University of British Columbia Statistics Department
Technical Report # 179

Seals, talk to us and we will identify you.

Keith Chan
Jaroslaw Harezlak
Robert St-Aubin
Marc Vallée

Department of Statistics
6356 Agricultural RD # 333
University of British Columbia
Vancouver, B.C.
Canada
V6T 1Z2

December 1998 (from February 24, 1998 manuscript)

1 Introduction

The purpose of our analysis is to find the statistical method which best discriminates and classifies the calls of the 3 herds of seals. If the calls can be discriminated, this would suggest no inter-breeding between the herds.

The data contains continuous, nominal and discrete variables. This led to a search for a method that suits this type of data. Krzanowski (1975) introduced a method of discrimination based on a mixture of indicator and continuous variables. We can incorporate this method by coding the nominal variables having k states into $(k-1)$ indicator variables. We used this idea in transforming the nominal variables.

Section 2 describes the initial data analysis, where the distributions of each variable for each herd are compared using descriptive statistics. This gave an idea of which variables would be useful in differentiating the herds. The following Section outlines the different methods used and their properties. The last two Sections report the results of the analysis and the evaluation of the methods.

2 Initial Data Analysis

The data set consists of 7 variables on 3000 calls equally divided among the three herds of seals. These 7 variables are a mixture of data types: element duration (V1), time between elements (V2), pitch at start of call (V4) and pitch at end of call (V5) being continuous variables while number of elements (V3) is discrete and waveform shape (V6) and call shape (V7) are nominal. The question here is which variables will be helpful in differentiating the herds?

For each of the three herds, histograms and boxplots for element duration were constructed to see if the distribution differs from herd to herd. If the distributions do differ, this may be an indicator that the variable will be useful for discrimination. The distribution for each herd appeared skewed to the right with that of the Jan Mayen herd being more skewed than the other two; see Figure 1.

A similar analysis was done for number of elements and time between elements; see Figures 2 and 3. The distribution for number of elements appeared skewed to the right with the Jan Mayen herd being distributed less skewed than the two others.

The distribution of time between elements, or interduration, was slightly skewed to the right with the Front herds' distribution being less skewed than the other two. This variable only had an entry if the number of elements was greater than 1. This led to the consideration of two analyses; an analysis on just the multiple element calls and an analysis on all the calls. To include information on interduration in the latter analysis, we created an indicator variable for interduration being greater than 300 ms. We considered other splits as well but 300 ms seemed to differentiate the herds rather well; see Figure 4.

The number of multiple element calls from each herd was unequal as a consequence of differing number of single element calls from herd to herd. The number of multiple element calls was 482 from the Jan Mayen herd, 605 from the Gulf herd and 575 from the Front herd. Although unequal, these counts do not differ greatly.

Figure 1. Histograms for Element Duration (Note: The Histograms are for Element Durations less than 3000 ms. For each herd, less than 1% of the Element Durations are greater than 3000 ms.)

Figure 2. Histograms for Number of Elements.

Figure 3. Histograms for Time Between Elements.

Figure 4. Comparison of the Interduration of the Three Herds.

In order to reduce the skewness, the log transformation of element duration and time between elements and the square root transformation of number of elements were considered.

The frequencies of waveform shape are high for a few classes, {2,5,7}, and quite small for the remainder; see Figure 5. The proportion of waveform shape 2 for the Jan Mayen herd is much larger than that of the other herds, while the proportion of waveform shape 5 for the Jan Mayen herd is smaller than for the other herds, and the proportion of waveform shape 7 are roughly the same for each herd.

The frequently occurring call shapes were {0,5}; see Figure 6. The proportion of call shape 0 for the Jan Mayen herd is much larger than that of the other herds and the proportion of call shape 5 for the Jan Mayen herd is slightly less than for the other herds.

Figure 5. Comparison of Waveform Shape of the Three Herds.

Figure 6. Comparison of Call Shape of the Three Herds.

These two nominal variables can not be used in the analysis as they are presented. The classes have no obvious ordering and the indexing of the classes by integers is arbitrary. To incorporate these variables in our analysis, we considered applying the idea of Krzanowski (1980) but the number of variables in the analysis would then increase drastically as there are 8 classes for waveform shape and 12 classes for call shape. Hence, we decided to include only the high frequency classes of the two variables, i.e., indicator variables for waveform shapes {2,5,7} and call shapes {0,5}.

The correlation of pitch at start of call and pitch at end of call in the octave scale was 0.85. Due to this large value, the sum of the logs of the two variables (denoted as pitch.sum) and the difference of the logs of the two variables (denoted pitch.dif) were used rather than the two individual highly correlated variables. These two new variables had a low correlation of 0.18 and may be more useful for discrimination. The distribution of pitch.sum seems to differ for each herd and the distribution of pitch.dif seems to distinguish the Jan Mayen herd from the other two.

The selection of variables to be used in our final analysis was arrived at in a preliminary stage using different discrimination methods (to be discussed in Section 2) and different combinations of variables. The confusion matrix and error rate (to be discussed in Section 3) are the premiss on which these variables were deemed useful for discrimination and classification. The variables pitch.dif and wshape.7 were not selected as they had close to no effect on the error rate. The variables in the final analysis were as follows:

$$\text{elemdur.log} = \log_{10}(V1), \quad \text{numelem.sqrt} = \sqrt{V3}$$

$$\text{pitch.sum} = \log_2(V4) + \log_2(V5)$$

$$\text{wshape.i} = \begin{cases} 1, & \text{if waveform shape is i, i=2,5;} \\ 0, & \text{otherwise.} \end{cases}$$

$$\text{cshape.i} = \begin{cases} 1, & \text{if call shape is i, i=0,5;} \\ 0, & \text{otherwise.} \end{cases}$$

The analysis that included all calls used in addition:

$$\text{interdur.300} = \begin{cases} 1, & \text{if interduration} > 300; \\ 0, & \text{otherwise.} \end{cases}$$

The analysis that involved only multiple element calls used the above variables with interdur.300 replaced by interdur.log $= \log_{10}(V2)$.

3 Methodology

We have used four methods of classification in our analysis of the data. Our data are multivariate, therefore we wanted to find some combination of variables maximizing the between-class spread to within-class spread. We briefly describe each of the methods used.

- Linear Discriminant Analysis (LDA)

  Introduced by Fisher (1936), LDA seeks a linear combination of the variables which achieves the maximum ratio of the separation of the class means to the within-class variance. Computationally the problem may be solved using principal components of the within-class covariance matrix. LDA is an optimal procedure for classifying normal populations with the same covariance structure. Fisher's rule gives lines as separators between classified populations.

- Quadratic Discriminant Analysis (QDA)

  This method is a variation of LDA which does not require the equality of the covariance matrices of the populations considered. The rule obtained using this procedure gives quadratic separation curves.

Our hesitation with using one of the above procedures is based on the data on hand. Both procedures are optimal for multivariate normal data and intended for use with continuous responses only. However the seal data is mixed (some variables continuous and some categorical). Krzanowski (1977) used computer simulations to evaluate the performance of the Fisher's linear discriminant function. He noted: "A low correlation in one population but a high correlation in the other, or a change in the sign of the correlations between the two populations could indicate conditions unfavorable to Fisher's linear discriminant function."

Therefore, we tried to find more suitable procedures for producing appropriate classification rules. We have found two methods more suited to the seal data: Mixture Discriminant Analysis (MDA) and Classification and Regression Trees (CART).

- Mixture Discriminant Analysis (MDA)

  Hastie and Tibshirani (1994) introduced a whole new class of discrimination/classification methods, the main feature being the modeling of each class by the mixture of multivariate normals with common covariance matrix. MDA inherits some of the properties of LDA. It produces a hierarchy of coordinates in terms of their abilities to discriminate between the classes. An additional benefit of MDA is a greater number of coordinates on which to base the classification. MDA can be considered as a smooth version of Learning Vector Quantization - LVQ (Kohonen 1989). LVQ finds a set of cluster centers for each class, and the classification is performed by finding the closest cluster center and assigning a new observation to the associated class. The program implementing the MDA procedure can be thought of as a repeated-regression procedure using optimal scoring. Any method of regression can be used. Our analysis used linear regression and MARS (Multivariate Adaptive Regression Splines) - Friedman (1991). MARS enriches the procedure by transforming and expanding the original data set. MDA has two main advantages, one is the use of non-linear decision boundaries and the other is that it is more efficient than the classical methods for dealing with mixed (continuous-categorical) data.

  Please note that for our analysis, we used the MDA method with 2 and 3 subclasses in order to verify the sensitivity of the mixture discriminant analysis method to the numbers of subclasses used.

- Classification and Regression Trees (CART)

  This approach is very different from the others. The procedure starts by considering all the observations to be in a single group. At each successive step the probability distribution over the categories is computed for each observation and the Bayes decision rule is used to determine the category with highest probability. The implementation of CART used in our analysis employed the most common method called one-step lookahead. This method avoids the complexity of optimizing the performance of the whole tree by choosing the consecutive splits in an optimal way.

We tried to determine the most appropriate procedure for the classification task. Since optimality of most classification methods relies heavily on the normality assumption, we sought more flexible procedures suitable this data. As the results in the next section indicate, some of the methods were inappropriate for use with mixed data, and they gave poor classification rules.

4 Analysis and Results

The main goal of this analysis was to find an objective classification method using a few easily measured variables. As described in Section 2, we decided to analyze two different sets of data. The first one was called "Multiple Elements Calls" and the second one was referred to as "All Calls".

To determine which of the methods described in Section 3 performs best, we compared the performance of all four methods when applied to the two data sets mentioned above. To order the methods according to their performance, two tools were used: confusion matrix and error rate. As introduced by Massy (1965), the confusion matrix sets out the results of applying a discriminant analysis procedure. If there are k possible categories, it is a $k \times k$ matrix with columns denoting the true category and rows the category assigned. Confusion (misclassification) is represented by the off-diagonal elements. In our case, we will obtain a 3×3 matrix. However, this matrix does not use the information on proportions of individuals in each category. This is why we will also compute, for each method, an error rate that corresponds to the proportion of misclassified observations.

Our analysis was divided into two major parts: discrimination and classification. Discrimination was carried out on a random sample taken from the data. We called this sample the "training sample", because it is the one on which we built the model. The classification step consisted of applying the resulting model to the complement of the training sample (the "validation sample") which was intended to test the predictive performance of the discrimination method used. This was measured by computing the confusion matrix and the error rate based on these results.

In order to decide on the size of both the training and validation sample, we did a sensitivity analysis to investigate the stability of the confusion matrix and the error rate to the size of the validation sample. Those quantities appeared quite stable and not very dependent on the sample size, so we opted for a validation sample of size 300 (100 calls per herd) for the sake of simplicity. Having fixed the size of the sample to be used in the classification step, the training sample consisted of all the observations not in the validation sample.

We proceeded with this analysis for each of the four methods mentioned in Section 4. Moreover, the analysis was repeated 10 times for each method (ie. we randomly selected a new training sample in each of the ten repetitions) in order to get an understanding of the variability of the results. The average confusion matrices and error rates based on the ten repetitions for all five methods and for both data sets are presented in Tables 1 and 2.

- Tables 1-2 presented here.....
- Table 1 = Error Rates
- Table 2 = Confusion Matrices

The ordering of the methods are very similar for both data sets. However, it seems that the classification was superior in the "Multiple Elements Calls" set. Therefore, we will focus on those results . Nevertheless, it is to be noted that these conclusions apply as well to the other data set.

First, let us examine how well each technique of discrimination performed. From Table 1, it can be seen that of the five techniques used, CART performed the best while the LDA and QDA methods performed the worst. For CART we got an error rate as low as 29% on the validation sample. All other techniques gave error rates above 35%, with LDA and QDA being the highest at 44% and 40% respectively. It is also important to note that while the MDA method with 3 subclasses did pretty well, use of only two subclasses did not give very good results. Therefore we conclude that the MDA method is sensitive to the number of subclasses. It is also important to note that the standard deviations for each method are slightly higher for the validation sample than for the training sample. This is not surprising as the models were built using the training sample. Also, most methods have a SD around 20 for the validation sample except QDA and MARS which yield a SD of 35 and 9 respectively. The MARS method being much more stable than the other methods, including CART, may create some doubts about the predicting superiority of CART. However, using a one standard deviation confidence interval for the error rate does not give any overlap between MARS and CART. Thus based on these results, we will now consider CART as our method of choice for classification.

Secondly and most importantly, we will now try to answer the initial question that led to this analysis. Are the three herds of harp seals different in their vocalization patterns?

The diagonal of the CART confusion matrix, indicates a pretty high proportion of correct classification for each herd. Moreover, the first column of this confusion matrix shows that almost 80% of the time we classify correctly an observation from the Jan Mayen herd. Furthermore, adding lines 2 and 3 of the last two columns of the same confusion matrix shows that we are able to distinguish between the Jan Mayen herd and the joined Gulf/Front herd in 90% and 94% of the cases respectively. Thus, it appears that the underwater vocalizations of the Jan Mayen herd are different from those of the Gulf and Front herds.

Slightly weaker results arise when looking at the classification of the second and third herd. We still obtain a high number of correct classifications, but the distinction between the Gulf and Front herd is not as obvious as in the Jan Mayen herd case. It appears to be more difficult to distinguish between the last two herds, although there is some evidence of differences in the vocalization patterns.

We are then led to conclude that the vocalization of the Jan Mayen herd does differ from the two remaining herds. However, we cannot, based on our results, provide strong evidence that the herds from the Gulf and Front have different vocalization patterns.

5 Discussion and Conclusion

Our main objective was to see if there was a way of discriminating between the three herds of seals based on different characteristics of their vocalizations. We now summarize our findings from the previous sections. In the initial data analysis the production of histograms and box plots suggested that we would not encounter great difficulties in discriminating the Jan Mayen herd from the two others. The distinction between the Gulf and the Front herds was not obvious although the inter-duration time seemed to differentiate the two. Inter-duration being available only for multiple element calls led to two separate analyses: All Calls and Multiple Element Calls.

With regards to methods, our initial idea was to conduct a discriminant analysis, but due to the different types of variables in the dataset we had to find a particular version which would be appropriate in dealing with this kind of data. Even when considering the theoretical properties of each of the suggested methods, the best discrimination technique was not obvious. Hence, a measure of their individual performance had to be evaluated.

The performance of each method was compared using confusion matrices and error rates. Among the methods used in our analysis. CART had the best classification results for both sets of data. Our overall error rates ranged from 29% (using Multiple Elements Calls) to 36% (using All Calls). Considering the individual herd's proportion of correct classification we can see that the Jan Mayen herd differs from the two others with 78% (multiple elements calls) correct classification. The discrimination between the Front and Gulf herds was not as evident, although we still got a majority of correct classification in both cases.

Although the results were satisfactory, some issues about generalization to each herd arose. Firstly, we had limited information on the sampling scheme. The analysis should not be based on calls all coming from the same seal for each group. Secondly, the recording instruments were not the same for all three herds. Can this fact be ignored? Should there be any doubts about the instruments' reliability? Finally, we were concerned with the particular scale used to identify the call shape. This variable is presented on a nominal scale indexed by integers yet, in the data set, it also takes on non-integer values. We treated these non-integer values as additional classes of this variable.

To conclude, although it was later implied by Dr. Terhune in his initial address at the 1997 SSC Annual Meeting that the herds from the Front and the Gulf did in fact inter-breed, we do not believe that this undermines the fact that we did find a difference. Perhaps, even if they do inter-breed, there still exists a difference in their calls. Even if the distinction between these two herds is not as obvious as the one discriminating them from the Jan Mayen herd, the fact that we did obtain a fairly high proportion of correct classification should indicate that the variables do present information that makes it possible to distinguish between these two herds. As it is sometimes said: "If you look hard enough for something, chances are you will find it.". Is that what happened here? We think not. In fact, we strongly believe that what we did was listen to what the data, or in this case the seals, were saying.

REFERENCES

Friedman, J. (1991). Multivariate adaptive regression splines (with discussion). *Ann. Statist.*, 19, 1-141

Hastie, T., Tibshirani, R., Buja, A. (1994). Flexible discriminant analysis by optimal scoring. *J. Amer. Statist. Assoc.*, 89, 1255-1270

Hastie, T., Tibshirani, R., Buja, A. (1996). Discriminant analysis by gaussian mixtures. *J. Roy. Statist. Soc. Ser. B*, 58,

Johnson, R.A., Wichern, D.W. (1992). *Applied Multivariate Statistical Analysis*, Prentice Hall.

Kohonen, T. (1989). *Self-Organization and Associative Memory*, Springer-Verlag.

Krzanowski, W.J. (1975). Discrimination and classification using both binary and continuous variables. *J. Amer. Statist. Assoc.*, 70, 782-790

Krzanowski, W.J. (1977). The performance of Fisher's linear discriminant functtion under non-optimal conditions. *Technometrics*, 19, 191-200

Krzanowski, W.J. (1980). Mixtures of continuous and categorical variables in discriminant analysis. *Biometrics* 36, 493-499

Lachenbruch, P.A., Goldstein, M. (1979). Discriminant analysis. *Biometrics*, 35, 69-85

Massy, W.F. (1965). *J. Advert. Res.*, 5, 39-48

Terhune, J.M. (1994). Geographical variation of harp seal underwater vocalizations. *Can. J. Zool.*, 72, 892-897

Venables, W.N., Ripley B.D. (1994). *Modern Applied Statistics with S-PLUS*, Springer-Verlag

ACKNOWLEDGEMENTS

We wish to thank Dr. John Petkau for the very valuable advice and help provided throughout the redaction of this article.

Thanks to Harry Joe, Nancy E. Heckman, James V. Zidek and Paul Gustafson for financial support and helpful comments.

We would also like to thank all the other members of the Department of Statistics at U.B.C. who helped us participate in the SSC data analysis competition.

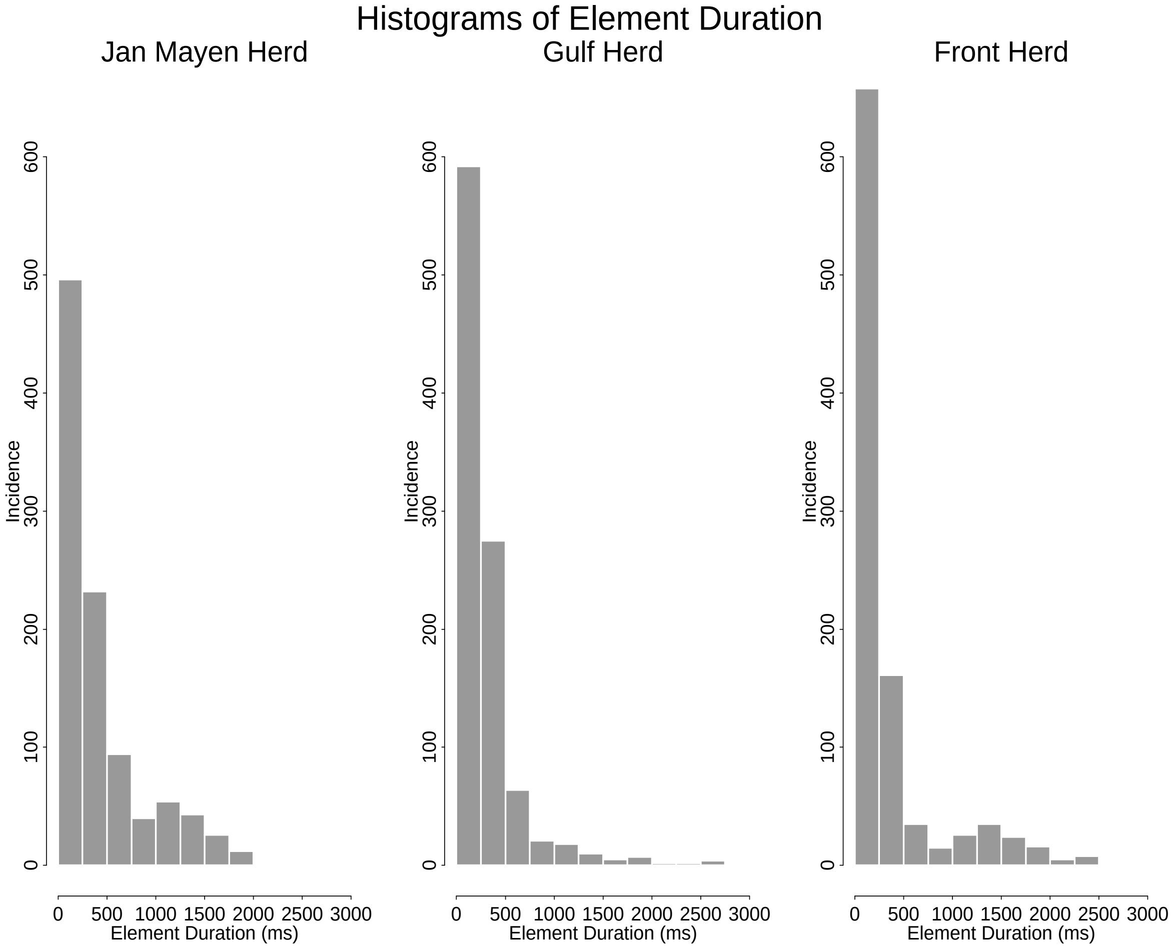

Figure 1: Histograms for Element Duration

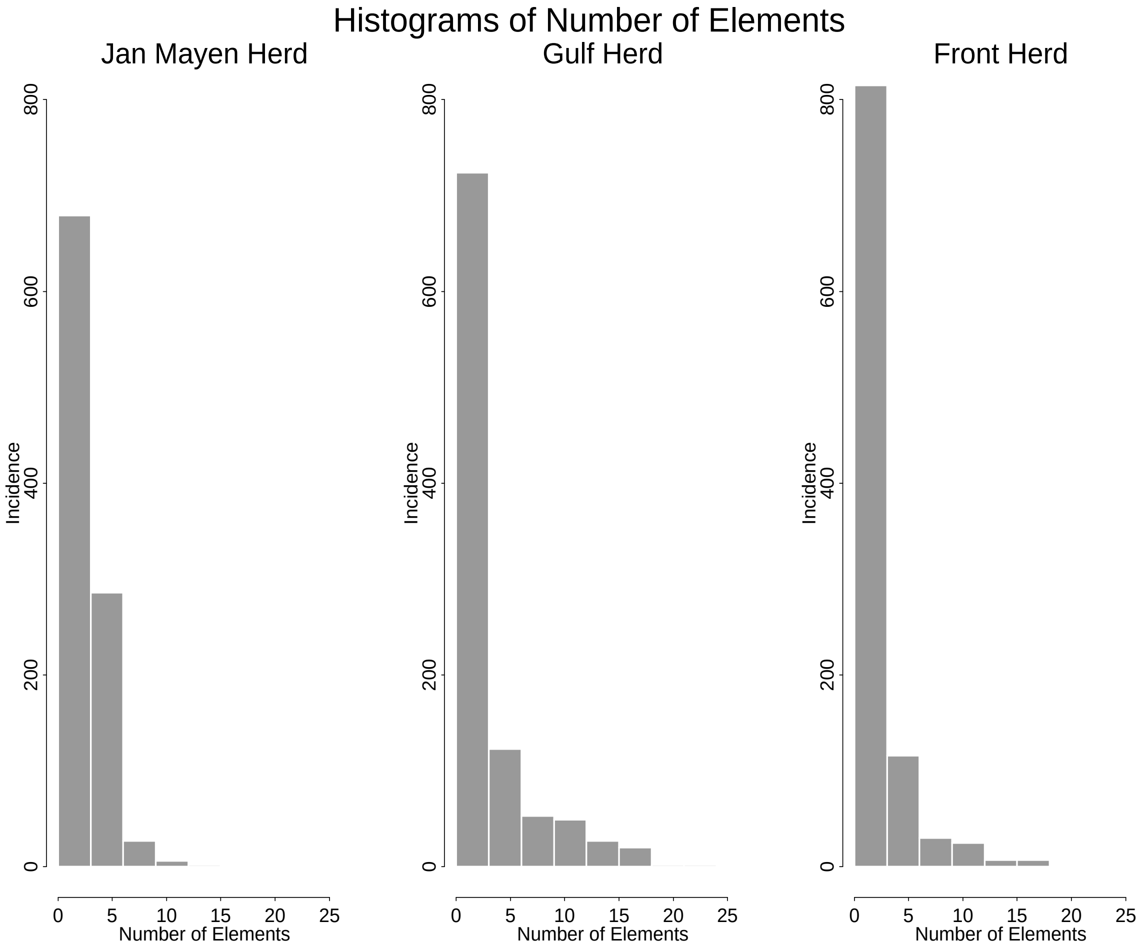

Figure 2: Histograms for Number of Elements

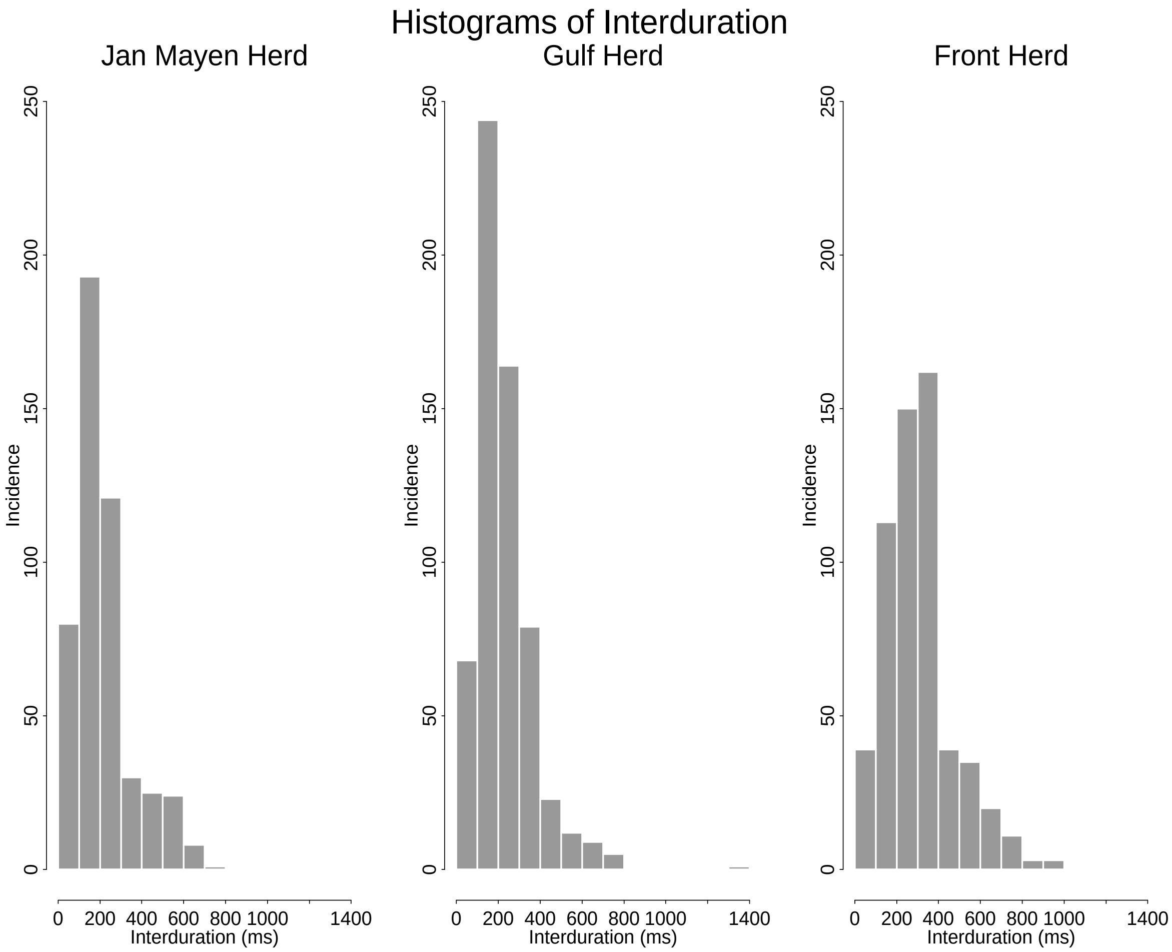

Figure 3: Histograms for Time Between Elements

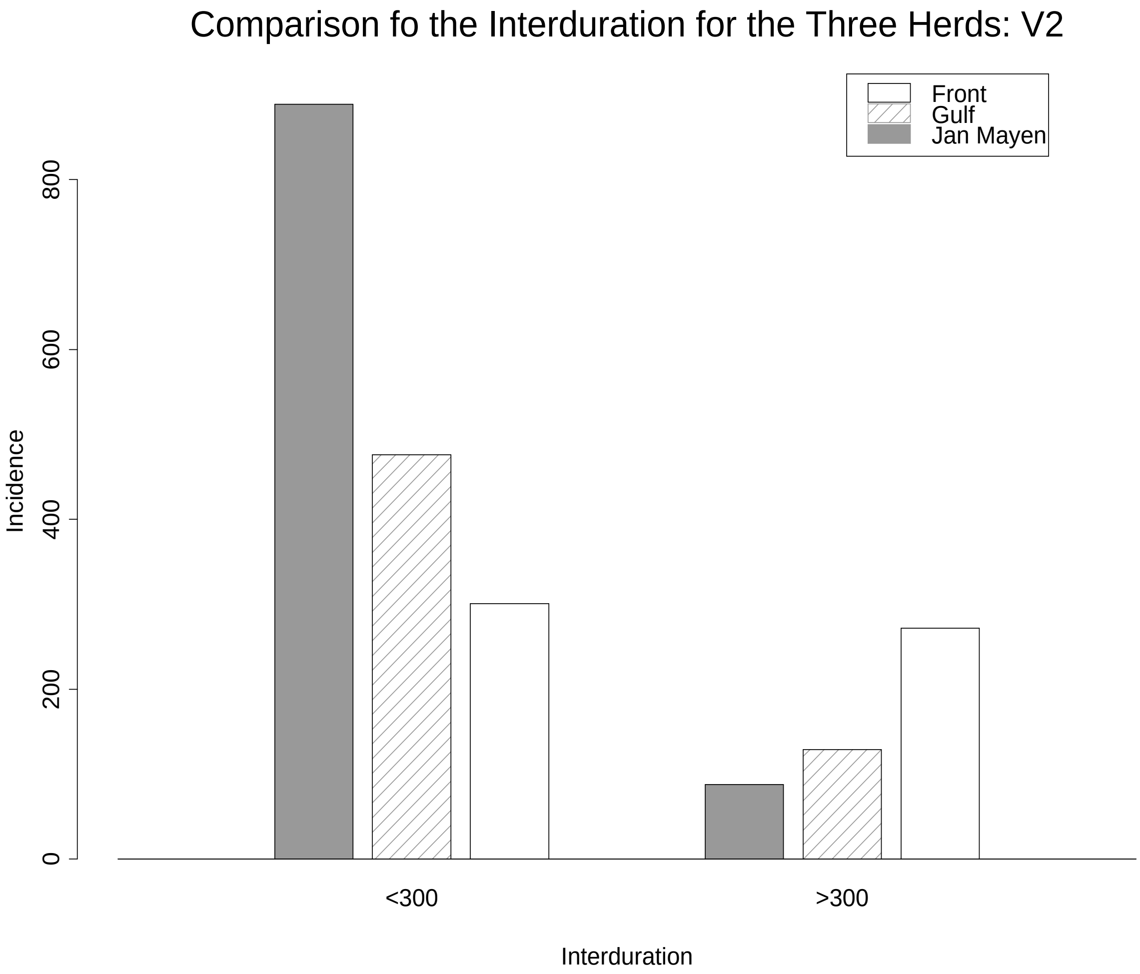

Figure 4: Comparison of the Interduration of the Three Herds

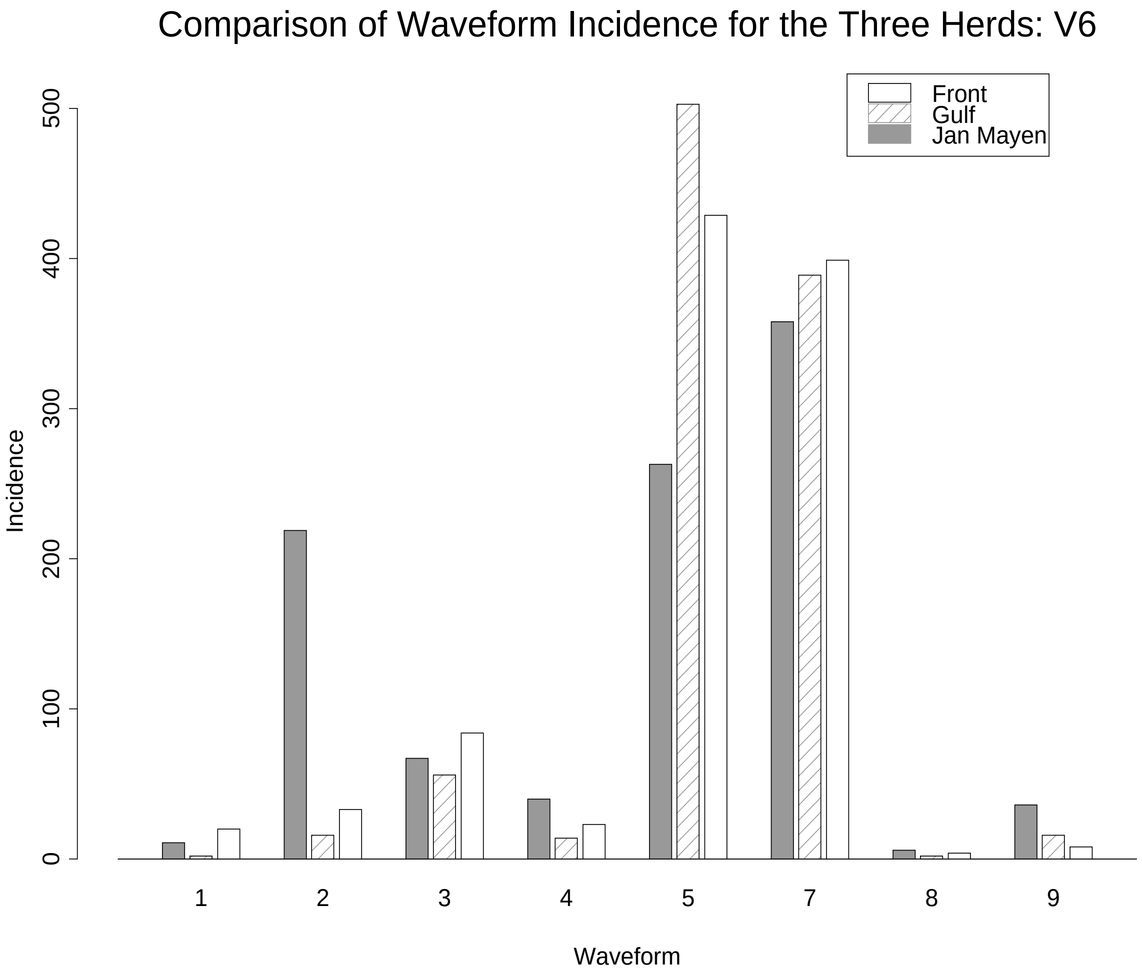

Figure 5: Comparison of Waveform Shape of the Three Herds

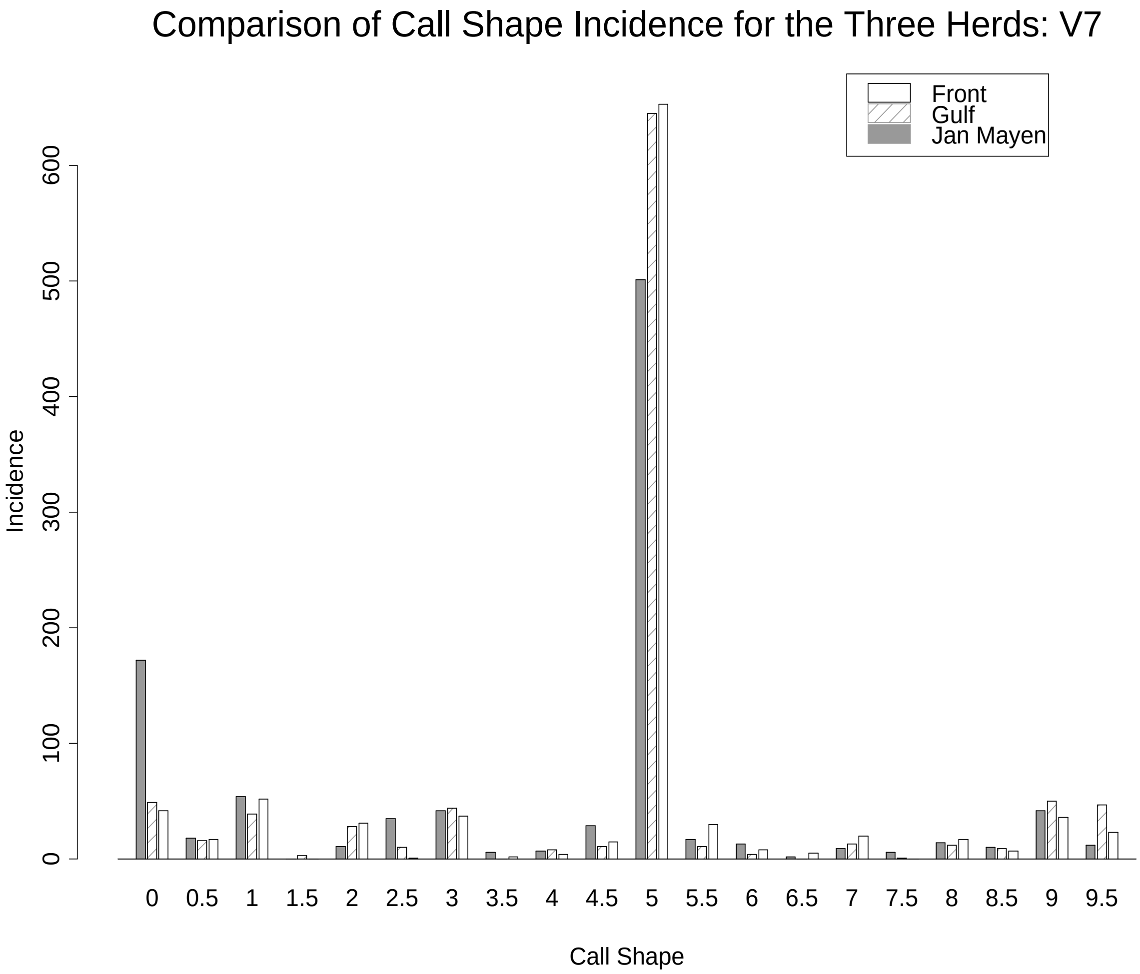

Figure 6: Comparison of Call Shape of the Three Herds

Error Rates $\times$ 1000 (SD)

Technique	**All Calls**		**Multiple Element Calls**	
	Validation	*Training*	*Validation*	*Training*
LDA	473 (22)	490 (3)	439 (23)	425 (5)
QDA	491 (27)	500 (5)	398 (35)	383 (6)
MDA (2 subclasses)	474 (21)	457 (4)	415 (16)	415 (5)
MDA (3 subclasses)	446 (26)	434 (4)	371 (20)	359 (8)
MARS	426 (24)	414 (4)	350 (9)	349 (5)
CART	362 (24)	306 (8)	291 (28)	220 (12)

Confusion Matrices

Technique	All Calls 1	2	3	Mult. Calls 1	2	3
LDA	**59**	15	27	**48**	4	9
	23	**58**	32	23	**61**	32
	18	27	**41**	29	35	**59**
QDA	**49**	13	20	**54**	7	8
	29	**61**	37	20	**58**	23
	22	26	**43**	26	35	**69**
MDA (2 subclasses)	**64**	23	29	**59**	13	12
	20	**54**	31	21	**58**	30
	16	23	**40**	20	29	**58**
MDA (3 subclasses)	**73**	27	34	**61**	6	10
	11	**51**	23	22	**64**	26
	16	22	**43**	17	30	**64**
MARS	**65**	20	18	**70**	10	10
	17	**58**	32	16	**65**	31
	18	22	**50**	14	25	**59**
CART	**70**	13	13	**78**	10	6
	16	**59**	25	13	**65**	24
	14	28	**62**	9	25	**70**