JMP056: Where Have All the Butterflies Gone?

Time Series Analysis, Forecasting and Generalized Linear Mixed Models

Produced by

Jennifer L. Verdolin, PhD
verdolin.jennifer@gmail.com

Where Have All the Butterflies Gone?

Time Series Analysis, Forecasting and Generalized Linear Mixed Models

Key Ideas

This case study uses time series analysis, generalized linear mixed models, and forecasting to evaluate how butterfly populations are being impacted by climate and land-use changes.

Background

(Captain-tucker, CC BY-SA 3.0)

Butterflies are insects that belong to the order Lepidoptera and represent an important group, many of which are pollinators for a wide variety of plants. There are currently about 17,500 described species. Although they occupy many different habitats, all species of butterfly go through four distinct life stages. Because they cannot regulate their own body temperatures, butterflies are sensitive to changes in temperature. Worldwide there has been a precipitous decline in butterfly populations, and scientists are concerned about the possible effects of climate change on these sensitive species. There is already some indication that a warming climate is affecting populations, but the effects may not be uniform for every species. Some species appear to be benefitting from warmer climates and expanding their ranges, while others are declining. Dr. Katy Prudic explains the complicated relationship between temperature and butterfly populations in her podcast interview, Butterflies: The Pandas of the Sky. Despite evidence that a changing climate is leading to population declines, there is still uncertainty over how much influence land-use changes such as increased urbanization or agriculture might also be contributing.

Since butterflies have an incredibly important ecological role in a diversity of ecosystems, population monitoring has been a priority for decades. In addition, there has been a concerted effort in the United States to recruit the public to assist scientists through several key citizen science initiatives. There are those that focus on specific species like the monarch butterfly and others that collect data on all species sightings. In this case study, we will focus on two key citizen science databases: iNaturalist and the North American Butterfly Association (NABA). iNaturalist is a database where thousands of nature enthusiasts contribute species identifications, which are then screened first by a machine learning algorithm and then by at least two human experts. The NABA data are collected by teams of volunteer or community

scientists and are centered on tracking butterfly populations over time. The observations reflect the number of butterflies seen by teams during a count. The iNaturalist dataset is a general species-identification database where each observation is submitted individually. The data collected by community scientists are invaluable in helping scientists track butterfly populations. As more researchers utilize these data, however, an important question to consider is how comparable the results are when using different databases. We will explore this by looking specifically at changes in the monarch butterfly over time with data from both iNaturalist and NABA datasets.

The Task

Use open-access data for butterflies from iNaturalist to investigate whether there is evidence of a change in the monarch butterfly population size in the western United States between 2001-2019. These data were filtered to include only data from western states in the US (California, Oregon, Washington, Idaho, Montana, Wyoming, Nevada, Utah, Arizona, New Mexico, and Colorado) and records after 2001 but before 2020 (6,364 records).

We will:

- Evaluate the change in monarch butterfly abundance over time
- Evaluate how temperature, precipitation, and urban and agricultural land conversion influence abundance
- Predict future abundance

Source: Forister, M. L., Halsch, C. A., Nice, C. C., Fordyce, J. A., Dilts, T. E., Oliver, J. C., ... & Glassberg, J. (2021). Fewer butterflies seen by community scientists across the warming and drying landscapes of the American West. *Science*, *371*(6533), 1042-1045.

The butterfly data are open access and available online at iNaturalist users, iNaturalist (2021). iNaturalist Research-grade Observations. iNaturalist.org. Occurrence dataset https://doi.org/10.15468/ab3s5x accessed via GBIF.org on 2021-11-22 and GBIF.org (22 November 2021). GBIF Occurrence accessed https://doi.org/10.15468/dl.cnv7bg.

The land-conversion data are open-access data and were retrieved from the Enhanced National Land Cover Dataset (2001-2019; https://www.mrlc.gov/eva/). The value of interest was the fraction of agricultural and urban land in the most recently sampled year in a 25-mile radius (Agricultural: 2019; Urban: 2016). Urban land cover comprised four categories: Developed Open Space, Developed Low Intensity, Developed Medium Intensity, and Developed High Intensity. All crop types were grouped together into a single class for agricultural land cover.

The Data

iNat Monarch butterfly.jmp and
iNat Monarch butterfly (by County).jmp

The first dataset contains yearly observations from 2001-2019 for the monarch butterfly across 11 states. The variables in the dataset include:

- Year the observation was made
- State in which the observation was made
- County in which the observation was made
- Occurrences where a single butterfly was observed

The second dataset contains the total number of occurrences per county from 2001-2019 for the monarch butterfly. The variables in the dataset include:

- County in which the observations were made
- Total number of occurrences
- Fraction of land changed to urban use in 2016 (from 2009 until last year sampled)
- Fraction of land changed to agricultural use in 2019 (from 2009 until last year sampled)
- Average annual precipitation
- Average annual maximum temperature
- Average annual minimum temperature

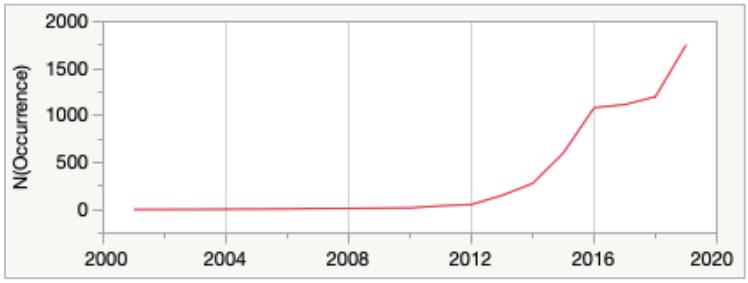
Data Exploration

Our first objective for analysis is to investigate rates of change through time for monarch butterflies.

Open the iNat Monarch butterfly.jmp dataset.

Exhibit 1 shows the time series plot for the occurrence of monarch butterflies in all the states combined between 2001-2019. We can see that the number of occurrences of monarch butterflies appears to be increasing over time. One question to consider is, does this vary by state?

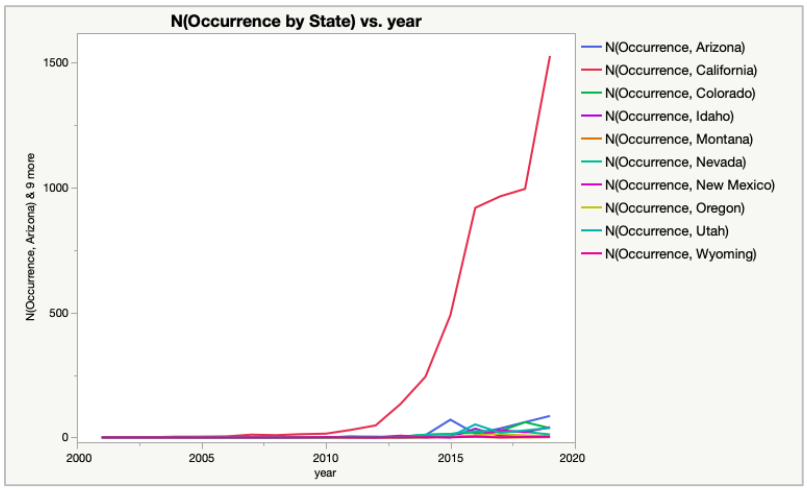
Exhibit 1



(Create a Summary Table: Table > Summary. Select Occurrences then Statistics drop down menu > N. Drag and drop Year into Group. Once you have the Summary Table, select Graph Builder. Choose the Line graph; drag and drop Year onto the X axis and Occurrence onto the Y axis.)

Exhibit 2 shows the time series plot for the occurrence of monarch butterflies for each of the western United States between 2001-2019. California has the highest number of recorded monarchs and may be driving the pattern observed. It is also the largest state and may have more records simply because of its size.

Exhibit 2



(Create a Summary Table of Occurrences by State: Table > Summary; Occurrences > Statistics > N; State >Subgroup; Year>Group. Click Ok. Graph > Graph Builder; drag and drop N(Occurrences) in Y, Year in X for Exhibit 1. For Exhibit 2, drag and drop N(Occurrences by State) in Y, Year in X. Click on the smoother icon at the top to remove the smoother, and click on the line icon. Or, right-click in the graph, and select Smoother > Remove, and Points > Change to > Line. Then, click Done.)

We can take a quick look to determine if California is disproportionately represented by showing the number of occurrences recorded for each state in a table. Exhibit 3 shows the number of records for each state. We can see that California has 5,410 records out of a total of 6,364, representing 85.009% of the data.

`Exhibit 3

Tabulate

State	N
Arizona	291
California	5410
Colorado	175
Idaho	54
Montana	5
Nevada	112
New Mexico	114
Oregon	43
Utah	149
Wyoming	11

(Analyze > Tabulate; drag states into the drop zone for rows.)

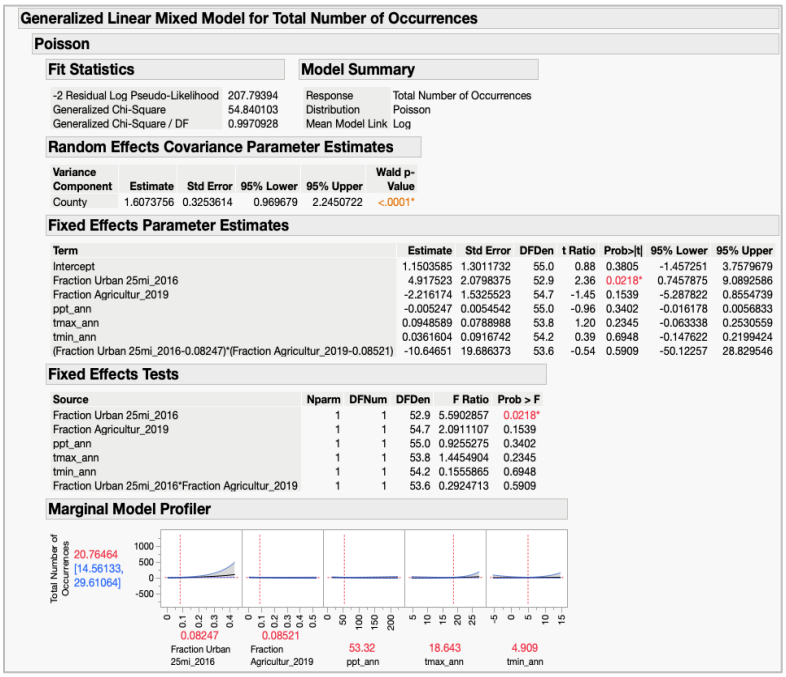
Now that we have spent some time exploring our data, we can begin thinking about investigating whether or not land-use and climate changes are impacting monarch butterflies.

Analysis

This problem involves one response variable, the occurrence of monarch butterflies over time, and various potential predictors of their presence. In this case, we have several predictor variables: average annual precipitation, average annual maximum temperature, average annual

minimum temperature, the fraction of land converted to urban use, and the fraction of land converted to agricultural use. We should not do a simple linear regression of each predictor variable. A Generalized Linear Mixed Model (GLMM), on the other hand, allows us to understand how these variables influence the presence of monarch butterflies while considering random effects such as county-by-county variability. For this analysis, we will use the iNat Monarch butterfly (by County).jmp dataset. Our response variable is the total number of occurrences, so we will select a Poisson distribution with a logarithmic link (typical for count data) and specify that this could vary randomly across counties. We also want to consider an interaction between the two land-use variables.

Exhibit 4



Generalized Linear Mixed Model for Total Number of Occurrences

Poisson

Fit Statistics

-2 Residual Log Pseudo-Likelihood	207.79394
Generalized Chi-Square	54.840103
Generalized Chi-Square / DF	0.9970928

Model Summary

Response	Total Number of Occurrences
Distribution	Poisson
Mean Model Link	Log

Random Effects Covariance Parameter Estimates

Variance Component	Estimate	Std Error	95% Lower	95% Upper	Wald p-Value
County	1.6073756	0.3253614	0.969679	2.2450722	<.0001*

Fixed Effects Parameter Estimates

Term	Estimate	Std Error	DFDen	t Ratio	Prob>|t|	95% Lower	95% Upper
Intercept	1.1503585	1.3011732	55.0	0.88	0.3805	-1.457251	3.7579679
Fraction Urban 25mi_2016	4.917523	2.0798375	52.9	2.36	0.0218*	0.7457875	9.0892586
Fraction Agricultur_2019	-2.216174	1.5325523	54.7	-1.45	0.1539	-5.287822	0.8554739
ppt_ann	-0.005247	0.0054542	55.0	-0.96	0.3402	-0.016178	0.0056833
tmax_ann	0.0948589	0.0788988	53.8	1.20	0.2345	-0.063338	0.2530559
tmin_ann	0.0361604	0.0916742	54.2	0.39	0.6948	-0.147622	0.2199424
(Fraction Urban 25mi_2016-0.08247)*(Fraction Agricultur_2019-0.08521)	-10.64651	19.686373	53.6	-0.54	0.5909	-50.12257	28.829546

Fixed Effects Tests

Source	Nparm	DFNum	DFDen	F Ratio	Prob > F
Fraction Urban 25mi_2016	1	1	52.9	5.5902857	0.0218*
Fraction Agricultur_2019	1	1	54.7	2.0911107	0.1539
ppt_ann	1	1	55.0	0.9255275	0.3402
tmax_ann	1	1	53.8	1.4454904	0.2345
tmin_ann	1	1	54.2	0.1555865	0.6948
Fraction Urban 25mi_2016*Fraction Agricultur_2019	1	1	53.6	0.2924713	0.5909

Marginal Model Profiler

(Analyze > Fit Model > Personality > Generalized Linear Mixed Model. Then under Distribution select Poisson. Add Total Number of Occurrences to the Y, add all the variables except County to Fixed Effect, and add an interaction term crossing Fraction of Urban and Fraction of Agriculture. Then click on Random Effects and add County. Then click Run.)

Exhibit 4 shows that, based on the Wald test result, we have significant heterogeneity across counties, and the fraction of urban land conversion had a significant influence on the total number of monarch butterflies observed between 2001-2019. This suggests the possibility that certain counties have higher numbers of monarch butterflies and may represent important migratory pathways. Because the analysis applies a log link, the parameter estimates are the marginal effects on Total# Occurrence in %. We can explore the effects on our response variable, Total# of Occurrences, using the Marginal Profiler (*Poisson > Marginal Model Inference > Profiler)*. From this, we observe that for the Fraction Unrbarn 25mi_2016 at 0.08247, Fraction Ag_2019 at 0.08521, ppt_ann at 53.32, tmax_ann at 18.643, and tmin-ann at 4.909, the predicted Total# Occurrence is 21 with a 95% confidence interval of [15, 30].

If you recall from Exhibit 3, a large portion of the data comes from California, and larger counties in California may be contributing to the county-by-county variability. The analysis also indicates that butterfly numbers are higher in areas that experienced a greater fraction of urban land conversion but lower in areas that recorded higher conversion to agricultural land use.

When it comes to conservation, what we really want to know is, will monarch butterfly densities increase in the future? To do that we can turn to Time Series Forecasting and use the data from 2001-2019 to predict what will happen to monarch butterfly numbers for the next 10 years. To do this we will use our summary table from Exhibit 1. One challenge we have is this data is highly skewed, which makes it more difficult to model. The first thing we want to do is apply a log transformation to the number of occurrences for each year. Once we have done that we can move forward with the Time Series Forecasting using the log N(Occurrences).

Exhibit 5

Time Series Forecast

Group Summary of Time

N Rows	Missing Value	Unique Timestamps	Start	End	Missing Record
18	No	Yes	2001	2019	Yes

Group Summary of Log[Sum(Occurrences)]

Min	Max	Length	N Obs	N Missing	All Positive	Comment
0	7.470224	19	18	0	No	

Model Reports

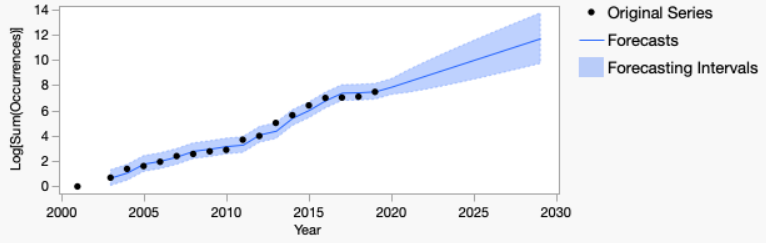
Log[Sum(Occurrences)]

Model Type

Model Abbreviation	AAN
Error Type	Additive
Trend Type	Additive
Damped	No
Season Type	None
Season	
Constrained	Yes

Model Report

Series and Forecasts



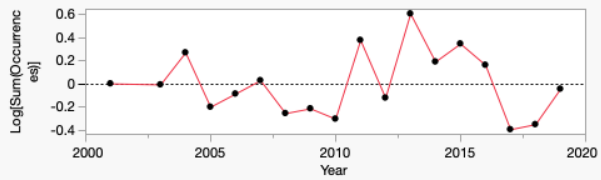
Model Summary

-2Loglikelihood	4.675229
AIC	14.67523
BIC	18.8413
Nparm	5
Sigma	0.318234

Parameter Estimates

Parameter	Estimate	Std Error	Lower 95%	Upper 95%
alpha	0.99990000	0.22345001	0.5619460	1.4378540
beta	0.00298495	0.18349011	-0.3566491	0.3626190
l0	0.27746608	0.29353422	-0.2978504	0.8527826
b0	0.42447483	0.07138328	0.2845662	0.5643835

One-Step-Ahead Forecasting Errors



(Using the log N(Occurrences) from the Summary Table from Exhibit 1 Analyze > Specialized Modeling > Time Series Forecast, drag and drop Log Sum(Occurrences) and Year into the Time. A dialogue box opens. Click on the Complete Specifications tab and click Select All for both the Additive Error Models and Multiplicative Error Models and check the Preserve Models Criterion box. JMP will automatically select the best-fitting model. Then click Run.)

What we can see in Exhibit 5 is that JMP selected the AAN model as the best fit based on the Akaike information criterion. The AAN model is an additive error model. In the parameter estimates table, b0 represents the growth. For this forecast model, the parameter estimate *beta* is close to 0, so *b0* can be interpreted as the constant growth rate over time; because we used the log transformation, the year-over-year growth rate on the original Total#Occurrence is *100% * (exp(0.42)-1)=58.* If we save the results from the Time Series Forecasting menu, we can check the One Step Ahead Predictions and see the log predicted number of butterflies. Then we can apply the antilog to see the number of butterflies predicted for the next 10 years. Based on the data from 2001-2019, it is expected that the monarch butterfly abundance will continue to increase into 2029.

Summary

When we explored our data, we saw that it appears butterfly abundance is increasing over time, especially in California. We saw in our analysis that the fraction of urban land conversion was positively correlated with monarch butterfly densities and that there was a great deal of variation depending on the county that was sampled. According to our forecast, monarch butterfly populations appear to be rebounding. Scientists aren't sure if this means the species is recovering after decades of decline or if it is an anomaly. Only time will tell.

Statistical Insights

In this case, we see the value of data collected by citizen scientists. The data on the occurrence of monarch butterflies across counties is variable, however, because the sampling effort is not homogenous. To address this, we can consider counties as a random effect in the GLMM. When it comes to forecasting future numbers of butterflies, the occurrence data were highly skewed, requiring a log transformation for proper analysis. We also learned that:

- Urban land conversion is positively correlated with increasing monarch numbers.
- Monarch butterfly populations are increasing year after year by an average of 44%.

The ability to predict future butterfly populations depends on many factors, including the sampling effort across their range. This effort is not evenly distributed, potentially reducing our confidence in the results.

JMP Features and Hints

This case study uses the Graph Builder to compare monarch butterfly occurrences across states and Tabulate to produce summary statistics across states, counties, and years. For data analysis, we used GLMM to evaluate the effect of fixed and random effects on the occurrence of monarch butterflies and Time Series Forecasting to predict future values.

Exercises

As we saw with the iNaturalist dataset, monarch butterflies appear to be increasing in abundance. As mentioned earlier, NABA performs annual counts over repeated areas. This dataset is very different from the iNaturalist dataset in some key ways. First, the iNaturalist relies on opportunities as they arise, whereas NABA data collection is systematic in nature.

Second, NABA data comprise systematic counts of repeated areas 15 miles in diameter, while the iNaturalist data are single observations at random locations. Scientists may use either of these datasets, and the question is whether or not they produce the same result given the differences in sampling. We will explore this using NABA data collected across the same western states from 2001-2019.

The Data NABA Monarch butterfly.jmp
NABA Monarch butterfly (by County).jmp

The butterfly data are open access and available online at: https://github.com/jcoliver/citsci-western-butterflies.

Source: Forister, M. L., Halsch, C. A., Nice, C. C., Fordyce, J. A., Dilts, T. E., Oliver, J. C., ... & Glassberg, J. (2021). Fewer butterflies seen by community scientists across the warming and drying landscapes of the American West. *Science*, *371*(6533), 1042-1045.

1. Exploring data is always a good first step. Explore how monarch butterfly abundance has changed over time.
2. Next, produce a time series plot for the occurrence of monarch butterflies in each of the western United States between 2001-2019. What do you notice?
3. Using a Generalized Linear Mixed Model, test the influence of land use and climate on monarch butterfly densities with county as a random effect.
4. After running these analyses, what can you say about the role of climate and land conversion on monarch butterfly densities?
5. Using JMP's Time Series Forecasting, what can you predict about monarch butterfly populations in the next 10 years? Is the outcome the same as what we saw with the iNaturalist dataset?

Bonus Question

1. Using the forecasted growth expected for monarch butterfly populations, download the data for 2020-2023 from iNaturalist and determine if the estimated growth is similar to the data reported in the iNaturalist database.