The Progress of Cancer Research: Analyzing the Amount of New Scientific Discoveries Published Over Time

Erick Lu

22026297

1 Background and Introduction

1.1 The Question

The scientific question that I am addressing in this paper is: *can I use ARIMA models to model the amount of scientific papers published in the field of cancer research over time?* Modeling the behavior of the academic publication process may provide insight and understanding on the progress of research, as well as reveal the productivity patterns of those who work the field.

1.2 The Data

The data we will be working with consists of monthly counts of the number of scientific papers published in the field of cancer research from January 1964 to December 2011. The source of the data is PubMed[1], which is an online database containing scientific articles from peer-reviewed journals in the life sciences. In order to obtain the data, I wrote a web scraping program in Python that searches PubMed for cancer research articles and downloads the abstracts directly. All 2,135,309 abstracts pertaining to cancer research were downloaded, and the month and year of each abstract was extracted into a text file. I then used R to count the number of scientific articles released under each month of each year to make the time series. There are a total of 576 equispaced data points.

Abstracts came from 508 peer-reviewed journals. These include journals specifically devoted to cancer research, such as *Journal of Clinical Oncology*, as well as journals that cover a broad variety of topics pertaining to immunobiology, such as *Cell Host and Microbe* and *PLoS Genetics*. Every month, only a fixed amount of articles can be published in any issue of any journal. A fraction of these publications will be in the field of cancer research. Thus, a time series analysis on the number of papers published in the field of cancer over time may yield many interesting conclusions, including when the optimal time of the year to submit a paper on cancer might be. This will be addressed in the Conclusion section.

[1]National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/pubmed

2 Exploratory Data Analysis: Gaining Insight on the Data

Figure 1 below is a plot of the raw data, along with a small subsection plotted with monthly symbols.

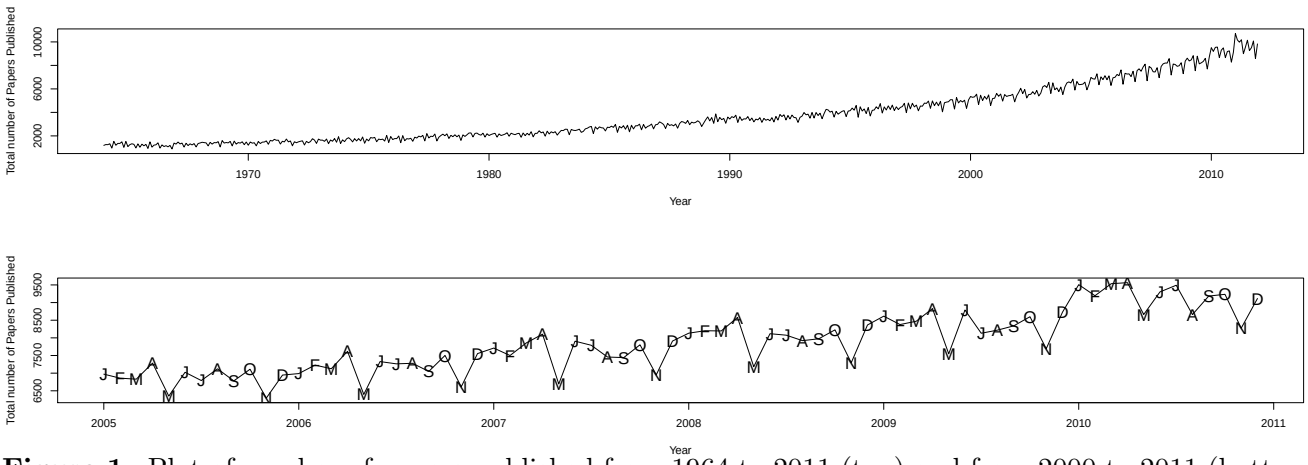

Figure 1: Plot of number of papers published from 1964 to 2011 (top) and from 2000 to 2011 (bottom).

Looking at the plot of the raw data, we observe an increasing trend as well as a seasonal pattern. Values for the number of papers published seem to be lowest in May and November of each year, and highest in the months preceeding them. Explanations for this behavior may be due to decreased productivity during the Summer and Winter holidays, and increased productivity due to PhD students who are pressured to publish before graduating. We can look at the underlying structure of the data in Figure 2.

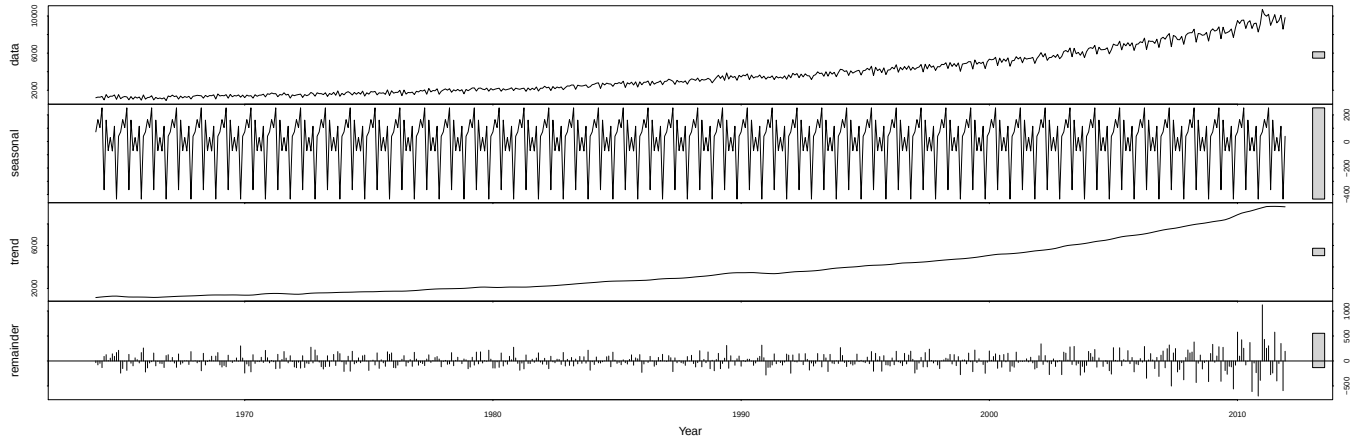

Figure 2: Seasonal decomposition of time series by Loess. Graph shows raw data (top), seasonal pattern (second row), trend (third), and residuals (bottom).

The repeating pattern in the second row of Figure 2 verifies the seasonal pattern we observered earlier in the raw data. We can also see that the variance in the residuals greatly increases from 2009 and onward. In order to satisfy the constant variance assumption when fitting ARIMA models, we will apply a transformation to the series to *stabilize the variance*. A popular technique is the power transformation, introduced by Box and Cox (1964), in which the data, y, is transformed to $g(y) = \frac{y^\lambda - 1}{\lambda}$ for $\lambda \neq 0$ and $\log(y)$ for $\lambda = 0$. We may construct a maximum likelihood plot for λ (Figure 3, on next page), in order to determine the order of the power transformation. The maximum likelihood estimate for lambda based

on our time series is 0.5, suggesting a square root transformation of the data.

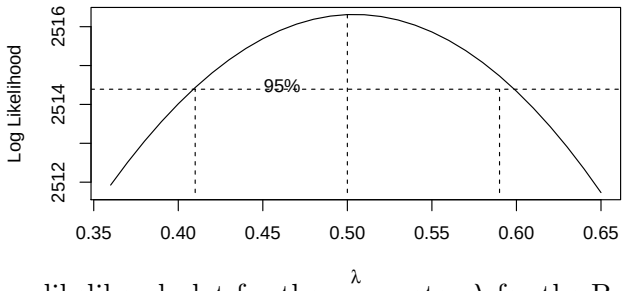

Figure 3: Maximum likelihood plot for the parameter λ for the Box-Cox transformation.

2.1 Acquiring an approximately stationary series

Now that we have observed the behavior of the data, we will attempt to transform the series into something approximately stationary, in order to be able to satisfy the stationarity required for fitting ARIMA models. In addition to stabilizing the variance with the square root transformation, we will remove the mean level of the process by taking the first difference of the Box-Cox transformed data.

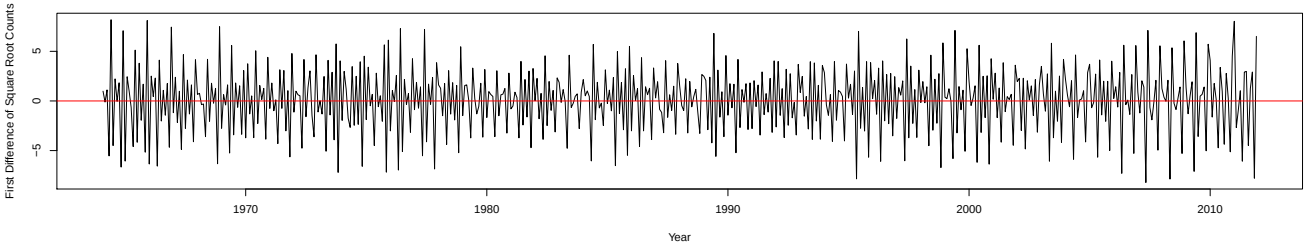

Figure 4: Plot of first difference of the square root of the data over time.

The time series of the first difference of the square root seems to have zero mean and have roughly constant variance. However, strong seasonality is sill present, as evidenced by the sample autocorrelation function below in Figure 5. This suggests that seasonal differencing may be required to achieve an approximately stationary series. The sample ACF displays a repeating pattern of peaks at multiples of 6.

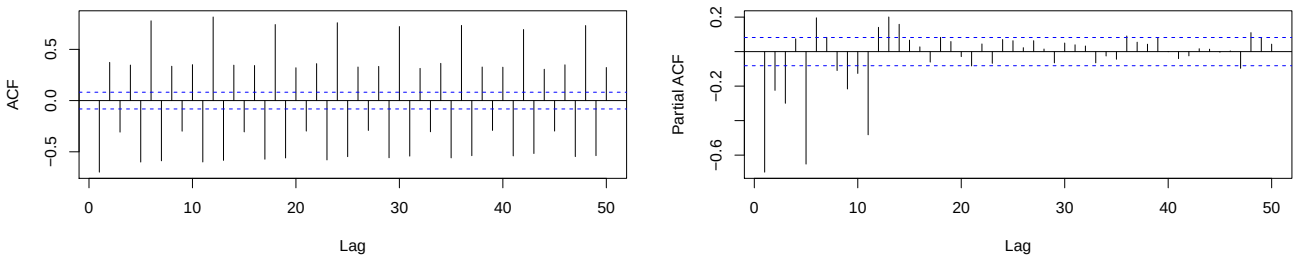

Figure 5: Sample ACF (left) and PACF (right) plots of the first difference of the square root data.

This can be explained by the biannual oscillations seen each year in the raw data. However, if we observe the raw data closely, as well as the seasonal decomposition in Figure 2, the two peaks have slightly different behavior, in which values from February to April behave differently than values from August to October. Thus, for future analysis, I will be using a seasonal lag of 12 instead of six, to account for the different behavior. The information we have discovered so far points toward fitting a seasonal ARIMA model to

the data. The effects of seasonal differencing (s = 12) are shown in the plots in Figure 6 below:

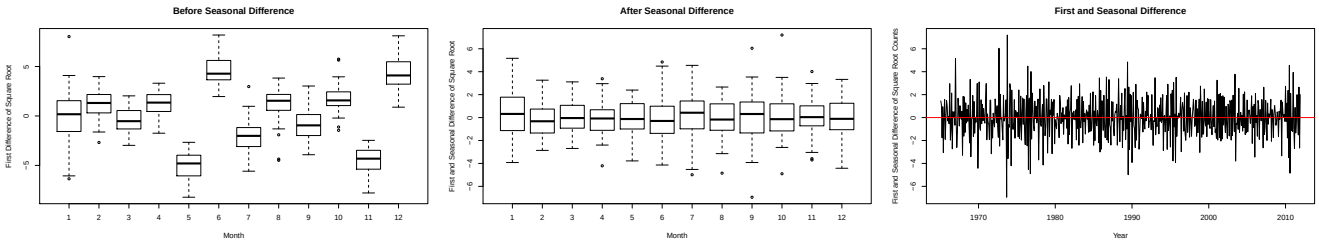

Figure 6: Plots of first diff. of square root before (far left) and after (right 2) seasonal differencing.

After taking the seasonal difference of the data with period 12, a plot of the data (far right) looks approximately stationary. Judging by the box plots, the seasonal difference has removed most of the variability between months in the data. However, the sample ACF and PACF of the data after first and seasonal differencing shows that seasonal autocorrelation still exists at multiples of lag 12, shown below.

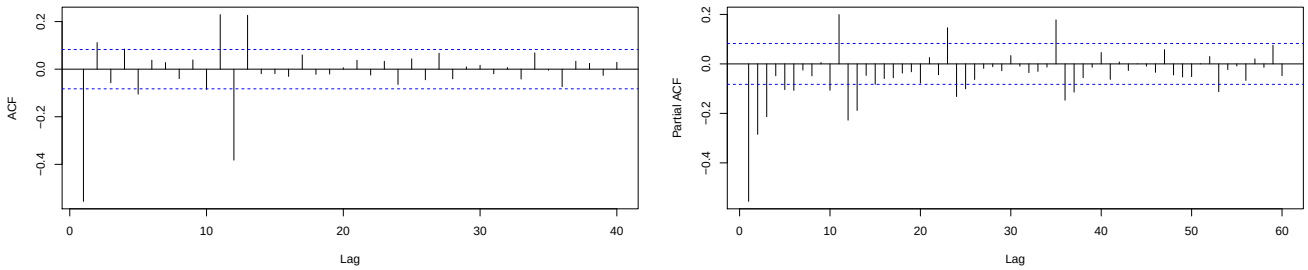

Figure 7: Sample ACF (L) and PACF (R) of the transformed series after first and seasonal differencing.

The sample ACF and PACF of the first and seasonal difference both exhibit significant nonseasonal and seasonal autocorrelation, suggesting that the data behave according to a *multiplicative* seasonal ARIMA model with period 12. The nonseasonal and seasonal autocorrelation cuts off after 1 and 12 respectively, whereas the nonseasonal and seasonal partial autocorrelation cuts off at lags 3 and 36, respectively.

3 Fitting the Seasonal ARIMA Model

Let y_1,..., y_t be a discrete equispaced time series. A seasonal ARIMA $(p,d,q)\times(P,D,Q)_s$ model for y_t is

$$\phi_p(B)\,\Phi_P(B^s)W_t = \theta_q(B)\Theta_Q(B^s)e_t \qquad (1)$$

where $W_t = \nabla^d\nabla_s^D y_t^{(\lambda)}$, $y_t^{(\lambda)}$ is an appropriate transformation of y_t, and e_t is normal white noise. $\phi_p(B)$, $\Phi_P(B^s)$, $\theta_q(B)$, and $\Theta_Q(B^s)$ are nonseasonal and seasonal AR and MA polynomials of order p,P, q and Q, respectively. In order to find the optimal model, we will use R to fit all combinations of models with nonseasonal orders $0\le p\le 4$, $d=1$ (for first difference), $0\le q\le 4$, and seasonal orders $0\le P\le 4$, $D=1$ (for seasonal difference), $0\le Q\le 4$, and $s=12$. The 4 models with the *lowest* AIC values are:

ARIMA $(2,1,2)\times(2,1,3)_{12}$	AIC = 1717.015	ARIMA $(3,1,1)\times(2,1,3)_{12}$	AIC = 1717.473
ARIMA $(4,1,3)\times(3,1,1)_{12}$	AIC = 1720.384	ARIMA $(4,1,3)\times(3,1,3)_{12}$	AIC = 1721.304

Our final model for the box-cox transformed series will be an ARIMA $(2,1,2)\times(2,1,3)_{12}$, because it has the *fewest predictors* as well as the *lowest AIC*. The equation for for this model is on the next page.

By equation (1), the equation for an ARIMA $(2,1,2) \times (2,1,3)_{12}$ model is:

$$(\mathbf{1} - \phi_\mathbf{1}\mathbf{B^1} - \phi_\mathbf{2}\mathbf{B^2})(\mathbf{1} - \mathbf{\Phi_1}\mathbf{B^{12}} - \mathbf{\Phi_2}\mathbf{B^{24}})\mathbf{W_t} = (\mathbf{1} - \theta_\mathbf{1}\mathbf{B} - \theta_\mathbf{2}\mathbf{B^2})(\mathbf{1} - \mathbf{\Theta_1}\mathbf{B^{12}} - \mathbf{\Theta_2}\mathbf{B^{24}} - \mathbf{\Theta_3}\mathbf{B^{36}})\mathbf{e_t}$$

Solving for W_t results in:

$$W_t = \phi_1 W_{t-1} + \phi_2 W_{t-2} + \Phi_1 W_{t-12} - \phi_1\Phi_1 W_{t-13} - \phi_2\Phi_1 W_{t-14} + \Phi_2 W_{t-24} - \phi_1\Phi_2 W_{t-25} -$$

$$\phi_2\Phi_2 W_{t-26} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \Theta_1 e_{t-12} + \theta_1\Theta_1 e_{t-13} + \theta_2\Theta_1 e_{t-14} - \Theta_2 e_{t-24} + \theta_1\Theta_2 e_{t-25} +$$

$$\theta_2\Theta_2 e_{t-26} - \Theta_3 e_{t-36} + \theta_1\Theta_3 e_{t-37} + \theta_2\Theta_3 e_{t-38} \qquad where \;\; W_t = \nabla^d \nabla_s^D y_t^{(\lambda)}$$

Fitting the SARIMA model using $\mathbb{R}$, the values of the estimated coefficients are as follows. The standard errors of the coefficients are all relatively small, and all p-values were significant at the 1% level.

Coefficient	ϕ_1	ϕ_2	θ_1	θ_2	Φ_1	Φ_2	Θ_1	Θ_2	Θ_3
Estimate	0.4854	0.2117	-1.3663	0.4006	0.7712	-0.9929	-1.5282	1.5405	-0.7428
St. Error	0.0767	0.0620	0.0924	0.0893	0.0260	0.0014	0.0547	0.0431	0.0411

4 Model Diagnostics: Checking the Assumptions

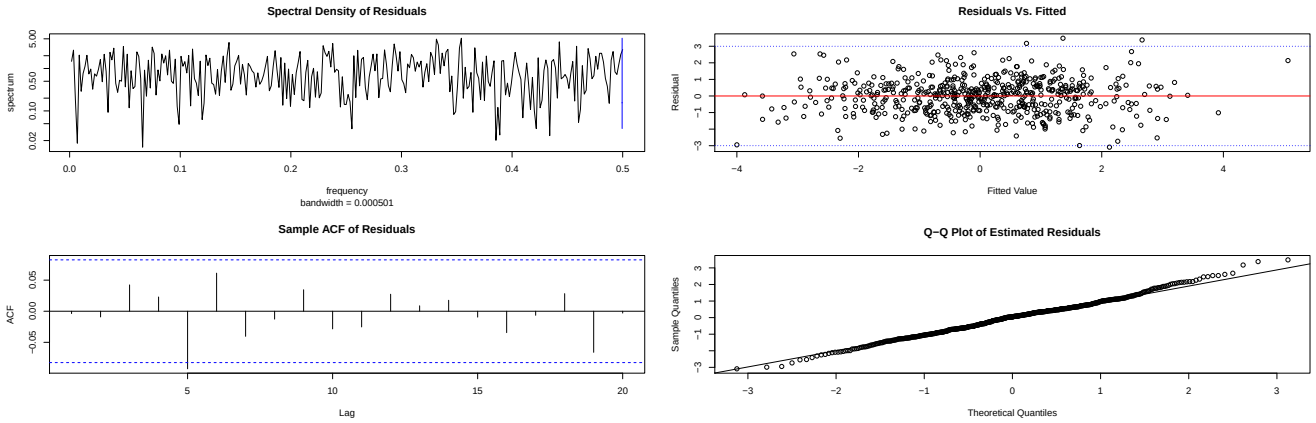

Figure 8: Graphs for model diagnostics. Raw periodogram of Residuals (top left), Residuals Vs Fitted (top right), Sample ACF of Residuals (bottom left) and Normal QQ plot of Residuals (bottom right).

Residual analysis will be used to assess the fit of the model. We must check the assumption that the residuals are *uncorrelated and normally distributed*. The raw periodogram of the residuals does not contain any large signal peaks and has *no apparent pattern*, and the majority of the peaks are within the 95% confidence interval, suggesting that the model may fit well. In addition, there is *no apparent pattern* in the residuals vs. fitted plot, and *very little autocorrelation in the residuals*. The autocorrelation at lag 5 is slightly outside the confidence interval, but this may be due to the fact the the bounds are defined as $\pm\frac{2}{\sqrt{n}}$. P-values for the Ljung-Box statistic for lags 1 through 20 *all faill to reject the null hypothesis* that the error terms are uncorrelated. The Normal QQ plot shows that the residuals follow roughly a normal distribution. This is confirmed by the Shapiro-Wilk test for normality which results in a p-value of 0.1219, failling to reject the null hypothesis that the residuals follow a normal distribution. Overall, it seems that the model satisfies the assumptions and is a good fit.

5 Some Spectral Analysis

In order to provide additional information on whether a seasonal ARIMA model is appropriate for the data, I will compare the sample spectral density of the first and seasonal difference of the transformed data to the theoretical spectral density of the ARIMA $(2,0,2) \times (2,0,3)_{12}$ model using the coefficients provided on the previous page. The sample spectral density was smoothed with Daniell window m = 8.

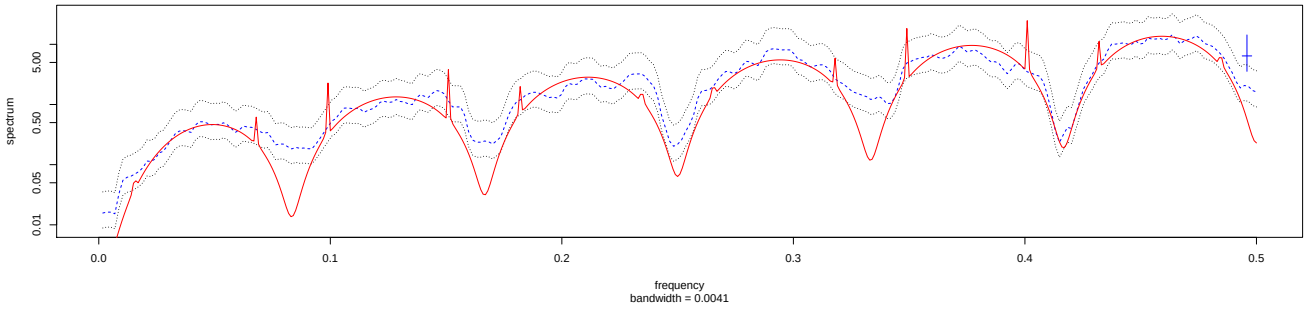

Figure 9: Smoothed sample spectral density (blue-dashed line) with 95% confidence interval. The theoretical spectral density of an ARIMA $(2,0,2) \times (2,0,3)_{12}$ model is overlaid (solid red line).

The theoretical spectral density lies within the 95% confidence interval for a vast majority of the frequencies, except for the troughs at frequecies of multiples of $\frac{1}{12}$. This may be because the sudden drops were covered up by smoothing. These results further confirm that the model is fitting the data nicely.

6 Discussion and Conclusions

My answer to the question posed at the begnning of the report is: *yes, the number of papers published in the field of cancer research over time can be modeled, using a seasonal ARIMA* $(2,1,2) \times (2,1,3)_{12}$ *model on the square root of the counts.* The equation of the final model, along with estimated coefficients, were presented on the previous page. We may use this model to predict the number of cancer research articles that will be published in a given month of a given year. This has several implications.

One practical use is to increase the chances of getting your paper published. Often times, many papers that have quality research are delayed by several months in the publication process. Submitting a paper related to cancer research during a month in which many papers are expected to be published on the topic may increase your chances of being selected. One setback to this interpretation is that we do not have the data for the total number of papers *submitted*. Without this information, we do not know whether or not the real reason less papers are being published is because less are being submitted. If this data can be obtained, future analysis on the *proportion* of papers accepted and published may yield results with better interpretive value. Other factors may include previously discussed seasonal productivity patterns.

6.1 References

Cryer, J.D. and Chan, K. Time Series Analysis With Applications in R. Springer Texts in Statistics. 2008.

Hipel KW, et. al. "Advances in Box-Jenkins Modeling." Water Resources Research Vol. 13 No. 3. (1977).