O Ye of Little Faith
- Forecasting Old Faithful

Chang Shen & Dan Zhao

12/13/2019

Contents

Introduction **1**
- Background . . . 1
- Data Resource . . . 1
- Data Manipulation . . . 1
- Exploratory Data Analysis . . . 3
- Multivariate Analysis . . . 7

Appendix **11**
- Reference . . . 13

Introduction

Background

Old Faithful is a cone geyser located in Yellowstone National Park in Wyoming, United States. It was named Old Faithful in 1870 during the Washburn-Langford-Doane Expedition and was the first geyser in the park to receive a name. The analysis presented here attempts to forecast eruption durations, eruption occurance duration, and eruption intervals (i.e. intervals between consecutive eruptions). Generally speaking, how eruption time, year, and duration will influence the duration still a mystery.

Data Resource

From 1970 onwards, with somewhat irregular times for a few years in between, the Geyser Observation and Study Association (GOSA) began collecting and reporting data on geysers and other geothermal phenomena in Yellowstone National Park and elsewhere("Geyserstudy" n.d.). To study and explore dynamics behind driving , we start with the original data that records eruption, duration, etc. (Old Faithful Visitor Center Logs from (http://www.geyserstudy.org/ofvclogs.aspx). These data contain recorded observations in the form of logs from 1970 to 1981 and are compiled by Marion Powell and Mary Beth Schwarz while logs from 1981 to 2012 are compiled by Lynn.

Data Manipulation

The raw data has 460,335 observations. A preview of the data is presented below:

x

Date Geyser Time ie VR Interval Duration Preplay Height Predict Bar. Pres OF Com
1/3/1970 Lion 9:36
1/3/1970 Lion 11:05
1/3/1970 Lion 13:20
1/3/1970 Lion 15:45
1/3/1970 Vault 8:41 ~ 20 feet

x
1/5/1970 Giantess noise/ofl
1/7/1970 Beehive 8:30
1/8/1970 Round 10:05 lovely-15
1/11/1970 Silex “dwn~20””””

From the preview above, it can be seen that the data are unformated with several features containing a mix of alpha-numeric data which make feature extraction and modeling difficult. As such, quite a bit of data structuring is required.

1. Defining & Structuring Features

The raw file contains not only data pertaining to Old Faithful but also to other geysers in the Yellow Stone. We first use filters based upon regular expressions the oldfaiful related observations and the seperator()ˆ to determine the column variables. The orignal data come with several variables (Date/Geyser/Time/ie VR/Interval/Duration/Preplay Height/Predict/Bar. Pres/OF/ Com). For our purposes, we reduce the number of relevant features to the following four:

- **Date**: The date of an eruption(formated to “yyyy-mm-dd”)
- **Time**: The start time of an observed eruption
- **Interval**: The time interval between this eruption and the last observed eruption.
- **Duration**: The eruption’s duration

2. Data cleaning

After familiarizing ourselves with the data, we proceeded by structuring and cleaning the data as follows:

- Missing observations
  - Remove rows where all variable values are NA
  - Date imputation: if a time exists without a date value, the date is imputed by looking at nearest date that is missing one of these observation times (i.e. observation time is in 24 hour military time so it is easy to tell if a date is missing a morning or evening time, for example)
- Cleaning the ‘Date’ variable
  - Remove irrelevant non-numeric characters from date values (e.g. ‘~’, ‘-’, etc.) with the exception of the colon (‘:’) to differentiate hours from minutes
  - Correct the time portion of the date value to a standard format (e.g ‘17;00’ to ‘17:00’)
  - If a date value had a field with an irrelevant sentence or string, the observation was deleted
  - Standardize all date values in date field to ‘yyyy-mm-dd’
  - Check if dates are assorted in the correct chronological order in terms of year, months, and days; also as a double-check for mis-prints in date (e.g. ‘2011-09-01’, ‘1011-09-02’, ‘2011-09-03’ should result in the second value being corrected to ‘2011-09-02’)
- Cleaning the ‘Duration’ variable
  - Remove all unnecessary spaces and blanks in each value (e.g. ’ ’)
  - Correct relevant punctuation (e.g. ‘;’ to ‘:’ for values in terms of time or ‘.’ to ‘:’ like ‘17.00’ to ‘17:00’)
  - Convert values from (hour:minute:second) to (minute:second); i.e. convert hours to minutes
  - For duration values that show a range, rather than numeric values, such as (‘2:00 - 3:30’), the median is taken as imputation
  - Convert duration from (minute:second) to just minutes
  - Format values to be consistent with numeric values (e.g. “4 1/2” to 4.5)

  - Convert values with any string portion into a compatible format (e.g. 3m to 3; 60s to 1)
  - convert the place holder(X/x) to 0(e.g. for 02:2x we convert the value to 02:20), this might casuse a bias however more information for better imputation(!!!)
  - For descriptive string values, strings that describe the duration in words, such as those in the set (e.g. 'Longer', 'Short' or 'S' etc.), these values are all standardized so as to belong into 'L', 'M', or 'S' for 'Long', 'Medium', or 'Short'
- Cleaning the 'Time' variable
  - Convert values from 12-hour time schedule to 24-hour clock schedule
  - Reformat values as h:m:s
  - Eliminate the duplcated Times in each day
- CLeaning the 'Interval' variable
  - Remove characters and strinsg to leave numeric values
  - Reformat into minutes (creating a new variable called 'Interval_lag')
  - Correct erroneous values due to row misalighment (e.g. if interval value is associated with row 1 but is in row 2 instead, it is corrected and placed back into row 2)
- Outliers Mark an observation as an outlier (via an indicator variable 'outlier') when duration is greater than 5 minutes and or interval is greater than 3 hours per John's past analysis. Keep outliers in the final data set, but we won't discuss them in later analysis part

Exploratory Data Analysis

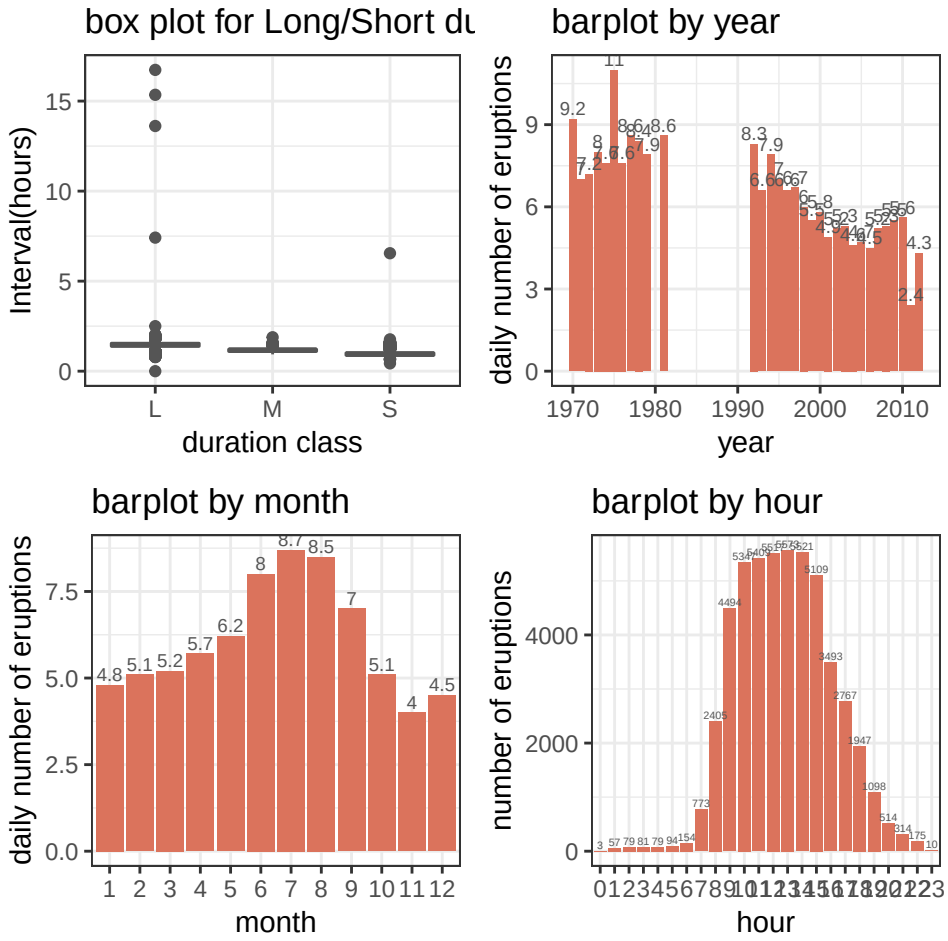
1. First Impressions

The plots below attempt to provide an initial description of how interval and duration relate to one another through our data depending on the 'size' of duration (L/M/S) as well as how interval and duration themselves evolve over time, across different time scales.

Starting from the top left and going clockwise, the first plot is a box plot of interval values (in hours) by duration class/size (i.e. whether a duration is classified as L/M/S or large, medium or small). This is done because comparing interval against duration based on their numeric values alone would: (i) obfuscate any clear trends in the data

The second plot plots the number of eruptions (i.e. the average number of eruptions in each day) over the years, the third shows the number of eruptions by hour (averaged by hour over all the years), and the last shows the number by month (averaged over each of the months).

Some initial impressions: - Looking at the triple boxplot in the top left for observations whose durations are marked as (L/M/S), we see that eruptions with longer durations (in the group 'L'), tend to be associated with longer intervals (until the next eruption). One can imagine that this may describe a geological process which can build up and accumulate energy for a long eruption but may take a long time to 'recharge' before the next one. However, it seems that some interval values associated with 'S', or small, durations exhibit more variance in its values than ones in the 'M' or medium class. - We can see that over the years, the average number of eruptions within a day tend to follow a decreasing trend overall. Meanwhile, we see that, in the third plot, the average number of eruptions tend to follow a slightly left-skewed distribution when it comes to a 24-hour scale where most of the eruptions in our data tend to occur between 7:00 and 19:00. Finally, the last plot shows that most of our eruptions, on average, tend to occur in the summer months of April to August/September before turning less active during the winter months.



2. Overall Trends

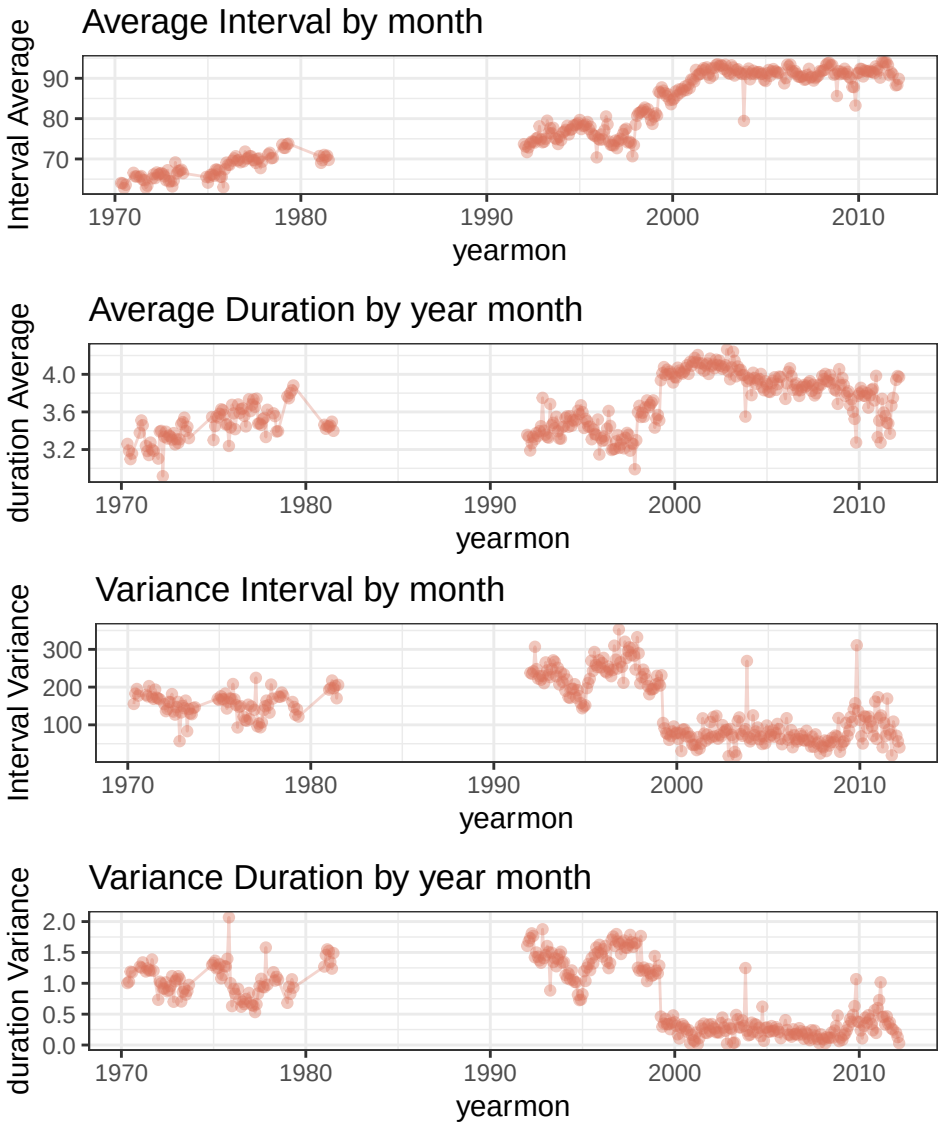
To see whether there may be distinct trends in the data and get a feel for the dynamics behind duration and interval, we average the data for each month and then plot the month-year values for all years in the data (i.e. 12 values for 1970, 12 values for 1971, etc. with each of the 12 values of a year being the average of the values in the month of that year). Note that there is a break between the early 1980s to 1990s due to missing data on interval and, as a result, duration.

From around 1995 to 2004, we see an uptick in both average interval and average duration; the average monthly interval stabilizes afterwards but the average duration trends downward. As for the sample variance (within month), we see both duration and interval seeing a significant drop in their respective sample variance from late 1990s onward with the exception of 2 to 3 outliers. As such, there are roughly two regimes in variance for both duration and interval: pre-1990 and post-(late 1990s).

From this alone, given that average interval is much higher from 2000 onward but the variance is somewhat lower, suggesting that the trend is solidifying around a higher average interval, at least on average from month to month. Similarly, average duration is higher from 2000 onward than from its regime pre-1990 but has shown a slight dip starting from the early 2000s despite still being higher than average duration in 1990 and earlier.

Consequentially, this preliminary analysis might tell us that tourists who want to visit Old Faithful now, they may see eruptions which last longer than before 2000, but this comes at having to wait, on average, a longer interval before the next eruption if they miss a preceding eruption. The relativey low variance from 2000 onward suggests that this analysis or rules of thumb for duration and interval are not likely to undergo too much of a drastic change.

However, there are only marginal analyses–to see how duration and interval may co-vary (to leverage for predicting eruption timing etc.), we move onto the covariance/correlation analysis below. But first, we examine if there may be seasonality for duration and interval; we suspect that if seasonality were to exist, it would likely exist on the monthly level given the geological and geophysical nature of geyser eruptions.



3. Seasonality

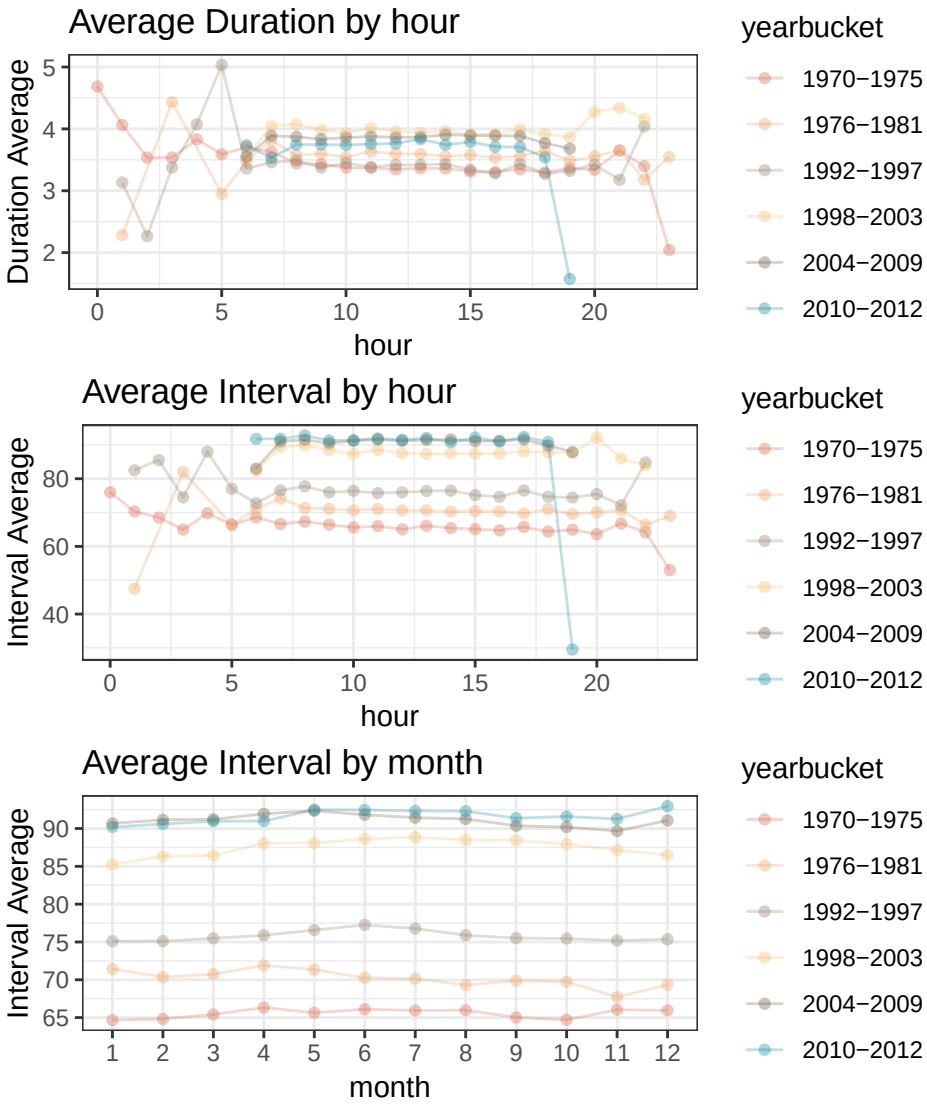
To survey the data for seasonal patterns, we summarized the interval and duration of geyser eruptions by month for every five years (bottom two plots). We do this in order to better segment the data and focus on whether monthly or seasonal trends over the years have strengthened or weakened. We repeat these visualizations but on an hourly level too; this is to see if any patterns persist on a daily 24 hours scale (top two charts).

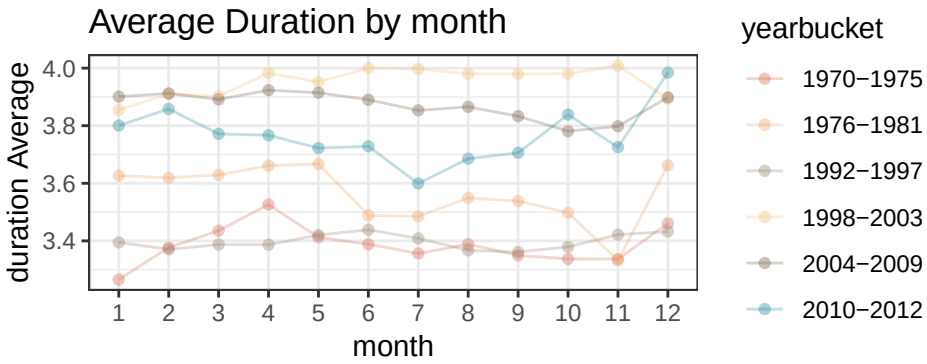
In the top two plots, we see that the average hourly duration values across the years tend to be relatively consistent: from 5:00 to around 18:00, the average duration of about 3 to 4 minutes has remained roughly the same with no trend over time. However, outside these hours, we see considerable variation in average duration by hour. On the other hand, verage interval length has seen a gradual increase over the years over the hours from 5:00 to 17:00, rising from about 65 minutes to about 88 minutes over the years. At least on the hourly scale, it appears that the average interval has increased over the years across the hours of a

workign day while average duration has roughly stayed the same in a 3 to 4 minute band within the same hours.

For average interval in the bottom two plots, we see that over the years, an overall trend has been that the average interval is increasing across all months which corroborates our findings in the section earlier above. We also see in more recent years (past 15 years), there is a slight seasonal bump in average duration from April to August as well as from November to December (past 10 years).

For average duration, the situation is more complicated. There does not seem to be any clear rhyme or reason over the years in terms of monthly trends. This may be a clue for where difficulties in prediction may lie in later analysis. The only clear sign of potential seasonality in duration is the spike from November to December across almost all years.





Multivariate Analysis

1. Visualizing Joint Dependency

Since there are many time gaps within the data (e.g. gaps of 6 months/1 year/10 year without any data), conventional time series analysis for modeling marginal or joint relationships is unlikely to be helpful. Imputation of said gaps would also be unhelpful due to the long gaps of unobservables. But because the data still constitute a time-series (with significant time gaps and irregular reporting intervals), we still hope to incorporate some aspect of temporal structure into our modeling efforts. Our end goal is to model future interval length (interval[t]) as a function of the duration of the most recent eruption (duration[t]), the duration of the eruption before the most recent (duration[t-1]), and the interval length between these two eruptions (duration[t-1], i.e. the wait between the most recent eruption and the preceding eruption). In essence, what this means is that we aim to predict when the next eruption will be.

A note on the potentially confusing time indices on the variables. For some fixed time, say t, interval[t] and duration[t] may share the same time index, but the value of interval[t] is only realized when the next geyser erupts at some t+1—because only then can the interval be calculated via taking the difference of t+1 geyser's start time and the t geyser's end time. Interval, on the other hand, only tracks the length of the geyser eruption when it happens at/within time t. Similarly, this means that interval[t-1] is the interval between geyser at time t and the geyser at time t-1. As a result, despite the same time indexing, interval effectively leads duration by a lag of 1. Therefore, forecasting when the next eruption will occur is equivalent to our stated goal: forecasting interval[t] based on duration[t], duration[t-1], and interval[t-1].

We first look to see what the bivariate relationships between some of the relevant variables may look like. Below are bivariate scatterplots with each variable's estimated marginal density drawn on its opposite axis (i.e. if interval is on the y-axis then its estimated marginal density is outlined on the opposite side, the right axis). The marginal densities are estimated via kernel density estimation using a standard Gaussian kernel and default bandwidth parameters.

Starting in the top row moving from left to right, we have plots of:

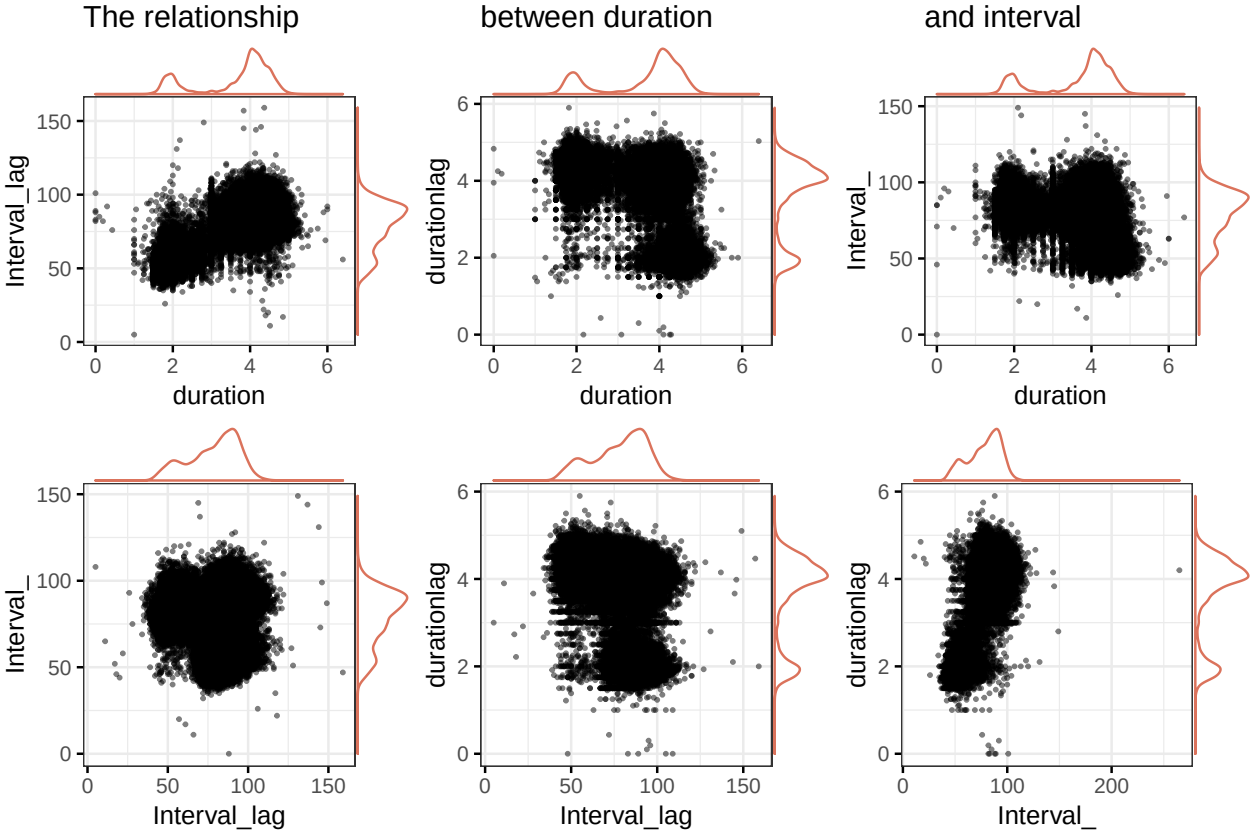
- duration[t] vs. interval[t-1]
- duration[t] vs. duration[t-1]
- duration[t] vs. interval[t]

Lastly, in the bottom row from left to right:

- interval[t-1] vs. interval[t]
- interval[t-1] vs. duration[t-1]
- interval[t] vs. duration[t-1]

Overall, from the plots below, we clearly see several centers of mass in each pair of variable visualizations. Additionally. all the marginals on the axes of the plots exhibit bi-modality (having more than one peak with usually two defined peaks). This highly suggests the presence of multiple distributions. A mixture model or cluster-based model is probably best to characterize the joint distribution between these variables–moreover,

this also suggests that the best way to model the trivariate distribution of duration[t], duration[t-1], and interval[t-1] to forecast interval[t] is through a mixture model of some sort.



2. Gaussian Mixture Model

To model the joint dependence between duration and interval in the presence of several regimes and multiple modes, we employ a Gaussian mixture model (Titterington, Smith, and Makov 1985). A Gaussian Mixture Model (GMM) is an unsupervised mixture model which parameterizes the observed variables as normal random variables. The parameters and weights for the individual components and the weights, respectively, are typically estimated via some form of iterative Expectation Maximization (EM). Visual inspection of the scatterplots show about 3 distinct clusters and, as such, the number of components in the mixture was decided to be set to three.

Initial inspection of the scatterplots reveal several distinct 'regimes' or clusters while the estimated marginals are usually bi-modal. Typically, these signs are telling of multiple population/segment overlap. Methods to resolve this issue usually boil down to either an indirect method—finer segmentation/weighting of the underlying sample—or direct method–modeling the mixed distribution directly via a mixture model or hierarchical model. Under ideal circumstances, we would model a multivariate distribution (or mixture model) between all four variables of interest based on past data: duration[t], duration[t-1], interval[t-1], and interval[t]. To make forecasts, we would make a new observation of duration[t], duration[t-1], and interval[t-1] before looking to the joint distribution to see the conditional probability distribution of interval[t] given the realized values of duration[t], duration[t-1], and interval[t-1]. From there, we could make probabilistic forecasts with confidence intervals around what range of values we would expect for interval[t] and, therefore, when the next eruption may occur.

However, sometimes we may not observe interval[t] if we start extrapolating/inferring away from relying on past interval values. Instead, we fit a mixture model on three dimensions: duration[t], duration[t-1], and interval[t-1]. After training our model, we then use the fitted model to characterize our existing data into different regimes/clusters. We then compare plot each dimension of our mixture model against interval[t]

with values color-coded by the cluster/mixture component that they are classified. The visualized plots are below along with the fitted parameters.

Using our GMM model with the variables duration[t-1], interval[t-1], duration[t] with a pre-determined cluster number 3, the final model on the best iteration achieved a BIC -4.5114933×10^5 and a Log likelihood -2.2542026×10^5.

+ Fitted Parameters

Mean of each component/cluster The means of each dimension/each variable within each component.

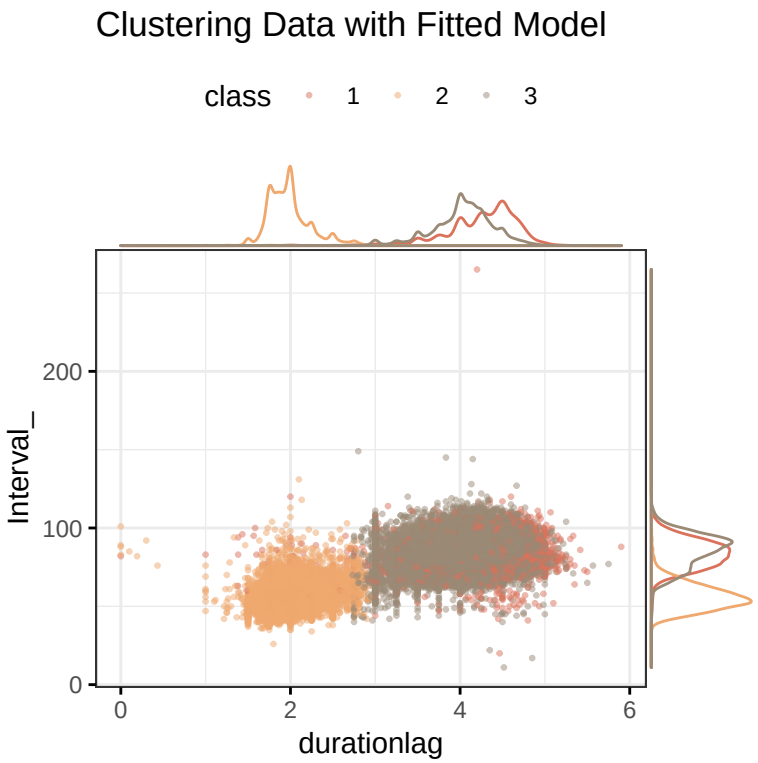
model	cluster 1	cluster 2	cluster 3
Interval_lag	85.10	56.13	84.88
duration	4.00	1.99	4.41
durationlag	4.04	4.27	1.984

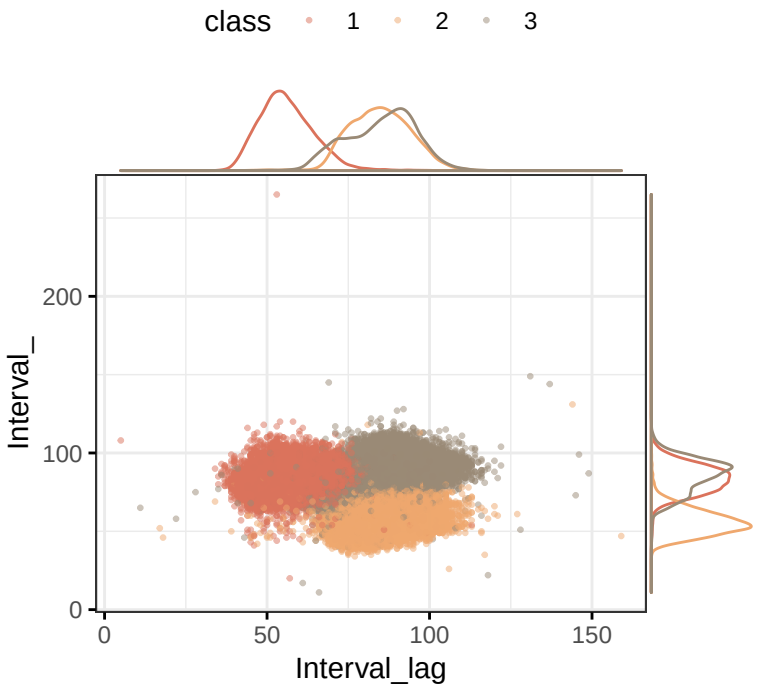
While cluster 1 and cluster 3 are quite similar in their means of interval[t-1] and duration[t], their means are quite different for duration[t-1]. Cluster 2 also stands out quite a bit in its means for interval[t-1] and duration[t].

Fitted component weights

model	w_1	w_2	w_3
Value	0.525	0.241	0.234

Note that cluster 1 has almost twice the weight of cluster 2 or 3 and that cluster 2 and 3 have about the same weight.





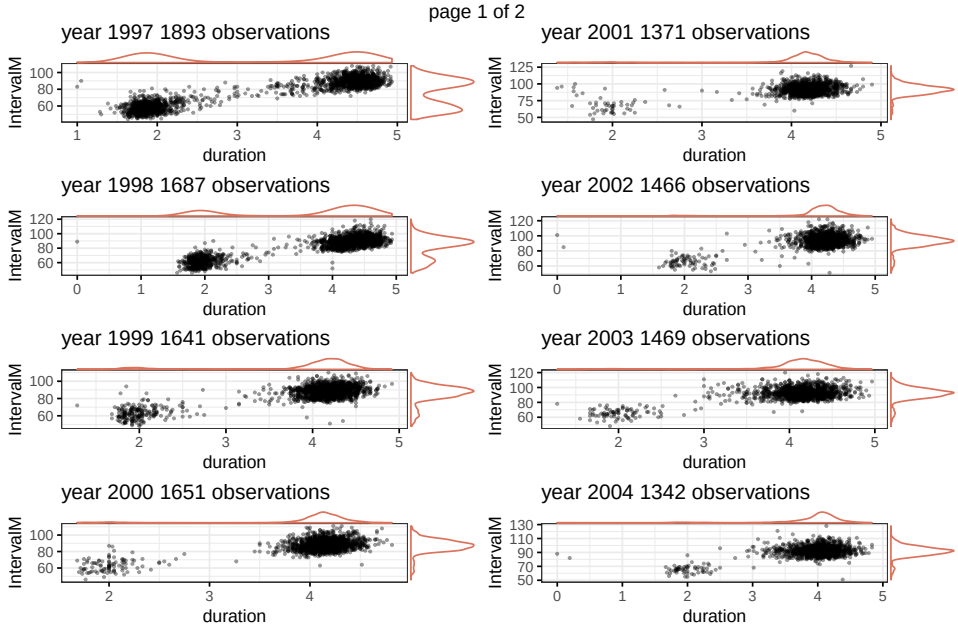
As mentioned earlier, after fitting the mixture model with three components/dimensions (duration[t-1], interval[t-1], duration[t]), we use the model to categorize our existing data points as belonging into one of the three fitted components. We then plot each variable used in the mixture model against our main variable of interest, interval[t], but color code the values based on component/cluster membership.

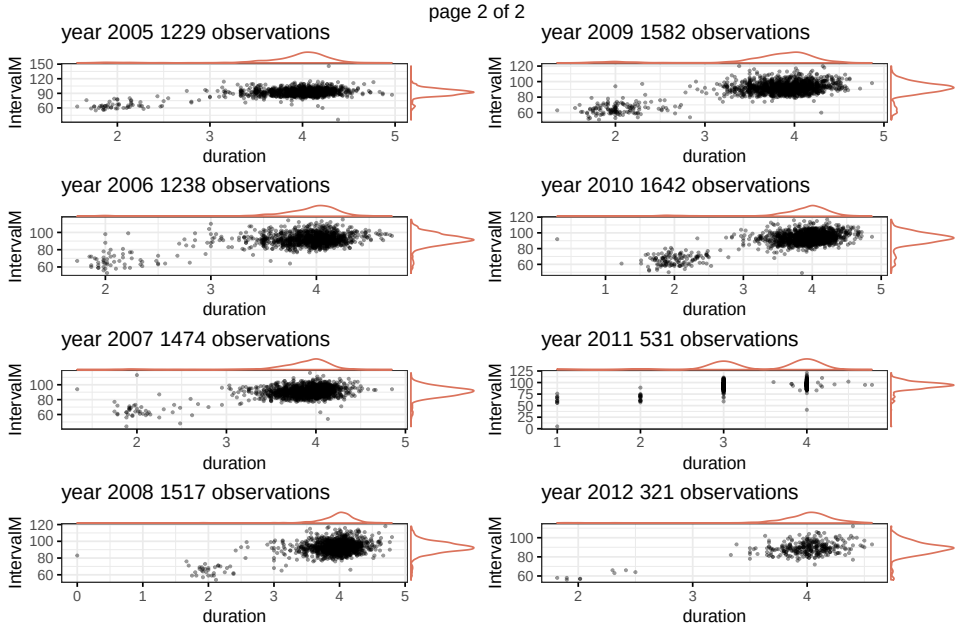
For instance, for all three plots above show interval[t] plotted against each one of the three variables in

the mixture model, like in the plots earlier above, but now the values are color-coded to indicate which cluster/component they each belong to. As an example for forecasting with this model, suppose we would obtain a new observation that had realizations of (4, 4, 100) for (duration[t], duration[t-1], interval[t-1]). Then, based on the plots, we see that the first plot a value of 4 is associated with clusters 1/2, the second with 1/3, and the last with 1/2. Then by majority vote, we see that cluster 1 wins and the forecast for interval[t] will be derived from cluster/component 2 by looking at the range of interval[t] values associated with cluster 2 and taking an creating a statistic/estimator based off of that. Many other methods exist for forming forecasts off this mixture model due to the GMM originally being an unsupervised model, but this is just an example we have employed. With further work and time, we would have liked to explore other ways of transforming this unsupervised model into a supervised one and compare accuracy rates.

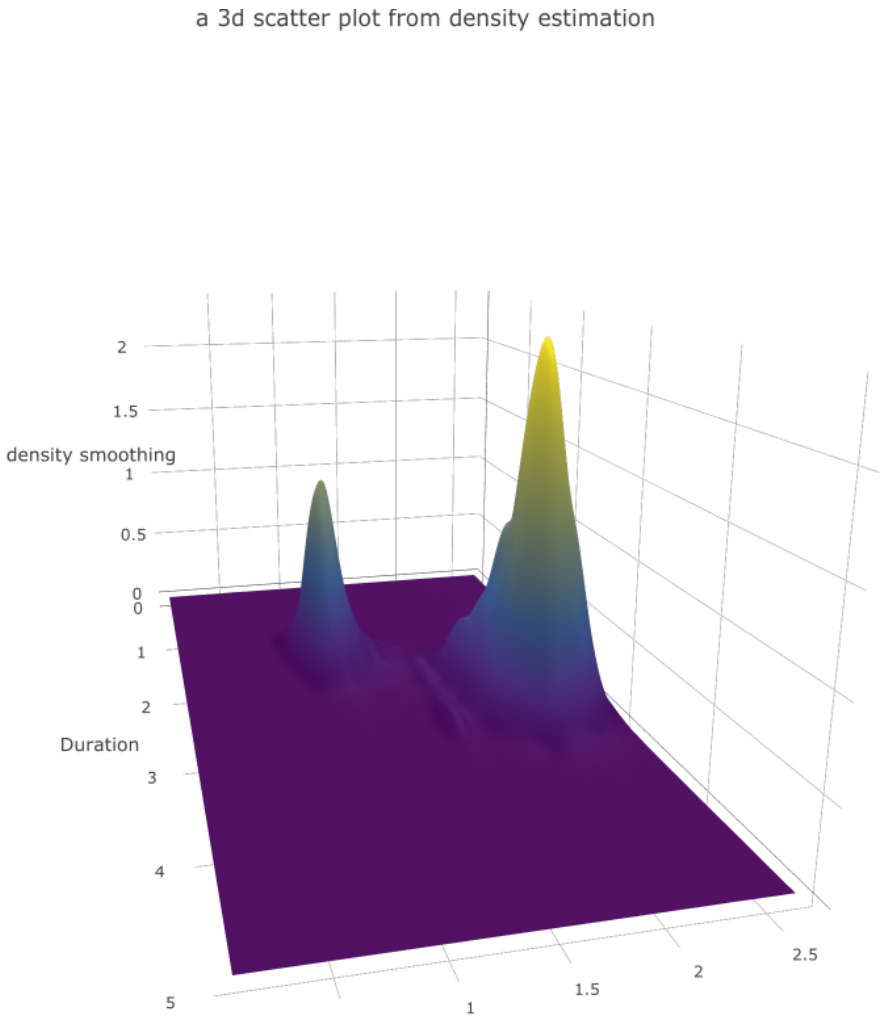
Appendix

Relationship by year




year 2005 1229 observations
year 2009 1582 observations
year 2006 1238 observations
year 2010 1642 observations
year 2007 1474 observations
year 2011 531 observations
year 2008 1517 observations
year 2012 321 observations
IntervalM
duration

A 3D density of the Interval and Duration varaibles(Using multivariate kernel density estimation)



Reference

"Geyserstudy." n.d. Accessed December 13, 2019. http://www.geyserstudy.org/ofvclogs.aspx.

Titterington, D., A. Smith, and U. Makov. 1985. *Statistical Analysis of Finite Mixture Distributions*. Wiley.