Comparing methods to predict baseline mortality for excess mortality calculations – unravelling 'the German puzzle' and its implications for spline-regression

Tamás Ferenci*

Contents

Abstract

Introduction: The World Health Organization presented global excess mortality estimates for 2020 and 2021 on May 5, 2022, almost immediately stirring controversy, one point of which was the suspiciously high estimate for Germany. Later analysis revealed that the reason of this – in addition to a data preparation issue – was the nature of the spline-model underlying WHO's method used for excess mortality estimation. This paper aims to reproduce the problem using synthetic datasets, thus allowing the investigation of its sensitivity to parameters, both of the mortality curve and of the used method, thereby shedding light on the conditions that gave rise to this error and its possible remedies.

Material and Methods: A negative binomial model was used with constant overdispersion, and a mean being composed of three terms: a long-term change (modelled with a quadratic trend), deterministic seasonality, modelled with a single harmonic term and random additional peaks during the winter (flu season) and during the summer (heat waves). Simulated mortality curves from this model were then analyzed with naive methods (simple mean of the latest few years, simple linear trend projection from the latest few years), with the WHO's method and with the method of Acosta and Irizarry. Four years of forecasting was compared with actual data and mean squared error (MSE), mean absolute percentage error (MAPE) and bias were calculated. Parameters of the simulation and parameters of the methods were varied.

Results: Capturing only these three characteristics of the mortality time series allowed the robust reproduction of the phenomenon underlying WHO's results. Using 2015 as the starting year – as in the WHO's study – results in very poor performance for the WHO's method, clearly revealing the problem as even simple linear extrapolation was better. However, the Acosta-Irizarry method substantially outperformed WHO's method despite being also based on splines. In certain – but not all – scenarios, errors were substantially affected by

*Physiological Controls Research Center, Óbuda University and Department of Statistics, Corvinus University of Budapest, ferenci.tamas@nik.uni-obuda.hu

the parameters of the mortality curve, but the ordering of the methods was very stable. Results were highly dependent of the parameters of the estimation procedure: even WHO's method produced much better results if the starting year was earlier, or if the basis dimension was lower. Conversely, Acosta-Irizarry method can generate poor forecasts if the number of knots is increased. Linear extrapolation could produce very good results, but is highly dependent on the choice of the starting year, while average was the worst in almost all cases.

Discussion and conclusion: The performance of the WHO's method with its original parametrization is indeed very poor as revealed by extensive simulations, i.e., the "German puzzle" was not just an unfortunate mishap, however it can be profoundly improved by a better choice of parameters. After that, its performance is similar to that of Acosta-Irizarry method, with WHO dominating for longer fitting periods, Acosta-Irizarry in the shorter ones. Despite simplicity, linear extrapolation could exhibit a good performance, but it is highly dependent on the choice of the starting year; in contrast, Acosta-Irizarry method exhibits a relatively stable performance (much more stable than WHO's method) irrespectively of the starting year. Using the average method is almost always the worst except for very special circumstances. This proves that splines are not inherently unsuitable for predicting baseline mortality, but care should be taken, in particular, these results suggest that the key issue is that the structure of the splines should be rigid. No matter what approach or parametrization is used, model diagnostics must be performed before accepting the results, and used methods should be preferably validated with extensive simulations on synthetic datasets or time series cross validation. Further research is warranted to understand how these results can be generalized to other scenarios.

Introduction

Excess mortality is the difference between the actual mortality (number of deaths) of a time period in a given country or (sub- or supranational) region and its "expected" mortality, defined as the mortality statistically forecasted from the region's historical data. Calculation of excess mortality can be used to characterize the impact of an event on mortality if the historical data is prior to the onset of the event, therefore the prediction pertains to a counterfactual: mortality that would have been observed without the event [1]. Thus, the difference, i.e., excess mortality indeed measures the impact of the event, assuming the prediction was correct.

Calculating excess mortality is particularly useful if the event's impact on mortality is hard to measure directly, for instance, one of the typical applications is characterizing the mortality associated with natural disasters [2–4], but is is also used for epidemics where direct mortality registration is missing, incomplete or unreliable, a prime example being the seasonal flu [5,6].

COVID-19 is no exception. While mortality is reported in developed countries, typically daily or weekly, this suffers from two drawbacks, first, the number of reported deaths is – while to much less extent than the number of reported cases – but still contingent on testing activity, which might be vastly different between countries or time periods, and second, despite efforts at standardization the criteria for the certification of deaths might be different between countries [7]. Excess mortality resolves both of these problems as it is completely immune to differences in testing intensity and cause of death certification procedure. This makes it especially suitable for between-country comparisons, which is a critical issue to better understand the pandemic, in particular, to evaluate different control measures and responses [8].

This, however, comes at a price. First, and perhaps most importantly, excess mortality is inherently a gross metric, measuring both the direct and the indirect effects, the latter of which can be both positive (e.g., COVID control measures also provided protection against flu) or negative (e.g., the treatment of other diseases became less efficient) [9]. Second, the excess mortality is the slowest indicator: the necessary data, that is, the number of deaths usually becomes available with a 4 week lag (and even that is typically revised to some extent later) even in the developed countries. Finally, the whole calculation depends on how accurate the forecast was.

The last of these issues will be the focus now: given the importance of cross-country comparisons, it is crucial the results indeed reflect differences and are not too sensitive to the used prediction method.

Only those methods will be considered now that use traditional regression approach, i.e., methods using

ARIMA models [10–13], Holt-Winters method [14] or based on Gaussian process [15] won't be considered, just as ensemble methods [16,17]. Question concerning age- or sex stratification or standardization [18], small area estimation [19,20] and inclusion of covariates, such as temperature, to improve modelling [16,17,19] will also not be considered here.

This leaves us with two questions, the handling of seasonality and the handling of long-term trend. For the latter, these are the typical solutions in the literature concerning COVID-19:

- Using the last pre-pandemic year [16,21]. This is good – even if not perfect – considering the long-term trends, as it uses data closest to the investigated period, but it has a huge variance, due to the natural year-to-year variability of mortality.
- Using the average of a few pre-pandemic years (typically 5) [22–28]. This is more reliable as averaging reduces variability, however, it is even more biased in case the mortality has a long-term trend (which it almost always has), for instance, if mortality is falling, this provides an overestimation, thus, excess mortality is underestimated.
- Using a linear trend extrapolation [29–31]. This accounts for the potential trends in mortality, removing the bias of the above methods, at least as far as linearity is acceptable, but it depends on the selection of the starting year from which the linear curve is fitted to the data. It also has higher variance than the averaging approach, but is is usually less of concern, given the huge amount of data typically used (unless a small country, and/or age or sex strata is investigated).
- Using splines [32,33]. The method of Acosta and Irizarry [34,35] is based on splines, just as many other custom implementation [16,36], which, crucially, includes the model used by the World Health Organization (WHO) [37].

While this issue received minimal public attention, the choice of the method (and its parameters) to handle long-term trend might have a highly relevant impact on the results of the calculation, as evidenced by the case of the excess mortality estimation of the WHO. On May 5, 2022, WHO published its excess mortality estimates [38], which immediately raised questions: among others, the estimates for Germany were surprisingly high [39]. The reason why it came as a shock is that WHO's estimate, 195 000 cumulative excess death for Germany in the years 2020 and 2021, was inexplicably outlying from every previous estimate, for instance, World Mortality Dataset reported 85 123 deaths for the same period [1].

The case was so intriguing, that one paper termed it the "German puzzle" [39]. Figure 1 illustrates the "German puzzle" using actual German data, with the curves fitted on 2015-2019 data and extrapolated to 2020 and 2021 (as done by the WHO): while the dots visually indicate rather clearly a simple upward trend (as shown by the linear extrapolation), the spline prediction turns back.

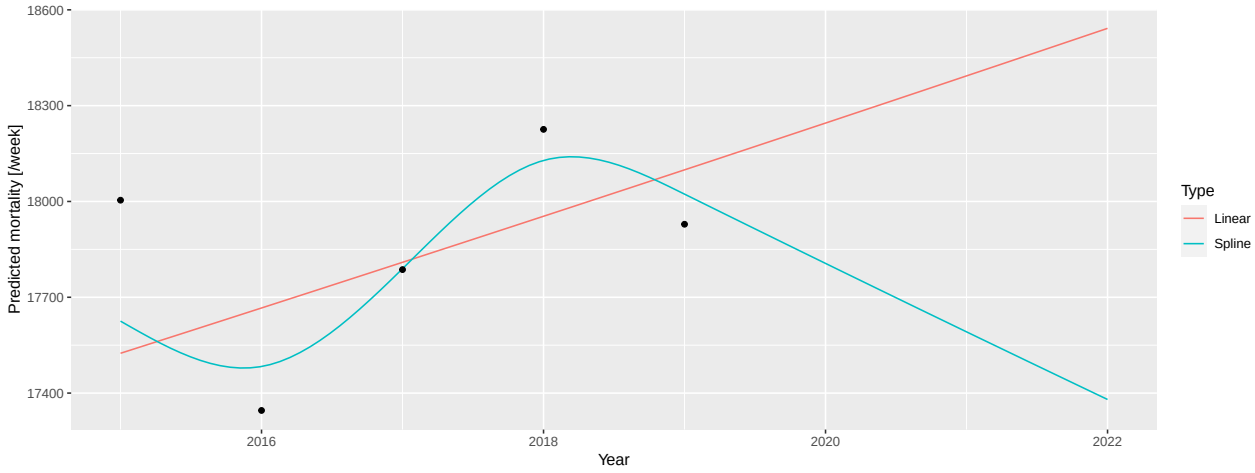

Figure 1: A linear trend and a spline fitted on German mortality data 2015-2019 and extrapolated to 2020 and 2021.

The explanation later provided by the WHO [37] stated that the problem was due to two issues, first, the

WHO applied a rescaling method to the raw data to compensate for underreporting (due to late registration, for instance), but this was unnecessary in case of Germany, with excellent death registration. Figure 1 shows the unadjusted German data avoiding this problem, so that focus can be placed on the second issue that will be the subject of investigation now: the usage of splines.

As described above, WHO's method uses a spline to capture the long-term trend, and it seems that the problem is that the lower data of 2015 had a very high impact on the spline, with that single observation turning the entire spline, despite earlier points showing an upward trend. It seems to much weight is put on this – likely short-term, random, noise-like – fluctuation, i.e., the extrapolation was too sensitive to this. The culprit is quickly identified as spline-regression itself, with one commentator saying "[e]xtrapolating a spline is a known bad practice" [39].

(Interestingly enough, the problem only exists in this particular case if year is used as a predictor. WHO's paper suggests just this [37], but this might be only a typo, as the long-term trend should be represented not by the – abruptly changing – year indicator, but rather a continuously changing indicator of time, such as days since a given date.)

But really splines are to be blamed? Motivated by obtaining a better understanding of the "German puzzle", this paper aims to investigate the following questions: 1) Really splines *per se* were the culprit? 2) What were the particlar characteristics, both of the scenario and of the used spline-regression, that gave rise to the problem? 3) Is there a better way to predict baseline for excess mortality calculation avoiding this problem?

To answer these questions, first a model will be devised that is able to generate mortality curves that capture the relevant features exhibited by the real-life German example. Thus, it'll be possible to calculate the accuracy of a forecast (as the ground truth is now known), and also to investigate how parameters of the simulation influence it. With averaging several simulations, the mean accuracy can be approximated, allowing the comparison of the methods, and investigating its dependence on the parameters – both of the mortality curve and of the parameters of the method – thereby hopefully resolving the "German puzzle".

Material and Methods

Data source

Weekly all-cause mortalities for Germany were obtained from the European Statistical Office (Eurostat), database `demo_r_mwk_ts` [40]. No additional preprocessing or correction was applied such as that for late registration, i.e., the part of the problem with the WHO's approach due to upscaling was avoided, so that the focus is now solely on the modeling aspect. A detailed comparison of the possible data sources can be found in Supplementary Material 1.

Basic properties (raw weekly values, yearly trend, seasonal pattern) are shown on Figure 2.

Simulation model

Based on the patterns that can be observed on Figure 2, the following three components will be used to create synthetic datasets:

- Long-term change, modelled with quadratic trend; described by three parameters (constant, linear and quadratic term)
- Deterministic seasonality, modelled with a single harmonic (sinusoidal) term; described by two parameters (amplitude and phase)
- Random additional peaks during the winter (i.e., flu season) and during the summer (i.e., heat waves); described in each season by five parameters (probability of the peak, minimum and maximum value of the peak height, minimum and maximum value of the peak width)

These govern the expected value; the actual counts are obtained from a negative binomial distribution with constant size. Detailed description of how the above model is built, and what parameters are used can be found in Supplementary Material 2.

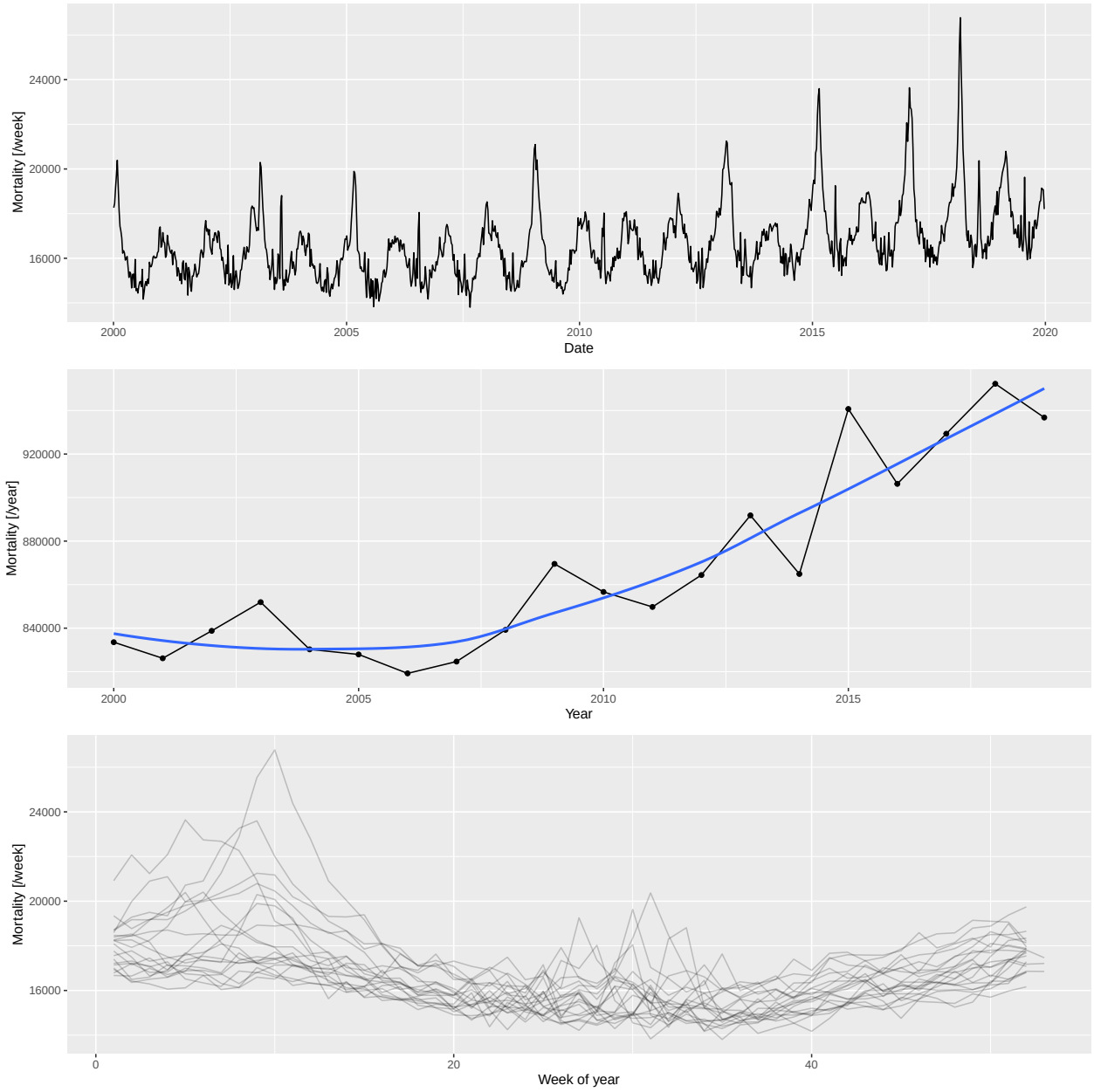

Figure 2: Weekly mortalities (upper), yearly mortalities with LOESS-smoother (middle) and seasonal pattern (bottom) of the German mortality data, 2000-2019.

Baseline mortality prediction

Four methods will be used for predicting mortality, including the WHO's method, an advanced alternative method that also uses splines, developed by Acosta and Irizarry in 2020 [34], and two naive methods as a comparison. These cover the widely used, classical statistical methods used for predicting baseline mortality in excess mortality studies.

- Average: after accounting for seasonality with a single cyclic spline, the average of the preciding years will be used as the – constant – predicted value. Parameter: starting year (i.e., how many previous year is used for averaging). Some studies used the last pre-pandemic year (2019) as the predicted baseline mortality, this is just the special case of this method, with the starting year set to 2019.
- Linear: after accounting for seasonality with a single cyclic spline, the long-term trend is modelled with a linear trend, that is extrapolated. Parameter: starting year (from which the model is fitted).
- WHO's method: the method is reconstructed from the description provided in [37]. In brief, seasonality is accounted with single a cyclic spline (as done in the previous cases), and the long-term trend is accounted with a thin plate regression spline. The only deviation compared to WHO's paper is that – as noted above – the actual time (number of days since 1970-01-01) is used as the predictor for long-term trend, not the abruptly changing year. The model is estimated with restricted maximum likelihood (REML). Parameters: starting year (from which the model is fitted) and k, the dimension of the basis of the spline used for capturing the long-term trend.
- Acosta-Irizarry (AI) method: the method described in [34] using their reference implementation. Parameters: starting year (from which the model is fitted) and *tkpy*, the number of trend knots per year; other parameters are left on their default values (i.e., two harmonic term is used).

Validation through simulation

First, a synthetic dataset is randomly generated from the model described above using the investigated parameters (parameters of the scenario). This dataset simulates mortalities from the beginning of 2020 to the end of 2023. Then, the investigated prediction method with the investigated parametrization (parameters of the method) is applied, and after fitting, a prediction is obtained for 2020 to 2023. The goodness of prediction is quantified as mean squared error (MSE), mean absolute percentage error (MAPE) and bias in those 4 years based on 1000 replications of this simulational procedure. This is repeated for all parameters of the scenario, all prediction methods and all parameters of the method.

Investigated parameters of the prediction methods were the following:

- Average: starting year 2000, 2005, 2010, 2015, 2019
- Linear: starting year 2000, 2005, 2010, 2015
- WHO's method: all possible combination of starting year 2000, 2005, 2010, 2015 and k (basis dimension) 5, 10, 15, 20
- Acosta-Irizarry method: all possible combination of starting year 2000, 2005, 2010, 2015 and *tkpy* (trend knots per year) 1/4, 1/5, 1/7, 1/9, 1/12

For the scenario, simulations were run with the optimal parameters discussed in Supplementary Material 2 (base case scenario) and then all parameters were varied in 5 steps from half of the base case value to twice of that, with two exception: the constant term of the trend is varied only from 90% to 110% (to avoid irrealistic values) and probabilities are prevented from going above 100%. Only one parameter was varied at a time, the others being fixed at their base case value. That is, no interaction was investigated: while this could be potentially interesting, it has a very high computational burden.

Details are provided in Supplementary Material 3.

Programs used

All calculations are carried out under the R statistical program package version 42.0 [41] using packages `data.table` [42] (version 1.14.2) and `ggplot2` [43] (version 3.3.6), as well as `excessmort` (version 0.6.1), `mgcv` (version `rpackageVersion("mgcv")`), `scorepeak` (version 0.1.2), `parallel` (version 4.2.0) `lubridate` (version 1.8.0), `ISOweek` (version 0.6.2) and `eurostat` (version 3.7.10).

Full source code allowing complete reproducibility is openly available at https://github.com/tamas-ferenci/MortalityPrediction.

Results

Figure 3 illustrates the predictions (for 2020-2023) for the base case scenario by showing the estimated yearly deaths for 200 randomly selected simulation together with the ground truth for all 4 method with all possible parameters.

Figure 3 already strongly suggests some tendencies, but to precisely evaluate it, error metrics have to be calculated. Figure 4. shows all three error metrics for WHO and Acosta-Irizarry methods, for all possible parametrizations.

As already suggested by Figure 3, this confirms that $k = 3$ (WHO) and $tkpy = 1/7$ (Acosta-Irizarry) are the best parameters in this particular scenario. Note that the default value in the method used by the WHO is $k = 10$, but it is just $tkpy = 1/7$ for the Acosta-Irizarry method. It worth comparing all four methods with different starting years, but with the remaining parameters (k for WHO's method, $tkpy$ for the Acosta-Irizarry method) set to the values that are optimal in this particular scenario (Figure 5).

All the above investigations used the base case scenario for the simulated mortality curve. Figure 6. shows the best error metrics achievable with each method in a given scenario.

Finally, note that as different methods were evaluated on the same simulated dataset for each simulation, it is possible to compare not only the averages, but directly compare the errors themselves. Figure 7 shows direct comparison between the best parametrization of the WHO's method and the Acosta-Irizarry method for 200 randomly selected simulations in each scenario, with 2015 as the starting year. In the base case scenario, Acosta-Irizarry performed better in 59.8% of the cases.

Discussion

These results demonstrate that we were able to reliably reproduce the "German puzzle" using synthetic datasets. This approach allowed a deep investigation of how the results depend on the used method, its parameters and on the parameters of the scenario.

As expected, prediction with average has the highest error, and is highly biased (but is improved by shortening the fitting dataset). This of course depends on the form of the historical mortality curve, theoretically, for a more or less constant curve even this prediction can work better.

Linear extrapolation seems to be a very promising alternative, the only problem being that it is very sensitive to the appropriate choice of the starting year: that phase should be covered where the change in historical mortality is linear. This is prone to subjectivity and might not work at all if the linear phase is too short (limiting the available information), i.e., it depends on how wiggly is the historical curve.

Splines in contrast can work theoretically well even in those cases: it can use all historical data, i.e., it is not abruptly cut off as with linear extrapolation, but more weight is placed on the trends suggested by the recent observations. At first glance, this seems to be the ideal solution, but as this investigation reveals, what is meant by "more weight" and "recent" is crucial, and certain choices can results in very poor extrapolations, despite the tempting theoretical properties.

The overall picture in selecting the optimal parameters, confirmed by the results of both spline-based methods, is that splines should be quite rigid in baseline mortality prediction for excess mortality calculation. This is the concordant conclusion from the experiences both with the WHO method (where increasing basis dimension decreased performance) and the Acosta-Irizarry method (where increasing trend knots per year decreased performance).

The likely explanation is that mortality curves exhibit only slow changes, so high flexibility is not required, and – as with any regression model with too high model capacity – can be downright detrimental, as it allows the model to pick up noise, i.e., can result in overfitting.

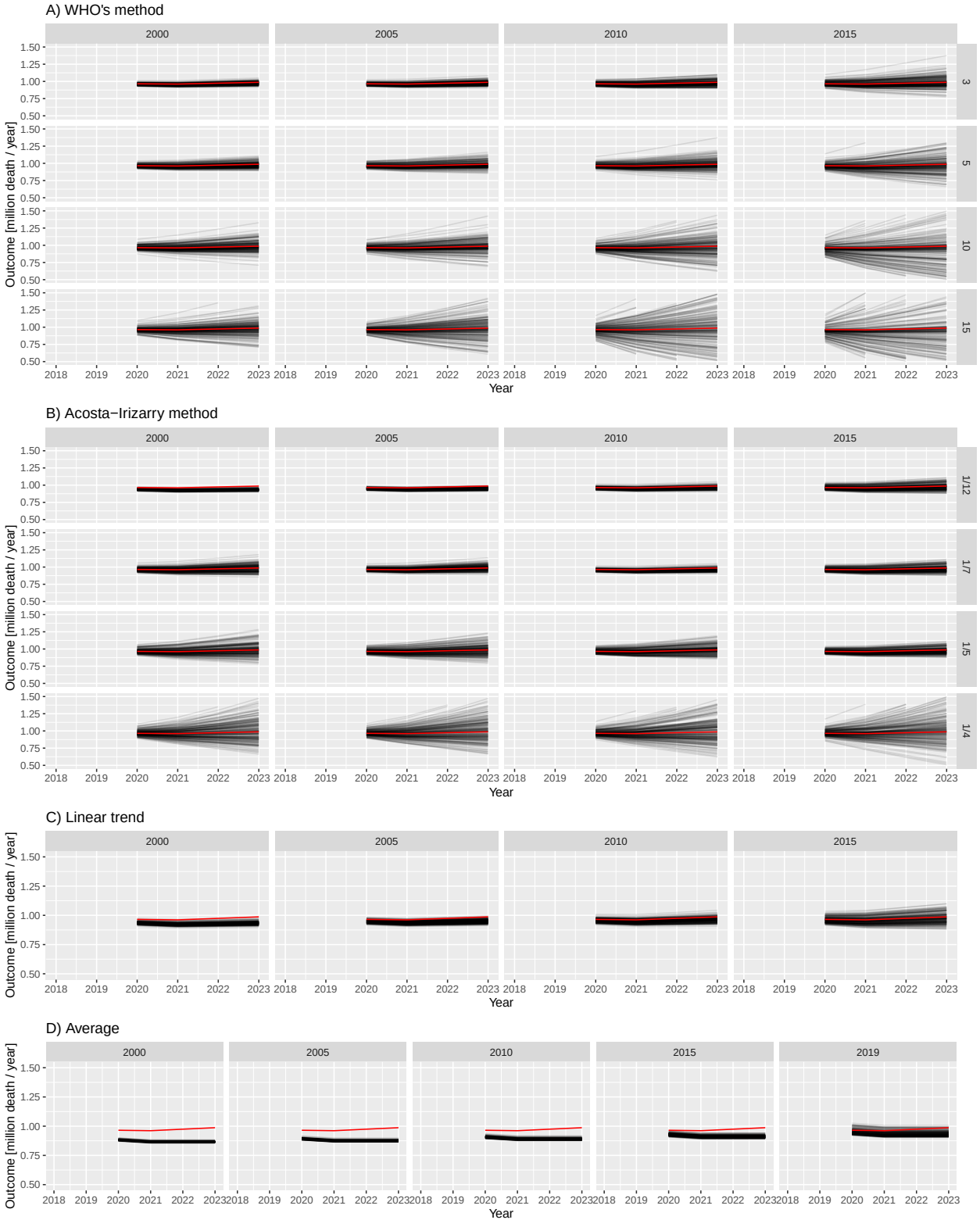

Figure 3: Estimated yearly deaths (for 2020-2023) for 200 randomly selected simulation together with the ground truth. A) WHO's method, B) Acosta-Irizarry method, C) Linear trend, D) Average. Parameters of the methods are shown in column and row headers. Parameters of the scenario are set to the base case values.

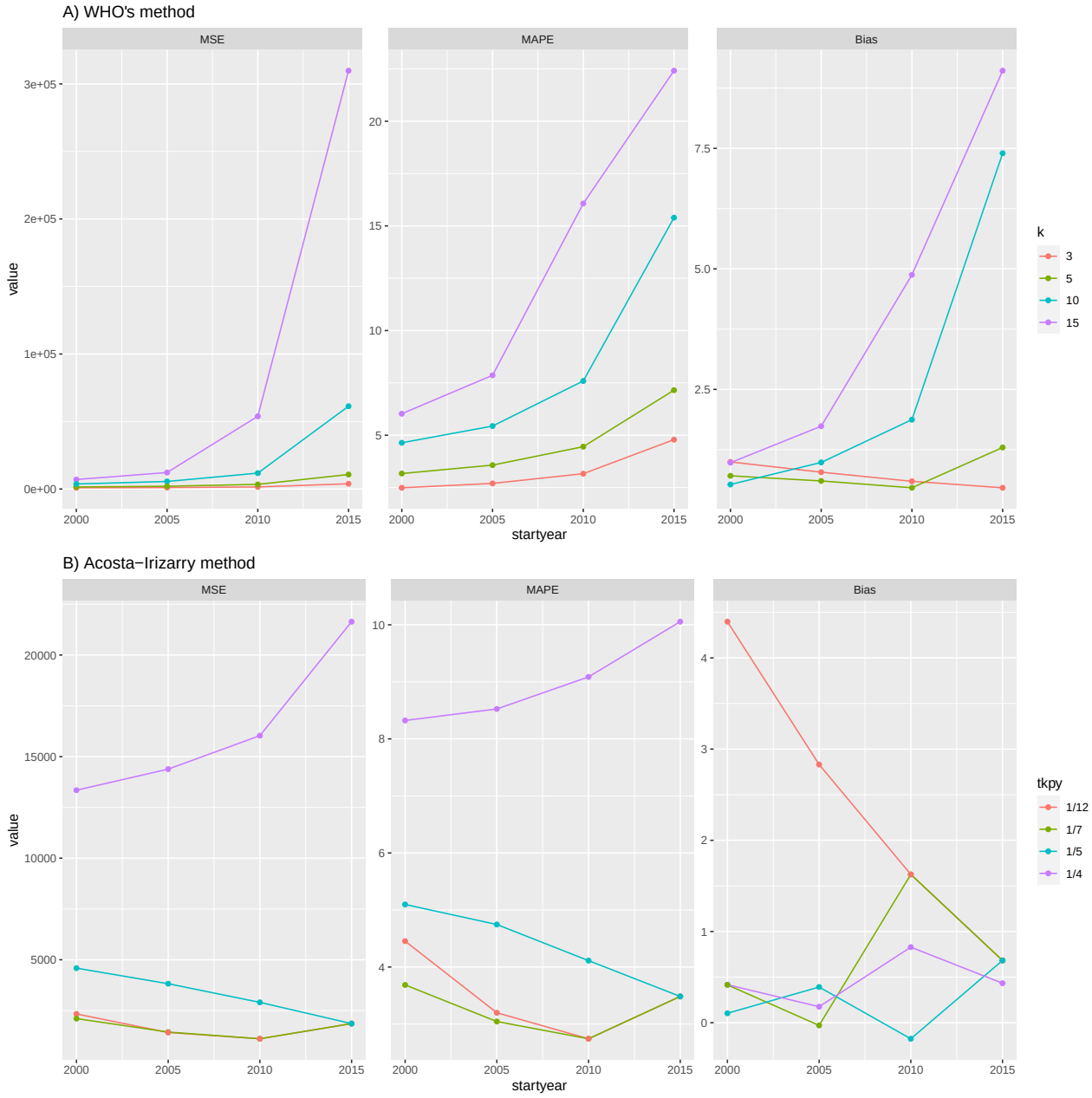

Figure 4: Different error metrics (MSE, MAPE, Bias) for the WHO's method (above) and the Acosta-Irizarry method (below) for all possible parameter combinations of these two methods. Parameters of the scenario are set to the base case values.

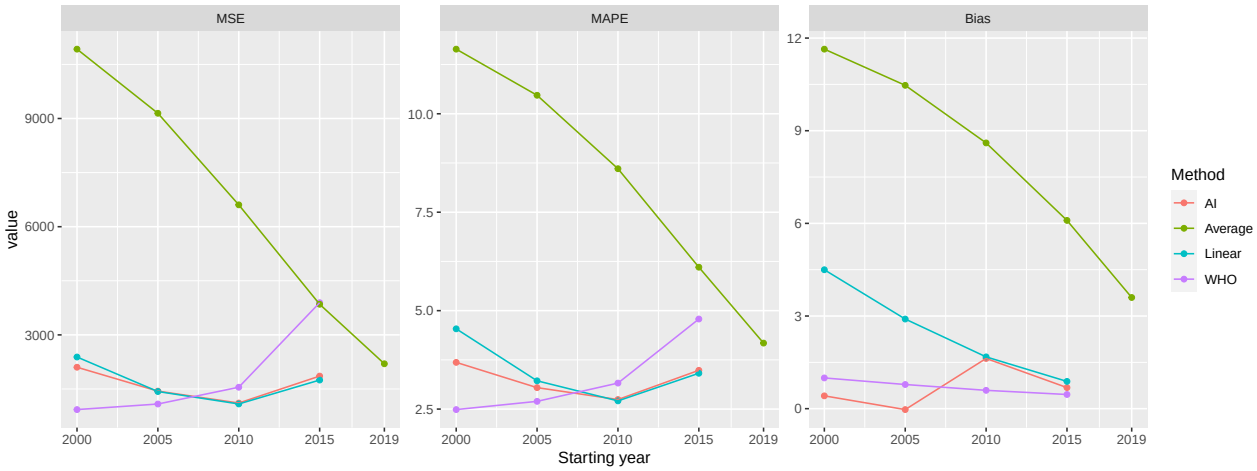

Figure 5: Different error metrics (MSE, MAPE, Bias) for all method, with different starting years, but remaining parameters set to optimal values for this particular scenario. Parameters of the scenario are set to the base case values.

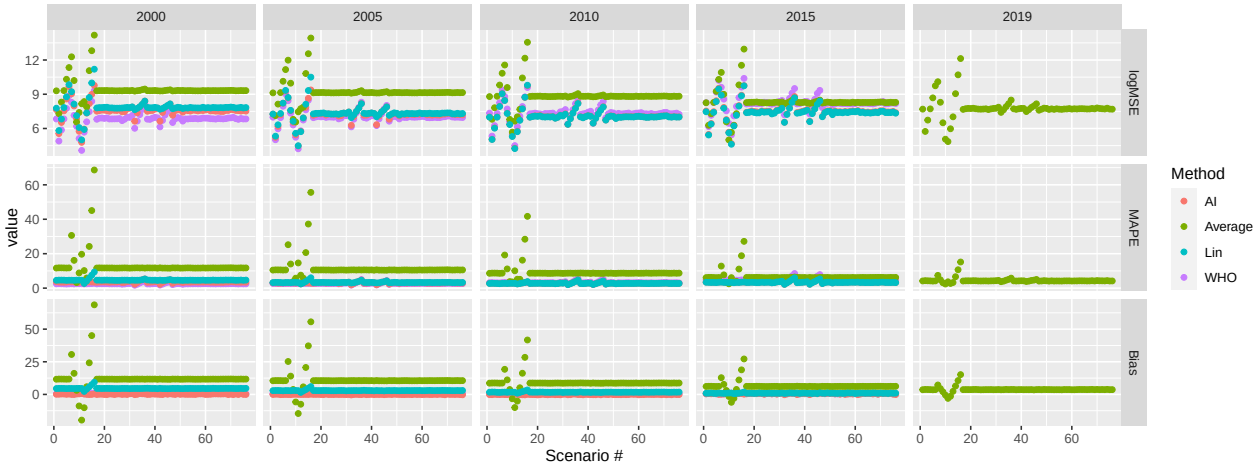

Figure 6: Best achievable error metrics with each method in each simulational scenario of the mortality curve (#1 is the base case scenario).

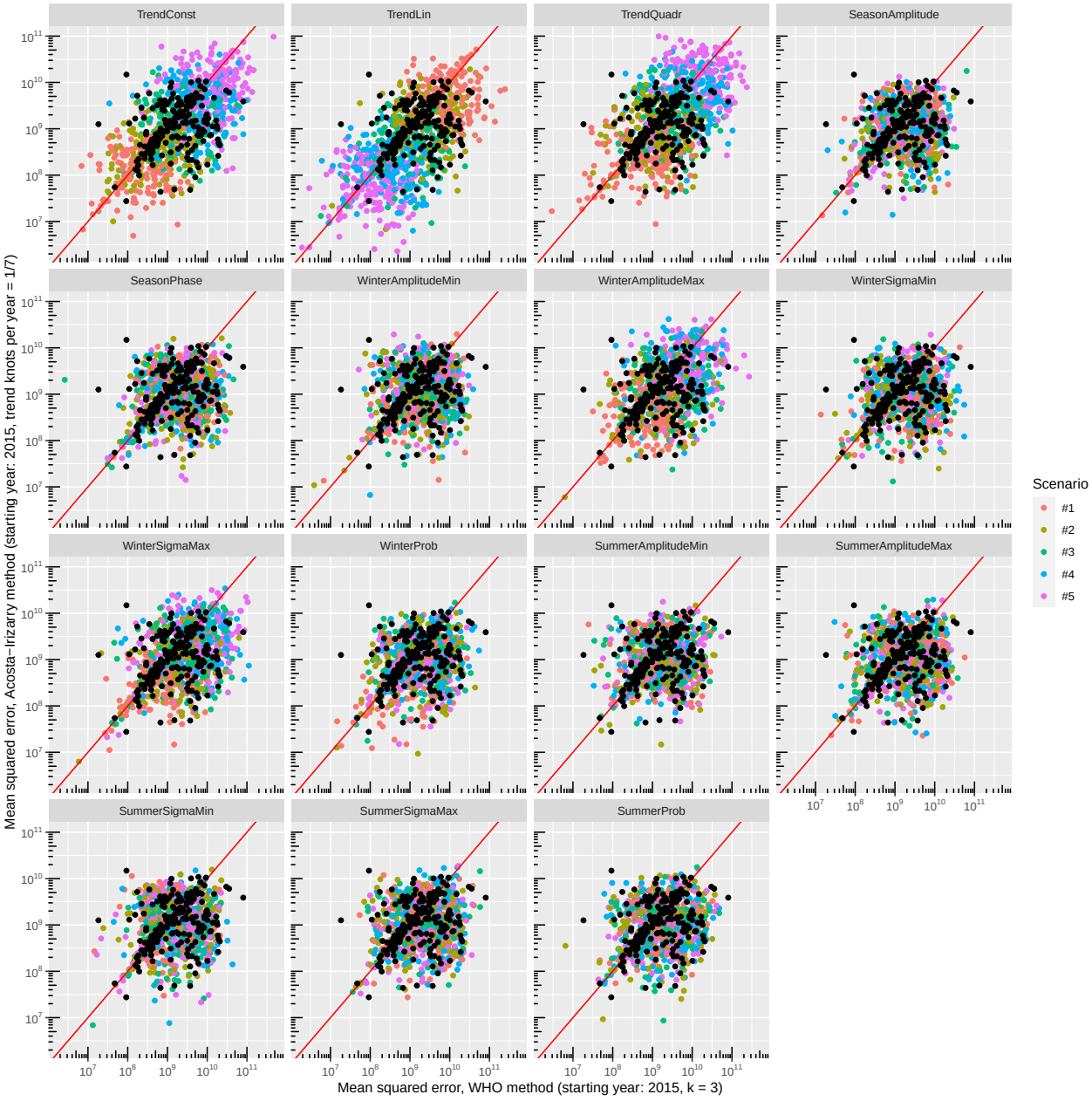

Figure 7: Errors – squared distance of the predicted outcome from its true value – of the WHO's method and the Acosta-Irizarry method (under best parametrization) on the same simulated datasets for 200 randomly selected simulations with different scenarios. Black dots indicate the base case scenario, scenario #1 to #5 represent varying the parameter shown on the panel from half of its base case value to twice (with the exception of the constant term where it is varied from 90% to 110%). Probabilities are limited to be below 100%.

In Germany, the data for 2019 was somewhat lower, likely due to simple random fluctuation, but unfortunately the spline was flexible enough to be "able to take this bend". Note that data are presented using the ISO 8601 year defintion meaning that year can be either 52 or 53 weeks long [44]; from 2015 to 2019 every year is 52 weeks long except 2015 which is one week longer. This adds to the reasons why the value of 2015 is higher, increasing the wiggliness of the German data and thereby potentially contributing to the problem, as the increased wigliness of the data forces the thin plate regression spline used in the WHO's method to be more flexible.

The WHO method is only acceptable with $k \leq 5$ (but even that requires longer observation than starting from 2015, as was done by the WHO), not higher. For the Acosta-Irizarry method, 1/4 trend knots per year was definitely too flexible, perhaps even 1/5 is too high. Note that the default value in the reference implementation of the Acosta-Irizarry method is 1/7, and authors in fact do not recommend using a value much higher. The WHO's paper unfortunately does not specify what basis dimension was used [37], but the default of the package used there is $k = 10$, so even $k = 5$ is substantially lower, not to speak of $k = 3$. This is likely a crucial component in WHO's experience, where the starting year was 2015 (and probably $k = 10$ was used).

Among the two spline-based methods, when the rigidity parameters were used that are optimal in this particular scenario, WHO's method performed better with longer fitting periods, Acosta-Irizarry performed better with shorter ones. However, the performance of the Acosta-Irizarry method was much less dependent on the starting year (i.e., its disadvantage compared to WHO's method was less for earlier starting year than that of the WHO's method for later starting years).

We are aware of two previous works from the literature that are comparable to the present investigation. Both Nepomuceno et al [45] and Shöley [46] is similar to ours in a sense that they used – among others – several models, partly overlapping those presented here, but, importantly, neither of them considered splines for long-term trend at all. Nepomuceno et al did not try to evaluate the methods, it compared them to each other without having ground truth, i.e., the aim was to investigate concordance. In contrast, Shöley did try to give an objective evaluation, but in contrast to the synthetic dataset simulation approach used here, it applied time series cross-validation with historical data to measure accuracy. Cross-validation has the advantage that is guaranteed to be realistic – as opposed to a simulation – but there is less freedom, as investigators are bound to empirical data, with limited possibility in varying the parameters.

Thus, we believe that this is the first systematical study to investigate the application of splines for the long-term trend component in mortality prediction for excess mortality calculation, the first to investigate the impact of their parametrization, and the first to use synthetic dataset validation in general for that end.

The most important limitation of the present study is that every simulation is committed to the same model structure, for instance, long-term trend is limited to be quadratic. It would be important to extend the present investigation to other long-term trends, and to other models in general (such as those with different seasonality, or interaction between seasonality and trend etc.).

We did not investigate the impact of using the population and modelling death rates versus modelling death counts directly, nor did we examine the potential impact of the frequency of the data. Weekly data was used throughout this study: this might not be available for developing countries, but it is almost universally available for developed countries, in which case, the use of the most frequent data seems to be logical (with the appropriate handling of seasonality). In the same vein, adjustment for late registration and imputation of missing data, which might be needed where full data is not available, is not considered here, as the focus is on developed countries.

Conclusion

The performance of the WHO's method with its original parametrization is indeed very poor as revealed by extensive simulations, i.e., the "German puzzle" was not just an unfortunate mishap, however it can be profoundly improved by a better choice of parameters. After that, its performance is similar to that of Acosta-Irizarry method, with WHO dominating for longer fitting periods, Acosta-Irizarry in the shorter ones. Despite simplicity, linear extrapolation could exhibit a good performance, but it is highly dependent on the

choice of the starting year; in contrast, Acosta-Irizarry method exhibits a relatively stable performance (much more stable than WHO's method) irrespectively of the starting year. Using the average method is almost always the worst except for very special circumstances.

This proves that splines are not inherently unsuitable for predicting baseline mortality, but care should be taken, in particular, these results suggest that the key issue is that the structure of the splines should be rigid. No matter what approach or parametrization is used, model diagnostics must be performed before accepting the results, and used methods should be preferably validated with extensive simulations on synthetic datasets or time series cross validation. Further research is warranted to understand how these results can be generalized to other scenarios.

Supplementary Material 1: Comparison of data sources

For a country like Germany, four data sources come into consideration for weekly mortality data: Eurostat [40], the Short-Term Mortality Fluctuations (STMF) dataset of the Human Mortality Database (WMD) [47], the World Mortality Database [1] and the national data provider (in this case, the Federal Statistical Office of Germany). The last is usually more complicated, limits extension to other countries and is unnecessary for developed countries, so it'll be avoided in this case. Also, for Germany, WMD simply copies the data of the STMF ("We collect the weekly STMF data for the following countries: [...] Germany, [...]") leaving us with two options.

We shall compare whether these two report identical data (Figure 8).

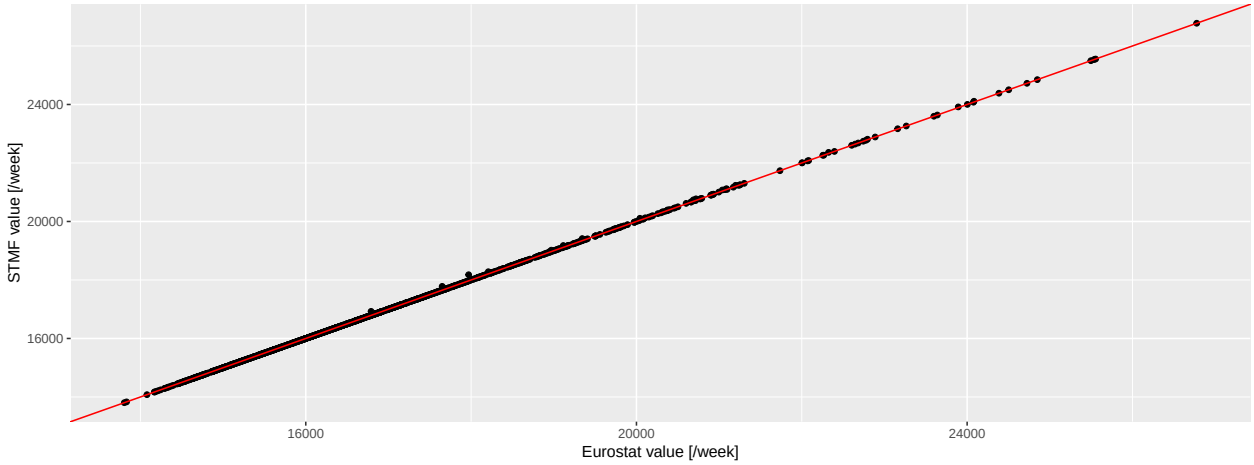

Figure 8: Supplementary Figure 1: Weekly number of deaths according to the Eurostat (horizontal axis) and the STMF database (vertical axis) in Germany. Red line indicates the line of equality.

The two are almost identical (with a correlation of 0.9999863), with differences only occuring for the latest data and of minimal magnitude, so we can safely use the Eurostat database.

Supplementary Material 2: Generating realistic synthetic datasets

First, Figure 2 should be inspected, as it already gives important clues on the setup of a realistic model from which synthetic datasets could be generated. Figure 9 gives further insight by plotting each year separately.

The following observations can be made:

- There is a long-term trend, seemingly quadratic.
- There is a strong seasonality with winter peak and summer trough.
- There are peaks – in addition to the seasonality – in the winter and also during the summer (although the shape seems to be different, with winter peaks seeming to be broader and higher).

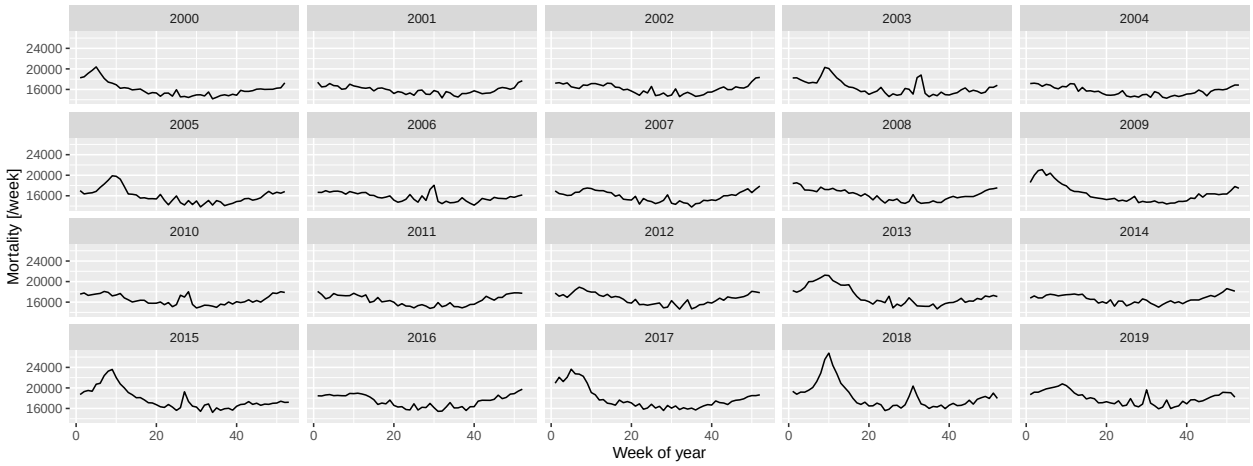

Figure 9: Weekly number of deaths in Germany, separated according to year.

To investigate these, first a spline-smoothing – with thin plate regression spline [32] – is applied to obtain the long-term trend, and a single harmonic term is included as a covariate to remove seasonality. No interaction is assumed between the two, i.e., it is assumed that the seasonal pattern is the same every year. (The peaks are not accounted for at this stage which means that the curve is above the true one, but the difference is likely has minimal due to the rarity and short duration of the peaks. This will be later corrected, after the peaks were identified.) All analysis will be carried out on the log scale (meaning the effect of covariates is multiplicative) using negative binomial response distribution to allow for potential overdispersion [48].

Figure 10 shows the results, overplotted with the model where the long-term trend is a completely parametric quadratic trend. One can observe very good fit between the two, so all subsequent investigation will use the quadratic trend which is much easier to handle. This is only meaningful for short-term extrapolation, but this is what will be needed now (two years of extrapolation will be used in the present study); also it is not possible to better differentiate between functional forms at this sample size.

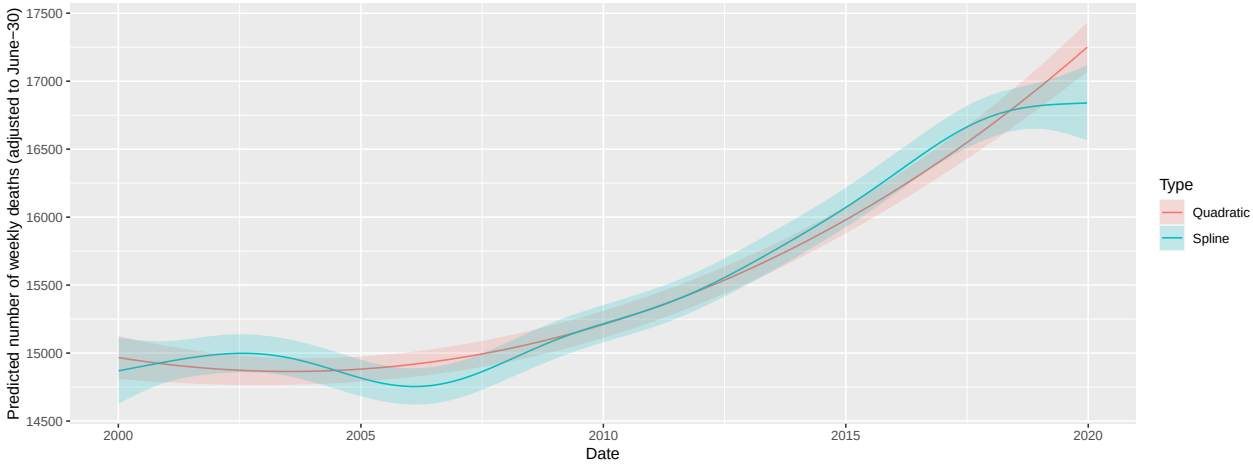

Figure 10: Fitting long-term trend as spline (black) and as quadratic trend (red); shaded area indicated 95% confidence interval. Seasonality is removed by including a single harmonic term in the regression in both cases.

The coefficients can be transformed to equivalent forms that are more meaningful. Thus, the three parameters of the quadratic trend can be expressed as a minimum point (2003-07-15), value at the minimum (15918.54) and value at the end of 2020 (18825.83). (This differs from the value seen on Figure 10, as that also includes

the effect of the harmonic term.) The two parameters of the harmonic regression can be expressed as an amplitude, a multiplier (9.4%) and a phase shift (-0.7, i.e., minimum at week 32 of the year).

Figure 11 shows the predictions of the above model.

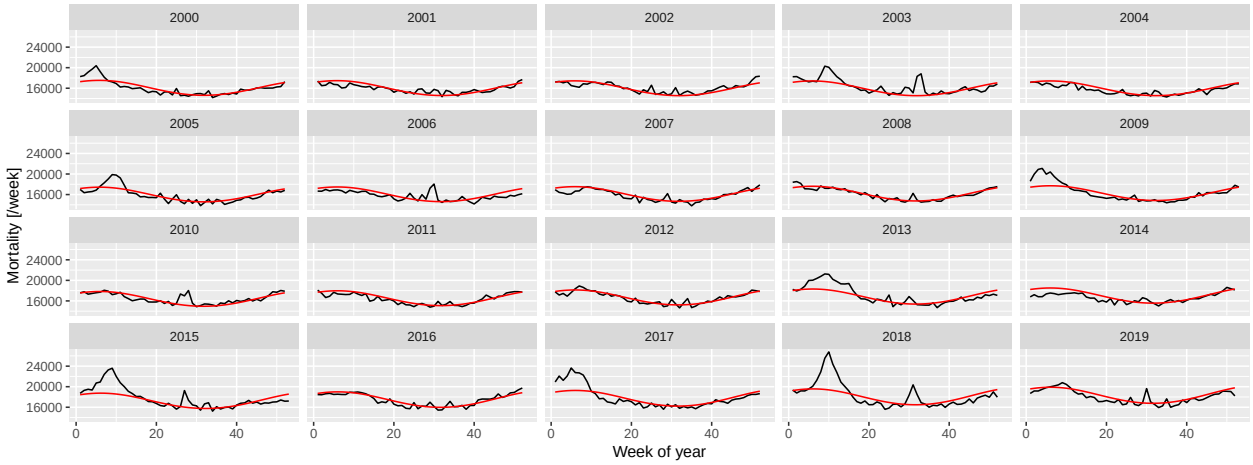

Figure 11: Weekly number of deaths in Germany, separated according to year, showing the predictions from the model with quadratic long-term trend and a single, fixed harmonic term.

A good fit can be observed, apart from summer and winter peaks. Thus, to capture them, the predictions are subtracted; with the results shown on Figure 12.

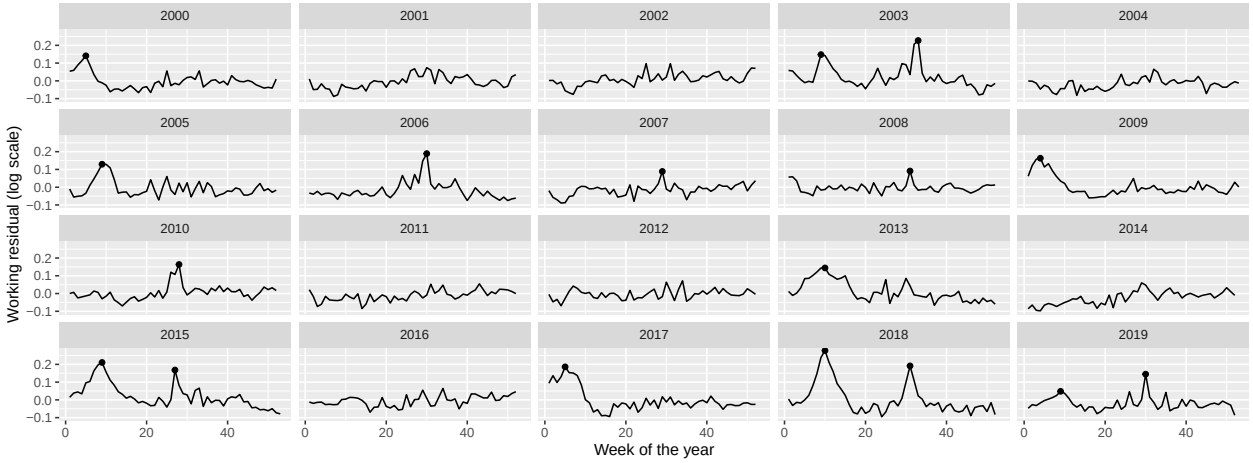

Figure 12: Residuals of the fitted model with quadratic long-term trend and a single, fixed harmonic term. Dots indicate identified peaks.

Peaks in the residuals were identified with the peak detector of Palshikar [49] using parameters that were empirically tuned to identify to visually clear peaks. Results are shown on Figure 12 with black dots; an indeed good identification of the unequivocal peaks can be seen.

Figure 13 shows the peaks themselves with a $\pm$ 100 days neighbourhood. This reinforces the idea that summer and winter peaks are somewhat different, but more importantly, suggests that the rescaled probability density function of the Cauchy distribution, i.e., $\frac{a}{\pi s}\frac{1}{1+\left(\frac{x-x_0}{s}\right)^2}+b$ might be a good – and parsimonious – function form to capture the shape of the peaks.

To check this theory, the best fitting function was found for each peak individually using the Nelder-Mead method [50] with mean squared error objective function. Results are shown on 13 as red lines; an almost

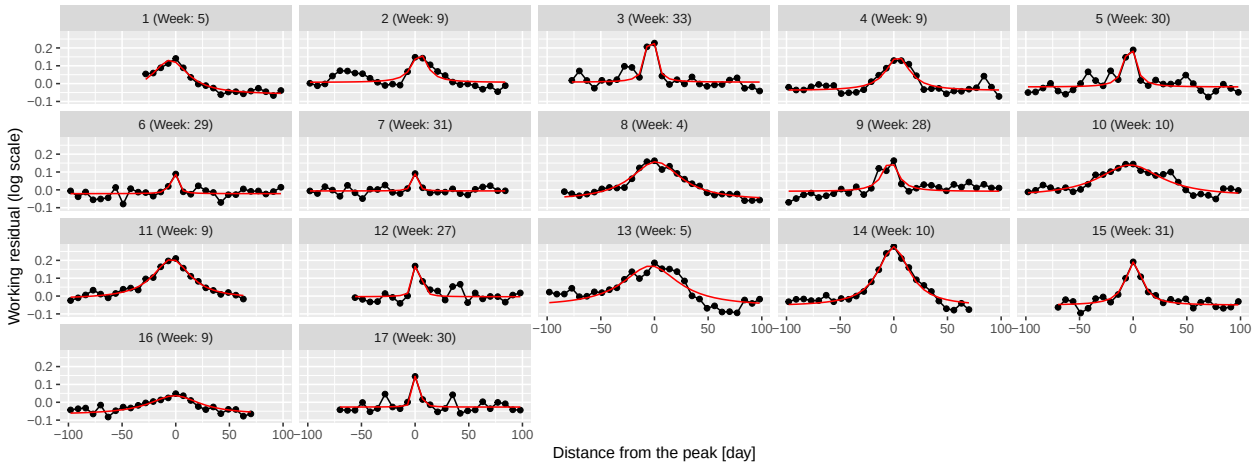

Figure 13: 100-day width neighbourhood of the identified peaks. Red line indicates the best fitting rescaled Cauchy density.

perfect fit can be observed for all peaks confirming the initial idea of using Cauchy density.

This now puts us in a position to investigate the distribution of the parameters (i.e., a, b, s and x_0), which is shown on Figure 14 separated according to whether the peak is during the summer or not. Peak height is also calculated, defined as height at zero (which is $\frac{a}{\pi s\left(1+\frac{x_0^2}{s^2}\right)}$) not the actual maximum height (which is at x_0) to avoid extremely large heights – which were never actually observed – due to peaks with small s, i.e., very narrow peaks.

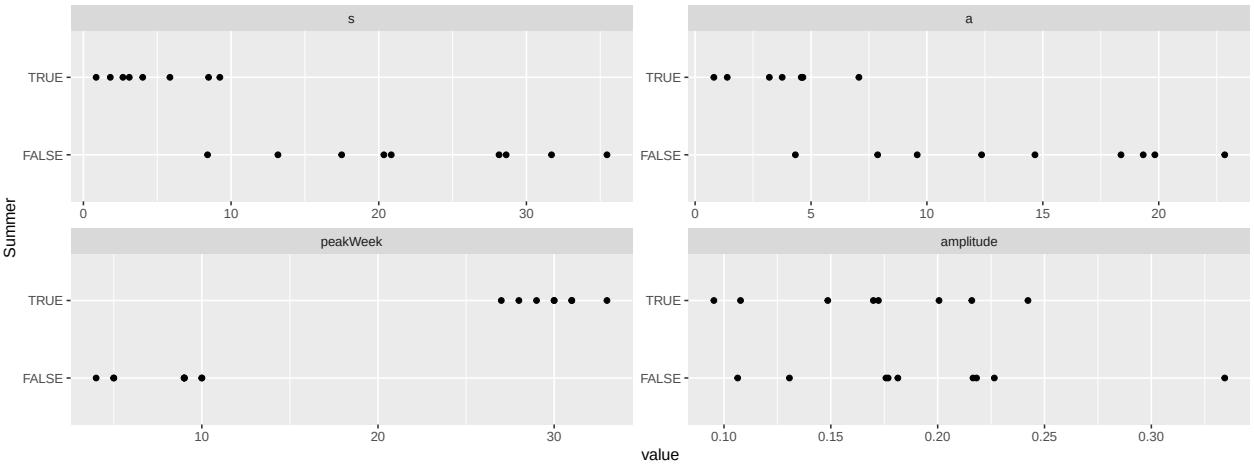

Figure 14: Distribution of the parameters of the best fitting rescaled Cauchy densities for each peak, separated according to whether the peak is during the summer.

This verifies that the width is indeed different, with the s of the summer peaks being below 10, and the winter peaks being above, i.e., summer peaks are shorter in duration, raise and fall faster. Interestingly, the peak heights are not substantially different between winter and summer. Also note that the probability of having a peak at all is different: there are 8 summer peaks and 9 winter peaks (from 20 years). Winter peaks occur between weeks 4 and 10, summer peaks occur from weeks 25 to 35.

This allows the removal of the peaks (Figure 15), and, after these peaks are removed, it is possible to re-estimate trend and seasonality, now without the biasing effect of the peaks. This "bootstrap" procedure is adequate after this second iteration, as no further peaks can be seen after the removal of the re-estimated

trend and seasonality.

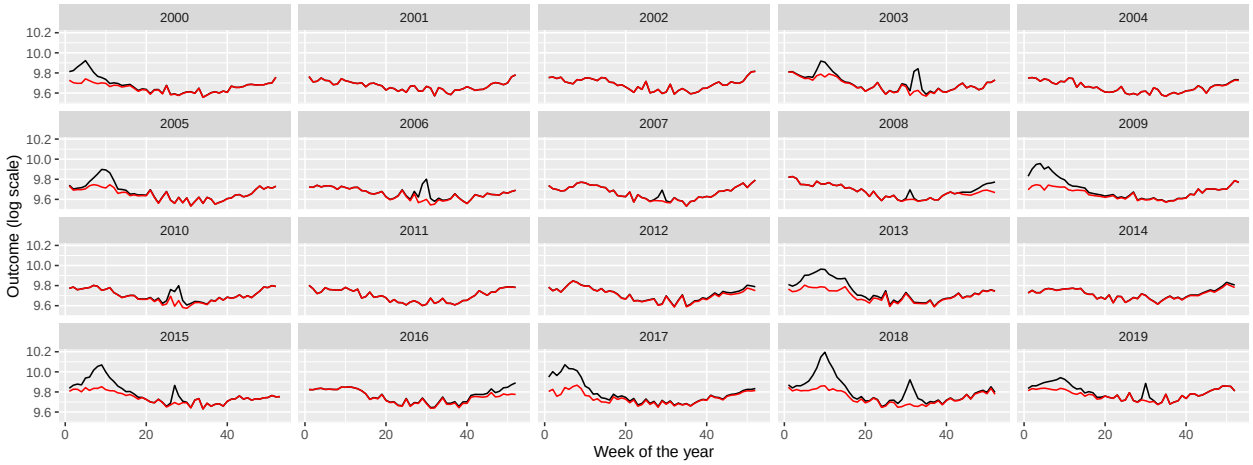

Figure 15: Weekly number of deaths in Germany, separated according to year (black) and with peaks removed (red).

Creating the appropriate model to simulate such peaks is not straightforward: there is stochasticity in the position of the peaks and in its shape, i.e., height and width. (Actually, even the presence of a peak is stochastic.) The following procedure will be used: the presence is generated as a Bernoulli random variate (with the probabilities described above, different for summer and winter), the onset date is uniformly distributed (from 0 to 0.2 in scaled weeks for the winter peak and from 0.5 to 0.7 for the summer peak), i.e., the position itself is random, but the parameters of the underyling distribution are fixed. The b parameter is set to zero (irrespectively of its estimated value, to really capture only the peak, locally – a non-zero b would mean a non-local effect), while s and a are randomly generated for each peak, again, separately for summer and winter peaks. Given the high correlation between s and a, not these, but rather s (width) and the amplitude will be generated as a random variate from – independent – uniform distributions. The parameters (minimum and maximum) of the uniform distributions both for s and the amplitude will be considered as a parameter (hyperparameter) of the simulational procedure, just as the p probability of the Bernoulli distribution, all different for summer and winter.

The following table summarizes the parameters:

Parameter	Value
Linear term of the trend	10.11
Constant term of the trend	0.00
Quadratic term of the trend	0.00
Amplitude of seasonality (log scale)	0.07
Phase of seasonality	-0.61
Minimum of winter peak amplitude (log scale)	0.11
Maximum of winter peak amplitude (log scale)	0.33
Minimum of winter peak width	8.41
Maximum of winter peak width	35.46
Probability of winter peak	0.45
Minimum of summer peak amplitude (log scale)	0.10
Maximum of summer peak amplitude (log scale)	0.24
Minimum of summer peak width	0.86
Maximum of summer peak width	9.24
Probability of summer peak	0.40

Given the mechanism described above, the procedure to simulate synthetic datasets can easily be created. Of course, as every calculation is carried out on the log scale, the mean should be exponentiated at the last step.

Figure 16 illustrates the synthetic data set creation with a single simulation. (Of course, to assess the correctness of the simulation, several realizations have to be inspected.) In addition to the plots of 2, it also gives the autocorrelation function so that it can also be compared. (The simulated outcomes are themselves independent – meaning that effects like the increased probability of a flu season if the previous year did not have one are neglected –, but the trend, seasonality and the peaks induce a correlation structure.)

Several simulations confirm an overall good fit between the actual data set and the simulated ones. Thus, it is now possible to investigate the properties of the mortality prediction on algorithms using simulated datasets, where the actual outcome is known, and the parameters can be varied.

Supplementary Material 3: Validation through simulation

Two sets of parameters have to be set up: parameters of the simulation (i.e., parameters of the scenario, on which the methods will be run) and parameters of the methods. They're set up as given in the main text.

One thousand simulation will be run for each parameter of the scenario, and for each of the 1000 simulated data set all 4 methods with all possible parameters of the methods will be evaluated. To increase the speed, simulations will be run in parallel. (The problem is embarrassingly parallel, as different simulations are completely independent of each other [51].)

Acknowledgement

On behalf of Project KOMPLEXEPI we thank for the usage of ELKH Cloud (https://science-cloud.hu/) that significantly helped us achieving the results published in this paper.

References

1. Karlinsky A, Kobak D. Tracking excess mortality across countries during the COVID-19 pandemic with the World Mortality Dataset. eLife [Internet]. 2021 [cited 2022 Jun 28];10:e69336. Available from: https://elifesciences.org/articles/69336
2. Santos-Burgoa C, Sandberg J, Suárez E, Goldman-Hawes A, Zeger S, Garcia-Meza A, et al. Differential and persistent risk of excess mortality from Hurricane Maria in Puerto Rico: A time-series analysis. The Lancet Planetary Health [Internet]. 2018 [cited 2022 Jul 13];2:e478–88. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2542519618302092
3. Rivera R, Rolke W. Modeling excess deaths after a natural disaster with application to Hurricane Maria. Statistics in Medicine [Internet]. 2019 [cited 2022 Jul 13];38:4545–54. Available from: https://onlinelibrary.wiley.com/doi/10.1002/sim.8314
4. Morita T, Nomura S, Tsubokura M, Leppold C, Gilmour S, Ochi S, et al. Excess mortality due to indirect health effects of the 2011 triple disaster in Fukushima, Japan: A retrospective observational study. J Epidemiol Community Health [Internet]. 2017 [cited 2022 Jul 13];71:974–80. Available from: https://jech.bmj.com/lookup/doi/10.1136/jech-2016-208652
5. Simonsen L, Clarke MJ, Williamson GD, Stroup DF, Arden NH, Schonberger LB. The impact of influenza epidemics on mortality: Introducing a severity index. Am J Public Health [Internet]. 1997 [cited 2022 Jul 13];87:1944–50. Available from: https://ajph.aphapublications.org/doi/full/10.2105/AJPH.87.12.1944
6. Rosano A, Bella A, Gesualdo F, Acampora A, Pezzotti P, Marchetti S, et al. Investigating the impact of influenza on excess mortality in all ages in Italy during recent seasons (2013/14–2016/17 seasons). International Journal of Infectious Diseases [Internet]. 2019 [cited 2022 Jul 13];88:127–34. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1201971219303285
7. Leon DA, Shkolnikov VM, Smeeth L, Magnus P, Pechholdová M, Jarvis CI. COVID-19: A need for real-time monitoring of weekly excess deaths. The Lancet [Internet]. 2020 [cited 2022 Jul 13];395:e81. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0140673620309338
8. Pearce N, Lawlor DA, Brickley EB. Comparisons between countries are essential for the control of COVID-19. International Journal of Epidemiology [Internet]. 2020 [cited 2022 Jul 13];49:1059–62. Available from: https://academic.oup.com/ije/article/49/4/1059/5864919
9. Beaney T, Clarke JM, Jain V, Golestaneh AK, Lyons G, Salman D, et al. Excess mortality: The gold standard in measuring the impact of COVID-19 worldwide? J R Soc Med [Internet]. 2020 [cited 2022 Jul

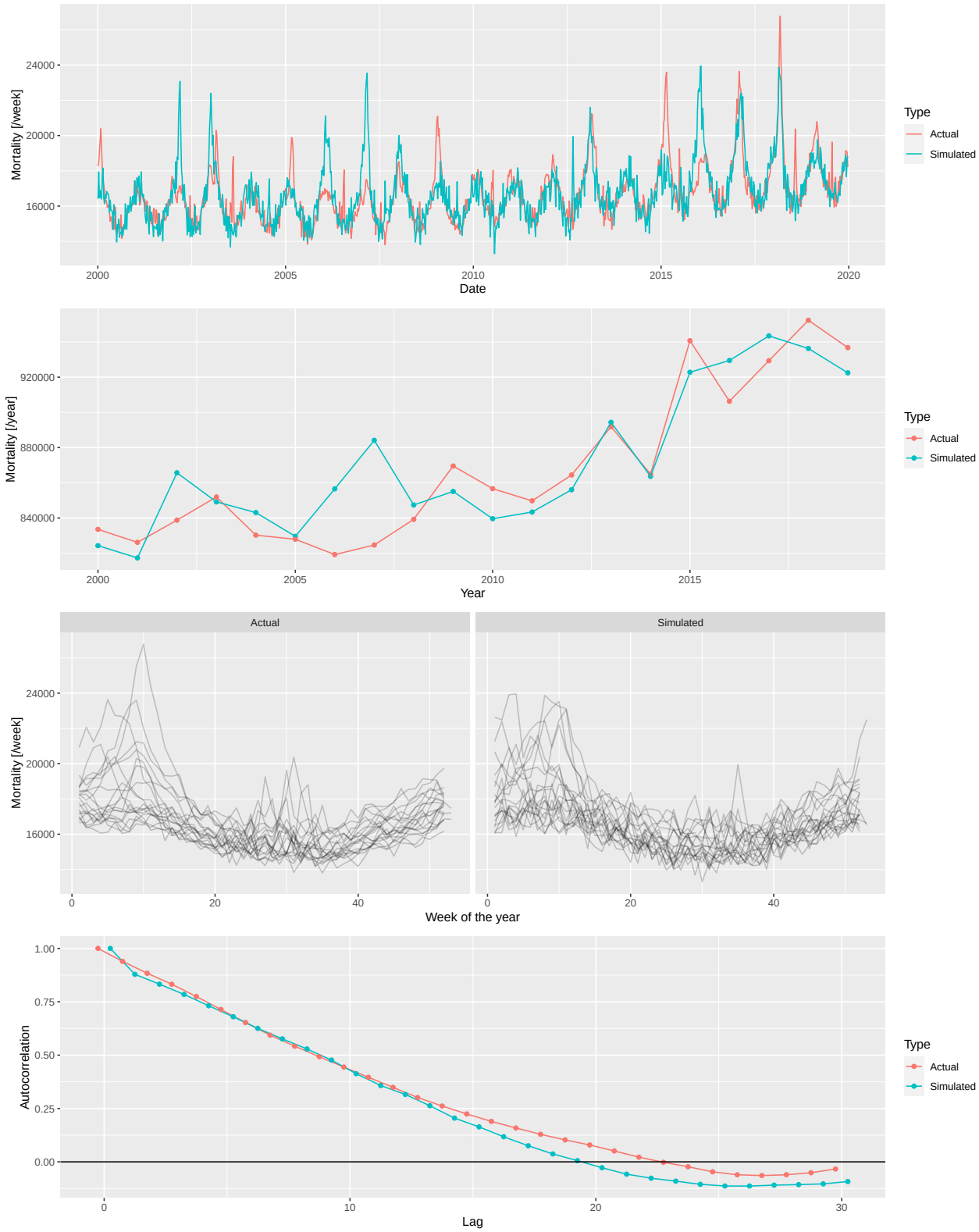

Figure 16: From top to bottom: weekly mortalities, yearly mortalities, seasonal pattern and autocorrelation function of the actual German mortality data and a single simulated dataset, 2000-2019.

13];113:329–34. Available from: http://journals.sagepub.com/doi/10.1177/0141076820956802
10. Faust JS, Krumholz HM, Du C, Mayes KD, Lin Z, Gilman C, et al. All-Cause Excess Mortality and COVID-19–Related Mortality Among US Adults Aged 25-44 Years, March-July 2020. JAMA [Internet]. 2021 [cited 2022 Jul 14];325:785. Available from: https://jamanetwork.com/journals/jama/fullarticle/2774445
11. Kirpich A, Shishkin A, Weppelmann TA, Tchernov AP, Skums P, Gankin Y. Excess mortality in Belarus during the COVID-19 pandemic as the case study of a country with limited non-pharmaceutical interventions and limited reporting. Sci Rep [Internet]. 2022 [cited 2022 Jul 14];12:5475. Available from: https://www.nature.com/articles/s41598-022-09345-z
12. Faust JS, Du C, Liang C, Mayes KD, Renton B, Panthagani K, et al. Excess Mortality in Massachusetts During the Delta and Omicron Waves of COVID-19. JAMA [Internet]. 2022 [cited 2022 Jul 14];328:74. Available from: https://jamanetwork.com/journals/jama/fullarticle/2792738
13. Rossen LM, Ahmad FB, Anderson RN, Branum AM, Du C, Krumholz HM, et al. Disparities in Excess Mortality Associated with COVID-19 — United States, 2020. MMWR Morb Mortal Wkly Rep [Internet]. 2021 [cited 2022 Jul 14];70:1114–9. Available from: http://www.cdc.gov/mmwr/volumes/70/wr/mm7033a2.htm?s_cid=mm7033a2_w
14. Bradshaw D, Dorrington RE, Laubscher R, Moultrie TA, Groenewald P. Tracking mortality in near to real time provides essential information about the impact of the COVID-19 pandemic in South Africa in 2020. S Afr Med J [Internet]. 2021 [cited 2022 Jul 14];111:732. Available from: http://www.samj.org.za/index.php/samj/article/view/13304
15. Modi C, Böhm V, Ferraro S, Stein G, Seljak U. Estimating COVID-19 mortality in Italy early in the COVID-19 pandemic. Nat Commun [Internet]. 2021 [cited 2022 Jul 14];12:2729. Available from: http://www.nature.com/articles/s41467-021-22944-0
16. Wang H, Paulson KR, Pease SA, Watson S, Comfort H, Zheng P, et al. Estimating excess mortality due to the COVID-19 pandemic: A systematic analysis of COVID-19-related mortality, 2020–21. The Lancet [Internet]. 2022 [cited 2022 Jul 14];399:1513–36. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0140673621027963
17. Kontis V, Bennett JE, Rashid T, Parks RM, Pearson-Stuttard J, Guillot M, et al. Magnitude, demographics and dynamics of the effect of the first wave of the COVID-19 pandemic on all-cause mortality in 21 industrialized countries. Nat Med [Internet]. 2020 [cited 2022 Jul 14];26:1919–28. Available from: https://www.nature.com/articles/s41591-020-1112-0
18. Stang A, Standl F, Kowall B, Brune B, Böttcher J, Brinkmann M, et al. Excess mortality due to COVID-19 in Germany. Journal of Infection [Internet]. 2020 [cited 2022 Jul 16];81:797–801. Available from: https://linkinghub.elsevier.com/retrieve/pii/S016344532030596X
19. Konstantinoudis G, Cameletti M, Gómez-Rubio V, Gómez IL, Pirani M, Baio G, et al. Regional excess mortality during the 2020 COVID-19 pandemic in five European countries. Nat Commun [Internet]. 2022 [cited 2022 Jul 14];13:482. Available from: https://www.nature.com/articles/s41467-022-28157-3
20. Davies B, Parkes BL, Bennett J, Fecht D, Blangiardo M, Ezzati M, et al. Community factors and excess mortality in first wave of the COVID-19 pandemic in England. Nat Commun [Internet]. 2021 [cited 2022 Jul 14];12:3755. Available from: http://www.nature.com/articles/s41467-021-23935-x
21. Joy M, Hobbs FR, Bernal JL, Sherlock J, Amirthalingam G, McGagh D, et al. Excess mortality in the first COVID pandemic peak: Cross-sectional analyses of the impact of age, sex, ethnicity, household size, and long-term conditions in people of known SARS-CoV-2 status in England. Br J Gen Pract [Internet]. 2020 [cited 2022 Jul 14];70:e890–8. Available from: https://bjgp.org/lookup/doi/10.3399/bjgp20X713393
22. Alicandro G, Remuzzi G, La Vecchia C. Italy's first wave of the COVID-19 pandemic has ended: No excess mortality in May, 2020. The Lancet [Internet]. 2020 [cited 2022 Jul 13];396:e27–8. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0140673620318651
23. Haklai Z, Aburbeh M, Goldberger N, Gordon E-S. Excess mortality during the COVID-19 pandemic in Israel, March–November 2020: When, where, and for whom? Isr J Health Policy Res [Internet]. 2021 [cited 2022 Jul 14];10:17. Available from: https://ijhpr.biomedcentral.com/articles/10.1186/s13584-021-00450-4
24. Modig K, Ahlbom A, Ebeling M. Excess mortality from COVID-19: Weekly excess death rates by age and sex for Sweden and its most affected region. European Journal of Public Health [Internet]. 2021 [cited 2022 Jul 14];31:17–22. Available from: https://academic.oup.com/eurpub/article/31/1/17/5968985
25. Krieger N, Chen JT, Waterman PD. Excess mortality in men and women in Massachusetts during the COVID-19 pandemic. The Lancet [Internet]. 2020 [cited 2022 Jul 14];395:1829. Available from:

https://linkinghub.elsevier.com/retrieve/pii/S0140673620312344
26. Michelozzi P, de'Donato F, Scortichini M, Pezzotti P, Stafoggia M, De Sario M, et al. Temporal dynamics in total excess mortality and COVID-19 deaths in Italian cities. BMC Public Health [Internet]. 2020 [cited 2022 Jul 14];20:1238. Available from: https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-020-09335-8
27. Vieira A, Peixoto VR, Aguiar P, Abrantes A. Rapid Estimation of Excess Mortality during the COVID-19 Pandemic in Portugal -Beyond Reported Deaths: JEGH [Internet]. 2020 [cited 2022 Jul 14];10:209. Available from: https://www.atlantis-press.com/article/125941684
28. Statistics | Eurostat [Internet]. [cited 2022 Jul 14]. Available from: https://ec.europa.eu/eurostat/databrowser/view/demo_mexrt/default/table?lang=en
29. Ghafari M, Kadivar A, Katzourakis A. Excess deaths associated with the Iranian COVID-19 epidemic: A province-level analysis. International Journal of Infectious Diseases [Internet]. 2021 [cited 2022 Jul 13];107:101–15. Available from: https://linkinghub.elsevier.com/retrieve/pii/S120197122100326X
30. Kobak D. Excess mortality reveals Covid's true toll in Russia. Significance [Internet]. 2021 [cited 2022 Jul 13];18:16–9. Available from: https://onlinelibrary.wiley.com/doi/10.1111/1740-9713.01486
31. Scortichini M, Schneider dos Santos R, De' Donato F, De Sario M, Michelozzi P, Davoli M, et al. Excess mortality during the COVID-19 outbreak in Italy: A two-stage interrupted time-series analysis. International Journal of Epidemiology [Internet]. 2021 [cited 2022 Jul 14];49:1909–17. Available from: https://academic.oup.com/ije/article/49/6/1909/5923437
32. Wood SN. Generalized additive models: An introduction with R. Second edition. Boca Raton: CRC Press/Taylor & Francis Group; 2017.
33. Jaroslaw H, David R, Matt W. Semiparametric regression with R. New York, NY: Springer Science+Business Media; 2018.
34. Acosta RJ, Irizarry RA. A Flexible Statistical Framework for Estimating Excess Mortality. Epidemiology [Internet]. 2022 [cited 2022 Jul 13];33:346–53. Available from: https://journals.lww.com/10.1097/EDE.0000000000001445
35. Islam N, Shkolnikov VM, Acosta RJ, Klimkin I, Kawachi I, Irizarry RA, et al. Excess deaths associated with covid-19 pandemic in 2020: Age and sex disaggregated time series analysis in 29 high income countries. BMJ [Internet]. 2021 [cited 2022 Jul 13];n1137. Available from: https://www.bmj.com/lookup/doi/10.1136/bmj.n1137
36. Rivera R, Rosenbaum JE, Quispe W. Excess mortality in the United States during the first three months of the COVID-19 pandemic. Epidemiol Infect [Internet]. 2020 [cited 2022 Jul 14];148:e264. Available from: https://www.cambridge.org/core/product/identifier/S0950268820002617/type/journal_article
37. Knutson V, Aleshin-Guendel S, Karlinsky A, Msemburi W, Wakefield J. Estimating Global and Country-Specific Excess Mortality During the COVID-19 Pandemic. arXiv; 2022 [cited 2022 Jul 14]; Available from: https://arxiv.org/abs/2205.09081
38. Adam D. 15 million people have died in the pandemic, WHO says. Nature [Internet]. 2022 [cited 2022 Jul 15];605:206–6. Available from: https://www.nature.com/articles/d41586-022-01245-6
39. Van Noorden R. COVID death tolls: Scientists acknowledge errors in WHO estimates. Nature [Internet]. 2022 [cited 2022 Jul 15];606:242–4. Available from: https://www.nature.com/articles/d41586-022-01526-0
40. Eurostat - Data Explorer [Internet]. [cited 2022 Jun 28]. Available from: https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=demo_r_mwk_ts&lang=en
41. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2022. Available from: https://www.R-project.org/
42. Dowle M, Srinivasan A. Data.table: Extension of 'data.frame' [Internet]. 2021. Available from: https://CRAN.R-project.org/package=data.table
43. Wickham H. Ggplot2: Elegant Graphics for Data Analysis [Internet]. Springer-Verlag New York; 2016. Available from: https://ggplot2.tidyverse.org
44. International Organization for Standardization. ISO 8601. 2019.
45. Nepomuceno MR, Klimkin I, Jdanov DA, Alustiza-Galarza A, Shkolnikov VM. Sensitivity Analysis of Excess Mortality due to the COVID-19 Pandemic. Population & Development Rev [Internet]. 2022 [cited 2022 Jul 17];48:279–302. Available from: https://onlinelibrary.wiley.com/doi/10.1111/padr.12475
46. Schöley J. Robustness and bias of European excess death estimates in 2020 under varying model specifications [Internet]. Epidemiology; 2021 Jun. Available from: http://medrxiv.org/lookup/doi/10.1101/2

021.06.04.21258353
47. Jdanov DA, Galarza AA, Shkolnikov VM, Jasilionis D, Németh L, Leon DA, et al. The short-term mortality fluctuation data series, monitoring mortality shocks across time and space. Sci Data [Internet]. 2021 [cited 2022 Jun 28];8:235. Available from: https://www.nature.com/articles/s41597-021-01019-1
48. Hilbe JM. Negative Binomial Regression [Internet]. 2nd ed. Cambridge University Press; 2011 [cited 2022 Jul 15]. Available from: https://www.cambridge.org/core/product/identifier/9780511973420/type/book
49. Palshikar G. Simple algorithms for peak detection in time-series. Proc 1st Int Conf Advanced Data Analysis, Business Analytics and Intelligence. 2009.
50. Nelder JA, Mead R. A Simplex Method for Function Minimization. The Computer Journal [Internet]. 1965 [cited 2022 Jul 15];7:308–13. Available from: https://academic.oup.com/comjnl/article-lookup/doi/10.1093/comjnl/7.4.308
51. Matloff NS. The art of R programming: Tour of statistical software design. San Francisco: No Starch Press; 2011.