Spring 2024

Time Series Modeling to Ascertain Age in Fisheries Management

Kathleen Sue Kirch
Old Dominion University

TIME SERIES MODELING TO ASCERTAIN AGE IN FISHERIES MANAGEMENT

by

Kathleen Sue Kirch
B.S. May 2012, Old Dominion University
M.S. May 2015, Old Dominion University
M.S. December 2016, Old Dominion University

A Dissertation Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

OCEANOGRAPHY

OLD DOMINION UNIVERSITY
May 2024

Approved by:

John Klinck (Director)

Cynthia Jones (Member)

Norou Diawara (Member)

Kent Carpenter (Member)

ABSTRACT

TIME SERIES MODELING TO ASCERTAIN AGE IN FISHERIES MANAGEMENT

Kathleen Sue Kirch
Old Dominion University, 2024
Director: Dr. John Klinck

The ability to assign accurate ages of fish is important to fisheries management. Accurate ageing allows for the most reliable age-based models to be used to support sustainability and maximize economic benefit. Structures used to age include bones, scales, and most commonly ear bones (otoliths). Assigning age relies on validating putative annual marks by evaluating accretional material laid down in patterns in fish otoliths, typically by marginal increment analysis. These patterns often take the shape of a sawtooth wave with an abrupt drop in accretion yearly to form an annual band and are typically validated qualitatively. Researchers have shown keen interest in modeling marginal increments to verify the marks do, in fact, occur yearly. However, it has been challenging in finding the best model to predict this sawtooth wave pattern. I propose three new applications of time series models to validate the existence of the yearly sawtooth wave patterned data: autoregressive integrated moving average (ARIMA), unobserved components model (UCM) and copula. These methods are expected to enable the identification of yearly patterns in accretion. ARIMA and UCM account for the dependence of observations and error, while copula incorporates a variety of marginal distributions and dependence structure. Results indicate that all three models are valid to predict annuli formation. ARIMA works best with a sharp, distinct sawtooth wave while copula is best for data without the sharp drop.

This dissertation is dedicated to Isabelle and Brett Kirch for all their support and encouragement.

ACKNOWLEDGEMENTS

I would first like to thank my committee Chair, Dr. John Klinck. I appreciate the leadership and support. Special thanks to Dr. Cynthia Jones. Without her constant support and encouragement, this dissertation would not have been possible. She has been my motivating champion for my success, from raising my confidence to invaluable guidance and support. My gratitude goes to you. Many thanks to Dr. Norou Diawara. I have enjoyed learning so much on the statistical concepts. I would also like to thank Dr. Kent Carpenter. His assistance has been invaluable. I feel privileged to have him as part of my dissertation. While overseas, he has taken precious time to connect and send diligent feedback. It was a joy to be part of this outstanding yet patient group of scholars. Thank you very much.

I would also like to thank Dr. Hongsheng Liao, Jessica Branscome, Emily Davis, Dr. Michael Schmidtke, and all the lab technicians at the Center for Quantitative Fisheries Ecology for their otolith sectioning instructions, scientific discussions, and everything else they have done. Thank you to Christine Sookdheo for her weekly writing reminders and conversations.

I would like to thank the statistics department at Old Dominion University for being so helpful and kind. A special thanks to Mark Ledbetter for all his help during mathematical statistics.

And finally, but in no way least, my deepest thanks to my two children, Isabelle and Brett Kirch, for their undying support throughout this process.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

CHAPTER 1

INTRODUCTION

Understanding age of fish is crucial for marine ecosystem management as an integral part of fishery science research centers. The Magnuson-Stevens Fishery Conservation and Management Act requires that all fisheries in the United States have a management plan to ensure sustainability and maximum economic yield (2007). Poor management can lead to closure of the fishery or increased regulation such as size or number limits in the fish caught (Abbott and Haynie, 2012; Abesamis et al., 2014; Anderson, 2014; Arostegui et al., 2021). These increased regulations can have a detrimental effect to fishermen and local economies. For example, for the year 2020, $253 billion in sales were generated from U.S. commercial and recreational saltwater fishing. Fisheries contributed $117 billion to gross domestic product and provided for 1.7 million jobs (2023).

To create the best management plans, accurate ageing of fishes is crucial (Campana 2001). Age-based models allow for the most accurate analysis leading to sustainability with maximum economic benefit (Flinn and Midway, 2021; Haddon, 2011; Quinn and Deriso, 1999). One of the first age structured models was proposed by Beverton and Holt in 1957 and is still actively used in research today (Quinn and Deriso, 1999; Thorson, 2019).

Beverton and Holt relied on scale annuli to base their assessments of age. Since then, otolith-based ages have proven to be more reliable because the structure do not degrade like some scales. Otoliths are often sectioned to reveal double banded rings, one opaque and one translucent, forming an annual ring, and can be counted in a method similar to counting the rings

of a tree (Black et al., 2005; Godet et al., 2020). But first, it must be validated that the rings occur yearly, and not more or less often.

A *first* method of validating ages is the release of marked fish of a known age, the most robust method of age validation. A *second* method of validating ages is using radiochemical dating, measuring the radioactive decay of certain elements found within the fish, typically in otoliths. A *third* method of validating ages is the mark and recapture of fish tagged chemically, the most rigorous method when the fish is tagged at a young age and then recaptured at an older age. A *fourth* method of validating ages is with discrete length mode samples for age proxies. This method follows the modes of length over a given time and does not validate absolute age or annuli periodicity. A *fifth* method of validating ages is the use of natural date specific markers. This is similar to radiochemical dating in that it uses major events that chemically marks an entire fish population. A *sixth* method of validating ages is the use of captive fish raised from birth. The main disadvantage of this method is that annuli formation depends on environmental factors, so an artificial environment could produce artificial annuli. A *seventh* method of validating ages is marginal increment analysis (MIA). MIA is the most commonly used method because it is economical and efficient. It involves collecting otoliths for each month throughout the year (Campana, 2001).

MIA data is typically analyzed qualitatively through graphical and visual methods. One method involves graphing the monthly average measurement and visually inspecting the graph for a drop in accretion following a sawtooth pattern (Barbieri et al., 1994; Rodriguez-Marin et al., 2021). Edge analysis involves comparing the month of capture to whether the band at the edge of the otolith is opaque or translucent (Gebremedhin et al., 2021; Okamura and Semba, 2009).

However, there are studies using quantitative methods. One such quantitative model is based on analysis of variance (ANOVA) (Gumus, et al., 2007; Phelps, 2019; Williams et al., 2005). This method compares the growth measurements between months and looks for differences in the growth from one month to the next by reading the width of the ring. But, if the data exhibits outliers, conventional statistics can lack power and analysis may fail (Okamura et al., 2013).

Okamura et al. (2013) used circular statistics to model a single year's pattern in *Bathyraja parmifera*, a ray, in the eastern Bering Sea. Their idea was to use a circular-linear regression model with truncated wrapped Cauchy distribution for random effects, breaking the annual cycle into months. They also assumed that the estimated periodicity is invariant with age.

Both circular analysis and analysis of variance methods ignore the multi-year effects of continued growth. It is logical to build the model based on a time series that does not just include one year, but multiple years, with data partitioned by month capturing seasonality. A simple graphical display of that data shows patterns of the growth that mimic a sawtooth wave with an abrupt drop in accretion yearly to form an annual band (de Alaiza Martinez et al., 2015; Okamura et al., 2013). There are many limitations in the sawtooth related to model assumption, validation, and temporal dependence associated with MIA. Moreover, it has been challenging in finding the best model to predict this sawtooth wave pattern, due to the scarcity of fitted models available in statistical literature.

In addition to the sawtooth wave, fishery data is rife with missing and incomplete data sets (Afrifa-Yamoah et al., 2020; Afrifa-Yamoah et al., 2021). This missing data can be dealt with in numerous ways. It may be ignored, and the dataset will be incomplete; proxies can be used to estimate missing data; or it can be imputed. An important aspect of imputation in this

case is the ability to capture the time series setting of the data. Multiple imputation is much more preferred than single imputation (Azur et al., 2011; Donders et al., 2006).

Afrifa-Yamoah et al. (2021) analyzed recreational power boat launches from March 2011 to February 2012 and May 2013 to April 2014 in two different locations of Western Australia. Precipitation, temperature, humidity, wind speed, gust and direction, sea level pressure, and day of the week were used as covariates. They used Bayesian regression to impute missing data ranging from March 2012 to April 2013 and May 2014 to July 2014. With the use of posterior probabilities based on covariates, the analysis was able to determine the most important covariates for each location, and forecasted estimates were well within confidence intervals of observed data. So instead of ignoring missing data, recent and valid imputation methods become an attractive alternate to missing data.

In this dissertation, statistical aged models, time series data will be built on two kinds of fish: Atlantic croaker (*Micropogonias undulatus*) and blackbelly rosefish (*Helicolenus dactylopterus*) and I demonstrate the value of the statistical approach to evaluating marginal increments (MIs) for plausible temporal behaviors. Some of the Atlantic croaker data has been presented in Barbieri et al. (1994), and the dissertation will use that data as a proof of concept in Chapter 3. Atlantic croaker and blackbelly rosefish data for the next studies in Chapter 4 and 5 were collected by the Center for Quantitative Fisheries Ecology. These chapters extend our ability to analyze MI when following cohorts over time and when missing data are present for the first use of such statistical models in the validation of annuli through MIA.

CHAPTER 2

REVIEW OF LITERATURE FOR TIME SERIES

Time series is the analysis of data collected over time and MIA is an example of time series data. The MIA describes the changes in otolith accretion over months and years. However, MIA has never been analyzed with time series methods.

2.1 Classical time series models

Time series is a sequence of observations taken over n distinct times, $t_1, t_2, \ldots, t_n$, with $t_1 < t_2 < \ldots < t_n$, with mean of $\mu_y(t) = E(y_t)$ where the y_t describes the distribution of the observations at time t. For example, the monthly catch of fish and monthly fishing effort are examples of time series (Elmezouar et al., 2021; Saila et al., 1980). The autocovariance between two time periods, t and s, is $\gamma_y(t,s) = Cov(Y_t, Y_s) = E[(Y_t - \mu_t)(Y_s - \mu_s)]$ (Brockwell and Davis, 2002). The covariance is used to measure of correlation between the observations at different time periods. Typically, the further away the time periods, the more likely the observations are independent. However, when the time periods are close to each other, the dependence needs to be included in the modeling of the observations. In most cases, time is discrete and equidistant. If the times are equidistant, then the time series model can be considered consistent, and properties associated with that time series are proposed in the literature. If not equidistant, imputation may be used to regain consistency over time.

Nonconsistency in the data can be caused by lack of equidistant measurement. It is common in fishery science and may be linked to different scenarios such as, but not limited to, collection being cost prohibitive, limited time or resources, or adverse weather conditions making specimen collection impossible (Benavides et al., 2023). Benavides et al. (2023) used

imputation on a dataset with missing values. They found a seasonal decomposition with Kalman filters produced the lowest mean squared errors. Further methods to overcome lack of consistency in collection of the data will be described in Chapter 4.

The time series must also be stationary to be able to build predictions. Stationary means (i) $\mu_Y(t)$ is independent of t, (ii) autocovariance, $\gamma_Y(t+h,t)$, is independent of t for each h and (iii) $(Y_1, \ldots, Y_n)$ has the same distribution as $(Y_{1+h}, \ldots, Y_{n+h})$ for any choice of n and for any lag h (Brockwell and Davis, 2002).

An example of a stationary time series process is the autoregressive model (AR) of order 1. An AR(1) is defined as

$$Y_t = \phi Y_{t-1} + \varepsilon_t, \tag{1}$$

where $\varepsilon_t \sim N(0, \sigma^2)$ and ϕ is a constant so that the time series process is stationary (Brockwell and Davis, 2002). The autocovariance function for an AR(1) is

$$\gamma_o = \frac{\sigma^2}{1-\phi^2}. \tag{2}$$

This leads to the assumption that $|\phi| < 1$ (Cryer and Chan, 2008). Simulated AR(1) data where $\hat{\phi} = 0.9$ is shown in Figure 1 with the autocorrelation function (ACF) and partial autocorrelation function (PACF).

An AR(2) model is defined

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \varepsilon_t, \quad (3)$$

where $\varepsilon_t \sim N(0, \sigma^2)$, $\phi_2 + \phi_1 < 1$, $\phi_2 - \phi_1 < 1$, and $-1 < \phi_2 < 1$ (Brockwell and Davis, 2002). Simulated AR(2) data where $\hat{\phi}_1 = 1.5$ and $\hat{\phi}_2 = -0.75$ is shown in Figure 2 with the ACF and PACF.

The model can be extended to an autoregressive order p, having the form

$$Y_t = \sum_{k=1}^{p} \phi_k Y_{t-k} + \varepsilon_t, \quad (4)$$

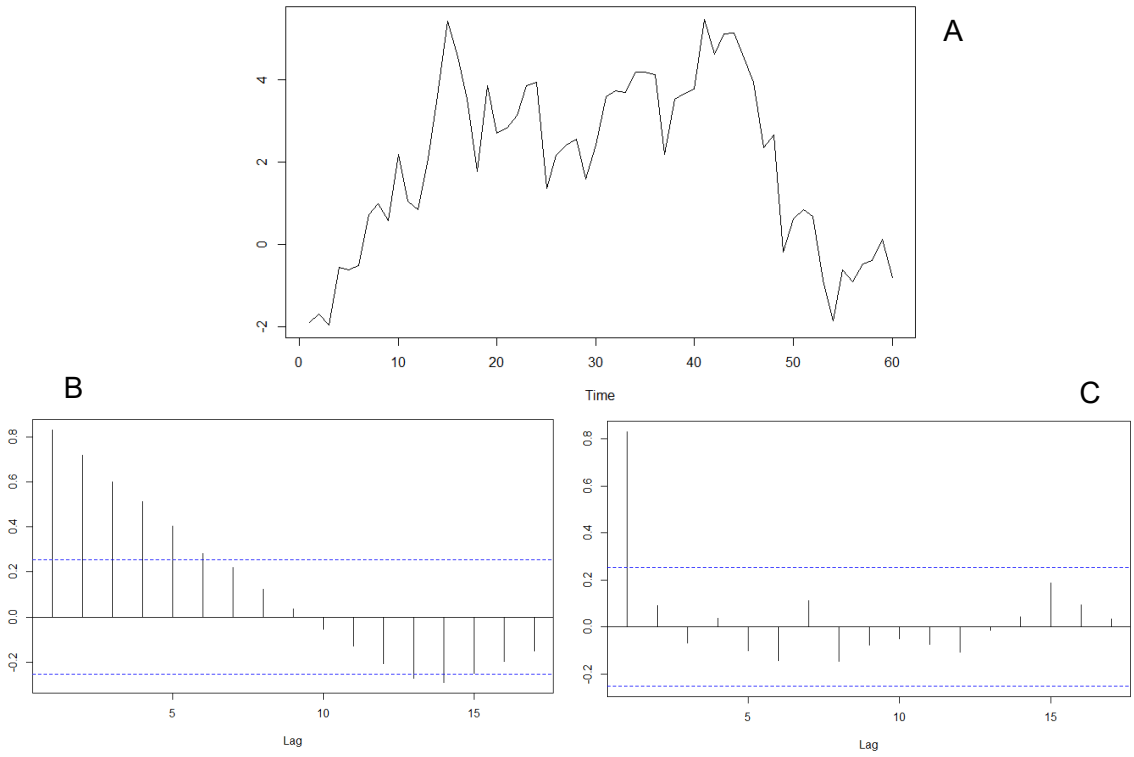

Fig. 1: A) Plot of simulated AR(1) data where $\phi = 0.9$. B) ACF of simulated AR(1) data where $\phi = 0.9$. C) PACF of simulated AR(1) data where $\phi = 0.9$. Reproduced from [33].

where $\varepsilon_t \sim N(0, \sigma^2)$ and the characteristic polynomial equation associated with AR(p) is defined as

$$1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p = 0. \quad (5)$$

The equation has p roots, and the time series is stationary if the roots are such that their norm is greater than 1, i.e. $|z| > 1$ (Brockwell and Davis, 2002). When considering an AR(1) model, the autocovariance function of order k is defined as

$$\gamma_k = \phi^k \frac{\sigma_e^2}{1-\phi^2}, \quad k = 1, 2, 3, \ldots, p, \quad (6)$$

while the autocorrelation function is

$$\rho_k = \frac{\gamma_k}{\gamma_o} = \phi^k, \quad k = 1, 2, 3, \ldots, p. \quad (7)$$

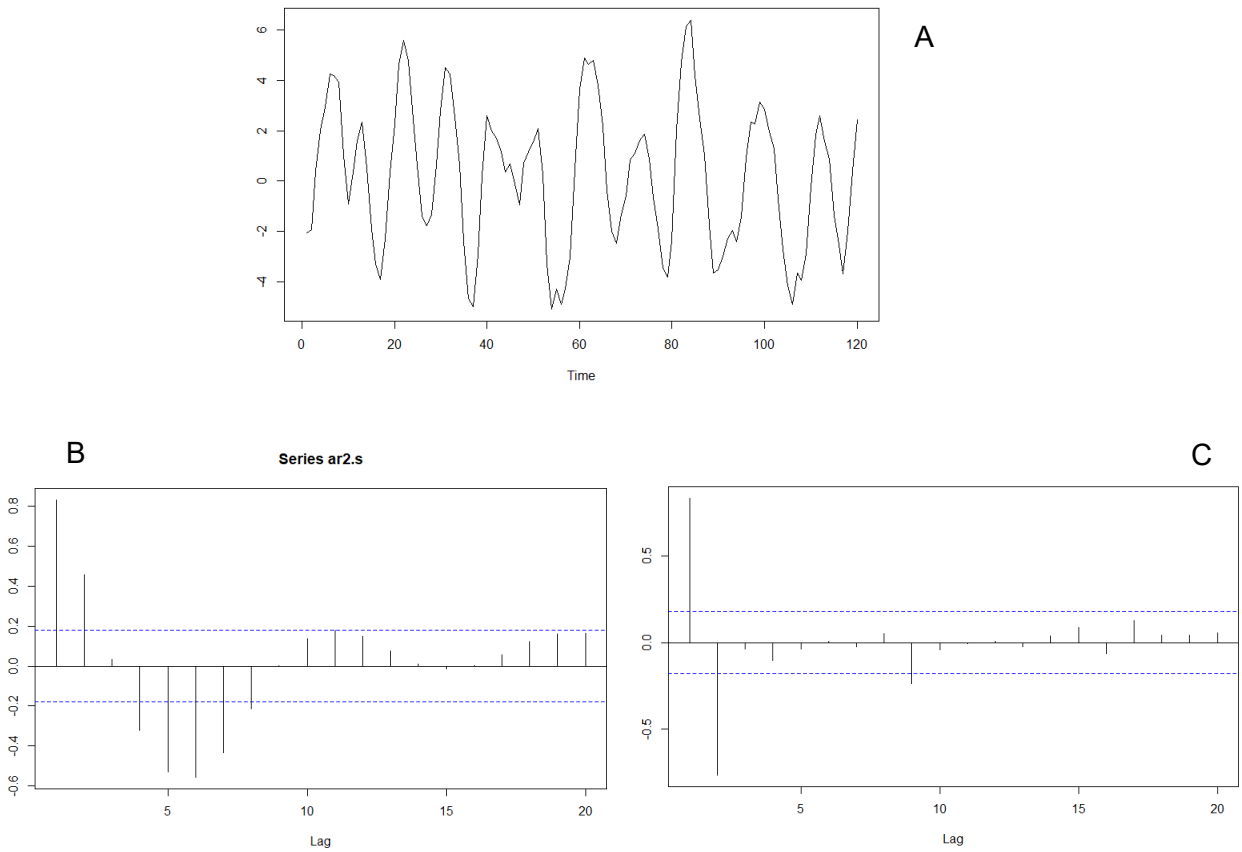

Fig. 2: Plot of simulated AR(2) data where $\phi_1 = 1.5, \phi_2 = -0.75$. B) ACF of simulated AR(2) data where $\phi_1 = 1.5, \phi_2 = -0.75$. C) PACF of simulated AR(2) data where $\phi_1 = 1.5, \phi_2 = -0.75$. Reproduced from [33].

In a given dataset, the choice of the order p is associated with the PACF (Brockwell and Davis, 2002). The equation for the PACF of a process is defined as

$$\rho_k = \phi_{p1}\rho_{k-1} + \cdots + \phi_{p(p-1)}\rho_{k-p+1} + \phi_{pp}\rho_{k-p}, \; k = 1, 2, \ldots, p \tag{8}$$

This formulation leads to the Yule-Walker equation, written as

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{p-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{p-2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \rho_{p-1} & \rho_{p-2} & \rho_{p-3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \phi_{p1} \\ \phi_{p2} \\ \vdots \\ \phi_{pp} \end{bmatrix} = \begin{bmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_p \end{bmatrix}, \tag{9}$$

or simply

$$\boldsymbol{P_p \phi_p = \rho_p}. \quad \text{(Box et al., 2016)} \tag{10}$$

Another case of a stationary process is obtained from the moving average (MA) model of order 1. A moving average process acts on the errors of the model and a MA (1) has an equation of

$$Y_t = \varepsilon_t - \theta\varepsilon_{t-1}, \tag{11}$$

where $\varepsilon_t \sim N(0, \sigma^2)$, θ is a constant and $|\theta| < 1$ such that the process is stationary (Brockwell and Davis, 2002). Simulated MA(1) data where $\hat{\theta} = -0.9$ is shown in Figure 3 with the ACF and PACF. The MA(1) process has an autocovariance functions given as

$$\begin{cases} \gamma_o = \sigma^2(1+\theta^2) \\ \gamma_1 = -\theta\sigma^2 \\ \gamma_k = 0, \ \ for\ k \geq 2 \end{cases}, \tag{12}$$

and autocorrelation

$$\begin{cases} \rho_1 = \frac{-\theta}{1+\theta^2} \\ \rho_k = 0, \ \ for\ k \geq 2 \end{cases}. \tag{13}$$

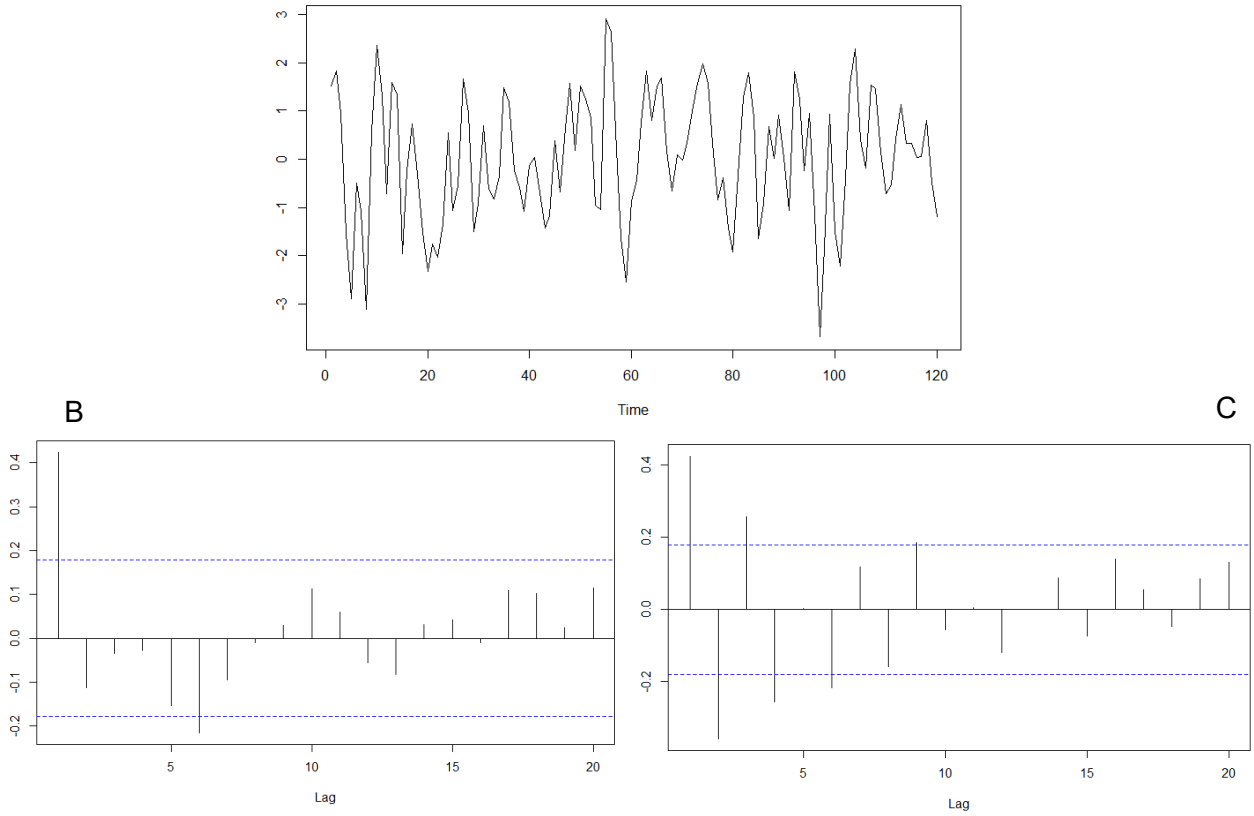

Fig. 3: A) Plot of simulated MA(1) data where $\hat{\theta} = -0.9$. B) ACF of simulated MA(1) data where $\hat{\theta} = -0.9$. C) PACF of simulated MA(1) data where $\hat{\theta} = -0.9$. Reproduced from [33].

An MA(2) process is defined as

$$Y_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2}, \tag{14}$$

where $\theta_2 + \theta_1 < 1, \theta_2 - \theta_1 < 1$, and $-1 < \theta_2 < 1$. Simulated MA(2) data where $\hat{\theta}_1 = 1$ and $\hat{\theta}_2 = -0.6$ is shown in Figure 4 with the ACF and PACF.

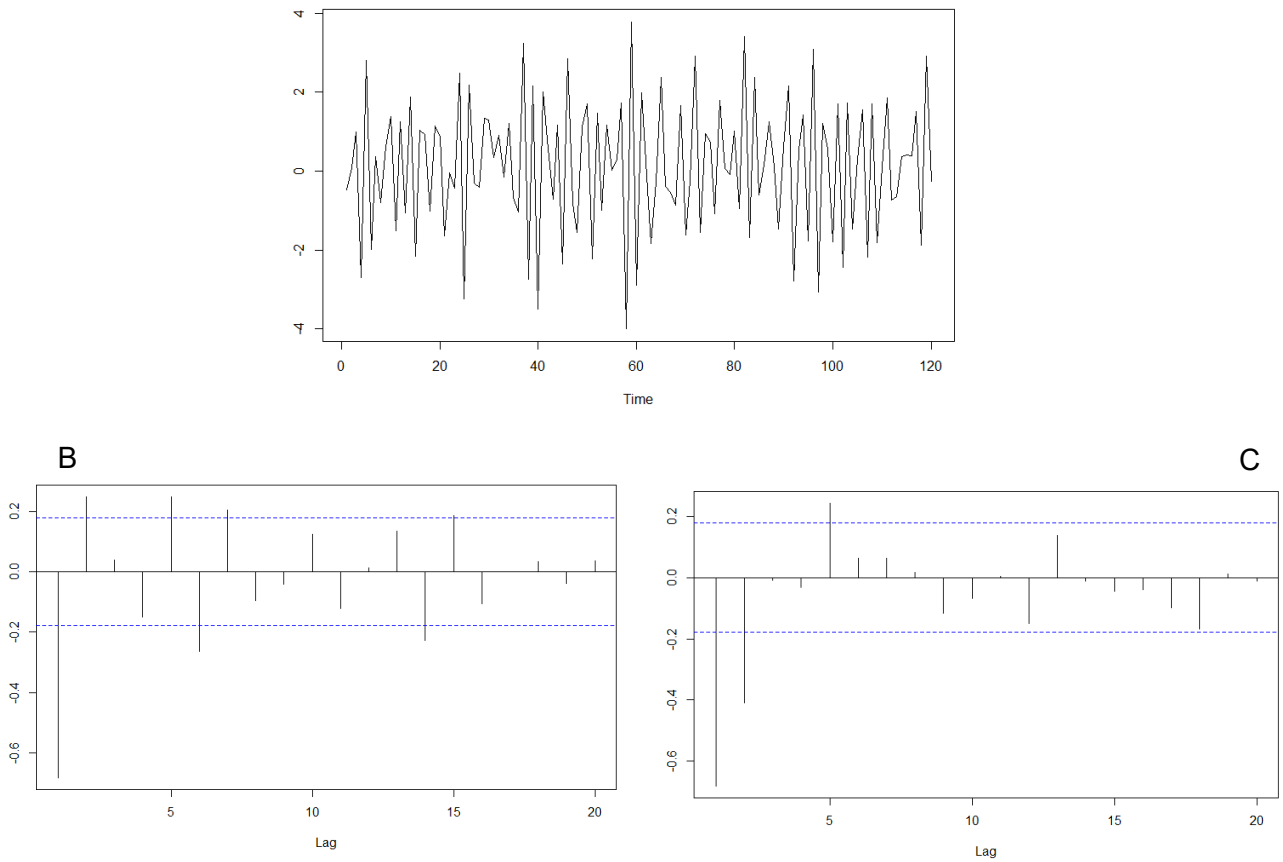

Fig. 4: Plot of simulated MA(2) data where $\hat{\theta}_1 = 1, \hat{\theta}_2 = -0.6$. B) ACF of simulated MA(2) data where $\hat{\theta}_1 = 1, \hat{\theta}_2 = -0.6$. C) PACF of simulated MA(2) data where $\hat{\theta}_1 = 1, \hat{\theta}_2 = -0.6$. Reproduced from [33].

The moving average formula can be extended to order q and has the following formulation

$$Y_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}\,, \tag{15}$$

where $\varepsilon_t \sim N(0, \sigma^2)$ (Brockwell and Davis, 2002).

The characteristic polynomial equation is

$$1 - \theta_1 z - \cdots - \theta_q z^q = 0. \tag{16}$$

The equation has q roots, and the time series is stationary if the roots of the characteristic polynomial have norm greater than one, i.e. $|z| > 1$.

An MA (q) has an autocovariance function of

$$\gamma_o = \left(1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2\right)\sigma_e^2. \quad (17)$$

To help determine the order q, the autocorrelation function (ACF) can be used and is defined as:

$$\rho_k = \begin{cases} \frac{-\theta_k + \theta_1\theta_{k+1} + \cdots + \theta_{q-k}\theta_q}{1+\theta_1^2+\cdots+\theta_q^2}, for\ k = 1,2,\ldots,q \\ 0, \quad for\ k > q \end{cases}. \quad (18)$$

Combining the two ideas of autoregressive and moving average, we obtain an autoregressive-moving average model (ARMA). For example, the ARMA (1,1) process is defined as:

$$Y_t = \phi Y_{t-1} + \varepsilon_t + \theta\varepsilon_{t-1}, where\ \phi + \theta \neq 0, \quad (19)$$

with $\varepsilon_t \sim N(0, \sigma^2)$, an autocovariance function

$$\gamma_o = \frac{1-2\phi\theta+\theta^2}{1-\phi^2}\sigma^2, \quad (20)$$

and an autocorrelation function

$$\rho_k = \frac{(1-\theta\phi)(\phi-\theta)}{1-2\theta\phi+\theta^2}\phi^{k-1}. \quad (21)$$

Simulated ARMA(1,1) data where $\hat{\phi} = 0.6$ and $\hat{\theta} = -0.3$ is shown in Figure 5 with the ACF and PACF.

In the general ARMA(p, q), the model becomes :

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \cdots - \theta_q\varepsilon_{t-q}, \quad (22)$$

with an autocovariance function

$$\gamma_k = \phi_1\gamma_{k-1} + \cdots + \phi_p\gamma_{k-p}, \ \ k \geq q+1, \quad (23)$$

and an autocorrelation function

$$\rho_k = \phi_1\rho_{k-1} + \cdots + \phi_p\rho_{k-p}. \quad (24)$$

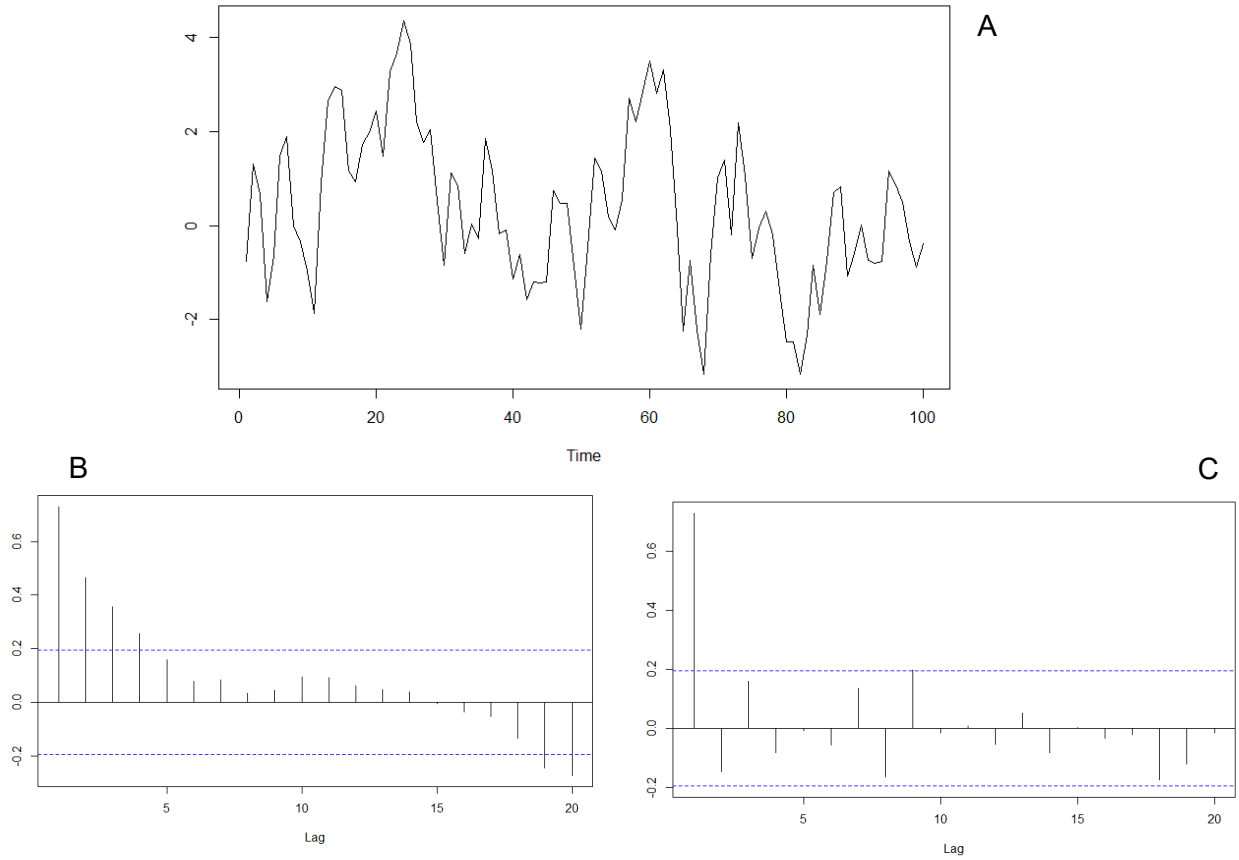

Fig. 5: Plot of simulated ARMA(1,1) data where $\hat{\phi} = 0.6,\ \theta = -0.3$. B) ACF of simulated ARMA(1,1) data where $\hat{\phi} = 0.6, \theta = -0.3$. C) PACF of simulated ARMA(1,1) data where $\hat{\phi} = 0.6, \theta = -0.3$. Reproduced from [33].

To achieve stationarity, (integrated) differences can be used between measurements at selected time points, creating an autoregressive integrated moving average process (ARIMA(p, d, q), where d is the differenced order), defined as:

$$\phi(B)(1 - B)^d Y_t = \theta(B)\varepsilon_t, \tag{25}$$

where $\varepsilon_t \sim N(0, \sigma^2)$ and B is the backshift operator; for example, $(1 - B^d)y_t = y_t - y_{t-d}$.

Also, $\phi(z) = 1 + \phi_1 z + \cdots + \phi_p z^p$ and $\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$ (Brockwell and Davis,

2002). Simulated ARIMA(1,1,1) data where $\hat{\phi} = 0.7$ and $\hat{\theta} = 0.7$ is shown in Figure 6 with the ACF and PACF.

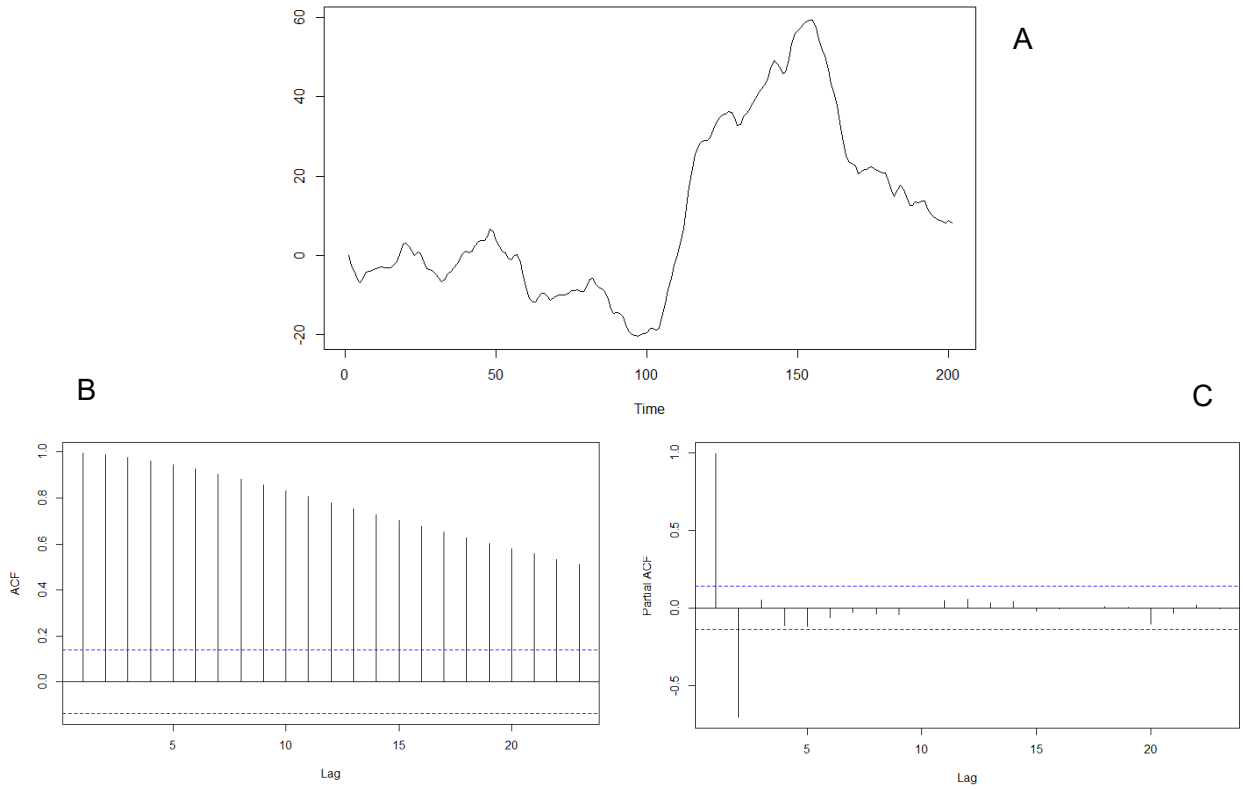

Fig. 6*:* Plot of simulated ARIMA(1,1,1) data where $\hat{\phi} = 0.7,\ \hat{\theta} = 0.7$. B) ACF of simulated ARIMA(1,1,1) data where $\hat{\phi} = 0.7, \hat{\theta} = 0.7$. C) PACF of simulated ARIMA(1,1) data where $\hat{\phi} = 0.7, \hat{\theta} = 0.7$. Reproduced from [33]p.

Mehmood et al. (2020) used nonstationary data from Pakistani fisheries from 1947-1997. They took the log of the differences to create stationarity. ARIMA was used to forecast the fisheries from 2017-2026. It was found that ARIMA (2,1,3) modeled the forecasting best.

Another method of analyzing time series is with trigonometric functions, also known as harmonic regression. Harmonic regression is used on data that is stationary and has longer periodic cycles than traditional time series models. Any non-zero mean stationary process can be

expressed as a collection of uncorrelated sinusoids. A function to express such an example of a trigonometric function is:

$$y_t = \sum_{j=1}^{k}\left[a_j \cos\left(2\pi\lambda_j t\right) + b_j \sin\left(2\pi\gamma_j t\right)\right], \quad (26)$$

where k is the number of harmonic functions, a_j and b_j are the amplitudes of their respective trigonometric functions, and λ_j and γ_j are the respective frequencies (Brockwell and Davis, 2002; Chesneau, 2020; Chesneau, 2021). The use of the tangent function is also possible, but less common (Figures 7 and 8).

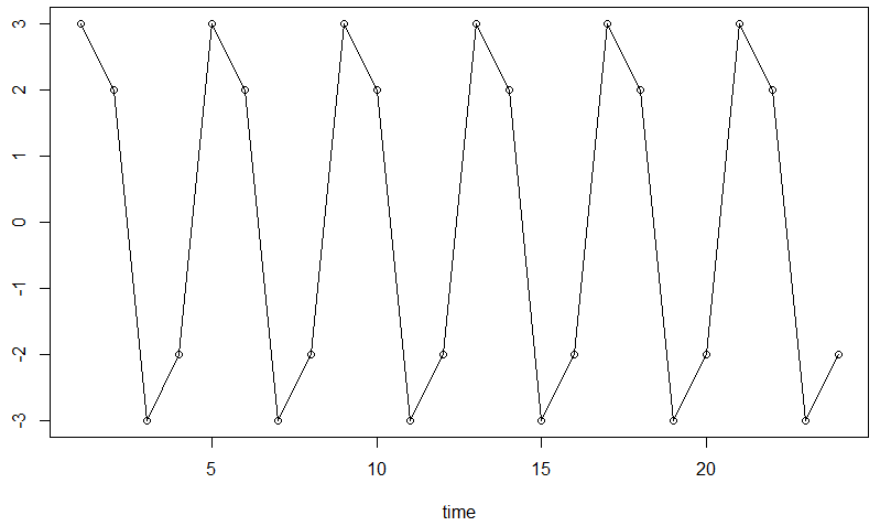

Fig. 7: Plot of $Y_t = -2cos\left(\frac{2\pi t}{4}\right) + 3sin\left(\frac{2\pi t}{4}\right)$.

Ashrafi et al. (2020) uses harmonic regression to analyze the seasonality of catch per unit effort in the near shore and off-shore Norwegian cod fishery. They used a fast Fourier transformation to determine the number of harmonics; they found two harmonic functions to fit the data. Their result concluded that the near shore fishery had a seasonal component to the catch per unit effort, while the offshore fishery did not.

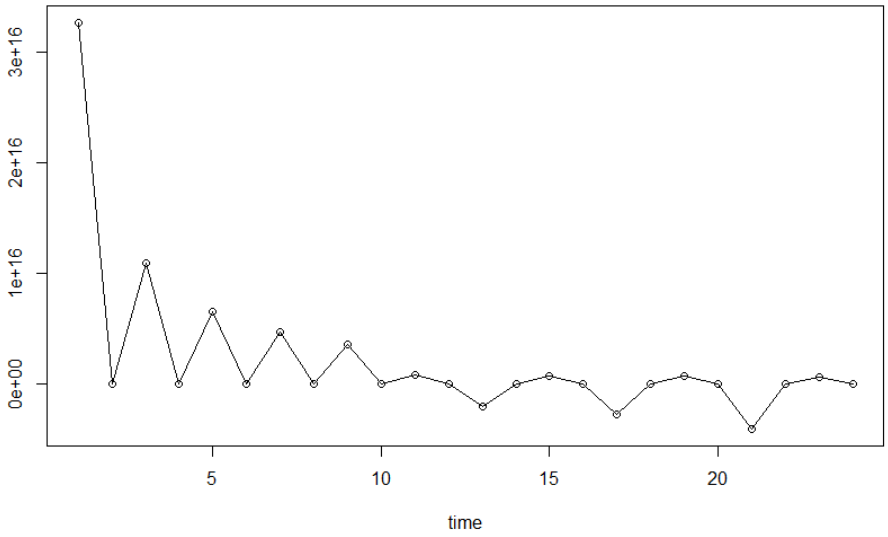

Fig. 8: Plot of $Y_t = -2cos\left(\frac{2\pi t}{4}\right) + 3sin\left(\frac{2\pi t}{4}\right) + 2tan\left(\frac{2\pi t}{4}\right)$.

2.2 Unobserved Component Models (UCMs)

Although time series models described in Section 2.1 are efficient, issues arise when stationarity is not met. Under stationarity, the assumption that the variance is the same may not be valid. Research has proposed another type of modeling, the UCMs. The novelty resides in the decomposition of the series capturing its different patterns. The time series under UCM is decomposed into a model with trend, seasonality, cycle, and regression elements. To get a clear picture of the underlying process, the trend is the first characteristic one may look for. If the trend is monotonic, differencing was used in the usual time series model from the previous section. In this new formulation, the trend is described by its mean long-term change. It gives an idea of the general tendency associated with the time series.

Segregating the seasonality, we get a clear understanding of periodicity of the process. Seasonality exists when there are consistent patterns over a fixed period of time. The cycle describes periods much longer than that of the seasonality, and typically uses a combination of sine and cosine functions.

The equation for this decomposition is written as

$$y_t = \mu_t + \gamma_t + \psi_t + \sum_{j=1}^{m} \beta_j x_{jt} + \varepsilon_t, \quad (27)$$

where μ_t represents the trend, γ_t represents the seasonal, ψ_t represents the cyclical component, $\sum_{j=1}^{m} \beta_j x_{jt}$ is the regression term with x_{jt} covariates, and $\varepsilon_t \sim iid\ N(0, \sigma^2)$ (SAS Institute, 2014). The regression term captures the effects of any covariates to the model.

Bian et al. (2019) used UCM to predict monthly car volume in a New Jersey highway from January 2006 to October 2016. The UCM used trend, consisting of a level with no slope and no cyclical component and no covariates. The data were stationary after looking at the ACF and PACF (there were no spikes in the graphs other than the first). A seasonal component was also added because of an obvious pattern in the car volume. They calculated the mean squared deviation, mean absolute deviation, and mean absolute percentage error between their observed and predicted values over a year. UCM was tested compared to simple linear regression, ARIMA, support vector machine, and artificial neural networks. It was found that the UCM performed better on the data than all the other models.

Haimerl and Hartle (2023) used UCMs to analyze COVID-19 daily cases reported by Johns Hopkins University from January 2020 to December 2022. Their data was non-stationary with changing variance over time. The model included trend that was volatile depending on the status of the disease at the time. The seasonality was described as six days, and cyclical component was captured by an AR(2). They found UCMs performed better than deterministic models with and without seasonality.

2.3 Copula

Normality and independence assumptions of the errors in the transformed time series model may be violated. Also, untransformed data may not be stationary. A more flexible

alternative is offered by the use of copula functions. A copula is a joint distribution of random variables whose marginal distributions have been transformed into their uniform cumulative distribution function (CDF) version. Kim et al. (2011) considered the joint distribution of bivariate time series model using copula.

A bivariate copula function is defined as

$$C: [0,1]^2 \rightarrow [0,1] , \tag{28}$$

which satisfies two sets of requirements. The first set being that:

$$C(u,0) = C(0,v) = 0, C(u,1) = u \; and \; C(1,v) = v \; for \; 0 \leq u, v, \leq 1, \tag{29}$$

or that the marginal distributions follow a uniform [0,1] distribution, u and v are the CDFs of some datasets. The second requirement set is

$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0 \tag{30}$$

$$for \; 0 \leq u_1 \leq u_2 \leq 1 \; and \; 0 \leq v_1 \leq v_2 \leq 1, \tag{31}$$

meaning the probability mass is non-negative (Sun et al., 2020).

Sklar's theorem states that for any two continuous functions X and Y, the joint distribution can be formulated from its marginal CDFs F(x) = Pr(X≤x) and F(y) = Pr(Y≤y), and there exists a copula C such that $f(X,Y) = C(F_x, F_y|\theta)$. Moreover, the copula C is unique if X and Y are continuous (Sun et al., 2020). However, when $F_x$and $F_y$are discrete or mixed, the uniqueness of the copula is within a suitable range.

There are many different measures of goodness of fit associated with copula models. One of those measures of goodness of fit is the Kendall's tau statistic. It measures the difference in concordance and discordance between the two variables. The Kendall's tau is defined as the partial derivative of C with respect to u and v, or

$$\tau = 1 - 4 \iint \partial_u C(u,v) \partial_v C(u,v) du \; dv, \tag{32}$$

where $-1 < \tau < 1$ (Chesneau, 2022).

Marsh et al. (2015) used the bivariate Gaussian copula to analyze habitat preferences and spatial distributions of Albacore tuna in New Zealand's Exclusive Economic Zone. Data were collected from 2003 to 2012 and included catch per unit effort based on four seasons in a year. Sea surface temperature, sea surface height, sea surface gradient, net primary productivity, and depth were used as covariates. They found that the catch per unit effort over the time period modeled using Gaussian copula provided a better model fit than the traditional models. The traditional models assume independence of habitat conditions, whereas the copula modeled them jointly.

Rautureau et al. (2010) studied copulas to test the feasibility of a French monkfish derivatives market. They studied daily average prices from 44 seafood auction markets from January 1994 to December 2006. Size, presentation, and quality were used as covariates. Volatility was observed in the maximum price, but not with very low prices. Spearman's rho and Kendall's tau were calculated to check for associations between sizes. Spearman's rho ranged from 0.188 to 0.705 while Kendall's tau ranged from 0.129 to 0.529. They concluded that the bivariate copula modeled the non-linear dependence relationship between price indices and sizes of the monkfish suitably well. A derivative market for French monkfish was then a feasible price risk management.

2.4 Imputation

In chapter four, following a cohort of Atlantic croaker, and chapter five, time series with blackbelly rosefish, I consider datasets with missing data. There are three types of missing data, missing completely at random (MCAR), missing at random (MAR), and missing not at random

(MNAR). The probability that an MCAR observation is missing does not depend on anything observed or unobserved. The missingness can be ignored and the data can be handled as a simple random subsample for analysis (Baraldi and Enders, 2010; Little et al., 2014). The probability associated with MAR observation is based on values of observed data. The missingness is correlated to other collected variables. The probability associated with a MNAR observation is missingness may depend on something that has not been measured or is related to the missing values themselves (Baraldi and Enders, 2010; Little et al., 2014). The lack of data is informative to the analysis (Davey and Savla, 2010; Little et al., 2014). Many different probability distribution functions can be used. The choice depends on the type of missingness at hand.

There are six strategies for handling missing data. The first is the *complete case method* where missing values are ignored. This method uses only complete observations (Davey and Savla, 2010). The second method is called *available case method*. This method uses information from both complete and incomplete observations. For example, in the case of test scores, where we have a pre and post evaluation; if one score is missing, it is ignored (Davey and Savla, 2010). Available case method works better at correcting bias with MAR data. A third method is *list-wise deletion*. This method removes observations at the multivariate level that have one or more data points missing. It should only be used with MCAR data or else it leads to biased estimates (Baraldi and Enders, 2010). The fourth method is called *list-wise deletion with weighting*. This method is similar to list-wise deletion but gives weight to some observations to reduce bias. For example, in a gender study, if female subjects are twice as likely to complete the survey, male subjects have their responses counted two times (Davey and Savla, 2010). A fifth method is *pair-wise deletion*. In this method, sample moments are calculated based on pairs of available variables (Baraldi and Enders, 2010). A sixth method is based on the *expectation maximization*

algorithm. In this method, sample moments of the variable of interest are estimated with a two-step iterative process. In the expectation step, expected values of the variable of interest replace missing values; in the maximization step, maximum likelihood estimates are made of the mean and covariance. This iteration repeats until data convergence is achieved (Davey and Savla, 2010).

All of the above-described methods are well documented; however, another alternative method to deal with missing data is imputation. *Single imputation* replaces the missing data point with a plausible value such as the mean or estimates from a regression equation (Baraldi and Enders, 2010). This method is known to produce unbiased estimators in MCAR data but when used with MAR data, it produces standard errors that are too small (Davey and Savla, 2010). Such estimates lead to lower bias estimates that are misleading. Another type of imputation, *multiple imputation,* replaces missing data with multiple plausible values, based on the distribution, the covariates, and the frequency of the missingness in the dataset. Since multiple choices are available for the missing data, the variability in parameter estimates is only due to the uncertainty from the imputation process used (Davey and Savla, 2010; Little et al., 2014).

In the cases of chapters four and five, the data are missing not at random. Data collection was limited due to adverse weather conditions. Due to the missing not at randomness of the data I chose to use multiple imputation. A type of multiple imputation is multiple imputation by chained equation (MICE). Three types of MICE are multivariate normal model, both Bayesian and non-Bayesian, and predictive mean matching (PMM). Multivariate normal imputation tends to require a large sample size to compensate for a general underestimation of variance. They also perform poorly when model assumptions, such as normality and homoscedasticity, are violated. PMM tends to be more robust to model misspecifications such as heteroscedasticity and non-

normality since the predicted values are limited to the range of observed values which can preserve any non-linear relationships (Kleinke 2017). Kleinke (2017) conducted simulations with varying sample sizes, donor pool sizes, missing data percentages, and distribution skewness. PMM consistently provided better results than multivariate normal imputation in all the scenarios tested (Kleinke 2017).

PMM takes several steps. First, for the entries with no missing data, marginal increment is regressed on time, producing a set of coefficients. Second, coefficients are randomly drawn from those estimated. Third, the randomly drawn coefficients are used to predict the marginal increment ratio of all data points. Fourth, missing points are matched with non-missing points whose predicted values are close. Fifth, from the close predictions, one is randomly chosen and assigned to the missing value. Sixth, steps two to five are repeated a set number of times. The mean of these repeats is used as the imputed value (Allison, 2015). In other words, for the 2008 cohort that there is no sample in February 2010 but there is in Feb 2009, because values are standardized, the use of PMM to fill in the gap of the value in 2010 should lie somewhere near the 2009 MI proportion. In contrast to PMM, the MICE generates multiple data values, and captures the potential heterogeneity, or variance, as needed when building confidence interval.

CHAPTER 3

Fitting Time Series Models To Fisheries Data To Ascertain Age

3.1 Introduction

Sustainable fisheries rely on the use of age-based demographic models that account for reproduction and growth to maximum economic benefit. To achieve this goal requires that fish be ascribed correct ages (Gebremedhin et al., 2021). Ascribing correct ages relies on validating putative annual marks on fish bones (Campana, 2001). The most commonly used technique to validate marks is marginal increment analysis (MIA) (Campana, 2001; Foster, 2001). MIA evaluates the width of newly accreted material laid down on the edge of fish ear bones measured over a year compared to material accreted over the previous year (Campana, 2001) (Figure 9). The ratio is typically averaged by month over all age groups and plotted over a single year. The plot shows a single dip if one band is produced yearly (Figure 10). This dip validates the pattern of a putative annual ring, allowing for accurate ageing of the fish to the nearest year.

For this pattern verification, few instances exist of statistically validated quantitative approaches because of irregular patterns observed due to changing environmental conditions. One such irregular pattern often looks like a sawtooth wave with an abrupt drop in accretion yearly to make an annual band (de Alaiza Martinez et al., 2015; Okamura et al., 2013; Phelps et al., 2019). Such sawtooth wave patterns provide challenges in fitting data to statistical models.

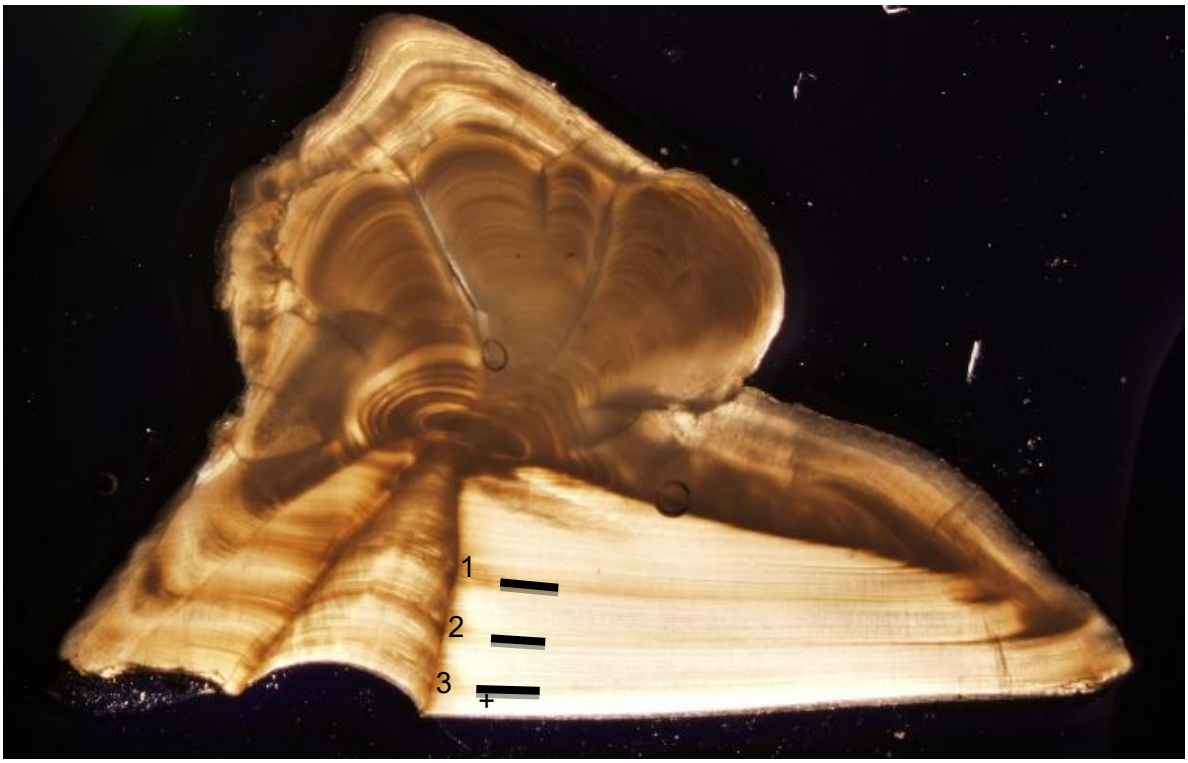

Fig. 9: This is the microscopic cross-section of the ear bone of a 3-year-old Atlantic croaker. The age is evidenced by the three numbered dark bands. Year 1 begins at the nucleus and continues down the succal groove (dashed line) to band 1. Year 2 begins at band 1 and continues to band 2. Year 3 begins at band 2 and ends at band 3. Accretional growth during year 3 extends to the edge of the ear bone as indicated by the "+". Inset: the marginal increment ratio is b/a.

Attempts have been made to overcome the difficulty of modeling the sawtooth wave pattern. Okamura et al. (2013) used circular statistics and conversion of the month of capture to the median of the month, and then another conversion of the median of the month to radians with accretion ratios as the linear vector. The circular method provides a short-term forecast that assumes the data is uniformly distributed around one preferred direction (Landler et al., 2018). In contrast, Phelps et al. (2019) used analysis of variance to identify significant differences in increment widths between months, but not in the years' trend. Analysis with a time series model

is appropriate for yearly patterned data because time series models identify seasonal patterns and should detect a yearly accretion pattern.

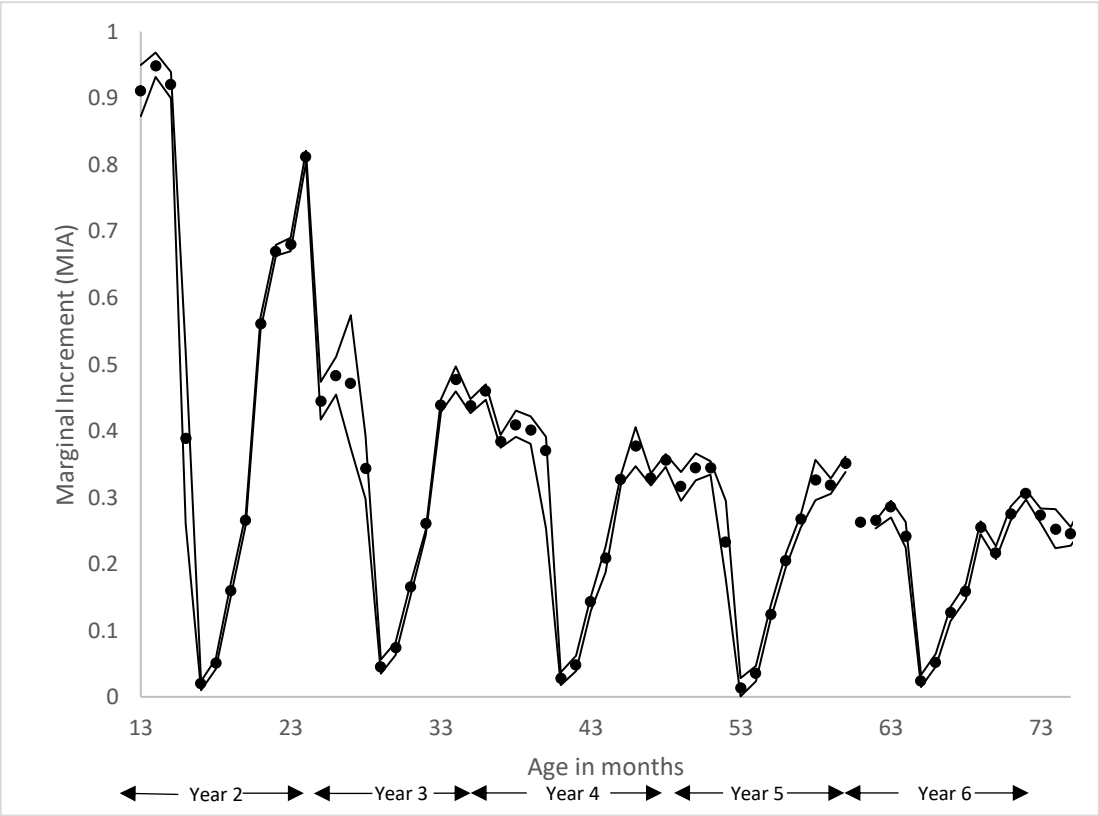

Fig. 10: Time series of marginal increment analysis (MIA) data with one standard error interval from Barbieri et al. (1994). Notice the dip in May of each year. By fisheries convention, age in months is arbitrarily assigned a birthdate of January 1 to all fishes born in the northern hemisphere. The x axis has two legends. The first shows age in months, the second shows the distribution of months as years.

The advantage of time series analysis is that it accounts for the dependence of the observations and errors in the dataset (Makridakis et al., 1983). Time series analysis of marginal increments can allow for an objective determination of the temporal, seasonal, and cyclic nature of the growth accretion. Time series models, such as autoregressive integrated moving average

(ARIMA), unobserved components models (UCM) and Gaussian copula can pinpoint seasonal changes through time.

A time series approach works especially well when data is complete and abundant. Other assumptions of ARIMA include the independence and identical distribution under the Gaussian nature of the error terms in the model statement, as well as the constant variance assumption, or stationarity. Unfortunately, the sawtooth wave pattern in MIA data complicates finding a well-fitted time series model.

The UCM can break a time series of data into trend, cyclical, seasonal, autoregressive, and lagged components (Koopman and Ooms, 2004; Bian et al., 2019). Bian et al. (2019) gave a summary of UCM applications but did not include analysis in fisheries research. To our knowledge, this paper is the first to use UCM to model MIAs. UCMs use smoothing models that allow for improved analysis and put more weight onto observations that are closer to each other due to the non-ignorable correlation of time series measurements (Yang and Zhang, 2019). The single error model is in fact broken up into its component calibration.

The Gaussian copula construction emanates from Joe (2015) and other authors (Alqawba et al., 2019; Sun et al., 2020). However, just as in the UCM case, its use in fisheries research is still in its infancy (Hosack et al., 2014; Marsh et al., 2015). Copula models easily incorporate a variety of marginal distributions and different dependence structures, unlike ARIMA whose assumptions rely on identical joint distributions at all time values and identical Gaussian marginal distributions (Makridakis et al., 1983). As a first advantage, the copula approach for parameter estimation in multi-stage models captures the dependence from the temporal samples. A second advantage is that the copula circumvents the non-trivial computations in the variances of the estimates. Gaussian copulas are very common in economics, epidemiology, time series

analysis, and climate change (Alanazi , 2021; Patton 2012; Alqawba et al., 2021; Yin et al., 2022; Yin et al., 2023).

We propose the comparison of the ARIMA, UCM, and copula-based statistical time-series methods while incorporating serial dependence. The goal is to evaluate the feasibility and preliminary efficacy of the models on fish ear bone accretion data. We are fortunate in this first test to have an unusually complete dataset without the often-missing seasonal components due to conditions that prohibit fishing such as adverse weather or lack of fish availability.

The paper is organized as follows. In Section 2, the data are described with the ARIMA, UCM, and copula models, and the statistical analysis methods are presented. In Section 3, the solutions are presented. We end with a discussion of the performance of these models using our complete dataset.

3.2 Materials and Methods

3.2.1 Data Collection

The data for this study were extracted from the data included in Barbieri et al. (1994) using DataThief III, Version 1.7 (2015) to extract the data points from a graph. According to the authors' methods, fish were collected each month between June 1988 and June 1991 from commercial fisheries in the Chesapeake Bay and brought to their laboratory. Next, individual biometrics were recorded, and ear bones excised.

Fish ear bones (sagittal otoliths) were sectioned, and then mounted on microscope slides and magnified to a range from 1x to 8x and photographed using an Olympus DP71 camera and the program Cells Sense (Figure 9). The images were then uploaded to the program Image-Pro Plus v. 6.2.0.424 (Media Cybernetics Inc.) for marginal increment measurement, in micrometers.

The first marginal increment was measured from the last dark band to the outer edge of the ear bone. The second is measured from the last dark band to the previous dark band. These measurements are then combined as a ratio of the last partial accretion to the last full accretion band (Figure 9 inset). The average ratio was then calculated for each age class by its month of capture. The average was plotted by month and visually inspected for the drop in the sawtooth wave. Fish were categorized by the age at capture, regardless of year of capture. The youngest age caught was 1 year.

3.2.2 Model Diagnostics and Goodness of Fit Measurements

Consider the data as n observations of type $y_1, \dots, y_n$; the likelihood function can be written as

$$L(\Theta; y_1, \dots, y_n) = \prod_{i=1}^{n} f_y(y_i; \Theta), \tag{33}$$

where,

$$\Theta = \left[\theta_1, \dots, \theta_p\right]^T \tag{34}$$

is the vector of *p* parameters associated with the model equation.

The log likelihood equation can then be formulated as

$$log(L) = \sum_{i=1}^{n} log\left\{f_y(y_i; \Theta)\right\}. \tag{35}$$

Akaike's information criterion is then calculated by

$$AIC = -2log(L) + 2K, \tag{36}$$

where $K = p$ is the number of parameters in the model (Burnham et al., 2011; Piegorsch and Bailer, 2005).

The mean squared error is formulated as

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2, \tag{37}$$

and the Kendall's tau that measures the difference in the measure of concordance and discordance between the marginal CDFs is:

$$\tau = 1 - 4 \iint \partial_u C(u,v) \partial_v C(u,v) du\, dv \quad (38)$$

as in Sun et al. (2020), where *C* represents the copula function that will be described in the "Copula" section, with *u* and *v* capturing the marginal distribution of X and Y, respectively.

3.2.3 Time Series Model

The data from the time series process X_t is described as an ARIMA (*p,d,q*) model and is defined as:

$$\phi(B)X_t = \theta(B)Z_t, \quad (39)$$

where, *p*, *d*, and *q* are non-negative integers, based on a sample size of n, with:

$$X_t = (1-B)^d Y_{t,} \quad (40)$$

$$Z_t \sim WN(0, \sigma^2), \quad (41)$$

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p, \quad (42)$$

$$\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q. \quad (43)$$

where Y_t is the transformed X_t series with mean retrieved, and B is the backward shift operator (Brockwell and Davis, 2002).

The parameter *p* is associated with the autoregressive portion of the process while *q* is associated with the moving average portion of the process. To help estimate the autoregressive parameter (*p*), the sample autocorrelation function can be used once graphed,

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}, \quad -n < h < n, \quad (44)$$

where γ is the sample autocovariance function of the lag h and is defined as

$$\hat{\gamma}(h) := n^{-1} \sum_{t=1}^{n-|h|} \left(x_{t+|h|} - \overline{x}\right)(x_t - \overline{x}), \quad -n < h < n. \tag{45}$$

To help estimate the moving average parameter (q), the sample partial autocorrelation function can be used once graphed,

$$\alpha(0) = 1, \tag{46}$$

$$\alpha(h) = \phi_{hh}, \ h \geq 1, \tag{47}$$

where ϕ_{hh} is the last component of

$$\phi_h = \Gamma_h^{-1} \gamma_h, \tag{48}$$

where

$$\Gamma_h = [\gamma(i-j)]^h, i, j = 1 \ldots T \text{ , } h \geq 1 \quad \text{and} \tag{49}$$

$$\gamma_h = \gamma(h). \tag{50}$$

The likelihood function is (Almetwally et al., 2022):

$$L(\theta|Y_t) = \prod_{n=2}^{n} \frac{1}{(2\pi^{n/2})|\Sigma|^{1/2}} exp\left(\frac{-1}{2} Y' \Sigma^{-1} Y\right), \tag{51}$$

where Σ is the variance/covariance matrix of the observed time series data.

R package “astsa” is used to loop through possible combinations of p and q ranging from zero to three. (Hyndman and Killick, 2022). The AIC, log likelihood, and MSE were used to decide the best model.

3.2.4 Unobserved Components Model

Components of the models can be assumed unobserved and must be estimated under a time series model. Unobserved components model (UCM) offers such flexibility. Unobserved components can be modeled using the equation:

$$Y_t = \mu_t + \gamma_t + \psi_t + r_t + \sum \varphi_i Y_{t-i} + \varepsilon_t, \tag{52}$$

where

μ_t represents the trend component,

γ_t represents the seasonal component,

ψ_t represents the cycle trend,

r_t is the autoregressive term,

$\sum \varphi_i Y_{t-i}$ is a regressive term involving the lagged dependent variables,

ε_t is the error term assumed to be independent with identical Gaussian distribution,

and μ_t, γ_t, ψ_t, and r_t are assumed to be independent of each other (Yang and Zhang 2019).

The likelihood function is (Almetwally et al., 2022):

$$L(\theta|Y) = p_1(y_1) \prod_{t=2}^{n} p_t(y_t|y_{t-1};\theta), \quad (53)$$

where $Y = (y_1, \ldots, y_n)'$.

Analysis is conducted using SAS "proc ucm." Six models are tested using various combinations of level, slope, cycle, and season, and described in Table 1 as follows.

Table 1: Combination of all possible methods used to test the unobserved components models (UCMs).

	Level	**Slope**	**Cycle**	**Season**
Model 1	Stochastic	Stochastic	Stochastic	Stochastic
Model 2	Stochastic	Stochastic	--	Stochastic
Model 3	Stochastic	Fixed	Stochastic	Stochastic
Model 4	Stochastic	--	Stochastic	Stochastic
Model 5	Stochastic	Fixed	--	Stochastic
Model 6	Stochastic	--	--	Stochastic

3.2.5 Copula Model

A copula is defined as

$$P(Y_t \leq y_t, Y_{t-1} \leq y_{t-1}) = C\{P(Y_t \leq y_t), P(Y_{t-1} \leq y_{t-1})\}, t = 1,2, \ldots, T, \quad (54)$$

$$C: [0,1]^2 \rightarrow [0,1], \quad (55)$$

which satisfies the requirement

$$C(u,0) = C(0,v) = 0, C(u,1) = u, C(1,v) = v, for 0 \leq u,v \leq 1 \text{ and} \quad (56)$$

$$C(u_2,v_2) - C(u_2,v_1) - C(u_1,v_2) + C(u_1v_1) \geq 0 for 0 \leq u_1 \leq u_2 \leq 1 \wedge 0 \leq v_1 \leq v_2 \leq 1. \quad (57)$$

In addition, Sklar's theorem states that for the equation

$$P(X \leq x, Y \leq y) = C\{F(x), G(y)\} \quad (58)$$

a unique function of C can be found if F and G are continuous (Sun et al., 2020; Alanazi , 2021). This paper will use beta marginals described as

$$G(y) = \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha,\beta)}, \quad (59)$$

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \quad (60)$$

$$\mu = E[Y_t], \quad (61)$$

$$\sigma^2 = Var(Y_t), \quad (62)$$

Where α and β are the non-negative shape and scale parameters, and Γ is the Gamma function. Beta marginals are used since the data is limited to values ranging between zero and one. The likelihood function is (Guolo and Varin, 2014):

$$L(\theta|Y) = p(y_1,\theta)\prod_{t=2}^{n} p_t(y_t|y_{t-1},\dots y_1;\ \theta). \quad (63)$$

The equation used in the copula analysis was

$$MIA_t = tan\left(\frac{2\pi t}{12} - 4\right)\left(sin\left(\frac{2\pi t}{12} - 4\right) + \frac{2}{\pi^2}cos\left(\frac{2\pi t}{12} - 4\right)\right) + \varepsilon_t, \quad (64)$$

where sin, cos and tan are the trigonometric functions, and $\varepsilon_t \sim iid\ N(0,\sigma^2)$ are the error terms.

The R package "gcmr" was used to fit the copula models (Masarotto and Varin, 2017). For the sake of interpretability of smoothness, ARIMA (1,1,0) was selected. Due to the time dependency, trigonometric functions were used in the model equation (Chesneau 2021a,b, 2022).

The copula is more parsimonious than UCM or time series. It includes only the time series parameters in a more succinct form based on the marginal distributions of the Y_t's. The copula equation was estimated and reported along with the log likelihood and AIC. Kendall's tau was calculated to estimate correlation between marginal accretion width over consecutive time periods.

3.3 Results

3.3.1 Data

Ear bones from a total of 1185 fish were prepared and increments measured by Barbieri et al. (1994). The monthly average increment by age is graphed in Figure 10. A clear decrease in accretion width can be observed in May of each year, as well as a decreasing variance over time.

3.3.2 ARIMA Model

The model with the highest log likelihood (65.6458), and lowest AIC (-117.292), and MSE (0.00805) had an autoregressive order of 2 and a moving average order of 3 (Table 2). All the estimated parameters were significant except the third moving average term, and when this third term is removed, the constant term becomes insignificant (Table 3). The QQ plots of these two models show the majority of the data with a normal distribution (Figure 11).

3.3.3 Unobserved Component Model

Table 2: Model diagnostics for the ARIMA (p,1, q) model. The bold values show the best model parameters, while the italic values show the second-best model parameters.
A shows the AICs for selected ARIMA models.
B shows the MSE for selected ARIMA models.
C shows the log likelihood for selected ARIMA models.

A

p\q	**0**	**1**	**2**	**3**
0	-80.6663	-83.841	-81.8519	-88.0023
1	-83.211	-81.8484	-90.054	-89.9316
2	-82.1529	-96.3694	*-83.3204*	**-117.292**
3	-80.8888	-99.4612	-107.115	-104.951

B

p\q	0	1	2	3
0	0.0178	0.0165	0.0165	0.0140
1	0.0167	0.0165	0.0137	0.0134
2	0.0164	0.0124	*0.0144*	**0.0081**
3	0.0162	0.0114	0.0100	0.0100

C

p\q	0	1	2	3
0	42.33315	44.9205	44.9259	49.0012
1	44.6055	44.9242	50.0270	50.9658
2	45.0764	53.1847	*47.6602*	**65.6458**
3	45.4444	55.7306	60.5575	60.4754

The six models tested using various combinations are presented in Table 4. Model 4 provided the highest log likelihood (67.874) and lowest AIC (-123.7), while Model 1 produced the lowest MSE (0.00626). The AIC and MSE were smaller than the ARIMA model, while the log-likelihood is larger than the best of the ARIMA models (Table 4). All models produced an

identical graph with an extremely narrow confidence interval (Figure 12). The tails do deviate from the 45^{o} line.

The residuals for Models 2,3, and 5 had a distinct sinusoidal pattern while the others had a slightly more random appearance (Figure 13). The QQ plots of the residuals show most of the data follows a normal distribution with the tails deviating from the 45^{o} line (Figure 14). Models 2, 5 and 6 had a sinusoidal pattern in the ACF graphs (Figure 15).

Table 3: Parameter estimates for best ARIMA models.
A shows parameter estimates for best time series model, ARIMA (2,1,3) (bold in Table 2).
B shows Parameter estimates for second best time series model, ARIMA (2,1,2) (italic in Table 2).

	Estimate	Standard Error	t value	p-value
AR 1	1.6169	0.0559	28.932	< 0.0001
AR 2	-0.9253	0.0506	-18.2994	< 0.0001
MA 1	-2.063	0.1885	-10.9446	< 0.0001
MA 2	1.2442	0.3565	3.4899	0.0009
MA 3	-0.1262	0.1803	-0.7	0.4864
Constant	-0.0051	0.0019	-2.6825	0.0093

AR 1	0.9625	0.0715	13.4632	< 0.0001
AR 2	-0.9573	0.0561	-17.073	< 0.0001
MA 1	-0.8789	0.0877	-10.0175	< 0.0001
MA 2	1	0.0861	11.6125	< 0.0001
Constant	-0.0083	0.0161	-0.5134	0.6094

Table 4: Summary statistics for each of the six UCMs. Model 4 has the lowest AIC and the largest log likelihood. Model 1 has the lowest MSE.

Model	AIC	MSE	Loglikelihood
1	-121.2	**0.0063**	67.594
2	-91.53	0.0100	49.766
3	-118	0.0065	65.014
4	**-123.7**	0.0084	**67.874**
5	-93.53	0.0100	49.766
6	-99.68	0.0128	52.842

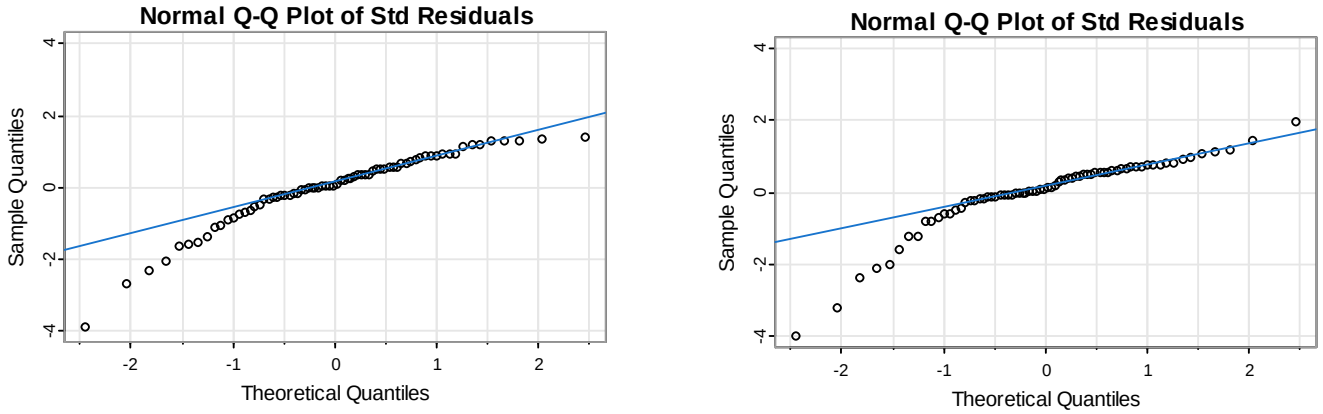

Fig. 11: QQ plots of the best two ARIMA models. A is (2,1,3) and B is (2,1,2).

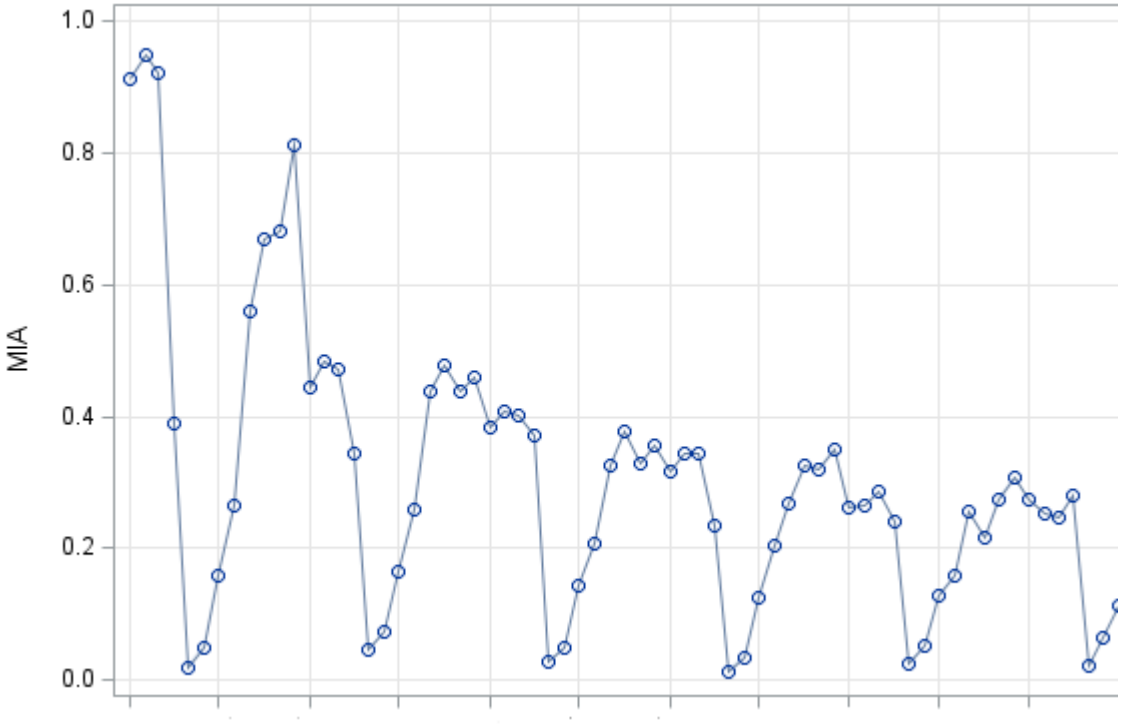

Fig. 12: MIA data fit from each of the six unobserved components models (UCMs). The same graph was produced for all six models. The confidence intervals are very narrow. Notice the dip in May of each year described with a red line for each May.

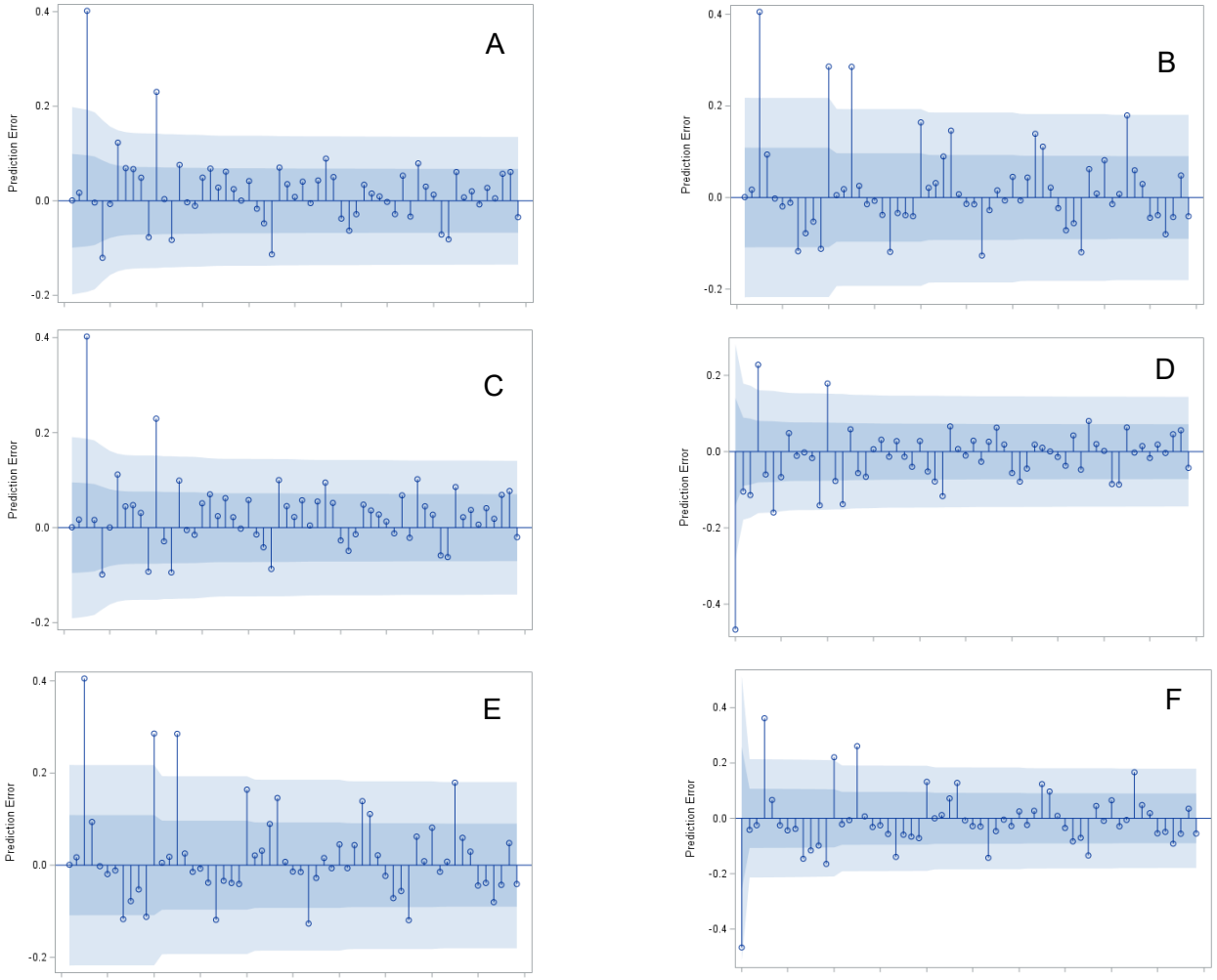

Fig. 13: Residual plot of the six UCMs. Dark blue indicates one standard error while light blue indicates two standard errors. Note: Models D and F are displayed on a different y-scale than the other models to highlight the modeling of the MIA at early ages; also, x-scale indicates months i.e., 13 = January, 19 = July, etc.

A) shows Model 1.
B) shows Model 2.
C) shows Model 3.
D) shows Model 4.
E) shows Model 5.
F) shows Model 6.

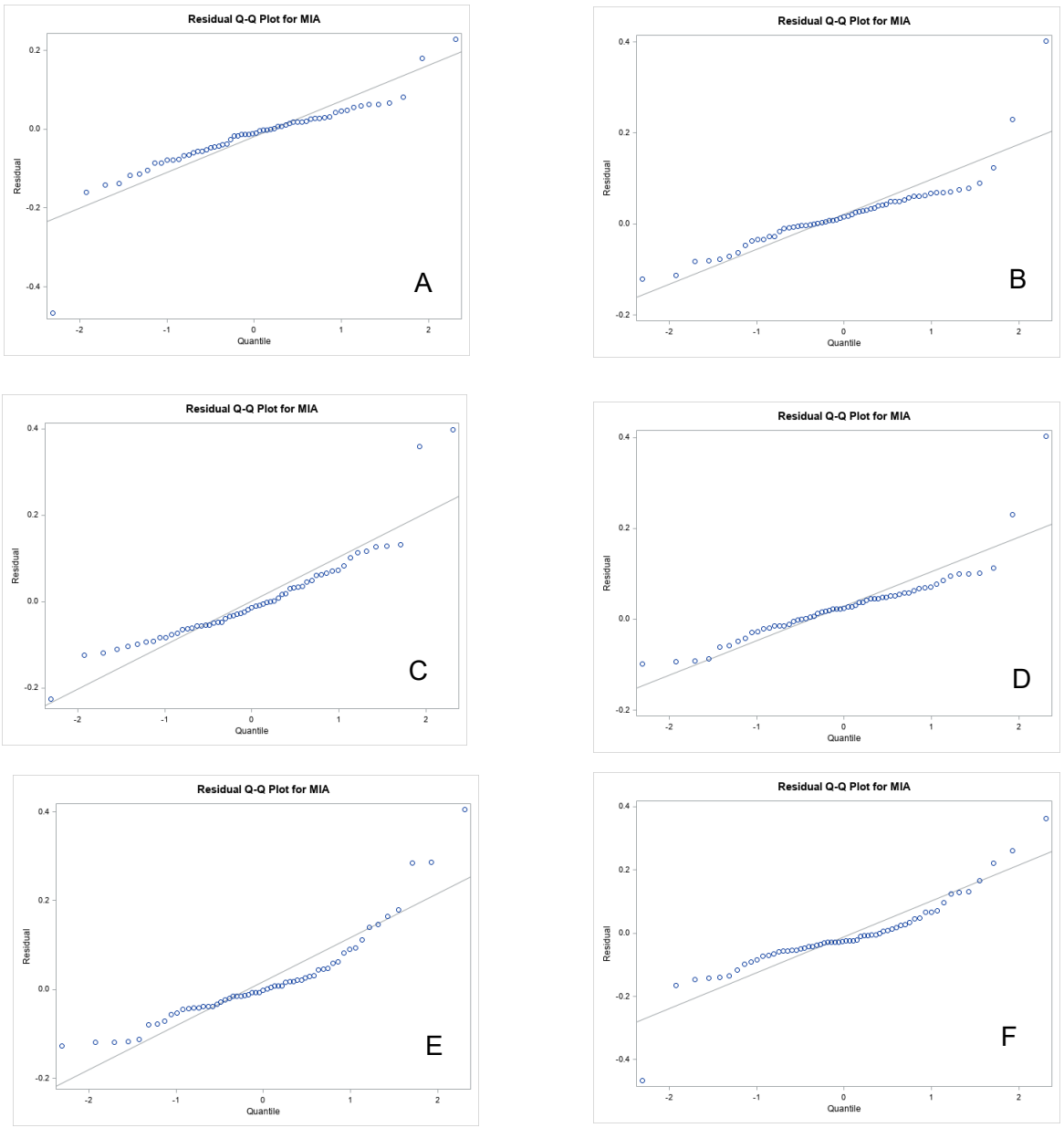

Fig. 14: QQ plots of the six UCMs.
A) shows Model 1. B) shows Model 2. C) shows Model 3.
D) shows Model 4. E) shows Model 5. F) shows Model 6.

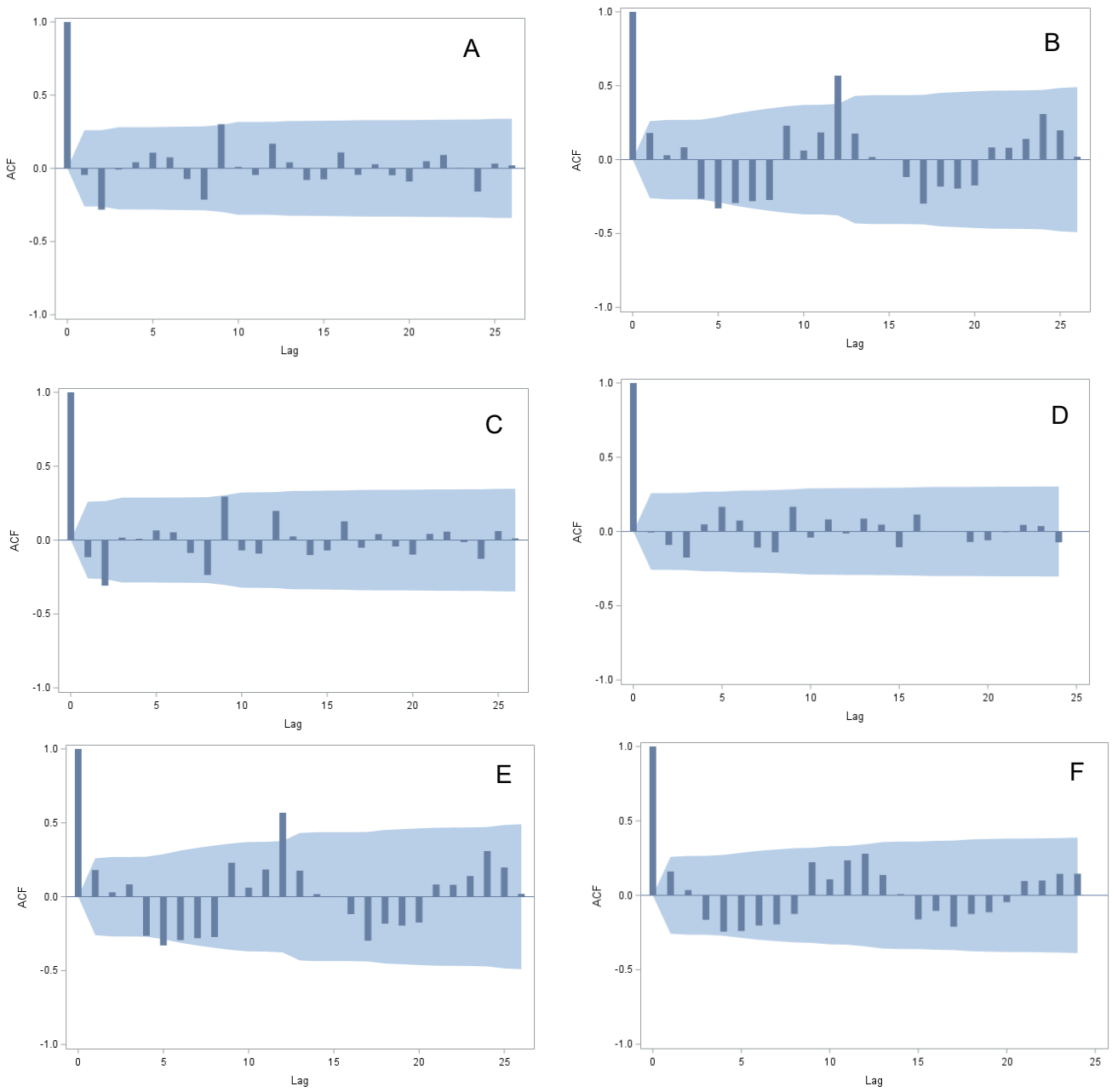

Fig. 15: Autocorrelation function plot of the six UCMs. Dark blue indicates two standard errors from mean.
A) shows Model 1.
B) shows Model 2.
C) shows Model 3.
D) shows Model 4.
E) shows Model 5.
F) shows Model 6.

3.3.4 Copula Model

Figure 16 describes the graph of the copula function. Tangent was added to the equation to properly model the drop of the sawtooth wave. The copula produced a loglikelihood, AIC, and Kendall's tau of -52.31, -188.9633, and -0.5503, respectively. Kendall's tau is negative since as time increases the marginal increment decreases. Such a result has not been captured under ARIMA or UCM. The AIC and log likelihood were smaller than the ARIMA or UCM. The variance of the copula model was larger than either the ARIMA or UCM (Figure 17). The conditional residual plot showed dips corresponding to May of each year, and QQ plots showed the majority of the data within a normal distribution (Figure 17). The marginal residual plot showed dips corresponding to May of each year, and QQ plot showed the majority of the data within a normal distribution (Figure 17). The pattern in the MIA residuals shows that ARIMA or UCM models are better fits.

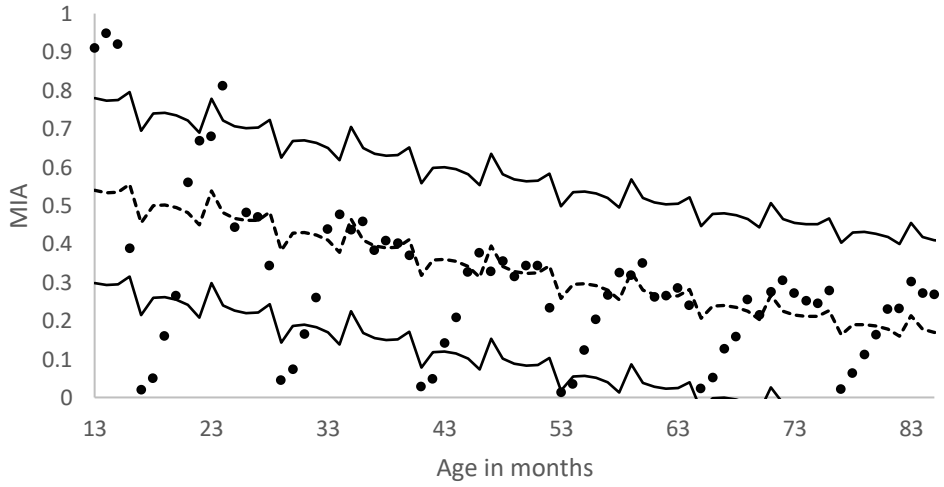

Fig. 16: Graph of Gaussian copula with beta marginal fit to model MIA data. Copula estimate indicated by the dashed line and shows the model fit; 95% confidence interval is indicated by solid lines.

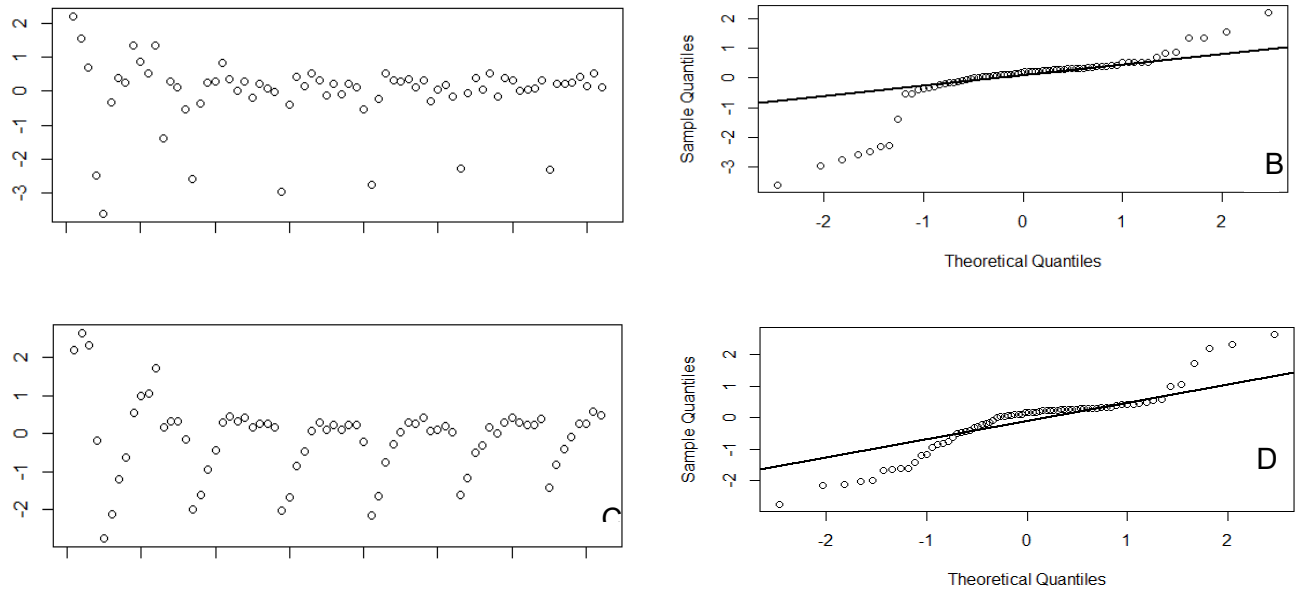

Fig. 17: A shows the conditional residuals of copula model.
B shows the QQ plot of the conditional residuals of the copula model.
C shows the marginal residuals of copula model.
D shows the QQ plot of the marginal residuals of the copula model.

3.4 Discussion

The sawtooth wave pattern of MIA data proved challenging to model, but when analyzed with ARIMA, UCM, and copula, these methods provided precise timing of accretional patterns in fish ear bones. The models demonstrated that, for Atlantic croaker, dark bands had formed by May and occurred only once during the year. Thus, providing a model to validate the formation of dark bands were evident after a sharp drop in accretion while also providing statistical metrics of model fit – attributes missing from qualitative measures.

The first step in our approach was to formulate more flexible assumptions about the dependence structure of the process. More precisely, the joint density of the accretion process could directly and conveniently describe the stochastic process. Using the dependence structure

of ARIMA, the conditional probability of the saw-tooth pattern was well described. The UCM model matched the results from the ARIMA as the autocorrelation was captured under Gaussian white noise and is also extendable when one considers the copula types of distributions. In this paper, we applied the copula ideas of Salinas-Gutierrez et al. (2009) and Alqawba and Diawara (2020), with beta marginals.

The ARIMA and UCM produced the best results. ARIMA is well established and widely used in analysis of complete datasets. The UCMs are valuable extensions of the time series with variance decreasing over time. The confidence intervals in both contain most of the recorded data and matched the seasonal pattern well. In both models, parsimony is obtained mainly because of complete data. The cycle was automatically obtained giving us close to perfect time-varying predictions.

Although the copula did not perform as well, the results were still acceptable for its use in MIA. The variance estimate is higher in this case than the ARIMA and UCM. The challenge may be due to the choice of the marginal distribution not being as good a fit to these data. The way ARIMA and UCM components were captured, as regularity of cycle and seasonality, provides less of an emphasis on sampling. The strength of copulas is in capturing a flexible correlation structure when additional variables, such as temperature or length, are measurable. In conclusion, all three models are effective tools to validate yearly accretional patterns in fish ear bone despite their differences in constraints and assumptions.

Overall, we validated the annual pattern of the MIA data with all three models (ARIMA, UCM, and copula). In this case we had a full data set and have shown we can use these three methods with quantitative results that validate the qualitative visual results seen in Barbieri et al. (1994) and Foster (2001). We anticipate that copulas will outperform ARIMA and UCMs when

challenged with incomplete data where imputation is necessary, when some variable transformations are not recommended, when missingness cannot be avoided, and the effects of covariates are not removable. In the future, we will test the performance of copulas when challenged with both the sawtooth wave pattern and incomplete datasets (Kirch et al., in draft). We hope to increase the use of copula in this field and generalize this method of estimation for higher dimension problems. In further research, the exploration of incomplete datasets and other copulas should bring in interesting results.

CHAPTER 4

Marginal Increment Analysis On A Single Cohort With Imputed Data Using Time Series Models

4.1 Introduction

Marginal increments (MI) have been widely used in fisheries to validate putative annual bands in fish hard parts, typically otoliths (Barbieri et al., 1994 and Huang et al., 2022). Marginal increment analysis (MIA) measures the ratio of newly accreted material on fish hard parts and is often done qualitatively (Barbieri et al., 1994; Huang et al., 2022; Smith, 2014). Qualitative analysis uses plots of the monthly mean increment width beyond the last observed annulus (the accretion width) and the plot is then visually inspected for the drop in accretion widths (Barbieri et al., 1994; Huang et al., 2022; Smith, 2014). Attempts to provide quantitative analysis have been limited to analysis of variance (ANOVA), wherein data are tested to find differences in mean width of new accretional material between months. However, ANOVA cannot model a multi-year pattern (Phelps et al., 2019); it assumes independence from one monthly observation to the next. A more recent quantitative approach is built from circular analysis which converts the date of capture to the middle date of its month, and then converts this date to radians with marginal increment ratios as linear vectors (Okamura et al., 2013).

Another quantitative approach to analyze MI was presented in Kirch et al. (2023), who applied novel time series models: the autoregressive integrated moving average (ARIMA), the unobserved components model (UCM), and the copula model, to an Atlantic croaker dataset that contained multiple cohorts with no missing data to demonstrate the efficacy of these approaches. Because MIA data typically violate the ARIMA assumption of constant variance over time, Kirch et al. (2023) also used a UCM, which is time series with added trend, cyclical, seasonal,

autoregressive, and lagged components. Finally, Kirch et al. (2023) applied copula models which handles the violation of the ARIMA assumptions of constant variance over time and may be a better fit for MIA. They found that ARIMA and UCM modeled MIA well when using data taken across cohorts (ignoring birth year), whereas copulas did not perform as well. Time series models allow for the evaluation of timing of annulus formation through time, unlike analysis of variance or circular analysis that compress all data into a single year.

There are two methods of data collection that validate the time course of annuli formation. Most often data coverage is optimized by collecting fish of various ages, regardless of birth-year cohort (Barbieri et al., 1994). For example, age-one fish can be born in different years and so on. Such an approach provides a more complete dataset with a larger sample size and better monthly coverage than following a single cohort. A drawback of this method is that incremental patterns can be obscured by different cohort growth conditions due to interannual changes in the environment. A second method of data collection, which is rarely used, is to follow a single cohort over time. This method has the advantage of following a strong year class to see if seasonal deposition patterns shift with age. However, this method is more likely to have missing data due to changes in availability and sparser data as specimens age and abundance declines, hence its difficulty to model (Benavides et al., 2023).

There are other challenges in modeling marginal increment data, regardless of which approach is used. MIA typically shows a sawtooth wave pattern with a steep drop when a putative annulus is formed (de Alaiza Martinez et al., 2015). This pattern can complicate model fitting because the assumptions of normality made in ARIMA and UCM models are not met. Furthermore, the sawtooth wave-like seasonality with its smooth increase and sudden drop is difficult to predict.

In addition to the challenge of the sawtooth wave pattern of marginal increment deposition, the field of fisheries can be challenged with missing data, particularly for validating age with MIA (Quinn and Deriso, 1999) . Data can be missing for a myriad of reasons, such as cost to obtain it, dearth of fish availability, or adverse environmental conditions that make collection difficult or impossible. These missing data are typically dealt with by using proxy measurements with extrapolation (Post et al., 2008; Read et al., 2006; Zainuddin and Saitoh, 2004). However, other methods not often used in MIA are formal imputation techniques. The types of imputation in this study include multiple imputation by chained equations (MICE) using a multivariate normal model, both Bayesian and non-Bayesian, and predictive mean matching (PMM). Multiple imputation methods tend to produce less biased estimators than either maximum likelihood or weighting approaches (Kleinke 2017).

Our paper explores the use of ARIMA, UCM, and copula, to quantitatively model marginal increment patterns where specimens are taken from a single cohort and data are sparse. We have previously shown that time series and UCM provide direct statistical methods to model MIA when specimens are taken over multiple cohorts (Kirch et al., 2023), but MIAs taken from single cohorts provide further challenges. There are two challenges addressed in this paper. The first is model building when missing data must be accounted for and the second is the estimation of missing data when following a single cohort. The goal is to evaluate the feasibility and efficacy of ARIMA, UCM, and copula modeling approaches when data are limited due to following a single cohort.

4.2 Materials and Methods

4.2.1 Data Collection

Fish were collected by the Virginia Marine Resource Commission in their yearly commercial survey from 2010 to 2014. The fish were collected from the Chesapeake Bay and Atlantic Ocean by trawler, seine haul, pound net, and gill net. This data set consists of the 2009 cohort of Atlantic croaker (*Micropogonias undulatus*) born during 2008. Specimens were measured, weighed, and sagittal otoliths extracted.

Extracted otoliths were sectioned with a low-speed isomet saw and then mounted onto glass slides. The sections were then magnified in a range from1x to 8x and photographed using an Olympus DP71 camera and the program Cells Sense. The images were then uploaded to the program Image-Pro Plus v. 6.2.0.424 (Media Cybernetics Inc.) and measured digitally in micrometers. The distances measured were the width of the last complete annulus and the distance of the newly accreted material, from the edge to the last completed annulus.

The raw data was then min-max scaled by taking the raw measurement, subtracting the minimum width for the age/year combination and dividing by the range of measurements for the same age/year combination.

$$Scaled\ measurement = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{65}$$

4.2.2 Imputation

Three methods of imputation are used in this paper, multivariate normal imputation, both Bayesian and non-Bayesian, and PMM. Multivariate normal imputation models tend to require a large sample size to compensate for a general underestimation of variance (Kleinke, 2017). They also perform poorly when model assumptions, such as normality and homoscedasticity, are violated. In contrast, PMM tends to be more robust to model misspecifications (Kleinke, 2017). Kleinke (2017) conducted simulations with varying sample sizes, donor pool sizes, missing data

percentages, and distribution skewness and found that PMM consistently provided better results than multivariate normal models .

Missing data was imputed using the R package MICE with three different imputation methods, multivariate normal models, both Bayesian, and non-Bayesian, and PMM (van Buuren et al., 2022).

4.2.3 Goodness of fit tests

We used the log likelihood to formulate Akaike's information criterion as

$$AIC = -2log(L) + 2K, \tag{66}$$

where $K = p$ is the number of parameters in the model and L is the corresponding log likelihood associated with the model (Burnham et al., 2011; Piegorsch and Bailer, 2005).

The mean squared error is formulated as

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2, \tag{67}$$

where Y_i is the observed value and $\hat{Y}_i$ is the predicted value.

The Kendall's tau measures the difference in the measure of concordance and discordance between the marginal CDFs is:

$$\tau = 1 - 4\int_0^1 \int_0^1 C(u,v)dC(u,v), \tag{68}$$

as in Sun et al. (2020), where *C* represents the copula function that will be described in the 'Copula' section, with *u* and *v* capturing the marginal distributions of Y_t and Y_{t-1}, respectively.

4.2.4 Time Series

An ARIMA (p, d, q) process is defined as:

$$\phi(B)(1-B)^d Y_t = \theta(B)\varepsilon_t, \tag{69}$$

where ϕ and θ are functions of the autoregressive (AR) and moving average (MA) parts, respectively, $\varepsilon_t \sim N(0, \sigma^2)$, B is the backshift operator, and d is the differencing parameter (Brockwell and Davis, 2002).

The autocorrelation function gives a measure of dependence between values at different time points and is used to estimate the order of the AR parameter (p). The sample partial autocorrelation function is used to estimate the order of the MA parameter (q) (Brockwell and Davis, 2002).

R package, Applied Statistical Time Series Analysis, or "astsa" was used to loop through possible combinations of p and q ranging from zero to three (Hyndman and Killick, 2022). The AIC, log likelihood, and MSE were used to determine the best model.

4.2.5 Unobserved components model

Unobserved components model is modeled using the equation:

$$Y_t = \mu_t + \gamma_t + \psi_t + r_t + \sum \beta_j X_{jt} + \varepsilon_t, \quad (70)$$

where,

μ_t represents the trend component,

γ_t represents the seasonal component,

ψ_t represents the cycle trend (one that is longer than seasonal),

r_t is the autoregressive term,

$\sum \beta_j X_{jt}$ is a regressive term on the independent variables where j is the number of covariates. The error term is ε_t and assumed to be independent and have identical Gaussian distributions, and μ_t, γ_t, ψ_t, and r_t are assumed to be independent of each other (Yang and Zhang, 2019).

Analysis was conducted using SAS "proc ucm." Six models were tested for each dataset, using various combinations of level (y-intercept), slope, cycle, and season and shown in table 5.. Fixed and stochastic refer to the variance in the model component. A fixed variable has a fixed variance of zero while a stochastic variable has an estimated variance in the model fit.

Table 5: Combination of all possible methods used to test the unobserved components models. Level refers to the y-intercept, Slope refers to the change in mean over time, Cycle refers to time components on time scales larger than seasonal and Season refers to the seasonal component.

	Level	**Slope**	**Cycle**	**Season**
Model 1	Stochastic	Stochastic	Stochastic	Stochastic
Model 2	Stochastic	Stochastic	--	Stochastic
Model 3	Stochastic	Fixed	Stochastic	Stochastic
Model 4	Stochastic	--	Stochastic	Stochastic
Model 5	Stochastic	Fixed	--	Stochastic
Model 6	Stochastic	--	--	Stochastic

4.2.6 Copula

We defined a copula as

$$P(Y_t \leq y_t, Y_{t-1} \leq y_{t-1}) = C\{P(Y_t \leq y_t), P(Y_{t-1} \leq y_{t-1})\},\ t = 1,2,\dots,n, \quad (71)$$

with the following requirements: the marginal distribution follows a uniform [0,1] and the probability mass is non-negative (Sun et al., 2020).

This paper used beta marginals described as

$$G(y) = \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha,\beta)}, \quad (72)$$

where y is the observed value of the marginal increment and α and β are shape parameters. Beta marginals are used since they are limited in range from zero to one which matches the range for the min/max scaled marginal increment data.

The R package "gcmr" was used to fit the copula models (Masarotto and Varin, 2017). This package has been used to capture dependence between consecutive observations using Gaussian copulas. Different choices of *p* and *q* were considered, but for the sake of simplicity ARIMA (1,0,0) was selected. Due to the time dependency, trigonometric functions were also used in the model equation (Chesneau, 2021a,b, 2022). The copula parameters were estimated and reported along with the log likelihood and AIC. Kendall's tau (τ) was calculated to estimate correlation between accretion width and time.

4.3 Results

4.3.1 Data

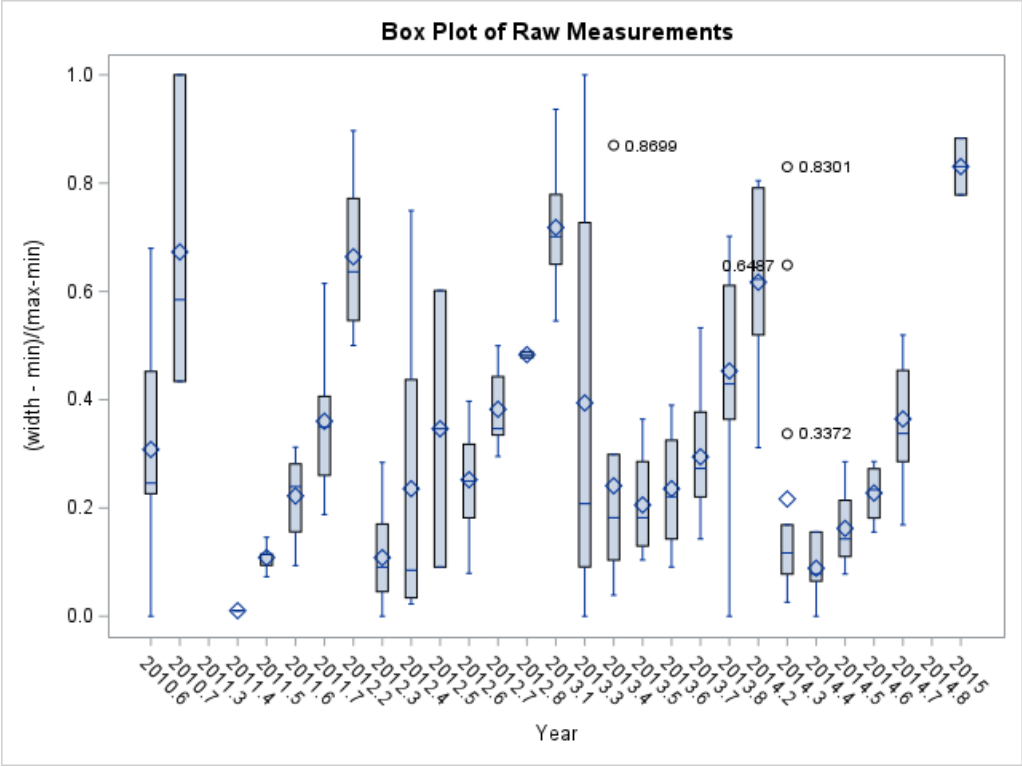

Fig. 18: Box plot of min-max marginal-increment scaled data. Decimals represent fractions of year i.e. July of 2010 = 2010.6.

Otoliths from 377 fish from the 2008 cohort were collected and increments measured covering the time span from January of 2010 through December 2014. Data is available for 32 of the months during this time period. Missing data included 28 months' worth of imputed average monthly widths (46.67%). The collected data had a mean min/max scaled value of 0.3908 and a variance of 0.03991. The PMM imputed data had a mean min/max scaled value of 0.4345 and a variance of 0.0103. The Bayesian multivariate normal imputed data had a mean of 0.4658 and a variance of 0.0125. The non-Bayesian multivariate normal imputed data had a mean value of 0.4616 and a variance of 0.0118. Data from winter months, October to January, were most often missing.

4.3.2 Imputation

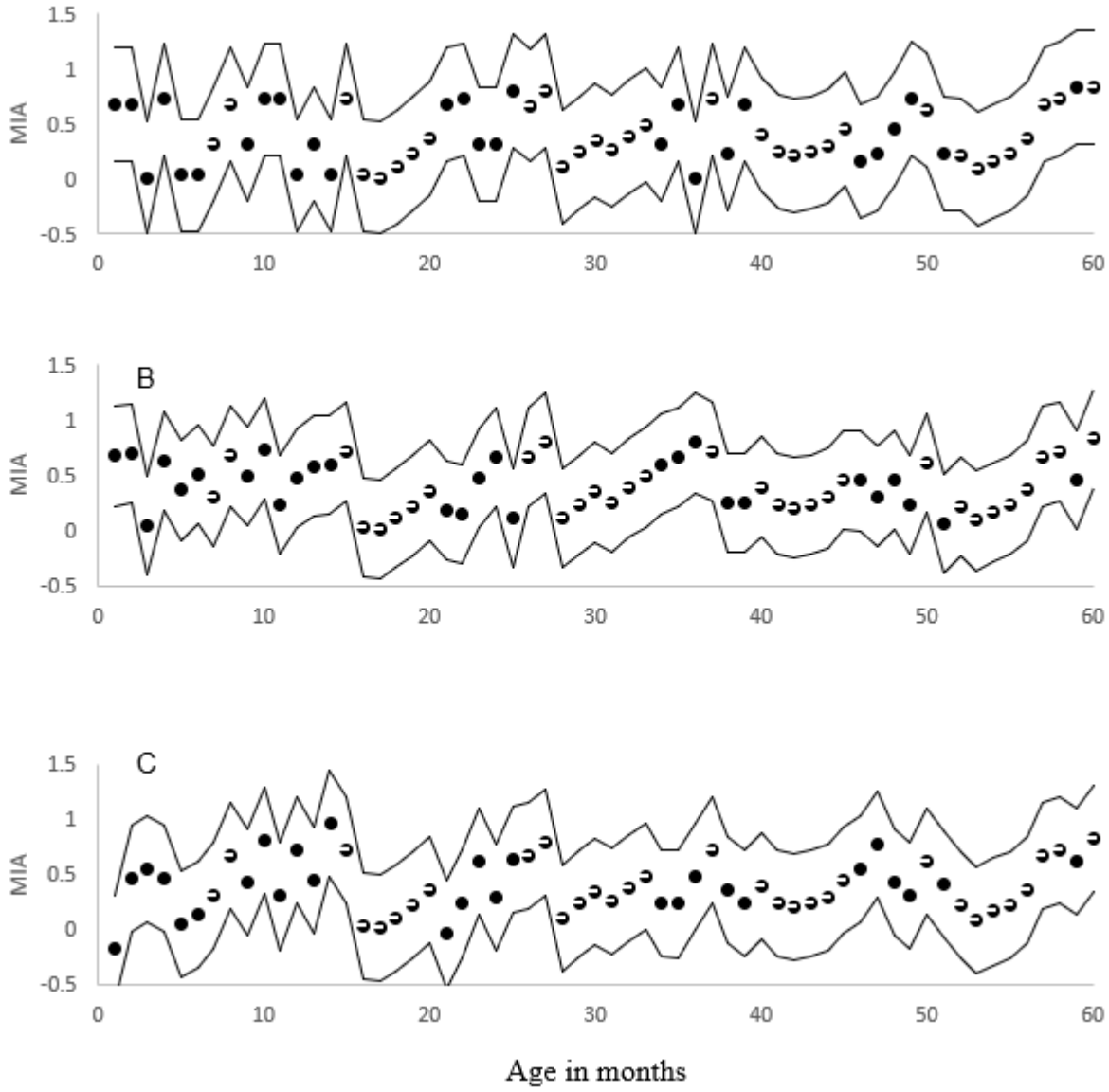

Fig. 19: Time series graph of true and imputed values with 95% confidence intervals. Solid black circles indicate imputed values. A: Predictive mean model (pmm), B: Bayesian multivariate normal model, C: non-Bayesian multivariate normal model

Both the unimputed raw and min-max feature scaled data showed a steep drop in slope around the months of April and May (Figure 18).

It is difficult to observe the seasonality in the time series graphs of the imputed data, which is evidence that no of the imputation method was visually better than another (Figure 19).

4.3.3 Time Series

The PMM imputed data had the lowest AIC value of 19.0968 with an ARIMA (3,1,2). The lowest MSE (0.05476) was an ARIMA (3,1,2), and the highest log likelihood (-2.5208) was an ARIMA (3,1,3) (Table 6).

Table 6: Model diagnostics for AIC, MSE, and log likelihood for the time series with predictive mean matching imputed data. The bolded number is the best value of the models.
A) shows AIC for selected time series models. B) shows MSE for selected time series models. C) shows Log likelihood for selected time series models.

A

p\q	0	1	2	3
0	44.3356	19.2097	20.3361	19.7689
1	27.6618	20.0343	21.6667	21.5992
2	28.9819	20.6808	21.9652	23.5982
3	28.8551	21.1578	**19.0968**	21.0416

B

p\q	0	1	2	3
0	0.1160	0.0683	0.0675	0.0650
1	0.0841	0.0673	0.0671	0.0647
2	0.0831	0.0661	0.0653	0.0647
3	0.0800	0.0640	**0.0548**	0.0549

C

p\q	0	1	2	3
0	-20.1678	-6.60483	-6.16804	-4.88446
1	-10.8309	-6.01716	-5.83335	-4.79958
2	-10.4909	-5.34041	-4.98259	-4.79911
3	-9.42757	-4.5789	-2.54838	**-2.52078**

The Bayesian multivariate normal imputed data had the lowest AIC value (0.1314), the lowest MSE (0.0391), and the highest log likelihood (7.9343) all with an ARIMA (3,1,3) (Table 7).

Table 7: Model diagnostics for the time series model with Bayesian multivariate normal imputed data. The bolded number is the best value of the models.
A) shows AIC for selected time series models.
B) shows MSE for selected time series models.
C) shows Log likelihood for selected time series models.

A

p\q	0	1	2	3
0	20.6640	4.3431	2.9163	3.6957
1	9.8182	2.0819	3.3866	5.0619
2	8.6646	3.1803	4.4459	6.0477
3	9.5683	11.3273	7.1510	**0.1314**

B

p\q	0	1	2	3
0	0.0777	0.0563	0.0504	0.0495
1	0.0622	0.0498	0.0495	0.0492
2	0.0589	0.0493	0.0477	0.0476
3	0.0577	0.0575	0.0493	**0.0391**

C

p\q	0	1	2	3
0	-8.3320	0.8284	2.5418	3.1522
1	-1.9091	2.9590	3.3067	3.4690
2	-0.3323	3.4099	3.7770	3.9762
3	0.2158	0.3364	3.4245	**7.9343**

The non-Bayesian multivariate imputed data had the lowest AIC value (5.7544) with a ARIMA (1,1,1). Its lowest MSE (0.0486) and highest log likelihood (3.3092) were obtained with an ARIMA (3,1,3) (Table 8).

Table 8: Model diagnostics for the time series model with non-Bayesian multivariate normal imputed data. The bolded number is the best value of the models.
A shows AIC for selected time series models.
B shows MSE for selected time series models.
C shows Log likelihood for selected time series models.

p\q	0	1	2	3
0	18.6800	12.6983	6.8417	6.9564
1	16.2172	**5.7544**	7.7167	8.3577
2	16.7049	7.6928	8.9880	8.8365
3	13.8461	9.0022	10.0857	9.3816

A

p\q	0	1	2	3
0	0.0751	0.0652	0.0540	0.0525
1	0.0695	0.0533	0.0532	0.0519
2	0.0677	0.0532	0.0526	0.0500
3	0.0621	0.0523	0.0515	**0.0486**

B

p\q	0	1	2	3
0	-7.3400	-3.3491	0.5791	1.5218
1	-5.1086	1.1228	1.1417	1.8212
2	-4.3524	1.1536	1.5060	2.5817
3	-1.9231	1.4989	1.9571	**3.3092**

C

4.3.4 Unobserved component model

Model one failed to converge for all three imputation methods. The model was possibly overparameterized with not enough data to capture the seasonality, trend, and cycle.

At least ten percent of the data was found outside the 95% confidence interval. *Models two* (no long-term cycle) and *four* (no slope) had 11 points (18.3%) outside the confidence interval. *Model three* (fixed slope) had eight points (13.3%) outside the confidence interval.

Model five (no long-term cycle and fixed slope) had nine points (15%). *Model six* (no long-term cycle and no slope) had 14 points (23.3%) outside the confidence interval (Figure 3). *Model six* had the lowest AIC value of 26.741, while *model four* had the lowest MSE (0.2927) and highest log likelihood (-9.667) (Table 9). Overall, *models six* and *four* seem to be the most suitable for this MIA data. Both models have no slope component meaning there is no linear trend to the data.

Table 9: Summary statistics for each of the six UCM models using AIC, root mean squared error (RMSE) and log likelihood.
A) shows predictive mean matching
B) shows Bayesian multivariate normal model
C) shows non-Bayesian multivariate normal model

Model	AIC	RMSE	Log likelihood	
1	Failed to converge			A
2	38.917	0.3270	-15.46	
3	41.729	0.3239	-14.86	
4	31.335	**0.2927**	**-9.667**	
5	36.917	0.3270	-15.46	
6	**26.741**	0.2956	-10.37	
1	Failed to converge			B
2	26.247	0.2832	-9.124	
3	24.836	0.2763	-6.418	
4	20.978	**0.2642**	**-4.489**	
5	24.247	0.2832	-9.124	
6	**15.169**	0.2648	-4.584	
1	Failed to converge			C
2	30.626	0.2943	-11.31	
3	33.369	0.2883	-10.68	
4	18.375	**0.2557**	**-3.188**	
5	28.626	0.2943	-11.31	
6	**18.348**	0.2726	-6.174	

In all the model fits for the Bayesian multivariate normal imputed data, at least twenty percent of the data was found outside the 95% confidence interval, *model two* (no cycle) had 12 points (20%) outside the confidence interval, *models three* (fixed slope) and *five* (no cycle with a

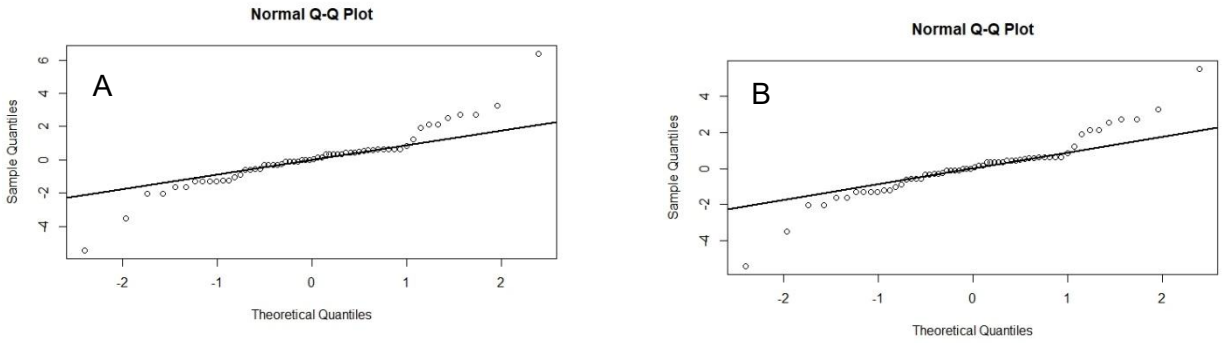

Fig. 20: QQ Plots of the predictive mean matching copula model. A shows the marginal residuals, B shows the conditional residuals.

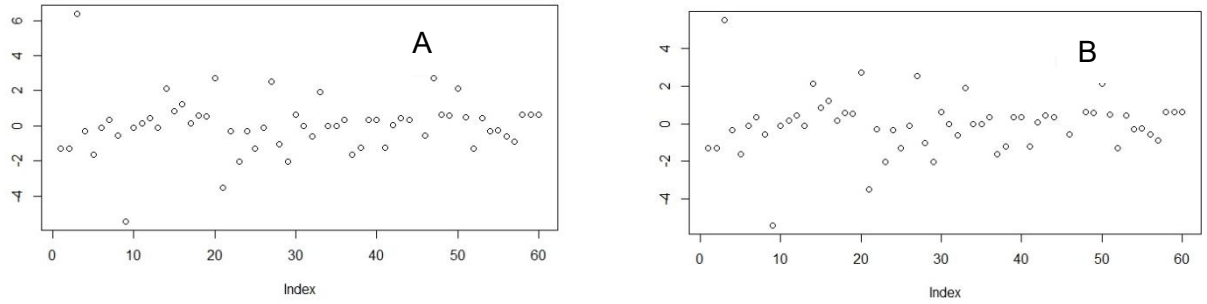

Fig. 21: Residual plots for predictive mean matching copula model. A shows the marginal residuals, while B shows the conditional residuals.

fixed slope) had 13 points (21.67%) outside the confidence interval, and *models four* (no slope) and *six* (no cycle and no slope) had 21 points (35%) outside the confidence interval (Figure 4). *Model six* had the lowest AIC value (15.169), while *model fou*r had the lowest MSE (0.2642) and highest log likelihood (-4.489) (Table 9). Both models have no slope component meaning there is no linear trend to the data.

In all the model fits for the non-Bayesian multivariate normal imputed data, at least ten percent of the data was found outside the 95% confidence interval, *model two* (no cycle) had 17 points (28.3%) outside the confidence interval, *model three* (fixed slope) had six points (10%) outside the confidence interval, *model four* (no slope) had 11 points (18.3%) outside the confidence interval, *model five* (no cycle fixed slope) has 16 points (26.7%) outside the confidence interval, and *model six* (no cycle no slope) has 18 points (30%) outside the confidence interval. (Figure 5). *Model six* had the lowest AIC value (18.348), while *model four* had the lowest MSE (0.2557) and highest log likelihood (-3.188) (Table 9). Both models have no slope component meaning there is no linear trend to the data.

4.3.5 Copula

Using the PMM and differencing, the AIC, log likelihood, and Kendall's tau were -97.459, -52.73, and 0.1464, respectively (Table 10). The conditional residual plot and QQ plot showed no obvious patterns with a well-fitting normal distribution (Figures 20 and 21). The marginal residual plot and QQ plot showed no obvious pattern in the residual value over time with a well-fitting normal distribution (Figures 20 and 21). The best fit equation is given as

$$MIA(t) = -tan\left(\frac{2\pi t}{12}\right)\left(sin\left(\frac{2\pi t}{12}\right) - \frac{2}{\pi^2}cos\left(\frac{2\pi t}{12}\right)\right) + t, \quad (74)$$

where the sinusoidal functions are used with a frequency of $\frac{2\pi}{12}$ and described in Figure 22A.

Table 10: AIC, log likelihood and Kendall's tau for the copula models with imputation methods of predictive mean model, Bayesian multivariate normal, and non-Bayesian multivariate normal.

	AIC	**Log likelihood**	**Kendall's tau**
Predictive mean matching	**-97.459**	**-52.73**	0.1464
Bayesian multivariate normal	-88.829	-48.414	-0.1324
Non-Bayesian multivariate normal	-90.108	-49.054	-0.0441

The Bayesian multivariate normal imputation and differencing model had an AIC, log likelihood, and Kendall's tau of -88.829, -46.414 and -0.1324, respectively (Table 10). The conditional residual plot and QQ plot showed no obvious patterns with a well-fitting normal

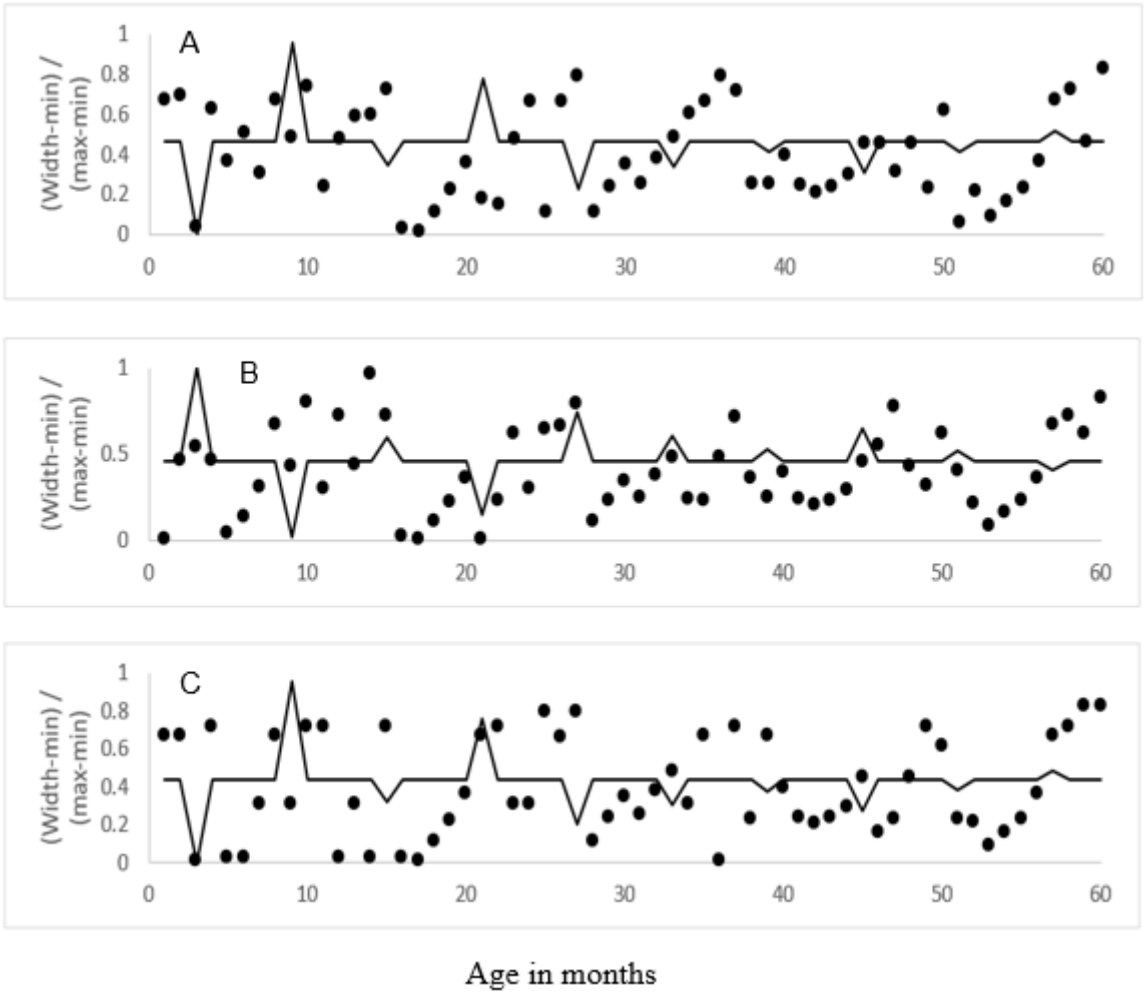

Fig. 22: Time series graphs of copula results. A shows predictive mean matching, B shows Bayesian multivariate normal model, and C shows non-Bayesian multivariate normal model. Solid circles represent data while the solid line is the best fit equation.

distribution (Figures 23 and 24). The marginal residual plot and QQ plot showed no obvious pattern in the residual value over time with a well-fitting normal distribution (Figures 23 and 24). The best fit equation is given as

$$MIA(t) = -tan\left(\frac{2\pi t}{12}\right)\left(sin\left(\frac{2\pi t}{12}\right) - \frac{2}{\pi^2}cos\left(\frac{2\pi t}{12}\right)\right) + t, \quad (75)$$

where the sinusoidal functions are used with a frequency of $\frac{2\pi}{12}$ and described in Figure 22B.

The non-Bayesian multivariate normal imputed with differencing had an AIC, log likelihood, and Kendall's tau of -90.108, -49.054, and 0.0441, respectively (Table 10). The conditional residual plot and QQ plot showed a slight sinusoidal pattern with a well-fitting normal distribution (Figures 25 and 26). The marginal residual plot and QQ plot showed a slight sinusoidal pattern over time with a well-fitting normal distribution (Figures 25 and 26). The best fit equation is given as

$$MIA(t) = -tan\left(\frac{2\pi t}{12}\right)\left(sin\left(\frac{2\pi t}{12}\right) - \frac{2}{\pi^2}cos\left(\frac{2\pi t}{12}\right)\right) + t, \quad (76)$$

where the sinusoidal functions are used with a frequency of $\frac{2\pi}{12}$ and described in Figure 22C.

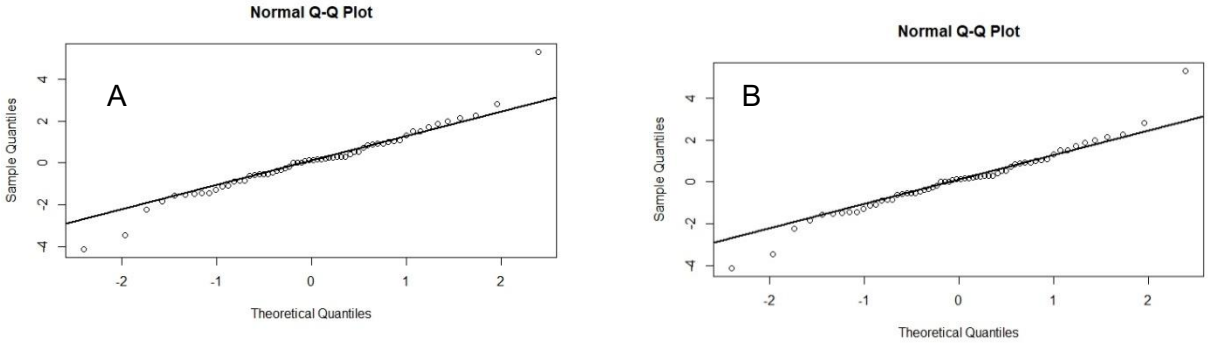

Fig. 23: QQ plot of Bayesian multivariate copula model. A shows the marginal residuals and B shows the conditional residuals.

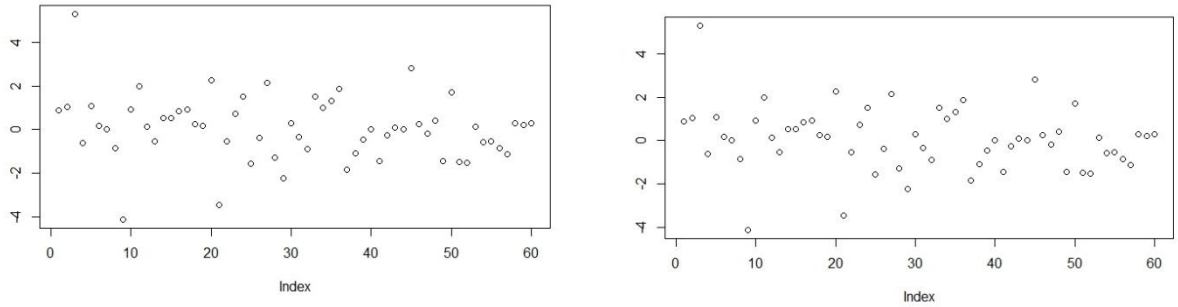

Fig. 24: Residual plot of Bayesian multivariate normal copula model. A shows the marginal residuals while B shows the conditional residuals.

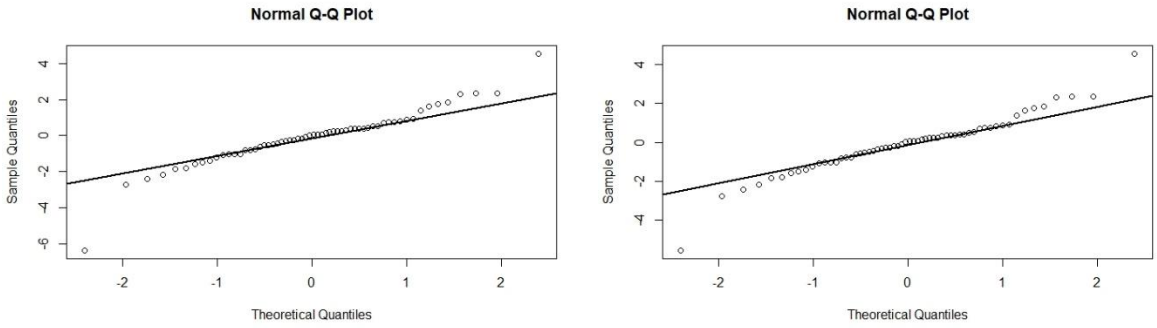

Fig. 25: QQ plot of the non-Bayesian multivariate normal copula model. A shows the marginal residuals and B shows the conditional residuals.

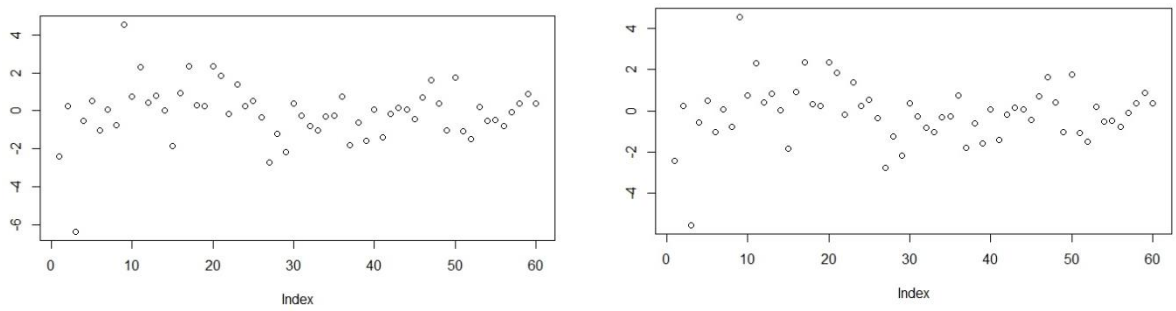

Fig. 26: Residual plot for non-Bayesian multivariate normal copula model. A shows marginal residuals and B shows conditional residuals.

4.4 Discussion

The analysis of single cohort MIA for Atlantic croaker is very valuable but challenging. We have proposed modeling data on Atlantic croaker MI using three time series models (ARIMA, UCM, and copula) while capturing missing data that is unavoidable in this single cohort data collection. The models have shown that, when following a single cohort through time, annuli in Atlantic croaker are consistently formed in March until age six with no seasonal progression in formation as the fish age (Figure 18 and Figure 22). The formation agrees with

other findings of Barbieri et al. (1994) and Kirch et al. (2023) that are based on multiple cohorts. Marginal increment data from a single cohort can now be used to validate annuli formation seasonally with age using time series models. The seasonal pattern confirms the conclusion shared in Barbieri et al. (1994) that only one annulus is formed each year for Atlantic croaker regardless of age. This yearly growth verification now comes with statistical significance with the model-based analyses (ARIMA, UCM, and copula) in contrast to models based on analysis of variance (Phelps et al., 2019). The analysis of variance pools all the yearly age information and then compresses age data into a single sawtooth wave. Our results also show that time series have advantages over the circular methods which compresses data into a single wave and shows poor fitting when uncertainty in band formation timing is high (Okamura et al., 2013). The observations of the MI are correlated and are not independent over time.

Unlike ANOVA and circular statistics, time series models have a variance/covariance matrix. This matrix allows analysis of the sawtooth itself through the variance of the specific month of the drop. A higher variance over a short time period is indicative of a steep drop, an indication of the beginning of annuli formation of a cohort occurring over a short period of time. On the other hand, a high variance over several months indicates annuli formation taking place over that longer time period.

Indeed, the copula with PMM imputation captures the seasonal pattern while compensating for the lack of representation of potential extreme values in the dataset. It also handles the imputation technique as it has the lowest AIC value compared to other time series models. The ARIMA model with Bayesian multivariate normal imputation was the next best model according to AIC, MSE, and log likelihood. UCM produced the highest AIC and lowest log likelihood with a mixture of Bayesian and non-Bayesian multivariate normal imputation.

When missing data is present, copulas are very suitable choice models. They produce the same pattern as the ground truth found in Kirch et al. (2023) and Barbieri et al. (1994).

Although temperature and environmental conditions are thought to affect otolith growth, little is known about how this would influence a cohort from year to year in timing of annulus formation. The copula and UCM models can easily add climatological variables to the model, allowing for such analysis. Specifically, variables such as temperature, salinity, or oxygen levels could add valuable understanding of changes in annuli formation. These analyses can add to a better understanding of otolith growth and validating annuli growth. Although I did not have those auxiliary data for Atlantic croaker, when available it may add insights when used with these methods.

CHAPTER 5

Using Time Series Analysis To Validate Age For Blackbelly Rosefish Off The Coast Of Virginia

5.1 Introduction

Validating annuli formation on otoliths allows fish to be accurately aged for understanding their demography. As far as I know, nothing has been done to validate blackbelly rosefish (*Helicolenus dactylopterus*) ageing under statistically robust models. Lately, Kirch et al. (2023; in draft) tested three time series models: autoregressive integrated moving average (ARIMA), unobserved components models (UCMs) and copulas to validate marginal increment analysis (MIA) on Atlantic croaker (*Micropogonias undulatus)* with missing data. However, there is a higher presence of missing data due to the cost and time required to collect blackbelly rosefish; they live farther offshore and in deeper water. Moreover, the pattern in marginal increment (MI) accretion using otolith measurement is less certain and harder to model than with Atlantic croaker.

Blackbelly rosefish are a deep-sea demersal species found in the Northeast Atlantic Ocean, Mediterranean Sea, the Gulf of Guinea, off the coast of South Africa, and the western Atlantic Ocean along the entire United States coast, Gulf of Mexico, and the Caribbean Sea (Chamberlin et al., 2023; Eschmeyer, 1969). They inhabit continental slopes at depths ranging from 200 to 1000 m (Chamberlin et al., 2023; Barsukov, 1980). Because of this depth and the near constant environmental conditions in which they live, blackbelly rosefish otoliths are harder to age and have greater variance in ageing and marginal increment measurements. Furthermore, ageing of blackbelly rosefish has been of interest to fishery scientists because of their life history and recent exploitation (Kelly et al., 1999; Chamberlin et al., 2023). Like other deep sea fishes,

blackbelly rosefish putatively have long life spans, slow growth rates, and late maturity throughout their geographic range and across subpopulations (Kelly et al., 1999).

Blackbelly rosefish are typically aged by reading annuli of whole or sectioned otoliths. Abecasis et al. (2006) found that sectioned otoliths are better for ageing older specimens due to the annuli being harder to read in older ages. Recently, Chamberlain et al. (2023) validated the yearly annuli pattern using carbon-14 analysis of eye lenses, but did not confirm the yearly pattern with a statistical model. Moreover, carbon 14 analysis confirms only total age, not the timing of seasonal deposition. In addition, White et al. (1998) studied the population off the coast of South Carolina. They had difficulty in modeling a growth curve and performed marginal increment on mostly juvenile specimens, finding annuli formation occurred from March to May. In contrast, studies in Europe show MI formation occurred between June and August in their analyses (Abecasis et al., 2006; Sami et al., 2016). There has been no resolution of this discrepancy.

The purpose of this paper is to apply statistical models to MI of blackbelly rosefish through the lifespan of fish collected of the U.S. East Coast to resolve annuli formation there. We use three approaches: ARIMA, UCM, and copula.

5.2 Materials and Methods

5.2.1 Data Collection

With support from the Center for Quantitative Fisheries Ecology (CQFE), boats were chartered to collect samples from the Norfolk Canyon in the Atlantic Ocean between 2008 and 2015. Because of the difficulty in collecting blackbelly rosefish, the Virginia Marine Resource

Commission helped supplement the blackbelly rosefish samples by reaching out to local anglers for donations. All samples were then forwarded to the CQFE for processing.

Fish characteristics were measured by recording the total length (in mm), fork length (in mm), sex, and reproductive stage according to Brown-Peterson et al., (2011) guidelines. Both sagittal otoliths were removed, cleaned, and stored dry in coin envelopes.

For each sample, one sagittal otolith was chosen at random (left vs right) (as suggested in Isidro, 1987) and placed into a Thermolyne 1440 furnace at 400ºC and baked for approximately one and a half minutes and then encased in an epoxy resin. A 0.3 mm section was then cut using a Buehler IsoMet low-speed saw, placed on a glass slide, and covered with Flotexx.

Afterward, the sections were magnified in a range from 1 to 8 times and photographed using an Olympus DP71 camera and the program Cells Sense. The images were then uploaded to the program Image-Pro Plus v. 6.2.0.424 (Media Cybernetics Inc.) and measured digitally in micrometers. The distance of the partial annulus and last completed annulus were measured. The marginal increment was calculated by taking the last incomplete annulus and dividing by the width of the last complete annulus (Hyndes, 1992).

The data of the marginal increment was then min/max transformed by taking the marginal increment reading, subtracting the minimum increment for the age, and dividing by the range of measurements for the same age. This transformation keeps the measurement between zero and one (0-100%), as it rescales numbers that otherwise could not be used, such as those greater than 100%.

To build the ARIMA, UCM, and copula models, due to the sparsity of specimens, the data has been grouped into younger fish less than ten years of age, ten years, eleven years, etc. to twenty years and then an older than twenty years group (Magnusson and Hilborn, 2007). With

this classification, the estimation of missing values/imputation technique is made reliable. The missingness in the blackbelly rosefish is well accounted for and the pattern associated with the MI is then predictable.

5.2.2 Imputation

MI data was filled with missing values. The missingness in blackbelly rosefish is unavoidable because blackbelly rosefish are hard to collect in terms of cost and time required for collection. Even with the help of volunteer donations, there were not sufficient fish to guarantee readings at each age and month category of the ageing process, hence grouping of ages. Imputing the missing values aids in validating annulus formation and allows for age-based models to be used by researchers and management teams.

Of the multiple methods of imputation, predictive mean matching (PMM) using multiple imputation by chained equations (MICE) has been found to be more robust to model misspecifications, heteroscedasticity and non-normality, since the predicted values are limited within the range of observed values which can preserve any non-linear relationships (Kleinke, 2017; Kleinke, 2018). Kleinke (2017) conducted simulations with varying sample sizes, donor pool sizes, missing data percentages, and distribution skewness and found that PMM consistently provided better results than multivariate normal models in all the scenarios tested. Kirch et al. (in draft) found the best results with PMM imputation using MICE.

Missing data were then imputed using the R package MICE with PMM (van Buuren et al., 2022).

5.2.3 Traditional graphical marginal increment analysis

Traditional marginal increment analysis is conducted by graphing the mean monthly average of the min/max converted data and then visually inspecting any pattern. This was done for the entire dataset. Since blackbelly rosefish mature at fourteen years (Kelly et al., 1999), the data was split into two groups, less than fourteen years, and fourteen and greater, to see if there was a better pattern to the data based on reproductive status. Adding these statistical tools to the blackbelly rosefish growth becomes necessary.

5.2.4 Time Series

An ARIMA(p, d, q) process is defined as:

$$\phi(B)(1 - B)^d Y_t = \theta(B)\varepsilon_t, \tag{77}$$

where ϕ and θ are functions of the autoregressive (AR), and moving average (MA), parts, respectively, $\varepsilon_t \sim N(0, \sigma^2)$, B is the backshift operator, and d is the differencing parameter (Brockwell and Davis, 2002). Without the differencing, the ARIMA turns out to be a simple ARMA(p, q) model.

We used the autocorrelation function to estimate the AR parameter (p); and the sample partial autocorrelation function to estimate the MA parameter (q).

R package "astsa" will be used to loop through possible combinations of p and q ranging from zero to three (Hyndman and Killick, 2022). The AIC, log likelihood, and variance will be used to decide the best model.

5.2.5 Unobserved Components Model

Unobserved components models are described using the equation:

$$Y_t = \mu_t + \gamma_t + \psi_t + r_t + \sum \varphi_i Y_{t-i} + \sum \beta_j X_{jt} + \varepsilon_t, \quad (78)$$

where

μ_t represents the trend component,

γ_t represents the seasonal component,

ψ_t represents the cycle trend,

r_t is the autoregressive term,

$\sum \varphi_i Y_{t-i}$ is a regressive term involving the lagged dependent variables,

$\sum \beta_j X_{jt}$ is a regressive term on the independent variables,

ε_t is the error term assumed to be independent and have identical Gaussian distributions,

and μ_t, γ_t, ψ_t, and r_t are assumed to be independent of each other (Yang and Zhang 2019).

With the addition of cycle and seasonal components, UCMs are another way to describe time series data. Analysis is conducted using SAS® version 9.4 "proc ucm." A description of the six chosen models is given in Table 11.

Table 11: Combination of all possible methods used to test the unobserved components models (UCMs).

	Level	**Slope**	**Cycle**	**Season**
Model 1	Stochastic	Stochastic	Stochastic	Stochastic
Model 2	Stochastic	Stochastic	--	Stochastic
Model 3	Stochastic	Fixed	Stochastic	Stochastic
Model 4	Stochastic	--	Stochastic	Stochastic
Model 5	Stochastic	Fixed	--	Stochastic
Model 6	Stochastic	--	--	Stochastic

5.2.6 Copula

When ARIMA assumptions are violated, such as normally distributed errors, stationary data, and equal variance across time, copulas may be a better method for analysis.

We defined a copula is as

$$P(Y_t \leq y_t, Y_{t-1} \leq y_{t-1}) = C\{P(Y_t \leq y_t), P(Y_{t-1} \leq y_{t-1})\}, t = 3, \dots, n. \quad (79)$$

Where C is some joint density obtained from the CDFs of the consecutive MIA time growth (Sun et al., 2020).

This paper will use Gaussian copula and marginals, the later described as

$$G(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}, y \in \mathbb{R}. \quad (80)$$

The R package "gcmr" was used to fit the copula models (Masarotto and Varin, 2017). The copula will be fitted using the consecutive observations of the MIA. The first two measurements were removed so that a burn-in period with several non-imputed values could be used. The copula can include different types of autoregressive and moving average structures; ARMA (3,3) is selected as a suitable choice for the gcmr package copula function. Copula can also handle trigonometric functions with time dependence. Hence, trigonometric functions will also be used in the model equation (Chesneau 2021a,b, 2022; Kirch et al., 2023). The copula parameters are estimated and reported along with measures of goodness of fit, the log likelihood and Akaike information criterion (AIC). Kendall's tau (τ) are calculated to estimate correlation between accretion width over time.

5.2.7 Goodness of fit tests

We considered three models: the ARIMA, UCM, and copula. All three time series models were run and measures of goodness of fit were captured using the log likelihood to formulate Akaike's information criterion as

$$AIC = -2log(L) + 2K, \tag{81}$$

where $K = p$ is the number of parameters in the model (Burnham et al., 2011; Piegorsch and Bailer, 2005).

The mean squared error is formulated as

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2, \tag{82}$$

where Y_i is the observed observation and $\hat{Y}_i$ is the predicted observation.

The Kendall's tau that measures the difference in the measure strength and direction of association between the time points:

$$\tau = 1 - 4\int_0^1 \int_0^1 C(u,v)dC(u,v), \tag{83}$$

as in Sun et al. (2020), where *C* represents the copula function, with *u* and *v* capturing the marginal distribution of the consecutive time points.

5.3 Results

5.3.1 Data

A total of 802 fish were collected from the Norfolk Canyon between 2008 and 2015. Out of the 802, 452 fish were collected by charter boat and 350 by angler donations. The largest sample sizes were collected in June and July with over 200 each month. None were captured in February and December, and all other months had samples below 100.

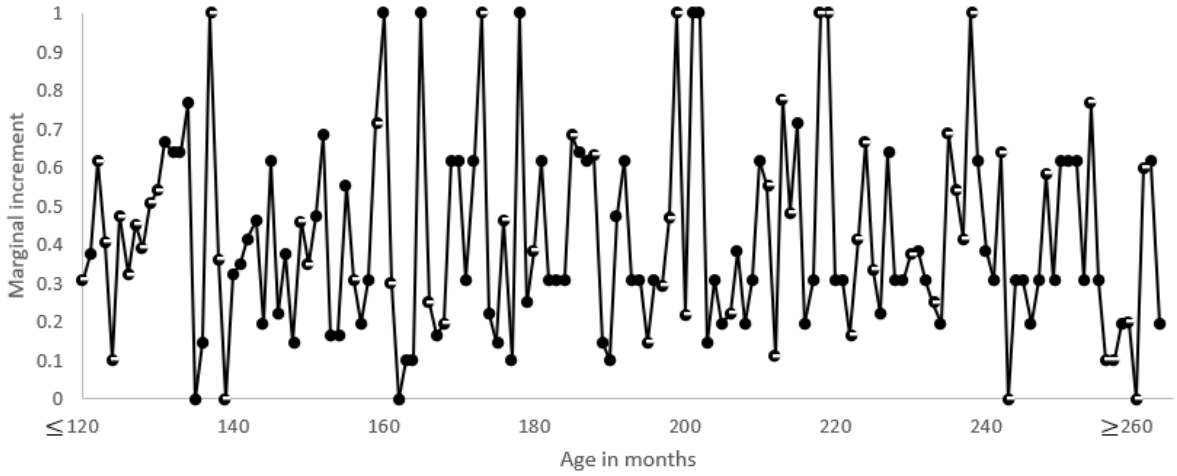

Fig. 27: Observed and imputed increment width plotted over age in months. Observed data are hollow circles.

Due to time and cost constraints in processing, otoliths from 199 fish were randomly selected from the dataset for MIA. The mean of the marginal increment of the raw data was 0.4527 with a standard deviation of 0.2747. The mean of the predictive mean matching marginal increment data was 0.4175 with a standard deviation of 0.2537. Figure 27 shows the graph of the time series of data.

5.3.2 Traditional graphical marginal increment analysis

Traditional MIA relies on the qualitative review of the graph of MI by month. For the entire sample, there is no visibly discernable pattern to the data over all ages. There is a decrease in the marginal increments for the under fourteen-year-olds in May, compared to those fish over fourteen years old (Figure 28). This indicates that one annulus is produced per year in this age

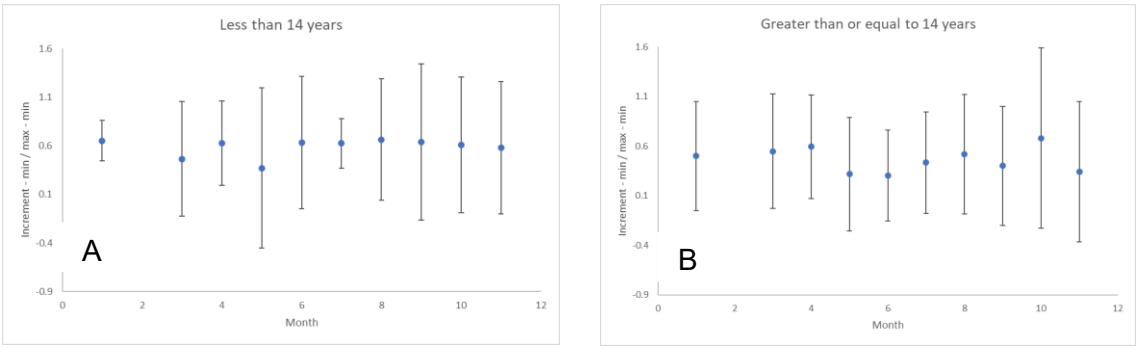

Fig. 28: Traditional marginal increment analysis. A shows the graph for blackbelly rosefish less than 14 years. B shows the graph for blackbelly rosefish 14 years and older.

group. Under this traditional method, the variance is too large for us to make any meaningful description of the change in MI even when breaking up the data in terms of age group. To add statistical significance to the analysis, ARIMA, UCM, and copulas are used in the next sections.

5.3.3 ARIMA

The model (3,0,3) has the lowest AIC, 12.21; MSE, 0.0656; and highest log likelihood, 1.9. (Table 12). All the parameters were considered significant by an alpha value of 0.05 in the (3,0,3) model.

Measures of goodness of fit showed the residuals have a zero mean and have no particular pattern outside of the twelve months spike associated with the yearly annuli formation (Figure 29). The Ljung-Box statistics were all above 0.05 meaning the observations were not correlated, indicating a good fit to the ARIMA (3,0,3).

Table 12: Model diagnostics for the ARMA (p, q) model. The bold values show the best model parameters.
A) shows the AICs for selected ARIMA models.
B) shows the MSE for selected ARIMA models.
C) shows the log likelihood for selected ARIMA models.

A

p/q	0	1	2	3
0	16.638	17.9523	14.7897	16.7883
1	18.1919	16.8829	16.783	14.1231
2	15.7735	17.1037	14.5534	15.6319
3	17.5654	18.9834	15.5845	**12.2096**

B

p/q	0	1	2	3
0	0.0639	0.0636	0.0613	0.0613
1	0.0609	0.0637	0.0623	0.0613
2	0.0618	0.0615	0.0594	0.0583
3	0.0617	0.0614	0.0582	**0.0545**

C

p/q	0	1	2	3
0	-6.319	-5.9761	-3.3949	-3.3941
1	-6.096	-4.4414	-3.3915	-1.0615
2	-3.8867	-3.5519	-1.2767	-0.8159
3	-3.7827	-3.4917	-0.7922	**1.8952**

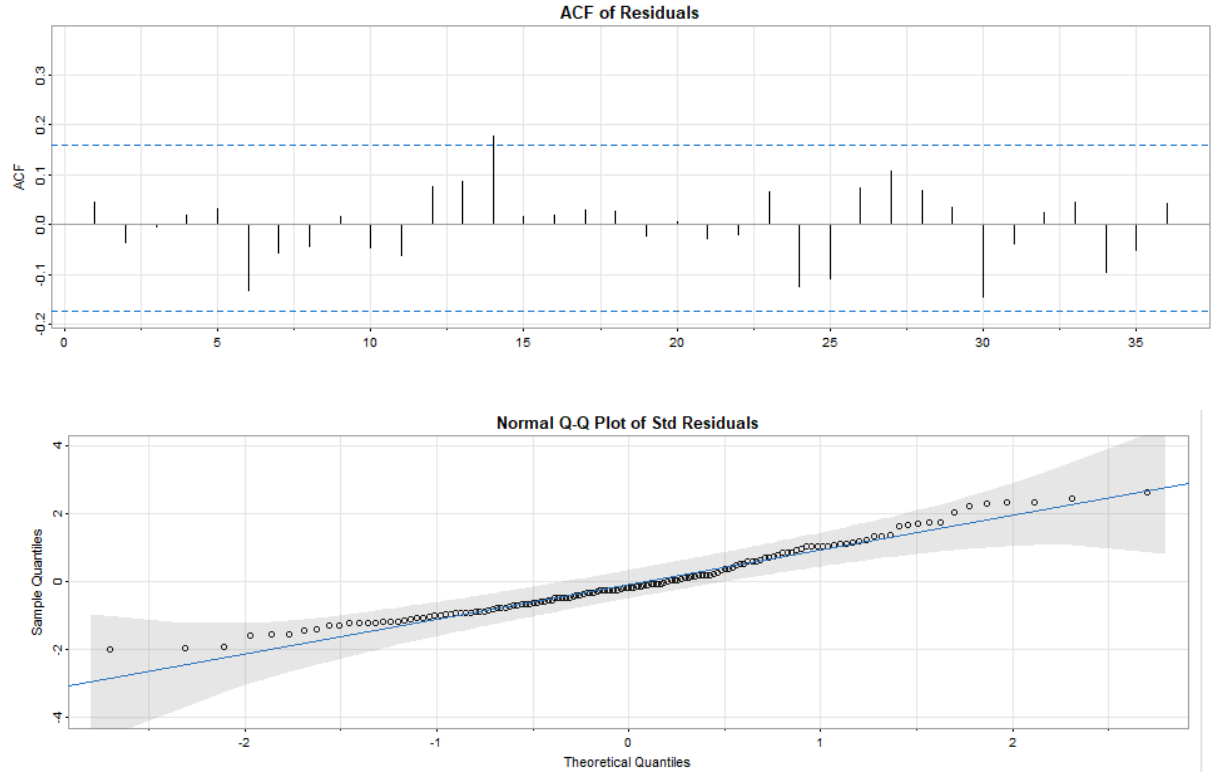

Fig. 29: ACF of residuals (A) and QQ plot (B) of the time series model.

5.3.4 UCM

Model six has the lowest AIC of 53.81 while model 4 has the lowest MSE of 0.0808 and highest log likelihood of -23.66 (Table 13). QQ plots show the tails of the data violate the normality assumption (Figure 30). All estimates for seasonality are zero.

Table 13: Summary statistics for each of the six UCM models. The lowest AIC is model 6, the lowest MSE is model 4, and the largest log likelihood is model 4.

Model	AIC	MSE	Loglikelihood
1	75.031	0.0870	-30.52
2	69.031	0.0870	-30.52
3	73.031	0.0870	-30.52
4	59.322	**0.0808**	**-23.66**
5	67.031	0.0870	-30.52
6	**53.81**	0.0811	-23.9

5.3.5 Copula

The best fit equation is:

$$MIA_t = ln\left(-tan\left(\frac{2\pi t}{12}\right)\left(sin\left(\frac{2\pi t}{12}\right) + \frac{2}{\pi^2}cos\left(\frac{2\pi t}{12}\right)\right) + t\right). \quad (84)$$

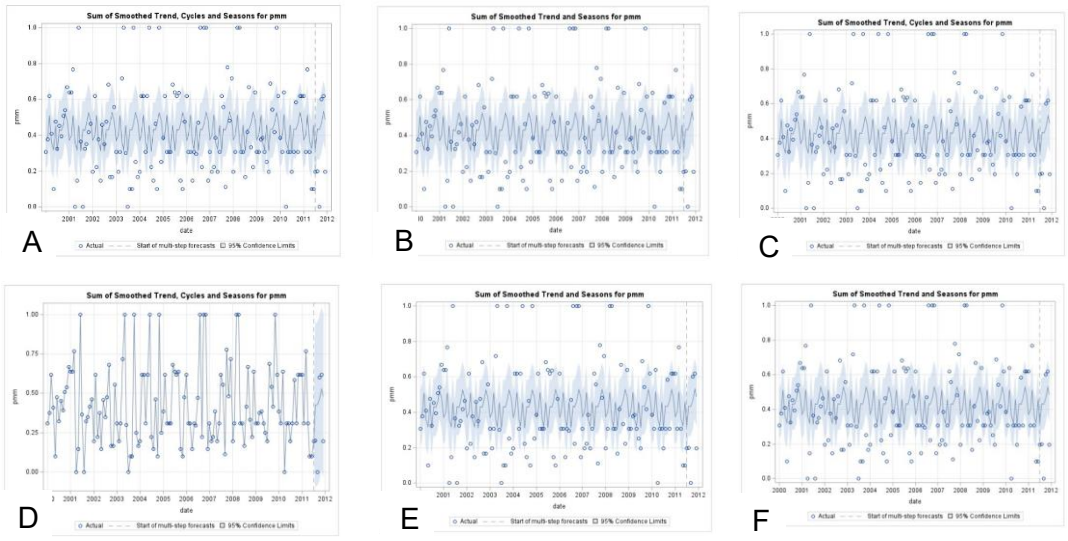

Fig. 30: MIA data fit from each of the six unobserved components models (UCMs). The light blue area is the 95% confidence interval.

This can include different types of autoregressive, moving average structures. ARMA (3,3) was selected to build upon the time series analysis. The AIC is 18.8, while the log likelihood is 0.0400. Kendall's tau was -0.0292.

The copula fit to the data is not very good (Figure 31). But, The QQ plot of the conditional residuals shows a close to normal distribution while the marginal residuals show a poor normality fit (Figure 32).

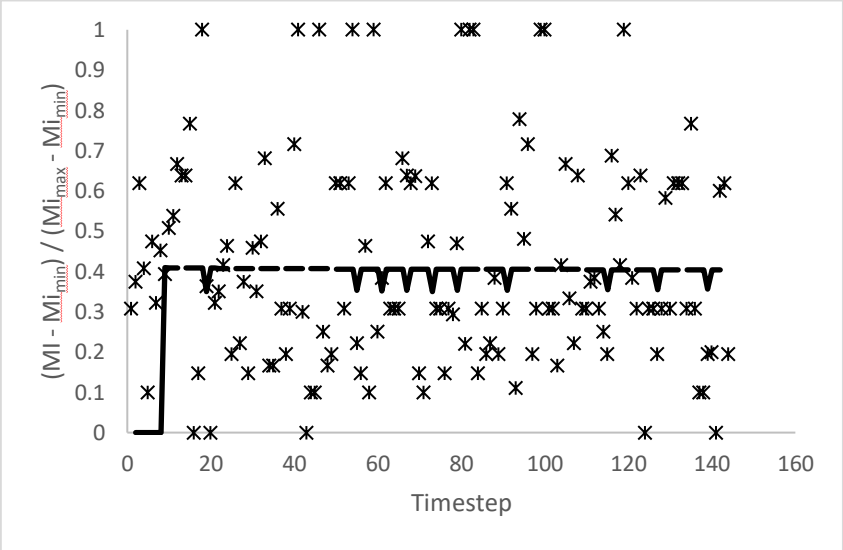

Fig. 31: Graph of copula fit to imputed data.

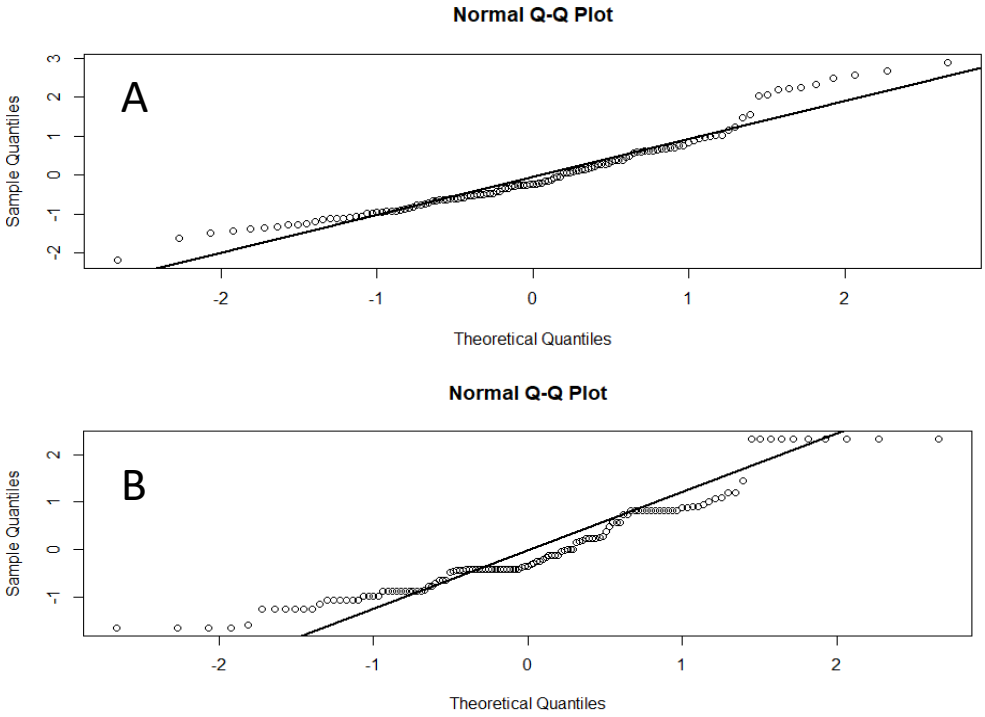

Fig. 32: Conditional (A) and marginal (B) residual QQ plot of the copula model.

5.4 Discussion

As a deep-sea fish, blackbelly rosefish have proven to be difficult in validating ages because of their longevity, difficulty in collecting specimens, and lack of strong environmental seasonal signature. Studies in Europe show annuli formation in June to August (Abecasis et al., 2006; Isidro 1987; Sami et al., 2016), while the study in South Carolina (White et al., 1998) limited their results to mostly juvenile specimens with a annulus formation in May. Empirical methods for ageing blackbelly rosefish were based on visually observing a sectioned otolith graph of the traditional MIA, and no statistical model was proposed to describe the autoregressive MI of the age. In this paper, I was able to fit time series statistical models to MIs of blackbelly rosefish from the Norfolk Canyon area off the coast of Virginia. I confirm a May timing of forming their annulus over the entire lifetime.

Of the statistical models, ARMA provided the best fit. By integrating the time dependence with MI**,** we did not need to separate the data into two groups, juveniles and adults. The estimates of the MI distributions under ARMA, UCM, and copula were viable options in modeling MIA.

The ARMA fit the data and captured the yearly growth because of the dependence of the data in the MI. The visual inspection of the graphs has limitations that the ARMA does not. The value of time series is that there is less variance in the monthly observations when the data is spread over several years instead of compressed into one year in traditional MIA. In other words, the between year variance is lower than the within year variance.

The copula model produced the second best model. It accounted for the huge variance in the dataset and the violation of the constant normality assumption of the errors. The copula relaxes the assumption of normality in the errors while capturing the time dependence under

ARMA (3,3). The ARMA model works better since the sawtooth pattern is more noticeable (Kirch et al., 2023) in contrast to the copula model. We suspect that the copula may be over parameterizing the model, indicating a fit not as appropriate as ARMA. Still, these two models allow us to provide the age distribution function for the MI of blackbelly rosefish. UCM produced the worst results with the highest AIC value. This is likely due to the large amount of missing data in the set which UCM does not handle as well as the other two models even with imputation.

These analyses give a framework for MI in the biological synthesis associated with the blackbelly rosefish with a preponderance of missing values and sawtooth pattern. To my knowledge, I am the first to implement and give a statistical technique to assess the MIA change over the lifetime for blackbelly rosefish.

According to the data measurements, the growth of the MIA is governed by: i) the periodicity with significant growth that is noticeable every year; ii) at the month of May, there is a dominant, non-ignorable sawtooth pattern; iii) growth is dictated by age. Such findings lead to great tools for blackbelly rosefish conservation. Although these cannot reveal the difference in annulus timing between East Coast and European stocks, I have shown that timing of annulus formation is not a factor of age.

CHAPTER 6

Conclusion

There are many issues with ageing fish; techniques have been studied for over one hundred years. Finding solutions for accurate ageing and age estimation are critical to the conservation of fishes and fisheries. Previous studies have used qualitative analyses or quantitative analyses that violate underlying stated assumptions. This dissertation has used quantitative statistical methods that properly handle model assumptions, instead of the traditional qualitative method of ageing fish through MIA.

Time series models can now be used to help validate annuli formation in multiple fish species with varying life histories. ARIMA, UCMs, and copulas provide a method to model MI statistically. In general, the best methods to use depends upon the shape and characteristics of the data. The data are identified with a sawtooth pattern. The ARIMA model worked better when the sawtooth pattern was evidenced in the data. When the sawtooth was not as distinct, the copula provided the best fit. In addition, all methods can handle the missingness associated with the data by estimation through PMM by way of MICE.

ARIMA requires the assumption of equal variance across time and stationarity of the data. While UCMs can handle nonstationary data with a trend component it can also incorporate covariates that ARIMA cannot. Copula relaxes the assumptions of both models and is built from the data's CDF while also allowing for covariate addition.

In this work, time series models proved to be better methods to quantify MIA than either ANOVA or circular statistics, the other two methods used in the literature. ANOVA analyzes the data under the assumptions of Gaussian errors with constant variance and each month is

independent of other months. Then it compares the months to identify significant difference among the MI measurements, and cannot analyze data longer than one year. Circular statistics assumes normality of the data, and requires conversion of the date to radians, losing the ability for easy interpretation. Neither of these methods are able to fit the sawtooth pattern explicitly.

The time series methods allow for proper analysis even when a pattern is not apparent visually in the graph. The MI graphs of the blackbelly rosefish did not have an apparent pattern, but the ARIMA was able to discern the yearly pattern because the between year variance was lower than the within year variance. The spread of the data in each month is lower when looking at the data spread over years rather than just one year.

MIA can now be conducted on a single cohort through time series analysis. Imputation replaces the missing values, allowing for the time series analysis. Because of the missing data, a single cohort is rarely followed. Following a single cohort allows for data that is more standardized across samples. The specimens have lived similar life histories with the same environmental conditions. Hence, the effects of time can be examined.

This work can be extended in multiple ways. The addition of covariates with the UCM and copula models can help with model fitting and provide for deeper analysis. It is believed that environmental conditions influence otolith growth. Having these extra data points can add further insight into otolith growth and hence fish growth. Different methods of imputation could also be tested. Many methods are not appropriate for missing not at random data. Further study could find an imputation method that works better for this type of missingness.

References

(2007). Magnuson-Stevens Fishery Conservation and Management Act.

(2023). Fisheries economics of the United States 2020. D. o. Commerce.

Abbott, J. K., Haynie, A.C. (2012). "What are we protecting? Fisher behavior and the unintended consequences of spatial closures as a fishery management tool." Ecological Applications 22(3): 762-777.

Abecasis, D., Costa, A.R., Perreira, J.G., Pinho, P.M.R. (2006). "Age and growth of bluemouth, *Helicolenus dactylopterus* (Delaroche, 1809) from the Azores." Fisheries Research 79: 148-154.

Abesamis, R. A., Green, A.L., Russ, G.R., Jadloc, C.R.L. (2014). "The intrinsic vulnerability to fishing of coral reef fishes and their differential recovery in fishery closures." Rev Fish Biol Fisheries.

Afrifa-Yamoah, E., Mueller, U.A., Taylor, S.M., Fisher, A.J. (2020). "Missing data imputation of high-resolution temporal climate time series data." Meteorological Applications.

Afrifa-Yamoah, E., Taylor, S.M., Mueller, U. (2021). "Modeling climatic and temporal influences on boating traffic with relevance to digital camera monitoring of recreational fisheries." Ocean and Coastal Management 215.

Alanazi, F. (2021). The spread of COVID-19 at hot-temperature places with different curfew situations using copula models. 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA).

Allison, P. (2015). "Imputation by Predictive Means Matching: Promise and Peril." Retrieved 3/10/24, 2024, from https://statisticalhorizons.com/predictive-mean-matching/.

Almetwally, E. M., Dey, S., Nadarajah, S. (2022). "An overview of discrete distributions in modelling COVID-19 data sets." Sankhya A: The Indian Journal of Statistics: 1-28.

Alqawba, M., Diawara, N. (2020). "Copula-based Markov zero-inflated count time series models with applications." Journal of Applied Statistics: 1-18.

Alqawba, M., Fernando, D., Diawara, N. (2021). "A class of copula-based bivariate Poisson time series models with applications." Computation 9(108): 1-12.

Alqawba, M. S., Diawara, N., Chaganty, N.R. (2019). "Zero-inflated count time series models using Gaussian copula." Sequential analysis 38(3): 342-357.

Anderson, L. G. (2014). "An economic analysis of highgrading in ITQ fisheries regulation programs." Marine Resource Economics 9: 209-226.

Arostegui, M. C., Anderson, C.M., Benedict, R.F., Dailey, C., Fiorenza, E.A., Jahn, A.R. (2021). "Approaches to regulating recreational fisheries: Balancing biology with angler satisfaction." Rev Fish Biol Fisheries.

Azur, M. J., Stuart, E.A., Frangakis, C., Lead, P.J. (2011). "Multiple imputation by chained equation: What is it and how does it work?" International Journal of Methods in Psychiatric Research 20(1): 40-49.

Baraldi, A. N., Enders, C.K. (2010). "An introduction to modern missing data analyses." Journal of School Psychology 48: 5-37.

Barbieri, L., Chittenden, M.E., Jones, C.M. (1994). "Age, growth, and mortality of Atlantic croaker, *Micropogonias undulatus*, in the Chesapeake Bay Region, with a discussion of apparent geographic changes in population dynamics." Fishery Bulletin 92(1): 1-12.

Barbieri, L., Chittenden, M.E., Lowerre-Barbieri, S.K. (1994). "Maturity, spawning, and ovarian cycle of Atlantic croaker, *Micropogonias undulatus*, in the Chesapeake Bay and adjacent coastal waters." Fishery Bulletin 92: 671-685.

Barsukov, V. V. (1980). "Subspecies of the Atlantic blackbelly rosefish, Helicolenus dactylopterus (Dela Roche, 1809)." Journal of Ichthyology 19: 1-17.

Benavides, I. F., Santacruz, M., Romero-Leiton, J.P., Barreto, C., Selvaraj, J.J. (2023). "Assessing method for multiple imputation of systematic missing data marine fisheries time series with a new validation algorithm." Aquaculture and Fisheries 8: 587-599.

Bian, Z., Zhang, Z., Liu, X., Qin, X. (2019). "Unobserved component model for predicting monthly traffic volume." J. Trans. Eng. 145(12): 1-10.

Black, B. A., Boehlert, G.W., Yaklavich, M.M. (2005). "Using tree-ring cross dating techniques to validate annual growth increments in long-lived fishes." Canadian Journal of Fisheries and Aquatic Sciences 62: 2277-2284.

Box, G. E. P., Jenkins G.M., Reinsel, G.C., Ljung, G.M. (2016). Time Series Analysis: Forecasting and Control 5th edition. Hoboken, NJ, John Wiley & Sons, Inc.

Brockwell, P. J., David, R.A. (2002). Introduction to Time Series and Forecasting. New York, Springer Science+Business Media, Inc.

Brown-Peterson, N. J., Wyanski, D.M., Saborido-Rey, F., Macewicz, B.J., Lowerre-Barbieri, S.K. (2011). "A standardized terminology for describing reproductive development in fishes." Marine and Coastal Fisheries 3(1): 52-70.

Burnham, K. P., Anderson, D.R., Huyvaert, K.P. (2011). "AIC model selection and multimodal inference in behavioral ecology: Some background, observations, and comparisons." Behav Ecol Sociobiol 65(23): 23-35.

Campana, S. (2001). "Accuracy, precision and quality control in age determination, including a review of the use and abuse of age validation methods." Journal of fish biology 59(2): 197-242.

Chamberlin, D. W., Siders, Z.A., Barnett, B.K., Patterson, W.F. (2023). "Eye lens-derived delta carbon 14 signatures validate extreme longevity in the deepwater scorpaenid blackbelly rosefish (*Helicolenus dactylopterus*)." Scientific Reports 13(7438): 1-11.

Chesneau, C. (2021a). "On new types of multivariate trigonometric copulas." Applied Math: 3-17.

Chesneau, C. (2021b). "A study of the power-cosine copula." Open J. Math. Anal. 5(1).

Chesneau, C. (2022). "A note on a simple polynomial-sine copula." Asian J. Math. Appl.: 1-14.

Cryer, J. D., Chan, K. (2008). Time Series Analysis: With Applications in R 2nd Edition. New York, NY, Springer Science + Business Media, LLC.

Davey, A., Savla, J. (2010). Statistical Power Analysis with Missing Data: A Structural Equation Modeling Approach. New York, NY, Routledge Taylor & Francis Group.

de Alaiza Martinez, P. G., Babushkin, I., Berge, L., Skupin, S., Caberera-Granado, E., Kohler, C., Morgner, U., Husakou, A., Hermann, J. (2015). "Boosting terahertz generation in laser-field ionized gases using a sawtooth wave shape." Physical Review Letters(114).

Donders, A. R. T., van der Heijden, G.J.M.G, Stijnen, T., Moons, K. G. M. (2006). "Review: A gentle introduction to imputation of missing values." Journal of Clinical Epidemiology 59(10): 1087-1091.

Elmezouar, Z. C., Almanjahie, I.M., Ahmad, I., Laksaci, A. (2021). "Comparison of the ARFIMA, ARIMA, and artificial neural models to forecast the total fisheries production in India." Journal of Animal and Plant Sciences 31(5): 1477-1484.

Eschmeyer, W. N. (1969). "A systematic review of the scorpionfishes of the Atlantic Ocean (Pisces: Scorpaenidae)." Occas Pap Calif Acad Sci 79: i-iv+ 1-143.

Flinn, S. A., Midway, S.R. (2021). "Trends in growth modeling in fisheries science." Fishes 6(1): 1-18.

Foster, J. R. (2001). Age, growth, and mortality of Atlantic croaker, *Micropogonias undulatus*, in the Chesapeake Bay region.

Gebremedhin, S., Bruneel, S., Getahun, A., Anteneh, W., Goethals, P. (2021). "Scientific methods to understand fish population dynamics and support sustainable fisheries management." Water Research 13(574): 1-20.

Godet, L., Decaulne, A., Poirier, C. (2020). "Dating saltmarshes using tree rings on a halophilous plant." Wetlands Ecol Manage.

Gumus, A., Bostanci, D., Yilmaz, S., Polat, N. (2007). "Age determination of *Scardinius erythrophthalmus* (Cyprinidae) inhabiting Befra Fish Lakes (Samsun, Turkey) based on otolith readings and marginal increment analysis." Cybium 31(1): 59-66.

Guolo, A. and Varin, C. (2014). "Beta regression for time series analysis of bounded data, with applications to Canada Google flu trends." The Annals of Applied Statistics 8(1): 74-88.

Haddon, M. (2011). Modelling and Quantitatibe Methods in Fisheries. Boco, Raton, FL, Chapman & Hall.

Haimerl, P. and Hartl, T. (2023). "Modeling COVID-19 infection rates by regime-switching unobserved components models." Econometrics 11(10): 1-15.

Hosack, G. R., Peters, G.W., Ludsin, S.A. (2014). "Interspecific relationships and environmentally driven catchabilities estimated from fisheries data." Can. J. Fish. Aquat. Sci. 71: 447-463.

Huang, S., Chang, S., Lai, C., Yuan, T., Weng, J., He, J. (2022). "Length-weight relationships, growth models of two croakers (Pennahia *macrocephalus* and *Atrobucca nibe*) off Taiwan and growth performance indices of related species." Fishes 7(281).

Hyndes, G. A., Loneragan, N.R., Potter, I.C. (1992). "Influence of sectioning otoliths on marginal increment trends and age and growth estimates for the flathead Platycephalus speculator." Fishery Bulletin 90: 276-284.

Hyndman, R. J., Killick, R. (2022). CRAN Task View: Time Series Analysis. Version 2022-10-23.

Isidro, E. J. (1987). Age and growth of the blumouth, Helicolenus dactylopterus dactylopterus (De la Roche, 1809) off the Azores. International Counsil for the Exploration of the Sea.

Joe, H. (2015). Dependence modeling with copulas. Boca Raton, FL, CRC Press.

Kelly, C., Connolly, P., Bracken, J. (1999). "Age estimation, growth, maturity, and distribution of the bluemouth rockfish *Helicolenus d. dactylopterus* (Delaroche 1809) from the Rockall Trough." ICES Journal of Marine Science 56(1): 61-74.

Kim, J., Sungur, E.A., Choi, T., Heo, T. (2011). "Generlized bivariate copulas and their properties." Model Assisted Statistics and Applications 6: 127-136.

Kirch, K. S., Diawara, N. Jones, C.M. (2023). "Fitting time series models to fisheries data to ascertain age." Journal of Probability and Statistics 2023: 1-10.

Kleinke, K. (2017). "Multiple imputation under violated distributional assumptions: A systematic evaluation of the assumed robustness and predictive mean mataching." Journal of Educational and Behavioral Statistics 42(4): 371-404.

Kleinke, K. (2018). "Multiple imputation by predictive mean matching when sample size is small." Methodology 14(1): 3-15.

Koopman, S. J., Ooms, M. (2004). Forecasting daily time series using periodic unobserved components time series models. Tinbergen Instititue Discussion Paper Tinbergen Institute, Amsterdam and Rotterdam. No -4-135/4.

Landler, L., Ruxton, G.D., Malkemper, E.P. (2018). "Circular data in biology: Advice for effectively implementing statistical procedures." Behavioral Ecology and Sociobiology 72(128): 1-10.

Little, T. D., Jorgensen, T.D., Lang, K.M., Moore, E.W.G. (2014). "The joys of missing data." Journal of Pediatric Psychology 39(2): 151-162.

Magnusson, A., Hilborn, R. (2007). "What makes fisheries data informative?" Fish and Fisheries 8: 337-358.

Makridakis, S., Wheelwright, S.C., McGee, V.E. (1983). Forecasting: Methods and Applications. New York, John Wiley & Sons.

Marsh, C., Sibanda, N., Dunn, M., Dunn, A. (2015). "A copula-based habitat preference index in fish spatial population modelling." Procedia Environmental Sciences 27: 2-5.

Masarotto, G., Varin, C. (2017). "Gaussian Copula Regression in R." Journal of Statistical Software 77(8): 1-26.

Mehmood, Q., Sial, M.H., Sharif, S., Hussain, A., Riaz, M., Shaheen, N. (2020). "Forecasting the fisheries production in Pakistan for the year 2017-2026, using Box-Jenkin's methodology." Pakistan Journal of Agricultural Research 33(1): 140-145.

Okamura, H., Punt, A.E., Semba, Y., Ichinokawa, M. (2013). "Marginal increment analysis: a new statistical approach of testing for temporal periodicity in fish age verification." Journal of Fish Biology 82(4): 1239-1249.

Okamura, H. and Semba, Y. (2009). "A novel statistical method for validating the periodicity of vertebral growth band formation in eslasmobranch fishes." Canadian Journal of Fisheries and Aquatic Sciences 66: 771-780.

Patton, A. J. (2012). "A review of copula models for economic time series." Journal of Multivariate Analysis 110: 4-18.

Phelps, Q. E., Kim, J.H., Lim, C., Odenkirk, J.S. (2019). "Marginal increment analysis of northern snakehead otoliths." American Fisheries Society Symposium 89: 1-6.

Piegorsch, W. W., Bailer, A.J. (2005). Analyzing Environmental Data. West Sussex, England, John Wiley and Sons Ltd. .

Post, J. R., Persson, L., Parkinson, E.A., van Kooten, T. (2008). "Angler numerical response across landscapes and the collapse of freshwater fisheries." Ecological Applications 18(4): 1038-1049.

Quinn, T., Deriso, R. (1999). Quantitative Fish Dynamics. Oxford, Oxford University Press.

Rautureau, N., Houakmi-Royer, Z., Perraudeau, Y. (2010). "Feasibility of derivatives for monkfish in France: A dependence analysis using the empirical copula." European Review of Agricultural Economics 37(3): 209-229.

Read, A. J., Drinker, P., Northridge, S. (2006). "Bycatch of marine mammals in U.S. and global fisheries." Conservation Biology 20(1): 163-169.

Rodriguez-Marin, E., Busawon, D., Luque, P.L., Castillo, I., Stewart, N., Krusic-Golub, K., Parejo, A., Hanke, A. (2021). "Timing of increment formation in Atlantic bluefin tuna (*Thunnus thynnus*) otoliths." Fishes 7(227): 1-14.

Saila, S. B., Wigbout, M., Lermit, R.J. (1980). "Comparison of some time series models for the analysis of fisheries data." J. Cons. int. Explor. Mer 39(1): 44-52.

Salinas-Gutierrez, R., Hernandez-Aguirre, A. Villa-Diharce, E.R. (2009). Using copulas in estimation of distribution algorithms. MICAI 2009: Advances in Artificial Intelligence. A. H. Aguirre, R. M. Borja and C. A. R. Garcia. Berlin, Heidelberg. 5845.

Sami, M., Rym, E., Fatma, E., Bachra, C., Hechmi, M. (2016). "Age and growth of bluemouth Helicolenus dactylopterus (Delarocha, 1809) in the northern waters of Tunisia (Central Mediterranean)." Greener Journal of Life Sciences.

Sas Institute, Inc. (2014). SAS/ETS 13.2 User's Guide. Cary, NC, SAS Institute Inc.

Smith, J. (2014). "Age validation of lemon sole (Microstomus kitt), using marginal increment analysis." Fisheries Research 157: 41-46.

Sun, L., Huang, X., Alqawba, M.S., Kim, J., Emura, T. (2020). Copula-based Markov Models for Time Series: Parametric Inference and Process Control. Singapore, Springer Nature Singapore Pte Ltd.

Thorson, J. T. (2019). "Predicting recruitment density dependence and intrinsic growth rate for all fishes worldwide using data-integrated life-history model." Fish and Fisheries: 237-251.

Van Buuren, S., Groothuis-Oudshorn, K., Vink, G., Schouten, R., Robitzsch, A., Rockenschaub, P., Doove, L., Jolani, S., Moreno-Betancur, M., White, I., Gaffert, P., Meinfelder, F., Gray, B., Arel-Bundock, V., Cai, M., Volker, T., Lissa, C., Oberman, H. (2022). Multivariate Imputation by Chained Equations. Version 3.15.0.

White, D., Wyanski, D., Sedberry, G. (1998). "Age, growth, and reproductive biology of the blackbelly rosefish from the Carolinas, USA." Journal of Fish Biology 53(6): 1274-1291.

Williams, A. J., Davies, C.R., Mapstone, B.D. (2005). "Variation in the periodicity and timing of increment formation in red throat emperor (Lethrinus miniatus) otoliths." Marine and Freshwater Research 56(5): 529-538.

Yang, Y., Zhang, H. (2019). "Spatial-temporal forecasting of tourism demand." Annals of Tourism Research 75: 106-119.

Yin, J., Gentine, P., Slater, L., Gu, L., Pokhrel, Y., Hanasaki, N., Guo, S., Xiong, L., Schlenker, W. (2023). "Future socio-ecosystem productivity threatened by compound drought-heatwave events." Nature Sustainability: 1-31.

Yin, J., Slater, L., Gu, L., Liao, Z., Guo, S., Gentine, P. (2022). "Global increases in lethal compound heat stress: Hydrological drought hazards under climate change." Geophysical Research Letters 49: 1-23.

Zainuddin, M., Saitoh, S. (2004). "Detection of potential fishing ground for albacore tuna using synoptic measurements of ocean color and thermal remote sensing in the northwestern North Pacific." Geophysical Research Letters 31: 1-4.

VITA

Kathleen Sue Kirch
Department of Ocean and Earth Sciences
Old Dominion University
Norfolk, VA 23529

EDUCATION

B.S. (May 2012) in Biology, Old Dominion University, Norfolk, VA 23529
M.S. (May 2015) in Ocean and Earth Sciences, Old Dominion University, Norfolk, VA 23529
M.S. (Dec 2016) in Statistics, Old Dominion University, Norfolk, VA 23529
Ph.D. (Expected May 2024) in Ocean and Earth Sciences, Old Dominion University, Norfolk, VA 23529

PROFESSIONAL EXPERIENCE

2019-Present, Statistician, Leidos, Reston, VA 20190

PUBLICATIONS

Kirch, K.S., Diawara, N., Jones, C.M. Fitting time series models to fisheries data to ascertain age. Journal of Probability and Statistics. 1-10. https://doi.org/10.1155/2023/9991872
.