Ebola Forecasting - Error Analysis

Frederic Schoenberg, Sarita Lee, Andy Shen

1 Data Input and Cleaning

We assume the most accurate dataset is the most recent dataset of the outbreak. We tally the cases such that there is a running total of infections at each date. This dataset represents the true number of cases at any given point during the outbreak.

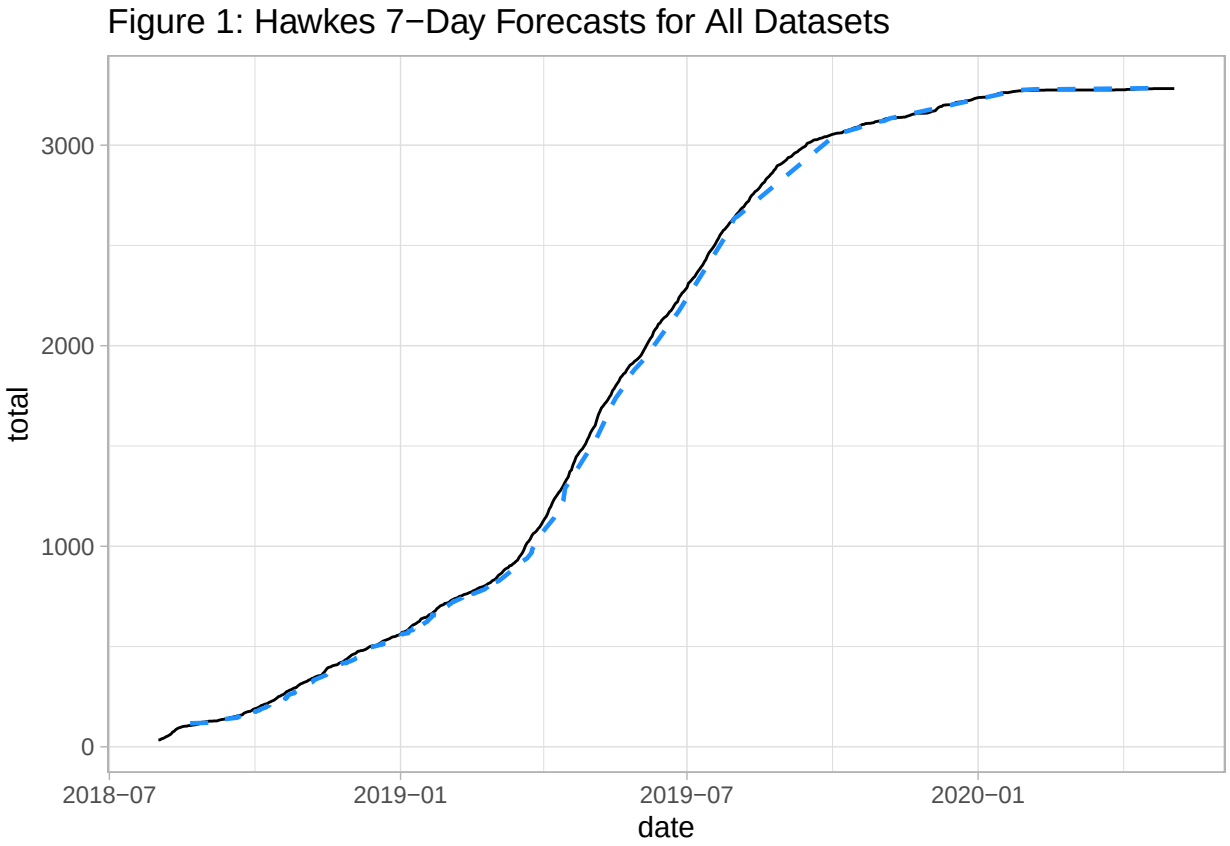
We then import the projections from the Hawkes and Recursive models. For these predictions, the date preceding the forecasts is the last date of that dataset with at least one case. The forecasted numbers then predict the additional number of infections 7, 14, and 21 days after that date, respectively.

Some dates in the Hawkes forecast models do not have a corresponding Recursive model forecast, so we omit those values from our analysis.

2 Hawkes Complete Outbreak Analysis

2.1 7-Day Forecast Analysis

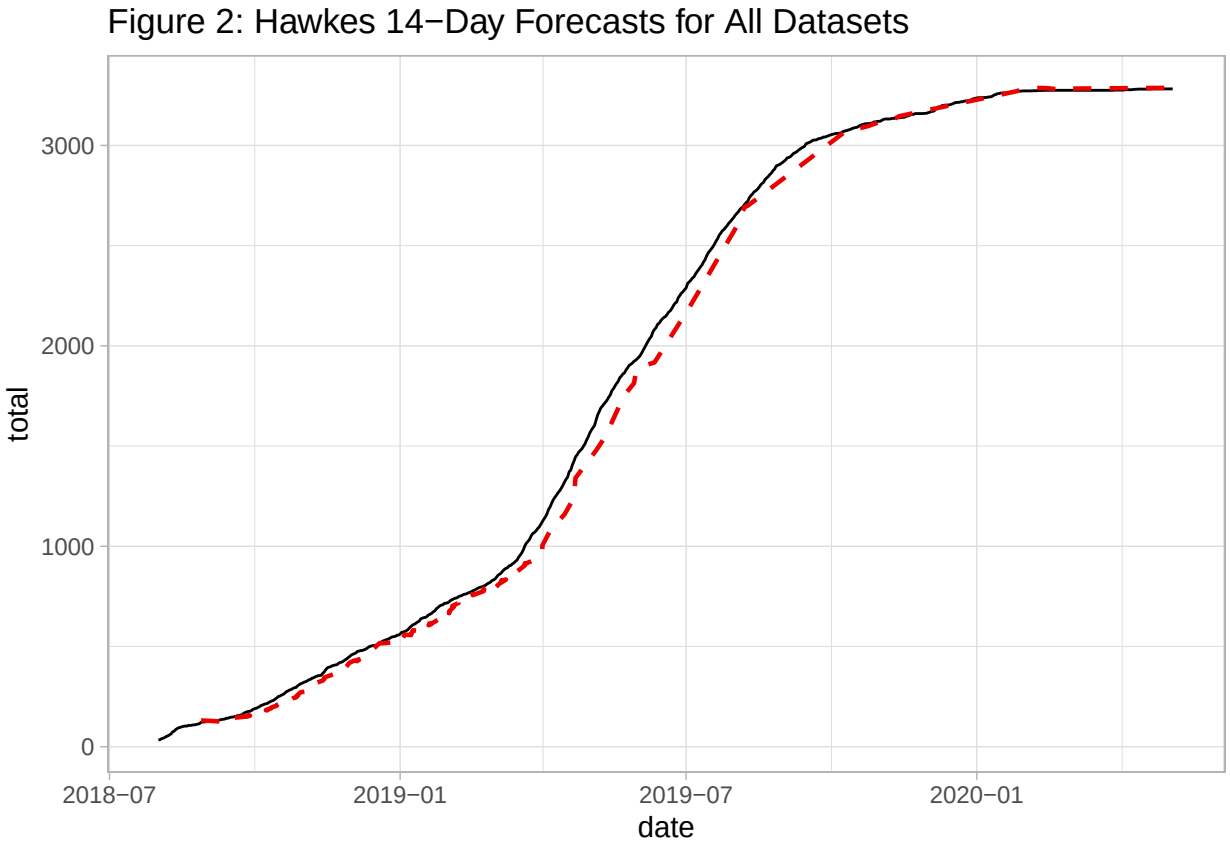
Figure 1 below shows the Hawkes 7-Day Forecasts for all recorded simulations with respect to the recorded number of infections. The RMSE values for all forecasts are included in Table 1.



The 7-day Hawkes model generally shows accurate predictions throughout the duration of the pandemic, with slight under-prediction in 2019 during the middle of the pandemic. The Hawkes 7-day forecasts appear to predict the case counts at the beginning and end of the pandemic quite accurately.

2.2 14-Day Forecast Analysis

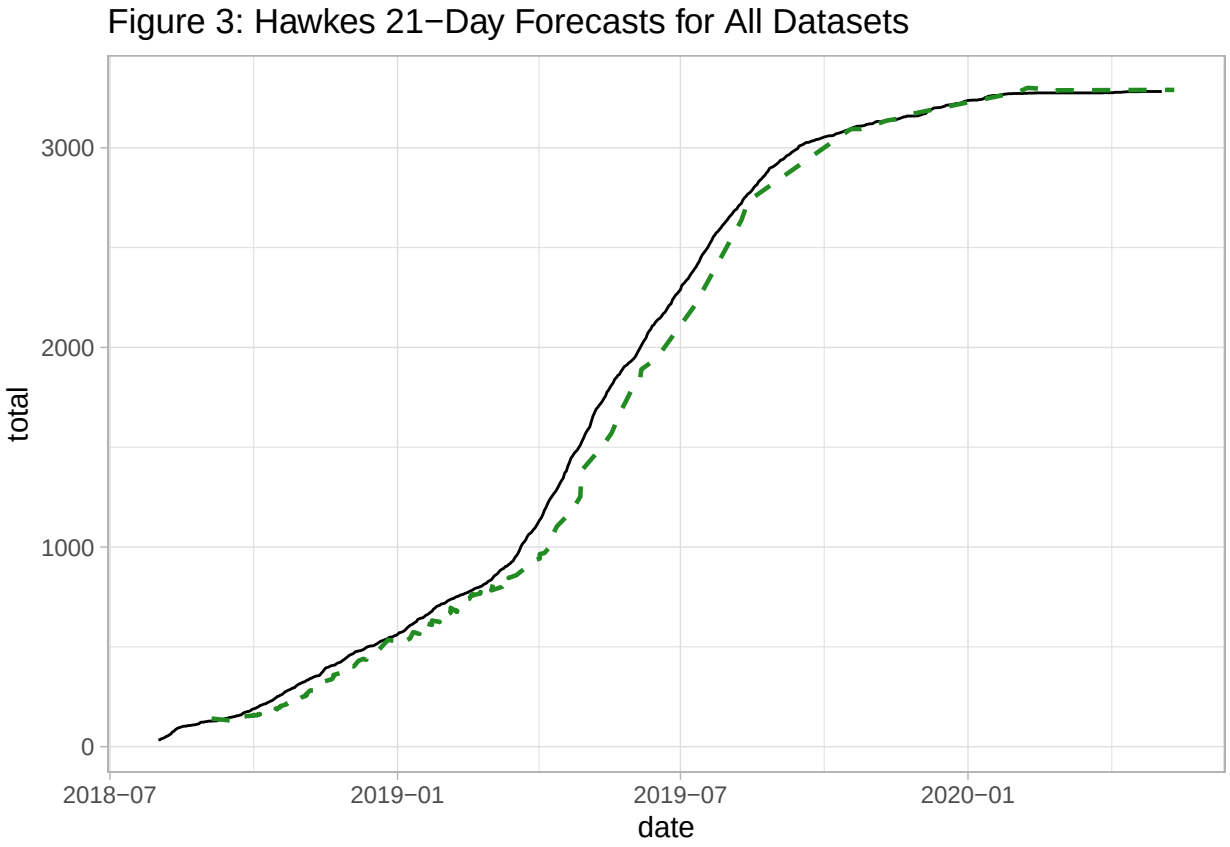
Figure 2 below shows the Hawkes 14-Day Forecasts for all recorded simulations with respect to the recorded number of infections.



The 14-day Hawkes predictions tend to slightly under-predict the true case counts in the beginning and middle portions of the pandemic, with more accurate prediction in 2020 as the pandemic comes to an end. The largest prediction discrepancy is during 2019.

2.3 21-Day Forecast Analysis

Figure 3 below shows the Hawkes 21-Day Forecasts for all recorded simulations with respect to the recorded number of infections.

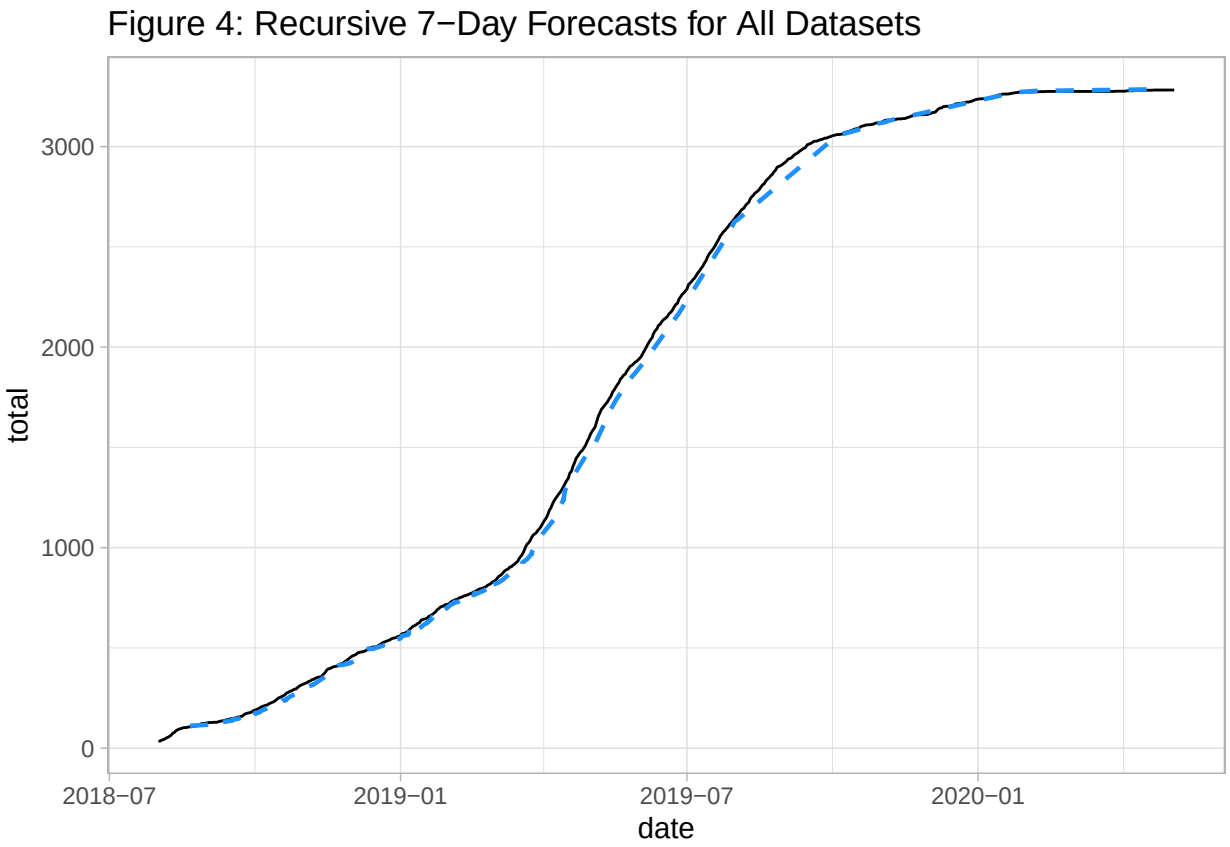


The 21-day Hawkes projections follow a similar pattern as 14-day Hawkes projections, with large under-predictions in the middle of the pandemic and more accurate projections in 2020 towards the end.

3 Recursive Complete Outbreak Analysis

3.1 7-Day Forecast Analysis

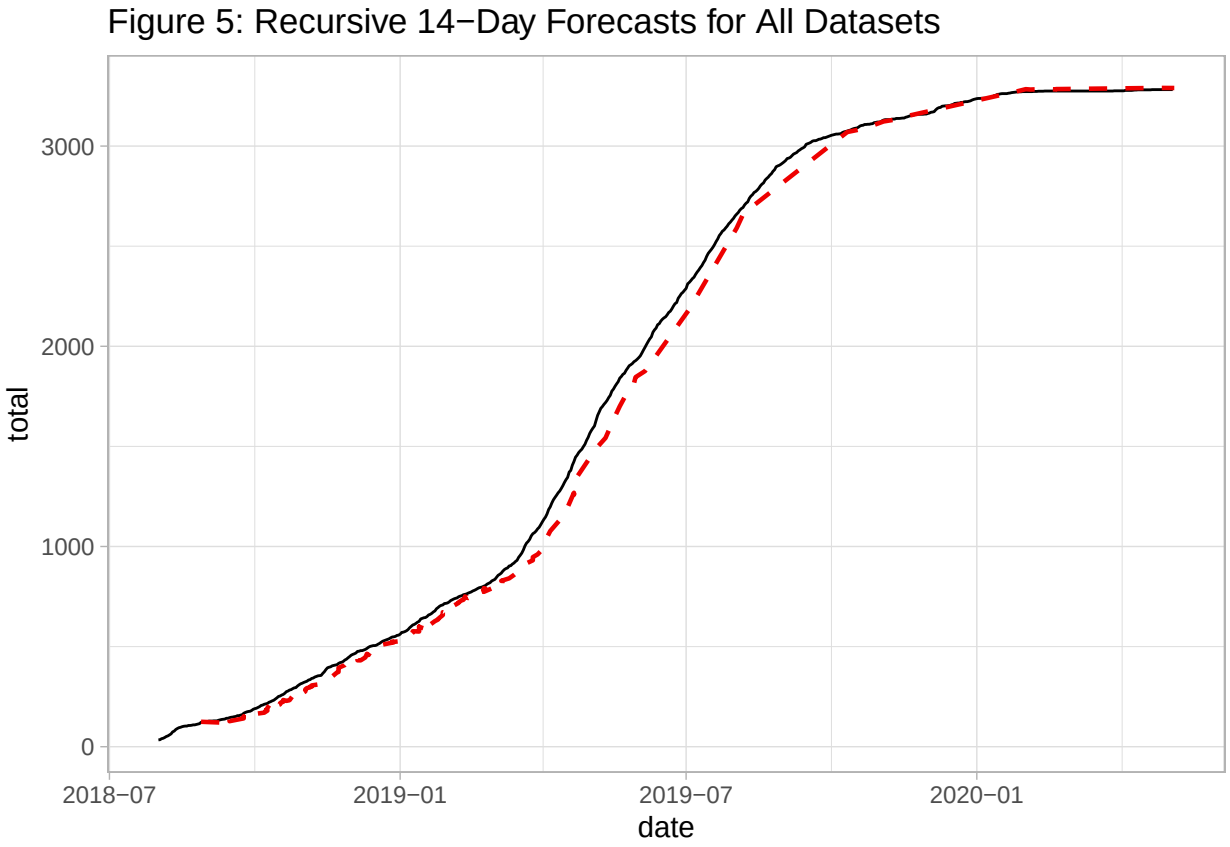
Figure 4 below shows the Recursive 7-Day Forecasts for all recorded simulations with respect to the recorded number of infections.



In general, the 7-day Recursive projections tend to under-predict the actual case counts. The largest errors generally occur during mid-2019, which is in the middle of the pandemic, whereas the model tends to have better prediction in the beginning and towards the end of the pandemic.

3.2 14-Day Forecast Analysis

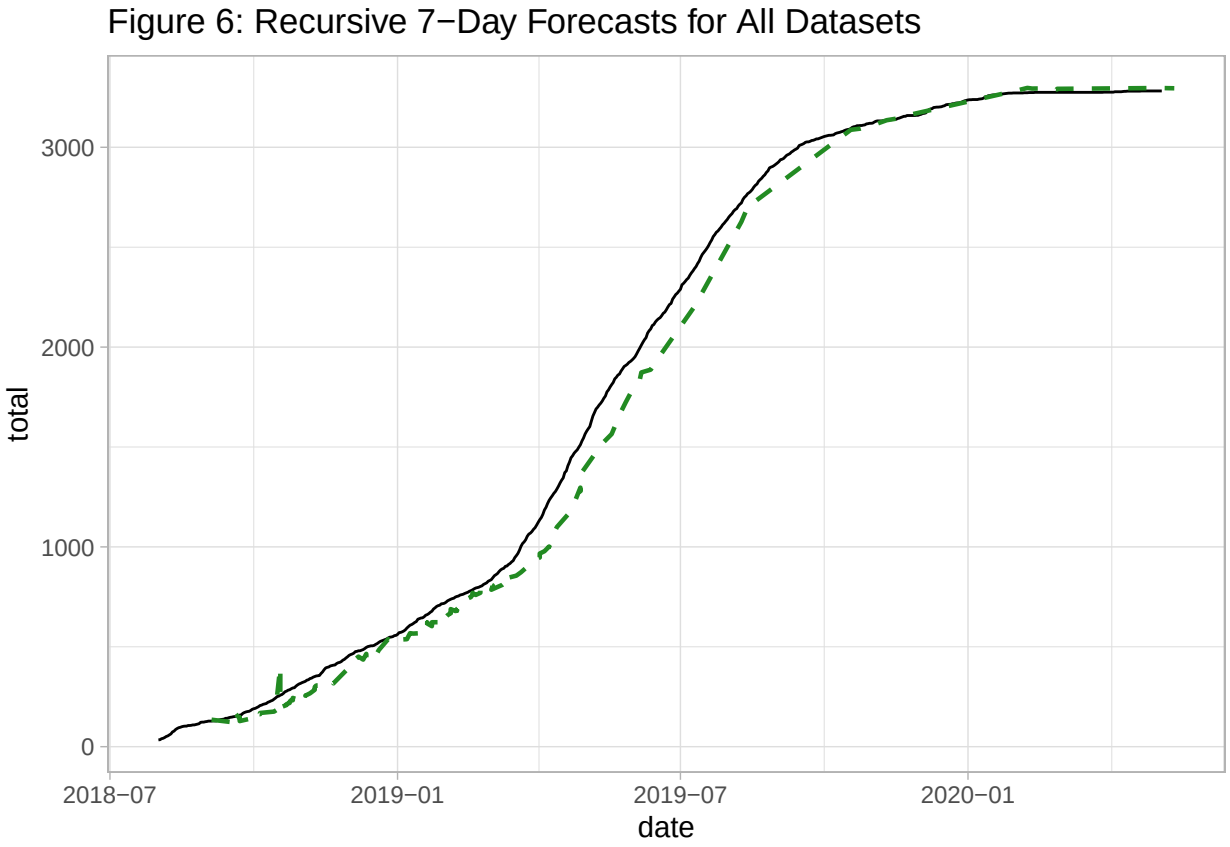
Figure 5 below shows the Recursive 14-Day Forecasts for all simulations with respect to the recorded number of infections.



The 14-day Recursive forecasts tend to under-predict the actual case counts during most of the pandemic, and has better prediction at the end of the pandemic.

3.3 21-Day Forecast Analysis

Figure 6 below shows the Recursive 21-Day Forecasts for all recorded simulations with respect to the recorded number of infections.



Similar to the 14-day forecasts, the 21-day Recursive forecasts tend to under-predict the true case counts in the beginning and middle of the pandemic, but is more accurate towards the end of the pandemic with slight over-prediction.

4 RMSE for Full Hawkes and Recursive Datasets

We compute the Root Mean Square Error (RMSE) of the 7, 14, and 21-day forecasts for both the Hawkes and Recursive models. The RMSE is computed as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}}$$

where N is the total number of observations.

The table below (Table 1) shows the RMSE values for the Hawkes and Recursive models, with respect to every simulated forecast during the outbreak.

Table 1: RMSE values for Hawkes and Recursive Models for all datasets.

	Hawkes	Recursive
7-day	28.75	29.67
14-day	60.43	60.70
21-day	90.87	92.22

We see that the Hawkes model forecasts have a consistently lower RMSE than those of the Recursive model for all three prediction days, when looking at all simulations.

5 Omission of Repeated Entries

Many of the forecasts were run with the same date. These extra runs are likely due to minor adjustments in the previously recorded data, so we refine our data to omit any repeated forecasts and only consider the most recent forecast with a repeated date. Therefore, for multiple forecasts that ended on the same date, we select the entry furthest down in the dataset, as it denotes the set with the most recent numbers.

The trend for this analysis very closely mirrors that of the previous analysis in terms of prediction, as evidenced in figures 7-12 below.

5.1 Hawkes Analysis

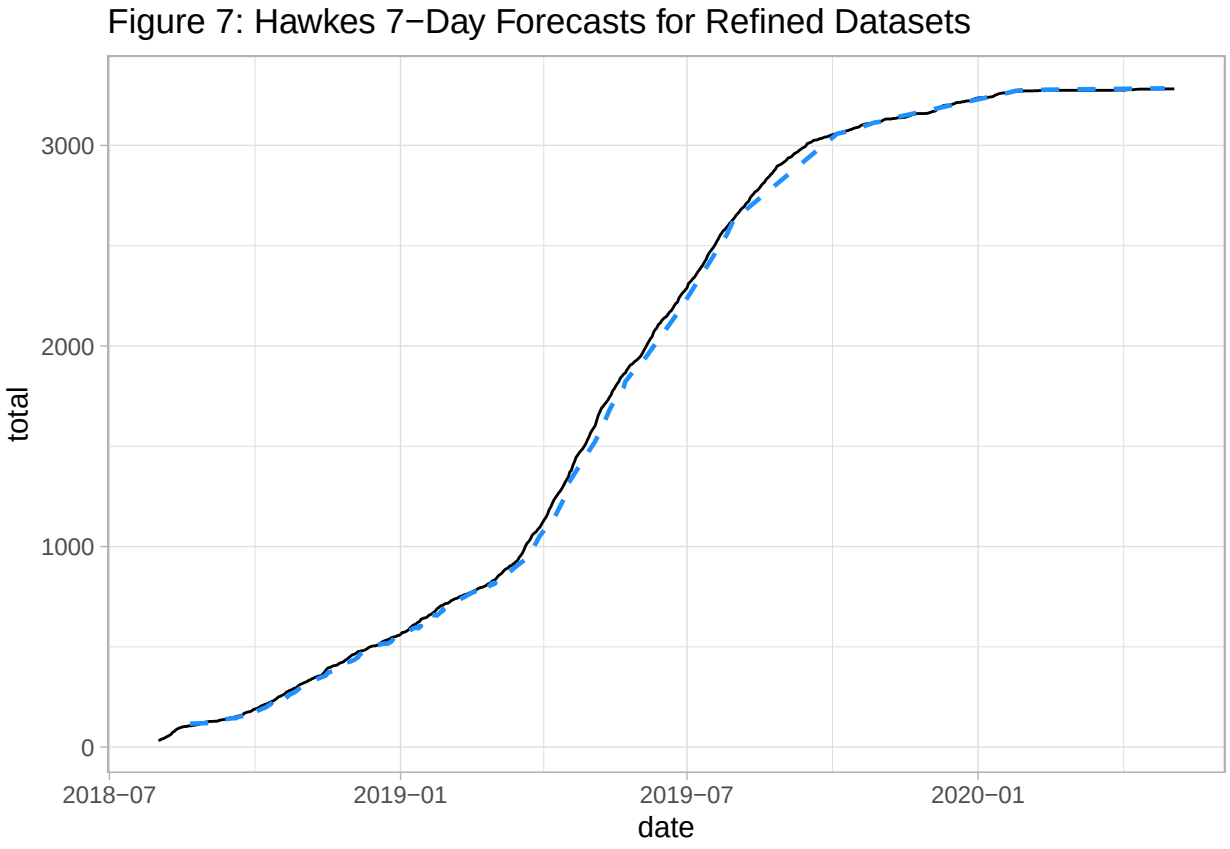


Figure 8: Hawkes 14−Day Forecasts for Refined Datasets

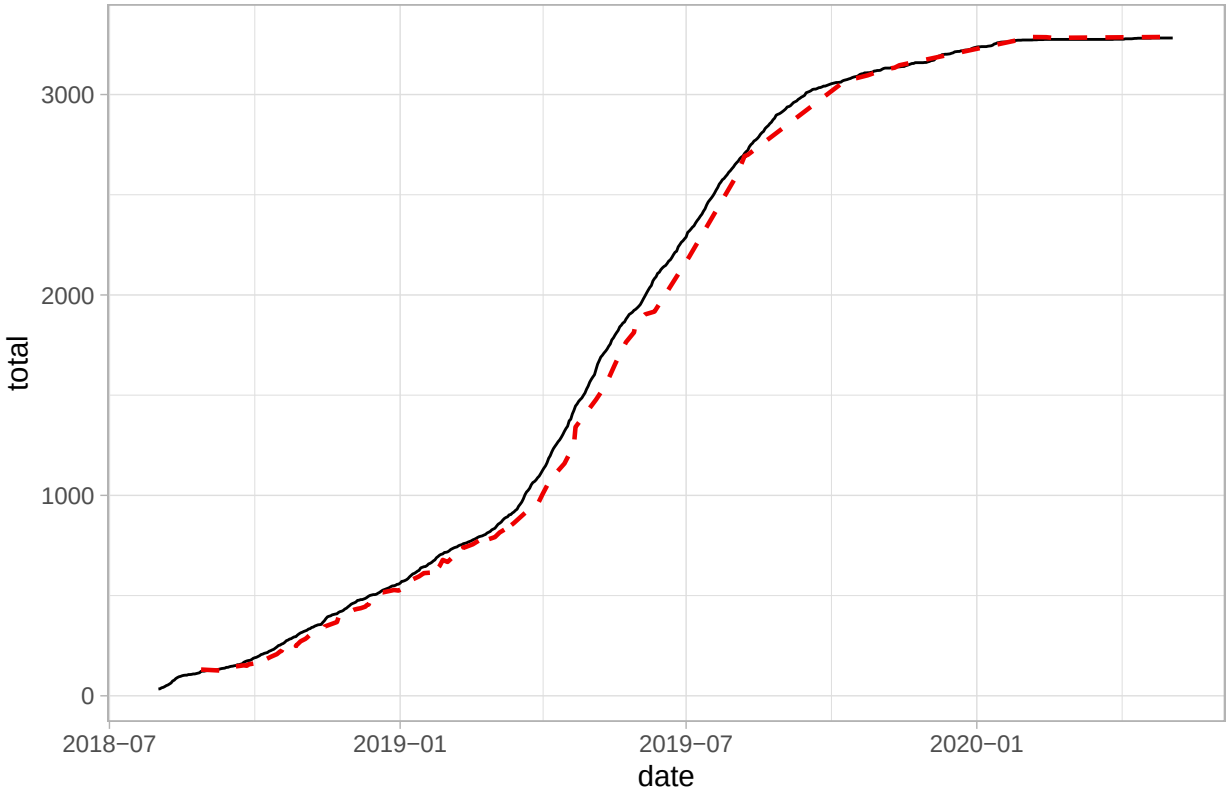
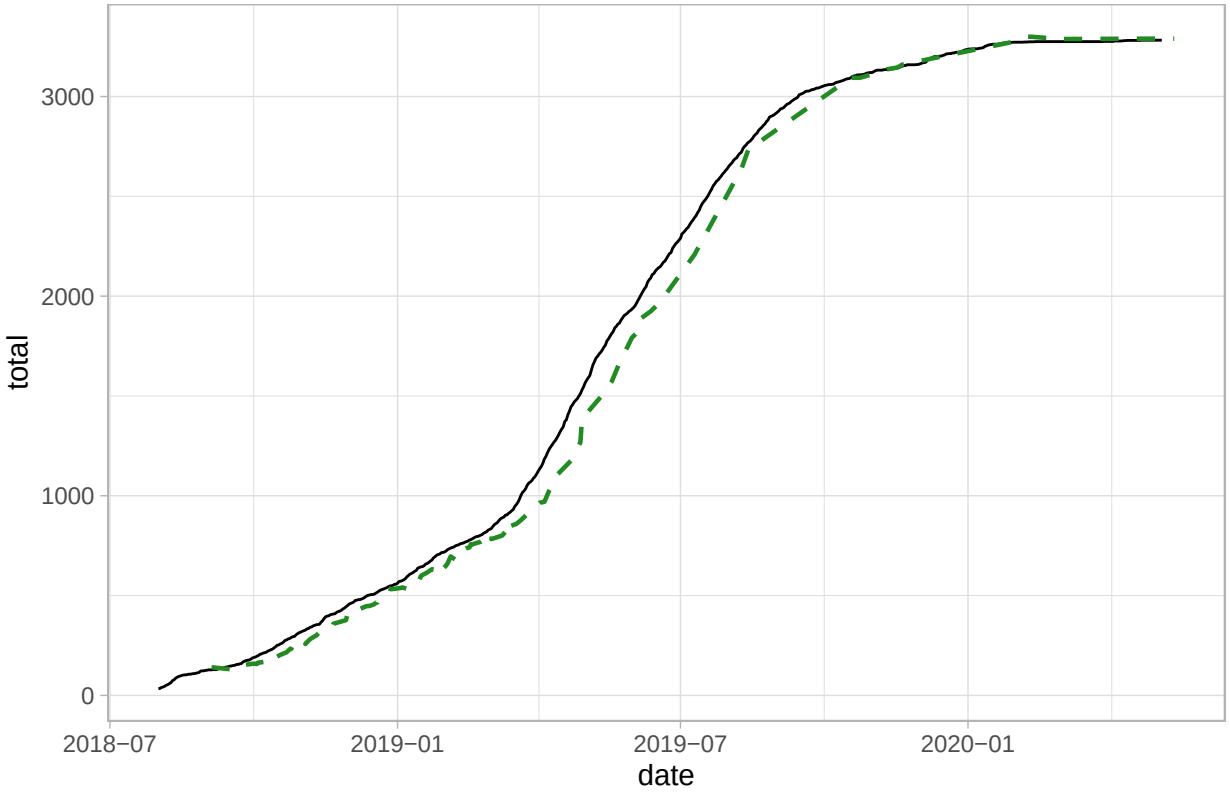


Figure 9: Hawkes 21−Day Forecasts for Refined Datasets



5.2 Recursive Analysis

Figure 10: Recursive 7−Day Forecasts for Refined Datasets

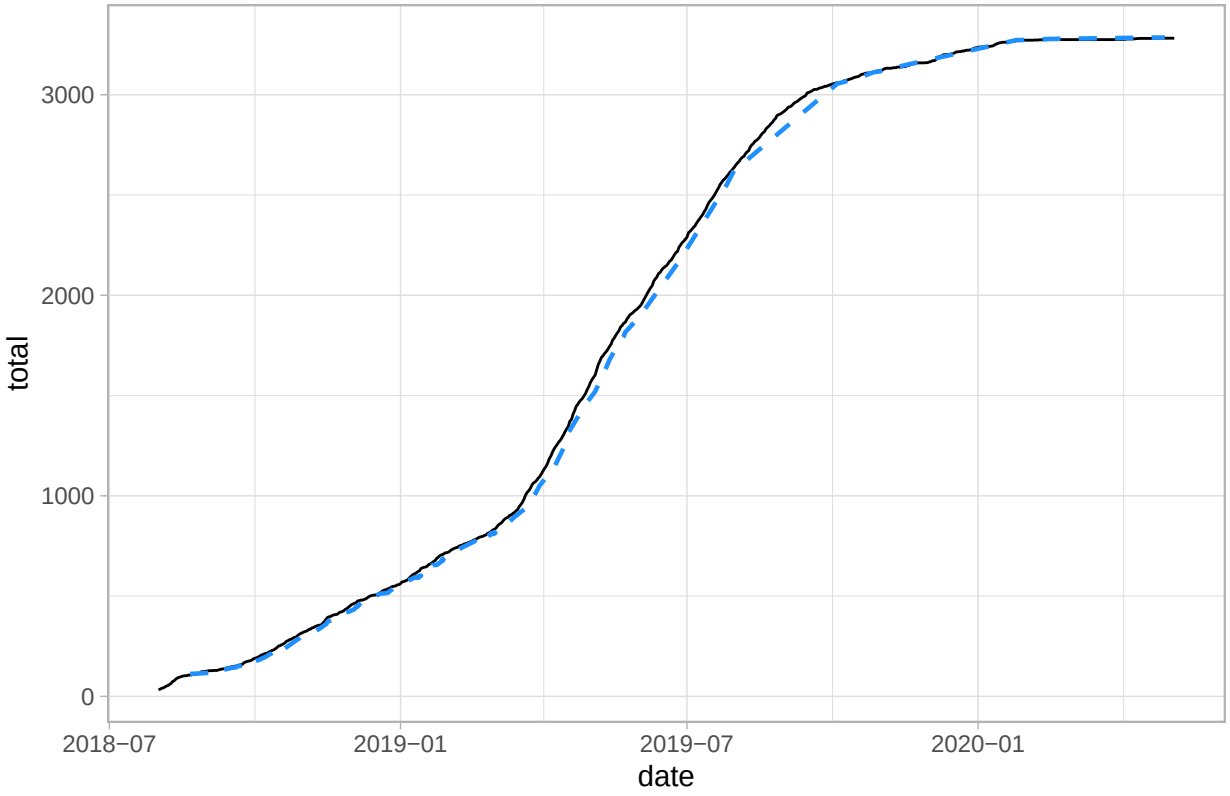


Figure 11: Recursive 14−Day Forecasts for Refined Datasets

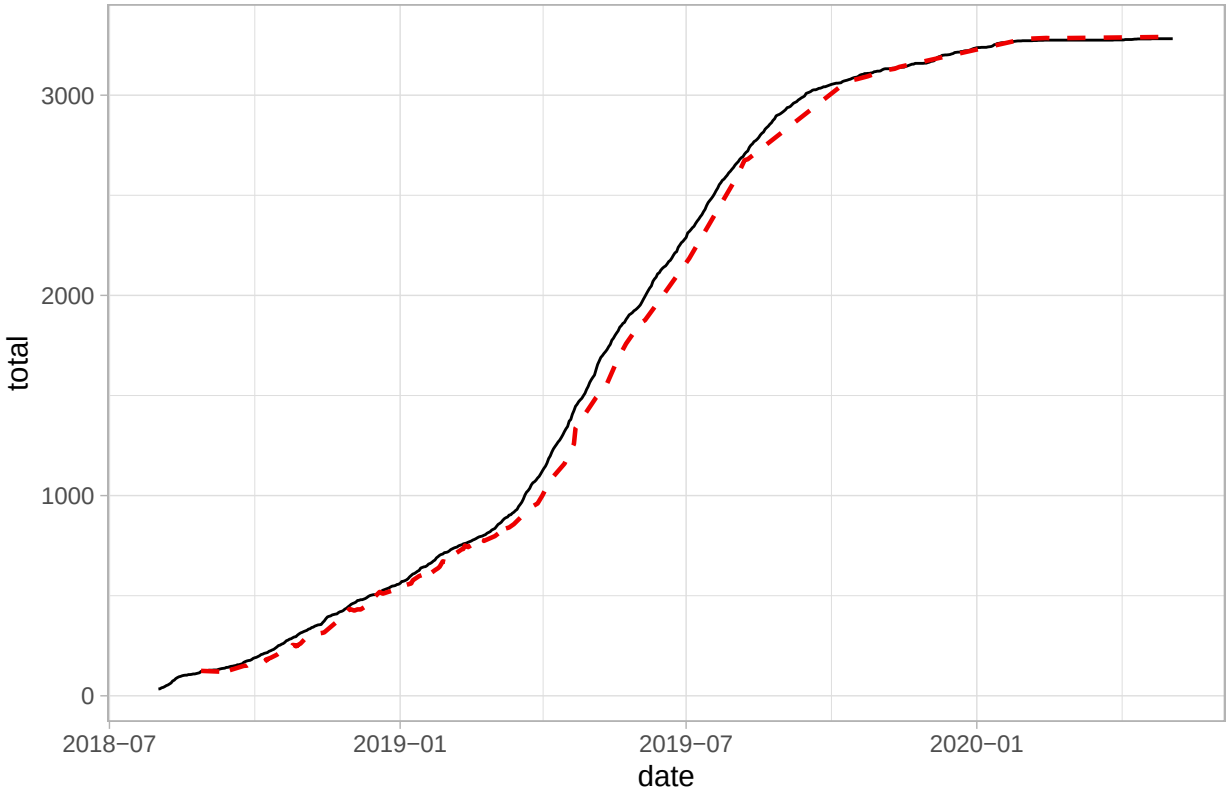


Figure 12: Recursive 21−Day Forecasts for Refined Datasets

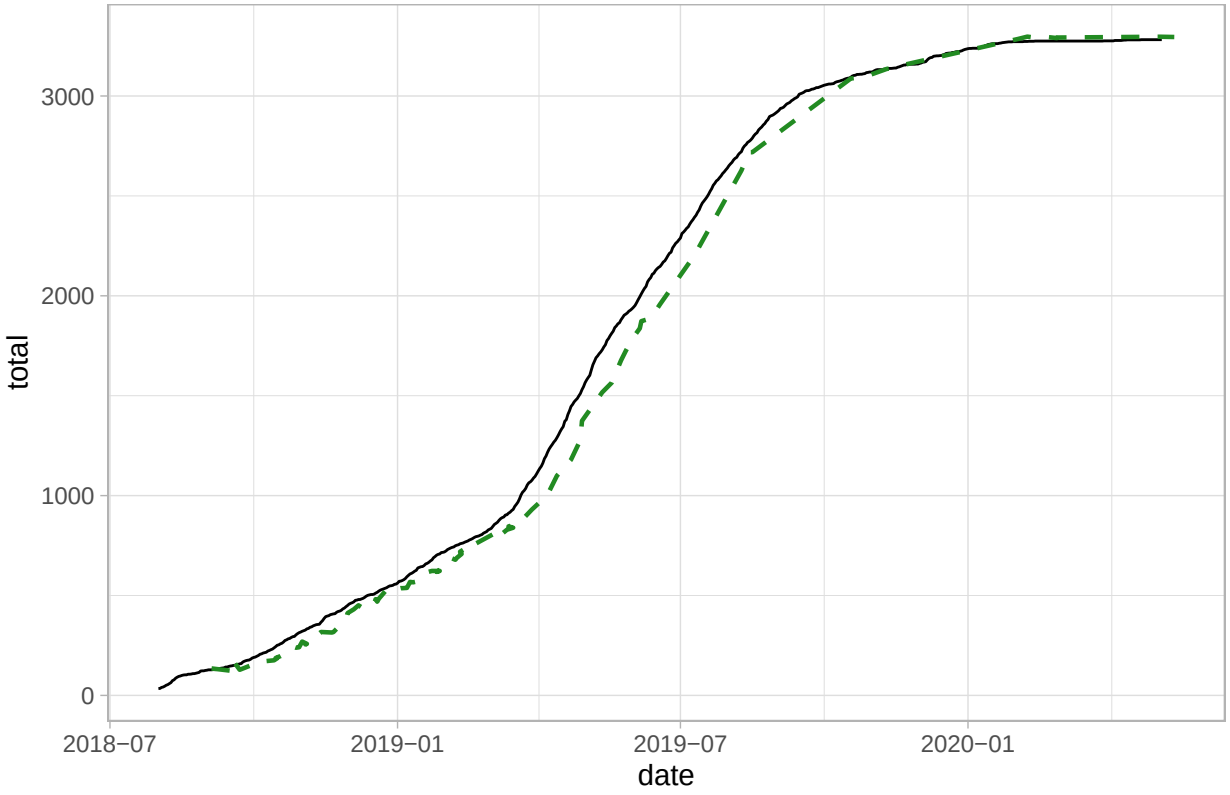


Table 2: RMSE values for Hawkes and Recursive Models for refined dataests.

	Hawkes	Recursive
7-day	29.47	30.71
14-day	61.12	62.09
21-day	90.86	92.99

There is not a large difference in RMSE of the full forecast analysis compared with that from the refined forecasts with the repeated dates removed.