Post-Processing COVID-19 Forecasts

- First Presentation -

Matthias Herp & Joel Beck

10.12.2021

Motivation:
UK Covid-19 Crowd Forecasting Challenge

- Predict Number of Covid-19 Cases and Deaths for the next 4 weeks in the United Kingdom

https://www.crowdforecastr.org/2021/05/11/uk-challenge/
https://epiforecasts.io/uk-challenge/

Motivation:
UK Covid-19 Crowd Forecasting Challenge

- Predict Number of Covid-19 Cases and Deaths for the next 4 weeks in the United Kingdom
- Submission of weekly predictions via an interactive web application

https://www.crowdforecastr.org/2021/05/11/uk-challenge/
https://epiforecasts.io/uk-challenge/

Motivation:
UK Covid-19 Crowd Forecasting Challenge

- Predict Number of Covid-19 Cases and Deaths for the next 4 weeks in the United Kingdom
- Submission of weekly predictions via an interactive web application
- Part of ongoing research project by the epiforecasts group at the London School of Hygiene & Tropical Medicine where our project supervisor Nikos Bosse is engaged as a doctoral candidate

https://www.crowdforecastr.org/2021/05/11/uk-challenge/
https://epiforecasts.io/uk-challenge/

Motivation:
UK Covid-19 Crowd Forecasting Challenge

- Idea: Compare forecasts from humans with model-based predictions

https://www.crowdforecastr.org/2021/05/11/uk-challenge/
https://epiforecasts.io/uk-challenge/

Motivation:
UK Covid-19 Crowd Forecasting Challenge

- Idea: Compare forecasts from humans with model-based predictions
- Empirically human forecasts are surprisingly competitive and in some cases even better than statistical models

https://www.crowdforecastr.org/2021/05/11/uk-challenge/
https://epiforecasts.io/uk-challenge/

Motivation:
UK Covid-19 Crowd Forecasting Challenge

- Idea: Compare forecasts from humans with model-based predictions
- Empirically human forecasts are surprisingly competitive and in some cases even better than statistical models
- This is mostly true for point forecasts, prediction intervals are often chosen too narrow, i.e. humans tend to be too confident in their own predictions

https://www.crowdforecastr.org/2021/05/11/uk-challenge/
https://epiforecasts.io/uk-challenge/

Motivation:
UK Covid-19 Crowd Forecasting Challenge

- Idea: Compare forecasts from humans with model-based predictions
- Empirically human forecasts are surprisingly competitive and in some cases even better than statistical models
- This is mostly true for point forecasts, prediction intervals are often chosen too narrow, i.e. humans tend to be too confident in their own predictions
- Goal: Use valuable information from point forecasts and adjust prediction intervals / quantile forecasts with an appropriate correction procedure

https://www.crowdforecastr.org/2021/05/11/uk-challenge/
https://epiforecasts.io/uk-challenge/

Motivation:
UK Covid-19 Crowd Forecasting Challenge

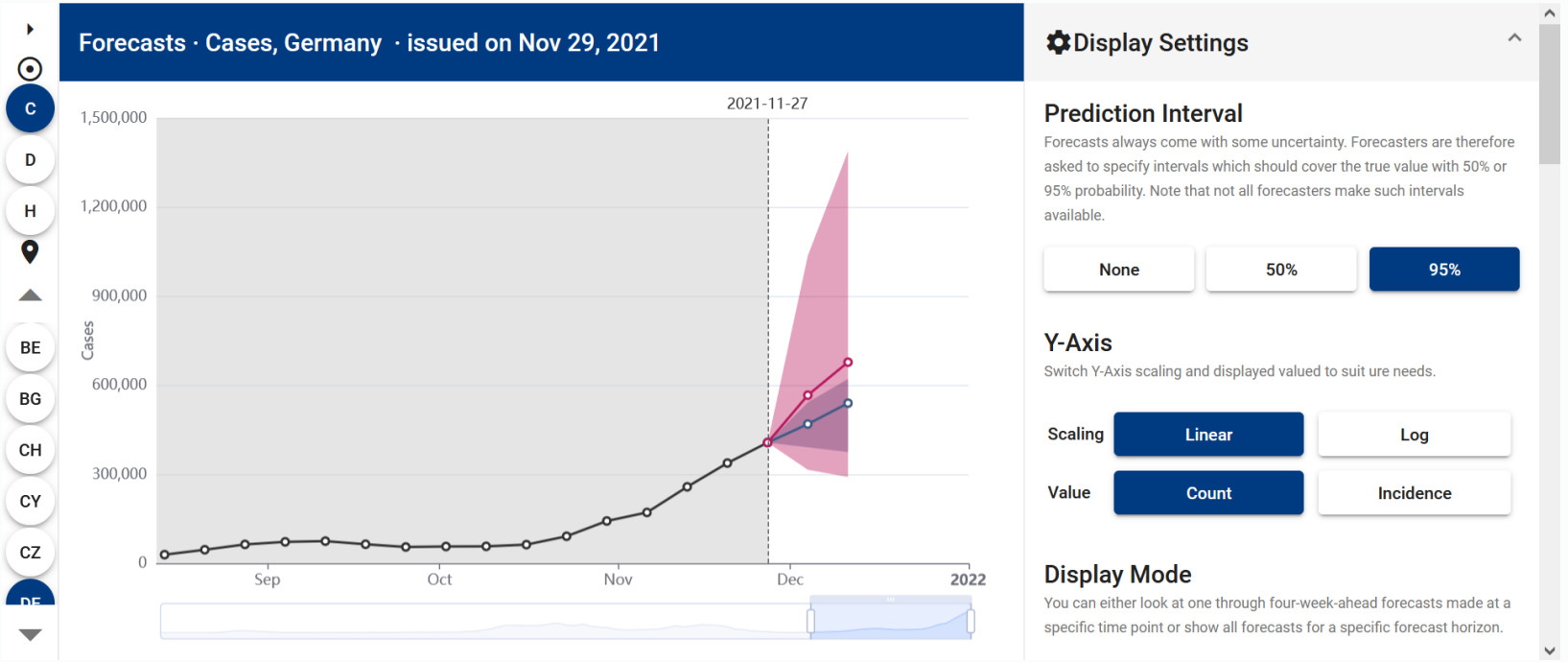
Show 4 entries Search:

location	location_name	target_end_date	target_type	true_value	population	forecast_date	quant
All	All	All	All	All	All	All	All
GB	United Kingdom	2021-06-19	Cases	62474	66022273	2021-06-14	
GB	United Kingdom	2021-06-12	Deaths	60	66022273	2021-06-07	
GB	United Kingdom	2021-06-19	Cases	62474	66022273	2021-05-24	
GB	United Kingdom	2021-08-14	Deaths	613	66022273	2021-07-19	

Showing 1 to 4 of 50 entries Previous 1 2 3 4 5 … 13 Next

European Forecast Hub

- UK Data only contains few observations over time span of 13 weeks
- European Forecast Hub provides data with forecasts from international research groups for many European Countries over a longer time horizon



https://covid19forecasthub.eu/index.html

Post Processing

Idea

- Adjust Forecasts based on performance metrics for out-of-sample data

Idea

- Adjust Forecasts based on performance metrics for out-of-sample data
- Split in 3 separate data sets:
  - **Training:** Build quantile predictions model
  - **Validation:** Determine hyperparameters of post-processing method
  - **Test:** Evaluate adjusted predictions

Idea

- Adjust Forecasts based on performance metrics for out-of-sample data
- Split in 3 separate data sets:
  - **Training:** Build quantile predictions model
  - **Validation:** Determine hyperparameters of post-processing method
  - **Test:** Evaluate adjusted predictions
- **Important:** In our project we do **not** build a prediction model, we merely adjust forecasts of existing ones.

Evaluation

- Based on Weighted Interval Score[1]

$$\text{WIS} = \text{Sharpness} + \text{Overprediction} + \text{Underprediction}$$

[1] https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008618

Evaluation

- Based on Weighted Interval Score[1]

$$\text{WIS} = \text{Sharpness} + \text{Overprediction} + \text{Underprediction}$$

- For a given quantile level α, true observed value y as well as lower bound l and upper bound u of the corresponding $(1-\alpha)\cdot 100\%$ prediction interval, the score is computed as

$$Score_\alpha(y) = (u-l) + \frac{2}{\alpha}\cdot(l-y)\cdot\mathbf{1}(y\leq l) + \frac{2}{\alpha}\cdot(y-u)\cdot\mathbf{1}(y\geq u)$$

[1] https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008618

Evaluation

- Based on Weighted Interval Score[1]

$$\text{WIS} = \text{Sharpness} + \text{Overprediction} + \text{Underprediction}$$

- For a given quantile level α, true observed value y as well as lower bound l and upper bound u of the corresponding $(1-\alpha)\cdot 100\%$ prediction interval, the score is computed as

$$Score_\alpha(y) = (u-l) + \frac{2}{\alpha}\cdot(l-y)\cdot\mathbf{1}(y\leq l) + \frac{2}{\alpha}\cdot(y-u)\cdot\mathbf{1}(y\geq u)$$

- The Score of the entire model can be obtained from a weighted sum over all (included) quantile levels α

[1] https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008618

Evaluation

- Based on Weighted Interval Score[1]

$$\text{WIS} = \text{Sharpness} + \text{Overprediction} + \text{Underprediction}$$

- For a given quantile level α, true observed value y as well as lower bound l and upper bound u of the corresponding $(1-\alpha)\cdot 100\%$ prediction interval, the score is computed as

$$Score_\alpha(y) = (u-l) + \frac{2}{\alpha}\cdot(l-y)\cdot\mathbf{1}(y\leq l) + \frac{2}{\alpha}\cdot(y-u)\cdot\mathbf{1}(y\geq u)$$

- The Score of the entire model can be obtained from a weighted sum over all (included) quantile levels α

- Implemented in the **`scoringutils`** R package written by Nikos

[1] https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008618

Conformalized Quantile Regression

Conformalized Quantile Regression

Theory based on Paper Romano Y., Patterson E., and Candès E. (2019): *Conformalized Quantile Regression*

Central Theorem:

If $(X_i, Y_i), i = 1, \ldots, n+1$ are exchangeable, then the $(1-\alpha) \cdot 100\%$ prediction interval $C(X_{n+1})$ constructed by the CQR algorithm satisfies

$$P\left(Y_{n+1} \in C(X_{n+1})\right) \geq 1 - \alpha \quad (\text{coverage}).$$

Moreover, if the conformity scores E_i are almost surely distinct, then the prediction interval is nearly perfectly calibrated:

$$P\left(Y_{n+1} \in C(X_{n+1})\right) \leq 1 - \alpha + \frac{1}{|I_2| + 1} \quad (\text{precision}).$$

CQR Algorithm

Step 1:
Split the data into a training and validation (here called *calibration*) set, indexed by I_1 and I_2, respectively

CQR Algorithm

Step 1:
Split the data into a training and validation (here called *calibration*) set, indexed by I_1 and I_2, respectively

Step 2:
For a given quantile α and a given quantile regression algorithm $\mathcal{A}$, calculate lower and upper interval bounds on the training set:

$$\left\{\hat{q}_{\alpha,low},\ \hat{q}_{\alpha,high}\right\} \leftarrow \mathcal{A}\left(\{(X_i, Y_i) : i \in I_1\}\right)$$

CQR Algorithm

Step 1:
Split the data into a training and validation (here called *calibration*) set, indexed by I_1 and I_2, respectively

Step 2:
For a given quantile α and a given quantile regression algorithm $\mathcal{A}$, calculate lower and upper interval bounds on the training set:

$$\left\{\hat{q}_{\alpha,low},\ \hat{q}_{\alpha,high}\right\} \leftarrow \mathcal{A}\left(\{(X_i, Y_i) : i \in I_1\}\right)$$

Step 3:
Compute **conformity scores** on the calibration set:

$$E_i := \max\left\{\hat{q}_{\alpha,low}(X_i) - Y_i,\ Y_i - \hat{q}_{\alpha,high}(X_i)\right\} \quad \forall\, i \in I_2$$

For each i, the corresponding score E_i is **positive** if Y_i is **outside** the interval $\left[\hat{q}_{\alpha,low}(X_i),\ \hat{q}_{\alpha,high}(X_i)\right]$ and **negative** if Y_i is **inside** the interval.

CQR Algorithm

Step 4:
Compute the **margin** $Q_{1-\alpha}(E, I_2)$ given by the $(1-\alpha)(1+\frac{1}{1+|I_2|})$-th empirical quantile of the scores E_i in the calibration set.

CQR Algorithm

Step 4:
Compute the **margin** $Q_{1-\alpha}(E, I_2)$ given by the $(1-\alpha)(1+\frac{1}{1+|I_2|})$-th empirical quantile of the scores E_i in the calibration set.

Step 5:
On the basis of the original prediction interval bounds $\hat{q}_{\alpha,low}(X_i)$ and $\hat{q}_{\alpha,high}(X_i)$, the new *post-processed* prediction interval for Y_i is given by

$$C(X_{n+1}) = \left[\hat{q}_{\alpha,low}(X_i) - Q_{1-\alpha}(E, I_2),\ \hat{q}_{\alpha,high}(X_i) + Q_{1-\alpha}(E, I_2)\right].$$

CQR Algorithm

Step 4:
Compute the **margin** $Q_{1-\alpha}(E, I_2)$ given by the $(1-\alpha)(1+\frac{1}{1+|I_2|})$-th empirical quantile of the scores E_i in the calibration set.

Step 5:
On the basis of the original prediction interval bounds $\hat{q}_{\alpha,low}(X_i)$ and $\hat{q}_{\alpha,high}(X_i)$, the new *post-processed* prediction interval for Y_i is given by

$$C(X_{n+1}) = \left[\hat{q}_{\alpha,low}(X_i) - Q_{1-\alpha}(E, I_2),\ \hat{q}_{\alpha,high}(X_i) + Q_{1-\alpha}(E, I_2)\right].$$

- Note that the **same** margin is subtracted/added for the lower and upper bound, which limits the flexibility

CQR Algorithm

Step 4:
Compute the **margin** $Q_{1-\alpha}(E, I_2)$ given by the $(1-\alpha)(1+\frac{1}{1+|I_2|})$-th empirical quantile of the scores E_i in the calibration set.

Step 5:
On the basis of the original prediction interval bounds $\hat{q}_{\alpha,low}(X_i)$ and $\hat{q}_{\alpha,high}(X_i)$, the new *post-processed* prediction interval for Y_i is given by

$$C(X_{n+1}) = \left[\hat{q}_{\alpha,low}(X_i) - Q_{1-\alpha}(E, I_2),\ \hat{q}_{\alpha,high}(X_i) + Q_{1-\alpha}(E, I_2)\right].$$

- Note that the **same** margin is subtracted/added for the lower and upper bound, which limits the flexibility
- Possible extensions could use different margins and/or **multiplicative** correction terms

The `postforecasts` package 📦

Core Idea

- Structured and unifying framework for implementing various post-processing techniques

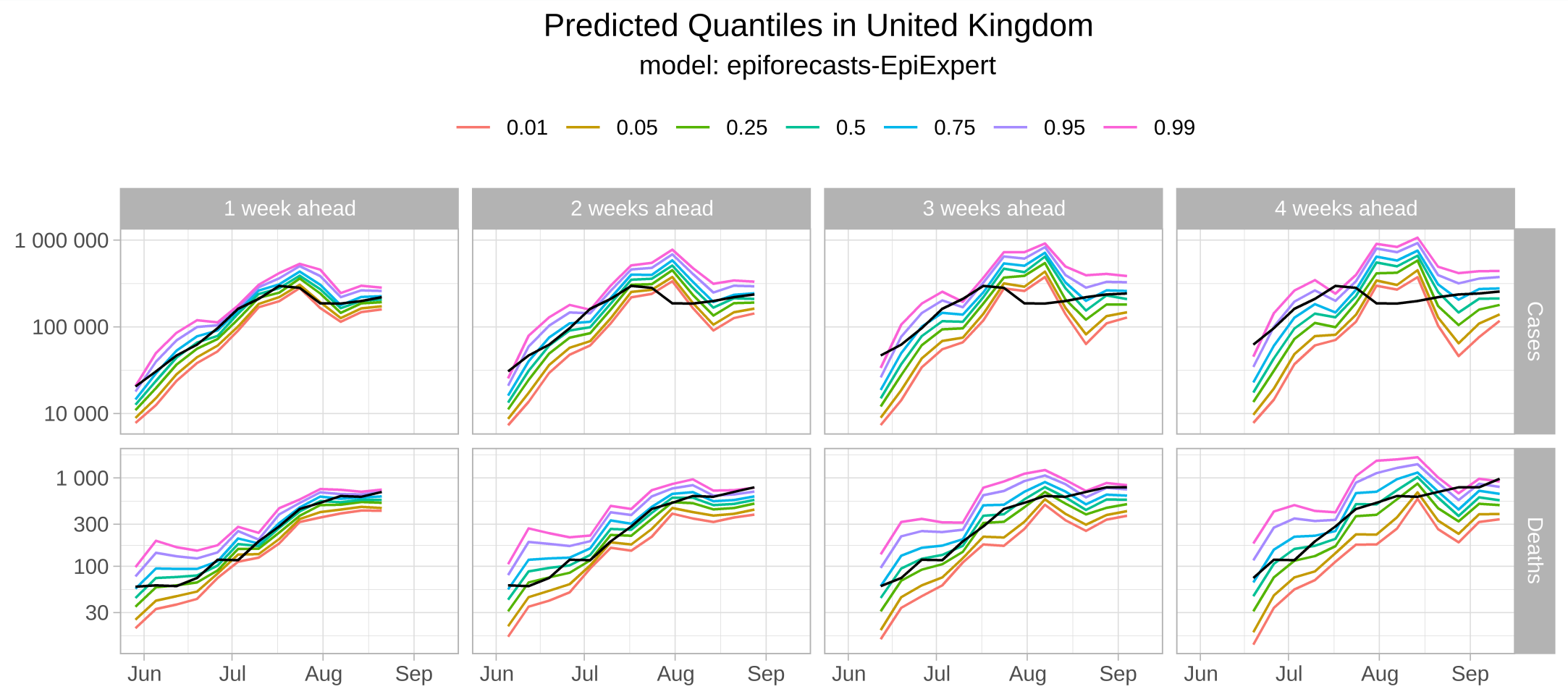
Core Idea

- Structured and unifying framework for implementing various post-processing techniques
- Aims to establish a consistent workflow for a collection of post-processing methods

Core Idea

- Structured and unifying framework for implementing various post-processing techniques
- Aims to establish a consistent workflow for a collection of post-processing methods
- Allows for convenient comparisons between methods for the data of interest

Overview of original Data



Update Predictions with CQR

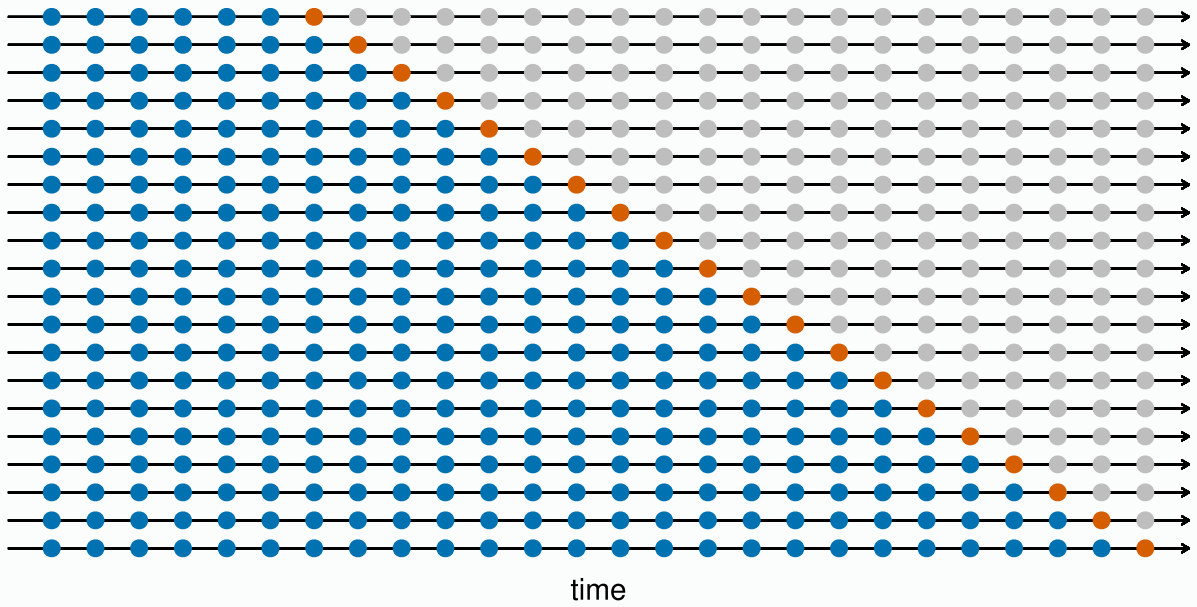
- The interval score for the entire data set improves after applying the CQR method
- This result holds in the aggregate over all horizons and quantiles
- They also hold in both target types

```
df_updated <- update_predictions(
  df,
  method = "cqr",
  models = "epiforecasts-EpiExpert",
  locations = "GB"
)
```

method	target_type	interval_score	sharpness	underprediction	overprediction
original	Cases	55982.368	11210.563	10649.916	34121.890
cqr	Cases	50795.905	22053.565	4252.159	24490.181
original	Deaths	55.792	19.285	28.865	7.642
cqr	Deaths	54.414	24.199	23.859	6.356

Time Series Cross-Validation

- Time series cross-validation iterates through the data along the time dimension
- At each time point the test set is composed of the one step ahead prediction
- The algorithm typically starts with a minimum number of observation as the initial training set



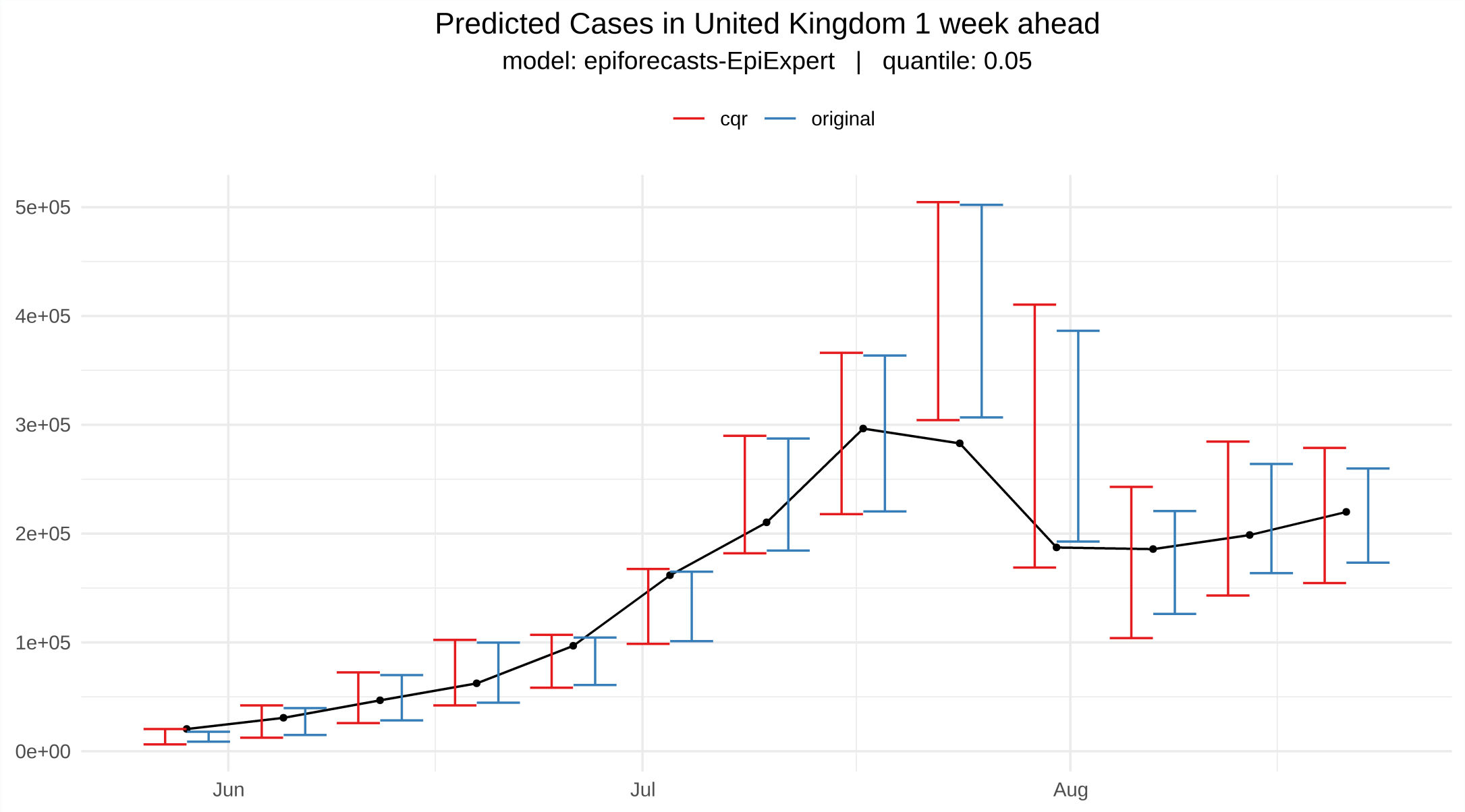
https://otexts.com/fpp3/tscv.html

Update and Evaluate Forecasts for a fixed time horizon

- For the validation set the interval score improvement is much lower
- Under- and overprediction decrease while sharpness increases
- This indicates an overall increase in prediction intervals

```
df_updated <- update_predictions(
  df,
  method = "cqr",
  models = "epiforecasts-EpiExpert",
  locations = "GB",
  cv_init_training = 5
)
```

method	target_type	interval_score	sharpness	underprediction	overprediction
original	Cases	72069.848	14211.324	11169.036	46689.488
cqr	Cases	68141.683	26533.751	5044.787	36563.145
original	Deaths	72.533	23.779	38.785	9.969
cqr	Deaths	72.080	24.225	37.367	10.488

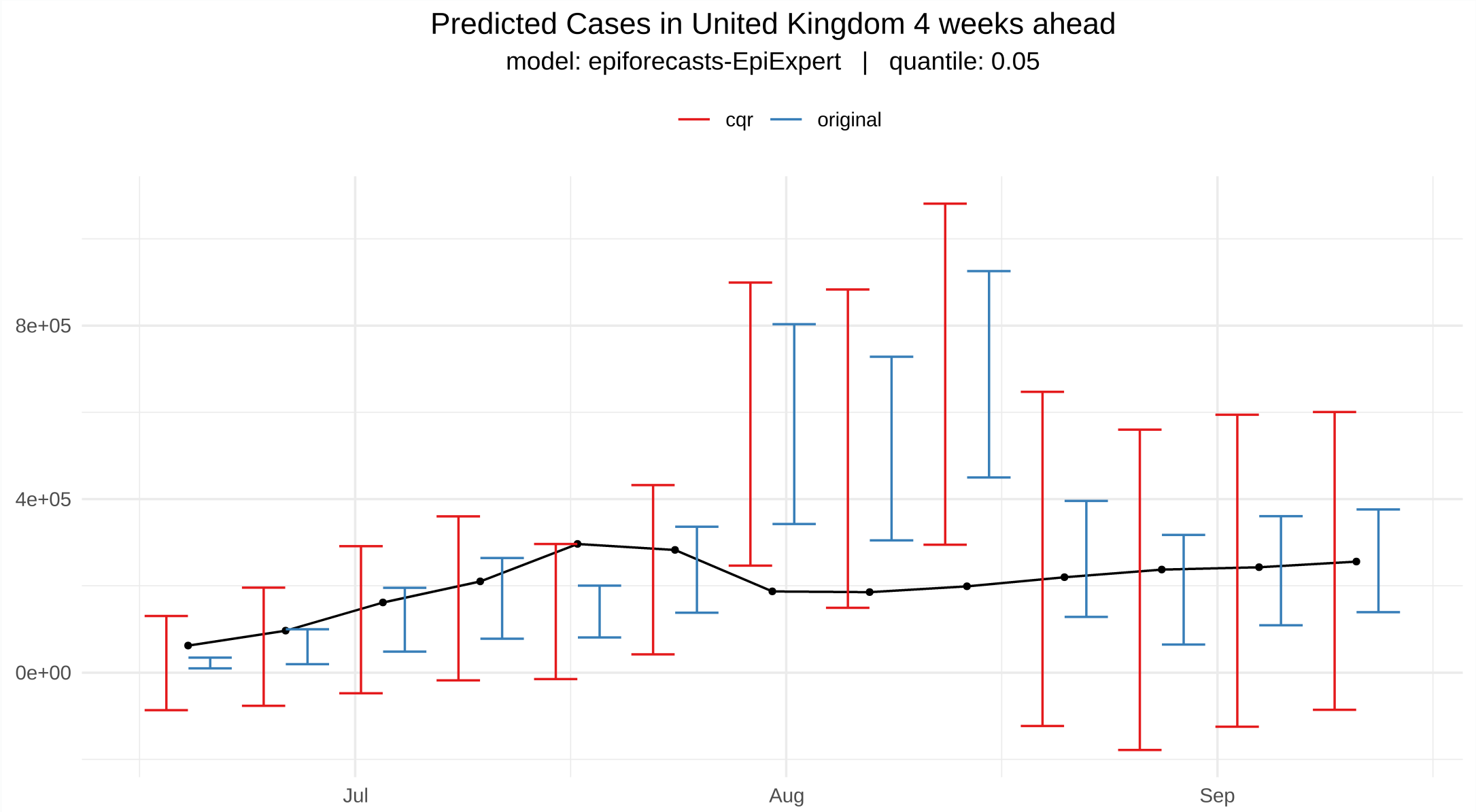

Predicted Cases in United Kingdom 1 week ahead
model: epiforecasts-EpiExpert | quantile: 0.05
cqr
original
5e+05
4e+05
3e+05
2e+05
1e+05
0e+00
Jun
Jul
Aug

Evaluate Predicted Cases stratified by time horizon

- CQR provides a benefit in forecasting at larger horizons
- In the aggregate results for larger horizon tend to have a stronger weight on the scores
- Similar results are found for the prediction interval size, larger quantiles tend to benefit more

```
df_updated <- update_predictions(
  df,
  method = "cqr",
  models = "epiforecasts-EpiExpert",
  locations = "GB",
  target_types = "Cases", # if not specified, all time horizons are included
  cv_init_training = 5
)
```

method	horizon	interval_score	sharpness	underprediction	overprediction
original	1	23469.36	7907.668	1985.659	13576.03
cqr	1	23599.42	9410.678	1489.086	12699.66
original	2	53316.51	11982.427	5988.121	35345.96
cqr	2	53454.37	16737.482	4172.729	32544.16
original	3	88343.02	15967.668	13155.618	59219.74
cqr	3	83473.54	30550.912	6320.756	46601.88
original	4	107965.50	19022.767	20280.802	68661.93
cqr	4	98614.89	43350.061	7184.281	48080.55


Predicted Cases in United Kingdom 4 weeks ahead
model: epiforecasts-EpiExpert | quantile: 0.05
cqr
original
8e+05
4e+05
0e+00
Jul
Aug
Sep

What's next?

Many possible directions

- Extension / Refinement of the CQR Method as well as the implementation of further Post-Processing methods

Many possible directions

- Extension / Refinement of the CQR Method as well as the implementation of further Post-Processing methods
- Analysis of method performances in relationship characteristics such as the sample size, forecast horizon, interval width and prediction model

Many possible directions

- Extension / Refinement of the CQR Method as well as the implementation of further Post-Processing methods
- Analysis of method performances in relationship characteristics such as the sample size, forecast horizon, interval width and prediction model
- Construct new Post-Processing method as Ensemble Model of individual processing techniques

References

Traditional CQR Method

Romano Y., Patterson E., and Candès E. (2019). Conformalized Quantile Regression. *NeurIPS Annual Conference on Neural Information Processing Systems.*

- Paper: https://proceedings.neurips.cc/paper/2019/file/5103c3584b063c431bd1268e9b5e76fb-Paper.pdf
- Poster: https://github.com/yromano/cqr/blob/master/poster/CQR_Poster.pdf

Variations and Extensions of CQR

Tibshirani R. (2019). Advances and Challenges in Conformal Inference. *Carnegie Mellon University.*

- Slides: www.stat.cmu.edu/~ryantibs/talks/conformal-2019.pdf

References

Theoretical Foundation of Weighted Interval Score

Bracher et al. (2021). Evaluating epidemic forecasts in an interval format.

- Paper: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008618

Time Series Cross Validation

Hyndman R., Athanasopoulos G. (2021). Forecasting: principles and practice, 3rd edition. *OTexts: Melbourne, Australia*.

- Online Version: https://otexts.com/fpp3/

References

More Information about the UK Covid-19 Forecasting Challenge

- Website: https://www.crowdforecastr.org/2021/05/11/uk-challenge/
- Evaluation & Ranking: https://epiforecasts.io/uk-challenge/

More Information about the European Forecasting Hub

- Website: https://covid19forecasthub.eu/index.html
- GitHub: https://github.com/epiforecasts/covid19-forecast-hub-europe