Fruit Tree Pollination

Analysing existing study data to see how bees and other pollinators contribute to crop success

Paul Sabin   16 February 2024

Contents

01

Introduction

Client - Research Topic - Task

Client

Start-up company in a consortium with:

- Universities
- Beekeeper associations
- Farming associations

Wants to conduct research on pollinators and crop interactions.

For this project I worked with a German start-up that provides digital services for beekeepers and fruit farmers that rely on pollination from insects for their crops. This company has teamed up with a number of European universities, beekeeper associations and farming associations and is applying for EU funding to carry out a major research project. Ultimately, this consortium aims to develop sophisticated AI algorithms that can analyse and predict the interactions between specific insects and a selection of pollinator-dependent fruit crops.

Role of pollinators
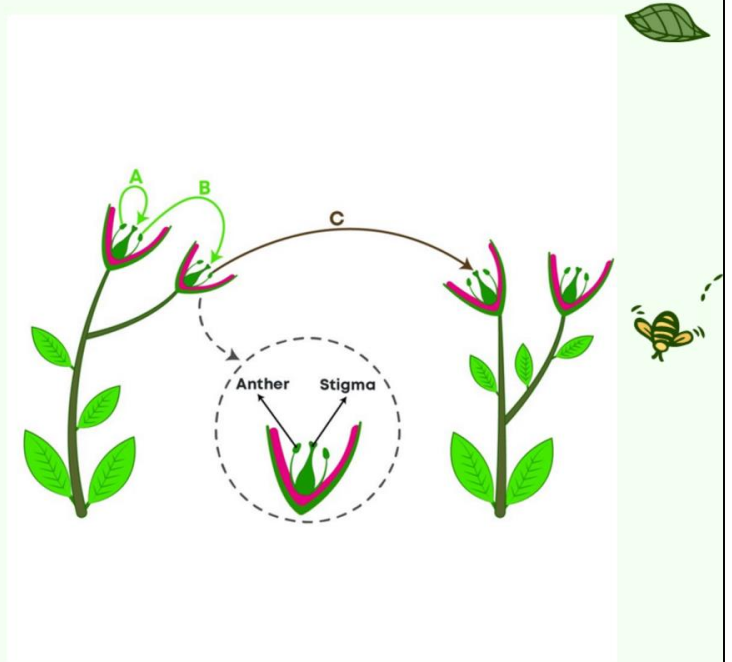
A quick reminder of the biology: pollination is when pollen is transferred from the male to the female reproductive organs of plants, which is essential for fertilisation and seed production. While some species can self-pollinate, fruit trees typically require cross-pollination between different varieties, so they rely on insects such as bees to transfer pollen from one flower to another as they forage for nectar in the orchard.
In this context, the global decline in bee populations – owing to climate change, agricultural practices and other factors – presents a risk to successful fruit cultivation.

So when we say pollinators, which insects are we talking about? We all know about honey bees, and beekeepers have a mutually beneficial relationship with farmers, because when they bring their hives to an orchard, for example, the farmer is receiving a pollination service as the bees transfer pollen between the flowers, while the bees collect plenty of nectar for producing honey.

However, honey bees are not the only managed pollinator species. Bumble bees are also bred and sold to farmers. Mason bees, which forage individually rather than as a colony, can be effective pollinators even in smaller numbers. Farmers can create structures they like to nest in to support mason bee reproduction, indeed one of the consortium partners of my client is a company that manufactures these structures for mason bees. Finally, we have wild pollinators, which include bees, hoverflies, moths and butterflies, beetles and other insects.
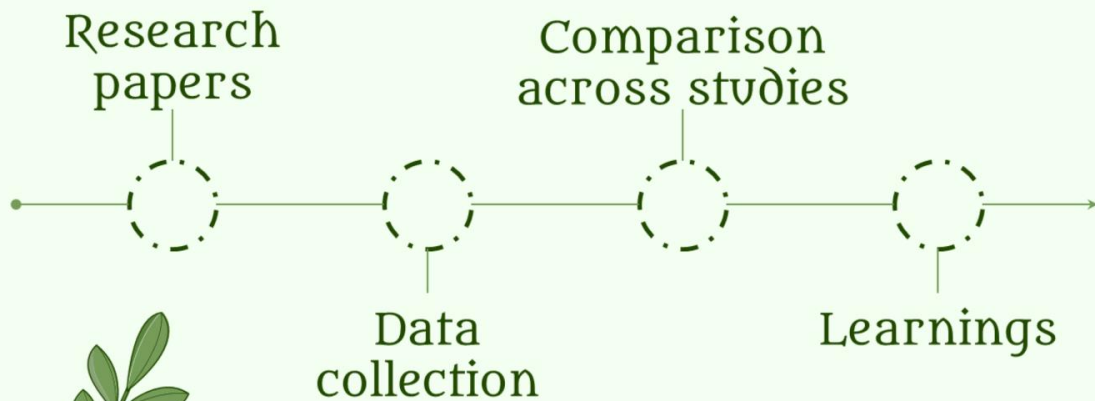
So I was tasked with performing a meta-analysis of existing studies. Specifically, my client wanted to know:

- What academic research already exists on pollinators & crop interaction?
- How do the respective researchers observe and record data in the field?
- Can we gain some insights by analysing data from multiple studies?
- How can these findings inform future research by the consortium partners?

02 Finding Data

Sources – Criteria – Selection

Google Scholar was my main starting point for finding sources, using keywords to search among hundreds of millions of scientific papers. I often found that, once I'd discovered one relevant study on crop pollination, I could find more in the same field by following the citations.
In this way I came across CropPol, a global database on crop pollination. It contains measurements recorded from 202 crop studies worldwide. Finally, an AI-assisted platform called Consensus helped me to track down relevant studies.

"Results from one crop can hardly be transferred to another crop"

—(Wilson and Thomson, 1991)

I've highlighted a quote here because the point was repeatedly emphasised in the academic literature I came across: a certain insect, such as the honey bee, might be great at pollinating one particular crop, but far less effective for another crop. Reasons for this include the respective size and shape of insects and flowers, movements between flowers and tree rows, pollinator preference for flowers of a certain colour and scent, and pollination mechanisms: tomatoes, for example, are served very well by the buzz pollination of bumble bees.

Inclusion Criteria

Crops
Apple, cherry, almond

Location
Crops grown in Europe

Recency
Data collected since 2010

Target
Fruit set / yield

Prunus dulcis

Prunus avium

Malus domestica

With that in mind, let's have a look at the criteria I used to select my datasets.
The client specified three crops to focus on – apple, cherry and almond. As you can
see from the photos below, their flowers are quite similar in appearance.
I decided to narrow my search to European locations in order to limit the amount of
variation in terms of insect species and climate.
Although field data was available going back many decades in some cases, I looked
for data collected since 2010 to capture a recent state of biodiversity and more
comparable farming practices.
Knowing that I wanted to make predictions on the data using machine learning, I
wanted to define a target variable. One common theme among the studies was to
look at the affect of various factors on the success of the crops. This was commonly
measured in terms of "fruit set", which means the percentage of flowers that start to
develop into fruits after successful pollination, or yield, which is the mass or weight of
fruit that is later harvested from a certain field.
So this is what I chose to try to predict. However, it's not simply a given that crop
success should be the target variable, because in these complex ecosystems, effects
can be observed in multiple directions. For instance, the choice of farming practices
can affect insect populations; or the presence of managed honeybees might affect the
visitation rate of wild bees. So there are multiple approaches you could take to
analysing and predicting this crop pollination data.

For this project it was essential to compare data across different studies. I selected 17 studies of apple, cherry and almond crops that were all included in the CropPol database, as this provided some degree of standardization in the data types and collection methods used. This gave me a total of 259 sites where observations were recorded.

Key Features

Fruit set
Reproductive success

Yield
Final crop success

Richness
Number of different pollinators observed

Abundance
How many of each pollinator were present

Visitation rate
Pollinator visits per hour

The CropPol database includes roughly 40 numerical features along with descriptive labels of how the data in the respective study was collected. For instance, some researchers walked straight lines – called "transects" – through a field and recorded all the pollinators they observed on either side, while others used a sweep net to capture pollinators while in flight or resting on plants.

The most important features for my research were the fruit set and yield, which I already mentioned as my respective target variables, along with Richness – the number of different pollinators observed, so you can also think of richness as the diversity of insects in the field – Abundance – How many of each pollinator were present – and Visitation rate – the number of visits each pollinator species made to a flower or a group of flowers in the space of one hour.

Unfortunately, not every study recorded all these features, so I split my data into groups according to the three target crops – apple, cherry and almond – and the presence of key features for comparison.

Correlation

Apple
(Spain)

Fruit set correlated with:
Visitation rate (0.24)
But negative correlation with abundance and richness

Cherry
(Belgium)

Fruit set correlated with:
Richness (0.45)
Abundance Syrphids (0.34)
Abundance Bombus (0.30)
But negative correlation with honey bee abundance

Almond
(Spain)

Yield (fruits/plant) correlated with:
Abundance honey bees (0.53)
But negative correlation with honey bee visitation

Next I took an initial look at which features were correlated with the target variables [see slide].
So my overall impression here was that the data was giving a very mixed picture, with some expected correlations and others that were counter-intuitive.

04 ML results

Simple linear regression

Linear Regression: fruits_per_plant vs ab_honeybee
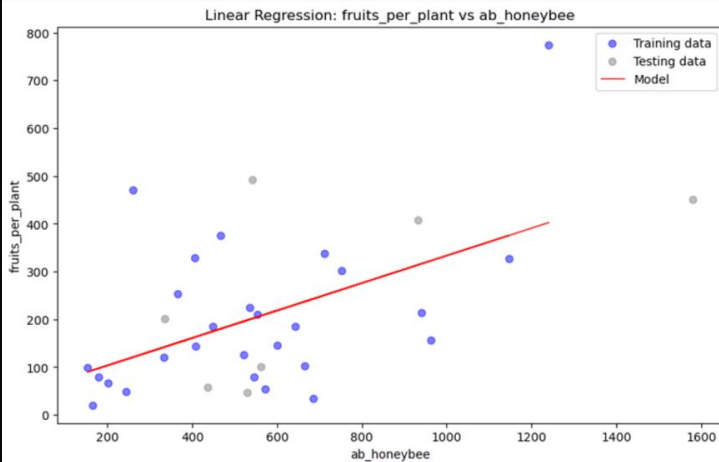
Almonds in Spain, 2 studies with 35 sites

- Coefficient: 2.53
- MAE: 123
- R-squared 0.36

I began the machine learning part of the project by training simple linear regression models on the individual features that correlated with my target variable for each dataset. This is a typical result from my data on almonds. You can see that a higher abundance of honey bees does correlate with a higher number of fruits per plant. However, the large mean absolute error value and modest R-squared value show that the model's ability to predict this crop's yield based on honey bee abundance is seriously limited.

Simple linear regression

Linear Regression: fruit_set vs observed_pollinator_richness
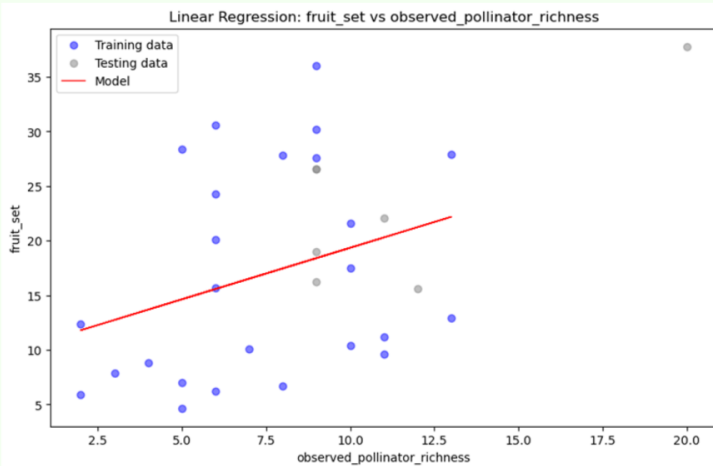
Cherries in Belgium, 4 studies with 32 sites

- Coefficient: 0.945
- MAE: 5.09
- R-squared 0.29

Here is a similar result that attempts to predict cherry fruit set based on the overall pollinator richness. Although we can observe a broad upward trend, the error sizes are very large and the R-squared value shows that only a small proportion of the variance in fruit set can be explained by pollinator richness. This should not come as a complete surprise, since we assume that crop success is influenced by numerous factors simultaneously. To try to capture this in my predictions, I next tried multiple linear regression models using different combinations of features.

Multiple linear regression

Apples in Spain, 2 studies with 46 sites

- Features: richness, abundance, visitation rate, honey bee visits
- Target: fruit set
- Coefficients: 0.33 -0.30, 2.46, 0.89
- Mean Absolute Error: 15.09
- R-squared score: - 0.44

Cherries in Switzerland, 1 study with 32 sites

- Features: abundance & visitation rate
- Target: fruit set
- Coefficients: 1.50, -0.017
- Mean Absolute Error: 20.87
- R-squared score: 0.22

In fact, this led to lower model accuracy in most cases. For example, predicting the fruit set of apples based on a combination of four features resulted in a negative R-squared score, which means this model is performing worse than simply using the mean fruit set value as a prediction every time. The highest R-square score I achieved using multiple linear regression was a modest 0.22, on the data from a Swiss study of cherry orchards.

Random Forest

Almonds in Spain, 2 studies with 35 sites

- Features: richness, abundance, honey bee visits
- Target: fruits per plant
- Mean Absolute Error: 216
- R-squared score: - 0.28

Cherries in Belgium, 4 studies with 32 sites

- Features: richness, abundance, honey bee abundance, bumble bee abundance, wild bee abundance
- Target: fruit set
- Mean Absolute Error: 6.13
- R-squared score: 0.34

As an alternative to the linear approach, I tried training a random forest model on my data sets, again experimenting with different combinations of features to try to improve the model accuracy. The results were very mixed, again resulting in negative R-squared scores in some cases. The most accurate random forest model showed an R-squared score of 0.34, predicting the fruit set of cherries based on 5 features of pollinator richness and abundance.
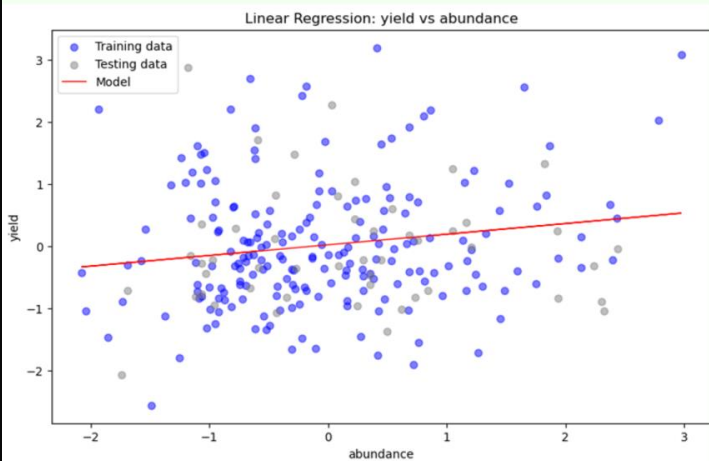
Exploring
further afield

By this stage, the limitations of my analysis are becoming increasingly clear. By dividing up my data into the most directly comparable groups, I've reduced the number of data points for training and testing my models in each case. Correlation with my target variable is fairly weak to begin with, and more sophisticated machine learning models are not proving any more accurate with their predictions of crop success.

This is when I decide to try an approach I'd seen in a frequently cited paper by Lucas Garibaldi, who standardized the data from numerous crop pollination studies using z-scores prior to his analysis. Following this method allowed me to compare data across a larger number of studies.

Normalized data

Linear Regression: yield vs abundance

Legend:
- Training data
- Testing data
- Model

(x-axis: abundance, y-axis: yield)

Apples worldwide, 17 studies with 248 sites

- Coefficient: 0.172
- MAE: 0.73
- R-squared - 0.05

So this is my normalized data. Here you can see a simple linear regression model of total pollinator abundance versus yield using all 17 studies of apples worldwide from the CropPol database. As you can see, the axes now range from 0, which is the mean value for that feature within the individual study, and up to three standard deviations either side of the mean. Despite having a much larger set of training data, the correlation remains weak and the model accuracy is very low.

Normalized data

Linear Regression: yield vs visitation_rate
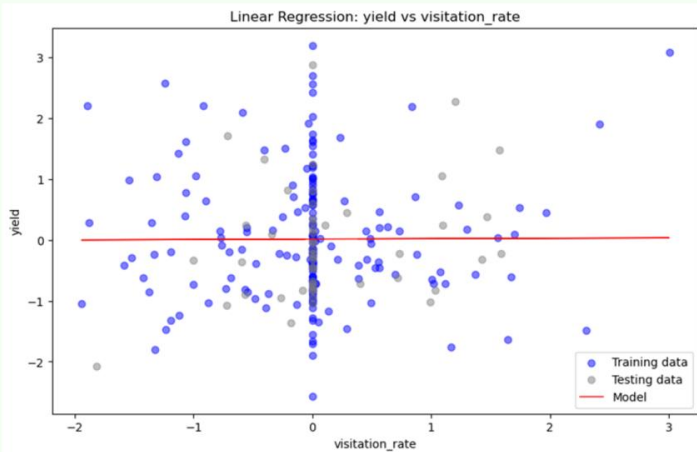
Apples worldwide, 17 studies with 248 sites

- Coefficient: 0.007
- MAE: 0.73
- R-squared - 0.00

What's more, normalizing the data doesn't help me tackle the issue of missing data. In fact, the more studies I compare together, the greater the problem of data sparsity on individual features. You can see that illustrated on this plot of overall visitation rate versus yield. A large proportion of data points are sitting here on the mean line because the respective studies only collected data on pollinator abundance, not visitation rate.

I'll round off the presentation by summarising some of the challenges that arose in this project and giving some key takeaways.

One major challenge is gathering enough data to perform a meta-analysis. The fact that individual crops require separate analysis already restricts the availability of data for comparison.

What's more, merging data from a greater number of studies leads to more missing features as well as variation in the field observation methods that may obscure patterns in the data. And this is even despite the efforts of the CropPol database creators to bring a degree of standardization to the features in crop pollination studies.
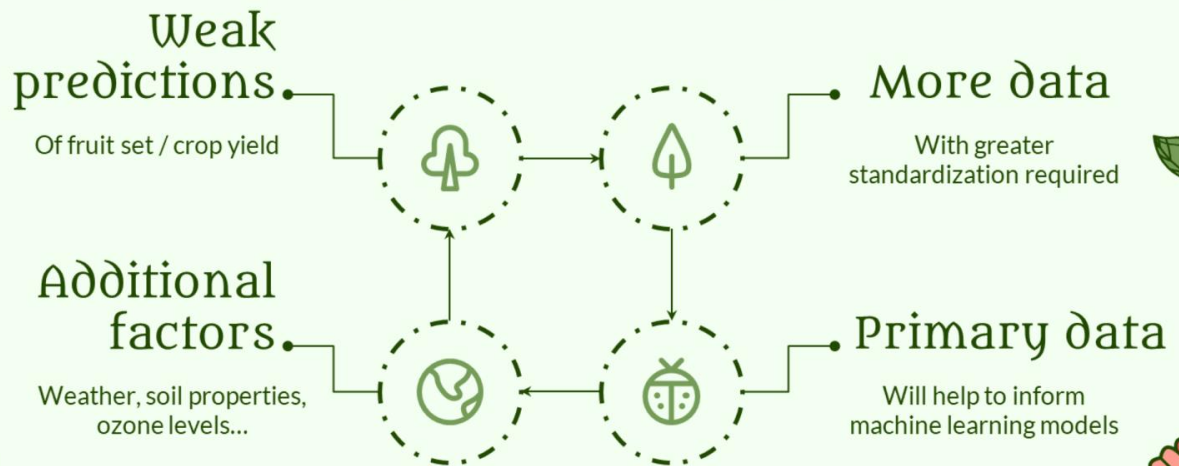
Although normalization using z-scores enabled me to increase the sample size for each crop by integrating data from more varied studies, this approach can lead to conclusions that don't reflect the diversity of the original data.

Finally, I'll leave you with some key takeaways.

Overall the data showed some correlation between pollinator abundance, species richness and the target feature of fruit set or yield. However, they were too weakly related to enable accurate predictions of crop success. Insufficient data was available to determine the influence of individual pollinators, such as honey bees versus bumble bees or wild bees, on these fruit crops.

To merit further exploration, particularly with more sophisticated machine learning models, it will be necessary to gather not only more data, but more data sets that share the same features and observation methods.

It's worth recalling that the presence and activity of pollinators is only one part of the story. Fruit trees also depend on weather conditions, soil properties and other factors to reproduce and grow fruit to maturity. When comparing data across studies, particularly from different countries, there is a risk of masking these influences.

Finally, in the context of my client's wider project, a meta-analysis of existing studies is one of numerous elements that will feed into the development of machine learning models in an iterative process. This will also include primary data on the specific crops and pollinators of interest conducted by the university partners in the consortium. Together, the combination of primary and secondary data will help to build a more complete picture of plant-pollinator interactions and their importance for the success of fruit crops.

Thanks!

Any questions or feedback?
paul@sabinwords.com