# PRS Analysis for WSL

Lukas Graz

2025-02-25

For the **release notes** see the corresponding GitHub page

## Data Preparation

### Train Test Split for Inference

Data was split into training and test sets (50/50) for hypothesis testing to ensure valid inference after feature selection.

### Missing Values

Missing value imputation was performed using MissForest doi:10.1093/bioinformatics/btr597. This method leverages conditional dependencies between variables to predict missing values through an iterative random forest approach.

To avoid introducing spurious correlations between different variable sets, we imputed the following data groups separately:

- PRS variables on the complete dataset
- Mediators on training data only
- GIS variables on training data only
- Mediators for prediction analysis
- GIS variables for prediction analysis
- PRS variables for prediction analysis

Mediators and GIS variables were intentionally not imputed on the test set to maintain valid inference, as MissForest does not provide a mechanism to propagate imputation uncertainty. An alternative would be the `mice`-routine, which could be implemented in future analyses. Missing values in the test set predictors remained untreated, which is justified under the missing

completely at random (MCAR) assumption—where missing values occur independently of all other variables.

For the prediction analysis, fewer statistical assumptions are required, so using the MissForest approach does not violate any assumptions.

PRS variables could have been imputed separately for training/test sets and prediction analysis, but we prioritized simplicity as these variables serve only as response variables.

Additionally, we compared MissForest with simpler imputation methods (variable-wise and observation-wise mean imputation) for the PRS variables. Results confirmed that MissForest consistently outperformed these alternatives.

## Main Analysis

### Response Variable Selection

- Aggregated mean
- LA (Fascination)
- BA (Being Away)
- EC (Extent Coherence)
- ES (Compatibility)

**PCA Verification** of this approach. Key findings:

- Data can be well approximated with 3-4 dimensions
- First dimension is close to weighted average of all variables (correlation >0.99)
- EC (Extent Coherence) shows most divergence (see PC2)
- LA (Fascination) and BA (Being Away) show similarity (see PC1-PC3)
- Aggregated PRS variables justified by PCA results (similar rotation values), supporting use of mean

### Prediction Analysis with Machine Learning Methods

Details and results in the notebook.

This section investigates predictive relationships between Perceived Restorativeness Scale (PRS) variables, mediator variables, and Geographical Information System (GIS) variables using various machine learning approaches. We employed a systematic methodology to quantify the predictive power of different variable combinations.

### Methodological Approach

We evaluated multiple machine learning models using the mlr3 framework (cite doi:10.21105/joss.01903) :

- Linear models (baseline)
- XGBoost (gradient boosting with tree-based models and hyperparameter tuning for learning rate and tree depth) (cite arxiv:1603.02754)
- Random Forests (with default parameters) (cite doi:10.1023/A:1010933404324)

Performance was measured as percentage of explained variance on hold-out data, calculated as (1 - MSE/Variance(y)), where MSE represents mean squared error.

### Model Combinations

To systematically explore predictive relationships, we tested four model configurations:

1. PRS ~ GIS: Predicting PRS variables using only GIS variables
2. PRS ~ GIS + Mediators: Predicting PRS variables using both GIS and mediator variables
3. PRS ~ Mediators: Predicting PRS variables using only mediator variables
4. Mediators ~ GIS: Predicting mediator variables using GIS variables

### Results

- GIS shows limited predictive power for PRS on ES (5% variance explained)
- GIS + Mediators explain 25% of PRS variance
- Mediators alone explain majority of PRS variance
  - GIS primarily helps with ES through tree-based methods
  - Suggests GIS effect is more interaction-based than direct
  - Similar reduction in tree-based methods observed in BA

## Hypothesis Testing: Investigation of Variable Effects on Perceived Restorativeness Scale

Details and results in the notebook.

Here we investigated which variables (including their interactions) influence PRS variables using multiple linear regression. With 190 variables (counting interactions), the variance inflation factor (VIF) was high and the multiple testing problem severe. We therefore implemented a stepwise feature selection using Bayesian Information Criterion (BIC) on the training data,

starting with an empty model to help computational complexity. Selected features were subsequently used to fit models on the test set to obtain valid p-values. To keep the coefficients interpretable in the presence of interactions, each variable is scaled to mean 0 and standard deviation 1.

### Model Specification and Analysis

The analysis systematically explored two key relationship pathways:

1. Mediators ~ $(\text{GIS})^2$ - examining how environmental features predict psychological mediators
2. PRS ~ $(\text{Mediators} + \text{GIS})^2$ - investigating how both environmental features and psychological mediators contribute to perceived restorativeness

For each target variable, we constructed a separate model using stepwise selection and evaluated it on the test dataset.

### Results

- For HM_Noise (now removed): Continuous mediator outperforms categorical (scaled to mean 0, sd 1)
- Full `mice` NA-handling likely unnecessary
  - Models use few variables
  - Only LNOISE shows high NA count
  - Information detection still fails
- Significant edges remain in SEM (see all interactions)