Analyzing and Forecasting Monthly Red Sea Urchin Landings in Santa Barbara from 2008 to 2019

Callum Weinberg: 174 Final Project

December 3rd 2021; Edited January 2023

Abstract

The red sea urchins is an environmentally and economically important species that is fished and landed around Santa Barbara county. A time series model is developed to describe landings from 2008 to 2018, and predict 2019 landings. The Box-Jenkins methodology was implemented, and a $SARIMA(1,0,1) \times (1,1,1)_{12}$ was fit to the data. The model fit the data well, and predicted 2019 landings with a reasonable degree of accuracy. Future work should extend this model to one that additionally accounts for practical lower bounds on sea urchin hauls.

Introduction

Background

The red sea urchin is an important species in Southern California, both for its economic value and for its role in coastal-marine ecosystems. Between 2005 and 2014 red sea urchins landings for California were valued between 5.1 to 14.9 million dollars per year.[11, Section 2.3.2] Since the 1970s, red sea urchins have been primarily fished off the coast of Mendocino and Sonoma counties in Northern California, and between Santa Barbara County[1] (Point Conception) and San Diego County in Southern California. Santa Barbara Port accounted for over 60 percent of landings between 1971 and 2018.[11, Section 2.4] In 2014, red sea urchin was the species with second highest landings (in pounds) in The Santa Barbara area, at about 6.5 million pounds.[2][2]

Red sea urchins play an important roll in coastal marine ecosystems as well. They live in kelp forests and their abundance and population health affect the kelp and species that depend on the kelp, and also species that depend on the sea urchins for food.[11, Section 1.4]

Red sea urchin landings have been on the decline since 2014.[11, Section 2.3.2] The California Department of Fish and Wildlife (CDFW) attributes this to a combination of weather patterns (El Niño effect in 2015), disease, purple sea urchins out-competing red sea urchins, and over-fishing.[11, Section 2.3.2] Red sea urchin populations have also been declining, although not at the same rate as landings.[11, 1.2.1] CDFW hypothesizes the difference is possibly due to the amount of consumable meat per urchin declining (due to lower kelp levels and more purple sea urchins), so there is less demand for "low quality" urchins.[11, Overview]

Purpose

The purpose of this report is to develop a statistical, time series model that describes red sea urchin landings in the Santa Barbara area, and then to use the model to predict 2019 landings. While

[1] This includes the Channel Islands, which historically have been some of the most productive red sea urchin fishing grounds.[11, Section 2.1]

[2] Squid was the most landings species, at over 70 million pounds in 2014. Besides squid, no other species comes particularly close to red sea urchin in Santa Barbara. The third highest species landings were red rock crab, with just under a million pounds of red rock crab in 2014.[2]

population levels of urchins are a motivation for this analysis, only landings (commercial fishing) data are analyzed and modeled. Therefore the scope of this report is modeling and prediction of landings.

Data

This report makes use of monthly red sea urchin landings for the Santa Barbara region, from 2008 until 2019. 2008 through 2018 are treated as the training data, and 2019 data is used to test the model. The data is recorded by CDFW and is available in table 12 of each annual report.[3] Landings are measured and reported in pounds.

Methodology

The time series model in this report makes use of the Box-Jenkins methodology to choose a model.[3] Box-Jenkins Models are built from autoregressive terms and moving average terms. A seasonal autoregressive integrated moving average (SARIMA) model can be developed given the data is stationary (the seasonal portion models any seasonal trend, and the integrated portion models any non-zero linear trend).[8, 6.4.4.5] No covariates are considered with this methodology, rather a model is built only from monthly red sea urchin landings.

Data Processing

The monthly landings data was retrieved from Table 12 of each annual landings report[3] and entered into a csv file. Minimal data cleaning was required: a date variable was created for each month and year, and the landings-weight data was used as reported by CDFW.

Statistical Code and Software

R and RStudio software are used to perform the analysis and create the visuals shown in this report. Appendix C includes a copy of the R code used in this report. Appendix B includes a description of some key aspects of the code, as referenced in the following sections. The following packages were used in addition to base R: tidyr[14], dplyr[15], knitr[17], lubridate[6], ggplot2[13], qpcR[9], forecast[7], MASS[12], cowplot[16], GeneCycle[1], and TSA[4].

Results

A $SARIMA(1,0,1)\times(1,1,1)_{12}$ model was determined to be the best model after seasonally differencing the data. The residuals of the model are uncorrelated and roughly Gaussian. Spectral analysis also suggests the model is appropriate. The model forecasts 2019 red sea urchin landings well. While the actual 2019 landings are inside the forecast's confidence interval, the values are negative (and the confidence interval covers negative values).

Model Development and Evaluation

2008 to 2018 Monthly Landing Data

Figure 1 below shows the monthly landings of red sea urchin, in thousands of pounds, for 2008 through 2018 (the 2019 data is reserved to test the model, and can be seen in figure 9 below). The landings appear roughly flat between 2008 and 2014, and then clearly decrease starting around 2014, as CDFW reported[11]. There may be some periodicity to the data as well. Landings are lowest around late winter or early spring, and then increase until November. Figure A1 in Appendix A shows a periodogram of the data. The periodogram indicates any cycles in a time series, with higher values indicating a period. The frequency .8 has a high periodogram in Figure 2, and a frequency of .8 corresponds to a period of approximately 12 months. This confirms that there is likely an annual seasonal component to the data.

[3] PSTAT 274 final project required the use of the Box-Jenkins approach, and other time-series models were not considered.

The time series in Figure 1 does not show any signs of sharp or non-symmetric behavior. Such a series can be difficult to model using the Box-Jenkins methodology. Finally, the variance looks somewhat constant for the series. This is considered in more detail in the next section.

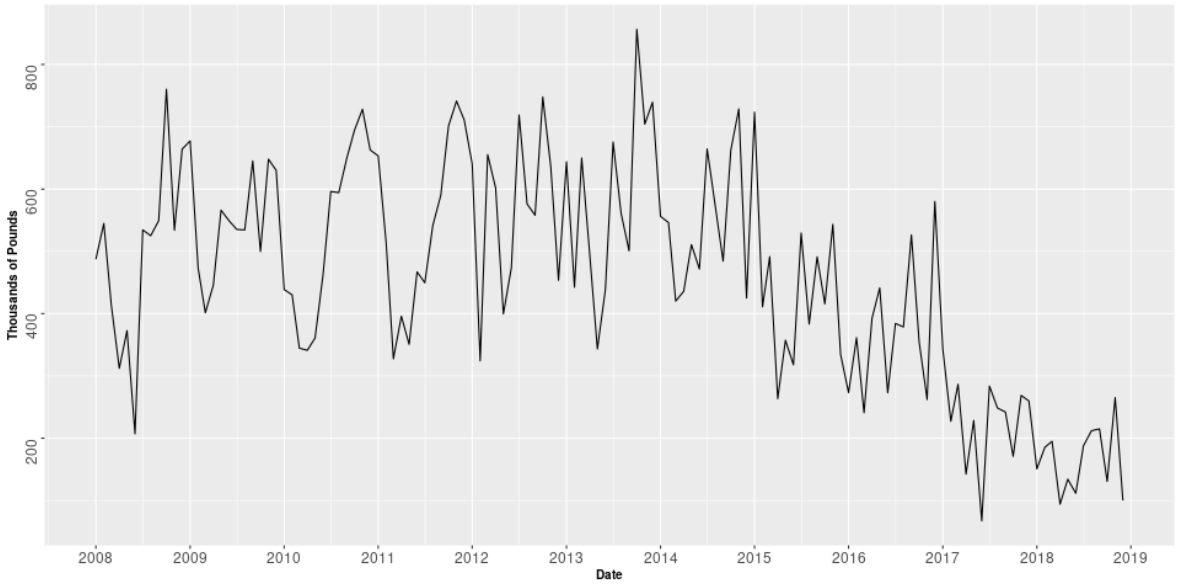

Figure 1: Red sea urchin landings in the Santa Barbara area, measured in thousands of pounds. Monthly, 2008 to 2018.

Transforming the Data

First, it is considered whether the data needs to be transformed to make the variance stationary. Log and square root transformations were considered and did not improve any possible non-constant variance. A Box-Cox transformation was considered as well. The result was that no transformation was needed, and that the data is roughly normally distributed. See Appendix A2 for the results of the Box-Cox transformation. Figure 2 shows a histogram of the raw, 2008 to 2018 data. This distribution is roughly normally distributed, and is not particularly skewed and does not have any clear outliers. The variance of this time series is 3.14×10^{10}.

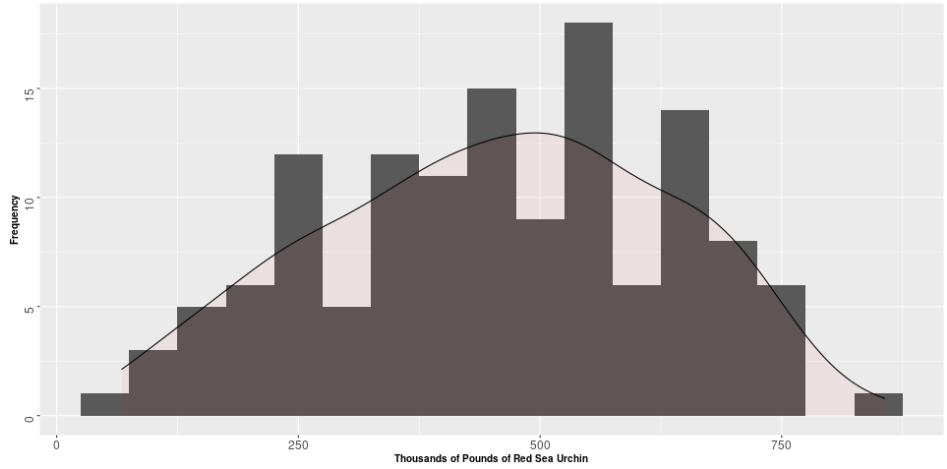

Figure 2: Histogram of the monthly red sea urchin landings (in Thousands of Pounds)), 2008-2018.

Given the data shows both trend and seasonality, some form of differencing must be implemented to make the data stationary. Appendix A3 shows the ACF (autocorrelation function) for the original

time series, which clearly indicates non stationary data (the ACF decays slowly and appears periodic).

Three options were explored to make the data stationary. A) Differencing at lag 1 once, B) differencing at lag 12 and then at lag 1, and C) Only differencing at lag 12. Option B was found to increase the overall variance of the data to 3.41×10^{10}, and preliminary models showed signs of overdifferencing. Option A was considered in more detail. Ultimately, the modeler decided that option C) was most appropriate given the time series has a clear period (see Figure A1) and that a stationary time series could be achieved just from option 3. See appendix B1 for a more detailed discussion of this modeling decision.

After differencing at lag 12 to remove seasonality, the variance was reduced to 2.03×10^{10} and the data appears Gaussian (see Figure 3, right). Additionally the seasonal non-stationarity has been removed (see Figure 3, left). As indicated by the trend-line in Figure 3, there is possibly still some trend non-stationarity. However, as discussed in the next section, the ACF and Partial ACF (PACF) both appear to be non-stationary and match a seasonal model. Therefore this data will be assumed stationary for the remainder of this analysis section.

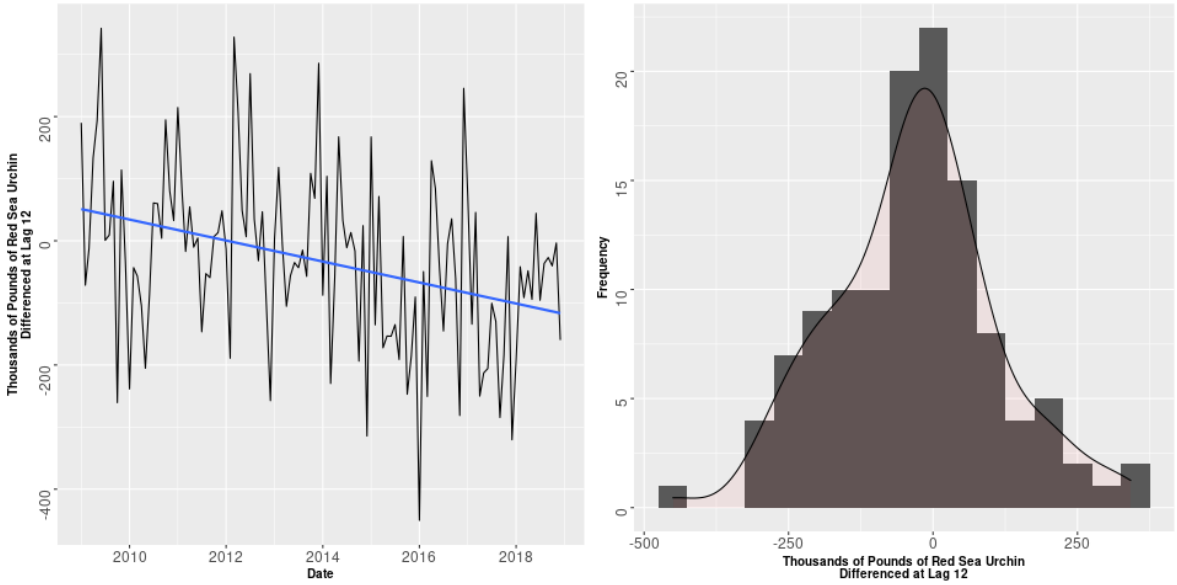

Figure 3: Left: 2008 to 2018 red sea urchin landings after differencing at lag 12. Right: Histogram of the differenced data.

ACF and Partial ACF

Figure 4 shows the ACF and PACF of the differenced sea urchin landings data. The ACF indicates autocorrelations at a given lag. The PACF indicates partial autocorrelations at a given lag. The blue-dotted lines are 95 percent confidence intervals for the ACFs and PACFs. The typical interpretation is that if an ACF or PACF is within the confidence interval, the ACF and PACF are not significant at that given lag.[4]

Note that as compared with A3, Figure 4 (left) does not show slow decay in autocorrelations. Instead ACFs at lags 2 and 3 may be significant, with ACFs at lags 15 and 21 possibly significant. ACF at lag 12 is very likely significant, and suggests a seasonal model is appropriate. No ACFs appear significant after lag 21.

The PACF produces a similar result. The PACF at lags 2 and 3 may be significant. The PACF at lag 12 is very likely significant, as is the PACF at lag 15. There are a few more larger lags outside the confidence interval. Since the confidence interval is 95 percent, it would be expected that if this process was repeated (for similar data) many times, one out of every twenty PACFs that are in fact

[4]The ACF and PACF at lag 0 are always 1, and they are not shown in the graph as it makes the visual harder to interpret.

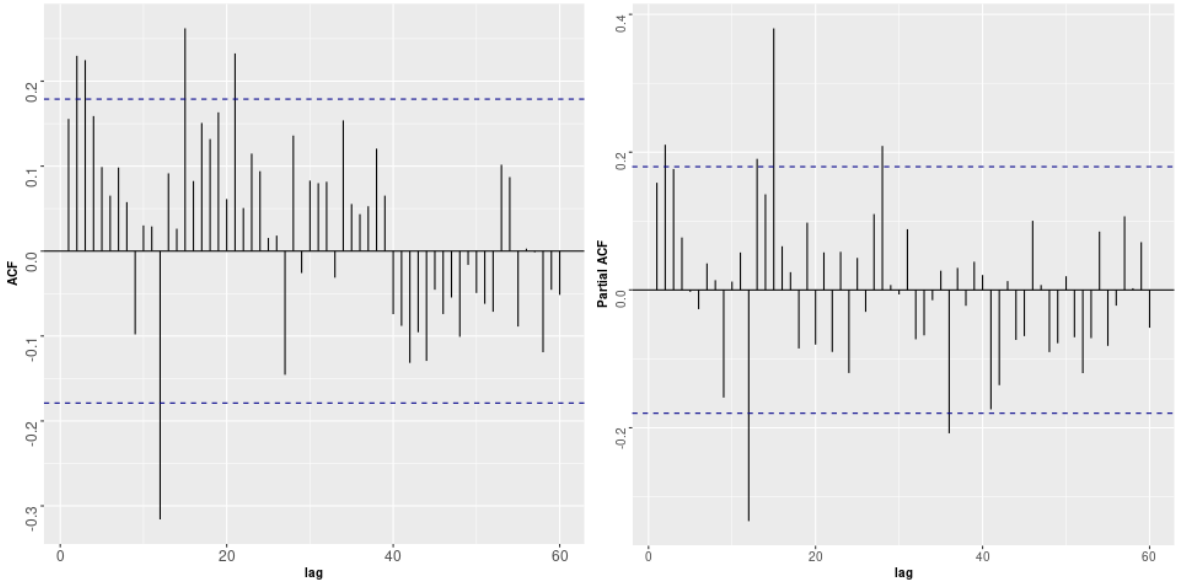

Figure 4: Left: ACF for time series after differencing at lag 12. Right: PACF for time series after differencing at lag 12.

not significantly different than zero appear outside the confidence interval. In other words, it is not unusual that a few higher lag-PACFs (or ACFs) would be outside the confident interval, and are likely not indicative of the actual correlations in the time series.

For constructing the time series model, the ACF is indicative of the degree of moving average terms that should be used in the model (denoted q), while the PACF is indicative of the number of autoregressive terms that should be used (denoted p). Additionally, given a seasonal model is considered, the seasonal portion has its own autoregressive and moving average terms, denoted (P) and (Q) respectively.

Both the ACF and PACF indicate that there is significant autocorrelation/partial autocorrelation at lag 12, but not at further multiples of 12. Therefore $P = 1$ and $Q = 1$ are good choices for fitting the model. Because the ACF/PACF at lags 15 appear significant, $p = 3$ and $q = 3$ seem like good initial choices for the number of autoregressive and moving average terms ($12 + 3 = 15$). This is supported by ACF lags 2 and 3 and PACF lag 2 appearing significant. $p = 2$ and $q = 2$ will be considered as well in selecting models to fit. The result is the initial model to attempt to be fitted is a $SARIMA(3, 0, 3) \times (1, 1, 1)_{12}$ model.

Fitting Models

Given X_t corresponds to red sea urchin landings at time t. Then a SARIMA model can be expressed in the following form:

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D X_t = \theta(B)\Theta(B^s)Z_t; \;\; Z_t \sim WN(0, \sigma_Z^2)$$

For which B represents the backshift operator (such that $B^k X_t = X_{t-k}$ where k is the lag difference between times t and another time, s). $\phi()$ corresponds to the autoregressive terms, $\Phi()$ corresponds to the seasonal autoregressive terms, $\theta()$ corresponds to the moving average terms, and $\Theta()$ corresponds to the seasonal moving average terms. s corresponds to the seasonal difference, which is 12 since the data was differenced at lag 12. d is the number of times the data was differenced at lag 1, which is 0 for this seasonal-only differenced data. D is the number of times the data was differenced at lag s, which is 1. Finally, Z_t is represents a White-Noise random variable at time t with mean 0 and variance σ_Z^2.

A number of different models were initially considered through a process of making changes to the

$SARIMA(3,0,3) \times (1,1,1)_{12}$ model. Candidate are not mentioned or explored in more detail if they were found to be non-invertible or non-stationary. The $SARIMA(3,0,3) \times (1,1,1)_{12}$ was itself found to be non-invertible, given $\Theta_1 = 1$. Considering the ACF and PACF again, it is not clear lag 3 is significant, so $p = 2$ and $q = 2$ were were considered as well, along with variations.

Three models were found to be suitable candidate models. Models 1 and 2 have terms fixed to 0.[5]

- Model1: $SARIMA(2,0,2) \times (1,1,1)_{12}$ with $\theta_1 = 0$
- Model2: $SARIMA(2,0,3) \times (1,1,1)_{12}$ with $\phi_1 = \theta_1 = \theta_2 = 0$
- Model3: $SARIMA(1,0,1) \times (1,1,1)_{12}$

Table 1 shows the coefficients for each of the three models. The coefficients for the terms of the model are estimated using Maximum Likelihood estimation.[6] Table 2 shows the values of the complex roots related to each term. All of the roots need to be greater than 1.[7] for the model to be stationary and invertible. All of the roots are greater than 1, though it is worth noting that models 1 and 3 have roots very close to 1, which all else equal would suggest model 2 is preferable.

Model	ϕ_1	ϕ_2	θ_1	θ_2	θ_3	Φ_1	Θ_1
Model 1	-.2099	-.7849	NA	-.6333	NA	-.0408	-.8901
Model 2	NA	-.3778	NA	NA	.3506	.1782	-.3802
Model 3	-.9970	NA	-.7925	NA	NA	-.04517	-.8909

Table 1: Estimated Model Coefficients for Models 1, 2, and 3.

Model	$z^*_{\phi_1}$	$z^*_{\phi_2}$	$z^*_{\theta_1}$	$z^*_{\theta_2}$	$z^*_{\theta_3}$	$z^*_{\Phi_1}$	$z^*_{\Theta_1}$
Model 1	1.002918	1.270341	1.256595	1.256595	NA	24.509804	1.123469
Model 2	1.626931	1.626931	1.418173	1.418173	1.418173	5.611672	2.630195
Model 3	1.003009	NA	1.261830	NA	NA	23.980815	1.122460

Table 2: Value of Roots (z^*) corresponding to each coefficient for Models 1, 2, and 3

Akaike's Information Criterion (AIC)[8] is a diagnostic which is used to compare the quality of models (that come from the same data). It measures the fit of the model, while penalizing for more terms being included in the model. The AIC for Model 1 is 3144.118, the AIC for Model 2 is 3166.083, and the AIC for Model 3 is 2885.948.

Diagnostic Checking

Along with AIC, a number of other diagnostic checks are performed on these three models to determine whether they are an appropriate fit, and if so which model is optimal. Specifically, the residuals of the model should be similar to Gaussian white noise (that is, uncorrelated). Two approaches are used to determine whether the residuals are appropriate. First, ACF, PACF, histogram, and quantile-quantile plots of the residuals are used to determine if the residuals behave like white-noise (the first two plots) and are approximately Gaussian (the second two plots). The residual plots for Model 3 are shown in Figure 5, and the residual plots for Models 1 and 2 are shown in Figure 6.

[5] In appendix C, Model 1 is referred to as Model40, Model 2 is referred to as Model44, and Model 3 is referred to as Model43.

[6] The built-in R package "stats" has a function called arima(), which was used here for coefficient estimation for the SARIMA models.

[7] For a complex root, $z^* = x + iy$, the condition $|z^*| = |\sqrt{x^2 + y^2}| > 1$ is the equivalent of this. condition[5, Slide 24]

[8] Technically Akaike's Corrected Information Criterion is implemented here, and in the statistical code this is done with the AICc() function from the qpcR package[9].

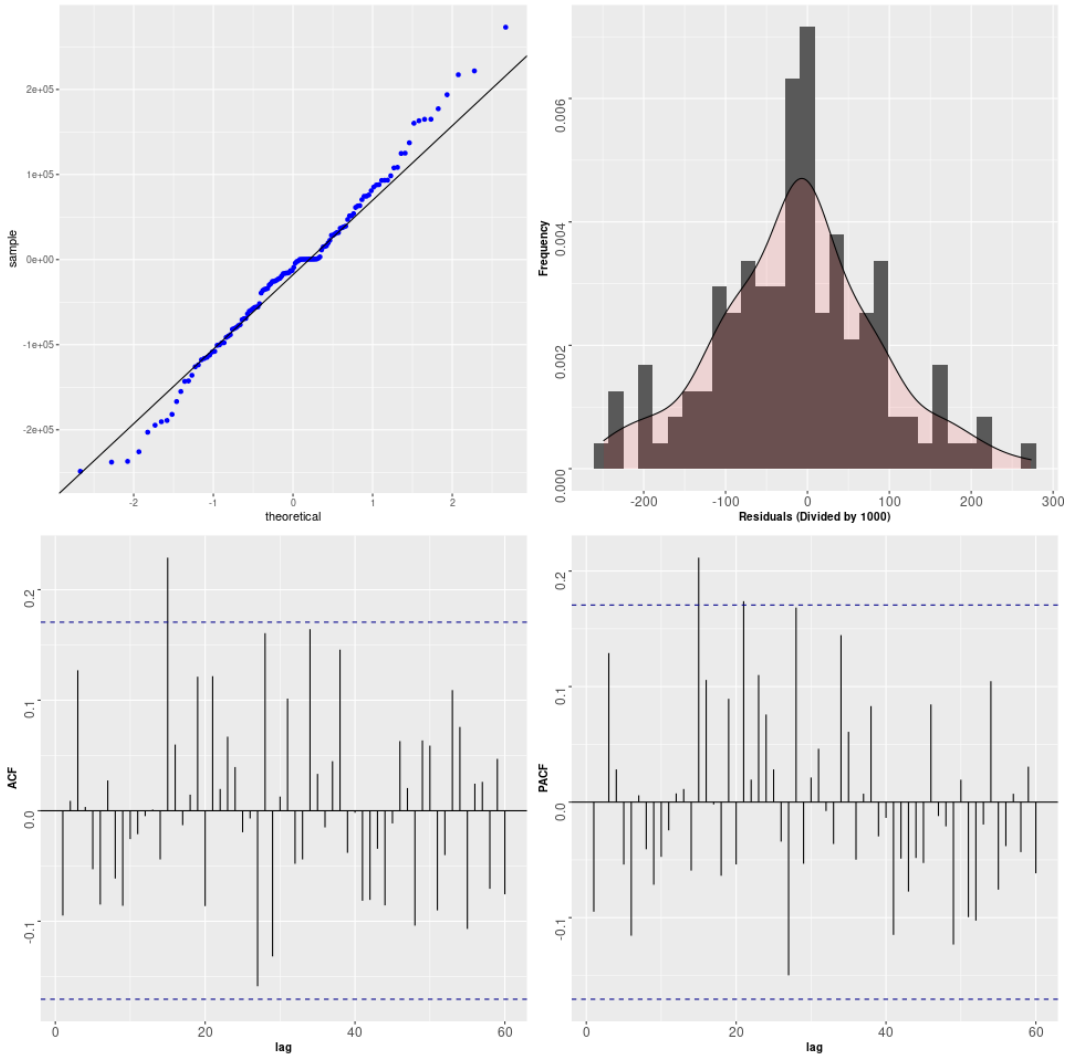

Figure 5: Left: ACF for time series after differencing at lag 12. Right: PACF for time series after differencing at lag 12.

In Figure 5, the residual plots for Model 3 indicate the residuals are approximately Gaussian. There is some slightly non-Gaussian behavior in the tails (potential outliers), but ultimately the model is a reasonable fit. The ACF and PACF show non-significant lags mostly, except for lag 15 for both graphs.

Figure 6 shows the residuals plots for Model 1 are fairly similar to those of Model 3. The quantile-quantile plot is a slightly worse fit, but otherwise the graphs are similar. Model 2 is a worse fit than Models 1 and 3 based on the residual plots. The quantile-quantile plot shows more deviation from normality, and the autocorrelation and partial autocorrelation at lag 15 are larger and significant.

Along with the graphical analysis, a number of hypothesis tests are performed to check for linear and non-linear independence in the residuals (the residuals should be independent to meet the white-noise criteria). Box-Pierce and Ljung-Box tests are performed, with the null hypothesis being the data is white noise. McLeod-Li tests for non-linear dependence. Finally a Shapiro-Wilk test is performed to test for normality of the residuals. The results of the tests are in Table 3.

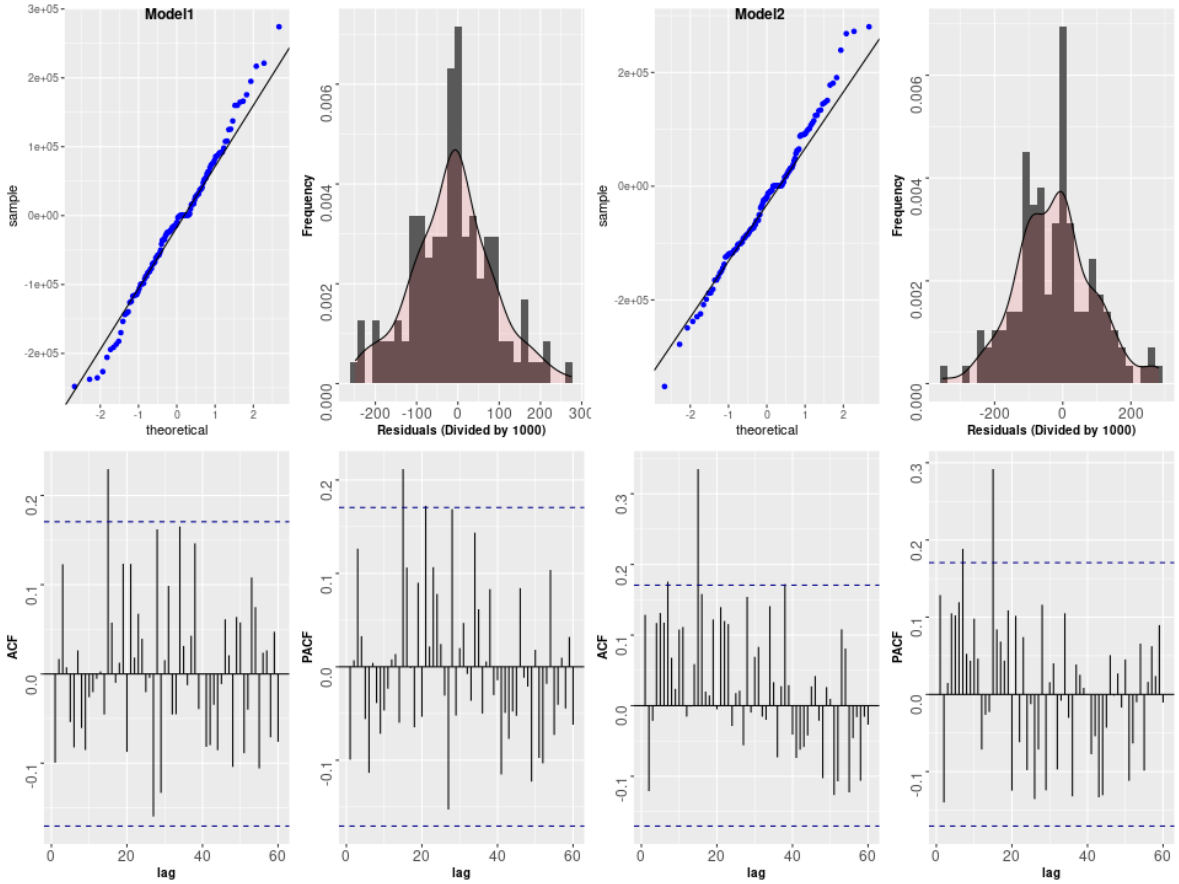

Figure 6: Left 4 Plots: quantile-quantile Plot, Histogram, ACF, and PACF for Model 1. Right 4 Plots: quantile-quantile Plot, Histogram, ACF, and PACF for Model 2.

Model	Box-Pierce	Box-Ljung	Mcleod-Li	Shapiro-Wilk
Model 1	0.3902	0.3560	0.6268	0.4417
Model 2	0.01192	0.0078	0.3555	0.4370
Model 3	0.4973	0.4596	0.6175	0.4443

Table 3: Results of diagnostic tests. All reported values are p-values. Statistical code for these tests can be found in Appendix C.

Based on the results of the McLeod-Li test, none of the models show any signs of non-linear dependence. However he Box-Pierce and Box-Ljung tests indicate the White Noise assumption is inappropriate for Model 2. The Shapiro-Wilk test is not significant for any of the Models, and therefore no evidence of non-Gaussian residuals is found.

Selected Model

Visual inspection of the residuals (see Figures 5 and 6) and the results of the diagnostic tests (see Table 3) suggest that Model 1 or Model 3 should be chosen. Model 3 has lower AIC than Model 1. Additionally, Model 3 has one less term for the model than Model 1 (all else equal, simpler models are preferable). Therefore Model 1 is the chosen model for fitting the red sea urchin time series data, and for forecasting 2019 data.

Model 3 is defined as follows:

$$(1 - .970B)(1 - .0452B^{12})(1 - B^{12})X_t = (1 - 0.7925B)(1 - 0.8909B^{12})Z_t; \;\; Z_t \sim WN(0, \sigma_Z^2)$$

This model appears mostly satisfactory. There is a slight issue in that there appears to be a significant autocorrelation/partial autocorrelation for the residuals at lag 15. However it was the best model developed, and will be used for forecasting. Additionally, it should be noted that this model only contains terms related to $p = 1$ and $q = 1$. So even though it was the best fitting model, it is somewhat difference than what was expected based on the ACF and PACF.

Spectral Analysis of Model 3

Spectral analysis of the model is considered. A seasonal time series can be expressed as a summation of sine and cosine waves, and periodogram indicates what periods are found in the data. A periodogram was already analyzed when considering whether seasonal differencing was appropriate for the data set, see Appendix A1. There is a spike at frequency .0815, suggesting a period of roughly 12 months. As mentioned before, there seems to be an annual period, which is not surprising given the data is monthly. Additionally, there is a smaller spike at frequency .015, which corresponds to a period of 67.5 months or 5.5 years. Given the training data is only 11 years long, it is difficult to assess the significance of this period.

Spectral analysis is also applied to the residuals of Model 3. If the residuals are White Noise, no dominant period should be found in the periodogram. Figure 7 shows the periodogram for the residuals of Model 3. There is no clear dominant period for the residuals, supporting the conclusion that the residuals are White Noise.

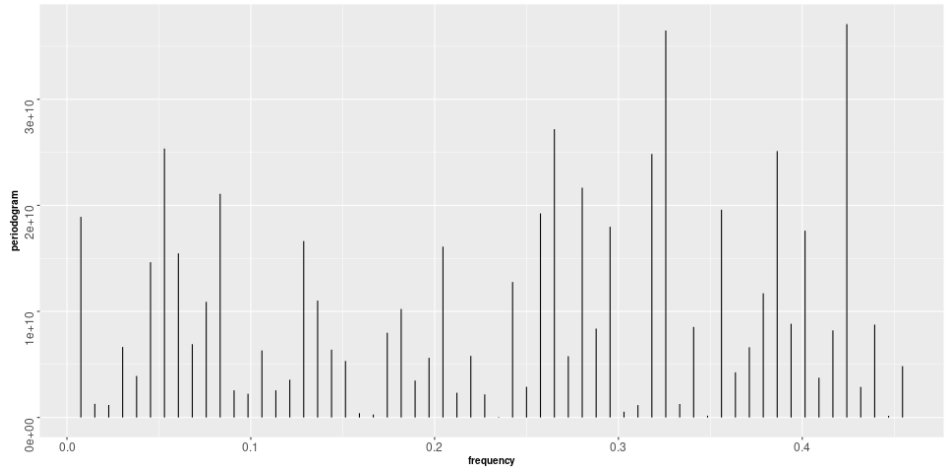

Figure 7: Periodogram of the residuals of Model 3.

A Fisher's test for periodicity is performed. The null hypothesis is that the data is White Noise, and the alternative hypothesis is that the data is not White Noise. The test results in a p-value of .833 for the residuals of Model 3[9], providing more evidence that the residuals are White Noise. Finally, a Kolmogorov-Smirmov Test for Periodicity is performed. The results are shown in Figure 8. The cumulative periodogram is entirely within the boundaries: therefore the null hypothesis is not rejected, which again provides evidence the residuals of Model 3 are White Noise.

[9]See Appendix C

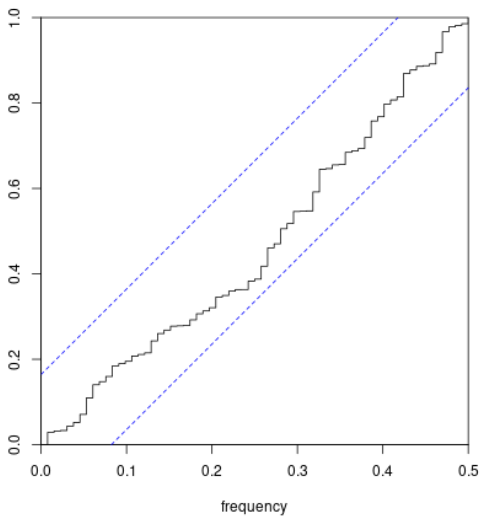

Figure 8: Kolmogorov-Smirmov Test for the residuals of Model 3.

The result of the spectral analysis is that there is clearly periodicity in the red sea urchin landings time series, and there is no periodicity in the residuals of the model chosen to forecast the data.

Forecasting 2019 Landings

Model 3 is used to forecast the 2019 landings of red sea urchins in Santa Barbara. The forecasted data is shown in Figure 9 with 95 percent confidence intervals Figure 9 includes the real 2019 data as well.

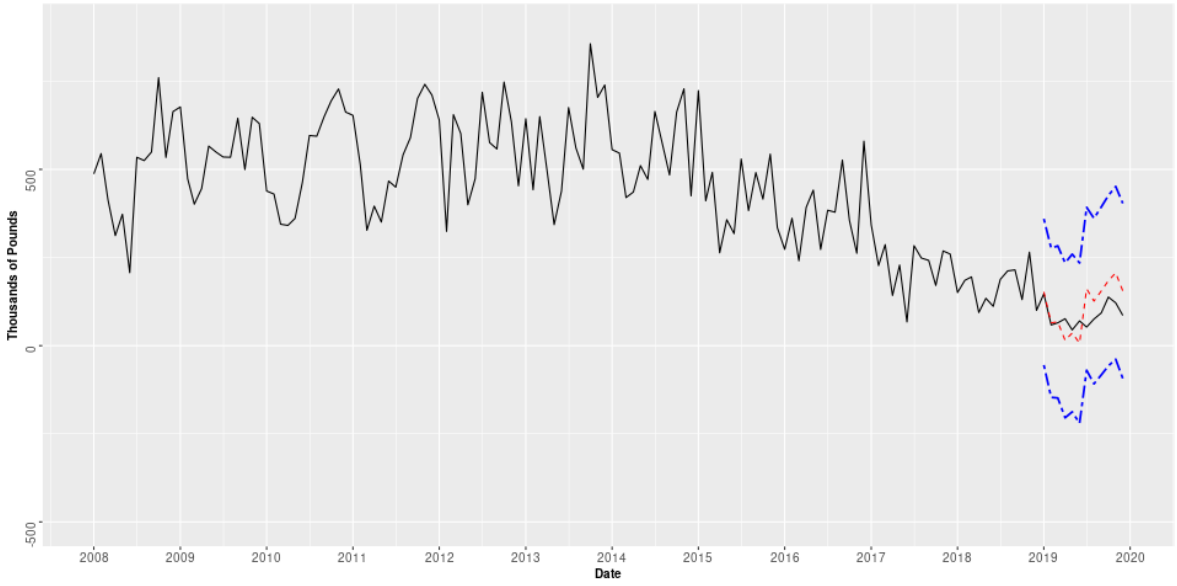

Figure 9: Black line indicates actual red-sea-urchin landings. Red line indicates the predicted landings for 2019. The blue lines are the bounds for the 95 percent confidence interval for the prediction.

As can be seen in Figure 9, the forecasted halls track closely to the actual hauls in 2019. Noticeably, there is a dip in the first part of 2019 which the forecast predicts (although it predicts it to be closer to 0 than the actual observed numbers), and then predicts a rise in sea urchin hauls in the second

half of the year, which also occurs in the real data. The 95% confidence intervals do prove limited in use-fullness. In particular, the bottom of the prediction is logically 0 (cannot have negative hauls). None of the predicted monthly hauls are negative (the smallest is June, for which the prediction is about 7 thousand pounds). Still, the confidence interval indicates an issue with the choice of model, namely negative predictions are possible in this SARIMA fit.

Conclusion

The purpose of this report was to first develop a time series model of red sea urchin landings in Santa Barbara using the Box-Jenkins methodology and then use that model to forecast 2019 landings. The model development was successful, and the forecast is proven predictive. But improvements can be made.

The fitted model is a $SARIMA(1,0,1) \times (1,1,1)_{12}$ model. With the coefficients estimated by the Maximum Likelihood method, the model is defined as:

$$(1 - .970B)(1 - .0452B^{12})(1 - B^{12})X_t = (1 - 0.7925B)(1 - 0.8909B^{12})Z_t; \quad Z_t \sim WN(0, \sigma_Z^2)$$

The model satisfied diagnostic tests, spectral analysis tests, and was the best fitting of all the models considered. The seasonal terms were as expected based on the ACF and PACF of the seasonally-differenced data, although the non-seasonal terms were slightly different[10].

The forecast performed well. However, since hauls can not be less than 0, a better forecast would involve some sort of decay in rate-of-decline as the landings value approaches zero. Future work should focus on a modeling approach that can account for this phenomena.

I would like to acknowledge the help I received on this project from Professor Feldman, who taught PSTAT 274 (in which I am enrolled in the Fall of 2021). I had a meeting with Professor Feldman in which she helped me interpret the seasonal differencing and select an initial SARIMA model based on the ACF and PACF of the differenced data. I would also like to thank Laurel Abowd (Master's student in Bren School of Environmental Science and Management, Class of 2022) who gave me the idea to look at fisheries data for this report. I also bounced modeling ideas off her.

This report was written as the final project for PSTAT 274 for the Fall of 2021. All information on the Box-Jenkins Methodology, Spectral Analysis, and time series modeling comes from information provided by the course, Professor Feldman, and Professor Feldman's materials unless otherwise specified above.

Note: this report was re-edited and republished on my Github page in January 2023, to account for some coding errors. The majority of the work did not change (including the choice of models), but the model fits changed slightly and Tables 1, 2, and 3 and Figures 5, 6, 7, 8, and 9 all changed. The major change was fixing an mistake the Author made in the use of the R arima() function. The mistake was indicated by Professor Feldman to the Author after submission of the report.

[10] We were expecting $p = 2, 3$ and $q = 2, 3$

References

[1] Miika Ahdesmaki, Konstantinos Fokianos, and Korbinian Strimmer. *GeneCycle: Identification of Periodically Expressed Genes*, 2021. R package version 1.1.5.

[2] CDFW. Table 12 - monthly landings in pounds in the santa barbara area during 2014. Technical report, California Department of Fish and Wildlife, 2015. Red sea urchin and red rock crab hauls appear on page 4. Squid hauls appear on page 5. Available at $https://nrm.dfg.ca.gov/FileHandler.ashx?DocumentID = 105680\&inline$.

[3] CDFW. Final california commercial landings. Technical report, California Department of Fish and Wildlife, 2019. Table 12 contains data for Santa Barbara area landings for each of the years. The data is available at https://wildlife.ca.gov/Fishing/Commercial/Landings.

[4] Kung-Sik Chan and Brian Ripley. *TSA: Time Series Analysis*, 2020. R package version 1.3.

[5] Raya Feldman. week 2 slides pstat 174/274. Technical report, University of California Santa Barbara, 2021. Available through instruction in PSTAT 174/274 at UCSB.

[6] Garrett Grolemund and Hadley Wickham. Dates and times made easy with lubridate. *Journal of Statistical Software*, 40(3):1–25, 2011.

[7] Rob J Hyndman and Yeasmin Khandakar. Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, 26(3):1–22, 2008.

[8] Jack Prins. *NIST/SEMATECH e-Handbook of Statistical Methods*. NIST/SEMATECH, 2013. Available at https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc445.htm.

[9] Andrej-Nikolai Spiess. *qpcR: Modelling and Analysis of Real-Time PCR Data*, 2018. R package version 1.4-1.

[10] Stackoverflow. *qqnorm and qqline in ggplot2: Stackoverflow*, 2011. https://stackoverflow.com/questions/4357031/qqnorm-and-qqline-in-ggplot2'.

[11] Derek Stein. Red sea urchin, mesocentrotus franciscanus, enhanced status report. Technical report, California Department of Fish and Wildlife, 2019. https://marinespecies.wildlife.ca.gov/red-sea-urchin/undefined/; https://marinespecies.wildlife.ca.gov/red-sea-urchin/the-fishery/; https://marinespecies.wildlife.ca.gov/red-sea-urchin/the-species/.

[12] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.

[13] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[14] Hadley Wickham. *tidyr: Tidy Messy Data*, 2021. R package version 1.1.3.

[15] Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *dplyr: A Grammar of Data Manipulation*, 2021. R package version 1.0.6.

[16] Claus O. Wilke. *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*, 2020. R package version 1.1.1.

[17] Yihui Xie. knitr: A comprehensive tool for reproducible research in R. In Victoria Stodden, Friedrich Leisch, and Roger D. Peng, editors, *Implementing Reproducible Computational Research*. Chapman and Hall/CRC, 2014. ISBN 978-1466561595.

Appendix A

A1: Periodogram of Data

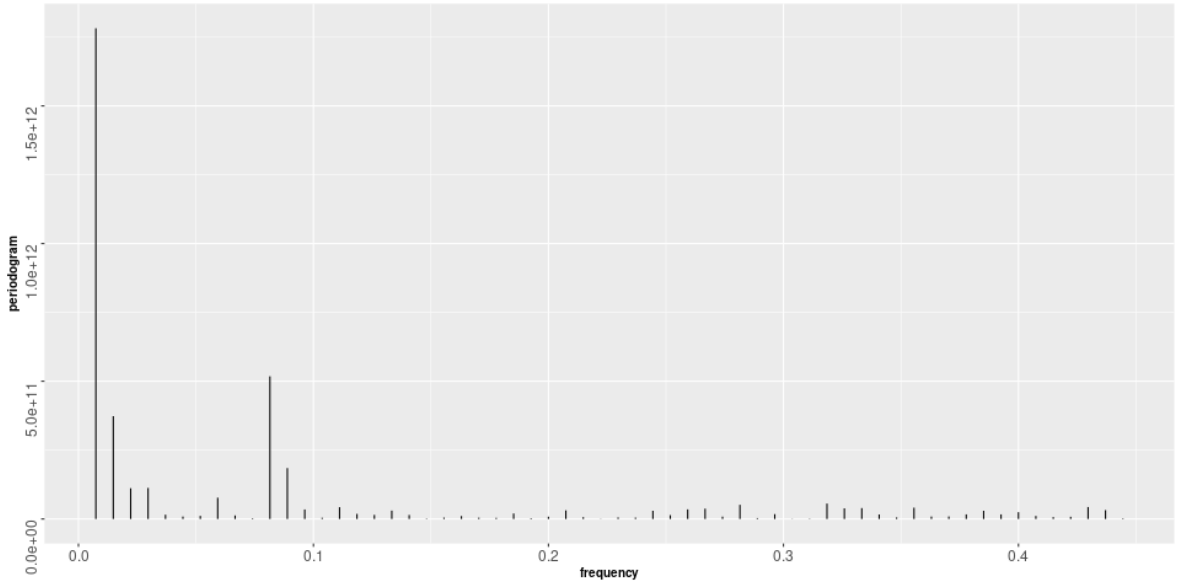

Figure 10: Periodogram of red sea urchin landings in the Santa Barbara area from 2008 to 2018. There is a clear spike around a frequency of .8, which corresponds to a period of 1.2. For Annual Data this indicates seasonality (12 months in the year).

A2: Box Cox Transformation

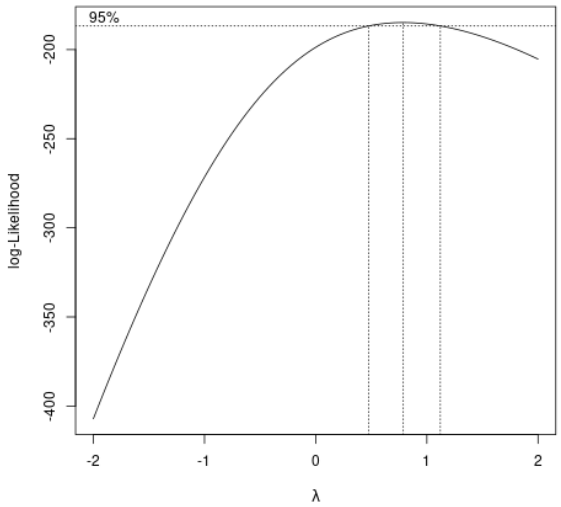

Figure 11: The 95 percent confidence interval for the optimal λ for the Box-Cox transformation includes 1, indicating a transformation is not necessary, and the data is approximately Gaussian.

A3: ACF of Original 2008-2018 Data

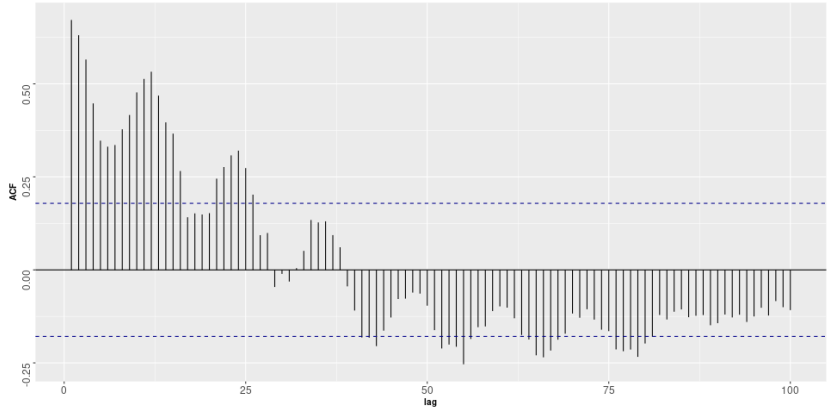

Figure 12: The ACF decays slowly and shows a clear period, indicating differencing is necessary, and that seasonal differencing may be appropriate.

Appendix B

The statistical code for this report is provided in Appendix C. The code for this report is comprised of 7 files (and one csv file containing the landings data). 01_*Clean_Data_Plot* loads the data from the csv file, performs some basic data cleaning, and plots the time series. Additionally an ACF of the original data series is plotted.

02_$Variance_Stabilization_and_Differencing$ determines whether a transformation of the data is necessary to remove heteroskedasticity (it was determined that it was not necssary). Then different differencing strategies are considered. The final choice was to only difference at lag 12. Differencing at lag 12 then lag 1 led to overdifferencing, which was explored in a few of the models in the fourth file. Differencing at lag 1 only led to a model that appeared stationary at first. Ultimately though, it was difficult to fit a model to it. An AR(15,1,0) model was fit and was found to be stationary and invertible (the code for this is in the fourth file). Ultimately, a modeling decision was made that the periodicity needed to be accounted for in the model, and a complex AR(15,1,0) model was not a good substitute (the decision might have been different if the model was much less complex).

03_ACF_PACF_Plots includes code to plot the ACFs and PACFs of the differenced data for all three differencing options.

04_$Fitting_Models$ includes the code to fit the models. A number of models were attempted for the different differencing strategies. The code was left in for the trend difference data and seasonal-trend differenced data to demonstrate that these models were considered (Model 22 was the AR(15,1,0) model. However, only the seasonal-only differenced models were fully vetted.

05_*Diagnostic_Tests* Performs the diagnostic tests and plots the residual plots. As with 04, some diagnostics were tested for some of the other differencing models, but the code is not fully complete for those models. Additionally, the plotting QQ plot in GGplot borrowed an idea from this source on Stack Overflow [10].

06_*Forecasting* includes code to forecast Model 3, as well as Models 1 and 2. The AR(15,1,0) model from the the lag-1 only differenced model is plotted here as well.

07_*Spectral_Analysis* Includes code to plot the periodogram of the original data, and the spectral analysis tests for the residuals.

Appendix C

The code for this project can be found on my Github page. Please visit https://leoncw.github.io/.