Predicting Vinyl Sales in the 21st Century

Math 443 Final Project

Nicholas Kane

December 2025

1 Introduction

The vinyl record is one of the first mediums for recorded sound. Although the 2020s are the digital streaming era, there is increasing demand for vinyl records, and the data to prove it. Over the past two decades, the sales of vinyl records have been increasing and show no signs of stopping. Visually the growth seems exponential, but some time series analysis will find a model to represent this growth.

2 Objective and Data

The objective of this project is to use the RIAA vinyl sales by year to predict what the next 5-10 years will look like for vinyl. As collecting records is a passion of mine, I wanted to see if I could use what I learned in Math 443 to make informed predictions, and see if there were any surprising details I could uncover. I pulled the data from the public data for yearly sales from 1973-2023, then got the number from the 2024 RIAA end of year report. Although the data exists from 1973, my interest lies in 21st century vinyl sales.

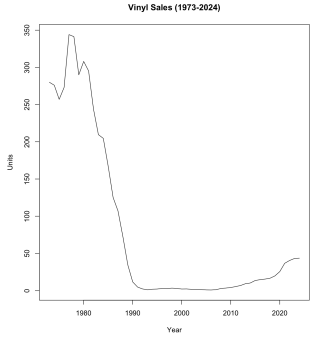

Figure 1: Vinyl Sales from 1973 to 2024

As observes in Figure 1, the sales are significantly higher in the 1970s and 1980s, but decline around the 1990s. Due to this, I have truncated the data and only included years from 2000 to 2024. Since there has been a noticeable trend, I am interested if any time series models can make predictions or suggest additional information about what the next few years look like for this medium. Figure 2 contains the years I used in my exploratory analysis and models.

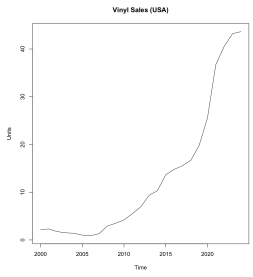

Figure 2: Vinyl Sales from 2000 to 2024

3 Exploratory Data Analysis

The first step of exploratory data analysis was to look at the raw plot of my time series and extract important information. Table 1 contains a summary of the data. It can be observed that there is a positive trend that may be either quadratic or exponential since the data has grown much faster over the past 12 years.

Minimum	Q1	Median	Mean	Q3	maximum
0.9	2.2	6.9	13	16.7	43.6

Table 1: Summary of the Time Series

Additionally, the ACF and PACF can be plotted to observe trend and seasonality, as well as identify AR, I, and MA parts. Figure 3 shows the raw ACF and PACF, and there is a slow decay on the ACF implying there will be some AR component, or that I will need to do some differencing. The PACF shows one value which leads believe an AR component of 1 or 2.

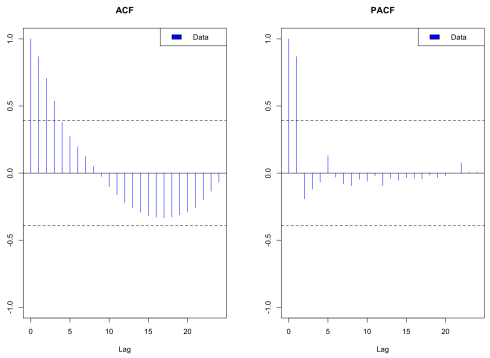

Figure 3: Vinyl ACF/PACF

I also attempted some MA smoothing to see if we could obtain a smoother curve, however since there are only 25 years of data, the higher of a value selected for q will start to eliminate too much of our data. The plots are below in Figure 4.

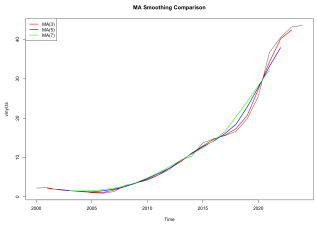

Figure 4: MA Smoothing

4 Stationary Testing

Now that a bit of exploratory analysis has been completed, I can check the time series for stationary properties. This will confirm if we need differencing or not. Performing difference once allowed a time series to be obtained. These results can be found in Figure

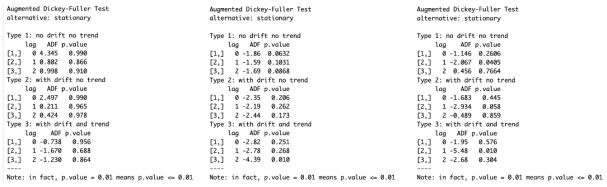

Figure 5: Dickey-Fuller Test for Vinyl Data

5 Model Selection

For this time series, it is clear there is no seasonal component, so we will ignore the SARIMA model, but still attempt exponential, ARIMA, and use Auto ARIMA to find a hypothetical best model.

5.1 Exponential Smoothing

For exponential smoothing, I tried four different models: AAN (Additive Error, Additive Trend, No Seasonality), MAN (Multiplicative Error, Additive Trend, No Seasonality), AANd (Additive error, Additive Trend, No Seasonality, Damping), and MANd (Multiplicative Error, Additive Trend, No Seasonality, Damping).

Test	**AAN**	**MAN**	**AANd**	**MANd**
Ljung–Box Q	0.943	0.7251	0.9789	0.7842
McLeod–Li Q	0.2226	1	0.968	0.9977
Turning points T	0.8696	0.1006	0.5114	0.5114
Diff signs S	0.0066*	0.4969	0.0415*	0.0415*
Rank P	0.9627	0.3502	0.7793	0.1611

Table 2: p-values for Residual Tests

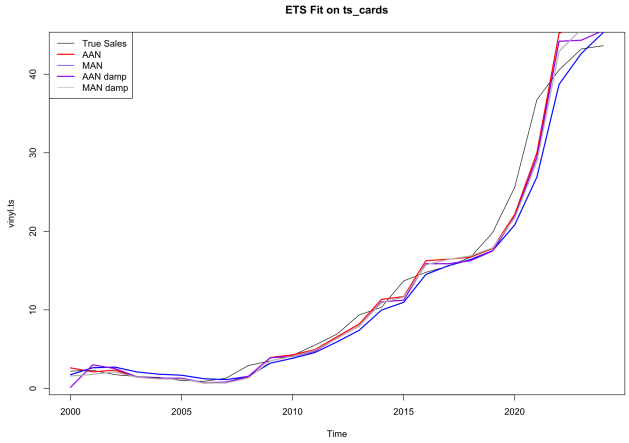

Figure 6: Exponential Models

As seen in table 2, all exponential models pass the Ljung-Box, the McLeod-Li, the Turning Points, and Rank tests. However, the MAN model is the only one which passes the diff signs test, meaning it is the best of these four models, and the exponential model I will choose it as my forecasting model.

5.2 ARIMA

For manual ARIMA models, I tried three. ARIMA(1,1,0), ARIMA(1,1,1), and ARIMA(2,1,0). These will be contrasted with all models in the next section.

5.3 Auto ARIMA

When using the auto.arima function I found that the best model was actually ARIMA(0,2,0). So there was no AR/MA component, and actually relied fully on differencing twice.

5.4 AIC Comparison

Table below shows the AIC for each model tested.

AAN	MAN	AANd	MANd	ARIMA(1,1,0)	ARIMA(1,1,1)	ARIMA(2,1,0)	ARIMA(0,2,0)
129.1	117.8	129.5	126.2	107.7	109.7	109.7	105

Table 3: AIC Comparison

It is obvious from this table that all ARIMA models outperform the exponential models. The three best models are ARIMA(1,1,0), ARIMA(2,1,0), and ARIMA(0,2,0). I will forecast future values with these three models, as well as the MAN exponential model for comparison.

6 Forecasting

6.1 Short Term Forecasting (5 years)

I will first attempt to forecast the next 5 years with the three best models and the exponential model. The results can be observed in Figure 6, and the values of the forecast for 2029 with the 95% confidence interval can be found in Table 3. As you can observe, both the ARIMA(1,1,0) and ARIMA(2,1,0) have the best predictions, and a much smaller confidence interval. The ARIMA(0,2,0) has a larger window, and the exponential model has the largest window, even displaying bounds that are impossible (negative). Although (0,2,0) model performed better with AIC, the ARI models have a smaller margin of error.

Model	Forecast	Lower	Upper
ARIMA(1,1,0)	44.45	25.04	63.87
ARIMA(2,1,0)	44.65	25.35	63.96
ARIMA(0,2,0)	45.72	12.68	78.75
ETS(M,A,N)	53.71	-26.13	133.55

Table 4: 2029 Forecasts

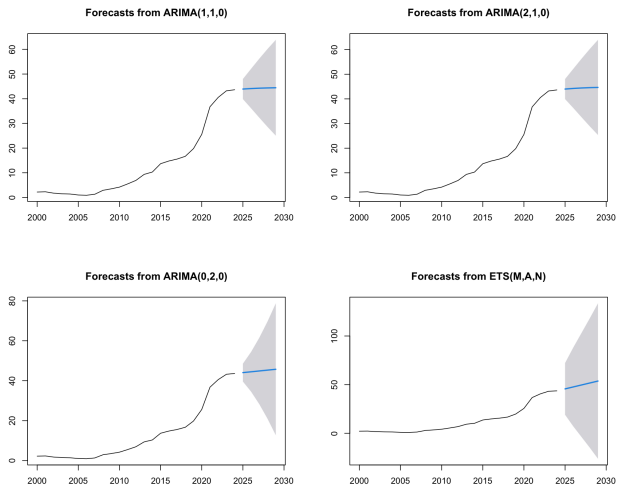

Figure 7: Short-term Forecasts

6.2 Long Term Forecasting (10 years)

Next, I will first attempt to forecast the next 10 years with the three best models and the exponential model. The results can be observed in Figure 7, and the values of the forecast for 2029 with the 95% confidence interval can be found in Table 3. As you can observe, both the ARIMA(1,1,0) and ARIMA(2,1,0) have the best predictions, and a much smaller confidence interval. The ARIMA(0,2,0) has a larger window, and the exponential model has the largest window, even displaying bounds that are impossible (negative). Although (0,2,0) model performed better with AIC, the ARI models have a smaller margin of error.

Model	Forecast	Lower	Upper
ARIMA(1,1,0)	44.58	10.20	78.98
ARIMA(2,1,0)	44.84	10.27	79.42
ARIMA(0,2,0)	47.81	-39.59	135.21
ETS(M,A,N)	63.78	-97.10	224.64

Table 5: 2034 Forecasts

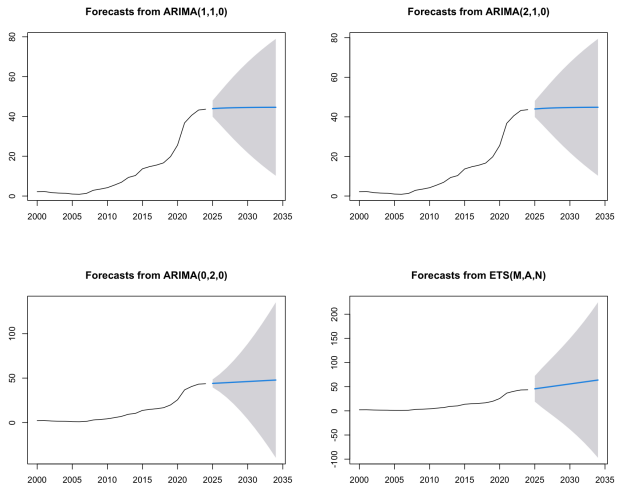

Figure 8: Long Forecasts

7 Conclusion

Overall, found that ARIMA was the best way to model the time series, and if used for prediction, a model of (1,1,0) or (2,1,0) would be the best. However, if one wants to optimize AIC, then (0,2,0) should be used. As the models forecasted, vinyl records sales will continue on an upward trend, though it may slow down. This was an interesting project. However, the results I got were not too satisfying, and the data was yearly, and only over a small range of 25 years, so I did not get to work with the seasonal modeling techniques. If I were to expand upon this project, I would look at more granular data, maybe monthly sales, to determine if there is a seasonality in addition to the trend.

8 Appendix: Code

```
library(itsmr)
library(tseries)
library(aTSA)
library(forecast)

vinyl <- read.csv("Vinyl_Units_USA_1973_2024.csv", header = TRUE)
plot(vinyl, type='l', main = "Vinyl Sales (1973-2024)",xlab="Year", ylab="Units")

vinyl.mod <- vinyl[28:52,]
vinyl.ts <- ts(vinyl.mod[,2], start=2000, frequency=1)
plot(vinyl.ts, main = "Vinyl Sales (USA 2000-2024)", ylab = "Units")
summary(vinyl.ts)
plot(vinyl.ts)
plota(vinyl.ts)
adf.test(vinyl.ts)

plot(log(vinyl.ts))
plota(log(vinyl.ts))

y2 <- ts(trend(vinyl.ts, 2), start=2000, frequency=1)

plot(vinyl.ts)
lines(y2)

y3 <- ts(trend(vinyl.ts, 3), start=2000, frequency=1)

plot(vinyl.ts)
lines(y3)

ma.fit3 <- ma(vinyl.ts, order = 3)
ma.fit5 <- ma(vinyl.ts, order = 5)
ma.fit7 <- ma(vinyl.ts, order = 7)
plot(vinyl.ts, main="MA Smoothing Comparison")
lines(ma.fit3, col="red", lwd=2)
lines(ma.fit5, col="blue", lwd=2)
lines(ma.fit7, col="green", lwd=2)
legend("topleft", c("MA(3)", "MA(5)", "MA(7)"),
       col=c("red","blue","green"), lwd=2)

par(mfrow=c(1,1))
ets.fit <- ets(vinyl.ts)
summary(ets.fit) # not good, no trend


plot(vinyl.ts, main = "ETS Fit on ts_cards")
lines(fitted(ets.fit), col="red", lwd=2)
legend("topleft", c("Original", "ETS fit"),
       col=c("black","red"), lwd=c(1,2))

ets.add <- ets(vinyl.ts, model="AAN")
ets.mult <- ets(vinyl.ts, model="MAN")
ets.dampa <- ets(vinyl.ts, model="AAN", damped = TRUE)
ets.dampm <- ets(vinyl.ts, model="MAN", damped = TRUE)

test(residuals(ets.add))
#adf.test(residuals(ets.add))
test(residuals(ets.mult))
#adf.test(residuals(ets.mult))
test(residuals(ets.dampa))
```

```
#adf.test(residuals(ets.dampa))
test(residuals(ets.dampm))
#adf.test(residuals(ets.dampm))


ets.aic <- c(AIC(ets.add),AIC(ets.mult),AIC(ets.dampa),AIC(ets.dampm))
names(ets.aic) = c("AAN","MAN","AAN(D)", "MAN(D)")
ets.aic

plot(vinyl.ts, main = "ETS Fit on vinyl")
lines(fitted(ets.add), col="red", lwd=2)
lines(fitted(ets.mult), col="blue", lwd=2)
lines(fitted(ets.dampa), col="purple", lwd=2)
lines(fitted(ets.dampm), col="gray", lwd=2)
legend("topleft", c("True Sales", "AAN", "MAN", "AAN damp", "MAN damp"),
       col=c("black","red","blue","purple","gray"), lwd=c(1,2))

#MAN without damping clearly the best
dif1 <- diff(vinyl.ts,1)
plota(dif1)
adf.test(dif1)
dif2 <- diff(vinyl.ts, 2)
plota(dif2)
adf.test(dif2)

#ARIMA
arima110 <- arima(vinyl.ts, order=c(1,1,0))
test(residuals(arima110))
adf.test(residuals(arima110))
arima111 <- arima(vinyl.ts, order=c(1,1,1))
test(residuals(arima111))
adf.test(residuals(arima111))
arima210 <- arima(vinyl.ts, order=c(2,1,0))
test(residuals(arima210))
adf.test(residuals(arima210))

#AutoArima
set.seed(1)
best.arima <- auto.arima(vinyl.ts)
best.arima
test(residuals(best.arima))
adf.test(residuals(best.arima))

arima.aic <- c(AIC(arima110), AIC(arima111), AIC(arima210), AIC(best.arima))
names(arima.aic) <- c("ARIMA 110", "ARIMA 111", "ARIMA 210", "ARIMA 020")

#AIC Comparison
final.aic <- c(ets.aic, arima.aic)
final.aic #beat are 110, 210, 020

#Forecasting
par(mfrow=c(2,2))
forecast.1.s <- forecast::forecast(arima110, h=5, level=95)
forecast.2.s <- forecast::forecast(arima210, h=5, level=95)
forecast.3.s <- forecast::forecast(best.arima, h=5, level=95)
forecast.4.s <- forecast::forecast(ets.mult, h=5, level=95)
plot(forecast.1.s)
plot(forecast.2.s)
plot(forecast.3.s)
plot(forecast.4.s)
```

```
par(mfrow=c(2,2))
forecast.1.l <- forecast::forecast(arima110, h=10, level=95)
forecast.2.l <- forecast::forecast(arima210, h=10, level=95)
forecast.3.l <- forecast::forecast(best.arima, h=10, level=95)
forecast.4.l <- forecast::forecast(ets.mult, h=10, level=95)
plot(forecast.1.l)
plot(forecast.2.l)
plot(forecast.3.l)
plot(forecast.4.l)
```