Crested Coralroot Orchid
© McCool-Grime

FLOWER HUNTING WITH INATURALIST

A BRAINSTATION CAPSTONE
BY LISA MCCOOL-GRIME

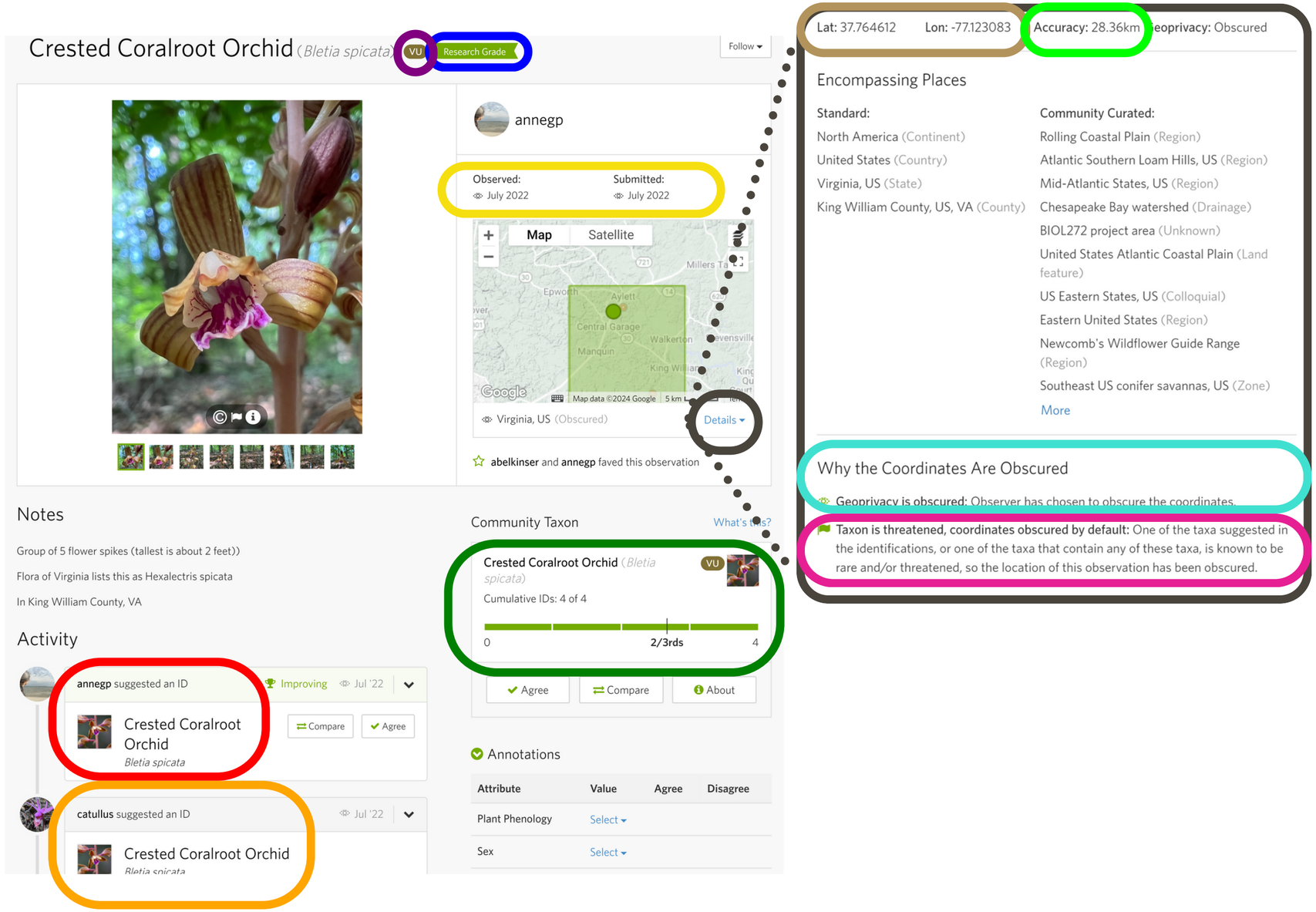
INTRO TO INATURALIST

iNaturalist is a "crowdsourced species identification system and organism occurrence recording tool".

Most often users upload pictures with timestamped and geolocated information about non-human living objects in the natural world. The user can **identify their own taxon guess**. Other users are able to see these observations and **add their own guesses about the taxon**. If two or more users ID one observation at the species level and **2/3s of the count is in agreement**, an observation becomes **research grade**.

If the ID has a **conservation status** that flags the taxon as vulnerable in the area it has been located, iNaturalist will flag the post as requiring a **"taxon geoprivacy" of obscured**. Individual users can also **flag any of their own posts as obscured** for personal reasons. If an observation is flagged as requiring an obscured geoprivacy, the publicly available **latitude and longitude** will be moved (around 26km or more) and the **public positional accuracy** of the observation will reflect this.

iNaturalist will also **granularize time data** to only month and year on their public-facing website when an observation has an obscured geoprivacy.



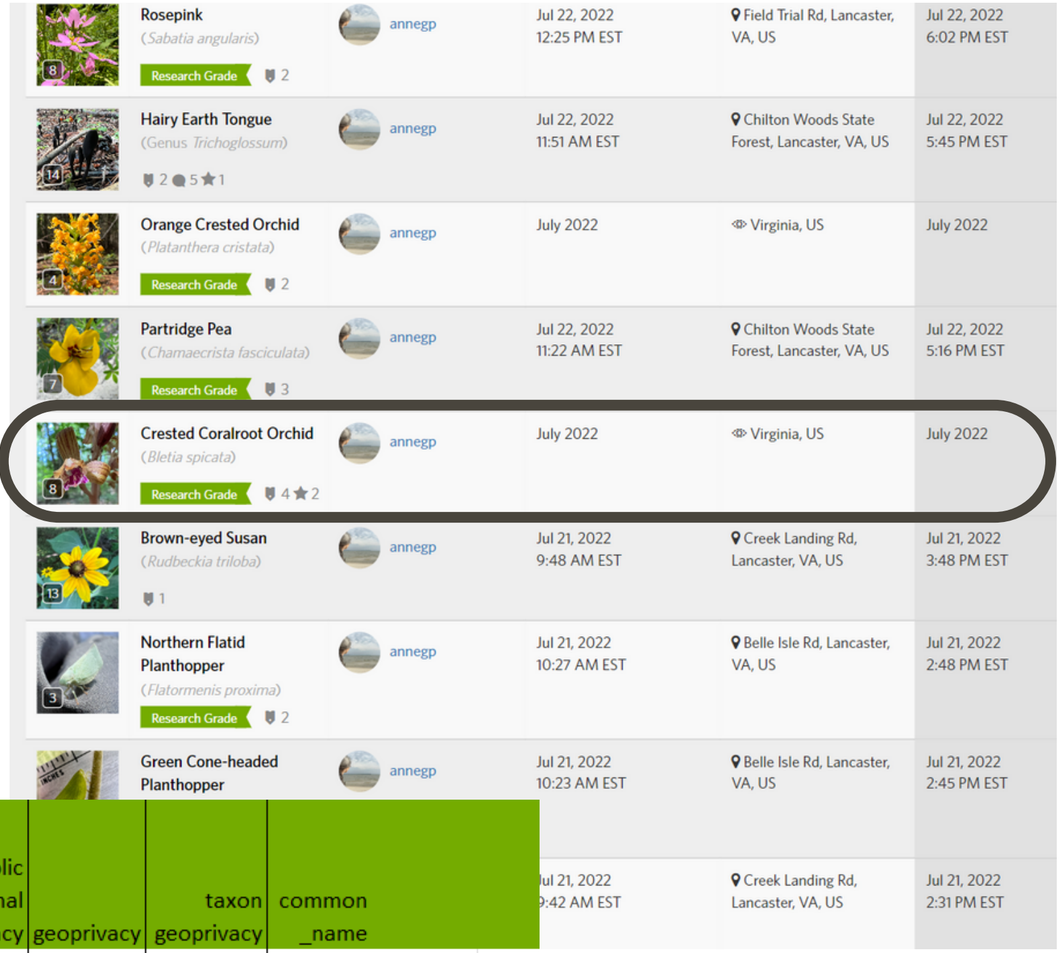
THE PROBLEM

Knowing the user and month gives access to a large amount of surrounding data that is full of clues about the location of the vulnerable species.

Even on the public-facing website, there is enough information to give a good guess, but downloading the data for this particular user gives access to even more information, including the more specific timestamps related to the observation.

More measures would need to be taken to protect vulnerable species from poaching. The question I am exploring with this capstone is: ***What can data science and machine learning tell us about the best practices for obscuring the geolocation of at-risk plants and the geolocation or other information of those observations made near it in time by the same user?***

Ethical note: While I do not imagine this project will have much reach beyond class, I do recognize that, in the wrong hands, this project could make vulnerable species more vulnerable. So just in case, I do ask others keep species protection in mind before sharing.



id	time_observed_at	created_at	place_guess	latitude	longitude	public positional accuracy	geoprivacy	taxon geoprivacy	common _name
127257452	2022-07-21 13:42:3	2022-07-21 18:31:4	Creek Landing Rd, Lancaster, VA, US	37.7805	-76.5935	17			buttonweed
127259074	2022-07-21 14:23:3	2022-07-21 18:45:1	Belle Isle Rd, Lancaster, VA, US	37.7791	-76.5939	4			Green Cone-headed Planth
127259668	2022-07-21 14:27:2	2022-07-21 18:48:4	Belle Isle Rd, Lancaster, VA, US	37.7794	-76.5934	3			Northern Flatid Planthoppe
127267270	2022-07-21 13:48:2	2022-07-21 19:48:5	Creek Landing Rd, Lancaster, VA, US	37.7801	-76.5925	9			brown-eyed Susan
127405843	2022-07-22 13:27:0	2022-07-22 19:58:5	Virginia, US	37.7646	-77.1231	28359	obscured	obscured	crested coralroot orchid
127415442	2022-07-22 15:22:4	2022-07-22 21:16:5	Chilton Woods State Forest, Lancas	37.8234	-76.5351	4			partridge pea
127417825	2022-07-22 15:40:1	2022-07-22 21:37:4	Virginia, US	37.9764	-76.5432	28329		obscured	orange crested orchid
127418944	2022-07-22 15:51:5	2022-07-22 21:45:3	Chilton Woods State Forest, Lancas	37.8217	-76.5314	4			Hairy Earth Tongue
127421140	2022-07-22 16:25:1	2022-07-22 22:02:5	Field Trial Rd, Lancaster, VA, US	37.8246	-76.5514	4			Rosepink

CHOOSING A TARGET SPECIES

Conservation Status: Obscured Taxon Geoprivacy in Many Places

Yellow Lady's Slipper | United States | Go | Filters

5,594 OBSERVATIONS	1 SPECIES	957 IDENTIFIERS	2,679 OBSERVERS

Conservation Status: Obscured Taxon Geoprivacy in Many Places

Spring Ladies' Tresses | United States | Go | Filters

4,182 OBSERVATIONS	1 SPECIES	509 IDENTIFIERS	2,135 OBSERVERS

Conservation Status: Open Taxon Geoprivacy in All Places Except Quebec

Pale-spiked Lobelia | United States | Go | Filters

4,513 OBSERVATIONS	1 SPECIES	547 IDENTIFIERS	2,311 OBSERVERS

To train and test models, I need validation data that has locations given at a research level of public positional accuracy (30m or less). This means I can't use data from vulnerable species to build my models.

Pale-Spiked Lobelia (Lobelia spicata) has similar overview data inside the United States, but is not a protected species, which means I can use it as a target species with validataion data toward the following question:

If iNaturalist's obscuration of the location is the ONLY protection measure taken to hide an observation's location, how well (or not) does it, on its own, protect the target from being found?

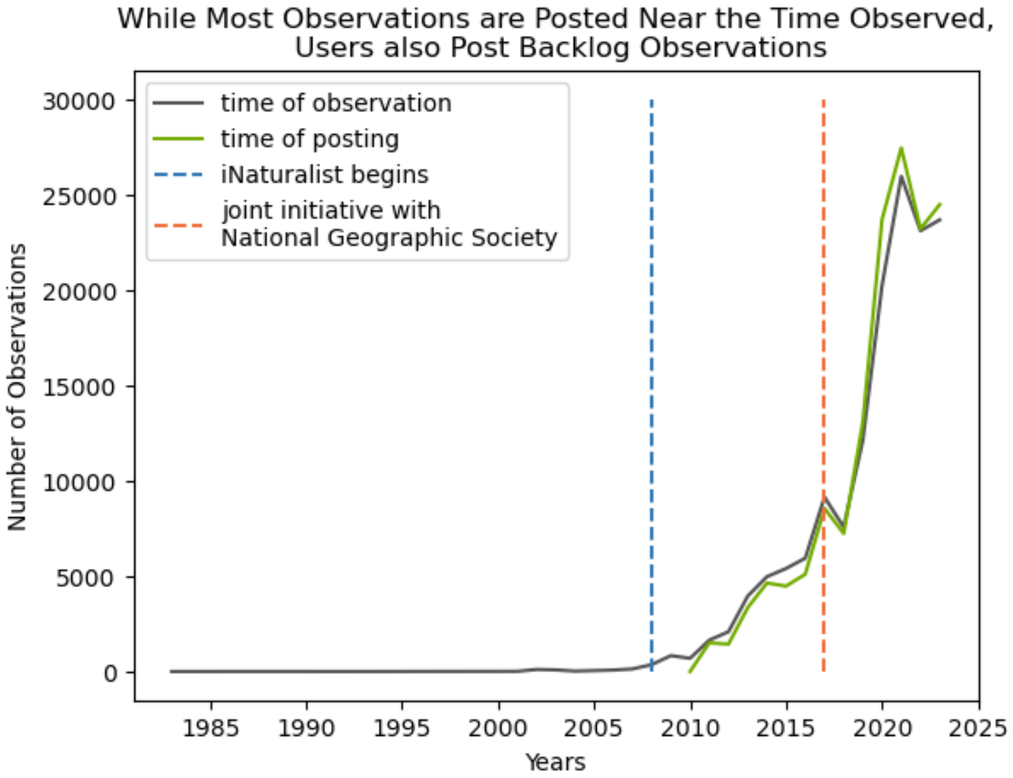
Assumptions: a) User data surrounding Lobelia spicata will look similar to user data surrounding vulnerable species. (This is actually not true, however, the assumption helps test the question "if the system's measure is the only one used..."). b) User data surrounding Lobelia spicata will look similar to user data surrounding other non-vulnerable species. (This can and should be tested).

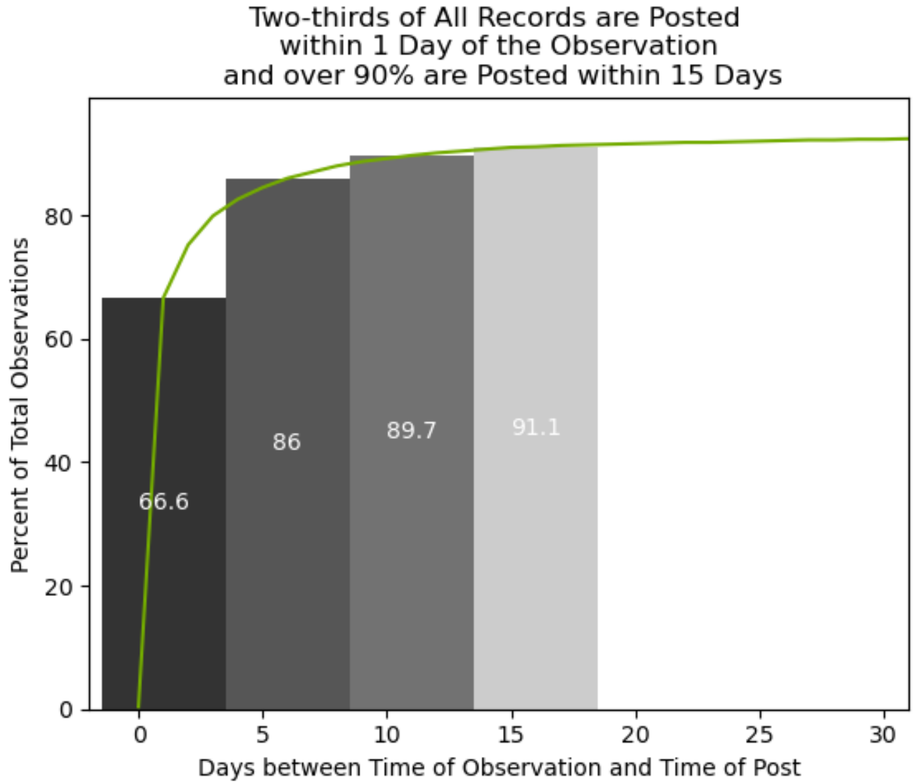
CONSTRUCTION OF DATASET FOR STUDY

Pale-Spiked Lobelia
© McCool-Grime

- Download all Lobelia spicata observations in the United States
- Filter for only those with public positional accuracy of 30m or less.
- Create a list of all unique users in the filtered dataset (returned 1245 users).
- Ideally, I would then be able to create a larger set from all of those 1245 users combined. However, the mechanism that was available to me for querying the iNaturalist database (with my current skillset), only allowed me to query for one user at a time. Given this restriction, I chose the following methodology:
  - Use random shuffle to pick 10 users from the larger set at a time.
  - Query, download and concat these sets of 10 into a larger dataset, until I have reached at least 100,000 total entries AND at least 100 Lobelia spicata observations with public positional accuracy of 30 m or less.
  - This method created a dataset that, once cleaned, had exactly 100 qualifying Lobelia spicata entries AND overall 148243 rows.

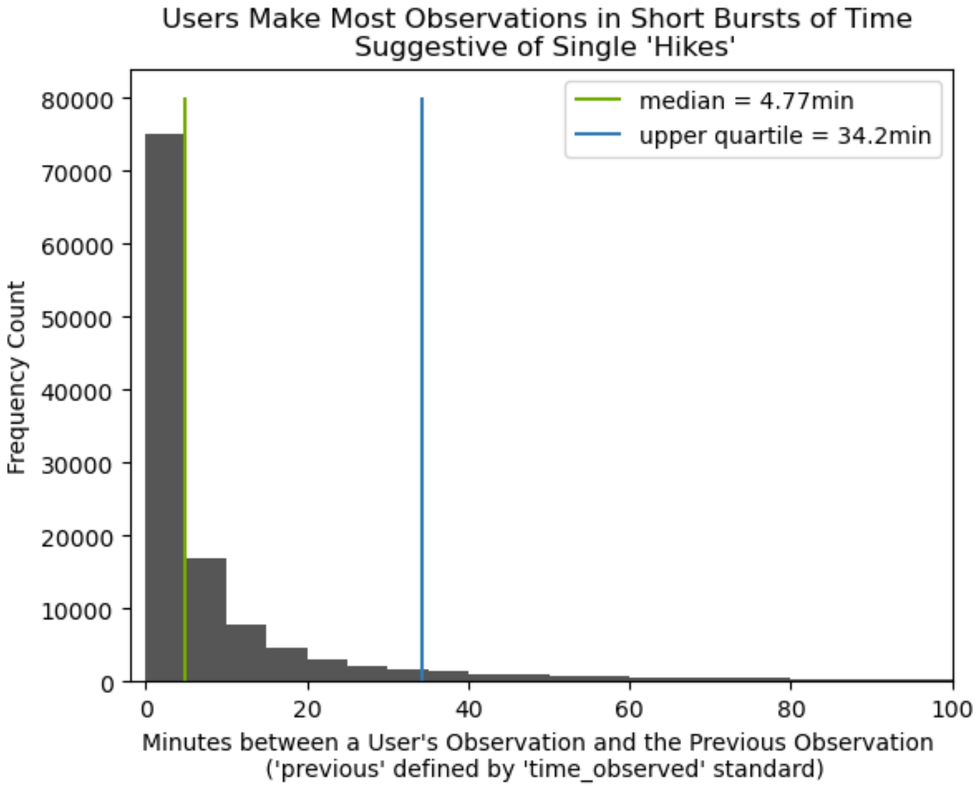
RELEVANT INSIGHTS: TIME-RELATED CATEGORIES

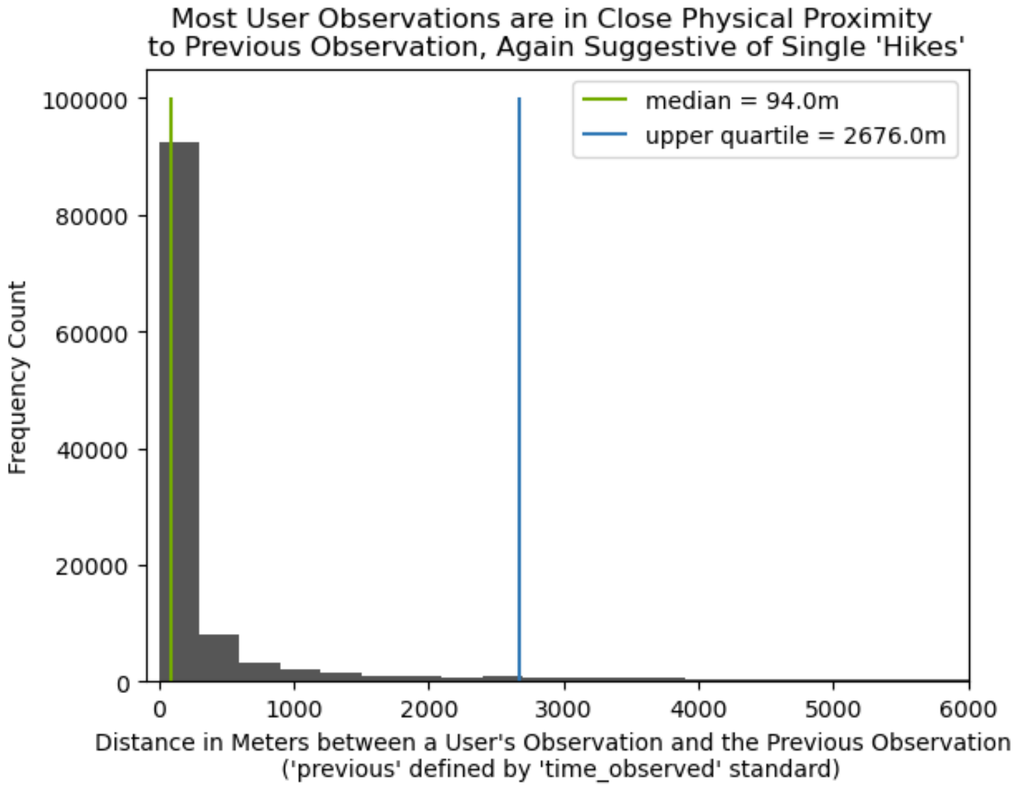




Impact: An important question for study will be: “if ”time_observed_at” is removed (obscured by iNaturalist), how well can other categories predict a target’s location? Here we see “created_at” makes a likely alternative category to “time_observed_at” for ordering data for interpolation. Additionally, iNaturalist’s system for creating unique observation IDs is to consecutively assign new IDs as observations are created. Therefore the ID category makes for an alternative as well.

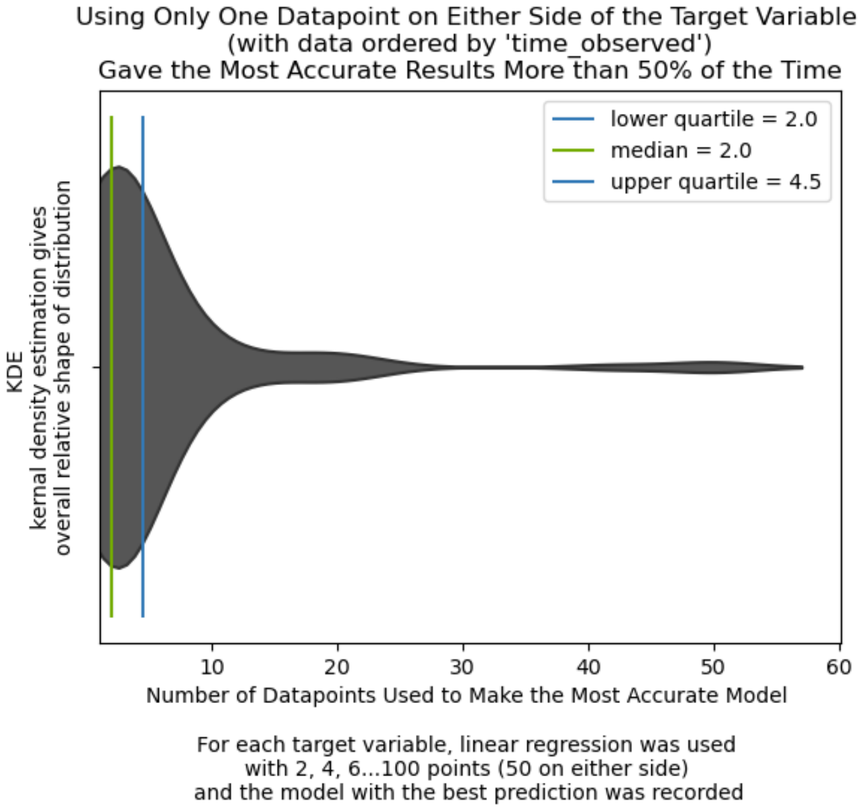
RELEVANT INSIGHTS: PROXIMITY



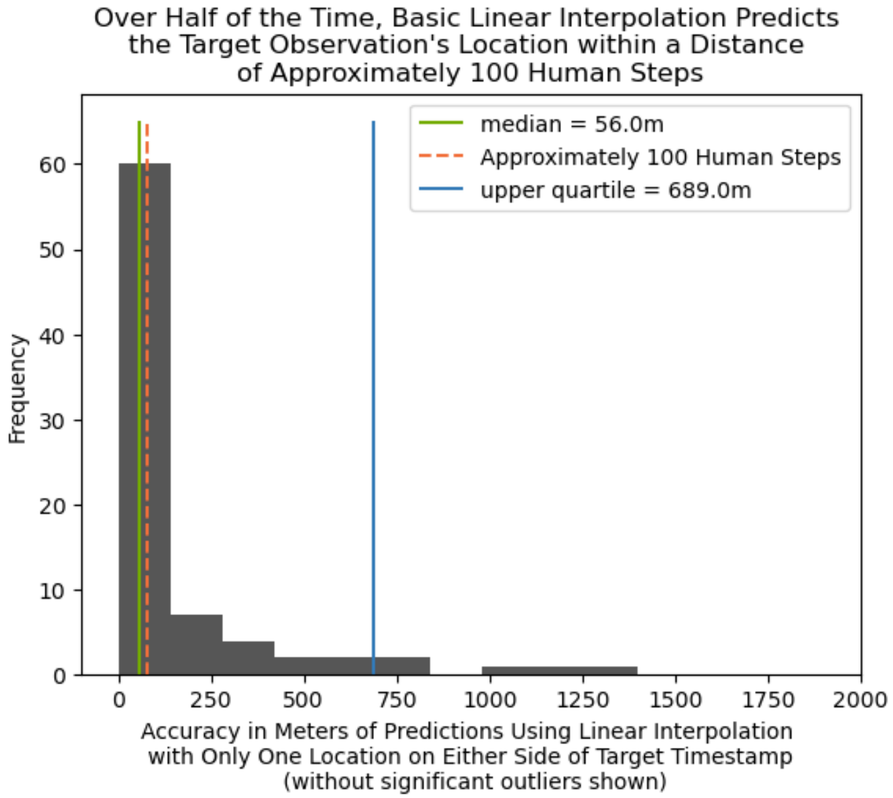


Impact: The outlier vs. mode behavior for both distance and time suggests there are time and location clusters among the data (mode) with single large gaps in time (outliers). If the observation time of a target species with an obscured location falls inside of a time/location cluster, the odds for finding the location of the target species becomes much greater. I don't yet know what kinds of modeling would identify these clusters, but if they exist, models that can find time/location clusters inside of the data will be useful for later analysis.

BASIC INTERPOLATION: INITIAL INSIGHTS



Impact: After this exploration of basic interpolation, I chose to use only the two datapoints on either side of the target timestamp to see how well it would perform at predicting the target locations. The results of the accuracy of that model are to the right.



Impact: It is clear from this exploration of linear interpolation at its most basic that the current methods used by iNaturalist for protecting the location of vulnerable species isn't sufficient. However, there are outliers in this dataset--and those outliers have clues as to what circumstances DO protect the location.

QUESTIONS TO PURSUE FROM HERE

- Assuming iNaturalist takes the measure to better obscure the “time_observed_at” and “created_at” data, how well does the basic interpolation model perform when the dataset is ordered by observation ID?
- What does a train/test split look like for my current modeling process?
- What models can be created to identify time and location clusters? How can they be used to refine the success of a finder model? What can we learn from this about protecting a target species?
- What can we learn from the outliers in the last dataset about how to better protect the location of target species?
- What measures do individual users take in addition to iNaturalist when trying to protect the location of a vulnerable species? Can those methods be systematized and if so, how well do they work?
- Does the second assumption I made about Lobelia spicata hold? Do I get similar results if I chose a different target species?

Yelllow Lady's Slipper
© McCool-Grime