3B Forecasting Methods

19 May, 2023

Contents

Overview

The Triple Billion Indices are calculated as a function of 46 outcome indicators. To forecast the Triple Billions to 2030 each tracer indicator must be forecast to 2030. For the few indicators where data is available the forecast captures the impact of the COVID-19 pandemic.

Forecasts are generally produced in a two step process:

1. A "baseline" model is fit using 2000-2019 data to capture pre-pandemic trends and project to 2030 what might have occurred without the COVID-19 pandemic.
2. The baseline projections are then adjusted during pandemic years to account for disruption due to the COVID-19 pandemic.

Baseline Projections

For each of the tracer indicators one of the following modeling strategies is used to produce baseline projections to 2030. Table 2 lists the modeling strategy used for each indicator.

1. **Constant**: The latest data point is held constant until 2030 and all draws are equal to this constant value.

2. **AutoRegressive Integrated Moving Average (ARIMA)**: ARIMA time series models are fit for each indicator-location to project to 2030.

AutoRegressive Integrated Moving Average (ARIMA)

Without shocks like the COVID-19 Pandemic, each tracer indicator in each country should roughly follow previous time trends. One of the most common time series forecasting methods is the AutoRegressive Integrated Moving Average (ARIMA) model. ARIMA models can be used to forecast each indicator for each country to 2030.

See Chapter 9 of 'Forecasting: Principles and Practice' for a brief description of ARIMA time series models. A basic subtype of ARIMA models is the random walk model with drift, also known as ARIMA$(0, 1, 0)$. This can be written as

$$y_t = c + y_{t-1} + \epsilon_t$$

where y_t is the sum of the previous data point y_{t-1}, the average trend or drift c, and random noise e_t.

A non-seasonal ARIMA model can generally be defined by three parameters p, d, and q where

$$p = \text{the autoregressive order}$$
$$d = \text{the order of differencing}$$
$$q = \text{the moving average order}$$

The non-seasonal ARIMA model can then be written as

$$y'_t = c + \phi_1 y'_{t-1} + ... + \phi_p y'_{t-p} + \theta_1 \epsilon_{t-1} + ... + \theta_q \epsilon_{t-q} \epsilon_t$$

where y'_t is the series of values differenced by order d. Differencing a time series is simply calculating the difference between consecutive observations.

Another common method for time series modeling are exponential smoothing models (ETS). ETS(., ., .,) models can be classified by the error (E), trend (T) and seasonal (S) components. For this work we only consider additive error and non-seasonal ETS models which can be also be written as ARIMA models. Chapter 8 of 'Forecasting: Principles and Practice' again provides descriptions of ETS models and Table 9.4 from the textbook is reproduced below to show equivalency between additive errors & non-seasonal ETS models and ARIMA models.

Table 1: 'Equivalence relationships between ETS and ARIMA models'. From Table 9.4 of 'Forecasting: Principles and Practice'.

ETS model	ARIMA model	Description
ETS(A,N,N)	ARIMA(0,1,1)	Simple exponential smoothing
ETS(A,A,N)	ARIMA(0,2,2)	Holt's linear trend method
ETS(A,Ad,N)	ARIMA(1,1,2)	Damped trend method

The `fable` R package provides functions to fit ARIMA & ETS models, and do model selection and validation for each model class. `fable` uses the Hyndman-Khandakar algorithm to automatically iterate through the space of possible ARIMA models and uses corrected Akaike's Information Criterion (AICc) to select the best $\text{ARIMA}(p, d, q)$. For ETS models the `fable` package also uses AICc to select the best ETS model. This procedure as implemented in the `fable` package is described in more detail at the following links for ARIMA and ETS models.

Most tracer indicators are bounded between a minimum and maximum value. For indicators measured as percentages the logit transformation is used to constrain values between 0 and 100. For indicators that are positively constrained the scaled logit transformation is used to constrain values between 0 and twice the maximum observed value in the time series across all countries.

The general algorithm used to produce a baseline time series forecast for each indicator-location is:

- For most indicators only data points in 2019 and before are used to fit the ARIMA models. See table 3 for a list of indicators where all available data is used.
- Fit the random walk model with drift `ARIMA(0,1,0)`.
- In specific cases, the forecasts from the random walk model with drift are inconsistent with prior expectations about future trends. For all locations `fable` is also used to fit and select a best:
    1. non-seasonal ARIMA model.
    2. additive & non-seasonal ETS model.
- By default the random walk model with drift is used to make final baseline projections to 2030. Visual review is used to select for which indicator-locations the automatically selected model from `fable` should be used instead. In total 96.5% of indicator-locations did not use the default random walk model with drift `ARIMA(0,1,0)` or the default constant model. Table 2 shows this percentage separately for each indicator.
- 1000 draws of the baseline forecast are then sampled from the selected model to feed into the next covid adjustment step.

COVID-19 Pandemic Adjustment

A subset of 3B tracer indicators account for the COVID-19 Pandemic in input estimates for the year 2020 and 2021.

There are three different scenarios for whether or not a tracer indicator has been adjusted for the impacts of the COVID-19 Pandemic.

1. The tracer indicator does not include any estimates after 2019. In this case we do not attempt to predict the impact of the COVID-19 Pandemic in forecasts. Table 3 column '2020-2021' shows the percentage of location-years with data.

2. The tracer indicator accounts for the COVID-19 Pandemic in a very small subset of locations. For example the anc4 tracer indicator included data points for 2.8% of location-years in 2020 and 2021. In these cases we adjust forecasts to exactly match input estimates in 2020 and 2021 where they are available but do not attempt to predict the impact of the COVID-19 Pandemic in other locations.

3. The tracer indicator accounts for the COVID-19 Pandemic in a substantial number of locations that allows us to predict potential impacts of the COVID-19 Pandemic in other locations where input estimates are not included. Table 3 column 'COVID Adjustment' shows for which indicators this adjustment is applied

This adjustment predicts the difference between the baseline forecast and input estimates given the estimated per capita number of covid infections.

$$\delta_{l,t} = \hat{y}_{c,t} - y_{c,t}$$

where :

$\hat{y}_{c,t}$ =baseline prediction of indicator

in transformed space in country c and time t

$y_{c,t}$ =input estimate of indicator

in transformed space in country c and time t

$\delta_{l,t}$ =residual difference between baseline prediction and input estimate

in transformed space in country c and time t

$$\delta'_{l,t} = \alpha_0 + \beta X_{c,t} + \epsilon_{c,t}$$
$$\epsilon_{c,t} \sim Normal(0, \sigma)$$

In scenario 3 the predicted difference from the baseline forecast is used for location-years where data does not exist for 2020 or 2021. Where input data does exist for 2020 or 2021 the forecasts are adjusted to exactly match the input data point. Indicator specific assumptions are then made about when indicator forecasts will return to baseline forecast trends. The 'Return to Baseline Year' column in Table 3 shows the year that each covid adjusted indicator is assumed to return to baseline forecast trends after 2021 data points or predicted 2021 covid adjusted estimates. In cases where it is assumed that the indicator will return to baseline levels in 2024, 2022 and 2023 are filled in using linear interpolation between 2021 and 2024 estimates.

References

Hyndman, R.J., & Athanasopoulos, G. (2021) Forecasting: principles and practice, 3rd edition, OTexts: Melbourne, Australia. OTexts.com/fpp3.

O'Hara-Wild M, Hyndman R, Wang E (2023). *fable: Forecasting Models for Tidy Time Series.* R package version 0.3.3, https://CRAN.R-project.org/package=fable.

Tables

Table 2: The percentage of countries for each indicator that use the specified baseline modeling strategy.

ind	Constant	ARIMA(0,1,0)	fable ARIMA search	fable ETS search	Manual ARIMA specification
stunting	0%	100%	0%	0%	0%
wasting	100%	0%	0%	0%	0%
overweight	0%	100%	0%	0%	0%
devontrack	100%	0%	0%	0%	0%
child_obese	0%	98%	1%	1%	1%
ipv	100%	0%	0%	0%	0%
child_viol	100%	0%	0%	0%	0%
suicide	0%	97%	2%	1%	1%
road	1%	99%	0%	0%	0%
water	0%	92%	7%	1%	0%
water_urban	0%	100%	0%	0%	0%
water_rural	0%	100%	0%	0%	0%
hpop_sanitation	0%	99%	1%	0%	0%
hpop_sanitation_urban	0%	100%	0%	0%	0%
hpop_sanitation_rural	0%	99%	0%	0%	1%
fuel	0%	92%	6%	2%	0%
alcohol	0%	90%	1%	8%	1%
pm25	0%	100%	0%	0%	0%
hpop_tobacco	0%	100%	0%	0%	0%
adult_obese	0%	100%	0%	0%	0%
transfats	100%	0%	0%	0%	0%
fp	0%	100%	0%	0%	0%
anc4	0%	97%	2%	0%	1%
dtp3	0%	67%	29%	1%	3%
pneumo	0%	100%	0%	0%	0%
tb	0%	98%	1%	0%	1%
art	0%	97%	0%	3%	0%
itn	0%	100%	0%	0%	0%
uhc_sanitation	0%	98%	1%	0%	1%
bp	0%	0%	0%	100%	0%
fpg	0%	100%	0%	0%	0%
uhc_tobacco	0%	100%	0%	0%	0%
beds	0%	100%	0%	0%	0%
doctors	0%	100%	0%	0%	0%
nurses	0%	100%	0%	0%	0%
espar	14%	79%	6%	0%	0%
polio_routine	0%	80%	16%	0%	3%
measles_routine	0%	79%	18%	0%	3%
yellow_fever_routine	100%	0%	0%	0%	0%
meningitis_routine	100%	0%	0%	0%	0%
detect	0%	93%	5%	2%	0%
notify	1%	95%	1%	3%	0%
respond	0%	97%	1%	2%	0%

Table 2: The percentage of countries for each indicator that use the specified baseline modeling strategy. *(continued)*

ind	Constant	ARIMA(0,1,0)	fable ARIMA search	fable ETS search	Manual ARIMA specification
detect_respond	0%	96%	2%	3%	0%

Table 3: '2000-2019': Percent of location years between with data. '2020-2021': Percent of location years between with data. 'Baseline Data Year Cutoff': The maximum year of data used in the baseline forecasting model. 'None' if no cutoff is applied and all data is used. 'COVID Adjustment': Whether the covid adjustment model is applied. 'Return to Baseline Year': The year covid adjusted estimates return to match baseline forecast trends.

ind	2000-2019	2020-2021	% Missing Locations	Baseline Data Year Cutoff	COVID Adjustment	Return to Baseline Year
stunting	80.9%	80.9%	19%	2019	FALSE	N/A
wasting	23.6%	9.8%	24%	None	FALSE	N/A
overweight	81.4%	81.4%	19%	2019	FALSE	N/A
devontrack	1.5%	0%	69%	None	FALSE	N/A
child_obese	88.1%	0%	2%	2019	FALSE	N/A
ipv	3.9%	0%	22%	None	FALSE	N/A
child_viol	2.2%	0%	57%	None	FALSE	N/A
suicide	94.3%	0%	6%	2019	FALSE	N/A
road	94.3%	0%	6%	2019	FALSE	N/A
water	58.8%	29.4%	41%	2019	FALSE	N/A
water_urban	6.5%	3.4%	93%	2019	FALSE	N/A
water_rural	1%	0.5%	99%	2019	FALSE	N/A
hpop_sanitation	60.8%	29.4%	39%	2019	FALSE	N/A
hpop_sanitation_urban	2.1%	1%	98%	2019	FALSE	N/A
hpop_sanitation_rural	3.6%	1.8%	96%	2019	FALSE	N/A
fuel	98.5%	98.5%	2%	2019	FALSE	N/A
alcohol	96.5%	0%	3%	2019	FALSE	N/A
pm25	49.5%	0%	1%	2019	FALSE	N/A
hpop_tobacco	79.5%	1.3%	15%	2019	FALSE	N/A
adult_obese	88.1%	0%	2%	2019	FALSE	N/A
transfats	67%	67%	33%	None	FALSE	N/A
fp	99.5%	94.8%	0%	2019	FALSE	N/A
anc4	91.5%	2.8%	0%	2019	FALSE	N/A
dtp3	100%	100%	0%	2019	TRUE	2024
pneumo	91.1%	2.3%	0%	2019	FALSE	N/A
tb	99.7%	97.7%	0%	2019	TRUE	2022
art	98.9%	88.7%	0%	2019	FALSE	N/A
itn	20.6%	20.6%	79%	2019	FALSE	N/A
uhc_sanitation	99.4%	45.4%	0%	2019	FALSE	N/A
bp	99.9%	0%	0%	2019	FALSE	N/A
fpg	90%	0%	0%	2019	FALSE	N/A
uhc_tobacco	98.5%	42.3%	0%	2019	FALSE	N/A

Table 3: '2000-2019': Percent of location years between with data. '2020-2021': Percent of location years between with data. 'Baseline Data Year Cutoff': The maximum year of data used in the baseline forecasting model. 'None' if no cutoff is applied and all data is used. 'COVID Adjustment': Whether the covid adjustment model is applied. 'Return to Baseline Year': The year covid adjusted estimates return to match baseline forecast trends. *(continued)*

ind	2000-2019	2020-2021	% Missing Locations	Baseline Data Year Cutoff	COVID Adjustment	Return to Baseline Year
beds	96.5%	31.2%	0%	2019	FALSE	N/A
doctors	95.9%	35.6%	0%	2019	FALSE	N/A
nurses	96.3%	32%	0%	2019	FALSE	N/A
espar	9.2%	90.5%	1%	None	FALSE	N/A
polio_routine	100%	100%	0%	2019	TRUE	2024
measles_routine	100%	100%	0%	2019	TRUE	2024
yellow_fever_routine	18.6%	18.6%	81%	None	FALSE	N/A
meningitis_routine	18.1%	7.5%	81%	None	FALSE	N/A
detect	16%	92.3%	7%	None	FALSE	N/A
notify	16%	92.3%	7%	None	FALSE	N/A
respond	16%	92.3%	7%	None	FALSE	N/A
detect_respond	16%	92.3%	7%	None	FALSE	N/A

Table 4: 3B tracer indicator descriptions

ind	description	unit_raw	hpop	uhc	hep
stunting	Prevalence of stunting in children under 5 (%)	%	1	0	0
wasting	Prevalence of wasting in children under 5 (%)	%	1	0	0
overweight	Prevalence of overweight in children under 5 (%)	%	1	0	0
devontrack	Proportion of children under 5 developmentally on track (health, learning and psychosocial well-being) (%)	%	1	0	0
child_obese	Prevalence of obesity among children and adolescents (aged 5-19) (%)	%	1	0	0
ipv	Proportion of women (15-49) subjected to violence by current or former intimate partner (%)	%	1	0	0
child_viol	Proportion of children (aged 1-17) experiencing physical or psychological aggression (%)	%	1	0	0
suicide	Suicide mortality rate (per 100 000 population)	%	1	0	0
road	Road traffic mortality rate (per 100 000 population)	per 100 000 population	1	0	0
water	Proportion of population using safely managed drinking water services (%)	per 100 000 population	1	0	0

Table 4: 3B tracer indicator descriptions *(continued)*

ind	description	unit_raw	hpop	uhc	hep
water_urban	Proportion of urban population using safely managed drinking water services (%)	%	1	0	0
water_rural	Proportion of urban population using safely managed drinking water services (%)	%	1	0	0
hpop_sanitation	Proportion of population using safely managed sanitation services (%)	%	1	0	0
hpop_sanitation_urban	Proportion of urban population using safely managed sanitation services (%)	%	1	0	0
hpop_sanitation_rural	Proportion of rural population using safely managed sanitation services (%)	%	1	0	0
fuel	Proportion of population with primary reliance on clean fuels (%)	%	1	0	0
alcohol	Total alcohol per capita consumption in adults aged 15+ (litres of pure alcohol)	litres of pure alcohol	1	0	0
pm25	Annual mean concentrations of fine particulate matter (PM2.5) in urban areas (Âµg/m3)	µg/m3	1	0	0
hpop_tobacco	Prevalence of tobacco use in adults aged 15+ (age-standardized) (%)	%	1	0	0
adult_obese	Prevalence of obesity among adults (aged 18+) (%)	%	1	0	0
transfats	Best practice policy implemented for industrially produced trans fatty acids (Y/N)	Yes/No	1	0	0
fp	Demand satisfied with modern methods (married women or in-union) 1 (%)	%	0	1	0
anc4	Antenatal care coverage (+4 visits) 2 (%)	%	0	1	0
dtp3	DPT3 Immunization coverage (%)	%	0	1	0
pneumo	Care seeking for suspected pneumonia (%)	%	0	1	0
tb	TB treatment coverage	%	0	1	0
art	HIV ART coverage 4	%	0	1	0
itn	ITN use 5	%	0	1	0
uhc_sanitation	Use of basic sanitation	%	0	1	0
bp	Hypertension treatment coverage	%	0	1	0
fpg	Raised fasting blood glucose	%	0	1	0
uhc_tobacco	Tobacco use prevalence* 3	%	0	1	0
beds	Hospital beds density*	per 10 000 population	0	1	0
doctors	Density of doctors	per 10 000 population	0	1	0
nurses	Density of nurses/midwives	per 10 000 population	0	1	0

Table 4: 3B tracer indicator descriptions *(continued)*

ind	description	unit_raw	hpop	uhc	hep
espar	IHR core capacity index*	%	0	1	1
polio_routine	Polio 3 - Routine	%	0	0	1
measles_routine	Measles (MCV1) - Routine	%	0	0	1
yellow_fever_routine	Yellow Fever - Routine	%	0	0	1
meningitis_routine	Meningitis - Routine	%	0	0	1
detect	Time to detect	days	0	0	1
notify	Time to notify	days	0	0	1
respond	Time to respond	days	0	0	1
detect_respond	Time to detect and respond	days	0	0	1