# Introduction

### CNDC missing values

Figure : Part of CNDC plotted to show the scale of missing data.

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 was to evaluate data imputation on oceanographic time series by the use of traditional time series forecasting methods (ARIMA) and exogenous variables. Additionally, the research exploited models based exclusively on the exogenous variables to evaluate how effective a data imputation method could be on the absence of time series data. The models to be evaluated are ARIMA with exogenous variables, Generalised Linear Models (GLM), And Multi Layer Perceptron Regression (MLP) models. ARIMA models have been analysed using 1 period ahead forecasting. 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

Data was sourced from the Integrated Marine Observing System (IMOS) - IMOS is a national collaborative research infrastructure, supported by the Australian Government. 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 which can be obtained from <https://github.com/santinoalves/timeseries> in the file adjustTimeSeries.py along with all other code and data files. 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. The data used was taken from 23/1/2014 – 9/4/2018 resulting in between 18000 - 140000 data points depending on the variable.

Variable selection

The variables were ranked against the target time series using three methods to determine their suitability to be used as exogenous variables. The top two variables from each method were then considered for analysis. The methods used were Dynamic Time Warp, Cross Correlation Function and Pearson’s r coefficient.

Three methods of variable selection have been tested:

Dynamic time warp , (2007) Dynamic Time Warping. In: Information Retrieval for Music and Motion. Springer, Berlin, Heidelberg, 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,Stan Salvador, and Philip Chan. “FastDTW: Toward accurate dynamic time warping in linear time and space.” Intelligent Data Analysis 11.5 (2007): 561-580.

Cross Correlation function (CCF), R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>, 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. Time series were made stationary through the use of differencing. 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

Adjusted r squared is a measure of the goodness of fit of a model to the data it is modelling adjusted for the numbers explanatory variables used in the model. A value of 1 means the model perfectly models the data. Miles, J. (2014). R Squared, Adjusted R Squared. In Wiley StatsRef: Statistics Reference Online (eds N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri and J.L. Teugels). doi:[10.1002/9781118445112.stat06627](https://doi.org/10.1002/9781118445112.stat06627)

Root mean square error (RMSE)

Root mean square error produces an error value in the same scale of the data. This allows it to be easily seen how much error there is in a more practical term. Barnston, A., (1992). “Correspondence among the Correlation [root mean square error] and Heidke Verification Measures; Refinement of the Heidke Score.”

Weight Mean Absolute Percentage Error (WMAPE)

WMAPE produces an error that describes the percentage of difference between the data and the error.

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**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| TimeWarp | |  | CCF | |  | Pearson r | |
| darwinBiochem\_TURB\_3 | 1319.172 |  | darwinWind\_WSPD\_30min\_0 | 0.312919 |  | darwinWind\_WSPD\_30min\_1 | 0.659016 |
| darwinBiochem\_PAR\_2 | 1620.769 |  | darwinBiochem\_CNDC\_3 | 0.116793 |  | darwinBiochem\_TEMP\_7 | 0.12286 |

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.

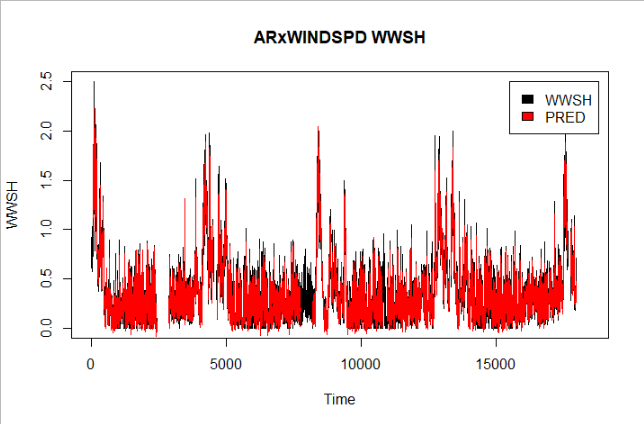
Timewarp suggests that TURB and PAR are the most similar time series and thus may be the best predictors.

## Results

ARIMA

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **AR** |  |  |  | **MA** |  |  |  |
| **No exo** | Error | Order | Standard Deviation | | **No Exo** | Error | Order | Standard Deviation |
| arsq | 0.879819 | 14 | 0.021953 |  | arsq | 0.876226 | 15 | 0.021387 |
| WMAPE | 0.23007 | 13 | 0.012907 |  | WMAPE | 0.234632 | 15 | 0.013989 |
| rmse | 0.113255 | 14 | 0.003949 |  | rmse | 0.115011 | 15 | 0.003975 |
| reler | 0.906988 | 13 | 0.024092 |  | reler | 0.924673 | 15 | 0.01625 |
| **CNDC** | Error | Order | Standard Deviation | | **CNDC** | Error | Order | Standard Deviation |
| arsq | 0.875863 | 10 | 0.02147 |  | arsq | 0.869308 | 9 | 0.02195 |
| WMAPE | 0.234945 | 10 | 0.013143 |  | WMAPE | 0.241781 | 9 | 0.015153 |
| rmse | 0.115288 | 10 | 0.003898 |  | rmse | 0.118335 | 9 | 0.00367 |
| reler | 0.92615 | 10 | 0.021474 |  | reler | 0.952634 | 9 | 0.011188 |
| **TEMP** | Error | Order | Standard Deviation | | **TEMP** | Error | Order | Standard Deviation |
| arsq | 0.875933 | 10 | 0.021548 |  | arsq | 0.869165 | 8 | 0.021939 |
| WMAPE | 0.234777 | 10 | 0.013251 |  | WMAPE | 0.24175 | 9 | 0.01535 |
| rmse | 0.115252 | 10 | 0.00401 |  | rmse | 0.118409 | 9 | 0.00387 |
| reler | 0.925447 | 10 | 0.020468 |  | reler | 0.952458 | 9 | 0.009417 |
| **CPHL** | Error | Order | Standard Deviation | | **CPHL** | Error | Order | Standard Deviation |
| arsq | 0.879134 | 10 | 0.024736 |  | arsq | 0.872697 | 8 | 0.025368 |
| WMAPE | 0.227399 | 10 | 0.021053 |  | WMAPE | 0.234032 | 9 | 0.022842 |
| rmse | 0.115456 | 10 | 0.004003 |  | rmse | 0.118552 | 9 | 0.003948 |
| reler | 0.89534 | 10 | 0.051076 |  | reler | 0.920951 | 9 | 0.048354 |
| **PAR** | Error | Order | Standard Deviation | | **PAR** | Error | Order | Standard Deviation |
| arsq | 0.875858 | 10 | 0.021881 |  | arsq | 0.869119 | 8 | 0.022259 |
| WMAPE | 0.234855 | 10 | 0.013209 |  | WMAPE | 0.241811 | 9 | 0.015279 |
| rmse | 0.115271 | 10 | 0.004034 |  | rmse | 0.118415 | 8 | 0.003939 |
| reler | 0.92577 | 10 | 0.020838 |  | reler | 0.952714 | 9 | 0.00979 |
| **TURB** | Error | Order | Standard Deviation | | **TURB** | Error | Order | Standard Deviation |
| arsq | 0.884397 | 10 | 0.026487 |  | arsq | 0.877833 | 8 | 0.028192 |
| WMAPE | 0.21867 | 10 | 0.041293 |  | WMAPE | 0.225377 | 8 | 0.044104 |
| rmse | 0.116843 | 10 | 0.006602 |  | rmse | 0.120055 | 8 | 0.006191 |
| reler | 0.85641 | 10 | 0.123496 |  | reler | 0.882351 | 8 | 0.13332 |
| **WIND** | Error | Order | Standard Deviation | | **WIND** | Error | Order | Standard Deviation |
| arsq | 0.897329 | 10 | 0.016206 |  | arsq | 0.892448 | 7 | 0.016129 |
| WMAPE | 0.213774 | 10 | 0.018478 |  | WMAPE | 0.218802 | 7 | 0.020613 |
| rmse | 0.107392 | 10 | 0.005822 |  | rmse | 0.109992 | 7 | 0.006391 |
| reler | 0.84231 | 10 | 0.047715 |  | reler | 0.861692 | 7 | 0.049236 |

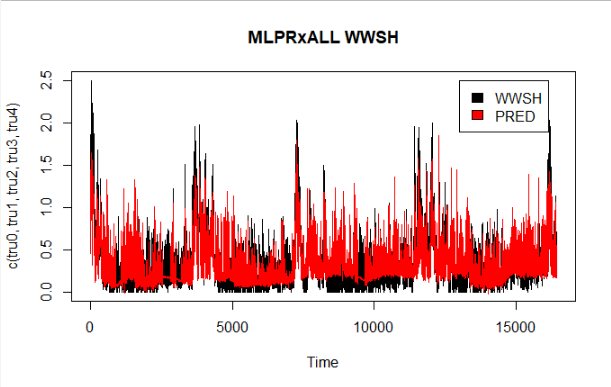
Looking at each of the results its clear WindSpd is the best exogenous variable for WWSH prediction as predicted by CCF and Pearson’s r. TURB was the next best predictor and provided some improvement but not as significant as WindSpd. CNDC and TEMP provided no improvement over an AR or MA model with 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 however it is surprising they were chosen over the other variables. In this case CCF and Pearson r coefficient were the best predictors of the variables.



MLPR

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **WIND+CNDC+TEMP** | Error | Standard deviation | | **PAR** | Error | Standard deviation |
| arsq | 0.58135 | 0.04436 |  | arsq | -0.00315 | 0.00797 |
| WMAPE | 0.443756 | 0.04186 |  | WMAPE | 0.68758 | 0.06048 |
| rmse | 0.21357 | 0.01783 |  | rmse | 0.33003 | 0.02889 |
|  |  |  |  |  |  |  |
| **CNDC** | Error | Standard deviation | | **TURB** | Error | Standard deviation |
| arsq | -0.06019 | 0.05387 |  | arsq | -0.1562 | 0.39312 |
| WMAPE | 0.71062 | 0.09845 |  | WMAPE | 0.74818 | 0.19969 |
| rmse | 0.33944 | 0.03358 |  | rmse | 0.35783 | 0.06204 |
|  |  |  |  |  |  |  |
| **TEMP** | Error | Standard deviation | | **WIND** | Error | Standard deviation |
| arsq | 0.03236 | 0.0864 |  | arsq | 0.55315 | 0.06858 |
| WMAPE | 0.68891 | 0.115822 |  | WMAPE | 0.45633 | 0.03726 |
| rmse | 0.32451 | 0.03849 |  | rmse | 0.22043 | 0.01812 |
|  |  |  |  |  |  |  |
| **CPHL** | Error | Standard deviation | |  |  |  |
| arsq | -0.1719 | 0.35579 |  |  |  |  |
| WMAPE | 0.74804 | 0.17779 |  |  |  |  |
| rmse | 0.35582 | 0.0536 |  |  |  |  |

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **WIND+CNDC+TEMP** | Error | Standard deviation | | **TURB** | Error | Standard deviation |
| arsq | 0.454903 | 0.146489 |  | arsq | -0.1106 | 0.38401 |
| WMAPE | 0.502732 | 0.133917 |  | WMAPE | 0.70297 | 0.20438 |
| rmse | 0.258831 | 0.04125 |  | rmse | 0.35925 | 0.05724 |
|  |  |  |  |  |  |  |
| **CNDC** | Error | Standard deviation | | **WIND** | Error | Standard deviation |
| arsq | 0.0155 | 0.00828 |  | arsq | 0.47419 | 0.06701 |
| WMAPE | 0.69135 | 0.05937 |  | WMAPE | 0.50618 | 0.08284 |
| rmse | 0.33263 | 0.02999 |  | rmse | 0.24418 | 0.02642 |
|  |  |  |  |  |  |  |
| **TEMP** | Error | Standard deviation | |  |  |  |
| arsq | -0.0026 | 0.03967 |  |  |  |  |
| WMAPE | 0.69009 | 0.07483 |  |  |  |  |
| rmse | 0.3305 | 0.03071 |  |  |  |  |

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

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | TimeWarp | | CCF | | | Pearson r | |
| CNDC | 0 |  | CNDC | 1 |  | CNDC | 1 |
| TEMP | 10110.08 |  | TEMP | 0.289773 |  | TEMP | 0.789107 |
| TURB | 51950.78 |  | TURB | 0.012316 |  | PAR | 0.086368 |
| CPHL | 64894.67 |  | CPHL | 0.009168 |  | CPHL | 0.036445 |
| PAR | 73330.46 |  | PAR | 0.003839 |  | TURB | 0.029396 |

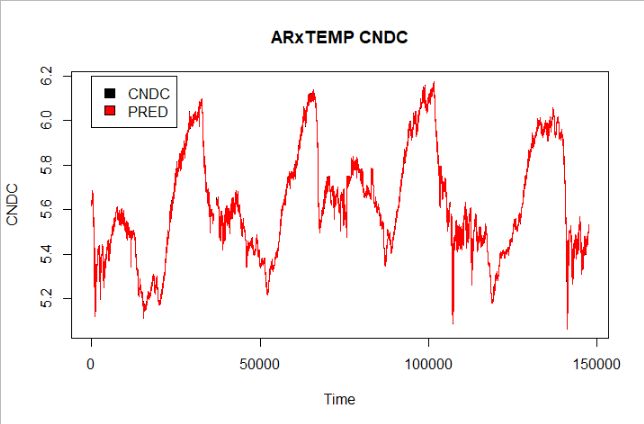
Observing the 3 methods we can see all choose temperature as the best predictor by a wide margin and CCF and time warp rank the differing variables identically. Pearsons r coefficient although having a different ordering still follows the same trend as the 3 predictors provide similarly bad results.

Results

ARIMA

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| AR |  |  |  | MA |  |  |  |
| No exo | Error | Order | Standard Deviation | No exo | Error | Order | Standard Deviation |
| **arsq** | 0.9996791 | 3 | 7.62E-05 | **arsq** | 0.9996792 | 4 | 7.60E-05 |
| **WMAPE** | 0.0002839 | 4 | 3.25E-05 | **WMAPE** | 0.0002841 | 4 | 3.24E-05 |
| **rmse** | 0.0039933 | 4 | 9.46E-04 | **rmse** | 0.0039925 | 4 | 9.45E-04 |
| **reler** | 0.9106141 | 4 | 1.44E-02 | **reler** | 0.9110674 | 4 | 1.42E-02 |
| **TEMP** | Error | Order | Standard Deviation | **TEMP** | Error | Order | Standard Deviation |
| **arsq** | 0.9997039 | 3 | 7.18E-05 | **arsq** | 0.9997039 | 4 | 7.16E-05 |
| **WMAPE** | 0.0002324 | 8 | 3.59E-05 | **WMAPE** | 0.000232 | 10 | 3.53E-05 |
| **rmse** | 0.0038353 | 4 | 9.17E-04 | **rmse** | 0.0038353 | 4 | 9.16E-04 |
| **reler** | 0.7431453 | 8 | 4.22E-02 | **reler** | 0.7421107 | 10 | 4.19E-02 |
| **TURB** | Error | Order | Standard Deviation | **TURB** | Error | Order | Standard Deviation |
| **arsq** | 0.9996835 | 3 | 7.35E-05 | **arsq** | 0.9996837 | 4 | 7.34E-05 |
| **WMAPE** | 0.0002735 | 4 | 1.59E-05 | **WMAPE** | 0.0002737 | 4 | 1.59E-05 |
| **rmse** | 0.0041361 | 4 | 1.10E-03 | **rmse** | 0.0041353 | 4 | 1.10E-03 |
| **reler** | 0.8828291 | 4 | 7.08E-02 | **reler** | 0.883316 | 4 | 7.11E-02 |
| **PAR** | Error | Order | Standard Deviation | **PAR** | Error | Order | Standard Deviation |
| **arsq** | 0.9996791 | 3 | 7.618E-05 | **arsq** | 0.9996792 | 4 | 7.60E-05 |
| **WMAPE** | 0.000284 | 4 | 3.245E-05 | **WMAPE** | 0.0002841 | 4 | 3.24E-05 |
| **rmse** | 0.0039933 | 4 | 0.0009465 | **rmse** | 0.0039925 | 4 | 9.45E-04 |
| **reler** | 0.9107014 | 4 | 0.0142685 | **reler** | 0.9111549 | 4 | 1.41E-02 |
| **CPHL** | Error | Order | Standard Deviation | **CPHL** | Error | Order | Standard Deviation |
| **arsq** | 0.9996826 | 4 | 7.78E-05 | **arsq** | 0.9996828 | 4 | 7.76E-05 |
| **WMAPE** | 0.0002774 | 4 | 3.69E-05 | **WMAPE** | 0.0002775 | 4 | 3.68E-05 |
| **rmse** | 0.0040328 | 4 | 9.13E-04 | **rmse** | 0.0040319 | 4 | 9.12E-04 |
| **reler** | 0.8881769 | 4 | 2.46E-02 | **reler** | 0.88868 | 4 | 2.47E-02 |

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.



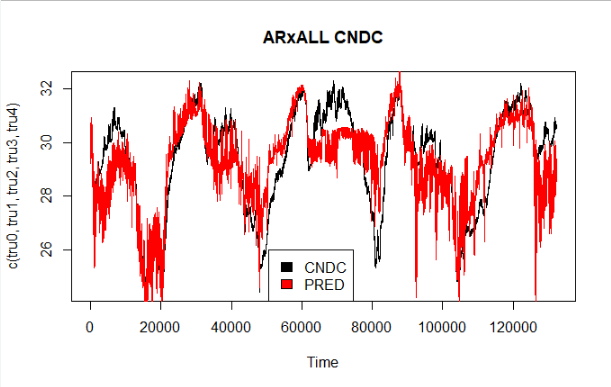
GLM

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TEMP+TURB+PAR+CPHL** | Error | Standard deviation | | **TURB** | Error | Standard deviation |
| arsq | 0.51999 | 0.26247 |  | arsq | -0.2909 | 0.57759 |
| WMAPE | 0.02104 | 0.00235 |  | WMAPE | 0.03536 | 0.00844 |
| rmse | 0.15454 | 0.01679 |  | rmse | 0.24992 | 0.02618 |
|  |  |  |  |  |  |  |
| **CPHL** | Error | Standard deviation | | **PAR** | Error | Standard deviation |
| arsq | -0.2371 | 0.37436 |  | arsq | -0.31 | 0.32322 |
| WMAPE | 0.03583 | 0.00874 |  | WMAPE | 0.03726 | 0.00706 |
| rmse | 0.2477 | 0.03676 |  | rmse | 0.25116 | 0.02937 |
|  |  |  |  |  |  |  |
| **TEMP** | Error | Standard deviation | |  |  |  |
| arsq | 0.51623 | 0.14894 |  |  |  |  |
| WMAPE | 0.02305 | 0.00375 |  |  |  |  |
| rmse | 0.15193 | 0.02204 |  |  |  |  |

Temperature provides the best results once again while each other variable provides similar result however including more variables increased the deviation in error between each test set. This means just temperature is the likely the best model as it appears more consistent compared to a model with all parameters.

MLPR

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TEMP+TURB+PAR+CPHL** | Error | Standard deviation | | **TURB** | Error | Standard deviation |
| arsq | 0.5222 | 0.33824 |  | arsq | -0.343 | 0.44259 |
| WMAPE | 0.020569 | 0.00520 |  | WMAPE | 0.03693 | 0.00583 |
| rmse | 0.13892 | 0.02922 |  | rmse | 0.24726 | 0.02355 |
|  |  |  |  |  |  |  |
| **CPHL** | Error | Standard deviation | | **PAR** | Error | Standard deviation |
| arsq | -0.3195 | 0.31529 |  | arsq | -0.3007 | 0.31897 |
| WMAPE | 0.03762 | 0.00678 |  | WMAPE | 0.0369 | 0.00715 |
| rmse | 0.25019 | 0.02973 |  | rmse | 0.25037 | 0.02993 |
|  |  |  |  |  |  |  |
| **TEMP** | Error | Standard deviation | |  |  |  |
| arsq | 0.52092 | 0.38621 |  |  |  |  |
| WMAPE | 0.02119 | 0.0051 |  |  |  |  |
| rmse | 0.14146 | 0.03428 |  |  |  |  |



MLPR provides similar but slightly better result then GLM with temperature providing the best results. Unlike the GLM the MLPR provides a better result using all variables as predictor variables in both error and standard deviation between folds.

TEMPERATURE

Model Selection

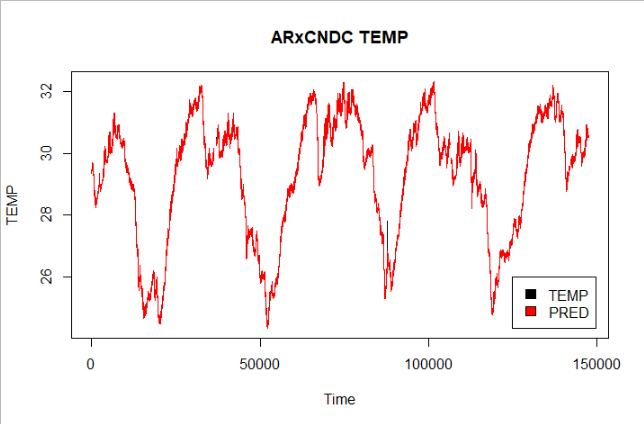
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Time Warp | | CCF |  | Pearson's r | |
| TEMP | 0 | **TEMP** | 1 | **TEMP** | 1 |
| CNDC | 10110.08 | **CNDC** | 0.2897728 | **CNDC** | 0.7891069 |
| TURB | 61832.26 | **PAR** | 0.1835005 | **TURB** | 0.1171451 |
| CPHL | 72948.56 | **CPHL** | 0.0090274 | **CPHL** | 0.0796872 |
| PAR | 88800.53 | **TURB** | 0.0081734 | **PAR** | 0.0133752 |

Dynamic Time Warp and Pearsons’s r coefficient provide the same ranking in variables whole CCF also chooses the same best variable of CNDC it has a different ordering for the other variables.

ARIMA

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| AR |  |  |  | MA |  |  |  |
| No exog | Error | Order | Standard Deviation | No exog | Error | Order | Standard Deviation |
| **arsq** | 0.9999307 | 6 | 4.13E-05 | **arsq** | 0.9999308 | 10 | 4.13E-05 |
| **WMAPE** | 0.0002708 | 7 | 2.17E-05 | **WMAPE** | 0.0002708 | 9 | 2.17E-05 |
| **rmse** | 0.0143743 | 7 | 2.41E-03 | **rmse** | 0.0143716 | 10 | 2.40E-03 |
| **reler** | 0.821519 | 1 | 1.39E-02 | **reler** | 0.82168 | 9 | 1.46E-02 |
| **CNDC** | Error | Order | Standard Deviation | **CNDC** | Error | Order | Standard Deviation |
| **arsq** | 0.9999393 | 9 | 3.32E-05 | **arsq** | 0.9999393 | 10 | 3.31E-05 |
| **WMAPE** | 0.0002541 | 1 | 2.35E-05 | **WMAPE** | 0.0002549 | 9 | 2.26E-05 |
| **rmse** | 0.0135596 | 6 | 2.02E-03 | **rmse** | 0.0135561 | 10 | 2.02E-03 |
| **reler** | 0.7702271 | 1 | 1.04E-02 | **reler** | 0.7728237 | 9 | 1.10E-02 |
| **TURB** | Error | Order | Standard Deviation | **TURB** | Error | Order | Standard Deviation |
| **arsq** | 0.9999393 | 6 | 3.01E-05 | **arsq** | 0.9999393 | 10 | 3.00E-05 |
| **WMAPE** | 0.0002556 | 7 | 3.31E-05 | **WMAPE** | 0.0002557 | 9 | 3.31E-05 |
| **rmse** | 0.0142455 | 6 | 2.09E-03 | **rmse** | 0.0142426 | 10 | 2.09E-03 |
| **reler** | 0.7765169 | 1 | 8.76E-02 | **reler** | 0.7767526 | 9 | 8.91E-02 |
| **PAR** | Error | Order | Standard Deviation | **PAR** | Error | Order | Standard Deviation |
| **arsq** | 0.999931 | 4 | 4.00E-05 | **arsq** | 0.9999308 | 10 | 4.04E-05 |
| **WMAPE** | 0.0002729 | 8 | 1.98E-05 | **WMAPE** | 0.000273 | 9 | 1.98E-05 |
| **rmse** | 0.0143842 | 4 | 2.26E-03 | **rmse** | 0.0144007 | 10 | 2.29E-03 |
| **reler** | 0.8285641 | 8 | 1.83E-02 | **reler** | 0.8287301 | 9 | 1.82E-02 |
| **CPHL** | Error | Order | Standard Deviation | **CPHL** | Error | Order | Standard Deviation |
| **arsq** | 0.9999362 | 6 | 3.51E-05 | **arsq** | 0.9999362 | 10 | 3.51E-05 |
| **WMAPE** | 0.0002611 | 9 | 2.41E-05 | **WMAPE** | 0.0002612 | 9 | 2.41E-05 |
| **rmse** | 0.0141304 | 6 | 1.92E-03 | **rmse** | 0.0141275 | 10 | 1.92E-03 |
| **reler** | 0.7923168 | 7 | 4.04E-02 | **reler** | 0.7924922 | 9 | 4.04E-02 |

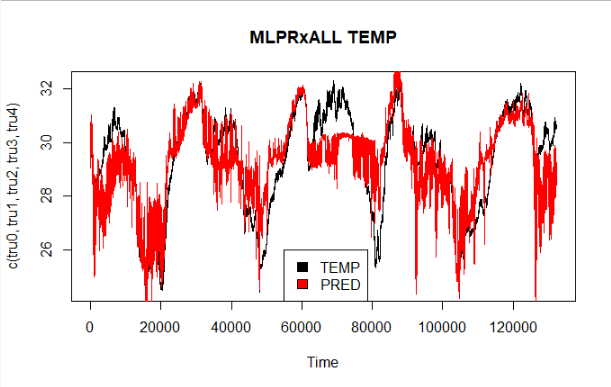
Looking at the AR and MA results it appears that CNDC is the best exogenous variable with TURB close behind. However these results are only a minor improvement over an AR or MA model with no exogenous variables. Dynamic Time Warp and Pearson’s r coefficient appear to have best predicted the exogenous variables.



MLPR

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **CNDC+TURB+PAR+CPHL** | Error | Standard deviation | | **TURB** | Error | Standard deviation |
| arsq | 0.58072 | 0.20179 |  | arsq | -0.2239 | 0.23777 |
| WMAPE | 0.02918 | 0.00324 |  | WMAPE | 0.05641 | 0.01006 |
| rmse | 1.07419 | 0.11331 |  | rmse | 1.97626 | 1.97626 |
|  |  |  |  |  |  |  |
| **CPHL** | Error | Standard deviation | | **PAR** | Error | Standard deviation |
| arsq | -0.1602 | 0.18417 |  | arsq | -0.2177 | 0.17556 |
| WMAPE | 0.0546 | 0.00716 |  | WMAPE | 0.0583 | 0.00797 |
| rmse | 1.93974 | 0.30204 |  | rmse | 2.03916 | 0.29923 |
|  |  |  |  |  |  |  |
| **CNDC** | Error | Standard deviation | |  |  |  |
| arsq | 0.56818 | 0.11973 |  |  |  |  |
| WMAPE | 0.03198 | 0.00401 |  |  |  |  |
| rmse | 1.19204 | 0.05763 |  |  |  |  |

CNDC is once again the best exogenous variable however there is a significantly larger difference in the quality of predictions by the other variables. Using multiple variables also provides a significant improvement over just CNDC and lowers some of the standard deviation between test sets.



GLM

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **CNDC+TURB+PAR+CPHL** | Error | Standard deviation | | **TURB** | Error | Standard deviation |
| arsq | 0.58785 | 0.1474 |  | arsq | -0.0886 | 0.27977 |
| WMAPE | 0.03091 | 0.00367 |  | WMAPE | 0.05438 | 0.01485 |
| rmse | 1.21565 | 0.05962 |  | rmse | 1.98092 | 0.34882 |
|  |  |  |  |  |  |  |
| **CPHL** | Error | Standard deviation | | **PAR** | Error | Standard deviation |
| arsq | -0.081 | 0.23179 |  | arsq | -0.2084 | 0.20211 |
| WMAPE | 0.05305 | 0.00671 |  | WMAPE | 0.05809 | 0.00845 |
| rmse | 1.94392 | 0.28078 |  | rmse | 2.02951 | 0.29635 |
|  |  |  |  |  |  |  |
| **CNDC** | Error | Standard deviation | |  |  |  |
| arsq | 0.51859 | 0.16241 |  |  |  |  |
| WMAPE | 0.03637 | 0.00333 |  |  |  |  |
| rmse | 1.24932 | 0.05787 |  |  |  |  |

A GLM appears to produce similar but slightly worse results then the MLPR, specifically when all variables are included the GLM only slightly improves over the CNDC model whereas the MLPR greatly improved.

WINDSPD

Model Selection

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Time warp |  | CCF |  | Pearson's r |  |
| darwinWind\_WSPD\_30min\_0 | 0 | darwinWind\_WSPD\_30min\_0 | 1 | darwinWind\_WSPD\_30min\_0 | 1 |
| darwinBiochem\_TURB\_1 | 9296.714 | darwinBiochem\_PAR\_0 | 0.02313 | darwinBiochem\_CNDC\_0 | 0.079564 |
| darwinBiochem\_TURB\_0 | 9816.589 | darwinBiochem\_TEMP\_0 | 0.021484 | darwinBiochem\_CNDC\_1 | 0.079325 |
| darwinBiochem\_CNDC\_0 | 10613.59 | darwinBiochem\_TEMP\_1 | 0.020509 | darwinBiochem\_TURB\_0 | 0.047287 |
| darwinBiochem\_CNDC\_1 | 10637.48 | darwinBiochem\_PAR\_1 | 0.014365 | darwinBiochem\_TURB\_1 | 0.044222 |
| darwinBiochem\_PAR\_0 | 10689.19 | darwinBiochem\_CPHL\_1 | 0.014356 | darwinBiochem\_PAR\_0 | 0.040754 |
| darwinBiochem\_CPHL\_1 | 11986.52 | darwinBiochem\_CNDC\_0 | 0.012893 | darwinBiochem\_PAR\_1 | 0.033162 |
| darwinBiochem\_CPHL\_0 | 12002.19 | darwinBiochem\_TURB\_0 | 0.0125 | darwinBiochem\_CPHL\_1 | 0.022211 |
| darwinBiochem\_TEMP\_1 | 12440.17 | darwinBiochem\_CNDC\_1 | 0.010315 | darwinBiochem\_TEMP\_1 | 0.022013 |
| darwinBiochem\_TEMP\_0 | 12445.92 | darwinBiochem\_CPHL\_0 | 0.009205 | darwinBiochem\_CPHL\_0 | 0.021729 |
| darwinBiochem\_PAR\_1 | 13526.34 | darwinBiochem\_TURB\_1 | 0.008021 | darwinBiochem\_TEMP\_0 | 0.021663 |

CNDC and TURB appear to be the best predictors by TIME Warp and Pearon’s r, CCF however suggests PAR and TEMP as the best predictor.

ARIMA

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **AR** |  |  |  | **MA** |  |  |  |
| **No exo** | Error | Order | Standard Deviation | | **No Exo** | Error | Order | Standard Deviation |
| **arsq** | 0.8464925 | 1 | 0.0398429 |  | **arsq** | 0.8464925 | 1 | 0.0398429 |
| **WMAPE** | 0.1793231 | 1 | 0.0220561 |  | **WMAPE** | 0.1793231 | 1 | 0.0220561 |
| **rmse** | 3.4448624 | 1 | 0.089413 |  | **rmse** | 3.4448624 | 1 | 0.089413 |
| **reler** | 1.1238948 | 1 | 0.0379701 |  | **reler** | 1.1238948 | 1 | 0.0379701 |
| **CNDC** | Error | Order | Standard Deviation | | **CNDC** | Error | Order | Standard Deviation |
| **arsq** | 0.8655998 | 10 | 0.0331988 |  | **arsq** | 0.8683633 | 10 | 0.0320189 |
| **WMAPE** | 0.1575427 | 7 | 0.01527 |  | **WMAPE** | 0.15745 | 1 | 0.0148609 |
| **rmse** | 3.2349475 | 10 | 0.100242 |  | **rmse** | 3.2027603 | 10 | 0.1026423 |
| **reler** | 0.9891236 | 2 | 0.0087832 |  | **reler** | 0.988717 | 1 | 0.0066834 |
| **TEMP** | Error | Order | Standard Deviation | | **TEMP** | Error | Order | Standard Deviation |
| **arsq** | 0.8655852 | 10 | 0.0332044 |  | **arsq** | 0.8683413 | 10 | 0.0320311 |
| **WMAPE** | 0.1575875 | 2 | 0.0153179 |  | **WMAPE** | 0.1574962 | 1 | 0.0148885 |
| **rmse** | 3.2350995 | 10 | 0.1001522 |  | **rmse** | 3.2029895 | 10 | 0.1024876 |
| **reler** | 0.9893897 | 2 | 0.0089203 |  | **reler** | 0.9889959 | 1 | 0.0070142 |
| **CPHL** | Error | Order | Standard Deviation | | **CPHL** | Error | Order | Standard Deviation |
| **arsq** | 0.8694637 | 10 | 0.0356082 |  | **arsq** | 0.8722118 | 10 | 0.034385 |
| **WMAPE** | 0.1518348 | 7 | 0.0178875 |  | **WMAPE** | 0.1517496 | 1 | 0.0175685 |
| **rmse** | 3.2427665 | 10 | 0.0925093 |  | **rmse** | 3.2098474 | 10 | 0.0968168 |
| **reler** | 0.9519675 | 7 | 0.040561 |  | **reler** | 0.9516309 | 1 | 0.0419878 |
| **PAR** | Error | Order | Standard Deviation | | **PAR** | Error | Order | Standard Deviation |
| **arsq** | 0.8655632 | 10 | 0.033089 |  | **arsq** | 0.8683449 | 10 | 0.0319653 |
| **WMAPE** | 0.1575404 | 2 | 0.0152885 |  | **WMAPE** | 0.1574515 | 1 | 0.0148495 |
| **rmse** | 3.2356392 | 10 | 0.1003862 |  | **rmse** | 3.2030899 | 10 | 0.1026262 |
| **reler** | 0.9891065 | 2 | 0.0089202 |  | **reler** | 0.988732 | 1 | 0.0069664 |
| **TURB** | Error | Order | Standard Deviation | | **TURB** | Error | Order | Standard Deviation |
| **arsq** | 0.874842 | 10 | 0.0304212 |  | **arsq** | 0.8774451 | 10 | 0.0293609 |
| **WMAPE** | 0.1480535 | 2 | 0.0280527 |  | **WMAPE** | 0.1479837 | 1 | 0.0276252 |
| **rmse** | 3.2535554 | 10 | 0.1118162 |  | **rmse** | 3.2205068 | 10 | 0.1122381 |
| **reler** | 0.9263472 | 2 | 0.1288874 |  | **reler** | 0.9260848 | 1 | 0.1269778 |

TURB can be seen to be the best predictior of WINDSPD, however none of the methods appear to make a significant improvement of the benchmark method and have high overall error.

MLPR

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **CNDC+TURB+PAR+CPHL+TEMP** | Error | Standard deviation | | **TURB** | Error | Standard deviation |
| arsq | -0.132 | 0.08247 |  | arsq | -0.0791 | 0.09612 |
| WMAPE | 0.59109 | 0.19718 |  | WMAPE | 0.58367 | 0.20153 |
| rmse | 9.67099 | 1.56845 |  | rmse | 9.43174 | 1.62542 |
|  |  |  |  |  |  |  |
| **CPHL** | Error | Standard deviation | | **PAR** | Error | Standard deviation |
| arsq | -0.0878 | 0.07538 |  | arsq | -0.032 | 0.02489 |
| WMAPE | 0.58164 | 0.19417 |  | WMAPE | 0.56074 | 0.17652 |
| rmse | 9.40898 | 1.57128 |  | rmse | 9.12852 | 1.35521 |
|  |  |  |  |  |  |  |
| **CNDC** | Error | Standard deviation | | **TEMP** | Error | Standard deviation |
| arsq | -0.0568 | 0.03228 |  | arsq | -0.0423 | 0.03816 |
| WMAPE | 0.56723 | 0.17767 |  | WMAPE | 0.56521 | 0.18639 |
| rmse | 9.229366 | 1.31252 |  | rmse | 9.18625 | 1.48626 |

We can see significantly higher error in the MLPR results compared to ARIMA. Surprisingly using multiple variables did not lower the error for WINDSPD unlike the other target variables tested.PAR, TEMP and CNDC gave the lowest errors with all having similar amounts of error however the error appears too high to be useful.

GLM

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **CNDC+TURB+PAR+CPHL+TEMP** | Error | Standard deviation | | **TURB** | Error | Standard deviation |
| arsq | 0.00892 | 0.10727 |  | arsq | -0.0369 | 0.13886 |
| WMAPE | 0.53054 | 0.20349 |  | WMAPE | 0.55309 | 0.21418 |
| rmse | 9.4518 | 1.42121 |  | rmse | 9.49629 | 1.56214 |
|  |  |  |  |  |  |  |
| **CPHL** | Error | Standard deviation | | **PAR** | Error | Standard deviation |
| arsq | -0.0107 | 0.03746 |  | arsq | -0.0341 | 0.02995 |
| WMAPE | 0.54628 | 0.1791 |  | WMAPE | 0.5623 | 0.18219 |
| rmse | 9.2319 | 1.44015 |  | rmse | 9.16578 | 1.41901 |
|  |  |  |  |  |  |  |
| **CNDC** | Error | Standard deviation | | **TEMP** | Error | Standard deviation |
| arsq | -0.0361 | 0.02082 |  | arsq | -0.0331 | 0.02758 |
| WMAPE | 0.5626 | 0.18045 |  | WMAPE | 0.56183 | 0.18048 |
| rmse | 9.16892 | 1.36648 |  | rmse | 9.16076 | 1.41128 |

The GLM produces similar results to the MLPR method which once again appear to have too much error to be useful.

# Discussion

Variable selection

The variable selections appeared to have similar accuracy, while CCF and Pearson’s r always predicted the best variable, Timewarp sometimes failed to pick the best variable. 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. Timewarp also took the longest to compute compared to the other methods. Pearson’s r coefficient overall appeared to produce the best results in ranking the prediction of exogenous variables.

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. 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, these models can be useful on longer periods of missing data.

CNDC and TEMP could be effectively predicted by all methods although ARIMA provided much better prediction. The MLP and GLM methods although producing higher error still produced usable results and decent predictions as the error for the models was still low.

WWSH and WINDSPD produced poor results from all methods. The MLP and GLM models for WWSH and WINDSPD failed to predict or model these target timeseries having WMAPE higher than 50% in most scenarios. The ARMA results were significantly better with WMAPE between 15~25% using most variables however looking at the relative error this was often not significantly better than the benchmark method for WINDSPD and only slightly better for WWSH.

Despite having an overall large error for prediction, MLP models provided a surprisingly good result. The model produced had about twice the error of the ARMA models which given there is no time dependence and can be used to replace larger sections of missing data is an acceptable trade off.

Future work

ARIMA models using both AR and MA

In this report due time restrictions only AR and MA models were tested not ARIMA models as the computational time would be too large for the scope of this project

Further lags of ARIMA tested

Most models were only tested up to 10 lags due to time constraints

Long term prediction

Only 1 step ahead prediction was tested for ARIMA models. For the purpose of prediction and replacing missing data longer period predictions could be useful and should be tested.

Models using previous lags of exogenous variables

Exogenous variables were only considered between the current lag and data points between the previous lag for higher frequency data. Previous lags of exogenous variables could also be used In ARIMA models.

Multiple exogenous timeseries in ARIMA

In ARIMA models only singular exogenous variables were used per model. Models could be tested incorporating multiple exogenous variables.

Different methods of dealing with mixed frequency data.

Only one method of managing mixed frequency data was implemented and other may be tested for example taking the mean of multiple data points when several occur between lower frequency data or re-sampling if the target variable is of higher frequency.

Non Linear models

Other types of models such as non linear and machine learning models need to be evaluated for time series forecasting.

References

Stats Package R

Used for ccf and diff

R Core Team (2017). R: A language and environment for statistical

computing. R Foundation for Statistical Computing, Vienna, Austria. URL

https://www.R-project.org/.

Forecast package R

Used for ARIMA.

Hyndman R, Bergmeir C, Caceres G, Chhay L, O'Hara-Wild M, Petropoulos F,

Razbash S, Wang E and Yasmeen F (2018). \_forecast: Forecasting functions for

time series and linear models\_. R package version 8.3, <URL:

http://pkg.robjhyndman.com/forecast>.

Hyndman RJ and Khandakar Y (2008). “Automatic time series forecasting: the

forecast package for R.” \_Journal of Statistical Software\_, \*26\*(3), pp.

1-22. <URL: http://www.jstatsoft.org/article/view/v027i03>.

Fast Dynamic Time Warp

(2007) Dynamic Time Warping. In: Information Retrieval for Music and Motion. Springer, Berlin, Heidelberg

Stan Salvador, and Philip Chan. “FastDTW: Toward accurate dynamic time warping in linear time and space.” Intelligent Data Analysis 11.5 (2007): 561-580.

Adjusted R squared

Miles, J. (2014). R Squared, Adjusted R Squared. In Wiley StatsRef: Statistics Reference Online (eds N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri and J.L. Teugels). doi:[10.1002/9781118445112.stat06627](https://doi.org/10.1002/9781118445112.stat06627)

RMSE

Barnston, A., (1992). “Correspondence among the Correlation [root mean square error] and Heidke Verification Measures; Refinement of the Heidke Score.”