Effect of Diurnal Data Gaps on Regression and FFT Analysis

Martin D. Guiles

SOEST, University of Hawaii at Manoa
1. Introduction

One advantage of using regression methods to analyze time series data, specifically in this case ADCP velocity, is the ability of the method to recover constituent amplitude from data that contains a significant number of missing points. However, in the case of the HOME moorings upper 75kHz ADCPs, we need to investigate the actual performance of the regression method due to the periodic nature of the missing data. It is a common signature of diel zooplankton migration to have data gaps during the night when there is few zooplankton in the ADCP’s range. This is the case for the A2 and C2 data. Some analysis has been done within this depth range of lossy data, but the viability of the method was not clear. We can look closely at this unique set of circumstances to determine what errors may be overlooked using complex demodulation.

It will be shown that the diurnal errors influence negatively estimates of $S_2$ amplitude in this specific regression analysis. Additionally, we compare band averaged fft spectral analysis with the regression analysis to determine the relative merits of each as pertaining to moving window amplitude reconstructions.
2. Modeled Data

A series was constructed consisting of a $M_2$ signal with 3.75cm/s amplitude and a $S_2$ signal with amplitude of 1.27cm/s. These amplitudes were extracted from the C2 75kHz ADCP of the HOME project. The depth bin used is 488m. The phase of the synthesized signal was matched to this ADCP record. Added to this was Gaussian white noise with a signal to noise ratio of 35. In figure 1 a portion of each data set is shown.

![Figure 1. Example of the velocity data with blue indicating the removed signal and green showing the piecewise Hermite polynomial replacement. On the left is the synthetic $M_2$ and $S_2$ record, on the right is ADCP data from the HOME C2 mooring at 488m.](image)

For the C2 data, we have extracted the pattern of missing data from a bin in the lossy upper range of the instrument. This pattern is then
applied to the depth we will analyze, a depth having very little missing data. In figure 2 a comparison of the missing data spectra is shown. The major features are replicated by the synthetic gaps, however it is of note that the semi-diurnal peak in the real missing data spectra is broader and exhibits a lower maximum. A possible explanation for the reduced semi-diurnal peak in the spectra of the real gaps is the presence of a second set of periodic gaps at near semi-diurnal frequencies and of a different phase than the diurnal harmonic. When constructing the synthetic gap series we chose a shorter duration (5hr) than many of the gaps in the real data at the worst depth bin, but the choice is representative of a larger depth range of the lossy data.

3. Analysis

To create a time series of amplitude (or other decomposition quantity like PSD), the length of window will determine not only the gravest frequency resolved but it will determine how well frequencies close together can be individually resolved (Rayleigh criteria). In this case we want to resolve $M_2$ and $S_2$, so for the fixed window test we have chosen a window near 15 days, which is near periodic for both frequencies. For the spectral
band averaged method we will apply a normalized Hanning window. This allows fine resolution of amplitude variations in time.

The initial comparison is between the complex demodulation of the lossy data vs. the original data (figure 3). In the lower panels the $S_2$ mean ratio is shown, representing the error introduced by the periodicity of the lossy data. Also indicated is the ratio of missing data. These panels show great similarity in the effects of analysis between the real and synthesized data sets. Clearly $M_2$ is recovered more accurately than $S_2$. We can observe that with the level of missing data near forty percent, the accuracy of the method is weak irregardless of the periodicity of the missing data.
However, with the modeled data having gaps of shorter duration, $M_2$ is recovered well, but the $S_2$ is being influenced by the periodicity of the gaps.

**Figure 3.** In the top panel $M_2$ amplitudes are shown for 15 day moving regression analysis. One series is on the original data, and the second is for data that has daily gaps (NaN) of five hours duration. The bottom panel is for the $S_2$ frequency, and reveals significant overestimation of amplitude for the lossy data series.

To examine how the tapered fft method compares with regression we constructed mean component amplitudes from each, with the fft method applied to each data set with the gaps represented by a piecewise Hermite polynomial interpolation. For the maximum velocity envelope of the modeled data a running mean was applied to the series prior to the addition of noise. To obtain a similar spring neap cycle representation for the actual
data a digital filter was used with two components near $M_2$ and two near $S_2$ then a running mean applied to the squared result. In figure 4 it is clear that the taper on the fft captures the spring neap cycle accurately in both cases. The amplitudes from the regression analysis are high from the periodicity of the gaps and additionally elevated in the case of the real data from the extent of missing data. Yet, for obtaining a correct amplitude representative of the long time period, the regression method is quite stable in the synthetic data case.
Figure 4. Reconstructed components are compared here between the regression and fft methods. On the left is the synthetic data with a moving average total velocity maximum signal depicting the spring neap cycle contained in the synthetic data. On the right are the semi-diurnal tidal components reconstructed from the C2 mooring. The spring neap cycle is reconstructed from a four component semi-diurnal band filter of the total record. The velocity envelopes are scaled to fit, and are not actual amplitudes. Actual mean component amplitude in indicated with black triangles.