UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

COVID-19 World Population Gap Analysis

Ondine JI, Olha ZHUK, Miguel FONSECA PRIETO, Rei PETI, Thomas RIGOU, Hugo MICHEL

Professor: Eric Vansteenberghe

Quantitative methods in finance

-

M2 Finance Technology Data - Sorbonne School Of Economics

Université Paris 1 Panthéon-Sorbonne

Paris, France, Fall 2023

Contents

List of Tables

1 Literature Review

For this specific context about the world's population gap analysis, the literature review was based on three distinct papers to provide insights and guidance into population forecasting and the importance of demographic factors and economic outcomes. The first paper Patricia E Beeson (2001), examines county-level population growth in the United States from 1840 to 1990, it provides useful methodologies, particularly the use of OLS regressions, for examining how the characteristics of regions influence their population growth rates. The second paper Adelman (1963), written by Irma Adelman in 1963, uses a similar OLS regression and provides a historical perspective on the complex relationship between demographic shifts and economic development. Finally, Heather Booth's 2006 paper Booth (2006) is an analysis about demographic forecasting methodologies, emphasizing the use of models such as AR(1), ARIMA and ARIMAX, which are commonly used for this type of context. As well, she mentions the limitations of forecasting methods when applied to demographic trends. Taking into account these three papers, we have a solid foundation for the study's approach and analysis by understanding the methodology for forecasting population in the context of COVID-19's impact on mortality and demographics. It is also worth mentioning that all of these papers are highly ranked in the CNRS journal category, indicating that they are trustworthy and reputable sources of information.

Taking into account the information provided by these three papers, the relevant data to use would be demographic data such as age, fertility and mortality. Macroeconomic data such as the GDP, unemployment rate and immigration are also crucial to perform this analysis. As there are many factors to take into account for a population forecast, this exogenous data are key to explain its dynamics.

	Population growth in U.S counties, 1840-1990	***An Econometric Analysis of Population Growth***	***Demographic forecasting: 1980 to 2005 in review***
Authors	Patricia E Beeson (Professor) a, David N DeJong (Professor) a, Werner Troesken (Associate Professor)	Irma Adelman	Heather Booth
Year of publication	2001	1963	2006
Journal	Regional Science and Urban Economics	The American Economic Review	International journal of forecasting
CNRS category of the journal	2	1e	2
Google Scholar citation count	258	336	494
Data	Data from 1840 and 1990 about location and growth of the U.S. population using county-level census data.	From 1947 to 1957 United Nations 37 countries for fertility, with annual per capita incomes data, wide geographical distribution but Africa and Asia are relatively underrepresented 34 countries for mortality data	N.A. as it's a review
Model	Ordinary Least Squares (OLS) regression analysis	2 Regressions on fertility and mortality	AR(1), ARIMA

Figure 1.1: Literature Review

2 Model Selection

Based on the literature review, extrapolation emerges as the most prevalent approach in demographic forecasting, operating under the core assumption that the future will, in some manner, extend the patterns of the past. Univariate ARIMA modeling stands out as the most frequently used method for such extrapolation.

Our approach involves applying the ARIMA model to demographic data from countries that do not show seasonality. Conversely, for data-sets exhibiting pronounced seasonality, we plan to use the SARIMA model, tailored to manage seasonal fluctuations, thereby enabling more accurate out-of-sample forecasting. Furthermore, we propose integrating relevant exogenous variables – such as birth rate, female fertility rate, life expectancy, mortality rate, and the survival rate at 65 years – to enhance the accuracy of our forecasts, particularly during periods marked by extreme values like the COVID-19 pandemic. This will be achieved through the application of the ARIMAX model, which incorporates these exogenous factors, as also discussed in Heather Booth's paper. For the World Population data, which shows seasonality, our literature review has guided us to choose the ARIMA model.

To improve upon our initial analysis and capture the non-linear relationships that our independent variables may have with population growth, we have incorporated an LSTM (Long Short-Term Memory) neural network. Indeed, artificial neural networks stand out from econometric models by the way they apply different data transformations during the forward propagation process, allowing them to capture non-linear relationships efficiently. In our case, the LSTM network will be fitted on the residuals of the econometric model. The use of LSTM layers are justified by the fact that they are a type of recurrent neural network, which unlike traditional feedforward neural networks, have feedback connections that allows them to process entire sequences of data making them very suitable for time-series analysis.

3 Data Collection

Considering the data used in the papers on population growth we analyzed, we conducted a comprehensive data collection on global population to assess the impact of COVID-19 on the population gap. Initially, we obtained monthly population data from Factset, a private data provider, covering the [1960 to 2023] period. Subsequently, we narrowed our analysis to the subperiod from 1994 to 2020 to enhance the model's effectiveness and align with data availability across various countries.

Acquiring monthly data was challenging since credible sources publish their data on different frequencies. Most national statistical bureaus provided annual data. In rare cases where monthly data were available, there were still some challenges due to the data provider's practice of duplicating information for the subsequent month in instances of missing data throughout the year. Consequently, we observe consistent figures for certain periods, such as in the case of Argentina. The approach to deal with missing values was employed. Since we expect the changes in birth from one month to the other to not be at great levels, each missing value was replaced by the last month's value (using a forward fill function). Every value that remained uncompleted after this approach consisted of 2023 observations that would not be a significant part of the data used in this study. Subsequently, we separately gathered monthly population data for the world and five specific countries:

- Argentina
- China
- France
- New Zealand
- Norway

What criteria have determined the selection ?

1. Geographical location

Our approach involves examining diverse countries across various geographical

regions to analyze the impact of COVID-19 globally. Specifically, we have chosen countries representing different regions: Norway for the Nordic region, China for Asia, New Zealand for Oceania, Argentina for South America, and France for Europe.

2. Different population sizes of the countries

We selected countries with varying population sizes—small, medium, and large—to assess their impact on population dynamics.

3. Government policy responses to the pandemic

Our analysis includes an examination of stay-at-home and vaccine policies, comparing nations such as Norway, which terminated its COVID-19 measures early, with countries like New Zealand and Argentina, which implemented some of the world's longest lockdowns. In constructing our model, we utilized exogenous variables within the SARIMAX framework to forecast population dynamics. These variables include birth rate, fertility rate among females, life expectancy, mortality rate, and survival age at 65 and were collected from World Bank Open Data source. By incorporating these factors into our model, we aim to provide a more comprehensive understanding of the potential population evolution, considering the impact of COVID-19.

4 Model Calibration

4.1 Exploratory Data Analysis

The data processing strategy we implement is as follows:

1. Import the Data
2. Data cleansing (data imputation and removing outliers)
3. Data visualization after cleansing process

Here is the cleaned dataframe.

Figure 4.1: Dataframe Monthly population

Month	Argentina	China	France	New Zealand	Norway
1994-01-01	34.846916	1198.50	59.070	3.6115	4.324815
1994-02-01	34.846916	1198.50	59.078	3.6115	4.324815
1994-03-01	34.846916	1198.50	59.090	3.6115	4.324815
1994-04-01	34.846916	1198.50	59.105	3.6200	4.324815
1994-05-01	34.846916	1198.50	59.122	3.6200	4.324815
...	...	...	...	...	...
2019-12-01	45.089493	1410.08	67.430	5.0404	5.328212
2020-01-01	45.479120	1412.12	67.442	5.0828	5.367580
2020-02-01	45.479120	1412.12	67.457	5.0828	5.367580
2020-03-01	45.479120	1412.12	67.473	5.0828	5.367580
2020-04-01	45.479120	1412.12	67.481	5.0902	5.367580

316 rows × 5 columns

Below is the time series of the evolution of the population for each country on the list:

Figure 4.2: Argentina population evolution over time

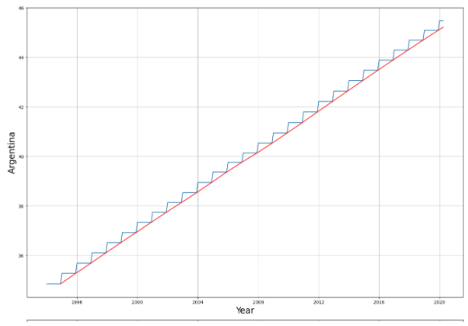


Figure 4.3: China population evolution over time

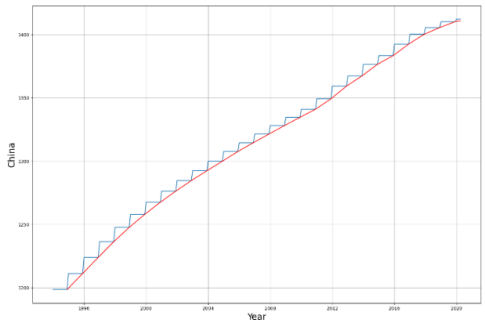


Figure 4.4: France population evolution over time

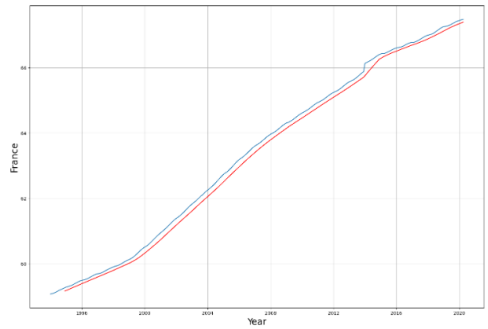


Figure 4.5: France population evolution over time

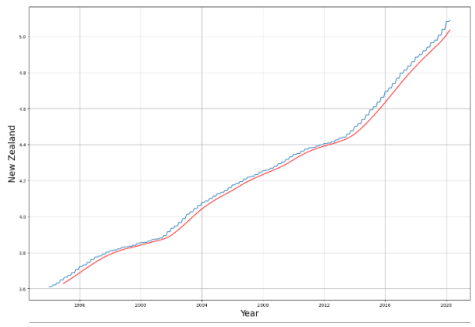
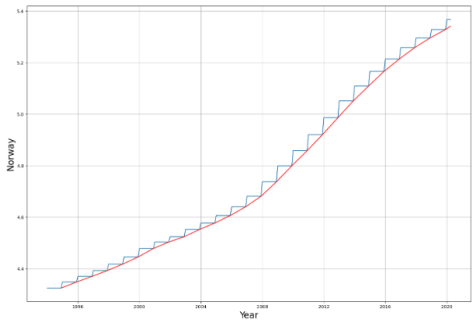
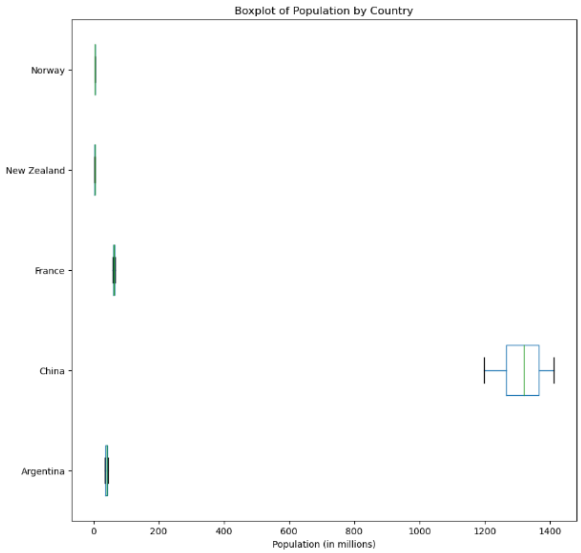


Figure 4.6: Norway population evolution over time



The dataset summary depicted through boxplot visualization.

Figure 4.7: Dataset Summary



As shown by the moving average (i.e. red line on each plot) we note that the evolution of the population of each country has a positive trend.

Moreover, we can see that Argentina, China, New Zealand and Norway have a high degree of seasonality (a pattern that repeats itself over time).

The period of seasonality seems to be one year because the pattern repeats itself every year.

As a result, we assume that this is due to the data provider's practice of duplicating information from the previous month for the subsequent month when data is missing.

Since the data provider wasn't able to collect sufficiently granular data (i.e., monthly data), the data was duplicated for several months of the year.

Otherwise, we can say that for the France population time series, the series doesn't seem to have seasonality; the values do not repeat in the subsequent months.

Overall, our time series demonstrate an upward trend with a positive drift.

In conclusion, we can state that our time series show a strong increasing trend with strong seasonality.

4.2 Decomposition of the monthly series

The time series can be decomposed into the following 3 components:

- Trend: The trend component represents the long-term, systematic, and often nonlinear movement in the data over time. It captures the underlying direction in the time series, whether it's increasing, decreasing, or remaining relatively constant. Trends can be caused by various factors, such as economic growth, population changes, or technological advancements. A time series with a clear trend component may exhibit a consistent upward or downward movement, which is not related to short-term fluctuations or seasonality.
- Seasonality: Seasonality refers to the regular, repeating patterns in the data that occur at fixed intervals. These intervals can be daily, weekly, monthly, quarterly, or any other specific time frame. Seasonality is often associated with external factors, such as the calendar (e.g., holidays), weather, or cultural events. Time series with seasonality will show periodic patterns that repeat within a particular time frame.
- Residuals: The residual component, also known as the irregular component, represents the unexplained or random variation in the time series that cannot be attributed to the trend or seasonality. It includes noise, unexpected events, and other random factors. Residuals are essentially what remains after removing the trend and seasonality components. Analyzing the residuals is important because they contain valuable information about the inherent uncertainty and unpredictability of the data.

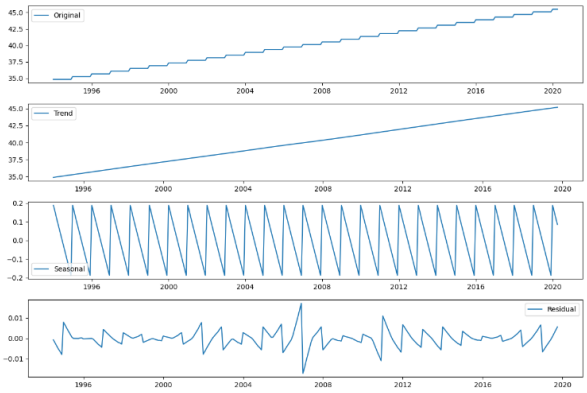
From a mathematical point of view, the decomposition of a time series is as follows:

$Y = drift + trends + seasonality + noise$

To decompose the series, we choose the following period = 12

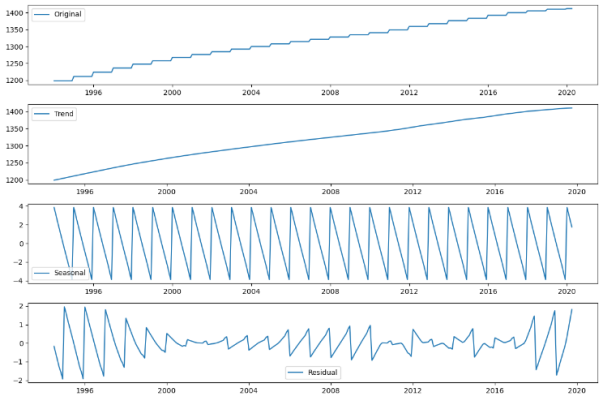
Argentina

Figure 4.8: Seasonal Decomposition for Argentina



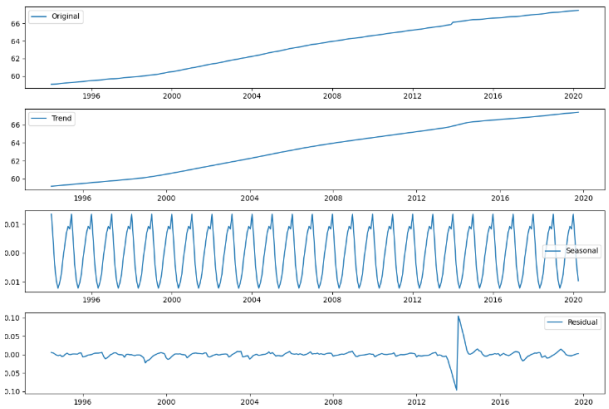
China

Figure 4.9: Seasonal Decomposition for China



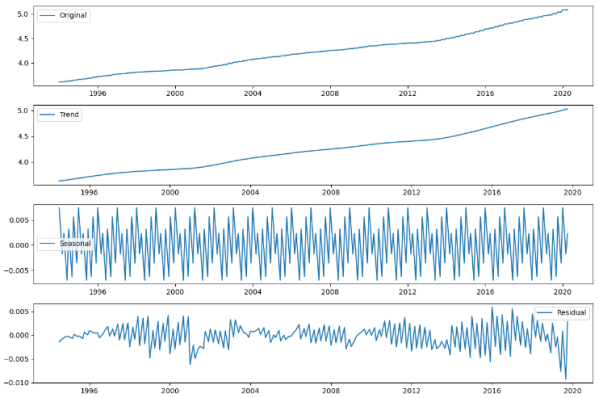
France

Figure 4.10: Seasonal Decomposition for France



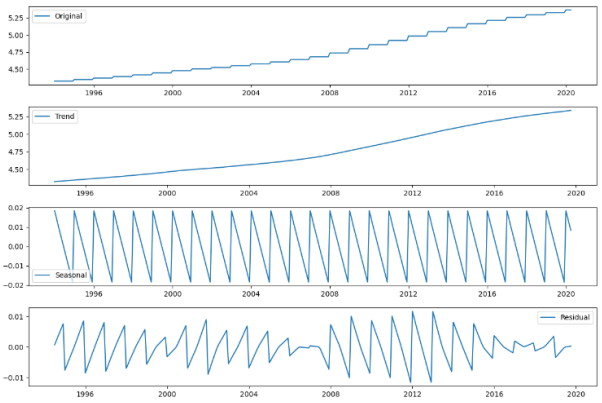
New Zealand

Figure 4.11: Seasonal Decomposition for New Zealand



Norway

Figure 4.12: Seasonal Decomposition for Norway



Analyzing trend

From the diagrams, we easily interpret that there is an upward trend for the evolution of population for each country.

Analysing Seasonality The above graphs clearly indicate a spike at the beginning of each year for Argentina, China and Norway.

4.3 Detrending Time Series

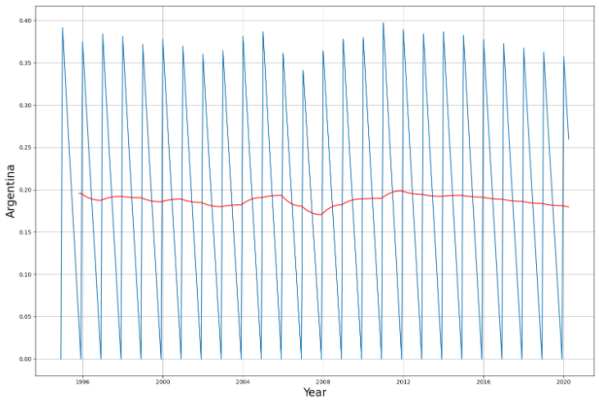
To remove the trend from the original time series, we subtract the annual moving average from the original time series.

$\text{Result}_t = Y_t - MA_t$

- Y_t denotes the original time series
- MA_t denotes the annual moving average

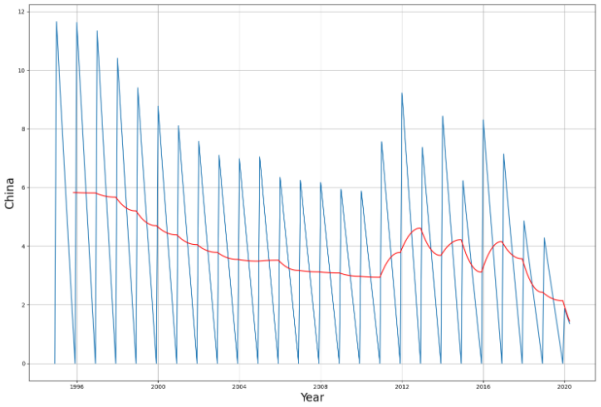
Trendless Argentina

Figure 4.13: Trendless Argentina



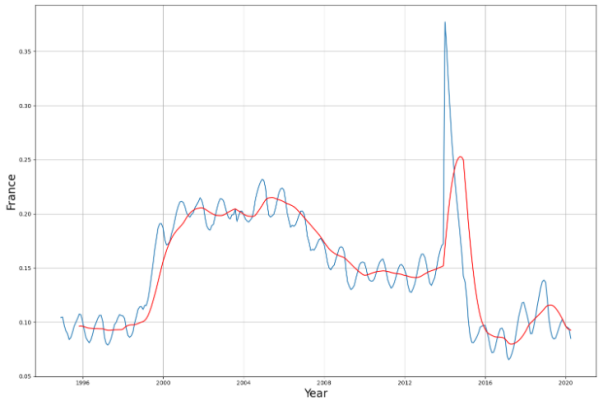
Trendless China

Figure 4.14: Trendless China



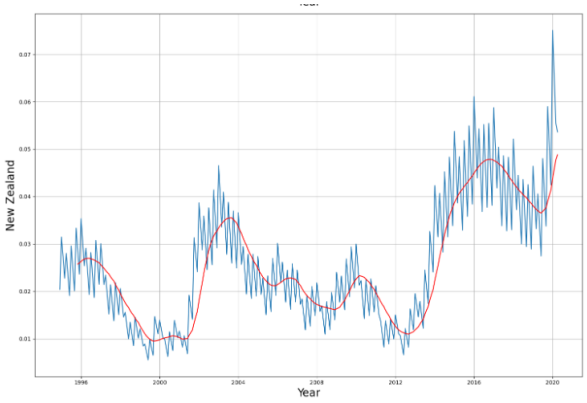
Trendless France

Figure 4.15: Trendless France



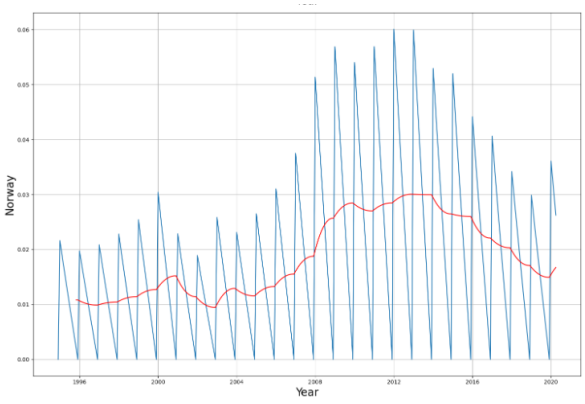
Trendless New-Zealand

Figure 4.16: Trendless New-Zealand



Trendless Norway

Figure 4.17: Trendless Norway



As we can see, even by removing the trend of our time series, the data seems to remain non-stationary. Hence, removing the trend of the time series doesn't appear to be enough

to transform our time series into a stationary time series.

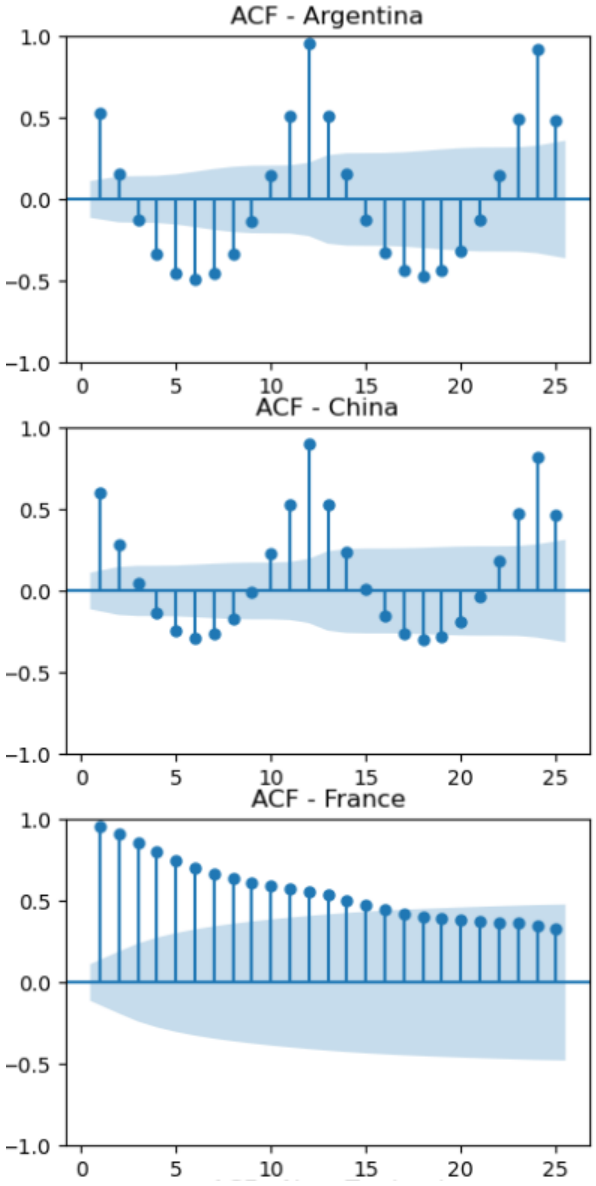
4.4 Identifying seasonal period using ACF

We plot the ACF of the detrended data in order to identify the period of seasonality of our series.

To do so we define a list of lags : $lags = [12, 24, 36, 48, 60, 72]$

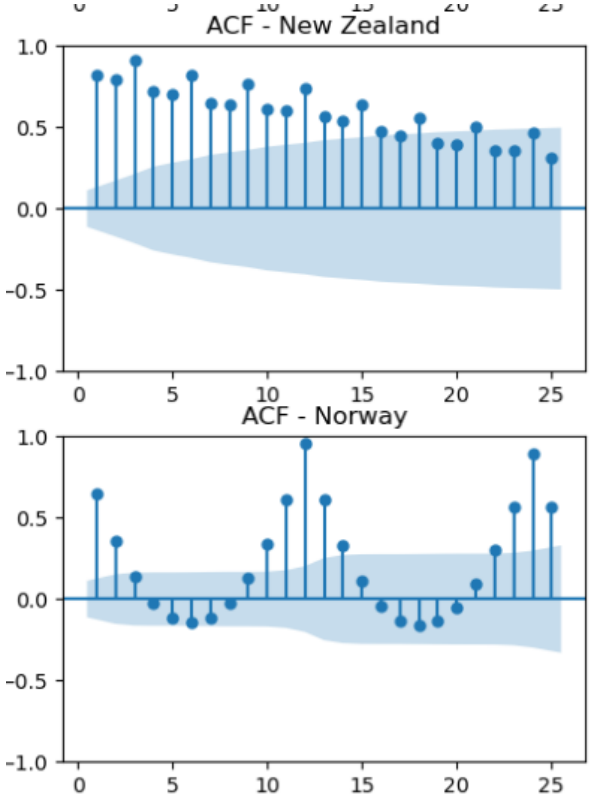
Seasonal ACF for Argentina, China, France

Figure 4.18: Seasonal ACF for Argentina, China, France



Seasonal ACF for New-Zealand, Norway

Figure 4.19: Seasonal ACF for New-Zealand, Norway



The ACF plot for Argentina, China, Norway shows that there is a seasonal component (we see a peak at lag 12). With this in mind, we can improve our predictions. Indeed, the time period of the seasonal component of the data is 12.

However, the time series for New Zealand and France doesn't seem to include a seasonal pattern contrary to what we noted previously. If we focus on the ACF plot for Norway, we observe that every lag that is a multiple of 3, holds greater significance. This is most likely caused by the data provider updating the country's information every quarter.

Hence, we have to deal with seasonal time series for the following countries:

- Argentina
- China
- Norway

Hence, we have to deal with non-seasonal time series for the following countries:

- France

- New Zealand

4.5 Check for the stationary

To check the stationarity of the population time series, we assume that the time series have a deterministic constant offset and a linear trend.

We perform an ADF test to examine the presence of a unit root for each series.

In this regard, we consider the following hypothesis in the right order:

$\Delta X_t = b_0 + \rho X_{t-1} + \sum_{j=1}^{p-1} \phi_j \delta X_{t-j} + \epsilon_t$

REGRESSION: CONSTANT AND TREND By performing this test, we assume that the time series has a deterministic constant offset and a linear trend.

Test for a deterministic trend:

- H_0 : The trend coefficient is not significant ($b_1 = 0$)
- H_1 : The trend coefficient is significant ($b_1 \neq 0$)

If H_0 is rejected, we accept $H1$ and check the presence of unit root Test for unit root with trend:

- H_0: There is a unit root (i.e. $\rho \neq 0$ non-stationary) with a significant trend. The time series is non-stationary with a deterministic trend.
- H_1: There is no unit root (i.e. $\rho = 0$ stationary), no stochastic trend, but a deterministic trend.

Given that we have seasonal and non-seasonal time series, we have to apply different approaches for both type of time series

4.5.1 Seasonal differencing

We will now focus on seasonal time series (Argentina, China, Norway)

There are many transformations we can apply to a time series to remove the seasonality and make it stationary.

As we know that our time series has a seasonal pattern and upward trend, we will try the following techniques:

- substract with moving average
- substract the time series value of one season ago

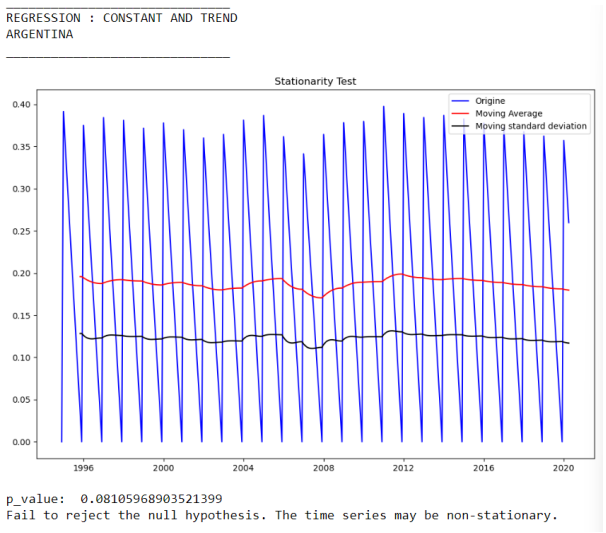
1st Technique : Subtract with the moving average

$\text{Result}_t = Y_t - MA_t$

- Y_t denotes the original time series
- MA_t denotes the annual moving average

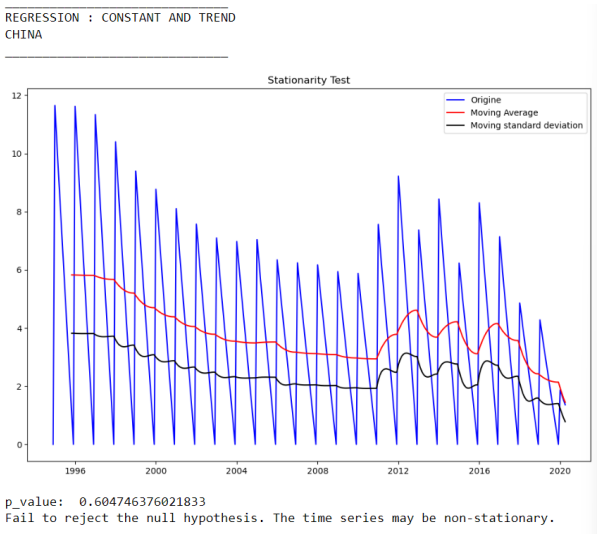
Argentina : 1st difference (First technique)

Figure 4.20: Argentina : 1st difference (First technique)



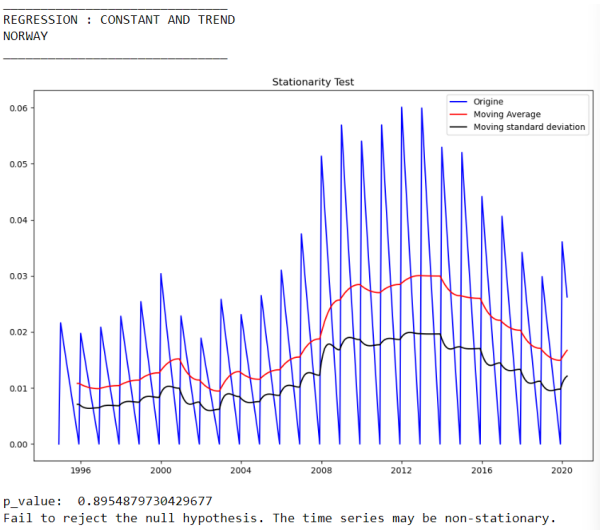
China : 1st difference (First technique)

Figure 4.21: China : 1st difference (First technique)



Norway : 1st difference (First technique)

Figure 4.22: Norway : 1st difference (First technique)



We can see that with this 1st technique of difference, our series remains non-stationary. Let's apply the second technique.

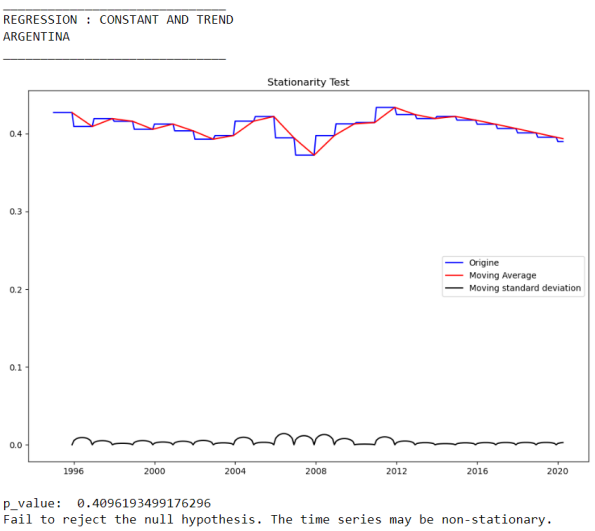
2nd technique : Subtract one season back

$\Delta_{y_t} = y_t - y_{t-S}$

To do so, we remove the seasonal pattern of seasonal time series by subtracting the seasonal component estimated from the year prior to the actual observation (S = months)

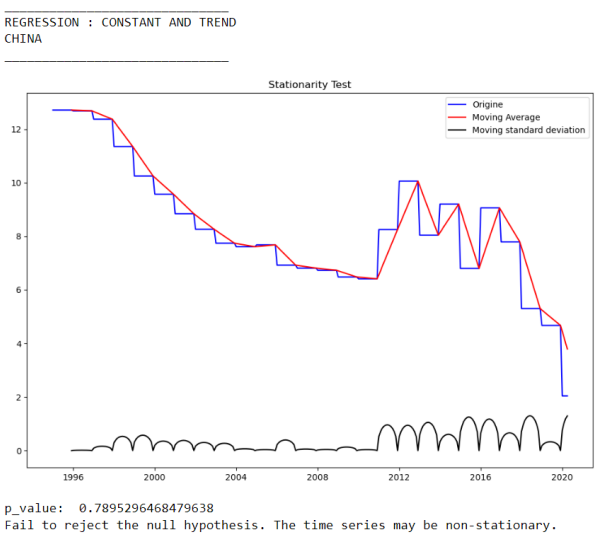
Argentina : 1st difference (Second technique)

Figure 4.23: Argentina : 1st difference (Second technique)



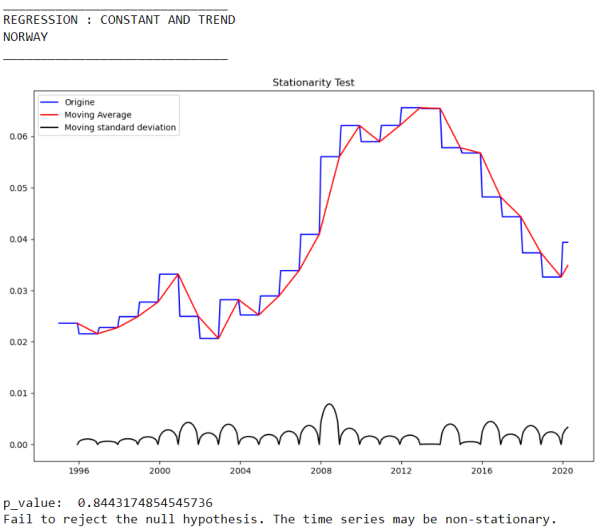
China : 1st difference (Second technique)

Figure 4.24: China : 1st difference (Second technique)



Norway : 1st difference (Second technique)

Figure 4.25: Norway : 1st difference (Second technique)



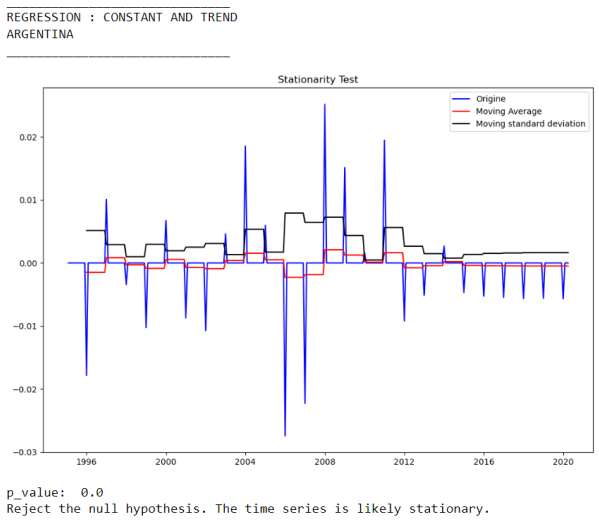
We can see that after taking the first difference with this second technique all of our time series remained non-stationary.

Therefore, we have to take the second difference.

Take the 2nd difference

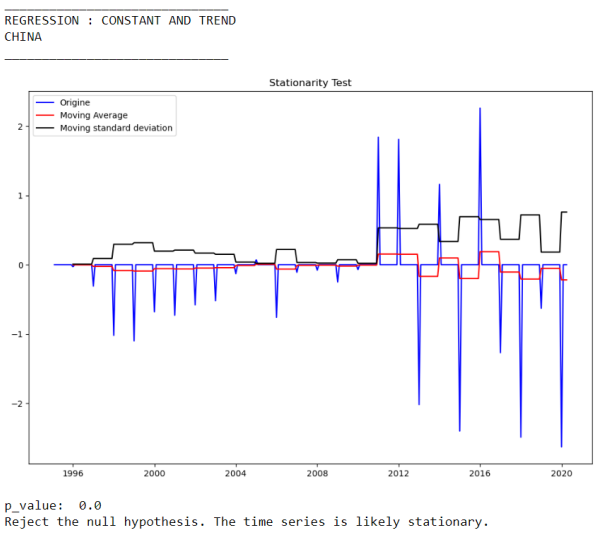
Let's take the second difference to transform our non-stationary series into stationary series

Figure 4.26: Argentina : 2nd difference (Second technique)



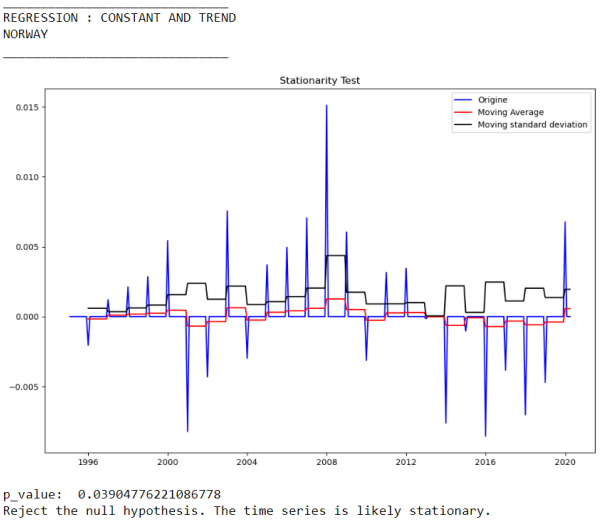
China : 2nd difference (Second technique)

Figure 4.27: China : 2nd difference (Second technique)



Norway : 2nd difference (Second technique)

Figure 4.28: Norway : 2nd difference (Second technique)



After taking the second difference, we observe that the series are now stationary

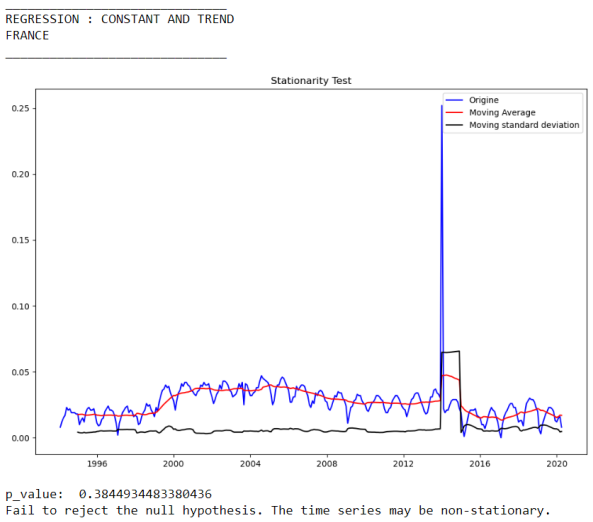
4.6 Non-seasonal differencing

To take the first difference of a non-seasonal time series we apply the following formula:

$\Delta_{y_t} = y_t - y_{t-1}$

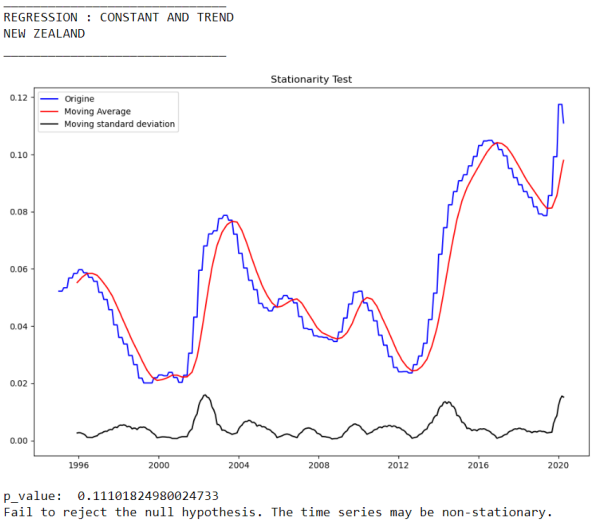
France : 1st difference

Figure 4.29: France : 1st difference



New Zealand : 1st difference

Figure 4.30: New Zealand : 1st difference

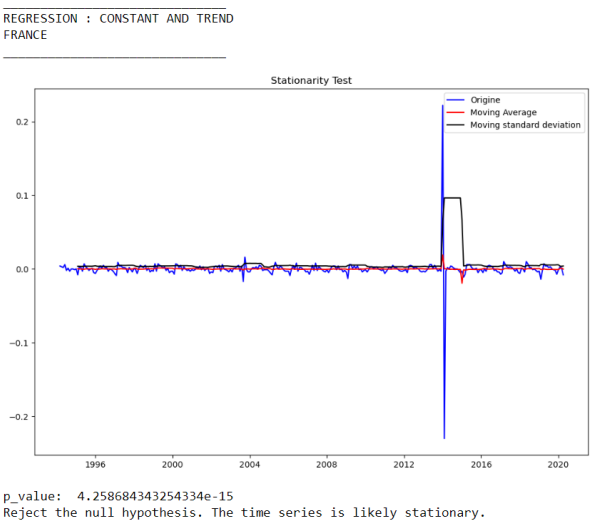


After taking the 1st difference for the non-seasonal time series, they remain non-stationary. Therefore, we need to proceed with the second difference.

Let's take the 2nd difference by applying the previously mentioned formula twice

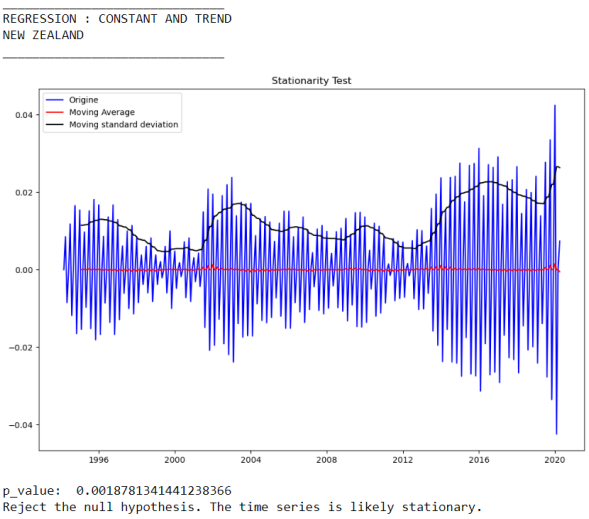
France : 2nd difference

Figure 4.31: France : 2nd difference



New Zealand : 2nd difference

Figure 4.32: New Zealand : 2nd difference



After taking the 2nd difference, the series are now stationary

4.7 Identifying orders of ARIMA model using ACF and PACF

To find seasonal orders, we plot the ACF and PACF of differentiated time series.

By comparing the ACF and PACF for a time series, we can deduce the order of the model. If the amplitude of the ACF decreases with increasing lag and the PACF cuts off after a certain lag p, then we have an AR(p) model.

If the ACF amplitude cuts off after a certain offset q and the PACF amplitude decreases, then we have an MA(q) model.

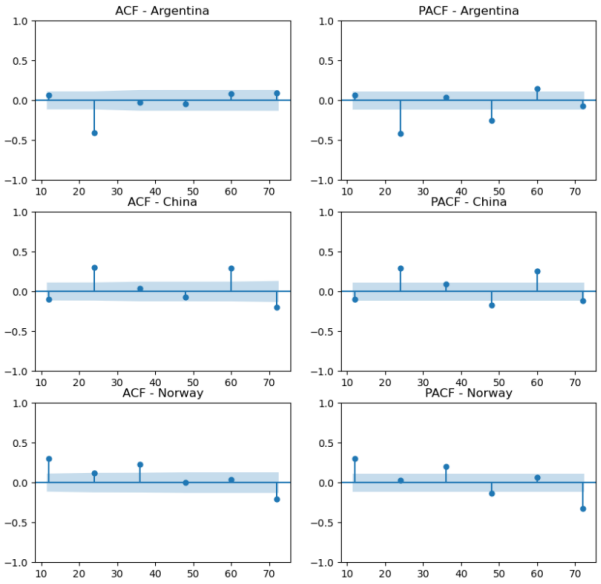
In summary:

AP(p)	**MA(q)**	**ARMA(p,q)**
Tails off	Cuts off after lag q	Tails off
Cuts off after lag p	Tails off	Tails off

Focus on seasonal time series

In order to plot the ACF and PACF of seasonal time series, we set the lags parameter to a list of offsets instead of a maximum. This traces the ACF and PACF to these specific offsets only.

Figure 4.33: ACF and PACF plot for seasonal timeseries



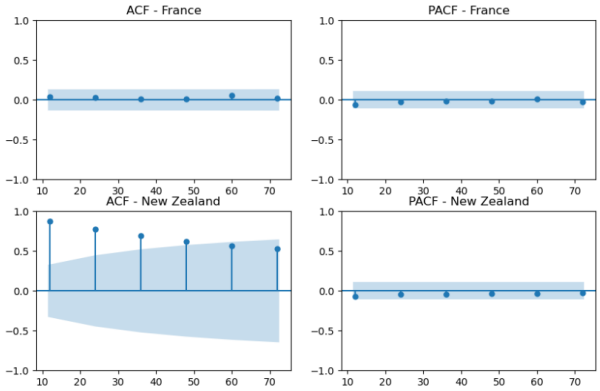
According to the ACF plot of Argentina, we can see that the ACF plot cuts off after lag 2. So it seems to be MA(2)

According to the ACF and PACF of China and Argentina, we notice that the autocorrelation at lag 1 is negative, potentially indicating that we've taken the difference too many times.

Otherwise, according to the ACF and PACF plots of China and Norway, it is difficult to identify the best order for an ARIMA model.

Focus on non-seasonal time series

Figure 4.34: ACF and PACF plot for non-seasonal timeseries



According to the ACF and PACF plots of France, all the lags are very small. Therefore, we can't identify the best order.

According to the ACF plot of New Zealand, the ACF values are high and decrease very slowly; this is a sign that the data may not be perfectly stationary.

4.8 Forecast the population

To forecast the population of each country, we apply an ARIMA model for non-seasonal timeseries and a SARIMA model for seasonal timeseries. We use the following methodology to forecast the population of each country:

The strategy to forecast the population is as follows:

1. Split data into training data (non-covid period) and test data (covid period)
2. Search the best orders over AIC and BIC criteria thanks to *auto-arima package*
3. Analyze the residuals of the fitted model
4. Forecast the population
5. Visualize the forecasting period
6. Evaluation of the forecasting performance via the MSE score

More precisely, in order to check the residuals of the fitted model we also check that the residuals of the model are not serially correlated. To do so, we perform an ACF and

Liung-Box-test and plot different graphics such as Normal Q-Q, Histogram + estimated density and correlogram

Residuals plot

This plot shows the standardized residuals. If our model works correctly, there should be no obvious structure in the residuals.

Histogram + estimated density

The histogram shows the measured distribution; the orange line shows a smoothed version of this histogram; and the green line shows a normal distribution. If our model is correct, these two lines should be almost identical.

Normal Q-Q

The normal Q-Q diagram is another way of showing how the distribution of model residuals compares to a normal distribution. If our residuals are normally distributed, all points should lie along the line.

Correlogram

The final graph is the correlogram, which is simply an ACF plot estimated on the model's residuals. 95% of correlations for a shift greater than zero should not be significant. If there is a significant correlation in the residuals, it means that there is information in the data that our model has not captured.

In summary:

- Standardized residual: There is no obvious pattern in the residuals
- Histogram + KDE estimate: The KDE curve should be very similar to the normal distribution
- Normal Q-Q: Most of the data points should be lie on the straight line
- Correlogram: 95% of correlations for lag grater than zero should be significant
- Ljung-Box test

In the Ljung-Box test, we evaluate the following hypothesis:

- H_0: The residuals are independently distributed

- H_1: The residuals are not independently distributed; they exhibit serial correlation.

To evaluate the performance of the ARIMA, we apply the MSE score by comparing the distance between our prediction and the ground truth.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where: * n is the number of observations, * y_i is the actual value for the i-th observation, * $\hat{y}_i$ is the predicted value for the i-th observation.

4.8.1 Fitting ARIMA for non-seasonal timeseries

- p**:** number of autoregressive terms (AR order)
- d**:** number of non-seasonal differences (differentiation order)
- q**:** number of moving average terms (MA order)

Given that all our series are differentiated two times, we set the order d at 2 because we took the second difference of our time series

Forecasting France population

Figure 4.35: Best orders selection with auto-arima

```
Performing stepwise search to minimize aic
 ARIMA(1,2,1)(0,0,0)[0] intercept   : AIC=-1377.951, Time=0.31 sec
 ARIMA(0,2,0)(0,0,0)[0] intercept   : AIC=-888.857, Time=0.08 sec
 ARIMA(1,2,0)(0,0,0)[0] intercept   : AIC=-1140.530, Time=0.06 sec
 ARIMA(0,2,1)(0,0,0)[0] intercept   : AIC=inf, Time=0.23 sec
 ARIMA(0,2,0)(0,0,0)[0]             : AIC=-890.857, Time=0.05 sec
 ARIMA(2,2,1)(0,0,0)[0] intercept   : AIC=-1507.578, Time=0.49 sec
 ARIMA(2,2,0)(0,0,0)[0] intercept   : AIC=-1274.072, Time=0.10 sec
 ARIMA(3,2,1)(0,0,0)[0] intercept   : AIC=-1549.531, Time=0.53 sec
 ARIMA(3,2,0)(0,0,0)[0] intercept   : AIC=-1356.008, Time=0.14 sec
 ARIMA(3,2,2)(0,0,0)[0] intercept   : AIC=-1500.853, Time=0.61 sec
 ARIMA(2,2,2)(0,0,0)[0] intercept   : AIC=-1447.519, Time=0.68 sec
 ARIMA(3,2,1)(0,0,0)[0]             : AIC=inf, Time=0.45 sec

Best model:  ARIMA(3,2,1)(0,0,0)[0] intercept
Total fit time: 3.729 seconds
```

The best orders selected by the package are ARIMA$(3, 2, 1)$

Figure 4.36: Auto-Arima summary

SARIMAX Results

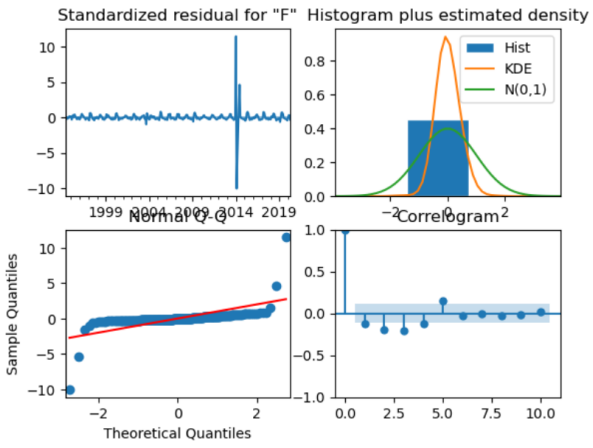
Dep. Variable:	y	**No. Observations:**	314
Model:	SARIMAX(3, 2, 1)	**Log Likelihood**	780.766
Date:	Sat, 11 Nov 2023	**AIC**	-1549.531
Time:	22:38:54	**BIC**	-1527.073
Sample:	03-01-1994	**HQIC**	-1540.555
	- 04-01-2020		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	-6.022e-06	3.69e-05	-0.163	0.870	-7.84e-05	6.63e-05
ar.L1	-1.1252	0.015	-76.876	0.000	-1.154	-1.097
ar.L2	-0.8617	0.024	-36.199	0.000	-0.908	-0.815
ar.L3	-0.3619	0.016	-22.037	0.000	-0.394	-0.330
ma.L1	-0.9849	0.024	-40.442	0.000	-1.033	-0.937
sigma2	0.0004	8.37e-06	44.392	0.000	0.000	0.000

Ljung-Box (L1) (Q):	7.16	**Jarque-Bera (JB):**	108291.01
Prob(Q):	0.01	**Prob(JB):**	0.00
Heteroskedasticity (H):	47.28	**Skew:**	1.17
Prob(H) (two-sided):	0.00	**Kurtosis:**	94.24

Figure 4.37: Check residuals for ARIMA$(3, 2, 1)$

```
##########################################
Check residuals of ARIMA for FRANCE
##########################################

Mean of Residuals: 0.00014418656652218548
We reject H0 and accept H1:The residuals are not independently distributed
Serial correlation detected in residuals.
```



The Q-Q plot looks heavy-tailed. This means that, compared with the normal distribution, there is much more data in the extremities than in the center of the distribution.

Furthermore, according to the density histogram, we can see that the green line showing a normal distribution is far from the orange line. This suggests, that our ARIMA model

can be improved upon.

Otherwise, according to the correlogram, there is no significant correlation in the residuals; our model seems to be a good fit of the information in the data

Figure 4.38: Forecasting population of France

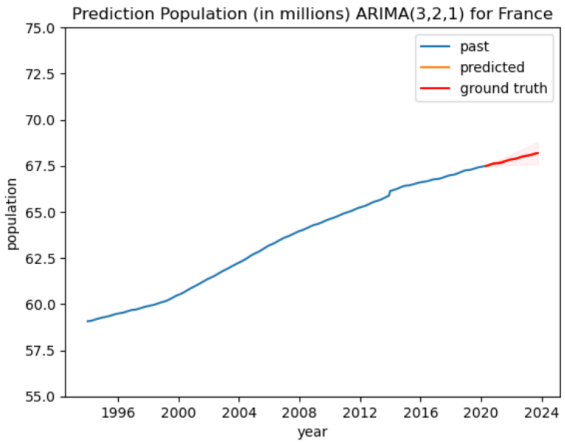
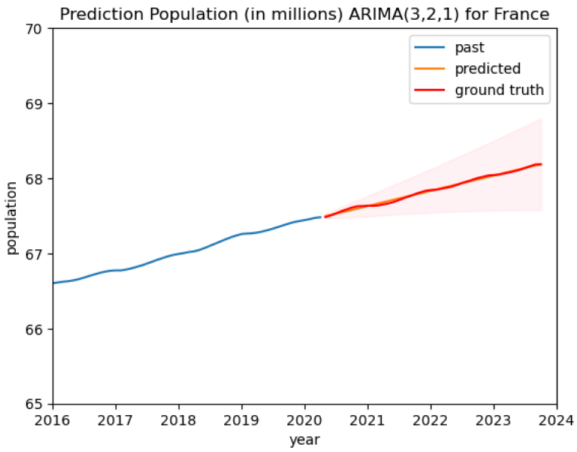


Figure 4.39: Zoom on the forecasting period



The MSE score is : 0.000241

The ARIMA model seems to forecast pretty well the population growth of France.

Forecasting New-Zealand population

Figure 4.40: Best orders selection with auto-arima

```
Performing stepwise search to minimize aic
 ARIMA(1,2,1)(0,0,0)[0] intercept   : AIC=-1535.324, Time=0.28 sec
 ARIMA(0,2,0)(0,0,0)[0] intercept   : AIC=-1123.212, Time=0.06 sec
 ARIMA(1,2,0)(0,0,0)[0] intercept   : AIC=-1221.176, Time=0.06 sec
 ARIMA(0,2,1)(0,0,0)[0] intercept   : AIC=inf, Time=0.22 sec
 ARIMA(0,2,0)(0,0,0)[0]             : AIC=-1125.212, Time=0.08 sec
 ARIMA(2,2,1)(0,0,0)[0] intercept   : AIC=-2358.560, Time=0.25 sec
 ARIMA(2,2,0)(0,0,0)[0] intercept   : AIC=inf, Time=0.08 sec
 ARIMA(3,2,1)(0,0,0)[0] intercept   : AIC=-2448.697, Time=0.53 sec
 ARIMA(3,2,0)(0,0,0)[0] intercept   : AIC=inf, Time=0.22 sec
 ARIMA(3,2,2)(0,0,0)[0] intercept   : AIC=-2369.699, Time=0.85 sec
 ARIMA(2,2,2)(0,0,0)[0] intercept   : AIC=-2416.677, Time=0.24 sec
 ARIMA(3,2,1)(0,0,0)[0]             : AIC=-2450.074, Time=0.40 sec
 ARIMA(2,2,1)(0,0,0)[0]             : AIC=-2423.172, Time=0.19 sec
 ARIMA(3,2,0)(0,0,0)[0]             : AIC=inf, Time=0.06 sec
 ARIMA(3,2,2)(0,0,0)[0]             : AIC=inf, Time=0.73 sec
 ARIMA(2,2,0)(0,0,0)[0]             : AIC=inf, Time=0.23 sec
 ARIMA(2,2,2)(0,0,0)[0]             : AIC=inf, Time=0.33 sec

Best model:  ARIMA(3,2,1)(0,0,0)[0]
Total fit time: 4.834 seconds
```

The best orders selected by the package are ARIMA$(3, 2, 1)$

Figure 4.41: Auto-Arima summary

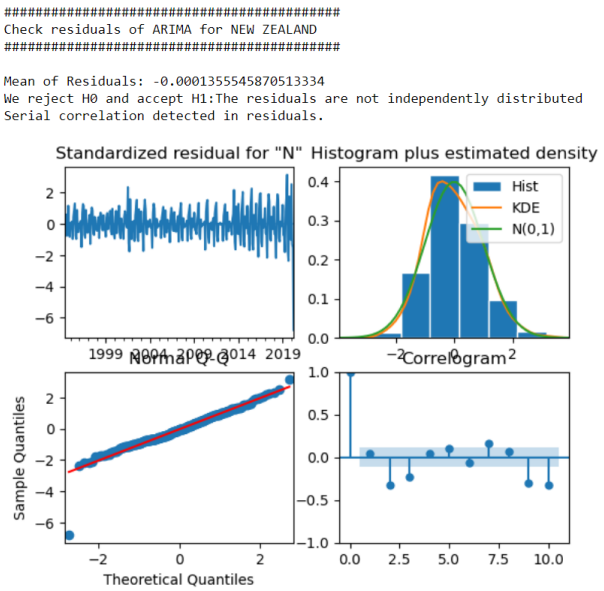
SARIMAX Results

Dep. Variable:	y	**No. Observations:**	314
Model:	SARIMAX(3, 2, 1)	**Log Likelihood**	1230.037
Date:	Sat, 11 Nov 2023	**AIC**	-2450.074
Time:	22:39:06	**BIC**	-2431.359
Sample:	03-01-1994	**HQIC**	-2442.594
	- 04-01-2020		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-1.6509	0.045	-36.950	0.000	-1.738	-1.563
ar.L2	-1.6215	0.047	-34.869	0.000	-1.713	-1.530
ar.L3	-0.6609	0.046	-14.260	0.000	-0.752	-0.570
ma.L1	-0.8472	0.061	-13.860	0.000	-0.967	-0.727
sigma2	2.127e-05	1.17e-06	18.106	0.000	1.9e-05	2.36e-05

Ljung-Box (L1) (Q):	0.58	**Jarque-Bera (JB):**	478.20
Prob(Q):	0.45	**Prob(JB):**	0.00
Heteroskedasticity (H):	3.07	**Skew:**	-0.72
Prob(H) (two-sided):	0.00	**Kurtosis:**	8.89

Figure 4.42: Check residuals for ARIMA$(3, 2, 1)$



The Q-Q plot looks pretty good. All the data is close to the red line.

Furthermore, according to the density histogram, we can see that the green line showing a normal distribution is close to the orange line. This suggest that our ARIMA model is pretty good

Otherwise, according to the correlogram, there seems to be a significant correlation in the residuals, and our model struggles to capture the information present in the data.

Figure 4.43: Forecasting population of New Zealand

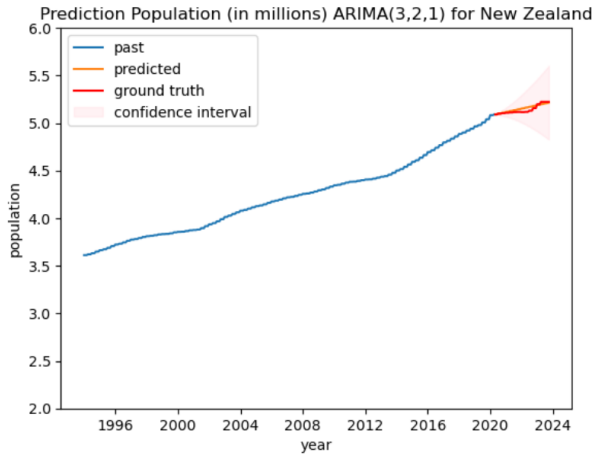
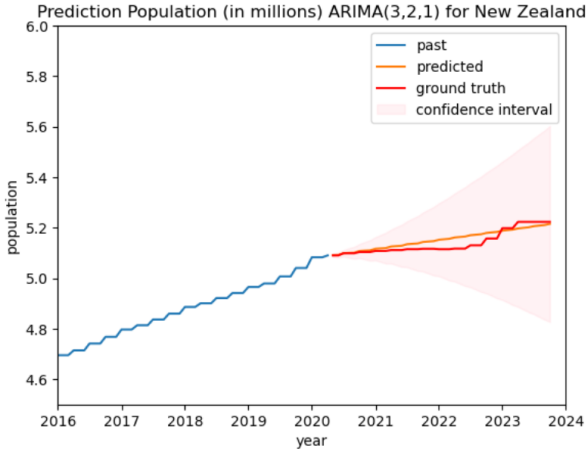


Figure 4.44: Zoom on the forecasting period



The MSE score is : 0.000609

4.8.2 Fitting SARIMA for seasonal timeseries

As for non-stationary series, we use the auto-arima package to find the best parameters for the SARIMA model.

$SARIMA(p,d,q)(P,D,Q)_s$

- P: seasonal AR order
- D: seasonal differencing order
- Q: seasonal MA order
- S: number of time step per cycle (i.e 12 in our case as we identify previously thanks to the seasonal ACF plot)

Given that we took the second difference to make our time series stationary we set the seasonal differencing parameter D at 2.

Forecasting Argentina population

Figure 4.45: Best orders selection with auto-arima

```
Performing stepwise search to minimize aic
 ARIMA(2,0,2)(1,2,1)[12]             : AIC=inf, Time=7.83 sec
 ARIMA(0,0,0)(0,2,0)[12]             : AIC=-1929.113, Time=0.14 sec
 ARIMA(1,0,0)(1,2,0)[12]             : AIC=-1956.174, Time=0.59 sec
 ARIMA(0,0,1)(0,2,1)[12]             : AIC=inf, Time=2.27 sec
 ARIMA(1,0,0)(0,2,0)[12]             : AIC=-1927.113, Time=0.22 sec
 ARIMA(1,0,0)(2,2,0)[12]             : AIC=-2108.054, Time=3.87 sec
 ARIMA(1,0,0)(3,2,0)[12]             : AIC=-2113.972, Time=4.09 sec
 ARIMA(1,0,0)(3,2,1)[12]             : AIC=-2176.188, Time=15.05 sec
 ARIMA(1,0,0)(2,2,1)[12]             : AIC=-2126.029, Time=3.91 sec
 ARIMA(1,0,0)(3,2,2)[12]             : AIC=inf, Time=19.21 sec
 ARIMA(1,0,0)(2,2,2)[12]             : AIC=inf, Time=10.88 sec
 ARIMA(0,0,0)(3,2,1)[12]             : AIC=-2172.255, Time=17.36 sec
 ARIMA(2,0,0)(3,2,1)[12]             : AIC=-2175.169, Time=16.46 sec
 ARIMA(1,0,1)(3,2,1)[12]             : AIC=-2181.262, Time=25.55 sec
 ARIMA(1,0,1)(2,2,1)[12]             : AIC=-2124.029, Time=8.33 sec
 ARIMA(1,0,1)(3,2,0)[12]             : AIC=-2111.972, Time=7.68 sec
 ARIMA(1,0,1)(3,2,2)[12]             : AIC=-2229.160, Time=29.60 sec
 ARIMA(1,0,1)(2,2,2)[12]             : AIC=-2204.637, Time=17.93 sec
 ARIMA(1,0,1)(3,2,3)[12]             : AIC=inf, Time=30.61 sec
 ARIMA(1,0,1)(2,2,3)[12]             : AIC=inf, Time=31.54 sec
 ARIMA(0,0,1)(3,2,2)[12]             : AIC=-2225.186, Time=13.42 sec
 ARIMA(2,0,1)(3,2,2)[12]             : AIC=inf, Time=18.63 sec
 ARIMA(1,0,2)(3,2,2)[12]             : AIC=-2207.263, Time=8.59 sec
 ARIMA(0,0,0)(3,2,2)[12]             : AIC=-2197.521, Time=5.78 sec
 ARIMA(0,0,2)(3,2,2)[12]             : AIC=-2209.257, Time=9.67 sec
 ARIMA(2,0,0)(3,2,2)[12]             : AIC=inf, Time=23.43 sec
 ARIMA(2,0,2)(3,2,2)[12]             : AIC=inf, Time=25.19 sec
 ARIMA(1,0,1)(3,2,2)[12] intercept   : AIC=-2203.234, Time=12.95 sec

Best model:  ARIMA(1,0,1)(3,2,2)[12]
Total fit time: 370.830 seconds
```

The best orders selected by the package are SARIMA$(1,0,1)(3,2,2)_12$

Figure 4.46: Auto-Arima summary

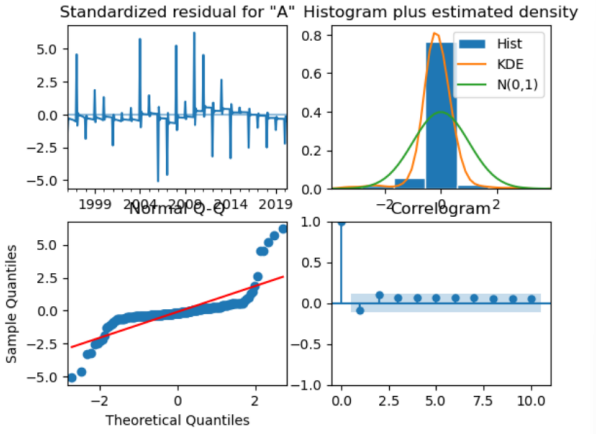
SARIMAX Results

Dep. Variable:	y	**No. Observations:**	303
Model:	SARIMAX(1, 0, 1)x(3, 2, [1, 2], 12)	**Log Likelihood**	1122.580
Date:	Sat, 11 Nov 2023	**AIC**	-2229.160
Time:	22:45:19	**BIC**	-2200.111
Sample:	02-01-1995	**HQIC**	-2217.507
	- 04-01-2020		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.85e-05	6.98e-08	264.994	0.000	1.84e-05	1.86e-05
ma.L1	-1.85e-05	6.48e-08	-285.487	0.000	-1.86e-05	-1.84e-05
ar.S.L12	-0.8567	0.058	-14.712	0.000	-0.971	-0.743
ar.S.L24	-0.7159	0.045	-16.050	0.000	-0.803	-0.628
ar.S.L36	-0.2177	0.048	-4.490	0.000	-0.313	-0.123
ma.S.L12	-0.0565	0.073	-0.775	0.438	-0.199	0.086
ma.S.L24	-0.7112	0.058	-12.201	0.000	-0.825	-0.597
sigma2	1.587e-05	8.7e-07	18.246	0.000	1.42e-05	1.76e-05

Ljung-Box (L1) (Q):	0.00	**Jarque-Bera (JB):**	18223.61
Prob(Q):	0.96	**Prob(JB):**	0.00
Heteroskedasticity (H):	0.38	**Skew:**	-0.73
Prob(H) (two-sided):	0.00	**Kurtosis:**	42.57

Figure 4.47: Check residuals for SARIMA$(1,0,1)(3,2,2)_12$



The Q-Q plot looks heavy-tailed. This means that, compared with the normal distribution, there is much more data in the extremes rather than at the center of the distribution.

Furthermore, according to the density histogram, we can see that the green line showing a normal distribution is far from the KDE distribution (i.e. orange line. This suggests that our ARIMA model can be improved.

Otherwise, according to the correlogram, there is no significant correlation in the residuals; our model seems to capture well the information in the data

Figure 4.48: Forecasting population of Argentina

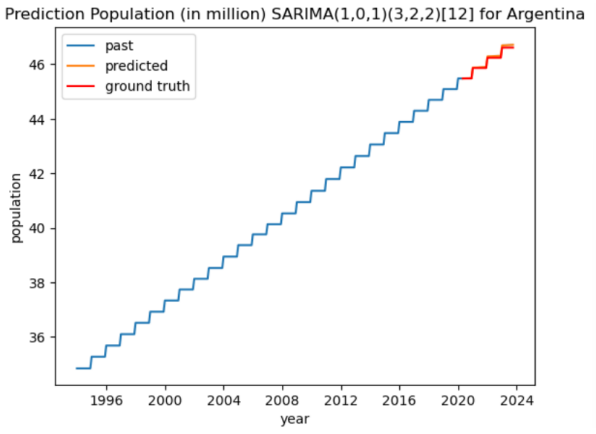
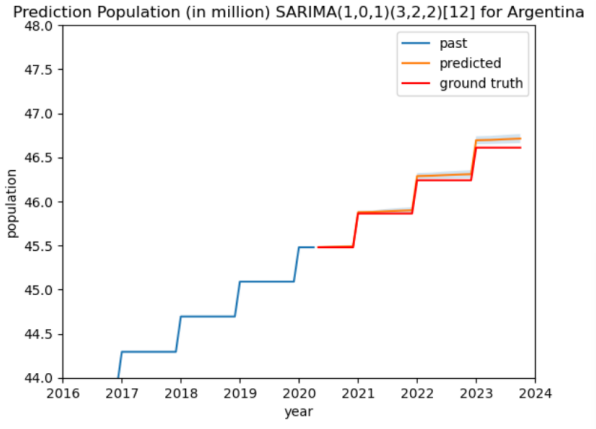


Figure 4.49: Zoom on the forecasting period



The MSE score is : 0.0032204

Forecasting China population

Figure 4.50: Best orders selection with auto-arima

```
Performing stepwise search to minimize aic
 ARIMA(2,0,2)(1,2,1)[12]             : AIC=inf, Time=8.62 sec
 ARIMA(0,0,0)(0,2,0)[12]             : AIC=799.875, Time=0.12 sec
 ARIMA(1,0,0)(1,2,0)[12]             : AIC=546.766, Time=0.35 sec
 ARIMA(0,0,1)(0,2,1)[12]             : AIC=inf, Time=2.15 sec
 ARIMA(1,0,0)(0,2,0)[12]             : AIC=801.875, Time=0.15 sec
 ARIMA(1,0,0)(2,2,0)[12]             : AIC=426.111, Time=1.51 sec
 ARIMA(1,0,0)(3,2,0)[12]             : AIC=409.897, Time=2.23 sec
 ARIMA(1,0,0)(3,2,1)[12]             : AIC=inf, Time=14.46 sec
 ARIMA(1,0,0)(2,2,1)[12]             : AIC=inf, Time=6.38 sec
 ARIMA(0,0,0)(3,2,0)[12]             : AIC=407.897, Time=1.34 sec
 ARIMA(0,0,0)(2,2,0)[12]             : AIC=424.111, Time=0.86 sec
 ARIMA(0,0,0)(3,2,1)[12]             : AIC=inf, Time=13.36 sec
 ARIMA(0,0,0)(2,2,1)[12]             : AIC=inf, Time=4.64 sec
 ARIMA(0,0,1)(3,2,0)[12]             : AIC=409.897, Time=1.78 sec
 ARIMA(1,0,1)(3,2,0)[12]             : AIC=411.897, Time=3.28 sec
 ARIMA(0,0,0)(3,2,0)[12] intercept   : AIC=409.896, Time=3.87 sec

Best model:  ARIMA(0,0,0)(3,2,0)[12]
Total fit time: 65.108 seconds
```

The best orders selected by the package are SARIMA$(0,0,0)(3,2,0)_12$

Figure 4.51: Auto-Arima summary

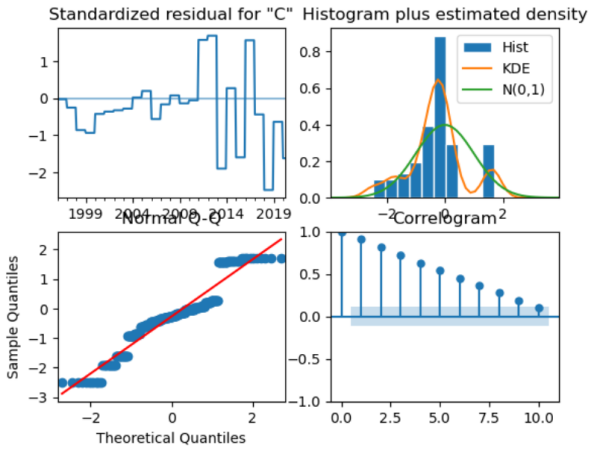
SARIMAX Results

Dep. Variable:	y	**No. Observations:**	303
Model:	SARIMAX(3, 2, 0, 12)	**Log Likelihood**	-199.949
Date:	Sat, 11 Nov 2023	**AIC**	407.897
Time:	22:46:44	**BIC**	422.422
Sample:	02-01-1995	**HQIC**	413.724
	- 04-01-2020		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.S.L12	-1.4046	0.034	-41.005	0.000	-1.472	-1.337
ar.S.L24	-0.9428	0.042	-22.680	0.000	-1.024	-0.861
ar.S.L36	-0.2759	0.022	-12.444	0.000	-0.319	-0.232
sigma2	0.2248	0.003	68.966	0.000	0.218	0.231

Ljung-Box (L1) (Q):	0.00	**Jarque-Bera (JB):**	55242.06
Prob(Q):	1.00	**Prob(JB):**	0.00
Heteroskedasticity (H):	33.20	**Skew:**	-2.03
Prob(H) (two-sided):	0.00	**Kurtosis:**	71.82

Figure 4.52: Check residuals for SARIMA$(0,0,0)(3,2,0)_12$



The Q-Q plot looks like its heavy tailed. This means that, compared with the normal distribution, there is much more data at the extremes than in the center of the distribution.

Furthermore, the density histogram tells us that the green line shows that the normal distribution is far from the KDE distribution (i.e. orange line). This suggests that our ARIMA model can be improved.

Otherwise, according to the correlogram, there is a significant correlation in the residuals; our model doesn't capture all the information in our data.

Figure 4.53: Forecasting population of China

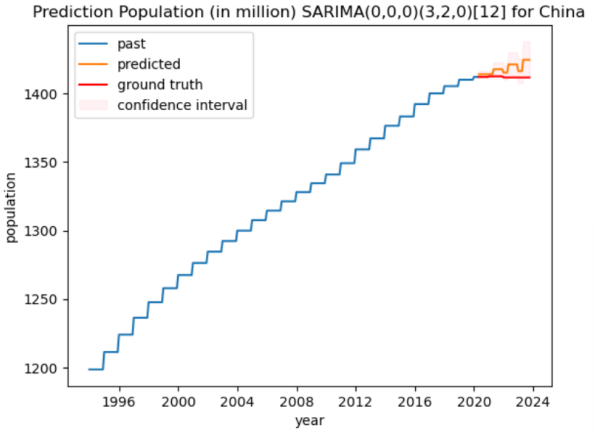
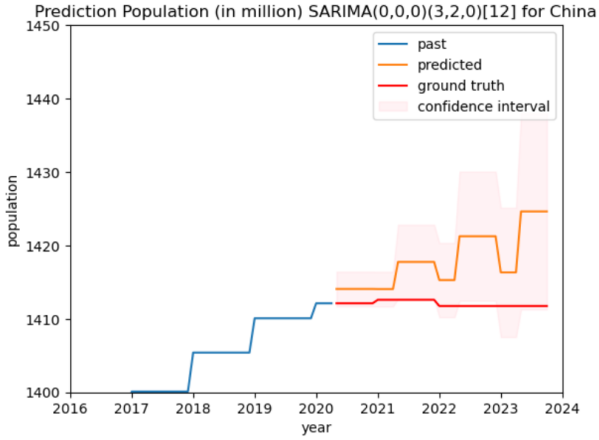


Figure 4.54: Zoom on the forecasting period



The MSE score is : 50.08134

Forecasting Norway population

Figure 4.55: Best orders selection with auto-arima

```
Performing stepwise search to minimize aic
 ARIMA(2,0,2)(1,2,1)[12]             : AIC=-2593.597, Time=1.26 sec
 ARIMA(0,0,0)(0,2,0)[12]             : AIC=-2414.156, Time=0.19 sec
 ARIMA(1,0,0)(1,2,0)[12]             : AIC=-2501.959, Time=1.34 sec
 ARIMA(0,0,1)(0,2,1)[12]             : AIC=-2410.157, Time=0.61 sec
 ARIMA(2,0,2)(0,2,1)[12]             : AIC=inf, Time=7.08 sec
 ARIMA(2,0,2)(1,2,0)[12]             : AIC=-2495.959, Time=1.70 sec
 ARIMA(2,0,2)(2,2,1)[12]             : AIC=-2637.858, Time=20.43 sec
 ARIMA(2,0,2)(2,2,0)[12]             : AIC=-2604.647, Time=2.79 sec
 ARIMA(2,0,2)(3,2,1)[12]             : AIC=-2654.965, Time=21.84 sec
 ARIMA(2,0,2)(3,2,0)[12]             : AIC=-2617.754, Time=4.87 sec
 ARIMA(2,0,2)(3,2,2)[12]             : AIC=-2650.100, Time=19.83 sec
 ARIMA(2,0,2)(2,2,2)[12]             : AIC=-2651.708, Time=17.95 sec
 ARIMA(1,0,2)(3,2,1)[12]             : AIC=-2657.095, Time=23.80 sec
 ARIMA(1,0,2)(2,2,1)[12]             : AIC=-2639.858, Time=8.06 sec
 ARIMA(1,0,2)(3,2,0)[12]             : AIC=-2619.754, Time=4.63 sec
 ARIMA(1,0,2)(3,2,2)[12]             : AIC=-2652.102, Time=32.27 sec
 ARIMA(1,0,2)(2,2,0)[12]             : AIC=-2606.647, Time=2.25 sec
 ARIMA(1,0,2)(2,2,2)[12]             : AIC=-2653.247, Time=7.16 sec
 ARIMA(0,0,2)(3,2,1)[12]             : AIC=-2659.090, Time=23.87 sec
 ARIMA(0,0,2)(2,2,1)[12]             : AIC=-2641.858, Time=4.64 sec
 ARIMA(0,0,2)(3,2,0)[12]             : AIC=-2621.754, Time=2.93 sec
 ARIMA(0,0,2)(3,2,2)[12]             : AIC=-2654.101, Time=25.74 sec
 ARIMA(0,0,2)(2,2,0)[12]             : AIC=-2608.647, Time=1.79 sec
 ARIMA(0,0,2)(2,2,2)[12]             : AIC=-2654.770, Time=17.50 sec
 ARIMA(0,0,1)(3,2,1)[12]             : AIC=-2647.207, Time=7.24 sec
 ARIMA(0,0,3)(3,2,1)[12]             : AIC=-2643.208, Time=11.85 sec
 ARIMA(1,0,1)(3,2,1)[12]             : AIC=-2645.207, Time=9.54 sec
 ARIMA(1,0,3)(3,2,1)[12]             : AIC=-2641.208, Time=15.07 sec
 ARIMA(0,0,2)(3,2,1)[12] intercept   : AIC=-2643.263, Time=16.49 sec

Best model:  ARIMA(0,0,2)(3,2,1)[12]
Total fit time: 314.774 seconds
```

The best orders selected by the package are SARIMA$(0,0,2)(3,2,1)_12$

Figure 4.56: Auto-Arima summary

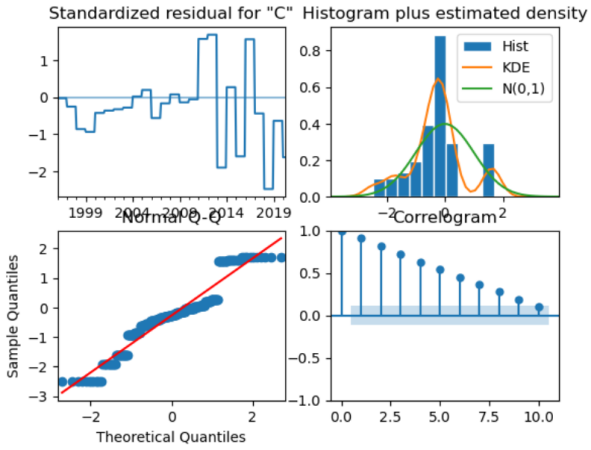
SARIMAX Results

Dep. Variable:	y	**No. Observations:**	303
Model:	SARIMAX(0, 0, 2)x(3, 2, [1], 12)	**Log Likelihood**	1336.545
Date:	Sat, 11 Nov 2023	**AIC**	-2659.090
Time:	22:52:03	**BIC**	-2633.671
Sample:	02-01-1995	**HQIC**	-2648.893
	- 04-01-2020		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-5.234e-06	4.06e-08	-128.985	0.000	-5.31e-06	-5.15e-06
ma.L2	-5.254e-06	4.11e-08	-127.955	0.000	-5.33e-06	-5.17e-06
ar.S.L12	-0.6114	0.045	-13.536	0.000	-0.700	-0.523
ar.S.L24	-0.4213	0.048	-8.809	0.000	-0.515	-0.328
ar.S.L36	-0.0233	0.034	-0.692	0.489	-0.089	0.043
ma.S.L12	-0.6936	0.042	-16.626	0.000	-0.775	-0.612
sigma2	3.682e-06	1e-07	36.780	0.000	3.49e-06	3.88e-06

Ljung-Box (L1) (Q):	0.00	**Jarque-Bera (JB):**	12413.41
Prob(Q):	1.00	**Prob(JB):**	0.00
Heteroskedasticity (H):	0.95	**Skew:**	-1.22
Prob(H) (two-sided):	0.80	**Kurtosis:**	35.59

Figure 4.57: Check residuals for SARIMA$(0,0,2)(3,2,1)_12$



We can draw the same conclusion as for the Chinese time series.

The Q-Q plot looks heavy-tailed and the density histogram is far from the KDE so there is room for improvement.

Otherwise, according to the correlogram, there is a significant correlation in the residuals, suggesting that our model is not a good fit.

Figure 4.58: Forecasting population of Norway

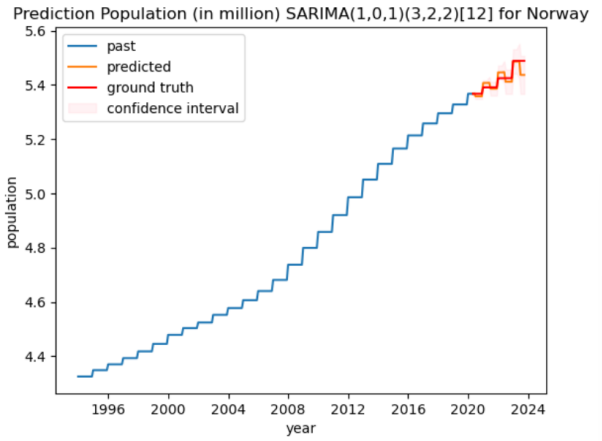
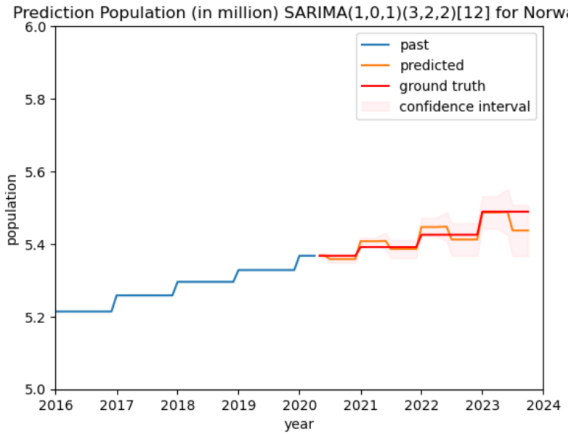


Figure 4.59: Zoom on the forecasting period



The MSE score is : 0.0004015

4.9 Discussion about the results

Generally, we obtain good results for the following countries: Argentina, France, New Zealand and Norway. The models are quite good at predicting the population outside the training dataset.

Conversely, we do not achieve high accuracy for China. Indeed, from 2019 onwards, China's evolution remains stable. This seems strange. The question then is: Has China been more affected by COVID than other countries in the world, or is this simply due to a data quality problem (the data provider may not have had access to China's population data for the COVID period)? China seems to be impacted by COVID because the population remains stable during this period. That's why the forecast of the SARIMA model did not fit well with the ground truth. As becomes evident, we have a structual break after 2019. Hence, the SARIMA model is not able to capture rare events such as COVID, leading to extreme values. In fact, past data are not suitable to explain the future when extreme behaviors occur.

4.10 ARIMAX on the Chinese Population

In the SARIMA part, we didn't manage to accurately forecast the population of China during COVID.

Hence, the goal is to see if the use of exogenous data such as mortality rate, life expectancy

birth rate, etc. to forecast China population will help to forecast more precisely the china population during the COVID period

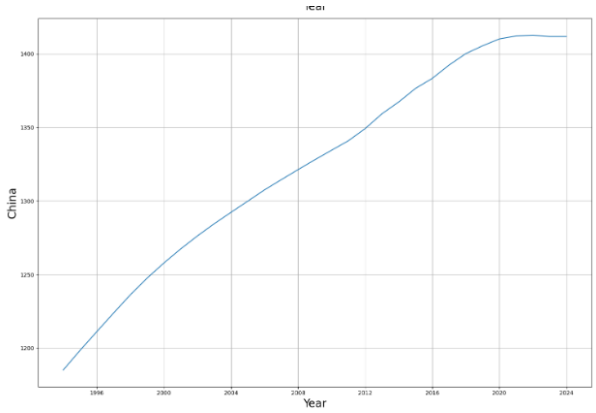
Given that all the exogenous data has an annual frequency, we have to resample our monthly population data to annual population data.

4.10.1 Data Processing

According the Data Processing strategy for exogenous data, we use the same Data Processing as mentioned previously (i.e. data imputation, removing outliers, etc.)

Annual China population evolution

Figure 4.60: Annual China population evolution



Here we can see the break change in the curve at the year 2020. Indeed, the series remains stable after 2020.

Overall, the growth of the Chinese population has a positive trend until 2020.

In our analysis, we will explore whether transitioning to ARIMAX allows us to better accommodate this observed structural break change.

Exogenous Data

We collect data on demographic indicators from a World Bank data source for several countries. In our case, we will focus on demographic indicators for China.

The collected indicators are:

- China birth rate

- China fertility rate
- China life expectancy
- China mortality rate
- China survival up to 65 years old

Figure 4.61: Plot of Birth rate, Fertility rate for women

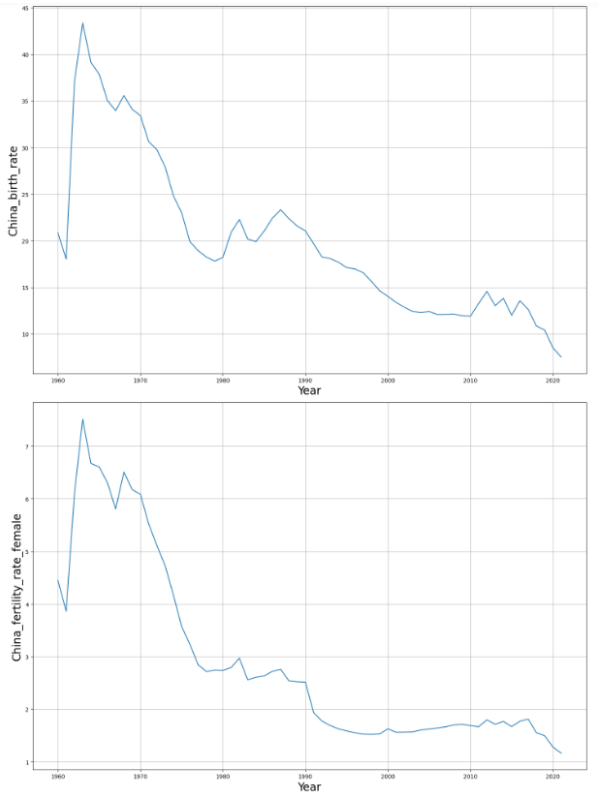


Figure 4.62: Plot of China Birth rate and Fertility rate for women

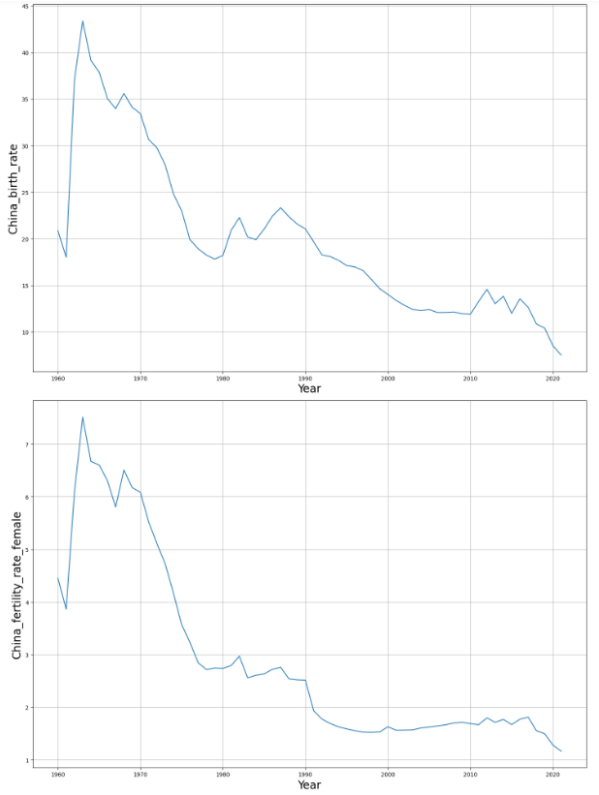


Figure 4.63: Plot of China Mortality rate and life expectancy

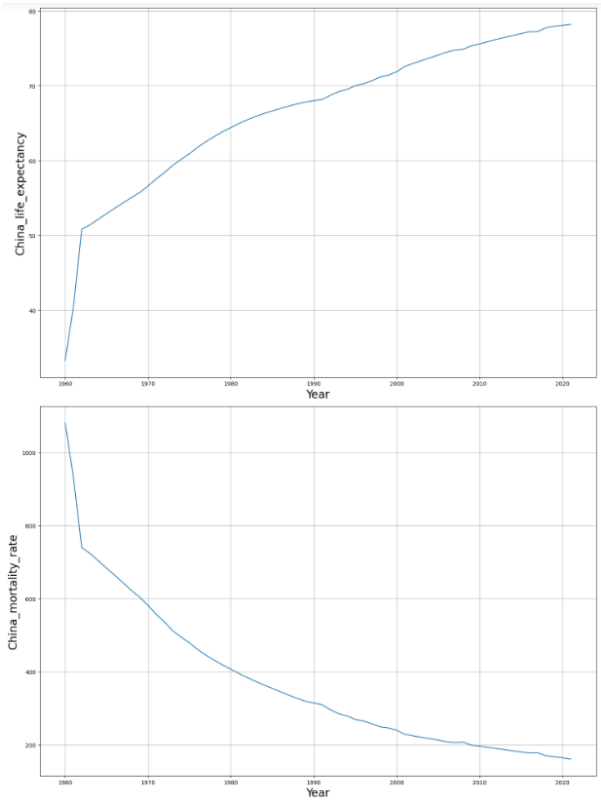
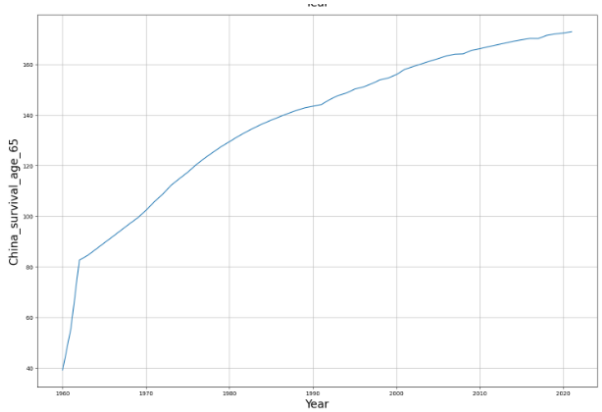


Figure 4.64: Plot of China survival age up to 65 years old



We observe a significant decline in both the birth rate and fertility rate since 2016 and during the COVID period. This can explain why the Chinese population remains stable.

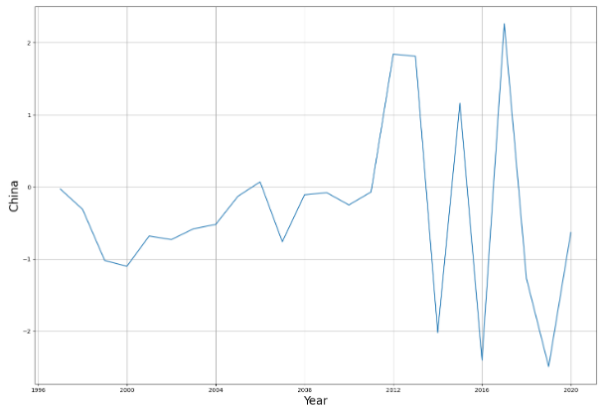
Conversely, we also note a decline in the growth of life expectancy over time, particularly during the COVID period, in contrast to the patterns observed in previous years.

4.10.2 Forecasting the China population with ARIMAX

To fine-tune the ARIMAX model to predict the Chinese annual population, we apply the same methodology as that used to forecast the monthly population of other countries (Argentina, France, Norway, etc.). Hence, we took the second difference to make the annual population of China stationary and then we used the auto-arima library to find the best parameters for our ARIMA model. Finally we include in our ARIMA model the exogenous variables mentioned above.

Take the 2nd difference

Figure 4.65: Plot 2nd difference China annual population



After taking the second difference the time series seems to be stationary. Let's verify that with ADF test.

Figure 4.66: ADF test

```
_____________________________
REGRESSION : CONSTANT AND TREND
ANNUAL CHINA

_____________________________
p_value:  6.317470172299791e-06
Reject the null hypothesis. The time series is likely stationary.
```

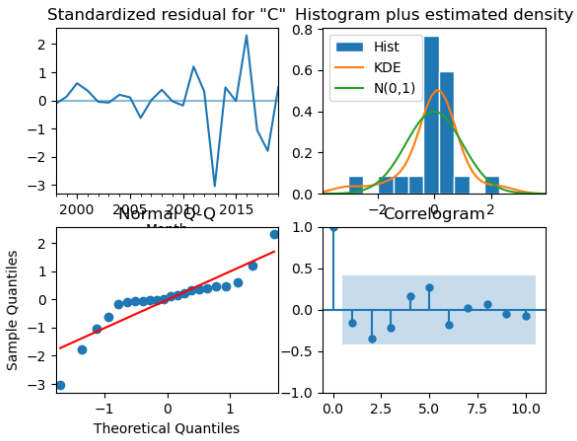
Figure 4.67: Best orders selection with auto-arima

```
Performing stepwise search to minimize aic
 ARIMA(1,2,1)(0,0,0)[0] intercept   : AIC=inf, Time=0.08 sec
 ARIMA(0,2,0)(0,0,0)[0] intercept   : AIC=121.601, Time=0.01 sec
 ARIMA(1,2,0)(0,0,0)[0] intercept   : AIC=104.085, Time=0.01 sec
 ARIMA(0,2,1)(0,0,0)[0] intercept   : AIC=inf, Time=0.06 sec
 ARIMA(0,2,0)(0,0,0)[0]             : AIC=119.618, Time=0.01 sec
 ARIMA(2,2,0)(0,0,0)[0] intercept   : AIC=96.223, Time=0.02 sec
 ARIMA(3,2,0)(0,0,0)[0] intercept   : AIC=97.427, Time=0.03 sec
 ARIMA(2,2,1)(0,0,0)[0] intercept   : AIC=inf, Time=0.08 sec
 ARIMA(3,2,1)(0,0,0)[0] intercept   : AIC=inf, Time=0.12 sec
 ARIMA(2,2,0)(0,0,0)[0]             : AIC=94.233, Time=0.02 sec
 ARIMA(1,2,0)(0,0,0)[0]             : AIC=102.115, Time=0.02 sec
 ARIMA(3,2,0)(0,0,0)[0]             : AIC=95.437, Time=0.02 sec
 ARIMA(2,2,1)(0,0,0)[0]             : AIC=inf, Time=0.07 sec
 ARIMA(1,2,1)(0,0,0)[0]             : AIC=inf, Time=0.19 sec
 ARIMA(3,2,1)(0,0,0)[0]             : AIC=inf, Time=0.08 sec

Best model:  ARIMA(2,2,0)(0,0,0)[0]
Total fit time: 0.820 seconds
```

The best orders selected by the package are ARIMAX$(0, 0, 2)$

Figure 4.68: Check residuals for SARIMAX$(0, 0, 2)$

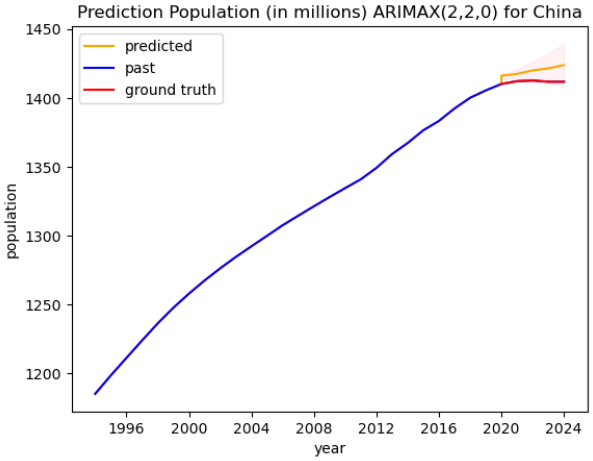
```
##########################################
Check residuals of ARIMA for CHINA
##########################################

Mean of Residuals: -0.05268491898061863
We fail to reject H0: The residuals are not independently distributed.
No significant serial correlation in residuals.
```



The Q-Q plot looks heavy tailed. Furthermore, according to the density histogram, we can see that the green line showing a normal distribution is not too far from the KDE distribution (i.e. orange line. This suggest that our ARIMA model fit well with the data.

Otherwise, according to the correlogram, there is no significant correlation in the residuals, our model seems to be a good fit for our data.

Figure 4.69: Forecasting annual population of China



The MSE score is : 71.6134

According to the MSE, the performance of the ARIMAX model doesn't look very good: on average, we have a 70 million gap between the true population value in China and the value we predict for the COVID year. In fact, adding exogenous variables does not seem to improve the performance of the ARIMA model. The results are also worse than the ARIMA model we used to predict China's monthly population. The granularity of the data seems to have an impact on the model's prediction performance.

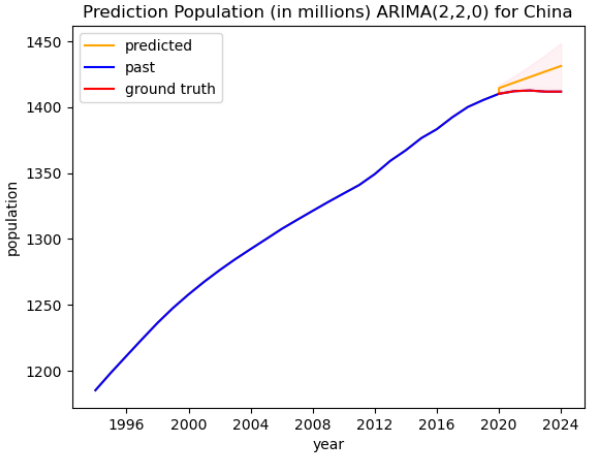
4.10.3 Forecasting the China population with ARIMA

The final step is to apply an ARIMA model on annual China population in order to compare the performance of ARIMAX vs. ARIMA model (without exogenous variables) . This allow us to compare the importance of exogenous variables for forecasting population when extreme event occurs such as COVID.

To do so we will fit an ARIMA model with the same order than the ARIMAX model

As we fit the same ARIMA model than previously we don't analyze the residuals of our model and dive directly to the forecasting process.

Figure 4.70: Forecasting annual population of China



Hence the orders are : ARIMA$(2, 2, 0)$

The MSE score is : 155.7852

4.10.4 Discussion: ARIMAX vs. ARIMA

According to the MSE score of both models (ARIMAX vs. ARIMA), the ARIMAX (doesn't take into account the exogenous variables) seems to perform better given its lower MSE.

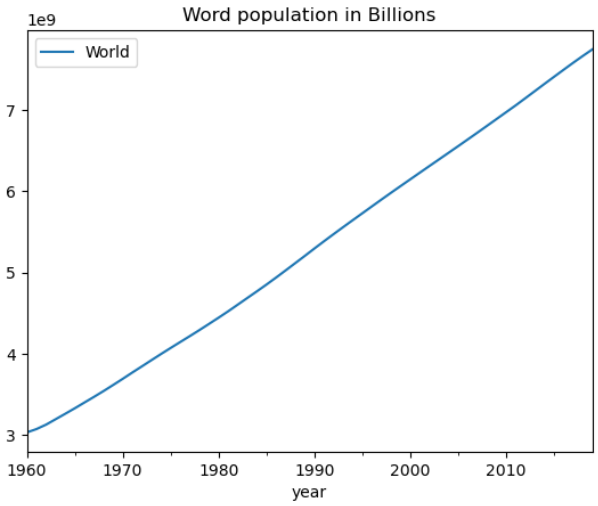
4.11 World population forecast

Now we will try to forecast the World population during the COVID period.

To do so, we use annual world population data collected from world bank data platform.

We apply the methodology to prepare our date and make our series stationary.

Figure 4.71: World population evolution



We can see that the time series have a positive trend and seems do not have to any seasonal pattern.

We fit a ARIMA model to select best orders thanks to auto-arima package The order of our best ARIMA model are : ARIMA$(4, 2, 1)$

Figure 4.72: Forecasting population of World population

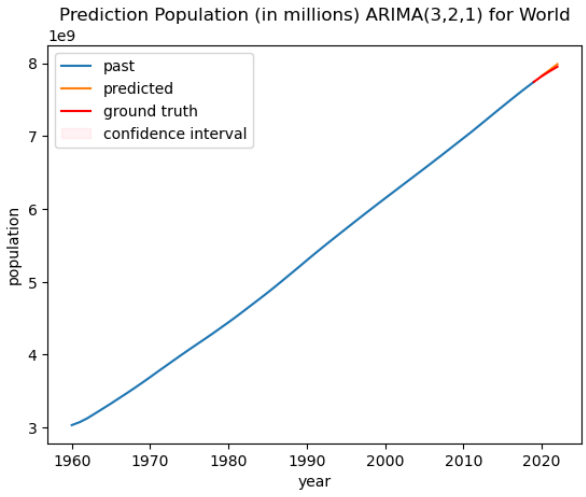


Figure 4.73: Zoom Forecasting period

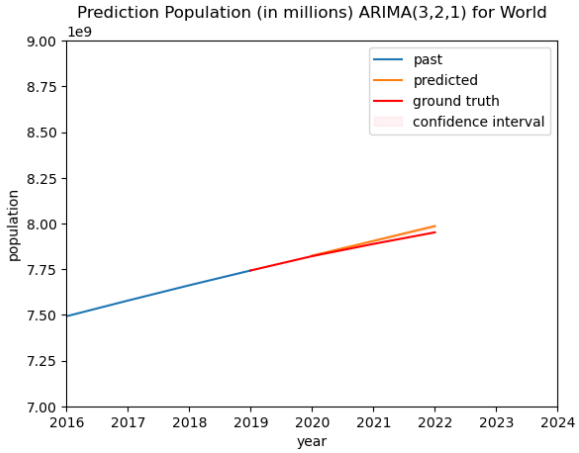


Figure 4.74: Ground truth vs. Predictions

	World	predicted_mean
0	7820963775.0	7.823624e+09
1	7888161297.0	7.904580e+09
2	7951149546.0	7.985538e+09

The MSE score is : 486402302130827.7

Even if is the forecast plot, our prediction seems to be not that bad, a small deviation from the ground truth on the graph actually corresponds to a huge deviation in reality, as we work with large numbers (10e9).

4.12 Hybrid LSTM ARIMA/SARIMA modelling

To go further in our analysis and allow our model to capture the non-linear relationship of our dependent variables on population growth, we now proceed to constructing the LSTM Neural Network to correct for the econometric models' residuals.

At this point of our analysis we decided to remove China from the Neural Network as the residuals from the econometric model for this particular country made all data points look like outliers and prevented the neural network from converging. We also standardized all of the model inputs using the RobustScaler from sklearn. Choosing the RobustScaler

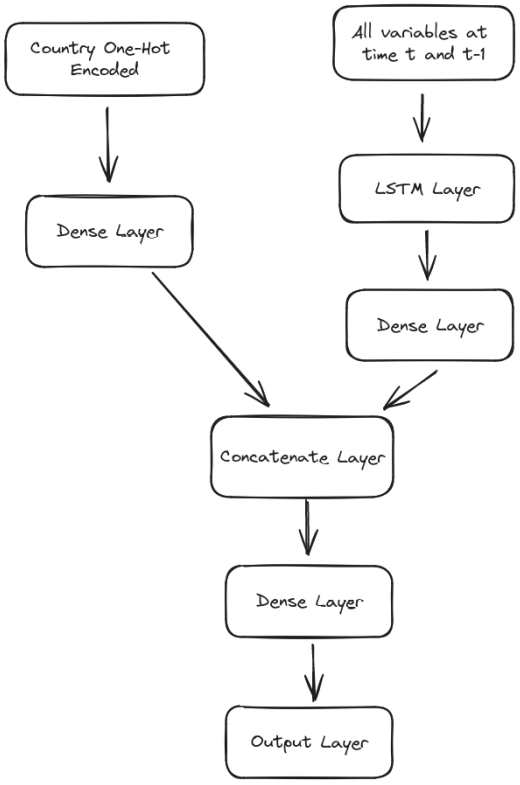
instead of the StandardScaler is justified by the small size of our dataset and the fact that we are studying multiple countries which could affect the distribution of our variables. Finally, the residuals from the Econometric model were rescaled by multiplying them by 100, otherwise the values were too small and the precision of the neural network did not allowed it to converge.

To increase the size of the dataset used for the neural network, we decided not to make one model for each country but rather to include country-fixed effects by One-Hot Encoding the countries' names in a separate branch. We then used a mix of activation function from gelu, tanh and linear as all of those layers are able to analyze both positive and negative values.

The training data is composed of annual data from Argentina, France, New Zealand and Norway from which we removed the latest 5 observations for each country making it an 80/20 split.

The final LSTM model is thus a two-branch neural network, as presented in the following figure:

Figure 4.75: LSTM model



The model managed to fit the residuals with high precision, thus improving upon the econometric model performance, as can be seen from the following graphs depicting the underlying econometric model residuals (our target) and the neural network predictions:

Figure 4.76: LSTM Predictions for Argentina

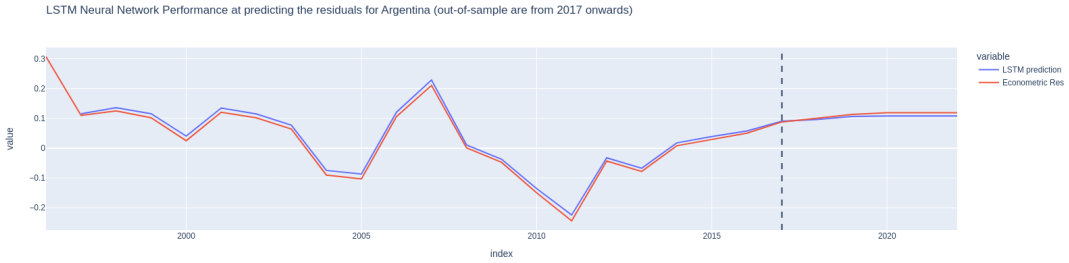


Figure 4.77: LSTM Predictions for France

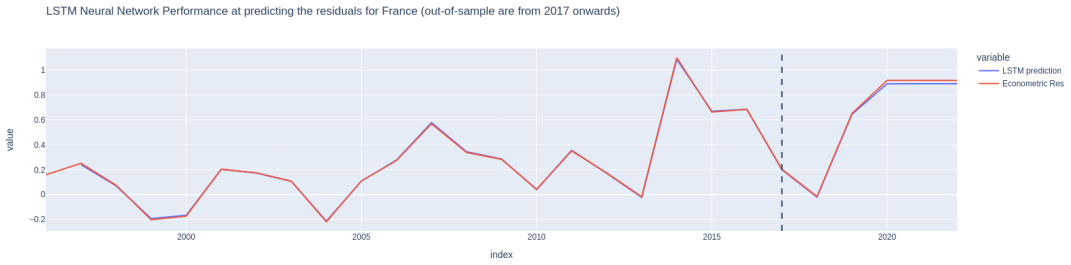


Figure 4.78: LSTM Predictions for New Zealand

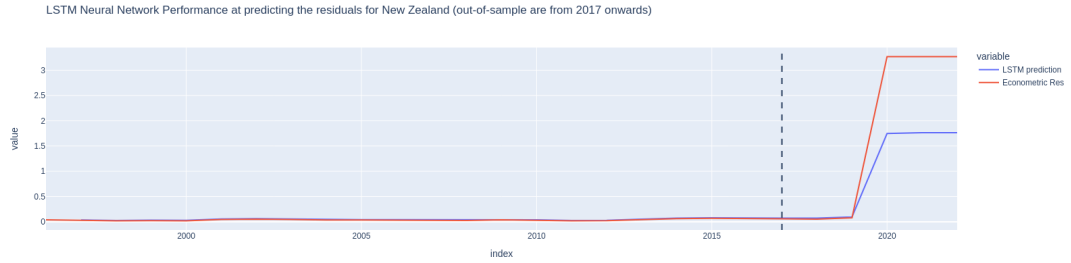
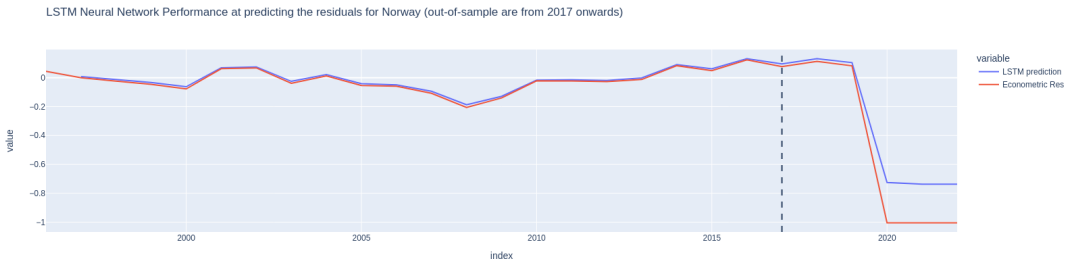


Figure 4.79: LSTM Predictions for Norway



Some of the latest predictions (2020 onward) made from New Zealand and Norway seem to have not fully captured the residuals. This difference can be partially explained by us missing some values for our latest observations and filling in NaN values using the fill-forward method, making the information used for the prediction outdated.

To see full python code of our analysis please refer to our GitHub repository with the following link QMF Project

5 Population Gap Analysis

Monthly Gap Analysis

01 June 2023 (in million hab)						
	China	Argentina	France	New Zealand	Norway	World*
Forecast	1424.63	46.702	68.12	5.201	5.4889	7904
Ground truth	1411.75	46.609	68.11	5.223	5.4883	7888
Difference	12.878	0.093	0.007	-0.022	0.0006	16
Percentage Error (%)	0.91	0.199	0.0103	0.421	0.011	0.2

Table 5.1: Population gap between the real data and the forecast in June 2023

* Note: World in June 2022. As we are restricted by the world population data for 2023 (because we are currently in 2023), so we have forecasted the world population until 2022 and compared our prediction with the real world population for 2022.

Based on the comparison between the projected and actual population figures, where the percentage error stays under 1% for each evaluated country, we can draw a conclusion that the forecasted populations are accurate and reflect the real populations of these countries, with the notable exception of China. This indicates that our SARIMA model is generally effective in predicting future population values.

MSE score performance

	Argentina	China	France	New Zealand	Norway	World
MSE	0.0032	50.0813	0.00060	0.00024	0.00040	4.864e9

Table 5.2: Forecast prediction performance

The Mean Squared Error (MSE) quantifies the average squared difference between predicted and true values, serving as a measure of the predictive model's accuracy. We employ MSE in our analysis to assess the precision of the predictions concerning the true values.

Given that a lower MSE means lower error and better accuracy in predictions, we can conclude that the predictions for New Zealand, Norway, France, and Argentina are accurate and precise and reflect the actual values. However, that's not the case for China where the MSE score is 50.08, signaling the larger error.

That's we apply ARIMAX on annual China population to improve the performance of our model by include exogenous variables.

ARIMAX vs. ARIMA performance on annual China population

	ARIMAX	ARIMA
MSE	71.631	155.785

Table 5.3: ARIMA vs. ARIMAX performance prediction

According to the MSE score of both models (ARIMAX vs. ARIMA), the ARIMA (that does not take into account the exogenous variables) seems to be forecast the population in a more accurate way. Indeed, the MSE score of the ARIMAX model (71.6314) is two time lower than the MSE score of the ARIMA model (155.7852). This suggest that exogenous variables are really useful for forecasting population. This finding are consistent with the literature review.

6 Conclusion

The significance of data granularity in forecasting population accurately is evident. Finer granularity enhances the model's predictive performance. For instance, our model's Mean Squared Error (MSE) for monthly data forecasting of China's population is lower (50.08) compared to using annual data (71.63).

The impact of COVID-19 on different countries varies. China, for example, exhibited stable population figures during the pandemic, reflecting a broader trend in populous Asian countries, which experienced more pronounced declines. This led to poorer out-of-sample performance for China in our model due to the pandemic's impact.

Moreover, incorporating exogenous variables proves beneficial for more accurate out-of-sample population predictions.

In summary, while using historical demographic trends to predict future trends offers valuable insights into population growth mechanisms, it has limitations. These include the inability to factor in structural shifts or external influences on demographic trends, potentially leading to missed significant trend changes or structural transformations in the future.

References

Adelman, I. (1963). An econometric analysis of population growth.

Booth, H. (2006). Demographic forecasting: 1980 to 2005 in review.

Patricia E Beeson, David N DeJong, W. T. (2001). Population growth in u.s counties, 1840-1990.