The current set-up of my research is as follows: I first split up the data in a 70/30 train-test set where none of the individuals in the train set occur in the test set, and vice versa. I then split the train set into four folds, such that we obtain four sets of 75/25 train-validation sets. Again, none of the individuals in the train set occur in the validation set of that fold, and vice versa.

Now, our mixed models can make two different types of prediction:

1. Predicting observations for new clusters/individuals, with no past observations;

2. Predicting future observations for clusters/individuals within the sample.

In case of predicting new individuals (scenario 1), the predicted value equals the estimated fixed-effects, as the random effects are unknown and set to zero. In contrast, when we predict future observations for individuals which we have already seen during model fitting (scenario 2), we set the prediction equal to the estimated fixed-effects plus the individual specific random effect.

The way our train-validation-test set-up is now, we are never in the second scenario, which is good, because our eventual tuned model needs to be implemented in the microsimulation model of EMC. Since the model is used to simulate data points, we can compare this most to scenario 1, where we predict new individuals only.

However, I would also be interested in seeing how the comparison between model performance across all models hold up when we consider the second scenario, since ‘regular’ machine learning models estimate scenario 2 and 1 in the same manner, whereas the MERF models exploit the random effect that we already previously estimated to predict new observations of an individual in scenario 2. This would be a secundary analysis, just to round out my research and shine light on the different abilities of each type of model.

To do so, I figured I might make a small extra\* test data set, consisting of approximately 5,000 individuals of which at least one observation occurs in the training set, and (at least) one of the last observations will need to be predicted. For illustration purposes, see the following figure:

Table

Description automatically generated

My main analysis will be on the dark blue non-test data and pink test data, but before I execute my analysis I create an extra test set denoted by \* in the salmon-y color. This test set includes:

1. INDIVIDUALS WITH TWO ROUNDS:
   1. Individuals with 1 observation in non-test data & 1 observation in the extra\* data set [ 3,000 individuals ]
2. INDIVIDUALS WITH THREE ROUNDS:
   1. Individuals with 1 observation in non-test data & 2 observations in the extra\* data set [ 750 individuals]
   2. Individuals with 2 observations in non-test data & 1 observation in the extra\* data set [ 1,500 individuals]
3. INDIVIDUALS WITH FOUR ROUNDS
   1. Individuals with have 1 observation in non-test data & 3 observations in the extra\* data set [ 333 individuals]
   2. Individuals with 2 observations in non-test data & 2 observation in the extra\* data set [ 500 individuals]
   3. Individuals with 3 observations in non-test data & 1 observation in the extra\* data set [ 1,000 individuals ]

I figured an extra\* test set of ±9,000 observations would be good. Specifically, I thought I would make the extra\* test set to contain 3,000 *observations* each for case 1, 2, and 3.

So:

* 3,000 individuals / 3,000 observations for case 1.a.
* 750 individuals / 1,500 observations of case 2.a.
* 1,500 individuals / 1,500 observations of case 2.b
* 333 individuals / 999 observations of case 3.a.
* 500 individuals / 1,000 observations of case 3.b
* 1,000 individuals / 1,000 observations of case 3.c.

But honestly the composure of this data set is rather arbitrary, I just want to get a wide arrangement of possibilities in the data set. If you have any recommendations, I’d be more than happy to hear them.