

## **CHAPTER EIGHT**

### **TWO INTERNATIONAL MATHEMATICAL MODELING QUESTIONS WITH EXEMPLARY SOLUTIONS**

These materials have been made available to the NSF Advisory Council, NSF liaison officers, and pilot project participants by the generosity of COMAP (Dr. Sol Garfunkel).

#### **Two Snow Plow Problem**

The solid lines on the map represent paved, two-lane roads in a snow-removal district in Wicomico County. The broken lines are state highways. After a snow fall, two plow-trucks were dispatched from a garage that is about 4 miles west of each of two points (\*) marked on the map. Find an efficient way to use two trucks to sweep snow from the county roads. The trucks may use the state highways to access the county roads.

Assume that the trucks neither break down nor get stuck and that the road intersections require no special plowing techniques.

Before you read the Outstanding (Best in World) solution to the Snow Plow Problem, review the topics of Graph, Tree, Directed Graph, Cycle, Euler Circuit, and Undirected Tree, covered in standard texts of *Discrete Mathematics*.

## OUTSTANDING SOLUTIONS

### A REAL SNOW JOB

Joel E. Atkins  
Jeffrey S. Dierckman  
Kevin O'Bryant

Rose-Hulman Institute of Technology, Terre Haute, IN 47803  
Advisor: A. W. Schurle

#### SUMMARY

We were asked to design optimal routes for two plow-trucks to clear the county roads in a snow removal district of Wicomico County, MD. We felt that an optimal route would minimize the time spent driving over already-plowed roads and highways. We also felt that an optimal route would end where it had started.

Drawing analogies to graph theory, we found a subgraph, which was a tree, of the graph representation of the snow-removal district. Traversing this tree in preorder yielded an Euler circuit. This led us to a route where neither plow-truck ever drove over highways or county roads that had already been cleared. This route also enabled both plow-trucks to finish where they started. Therefore, this route satisfies our definition of an optimal route.

The model we present has several strengths. First, it produces an optimal route for the two plow-trucks. Second, it is easily adaptable to changes in the snow-removal district. Third, it is easy to compensate for changes in the relative capabilities of the plow-trucks or the number of the plow-trucks. Lastly, it provides a simple algorithm for the plow-truck drivers to follow:

#### ASSUMPTIONS

1. Plow-trucks do not break down or get stuck.
2. Neither intersections nor dead ends require special snow removal techniques.

3. Plow-trucks travel in the right lane of the road.
4. The plow-trucks enter the district on the roads directly east of the starred locations on the map.
5. Snowfall is uniform throughout the district.
6. Both plow-trucks are identical in their snow-removal capabilities.
7. Both plow-trucks travel at the same speed.
8. A plow-truck plows exactly one lane at a time.

#### JUSTIFICATION OF ASSUMPTIONS

1. The problem statement says that plow-trucks do not break down or get stuck.
2. The problem statement also says that intersections do not require special plowing techniques. It seems reasonable to include dead ends in this assumption, because there will be no traffic at the ends of roads.
3. Plow-trucks must obey traffic laws.
4. Since the plow-trucks begin 4 mi west of the starred locations on the map, off the map, we assume that the plow-trucks can reach the district in the shortest time on Dagsboro Road and Ocean City Road.
5. Because of the small size (approximately 5.6 mi by 12.9 mi) of the snow-removal district and the relatively large size of weather fronts, we assume that snowfall is uniform throughout the district.
6. Since the problem states nothing about the relative capabilities of the plow-trucks, we had to make some assumptions. We assume equality to simplify the exposition of the model; however, the model can easily compensate for differences in the capabilities of the plow-trucks.

7. In at least one state, all county roads have the same speed limit. There is nothing on the map which would indicate that this is not the case in Wicomico County.  
Therefore, assume that any plow-truck in the district will be able to travel at that maximum speed unless excessive snow makes this impossible. If there is excessive snow, it will affect all plow-trucks in the district in the same way, because we assume uniform snowfall (assumption (5) above).
8. It is unreasonable to assume that a plow-truck would plow less than one lane, because to do so would require the plow-truck to traverse each road more than twice. It would be unsafe for a plow-truck to plow more than one lane at a time, because doing this would be hazardous to oncoming traffic. We assume, therefore, that a plow-truck plows exactly one lane at a time.

#### ANALYSIS AND DESIGN OF THE MODEL

The problem is to find the most efficient way for the two plow-trucks to clear the county roads. Our model measures the efficiency by the amount of snow that the plow-trucks are able to clear in a given period of time spent in the district. This time is affected only by the speed of the plow-trucks and the amount of time spent clearing county roads that have not yet been plowed. Thus, plow-trucks driving on highways and county roads that have already been cleared would reduce efficiency. Since the model cannot affect the speed of the plow-trucks, efficiency will be maximized when the ratio of time spent clearing county roads in the district to the total time spent in the snow-removal district is maximized. Therefore, the best possible result, which this model obtains, is for both plow-trucks to spend all of their time clearing roads now previously plowed.

In keeping with the desire to spend as much time as possible clearing new roads, we decided that turning around is preferable to driving on highways or already-cleared roads. It is beneficial to end the route where we enter the district, for two reasons. Primarily, it may be necessary to redo the route if snow has continued to fall. If not, the plow-trucks will need to return to their garage 4 mi west of the starred locations on the map.

Our model requires that the snow-removal district be divided into two areas, with one plow-truck assigned to each area. To accomplish this, the roads should be divided so that each plow-truck will have the same number of miles of roads to clear. The alternative, one plow-truck finishing before the other, requires more time to clear the entire district. Furthermore, both areas should be connected. If not, the plow-truck assigned to an area would need to travel on roads that the other plow-truck would be clearing anyway.

To split the district evenly, we measured the distances of all the county road segments. We enlarged the provided map with a photocopier and scaled distances from the enlargement. Our method was to conform a pliable string to the shape of the map road segment; the length of the string when taut determined the map length of the road segment. The map length was then scaled to give the length of the road segment. Having determined the lengths of all county road segments, we then split the district into two connected areas, with one having 0.06 more miles of county roads than the other (see Figure 1), a negligible difference compared to the 125.29 total miles of roads.

The problem is now reduced to finding an optimal route for each plow-truck to cover its area of the district. To solve this problem, it was helpful to view the map in the context of graph theory. Each area was represented by a directed graph, with each intersection and dead end a vertex, and each lane an edge. Both graphs are connected, since the roads in the snow district

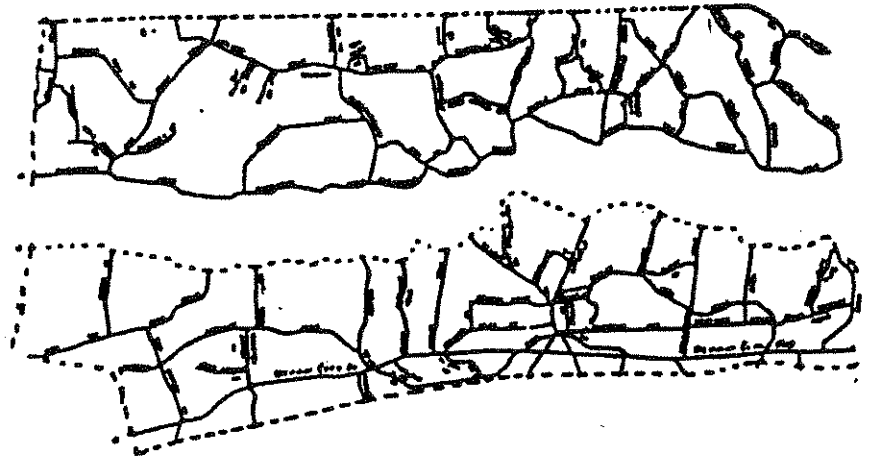


Figure 1. The snow-removal district split into northern and southern areas.

connected and each of the areas are connected. An optimal route for a plow-truck becomes an Euler circuit for the graph of that plow-truck's area.

If one of the graphs,  $A$ , has a tree structure, finding an Euler circuit would be simple. The associated undirected graph of  $A$  would be a rooted tree,  $G$ . A simple Euler circuit for  $A$  could then be obtained by following the preorder traversal [Grimaldi 1989, 487] of  $G$  and then returning to the root of  $G$ .

If A is not a tree graph, it has a subgraph that includes all the vertices of A and is a directed tree graph. To find a subgraph, S, of this form, one edge is eliminated from each loop until no loops remain. Since only edges from loops have been removed, S will still be connected; therefore, S is a tree. Figure 2 shows an example of a graph A and such a subgraph S.

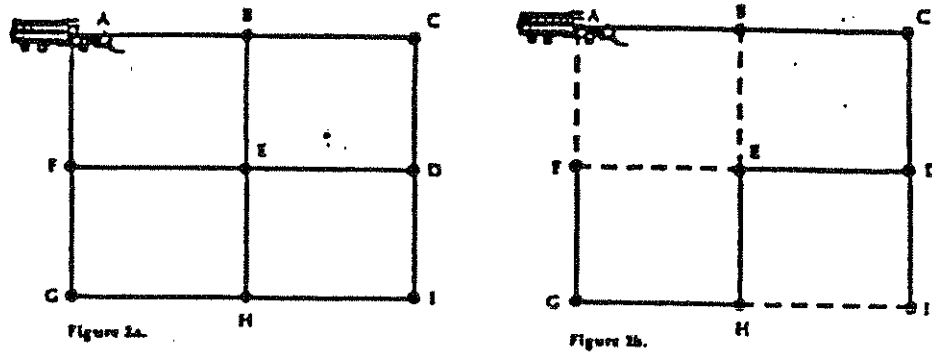


Figure 2a. A graph A

Figure 2b. A subgraph S of A that includes all the vertices of A and is a tree.

Let T be the undirected tree associated with S. For example, the graph in Figure 3a is the undirected tree associated with Figure 2b. The preorder traversal of Figure 3a is I, F, G, H, E, A, B, C, and D. The reader should note, for example, that to travel from I to F, it is necessary to pass through D, E, H, and G.

Let  $T'$  denote the pseudotree formed by adding to  $T$  edges that were in  $A$  but not in  $S$ . These edges will be added such that they are branches only of the first of their vertices to appear in the preorder. It is also necessary to require that these new edges be added only to original vertices of  $T$ . To finish constructing  $T'$ , each edge is made into two directed edges between the two vertices, so that  $T'$  will be a directed graph. Figure 3b shows the directed pseudotree for Figure 3a.

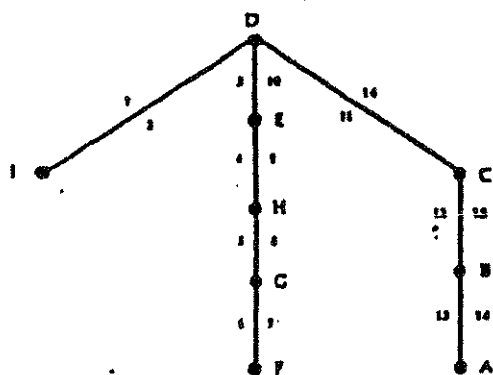


Figure 3a.

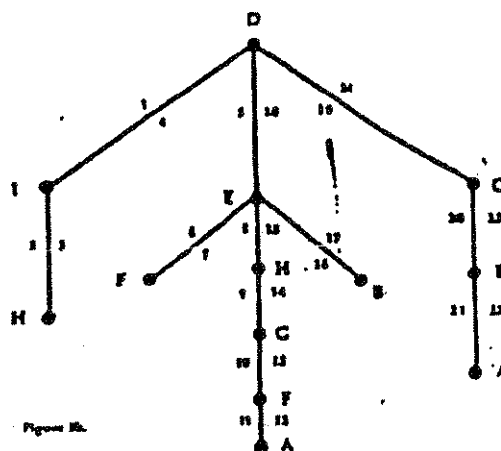


Figure 3b.

Figure 3a. An undirected tree  $T$  associated with the tree-subgraph  $S$  of Figure 2b, with preorder traversal indicated.

Figure 3b. A directed pseudotree  $T'$  for the undirected tree  $T$  of Figure 3a, formed by adding edges that were in  $A$  but not in  $S$ ; the numerals denote the preorder traversal that extends the preorder traversal of  $T$ .

$T'$  may be traversed by following the preorder of  $T$  and agreeing that any edge in  $T'$  connecting vertices that are not connected in  $T$  will be traversed at the first opportunity, to be followed immediately by returning on the other edge connecting those two vertices (see Figure 4).

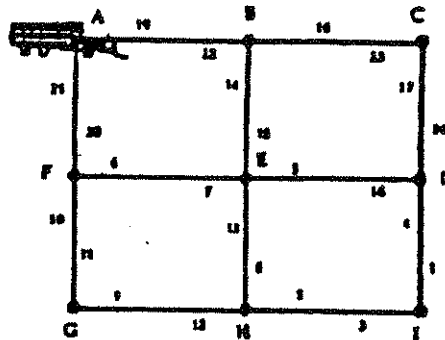


Figure 4. An Euler circuit for  $A$ , based on the traversal of  $T'$  in Figure 3b.

Call this completed circuit  $P$ . Thus  $P$  is also an Euler circuit for  $A$ , because there is a one-to-one correspondence between the edges in  $A$  and the edges in  $T'$ . Thus every edge in  $A$  is traversed the one time it is traversed in  $T'$  by following  $P$ .

#### APPLYING THE MODEL AND ANALYZING ERRORS

Applying the method presented above to the map given in the problem statement, we find optimal paths for the two areas shown in Figure 1. The roads that will not be in  $T$  are removed

from the maps in Figure 5a and 5b, which show subgraphs (which are trees) of the northern and southern areas.

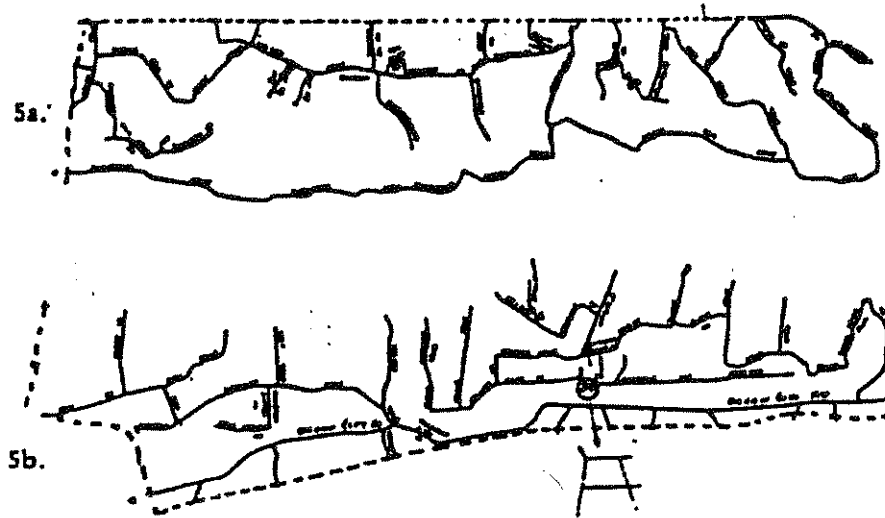


Figure 5. Tree-subgraphs for the northern and southern areas, with roads removed that will not be in the pseudotrees for the areas.

In Figure 6, these roads have been returned, but as dead ends. The optimal paths for the two areas are then obtained by traversing these maps in preorder. The precise paths are shown in an Appendix. [Editor's Note: Omitted for space reasons.]

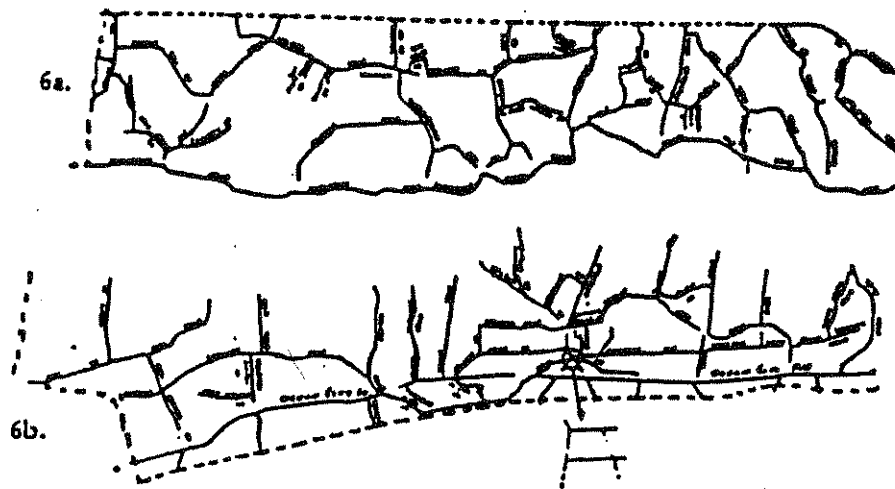


Fig. 6. Subgraphs for the northern and southern areas, with roads not in their pseudotrees reincluded as dead ends.

Now that optimal paths are known, they can be implemented. It is an important requirement for the model that the two areas take the same amount of time to complete. Unfortunately, there are a number of reasons why this might not be the case:

- *Minimal Error.* The scaling increments on the measuring device were equivalent to 0.04 mi. Therefore, 0.02 mi would be the minimal error on any measurement.

- *Errors in Measuring Roads.* The map length of the road was approximated with pliable string. This string was then measured with the measuring device, as was the scale of the enlargement. The inaccuracies involved make it doubtful that the minimal error was achieved.
- *Turning Around.* The plow-truck for the southern area will have to turn around 47 times, while that for the northern area will have to turn around 35 times. While this difference does not seem significant when compared to the total distance that the two plow-trucks will be driving, it will make some difference.
- *0.06 Miles.* The northern area has 0.06 mi more of roads.

None of these errors represent significant flaws in this model. To a large extent, these errors are random; therefore, the errors in the northern area are expected to be roughly equal to the errors in the southern area. Thus a cancellation effect may occur, and we can expect both areas to include nearly the same length of roads.

What errors remain can be eliminated in the following way. The first time this model is implemented, the difference in the time to complete the two areas can be found. If this difference is significant, it can be compensated for by transferring roads, or parts of roads, from the route taking longer to complete to the other route. As long as this is done so that both routes remain connected, it will be possible to use the algorithm above to produce two new optimal routes.

#### STRENGTHS OF MODEL

A good model can easily accommodate many variations of the original problem. Our model can handle a great variety of changes in the problem, including plow-trucks with different capabilities, different numbers of plow-trucks, even different maps.

To compensate for plow-trucks with differing capabilities, the boundaries between the areas can be moved to give more powerful or faster plow-trucks more road.

If Wicomico County decided to assign a different number of plow-trucks to this snow-removal district, the district should be divided into that number of connected areas. Each area should be assigned a plow-truck of appropriate size and contain a road providing quick access to its garage. Then the entire district will be cleared promptly, because the plow-trucks will finish at approximately the same time.

Our model can be applied to a different road system as well. The method, of dividing the system into areas for each plow-truck and using trees to find optimal paths through each area, can be applied equally well to any set of connected roads.

Furthermore, the model has several inherent strengths that arise from the underlying graph-theoretic structure. The drivers can follow simple instructions to complete optimal paths covering their routes; the plow-trucks finish their routes near their respective garages; a minimum amount of fuel is used, as there is no unnecessary driving.

#### DRIVER'S INSTRUCTIONS

The following instructions can be given to a driver who has a limited knowledge of the area:

1. Stay in your assigned area.
2. Always drive on lanes that have not been plowed.
3. If you come to an intersection that has already been plowed and the road behind you has had only one lane plowed, turn around and plow the other lane.
4. If it was not necessary to turn around and one or more completely unclear roads meeting at the intersection, turn onto the rightmost of these.

5. Otherwise, plow the rightmost road that does not have a plowed right lane.

These instructions are equivalent to traversing a pseudotree in preorder, since the driver will treat one edge of any loop as an edge unique to the pseudotree, by turning around at the end of the road representing that edge and then clearing the other lane of that road.

#### REFERENCE

Grimaldi, Ralph. 1989. *Discrete and Combinatorial Mathematics: An Applied Introduction*. New York: Addison-Wesley.

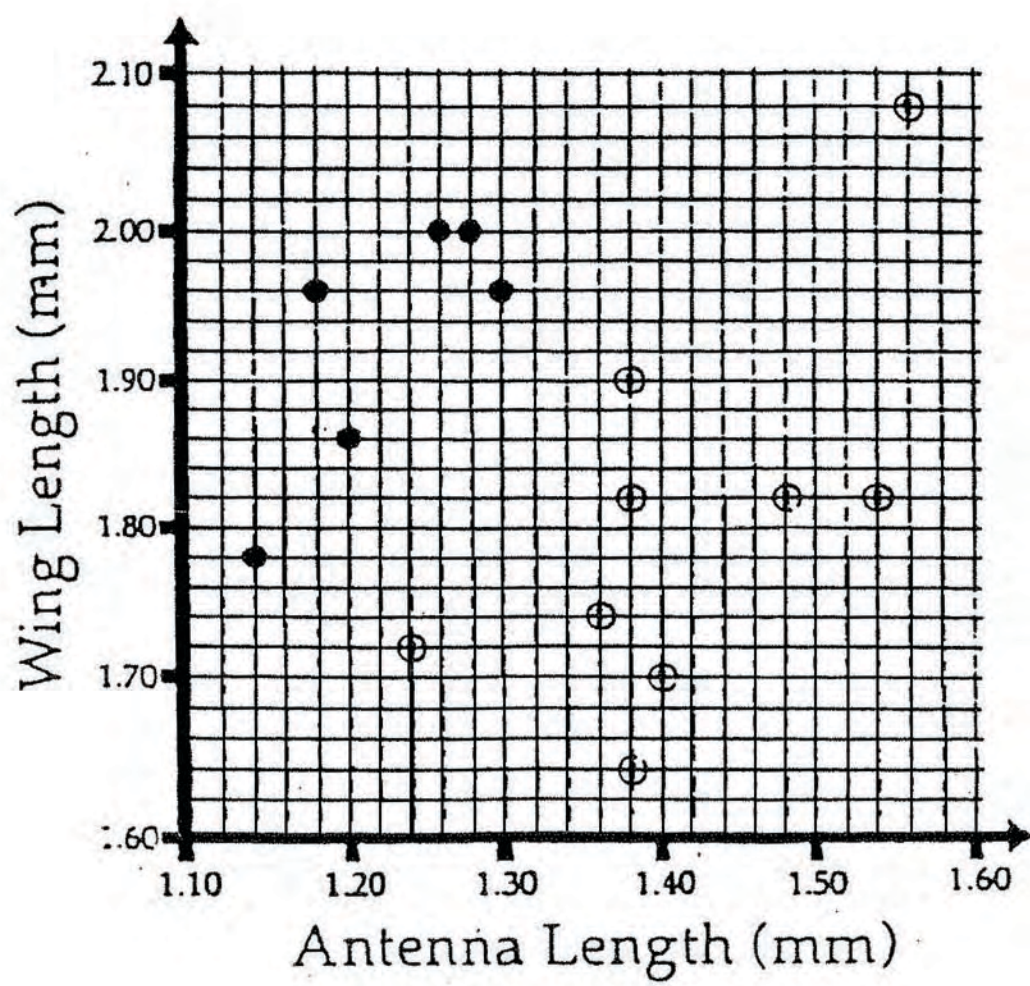
Let us move on to an exemplary math team paper in Biological Classification. Before you peruse the paper, please review Discriminant Analysis in SPSS (Advanced Student Guide, Chicago: SPSS Inc., 1990) or Multivariate Analysis by Maurice Tatsuoka (New York: Wiley, 1975).

Also review Numerical Analysis procedures to test for normal densities, bivariate normal densities, and the assumptions underlying Discriminant Analysis. The students from California Polytechnic State University (San Luis Obispo) submitted the best paper in the world. They demonstrated exceptional brilliance in correctly choosing Discriminant Analysis as the key to the solution. They also demonstrated mastery of the mathematics of Discriminant Analysis, the assumptions underlying the method, and clear explanation of the strengths and weaknesses of their approach. Remember, this was best in world. Your team wins by competing and most importantly working together to better understand the pivotal role of Mathematical Modeling in the 21st century.

### Midge Classification

Two species of midges, Af and Apf, have been identified by biologists W.L. Grogan and W.W. Wirth (1981) on the basis of antenna and wing length. Please refer to the diagram. Each of nine Af midges is denoted by " $\theta$ ", and each of six Apf midges is denoted by " $\bullet$ ". It is important to be able to classify a specimen as Af or Apf, given the antenna and wing length.

- 1) Given a midge that you know is species Af or Apf, how would you go about classifying it?
- 2) Apply your method to three specimens with antenna, wing lengths (1.24, 1.80), (1.28, 1.84), (1.40, 2.04).
- 3) Assume that species Af is a valuable pollinator, and species Apf is a carrier of a debilitating disease. Would you modify your classification scheme, and if so, how?



OPTIMAL CLASSIFICATION AND SEPARATION:  
INFERENCES ABOUT A MEAN VECTOR

Scott Guth  
Mike Kelleher  
Scott Langfeldt  
California Polytechnic State University  
San Luis Obispo, CA 93407

Advisor: T. O'Neil

SUMMARY

A common problem in numerical taxonomy is to find the optimal separation of populations and classify individuals into those populations. Often there are few data points, and perhaps there is even overlap among the different populations. In 1981 W.L. Grogan and W.W. Wirth identified 15 midges, 6 of type Apf and 9 of type Af. Our goal was to find the best possible dividing curve between the two types; achieving it required extensive use of multivariate statistical analysis and some creativity.

There are three essential parts to solving the problem. First, we describe algebraically the unique features (in terms of antenna and wing length) of the observed midges, to yield a discriminant that separates the populations. Second, we classify given midges of unknown species as Apf or Af. Finally, we take into consideration the total probability of misclassification and revise our discriminant to minimize the expected cost of misclassification.

We present Fisher's method of classification, with variations. We analyze its ability to separate the populations and classify new midges. In addition, we mention other methods and discuss why we found them unsuitable.

We use Fisher's method to classify the given specimen midges. We then adjust the method for different assumptions, such as the costs of misclassification and different population proportions. Finally, we assess the likelihood of misallocation.

We compute four lines of separation, based on various assumptions about population proportions and misclassification costs. Assuming equal population sizes and misclassification costs, the Fisher method gives the line  $y = 0.8883x + 0.6930$ . Using this line, we classify all the test specimens as Apf.

#### 1. STATEMENT OF THE PROBLEM

Figure 1 displays the data gathered by Grogan and Wirth.

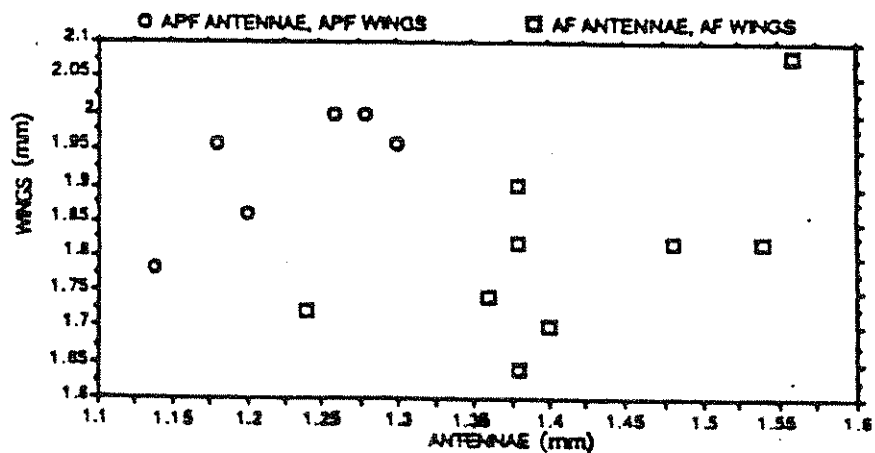


Figure 1. Display of data collected by Grogan and Wirth

The problem statement requested that we:

1. Find a procedure for classifying the two species from wing and antenna lengths.
2. Use that procedure to classify three given specimens.
3. Modify our procedure to take into account the assumptions that species Af is a valuable pollinator, and species Apf is a carrier of a debilitating disease.

## 2. ASSUMPTIONS

1. We assume homogeneity of gender of the specimens.
2. We assume that the data are from two different bivariate normal populations.
3. Since no direct information is given, we assume reasonably either that both populations are equal in size, or else that the populations are in the same proportions as the sample:  
6/15 Apf, 9/15 Af.
4. We assume that wing length is precisely as good (or bad) an indicator of species as antenna length.
5. We assume that wing length and antenna length are sufficient to determine the species of midge.

## 3. JUSTIFICATION OF ASSUMPTIONS

### 1. *Sexual homogeneity*

Since sex of the individual midge is not part of the data, we must assume either that there are no differences (which we know to be false) or that only one sex is represented in the data.

## 2. *Evaluating bivariate normality*

There are various methods to assess bivariate normality. One method is to transform the data points onto a 4-dimensional sphere; another method simply checks that the marginals are normal. Opting for a happy medium, we chose a chi-square method, which examines level curves of the probability surface.

For each species, approximately half of the data point pairs  $X$  should lie with the ellipse  $(X - \mu)' \Sigma^{-1} (X - \mu) \leq \chi^2_{2^2}(0.5)$ , where  $\mu$  is the vector of population means (wing length, antenna length) and  $\Sigma$  is the  $2 \times 2$  population covariance matrix.

We used sample values to estimate the parameters for each population of midges. Since there are so few data points, we calculated the left-hand side at each point in both samples to see if it lies with the corresponding ellipse. We found that all points lie with the desired regions [EDITOR'S NOTE: For space reasons, the tables of values are omitted.] Thus, we do not reject the assumption of bivariate normality for each population.

## 3. *Population proportions*

We must make some assumption about the population proportions. There is no reason to assume any particular proportion, so we pick two on which to focus attention:

$$P_{Apf} = P_{Af} = 1/2 \quad \text{and} \quad P_{Apf} = 6/15, \quad P_{Af} = 9/15$$

## 4. *Wing and antenna as indicators*

We were given no information about the significance of wing length being better or worse than antenna length, and entomological research gave us no reason to believe that they are not equal indicators.

#### 5. *Sufficient number of features*

We were given data on wing and antenna length only. We therefore must assume that there are no other data values necessary to determine the species.

#### 4. ANALYSIS OF THE PROBLEM: FISHER'S METHOD

One of our assumptions is that the data for the populations come from bivariate normal distributions.

The p-dimensional normal probability density function is

$$f(X) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp [-(1/2) (X - \mu)' \Sigma^{-1} (X - \mu)]$$

R. A. Fisher reduced this equation to give a classification function. The idea behind his function is that the two-dimensional variable  $X$  can be transformed into a one-dimensional variable  $y$  by projection onto a real line. The line is chosen such that the projection onto it yields the widest possible spacing between the  $y$ 's of one population and the  $y$ 's of the other. The distribution of the  $y$ 's on the line is normal. At the point  $m$  equidistant from the univariate means, a perpendicular is erected, which is our discriminator.

Important in this method are the sample covariance matrices  $s_{Apf}$  and  $s_{Af}$ . The four elements of each matrix are measures of variance in the sample. The entries are:

$$s_{11} = \frac{\sum_{i=1}^n (w_1 - \bar{w})^2}{n - 1}$$
$$s_{22} = \frac{\sum_{i=1}^n (a_1 - \bar{a})^2}{n - 1} \quad \text{and}$$

$$s_{12} = s_{21} = \sqrt{\frac{\sum_{i=1}^n (a_i - \bar{a})(w_i - \bar{w})}{n-1}}$$

where the antenna measurements are denoted by  $a_i$  and the wing measurements by  $w_i$ .

R.A. Fisher proposed the following discriminant:

Sample point  $x$  should be allocated to population 1 if

$$(x_1 - x_2)' S_{\text{pooled}}^{-1} x \geq (1/2) (x_1 - x_2)' S_{\text{pooled}}^{-1} (x_1 + x_2) + \ln [(c_{12}/c_{21}) (p_2/p_1)],$$

where  $S_{\text{pooled}}^{-1}$  is the inverse of the adjusted covariance matrix for the pooled data from both populations,  $c_{ij}$  is the cost of misassigning an observation of species  $j$  to population  $i$ , and  $p_i$  is the proportion of the size of the population  $i$  to the entire midge population size.

If the previous inequality is not true,  $x$  should be allocated to population 2.

Although the above criterion is not intuitive, the effects of changing  $c_{12}$  and  $c_{21}$  are.

Weighting the cost of misclassification for either population is equivalent to increasing the other population's size. In our situation, misclassifying an  $A_{pf}$  (from population 1) as an  $A_f$  (population 2) is more severe than the reverse. So we should have the cost ratio  $c_{12}/c_{21}$  less than 1.

Assuming equal population proportions, the natural log term will be negative, thereby decreasing the discriminant on the right-hand side and thus tending to classify more midges as  $A_{pf}$ .

Fisher assumed that the population sizes would be equal and that there is no difference in the cost of misclassification between the two populations. This simplification causes the log term to vanish. (Later we will consider the effects of the log term under assumptions different from Fisher's.) Fisher's allocation rule reduces to:

$$(x_1 - x_2)' S_{\text{pooled}}^{-1} x \geq (1/2) (x_1 - x_2)' S_{\text{pooled}}^{-1} (x_1 + x_2)$$

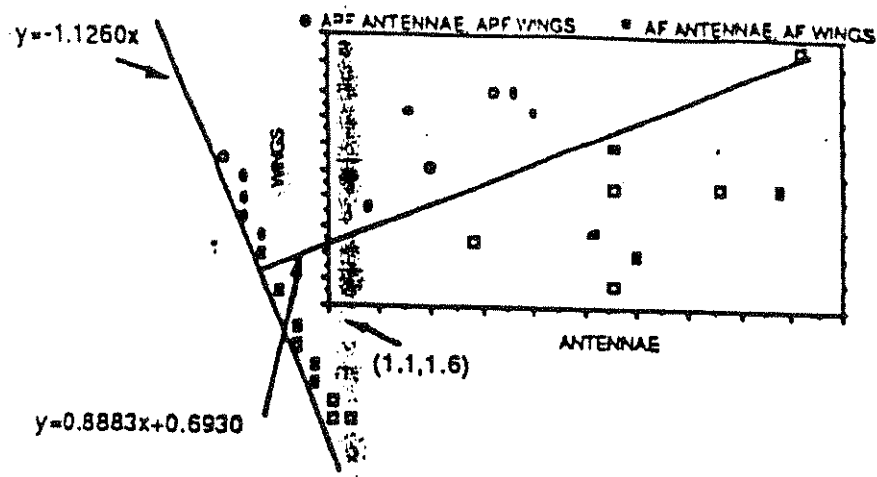


Figure 2. Projection onto Fisher's perpendicular.

This inequality describes the "northwestern" half-plane of midges that should be classified as  $A_{pf}$ . The line dividing the two populations has equation  $y = 0.8883x + 0.6930$  (see Figure 2).

A popular method for evaluating the discriminant, if the distribution of the populations is unknown (i.e., relying only on sample data), is known as the *apparent error rate* (APER) and is quite simple to compute, especially for small sample sizes. APER is simply the proportion of the specimens that are misclassified by the function. In our case,  $APER = 1/15$ . There are better ways of evaluating the discriminant, but they all involve having fairly large samples. We do not have this luxury with only 15 data points given.

Determining the likelihood that a point has been misclassified is quite easy. In Figure 3, we see shaded regions that are the overlap of the two normal curves. In these regions there is a reasonably high probability of misclassification. At a point in the shaded regions, the ratio of the two probability density values is equal to the ratio of  $p(R)$ , the probability of assessing correctly, to  $p(W)$ , the probability of assessing incorrectly.

Since  $p(W) + p(R) = 1$ , we have:

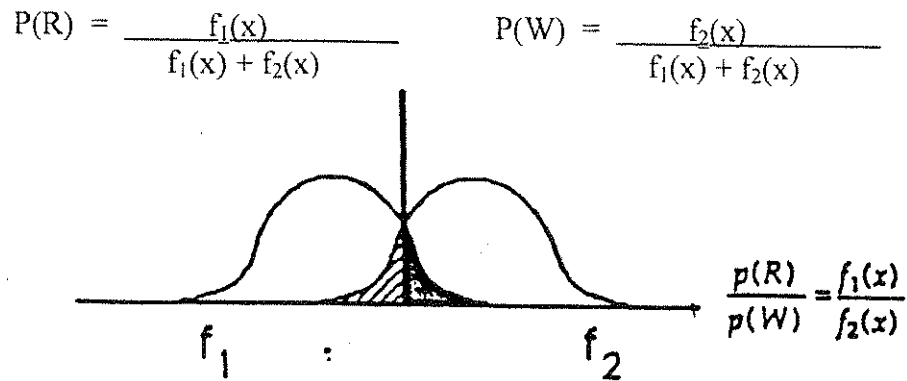


Figure 3. Overlap of two normal probability density functions.

However, if one does not assume a cost ratio of 1, the picture looks more like Figure 4. The adjusted Fisher discriminant no longer splits the curves at the point of equal probability. The probability of misclassifying an  $A_f$  as an  $A_{pf}$  has skyrocketed, while the probability of misclassifying an  $A_{pf}$  as an  $A_f$  has plummeted.

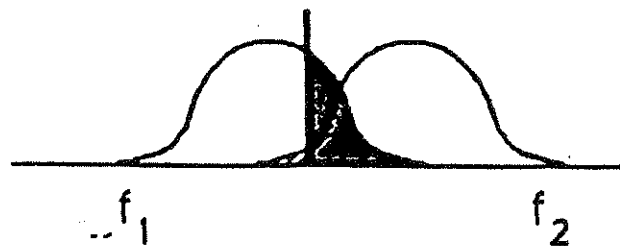


Figure 4. Overlap of normal probability density functions for non-unit ratio of misclassification costs.

Actually, Fisher's method for discriminating two populations is a simplification of another method, the *quadratic classification rule*:

$$(x'_1 S^{-1}_{Apf} - x'_2 S^{-1}_{Af}) x - 1/2 x'_1 (S^{-1}_{Apf} - S^{-1}_{Af}) x - k \geq \frac{[(c(1/2) (P_2))]}{(c(2/1) (P_1))}$$

where

$$k = \frac{1}{2} \ln \left( \frac{|S_{Apf}|}{|S_{Af}|} \right) + \frac{1}{2} (x_1' S_{Apf}^{-1} x_1 - x_2' S_{Af}^{-1} x_2)$$

If the two covariance matrices are equal, the middle term vanishes and the rule reduces to Fisher's.

The quadratic classification rule divides the plane with a conic section. For our data, with equal population proportions and misclassification cost ratio 1, we get an ellipse. The given points fall in the square window of Figure 5, where the ellipse has equation

$$-15.8829a^2 + 36.26357aw - 23.26571w^2 - 37.0697a + 52.00338w - 28.47765 = 0$$

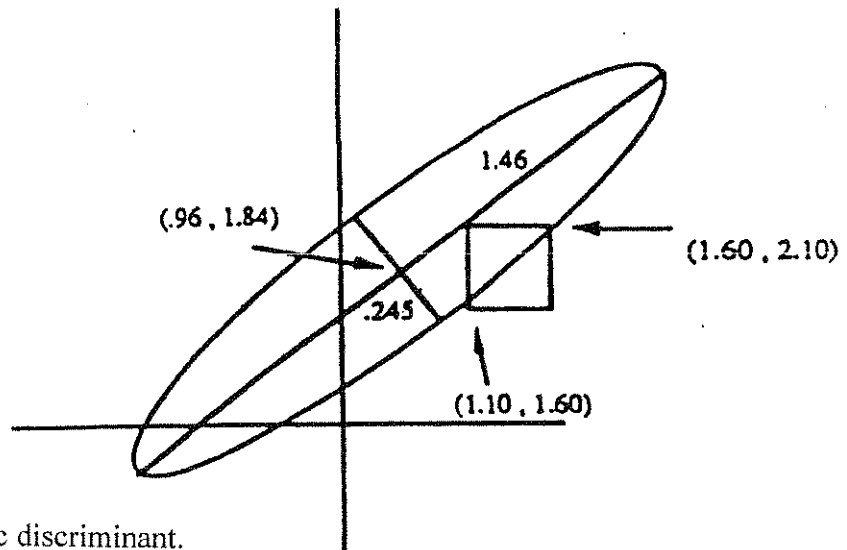


Figure 5. The quadratic discriminant.

The Fisher discriminant line deviates from the ellipse segment in the window by less than 1%. Thus, the two sample covariance matrices are close enough to being equal for Fisher's method to be valid. Computing the ellipses for unequal populations and unequal misclassification costs can be done easily, but little information is gained, since Fisher's method works so well.

## 5. RESULTS

We classified the three specimen midges and determined the likelihood of our classification being in error. We were able to make simple adjustments to consider different population proportions (see Table 1).

Table 1.

Classification of test specimens, with cost ratio 1. The function  $w$  is the Fisher discriminant function that gives the boundary value for wing length as a function of antenna length.

Equal population sizes $P_1 = P_2 = 0.5$ $W(\text{ant}) = 0.88830 \text{ ant} + 0.69300$				Unequal population sizes $P_1 = 0.4, P_2 = 0.6$ $w(\text{ant}) = 0.88830 \text{ ant} + 0.71524$		
Midge (ant,wing)	w(ant)	species	p(W)	w(ant)	species	p(W)
(1.24, 1.80)	1.794	$A_{pf}$	.27	1.816	$A_f$	.36
(1.28, 1.84)	1.830	$A_{pf}$	.26	1.852	$A_f$	.34
(1.40, 2.04)	1.937	$A_{pf}$	.12	1.959	$A_{pf}$	.17

In practice, one would want to set a threshold for the proportion of  $A_{pf}$  midges misclassified as  $A_f$  midges. For definiteness, we chose 0.5 for our cost ratio, which is equivalent to assuming that misclassifying an  $A_{pf}$  is twice as bad as misclassifying an  $A_f$ . This change in the cost assumption yields the results of Table 2.

Table 2.

Classification of test specimens, with cost ratio 0.5.

Equal population sizes $P_1 = P_2 = 0.5$ $W(\text{ant}) = 0.88830 \text{ ant} + 0.65497$				Unequal population sizes $P_1 = 0.4, P_2 = 0.6$ $w(\text{ant}) = 0.88830 \text{ ant} + 0.67721$		
Midge (ant,wing)	w(ant)	species	p(W)	w(ant)	species	p(W)
(1.24, 1.80)	1.756	$A_{pf}$	.27	1.779	$A_{pf}$	.64
(1.28, 1.84)	1.792	$A_{pf}$	.26	1.814	$A_{pf}$	.66
(1.40, 2.04)	1.899	$A_{pf}$	.12	1.921	$A_{pf}$	.17

Of note in Table 2 are two entries in the last column that indicate probability of misclassification greater than 0.5. These figures suggest that the midges in question should be classified in the other group. However, the cost ratio changes  $p(W)$  and  $p(R)$  so that there is no longer an upper limit of 0.5 on  $p(W)$ .

## 6. STRENGTHS AND WEAKNESSES

There are several strong points to using Fisher's method:

- It is relatively simple (compared to translating the data points to points on a higher-dimensional surface);
- It is well known;
- It makes no strong assumptions about the two populations, just a minor assumption about the structure of the covariance matrix;
- It gives a nice, believable result.

Methods that use training samples have some advantage over Fisher's, in that the data used to create the classification is not then used to evaluate the classification. Fisher's method, in using every data point, achieves the highest possible precision.

Another strong aspect of the Fisher method is the malleability of the discriminant function. This robustness may be hard for the casual observer to appreciate; but to the mathematician, it is a godsend. The simplicity of the model allowed for keeping the equations algebraic until we were ready to evaluate them. Also, changing the population proportions and the cost of a misclassification involved adding only a single term to the discriminant.

Unfortunately, Fisher's method also leaves some weak spots in the model. For example, it uses a pooled  $s$  value, which calls for near-equality in the variances; and it assumes normality. While one can simulate this normality with transformations, the less normal the probability distribution function is, the less likely it is that Fisher's method will give a reasonable separation and classification.

## 7. ALTERNATIVE SEPARATION MODELS

### 7.1 Regression line

The first method was a quick and dirty attempt to gain some basic knowledge about the two samples. Regression lines were calculated for the  $A_f$  and the  $A_{pf}$  data, and the average of the two lines was calculated. The idea was to allocate a midge to the population in whose half-plane its measurements fall.

$A_f$	$y = 0.825x + 0.637$	$\{r_{adj}^2 = 0.31\}$
$A_{pf}$	$y = 1.100x + 0.576$	$\{r_{adj}^2 = 0.52\}$
Average	$y = 0.9625x + 0.6075$	

We then observed that the  $A_f$  midge with antenna and wing lengths of 1.56 and 2.08 respectively was an odd midge, possibly an outlier. We decided to try this method without that data point, and then calculated the new  $A_f$  regression line and average:

$$\begin{array}{lll} A_f & y = 0.354x + 1.28 & \{r_{adj}^2 = 0.00\} \\ \text{Average} & y = 0.727x + 0.928 & \end{array}$$

The correlation is worse, and we began to doubt this method. After more study, we cast it out as worthless for this problem.

## 7.2 Centroid bisector

In the second method, we calculated the centroids (i.e., averages) for each sample. We then found the perpendicular bisector of the line connecting these points. The centroids of the  $A_{pf}$  and  $A_f$  samples are (1.22, 1.93) and (1.41, 1.80). The centroid bisector is  $y = 1.52576x - 0.14850$ .

This method had an incredible intuitive feel to it, though it was hard to see why it should be considered valid. In fact, we soon realized that this method was a beautiful solution to a problem in which the populations had equal standard deviations and sizes - not the case for our problem, so we rejected the method.

## 7.3 Equality of squares

In this method we calculated a line such that for each point on the line, the average of the sums of the squares of the distances to the points in one sample is equal to the same measurement for the other sample. In equation form:

$$\frac{\sum_{i=1}^N \{[x - x_i]^2 + [(mx + b) - y_i]^2\}}{N} = \frac{\sum_{j=1}^M \{[x - u_j]^2 + [(mx + b) - v_j]^2\}}{M}$$

We calculated  $b$  by setting  $x = 0$  and then calculated  $m$  by setting  $x = 1$ . The line computed was  $y = 1.5272x - 0.2074$ .

This model left the point at (1.24, 1.72) in the wrong population. We again turned our suspicions on the potential outlier, (1.56, 2.08). We omitted this point and calculated the line again. We came up with the line  $y = 1.0745x + 0.4297$ .

For standard deviations that are nearly the same, this is an acceptable model; but for substantially different variances in the two populations, the model is not suitable.

---

### References

- Devore, Jay L. 1982. *Probability and Statistics for Engineering and Statistics*. Belmont, CA: Wadsworth.
- Johnson, Richard A. and Dean W. Wichem, 1982. *Applied Multivariate Statistical Analysis*. Englewood Cliffs, NJ: Prentice-Hall.