Time Series Analysis

Dr. Sarath Dantu

25/11/2020

```
knitr::opts_chunk$set(echo = T,fig.align = TRUE)
```

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install("limma")
```

```
## Bioconductor version 3.12 (BiocManager 1.30.10), R 4.0.2 (2020-06-22)
```

```
## Installing package(s) 'limma'
```

```
##
## The downloaded binary packages are in
##  /var/folders/yl/tc243tgd34xbl27q0wgyvw640000gn/T//RtmpPAx1aE/downloaded_packages
```

```
library(limma)
```

```
## Warning: package 'limma' was built under R version 4.0.3
```

We are going to learn:

- Reading in time series data
- Plotting
- Trends

1. Dataset

You are going to work with a time series data from my research. In this dataset I was tracking movement of Benzene particle in eight simulations performed by me. You have eight files in the data folder (if you have not downloaded the data, please check the week9 folder on the blackboardlearn CS5701 module web page).

Each data file contains two columns: - Time (nanoseconds –> ns) - Distance in Angstroms (Å), Angstrom is $10^{-8}cm$

```
# I do not want to declare multiple objects to store data, so to make my life easy...
list_file_names <- c("DT-1","EY-2","EY-5","EY-6","GC-1","GC-2","SK-3","VN-1")
# declare an empty list
data <- list()
for (i in c(1:length(list_file_names)))
```

```
{
  # do ensure you have the data/ folder in the directory you are working in. Hint getwd()
  file_name<-paste("data/",list_file_names[i],".out",sep="")
  file_data<-read.table(file_name)

  # the time column is common in files and it is sufficient to store it once
  if(i==1) {
    data<-cbind(file_data$V1,file_data$V2)
  } else {
    # I do not want the time column so I am just storing the Benzene displacement
    data<-cbind(data,file_data$V2)
  }
}
# For plotting I would like to have convenient column names
colnames(data)<-c("Time",list_file_names)
```

How to access the data from the dataframe...

```
# Just making sure things are proper
head(data)
```

```
##       Time     DT-1     EY-2    EY-5    EY-6    GC-1    GC-2    SK-3    VN-1
## [1,] 0.002 10.14631  9.90610 5.78904 7.96696 4.30019 5.39233 3.63035 6.86475
## [2,] 0.004 10.38451 10.49575 5.89161 7.90949 4.55008 5.55319 2.62067 6.46629
## [3,] 0.006 10.51646 10.21704 6.93535 7.49162 4.69418 5.80869 2.49526 6.27893
## [4,] 0.008  9.74766  9.99790 6.37290 7.32073 4.73039 6.62391 2.57575 6.61082
## [5,] 0.010 10.40680  9.80587 5.60543 6.79244 4.21244 7.04065 3.13868 6.56787
## [6,] 0.012 10.11349  8.91882 6.87878 7.76503 4.88980 6.92569 3.86879 6.41133
```

```
head(data[,"Time"])
```

```
## [1] 0.002 0.004 0.006 0.008 0.010 0.012
```

```
summary(data[,"DT-1"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.203   6.762   7.676  14.747  22.852  57.881
```

2. Data visualisation

Declaring the color palette. I find it easier to declare things upfront so that I can use them multiple times and backmapping the data to color guide makes life easy.

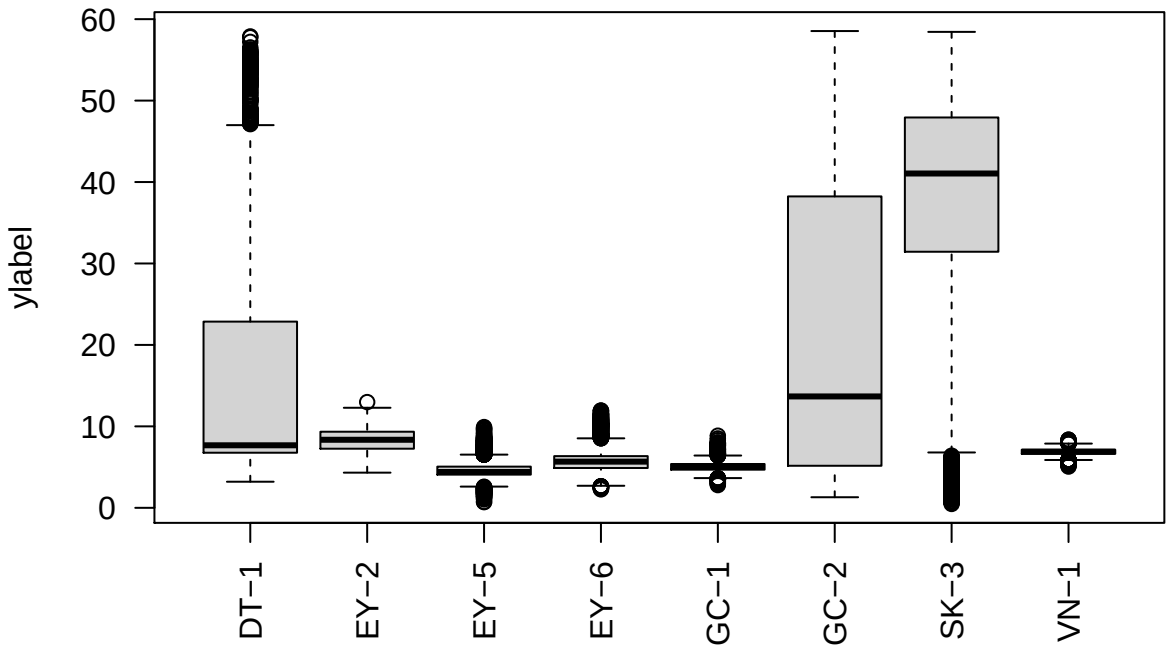
```
transparency<-0.5
data_color <- c(
  rgb(0.0,0.0,0.0,transparency), #black
  rgb(1.0,0.0,0.0,transparency), #red
  rgb(0.0,1.0,0.0,transparency), #green
  rgb(0.0,0.0,1.0,transparency), #blue
  rgb(1.0,0.5,0.0,transparency), #orange
```

```
  rgb(0.0,1.0,1.0,transparency), #cyan
  rgb(0.0,0.5,0.5,transparency), #teal
  rgb(0.28,0.24,0.2,transparency) #taupe
)
ylabel <- "Distance (Å)"
xlabel <- "Time (ns)"
plot_title <- "Benzene displacement"
```

Box plot

Can you identify any trends from the box plot?

```
par(las=2)
boxplot(subset(data,select=list_file_names),horizontal=F,ylab="ylabel")
```



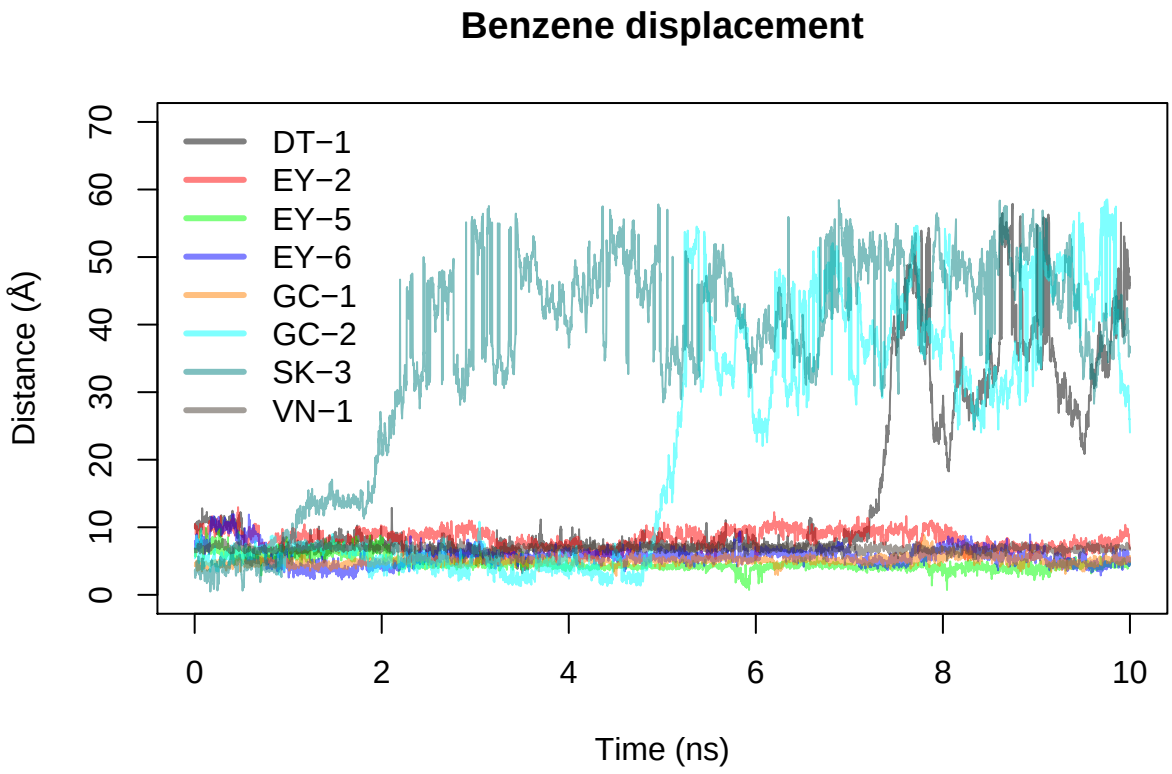
Time vs. Benzene displacement

What do you see in this time series data. Distance close to 0, Benzene stays where it has started, if the distance increases benzene is running away.

```
for (i in c(1:length(list_file_names)))
{
```

```
  # the time column is common in files and it is sufficient to store it once
    if(i==1) {

    plot(data[,"Time"],data[,list_file_names[i]],xlab=xlabel,ylab=ylabel,main=plot_title,col=data_color
  } else {
    lines(data[,"Time"],data[,list_file_names[i]],col=data_color[i])
  }
}
legend(x="topleft",legend=list_file_names,col=data_color,horiz=F,lwd=3,bty='n')
```



Histograms

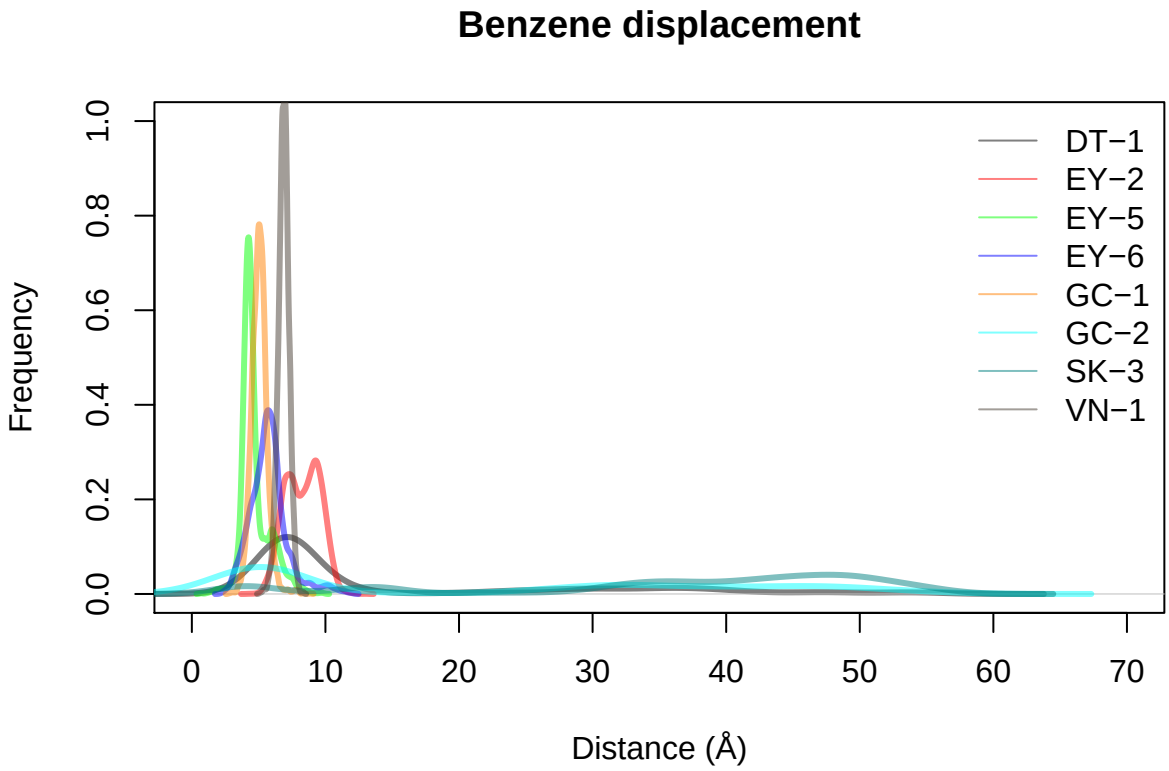
What does the distribution of the data tell us? Is it possible for Benzene to stay in multiple places?

```
ylabel <- "Distance (Å)"
xlabel <- "Time (ns)"
for (i in c(1:length(list_file_names)))
{

  # the time column is common in files and it is sufficient to store it once
    if(i==1) {

    plot(density(data[,list_file_names[i]]),xlab=ylabel,ylab="Frequency",main=plot_title,col=data_color
  } else {
```

```
    lines(density(data[,list_file_names[i]]),col=data_color[i],lwd=3)
  }
}
legend(x="topright",legend=list_file_names,col=data_color,horiz=F,lwd=1,bty='n')
```



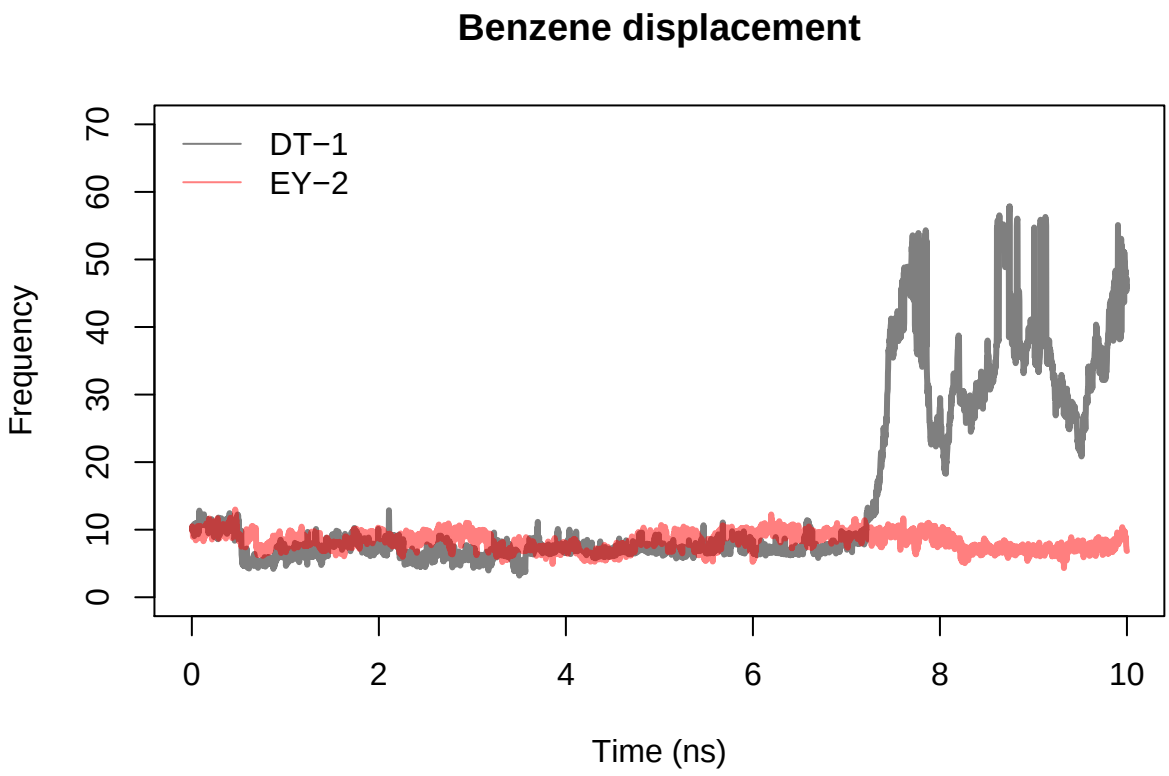
Looking at a single dataset

We are now going to take a closer look at DT-1 and EY-2. How would you identify outliers in this data?

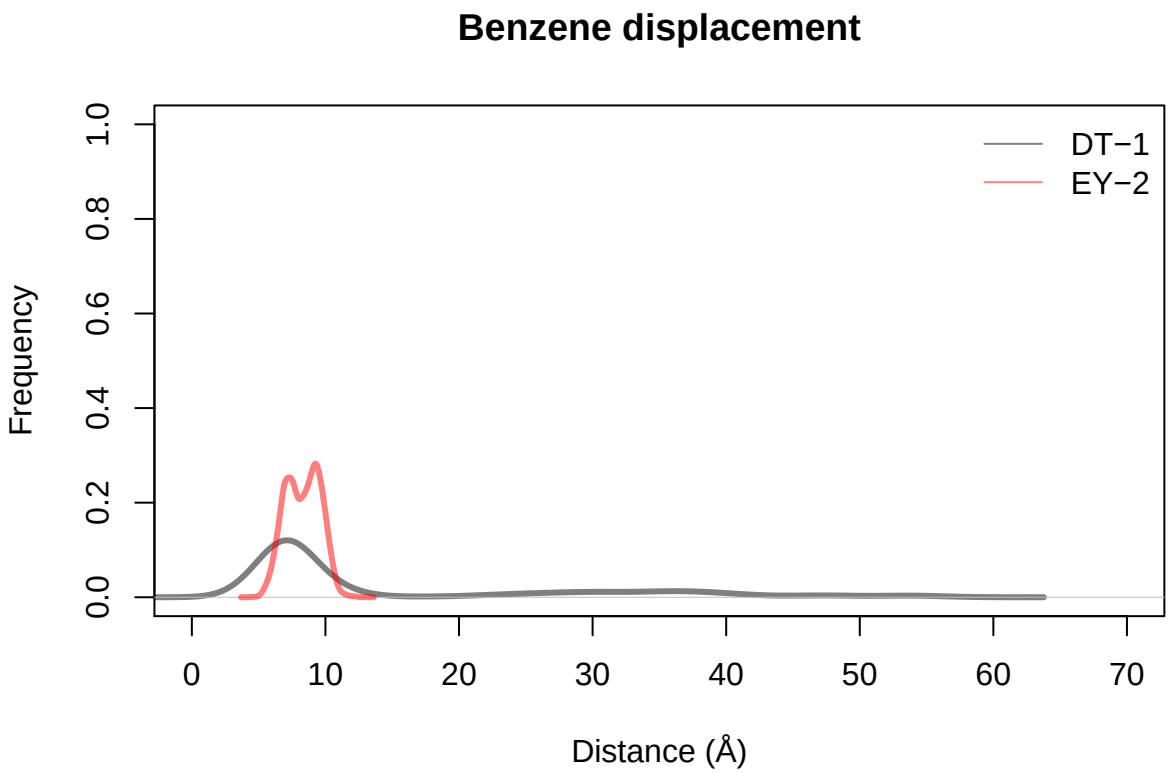
```
summary(cbind(data[,"DT-1"],data[,"EY-2"]))
```

```
##        V1               V2
##  Min.   : 3.203   Min.   : 4.318
##  1st Qu.: 6.762   1st Qu.: 7.259
##  Median : 7.676   Median : 8.354
##  Mean   :14.747   Mean   : 8.310
##  3rd Qu.:22.852   3rd Qu.: 9.344
##  Max.   :57.881   Max.   :12.982
```

```
plot(data[,"Time"],data[,list_file_names[1]],xlab=xlabel,ylab="Frequency",main=plot_title,col=data_colo
lines(data[,"Time"],data[,list_file_names[2]],col=data_color[2],lwd=3)
legend(x="topleft",legend=list_file_names[1:2],col=data_color,horiz=F,lwd=1,bty='n')
```



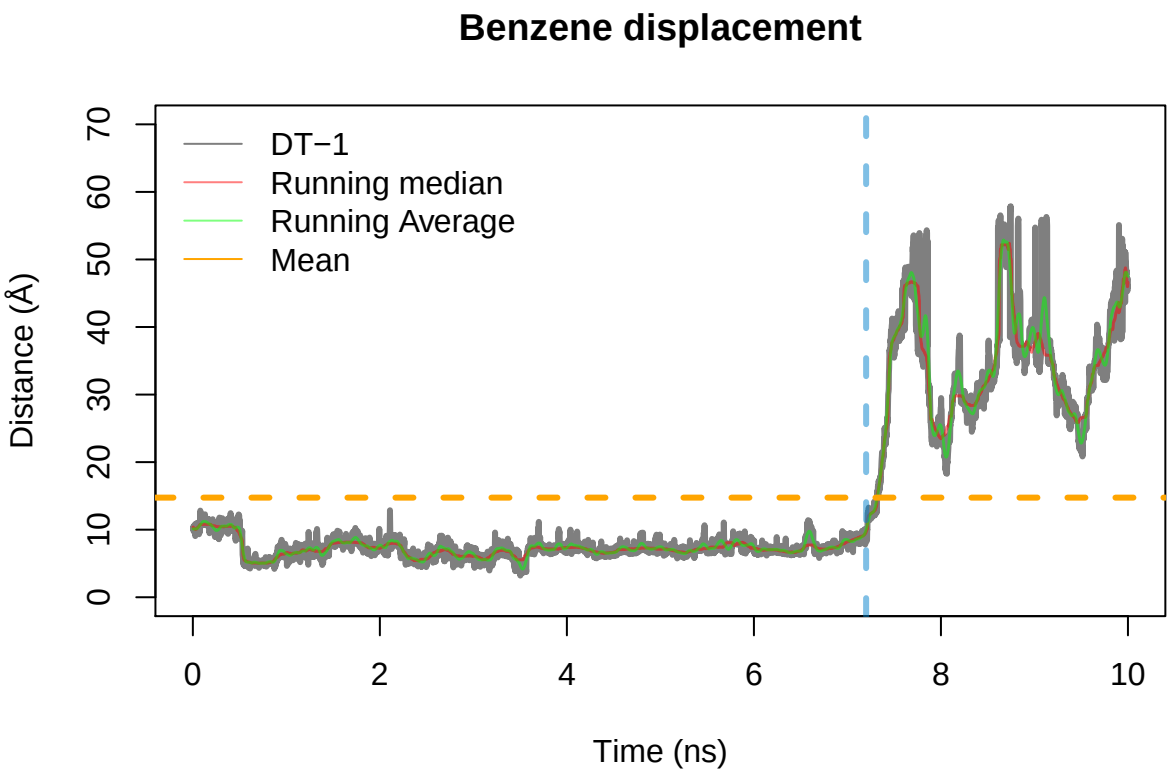

```
plot(density(data[,list_file_names[1]]),xlab=ylabel,ylab="Frequency",main=plot_title,col=data_color[1],
lines(density(data[,list_file_names[2]]),col=data_color[2],lwd=3)
legend(x="topright",legend=list_file_names[1:2],col=data_color,horiz=F,lwd=1,bty='n')
```



Trends

The data is very noisy. To understand the trends better, we can calculate the running average or running median, i.e. moving median of every 'n' points is calculated.

```
plot(data[,"Time"],data[,list_file_names[1]],xlab=xlabel,ylab=ylabel,main=plot_title,col=data_color[1],
lines(data[,"Time"],runmed(data[,list_file_names[1]],k=101),col=data_color[2],lwd=1.5)
lines(data[,"Time"],tricubeMovingAverage(data[,list_file_names[1]],span=0.01),col=data_color[3],lwd=1.5
abline(h=mean(data[,list_file_names[1]]),lty=2,lwd=3,col="orange")
abline(v=7.2,lty=2,lwd=3,col=rgb(0,0.5,0.8,0.5))
legend(x="topleft",legend=c("DT-1","Running median","Running Average","Mean"),col=c(data_color[1:3],"ora
```



Optional

Can you get rid of a subset (without deleting the data from the dataframe) and check what is going on with the remaining simulations?