GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN IN PUBLICA COMMODA SEIT 1737

Analyzing the Influence of Model and Ensemble Structure on Performance of Real-Time COVID-19 Forecasts

Master thesis for the Master of Science course "Applied Statistics"
at the University of Göttingen

Author:
Friederike Sarah Becker,
Student ID: 21914687,
born in Herten, Germany

Supervisors
Prof. Dr. Thomas Kneib
Nikos Bosse

Submitted on June 17, 2022
to the Faculty of of Business and Economic Sciences at Göttingen University

Contents

List of Figures

Acronyms

cdf cumulative distribution function

ECDC European Center for Disease Prevention and Control

1 Introduction

In recent years evidence in epidemiology - as well other fields - has accumulated that in order to obtain accurate and well calibrated forecasts of targets of interest, such as case numbers of a disease in a given week, it is often advisable to not rely solely on single model outputs, but to rather consider an aggregate of forecasts made by a group of models, widely referred to as ensemble forecasts (give some cites).

The recent COVID-19 epidemic has turned out to be no exception in this regard. In order to obtain an accurate picture of current disease dynamics for decision makers and following similar efforts in the US, in March 2021 the European Center for Disease Prevention and Control (ECDC) has instigated the European Forecast Hub, collating weekly real-time distributional forecasts for short-term incidence COVID-19 cases and deaths from independent modeling teams noa (2021). It was found that, in general, ensemble forecasts that aggregate all single model outputs into a single common forecast showed more consistent performance than any single model for both case and death incidence forecasts Sherratt and Gruson. (citecitecite-somemore).

However, while evidence for the advantages of employing an ensemble strategy is ubiquitous, a question that remains is which type of model or ensemble structure gives an edge over others, as well as exploring the interaction between ensemble and model structure. Hence, the question is whether we can identify certain individual modeling or ensemble strategies that consistently perform better than others within the European Forecast Hub. Or alternatively - in lieu of succeeding to establish universal dominance statements - whether it is possible to identify certain situations in which some models or ensemble compositions have an edge over others. In this thesis, the focus will lie on three dimensions of this issue:

First, for the goal of eliciting accurate short-term forecasts, it is an ongoing topic of research whether models that are to some extent explicitly epidemiological in nature, that is, seeking to model transmission dynamics in a population, are to be preferred over models that solely rely on past information of the target time series and are thus agnostic to the underlying transmission dynamics (Funk and Abbott). Different diseases and epidemics warrant/require varying approaches in modeling strategy and we thus aim to investigate whether, for the European Forecast Hub, definitive or (more likely) situation-dependent rankings can be established. For example, Bracher et al. (2021) identified that statistical models, which rely on past time series dynamics for prediction, are consequently slow to respond to changes in trends, raising the question whether compartmental models fare better in this regard. These types of results can then potentially also be leveraged for forecast composition - in fact, Taylor and Taylor (2021) conjecture that during low-incidence periods, compartmental models should perform better than statistical ones and that ensembles consisting purely of these models should consequently exhibit better performance. Lastly, one must not forget that there might be different require-

ments and goals for forecasts depending on who is using them and under which circumstances - these differing goals can be captured with the choice of a corresponding scoring rule. To this end, it is conceivable that, for instance, one model type fares better with regard to point accuracy, while another might exhibit better coverage - for example, it could be evident that one model type exhibits more overconfidence in certain situations. We will thus investigate how the preferred model type varies with the choice of scoring rule.

Another central question in past and current analysis of both European and US COVID Hub data has been whether ensembling procedures should discriminate between their potential member models based on merit. That is, if the ensemble should assign higher weights to models that, in terms of the scoring metric of interest, have performed better in the past than other models, as opposed to applying equal weighting for all models. Results on this procedure of performance-based weighting have been somewhat mixed. For predicting number of deaths in the US forecast hub data, Taylor and Taylor (2021) find that, especially for states that exhibit high mortality, performance-based weighting leads to higher prediction accuracy. (Also include paper by Brooks here, which found no advantage.) Conversely, in the European Forecast Hub, Sherratt and Gruson find no significant improvement of weighted methods in comparison with unweighted and, similarly, Bracher et al. (2021) also find no systematic benefits of the weighted approach for data from Germany and Poland. Taylor and Taylor (2021) identify a possible reason for the shortcomings of weighted approaches: they require comparable records of historical accuracy and are thus challenging to implement in datasets where model availability fluctuates - this is presumably especially an issue for the European Forecast Hub, where it is common to observe large participation gaps for models.

Lacking evidence for an alternative superior ensembling technique, the European Forecast Hub has thus relied on the unweighted mean ensemble, which was then superseded by the unweighted median ensemble due to its higher resistance to outliers (Sherratt and Gruson). Consequently, the question arises whether it is perhaps possible to establish some guidelines - for example pertaining to situational circumstances, model structure or relation between models - in a data setting where it is not feasible/beneficial to rely on hard and fast mathematical rules via weighted ensembles. The hope is that through the investigation of ensemble behavior in response to these issues, we can establish some heuristics for when it is useful to have certain models in an ensemble or, as a softer goal, to simply gain more insight into how ensemble performance varies in response to the aforementioned dimensions.

There are several possible dimensions to investigate with regard to this research question, with some lines of inquiry having come up before in earlier studies. For instance, a worthwhile question to investigate is whether a model can consistently be underperforming (by one or any scoring metric used) in comparison to other individual models and the ensemble, but can nevertheless provide occasional or consistent benefit when included in said ensemble. An easy example would be a model that consistently underpredicts a target: this is of course in

and of itself undesirable, but could provide great benefit to an ensemble that has a tendency to overpredict the target - a similar case was identified in Bosse and Abbott (2021), and it would be interesting to investigate whether such models can be found in the existing pool of the Hub models. Along these lines, there is also the question whether the choice of adding an additional model is somehow dependent on the summary function used to generate the ensemble: in the case of the models investigated by Bosse and Abbott (2021), it turned out to be somehow "safer" to add models in a median than a mean ensemble, presumably as it is more resistant to outliers.

With regard to the aforementioned wide success of ensembles in the forecasting realm, a notion that seeks to explain this success is that averaging over a number of separate models both acts as a mitigator for individual model bias and reduces overall performance variation (do the cites). It is thus conceivable that including models that are too similar and hence somehow make "the same type of mistake" (be it directional, or in relation to over-/ or underconfidence) could, in a sense, hijack/overpower/overtake/skew the ensemble and thus be detrimental to its performance. In turn, this would mean that establishing a notion of "too big" (find more succinct word) model similarity and consequently culling models based on this notion could be beneficial. To this end, we use the Wasserstein 2-metric/Cramer distance, as applied to a discrete set of quantiles and investigate whether excluding single or multiple models that form a sort of "model cluster" improves ensemble performance.

Another question we'd like to investigate with regard to ensemble composition is the consistency and variation of forecast performance in relation to the number of its member models. As we believe that including additional models lowers the variation of the ensemble and thus improves its performance, the expectation here is "more is always better" - nevertheless, more knowledge on how exactly ensemble performance relates to ensemble size could be very valuable, especially in situations where resources are limited and it's not immediately clear whether investing into additional models would be rewarding. To this end, we will randomly sample all sets of member models in the Hub, as an answer to the question of "what would have been if we'd have less models?". Furthermore, we consider whether ensemble performance on average declined in weeks where not a lot of forecasts were available.

As already mentioned, the entire procedure should, to some extent, be regarded more investigatively/inquisitively and with the aim to establish heuristics/soft guidelines rather than hard and fast rules. By nature, the methods described here have a certain ad-hoc character, in a way that having a mathematically formulated rule that "simply" weights by past performance is not. It is possible that no exact guidelines emerge from the analysis, or that emerging results will be very specific to the data at hand and not necessarily generalizable. Nevertheless, we still deem there to be value in this type of analysis, as it can lead to greater understanding of ensemble behavior - performing such an inquisitive deep dive can be regarded as the novel contribution of this thesis.

There are several dimensions one can investigate here: low-incidence periods, periods of exponential growth, periods where not a lot of models are available, model similarity. Finally (note: and if there is time), we want to consider whether tweaking the method of ensemble building might lead to an increase in performance. The aforementioned/currently used methods treat each quantile forecast as separate and thus build a common ensemble forecast by applying some sort of summary function (usually mean or median) to each separate quantile set. [1] However, a potentially sensible/worthwhile/viable alternative approach could be first building a common probability distribution from the set of available forecasts, then taking the quantiles from this aggregate/mixture distribution. To this end, we consider imputing a cumulative distribution function (cdf) for each forecast separately, then taking the ensemble's quantiles from the aggregate/mixture cdf. Put succinctly, we want to investigate whether aggregating the forecasts in cdf rather than quantile space could provide a benefit.
In a nutshell, the goal is of this thesis is to investigate ensemble behavior as it relates to its member models and the current epidemic circumstances, with the hope of potentially deriving some heuristic guidelines for ensemble composition from the findings.

[1] in the case of the median, as identified by Bracher et al. (2021)), this might lead to not so well formed distributions.

Bibliography

European Covid-19 Forecast Hub, 2021. URL https://covid19forecasthub.eu/papers.html.

Nikos Bosse and Sam Abbott. Comparing human and model-based forecasts of COVID-19 in Germany and Poland. 2021.

J. Bracher, D. Wolffram, J. Deuschel, K. Görgen, J. L. Ketterer, A. Ullrich, S. Abbott, M. V. Barbarossa, D. Bertsimas, S. Bhatia, M. Bodych, N. I. Bosse, J. P. Burgard, L. Castro, G. Fairchild, J. Fuhrmann, S. Funk, K. Gogolewski, Q. Gu, S. Heyder, T. Hotz, Y. Kheifetz, H. Kirsten, T. Krueger, E. Krymova, M. L. Li, J. H. Meinke, I. J. Michaud, K. Niedzielewski, T. Ożański, F. Rakowski, M. Scholz, S. Soni, A. Srivastava, J. Zieliński, D. Zou, T. Gneiting, M. Schienle, List of Contributors by Team, CovidAnalytics-DELPHI, Michael Lingzhi Li, Dimitris Bertsimas, Hamza Tazi Bouardi, Omar Skali Lami, Saksham Soni, epiforecasts-EpiExpert and epiforecasts-EpiNow2, Sam Abbott, Nikos I. Bosse, Sebastian Funk, FIAS FZJ-Epi1Ger, Maria Vittoria Barbarossa, Jan Fuhrmann, Jan H. Meinke, German and Polish Forecast Hub Coordination Team, Johannes Bracher, Jannik Deuschel, Tilmann Gneiting, Konstantin Görgen, Jakob Ketterer, Melanie Schienle, Alexander Ullrich, Daniel Wolffram, ICM-agentModel, Łukasz Górski, Magdalena Gruziel-Słomka, Artur Kaczorek, Antoni Moszyński, Karol Niedzielewski, Jedrzej Nowosielski, Maciej Radwan, Franciszek Rakowski, Marcin Semeniuk, Jakub Zieliński, Rafał Bartczuk, Jan Kisielewski, Imperial-ensemble2, Sangeeta Bhatia, ITWW-county repro, Przemyslaw Biecek, Viktor Bezborodov, Marcin Bodych, Tyll Krueger, Jan Pablo Burgard, Stefan Heyder, Thomas Hotz, LANL-GrowthRate, Dave A. Osthus, Isaac J. Michaud, Lauren Castro, Geoffrey Fairchild, LeipzigIMISE-SECIR, Yuri Kheifetz, Holger Kirsten, Markus Scholz, MIMUW-StochSEIR, Anna Gambin, Krzysztof Gogolewski, Błażej Miasojedow, Ewa Szczurek, Daniel Rabczenko, Magdalena Rosińska, MOCOS-agent1, Marek Bawiec, Marcin Bodych, Tomasz Ożański, Barbara Pabjan, Ewaryst Rafajłlowicz, Ewa Skubalska-Rafajłowicz, Wojciech Rafajłowicz, Agata Migalska, Ewa Szczurek, SDSC ISG-TrendModel, Antoine Flahault, Elisa Manetti, Christine Choirat, Benjamin Bejar Haro, Ekaterina Krymova, Gavin Lee, Guillaume Obozinski, Tao Sun, Dorina Thanou, UCLA-SuEIR, Quanquan Gu, Pan Xu, Jinghui Chen, Lingxiao Wang, Difan Zou, Weitong Zhang, USC-SIkJalpha, Ajitesh Srivastava, Viktor K. Prasanna, and Frost Tianjian Xu. A pre-registered short-term forecasting study of COVID-19 in Germany and Poland during the second wave. *Nature Communications*, 12(1):5173, December 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-25207-0. URL https://www.nature.com/articles/s41467-021-25207-0.

Sebastian Funk and Sam Abbott. Short-term forecasts to inform the response to the Covid-19 epidemic in the UK.

Katharine Sherratt and Hugo Gruson. (Draft) Predictive performance of multi-model ensemble forecasts of COVID-19 across European nations.

James W. Taylor and Kathryn S. Taylor. Combining probabilistic forecasts of COVID-19 mortality in the United States. *European Journal of Operational Research*, page S0377221721005609, June 2021. ISSN 03772217. doi: 10.1016/j.ejor.2021.06.044. URL https://linkinghub.elsevier.com/retrieve/pii/S0377221721005609.