Chapter 3

Material and Methods

**3.1 Software:**

Python and R have been used as the software languages.

We did our data pre-processing step on Python through Jupyter notebook. The reason is based on the convenience of Python modules. It can easily display data visualizations and impute missing values.

R will focus more on modelling and diagnosis as R can select and assess the model by using different libraries.

**3.2 Description of the Data:**

The given raw dataset is in one CSV file with 1.16 GB storage. There are 3,123,117 observations with 57 variables, including the dependent variable CPHL.

The dataset records values on seven distinct gliders distributed across Australia from 2013 to 2015. However, a large number of data at around 19% is missing that spreads unevenly across different variables. Take one feature, VCUR, which is the value of seawater velocity at northward as an example, a significant amount at 97% overall is missing.

**3.3 Pre-processing Steps:**

1. Import libraries and explore the raw data

First of all, we visualize the raw data by plotting different types of diagrams, like missing value tables, dendrogram, correlation diagram, histogram, distribution mapping and time-series. The above methods give insights into the distribution on each glider and feature. Then we filter data with the valid range to generate the legal dataset according to the standard variable instance table (AODN 2016)【0】. For individual value, whose missing percentage is significantly high at more than 90% like UCUR, ignoring the entire feature is a wise choice after comparing the variable’s representation and diagnose the missing data mechanism. Nevertheless, not all features can be deleted as straightforward. Since the raw dataset is enormous and contains lots of missing value that distributed unequally by gliders, splitting the glider allows us to have a better understanding of the missing value for the further imputation process.

2. Analyze features by features on each glider

After the split of raw data into seven subsets according to seven different gliders, we start to analyze every single feature one by one. Depending on the corresponding quality control types defined by the raw table which are eight categories in total, we firstly kept all the value under the good data type and then calculated the percentage of the remaining categories, such as bad or missing data. The data would either be dropped or imputed according to the importance and the missing percentage.

3. Imputation missing value

For the missing data regarding the feature that is expected to be imputed, we plot the time-series for the variable against the remaining features, finding the possible correlation and diagnosing the missing mechanisms. After compared different imputations methods, including KNN, mean and median, KNN has been selected to fill in the table as it gives the highest accuracy on the training model.

4. Explore data after imputation and compared the differences between each glider

Previous diagrams in step 1 have been plotted again to make sure there is no missing value for each variable. Further, the filled data on separate glider has been analyzed to compare the difference. Finally, we generate CSV files for the further modelling step.

**3.4 Data Cleaning:**

For missing data, we either delete them or impute them based on the percentage and the meaning of the missing data.

It is always better to keep data than to discard it, but if the missing data is limited to a small number of observations (less than 5%), we may delete those cases from the analysis.

If the percentage of missing data is more significant than 5%, but less than 40% and the data is neither missing entirely at random nor missing not at random, we should consider filling the missing data by imputation.

If the data is missing for more than 40% of observations, we should consider deleting the missing values.

If the data is missing for more than 90% of observations, this feature could be considered to delete.

For the method of missing value imputation, after comparing the strength and shortage of mean imputation, the most frequent value imputation, and KNN imputation, we decided to implement KNN imputation. In KNN imputation, the k nearest neighbors algorithm can be used for imputing missing data by finding the k closest neighbors to the observation with missing data and then imputing them based on the non-missing values in the neighbors. The reason for that is because we have high dimensional data, and KNN imputation is specifically designed for this situation, and it could also interpret this situation the most. Besides that, KNN is also the most widely used imputation method.

**3.5 Assumption:**

Assume the independence between each observation.

Assume the change of device has no impact on the measurement.

Assume there is no measurement bias between different gliders.

Assume there is no measurement bias due to the time change for each glider.

3.6 Modelling methods

After having a brief overview of our datasets, we would basically prepare to apply some modeling methods on these datasets.

3.6.1 Time-series Model method

The first method we would apply to our datasets is the time-series model according to Ronchi [1], since gliders collected data among time not just a time point and they might influence each other among time. However, the time-series model is suitable when variables are changing periodically according to Genchiro [2]. Therefore, we should reject this model if variables have little periodical features or the time range is too short to analyze.

3.6.2 Generalized Linear Model method (GLM)

Generalized Linear Model is very useful by building a relationship between the predictors and the response via some link functions. However, the Generalized Linear Model can only be applied when the response belongs to the exponential family distribution and the link function describes the relationship between the response and other linear combinations of predictors according to Julian.J [3]. Hence, for trying this method, we would plot the distribution of the response CPHL firstly to check whether this dataset could be applied to the Generalized Linear Model method. If possible, we will compare, fit, approve this model by analyzing residuals vs fitted diagrams, Q-Q plot, R^2, p-values for each predictor, and ANOVA table, then finally choose the best fit one.

3.6.3 Variance Inflation Factor method (VIF)

Catalina B states that the Variance Inflation Factor (VIF) is a useful tool to detect whether there is multicollinearity in our models [4]. Multicollinearity exists when there are at least two predictors that have a strong correlation with each other which will reduce the precision of the estimate coefficients and weaken the model’s accuracy. The existence of multicollinearity is troubling when VIF is greater than 10. Hence, we need to plot the correlation plot to have a brief understanding of correlation among variables and check the VIF when fitting our model, ensuring VIF values of all predictors in our final model are less than 10.

Reference:

1. <http://thredds.aodn.org.au/thredds/dodsC/IMOS/ANFOG/slocum_glider/Forster20170911/IMOS_ANFOG_BCEOPSTUV_20170911T071056Z_SL287_FV01_timeseries_END-20171002T010328Z.nc.html>

1.     <https://primoa.library.unsw.edu.au/primo-explore/fulldisplay?vid=UNSWS&docid=TN_cdi_scopus_primary_608024870&context=PC&query=any,contains,Basic%20Concepts%20and%20Modelling%20Methods&_ga=2.229935832.538925737.1604536476-972327823.1604536476>

2.     book: introduction to time series modelling with application in R

3.     <https://ebookcentral.proquest.com/lib/unsw/reader.action?docID=4711494>

4.     <https://arxiv.org/pdf/2005.02245.pdf>

**Reference:**

<https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-017-0442-1>