Executive Summary

As mathematicians making a model for field biologists, we needed a simple and statistically defensible way to identify which of the two morphologically similar species (Aa harmless, and Ax dangerous) flies we were catching in Guatemala. We needed to find a model that could be executed with low error risk. Our goal was to determine a rule that can be simply followed in the field that can classify flies using easily measured body lengths (wing, abdomen, and antenna).

We began by analyzing all the available measurements and found that there was not a single variable that provided a clear distinction between the species. When we explored two-variable relationships, only the abdomen-to-antenna ratio (Q) showed consistent separation. The relationships involving wing length showed us a high overlap and were therefore excluded. We chose a ratio 3 standard deviations below the mean of Q to create in our model a tendency to overclassify flies as the more dangerous Ax. We found this threshold to be 1.36. Flies above the threshold we categorized as Ax, and below we categorized as Aa.

We next validated this model by generating 100 pseudoflies (50 Aa, 50 Ax) to test our model's robustness. Each pseudofly approximately preserved the real-data correlation between abdomen and antenna by drawing one variable from the normal distribution of that species and predicting the other through species-specific linear regressions. Residuals from these regressions were also predicted well by a normal distribution, allowing us to generate synthetic noise for our predicted values through normally distributed random values. Both abdomen-first and antenna-first generation methods produced visually and statistically similar outcomes.

On this simulated dataset, our classifier performed with approximately 98% overall accuracy, missing very few Ax flies, a critical result for field safety. When tested on three new field species with Q of 1.45, 1.44, and 1.46, all exceeded our threshold and were therefore classified as Ax.

To make this process field-usable, we established a three-step guide:

1. Measure the abdomen and antenna.
2. Compute Q and compare to 1.36.
3. If the ratio exceeds the threshold, treat the fly as Ax. Otherwise, treat it as Aa.

Though it is possible to refine the determined threshold by adding additional data and recalculating the means of each variant, we would encourage biologists to exercise caution when feeding data classified by this model back into it. Doing so would reinforce existing errors, such as the bias toward Ax classifications, gradually bringing down accuracy. A safe way to improve the model is to add additional data only after it is verified by a trusted external method, such as genetic testing, when and where feasible.

Problem Restatement

Biologists collected three length measurements (wing, abdomen, and antenna) for two biting fly species found in Guatemala, labeled as Aa and Ax. Ax flies are vectors of disease, while Aa flies are comparatively harmless. Using the three simple measurements, we are tasked to:

1. Design a field classification rule that distinguishes Aa from Ax using wing, abdomen, and/or antenna lengths
2. Generate 100 pseudoflies as synthetic data and use them to test our model
3. Finally, test our procedure with the following three flies:

Table 1: Measurements of 3 testing flies			
Fly 1	2.81	1.80	1.24
Fly 2	2.65	1.84	1.28
Fly 3	3.61	2.04	1.40

We were given the following measurements for the flies:

Table 2: Given fly measurements					
Measurements of Aa flies			**Measurements of Ax flies**		
Wing	**Abdomen**	**Antenna**	**Wing**	**Abdomen**	**Antenna**
2.87	1.78	1.14	2.10	1.72	1.34
4.02	1.86	1.29	3.31	1.94	1.58
3.18	1.96	1.3	3.83	1.74	1.64
3.51	2.00	1.26	3.11	1.70	1.45
3.74	2.10	1.39	1.65	1.82	1.38
3.95	1.96	1.28	1.73	1.83	1.48
3.28	1.87	1.28	1.84	1.90	1.49
			2.49	1.82	1.54

	2.53	1.88	1.56
	2.99	1.87	1.65

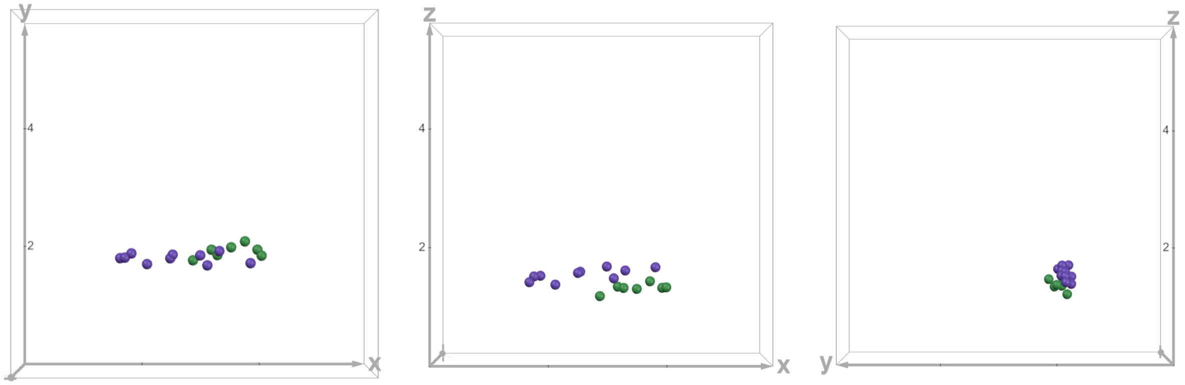
Assumptions

1. We assume that the data provided is a representative sample of flies in the wild.
2. We assume that abdomen, antennae, and wing length on flies are not independent of one another, as, for example, a larger fly will have both a larger abdomen and a larger antennae length.
3. We assume flies are independent samples, and our simulated “pseudoflies” are generated as independent draws from the fitted distributions/regressions.
4. We assume asymmetric costs: misclassifying Ax as Aa is worse than the reverse, so our decision rule is biased toward calling Ax when uncertain.

Determining Fly Type

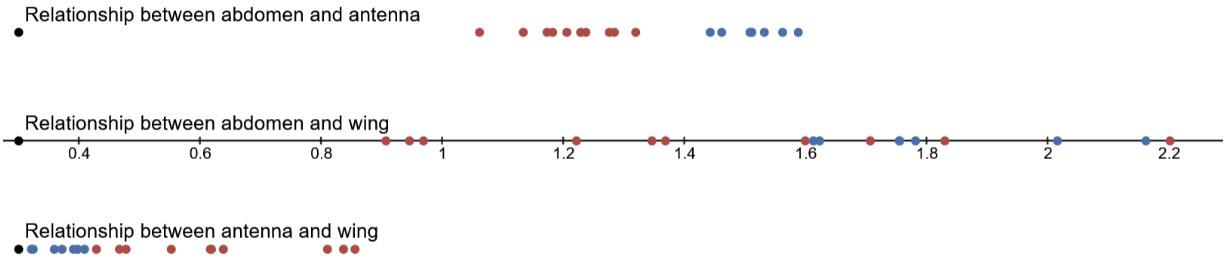
We began by examining a graph comparing the wing (x), abdomen (y), and antennae (z) length of the 2 sets of flies. Looking at the projections of the data (traces) onto the xy plane, yz plane, and xz plane, we can observe that none of the three projections reveal a visually significant difference between the two sets of data (Figure 1).

Figure 1: Traces of wing (x), abdomen (y), and antennae (z) for both fly species



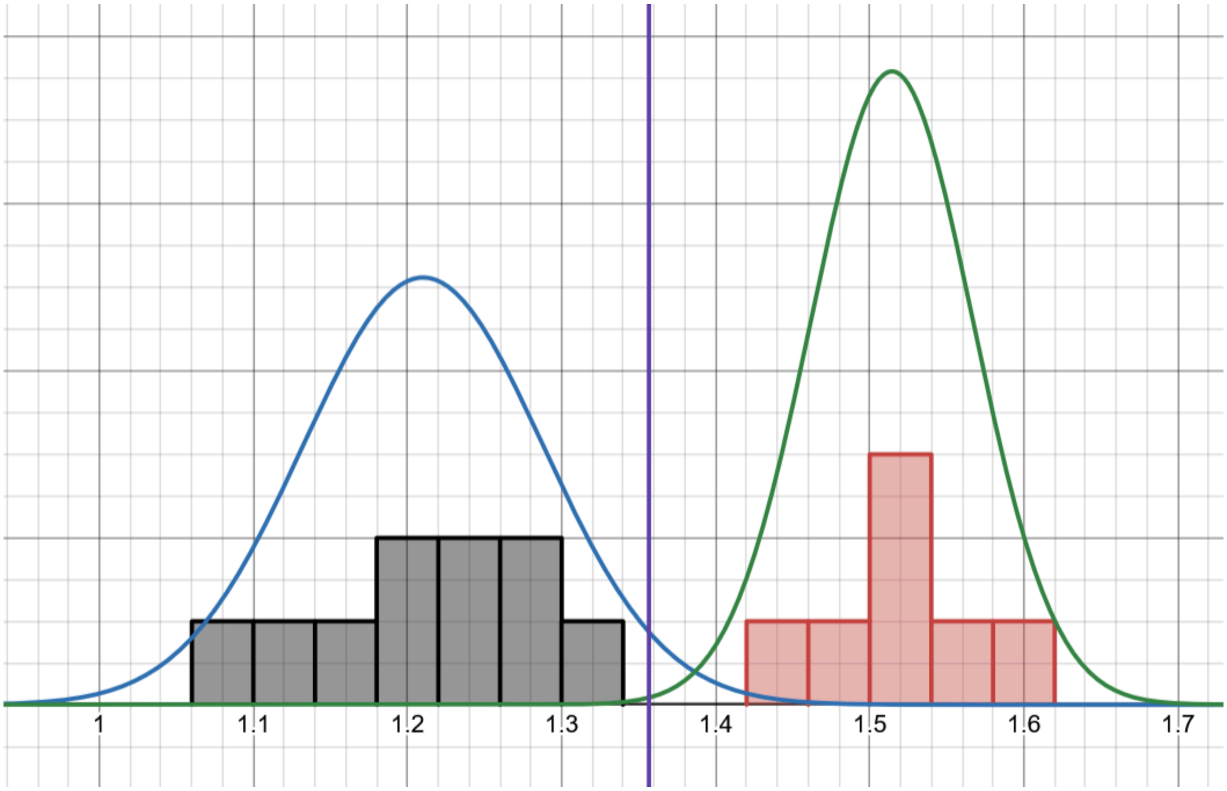
Since using a single measurement did not seem feasible, we then tested the relationships between any two of the fly measurements. We would prefer a method of testing that only uses two measurements instead of three so that field biologists would need to make fewer measurements, decreasing complexity and error from incorrect measurements. We plotted the relationships (abdomen/antenna, wing/abdomen, wing/antenna) for both data sets and observed the separation between the sets of data. We only observed a significant partitioning of data sets for the antenna/abdomen plot (Figure 2). We refer to the antenna/abdomen ratio as Q.

Figure 2: Visualization of ratios of fly measurements by species



We fit the data for each species of fly to two bell curves (Figure 3). Our field test for biologists to determine species is made to prioritize describing flies as Ax over Aa if there is uncertainty, since miscategorization of flies as Aa when they are Ax is considered more dangerous, due to their nature as vectors of disease. For that reason, we chose to measure 3 standard deviations below the mean of Q for the Ax flies, producing a 99.7% confidence interval. This means we are 99.7% sure the true population mean for the Ax flies appears above the classifier (3 standard deviations away from the sample mean). This classifier is shown as a vertical line in Figure 3.

Figure 3: Normal curves of current Ax and Aa abdomen/antenna data sets



Future data can be added to each data set to refine the model further. We recommend only adding data that is verified through a separate, trusted metric, since using our metric will reinforce the bias of our model (see discussion).

Guide for Biologists

This section is a short guide for biologists on how to use our model to identify fly species in the field. To classify a fly, first take its abdomen and antenna measurements. Then, divide the abdomen measurement by the antennae, and test if this value is above the classifier in our model (without any additional data, this classifier is 1.36). A value above the classifier indicates that the fly is likely of the Ax species, while a value below the classifier indicates that the fly is likely of the Aa species.

To refine the model, use additional data collected in the field. Do not use our metric to add data to our model, or model bias will be reinforced (see discussion). Recalculate the mean and standard deviations for each data set, and then calculate the classifier as 3 standard deviations below the mean of the Ax data set.

Generating Pseudoflies

We chose to generate values for pseudoflies using normally distributed random numbers. For these values to be reasonable, we must show that the data for the flies we have is reasonably well approximated by a normal distribution. This can be accomplished through examination of the normal probability plots of the flies' attributes. In a dataset well-approximated by a normal distribution, the expected z-value, z_{exp}, of each data point if it were normally distributed plotted against its value, x, should be well-approximated by

$$x(z) = sz + \bar{x}$$

where s is the standard deviation of the dataset and $\bar{x}$ is the mean of the dataset. The expected z-value of the ith smallest data point of a dataset is defined as the probit of the plotting position of that data point, where the probit is the inverse cumulative distribution of the normal distribution, and the plotting position is the approximate quantile of the ith smallest value of the normal distribution (here approximated by Blom's formula).

The normal probability plots for the antennae and abdomen measurements of each type of fly indicate that the measurements are well-approximated by a normal distribution (Figures 4-5); that is, the points (z_{exp}, x) are close to the line $x(z)$ for each dataset (Aa antennae, Aa abdomen, Ax antennae, Ax abdomen). The shape of the normal probability plot for Ax antennae (Figure 5) suggests possible skew by its "S" shape, but the correlation coefficient between z_{exp} and x, 0.9184, was high enough that we chose to proceed with the assumption that values generated by a normal distribution would still be good representations of realistic antennae values.

Figure 4. Normal probability plots for antennae and abdomen measurements of Aa flies

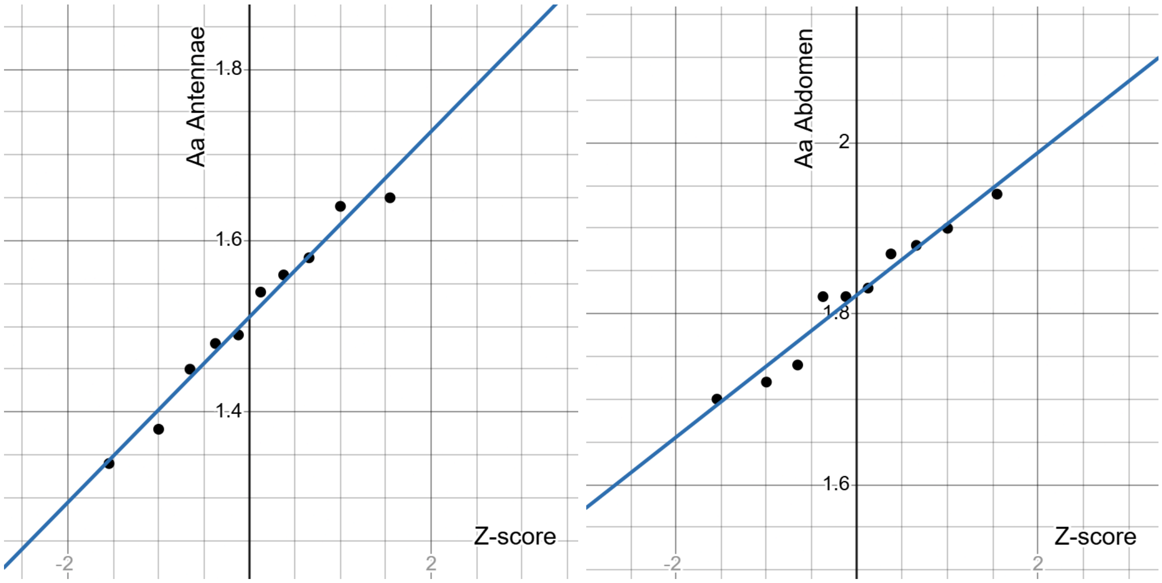
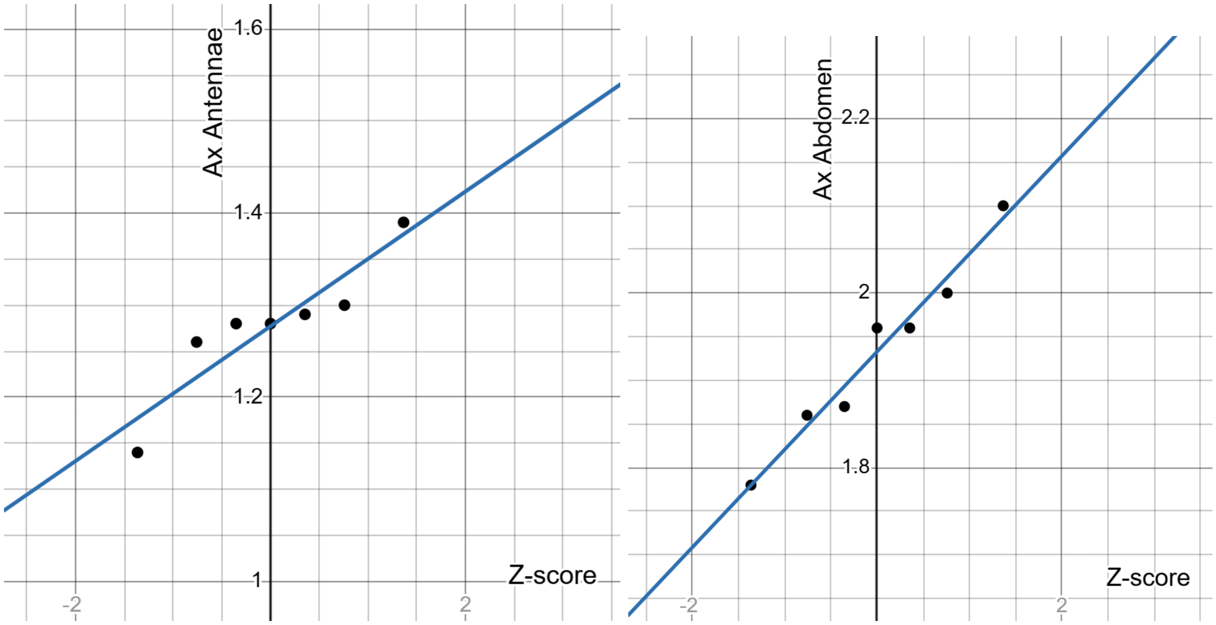
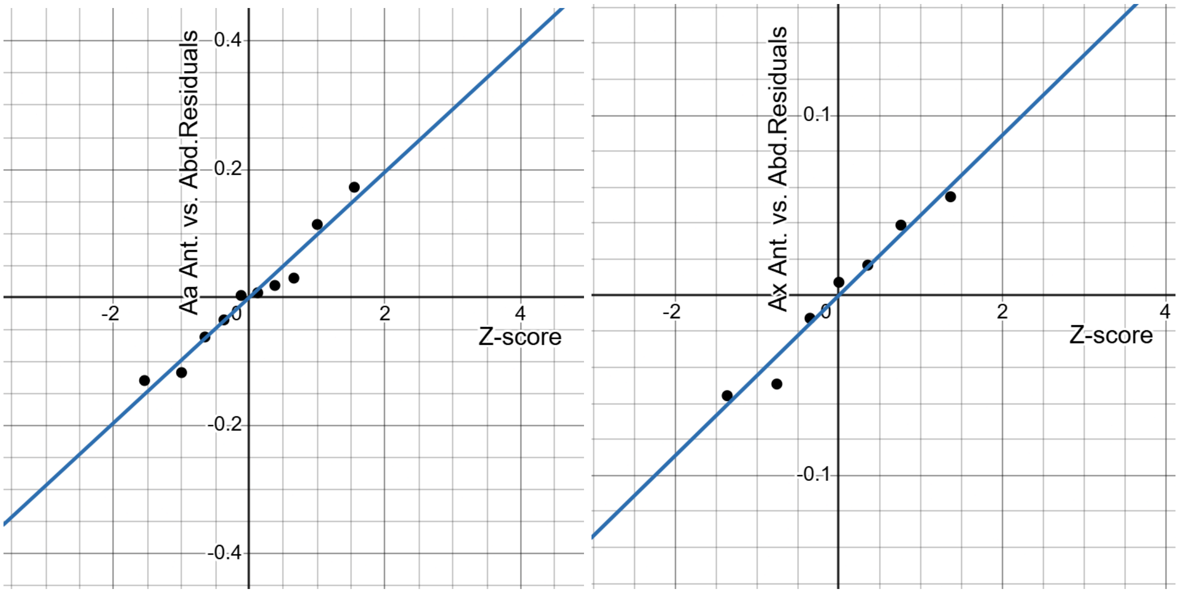


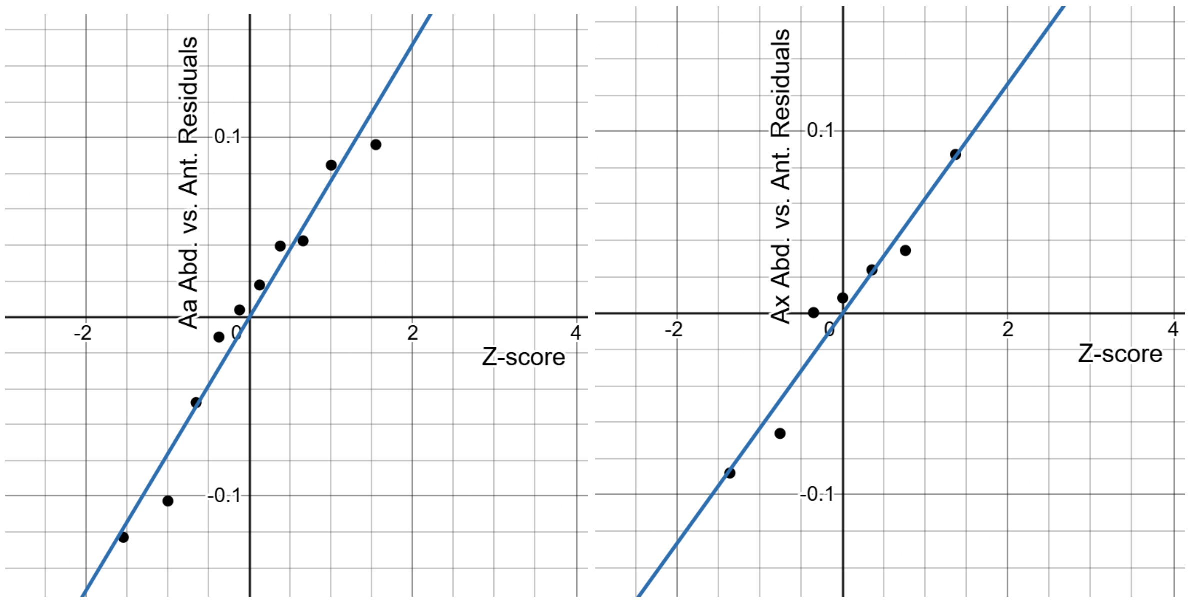
Figure 5. Normal probability plots for antennae and abdomen measurements of Ax flies



It is thus reasonable to generate pseudo-values for each dataset by generating normally distributed random numbers according to the normal distribution of that dataset (here accomplished via Excel's NORMINV function). Since our classification criterion necessitates a measure for both the antennae and abdomen of a fly, and since we assume that these values are correlated in some way, we used Desmos's regression tool to fit a linear regression model between the Aa abdomen and antennae datasets and between the Ax abdomen and antennae datasets. For half of the 100 generated flies (25 of which were Aa and 25 of which were Ax), we generated a pseudo-abdomen measurement using the method described above. Then, we used our Aa linear regression model to predict an antennae measurement for that abdomen measurement. However, it is unrealistic to assume that antennae measurements are perfectly predicted by abdomen measurements, as evidenced by the residual values of our regression models. To account for this, we can add synthetic residuals to our predicted values, which can, in turn, be generated by a normally distributed random number according to the normal distribution of the residuals for the regression model. We determined that this method of generating synthetic residuals is reasonable by using normal probability plots for each of our sets of residuals, which indicated that the residuals were again well-approximated by normal distributions (Figure 6). We called this group, with randomly generated abdomen measurements and predicted antennae measurements with added noise, abdomen-first pseudoflies.

Figure 6. Normal probability plots for residuals of abdomen-antennae linear regression models





We repeated this process for the remaining 50 flies (again, 25 Aa and 25 Ax), this time using the generated antennae values to predict abdomen values and adding normally-distributed random synthetic residuals as before. We called this group antennae-first pseudoflies.

It is important to distinguish between the abdomen-first and antennae-first pseudoflies because their method of generation is fundamentally different, despite producing similar results. The regression model of the abdomen-first pseudoflies attempts to minimize the vertical distance between the points (a_{bd}, a_{nt}) and the abdomen-first regression line. By contrast, the regression model of the antennae-first pseudoflies attempts to minimize the vertical distance between the points (a_{nt}, a_{bd}) and the antennae-first regression line, or, in other words, the horizontal distance between the points (a_{bd}, a_{nt}) and the inverse of the antennae-first regression line. This produces two distinct fits, neither of which will produce fully accurate pseudoflies given the limited sample data and inherent complexity of living organisms. To minimize bias, half of our pseudoflies are abdomen-first and the other half are antennae-first.

After creating our 100 pseudo flies, visual inspection showed no significant discrepancies between pseudo- and real flies (Figure 7) nor abdomen- and antennae-first pseudoflies (Figure 8). Table 2 shows sample values and calculations for each fly variant and method.

Table 3: Generation of Pseudoflies with Sample Values				
Variant	Abdomen	Antennae	Q	Method

Aa	1.89	1.58	1.20	Abdomen-first $\bar{x} = 1.822,\ \sigma_{x} = 0.080,\ \sigma_{err} = 0.094$ $a_{bd} = randNorm(\bar{x},\ \sigma_{x})$ $a_{nt} = 0.524x + 0.556 + randNorm(0, \sigma_{err})$
Aa	1.79	1.43	1.25	
Aa	1.71	1.42	1.48	Antennae-first $\bar{x} = 1.511,\ \sigma_{x} = 0.103,\ \sigma_{err} = 0.073$ $a_{nt} = randNorm(\bar{x},\ \sigma_{x})$ $a_{bd} = 0.317x + 1.343 + randNorm(0, \sigma_{err})$
Aa	1.75	1.40	1.57	
Ax	2.00	1.34	1.20	Abdomen-first $\bar{x} = 1.933,\ \sigma_{x} = 0.105,\ \sigma_{err} = 0.042$ $a_{bd} = randNorm(\bar{x},\ \sigma_{x})$ $a_{nt} = 0.575x + 0.166 + randNorm(0, \sigma_{err})$
Ax	1.75	1.11	1.26	
Ax	1.83	1.23	1.48	Antennae-first $\bar{x} = 1.277,\ \sigma_{x} = 0.074,\ \sigma_{err} = 0.060$ $a_{nt} = randNorm(\bar{x},\ \sigma_{x})$ $a_{bd} = 1.176x + 0.431 + randNorm(0, \sigma_{err})$
Ax	1.85	1.24	1.48	

Figure 7. Antennae vs. Abdomen of Pseudo- and Real Flies, where dots represent Aa flies and Xs represent Ax flies

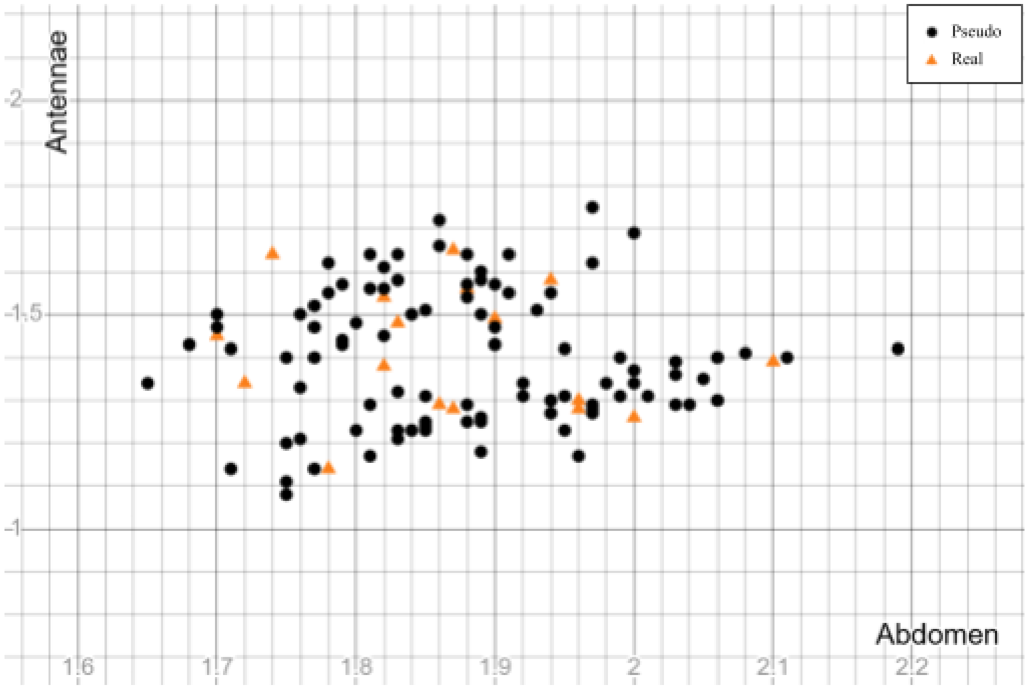
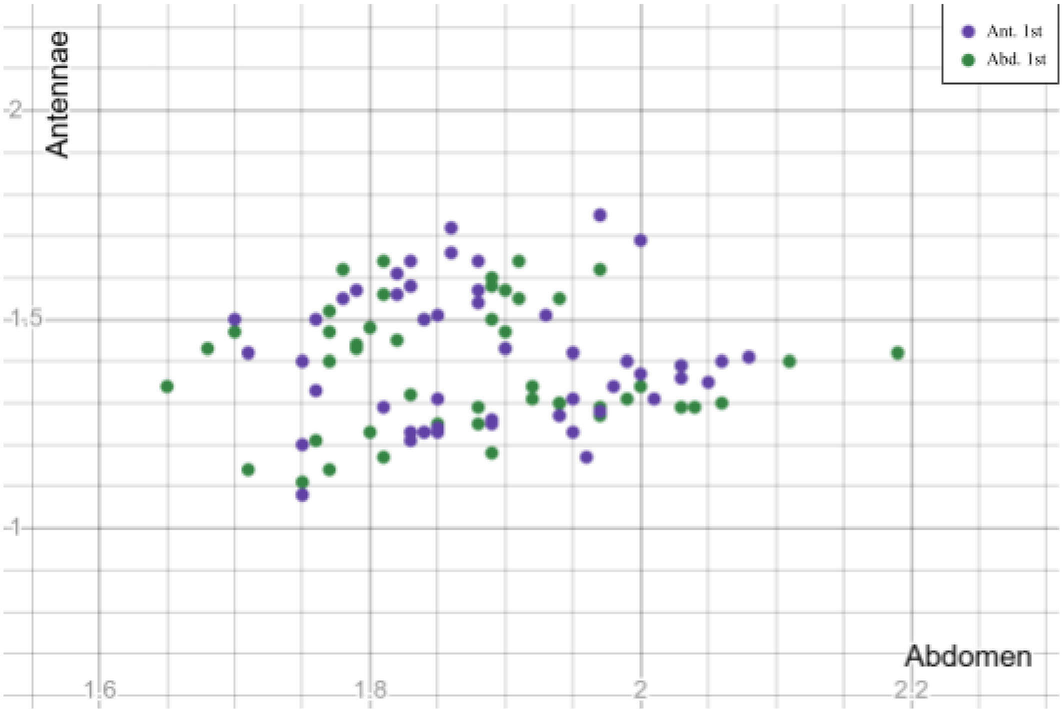


Figure 8. Antennae vs. Abdomen of Abdomen-First and Antennae-First Pseudoflies, where dots represent Aa flies and Xs represent Ax flies



Evaluating Model using Pseudoflies

Now that we have a reasonable method of generating pseudoflies, we can evaluate our classification method on a wider set of values than our original data. With an Ax fly being considered “positive” (for the disease-carrying gene), we can evaluate the performance of our model using true positive rate (TPR), false positive rate (FPR), precision (P), and accuracy (A). These metrics are defined as

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

$$P = \frac{TP}{TP + FP}$$

$$A = \frac{TP + TN}{TP + FP + TN + FN}$$

where TP is the number of true positives (Ax flies classified as Ax), FN is the number of false negatives (Ax flies classified as Aa), FP is the number of false positives (Aa flies classified as Ax), and TN is the number of true negatives (Aa flies classified as Aa). We are primarily concerned with the TPR of our model, as it is preferable to assume a fly is more dangerous than it really is and then take proper precautions rather than assume it is safe when it has the capacity to inflict disease. We are secondarily concerned with the accuracy of our model, as it is important that scientists have accurate data when studying the biting flies.

The performance of our threshold on the 100 generated pseudoflies is as follows:

Table 4. Performance of Model on 100 Generated Pseudoflies

	Actually Aa	Actually Ax	Total		TPR	FPR
Classified as Aa	47	0	47		1.0	0.06
Classified as Ax	3	50	53		Precision	Accuracy
Total	50	50			0.94	0.97

The model performs well according to our metrics on the generated pseudoflies. However, 100 flies is a small sample size and thus more likely to be biased. To obtain a fuller picture of our model’s performance, we generated 1000 more pseudoflies using the same method and evaluated them in the same way:

Table 5. Performance of Model on 1000 Generated Pseudoflies

	Actually Aa	Actually Ax	Total		TPR	FPR
Classified as Aa	479	1	480		0.998	0.042
Classified as Ax	21	499	520		Precision	Accuracy
Total	500	500			0.960	0.978

The model performs similarly on a larger set of pseudoflies. Our TPR and accuracy are very high, as desired, and our results suggest that very few Ax flies will be classified as Aa.

Classifying 3 new flies

Table 6: Measurements and analysis of 3 testing flies

	Wing	**Abd**	**Ant.**	**Abd/Ant**	**Species categorization**
Fly 1	2.81	1.80	1.24	1.45	Ax
Fly 2	2.65	1.84	1.28	1.44	Ax
Fly 3	3.61	2.04	1.40	1.46	Ax

We can now apply our threshold to the 3 flies for which we have been given the measurements. We can easily find Q for each, as shown in column 4 of Table 3. Since all 3 ratios fall above our threshold of 1.36, we categorize each of these flies as Ax flies (Table 6).

Discussion

Our model has some weaknesses that should be considered in its use:

1. We purposefully have biased our model at the border between Ax and Aa fly classifications. If our model is used to update itself—using our classification metric to

determine fly species and using those species to refine the model—it will reinforce the bias. Over time, this would make the model less and less accurate, and eventually lead to the subsumption of Aa flies into the Ax classification.

2. Our model may be overfit to our small data set. When we are trying to differentiate Ax and Aa flies, we may have some error because of the small data set size that doesn't accurately represent the full species range. To measure the level of this overfit, we calculated the standard error of the mean of each data set.

$$SEM = \frac{s}{\sqrt{n}}$$

$$SEM_x = \frac{0.0526}{\sqrt{7}} = 0.0199$$

$$SEM_a = \frac{0.0779}{\sqrt{10}} = 0.0246$$

   These measurements indicate the expected difference between our sample mean and the true population means. The lower value for the Ax flies indicates that our Ax mean is more accurate than our Aa mean, even though the Aa data set is larger.

3. The initial data we were provided may have been subject to error. Biologists may have only been able to catch some of the flies (such as the slowest), which may result in sampling bias and results from our model that do not reflect reality.

4. The distribution of ratios in our pseudoflies model was well-approximated by a normal curve, but was not exactly in a normal distribution. Our normal probability plots generally support the normal distribution, so the normally distributed random numbers reasonably reflect the data, but the pseudoflies' distribution will not exactly match that of the real flies. While this does not affect our threshold, which was determined by only the real flies, it may affect our model's performance on the pseudoflies (e.g., if the Ax flies were skewed to the left of the normal distribution, false positives would be more likely).

5. Other regressions than linear regression might be more accurate. We chose linear regression for our pseudoflies because it is simple and reasonably accurate; however,

other forms of regression may have performed better by, e.g., accounting for curvature in our data.

Conclusion

A single, field-ready ratio (Q) can be used to accurately classify the current Aa and Ax samples. We developed a simple classification rule to distinguish Ax and Aa. When $Q \geq 1.36$ (three SD below the Ax mean), the fly can be identified as Ax. Validation with 100 and 1000 pseudoflies confirmed robustness: accuracy around 98% with true positive rate of 1.00 and 0.998, respectively. Applied to three additional flies (Q=1.45, 1.44, 1.46), all were classified as Ax. Our model demonstrates a two-measurement, low cost classifier for field use. Future updates should add only independently verified labels, recompute per-species means and standard deviations, and reapply the same Ax-favoring threshold.

Contributions

Adam - Modeling classifier; Determining fly time, guide for biologists, classifying 3 new flies, and discussion sections; tables 1, 2, and 6 and figures 1-3; editing

Emme - Pseudofly generation & evaluation; figures 4-8, tables 3, 4, & 5; editing

Cynthia - Generating pseudoflies; assumptions; editing; conclusion

Hadi - Exec Summary, Assumptions, Conclusion, Editing

Lavigne’s Comments

You’ve presented a simple classifier on a single transformed feature that is reasonable and effective. You’ve taken care to respect the usability of the classifier in the field. You’ve validated the model using synthetic data that was very cautiously prepared with a novel approach. The discussions were very effective.

The writing here is very effective. You write with a mature voice, use vocabulary appropriately, and introduce new terminology that is clear and appropriate in register. Your results are communicated in figures and tables. An area for growth, if I had to point to one, would be to finesse the figures a bit more. The word “we” appears 46 times in this write-up.