

# Predicting Daily Smog by State Space Estimation

Dave Zes

November 4, 2009

ABSTRACT. In the following we use Adaptive Least Squares to perform one-step-ahead prediction for daily max one-hour average level of smog (ozone, or O3) in Lancaster, California (Site 2031). We initially consider a 4-lag model, then turn to a leaner 2-lag model. We also briefly consider an ARMA treatment. We find that among these models the small state space model boasts the smallest prediction error.

## 1 Introduction

The reader is directed to two other documents, both of which may be downloaded from the web. The first, titled *A Brief Heuristic Primer on Kalman Filtering* provides introductory concepts that underpin the work-ups that follow. The second document provides R source code to reproduce the analysis presented below. Directions to obtain these are at the end of this paper.

Our data are provided by the California Air Resources Board, Air Quality Data Branch. Our response of interest is local concentration of ozone in air in parts per million (ppm) as reported from a monitoring station in Lancaster, California. Specifically, our variable is the average level of concentration over the *worst* one-hour period over the course of each day. We indexed our data fixing Day 1 to the date of the first recorded datum, 01-Jan-1980; our last datum falls on 15-Feb-1990 — Day 3642. We are missing entries for 57 days. The data histogram, Fig. 1, reveals no surprises. Our variable is never negative, rarely zero, but commonly very near zero. Much in the Poisson spirit, we see occasional realizations of values well above the median.

Figure 2 shows Ozone as a function of time (top), where we see a strong seasonal component. The power spectrum (bottom) jibes nicely with our expectation. Power peaks sharply at the low frequency corresponding to a period of 364.5 days.

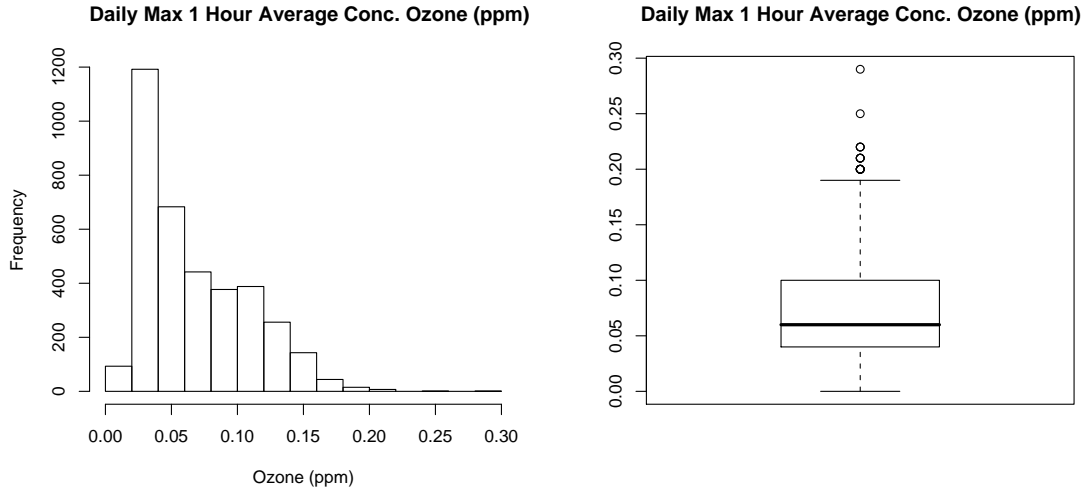


Figure 1: Univariate view

In preparation for autoregression we filled in missing values with the global mean, and appropriately re-indexed.

## 2 State Space Modeling

Our initial model is

$$\hat{y}_t = \beta_0 + \beta_1 \sin(2\pi t/365.4) + \beta_2 \cos(2\pi t/365.4) + \beta_3 y_{t-1} + \beta_4 y_{t-2} + \beta_5 y_{t-3} + \beta_6 y_{t-4} \quad (1)$$

We understand that if it appears the autoregressive component for our largest lag,  $k = 4$ , is strong, we may add more lag terms. If, on the other hand, we find near zero lag coefficient(s), (e.g.,  $\beta_3 \approx 0$  or  $\beta_4 \approx 0$ ), we can lean the model by discarding these terms or regularizing.

When we perform a Fourier Transform, or when we view a spectrogram, we must remove ourselves from the familiar time domain and move into the frequency domain. The State Space formulation of a sequence requires a spiritually similar transformation. We are no longer interested in the evolution of our familiar response,  $y_t$ , as a function of regressors. Rather, we are interested in the evolution of the parameter,  $\beta = (\beta_1, \beta_2, \dots, \beta_d)$ , over time. That is, we regard  $y_t$  and  $\mathbf{x}_t$  as fixed, and the parameter that maps one to the other motile over its supporting space, subject to disturbances. Since the parameter is ever-changing, it inherits the time subscript,  $\beta_t$ .

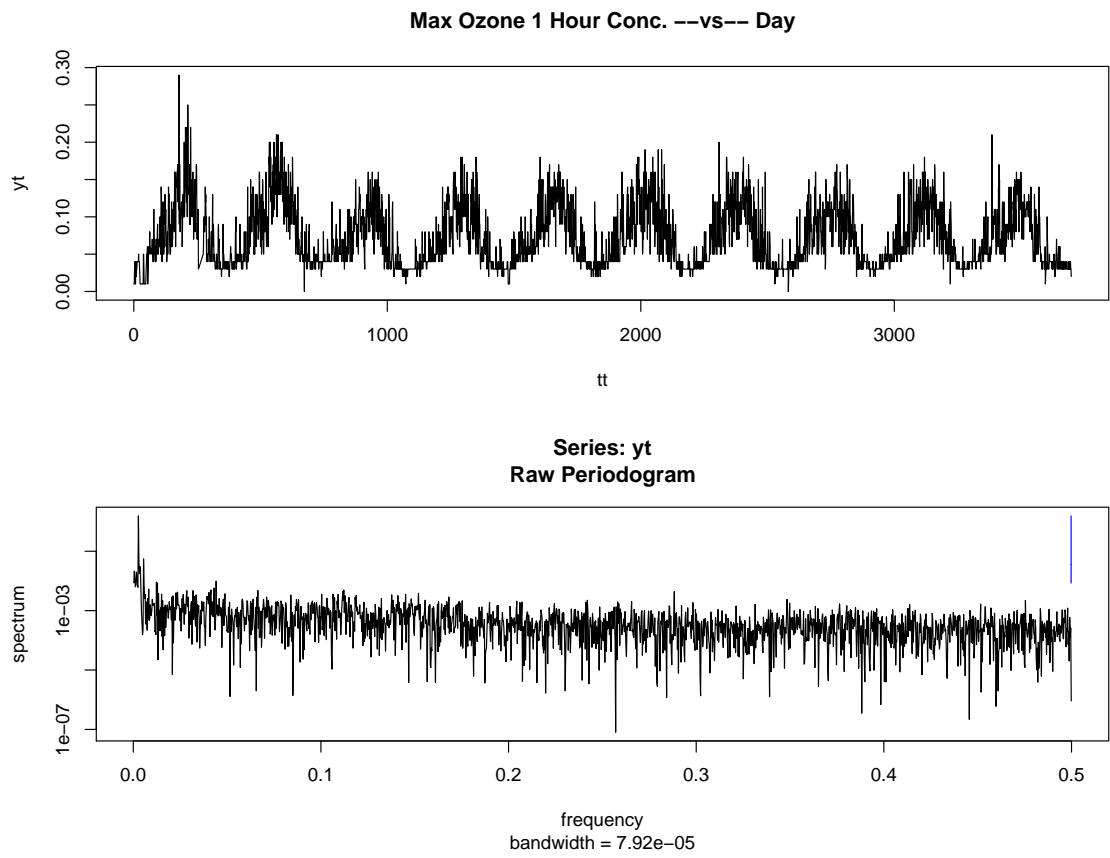


Figure 2: Daily 1Hr Max Ozone versus Days since 01-Jan-1980 & power spectrum

We have that

$$y_t = \mathbf{x}_t \boldsymbol{\beta}_t + \varepsilon_t \quad (2)$$

with the LS solution written as

$$\mathbf{L}_{\mathbf{xx}t}^{-1} \mathbf{l}_{\mathbf{xy}t} := \mathbf{E}[\mathbf{x}_t^T \mathbf{x}_t]^{-1} \mathbf{E}[\mathbf{x}_t^T y_t] \quad (3)$$

$$= \boldsymbol{\beta}_t \quad (4)$$

So, for our example,

$$\mathbf{L}_{\mathbf{xx}t} = \mathbf{E} \left\{ \begin{pmatrix} 1 & S(t) & C(t) & y_{t-1} & y_{t-2} & y_{t-3} & y_{t-4} \\ S(t) & S(t)^2 & S(t)C(t) & S(t)y_{t-1} & S(t)y_{t-2} & S(t)y_{t-3} & S(t)y_{t-4} \\ C(t) & C(t)S(t) & C(t)^2 & C(t)y_{t-1} & C(t)y_{t-2} & C(t)y_{t-3} & C(t)y_{t-4} \\ y_{t-1} & y_{t-1}S(t) & y_{t-1}C(t) & y_{t-1}^2 & y_{t-1}y_{t-2} & y_{t-1}y_{t-3} & y_{t-1}y_{t-4} \\ y_{t-2} & y_{t-2}S(t) & y_{t-2}C(t) & y_{t-2}y_{t-1} & y_{t-2}^2 & y_{t-2}y_{t-3} & y_{t-2}y_{t-4} \\ y_{t-3} & y_{t-3}S(t) & y_{t-3}C(t) & y_{t-3}y_{t-1} & y_{t-3}y_{t-2} & y_{t-3}^2 & y_{t-3}y_{t-4} \\ y_{t-4} & y_{t-4}S(t) & y_{t-4}C(t) & y_{t-4}y_{t-1} & y_{t-4}y_{t-2} & y_{t-4}y_{t-3} & y_{t-4}^2 \end{pmatrix} \right\} \quad (5)$$

where, to tighten up notation, we say  $S(t) = \sin(2\pi t/365.4)$  and  $C(t) = \cos(2\pi t/365.4)$ , and we have

$$\mathbf{l}_{\mathbf{xy}t} = \mathbf{E} \left\{ \begin{pmatrix} y_t \\ y_t S(t) \\ y_t C(t) \\ y_t y_{t-1} \\ y_t y_{t-2} \\ y_t y_{t-3} \\ y_t y_{t-4} \end{pmatrix} \right\} \quad (6)$$

Recall that the *state space system* is defined:

$$\boldsymbol{\beta}_t = \mathbf{F}_t \boldsymbol{\beta}_{t-1} + \varepsilon_t \quad (7)$$

$$\mathbf{y}_t = \mathbf{X}_t \boldsymbol{\beta}_t + \nu_t \quad (8)$$

We set  $\mathbf{F} = \mathbf{I}$ , and we have  $\mathbf{X}_t$  be our row vector (a record of data at  $t$ ), and we get

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \varepsilon_t \quad (9)$$

$$y_t = \mathbf{x}_t \boldsymbol{\beta}_t + \nu_t \quad (10)$$

The Kalman solution, as detailed in my other paper, comes by way of

$$\widehat{\boldsymbol{\beta}}_{t-1} = \widehat{\mathbf{L}}_{\mathbf{xx} \ t-1}^{-1} \widehat{\mathbf{l}}_{\mathbf{xy} \ t-1} \quad (11)$$

$$\widehat{\mathbf{L}}_{\mathbf{xx} \ t} = \widehat{\mathbf{L}}_{\mathbf{xx} \ t-1} + \mathbf{K}_t \cdot \left( \mathbf{x}_t^T \mathbf{x}_t - \widehat{\mathbf{L}}_{\mathbf{xx} \ t-1} \right) \quad (12)$$

$$\widehat{\mathbf{l}}_{\mathbf{xy} \ t} = \widehat{\mathbf{l}}_{\mathbf{xy} \ t-1} + \mathbf{K}_t \cdot \left( \mathbf{x}_t^T y_t - \widehat{\mathbf{l}}_{\mathbf{xy} \ t-1} \right) \quad (13)$$

$$\mathbf{P}_t^{(-)} = \mathbf{P}_{t-1}^{(+)} + \mathbf{Q}_t \quad (14)$$

$$\mathbf{K}_t = \mathbf{P}_t^{(-)} \left( \mathbf{P}_t^{(-)} + \mathbf{R}_t \right)^{-1} \quad (15)$$

$$\mathbf{P}_t^{(+)} = (\mathbf{I} - \mathbf{K}_t) \mathbf{P}_t^{(-)} \quad (16)$$

### 3 Estimation

To run our filter, we need five things. First, we need initial estimates of our 2 covariance objects. That is, we need  $\mathbf{L}_{\mathbf{xx}0}$  and  $\mathbf{l}_{\mathbf{xy}0}$ . We also have three filter tuning parameters, the prediction error (squared) at time zero,  $\mathbf{P}_0$ , and two sequences,  $\{\mathbf{Q}_t\}$ ,  $\{\mathbf{R}_t\}$ . *Fortunately* for our scenario, and just about any other like it, the amount of adjustment needed for each element of the covariance objects is already, in a sense, weighted by their own stochasticity. Hence, we may use a scalar gain,  $K$ , (same thing as a uniform diagonal matrix,  $\mathbf{K}$ ), and furthermore, we can set  $R = R_t, \forall t$  and  $Q = Q_t, \forall t$ .

We calculate  $\mathbf{L}_{\mathbf{xx}0}$  and  $\mathbf{l}_{\mathbf{xy}0}$  from the first 1000 data points. We then found  $P_0, R, Q$  by seeking values that minimized the square prediction error over the remaining 2698 data.

### 4 Results

For our larger 4-Lag Model, we found optimal controls to be  $P_0 = 1, R = 110, Q = 0.00017$ . Mean error over the last 2698 points is 0.02268995 ppm; over the last 365 points, 0.02175534 ppm. We obtained slightly improved results by discarding lag-3 and lag-4. Optimal controls for this leaner 2-Lag Model were found:  $P_0 = 1, R = 67, Q = 0.00017$ . Mean error over the last 2698 points is 0.02262347 ppm; over the last 365 points, 0.0217529 ppm. We were able to gain slight improvement by applying regularization, but details will not be provided here.

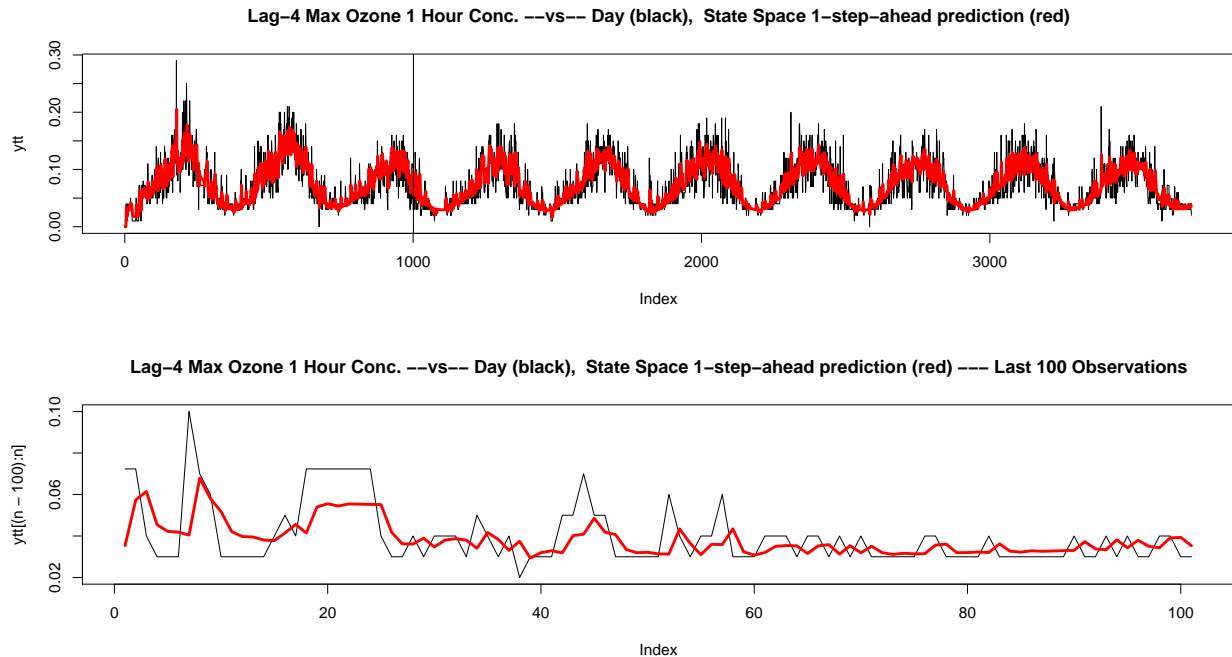


Figure 3: 4-Lag Model. Data (black) with 1-step-ahead Prediction (red)

Figures 3 & 4 show the original sequence with the prediction overlaid in red for the large and small models respectively. The top panel in each shows the full data, the bottom panel zooms us in to just the last 100 observations.

Figures 5 & 6 show our parameter estimates as they evolve through our filter over time.

Lastly, we consider an ARIMA analysis. In Figure 7 we revisit our ozone data. In the top panel we overlay an LS sinusoid of period 365.2 days. The residuals from this fit are shown in the bottom panel. There the points, shown as dots, reveal patterning. The appearance of a complementary cyclic pattern suggests that the original sequence contains many same or near-valued entries. The cause of this is not known. Possibly missing data, prior to our handling, was imputed with a constant, or the ozone sensor was biased. In any event, there's a strong suggestion of non-stationarity. Despite, we found the best predictive ARIMA model, looking for minimum mean prediction error over the last 365 points, was ARIMA(2,0,1). (We restricted our space to only the last 365 data because the prediction function for an ARIMA fit object runs very slowly in R). The mean prediction error was found to be 0.02190743 — slightly underperforming the 2-Lag State Space Model. Table 1 summarizes model performance.

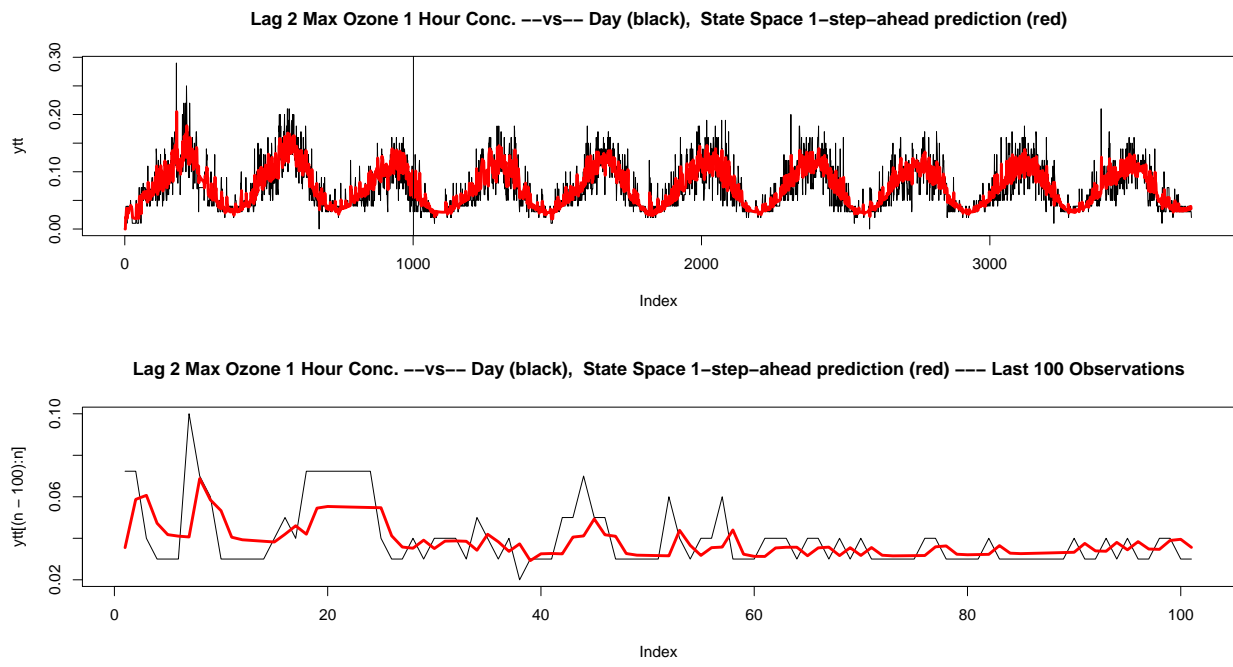


Figure 4: 2-Lag Model. Data (black) with 1-step-ahead Prediction (red)

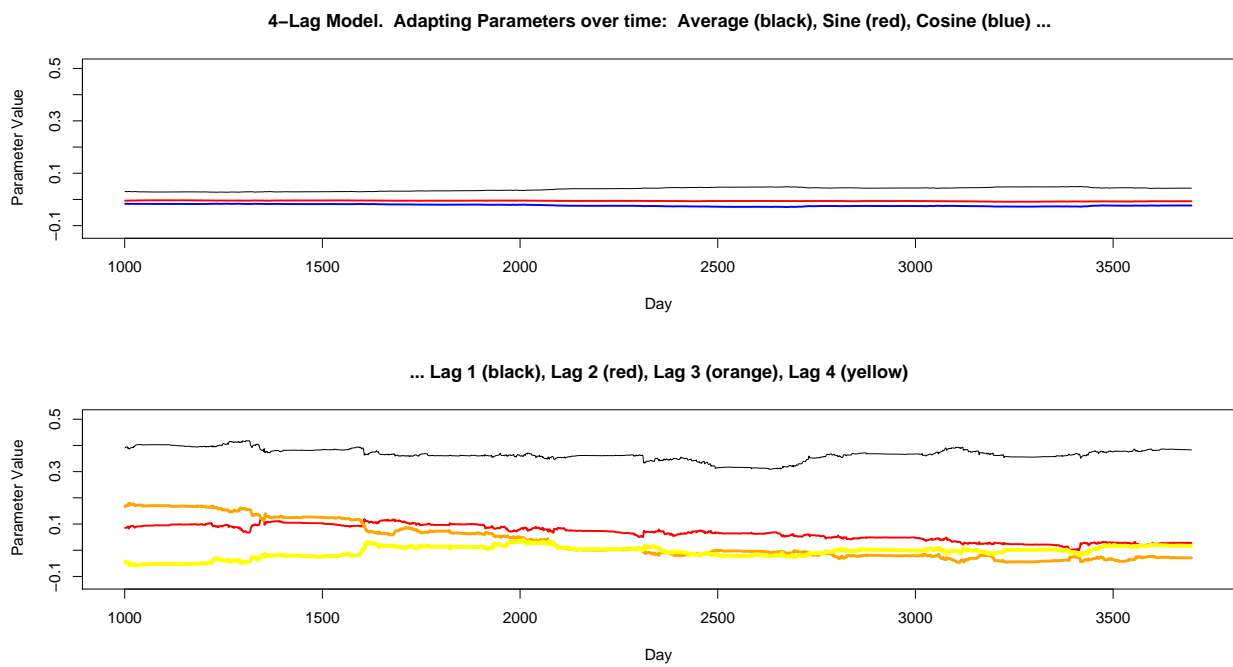


Figure 5: 4-Lag Model. Parameter values evolving — adapting — over time

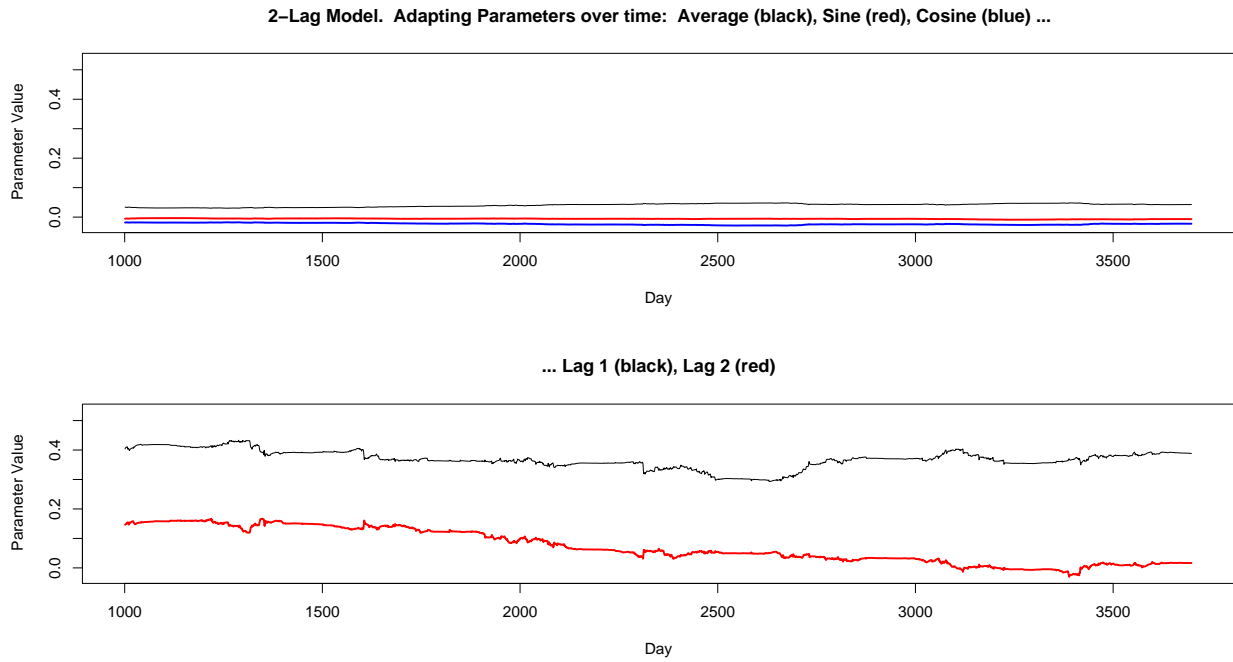


Figure 6: 2-Lag Model. Parameter values evolving — adapting — over time

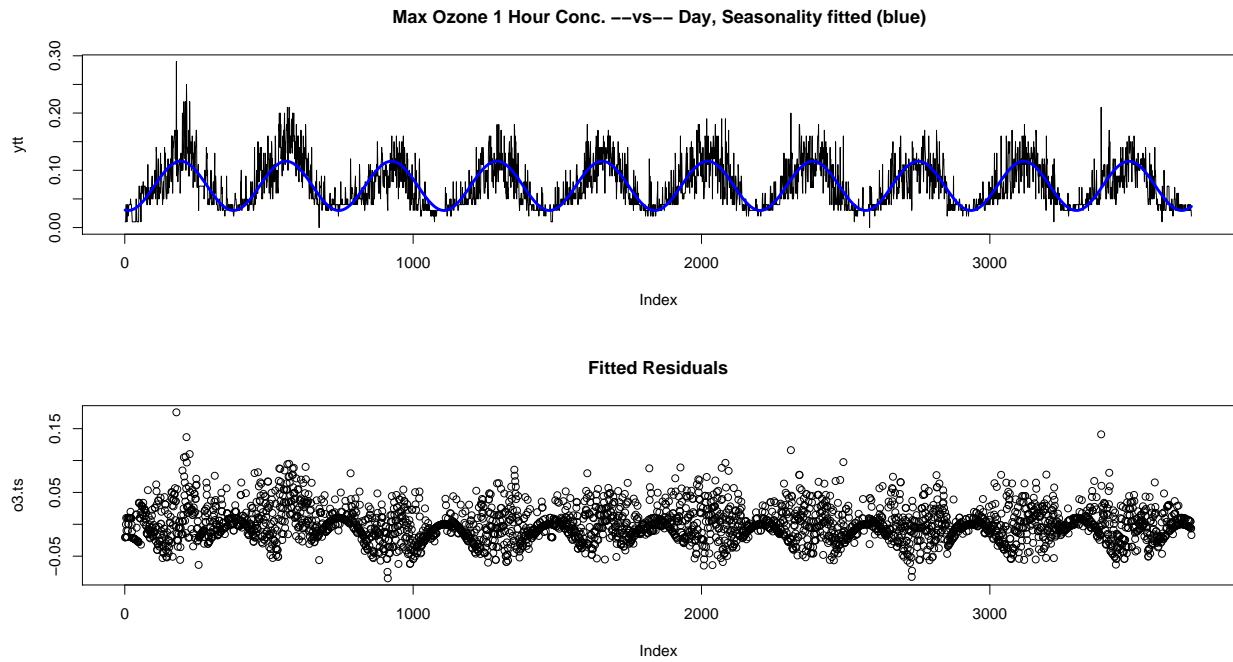


Figure 7: Our Ozone data fitted with cyclic component in blue (top). Fitted residuals versus Day (bottom)

Model	error
Lag-4 SS	0.0217553
Lag-2 SS	0.0217529
ARIMA(2,0,1)	0.0219074
ARIMA(2,0,2)	0.0219141
ARIMA(3,0,2)	0.0219103

Table 1: 1-step-ahead prediction performance for various models

## 5 Companion Documents

R code:

<http://www.stat.ucla.edu/~davezes/site2031stateSp.R>

Introductory material:

<http://www.stat.ucla.edu/~davezes/kalmanPrimer.pdf>

## References

- [1] Brian D.O. Anderson and John B. Moore, *Optimal Filtering*. Dover, Mineola, NY, 1979, republished in 2005.
- [2] Richard Berk, *Regression Analysis: A Constructive Critique*. Sage, Thousand Oaks, CA, 2004.
- [3] S. Gunnarsson, “Combining Tracking and Regularization in Recursive Least Squares Identification.” Proceedings of the 35th Conference on Decision and Control. Kobe, Japan. December 1996.
- [4] S. Gunnarsson, “On Covariance Modification and Regularization in Recursive Least Squares Identification.” In 10th IFAC Symposium on System Identification — SYSID’94, pages 661–666, Volume 2. Copenhagen, Denmark, July 1994.
- [5] E.J. Hannan and Manfred Deistler, *The Statistical Theory of Linear Systems*. Wiley, 1988.
- [6] Simon Haykin, *Adaptive Filter Theory*. Forth Edition. Prentice-Hall, Upper Saddle River, 2002.
- [7] Chi Sing Leung et. al, “On the Regularization of Forgetting Recursive Least Square.” IEEE Transactions on Neural Networks, Volume 10, Number 6. November, 1999.
- [8] Irwin Miller and Marylees Miller, *John E. Freund’s Mathematical Statistics*. Prentice-Hall, Upper Saddle River, 1999.
- [9] Robert H. Shumway and David S. Stoffer, *Time Series Analysis and Its Applications, With R Examples*. Second Edition. Springer, N.Y., 2006.
- [10] Mike West and Jeff Harrison, *Bayesian Forecasting and Dynamic Models*. Second Edition. Springer-Verlag, New York, 1997.