# Cancer Statistical Modeling

Dr. Vinodh Chellamuthu, Dr. Md Sazib Hasan,
Riley Morgan, and Gregory Schmidt

September 2021

## 1 Abstract

Lung cancer is one of the deadliest cancers in the state of Utah. According to the National Cancer Society's estimates in the United States, there were about 228,820 new cases of lung cancer and about 135,720 deaths from lung cancer in the United States in 2020. We obtained data from the Surveillance Epidemiology and End Results (SEER) database of the National Cancer Institute (NCI) and utilized JoinPoint Regression to predict the future trend. We developed a hybrid joinpoint regression model, ARIMA time series model and Long short term memory (LSTM) model to describe mortality trends in Utah, assuming that the observed counts are probabilistically characterized by the Poisson distribution. We will compare our results with each of techniques for better accuracy. Using these methods, we can identify common trends that will allow us to come to a conclusion about lung cancer mortality in Utah. Our simulation results predict the rise of new deaths per year and future implications to help the policymakers in decision making.

## 2 Introduction

Cancer is the second leading cause of death in the United States behind heart disease [3]. The most deadly cancers worldwide in 2020 include lung cancer (1.8 million deaths), colon and rectum cancer (916,000 deaths), liver cancer (830 000 deaths), and stomach cancer (769 000 deaths). In addition to being the most fatal cancer in the world, lung cancer has the second highest annual incidence rate of all cancers with 2.21 million cases worldwide [4]. The quantity and severity of lung cancer cases make forecasting future lung cancer incidences and mortality rates a significant area of study.

In order to predict future trends of cancer, researchers employ a wide number of forecasting and regression techniques. These forecasting methods will aid in providing accurate and reliable models to predict the trajectory of cases based on specific factors.

A notable forecasting method includes the ARIMA model, a Time-Series-Analysis technique. For example, ARIMA models have been used to predict annual incidence and mortality rates of prostate cancer in Australia [5]. Another study has used ARIMA models to forecast cancer trends within the United States [6]. Furthermore, ARIMA models have been used to predict $CO_2$ emissions in India [17].

Researchers also turn to many other machine learning techniques. One study used support vector regression, back propagation, as well as a Long-Short-Term-Memory (LSTM) model to predict lung cancer incidence rates in the United States [7]. LSTM has also been used to forecast COVID-19 confirmed, recovered, and death cases [8]. Another study utilized LSTM and T-LSTM models for weather forecasting data [16].

A notable study for our research utilized both LSTM and ARIMA models in order to compare the performance of both models when trained on economic time series data [18].

Joinpoint Regression is another useful tool in order to analyze cancer mortality trends over time. For instance, Kafle used Bayesian Joinpoint Regression as a means to find the Annual Percentage Change (APC) of brain cancer mortality rates in the United States [9]. Additionally, Joinpoint Regression has been used to find the Age-Standardized mortality rates of cancer in Japan [10]. Furthermore, another study used Joinpoint Regression to analyze breast cancer incidence trends from data gathered from Isfahan, Iran [11].

The goal of this study is to analyze lung cancer mortality trends in the state of Utah by using various machine learning and regression models. We will be focusing on using Time Series Analysis (ARIMA), LSTM models, and Joinpoint Regression for this study. Additionally, we will be comparing lung cancer mortality trends based on different biological and geographical characteristics such as: age, gender, ethnicity, and county location. Finally we will be discussing and comparing the quality of each model for every fit.

# 3 Methodologies

## 3.1 Data Description

The data used to perform our study on the trend analysis of lung cancer within the state of Utah was gathered from the National Cancer Institute's Surveillance, Epidemiology, and End Results Program (SEER)[1]. Using the SEER*Stat Software, we were able to extract lung cancer data from the year 1969 to the year 2019 in the state of Utah. The data provided the number of deaths caused by lung cancer in each year during that the time-frame. Each individual case provided data on the county, age, gender and ethnicity of the specific incident. Data gathered for county and age were converted into a crude rate per 100,000 people by using SEER's annual population data [2]. Using this information we were able to use several models to analyze mortality data to predict the future trends of lung cancer mortality and the factors that may effect the number of deaths of Lung Cancer within the state of Utah.

## 3.2 ARIMA

ARIMA method is a powerful and effective tool used to analyze and forecast time series data. ARIMA is an acronym that describes the features of the culmination of different time series models put together. It stands for Auto Regressive Integrated Moving Average. The Auto Regressive (AR) term specifies an output that depends on the order of lagged observations. The definition of an Autoregressive model of order $p$, AR($p$) is written as such:

$$X_t = c + \sum_{i=1}^{p} \phi_i X_{t-i} + \epsilon_t$$

Where $\phi_i$ is the $i$th parameter of the model, $c$ is a constant, and $\epsilon_t$ is Gaussian white noise with mean zero and a variance of $\sigma_\epsilon^2$. The Integrative term (I) describes the order of differencing in terms within the data. For every term $a_t$, we replace it with $z_t$ where $z_t = a_t - a_{t-1}$. This process is repeated $d$ times in order to transform the data to be stationary. Finally, the Moving Average (MA) term accounts for residual errors for the output. The MA($q$) model, of order $q$, can be defined as:

$$X_t = \mu + \sum_{i=0}^{q} \theta_i \epsilon_{t-i}$$

Where $\mu$ is the mean of the time series data, and $\theta_i$ is the coefficient parameter for the residual white noise error $\epsilon_{t-i}$. We can combine these equations to form a complete ARIMA($p, d, q$) model defined by:

$$X_t = c + \sum_{i=1}^{p} \phi_i X_{t-i} + \epsilon_t + \mu + \sum_{i=0}^{q} \theta_i \epsilon_{t-i},$$

where p is the number of autoregressive terms,d is the number of nonseasonal differences needed for stationarity, and q is the number of lagged forecast errors in the prediction equation.

## 3.3 Long Short Term Memory (LSTM)

Neural networks are a practical method used to fit trends within data. LSTM is a special kind of neural network called a Recurrent Neural Network (RNN) that can process sequences of data at a time in order to predict new values. The most basic kind of neural network is called an Artificial Neural Network (ANN) which consists of three parts: an input layer, a number of hidden layers, and an output layer. The size of the input layer is determined by the number of features in the data. Individual layers are made up of an arbitrary amount of perceptrons, and each perceptron in one layer connects to every perceptron in the next layer. In order to produce an output, all perceptrons take a weighted sum of every weighted input and applies it to an activation function. The activation function will then determine the output of the perceptron. Some common activation functions for regression learning include sigmoid, tanh, and

ReLu. In order to obtain an optimal weight for every connection, a technique called back propagation is used to adjust weights for every epoch.

RNN improves upon this idea by using a sequence of data in order to learn to predict the next data point in the sequence. In this network the hidden layers have an extra function by remembering information that was fed through the network in an earlier step. These kinds of networks suffer from a problem that arises when training data with long-term dependencies. This problem is called the vanishing gradient problem [12]. The vanishing gradient causes weights to either approach zero or infinity during back propagation which in term causes further adjustments to those weights to be very minimal. Therefore new methods to avoid the vanishing gradient were developed such as the LSTM network [13]

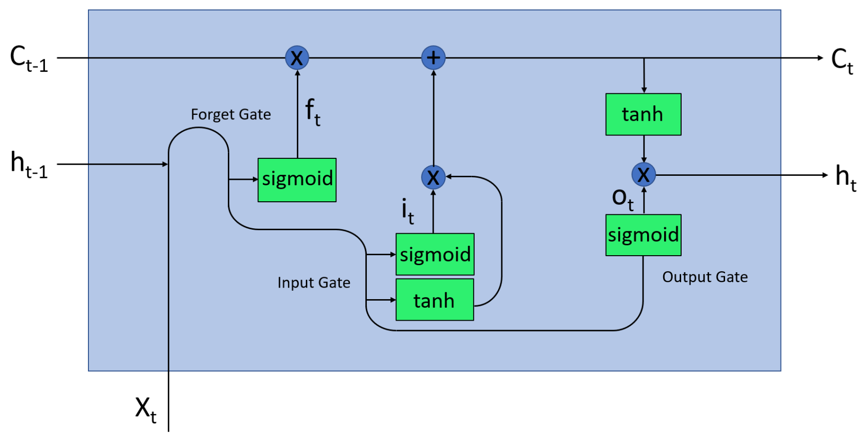



Figure 1: LSTM architecture

The LSTM network consists of cells that store information and data. In each cell there are three sequential gates: a forget gate, an input gate, and an output gate. The forget gate determines which information should be passed on and which information should be removed or "forgotten". The forget gate can be expressed mathematically as such:

$$f_t = \sigma(W_f h_{t-1} + U_f x_t + b_f) \tag{1}$$

where $f_t$ is a value ranging from 0 to 1. The input gate's function is to add new information to the cell state from the current input passed on by the forget gate. It's calculation is shown as follows:

$$i_t = \sigma(W_u h_{t-1} + U_i x_t + b_i) \tag{2}$$

$$\widetilde{C_t} = tanh(W_C h_{t-1} + U_C x_t + bC) \tag{3}$$

where $i_t$ is a value between 0 to 1. $C_t$ is the candidate cell state that will be used to update the current cell state. Together, equations 2 and 3 can be used to create an expression for updating the cell state:

$$C_t = C_t - 1 \odot f_t + i_t \odot \widetilde{C_t} \tag{4}$$

where $\odot$ is the Hadamard product of a matrix. Finally, the output gate $o_t$ simply determines the output hidden state $h_t$ based on the current cell state. The mathematical expressions for $o_t$ and $h_t$ are as such:

$$o_t = \sigma(W_o h_{t-1} + U_o x_t + b_o) \tag{5}$$

$$h_t = o_t \odot tanh(C_t) \tag{6}$$

where $o_t$ ranges from 0 to 1 and $h_t$ is the new hidden state created from the information of the current cell state [14].

It is common practice in machine learning to split up the data into 3 parts: training, validation and testing. This is done in order to avoid over-fitting a model. For this study, and for each model, we have set aside the first 49 percent of the data as training data, the next 21 percent of the data was set aside as validation data, and the remaining 30 percent of the data as testing data.

The neural network used for this study consists of over 297,000 parameters. The specific configuration of our neural network's layers are as follows: first the input layer, then two LSTM layers of sizes 200 and 100, followed by a series of dense layers of sizes 100 and 50, and finally an output layer.

In order to determine the best fit LSTM model, we have trained multiple models of varying window sizes for every demographic. We then choose the most balanced model between the training and testing RMSE loss for every window size. For every model we typically tried window sizes between 1 and 5, or between 1 and 3 if there is not enough data have a large enough validation set.

## 3.4 Joinpoint Regression

Joinpoint regression is a method that is used in different domains to assess the changes in time-series data. Scenarios in which joinpoint modeling is often used includes monitoring suicide rates, cancer mortality rates and driver deaths after implementing new traffic regulations [15]. To analyse the trends using joinpoint models, different connected lines are used to explain the changes in time based on different inflection points. In a joinpoint regression model, linear regression line segments are connected at joinpoints to show the difference in trend intervals over a period of time. The result of this provides a piecewise function in which multiple different linear regression lines combine to form the trend of the dataset. This allows for more accurate and detailed modeling dependent on the needs of the data.

The general form of a joinpoint regression model is given by a piecewise function (7). Each linear regression line that makes up the function y is connected at a point from the joinpoint vector $j = (x_1, x_2, \ldots, x_{k-1}, x_k)$ on the x axis.

$$y = \begin{cases} ax + c_1 & x < x_1 \\ bx + c_2 & x_1 \leq x < x_2 \\ \vdots & \vdots \\ dx + c_3 & x_{k-1} \leq x < x_k \\ ex + c_4 & x_k \leq x \end{cases} \tag{7}$$

Annual percentage change (APC) in one way to characterize the growth over time within certain years. APC states that the rates are assumed to be constant for each year in a certain period. A positive APC value describes growth within the time-frame, while a negative APC describes decay. To calculate the APC we fit the least square regression line to the natural logarithm of rates using the year as the regression variable. The equation to calculate the APC is as follows:

$$APC = 100 * (e^m - 1) \tag{8}$$

where m is the slope of the linear regression line. In order to compare the APCs of lstm and arima models, joinpoint regression was used on each model's predicted values. This was done in order to better compare the most recent APC trends of each model.

# 4 Results

## 4.1 Total Utah Deaths

Lung cancer deaths have steadily increased over time reaching 447 total deaths per year in 2018. By using Joinpoint Regression we have found that a model with one joinpoint at the year 1993 was the best fit. The APC of the model changed from 4.0 to 1.2 suggesting that the growth of the total number of deaths is slowing down. An arima model of order (4, 1, 1) was found to be the most accurate arima model. The arima model also showed a steady increase in total deaths. The APCs of the arima model were best calculated with two joinpoints changing from 3.60 to 2.03 to 0.69. Meanwhile the LSTM model showed a similar result. The most balanced and optimal LSTM model used had a window size of 4 which can be observed in table 1. The LSTM model's predictions also produced a 2 joinpoint model when calculating the APC. The individual APCs of each line segment went from 3.03 to 1.68 to 0.58. In each case every model suggests that the total growth of lung cancer deaths in Utah is slowing down in terms of APC. The LSTM model (window size 8) had a better testing loss of 22.2 compared to the ARIMA model's 24.59.

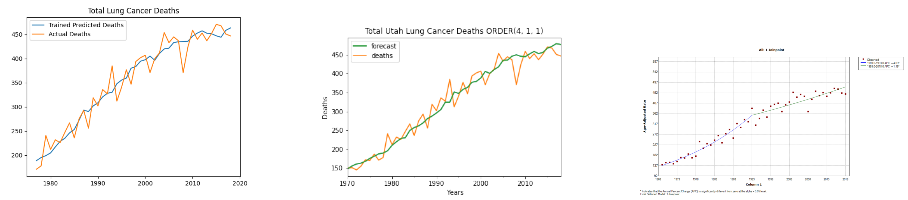



Figure 2: Total Utah Deaths: a. using ARIMA, b. Using LSTM and c. Using Joinpoint regression

Table 1: Total Utah Deaths

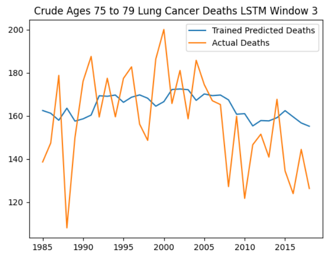
| Window Size | Training Loss | Validation Loss | Testing Loss |
|---|---|---|---|
| 1 | 36.38 | 26.24 | 27.29 |
| 2 | 62.24 | 21.12 | 27.63 |
| 3 | 46.75 | 19.61 | 29.51 |
| 4 | 30.61 | 23.37 | 29.88 |
| 5 | 35.03 | 22.10 | 26.68 |

## 4.2 Age

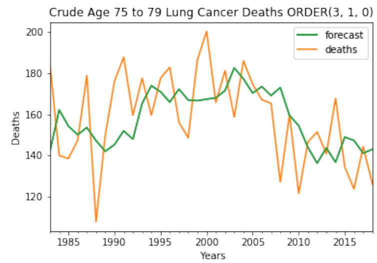
Age shows a significant correlation to lung cancer deaths and we have chosen to model the age groups that are most at risk for lung cancer related deaths. The three age groups that are most at risk are: ages 85+, ages 80-84, and ages 75-79. A crude death rate per 100,000 people in their respective age group will help us understand if the average amount of deaths in each age group is increasing or decreasing.

## 4.3 Age: 75-79

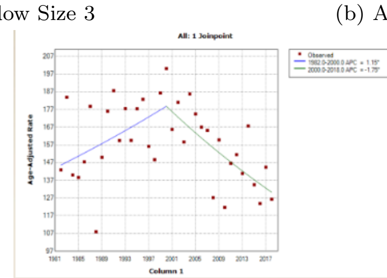
A joinpoint model with one joinpoint in the year 2000 showed that the APC changed from 1.15 to -1.75. This indicates that the overall crude death rate per 100,000 people is decreasing in this age group. The corresponding arima model of order (3, 1, 0) predicted a similar trend. The APCs of the ARIMA model used 2 joinpoints that change from 0.87 to -3.92 to 0.51. This suggests that the arima model predicts that the current trend is slowly growing. The LSTM model with the most balanced loss used a window size of 3. The LSTM model used one joinpoint for calculating the APCs. The APCs changed from 0.40 to -0.62. The LSTM model most closely matches the joinpoint model on the original data in terms of APC. The ARIMA model had a better fit with a testing loss of 15.44 as opposed to LSTM's most balanced model with a testing loss of 24.64.




(a) LSTM Window Size 3




(b) ARIMA model




(c) Joinpoint model

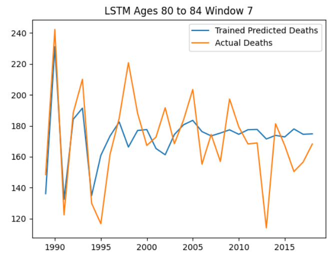
Figure 3: Crude rate per 100,000 people of ages 75 to 79

Table 2: Ages 75 to 79 models

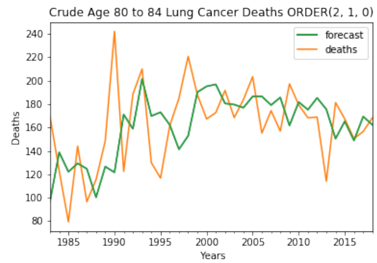
| Window Size | Training Loss | Validation Loss | Testing Loss |
|---|---|---|---|
| 1 | 20.02 | 6.18 | 33.27 |
| 2 | 22.32 | 9.98 | 25.95 |
| 3 | 21.43 | 10.71 | 24.64 |
| 4 | 21.84 | 19.36 | 17.56 |
| 5 | 23.42 | 18.60 | 16.41 |

## 4.4 Age: 80-84

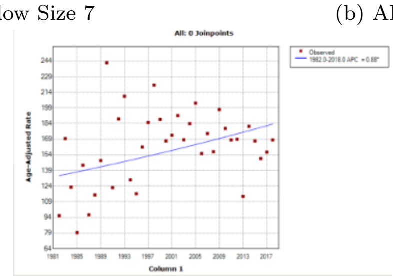
The original data used a joinpoint model with zero joinpoints with an overall APC of 0.88. Meanwhile the ARIMA model of order (2, 1, 0) used a joinpoint model with one joinpoint with the APC changing from 3.17 to -0.93. The LSTM model of window size 4 used a joinpoint model with one joinpoint as well. The APCs of the LSTM model changed from 0.98 to 0.32. Here, the ARIMA and LSTM models most closely match each other as opposed to the original data. These two models also indicate a decrease in the overall death rate as opposed to the original joinpoint model. The ARIMA model outperformed the LSTM model (window size 7) in terms of testing loss with 16.79 as opposed to 22.84.




(a) LSTM Window Size 7




(b) ARIMA model




(c) Joinpoint model

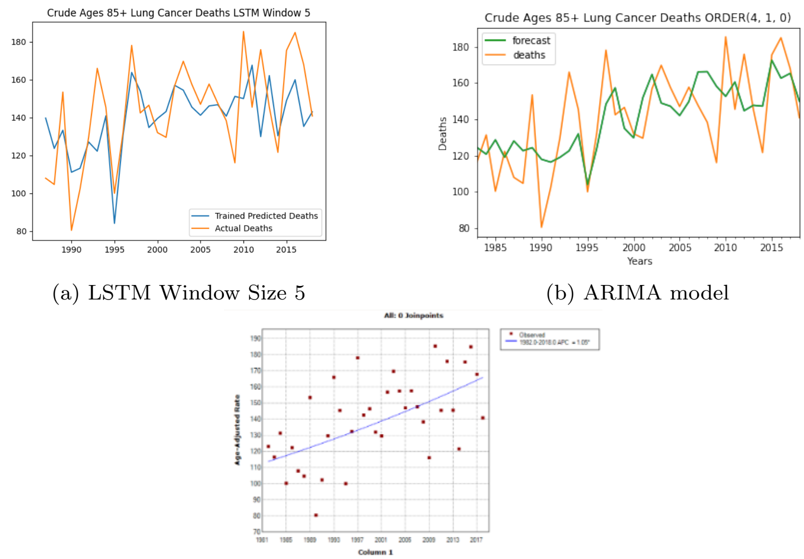
Figure 4: Crude rate per 100,000 people of ages 80 to 84

Table 3: Ages 80 to 84 models

| Window Size | Training Loss | Validation Loss | Testing Loss |
|---|---|---|---|
| 1 | 41.47 | 18.85 | 21.87 |
| 2 | 42.55 | 18.04 | 21.91 |
| 3 | 40.06 | 16.73 | 23.14 |
| 4 | 36.28 | 18.40 | 23.60 |
| 5 | 38.78 | 18.60 | 22.18 |

### 4.5 Age: 85 and Older

A joinpoint model on the original data was calculated with zero joinpoints and had an overall APC of 1.05. The ARIMA model of order (4, 1, 0) used a joinpiont model with two joinpoints at years 1995 and 1998. The APC changed from -0.54 to 7.78 to 0.41. For the LSTM model, a window size of 5 was most appropriate. The joinpoint model for the LSTM model's predictions most closely resembled the original data. The model used zero joinpoints with an overall APC of 0.73. The LSTM model (window size 5) had a better testing loss compared to the ARIMA model, 17.99 to 25.15.




(a) LSTM Window Size 5

(b) ARIMA model

(c) Joinpoint model

Figure 5: Crude rate per 100,000 people of ages 85+

Table 4: Ages 85+ models

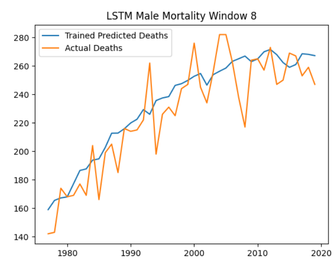
| Window Size | Training Loss | Validation Loss | Testing Loss |
|---|---|---|---|
| 1 | 29.22 | 8.96 | 26.50 |
| 2 | 28.72 | 11.83 | 26.95 |
| 3 | 28.47 | 12.46 | 26.83 |
| 4 | 25.56 | 7.39 | 17.16 |
| 5 | 18.76 | 7.97 | 17.99 |

### 4.6 Gender: Male

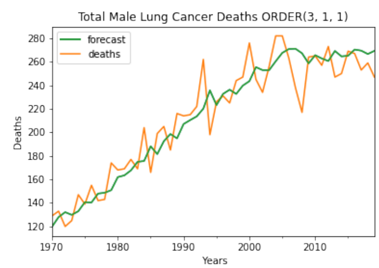
The overall total male deaths used a joinpoint model with one joinpoint in the year 1999. The APC decayed from 2.47 all the way to -0.01. The ARIMA model of order (3, 1, 1) used a joinpoint model with two joinpoints located at 1993 and 2006. The APC also showed a decay from 2.64 to 1.40 to 0.00. Similarly the LSTM model of window size 5 had a joinpoint model with two joinpoints. It showed a decay from 1.8 to 0.97 to -0.04. In each case, every model suggested that the APC of male deaths is stabilizing. The ARIMA model performed better on the testing data with a testing loss of 16.03 as opposed to LSTM's (window size 8) testing loss of 18.43.

### 4.7 Gender: Female

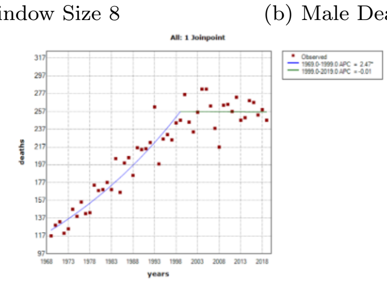
For total female deaths, a joinpoint model with one joinpoint in 1993 showed that the APC had changed from 7.61 to 2.05. The ARIMA model of order (1, 1, 1) used a joinpoint model with two joinpoints located at 1987 and 2000 and had an APC change of 6.35, 4.73, and 2.03. Meanwhile the LSTM model with window size 4 showed an APC change from 4.77 to 1.69 with only one joinpoint. Each of the models indicate that the APC of total female deaths is slowing down. The LSTM model slightly outperformed the ARIMA model on the testing data with a loss of 14.07 as opposed to 14.53.




(a) Male Deaths Window Size 8




(b) Male Deaths ARIMA model

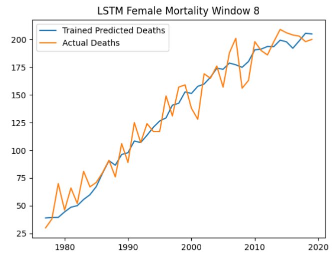



(c) Joinpoint model

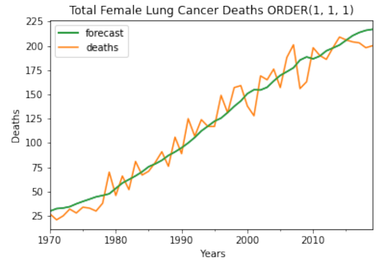
Figure 6: Total Male Deaths

Table 5: Male Demographic models

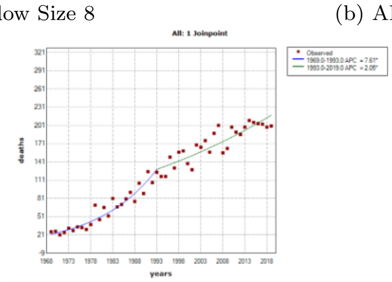
| Window Size | Training Loss | Validation Loss | Testing Loss |
|---|---|---|---|
| 1 | 37.55 | 20.47 | 16.61 |
| 2 | 33.78 | 20.45 | 22.41 |
| 3 | 26.88 | 19.59 | 19.62 |
| 4 | 25.34 | 19.82 | 21.79 |
| 5 | 21.28 | 18.51 | 19.65 |




(a) LSTM Window Size 8




(b) ARIMA model




(c) Joinpoint model

Figure 7: Total Female Deaths

Table 6: Female Demographic models

| Window Size | Training Loss | Validation Loss | Testing Loss |
|---|---|---|---|
| 1 | 19.31 | 19.15 | 17.88 |
| 2 | 33.17 | 17.88 | 20.12 |
| 3 | 30.10 | 17.50 | 17.02 |
| 4 | 14.65 | 17.65 | 14.16 |
| 5 | 12.50 | 16.82 | 15.15 |

## 4.8 County

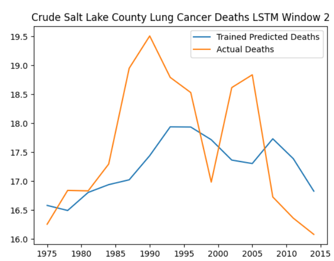
In order to have a good sample size, we have chosen to model the four most populated counties in utah: Salt Lake County, Utah County, Davis County, and Weber County. We have also converted the data into a crude death rate of deaths per 100,000 people in their respective counties. This is done to see if the overall death rate is increasing or decreasing.

## 4.9 County: Salt Lake County

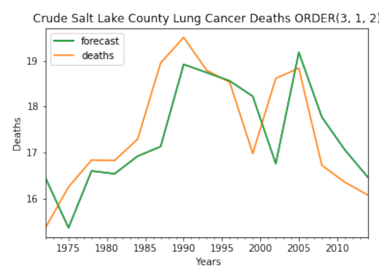
The Salt Lake County crude death rate joinpoint model used one joinpoint in the year 1993 and had an APC decrease of 1.02 to -0.63. The corresponding ARIMA model of order (3, 1, 2) also used a joinpoint model with one joinpoint. The APC changed from 0.86 to -0.46. Similarly the LSTM model with window size 2 produced a joinpoint model with one joinpoint with the APC changing from 1.25 to -0.77. Here, each model agrees that the overall lung cancer death rate in Salt Lake County is on a downward trend. The ARIMA model had a slightly better testing score of 0.63 compared to the LSTM model's (window size 2) testing score of 1.14.

### 4.10 County: Utah County

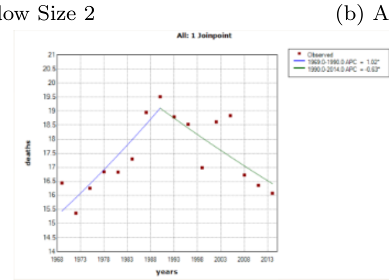
The Utah County joinpoint model did not use any joinpoints and had an APC of -0.43. The ARIMA model for Utah County with order (1, 1, 1) produced a joinpoint model with one joinpoint in 1972 and had an APC change from 0.8 to -1.28. The LSTM model with window size 4 also produced a joinpoint model with one joinpoint. The APC for this model decayed from 1.43 to -2.22. In each model the recent APC trend for Utah County is negative and ranges from -0.43 to -2.22. The ARIMA model had a better testing score of 0.58 compared to the LSTM model's (window size 1) of 1.40.




(a) LSTM Window Size 2




(b) ARIMA model

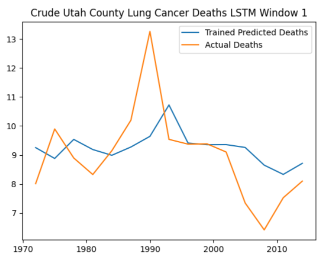



(c) Joinpoint model

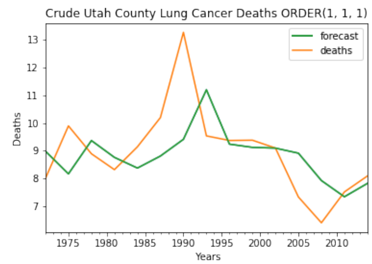
Figure 8: Crude death rate per 100,000 people in Salt Lake County

Table 7: Salt Lake County LSTM models

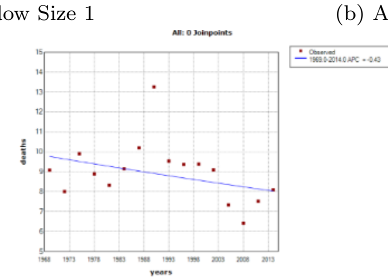
| Window Size | Training Loss | Validation Loss | Testing Loss |
|---|---|---|---|
| 1 | 1.01 | 0.72 | 1.07 |
| 2 | 1.14 | 0.67 | 1.14 |
| 3 | 0.95 | 1.04 | 1.26 |




(a) LSTM Window Size 1




(b) ARIMA model




(c) Joinpoint model

Figure 9: Crude death rate per 100,000 people in Utah County

Table 8: Utah County LSTM models

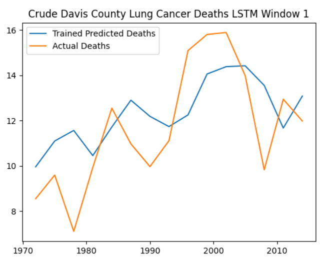
| Window Size | Training Loss | Validation Loss | Testing Loss |
|---|---|---|---|
| 1 | 1.55 | 0.03 | 1.40 |
| 2 | 2.67 | 1.56 | 3.93 |
| 3 | 3.11 | 2.97 | 6.16 |

## 4.11 County: Davis County

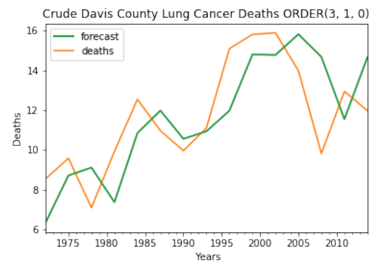
The Davis County model used one joinpoint and had an APC change from 2.60 to -1.87. The ARIMA model with order (3, 1, 0) produced a joinpoint model without any joinpoints. The overall APC of the ARIMA model was 1.74. The LSTM model of window size 1 also produced a joinpoint model without any joinpoints. The overall APC of the LSTM model was 1.83. In this county location the original joinpoint model greatly differed from each of the other trained models in terms of APC in the recent trend. The ARIMA model had a better training loss of 1.27 compared to the LSTM model's (window size 1) 1.96.

Table 9: Davis County LSTM models

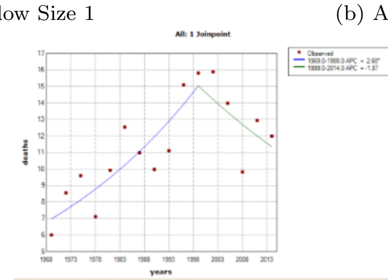
| Window Size | Training Loss | Validation Loss | Testing Loss |
|---|---|---|---|
| 1 | 2.07 | 2.36 | 1.96 |
| 2 | 2.27 | 2.48 | 2.57 |
| 3 | 2.90 | 0.94 | 5.41 |




(a) LSTM Window Size 1




(b) ARIMA model




(c) Joinpoint model

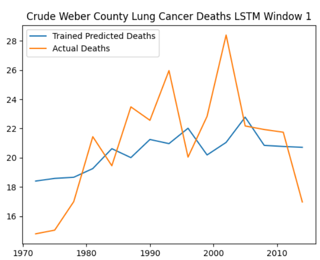
Figure 10: Crude death rate per 100,000 people in Davis County

## 4.12 County: Weber County

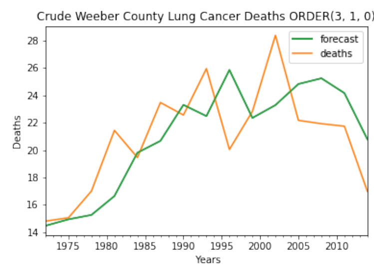
The Weber County joinpoint model used one joinpoint in 2002. This model showed an APC change from 1.83 to -3.34. The ARIMA model of order (3, 1, 0) produced a joinpoint model with one joinpoint with an APC change 2.47 to -0.60. The LSTM model with window size 1 also produced a joinpoint model with one joinpoint. This model had an APC change from 2.12 to -0.61. This indicates that the current overall death rate in Weber County is decreasing in each model. The ARIMA model had a better testing loss of 2.65 compared to the LSTM model's (window size 1) 3.76.

Table 10: Weber County LSTM models

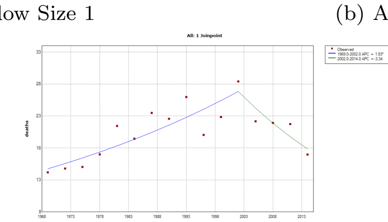
| Window Size | Training Loss | Validation Loss | Testing Loss |
|---|---|---|---|
| 1 | 3.03 | 2.33 | 3.76 |
| 2 | 2.67 | 1.56 | 3.93 |
| 3 | 3.11 | 2.97 | 6.16 |




(a) LSTM Window Size 1




(b) ARIMA model



(c) Joinpoint model

Figure 11: Crude death rate per 100,000 people in Weber County

A overall conclusion considering all the aspect has to be drawn here.

# 5 Conclusion

working on it

# References

[1] Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Mortality - All COD, Aggregated With State, Total U.S. (1969-2019) <Katrina/Rita Population Adjustment>, National Cancer Institute, DCCPS, Surveillance Research Program, released April 2021. Underlying mortality data provided by NCHS (www.cdc.gov/nchs)

[2] Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Populations - Total U.S.

(1969-2020) <Katrina/Rita Adjustment>- Linked To County Attributes - Total U.S., 1969-2020 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, released January 2022.

[3] Ahmad FB, Anderson RN. The Leading Causes of Death in the US for 2020. JAMA. 2021;325(18):1829–1830. doi:10.1001/jama.2021.5469

[4] World Health Organization (WHO). (2022, February 3). Cancer. Who. Retrieved July 4, 2022, from https://www.who.int/news-room/fact-sheets/detail/cancer

[5] Earnest, A., Evans, S. M., Sampurno, F., & Millar, J. (2019). Forecasting annual incidence and mortality rate for prostate cancer in Australia until 2022 using autoregressive integrated moving average (ARIMA) models. BMJ Open, 9(8), e031331. https://doi.org/10.1136/bmjopen-2019-031331

[6] Xie, L. (2017). Time Series Analysis and Prediction on Cancer Incidence Rates. Journal of Medical Discovery, 2(3). https://doi.org/10.24262/jmd.2.3.17030

[7] Tuncal, K., Sekeroglu, B., & Ozkan, C. (2020). Lung Cancer Incidence Prediction Using Machine Learning Algorithms. Journal of Advances in Information Technology, 11(2), 91–96. https://doi.org/10.12720/jait.11.2.91-96

[8] Absar N, Uddin N, Khandaker MU, Ullah H. The efficacy of deep learning based LSTM model in forecasting the outbreak of contagious diseases. Infect Dis Model. 2022 Mar;7(1):170-183. doi: 10.1016/j.idm.2021.12.005. Epub 2021 Dec 28. PMID: 34977438; PMCID: PMC8712463.

[9] Kafle, Ram C., "Trend Analysis and Modeling of Health and Environmental Data: Joinpoint and Functional Approach" (2014). Graduate Theses and Dissertations.

[10] Qiu, D., Katanoda, K., Marugame, T., & Sobue, T. (2009). A Joinpoint regression analysis of long-term trends in cancer mortality in Japan (1958-2004). International Journal of Cancer, 124(2), 443–448. https://doi.org/10.1002/ijc.23911

[11] Dehkordi ZF, Tazhibi M, & Babazade S. Application of joinpoint regression in determining breast cancer incidence rate change points by age and tumor characteristics in women aged 30-69 (years) and in Isfahan city from 2001 to 2010. J Edu Health Promot 2014;3:115.

[12] Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks, 5(2), 157–166. https://doi.org/10.1109/72.279181

[13] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[14] Wang, P., Zheng, X., Ai, G., Liu, D., & Zhu, B. (2020). Time series prediction for the epidemic trends of COVID-19 using the improved LSTM deep learning method: Case studies in Russia, Peru and Iran. Chaos, Solitons Fractals, 140(November). https://doi.org/10.1016/j.chaos.2020.110214

[15] Gillis, Daniel, & Brandon P.M. Edwards. (2019). The Utility of Joinpoint Regression for Estimating Population Parameters given Changes in Population Structure, https://www.sciencedirect.com/science/article/pii/S2405844019361754.

[16] Karevan, Z., Suykens, J. A. K. (2020). Transductive LSTM for time-series prediction: An application to weather forecasting. Neural Networks, 125, 1–9. https://doi.org/10.1016/j.neunet.2019.12.030

[17] Nyoni T, Bonga WG (2019) Prediction of CO2 emissions in india using arima models. DRJ-J Econ Finance 4(2):1–10

[18] Siami-Namini, S., Tavakoli, N., Siami Namin, A. (2018). A Comparison of ARIMA and LSTM in Forecasting Time Series. 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA). https://doi.org/10.1109/icmla.2018.00227