SMPE TP

Yangtao WANG

1/10/2020

Introduction of the subject

In 1854, the Soho quarter of London saw one of the worst cholera epidemics of the United Kingdom, leading to 616 deaths. This outbreak has become famous because the detailed analysis of its cases proposed by the physician John Snow. He showed in particular that cholera was transmitted through water, rather than through the air as it was commonly believed at the time.

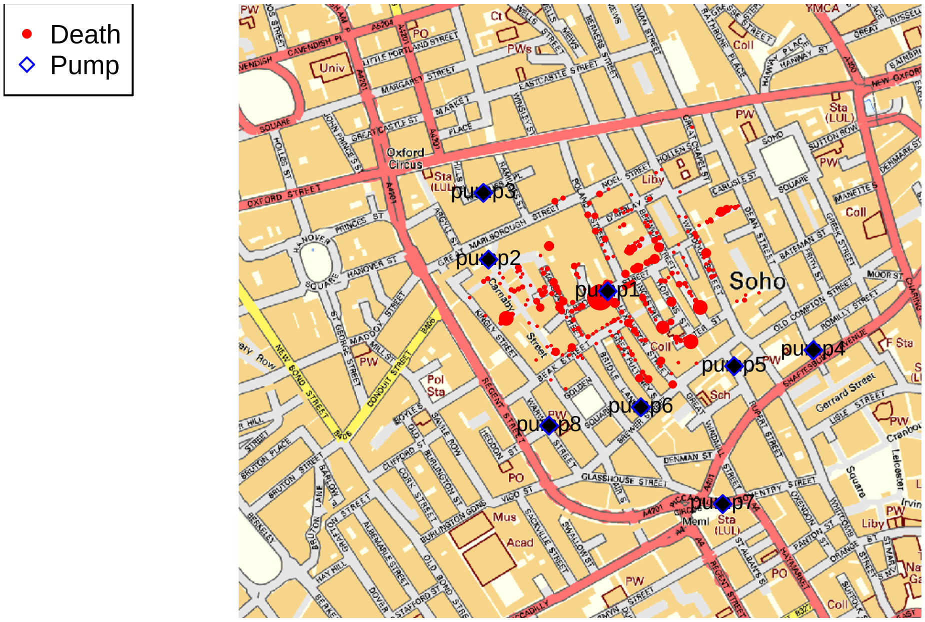
A key element of Snow's analysis was a map on which he had marked the places where people died and also the public water pumps. These data are today available on digital form. We ask you to use them to re-create John Snow's map in a computational document.

Mission

1. On the basis of the numerical data, produce a map in the spirit of John Snow's. Indicated the places where people dies with markers whose size indicates the number of deaths. Show the pumps on the same map, using a different symbol and/or color.
2. Try to find different ways to show that the Broad street pump is at the center of the outbreak.

Recreation of John Snow's map

We use the digital dataset from http://blog.rtwilson.com/john-snows-cholera-data-in-more-formats/ to reproduce the John Snow's map. The code of the reproduction is in the annexe.



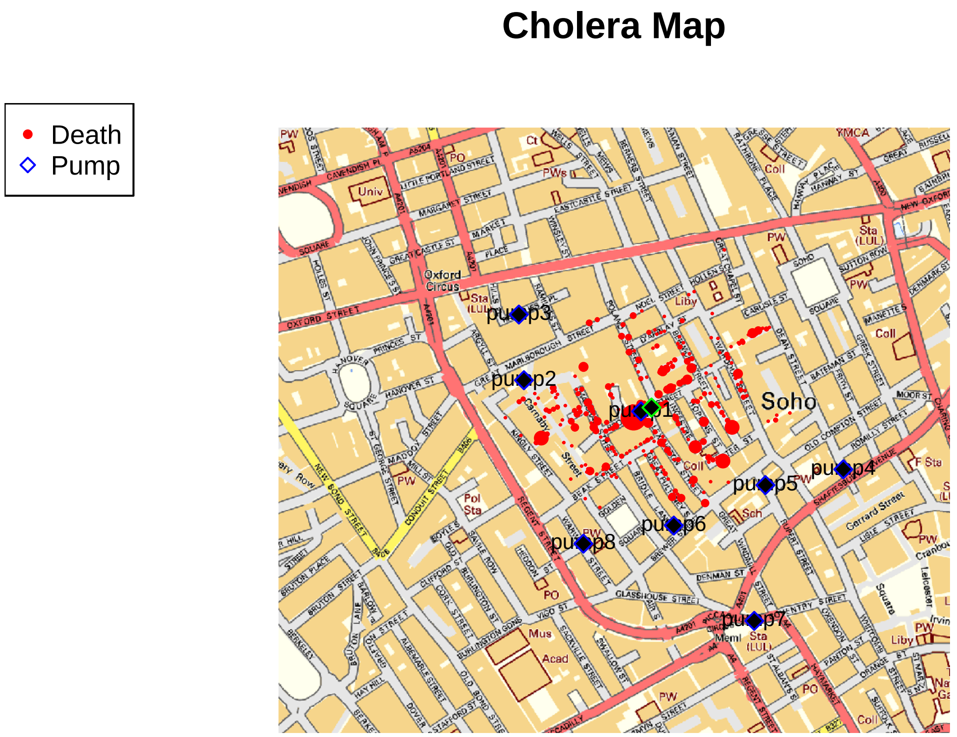
We can see from the map that there are 8 pumps(blue) in the quarter in total. They are numbered from “pump1” to “pump8”. The red points are the locations where people dies with markers whose size indicates the number of death. We can easily find that all the red points are surrounded the pump1, which is the Broad street pump. In the next section, we will use different methods to prove that it is at the center of the outbreak.

Method

In this section, we will use different way to prove that the Broad street pump is at the center of the outbreak.

Barycenter

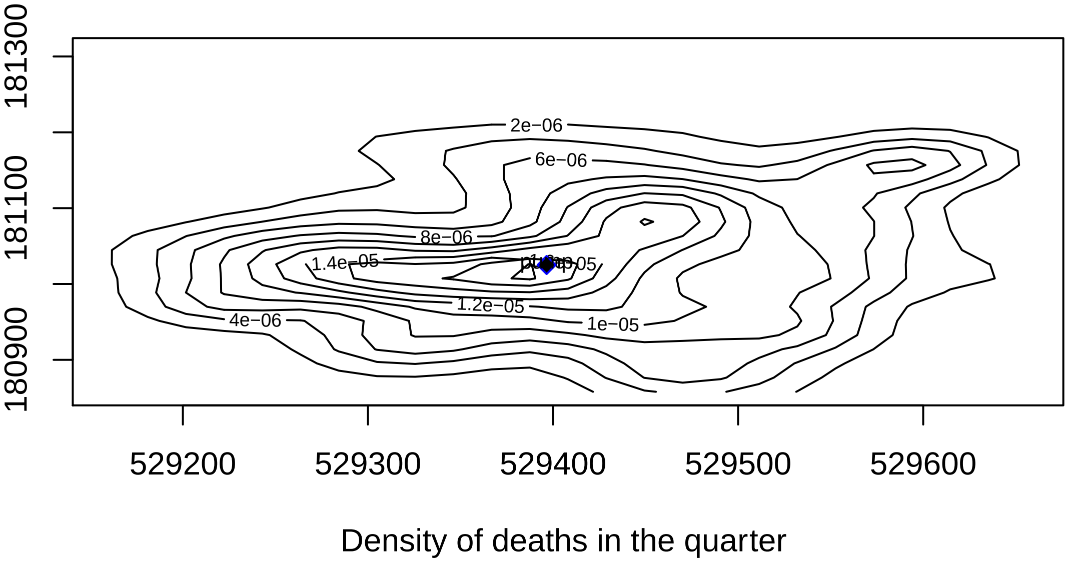
We can calculate the Barycenter of all the death people, which is marked as a green point in the following map. The Barycenter of the death people is very close to the ‘pump1, which is the Broad street pump.



By computing the Euclidean Distance, we find that the distance between barycenter of death people and the pump1 is the smallest, which means that the pump1 is cloest to the outbreak. From this comparison, we can prove that the Broad street pump is at the center of the outbreak.

```
##      pump1    pump2    pump3    pump4    pump5    pump6   pump7    pump8
## 1 19.45574 227.2384 282.0666 351.0775 240.0644 209.1758 412.307 265.0819
```

Density



We can also prove it by plot the density of the death people. The figure above is a contour plot. The highest contour line is at the center of the figure. We plot the location of the Broad Street Pump on the figure as well, it is very close to the hightest contour line, which also proves that it is as the center of the outbreak.

Discussion about the result

From the prespective of the statistics, it is proved in the previous section that the Broad Street Pump is at the center of the death people, which should be the outbreak of the cholera. However, as we don't know where this digital data comes from, so we have the reason to doubt on the reliability of the source, that means, the pump might not the be the reason for cholera. In addition, even though we prouve that the pump is at the center of the outbreak, but it may be only by chance, and the true source, for example, might be a resturant, a bakery etc.

As a conclusion, to better understand the true reason of the death, we need some more data about the death, for example, we need to collect data about the common things they did before death, in that way we can give a better inference about the source of chelora.

Annexe

Data reference

http://blog.rtwilson.com/john-snows-cholera-data-in-more-formats/

To install rgdal package

We need to install an dependence for exmaple in Mac Os :

```
brew gdal
```

In Ubuntu:

```
sudo apt-get install gdal-bin proj-bin libgdal-dev libproj-dev
```

Here are some reference for installing the gdal :

1. https://gist.github.com/dncgst/111b74066eaea87c92cdc5211949cd1e
2. https://stackoverflow.com/questions/15248815/rgdal-package-installation

Code

```
#Load data
library(maptools)
library(sp)
library(raster)
death<-rgdal::readOGR("./data/Cholera_Deaths.shp")
pump<-rgdal::readOGR("./data/Pumps.shp")
map<-raster("./data/OSMap.tif")

#Create map
plot(map, main="Cholera Map")
plot(death, add=T, col = "red",  pch=20, cex=death$Count/6)
pump_coord = coordinates(pump)
pump$Id <- paste("pump", 1:length(pump), sep ="")
text(pump_coord[,1], pump_coord[,2], labels=pump$Id, cex= 0.7)
plot(pump, add=T, col = "blue", pch=23 )
legend("topleft", legend=c("Death", "Pump"),
       col=c("red", "blue"), pch=c(20,23), cex=0.8)

plot(map, main="Cholera Map")
plot(death, add=T, col = "red",  pch=20, cex=death$Count/6)
text(pump_coord[,1], pump_coord[,2], labels=pump$Id, cex= 0.7)
plot(pump, add=T, col = "blue", pch=23 )
legend("topleft", legend=c("Death", "Pump"),
       col=c("red", "blue"), pch=c(20,23), cex=0.8)

barycenter_x = sum(
  death$Count * coordinates(death)[,1]
) / sum(death$Count)
barycenter_y = sum(
  death$Count * coordinates(death)[,2]
```

```r
) / sum(death$Count)
bar_x = c(barycenter_x)
bar_y = c(barycenter_y)
bar.df = data.frame(matrix(ncol=2, nrow=1))
colnames(bar.df) = c("x", "y")

bar.df$x = bar_x
bar.df$y = bar_y

barycenter = SpatialPointsDataFrame(
  coords=bar.df,
  data=bar.df,
  proj4string=death@proj4string)

plot(barycenter, add=T, col = "green", pch=23 )

#Compute distance between barycenter and pumps
distance<-data.frame(matrix(ncol = length(pump$Id)))

colnames(distance) = c(pump$Id)
index = 1
for(pumpID in pump$Id){
  distance[1,index] <- dist(rbind(coordinates(barycenter), coordinates(pump)[pump$Id == pumpID]))
  index = index + 1
}
print(distance)


#Contour
library(MASS)

all_death_coord = function(x,y,count){
  coords<-data.frame(matrix(ncol = 2, nrow = sum(count)))
  index = 1
  for(i in 1:length(x)){
    for(j in 1:count[i]){
      coords[index,1]=x[i]
      coords[index,2]=y[i]
      index = index + 1
    }
  }
  return(coords)
}

death_coordinate<-all_death_coord(coordinates(death)[,1],coordinates(death)[,2],death$Count)
f1<-kde2d(death_coordinate[,1], death_coordinate[,2])
contour(f1, xlab = "Density of deaths in the quarter ")
coord_center_pump = coordinates(pump)[pump$Id == "pump1"]
center_pump_df = data.frame(matrix(ncol=2, nrow=1))
colnames(center_pump_df) = c("x", "y")

center_pump_df$x = coord_center_pump[1]
center_pump_df$y = coord_center_pump[2]
```

```r
center_pump = SpatialPointsDataFrame(
  coords=center_pump_df,
  data=center_pump_df,
  proj4string=death@proj4string)
plot(center_pump, add=T, col = "blue", pch=23 )
text(center_pump$x, center_pump$y, labels="pump", cex= 0.7)
```