Investigating Factors into Squirrels' Attitudes towards Humans in New York

Team BCHZ - Nick Bayer, Richard Cui, Laura Han, Anna Zhang

2023-12-13

Introduction and Data

As a result of continuous and ever-expanding human development, animals must inevitably interact with humans more often. According to prevailing scientific theory, these interactions often disrupt animals' natural behaviors; since animals see humans as threats, it is no surprise that they would treat the presence of humans the same way they would the presence of other predators [1]. However, recent studies show that the squirrels may undergo the process of phenomenon called synurbization, or the process of adapting to an urbanized environment through changes in their natural behavior [2].

In an effort to investigate these two competing theories, and to better understand the dynamic between squirrels and humans, we carry out this research project to determine the factors affecting squirrel's indifference to human presence. From there, we will deduce whether the squirrels' attitude to humans are caused by human presence or other natural factors such as species or age.

We hypothesize that the age category, location, distance above ground when spotted, number of activities that the squirrel was doing, sound that the squirrel makes, and tail signs could have a relationship with the attitude of the squirrel (whether indifferent or not), providing valuable insight into the validity of "synurbanization".

Data Description

We are sourcing our data set from the TidyTuesday project on GitHub. The data originally came from The 2018 Squirrel Census, a project based on the sightings of Eastern gray squirrels (Sciurus carolinensis) in Central Park, New York City.

In October of 2018, the Squirrel Census Team and a group of over 300 volunteers collected the data based on squirrel sightings around Central Park. The data was collected between October

06, 2018 and October 20, 2018 during both the A.M. and P.M. across Central Park. Each observation is one squirrel, with a Unique ID of [hectare-shift-date-hectare squirrel number]. Almost every observation is a unique squirrel with the exceptions of a few duplicate entries, which is resolved in the data cleaning section below.

The data set has 3023 observations and 31 variables and contains a wide range of situational factors and squirrel characteristics. We are given the location, in both longitude and latitude, the hectare of the park the squirrel was located in, the date, and whether it was found in the AM or PM. In addition to assigning the squirrel a unique ID, the data also has information on whether the squirrel is an adult or a juvenile, its primary and highlight fur colors, and the sequence of sightings in one session. The dataset then records each squirrel's exact location, their distance from the ground, and the objects they were found on. Additionally, there is data for the activities the squirrel was observed doing, ranging from running to foraging, with a separate column for all other activities not specifically mentioned. It gives data on the sounds the squirrel made and tail movements, if any. Finally, it has 4 separate columns for the squirrel's response when approached by humans: either the squirrel approached, was indifferent, ran away, or any other action.

Data Cleaning and Initial Exploratory Data Analysis

Based on the data description, each squirrel is assigned a unique squirrel ID. We first check if any squirrel is observed or recorded multiple times and found that there are 5 squirrel ID's that appeared multiple times in the dataset (details in appendix). We then examined each of the five duplicated squirrel entries and found that they were simply duplicate observations with very slight differences (after 4 decimal places) in latitude and longitude (an example is shown in appendix). Since the difference is negligible and longitude and latitude are not one of our hypothesized predictors, we simply remove all the duplicate entries from the data set.

For variable `Age`, there is an unwanted level of "?", which is probably due to the error during data collection. Therefore, we will drop all observations with `Age == "?"` (a total of 125 observations, which is acceptable given our original data size of 3018).

We then create a numeric variable named `Activity_Score` that encapsulates the aggregate number of activities a squirrel is engaged in during the observation period. The distribution of the `Activity_Score` variable is unimodal, slightly right skewed, and has a median of 1. This means that most squirrels in the data set were only engaged in 1 activity during the span of observation.

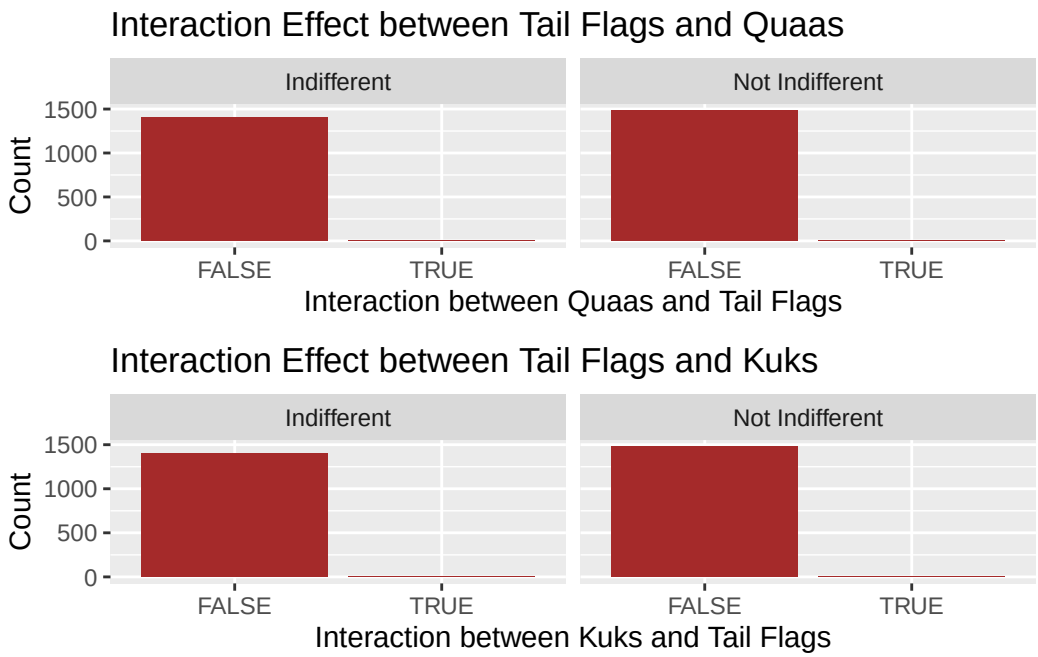
Our EDA shows that there seems to be a roughly equal number of squirrels that showed indifference to humans and those that didn't. However, there were far fewer squirrels that approached humans than squirrels that didn't. Thus, we decided that using Indifference vs. Not Indifferent would likely result in a model with a better model than using Approached vs. Not Approached.

The distribution suggests that squirrels could show that they are not as synurbanized as we thought.

The distribution of `Activity_Score` for squirrels that were Indifferent vs. Not Indifferent is roughly the same shape. Therefore, `Activity_Score` and `Indifferent` do not seem to be correlated.

Since there were so few squirrels that approached the researchers, it is hard to tell whether the graphs are very different in shape. The difference may be something that we explore further.

An interaction that we would like to examine further is whether the presence of `Tail flags`, combined with `Quaas` and `Kuks`, influences the likelihood that a squirrel is indifferent to humans. Tail flags are a waving motion that squirrel to exaggerate their size and confuse rivals or predators. `Quaas` are elongated vocal communications that squirrels use to indicate the presence of a ground predator. Meanwhile, kuks are a chirpy vocal communication used for a variety of reasons. Therefore, examining the interaction between tail flags and quaas, and more broadly, kuks, could help us understand how squirrels perceive humans and how synurbanization affects that perception. In both graphs below, there are more counts of FALSE for the interaction term between both `Tail Flags` and `Quaas` as well as between `Tail Flags` and `Kuks` (1600) than TRUE (1400). When we analyze the distribution of TRUE and FALSE for Indifferent vs. Not Indifferent, we see that the number of FALSE values is approximately the same, indicating that the interaction term may not be correlated with `Indifferent`.



Methodology

Analysis approach

Our response variable is `Indifferent`, which is a categorical variable that indicates whether or not the squirrel is indifferent to humans (meaning that the squirrel does not run away). Potential predictors include `Activity_Score` (a quantitative variable that records the number of activities the squirrel is observed doing), `Age` (a categorical variable that indicates whether the squirrel is adult or juvenile), `Primary Fur Color` (categorical variable), `Above Ground Sighter Measurement` (quantitative variable), sounds that the squirrels are making (categorical variables including `Kuks`, `Quaas`, and `Moans`), `Tail flags` and `Tail twitches` (also categorical).

To explore the relationship between whether or not the squirrel is indifferent and predictor variables, such as age, distance above ground when spotted, number of activities the squirrel is observed doing, sound that the squirrel makes, and tail signs, we plan to use a logistic regression. We are using logistic regression because our dependent variable, `Indifferent,` is categorical, and we believe that the log odds of a squirrel being indifferent has a linear relationship with the predictor variables we identified above. We will compare logistic regression models using AIC and BIC to evaluate the predictor variables and interaction term should be included in the model to best predict the attitude of the squirrel towards humans. We will also perform 10-fold cross-validation for model comparison.

Model 1

Model 1 includes all variables stated in our hypothesis.

Recipe for Model 1 steps: 1) Change response variable into factors. 2) Map all "FALSE" in `Above Ground Sighter Measurement` variable to 0, then convert the variable type to integer. 3) Create dummy variables for all nominal predictors. 4) Create interaction terms between `Quaas` and `Tail flags` 5) Remove all variables with zero variance.

.metric	.estimator	mean	n	std_err	.config
accuracy	binary	0.522	10	0.011	Preprocessor1_Model1
roc_auc	binary	0.529	10	0.012	Preprocessor1_Model1

mean_aic	mean_bic
2776.198	2842.858

Model 2

As the EDA has shown, `Activity_Score` does not seem to have a relationship with `Indifferent`, so we take it out for Model 2. We conduct a drop-in-deviance test that confirms that the coefficient of `Activity_Score` is not statistically significant from 0 because the p-value of the test (0.87) was greater than 0.05.

Recipe for Model 2 steps: 1) Change response variable into factors 2) Map all "FALSE" in `Above Ground Sighter Measurement` variable to 0, and not "FALSE" to 1, then convert the variable type to factors. 3) Create dummy variables for all nominal predictors 4) Remove all variables with zero variance.

As observed below, Model 2 has lower AIC ($2774.98 < 2776.20$) and BIC ($2836.04 < 2842.86$) than Model 1. It also has marginally higher accuracy ($0.528 > 0.522$) and higher AUC ($0.537 > 0.529$). Therefore, Model 2 is the better model and we will use it for the rest of our analysis.

.metric	.estimator	mean	n	std_err	.config
accuracy	binary	0.528	10	0.012	Preprocessor1_Model1
roc_auc	binary	0.537	10	0.013	Preprocessor1_Model1

mean_aic	mean_bic
2774.984	2836.042

Drop-in-deviance p-value
0.869

Multicollinearity

Table 6: VIF Table of Predictors

	VIF
`Above Ground Sighter Measurement`	1.011906
KuksTRUE	1.034798
QuaasTRUE	1.153590
MoansTRUE	1.000000
`Tail flags`TRUE	1.032303
`Tail twitches`TRUE	1.006639
Age_Juvenile	1.007001

	VIF
`Primary Fur Color_Cinnamon`	4.410060
`Primary Fur Color_Gray`	4.411834
`QuaasTRUE_x_\Tail flags‘TRUE‘`	1.162523

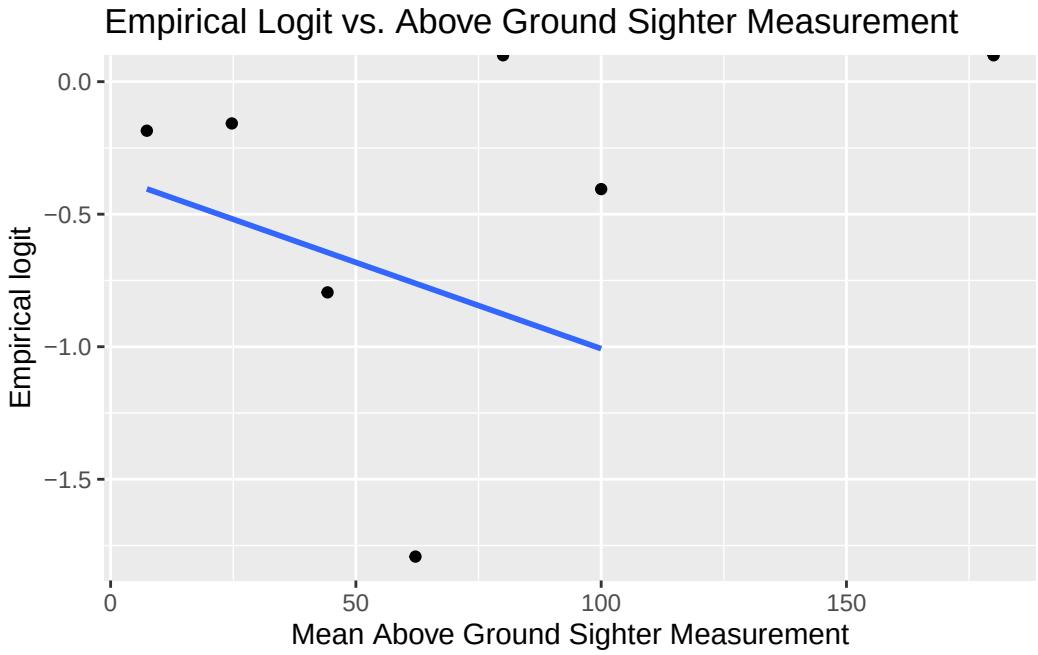
From the VIF table, there are no concerning VIF values (greater than 10), so we do not need to adjust for any multicollinearity issue in our model.

Model Conditions

We will check model conditions for Model 2 since this is the better performing model due to higher accuracy and AUC and lower AIC and BIC.

Linearity

There is one numeric variable in Model 2, `Above Ground Sighter Measurement`, so we check its linearity using empirical logit.



The linearity condition is satisfied. There is not an obvious non-linear relationship between the empirical logit and the predictor variable of `Above Ground Sighter Measurement`.

We could improve this condition if we had more specific data with fewer NA values. However, since there are currently no obvious non-linear patterns between `Above Ground Sighter`

`Measurement` and the empirical logit, it is reasonable to conclude that the linearity condition is satisfied.

Randomness

The data was collected from the sightings of squirrels in Central Park, NYC from a group of volunteers. Although the squirrels were not randomly sampled, the sample of squirrels can be considered as random since we do not have reason to believe that the characteristics of squirrels collected in this study differ systematically from squirrels in other urban areas in regards to their indifference to humans. Since the squirrels in the study are Eastern gray squirrels, the population would be squirrels living in major urban areas in the Eastern half of the United States (the squirrel's natural habitat range). Therefore, randomness is satisfied.

Independence

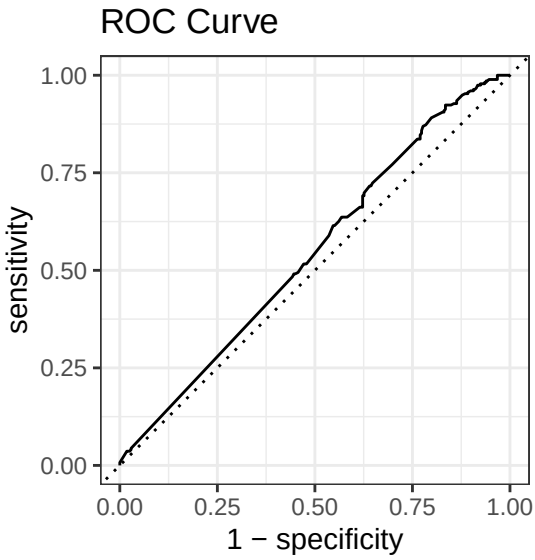
The data are not spatially or time-correlated since the data were collected in one-time snapshot and in one location. Therefore, we have no reason to believe that independence was violated.

Thus, all model conditions for logistic regression are satisfied.

Results

Since Model 2 has a higher accuracy and AUC, as well as lower AIC and BIC, we will choose Model 2. Next, we fit the model to the entire squirrels dataset and interpret the coefficients.

term	estimate	std.error	statistic	p.value
(Intercept)	-0.190	0.237	-0.802	0.423
`Above Ground Sighter Measurement`	-0.009	0.005	-1.976	0.048
KuksTRUE	-0.378	0.256	-1.474	0.141
QuaasTRUE	-0.584	0.414	-1.410	0.158
MoansTRUE	-12.315	324.744	-0.038	0.970
`Tail flags`TRUE	0.216	0.196	1.100	0.271
`Tail twitches`TRUE	-0.044	0.119	-0.371	0.711
Age_Juvenile	-0.229	0.137	-1.668	0.095
`Primary Fur Color_Cinnamon`	0.046	0.262	0.175	0.861
`Primary Fur Color_Gray`	0.275	0.239	1.150	0.250
`QuaasTRUE_x_`\Tail flags‘TRUE‘	1.892	1.210	1.563	0.118



.metric	.estimator	.estimate
roc_auc	binary	0.546

```
          Truth
Prediction FALSE TRUE
     FALSE   116  100
     TRUE    162  175
```

Of all the variables we examined, `Above Ground Sighter Measurement` was the only term that had coefficients with significant p-values at $\alpha = 0.05$ significance level. For every additional 1 meter in the squirrel's location above ground, the odds that it is indifferent is multiplied by a factor of 0.991 ($\exp(-0.009)$), holding all else constant. Although `Adult_Juvenile` is not significant at the $\alpha = 0.05$ level, it is significant at the $\alpha = 0.10$ level. This means that the odds that a juvenile squirrel is indifferent to a human is expected to be 0.795 ($\exp(-0.229)$) time the odds for an adult squirrel, holding all else constant. In other words, being a juvenile squirrel and being higher off the ground decrease the odds that a squirrel is indifferent to humans.

The model does not have a high predictive power as shown by the low AUC (0.546) and accuracy (0.526, as calculated by $\frac{116+175}{116+100+162+175}$). This is somewhat expected given what we saw in the EDA section, where the distribution of the predictors is either similar across the two levels of response or very imbalanced.

Discussion and Conclusion

In this paper, we have investigated which factors significantly affect whether a squirrel is indifferent to human presence, using a logistic regression model and a 10-fold cross-validation for model selection. We concluded that based on our logistic model, there is only one significant predictor of `Indifferent` at $\alpha = 0.05$ significant level, which is `Above Ground Sighter Measurement`, and another significant predictor at $\alpha = 0.10$ significance level, which is `Age_Juvenile`. Putting this into a larger context, this shows that the squirrels' attitude to humans are largely caused by natural factor since `Above Ground Sighter Measurement` is considered a "neutral" variable (i.e. not particularly correlated with either human presence or natural factors) and `Age`, which is believed to be a natural factor of each squirrel, turned out to be a significant predictor.

Surprisingly, both `Tail Flags`, which occurs when squirrels identify a threat and seek to confuse it, and `Tail Twitches`, which indicate curiosity, were not significant in the model. The insignificance of these variables may indicate that the squirrels do not see humans as a threat or threat do not contribute much to their attitude of being indifferent or not.

Although `Above Ground Sighter Measurement` had a statistically significant coefficient at $\alpha = 0.05$, the actual value of the coefficient is near-zero. The fact that it is negative is difficult to interpret because it suggests that the higher the squirrel is from the ground, the lower the odds of it being indifferent to the researchers. However, logically, we might argue that the higher the squirrel is off the ground, the safer it should feel from the researchers. Or perhaps, there are confounding factors: maybe the squirrels on trees are skittish due to another threat.

One important limitations is the concerns about the way data is collected. For example, based on the data description, the data is collected from volunteer sighting in NYC. This might introduce some lack of rigorousness, especially when it comes to numeric variables like `Above Ground Sighter Measurement` since the volunteers might give a very subjective estimation of the height above the ground when standing and observing from distance. We have considered turning `Above Ground Sighter Measurement` into a categorical variable, but we have found that this further decreased our model's predictive power, potentially due to the fact that we are losing information when mapping this variable to a true/false variable. In addition, there are a lot of NA's in the dataset, which to some extent limits our ability to do analysis.

For future work, we will try to extract more useful predictors from feature selection and feature engineering steps in order to increase the predictive power of our model. For example, instead of simply dropping the longitude and latitude from the dataset, we could potentially search the longitude and latitude of downtown NYC and use this information to calculate the Euclidean distance (or perhaps Manhattan distance since it's in NYC) between each observation and the city center for the model. In addition, we could try some non-linear machine learning models, such as random forest and boosting, since the linear model seem not to be able to give accurate predictions in this case.

Appendix

Data Cleaning for Duplicate Squirrels

The squirrels with repeated observations are summarized in the following table.

Table 9: Squirrels Observed More Than Once

Unique Squirrel ID	count
1F-AM-1010-04	2
37E-PM-1006-03	2
40B-AM-1019-06	2
4C-PM-1010-05	2
7D-PM-1010-01	2

The duplicate rows for one such squirrel is reported below (results are truncated to only show the first 6 columns due to space limitations, but we have checked that all variables are the same besides the slight latitude and longitude difference described above).

Table 10: Duplicate Entries for Squirrel 1F-AM-1010-04

X	Y	Unique Squirrel ID	Hectare	Shift	Date
-73.97662	40.76619	1F-AM-1010-04	01F	AM	10102018
-73.97659	40.76609	1F-AM-1010-04	01F	AM	10102018