Feature Selection Methods for K-Nearest Neighbor Discriminant Analysis of Acoustic Resonance Signatures

Ben Peressini
Montana State University
Fall 1994

INTRODUCTION

Non-intrusive, non-destructive identification methods of fill materials in various containers have important applications. Such methods are of use when contact with or destruction of the containers is expensive or dangerous. Some applications include quality inspection, treaty verification, and facilitation of efficient and safe disposal of hazardous munitions.

One non-intrusive, non-destructive method of fill identification is based on examining the container's resonance properties. In this method, a broad band acoustic source generates vibrations on the object's surface. The vibrations induce natural resonance modes on the container which are measured with a laser vibrometer, producing a resonance spectrum or signature. If the fill affects the vibration characteristics of its container in a detectable and consistent manner, the vibration characteristics should be apparent in the signature, thus providing the necessary information for fill material classification. In particular, fill density and viscosity may affect the vibration resonance frequencies and their amplitudes.

Classification is obtained through a *k*-nearest neighbor discriminant statistical analysis of the resonance spectra. In this method, a vector of spectrum features from an unknown signature is compared to similar vectors from a group of training signatures with known fill characteristics. The identities of the training signatures that are the *k* nearest neighbors (in the vector space) to the unknown signature are used to classify the unknown signature. Proximity in the vector space is determined using Mahalanobis distance or a similar metric. To the degree that the feature vectors cluster the signatures in distinct regions of the vector space, successful classification is achieved.

The key to successful classification is in determining an appropriate vector of variables or features to use in the discriminant analysis. The acoustic signature raw data consist of a large number of frequencies and their associated resonance amplitudes. It is theoretically possible to perform a *k*-nearest neighbor discriminant analysis based on all of the raw data. Using so many variables, however, would lead to slow implementation and may not be possible except with very large computing resources. More importantly, using all of the variables would not necessarily produce the best classification results.

The purpose of this report is to investigate which of several feature selection methods for acoustic data will yield the best set of discriminant variables. The feature selection methods that were examined are based on one of two approaches. In the first approach, frequencies having the maximum amplitudes are identified. The second approach is based on one-to-one pattern matching of all peak locations within a selected frequency range. A peak is an amplitude larger than each of its two immediate neighbors.

The feature selection methods were applied to data pertaining to the classification of simulated chemical weapon fills. Three data sets of vibration resonance signatures from two container types were examined. The container types consisted of 155 mm artillery munitions and capped pipes. They were filled with varying amounts and types of chemical fills. These chemical fills were chosen to simulate the density and viscosity characteristics of agents found in chemical weapons. In addition, the focal location of the measurement laser was varied. The objects were classified according to both the type and amount of fill they contained.

This report briefly describes the theory behind the inspection method as well as the laser-acoustic system used to determine fill particulars. Details about the data set details also are provided. The method of *k*-nearest neighbor discriminant analyses is described and the manner in which the degree of success was judged is explained. The remainder of the report is devoted to an exploratory investigation of the feature selection methods and a comparison in terms of successful classification of the various data sets.

INSPECTION SYSTEM

To obtain an object's resonance signature, the object is first made to vibrate in response to broad band white noise emitted by a loudspeaker. The object's acoustic excitation is measured with a non-contacting laser vibrometer focused at a point on the object. Doppler shifting in the laser light is sensed by the instrument and converted to instantaneous surface velocity measures which are in turn digitized. Fast Fourier transforms are calculated from the digital data and signatures are estimated by averaging the fast Fourier transforms. For more detail on the laser-acoustic inspection system see Blackwood, et al (1994).

OVERVIEW OF DATA

The three data sets examined are referred to as uncentered bullet data, centered bullet data, and pipe data. Both bullet data sets were measurements made on the same set of 155 mm artillery shells, 23 inches in length. The shells were measured standing upright on a concrete surface. For the uncentered bullet data the laser was focused off-center at a point 14 inches up from the shell bottom. For the centered bullet data the laser was focused near the shell center at a point 11 inches from the shell bottom. The pipe data set measurements were made on 18 inch long pipes sealed with rounded caps. A table-mounted stand allowed the pipes to be placed upright while measurements were taken. For the pipe data the laser was focused at a point 9 inches from the pipe bottom.

For several of the measured pipes and bullets there were an insufficient number of replications for analysis purposes. In addition, several pipes and bullets had duplicate measurement replications that would complicate analysis. After such unusable observations were removed, the uncentered bullet data and centered bullet data each include 76 distinct objects while the pipe data contain measurements on 75 distinct objects. The measurement process was repeated 3 times on each object resulting in 228 total observations in both the uncentered and centered bullet data, and 225 in the pipe data.

Of the three replications made on each object in all data sets, two were randomly assigned to a training set while the remaining was assigned to a test set. Distinct training and test sets allow for pseudo-jackknife validation using only the training set as well as true cross-validation by using the training set to classify the test set. Table 1 summarizes the data.

Table 1: Data summary.

Data set	Number of observations in training set	Number of observations in test set	Total number of observations in data set
uncentered bullet	152	76	228
centered bullet	152	76	228
pipe	150	75	225

Two types of variables were considered for each object, classification variables and classifier variables. The classification variables define the characteristics to which we hope to classify the objects (i.e., fill level and fill type). The classifier variables consist of the resonance amplitude information at the sampled frequencies, in effect encoding the signature information. The classifier variables make up the set of information from which features are selected for input into the *k*-nearest neighbor analysis.

Classification Variables

Each object (bullet or pipe) was filled with a particular type and amount of chemical. The variables identifying the object's fill include the fill level, the chemical simulant, and the more general agent group to which the chemical simulant belongs. There were 4 possible fill levels consisting of 100, 75, 50, and 25 percent full. Nineteen distinct chemical simulants were divided into 8 non-overlapping groups based on the agents they simulate. The agent groups are GA, GD, VX, HD, HN3, L1, QL, and DF. Within each agent group the different chemicals were intended to correspond to different levels of agent purity. For example, dimethyl sulfoxide simulates the

DISCRIMINANT ANALYSIS METHOD

K-Nearest Neighbor Discriminant Analysis

Suppose an object belongs to one of a set of m mutually exclusive populations $P_1, P_2, \ldots, P_m$ (e.g. fill levels). Given a vector of measured variables of the object (e.g. spectrum features), y, we wish to classify the object to one of the populations. Assuming all misclassifications are equally costly, the k-nearest neighbor approach is to classify the object to population P_j if for a training set of data

$$\pi_j (k_j/n_j) > \pi_i (k_i/n_i) \text{ for all } i \neq j$$

where k_i is the number of k-nearest neighbors to y that are members of P_i, n_i is the number of training set members that belong toP_i, and π_i is the prior probability of membership in P_i. The choice of the number of neighbors considered, k, is typically small. For the following applications k was chosen to be 1 or 3.

Determining proximity to y is done using a Mahalanobis distance metric. The Mahalanobis distance between y from an unknown population and x_i, a vector of known population membership P_i, is

$$D(y,x_i) = [(y - x)' V_i^{-1} (y - x)]^{½}$$

where V_i is the covariance matrix of P_i. The covariance matrix is generally unknown and may be approximated by one of several methods including

- a pooled covariance matrix
- a diagonal matrix of the pooled covariance matrix,
- a covariance matrix within population i
- a diagonal of the covariance matrix within population i
- an identity matrix

For purposes of this analysis, a pooled covariance matrix or an identity matrix was typically used. Appendix 1 outlines the relationship between k-nearest neighbor discriminant analysis and traditional discriminant analysis.

Measures of Success

Both fill level and agent group classification results were examined to judge the discrimination methods' effectiveness. K-nearest neighbor discriminant analysis classifies an observation based on its proximity to members of the training set. Jackknife validation examines how well the selected variables enable the training data to classify its own members. Each training set observation is classified using the remaining training observations. For true jackknife validation of an observation, the covariance matrix used in calculating the Mahalanobis distances is estimated without using that observation, necessitating as many covariance matrix estimates as there are members in the training set. For computational convenience, pseudo-jackknife validation was used. For pseudo-jackknife validation, only one covariance matrix estimate based on all the training observations is used in calculating the Mahalanobis distances. The percentage of the training set correctly classified was calculated as a measure of success. The pseudo-jackknife validation results tend to be optimistic.

Successful classification was also examined using true cross-validation. Each test set observation was classified using the training observations. As before, the percentage of the test set correctly classified was calculated as a measure of success.

FEATURE SELECTION METHODS

Frequencies of Maximum Amplitudes

Suppose a particular peak appears in all signatures for a specific type of object, but the exact peak frequency shifts depending on the contents of the measured objects. The frequency location of this "moving peak" may be used in fill discrimination. For example, consider the fill level for a set of 155 mm bullets. Suppose a noticeably large peak occurs at approximately 2,600 Hz for all the measured bullets, with the exact peak frequency shifting between 2,500 Hz and 2,750 Hz. Now, consider the 25, 50, 75, and 100% fill level groups separately. If the peaks cluster in the range 2,225-2,575 Hz for the 100% fill level bullets, but in the 2,575-2,625 Hz for the 75% fill level bullets and so forth, then the frequency of the peak will be a feature with substantial discriminatory power. See Figure 2 for a simple illustration of this moving peak. Since there will be numerous smaller peaks in the range 2,500-2,750 Hz that are data noise, the key peak may be isolated by searching for the frequency of the maximum amplitude in that range. This is the frequency of maximum amplitude method of feature selection.

To select the moving peak frequency features, certain frequency ranges are examined for their maximum amplitudes and the frequencies associated with them. By comparing these frequency locations relative to the characteristic of interest (i.e., fill level or agent group) the importance of a particular peak is determined. Although not directly a part of the moving peak pattern, the actual maximum amplitudes can also be used in the discriminations if they are found useful.

The various criteria of range selection that were investigated consisted of the following: arbitrary contiguous spanning ranges, ranges indicated by nonparametric Kruskal-Wallis ANOVA tests, and visually selected ranges. These three methods were implemented for both fill level and agent group classification using both of the bullet data sets and the pipe data set.

Arbitrary Contiguous Spanning Ranges

The 0 to 12,000 Hz frequency range was divided into sixteen contiguous spanning frequency ranges. The first eight ranges were approximately 500 Hz wide while the last eight were approximately 1,000 Hz wide. This partitioning was motivated by the considerably higher amount of amplitude activity observed in the lower frequency ranges. Although this may not be the best method of isolating moving peaks, it should establish to some extent their shifting effects.

Recall that different sampling frequencies were used in gathering the bullet data and the pipe data. For the two types of objects this results in a different number of amplitude variables spanning the same frequency range. For the same frequency range there are approximately four times as many bullet amplitude variables as there are pipe amplitude variables.

The frequency of the maximum amplitude in each of the sixteen ranges were calculated for each resonance spectrum in each of the three data sets. For each of the three data sets, the sixteen variables were used in distinct k-nearest neighbor discriminant analyses classifying the observations separately to fill level and agent group.

Nonparametric Kruskal-Wallis ANOVA

A nonparametric approach to range selection was also pursued. To determine which ranges would be beneficial to use, a sequence of tests using Kruskal-Wallis analysis of variance was employed. This analysis looks for important frequency ranges by identifying frequencies at which there are significant differences in amplitudes between classification groups in the training data set. To the extent that it correctly identifies the range over which a peak moves, this should yield superior classification than the arbitrary contiguous spanning frequency ranges approach which could inadvertently split or merge moving peak shifting ranges.

The Kruskal-Wallis ANOVA tests are calculated as follows. Consider M groups with a variable of interest associated with each observation. Let n_m be the number of observations associated with the m^{th} group where $N = n_1 + n_2 + ... + n_M$. The variable values from all groups are pooled, arranged in increasing order, and then assigned ranks. The sum of ranks for each of the groups is found and denoted T_m for $m = 1, 2, ..., M$. To test the hypothesis that the distributions of the variable for the M groups are identical, the Kruskal-Wallis test statistic is

$$H = 12/[N(N+1)]\sum_m (T_m^2/n_m) - 3[N+1].$$

Under a true hull hypothesis, the above test statistic approximately follows a chi-squared distribution with $M-1$ degrees of freedom (Montgomery, 1991).

For example, we can examine one of our data sets to test if the distribution of each amplitude variable is identical for the four fill levels. For every amplitude variable that is found significant at the 0.05 level, the fill level with the largest mean amplitude is recorded. A frequency range has substantial discriminatory power if it records a relatively long series of significant amplitude variables that in turn clusters by fill level. That is, an informative range would be one that could be further divided into smaller distinct subranges, each of which is associated with fill level 25, 50, 75, or 100 percent full (i.e. the largest mean value for every amplitude variable in each subrange is 25, 50, 75, and 100% fill level respectively). In such a range, the shifting pattern of the moving peak that is associated with the various fill levels is clearly visible and the frequency location of the peak can be used to discriminate between fill levels.

The above procedure results in considerable data reduction. Suppose that the Kruskal-Wallis ANOVA for the uncentered bullet data identified four significant ranges for discriminating between fill level. After determining the maximum amplitudes and their associated frequencies in each of those subranges, each spectrum can now be characterized by 8 parameters: 4 amplitudes and their 4 frequencies, rather than the original 1,180 amplitude variables. A *k*-nearest neighbor discriminant analysis would then be implemented based on those variables.

Visually Selected Frequency Ranges

The Kruskal-Wallis analysis may not successfully identify appropriate frequency ranges. It could cut off the extreme tail ends of the ranges which may not show up as significant but are still important in discrimination. Or it may find all frequency variables significant. Therefore it may be beneficial to examine representative acoustic spectra and see if important frequency ranges can be visually selected.

The visually selected ranges were the same for the two bullet data sets and somewhat different for the pipe data. They are given in Table 3.

Table 3: Visually selected ranges.

Bullet data :		**Pipe data:**	
Frequency range (Hz)	**Variable numbers**	**Frequency range (Hz)**	**Variable numbers**
1,500-2,000	150-200	750-1,750	18-43
2,000-2,800	200-280	1,750-2,500	44-61
2,800-3,200	280-320	2,500-3,250	62-80
3,200-4,000	320-400	3,250-4,250	80-105
5,500-6,500	550-650	5,250-6,250	129-154
7,000-8,200	700-820		

K-nearest neighbor discriminant analyses were employed using the frequencies of the maximum amplitudes within those ranges.

Pattern Matching Methods

The frequency of the maximum amplitude approach deals only with major amplitudes in certain frequency ranges of an object's resonance spectrum. It may be that even secondary resonance peaks are characteristic to the object's fill qualities. In that case, spectra could be compared by matching more detailed amplitude patterns. What is desired is a sensitive direct one-to-one matching between signatures over a reduced frequency range, aiming to detect more subtle differences than the frequency of maximum amplitude approach. Therefore, the only data reduction implemented is the reduction of the frequency range over which signatures are compared. Several approaches to pattern matching were pursued including the following: pure amplitude matching, standardized amplitude matching, and binary peak vector matching. For all three methods, the reduced frequency ranges considered were determined by observed resonance activity and limitations of computing resources. The ranges chosen were also adjusted in response to empirical results.

Pure Amplitude Matching

One simple way to compare objects' acoustic patterns is to examine their resonance amplitudes over a range of frequencies in which dominant peaks are occurring. By doing so, the acoustic spectra are assessed for similarity both in pattern of peak appearance and for the comparative magnitude of those peaks.

The selection of the frequency range was based on overall level of peak activity. Subject to computing restrictions, preliminary examination indicated that the best frequency range to use was 2,000-3,500 Hz.

Standardized Amplitude Matching

Examination of individual spectra indicates that some acoustic signatures' amplitudes differ from others by very large orders of magnitude. These differences appear to be unrelated to classification groups. Since magnitudes of peaks would affect the discriminant analyses using pure amplitude matching, analyses were also performed using standardized amplitude matching. Two types of standardization were investigated. For the first standardization, each amplitude in a signature from the frequency range 2,000-3,500 Hz was divided by the maximum amplitude for the signature in the frequency range 2,000-5,500. For the second standardization, each amplitude in the frequency range 2,000-3,500 Hz was divided by the mean amplitude in the frequency range 2,000-5,500. The slightly larger frequency range of 2,000-5,500 Hz was used for the selection of the mean and maximum amplitudes to obtain a better measure of average amplitude.

Binary Peak Vector Matching

The previous two pattern matching approaches involve comparing both peak locations and amplitudes. It is possible, however, to compare signatures based only on the location of the peaks. To do so, each signature was reduced to a binary vector indicating the pattern of peaks by examining each amplitude variable in sequence. If an amplitude was larger than both of its neighbors it was considered a peak and recorded as a "1". Otherwise, it was not considered a peak and recorded as a "0". K-nearest neighbor analyses were run based on selected frequency ranges of the binary peak vectors.

FILL LEVEL CLASSIFICATION RESULTS

Uncentered bullet data

Table 4 lists the fill level classification results for the uncentered bullet data. Typically only the frequency variables and not the actual amplitudes were found useful. The exceptions are indicated below. The Kruskal-Wallis ANOVA tests on the uncentered bullet data set identified four significant frequency ranges including 2,420-2,760 Hz, 3,510-3,830 Hz, 4,960-5,050 Hz, and 6,870-7,210 Hz. Both the frequencies of the maximum amplitudes and the actual maximum amplitudes from each range were found important in the discriminant analysis. For the visually selected ranges, only the frequency of the maximum amplitude in each of the 6 ranges given previously in Table 3 were found important in the discriminant analysis. The binary peak vector matching discriminant analysis was run on the frequency range 1,500-3,000 Hz (approximately 150 amplitude variables). For reference, note that a random allocation of observations to fill level would result in about 25% correct classification.

Table 4: Fill level classification results for uncentered bullet data.

Spectrum features	Percent of training set correctly classified	Percent of test set correctly classified
Frequencies of maximum amplitudes		
• Arbitrary contiguous spanning ranges	58	45
• Kruskal-Wallis ANOVA	85	71
•Visually selected frequency ranges	68	53
Pattern matching		
• Pure amplitude matching	30	38
• Standardized amplitude matching		
• By maximum amplitude	33	45
• By mean amplitude	35	34
• Binary peak vector matching	32	42

Centered bullet data

Table 5 lists the fill level classification results for the centered bullet data. For all of the frequency of maximum amplitude methods, only the frequencies and not the actual amplitudes were used. The Kruskal-Wallis ANOVA tests on the centered bullet data set identified four significant frequency ranges including 2,100-2,800 Hz, 3,490-3,830 Hz, and 5,620-6,600 Hz that exhibited the moving peak pattern. Two other interesting frequency ranges were 4,880-4,930 Hz and 7,500-7,560 Hz. The former was associated exclusively with fill level 100 (i.e. the mean value for 100% fill level was always largest) and the latter with fill level 50. The binary peak vector matching discriminant analysis was run on the frequency range 1,500-3,000 Hz (approximately 150 amplitude variables).

Table 5: Fill level classification results for centered bullet data.

Spectrum features	Percent of training set correctly classified	Percent of test set correctly classified
Frequencies of maximum amplitudes		
• Arbitrary contiguous spanning ranges	69	45
• Kruskal-Wallis ANOVA	79	79
• Visually selected frequency ranges	84	86
Pattern matching		
• Pure amplitude matching	62	64
• Standardized amplitude matching		
• By maximum amplitude	66	63
• By mean amplitude	63	63
• Binary peak vector matching	36	34

Pipe data

Table 6 lists the fill level classification results for the pipe data. For all of the frequency of maximum amplitude methods, only the frequencies and not the amplitudes were used. The Kruskal-Wallis ANOVA tests on the pipe training set indicated that at the .05 level virtually all of the amplitude variables were significant, thus were unsuccessful in isolating important frequency ranges. However, several ranges were distinct in that they did exhibited the moving peak pattern associated with different fill levels. Three ranges so identified were 1,139-1,667 Hz, 2,603-3,782 Hz, and 5,165-5,816 Hz. An additional range was considered, 5,856-6,507 Hz, that was associated almost exclusively with the 100 fill level. The binary peak vector matching discriminant analysis was run on the frequency range 1,000-4,000 Hz (approximately 74 amplitude variables).

Table 6: Fill level classification results for pipe data.

Spectrum features	Percent of training set correctly classified	Percent of test set correctly classified
Frequencies of maximum amplitudes		
• Arbitrary contiguous spanning ranges	90	95
• Kruskal-Wallis ANOVA	87	92
• Visually selected frequency ranges	92	92
Pattern matching		
• Pure amplitude matching	80	88
• Standardized amplitude matching		
• By maximum amplitude	73	78
• By mean amplitude	74	80
• Binary peak vector matching	80	84

Robustness

Ideally the fill level classification method will be robust in that classification will be good in situations that may be less than optimal. One such situation occurs if there are differences in the point of measurement among observations. If we obtain good classification results in such a situation it speaks well for the robustness of our method.

The uncentered bullet training data was used to classify the centered bullet data. (The training and test centered data sets were combined.) The goal was to classify the observations by fill level using some subset of the frequency subranges previously found important by the Kruskal-Wallis ANOVA tests on the *uncentered* training set (2,420-2,760 Hz, 3,510-3,830 Hz, 4,960-5,050 Hz, 6,870-7,210 Hz). Using all 4 of the frequency ranges led to a successful classification rate of 78%. Better results were found using frequencies of maximum amplitudes in the first and third ranges only, yielding a successful classification rate of 80%. These classification results are better than the results for the uncentered bullet data. This certainly attests to a degree of robustness in the frequency of maximum amplitude approach to fill level classification.

AGENT GROUP CLASSIFICATION RESULTS

The Kruskal-Wallis ANOVA tests were unsuccessful in indicating important frequency ranges for discrimination between agent group. Instead they found many very short runs of frequencies with little group clustering, leading to no clear choice of significant frequency ranges. Thus, no results for this method are presented.

Uncentered bullet data

Table 7 lists the agent group classification results for the uncentered bullet data. For both of the frequency of maximum amplitude methods applied, only the frequencies were used. The binary peak vector matching routine was run on the frequency range 1,500-3,000 Hz (approximately 150 amplitude variables). A random allocation of observations to agent group would result in about 12.5% correct classification.

Table 7: Agent group classification results for uncentered bullet data.

Spectrum Features	Percent of training set correctly classified	Percent of test set correctly classified
Frequencies of maximum amplitudes		
• Arbitrary contiguous spanning ranges	35	36
• Kruskal-Wallis ANOVA	-	-
• Visually selected frequency ranges	41	54
Pattern matching		
• Pure amplitude matching	61	63
• Standardized amplitude matching		
• By maximum amplitude	57	62
• By mean amplitude	69	72
• Binary peak vector matching	76	86

Centered bullet data

Table 8 lists the agent group classification results for the centered bullet data. For both of the frequency of maximum amplitude methods applied, only the frequencies were used. The binary peak vector matching routine was run on the frequency range 1,500-3,000 Hz (approximately 150 amplitude variables).

Table 8: Agent group classification results for centered bullet data.

Spectrum Features	Percent of training set correctly classified	Percent of test set correctly classified
Frequencies of maximum amplitudes		
• Arbitrary contiguous spanning ranges	18	25
• Kruskal-Wallis ANOVA	-	-
• Visually selected frequency ranges	45	36
Pattern matching		
• Pure amplitude matching	20	26
• Standardized amplitude matching		
• By maximum amplitude	21	22
• By mean amplitude	18	29
• Binary peak vector matching	82	74

Pipe data

Table 9 lists agent group classification results for the pipe data. For both of the frequency of maximum amplitude methods applied, only the frequencies were used. The binary peak vector matching routine was run on the frequency range 1,000-4,000 Hz.

Table 9: Agent group classification results for pipe data.

Spectrum features	Percent of training set correctly classified	Percent of test set correctly classified
Frequencies of maximum amplitudes		
• Arbitrary contiguous spanning ranges	31	33
• Kruskal-Wallis ANOVA	-	-
• Visually selected frequency ranges	38	35
Pattern matching		
• Pure amplitude matching	38	41
• Standardized amplitude matching		
• By maximum amplitude	31	32
• By mean amplitude	34	32
• Binary peak vector matching	28	56

DISCUSSION

The successful classification rate varied depending on the classification variable investigated, the feature selection method employed, and the type of object examined. Because there were several factors differing between the data sets, there may be several factors contributing to the varying classification success rates. For example, the differing sampling frequencies between bullet and pipe measurements, their substantial difference in design and construction, and the focus location of the measurement laser may all contribute to classification disparities.

For fill level classification of the uncentered bullet data, only the Kruskal-Wallis ANOVA approach to range selection for the frequencies of maximum amplitudes successfully classified more than 70% of the test set. However, for agent group classification of the uncentered data, both the binary peak vector and mean standardized amplitude approaches to pattern matching successfully classified more than 70% of the test set. For fill level classification of the centered bullet data, both the Kruskal-Wallis ANOVA and visual selection approaches of frequency range selection successfully classified more than 70% of the test set. In contrast, for agent group classification of the centered data, only the binary peak vector method of pattern matching successfully classified more than 70% of the test data set. For fill level classification of the pipe data, every method successfully classified more than 90% of the test data. For agent group classification of the pipe data, none of the methods successfully classified the test set.

In general, the uncentered bullet data had superior agent group classification results compared to the centered bullet data. In contrast, the centered bullet data had superior fill level classification compared to the uncentered data. These classification differences could be due to the focal location of the measurement laser. Focusing the laser near the top of the inspected objects may emphasize agent group differences while muting fill level differences. Similarly, focusing the laser near the top of the inspected objects may emphasize fill level differences while muting agent group differences. If so, this is consistent with the pipe data's classification results since the pipes were measured with the laser centered. This is particularly true if it is reasonable to suppose the muting and emphasizing effects of the laser focal location may be exaggerated by the simple pipe construction.

The wide pipe sampling frequency of 40.67 Hz may also explain the pipe data's disparity between fill level and agent group classification. The moving peak patterns for the bullets suggest that peak shifts due to fill level difference were well over 40 Hz in magnitude. Thus, sampling at 40 Hz rather than 10 Hz should not affect fill level discrimination for the pipe data. In contrast, the binary peak vector matching routine that worked well for the bullet data makes use of peak patterns that are much more subtle, suggesting that differences in peak patterns due to agent group occur over frequency ranges smaller than 40 Hz. If so, agent group classification with 40 Hz data would not be expected to produce particularly good results.

For classification of the bullet data to fill level, the most successful method was a *k*-nearest neighbor discriminant analysis based on the frequency of the maximum amplitude in several important ranges of frequencies. For classification of the bullet data to agent group, the most successful method was a k-nearest neighbor discriminant analysis based on matching signature patterns within a specific frequency range. Since the two feature selection methods emphasize different signature attributes, apparently agent group and fill level affect different signature attributes. In particular, agent group membership may determine *where* peaks occur, while fill level may determines the relative amplitude of those peaks. Members of an agent group may always have a particular pattern of peaks that are present regardless of fill level. However, fill level may influence the amplitudes of those peaks particular to the agent group. See Figure 3 for an illustration of this possibility. In the figure, the frequency of the maximum amplitude is the same for common fill levels, but different for common agent groups. At the same time, the patterns of peak occurrence (frequencies at which they occur) are the same for common agent groups but different for common fill levels.

Although several methods show promise in agent group and fill level discrimination, a higher level of successful classification would be preferable before a broad implementation of the methods was pursued.

FUTURE STUDY

Much of the analysis done to date was exploratory in nature. Verification of the ideas suggested by the previous discussion would be useful. In particular, new acoustical signature bullet data gathered with the 40.67 Hz sampling frequency would be informative to analyze to determine if the wider sampling frequency leads to lower successful agent classification rates. For the same reason, acoustical signature pipe data with a sampling frequency of 10.17 Hz would be beneficial. It would also be informative to make pipe measurements with the measurement laser uncentered to see if this improves agent group classification.

Optimization of the more useful feature selection methods would also be beneficial. While considerable care was taken in the range selection for the frequencies of maximum amplitudes method, the range selection for the pattern matching methods were less rigorously pursued. Effort could be made to optimize these range selections. In addition, the binary peak vector matching method could be fine tuned. The current routine compares only exact peak matches. A way of assessing proximity of peaks could be implemented by weighting amplitude variables for closeness. This might result in more complete comparisons and better classification.

Some preliminary investigation was made into the possibility of a classification tree approach to classifying fill characteristics (Blackwood, et al, 1994). If a classification tree approach is pursued, it might be of interest to see how well we can classify observations to chemical simulant type *within* agent group. The particular chemical simulant types and their agent groups are listed in Table 2. Table 10 lists the percentages of each agent group from the centered bullet data set correctly classified to chemical simulant type using binary peak vector matching over the 1,500-3,000 Hz range.

Table 10: Chemical simulant type classification results for centered bullet data.

Within agent group	Percent of training set correctly classified	Percent of test set correctly classified
GA	69	75
GD	87	87
HD	71	75
HN3	91	81
L1	81	87
QL	96	92
VX	75	100

These results compare to successful classification rates of 61% and 51% when classified by chemical simulant type without first separating into agent groups. These classification rates speak favorably for a tree approach. Further investigation of this possibility would be beneficial.

Appendix 1: K-nearest Neighbor and Traditional Discriminant Analysis.

The following discussion assumes that an object is to be classified to one of two populations. Extensions to more than two populations are straightforward.

Assume y is a $d \times 1$ vector of measures from an object belonging to one of two populations, P_1 or P_2. Given y we wish to classify the object into the correct population. The approach often pursued is discriminant analysis. As an initial step, a decision rule is found. If R is the sample space of y, partition R into R_1 and R_2 such that $R_l \cup R_2 = R$ and $R_l \cap R_2 = \varnothing$. The classification decision is to assign the object to P_1 if $y \in R_1$. Thus our goal is some optimal partitioning of R.

As an unavoidable consequence of any such classification scheme, there will be incorrect decisions made and costs associated with them. Let $c(i|j)$ be the cost of classifying to the i^{th} population an object that is in fact a member of population j. For simplicity assume that misclassification of a P_1 member to P_2 is as equally costly as a misclassification of a P_2 member to P_1. In terms of cost functions, $c(i|j) = c(j|i)$. Denote the probability that a randomly selected object belongs to population i as π_i. If we define risk as the expected loss of a classification decision, we can denote the risk of a decision rule as

$$r(R_1, R_2) = c(2|1)\, pr(2|1, R_1, R_2)\, \pi_1 + c(1|2)\, pr(1|2, R_1, R_2)\, \pi_2$$

where $pr(i|j, R_1, R_2)$ is the probability of classifying an observation from P_i into P_j using the partition $R = R_l \cup R_2$ and $R_1 \cap R_2 = \varnothing$.

The traditional approach to discriminant analysis is to assume the form of the densities $f_i(y)$ and minimize the risk function with respect to the partitioning $R = R_l \cup R_2$. The decision rule so found is to classify the object to P_1 if $y \in R_1$ where

$$R_1 = \{y; f_1(y)\, c(2|1)\, \pi_1 \geq f_2(y)\, c(1|2)\, \pi_2\}$$

otherwise classify the object to P_2. R_1 can be expressed as

$$R_1 = \{y; f_1(y)/f_2(y) \geq [c(1|2)\, \pi_2]/[c(2|1)\, \pi_1]\} \tag{1}$$

if $f_2(y) \neq 0$ and costs are equal. This decision rule is based on a comparison of the density ratio to the prior ratio.

An important restriction of traditional discriminant analysis is that it requires knowledge of the populations' density functions. It is often unrealistic to expect such exacting knowledge about the populations. In addition, if an incorrect distribution is assumed classification performance may suffer.

K-nearest neighbor discriminant analysis does not require knowledge of the densities. The theory behind k-nearest neighbor discriminant analysis is identical to that of the traditional method. The decision rule of interest remains to classify y to P_1 if $y \in R_1$ where R_1 is defined as in equation (1). However, now the densities are unknown and must be estimated. An intuitive estimate of $f_i(y)/[f_2(y)+f_2(y)]$ is k_i/n_i, the number of k-nearest neighbors of y that are members of P_i divided by the total number of members of P_i. We will use the previous ratio to estimate the ratio of densities. That is, the ratio

$$(k_1/n_1)/(k_2/n_2) \text{ is used as an estimator of } f_1(y)/f_2(y).$$

Therefore our classification rule is to classify the object to P_1 if $y \in R_1$ where

$$R_1 = \{y;\ (k_1/n_1)/(k_2/n_2) \geq \pi_2/\pi_1\}$$

REFERENCES

1. L. G. Blackwood, Personal communications, Summer 1994.

2. L. G. Blackwood, J. G. Rodriguez, D. M. Tow, "Discrimination of Munition Fill Types by K-Nearest Neighbor Classification Tree Analysis of Acoustic Resonance Signatures," unpublished internal document, Idaho National Engineering Laboratory.

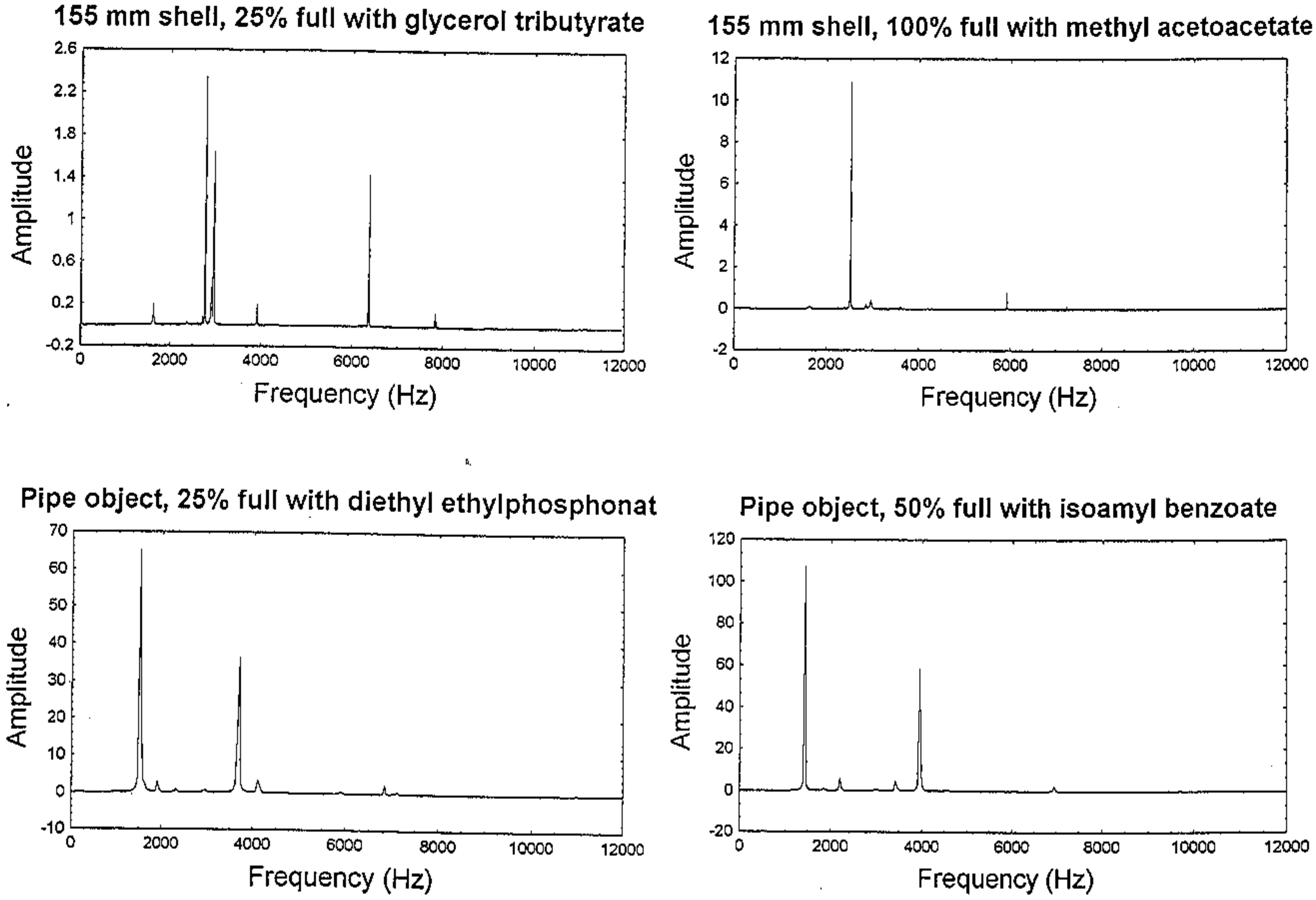
3. R. J. Boik, Personal communications, Spring and Fall 1994.

4. D. C. Montgomery, *Design and Analysis of Experiments, Third Edition*. New York: Wiley (1991).

5. SAS Institute Inc., SAS/STAT *User's Guide, Version 6, Fourth Edition, Volume 1.* PROC DISCRIM. SAS Institute Inc., Cary, NC (1989).

6. G. A. F. Seber, *Multivariate Observations.* New York: Wiley (1984).

Figure 1: Sample acoustic signatures.



155 mm shell, 25% full with glycerol tributyrate

155 mm shell, 100% full with methyl acetoacetate

Pipe object, 25% full with diethyl ethylphosphonat

Pipe object, 50% full with isoamyl benzoate

Figure 2: Moving peak illustration.

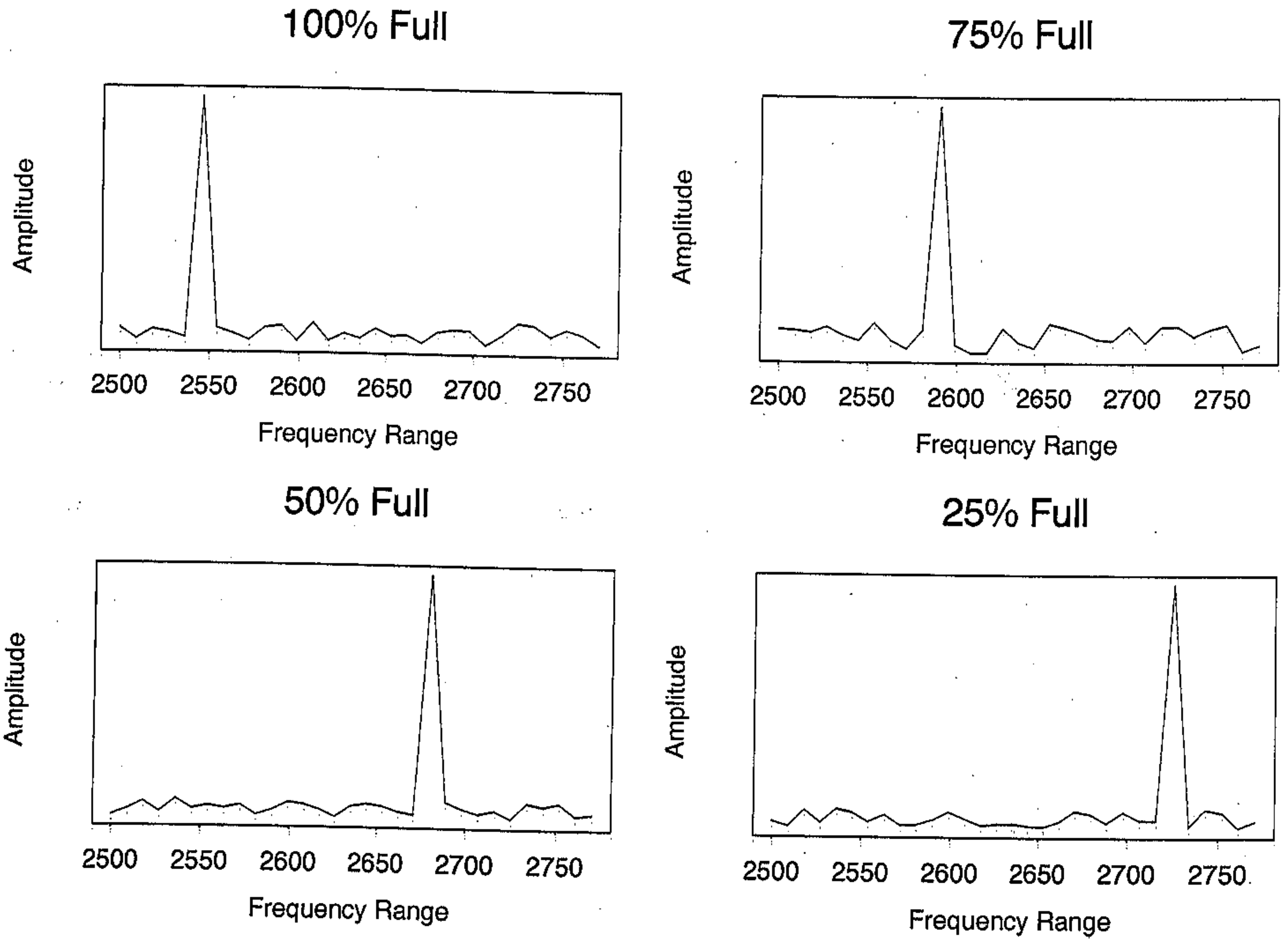
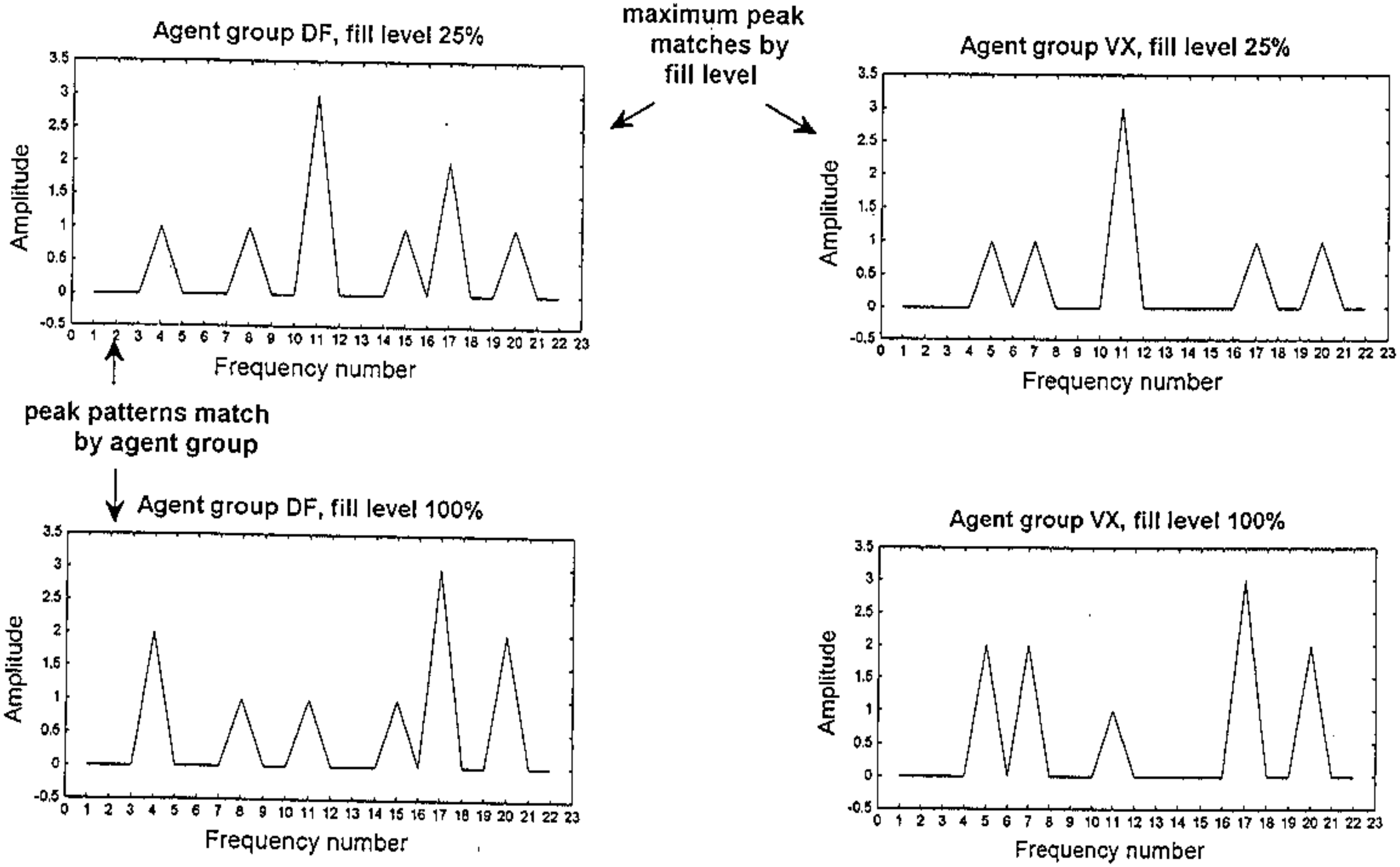


Figure 3: Possible pattern differences between agent group and fill level.


Agent group DF, fill level 25%
maximum peak matches by fill level
Agent group VX, fill level 25%
peak patterns match by agent group
Agent group DF, fill level 100%
Agent group VX, fill level 100%
Amplitude
Frequency number