ANALYSIS OF THE ALLELOPATHIC POTENTIAL OF RICE USING K-MEANS CLUSTERING OF HPLC CHROMATOGRAMS

Edward E. Gbur

John D. Mattice

Robert H. Dilday

ANALYSIS OF THE ALLELOPATHIC POTENTIAL OF RICE USING K-MEANS CLUSTERING OF HPLC CHROMATOGRAMS

Edward E. Gbur[1], John D. Mattice[1], and Robert H. Dilday[2]

[1]University of Arkansas
[2]USDA-ARS National Rice Research Center

Abstract

Allelopathy is the ability of an organism to affect the growth of another organism through the introduction of chemical compounds into the environment. Several researchers have reported rice inhibition of the growth of weed species such as barnyard grass and ducksalad. The objective of this study was to relate patterns found in HPLC chromatograms for leaf extracts of different rice accessions to their weed control activity. K-means cluster analysis was performed on 20 peak heights from chromatograms from 40 rice accessions. The resulting clusters corresponded to observed behavior of the accessions reported in other sources. Stepwise discriminant analysis was used to determine if the number of peaks needed to separate accession types could be reduced.

Keywords: Discriminant analysis; Principal components

1. Introduction

Several researchers have observed that some rice varieties have the ability to inhibit the growth of certain weed species. Early reports of this phenomenon include those of Dilday et al. (1989) for inhibition of ducksalad and Navarez and Olofsdotter (1996) for barnyard grass. If this inhibitory trait can be incorporated into agronomically useful rice varieties, then growers would have options which would require fewer herbicide applications and/or reduced application rates for weed control.

There are two mechanisms for interaction among organisms in a plant community, allelopathy and competition (Putnam and Tang, 1986). Allelopathy occurs when one plant introduces chemicals into the environment which affect another plant. The effect may be negative (interference or suppression) or positive (stimulation or attraction). Competition occurs when, given a limited resource used by both plants, one plant is more efficient in its use of the resource than the other, to the disadvantage of the latter. Proof of allelopathy requires the identification of the chemical compounds produced by the plant and the demonstration that the presence of these compounds affect the other plant in the absence of the producer plant.

In the case of rice and weed species, the effect observed by researchers may be the result of allelopathy or competition or both. The overall goal of the study reported here is to find a non-destructive method to predict whether or not a young rice plant has the potential to inhibit weed growth. Ideally the method would require a minimum amount of space, be relatively

inexpensive, could be completed in a relatively short period of time, and could be carried out without resorting to field trials.

The use of high performance liquid chromatography (HPLC) as the basis for such a method was considered. If allelopathy is a factor in the inhibitory behavior, then the chromatographs of rice accessions with and without inhibitory ability should be different. The differences should ultimately lead to the identification of the compounds involved in the allelopathy. More specifically, the study attempted to determine if HPLC can be used to separate rice accessions possessing inhibitory properties from those which do not. If so, then other chemical techniques can be used to identify the compounds associated with each peak.

2. Experiment and Data

The experiment to evaluate the use of HPLC involved 40 rice accessions. Because of space limitations, the experiment was run in four groups of accessions over time, with several key accessions placed in more than one group. For purposes of our analysis, the same accessions in different groups were treated as different accessions. Thus, the data were labeled as if they came from 48 accessions.

Rice plants were grown in the greenhouse for 10 days, 10 plants per pot, and three replications per accession. After 10 days, samples of fresh leaf tissue were obtained and placed in a solution of methanol and deionized water. The solution was analyzed by HPLC. Additional details can be found in Mattice et al. (2001).

Examination of the chromatograms yielded 20 peaks whose heights were included in the statistical analysis. Chromatograms from some accessions contained all 20 peaks while others did not. Missing peaks were assigned a height of zero. Peaks were identified by number rather than by time to facilitate uniform labeling in future experiments. The chemical compounds associated with each peak had not been identified. An example of chromatograms which illustrate the differences among accessions is shown in Figure 1. In other studies, PI312777 has shown weed control activity and Rexmont has not.

Of the 40 accessions, 19 are listed in the USDA-ARS Germplasm Resource Information Network (GRIN). Of these, 10 have been reported to have shown weed control activity and 9 have not. Bioassay data from other experiments conducted by one of the authors have shown that the remaining 21 accessions were not successful in inhibiting the growth of barnyard grass.

3. Statistical Analysis

Since one of the objectives of the study was to develop a methodology which can be used for screening in the early stages of a breeding program where independent verification of weed control activity would be impossible, the activity information on the 40 accessions in the experiment was used only to check the success of the methodology and not as an integral part of the analysis.

Our emphasis was on an exploratory analysis which described the relevant features of the data. The basic steps in our analysis were as follows:

(1) Use k-means clustering on the 20-dimensional vectors of peak heights.
(2) Compare cluster membership to the known weed control activity information from GRIN and the bioassay data to ascertain if clustering successfully separated the accessions.
(3) Assume that the data are samples from populations defined by the clusters from (1). Use stepwise discriminant analysis to determine if a smaller number of peaks can separate the populations.
(4) Rerun k-means clustering using a subset of peaks based on the results from (3) and compare the results to the known activity information.

It is common in multivariate analysis to begin by using principal components to reduce the dimensionality of the data. When principal components were calculated for our data, it was found that the first three principal components accounted for 60.4% of the variability in the data. Nine principal components were needed to explain approximately 90% of the variation and 12 were needed to explain 95% of the variation. Unfortunately, the principal components were not interpretable in light of the objective of identifying specific compounds (individual peaks) for further study by other chemical techniques. Hence, the original data were used in all subsequent analyses.

Krzanowski and Marriott (1995) provide an overview of cluster analysis, including optimization methods. K-means clustering was introduced by MacQueen (1967) as an example of a more general k-means problem. It is a non-hierarchical method in which the final set of clusters minimizes the sum of squared Euclidean distances of observations within clusters from their respective cluster centroids (mean vectors). Clustering is accomplished through an iterative algorithm which can be described roughly as follows.

(1) For a specified value of k, randomly select k observations as cluster centroids.
(2) Sequentially assign each remaining observation to the cluster whose centroid is closest to it in terms of Euclidean distance. Recalculate the centroid of the cluster to which the observation was added.
(3) When all observations have been assigned to a cluster, go back through the data and check each observation to determine if it is closer to another cluster centroid than to that of the cluster to which it has been assigned. If it is closer to another centroid, reassign the observation and recalculate the affected centroids.
(4) Repeat (3) until the cluster centroids change by less than a prespecified amount or no change occurs.

The advantage of k-means clustering over a hierarchical method is that observations can be reassigned to different clusters during the clustering process. The disadvantage of having to specify the number of clusters k a priori can be overcome by trying a series of values of k and observing changes in the structure of the resulting sets of clusters. Methods have been developed which attempt to select the "best" value of k but were not used in our analysis.

4. Analysis of the Data

The data were subjected to k-means clustering for k = 2, ..., 7 clusters using SAS's PROC FASTCLUS. For k = 2, "active" accessions (those with known weed inhibitory activity) were found in both clusters. For k = 3, ..., 6, all active accessions were in a cluster by themselves. For k = 7, the active accession cluster was divided into two clusters, each containing only observations from active accessions. For k = 3, a plot of the first three canonical variates is shown in Figure 2 along with a plot of the first two canonical variates. Figure 3 shows plots of the first two canonical variates for k = 4, 5, 6, and 7. The figures show that the active accession cluster is clearly separated from the other clusters. Moreover, regardless of the value of k > 2, the other clusters appear more as a division of a data cloud rather than as well separated clusters. In general, all replications of an accession were placed in the same cluster.

Since k = 3 successfully separated the active accessions from the nonactive accessions and, for this study, we were not interested in the structure of the nonactive accessions, the primary focus of the remaining analyses was on k = 3. Except as noted below, the same set of random seeds (initial set of observations for centroids) were used.

Since k-means clustering is an iterative algorithm for solving an optimization problem, a poor choice of seeds can lead to a local rather than global optimum. For k = 3, the clustering procedure was run with six distinct sets of seeds. For five of these sets, the final clusters were essentially the same with only one or two observations from nonactive accessions switching clusters. The remaining set of seeds gave two clusters containing five and 46 observations from nonactive accessions, respectively. The third cluster was composed of all 36 observations from active accessions and 57 observations from nonactive accessions. The value of the sum of squared distances at convergence was 1278 for this set of seeds as compared to 1160 for the other five sets of seeds, indicating a local rather than global minimum.

The next step was to attempt to identify statistically important peaks used to define the k = 3 clusters. If successful, this will reduce the amount of laboratory work required to identify the chemical compounds potentially associated with the allelopathy. Preliminary plots of peak height against accession number for each peak number indicated differences among clusters for some peaks. Figure 4 illustrates the range of patterns. For peaks 25 and 36, the cluster of active accessions has peak heights which are different from the nonactive accession clusters. For peak 25, the peaks are larger, indicating more of the compound, while for peak 36, the peaks are much smaller than those of the nonactive accessions. Peak 21 appears to distinguish one of the nonactive accession clusters from the remaining clusters. In contrast, peak 20 does not appear to provide any information for cluster separation. While these plots indicate the feasibility of reducing the number of peaks needed, they represent 1-dimensional projections of 20-dimensional data and as a result, may be misleading.

To identify statistically important peaks, we assumed that the clusters represented random samples from known populations defined by the clusters. Stepwise discriminant analysis was carried out using SAS's PROC STEPDISC with $\alpha_{enter} = 0.15$ and $\alpha_{stay} = 0.15$. A common multivariate normal distribution across populations was assumed. Ten peaks were selected. A summary of the results are given in Table 1.

The next step in the analysis was to determine if the reduced set of peaks could separate the active and non-active accessions. The data from the ten selected peaks were subjected to k-means clustering for k = 3, ..., 7. The results were generally the same as those using all 20 peaks except that the active accessions cluster contained two to four observations from nonactive accessions. For k = 6, three active accession observations were placed in one of the non-active clusters. We did not investigate further to determine if this represented a local minimum. Although the results are not as definitive as those using all 20 peaks, coupled with the discriminant analysis results, they do indicate that not all 20 compounds need to be studied further.

To determine if even fewer peaks can be used to produce clusters in which the active accessions are isolated from the non-active accessions, peaks were eliminated sequentially in the reverse order that they entered the stepwise discriminant analysis. K-means clustering with k = 3 was applied to the reduced data vectors at each step. When peak 28 was eliminated, the remaining nine peaks produced the same clusters as when it was included. When peaks 28 and 29 were eliminated, the active accession cluster contained four observations, each from a different nonactive accession. The same result was achieved when peaks 28, 29, and 23 were eliminated. Adding peak 31 to the list of eliminated peaks lead to a major rearrangement of cluster contents and the active accessions were no longer all in the same cluster. These results indicate that, at least initially, further laboratory work needed to identify and isolate chemical compounds can be concentrated on a relatively small number of peaks.

5. Conclusion

The results of this study demonstrate that it is feasible to separate rice accessions with and without weed control activity by analyzing HPLC chromatograms using standard multivariate analysis techniques. Moreover, some reduction in the number of compounds (chromatogram peaks) which need to be studied further is possible.

Bibliography

Dilday, R. H., P. Nastasi and R. J. Smith (1989). Allelopathic observation in rice (*Oryza sativa* L.) to ducksalad (*Heteranthera limosa*). *Proceedings of the Arkansas Academy of Science*. 43, 21-22.

Krzanowski, W. J. and F. H. C. Marriott (1995). *Multivariate analysis. Part 2: Classification, covariance structures and repeated measurements*. London: Edward Arnold.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley Symposium. Volume 1*. University of California Press. 281-297.

Mattice, J. D., R. H. Dilday, E. E. Gbur and B. W. Skulman (2001). Inhibition of barnyardgrass (*Echinochloa crus-galli*) growth with rice (*Oryza sativa* L.). *Agronomy Journal*, 93, Jan-Feb issue.

Navarez, D. C. and M. Olofsdotter (1996). Relay seeding technique for screening allelopathic rice (*Oryza sativa* L.). *Proceedings of the Second International Weed Control Congress, Copenhagen*. 1285-1290.

Putnam, A. R. and C.-S. Tang (1986). Allelopathy: State of the science. In *The science of allelopathy*. A. R. Putnam and C.-S. Tang, editors. 1-19.

Table 1. Results of the stepwise discriminant analysis to select statistically important chromatogram peaks.

Step	Number of peaks selected	Peak entered	Peak removed	α to enter or remove
1	1	25		<0.0001
2	2	21		<0.0001
3	3	22		<0.0001
4	4	15		<0.0001
5	5	32		<0.0001
6	6	16		<0.0001
7	7	36		0.0002
8	8	31		0.0014
9	9	23		0.0085
10	10	29		0.0038
11	9		32	0.3181
12	10	28		0.0310

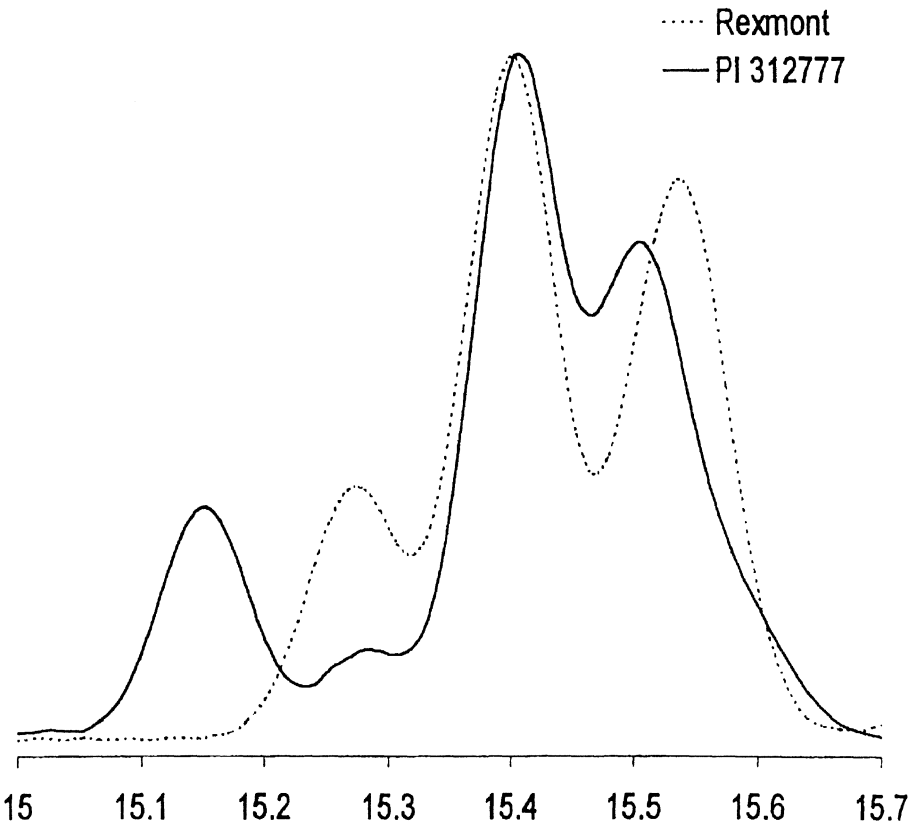

Figure 1. Portions of the HPLC chromatograms of rice leaf tissue from PI312777 and Rexmont.

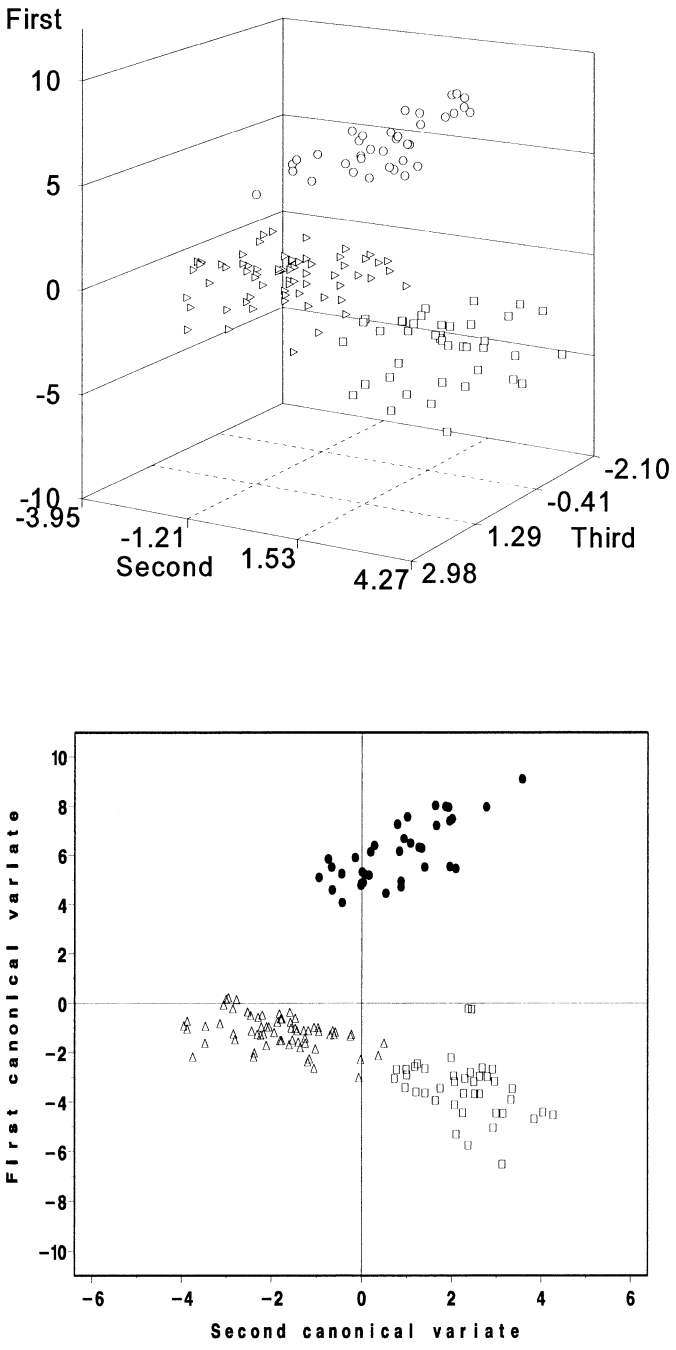

Figure 2. Plots of the canonical variates based on k-means clusters for k = 3. The observations in the active accession cluster are represented by circles.

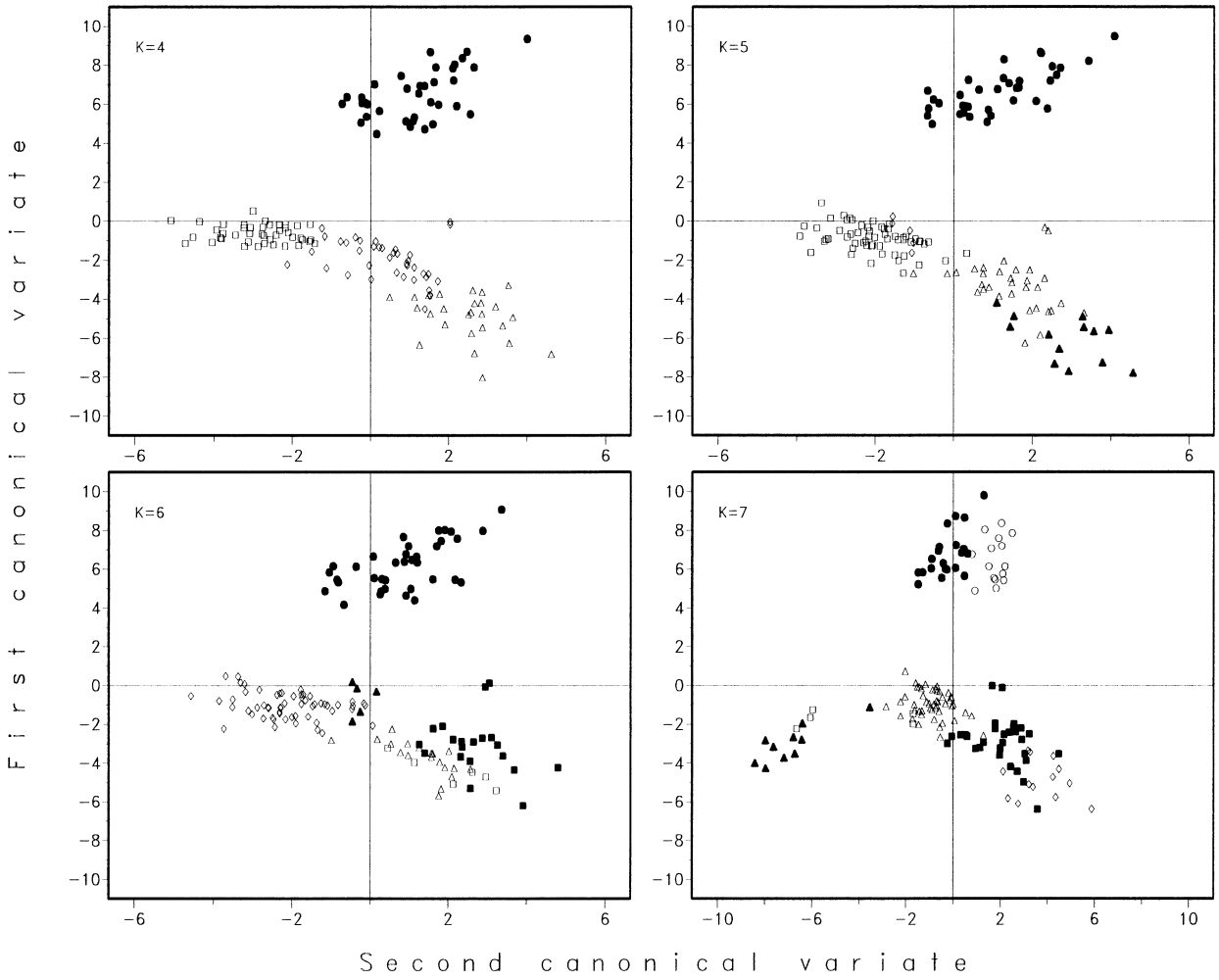

Figure 3. Plots of the first and second canonical variates based on k-means clustering for k = 4, 5, 6, and 7. The observations in the active accession cluster(s) are represented by circles having positive values of the first canonical variate.

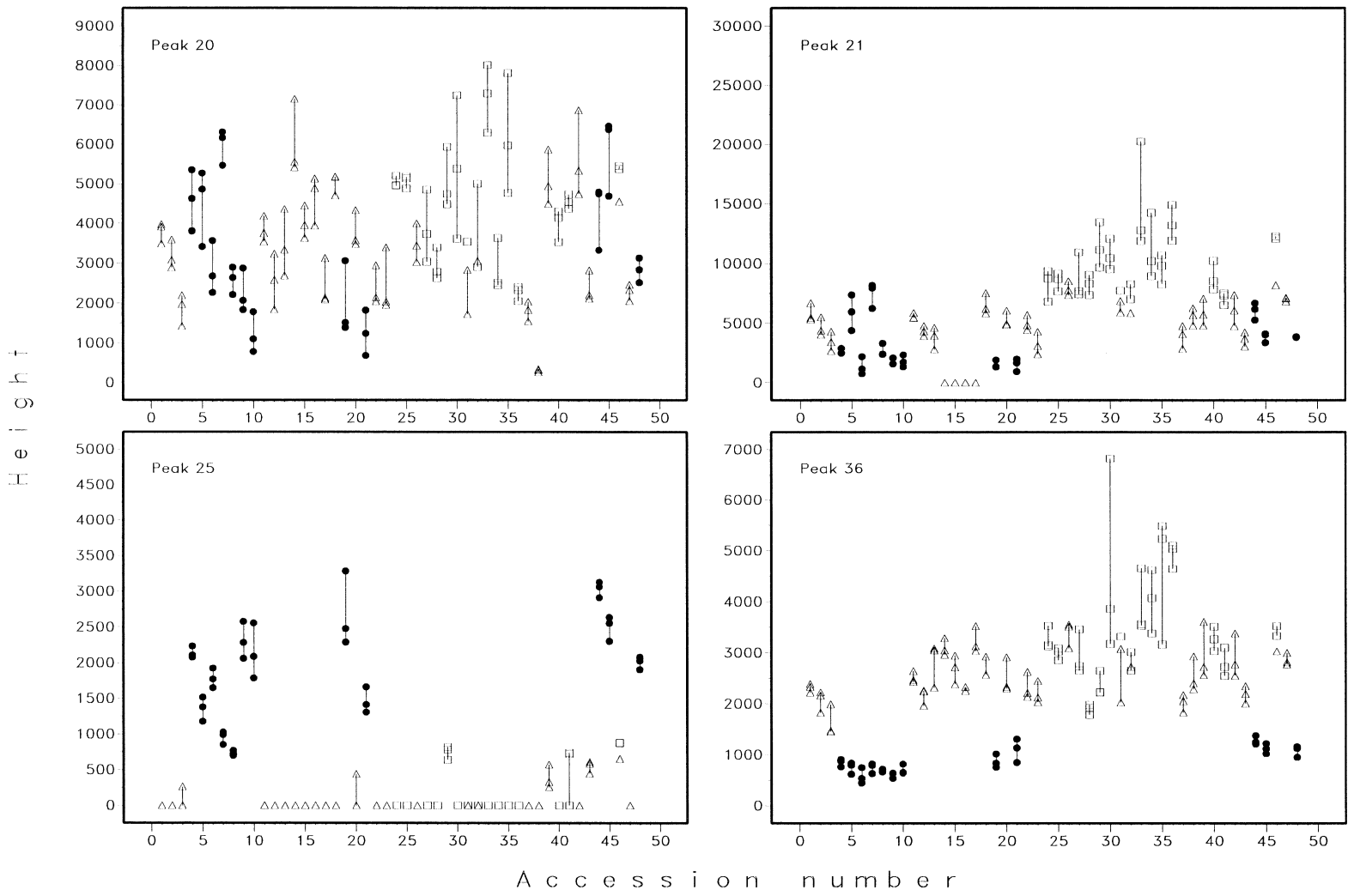

Figure 4. Plots of peak heights for selected peaks based on k-means clustering for k = 3. The observations in the active accession cluster are represented by circles.