PHP2650: Statistical Learning and Big Data

Assignment 1 - Large Data in R

Antonella Basso

Febuary 19, 2022

Problem 1: MapReduce

Consider a set of files that each contains a column for a variable x that takes on integer values between some minimum and maximum value. The goal is to find the median of x across all files. Explain how you can find the median using a MapReduce approach. Under what circumstances would you consider this approach? Explain why.

Solution

To execute the desired task using a MapReduce approach, we would first have to map the specified columns in all files onto files that arrange the integer values of x in ascending order. That is, the map operation would assign temporary keys to these values to create several ordered lists. The reduce operation would then take this set of files and combine them such that they maintained their ordered structure (while preserving possible repeated values). The resulting file would thus contain an ordered list of integers with corresponding place values that could be used to assign the median of x to the middle value if there are an odd number of integers and to the mean of the two middle values if there are an even number of integers. Given that we seek to obtain a piece of information from a subset of the data which all files have in common, a MapReduce approach is the best and most efficient way of tackling this problem. Primarily, it allows us to extract and manipulate only the parts of the data that we need without having to download and combine every file. This approach saves us both the time and the memory it would take to gather a single piece of information from massive amounts of data with traditional programming. For this reason, establishing what we want to obtain from our data before loading it into memory can prove to be vastly cost effective in situations like this (namely, when a task is simple and we're dealing with multiple large files that share similar traits).

Problem 2: Managing Files Outside R

The file 'breaths.zip' in the Data folder on Canvas contains multiple files, each containing data for a single breath of an artificial lung on a ventilator. We will focus on the inspiratory phase of the breath when air is let into the lung. Use unix commands to combine the files into a single csv file containing only entries for which 'u_out' is equal to 0. Give the commands you used in your solution below and explain what each piece is doing. Explain why this approach may be preferable compared to loading all files into R and then filtering.

Variable	Description
"id"	Identifier unique to each observation.
"breath_id"	Identifier unique to each breath.
"R"	Lung attribute indicating how restricted the airway is (cmH2O/L/S).
"C"	Lung attribute indicating how compliant the lung is (mL/cmH2O).
"time_step"	Actual time stamp.
"u_in"	Control input for the inspiratory solenoid valve (0 to 100).
"u_out"	Control input for the expiratory solenoid valve (0 or 1).
"pressure"	Airway pressure measured in the respiratory circuit (cmH2O).

Solution

```
cat /Users/antonellabasso/Desktop/PHP2650/DATA/breaths/*.csv |
  awk -F\, '{if (NR==1 || $7==0) print $0;}' > breaths_ip.csv
head breaths_ip.csv
```

The ‘cat’ command used in the first line of this code concatenates all (*) csv files in the given directory (/Users/antonellabasso/Desktop/PHP2650/DATA/breaths/). That is, all files found in the ‘breaths’ unzipped folder within the specified path. The pipe command (|) then sends this output (the concatenated files) to the subsequent command ‘awk’, which pattern matches for given strings. Specifically, the conditional (if) statement in quotes defines the pattern to be searched for in each line of the input document (in this case, either that the line is the first row (NR==1), or (||) that the 7th column, “u_out”, is equal to 0 ($7==0)) and the action following it indicates what will be done when a match is found (in this case, that the entire line be printed (‘print $0;’). The ‘-F’ command used before this statement specifies the delimiter to look for which separates columns (here, it is a comma, so we write’-F,‘). Finally, the’>’ command sees the output (all selected lines) into a new csv file called ‘breaths_ip.csv’. We include a ‘head’ command followed by this new file name in the last line to give the first 10 rows/lines of the document and ensure that it matches the desired format.

Problem 3: Exploratory Data Analysis (EDA)

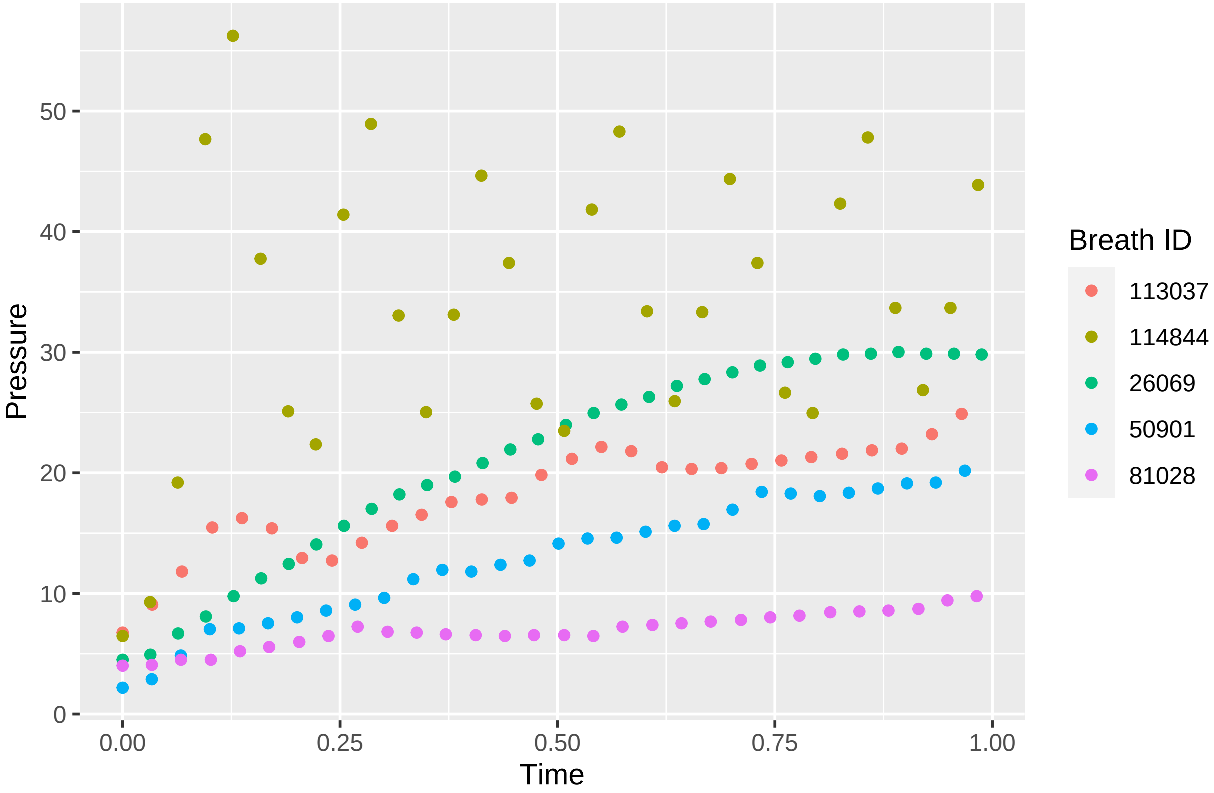
Using the csv file you created in Problem 2, conduct an exploratory analysis to understand the data distributions and relationships between different variables. You should first load your data into a spark table before conducting your analysis. Write up your findings in report format including appropriate tables and figures. As part of your analysis, you should:

a) Use the mclapply function to find the median time pressure peaks in a breath.
b) Plot the average breath’s pressure and air intake over time.
c) Repeat the plot above but for different values of R and C.

Solution

One of the main focuses of this EDA is to identify airway pressure changes in breaths over time. To get a glimpse of our data in this respect we select five breaths at random (by breath ID) and plot their individual air pressure trajectories over time. Figure 1 below shows us that air pressure generally tends to increase over the course of a single breath (in the inspiratory phase). However, it is evident that not all breaths have the same rate of air pressure change. And, that some pressure patterns may even be sporadic. This difference in rates can be observed below, as trajectories appear to branch out from a similar starting points. That is, while these breaths all begin with relatively similar airway pressures, they end with vastly different ones, hinting at a possible increase in air pressure variance with time.

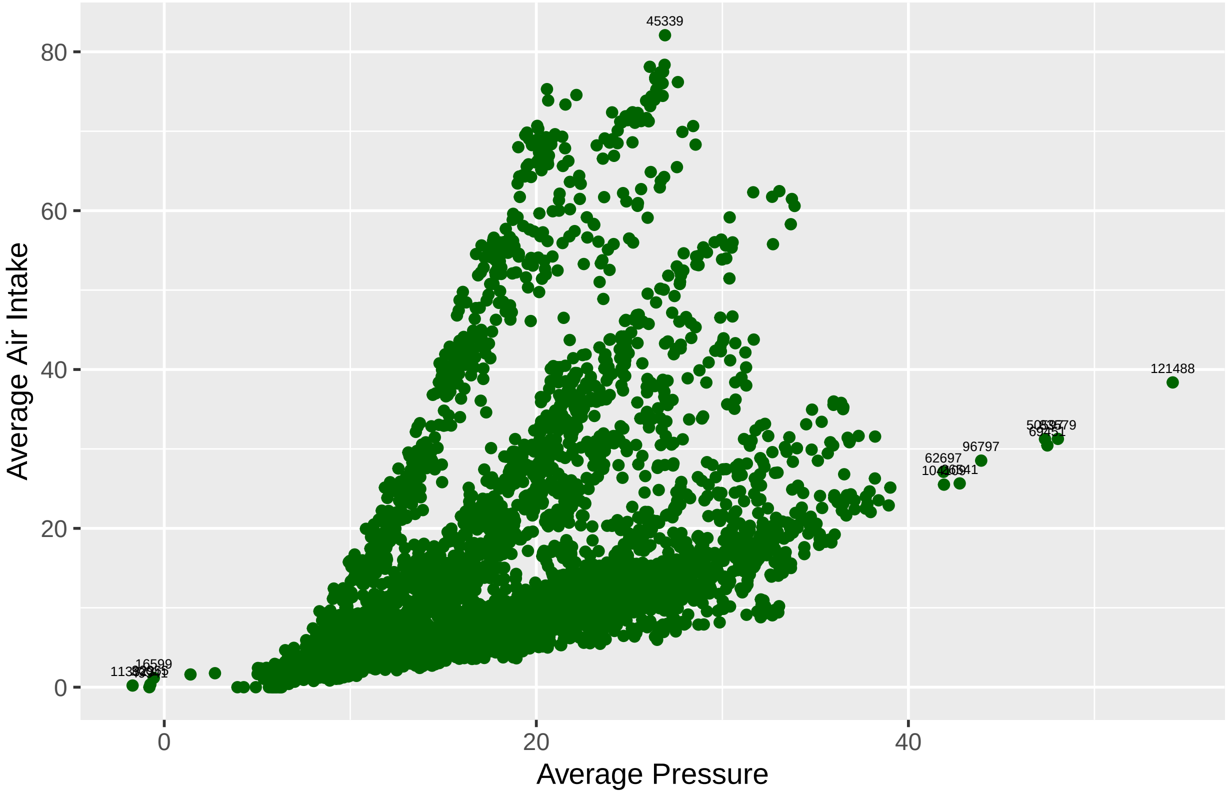
Figure 1: Breath Pressures Over Time



We might also be interested in observing pressure peaks across breaths. To get a sense for where we may see a pressure peak in a breath based on the data, we compute the median pressure peak time via parallel computing. That is, using the “mclapply” command from the “parallel” library, we apply a function to each breath ID that finds the time step corresponding to its maximum airway pressure value in the data. Using “mclapply”, this function is applied simultaneously to all 5,000 breath ID’s to increase efficiency, allowing us to then compute our desired statistic, from the resulting list of time steps, more quickly. Through this process, we find that the median pressure peak time is roughly 0.8282. Since time in our data is a continuous random variable ranging from 0 to 1 (for each breath), we may deduce that, out of all breaths recorded, the median airway pressure peak occurred around 83% of the way through the breath.

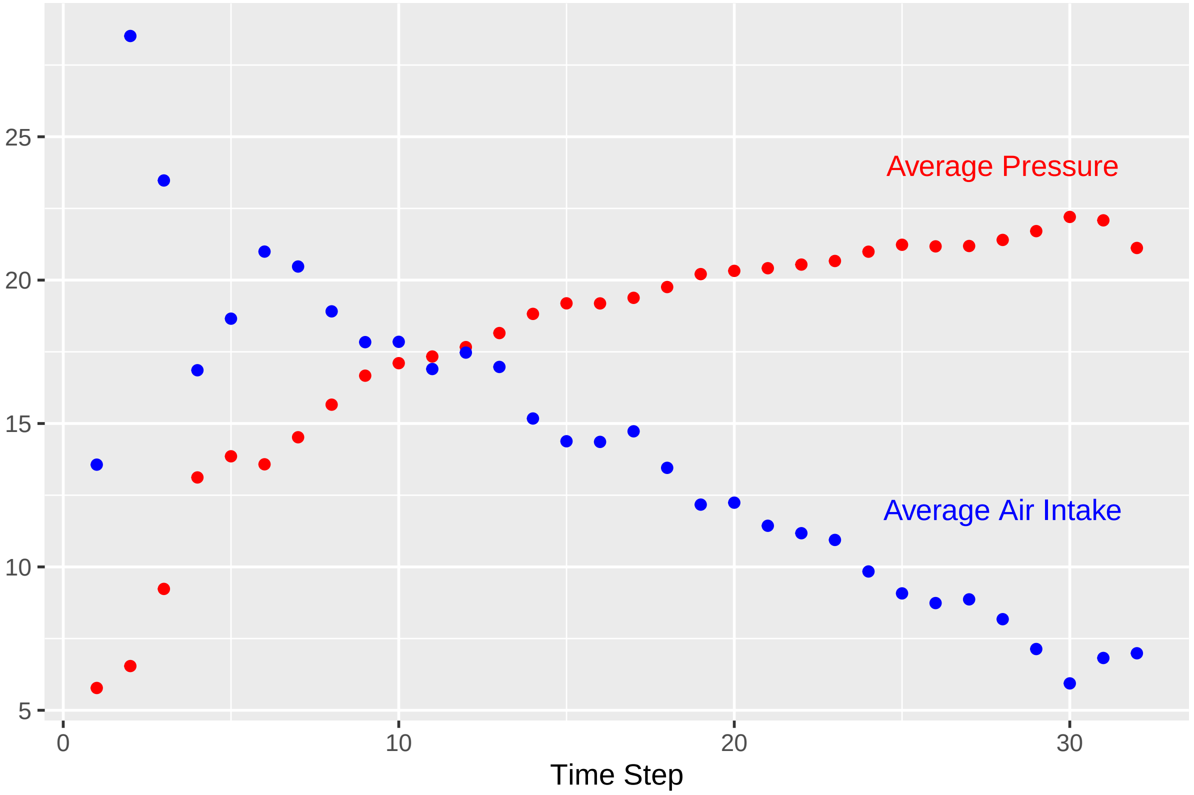
Moreover, given that the data we are working with involves the inspiratory breath phase, it may be of interest to analyze air intake ("u_in") patterns as well and observe its association with airway pressure and other variables such as levels of airway restriction ("R") and lung compliance ("C"). To first gain some intuition about the relationship between airway pressure and air intake, as well as their trends in the data, we construct a scatter plot of average airway pressures (x-axis) against average air intakes (y-axis) across all breaths. Figure 2 below specifically demonstrates the positive and perhaps exponential association between both variables. That is, it appears that large mean breath pressures tend to coincide with larger air intake values. Further, with the exception of some potential outliers (marked by their corresponding breath ID's), there appear to be clusters of breaths whose mean airway pressures and air intakes form different (seemingly) exponential trends. Whether or not this is truly so, it is clear from this graph that the data generally becomes more spread out with larger average values of air intake and pressure. Namely, we notice that, despite the positive trend, greater airway pressure averages have a much larger range of corresponding average air intake values.

Figure 2: Average Pressures vs. Average Air Intakes



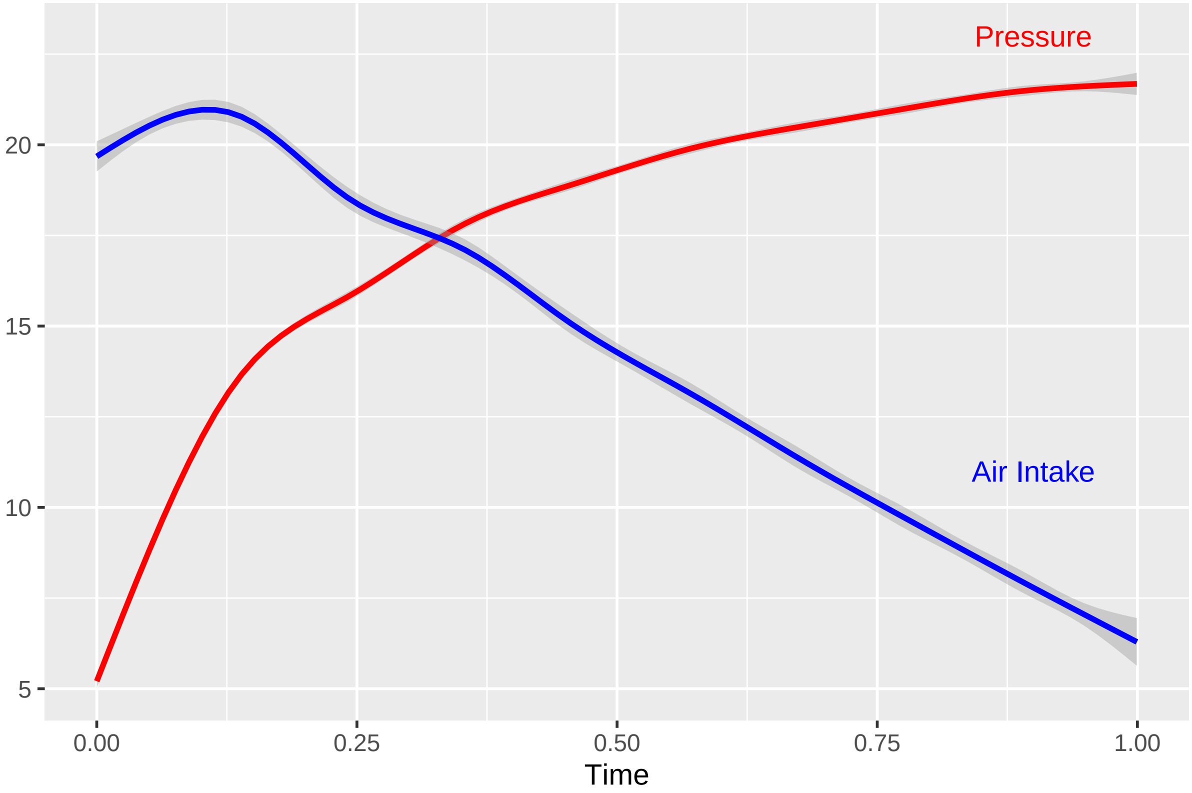
To gain some sense for how the two variables behave over time, we rank time steps across breath ID's, giving each a integers in ascending order starting at 1 (as they are similar but not exact across breath ID's), to compute mean breath airway pressure and air intake at each time step (over a whole breath). Figure 3 below depicts these averages over time, which have clearly opposite trends.That is, ignoring the air intake inconsistencies seen between time steps 0 and 5, this plot tells us that airway pressure increases with time, while air intake tends to decrease.

Figure 3: Average Breath's Pressure and Air Intake Over Time



These behaviors are also captured by the smooth plot (Figure 4) below, which gives us a more clear, yet generalized visual. We notice that airway pressure is always increasing with time, although the rate of increase becomes smaller with each time step (similar to a logarithmic function), which we saw in Figure 3 above. On the other hand, we notice that air intake, although perhaps more inconsistent at the start of the breath, decreases over time at a more constant rate. From these graphs, we may infer that over the course of a single breath at inspiratory phase, it is generally the case that less and less air is introduced into the lung with each time step, while pressure increases at a declining rate and remains relatively high toward the end of the breath.

Figure 4: Breath Pressure and Air Intake Over Time



Exploring these trends in airway pressure and air intake with respect to airway restriction ("R") and lung compliance ("C") levels also contributed to obtaining a more exhaustive understanding of their relationship and the overall data. Specifically, the two tables below give us the mean airway pressures and air intakes for each unique value of airway restriction and lung compliance level. Here we observe that as airway restriction increases, mean airway pressure either increases slightly or remains constant, and mean air intake decreases significantly. In contrast, as lung compliance increases, mean pressure either slightly decreases or remains constant, and mean air intake increases significantly.

R	Mean Pressure	Mean Air Intake
5	16.44499	20.887355
20	18.07512	15.768438
50	18.27938	8.592889

C	Mean Pressure	Mean Air Intake
10	18.56598	7.706806
20	17.86927	14.933700
50	16.36069	21.687386

To visualize these patterns, smooth plots like that shown in Figure 4 were created for different values of airway restriction and lung compliance. Particularly, Figures 5, 6 and 7 show changes in airway pressure and air intake over time for low (5), medium (20), and high (50) values of airway restriction, respectively. Similarly, Figures 8, 9 and 10 depict these changes for low (10), medium (20), and high (50) values of lung compliance, respectively. We notice in all plots, as with the means in the tables above, that airway pressure doesn't change significantly, while the rate (of change/decrease) of air intake decreases with growing airway restriction and increases with growing lung compliance. This implies that airway restriction and lung compliance are more highly correlated with air intake than with airway pressure. Moreover, we see that graphs for low airway restriction and high lung compliance appear similar, as do graphs for high airway restriction and low lung compliance, which hints at a correspondence between these values (supported by the values in the tables above). And, it would not be unreasonable to infer, independent of these results, that more restricted airways lead to less compliant lungs. However, whether airway restriction and lung compliance are actually associated with air intake, or even whether there exits a true relationship between airway pressure and air intake, are things that would require more extensive analysis to confirm.

Figure 5: Breath Pressure and Air Intake Over Time: R=5

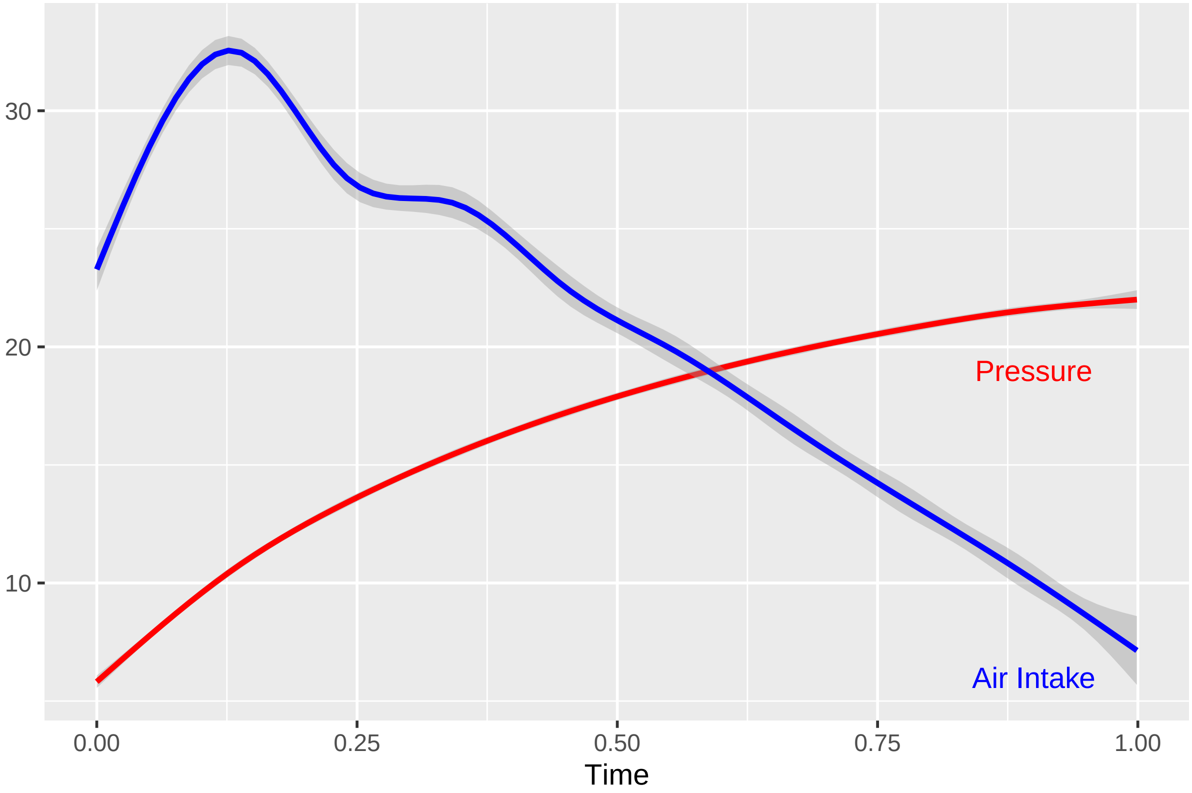


Figure 6: Breath Pressure and Air Intake Over Time: R=20

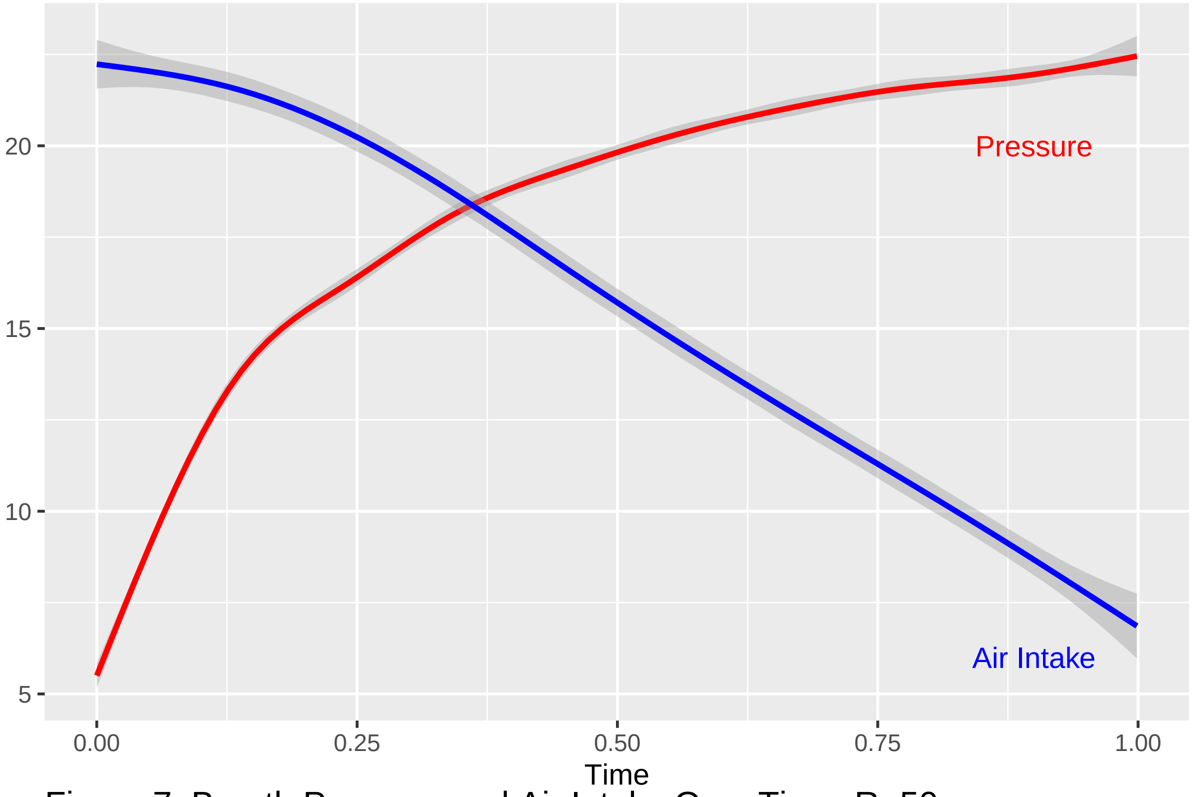


Figure 7: Breath Pressure and Air Intake Over Time: R=50

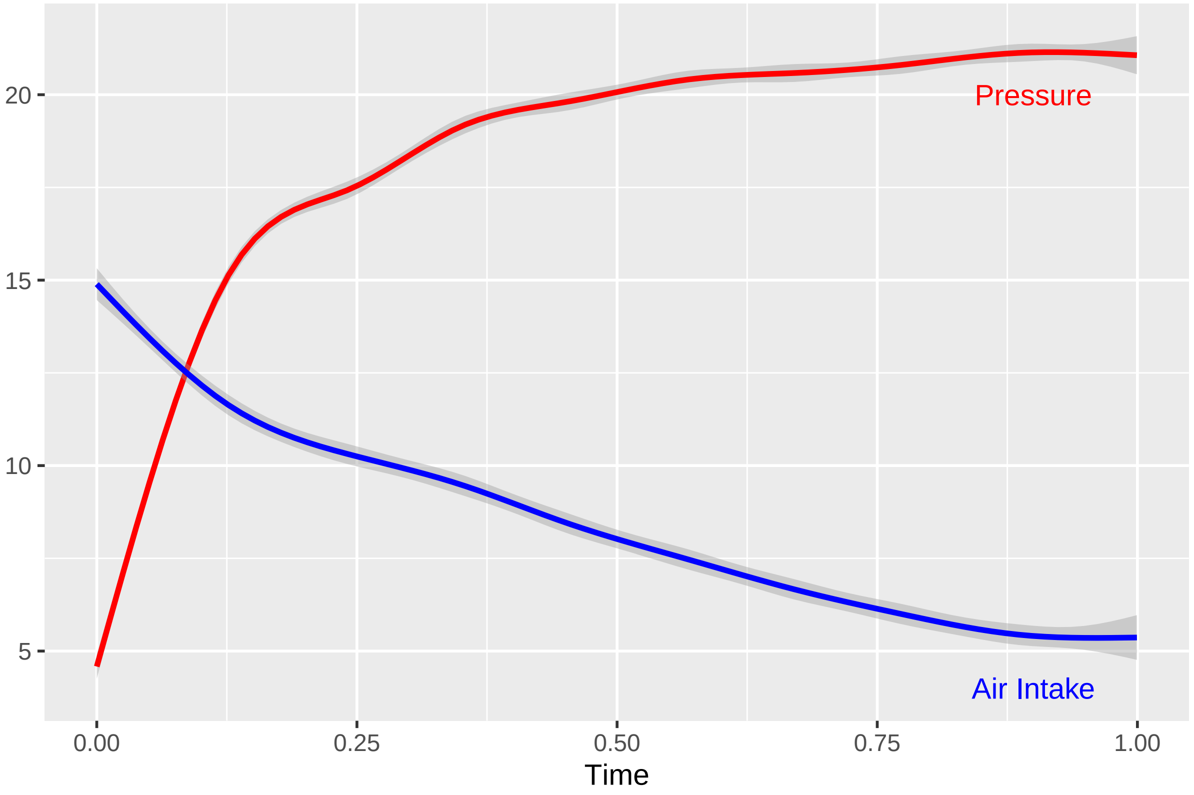


Figure 8: Breath Pressure and Air Intake Over Time: C=10

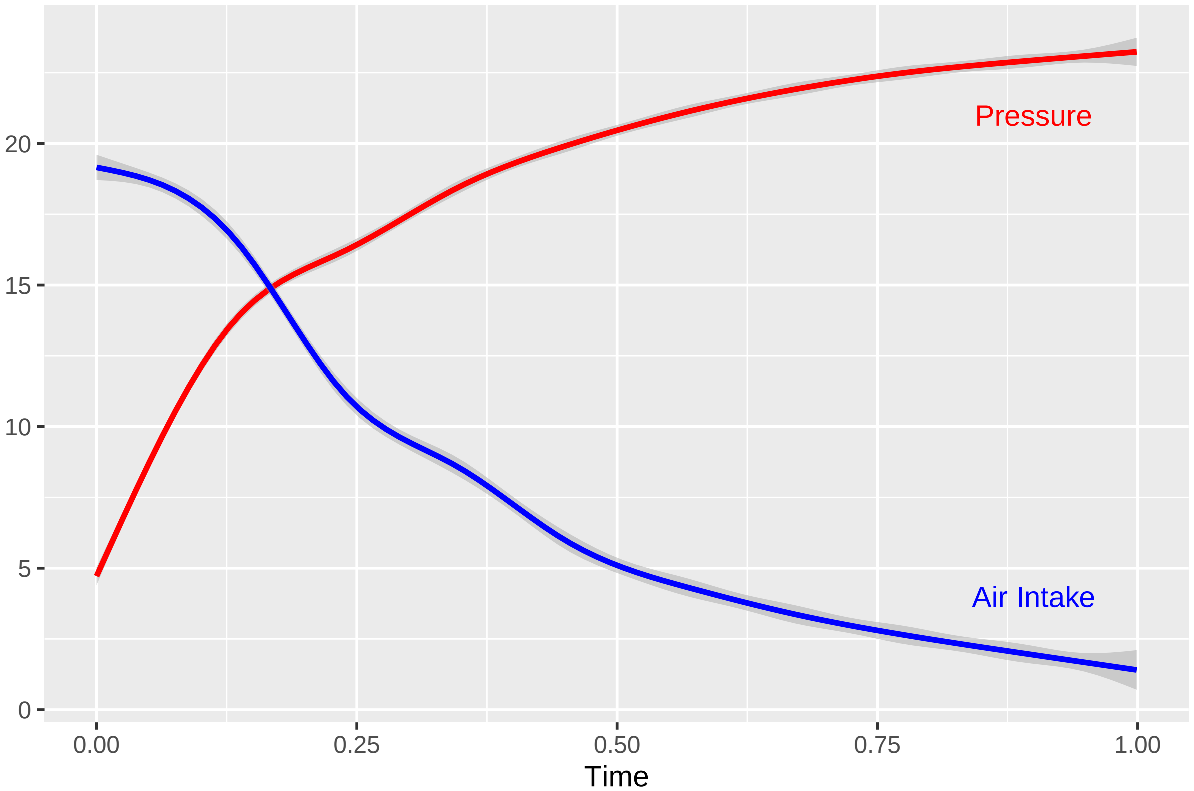


Figure 9: Breath Pressure and Air Intake Over Time: C=20

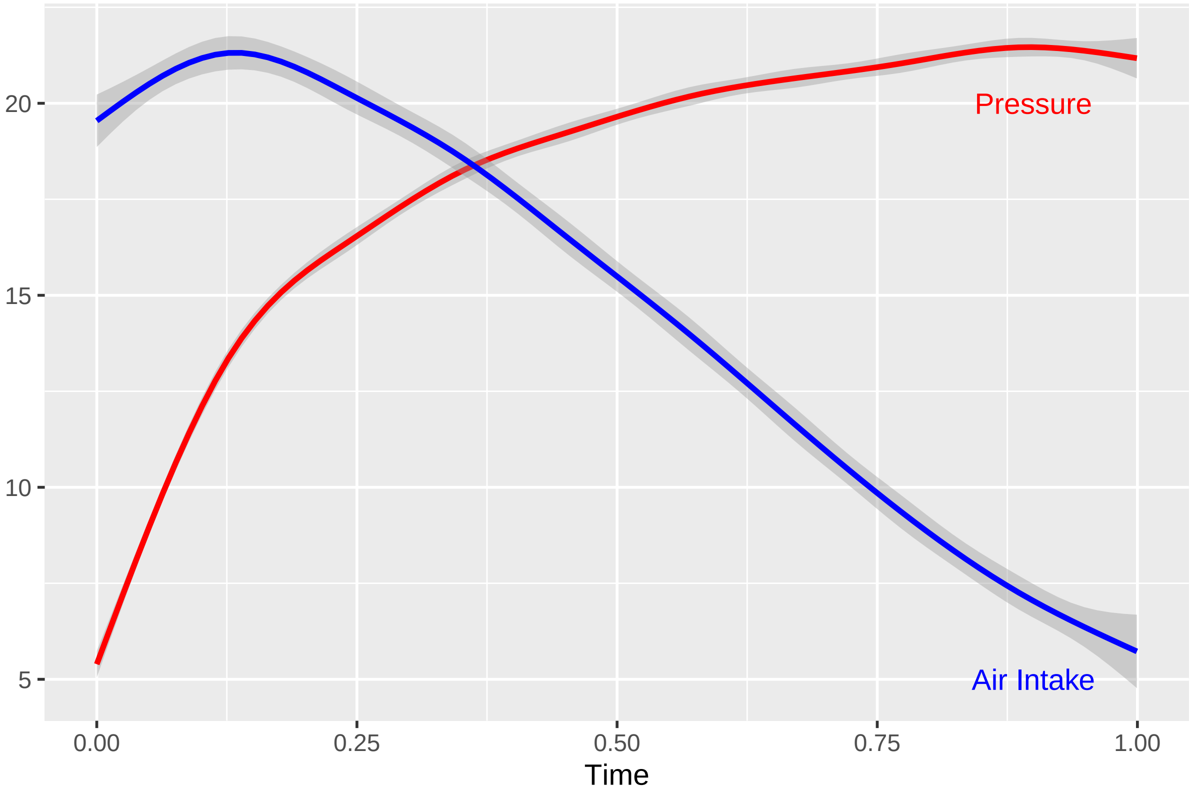
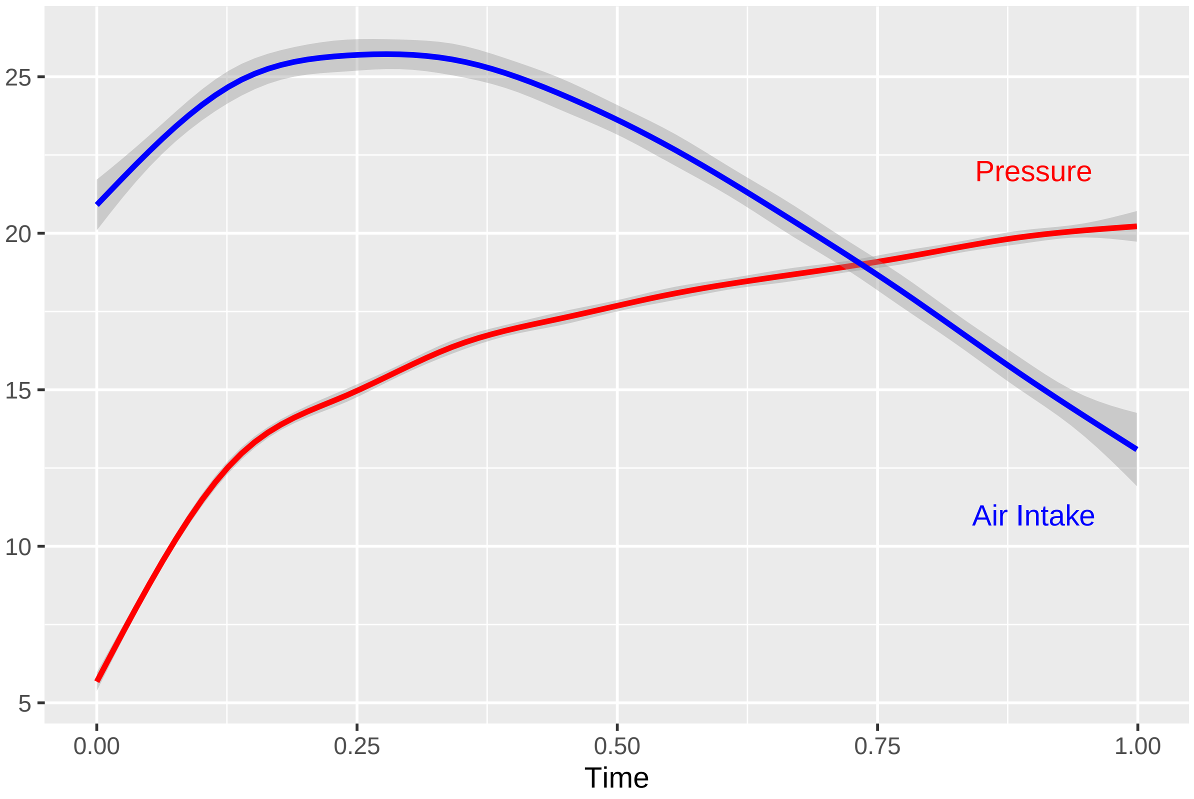


Figure 10: Breath Pressure and Air Intake Over Time: C=50



Code

```
#loading libraries
library(tidyverse)
library(dplyr)
library(sparklyr)

#reading data into spark
connect <- spark_connect(master="local")
file_path <- "/Users/antonellabasso/Desktop/PHP2650/DATA/breaths_ip.csv"
breaths <- spark_read_csv(connect, "breaths", file_path, header=T)
src_tbls(connect)

#seeing data layout
class(breaths)
colnames(breaths)
sdf_dim(breaths)
head(breaths)
```

```
#loading library for "mclapply"
library(parallel)

#loading new csv (breath_id, time_step, pressure) made with unix
breaths_press <- read.csv("/Users/antonellabasso/Desktop/PHP2650/DATA/breaths_press_time.csv")

#unique breath ids (5000)
breath_ids <- unique(breaths_press$breath_id)
```

```
#Figure 1 - Scatter Plot: Plotting Breath Pressure Over Time for 5 Random Breaths

#sampling 5 breaths
set.seed(4)
random_breaths <- sample(breath_ids, 5)

#function to get data for specific breath id
get_breaths <- function(i){
  rand_breaths <- breaths_press %>%
    filter(breath_id==i) %>%
  return(rand_breaths)
}

#applying function to each sampled breath id and merging data frames
random_breaths_df <- mclapply(random_breaths, get_breaths) %>%
  reduce(full_join, by = c("breath_id", "time_step", "pressure"))

#scatter plot
random_breaths_df %>%
  group_by(breath_id) %>%
  ggplot(aes(time_step, pressure, color=as.character(breath_id))) +
  geom_point() +
  labs(x="Time", y="Pressure", color="Breath ID") +
  ggtitle("Figure 1: Breath Pressures Over Time")
```

```
#Finding Median Pressure Peak Time with "mclapply" (parallel computing)
```

```r
#function to find time step of pressure peak given breath id
med_time <- function(i){
  x <- breath_ids[i]
  time <- breaths_press %>%
    filter(breath_id==x) %>%
    filter(pressure==max(pressure, na.rm=TRUE)) %>%
    select(time_step) %>%
    pull()
  return(time)
}

#applying function to each breath id
times <- mclapply(1:5000, med_time)
median(unlist(times)) #getting median time step
```

```r
#Figure 2 - Scatter Plot: Average Pressure and Air Intake by Breath

breaths %>%
  group_by(breath_id) %>%
  summarize(mean_pressure=mean(pressure), mean_u_in=mean(u_in)) %>%
  ggplot(aes(mean_pressure, mean_u_in, label=breath_id)) +
  geom_point(color="darkgreen") +
  geom_text(aes(label=ifelse(mean_pressure>40|mean_pressure<0|mean_u_in>80,
                             as.character(breath_id),
                             '')),
            hjust=0.5, vjust=-1, size=1.75) +
  labs(x="Average Pressure", y="Average Air Intake") +
  ggtitle("Figure 2: Average Pressures vs. Average Air Intakes")
```

```r
#Figure 3 - Scatter Plot: Average Breath's Pressure and Air Intake Over Time

#computing average breath's pressure and air intake at each time step
#ranking time steps (as they are similar but not exact across breath ids)
time_rank <- breaths %>%
  group_by(breath_id) %>%
  mutate(rank = rank(time_step))  %>%
  select(breath_id, u_in, pressure, rank)

#taking averages over each time step
breaths_avgs <- time_rank %>%
  group_by(rank) %>%
  summarize(mean_pressure=mean(pressure), mean_u_in=mean(u_in)) %>%
  arrange(-desc(rank))

#scatter plot
breaths_avgs %>%
  ggplot() +
  geom_point(aes(rank, mean_pressure), color="red") +
  geom_point(aes(rank, mean_u_in), color="blue") +
  labs(x="Time Step", y="") +
  annotate(geom="text", x=28, y=24, label="Average Pressure", color="red") +
  annotate(geom="text", x=28, y=12, label="Average Air Intake", color="blue") +
  ggtitle("Figure 3: Average Breath's Pressure and Air Intake Over Time")
```

```
#Figure 4 - Smooth Curve/Plot: Breath Pressure and Air Intake Over Time
#gives us a similar visual of trends as the previous graph

breaths %>%
  ggplot() +
  geom_smooth(aes(time_step, pressure), color="red",
              method="gam", formula=y~s(x, bs="cs")) +
  geom_smooth(aes(time_step, u_in), color="blue",
              method="gam", formula=y~s(x, bs="cs")) +
  labs(x="Time", y="") +
  annotate(geom="text", x=0.9, y=23, label="Pressure", color="red") +
  annotate(geom="text", x=0.9, y=11, label="Air Intake", color="blue") +
  ggtitle("Figure 4: Breath Pressure and Air Intake Over Time")
```

```
#Tables 1 & 2: Average Pressure and Air Intake Given Airway Restriction (R) and Lung Compliance (C)

#table 1
table_1 <- breaths %>%
  group_by(R) %>%
  summarize("Mean Pressure"=mean(pressure), "Mean Air Intake"=mean(u_in)) %>%
  arrange(-desc(R))

#table 2
table_2 <- breaths %>%
  group_by(C) %>%
  summarize("Mean Pressure"=mean(pressure), "Mean Air Intake"=mean(u_in)) %>%
  arrange(-desc(C))

knitr::kable(table_1, format="markdown")
knitr::kable(table_2, format="markdown")
```

```
#Figures 5, 6, & 7 - Smooth Curve/Plot: Breath Pressure and Air Intake by Airway Restriction (R)

#Low R
breaths %>%
  filter(R==5) %>%
  ggplot() +
  geom_smooth(aes(time_step, pressure), color="red",
              method="gam", formula=y~s(x, bs="cs")) +
  geom_smooth(aes(time_step, u_in), color="blue",
              method="gam", formula=y~s(x, bs="cs")) +
  labs(x="Time", y="") +
  annotate(geom="text", x=0.9, y=19, label="Pressure", color="red") +
  annotate(geom="text", x=0.9, y=6, label="Air Intake", color="blue") +
  ggtitle("Figure 5: Breath Pressure and Air Intake Over Time: R=5")

#Med R
breaths %>%
  filter(R==20) %>%
  ggplot() +
  geom_smooth(aes(time_step, pressure), color="red",
              method="gam", formula=y~s(x, bs="cs")) +
  geom_smooth(aes(time_step, u_in), color="blue",
              method="gam", formula=y~s(x, bs="cs")) +
  labs(x="Time", y="") +
  annotate(geom="text", x=0.9, y=20, label="Pressure", color="red") +
  annotate(geom="text", x=0.9, y=6, label="Air Intake", color="blue") +
  ggtitle("Figure 6: Breath Pressure and Air Intake Over Time: R=20")

#High R
breaths %>%
  filter(R==50) %>%
  ggplot() +
  geom_smooth(aes(time_step, pressure), color="red",
              method="gam", formula=y~s(x, bs="cs")) +
  geom_smooth(aes(time_step, u_in), color="blue",
              method="gam", formula=y~s(x, bs="cs")) +
  labs(x="Time", y="") +
  annotate(geom="text", x=0.9, y=20, label="Pressure", color="red") +
  annotate(geom="text", x=0.9, y=4, label="Air Intake", color="blue") +
  ggtitle("Figure 7: Breath Pressure and Air Intake Over Time: R=50")
```

```
#Figures 8, 9, & 10 - Smooth Curve/Plot: Breath Pressure and Air Intake by Lung Compliance (C)

#Low C
breaths %>%
  filter(C==10) %>%
  ggplot() +
  geom_smooth(aes(time_step, pressure), color="red",
              method="gam", formula=y~s(x, bs="cs")) +
  geom_smooth(aes(time_step, u_in), color="blue",
              method="gam", formula=y~s(x, bs="cs")) +
  labs(x="Time", y="") +
  annotate(geom="text", x=0.9, y=21, label="Pressure", color="red") +
  annotate(geom="text", x=0.9, y=4, label="Air Intake", color="blue") +
  ggtitle("Figure 8: Breath Pressure and Air Intake Over Time: C=10")

#Med C
breaths %>%
  filter(C==20) %>%
  ggplot() +
  geom_smooth(aes(time_step, pressure), color="red",
              method="gam", formula=y~s(x, bs="cs")) +
  geom_smooth(aes(time_step, u_in), color="blue",
              method="gam", formula=y~s(x, bs="cs")) +
  labs(x="Time", y="") +
  annotate(geom="text", x=0.9, y=20, label="Pressure", color="red") +
  annotate(geom="text", x=0.9, y=5, label="Air Intake", color="blue") +
  ggtitle("Figure 9: Breath Pressure and Air Intake Over Time: C=20")

#High C
breaths %>%
  filter(C==50) %>%
  ggplot() +
  geom_smooth(aes(time_step, pressure), color="red",
              method="gam", formula=y~s(x, bs="cs")) +
  geom_smooth(aes(time_step, u_in), color="blue",
              method="gam", formula=y~s(x, bs="cs")) +
```

```r
  labs(x="Time", y="") +
  annotate(geom="text", x=0.9, y=22, label="Pressure", color="red") +
  annotate(geom="text", x=0.9, y=11, label="Air Intake", color="blue") +
  ggtitle("Figure 10: Breath Pressure and Air Intake Over Time: C=50")
```