4-2018

Probabilistic Modeling of Student Interactions during a Passing Period at the University of Dayton

Allyson Pacifico

Probabilistic Modeling of Student Interactions during a Passing Period at the University of Dayton

Honors Thesis

Allyson Pacifico

Department: Mathematics

Advisor: Peter Hovey, Ph.D.

April 2018

Probabilistic Modeling of Student Interactions during a Passing Period at the University of Dayton

Honors Thesis

Allyson Pacifico

Department: Mathematics

Advisor: Peter Hovey, Associate Professor Mathematics

April 2018

Abstract

The University of Dayton is composed of five colleges and schools: College of Arts and Sciences, School of Law, School of Business Administration, School of Education and Health Sciences, and School of Engineering. The University of Dayton is composed of about 11,000 students on campus who all have distinct class schedules and paths they take between their classes. In this study, I wanted to know the probability of meeting my friends with a different class schedule as I walk between classes. The data consisted of one to two students from each college, except for the School of Law, who documented their paths on a modified campus map for a week. Using R, the simulation randomly selects two paths from the data, generates a random time between each node, and compares the time of the identical nodes to see if the students were to have met.

Dedication or Acknowledgements

Dr. Hovey was my thesis advisor. Dr. Chen helped me clean up my R code. The Undergraduate Honors Thesis Research Grant Committee for their grant to help me conduct my research and Vicki Winthrow for helping me during the reimbursement process.

UNIVERSITY of DAYTON

Table of Contents

Introduction

College is a great place to build both hard and soft skills through academic classes and experiential learning opportunities outside of the classroom. Nested in Dayton, Ohio, the University of Dayton shares the spirit of the local community with its focus on ingenuity and innovation. The Univeresity of Dayton, henceforth referred to as UD, builds upon these skills as a top-tier Catholic research university "committed to educating the whole person and linking learning and scholarship with leadership and service" (website). UD encourages its members to recognize their individual talents, to employ their skills to meet human needs, and to collaborate in building community. This 373-acre estate includes on-campus university housing ranging from dorms to houses, academic buildings, and research centers.

UD has a strong sense of community within the institution and the students themselves. Approximately 90 percent of UD's undergraduates living on campus (https://www.udayton.edu/studev/housing/), students are constantly interacting with each other inside and outside of their academic classes. The University For The Common Good enrolled 8,096 undergraduates as of Fall 2017 (UD Factbooks). The students can be enrolled in at least one of the four college or school at UD: College of Arts and Sciences, School of Business Administration, School of Education and Health Sciences, and School Engineering. Table 1 is from the official 2017 UD Fact Book which outlines the percent of students enrolled in each college or school.

	Full-time Undergraduate Students
Total Enrollment	**8,096**
College of Arts and Sciences (CAS)	39.22%
School of Business Administration (SBA)	24.85%
School of Education and Health Sciences (SEHS)	11.86%
School of Engineering (SE)	23.69%

Table 1. Outline of Undergraduate Students, UD Factbooks 2017

On weekdays, students flood the main campus to attend their classes, communicate with their peers, go to the dining hall for food, or find a spot on campus to do homework throughout the day. This research focuses particularly on the interactions that occur during passing periods, which are defined as the fifteen minute blocks in between standard class sessions. The research was restricted to University of Dayton's main academic campus for simplification. It was assumed that students tend to follow the most direct set routes on campus based on their schedules rather than selecting random routes each time. After collecting data on the student's paths, R was used to analyze two paths randomly selected from a pool of fixed paths and to model time progressions between nodes via gamma distribution. This research reveals interesting information regarding the probability of meeting different college undergraduates by running a simple simulation model based on real world data.

Definition of Terms

Passing Period: 15 minute block between each scheduled class session

Node: position on campus which were arbitrarily chosen via observation

Node Pair: any two consecutive nodes within a passing period

Path: a path p_i consists of a sequence of nodes n_1, n_2,...n_k

Time progression: elapsed time between nodes $t(p_i) = t_1 = 0, t_2, t_3, \ldots t_k$ (in seconds)

Interaction: when two paths p_1 & p_2 have a common n_k and $t_{nk}(p_1)$ is within +/- 10 seconds of the elapsed time at $t_{nk}(p_2)$

Methodology

My research required gathering data from University of Dayton Undergraduate students in order to test the assumption and calculate the probabilities. Data was collected using Google Sheets to document and centralize information.

Survey Development & Data Collection

To start, a map of the main campus was created and nodes were identified that represented high density areas on campus such as entry/exit points to campus, academic buildings, and major intersections. The buildings monitored in the study were Kettering Labs (KL), Humanities(HM), St. Joseph's Hall (SJ), Zheler's Hall(ZH), Science Center (SC), Anderson (AN), and Miriam Hall (MH). The first version of the map was used for the pilot study. It was conveniently given to ten friends along with an informal discussion on how to document use the map and a standard clock to document their time progression for each path for one week. Figure 1 shows the diagram of the main campus and form for each student to fill out.

Table 1. Pictorial representation of UD's main academic class room

After collecting feedback from the pilot study, there were a multitude of errors on the map that needed to be fixed such as missing or duplicate nodes. The collected data also lacked accurate results as it did not represent all schools and student's documentation lacked consistency and accuracy. This meant two things: (1) the map was not clear (2) standard clock was not accurate enough to identify an interaction.

A revised map and table were developed for students to provide better user experience. In the second phase of data collection, two students were randomly selected from each college to participate in the compensated research. Students were required to attend a formal information session that included the purpose of the research. Students were also informed on where and how to document their paths to ensure consistency within the centralized data. Documentation included training students on which nodes to include in their path, starting and ending their paths with either an entry or exit to campus or an entry or exit into a building, and indicating their arrival at the first node to be 0 (i.e. $t_1(p) = 0$). Lastly, students practiced using a stopwatch that has minutes, seconds, and

milliseconds to record their time progressions between each node to provide accurate measures of elapsed time. A total of 98 paths were collected from the subjects. Table 2 shows an excerpt of the centralized Google Sheet where students documented their information. Node_id indicates the order of the nodes within a path.

path_id	**node_id**	**node_name**	**elapsed_time**	**college**
...	...	...	...	...
3	3	M3	00:27.81	sehs
3	4	B29	00:32.54	sehs
4	1	B29	0:00.00	sehs
4	2	M2	00:04.21	sehs
4	3	M9	00:11.14	sehs
4	4	B19	00:18.27	sehs
5	1	B19	0:00.00	sehs
...	...	...	...	...

Table 2. Second data collection method

Data Analysis

It was crucial to transform elapsed time for all 399 nodes into one unit (seconds) to maintain consistency within the analysis. The data includes a minimum of 4.2 seconds between two nodes, specifically B29 to M2, and a maximum time of 676.4 seconds or roughly 11 minutes between two nodes, namely B29 to C12. This elongated time progression could have been caused by the student stopping to chat with a friend along the path or even getting food on the way to his or her own next class. The average time it took to get from one node to the next is just under a minute at 49.40 seconds. Initially, the interaction window was set to be +/- two minutes but it is obvious that the descriptive statistics suggest modifying it. Figure 2 shows the distribution of the time progressions and identifies the shape of the data to use the right indicators when reducing the interaction window.

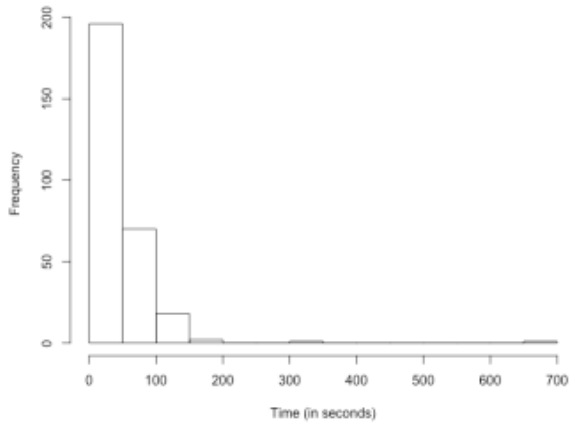

Figure 2. Histogram of Time Progressions for Every Node Except Starting Nodes

Figure 2 identifies that the data is skewed to the right, or positively skewed. Therefore, the mean would best describe the central tendency of the data. Given that the average time was about 49.5 seconds, the interaction window was reduced to +/- 10

seconds to give adequate enough time to for an interaction to occur at a designated node but also allow time for travel to the next node within the time progression.

Figure 3 shows a histogram of the data without the upper outliers to focus on the overall trend of the data. The descriptive statistics did not vary too much with the average seconds between two nodes without outliers being at 41.79 seconds. Using Figure 3, the visual trend in the data suggested the gamma distribution to fit the overall data.

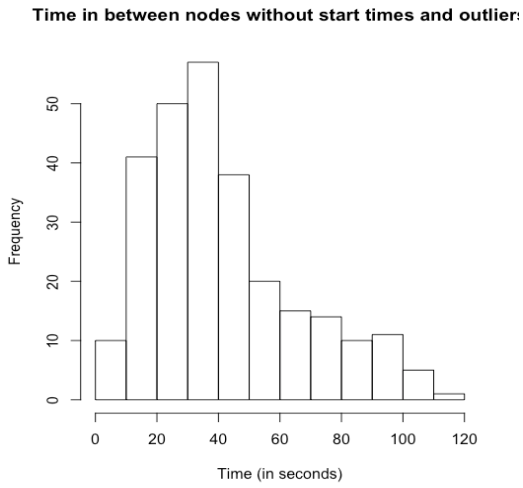

Figure 3. Histogram of Time Progressions for Every Node Except Starting Nodes and Outliers

The researchers found that individual node pairs lacked sufficient data for themselves. That is, if a random time was generated using the entire dataset, node pairs that were farther apart in time and distance would impact node pairs that were closer together and vise versa. To address this issue, each node pair was filtered by time which also served as an indicator of distance. Identical node pairs and node pairs with similar time progressions were grouped together. Both the average and standard deviations were obtained from each group to ultimately find the gamma function for each group. Unique

gamma functions for each group improves the accuracy of a randomly generated time progression for a node pair.

A simulation was created using R to test the data 100,000 times. First, two paths were selected with replacement. For each node in the path, a random time progression between each node was generated using the gamma distribution of the assigned group. The simulation then updates the elapsed time as $t_k = t_k + t_{k-1}$. Lastly, it would compare the time progression for common nodes to see if the students on the chosen paths would be considered to have an interaction. This process would repeat 100,000 times.

Paths obtained from the same college are half as likely to be selected for the simulation as to paths from different colleges. A second simulation was conducted for two paths obtained from the same college. This ensured the final results would be comparable between each combination of schools.

Results

The results obtained after running the simulation for 1000,000 times and running same colleges pairs an extra 5,500 times are displayed in Table 3.

College 1	College 2	Interaction	Total	P(Interaction | K = 100,000 & k = 5,500)
CAS	CAS	3,757	10,468	35.9%
CAS	SBA	1,060	11,891	8.9%
CAS	SE	1,735	12,427	14.0%
CAS	SEHA	633	10,629	6.0%
SBA	SBA	5,867	12,467	47.1%
SBA	SE	1,633	14,644	11.2%
SBA	SEHS	400	12,511	3.2%
SE	SE	1595	13,098	12.2%
SE	SEHS	872	12,823	6.8%
SEHS	SEHS	1,907	11,042	17.3%

Table 3. Result output from R

For selected paths from SBA and SEHS are the least likely to interact during a passing period at 3.2% of the time. Paths selected from SBA and SBA have the highest likelihood for interaction at 47.1% of the time. Paths chosen from same colleges have a significantly higher likelihood of having an interaction. Overall, the average interaction happens 16.2% of the time.

It is important to note the limitations of this study. Since there was an assumption that a student was most likely going to take the same path to their classes, it was given that there would be repeated paths within the data. Although this was helpful in finding the average time progression between a pair of nodes within a path. Another limitation was that four first years and two seniors were involved in the study. A traditional progression of academics at UD show that upperclassmen are more likely to take specialized courses rather than first years. Thus, upperclassmen are more likely to stay within their respective academic building rather than attend classes in other academic buildings to satisfy general education classes. SEHS paths were underrepresenting since their main academic building was not on the main campus and therefore not on the map. As mentioned in the last section, paths obtained from the college have a significantly smaller chance of being selected for the simulation as to paths from different colleges. Although a second simulation was created to address this issue, the results lack reliable verdict for paths chosen from the same college.

Conclusion

To recap, interactions occured an average of about 16% of the time. SBA students interact with other SBA students almost half of the time and SBA and SEHS students will rarely meet. Miriam Hall, where SBA classes are predominantly in, is the most consolidated academic building which could affect interaction. Conversely, the academic buildings and course load for students in SBA and SEHS are vastly different which supports the lack of interaction between them. The most prominent finding was that students are more likely to meet students within their college. This could be due to the fact that most colleges have their own designated building therefore students within the same college are likely to meet during their walk to class or are even in the same class. The limitations of this study could have also influenced this finding since the same exact path could have been chosen twice multiple times.

It is advised for future studies to use a mobile application to automatically collect location and elapsed time to enhance convenience and reduce time. If someone were to adopt this as their research in the future, it would be beneficial to collect a wider variation of paths to get a better representation of the student population during passing periods. Creating a mobile application would make documentation easier for the subjects, more accurate, and easier for the researcher when centralizing the data. One could also test if upperclassmen were more likely to interact with other upperclassmen due to higher level and specialized courses in specific academic buildings in comparison to first year students who are taking general education courses all over campus. One could also consider the impact of the interaction time window (20, 10, or 5 seconds) to allow for

more variability of interaction. Lastly, to add another math component, the Markov Chain approach could be used to build transitional probabilities between nodes. That is, the simulation builds a random path based on the probabilities of the next node.

The findings in this study and future findings could go further as to suggest new methods for increasing interaction between students from different colleges such as urban planning and transdisciplinary academic requirements. The implications overall create an open and welcoming space for students to connect and build relationships with one another outside of their major. As colleges and universities across the nation move towards a holistic and integrative education, it is beneficial to consider how everyday student interactions on campus build soft skills, extend professional networks, and boost university morale.

Appendix

R Code for Descriptive Statistics

```
# import fall 2017 path data
pathdf <- read.csv("Documents/Thesis/fa2017data.csv")

#create summary of pathdf
summary(pathdf, maxsum = 50)

#change names of columns
colnames(pathdf) <-
 c("path_id","node_id","node_name","time","time2",
                       "college","time_between","elapsed_time")

#check class of columns.
#(integer) path_id, node_id (numeric) elapsed time
#(factor) node_name, time, time2, college, time_between
sapply(pathdf, class)

#create int between for time_between (in seconds)
#    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#  0.00    0.00   26.99   37.22   49.08  676.41
pathdf$time_between <- as.numeric(pathdf$time_between)
print(pathdf$time_between)
summary(pathdf$time_between)
hist(pathdf$time_between, main="Time in between nodes for all
 subjects with start times",
     xlab= "Time (in seconds)", las=1)

#take out values that are 0:00 which indicate the start of a path
# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#  4.20   23.54   37.19   49.40   60.20  676.41
between <- pathdf$time_between[pathdf$time_between != 0.00]
print(head(between))
summary(between)
hist(between, main="Time in between nodes for all subjects
 without start times",
     xlab= "Time (in seconds)")

# histogram for each
hist(pathdf$time_between,data = pathdf$path_id, main="Time in
 between nodes for all subjects",
     ylab="Frequency", xlab="Time (seconds)")
hist(pathdf$time_between[pathdf$college == "cas"],data =
 pathdf$path_id, main="Time in between nodes for CAS subjects",
     ylab="Frequency", xlab="Time (seconds)")
hist(pathdf$time_between[pathdf$college == "se"],data =
```

```
 pathdf$path_id, main="Time in between nodes for SE subjects",
     ylab="Frequency", xlab="Time (seconds)")
hist(pathdf$time_between[pathdf$college == "sehs"],data =
 pathdf$path_id, main="Time in between nodes for SEHS subjects",
     ylab="Frequency", xlab="Time (seconds)")
hist(pathdf$time_between[pathdf$college == "sba"],data =
 pathdf$path_id, main="Time in between nodes for SBA subjects",
        ylab="Frequency", xlab="Time (seconds)")

#take out outliers for the sake of making it look nicer
remove_outliers <- function(x, na.rm = TRUE, ...) {
  qnt <- quantile(x, probs=c(.25, .75), na.rm = na.rm, ...)
  H <- 1.5 * IQR(x, na.rm = na.rm)
  y <- x
  y[x < (qnt[1] - H)] <- NA
  y[x > (qnt[2] + H)] <- NA
  y
}
between2 <-
 remove_outliers(between)[!is.na(remove_outliers(between))]
summary(between2)
hist(between2, main="Time in between nodes without start times
 and outliers",
     xlab= "Time (in seconds)")

#get gamma function of between (without start times) times
xbar <- mean(between)
print(xbar) #xbar = 49.40 =
variance <- var(between)
print(variance) #variance = 2567.653
sd <- sd(between)
print(sd) #sd = 50.67201

beta <- (variance/xbar)
print(beta) #beta =51.97418 = scale
alpha <- (xbar^2)/variance
print(alpha) #alpha =  0.9505194 = shape

#get gamma function of between2 (without outliers, without start)
 times
xbar <- mean(between2)
print(xbar) #xbar = 41.79309
variance <- var(between2)
print(variance) #variance = 587.4156
sd <- sd(between2)
print(sd) #sd = 24.23666

beta <- (variance/xbar)
print(beta) #beta =14.05533 = scale
alpha <- (xbar^2)/variance
```

```
print(alpha) #alpha =  2.973469 = shape


# create gamma function, saves random gamma values (times in
 seconds) in y
# create probability function for gamma function
# print out probabilities (probability of times)
y=rgamma(15, alpha, scale = beta)
print(y)
Y <- function(y)  {y^(alpha-1) * exp(-y/beta)  /  ((beta^alpha) *
 factorial(alpha-1))}
print(Y(y))
```

R Code for Simulation

```
# ---------------------------------------------
# Obtain data
# ---------------------------------------------
pathdf <- read.csv("Documents/Thesis/fa2017data.csv")[,1:8]
colnames(pathdf) <-
 c("path_id","node_id","node_name","time","time2",
                      "college","time_between","elapsed_time")

# ---------------------------------------------
# Obtain gamma function variables for times
# ---------------------------------------------

#get means and std for each group
#then get gamma distr
groups <-read.csv("Documents/Thesis/groups.csv")[,1:3]
groupmean <-c(mean(groups$time[groups$group == "1"]),
 mean(groups$time[groups$group == "2"]),
              mean(groups$time[groups$group == "3"]),
 mean(groups$time[groups$group == "4"]),
              mean(groups$time[groups$group ==
 "5"]),mean(groups$time[groups$group == "6"]),
              mean(groups$time[groups$group ==
 "7"]),mean(groups$time[groups$group == "8"]),
              mean(groups$time[groups$group ==
 "9"]),mean(groups$time[groups$group == "10"]))
groupvar <- c(sd(groups$time[groups$group == "1"],na.rm=
 FALSE)^2, sd(groups$time[groups$group == "2"],na.rm= FALSE)^2,
              sd(groups$time[groups$group == "3"],na.rm=
 FALSE)^2, sd(groups$time[groups$group == "4"],na.rm= FALSE)^2,
              sd(groups$time[groups$group == "5"],na.rm=
 FALSE)^2,sd(groups$time[groups$group == "6"],na.rm= FALSE)^2,
              sd(groups$time[groups$group == "7"],na.rm=
 FALSE)^2,sd(groups$time[groups$group == "8"],na.rm= FALSE)^2,
```

```
             sd(groups$time[groups$group == "9"],na.rm=
 FALSE)^2,sd(groups$time[groups$group == "10"],na.rm= FALSE)^2)
gam <- data.frame(group=1:10, beta = groupvar/groupmean,
                  alpha = (groupmean^2)/groupvar, groupmean,
 groupvar)


# ---------------------------------------------
# import unique pairs for group look up function
# ---------------------------------------------
uniquepair <- read.csv("Documents/Thesis/uniquepair.csv")[,1:2]

# ---------------------------------------------
# create meet data frame
# ---------------------------------------------
schools <- c("cas","sba","se","sehs")
collegepairs<- data.frame("College1"=character(),"College2"
 =character())

for(ischool1 in schools){
  for(ischool2 in schools){
    if(!(ischool1 %in% collegepairs$College2 & ischool2 %in%
 collegepairs$College1)){
      tempcollegepairs <- data.frame("College1" =
 ischool1,"College2" =ischool2)
      collegepairs <- rbind(collegepairs,tempcollegepairs)
    }
  }
}

meetdf <- data.frame(collegepairs,
                     "Met" = 0,
                     "Total" = 0
)
#print(meetdf)

# ---------------------------------------------
# pick two paths at random regardless of school
# ---------------------------------------------


k <- 100,000 # number of times to run

# simulation for k times
for(isimulation in 1:k){
  #p <- sample(pathdf$path_id,2)
  p <- sample(pathdf$path_id, 2)
  testdf <- list(NA, NA)
  col <- rep(NA, 2)
```

```
 # for each path, obtain the nodes in each path (path) and
 # create a temp dataframe (tdf)

 for(ipath in 1:length(p)){
   path <- pathdf$node_name[pathdf$path_id %in% p[ipath]]
   col[ipath] <- as.character(pathdf$college[pathdf$path_id %in%
p[ipath]][1])

   # for each node in the path, obtain the actual time t.
   for (inode in 1:length(path)){

     # if first node, set first element in vector
     # time and elapsed to 0
     if (inode == 1){
       time <- 0
       elapsed <- 0
     }else{
       # if not first node, obtain random gamma time
       # then add the time to vector time and elapsed

       pairlookup1 <- paste(path[inode],path[(inode -
1)],sep='')
       pairlookup2 <- paste(path[(inode - 1)],
path[inode],sep='')
       pairlookup1 <- gsub(" ","", pairlookup1)
       pairlookup2 <- gsub(" ", "", pairlookup2)

       index <-
ifelse(length(which(uniquepair$pair==pairlookup1))==0,
                      which(uniquepair$pair==pairlookup2),
                      which(uniquepair$pair==pairlookup1))


       g <- uniquepair$group[index]
       g <- as.integer(gsub(" ", "", g))
       randomt <- rgamma(1, gam$alpha[gam$group==g], scale =
gam$beta[gam$group==g])

       time <- c(time, randomt)
       elapsed <- c(elapsed, sum(time))
     }# end if else
   }#end for loop inode

   # compile data frame with i = path_id,
   # path = nodes in the path, time, and elapsed
   testdf[[ipath]]  <- data.frame(p[ipath], path, time, elapsed)
 }#end for each node in  path

 # create vector with the same nodes
 testdf1 <- testdf[[1]]
```

```
  testdf2 <- testdf[[2]]
  rowcol <- which(((meetdf$College1 == col[1]) & (meetdf$College2
== col[2]))|((meetdf$College1 == col[2]) & (meetdf$College2 ==
col[1])))
  meetdf[rowcol,4] <- meetdf[rowcol,4] + 1

  samenode <- intersect(testdf1$path, testdf2$path)
  #print(samenode)

  # if samenode is not 0,
  # check to see if they are said to have met
  # (elapsed time difference is less than 10 seconds)
  if(length(samenode) != 0){

    # for each element(node) in samenode
    for(isame in 1:length(samenode)){
      # collect elapsed time for n1 in first path
      t1 <- testdf1$elapsed[testdf1$path==samenode[isame]]
      # collect elapsed time for n2 in second path
      t2 <- testdf2$elapsed[testdf2$path==samenode[isame]]

      # check to see if they are said to have met
      if(abs(t1-t2) < 10){
        #  match colleges & update counts in the meetdf data
frame
        meetdf[rowcol,3] <- meetdf[rowcol,3] + 1
        break
      } # end if
    }# end for each isame
  }# end if
}# end for each isimulation

meetdf$conditional <- meetdf$Met/meetdf$Total  # conditional
 probabilities
print(meetdf)
```

Example R Output for Simulation (meetdf)

```
  College1 College2 Met Total conditional
1       cas      cas 118   297  0.39730640
2       cas      sba  67   673  0.09955423
3       cas       se  97   678  0.14306785
4       cas     sehs  30   582  0.05154639
5       sba      sba 158   346  0.45664740
6       sba       se  88   748  0.11764706
7       sba     sehs  22   668  0.03293413
8        se       se  41   432  0.09490741
9        se     sehs  57   768  0.07421875
10     sehs     sehs  77   308  0.25000000
```

Bibliography

"University of Dayton Factbook Fall 2017." University of Dayton Factbook Prepared by Office of Institutional Reporting.

Lee, S. and Jo, Min Gu. "Probabilistic Modeling of Interactions: An Empirical Analysis of Simulation Using R." UC Berkeley.

The R Project for Statistical Computing, https://www.r-project.org/.