# Investigating Occurrence Data and Range Shifts Using GBIF, an Online Biological Collection

**Learning Outcomes:** Students will be able to…

- identify available data sets by doing quick surveys of an online portal to ensure sufficient data exists,
- download species records for chosen species from the GBIF data portal,
- clean downloaded data sets and merge data sets from different sources and for varying taxonomic synonyms,
- extract data from a pivot table to use in a regression of latitude vs. year,
- create data visualizations of species distributions over time,
- and accurately interpret graphs for evidence of range shifts over time.

## Introduction

You will closely examine one to several of the species that you and your team recorded in iNaturalist. For that, you will use GBIF—the Global Biodiversity Information Facility, (https://www.gbif.org), which is an international network and research infrastructure funded by the world's governments aimed at providing open access to data about all known species on Earth, at least if they have occurrence records that have been uploaded.

GBIF works with participating organization and data-holding institutions (e.g., museums) around the world to provide common standards and open-source tools, enabling researchers and collectors to share information about where and when species have been recorded. These occurrence records, or observations of date and location, derive from many sources, including everything from museum specimens collected in the 18th and 19th century to geotagged smartphone photos shared by amateur naturalists (you!) in recent days and weeks (and uploaded through apps such as iNaturalist).

The GBIF network combines all these sources together, resulting in hundreds of millions of species occurrence records. Publishers provide open access to their datasets using machine-readable Creative Commons license designations, allowing scientists, researchers and others to apply the data in hundreds of peer-reviewed publications and policy papers each year. Many of these analyses—which cover topics from the impacts of climate change, the spread of invasive species and alien pests, priorities for conservation and protected areas, food security, and human health—would not be possible without this facility.

## A note on taxonomy: synonyms

You have likely observed that when you are uploading images to iNaturalist, the system suggests a scientific name for the species. That name consists of two italicized words, the genus and the species. This is the accepted system of nomenclature in which two terms are used to denote a species of living organism, the first one indicating the genus and the second the

specific epithet (epithet = name). Common names, like red oak, are quite variable over a species range, but the scientific name is fixed throughout a species' range. However, names can change, as I explain below. But first, let's explore the full name of a species, using a common bee in North Carolina, the eastern carpenter bee, *Xylocopa virginica* Linnaeus, 1771. *Xylocopa* is the genus name, *virginica* is the species name, Linnaeus is the "taxonomic naming authority," that is the scientist who named the species, and he published that name in 1771. The fact that "Linnaeus 1771" is not in parentheses indicates that the name has not changed since Linnaeus named it.

Why would a taxonomic name change? There are several reasons, and while annoying, these changes are an attempt to reflect more accurately both the historical details of the naming of taxa (the nomenclature) and the increase in knowledge about the evolutionary relationships between organisms (the taxonomy). One reason for a name change is that the species had already been discovered and name, so the earlier valid name must be used for the "new" discovery, too. Sometimes it is discovered that a species needs to be moved to another genus, or even to a brand-new genus as a result of research that indicates new evolutionary relationships.

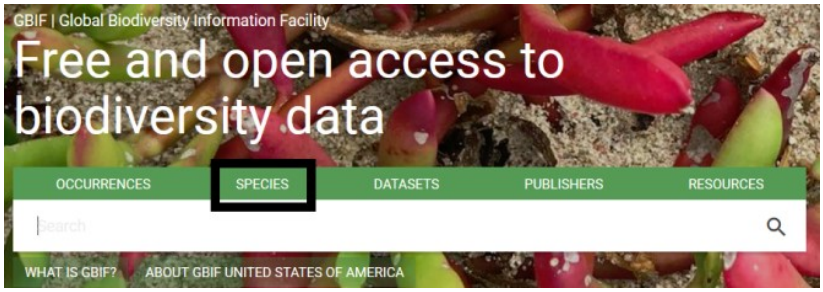
An example is a butterfly with the common name of great blue mime, which was originally named as *Chilasa paradoxa* Zincken 1831. Scientists later realized that it actually is a species of swallowtail butterfly and more appropriately belongs in the *Papilio* genus. Thus, the name changed to *Papilio paradoxa* (Zincken 1831). The scientist who moved it is not listed, but note that the original authority and date is now in parentheses. If you see that, you automatically know the name is not the original name. This is critical in our examination of museum records found on GBIF – we must search for all records of the current name as well as any synonyms. *Chilasa paradoxa* is a synonym of *Papilio paradoxa*, so we must search for all records of both names. We will practice how to search for whether species have synonyms and how to download all the records we need for the species we're searching for.

**GBIF.org: Investigating Changes in Species' Ranges or Impacts of Invasive Species Over Time Using Digitized Biological Collections (adapted from Linton et al. 2019)**

1. **NOTE: You will only perform this exercise for <u>some of the specimens</u> that you (or the iNaturalist community) have identified to the species level. Choose species that have something in common – for instance, all insects, or all trees, or all invasive species – so that you have a theme or broad question to investigate and look for consistent pattern.**
2. Once you have a confirmed identification (Research Grade observation, to the species), you will investigate your chosen species' North American distributions using GBIF. You must first do a little research on the species to verify the occurrence of taxonomic synonyms.
3. GBIF includes both specimen-based and observation-based biological collection data. You will use GBIF records to estimate species ranges and range changes over time.
4. NOTE: the absence of a species record from a specific place at a specific time cannot be considered proof that it wasn't present in that location in that year, so this is only an

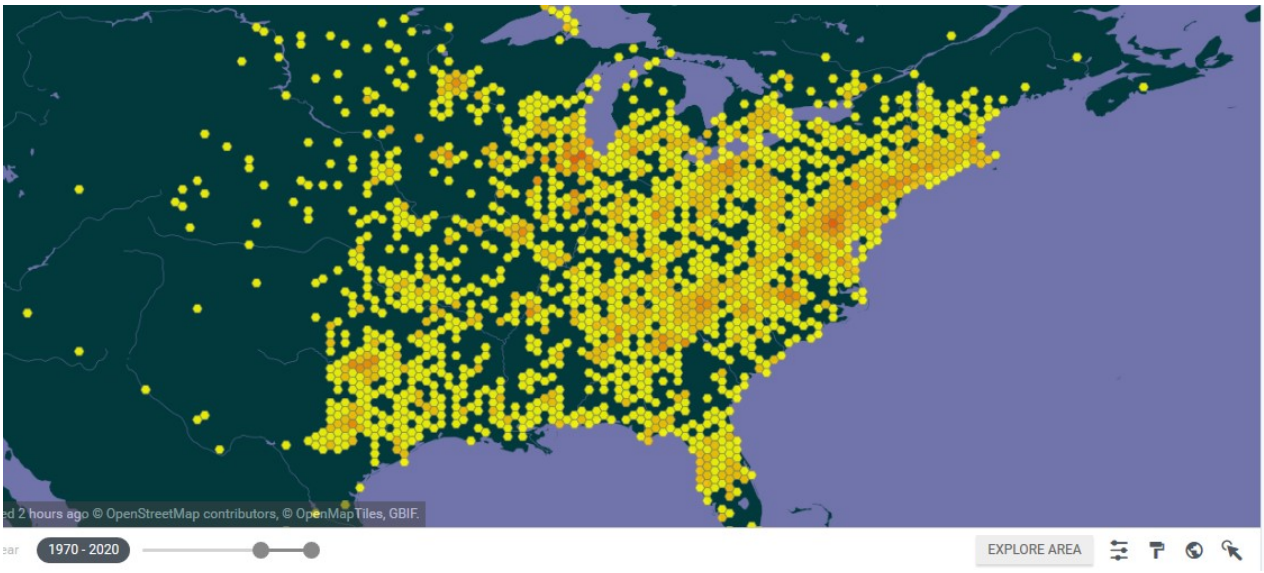
*estimate* of species ranges that we can use. For many species these are the only data available to make these range change estimates. In addition, while you may be looking for or interested in possible range shifts correlated with global climate change, there are other factors that might lead to changes in ranges over time, such as habitat destruction, introduction of a species, and more.

5. Navigate to the GBIF website @ https://www.gbif.org and create a user account if you don't already have one. [The screenshots below are specific to Firefox and Excel on a PC.]
6. Click on "Species" above the search box, and then enter the correct spelling of the species for which you want data in the search box. You can use common name or scientific binomial.




7. Search results will appear to the right of your search panel. Click on the one that best matches what you are looking for. Be cautious here, as some species that are very common will have huge datasets that will take a long time to download and may prove unwieldy to work with in Excel.

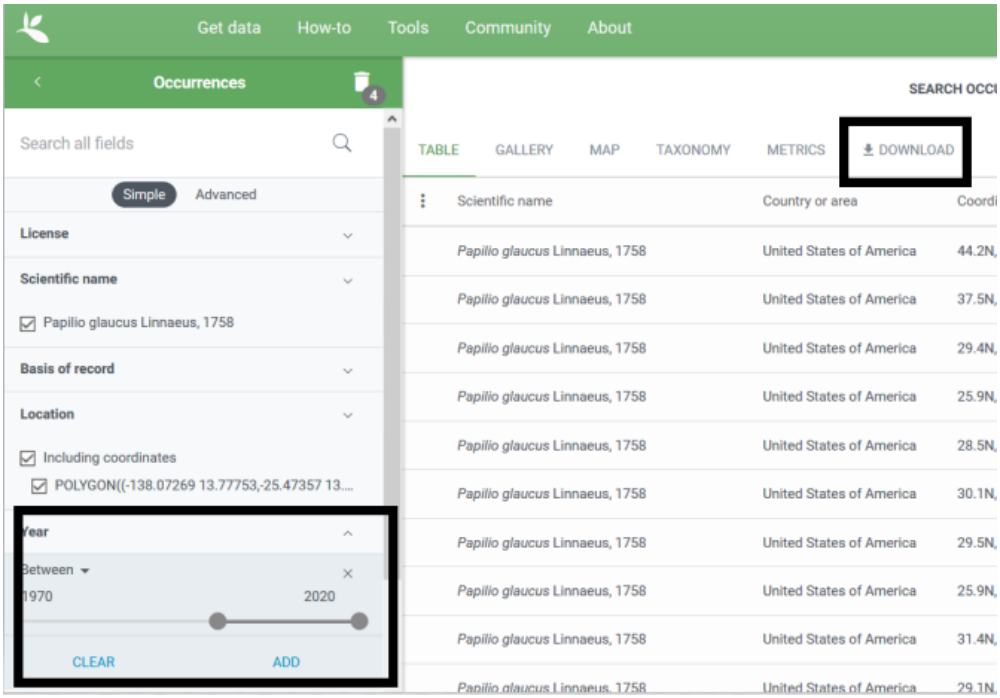

8. One approach you can take is to examine distribution/collection points over time. Directly below the map will be a slider for time of collection ranging anywhere between 1600 and the current date. You can take screen shots of maps in 20-year increments beginning with 1920-1920 and ending with 2000-2020. Note: As some species do not have data on specific years, adjust your ranges so that you have data in logical time ranges. Take screenshots of each of your maps. Paste your maps into a document that is associated with that species for later inclusion in your lab report. While I suggest starting in 1920, that could become labor intensive, so you may want to investigate a more recent range of time. In that case, I suggest examining at the very least the last 50 years (1970-2020 – so do two 25-year increments).

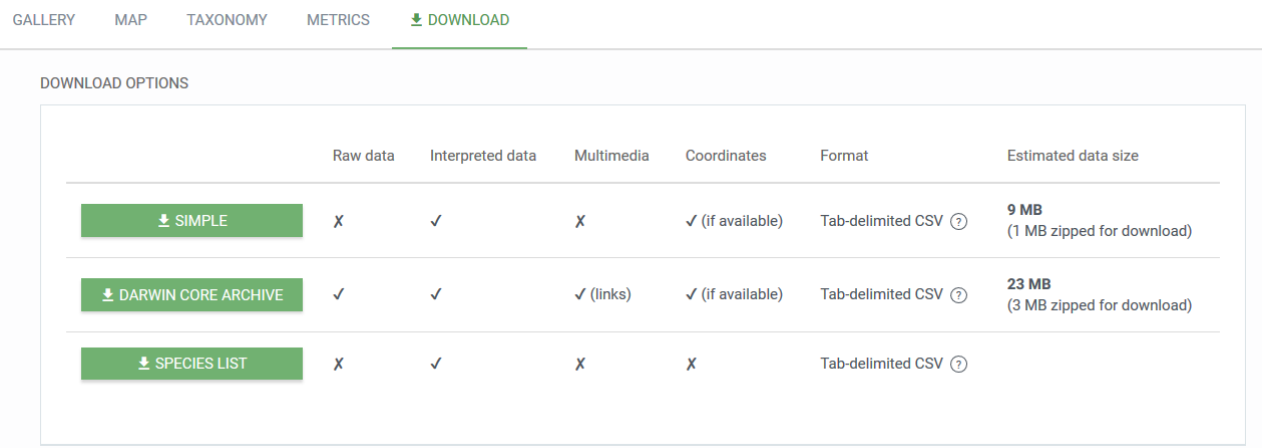



Example of a screenshot of a distribution map for *Papilio glaucus,* the eastern tiger swallowtail from 1970-2020. Note the year slider on the bottom left and the "explore area" button on the bottom right.

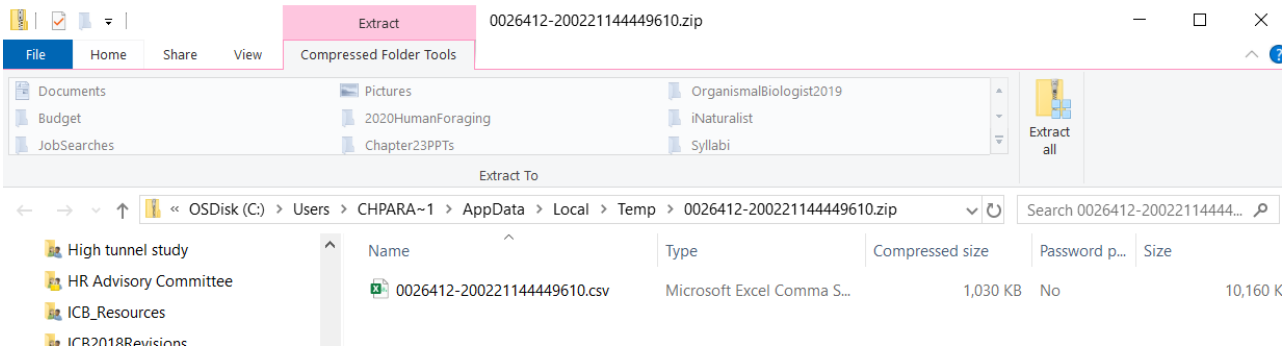
Female *Papilio glaucus,* the eastern tiger swallowtail.

9. Another approach is to examine latitudinal changes over time. For that, return to your GBIF search results and click on the "Explore" or "Explore Area" (if you've zoomed in) button below the map. In the menu section to the left, open the "year" tab and set the range to 1920 – present (or whatever your desired year range is). Select "Download" from the menu above the specimen records, then "SIMPLE" to download your data set. It will give you a popup where you have to click "Understood" and then another popup that gives you status of the download. Don't worry that it states it might take 3 hours to download – it usually takes only a couple of minutes.

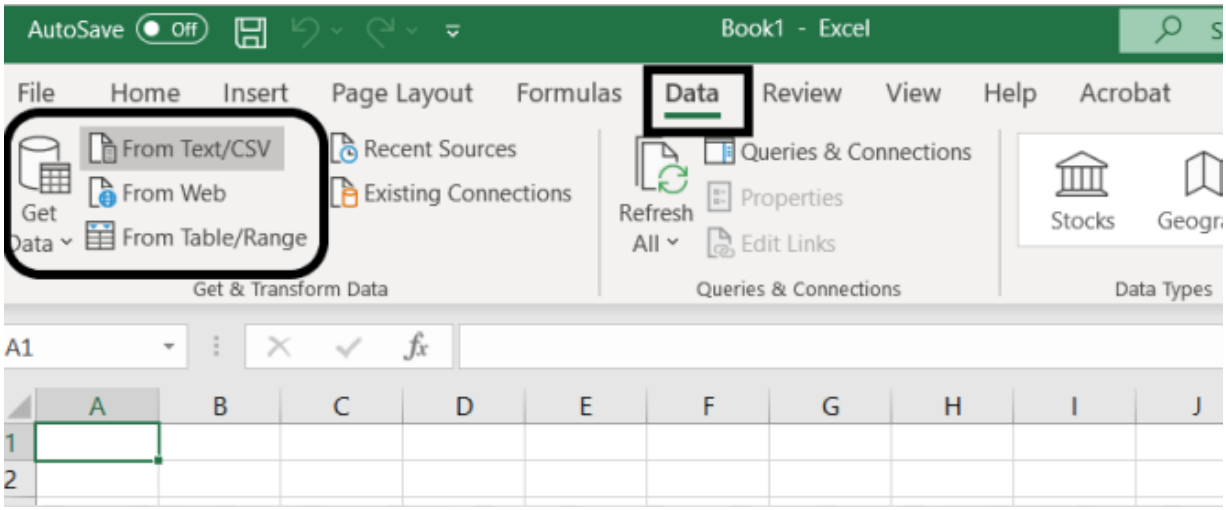


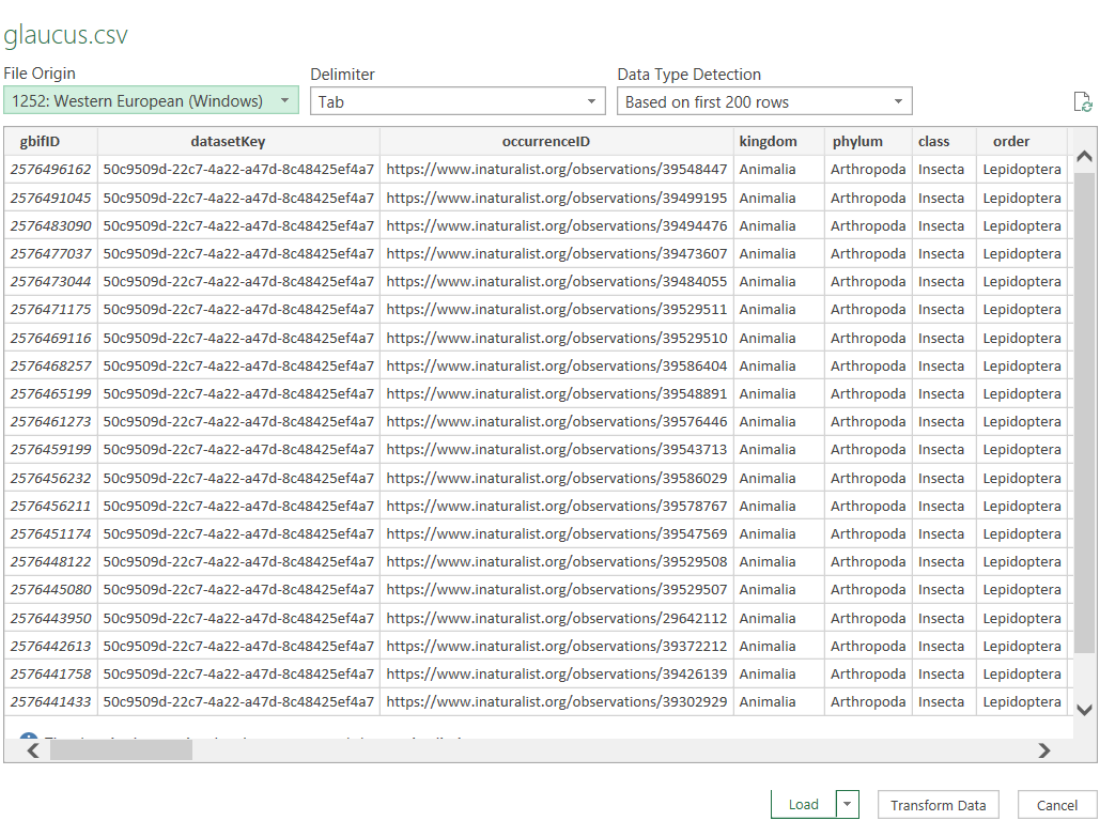

10. After the download is complete you will receive an email from GBIF. Click the link in the email to access your download. I usually choose to open the folder. Note that it is a zipped folder with the CSV compressed within. At this point you will need to extract the file (see the "Extract all" button in the upper right; save it in a folder or on your desktop, but **do not** double click to open it! NOTE: at this point, those using MACs will have a different experience for extraction of zip files. If we work together, we will be able to figure out how to handle the extraction; I can't currently help with that.




11. To open your dataset in Excel, you will need to **import** the CSV file using the **"get external data"** feature of Excel, found in the Data tab menu. Again, with MACs, Excel will look a little different, but there should be an option under the "Data" tab to "Get Data" or "From Text/CSV." Those are the commands to look for.




12. Once you select your folder to be imported, you'll get a popup window that will show the first 20 lines of the file. Inspect it to make sure it shows multiple columns, and not all data in one column. If it does, then select "Load."

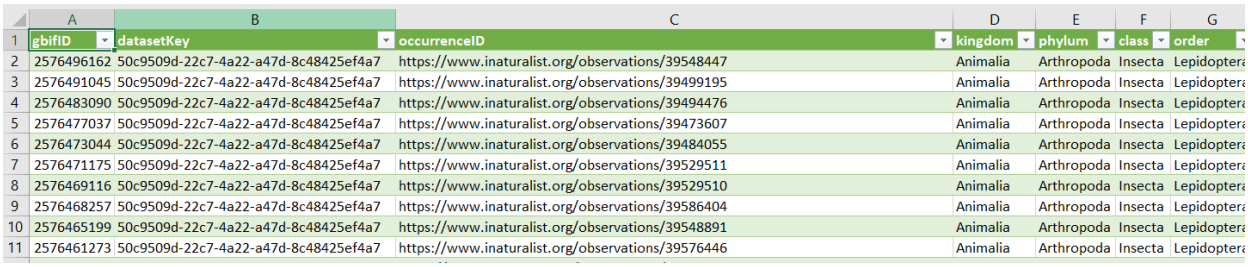

glaucus.csv

File Origin: 1252: Western European (Windows) | Delimiter: Tab | Data Type Detection: Based on first 200 rows

| gbifID | datasetKey | occurrenceID | kingdom | phylum | class | order |
|---|---|---|---|---|---|---|
| 2576496162 | 50c9509d-22c7-4a22-a47d-8c48425ef4a7 | https://www.inaturalist.org/observations/39548447 | Animalia | Arthropoda | Insecta | Lepidoptera |
| 2576491045 | 50c9509d-22c7-4a22-a47d-8c48425ef4a7 | https://www.inaturalist.org/observations/39499195 | Animalia | Arthropoda | Insecta | Lepidoptera |
| 2576483090 | 50c9509d-22c7-4a22-a47d-8c48425ef4a7 | https://www.inaturalist.org/observations/39494476 | Animalia | Arthropoda | Insecta | Lepidoptera |
| 2576477037 | 50c9509d-22c7-4a22-a47d-8c48425ef4a7 | https://www.inaturalist.org/observations/39473607 | Animalia | Arthropoda | Insecta | Lepidoptera |
| 2576473044 | 50c9509d-22c7-4a22-a47d-8c48425ef4a7 | https://www.inaturalist.org/observations/39484055 | Animalia | Arthropoda | Insecta | Lepidoptera |
| 2576471175 | 50c9509d-22c7-4a22-a47d-8c48425ef4a7 | https://www.inaturalist.org/observations/39529511 | Animalia | Arthropoda | Insecta | Lepidoptera |
| 2576469116 | 50c9509d-22c7-4a22-a47d-8c48425ef4a7 | https://www.inaturalist.org/observations/39529510 | Animalia | Arthropoda | Insecta | Lepidoptera |
| 2576468257 | 50c9509d-22c7-4a22-a47d-8c48425ef4a7 | https://www.inaturalist.org/observations/39586404 | Animalia | Arthropoda | Insecta | Lepidoptera |
| 2576465199 | 50c9509d-22c7-4a22-a47d-8c48425ef4a7 | https://www.inaturalist.org/observations/39548891 | Animalia | Arthropoda | Insecta | Lepidoptera |
| 2576461273 | 50c9509d-22c7-4a22-a47d-8c48425ef4a7 | https://www.inaturalist.org/observations/39576446 | Animalia | Arthropoda | Insecta | Lepidoptera |
| 2576459199 | 50c9509d-22c7-4a22-a47d-8c48425ef4a7 | https://www.inaturalist.org/observations/39543713 | Animalia | Arthropoda | Insecta | Lepidoptera |
| 2576456232 | 50c9509d-22c7-4a22-a47d-8c48425ef4a7 | https://www.inaturalist.org/observations/39586029 | Animalia | Arthropoda | Insecta | Lepidoptera |
| 2576456211 | 50c9509d-22c7-4a22-a47d-8c48425ef4a7 | https://www.inaturalist.org/observations/39578767 | Animalia | Arthropoda | Insecta | Lepidoptera |
| 2576451174 | 50c9509d-22c7-4a22-a47d-8c48425ef4a7 | https://www.inaturalist.org/observations/39547569 | Animalia | Arthropoda | Insecta | Lepidoptera |
| 2576448122 | 50c9509d-22c7-4a22-a47d-8c48425ef4a7 | https://www.inaturalist.org/observations/39529508 | Animalia | Arthropoda | Insecta | Lepidoptera |
| 2576445080 | 50c9509d-22c7-4a22-a47d-8c48425ef4a7 | https://www.inaturalist.org/observations/39529507 | Animalia | Arthropoda | Insecta | Lepidoptera |
| 2576443950 | 50c9509d-22c7-4a22-a47d-8c48425ef4a7 | https://www.inaturalist.org/observations/29642112 | Animalia | Arthropoda | Insecta | Lepidoptera |
| 2576442613 | 50c9509d-22c7-4a22-a47d-8c48425ef4a7 | https://www.inaturalist.org/observations/39372212 | Animalia | Arthropoda | Insecta | Lepidoptera |
| 2576441758 | 50c9509d-22c7-4a22-a47d-8c48425ef4a7 | https://www.inaturalist.org/observations/39426139 | Animalia | Arthropoda | Insecta | Lepidoptera |
| 2576441433 | 50c9509d-22c7-4a22-a47d-8c48425ef4a7 | https://www.inaturalist.org/observations/39302929 | Animalia | Arthropoda | Insecta | Lepidoptera |

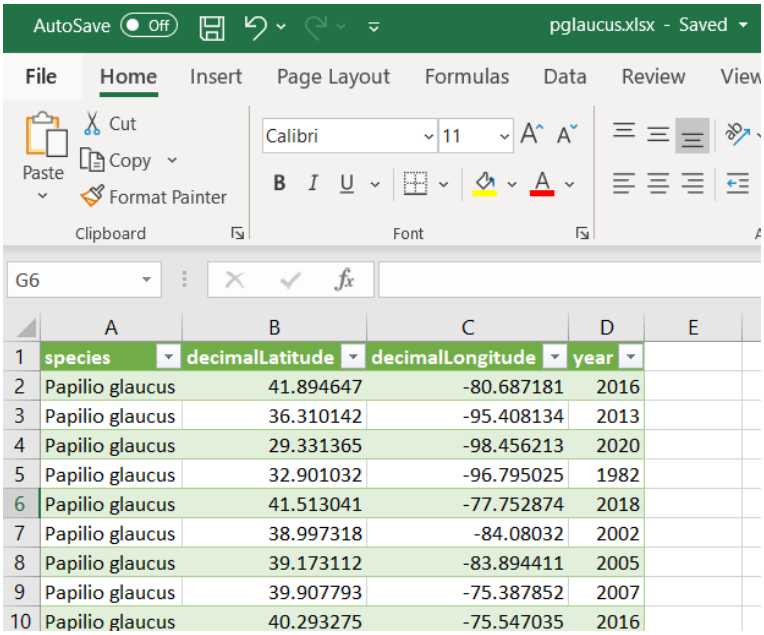
Load | Transform Data | Cancel

13. Once you have your data imported, clean it by removing unnecessary columns. The only essential columns are “Latitude” and “Year,” but you will want to keep at least “Longitude” and possibly “Basis of Record” to provide some context for your data. Everything else is important to other scientists, but not to us and just makes your database excessively large. What you see below is the first few columns of the file I imported, and none of those columns are important to us. However, as a best practice, when you delete columns always save the modified file under a new name.



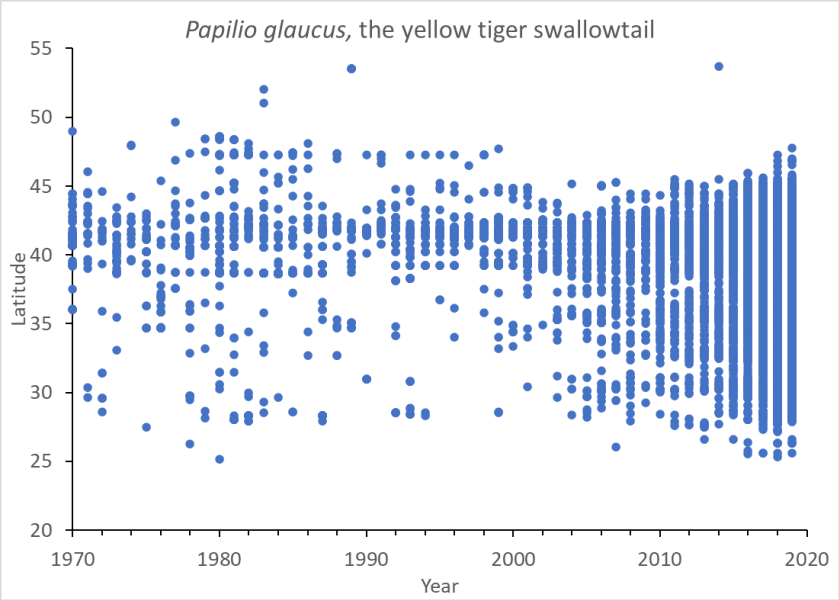
| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | gbifID | datasetKey | occurrenceID | kingdom | phylum | class | order |
| 2 | 2576496162 | 50c9509d-22c7-4a22-a47d-8c48425ef4a7 | https://www.inaturalist.org/observations/39548447 | Animalia | Arthropoda | Insecta | Lepidoptera |
| 3 | 2576491045 | 50c9509d-22c7-4a22-a47d-8c48425ef4a7 | https://www.inaturalist.org/observations/39499195 | Animalia | Arthropoda | Insecta | Lepidoptera |
| 4 | 2576483090 | 50c9509d-22c7-4a22-a47d-8c48425ef4a7 | https://www.inaturalist.org/observations/39494476 | Animalia | Arthropoda | Insecta | Lepidoptera |
| 5 | 2576477037 | 50c9509d-22c7-4a22-a47d-8c48425ef4a7 | https://www.inaturalist.org/observations/39473607 | Animalia | Arthropoda | Insecta | Lepidoptera |
| 6 | 2576473044 | 50c9509d-22c7-4a22-a47d-8c48425ef4a7 | https://www.inaturalist.org/observations/39484055 | Animalia | Arthropoda | Insecta | Lepidoptera |
| 7 | 2576471175 | 50c9509d-22c7-4a22-a47d-8c48425ef4a7 | https://www.inaturalist.org/observations/39529511 | Animalia | Arthropoda | Insecta | Lepidoptera |
| 8 | 2576469116 | 50c9509d-22c7-4a22-a47d-8c48425ef4a7 | https://www.inaturalist.org/observations/39529510 | Animalia | Arthropoda | Insecta | Lepidoptera |
| 9 | 2576468257 | 50c9509d-22c7-4a22-a47d-8c48425ef4a7 | https://www.inaturalist.org/observations/39586404 | Animalia | Arthropoda | Insecta | Lepidoptera |
| 10 | 2576465199 | 50c9509d-22c7-4a22-a47d-8c48425ef4a7 | https://www.inaturalist.org/observations/39548891 | Animalia | Arthropoda | Insecta | Lepidoptera |
| 11 | 2576461273 | 50c9509d-22c7-4a22-a47d-8c48425ef4a7 | https://www.inaturalist.org/observations/39576446 | Animalia | Arthropoda | Insecta | Lepidoptera |

14. This shows the file after I removed unnecessary columns and **saved** it under a new name:




| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | species | decimalLatitude | decimalLongitude | year | |
| 2 | Papilio glaucus | 41.894647 | -80.687181 | 2016 | |
| 3 | Papilio glaucus | 36.310142 | -95.408134 | 2013 | |
| 4 | Papilio glaucus | 29.331365 | -98.456213 | 2020 | |
| 5 | Papilio glaucus | 32.901032 | -96.795025 | 1982 | |
| 6 | Papilio glaucus | 41.513041 | -77.752874 | 2018 | |
| 7 | Papilio glaucus | 38.997318 | -84.08032 | 2002 | |
| 8 | Papilio glaucus | 39.173112 | -83.894411 | 2005 | |
| 9 | Papilio glaucus | 39.907793 | -75.387852 | 2007 | |
| 10 | Papilio glaucus | 40.293275 | -75.547035 | 2016 | |

15. Now sort by "Year" and remove any rows that do not have the year recorded and remove any other obvious outliers or duplicates. You can also remove all rows outside the date range you are examining. Once you have done all that, create a high-quality scatterplot of latitude vs. year (with axes labeled). Copy and paste your graph into the document that is associated with that species.




Example of how to plot latitude over time to examine potential expansion of a species range over time. Note that there are many more data points after 2000, due to the development of iNaturalist and other apps that allow users to upload occurrence records. Despite that, it does not appear that the eastern tiger swallowtail has expanded its range over the last 50 years.

**A few key points to keep in mind as you prepare your research report**

1. You have selected your species. Consider the rationale for your choice or choices – e.g., they belong to the same taxonomic category, or they are all invasive, or they may all be sensitive to human activities. Contact me if you have issues with downloads, data visualization, or anything else related to your analysis.
2. Prepare your summary figures and tables. **Identify the key points that you wish to make and structure your paper around them. What is the story you wish to tell?**
3. For the **Literature Cited section, r**emember to **cite your data,** using the format that GBIF provides when you download your data. Include complete citations of any other works you cite, including any sources of information on the species you researched. For this study, I am fine with sources such as Wikipedia, but there are also many online sources with information on a variety of species, such as Encyclopedia of Life (https://eol.org/).

**This is what a GBIF citation should look like:**

GBIF.org (17 March 2021) GBIF Occurrence Download https://doi.org/10.15468/dl.38smuc (but if you have more than one species, be sure to add in the species name).

**Literature Cited**

Linton D, Monfils A, Ellwood L, Phillips M (2019) Species Range Over Space and Time. Biodiversity Literacy in Undergraduate Education, QUBES Educational Resources. doi:10.25334/Q47T7F.