Università degli Studi della Campania “Luigi Vanvitelli”

Internship Host: ISTAT (Istituto Nazionale di Statistica)

Master’s in Data Science

Where You Live, How You Live: Spatial Patterns of Health Inequality Across EU Regions

Internship Report

Student: Mahdi Mohammadzadeh
Supervisor: Simona Cafieri, Prof. Rosanna Verde
Host: ISTAT — Istituto Nazionale di Statistica

A.Y. 2024–2025

Contents

List of Figures

List of Tables

Abstract

This report documents an internship project carried out with ISTAT, focused on building a reproducible pipeline that makes Eurostat regional data easier to use for territorial analysis. Using Eurostat indicators at NUTS-2 level, we harmonise data on *self-reported unmet medical need* (2021–2023) and three socio-economic correlates: GDP per inhabitant (PPS, EU=100), unemployment rate, and at-risk-of-poverty rate. The resulting master dataset covers 135 regions and 379 region–year observations and is accompanied by a transparent raw-to-processed workflow and report-ready tables and figures.

Descriptive evidence indicates that unmet medical need is spatially concentrated in a limited set of high-burden territories and differs systematically across broad welfare-regime groupings. Baseline OLS models (2023) and a short-panel fixed-effects robustness check (2021–2023) show a stable sign pattern: unemployment is the most consistent positive correlate of unmet need, while GDP per inhabitant is consistently negative. An optional 2021 extension includes a mortality indicator (available only for 2021) to provide additional health context. Results are interpreted as descriptive associations rather than causal effects.

1 Context & Motivation

1.1 Why unmet medical need matters

Reducing avoidable health inequalities is a central objective of European social and public health policy. A particularly policy-relevant dimension is *access*: even within the same country, the ability to obtain timely medical care can differ substantially across territories due to differences in income, labour-market conditions, local service capacity, and institutional settings.

This report uses the Eurostat indicator *self-reported unmet medical need* as a practical proxy for access barriers. The measure captures the share of people who report needing a medical examination or treatment but not receiving it, typically because of *cost*, *distance/transport*, or *waiting lists*. While self-reported indicators have limitations, they provide timely information on perceived constraints and are commonly used in monitoring access and equity.

1.2 Internship objectives and deliverables

The internship objective was to build a clean and reproducible regional dataset and a transparent analysis workflow that can be extended in future thesis work. Concretely, the deliverables are:

- a raw-to-processed data pipeline that downloads and cleans Eurostat SDMX-CSV extracts into tidy regional tables;
- a harmonised master dataset for 2021–2023 with consistent region identifiers and documented transformations;
- exploratory analysis, baseline econometric models, and complementary predictive models to identify key correlates;
- report-ready figures and regression/ML tables to support communication and iteration.

1.3 Aim and scope

The empirical focus is on NUTS-2 regions in the European Union over 2021–2023 (post-pandemic recovery years, subject to data availability in the extracted files). The main aim is descriptive: to quantify regional disparities in unmet medical need and to assess how unmet need co-varies with socio-economic conditions, notably GDP per inhabitant, unemployment, and poverty risk.

Importantly, the models in this report are not intended to provide causal estimates. Instead, they provide robust *associational* evidence and a compact, reproducible framework for identifying high-burden areas and key correlates that may inform more in-depth work.

1.4 Research questions

1. How is unmet medical need distributed across EU NUTS-2 regions in 2021–2023?
2. How are GDP, unemployment, and poverty associated with unmet medical need in a recent cross-section (2023)?
3. Do patterns remain similar when using within-region variation in a short panel (2021–2023)?
4. Do simple contextual groupings (e.g., welfare regime; high-burden regions) show systematic differences in unmet need?
5. As an extension, does adding a mortality indicator (available in 2021) materially change cross-sectional associations?

1.5 What this report contributes

Beyond the substantive findings, the main contribution is methodological: a reproducible pipeline that harmonises heterogeneous Eurostat regional extracts into a single analysis-ready panel. The workflow includes explicit checks for one-to-many merge risks, documented coverage by region and year, and diagnostic steps such as multicollinearity assessment and sensitivity checks. These choices reflect good practice in official-statistics environments where transparency and auditability are essential.

2 Data Sources

2.1 Data sources

The core dataset is built from official Eurostat regional indicators, accessed through the Eurostat Data Browser and exported in SDMX-CSV format [1]. Raw extracts are stored under `data/raw/` and cleaned versions under `data/processed/`. The analysis is conducted at NUTS-2 level to balance territorial detail with cross-country comparability.

2.2 Core variables and definitions

Four core series are used throughout the 2021–2023 panel:

- **Unmet medical need (%)**: `hlth_silc_08b_r`. Share of people reporting unmet need for a medical examination or treatment (regional breakdown).
- **GDP per inhabitant (PPS, EU=100)**: `tgs00006`. Regional GDP per inhabitant in purchasing power standards, indexed to EU average.
- **Unemployment rate (%)**: `lfst_r_lfu3rt`. Regional unemployment rate for the total population (working age).
- **At-risk-of-poverty rate (%)**: `ilc_li41`. Share of people at risk of poverty (regional breakdown).

All percentage variables are expressed in percentage points.

2.3 Optional extension: mortality

To provide an additional health-outcome perspective, we extracted a mortality indicator:

- **Standardised death rate (all causes, per 100,000)**: extracted at NUTS-2 level. In the available extracts, this series is observed only for 2021. For this reason, mortality is used as an *optional* extension on the 2021 cross-section and is not included in the main 2021–2023 panel specifications.

2.4 Contextual lookup

A separate lookup table (`region_lookup.csv`) maps region codes to human-readable labels and contextual categories used for grouping and interpretation. In particular, the report uses a `welfare_model` classification (e.g., Continental, Mediterranean, Nordic)

and a region label to highlight high-burden territories in descriptive figures.

2.5 Final analysis table and coverage

Each indicator is cleaned into a tidy long format and harmonised on `(geo, time)`, where `geo` is a NUTS-2 code and `time` is the calendar year. To avoid one-to-many merge issues, each source is collapsed to a single value per `(geo, time)` before merging.

The resulting master dataset contains

$$\{\texttt{unmet\_need\_pct}, \texttt{gdp\_pps\_euavg}, \texttt{unemp\_rate}, \texttt{poverty\_rate}\}$$

for 2021–2023, covering 135 NUTS-2 regions and 379 region–year observations. Mortality is merged with a left join and is available only for 2021, resulting in missing values for 2022–2023 (270 missing mortality values in the final master table). The report therefore presents mortality-inclusive models on the 2021 cross-section and panel models without mortality.

3 Data Preparation & Methods

3.1 Workflow overview

The workflow follows a reproducible *raw* → *processed* → *analysis* structure:

1. download/collect SDMX-CSV extracts and store under `data/raw/`;
2. clean each source into a tidy table under `data/processed/`;
3. harmonise identifiers and merge into a master dataset;
4. produce exploratory figures and estimation tables for reporting.

All steps are implemented in Python notebooks/scripts with saved intermediate outputs, enabling re-runs and auditability.

3.2 Cleaning and harmonisation

Eurostat SDMX exports can differ in column naming and may include multiple dimensions (e.g., sex, age group, unit). Cleaning therefore applies a consistent set of operations:

- **Column harmonisation**: map alternative SDMX column names to standard names (`geo`, `time`, value column).
- **Type coercion**: ensure `time` is numeric and values are floats; drop non-numeric artefacts.
- **NUTS-2 filtering**: restrict to NUTS-2 regions using either an explicit `nuts_level_guess` field (when present) or the standard NUTS-2 code length heuristic (4-character codes).
- **One value per region–year**: collapse each dataset to a single observation per `(geo, time)` (mean aggregation when necessary) to prevent merge explosions.

3.3 Merging strategy and quality checks

The master dataset is built by merging the four core indicators using inner joins on `(geo, time)` after restricting to the common period across core series (2021–2023). This yields a complete-case panel for the main analysis.

Mortality is merged afterwards using a left join because it is only available for 2021 in the extracted files. We explicitly report the number of missing mortality values after the merge so that the analyst can decide whether to (i) run a 2021-only specification including mortality, or (ii) retain the 2021–2023 panel without mortality.

As basic validation, the pipeline reports:

- row counts and unique region counts before/after merging;
- duplicates in `(geo, time)` (should be zero);
- missingness by column in the final master dataset.

3.4 Econometric specifications

Two complementary econometric approaches are used.

3.4.1 Cross-sectional OLS

For a given year (e.g., 2023), we estimate:

$$\texttt{unmet\_need\_pct}_i = \alpha + \beta_1\,\texttt{gdp\_pps\_euavg}_i + \beta_2\,\texttt{unemp\_rate}_i + \beta_3\,\texttt{poverty\_rate}_i + \varepsilon_i,$$

with heteroskedasticity-robust standard errors (HC family) [3, 2]. A 2021 extension adds mortality as an additional covariate where available.

3.4.2 Short-panel fixed effects

To exploit within-region changes over 2021–2023, we estimate a region fixed-effects model:

$$\texttt{unmet\_need\_pct}_{it} = \alpha_i + \gamma_t + \beta_1\,\texttt{gdp\_pps\_euavg}_{it} + \beta_2\,\texttt{unemp\_rate}_{it} + \beta_3\,\texttt{poverty\_rate}_{it} + u_{it},$$

where α_i captures time-invariant regional heterogeneity and γ_t captures common year shocks. Given the short horizon, this model is interpreted cautiously and primarily as a robustness check against purely cross-sectional comparisons.

3.5 Predictive models as complements

To complement interpretability-focused regressions, we also fit predictive models:

- **Random Forest**: flexible non-linear baseline; feature importances are used for descriptive ranking.
- **Lasso**: sparse linear model with standardised predictors, useful for directionally interpretable feature selection.

Validation uses cross-validation strategies designed to reduce optimistic bias, including group-based splits (by region) and leave-one-year-out checks for the short panel. ML results are presented as descriptive complements and do not carry a causal interpretation.

3.6 Reproducibility and reporting

Outputs are saved as CSV tables and PNG figures and included in this report via LaTeX inputs. The repository structure separates raw data, processed data, code, tables, and images, making the workflow easier to audit and extend. The report explicitly documents the final sample size, year coverage, and any variables with limited availability (notably mortality).

4 Exploratory Findings

4.1 Exploratory analysis strategy

Exploratory analysis serves two purposes: (i) to describe the spatial distribution of unmet medical need and identify high-burden territories, and (ii) to check whether simple linear associations align with later modelling results. Because the dataset is cross-country and regional, descriptive plots are preferred over country-specific narratives to maintain comparability.

4.2 Correlation structure

Figure 4.1 summarises pairwise linear correlations among the core variables in the pooled 2021–2023 sample. As expected, socio-economic indicators are correlated: GDP tends to be negatively related to unemployment and poverty, while unemployment and poverty are positively related. This motivates explicit multicollinearity checks in the regression section.

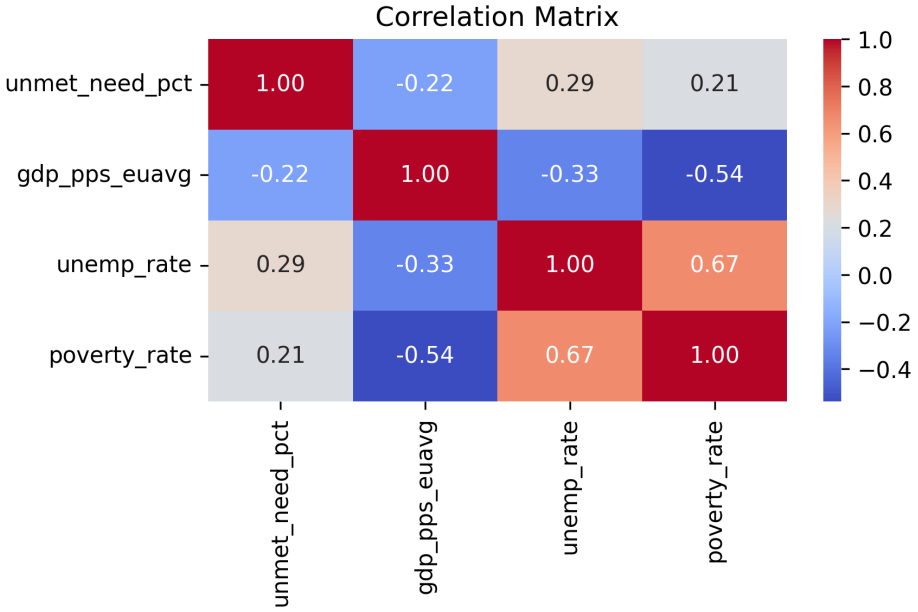

Figure 4.1: Correlation heatmap of the core indicators (pooled 2021–2023).

4.3 Distribution of unmet need

Figure 4.2 shows the distribution of unmet medical need across regions and years. The distribution is right-skewed: most regions cluster at relatively low unmet need, while a smaller set exhibits substantially higher levels. This pattern is consistent with the presence of geographically concentrated access barriers.

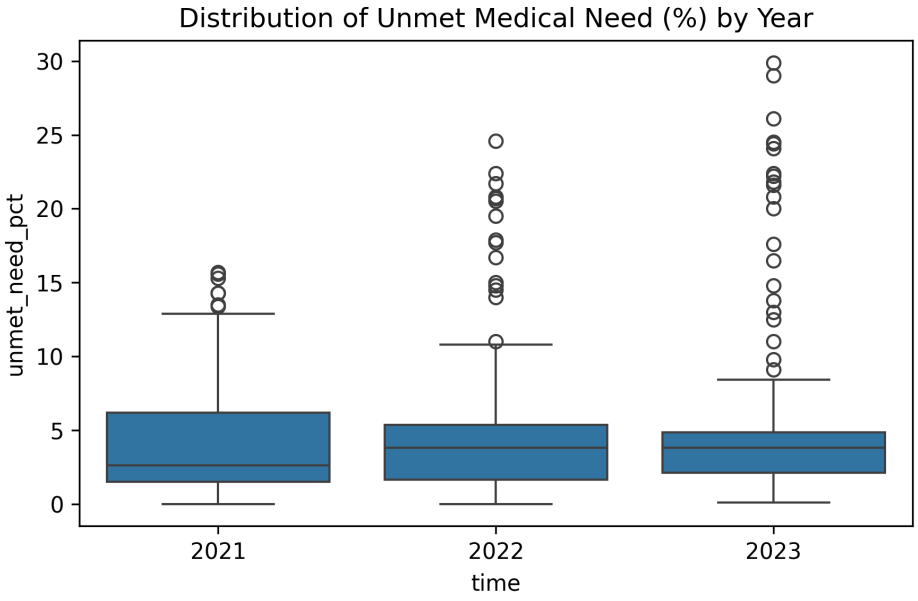

Figure 4.2: Distribution of unmet medical need by year (NUTS-2).

4.4 Bivariate relationships

Figure 4.3 visualises bivariate scatter relationships between unmet need and the three socio-economic indicators. The plots suggest a negative association with GDP and positive associations with unemployment and poverty, though relationships are noisy and potentially non-linear. These patterns motivate the combination of OLS (for transparent associations) and flexible ML models (to check robustness of ranking and importance).

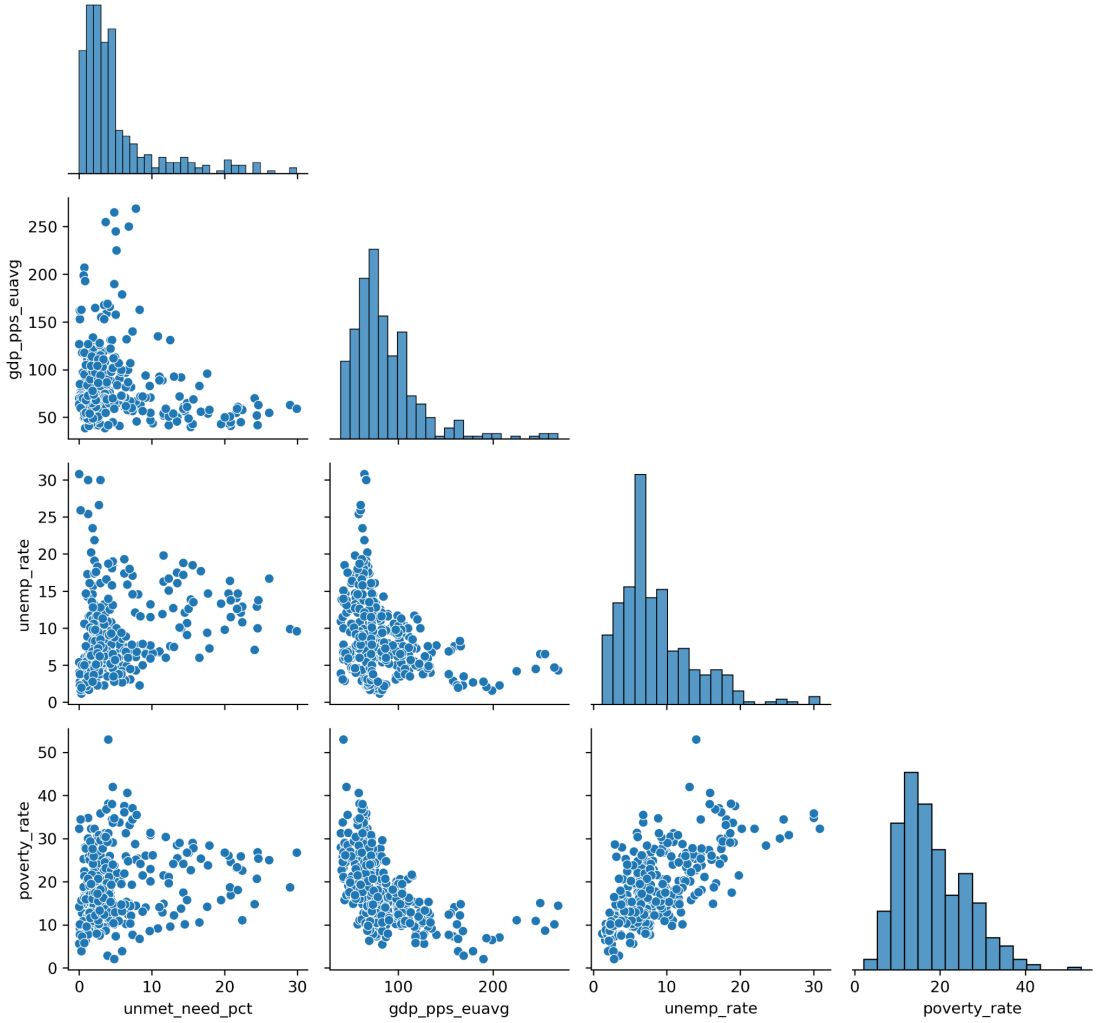

Figure 4.3: Bivariate relationships between unmet need and socio-economic indicators (pooled 2021–2023).

5 Results

5.1 Reading guide

This chapter presents results in three layers. First, we summarise descriptive group differences (welfare regimes and high-burden regions). Second, we report baseline econometric associations, starting with a recent cross-section (2023) and then a short-panel fixed-effects check (2021–2023). Third, we provide complementary predictive results from ML models to assess whether the same predictors emerge as important under flexible specifications.

5.2 Cross-sectional association in 2023 (OLS with robust SE)

Table 5.1 reports a baseline OLS regression for 2023, relating unmet medical need to GDP, unemployment, and poverty risk. Robust standard errors are used to reduce sensitivity to heteroskedasticity.

Table 5.1: OLS (2023) with heteroskedasticity-robust (HC1) standard errors. Significance: *** p<0.001, ** p<0.01, * p<0.05, † p<0.1.

Variable	Coef.	Std. Err.	p-value
const	7.801**	2.665	0.003
gdp_pps_euavg	-0.041*	0.018	0.026
unemp_rate	0.443*	0.220	0.044
poverty_rate	-0.117	0.110	0.284

Across specifications, the sign pattern is stable: higher unemployment is associated with higher unmet need, while higher GDP is associated with lower unmet need. Poverty risk tends to move in the expected direction but is less precisely estimated once unemployment is included, consistent with shared information between labour-market and poverty indicators.

5.3 Panel robustness check (2021–2023 fixed effects)

Because levels differ strongly across countries and regions, we also estimate a short-panel fixed-effects model to focus on within-region changes over 2021–2023. The goal is not to make strong dynamic claims (the window is short), but to check whether the

cross-sectional associations are broadly consistent when time-invariant regional factors are absorbed.

Table 5.2: Two-way fixed effects panel model (2021–2023) with region and year fixed effects, clustered standard errors by region. Predictors are standardised (z-scores). Significance: *** p<0.001, ** p<0.01, * p<0.05, † p<0.1.

Variable	Coef.	Std. Err.	p-value
Constant	1.997	2.339	0.393
Year 2022 (vs 2021)	0.144	0.391	0.712
Year 2023 (vs 2021)	0.669	0.452	0.139
GDP (z-score)	-0.593	1.811	0.743
Unemployment (z-score)	-3.099**	1.128	0.006
Poverty (z-score)	0.503	0.739	0.496

The fixed-effects estimates are generally weaker in magnitude, as expected in a short panel with limited within-region variation, but the direction of association for unemployment and GDP remains informative. Results should be interpreted as descriptive robustness checks rather than causal estimates.

5.4 Mortality extension on the 2021 cross-section

Mortality is only available for 2021 in the extracted files, so it is analysed as an extension on the 2021 cross-section. We first re-estimate the baseline 2021 model (Table 5.3) and then add mortality (Table 5.4).

Table 5.3: OLS (2021) baseline model estimated on the subset of regions with available mortality data (robust HC1).

Variable	Coef.	Std. Err.	p-value
const	3.564*	1.770	0.044
gdp_pps_euavg	-0.014	0.011	0.214
unemp_rate	0.223†	0.125	0.074
poverty_rate	-0.004	0.074	0.956

Table 5.4: OLS (2021) extended model including mortality rate (robust HC1).

Variable	Coef.	Std. Err.	p-value
const	5.540†	3.071	0.071
gdp_pps_euavg	-0.019	0.014	0.169
unemp_rate	0.160	0.161	0.322
poverty_rate	0.017	0.082	0.839
mortality_rate	-0.001	0.001	0.356

Including mortality does not qualitatively overturn the core socio-economic associations, but it provides an additional health context. Given the limited time coverage, these

results are best seen as an enrichment rather than a central pillar of the 2021–2023 panel analysis.

5.4.1 Multicollinearity diagnostics

Socio-economic predictors are correlated, so we compute Variance Inflation Factors (VIFs) for the 2021 specification (Table 5.5). The values help interpret coefficient instability and guide careful reporting (e.g., focusing on sign and robustness rather than over-interpreting small changes in magnitude).

Table 5.5: Variance Inflation Factors (VIF) for the 2021 model including mortality. Values below 5 indicate low multicollinearity.

variable	VIF
const	58.260000
gdp_pps_euavg	1.740000
unemp_rate	2.860000
poverty_rate	2.500000
mortality_rate	1.880000

5.5 Contextual patterns: welfare regimes and high-burden regions

Beyond regression coefficients, descriptive group comparisons help communicate the magnitude of disparities.

5.5.1 Welfare regime differences

Figure 5.1 compares average unmet need across welfare regimes (pooled 2021–2023), and Figure 5.2 shows how these averages evolve over time. The plots highlight systematic differences across regimes, suggesting that institutional context may shape perceived access barriers.

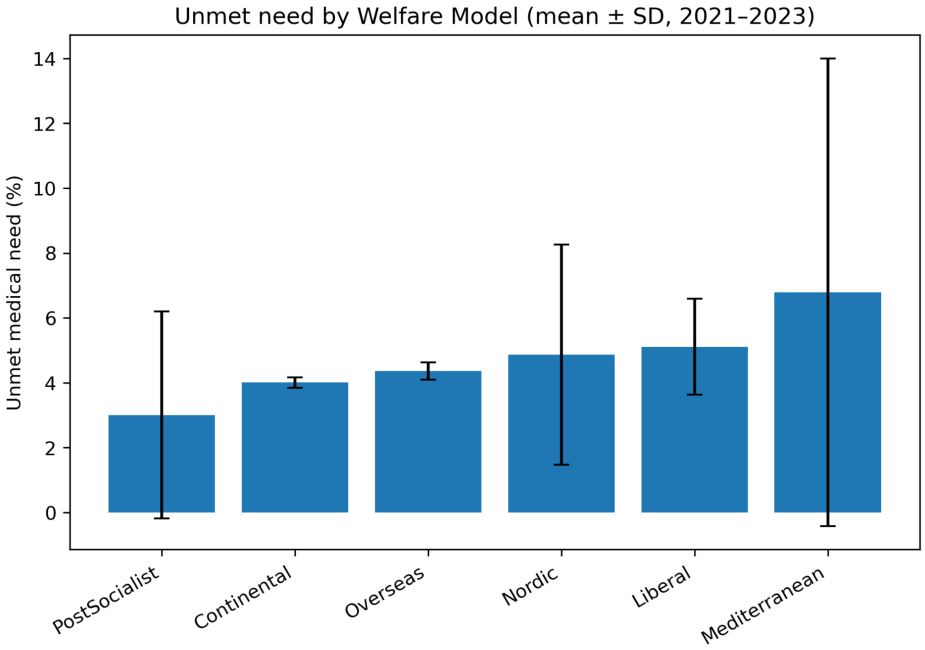

Figure 5.1: Average unmet medical need by welfare regime (mean ± SD), pooled 2021–2023.

Note. The first plot summarises importances in the multi-view setting; the next figure focuses on the lag-feature specification in the short panel. To keep each figure readable, we place them on separate pages.

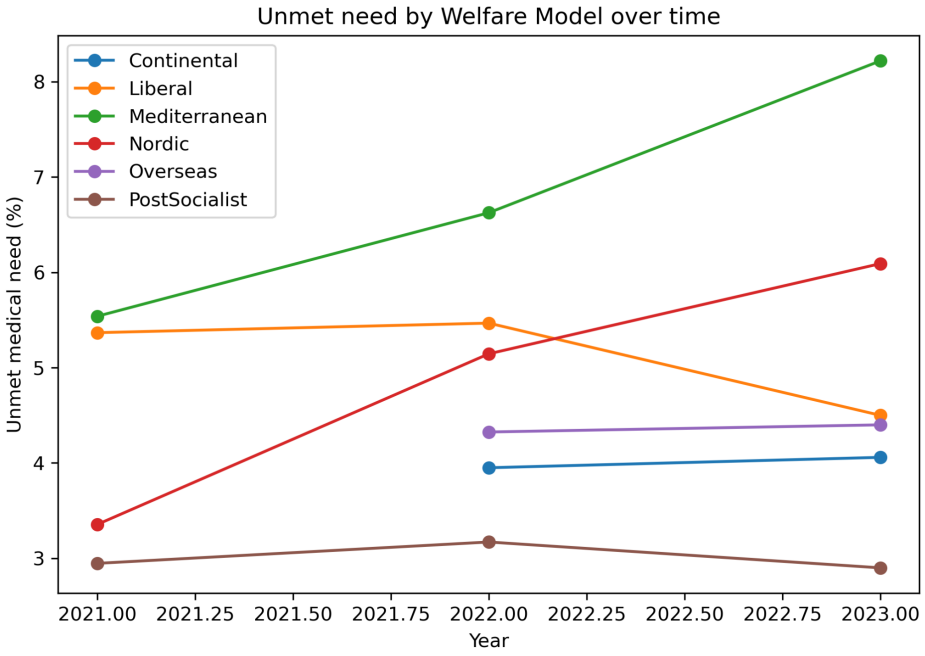

Figure 5.2: Unmet medical need over time by welfare regime (mean by year).

5.5.2 High-burden regions

Figures 5.3 and 5.4 focus on the top 10 high-unmet-need regions (pooled and over time). The purpose is to show whether high burden is broad-based or concentrated in a small number of territories.

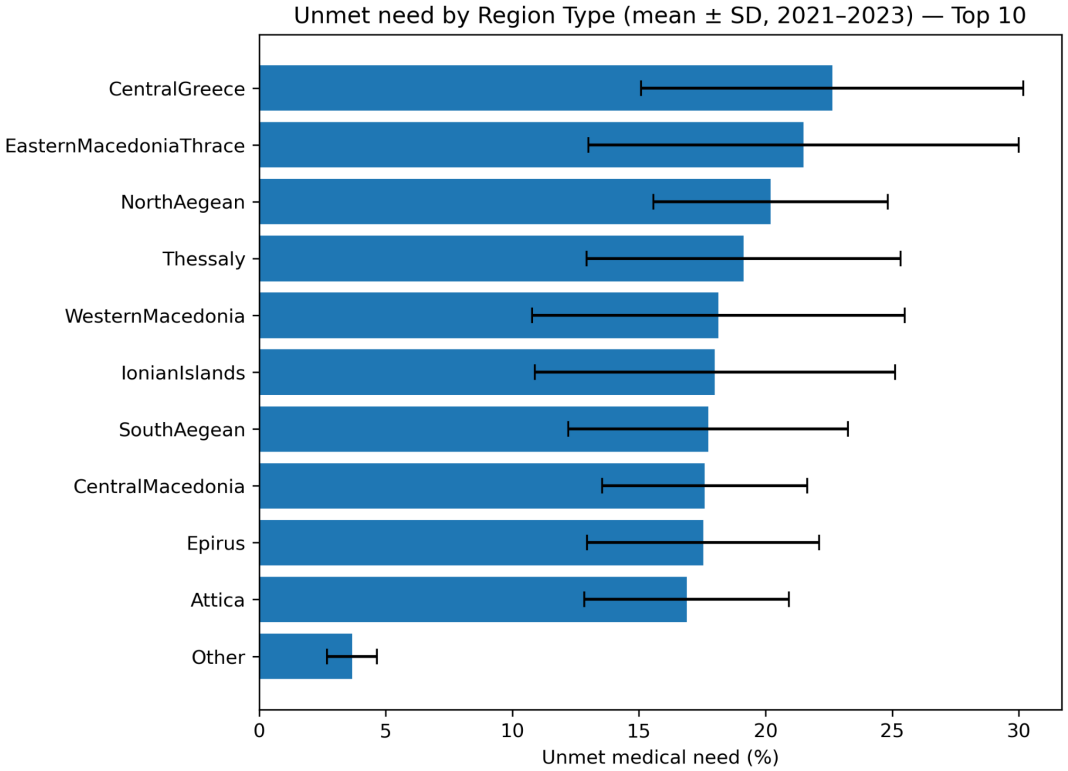

Figure 5.3: Top 10 regions by unmet medical need, pooled 2021–2023. Remaining regions are grouped as "Other".

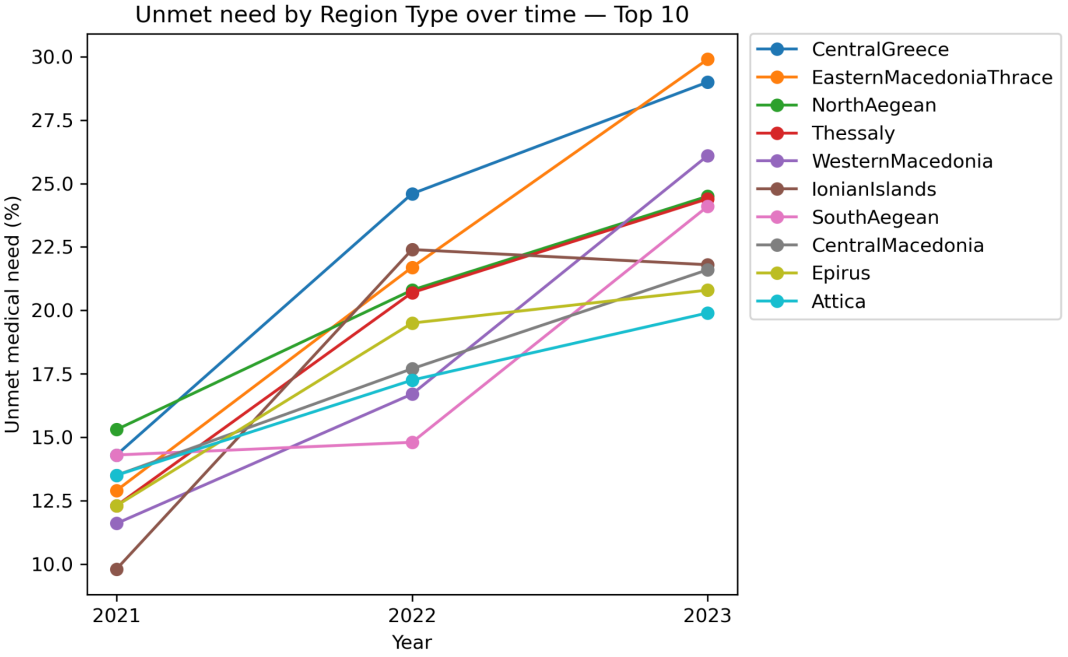

Figure 5.4: Trajectories of unmet medical need for the top 10 regions (2021–2023).

5.6 Complementary predictive models

Predictive models are used to check whether the same covariates emerge as important when we allow for non-linearities and interactions. We emphasise that these are *predictive associations*, not causal effects.

5.6.1 Multi-view cross-validation

Table 5.6 summarises multi-view models combining numeric predictors with contextual groupings. These results are mainly used to validate whether adding context improves predictive performance in a way consistent with the descriptive patterns.

Table 5.6: Grouped cross-validation performance for the multi-view random-forest model (numeric + contextual predictors). Negative R^2 indicates worse-than-mean prediction on some folds.

CV folds	R2_mean	R2_std	MAE_mean	MAE_std
5	-1.830	1.733	4.780	2.648

5.6.2 Group-aware validation for 2023

To reduce leakage from region-specific patterns, we also evaluate models with group-aware splits in 2023 (Table 5.7). This helps assess generalisation to unseen regions.

Table 5.7: Predictive performance on 2023 with country-wise grouped cross-validation.

model	R2_mean	R2_std	MAE_mean	MAE_std
Lasso	-51.544	64.945	5.558	3.030
Ridge	-52.634	66.480	5.599	3.043
RandomForest	-70.452	125.821	4.616	2.740

5.6.3 Leave-one-year-out checks in the short panel

Finally, we evaluate whether models trained on two years generalise to the third (Table 5.8). Given the short horizon, this is a demanding but informative check.

5.6.4 Feature importance summaries

Figures 5.5 and 5.6 report feature importances from random forest specifications. The rankings generally confirm that unemployment and GDP are among the strongest predictors, consistent with the OLS results.

Table 5.8: Panel predictive performance using leave-year-out evaluation (train on one year, test on another).

test_year	R2	MAE
2022	0.583	2.842
2023	0.470	2.726

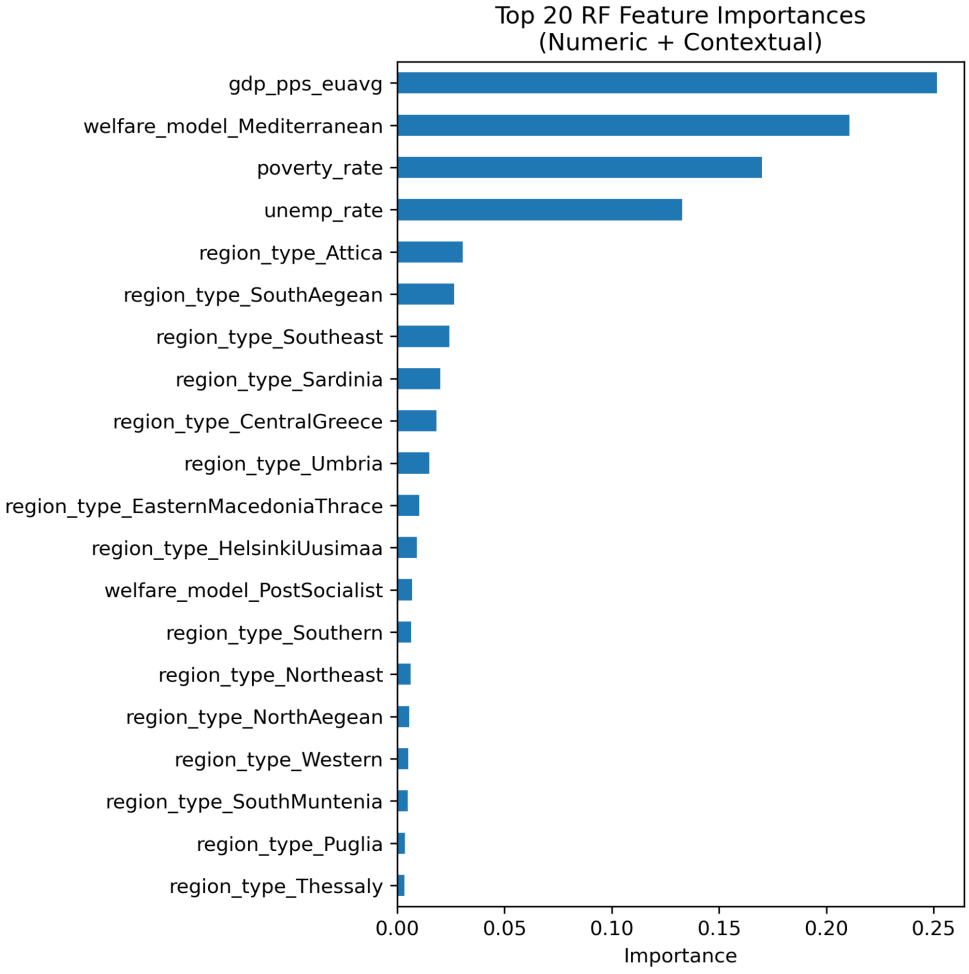

Figure 5.5: Top 20 feature importances from a random forest combining numeric and contextual features.

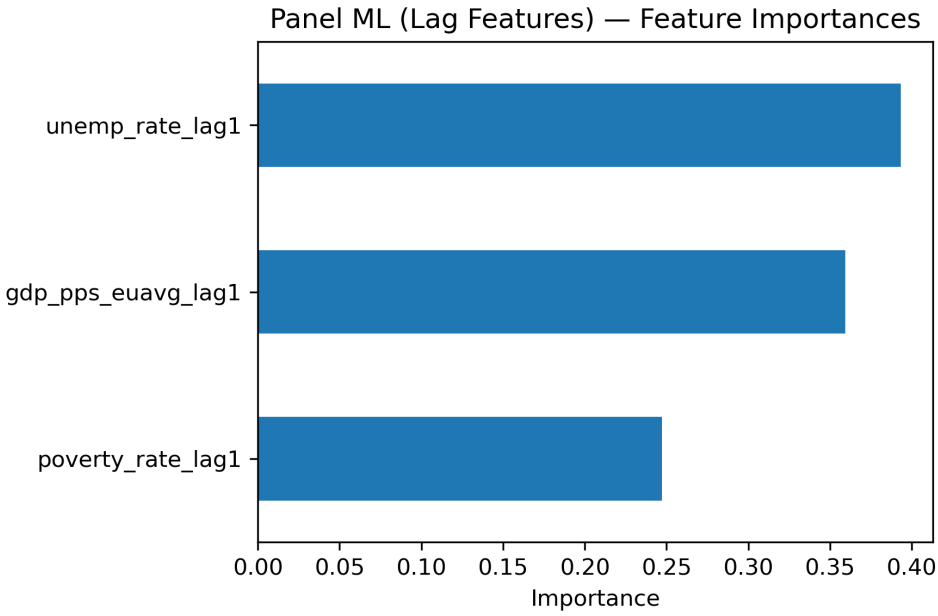

Figure 5.6: Feature importances from a lag-feature random forest.

5.6.5 Lasso (standardised predictors)

Figure 5.7 reports coefficients from a Lasso model with standardised predictors (2023). The sign pattern mirrors the OLS results: unemployment loads positively, GDP loads negatively, and poverty is comparatively weaker once unemployment is included. As with the other ML models, coefficients should be interpreted as predictive associations rather than causal effects.

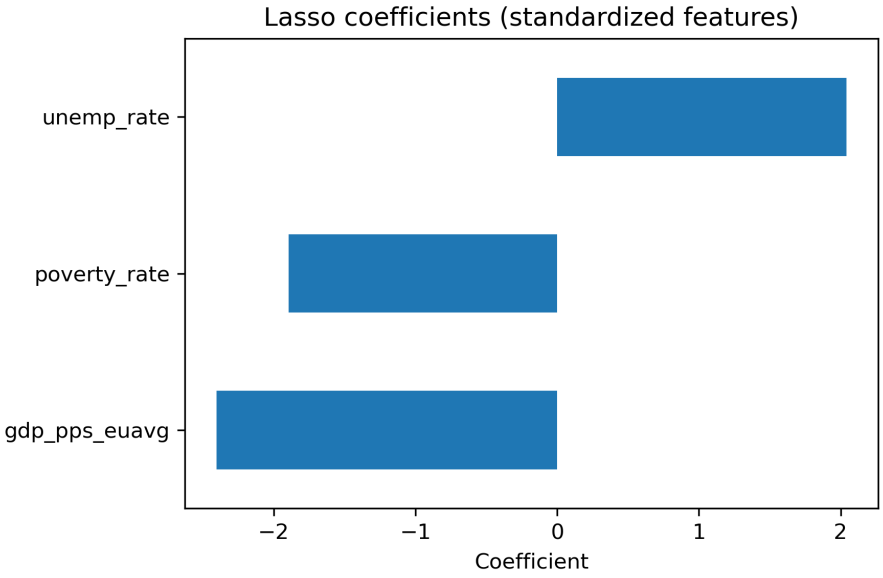

Figure 5.7: Lasso coefficients using standardised predictors (2023).

5.7 Summary of results

Across descriptive and modelling approaches, unmet medical need appears spatially concentrated in a limited set of high-burden regions, with systematic differences by broad welfare-regime groupings. Unemployment is the most consistently positive correlate, while GDP per inhabitant is consistently negative. Poverty risk is directionally positive but overlaps with unemployment in explanatory power.

Mortality provides useful context in 2021 but is not available for panel estimation in the extracted files, so it is treated as an optional 2021-only extension. Overall, the consistency of descriptive patterns, regression signs, and ML feature rankings supports the view that labour-market conditions and regional prosperity are tightly linked to perceived access barriers in the observed period.

6 Conclusion, limitations and next steps

6.1 Key findings

This internship produced a reproducible workflow to analyse regional inequalities in healthcare access using Eurostat indicators at NUTS-2 level. The final deliverable is an analysis-ready dataset covering 135 regions over 2021–2023 (379 observations), with a clear and auditable raw-to-processed pipeline.

Substantively, three findings stand out:

- **Spatial concentration:** unmet medical need is not evenly distributed, but concentrated in a limited set of high-burden territories.
- **Socio-economic correlates:** unemployment is the most consistent positive correlate of unmet need across specifications, while GDP per inhabitant is a consistent negative correlate.
- **Context matters:** broad contextual groupings (e.g., welfare regime) display systematic differences, supporting the value of combining numeric indicators with institutional context for descriptive monitoring.

6.2 Interpretation and limits

Results are descriptive associations, not causal effects. Cross-sectional models may reflect unobserved differences across countries and regions (health-system capacity, eligibility rules, service supply). The fixed-effects model mitigates time-invariant heterogeneity, but the 2021–2023 horizon is short and within-region variation is limited; therefore, coefficients should be interpreted cautiously.

Data limitations also matter. Most indicators are available for 2021–2023 in the extracted files, but mortality is only available for 2021 and is therefore used as an optional extension on the 2021 cross-section. Self-reported unmet need may capture perception and reporting behaviour in addition to objective access constraints.

6.3 Practical next steps

Without adding new data sources, the workflow can be extended in three high-value directions:

1. **Richer access barriers:** separate unmet need by reported reason (cost, distance, waiting list) to distinguish financial vs. capacity constraints.

2. **Broader covariate set:** add supply-side indicators where available (e.g., physicians per capita, hospital beds) to better characterise service capacity.
3. **Targeted monitoring outputs:** produce a compact dashboard-style set of tables/figures for regular territorial monitoring (e.g., top/bottom regions, changes over time, regime comparisons).

6.4 Skills and professional learning

From a professional-statistics perspective, the project strengthened skills in (i) harmonising heterogeneous SDMX extracts, (ii) building merge-safe tidy datasets with explicit coverage checks, and (iii) communicating results with transparent assumptions and limitations. These practices align with the quality and reproducibility standards expected in official-statistics environments.

Bibliography

[1] Eurostat. Eurostat data browser. https://ec.europa.eu/eurostat/, 2025. Accessed 2025.

[2] James G. MacKinnon and Halbert White. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29(3):305–325, 1985.

[3] Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838, 1980.