Abstract

* Develop cross validation method for determining the best model for predicting missing data
* Includes selection of model eg AR lags and potential variables

Introduction

* Description of purpose for trying to predict missing data from these sensors
* Goal of results

Method

* Method for performing cross validation
* Description of error terms used
* Description of data management eg, how differing frequencies is dealt with, normalisation where needed
* Description of methods for variable choice

Results

* Tables of error results and determine the best model from each
* Different methods for variable selection shown

Discussion

* Compare results from different types of models
* Compare variable selection from different methods
* Comment on effectiveness of models in predicting the data.

# Introduction

The Australian Integrated Marine Observing System (IMOS) consists of an array of sensors providing a variety of monitoring data. For various reasons such as faulty sensors there may be missing or inaccurate data for various time periods. These periods can range from singular missing data points to weeks or months of missing data.

The goal of this project is to compare and analyse a variety of methods for predicting sections of missing data. The methods discussed in this will focus on Auto Regressive Moving Average (ARIMA), Generalised Linear Models (GLM), And Multi Layer Perception Regression (MLP) models. These models are to be compared for their general effectiveness at predicting various target timeseries, their ability for short step forecasting eg 1 period ahead, and for large period forecasting like the weeks of missing data mentioned above. A variety of methods are to be discussed for use in determining the ideal prediction variables for each of these methods, such as Dynamic time warp, Cross Correlation and Pearson’s r coefficient.

# Method

Data Gathering

The URL of the desired timeseries are taken from the IMOS database and downloaded into python. Multiple files are then combined together to generate one timeseries of each desired variable using a python script ats.readTimeSeries. For the desired target variable suitable timeseries to perform analyses with are chosen based on suitability depending on the nature of a system (eg relationship between variables) and their frequency. Variables of a higher frequency are split into multiple variables for each occurrence between the largest frequency. These are then used to generate a single file of the target variable and exogenous variables. Variables of a lower frequency then the target variable have not been used for this report.

The Data

The variables that will be used for this analysis are taken from Darwin sensors in particular, the biochemical data using variables TEMP,CNDC,TURB,PAR and CPHL taken in a 15min period, the wave data in particular WWSH taken in a 2hr period, and the wind speed taken in a 30min period.

Variable selection

Three methods of variable selection have been tested:

Dynamic time warp is one method, it is implemented by first normalising the timeseries beteen 0 and 1 before being run through the dynamic time warp. This determines the similarity of two timeseries by measuring the “distance” between two timeseries. It was implemented using fast dtw in python.

Cross Correlation function (CCF) is the second method used to determine suitable variables. The Time series were checked to be stationary and made stationary if required for CCF to be used. This calculates the Cross Correlation between different lags of two time series to determine one series ability to predict another.

Pearsons r coefficient was calculated between each variable to determine the variables with the highest linear correlation.

Cross Validation

K fold HV-blocked Cross Validation as described in J. Racine, Consistent cross-validatory model-selection for dependent data: hv-block cross-validation, Journal of Econometrics 99 (1) (2000) 39–61 and implemented in On the use of cross-validation for time series predictor evaluation Christoph Bergmeir ⇑ , José M. Benítez

This has been implemented by using the amount of k folds to determine the size of v for the hv-blocks.

The data is then split into these k folds where each k fold is used as the test set once, and for each test set all data not within the test set is used for training. From each side of the test set H data points are removed from the training set to ensure independence from the test set. H is determined by the amount lags in the dataset with significant autocorrelation and partial auto correlation.

ARIMA

For use with ARIMA models all data must be stationary and is checked to be using both KPSS and AdFuller tests to ensure stationarity. Differencing is used to make non stationary data stationary before testing. An ARMA model is then trained using each training set, and then used to predict the test set using one step ahead forecasts which is then un-differenced using the original timeseries. This is then compared to the original values of the test set and error measurers are produced. This is then repeated for a suitable range of AR and MA orders and the values recorded. The order with the lowest mean error is then chosen as the ideal order for the ARMA model.

MLP regression

For use with Multi-layer Perception regression (MLPR) the data is first normalised between 0 and 1 before being separated into folds for cross validation. Null values of the data must then be removed by row through all variables. The training sets are then used to produce an MLPR model which then predicts the test set. This test set is then un-normalised and used to produce error measures for each model.

GLM

For GLM’s the data is kept in its original form unlike the other methods. The same cross-validation method is used to separate the data, train a model using each training set then create predicted values using the test sets. These predicted values are then used to generate the error measures.

Error measures

Adjusted R Squared

Root mean square error (RMSE)

Weight Mean Absolute Percentage Error (WMAPE)

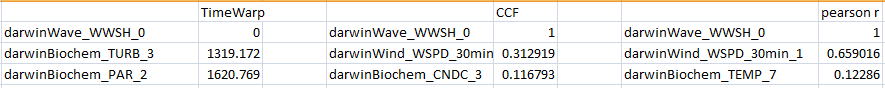
Relative Error

Relative error compares the error of the prediction to that of a benchmark method. For the purposes of this report the benchmark was taken as the previous value in the timeseries. Due to the nature of this benchmark was only used for ARIMA models as GLM and MLPR are not time dependant.

# Results

### WWSH

**Model Selection**

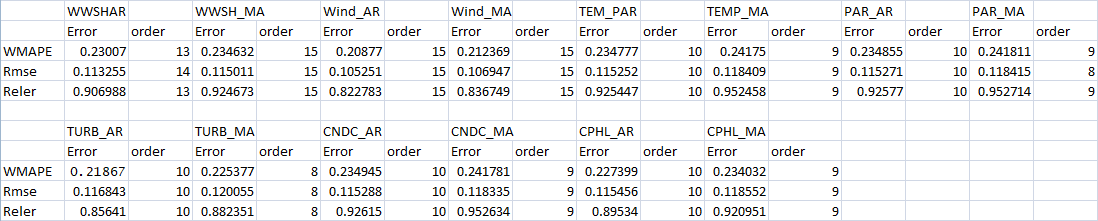


Each method appears to nominate differing variables for use as exogenous variables however CCF and Pearson r give more similar results both highlighting WSPD\_30min as the best predictor however CCF suggests CNDC will be the next best predictor while Pearson r suggests TEMP.

Timewarp suggests that TURB and PAR are the most similar time series and thus may be the best predictors.

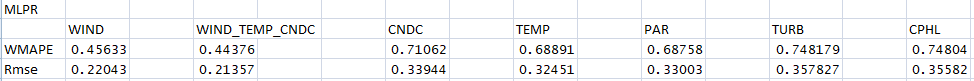
Results

ARIMA



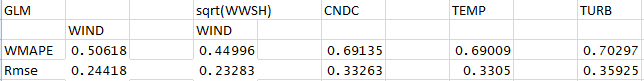
Looking at each of the results its clear WindSpd is the best exogenous variable for WWSH prediction as predicted by CCF and Pearson’s r. TURB was the next best predictor and provided some improvement but not as significant as WindSpd. CNDC and TEMP provided no improvement over an AR or MA model with no exogenous variables and may have even increased the error.  
Looking at the measures for CCF and Pearson r it is not surprising that CNDC and TEMP were poor predictors as their values are significantly lower than that of WSPD.

MLPR



Using a MLP model WINDSPD was once again the best predictor for WWSH, followed by TEMP then CNDC and finally TURB. WINDSPD was a significantly better predictor while the other predictors all provided similar results. It can also be seen that ARIMA models provided much lower error values then the MLP models however this is unsurprising

GLM



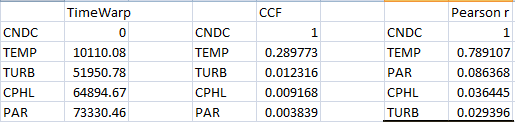
A Generalised Linear model provides a similar result to MLPR however with slightly higher errors.

Looking at residuals for WWSH and WindSPD a slight nonlinear trend could be seen and a sqrt transformation was used on WWSH, This provided significantly better results however still higher error then a standard MLPR

It is clear from these various methods that WindSPD is the best predictor of WWSH, and the other variables tested here provide poor results. CCF and pearsons r coefficient were significantly better then time warp.

CNDC

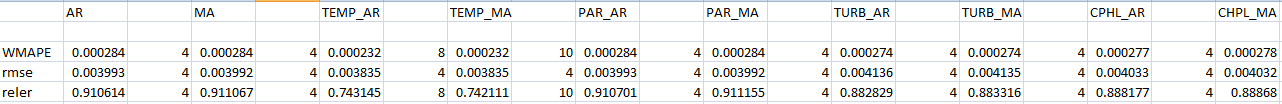
Model Selection



Observing the 3 methods we can see all choose temperature as the best predictor by a wide margin and CCF and time warp rank the differing variables identically. Pearsons r coefficient although having a different ordering still follows the same trend as the 3 predictors provide similarly bad results.

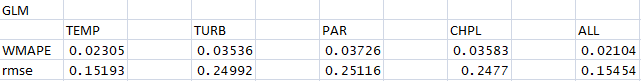
Results

ARIMA



Temperature was the variable that provided the best result as predicted by all 3 variable selection methods. The other variables did not provide any significant improvement over an ARMA model with no exogenous variables.

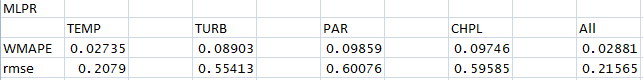
GLM

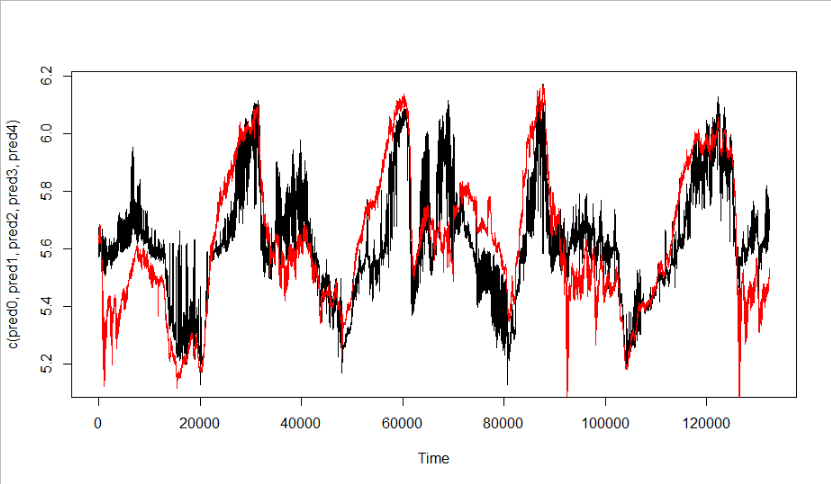


Temperature provides the best results once again while each other variable provides similar result.

Using all exogenous variables in a model provides a slight improvement over just temperature.

MLPR



Figure MLPR With All exogenous variables, Red= actual Values. Black = prediction

MLPR provides similar but slightly worse result then GLM with temperature providing the best results. However here using all variables results in higher error compared to just temperature.

# Discussion

Variable selection

The variable selections appeared to have similar accuracy, while CCF and Pearson’s r always predicted the best variable, Timewarp often also picked the same variable or often the 2nd best. Timewarps is more reliant on large amounts of data as in tests with smaller samples its accuracy was less reliable while CCF and pearsons’s r stayed consistent.

Models

As expected ARIMA models produced significantly less error then GLM and MLPR models producing less than half the error. This makes the ARIMA models incredibly efficient at short horizon predictions however due to the nature of the ARIMA models they are not suitable for large predictions. Observing the data we can see at times upwards of 1000 data points can be missing which could be justify the compromise in accuracy. It can also be seen ARIMA is a viable predictor for timeseries without exogenous variables while MLPR and GLM require at least one exogenous variables and improve significantly more. ARMA models also took significantly longer to run the more variables and higher orders of AR and MA over other methods.

Comparing MLPR and GLM both generally produced similar results however MLPR generally provided a slightly better result. Both methods can be seen to produce best results when using multiple exogenous variables and poor results with just one variable. As these methods are not time dependant and use no data from the target time series they are ideal for longer period forecasts which will be useful for longer periods of missing data.

Overall the models effectiveness of prediction varies by each variable.

CNDC and TEMP could be effectively predicted by all methods although ARIMA provided much better prediction.

WWSH and WINDSpd were less successful in their predictions with error roughly 20% the size of the values with arima and 50% using linear regression methods. These methods have shown potential and only used a limited selection of variables.