8-2011

Comparison of Time Series and Functional Data Analysis for the Study of Seasonality.

Jake Allen
East Tennessee State University

Comparison of Time Series and Functional Data Analysis for the Study of Seasonality

---

A thesis

presented to

the faculty of the Department of Mathematics

East Tennessee State University

In partial fulfillment

of the requirements for the degree

Master of Science in Mathematical Sciences

---

by

Jake Allen

August 2011

Edith Seier, Ph.D., Chair

Robert Price, Ph.D.

Robert Gardner, Ph.D.

Keywords: phase-plane plots, X-11, French Broad River, moving averages

ABSTRACT

Comparison of Time Series and Functional Data Analysis for the Study of Seasonality

by

Jake Allen

Classical time series analysis has well known methods for the study of seasonality. A more recent method of functional data analysis has proposed phase-plane plots for the representation of each year of a time series. However, the study of seasonality within functional data analysis has not been explored extensively. Time series analysis is first introduced, followed by phase-plane plot analysis, and then compared by looking at the insight that both methods offer particularly with respect to the seasonal behavior of a variable. Also, the possible combination of both approaches is explored, specifically with the analysis of the phase-plane plots. The methods are applied to data observations measuring water flow in cubic feet per second collected monthly in Newport, TN from the French Broad River. Simulated data corresponding to typical time series cases are then used for comparison and further exploration.

ACKNOWLEDGMENTS

First, I would like to thank Dr. Edith Seier for her guidance throughout this process. Her extreme patience was necessary and her enthusiasm was motivational. Thanks also to Dr. Bob Price and Dr. Bob Garnder for accepting the challenge to be on my committee. The ETSU family has been very courteous and encouraging, and I'm thankful for all of the new faces that I have encountered over the past two years.

Finally, thanks to my family and friends for constantly reminding me to stay on the path. Without the encouragement, and sometimes aggravation, this accomplishment may not have been possible. With the path continuing, it's comforting to know who will be there as I seek out God's will.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

A common question one may ask while conducting statistical analysis is: how do the data change over time? This exact question has peaked the interests of many statisticians and other inquiring minds while exploring potential research topics. However, the topic intended for research needed to involve some environmental characteristic in order to integrate the researcher's love for the outdoors. Thankfully, time series analysis is often used to analyze environmental topics such as climate, hydrology, and ozone level, just to name a few.

However, since time series plots often resemble combined curves of acceleration and deceleration in either a positive or negative manner, one can conclude that some of the variation among curves could be explained at some level by derivatives [9]. Now, if one becomes interested in derivatives, this gives good evidence to work with the data as a function rather than simply vectors of measurements over time [9]. But, what is the approach to visually analyze these derivatives? Functional data analysis provides an extremely useful plot of velocity versus acceleration: phase plane plots.

Considering this new analytical approach raises a variety of questions, particularly: what type of information do phase plane plots offer that traditional time series analysis does not offer? Answering this, and questions about integrating both time series and functional data analysis, is the primary goal of this work. Do the two methods complement each other, and can they be used simultaneously?

Background information is the first essential component of this study. The French Broad River was chosen as an environmental resource to provide certain seasonal data for analysis. This river holds great value to people across western North Carolina

and east Tennessee and its geographical implications, regional impacts, and physical layout are all informatively mentioned.

Now, the data itself are expressed as a time series to prepare for analysis. From given plots, a seasonal component is obviously present and thus extracted for study. Also, an ARIMA model is discussed and a SARIMA model is presented and its parameters are stated, only to be followed by a decomposition approach to this analysis. Using SAS, the X-11 method is then described and implemented as the considerable choice for the decomposition.

The next step is a basic introduction into the inner workings of functional data analysis–particularly with phase-plane plotting. Functional data analysis is a mathematical approach for which some statistical questions have not been addressed yet. In this work, a typical statistical question such as sensitivity with respect to noise is addressed with regard to the phase plane plots. Statistical software R is used and thus explanations are given for command arguments. These plots are an important graphical representation for this research and are integral to its conclusions.

Since functional data and time series analysis provide the bulk of statistical work, it is only appropriate to combine the two and look at phase-plane plots of components extracted from a time series approach. Comparatively, what conclusions can be drawn and are there any implications from the findings? In the detection of seasonality, what does each method offer analytically? Answering questions like these will conclude this research.

2 THE CASE STUDY

2.1 The French Broad River

Known by the Cherokee Indians as "Agiqua", translated "Long Man", the French Broad River garners respect and thanksgiving from the inhabitants that line its flowing 213 miles [7]. It provides a variety of resources to mostly rural communities as it winds through a heavily forested yet breathtaking landscape. Many hours have been spent treading the river's banks or canoeing its rapids in search of the plentiful resources that it contains. No doubt that the French Broad River also contains valuable information to modern investigators as well.

Running into then French Territory, the name "French Broad" caught on quickly to this wide river with substantial streams and watersheds [7]. It begins its journey in Transylvania County, NC and winds through eight counties in North Carolina and Tennessee, finally being dammed just before reaching Knoxville, TN by the Douglas Dam [10]. Because of the regions' diverse environmental impacts, the flow of the French Broad River varies from year to year and month to month, providing an essential data set to this work.

2.2 Exploratory Analysis

One of the most useful tools for statisticians is being allowed to view the behavior of a variable over time. Time series applications are often found conveniently within nature–one contributing factor to the choosing of this topic. The analysis begins with the simplest of tasks: plotting.

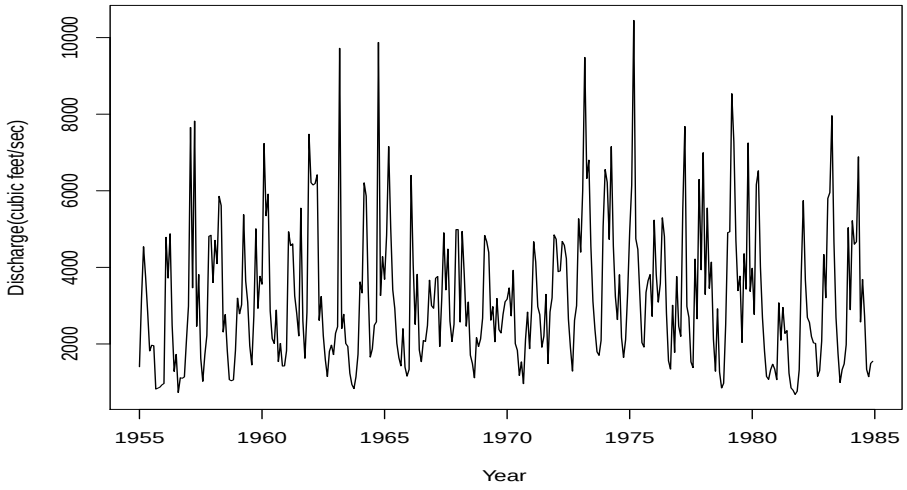

Figure 1: Time Series of French Broad River Flow

A time series plot such as that in Figure 1 can reveal significant information quickly and efficiently. The United States Geological Survey collects data about water flow, water quality, and many other factors from rivers all across the nation. Data from the French Broad River at a site in Newport, TN has been collected since 1920. The data being used here is from January 1955 to December 1984 in monthly averages of water discharge measured in cubic feet per second. These years were chosen because within them there was no missing data and these years are fairly recent. Missing data would need to be replaced based on averages, a task worth avoiding if possible. The most recent years had missing data from a few months and would also not allow for such a large data set. By using 30 years worth of information, a greater diversity of information is collected because of historical variation.

2.3 Seasonal Behavior

The next big task in analyzing this series is looking at variation within each year. A time series is said to have seasonal behavior if there is a pattern that lasts one year and repeats, with some variations of course, year after year [6]. Seasonality stems from changes in perhaps temperature, precipitation, etcetera [6]. Because our series exhibits annual periodicity, we can further analyze it's seasonal component to fully define a specific model. By controlling for seasonality, it can be determined what effects are confounded within this series' seasonal attributes [6].

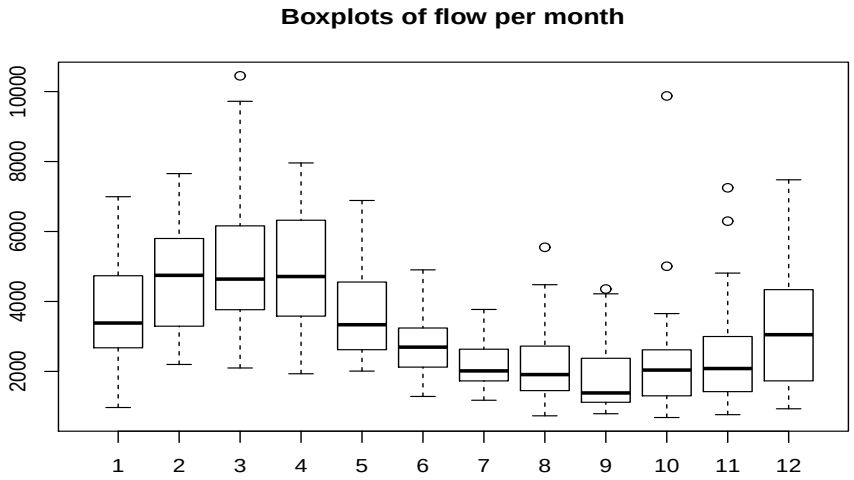

Figure 2: Seasonal pattern

Graphical representation is particularly helpful in spotting seasonality–either with the time series plot itself, boxplots per month, or a correlogram [6]. Figure 2 displays twelve boxplots, one for each one of the twelve months. The medians of the boxplots reveal the seasonal pattern of the river flow. A correlogram is simply the plot of

autocorrelations against lag [5]. Autocorrelation of lag k is the correlation between observations that are k units of time apart. The autocorrelation function can be summarized as a way to identify explanatory relationships within a single time series, and is seen in figure 3 [5]. For instance, the river flow time series is composed of 360 observations from January 1955 to December 1984. Let's call observations $Y_1, Y_2, ..., Y_{360}$ the observed points at time periods 1, 2, ..., 360 respectively. Lagging the series by one period creates 355 pairs of overlapping observations to compare in order to calculate the autocorrelation of order 1 [5]. Now, we can compute the correlation as if it were the comparison between two sets of quantitative data; but coming from one time series, these statistics are known as autocorrelation [5].

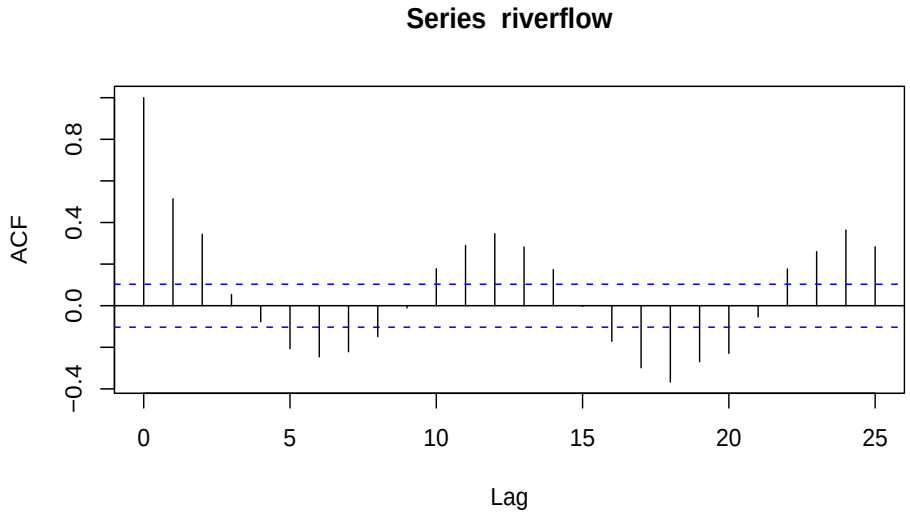

Figure 3: Autocorrelation Function

Together this plot calculates autocorrelations at lags 1, 2, ..., to make up the autocorrelation function (or ACF) [5]. Notice that the autocorrelation at lag 12 is

much higher than that of others. This is a result of the seasonal pattern in the data, showing that high (and thus low) levels of water flow in the French Broad River tend to be 12 months apart. In other words, the observations that are 12 months apart are correlated because they were done in similar times of different years. Figure 2 shows how the river flow changes from January to December, with the wave-like appearance indicating annual seasonal changes. Notice also the substantial variability among years for the same month, something to be explored in functional data analysis as well.

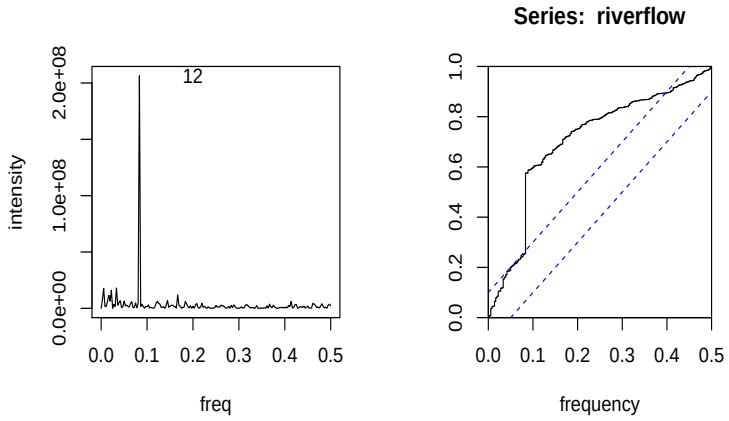

Figure 4: Periodogram

Another useful display that extracts the possibility of a seasonal component is a periodogram (Figure 4). The periodogram is a tool defined by Schuster in 1898 to look for hidden periodicities [12]. More recent references to the periodogram include Cryer and Chan [1]. The periodogram considers the angular frequencies in the interval $[0, \pi]$, or equivalently the frequencies in the interval $[0, \pi]$. The intensity of each angular

frequency w is defined as $I(w)$

$$I(w) = [a(w)]^2 + [b(w)]^2 \tag{1}$$

where $a(w) = \frac{2}{N}\sum y_t^* cos(tw), \quad b(w) = \frac{2}{N}\sum y_t^* sin(tw)$ are basically the covariances of the time series (y_t^*), already adjusted for the mean, with the frequencies. The periodogram is the plot of $I(w)$ versus w, where w takes values from 0 to π. The left side of Figure 4 displays the periodogram of the river flow using the frequencies f in the interval [0,0.5] instead of the angular frequencies. The highest intensity corresponds to the frequency $0.08 = 1/12$ associated with seasonal behavior. Figure 4 also includes the cumulative periodogram on the right that indicates approximately 35% of the variability in the river flow is associated with seasonality.

3 CLASSICAL TIME SERIES APPROACH

3.1 The Additive and Multiplicative Models

To fully understand time series analysis, it is crucial to appreciate the workings of its component parts. Early studies of time series were often done by economists interested in business cycle changes related to calendar effects, trends, cycles, seasonal, and irregular components [6]. One way of compiling these components is to add them. In our case, however, there is no need to analyze calendar effects components and so the decomposition can be represented by

$$Y_{a_t} = S_t + T_t + I_t$$

where Y_{a_t} = additive time series, T_t = trend, S_t = seasonality, and I_t = irregularity [6].

The additive model is more appropriate if seasonal fluctuations do not change throughout the series [5]. If these fluctuations occur proportionally with increases or decreases in the level of the series, then the multiplicative model would be more appropriate [5]. Occasionally, one could use a transformation of the data rather than having to choose between an additive or multiplicative model. Specifically, taking logarithms turns a multiplicative relationship into an additive relationship, because

$$Y_{a_t} = S_t \times T_t \times I_t$$

implies

$$\log Y_{a_t} = \log S_t + \log T_t + \log I_t.$$

3.2 Decomposing the Time Series

Due to the importance of seasonality in the river flow, it is useful to apply decomposition methods that allow observations of the trend and seasonal components of the time series separately. The X-11 process can be summarized in five stages: (1) trading day adjustment, (2) trend cycle estimation, (3) preparation of seasonal adjustment factors (4) treatment of extremes, and (5) create component tables and summary statistics [6]. With the particular time series interest only in seasonal and trend components, the trading day adjustment is of no concern since the nature of a business cycle is irrelevant. Basically, the method iterates through the trend, seasonal, and irregular components and smooths the data at each iteration, estimating the trend component and dividing the data by the trend in order to estimate the seasonal and irregular components [6]. More simply, the following equation is used as explanation:

$$\hat{S}_t \hat{I}_t = \frac{Y_t}{\hat{C}_t}. \tag{2}$$

Note that the hats over the terms indicate estimates.

Figure 5 is the resulting plot of the original time series and seasonal, trend, and irregular components extracted from the X-11 process. The seasonal plot shows obvious annual change, however also displaying significant variability among years. Environmental impacts vary from year to year, and this is easily seen in this plot. The seasonal component also indicates that the difference between low and high months was highest in the late 1950s and early 1960s. The seasonal component seems to diminish in the central observations, but eventually picks up pace in the 1980s. If a longer data set was used, perhaps the recorded seasonality in the 1990s would

eventually reach that of the 1950s. The trend cycle analysis shows that some trend cycles last longer than one year. For instance, there appears to be a upward cycle of at least two or three years in the early 1970s, perhaps because of drought in the late 1960s requiring the river to recover from low water flow [4]. And the irregular component simply shows that there is variability or noise from year to year and month to month. Notice that the irregular component picks up the highest recorded crest, occurring in 1977 [4].

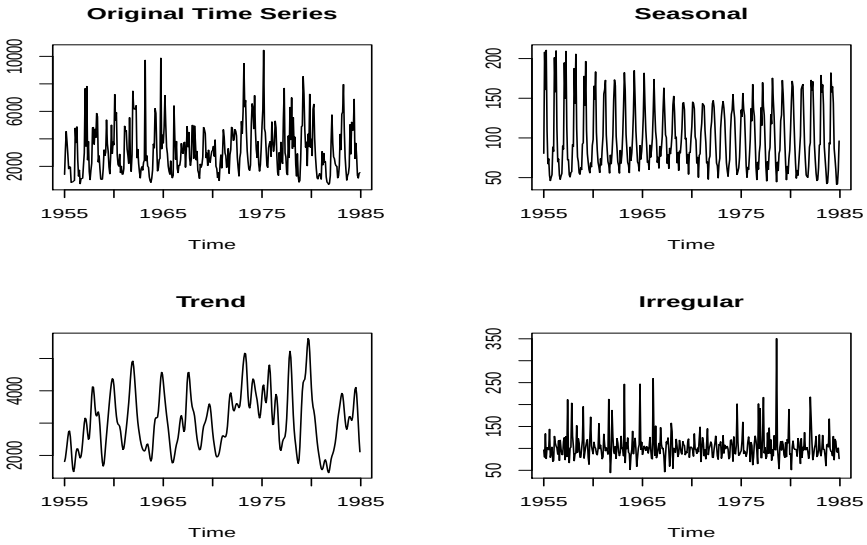

Figure 5: Components estimated with the X-11 method

3.3 A SARIMA Model

ARIMA (autoregressive integrated moving averages) and SARIMA models were developed by G.E.P. Box and G. Jenkins in 1968 [1]. The basic idea behind Box & Jenkins' models is that the current value of a time series can be written in terms of the

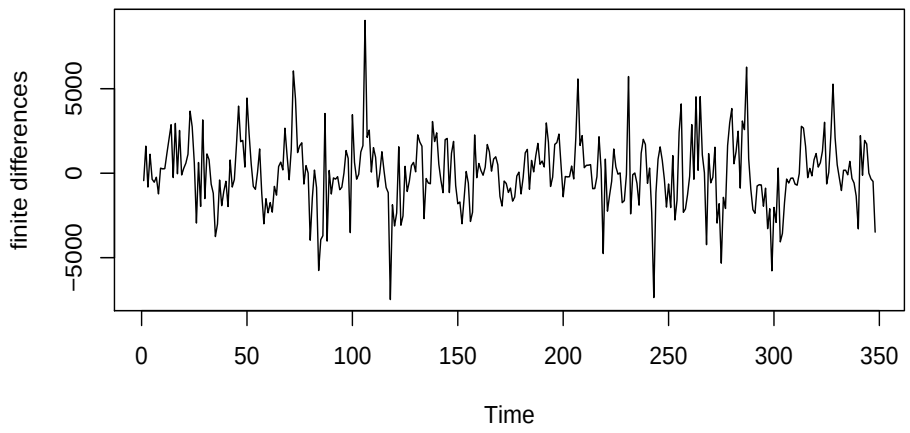

Figure 6: Finite differences of order 12

previous values. In the case of the French Broad river, due to the presence of seasonality, a SARIMA model is needed. The identification of the specific SARIMA model is done by looking at the pattern described by the serial autocorrelation function [1]. Figure 3 (autocorrelation function) indicates that it is necessary to calculate finite differences of order 12 because the autocorrelations of order multiple of 12 do not decline quickly enough. Figure 6 displays the finite differences of order 12 of the river flow time series. Figure 7 contains the autocorrelations and partial autocorrelations of the finite differences. The patterns observed in Figure 7 indicate that a SARIMA model $(2,0,0)(0,1,1)_{12}$ could be a candidate to represent the behavior of the river flow under the Box & Jenkins approach. The evaluation of the model indicates that it is an acceptable model and all the parameters in the model are necessary. The p-values corresponding to the null hypotheses that the parameters can be made equal

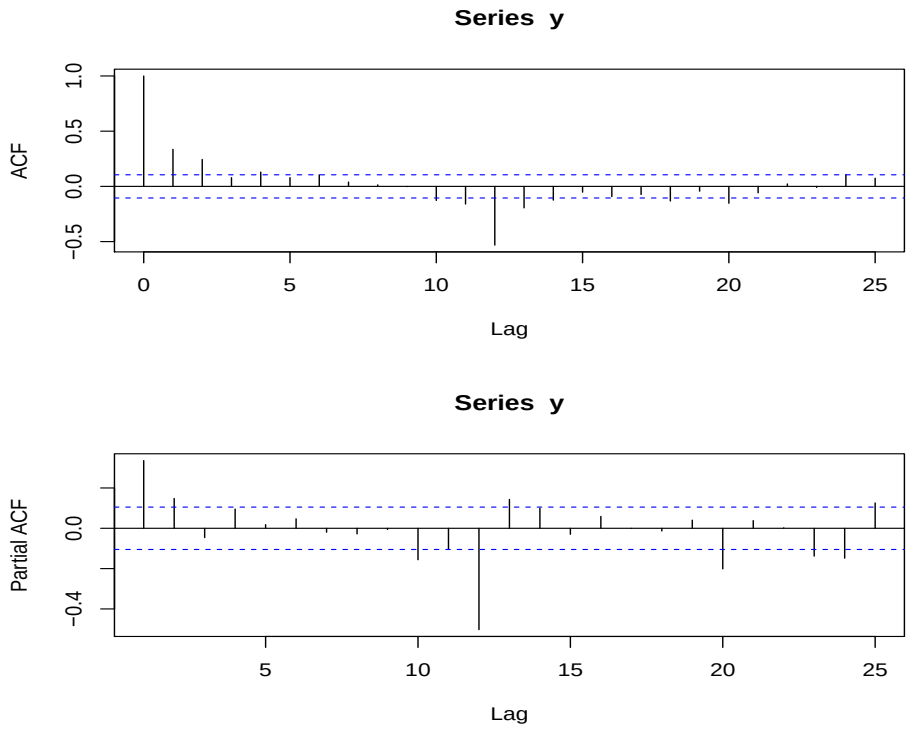

Figure 7: Autocorrelation and partial autocorrelation of finite differences

to 0 are small and those null hypotheses are rejected (Table I). The p-values of the Box-Pierce statistics are small thus the null hypothesis that the residuals come from a white noise process is not rejected (Table I).

Table 1: Minitab Output for ARIMA Modeling

(a) Final Estimates of Parameters

Type	Coef	SE Coef	T	P
AR 1	0.2613	0.0527	4.96	0.000
AR 2	0.2048	0.0528	3.88	0.000
SMA 12	0.9428	0.0253	37.22	0.000

```
Differencing: 0 regular, 1 seasonal of order 12

Number of observations:  Original series 360, after differencing 348

Residuals:    SS =  640783175 (backforecasts excluded)

              MS =  1857343  DF = 345
```

(b) Modified Box-Pierce Chi-Square Statistic

Lag	12	24	36	48
Chi-Square	8.6	28.5	42.7	55.8
DF	9	21	33	45
P-Value	0.477	0.127	0.120	0.129

Using coefficients from the MINITAB output, the SARIMA model for the river flow is written as:

$$(1 - 0.2613B - 0.2048B^2)(1 - B^{12})y_t = (1 - 0.9428B^{12})a_t. \quad (3)$$

Doing the algebra, the following expression is obtained for the current value of the river flow:

$$y_t = 0.2613y_{t-1} + 0.2048y_{t-2} + y_{t-12} - 0.2613y_{t-13} + 0.2048y_{t-14} + a_t - 0.9428a_{t-12}. \quad (4)$$

Notice that the current value of the river flow can be expressed in terms of the river flow that was observed months before plus certain purely random components.

4 FUNCTIONAL DATA ANALYSIS

4.1 Defining Functional Data Analysis

The concept of functions is something learned during the early algebra years of mathematical education. However, the simple idea of plotting a function based on some input/output rule should be further expanded for this case. Values that reflect smooth variation can be expressed as functions, such as with the time series case. In other words, consider the observation for each individual to be values through time or space from an underlying stochastic process. Considering the case with river flow, there is an uncertainty or noise in the measurement of cubic feet per second, and although the measurements are discrete values, they reflect a smooth variation in that measurement and could be assessed as a function of river flow [9].

Additionally, referencing Figure 1, it is easy to see strong evidence of acceleration in the river's flow followed by sharp deceleration. The conclusion can thus be made that some of the variation between years can be explained at a level of derivatives [9]. Any time that derivatives play a role in analysis, it gives the researcher evidence to think of the variable as a function rather than vectors of data in discrete time [9].

As in this case, the data work as a single long record measured from 1955 to 1984. As such, these data can show variation at several levels [9]. Many times there is a tendency for a record to show exponential (or perhaps geometric) increase or decrease over time [9]. This is particularly true for economic time series, which are known to transition well into functional data analysis. However, in the case of the French Broad River flow, a finer scale can be analyzed to notice departures from this

trend attributed to years of drought, flooding, etcetera. Even more specifically, and beneficial to the analysis of this research, a marked annual variation raises questions of whether or not a seasonal trend shows some longer term changes [9].

Since it has been shown that the given data comes through a process that is naturally described as functional, the next analytic question becomes: what are the goals of analyzing functional data? Ramsay [9] summarizes them as follows:

- to represent the data in ways that aid further analysis
- to display the data so as to highlight various characteristics
- to study important sources of pattern and variation among the data
- to explain variation in an outcome or dependent variable by using input or independent variable information
- to compare two or more sets of data with respect to certain types of variation, where two sets of data can contain different sets of replicates of the same functions, or different functions for a common set of replicates.

4.2 Phase-Plane Plots

Much of the concentration for this research will be on the goal of displaying the data so as to highlight various characteristics. One useful way to accomplish this in functional data analysis is with phase-plane plotting [9]. Simply put, a phase-plane plot is a depiction of acceleration versus velocity as a reflection of energy transfer, with energy being the effort or work required to show change within a system over

time [9]. Where the concepts of energy and functional data vary on more than one time scale lead to this graphical technique [9]. The specified river flow data exhibits variation on different time scales:

- The longest scale is the thirty year progression of the river's water flow
- There are events that last a few years because of extended drought or flooding or some other lengthy environmental factor
- The shortest scale shows seasonal variation over an annual cycle that typically repeats

Keep in mind how the sine function changes over a one year period in Figure 8 before considering the phase-plane plot in Figure 9. Notice the similarities between the medians of the boxplots in Figure 2 that suggested a seasonal pattern similar to

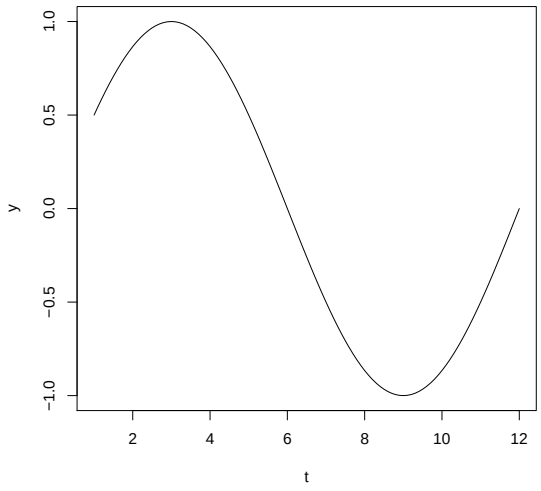

Figure 8: Sine Function Wave for one Year

what is seen here. This "smoothness" will be of particular interest when choosing a smoothing parameter for later phase-plane plots.

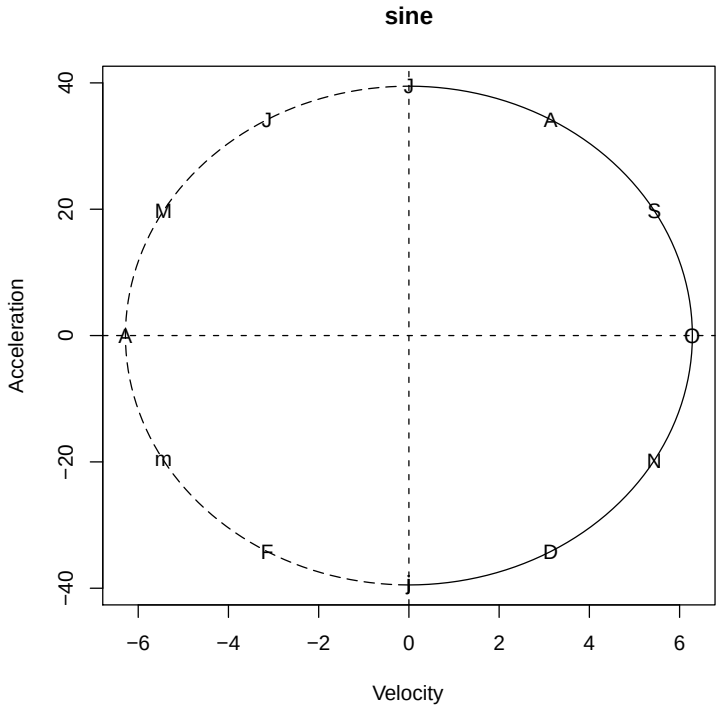

Figure 9: Sine Function Phase-Plane Plot

J. O. Ramsay depicts a very helpful image to describe how phase-plane plots show energy transfer. Since the tool is a plot of acceleration against velocity, consider the phase-plane plot of the function $\sin(2\pi t)$ in Figure 9 [9]. Note that since this study compares time series, the months are labeled on the figure starting with January (lowercase j) at the bottom and rotating clockwise. Ramsay describes this simple function as a basic *harmonic* process that can be compared to the vertical position of the end of a suspended spring [9]. The spring bounces with a period of one time unit and starts at time $t = 0$, where the spring oscillates because there is an exchange

between potential and kinetic energy [9]. When the spring is momentarily still at either end of its trajectory its potential energy is maximized [9]. Vice versa, when the spring passes through the position 0, the velocity is greatest but acceleration is zero [9]. So it can be shown that potential energy is associated with acceleration and kinetic energy with velocity [9]. For the purposes of this study, the involvement of kinetic versus potential energy will focus the attention on the dynamics of the seasonal component.

4.3 Obtaining the Phase-Plane Plots

The first step in obtaining these useful plots is to define a set of functional building blocks called basis functions [3]. They can be created with mathematical softwares R and MATLAB. R is chosen to do this analysis. The spline basis system (and bsplines in particular) is what will be used for the creation of these phase-plane plots. Start by using a set of these functional building blocks $\phi_k, k = 1, ..., K$ called *basis functions*, combined linearly [3]. Now, each basis system requires a specific set of K basis functions ϕ_k's. Now, creating a basis object *flowbasis* by recalling the `create` function in R yields:

```
flowbasis <- create.bspline.basis(rangeval, nbasis, norder, breaks)
```

Now, in each create function there is a set of arguments that must be clarified: rangeval, nbasis, norder, and breaks. `Rangeval` specifies the lower and upper limits of the values of the argument t and is a vector object of length 2 [3]. The values of interest are the years in which data was collected, so `rangeval=c(1955,1985)`.

Next, `nbasis` is an integer specifying the number of K basis functions and `norder` is an integer specifying the order of b-splines. So how are these values found?

Well, if the interval of observation is broken into subintervals, with boundary points called *breaks*, splines can be more easily constructed [3]. And over these subintervals, the spline function can be described as a polynomial of fixed degree (or order), where the order of a polynomial is one higher than its degree or highest power within the polynomial [3]. More precisely, a spline basis is defined in terms of a set of *knots* [3]. Note that every knot has the same value as a break point, but multiple knots can occur at certain breakpoints [3]. However, for this study, only one knot is placed at a break point, forcing the number of derivatives to be two less than its order, ensuring that the splines will be seen as smooth [3].

Now, the number K of basis functions in a spline basis system is given by the relation *number of basis functions = order + number of interior knots* where interior knots are those at break points not located at the beginning or end of the domain defined by the function [3]. For this study, `norder` is chosen to be 8 because of the smoothness it presents, and thus will consist of seventh degree polynomial segments. Also, the number of interior knots will be like the majority of applications with only a single knot at every breakpoint. There are 360 observations of data, simplifying the equation to $366 = 8 + 358$.

The basis object is then created:

```
flowbasis <- create.bspline.basis(rangeval=c(1955,1985), nbasis=366, norder=8)
```

The breaks command can be inferred from R since nbasis = nbreaks + norder - 2, where nbreaks = length(breaks)[8].

Next, a function is created to turn an integer specifying an order of a derivative into the equivalent linear differential operator object, call it m [8]. Use $m = 4$ when the goal is to study velocity and acceleration, yielding LfdobjWater = int2Lfd(4).

Finally, the last steps create a smoothing parameter to be examined:

```
WaterSm <- smooth.basisPar(argvals=index(riverflow),
y=(coredata(riverflow)), fdobj=flowbasis,
Lfdobj=LfdobjWater, lambda=1e-11)
```

Recalling the function `smooth.basis` is done with arguments for argvals, y, fdobj, Lfdobj, and the smoothing parameter lambda. `argvals` is simply the index of the data itself and y is an array with values of curves at sampling points or argument values [3]. `coredata` is used in this array to stip off the index/time attributes and recall only the observations [8]. The basis object flowbasis will now be used to define a functional data object–so a basis was first created and then recited in the form of functional data. The linear differential operator object `Lfdobj` was created earlier as `LfdobjWater` and is now replaced.

Finally, the smoothing parameter λ can be chosen. One way to accomplish this is to minimize the *generalized cross-validation* measure GCV developed by Craven and Wahba (1979) [3]. The criterion is

$$GCV(\lambda) = \left(\frac{n}{n - df(\lambda)}\right)\left(\frac{SSE}{n - df(\lambda)}\right). \tag{5}$$

The `lambda2gcv` function in R can be used to return the minimizing value for certain values of lambda [3]. Instead of listing these values, a plot can be more helpful in finding the proper value of lambda.

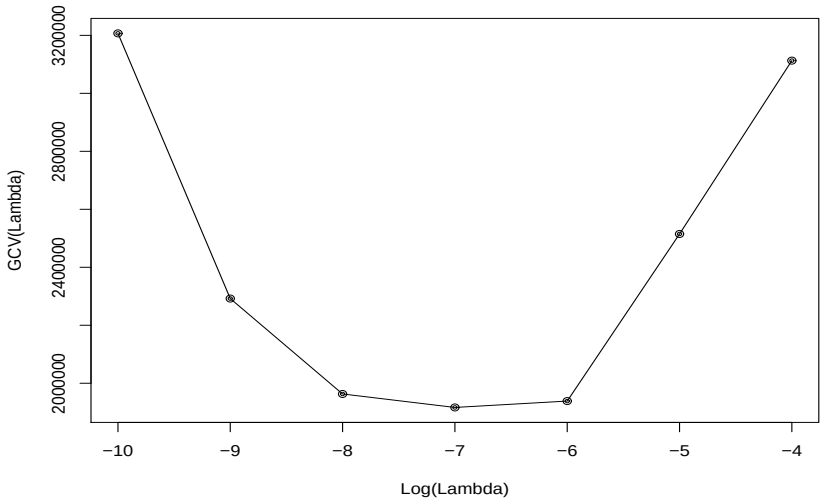

Figure 10: GCV to find the best λ

From Figure 10 it can be seen that the minimizing value of λ is $1e-7$ or 0.0000001. However, Ramsay recommends caution when choosing the smoothing parameter, and encourages an enlightenend approach to the decision rather than choosing λ based solely on automatic methods like GCV minimization [3]. A look at phase-plane plots with the GCV selected smoothing parameter will show why this is the case.

Notice from Figure 11 how, with this value of λ, that the phase plane plots show very little information or exchange in energy. Obviously this is not a true depiction, since it is known that natural water flow exhibits consistent change from a seasonal standpoint. Perhaps this could be explained because this particular value of λ smooths the series too much. Figure 12 shows that fits estimates to the series reveals just how extreme the jumps are from one month to the next, and how over-smoothing is a potential hazard.

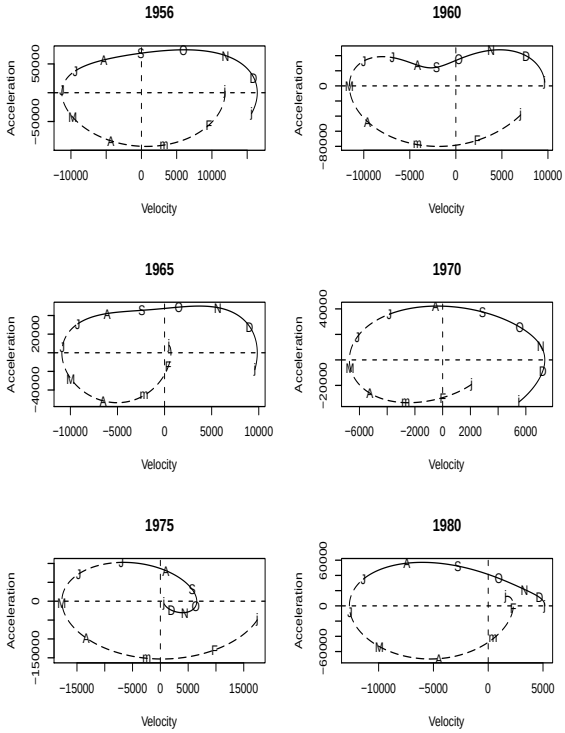

Figure 11: $\lambda = 1e - 7$

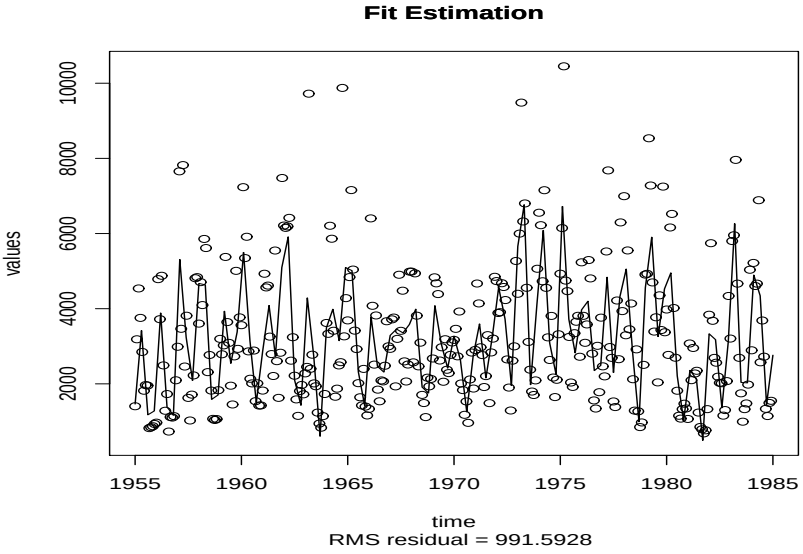

Figure 12: Fitted Values

If these jumps are smoothed over, then the phase plane plots will have no room to show the dynamics of the series. Therefore, Ramsay's cautions are confirmed from the phase-plane plots and a more reliable value for λ should be chosen. After much experimentation, the value that reveals the most information and the best smoothing procedure within the phase-plane plots is $1e-11$ or 0.00000000001. Values smaller than this that approach zero will fit nearly exact values from the data, causing the plots to be under smoothed and over complicated.

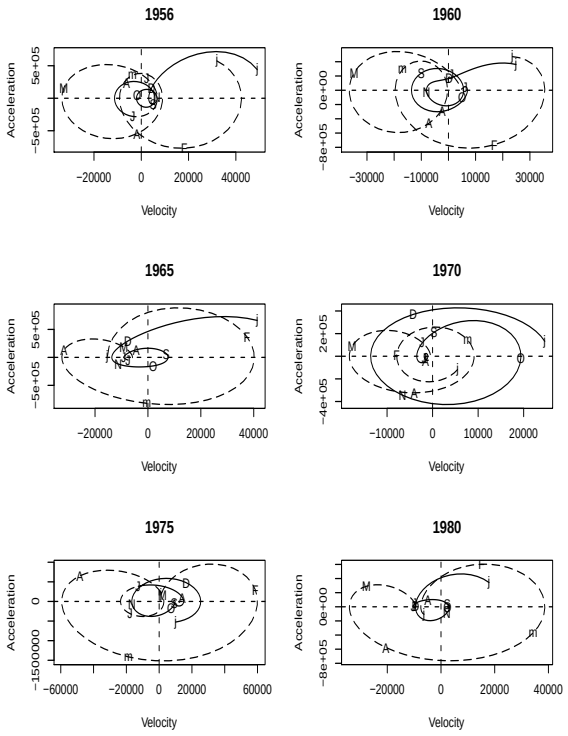

Figure 13: $\lambda = 1e-11$

Now in Figure 13 with $\lambda = 1e-11$ the amount of smoothing seems to be appropriate for allowing further analysis. In terms of detecting seasonality, first notice how each year contains cycles during particular seasons. Additionally, notice that cycles

tend to occur in similar locations and with similar intensity, a pattern indicative of seasonality. So noticing seasonality is fairly easy, but each individual year will not be analyzed here. Rather, these plots will be used as reference or comparison with the phase-plane plots obtained later.

So the question remains: do phase-plane plots offer something that traditional time series plots do not offer? While many plots have been presented, a look at a single year may be a good approach to fully answering this question.

The French Broad River had significant crests in 1977 in North Carolina and eventually flowed into the measurement area in Newport [4]. However, this individual year would be difficult to observe as a flood year from the time series plot. The time series plot easily outlines severe crests and troughs, but causes difficulty spotting entire years displaying uncommon river flow behavior. On the other hand, Figure 14 indicates more intensity on the velocity and acceleration scales, a good sign of large amounts of water flow. Additionally, the phase-plane plot of 1977 seems to loop more than other years starting in June, probably because of severe sporadic rainfall within each month uncommon in the drier summer season. One large loop running from January to May could be an effect of long, consistent snow melt in the mountain regions eventually slowed by the coming of spring. This monthly, seasonal analysis is much more vivid in the phase-plane plots.

Conversely, 1969 is known to be a year of drought in the region [4]. In this case, the time series plot easily displays that this is a year of drought, with no small summer cycles and a severe decrease in intensity. But unless the scale is changed to examine this one year, phase-plane plots (Figure 15) do the analysis just as well.

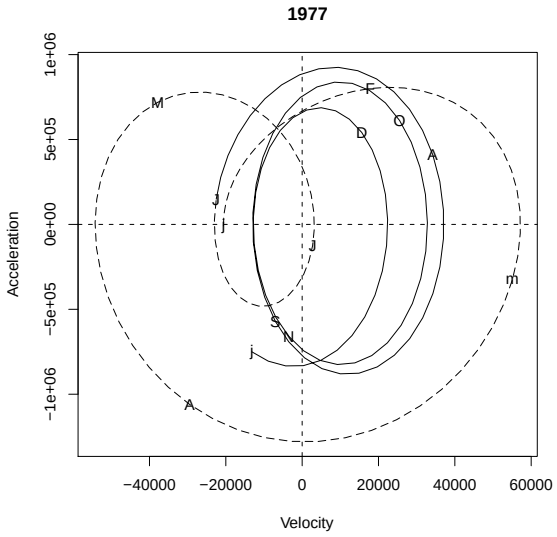

Figure 14: Flood Year

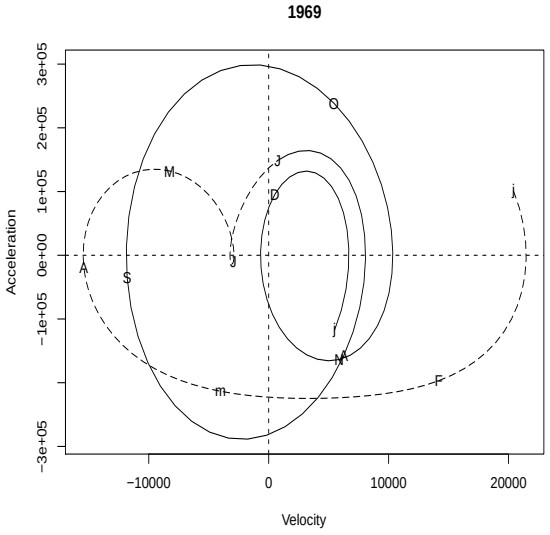

Figure 15: Drought Year

5 WORKING WITH SIMULATED DATA

5.1 The Additive Model Case

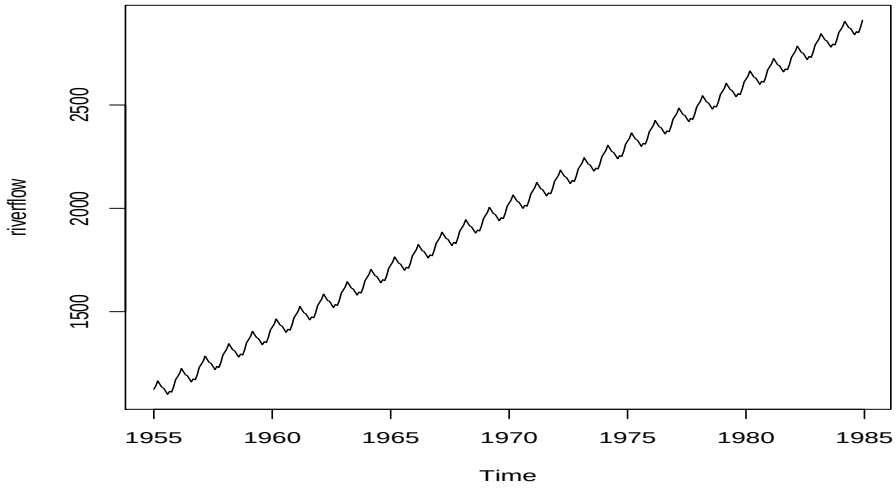

Figure 16: Simulated Additive Model Time Series

Now the question becomes: what do the phase plane plots look like when the data are simulated with a known seasonal pattern? The simulated data represent regular patterns over the years without noise, and so one would expect each year to have similar cycles within the plots. Examining the location and intensity of those cycles will prove helpful in the interpretation of the phase plane plots. Let's start with the additive model case represented by

$$Y_{a_t} = S_t + T_t + I_t$$

as mentioned earlier and simulated with Figure 16. Figure 17 shows by what coefficients that the simulated series has been changed, either with multiplication or

addition. Notice that these patterns resemble the seasonal patterns reflected by the original time series and similarities also exist with the sinusoidal pattern (although more rigid than the sine function).

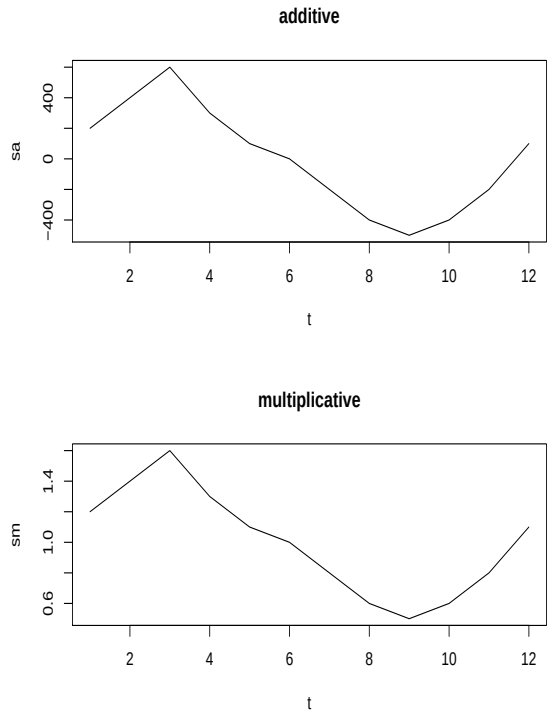

Figure 17: Seasonal Pattern for Simulations

Analyzing the phase-plane plots in Figure 18, the end years of 1955 and 1984 are included to show how the knot spacing affects these end values, however, our focus will be on the interior years. Now, comparing the years 1960, 1965, 1970, and 1975, each plot contains the same information. The intensity of each plot is basically the same, as are the locations of each respective seasonal cycle. For instance, two large cycles can be examined: the spring and fall cycles. Notice a large cycle beginning in January with positive velocity or kinetic energy and near zero acceleration. Velocity

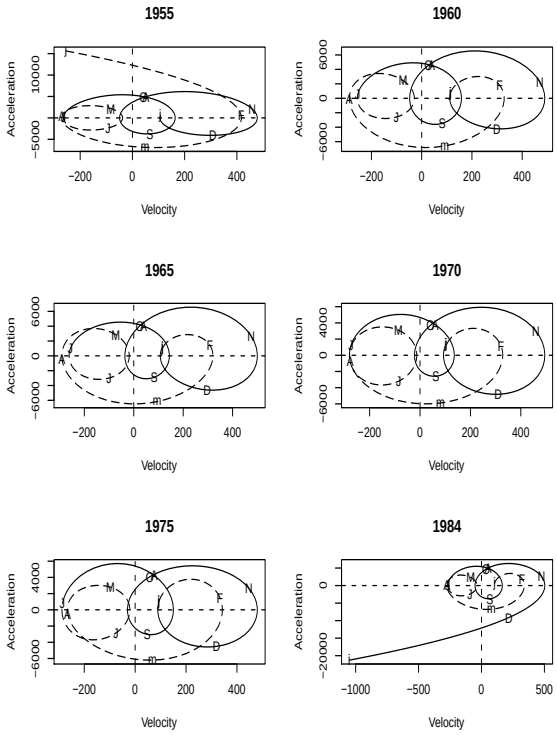

Figure 18: Additive Model Phase Plane Plot

then decreases for three months heading into April. Two smaller cycles follow: early summer with negative velocity and little acceleration, and late summer with the same phenomena. Finally, the potential energy is maximized at the beginning of fall.

Since this study is predominantly focused on the analysis of seasonal factors, what happens if the presence of a linear trend is eliminated? Will this change the phase-plane plots?

One would expect these additive model plots with no trend to look similar to Figure 18, however, this is not exactly the case. The resemblance between the additive model time series with no trend (Figure 19) and the multiplicative time series is probably the cause. Although the simulated additive model with a trend creates

a seasonal pattern, it is small in scale and with less drastic changes. The shorter periodicity of the additive series with no trend must cause the similarities between Figure 20 (No Trend Additive Model Phase-Plane Plots) and the multiplicative phase-plane plots, with the main difference persisting in the scales or intensity caused by the trend existing in the multiplicative case.

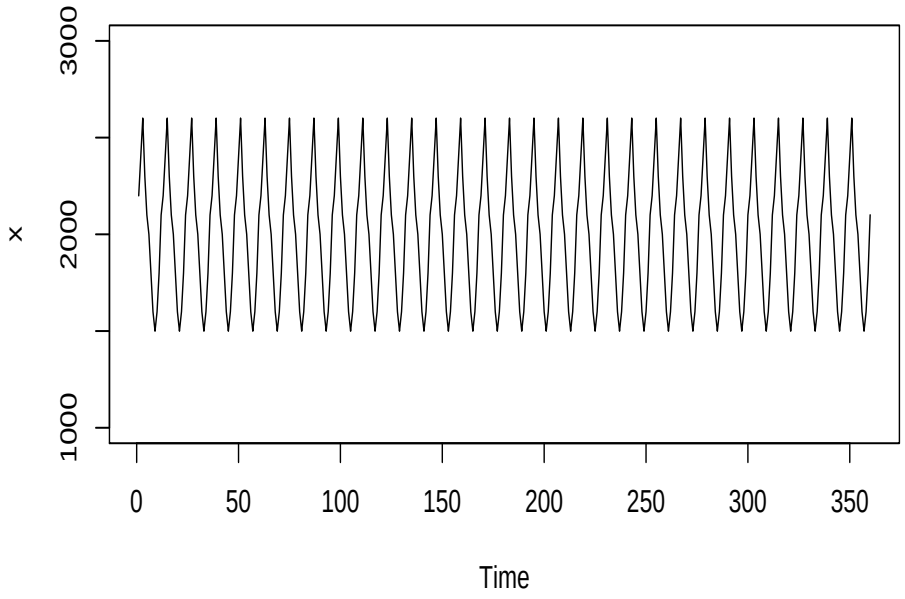

Figure 19: Simulated Additive Model with no Trend

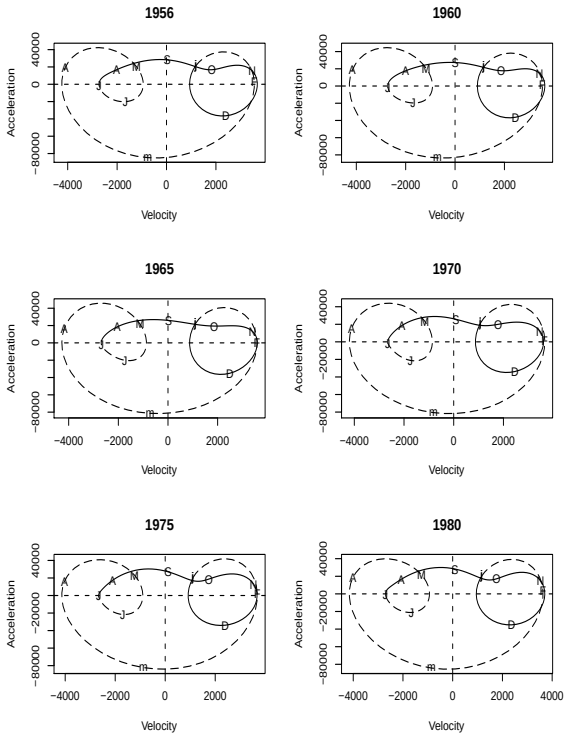

Figure 20: No Trend Additive Model Phase-Plane Plots

5.2 The Multiplicative Model Case

The multiplicative composed time series is simply a multiplication of the components previously discussed, and now represented by

$$Y_{a_t} = S_t \times T_t \times I_t$$

and having the same symbol representation. Our simulated case is seen in Figure 21. The U.S. Bureau of the Census usually uses this multiplicative process and it is often used within their research [6].

Again, the question arises concerning the effects of now multiplicative simulated data on phase plane plots. Since cycles expand in measure, one would expect the

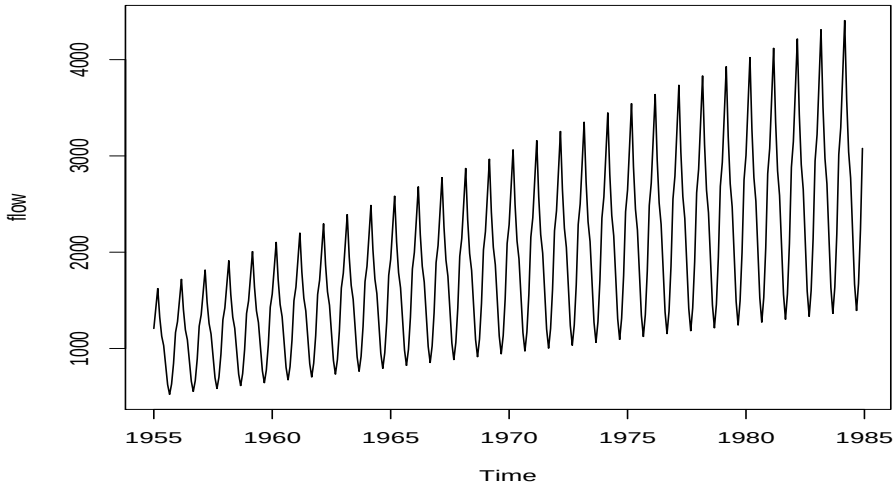

Figure 21: Simulated Multiplicative Model Time Series

intensity of the plots to change with time. However, it is suspected that the location of cycles will remain consistent throughout the years.

Examining, in Figure 22, the interior years of 1960, 1965, 1970, and 1975, the greatest visible effect is the intensity seen in the scale of each plot. While loop locations stay steady, the scale values seem to grow over the years, indicating the growth in measure of each loop and consequently the entire plot.

The first noticeable object is the large cycle similar to the simulated additive model phase-plane plot: positive velocity and zero acceleration in January followed by a large cycle for three months losing velocity and picking up acceleration into April and May. However, there is no small cycle indicating the end of summer, but rather a short transition increasing in velocity from September into the Fall months. Still yet, the similarities are striking.

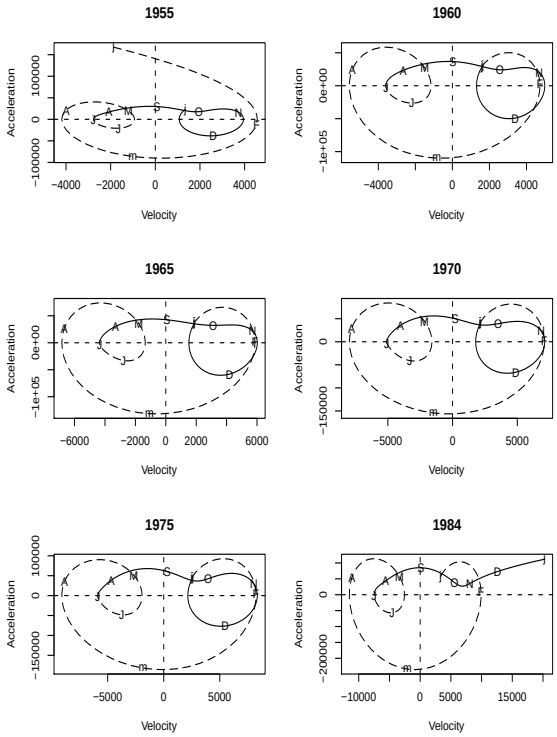

Figure 22: Multiplicative Model Phase Plane Plot

5.3 The Effect of the Smoothing Parameter Lambda

Another great way to explore the value of phase-plane plots is to play with values of the smoothing parameter λ. For instance, the value used in the simulated plots is $1e-11$, but what happens if lower or higher values for the parameter are used?

From Figure 23, with $\lambda = 1e-9$, it is easily seen that the series is still over smoothed and reveals little information compared to the plot with $\lambda = 1e-11$. Potential energy or acceleration appears to be greatest in October and kinetic energy or velocity is greatest around December–but past these facts not much analysis can be done. Now, what if the value of λ is decreased?

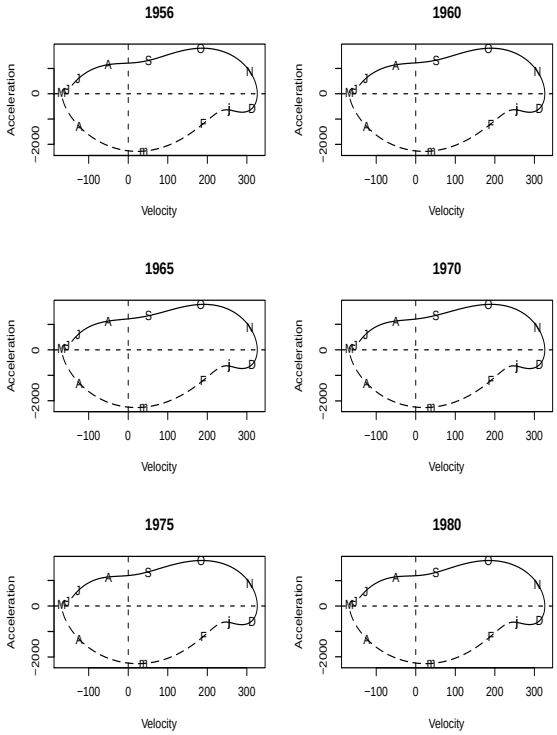

Figure 23: Simulated Additive Series, $\lambda = 1e - 9$

Not much change is observed from the plots in Figure 24 where $\lambda = 1e-13$ and the original plots where $\lambda = 1e - 11$. So as λ approaches zero, the change in smoothing does not seem beneficial.

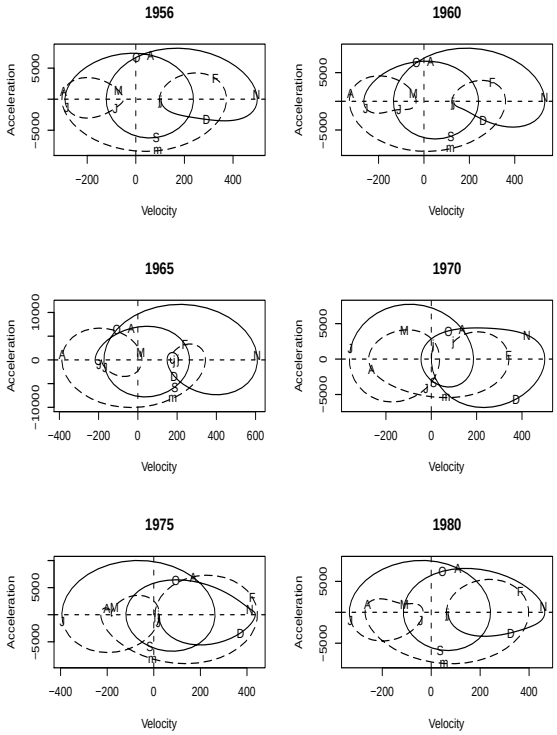

Figure 24: Simulated Additive Series, $\lambda = 1e - 13$

5.4 Effect of Noise

While the simulated data are useful for interpretation and understanding of such complex objects as the phase-plane plots, they do not represent realistic or real-world data. Therefore, by adding noise the plots of the contaminated data (Figure 25) can be easily compared to the noise free series. For simplicity's sake, only the multiplicative model will be analyzed here.

Now, analyzing the phase-plane plots in Figure 26 and starting from January, each year still indicates a large cycle extending from January to around April. Another consistency is the small cycle with negative velocity and little acceleration

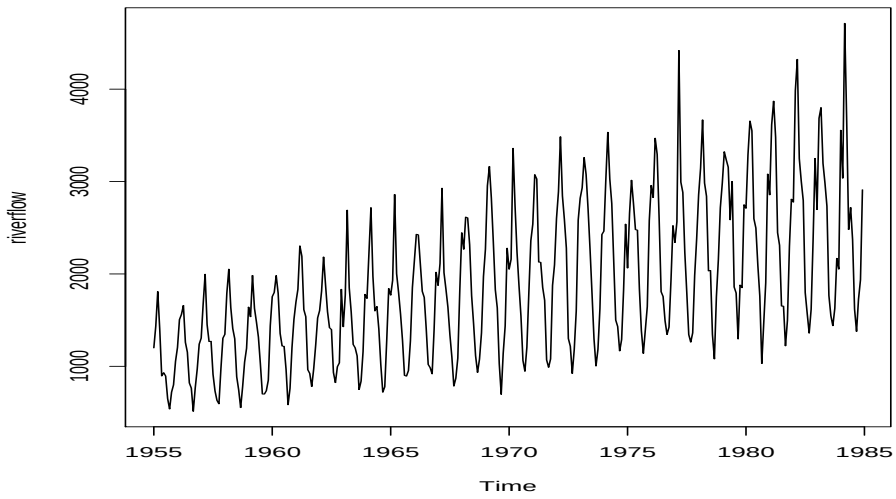

Figure 25: Simulated Multiplicative Model with Noise Phase-Plane Plot

explained by the summer or early summer months. The main difference is that the Fall months could create a large cycle of positive velocity because of environmental factors expected during that time of year. While the presence of noise does affect the phase-plane plots, overall it can be shown that even with noise there remains consistencies within the plots, and small changes from those expectations could be invaluably analyzed.

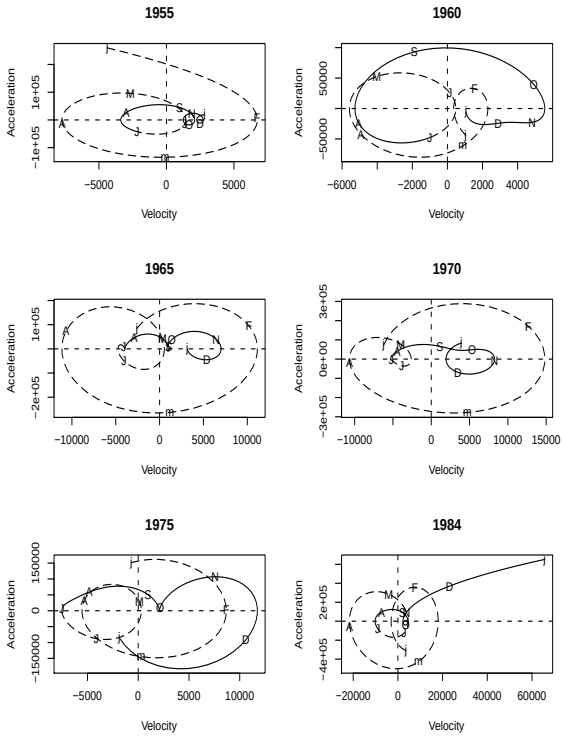

Figure 26: Noise Series Phase Plane Plots

5.5 Other Effects

In efforts to fully understand the relationship between the seasonal pattern and the shape of the phase plane plots, one must explore a variety of symmetric and non symmetric seasonal patterns. First, examine the effect when the months are switched (i.e. July-December and January-June flipped). Figure 27 gives resulting plots.

As compared with the years in its counterpart, Figure 22, the obvious difference is the switching of the months on the plots. The overall layout is still the same, but now the phase-plane plots have been flipped across the y-axis at velocity=0. Now

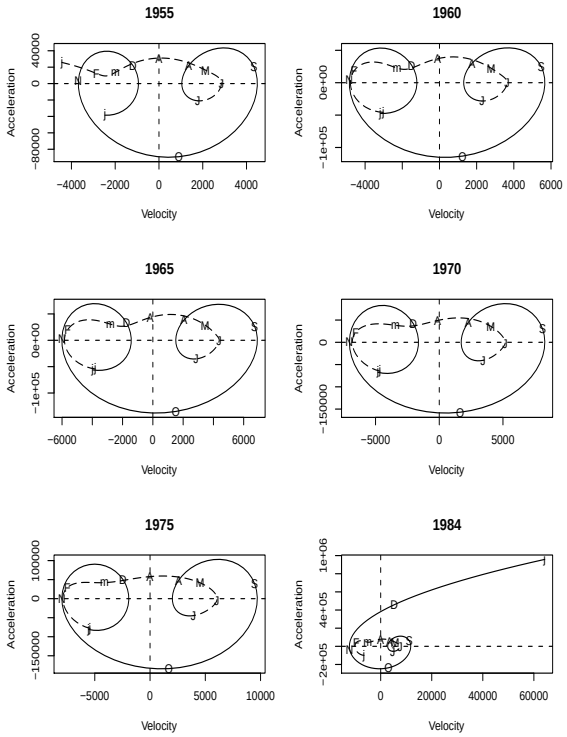

Figure 27: Flipped Months Multiplicative Phase Plane Plots

it is valuable to notice the symmetry within phase-plane plotting, especially in the "flipping" of data. So symmetry plays a rather small role concerning phase-plane plots.

6 COMBINING TIME SERIES AND FUNCTIONAL DATA ANALYSIS

The river flow of the French Broad river was analyzed in Chapter 3 using decomposition methods from classical data analysis. In Chapter 4 the analysis was done with the phase-plane plots from functional data analysis. Now the question is, what happens if these two approaches are combined.

6.1 Smoothing with Moving Average

The process of doing phase-plane plots involves a smoothing process. However, it has been seen in previous sections that phase-plane plots are pretty sensitive to noise. Time series offers several ways of smoothing data, one of them is the application of moving averages. The application of moving average becomes in this way a pre-smoothing of the time series before doing the phase-plane plots. A moving average is an average of a specified number of observations around each observation in that series [2]:

$$\bar{x} = \frac{\sum_{k=-1}^{1} x_{i+k}}{3} \tag{6}$$

For this monthly series seasonal effects must be estimated, so a moving average length of 3 is used. Figure 28 indicates how this procedure uses average values to almost smooth the series into one with less outliers or extreme values.

Now, the question remains, does this pre-smoothing procedure effect the original phase-plane plots? Since the moving averages method cuts off the first and last values, arbitrary values can be selected as placeholders for those positions as long as the phase-plane plots are not considered for the end years of 1955 and 1984.

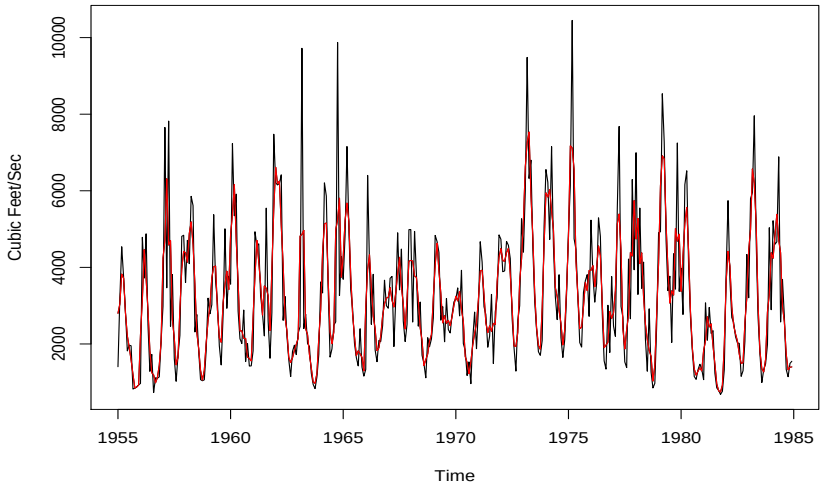

Figure 28: Time series smoothed with moving average of length 3

Comparing the phase-plane plots of Figure 29 with the original plots reveals both striking similarities and obvious contrasts. On one hand, since outliers were reduced, the intensity of the years (noticed in the scale of the plots) has been reduced for most. Also, particularly in the years of 1965, 1975, and 1980, the typical large cycle from January to May is not dramatically changed. An approach toward zero acceleration in 1960, 1965, and 1975 around February and March seems to be a common consequence of the moving averages within that large cycle, perhaps because crest periods can easily be observed in these months and then severely reduced because of moving averages. Also, the small cycle centered around zero acceleration and zero velocity occurring in the months of October and November appears to remain consistent despite the moving averages method. It seems almost certain that the method reduces the drama within the plots. Large cycles seem to be easily reduced and the plots,

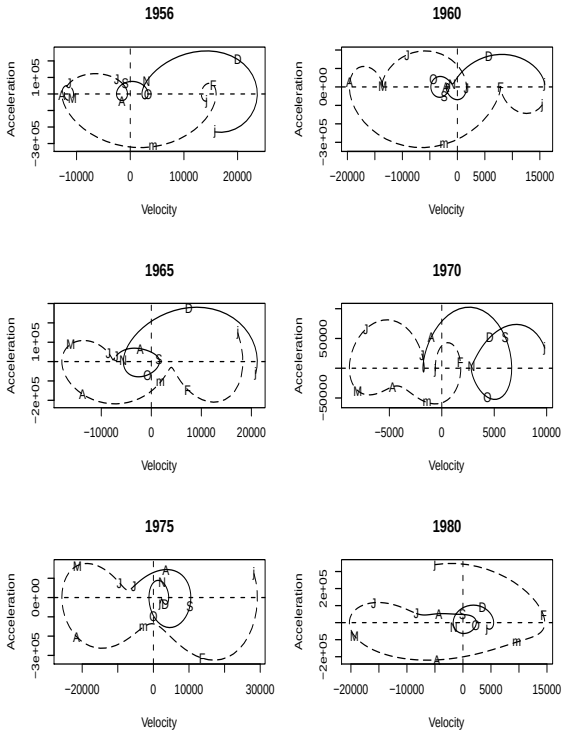

Figure 29: Moving Average Phase-Plane Plots

since less congested without extreme values, may be easier to read. However, if cycles and energy exchanges are reduced too much then the plots become simplistic and the true character of the data could be downplayed. It's important for the researcher to understand how smoothing affects the plots here.

6.2 Seasonal Component Phase-Plane Plots

After using X-11 decomposition, time series plots of seasonal, trend, and irregular components were obtained and examined earlier in Figure 5. So, to keep with common analysis, a look at the phase-plane plots of those components seems appropriate. The main focus is on the seasonal component, so the phase-plane plots in Figure 30 will

be restricted to seasonality. Keep in mind that before, phase-plane plots were used as a way to indicate or see seasonality. But now, the seasonal component is viewed to see how it changes annually.

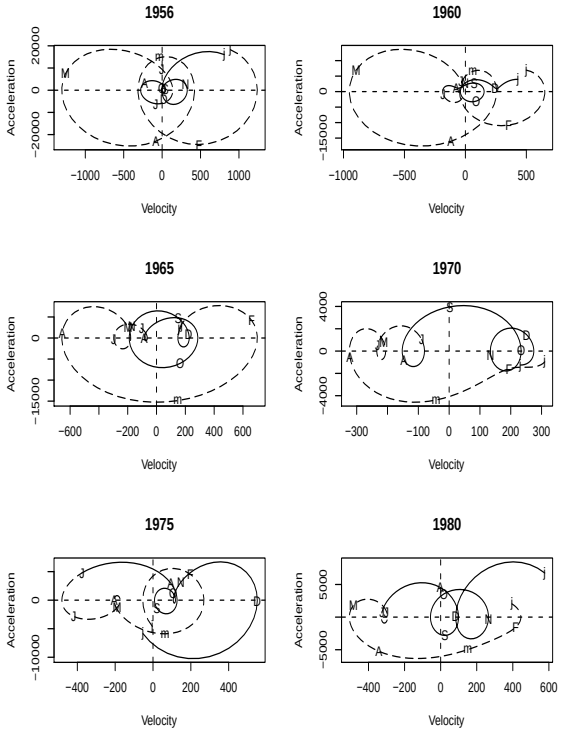

Figure 30: Seasonal Component Phase-Plane Plots

The first obvious note is that each plot is different, a consequence of the variability in river flow among the observed years. Because this is now a plot of the seasonal component itself and not the actual data, the expectations change a bit. For instance, the plot now tells when and with what intensity that the seasonality changes, not when the river flow fluctuates. One would expect that heavy rain and perhaps snow melt would contribute to large changing seasonal factors in winter and early spring. Summer often brings consistencies or little changes, and fall tends to bring potential

unpredictable changes that could contribute to large intensity seasonal factor within these plots.

From analyzing Figure 30, one can see that the summer months bring a small cycle with little to no acceleration and sometimes negative velocity. The seasonal component sees its smallest variation within the summer. February, March, and April seasonal factors seem to begin with positive velocity, decrease, and maintain negative acceleration. However, because they usually lie within a large cycle, this indicates large intensity or presence of the changes within seasonal component. Also, small cycles are seen with positive velocity around October and November, a similarity even to the phase-plane plots of simulated data.

So, although the goal is not to detect the seasonal factor, by plotting it we can see that it changes in a similar fashion as the time series object itself. However, the clarity of analysis within these phase plane plots is murky compared to Figure 5, where the time series plot of seasonality marked an obvious repeated annual function.

7 CONCLUSIONS

From a statistical standpoint, time series analysis is extremely beneficial. From a mathematical standpoint, functional data analysis adds a modern twist on typical analysis. While one method is not meant to replace the other, each has advantages over the other. From this research, it appears that functional data analysis (particularly with phase-plane plots) compliments time series analysis fairly well.

It should be noted that traditional time series plots do prove advantageous over a longer time scale, particularly in examining the long term trend of data. While phase-plane plots offer great yearly and monthly analysis, prediction based on past events becomes much more difficult. One can see how and when typical years change seasonally, but limitations arise concerning expectations of maximum and minimum water flow in this case.

Also, a time series plot easily exposes outliers and extreme values, but this becomes more difficult in phase-plane plotting. Since single values are not shown, only the scale or intensity of the plots can be affected–and even that can be affected by more than one outlier. This is where the moving averages method appears to prove useful. Although a smoothing parameter λ is chosen to smooth the data, moving averages reduces the scale and appears make the phase-plane plots easier to read, and with their interpretation being difficult, this could be a wonderful tool. However, what is lost with moving averages may be another area to explore.

One other explored area is the effect of noise on phase-plane plotting. A multiplicative model is simulated and phase-plane plots created, showing consistent changes in summer and late fall and slow changes in late winter and early spring. However, once

a slight amount of simulated noise is added the sensitivity of the phase-plane plots is revealed. Although the overall pattern of changes remains similar, the rate at which velocity changes and the rate of acceleration differs dramatically.

Phase-plane plots compliment the research behind the time series approach, but the plots cannot replace the valuable numerical information gained from that approach. Functional data analysis also provides methods for algebraic or numerical interpretation, and those advantages could be explored from the context of detecting seasonality by future researchers.

The pre-smoothing of our moving averages method may be a particular area for other researchers to explore. It appeared that the early smoothing reduced the drama and congestion within the phase-plane plots, but with occasional loops being minimized or even eliminated. This could present problems, depending on the questions of the researcher. However, the big picture of this analysis remained even with moving averages. It did not as drastically alter the phase-plane plots like the small changes in the smoothing parameter λ.

The analysis of the phase-plane plots of the seasonal component should be approached a little differently. First, the search was for obvious cycles indicating change in the factor. Then, the timing of those seasonal factors is compared with what is expected which corresponds to the phase-plane plots of the original data. The sinusoidal presence within the seasonal series plot caused resulting phase-plane plots similar to the original series. However, phase-plane plots of seasonal components do not present equal clarity of seasonal presence when compared to classical time series analysis.

Therefore, it can be shown that phase-plane plots are particularly useful in the detection of seasonality. However, obtaining the right plots with the right amount of detail is a very tedious task, more difficult than traditional time series plots. Choosing the smoothing parameter λ is is difficult and the answer ambiguous. Although functional data analysis alone can be useful for prediction and forecasting, the phase-plane plots themselves do not seem to offer an advantage in predictive modeling. However, the exchange in energy shown by these plots appears to be particularly useful in environmental/physical data analysis where energy transfer is constantly shifting, and its applications should be explored more thoroughly in the future.

BIBLIOGRAPHY

[1] K. Chan and J. Cryer, Time Series Analysis with applications in R, Second edition, Springer, New York (2008).

[2] P. Cowpertwait and A. Metcalfe, Introductory Time Series with R, Springer, New York (2009).

[3] S. Graves, G. Hooker, and J. Ramsay, Functional Data Analysis with R and MATLAB, Springer, New York (2009).

[4] HAMweather, a WeatherNation Company, [`http://weather.hamweather.com/rivers/gauge/AVLN7.html`] (2011). Accessed June 2011.

[5] R. Hyndman, S. Wheelwright, and S. Madridakis, Forecasting Methods and Applications, 3rd edition, John Wiley & Sons, Inc. (1998).

[6] M. McGee and R. Yaffee, Introduction to Time Series Analysis and Forecasting, Academic Press, Inc. (2000).

[7] N.C. Office of Environmental Education, [`http://www.ee.enr.state.nc.us/public/ecoaddress/riverbasins/frenchbroad.150dpi.pdf`] (2007). Accessed June 2011.

[8] R Development Core Team, R: A Language and Environment for Statistical Computing, Vienna, Austria, 2006. [`http://www.R-project.org`].

[9] J. Ramsay and B. Silverman, Functional Data Analysis, 2nd Edition, Springer, New York (2005).

[10] RiverLink, Inc, [http://seris.info/RiverLink/main.shtml] (2002). Accessed June 2011.

[11] United States Geological Survey, [http://waterdata.usgs.gov/tn/nwis/monthly/?referred_module=sw&site_no=03455000&por_03455000_1=1112466,00060,1,1900-11,2010-09&format=html_table&date_format=YYYY-MM-DD&rdb_compression=file&submitted_form=parameter_selection_list] (2011). Accessed June 2011.

[12] G. Yule, An Introduction to the Theory of Statistics, C. Griffin, London (1950).

APPENDICES

Appendix A: R Code

```
# Original Time Series Plot

river <- scan(‘‘f:/DATA.dat")

riverflow <- ts(river, start=c(1955,1), frequency=12)

ts.plot(riverflow)


# Periodogram


perioplot<-function(x){

adjx=x-mean(x);

tf=fft(adjx);

nf=length(tf); n2=nf/2+1;

 pritf<-tf[c(1:n2)];

intensity<-(abs(pritf^2))/nf;

nyquist=1/2; pfreq<-seq(0,nf/2,by=1);

 freq<-pfreq/(length(pfreq)-1)*nyquist;

intmax<-max(intensity)

posmax<-max.col(t(intensity))

 freqmax<-(freq[posmax])

 maxper<-1/freqmax

 plot(freq,intensity,type="l")
```

```
 text(0.2,intmax, label= maxper)}


perioplot(riverflow)


# Phase-Plane Plots (fda package required)


flowbasis <- create.bspline.basis(rangeval=c(1955,1985),
nbasis=366, norder=8)
LfdobjWater <- int2Lfd(4)



WaterSm <- smooth.basisPar(argvals=index(riverflow),
y=(coredata(riverflow)), fdobj=flowbasis,
Lfdobj=LfdobjWater, lambda=1e-11)



par(mfrow=c(3,2))
 phaseplanePlot(1956, WaterSm$fd, main='1956')
phaseplanePlot(1960, WaterSm$fd  , main='1960')
phaseplanePlot(1965, WaterSm$fd ,  main='1965')
 phaseplanePlot(1970, WaterSm$fd, main='1970')
phaseplanePlot(1975, WaterSm$fd,  main='1975')
phaseplanePlot(1980, WaterSm$fd,  main='1980')
```

```
# Simulated Sine Function


t<-seq(1,12,by=0.1)

y<-sin(2*t*pi/12)

plot(t,y,'l')
```

Appendix B: SAS Code

```
options ps=1000;

data jake ;

  infile 'f:\data.dat' ;

    input flow ;

    date=intnx('month', '01jan1955'd,_n_-1);

format date monyyyy;

proc print;

proc x11 data=jake ;

  monthly date=date;

  var flow;

  tables d10 d11 d12;

  output out=comps b1=series d10=season d11=adjust d12=trend d13=irre ;

proc print data=comps;

  run;
```

VITA

JAKE ALLEN

Education:	B.A. Mathematics, Carson-Newman College, Jefferson City, Tennessee 2009 M.S. Mathematics, East Tennessee State University Johnson City, Tennessee 2011
Professional Experience:	Graduate Assistant, East Tennessee State University, Johnson City, Tennessee, 2009–2011