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
- FA (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)
- FA (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 ~ (GIS)² - examining how environmental features predict psychological mediators
2. PRS ~ (Mediators + 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)

All Interactions: Mediators ~ (GIS)^2

Significant codes as usual: `0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1`

Covariate	FEELNAT	LNOISE
(Intercept)	0.062	-0.001
LCARTIF_sqrt	-0.152**	-0.124*
LCARTIF_sqrt:RL_NDVI	0.115**	
OVDIST_sqrt	0.027	
RL_NDVI	0.150***	
RL_NOISE		-0.242***

Source: Article Notebook

4

PRS ~ (Mediators + GIS)^2

Significant codes as usual: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Covariate	MEAN	FA	BA	EC	ES
(Intercept)	-0.005	0.006	-0.008	-0.006	0.027
DISTKM_sqrt	-0.019				0.073.
FEELNAT	0.252***	0.266***	0.252***	0.058	0.270***
FEELNAT:LNOISE	0.022	-0.036			
LCFOREST_sqrt	-0.066.			-0.114**	
LNOISE	0.212***	0.170***			0.142**
LNOISE:FEELNAT					-0.013
RL_NDVI		-0.105**			

Source:

Predict RL via HM

Details/Code and results in the notebook.

Procedure: Stepwise feature selection using BIC on training data and subsequent model fitting on test data. Performed seperately for `RL_NDVI` and `RL_NOISE`.

Predictors:: `HM_NDVI + HM_NOISE + LANG + AGE + SEX + SPEED_log + JNYTIME_sqrt` with all two-way interactions.

RL_NDVI

```
Call:
lm(formula = formula(step_ndvi), data = D_tst)

Residuals:
   Min     1Q Median     3Q    Max
-3.347 -0.426  0.178  0.682  1.874

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      -0.2041     0.0859   -2.37   0.0179 *
HM_NDVI           0.1572     0.0386    4.07  5.4e-05 ***
LANGGerman        0.2689     0.0972    2.77   0.0058 **
```

```
LANGItalian              -0.0768    0.1825    -0.42    0.6740
SPEED_log                -0.1112    0.0416    -2.67    0.0077 **
JNYTIME_sqrt              0.1059    0.0406     2.61    0.0094 **
HM_NDVI:SPEED_log        -0.1619    0.0399    -4.05  5.7e-05 ***
HM_NDVI:JNYTIME_sqrt     -0.1194    0.0395    -3.02    0.0026 **
SPEED_log:JNYTIME_sqrt   -0.0854    0.0394    -2.16    0.0308 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.951 on 601 degrees of freedom
Multiple R-squared:  0.108, Adjusted R-squared:  0.0963
F-statistic: 9.11 on 8 and 601 DF,  p-value: 7.29e-12
```

Source: Predict RL via HM

- $R^2 = 0.08$
- Higher HM_NDVI corresponds to slightly higher RL-NDVI
- higher JNYTIME_sqrt corresponds to slightly higher RL-NDVI
- The faster (or further) you travel to RL, the more the RL_NDVI differs from HM_NDVI (negative interaction effect)

RL_NOISE

```
Call:
lm(formula = formula(step_noise), data = D_tst)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9524 -0.7719 -0.0133  0.6588  2.8255

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           -0.000822   0.037211   -0.02    0.982
HM_NOISE               0.240824   0.037282    6.46  2.2e-10 ***
SPEED_log             -0.065753   0.037374   -1.76    0.079 .
JNYTIME_sqrt          -0.314663   0.037380   -8.42  2.8e-16 ***
HM_NOISE:JNYTIME_sqrt -0.037327   0.036309   -1.03    0.304
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.919 on 605 degrees of freedom
```

```
Multiple R-squared:  0.161, Adjusted R-squared:  0.156
F-statistic: 29.1 on 4 and 605 DF,  p-value: <2e-16
```

- R² = 0.184
- Participants can't completely escape HM_NOISE (HM_NOISE positive predictor)
- LANGItalians have it louder (than LANG de/fr)
- Longer JNYTIME_sqrt leads to lower NOISE

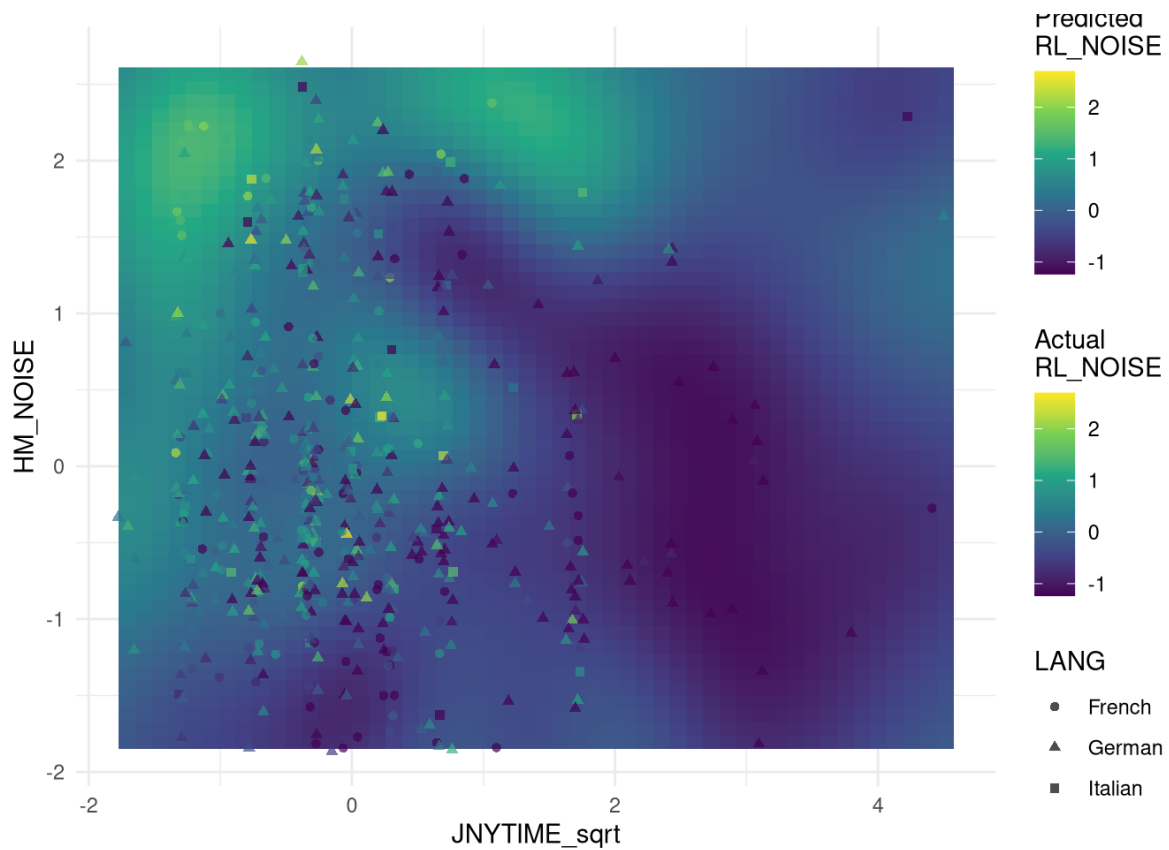Visualizing the effect of `HM_NOISE` and `JNYTIME_sqrt` on `RL_NOISE`:

Figure 1

Plots

see the notebook.