MSwM examples

Jose A. Sanchez-Espigares, Alberto Lopez-Moreno
Dept. of Statistics and Operations Research
UPC-BarcelonaTech

January 4, 2013

Abstract

Two examples are described to illustrate the use of the `MSwM` package. First, a simulated dataset is modeled in detail. Next, Markov Switching Models are fitted to a real dataset with a discrete response variable. The main methods and graphical representations are used to validate different approaches to model these datasets.

1 Simulated Example

The `example` data is a simulated data set to show how `msmFit` can detect the presence of two different regimes: one in which the response variable is highly correlated and other in which the response only depends on an exogenous variable `x`. The autocorrelated observations are in the intervals 1 to 100, 151 to 180 and 251 to 300. The real models for each regime are:

$$y_t = \begin{cases} 8 + 2x_t + \varepsilon_t^{(1)} & \epsilon_t^{(1)} \sim N(0,1) \quad t = 101:150, 181:250 \\ 1 + 0.9y_{t-1} + \varepsilon_t^{(2)} & \epsilon_t^{(2)} \sim N(0,0.5) \quad t = 1:100, 151:180, 251:300 \end{cases}$$

```
> data(example)
```

The plot in Fig.1 shows that in the intervals where does not exist autocorrelation the response variable `y` has a similar behaviour as the covariant `x`. A linear model is fitted to study how the covariate `x` explains the variable response `y`.

```
> mod=lm(y~x,example)
> summary(mod)

Call:
lm(formula = y ~ x, data = example)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.8998 -0.8429 -0.0427  0.7420  4.0337 
```

```
> plot(ts(example))
```

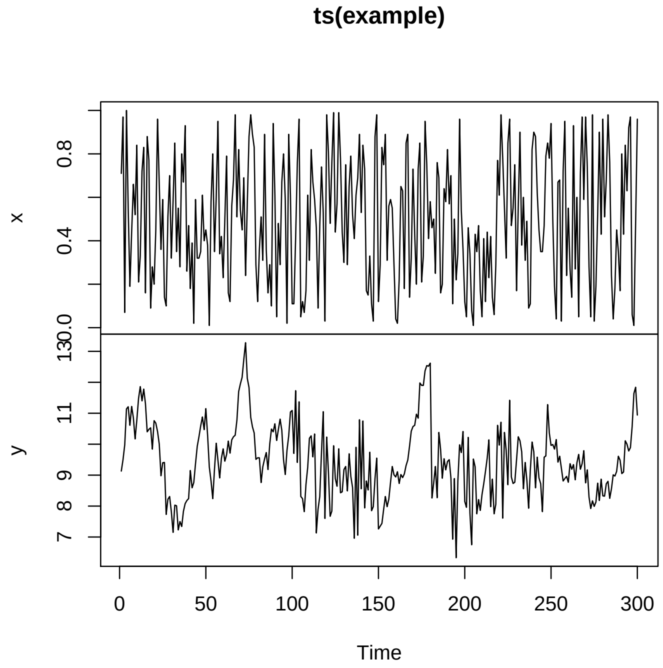

Figure 1: Simulated data. The y variable is the response variable and there are two periods in which this depends on the x covariate

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   9.0486     0.1398  64.709  < 2e-16 ***
x             0.8235     0.2423   3.398  0.00077 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.208 on 298 degrees of freedom
Multiple R-squared: 0.03731,        Adjusted R-squared: 0.03408 
F-statistic: 11.55 on 1 and 298 DF,  p-value: 0.0007701
```

The covariate is really significant but the data behaviour is very bad explained by the model. The plot of the linear model residuals in Fig.1 indicates that their autocorrelation is significant. The diagnostics plots for the residuals (Fig.2) confirm that they does not seem to be white noise and that they have autocorrelation. Next, a Autoregressive Markov Switching Model (MSM-AR) is fitted to the data. The order for the autoregressive part is set to one. In order to indicate that all the parameters can be different in both periods, the switching parameter (sw) is set to a vector with four components with value equal to TRUE. The last value when fitting a linear model is referred to the residual

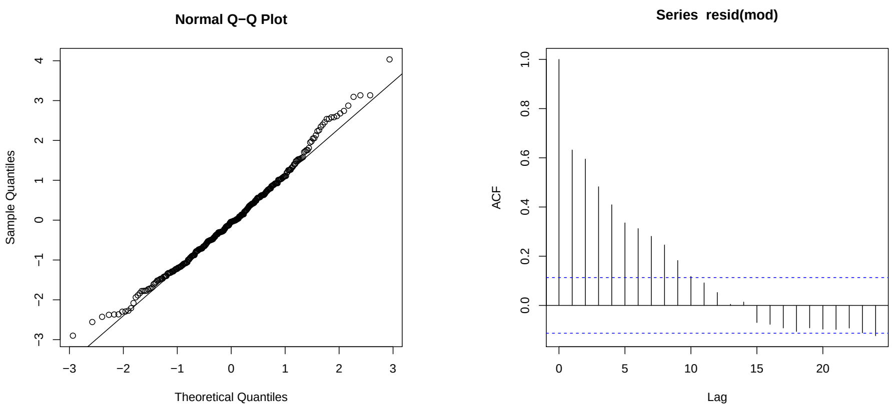

Figure 2: Normal Probability plot and Autocorrelation Function of the residuals from the linear model

standard deviation. There are some options to control the estimation process, like a logical parameter to indicate whether parallelization of the process is done or not.

```
> mod.mswm=msmFit(mod,k=2,p=1,sw=c(T,T,T,T),control=list(parallel=F))
> summary(mod.mswm)

Markov Switching Model

Call: msmFit(object = mod, k = 2, sw = c(T, T, T, T), p = 1, control = list(parallel = F))

       AIC     BIC    logLik
  637.0736 693.479 -312.5368

Coefficients:

Regime 1
---------
               Estimate Std. Error t value  Pr(>|t|)
(Intercept)(S)   0.8417     0.3026  2.7816  0.005409 **
x(S)            -0.0533     0.1348 -0.3954  0.692548
y_1(S)           0.9208     0.0306 30.0915 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5034675
Multiple R-squared: 0.8375

Standardized Residuals:
          Min            Q1           Med            Q3           Max
-1.5153666657 -0.0906543311  0.0001873641  0.1656717256  1.2020898986
```

```
Regime 2
---------
               Estimate Std. Error t value  Pr(>|t|)
(Intercept)(S)   8.6393     0.7256 11.9064 < 2.2e-16 ***
x(S)             1.8771     0.3107  6.0415 1.527e-09 ***
y_1(S)          -0.0569     0.0799 -0.7121    0.4764
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9339683
Multiple R-squared: 0.2408

Standardized Residuals:
        Min          Q1         Med          Q3         Max
-2.31102193 -0.03317756  0.01034139  0.04509105  2.85245598

Transition probabilities:
           Regime 1   Regime 2
Regime 1 0.98499728 0.02290884
Regime 2 0.01500272 0.97709116
```

The model `mod.mswm` has a regime where the covariant `x` is very significant and in the other regime the autocorrelation variable is very significant too. In both, the R-squared have high values. Finally, the transition probabilities matrix has high values which indicate that is difficult to change from on regime to the other. The model detect perfectly the periods of each state. The residuals look like to be white noise and they fit to the Normal Distribution. Moreover, the autocorrelation has disappeared.

```
>    plotProb(mod.mswm,which=1)
```

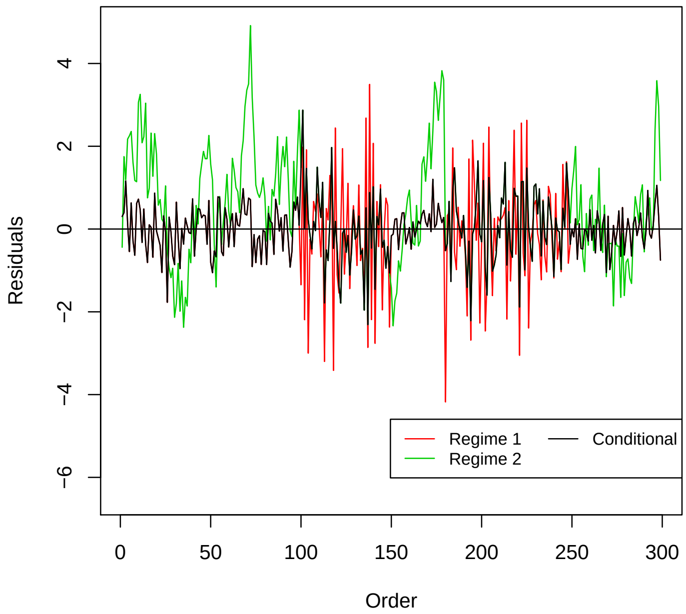

Figure 3: Residuals form the Autoregressive MSM model

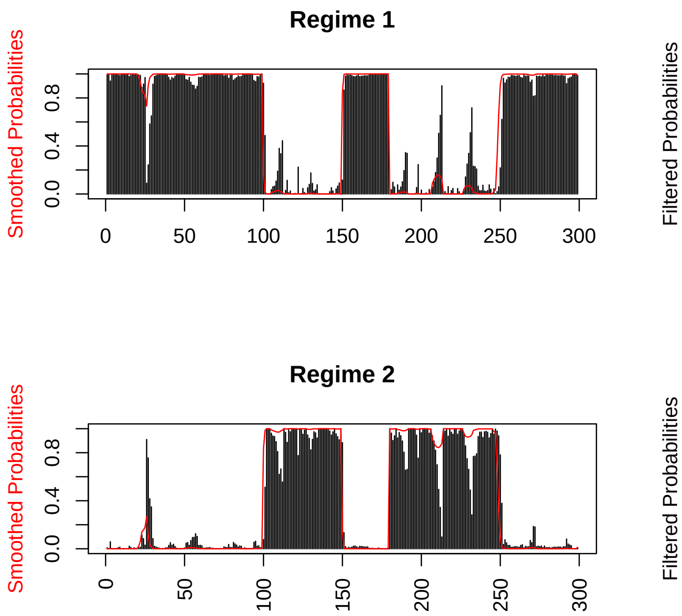

Figure 5: Filtered and Smoothed Probabilities for both regimes in the MSM-AR model

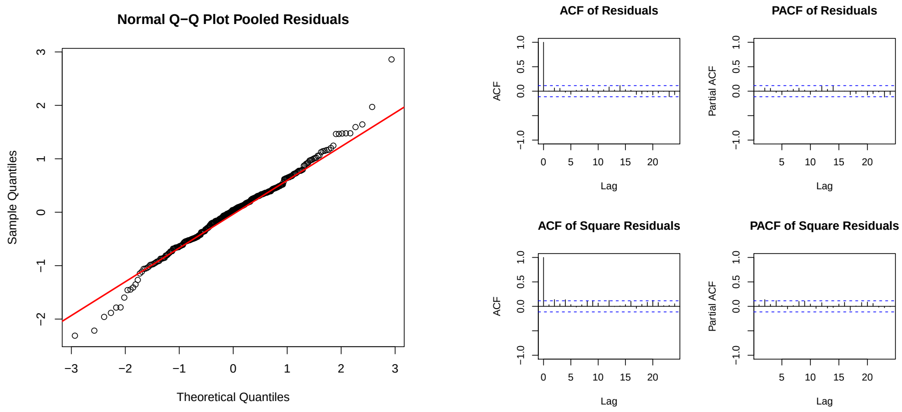

Figure 4: Normal Probability plot and Autocorrelation Function of the residuals from the Complete MSwM model. They are obtained by using the `plotDiag` method

```
>     plotProb(mod.mswm,which=2)
```

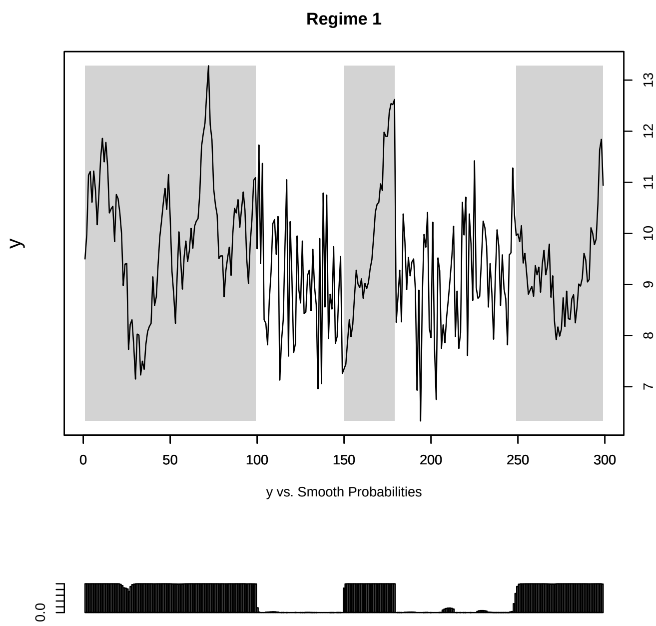

Figure 6: Response variable indicating which observations are associated to regime 1

The graphics show that the periods for each regime have been detected perfectly.

```
>     plotReg(mod.mswm,expl="x")
```

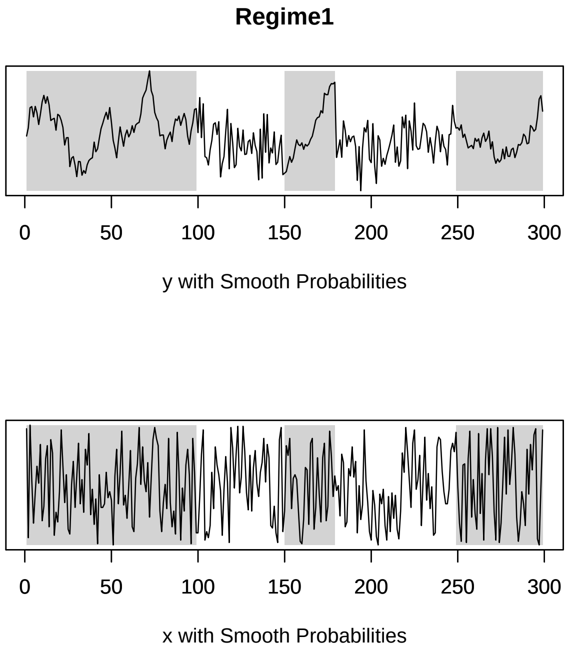

Figure 7: Relationship between x and y locating the two regimes

2 Daily Traffic Casualties by car accidents in Spain

The `traffic` data (Fig.8) contains the daily number of deaths in traffic accidents in Spain during the year 2010, the average daily temperature and the daily sum of precipitations. The interest of this data is to study the relation between the number of deaths with the climate conditions. We illustrate the use of a Generalized Markov Switching Model in this case because there exists a different behaviour between the variables during weekends and working days. To avoid a long example, the explanations of how the functions work and repeated results are skipped.

In this example, the response variable is a counting variable. For this reason we fit a Poisson Generalized Linear Model.

```
> model=glm(NDead~Temp+Prec,traffic,family="poisson")
> summary(model)

Call:
glm(formula = NDead ~ Temp + Prec, family = "poisson", data = traffic)
```

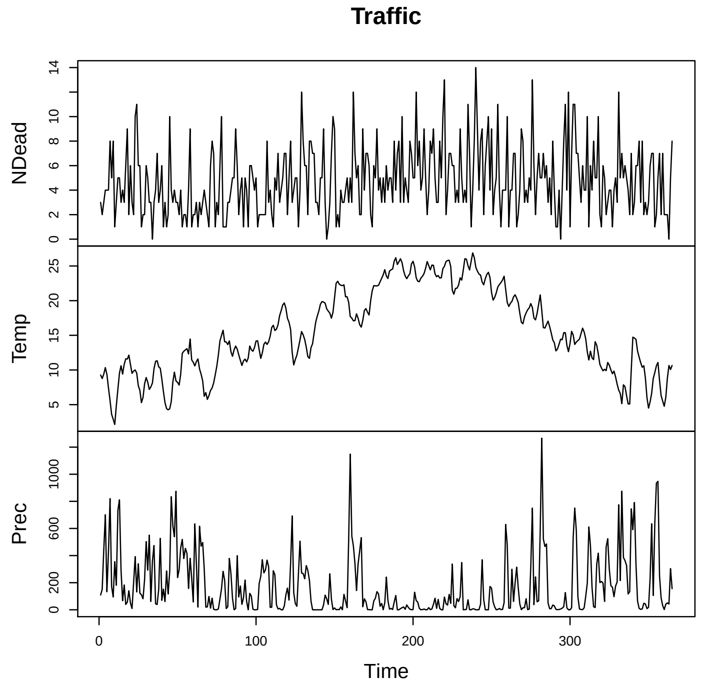

Figure 8: Traffic data: Daily traffic casualties in Spain and climate variables

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1571  -1.0676  -0.2119   0.8080   3.0629

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.1638122  0.0808726  14.391  < 2e-16 ***
Temp        0.0225513  0.0041964   5.374  7.7e-08 ***
Prec        0.0002187  0.0001113   1.964   0.0495 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 597.03  on 364  degrees of freedom
Residual deviance: 567.94  on 362  degrees of freedom
AIC: 1755.9

Number of Fisher Scoring iterations: 5
```

In the next step, the Markov Switching Model is fitted using `msmFit`. To fit a Generalized Markov Switching Model, the family parameter has to be included.

Moreover, the glm's don't have the standard deviation parameter and, because of this, the sw parameter doesn't contain its switching parameter.

```
> m1=msmFit(model,k=2,sw=c(T,T,T),family="poisson")

R Version:  R version 2.15.2 (2012-10-26)

> summary(m1)

Markov Switching Model

Call: msmFit(object = model, k = 2, sw = c(T, T, T), family = "poisson")

       AIC      BIC    logLik
  1713.877 1772.676 -850.9387

Coefficients:

Regime 1
---------
               Estimate Std. Error t value Pr(>|t|)
(Intercept)(S)   1.5658     0.1587  9.8664  < 2e-16 ***
Temp(S)          0.0194     0.0080  2.4250  0.01531 *
Prec(S)          0.0004     0.0002  2.0000  0.04550 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Regime 2
---------
               Estimate Std. Error t value  Pr(>|t|)
(Intercept)(S)   0.7649     0.1814  4.2166  2.48e-05 ***
Temp(S)          0.0288     0.0083  3.4699 0.0005207 ***
Prec(S)          0.0002     0.0002  1.0000 0.3173105
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Transition probabilities:
          Regime 1  Regime 2
Regime 1 0.5086262 0.2712413
Regime 2 0.4913738 0.7287587
```

Both states have significant covariates, but the precipitation covariate is only significant in one of the two.

```
> intervals(m1)

Aproximate intervals for the coefficients. Level= 0.95

(Intercept):
```

```
                Lower Estimation    Upper
Regime 1 1.2547681   1.565849 1.876929
Regime 2 0.4092239   0.764852 1.120480


Temp:
                  Lower Estimation      Upper
Regime 1 0.003667658 0.01939759 0.03512753
Regime 2 0.012627702 0.02884962 0.04507153


Prec:
                    Lower   Estimation        Upper
Regime 1 -4.927615e-05 0.0004105931 0.0008704623
Regime 2 -1.865816e-04 0.0001846775 0.0005559366
```

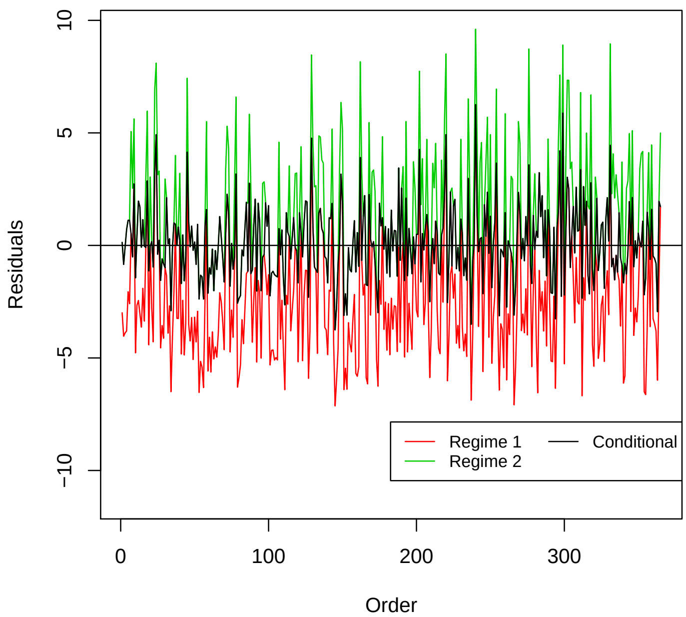

Figure 9: Residuals form the Autoregressive MSM model

The Pearson residuals from Fig. 9 are calculated from an object of class 'MSM.glm' because the model is an extension of a General Linear Model. The residuals have the classical structure of white noise. The residuals aren't autocorrelated but they don't fit very well to a Normal Distribution. However, normality of the Pearson residuals is not a critical condition for generalized linear model validation.

```
>     plotProb(m1,which=2)
```

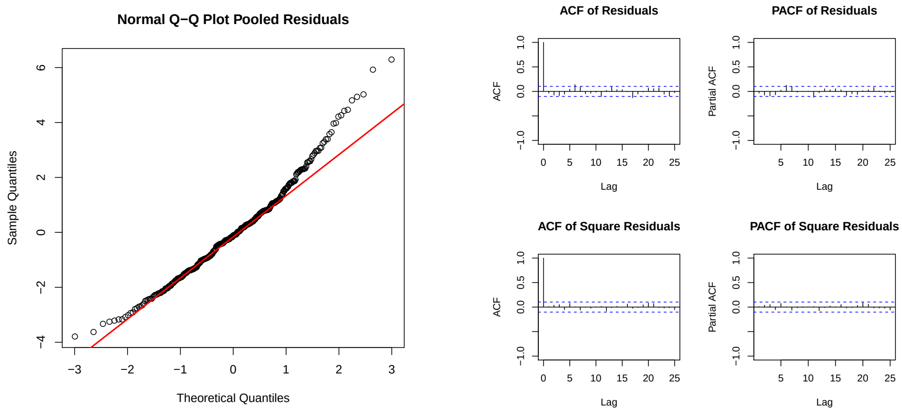

Figure 10: Normal Probability plot and Autocorrelation Function of the residuals from the Autoregressive MSwM model. They are obtained by using the `plotDiag` method

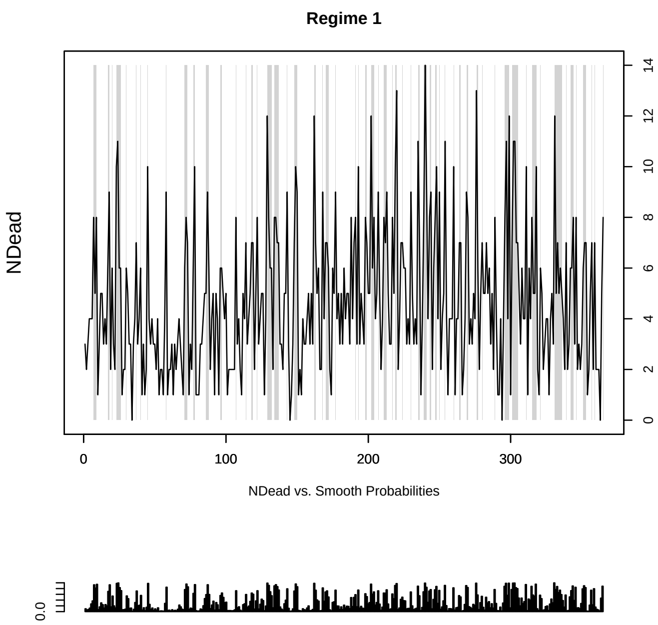

Figure 11: Response variable indicating which observations are associated to regime 1

Using the function `plotProb` we can see how the regimes are distributed in shorts periods because the bigger one contains basically working days.