An Ensemble Method for Disease Surveillance

Published Version:
Vesely, A., Roli, G., Scagliarini, M., Miglio, R. (2025). An Ensemble Method for Disease Surveillance. Springer Cham [10.1007/978-3-031-64447-4_110].

Availability:
This version is available at: https://hdl.handle.net/11585/1004190 since: 2025-02-05

Published:
DOI: http://doi.org/10.1007/978-3-031-64447-4_110

An ensemble method for disease surveillance

Anna Vesely, Giulia Roli, Michele Scagliarini, and Rossella Miglio

Department of Statistical Sciences, University of Bologna, Bologna, Italy

Abstract. Healthcare heavily relies on effective disease surveillance, particularly during epidemiological emergencies. Predicting trends is challenging but crucial for effective policies. We propose an ensemble modelling framework to forecast time series of counts, such as the number of cases or hospitalizations. The method relies on multiple models, including flexible regressions with cubic splines, generalized additive models, Bayesian structural time series, and integer-valued generalized autoregressive conditional heteroscedasticity models. Predictions are computed as weighted averages on some of the individual models' predictions, where model selection and weights are based on performance on past observations. This approach allows to exploit diverse model strengths and compensate for individual weaknesses, ensuring more stable predictions. The behavior of the proposed method is illustrated on daily COVID-19 hospitalization counts in Emilia-Romagna in 2022, with a focus on periods showing changing trends.

Keywords: disease surveillance, ensemble, forecasting, COVID-19

1 Introduction

Disease surveillance is a crucial task in healthcare and life sciences, especially during periods of a health emergency, such as those occurred in the recent pandemic. A good surveillance system supports decisions towards an optimal allocation of human, technical and economic resources, allows the development of better and innovative health policies, and prevents critical situations during waves of emergency. Under this framework, forecasting trends is a challenging objective, as it can anticipate a potential change and allows to support better decision-making processes towards early and more effective health policy solutions.

In this work, we consider the problem of predicting time series of counts, such as the number of cases or hospitalizations. Several works on infectious diseases and other fields showed that ensemble approaches, which exploit and combine predictions from multiple models, are often more accurate than any forecast from a single model (see, e.g., [2]). Ensemble methods combine the strengths of different models, while compensating for the weaknesses of individual models and producing more robust predictions. Indeed, if one model performs poorly in a certain scenario, its impact can be mitigated by the others in the ensemble.

We propose a general framework for an ensemble modelling approach, which provides predictions computed as a weighted average of different models' predictions. Candidate models are first analyzed in terms of performance on past observations, i.e., on how precisely they predict the last available observations. The accuracy of this prediction determines if and with which weight the model contributes to the ensemble. The method allows for any suitable choice of the prediction window and to include eventual covariates.

The behavior of the proposed approach is examined in the study of daily counts of COVID-19 hospitalizations in Emilia-Romagna. This analysis is the result of a long experience conducted on these data in collaboration with the Local Health Agency of Bologna, which suggested the importance of daily predictions based on a weekly frequency due to the organisational, managerial and executive set-up of hospitals. The effectiveness of an ensemble approach in this setting was illustrated in [6].

The manuscript is organized as follows. We give the general structure of the proposed ensemble approach in Sect. 2, then we introduce the contributing models in Sect. 3. The application on the COVID-19 data is shown in Sect. 4.

2 Ensemble model

Suppose that we are interested in studying a count time series $y_t \in \mathbb{N}$, with $t = 1, \ldots, T$, possibly in presence of a vector of covariates $\mathbf{x}_t \in \mathbb{R}^n$. The goal of the analysis is forecasting the next Q counts, i.e., computing predictions $\hat{y}_t$ for $t = T+1, \ldots, T+Q$. We assume that the predictions are calculated using data from the last W available time points ($t = T-W+1, \ldots, T$), where W is chosen according to the nature of the problem.

Each count y_t may be regarded as the realization of a Poisson random variable Y_t with parameter μ_t, where μ_t can be modelled in different ways. We propose an ensemble approach that takes into account multiple models and determines the desired predictions $\hat{y}_t$ as weighted averages of the individual models' predictions. The main idea is to fit a number K' of reasonable models, select $K \leq K'$ of them based on their performance on past observations, and finally compute

$$\hat{y}_t = \sum_{k=1}^{K} w_k \hat{y}_t^{(k)} \qquad (t = T+1, \ldots, T+Q) \tag{1}$$

where $\hat{y}_t^{(k)}$ is the prediction given by the k-th model, and $w_1, \ldots, w_K \in [0, 1]$ are some weights such that $\sum w_k = 1$.

The K selected models and the weights w_k are determined by shifting the focus back by Q time points. This way, each individual model is used to predict the last Q available counts ($t = T-Q+1, \ldots, T$), and then these predictions are compared to the observed data. Higher weights are given to the models that make the most accurate predictions, while those with the worst performance are discarded and do not contribute to the ensemble. The pipeline of the proposed analysis is as follows.

1. **Fitting.** Fit all K' candidate models on the last W data points ($t = T - W + 1, \ldots, T$). For any model choice, if different settings (e.g., pre-specified parameters) are possible, select the settings that lead to the lowest Bayesian Information Criterion (BIC) when appropriate.
2. **Selection.** Consider the last Q available counts. Fit all K' models (with the eventual settings as determined in the previous point) on $t = T - W - Q + 1, \ldots, T - Q$. For each model, compute the predictions $\hat{y}_t^{(k)}$, $t = T - Q + 1, \ldots, T$, and compare them to the observed counts y_t through the Mean Squared Deviation (MSD). Keep the K models that lead to the lowest MSD.
3. **Weight calibration.** Consider the last Q available counts and the selected models. Determine the weights that minimize the sum of squared differences between the ensemble predictions $\hat{y}_t$ and y_t, $t = T - Q + 1, \ldots, T$. The minimization may be carried out through the Augmented Lagrangian method for equality constraints, implemented in the `R` package `nloptr`.
4. **Prediction.** Apply the selection from 2. and the weights from 3. to the models in 1. to calculate the desired predictions $\hat{y}_t$, $t = T, \ldots, T + Q$.

Confidence intervals can be computed analogously to the ensemble predictions in (1), taking a weighted average of the bounds given by the selected models.

3 Contributing models

In our proposal, we take into account a set of $K' = 7$ candidate models, $K = 5$ of which are then selected for the ensemble. The first, which we will refer to as Flexible [1], is a Generalized Linear Model (GLM) where the logarithm of the Poisson parameter μ_t is taken as a function of the covariates $\mathbf{x}_t$ and restricted cubic splines of the time t. The GLM and the splines function can be implemented using the `R` packages `stats` and `rms`, respectively.

The second model that is considered is the Generalized Additive Model (GAM) [5], which is similar to the Flexible but allows for a more general smooth function of time. It is implemented in the `mgcv` package.

Subsequently, we take the Bayesian Structural Time Series (BSTS) model [7], available in the package `bsts`. This method combines Kalman filtering to structure the time series, spike-and-slab regression to select regression components, and Bayesian model average to convert samples from repeated Markov Chain Monte Carlo (MCMC) draws into the final outputs. The model takes into account the covariates $\mathbf{x}_t$ and an autoregressive component.

The final block of models is identified in the INteger-valued Generalized AutoRegressive Conditional Heteroscedasticity (INGARCH) models [3, 4], where the parameter μ_t depends on the covariates $\mathbf{x}_t$, past observations $y_{t-1}, \ldots, y_{t-p}$, past expectations $\mu_{t-1}, \ldots, \mu_{t-q}$, and powers of time $t^0, t^1, \ldots, t^r$. In particular, we choose four models, having $r = 0$ (Stationary), $r = 1$ (Linear), $r = 2$ (Quadratic), and $r = 3$ (Cubic). This family of models is implemented in the `tscount` package. If no prior information is available, the parameters corresponding to autoregression p and moving average q can be chosen as suggested in point

1. of the pipeline, considering all possible combinations of $p, q \in \{1, \ldots, Q\}$ and identifying the lowest BIC.

4 Application to COVID-19 hospitalizations

We examine the performance of the proposed ensemble method on data from the Local Health Agency on the daily number of hospitalizations for COVID-19, available at https://github.com/pcm-dpc/COVID-19. In particular, we focus on the non-intensive care unit (non-ICU) hospitalizations in Emilia-Romagna in 2022. The goal of the analysis is forecasting the counts on one week ($Q = 7$), using data from the previous three weeks ($W = 21$). As covariate, we take a factor variable that describes the day of the week (Sunday, Monday, etc.); holidays are coded as Sundays.

In disease surveillance interest usually lies in quickly detecting a change of trend, therefore we investigate if the method is able to catch such changes by comparing its predictions with the observed counts in four given periods. We examine two weeks where the number of hospitalizations starts to increase (Period 1: March 22-28, Period 3: June 21-27) or decreases after a peak (Period 2: May 3-9, Period 4: August 2-8). Fig. 1 shows the time series under analysis, highlighting the considered time periods.

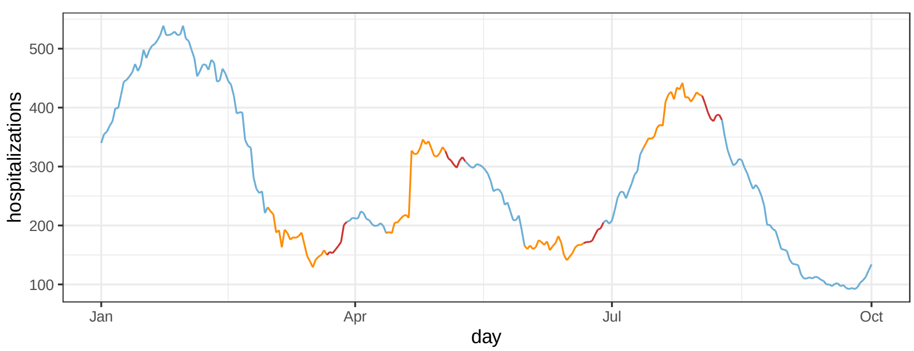

Fig. 1. Number of daily non-ICU hospitalizations for COVID-19 in Emilia-Romagna in 2022. The analysis focuses on four time periods: for each, the data from three weeks (orange) are employed to forecast the number of hospitalizations in the next week (red).

Predictions given by the ensemble method, as well as the $K = 5$ models contributing to it, are shown in Fig. 2. Furthermore, Table 1 displays the MSD used in the model selection step (i.e., shifting the focus back by one week) and the MSD computed on the prediction week of interest. We observe that the performance of individual models and the selection for the ensemble vary significantly over the four considered periods. However, the GAM is always selected and tends to provide accurate predictions. The ensemble predictions are, by construction, an average of the others; this mitigates extreme behaviours (see, e.g., the INGARCH cubic in Period 3) and tends to produce more stable results. The performance is particularly promising in Periods 2 and 3, while the method is not able to capture the newly increasing trend in Period 1.

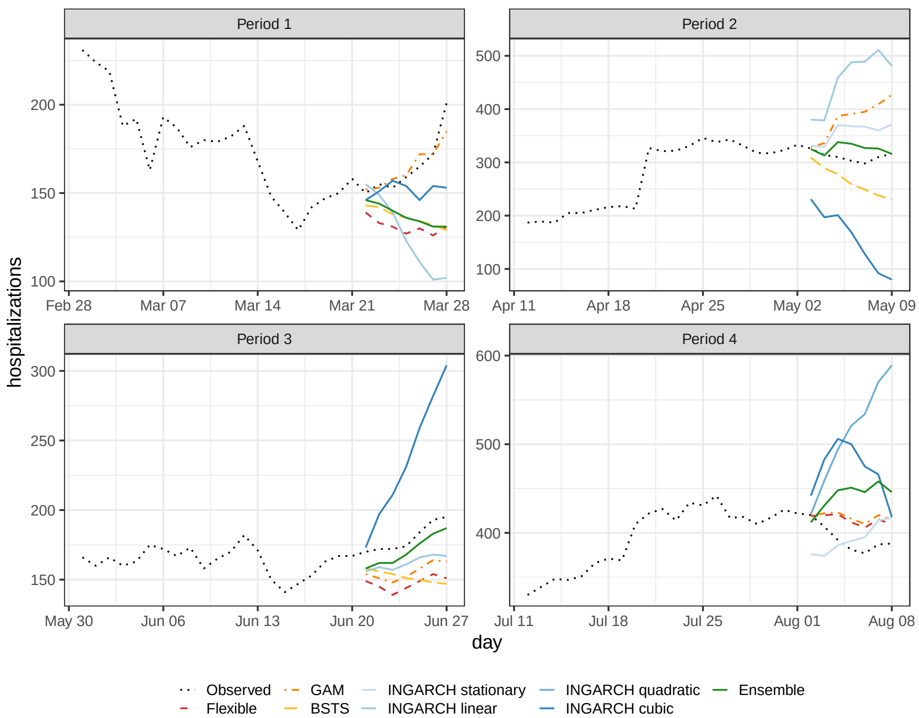

Fig. 2. Number of hospitalizations in four time periods: observed time series over four weeks, and predictions from different models in the last week. For each period, only the proposed ensemble model and the five individual models that contribute to the ensemble are displayed.

Future analyses may consider additional covariates other than the day of the week, such as the temperature and levels of air pollutants. Moreover, further research is needed on the comparison between the individual models and the ensemble. A high variability, e.g., expressed in terms of Coefficient of Variation (CV), may identify a critical situation, where the single models tend to be discordant, and thus the predictions become unstable, uncertain and less reliable. Preliminary analyses show that the highest CV values are obtained when the

Table 1. MSD between observed data and predictions of different models, in four time periods, for the selection step and the prediction. For each period, the two models with the highest selection MSD (in gray) are discarded.

	Period 1		Period 2		Period 3		Period 4	
Model	sel.	pred.	sel.	pred.	sel.	pred.	sel.	pred.
Flexible	23.2	23.2	59.2	10.4	18.7	18.6	14.8	6.2
GAM	22.0	4.2	21.7	26.0	9.9	13.8	17.6	7.0
BSTS	11.6	21.4	17.4	16.7	9.2	17.2	23.1	13.2
I. stationary	65.7	13.0	17.8	16.5	22.3	4.4	17.5	6.8
I. linear	15.2	31.9	6.3	49.7	12.9	10.4	20.1	39.8
I. quadratic	50.8	10.9	99.3	6.9	22.9	9.0	13.9	34.9
I. cubic	46.7	12.8	35.3	52.3	7.3	36.9	3.7	21.5
Ensemble	13.4	21.0	4.1	6.6	8.0	5.2	2.4	14.3

trends of occupancies change in the evolution of the phenomenon. Therefore, these CVs can be employed to detect time location of change points.

5 Acknowledgements

This research is co-funded by the Ministry of University and Research within the Complementary National Plan PNC-I.1 “Research initiatives for innovative technologies and pathways in the health and welfare sector”.

References

1. Durrleman, S., Richard, S.: Flexible regression models with cubic splines. Stat. Med. 8, 551–561 (1989). https://doi.org/10.1002/sim.4780080504
2. Farcomeni, A., Maruotti, A., Divino, F., Jona-Lasinio, G., Lovison, G.: An ensemble approach to short-term forecast of COVID-19 intensive care occupancy in Italian regions. Biom. J. 63, 503–513 (2021). https://doi.org/10.1002/bimj.202000189
3. Fokianos, K.: Some recent progress in count time series. Statistics 45, 49–58 (2011). https://doi.org/10.1080/02331888.2010.541250
4. Fokianos, K., Rahbek, A., Tjostheim, D.: Poisson autoregression. J. Am. Stat. Assoc. 104, 1430–1439 (2009). https://doi.org/10.1198/jasa.2009.tm08270
5. Hastie, T., Tibshirani, R.: Generalized additive models: some applications. J. Am. Stat. Assoc. 82, 371–386 (2014). https://doi.org/10.1080/01621459.1987.10478440
6. Miglio, M., Roli, G., Scagliarini, M.: A joint use of monitoring and forecasting methods to detect change points in daily hospitalizations. In: Programme and Abstracts, 24th International Conference on Computational Statistics, COMPSTAT 2022, pp. 22–23 (2022). http://www.compstat2022.org
7. Scott, S., Varian, H.: Predicting the present with Bayesian structural time series. Int. J. Math. Model. Numer. Optim. 5 (2014). https://doi.org/10.1504/IJMMNO.2014.059942