Analysis of CO_2 data at Mauna Loa

A. C. Davison*

10th March 2017

Summary[1]

This report[2] describes a time series analysis of the CO_2 data from the Mauna Loa observatory, the so-called Keeling curve, which was one of the earliest measured signs of the effect of human activity on the global environment. The data show strong slightly super-linear trend and seasonality, rising over the period 1958–2008 from 315–385ppmv, and marked annual variation of around 5ppmv. A seasonal ARIMA model fits the data well. Predictions from it suggest that the level 400ppmv will be breached in around 2015, or perhaps slightly earlier.

*Institute of Mathematics, IMA-FSB-EPFL, Station 8, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland, `Anthony.Davison@epfl.ch`

[1]Summary: a short summary of the report, with approximately one sentence for each section. Should not contain formulae

[2]Use of footnotes should be avoided in scientific writing: if something is important enough to be included, it should be in the text. However in this document footnotes are used to highlight a few presentational points.

1 Initial data analysis

The data[3] we consider are monthly mean atmospheric carbon dioxide at Mauna Loa Observatory, Hawaii. These data, measured as the dry mole fraction—defined as the number of molecules of carbon dioxide divided by the number of molecules of dry air—and given in units of parts per million by volume (ppmv), constitute the longest record of direct observation of CO_2 in the atmosphere. Measurements were started in March of 1958 and are currently available to the end of November 2008. A graph of the data, the so-called Keeling curve, is considered as iconic in climate change, as it was one of the earliest indications of the effect of human activities on the global climate. The data may be downloaded from a link from the page

`http://www.esrl.noaa.gov/gmd/ccgg/trends/co2_data_mlo.html`

where further details may be found; see also the Wikipedia article on the 'Keeling curve' and the Scripps Institute of Oceanography website,

`http://scrippsco2.ucsd.edu/`

In this report we analyse these observations with a view towards prediction of the values to the year 2020.

Figure 1 shows the data.[4] The two most striking features are the strong upward trend, as the CO_2 level rises from around 315ppmv to around 385ppmv over the 50 years, and the pronounced seasonal pattern. The trend is essentially piecewise linear, but apparently slightly convex, suggested a possible acceleration of the trend over the period of the data, and particularly perhaps from the late 1990s. The seasonality accounts for variation of around 5ppmv, and seems to vary slightly from year to year.

STL decomposition (Cleveland *et al.*, 1990)[5] of the data, shown in Figure 2, confirms these remarks. The seasonality is due to annual changes in vegetation in the northern hemisphere: the level of carbon dioxide in the atmosphere decreases from northern spring onwards as new plant growth takes the gas out of the atmosphere through photosynthesis, and rises again in the northern autumn when plants and leaves die off and decay to release the gas. When plotted with a different aspect ratio, the cycle is clearly non-sinusoidal, and appears to fluctuate somewhat in magnitude, broadly being smaller from 1960–1975 and around the year 2000 than in the 1990s and at present. This variation is difficult to explain. If it is not random fluctuation, it may indicate changes in the carbon cycle due to a gradual increase in global population on which is superimposed the destruction of rainforests, the collapse of heavy industry in Eastern Europe after 1990, and the more recent rapid industrialisation of China. It might also be due partly to volcanic effects. The trend in the STL plot shows more clearly a broad convex form, and the residuals show some positive autocorrelation.

Figure 3 shows the periodogram of the data, without and with three different degrees of smoothing. The anticipated peak at one cycle/year stands out very clearly, as do other peaks

[3] Introductory section: presentation of the data; exploratory data analysis (trend, seasonality, correlograms, periodograms); purpose of the analysis.

[4] The caption for the figure should give the units of measurement and enough detail for the figure to be read without reference to the text. Both axes should be labelled. However, interpretation of the plot is left to the text.

[5] Note the format for references.

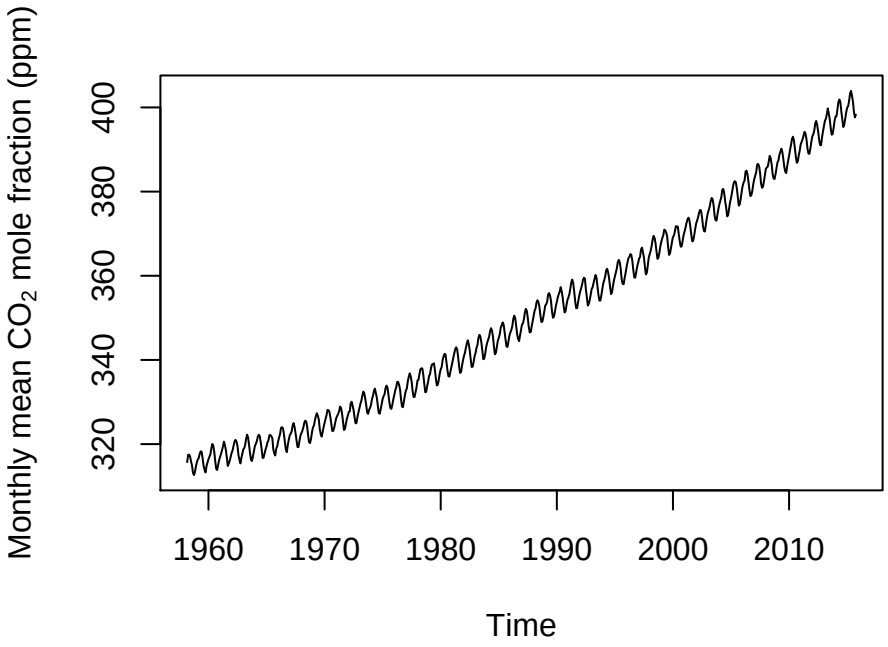

Figure 1: Mean monthly atmospheric CO_2 (ppmv) at Mauna Loa, Hawaii, from March 1958 to November 2008.

at 2, 3, and perhaps 4 cycles/year, which are harmonics of the annual cycle. There is also a peak on the left due to the overall trend, indistinguishable from very low-frequency variation, but apart from the harmonics and a dip at 3.5 cycles/year, the spectrum seems to be essentially flat from 2 cycles/year upwards.

With a view to the possible fitting of SARIMA models, Figure 4 shows the correlogram and partial correlogram of the data differenced at one lag to remove the trend and at 12 lags to remove the seasonal component, that is, of[6]

$$x_t = (I - B)(I - B^{12})y_t, \tag{1}$$

where $\{y_t\}$ represents the original data and B the backshift operator. Of the two plots the correlogram seems simpler: the strong negative correlation at lag 12, but not at lag 24, 36, ..., suggests a seasonal moving average component of order $Q = 1$, and the significant correlations at lags $h = 1, 11, 13$ suggest a moving average component of order $q = 1$.

An STL decomposition of the series $\{x_t\}$, not plotted here, shows no remaining systematic trend and only tiny seasonal variation, strongly suggesting that the differenced series is essentially stationary.

[6] Any displayed equations should be punctuated as parts of sentences. Punctuation rules of English grammar should be used for a report written in English, and of French grammar for a report written in French.

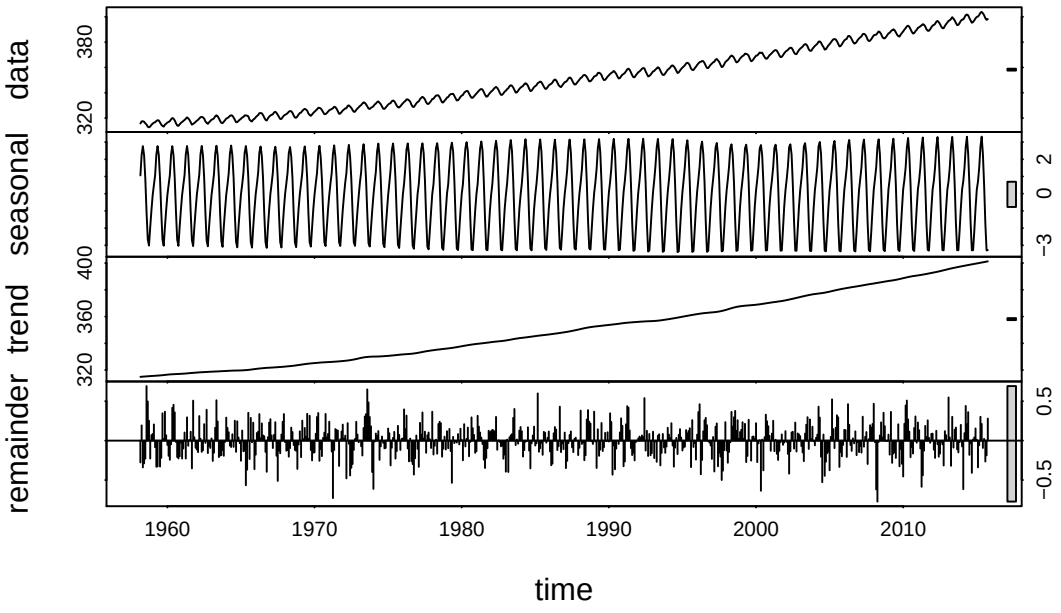

Figure 2: Mauna Loa data and its STL decomposition. Top panel: data. Second panel: seasonal effect. Third panel: trend. Fourth panel: residuals. The grey bar on the right is for comparison of the sizes of the different components.

2 Model-fitting

We now[7] try to fit to the data some seasonal ARIMA, $(p, d, q) \times (P, D, Q)_s$, models, using the usual notation; clearly for monthly data with an annual cycle, $s = 12$. In view of the trend and seasonality, and of the contents of Figure 4, it seems best to try fitting low-order ARMA models to the series x_t defined at (1), so $d = D = 1$. Table 1 shows the maximised log likelihood and AIC values for the models up to orders $(1, 1, 1) \times (1, 1, 1)_{12}$. The increases in log likelihood and corresponding drops in AIC are largest when moving average components are added to the model. The model with the smallest AIC among those fitted is $(1, 1, 1) \times (0, 1, 1)_{12}$, but three other ones are quite close: $(0, 1, 1) \times (1, 1, 1)_{12}$, $(1, 1, 1) \times (1, 1, 1)_{12}$, and the $(0, 1, 1) \times (0, 1, 1)_{12}$ that seemed indicated by the correlogram and partial correlogram. This last model and the best may be written as

$$x_t = (1 - 0.368_{0.042}B)(1 - 0.866_{0.022}B^{12})\varepsilon_t, \tag{2}$$

$$x_t = 0.186_{0.107}X_{t-1} + (1 - 0.530_{0.093}B)(1 - 0.866_{0.022}B^{12})\varepsilon_t, \tag{3}$$

with $\varepsilon_t \overset{\text{iid}}{\sim} (0, 0.3^2)$ in both cases, where B is the backshift operator and standard errors for coefficients appear as subscripts. The AR(1) coefficient of the model (3) is not quite significant

[7]Following sections: description of the statistical model(s); application to the data; results and interpretation of the chosen model(s).

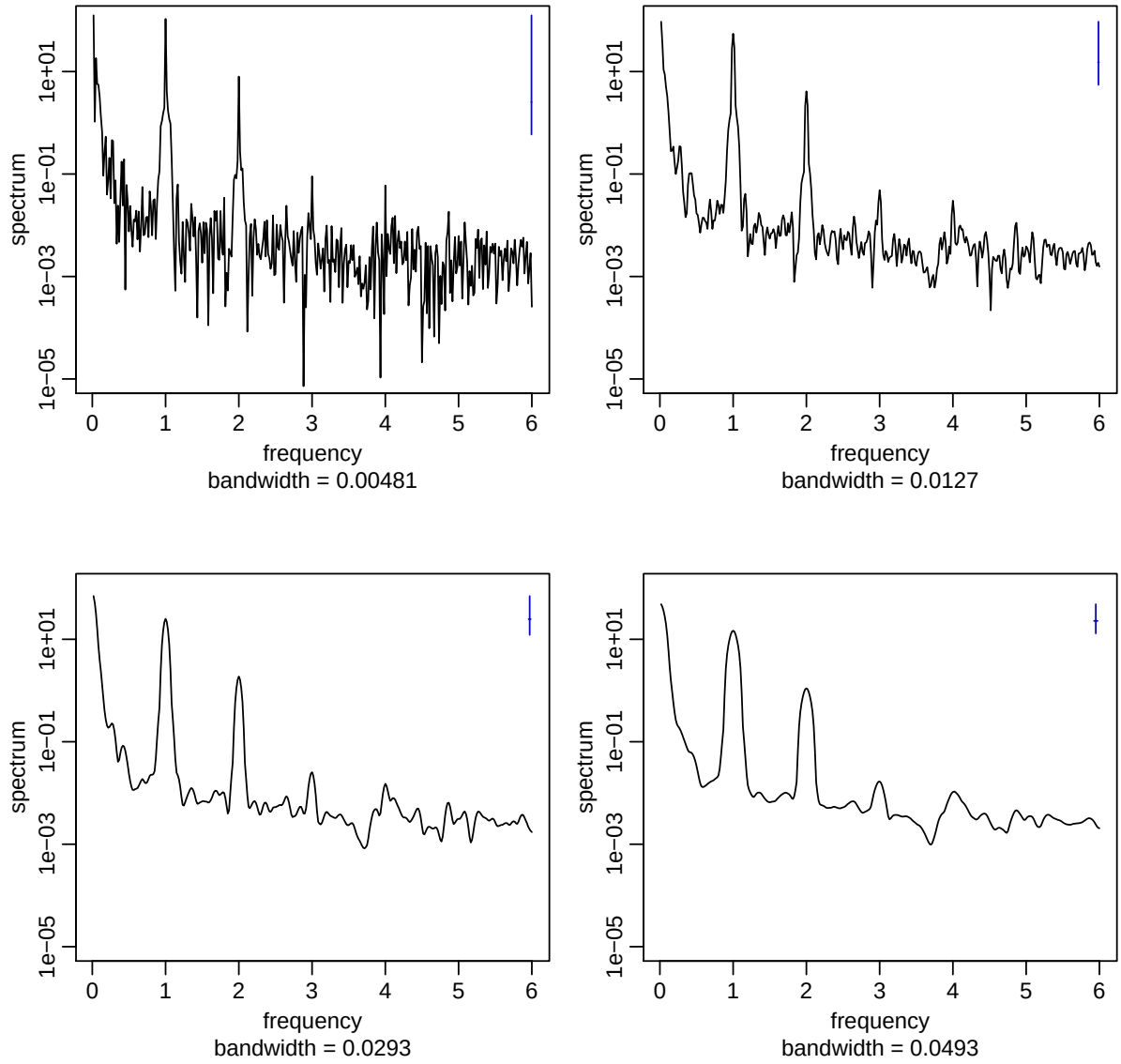

Figure 3: Frequency domain analysis of Mauna Loa data. Top left: raw periodogram. The remaining panels show successively more smoothed periodograms, with the effective bandwidth given under each panel and shown as the center line of the calibration bar at the top right of each panel. The height of the calibration bar shows significant variation in the periodogram.

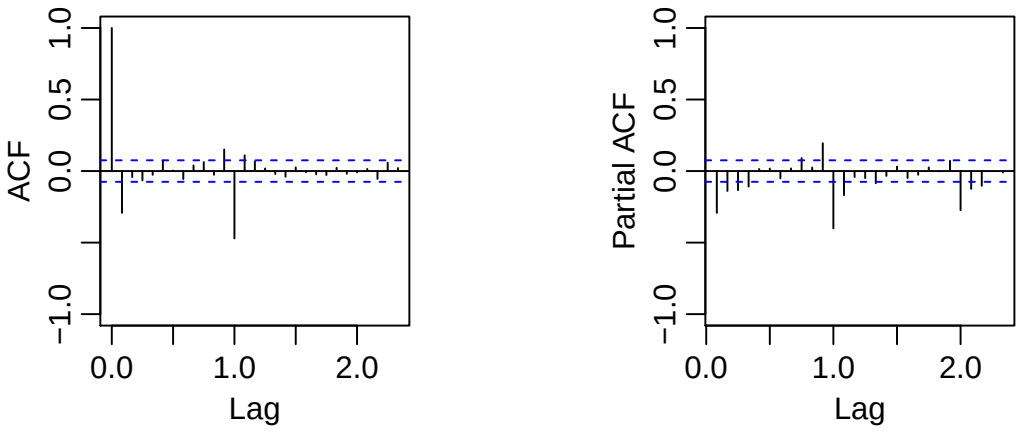

Figure 4: Correlogram and partial correlogram for the CO_2 data.

at the 95% level, and including it strongly affects the MA(1) coefficient: in fact $-0.530+0.186 = -0.344 \approx -0.368$, so that the effect of adding the autoregressive term is to dilute the moving average term, because $(1+\alpha B)^{-1} \approx 1-\alpha B$ for small α. As AIC has a tendency to suggest overfitting, and as the likelihood ratio statistics for comparisons among these four models are barely significant, if at all, parsimony pushes us to favour the simplest of them, (2). This has just two parameters, but the differencing that produced (1) has removed a trend parameter and the month effects. Thus expression (2) implies that the corresponding model for the original data y_t uses 15 parameters to explain the variation in 609 (correlated) observations; this seems a reasonable summary.

The interpretation of (2) in terms of the original data is complicated by the differencing. It is interesting, though perhaps a coincidence, that the variation in the data may be explained in terms of a moving average, which suggests that the CO_2 level at time t is related to the quantity added to the atmosphere over the past year. This interpretation has to be tempered by the facts that the moving average model applies to the differenced data X_t rather than to the original data, and that the moving average coefficients are negative, though this may be because of the differencing.

Plots of the usual time series diagnostics and of the cumulative periodogram, not shown here, confirm that the residuals from model (2) seem to be white noise, and a normal probability plot shows that the residuals are very close to normality.

3 Prediction

In this section we discuss prediction of future ozone levels to the year 2020, based on the models fitted above. One issue that often arises is that different almost equally plausible models may give rather different predictions, and then some sort of weighting may be required to produce

p	d	q	P	D	Q	Log likelihood	AIC
0	1	0	0	1	0	−375.55	753.1
1	1	0	0	1	0	−345.18	694.36
0	1	1	0	1	0	−334.6	673.21
0	1	0	1	1	0	−288.07	580.13
0	1	0	0	1	1	−192.4	388.79
1	1	0	1	1	0	−256.29	518.59
1	1	0	0	1	1	−164.07	334.13
0	1	1	1	1	0	−245.34	496.68
0	1	1	0	1	1	−154.47	314.93
1	1	1	1	1	0	−242.48	492.96
1	1	1	0	1	1	−151.47	310.95
0	1	1	1	1	1	−154.46	316.93
1	1	0	1	1	1	−164.05	336.11
1	1	1	1	1	1	−151.47	312.95

Table 1: Log likelihood and AIC values for SARIMA $(p, d, q) \times (P, D, Q)_{12}$ models fitted to the CO_2 data.

an overall predictor taking this into account. Often the weighting will be based on the relative plausibility of the different models, as judged by their relative likelihoods, analogous to Bayesian model averaging. Since the most plausible model in Table 1 has maximised log likelihood $\hat{\ell}$ of around -131.78, a model with $\hat{\ell} = -140$ would be roughly $\exp(140 - 132) \doteq e^8 = 2980$ times less likely than the best model, and it therefore seems reasonable to exclude models for which $\hat{\ell} < -140$.

Figure 5 shows the data from the year 2001 onwards, with predictions and their 95% confidence limits, for the six models in Table 1 for which the log likelihood exceeds -140. There is very little difference between them. The level of 400ppmv is predicted to be breached for the first time in winter 2014–15, but the confidence intervals suggest that this might occur two years earlier, and by 2020 it seems very likely that the level will exceed 400ppmv throughout the year. Closer inspection of the graphs suggests that these forecasts underestimate the rate of increase: the linear trend for the prediction period is below the average trend shown by the data over 1958–2008, because differencing to stabilise the series mean has not fully accounted for the slight convexity in the trend, which is therefore underestimated beyond the range of the data.

The confidence bands shown in the figures account for the variability of the innovations, but not for model selection uncertainty or for the estimation uncertainty for the parameters. Figure 5 suggests that model selection uncertainty is small, and as there are roughly 40 parameters for each observation, estimation uncertainty is small also. The biggest source of prediction uncertainty is likely to be the failure of the model to match the trend immediately before the prediction period. An attempt to account for this through a second differencing operation gave a

worse fit, so within the SARIMA framework there seems to be little that can be done to remedy this.

4 Discussion

The[8] seasonal ARIMA model (2) seems to fit the data well, and its closest competitors among the SARIMA models fitted give very similar predictions, so one can have high confidence in them. They are, however, compromised by the failure of the trends for the predictions to match the observed trend for the period immediately before 2008. Fitting the SARIMA model from 1995 or 2000 onwards, not shown, does not seem to remedy this by giving a higher trend. If more time were available, it would be worthwhile to investigate this more fully.

A alternative and perhaps preferable approach would be to construct a suitable structural model (Durbin and Koopman, 2001).

References

Cleveland, R. B., Cleveland, W. S., McRae, J. E. and Terpenning, I. (1990) STL: A seasonal-trend decomposition procedure based on Loess. *Journal of Official Statistics* **6**, 3–73.

Durbin, J. and Koopman, S. J. (2001) *Time Series Analysis by State Space Methods.* Oxford University Press. ISBN 0-19-852354-8.

[8]Final section: summary of the results; discussion of the advantages and limitations of the applied method; what else might be done, if time/data were available.

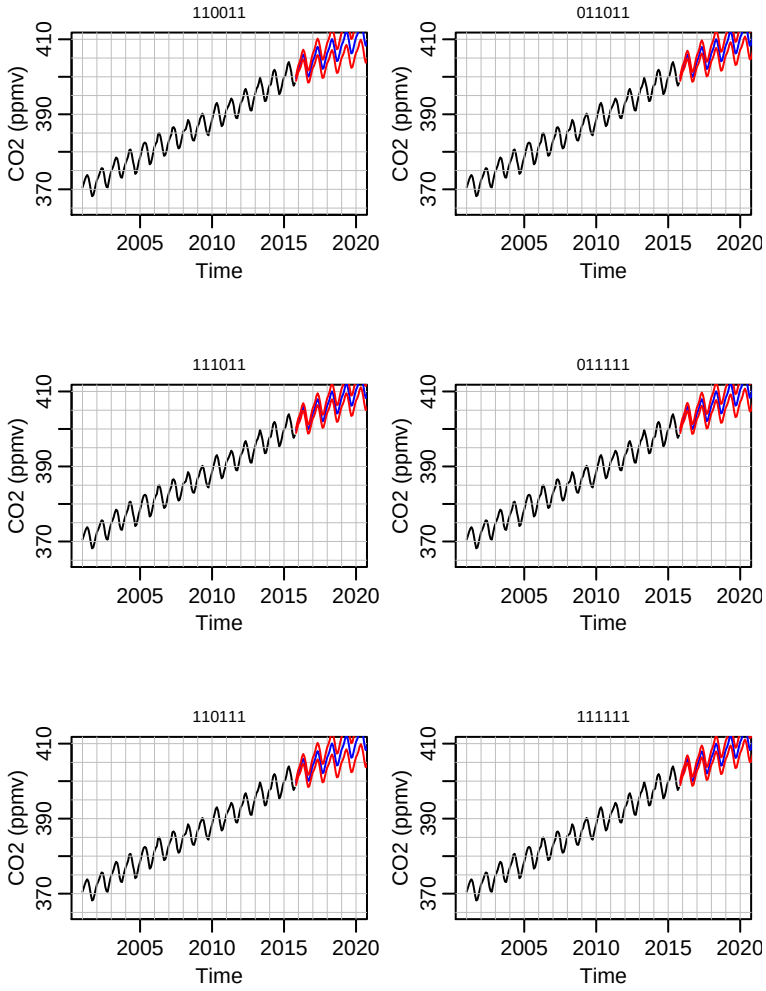

Figure 5: Comparison of predictions for levels of CO_2 (ppmv) for the period 2009–2020, for the six best models in Table 1. The data are shown in black, the predictions in blue, and 95% pointwise prediction bands in red. The grey grid is intended to aid comparison of the panels.