A magnifying glass with a snake

Description automatically generated

Predicting Reptile, Amphibian, and Arachnid Activity with Observational Data from the “iNaturalist” Application

Matthew Zingarella

Quinnipiac University

Revised 05/18/2023

Purpose of Document

This report describes the use of data on reported encounters with animals retrieved from the iNaturalist website/application in conjunction with weather data retrieved from the National Oceanic and Atmospheric Administration (NOAA) with the goal of gleaning insights on animal activity patterns and factors affecting the likelihood of observations. Specifically, the likelihood of observation of reptiles, amphibians, and arachnids in the New England area is examined with consideration to temperature, rainfall, and seasonal patterns. A metric, observation score, is developed to make these assessments and evidence for the efficacy of the metric is presented.

Outline of Document

Executive Summary………………………….……………………………………………………2 Section One: Description of the Datasets Utilized…………………………………………….….3 Section Two: Description of the ‘Observation Score’ Metric……………….…………………......5 Section Three: Assessing the Reliability and Limitations of ‘Observation Score’………………..5 Section Four: Future Directions for Project…………………………………………...…………8 Appendix…………………………………………………………………………………………..9….. All Stored Procedures……………..….............……...…………..............................................9 ………Procedures for Iconic Taxa…...……………………………………………………………9 …………sp\_GetTaxonObsByDate…………………………………………………………..……9…………sp\_GetTaxonObsByRain………………………………………………………..………9…………sp\_GetTaxonObsByTemp………………………………………………………..……10…………sp\_GetTaxonObsScoreByDate……………………………………………………..…10…………sp\_GetTaxonObsScoreByRain…………………………….………………………..…11…………sp\_GetTaxonObsScoreByTemp…………………………………………….…………12………Procedures for Species..…………………………………………………………………12 …………sp\_GetSpeciesImage………………………………………………………………..…12 …………sp\_GetSpeciesObsByDate………………………….………………...……….………13 …………sp\_GetSpeciesObsByRain………………………………………………………….…13…………sp\_GetSpeciesObsByTemp……………………………………………………………14…………sp\_GetSpeciesObsScoreByDate………………………………………………………14…………sp\_GetSpeciesObsScoreByRain………………………………………………………15…………sp\_GetSpeciesObsScoreByTemp…………………………….…….………………….16………Misc..Procedures…….……………………..…………………………………….....……17…………sp\_ChangeObsScores………………...………………….……………………….........17…………sp\_GetTotalObservations………...…….……………………………………………...18 …………sp\_GetObsByDayOfWeek…………………………………………………...….…….18 …………sp\_GetObsByYear…………………………………………………………….……….18 ………Functions…………………………………………………………………………………19 …………GetRainStats…………………………………………………………………………..19 …………GetTempStats………………………………………………………………………….19 User Privileges……………………………………………………………..………………...…..20

Executive Summary

Different species of animals tend to show very different patterns of activity based on seasonal factors and the weather conditions of a given day. This information is of interest not only to ecologists, but also the many hikers and naturalists who take enjoyment from the observation of nature. There is therefore value in the ability to attain data-driven insights into the impact of different weather conditions on the likelihood of observing a particular species of animal on a given day.

This report describes a method to generate such insights utilizing species observation data from iNaturalist and weather data from the National Oceanic and Atmospheric Administration (NOAA). iNaturalist is a website/mobile application in which users, typically hikers/nature enthusiasts, upload pictures of wildlife along with the date/location at which the observation occurred. Other users then receive the picture, and if they believe they know the species pictured, they submit an identification. If enough users submit an identification without disagreements, the identification is deemed “research-grade.” The iNaturalist dataset utilized in this report contained approximately 150,000 research-grade observations of reptile, amphibian, and arachnid species made in the New England area.

In section one I describe the datasets utilized and discuss the difficulty in inferring animal activity from this dataset, which is heavily biased by the patterns of those who submit observations. For example, the dataset contains a much larger proportion of observations on Saturdays and Sundays than it does on weekdays. When observations are weighed against factors like daily high temperatures, little variation is seen between reptiles, amphibians, and arachnids despite the undoubtedly different activity patterns of the taxa, which this data suggested all seem to prefer days that are sunny, 75 degrees, and not raining. From this information, it was inferred that the number of observations alone is not representative of animal activity, but the activity of the human observers reporting the encounters.

In section two I discuss a metric, coined “Observation Score (OS),” developed to address this bias and glean meaningful insights from the data. OS is calculated by coming up with an estimation of observer activity based on the number of unique users that submit observations on a given day, then weighing the significance of individual observations against the estimated level of observer activity on that day.

In section three I argue for the legitimacy of the OS metric by discussing the observable differences between the relation of raw observations to various factors and OS to the same factors. In general, more distinct patterns were seen between taxa and species in varying weather/seasonal conditions when OS is used.

In section four I discuss ways the calculation of OS might be improved in future versions of this database. I also discuss ways in which more weather data can be included to the database to improve the insights.

Section One: Describing the Datasets Utilized

The main dataset utilized to generate the insights in this report was composed of approximately 150,000 observations of reptiles, amphibians, and arachnids in the New England Area. The years the observations took place can be seen in Figure 1, below.

A picture containing text, screenshot, line, plot

Description automatically generated

**Figure 1. The total number of observations of reptiles, amphibians, and arachnids in the New England area on the iNaturalist app from 2007 to 2022. Dates before 2007 have been excluded as they contained under 5 observations.**

The bulk of the observations occurred after the year 2020, with the app spiked in popularity in late 2019. Interestingly and perhaps not surprising, this corresponds with the crowding of state parks that occurred during the COVID pandemic. In addition to having less data available, the data contained in earlier years was also less robust, having greater uncertainty in locational accuracy and many null values. For this reason, everything presented from this point forward contains only data from the 2018-2022 period.

An interesting observation made on the iNaturalist dataset, and the main obstacle that the later half of this report attempts to overcome, is that the data is heavily influenced by the activity patterns of the humans recording the encounters. Take Figure 2, below.

A picture containing text, screenshot, number, font

Description automatically generated

**Figure 2. Observations of reptiles, amphibians, and arachnids in the New England area on the iNaturalist app grouped by the day of the week**

A picture containing text, screenshot, font, plot

Description automatically generatedA picture containing text, screenshot, line, plot

Description automatically generatedAssuming that animal activity is not in fact influenced by whether it is a week day or weekend, this figure demonstrates the importance of considering the bias introduced to this publicly gathered data. In Figure 3, below, this bias presents itself more subtlety.

**Amphibians**

**Reptiles**

**Arachnids**

A picture containing text, screenshot, line, plot

Description automatically generated

**Figure 3. Observations of reptiles, amphibians, and arachnids in the New England area on the iNaturalist app grouped by the frequency with which they were observed on days with daily high temperatures between 0 and 100 degrees Fahrenheit in 5 degree increments**

Despite that reptiles, amphibians, and arachnids are animals with distinctly different patterns of activity at different daily temperatures, if one looks at the number of observations in iNaturalist alone they will see approximately the same patterns for each iconic taxon when compared to temperature ranges, with the slight exception of amphibians which are a little less affected by mid-low temperatures than the other two taxa.

Working with the hypothesis that this lack of variation was the result of observer activity patterns, a metric was developed to try to overcome this and achieve meaningful insights on animal activity patterns and the relative likelihood of observation under specific conditions.

Section Two: Description of the ‘Observation Score’ Metric

The idea behind the metric developed to account for observer behavior influencing the dataset, observation score (OS), is to look at each individual observation and assess how heavily observer behavior has contributed to the probability of the observation occurring.

To do this, the number of unique iNaturalist user\_id’s that reported observations for each day over the 2018 to 2022 period was counted. For each observation an OS was then assigned by dividing 1 by the number of observed that contributed to the data that day. If 100 users contributed data on a given day, and 20 of those observations were tree frogs, each tree frog observation would have an OS of 1/100 for a total OS of 20/100 for the day. Additionally, an OS of 0 was assigned for any observation at occurred on a day when less than 3 users submitted data. This, for the most part, prevented the metric from resulting in wild outliers on the occasional days when only 1 user would submit data for the day (usually on winter days when it is highly unlikely anyone will be submitting observations, but someone finds a dead frog and uploads it to the app).

Section Three: Assessing the Reliability and Limitations of Observation Score

To assess whether OS is reliable for gleaning insights form the iNaturalist data, comparing figures that look at the same factors (temperature, rain, etc) for observation data and OS provides some context. Take Figure 4, below.

**Observation Score**

**Raw Observations**

A close-up of a graph

Description automatically generated with low confidence

**Figure 4. Comparison of the distribution of observation frequency over a range of daily high temperatures for amphibians when compared to the number of observations that occur on days in each temperature range vs the cumulative observation score for days in that temperature range**

By generating a plot that looks at the OS instead of just the number of raw observations a very different distribution was generated for amphibians. The most frequent temperature range decreased from 75-80 degrees to 55-60 degrees, which intuitively makes more sense for amphibians, many of which do not do well when conditions are hot and dry. The transformation was even more dramatic for arachnids as seen in Figure 5, below.

**Observation Score**

**Raw Observations**

A close-up of a graph

Description automatically generated with low confidence

**Figure 5. Comparison of the distribution of observation frequency over a range of daily high temperatures for arachnids when compared to the number of observations that occur on days in each temperature range vs the cumulative observation score for days in that temperature range**

While the one data point in the 15 to 20 degree temperature range is a clear outlier, probably as a result of OS giving too much weight to a low probability event, the temperature ranges for arachnids became much more evenly distributed. This figure reveals something important to note about arachnids in the dataset, which is that many of the observations submitted occur indoors. It is interesting that, excluding the outlying point, the distribution looks slightly bimodal, this might indicate that the first peak is the temperature at which spiders are most commonly seen indoors and the second peak is when they are most commonly seen outdoors.

Most notably, implementing the OS metric resulted in very different distributions between taxa in stark contrast to the similar distributions that resulted from raw observations. Combined with he fact that the resulting distributions seem logical, this suggests that observer behavior was indeed interfering with the observation data and OS was effective in negating it.

While it is interesting to look at whole taxa with OS, it also produces some interesting results with individual species. Take Figure 6, below, which is a report generated from the database utilizing OS to produce insights on the Red-Backed Salamander (*Plethodon cinereus*).

**Figure 6.** A screenshot of a graph

Description automatically generated with low confidence**A report on the Red-backed Salamander (*Plethodon cinereus*) that uses Observation Score and iNaturalist data to provide information on how daily high temperatures, rainfall, and the day of the year affect the likelihood of observation**

By using the OS, one can predict that they are more likely to encounter a red-backed salamander on a cooler day in the spring or fall. There is also a greater probability to see them on days when it is raining.

Overall, OS used in conjunction with the iNaturalist data appears to provide meaningful insights into the activity patterns of animals, although it does occasionally result in data points that are far outliers as was seen in Figure 5.

Section Four: Future Directions for Project

While the results generated by applying OS to the iNaturalist dataset certainly produce meaningful insights, there is still room to improve the way the metric is calculated. For one, the method by which outlying data points are prevented, which is by discounting days when there weren’t many users providing data, is only somewhat effective and ends up discarding a lot of data that might be useful. The charts presented in this report were generated from species/taxa that had thousands of entries in the database. When less observations are available the results tend to be much less meaningful. therefore, it would be better to calculate OS in a way that works without needing to discard data.

Furthermore, the method by which temperatures at which observations occurred was estimated also has a lot of room for improvement. The database contains two datasets of temperature observations between 2018 and 2022 recorded at a single point by two different weather stations. One of the stations was in Connecticut, the other in New Hampshire. The latitude/longitude data available for the observations as well as the weather stations was used to determine which station was closer to the observation, then that data was used. Obviously, this leaves a lot of room for error since there could be large differences in rainfall and temperature in areas distant from the stations. Incorporating more weather datasets would certainly refine the results generated.

Appendix

All Stored Procedures

*Procedures for Iconic Taxa*

**sp\_GetTaxonObsByDate**: Returns the cumulative number of observations that occurred on each day of the year (MM/DD) between 2018 and 2022.

A screenshot of a computer

Description automatically generated with low confidence

**sp\_GetTaxonObsByRain**: Returns the number of observations that occurred on days with above 0.1 inches of rainfall, and below 0.1 inches of rainfall, divided by the number of days above or below 0.1 inches of rainfall, respectively. Functions are shown in next section of appendix.

A screenshot of a computer program

Description automatically generated with low confidence

**sp\_GetTaxonObsByTemp**: Returns the cumulative number of observations that occurred on each day within 5 degree daily-high-temperature ranges spanning from 0 to 100 degrees divided by the number of days that occurred in each range between 2018 and 2022.

A picture containing text, screenshot, font

Description automatically generated

**sp\_GetTaxonObsScoreByDate**: Returns the cumulative number of observations that occurred on each day of the year (MM/DD) between 2018 and 2022. Dates without enough data (BadDates.too\_low = ‘Y’) are not included.

A screenshot of a computer code

Description automatically generated with low confidence

**sp\_GetTaxonObsScoreByRain**: Returns the cumulative OS that occurred on days with above 0.1 inches of rainfall, and below 0.1 inches of rainfall, divided by the number of days above or below 0.1 inches of rainfall, respectively. Functions are shown in next section of appendix.

A screenshot of a computer program

Description automatically generated with low confidence

**sp\_GetTaxonObsScoreByTemp**: Returns the cumulative OS that occurred within 5 degree daily-high-temperature ranges spanning from 0 to 100 degrees divided by the number of days that occurred in each range.

A screenshot of a computer program

Description automatically generated with medium confidence

*Procedures for Species*

**sp\_GetSpeciesImage:** Returns the url for one of the images submitted to iNaturalist for the given species

A screenshot of a computer code

Description automatically generated with medium confidence

**sp\_GetSpeciesObsByDate:** Returns the cumulative number of observations that occurred on each day of the year (MM/DD) between 2018 and 2022.

A picture containing text, font, line, screenshot

Description automatically generated

**sp\_GetSpeciesObsByRain:** Returns the number of observations that occurred on days with above 0.1 inches of rainfall, and below 0.1 inches of rainfall, divided by the number of days above or below 0.1 inches of rainfall, respectively. Functions are shown in next section of appendix.

A screenshot of a computer program

Description automatically generated with low confidence

**sp\_GetSpeciesObsByTemp:** Returns the cumulative number of observations that occurred on each day within 5 degree daily-high-temperature ranges spanning from 0 to 100 degrees divided by the number of days that occurred in each range between 2018 and 2022.

A picture containing text, font, screenshot

Description automatically generated

**sp\_GetSpeciesObsScoreByDate**: Returns the cumulative number of observations that occurred on each day of the year (MM/DD) between 2018 and 2022. Dates without enough data (BadDates.too\_low = ‘Y’) are not included.

A picture containing text, screenshot, font

Description automatically generated

**sp\_GetSpeciesObsScoreByRain**: Returns the cumulative OS that occurred on days with above 0.1 inches of rainfall, and below 0.1 inches of rainfall, divided by the number of days above or below 0.1 inches of rainfall, respectively. Functions are shown in next section of appendix.

A screenshot of a computer program

Description automatically generated with low confidence

**sp\_GetSpeciesObsScoreByTemp**: Returns the cumulative OS that occurred within 5 degree daily-high-temperature ranges spanning from 0 to 100 degrees divided by the number of days that occurred in each range.

A screenshot of a computer code

Description automatically generated with low confidence

*Misc. Procedures*

**sp\_ChangeObsScores**: This procedure contains the current method by which OS are calculated. Changing the method and then running the procedure will update OS in the observations table.

A picture containing text, screenshot, font

Description automatically generated

**sp\_GetTotalObservations:** Returns the total number of observations for a species

A screenshot of a computer code

Description automatically generated with medium confidence

**sp\_GetObsByDayOfWeek:** Returns the total observations in dataset grouped by the day of the week.

A screenshot of a computer program

Description automatically generated with low confidence

**sp\_GetObsByYear:** Returns the total count of observations in the dataset between 2007 and 2022 grouped by the year of the observation.

A screenshot of a computer program

Description automatically generated with low confidence

*Functions*

**GetRainStats:** Returns the number of days it was either raining or not raining for the given station and condition (raining/not\_raining) that is input between 2018 and 2022.

A screenshot of a computer program

Description automatically generated with medium confidence

**GetTempStats:** takes in a weather station name and the floor of a temperature range in 5 degree increments (10, 15, 20, etc) and returns how many days were observed within that 5 degree range at that weather station between 2018 and 2022.

A picture containing text, screenshot, font

Description automatically generated

User Privileges

Creates one user, ReportGenerator, who receives only the stored procedures necessary for generating the reports I made in SSRS. A second user, ObsScoreModifier, is something that would hypothetically help me look for new way to calculate ObsScore, and are thus granted a more generous range of permissions for looking at/changing certain tables

A screenshot of a computer

Description automatically generated with medium confidence