

Part 1 Core: Evidence related to fish stocks in New Zealand

1 Select a fish data set and use a pipeline

The first part of this experiment primarily leverages the public data containing information of "fish" provided by the given website (<http://www.data.govt.nz/>). Import the data to provide data manipulation, visualization, pipeline modelling functions, and model output functions.

Data set 1:

Fish monetary stock account, 1996–2018

<From: Environmental-economic accounts: 2019 – tables >

Two techniques:

Clustering (K-means) and Linear Regression

***Business understanding**

Experts predict that most of the world's fishery resources are in a state of overload and are on the verge of collapse, with a sharp decline in the number of marine fish resources. As New Zealand's export industry, fisheries have contributed to the domestic economy. However, the impact of large-scale fishing on fish resources cannot be ignored. Through the analysis and evaluation of fish data, the government fishery management department can formulate protection and restriction measures according to the status and trend of fish resources, make rational use of fishery resources, and achieve long-term sustainable development.

***Data understanding**

This data set contains a large amount of fish characteristic information, and describes value and quantity of the fish stocks in commerce.

We just choose part of this data set to illustrate the issue only by treating all fish species as one object to analyse and evaluate different value. The data set contains the 6 attributes and 23 instances, including TACC (Total allowable commercial catch), Asset value of the commercial fish resource, Seafood exports value and quantity, Total catch and year (1996-2018). All attributes are numeric-valued. The catch and fishing effort time series are used by managers to safeguard the availability of resources in the future. Though the instances in the fish data set is bit limited, experiment aims to observe and predict fish stocks status and economic growth based on these indicators.

***Data preparation**

At this stage, we extract the data into usable formats (including csv, arff etc.), check and deal with the data in order to obtain quality data, including ignoring the non-contributory data and just using the total metric. There are 20 missing values in this

data set. Replace them with the mean. Apply Normalize in attributes to reduce the higher mean error due to different measuring unit (millions, dollars & tonnes).

※Modelling

Select K-Means and Linear Regression and compare these two very different machine learning models on the fish data set for predicting the status of fish stocks Catch. Regression is suitable to predict a continuous variable based on learning from a known data set, whereas the experiment is to analyse the correlation degree between fishing quantity and other attributes, and the principle of linear regression is to judge the relationship between variables and predict dependent variables with independent variables. The clustering algorithm calculates the similar density of samples in the fish data, which would analyse the correlation degree between fishing quantity and other attributes. The data sets are all numerical values, meeting the requirements of the two methods.

※Evaluation

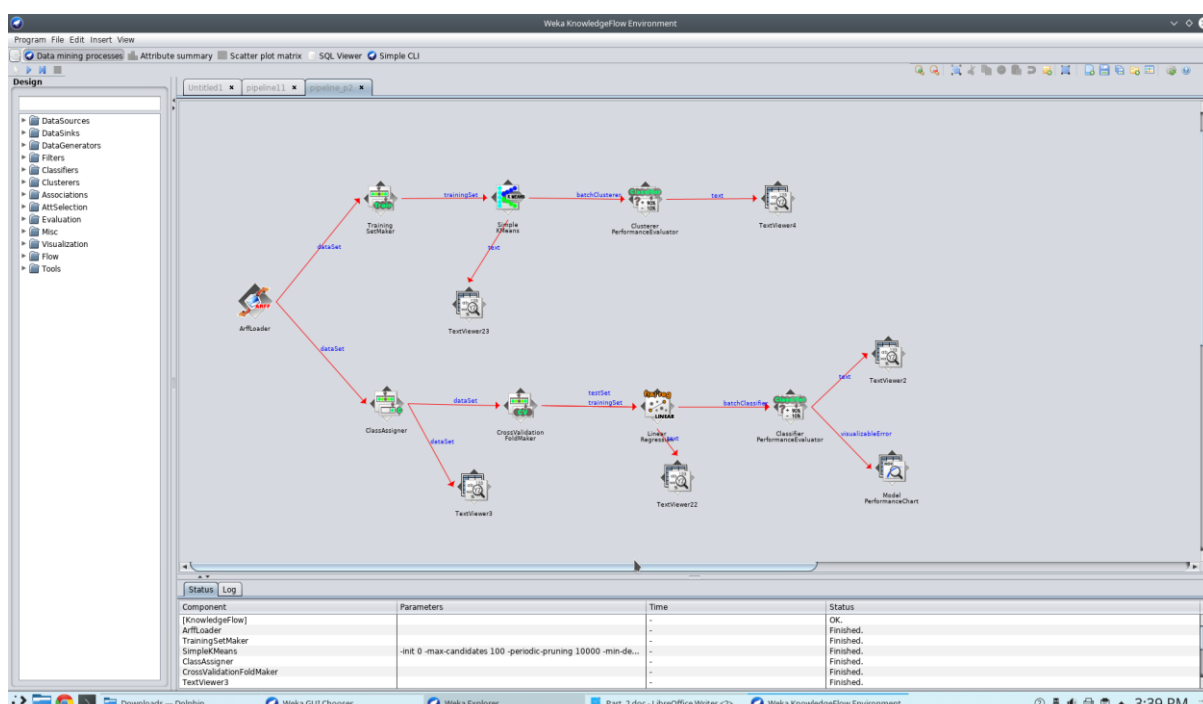
This pipeline supports two Algorithm to evaluate a solution. Check the accuracy and measure the availability of the model. Then, we would analyse the results and conduct an overall comparison and adjust those algorithms that perform well in order to determine the best tool for the clustering or prediction purposes

※Deployment

The accuracy of a method on a given test set is used to confirm whether the algorithm can be considered acceptable and clustering predict future data for which the label or trend is not known. We also need to additional effort to optimize the model, such as change parameter, etc.

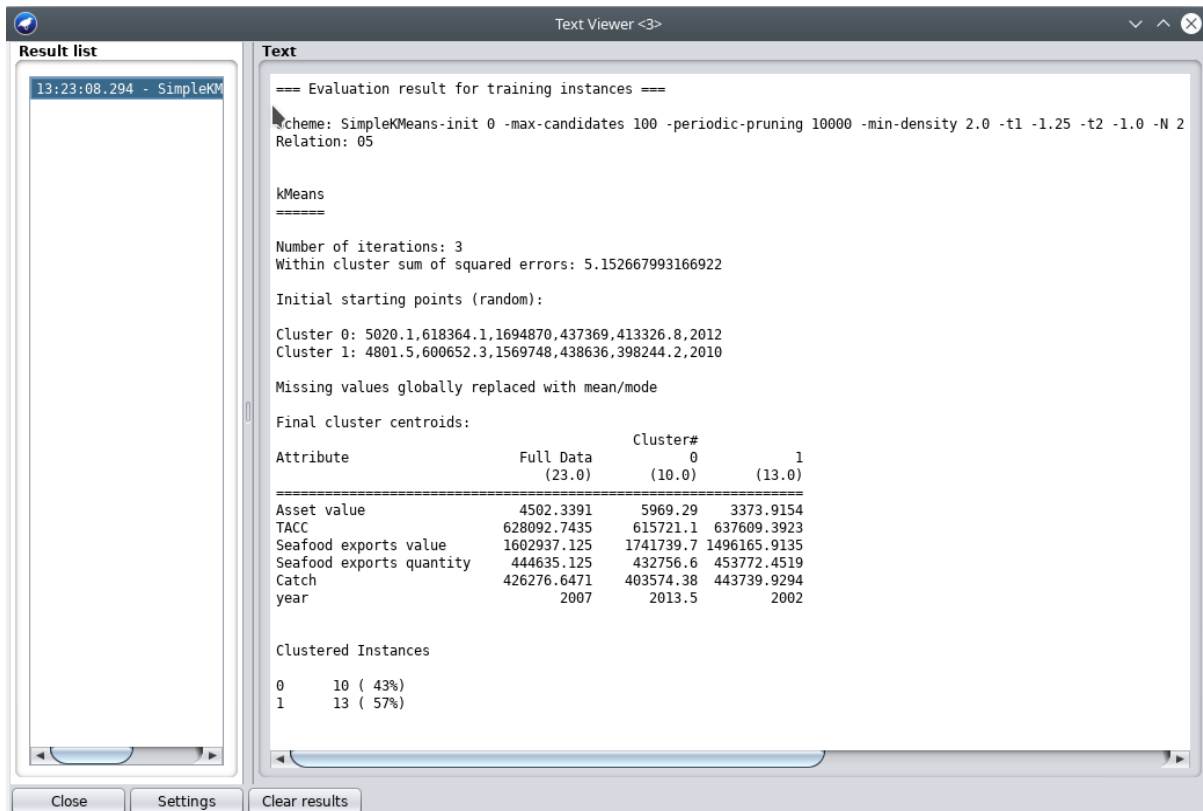
---Create a pipeline using K-means and Linear Regression

---Load fish data set, run the information and the result as follows:



2 Describe results of two techniques

2.1 Clustering (K-means)



The screenshot shows a 'Text Viewer' window with the following content:

```
=== Evaluation result for training instances ===
Scheme: SimpleKMeans-init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2
Relation: 05

kMeans
=====

Number of iterations: 3
Within cluster sum of squared errors: 5.152667993166922

Initial starting points (random):
Cluster 0: 5020.1,618364.1,1694870,437369,413326.8,2012
Cluster 1: 4801.5,600652.3,1569748,438636,398244.2,2010

Missing values globally replaced with mean/mode

Final cluster centroids:
Attribute          Full Data          Cluster#
                   (23.0)             0             1
                   (10.0)          (13.0)
-----
Asset value        4502.3391          5969.29       3373.9154
TACC               628092.7435        615721.1      637609.3923
Seafood exports value 1602937.125        1741739.7     1496165.9135
Seafood exports quantity 444635.125         432756.6     453772.4519
Catch              426276.6471        403574.38     443739.9294
year               2007               2013.5        2002

Clustered Instances
0      10 ( 43%)
1      13 ( 57%)
```

At the bottom of the window are three buttons: 'Close', 'Settings', and 'Clear results'.

The K-means algorithm first randomly assign K cluster centres, then:

---Assigning each instance to the cluster centre nearest to it to obtain K clusters.

---It calculates the mean value of all instances in each cluster separately and take them as the centre of each cluster.

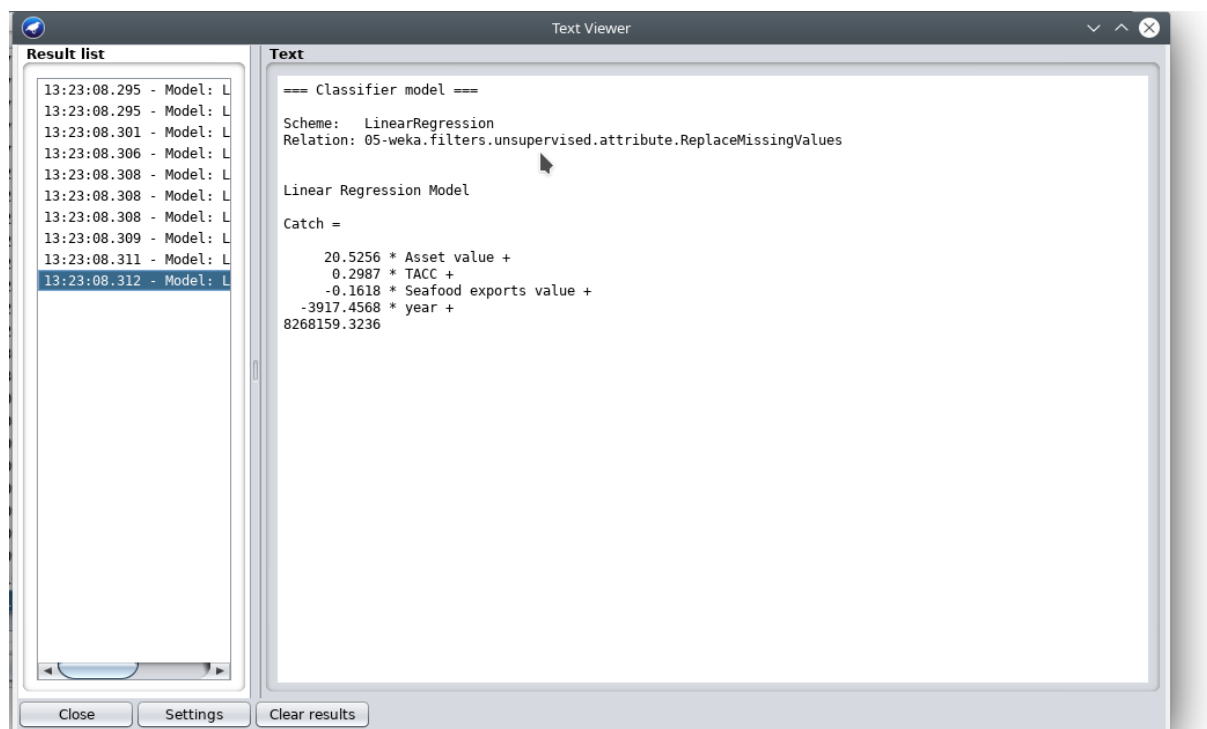
---Repeat the second and third steps until the centres of the K clusters are fixed and the distribution of clusters is fixed.

The above result shows that 23 instances are divided into 2 clusters, containing 10 and 13 instances respectively. The result is also shown as percentage.

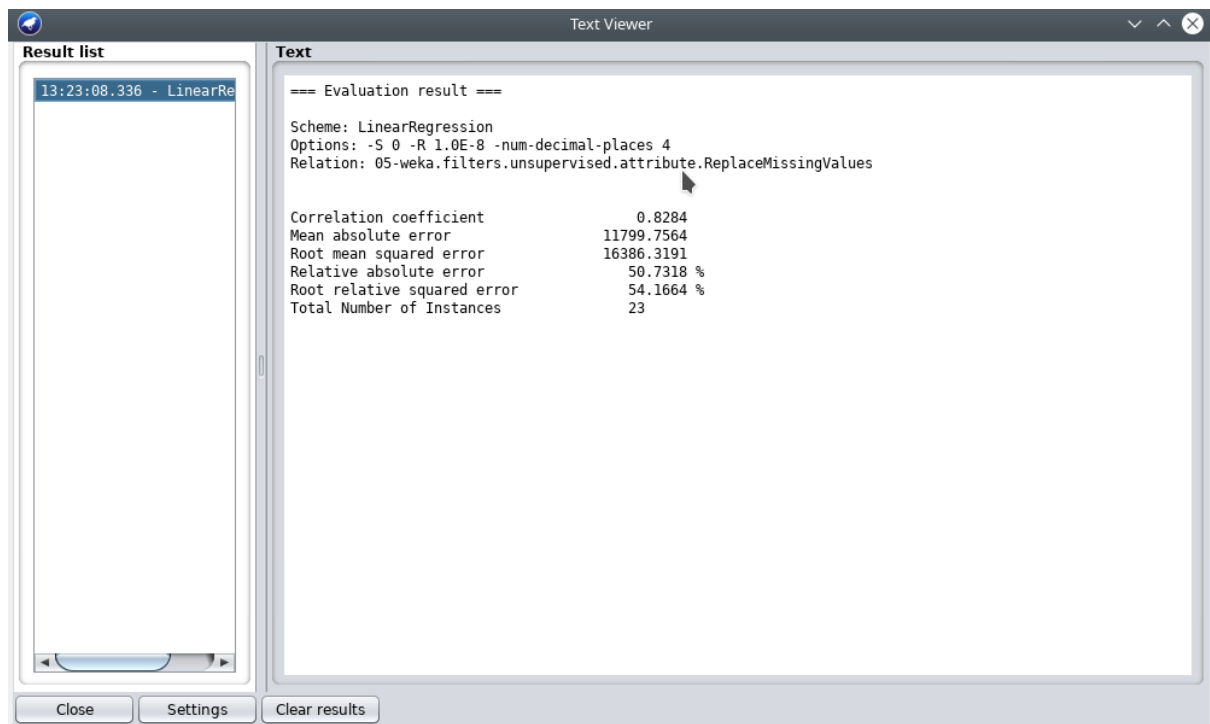
“With cluster sum of squared errors” is the standard to judge whether clustering is well or poor. The smaller the value, the smaller the distance between instances of the same cluster.

After cluster, the location of each cluster centre is listed, for numerical results, cluster centre is its mean value.

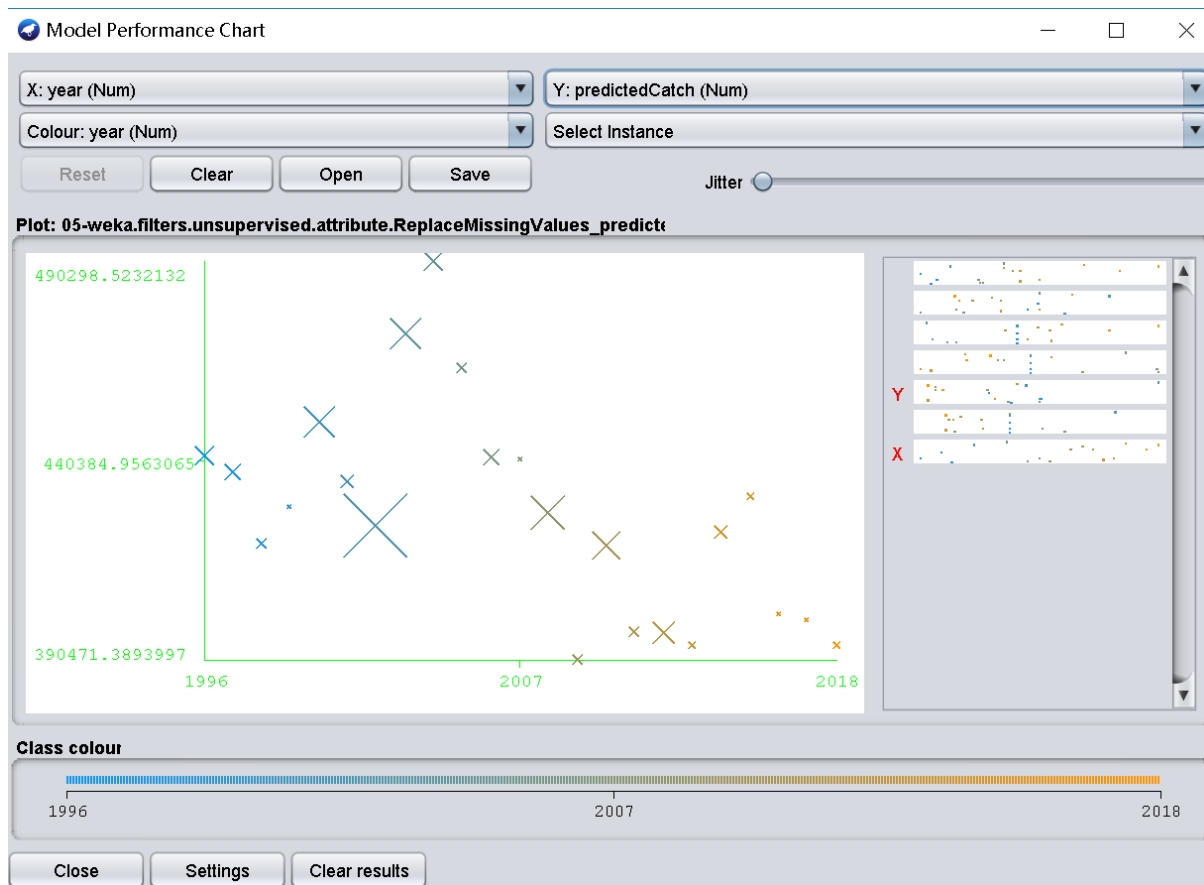
2.2 Linear Regression



From the results of the model, we can see that catch is mainly related to Asset value, and has little to do with seafood exports quantity. Also, Catch is positively corrected with Asset value and TACC, and negatively correlated with seafood exports quantity and year. The “-” in front of the number indicates that catch decrease as year increases.



The evaluation results of the model have the above indexes, and the accuracy is expressed by correlation coefficient, 0.8284. The model has good performance. Also, it can be seen that the mean squared error deviation is small whereas relative absolute error is large.



In the regression model performance chart, the target predicted value is displayed (predicted Catch). X represents the error between the actual value and the predicted value of each instance. The larger the size of X, the greater the error.

3 Identify the aspects of two techniques

	Clustering K-means	Regression Linear Regression
Machine learning	unsupervised learning	supervised learning
Use for	categorization and impact of number of attributes	evaluate the trends, making estimates, and forecast
Result	K-cluster (probability)	mathematical notations

Regression method is a supervised learning algorithm for predicting and modelling numerical continuous random variables. The output result is that the mathematical formula expresses the relationship between the dependent variable catch and other independent variables. As Catch is related to assets and other factors, the linear regression result shows a good fit.

Clustering does not know the attribute range of all kinds of fish samples in advance. It is an unsupervised learning (i.e. the data are not labelled). It can only analyse the attributes of the samples based on the distribution of the samples in the feature space, and the output results show the groups surrounding the clustering centre. The algorithm is fast and simple enough, but the number of clusters needs to be specified. The selection of K value is not easy to determine, and some relatively poor clusters will be obtained.

Judging from the accuracy of the results, the regression performed better. However, each missing value removes one data point that could optimize the regression. In simple linear regression, outlier can significantly disrupt the outcomes. Each cluster is presented by a representative feature. Different sets of features may produce different clustering.

4 revisit the business understanding

The other two questions could be asked of the fish data:

Question 2: What are the impacts of fishing on the ecological environment?

Question 3: According to the current fish catch, how to predict the future trend of fish resources (e.g. TACC, the quantity or value)?

※Business understanding (Question 2)

Experts predict that most of the world's fishery resources are in a state of overload and are on the verge of collapse, with a sharp decline in the number of marine fish resources. Overfishing will make the fishing effort exceed a reasonable level, resulting in increased environmental pressure. Studying the long-term sequence effect of fishing action on environment can understand the trend and changes, and judge the correlation between fish catch and other factors, which can provide reliable indicator for fishery manager to plan and utilize rationally and maintain sustainable development of ecology and fishery.

※Business understanding(Question 3)

As New Zealand's export industry, fisheries have contributed to the domestic economy. However, the impact of large-scale fishing on fish resources cannot be ignored. Through the analysis and evaluation of fish data, the government fishery management department can formulate protection, judge the correlation between fish catch and other factors and restriction measures, such as commercial allowable fishing restrictions. According to the status and trend of fish resources, they would make rational use of fishery resources, and achieve long-term sustainable development.

Part 2 Completion: Feature importance to Fish stocks in New Zealand

1 Business understanding with questions and datasets

Question 3: According to the current fish catch, how to predict the future trend of fish resources (e.g. TACC, the quantity or value)?

Data set 2:

Landings from stocks meeting or exceeding performance thresholds (2009–14)

Data set 3:

Fishing effort (number of trawl tows) by year (1990–2014)

Data set 2 description: Our fish stocks are affected by commercial, customary, and recreational fishing, and environmental pressures (e.g. ocean temperature, acidity, and productivity). The Ministry for Primary Industries uses three performance measures to assess influences on fish stocks: a soft limit (below which a rebuilding plan is required), a hard limit (below which closing a fishery should be considered), and an overfishing threshold (where the rate of extraction is higher than the rate of replenishment). This dataset relates to the "State of fish stocks" measure on the Environmental Indicators.

Data set 3 description: Seabed trawling is the practice of towing fishing nets near or along the ocean floor. The towing process can physically damage seabed (benthic) habitats and species.

Commercial fishers use CELR, TCER, and TCEPR forms to record trawl fishing.

CELR – catch effort landing return was used by vessels smaller than 28m .

TCER – trawl catch effort return replaced the CELR form in 2008. vessels (6–28m long)

TCEPR – trawl catch effort processing return is used primarily by vessels longer than 28m operating in waters deeper than 200m.

These two data sets contain important indicators (restrictions and trawls) related to fish catch, which are of great significance for analysing whether it is considered overfishing and the degree of possible serious consequences, and also provide the basis for taking corresponding protection and restriction measures.

2 Merge the datasets using techniques

Two additional data sets are selected get more features to support the finding in this task.

Data set 2:

Landings from stocks meeting or exceeding performance thresholds (2009–14).csv

Data set 3:

Fishing effort (number of trawl tows) by year (1990–2014).csv

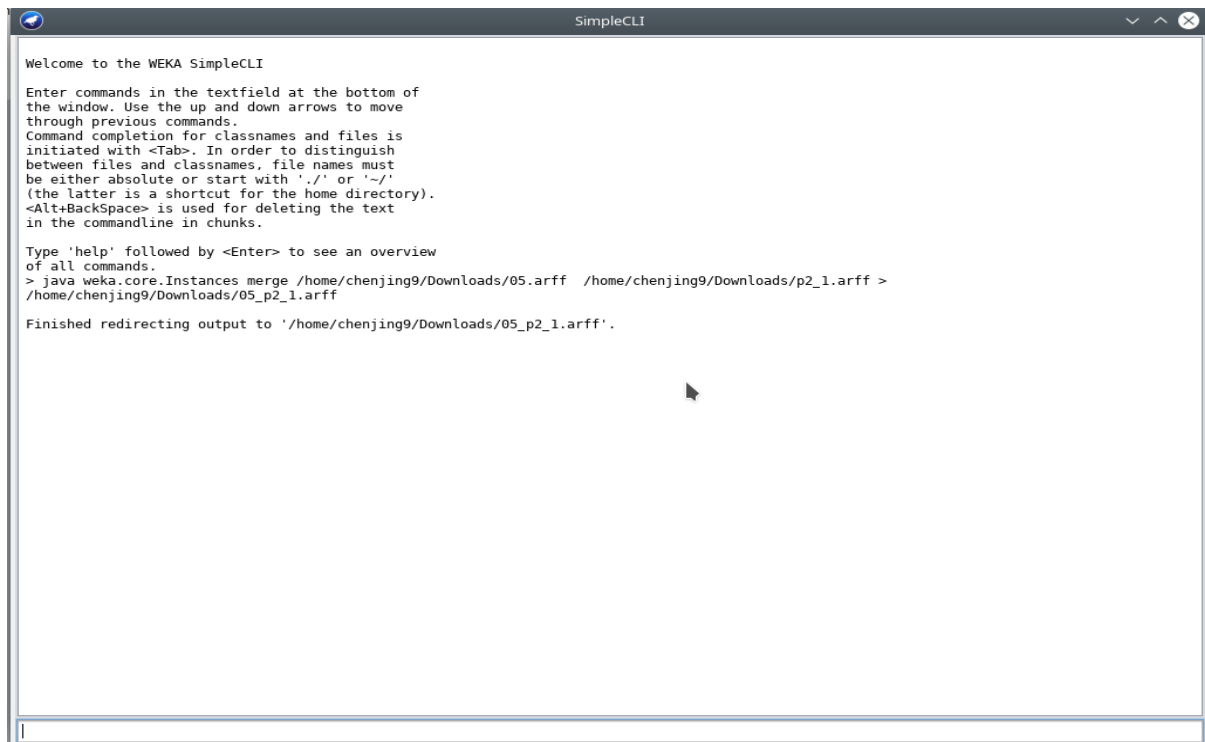
---format three data set into .arff file

---use command-line in weka SimpleCLI as follows, and then merge Data set 1 (05.arff)and Data set 2(p2_1.arff) together to generate the new file(05_p2_1.arff.)

---use join to merge the new file and Data set 3 together in the pipeline.

Use command-line:

```
java weka.core.Instances merge /home/chenjing9/Downloads/05.arff  
/home/chenjing9/Downloads/p2_1.arff > /home/chenjing9/Downloads/05_p2_1.arff
```



Result list

15:50:45.337 - SimpleKM

Text

Initial starting points (random):

Cluster 0: 5020.1,618364.1,1694870,437369,413326.8,2012,247400,8600,275600,1400,228900,9700,2012

Cluster 1: 4801.5,600652.3,1569748,438636,398244.2,2010,228300,12600,251200,2400,162700,20500,2010

Missing values globally replaced with mean/mode

Final cluster centroids:

Attribute	Full Data (23.0)	Cluster# 0 (20.0)	1 (3.0)
Asset value	4502.3391	4501.975	4504.7667
TACC	628092.7435	633019.86	595245.3
Seafood exports value	1602937.125	1599812.6438	1623767
Seafood exports quantity	444635.125	445948.8937	435876.6667
Catch	426276.6471	430876.9891	395607.7
year	2007	2006.55	2010
From_stocks_above_soft_limit	242283.3333	244375.8333	228333.3333
From_stocks_below_soft_limit	11283.3333	11035.8333	12933.3333
From_stocks_above_hard_limit	265816.6667	268124.1667	250433.3333
From_stocks_below_hard_limit	2566.6667	2386.6667	3766.6667
From_stocks_without_overfishing	210433.3333	214758.3333	181600
From_stocks_with_overfishing	14433.3333	13853.3333	18300
Fishing_year	2007	2006.55	2010

Clustered Instances

0 20 (87%)

1 3 (13%)

Close

Settings

Clear results

The two data sets are merged into a new file according to the common key_field of the year. The first data set has 6 attributes and 23 instances (1996-2018). The second data table has only 7 attributes and 6 instances. The missing values of the year (2009-2014) are processed by "Replace missing value" to obtain 23 instances. The new file obtained by merging the two tables horizontally contains 14 attributes and 23 instances.

Result list

15:50:45.339 - Model: L
15:50:45.348 - Model: L
15:50:45.348 - Model: L
15:50:45.348 - Model: L
15:50:45.352 - Model: L
15:50:45.354 - Model: L
15:50:45.354 - Model: L
15:50:45.354 - Model: L
15:50:45.354 - Model: L
15:50:45.362 - Model: L
15:50:45.376 - Model: L

Text

```

=== Classifier model ===

Scheme:  LinearRegression
Relation: 05_p2_1

Linear Regression Model

Catch =
    17.4063 * Asset value
    0.3589 * TACC +
    -0.0548 * Seafood exports value +
    -4049.591 * year +
    -4049.5911 * Fishing_year +
    16483962.4841

```

Close

Settings

Clear results

15:50:45.380 - LinearRe

Close

Settings

Clear results

Text Viewer

=== Evaluation result ===

Scheme: LinearRegression

Options: -S 0 -R 1.0E-8 -num-decimal-places 4

Relation: 05_p2_1

Correlation coefficient

Mean absolute error

Root mean squared error

Relative absolute error

Root relative squared error

Total Number of Instances

Ignored Class Unknown Instances

0.8615

15162.6414

17531.6366

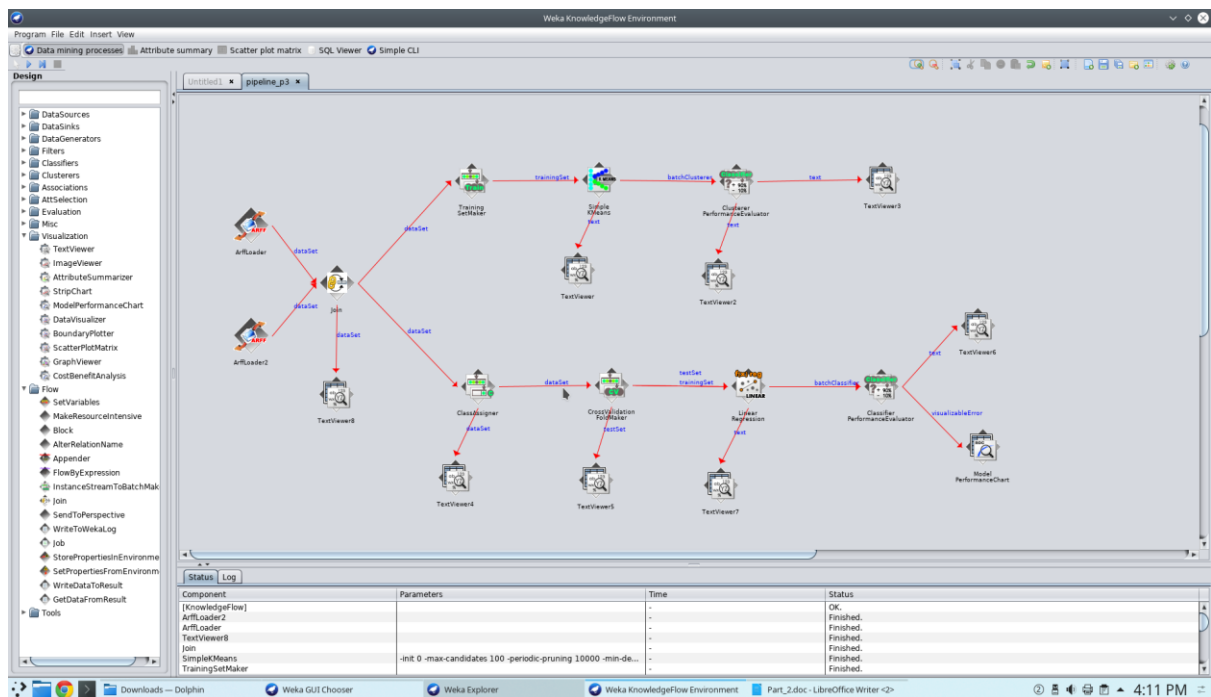
48.295 %

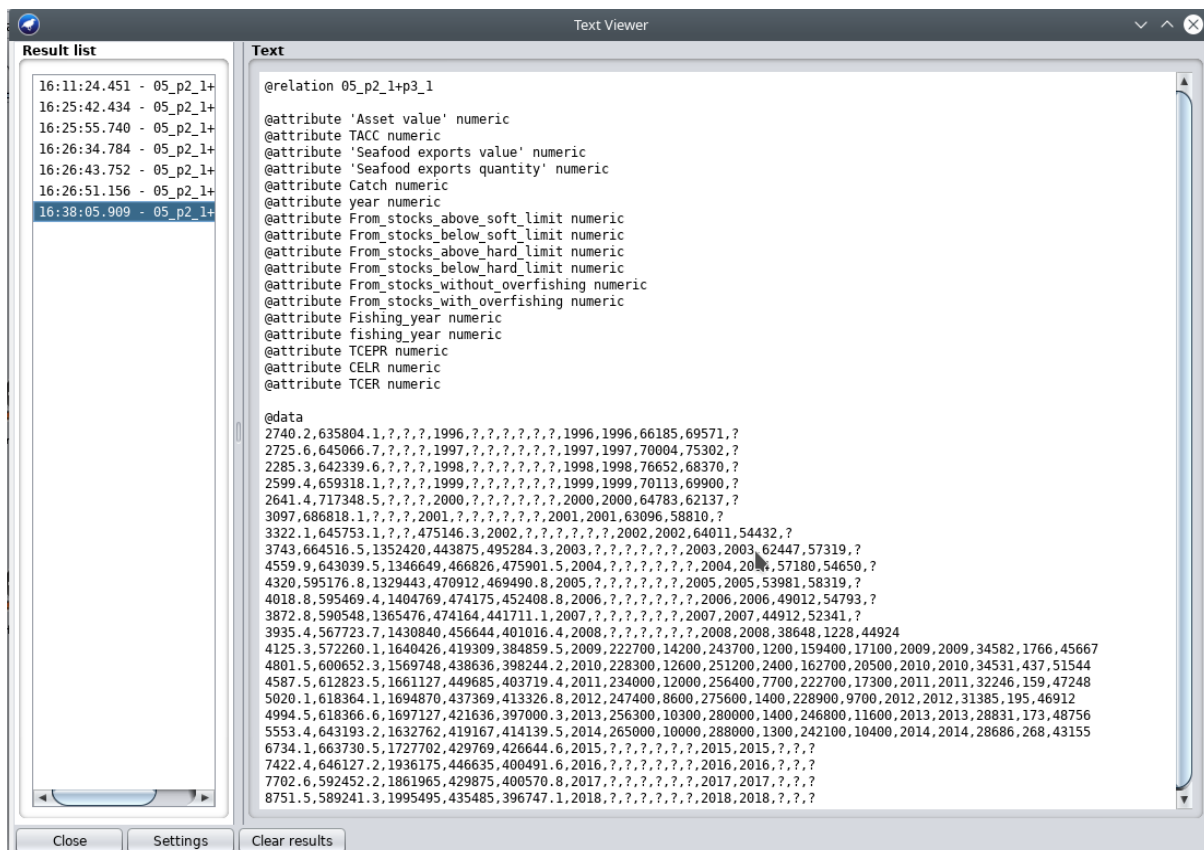
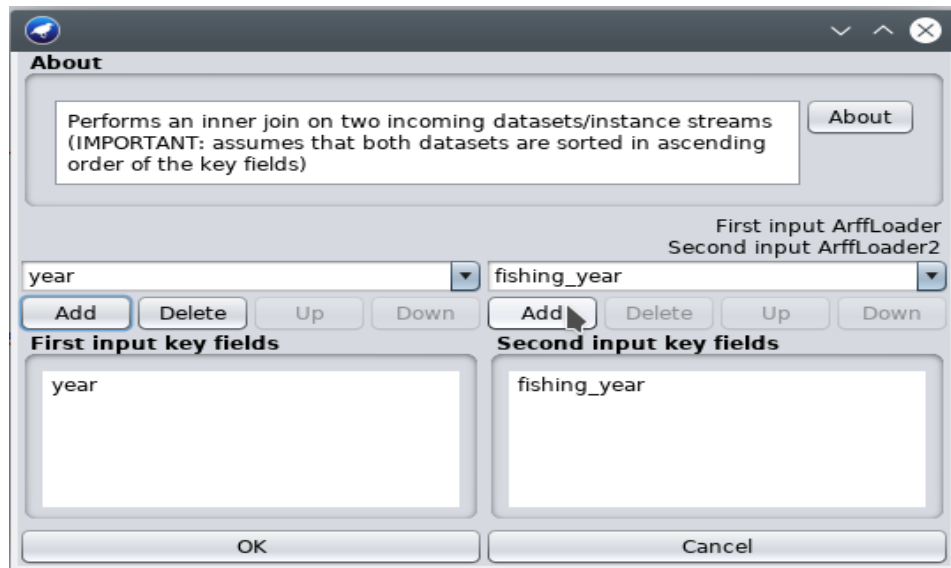
49.1276 %

17

6

Use join option to merge the new file and Data set 3 together in the pipeline





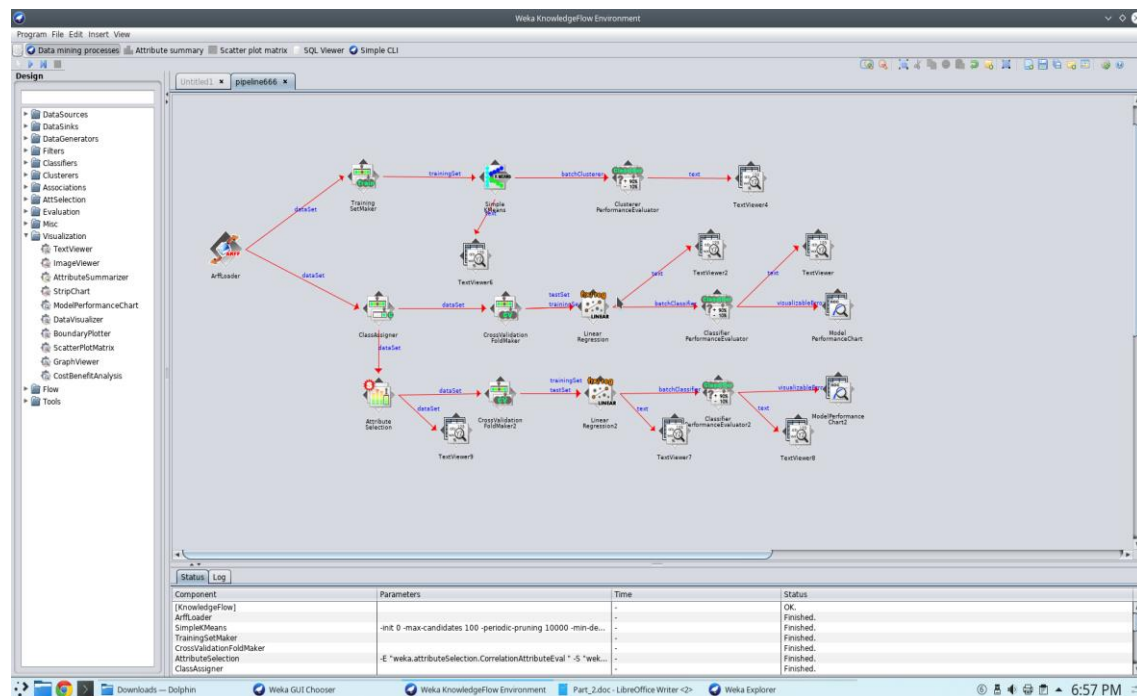
When two data sets have the same number of instanced and at least one common variable value (a common attribute), the two tables can be associated by merging.

From the output results, it could be seen that the attributes in both files have been added into the third file by merging.

3 Utilise dimensionality reduction techniques

Add Attribute Selection function to pipeline, and compare before and after using the preparation.

Dimensionality reduction has many important applications. Too high dimensionality of features will increase the training burden and storage space. Dimensionality reduction is to remove redundancy of features and express features with fewer dimensions.



Weka Explorer

Preprocess | Classify | Cluster | Associate | **Select attributes** | Visualize

Attribute Evaluator

Choose **CorrelationAttributeEval**

Search Method

Choose **Ranker -T 1.7976931348623157E308 -N 1**

Attribute Selection Mode

☐ Use full training set

☒ Cross-validation Folds **10** Seed **1**

(Num) Catch

Start Stop

Result list (right-click for options)

- 17:26:29 - Ranker + CorrelationAttributeEval
- 17:46:58 - Ranker + CorrelationAttributeEval
- 17:49:04 - Ranker + CorrelationAttributeEval
- 17:57:51 - Ranker + CorrelationAttributeEval

Attribute selection output

Fishing_year
fishing_year
TCEPR
CELR
TCER

Evaluation mode: 10-fold cross-validation

==== Attribute selection 10 fold cross-validation seed: 1 ====

average merit	average rank	attribute
0.867 +- 0.01	1.2 +- 0.4	15 TCEPR
0.866 +- 0.012	1.8 +- 0.4	16 CELR
0.568 +- 0.041	3.2 +- 0.4	4 Seafood exports quantity
0.468 +- 0.065	3.8 +- 0.4	2 TACC
0.129 +- 0.028	5 +- 0	9 From_stocks_above_hard_limit
0.121 +- 0.028	6.3 +- 0.46	7 From_stocks_above_soft_limit
0.121 +- 0.03	6.7 +- 0.46	11 From_stocks_without_overfishing
0.011 +- 0.018	8 +- 0	10 From_stocks_below_hard_limit
-0.049 +- 0.024	9.2 +- 0.6	17 TCER
-0.11 +- 0.023	9.9 +- 0.3	12 From_stocks_with_overfishing
-0.148 +- 0.026	10.9 +- 0.3	8 From_stocks_below_soft_limit
-0.494 +- 0.041	12 +- 0	1 Asset value
-0.679 +- 0.028	13 +- 0	3 Seafood exports value
-0.782 +- 0.026	14.2 +- 0.6	6 year
-0.782 +- 0.026	15 +- 0	14 Fishing_year
-0.782 +- 0.026	15.8 +- 0.6	13 Fishing_year

Status

OK Log

The above figure shows the attribute ranking related to 'catch' and the positive and negative correlation degree. It calculates the correlation between each attribute and the output variable.

We select only those attributes that have a moderate-to-high positive or negative correlation (close to -1 or 1) and drop those attributes with a low correlation (value close to zero).

According to the output results, we need to integrate the following data:

Remove:

2 redundant features: fishing year and Fishing year

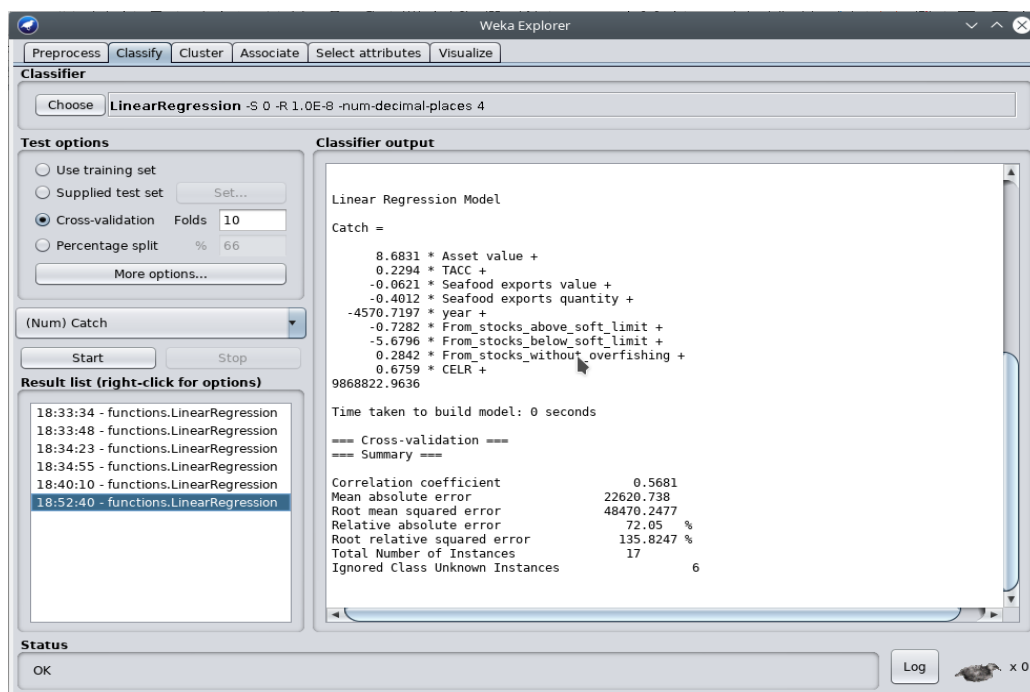
2 Irrelevant features:

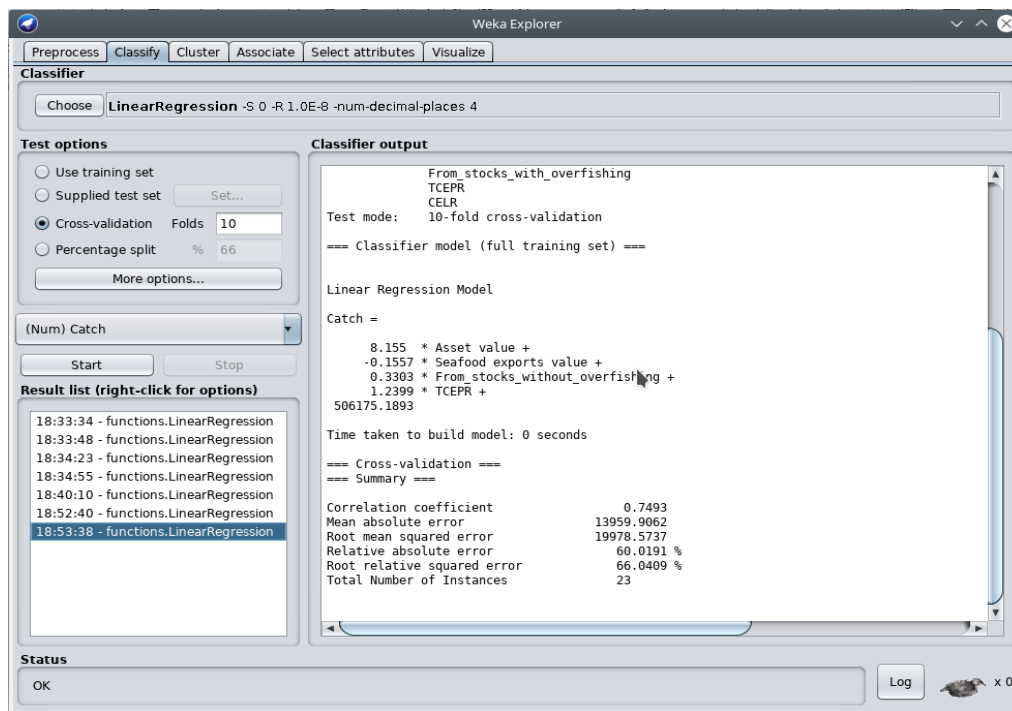
0.011 +- 0.018	8	+- 0	10 From_stocks_below_hard_limit
-0.049 +- 0.024	9.2 +- 0.6		17 TCER

Handle missing data:

Choose replace with missing value, and missing value are replaced by average values, and the result before and after treatment are compared.

4 Analyse the output





After dealing with irrelevant features and missing value, the regression model performed better than before (0.7493 and 0.5681). We could also see the root mean Squared error is still very high, which is due to the different sample measurement units and huge number gap.

The most important feature is Asset value. Also, TCEPR and From stocks without overfishing are the positive relevant attributes. Seafood exports value is negative relevant.

However, the results showed that some important indicators become irrelevant, such as From stocks below soft limit. This is probably due to the small number of data sets, replacing some attributes with values, with affects the model analysis and prediction.

In conclusion, fish asset value in New Zealand , which has the highest weights in the regression, accounts for a large proportion of fish catch. The decrease in fish catch is related to trawling and without fishing, whereas seafood exports value does not decrease .