

Time Series HW 7

Kenny Flagg and Andrea Mack

October 27, 2016

Revisit your CO₂ concentration time series from HW 4. If you want to switch groups from that assignment you can or you can work in the same ones. If you switch groups, pick a time series to analyze from those you worked with. Groups of up to 3. And just one time series per group.

As in HW4, we will use the High Altitude Global Climate Observation Center, Mexico (MEX) dataset because we thought the High Altitude aspect may show interesting features of CO₂ concentrations not available in other datasets. The site is located at the coordinates 18.984, -97.311 near the summit of a 15,000 ft mountain.

The information page on these data indicates measured responses are on the X2007 CO₂ mole fraction scale. The excerpt from the information page gives insightful information about CO₂ and the data:

Carbon dioxide (CO₂) in ambient and standard air samples is detected using a non-dispersive infrared (NDIR) analyzer. The measurement of CO₂ in air is made relative to standards whose CO₂ mole fraction is determined with high precision and accuracy. Because detector response is non-linear in the range of atmospheric levels, ambient samples are bracketed during analysis by a set of reference standards used to calibrate detector response. Measurements are reported in units of micromol/mol (10^{-6} mol CO₂ per mol of dry air or parts per million (ppm)). Measurements are directly traceable to the WMO CO₂ mole fraction scale.

Uncertainty in the measurements of CO₂ from discrete samples has not yet been fully evaluated. Key components of it are our ability to propagate the WMO XCO₂ scale to working standards, the repeatability of the analyzers used for sample measurement, and agreement between pairs of samples collected simultaneously. Zhao and Tans (2006) determined that the internal consistency of working standards is +/- 0.02 ppm (68% confidence interval). The typical repeatability of the analyzers, based on repeated measurements of natural air from a cylinder, is +/- 0.03 ppm. Average pair agreement across the entire sampling network is +/- 0.1 ppm.

The Pacific Ocean Cruise (POC, travelling between the US west coast and New Zealand or Australia) data have been merged and grouped into 5 degree latitude bins. For the South China Sea cruises (SCS) the data are grouped in 3 degree latitude bins.

Sampling intervals are approximately weekly for the fixed sites and average one sample every 3 weeks per latitude zone for POC and about one sample every week per latitude for SCS.

Historically, samples have been collected using two general methods: flushing and then pressurizing glass flasks with a pump, or opening a stopcock on an evacuated glass flask; since 28 April 2003, only the former method is used. During each sampling event, a pair of flasks is filled.

1. *We are going to be exploring a more complete model that includes a long-term trend and a seasonal component. We need to pick the type of trend and the form of the seasonal component. For the long term trend, consider the following options: no trend, linear trend, or quadratic trend. For the seasonal component, consider no seasonal component, seasonal means, single harmonic pair ($m=1$), and 5th order harmonic ($m=5$). Consider all combinations of these components, fit using “lm”. Create a table that contains the model description, model used DF (so the count of free parameters), AICs, and Δ AICs, sorting the table by AICs. Use this information to discuss the top model selected (what was in it), the strength of support for that model versus the others, and the strength of evidence for a long-term trend and seasonal component (of the type selected) versus not including them in the model.*
 - *You should be creating a table that contains twelve models.*
2. *Now fit a “gam” from the “mgcv” package that includes a long-term trend based on a thin-plate spline with shrinkage that uses ‘k=years,bs=”ts”’ from the fractional year variable and a cyclic spline seasonal component. To build the cyclic spline component, use the numerically coded month variable that goes from 1 to 12 and k=12,bs=”cc”. Fit the model, plot the long-term trend and the seasonal component (use “plot(gam_model)”), and discuss the estimated components, using both the plots and the EDF of each term.*
3. *Calculate the AIC of the GAM using the “AIC” function and discuss how that result compares to your AICs in 1. How is it similar or different in terms of information (degrees of freedom) used?*
4. *Compare the fitted values of your GAM to those from your top model, plotting the two models’ results and the responses vs time on the same plot.*

A simulation study with autocorrelation present

5. *Revisit your simulation with an AR(1) from HW 6 10. Consider fitting a model with autocorrelation in it using “gls” from the “nlme” package that accounts for an MA(1) error and another that accounts for an AR(1) error. Run your simulation code, extracting the p-values from the two model summaries and estimate the type I error rate in each situation and compare it to what you get from the regular linear model.*
 - *In a “gls” model summary, the estimates, SEs, test statistics, and p-values are contained in the ‘tTable’ part of the summary. It has a similar layout to ‘coef’ that we pulled the [2,4] element from to get the p-value from “lm”.*

Some derivation practice (these can be handwritten).

If you have not completed STAT 421 or equivalent, please try the problem and then take advantage of advanced help by stopping by to chat about your answer.

6. Answer Cryer and Chan question 2.4 (page 20)
7. Suppose that we are interested in the properties of a local average (linear filter) of two observations from an original time series, x_t . The new series is $y_t = (0.5) * (x_{t-1} + x_t)$. The mean of x_t is 3, the variance of x_t is 4, and the correlation between any neighboring x_t 's is 0.5 (so $\text{cor}(x_t, x_{t-1}) = 0.5$). x_t 's more than two time points apart are uncorrelated (correlation is 0). Use the rules for means and variances of linear combinations to find $E(y_t)$, $\text{Var}(y_t)$, and $\text{Cov}(y_t, y_{t-1})$. Do not worry about what happens at the edges of the time series (for $t=1$ or $t=n$), only worry about t in general.
 - Note that you have some preliminary work to complete to go from the provided information to what you need to work on the three derivations requested.

References

R Code