Handling an Inconsistently Coded Categorical Variable in Longitudinal Data: The cat2cat Method and Its Application

VI Congress of Polish Statistics, 2026

Maciej Nasiński, PhD

Faculty of Economic Sciences, University of Warsaw

2026-07-01

Motivation

Research problem

Longitudinal analysis of categorical variables requires **comparability** of category definitions across waves. In practice, occupational, industrial, health, and education classifications are frequently **revised**.

- Classification updates create **non-comparable** categories across waves.

Classification Change: Occupations in Poland

KZIS revision in 2010

The Polish occupational classification (KZIS, aligned with ISCO) was substantially redefined.

- ~**10%** of codes on the 1-digit level changed.
- ~**80%** of codes on the 3-digit level changed.
- ~**85%** of codes on the 4-digit level changed.
- Many occupations were **split**, **merged**, or **redescribed**.

Illustrative split

Building caretakers (9141) became multiple categories, including housekeeper (515301), churchkeeper (515302), maintenance worker (515303), and rest of building caretakers (515390).

Common Fixes

Approach	Benefit	Cost
Drop variable	Avoids coding conflict	Narrows the scope of analysis
Aggregate using n-digits	Restores comparability	More bias with each next digit
Specialist recoding	Uses domain expertise and rules	One-to-one category mapping causes information loss

Need

A method that keeps **admissible transitions** and carries **uncertainty** explicitly into estimation.

Note

In a complete panel with known target-period categories, direct ID-based linkage may be sufficient. Probabilistic harmonisation is most useful when some records cannot be directly linked, for example in repeated cross-sections or rotational panels.

cat2cat method: Core Mechanism

Interpretation

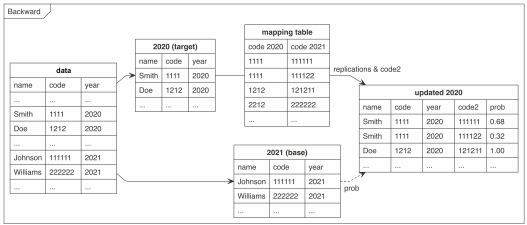
The method converts a deterministic coding problem into probabilistic harmonisation.

- If source category c can map to several targets, **replicate** the observation across all admissible targets.
- Assign a **probability weight** to each replication.
- Replication probabilities for the original record **sum to one**.
- Run standard **weighted analyses** on the transformed data (e.g., regression or trend analysis).

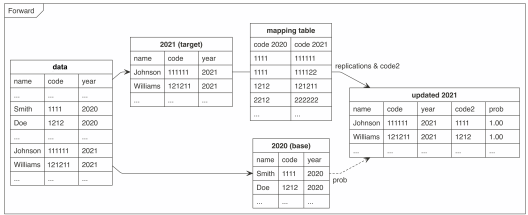
Mapping table for KZIS

OLD	NEW
1111	111101
1111	111102
1111	111103
1112	111201
1112	111202
7124	711601
7124	711603
7124	711605
7124	931290
...	...

Backward and Forward Harmonisation



Backward mapping: old waves are expressed in the newer coding system.



Forward mapping: newer waves are expressed in the older coding system.

Direction depends on research objective, reporting standard, and information content in mapping tables.

Weights and Statistical Intuition

For observation i with source category c, let $K(c)$ denote the set of admissible target categories.

$$\sum_{k \in K(c)} p_{ik} = 1, \qquad 0 \leq p_{ik} \leq 1$$

The harmonised dataset contains replicated records (i, k), each weighted by p_{ik}.

How probabilities are obtained

By using **frequencies** or **machine learning** on subject features from the base-period data.

A property

For variables that are not harmonised, **weighted means and variances are preserved** after replication.

Data and Substantive Question

Data

Structure of wages and salaries by occupations (Poland), waves: **2006, 2008, 2010, 2012, 2014**.

- ~700,000 observations per wave.
- Survey weights retained.
- Occupations harmonised with **cat2cat**.

Question

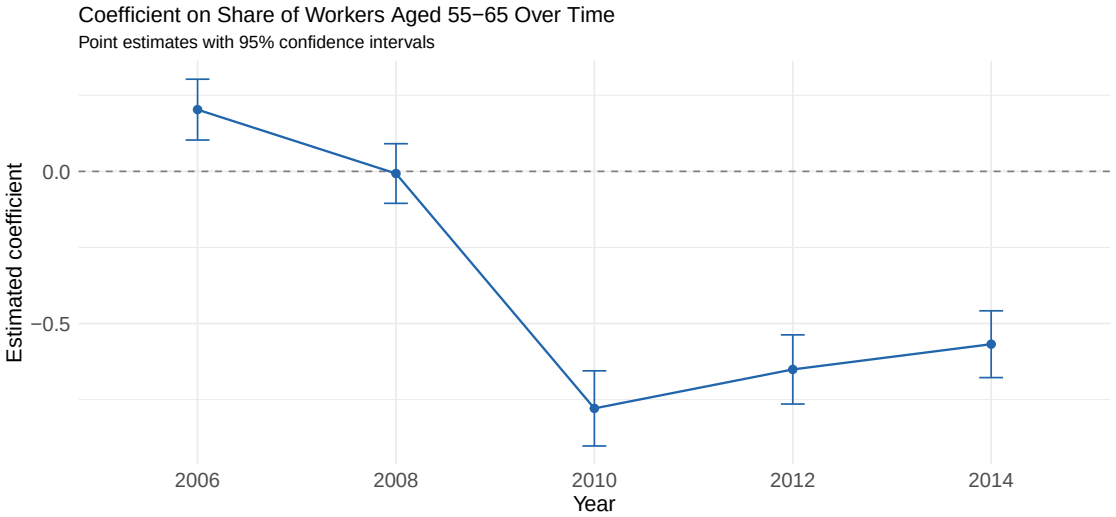
How does the share of **older workers (55–65)** relate to individual **real hourly wages**?

Empirical Model

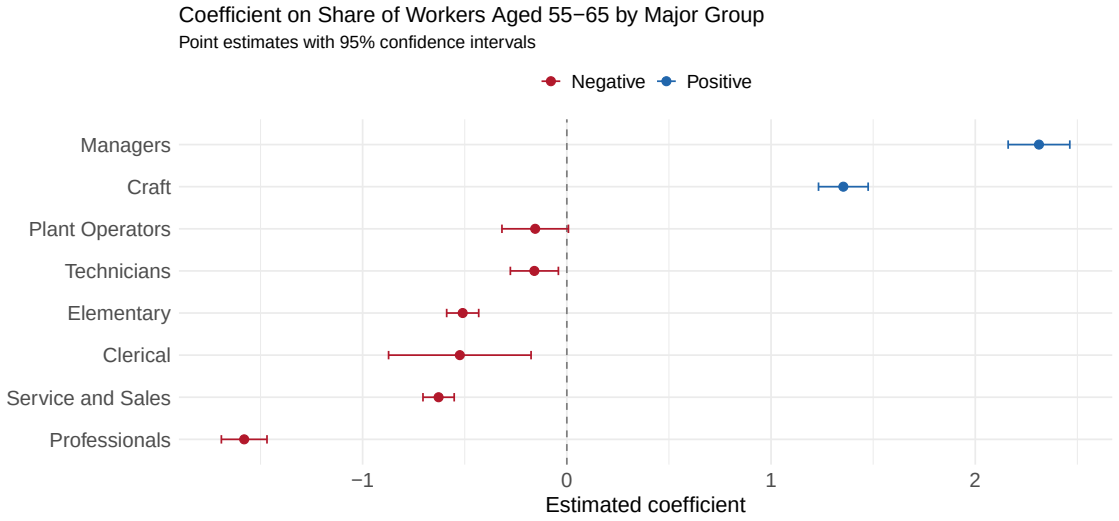
$$
\begin{aligned}
\ln w_{i,j} = \beta_0 &+ \beta_1 \text{edu}_{i,j} + \beta_2 \text{experience}_{i,j} + \beta_3 \text{experience}^2_{i,j} \\
&+ \beta_4 \textbf{share55_65}_j + \beta_5 \text{gender}_{i,j} + \beta_6 \text{work_hours}_{i,j} \\
&+ \beta_7 \text{contract_type}_{i,j} + \beta_8 \ln(\text{firm_size}_{i,j}) + \beta_9 \text{sector}_{i,j} \\
&+ \beta_{10} \text{negotiations}_{i,j} + \sum_m \beta_m \text{year}_m + \sum_n \beta_n X^{\text{NACE}}_{n,i,j} + \varepsilon_{i,j}
\end{aligned}
$$

- Estimation approach: **weighted linear regression** (Mincer-type specification) on harmonised data. Weights are **survey weights multiplied by transition probabilities**.
- w_{ij}: real hourly wage of worker i in occupation j.
- **share55_65**$_j$: frequency of workers aged 55–65 in 3-digit occupation j.

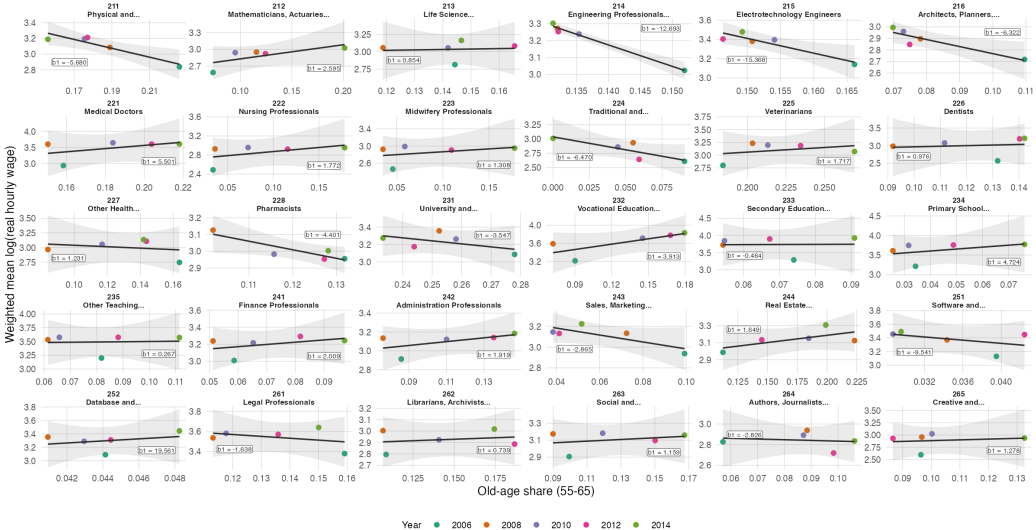
Results by Year



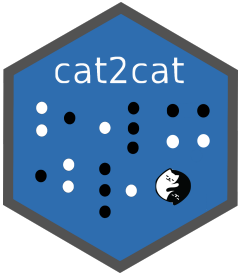
Results by Major Occupational Groups



Professionals: Heterogeneity in Coefficients Across 3-Digit Groups



Package Website and Code



- Full method description and implementation details are documented online.
- **Reproducible** examples include mapping tables, diagnostics, and estimation workflow.
- Source code (**R/Python**) and usage instructions are available for replication.

Package website and repository

Contributions, Limits, and Takeaway

Contributions

- Method: probabilistic harmonisation for evolving categorical codes.
- Empirics: improved understanding of the relationship between age structure and wages across occupational groups in Poland.
- Practice: reproducible implementation in R and Python.

Limits

- Quality depends on mapping table information and amount of data in the base period.
- Transition probabilities require external assumptions.
- Inference must include sensitivity analyses.

Takeaway

- Classification changes are an issue in longitudinal inference.
- cat2cat preserves detail while making uncertainty explicit.
- The framework generalises to ISCO, ICD, NACE/PKD.

Thank You

Thank you for your attention.

Contact: nasinski.maciej@gmail.com

Questions are welcome.