Time series of air pollutants

Jiri Neubauer, Jaroslav Michalek, Frantisek Bozek

The topic of the contribution is the description of time series of air pollutants. These time series contain values of carbon monoxide (CO), nitrogen oxides (NO_x), sulphur dioxide (SO_2) and suspended particulate matter (dust). The necessary pollutant concentrations were obtained from the measuring station with automatic emission monitoring in Vyškov, the Czech Republic, for the period 1997-2001. The assessment of the air pollution degree was based on monitoring concentrations of pollutants in the ground atmospheric layer.

For the purpose of elaborating time series these three method were used:

1. The SARIMA process
2. The process of hidden periods
3. The exponential smoothing

These statistical tests of randomness were calculated for verifying of the suitability of the model:

A) The test based on signs of differences
B) The test based on break points
C) The test based on the Kendall's coefficient t
D) The test based on the Spearman's coefficient r
E) The median test

In addition the portmanteau test was used for verifying of the ARMA process.

You can see on the transparency films the particular models (on all data series or in detailed view) and you can visually consider suitability of each models. It seems to be interesting to calculate the forecasting of 7 future values and compare with the measured data.

The second part is dedicated to the problem of exceeding the level. I demonstrate this problem on the time series of DUST. This series were found as the ARMA process – A(7). The estimation of the spectral density function were calculate using the expression of the spectral density function of the ARMA process and after it were computed estimations of moments l_0 and l_2. You can see in the table expectation of theoretical values of the random variable *Cu* based on calculated spectral density function and the real measured values from the series. The difference between this magnitudes is shown on the graph. Next table contains expectation of the random variable $Z(T)$ (total time over the level u) and measured values (see. figure). The anticoincidence between results calculated using given formulas and really measured values needs to be seeking in non-performance of the conditions a) – d) (especially condition c) was not satisfied (Kolmogorov – Smirnov test, chi square test, A test normality – reject the hypothesis of normality). The solution of this could be possible to find in any transformation of the series.

For the future, it is counted on the development of a model which would enable to predict the concentration dependence of selected pollutants on temperature, pressure and humidity, air, wind intensity, inversion and other climate conditions. Further, the observation of the dependence of concentration development as for particular pollutants in interrelation and the attempt of possible determination of their synergetic or antagonistic effects is being considered.

ARIMA

ARMA(p, q)

Let $\{e_t\}$ be white noise,

$$y_t = f_1 y_{t-1} + f_2 y_{t-2} + ... + f_p y_{t-p} + e_t + q_1 e_{t-1} + q_2 e_{t-2} + ... + q_q e_{t-q}$$

is a mixed autoregressive moving average process ARMA(p, q).

In lag operator form ($Ly_t = y_{t-1}, L(Ly_t) = L^2 y_t = y_{t-2}$)

$$(1 - f_1 L - f_2 L^2 - ... - f_p L^p) y_t = (1 + q_1 L + q_2 L^2 + ... + q_q L^q) e_t$$

$$f(L) y_t = q(L) e_t$$

ARIMA(p, d, q)

This model is convenient for description of nonstationary time series.

1-st difference … $\Delta y_t = y_t - y_{t-1} = y_t - Ly_t = (1 - L) y_t$

2-nd difference … $\Delta^2 y_t = \Delta(\Delta y_t) = \Delta(y_t - y_{t-1}) = \Delta y_t - \Delta y_{t-1} = y_t - 2y_{t-1} + y_{t-2} =$

$$= (1 - 2L + L^2) y_t = (1 - L)^2 y_t$$

ARIMA(p, d, q) is defined as

$$f(L) w_t = q(L) e_t ,$$

where

$$w_t = \Delta^d y_t .$$

We can write

$$f(L)(1 - L)^d y_t = q(L) e_t .$$

SARIMA(p, d, q)(P, D, Q) lag s

This model is convenient for description of nonstationary seasonal time series.

The seasonal lag operator… $L^s y_t = y_{t-s}, L^s (L^s y_t) = L^{2s} y_t = y_{t-2s}$

The seasonal difference operator … $\Delta_s = 1 - L^s$

$$\Delta_s y_t = (1 - L^s) y_t = y_t - y_{t-s}$$

$$\Delta_s^2 y_t = (1 - L^s)^2 y_t = (1 - 2L^s + L^{2s}) y_t = y_t - 2y_{t-s} + y_{t-2s}$$

We can write

$$f(L) \Phi(L^s)(1 - L)^d (1 - L^s)^D y_t = q(L) \Theta(L^s) e_t .$$

Properties of the autocorrelation function and the partial autocorrelation function and criteria AIC, FPE were used for determining the order of the model. Program STATISTICA and MATLAB were used for estimation of coefficients $f_1,...,f_p,q_1,...,q_q$, $\Phi_1,...,\Phi_P,\Theta_1,...,\Theta_Q$ (Maximum likelihood estimation).

Model of Hidden Periods

Periodogram

The periodogram $I(w)$ of time series $y_1, \ldots, y_n$ is defined as a function of variable w

$$I(w)=\frac{1}{4p}\left(a^2(w)+b^2(w)\right),\ -p \le w \le p\ ,$$

where

$$a(w)=\sqrt{\frac{2}{n}}\sum_{t=1}^{n} y_t \cos(wt)\ ,$$

$$b(w)=\sqrt{\frac{2}{n}}\sum_{t=1}^{n} y_t \sin(wt)\ .$$

Using the periodogram we can find significant periods in time series $y_1, \ldots, y_n$. We are trying to construct model of this series

$$y_t = m+\sum_{j=1}^{p}(a_j \cos(w_j t)+b_j \sin w_j t)+e_t,\ t=1,...,n.$$

m, a_j, b_j are unknown parameters that is necessary to estimate, w_j are different frequencies ($w_j \in (0,p)$), e_t is a normally distributed white noise ($N(0,s_e^2)$). The significant frequencies is possible to determine using a Fisher's test. A null hypothesis in the Fisher test is

$$y_t = e_t\ ,$$

which suppose that series do not contains any significant periods. An alternative hypothesis supposes that there are some significant periods in this series. The Fisher's test is based on values of the periodogram computing for frequencies

$$w_j^* = \frac{2pj}{n},\ j=1,...,m,\quad m=\left[\frac{n-1}{2}\right]$$

High values of the periodogram determine significant frequencies.

$$W = \max_{j=1,...,m} \frac{I(w_j^*)}{\sum_{i=1}^{m} I(w_j^*)}$$

The hull hypothesis is rejecting if

$$W > g_F\ ,$$

where g_F is a critical value of the Fisher's test (critical values are tabulated).

Parameters m, a_j, b_j we can estimate using linear regression model

$$y_t = \mathbf{A}\mathbf{x}_t + e_t,$$

where

$$\mathbf{x}_t = \begin{bmatrix} 1 & \cos(w_1 t) & \sin(w_1 t) & \mathbf{L} & \cos(w_M t) & \sin(w_M t) \end{bmatrix},$$

$w_1, \ldots, w_M$ are significant frequencies,

$$\mathbf{A} = \begin{bmatrix} m & a_1 & b_1 & \mathbf{L} & a_M & b_M \end{bmatrix}.$$

Exponential Smoothing

Brown's Exponential Smoothing

An exponential weights moving average is an average that weights the observed time series values unequally, with more recent observations being weighted more heavily than older observations. This unequal weighting is achieved through smoothing constants that determine how much weight is given to each observation.

If m_{t-1} is the moving average calculated for the first t-1 points in the series y_t then given the value y_t, the new moving average is found as

$$m_t = a\, y_t + (1-a)m_{t-1},$$

where a is the smoothing constant $(0 < a < 1)$.
For a data series $y_1, \ldots, y_n$ forecasts are given by

$$\hat{y}_t = m_{t-1}.$$

The initial value m_0 is calculated as the average level in the first quarter of the series.

Holt's Linear Smoothing

This adds a trend component to Brown's Exponential method. For a data series y_t forecasts are given by

$$\hat{y}_t = m_{t-1} + b_{t-1},$$

where

$m_t = a\, y_t + (1-a)(m_{t-1} + b_{t-1})$ is the level at time t,

$b_t = g\,(m_t - m_{t-1}) + (1-g)b_{t-1}$ is the trend at time t,

a is the level smoothing constant $(0 < a < 1)$,

g is the trend smoothing constant $(0 < g < 1)$.

The initial values m_0 and b_0 are calculated by a linear regression on the first half of the series.

Winter's Seasonal Smoothing (Additive)

This adds an additive seasonal component to Holt's Linear method. For a data series y_t forecasts are given by

$$\hat{y}_t = m_{t-1} + b_{t-1} + c_{t-s},$$

where

$m_t = a\,(y_t - c_{t-s}) + (1-a)(m_{t-1} + b_{t-1})$ is the level at time t,

$b_t = g\,(m_t - m_{t-1}) + (1-g)\,b_{t-1}$ is the trend at time t,

$c_t = d\,(y_t - m_t) + (1-d)\,c_{t-s}$ is the seasonal component at time t,

a is the level smoothing constant ($0 < a < 1$),

g is the trend smoothing constant ($0 < g < 1$),

d is the seasonal smoothing constant ($0 < d < 1$),

s is the season period.

Winter's Seasonal Smoothing (Multiplicative)

This adds a mulitiplicative seasonal component to Holt's Linear method. For a data series y_t forecasts are given by

$$\hat{y}_t = (m_{t-1} + b_{t-1})c_{t-s},$$

where

$m_t = a\,\frac{y_t}{c_{t-s}} + (1-a)(m_{t-1} + b_{t-1})$ is the level at time t,

$b_t = g\,(m_t - m_{t-1}) + (1-g)\,b_{t-1}$ is the trend at time t,

$c_t = d\,\frac{y_t}{m_t} + (1-d)\,c_{t-s}$ is the seasonal component at time t.

a is the level smoothing constant ($0 < a < 1$),

g is the trend smoothing constant ($0 < g < 1$),

d is the seasonal smoothing constant ($0 < d < 1$),

s is the season period.

Problems of Exceeding the Level *u*

Let $\{X_t\}$ is a random process defined in a interval $\langle 0,T\rangle$. We assume that $\{X_t\}$ possess the following properties:

a) $\{X_t\}$ is real $EX_t^2 < \infty$, $EX_t = 0$;

b) $\{X_t\}$ is continuous in the mean ($E|X_t - X_{t0}|^2 \to 0$ for $t \to t_0$) and strictly stationary (for any admissible $t_1, \ldots, t_n$ and any h is $F_{t_1,\ldots t_n}(x_1,\ldots,x_n) = F_{t_1+h,\ldots t_n+h}(x_1,\ldots,x_n)$)

c) $\{X_t\}$ is gaussian (all distribution functions are normal);

d) realizations of the process $\{X_t\}$ are continuous with probability 1.

Let $R(t)$ be an autocovariance function of the process $\{X_t\}$ and $f(l)$ be a spectral density function.

$$l_k = \int_{-\infty}^{\infty} l^k f(l)dl \ , \ k = 0,1,2,...$$

It is easy to proof that $l_0 = R(0)$, because $R(t) = \int_{-\infty}^{\infty} \exp(itl) f(l)dl$. If $\int |l| f(l)dl < \infty$, then $l_1 = 0$ (the spectral density function is axially symmetric).

Let u be a real number, G_u be a set of continuous functions in $\langle 0,T\rangle$ that are not identically equal u in any open subinterval of $\langle 0,T\rangle$ and $g(0) \neq u$, $g(T) \neq u$. The function $g \in G_u$ crosses the level u in a point $t_0 \in (0,T)$ if there are points t_1 and t_2 in arbitrary neighbourhood of t_0 that

$$[g(t_1) - u][g(t_2) - u] < 0 .$$

The number of this points is C_u.

Let $\{X_t\}$ fulfils conditions a) – d),

$$E[C_u\langle 0,T\rangle] = \begin{cases} \dfrac{T}{p}\left(\dfrac{l_2}{l_0}\right)^{1/2} \exp\left(-\dfrac{u^2}{2l_0}\right) & \text{for} \quad l_2 < \infty \\ \infty & \text{for} \quad l_2 = \infty \end{cases}$$

Let $\{X_t\}$ fulfils conditions a) – d), u be a real number. We define

$$Y(t) = \begin{cases} 1 & \text{when} \quad X(t) \geq u \\ 0 & \text{when} \quad X(t) < u \end{cases},$$

$$Z(t) = \int_0^T Y(t)dt .$$

Variable $Z(t)$ indicates total time when the process $\{X_t\}$ is over the level u.

$$EZ(T) = T\left[1 - f\left(\frac{u}{s}\right)\right],$$

where $s^2 = R(0)$, $f\left(\frac{u}{s}\right)$ is a distribution function of $N(0,1)$.

If we describe time series by ARMA(p, q) process

$y_t = f_1 y_{t-1} + f_2 y_{t-2} + ... + f_p y_{t-p} + e_t + q_1 e_{t-1} + q_2 e_{t-2} + ... + q_q e_{t-q}$,
then the spectral density function is

$$f(l) = \frac{s^2}{2p} \frac{\left|1 + q_1 \exp(-il) + q_2 \exp(-2il) + ... + q_q \exp(-iql)\right|^2}{\left|1 - f_1 \exp(-il) - f_2 \exp(-i2l) - ... - f_p \exp(-ipl)\right|^2},$$

where s^2 is $\mathrm{var}(e_t)$.

Series of DUST

Transformation: $y_t^{TR} = y_t - c$, where c = 56,57 (linear regression)

The series y_t^{TR} - model AR(7)

p(1)	p(2)	p(3)	p(4)	p(5)	p(6)	p(7)
0,5782	-0,0244	0,0035	0,0634	0,0019	0,0496	0,0893

1. The number C_u

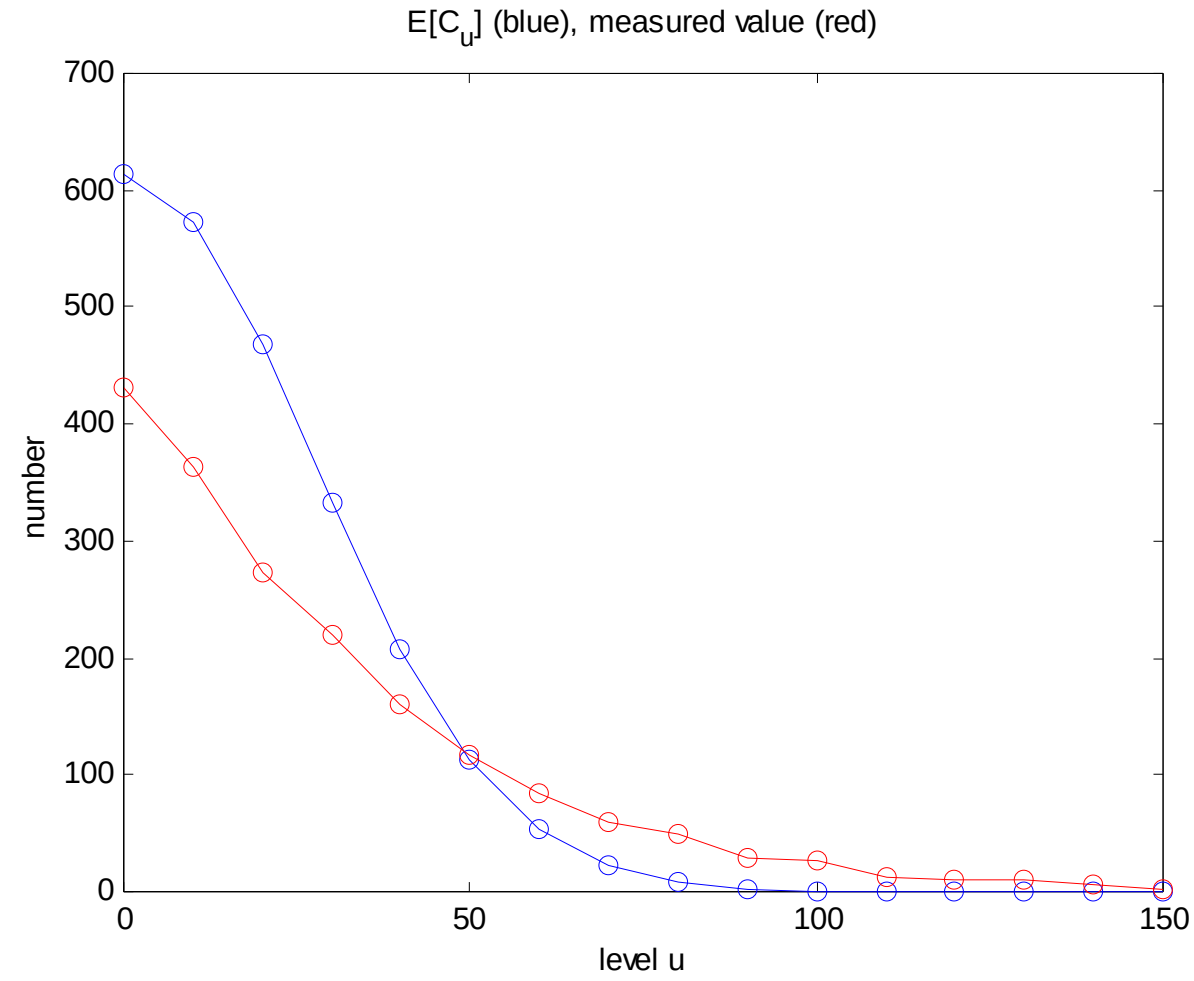
$l_0 = 739{,}0761$

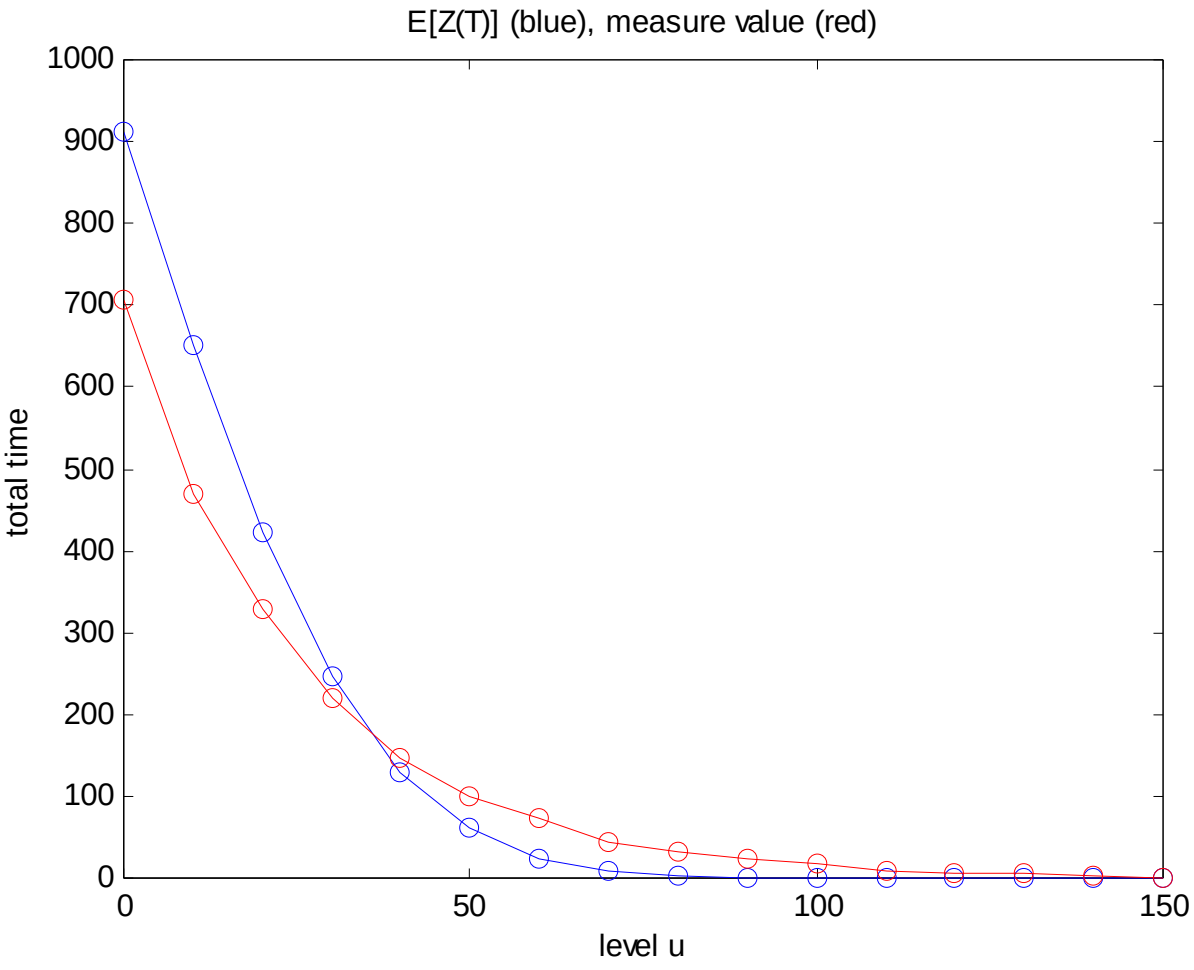
$l_2 = 823{,}2549$

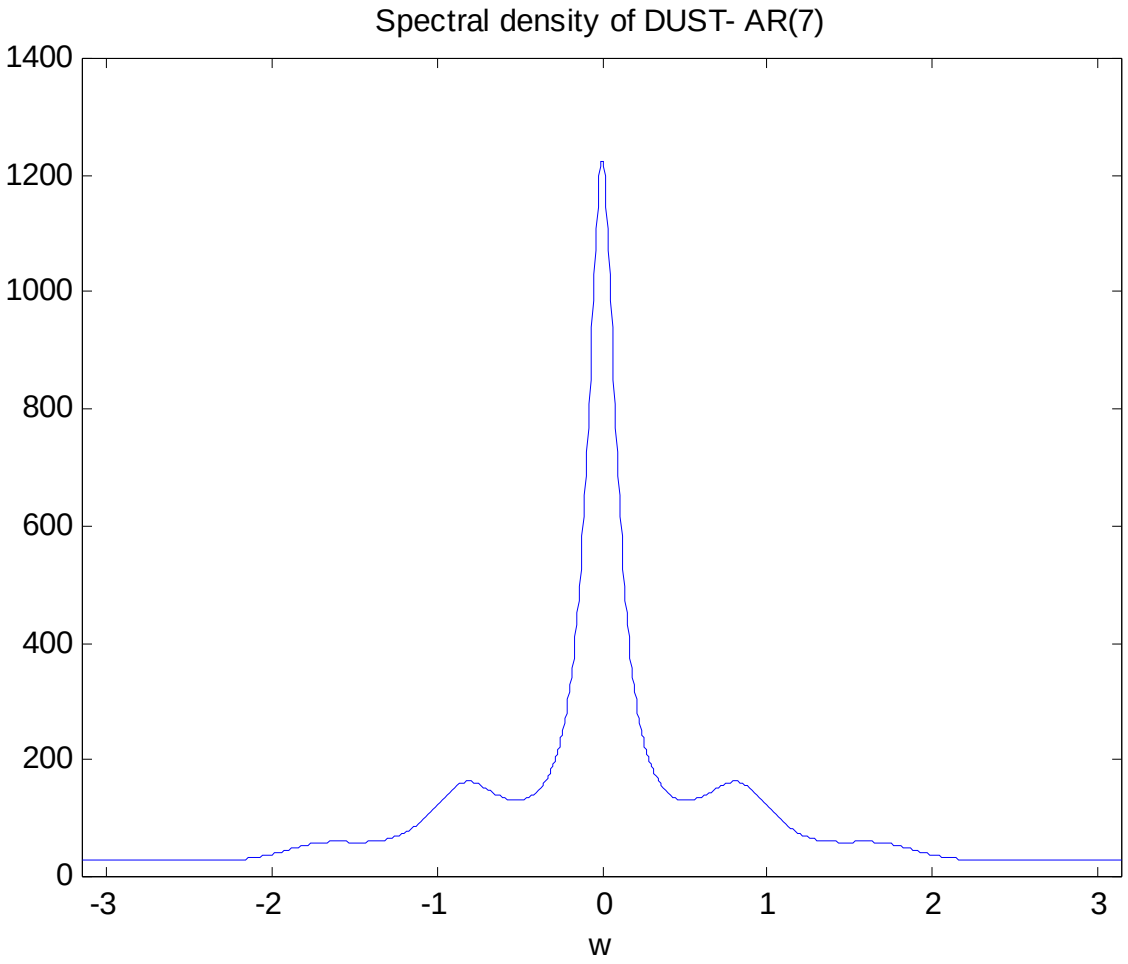
Level u	$E[C_u\langle 0,T\rangle]$	Measured value
0	613,1	432
10	573,0	363
20	467,7	273
30	333,5	219
40	207,7	161
50	113,0	117
60	53,7	85
70	22,3	59
80	8,1	49
90	2,6	29
100	0,7	27
110	0,2	12
120	0	10
130	0	10
140	0	6
150	0	2

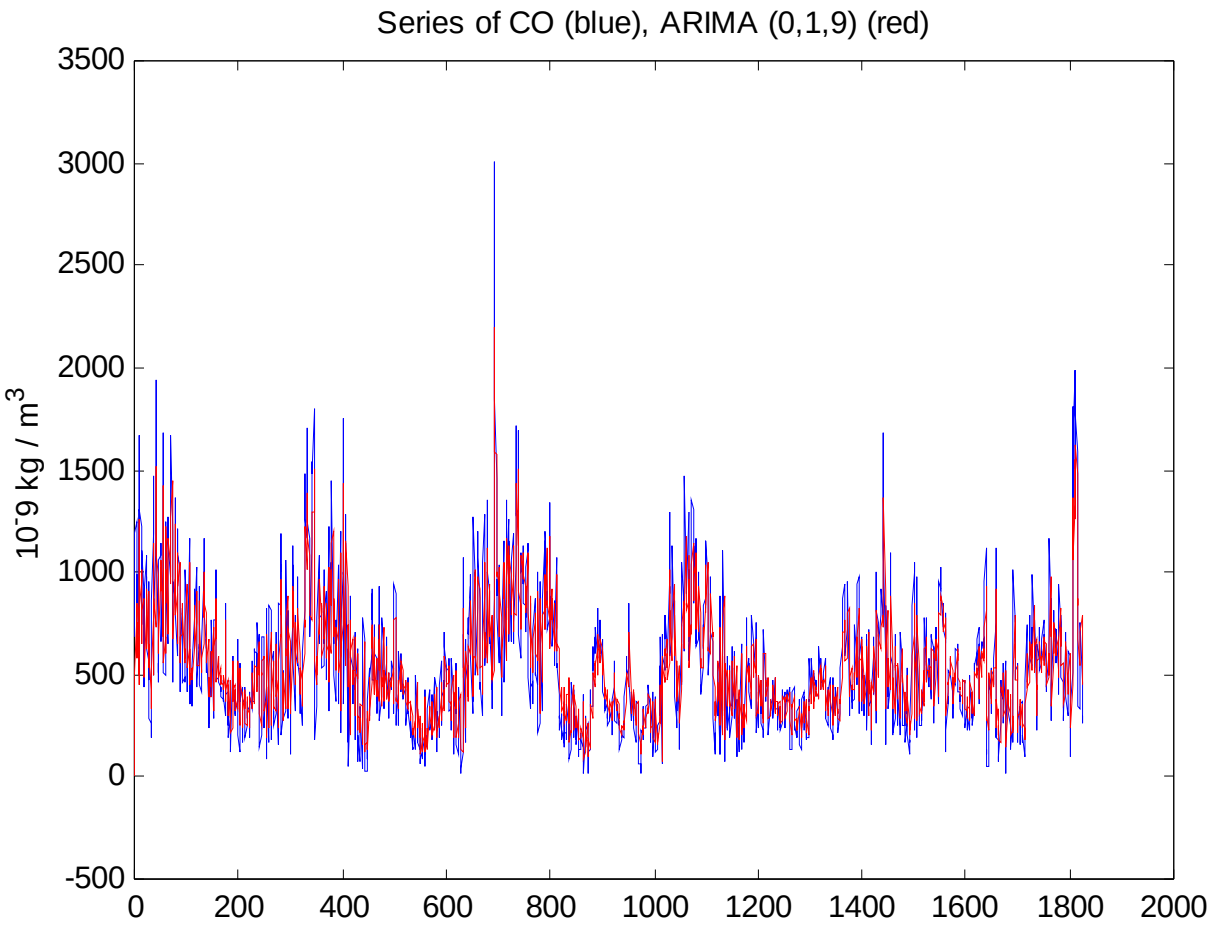
2. Total time over the level – $Z(T)$

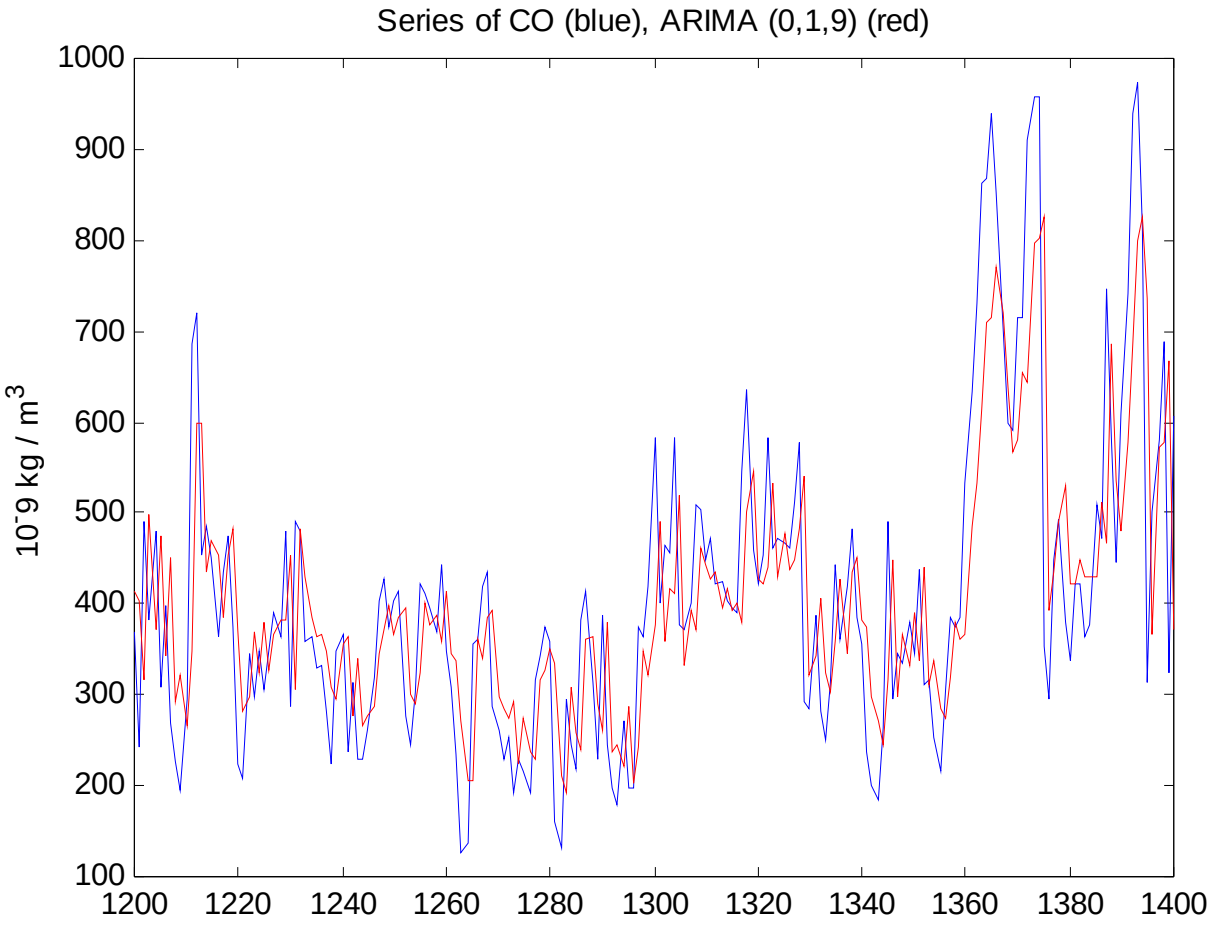
Level u	$E[Z(T)]$	Measured value
0	912,5	707
10	605,6	468
20	421,5	329
30	249,2	222
40	128,8	146
50	60,1	100
60	24,9	73
70	9,2	45
80	3,0	33
90	0,8	23
100	0,2	18
110	0,1	8
120	0	6
130	0	5
140	0	3
150	0	1

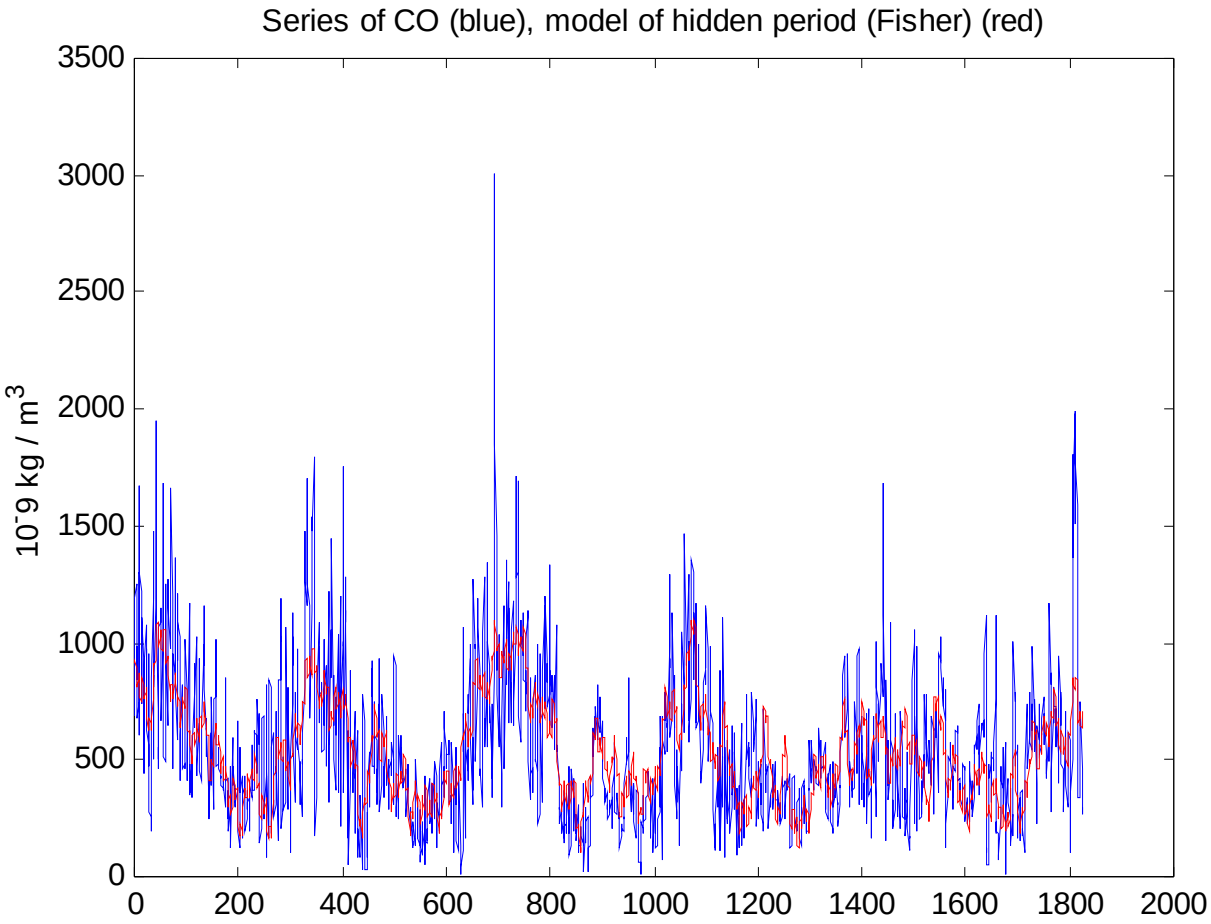


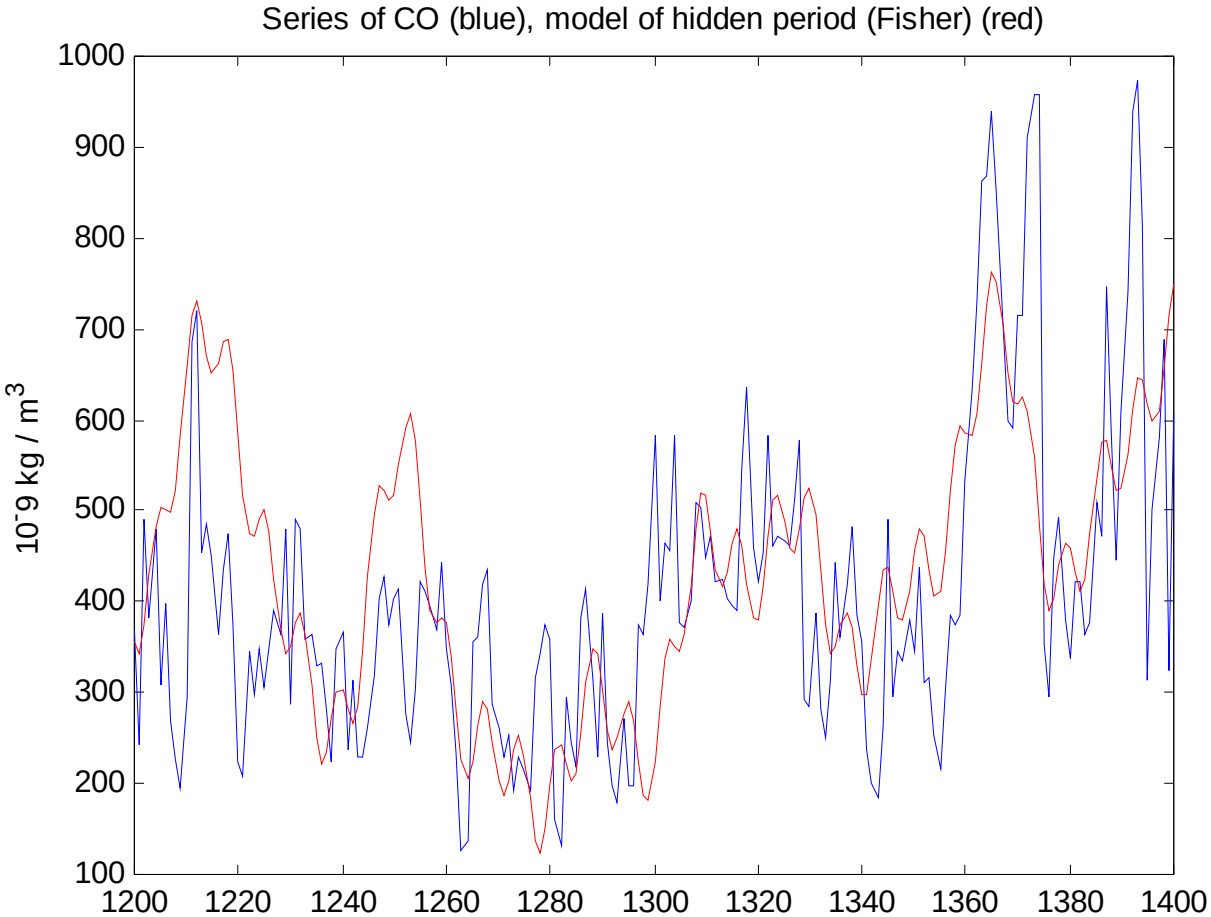


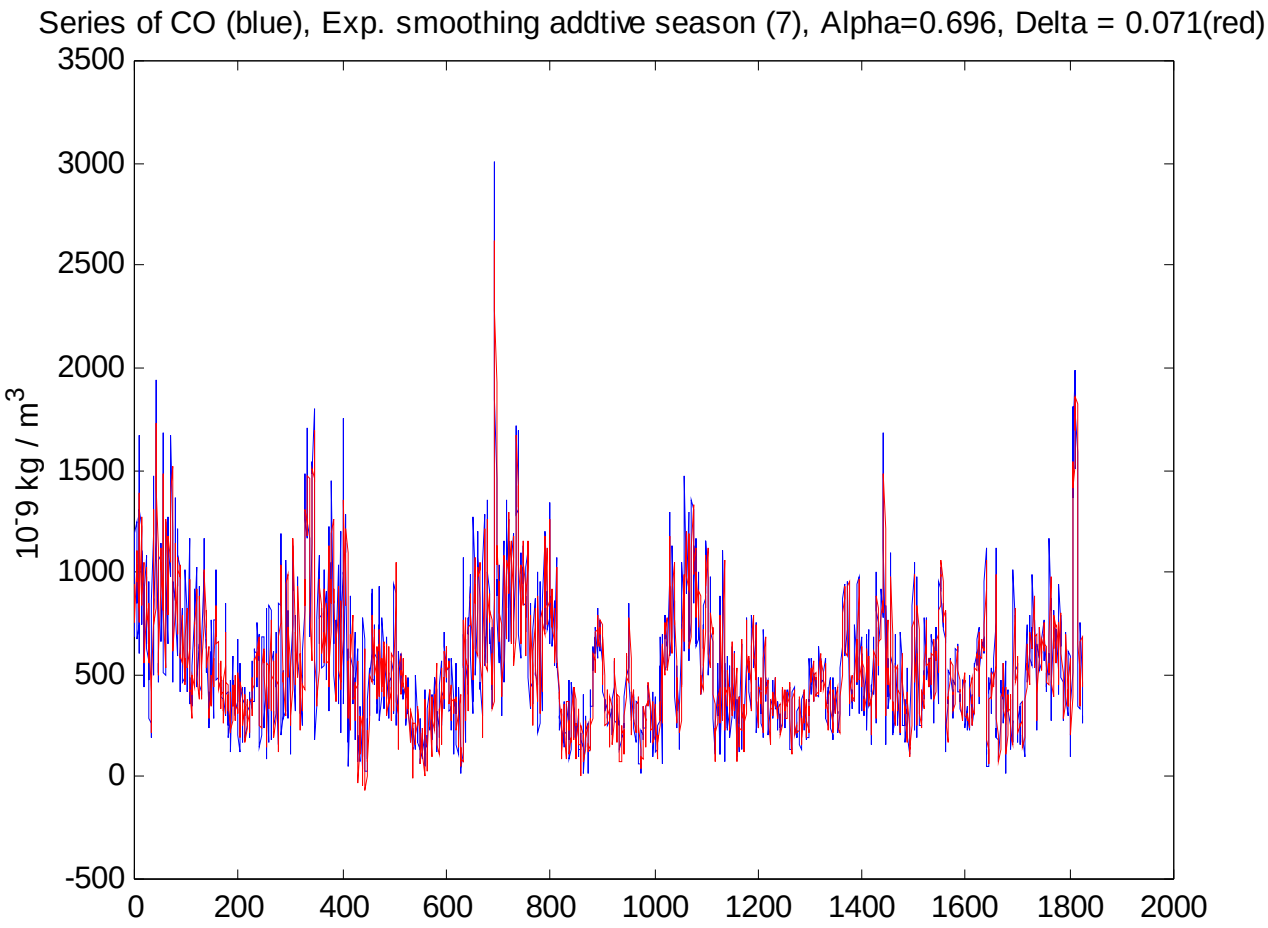


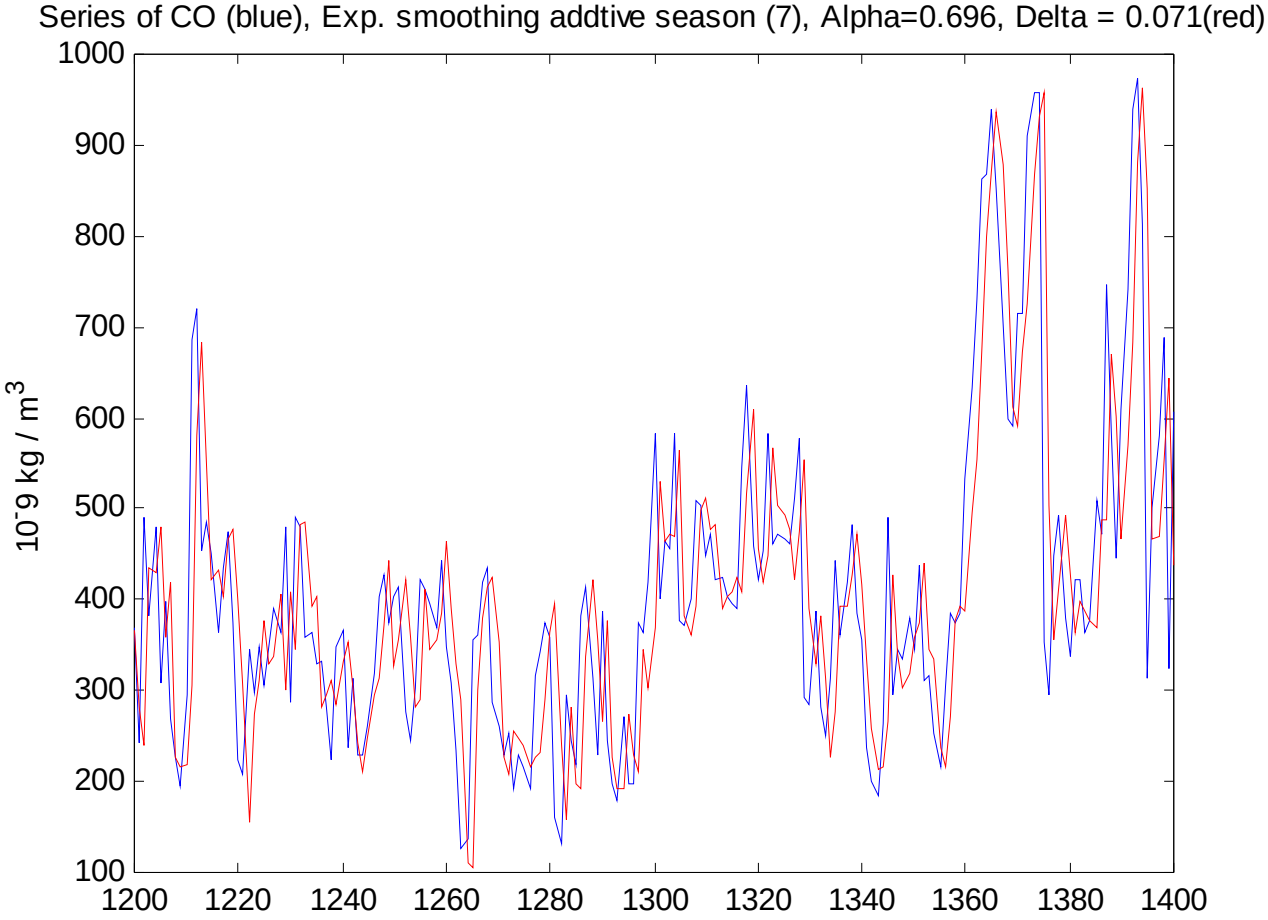


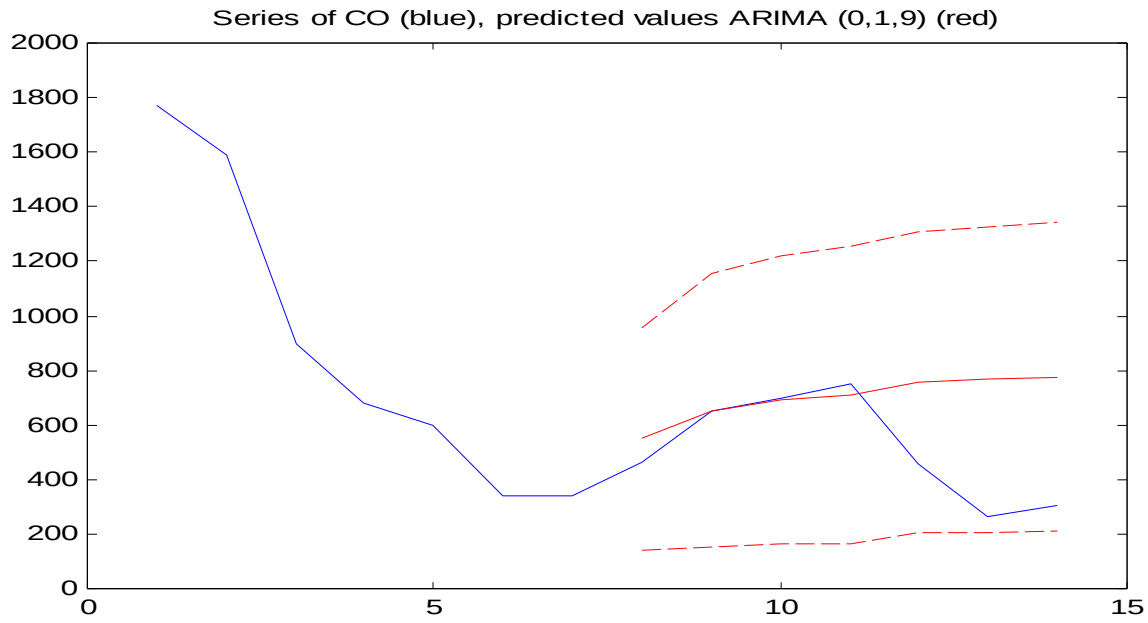


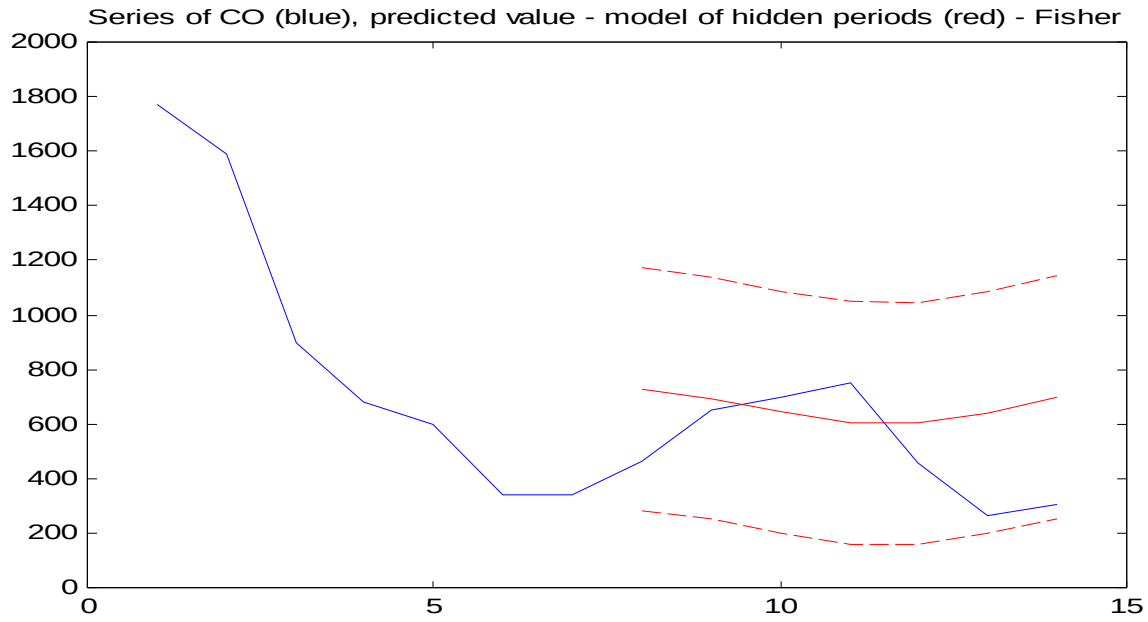


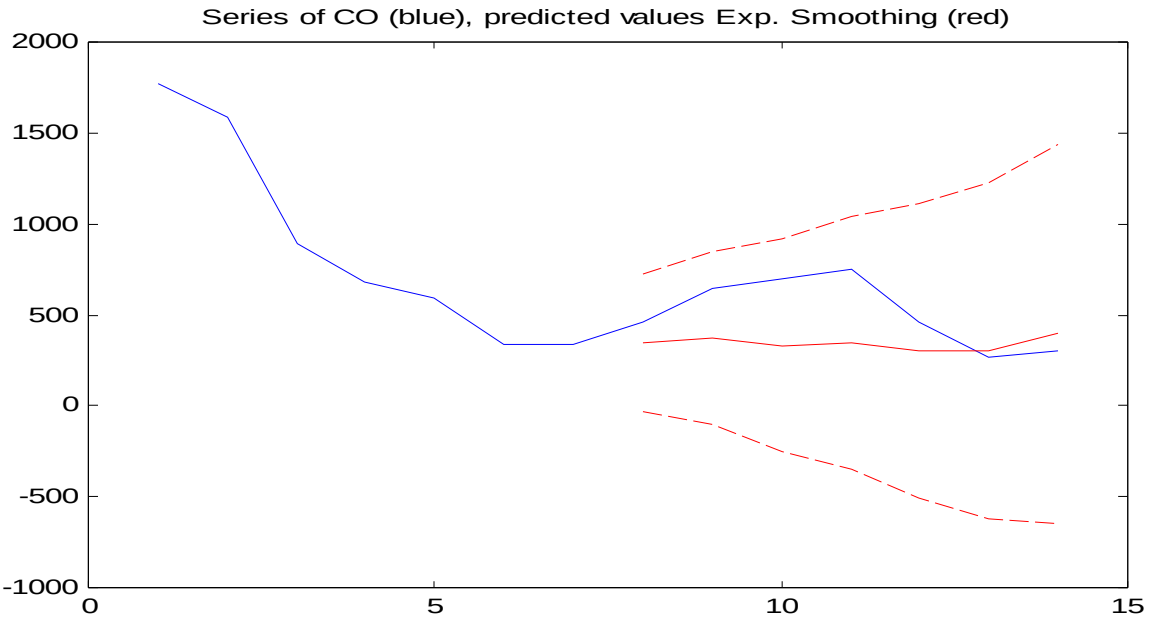


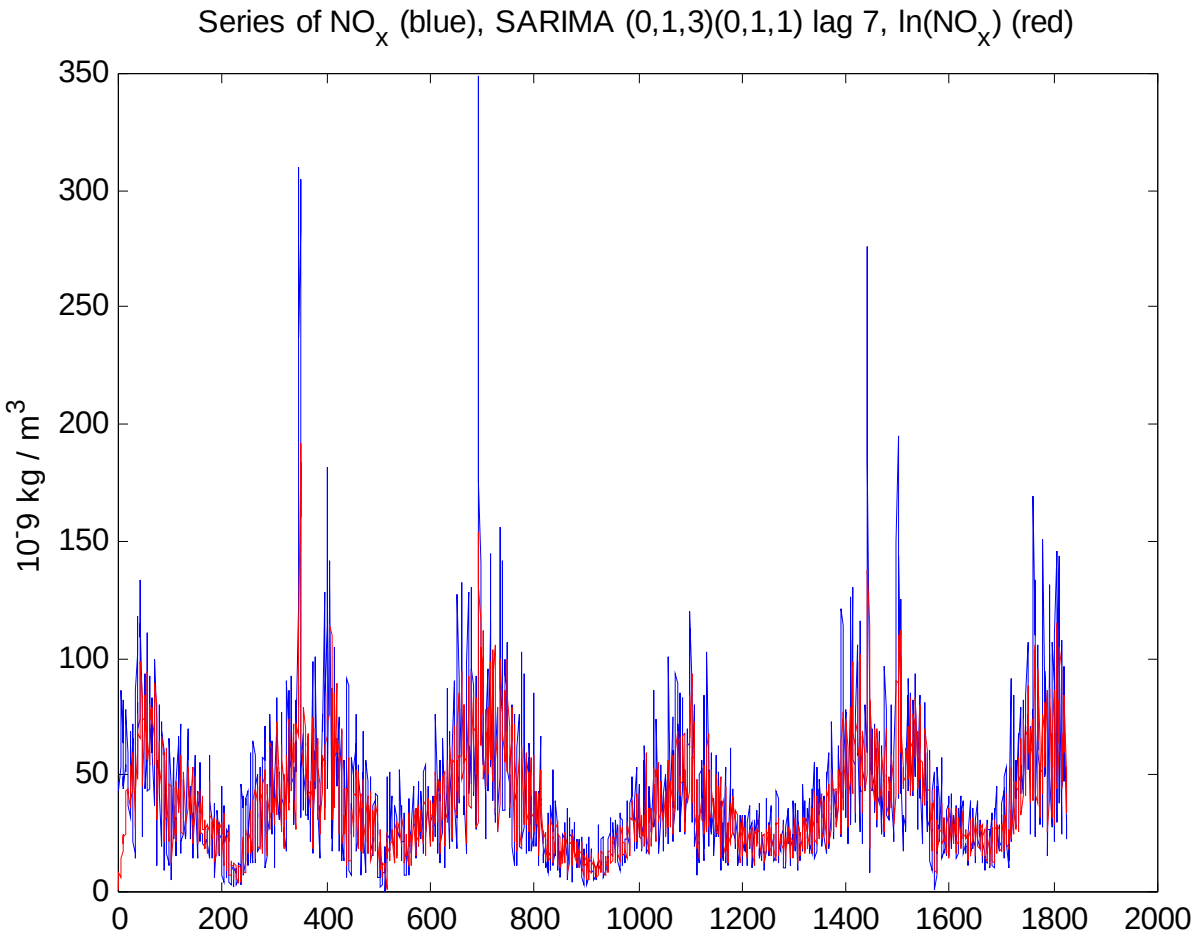


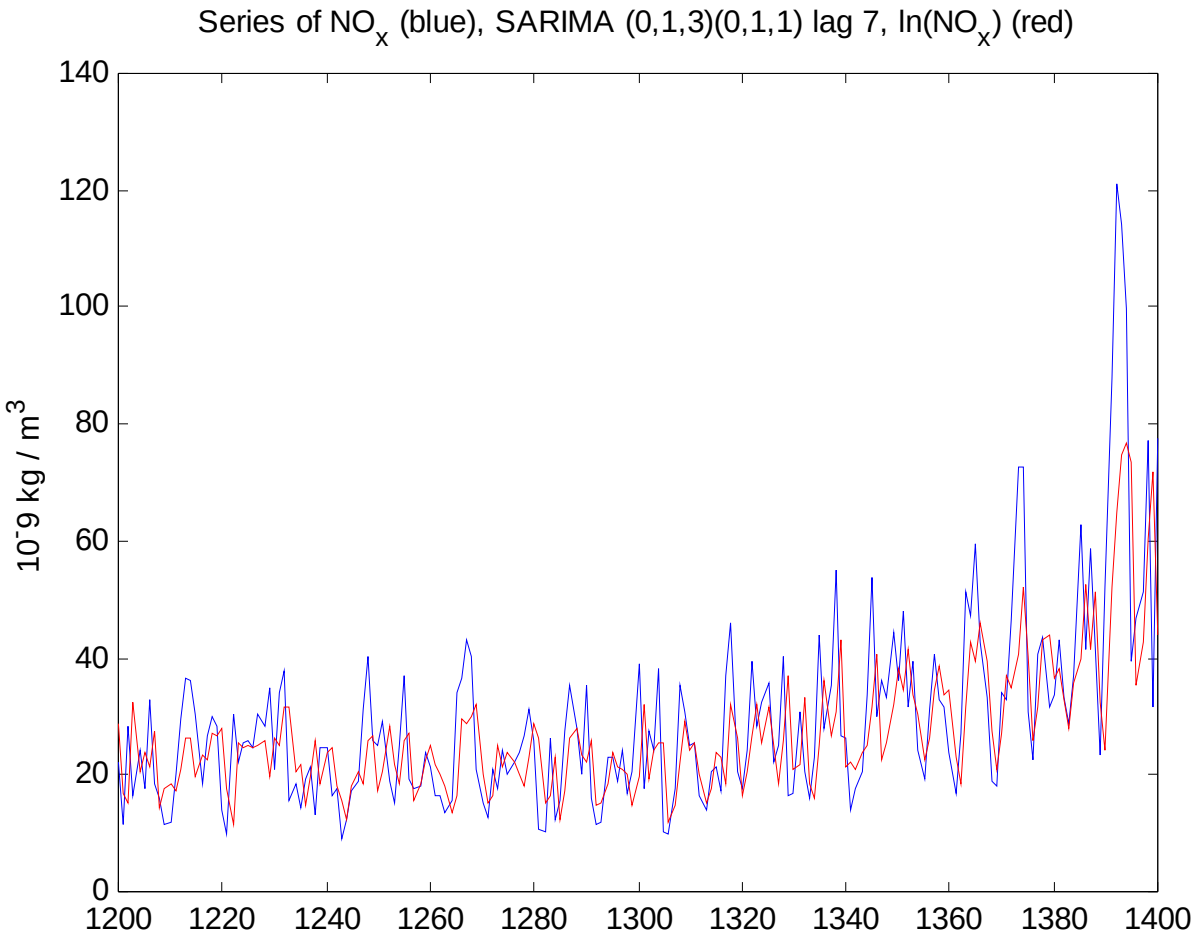


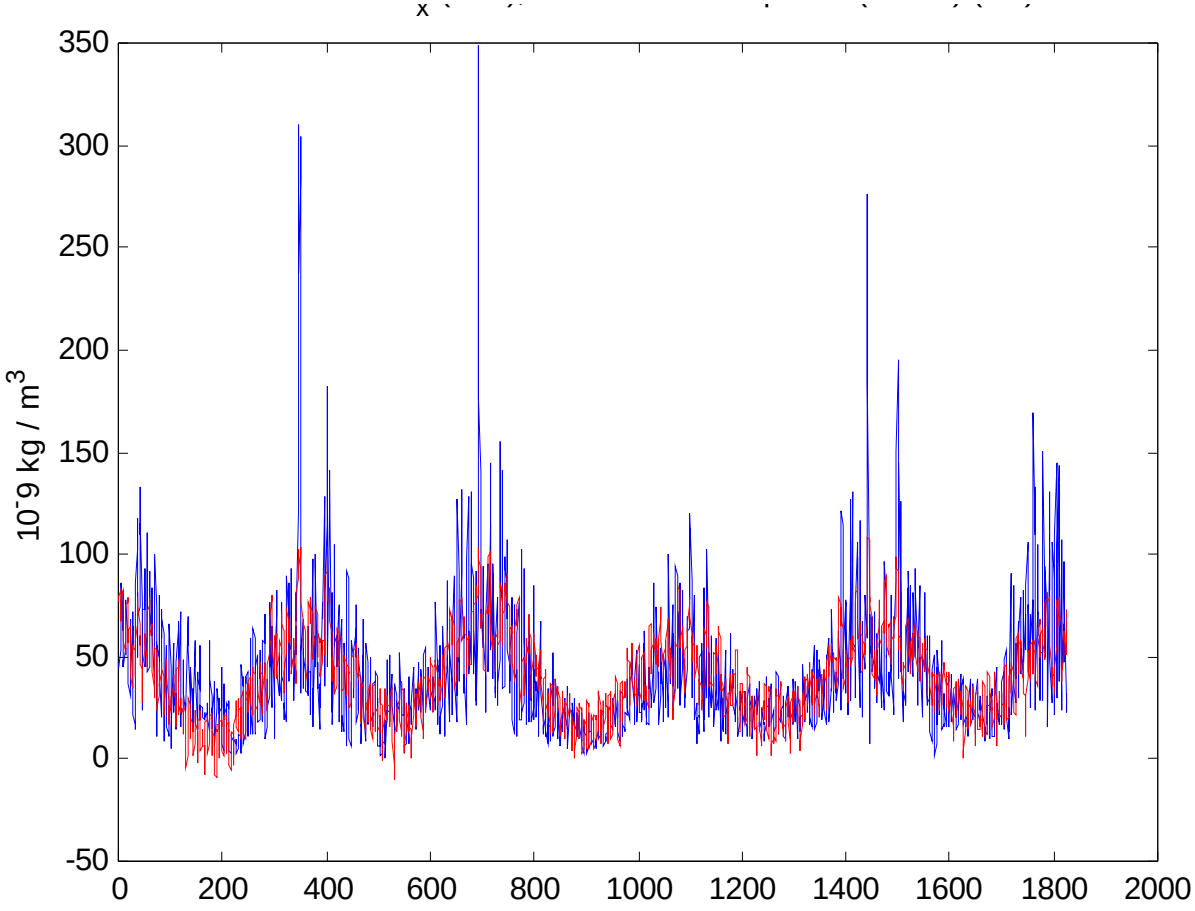












Series of NO_x (blue), model of hidden period (Fisher) (red)