- r: region
- c: country
- t: time (year)
- $e(u)r_{rt}$: employment (or unemployment) rate

Data

This research uses a multi-step imputation procedure to address missing values in the variable $e(u)r_{r,t}$, representing regional ($r \in \{1, \ldots, R\}$) employment or unemployment rates over the years ($t \in \{t_1, \ldots, T\}$) across EU27 countries ($c \in \{1, \ldots, C\}$). The methodology is designed to balance between within-region temporal coherence and broader cross-sectional consistency across countries. It proceeds in four core stages: temporal interpolation, model-based extrapolation using region-specific trends, and Empirical Bayes shrinkage toward country-level aggregates, followed by model-based extrapolation using RE.

The first stage addresses internal gaps in each regional time series using linear interpolation by filling missing values that are bounded by non-missing observations both before and after in the time dimension, using the following formula:

$$\widehat{e(u)r}_{rt}^{(1)} = e(u)r_{rt_0} + \frac{e(u)r_{rt_1} - e(u)r_{rt_0}}{t_1 - t_0} \cdot (t - t_0), \text{ for } t_0 < t < t_1$$

This is a standard method in time-series interpolation (Shumway & Stoffer, 2017), when trends are smooth and noise is limited. Since linear interpolation cannot extrapolate beyond the data range, it is used only where internal gaps exist.

The second stage imputes values at the beginning or end of the time series (i.e., outside observed bounds). We apply region-specific forecasting models that exploit within-region temporal dynamics. Each region r has its own intercept and trend, and may share common business cycle curvature with other regions. This approach generalises standard fixed effects estimation by allowing region-specific trends (Angrist & Pischke, 2009), while remaining computationally feasible through reduced interaction dimensionality. Three models are estimated:

- Linear: $e(u)r_{rt} = \alpha_r + \gamma_r t + \varepsilon_{rt}$
- Quadratic: $e(u)r_{rt} = \alpha_r + \gamma_r t + \gamma_2 t^2 + \varepsilon_{rt}$
- Cubic: $e(u)r_{rt} = \alpha_r + \gamma_r t + \gamma_2 t^2 + \gamma_3 t^3 + \varepsilon_{rt}$

Time variables are centred to improve numerical stability and mitigate collinearity (Wooldridge, 2010). Each model is estimated on a subsample of complete regional panels to ensure consistent

parameter estimation across all models. The best-fitting model is selected using the Akaike Information Criterion (AIC) (Akaike, 1974) and the Bayesian Information Criterion (BIC) (Schwarz, 1978), calculated as:

$$\text{AIC} = \text{N} \cdot \ln\left(\frac{\text{RSS}}{\text{N}}\right) + 2\text{k}, \quad \text{BIC} = \text{N} \cdot \ln\left(\frac{\text{RSS}}{\text{N}}\right) + \text{k} \cdot \ln(\text{N})$$

where N is the number of observations, k the number of estimated parameters, and *RSS,* the residual sum of squares. The selected model is then applied to all observed data, and missing values are forecasted and denoted as $\widehat{e(u)r}_{rt}^{(2)}$

To mitigate overfitting and improve predictive accuracy, particularly in regions with volatile estimates, **at the third stage,** we apply Empirical Bayes Shrinkage Toward Country-Year Averages (Efron & Morris, 1975; Rao & Molina, 2015). This method combines the region-specific forecast from the second stage with the corresponding country-year average, yielding a weighted imputed value:

$$\widehat{e(u)r}_{rt}^{\text{final}} = w_{c,t} \cdot \overline{e(u)r}_{c,t} + \left(1 - w_{c,t}\right) \cdot \widehat{e(u)r}_{rt}^{(2)}.$$

The weight $w_{c,t} \in [0,1]$ reflects the degree of uncertainty in the regional prediction relative to its national context:

$$w_{ct} = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_{c,t}^2},$$

where $\overline{e(u)r}_{ct}$ is the country-year average, σ_{ct}^2 the cross-sectional variance of regional forecasts within the country c and year t, and σ_t^2 the between-country-year variance in country means across all periods. When within-country variation is low, forecasts are shrunk more heavily toward the national mean, while higher dispersion leads to greater trust in the region-specific estimate. This reflects principles of hierarchical modelling and small area estimation, and has been shown to outperform pure unit-level forecasts in data-sparse environments (Fay & Herriot, 1979; Ghosh & Rao, 1994).

While region-specific fixed-effects trend models provide flexible and unbiased estimates for regions with sufficiently long time series, their performance deteriorates sharply when a region has only a small number of observations. In these short panels, FE estimation requires estimating multiple region-specific coefficients – intercepts and polynomial trend terms – with limited data, resulting in high-variance estimates, unstable fitted values, and unreliable extrapolations (Wooldridge, 2010; Greene, 2012). The problem is exacerbated in polynomial specifications, where higher-order terms amplify noise when degrees of freedom are low. This small-sample instability is a well-known issue in FE models, often referred to as the “incidental parameters” problem (Neyman & Scott, 1948), whereby unit-specific parameters cannot be estimated precisely due to insufficient within-unit observations. For such low-observation regions,

attempting FE-based trend extrapolation would produce artefacts that contradict the goal of constructing smooth, structurally meaningful imputed series.

To address limitations of fixed effects (FE) forecasting in short panels, we introduce a complementary random effects (RE) forecasting procedure as the **fourth stage** of the imputation process, specifically for regions with fewer than one-third non-missing observations. In contrast to FE models, which rely solely on within-unit variation, RE models partially pool information across all regions by assuming that unobserved heterogeneity is randomly drawn from a common distribution. This allows regions with sparse data to "borrow strength" from the broader panel, improving forecast stability and reducing variance.

The RE model used for forecasting is specified as:

$$e(u)r_{rt} = \alpha + \gamma \cdot t + u_r + \varepsilon_{rt}$$

where:

- $e(u)r_{rt}$ denotes the employment rate in the region r at time t,
- α is the global intercept,
- $\gamma \cdot t$ captures the common linear time trend,
- $u_r \sim \mathcal{N}(0, \sigma_u^2)$ is a region-specific random intercept,
- $\varepsilon_{rt} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ is the idiosyncratic error term.

In this framework, the region-specific deviations u_r are shrunk toward the overall mean intercept. This shrinkage is **data-driven**, with the extent of pooling governed by the relative magnitudes of σ_u^2 and σ_ε^2. When individual regional series are short or noisy, the model places greater emphasis on the overall trend, resulting in more stable and efficient forecasts – similar to **Empirical Bayes predictions** (Raudenbush & Bryk, 2002; Gelman & Hill, 2007).

This RE forecasting approach is particularly well-suited for **sparse panels**, where classical FE estimators are unreliable or undefined due to insufficient degrees of freedom. Moreover, the RE model has been shown to outperform FE in such cases in terms of mean squared error and predictive accuracy (Searle, Casella, & McCulloch, 2009; Hsiao, 2014).

As a result, the final imputation framework adopts a **hybrid forecasting strategy**:

- Regions with sufficient time depth are forecasted using flexible FE models with selected linear, quadratic, or cubic time trends.

- Regions with insufficient data are imputed using RE forecasts that combine population-wide patterns with individual historical information.

This dual strategy ensures **robustness, efficiency, and structural coherence** across the imputed dataset, preserving idiosyncratic regional dynamics where feasible while leveraging cross-sectional information where necessary.

We incorporated **bootstrapping** into the imputation routine to add a crucial layer of inferential robustness by allowing for the quantification of uncertainty around the imputed values. Rather than relying on a single deterministic imputation, this approach repeatedly resamples the panel data (with replacement) and re-applies the full imputation pipeline – linear interpolation, trend-based forecasting, Empirical Bayes shrinkage, and random effects estimation – to generate multiple plausible imputed datasets. The resulting distribution of imputed values for each missing observation enables the computation of standard errors and confidence intervals, reflecting the variability inherent in the imputation process itself. This is particularly important when imputation results are used in downstream statistical analyses, where ignoring imputation uncertainty can lead to underestimated standard errors and overconfident inferences (Little & Rubin, 2020). Formally, for each missing (un)employment rate $e(u)r_{rt}$, we obtain $\{\widehat{er}_{rt}^{(1)}, \widehat{er}_{rt}^{(2)}, \dots, \widehat{er}_{rt}^{(B)}\}$ from B bootstrap replications and summarise the imputation as the mean $\overline{er}_{rt}$ which is retained as the final imputed value. This integration of non-parametric resampling with a model-based imputation framework ensures that final estimates reflect both observed data structure and inferential variability.

This multi-stage imputation framework is especially well-suited for subnational labour market datasets, where short panels and data sparsity pose challenges to conventional imputation approaches.

Advantages of the Imputation Framework

- Flexibility and parsimony: The use of AIC/BIC ensures that the simplest adequate trend model is selected for extrapolation.
- Temporal and spatial coherence: The method captures local regional dynamics while regularising extreme forecasts through hierarchical shrinkage.
- Statistical robustness: The Empirical Bayes component offers a principled form of regularisation, reducing variance without assuming full pooling.

- Extensibility: The framework is modular and can be adapted to alternative specifications (e.g., splines, mixed models, or multiple imputation).

Conclusion

The proposed multi-step imputation methodology provides a rigorous approach to reconstructing missing regional employment (or unemployment) rates. It combines time-series continuity, flexible trend modelling, and cross-sectional regularisation to ensure that imputed values are both data-driven and structurally consistent. This makes it well-suited for empirical research where panel completeness is critical.

References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723.
- Angrist, J. D., & Pischke, J. S. (2009). *Mostly Harmless Econometrics*. Princeton University Press.
- Efron, B., & Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70(350), 311–319.
- Fay, R. E., & Herriot, R. A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74(366), 269–277.
- Ghosh, M., & Rao, J. N. K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9(1), 55–76.
- Little, R. J. A., & Rubin, D. B. (2020). *Statistical Analysis with Missing Data* (3rd ed.). Wiley.
- Rao, J. N. K., & Molina, I. (2015). *Small Area Estimation* (2nd ed.). Wiley.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2), 461–464.
- Shumway, R. H., & Stoffer, D. S. (2017). *Time Series Analysis and Its Applications* (4th ed.). Springer.
- Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data* (2nd ed.). MIT Press.
- Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

- Greene, W. H. (2012). *Econometric Analysis* (7th ed.). Pearson.
- Hsiao, C. (2014). *Analysis of Panel Data* (3rd ed.). Cambridge University Press.
- Neyman, J., & Scott, E. (1948). “Consistent Estimates Based on Partially Consistent Observations.” *Econometrica*, 16(1), 1-32.
- Raudenbush, S., & Bryk, A. (2002). *Hierarchical Linear Models* (2nd ed.). Sage.
- Searle, S. R., Casella, G., & McCulloch, C. E. (2009). *Variance Components*. Wiley.
- Wooldridge, J. M. (2010). *Econometric Analysis of Cross-Section and Panel Data* (2nd ed.). MIT Press.
- Little, R. J. A., & Rubin, D. B. (2020). *Statistical Analysis with Missing Data* (3rd ed.). Wiley.