Title: Empirical prediction intervals applied to short term mortality forecasts and excess deaths

Authors: Jonas Schöley[1], Ricarda Duerst[1]

Affiliations:

[1] Max Planck Institute for Demographic Research; Rostock, Germany

Corresponding authors:

Jonas Schöley `schoeley@demogr.mpg.de`

Abstract: We propose empirical prediction intervals for the study of weekly expected and excess deaths and demonstrate the superior coverage and generality of these intervals compared with conventional parametric intervals. Instead on relying on the suitability of parametric assumptions or the magnitude of errors over the fitting period, empirical prediction intervals are estimated from cross-validated time series of past prediction errors, reflecting the intuitive notion that a forecast is only as precise as similar forecasts in the past turned out to be. We further employ empirical prediction intervals to assess the level of excess mortality which can reliably be detected two years into the pandemic, given the error of our methods.

Background

We propose empirical prediction intervals for the study of weekly expected and excess deaths and demonstrate the superior coverage and generality of these intervals compared with conventional parametric intervals.

Mortality forecasts on a sub-annual timescales have gained relevance as the basis for excess deaths calculations during the COVID-19 pandemic. Framed as a forecasting problem, one aims to predict the weekly deaths which would have happened without COVID-19 by forecasting deaths over the pandemic period based on pre-pandemic trends. Those forecasted "expected deaths" are associated with an error which can probabilistically be expressed as a "prediction interval" within which the true expected deaths are to be found with a given probability.It is well known that prediction intervals around forecasted values tend to be too narrow. In the context of COVID-19 excess death modeling this phenomenon may lead to wrong conclusions regarding the impact of the pandemic on population mortality, by giving an overly optimistic picture on how precise one can actually measure excess deaths more than two years into the pandemic.

Empirical prediction intervals, instead on relying on correct parametric specifications, estimate the distribution of error around a forecasted value from actual past, out-of-sample forecasting errors. The idea is simple: One quantifies the distribution of error around a forecast by applying the forecasting method to an equivalent forecasting problem where the true value is known. In contrast to the cross-validation method of forecast assessment one seeks to estimate the full distribution of forecast error indexed over all desired forecasting strata and time points. In order to do so, a statistical model is fitted to known out-of-sample forecasting errors.

Research on empirical uncertainty intervals has been ongoing for at least half a century [1]. The 1980s saw active demographic reasearch on empirical intervals and the authors generally found that the intervals based on past error have better coverage than model based intervals, see e.g. [2], [3], [4].

In the machine learning literature there is growing interest in "distribution free uncertainty quantification", given that popular techniques do not provide an intrinsic estimate of uncertainty. The community has build a theory on empirical uncertainty intervals under the term "Conformal prediction". Here too, the idea is to look at the distribution of historical error measures to assess how likely any given future deviation from the prediction would be.

Interest in empirical prediction intervals saw a recent resurgence in the context of algorithmic prediction methods, for which probabilistic errors are hard or impossible to derive. The community has build a theory on empirical uncertainty intervals under the term "Conformal prediction". The idea is to look at the distribution of historical error measures to assess how likely any given future deviation from the prediction would be [5].

We demonstrate the construction and coverage of empirical prediction intervals via the example of weekly death counts in Denmark. The task is to estimate expected weekly death counts and prediction intervals over a period of 2 years starting 2018. We compare the coverage of the empirical prediction intervals with conventional parametric intervals.

Data and Methods

Data on weekly deaths was sourced from the Short Term Mortality Fluctuations Database (STMF)[6]. The data was split into four partially overlapping data series with 7 years of training data followed by a 2 year forecasting horizon each. While the first three data series (Figure 1A) are used for the estimation of the empirical forecasting error, the fourth series (Figure 1C) is reserved for the application and validation of the empirical uncertainty interval method – in an applied scenario, this would be the forecast of interest.

A First, forecast weekly deaths over past data

7 year training period   2 year forecasting period

B Second, estimate the time-indexed distribution of the observed forecasting error

C Third, derive quantiles of the estimated error distribution and apply to novel forecasts to get empirical prediction intervals

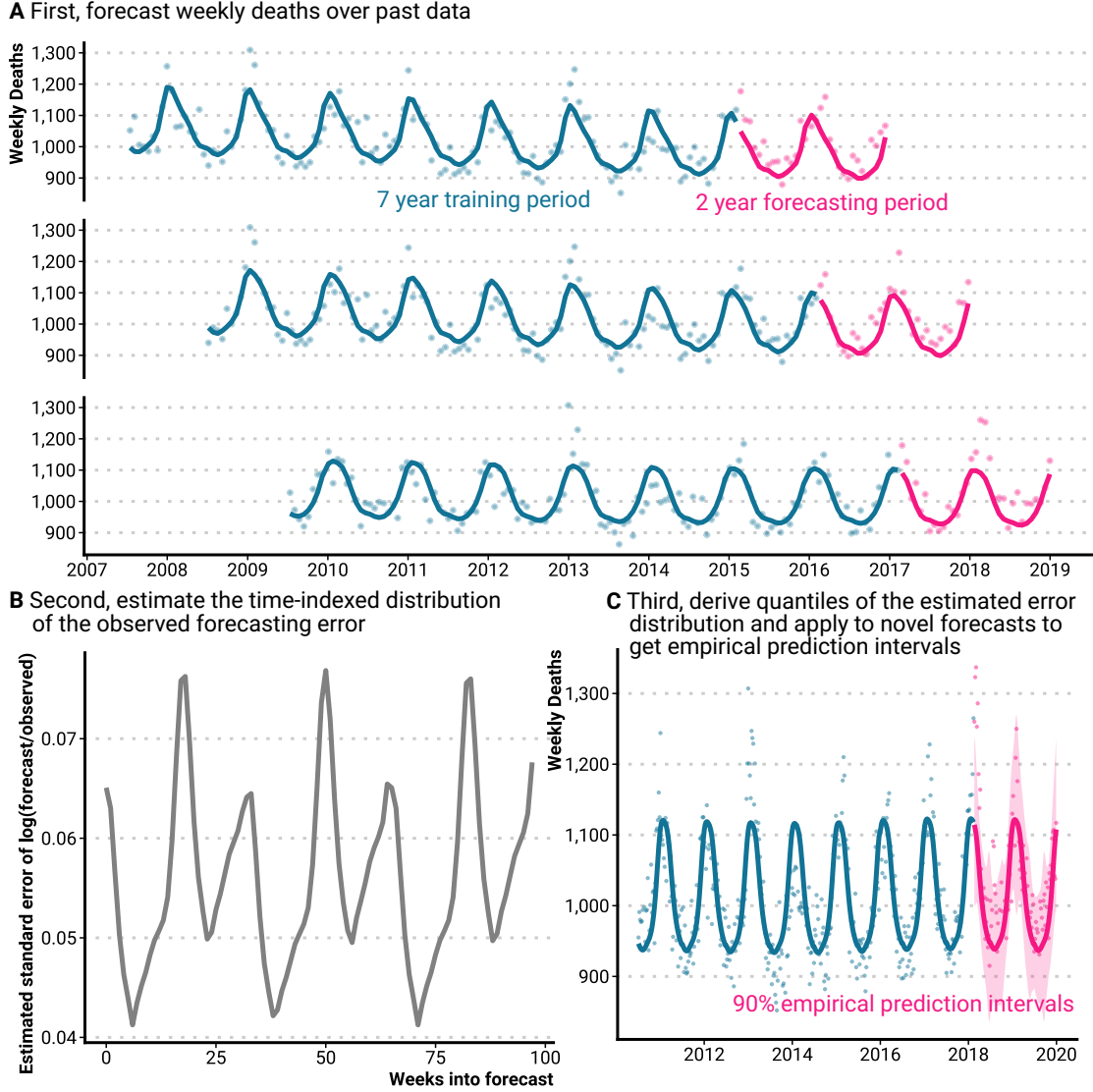90% empirical prediction intervals

Figure 1: Empirical prediction intervals are derived from the time-indexed distribution of past forecasting errors. Shown here is the example for expected Danish weekly death counts for the years 2018 to 2020.

First, we modelled the expected deaths at time t (measured in weeks since some origin), $\hat{y}_t$, as

$$\log \hat{y}_t = \alpha + \beta_t t + s_w(w[t]),$$

where $\beta_t t$ captures a long term trend and $s_w(w[t])$ is a cyclical penalized spline over week-

of-year to reflect the seasonality of death counts. The weekly deaths were assumed to be distributed as

$$Y_t \sim \text{Neg.Binomial}(\hat{y}_t, \phi).$$

Parametric prediction intervals for this model were derived via sampling of the posterior predictive distribution.

This model is representative of much of the applied work in excess death modeling, where regression approaches are prominent, but note that any other forecasting algorithm could have been chosen instead to demonstrate the method of empirical prediction intervals. We fitted this model over the 7 years fitting period in each data series and predicted over a two year horizon following the training data (Figure 1AC).

As a measure of forecasting error we defined ϵ_t to be the logged ratio between forecasted and observed deaths at time t.

$$\epsilon_t = \log \frac{\hat{y}_t}{y_t}$$

We modelled this error via a normal distribution with zero mean[1] and time-varying standard-deviation σ_t,

$$\epsilon_t \sim \text{Normal}(0, \sigma_t)$$
$$\log \sigma_t = \alpha + \beta_t t + s_w(w[t]),$$

where $\beta_t t$ captures a linear trend in the variance of the log relative forecasting error over time (i.e. increasing uncertainty over the forecasting horizon), and $s_w(w[t])$ captures the seasonality of the error variance as a function of week-of-year. In the context of expected deaths modeling this seasonality is pronounced due to the challenges in predicting the severity of flu-waves during winter. Other models, such as a non-parametric quantile regression, are possible as well. We fitted this model to the out-of-sample forecasting errors of the first three data series (Figure 1B).

Once estimates for σ_t were found we derived the quantiles around the forecasted expected deaths $\hat{y}_t$ of the fourth data series with the help of the standard-normal distribution (Figure 1C):

$$Q_p(\hat{y}_t) = \exp[\log \hat{y}_t + \sigma_t \Phi^{-1}(p)].$$

---

[1]Additionally, we could have modelled the mean error over time in order to estimate and correct for the biases of a forecasting model

Preliminary Results

The empirical prediction intervals display a substantially better coverage compared to the parametric intervals, which are too narrow (Table 1). In the final publication we will demonstrate the consequences of these wider prediction intervals for the analysis of excess deaths during the COVID-19 pandemic. Given the empirical error of our methods for calculating expected deaths, which level of excess deaths can we reliably detect two years into the pandemic?

| | Nominal coverage | | |
	50%	90%	95%
Actual coverage of			
Parametric intervals	30.6	66.3	77.6
Empirical intervals	**49.0**	**87.8**	**90.8**

Table 1: Comparing the nominal and actual coverage of parametric vs. empirical prediction intervals over the forecasting period. Note that the coverage of the empirical prediction intervals is evaluated on a separate validation set from the one used to construct the interval.

References

[1] W. H. Williams and M. L. Goodman. A simple method for the construction of empirical confidence limits for economic forecasts. *Journal of the American Statistical Association*, 66(336):752–754, dec 1971. doi:10.2307/2284223.

[2] Michael A. Stoto. The accuracy of population projections. *Journal of the American Statistical Association*, 78(381):13–20, mar 1983. doi:10.1080/01621459.1983.10477916.

[3] Joel E. Cohen. Population forecasts and confidence intervals for Sweden: a comparison of model-based and empirical approaches. *Demography*, 23(1):105–126, feb 1986. doi:10.2307/2061412.

[4] Stanley K. Smith and Terry Sincich. Stability over time in the distribution of population forecast errors. *Demography*, 25(3):461–474, aug 1988. doi:10.2307/2061544.

[5] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *J. Mach. Learn. Res.*, 9:371–421, June 2008. ISSN 1532-4435. doi:10.5555/1390681.1390693.

[6] Dmitri A. Jdanov, Ainhoa Alustiza Galarza, Vladimir M. Shkolnikov, Domantas Jasilionis, László Németh, David A. Leon, Carl Boe, and Magali Barbieri. The short-term mortality fluctuation data series, monitoring mortality shocks across time and space. *Scientific Data*, 8(1), sep 2021. doi:10.1038/s41597-021-01019-1.