I'm already on the chapter 5 of the book and starting to try a couple of use case scenarios.

I tried increasing the validation period to 50 to get to a point where the data is on a “before” spike trend or to simulate something where the series suddenly increase.

When I did that the whole forecast just suddenly didn't work. It seems like it just goes sideways and didn't recognize the new data.

I remember on the previous email thread with the book authors that the STLM was mentioned because it is more robust to outliers and so I did a little digging and I found some references about it here <http://stats.stackexchange.com/questions/140163/timeseries-analysis-procedure-and-methods-using-r>  and here <http://wombat2016.org/slides/rob.pdf>

Then I gave it a try and STLM seems like a better model for sudden spikes. I tried it on both 36 months and 50 months and STLM seems to be giving the better and tighter trend. The book examples on exponential smoothing uses the ETS function. Both STLM and ETS are from the Forecast package.

Also, all the graphs have the trailing (dashed line) and center (solid line) moving averages.

See the examples below on comparing the ETS and STLM with 36 and 50 months of validation period.

# ETS (Holt Winter’s) vs STLM

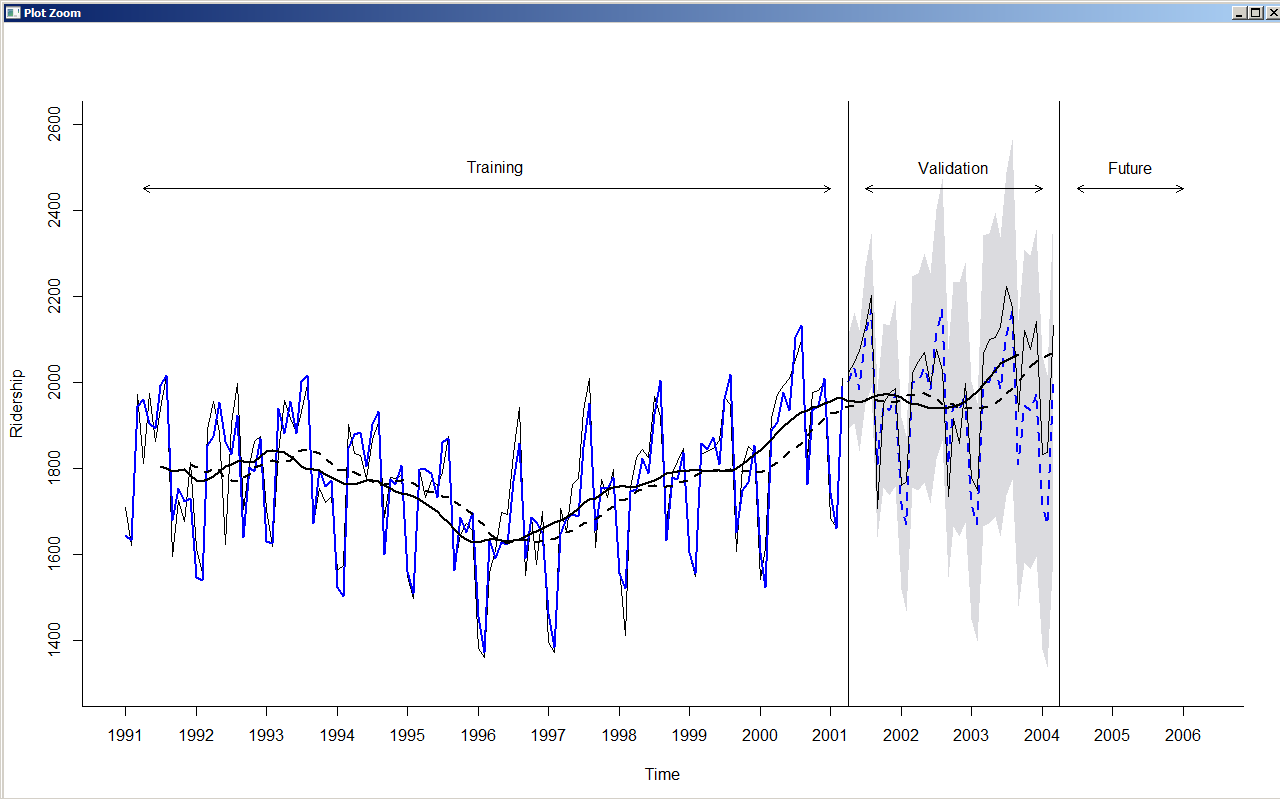
## Code

* Just change the nValid to 36 and 50 to get the 36 and 50 months graphs

|  |
| --- |
| library("forecast")  library("zoo")  # import data and create time series object  Amtrak.data <- read.csv("Amtrak data.csv")  ridership.ts <- ts(Amtrak.data$Ridership, start = c(1991, 1), end = c(2004, 3), freq = 12)  # separate training and validation data  nValid <- 36 # set to 36 or 50 months  nTrain <- length(ridership.ts) - nValid  train.ts <- window(ridership.ts, start = c(1991, 1), end = c(1991, nTrain))  valid.ts <- window(ridership.ts, start = c(1991, nTrain + 1), end = c(1991, nTrain + nValid))  # holt winters exponential smoothing  hw <- ets(train.ts, model = "AAM", restrict = FALSE)  hw.pred <- forecast(hw, h = nValid, level = 95) # set 0 for no confidence band  # stlm  mod\_stl = stlm(train.ts)  mod\_stl.pred = forecast(mod\_stl, h = nValid)  # trailing and center moving average  ma.trailing <- rollmean(ridership.ts, k = 12, align = "right")  ma.centered <- ma(ridership.ts, order = 12)  ### ETS (Holt Winters)  plot(hw.pred, ylim = c(1300, 2600), ylab = "Ridership", xlab = "Time", bty = "l", xaxt = "n", xlim = c(1991,2006.25), main = "", flty = 2)  axis(1, at = seq(1991, 2006, 1), labels = format(seq(1991, 2006, 1)))  lines(hw.pred$fitted, lwd = 2, col = "blue")  lines(valid.ts)  lines(c(2004.25 - 3, 2004.25 - 3), c(0, 3500))  lines(c(2004.25, 2004.25), c(0, 3500))  text(1996.25, 2500, "Training")  text(2002.75, 2500, "Validation")  text(2005.25, 2500, "Future")  arrows(2004 - 3, 2450, 1991.25, 2450, code = 3, length = 0.1, lwd = 1,angle = 30)  arrows(2004.5 - 3, 2450, 2004, 2450, code = 3, length = 0.1, lwd = 1,angle = 30)  arrows(2004.5, 2450, 2006, 2450, code = 3, length = 0.1, lwd = 1, angle = 30)  lines(ma.trailing, lwd = 2, lty = 2)  lines(ma.centered, lwd = 2)  ### STLM  plot(mod\_stl.pred, ylim = c(1300, 2600), ylab = "Ridership", xlab = "Time", bty = "l", xaxt = "n", xlim = c(1991,2006.25), main = "", flty = 2)  axis(1, at = seq(1991, 2006, 1), labels = format(seq(1991, 2006, 1)))  lines(mod\_stl.pred$fitted, lwd = 2, col = "red")  lines(valid.ts)  lines(c(2004.25 - 3, 2004.25 - 3), c(0, 3500))  lines(c(2004.25, 2004.25), c(0, 3500))  text(1996.25, 2500, "Training")  text(2002.75, 2500, "Validation")  text(2005.25, 2500, "Future")  arrows(2004 - 3, 2450, 1991.25, 2450, code = 3, length = 0.1, lwd = 1,angle = 30)  arrows(2004.5 - 3, 2450, 2004, 2450, code = 3, length = 0.1, lwd = 1,angle = 30)  arrows(2004.5, 2450, 2006, 2450, code = 3, length = 0.1, lwd = 1, angle = 30)  lines(ma.trailing, lwd = 2, lty = 2)  lines(ma.centered, lwd = 2) |

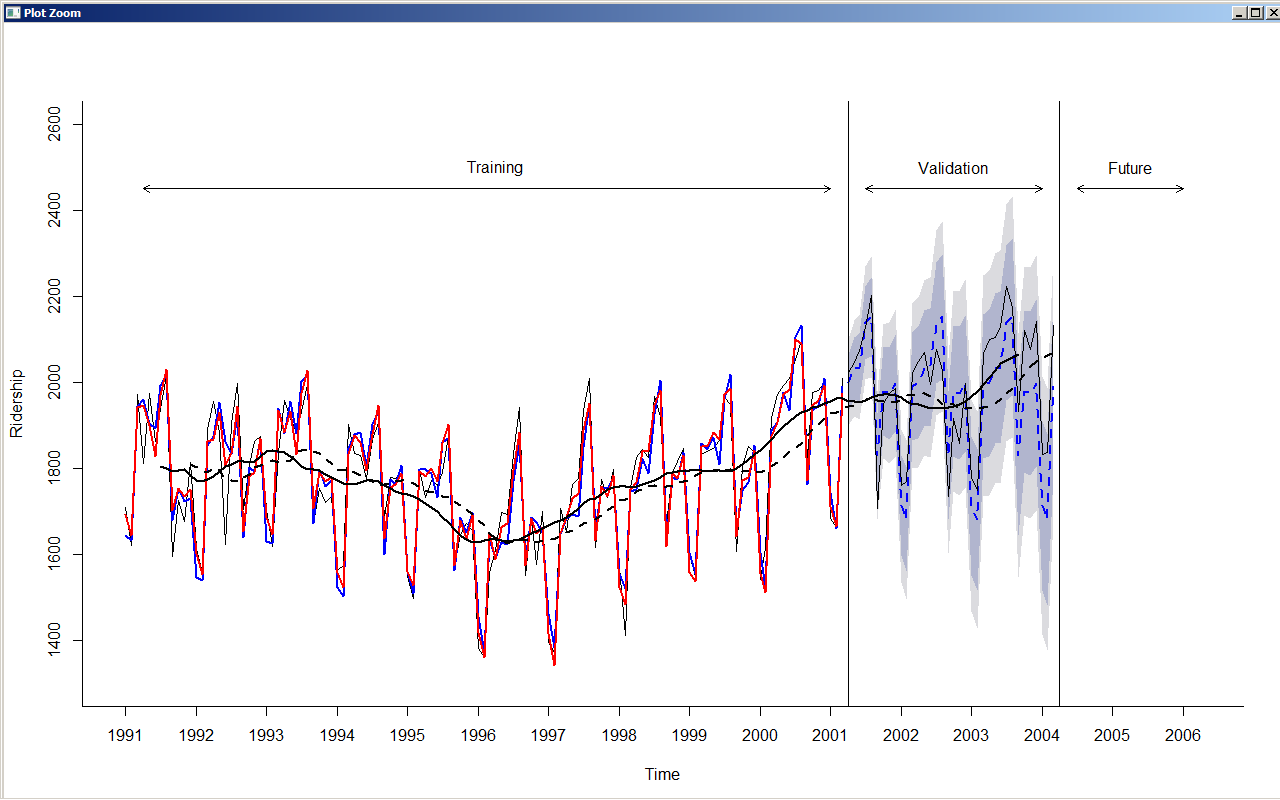
## 36 months validation

ETS (Holt Winter’s)



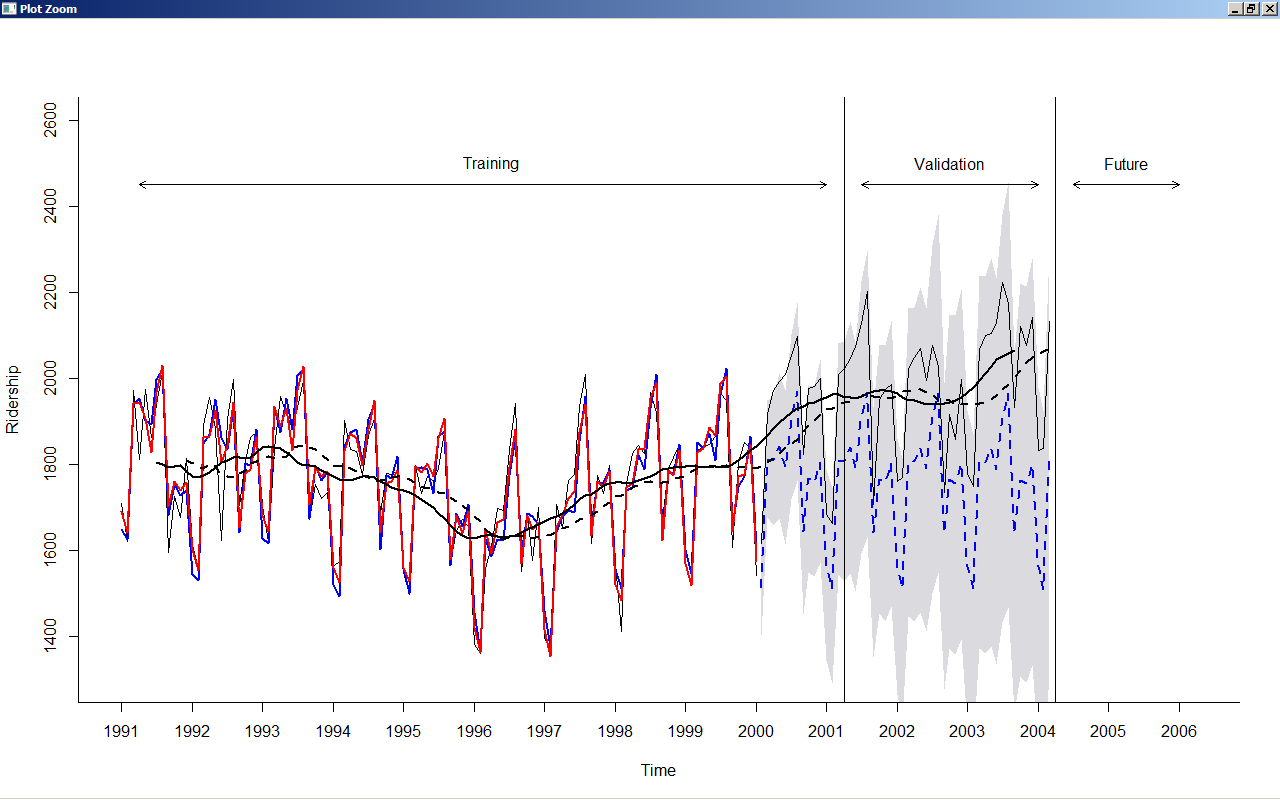
Vs

STLM



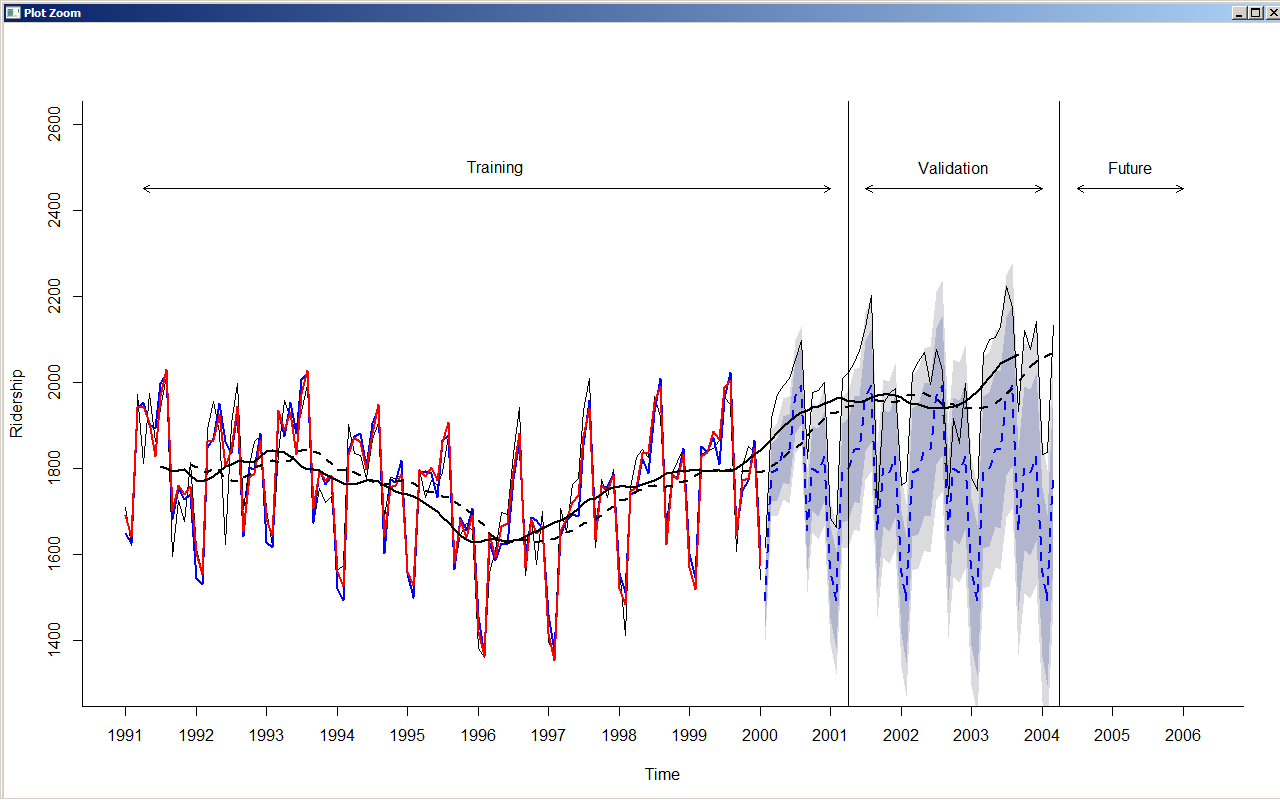
## 50 months validation

ETS (Holt Winter’s)



Vs

STLM



# But then…

Is there a model that could behave like the "trailing moving average" line? I mean here the "trailing moving average" clearly follows the trend of the series but of course the data is known. But we want our forecast to be as accurate like that. Is there such a model? or am I missing another step here to address the sudden spike?

# Time Series Linear Regression

I played around with the code on chapter 6 of the book (Regression based models – table 6.5) and it seems to work, it followed the trend when I input 50 months on the validation period. So the regression based model seems to be the one. That magic formula highlighted below seems to do the trick on sudden spikes on series.

And I also tried putting 1month, 6months, and then 12months and they all seem to behave the same although the confidence bands are different.

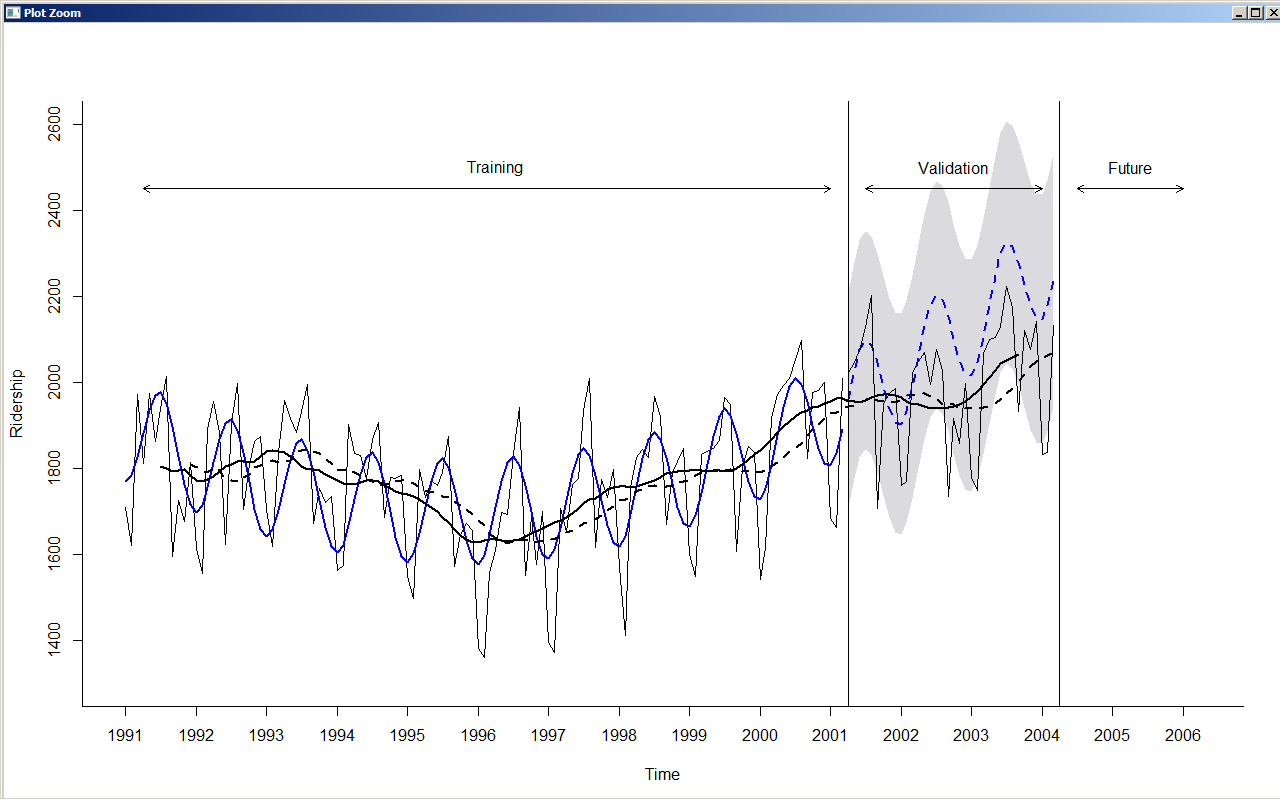
## Code

* Just change the nValid to 36 and 50 to get the 36 and 50 months graphs
* The “magic” here is the formula used to create the “train.lm.trig”

|  |
| --- |
| library("forecast")  library("zoo")  # import data and create time series object  Amtrak.data <- read.csv("Amtrak data.csv")  ridership.ts <- ts(Amtrak.data$Ridership, start = c(1991, 1), end = c(2004, 3), freq = 12)  # separate training and validation data  nValid <- 36 # set to 36 or 50 months  nTrain <- length(ridership.ts) - nValid  train.ts <- window(ridership.ts, start = c(1991, 1), end = c(1991, nTrain))  valid.ts <- window(ridership.ts, start = c(1991, nTrain + 1), end = c(1991, nTrain + nValid))  ### Time Series Linear Regression  train.lm.trig <- tslm(train.ts ~ trend + I(trend^2) + I(sin(2\*pi\*trend/12)) + I(cos(2\*pi\*trend/12)))  #summary(train.lm.trig)  train.lm.trig.pred <- forecast(train.lm.trig, h = nValid, level = 95)  # trailing and center moving average  ma.trailing <- rollmean(ridership.ts, k = 12, align = "right")  ma.centered <- ma(ridership.ts, order = 12)  ### Time Series Linear Regression  plot(train.lm.trig.pred, ylim = c(1300, 2600), ylab = "Ridership", xlab = "Time", bty = "l", xaxt = "n", xlim = c(1991,2006.25), main = "", flty = 2)  axis(1, at = seq(1991, 2006, 1), labels = format(seq(1991, 2006, 1)))  lines(train.lm.trig.pred$fitted, lwd = 2, col = "blue")  lines(valid.ts)  lines(c(2004.25 - 3, 2004.25 - 3), c(0, 3500))  lines(c(2004.25, 2004.25), c(0, 3500))  text(1996.25, 2500, "Training")  text(2002.75, 2500, "Validation")  text(2005.25, 2500, "Future")  arrows(2004 - 3, 2450, 1991.25, 2450, code = 3, length = 0.1, lwd = 1,angle = 30)  arrows(2004.5 - 3, 2450, 2004, 2450, code = 3, length = 0.1, lwd = 1,angle = 30)  arrows(2004.5, 2450, 2006, 2450, code = 3, length = 0.1, lwd = 1, angle = 30)  lines(ma.trailing, lwd = 2, lty = 2)  lines(ma.centered, lwd = 2) |

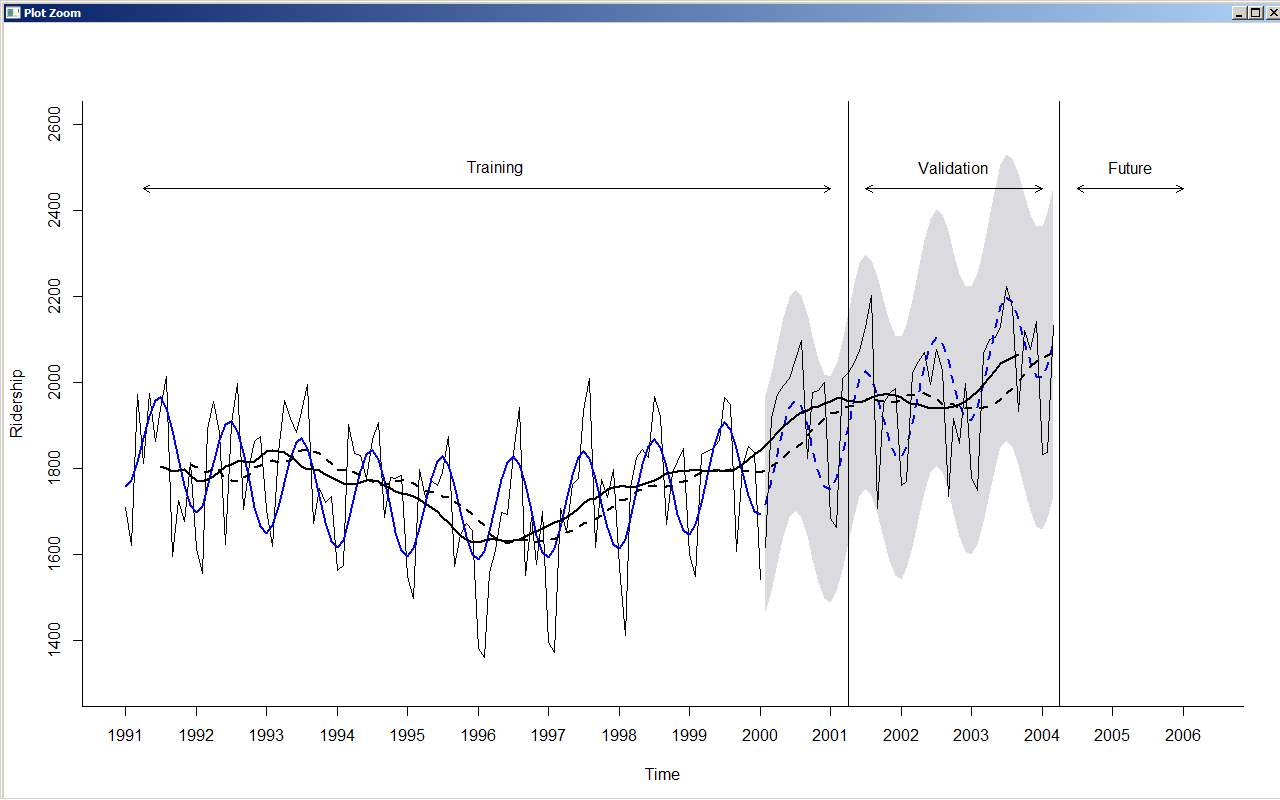
## 36 months validation

12 months cycle

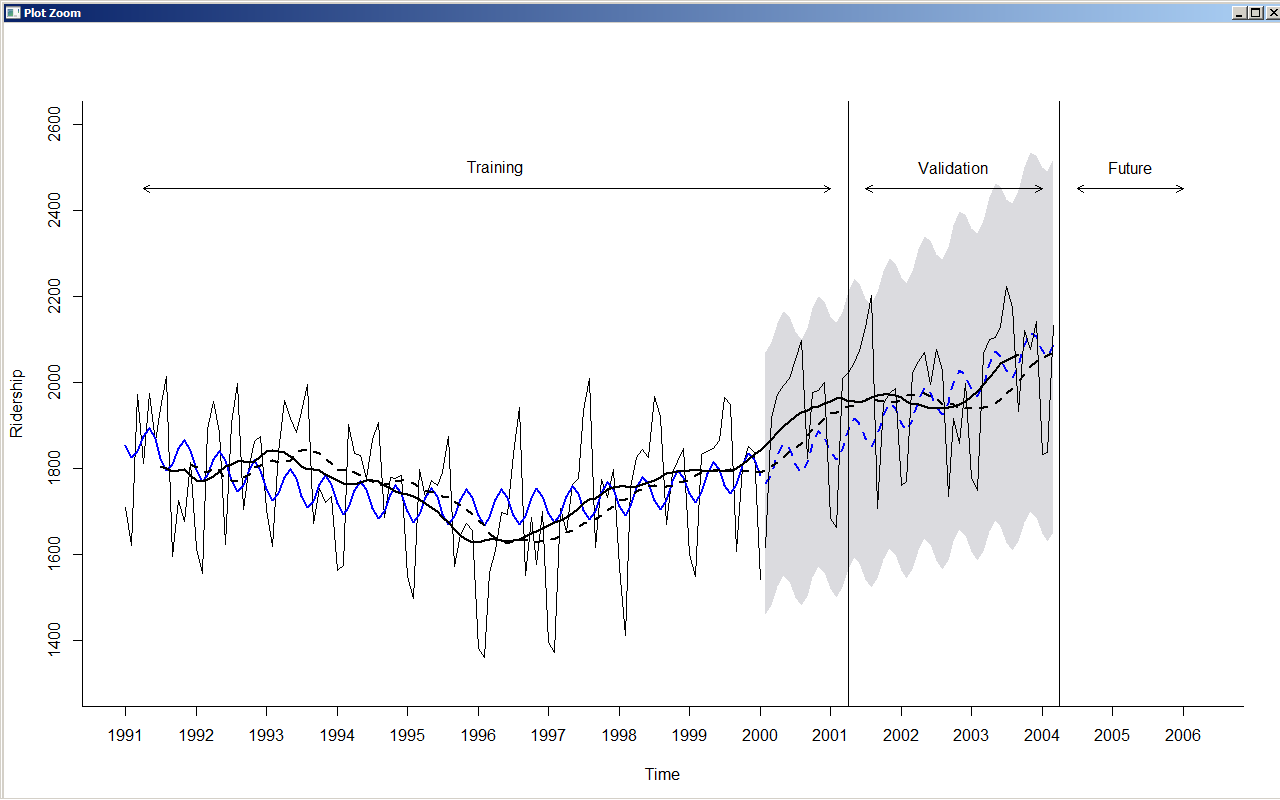


## 50 months validation

12 months cycle



6 months cycle



1 month cycle

