Unified-Year Data Construction Using Multiple & Hierarchical Imputation

Sainath Raja

Biomedical Engineering, NIT Raipur

Problem Background

Dataset Issue

In my dataset:

- Each feature column contains data from **only one specific year**.
- Examples:
  - Life Expectancy $\rightarrow$ 2019
  - GNI per capita $\rightarrow$ 2020
  - Diabetes rate $\rightarrow$ 2021
  - Obesity $\rightarrow$ 2018
- Therefore, features correspond to **different years**, not aligned in time.

Challenge

Machine learning models assume all features represent the **same time period**. But my dataset mixes values from 2018–2022.

Why This is a Problem?

Meaninglessness in Prediction

If feature values represent different years:

- HDI (2022)
- Diabetes (2019)
- Physicians (2021)

The ML model combines signals that **did not occur together**, causing:

- Wrong correlations
- Biased predictions
- Invalid assumptions

Goal

Convert all features to a **common target year** (e.g., 2020).

Research Paper Inspiration

Paper Context

The referenced research paper works with:

- Longitudinal medical data
- Patients observed at multiple yearly waves
- Missing years for some patients

How Missing Years Are Filled

They use:

- Multiple Imputation (MI)
- Hierarchical / Multilevel Imputation

Example:

- Data available: 2019, 2020, 2022
- Missing: 2021

MI predicts the missing 2021 value using patterns from the other years.

Adapting the Paper to My Dataset

My Scenario

Unlike the paper:

- I do not have multi-year values for each feature.
- Instead, each feature has only **one year of data**.

Key Insight

If I choose a target year (e.g., 2020), then:

- Treat each feature's 2020 value as **missing**.
- Use its available-year value as a predictor (e.g., 2019).
- MI imputes the expected value for 2020.

This Works!

MI does not require temporal sequences. It learns relationships **across features** to estimate the missing target-year values.

My Proposed Solution

Step-by-Step Approach

1. Choose a target year (example: 2020).
2. For each feature:
   - Keep its available-year value.
   - Create a new column for the target-year (2020).
   - Fill it with **missing values (NA)**.
3. Combine all features into one dataset.
4. Apply **Multiple Imputation (MICE)**.
5. The model predicts each feature's value as it would be in 2020.

Outcome

A fully unified dataset where:

- All features are in the same year
- Ready for modeling
- Free from temporal conflict

Why MI Works for This Task?

- MI models complex nonlinear relationships.
- It uses all available features as predictors.
- It produces statistically valid estimates.
- It is widely trusted in health and epidemiological studies.

Therefore:

Even if input years differ, MI can harmonize them to a single target year.

Conclusion

Summary

- My dataset contains features recorded in different years.
- For valid ML modeling, a unified-year dataset is essential.
- The research paper inspires an MI-based solution.
- MI predicts each feature's value for a chosen target year.
- Resulting dataset becomes consistent and meaningful.

Final Output = All Features Converted to the Same Year