Analysis of UK Drivers Deaths using Splines

Zubia Mansoor

02/12/2020

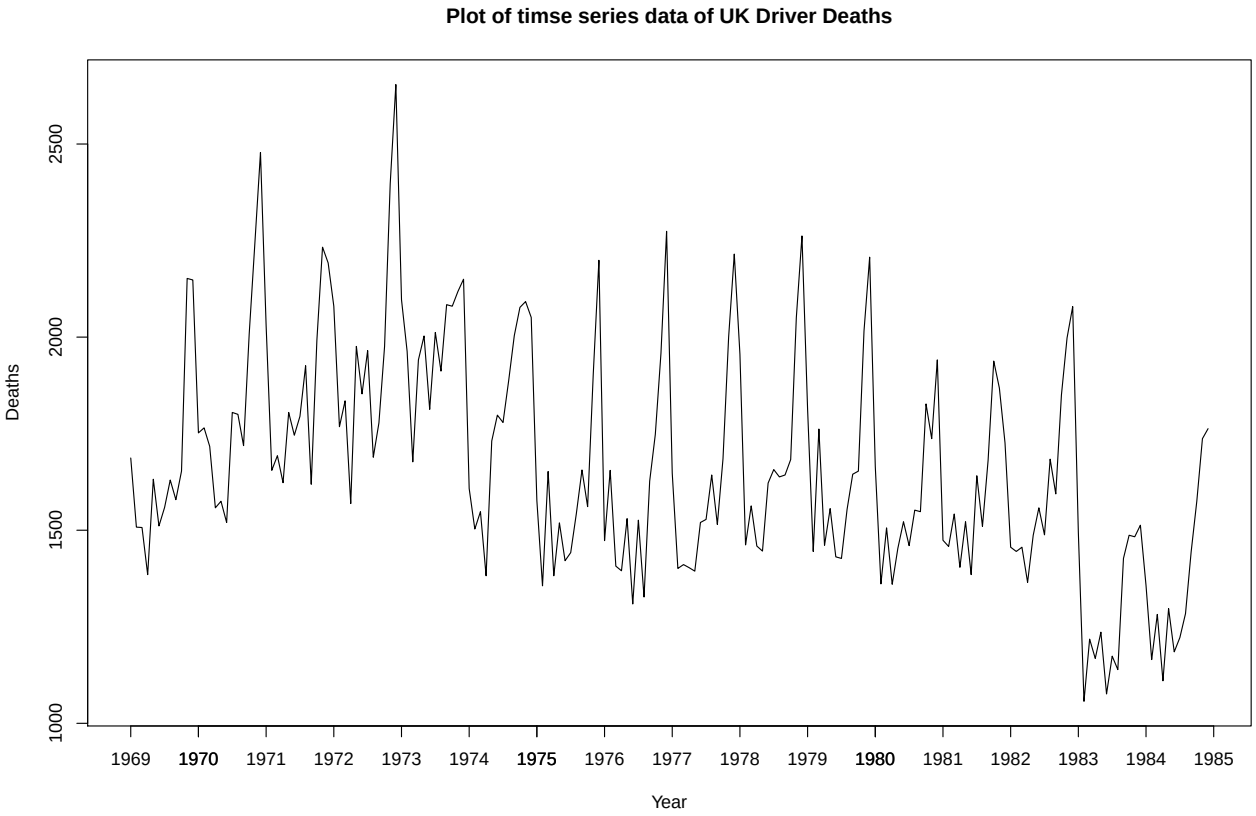
Exploratory Data Analysis

The `datasets` package has a set called `UKDriverDeaths`. The data contains monthly counts of automobile drivers killed in the UK from 1969 through 1984.

Let us view the first 10 rows of the dataset `UKDriversDeath`.

```
##   Deaths
## 1   1687
## 2   1508
## 3   1507
## 4   1385
## 5   1632
## 6   1511
```

The next step is to plot the time series data as below:

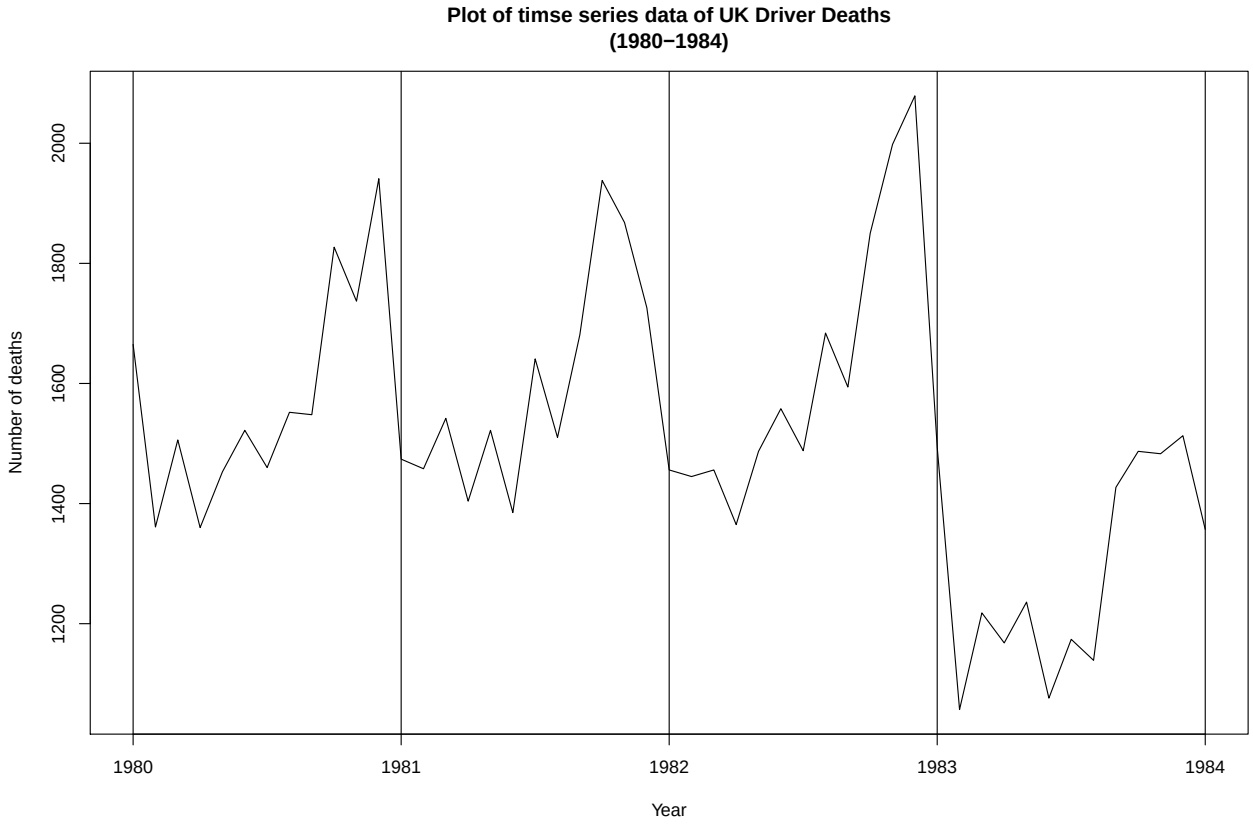


We observe that certain months of the year have generally higher death rates than others, creating a cyclic spiky pattern. We also see that the longer-term multi-year trends do not depict a flat line. Here, we are interested in the main mean trend ignoring the seasonal fluctuations.

Overall, we see an increasing death toll for the years 1969-1973. There is a steep decrease in deaths at the beginning of 1974. We then see a more or less similar pattern for the years 1974-1980. Following this, we

again see a decrease in deaths most pronounced in the 1983-1984 bracket.

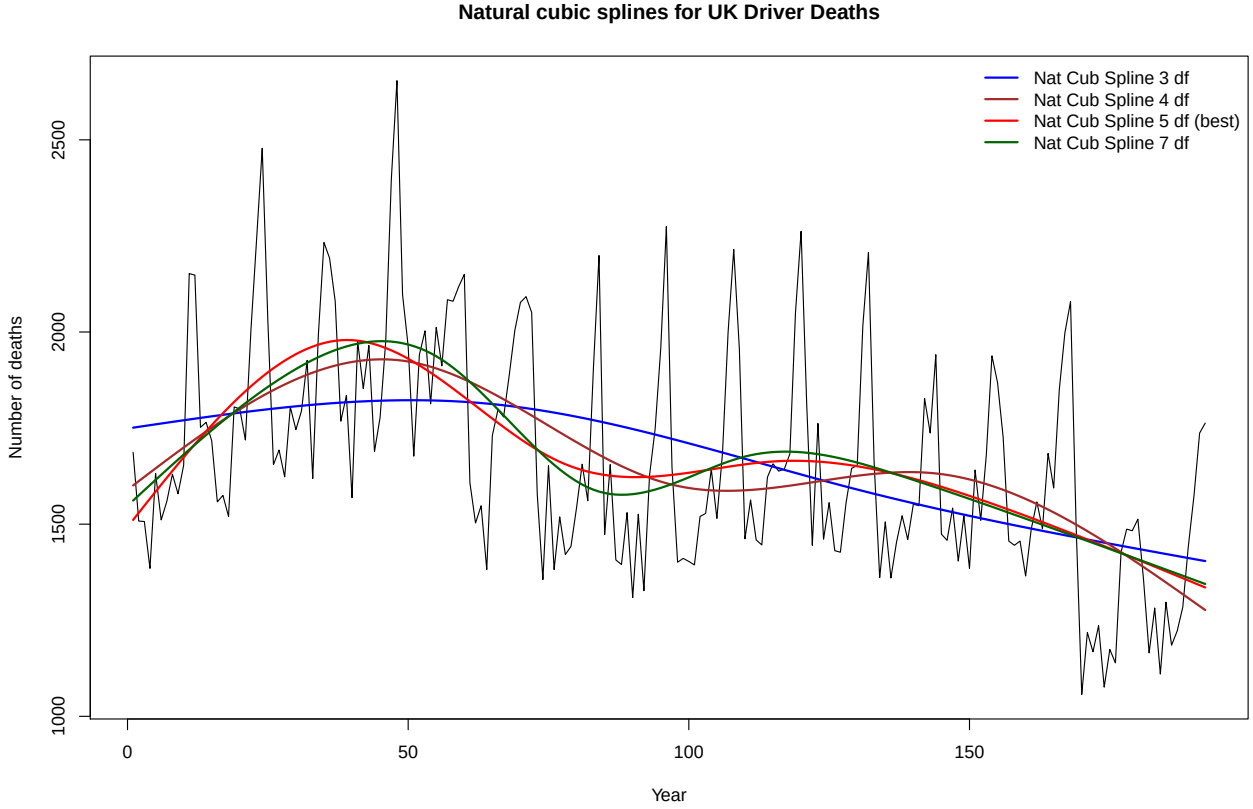
Let us zoom in to look at the trend between the years 1980 and 1984.



Yes, there clearly appears to be a significant decrease in deaths between 1983 and 1984 as compared to the other years.

Natural cubic splines

Here, we will use `natural cubic splines` with different degrees of freedom to capture the main mean trend while smoothing over the monthly spiky-ness.




```
## [1] "The MSEs for natural cubic splines with 3, 4, 5, 6, 7 DFs are:"
```

```
## [1] 63938.87 57824.47 56028.59 56252.61 55311.58
```

We fit natural cubic splines with degrees of freedom ranging from 3 to 7. For visual inspection, we plot the trend achieved by the natural cubic splines using 3, 4, 5 and 7 degrees of freedom. We observe that the cubic spline with `3 DF` is almost a straight line and fails to capture the underlying trend. Using `5 DF` appears to model the trend well without chasing the monlthy wiggliness too much and exhibits similar performance to a `7 DF` spline. Since we do not gain much by increasing the degrees of freedom, *the natural cubic spline with 5 DF is our best chosen curve*. If we go futher than that, we run the risk of chasing the monthly cycles.

We also take a look at their MSEs and the model with 5 DFs has the second lowest error. We note that the MSE goes down as we use 7 DFs or higher, possibly as the model starts chasing the monlthy trend rather than the main mean trend.

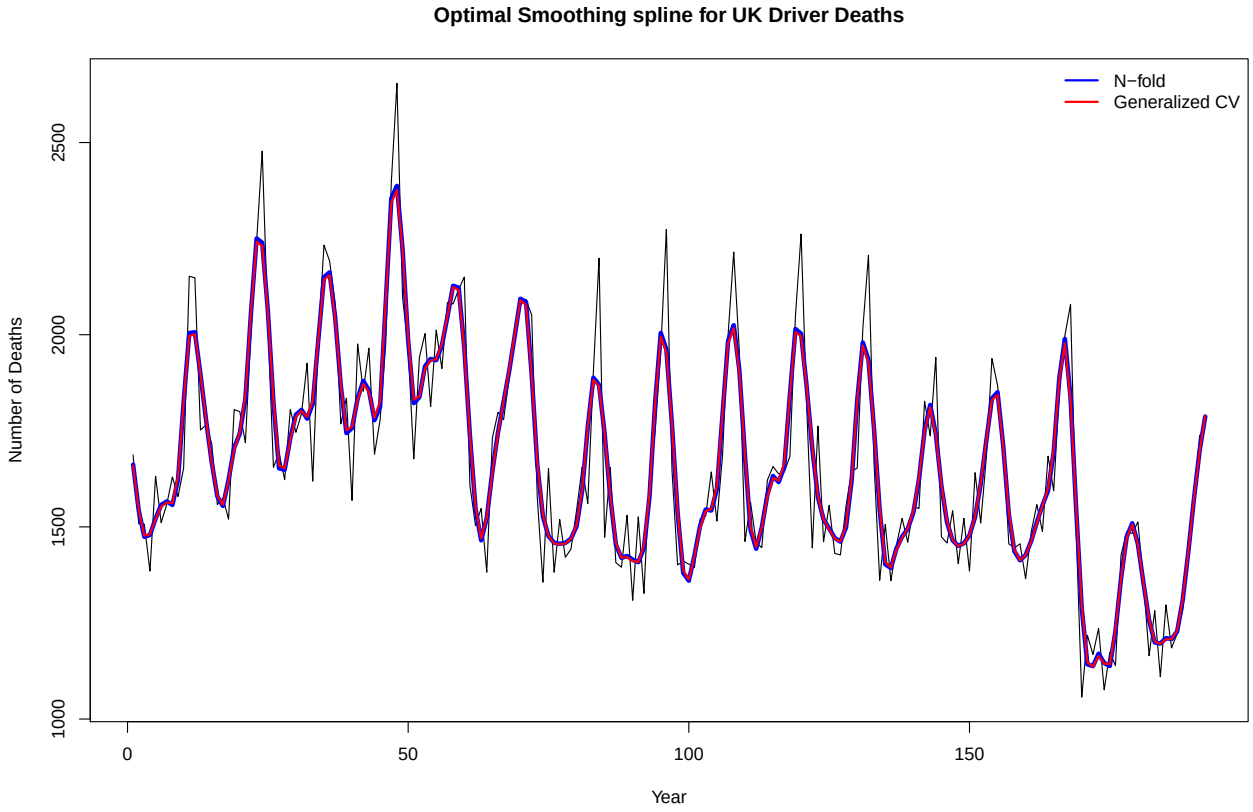
Optimal Smoothing splines

In this case, we will use `optimal smoothing splines` obtained using N-fold and Generalized cross-validation to model the `UKDriverDeaths`. Following are the results from each method.

```
## Call:
## smooth.spline(x = death.data$time, y = death.data$Deaths, cv = TRUE)
##
## Smoothing Parameter  spar= 0.2093414  lambda= 5.230072e-08 (10 iterations)
## Equivalent Degrees of Freedom (Df): 75.08636
```

```
## Penalized Criterion (RSS): 2082614
## PRESS(l.o.o. CV): 27826.57
```

```
## Call:
## smooth.spline(x = death.data$time, y = death.data$Deaths, cv = FALSE)
##
## Smoothing Parameter  spar= 0.2208054  lambda= 6.321932e-08 (12 iterations)
## Equivalent Degrees of Freedom (Df): 72.90066
## Penalized Criterion (RSS): 2157204
## GCV: 29199.38
```



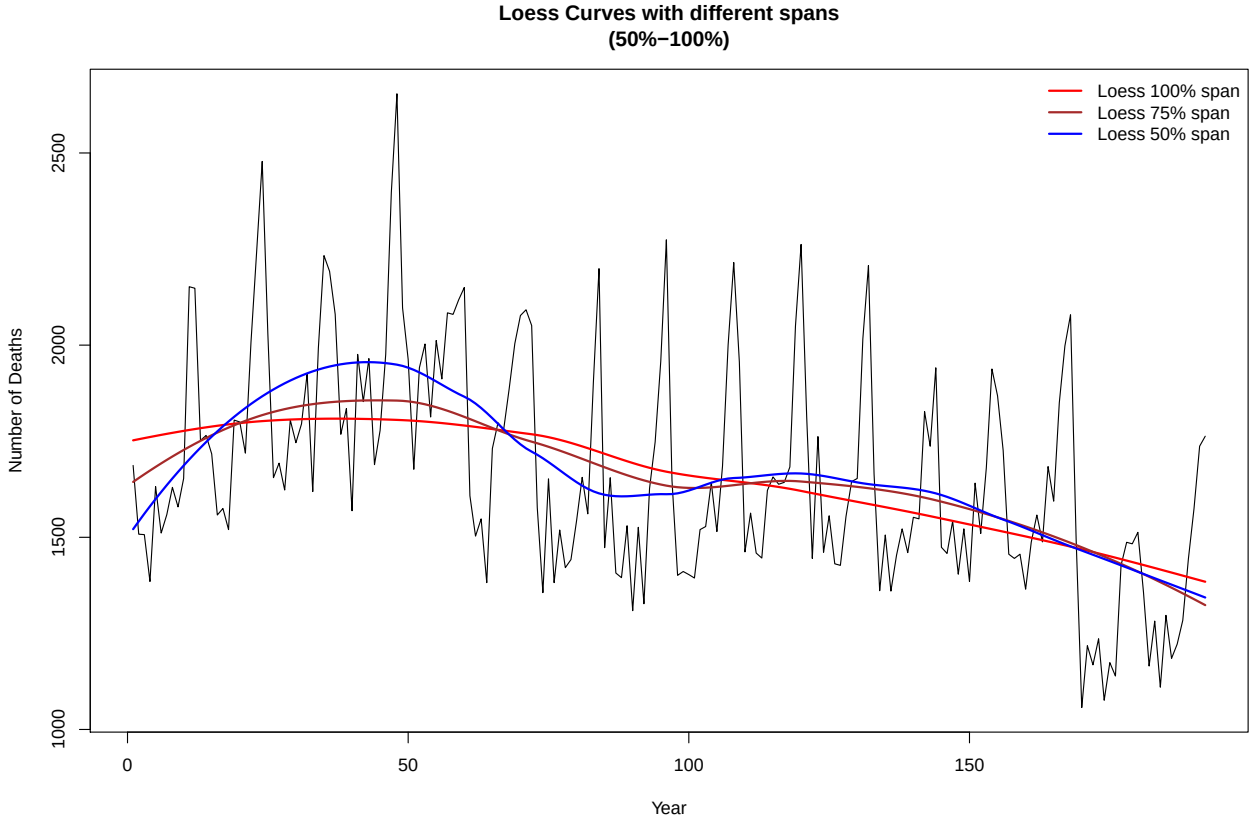
(a) Our goal is to achieve the main mean trend while smoothing over the monthly wiggliness. As is evident, the optimal smoothing splines obtained through `N-fold` and `Generalized` cross validation heavily chases the monthly fluctuations. Hence, it does a poor job of achieving our goal.

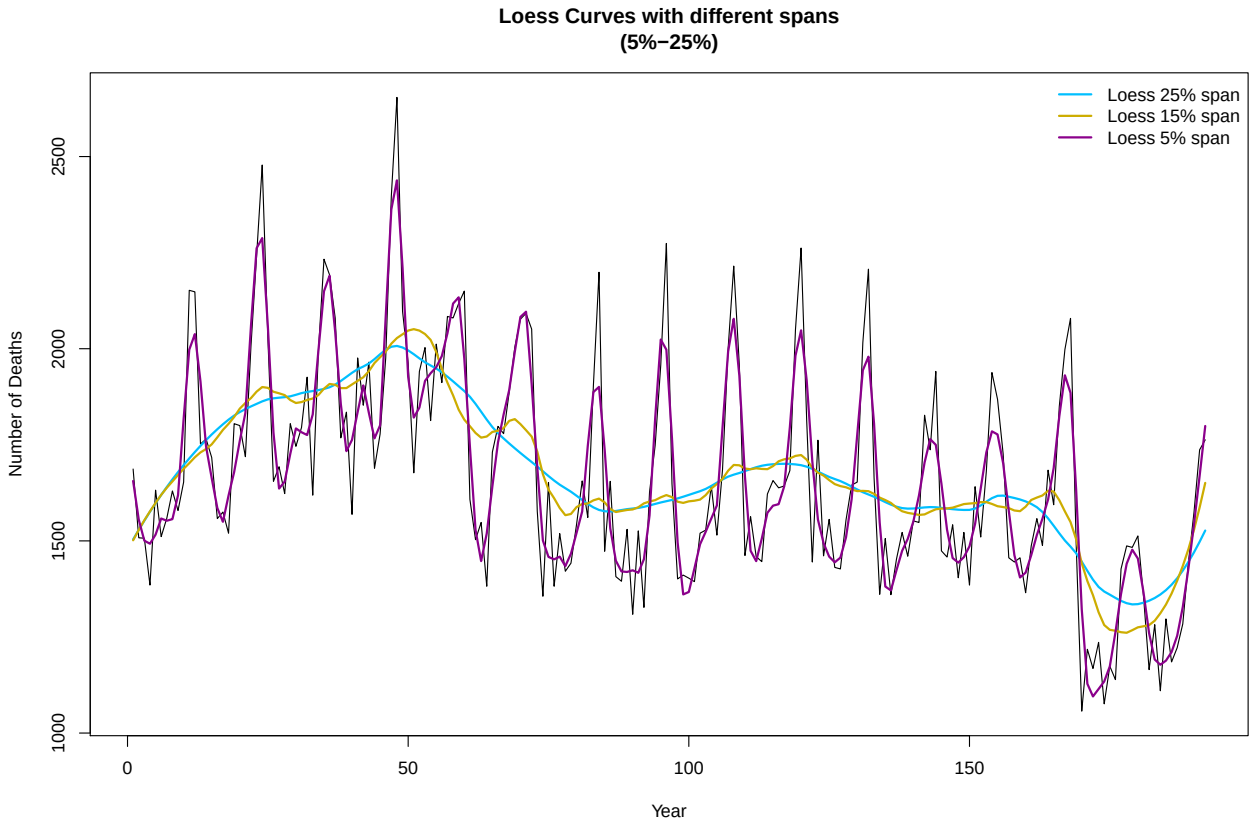
(b) The degrees of freedom obtained using `N-fold` and `Generalized` cross validation are 75 and 73 respectively. These are very high and hence very different from what we obtained using natural cubic splines (5 DF).

Loess curves

We will now fit `loess` curves with varying spans and degrees of freedom to the `UKDriverDeaths`. First, we try out different spans in the set {100%, 75%, 50%, 25%, 15%, 5%}. After selecting the best span, we try out the different degrees of freedom associated with it. Note that the `loess` in `R` uses a tri-cube kernel.

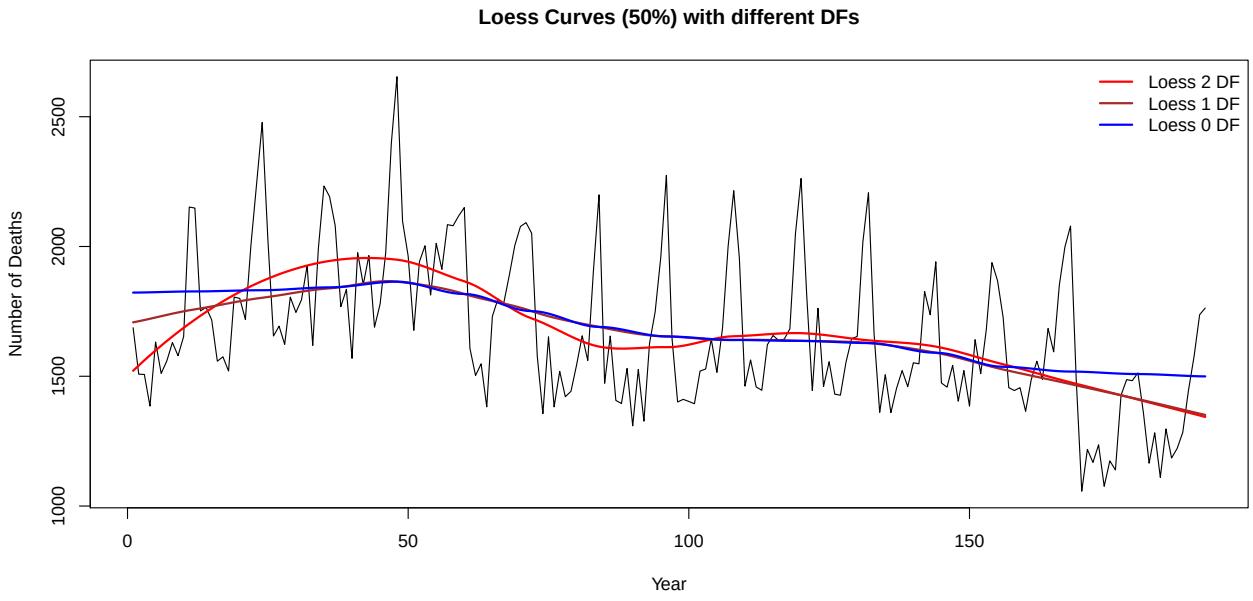
Let us first inspect the performance of `loess` curves over different spans visually.





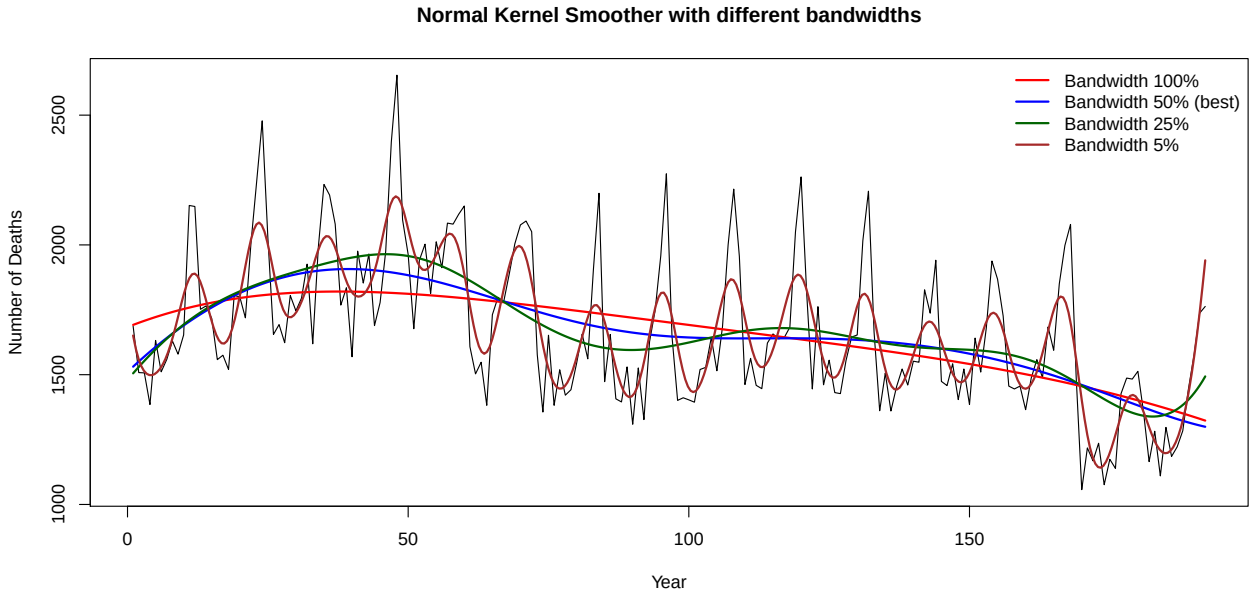
Looking at the first plot, we can say that loess curves with 100% and 75% do not capture the mean trend adequately and is almost a flat line. The second plot shows us that 5%, 15% and 20% span start overfitting the data and chasing the monthly patterns. Based on the graphs, the loess curve with 50% span appears to perform fairly well and captures the overall trend.

For the loess curve with 50% span, we try out the different degrees of freedom in the set $\{0, 1, 2\}$. Below is the plot showing the results.



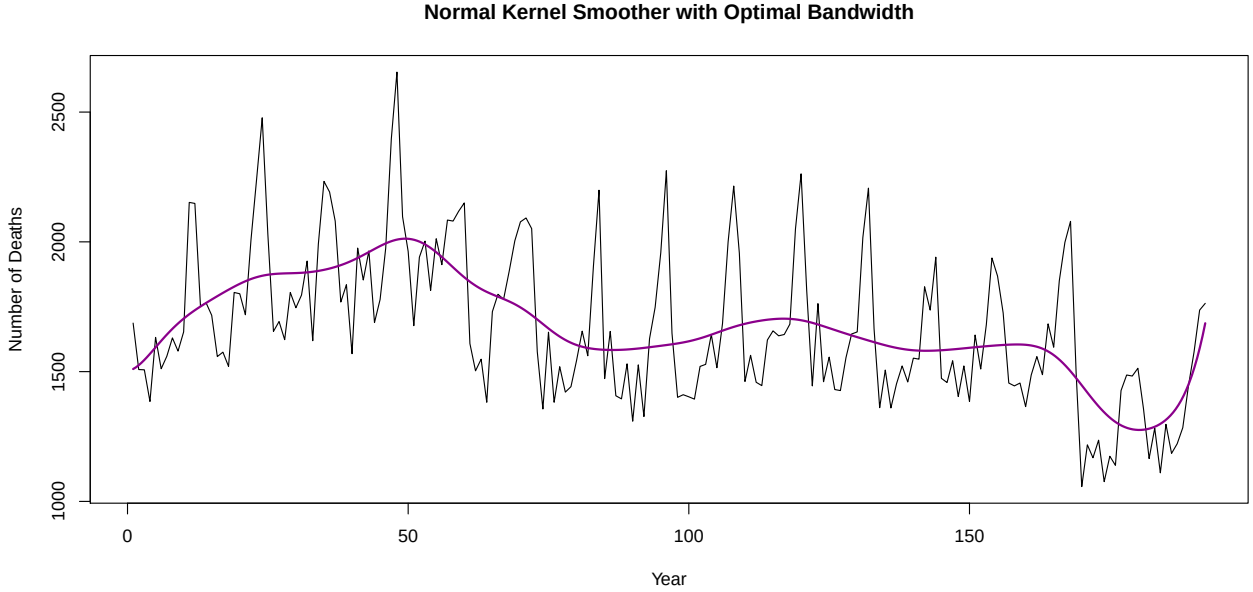
Thus, it is evident that the *loess curve with 50% span and 2 DFs is our best chosen curve.* The other two models have similar performance in the sense that they are unable to represent the main mean trend well.

(a) We repeat the same analysis, now with a `normal kernel`. We try different bandwidths in the range {100%, 50%, 25%, 5%} scaled by the standard deviation of X-values. Below is the plot depicting the different curves.



The figure above shows that 100% bandwidth is unable to capture the trend whereas 5% captures the seasonal patterns as well. Somewhere in between, 25% appears to model the trend moderately except at the very end where we expect the mean death rate to go down. Hence, the *normal kernel with 50% bandwidth (≈ 27 months) is our best chosen curve.*

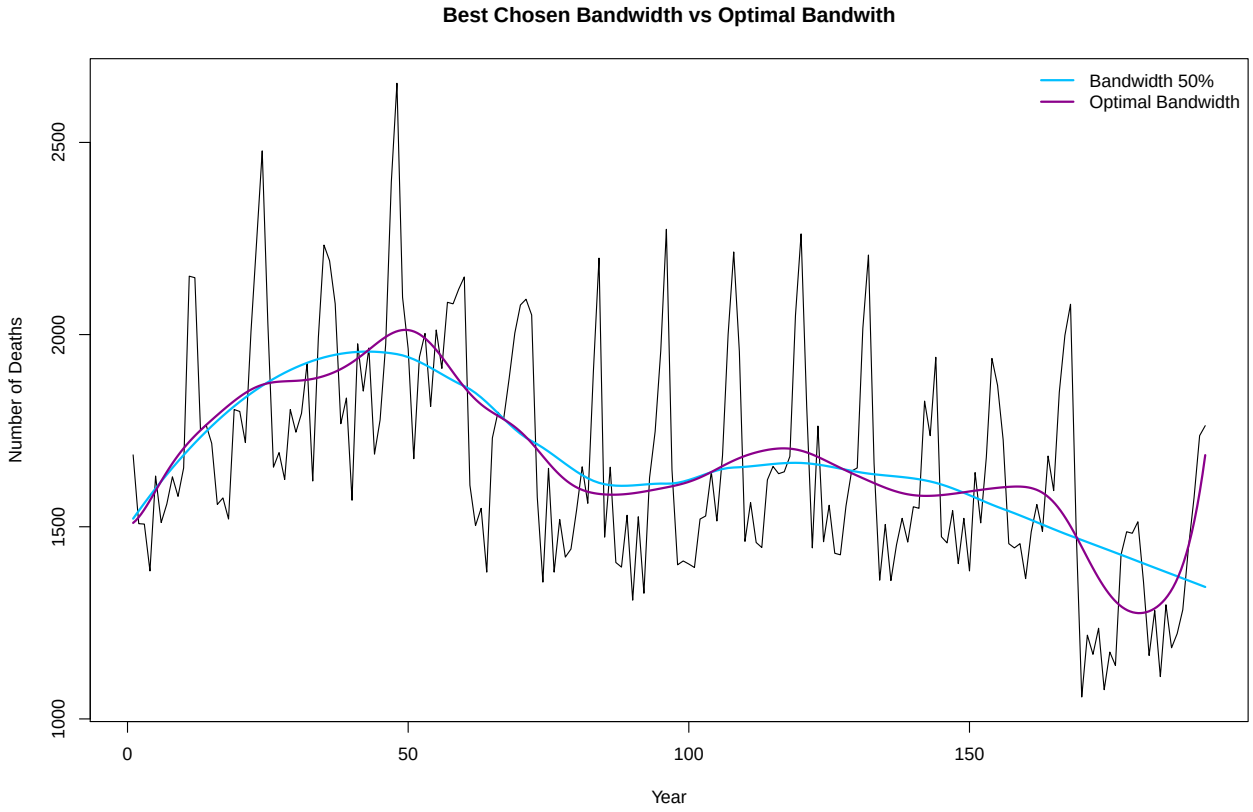
(b) Again, we repeat the analysis with a `normal kernel` this time using an optimal bandwidth as shown below.




```
## [1] "The optimal bandwidth is: 7.476"
```

The normal kernel smoother with optimal bandwidth (≈ 7 months) appears to be chasing the monthly wiggliness, especially at the tail where we expect a fall in overall death rate.

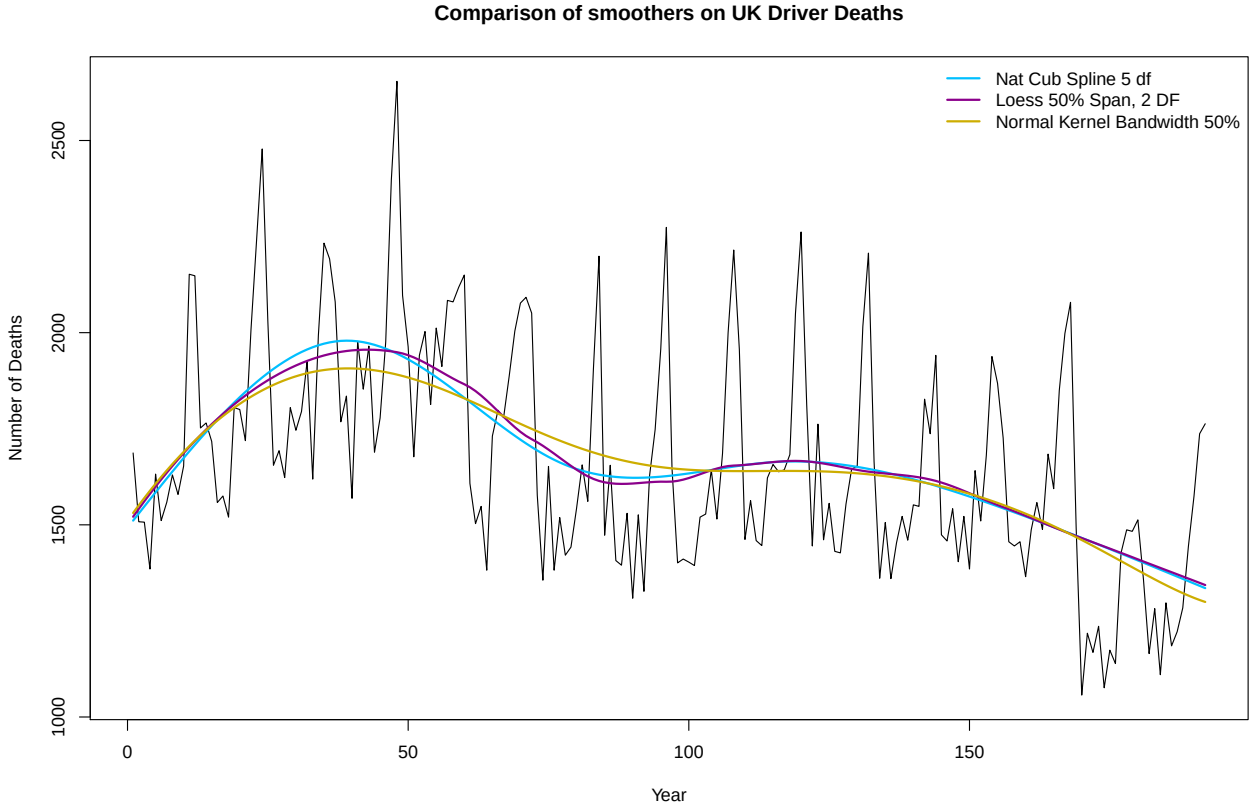
Comparison between best chosen normal kernel and optimal kernel



Comparing the normal kernel with best chosen and optimal bandwidths, we can say that the former seems to be doing a much better job at capturing the overall mean death trend while the latter succumbs to seasonal fluctuations.

Comparison between different splines

Finally, we compare the best results from each type of smoother namely `natural cubic splines`, `loess curves` and `normal kernel smoothers`.



Comparing the best results from each type of smoother, we observe that they all exhibit quite similar performance. Hence, we cannot objectively label one class of smoothers as 'best' in this case. It is important to note that for both smoothing splines and normal kernel smoothers, the optimal setting seems to model the monthly cycles better than the long-term multi-year trend.