The butterfly effect

A Case for Biodiversity using Machine Learning

Capstone Project - Sprint 2 - Apr 2, 2024 - Chloe Cousin

What are we looking at?

6,392,186
butterfly sightings

2001 - 2020

United Kingdom

6,392,186
butterfly sightings

2001 - 2020

United Kingdom

Dataset 1 is merged with:

Dataset 2

Butterflies Traits
(wingspan, flight duration...)

Dataset 3

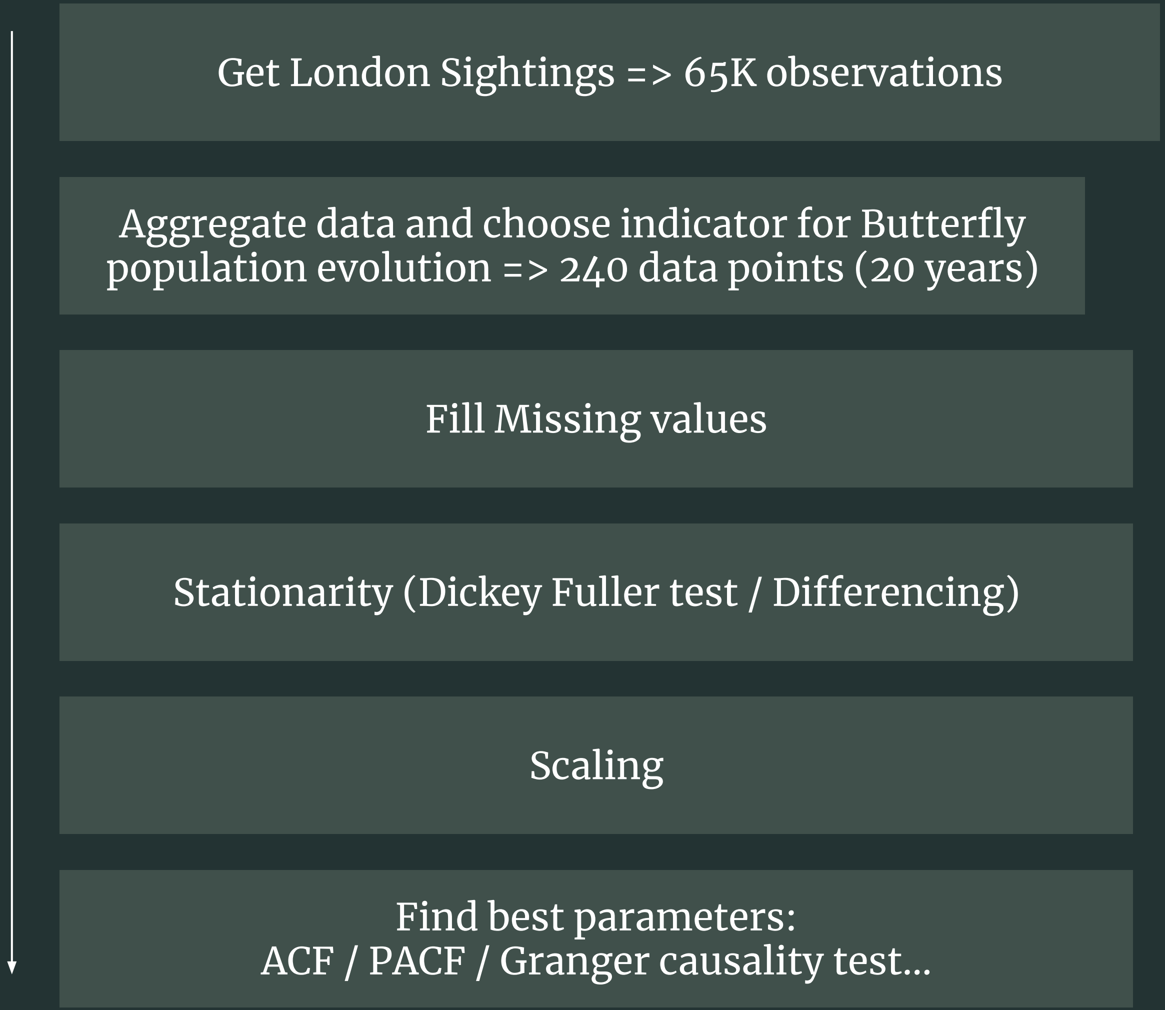
List of endangered
butterfly species in the UK

Dataset 4

External factors:
weather & air quality data

Preprocessing steps

(Depending on the model)



Get London Sightings => 65K observations

Aggregate data and choose indicator for Butterfly population evolution => 240 data points (20 years)

Fill Missing values

Stationarity (Dickey Fuller test / Differencing)

Scaling

Find best parameters:
ACF / PACF / Granger causality test...

What do we want to predict?

Using a **Time Series model**,

an **estimation of the butterfly population evolution** over the next years

In the context of **climate change**

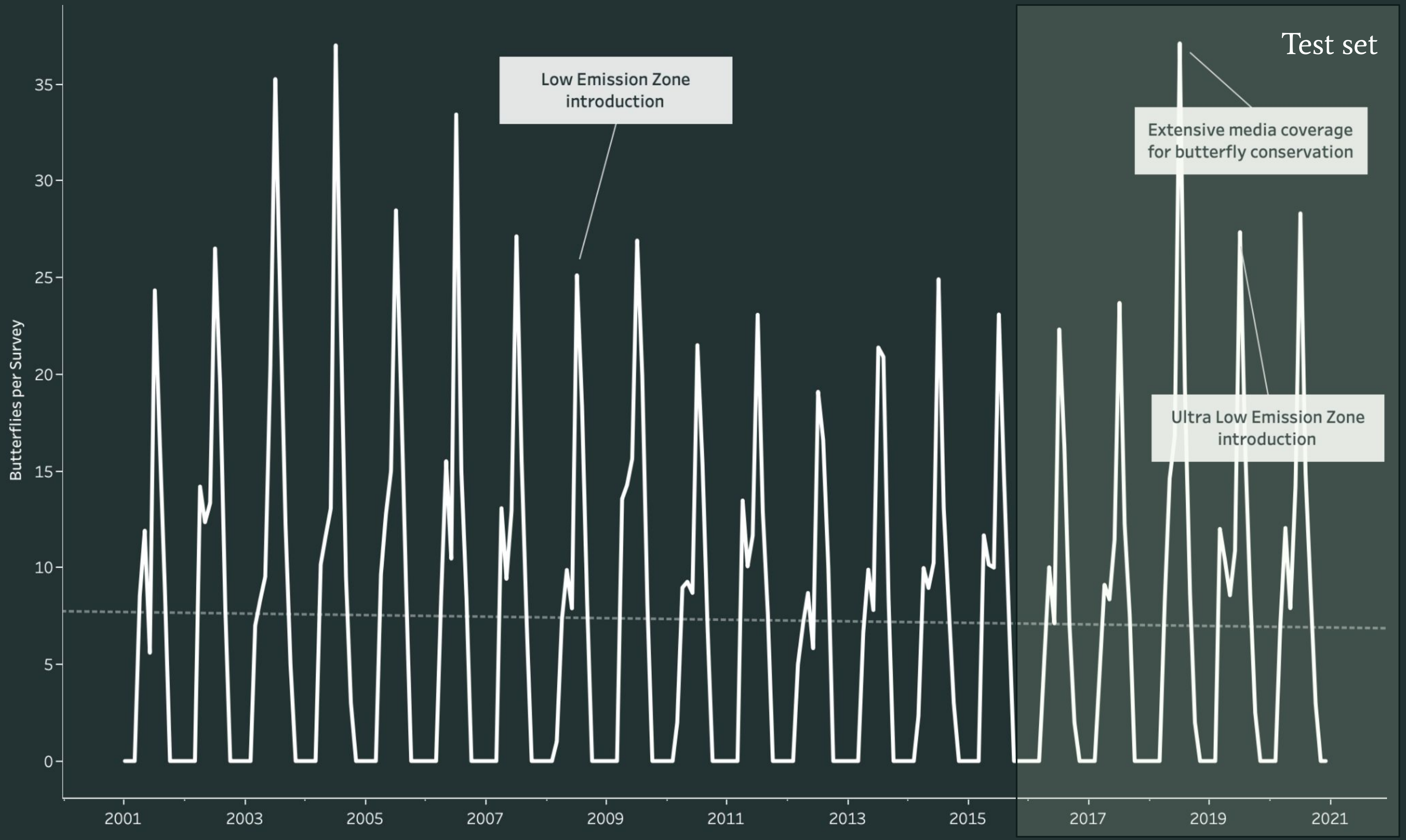
Why should we care?

The World Economic Forum states that **Biodiversity is 'critically important'** for 5 reasons, as it:

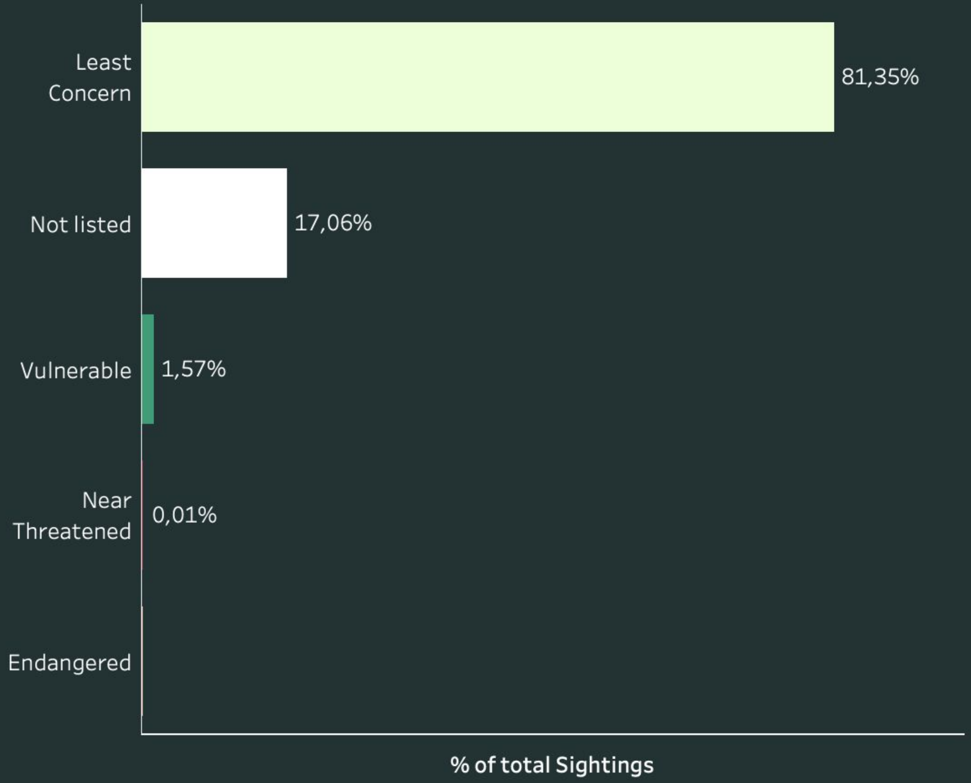
1. Ensures health and food security
2. Helps fight disease
3. Benefits business
4. Provides livelihood
5. Protects us

Why butterflies? The short life cycles are thought to be one of the best indicators of how healthy an environment is.

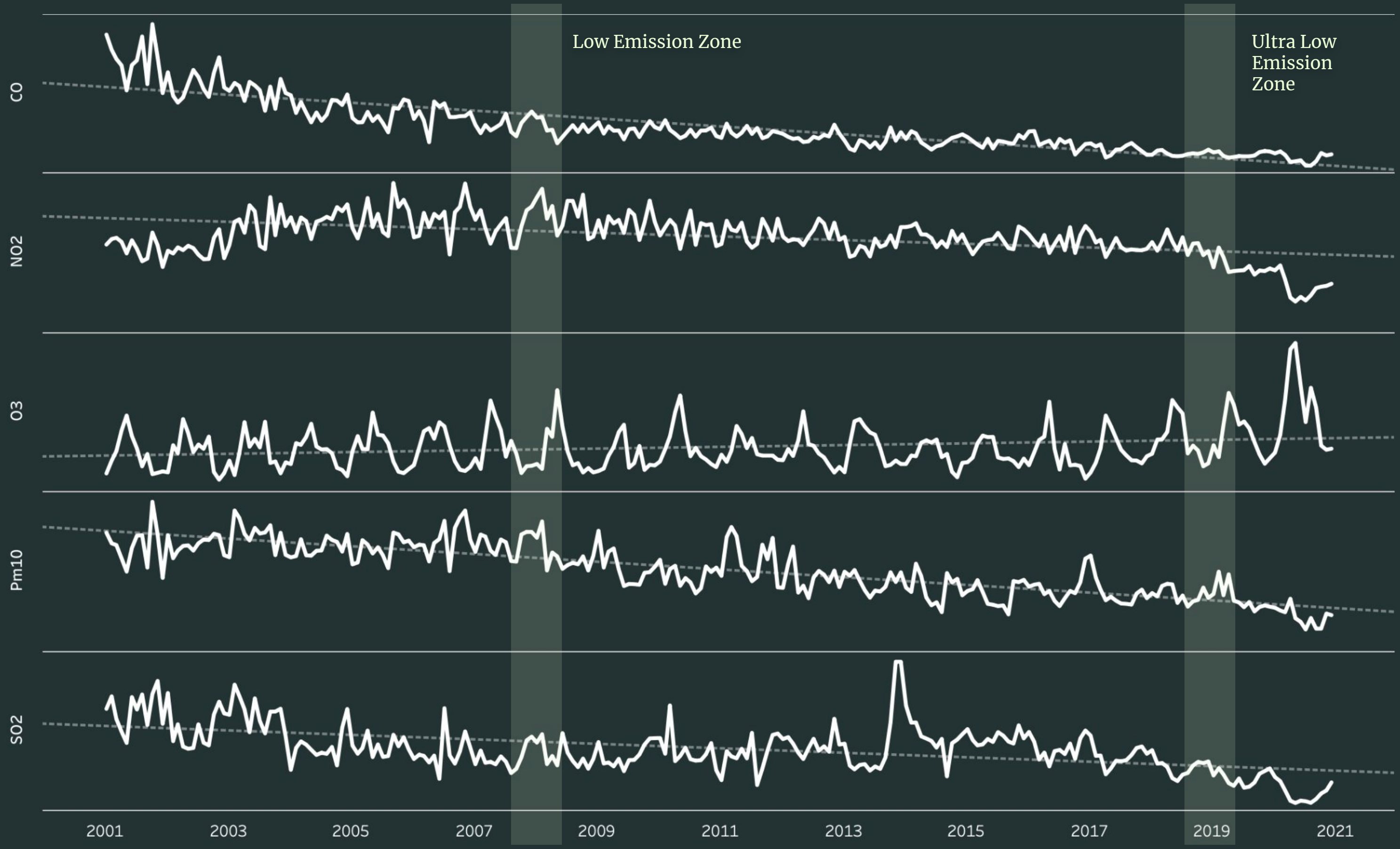
EDA initial findings


Test set
Low Emission Zone introduction
Extensive media coverage for butterfly conservation
Ultra Low Emission Zone introduction
Butterflies per Survey
35
30
25
20
15
10
5
0
2001
2003
2005
2007
2009
2011
2013
2015
2017
2019
2021

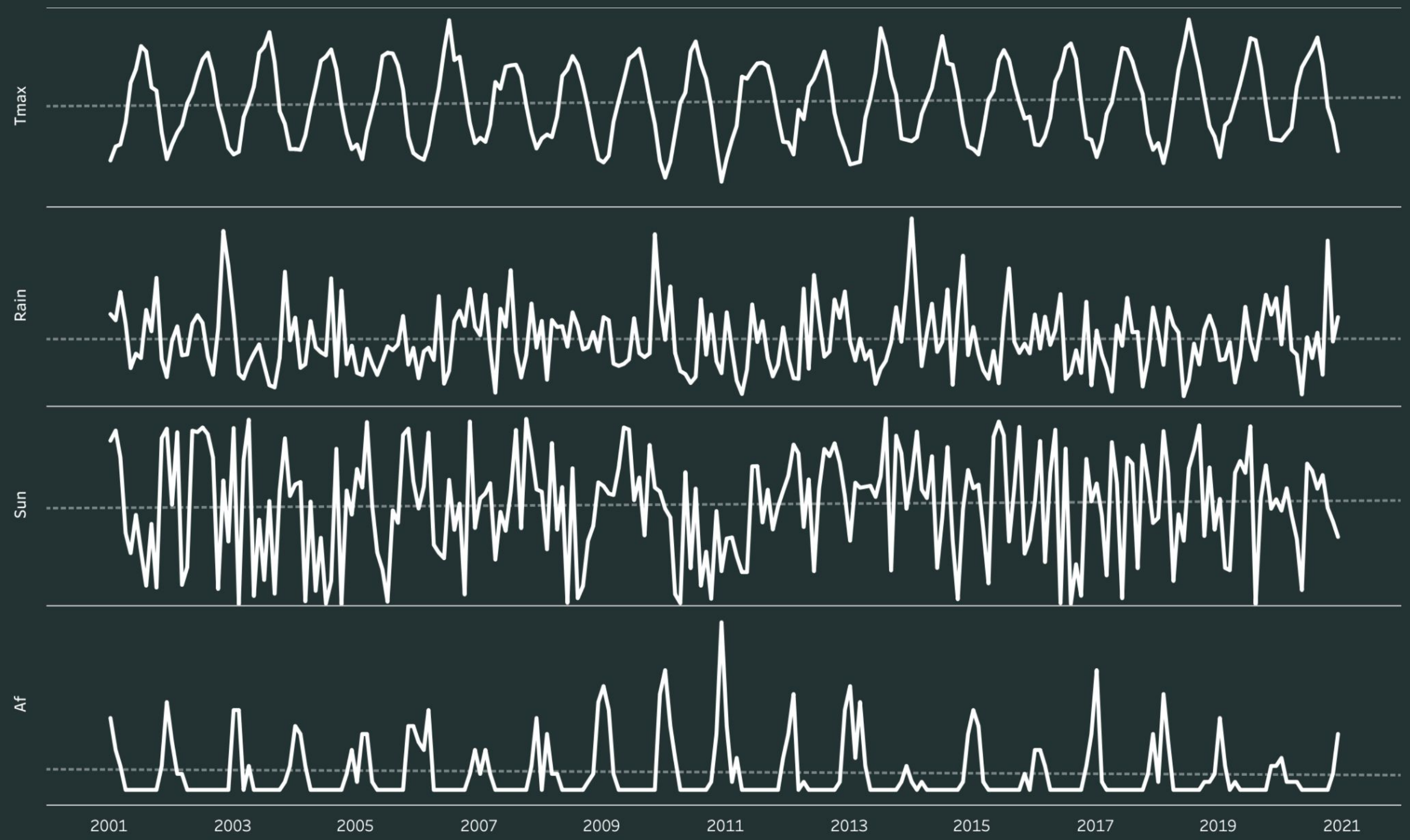
Most of the butterflies seen in London are not considered vulnerable



Air quality: most trends are down



Weather: temperatures are up



Complexities of Predicting Biodiversity Population Evolution

COLLECTING DATA

- Counts: Warm weather leads to some butterfly species being past their peak in numbers by the time the surveys starts each year

- Population estimation: impossible to account for all butterflies

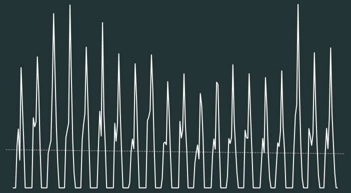
- Human observation: the methods for collecting data are very thorough at the UKBMS (Butterfly Monitoring Scheme), but approximations due to human action are still possible

UNDERSTANDING GLOBAL DYNAMIC FROM DATA

- External factors: not only weather and air quality factors, but also conservation efforts, population migration, plants density...

Baseline models and evaluation

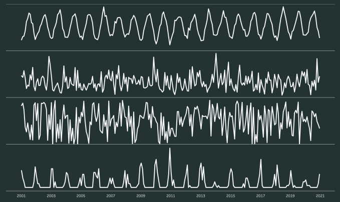
Univariate models



1 variable impact the predictions

> Butterflies per Survey evolution <

Multivariate models



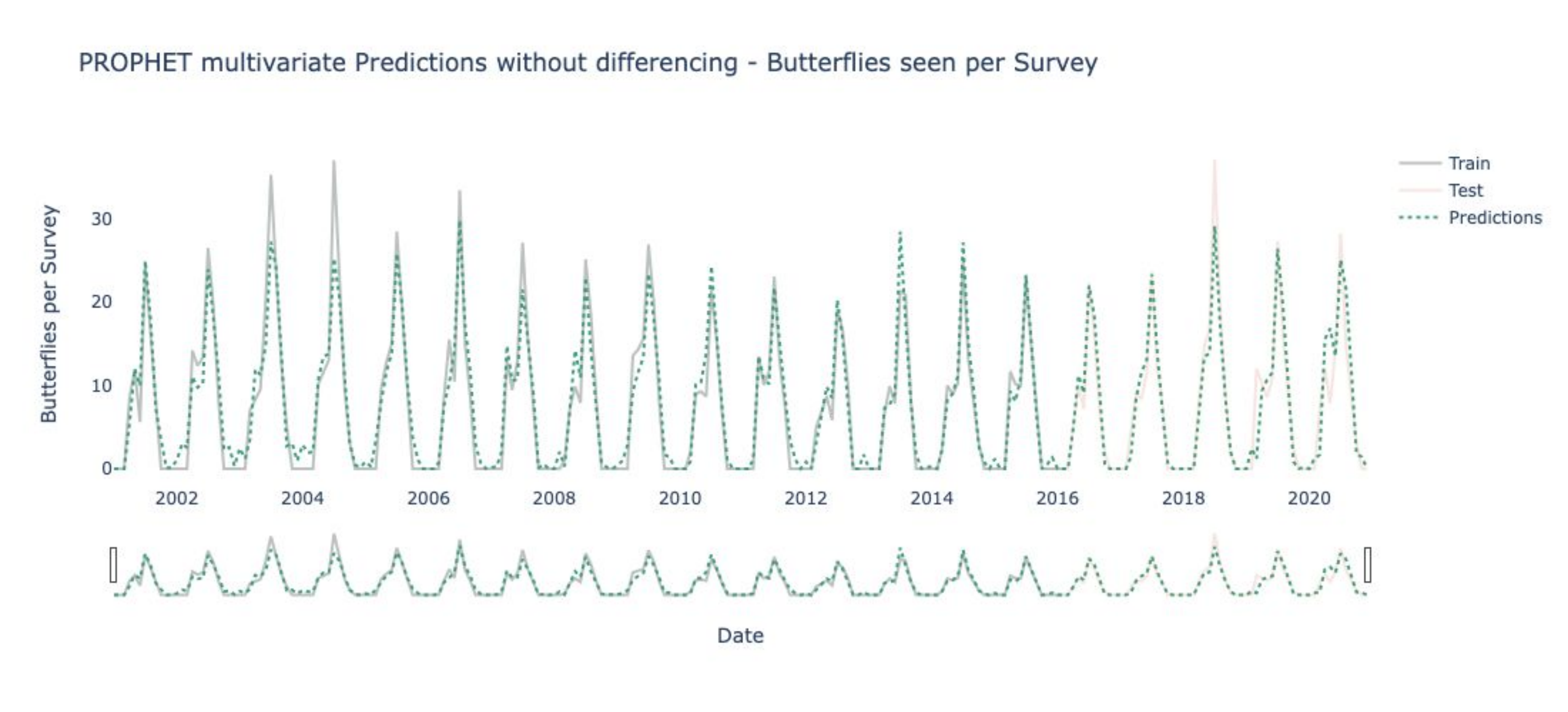
Several variables impact the predictions

> Butterflies per Survey evolution <

in the context of weather and air quality data

Models results

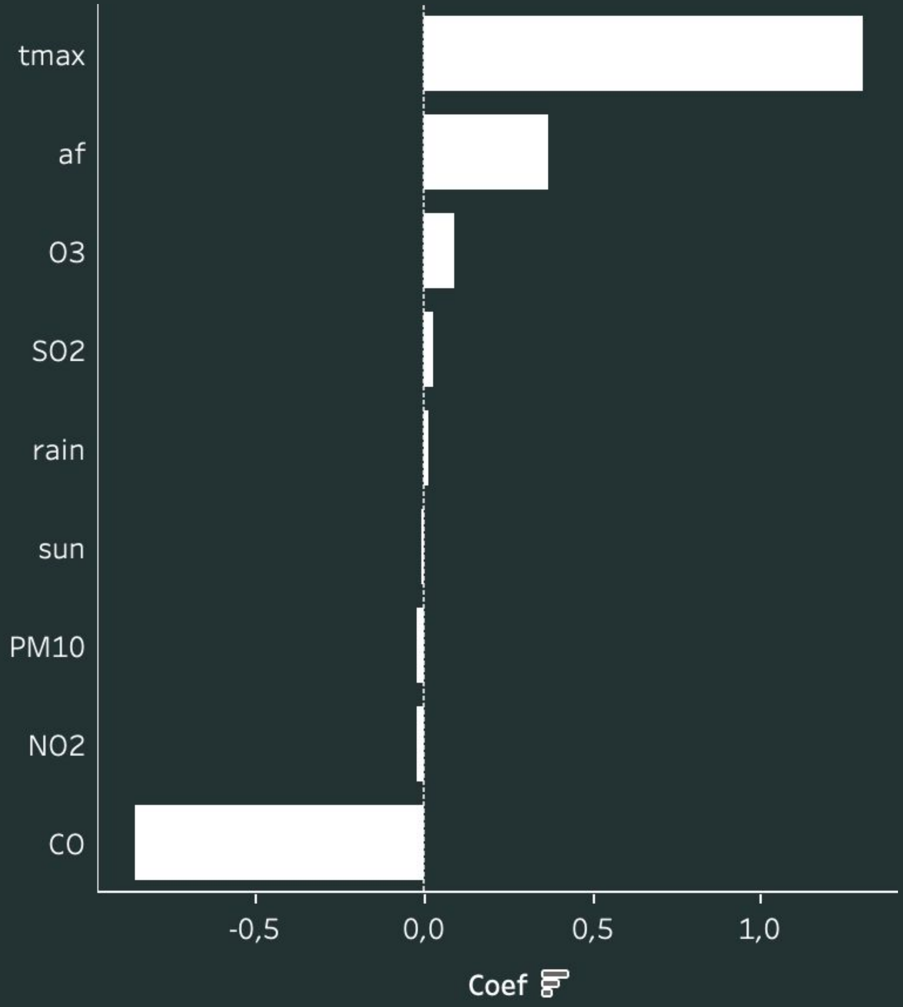
	Univariate models		Multivariate models		
	SARIMA	Prophet	VAR	Prophet No differencing	Prophet With differencing
MAPE train	15.3%	11.2%	16%	10%	9.3%
MAPE test	20.8%	21.8%	24%	17.9%	16.6%
Delta MAPE train/test	5.5%	10.6%	8%	7.8%	7.3%
Independent variables with highest impact	-	-	Rain Ground level ozone	Temperatures, Air frost, CO	Temperatures, frost, Particles, SO2, NO2
Notes	Predictions on test set seem to pick up only seasonality		Predictions on train and test seem to pick up only seasonality	Picks up sudden increase in 2018	12 months shift for all variables except the target one (see impact of external factors on the next generation) Picks up most trend movements



Prophet - multivariate 1

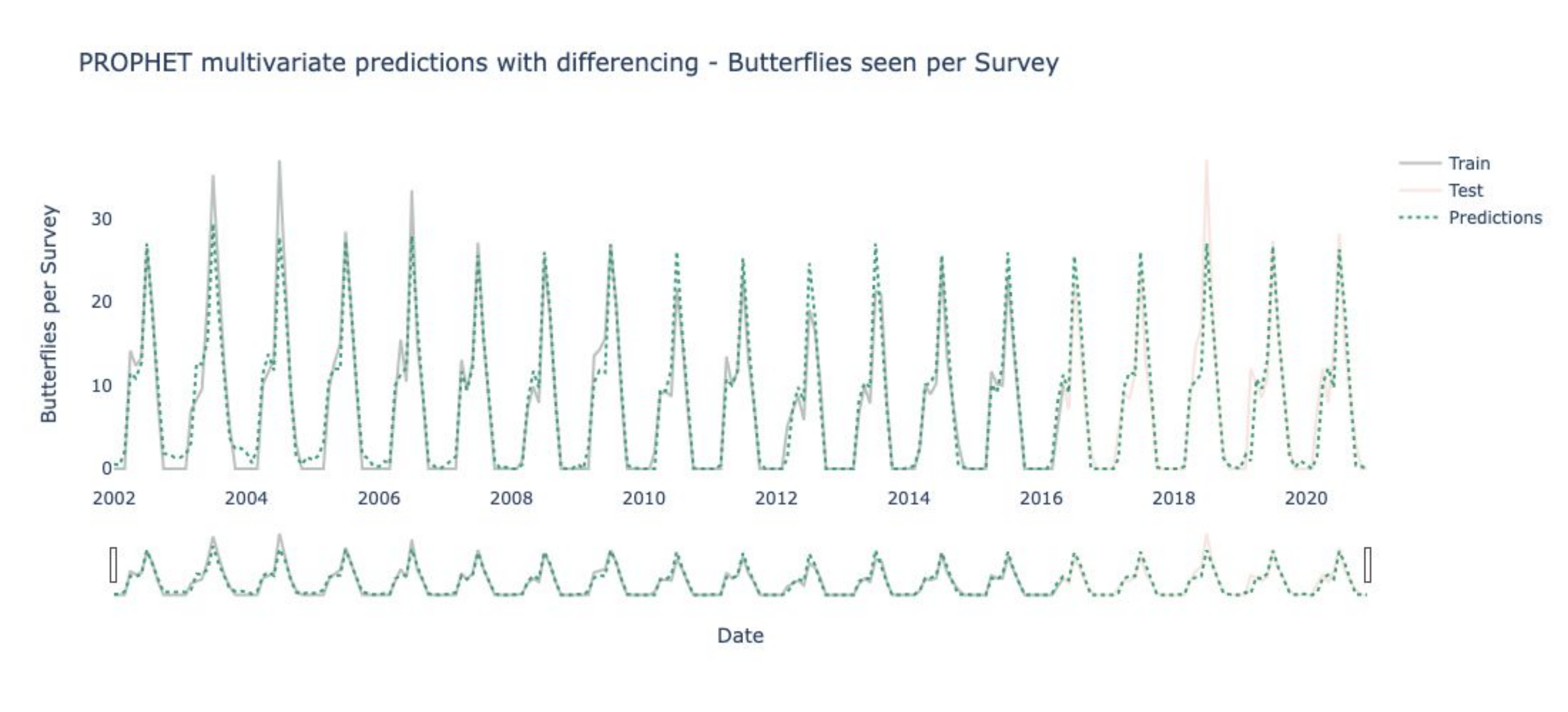
MAPE Train set: 10%

MAPE Test set: 17.9%



Apart from temperatures which follow the butterflies seasonality,

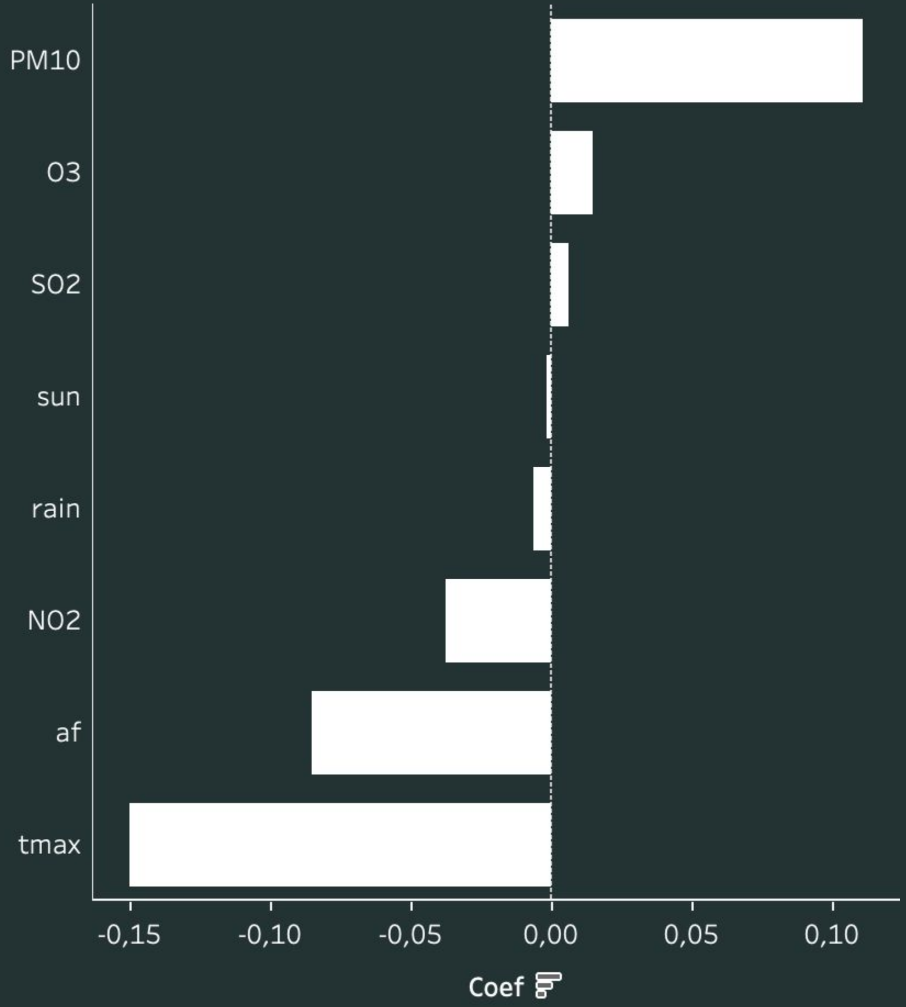
the carbon monoxide seems to also be predictive of the butterflies presence



Prophet - multivariate 2

MAPE Train set: 9.3%

MAPE Test set: 16.6%



Impact of the external factors on the following year:

When the temperature has been high in the previous year, there seems to be fewer butterfly observations.

PM10 and SO2: follow the same downward trend as the butterflies, so as the values are high, so are the butterflies'.

Interesting insights which will be interesting to explore further in sprint 3.

Next steps

Next steps

> Try other models

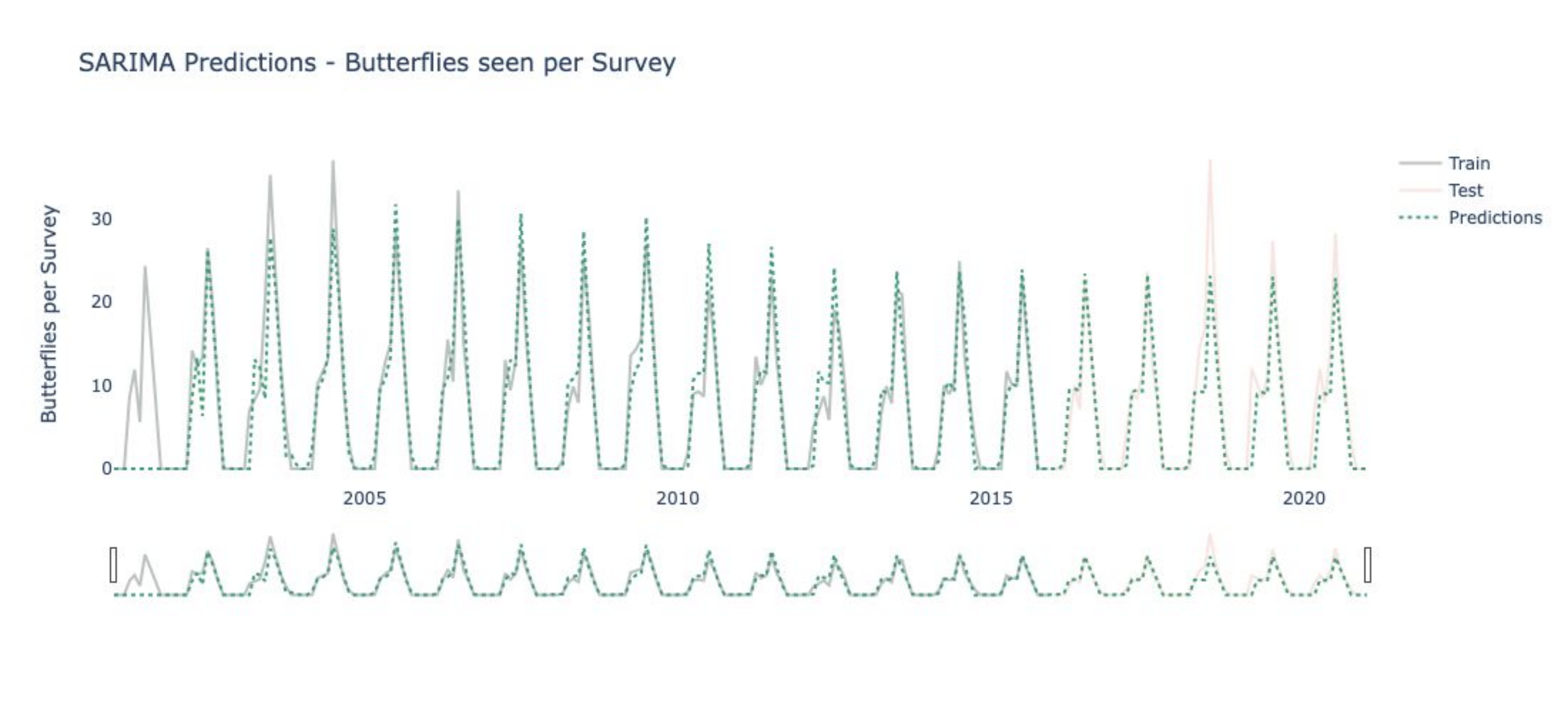
> Hyperparameters tuning

> Try trend-only multivariate analysis (without seasonality)

> Add predictions for weather and air quality data from official sources + good/bad scenarios, see how this affect the butterflies

Thanks!

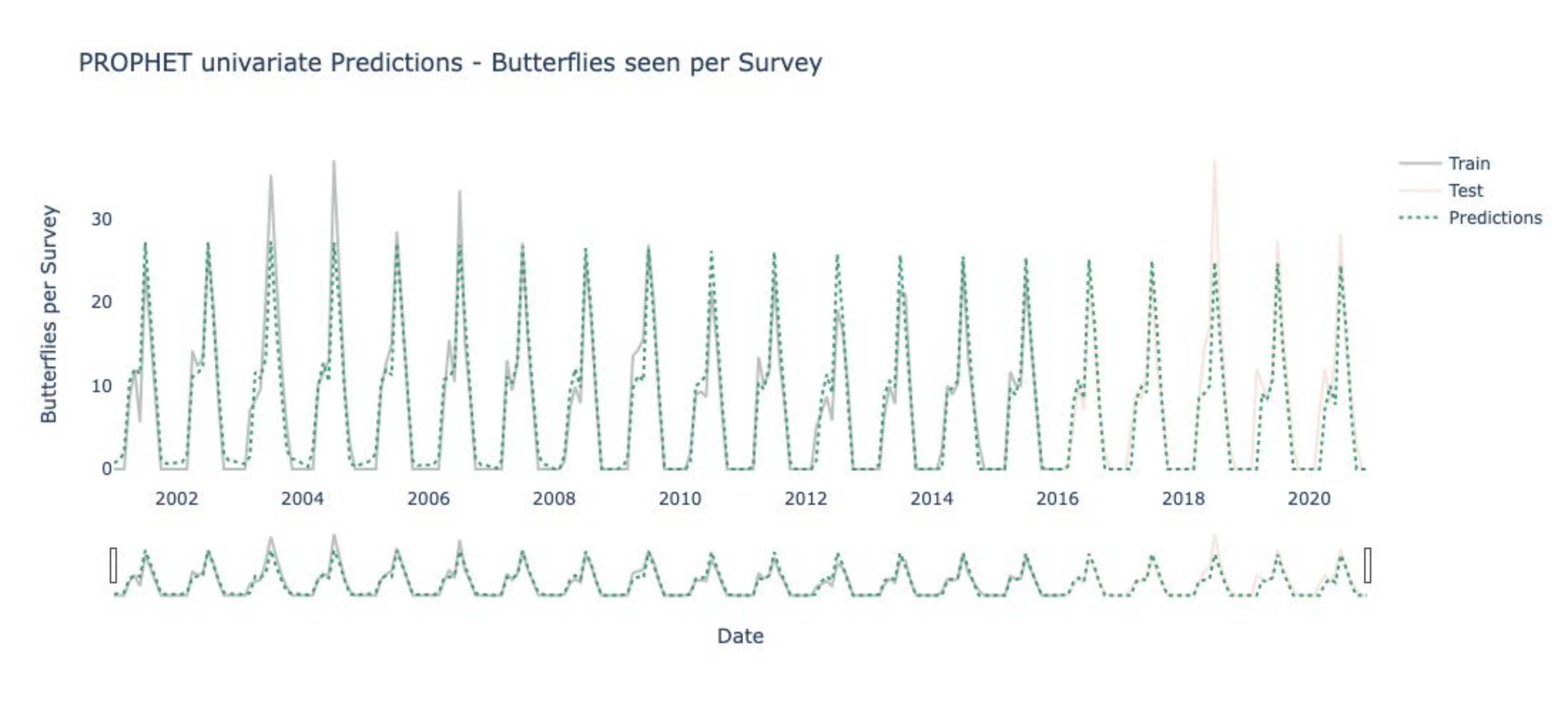
Appendix: Other models results



SARIMA - univariate

MAPE Train set: 15.5%

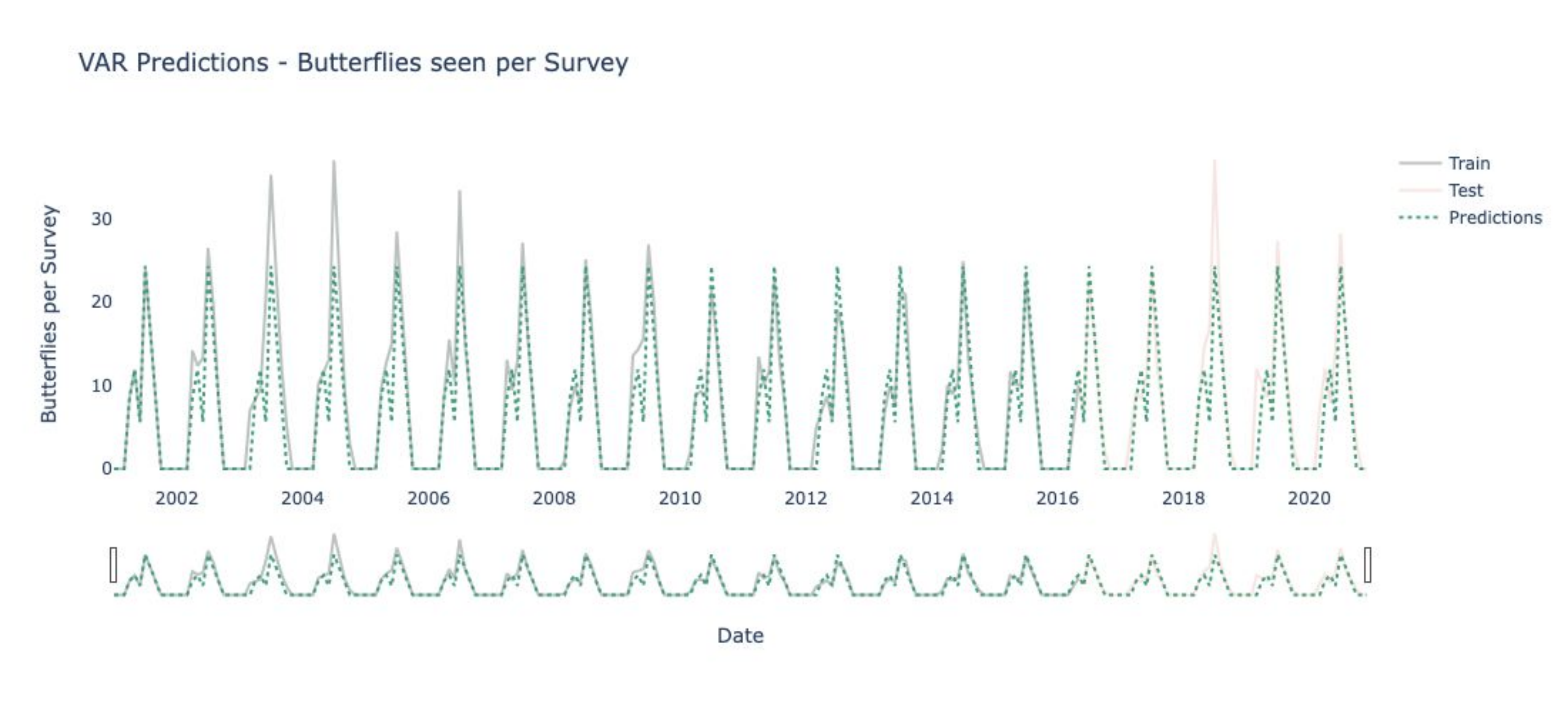
MAPE Test set: 21.5%



Prophet - univariate

MAPE Train set: 11.2%

MAPE Test set: 25.5%



VAR - multivariate

MAPE Train set: 16%

MAPE Test set: 24%