LEBEC PM-2.5 DATA WITH AR RESIDUALS

JAN DE LEEUW

ABSTRACT. Meet the abstract. This is the abstract.

1. INTRODUCTION

The data are hourly PM-2.5 measurements at the Lebec Air Monitor, which was installed and operated by ARB from 02-21-06 to 02-07-07. Data and auxiliary files can be downloaded from

http://idisk.mac.com/jdeleeuw-Public/lebec

The file `lebpm.R` contains the data in the form of a 352×24 dataframe, with the dates as row labels.

```
> source("lebpm.R")
> dim(lebpm)
[1] 352  24
> typeof(lebpm)
[1] "list"
```

To give an idea how the data look, we list the first five rows.

```
> lebpm[1:5,]
          0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
02-21-06 NA NA NA NA NA NA NA NA NA NA NA NA NA  7  6  6  9  9 12 13 11 18  5 15
02-22-06 13  9  7 21 20 17 19 15 15 10 10  7  7  7  6  6 14 31 17 20 13 11 10 13
02-23-06 13 13 11  9 12 12 17 15 15  8  8  7  4  6  7  4  6 27 25 28  8 22 13 10
02-24-06  9 11 12 12  9 19 10 15  8 21  9  7  4  6 11 17 16 25 46 32 16 19 17 15
02-25-06 18 21 11 12 16 16 12 11 11 14  9  7  6  1  3  4 10 10  9 10 10 12 11  7
```

About 13% of the data are missing.

```
> length(which(is.na(lebpm)))/(352*24)
[1] 0.1304451
```

Date: June 1, 2008 — 20h 11min — Typeset in TIMES ROMAN.

There are 9 days which are missing completely and 182 days for which all 24 hours are available.

```
> length(which(rowSums(ifelse(is.na(lebpm),1,0))==24))
[1] 9
> length(which(rowSums(ifelse(is.na(lebpm),1,0))==0))
[1] 182
```

It is convenient to make the data into a vector (of length 8448).

```
y<-as.vector(t(as.matrix(lebpm)))
```

Moreover, we convert the dates and times to `POSIXlt` format.

```
hrs<-formatC(0:23,flag="0",width=2)
dt<-as.vector(t(outer(rownames(lebpm),hrs,function(x,y) paste(paste(x,y),":00:00",sep=""))))
x<-strptime(dt,format="%m-%d-%y %H:%M:%S",tz="PST")
```

Now `x` is a list with 9 elements, which are vectors of length 8448. The nine elements (vectors) are

```
> names(x)
[1] "sec"   "min"   "hour"  "mday"  "mon"   "year"  "wday"  "yday"  "isdst"
```

Note that we have converted all dates and times to Pacific Standard Time, so we can ignore problems with daylight savings time (and thus `isdst=0` throughout).

To make a plot, we use the `zoo` package. For this purpose it is convenient to convert the times to `POSIXct` format, the number of seconds since the beginning of 1970.

```
> library(zoo)
> z<-zoo(y,as.POSIXct(x))
> pdf("pmplot_zoo.pdf")
> plot(z)
> abline(h=65,col="RED")
> dev.off()
quartz
     2
```

Insert Figure 1 *about here*

Clearly there are many outliers, which are at least partially caused by nearby wildfires. Note that the Day Fire started September 4, 2006, and lasted for about a month. It burned 162,702 acres, but was most of the time at least 20 miles from Lebec. The Quail Fire was from August 13 to August 16, 2006. It was in Lebec, and burned 4,864 acres.The Mt. Pinos Lightning Complex Fire was from July 23 to July 30, 2006 in Frazier Park. It burned 3,179 acres. It is unclear what caused the huge spikes in December and January.

As we would expect or hourly data, the autocorrelations are high. Note the slight bumps around lag 24 (which could be the day effect) and around lag 12 (which could be the commute effect).

```
> acf(coredata(z),na.action=na.pass)
```

A similar command gives the partial autocorrelations, which emphasize the same bumps.

Insert Figure 2 *about here*

Insert Figure 3 *about here*

The next step is to fit some `arima` models. We will limit ourselves to AR models of orders $0, 1, \cdots$. We give the `tsdiag` plots for orders 0, 1, and 6. The AIC is still decreasing at order 6, although slowly, and the AR coefficients are still significant. Plots are made by

```
> tsdiag(arima(coredata(z),order=(k,0,0))
```

Insert Figure 8 *about here*

Insert Figure 9 *about here*

Insert Table 1 *about here*

Next, AR models with regressors. We make dummies for hour, month, and weekday, using the dates in `POSIXlt` format that we already have. The instructions are

```
ghr<-ifelse(outer(x[["hour"]],0:23,"=="),1,0)
```

```
gmn<-ifelse(outer(x[["mon"]],0:11,"=="),1,0)
gwk<-ifelse(outer(x[["wday"]],0:6,"=="),1,0)
gg<-cbind(gmn[,1:11],gwk[,1:6],ghr[,1:23])
ar0<-arima(coredata(z),order=c(0,0,0),xreg=gg)
ar1<-arima(coredata(z),order=c(1,0,0),xreg=gg)
> sink("ar.txt")
> ar0
> ar1
> sink()
> pdf("tsdiagr0.pdf")
> tsdiag(ar0)
> dev.off()
quartz
     2
> pdf("tsdiagr1.pdf")
> tsdiag(ar1)
> dev.off()
quartz
     2
```

The text output file is

```

Call:
arima(x = coredata(z), order = c(0, 0, 0), xreg = gg)

Coefficients:
      intercept     gg1     gg2     gg3     gg4      gg5     gg6     gg7
         5.3820  1.1054  5.3825  4.7554  7.5084  12.9662  9.7410  9.7574
s.e.     0.9396  0.7031  0.8654  0.6883  0.6934   0.7088  0.7279  0.7465
         gg8      gg9    gg10    gg11     gg12    gg13    gg14    gg15    gg16
      6.8243  14.4707  1.6462  1.9399  -0.3214  0.9783  1.8650  2.9329  2.2313
s.e.  0.7002   0.7269  0.7097  0.7372   0.5657  0.5624  0.5573  0.5505  0.5531
        gg17     gg18    gg19     gg20     gg21     gg22     gg23     gg24
      3.3132  -0.3527  1.3220  -0.9829  -0.5350  -1.1456  -0.8413  -0.4379
s.e.  0.5536   1.0272  1.0247   1.0281   1.0281   1.0264   1.0281   1.0352
        gg25    gg26     gg27     gg28     gg29     gg30     gg31     gg32
      0.7773  0.1211  -0.7930  -1.5994  -1.5240  -1.4817  -2.1421  -1.0603
s.e.  1.0264  1.0142   1.0134   1.0198   1.0291   1.0265   1.0231   1.0264
        gg33    gg34    gg35    gg36    gg37    gg38    gg39    gg40
      2.1225  0.1886  0.9467  1.3845  1.4041  1.1828  0.6219  0.1612
s.e.  1.0265  1.0198  1.0223  1.0247  1.0205  1.0197  1.0197  1.0205

sigma^2 estimated as 160.6:  log likelihood = -29079.02,  aic = 58242.03
```

```

Call:
arima(x = coredata(z), order = c(1, 0, 0), xreg = gg)

Coefficients:
         ar1  intercept     gg1     gg2     gg3     gg4      gg5      gg6
      0.6166     5.1422  1.5193  5.6495  5.2541  7.9227  13.4311  10.2074
s.e.  0.0093     1.3528  1.3967  1.7319  1.3927  1.4031   1.4173   1.4510
          gg7     gg8      gg9    gg10    gg11     gg12    gg13    gg14    gg15
      10.2547  7.2512  15.2951  2.0617  2.0114  -0.5370  0.4195  1.5080  2.4535
s.e.   1.4710  1.4028   1.4619  1.4403  1.4612   0.9486  1.0055  1.0048  0.9957
         gg16    gg17     gg18    gg19     gg20     gg21     gg22     gg23
      2.0425  2.9573  -0.3972  1.2835  -0.1565  -0.5364  -1.1165  -0.8642
s.e.  0.9939  0.9360   0.6541  0.8253   0.9158   0.9659   0.9936   1.0115
         gg24    gg25    gg26     gg27     gg28     gg29     gg30     gg31
      -0.4642  0.9124  0.2157  -0.6445  -1.4897  -1.5559  -1.5820  -2.2163
s.e.   1.0248  1.0248  1.0210   1.0223   1.0278   1.0340   1.0322   1.0292
         gg32    gg33    gg34    gg35    gg36    gg37    gg38    gg39    gg40
      -0.8270  2.3313  0.4681  1.2104  1.4593  1.4239  1.2890  0.7552  0.2528
s.e.   1.0292  1.0259  1.0166  1.0071  0.9906  0.9591  0.9091  0.8202  0.6477

sigma^2 estimated as 104.2:  log likelihood = -27540.25,  aic = 55166.5
```

We now plot the regression coefficients to see the effects of month, weekday, and hour.

```
> b0<-coef(ar0)
> b1<-coef(ar1)
> mo0<-c(0,b0[2:12])
> mo1<-c(0,b1[3:13])
> we0<-c(0,b0[13:18])
> we1<-c(0,b1[14:19])
> hr0<-c(0,b0[19:41])
> hr1<-c(0,b1[20:42])
> pdf("month_effect.pdf")
> plot(mo0,type="l",col="BLUE")
> lines(1:12,mo1,col="GREEN")
> dev.off()
> pdf("weekday_effect.pdf")
> plot(we0,type="l",col="BLUE")
> lines(1:7,we1,col="GREEN")
> dev.off()
> pdf("hour_effect.pdf")
> plot(hr0,type="l",col="BLUE")
> lines(1:24,hr1,col="GREEN")
> dev.off()
```

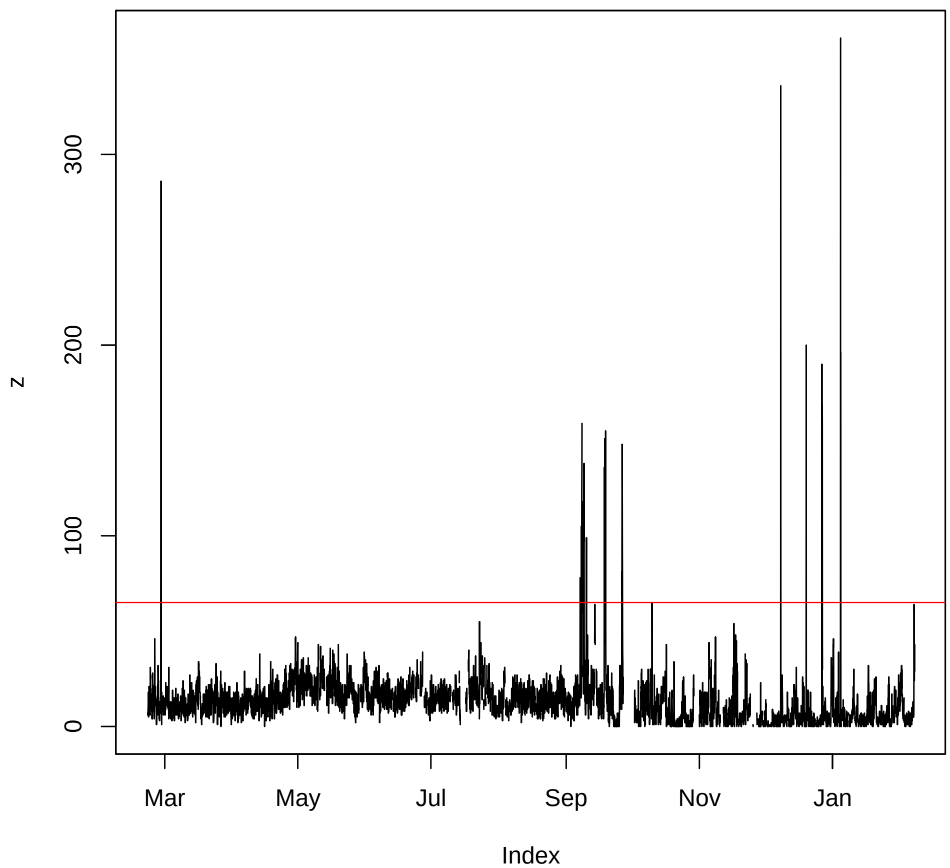

FIGURE 1. Raw Data

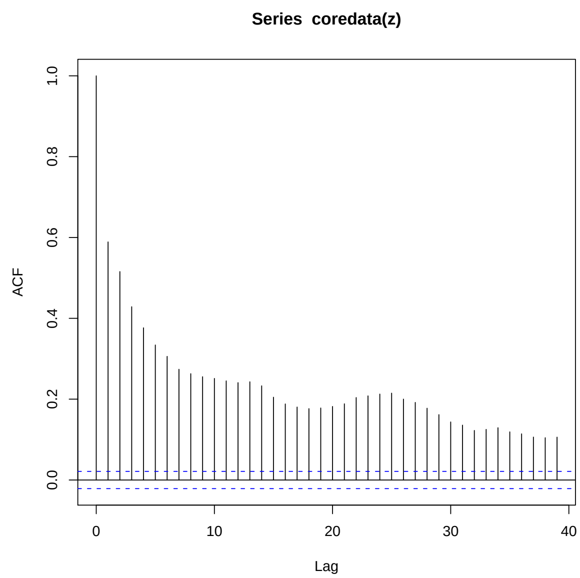

FIGURE 2. ACF

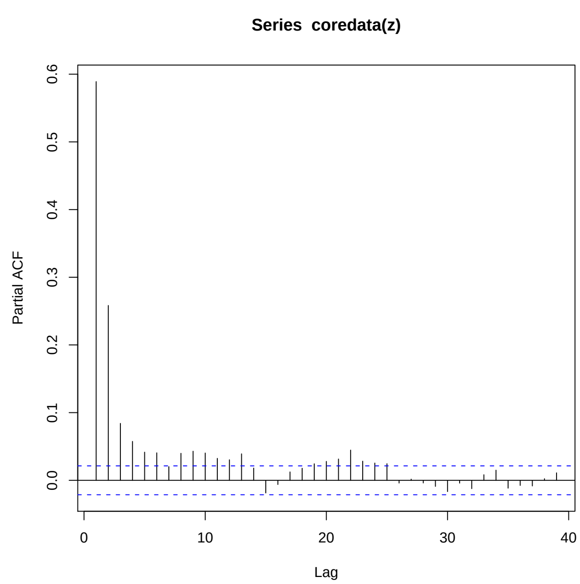

FIGURE 3. PACF

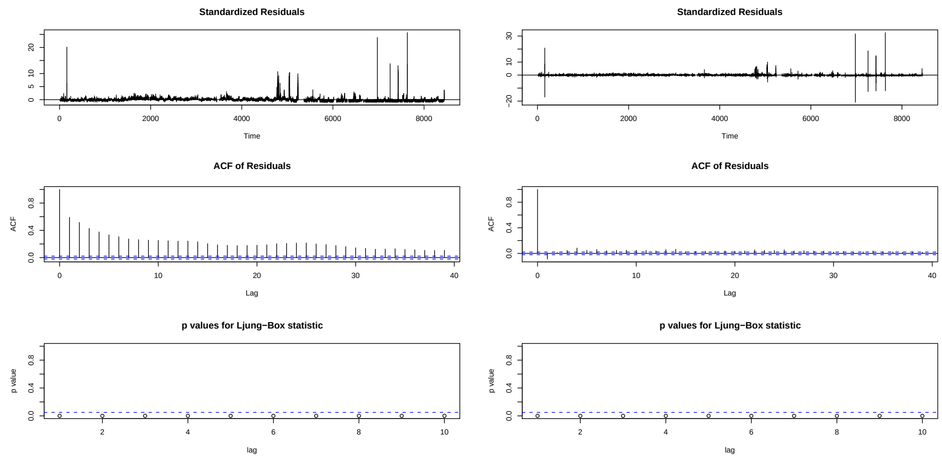

FIGURE 4. Residuals for AR(0) and AR(1)

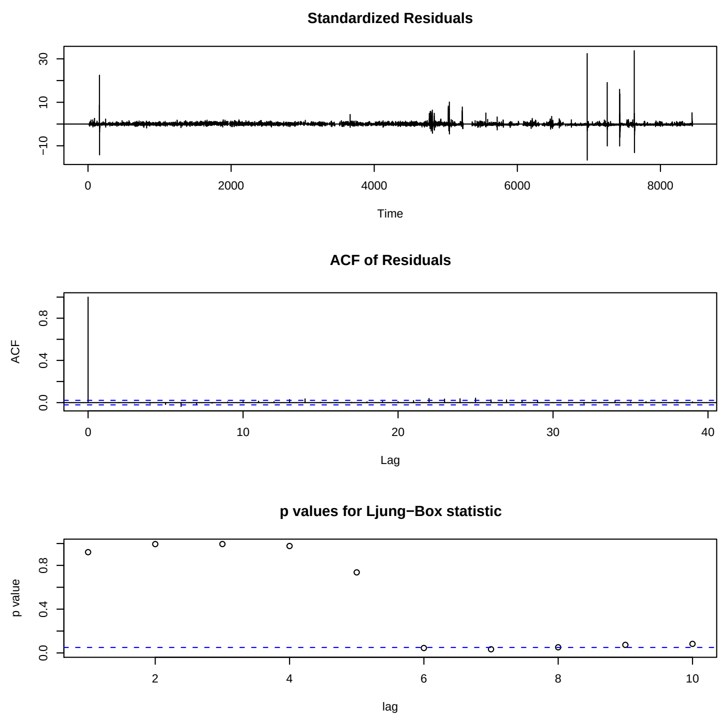

FIGURE 5. Residuals for AR(6)

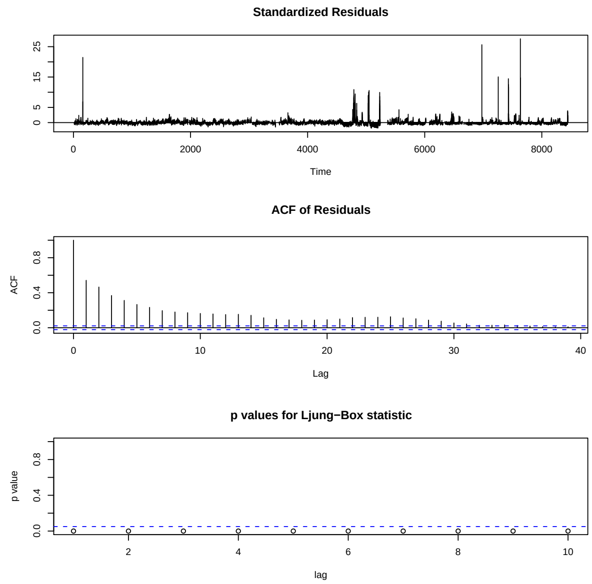

FIGURE 6. Residuals for AR(0) with Regression

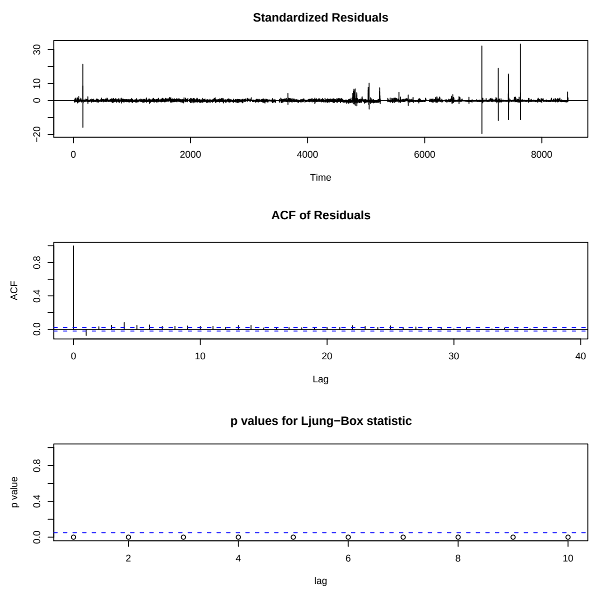

FIGURE 7. Residuals for AR(1) with Regression

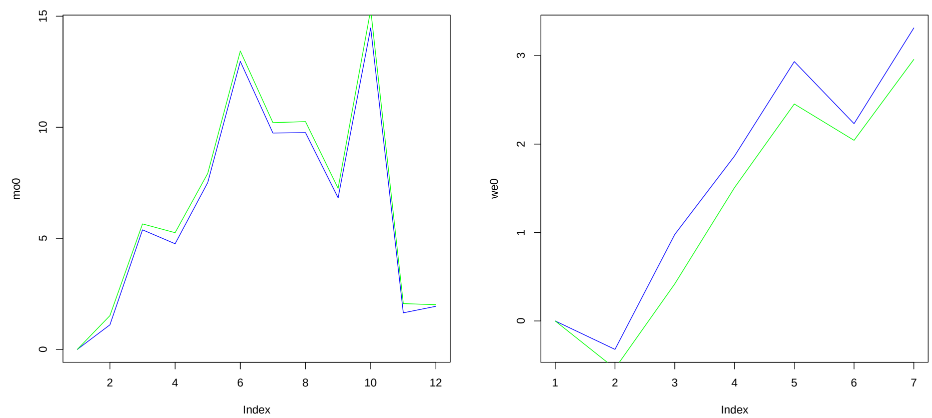

FIGURE 8. Effects of Month and Weekday

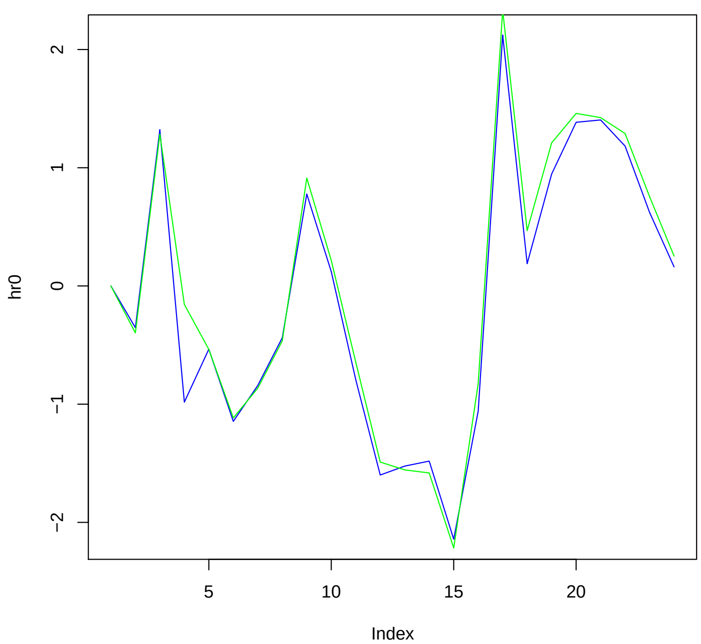

FIGURE 9. Effects of Hour

	a_1	a_2	a_3	a_4	a_5	a_6	σ^2	AIC
AR(0)							183.6	59145
AR(1)	0.6650						108.1	55374
AR(2)	0.5488	0.1625					106.1	55225
AR(3)	0.5365	0.1059	0.0968				105.2	55158
AR(4)	0.5298	0.0958	0.0504	0.0857			104.4	55104
AR(5)	0.5264	0.0947	0.0463	0.0665	0.0359		104.3	55097
AR(6)	0.5256	0.0912	0.0444	0.0627	0.0138	0.0417	104.1	55086

TABLE 1. AR Models for Raw Data

	a_1	σ^2	AIC
AR(0)		160.6	58242
AR(1)	0.6650	104.2	55167

TABLE 2. AR Models for Regression Residuals

DEPARTMENT OF STATISTICS, UNIVERSITY OF CALIFORNIA, LOS ANGELES, CA 90095-1554

E-mail address, Jan de Leeuw: deleeuw@stat.ucla.edu

URL, Jan de Leeuw: http://gifi.stat.ucla.edu