Nevertheless, we acknowledge that European indicators may not fully reflect local housing conditions.

Table 4 – *Coefficients for European Rural Areas.*

Variable	Cluster 1	Cluster 2	Cluster 3
Low Education	0	0.20	0
Education Rate	-0.19	0	-0.33
Low-Level Education Population	-0.07	0	0
Medium-Level Education Population	0	0	-0.07
High-Level Education Population	0	-0.04	0
Employment Rate	0	0	-0.14
NEETs 15-24	0.19	0	0
NEETs 25-34	0	0.20	0
Youth Employment (Level 3-4)	0	0	0
Youth Employment (Level 5-8)	-0.11	-0.44	0
Housing Cost Overburden	0.04	0	0.09
Median Income	-0.37	0	0
Severe Material Deprivation	0	0	0.32
High Income Share	0	-0.22	0
Under-occupied Housing	0.65	0	0
Part-Time Employment (Level 0-2)	0	0	0
Part-Time Employment (Level 3-4)	0	-0.18	0
Part-Time Employment (Level 5-8)	0	0	0
Self-Employed Percentage	0	0	0
Unemployment Rate 15-64	0.18	0	0.30

5. Limitations and Future Research

While the proposed methodology captures spatial heterogeneity and improves model interpretability, it has some limitations. The cross-sectional nature for 2023 data limits temporal inference; future work could explore dynamic clustering or longitudinal extensions. In addition, supplementing the statistical analysis with case studies or qualitative data may offer additional insights.

Another limitation is the potential endogeneity of some explanatory variables, which may generate reverse causality; addressing this issue would require causal inference or instrumental variable approaches in future research.

6. Conclusion

This work demonstrates the effectiveness of Clusterwise Lasso Regression in analyzing spatial inequalities in poverty risk across urban, suburban, and rural areas in Europe. The results reveal significant spatial heterogeneity: different area types show distinct socio-economic profiles and sensitivities to key variables such as education, youth employment, NEET rates, and housing conditions. By combining clustering and regularized regression, the method captures latent structures and enables the identification of context-specific predictors, highlighting the limits of uniform modelling approaches to poverty analysis.

The findings support the design of differentiated, evidence-based policies aligned with the principles of the European Pillar of Social Rights3 and the EU's cohesion policy framework4. For instance, urban areas may benefit from actions supported by the European Social Fund Plus[5] (ESF+) aimed at improving access to education, promoting youth employment, and enhancing social inclusion. In contrast, rural and peripheral regions—where infrastructure gaps and material deprivation are more pronounced—could be targeted through place-based interventions financed by instruments like the Just Transition Fund[6] or the European Regional Development Fund (ERDF), including investments in mobility, broadband access, and support for sustainable local economies. The proposed analytical framework can support policymakers in designing tailored interventions. These should account for the structural characteristics of each territory.

Future developments may include the adoption of non-linear modelling approaches, the temporal extension of the analysis to capture trends and dynamics, and the integration of spatial methods to account for geographical interdependencies and regional spillover effects.

References

BAGIROV A. M., MAHMOOD A., BARTON A. 2017. Prediction of monthly rainfall in Victoria, Australia: Clusterwise linear regression approach. *Atmospheric research*, Vol. 188, pp. 20-29.

COPUS A., MELO P. C., KAUP S., TAGAI G., ARTELARIS P. 2015. Regional poverty mapping in Europe–Challenges, advances, benefits and limitations. *Local Economy*, Vol.30, No. 7, pp. 742-764.

CHIRCOP D., MÜLLER K., NAVARRA C., PASIKOWSKA-SCHNASS M. 2022. *EU welfare systems and the challenges of poverty and inequality.*

EUROFOUND. 2023. *Unaffordable and inadequate housing in Europe*, Publications Office of the European Union, Luxembourg.

EUROSTAT. 2021. Glossary: At risk of poverty or social exclusion (AROPE).

EUROSTAT. 2024. Latest developments in income dynamics and poverty.

EUROSTAT. 2025. Living conditions in Europe - poverty and social exclusion.

[3] https://ec.europa.eu/regional_policy/funding/erdf_en

[4] https://www.eib.org/en/projects/topics/regionaldevelopment/index#:~:text=The%20European%20Union's%20Chesion%20Policy,and%20finance%20cohesion%20throughout%20Europe.

[5] https://commission.europa.eu/funding-tenders/find-funding/eu-funding-programmes/european-social-funden

[6] https://commission.europa.eu/funding-tenders/find-funding/eu-funding-programmes/just-transition-fund_en

GANJIGATTI J. P., PRATIHAR D. K.,2007. Global versus cluster-wise regression analyses *Journal of materials processing technology*, Vol. 189, No.1-3, pp. 352-366.

HASTIE T. 2009. *The elements of statistical learning: data mining, inference, and prediction.*

HICK R., POMATI M., STEPHENS M. 2022. Severe housing deprivation in the

European Union: A joint analysis of measurement and theory. *Social Indicators Research*, Vol. 164, No. 3, pp. 1271-1295.

KARMITSA N., TAHERI S., BAGIROV A., MAKINEN P. 2020. Missing value imputation via clusterwise linear regression. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 34, No. 4, pp. 1889-1901.

KHADKA M., PAZ A. 2017. Comprehensive clusterwise linear regression for pavement management systems. *Journal of Transportation Engineering, Part B: Pavements*, Vol. 143, No. 4.

OPENSHAW S. 1977. A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modelling. *Transactions of the institute of british geographers*, pp. 459-472.

POGGI J. M., PORTIER B. 2011. PM10 forecasting using clusterwise regression. *Atmospheric Environment*, Vol. 45, No. 38, pp. 7005-7014.

RELUGA K., KONG D., RANJBAR S., SALVATI N., VAN DER LAAN M.

2025. The impact of job stability on monetary poverty in Italy: causal small
area estimation.

RENAHY E., MITCHELL C., MOLNAR A., MUNTANER C., NG E., ALI F., O’CAMPO P. 2018. Connections between unemployment insurance, poverty and health: a systematic review. *The European Journal of Public Health*.

ROBERTS S., NOWAK G. 2014. Stabilizing the lasso against cross-validation variability. *Computational Statistics & Data Analysis*, Vol. 70, pp.198-211.

SERBAN A. M., BRAZIENĖ R. 2021. *Young People in Rural Areas: Diverse, Ignored and Unfulfilled.*

SPATH H. 1979. Algorithm 39 Clusterwise linear regression. *Computing.* Vol. 22.

SPATH H. 1982. A fast algorithm for clusterwise linear regression. *Computing.* Vol. 29.

Simona CAFIERI, Istat, cafieri@istat.it
Gianmarco BORRATA, Università Federico II, gianmarco.borrata@unina.it