CREATION OF A GAMEFISH OCCURRENCE DATASET FROM PUBLIC-FOCUSED INFORMATIONAL NEWSLETTERS

BY

REBECCA HEDREEN

A Thesis
Submitted to the School of Graduate and Professional Studies
in Partial Fulfillment of the Requirements for the
Degree of Master of Science in Biology

Southern Connecticut State University
New Haven, Connecticut
December 2021

ABSTRACT

Author:	Rebecca Hedreen
Title:	Creation of a gamefish occurrence dataset from public-focused informational newsletters
Thesis Sponsor:	Dr. Sean Grace, Biology
Institution:	Southern Connecticut State University
Year:	2021

In order to properly assess current ecological conditions, we need long-term ecological data. Historical ecology focuses on that long term, including the need to synthesize data from diverse sources. In the Long Island Sound, the Connecticut Department of Energy and Environmental Protection has been collecting data for both scientific and recreational purposes for decades, but the format of the recreational data (narrative) is not suitable for scientific analysis. This project is to collate and annotate game fish occurrence data from the Fishing Report newsletters put out by DEEP every week during the fishing season and the DEEP Trophy Fish annual reports, over a 12-year period. Species, location, and measurement data (as available) have been compiled into a data set, with geolocation coordinates added for the identifiable locations. This thesis consists of the machine-readable dataset, the protocol for collating this data, and an assessment of the suitability of the data for different kinds of analysis. The dataset will be published openly for reuse, reanalysis, and collaborative additions.

DEDICATION

To my husband and cats, who put up with frustration and foot-dragging, while always being willing to lend a hug and/or purr.

ACKNOWLEDGEMENTS

First, thanks to my advisor Dr. Sean Grace, for giving me the inspiration for this project (even if he didn't know it at the time,) and to him and my committee for being willing to take on a librarian with a very librarian-ish biology project. Also, to the dedicated librarians and library staff of the Connecticut State Library who archive things like the Weekly Fishing Report. This project wouldn't be remotely possible with archives like these. And finally, to the unnamed DEEP employees who have so regularly collected information for the benefit of the public. State employees rock!

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

> "For years I have been collecting information on the history of the Passenger Pigeon in Wisconsin. The data in the formal literature were disappointingly meagre. Little progress was made until the thought occurred that any nesting or trapping of consequence should receive mention in the local papers. An examination of the files of the Wisconsin newspapers provided information that exceeded all expectations." (Schorger 1939, p. 3)

A good deal of ecological work is limited by the amount and the longitudinal extent of existing data. It is hard to compare current data to historical trends if those trends only go back a few years. Limitations in historical data are one of the prime foci of the field of Historical Ecology. On the widest level, historical ecology is the study of human-environmental interactions across time; marine historical ecology is specifically that study within coastal and marine ecosystems (John N. Kittinger et al. 2015). Because humans have been fishing and gathering in coastal and marine ecosystems for a very long time, well before the ecologists started gathering systematic data, it is often unclear what comparison we should be making when we look to our recent data collection.

The history of the striped bass (*Morone saxatilis,* one of the species in my own study) on the eastern coast of the United States is a good example. One of the first population studies of the eastern populations was D. Merriman's Fish and Wildlife Service study which was published in 1941 (Merriman 1941). Previous studies had mostly focused on the life history and migratory habits of local populations and some population studies on the introduced species in California.

Merriman somewhat vaguely refers to early settlers' mentions of the abundance of the species, but the first direct measurement is from 1852 (a one-night catch of 9,900 pounds in the Niantic River, CT). Local catches are recorded as declining through the end of the 19th century and into the 20th, though populations are at this point known to fluctuate widely on a yearly and seasonal basis. Merriman also makes estimates as to the age of the population, based on the average weight of catches, which were declining over the period of this study. He concluded that the populations were declining even when the common yearly swings in age group size were taken into account.

And yet one of the first reports in the 1980s, during a period of study that led to fishing restrictions, speaks of the population declining from a high in the 1950s through 1970s (Tidal Fisheries Division, State of Maryland Department of Natural Resources 1981, Oct). On the page describing the current status of the striped bass for the Atlantic State Marine Fisheries Commission, the graph only goes back to those early 1980s reports (Atlantic States Marine Fisheries Commission, 2021). Which does show the immense recovery of the population since then, but does not compare it to any historical population even in the 1930s and 1940s when those early studies were done, much less any estimated populations of earlier centuries.

This is generally known as the Shifting Baselines problem, introduced within fisheries science by Daniel Pauly (1995) and further developed in the book of that title (Jackson et al. 2011). When comparisons are done to the most available research, available either in time or effort, then a slanted picture of the trends in population is produced. This is especially a problem in conservation and restoration efforts, as without an accurate picture of historical baselines it is difficult to know to what a current ecosystem should be compared to determine success or failure. Inaccurate comparisons also affect the management of commercial and recreational

fisheries. This has led to the efforts within historical ecology to find new (old) sources of data outside of the published scientific literature. Ecological data has been extracted from (as a few examples): newspapers (Schorger 1939; Vuorisalo et al. 2001; Cochran and Elliott 2012; Vuorisalo et al. 2014; Cochran 2015), diaries and logs (Primack et al. 2009; Claesson and Rosenberg 2010; Primack and Miller-Rushing 2012), field notes (Heberling et al. 2019 Feb), photographs (Zier and Baker 2006; McClenachan 2009; Bonfil et al. 2017), maps (Bromberg and Bertness 2005; Sanderson and Brown 2007; Sanderson 2009), specimen collections (Foster et al. 2002; Hoving et al. 2003; Thurstan et al. 2015), social media (Martino et al. 2021) and even menus (Van Houtan et al. 2013).

Trophy fish reports have been a fruitful source of non-commercial historical data. Richardson, et al. (2006), used data from an angling federation and a fishing magazine, paired with a survey, to study trends in recreational fishing in Wales from 1976 to 2002. They found decreases in size and numbers of many species, but found it difficult to correlate their data with commercially reported catch rates. McClenachan (2009) found declines in the size of trophy fish photographed in the Florida Keys from 1956 to 2007. Pita and Friere (2014) analyzed a historic archive of fishing competition data in Spain, finding that the estimated populations from the competition data matched those from other studies. Bonfil, et al. (2017), found the literal trophy fish (preserved specimens) not just the trophy records, as well as photographs of Mexican sawfish. Again, they were able to document a decline in numbers and sizes. Hiddink, et al. (2019), included trophy records in their study of angelsharks in Wales, which concluded that the angelshark had rapidly declined in the area, nearly unnoticed, from the 1980s to 2017. In a different format of records, Francis, et al. (2019), looked at the size of fish deemed large enough to be reported in newspapers. They found that over all species, the size of a fish large enough to

be reported declined slightly from 1869 to 2015. Some subgroups of species showed stronger declines, especially when catch effort and technology changes are taken into consideration.

At the same time, the limitations of the data sources must be acknowledged. Changes in focus of collection efforts, geographical concentrations and gaps, and deteriorating data are the key issues and priorities of a number of projects (Boakes et al. 2010; South Atlantic Fishery Management Council 2021). Recreational fishing resources are often legitimately considered anecdotal in the worst sense, and care must be taken to avoid "fish stories[1]."

Introduction to the project

This project explores whether a set of narrative reports, originally aimed at the public, can be analyzed to produce actual data of a sort that could be compared to other scientific data collections. A pilot dataset and basic protocol are the products of this project.

The impetus for this project was a question about tournament fish data and whether any trends related to size or population could be extracted from such records. However, few fish tournaments in the Sound have lasted long enough to have useful data, and, because each tournament has been produced by a different organization, collecting that data would be extremely difficult. In the process of exploring what data does exist for the Sound, I uncovered the weekly Fishing Report.

[1] For those unfamiliar with the classic "fish story," this phrase usually refers to a story where the fish gets bigger and bigger as the story is told. This is not considered reliable data.

The Department of Energy and Environmental Protection (the DEEP) has been putting out a weekly report since at least the mid-1990s about "what's biting where" in the Sound and surrounding areas - listing species, locations, and sometimes sizes. The collection of this information is standardized - every week during the fishing season, roughly April through October, a DEEP employee goes out and talks to bait shop proprietors, charter boat crews, and individual fishers to find out what they have been catching and seeing. Since 2006, the DEEP has produced the report as a newsletter, and before that the report was published in the Hartford Courant every week during the fishing season. The newsletters seemed to provide a good pilot set to see if meaningful data could be extracted from this sort of narrative text, giving me slightly more than a decade of data if it worked. The Fishing Report covers both marine and freshwater fish, however the freshwater fish are mostly sourced from stocking, not wild populations, and so are more restricted in numbers and locations. Also, there is some existing data related to marine species populations, so using the marine species for this project will allow for future comparisons with existing data.

It seemed unlikely that no longitudinal species and/or occurrence data existed for the Sound, but in fact the data are extremely limited. Constance Cook (1995) illustrated the LIS data problem by tracking down the data behind 3 large projects. She discovered a pattern of data collection, organization, and storage by a dedicated organization, which then dissolves, merges with another organization, or switches focus, and the data and other primary documentation are archived or given away without documentation, often to other agencies, but sometimes to individual researchers. At that point the data is extremely vulnerable to being lost, and at any rate becomes essentially inaccessible.

My own attempts to access commercial and recreational catch data resulted in the discovery that these are almost entirely aggregated by year and so show nothing about seasonal trends (Klee 2015). The twice-yearly trawl surveys are done over the course of months, resulting in 2 trawls for any given location that are not directly comparable in date to other areas (CT Department of Energy & Environmental Protection 2021). So this project also seemed to have the potential to provide a unique view of fish populations: some sort of week to week seasonal data that were not recorded (or at least not published) elsewhere. One difficulty with this stage of the project is that while there is obviously a large amount of data collected, those datasets are, for the most part, entirely inaccessible. For example, commercial catch data must be collected on a regular basis, but only annual aggregate data are published. Those data presumably exist somewhere, but they are not readily available.

Part of the expansion of open science has included the recognition that for data to be open for evaluation and reuse, that data must be Findable, Accessible, Interoperable, and Reusable (Wilkinson et al. 2016). 'Findable' includes data and metadata, including persistent IDs (DOIs, etc.) and good indexing. 'Accessible' refers to making the dataset available to the extent possible within the bounds of legal and privacy ethics. 'Interoperable' is making sure that the data are usable, especially in terms of file format, but also in terms of column headings and notes. 'Reusable' is often addressed by license, but also refers to community and disciplinary standards and conventions. My project addresses 'Interoperable' most strongly, since the focus of the project is to get narrative, non-machine-readable data into a dataset format which could be used more widely. The FAIR Principles include clear identification and description of data, and this is where Long Island Sound fisheries data fail the most. It is quite unclear what datasets exist (or still exist), exactly what those datasets cover, where the datasets are located, and who has access

to them. Dr. Grace once described 'notebooks and binders full of reports' just lying around the shelves of DEEP offices (personal communication). Hardly findable and accessible, much less interoperable and reusable. It became clear to me that one part of this demonstration project would have to be to provide whatever data I could extract in as FAIR a format as I could manage.

METHODOLOGY

I briefly looked into text-mining software, but the capabilities of at least the inexpensive software platforms are not adequate for a project like this. I could extract species and location names, but connecting them properly would require a lot of review, so I decided just to do it manually.

Starting with the 2006 newsletters, I read the Marine Fish reports looking for species names and locations. As shown in Figure 1, the newsletters are generally organized by species (by common name) with locations and conditions following. Each entry consists of: the date of the newsletter, common name, location as given in the newsletter, weight & length if mentioned, water temperature (usually a range, given at the beginning of the marine fish report), and the source document. (Species names were added later and geocoding were added later.) Some newsletters included notable catches, either within the narrative or in a separate listing. These often did not include locations (Figure 1.)

HICKORY SHAD fishing has really picked up. Schools can be seen on the surface feeding in lower rivers and estuaries. Your best chances are in the lower Connecticut River (DEEP Marine Headquarters fishing pier), Housatonic River, East and West River, Thames River, Lieutenant River, and Black Hall River.

NOTABLE MARINE CATCHES –

Species	Length (in.)	Weight (lbs)	Angler
Tautog	22.2"	8 lbs. 2 oz.	Sebastian V.
Tautog	24"	8 lbs. 2 oz.	Sarah Roman
Oyster Toadfish	14"	2 lbs. 12 oz.	William K.

Dataset entries:

Year	Month	Day	Name	Species	Location	Weight	Length	Temp	Source	Latitude	Longitude
2018	11	15	Hickory sha	Alosa mediocris	Black Hall River			50-55	CT Fishing F	41.28635	-72.3177
2018	11	15	Hickory sha	Alosa mediocris	CT River (DEP/DEEP pier)			50-55	CT Fishing F	41.31141	-72.3474
2018	11	15	Hickory sha	Alosa mediocris	East River, Guilford			50-55	CT Fishing F	41.2689	-72.6608
2018	11	15	Hickory sha	Alosa mediocris	Housatonic River			50-55	CT Fishing F	41.18237	-73.1236
2018	11	15	Hickory sha	Alosa mediocris	Lieutenant River			50-55	CT Fishing F	41.30862	-72.3417
2018	11	15	Hickory sha	Alosa mediocris	Thames River			50-55	CT Fishing F	41.31533	-72.0844
2018	11	15	Hickory sha	Alosa mediocris	West River			50-55	CT Fishing F	41.2805	-72.9397
2018	11	15	Oyster toac	Opsanus tau	not specified	2.75	14	50-55	CT Fishing Report		
2018	11	15	Tautog (Bla	Tautoga onitis	not specified	8.1	22.2	50-55	CT Fishing Report		
2018	11	15	Tautog (Bla	Tautoga onitis	not specified	8.1	24	50-55	CT Fishing Report		

Figure 1. Example of narrative and table from newsletter and the resulting data

Trophy Fish reports

I also added in Trophy Fish Reports from 2009-2018, which were slightly more amenable to machine reading, being published in table form (CT Department of Energy and Environmental Protection 2020). The Trophy Fish Award program is a program by the CT DEEP to encourage recreational fishing by rewarding angling skill. Any legally caught fish beyond minimum weight and/or length measurements is eligible, and can also be entered for State Records. Weights and/or lengths must be certified and the fish photographed and properly identified. Because of the certification process, Trophy Fish reports can be used as occurrence data. The reports themselves are published annually with name, award, data, and species in table form. I used Tabula, a software tool that extracts tabular data from PDFs and converts it to a spreadsheet (Aristarán et al. 2013). Some clean-up was needed because variations in the layout of the PDFs often resulted in irregular columns. I then deleted the fisher's name, award, and other extra data to produce date-species-measurement entries. Sometimes location was given, but not always.

These are true occurrence data of the classic type. There are slightly over 900 entries and because the source is included for each entry it is possible to separate these entries out for separate analysis.

Data cleanup

Data cleanup is one of those cases where Excel was NOT the best choice: number strings, most notoriously gene identifiers, are often converted into other formats, such as dates (Rijk et al. 2019 May). In my case, geo-coordinates were sometimes converted into other formats or mangled while copying. I eventually switched to OpenRefine, which is a much better system for cleaning and arranging data. Within OpenRefine, locations and common names were standardized by using the text faceting feature, merging and replacing names as needed. Latin species were added via FishBase (Froese and Pauly 2019). OpenRefine's text faceting also allowed review of all text fields to check for typos and other erroneous or outlying data for verification.

Geolocation provided different challenges. There are a number of programs and scripts that will take a named location and produce longitude and latitude coordinates, but they are all strictly land-based. For the first pass I used Google Maps, which gives a longitude/latitude coordinate in the metadata and URL for any point selected. While Google Maps provides coordinates with a high degree of specification, given that these locations are approximate, and that the accuracy of Google Maps coordinates is unknown. I have reduced the decimal places to 3 (Chapman and Wieczorek 2020). I pasted those into the dataset and used OpenRefine's fill command to add the coordinates to all identical places. For places that were not identifiable by name in Google Maps, I did searches on the internet and in the Coast Guard Pilot documents to

identify the places. At this stage I also corrected any obvious misspellings of place names. Because the narrative text generally grouped locations that were near each other, I could also assume that if a placename was duplicated on maps I could choose the one that was nearest the other sites. I was able to find all but one named place: Rugville Reef.

RESULTS

The current dataset includes nearly 23,000 data points consisting of, at minimum, a common name and date, but with a large percentage with locations specific enough to be geo-codable and a solid fraction with some measurement data as well. 63 species and 329 places were identifiable. Over a thousand points have locations, weights, and lengths (Table 1).

The most common species mentioned were, in descending order, striped bass, bluefish, summer flounder, and scup (porgy) (Figure 2). Rare species included several sharks, rays, triggerfish, eels, and a smooth pufferfish (Figure 3). Given the newsletter process, commonality and rarity must be assumed to be due to reporting, rather than actual populations. Because striped bass and bluefish are two of the most popular gamefish, they are mentioned in every possible newsletter and given extensive coverage. The rare species are mostly reported in the Trophy Fish Report, specifically because they are rare and unusual to catch. Species of little interest to recreational fishers are rarely mentioned. For example, menhaden are only mentioned when schools are large and noticeable enough to attract larger, more interesting, predator fish.

Table 1. Dataset descriptive statistics

Basic statistics	
Number of records	22,981
Fields in Dataset	Year, Month, Day, Common name, Species name, Type, Location, Weight, Length, Surface Temp, Source, Latitude, Longitude
Number of species	63
Number of locations	329
Range of locations	Latitude: 40.03 : 41.99; Longitude: -73.78 : -70.18
Most common species	Striped bass, bluefish, summer flounder, scup
Number of entries with	
- Latin name	22,974
- Geocoded Locations	21,269
- Weights	3,327
- Lengths	2,808
- Location & Weight	2,924
- Location & Length	2,214
- Length & Weight	1,396
- Location, Length, & Weight	1,121
- Location & Measurement (Length or Weight)	4,017
Source	
- Weekly Fishing Report	21,927
- Trophy Fish Report	1,054

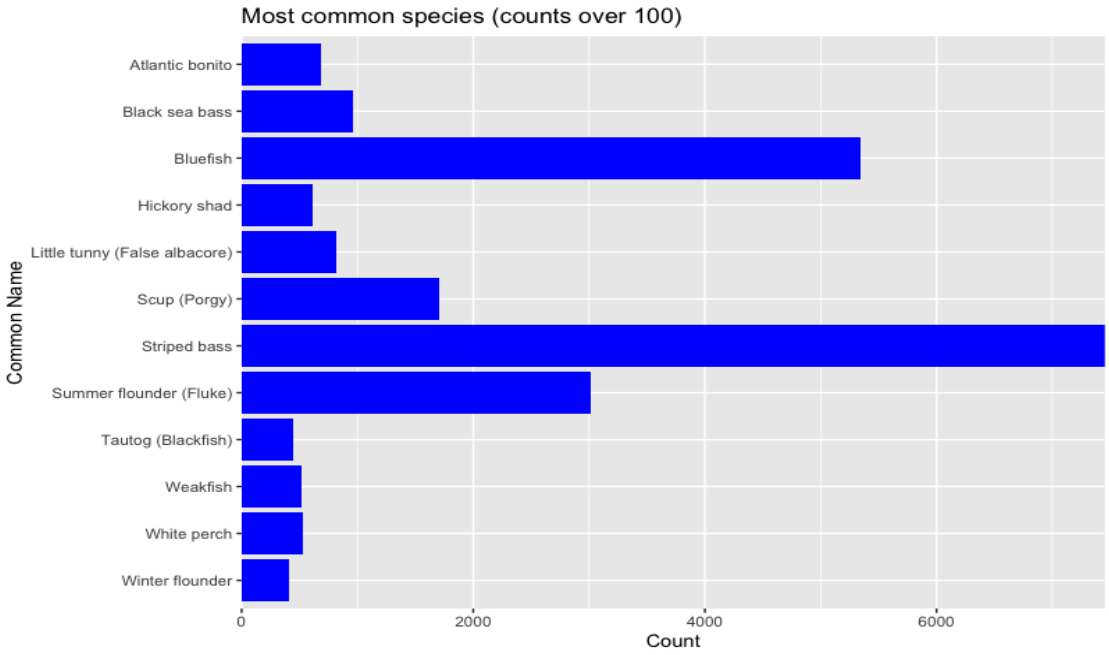

Figure 2. Most common species in the dataset

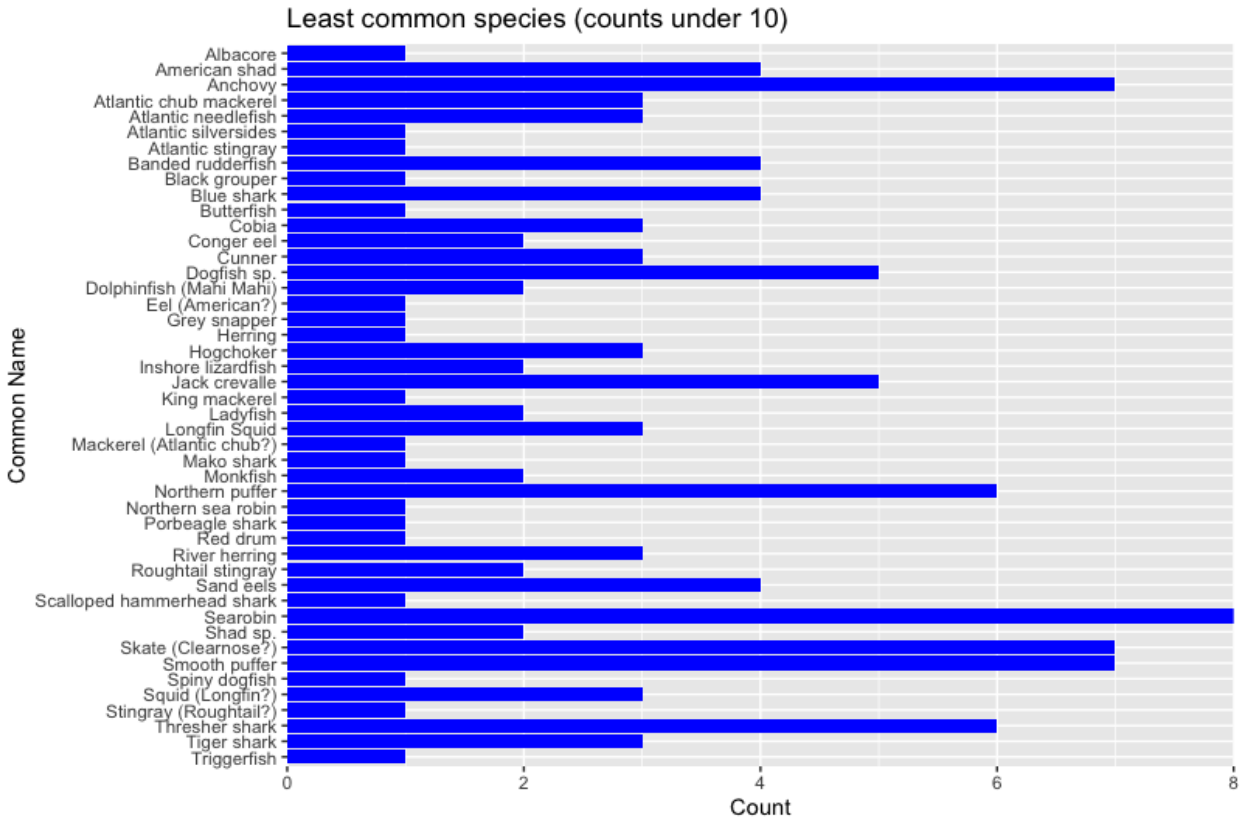

Figure 3. Least common species in the dataset

Locations identified cover the Connecticut shore, extend up the Connecticut River, and extend out to sea (Figure 4). An additional catch not represented on the map is in Maine, as the Trophy Fish reporting includes any fish caught by a Connecticut fisher, no matter where the fish was actually caught.

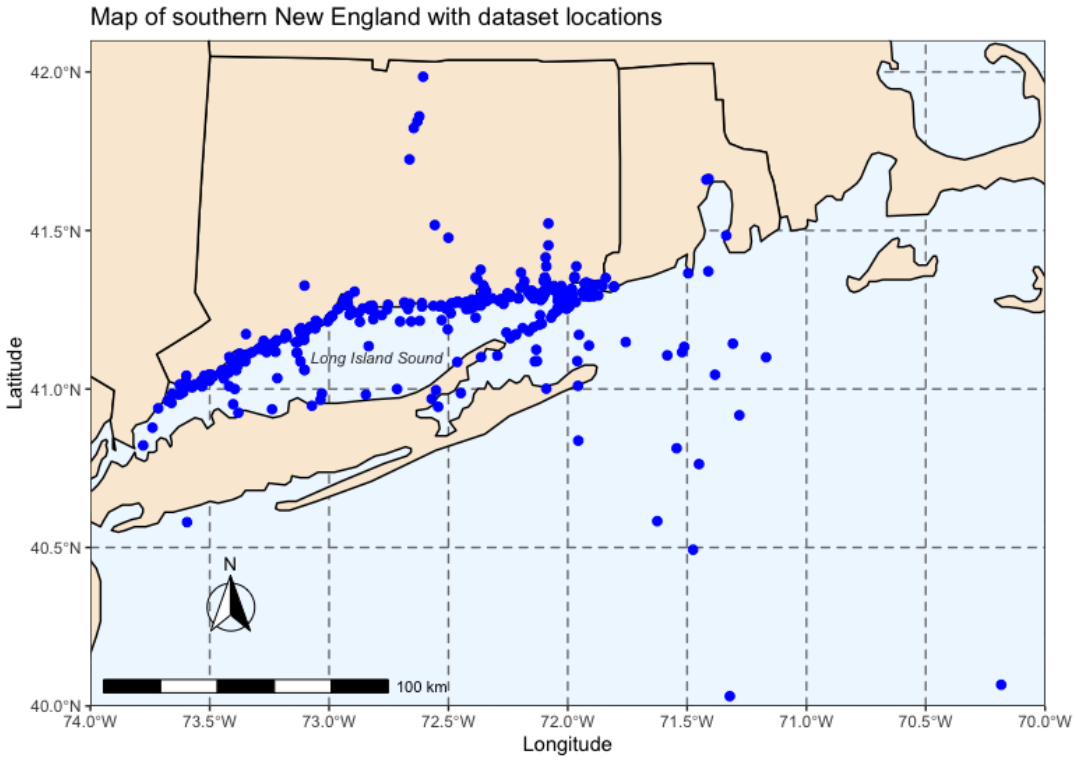

Figure 4. Map of dataset locations in Southern New England

Limitations

My conclusion is that this sort of data extraction is possible and potentially useful, but a number of criteria and limitations should be taken into account. As with many narrative projects, at least a sampling of the data should be cross-checked by another reviewer. Right now, there is no check on observer bias, as it is usually called, or narrative interpretation bias, to be more precise in this case. While most of the observations are quite straightforward in the texts - a

species is mentioned as present at a particular place - there are some vagaries in the natural language, such as referring to previously mentioned places with a new species, and more decision-making is involved in the locations. There are multiple measures for identifying and reducing observer bias, but they all require that the decisions of multiple observers be compared (Byrt et al. 1993).

I also have not figured out the best use for the locations that are vague or unidentifiable. Variations on "shoreline locations" include 78 entries, "rocky reefs" 58 entries, unnamed "reefs" 21 entries, tidal creeks and rivers 143 entries, and 19 entries only listed as "throughout LIS". If specific locations could be determined - for instance, by adding entries for all the reefs already named in the dataset to "rocky reefs" - then thousands of additional entries would be added. If only the existing named locations (329) are included in "throughout LIS" then that alone results in 6,251 entries.

In the end I only had one truly unidentifiable place, Rugville Reef, but I did make some assumptions about other locations, such as merging presumed misspellings ("Gardiners", "Gardeners", and "Gardners") and locating common names like "Sandy Point" based on the other locations in the particular text. Because the named locations are often grouped or ordered, commonly named places can often be assumed to be the nearest instance. For instance, if New Haven Harbor is mentioned just before "Sandy Point" then the mostly likely location is area near the Sandy Point Bird Sanctuary in West Haven, rather than Sandy Point Island on the border of Connecticut and Rhode Island near Stonington.

There is also simply the fact that as a recreational fishing resource, the weekly Fishing Reports are primarily concerned with fish species that are considered targets for recreational

fishing that year. Sometimes the species of interest change. For example, the oyster toadfish was not mentioned in earlier years at all (Table 2).

Table 2. Selected species with counts by year. Even popular fish such as bonito and weakfish are not represented every year

Year/ Species	Atlantic bonito	Northern kingfish	Oyster toadfish	Shortnose sturgeon	Smooth dogfish	Striped sea robin	Weakfish	White perch
2006	30						4	
2008	19						6	
2009	92						1	1
2010			6				1	
2011	10		3					
2012	73	2	3			2	5	2
2013	47	1	3				26	
2014	46	1	3				19	2
2015	61					19	41	158
2016	131	3	5	4	4	34	130	106
2017	75	10	2	2	7	23	122	60
2018	100		4	6	2	14	157	205

The biggest limitation is that I did not anticipate, and therefore document, true occurrence data and predictive occurrence data. "Bluefish were seen in New Haven harbor" vs. "Bluefish should be moving up the Sound to New Haven this week." Measurement data could be a proxy for true occurrence data in the current dataset, but that is not entirely accurate either, as some measurements given are predictive, if soundly based on previous weeks' measurements and historical trends.

DISCUSSION

While the project was strictly to evaluate the potential of the method, even the preliminary data are interesting. Basic trends in the growth of juvenile bluefish could be used to compare cohorts from year to year (Figure 5). Maximum recorded weights did not change over

the course of the decade of data though the data is sparse for all but the most popular fish (Figure 6).

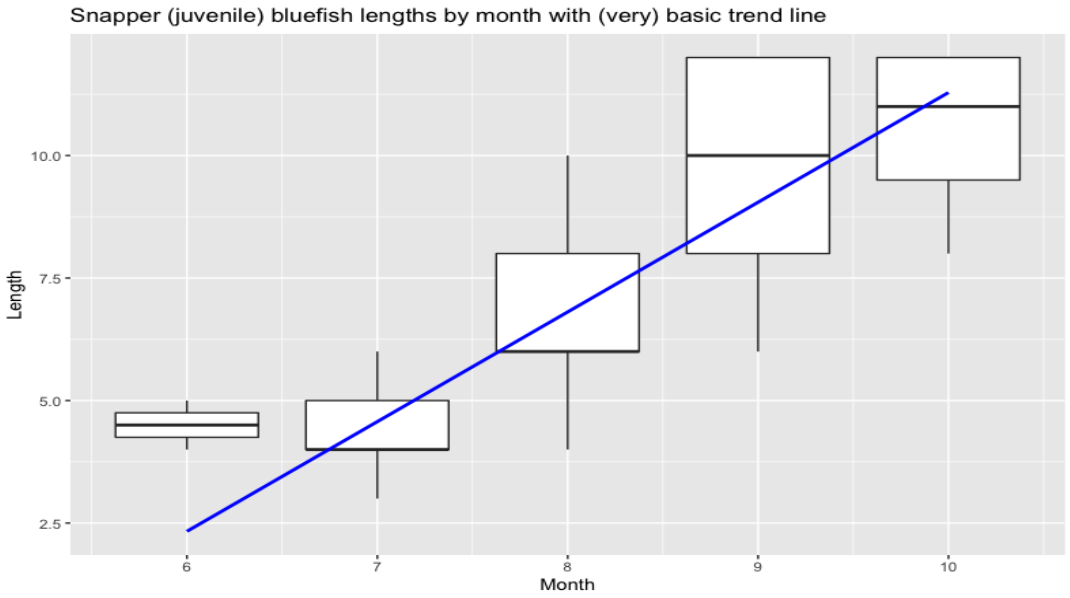

Figure 5. The dataset allows comparison of data such as juvenile bluefish growth across years

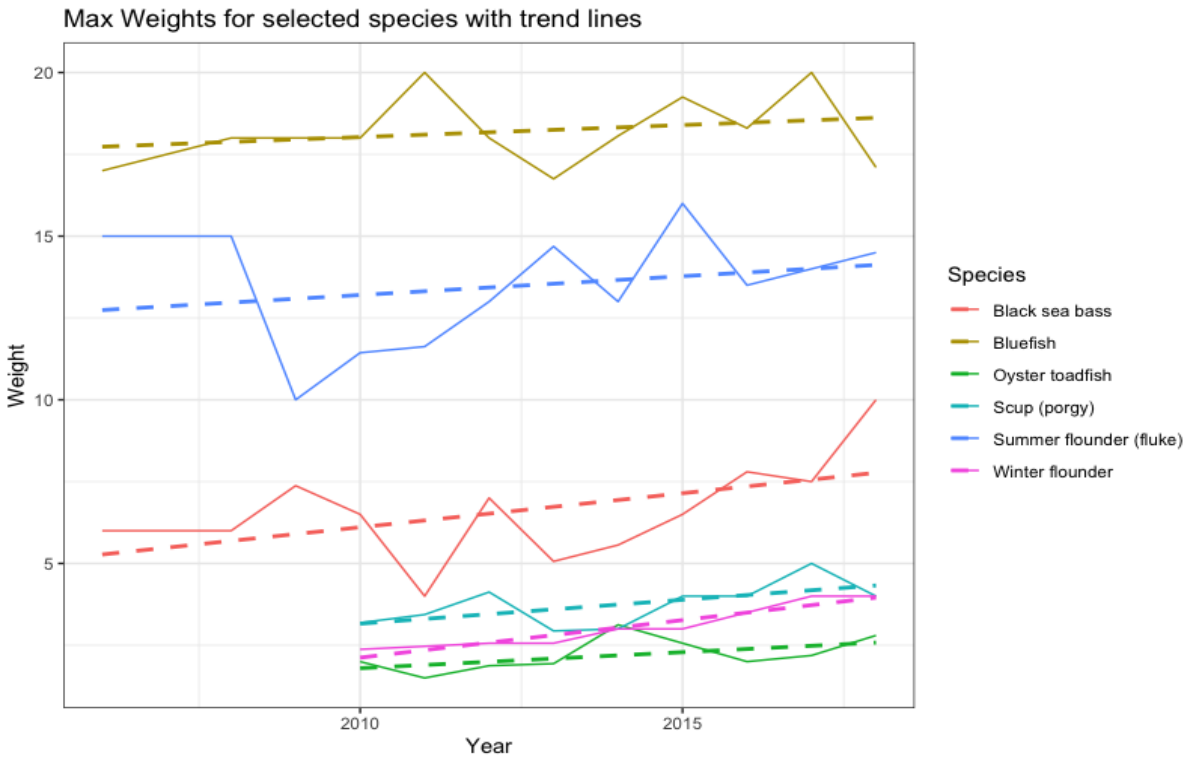

Figure 6. The dataset allows comparisons of data such as maximum weights observed over years.

Local Ecological Knowledge

The data in this study represents a form of Local Ecological Knowledge (LEK), non-scientific observational knowledge gained by locals by living and working in the local environment with the local species that has value in a scientific context. LEK is a form of anecdotal evidence, which is rightly viewed cautiously by scientists (McKelvey et al. 2008). However, it is certainly not useless. LEK or Fishers Knowledge (FK) can be used to fill in gaps in scientific knowledge and to suggest hypotheses and models to be tested in a more rigorous way (Silvano and Valbo-Jørgensen 2008; Bevilacqua et al. 2016). Because the information in the newsletters is collected from local fishers, charter boat crew, bait shop staff, and similar local experts, this data is definitely a form of LEK. The fact that this information has been collected so systematically for so long adds to its value. Irregularity in sampling, either in space or time, is an

ongoing problem in ecological work, and while there have been many efforts to address this (Hill 2012; Merow et al. 2017) adding additional sampling is almost always useful (as long as sampling methods are taken into consideration, especially when combining data from different sources.)

Future possibilities

The CT reports go back at least another decade and similar New York reports covering the Sound go back into the 1960s. The "Weekend Fishing and Boating" column goes back at least to 1965 (1965, May 28) and the "Wood, Field and Stream" column, which primarily focuses on hunting and freshwater fishing but occasionally mentions salt-water fishing as well, goes back at least to 1934 (Greenfield 1934, Dec. 8). Similar reports exist for other areas of the eastern seaboard, such as the Chesapeake Bay (Maryland Department of Natural Resources). So, there is plenty of potential for this project to expand. There is also an intriguing series of reports in the New York Times during the 1850's (then known as the New York Daily Times) regarding the conflicts between English/Canadian fishing fleets and fishers from the United States. While mostly concerned with the conflict, seizure, and loss of the ships and boats themselves, occasional mention is made of species, mostly mackerel, caught "From the Fishing Grounds" (1852).

Within the current dataset, I would like to see the data combined with additional data such as water temperature and oxygenation data, which exist for many parts of the Sound during the date range. Comparison with the twice-yearly trawl surveys would provide some validation of the occurrence data, but devising a protocol for that comparison may prove tricky, given the limited dates of the trawl data and the limited species of my data. The Trawl Survey data is

published annually with aggregates focusing on count and weight as proxies for abundance of over 100 species found in the Sound. While the trawls are conducted in a staggered fashion across the sound during spring and fall (so that 2 trawls are done in each area over the course of the year) the data are not published that way, but rather in seasonal aggregate. So, there is no way with published data to compare, for instance, the presence or absence of a species according to the trawl site 1428 to the newsletter locations around Guilford. The locations are identifiable but the species collections are not published by individual trip, so only presence/absence in the entire Sound is comparable. Calculated age indices, based on size, are given for specific species and identified by month ranges (bluefish for September/October and scup for May/June and September/October, for example) (Marine Fisheries Division, Bureau of Natural Resources 2019). A two-month range covers 8-9 newsletters, so a comparison of average size/age would be possible across the Sound, once newsletter sizes were transformed into ages using the same procedures used for the survey. This is a case where having the data in machine readable formats, as well as having access to the original, complete data, would greatly enhance the analysis potential.

This dataset could be used as a training set for a text-mining algorithm, which would make expansions of the project more feasible. One of the limits of text-mining is the need for examples of what you are looking for, usually referred to as "training" the algorithm. Because of the need for a training set of human extracted text associations, someone has to do the work of pulling those associations out before the algorithm can learn the rules (Ananiadou et al. 2006). It might be possible to use the current dataset to train an algorithm to recognize species-location pairs from the texts.

Publication

The dataset, protocol, and associated publications are being published openly in formats suitable for review and reanalysis. The current version of the dataset is available at https://doi.org/10.5281/zenodo.5532551, and the permanent link to all versions is https://doi.org/10.5281/zenodo.5532550. The OSF site set up for this project contains or links to: a description of the project, links to source material, the data set itself, this thesis and other publications related to the project, and a Zotero citation collection. That site is available at https://osf.io/sm8bv/.

CONCLUSION

In conclusion:

- Data extraction of this sort does work.
- The data are limited, but not useless.
- Some improvements in the data extraction protocol are needed.
- Figuring out how to compare the current data to existing data would be valuable but may be complicated.
- Archival/historical ecological data exist, but often in highly labor-intensive formats.

PLAIN LANGUAGE SUMMARY

Humans are a keystone species, meaning that we profoundly affect the ecosystems in which we live. We build, and grow, and harvest, and modify, and pollute, and destroy. And we've been doing that long before the development of the scientific method and the study of the science of ecology. So how do we study the long-term relationship between humans and our environment? That is the realm of historical ecology.

One of the aspects of historical ecology is devising ways to extract data from non-scientific historical sources, including newspapers, diaries and ships' logs, photographs, maps, social media, and menus, and from scientific work that wasn't formatted as ecological data, like field notes and specimen collections. (See the Introduction and References for details on other studies.) Part of the work of historical ecology is figuring out what kind of data can be found in these historical sources and how to make sure that the data is accurate and comparable to modern, scientifically collected data. This is especially important for conservation and habitat restoration, because if you can't figure out what the ecosystem looked like before we got involved, how can we "restore" it, or figure out how we can work with it to preserve it in the best shape it can be.

In this project, I used a source that is halfway between scientifically collected data and a historical narrative. The Weekly Fishing newsletters produced by the CT Department of Energy and Environmental Protection (DEEP) give suggestions for what recreational game fish are likely to be where during the fishing season in the Long Island Sound. They are produced using information that is regularly and systematically collected, but are in a form that is not very useful for scientific comparison purposes. But that information has been documented for decades, longer than most scientific studies can hope to run, and could provide a wonderful long-term

record of what's been seen, what's been caught, and when. If that information can be transformed into a format that can be compared year to year and compared to other sources of data.

There's a lot of this data, too. For this project I looked at slightly more than a decade of newsletters. But prior to that, there was a weekly column in the Hartford Courant back to the mid-1990s, and a similar program in NY State ran a weekly column in the New York Times back into the 1960s. For the 12 years of newsletters that I looked at (2006-2018, with 2007 mysteriously missing from the archives of the CT State Library) I was able to pull almost 23,000 sightings and predictions for over 60 species of fish, with the potential for even more if some locations can be verified.

That data can be compared across the decade, showing, for example, a clear pattern of when juvenile bluefish are spotted in the tidal rivers of CT. As the older data are added (I do intend to continue this project) we can see if there are trends in sightings or size of the fish as the Sound has warmed, has gotten more polluted, or started the summer low oxygen cycles (hypoxia) that has led to summer fish kills since the early 90s. By matching the dates of weather events like hurricanes, we can see how those major disruptions affect fish in the Sound. And by comparing my weekly data to data from other scientific studies, like the twice annual trawl surveys, we can see if there are trends that would let us predict seasonal changes in fish populations.

Because the Long Island Sound has such a long history of human use and exploitation, and because the modern science is mostly divided by sponsorship of 2 states (CT and NY), long term data is surprisingly hard to find. It's crucial for the future management of the Sound, in these days of climate change and ocean warming, to learn as much as we can about the ecosystem that provides so much pleasure and sustenance. Historical ecology provides one way

of doing that, and opens up the data collection to so many people as well. Who knows what set of diaries, photographs, or family histories might include important information about the ecology of the Sound, if we can learn how to use them?

REFERENCES

Ananiadou S, Kell DB, Tsujii J-i. 2006. Text mining and its potential applications in systems biology. Trends in Biotechnology. 24(12):571–579. doi:10.1016/j.tibtech.2006.10.002.

Aristarán M, Tigas M, Merrill JB. 2013. Tabula: Extract Tables from PDFs. http://tabula.technology/

Atlantic States Marine Fisheries Commission. 2021. Atlantic Striped Bass. http://www.asmfc.org/species/atlantic-striped-bass

Bevilacqua AHV, Carvalho AR, Angelini R, Christensen V. 2016. More than Anecdotes: Fishers' Ecological Knowledge Can Fill Gaps for Ecosystem Modeling. PLOS ONE. 11(5):e0155655. doi:10.1371/journal.pone.0155655.

Boakes EH, McGowan PJK, Fuller RA, Chang-qing D, Clark NE, O'Connor K, Mace GM. 2010. Distorted Views of Biodiversity: Spatial and Temporal Bias in Species Occurrence Data. PLOS Biology. 8(6):e1000385. doi:10.1371/journal.pbio.1000385.

Bonfil R, Mendoza-Vargas OU, Ricaño-Soriano M, Palacios-Barreto PY, Bolaño-Martínez N. 2017. Former widespread abundance and recent downfall of sawfishes in Mexico as evidenced by historical photographic and trophy records. Fisheries. 42(5):256–259. doi:10.1080/03632415.2017.1276339.

Bromberg KD, Bertness MD. 2005. Reconstructing New England salt marsh losses using historical maps. Estuaries and Coasts. 28(6):823–832. doi:10.1007/bf02696012.

Byrt T, Bishop J, Carlin J. 1993. Bias, Prevalence and Kappa. Journal of clinical epidemiology. 46:423–9. doi:10.1016/0895-4356(93)90018-v.

Chapman A, Wieczorek J. 2020. Georeferencing Best Practices. doi:10.15468/doc-gg7h-s853.

Claesson SH, Rosenberg AA. 2010. Stellwagen Bank Marine Historical Ecology: Final Report. Silver Spring, MD: Office of National Marine Sanctuaries Report No.: ONMS-10-02. http://stellwagen.noaa.gov/library/pdfs/sbnms_mhe_report.pdf

Cochran P. 2015. Big fish stories: Analysis of historical newspaper data on size of lake sturgeon (Acipenser fulvescens) in the Lake Michigan basin. Michigan Academician. 42(1):26–39. doi:10.7245/0026-2005-42.1.26.

Cochran PA, Elliott RF. 2012. Newspapers as sources of historical information about lake sturgeon (Acipenser fulvescens Rafinesque, 1817). Archives of Natural History. 39(1):136–146. doi:10.3366/anh.2012.0066.

CT Department of Energy & Environmental Protection. 2021. Long Island Sound Trawl Survey. https://portal.ct.gov/DEEP/Fishing/Fisheries-Management/Long-Island-Sound-Trawl-Survey.

CT Department of Energy and Environmental Protection. 2020. Trophy Fish Award Program. https://portal.ct.gov/DEEP/Fishing/General-Information/Trophy-Fish-Award-Program.

Cooke CB. 1995. The Long Island Sound studies: Where have all the data gone? In: Proceedings of the 20th Annual [1994] Conference of the International Association of Aquatic and Marine Science Libraries and Information Centers (IAMSLIC), October 9-13, 1994, in Honolulu, HI. Fort Pierce, FL: IAMSLIC. https://darchive.mblwhoilibrary.org/bitstream/handle/1912/824/proc94085.pdf

Foster DR, Motzkin G, Bernardos D, Cardoza J. 2002. Wildlife dynamics in the changing New England landscape. Journal of Biogeography. 29(10/11):1337–1357. doi:10.1046/j.1365-2699.2002.00759.x.

Francis FT, Howard BR, Berchtold AE, Branch TA, Chaves LC, Dunic JC, Favaro B, Jeffrey KM, Malpica-Cruz L, Maslowski N, et al. 2019. Shifting headlines? Size trends of newsworthy fishes. PeerJ. 7:e6395. doi:10.7717/peerj.6395.

Froese R, Pauly D. 2019. FishBase. https://www.fishbase.se.

Greenfield G. 1934, Dec. 8. Wood, Field and Stream. New York Times:22.

Heberling JM, McDonough MacKenzie C, Fridley JD, Kalisz S, Primack RB. 2019 Feb. Phenological mismatch with trees reduces wildflower carbon budgets. Maherali H, editor. Ecology Letters. doi:10.1111/ele.13224.

Hiddink JG, Shepperson J, Bater R, Goonesekera D, Dulvy NK. 2019. Near disappearance of the Angelshark Squatina squatina over half a century of observations. Conservation Science and Practice. 1(9):e97. doi:10.1111/csp2.97.

Hill MO. 2012. Local frequency as a key to interpreting species occurrence data when recording effort is not known. Methods in Ecology and Evolution. 3(1):195–205. doi:10.1111/j.2041-210x.2011.00146.x.

Hoving CL, Joseph RA, Krohn WB. 2003. Recent and historical distributions of canada lynx in maine and the northeast. Northeastern Naturalist. 10(4):363–382. doi:10.1656/1092-6194(2003)010[0363:RAHDOC]2.0.CO;2.

Jackson JBC, Alexander K, Sala E. 2011. Shifting baselines: The past and the future of ocean fisheries. Washington, DC: Washington, DC: Island Press.

Kittinger JN, McClenachan L, Gedan KB, Blight LK. 2015. Marine historical ecology in conservation: applying the past to manage for the future. Oakland, California: University of California Press.

Klee R. 2015. Interstate Marine Fisheries Management. State of CT Department of Energy and Environmental Protection Report No.: 3- ACA Final Report. https://portal.ct.gov/-/media/DEEP/fishing/performance_reports/InterstateMarineFisheriesManagementpdf.pdf

Marine Fisheries Division, Bureau of Natural Resources. 2019. Job 5: Marine finfish survey. https://portal.ct.gov/DEEP/Fishing/Fisheries-Management/Long-Island-Sound-Trawl-Survey.

Martino S, Pace DS, Moro S, Casoli E, Ventura D, Frachea A, Silvestri M, Arcangeli A, Giacomini G, Ardizzone G, et al. 2021 Mar. Integration of presence-only data from several sources. A case study on dolphins' spatial distribution. arXiv [q-bio, stat].:2103.16125. http://arxiv.org/abs/2103.16125.

Maryland Department of Natural Resources. Maryland Fishing Report. https://dnr.maryland.gov/Fisheries/Pages/fishingreport/index.aspx.

McClenachan L. 2009. Documenting loss of large trophy fish from the florida keys with historical photographs. Conservation Biology. 23(3):636–643. doi:10.1111/j.1523-1739.2008.01152.x.

McKelvey KS, Aubry KB, Schwartz MK. 2008. Using Anecdotal Occurrence Data for Rare or Elusive Species: The Illusion of Reality and a Call for Evidentiary Standards. BioScience. 58(6):549–555. doi:10.1641/B580611.

Merow C, Wilson AM, Jetz W. 2017. Integrating occurrence data and expert maps for improved species range predictions: Expert maps & point process models. Global Ecology and Biogeography. 26(2):243–258. doi:10.1111/geb.12539.

Merriman D. 1941. Studies on the striped bass (Roccus saxatilis) of the Atlantic coast. Fishery Bulletin. 50(1):1–77. https://spo.nmfs.noaa.gov/sites/default/files/pdf-content/fish-bull/fb50.1.pdf.

News from the Fishing Grounds–Dinner to Thomas Baring. 1852. New York Daily Times:1.

Pauly D. 1995. Anecdotes and the shifting baseline syndrome of fisheries. Trends in Ecology & Evolution. 10(10):430. doi:10.1016/S0169-5347(00)89171-5.

Pita P, Freire J. 2014. The use of spearfishing competition data in fisheries management: Evidence for a hidden near collapse of a coastal fish community of Galicia (NE Atlantic Ocean). Fisheries Management & Ecology. 21(6):454–469. doi:10.1111/fme.12095.

Primack RB, Miller-Rushing AJ. 2012. Uncovering, collecting, and analyzing records to investigate the ecological impacts of climate change: A template from Thoreau's Concord. BioScience. 62(2):170–181. doi:10.1525/bio.2012.62.2.10.

Primack RB, Miller-Rushing AJ, Dharaneeswaran K. 2009. Changes in the flora of Thoreau's Concord. Biological Conservation. 142(3):500–508. doi:10.1016/j.biocon.2008.10.038.

Richardson EA, Kaiser MJ, Edwards-Jones G, Ramsay K. 2006. Trends in sea anglers' catches of trophy fish in relation to stock size. Fisheries Research. 82(1):253–262. doi:10.1016/j.fishres.2006.05.014.

Rijk PD, D'Hert S, Strazisar M. 2019 May. Opentsv prevents the corruption of scientific data by Excel. bioRxiv.:497370. doi:10.1101/497370.

Sanderson EW. 2009. Mannahatta: A natural history of New York City. New York: New York: Abrams.

Sanderson EW, Brown M. 2007. Mannahatta: An Ecological First Look at the Manhattan Landscape Prior to Henry Hudson. Northeastern Naturalist. 14(4):545–570. doi:10.1656/1092-6194(2007)14[545:MAEFLA]2.0.CO;2.

Schorger AW. 1939. The great Wisconsin passenger pigeon nesting of 1871. The passenger pigeon. I(1):3–11. http://digital.library.wisc.edu/1711.dl/EcoNatRes.pp01n01.

Silvano RAM, Valbo-Jørgensen J. 2008. Beyond fishermen's tales: Contributions of fishers' local ecological knowledge to fish ecology and fisheries management. Environment, Development and Sustainability. 10(5):657. doi:10.1007/s10668-008-9149-0.

South Atlantic Fishery Management Council. 2021. SAFMC FISHstory. https://safmc.net/safmc-fishstory/.

Thurstan RH, McClenachan L, Crowder LB, Drew JA, Kittinger JN, Levin PS, Roberts CM, Pandolfi JM. 2015. Filling historical data gaps to foster solutions in marine conservation. Ocean & Coastal Management. 115:31–40. doi:10.1016/j.ocecoaman.2015.04.019.

Tidal Fisheries Division, State of Maryland Department of Natural Resources. 1981, Oct. Interstate fisheries management plan for the striped bass of the Atlantic Coast from Maine to North Carolina. Report No.: 1.

Van Houtan KS, McClenachan L, Kittinger JN. 2013. Seafood menus reflect long-term ocean changes. Frontiers in Ecology and the Environment. 11(6):289–290. doi:10.1890/13.wb.015.

Vuorisalo T, Lahtinen R, Laaksonen H. 2001. Urban biodiversity in local newspapers: A historical perspective. Biodiversity & Conservation; Dordrecht. 10(10):1739–1756. doi:10.1023/A:1012099420443.

Vuorisalo T, Talvitie K, Kauhala K, Bläuer A, Lahtinen R. 2014. Urban red foxes (Vulpes vulpes L.) In Finland: A historical perspective. Landscape and Urban Planning. 124:109–117. doi:10.1016/j.landurbplan.2013.12.002.

Weekend Fishing and Boating. 1965, May 28. New York Times: 27.

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 3(1):160018. doi:10.1038/sdata.2016.18.

Zier JL, Baker WL. 2006. A century of vegetation change in the San Juan Mountains, Colorado: An analysis using repeat photography. Forest Ecology and Management. 228(1):251–262. doi:10.1016/j.foreco.2006.02.049.

APPENDIX

Readme file for dataset, available at https://doi.org/10.5281/zenodo.5532550.

This CTFishReportOccurrencesReadMe.txt file was generated on 2021-09-23 by Rebecca Hedreen

GENERAL INFORMATION

1. Title of Dataset: CT Fishing Report Occurrences
2. Author Information

A. Principal Investigator Contact Information

Name: Rebecca Hedreen

Institution: Southern Connecticut State University

Address: 501 Crescent St. New Haven, CT 06515

Email: hedreenr1@southernct.edu

B. Associate or Co-investigator Contact Information

Name: Sean Grace

Institution: Southern Connecticut State University

Address: 501 Crescent St. New Haven, CT 06515

Email: graces2@southernct.edu

C. Alternate Contact Information

Name: Rebecca Hedreen

Email: bioscilibrarian@posteo.net, delibrarian@gmail.com

3. Date of data collection: 2017-2019
4. Geographic location of data collection: Long Island Sound (CT and NY, USA)
5. Information about funding sources that supported the collection of the data: No funding

SHARING/ACCESS INFORMATION

1. Licenses/restrictions placed on the data: CC-By 4.0

2. Links to publications that cite or use the data: https://osf.io/sm8bv/, DOI 10.17605/OSF.IO/SM8BV

3. Links to other publicly accessible locations of the data: none

4. Links/relationships to ancillary data sets: none

5. Was data derived from another source? yes

A. If yes, list source(s):

CT Dept. of Energy & Environmental Protection Weekly Fishing Reports, 2006-2018

CT Dept. of Energy & Environmental Protection Trophy Fish Reports, 2009-2018

6. Recommended citation for this dataset:

Hedreen R & Grace S. 2021. CT Fish Report Occurrences [data set]. DOI: 10.5281/zenodo.5532550

DATA & FILE OVERVIEW

1. File List:

CTFishReportOccurrencesReadme.txt (readme)

CTFishReportOccurrencesData.csv (data)

2. Relationship between files, if important: Data and readme.

3. Additional related data collected that was not included in the current data package: none

4. Are there multiple versions of the dataset? no

A. If yes, name of file(s) that was updated:

5. Why was the file updated?

6. When was the file updated?

METHODOLOGICAL INFORMATION

1. Description of methods used for collection/generation of data:

Manual extraction of species, location, and measurement data from PDF reports.

See https://osf.io/sm8bv/ for details.

2. Methods for processing the data: Cleanup using OpenRefine, v. 3.4.

3. Instrument- or software-specific information needed to interpret the data: none
4. Standards and calibration information, if appropriate: none
5. Environmental/experimental conditions: NA
6. Describe any quality-assurance procedures performed on the data: NA
7. People involved with sample collection, processing, analysis and/or submission: Rebecca Hedreen

DATA-SPECIFIC INFORMATION FOR: CTFishReportOccurrencesData.csv

1. Number of variables: 13
2. Number of cases/rows: 22981
3. Variable List:

    Year (integer)

    Month (integer)

    Day (integer)

    Common name (text)

    Species name (text, referenced to FishBase)

    Type (text, age or other sub-type of species)

    Location (text, named places)

    Weight (lbs)

    Length (in)

    Surface Temp (text, range or specific as reported in newsletter)

    Source (text, CT Fishing Report or CT Trophy Report)

    Latitude (decimal coordinates)

    Longitude (decimal coordinates)

4. Missing data codes: NA
5. Specialized formats or other abbreviations used: none