CBS COPENHAGEN BUSINESS SCHOOL

MSc Business Administration and Data Science

Chair of Predictive Analytics (CDSCO1005E)

“Forecasting Tuberculosis Incidence with ETS, ARIMA, and Dynamic Regression Models: Global and High-Burden Country Analyses”

Alessio Desideri - 176184

Candidate

Thomas Einfeldt

Examiner

Herdis Steingrimsdottir

Examiner

Number of characters: 19630

Number of pages: 10

Academic Year 2024/2025

Table of Contents

1. Introduction

Tuberculosis (TB) remains one of the most persistent infectious diseases worldwide, despite decades of medical advances and public health interventions. According to the World Health Organization (2024), about 10.6 million people developed TB in 2023, with the vast majority of cases occurring in low- and middle-income countries. While global incidence has been declining slowly, the rate of progress varies substantially across regions, reflecting differences in socio-economic conditions, healthcare capacity, and the burden of co-morbidities such as HIV.

This project develops along two main analytical perspectives. The first is to assess global TB trends using a range of statistical models. This involves examining aggregated global case data as well as applying the same modelling techniques to a narrower set of countries. Specifically, the ten nations with the highest reported cases over the past decade. The second perspective focuses in detail on a specific country, Lesotho, chosen based on both their TB burden and the presence of unstable economic and geopolitical conditions. This dual approach allows for a richer understanding of TB dynamics, combining a broad comparative framework with targeted country-level insights.

The primary aim of this work is to develop and compare statistical models capable of forecasting TB incidence in the selected countries computed. The analysis combines classical time series methods: Exponential Smoothing (ETS) and Autoregressive Integrated Moving Average (ARIMA) models. The latter one has been also developed as a dynamic regression approach that incorporate relevant exogenous variables, such as GDP per capita and epidemiological indicators. By systematically evaluating model performance and forecast accuracy, the study seeks to identify modelling strategies that can support better resource planning and policy design.

2. Data description

The primary dependent variable is annual TB incidence, measured as the number of new cases per 100,000 population, sourced from the WHO Global Health Observatory. Two sets of analyses were performed. For the global and “Top-10” country aggregates, annual means of TB incidence were computed across all countries and across the ten countries with the highest average incidence over the last decade. For these aggregates, dynamic regression models included two epidemiological regressors from WHO data: TB incidence among HIV-positive individuals and the number of new or relapse TB cases. For the country-specific analysis, the focus was on Lesotho for the period 2000–2022. The exogenous variables in this model were the same epidemiological indicators used in the aggregate analysis, plus GDP in current US dollars sourced from the World Bank. All series were annual and non-seasonal. Aggregations were computed using `na.rm=TRUE` for missing values, and no interpolation was required for the country-level series in the selected time frame. Stationarity issues were addressed via first differencing (d=1) as indicated by the Augmented Dickey–Fuller (ADF) and KPSS tests. Potential structural breaks were explored using the breakpoint methodology in the *strucchange* package.

Pearson/Spearman correlations and cross-correlations (CCFs) between TB incidence and candidate regressors were computed. Global and Top-10 aggregates show a strong contemporaneous association with epidemiological indicators (TB among HIV-positive and new/relapse cases), with CCFs peaking at lag 0. In Lesotho, (log)GDP is moderately negatively correlated with TB in levels but not in first differences, suggesting trend-driven spurious correlation; the CCF peaks at lag +1 (negative), consistent with macroeconomic conditions leading a reduction in TB incidence by about one year. Given the high collinearity between HIV and new/relapse (VIF ≅ 22–23), parsimonious specifications or regularization have been favored when combining these regressors.

3. Methodology

The modelling strategy integrates both univariate and multivariate approaches in order to evaluate the predictive performance of different time series techniques on TB incidence data and to assess the added value of incorporating external drivers.

3.1. Exploratory Analysis and Pre-Processing

The analysis began with visual inspection of the TB incidence series at global, "Top-10" country aggregate, and individual country (Lesotho) levels. This step aimed to detect underlying trends, sudden structural changes, or changes in variance that could influence model selection. For the aggregated datasets, TB incidence was computed as the mean annual value across relevant countries. While for the country-level datasets (2000–2022), TB incidence values were directly sourced from WHO and merged with GDP data from the World Bank.

3.2. Stationarity and Model Identification

Stationarity was assessed using both the Augmented Dickey–Fuller (ADF) and the KPSS tests to account for the possibility of Type I/II errors in a single test. In cases where the series showed evidence of non-stationarity, first differencing (d=1) was applied. Autocorrelation and partial autocorrelation plots (ACF, PACF) were used as specification tools to guide the selection of autoregressive (p) and moving average (q) terms. These graphical diagnostics, in combination with the information criteria (AIC, BIC), provided a foundation for model order selection.

3.3. Univariate Modelling

Two univariate approaches were applied: the first one is an ETS model, automatically selected to fit the level and trend components without seasonality (given the annual frequency). Parameters were estimated via maximum likelihood. On the other hand I used an ARIMA model: implemented via `auto.arima()` with stepwise search disabled and approximation turned off to ensure a more exhaustive model space exploration. Orders p, d, q were determined by minimizing the AIC while ensuring residual adequacy. Both global and "Top-10" series were split into training (80%) and testing (20%) sets. The model accuracy was assessed on the test set using RMSE, MAE, MAPE, and MASE. While the residuals were evaluated for autocorrelation via the Ljung–Box test and visually inspected for randomness and normality.

3.4. Dynamic Regression (ARIMAX)

Dynamic regression models were used to capture the influence of relevant exogenous variables while modelling autocorrelation in the error term. The general form is:

$$y_t = \beta_0 + \sum_{k=1}^{K} \beta_k x_{k,t} + \phi(B)^{-1}\theta(B)\varepsilon_t$$

where y_t is TB incidence, $x_{k,t}$ are the exogenous regressors, *$\phi(B)$* and *$\theta(B)$* are the AR and MA operators, and ε_t is white noise. For the aggregate analysis the features used are HIV-positive individuals and the number of relapse TB case. For the specific investigation on Lesotho, the same epidemiological regressors were used together with GDP in current US dollars. Three forecasting strategies were applied for the exogenous variables. The first consisted of a static hold, where regressors were kept fixed at their last observed values. The second relied on ARIMA-forecasted regressors, with each regressor modelled individually using ARIMA and the resulting forecasts incorporated into the main ARIMAX model. The third approach involved a scenario analysis, where regressor values were manually adjusted by ±5% to simulate potential future changes.

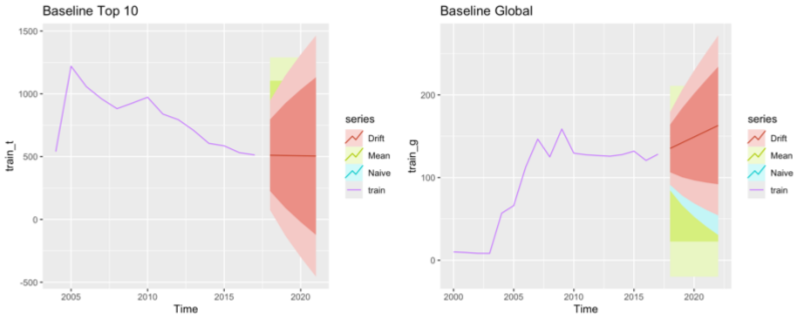
3.5. Model comparison and Forecasting generation

Model comparison was based on both in-sample and out-of-sample criteria: AIC for model (in sample) parsimony, and RMSE, MAE, MAPE, MASE on the hold-out set (out of sample). Forecasts were generated for a 3-year horizon for all models. Where relevant, 80% and 95% prediction intervals were produced to quantify uncertainty. For scenario-based forecasts, results were compared to the baseline static-hold forecasts to illustrate the sensitivity of TB incidence projections to changes in external drivers.

4. Results

4.1. Global and Top-10 Aggregates

The analysis considered two aggregate time series: the global average TB incidence per 100,000 population and the average incidence for the ten highest-burden countries over the last decade. In both cases, statistical testing confirmed non-stationarity in levels: Augmented Dickey–Fuller (ADF) p-values were above 0.49, while KPSS results were significant at the 5% level, indicating the need for differencing prior to modelling. ACF and PACF plots further revealed persistent autocorrelation in the undifferenced series, consistent with the test results. Structural break analysis using the Quandt Likelihood Ratio test identified multiple significant change points. In the global series, relevant break dates emerged in 2003, 2006, 2009, 2015 and 2018, while in the top-10 series breaks were detected in 2005, 2007, 2010, between 2012 and 2016, and in 2018. These dates align with periods of major public health interventions and possible changes in reporting methodology.

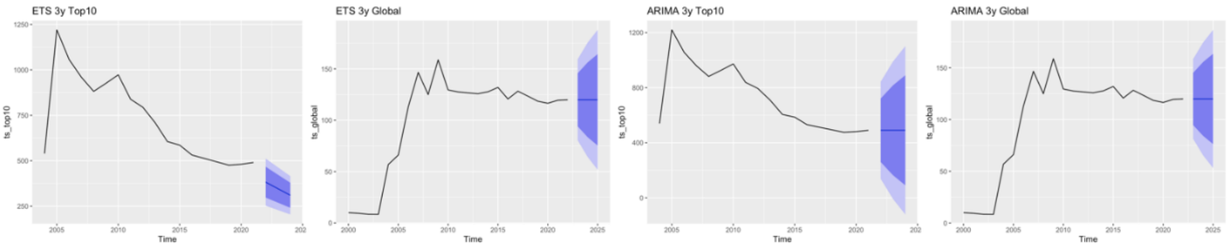


When the models were trained on 80% of the data and evaluated on the remaining 20%, ETS slightly outperformed ARIMA for the global series, whereas ARIMA clearly dominated in the high-burden group. The table below summarizes the out-of-sample accuracy for each model and dataset. Residual diagnostics, including the Ljung–Box test, returned p-values well above 0.1 for all fitted models, indicating no evidence of residual autocorrelation and supporting the adequacy of model specification.

Dataset	Model	RMSE	MAE	MAPE (%)	MASE	AIC

Global	ETS	8.95	8.66	7.28%	0.581	168.75
Global	ARIMA	8.99	8.69	7.31%	0.583	155.79
Top-10	ETS	79.94	64.19	13.22%	0.532	182.76
Top-10	ARIMA	28.57	27.56	5.71%	0.228	177.38

Following model selection, both ETS and ARIMA were refitted on the complete series to generate three-year forecasts. For the global series, the "ETS 3y Global" and "ARIMA 3y Global" plots both show a gradual, steady decline in TB incidence. In contrast, the high-burden group ("ETS 3y Top10" and "ARIMA 3y Top10") displays a steeper projected reduction, albeit with wider confidence intervals, reflecting higher volatility.



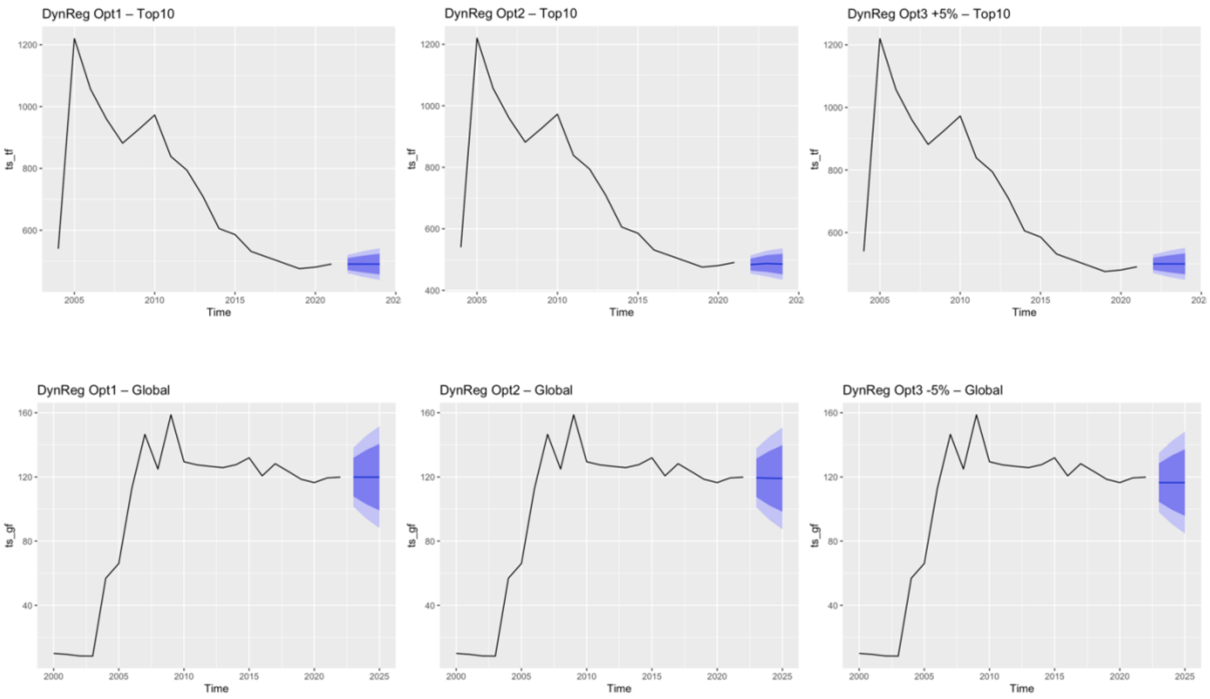
Both ETS and ARIMA models provided satisfactory out-of-sample accuracy, with MAPE values below 10% for the global series and below 6% for the Top-10 series. The minimal performance gap between the two methods indicates that both adequately captured the underlying temporal structure, supporting their use for short-term forecasting.

4.2. Dynamic Regression – Aggregates

To assess whether incorporating epidemiological regressors improves forecast accuracy, dynamic regression models (ARIMAX) were estimated for both the global and top-10 series. The exogenous variables included the mean annual TB incidence among HIV-positive individuals and the mean annual number of new or relapse TB cases. The in-sample fit showed substantial reduction in error metrics relative to the univariate ARIMA/ETS benchmarks, particularly in the global series, where the RMSE on the test set fell to nearly one-third of the ARIMA baseline. In the top-10 series, improvements were more modest but still notable, especially in MAPE.

Dataset	Model	RMSE	MAE	MAPE (%)	MASE	AIC
Global	Dyn. Reg.	3.56	3.24	2.71	0.218	164.72
Top-10	Dyn. Reg.	7.01	6.53	1.35	0.054	144.32

The “DynReg Opt1 – Global” and “DynReg Opt1 – Top10” plots illustrate the three-year forecasts obtained under the assumption that regressors remain constant at their latest observed values. For both series, the models anticipate a continued decline in TB incidence, with narrower prediction intervals compared to the univariate approaches. Residual diagnostics, including the Ljung–Box test, produced p-values well above conventional significance thresholds, and ACF/PACF plots of residuals showed no meaningful autocorrelation, supporting the adequacy of the model specification. Alternative scenarios were also simulated. In “DynReg Opt3 ±5% – Global” and “DynReg Opt3 ±5% – Top10”, the incidence in HIV-positive individuals and the number of new/relapse cases were adjusted by ±5% to explore sensitivity. Positive shocks lead to a mild flattening of the downward trend, whereas negative shocks accelerate the decline, with effects more pronounced in the global series.



Incorporating epidemiological regressors improved predictive accuracy, particularly for the Top-10 series where the MASE dropped below 0.06. The diagnostic evidence confirms that exogenous factors provided additional explanatory power without introducing serial correlation into the residuals.

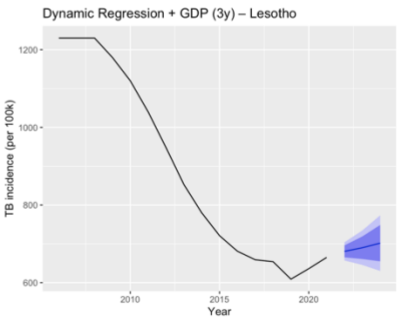
4.3. Country-level Dynamic Regression with GDP

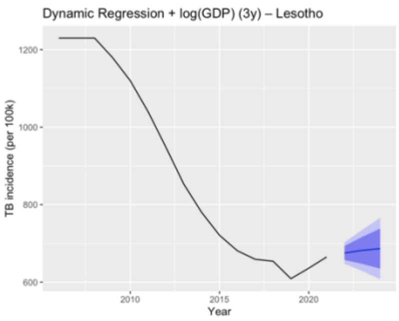
For Lesotho, a country with one of the world’s highest TB incidence rates, the dynamic regression model was extended to include GDP as an additional exogenous regressor alongside HIV-associated

TB incidence and new/relapse cases. The objective was to assess whether macroeconomic conditions provide additional explanatory power in a high-burden, resource-constrained setting. Two model specifications were estimated: using GDP in level form, and using the natural logarithm of GDP to reduce skewness and stabilize variance. In the baseline model with GDP in levels, the RMSE and MAE were high, and the MASE exceeded 1, indicating lower accuracy than a naïve forecast. Applying the log transformation improved out-of-sample performance: RMSE fell from 104.11 to 69.59, MAE from 98.76 to 66.11, and MAPE from 15.41% to 10.32%, while MASE decreased from 1.903 to 1.274. These improvements suggest that the log transformation yielded a more stable relationship between GDP and TB incidence, potentially mitigating the influence of extreme GDP values.

Dataset	Model	RMSE	MAE	MAPE (%)	MASE	AIC
Lesotho - no Log	Dyn. Reg. + GDP	104.11	98.76	15.41	1.903	127.61
Lesotho - Log	Dyn. Reg. + GDP	69.59	66.11	10.32	1.27	94.04

The “Dynamic Regression + GDP (3y) – Lesotho” plot for the level model showed a forecasted gradual decline in TB incidence, with wide prediction intervals reflecting the instability of the historical data. The log-transformed model retained a similar declining trend but with narrower prediction intervals, consistent with the lower forecast errors. Residual diagnostics in both cases (Ljung–Box p-values > 0.1 and absence of significant autocorrelation in ACF/PACF plots) suggested that model specification was adequate.





Overall, while including GDP in level form did not enhance forecast accuracy for Lesotho, the log-transformed GDP specification delivered moderate gains, improving forecast precision, and reducing error metrics. These results highlight the potential benefits of transformation when integrating macroeconomic variables into predictive models for small, high-burden countries, though findings remain sensitive to data quality and should be interpreted with caution.

4.4. Other analysis

In addition to the main forecasting models, several supporting analyses were conducted to better understand the statistical properties of the data and to establish performance benchmarks.

First, autocorrelation and partial autocorrelation plots (ACF/PACF) were examined for both the global and Top-10 series. These plots revealed strong positive autocorrelation at low lags, with a gradual decay, suggesting that differencing would be necessary to achieve stationarity, which is consistent with the results of the ADF and KPSS tests reported in Section 4.1. To further assess data characteristics, a Box–Cox transformation was applied to stabilize variance. Estimated λ values for both series were close to 1, indicating minimal variance instability. Additionally, STL decomposition was performed to explore potential seasonality. As the series are annual, no genuine seasonal pattern was found; the “quarterly” decomposition for the global series was a purely illustrative resampling to visualize cyclical components. The baseline models, Mean, Naïve, and Drift, were also evaluated as simple forecasting benchmarks. For the global series, the Mean forecast achieved the lowest error on the test set (RMSE ≅ 8.99, MAPE ≅ 7.31%), while for the Top-10 series the Drift model performed best (RMSE ≅ 23.61, MAPE ≅ 4.63%). Additional model diagnostics included the Normalized RMSE (NRMSE), computed over the full series, which provided scale-independent error measures broadly consistent with the RMSE rankings above. Ljung–Box tests on residuals for all fitted models (ETS, ARIMA, and dynamic regression variants) did not reject the null hypothesis of no autocorrelation at the 5% level, indicating that the models adequately captured the temporal dependence in the data. These supplementary analyses confirmed that the main modelling choices, differencing for stationarity, the absence of seasonal terms, and the selection of ARIMA, ETS and dynamic regression as primary forecasting tools, were well supported by the statistical evidence.

5. Discussion

The application of ETS and ARIMA models to TB incidence forecasting aligns with previous literature. Wang et al. (2018) modelled monthly TB incidence in China from 2005 to 2017 using SARIMA and a hybrid SARIMA–GRNN model, achieving substantial accuracy improvements with the hybrid approach. Their results confirm the suitability of univariate time-series methods for short-term TB forecasting when seasonality is present and data quality is high. Incorporating epidemiological regressors proved beneficial in the present study, particularly for the Top-10 series where the MASE fell below 0.06. This is consistent with findings by Tsan et al. (2022), who demonstrated that including environmental and epidemiological variables in LSTM and ARIMA models significantly reduced forecast error in predicting influenza-like illness and respiratory diseases. The weaker results for Lesotho when adding GDP reflect the observation that unstable or noisy regressors can degrade model accuracy. While this work focused on interpretable statistical models, evidence from Tsan et al. (2022) and Wang et al. (2018) suggests that hybrid approaches combining statistical and machine-learning methods may capture both linear and non-linear patterns, potentially improving forecasts in highly volatile settings.

6. Conclusion

This study applied ETS, ARIMA, and dynamic regression models to forecast TB incidence at global, high-burden country group, and country-specific levels. Both ETS and ARIMA achieved high short-term accuracy, with MAPE values below 10% for the global series and below 6% for the Top-10 group. Incorporating epidemiological regressors (HIV-associated TB incidence and new/relapse cases) substantially improved predictive performance, particularly for the Top-10. In Lesotho, adding GDP in its original level form did not yield benefits due to data volatility, with error metrics exceeding those of a naïve benchmark. When the GDP variable was log-transformed to stabilize variance, out-of-sample accuracy improved, reducing RMSE, MAE, and MAPE, yet performance remained weaker than for aggregate-level models. These results confirm that statistical time-series models remain effective tools for short-term infectious disease forecasting, especially when combined with relevant epidemiological covariates. However, model performance depends on data quality, stability, and update frequency. Future work could explore hybrid statistical–machine learning approaches, as successfully applied by Tsan et al. (2022) and Wang et al. (2018), to better capture both linear and non-linear dynamics in highly variable settings.

References

Tsan, Y. T., Chen, D. Y., Liu, P. Y., Kristiani, E., Nguyen, K. L. P., & Yang, C. T. (2022). The Prediction of Influenza-like Illness and Respiratory Disease Using LSTM and ARIMA. *International journal of environmental research and public health*, *19*(3), 1858.
Retrieved from https://doi.org/10.3390/ijerph19031858

Wang, H., Tian, C. W., Wang, W. M., & Luo, X. M. (2018). Time-series analysis of tuberculosis from 2005 to 2017 in China. *Epidemiology and infection*, *146*(8), 935–939.
Retrieved from https://doi.org/10.1017/S0950268818001115

Rob J. Hyndman & George Athanasopoulos. (2021). Forecasting: principles and practice, 3rd edition, OTexts: Melbourne, Australia.
Retrieved from https://otexts.com/fpp3

Hyndman, R. J., & Kakar, Y. (2008). Automatic Time Series Forecasting: The forecast Package for R. *Journal of Statistical Software*, *27*(3), 1–22.
Retrieved from https://doi.org/10.18637/jss.v027.i03

World Health Organization. (2024). *Global tuberculosis report 2024*.
Retrieved from https://www.who.int/teams/global-programme-on-tuberculosis-and-lung-health/tb-reports/global-tuberculosis-report-2024

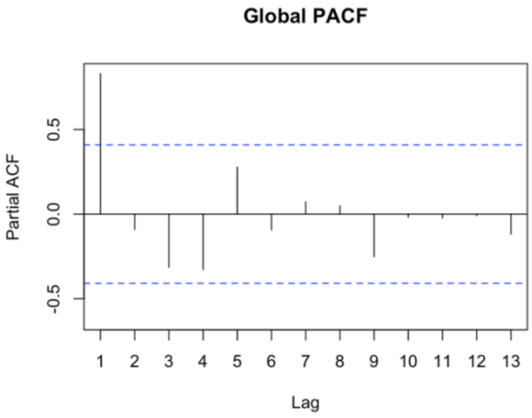
World Health Organization Indicators https://www.who.int/data/gho/data/indicators

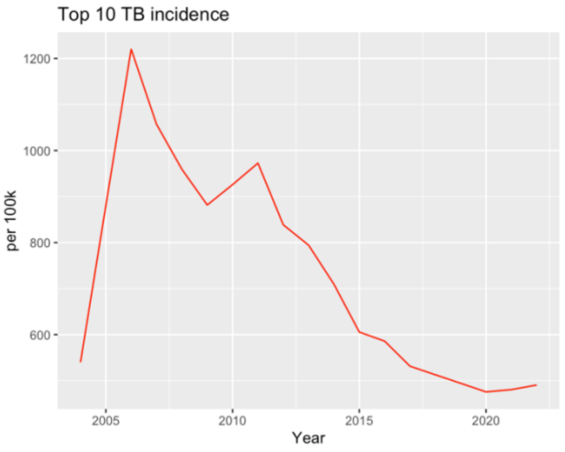
World Bank GDP Indicators https://data.worldbank.org/indicator/NY.GDP.MKTP.CD

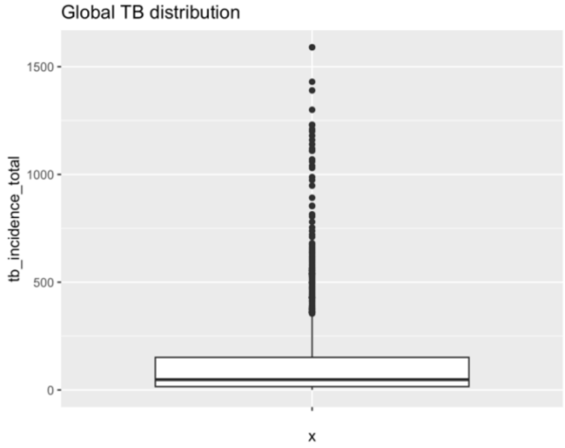
Appendix

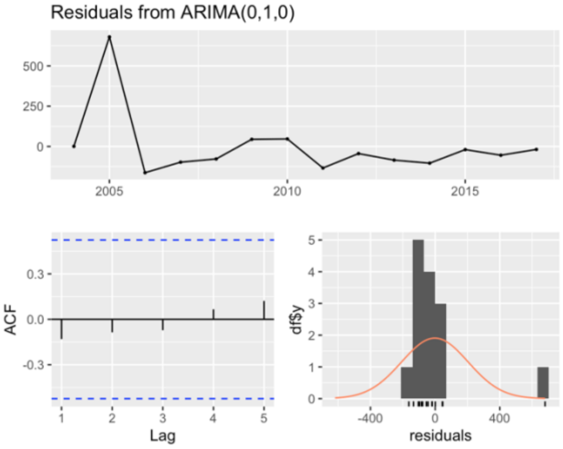
Dataset	Model	RMSE	MAE	MAPE (%)	MASE
Global	Mean	8.99	8.69	7.31	0.58
Global	Naïve	21.18	16.22	29.35	1.09
Global	Drift	31.52	29.56	24.83	1.98
Top-10	Mean	28.57	27.56	5.71	0.23
Top-10	Naïve	205.73	119.57	12.71	0.99
Top-10	Drift	23.61	22.33	4.63	0.19

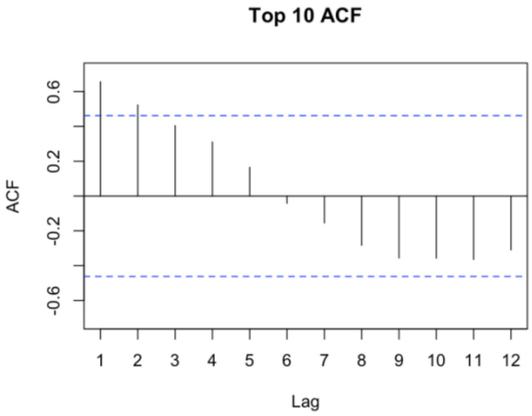
Dataset	Feature	r (Levels)	p (Levels)	r (Diff)	p (Diff)	CCF Lag	CCF Value
Global	TB incidence (HIV+)	0.870	7.1e-08	0.871	1.38e-07	0	0.870
Global	New/Relapse	0.907	2.32e-09	0.441	3.98e-02	0	0.907
Top-10	TB incidence (HIV+)	0.996	<1e-12	0.995	<1e-12	0	0.996
Lesotho	log(GDP)	-0.529	3.49e-02	0.119	0.673	+1	-0.561

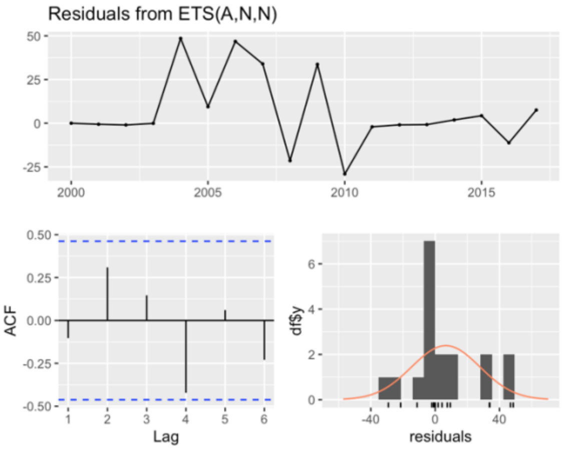


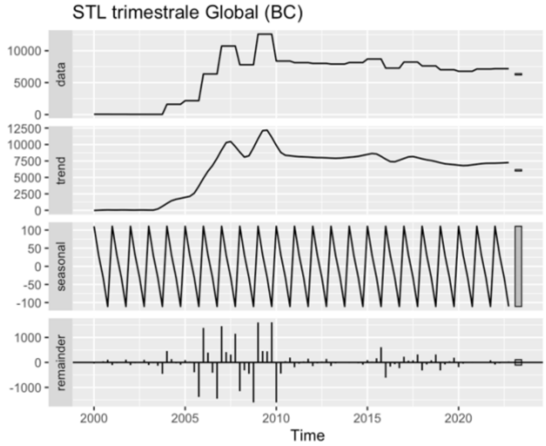


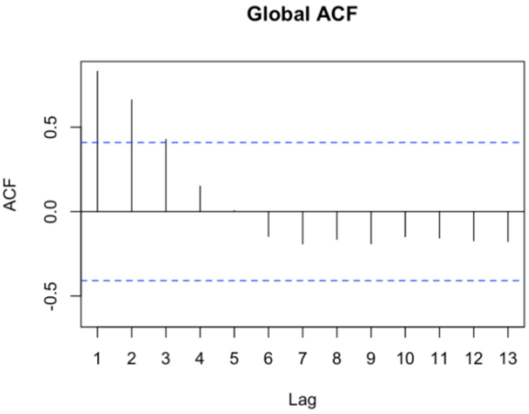

Global TB distribution
tb_incidence_total
1500
1000
500
0
x

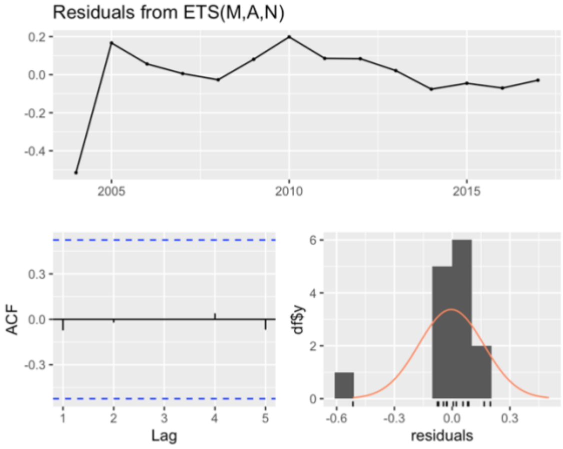

Residuals from ARIMA(0,1,0)
500
250
0
2005
2010
2015
ACF
0.3
0.0
-0.3
1
2
3
4
5
Lag
df$y
5
4
3
2
1
0
-400
0
400
residuals

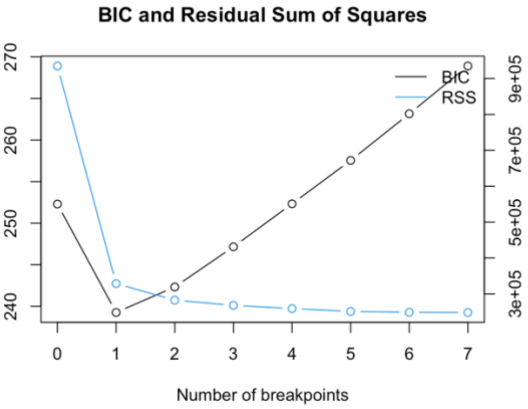

Top 10 ACF
ACF
0.6
0.2
-0.2
-0.6
1
2
3
4
5
6
7
8
9
10
11
12
Lag

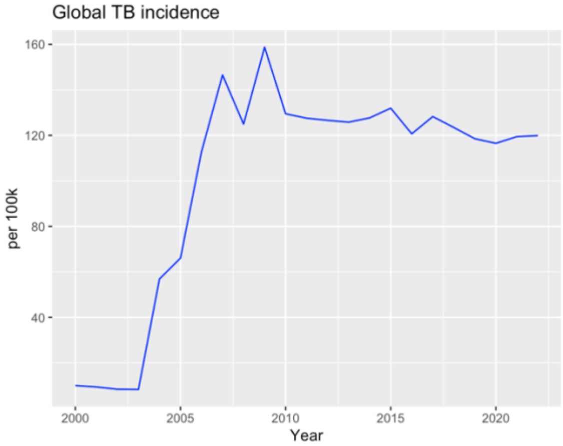

Residuals from ETS(A,N,N)
50
25
0
-25
2000
2005
2010
2015
ACF
0.50
0.25
0.00
-0.25
-0.50
1
2
3
4
5
6
Lag
df$y
6
4
2
0
-40
0
40
residuals

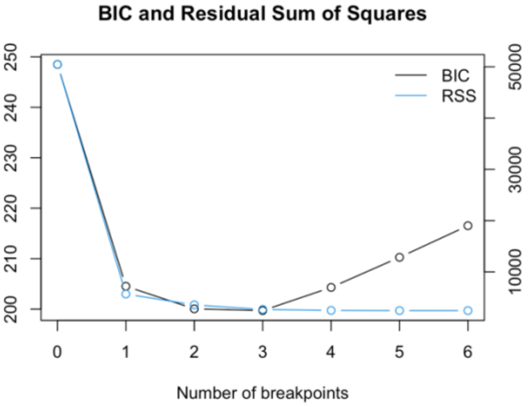

STL trimestrale Global (BC)
data
10000
5000
0
trend
12500
10000
7500
5000
2500
0
seasonal
100
50
0
-50
-100
remainder
1000
0
-1000
2000
2005
2010
2015
2020
Time

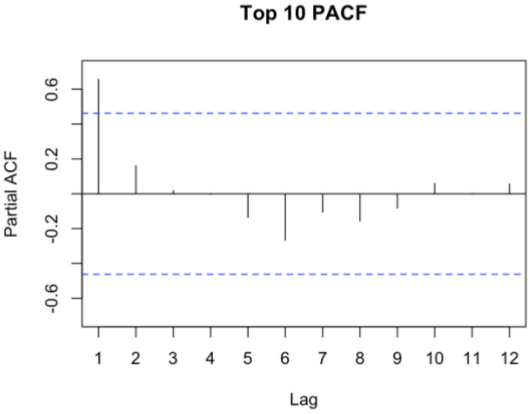

Global ACF
ACF
0.5
0.0
-0.5
1
2
3
4
5
6
7
8
9
10
11
12
13
Lag

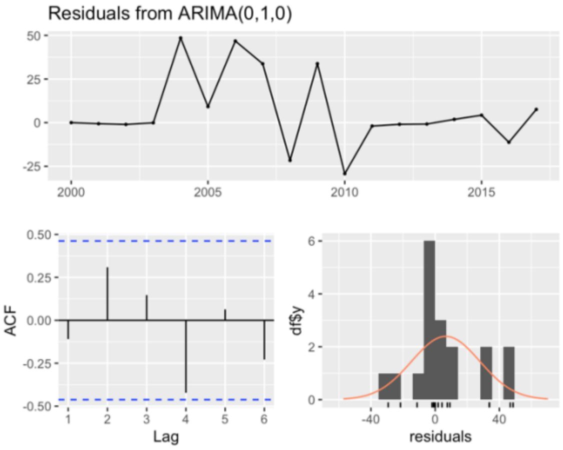

Residuals from ETS(M,A,N)
0.2
0.0
-0.2
-0.4
2005
2010
2015
ACF
0.3
0.0
-0.3
1
2
3
4
5
Lag
df$y
6
4
2
0
-0.6
-0.3
0.0
0.3
residuals

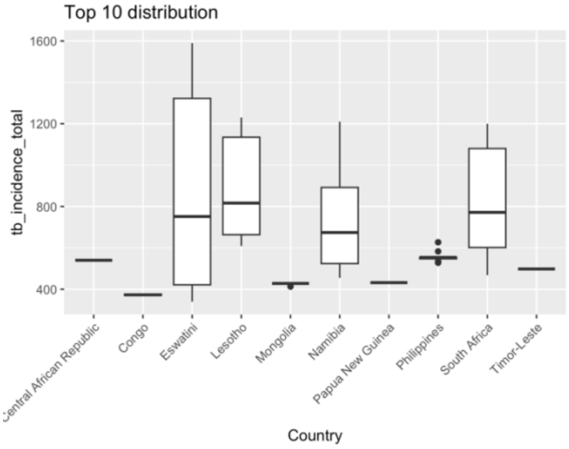

BIC and Residual Sum of Squares
270
260
250
240
BIC
RSS
9e+05
7e+05
5e+05
3e+05
0
1
2
3
4
5
6
7
Number of breakpoints

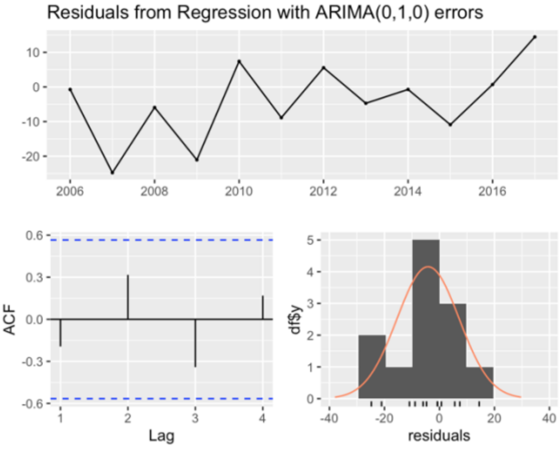

Global TB incidence
160
120
80
40
per 100k
2000
2005
2010
2015
2020
Year

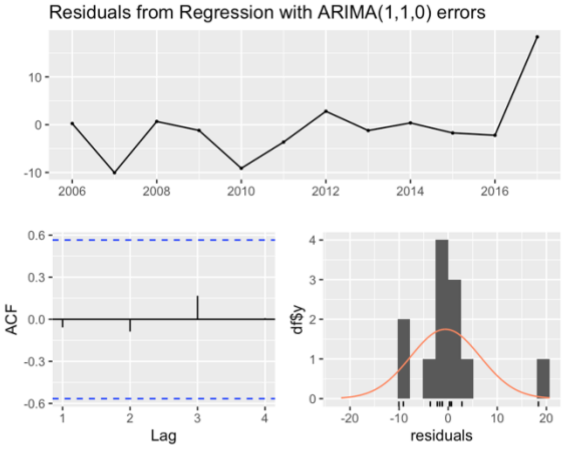

BIC and Residual Sum of Squares
250
240
230
220
210
200
BIC
RSS
50000
30000
10000
0
1
2
3
4
5
6
Number of breakpoints


Top 10 PACF
0.6
0.2
-0.2
-0.6
Partial ACF
1
2
3
4
5
6
7
8
9
10
11
12
Lag


Residuals from ARIMA(0,1,0)
50
25
0
-25
2000
2005
2010
2015
ACF
0.50
0.25
0.00
-0.25
-0.50
1
2
3
4
5
6
Lag
df$y
6
4
2
0
-40
0
40
residuals


Top 10 distribution


Residuals from Regression with ARIMA(0,1,0) errors


Residuals from Regression with ARIMA(1,1,0) errors