# AUTOMATED PLAUSIBILITY CHECK FOR BIRD OBSERVATIONS

HAW HAMBURG · MARS GROUP Multi-Agent Research & Simulation · DDA · VOGELWARTE.CH

Marina Siebold
marina.siebold@haw-hamburg.de

## MOTIVATION & PROBLEM STATEMENT

The growing reliance on citizen science datasets for ecological studies, such as bird sightings, demands robust quality assurance to ensure the integrity of research outcomes. Central to this process is the evaluation of the plausibility of each record based on historical patterns.
We present statistical and Machine Learning Outlier Detection methods to identify implausible records within bird sightings provided by DDA and Vogelwarte Switzerland. By flagging these outliers, our methods aim to streamline the manual review process for reviewers, facilitating the identification of potential errors or inaccuracies in the data.

## RELATED WORK

- A meta-analysis from Baker et al. found that most ecological citizen science projects rely on manual expert verification exclusively [1].
- Only a minority applies algorithmic approaches, with even less relying on Machine Learning methods [2].
- Growing data volume highlights the necessity of automated approaches to minimize required expert verification [1].
- *eBird*, a similar initiative, showcased reliable outcomes using *Emergent Filters* [3], user experience evaluation, and expert review [4].

## ABOUT THE DATA

The dataset utilized encompasses bird sightings gathered by volunteers in Switzerland and Germany from 2018 to 2022. Each record includes information on the observed species, date, location (transformed to uniform 50x50km grids), and post-generated land cover and altitude; visualized in **Figure 1**.
The record distribution, as shown in **Figure 2**, indicate a comprehensive coverage of Switzerland and Germany.

**Species:** Robin
**Date:** 10th July '23
**Location:** 50kmE4350N3250
**Altitude:** 50m
**Land Cover**: Agriculture

**Figure 1.** Features of a sample record. Pictures are optional and not used in this project.

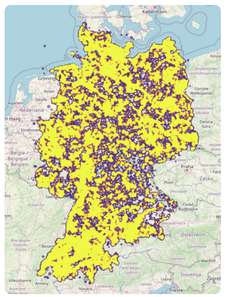

**Figure 2.** Distribution of recorded data points 2018-2022.

## SYSTEM EXTENSION

Our models will be integrated into the current validation process. They will precede the manual review to reduce the review amount and present anomalous data points sorted by plausibility.

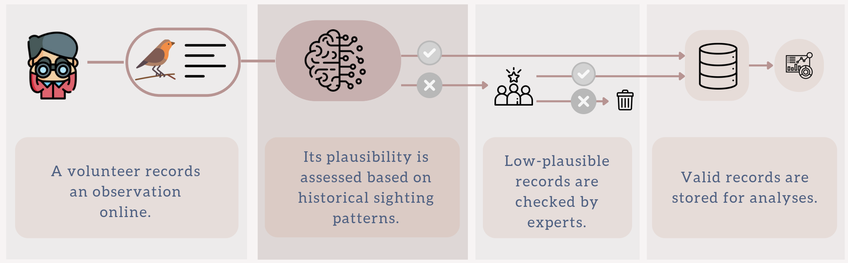


**Figure 3.** Extended review process at DDA and Vogelwarte Switzerland.

## METHODS

### 1 EMERGENT FILTERS

Based on the preceding year's records, the occurrence probability of a bird species can be computed for each day of the following year and each spatial 50x50km grid. Records are flagged for review if their occurrence probability falls below a predefined threshold (see **Figure 4**). While this approach solely considers date and location, it proved to be effective in other bird observation projects [3].

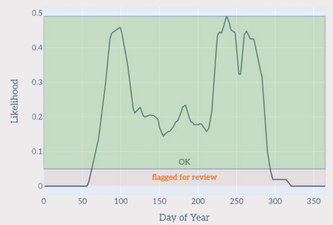


**Figure 4.** Emergent Filter plausibilities of garganays near Hannover 2023.

### 2 OUTLIER DETECTION

Three Outlier Detection methods, namely Isolation Forest, DBSCAN, and AutoEncoder, were employed to flag atypical bird sightings based on historic sighting patterns.
Recognizing the diversity of behaviours and preferences, one model is developed for each species, ensuring accurate conclusions about bird sightings plausibilities.
Shapley-analyses can assist the reviewers in pinpointing the most implausible feature (see **Figure 5**).

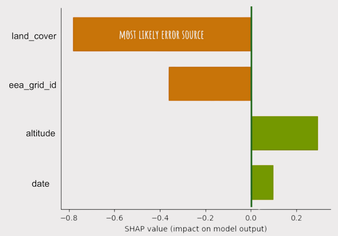


**Figure 5.** DBSCAN Shap values of a water pipit observed above Lago Maggiore.

## RESULTS

To assess the models' capabilities to detect implausible sightings, 2023 data was enriched with manually falsified records. We assessed:

- Sensitivity: Percentage of implausible records that the model identified
- Precision: Percentage of implausible records in the review pile

**Emergent Filters and DBSCAN were the most effective at identifying implausible records, with DBSCAN resulting in the fewest plausible records in the review pile.**

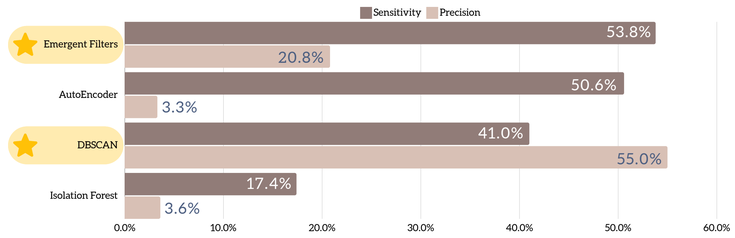


**Figure 6.** Evaluation results.

## FUTURE DIRECTIONS

1. **Hyperparameter optimization.** Aims to improve Sensitivity and Precision while considering review capacity.
2. **Feature value analysis.** An optimal combination of informative features and an ideal spatial representation will be identified to improve Outlier Detection.
3. **Concept Drift accommodation.** Concept drifts, involving shifts in population, migration, or habitats, challenge the models' accuracies. Drifts will be identified and strategies will be revised to adapt the models accordingly.

## REFERENCES

[1] Baker, Emily, et al. "The verification of ecological citizen science data: current approaches and future possibilities." Citizen Science: Theory and Practice 6.1 (2021).
[2] Wiggins, Andrea, et al. "Mechanisms for data quality and validation in citizen science." 2011 IEEE seventh international conference on e-Science Workshops. IEEE, 2011.
[3] Kelling, Steve, et al. "Emergent filters: automated data verification in a large-scale citizen science project." 2011 IEEE Seventh International Conference on e-Science Workshops. IEEE, 2011.
[4] Yu, Jun, et al. "Automated data verification in a large-scale citizen science project: a case study." 2012 IEEE 8th International Conference on E-Science. IEEE, 2012.