SANDIA REPORT

SAND2019-0854
Unlimited Release
Printed January, 2019

Conditioning multi-model ensembles for disease forecasting

Jaideep Ray, Katherine Cauthen, Sophia Lefantzi and Lynne Burks

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Approved for public release; further dissemination unlimited.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online ordering: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2019-0854
Unlimited Release
Printed January, 2019

Conditioning multi-model ensembles for disease forecasting

Jaideep Ray
Sandia National Laboratories, P.O. Box 969, Livermore, CA 94550-0969,

Katherine Cauthen & Sophia Lefantzi
Sandia National Laboratories, P.O. Box 5800, Albuquerque, NM 87185-5800,

and

Lynne Burks
One Concern, Inc, 169 University Ave., Palo Alto, CA 94301

Abstract

In this study we investigate how an ensemble of disease models can be conditioned to observational data, in a bid to improve its predictive skill. We use the ensemble of influenza forecasting models gathered by the US Centers for Disease Control and Prevention (CDC) as the exemplar. This ensemble is used every year to forecast the annual influenza outbreak in the United States. The models constituting this ensemble draw on very different modeling assumptions and approximations and are a diverse collection of methods to approximate epidemiological dynamics. Currently, each models' predictions are accorded the same importance, or weight, when compiling the ensemble's forecast. We consider this equally-weighted ensemble as the baseline case which has to be improved upon. In this study, we explore whether an ensemble forecast can be improved by "conditioning" the ensemble to whatever observational data is available from the ongoing outbreak. "Conditioning" can imply according the ensemble's members different weights which evolve over time, or simply perform the forecast using the top k (equally-weighted) models. In the latter case, the composition of the "top-k-set" of models evolves over time. This is called "model averaging" in statistics. We explore four methods to perform model-averaging, three of which are new.. We find that the CDC ensemble responds best to the "top-k-models" approach to model-averaging. All the new MA methods perform better than the baseline equally-weighted ensemble. The four model-averaging methods treat the models as black-boxes and simply use their forecasts as inputs i.e., one does not need access to the models at all, but rather only their forecasts. The model-averaging approaches reviewed in this report thus form a general framework for model-averaging *any* model ensemble.

Acknowledgment

This report documents work performed for the US Department of Defense, Defense Threat Reduction Agency, under contracts IA DTRA 10027-22678 and IA DTRA 10027-25894. This report describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the report do not necessarily represent the views of the U.S. Department of Energy or the United States Government. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Contents

Executive Summary **13**

1 Introduction **15**

1.1 Literature Review . . . 17

1.1.1 Model-Averaging Methods . . . 17

1.1.2 Model-Averaging with Disease Models . . . 18

1.2 Data and Model Ensemble . . . 20

2 Dynamic Model Averaging (DMA) **21**

2.1 Formulation . . . 21

2.2 Results . . . 23

2.3 Comparison with other MA Methods . . . 25

2.3.1 Bayesian Model Averaging . . . 27

2.3.2 Ranking-based Model Averaging . . . 28

2.4 Limits of Applicability . . . 32

2.4.1 Non-transferability of Model Weights . . . 32

2.4.2 Non-transferability of Model Rankings . . . 33

2.4.3 Impact of Optimized α . . . 35

3 Stacking **39**

3.1 Meta-Feature Set 1 . . . 41

3.2 Meta-Feature Set 2 . . . 47

3.3 Feature Pruning . . . 51

3.3.1 Correlation Coefficient . . . 51

3.3.2 Mutual Information (MII) . . . 54

3.4 Performance Prediction using a Decision Tree . . . 60

4 Conclusions **65**

References **68**

List of Figures

2.1 Left: 3-week-ahead forecasts of ILI levels (percentage of physician visits exhibiting ILI symptoms) generated by the 28 model ensemble for the 2016-2017 influenza season, plotted using dotted lines. The CDC FluView data is plotted using symbols. Middle: A comparison of DMA and raw ensemble results. The open circles plot the median DMA predictions, while the error bars are the tenth and ninetieth percentile predictions. The solid red line is the median prediction from the raw ensemble, while the dashed red line are the tenth and ninetieth percentiles. The filled symbols are the CDC FluView data. Right: The evolution of the weights of the top 3 models. All results are for the 2016-2017 influenza season, at the National level. 24

2.2 Left: Probability density functions (PDFs) of 3-week-ahead-forecasts at Week 13, 19, 25 and 32 of the 2016-2017 influenza season, as performed using DMA and the raw ensemble. The CDC FluView data ("observations") are plotted using the vertical line. Right: CRPS for the raw and DMA-ed ensemble predictions, computed over increasing durations. We see that DMA has smaller CRPSs, indicating better predictive accuracy. 24

2.3 Top left: Plot of the model ID of the best model for each week of the 2016-2017 influenza season, at the National level. No model is seen to provide consistently good 3-week-ahead forecasts for any sustained duration. Top right: Plot of the 3-week-ahead forecasts for the top 3 models, whose identities were determined at the end of the season. The CDC FluView data is plotted using filled symbols for comparison. The solid lines are the median forecasts and the dotted lines are the tenth and ninetieth percentiles of the forecast. We do not see a clear "winner" in the forecasts. Bottom left: $CRPS_t$ of all models, along with the CRPS of the best model and DMA. There is hardly any difference between them. Bottom right: The percentage difference in $CRPS_t$ between the best model's forecast and DMA's. The rank of the DMA forecast, as judged by $CRPS_t$ is plotted using the right-hand vertical axis. The DMA forecast is generally in the top 4. 26

2.4 Left: Ensemble of model predictions of 3-week-ahead ILI activity for the 2017-2018 influenza season, along with CDC's FluView data from syndromic surveillance at the National level. Middle: CRPS computed for the entire 2017-2018 influenza season, computed for the Nation as well as 10 HHS regions. DMA performs better. Right: The number of weeks that DMA and the raw ensemble provided better predictions, as gathered at the national level and the HHS regions. 27

2.5 Top left: 3-week-ahead forecasts of National ILI activity for the 2017-2018 influenza season and computed using BMA and DMA, along with the CDC FluView data plotted with symbols. The solid lines are the mean forecasts and the dashed lines, the $\pm 2\sigma$ bounds from the ensemble forecasts. The green vertical lines marks when the BMA forecast can be stably performed. Top right: $CRPS_t$ for BMA and DMA forecasts. Bottom left: The evolution of model weights when DMA is used. Bottom right: The evolution of BMA weights. 29

2.6 Comparison of ensemble predictions when the top three models are chose according to simple ranking (left column) and rank aggregating (right column). The filled symbols are CDC FluView data whereas the error bars plot the mean and $\pm 2\sigma$ forecasts. The top row contains the three-weak-ahead forecasts of ILI activity, the middle row contains the predictions of the peak ILI incidence and the bottom row predicts the week during which the peak ILI activity would be achieved. All forecasts are for the Nation as a whole, for the 2017-2018 influenza season. 31

2.7 Predictions of the peak ILI indicence week, for the 2017-2018 influenza season, using DMA, BMA and the raw ensemble. Note, however, the weights were computed using the three-week-ahead forecast as the QoI. Top left: The mean predictions at the National level. Top right: The $\pm 2\sigma$ bounds on the predictions. The horizontal line is the true value. Bottom: Results for HHS05. 34

2.8 Number of weeks where DMA, operated with the default value of α, looses to a different α value. The results are tallied over all HHS regions and weeks, for the 2017-2018 influenza season. 37

3.1 Regression coefficients on the mean 3-week ahead predictions from each flu model. . . . 42

3.2 Predicted ILI from stacking (blue circles) in each HHS region compared to the FluView ground truth (red line). 43

3.3 Average mean square error for each HHS region and each ensemble method. 44

3.4 Histogram of model coefficients from the stacking and DMA ensemble methods. 45

3.5 Gini coefficient of the model weights for each week for stacking and DMA performed on national data. 46

3.6 Spearman rank correlation of the model weights for each week for stacking and DMA performed on national data. 47

3.7 Regression coefficients on the mean and standard deviation of the 3-week ahead predictions from each flu model. 48

3.8 Average mean square error for each HHS region and each ensemble method. 49

3.9 Histogram of the ratio of the linear regression coefficient for the standard deviation of the model prediction to the coefficient for the mean prediction. Ratios larger than 1 indicate that the standard deviation is more important than the mean for the final prediction. 50

3.10 Pearson correlation coefficient between the mean 3-week ahead prediction of each flu model with the observations. 52

3.11 Training and testing error for the stacking model, averaged across all HHS regions and weeks, as a function of the number of features. 53

3.12 The number of features that resulted in minimum error for each week of the flu season. 54

3.13 Ranking of the features for each week of the flu season using the mutual information (MI) criteria. 56

3.14 Training and testing error for the stacking model, averaged across all HHS regions and weeks, as a function of the number of features. 57

3.15 The number of features that resulted in minimum error for each week of the flu season. 58

3.16 Difference in minimum error across number of features using the MI criteria and the correlation criteria for feature selection. Differences below 0 indicate that MI criteria performed better and differences above 0 indicate that correlation criteria performed better. 59

3.17 Decision tree for predicting whether stacking or the raw ensemble will have smaller error in a given week of the flu season. 61

3.18 Decision tree for predicting whether DMA or the raw ensemble will have smaller error in a given week of the flu season. 62

3.19 Decision tree for predicting whether stacking, DMA, or the raw ensemble will have smaller error in a given week of the flu season. 63

List of Tables

2.1 CRPS of the raw and DMA ensemble, averaged over the entire 2016-2017 influenza season. Results are provided at the national level as well as the 5 US Dept. of Health and Human Services regions. 25

2.2 CRPS of the raw and DMA ensemble, averaged over the entire 2016-2017 influenza season. Results are provided for 5 US Dept. of Health and Human Services regions. . 25

2.3 CRPS of the DMA and BMA ensemble, averaged over the entire 2017-2018 influenza season. Results are provided at the national level as well as the 5 US Dept. of Health and Human Services regions. 30

2.4 CRPS of the raw and DMA ensemble, averaged over the entire 2016-2017 influenza season. Results are provided for 5 US Dept. of Health and Human Services regions. . 30

2.5 Comparison of three-week-ahead forecast for 2017-2018 influenza season for all regions, performed using the two ranking-based methods for MA, DMA and the raw ensemble. We see that the ranking-based methods are better. The poor performance of DMA is partially due to all QoIs not being available for the late epoch of the outbreak, where DMA performs well. 32

2.6 CRPS of predictions of the peak epidemic week using weights computed for the three-week-ahead forecasts of ILI activity. Results are plotted for the raw, BMA and DMA ensemble, averaged over the entire 2016-2017 influenza season. Results are provided at the national level as well as the 5 US Dept. of Health and Human Services regions. 33

2.7 CRPS of predictions of the peak epidemic week using weights computed for the three-week-ahead forecasts of ILI activity. Results are plotted for the raw, BMA and DMA ensemble, averaged over the entire 2016-2017 influenza season. Results are provided for 5 US Dept. of Health and Human Services regions. 33

2.8 Comparison of forecasts of the held-out QoI in the 2017-2018 influenza season when models are ranked using rank-aggregation. We see keeping ranking based on all seven QoIs provide better forecasts 5 / 11 times, but the difference in the accuracy of the predictions follow no discernible trend. 36

2.9 Comparison of forecasts of the held-out QoI in the 2017-2018 influenza season when models are ranked using simple ranking. We see keeping ranking based on all seven QoIs provide better forecasts 5 / 11 times, but the difference in the accuracy of the predictions follow no discernible trend. 36

2.10 Proportion of weeks (as percentages), in the 2016-2017 influenza season, where α_{opt} yields a better $CRPS_t$ than the default value of 0.99. The performance of the raw ensemble is also provided for a comparison. 37

3.1 Flu models from the CDC forecasting challenge included in the stacking model. These models have predictions for all 11 HHS regions and all 25 weeks of the flu season. . . . 40

3.2 Average mean square error (MSE) for each ensemble method across all HHS regions and weeks. . . . 41

3.3 Average mean square error (MSE) for each ensemble method across all HHS regions and weeks. . . . 49

3.4 Confusion matrix for the stacking decision tree. . . . 64

3.5 Confusion matrix for the DMA decision tree. . . . 64

3.6 Confusion matrix for the multi-class decision tree. . . . 64

Executive Summary

In this study we investigate how an ensemble of disease models can be conditioned to observational data, in a bid to improve its predictive skill. The typical method to do so is model-averaging (MA; also called stacking), which weighs each model's predictions based on the model's plausibility, as conditioned on historical data. The challenge to date has been the sparsity of outbreak data - given the tremendous variability of outbreaks of the same disease and the fact that models are in constant flux, the data from multiple outbreaks cannot simply be concatenated together to set up a conventional MA problem. In this study, we develop three new MA methods that are robust to sparse data, and compare their performance to existing methods. We also investigate how we may construct a shrunk ensemble that can be conditioned by the sparse data.

We use the seasonal influenza outbreak in the US as the exemplar, due to the easy availability of influenza forecasts from CDC's FluSight ensemble (`https://predict.cdc.gov`) and CDC's syndromic surveillance FluView datasets (`https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html`). Previous work on MA (but using smaller, less diverse ensemble of influenza models) has shown that the primary benefit that conventional MA algorithms bestow is a consistency of ensemble forecasts. Further, a raw ensemble is found to be less predictive than a MA one.

Our first method is an adaptation of Dynamic Model Averaging (DMA). DMA sequentially assimilates CDC FluView data and updates the model weights. The second and third methods rely on using CDC FluView data to rank the models; the top three models' forecasts are simply combined in an equally weighted manner. All three methods are robust to sparse data. We compare them with Bayesian Model Averaging (BMA), the "standard" MA technique and an equally weighted ensemble ("raw" ensemble). BMA is a "batch" method that uses all the data available to compute model weights. If the ensemble is large and the data is small, BMA becomes unstable.

We find:

- The model-averaged predictions are more consistent than the forecasts from individual models.
- The DMA adaptation is better than the raw ensemble, indicating that it is a bona-fide MA method.
- The two ranking-based methods are almost equal in their predictive skill.
- The ranking-based MA algorithms are far more predictive than the others.
- When constructing the shrunk ensemble, it is better to choose models based on their diversity.
- The shrunk ensemble has similar predictive skill as DMA.
- The BMA algorithm performed the worst. This was entirely due to the EM algorithm becoming unstable for sparse data.

Looking forward, the study could be made more competitive by enhancing BMA with the *degenerate* EM algorithm which is has been shown to be somewhat robust to sparse data. We also discovered that the sole tuning parameter in DMA does not influence predictions much and and is not an substantial source of improvement in predictive skill.

Chapter 1

Introduction

Using an ensemble of models is a common tactic when attempting to forecast multi-phenomenological processes. In such cases, not only must the models contain representations of the individual phenomena, they also have to capture the interactions between the phenomena. If the process is such that it repeats e.g., weather or annual outbreaks of endemic diseases, the interactions between the phenomena may vary over time or space, creating a moving target for modelers attempting to construct predictive models. This is the case in epidemiology, where annual outbreaks are far from being identical. In such a case, it may be feasible to construct models that are predictive only under specific conditions. However, an ensemble of models could, in principle, capture all the relevant phenomena and their interactions, provided the ensemble's members are diverse (i.e., no duplicates) and the ensemble is large.

In some fortunate cases, the lack of predictive skill of a model may be due to uncertainty in the model inputs e.g., parameters, initial and boundary conditions. However, if we have a prior belief (in the form of a probability distribution) regarding what values the model inputs may assume, it may be possible to sample realizations of the uncertain model inputs, use them to seed multiple realizations of the same conceptual model and create a model ensemble. Note that the ensemble's members are structurally identical. Further, if observations of the process are available, it is possible to update the sampled realization so that the ensemble's predictions draw closer to the observed data. This is the principle behind modeling paradigms as diverse as ensemble Kalman filters [8] and random forests (Chp. 15 in Ref [16]).

In contrast to model input uncertainty, it is also possible that the lack of predictive accuracy of a model stems from the approximations inherent in it i.e., it is structurally deficient. In such a case, it is necessary that the ensemble contain models of very different structures. Further, the ensemble's members will have parameters and other inputs that have nothing in common with each other, and the modeling paradigms mentioned above are irrelevant when it comes to enhancing the predictive skill of an ensemble of *structurally diverse* models. This is the case with epidemiological models. The US Centers of Disease Control and Prevention (CDC) has assembled such an ensemble of influenza models. The models are used to generate forecasts of the annual influenza outbreak. The model forecasts do not, of course, agree, and the spread in forecasts is taken as measure of the ensemble's predictive uncertainty. This approach accords equal weight to each of the ensemble members. We refer to such an unweighted ensemble as a *raw ensemble.*

However, in most cases e.g., epidemiology of influenza, weather etc., one does have current information from the outbreak, and it should be possible, in principle, to identify the models in an ensemble that provide the best approximation to the data. Thus it should be possible to construct a *weighted* ensemble, where the weights could evolve in time (or space) to preserve the predictive skill of the ensemble. We call the process of *estimating* time-varying model weights from observational data *conditioning* an ensemble. In statistics, it is also called *model averaging* (MA). MA weighs the predictions of models, does not modify any model inputs and apart from the predictions, requires no other information e.g., gradients. MA is a class of "black-box" methods in the sense that the details of the models do not need to be exposed to the MA algorithm. These properties of MA allow it to be used with structurally dissimilar models in an ensemble.

There are many methods to perform MA (Chp. 8 in Ref. [16]). Technical challenges arise when the ensemble is large and the observational data available for MA is small. This is the case addressed in this study. One approach is to pick the best k models and simply use the mean forecast obtained from this *predictive cohort*. This requires one to select k *a priori*, which can be subjective. It also pose a challenge when a model provides multiple predictions (henceforth, quantities of interest or *QoI*s) e.g., in epidemiology, a model might provide a 1-week-ahead forecast of disease incidence, a forecast of the week when the outbreak will peak and what that peak might be. Each QoI might provide a different set of top-k models, making the construction of a predictive cohort impossible. The second approach assumes that the weights of models do not change rapidly in time and/or space. In such a case, one may use spatial/temporal correlations to provide a constraint on the estimation of weights (that the data fails to provide because of its sparsity). However, if the assumption of correlation is false, this approach can lead to erroneous results and it might even be advantageous to simply use the raw ensemble. The third approach computes the weights of models using a shrinkage method such as LASSO (Chp. 3 in Ref. [16]). In the process of doing so, it performs a data-driven pruning of the ensemble to recover a predictive cohort, without needing to specify a subjective k. This approach is called *stacking*.

In this study, we will explore four different ways of MA the CDC ensemble. One of them, Dynamic Model Averaging (DMA; [24]), estimates time-varying weights for all the models in the ensemble. The second, Bayesian Model Averaging (BMA; [23]), generates static weights, but is applied repeatedly with increasing amounts of data (as they become available during an outbreak), to generate a time-series of model weights. The other two address the problem of sparse data by choosing a predictive cohort of the three best models from the ensemble, and computing forecasts by according them equal weights. The performance of these four methods are compared against the predictive skill of the raw ensemble, with the figure of merit being whether they are better than the raw ensemble. This result should generalize across other model ensembles. We also compare the relative advantages of the four MA methods; however, this result will be specific to the CDC ensembles used for the influenza seasons of 2016-2017 and 2017-2018.

We also investigate whether we can shrink an ensemble of models into a smaller one that can be conditioned by the sparse data. We do so using elastic net regression and feature pruning, where the features are model predictions. We explore two opposing way of selecting features to be retained in the shrunk ensemble. The first one select features that correlate well with the observational data. The second ensures that the shrunk ensemble has a diversity of models. The approach that provides a weighted, but shrunk, ensemble with the higher predictive skill is deemed preferable.

This report is structured as follows. In the rest of this chapter, we will review the literature on MA and picking of a predictive cohort. We will also describe the observational data and CDC's model ensemble. In Chp. 2, we will develop the modification to DMA that converts it to a "black-box" method and compare its perfromance to the raw ensemble and other MA and predictive cohort methods. We also identify ensemble characteristics where a raw ensemble would be preferable to a DMA-ed one. In Chp. 3 we compare the efficacy of stacking versus the raw ensemble and DMA, and identify the feature pruning approach that yield a better shrunk ensemble. We conclude in Chp. 4.

1.1 Literature Review

1.1.1 Model-Averaging Methods

Model-averaging (MA) is a technique that is commonly used to improve the predictive skill of an ensemble of models which are structurally dissimilar i.e., they have unequal number of parameters or their parameters and inputs do not share a common interpretation. MA requires the use of observational data, against which the predictive skill of the ensemble is judged. Most commonly, one combines the predictions of the individual models linearly, in a weighted manner, in an attempt to approximate the observational data; the weights are estimated by minimizing a squared-error loss function (see Chp. 8.8 in Ref. [16]). The weights also sum to 1. It can be proven that in the limit of infinite observational data (or training data), the weights estimated will yield an ensemble prediction that is better than any of the individual models. If the ensemble contains a "true" model that can reproduce the observational data perfectly (modulo any measurement noise), MA will yield a weight of 1 for the true model, and zero for everything else.

However, observational data is limited and noisy, and ensembles can quite easily contain highly parameterized/complex models that will overfit the observational data. Thus, in practice, the estimation of weights is performed using cross-validation. A point estimate of the weights is obtained, which is then used to combine the forecasts of the individual models into an ensemble forecast. This method is called *stacking*. Stacking does not yield any uncertainty bounds for the ensemble forecasts, even if the individual models' forecasts have them.

Bayesian Model Averaging (BMA) is a method that allows a probabilistic ensemble forecast i.e., with forecasting uncertainty bounds. If the individual models in an ensemble provide probabilistic forecasts themselves, then BMA can be treated as estimating weights for a mixture of distributions (and, perhaps, the models' parameters too, if there is sufficient data [18]). The methods for estimating the weights can be involved if we assume that the model forecasts are arbitrary distributions. If the models provide point forecasts only, one may consider them to be the means of a mixture of Gaussians whose variance (or covariance) is unknown. The MA averaging problem then reduces to the estimation of the mixture weights and the unknown variance so that the resultant mixture maximizes the likelihood (or expectation) of the observational data. This method is described in Ref. [23], where the estimation of weights and variance is performed using Expectation-Maximization (EM). Couched in the language of stacking, the predictions by the individual models in the ensemble are "features" which are combined, in a weighted manner, to yield the ensemble forecast. Further, the individual model forecasts (features) are subjected to additive and multiplicative bias correction before being stacked. However, unlike stacking, this method also estimates a forecast uncertainty. This is one of the methods that we will employ in our study. This method could have two shortcomings. First, it ignores the forecasting uncertainties of the models in the ensemble. If this uncertainty is available (it is in our case; see Sec. 1.2), the ensemble forecast will most certainly be underestimated. Secondly, the method assumes that there is sufficient data to estimate all the weights. If the data is sparse (or the ensemble contains a large number of models), then the weights estimated by EM could be very inaccurate; EM does not allow one to provide a prior belief of what the weights could be.

Dynamic Model Averaging (DMA) is a method that allows MA when the models provide their forecasts as a Gaussian distribution. The method is described in Ref. [24], within the context of forecasting with an ensemble of three models. The models provide forecasts as Gaussian distributions. The ensemble forecast is assembled as a weight mixture of the individual models' Gaussian forecasts. The mean of the Gaussian is a *linear* function of certain time-varying predictors, which are continuously observed. The coefficients of this linear function, which are estimated as part of the algorithm, are also allowed to evolve in time. The quantity being predicted is also a function of time, and it is assumed that at different points in time, certain models may be more predictive than others i.e., the weights are allowed to evolve in time also. The three models in the ensemble are nested i.e., the simpler models can be recovered from the most complex model by ignoring the some of the model inputs/predictors. The estimation procedure is extremely high-dimension in the sense that both model weights *and* the individual models' parameters (in the linear relation for the Gaussian's mean) have to be estimated. The method is formulated as one of sequential data assimilation using online, continuously observed data y_t (i.e., the response) conditional on time-varying (observed predictors) x_t. Thus, the method is a filter. It specifies how to obtain a forecasts of a model's parameters θ'_t, given finalized parameters from a previous timestep $\hat{\theta}_{t-1}, y_{t-1}$, and how θ'_t may be updated to $\hat{\theta}_t$ using y_t. There are also expressions for doing the same for the model's weight w_t. The method is initiated with model parameters that are drawn from a prior distribution and equal weights. Being a sequential data assimilation method, the weights do not fail catastrophically in the absence of observational data y_t; the weights and model parameters remain close to their starting values (which may be very inaccurate). The method has two shortcomings: (1) it is, as formulated, limited to an ensemble of linear models whose parameters it attempts to estimate and (2) it cannot be adapted, in a straightforward manner, to be a "black-box" method, applied to an ensemble of nonlinear models whose model parameters are not accessible and are not meant to be adjusted using the observed data. In Chp. 2, we will modify this algorithm into a "black-box" one and assess whether it can weigh an ensemble to improve its predictive skill i.e., whether the modified DMA is better than the raw ensemble.

Ranking-based ensemble methods are an alternative method to condition an ensemble. In essence, the models in the ensemble are ranked based on their predictions, and the predictions from the k best models are simply averaged (or combined into an equivalent Gaussian distribution) to compute the ensemble forecast. The method is straightforward if the observed data or the model predictions are limited to one variable. If the data and model predictions span multiple variables / observables / QoIs, then one obtains a separate ranking of models using each QoI. For convenience, it is best to "merge" the separate ranking lists into a single one, from which we can choose the k best models. Such a way of merging ranking lists is described in Ref. [21], where the final list has a model ranking that minimizes the movement/changes of model ranks from the ranking lists developed using individual QoIs. We will check how ranking-based methods, BMA and DMA compare with each other and against the raw ensemble.

1.1.2 Model-Averaging with Disease Models

Over the last five years, there has been much interest in creating ensembles of disease models and MA them to improve the performance characteristics vis-á-vis the individual models in the ensemble. If the models provide deterministic forecasts, and training data for MA exists, one can combine the model forecasts using weights that can be directly computed from the Bayesian Information Criterion (BIC) of the individual forecasts (Chp. 8, Ref. [16]). In Ref. [27] a similar method was used to stack 6 forecasters of Zika incidence in Colombia. The models were fitted to cumulative case counts. The model weights were computed using the Akaike Information Criterion (AIC) of the resulting models. The individual models' forecast mean and variance were MA to provide the corresponding quantities for the ensemble. The ensemble was shown to predict the turning point and final size similar to the best model; however, the best model was not known *a priori*. Further, a different model was found to be the most predictive for the four cities where the outbreak was studied. An even simpler approach was adopted for dengue forecasting in Ref. [3]. An ensemble of 300 models (falling into three distinct types - method of analogues, Holt-Winters smoothing and SARIMA) was created and their fitting errors computed. The worst model was give one "vote", and the rest assigned "votes" in proportion to their training errors compared to the worst model. The "votes" were used to construct a probability density function (PDF) of the forecasts generated by the ensemble. It was found that the ensemble was not more predictive than the individual models but rather more consistent in predicting different QoIs, as well as from outbreak-to-outbreak. In contrast, models could fail unpredictably.

BMA has been used widely in combining the forecasts produced by an ensemble of disease models [29, 30]. A common approach is to use a diverse set of mechanistic disease models e.g., SEIR, SEIS etc., and data assimilation methods e.g., Ensemble Kalman filters, Ensemble Adjustment Kalman filters etc., and by mixing and matching them, create a large ensemble of forecasters. Each forecast is provided as a Gaussian PDF, and the ensemble forecast is a weighted mixture of Gaussians, computed using the BMA method described above (i.e., EM). The ensemble so obtained (where the ensemble's weights adapted with time and location) was compared against the raw ensemble, one where the weights for the individual forecasters were different but were held constant over time and locations. It was found the the ensemble with adaptive weights performed the best and the raw ensemble the worst, on average. Further, the ensemble did not necessarily perform better than the best model; rather it performed much more consistently across location, time and the QoIs.

In Ref. [25] the authors explored an ensemble of 3 statistical ILI (influenza-like illness) models but six different MA methods. This included equal weights, variable weights, and what they termed "feature-dependent" weights. In the latter case, weights for the individual models were deemed to be functions of time, model forecasting uncertainty and the level of ILI incidence. The functional dependence was cast as a regression tree which was estimated using gradient tree boosting. They found that the weights computed using features followed trends of individual models' forecasting accuracy. In addition, in their ensemble, the three models had comparable accuracies. The main outcome of the MA exercise was increased consistency in ensemble predictions rather than any significant improvement forecasting accuracy. An identical outcome was observed in Ref. [2], which used the same data, but a different of statistical models. The weights were computed with EM, using three different measures of forecasting accuracy.

The literature reviewed above concerned forecasting the evolution of ILI or dengue outbreaks. MA has also be used to combine an ensemble of anomaly detectors, as used in syndromic surveillance. In Ref. [20], the authors combined three time-series models for forecasting syndromes. The forecasts were compared with data and large discrepancies were flagged as anomalies. The ensemble method consisted of weighing the model forecasts so that the false positive rate was minimized and true positive rate enhanced. The weights were estimated via mixed integer programming. In contrast to the literature reviewed above, where MA essentially improved consistency but not the forecasting accuracy, the "ensembled" anomaly detector reduced false postive rates.

1.2 Data and Model Ensemble

The observational data for this study is obtained from CDC's influenza surveillance program. The CDC has a network of sentinel physicians who report the fraction of their outpatients exhibiting symptoms of influenza-like illnesses (ILI). Note that only a fraction of ILI patients are infected with influenza. It takes about two weeks to collect this data and display it on the influenza surveillance program website [10, 9]. This provides a sampling of ILI and its dynamics for the entire US. The data is available as a weekly time-series, lagged by two weeks, for the entire nation and individually for the 10 DHHS (Department of Health and Human Services) regions. New data is released every week. At any given moment of time, the time-series values for the last three weeks are rather uncertain and are updated when new data is released the subsequent week. In this study, we use the ILI data (also referred to as FluView data in this study) for the influenza seasons spanning 2016-2017 and 2017-2018. The new "surveillance" year starts on the 40th week of each year and continues for 52 weeks. The data is used after the annual outbreak ends and all the updates to the observational data are complete. While the surveillance year starts on Week 40, the influenza season starts when the *weighted* ILI time-series value surpasses a threshold for three consecutive weeks. The threshold varies for DHHS regions; see Refs. [13, 14] for historical thresholds and how they are calculated.

CDC has assembled a collection of influenza forecasting models and has conducted annual competitions to predict the influenza outbreak. The first such competition, which involved 16 teams, was for the 2013-2014 influenza season and is documented in Ref. [1]. Thereafter, the competition has been held every year; the ensemble of models has grown. An account of the predictive skill of the ensemble, spanning mid-2013 to mid-2017 can be found in Ref. [26]; it also contains citations to papers and reports that describe the models. The models predict seven Quantities of Interest (QoI): the 1-, 2-, 3- and 4-week-ahead forecasts of ILI incidence, the week of the influenza season onset, the peak incidence of the outbreak and the week when the outbreak peaks. These QoIs are predicted as probability densities (strictly, as binned probability masses), thus capturing the uncertainties in the forecasts. These forecasts are produced at the National level, as well as the 10 DHHS regions. These forecasts have been archived for the influenza seasons of 2015-2016, 2016-2017, 2017-2018 and during the time of writing the report, the incomplete 2018-2019 season and can be found in a GitHub repository [12]. Information on CDC's disease forecasting challenges can be found in Ref. [11]. The models used in the forecasts are under continuous development and consequently cannot be compared across seasons. Further, the composition of the ensemble changes every year and it has steadily become larger. We will use the ensembles from the 2016-2017 and 2017-2018 seasons, as they are rather large and have a sufficient diversity of models in them. The ensemble for 2016-2017 has 29 models and the 2017-2018 ensemble has 35 models.

Chapter 2

Dynamic Model Averaging (DMA)

In this chapter, we develop a method to model average an ensemble of K probabilistic forecasters. Each forecast is made in the form of $\mathcal{N}(\mu_t, \nu_t)^{(k)}, t = 1 \ldots T, k = 1 \ldots K$. The MA forecast is cast as $\mathcal{N}(\mu_t^*, \nu_t^*)$, and the formulation below develops expressions for (μ_t^*, ν_t^*) in terms of $(\mu_t, \nu_t)^{(k)}$ and $\{w_{t,k}\}$. Here $\{w_{t,k}\}$ is the probability that at model k represents epidemiological dynamics most closely at time t. The formulation closely follows DMA as described in Ref. [24] and adapts it to address a collection on black-box models whose parameters do not have to be estimated and which are *not* linear models.

The evolution of $(\mu_t, \nu_t)^{(k)}$ over time is performed by the forecasters themselves, often by using observational data $Y_t = \{y_l\}, l = 1 \ldots t$. The evolution of (μ_t^*, ν_t^*) requires one to evolve $\{w_{t,k}\}$ using Y_t, which is performed by the method described below. In essence, the method updates $\{w_{t-1,k}\}$ to $\{w_{t,k}\}$ using y_t. The weights at initiation $t = 0$ are given by $\{w_{t=0,k}\} = 1/K$ i.e., all models are equally probable. If the data is informative, and sufficient data has been assimilated, $\{w_{t,k}\}$ is quite different from $1/K$. This incremental updating of $\{w_{t,k}\}$ ensures that the algorithm delivers a set of weights for the models every time. This can be a challenge when the ensemble is large and the data Y_t limited.

In Sec. 2.1 we develop the new MA method and an "error measure" that quantifies the degree of disagreement between a probabilistic forecast and an observation y_t. The algorithm's performance is compared against individual models and the raw ensemble in Sec. 2.2. In Sec. 2.3, we compare our method (which we will loosely call DMA) with other MA methods. Finally, in Sec. 2.4, we investigate the limitations of the MA methods discussed here.

2.1 Formulation

Assumptions:

1. We are given K forecasters, that produce their one-step-ahead forecasts as a normal distribution $\mathcal{N}(\mu_t, \nu_t)^{(k)}$, where μ_t and ν_t are the mean and variance of the forecast at time t.

2. At time $t = 0$ there are no observations and our prior belief regarding the applicability of the K models is $1/K$.

3. Observations up to and including time t are given by $Y_t = \{y_l\}, l = 1 \ldots t$, Here y_l is obtained from CDC's Flu View dataset of season influenza incidence.

We desire a way of combining $(\mu_t, \nu_t)^{(k)}$ to produce $(\mu_t, \nu_t)^*$ using Y_t to compute and evolve $\{w_{t,k}\}$.

Let $\pi_{t-1|t-1,k}$ be the probability that model k represents the epidemiological dynamics given observations Y_{t-1} i.e.,

$$\pi_{t-1|t-1,k} = P(L_{t-1} = k \mid Y_{t-1}).$$

Here $P(L_{t-1} = k \mid Y_{t-1})$ refers to the probability that at $t-1$ the most accurate model was model k, given observations of epidemiological dynamics Y_{t-1}. Let q_{kl} be the probability that the best model for epidemiological dynamics, at time t, is model l, given that it was model k at time $t-1$. We can then predict the probability of the models as such

$$\pi_{t|t-1,k} = \sum_{l=1}^{K} \pi_{t-1|t-1,l} \times q_{kl} = \frac{\pi^{\alpha}_{t-1|t-1,k} + c}{\sum_l^K \pi_{t-1|t-1,l} + c}, \qquad \text{where} \qquad \alpha = 0.99, c = 0.001/K. \tag{2.1}$$

Here we have used the idea of a “forgetting” process [24] i.e., the posterior probability distribution over K models can be broadened and used as the prior for predicting the next step. This modeling step thus frees us from modeling q_{kl}. The predicted model probabilities have to be updated with observation y_t to arrive at the final version of the model probabilities $\pi_{t|t,k}$.

We compute the likelihood of producing the observation y_t using model k and update the predicted model probability i.e.,

$$L_t^{(k)} = \frac{1}{\sqrt{2\pi\nu_t^{(k)}}} \exp\left(-\frac{(y_t - \mu_t^{(k)})^2}{2\nu_t^{(k)}}\right), \qquad u_{t,k} = \pi_{t|t-1,l} L_t^{(k)} / \nu_{t-1}^{(k)} \qquad \text{and} \qquad \pi_{t|t,k} = \frac{u_{t,k}}{\sum_l u_{t,l}}. \tag{2.2}$$

Note that we penalize $u_{t,k}$ by the model's predictive variance. This makes $u_{t,k}$ dimensional. Note that $\pi_{t|t,k}$ is non-dimensional. The ensemble forecasts based on Y_{t-1} are given by

$$\begin{aligned} \mu_t^* &= \pi_{t|t-1,k}\mu_t^{(k)} \\ \nu_t^* &= \mathrm{Var}(y_t \mid Y_{t-1}) = \sum_k \left(\nu_t^{(k)} + (\mu_t^{(k)})^2\right) \pi_{t|t-1,k} - (\mu_t^*)^2 \\ &= \sum_k \nu_t^{(k)} \pi_{t|t-1,k} + \sum_k (\mu_t^{(k)})^2 \pi_{t|t-1,k} - \left(\sum_k \mu_t^{(k)} \pi_{t|t-1,k}\right)^2. \end{aligned} \tag{2.3}$$

Any forecast in the form of $\mathcal{N}(\mu_t, \nu_t)$ can be compared to an observation y_t using the continuous rank probability score (CRPS, Ref. [17]). Let $F_t(y)$ be the cumulative distribution function corresponding to the prediction $\mathcal{N}(\mu_t, \nu_t)$. Then CRPS at time t is given by

$$CRPS_t = \int_{-\infty}^{\infty} \left(F_t(y) - \mathcal{H}(y - y_t)\right)^2 dy, \tag{2.4}$$

where $\mathcal{H}(x)$ is the Heaviside function. We will refer to the mean of $CRPS_t$ over $t = 1 \ldots T$ as CRPS.

2.2 Results

We investigate the behavior of DMA as it MA the forecasts from the 2016-2017 ensemble at the National level. In Fig. 2.1 (left) we plot the 3-week-ahead forecasts from the ensemble of 28 models (which are plotted with dotted lines). The CDC FluView data is plotted with symbols. All results are at the National level. It is clear that there are many ensemble members which do not provide good forecasts. Further, the forecasts start only about 7 weeks into the influenza season. In Fig. 2.1 (middle) we plot the 3-week-ahead predictions obtained by MA the ensemble using DMA (circles and error bars) and the raw ensemble. We plot the median prediction, as well as the tenth and ninetieth percentiles. It is clear that the uncertainty bounds for the raw ensemble are wider than those computed using DMA. Due to the wider uncertainty bounds, the raw ensemble brackets more of the CDC data, but that does not make it a more accurate forecast. Further, after 25 weeks of observations, DMA provides extremely accurate and specific forecasts whereas the raw ensemble is far from doing so. In Fig. 2.1 (right) we plot the time-evolution of the weights of the models. We see that ultimately, the best model ends up getting a weight of 1 which persists till the end of the outbreak.

In Fig. 2.2 (left) we plot the probabilistic predictions produced by the MA and raw ensembles at four different times in the 2016-2017 outbreak. The MA predictions are Gaussians. We see that the DMA predictions' probability densities are narrower. Early in the outbreak DMA strives to match the mode with the observations. Halfway through the outbreak, the modes for both the DMA and raw ensemble match the CDC FluView data, though DMA provides forecasts with tighter bounds vis-á-vis the raw ensemble. In Fig. 2.2 (right) we plot the CRPS (averaged over increasing durations) for the raw ensemble and DMA forecasts. The vertical green lines, which correspond to the times plotted on the left, show that the CRPS of the DMA predictions are smaller than that of the raw ensemble.

In Fig. 2.3 (top left) we plot the model identity number (ranging from 1-28) of the best model (as determined by the $CRPS_t$) for each week. We see that a given model does not seem to perform well for more than 4 weeks at a stretch i.e., it would be unwise to attempt to pick a "best" model. In Fig. 2.3 (top right) we plot the predictions for the top three models, as evaluated at the end of the 2016-2017 season. The CDC FluView data is also plotted as is the DMA-ed predictions of 3-week-forecasts. The pattern is complex and there does not seem to be any forecast that is obviously better. Averaged over the entire season, the CRPS of the top three models are 0.366, 0.3733, and 0.376. The corresponding value for the DMA forecast is 0.3734. Clearly the difference is small, but *MA did not yield a better forecast.* In Fig. 2.3 (bottom left) we plot the PDF of $CRPS_t$ of all the models for the entire 2016-2017 season, and we see that there is substantial variability in the accuracy of forecasts; indeed, given the scatter seen in Fig. 2.1 (left) this is expected. We also plot the CRPS of the best model and the DMA forecast. We see that the difference between the best model and the DMA forecast, averaged over the entire 2016-2017 season is small, though DMA is less predictive. In Fig. 2.3 (bottom right) we plot the difference (as a percentage) between the $CRPS_t$ of the best model (which changes from week to week) and the DMA forecast as a solid line. It is seen to be below 10%. On the right hand vertical axis, we plot the rank of the DMA forecast in comparison with all the models. We see that the DMA forecast is usually ranked third or fourth, and its predictive accuracy, vis-á-vis the best model, is generally within 10%.

In Tables 2.1 and 2.2 we extend the comparison performed in Fig. 2.2 to all the DHHS (Dept. of Health and Human Services) regions. We tabulate the CRPS computed for the entire 2016-

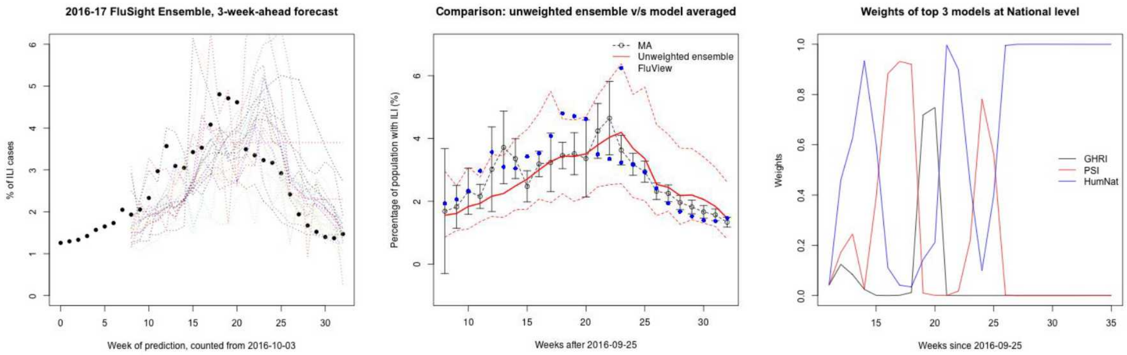


Figure 2.1: Left: 3-week-ahead forecasts of ILI levels (percentage of physician visits exhibiting ILI symptoms) generated by the 28 model ensemble for the 2016-2017 influenza season, plotted using dotted lines. The CDC FluView data is plotted using symbols. Middle: A comparison of DMA and raw ensemble results. The open circles plot the median DMA predictions, while the error bars are the tenth and ninetieth percentile predictions. The solid red line is the median prediction from the raw ensemble, while the dashed red line are the tenth and ninetieth percentiles. The filled symbols are the CDC FluView data. Right: The evolution of the weights of the top 3 models. All results are for the 2016-2017 influenza season, at the National level.

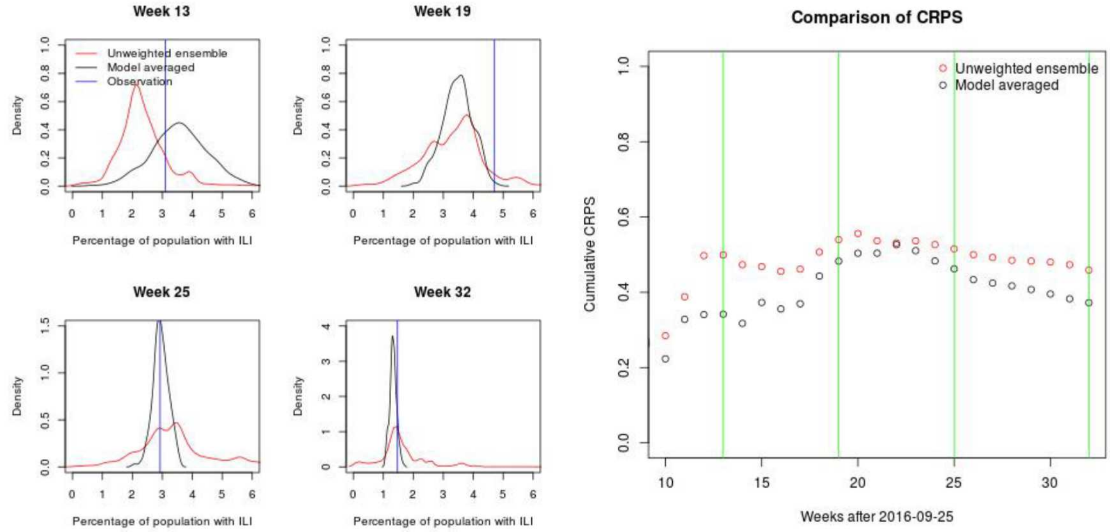


Figure 2.2: Left: Probability density functions (PDFs) of 3-week-ahead-forecasts at Week 13, 19, 25 and 32 of the 2016-2017 influenza season, as performed using DMA and the raw ensemble. The CDC FluView data ("observations") are plotted using the vertical line. Right: CRPS for the raw and DMA-ed ensemble predictions, computed over increasing durations. We see that DMA has smaller CRPSs, indicating better predictive accuracy.

Table 2.1: CRPS of the raw and DMA ensemble, averaged over the entire 2016-2017 influenza season. Results are provided at the national level as well as the 5 US Dept. of Health and Human Services regions.

Method	National	HHS01	HHS02	HHS03	HHS04	HHS05
Raw ensemble	0.46	0.30	0.60	0.60	0.74	0.56
DMA ensemble	0.37	0.27	0.57	0.64	0.77	0.51

Table 2.2: CRPS of the raw and DMA ensemble, averaged over the entire 2016-2017 influenza season. Results are provided for 5 US Dept. of Health and Human Services regions.

Method	HHS06	HHS07	HHS08	HHS09	HHS10
Raw ensemble	0.80	0.75	0.40	0.45	0.48
DMA ensemble	0.63	0.76	0.35	0.34	0.46

2017 influenza season using the raw ensemble and the DMA-ed one. We see that apart from HHS Regions 3, 4, and 8, DMA provides more accurate predictions; further, in the HHS Regions where DMA performs worse, the margin by which is under-performs is quite narrow. Thus we see that DMA tends to perform better than the raw ensemble. However, this does not imply that other MA techniques could not better than DMA in their predictive skill.

Comparing this behavior with the rapid change in the identity of the best model (top left), we come to the following conclusions:

1. The DMA forecast is never the most accurate forecast, but it is never very far from the best model.
2. The identity of the best model is unknown beforehand, and in any case, it changes from week to week, making reliance on *a* model unadvisable. In contrast, DMA provides a consistent way of providing a *"good enough"* forecast.
3. The DMA-ed ensemble tends to provide more accurate predictions than the raw ensemble in most of the HHS regions. In the few regions where it under-performs the raw ensemble, the margin is rather minor.

This is very similar to the conclusions that have been reached in literature (see Sec. 1.1). MA improves consistency in forecasts rather than any enhanced accuracy (over a model). Thus selecting a model / forecaster for further use becomes unnecessary, and we allow the observations from an outbreak to stack the models in a data-driven manner. The model-averaged predictions are generally more accurate than the raw ensemble.

2.3 Comparison with other MA Methods

In this section, we compare the performance of DMA against BMA as well as ranking-based method for MA. We use the CDC ensemble of influenza models for the 2017-2018. As a first check, in Fig. 2.4 (left), we plot the 3-week-ahead predictions of ILI activity from 35 models (in dots) as well as the

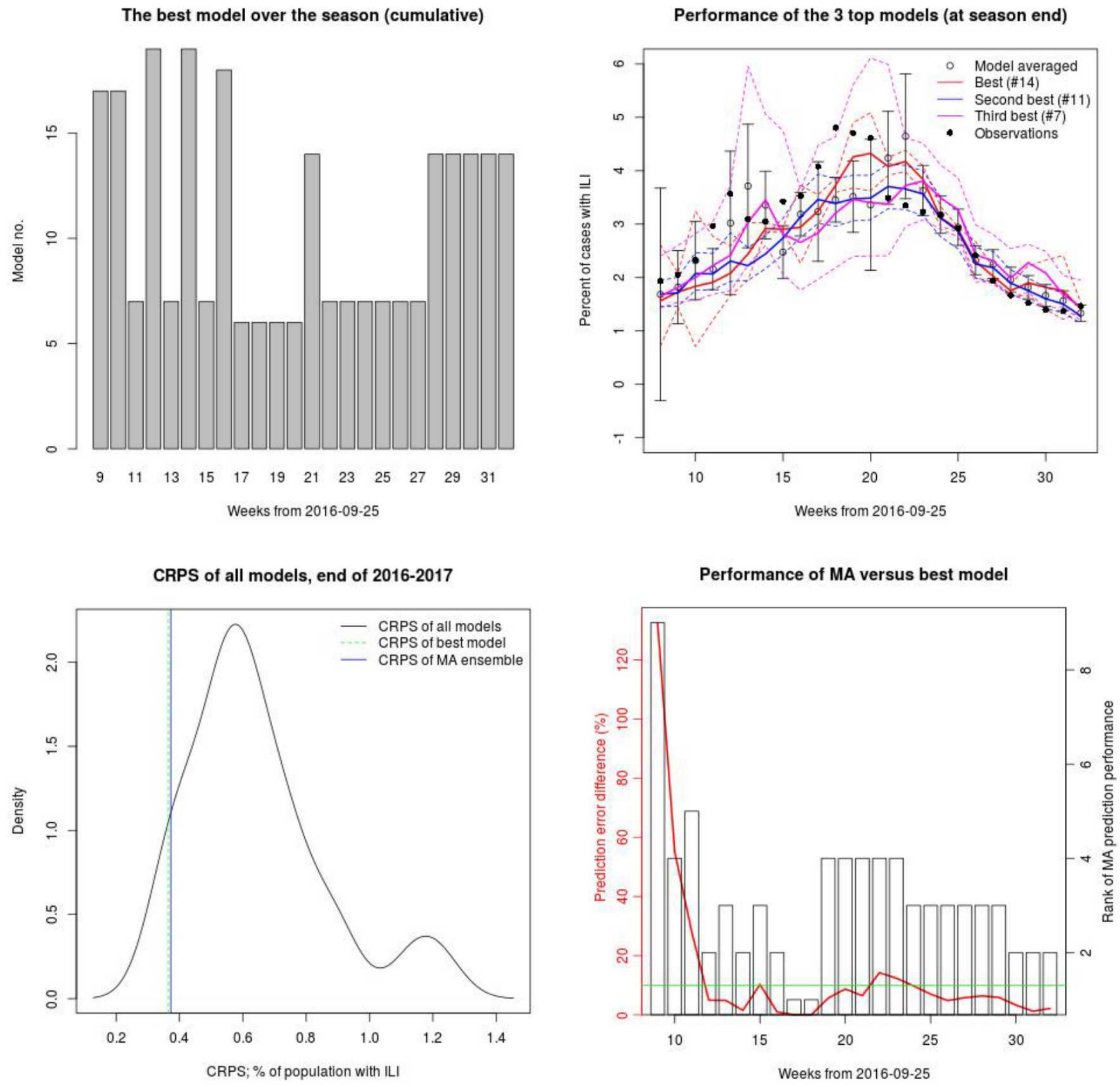


Figure 2.3: Top left: Plot of the model ID of the best model for each week of the 2016-2017 influenza season, at the National level. No model is seen to provide consistently good 3-week-ahead forecasts for any sustained duration. Top right: Plot of the 3-week-ahead forecasts for the top 3 models, whose identities were determined at the end of the season. The CDC FluView data is plotted using filled symbols for comparison. The solid lines are the median forecasts and the dotted lines are the tenth and ninetieth percentiles of the forecast. We do not see a clear "winner" in the forecasts. Bottom left: $CRPS_t$ of all models, along with the CRPS of the best model and DMA. There is hardly any difference between them. Bottom right: The percentage difference in $CRPS_t$ between the best model's forecast and DMA's. The rank of the DMA forecast, as judged by $CRPS_t$ is plotted using the right-hand vertical axis. The DMA forecast is generally in the top 4.

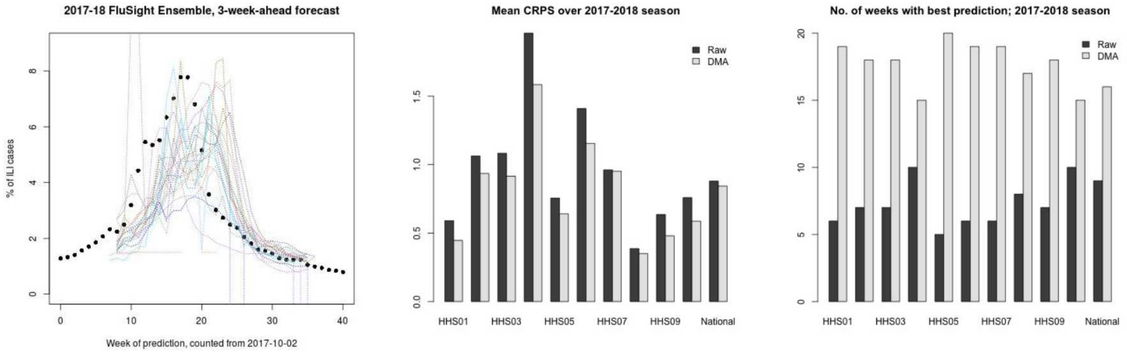


Figure 2.4: Left: Ensemble of model predictions of 3-week-ahead ILI activity for the 2017-2018 influenza season, along with CDC's FluView data from syndromic surveillance at the National level. Middle: CRPS computed for the entire 2017-2018 influenza season, computed for the Nation as well as 10 HHS regions. DMA performs better. Right: The number of weeks that DMA and the raw ensemble provided better predictions, as gathered at the national level and the HHS regions.

CDC syndromic surveillance data (in symbols). The season was a severe one and the predictions are all higher than those for the 2016-2017 season (Fig. 2.1). In Fig. 2.4 (middle), we compute the CRPS for the raw and DMA-ed ensemble, for all HHS regions. Clearly, in every case, the raw ensemble has a larger CRPS indicating that DMA provides a better prediction. In Fig. 2.4 (right), we plot the number of weeks DMA and raw ensemble provided the better prediction, over all HHS regions and the nation. Clearly, again, DMA outperforms the raw ensemble. However, comparing the middle and right figures, we see that while DMA might provide a better prediction on a large majority of the weeks, the small difference in the CRPS indicates that the difference in predictive skill is not much. However, in general, DMA performs much the same as for the 2016-2017 ensemble, providing a better prediction than the raw ensemble, not quite performing as well as the best model, but far more consistently.

2.3.1 Bayesian Model Averaging

BMA, as described in Ref. [23], is a method that can be used to stack an ensemble of deterministic models, and yet produced an ensemble forecast that is probabilistic. Consider K models that produce a forecast of ILI incidence at time $t, k = 1 \dots K$. Then the probability density of the ensemble forecast can be given as

$$p(y) = \sum_k p(y_t \mid M_k) p(M_k \mid y_t) = \sum_k w_k p(y_t \mid M_k), \tag{2.5}$$

where $w_k = p(M_k \mid y_t)$ is the posterior probability of model M_k being correct given observational data y_t. We also assume $\sum_k w_k = 1, w_k \geq 0$, making w_k weights, and $p(y_t \mid M_k)$ is the likelihood that the observation was generated by model M_k. Consider that the models are deterministic and produce their forecasts as $\mu_t^{(k)}$. Assume that there is a conditional probability that relates an ensemble forecast with a model one i.e., $g(y \mid \mu_t^{(k)})$. Then the ensemble forecast can be written as

$$p(y \mid \mu_t^{(1)}, \mu_t^{(2)}, \dots \mu_t^{(K)}) = \sum_k w_k g(y \mid \mu_t^{(k)}).$$

We will model $g(y \mid \mu_t^{(k)})$ as a normal distribution and rewrite the expression for the ensemble forecasts as

$$y \mid \mu_t^{(k)} \sim \mathcal{N}(a_k + b_k \mu_t^{(k)}, \nu), \tag{2.6}$$

where (a_k, b_k) are additive and multiplicative biases in model k's predictions and can be estimated from data by simple regression. The ensemble prediction is modeled as a Gaussian and given by

$$\begin{aligned} E[y \mid \mu_t^{(1)}, \mu_t^{(2)}, \ldots \mu_t^{(K)}] &= \sum_k (a_k + b_k \mu_t^{(k)}) \\ \mathrm{Var}(y \mid \mu_t^{(1)}, \mu_t^{(2)}, \ldots \mu_t^{(K)}) &= \sum_k w_k \left((a_k + b_k \mu_t^{(k)}) - \sum_i w_i (a_i + b_i \mu_t^{(i)})^2 \right)^2 + \nu. \end{aligned} \tag{2.7}$$

The values of (w_k, ν) are estimated via expectation-maximization, as described in Ref. [23]. Assuming that we are given a set of observations up to time t, Y_t, and the forecasts are independent, we can write a log-likelihood

$$\mathcal{L}(Y_t \mid w_1, w_2, \ldots w_K, \nu) = \sum_k \log \left(\sum_k w_k g_k (y_t \mid \nu^{(k)}) \right). \tag{2.8}$$

Note that this method is a "batch" rather than a "dynamic" method i.e., conditional on Y_t we estimate (w_k, ν) that are not functions of time. However, as Y_t changes with time, we get solve the BMA problem repeatedly and obtain a set of BMA parameters (w_k, ν) which *do* evolve in time.

In Fig. 2.5 (top left) we plot the 3-week-ahead forecasts of (National level) ILI activity as computed using DMA and BMA. We see that BMA forecasts being later than DMA. This is a consequence of the EM algorithm used to estimate (w_k, ν). The ensemble for 2017-2018 has 35 models, and the problem is under-constrained for a longer period of time. Even when the EM algorithm yields a result, it is not very good, with very large ν. However, as the season progresses, and we accumulate CDC FluView data, BMA forecasts improve, but never quite approach DMA's quality. CDC's FluView data is plotted using filled symbols. In Fig. 2.5 (top right) we plot the $CRPS_t$ for both the MA methods and DMA is seen to be better. In Table 2.3 and 2.4 we compare the CRPS of the DMA-ed and BMA-ed ensemble for all the HHS regions and see that DMA performs much better. This is not due to any particular strength of DMA, except its robustness to sparse observational data; the weights of the models, in DMA, incrementally evolve away from uniform weighting as data is incrementally assimilated. In Fig. 2.5 (bottom left and right) we plot the evolution of weights of a few models over time. We see that the evolutions are very different, with DMA providing smooth evolutions and BMA, which is performed anew every time, showing sharp changes. Identical behavior is seen in tests performed for other HHS regions.

2.3.2 Ranking-based Model Averaging

In the methodological developments above, we have concentrated on the three-week-ahead forecast as the QoI for ensemble predictions. However, as mentioned in Sec. 1.2, the CDC ensemble of models provide 7 QoIs. Per the literature review in Sec. 1.1, we see that it is preferable to MA for each individual QoI. In this section, we question this assumption. We will choose the top three models, that, on average, are most predictive across all 7 QoIs, and use them for forecasting. Each

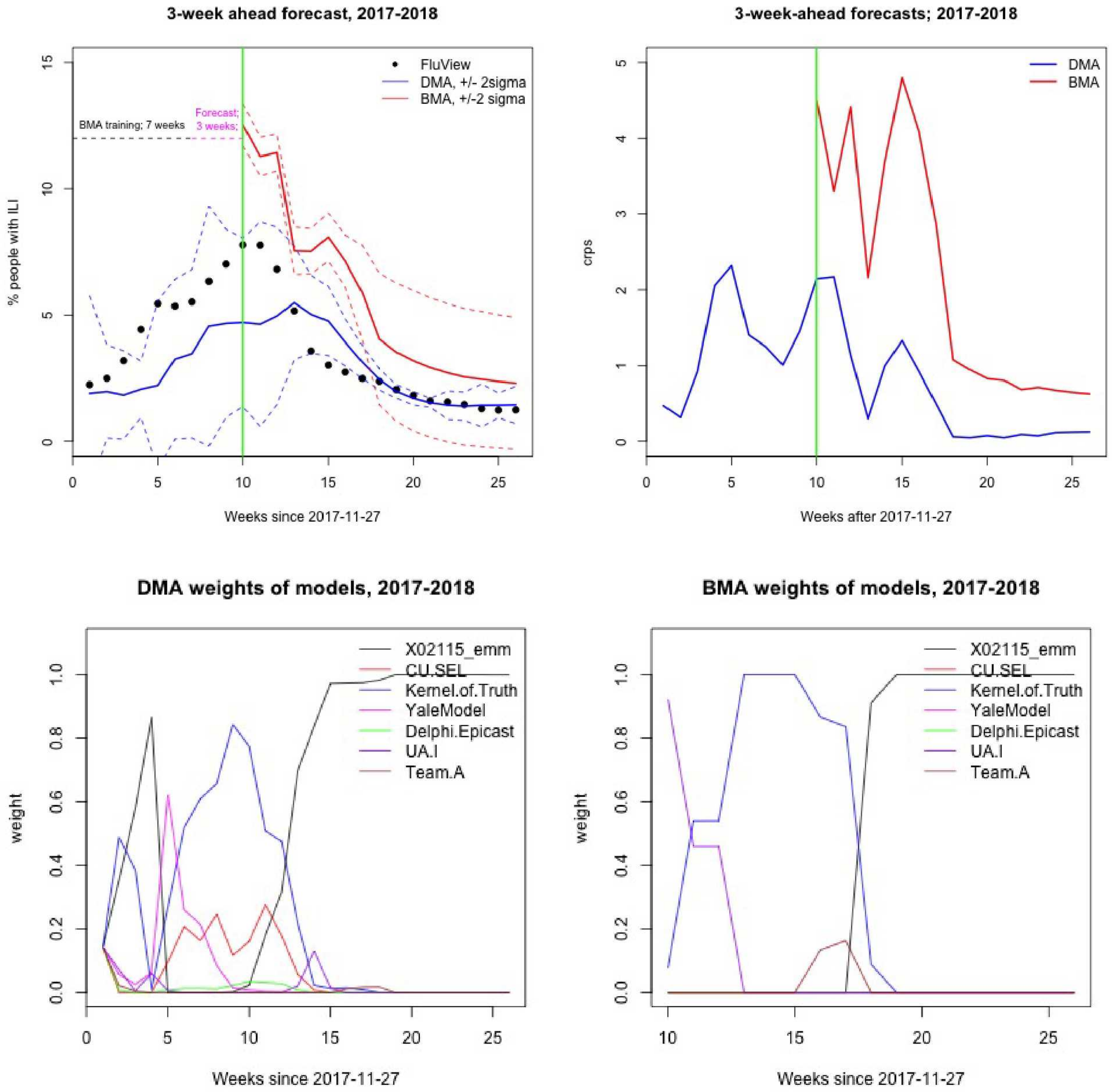


Figure 2.5: Top left: 3-week-ahead forecasts of National ILI activity for the 2017-2018 influenza season and computed using BMA and DMA, along with the CDC FluView data plotted with symbols. The solid lines are the mean forecasts and the dashed lines, the $\pm 2\sigma$ bounds from the ensemble forecasts. The green vertical lines marks when the BMA forecast can be stably performed. Top right: $CRPS_t$ for BMA and DMA forecasts. Bottom left: The evolution of model weights when DMA is used. Bottom right: The evolution of BMA weights.

Table 2.3: CRPS of the DMA and BMA ensemble, averaged over the entire 2017-2018 influenza season. Results are provided at the national level as well as the 5 US Dept. of Health and Human Services regions.

Method	National	HHS01	HHS02	HHS03	HHS04	HHS05
BMA ensemble	2.6	1.1	2.8	3.1	4.6	4.8
DMA ensemble	0.8	0.5	1.0	0.8	1.9	0.7

Table 2.4: CRPS of the raw and DMA ensemble, averaged over the entire 2016-2017 influenza season. Results are provided for 5 US Dept. of Health and Human Services regions.

Method	HHS06	HHS07	HHS08	HHS09	HHS10
Raw ensemble	2.3	1.44	0.72	2.3	1.4
DMA ensemble	1.4	0.87	0.40	0.6	0.61

of these three models provide their forecasts as $\mathcal{N}(\mu_t^{(k)}, \nu_t^{(k)}), k = 1 \ldots K = 3, t = 1 \ldots T$. These three forecasts are combined into an ensemble forecast using Eq. 2.3 with $\pi_{t|t-1,k}$, the weight of model k, replaced by $1/K = 1/3$. We consider two ways of arriving at a the top three models "that, on average, are most predictive across all 7 QoIs". We call them *simple ranking* and *rank aggregation*.

At any time l, for any QoI q, we can compute the relative prediction error $\epsilon_{l,q}^{(k)} = (y_l - \mu_l^{(k)})/y_l$ of any model k. One can compute the root mean square relative error $\tilde{\epsilon}_{t,q}^{(k)}$ over time steps $l = 1 \ldots t$. In *simple ranking* we compute the mean root-mean-square-relative-error over all $q = 1 \ldots Q = 7$ QoIs to obtain a mean predictive error $\overline{\tilde{\epsilon}_t^{(k)}}$. At any time t, $\overline{\tilde{\epsilon}_t^{(k)}}$ allows one to rank the models, and we choose the top 3.

In *rank aggregation* we do not average $\tilde{\epsilon}_{t,q}^{(k)}$ over the various QoIs. Instead, using $\tilde{\epsilon}_{t,q}^{(k)}$, we develop $t \times Q$ ranked lists $L_{tq}, t = 1 \ldots T, q = 1 \ldots Q = 7$ of models, one per QoI per timestep. At each timestep, we then "merge" the Q lists together using the combinatorial algorithm in Ref. [21] to obtain the "ensemble" list L_{ens}, from which we can pick the top three models. This method yields a different set of models than *simple ranking*, but Eq. 2.3 can be used to generate forecast using the *rank aggregated* list.

We next investigate whether the two ranking methods provide very different predictive accuracies and how they compare to BMA, DMA and the raw ensemble. In Fig. 2.6 we plot predictions of three-week-ahead ILI activity (top row), the peak ILI activity (middle row) and the week when the peak is reached (bottom row) using simple ranking (left column) and rang aggregation (right column). The CDC FluView data for the 2017-2018 influenza season is plotted with symbols and the error bars provide $\pm 2\sigma$ predictions. The results are for the National level. We see that apart from predictions of the peak week, which both methods predict poorly, the forecasts invariably bracket the actual data. This is quite different from the forecasts generated using DMA and BMA, where there were weeks that neither managed to bracket CDC FluView data (though that was only for the three-week-ahead QoI). This runs counter to accounts in literature indicating that the CDC FluSight ensemble of models may be different from the ensembles used in literature.

To investigate this further, in Table 2.5, we compare the predictions of three-week-ahead forecasts

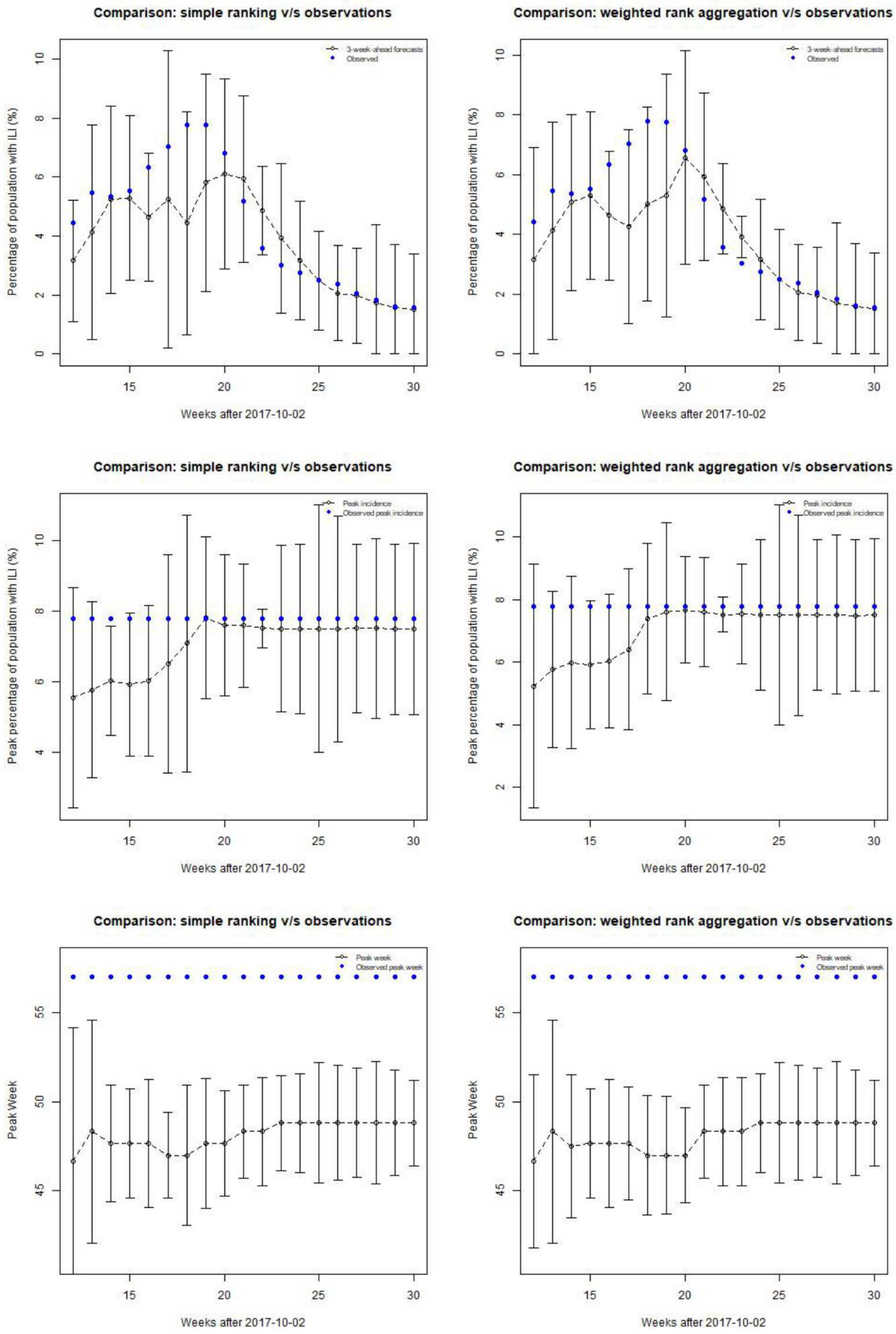


Figure 2.6: Comparison of ensemble predictions when the top three models are chose according to simple ranking (left column) and rank aggregating (right column). The filled symbols are CDC FluView data whereas the error bars plot the mean and $\pm 2\sigma$ forecasts. The top row contains the three-weak-ahead forecasts of ILI activity, the middle row contains the predictions of the peak ILI incidence and the bottom row predicts the week during which the peak ILI activity would be achieved. All forecasts are for the Nation as a whole, for the 2017-2018 influenza season.

Table 2.5: Comparison of three-week-ahead forecast for 2017-2018 influenza season for all regions, performed using the two ranking-based methods for MA, DMA and the raw ensemble. We see that the ranking-based methods are better. The poor performance of DMA is partially due to all QoIs not being available for the late epoch of the outbreak, where DMA performs well.

Region	Simple ranking	Rank aggregation	DMA	Raw ensemble
National	0.63	0.69	0.82	1.04
HHS01	0.57	0.56	0.54	0.64
HHS02	0.76	0.71	1.2	0.84
HHS03	0.92	0.91	1.14	0.8
HHS04	1.65	1.67	1.92	1.42
HHS05	0.51	0.51	0.72	0.82
HHS06	1.53	1.42	1.35	1.22
HHS07	0.91	0.93	1.2	1.31
HHS08	0.4	0.44	0.42	0.51
HHS09	0.55	0.59	0.58	0.68
HHS10	0.59	0.58	0.68	0.77

of ILI activity, for all HHS regions for the 2017-2018 season. These predictions are generated using the two ranking based methods, DMA, and the raw ensemble. The comparison is performed using CRPS computed over the entire 2017-2018 influenza season. We do not include BMA in the comparison due to its poor performance in Tables 2.3 and 2.4. The ensemble does not provide forecasts for all the QoIs for the entire season so the comparison in Table 2.5 covers a smaller duration than the one used to perform the BMA versus DMA comparison in Tables 2.3 and 2.4. This is also the reason why the CRPS values generated for the DMA-ed ensemble differ between the two tables; DMA becomes very accurate in the later half of the outbreak. Looking at the columns in Table 2.5, it is clear that

- The two ranking methods for conditioning the 2017-2018 CDC FluSight ensemble perform almost identical to each other, though the *rank aggregation* algorithm is far more computationally expensive, and,
- The two rank-based methods are far more predictive than the raw ensemble and the DMA-ed one.

This is quite at variance with the finding in literature with smaller and more homogeneous ensembles of models.

2.4 Limits of Applicability

2.4.1 Non-transferability of Model Weights

In Sec. 2.3.1 and 2.1 we developed expressions for model weights $\pi_{t|r,k}$ and demonstrated them on the forecasting of three-week-ahead predictions of ILI activity. Presumably, these weights are

Table 2.6: CRPS of predictions of the peak epidemic week using weights computed for the three-week-ahead forecasts of ILI activity. Results are plotted for the raw, BMA and DMA ensemble, averaged over the entire 2016-2017 influenza season. Results are provided at the national level as well as the 5 US Dept. of Health and Human Services regions.

Method	National	HHS01	HHS02	HHS03	HHS04	HHS05
BMA ensemble	49	52	51.5	51.8	47.8	53.4
DMA ensemble	49.4	55.2	52.2	51.9	47.8	51.6
Raw ensemble	49.5	51.9	50.4	51.0	48.0	52.3

Table 2.7: CRPS of predictions of the peak epidemic week using weights computed for the three-week-ahead forecasts of ILI activity. Results are plotted for the raw, BMA and DMA ensemble, averaged over the entire 2016-2017 influenza season. Results are provided for 5 US Dept. of Health and Human Services regions.

Method	HHS06	HHS07	HHS08	HHS09	HHS10
BMA ensemble	44	48.3	50.8	49.7	51.1
DMA ensemble	47	50.3	52.8	51	51.2
Raw ensemble	46.3	49.6	52.7	50.1	51.3

related to model plausibility, given data. If all models are equally plausible in their predictions of the 7 QoIs they forecast, then weights computed for one QoI should, in principle, be useable for other QoIs, which would be immensely helpful for MA. We investigate this hypothesis. It does run against the findings in literature that MA should be performed for individual QoIs.

In Fig. 2.7 we plot predictions of the week of peak ILI activity, but model-average the ensemble using weights computed using three-week-ahead predictions of ILI activity as the QoI. Results are plotted for BMA, DMA and the raw ensemble. On the top row, we plot predictions at the National level; in the bottom row, results for HHS Region 5. We see that the mean predictions are similar while BMA seems to have much tighter prediction bounds. For HHS05, there is no such obvious improvement in predictions. To investigate this lack of consistency further, we perform ensemble predictions for all HHS regions, compute their CRPS over the entire 2017-2018 season, and tabulate them in Tables 2.6 and 2.7.

The results in the tables show that (1) the prediction errors are large and (2) all the methods have the same predictive skill as the raw ensemble. Thus while transferability of model weights might succeed when we have related QoIs e.g., the one-week-ahead and three-week-ahead forecasts, in general, it is preferable to MA QoIs individually. This has been observed in literature for BMA, but we verify that the same holds for DMA.

2.4.2 Non-transferability of Model Rankings

In Sec. 2.3.2 we showed how ranking-based method for model-averaging an ensemble performed very well for the collection of influenza models being considered in this study. These rankings were generated using all the QoIs. In Sec. 2.4.1 we showed that the model weights computed using a QoI cannot be used to accurately forecast any other. We check whether a similar non-transferability of

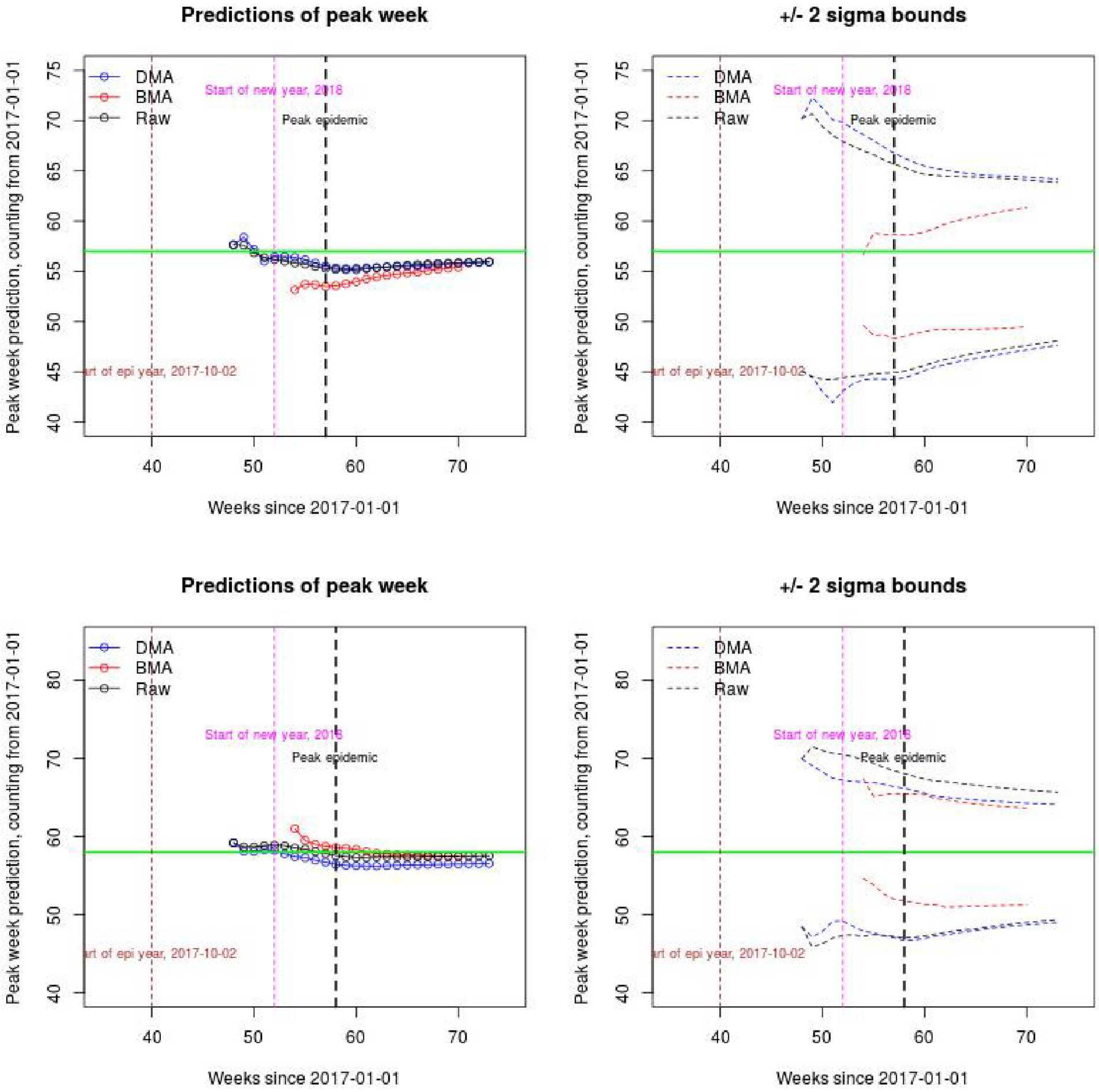


Figure 2.7: Predictions of the peak ILI indicence week, for the 2017-2018 influenza season, using DMA, BMA and the raw ensemble. Note, however, the weights were computed using the three-week-ahead forecast as the QoI. Top left: The mean predictions at the National level. Top right: The $\pm 2\sigma$ bounds on the predictions. The horizontal line is the true value. Bottom: Results for HHS05.

rankings hold for the ranking-based MA methods tested in Sec. 2.3.2, as there is no literature on the use of ranking-based methods to MA ensembles of disease models.

The models in the ensemble, as described in Sec. 2.3.2, predict 7 QoIs. We can choose, at random, a QoI to hold back, and devise a ranking of the models, using both rank-aggregation and simple ranking. The top three model can then be used to forecast the held-back QoI, and the accuracy compared with MA performed in the conventional manner i.e., when all the QoIs are used in the forecasting. The comparison can be performed using CRPS as the metric of accuracy across predictions performed at the National level and the 10 HHS regions. We assign a QoI from our set of seven (without replacement) to each of the 11 regions (10 HHS regions and the National level). We hold that QoI back when performing the ranking of models, and then forecast it using the top three models. The results for rank-aggregation are in Table. 2.8 and in Table 2.9 for simple ranking. We tabulate the CRPS of the forecast for 2017-2018 season for all the HHS regions and the National level, for all the QoIs. The CRPS for forecast performed with model rankings achieved with all the models are also provided.

We consider three conditions:

1. If the rankings are transferable, then the difference between CRPS will be small, but in general ranking with all seven QoIs will provide better forecasts.
2. If the rankings are only somewhat transferable, then omitting the QoI being forecast from the ranking process will result in inaccurate forecasts.
3. If the QoIs used in the ranking have no information to contribute towards the QoI that was held-back and is being forecast, then the results will be random.

The results in Table. 2.8 and 2.9 show that forecasting performed with models ranked using all seven QoIs turns out to be better five out of eleven times, regardless of the ranking algorithm. Further, the HHS regions for which ranking using all QoIs performed better are identical for the two ranking-based MA methods. This is equivalent to a random performance, indicating that holding back the QoI resulted in a ranking of models that had no information on that QoI. Thus rankings of models cannot be transfered across QoIs, no different from the results in Sec. 2.4.1.

2.4.3 Impact of Optimized α

The formulation in Sec. 2.1 contains a free parameter α that is set to 0.99 by default. It governs the degree of temporal correlation between weights $\pi_{t-1|t-1,k}$ and $\pi_{t|t-1,k}$, which is critical for stability as we evolve them with sparse data. However, if the temporal correlation is too strong (α near 1), and the ensemble evolves quickly in time, the unnecessary correlation can impair predictive skill. Thus it might be worthwhile to tune α.

We iterate through $\alpha \in [0.5, 0.99]$ in steps of 0.01 and perform DMA for 2017-2018, at the National levels and all other HHS regions. This is done for the three-week-ahead prediction of ILI activity and we compute the $CPRS_t$. This is compared to the $CRPS_t$ computed with the default value of $\alpha = 0.99$ and we determine the proportion of weeks where the alternative value of α performed better. This is plotted in Fig. 2.8 and we see that a value of $\alpha_{opt} = 0.77$ is optimal for 2017-2018. An identical study performed for 2016-2017 revealed $\alpha_{opt} = 0.73$. However, in both cases

Table 2.8: Comparison of forecasts of the held-out QoI in the 2017-2018 influenza season when models are ranked using rank-aggregation. We see keeping ranking based on all seven QoIs provide better forecasts 5 / 11 times, but the difference in the accuracy of the predictions follow no discernible trend.

Region	Held-out QoI	CRPS; QoI held out	CRPS; all QoIs
National	4-wk-ahead ILI forecast	0.77	0.71
HHS01	1-wk-ahead ILI forecast	0.44	0.56
HHS02	Season onset	4.25	7.19
HHS03	Season peak week	5.22	8.00
HHS04	3-wk-ahead ILI forecast	1.73	1.69
HHS05	2-wk-ahead ILI forecast	0.55	0.64
HHS06	Season peak incidence	2.86	0.59
HHS07	1-wk-ahead forecast	0.62	0.56
HHS08	3-wk-ahead ILI forecast	0.40	0.44
HHS09	2-wk-ahead ILI forecast	0.57	0.64
HHS10	Season onset	7.73	7.19

Table 2.9: Comparison of forecasts of the held-out QoI in the 2017-2018 influenza season when models are ranked using simple ranking. We see keeping ranking based on all seven QoIs provide better forecasts 5 / 11 times, but the difference in the accuracy of the predictions follow no discernible trend.

Region	Held-out QoI	CRPS; QoI held out	CRPS; all QoIs
National	4-wk-ahead ILI forecast	0.65	0.64
HHS01	1-wk-ahead ILI forecast	0.44	0.53
HHS02	Season onset	4.29	7.14
HHS03	Season peak week	4.96	7.88
HHS04	3-wk-ahead ILI forecast	1.65	1.65
HHS05	2-wk-ahead ILI forecast	0.54	0.60
HHS06	Season peak incidence	2.95	0.60
HHS07	1-wk-ahead forecast	0.66	0.53
HHS08	3-wk-ahead ILI forecast	0.39	0.40
HHS09	2-wk-ahead ILI forecast	0.51	0.60
HHS10	Season onset	7.53	7.14

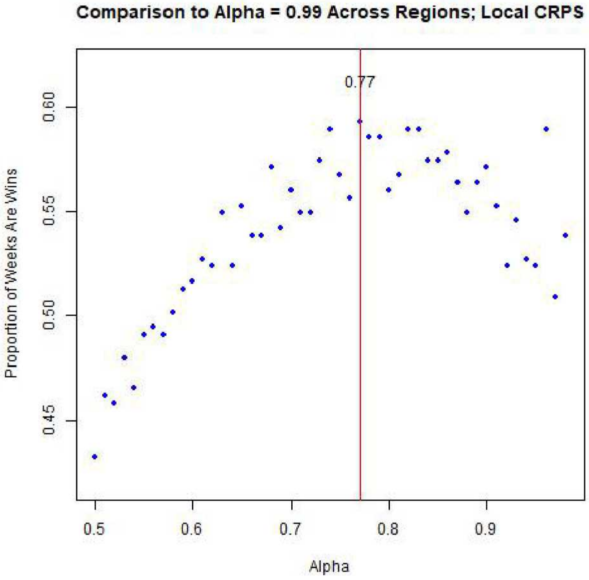


Figure 2.8: Number of weeks where DMA, operated with the default value of α, looses to a different α value. The results are tallied over all HHS regions and weeks, for the 2017-2018 influenza season.

we see that α_{opt} outperforms the default value about 60% of the time i.e., it is not an impressive improvement, even though the default value of α is quite different from the optimal one. However, it does not say how much of a difference α_{opt} makes for the predictions for each HHS region.

Table 2.10: Proportion of weeks (as percentages), in the 2016-2017 influenza season, where α_{opt} yields a better $CRPS_t$ than the default value of 0.99. The performance of the raw ensemble is also provided for a comparison.

Region	$\alpha = 0.99$	$\alpha = 0.73$	Raw ensemble
National	29	54	17
HHS01	46	29	25
HHS02	38	29	33
HHS03	38	33	29
HHS04	25	37.5	37.5
HHS05	29	50	21
HHS06	37.5	37.5	25
HHS07	29	42	29
HHS08	50	29	21
HHS09	58	29	13
HHS10	42	33	25

In Table 2.10 we tabulate the number of winning weeks when using $\alpha_{opt} = 0.73$ compared to the default $\alpha = 0.99$ for each HHS region. Results are computed for 2016-2017 for three-week-ahead predictions of ILI activity. We see that the "optimized" DMA outperforms the "default" DMA in 5 of the HHS regions, and in some cases, it ties with the raw ensemble. Thus while we may have tuned α, it has not resulted in any impressive improvement in predictions. This is because the optimization of α yields very different values for α_{opt} for the different HHS regions. We have, therefore, performed the study with $\alpha = 0.99$.

Chapter 3

Stacking

Stacking is an ensemble learning technique that combines predictions generated by different learning algorithms or models by using those predictions as inputs to a second-level learning algorithm [7]. For example, suppose you have n models, $M_1, M_2, ..., M_n$, and each generates a prediction, $y_1, y_2, ..., y_n$. A stacking framework uses another learning algorithm, called a meta-learning algorithm, to combine the predictions y_1 to y_n into one final prediction, y. Inputs to the meta-learning algorithm are called meta-features and typically include the model predictions (i.e., y_1 to y_n), but can also include additional values such as standard deviation, maximum probability, entropy, etc.

In this study, we trained a stacking model to combine the predictions from the flu models in the CDC forecasting challenge from the 2016−2017 season. The CDC challenge includes predictions from 29 models, but only 22 of those models have predictions for all 11 Health and Human Services (HHS) regions and all weeks of the flu season. Only those 22 models were included in the stacking model (see Table 3.1).

Model predictions for the 2016−2017 flu season started on 11/28/2016 (week 8) and continued once a week until 5/15/2017 (week 32). For each week of the flu season, starting with week 9, we trained a new stacking model using data from all previous weeks and all HHS regions. For example, in week 9, a stacking model was trained using data from week 8, resulting in a training dataset with 11 samples (one from each HHS region) and 22 features (one from each flu model). In week 32, a stacking model was trained using data from weeks 8 to 31, resulting in a training dataset with 264 samples (from 11 HHS regions for 24 weeks) and 22 features (one from each flu model). The final result was a set of 24 stacking models, one for each week of the flu season, where a stacking model for a given week was used to make predictions for all of the HHS regions.

The training datasets have relatively few samples compared to the number of features, especially in the early weeks of the flu season, leading to a risk of overfitting. To reduce this risk, we used linear regression with regularization as the meta-learning algorithm. Regularization is a penalty applied to the optimization in linear regression, defined as follows:

$$\lambda \cdot \left[\left(\frac{1-\alpha}{2} \right) \|W\|_2^2 + \alpha \cdot \|W\|_1 \right] \tag{3.1}$$

where W is the vector of regression weights, α is the elastic net mixing parameter (i.e., $\alpha = 1$ is the lasso penalty and $\alpha = 0$ is the ridge penalty), and λ is the regularization parameter. Larger λ leads to smaller regression weights and less overfitting.

To perform linear regression with regularization, we used the `glmnet` package in `R` [15, 22]. We performed cross-validation over λ and α to find the model with smallest mean square error.

Table 3.1: Flu models from the CDC forecasting challenge included in the stacking model. These models have predictions for all 11 HHS regions and all 25 weeks of the flu season.

1	4Sight
2	CU1
3	CU2
4	CU3
5	CU4
6	Delphi Epicast
7	Delphi Stat
8	GHRI
9	HumNat
10	ICS
11	ISU
12	KBSI
13	KOT Dev
14	KOT Stable
15	LANL
16	NEU
17	PSI
18	TeamA
19	TeamB
20	TeamC
21	Yale1
22	Yale2

We also explored two sets of meta-features: 1) the mean 3-week ahead prediction from each flu model, and 2) the mean and standard deviation of the 3-week ahead prediction from each flu model.

3.1 Meta-Feature Set 1

For the first set of meta-features, we used the mean 3-week ahead prediction from each of the 22 flu models in Table 3.1. We performed linear regression with regularization, as defined in Eq. 3.1, and the resulting linear coefficients after cross-validation on α and λ are shown in Fig. 3.1. During the start of the flu season, the model coefficients are positive and constantly changing, meaning no model or set of models dominates the prediction. Just before the flu season peak, near weeks 15 and 16, the few models that predict the peak well become very important and other model coefficients drop to zero. After the season peak, the model coefficients stabilize and the group of models with the most accurate predictions consistently have the largest coefficients.

The prediction made by the stacking model in each week and HHS region is shown in Fig. 3.2. For each week of the flu season, a single stacking model was trained using data for all HHS regions. That model was then used to make predictions for all regions. Ideally, a unique model for each HHS region and each week would be trained, but there was not enough training data to do so.

The mean square error from stacking was compared to Dynamic Model Averaging (DMA) and the raw ensemble. DMA is described in a previous section of this report and the raw ensemble is simply the unweighted average of the mean predictions from all flu models. Fig. 3.3 shows the mean square error for each method in each region. The raw ensemble has significantly larger mean square error in HHS regions 1, 6, and 7 due to outliers. Stacking and DMA have similar performance and neither is consistently better than the other. See Table 3.2 for the average mean square error across all regions for each method.

Table 3.2: Average mean square error (MSE) for each ensemble method across all HHS regions and weeks.

Method	MSE
Raw	4.4074
DMA	0.9585
Stacking	0.9393

We also compared the structure of the ensemble from stacking to DMA. Fig. 3.4 shows a histogram of the model coefficients for stacking and DMA for all weeks of the flu season. The DMA coefficients tend to cluster around 0 and 1, indicating that DMA tends to select a small group of models with the best predictions. In contrast, the stacking coefficients are more evenly distributed, indicating that stacking tends to use information from all the models.

These findings are further reinforced by computing and plotting the Gini coefficient of the model weights [6]:

$$G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2n^2\mu} \tag{3.2}$$

where n is the number of data points and μ is the average of x.

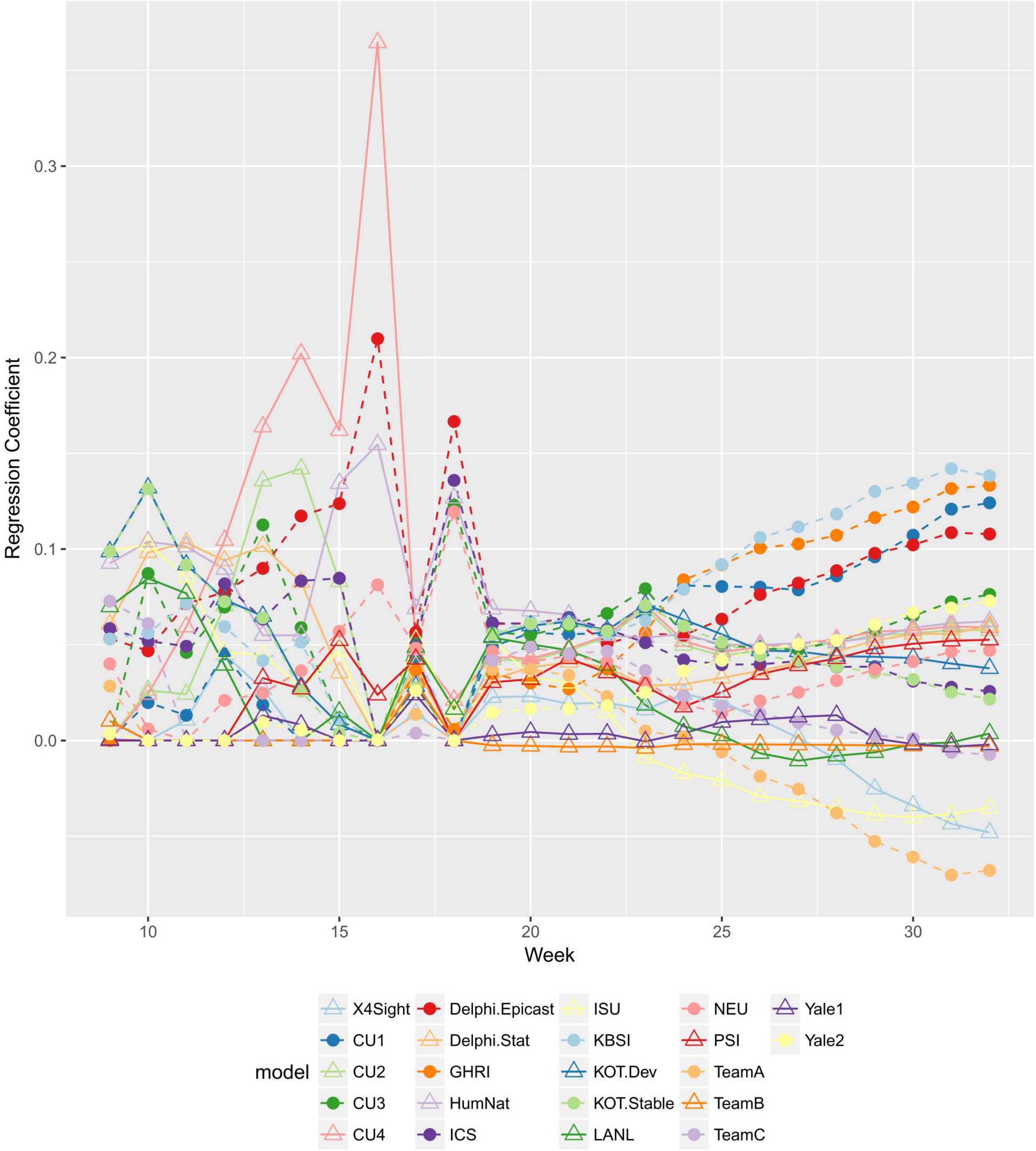


Figure 3.1: Regression coefficients on the mean 3-week ahead predictions from each flu model.

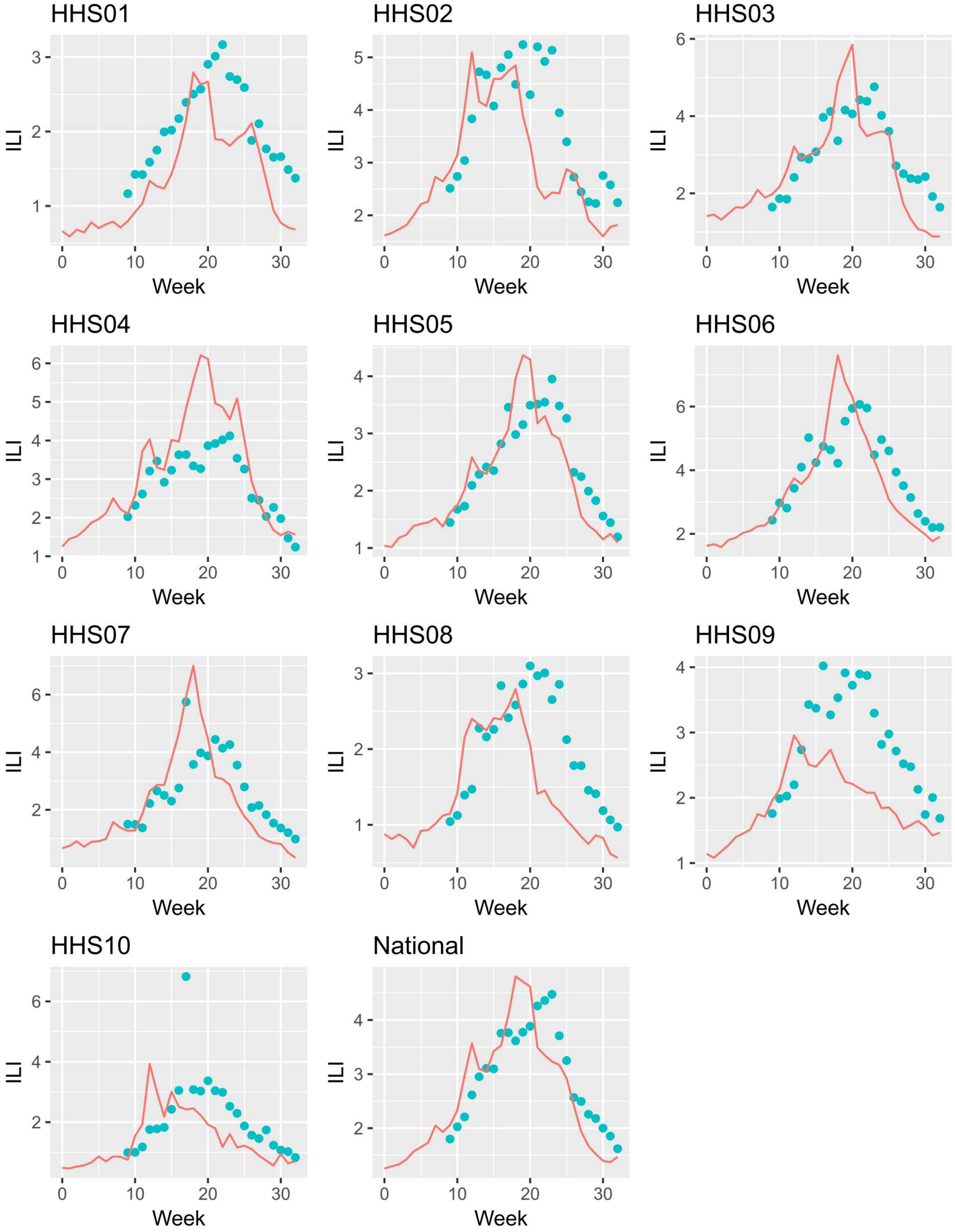


Figure 3.2: Predicted ILI from stacking (blue circles) in each HHS region compared to the FluView ground truth (red line).

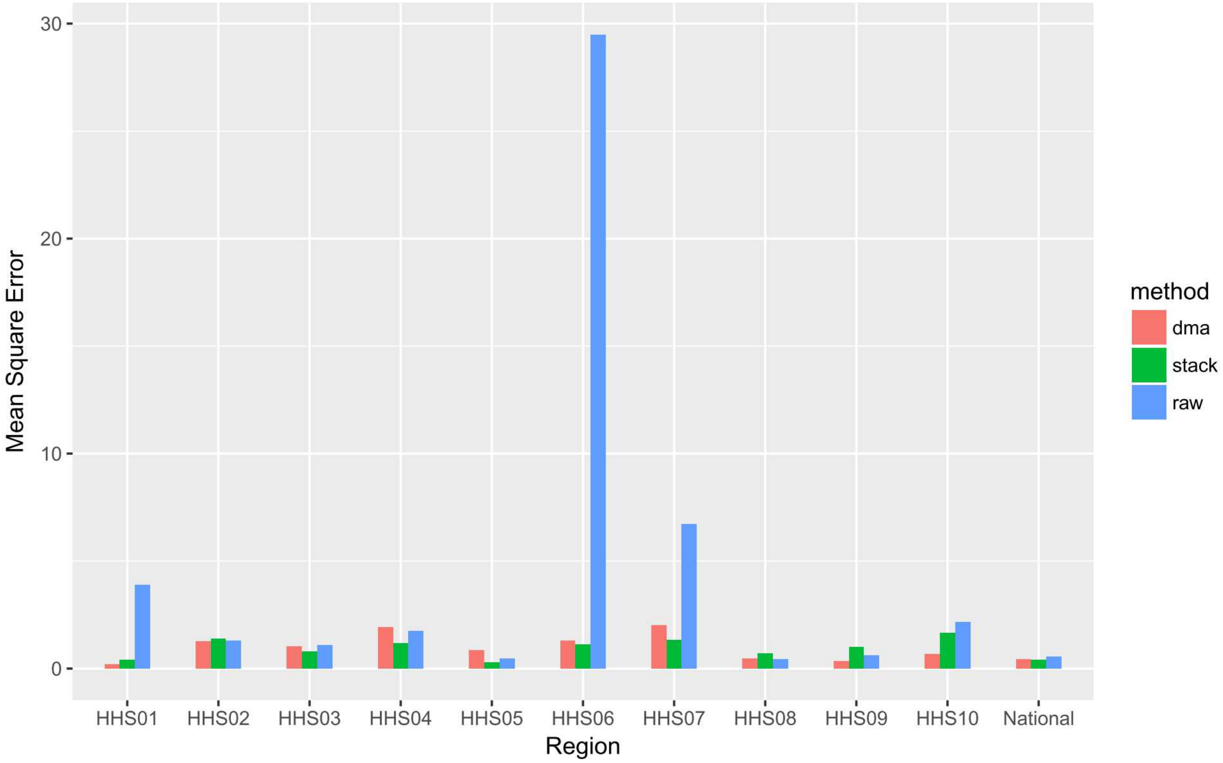


Figure 3.3: Average mean square error for each HHS region and each ensemble method.

The Gini coefficient provides a quantitative summary of the dispersion of a set of numbers. A large Gini coefficient (close to one) indicates that all values except one are equal to zero. A small Gini coefficient (close to zero) indicates that all values in the distribution are equal. Fig. 3.5 shows the Gini coefficients of the model weights for each week for stacking and for DMA performed on National data. As expected based on Fig. 3.4, DMA has a large Gini coefficient and stacking has a smaller Gini coefficient. The Gini coefficient for stacking increases just before the peak of the flu season around week 15, indicating that during this time, stacking tends to highly-weight a few models rather than uniformly weight all models.

The Spearman rank correlation was computed for the model weights in each week for stacking and DMA as shown in Fig. 3.6. Spearman rank correlation is defined as:

$$r = 1 - 6 \cdot \frac{\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{3.3}$$

where n is the number of data points and d_i is the difference in rank between the variables of interest.

A Spearman rank correlation near one indicates that the ordering of the coefficients on the models do not change significantly week-to-week, meaning the same models are most predictive each week. A small Spearman rank correlation indicates that the ordering of the model coefficients changes from week-to-week and therefore there is no group of models that performs best. The Spearman rank correlation is similar for stacking and DMA, except in the final few weeks of the flu season when the Spearman rank correlation is consistently near one for stacking. This indicates

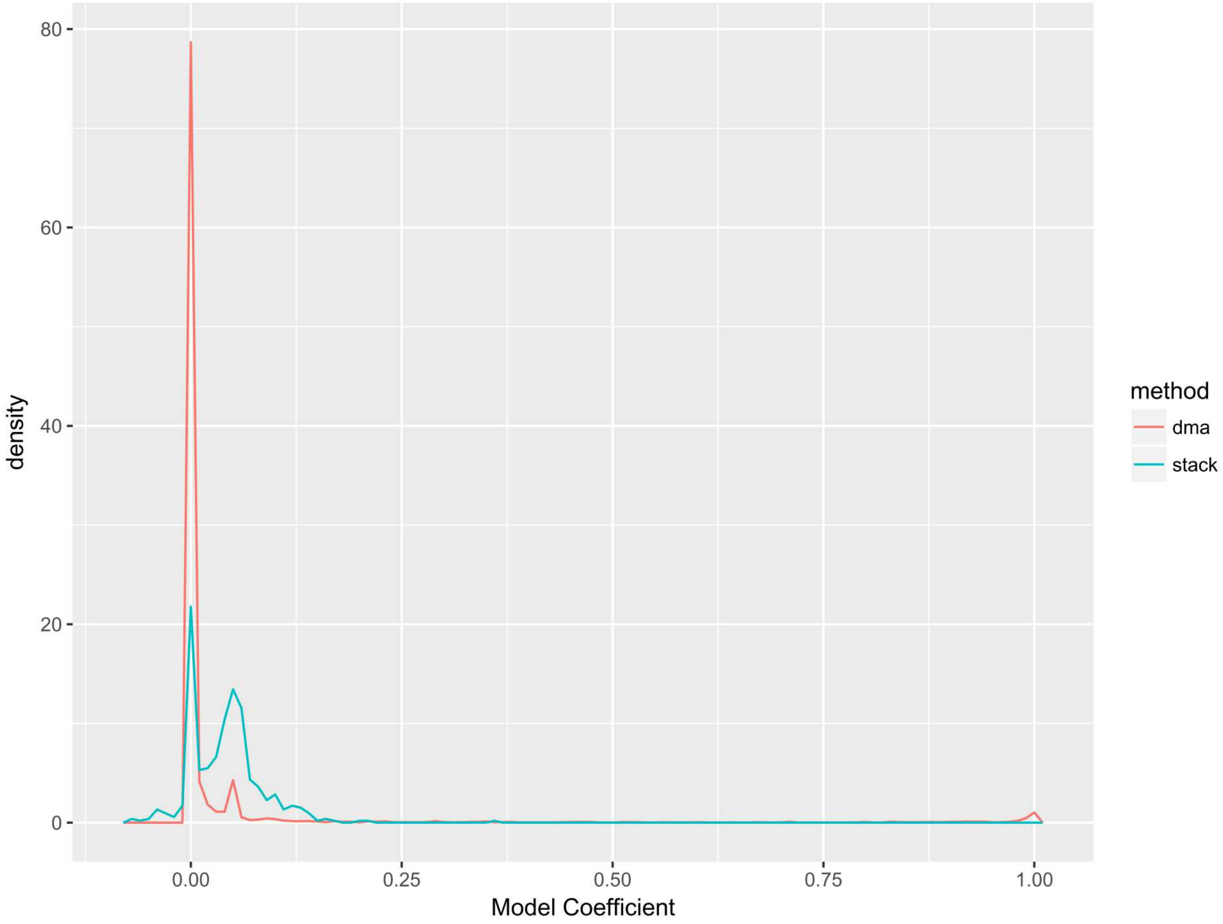


Figure 3.4: Histogram of model coefficients from the stacking and DMA ensemble methods.

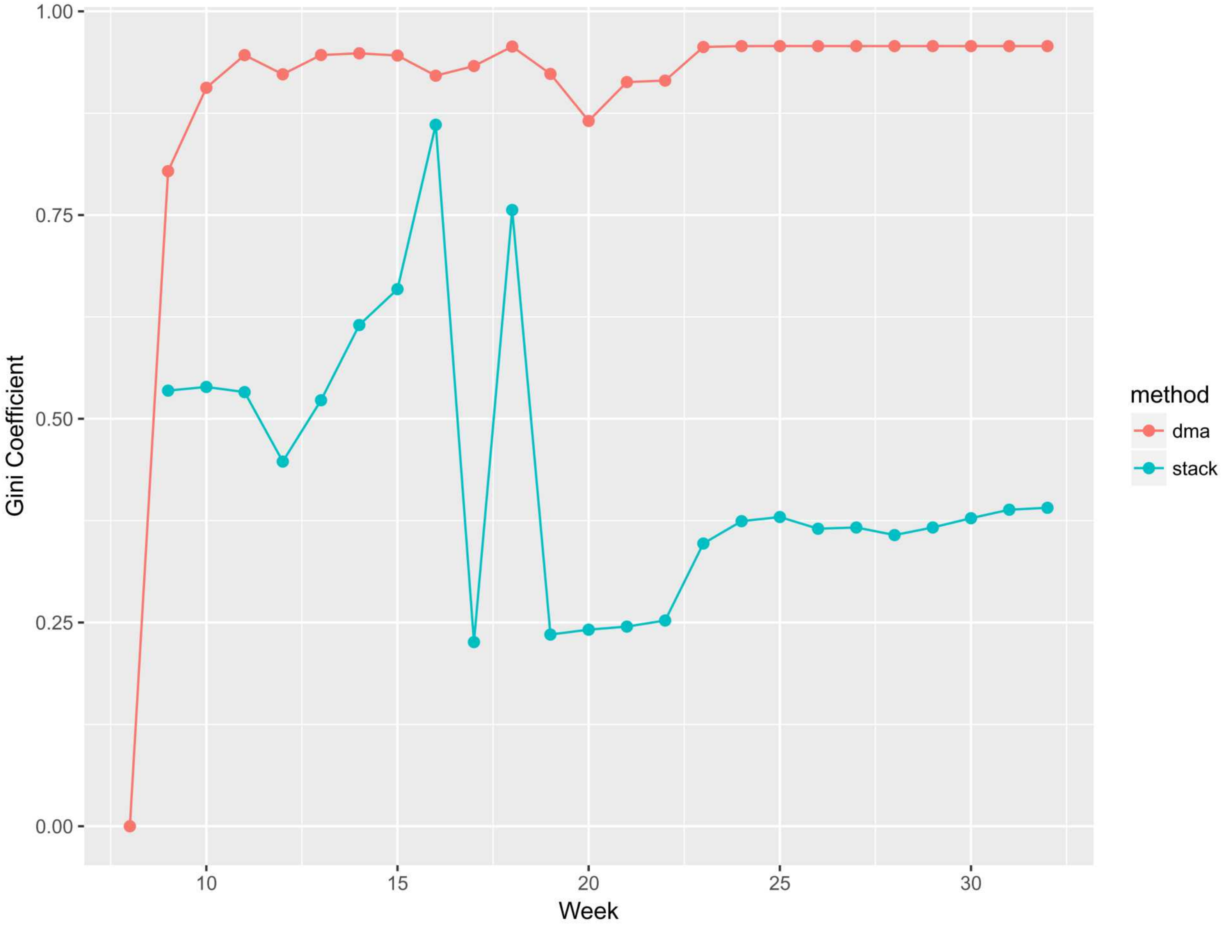


Figure 3.5: Gini coefficient of the model weights for each week for stacking and DMA performed on national data.

that once the peak of the flu season has passed, stacking finds a consistent ranking of the models.

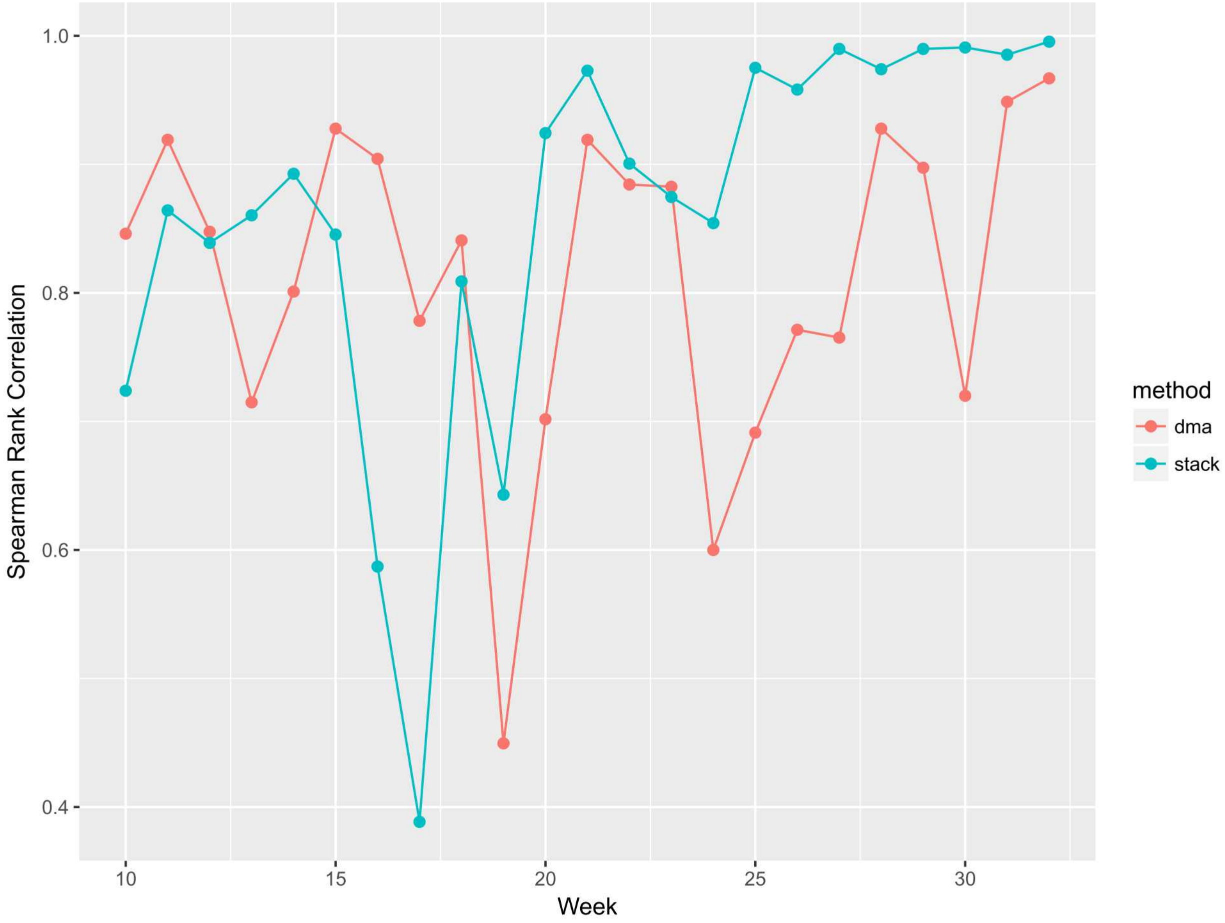


Figure 3.6: Spearman rank correlation of the model weights for each week for stacking and DMA performed on national data.

3.2 Meta-Feature Set 2

For the second set of meta-features, we used the mean and standard deviation of the 3-week ahead prediction from each of the 22 flu models in Table 3.1. This results in 44 features for the meta-learning algorithm, making overfitting an even greater potential issue. After performing linear regression with regularization and cross-validation on α and λ, the resulting coefficients are shown in Fig. 3.7. The coefficients are generally much larger than those using the means only as meta-features. Coefficients are especially large for some of the model standard deviations, which is unexpected.

The mean square error was compared to Dynamic Model Averaging (DMA) and the raw ensemble (see Fig. 3.8). Similar to results from meta-features set 1, the raw ensemble suffers from outliers while stacking and DMA have similar performance and neither is consistently better than the other. See Table 3.3 for the average mean square error across all regions for each method. The overall MSE is smaller for stacking using meta-feature set 2 than meta-feature set 1, but this is most likely

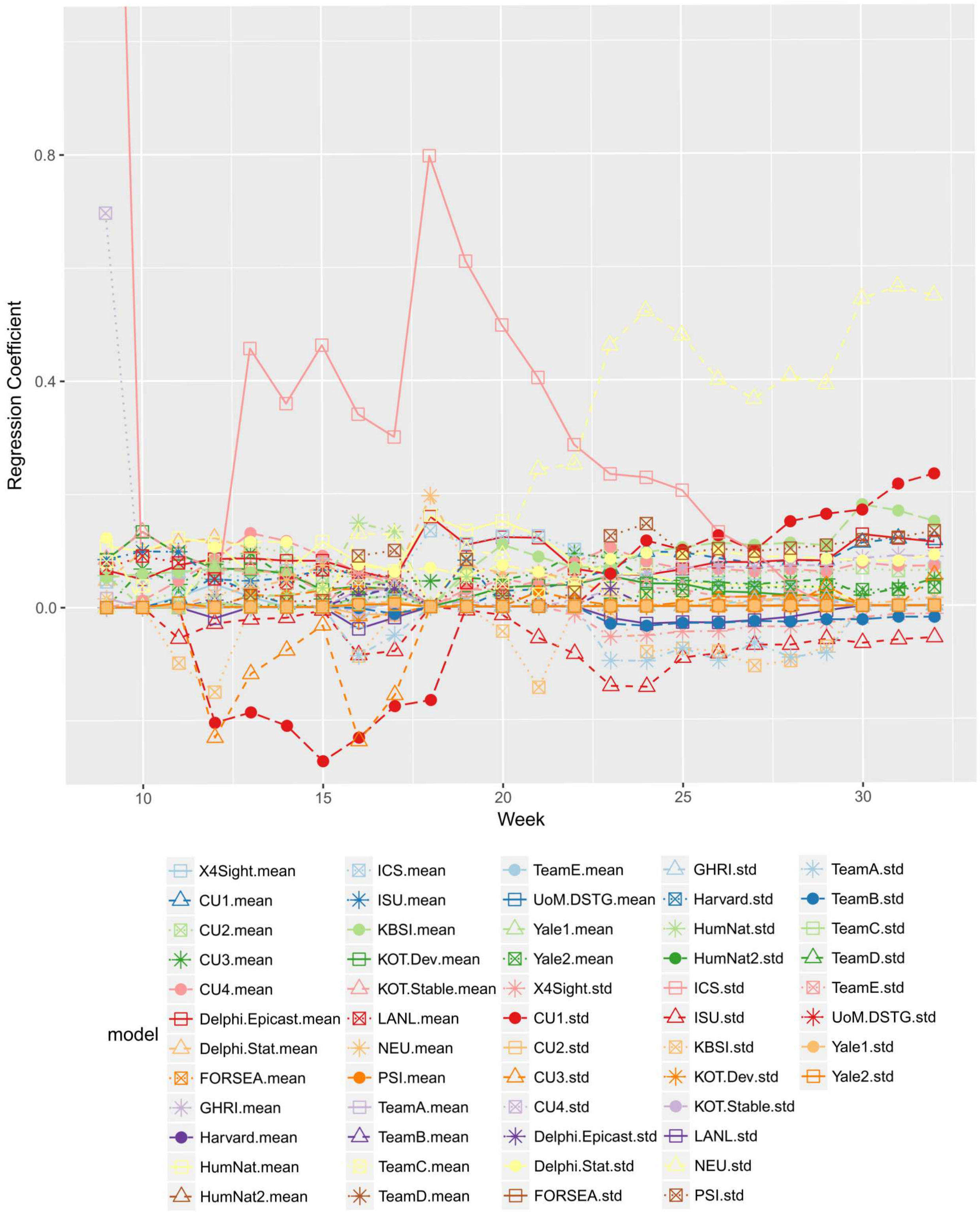


Figure 3.7: Regression coefficients on the mean and standard deviation of the 3-week ahead predictions from each flu model.

due to overfitting as described in the following paragraph.

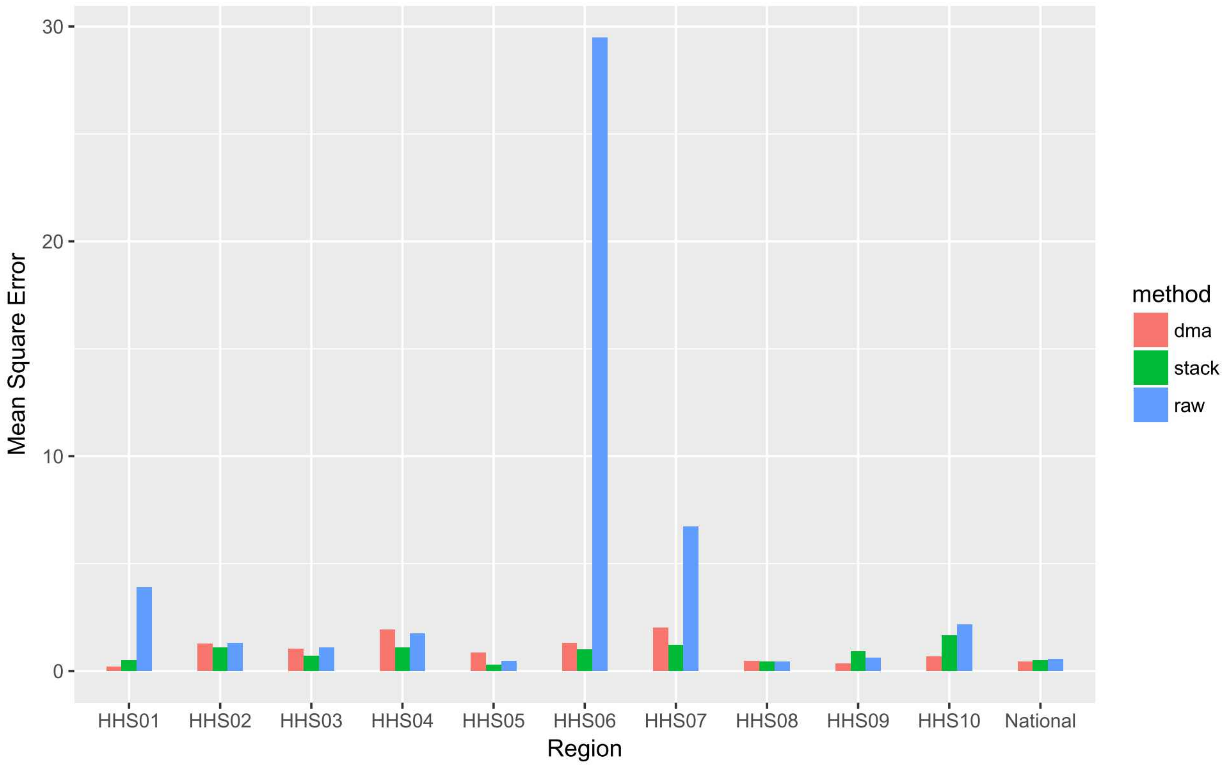


Figure 3.8: Average mean square error for each HHS region and each ensemble method.

Table 3.3: Average mean square error (MSE) for each ensemble method across all HHS regions and weeks.

Method	MSE
Raw	4.4074
DMA	0.9585
Stacking	0.8569

For each flu model, we computed the ratio of the linear regression coefficient on the standard deviation of the prediction to the coefficient on the mean prediction. A histogram of these ratios over all flu models in all HHS regions and all weeks of the flu season is shown in Fig. 3.9. A ratio larger than 1 indicates that the linear regression coefficient on the standard deviation of the prediction is larger than the mean, and therefore the standard deviation of the model prediction is more important to the final prediction than the mean prediction. This conclusion does not make intuitive sense and is therefore likely the result of overfitting. Since more than half of the ratios are greater than 1 (as shown in Fig. 3.9), we conclude that using the standard deviation of the model prediction as a meta-feature leads to significant overfitting. Therefore we use only the mean predictions as meta-features in the following sections. This overfitting also highlights the need for additional methods to reduce overfitting, such as feature pruning.

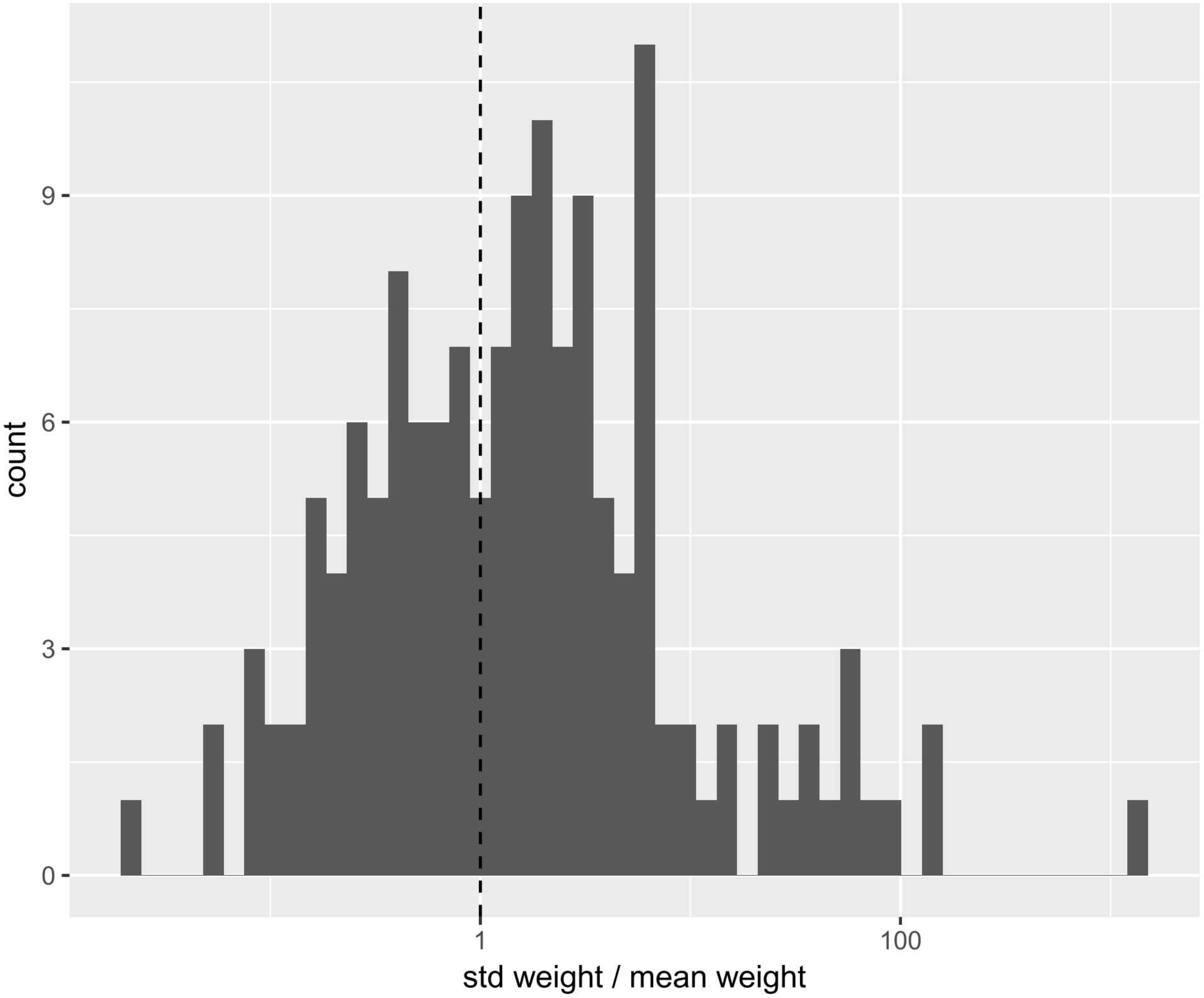


Figure 3.9: Histogram of the ratio of the linear regression coefficient for the standard deviation of the model prediction to the coefficient for the mean prediction. Ratios larger than 1 indicate that the standard deviation is more important than the mean for the final prediction.

3.3 Feature Pruning

Feature pruning is a technique to decrease the number of features in a learning algorithm in order to reduce redundancy, dimensionality, computational complexity, and error [4]. Many feature pruning techniques exist in literature, and here we applied two pruning methods to stacking. Both methods are greedy, meaning they add one feature to the meta-learning algorithm at a time and see how performance is affected. However, they differ in the criteria used to greedily select the next feature. The first method uses the Pearson correlation coefficient and the second method uses mutual information (MI).

3.3.1 Correlation Coefficient

The Pearson correlation coefficient is a simple selection criteria defined as [4]:

$$\rho = \frac{cov(x_i, Y)}{\sqrt{var(x_i) \cdot var(Y)}} \tag{3.4}$$

where x_i is the i^{th} feature, Y is the observation, $cov()$ is the covariance, and $var()$ is the variance.

In this study, feature pruning was applied to each week of the flu season. For a given week, the training dataset was the mean 3-week ahead prediction from each of the 22 flu models in Table 3.1 from all previous weeks. The Pearson correlation coefficient was computed between each feature and the observations (see Fig. 3.10). Some models consistently have the largest correlation with observations, such as Delphi Epicast and KBSI, while others consistently have low correlation, such as TeamB and Yale2. However, the rank of models changes over time and no single model is always ranked best or worst.

For each week in the flu season, the two features with the largest correlation with the observations were chosen and used to train the stacking model. Features were then added to the stacking model one at-a-time in ranked order based on their correlation coefficient with the observations. After each feature was added to the stacking model, both the training and testing error were computed. The training error was computed by performing 3-fold cross-validation during training and taking the average of the relative error across folds. The testing error was computed by using the trained model to predict the observation for the current week and taking the relative error between the prediction and observation.

Fig. 3.11 shows the training and testing error averaged across all HHS regions and weeks of the flu season as a function of the number of features. The training error increased with number of features up to about 10 features, then leveled off. In contrast, the testing error was nearly constant up to about 10 features, then it steadily declined. This behavior indicates that in general, more features lead to a more accurate prediction.

Fig. 3.12 shows the number of features that resulted in the stacking model with minimum error for each week of the flu season. Near the start of the flu season, using fewer features was best because the individual flu models are not well calibrated for the flu season yet. The number of useful features generally increased as the flu season progressed, until the flu season peak around weeks 16 to 20. During the flu season peak, the use of fewer features was better because relatively few flu models correctly predicted when the peak would occur. As the season progressed beyond

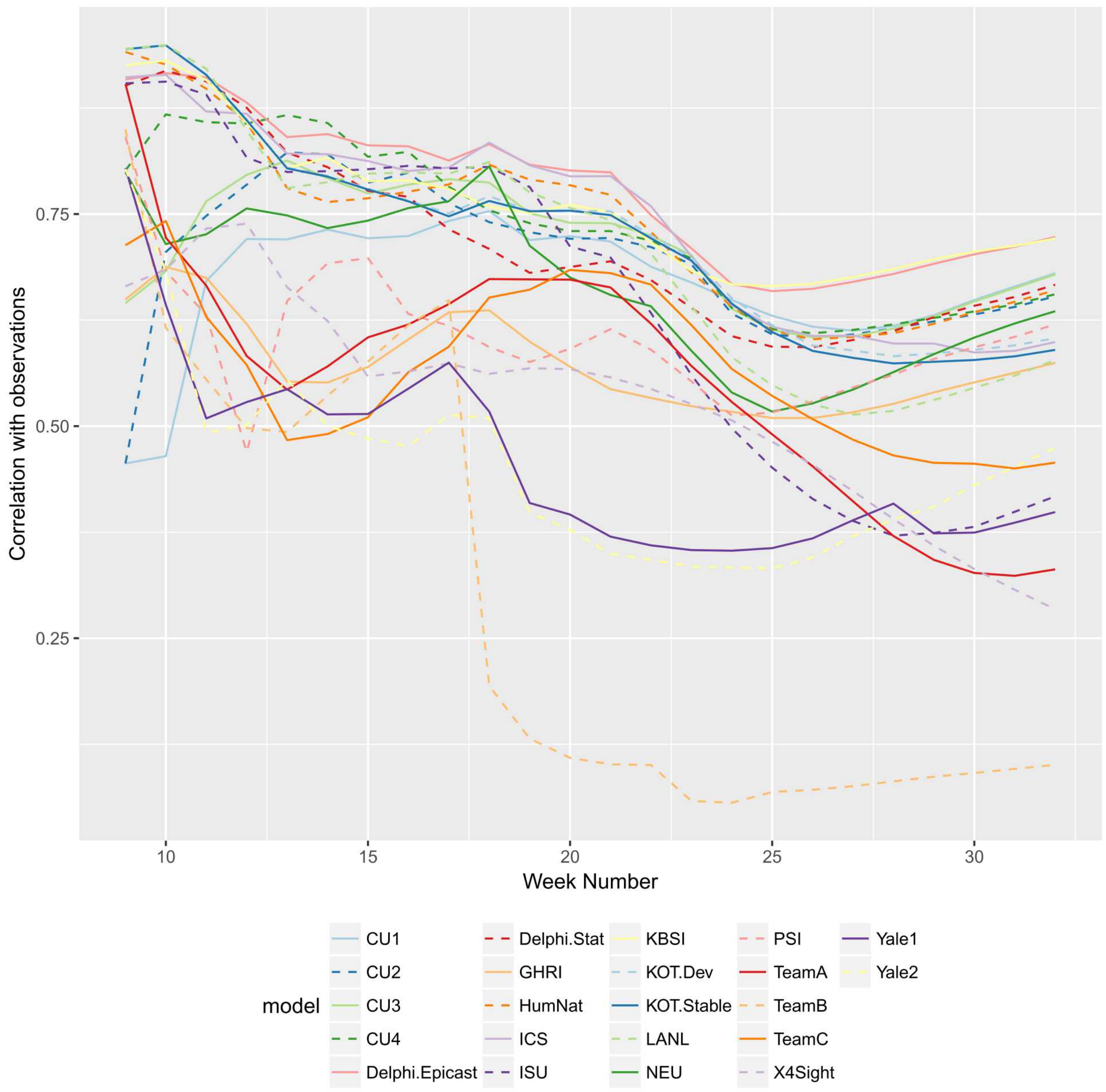


Figure 3.10: Pearson correlation coefficient between the mean 3-week ahead prediction of each flu model with the observations.

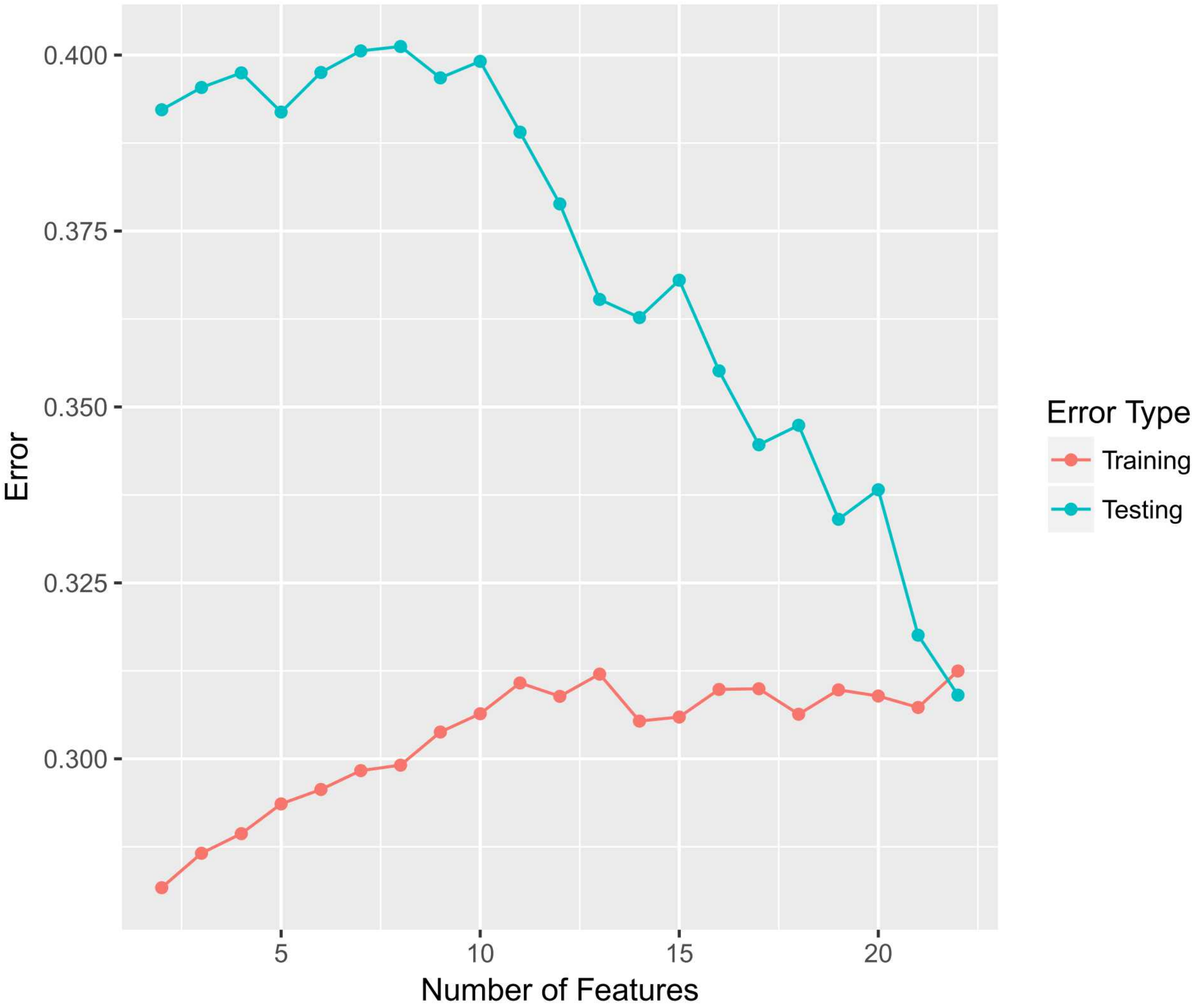


Figure 3.11: Training and testing error for the stacking model, averaged across all HHS regions and weeks, as a function of the number of features.

the peak, the training error resulted in a stacking model with fewer features than the testing error. This result indicated that using more features does not necessarily lead to overfitting, but might actually improve performance.

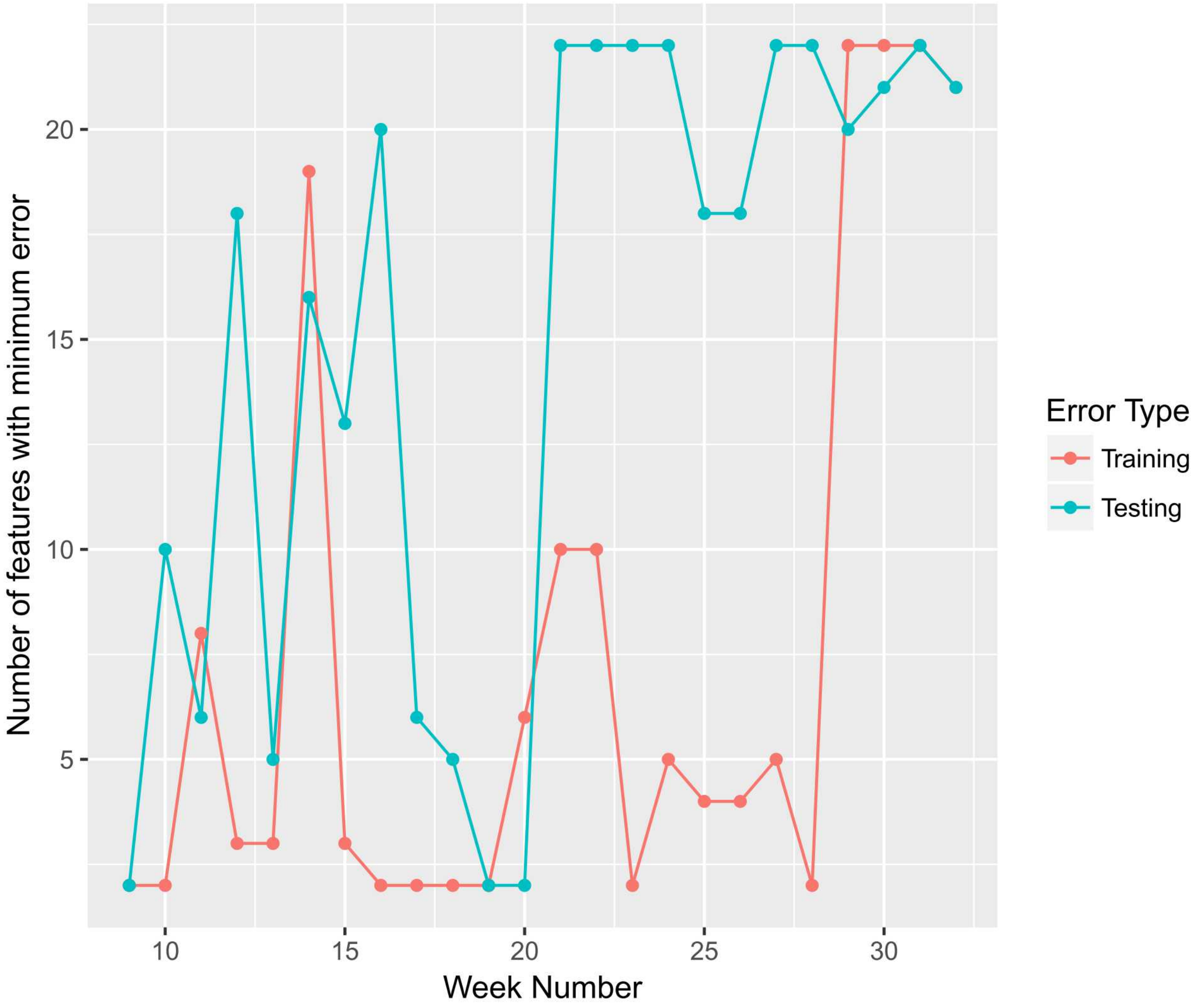


Figure 3.12: The number of features that resulted in minimum error for each week of the flu season.

3.3.2 Mutual Information (MII)

Mutual information (MI) is a concept from information theory that can also be used as criteria for feature ranking. The mutual information between two variables is the amount by which knowledge of one variable decreases uncertainty in the other. MI is defined as [5]:

$$I(X,Y) = \sum_{x\in X}\sum_{y\in Y} P(x,y)\log\frac{P(x,y)}{P(x)P(y)} \tag{3.5}$$

where $P(x,y)$ is the joint probability distribution, and $P(x)$ and $P(y)$ are the marginal probability distributions.

We implemented a greedy feature selection method that chooses new features by maximizing the MI with observations (i.e., relevancy of the feature) while minimizing MI with current features (i.e., redundancy of the feature). For each week of the flu season, we started with the feature that has the largest MI with the observations. Then we added one feature at a time by choosing the feature that maximizes this formula:

$$I(Y, f) - \beta \sum_{s \in S} I(s, f) \tag{3.6}$$

where $I()$ is the mutual information (see Eq. 3.5), Y is the observation, f is the new candidate feature, S is the subset of features already in use, and β is a parameter to control the trade-off between relevancy and redundancy. We used a β value of 0.5.

Using the optimization in Eq. 3.6, we ranked all the features for each week of the flu season, as shown in Fig. 3.13. Features were selected in order of rank starting with rank 1, so features with smaller ranks were selected first. No single feature is always selected first or last, but some features are consistently ranked better than others. For example, Delphi Epicast has a high ranking for most weeks, which is consistent with the fact that Delphi Epicast generally has large correlation with the observations (as shown in Fig. 3.10). But interestingly, some features that were shown to have low correlation with observations have a relatively high ranking, such as TeamB and 4Sight. This is because despite those features having low MI with the observations, they have even lower MI with the existing feature set and therefore might offer unique information for the prediction.

Fig. 3.14 shows the training and testing error averaged across all HHS regions and weeks of the flu season as a function of the number of features. The same general trends held for the MI criteria as the correlation criteria (shown in Fig. 3.11). The training error increased with number of features and testing error decreased. However, the testing error for the MI criteria dropped faster with fewer features than the correlation criteria, indicating that selecting features based on their diversity can help achieve better performance faster.

Fig. 3.15 shows the number of features that resulted in the stacking model with minimum error for each week of the flu season. Again, the same general trends held for the MI criteria as the correlation criteria (as shown in Fig. 3.12). The ideal number of features fluctuated early in the season and dropped near the flu season peak. However, toward the end of the season, the MI criteria tended to favor fewer features than the correlation criteria.

See Fig. 3.16 for a direct comparison of error between the MI and correlation criteria for feature selection. For each week of the flu season, we took the minimum error for each selection method across number of features. The difference in minimum error is shown. A difference below 0 indicates that the MI selection criteria performed better and a difference above 0 indicates that the correlation selection criteria performed better. For most weeks, the MI criteria had smaller training and testing error than the correlation method, although the difference is nearly negligible. This indicates that it was important to select features based on diversity.

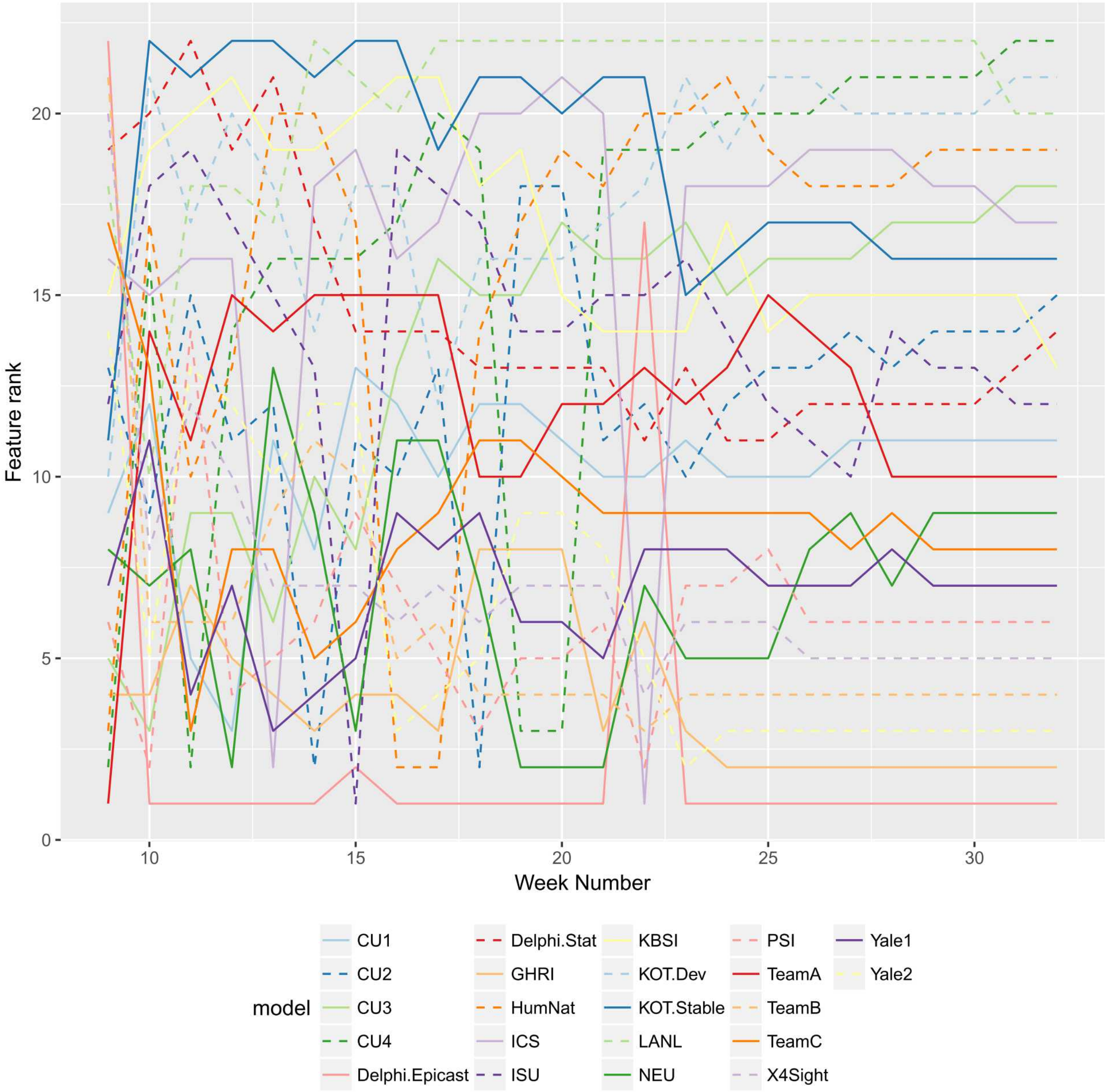


Figure 3.13: Ranking of the features for each week of the flu season using the mutual information (MI) criteria.

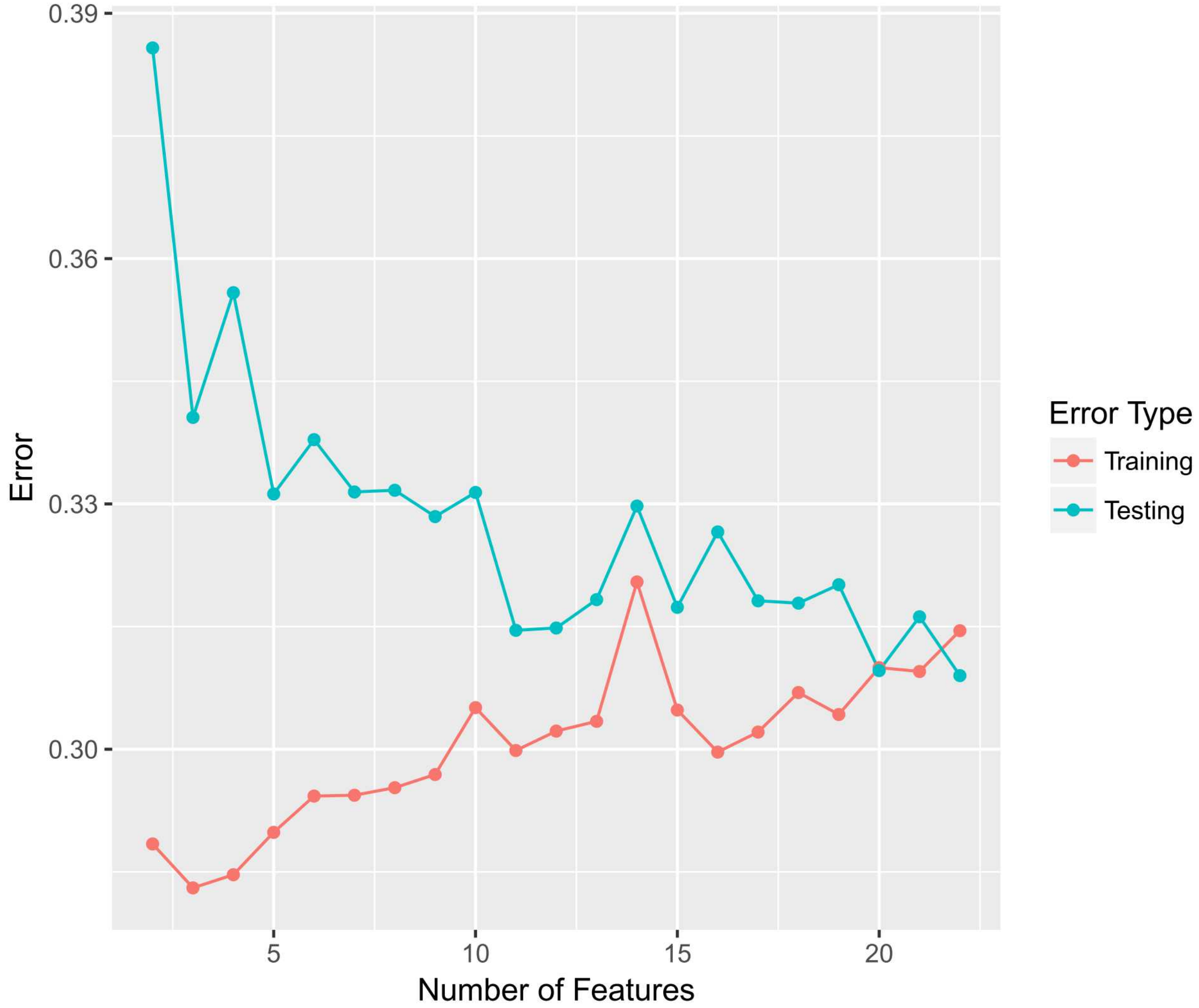


Figure 3.14: Training and testing error for the stacking model, averaged across all HHS regions and weeks, as a function of the number of features.

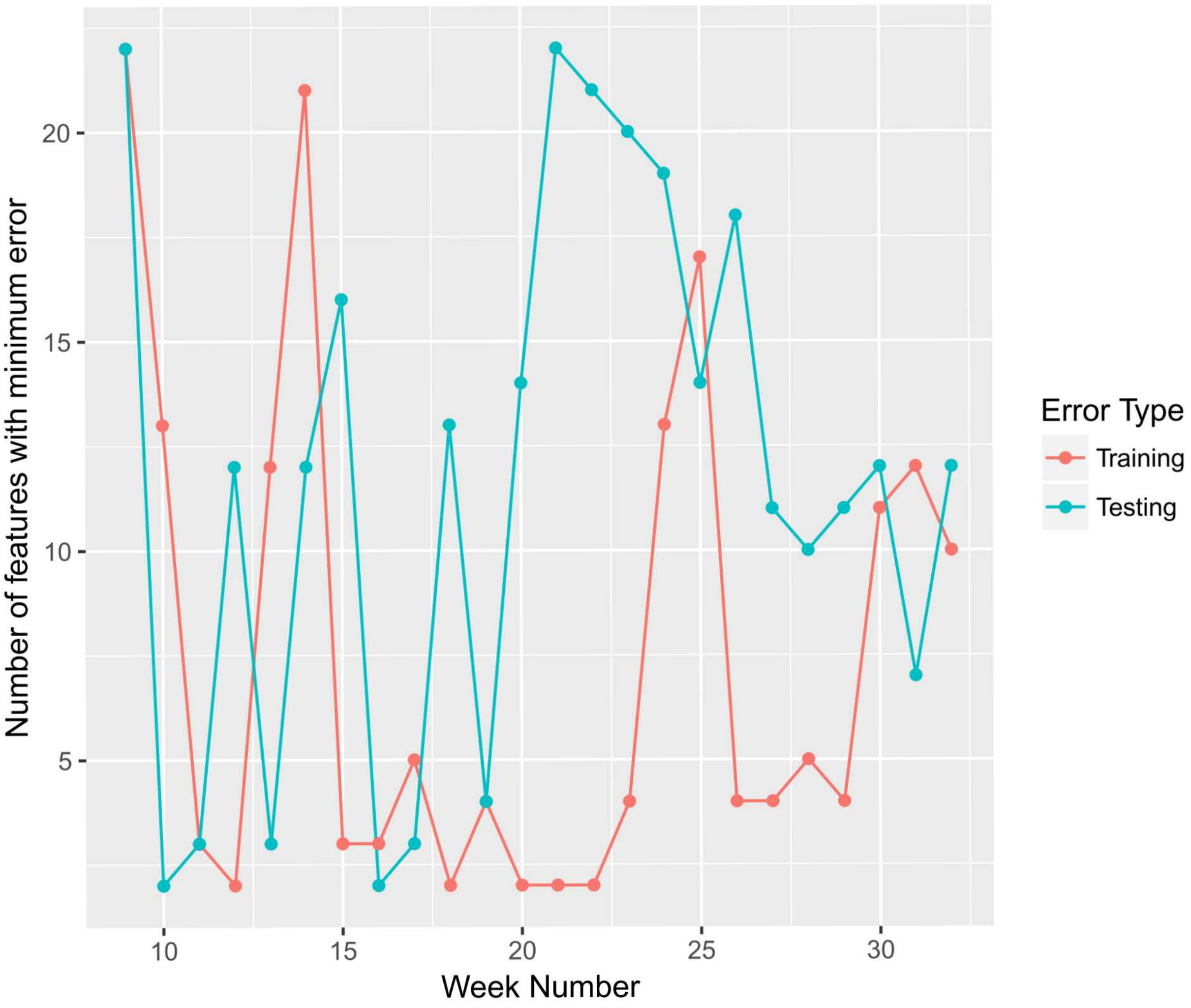


Figure 3.15: The number of features that resulted in minimum error for each week of the flu season.

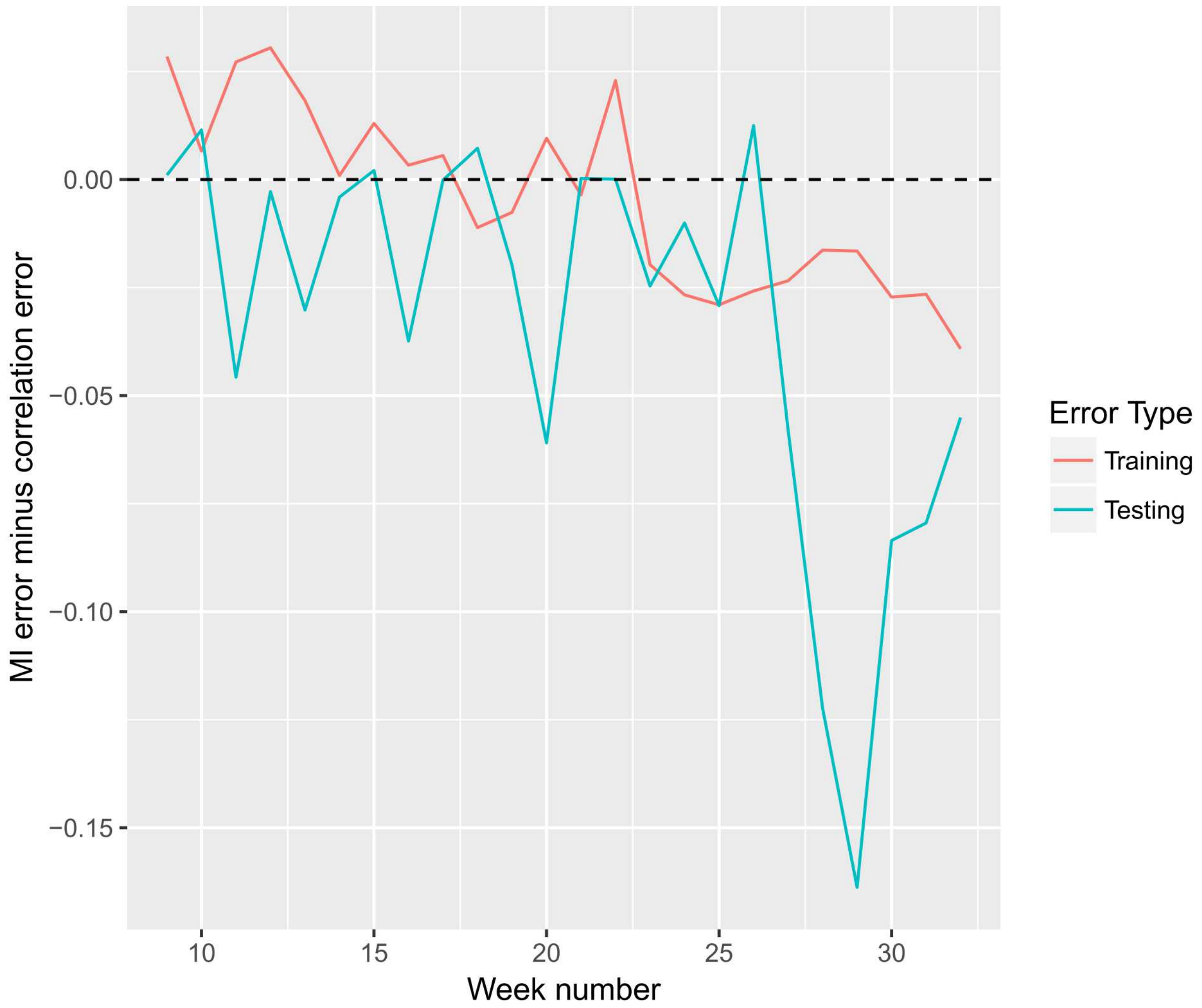


Figure 3.16: Difference in minimum error across number of features using the MI criteria and the correlation criteria for feature selection. Differences below 0 indicate that MI criteria performed better and differences above 0 indicate that correlation criteria performed better.

3.4 Performance Prediction using a Decision Tree

The ultimate goal of stacking and DMA is to improve predictive performance over the raw ensemble, where the raw ensemble is simply the average of the individual model predictions. In order to measure ensemble performance, we used the continuous rank probability score (CRPS) and absolute error as metrics. The CRPS is a measure of error between a probability distribution and a single data point, and is therefore an effective metric for comparing DMA to observations and the raw ensemble to observations. But because the stacking prediction is a single point and not a probability distribution, we used absolute error to compare stacking to observations.

Using CRPS and absolute error as metrics, either stacking or DMA has smaller error than the raw ensemble roughly 85% of the time during the 2016 – 2017 flu season. However, DMA and stacking are not better than the raw ensemble in the same weeks, indicating that neither method is universally better and instead, some kind of hybrid approach might have the best overall performance.

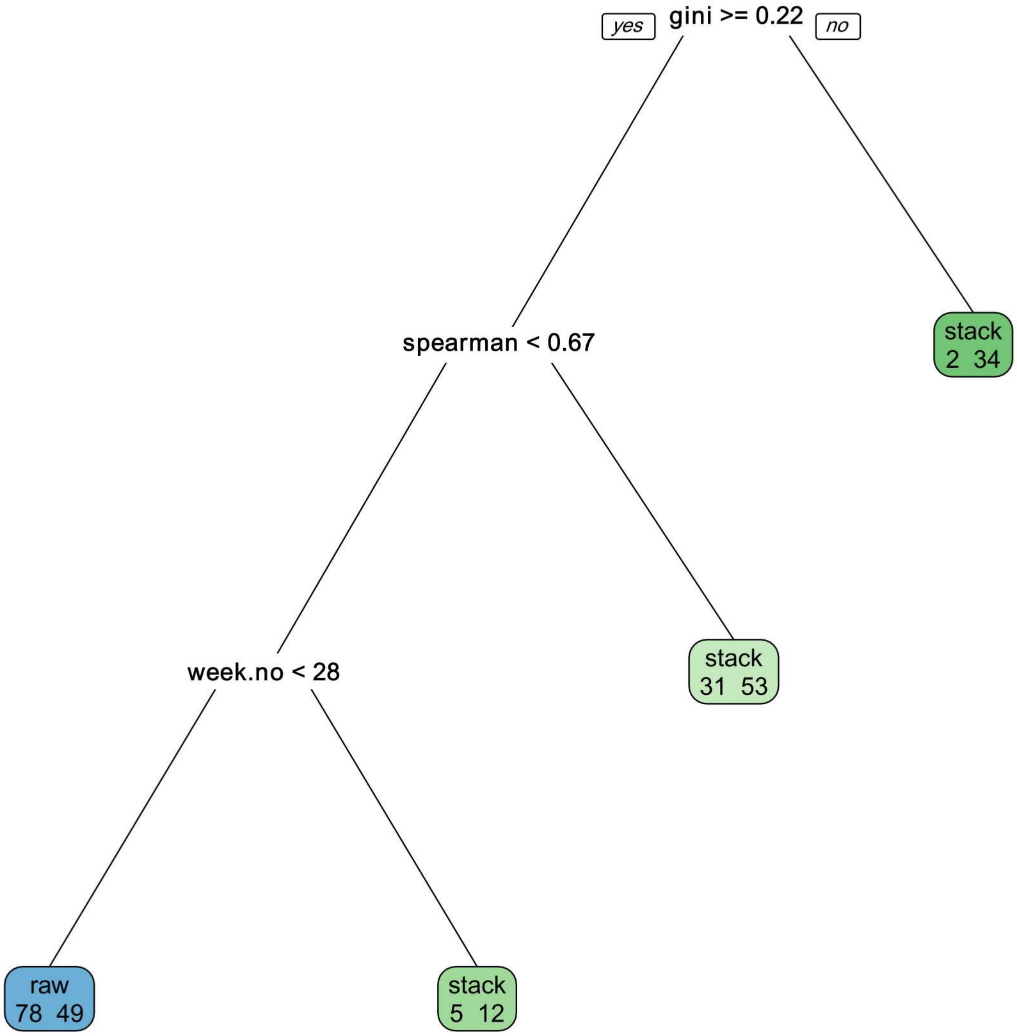
In order to predict which ensemble method will perform best in a given week, we trained three decision trees using the `rpart` package in `R` [28, 22]:

1. Predict whether stacking or the raw ensemble will perform better in a given week, aka, the stacking tree.
2. Predict whether DMA or the raw ensemble will perform better in a given week, aka, the DMA tree.
3. Predict whether stacking, DMA, or the raw ensemble will perform best in a given week, aka, the multi-class tree.

For features, we used characteristics of the raw ensemble rather than the stacking or DMA ensemble. This is because the raw ensemble will be computed for each week of the flu season and the decision tree can then be used to predict whether stacking or DMA should also be performed to improve performance. Specifically, the decision trees used the following features:

1. Gini coefficient (see Eq. 3.2) of the CRPS between the raw ensemble and observations. The Gini coefficient is a measure of dispersion in a set of numbers. A large Gini coefficient (near one) would indicate a small set of predictive models.
2. Spearman rank correlation coefficient (see Eq. 3.3) of the CRPS between the raw ensemble and observations. The Spearman rank is a measure of ordering over time, and a large Spearman rank (near one) would indicate that models have consistent performance week-to-week.
3. Week number of the flu season.
4. Standard deviation of the model predictions, where the standard deviation of all models were combined as variances, e.g., square root of the sum of squares.

The decision trees were trained using 3-fold cross-validation. We performed a sensitivity analysis over the depth of the trees and pruned them to the depth with minimum cross-validation error. The resulting decision trees are shown in Figures 3.17 for stacking, 3.18 for DMA, and 3.19 for multi-class.



yes
gini >= 0.22
no
spearman < 0.67
stack
2 34
week.no < 28
stack
31 53
raw
78 49
stack
5 12

Figure 3.17: Decision tree for predicting whether stacking or the raw ensemble will have smaller error in a given week of the flu season.

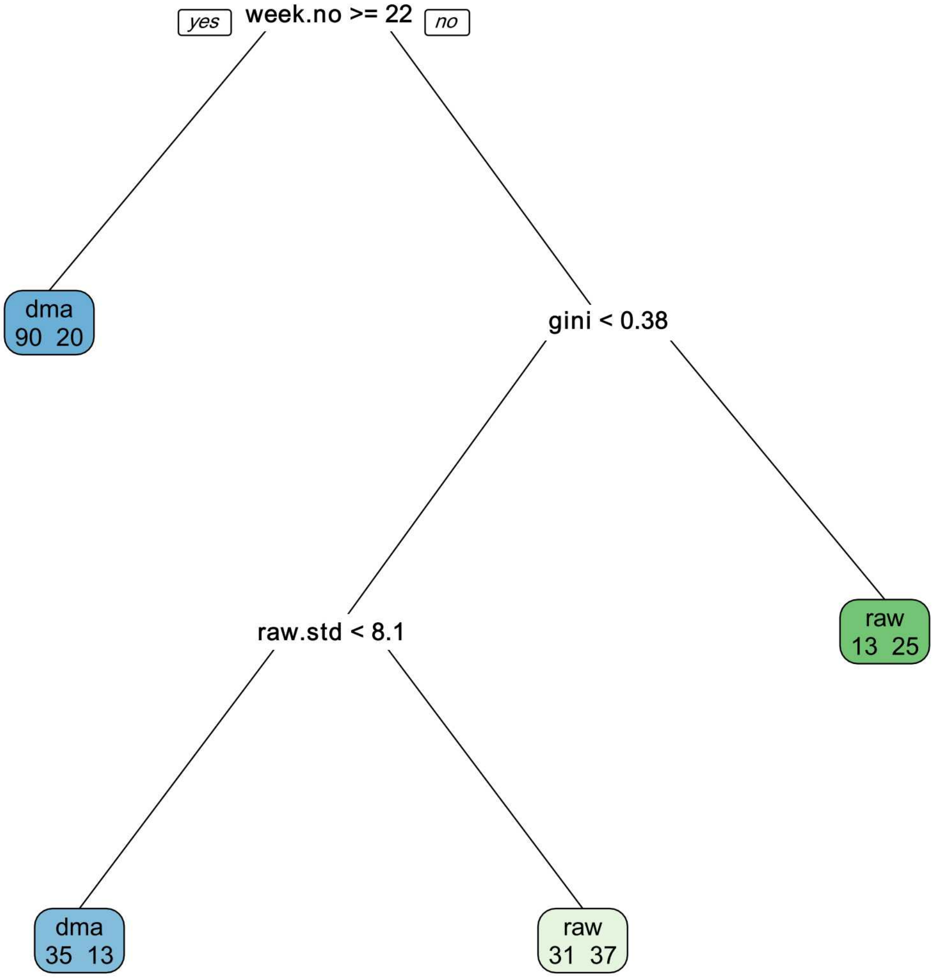

yes
week.no >= 22
no
dma
90 20
gini < 0.38
raw.std < 8.1
raw
13 25
dma
35 13
raw
31 37

Figure 3.18: Decision tree for predicting whether DMA or the raw ensemble will have smaller error in a given week of the flu season.

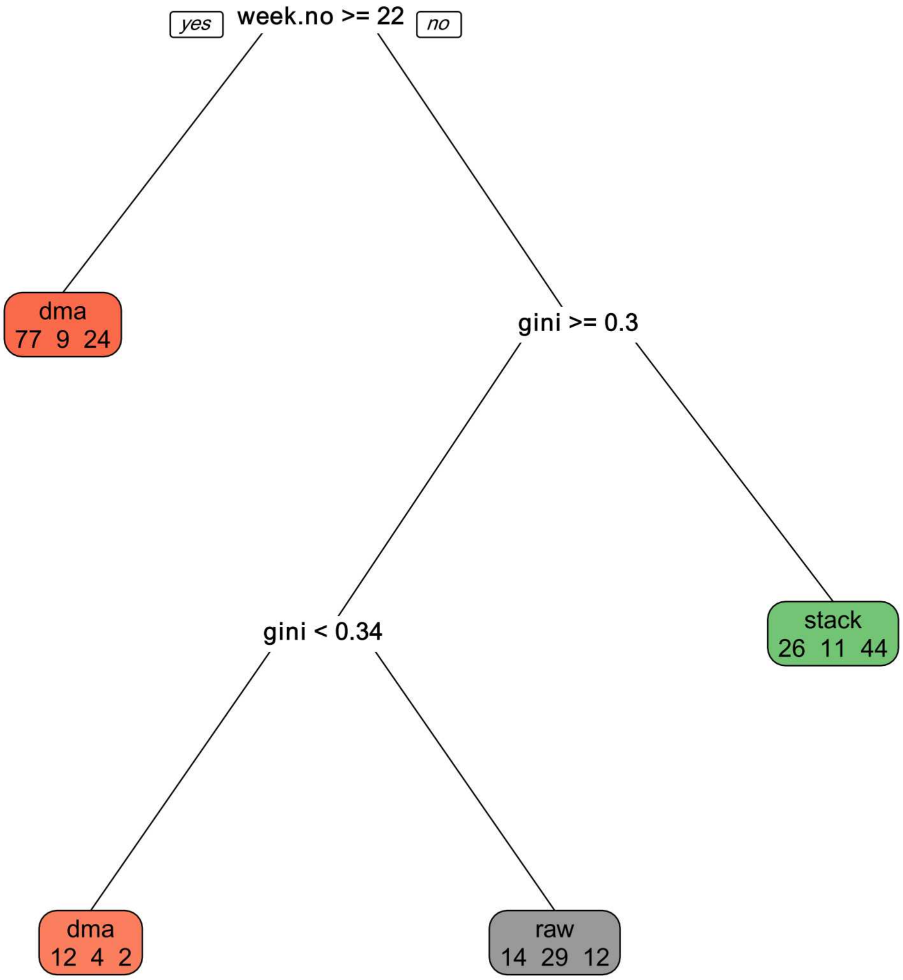

yes
week.no >= 22
no
dma
77 9 24
gini >= 0.3
gini < 0.34
stack
26 11 44
dma
12 4 2
raw
14 29 12

Figure 3.19: Decision tree for predicting whether stacking, DMA, or the raw ensemble will have smaller error in a given week of the flu season.

All 3 pruned decision trees contain the week of the flu season as a feature. Each case also predicted that stacking or DMA was more likely to perform better than the raw ensemble later in the flu season. This is likely because more data is available later in the flu season, leading to better training of stacking and DMA.

All 3 pruned decision trees also indicated that a small Gini coefficient leads to stacking, a large Gini coefficient leads to the raw ensemble, and somewhere in the middle leads to DMA. This was unexpected behavior because a large Gini coefficient indicates that a small group of highly-predictive models exists, and stacking or DMA could exploit that group by weighing them heavily. Instead, this result indicated that the raw ensemble performed best in this scenario, even though it gives an equal weight to all models.

While this result was unexpected, it may not be significant because predictions from the decision tree were not very accurate, as shown in the confusion matrices in Tables 3.4 for stacking, 3.5 for DMA, and 3.6 for multi-class. The average f1-score across classes was only 0.67 for stacking, 0.61 for DMA, and 0.58 for multi-class.

Table 3.4: Confusion matrix for the stacking decision tree.

	Prediction: Raw	Prediction: Stack
Actual: Raw	78	38
Actual: Stack	49	99

Table 3.5: Confusion matrix for the DMA decision tree.

	Prediction: Raw	Prediction: DMA
Actual: Raw	62	33
Actual: DMA	44	125

Table 3.6: Confusion matrix for the multi-class decision tree.

	Prediction: Raw	Prediction: DMA	Prediction: Stack
Actual: Raw	28	13	11
Actual: DMA	14	89	26
Actual: Stack	12	26	44

Using the current set of features, decision trees do not distinguish well between stacking, DMA, and the raw ensemble. More analysis is needed to investigate the current features and find more distinguishing features.

Chapter 4

Conclusions

In this study, we investigate methods that one could use to condition an ensemble of disease models to observational data, in a bid to improve its predictive skill. Outbreaks of diseases vary tremendously from instance to instance, and no individual model can capture this variability. In addition, models' predictive skills are inconsistent too, and they lose accuracy sporadically. Thus the individual models cannot be used for planning purposes. For the purposes of this study, we will use the annual influenza outbreak in the US as an exemplar, for reasons listed below.

The annual influenza outbreak in the US varies tremendously year-to-year, and there is no good influenza simulator that contains all the epidemiological processes that govern this phenomenon. To address this problem, the CDC has assembled an ensemble of influenza prediction models (called FluSight [11]) which, it is hoped, contain all the epidemiological processes of interest between all the models. Currently, the models are used to generate forecasts (which are archived [12]) which are then used as the ensemble prediction. Since the models are treated as equals, the uncertainty bounds on the predictions are unacceptably wide.We considered the ensemble for 2016-2017 and 2017-2018. The easily availability of model predictions and data allows us to concentrate on solving the conditioning problem, and were the primary reasons for choosing seasonal influenza as the exemplar problem.

The conventional solution of this problem is “model-averaging” (MA)’ or “stacking”. In MA, the forecasts (which are probability densities for the FluSight ensemble) are combined into a weighted mixture of densities. Typically, the model's forecasts are first converted into an approximate Gaussian density, so that the ensemble prediction is a weighted mixture of Gaussians. In stacking, we combine point forecasts from the models (mean or mode), in a linearly weighted manner, to obtain the forecast. The weights, in both cases, are computed using historical observational data. In our case, the observational data consists of CDC's FluView data, gathered weekly from its ILINet of sentinel physicians. The data consists of the percentage of physician visits where the patients exhibit ILI (influenza-like illness) symptoms.

The literature states that, in theory, MA (or stacking) should result in an ensemble that is more predictive than individual models. This requires one to have a diverse ensemble and lots of observational data to compute weights (Chapter 8 in Ref. [16]). In practice, its has been observed that MA delivers consistency to the forecasts, rather than accuracy; it is very possible that in a few instance a model will perform better, but never consistently so. Using an ensemble implies that one does not have to *select* a model; instead it is a risk-mitigation strategy at heart. Further, in case of influenza, the best model is not evident till very late in the outbreak, making model selection irrelevant for planning purposes. In addition, literature advises that if an ensemble can predict multiple Quantities of Interest (QoIs), it is best if they are MA individually.

The FluSight ensemble poses a few unusual challenges. It is far larger and far diverse than the disease ensembles that have been investigated in literature. In addition, the models are constantly under development and new ones are being added to the ensemble, implying that the archived forecasts cannot simply be concatenated together and used, with the historical CDC FluView archive, to compute weights with lots of observation data. Instead one must MA on a (influenza) season-by-season basis, using the ensemble that is specific to that season (during which the ensemble's composition and models remain constant). Thus one is faced with the prospect of MA a large ensemble with limited data that might be available during a season from CDC FluView. Thus it is not clear whether the MA methods used in literature will apply to the FluSight ensemble and whether the findings hold.

In this study we developed three new MA methods – an adaptation of dynamic model averaging (DMA), and two rank-based methods (simple ranking and rank aggregation) – and compared them with the conventional Bayesian Model Averaging (BMA) and raw ensemble (i.e., equally weighted) which have been used/studied to date. The new MA methods are designed to be usable when data is sparse, or specifically, when the number of weeks of CDC FluView data is smaller than the size of the ensemble (which has been the case after the 2015-2016 influenza season).The adaptation of DMA is a method that starts with an equally weighted ensemble and updates the weights by sequentially assimilating CDC FluView data. The ranking based methods attempt to find the three best models, *across all QoIs*, and then simply combine their forecasts (cast as Gaussian densities) using equal weights. They, too, can function with very large ensembles.

We also investigated, using elastic net regression and feature pruning, whether we could shrink the ensemble so that it could be conditioned using the sparse FluView data available. We also explored two opposing methods for selecting models to be retained in the shrunk ensemble. The first picks models based on the correlation between their predictions and FluView data. The second approach ensures model diversity in the shrunk ensemble. Our findings are:

- The MA predictions are more consistent than the forecasts from individual models. This agrees with literature.
- The DMA adaptation is better than the raw ensemble, indicating that it is a bona-fide MA method.
- The two ranking-based methods are almost equal in their predictive skill. However, one of the two algorithms (called rank aggregation) is extremely computationally expensive, without providing any advantage in ensemble's forecasting accuracy.
- The ranking-based "MA" algorithms are far more predictive than the others.
- When constructing the shrunk ensemble, it is better to choose models based on diversity.
- In the later part of the outbreak, when sufficient data led to models providing informative forecasts, the shrunk ensemble tended to be larger (as models provided diverse, rather than random, predictions).
- The predictive skill of the shrunk ensemble is similar to DMA. However, the weights of the models, and even the distribution of weights, is different.
- The BMA algorithm performed the worst. This was entirely due to the expectation-maximization (EM) algorithm used inside BMA to estimate model weights. In our case,

where we have less data than the number of models in the ensemble, EM is asked to solve an under-constrained estimation problem. EM was not designed to do so, and performs poorly.

We also encountered some features of DMA that are worth recording for posterity. We found that the weights estimated by DMA, for a give QoI, cannot be used to forecast other QOIs - the predictive skill is no different than that of a raw ensemble. In addition, the DMA algorithm has a free parameter, α that can be used to tune the temporal correlation of model weights; it is crucial for algorithmic stability when data is sparse and assimilated sequentially. We tuned α in the hope of improving DMA's forecasting skill significantly. While we did achieve some improvement, it was marginal. Our study was performed using the default value of $\alpha = 0.99$. We found that in certain cases, MA or stacking would provide worse predictions than the raw ensemble. We investigated whether certain characteristics of the raw ensemble could yield insight into when an ensemble should be conditioned on *sparse* data (note, it is *always* advantages to condition when observational data is abundant). The results were inconclusive.

Looking forward, the study could be improved, or at least made more competitive, by enhancing BMA. BMA is a "batch" method and has the potential to be more predictive than DMA. However, to do so, one would have to replace the EM algorithm that BMA uses with one that can accommodate under-constrained problems. The *degenerate* EM algorithm [19] is one such candidate.

References

[1] M. Biggerstaff, D. Alper, M. Dredze, S. Fox, I. C-H. Fung, K. S. Hickmann, B. Lewis, R. Rosenfeld, J. Shaman, M-H. Tsou, P. Velardi, A. Vespignani, L. Finelli, and for the Ifluenza Forecasting Contest Working Group. *BMC Infectious Diseases*, 16(1):357–347, 2016. 20

[2] L. C. Brooks, D. C. Farrow, S. Hyun, R. J. Tibshirani, and R. Rosenfeld. Nonmechanistic forecasts of seasonal influenza with iterative one-week-ahead distributions. *PLOS Computational Biology*, 14(6):1–29, 06 2018. 19

[3] A. L Buczak, B. Baugher, L. J Moniz, T. Bagley, S. M. Babin, and E. Guven. Ensemble method for dengue prediction. *PLOS ONE*, 13(1):1–23, 01 2018. 19

[4] Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers and Electrical Engineering*, 40:16–28, 2014. 51

[5] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. New York: Wiley, 1991. 54

[6] P.M. Dixon, J. Weiner, T. Mitchell-Olds, and R. Woodley. Bootstrapping the Gini coefficient of inequality. *Ecology*, 68:1548–1551, 1987. 41

[7] Saso Dzeroski and Bernard Zenko. Is combining classifiers with stacking better than selecting the best one? *Machine Learning*, 54:255–273, 2004. 39

[8] Geir Evensen. *Data assimilation*. Springer-Verlag, 2007. 15

[9] Centers for Disease Control and Prevention. CDC Flu View. `https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html`. 20

[10] Centers for Disease Control and Prevention. Weekly influenza surveillance report. `https://www.cdc.gov/flu/weekly/`. 20

[11] Centers for Disease Control and Prevention. Epidemic prediction initiative. `https://predict.cdc.gov`, 2018. 20, 65

[12] Centers for Disease Control and Prevention. Github repository of cdc influenza forecasting contest model predictions. `https://github.com/cdcepi/FluSight-forecasts/`, 2018. 20, 65

[13] Centers for Disease Control and Prevention. Overview of influenza surveillance in the united states. `https://www.cdc.gov/flu/weekly/overview.htm`, 2018. 20

[14] Centers for Disease Control and Prevention. Regional baseline values for influenza-like illness. `https://github.com/cdcepi/FluSight-forecasts/wILI_Baseline.csv`, 2018. 20

[15] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. 39

[16] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Spinger, 2009. 15, 16, 17, 18, 65

[17] H. Hersbach. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5):559–570, 2000. 22

[18] Jennifer A. Hoeting, David Madigan, Adrian E. Raftery, and Chris T. Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–401, 1999. 17

[19] X. Lin and Y. Zhu. Degenerate expectation-maximization algorithm for local dimension reduction. In D. Banks, F. R. McMorris, P. Arabie, and W. Gaul, editors, *Classification, Clustering, and Data Mining Applications. Studies in Classification, Data Analysis, and Knowledge Organisation*. Springer, Berlin, Heidelberg, 2004. 67

[20] I. Yahav T. Lotze and G. Shmueli. *Algorithm Combination for Improved Performance in Biosurveillance*, pages 173–189. Springer US, Boston, MA, 2011. 19

[21] V. Pihur, S. Datta, and S. Datta. Weighted rank aggregation of cluster validation measures: a monte carlo cross-entropy approach. *Bioinformatics*, 23(13):1607–1615, 2007. 18, 30

[22] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. 39, 60

[23] A. E. Raftery, T. Gneiting, F. Balabdaoui, and M. Polakowski. Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133:1155–1174, 2005. 16, 17, 27, 28

[24] A. E. Raftery, M. Karny, and P. Ettler. Online prediction under model uncertainty via dynamics model averaging: Application to a cold rolling mill. *Technometrics*, 52(1):52–66, 2010. 16, 17, 21, 22

[25] E. L. Ray and G. Reich N. Prediction of infectious disease epidemics via weighted density ensembles. *PLOS Computational Biology*, 14(2):1–23, 02 2018. 19

[26] Nicholas G Reich, Logan Brooks, Spencer Fox, Sasikiran Kandula, Craig McGowan, Evan Moore, Dave Osthus, Evan L Ray, Abhinav Tushar, Teresa Yamana, Matthew Biggerstaff, Michael A Johansson, Roni Rosenfeld, and Jeffrey Shaman. Forecasting seasonal influenza in the u.s.: A collaborative multi-year, multi-model assessment of forecast performance. *bioRxiv*, 2018. 20

[27] C. R. SEBRANGO-RODRIGUEZ, D. A. MARTINEZ-BELLO, L. SANCHEZ-VALDES, P. J. THILAKARATHNE, E. DEL FAVA, P. VAN DER STUYFT, A. LOPEZ-QUILEZ, and Z. SHKEDY. Real-time parameter estimation of zika outbreaks using model averaging. *Epidemiology and Infection*, 145(11):2313–2323, 2017. 18

[28] Terry Therneau and Beth Atkinson. *rpart: Recursive Partitioning and Regression Trees*, 2018. R package version 4.1-13. 60

[29] T. K Yamana, S. Kandula, and J. Shaman. Superensemble forecasts of dengue outbreaks. *Journal of the Royal Society, Interface*, 13(123), October 2016. 19

[30] T. K Yamanaa, S. Kandula, and J. Shaman. Individual versus superensemble forecasts of seasonal influenza outbreaks in the united states. *PLOS Computational Biology*, 13(11):1–17, 11 2017. 19

DISTRIBUTION:

1 Dr. Christopher Kiley, 8725 John J Kingman Road, Fort Belvoir, VA 22060-6201
1 Dr. Lynne Burks, One Concern, Inc, 169 University Ave., Palo Alto, CA 94301

1	MS 9152	Jaideep Ray, 08759
1	MS 1137	Katherine Cauthen, 08722
1	MS 0828	Sophia Lefantzi, 01516
1	MS 0899	Technical Library, 8944 (electronic copy)

Sandia National Laboratories