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.

# Method

Differing frequencies

As it is timeseries being analysed all data must be in the same frequency to be used.

This was achieved by separating higher frequency variables into several variables of the same frequency as the target lower frequency variable.

Data of a lower frequency was discarded to avoid manipulating data by either discarding or resampling existing data

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 and determines the similarity of two timeseries. It was implemented using fast dtw in python.

Cross Correlation function (CCF) is the second method used to determine suitable variables. Time series must be stationary before being compared by CCF.

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 which is then un-differenced. This is then compared to the original values of the test set and error measurers are produced.

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.

Error measures

Root mean square error (RMSE)

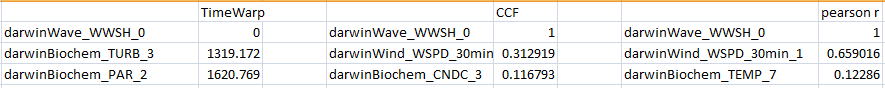
Weight Mean Absolute Percentage Error (WMAPE)

Relative Error

# Results

### WWSH

**Model Selection**

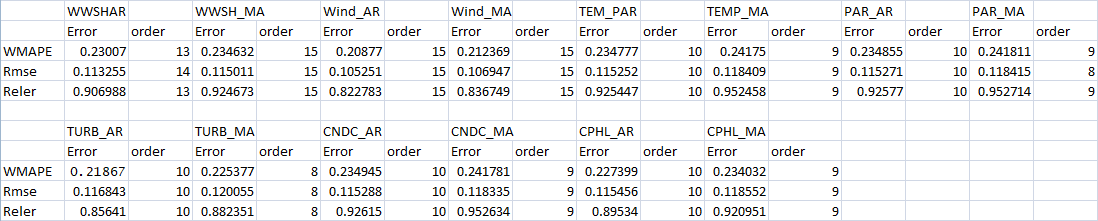


Each method appears to nominate differing variables for use as exogenous variables however CCF and Pearson r give more similar results both highlighting WSPD\_30min as the best predictor however CCF suggests CNDC will be the next best predictor while Pearson r suggests TEMP.

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

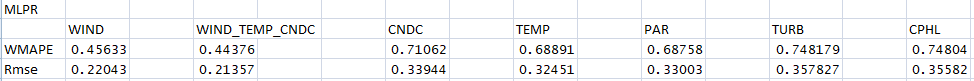
Results

ARIMA



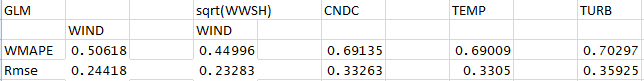
Looking at each of the results its clear WindSpd is the best exogenous variable for WWSH prediction as predicted by CCF and Pearson’s r. TURB was the next best predictor and provided some improvement but not as significant as WindSpd. CNDC and TEMP provided no improvement over an AR or MA model with 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



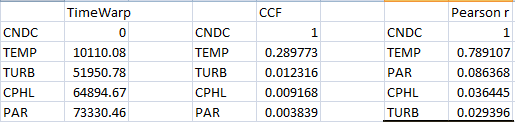
A Generalised Linear model provides a similar result to MLPR however with slightly higher errors.

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

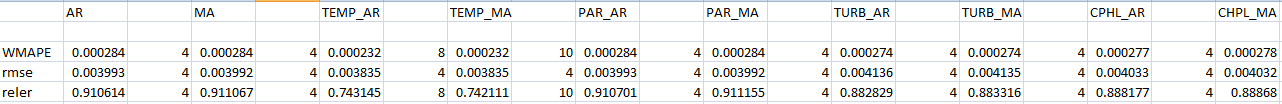
Model Selection



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.

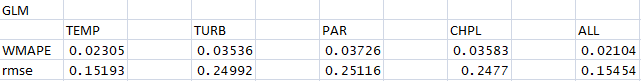
Results

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.

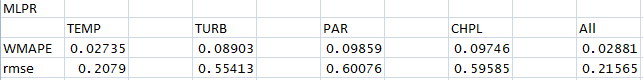
GLM



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



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.