Better Working World Data Challenge

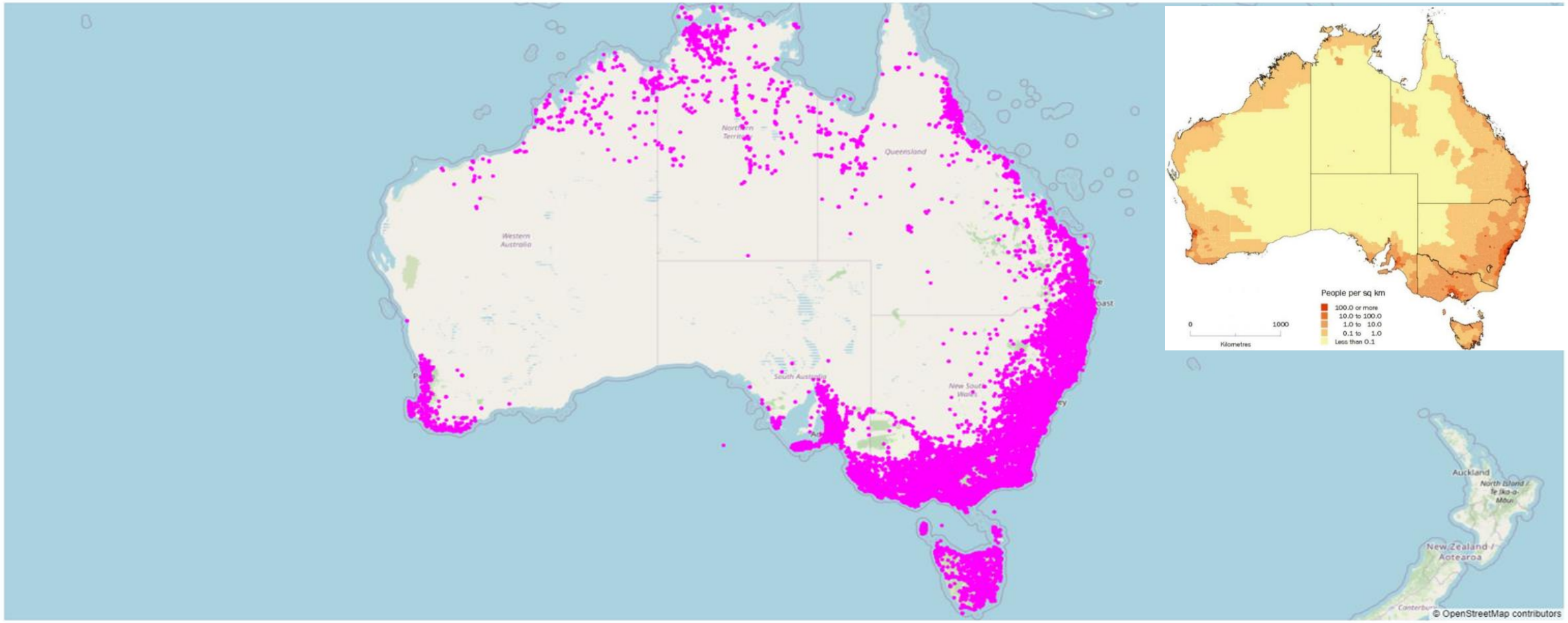
Biodiversity Study

TEAM # 24047
Ke Xu/Qiao Qin/Xueyan Geng/Muyan Cheng

Observation Data Selection

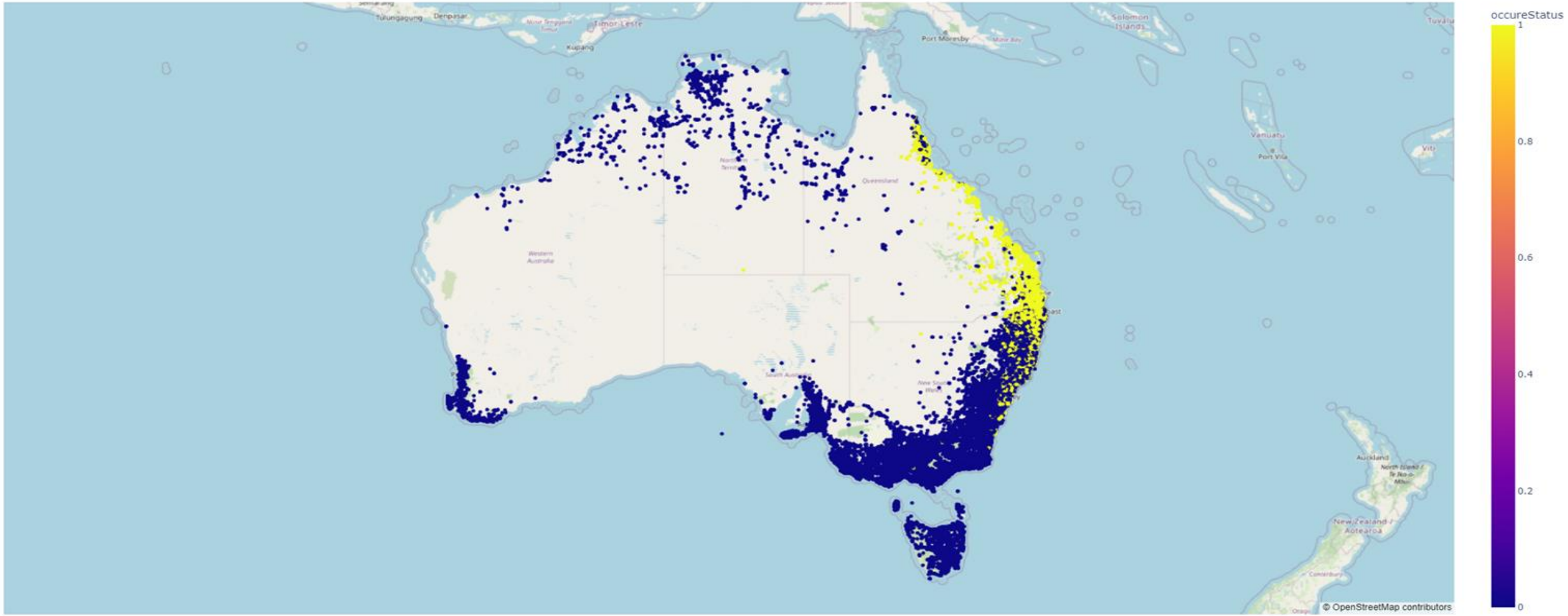
Observation Data First Impression

There are a total of 193,791 instances in the Frog occurrence dataset , of which 188,020 instances are from Australia (97.022%). We initially thought that using Australian data would effectively represent all the data.



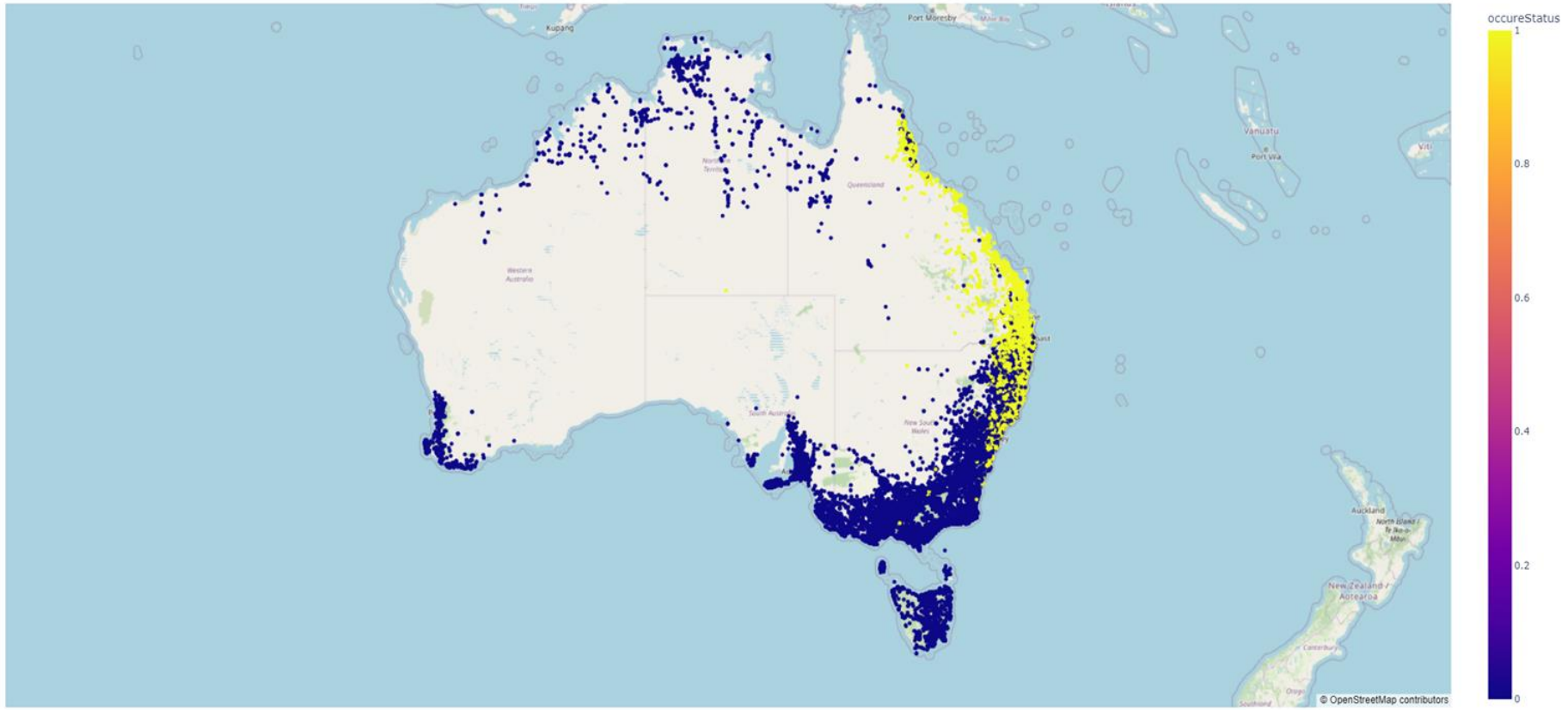
Observation Data Selection

There are 47,332 Litoria Fallax observations and 140,688 non-Litoria Fallax observations in our selected training data. The percentage of Litoria Fallax observations is about 33.64%.



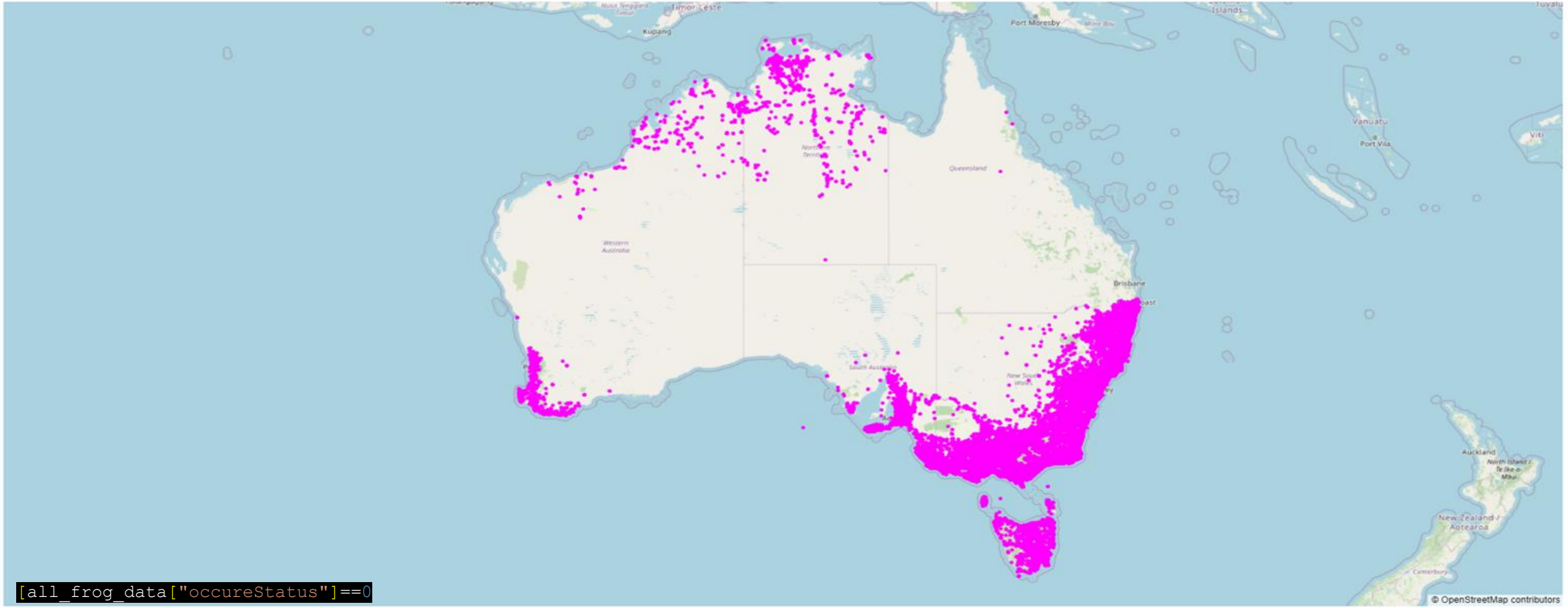
Sample Imbalance Resolution

- The low percentage of Litoria Fallax observations would cause serious overfitting, so we decided to remove a certain number of non-Litoria Fallax observations.
- The new dataset consists of 47,332 Litoria Fallax observations and 56,275 non-Litoria Fallax observations. The percentage of Litoria Fallax observations is about 45.68%.



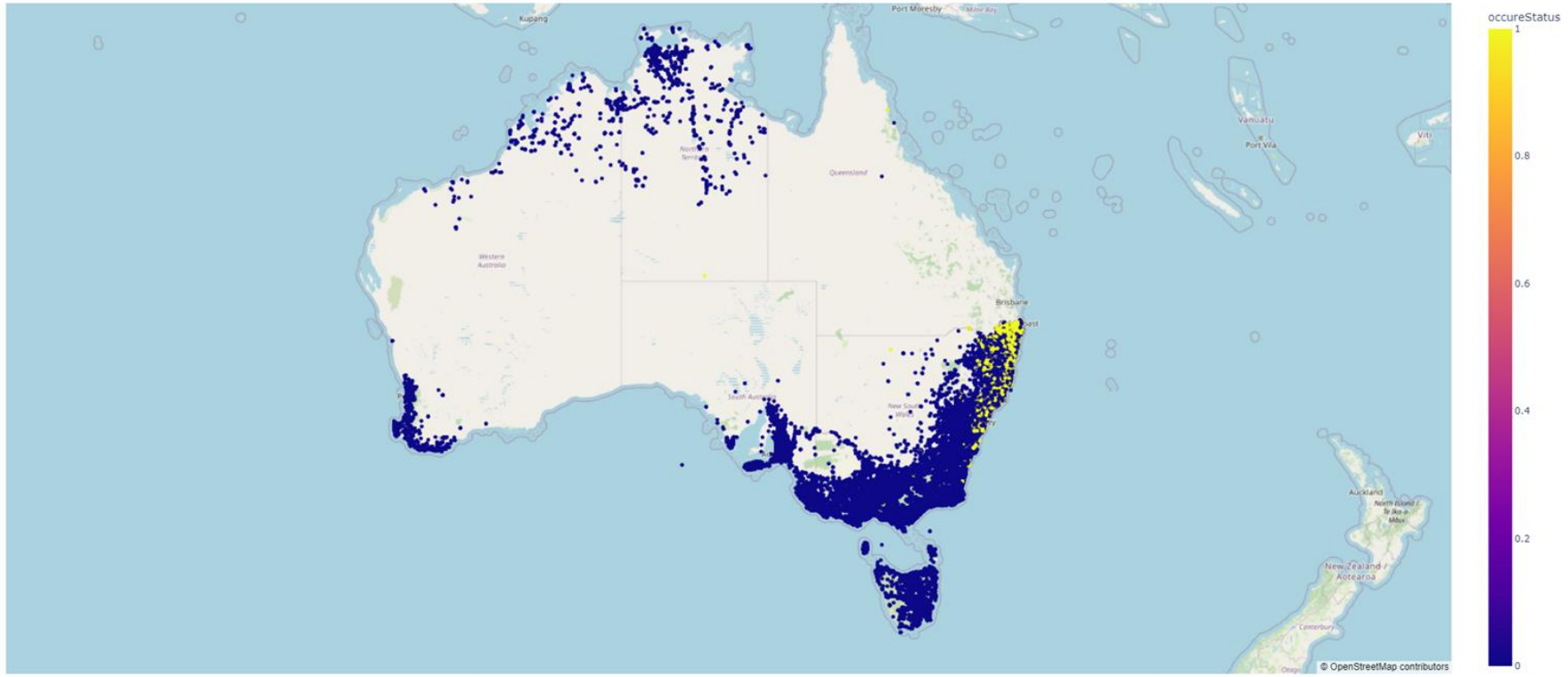
Queensland Data Removing

We have modified the total training set, and the new training set no longer includes Litoria Fallax observations from Queensland.



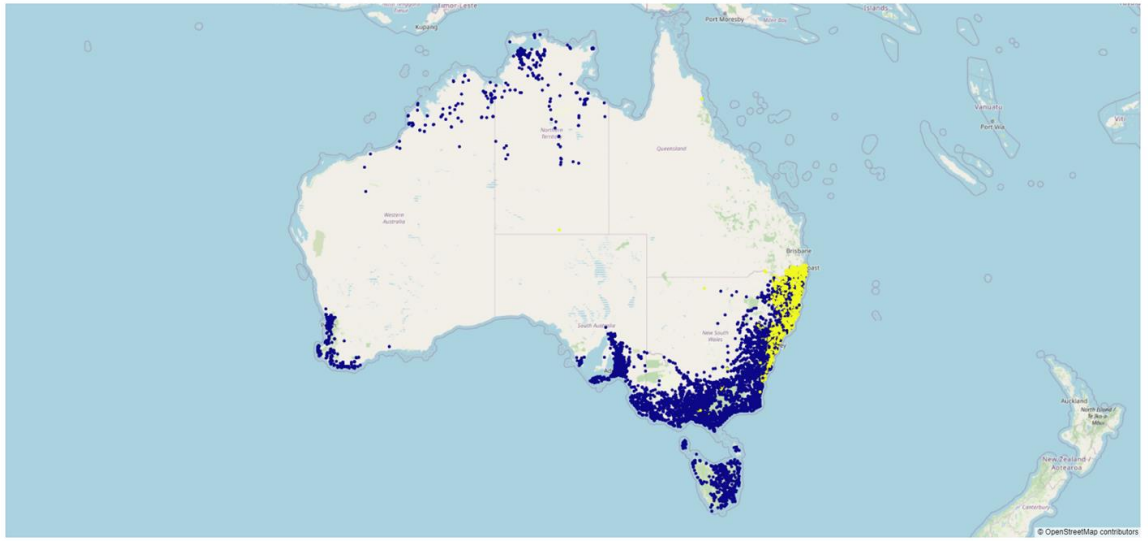
Increasing Sample Imbalance

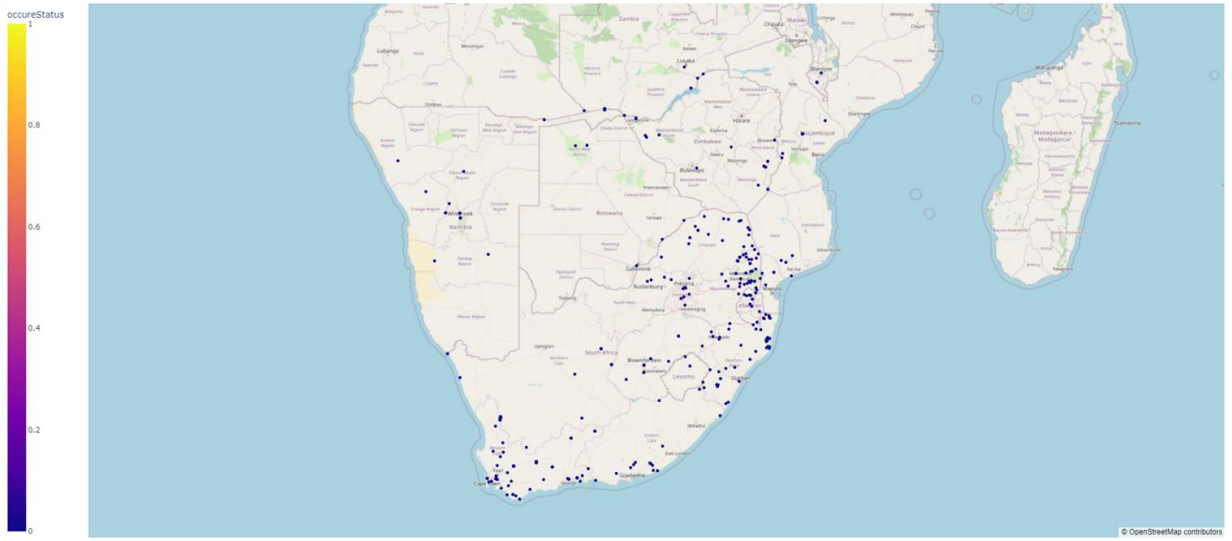
After removing the Litoria Fallax observations from Queensland, the percentage of Litoria Fallax observations in the total training set decreases by a large percentage. There are only 36,258 Litoria Fallax observations (20.01%) in the training set of 180,397, leaving 144,139 non-Litoria Fallax observations.

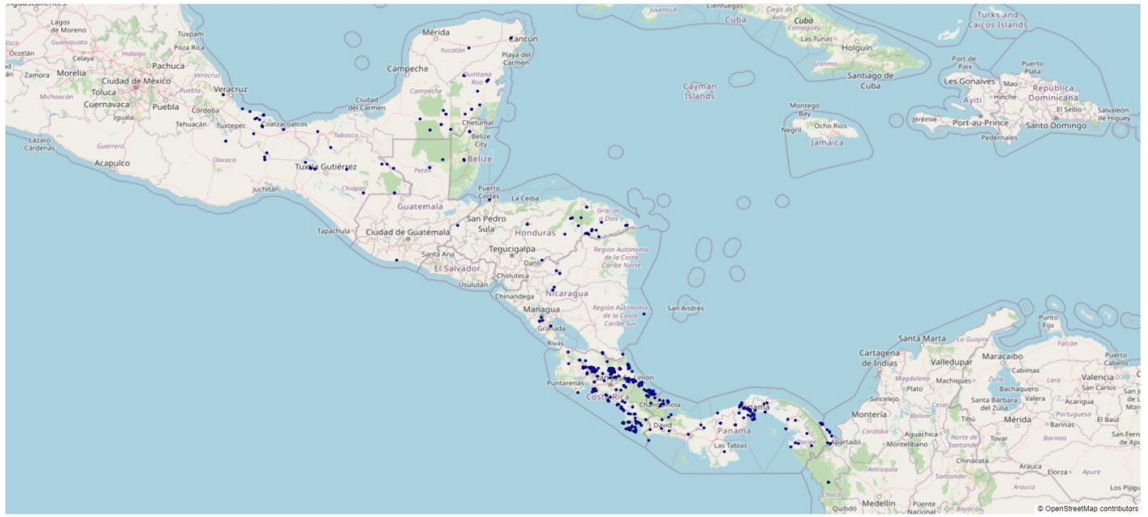


Given the uneven distribution of non-Litoria Fallax observations in Australia, we decided to bring in data from Costa Rica and South Africa to make our data more diverse.

Observation Data Selection







We have retained:
40% the Australian non-Litoria Fallax data;
50% non-Litoria Fallax data from Costa Rica;
50% non-Litoria Fallax data from South Africa.

The total processed training set had a total of 60,510 instances, of which 36,258 were Litoria Fallax observations (59.9%).

Predictor Variables

TerraClimate Data

Data Source

TerraClimate

- a dataset of monthly climate and climatic water balance for global terrestrial surfaces from 1958-2019.

Limits

Time Scale

- '2015-01-01' - '2019-12-31'

Localization

- Longitude, Latitude
- Australia, South Africa, Central America(Costa Rica)

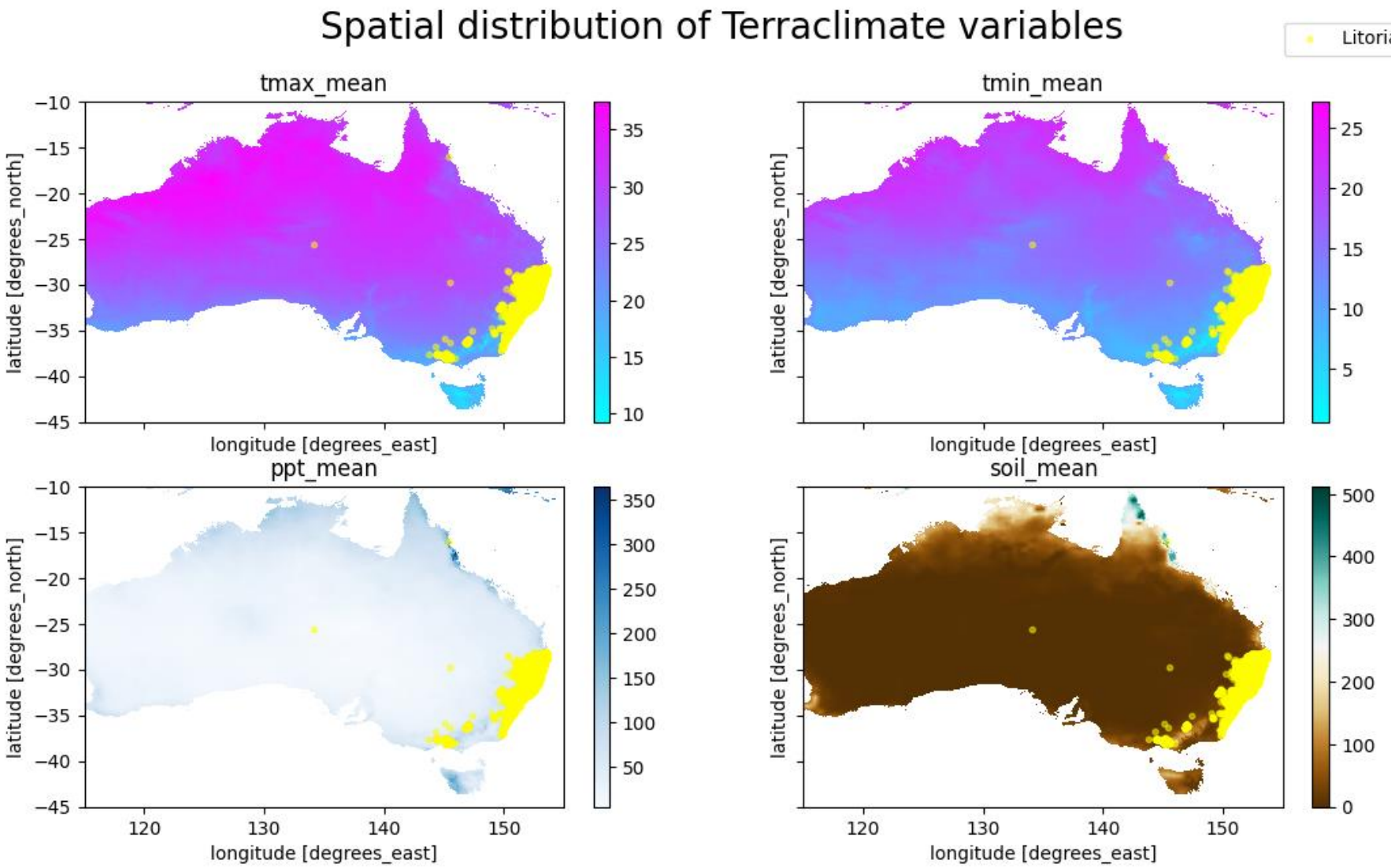
Variables

- maximum air temperature (tmax)
- minimum air temperature (tmin)
- accumulated precipitation (ppt)
- soil moisture (soil)

Preprocessing

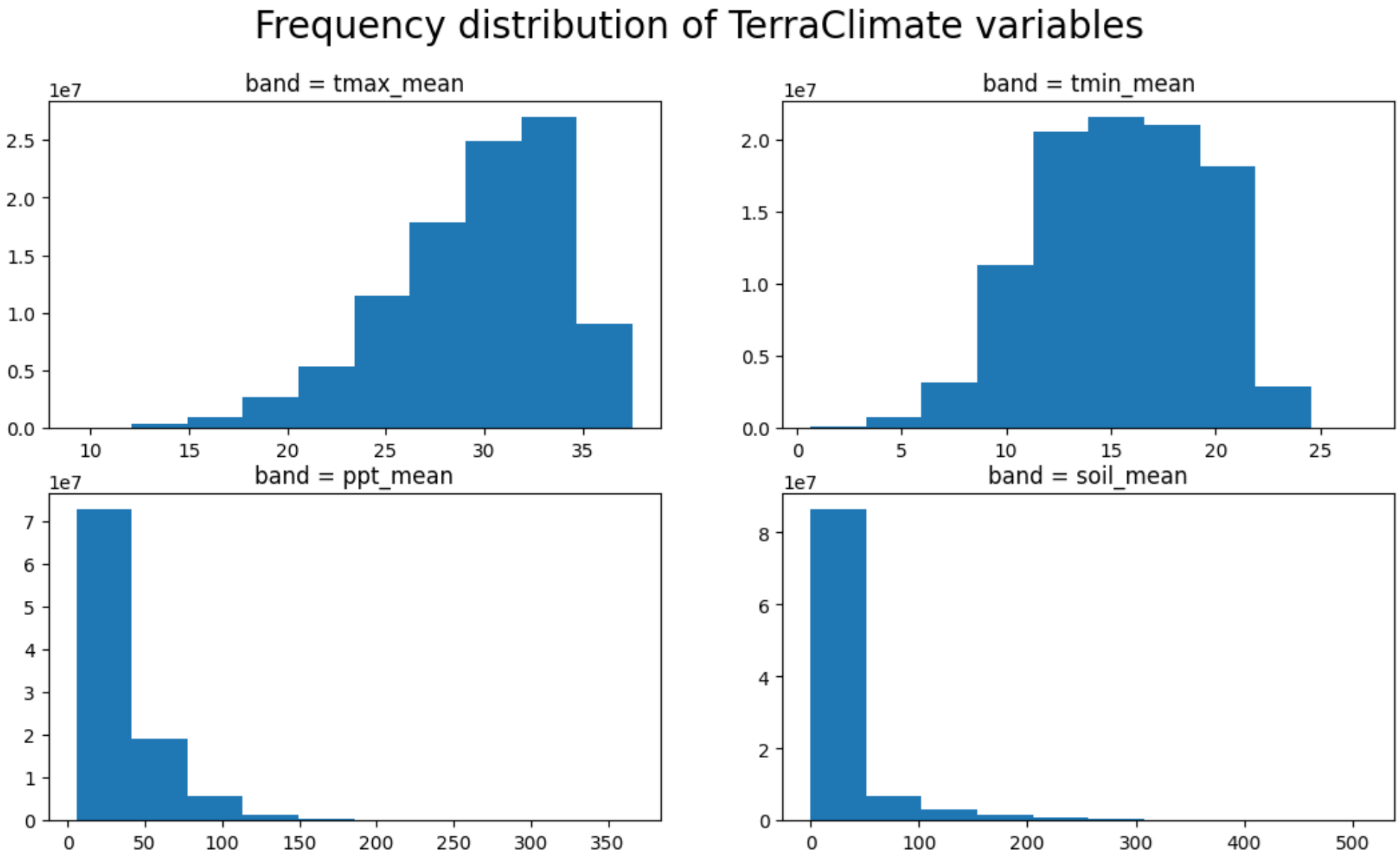
Replace all missing value with ‘0’

Data Visualization



- Climate Data Vs Frog Distribution
- Warmer, More Humid
- Useful for Classification

Data Visualization



- Asymmetrical Distribution
- Skewness

What We Get

01 The distribution of target frogs is linked to climate.

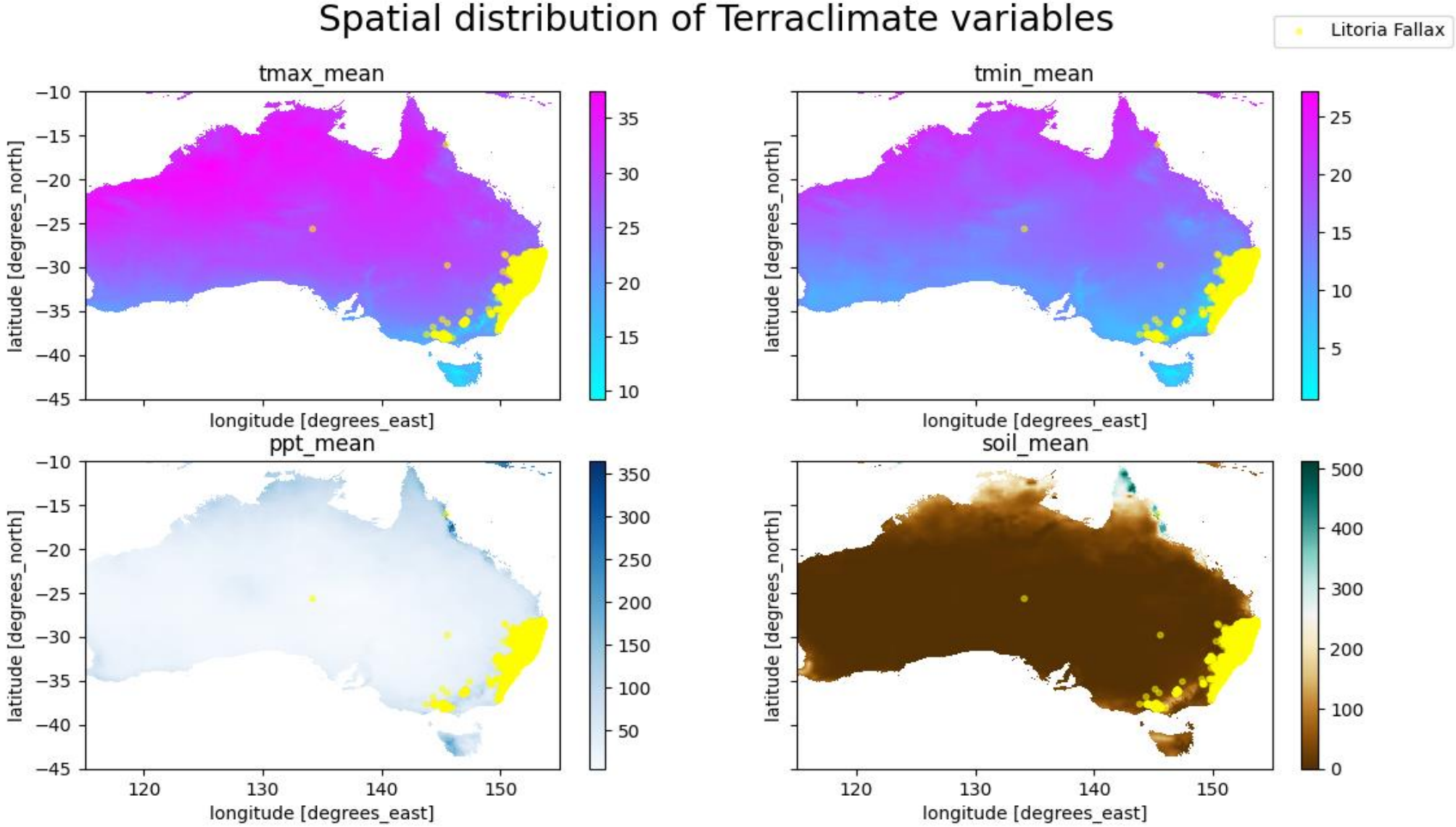
02 Further processing of the data is required, for example: standardisation, normalisation.

03 Next step: joining pretictors to the response variable.

Feature Engineering

PART 3

Feature Engineering



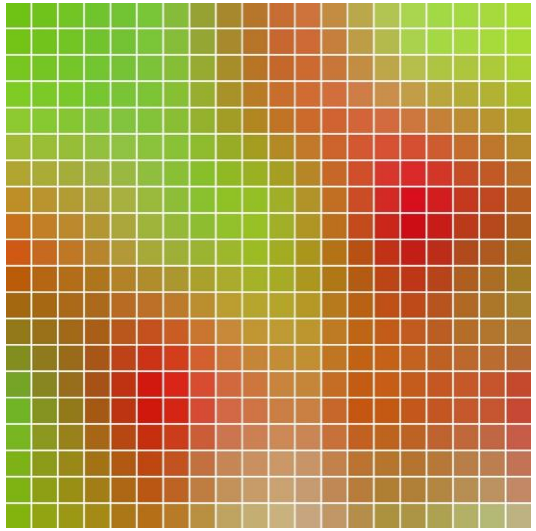
Feature Engineering

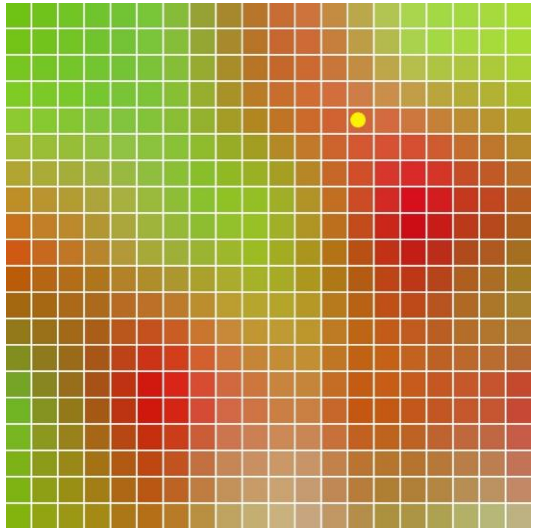
Divide the image into small boxes

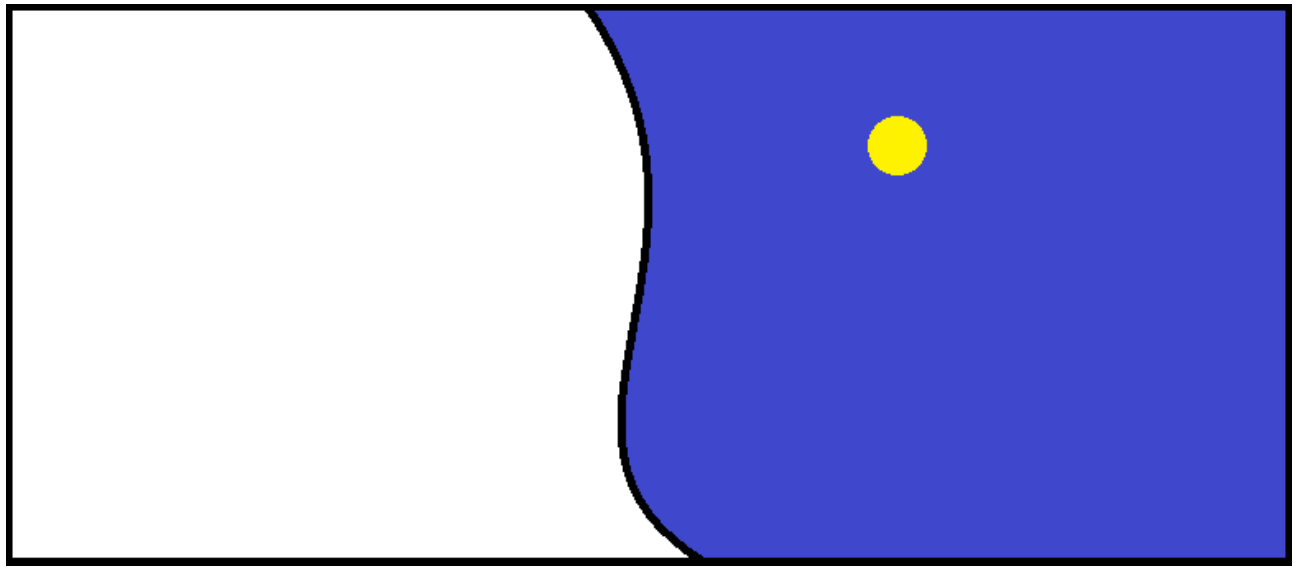
Get the mean of each box

An observation point land in a box

Use the measurement in that box for that observation point







Feature Engineering

	gbifID	eventDate	country	occureStatus	continent	stateProvince	decimalLatitude	decimalLongitude	species	key	ppt_mean	soil_mean	tmax_mean	tmin_mean
0	2626212822	2020-04-20 13:01:21	South Africa	0	Africa	Western Cape	-33.939515	23.445838	Xenopus Laevis	0	97.791667	75.511669	23.577166	13.778333
1	2429318559	2019-09-28 16:08:00	Malawi	0	Africa	Chikwawa	-16.232534	34.790058	Chiromantis Xerampelina	1	94.251667	88.056668	26.127166	15.201833
2	3456793261	2022-01-09 21:55:59	South Africa	0	Africa	Western Cape	-33.727559	21.163907	Xenopus Laevis	2	85.383333	77.081668	23.582333	13.792000
3	3468999853	2022-02-09 22:16:45	South Africa	0	Africa	KwaZulu-Natal	-27.932412	32.344295	Chiromantis Xerampelina	3	91.515000	95.556668	23.357666	10.982000
4	3117859060	2021-05-09 09:35:00	South Africa	0	Africa	Free State	-28.597582	26.429381	Xenopus Laevis	4	84.020000	73.623335	23.894999	13.936666

gbifID and **key** does not have any meaning for the prediction we are going to make

species and **occurStatus** are redundant

	occureStatus	decimalLatitude	decimalLongitude	ppt_mean	soil_mean	tmax_mean	tmin_mean
0	1	-32.719457	152.159267	97.280000	119.958335	23.788999	13.963000
3	0	-25.077627	32.065052	14.245000	0.100000	32.835333	17.347333
4	1	-33.693144	151.320884	97.791667	75.511669	23.577166	13.778333
7	1	-33.925746	151.164082	85.383333	77.081668	23.582333	13.792000
8	0	-26.102859	27.829833	14.471667	0.100000	31.494833	16.386833

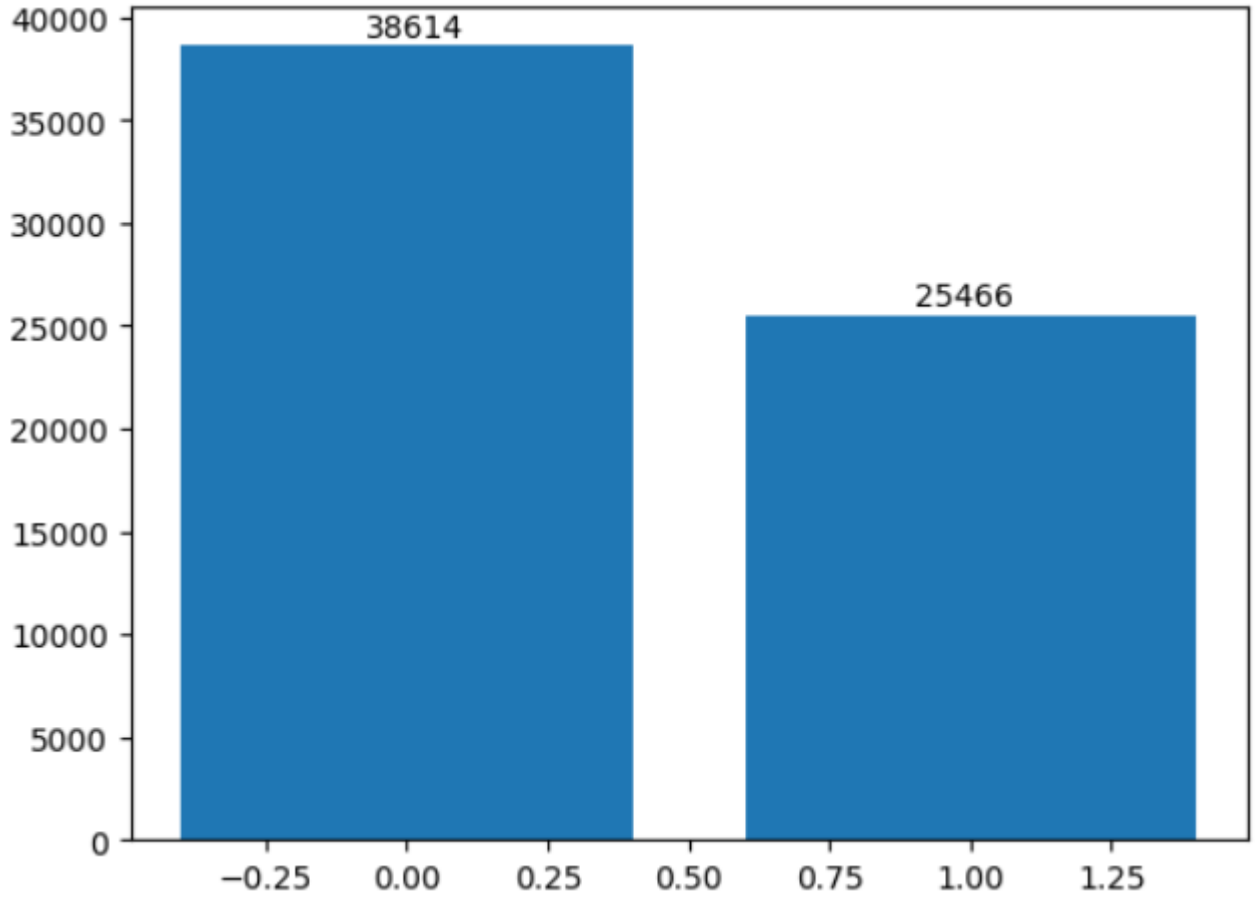
country, **continent** and **stateProvince** conveys the same but less accurate information with longitude and latitude

Model Performance

PART 4

Overall Information

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 91543 entries, 0 to 91604
Data columns (total 7 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   occureStatus      91543 non-null  int64  
 1   decimalLatitude   91543 non-null  float64
 2   decimalLongitude  91543 non-null  float64
 3   ppt_mean          91543 non-null  float64
 4   soil_mean         91543 non-null  float64
 5   tmax_mean         91543 non-null  float64
 6   tmin_mean         91543 non-null  float64
dtypes: float64(6), int64(1)
memory usage: 5.6 MB
```



Process

01

Train-validation split

02

Oversampling to handle the imbalance

03

Standard Normalization

04

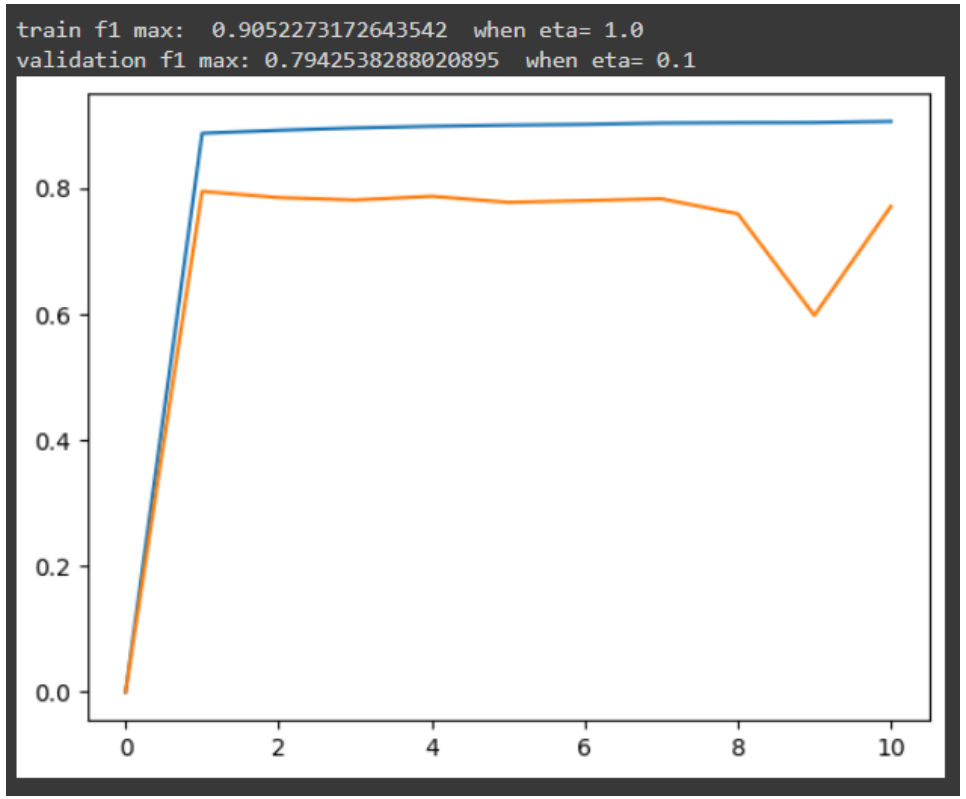
Building Model

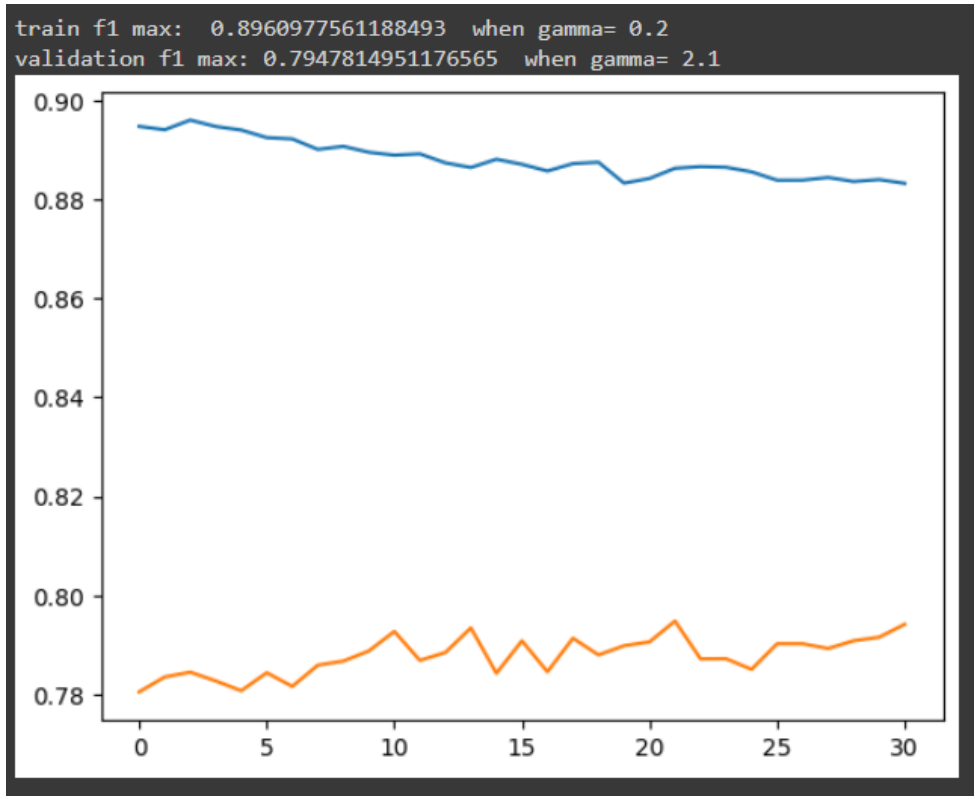
Model

XGBoost.

Hyperparameters:

eta,gamma,max_depth,min_child_weight,subsample,reg_lambda,reg_alpha





Model Performance

After GridSearch:

eta=0.3
gamma=1
max_depth=1
min_child_weight=0.2
subsample=0.81
reg_lambda=1.5
reg_alpha=2.1

F1 score:

Train: 0.852
Validation: 0.805
Test: 0.70

THANK YOU