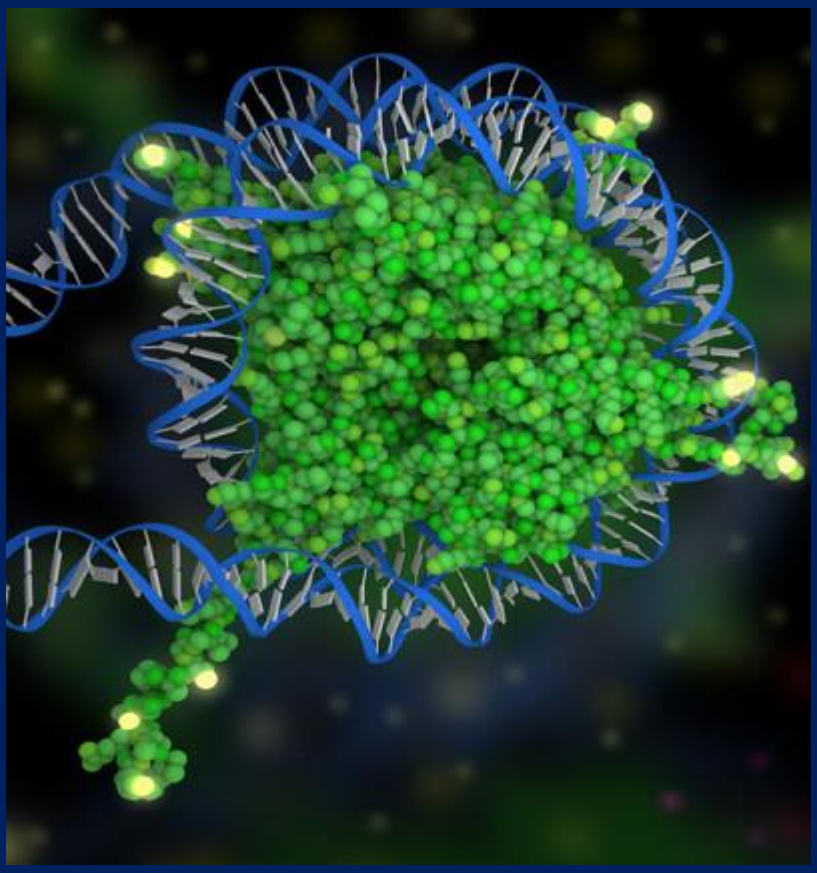


THE FUTURE OF CANCER

Analysis Report

Table of Contents

Executive Summary

All over the world, cancer has become a disease that seriously affects humans and wears people out financially, physically, and mentally. In almost every country in the world, there is a very serious increase in cancer cases as a result of the effects of the increasing population and changing human habits. Modeling the course of increasing cancer cases in an accurate and qualified manner will contribute positively to planning accordingly, taking the necessary precautions, and eventually reducing the rate of the increase in cancer cases.

From this point of view, as "The Healthy World Research Group", this analysis report has been prepared at the request of the Canadian Ministry of Health to help determine the public policy concerning cancer, and to take it into account in the government budget planning studies.

In our study, we tried to make sense of the course by making future predictions for 3 cancer types until 2025, using the number of cancer-related deaths reported across Canada between 1990 and 2019. When choosing these 3 cancer types, which are lung cancer, colon cancer, and pancreatic cancer, we took into account that the related cancer types are the most common. We analyzed the relevant data set using the Python program's time series analysis method.

Tracheal, Bronchus, and Lung Cancer: The number of reported deaths due to this cancer type - which were the most common types of cancer, was 15,026 in 1990. It reached 24,052 deaths in 2019. These figures correspond to approximately 27 percent of all cancer deaths. We anticipate a slight increase in the related cancer type by 2025.

Colon and Rectum Cancer: While the number of reported deaths due to colon cancer was 6,315 in 1990, it reached 11,616 in 2019, which shows that 13 percent of deaths due to cancer are related to this type of cancer. In our estimations, we expect a significant increase in colon cancer by 2025.

Pancreatic Cancer: In deaths related to this type of cancer, we saw those 2,775 deaths from 1990 reached 6,055 in 2019, which corresponds to 7 percent of all cancer deaths in proportion. By 2025, we can say that there will be a serious increase in pancreatic cancer. However, it should not go without saying that this type of cancer is the type of cancer that increases most proportionally compared to other types of cancer.

As a result, we can easily say that the remarkable increases in cancer cases will continue in the coming years, as can be seen in these 3 types of cancer. From this point of view, we think that to fight cancer more effectively, the public budget should be increased, health personnel planning should be reviewed, early diagnosis and diagnosis practices should be expanded, and studies should be carried out to raise public awareness.

The Problem

The Business Goal

The Healthy World Research Group is a private company founded in Canada in 2010 by a private initiative. The company in question provides support services to public institutions and organizations, pharmaceutical companies, and various private companies with various data in the field of health in line with their demands. The firm includes academics, independent researchers, senior healthcare professionals, and managers. This study was requested by the Canadian Ministry of Health to determine the public policies in the relevant field and to use the government budget effectively by making a prediction model for future cancer cases based on the data of the past years. In the period after people get cancer, the expenditures that have to be made by both individuals and the government are extremely expensive. The analysis in this study will be used by the relevant departments of the Canadian Government within the scope of policies for cancer-preventive measures and early detection, and a more effective public budget management will be realized. As a result of the action plan after this study, social health, peace, and a more livable economic order are envisaged.

The Data Analysis Goal

This analysis will be descriptive because we are trying to describe/interpret the past. In this case, a time frame of 30 years from 1990-2019. This is useful because it allows us to learn from past behaviors/trends, and understand how they may or may not influence future outcomes. We look at different cancer types and the most fatal of them (number of deaths recorded) to try and make sense of it. This analytics is forward-looking because our goal is to make a recommendation to the Canadian Ministry of Health and be able to infer which disease will be the most prevalent in the future. This will allow the Canadian government to adequately prepare and prevent a lot of deaths due to cancer in the future. It will also advise where and when to funnel health research grants in the near and distant future. The main outcome variables are the most frequent type of cancer, deaths recorded by type of cancer, and future deaths predicted.

About Data

The dataset for the analysis was retrieved from 'Our World in Data' (https://ourworldindata.org/cancer) and it is sourced from the Institute for Health Metrics and Evaluation, Global Burden of Disease (2019). It contains 6840 records along with 32 variables. Each record indicates the total number of deaths in a specific year, in a specific country, attributed to the range of different cancers. This includes 29 types of cancer deaths from 228 countries for 30 years which is from 1990 to 2019. Within this data, 3 types of cancer deaths in Canada were selected to perform our analysis.

	Entity	Year	Deaths - Tracheal	Pancreatic cancer	Colon and rectum cancer
2	Canada	1990	15026	2775	6315
3	Canada	1991	15297	2840	6375
4	Canada	1992	15573	2902	6444
5	Canada	1993	15994	2996	6574
6	Canada	1994	16143	3056	6646
7	Canada	1995	16620	3102	6731
8	Canada	1996	16764	3174	6805
9	Canada	1997	16941	3218	6897

Figure 1: The Dataset

Data Preparation

First, we made the titles of tumor names more understandable. There were no nulls or repetitive data in the data set. We changed our data from year to day, month to year, and moved on to the analysis part. Here, we will conduct our study only on cancer data in Canada. In any of the cancer types, rising or decreasing trends are almost nonexistent.

As a requirement of the study, we investigated which type of cancer we should focus on. When we analyzed 29 types, 3 types of cancer differed in incidence and slightly increasing trend. The first of these is Tracheal, bronchus, and lung cancer, which is the most common and has a rate of 27%. The second is Colon and rectum cancer with a 13% incidence. Finally, it is pancreatic cancer with an incidence of 7% with an increase of 2% from 1990 to 2019. We will take these 3 cancer types from our dataset and analyze them. In addition, the number of deaths from cancer has increased by 75% in 30 years. However, the Canadian population increased by 35% during the relevant period. Based on this, we are likely to encounter more cancer deaths in the future

Data Analysis Solution

3 types of cancer deaths with the highest mortality rate were selected for the analysis of predicting the number of deaths that could occur in the next 3 years.

Selection of Tools and Technology

Python language was selected to perform our analysis as it provides many easy-to-use libraries and tools to perform time series-related forecasting with a few lines of code. In addition to that, the stats library has tools for building models such as ARMA, ARIMA SARIMA, and SARIMAX models, which can help experiment with different models**.** Pandas library was used for data preprocessing purposes such as creating data frames and data manipulations. To create data visualizations and for graphical plotting, the Matplotlib library was used.

Model Selection

SARIMAX (Seasonal Auto-Regressive Integrated Moving Average with eXogenous factors) model was chosen as it is considered an updated version of the ARIMA model. Although our dataset is not seasonal, the SARIMAX model can be used in a non-seasonal way by not setting a value for seasonality.

Analysis

1. **Loading and Visualizing the existing data points**

read_csv() method in the pandas' package was used to read the cancer dataset as a pandas data frame and the cancer deaths from 1990 to 2019 were visualized using a line chart to distinguish patterns, trends, and noise. A non-seasonal upward trend is observed in all 3 cancer death counts.

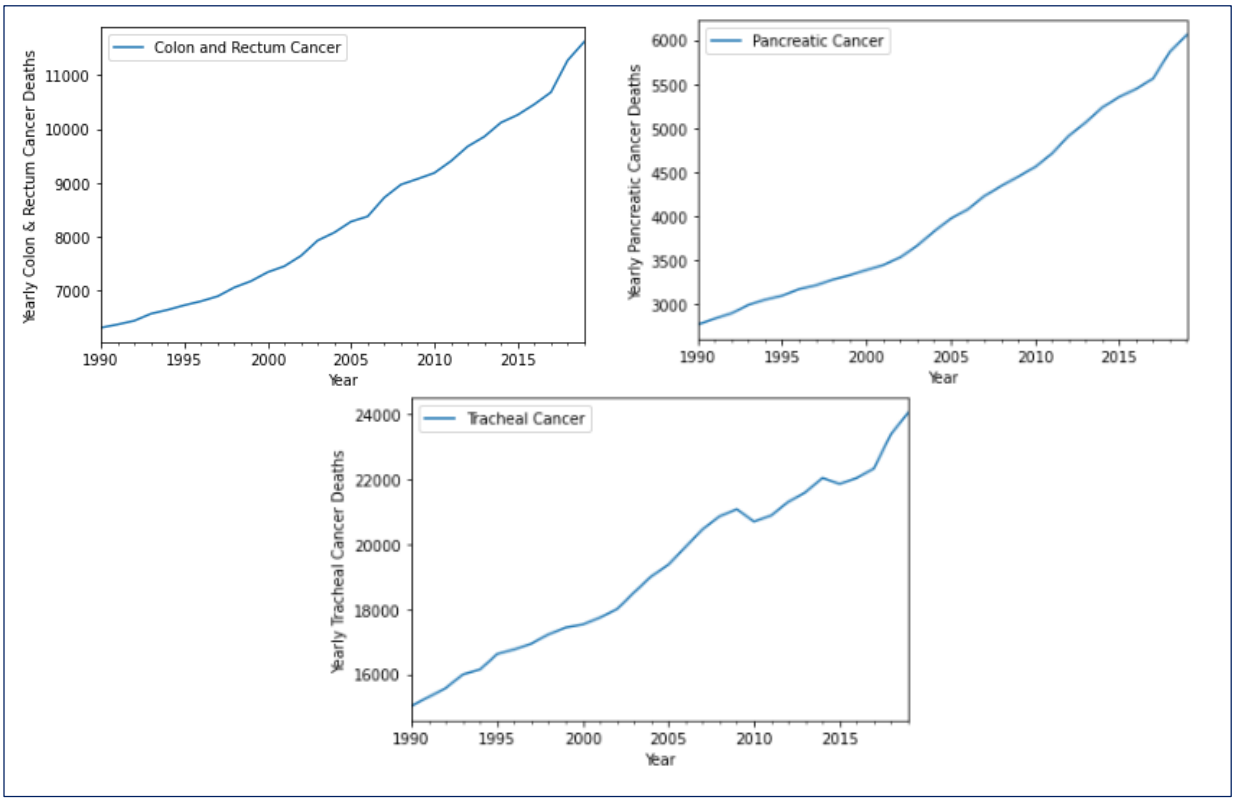


Figure 2: Line Graphs of 3 Types of Yearly Cancer Deaths

2. Defining the d, q, and p parameters

In the SARIMAX model, 3 parameters have been used to help model the major aspects of the time series trend and noise. These parameters are labeled p, d, and q. We have used p=4 (periods to lag for) which will use the four previous periods of our time series in the autoregressive portion of the calculation. It helps adjust the line that is being fitted to forecast the series. Variable d refers to the number of differencing transformations required by the time series to get stationary when the mean and variance are constant over time which makes it easier to predict. Variable q denotes the lag of the error component, where the error component is a part of the time series not explained by trend.

```
# Define the d and q parameters to take any value between 0 and 1
q = d = range(0, 2)

# Define the p parameters to take any value between 0 and 3
p = range(0, 4)

# Generate all different combinations of p, q and q triplets
pdq = list(itertools.product(p, d, q))

print('Examples of parameter combinations for SARIMAX...')
print('SARIMAX: {}'.format(pdq[1]))
print('SARIMAX: {}'.format(pdq[1]))
print('SARIMAX: {}'.format(pdq[2]))
print('SARIMAX: {}'.format(pdq[2]))
```

```
Examples of parameter combinations for SARIMAX...
SARIMAX: (0, 0, 1)
SARIMAX: (0, 0, 1)
SARIMAX: (0, 1, 0)
SARIMAX: (0, 1, 0)
```

Figure 3: Code Snippet of defining the d, q, and p parameters and the output

3. Creating Training and Test Datasets

Splitting data samples for training and testing helps to evaluate model performance and allows the model to generalize well to newer data. Our dataset was divided into 4:1 ratio to validate our model.

```
#Create Training and Test Datasets
train_data = data['1/1/1990':'1/1/2013']
test_data = data['1/1/2014':'1/1/2019']
```

Figure 4: Code Snippet of Assigning Training and Test Datasets

4. Building and fitting the model

After determining the p, d, and q values, the SARIMAX() implementation in the statsmodels package is used to build the model. The Akaike information criterion (AIC) was used as an estimator to determine the relative quality of each combination of p, q & d.

```
warnings.filterwarnings("ignore") # specify to ignore warning messages

AIC = []
SARIMAX_model = []
for param in pdq:
    try:
        mod = sm.tsa.statespace.SARIMAX(train_data,
                                        order=param,
                                        enforce_stationarity=False,
                                        enforce_invertibility=False)

        results = mod.fit()

        print('SARIMAX{} - AIC:{}'.format(param, results.aic), end='\r')
        AIC.append(results.aic)
        SARIMAX_model.append(param)
    except:
        continue

print('The smallest AIC is {} for model SARIMAX{}'.format(min(AIC,default=None),
                                                          SARIMAX_model[AIC.index(min(AIC,default=None))]))
```

```
The smallest AIC is 202.3797199571444 for model SARIMAX(3, 1, 0)
```

Figure 5: Code Snippet of building the SARIMAX Model

The above output suggests that SARIMAX(3, 1, 0) yields the lowest AIC value of 202.379. Therefore this combination is considered to be the optimal option.

5. Fitting the SARIMAX model

```
# Let's fit this model
mod = sm.tsa.statespace.SARIMAX(train_data,
                                order=SARIMAX_model[AIC.index(min(AIC))],
                                enforce_stationarity=False,
                                enforce_invertibility=False)

results = mod.fit()
```

Figure 6: Code Snippet of Fitting the Model

6. Review of Diagnostic plots

A model diagnostics were done to investigate any unusual behavior.

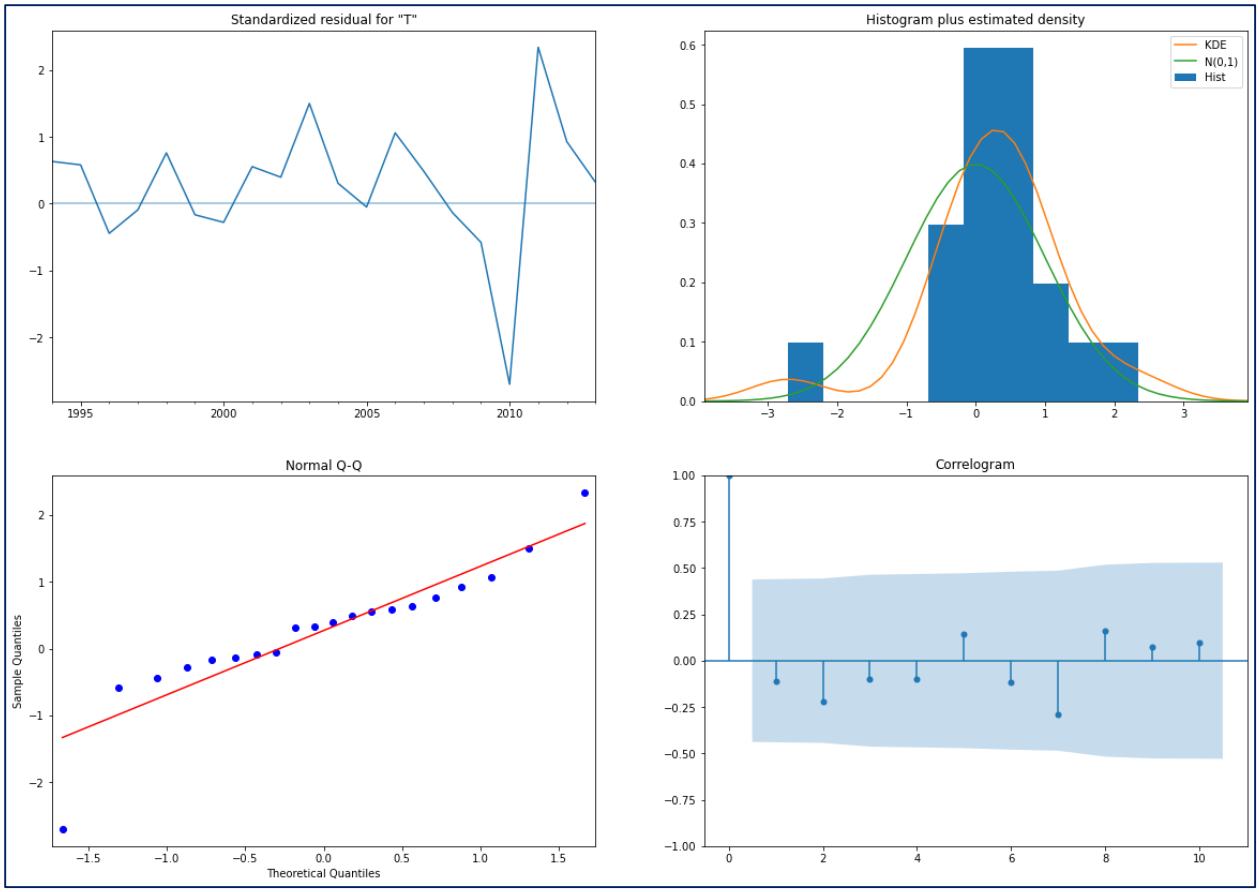


Figure 7: Code Snippet of Diagnostic Review Plots of Tracheal Cancer

Although it is not perfect, however, our model diagnostics suggest that the model residuals are near normally distributed.

7. Forecasting (Step ahead, Dynamic, True)

Step ahead forecasting & dynamic forecasting were done to help understand the accuracy of the forecasts.

```
In [20]: #1-step ahead forecasting implies that each forecasted point is used to predict the following one
         pred0 = results.get_prediction(start='01/01/2012', dynamic=False)
         pred0_ci = pred0.conf_int()
         pred0_ci.head()
```

Out[20]:

	lower Pancreatic Cancer	upper Pancreatic Cancer
2012-01-01	4792.356276	4914.674708
2013-01-01	5036.083421	5158.401853

```
In [21]: #Dynamic Forecasting
         pred1 = results.get_prediction(start='01/01/2012', dynamic=True)
         pred1_ci = pred1.conf_int()
         pred1_ci.head()
```

Out[21]:

	lower Pancreatic Cancer	upper Pancreatic Cancer
2012-01-01	4792.356276	4914.674708
2013-01-01	4873.236836	5118.243703

Figure 8: Output of 1 Step Ahead and Dynamic Forecasting

```
#True Forecasting - Colon
pred2 = results.get_forecast(steps=12)
pred2_ci = pred2.conf_int()
print(pred2.predicted_mean['1/1/2020':'1/1/2025'])
```

```
2020-01-01    11675.905631
2021-01-01    11972.619432
2022-01-01    12280.235390
2023-01-01    12598.715779
2024-01-01    12928.076184
2025-01-01    13268.855699
Freq: AS-JAN, Name: predicted_mean, dtype: float64
```

```
#True Forecasting - Pancreatic and Rectum Cancer
pred2 = results.get_forecast(steps=12)
pred2_ci = pred2.conf_int()
print(pred2.predicted_mean['1/1/2020':'1/1/2025'])
```

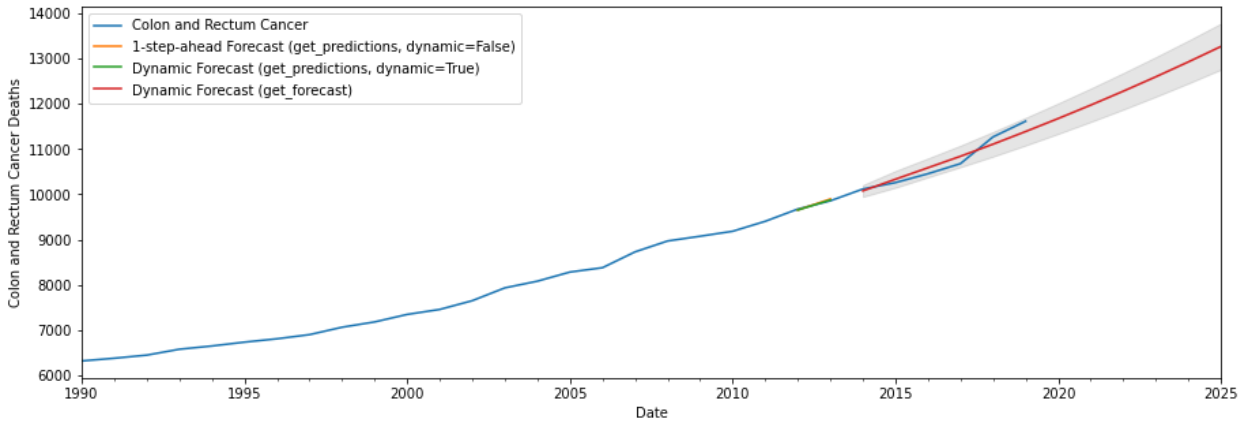
```
2020-01-01    6305.229068
2021-01-01    6499.074323
2022-01-01    6697.227674
2023-01-01    6899.766557
2024-01-01    7106.804499
2025-01-01    7318.440184
Freq: AS-JAN, Name: predicted_mean, dtype: float64
```

```
#True Forecasting - Tracheal
pred2 = results.get_forecast(steps=12)
pred2_ci = pred2.conf_int()
print(pred2.predicted_mean['1/1/2020':'1/1/2025'])
```

```
2020-01-01    22813.889977
2021-01-01    22890.601992
2022-01-01    22954.858100
2023-01-01    23008.201203
2024-01-01    23052.720743
2025-01-01    23089.759134
Freq: AS-JAN, Name: predicted_mean, dtype: float64
```

Figure 9: Outputs of True Forecasting

According to the below line graphs, it is clear the forecasting line is almost lying on the given values for this model. The model has predicted 7318 pancreatic, 1328 colon and rectum, and 23089 tracheal cancer deaths in 2025. Using this same model, we could predict the future death values as well.



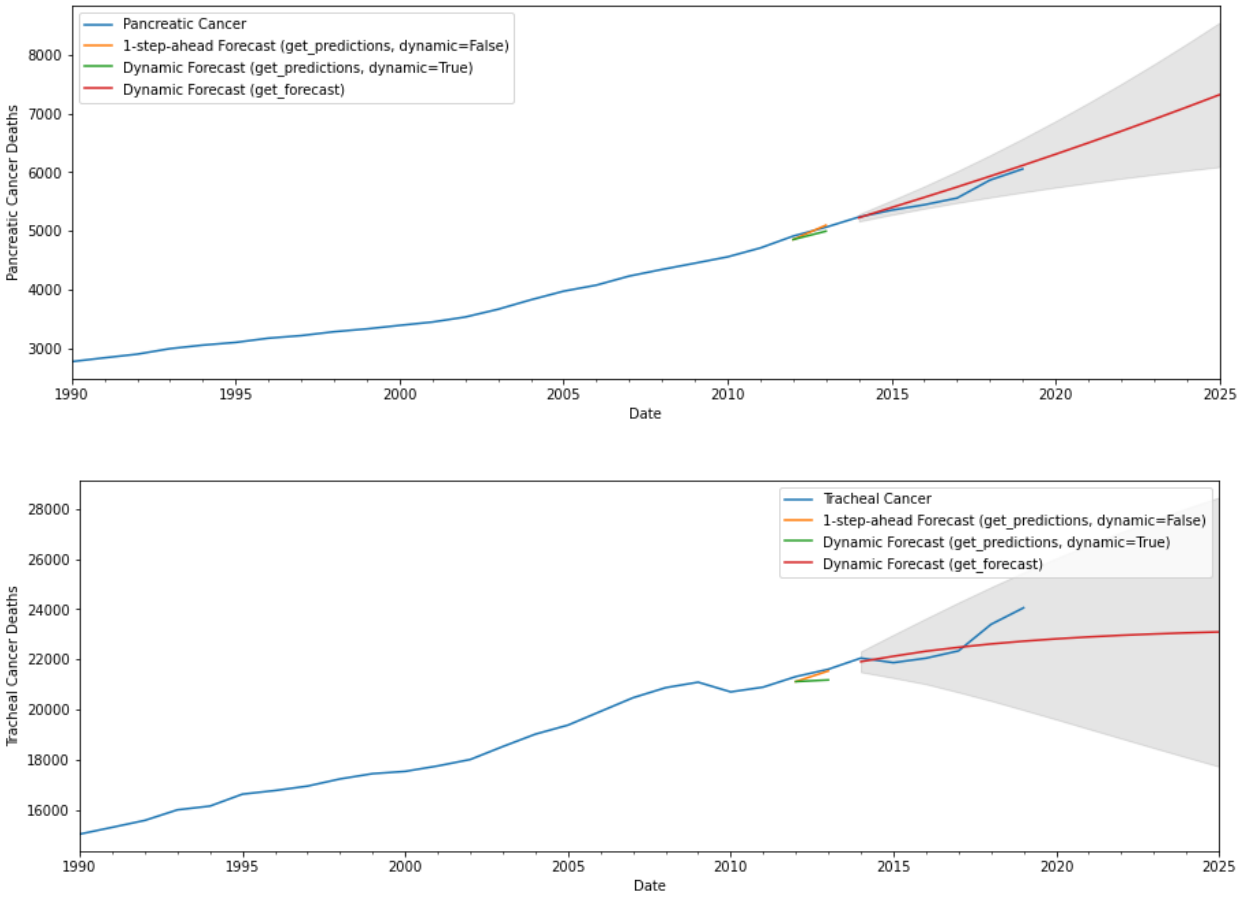


Figure 10: Line Graphs of forecasts predicted by the model for 3 cancer types

Performance Evaluation

To quantify the accuracy and the performance of the predictions, metrics such as Mean Absolute Error, Mean Squared Error and Root Mean Squared Error can be computed. Since these are absolute metrics and scale-dependent (expressed in units of the underlying data) these measures cannot be used to compare different time series. Percentage Error Metrics will be best suited to assess the performance of our outputs as they are scale independent and can be used to evaluate performances for time series-related models which are calculated by taking the average (mean) of the absolute difference between actuals and predicted values divided by the actuals. As per our model, the observations are as below,

```
Colon and Rectum Cancer
The Mean Absolute Percentage Error for the forecast in between 2014 &
2019 is 1.22%

Pancreatic Cancer

The Mean Absolute Percentage Error for the forecast in between 2014 &
2019 is 1.48%

Tracheal Cancer
The Mean Absolute Percentage Error for the forecast in between 2014 & 2
019 is 2.10%
```

Figure 11: Output of MAPE Values

All 3 of our observations have given a MAPE value of less than 5%, which is considered an indication that the forecast is acceptably accurate. Around 2.10% MAPE implies the model is about 97.9% accurate in predicting the next 15 observations.

Conclusion and Recommendation

The advantage of this data analysis report is it ensures that the health authorities in Canada are better equipped to deal with these types of cancer that are highly fatal. In the medium term, it will lead to a reduction in the number of deaths recorded by these cancer types.

The limitation of this analysis is it is fully concentrated on just the 3 top cancer types with the highest mortalities and does not speak to the rest of the cancer types.

From this point of view, we conclude our work with some recommendations:

- Based on the increase in all cancer cases in general, the government should increase the health budget and the budget allocated to the fight against cancer in particular.
- The number of qualified health personnel from all levels related to cancer should be increased, taking into account the increase in the following years.
- Treatments for active cancer diseases are very costly, therefore early diagnosis and screening practices should be expanded specifically for risk groups.
- We think that public service announcements should be broadcast to increase public awareness at various educational levels. They should also be broadcasted on social, print, and audio media, etc.

- With the help of this study, we think that we will direct the attention of the ministry to understand the causes of cancer cases and to reach more effective cancer treatments.
- And finally, we highly recommend more funding for data analysts in this field to enable them to expand their research to cover all cancer types and put an end to this fatal disease