National Research FLAGSHIPS
Wealth from Oceans

CSIRO

Cluster analysis: a novel approach to identify types of derelict nets that comprise ghost nets

Milestone Report to GhostNets Australia

Britta Denise Hardesty and Chris Wilcox

www.csiro.au

Enquiries should be addressed to:

Britta Denise Hardesty
CSIRO, Ecosystem Sciences PO Box 780 Atherton, QLD 4883

Distribution list

Riki Gunn	Electronic copy,

Copyright and Disclaimer

Important Disclaimer

CSIRO advises that the information contained in this publication comprises general statements based on scientific research. The reader is advised and needs to be aware that such information may be incomplete or unable to be used in any specific situation. No reliance or actions must therefore be made on that information without seeking prior expert professional, scientific and technical advice. To the extent permitted by law, CSIRO (including its employees and consultants) excludes all liability to any person for any consequences, including but not limited to all losses, damages, costs, expenses and any other compensation, arising directly or indirectly from using this publication (in part or in whole) and any information or material contained in it.

Introduction

We developed a statistical model that allows us to cluster nets into unique types or aggregates. Each of these unique types is known as a cluster. The model balances having a large number of types against the potential that there is variation within a type, either due to true variation or measurement error, to come up with the most likely number of net types (or clusters).

This analysis required the development of a customized mixture model that could handle both discrete and continuous data, as none were available. The model returns a mean and variance-covariance for each continuous variable (e.g. mesh size) for a cluster, which are assumed to be multivariate normally distributed. For each discrete variable in each cluster the model returns a probability of having each of the possible states of the variable. For instance, for knots, there are three categories in the data, yes, no, and NA (not applicable). For cluster (or equivalently net type) 1, the relative probabilities of each of these states are: 0.5181387 0.23540905 0.2464522. Therefore, a net with knots is twice as likely as a net without to be of net type 1, keeping in mind that we are ignoring its other characteristics.

The parameters from the cluster analysis can be used to assign additional nets to types, which will allow us to determine a range of things such as whether different types of nets arrive on different beaches or if more turtles appear in one type than another.

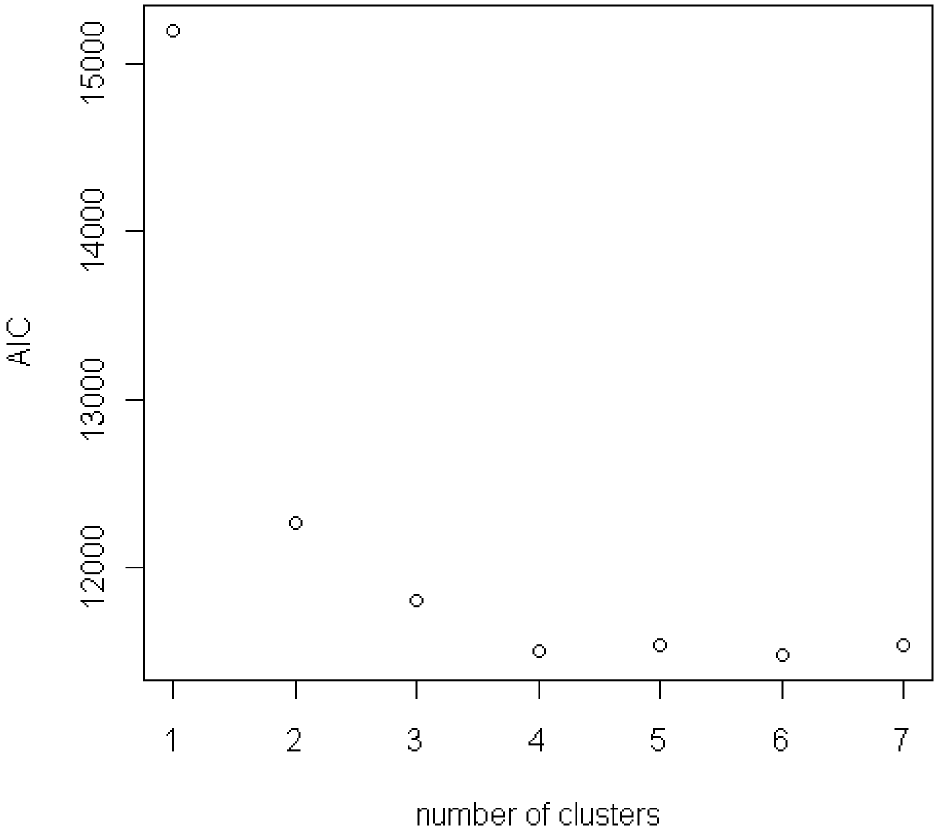
Cluster analysis is sensitive to outliers in the data, as these will appear to be singularities, i.e. clusters with a single observation. This results in the analysis method having numerical difficulties as it cannot estimate a variance for a cluster that has only a single observation. Due to this issue, it was necessary to carefully screen the data, and to remove a few values. Similarly, for estimating the number of clusters and their parameters, all of the observations need to have complete data. This requirement led to the need to exclude any observation with missing information. Assignment of new observations to clusters can still be done with missing data, although assignment will only be done probabilistically over the remaining data.

Results and Discussion

After cleaning the data there were 1333 observations that were usable. These observations each had 8 characteristics, including: color, strand type, whether it was braided or twisted, the number of strands (if twisted), double or single twine, mesh size, twine size, and net type (whether it was knotted or not). As some of these characteristics were only available if the net was of a particular type, for instance only twisted nets have a number of strands, categories with missing information of this type were coded as “other”. This allows the data to remain in the analysis, as there are no missing data per se, just irrelevant measurements.

We ran the analysis exploring models with numbers of clusters ranging from 1 (all nets are the same) up to 7.   The Akaike Information Criterion (AIC) is a statistical tool used to identify models that best fit the data while not containing too many parameters. In the context of cluster analysis the AIC provides a tool for identifying the best number of clusters

Figure 1. AIC values for cluster analysis of the current ghost net data.



to use in describing a particular data set. As the number of clusters increases from 1, the model with the number of clusters that produces the first minimum in AIC (i.e. models with 1 more or 1 less cluster both have higher AICs) is the optimal model. Figure 1 shows the results for the analysis, with 4 clusters giving the first minimum AIC value.

The attached excel workbook provides the original data, along with a cleaned format. The final sheet in the workbook contains the output from the cluster analysis. It gives the probabilities for each observation of belonging to each cluster and based on these probabilities provides the most likely cluster.

Cluster Number	Number of Observations
1	29
2	780
3	491
4	33

In summary, most of the nets in the dataset were identified as belonging to cluster 2, with the second most common cluster being 3 (Table 1). The two remaining clusters were relatively uncommon.

The following pages provide the parameter values for each of the clusters for each characteristic. The profile of a particular cluster can be found by looking down through the tables and constructing each characteristic. For instance, type 1 nets have a mesh size of 83 and a twine size of 34, they are mostly white or green with some blue or grey, generally twisted multistrand nets with 3 or 13 strands, generally have single twine, and more often knots.

One issue with this analysis is that the number of clusters that can be identified is limited by the number of observations available. With additional observations it might be possible to identify clusters that are present in the existing data, and there may also be other clusters that have not appeared in this data that could be found when more observations are accumulated.

The next step in the analysis is to evaluate whether the clusters identified make sense, if any additional information can be added to distinguish them, and then whether there is any additional data that can be added to assist in clarifying the cluster profiles.

Mean of continuous variables for each cluster

Cluster	Mesh Size	Twine Size
1	82.33982	34.502421
2	77.16413	1.939168
3	187.67565	3.483479
4	1355.73602	5.462768

Probability of each colour for each cluster

Cluster	green	white	blue	grey	light blue	black	yellow	red	orange	brown
1	0.2294	0.3154	0.1556	0.1018	0.0470	0.0001	0.0730	0.0775	0.0001	0.0000
2	0.4472	0.1178	0.2774	0.1110	0.0263	0.0059	0.0042	0.0062	0.0027	0.0013
3	0.4030	0.0676	0.4081	0.0714	0.0198	0.0031	0.0111	0.0083	0.0076	0.0000
4	0.4931	0.0827	0.1661	0.1342	0.0706	0.0001	0.0031	0.0498	0.0003	0.0000

Probability of each type of line for each cluster

Cluster	multistrand	mono
1	0.994	0.006
2	0.982	0.018
3	0.998	0.002
4	0.975	0.025

Probability of twisted or braided for each cluster

Cluster	twisted	other	braided
1	0.994	0.006	0.000
2	0.971	0.018	0.011
3	0.979	0.002	0.018
4	0.973	0.025	0.001

Probability of the number of strands for each cluster

	3		2		5	4	13
Cluster	strands	Other1	strands	Other2	strands	strands	strands
1	0.4119	0.0057	0.0005	0.0004	0.1667	0.1011	0.3137
2	0.9274	0.0182	0.0159	0.0112	0.0176	0.0071	0.0026
3	0.6753	0.0025	0.0117	0.0185	0.2289	0.0552	0.0079
4	0.6481	0.0255	0.0006	0.0012	0.0824	0.1835	0.0587

Probability of single or double twine for each cluster

cluster	single	other	Double
1	0.971	0.006	0.024
2	0.846	0.018	0.136
3	0.953	0.002	0.045
4	0.912	0.025	0.063

Probability of knots for each cluster

Cluster			
	knots	no knots	other
1	0.518	0.235	0.246
2	0.613	0.011	0.375
3	0.620	0.000	0.380
4	0.608	0.000	0.392

Contact Us

Phone: 1300 363 400
+61 3 9545 2176

Email: enquiries@csiro.au

Web: www.csiro.au

Your CSIRO

Australia is founding its future on science and innovation. Its national science agency, CSIRO, is a powerhouse of ideas, technologies and skills for building prosperity, growth, health and sustainability. It serves governments, industries, business and communities across the nation.