# Using Ensembles to Improve Forecast Performance

Exposé

Juan Sebastian Aristizabal Ortiz, Emmanuel Tchoumkeu-Ngatat

19 11 2021

## Problem:

Model Ensembles usually outperform individual models and forecasters in terms of predictive performance and therefore play an important role in any applied forecasting setting (e.g. epidemiology, finance, weather forecasting). Learned ensembles that adjust ensemble weights based on past performance hold great potential, but empirically it has proven surprisingly difficult to improve on simple mean or median ensembles. An important question therefore is what form of ensemble to choose i.e. based on *what criteria*. Often, researchers have to decide on what kind of ensemble to use and can only know much later whether their choice was good.

## Setting

We have two different data sources at our disposition:

- The first one consists of *human made forecasts* of COVID-19 arisen in the context of the UK COVID-19 Forecasting Challenge.
- The second one consists of *model based forecasts* postulated by research institutions for the European Forecast Hub. In this data set we have 31 different model types. True values are also included.
- Similar to the last one, the US Forecast Hub. A larger data set.

## Objective

This research project aims to investigate model ensembles in an epidemiological setting and tries to establish heuristics for when to use which ensemble type. Criteria for which heuristics/guidelines need to be developed are:

1. the *optimal* number of models to aggregate.
2. the *easiest* and *most efficient* weighting mechanism i.e. by e.g. ex-ante sub-setting the models considered by the ensemble by means of a goodness criteria.

3. the effect that the *similarity*, *difference*, *presence* of aggregated models may have on the ensemble output.

## Statistical Approach

As already mentioned, ensemble show to be highly difficult to improve empirically. We thus consider the "simple" *mean ensemble* and the *median ensemble* and build up from there.

**N**

To investigate the first criteria, we proceed as follows:

- We set the framework of "ensembling" and "comparing/evaluating" by means of a common function, that allows for efficient method usage and score calculation. For evaluating performance, we use WIS (Weighted Interval Score), a proper scoring rule.

- We first proceed by comparing untrained ensemble-performance on the European Forecast hub data set. Specifically, the effect of the *number of models aggregated*. To this end, for each value of $n = 1, ..., 31$ representing the number of models to be aggregated, we:
  i. Iteratively sample all possible combinations.
  ii. For each sample, we built an ensemble.
  iii. calculate the score for each ensemble.
  iv. average all scores.
  v. finally compare average performance for each $n$ value.
  vi. analyse results.

**Aggregation**

For the second criteria, consider now a training setting. We aim to compare

- *Weighted mean ensembles*, where a weight is assigned to every single component of the ensemble and then optimize train vs.
- *Ex-Ante selected mean ensembles* [1] where the researcher selects the variables to be considered based on prior judgment of their worthiness.

Hypothetically, the latter should be easier to compute. This should be assessed by * the training time needed to compute * Performance score in terms of WIS and bias.

**Variable influence and similarity; Wisdom of the crowds.**

Further analysis regards the influence the addition of a further model may have on overall performance. By considering the similarity/difference of models to aggregate we investigate furthermore

- How individual model influence depends on the homogeneity of the ensemble.
- Which (homogeneous or diverse) ensembles provide the best performance.
- If a "good" ensemble can arise from "bad" models.

To evaluate influence, we use leave-one-out-ensembles and compare by means of WIS and bias. To analyse similarity, we calculate pairwise Cramér-distance between any two member models and taking the average distance across all positive pairs.

We further consider Forecast calibration and PIT histograms to evaluate for under-over performance tendencies in our models.

## Conclusions and notes

We have presented the scope of our analysis. The priorities of the project follow the structure of this exposé in a hierarchical manner. Because of time constraints, it is expected that the analysis of three criteria may prove difficult to achieve. Our project partner is aware of this. We will work on the first criteria and the proceed and report accordingly.

[1] Name given by us; It is needed to look up the literature, if there is any