Nitrate Levels in Avon River Over Time

2023-05-29

The following is a multivariate analysis of Nitrate levels in the Avon river in Christchurch.

```
library(gam)
```

```
## Loading required package: splines
```

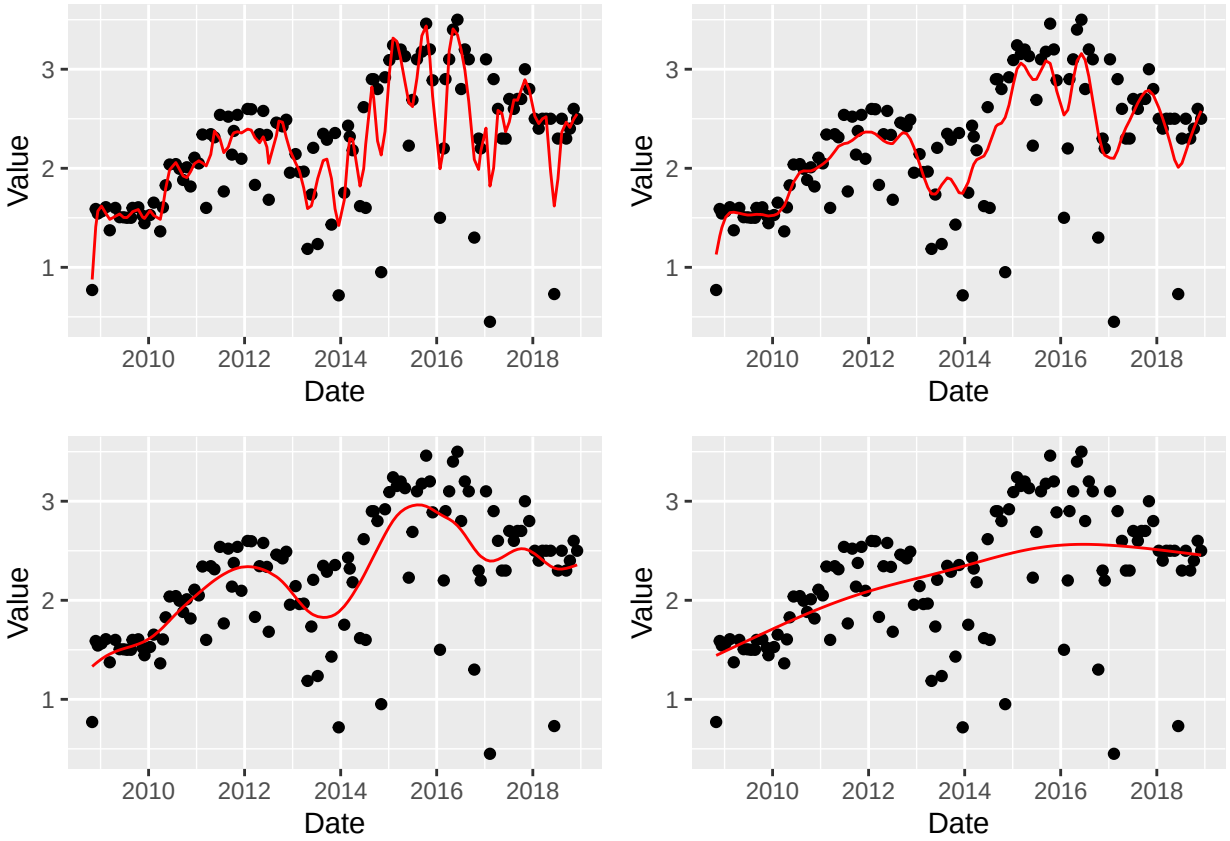
```
## Loading required package: foreach
```

```
## Loaded gam 1.22-2
```

```
library(ggplot2)
library(gridExtra)
ccc <- read.csv("CCC05.csv")
ecan <- read.csv("ECAN93.csv")
ccc$Date <- as.Date(ccc$Date, format="%d/%m/%Y")
mod1 <- gam(Value ~ s(Date, spar = 0.2), data=ccc)
mod2 <- gam(Value ~ s(Date, spar = 0.4), data=ccc)
mod3 <- gam(Value ~ s(Date, spar = 0.63), data=ccc)
mod4 <- gam(Value ~ s(Date, spar = 1), data=ccc)
ccc$predict1 <- predict(mod1)
ccc$predict2 <- predict(mod2)
ccc$predict3 <- predict(mod3)
ccc$predict4 <- predict(mod4)

plot1 <- ggplot(ccc, aes(x=Date, y=Value)) + geom_point() + geom_line(aes(x=Date, y=predict1),
plot2 <- ggplot(ccc, aes(x=Date, y=Value)) + geom_point() + geom_line(aes(x=Date, y=predict2),
plot3 <- ggplot(ccc, aes(x=Date, y=Value)) + geom_point() + geom_line(aes(x=Date, y=predict3),
plot4 <- ggplot(ccc, aes(x=Date, y=Value)) + geom_point() + geom_line(aes(x=Date, y=predict4),

grid.arrange(plot1, plot2, plot3, plot4, nrow=2)
```




```
summary(mod3)
```

```
##
## Call: gam(formula = Value ~ s(Date, spar = 0.63), data = ccc)
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.94583 -0.12562  0.09761  0.25403  0.74684
##
## (Dispersion Parameter for gaussian family taken to be 0.2292)
##
##     Null Deviance: 49.5862 on 122 degrees of freedom
## Residual Deviance: 25.3442 on 110.5726 degrees of freedom
## AIC: 181.6184
##
## Number of Local Scoring Iterations: NA
##
## Anova for Parametric Effects
##                           Df Sum Sq Mean Sq F value    Pr(>F)
## s(Date, spar = 0.63)    1.00 12.219 12.2190  53.309 4.654e-11 ***
## Residuals             110.57 25.344  0.2292
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Anova for Nonparametric Effects
##                       Npar Df Npar F    Pr(F)
```

```
## (Intercept)
## s(Date, spar = 0.63)    10.4 5.0305 3.79e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The AIC is not relevant because an over fitted model will have a low AIC

P-values are not relevant. An insignificant pvalue means a model may be over smoothed while a significant p value does not tell us if the model is over fitted or if it's a good fit.

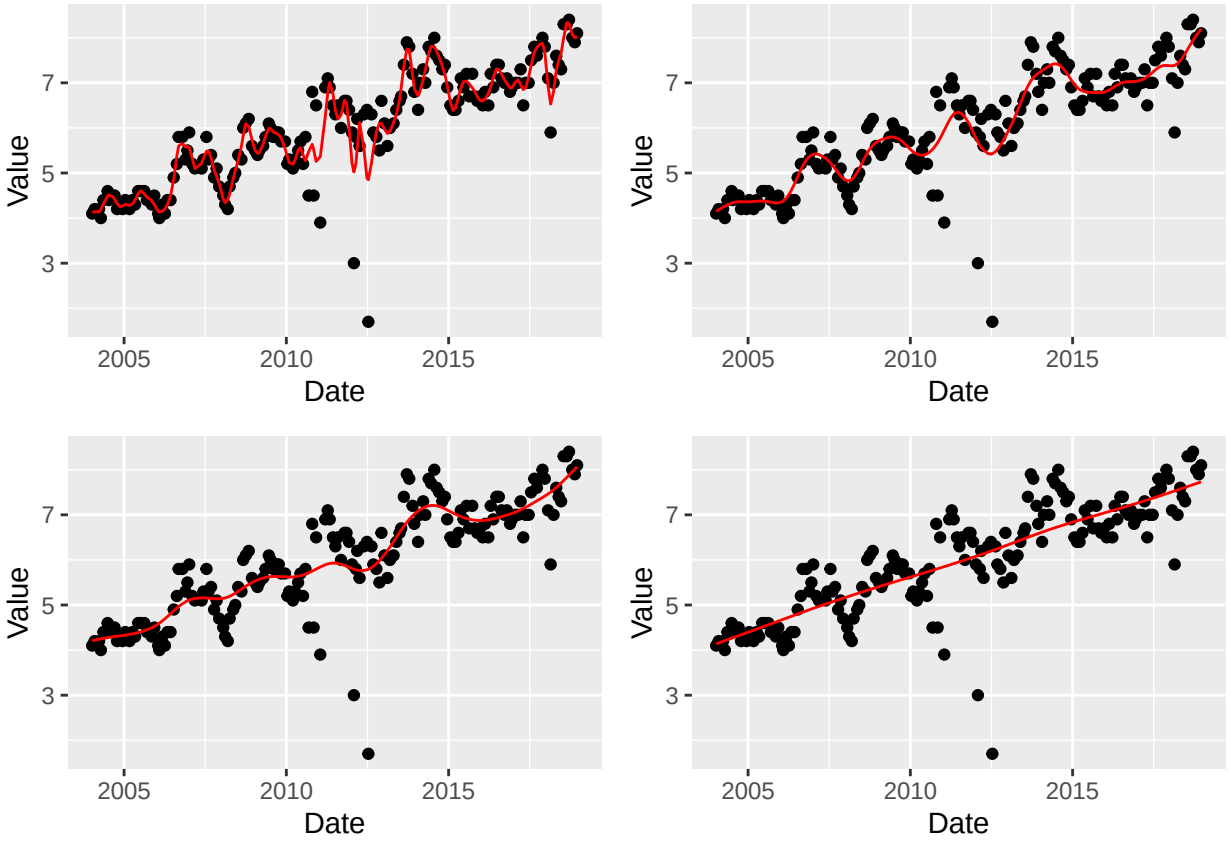
Mod1 and Mod2 over fit the data as the model captures features that may not actually be present, this causes a lot of noise in the data. Mod4 over smooths the data as the model does not capture the important features. Mod3 has a spar of 0.63. This looks like a good fit as it only captures the significant features in the model without capturing trends they may not actually be present in the data.

Mod3 shows the Nitrate levels in the Avon river are steadily increases over time. There appears to be a seasonal trend as there are clear peaks in 2012 and mid 2015. There are clear troughs in mid 2013 and mid 2018.

```
ecan$Date <- as.Date(ecan$Date, format="%d/%m/%Y")
mod5 <- gam(Value ~ s(Date, spar = 0.2), data=ecan)
mod6 <- gam(Value ~ s(Date, spar = 0.55), data=ecan)
mod7 <- gam(Value ~ s(Date, spar = 0.7), data=ecan)
mod8 <- gam(Value ~ s(Date, spar = 1), data=ecan)

ecan$predict5 <- predict(mod5)
ecan$predict6 <- predict(mod6)
ecan$predict7 <- predict(mod7)
ecan$predict8 <- predict(mod8)

plot5 <- ggplot(ecan, aes(x=Date, y=Value)) + geom_point() + geom_line(aes(x=Date, y=predict5),
plot6 <- ggplot(ecan, aes(x=Date, y=Value)) + geom_point() + geom_line(aes(x=Date, y=predict6),
plot7 <- ggplot(ecan, aes(x=Date, y=Value)) + geom_point() + geom_line(aes(x=Date, y=predict7),
plot8 <- ggplot(ecan, aes(x=Date, y=Value)) + geom_point() + geom_line(aes(x=Date, y=predict8),
grid.arrange(plot5, plot6, plot7, plot8, nrow=2)
```




```
summary(mod2)
```

```
##
## Call: gam(formula = Value ~ s(Date, spar = 0.4), data = ccc)
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.65041 -0.09580  0.07344  0.23888  0.99659
##
## (Dispersion Parameter for gaussian family taken to be 0.2128)
##
##     Null Deviance: 49.5862 on 122 degrees of freedom
## Residual Deviance: 19.8108 on 93.1011 degrees of freedom
## AIC: 186.2642
##
## Number of Local Scoring Iterations: NA
##
## Anova for Parametric Effects
##                          Df Sum Sq Mean Sq F value    Pr(>F)
## s(Date, spar = 0.4)   1.000 12.219 12.2190  57.423 2.544e-11 ***
## Residuals            93.101 19.811  0.2128
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Anova for Nonparametric Effects
##                      Npar Df Npar F     Pr(F)
```

```
## (Intercept)
## s(Date, spar = 0.4)    27.9 2.9573 5.109e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Mod5 overfits the data, it captures the main feautes in the data but it also captures features that may not actually exist. Mod7 and Mod8 both over-smooth the data as the main features are not captured. Mod6 has a spar of 0.55. This appears to be the best model as it captures the main features of the data without creating noise by capturing features that may not exist.

Mod6 shows the Nitrate levels in the Avon river are increasing over time on average. There appears to be a seasonal effect as the Nirtate levels in the river have a clear peak followed by a trough. This may be due to changing conditions in the environment such as temperature.