Poisson processes for the arrival of claims

Vermeir Jellen

December 21, 2015

Contents

1 Introduction

In this paper we study the claim arrival process of the famous 'Danish fire insurance' set in the period January 1st, 1980 until December 31st 1990. Our goal is to investigate the suitability of both homogeneous and in-homogeneous Poisson processes to model the claim arrivals. We first visualize the dataset and investigate the descriptive statistics of the arrival and inter-arrival times for the whole period. We initially assume that the claim arrival process follows a homogeneous Poisson process and subsequently calibrate the intensity parameter to the data.

Next, we investigate the corresponding statistics for each year separately and illustrate that the homogeneity assumption does not capture the data well: The intensity parameter of the Poisson process varies over time and the non-homogeneous intensity function $\lambda(t)$ contains non-stationary properties. However, in this context we also illustrate that a homogeneous Poisson process might be suitable to model claim arrivals for shorter time-intervals. We conclude the paper with an illustration on how to make an 'operational time change' from a non-homogeneous to a homogeneous model. We accomplish this by using the continuous mean value function of the corresponding non-homogeneous process to transform the inter arrival times.

2 Fitting claim arrivals to a poisson process

2.1 Danish fire insurance - data

The Danish fire insurance set contains 2167 claim counts. A visualization of the arrival times and the corresponding number of cumulative arrivals up to that time point are shown on the top graph, in figure 1. The x-axis represents the number of days passed since January 1, 1980. The y-axis represents the total number of claims that have been received up to this date. The bottom graph illustrates the frequency distribution of the inter-arrival times.

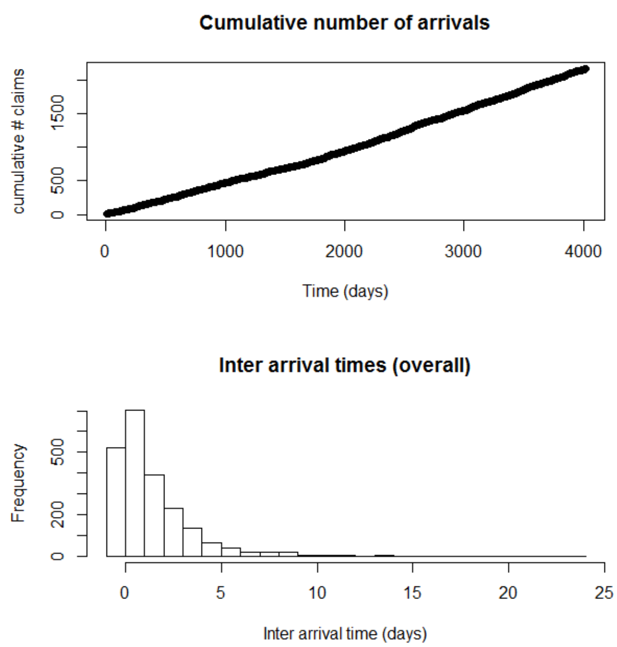

Figure 1: Danish fire insurance set - Arrival process and inter-arrival times

2.2 Calibrating a homogeneous Poisson process

Here, we argue that the inter-arrival time frequency distribution graph resembles the density function of an exponential distribution. If we consider the claim arrivals as a Poisson process with intensity rate λ then it can be shown that the inter-arrival times are indeed i.i.d Exp(λ). Hence, under this assumption we can determine the MLE $\hat{\lambda}$. As illustrated in the appendix A.1, to obtain $\hat{\lambda}$, we calibrate the claim arrival process to the poisson likelihood. Alternatively, we can also calibrate the inter-arrival times to the exponential distribution in order to obtain the same estimated value of 0.54. In figure 2. below, we compare the empirical inter-arrival times against the corresponding Exp(0.54) distribution. At first glance, we conclude that the data seems to fit the model relatively well.

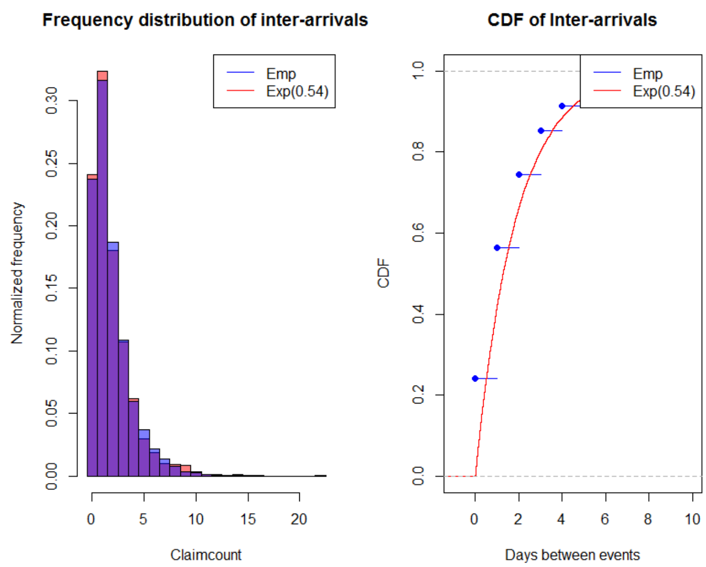

Figure 2: Frequency Distribution and CDF (empirical VS Exp(0.54))

2.2.1 Homogeneity assumption for whole data period

In the previous section we assumed the data to be homogeneous across the complete time period. However, in an insurance context, one will usually be faced with in-homogeneous claim arrival processes. In figure 3 below, the descriptive statistics of the data are shown for each year separately. Note that the MLE estimator for each individual period is provided as well. As first glance, the $\hat{\lambda}$ and corresponding annual mean inter-arrival times seem to fluctuate heavily across different time periods. This invalidates our homogeneity assumption.

	year	NrSamples	Minimum	First_Quartile	Median	Mean	Third_Quartile	Maximum	lambda
1	Overall	2167	0	1	1	1.854	3	22	0.539457
2	1980	166	0	1	2	2.199	3	11	0.454795
3	1981	170	0	1	2	2.147	3	12	0.465753
4	1982	181	0	1	1	2.017	3	10	0.49589
5	1983	153	0	1	2	2.379	3	22	0.42033
6	1984	163	0	1	2	2.252	3	16	0.444142
7	1985	207	0	1	1	1.754	2	14	0.570248
8	1986	238	0	0	1	1.538	2	14	0.650273
9	1987	226	0	0	1	1.619	2	9	0.617486
10	1988	210	0	0	1	1.719	3	12	0.581717
11	1989	235	0	0	1	1.574	2	15	0.635135
12	1990	218	0	0	1	1.674	2	9	0.59726

Figure 3: Overall and yearly data for inter-arrival times and claim-count

The distributions of the inter-arrival times are further illustrated on the left side in figure 4 below. The figure clearly demonstrates that a larger chunk of the data is centered closer around 0 as time goes by. For example, the first quartile and median values were 0 and 1 in 1986 while their respective counterparts were 1 and 2 in 1983. Furthermore, the plot on the right demonstrates that there are more claim counts during later time periods: This is consistent with the lower corresponding average inter-arrival times.

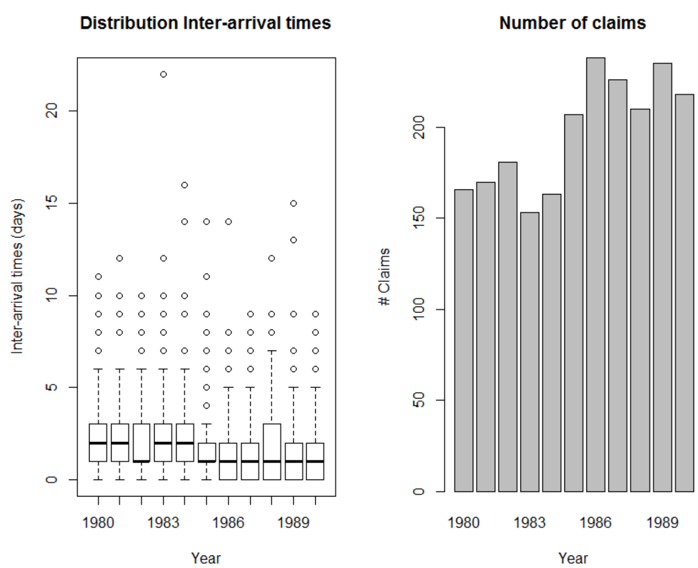

Figure 4: Yearly data for inter-arrival times and claim-count

Next, we investigate a moving average estimate of the intensity function $\lambda(t)$ of a non-homogeneous Poisson process. We illustrate the calibration of the Danish fire data to this intensity function in the appendix A.2. The result is shown in figure 5 below. The figure illustrates that the function is not stationary: The estimates of the the mean inter-arrival times move downwards over time which provides further evidence that the empirical data is indeed of in-homogeneous nature.

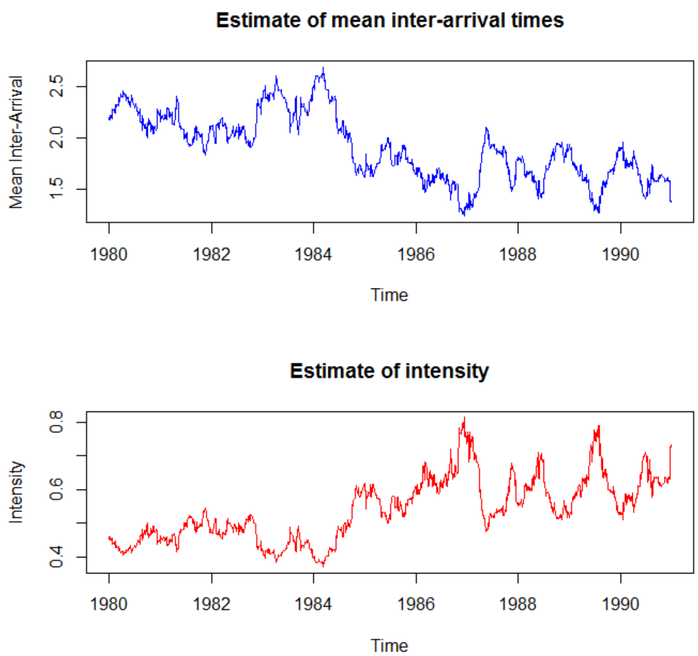

Figure 5: Moving mean inter-arrival times and intensity function

2.2.2 Homogeneity assumption for yearly time periods

In the previous subsection we concluded that the homogeneity assumption for the complete time period was not met. However, a homogeneous model assumption might be suitable for the arrival process during shorter time periods such as one year. We investigate this proposition for the 1980 Danish fire insurance data.

The top left graph in Figure 6. below shows the claim-count process $N(t)$ versus t for claims arriving in the year 1980, together with simulated sample paths from a Poisson process with intensity rate $\hat{\lambda}_{1980} = 0.455$. The frequency distributions and density plots for the empirical data are compared against an Exp(0.455) distribution in the top right and bottom left plots. The bottom right plot illustrates a QQ plot of the theoretical Exp(0.455) quantiles versus the empirical quantiles. We observe a few outliers in the tail, but in general the data matches the theoretical distribution fairly well. Hence, the claim arrival process can be considered homogeneous on this shorter time interval.

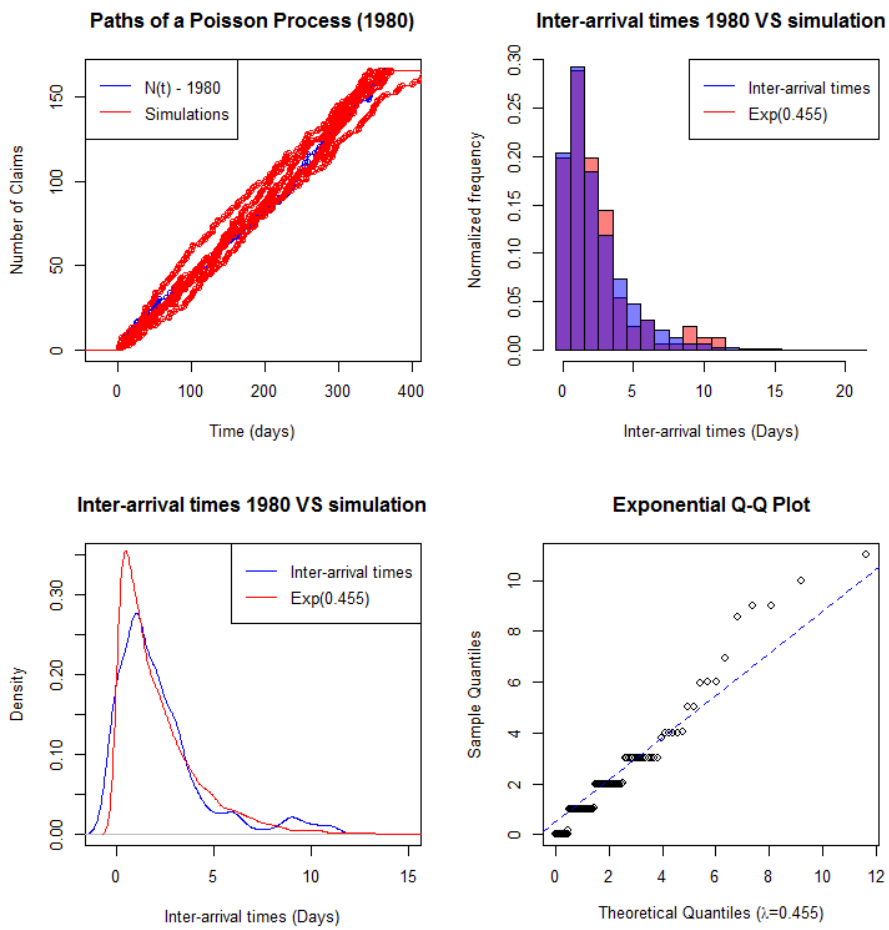

Figure 6: Claim-count process (1980) Versus Exp(0.455) simulations

2.3 Calibrating a non-homogeneous Poisson process

In this section we utilize an in-homogeneous Poisson process with continuous mean value function $\mu(.)$ to model the arrival process. In the last section we concluded that the claim arrival process for the Danish fires could be considered homogeneous during shorter yearly intervals. Hence, we can utilize the estimated yearly intensity rates -as reported in figure 3- to estimate $\mu(t)$ as a piecewise linear function with altering slopes during different years. We demonstrate the calculation in the appendix A.3 and figure 7 below shows a graph of $\mu(t)$ as a function of time t.

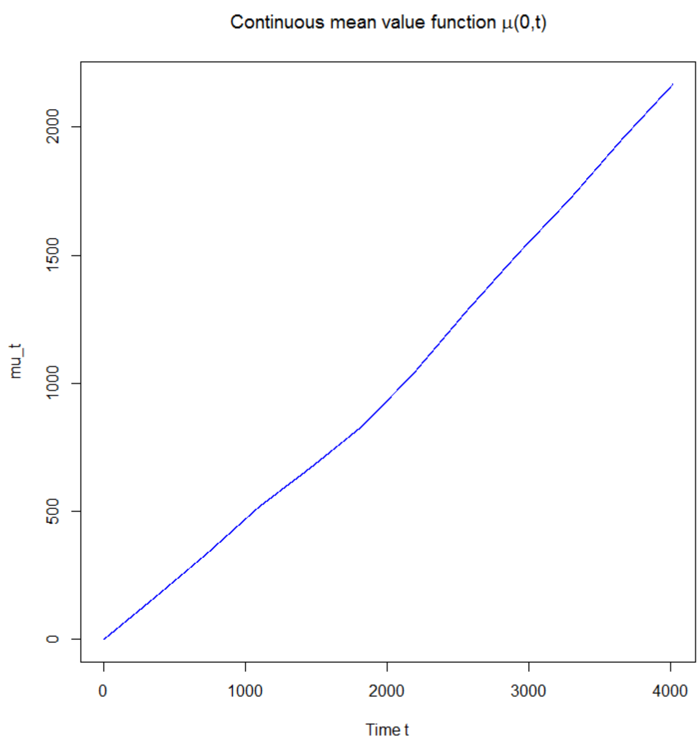

Figure 7: Graph of continuous mean value function $\mu(t)$

The proposition of operational time change implies that the non-homogeneous arrival times T_i can be transformed into arrival times of a homogeneous standard Poisson process. In figure 8. below we illustrate the moving average estimates of the mean inter-arrival time and intensity function of the transformed arrival times $\mu(T_i)$. Indeed, the graph illustrates that the intensity parameter moves around 1 in a stationary manner . In figure 9. we further verify that the inter-arrival times of $\mu(T_i)$ are i.i.d standard exponential. The theoretical quantiles of the standard distribution seem to match the sample quantiles fairly well and there is no autocorrelation present in the plots. An inverted methodology allows us to simulate sample paths of an in-homogeneous Poisson process from the paths of a homogeneous Poisson process.

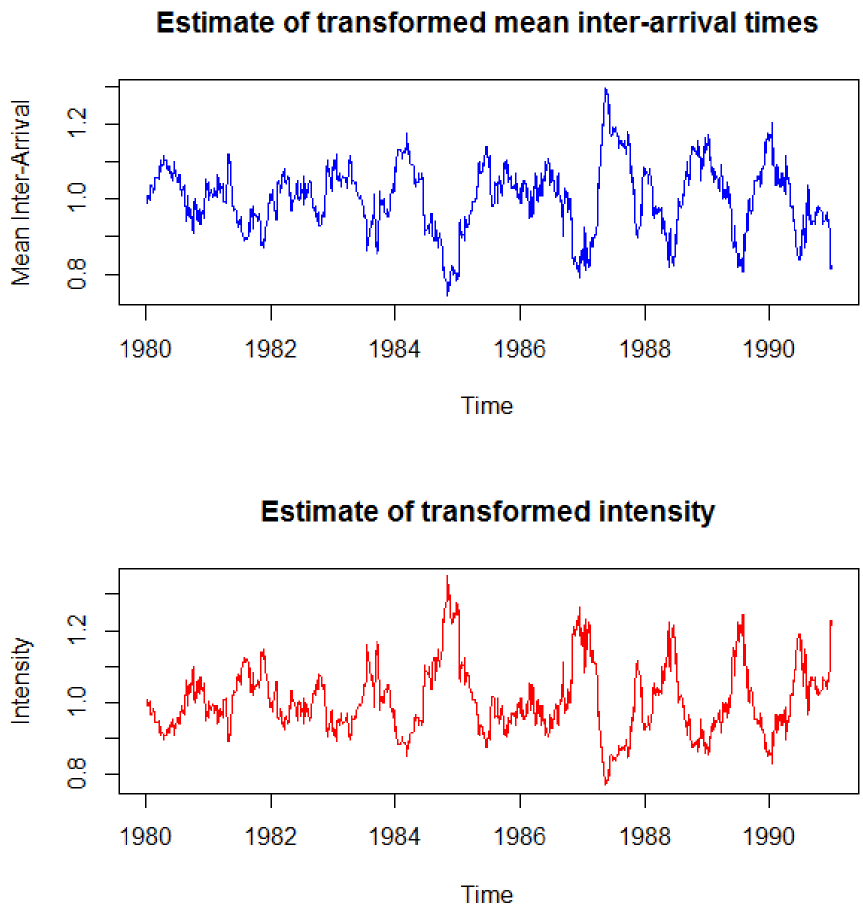

Figure 8: Moving mean inter-arrival time and intensity function of $\mu(T_i)$

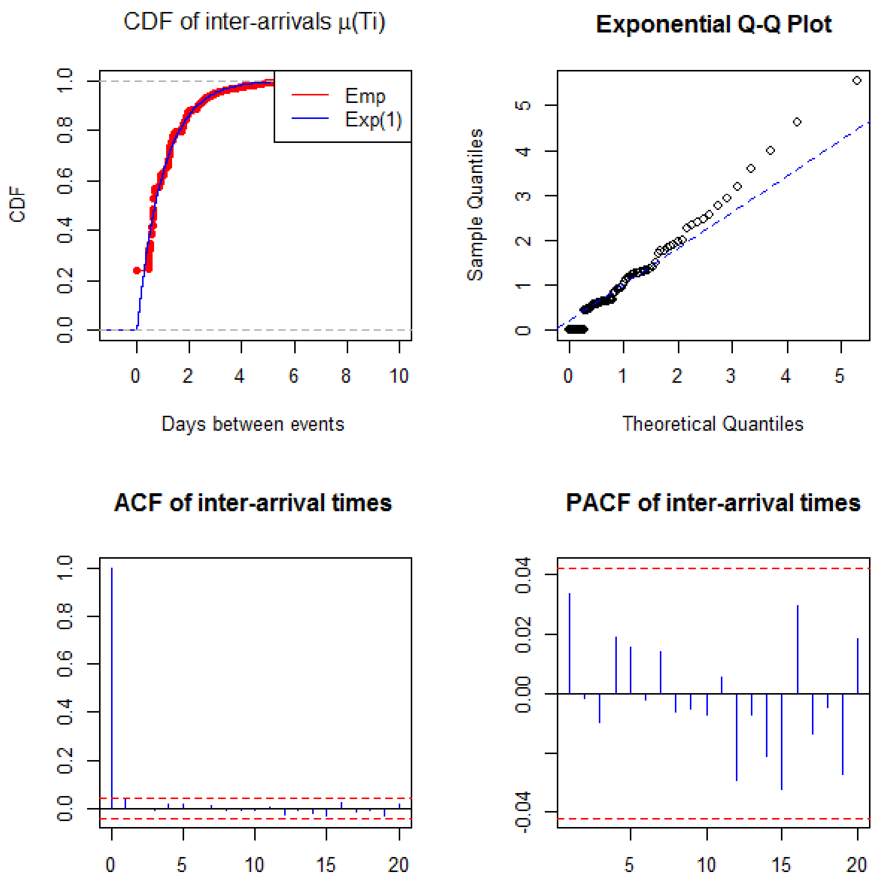

Figure 9: Inter-arrival times of $\mu(T_i)$ are $\sim$ Exp(1)

Appendices

A Code fractions

A.1 MLE calculation of homogeneous poisson process

```
startDate <- as.Date("01/01/1980",format="%d/%m/%Y")
arrivalDates <- as.Date(read.table(file="DanishData.txt",
                                   header=TRUE)$Date, format="%m/%d/%Y")
```

```
# Arrival times since start of the observation period,
# expressed in days
Ti <- as.numeric(difftime(arrivalDates,startDate,units="days"))
# Interarrival times, expressed in days
Wi <- diff(c(0,Ti))
# Number of claims at time T (1 claim for each arrival time)
claimCount <- rep(1,length(Ti))

library(plyr)
claims <- ddply(data[,c("time","Wi","count")],"time",
                                            numcolwise(sum))
claims$year <- format(claims$time,"%Y")

# Maximum likelihood for P[N(t+s)-N(s)=n]
Poi.lik = function(x,beta,expo){
  lambda = exp(beta) # transform unconstrained. Make lambda positive
  dens = exp(-lambda*expo)*(lambda*expo)^x/gamma(x+1)
  logl = sum(log(dens))
  return(-logl)
}
Poi.result = nlm(Poi.lik,1,hessian=TRUE,
                        x=claims$count,expo=claims$Wi)

lambda <- exp(Poi.result$estimate)
lambda
# -2*ln(L) + 2*k
Poi.AIC <- (+2)*Poi.result$minimum+2*length(Poi.result$estimate)
Poi.AIC

# Alternative
fm_pois <- glm(claims$count~1,family=poisson(link="log"),
                                            offset=log(claims$Wi))
lambda <- exp(fm_pois$coefficients[[1]])
lambda

# Or..
lambda = weighted.mean(claims$count/claims$Wi,claims$Wi)
lambda

# Or..
lambda = 1/mean(data$Wi)
```

A.2 Moving average estimates of intensity function $\lambda(t)$

```
centeredMean <- function(center=51,width=50,data)
{
  if(!is.numeric(data))
    stop("Input data should be numeric vector")

  sum <- 0;
  n <- length(data)

  minIndex <- center-width
  maxIndex <- center+width

  sum = sum(data[max(1,minIndex):min(n,maxIndex)])

  # Account for proportion of indices below 1
  if(minIndex <= 0)
    sum = sum + (abs(minIndex)+1)*data[1]
  # Account for proportion of indices above n
  if(maxIndex > n)
    sum = sum + (maxIndex-n)*data[n]

  return(sum/(2*width+1))
}

m <- 50
Wi <- data$Wi
movingMeanInterArrival <- sapply(1:length(Wi),
                         centeredMean, width=m, data=Wi)
movingLambda <- 1/movingMeanInterArrival

par(mfrow=c(2,1))
plot(data$time, movingMeanInterArrival, type="l", xlab="Time",
                         ylab="Mean Inter-Arrival", col="blue",
     main="Estimate of mean inter-arrival times")
plot(data$time, movingLambda, type="l", xlab="Time",
      ylab="Intensity", col="red", main="Estimate of intensity")
```

A.3 Calculation of the mean value function $\mu(t)$

```r
allDays <- seq(from=as.Date("1980/01/01"),
                        to=as.Date("1990/12/31"),by="days")
lambdas <- rep(0,length(allDays));
for(i in unique(data$year))
{
  lambda <- as.numeric(summaryStats$lambda[summaryStats$year==i])
  indices <- which(format(allDays,"%Y")==i)
  lambdas[indices] <- lambda
}

t <- 1:length(lambdas)
mu_t <- sapply(t, function(x) sum(lambdas[1:x]))

par(mfrow=c(1,1))
plot(t,mu_t,xlab="Time t",type="l",col="blue",
     main=expression(paste("Continuous mean value function ",
                                            mu,"(0,t)",sep="")))
```