HAW HAMBURG

Masterarbeit

Simon Birker

AI-driven validation of citizen science data: Anomaly Detection of Bird Sightings with Machine Learning Models and Statistical Approaches

Faculty of Engineering and Computer Science
Department Computer Science

Simon Birker

AI-driven validation of citizen science data: Anomaly Detection of Bird Sightings with Machine Learning Models and Statistical Approaches

Master thesis submitted for examination in Master´s degree
in the study course *Master of Science Informatik*
at the Department Computer Science
at the Faculty of Engineering and Computer Science
at University of Applied Science Hamburg

Supervisor: Prof. Dr. Thomas Clemen
Supervisor: Prof. Dr. Marina Tropmann-Frick

Submitted on: 19. Juni 2025

Simon Birker

Thema der Arbeit

AI-driven validation of citizen science data: Anomaly Detection of Bird Sightings with Machine Learning Models and Statistical Approaches

Stichworte

Maschinelles Lernen, Unüberwachtes Lernen, Emergente Filter, Ornithologie, Outlier-Erkennung

Kurzzusammenfassung Citizen-Science-Plattformen sind heute ein zentraler Bestandteil des Biodiversitätsmonitorings. Sie ermöglichen es Freiwilligen, Artenbeobachtungen mit bislang unerreichter räumlicher und zeitlicher Abdeckung zu erfassen. Die Verlässlichkeit dieser Daten hängt jedoch entscheidend von robusten Validierungsmechanismen ab. Da die Datenmengen kontinuierlich steigen, stößt die manuelle Expertenüberprüfung - die derzeit gängige Praxis - zunehmend an ihre Grenzen.

Diese Arbeit untersucht zwei automatisierte Ansätze zur Unterstützung der Datenvalidierung im Kontext der europäischen *Ornitho*-Plattform: (1) statistisch definierte *Emergent Filters* und (2) unüberwachte *Outlier Detection Models.* Die Filter erweitern klassische Plausibilitätsprüfungen um zeitliche, räumliche und habitatbezogene Dimensionen. Die eingesetzten Machine-Learning-Modelle-erstmals in diesem Kontext angewendet-lernen artspezifische Verteilungsmuster und erkennen Auffälligkeiten anhand ihrer Abweichung von typischen Merkmalskombinationen.

Beide Ansätze wurden an einem kuratierten Benchmark-Datensatz mit künstlich manipulierten Beobachtungen für 27 Arten evaluiert. Die quantitative Analyse zeigt, dass die unüberwachten Modelle die Filter in Bezug auf die F1-Scores klar übertreffen, insbesondere bei komplexeren Fehlertypen. Qualitatives Feedback von erfahrenen Ornithologen bestätigt zudem die Anwendbarkeit, Verständlichkeit und Nützlichkeit des Gesamtsystems. Die Ergebnisse zeigen außerdem artspezifische und merkmalsabhängige Unterschiede in der Detektionsleistung.

Diese Arbeit liefert sowohl eine fallbezogene Bewertung der Integration automatisierter Validierung in den Ornitho-Workflow als auch allgemeine Hinweise für die Entwicklung vertrauenswürdiger Entscheidungsunterstützungssysteme im Bereich der ökologischen Datenvalidierung.

Simon Birker

Title of Thesis

AI-driven validation of citizen science data: Anomaly Detection of Bird Sightings with Machine Learning Models and Statistical Approaches

Keywords

Machine Learning, Unsupervised Learning, Emergent Filters, Ornithology, Outlier Detection

Abstract Citizen science platforms have become a important part of modern biodiversity monitoring by enabling volunteers to submit species sightings at unprecedented spatial and temporal scales. However, the reliability of such data critically depends on robust validation mechanisms. Manual expert verification - currently the norm in most platforms - is increasingly reaching its limits as data volumes continue to grow.

This thesis investigates two automated approaches to support the validation of bird sighting data in the context of the European *Ornitho* platform: (1) statistically defined *Emergent Filters* and (2) unsupervised *Outlier Detection Models*. The Emergent Filters extend conventional ecological plausibility checks by incorporating temporal, spatial, and habitat-related dimensions. The Machine Learning models-applied for the first time in this context-learn species-specific distributions and identify implausible records based on their deviation from expected feature patterns.

Both methods were evaluated on a curated benchmark dataset containing artificially manipulated sightings for 27 species. Quantitative evaluation shows that unsupervised models consistently outperform statistical filters in terms of F1-score, especially for complex error types. Qualitative feedback from expert ornithologists confirms the practical relevance, interpretability, and perceived usefulness of the combined system. The results also highlight species-specific and feature-specific sensitivities that influence detection performance.

This work provides both a case study for the integration of automated validation into the Ornitho workflow and broader insights into designing trustworthy decision-support systems for ecological data validation.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Background

In recent years, citizen science initiatives have emerged as an increasingly important resource for large-scale ecological monitoring. By leveraging digital platforms and mobile applications, volunteers across the globe contribute millions of species observations, enabling datasets that far exceed the spatial and temporal resolution achievable by traditional field surveys. This development has notably transformed biodiversity monitoring, particularly for taxa such as birds, where citizen-contributed sightings have become a foundational component of national and international databases.

1.2 Problem Statement

Despite their broad utility, the quality and reliability of such datasets remain a central concern. The accuracy of species identification, correct geolocation, and plausibility of ecological context must be critically assessed before these data can be used in scientific analysis or policy decisions. Most citizen science platforms currently rely on manual expert review to ensure data quality (Baker et al. 2021). While this approach benefits from domain-specific ecological knowledge, it is increasingly reaching its limits due to the exponential growth in data volume. As highlighted by Baker et al. (2021), the scalability of expert-based validation is limited, creating a pressing need for complementary automated support mechanisms.

1.3 Research Gap

To address this challenge, automated validation methods offer a promising pathway for reducing the burden on human reviewers and improving the efficiency and consistency of quality assurance. However, only a minority of citizen science projects currently employ automated validation techniques, and most rely on basic statistical rules rather than data-driven or adaptive methods (Baker et al. 2021). The literature identifies two critical gaps: first, a lack of sophisticated, scalable detection algorithms; and second, a limited understanding of how domain experts perceive and interact with such automated tools in practical validation workflows.

1.4 Objectives and Approach

This thesis aims to close these gaps through the design, implementation, and evaluation of two complementary automated validation approaches in the context of the Ornitho platform - a collaborative citizen science network used in Germany, Switzerland, and other European countries. The focus lies on Outlier Detection in bird observation records, with the objective of flagging implausible entries prior to expert review.

The first approach builds on species-specific ecological knowledge by applying statistical *Emergent Filters*. These filters extend previous filtering strategies (e.g., those used in the eBird platform) by incorporating multiple environmental and spatial features, such as altitude, land cover, and seasonal timing, to detect deviations from expected ecological patterns. The second approach leverages unsupervised *Outlier Detection Models*, a novel addition to this domain, to identify records that deviate from typical sighting clusters in a multidimensional feature space.

To comprehensively assess these approaches, both quantitative and qualitative evaluation perspectives are employed. The quantitative analysis benchmarks performance across species and error types using precision, recall, and F1-score metrics. The qualitative component involves expert review of model predictions to assess ecological plausibility, interpretability, and perceived utility.

1.5 Research Questions

The thesis addresses the following research questions:

- **RQ1 – Comparative Model Performance:** To what extent do statistical (Emergent Filter) models and Machine Learning models differ in their accuracy for detecting user-generated errors, as measured by the F1-score? *Motivation:* Understanding the relative strengths of both approaches is essential for deciding which method should be prioritized or combined in practical validation workflows.
- **RQ2 – Error-type–specific Performance:** How does the detection performance of the investigated model approaches vary across the predefined error types (e.g., date errors, land-cover errors, altitude errors)? *Motivation:* Quantifying the detectability of each error type for both approaches helps reveal which anomalies are harder to detect and which methods are best suited for specific error types.
- **RQ3 – Species- or Guild-specific Performance:** To what degree does model performance differ among avian species or ecological guilds? *Motivation:* Quantifying detectability across species or guilds for both approaches reveals systematic performance differences and may indicate the need for method-specific adaptations.
- **RQ4 – Influence of Features:** Which features contribute significantly to the detection performance of the Machine Learning model? *Motivation:* Determining the most relevant features helps prioritize ecological variables that offer the greatest predictive value for future deployments.
- **RQ5 – Practitioner Preferences:** How do professional ornithologists evaluate the usability and reliability of statistical Emergent Filters compared with Machine Learning models for error detection? *Motivation:* Gaining expert feedback is crucial to ensure that automated tools are trusted, accepted, and actually useful in real-world review processes.

1.6 Contribution and Relevance

The results of this thesis are intended to serve two main purposes. First, they offer a case-specific assessment of the applicability and performance of automated Anomaly Detection in the Ornitho use case. Second, they contribute to the development of broader guidelines for integrating automated validation into citizen science workflows. These findings aim to inform researchers and practitioners on how to design, implement, and evaluate automated systems that are not only technically robust, but also trusted and usable by domain experts.

1.7 Project Context

The Ornitho platform, developed and operated by organizations such as the Federation of German Avifaunists (DDA) and the Swiss Ornithological Institute, has amassed over 90 million bird records in Germany and nearly 28 million in Switzerland. With over 330 million entries across 12 participating countries, it represents one of the most comprehensive repositories of avian observations in Europe. Given the accelerating data inflow, expert-only validation strategies are increasingly strained, prompting an urgent need for scalable support mechanisms.

This thesis has been conducted as a collaborative effort between the University of Applied Sciences Hamburg (HAW), the DDA, and the Swiss Ornithological Institute. It combines methodological expertise in data science with domain knowledge from professional ornithologists to develop, test, and evaluate practical solutions for real-world validation challenges in large-scale ecological monitoring.

2 Background

2.1 Citizen Science

Citizen science refers to the engagement of the public in scientific research activities, often by contributing data, observations, or analyses (Zhu & Newman 2024). The rise of digital platforms such as *ornitho.de*, *ornitho.ch*, *eBird*, and *Naturalist* has greatly expanded the scale and scope of citizen science initiatives in ecology and biodiversity monitoring (Johnston et al. 2023). These platforms enable volunteers to record wildlife observations via mobile or web-based interfaces, generating vast datasets that can address questions related to species distributions, phenology, and population trends. Citizen science efforts are also linked to help monitoring the sustainable development goals (SDGs) (De Sherbinin et al. 2021).

Volunteers in ecological citizen science commonly collect presence-only data by reporting sightings without noting absences-a straightforward approach that is employed by platforms like *eBird* and *Ornitho*. However, this data inherently contain errors due to observer variability (Johnston et al. 2023, Rempel et al. 2019), species misidentification, and spatial or temporal inaccuracies (Johnston et al. 2020, La Sorte & Somveille 2020). Collectively, these factors can significantly affect outcomes of ecological studies (Backstrom et al. 2025), requiring careful correction to avoid misleading inferences (Di Febbraro et al. 2023). In contrast, presence–absence data, collected via structured or semi-structured protocols, capture both detections and confirmed non-detections, enabling stronger statistical inference over time, although requiring rigorous survey design to distinguish true absences from non-detections (Parris et al. 2023).

2.2 Validation of Ecological Citizen Science Data

Reliable data are essential for monitoring ecological trends, informing conservation strategies, and guiding environmental management. However, the open-access nature of citizen science can lead to varying data quality. Therefore, ensuring the reliability of volunteer-collected data gathered through citizen science is a significant challenge, especially when these contributions by non-professionals are intended to support scientific research and policy decisions.

Without robust verification procedures, mistakes - whether due to misidentification, reporting errors, or other issues - can accumulate and potentially compromise the outcomes of analyses and decisions. Consequently, developing effective data validation methods is essential .

2.2.1 Approaches to Data Validation

Citizen science projects vary widely in their strategies for validating submissions, and many offer no explicit description of their verification processes (Baker et al. 2021, Cavadino et al. 2024). In a comprehensive review, Baker et al. (2021) identified four primary categories of validation approaches:

1. **Expert-Based Review:** In this method, trained experts or experienced scientists manually evaluate data submissions. Experts bring domain-specific knowledge to the task, allowing them to spot inconsistencies, errors, or outlier observations that automated systems might miss. Although time-consuming, expert reviews remain the gold standard for ensuring data reliability (Figuerola-Ferrando et al. 2024).
2. **Community Consensus:** This approach leverages the collective wisdom of the citizen science community. Through mechanisms such as peer review, rating systems, or discussion forums, community members can collectively validate observations. Community consensus often helps to democratize the validation process and can be particularly effective when a large number of participants are involved (Bourgeois et al. 2024).
3. **Automated Approaches:** Automated methods use computer algorithms, including Machine Learning techniques, to verify data submissions. These approaches are especially valuable when dealing with large datasets, as they can rapidly process

and flag potentially erroneous entries (Kessel et al. 2025). Despite their potential, Baker et al. (2021) found that automated approaches are rarely implemented in current projects, with documented use cases largely limited to projects with over 1,000 participants.

4. **Hybrid Systems:** Hybrid validation systems combine two or more of the above methods. For example, a project might initially screen submissions using automated algorithms and then have experts review the flagged entries. By integrating multiple validation layers, hybrid systems aim to balance scalability with the nuanced judgment that experts and community members provide.

Figure 2.1 illustrates the distribution of these verification strategies among the reviewed citizen science projects. The figure clearly shows that, irrespective of the project's size, expert-based review is dominantly used, followed by community consensus. Automated approaches are notably rare and are only integrated into larger projects. Specifically, in 256 investigated citizen science projects, only seven use cases incorporated an automated approach (Baker et al. 2021).

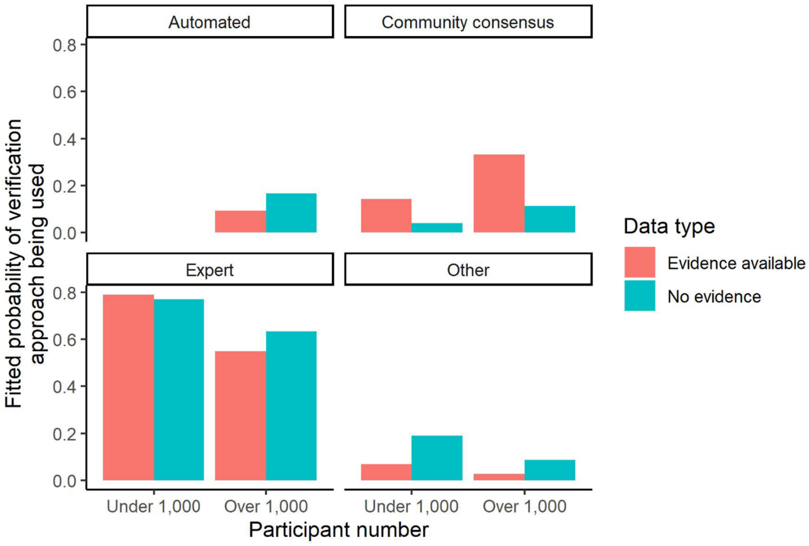


Figure 2.1: Overview of data verification strategies in ecological citizen science, adapted from Baker et al. (2021). Approaches range from manual expert review to fully automated methods.

This current underutilization of automated methods is not indicative of their unimportance. With the exponential growth of citizen science data, the scalability of traditional expert-based validation becomes a significant bottleneck (Baker et al. 2021). Automated approaches are likely to play an increasingly critical role by handling large volumes of data quickly and efficiently.

Furthermore, the integration of automated systems within hybrid frameworks can offer a balanced approach. For instance, initial automated screenings can filter out clearly erroneous data, while expert reviews can focus on borderline cases that require deeper analysis. This synergy promises to enhance the overall reliability and usability of citizen science data.

2.2.2 Framework for verifying ecological citizen science data

Recognizing this challenge, Baker et al. (2021) propose a structured framework for verifying ecological citizen science data that aims to balance efficiency and accuracy (see Figure 2.2). This framework is specifically designed to accommodate the increasing volume of data generated by citizen science initiatives while ensuring the reliability of the recorded observations.

The framework is organized hierarchically, with records undergoing different levels of verification based on the available information and the associated confidence in their accuracy. It comprises three key stages:

The first stage emphasizes the necessity of comprehensive data collection. Observers are encouraged to submit the maximum available evidence - such as photographs, audio recordings, and contextual metadata - alongside essential attributes such as date, location, and species identification. This approach enhances the verification process by leveraging a broader range of information. However, there is an inherent trade-off: increasing data requirements may discourage participation, as volunteers may be deterred by overly complex submission procedures (Baker et al. 2021).

To further improve data quality, each observation can be enriched with additional derived features. For example, species-specific characteristics can be considered: if a species is difficult to identify, frequently misclassified, or rarely recorded, the likelihood of misidentification increases. Similarly, sightings that occur outside the species' expected ecological context or are reported by an inexperienced observer with a history of errors may warrant greater scrutiny.

Following data collection and enrichment, records undergo a two-tiered verification process. The initial step involves automated verification, where algorithms assess the plausibility of records by integrating primary observation data with secondary metadata, such as historical occurrence records, species distribution patterns, temporal trends, and the observer's prior accuracy. These automated methods facilitate the efficient processing of large datasets and continuously refine verification criteria as new patterns and anomalies emerge. In cases where algorithmic verification yields inconclusive results, a community-based validation step is introduced, leveraging the collective expertise of a network of experienced contributors.

If neither automated nor community-based verification provides a sufficient level of certainty, records are escalated to expert review. This stage is essential for validating observations that deviate from established distributional, phenological, or ecological expectations, including rare, invasive, or otherwise anomalous sightings. Expert evaluation not only serves as the final validation step but also informs improvements to preceding verification stages by refining automated filters and community-based decision-making criteria.

Overall, this hierarchical framework optimizes data integrity while reducing the burden on expert reviewers, thereby establishing an efficient, scalable, and scientifically rigorous approach to verifying citizen science data.

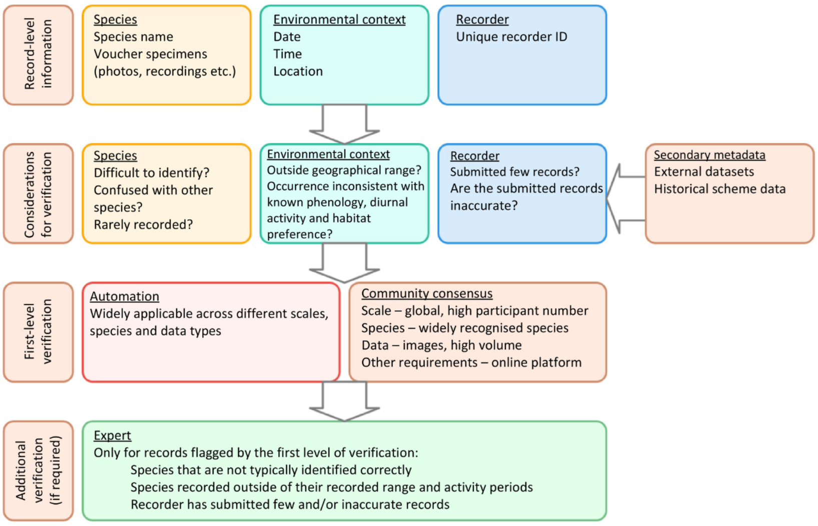

Record-level information
Species
Species name
Voucher specimens (photos, recordings etc.)
Environmental context
Date
Time
Location
Recorder
Unique recorder ID
Considerations for verification
Species
Difficult to identify?
Confused with other species?
Rarely recorded?
Environmental context
Outside geographical range?
Occurrence inconsistent with known phenology, diurnal activity and habitat preference?
Recorder
Submitted few records?
Are the submitted records inaccurate?
Secondary metadata
External datasets
Historical scheme data
First-level verification
Automation
Widely applicable across different scales, species and data types
Community consensus
Scale – global, high participant number
Species – widely recognised species
Data – images, high volume
Other requirements – online platform
Additional verification (if required)
Expert
Only for records flagged by the first level of verification:
Species that are not typically identified correctly
Species recorded outside of their recorded range and activity periods
Recorder has submitted few and/or inaccurate records

Figure 2.2: An idealized data verification pipeline for ecological citizen science, adapted from Baker et al. (2021). Automated Anomaly Detection, community feedback, and expert validation are combined in a feedback loop.

2.2.3 eBird's Emergent Filters

A notable automated approach used in the large-scale project *eBird* is the *Emergent Filters* method, initially introduced in Kelling et al. (2011) and later refined in Kelling et al. (2019). Emergent Filters integrate historical occurrence patterns and observer expertise to flag questionable records dynamically. This strategy balances two competing goals: retaining genuinely rare (but valid) sightings and weeding out clearly implausible submissions.

Conceptual Overview

Emergent Filters operate under the principle that historical patterns of species occurrence can inform the plausibility of new observations. If an incoming sighting deviates substantially from what has historically been reported for a given location, date, and species, it is flagged as potentially erroneous. Importantly, these thresholds for what constitutes an "outlier" are not static; they *emerge* and evolve over time as more data are validated, leading to progressively refined filter criteria.

Spatiotemporal Bin Definitions

A core step in implementing Emergent Filters involves partitioning the data space into spatiotemporal bins. For each species s, the platform defines a set of geographic and temporal "cells" (or bins):

$$B_s = \{(g, t) \mid g \in G,\, t \in T\},$$

where G is the set of relevant geographic regions (e.g., grid cells or administrative boundaries) and T is a set of time periods. For this, Kelling et al. (2011) apply day-of-year intervals. Each bin corresponds to a species s observed in region g during time window t.

Historical Frequency and Thresholds

For each bin (g, t), a historical frequency or distribution of observations for species s is computed:

$$F_s(g, t) \;=\; \frac{\text{Number of verified observations of species } s \text{ in bin } (g, t)}{\text{Total verified observations in bin } (g, t)}.$$

To account for fluctuations in reporting effort, Kelling et al. (2011) additionally introduce a rolling maximum frequency using a centered sliding 7-day window. For a given day t, the sliding window spans the interval

$$i = [t - 3, t + 3]$$

The highest observed frequency within that window is then assigned to day t, producing a smoothed daily probability of occurrence.

If this estimated likelihood for a given day t falls below a defined plausibility threshold (e.g., 5%), all corresponding sightings that are submitted on this day in the following year are flagged as potentially implausible.

Observer Reliability Adjustment

Beyond historical patterns, Kelling et al. (2011) also incorporate a weighting or "observer reliability score." Each participant u can be assigned a trust metric based on factors such as:

$$R(u) = \beta_1 \cdot \text{Expert Level} + \beta_2 \cdot \text{Past Accuracy} + \cdots$$

Lower $R(u)$ values for observers who repeatedly submit implausible sightings trigger more frequent flags, whereas highly reliable contributors can exceed typical thresholds with fewer questions raised. In practice, Kelling et al. (2011) discuss incrementally updating these reliability scores based on expert validation outcomes.

Workflow Integration

The final stage of Emergent Filters involves routing flagged observations to a review queue, where domain experts or experienced volunteers determine whether to confirm or reject them. Confirmed outliers lead to model updates, refining $\mu_s(g,t)$ and $\sigma_s(g,t)$ for future submissions. Rejected sightings adjust the observer's reliability score and further tighten the thresholds. Overall, this iterative framework can handle large, continuous data streams and dynamically adapt to changing patterns of species occurrence.

2.3 Unsupervised Machine Learning

Many automated verification methods rely on historical frequency analysis tailored to species and location. Another significant class of techniques is *Unsupervised Learning*, specifically Anomaly Detection[1]. Anomaly Detection is a branch of Machine Learning and statistical analysis aimed at identifying data points that significantly deviate from the majority of observations (Chandola et al. 2009). In ecological citizen science data, an *anomaly* might represent a rare but valid sighting (e.g., a bird seen outside its typical range) or an erroneous record (e.g., incorrect data entry or misidentification).

Unsupervised methods serve as flexible Anomaly Detectors that do not require labeled data for training, making them particularly useful for ecological data where outliers are seldom explicitly labeled. However, they may flag genuine but rare observations as anomalies, necessitating expert review or additional contextual validation.

Below, an intuitive overview of seven key unsupervised Anomaly Detection models applicable for in ecological and similar contexts is provided:

2.3.1 DBSCAN

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) detects anomalies by clustering densely packed data points (Ester et al. 1996). Points that fall in sparse regions are flagged as outliers. DBSCAN uses two parameters: the radius (eps) that defines a neighborhood around points, and the minimum number of points (MinPts) needed to form a cluster. It is effective at finding clusters of arbitrary shapes and identifying isolated points as anomalies but may struggle with varying density distributions across the data.

[1] Also referred to as Outlier Detection or Novelty Detection in some literature.

2.3.2 HDBSCAN

Hierarchical DBSCAN (HDBSCAN) states to improve on DBSCAN by building a hierarchy of clusters and extracting stable clusters based on density (Campello et al. 2013). Unlike DBSCAN, HDBSCAN adapts to variable density levels across the data. Important parameters include minimum cluster size and minimum samples, which influence the algorithm's sensitivity to outliers. This adaptive capability makes HDBSCAN especially suitable for ecological data, where density distributions can vary significantly.

2.3.3 Isolation Forest

Isolation Forest detects anomalies by randomly partitioning data into binary trees (Liu et al. 2008). Anomalies are points that require fewer partitions to isolate, indicated by shorter path lengths in these trees (Yepmo et al. 2024). Parameters include the number of trees and subsample size, which affect computational efficiency and detection accuracy. Isolation Forest is particularly effective for large datasets and handles high-dimensional data robustly, making it a practical choice for ecological data.

2.3.4 Autoencoder

An autoencoder is a neural network designed for Unsupervised Learning, encoding data into a lower-dimensional latent representation and reconstructing it back into its original form. Anomalies are identified by measuring the reconstruction error-the difference between the original data and its reconstruction. Higher reconstruction errors typically indicate potential anomalies. Autoencoders are particularly powerful for detecting complex, high-dimensional anomalies but require careful selection of architecture and parameters (Zhou & Paffenroth 2017).

2.3.5 Local Outlier Factor (LOF)

Local Outlier Factor (LOF) identifies anomalies by comparing the local density of points with that of their neighbors (Zhou et al. 2024). LOF assigns higher anomaly scores to points located in significantly less dense regions compared to their immediate neighborhood. The primary parameter, the number of neighbors, greatly influences sensitivity to local anomalies. LOF is especially valuable in ecological datasets, where locally anomalous observations may indicate important rare events or errors (Breunig et al. 2000).

2.3.6 k-Nearest-Neighbor (k-NN)

The k-Nearest Neighbor (k-NN) method identifies anomalies based on their distances to the k closest neighbors (Nizan & Tal 2024). Data points with significantly greater distances compared to typical data points are flagged as anomalies. The parameter k determines how many neighbors are considered, directly influencing Anomaly Detection sensitivity and specificity. While intuitive, the k-NN method may require computational optimizations to handle large datasets efficiently (Ramaswamy et al. 2000).

2.3.7 iNNE

Isolation using Nearest Neighbor Ensemble (iNNE) is an efficient, nearest neighbor-based method that isolates anomalies by constructing spherical boundaries around randomly sampled points Bandaragoda et al. (2014). Anomalies are points that are isolated by large hyperspheres. iNNE addresses key weaknesses of earlier isolation methods like Isolation Forest, including sensitivity to local anomalies and anomalies that exist near clusters of normal instances. With linear time complexity, iNNE is highly scalable and effective in large or high-dimensional datasets typically encountered in ecological studies (Bandaragoda et al. 2014).

3 Related Work

Citizen science has become a powerful tool for gathering large-scale biodiversity data, yet these open participation initiatives inevitably grapple with data quality control. Observations can include errors in species identification, spatiotemporal misplacements, or even fabricated submissions. Consequently, numerous verification strategies-ranging from manual expert reviews to fully automated Anomaly Detection-have been implemented to filter out erroneous records while retaining genuine rarities.

3.1 Current Solutions for Validation

This section provides an overview of existing approaches for verifying citizen science data in ecological research, with a particular emphasis on automated validation methods. Additionally, expert-based and community-driven verification strategies are examined to provide a comprehensive perspective on data quality assurance.

Traditionally, expert validation has served as the primary mechanism for ensuring the accuracy of citizen science data. In this approach, professional scientists manually assess the correctness of submitted observations to uphold scientific strictness (Pocock et al. 2024). While effective, this method presents significant challenges regarding scalability, as the increasing volume of submitted data surpasses the capacity of available experts (Baker et al. 2021). Furthermore, dependence on a limited number of specialists can lead to bottlenecks, delaying data availability and impeding timely decision-making. These limitations have driven the exploration of alternative or complementary verification strategies that maintain data integrity while accommodating large-scale datasets.

One such alternative is community-based validation, in which participants evaluate each other's submissions following predefined guidelines or through peer consensus. A prominent example is *Naturalist*, a platform that contributes to the dataset used in this study. *Naturalist* employs a consensus-based identification system, classifying observations as "research-grade" once a sufficient number of users agree on the species identity (Campbell et al. 2023). Although this approach alleviates the burden on experts, specialist oversight remains necessary for resolving uncertain or contentious observations, thereby balancing scalability with accuracy.

Recent advancements in citizen science verification have introduced automated methodologies. These approaches primarily function as pre-screening tools that identify potential misclassifications for expert review, thereby reducing manual validation efforts while preserving data reliability. Automated verification systems filter probable errors and facilitate real-time feedback loops, allowing contributors to refine system accuracy over time (Baker et al. 2021).

Automated approaches are particularly well - established in the context of image - based data. For example, Lotfian et al. (2019) present a Deep Learning model designed to classify images submitted by volunteers. However, for tabular datasets, automated validation remains relatively uncommon (Baker et al. 2021).

Nonetheless, certain platforms, such as *eBird* and *Project FeederWatch*, incorporate automated checks into their validation pipelines:

- ***Project FeederWatch*** combines automated Anomaly Detection with community-based review, delegating expert intervention to cases involving highly unusual reports (Bonter & Cooper 2012).
- ***eBird*** employs Emergent Filters to compare reported sightings against historical records and species distribution models, flagging anomalous records for further scrutiny (Kelling et al. 2019). *eBird* incorporates observer expertise as a weighting factor, assigning greater credibility to contributions from experienced users. The algorithmic details of the eBird approach is explained in detail in Chapter 2.2.3. Their approach is adopted in this thesis and motivated the exploration of more sophisticated filters.

These approaches are relatively dated, being incorporated in 2011 and 2012, respectively. However, they remain operationally relevant and align closely with the objectives of this study, as they focus on ornithological observations.

These established approaches primarily rely on rule-based heuristics and basic statistical filtering. Researchers therefore highlight a need for more sophisticated verification tools. For instance, Lotfian et al. (2021) advocate for the integration of Species Distribution

Models (SDMs) to gauge whether a reported sighting is ecologically plausible in a given location. While such models have demonstrated utility in assessing validity, they have yet to be systematically integrated into mainstream verification pipelines, leaving a gap between theoretical potential and practical implementation (Lotfian et al. 2021). Moreover, Sheard et al. (2024) highlight the potential of artificial intelligence-based validation systems for providing instantaneous feedback, underscoring the necessity for further innovation in automated Anomaly Detection.

4 Implementation

To address the research objectives set in Chapter 1, a multi-stage framework was developed to identify and investigate anomalous or outlier bird sightings. This framework encompasses an end-to-end pipeline-from raw data acquisition and feature generation to Emergent Filter design, Machine Learning design, and an interface for expert review. The primary goal is to capture unusual sightings that may indicate data-entry errors, biologically implausible records, or legitimate shifts in bird distribution patterns. The chapter begins with a detailed requirement analysis, specifying both functional and non-functional requisites. Subsequently, a modular software design is introduced, illustrating how the system's components are structured and how they interact. Finally, each module is described comprehensively, covering aspects of data ingestion, feature engineering, threshold tuning, model selection, and the Gradio-based reviewing interface employed to gather expert annotations.

4.1 Requirement Analysis

The requirement analysis serves to delineate the key operational and technical constraints necessary for building a robust Outlier Detection pipeline. In essence, it captures what the framework must accomplish (functional requirements) and how it must perform (non-functional requirements). These requirements ensure that the final implementation is both effective in detecting bird sighting anomalies and efficient in handling a large volume of geospatial data.

4.1.1 Functional Requirements

- **Data ingestion and preparation:** The system shall be capable of reading, merging, preprocessing and filtering bird sighting data from various sources. A taxonomic filtering mechanism shall limit the dataset to target species of interest.

- **Feature generation:** The framework shall produce enriched features (e.g., land cover and altitude) for each sighting to provide contextual information essential for Outlier Detection.
- **Emergent Filter computation:** Domain-inspired filters (e.g., date/location plausibility, habitat plausibility, altitude plausibility) shall flag suspicious records based on historical species-specific distributions.
- **Machine Learning implementation:** Multiple unsupervised ML models shall be implemented and optimized to systematically identify anomalous data points.
- **Reviewing interface:** An interface shall present the datapoint with the Emergent Filter and Machine Learning output to allow expert reviewers to label records as "Outlier" or "Not Outlier." If flagged as an outlier, reviewers shall be able to specify one or more reasons (e.g., implausible date, unusual altitude, etc.).
- **Result logging:** Reviewer decisions, along with relevant metadata, shall be logged for later evaluation of model performance.

4.1.2 Non-Functional Requirements

- **Scalability and performance:** The system shall handle thousands or even millions of sightings efficiently. Feature generation and model training processes shall be optimized to avoid excessive run times.
- **Modularity and maintainability:** The software shall be architected in a manner that separates core functionalities into distinct, reusable modules. This facilitates easier testing, debugging, and future extensions.
- **Data integrity:** Procedures for handling missing, invalid, or out-of-bounds values shall be strictly defined to maintain the reliability of outputs.
- **Usability:** The reviewer interface shall be straightforward enough to be operated by ornithologists with minimal technical expertise.
- **Reproducibility:** All data preprocessing and Outlier Detection procedures shall be clearly documented and version-controlled, supporting scientific validation and further development.

- **Robustness:** Mechanisms shall exist to detect and handle exceptions (e.g., unclassified habitat data, out-of-DEM range for altitude) so that the system's outputs remain meaningful even under imperfect data conditions.

4.2 Software Design

Given the breadth of requirements, a carefully structured software architecture is essential. A modular approach enables different parts of the system-such as feature extraction, Emergent Filtering, and Machine Learning to evolve independently while minimizing the ripple effects of changes. This approach is particularly advantageous in notebook-based development, where clear demarcations between data loading, processing steps, and result presentation can mitigate complexity.

A modular architecture was chosen, wherein each module encapsulates a specific functionality or set of related tasks. Figure 4.1 illustrates the top-level design, showing how data flows from the raw acquisition stage, through feature engineering and Emergent Filters, to the Machine Learning models and final reviewer interface.

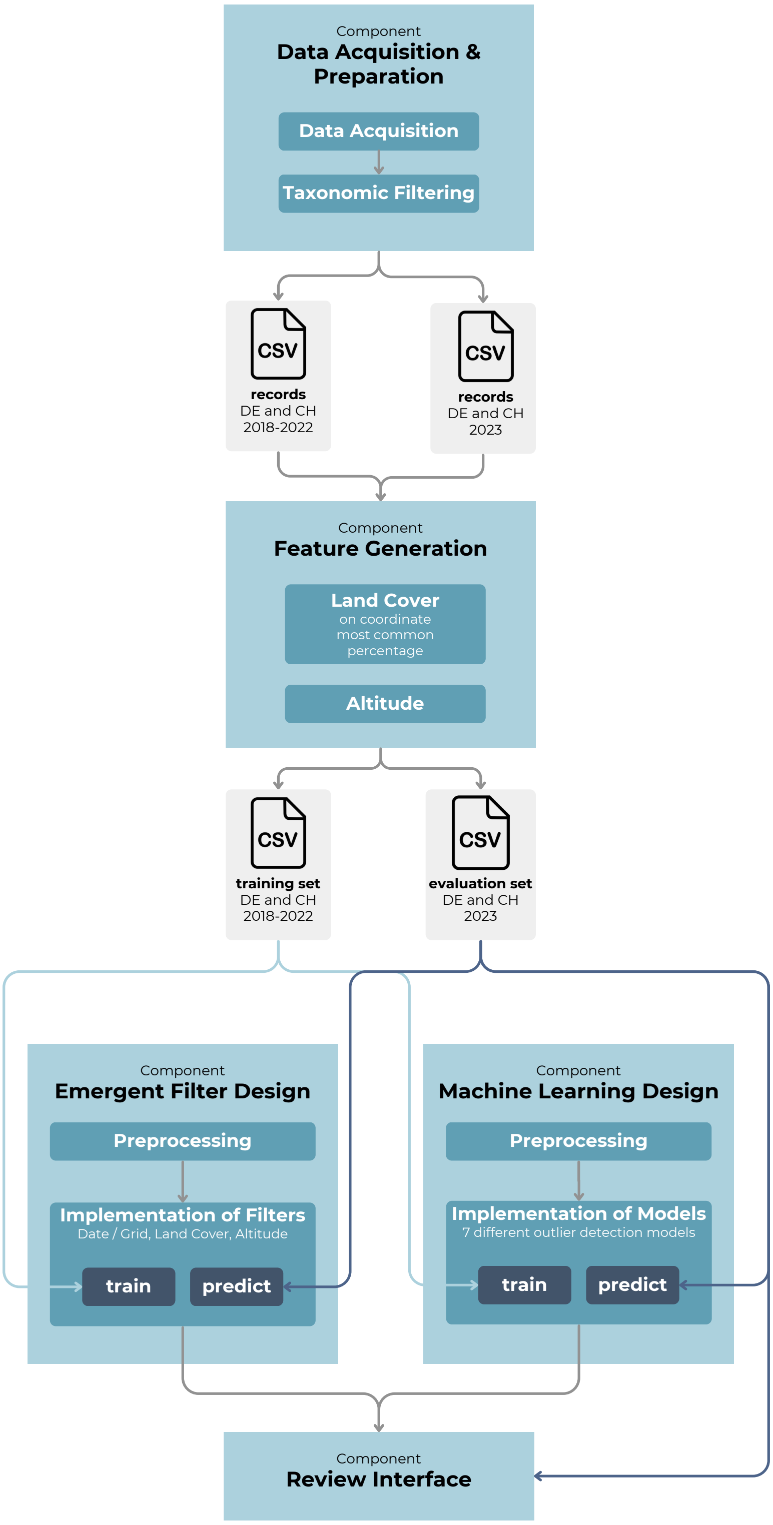


Figure 4.1: High-level overview of the software architecture, illustrating how each component is used.

To address the requirements outlined above, the system is divided into the following key components:

- **Data Acquisition & Preparation**
- **Feature Generation**
- **Emergent Filter Design**
- **Machine Learning Design**
- **Gradio Interface**

Each of these components is described in further detail below, providing insights into their respective functionalities, inputs, outputs, and internal processes.

4.3 Component Descriptions

This section details the core modules that make up the implemented framework, from the initial acquisition of bird sightings data to the final expert review of flagged anomalies. Each module addresses a specific part of the Outlier Detection workflow, enabling a structured, maintainable, and extensible system architecture.

4.3.1 Data Acquisition & Preparation

The data used in this thesis originates from systematic bird sighting records collected via the platforms *ornitho.ch* and *ornitho.de*, operated respectively by the Swiss Ornithological Institute (Schweizerische Vogelwarte) and the Dachverband Deutscher Avifaunisten (DDA). These citizen science platforms collect large volumes of ornithological observations across Switzerland and Germany, providing extensive datasets for ecological and temporal analyses.

The complete dataset is subdivided into two parts: a **training dataset** covering the years **2018 to 2022**, and an **evaluation dataset** corresponding to observations from **2023**. The current chapter focuses exclusively on the training data, which serves as the basis for model development and initial analysis. The evaluation dataset will be introduced separately in Chapter 5.

Data Source

Data were obtained from exports of *ornitho.ch* and *ornitho.de*. Each record corresponds to a single sighting event and contains the essential elements required for spatio-temporal modeling. Overall, the combined dataset initially comprised approximately **50 million individual sighting records** for **821 bird species**, providing a high-resolution view of bird distributions across Central Europe.

Before any processing, standardization procedures were performed to ensure consistency across both data sources. These included:

- Verification and harmonization of coordinate systems (EPSG:4326, WGS84 latitude/longitude).
- Standardization of date formats.
- Consistency checks to confirm that all required attributes (coordinates, species identifiers, and timestamps) were uniformly available across records.

Data Features

Each record in the dataset contains the following attributes:

- **Latitude** and **Longitude**: Geographical coordinates of the sighting (EPSG:4326).
- **Date**: The date when the observation was made (standardized as YYYY-MM-DD).
- **Atlas Code**: A code indicating bird breeding or migratory status (collected but not used in this project).
- **Bird Sighting Count**: The number of individuals observed (collected but not used in this project).
- **Altitude**: Altitude (in meters above sea level) of the sighting location.

Taxonomic Filtering

Given the extensive size and taxonomic breadth of the original dataset, which included records for 821 unique bird species, a strategic filtering step was necessary to ensure computational feasibility and to focus on ecologically significant species.

In collaboration with ornithologists from the Swiss Ornithological Institute and the DDA, a subset of **27 bird species** was selected. These species were identified as particularly interesting for scientific investigation, based on criteria such as ecological relevance, population trends, and spatial dynamics.

Thus, the training dataset used in this thesis consists exclusively of records of these 27 species, observed between 2018 and 2022. Table 4.1 provides an overview of the number of records per species within this filtered dataset.

Focusing the dataset on these 27 ornithologist-selected species ensures that the subsequent modeling steps remain both ecologically meaningful and computationally tractable.

4.3.2 Feature Generation

Augmenting the dataset with additional contextual attributes forms the foundation for more accurate Outlier Detection. Two primary types of features, *land cover* and *altitude*, were generated to reflect the environmental conditions at each sighting location. While some features (like altitude) were partially present in the input data, more advanced or fallback processes were implemented to handle missing or incorrect values. Weather features were considered but eventually excluded due to the high computational effort relative to the insights gained.

Land Cover

Land cover features capture the habitat types (e.g., forests, water bodies, agricultural areas) surrounding a bird sighting. Three distinct approaches were employed to integrate land cover information into the bird sightings data, each offering a different perspective on the spatial relationship between bird locations and their surrounding environments.

Species Name	**Number of Records**
Rock Ptarmigan	5933
Western Capercaillie	5242
Greater Scaup	15091
Twite	7756
Water Pipit	107835
Whinchat	129799
Three-toed Woodpecker	5180
Common Sandpiper	149326
Icterine Warbler	58037
Griffon Vulture	8593
Great Crested Grebe	428589
Common Rosefinch	7818
Garganey	86313
Middle Spotted Woodpecker	100175
Western Orphean Warbler	8898
Reed Bunting	277469
Ruddy Shelduck	99974
European Stonechat	262291
Black Kite	239725
White-tailed Eagle	139710
Whooper Swan	83481
Eurasian Pygmy Owl	13886
Northern Wheatear	120127
White-throated Dipper	88346
Meadow Pipit	196863
Citril Finch	9988
Eurasian Scops Owl	3598
Total	**2,660,043**

Table 4.1: Number of sighting records for the 27 selected bird species in the training dataset (2018-2022).

On-Coordinate Land Cover In this simplest approach, the exact coordinate of each sighting is matched to a land cover polygon from the Corine Land Cover (CLC) dataset. Formally, let

$$\mathcal{P} = \{\, P_k \mid k = 1, \ldots, n\}$$

be the set of land cover polygons (e.g., *forest*, *water*, etc.) that together form a tessellation of the study area. For a sighting with coordinates (x_i, y_i), that match the polygon $P_k \in \mathcal{P}$

such that

$$(x_i, y_i) \in P_k.$$

If such a polygon exists, the sighting is assigned the category corresponding to P_k. If (x_i, y_i) does not lie within any polygon in $\mathcal{P}$, a default label *unclassified* is assigned.

Percentage of Each Land Cover within 1 km^2 A more granular perspective computes the exact share (in percentage) of each land cover category inside a 1 km^2 buffer. Multiple columns (e.g., `forest_area_percent`, `urban_area_percent`, etc.) are added to the dataset, each reflecting the fraction of that cover type within the buffer. While this yields a rich habitat profile, it is also more computationally demanding.

Most Common Land Cover within 1 km^2 For this approach, each sighting is associated with the most dominant CLC polygon type in a 1 km^2 buffer. By analyzing the proportion of each land cover provided by CLC within this buffer, the algorithm identifies which category covers the largest area around the sighting. The most prevalent land cover category is assigned as the "most common" habitat descriptor. This approach simplifies multiple coverage percentages into a single, representative type, balancing granularity and interpretability.

Altitude

Altitude is another key ecological factor influencing where bird species typically forage or breed. Although many records already contained altitude information, additional steps were taken to address coverage gaps and potential outliers:

- **EU-DEM Integration:** The European Digital Elevation Model (EU-DEM) was locally processed, aligning each sighting coordinate with the corresponding raster cell to retrieve a reliable elevation value.
- **Imputation for Coastal Sightings:** Undefined elevations, which only occurred near coasts and marine areas were set to zero to correctly reflect these areas. This ensures that the completeness of the dataset is maintained without discarding valuable datasets.

4.3.3 Emergent Filter Design

Emergent Filters incorporate species-specific ecological knowledge directly into the Anomaly Detection process, complementing the general-purpose Machine Learning models. By constructing feature-specific lookup tables, Emergent Filters enable a biologically grounded plausibility assessment for each sighting.

Preprocessing

Before generating the Emergent Filters, the raw sighting data underwent basic preprocessing steps to ensure consistency and reliability:

- **Column Filtering:** Removal of irrelevant columns such as internal identifiers or fields not needed for the filter design.
- **Missing Value Check:** All rows containing missing values (NaNs) were removed to prevent propagation of errors during filter creation.
- **Data Type Standardization:** Relevant columns (e.g., species name, coordinates, date, altitude) were checked and converted to uniform data types.

This preprocessing ensured that the input data for all Emergent Filters was clean, consistent, and complete.

Implementation of Filters

Three types of Emergent Filters were developed, based on:

1. Date and location (grid cell and day-of-year),
2. Land cover types,
3. Altitude.

Each filter follows a two-stage process: *training* (construction of lookup tables) and *prediction* (application of thresholds).

Date / Location (Grid + Day-of-Year) **Training Phase:** For each species s, grid cell g, and day of year d, the relative plausibility of an observation was estimated as:

$$\text{frequency}(s, g, d) = \frac{\text{count sightings of } s \text{ on day } d \text{ in grid } g}{\text{total sightings in grid } g \text{ on day } d}$$

To smooth out sampling noise, a two-step circular rolling smoothing was applied: First, a 7-day centered maximum filter to account for temporal neighborhoods, then a 30-day centered moving average to produce the final plausibility value:

$$\text{plausibility}(s, g, d) = \text{mean}\left(\max\left(\text{frequency}(s, g, d-3 : d+3)\right) \text{ over } d-15 : d+15\right)$$

where wrap-around at year boundaries is handled (i.e., days 362-365 are neighbors to days 1-4).

Prediction Phase: Given a new sighting (s, g, d), the plausibility is looked up:

$$\text{plausibility}_{\text{new}} = \text{lookup}(s, g, d)$$

If $\text{plausibility}_{\text{new}} < \tau_{\text{date/grid}}$, the sighting is flagged as implausible. Here, $\tau_{\text{date/grid}}$ denotes the decision threshold (typically $\tau_{\text{date/grid}} = 0.05$).

Land Cover Filter **Training Phase:** For each species s, the mean share of each land cover type c_i was calculated:

$$\text{mean\_landcover}(s, c_i) = \frac{1}{N_s} \sum_{j=1}^{N_s} c_{i,j}$$

where N_s is the number of observations for species s and $c_{i,j}$ is the percentage of land cover type i for sighting j.

Prediction Phase: For a new sighting, the Euclidean distance between its observed land cover vector $\mathbf{c}_{\text{obs}}$ and the species' mean land cover vector $\mathbf{c}_{\text{mean}}$ is computed:

$$d_{\text{landcover}} = \sqrt{\sum_{i=1}^{k} (c_{\text{obs},i} - c_{\text{mean},i})^2}$$

If $d_{\text{landcover}} > \tau_{\text{landcover}}$, the sighting is flagged as anomalous. Here, $\tau_{\text{landcover}}$ represents the land cover deviation threshold.

Altitude Filter **Training Phase:** Altitude measurements were discretized into bins of size Δh (e.g., $\Delta h = 50$ m). For each species s and altitude bin b, the smoothed probability was estimated using Laplace smoothing:

$$P(s, b) = \frac{n(s, b) + \alpha}{N_s + \alpha \cdot B}$$

where:

- $n(s, b)$ = number of sightings of species s in bin b,
- N_s = total number of sightings of species s,
- B = number of altitude bins,
- α = smoothing parameter (typically $\alpha = 1$).

Prediction Phase: For a new sighting with altitude h_{new}, the corresponding bin b_{new} is identified. The plausibility is then:

$$P_{\text{altitude}} = P(s, b_{\text{new}})$$

If $P_{\text{altitude}} < \tau_{\text{altitude}}$, the sighting is considered implausible. The threshold τ_{altitude} typically corresponds to low-probability regions (e.g., $\tau_{\text{altitude}} = 0.05$).

4.3.4 Machine Learning Design

Alongside domain-inspired filters, a suite of unsupervised Machine Learning (ML) algorithms was employed to detect anomalies in the enriched dataset. A structured Machine Learning pipeline was implemented, including consistent preprocessing, model training, hyperparameter tuning, and prediction. Each algorithm was extensively tested and tuned through systematic optimization procedures. The implementations were performed using established Python libraries, as outlined below.

Preprocessing

Prior to model training, several preprocessing steps were conducted to standardize the feature space:

- **Coordinate Transformation:** Geographic coordinates (latitude, longitude) were projected into the `EPSG:3035` coordinate system (European-centric metric projection) to better preserve distances for distance-based algorithms.
- **Feature Engineering for Dates:** Day-of-year features were transformed into cyclical features using sine and cosine transformations:

  $$\mathrm{sin\_day} = \sin\left(2\pi\frac{\mathrm{day\_of\_year}}{365}\right), \quad \mathrm{cos\_day} = \cos\left(2\pi\frac{\mathrm{day\_of\_year}}{365}\right)$$

  to maintain continuity across year boundaries.
- **Feature Scaling:** Altitude and the reprojected (x, y) coordinates were standardized using a `StandardScaler`, ensuring that features had zero mean and unit variance.

These transformations enabled fair model comparisons and stable numerical behavior across the different Machine Learning algorithms.

Implementation of Models

Seven unsupervised Anomaly Detection algorithms were selected based on their effectiveness in previous research and their suitability for the spatial and environmental nature of bird observation data.

Model	**Source**	**Model Type**	**Optimized Parameters**
DBSCAN	sklearn	Density-based clustering	eps
HDBSCAN	hdbscan	Density-based clustering	min_cluster_size
Isolation Forest	pyod	Isolation-tree ensemble	n_estimators, max_samples
AutoEncoder	pyod	Neural network	hidden_neurons, activation_function
iNNE	pyod	Nearest-neighbor ensemble	max_samples
k-NN	pyod	Distance-based method	n_neighbors
LOF	pyod	Local density-based	n_neighbors

Table 4.2: Overview of Machine Learning models, source libraries, model types, and optimized parameters.

DBSCAN DBSCAN was implemented using `sklearn.cluster.DBSCAN`. *Reason for selection:* DBSCAN is a classic density-based clustering algorithm that identifies clusters as areas of higher point density and treats sparse regions as outliers, making it suitable for spatial Anomaly Detection tasks. *Optimized parameter:* The parameter `eps` defines the maximum distance between two samples for them to be considered neighbors. Optimizing `eps` is critical because it determines the scale at which clusters are formed. If `eps` is too small, true clusters may be split, incorrectly flagging normal points as anomalies. If `eps` is too large, distinct clusters may merge, causing true outliers to be missed. Thus, careful selection of `eps` balances sensitivity and specificity for Anomaly Detection.

HDBSCAN HDBSCAN was implemented using the `hdbscan` library. *Reason for selection:* HDBSCAN states to improve on DBSCAN by allowing clusters of varying density, making it particularly effective for datasets with heterogeneous sampling densities, such as bird sightings across diverse landscapes. *Optimized parameter:* The parameter `min_cluster_size` defines the smallest number of points needed to form a cluster. Optimizing `min_cluster_size` controls how tolerant the algorithm is to noise. Smaller values allow smaller clusters but risk treating noise as structure. Larger values suppress small patterns, potentially missing legitimate but rare occurrences. Therefore, tuning `min_cluster_size` is essential to distinguish between rare bird events and genuine outliers.

Isolation Forest Isolation Forest was implemented using `pyod.models.iforest`. *Reason for selection:* Isolation Forest is an ensemble-based method that isolates anomalies by randomly partitioning feature space, offering good performance even in high-dimensional data. *Optimized parameters:* The `n_estimators` indicates the number of isolation trees in the ensemble, while `max_samples` is the number of samples to draw for each tree. Increasing `n_estimators` generally improves model stability but with diminishing returns beyond a certain point. Adjusting `max_samples` affects how fine-grained the isolation becomes: smaller samples can highlight rare patterns better but may introduce noise. Thus, optimizing both parameters enhances the model's ability to robustly isolate rare or unusual bird observations.

AutoEncoder The AutoEncoder model was implemented using `pyod.models.auto_encoder`. *Reason for selection:* AutoEncoders can learn compressed internal representations of the input data, making them highly effective at capturing complex non-linear structures and highlighting deviations. *Optimized parameters:* The `hidden_neurons` sets the structure and size of hidden layers and the `activation_function` is the non-linearity used within layers, which in this project can be `relu` or `tanh`. The hidden layer structure determines how well the network can reconstruct normal observations while amplifying reconstruction errors for anomalies, while the choice of activation function influences the model's capacity to capture complex relationships: `relu` generally favors sparse representations, while `tanh` captures smoother transitions. Tuning these parameters is thus critical to ensure the AutoEncoder properly distinguishes between typical and anomalous bird sightings.

iNNE The iNNE model was implemented using `pyod.models.inne`. *Reason for selection:* iNNE combines the advantages of nearest-neighbor detection and ensemble methods, making it robust to noise and varying densities. *Optimized parameter:* `max_samples`, defining the maximum number of samples used per ensemble model. A lower `max_samples` value enhances sensitivity to local structure but risks instabilityand a higher `max_samples` smooths the model but may miss rare patterns. Optimizing `max_samples` balances these effects to maximize detection performance across diverse species and regions.

k-NN The k-NN model was implemented using `pyod.models.knn`. *Reason for selection:* k-NN is a simple yet powerful method that detects anomalies based on distances to neighboring points, which is intuitive and effective for spatial data. *Optimized parameter:* `n_neighbors`, the number of neighbors considered. Choosing a small `n_neighbors` value makes the model sensitive to local variations but increases false positives whereas choosing a large `n_neighbors` value smooths the distance metric, potentially missing true outliers. Thus, tuning `n_neighbors` is key to balancing local sensitivity and global stability.

Local Outlier Factor (LOF) The LOF model was implemented using `pyod.models.lof`. *Reason for selection:* LOF measures the local density deviation of a data point relative to its neighbors, making it particularly suited for detecting anomalies in regions with variable densities. *Optimized parameter:* The `n_neighbors` indicates the number of neighbors used for density estimation. Having a small value for `n_neighbors` highlights small-scale anomalies but increases sensitivity to noise and having a large value for `n_neighbors` focuses on broader trends but risks smoothing out interesting local deviations. Optimizing `n_neighbors` enables LOF to detect both subtle and pronounced anomalies effectively.

Model Training

During the training phase:

- All models except DBSCAN were trained using their `fit()` methods.
- DBSCAN, being a clustering method without a separate prediction phase, was fitted and labeled directly using `fit_predict()`.

The feature set included standardized x, y, and altitude values, as well as cyclical date features and land cover percentages.

Model Prediction

The prediction stage differed slightly across models:

- **HDBSCAN:** Used `approximate_predict()`, treating label -1 as anomalies.
- **DBSCAN:** Directly used labels from `fit_predict()` (label -1 = outlier).

- **All other models:** Anomaly scores were obtained via the `decision_function()` method. A sighting was classified as anomalous if the decision score fell below a species-specific threshold:

$$\text{anomaly if} \quad \text{decision\_score} < \tau$$

where τ was optimized per species to maximize F1 score, as described in the Evaluation Chapter (Section 5).

4.3.5 Review Interface

To complete the Anomaly Detection pipeline, a web-based reviewing interface was developed using `Gradio`. The main goal of this interface is twofold: first, to provide a **proof-of-concept** demonstrating how expert reviews could be integrated into an Anomaly Detection system; and second, to enable a **qualitative evaluation** of the predictions generated by Emergent Filters and Machine Learning models.

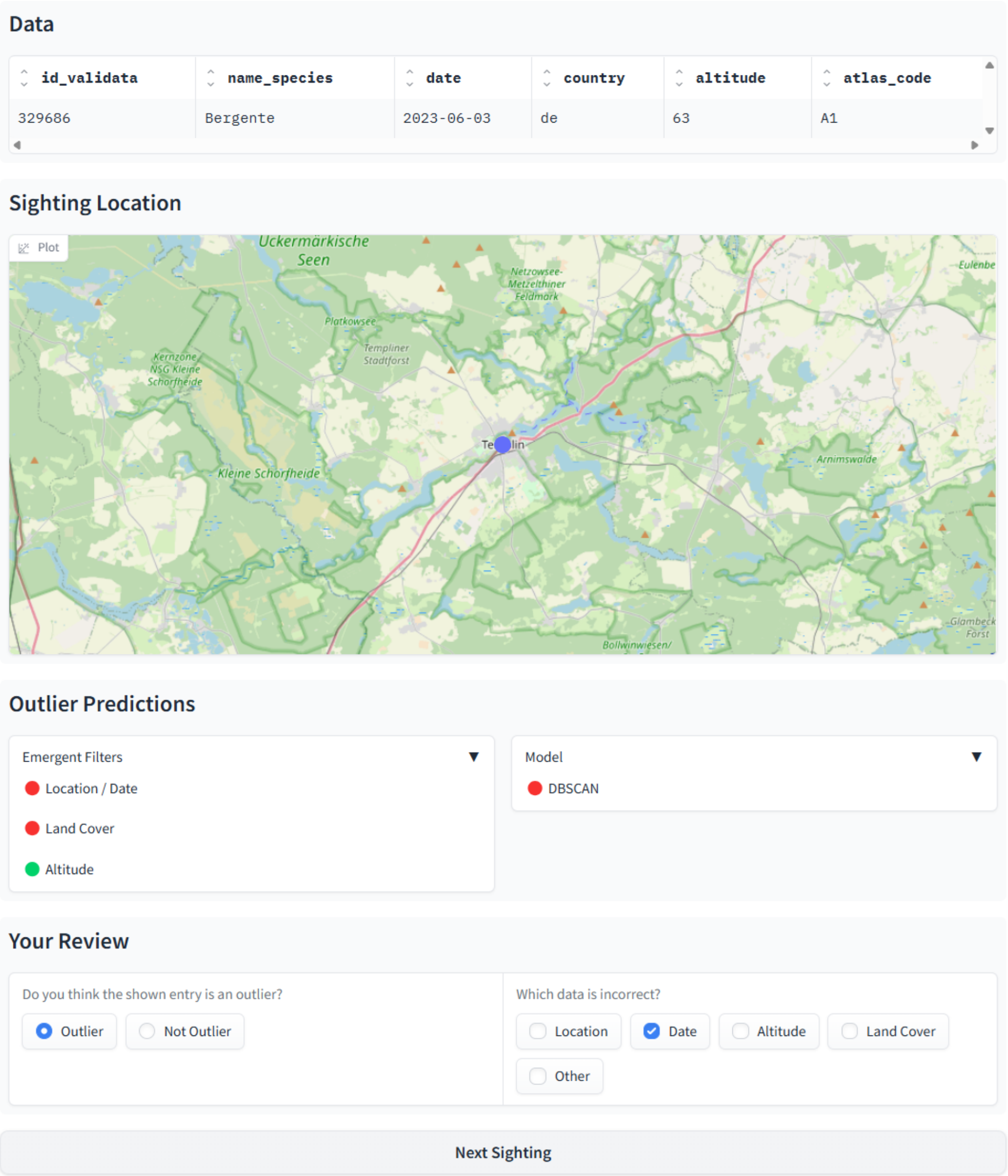
The interface was hosted on Hugging Face Spaces to ensure easy accessibility for participants via web browsers without requiring any local installations.

Explanation of Interface Components

Figure 4.2 shows the review interface. Each major component is briefly explained below.

Name Selection At the start of a session, the reviewer selects their name from a predefined list. This assignment ensures that each reviewer only accesses the correct subset of the validation data, enabling controlled experiment tracking and avoiding data duplication between reviewers.

Data Display The upper panel displays core sighting information, including the unique identifier (`id_validata`), species name, date of observation, country, altitude, and atlas code. This structured presentation ensures that the reviewer has immediate access to essential contextual information needed for decision-making.



Data

id_validata	name_species	date	country	altitude	atlas_code
329686	Bergente	2023-06-03	de	63	A1

Sighting Location

Outlier Predictions

Emergent Filters ▼

🔴 Location / Date

🔴 Land Cover

🟢 Altitude

Model ▼

🔴 DBSCAN

Your Review

Do you think the shown entry is an outlier?

● Outlier ○ Not Outlier

Which data is incorrect?

□ Location ☑ Date □ Altitude □ Land Cover □ Other

Figure 4.2: Screenshot of the developed review interface for bird sighting Anomaly Detection.

Sighting Location Below the data panel, an interactive map displays the exact geographic location of the sighting using OpenStreetMap layers. The map provides spatial context, helping reviewers judge the plausibility of sightings based on known species distributions and habitats.

Outlier Predictions In the “Outlier Predictions” section, the outputs of the Emergent Filters and the Machine Learning models are summarized:

- **Emergent Filters:** Location/Date, Land Cover, and Altitude plausibility checks are each visualized individually.
- **Model Predictions:** In this example the DBSCAN model output is displayed.

Each method outputs a simple green or red indicator: **green** indicates the method considers the sighting plausible, while **red** indicates the method flags the sighting as an outlier.

This visualization gives the reviewer an at-a-glance overview of which aspects of the sighting were considered unusual by the system.

Review Section In the “Your Review” panel, the reviewer provides their final judgment:

- They indicate whether they consider the sighting to be an **Outlier** or **Not Outlier**.
- If **Outlier** is selected, additional checkboxes appear, allowing the reviewer to specify which aspect(s) of the data they believe to be incorrect (e.g., Location, Date, Altitude, Land Cover, or Other).

This structured feedback enables fine-grained evaluation of model errors and strengths.

Next Sighting Button After completing the review for one sighting, the reviewer clicks the “Next Sighting” button to move to the next entry in their assigned dataset. Reviewer progress is automatically logged, and upon completion, a thank-you page with questions about the review process are displayed. Those questions will be further elaborated in Chapter 5.

Result Storage and Logging

All reviewer decisions are automatically logged into a CSV file, capturing:

- Entry ID
- Reviewer choice (Outlier/Not Outlier)
- If Outlier, specified reasons
- Timestamps

This structured review data forms the basis for qualitative evaluation of model and filter performance, enabling later analysis of reviewer agreement and error types.

4.3.6 Summary

This chapter presented the full implementation pipeline developed to detect anomalous bird sightings based on spatio-temporal and ecological patterns. Starting from a structured **requirement analysis**, both functional and non-functional needs were clearly defined to guide the system design. A **modular software architecture** was then introduced, enabling flexibility, maintainability, and targeted extension in the future.

Each major component of the system was subsequently detailed:

- **Data Acquisition & Preparation:** Raw bird sighting data from *ornitho.ch* and *ornitho.de* were collected, harmonized, and taxonomically filtered to focus on 27 expert-selected species of interest.
- **Feature Generation:** Environmental features such as land cover compositions and altitude were extracted and engineered to enrich the information available for each sighting.
- **Emergent Filter Design:** Species-specific, domain-inspired plausibility filters were developed for date-location relationships, habitat types, and altitudinal distributions, incorporating smoothing and probability modeling techniques.

- **Machine Learning Design:** Seven unsupervised Anomaly Detection algorithms (DBSCAN, HDBSCAN, Isolation Forest, AutoEncoder, iNNE, k-NN, LOF) were systematically implemented, including detailed preprocessing steps and implementation details.

- **Review Interface:** A Gradio-based web interface was created to enable expert reviewers to provide structured feedback on model outputs, generating a high-quality labeled dataset for evaluation.

Through this integrated approach, a flexible, transparent, and scalable framework was realized, setting the stage for the quantitative and qualitative evaluations discussed in the subsequent chapters.

5 Evaluation

5.1 Objective and Evaluation Strategy

The overarching goal of this chapter is to answer the research questions posed in Chapter 1 by evaluating the effectiveness of the developed Outlier Detection framework. The evaluation is conducted from both a *quantitative* and a *qualitative* perspective. While quantitative analyses provide objective measures of model and filter performance, the qualitative evaluation incorporates expert assessments to examine whether flagged anomalies are ecologically plausible or indicative of data issues.

This two-pronged strategy ensures a holistic understanding of the system's behavior: performance metrics capture predictive accuracy, while expert feedback sheds light on interpretability and usefulness.

5.2 Data Basis for Evaluation

5.2.1 The *validata* Dataset

The evaluation is based on the *validata* dataset, a curated subset of the bird sightings that was held out from the training data and augmented with controlled errors. It includes only the 27 bird species selected in collaboration with ornithologists (see Chapter 4), ensuring ecological relevance and feasibility for manual review.

In the dataset, each species is represented by a mix of unedited (plausible) and edited (implausible) sightings. The artificial errors were designed by the ornithologists to simulate realistic mistakes or rare phenomena, providing a suitable benchmark for model and filter evaluation.

5.2.2 Error Types and Distribution

Table 5.1 summarizes the total number of entries, the number of manipulated (edited) entries, and the amount of specific error types per species. In total, the dataset contains 3,902 edited sightings out of 417,886 total records.

Four main types of errors were introduced:

- **Date errors**: Observations shifted to biologically implausible dates (e.g., winter observations for a migratory species).
- **Distribution errors**: Coordinates and altitude moved far outside the species' typical range.
- **Habitat errors**: Sightings relocated to ecologically unsuitable habitats, which could include minor coordinate and altitude changes.
- **Count errors**: Unusual number of individuals recorded for a single sighting (this was not used in this project but is shown in the table).

These errors provide ground-truth labels against which the predictions of filters and models can be compared in subsequent evaluations.

5.3 Quantitative Evaluation

5.3.1 Objective

The aim of the quantitative evaluation is to systematically measure how effectively each Emergent Filter and unsupervised Machine Learning model detects artificially injected errors in the `validata` dataset. For this, the following sub-questions are relevant:

- How well does each method perform overall in identifying manipulated entries?
- Do different feature sets influence model performance?
- Are some species or error types more difficult to detect than others?

To answer these questions, a two-phase analysis is conducted:

name_species	total_entries	total_edited	count	date	distribution (coord)	habitat (coord)
Rock Ptarmigan	914	150	31	0	69	50
Western Capercaillie	790	149	19	0	69	61
Greater Scaup	2810	203	35	34	101	33
Twite	922	35	14	0	0	21
Water Pipit	12552	167	0	68	33	66
Whinchat	23706	150	19	35	63	33
Three-toed Woodpecker	1147	154	18	36	50	50
Common Sandpiper	24434	203	0	48	80	75
Icterine Warbler	15044	150	0	84	33	33
Griffon Vulture	1225	115	0	65	50	0
Great Crested Grebe	68101	151	0	0	50	101
Common Rosefinch	1900	135	0	34	68	33
Garganey	16239	199	49	84	33	33
Middle Spotted Woodpecker	23722	154	0	0	50	104
Western Orphean Warbler	1875	137	0	34	70	33
Reed Bunting	47408	151	0	34	33	84
Ruddy Shelduck	17454	120	0	0	70	50
European Stonechat	7335	100	0	0	50	50
Black Kite	49010	149	0	99	50	0
White-tailed Eagle	26050	149	0	34	82	33
Whooper Swan	12314	151	0	34	84	33
Eurasian Pygmy Owl	2746	150	0	0	50	100
Northern Wheatear	19304	150	0	84	33	33
White-throated Dipper	13616	153	0	0	0	153
Meadow Pipit	24778	155	0	34	33	88
Citril Finch	1571	110	0	34	43	33
Eurasian Scops Owl	919	112	0	34	45	33
All Species	417886	3902	185	909	1392	1416

Table 5.1: Key attributes for the 27 validata species, including total edited entries and associated error types.

1. **Performance Analysis**: Determine the optimal settings for filters and models and quantify their predictive performance.
2. **Pattern Analysis**: Identify systematic differences in performance across species and error types.

5.3.2 Evaluation Metric: F1 Score

Performance is measured using the F1 score, which balances **precision** (i.e., how many predicted anomalies are true anomalies) and **recall** (i.e., how many true anomalies are correctly predicted). This is particularly appropriate for imbalanced datasets like validata, where anomalous entries represent a small fraction of the total data.

Recall, also known as sensitivity, is a performance metric that measures the model's ability to correctly identify all relevant positive instances. It is defined as:

$$Recall = \frac{TP}{TP + FN}$$

where TP (True Positives) represents the number of correctly predicted positive cases, and FN (False Negatives) denotes the number of actual positive cases that the model failed to identify. A high recall indicates that the model successfully captures most of the true positive instances, which is especially important in scenarios where missing a positive case would be costly or critical.

Precision measures the accuracy of the positive predictions made by the model. It indicates how many of the instances predicted as positive are actually correct. Precision is defined as:

$$Precision = \frac{TP}{TP + FP}$$

with TP and FN as previously defined. A high precision indicates that when the model predicts a positive instance, it is usually correct, which is an important property in contexts where false positives should be minimized.

The F1-Score is the harmonic mean of precision and recall and provides a single metric that balances both. It is given by:

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

This metric ensures that models are rewarded only when they are both sensitive and specific, penalizing methods that generate too many false positives or miss too many true outliers.

5.3.3 Performance Analysis

Grid Search for Thresholds and Parameters All Emergent Filters and ML models were evaluated across a predefined set of thresholds or parameter configurations. These grid searches were executed per species to account for inter-species differences and to enable species-specific threshold optimization.

Emergent Filters The following thresholds and feature inputs were tested for each filter:

Filter	**Feature Inputs**	**Threshold Values**
Grid-Year Filter	Grid Cell ID, Day-of-Year	{1e-6, 1e-5, 0.0001, 0.001, 0.025, 0.05, 0.1, 0.135, 0.15, 0.25, 0.4, 0.5, 0.6}
Land Cover Filter	Land cover percentages (e.g., `forest_area_percent`)	{0.001, 0.1, 0.2, 0.3, ..., 1.4}
Altitude Filter	Altitude bins (50m)	{0.0001, 0.0005, 0.001, 0.005, 0.05, 0.1, 0.15, 0.3, 0.4}

Table 5.2: Emergent Filters with used feature inputs and threshold search spaces.

Machine Learning Models The following feature sets were used to train and test models with different environmental contexts:

Feature Set	**Included Features**
No Land Cover	Coordinates, Altitude, Date
No Altitude	Coordinates, Land Cover, Date
No Coordinates	Altitude, Land Cover, Date
No Date	Coordinates, Altitude, Land Cover
All Features	Coordinates, Altitude, Land Cover, Date

Table 5.3: Feature set configurations used for model training and evaluation.

Each model was evaluated using the hyperparameter search spaces defined in Table 5.4.

Evaluation Method For each species:

1. Emergent Filters and models generated binary predictions for all entries in `vali-data`.

2. Predictions were compared against ground-truth labels to compute confusion matrices.

3. F1 scores were calculated per species and averaged across all 27 species.

4. Best thresholds or hyperparameter configurations were selected for each species and model based on the highest F1 score.

5. The best filters and models were compared with one another.

Model	Fixed Parameters	Optimized Parameters (search space)
DBSCAN	`min_samples = 5`	`eps: {0.05-1.5, step=0.1}`
HDBSCAN	`prediction_data = True`	`min_cluster_size: {2, 25, 50, 100, 250, 450}`

Isolation Forest	`random_state = 42`	`n_estimators: {25, 50, 100, 200, 300, 400}`
		`contamination`
		`max_samples: {100, 1000, 10000, 50000, 100000}`

AutoEncoder	`epochs = 10`
	`batch_size = 32`
	`optimizer = adam`
	`preprocessing = False`
		`hidden_neuron_list: {(12,6), (12,8,6), (14,10,6), (16,12,8,6), (64,32)}`
		`hidden_activation: {relu, tanh}`
		`contamination`

iNNE	`random_state = 42`	`max_samples: {10-100, step=10}`
		`contamination`

k-NN	`method = mean`	`n_neighbors: {2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 15, 17, 20}`
		`contamination`

LOF	`random_state = 42`	`n_neighbors: {10, 20, 50, 100, 150, 200, 250, 300, 350, 400, 500, 600}`
		`contamination`

Table 5.4: Fixed and optimized parameters with respective search spaces used in hyperparameter tuning.

5.3.4 Pattern Analysis

Species Group Analysis

To evaluate whether performance varies systematically across ecological traits, the 27 species in the `validata` dataset were classified into predefined biological and ecological categories. The classification was provided by the DDA and includes both migration-related attributes and habitat associations.

The following five grouping criteria were used:

- **Migration Behavior**: Indicates whether a species is non-migratory, partially migratory, or fully migratory.
- **Migration Distance**: Categorized into short-, medium-, long-distance and non-migratory.

- **Breeding Habitat**: Main habitat type used for breeding, including alpine regions, forest, inland waters, open land, and others.
- **Feeding Habitat (Breeding Season)**: Main foraging environment during the breeding season.
- **Feeding Habitat (Winter)**: Main foraging environment during winter.

The full species-to-group mappings used for grouped performance analysis are presented in Table 5.5 and 5.6. Some categories require additional clarification:

- **Special Locations**: Refers to niche or artificial environments that do not fall into traditional habitat categories (e.g., gravel pits, rail embankments, construction zones).
- **Not Defined**: Indicates that habitat or migration traits could not be clearly assigned based on existing data.
- **Global Variables**: Captures habitat preferences that are either too broad or general to be assigned to a specific category.

Table 5.5: Species grouped by breeding habitat and migration traits.

Species	**Breeding habitat**	**Migration behavior**	**Migration distance**
Rock Ptarmigan	Alpine regions	Non Migratory	Non Migratory
Western Capercaillie	Alpine regions	Non Migratory	Non Migratory
Greater Scaup	Inland waters	Partial Migratory	Short-distance
Twite	Not defined	Migratory	Medium-distance
Water Pipit	Alpine regions	Migratory	Short-distance
Whinchat	Open land	Migratory	Long-distance
Three-toed Woodpecker	Alpine regions	Non Migratory	Non Migratory
Common Sandpiper	Inland waters	Migratory	Long-distance
Griffon Vulture	Not defined	Not defined	Non Migratory
Icterine Warbler	Multiple main habitat types	Migratory	Long-distance
Great Crested Grebe	Inland waters	Partial Migratory	Short-distance
Common Rosefinch	Multiple main habitat types	Migratory	Long-distance
Garganey	Inland waters	Migratory	Long-distance
Middle Spotted Woodpecker	Forest	Non Migratory	Non Migratory
Western Orphean Warbler	Special locations	Migratory	Long-distance
Reed Bunting	Inland waters	Migratory	Short-distance
Ruddy Shelduck	Inland waters	Partial Migratory	Short-distance
European Stonechat	Multiple main habitat types	Migratory	Long-distance
Black Kite	Forest	Non Migratory	Non Migratory
White-tailed Eagle	Inland waters	Non Migratory	Non Migratory
Whooper Swan	Inland waters	Partial Migratory	Short-distance
Eurasian Pygmy Owl	Forest	Non Migratory	Non Migratory
Northern Wheatear	Special locations	Migratory	Long-distance
White-throated Dipper	Inland waters	Non Migratory	Non Migratory
Meadow Pipit	Multiple main habitat types	Migratory	Short-distance
Citril Finch	Alpine regions	Migratory	Medium-distance
Eurasian Scops Owl	Multiple main habitat types	Migratory	Long-distance

Table 5.6: Species grouped by feeding habitat during breeding and winter seasons.

Species	**Feeding habitat (breeding)**	**Feeding habitat (winter)**
Rock Ptarmigan	Alpine regions	Alpine regions
Western Capercaillie	Alpine regions	Alpine regions
Greater Scaup	Inland waters	Coast and sea
Twite	Not defined	Coast and sea
Water Pipit	Alpine regions	Inland waters
Whinchat	Open land	Special locations
Three-toed Woodpecker	Alpine regions	Alpine regions
Common Sandpiper	Inland waters	Multiple main habitat types
Griffon Vulture	Not defined	Global variables
Icterine Warbler	Multiple main habitat types	Forest
Great Crested Grebe	Inland waters	Inland waters
Common Rosefinch	Multiple main habitat types	Global variables
Garganey	Inland waters	Inland waters
Middle Spotted Woodpecker	Forest	Forest
Western Orphean Warbler	Special locations	Forest
Reed Bunting	Inland waters	Multiple main habitat types
Ruddy Shelduck	Open land	Multiple main habitat types
European Stonechat	Multiple main habitat types	Multiple main habitat types
Black Kite	Inland waters	Multiple main habitat types
White-tailed Eagle	Inland waters	Open land
Whooper Swan	Inland waters	Open land
Eurasian Pygmy Owl	Forest	Forest
Northern Wheatear	Special locations	Special locations
White-throated Dipper	Inland waters	Inland waters
Meadow Pipit	Multiple main habitat types	Open land
Citril Finch	Alpine regions	Special locations
Eurasian Scops Owl	Open land	Forest

Error Type Analysis

Each manipulated sighting in `validata` was labeled with its error type. Predictions were grouped by error type for each filter and the best model. The F1 scores were calculated for each error type:

- **Date Errors**: Can be detected by Grid-Year Filter and all ML-Models.

- **Distribution Errors**: Can be identified via Grid-Year Filter and spatial trained ML-Models.
- **Habitat Errors**: Can be captured by the Land Cover Filter and all ML-Models which were trained with spatial features.

By analyzing performance across error types, it can be determined which filters or models are specialized and which are generalizable.

5.3.5 Summary

This quantitative evaluation establishes a structured framework to measure the effectiveness of each method and detect patterns across species and error types. Results of this evaluation will be presented and discussed in the Chapter 6 and 7 and are used to validate the system's accuracy, limitations, and ecological plausibility.

5.4 Qualitative Evaluation

5.4.1 Objective

The qualitative evaluation aims to assess the usability, interpretability, and ecological plausibility of the Outlier Detection system from a domain-expert perspective. While quantitative metrics reveal how well models and filters perform in terms of accuracy, they do not capture how these tools are perceived by practitioners in real-world applications. By incorporating the judgments and feedback of professional ornithologists, this section seeks to understand whether the outputs are trusted, actionable, and supportive of expert workflows. Additionally, the qualitative review highlights strengths and limitations in the current system design that are not evident through statistical evaluation alone. This expert feedback serves as a foundation for refining both the detection methods and the human–AI interface in future iterations.

5.4.2 Review Design

To complement the quantitative metrics with expert insight, a small-scale qualitative review was conducted. Two professional ornithologists, hereafter referred to as $\mathbf{R_1}$ and $\mathbf{R_2}$, interacted with the web-based review interface described in Chapter 4 to assess the plausibility of automatically flagged bird sighting records.

Sampling Strategy

For each reviewer, a stratified sample of **150** entries from the 2023 `validata` dataset was generated. The sampled records represented the same 27 species used in the training phase (Table 5.1), ensuring continuity and ecological relevance. Each reviewer received:

- **75 manipulated entries**, evenly distributed across three error types:
  - 25 *Date* anomalies (seasonal misplacements),
  - 25 *Distribution* anomalies (coordinate and altitude shifts),
  - 25 *Habitat* anomalies (inconsistent land cover).
- **75 unedited entries** serving as plausible controls.

This yielded a balanced design per reviewer, half anomalous, half control, and a total corpus of **300 reviewed sightings** (2×150), with **150 true anomalies** and **150 true negatives**.

Species selection followed the same constraints as the quantitative evaluation (Section 5), covering all 27 species in approximately equal proportions across both manipulated and control samples.

Annotation Workflow

Each entry was presented via a dedicated web interface (described in detail in Chapter 4), which included species metadata, an interactive map, and the outputs of all filters and the model. Reviewers assessed the plausibility of each sighting and provided binary judgments alongside explanatory labels.

Contrary to standard blind-review protocols, the outputs of the filters and the model were visible *before* the reviewer made their decision. This design choice reflects a realistic application context where expert judgment is intended to complement, not independently verify, model output. While this introduces the possibility of confirmation bias, it enables a more practical evaluation of interpretability and support value.

Logging and Storage

Each reviewer's actions were logged automatically into a CSV file, capturing:

- Reviewer ID and country
- Timestamp
- Entry ID
- Final judgment (Outlier / Not Outlier)
- Selected reason(s), if applicable

This dataset forms the ground-truth label set used for measuring reviewer-model agreement and supports qualitative interpretation in Chapter 6.

Post-review Feedback

At the end of each review session, a short open-form questionnaire was displayed, asking:

1. Did any filter or model stand out-positively or negatively?
2. Did the automated predictions influence your decision-making?
3. Did you find the predictions helpful?
4. Which do you prefer: the detailed filter outputs or the binary DBSCAN result?
5. Do you have any other comments or suggestions?

Those questions were chosen because they can capture the current state of the filters and models but also gather ideas on how to improve this approach. Answers were collected via email and coded thematically. These insights will be discussed in the qualitative results section to evaluate usability, trustworthiness, and potential improvements.

Limitations

This setup reflects a trade-off between experimental control and ecological validity. The simultaneous display of predictions may have shaped expert responses, but also simulates real-world use cases where decision support tools aim to inform rather than replace human judgment. While the sample size is limited, both reviewers were experienced ornithologists deeply familiar with the system, providing high-quality feedback for initial validation.

6 Results

6.1 Quantitative Evaluation

6.1.1 Performance Analysis

Emergent Filters

Figure 6.1 visualizes the F1 score distributions of the three Emergent Filters-Grid-Year, Land Cover, and Altitude across all 27 bird species. The Grid-Year filter yielded the highest average F1 score with the smallest interquartile range. The Altitude filter followed closely, with moderately higher variance and a slightly lower mean. The Land Cover filter exhibited the broadest spread and the lowest median performance, although outliers indicate that it performed very well for a few species.

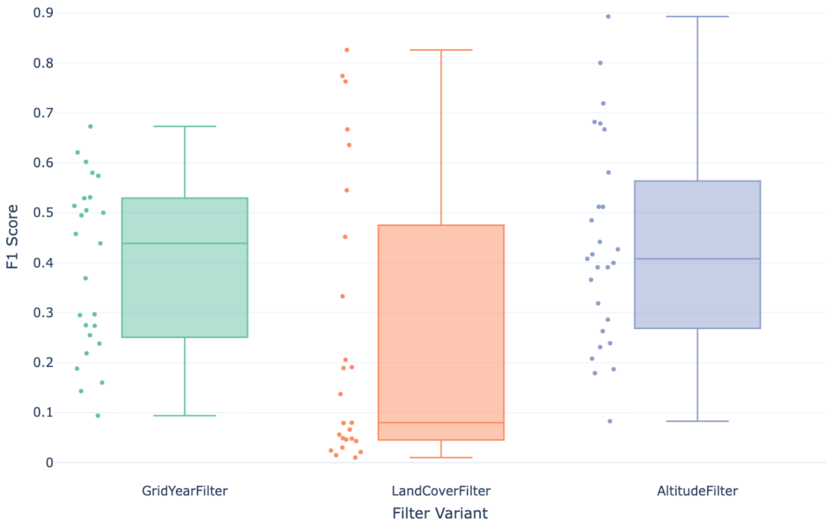


Figure 6.1: F1 score distribution for each Emergent Filter across 27 species.

Table 6.1 provides the exact threshold and best F1 score obtained for each species and each filter. The threshold yielding the best performance varied widely. For the Grid-Year filter, most species achieved optimal results at the lowest tested threshold (1e-06), with only a few preferring higher settings (0.001 or 0.025). For the Altitude filter, optimal thresholds clustered primarily around 0.0005 and 0.005. The Land Cover filter's optimal thresholds were more dispersed across the range from 0.001 to 1.1.

Species	**Grid-Year Filter**	**Land Cover Filter**	**Altitude Filter**
Rock Ptarmigan	0.495 (0.025)	0.333 (0.75)	0.800 (0.005)
Western Capercaillie	0.505 (1e-06)	0.636 (0.75)	0.485 (0.05)
Greater Scaup	0.219 (1e-06)	0.137 (0.6)	0.400 (0.005)
Twite	0.000 (1e-06)	0.046 (0.25)	0.083 (0.005)
Water Pipit	0.143 (0.025)	0.056 (1.05)	0.231 (0.0005)
Whinchat	0.500 (1e-06)	0.048 (1.1)	0.239 (0.0005)
Three-toed Woodpecker	0.673 (1e-06)	0.826 (0.75)	0.667 (0.005)
Common Sandpiper	0.275 (1e-06)	0.024 (1.0)	0.208 (0.0001)
Griffon Vulture	0.439 (0.025)	0.000 (0.001)	0.187 (0.05)
Icterine Warbler	0.602 (1e-06)	0.080 (1.1)	0.682 (0.0005)
Great Crested Grebe	0.188 (1e-06)	0.049 (1.0)	0.319 (0.0001)
Common Rosefinch	0.531 (1e-06)	0.206 (0.75)	0.427 (0.0005)
Garganey	0.369 (0.001)	0.030 (1.05)	0.512 (0.0005)
Middle Spotted Woodpecker	0.255 (1e-06)	0.189 (1.0)	0.263 (0.005)
Western Orphean Warbler	0.621 (1e-06)	0.452 (0.75)	0.366 (0.005)
Reed Bunting	0.295 (1e-06)	0.043 (1.0)	0.408 (0.001)
Ruddy Shelduck	0.274 (1e-06)	0.066 (1.0)	0.442 (0.001)
Black Kite	0.574 (1e-06)	0.000 (0.001)	0.893 (0.0001)
White-tailed Eagle	0.458 (1e-06)	0.015 (1.05)	0.512 (0.0005)
Whooper Swan	0.297 (1e-06)	0.010 (0.75)	0.391 (0.005)
Eurasian Pygmy Owl	0.094 (0.025)	0.763 (0.75)	0.179 (0.005)
Northern Wheatear	0.580 (1e-06)	0.021 (1.05)	0.391 (0.0001)
White-throated Dipper	0.000 (1e-06)	0.191 (1.0)	0.417 (0.001)
Meadow Pipit	0.160 (0.001)	0.079 (1.0)	0.286 (0.0005)
Citril Finch	0.529 (1e-06)	0.667 (0.75)	0.719 (0.005)
Eurasian Scops Owl	0.514 (0.025)	0.774 (0.75)	0.581 (0.005)
European Stonechat	0.238 (1e-06)	0.545 (1.0)	0.679 (0.0001)

Table 6.1: Best F1 scores and thresholds for each Emergent Filter per species. Format: `F1 (threshold)`

Machine Learning Models

Comparison of Feature Groups Figure 6.2 shows the average F1 score for all seven models - AutoEncoder, DBSCAN, HDBSCAN, Isolation Forest, KNN, LOF, and iNNE - under five feature variants: full feature set ("All"), exclusion of altitude, coordinates, date, and land cover.

KNN using the full feature set achieved the highest overall performance with an average F1 score of 0.65. The second-best configuration for KNN was the exclusion of land cover, which yielded nearly identical scores. Similarly, iNNE reached competitive performance with both the full feature set and the no land cover variant.

AutoEncoder, LOF, and DBSCAN performed best using all features, with slight decreases when excluding land cover. Isolation Forest showed a notable increase in performance when land cover was removed. In contrast, HDBSCAN underperformed across all feature configurations.

Across all models, excluding the date feature consistently led to the lowest average scores (except for HDBSCAN), followed by the exclusion of altitude, and then coordinates. This trend was consistent and observed in AutoEncoder, DBSCAN, Isolation Forest, KNN, LOF, and iNNE.

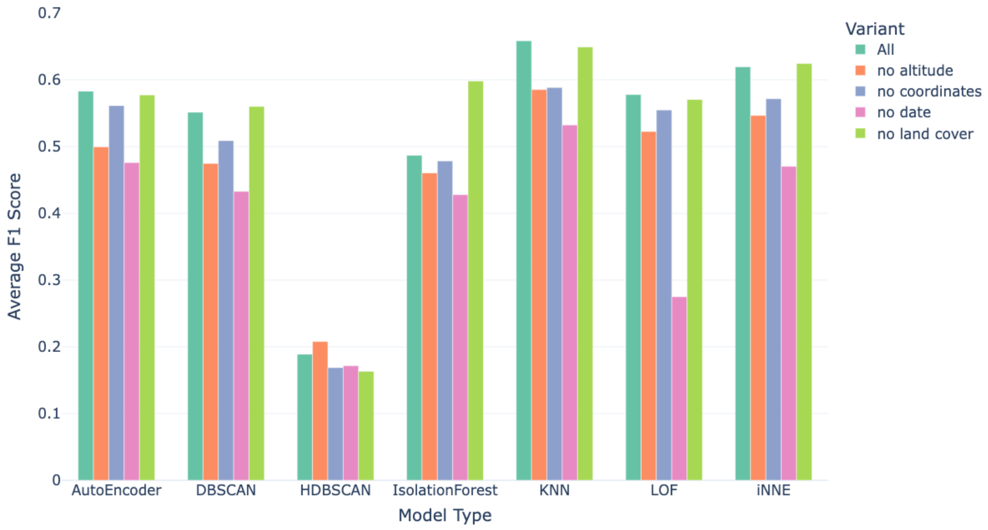


Figure 6.2: Average F1 score by model and feature removal variant.

Comparison of Algorithms Figure 6.3 displays the distribution of F1 scores for the best-performing variant of each model. The boxplot includes individual data points per species.

KNN with all features yielded the best overall distribution, exhibiting the highest median F1 score and the narrowest spread across species. iNNE and AutoEncoder followed closely behind, both showing high median performance and moderate variance. Isolation Forest and LOF demonstrated good average scores but greater variability, with LOF showing a wide interquartile range. DBSCAN showed a tighter spread but slightly lower median. HDBSCAN had the lowest scores overall, with the majority of species clustering around F1 scores below 0.3.

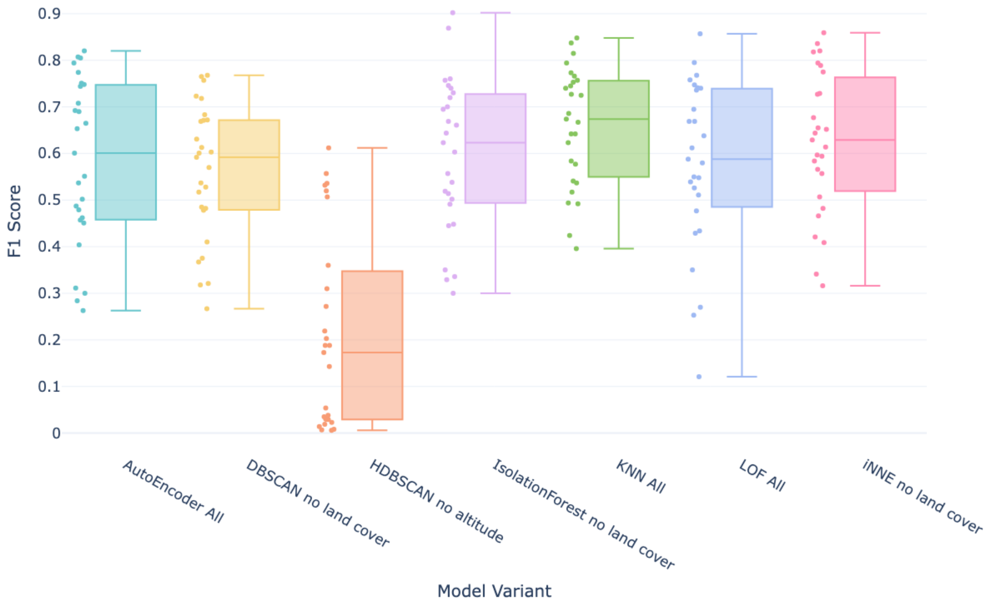


Figure 6.3: Best configuration per model: F1 score distributions across all species.

Table 6.2 summarizes the exact hyperparameters, thresholds, and F1 scores for the best-performing KNN model per species, using the full feature set. The number of neighbors and the decision thresholds varied considerably across species.

Species	Neighbors	Threshold	F1 Score
Rock Ptarmigan	15	0.794	0.794
Western Capercaillie	3	0.566	0.815
Greater Scaup	10	0.721	0.517
Twite	2	0.831	0.773
Water Pipit	7	0.656	0.642
Whinchat	4	0.521	0.577
Three-toed Woodpecker	6	0.592	0.837
Common Sandpiper	8	0.588	0.537
Griffon Vulture	6	0.629	0.492
Icterine Warbler	8	0.703	0.848
Great Crested Grebe	10	0.558	0.424
Common Rosefinch	7	0.632	0.667
Garganey	20	0.618	0.642
Middle Spotted Woodpecker	3	0.433	0.396
Western Orphean Warbler	9	0.541	0.740
Reed Bunting	3	0.597	0.674
Ruddy Shelduck	2	0.546	0.584
Black Kite	8	0.642	0.745
White-tailed Eagle	2	0.635	0.686
Whooper Swan	4	0.605	0.623
Eurasian Pygmy Owl	4	0.620	0.753
Northern Wheatear	2	0.491	0.725
White-throated Dipper	4	0.644	0.541
Meadow Pipit	8	0.610	0.494
Citril Finch	3	0.691	0.727
Eurasian Scops Owl	4	0.793	0.766
European Stonechat	2	0.509	0.757

Table 6.2: Best parameter settings and F1 scores for KNN model (all features) per species.

Emergent Filters vs. Machine Learning Models

Figure 6.4 compares the KNN model (all features) against the three Emergent Filters. KNN scored consistently higher across all species and demonstrated a smaller variance in F1 values. The Grid-Year and Altitude filters showed a broader distribution, while Land Cover had the lowest F1 values across most species.

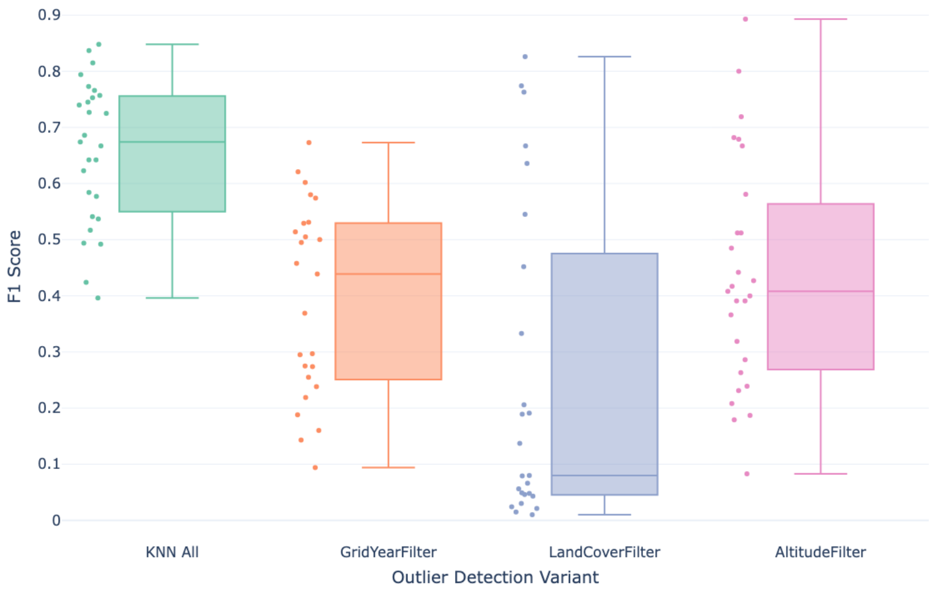


Figure 6.4: F1 scores of the best model (KNN) versus Emergent Filters.

6.1.2 Pattern Analysis

Species Performance Patterns

To evaluate model performance across different ecological traits, species-level F1 scores from the best-performing model configuration (KNN using all features) were grouped and analyzed along five classification axes: migration distance, migration behavior, breeding habitat, feeding habitat during breeding season, and feeding habitat during winter. Each group analysis includes 27 species.

Migration Distance Figure 6.5 shows the distribution of F1 scores grouped by migration distance. Long-distance migrants achieved relatively high F1 scores with moderate dispersion. Medium-distance migrants displayed consistently strong F1 values across only two species. Non-migratory species showed a broad range with a high upper bound. Short-distance migrants had the lowest median F1 scores among the defined categories.

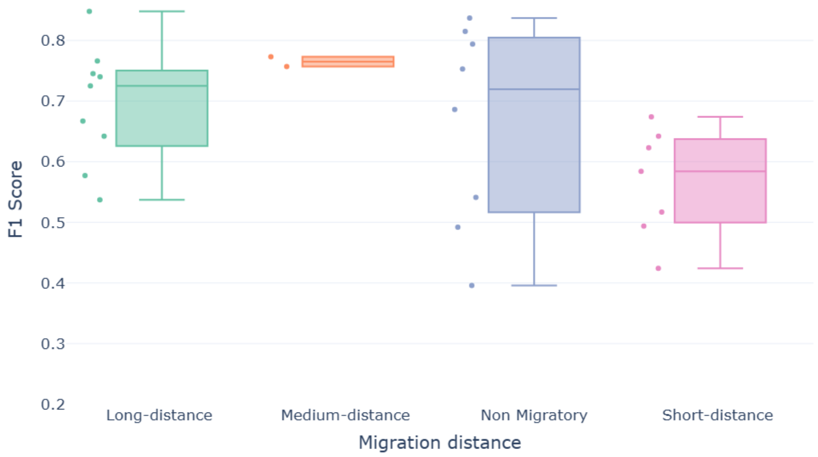


Figure 6.5: F1 scores grouped by migration distance.

Migration Behavior As shown in Figure 6.6, non-migratory species showed the highest median F1 scores although the widest variability. Fully migratory species followed closely in performance with slightly lower dispersion. Partial migrants had the lowest median values and displayed narrow distribution of F1 scores within a lower range.

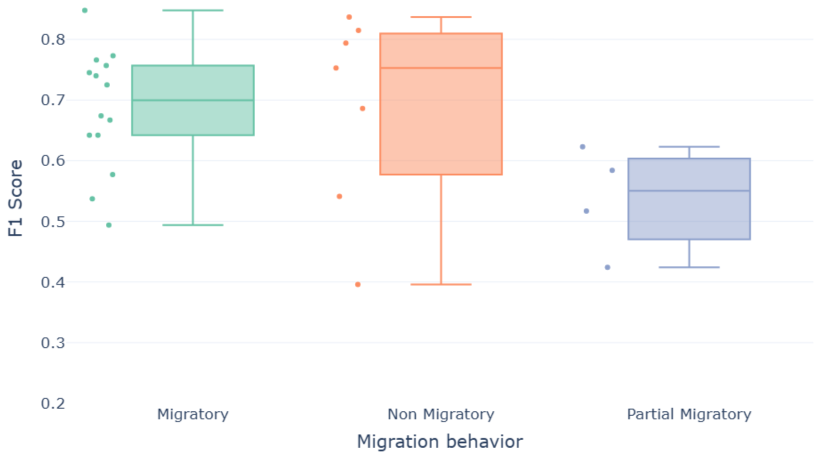


Figure 6.6: F1 scores grouped by migration behavior.

Breeding Habitat In Figure 6.7 the F1 score distributions are grouped by breeding habitat type. Species classified under “Alpine regions” achieved the highest overall scores and low variance. “Special locations” ranked second, with consistently high and stable F1 values, but with only a small sample size. “Multiple main habitat types” has a very comparable median score but more wide spread results than “Special locations”. Birds associated with forest habitats also showed broader variation, while inland water breeders had the lowest median F1 scores. Species from the “Not defined” category showed intermediate scores.

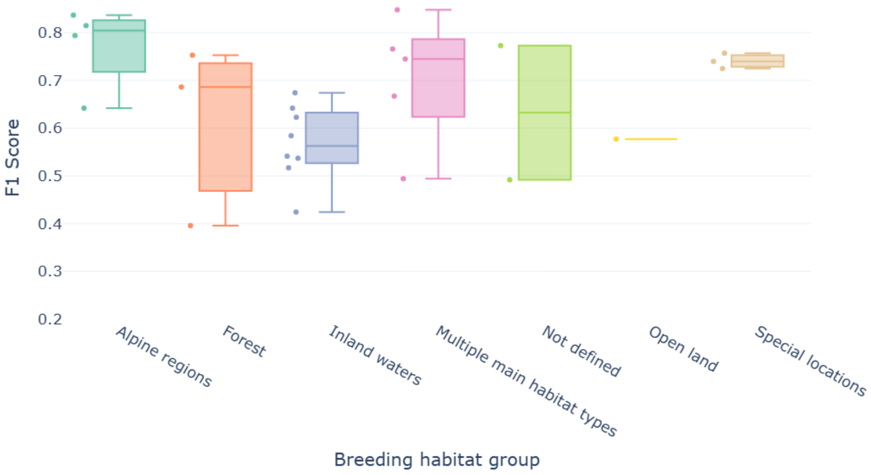


Figure 6.7: F1 scores grouped by breeding habitat.

Feeding Habitat During Breeding Season Figure 6.8 displays the performance grouped by feeding habitat used during the breeding season. Species feeding in “Alpine regions” showed the highest F1 values. “Special locations” followed closely with narrow dispersion. “Forest” and “Multiple main habit types” groups showed the most dispersed scores. Birds associated with “Open land”, “Not defined”, “Inland waters” and “Forest” categories had lower median values.

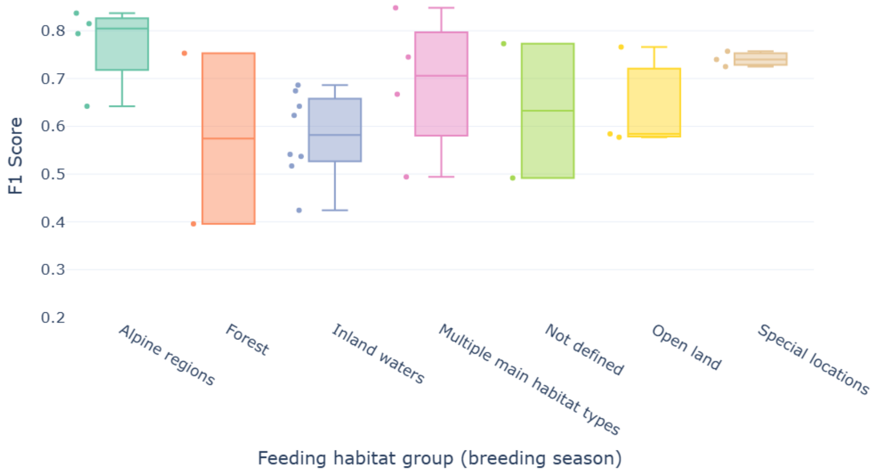


Figure 6.8: F1 scores grouped by feeding habitat during breeding season.

Feeding Habitat During Winter Figure 6.9 groups species by their feeding habitats during the winter season. The “Alpine regions” group maintained high scores with limited variability. The “Special locations” and “Forest” groups also showed elevated and stable scores. “Coast and sea” showed the broadest distributions. Birds grouped under “Global variables”, “Inland waters”, and “Open land” displayed wider variance and generally lower scores. “Multiple main habitat types” showed moderate performance.

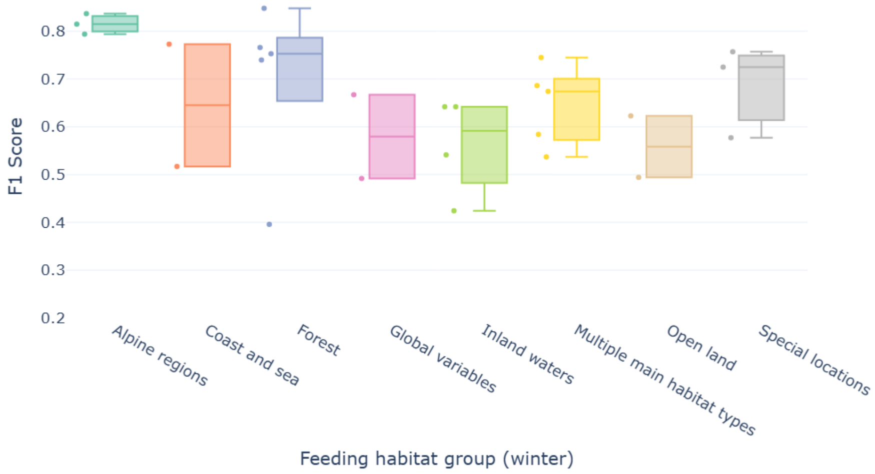


Figure 6.9: F1 scores grouped by feeding habitat during winter.

Error Type Performance

Figure 6.10 displays F1 scores by method and error type (habitat, spatial distribution, and date). The KNN model outperformed all Emergent Filters in every category. The Altitude filter achieved the second-best results for habitat and spatial errors. Grid-Year was third in performance across all types. The Land Cover filter yielded the lowest performance in all categories.

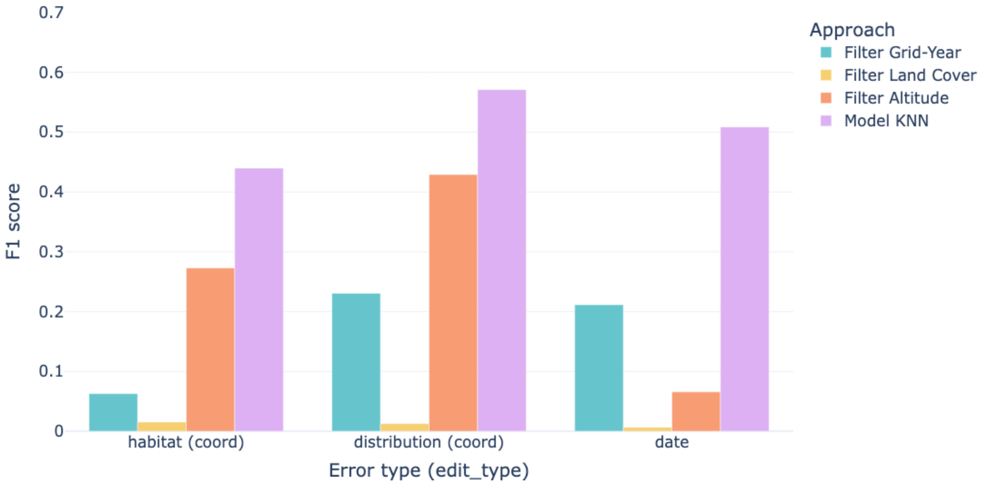


Figure 6.10: F1 scores by error type and method.

6.2 Qualitative Evaluation

6.2.1 Reviewer Participation

Two expert ornithologists, $\mathbf{R_1}$ and $\mathbf{R_2}$, participated in a structured qualitative review using the Gradio interface described in Chapter 4. Each reviewed a stratified sample of **150 records**, consisting of **75 true anomalies** (25 for each error type: date, distribution, and habitat) and **75 unaltered control records**. In total, **300 annotated cases** were produced.

6.2.2 Overall Impressions

Both reviewers provided generally positive feedback regarding the Outlier Detection tool. $\mathbf{R_1}$ expressed enthusiasm, stating they were *"pretty excited—the tool already flags many cases we would probably miss."* $\mathbf{R_2}$ highlighted the tool's potential in operational contexts, describing it as *"useful"* and *"time-saving."*

6.2.3 Per-Model Feedback

Emergent Land-Cover Filter This filter was praised for its ability to flag clear mismatches in forest versus open land habitats. However, both reviewers criticized the fixed 1 km buffer, noting that it was either too narrow (for wide-ranging species like *Black Kite*) or too coarse (for linear habitat specialists like *White-throated Dipper*). They recommended adapting the buffer size to the ecological profile of each species.

Altitude Filter Reviewers appreciated the filter's concept, but noted that many false positives stemmed from sighting imprecision. Since bird locations were approximated based on observer coordinates, birds spotted from a distance might appear in an ecologically implausible altitude zone and be therefore flagged. Hierarchical or more flexible spatial windows were suggested as improvements.

Grid-Year (Date/Location) Filter No major weaknesses were reported. The filter worked reliably for seasonal outliers, particularly when date and location combinations clearly contradicted known species behavior. Some anomalies (e.g. *Water Pipit* in summer) required context from additional cues like habitat.

Unsupervised Models Both experts singled out **DBSCAN** as particularly effective. It provided clear binary output while still aligning with their intuition in many cases.

6.2.4 Influence and Helpfulness of Predictions

- $\mathbf{R_1}$ consciously attempted to decide *before* reviewing model outputs and noted that *"the models and I often agreed."*

- $\mathbf{R_2}$ acknowledged that predictions *"partly influenced"* their final decisions.

Both reviewers agreed that the automated predictions were helpful and enhanced their confidence in labeling decisions.

6.2.5 Preference for Output Format

$\mathbf{R_1}$ valued the multi-dimensional input from the filters and model: *"If all filters and the models said the same, it was always an error."* $\mathbf{R_2}$ liked the transparency of the filter breakdowns but also appreciated the simplicity of DBSCAN's binary flag. A hybrid output—a binary alert with expandable detailed cues—was considered optimal by both reviewers.

6.2.6 User Interface Suggestions

1. **Spatial context.** Incorporate alternative basemaps such as SwissTopo with aerial imagery to assist with terrain interpretation. Show sighting accuracy (e.g., exact point vs. 1 km grid cell).
2. **Interface layout.** Minimize vertical scrolling by condensing information onto a single screen; a smaller, zoomable map is acceptable.
3. **Progress tracking.** $\mathbf{R_2}$ suggested adding a counter or progress bar to track remaining records.

6.2.7 Redefining "True Outliers"

$\mathbf{R_1}$ proposed expanding the gold standard definition to include both the synthetic errors and additional records that reviewers manually labeled as erroneous. This revised definition may serve as the foundation for future precision-recall evaluations.

6.2.8 Implications for Future Work

The qualitative feedback underscores that the current limitations are not conceptual but configurational. Key areas for enhancement include:

1. species-specific buffer sizes for land cover,
2. adaptive altitude windows to address spatial imprecision,
3. a redesigned user interface with improved usability and interpretability.

In conclusion, the qualitative evaluation supports the results of the quantitative assessment: automated Outlier Detection, when transparently communicated and ecologically contextualized, is a valuable addition to bird data validation workflows.

7 Discussion

RQ 1 - Comparative Model Performance

To what extent do statistical (Emergent Filter) models and Machine Learning models differ in their accuracy for detecting user-generated errors, as measured by the F1-score?

This research question evaluates the relative effectiveness of Emergent Filters and unsupervised Machine Learning (ML) models in detecting user-generated anomalies in bird sighting records. The comparison is centered on the F1-score, which provides a balanced metric combining both precision and recall. Quantitative results demonstrate substantial performance differences between the two model families, and qualitative expert feedback further contextualizes these findings.

Quantitative Comparison of Overall Performance

The results from the quantitative evaluation clearly indicate that ML models outperform Emergent Filters across all tested species and error types. As shown in Figure 6.4, the **K-Nearest Neighbors (KNN)** model using the full feature set yielded the highest average F1-score across all species, with a mean performance of 0.65 and minimal inter-species variance. In contrast, the best-performing Emergent Filter (Grid-Year) achieved substantially lower F1-scores, with a mean below 0.5 and a wider performance spread.

Further evidence is provided in Table 6.1, where most filters achieved only moderate F1-scores for individual species. Although certain filters such as the Altitude filter performed well for selected mountain and forest species (e.g., *Rock Ptarmigan*, *Boreal Owl*), these performances were isolated exceptions rather than indicative of consistent accuracy across species. Meanwhile, models such as **iNNE** and **AutoEncoder** exhibited robust results across diverse ecological contexts.

The superiority of ML models is further reinforced by Figure 6.3, where the F1-score distributions of the best variant of each model are shown. All ML models-with the exception of HDBSCAN-outperformed the Grid-Year, Altitude, and Land Cover filters. The statistical filters also displayed greater interquartile ranges and more frequent low outliers, highlighting their limited generalizability.

Qualitative Validation by Expert Reviewers

Qualitative evidence from expert ornithologists corroborates the numerical findings. Both reviewers, R_1 and R_2, described the automated predictions-particularly those of the ML models-as *helpful* and *time-saving*. R_1 remarked that *if all filters and the model agreed, it was always an error*, indicating that model agreement reinforced expert trust. Importantly, both reviewers highlighted the Machine Learning approach as particularly intuitive and valuable.

However, the filters were not dismissed outright. The Grid-Year filter was described as *reliable* for seasonal outliers and was seen as ecologically transparent. Reviewers appreciated the breakdown of filter-specific predictions, which helped them understand the rationale behind a flagged outlier.

Interpretation and Implications

The performance disparity between ML models and Emergent Filters can be attributed to several factors. Emergent Filters rely on static, species-specific lookup tables and plausibility thresholds. While biologically grounded, these filters are constrained by rigid assumptions (e.g., circular smoothing windows, fixed habitat buffers) and cannot easily adapt to subtle multidimensional outliers or species with irregular distributions. In contrast, ML models leverage the full joint feature space, capturing complex interactions between location, date, altitude, and land cover, thereby enabling superior generalization across varied species and habitats.

Despite this, filters offer higher transparency and interpretability. Their ecological specificity may make them more suitable for narrow use cases, such as preliminary filtering or use in regions with limited data availability. In operational settings, a hybrid approach-combining the flexibility of ML with the interpretability of filters-may offer the best trade-off.

Conclusion

In conclusion, ML models-particularly KNN, iNNE, and AutoEncoder-consistently outperformed statistical Emergent Filters in detecting manipulated bird sighting records. Their superiority was evident across overall accuracy metrics, interspecies robustness, and expert evaluations. However, the interpretability and domain specificity of Emergent Filters still provide valuable complementary insights. A combined workflow may offer optimal detection performance and user trust in applied biodiversity monitoring systems.

RQ 2 - Error-type-specific Performance

How does the detection performance of the investigated model approaches vary across the predefined error types (e.g., date errors, land-cover errors, altitude errors)?

This research question explores whether the detection accuracy of Emergent Filters and Machine Learning (ML) models varies depending on the specific type of artificially injected error. The three error types considered in this study were: **date errors**, **distribution errors** (spatial shifts in coordinates and altitude), and **habitat errors** (ecologically implausible land cover conditions). Each method's ability to capture these distinct error types was evaluated using F1-scores computed on stratified subsets of the `validata` dataset.

Quantitative Analysis of Error-type Detection

Figure 6.10 presents a breakdown of F1-scores across all models and filters by error type. The results show a clear and consistent pattern: the KNN model (with all features) achieved the highest F1 scores across all three error categories, confirming its robustness and generalizability. It was able to detect date errors, habitat errors, and distribution errors with high accuracy and low variance, making it the most reliable model for broad-spectrum error detection.

Among the Emergent Filters, performance was more heterogeneous and specialized:

- **Date Errors:** The Grid-Year filter, which directly encodes seasonal and spatial plausibility via lookup tables, showed the strongest performance among the Emergent Filters in this category. This confirms that such biologically-informed temporal filters are effective in capturing seasonal implausibilities.
- **Habitat Errors:** The Altitude filter outperformed the Land Cover filter and Grid-Year in identifying habitat-related anomalies. Although the Land Cover filter is designed for this purpose, its effectiveness was limited by rigid buffer configurations and high interspecies variability, as observed in Table 6.1.
- **Distribution Errors:** These were best detected by the Altitude filter among the statistical approaches. Grid-Year provided moderate performance, while the Land Cover filter underperformed significantly in all error categories.

Despite their ecological basis, none of the filters matched the F1-scores of the best ML model (KNN) for any error type. The performance gap was especially pronounced for habitat errors, where KNN nearly doubled the F1-score of the Altitude filter.

Qualitative Evaluation by Reviewers

Reviewers' responses provide nuanced insights into model effectiveness per error type. Both ornithologists confirmed that the Machine Learning model predictions often aligned with their own judgments across all categories. For date-related anomalies, R_1 explicitly noted that the Grid-Year filter *worked reliably*, especially when sightings contradicted known seasonal ranges.

In contrast, habitat-related errors presented more challenges for both filters and experts. R_2 remarked that false positives were common when birds were observed from a distance-making land cover and altitude misrepresentative of the actual habitat. R_1 criticized the fixed 1 km buffer for land cover calculations, stating that the filter's effectiveness *strongly depends on the species' ecology and mobility.*

Distribution anomalies were often successfully flagged by both filters and ML models. The Altitude filter was frequently cited as informative, though R_2 warned about imprecision due to the coordinate granularity of sightings. These qualitative observations align with the quantitative data and reinforce the importance of ecological nuance and spatial resolution in filter design.

Interpretation and Implications

The detection capabilities of different methods clearly vary by error type. Emergent Filters are constrained by their singular focus on one ecological axis-seasonality, altitude, or habitat-whereas ML models consider multidimensional patterns. This enables them to detect more subtle or overlapping anomalies. However, although not as successful as the ML model overall, the success of filters for specific error types (e.g., Grid-Year for date, Altitude for spatial shifts) demonstrates their utility in targeted contexts.

For applications where error types are known or can be predicted in advance, a hybrid strategy may prove advantageous: presenting outputs from both Emergent Filters and ML models can offer ornithologists multiple perspectives, enabling them to form a more informed judgment based on diverse model outputs.

Conclusion

Detection performance varies significantly by error type. The KNN model exhibited the best overall performance across all categories, confirming its suitability as a general-purpose detector. Emergent Filters demonstrated strengths in specific areas-particularly date and altitude plausibility-but were limited in scope and flexibility. Both quantitative and qualitative results support the conclusion that ML models, especially those using comprehensive feature sets, provide a more robust solution for diverse error detection in ornithological datasets.

RQ 3 - Species- or Guild-specific Performance

To what degree does model performance differ among avian species or ecological guilds?

This research question explores whether systematic differences exist in the ability of the Outlier Detection models and filters to identify erroneous sightings across different bird species and ecological groups. Understanding such variation is essential for designing model ensembles or filter strategies that generalize across avian taxa with differing ecological traits and movement patterns.

Quantitative Differences Across Species

Figure 6.3 and Table 6.2 reveal substantial inter-species differences in F1 scores, even when using the best-performing model (KNN with all features). For example, species such as *Icterine Warbler* (F1 = 0.85), *Three-toed Woodpecker* (F1 = 0.84), and *Western Capercaillie* (F1 = 0.82) achieved high detection scores, whereas *Great Crested Grebe* (F1 = 0.42) and *Middle Spotted Woodpecker* (F1 = 0.40) scored substantially lower. These differences likely reflect the ecological specificity and distributional consistency of each species, although interpretation of the causes is reserved for the next chapter.

Even the Emergent Filters (Table 6.1) showed uneven results across species. For instance, the Land Cover filter yielded an F1 score of 0.83 for *Three-toed Woodpecker* but only 0.02 for *Common Sandpiper*. Similar variability was found for the Altitude and Grid-Year filters. This emphasizes the importance of training models for each species individually and adjusting the parameters accordingly.

Species Grouping Patterns

To understand how ecological traits influence Anomaly Detection performance, species-level F1 scores from the best-performing model (KNN with all features) were grouped and analyzed along five groups. These included migration behavior, migration distance, breeding habitat, feeding habitat during the breeding season, and feeding habitat during winter. The analysis provides insights into which ecological traits may support or hinder detection. In this the small sample size of 27 species warrants caution in generalizing the results.

Species breeding or feeding in **alpine regions** consistently achieved the highest F1 scores across all habitat-related groupings. Their performance was not only strong but also stable, suggesting that the spatial and temporal consistency of alpine-associated sightings makes them more easily distinguishable from outliers. Similarly, species linked to **special locations**-such as gravel pits, rocky outcrops, or ruderal sites-performed surprisingly well across groupings. Their distinct spatial and isolated occurrence patterns may make them particularly useful for Anomaly Detection. However, it should be noted that the sample size for these classes is quite small, so caution should be exercised when generalizing these findings.

In contrast, species associated with **inland waters**, **open land**, or **undefined habitats** showed weaker and more variable performance. These habitat types may reflect broader or overlapping ranges, complicating the identification of outlier sightings.

Regarding migration traits, **non-migratory** and **long-distance migratory** species performed best. The strong performance of non-migratory birds likely stems from their restricted and stable spatial patterns. Long-distance migrants, often follow well-defined routes and timing, which could be making deviations easier to detect. **Short-distance migrants** and especially **partial migrants** yielded the lowest F1 scores. The variability in timing and behavior within these groups may result in blurred detection boundaries.

In summary, the grouping-based evaluation revealed that ecological consistency-whether in space (alpine breeders), behavior (non-migrants), or habitat specificity correlates with stronger Anomaly Detection performance. Conversely, flexible or poorly defined patterns in habitat use or migration behavior reduce model accuracy. While these trends are informative, they should be interpreted with care due to the limited number of species in the dataset.

Reviewer Observations on Species-specific Behavior

Reviewers also identified species-related variability in anomaly detectability. $\mathbf{R_1}$ and $\mathbf{R_2}$ both noted that certain wide-ranging or nomadic species, such as *Black Kite* or *Griffon Vulture*, generated higher rates of false positives in the filters due to their flexible habitat use and long-range movement. Conversely, sedentary species like *Middle Spotted Woodpecker* were easier to assess and classify.

Additionally, $\mathbf{R_2}$ emphasized that habitat-based filters were more effective for forest specialists than for birds inhabiting transitional or human-modified environments. Reviewers recommended tailoring filter behavior to species' ecological characteristics, which supports the idea of guild-aware modeling.

Interpretation and Implications

Both model- and filter-based detection performance is strongly species-dependent. Migratory status, habitat specificity, and ecological niche width all influence how easily anomalies can be detected. This highlights the need for adaptive strategies that consider species-specific data characteristics during both model training and threshold tuning.

These results also suggest that future implementations should allow for taxon-specific customization-such as adaptive buffer sizes for habitat features or flexible altitude bins-rather than applying uniform configurations across all species. In environments like citizen science platforms, where species differ widely in detectability and ecological behavior, such flexibility may be crucial for reliable Anomaly Detection.

Conclusion

Model performance varies considerably across species and ecological guilds. Species breeding or feeding in alpine regions or in structurally distinct habitats yielded the highest F1 scores. In contrast, species associated with inland waters, or partially migratory behavior performed generally less reliably. These findings highlight the importance of incorporating species ecology into model design and suggest that the most accurate Anomaly Detection frameworks will require adaptive or guild-specific configurations. The quantitative findings were in line with statements by the ornithologists.

RQ 4 - Influence of Features

Which features contribute significantly to the detection performance of the Machine Learning model?

This research question investigates the role of individual feature groups in enabling effective Anomaly Detection by Machine Learning models. The focus lies in identifying which spatial and ecological features-such as coordinates, altitude, land cover, or temporal attributes-are most critical for the models' ability to distinguish manipulated from valid records.

Quantitative Feature Ablation Results

To evaluate the contribution of each feature, models were trained and tested under five predefined feature configurations: (1) All Features, (2) No Land Cover, (3) No Altitude,

(4) No Coordinates, and (5) No Date. These ablations were applied uniformly across all models.

Figure 6.2 reveals consistent patterns across models. For nearly all algorithms, the highest average F1 scores were achieved when using the full feature set or when excluding only the land cover features. For example, KNN with all features yielded the best mean F1 score of 0.65, and removing land cover resulted in only a marginal drop in performance. iNNE showed a similar pattern, with very competitive performance in both the full and no-land-cover configurations.

AutoEncoder, LOF, and DBSCAN also benefited most from the full feature set. In these models, land cover appeared helpful but non-essential, as performance decreased only slightly when it was removed.

The Isolation Forest model, however, exhibited an exception: its performance improved noticeably when land cover was excluded. This may indicate sensitivity to noise introduced by high-dimensional habitat representations in tree-based models.

The removal of the date feature universally impaired performance across all models (with the exception of HDBSCAN, which performed poorly regardless of features). Date removal yielded the lowest scores, followed by exclusion of altitude and coordinates. These findings underscore the crucial role of temporal information in identifying seasonal anomalies.

Ranking of Feature Importance

Based on the aggregate patterns across models, the relative importance of feature groups can be ranked as follows:

1. **Date (Day-of-Year)**: Most important feature, with consistent performance drops when excluded. Encoded as cyclical sine/cosine features, date allowed models to detect seasonal mismatches.
2. **Altitude**: Second most important, particularly for distinguishing ecologically implausible elevations.
3. **Coordinates (X/Y)**: Provided spatial context essential for detecting geographical outliers.

4. **Land Cover**: Less critical than other features, but still provided useful habitat-based information for certain species.

These rankings hold across all models except HDBSCAN, which showed no clear benefit from any feature configuration, and Isolation Forest, which was adversely affected by land cover.

Reviewer Perspectives on Feature Utility

The qualitative evaluation provides further insight into the perceived utility of different feature types. Reviewer $\mathbf{R_1}$ particularly valued the Emergent Filter outputs based on date and altitude plausibility. These features appeared to match their own ecological intuitions and aligned with the known behavior of many target species.

For land cover, both reviewers expressed concerns about interpretability and resolution. The fixed $1\,\mathrm{km}^2$ buffer used for habitat composition was considered too coarse for linear habitat users (e.g., *White-throated Dipper*) and too narrow for wide-ranging species (e.g., *Black Kite*). This aligns with the modest quantitative contribution of land cover features observed in the model performance.

In contrast, the Grid-Year filter (which integrates date and location) was viewed as highly valuable and ecologically consistent, indirectly confirming the high importance of date-related features.

Conclusion

Temporal (date) features and spatial (altitude and coordinate) features are essential components for accurate Anomaly Detection. Their removal consistently degrades model performance. Land cover features provide additional ecological context but are less critical and may even hinder some models if not configured carefully. These results highlight the importance of thoughtful feature engineering and suggest that species-specific feature resolution-especially for land cover-may enhance performance further.

RQ 5 - Practitioner Preferences

How do professional ornithologists evaluate the usability and reliability of statistical Emergent Filters compared with Machine Learning models for error detection?

This research question explores the human-centered perspective: How do domain experts perceive the utility, reliability, and transparency of automated Anomaly Detection methods? Two professional ornithologists ($\mathbf{R_1}$ and $\mathbf{R_2}$) participated in a structured qualitative review of the model and filter outputs. Their feedback offers essential insight into which approaches align with expert expectations and practical needs.

General Perception of the System

Both reviewers evaluated the Anomaly Detection tool positively. $\mathbf{R_1}$ expressed enthusiasm, noting they were *"pretty excited-the tool already flags many cases we would probably miss"*. $\mathbf{R_2}$ viewed the tool as *"useful"* and *"time-saving"*, especially for pre-filtering large volumes of citizen science data.

These impressions confirm that automated systems can add real value in practice, particularly when they surface subtle errors that might otherwise go unnoticed.

Filter-Specific Feedback

Each Emergent Filter received distinct comments regarding its strengths and limitations:

Grid-Year Filter (Date + Location) This filter was widely appreciated for its ecological plausibility. Both reviewers confirmed that it reliably detected outliers in seasonal and spatial distributions. Anomalies like summer records for winter-only species (or vice versa) were flagged consistently, making this filter one of the most trusted components.

Altitude Filter The altitude filter was recognized as conceptually valid, but reviewers raised practical issues. Many false positives occurred when birds were observed from a distance, and their apparent coordinates (based on observer location) placed them in an implausible altitude zone. This spatial imprecision diminished trust in the filter's output. Both reviewers suggested adopting more flexible or hierarchical altitude windows to mitigate this issue.

Land Cover Filter The land cover filter generated the most criticism. While effective in some forest-vs-open land cases, its use of a fixed 1 km buffer was considered ecologically unrealistic for many species. For wide-ranging birds like *Black Kite*, the buffer was too narrow, while for habitat specialists like *White-throated Dipper*, it was too coarse. Both reviewers recommended species-specific tuning of the land cover resolution.

Model Output and Interpretability

Only the DBSCAN model was shown during the qualitative review to reduce reviewer workload and because it was the most readily available tuned model at the time.

$\mathbf{R_1}$ and $\mathbf{R_2}$ both found the models binary output simple and helpful. It provided an immediate impression of model confidence and aligned well with their own judgments in most cases. While the lack of explanation was noted as a potential limitation, the clarity of the output was appreciated.

Preference for Output Format

The reviewers articulated different but complementary preferences:

- $\mathbf{R_1}$ appreciated having multiple filters and a model side-by-side. When all flagged a sighting, the decision was clear.
- $\mathbf{R_2}$ favored the simplicity of binary alerts (as in the Model output) but still valued the transparency of the filters.

Both expressed interest in a hybrid interface: a simple alert (e.g., "Potential Outlier") supplemented with detailed breakdowns of the contributing filters or feature deviations. This structure would enhance trust while maintaining usability.

Influence of Model Predictions

R_1 made an effort to ignore the automated predictions initially, then compare their own decisions afterward. They reported a high degree of agreement with the models and filters.

R_2 acknowledged that the predictions influenced their decisions but found this helpful rather than misleading.

Both concluded that the predictions added confidence and improved consistency in decision-making, especially for borderline cases.

Recommendations and Interface Improvements

The reviewers proposed several specific improvements:

- Add aerial basemaps (e.g., SwissTopo) for improved terrain interpretation.
- Visualize sighting uncertainty (e.g., 1 km grid vs. exact point).
- Reduce vertical scrolling and condense the interface onto a single page.
- Include a progress bar or counter to show remaining records.

These suggestions focus on improving ecological context and usability without altering the underlying detection logic.

Redefining "True Outlier"

R_1 proposed broadening the ground-truth definition of outliers. In addition to the synthetically introduced errors, they suggested incorporating manually labeled cases flagged by experts as implausible. This expanded label set would reflect real-world plausibility more accurately than synthetic error types alone.

Conclusion

The qualitative evaluation demonstrates that professional ornithologists find both statistical filters and Machine Learning models useful, provided their outputs are interpretable and ecologically grounded. Among the filters, the Grid-Year filter earned the highest trust. The model (in this case DBSCAN) proved effective as a binary flagging tool. Hybrid systems, combining clear alerts with transparent breakdowns, are preferred. Usability concerns center around interface layout and ecological realism (e.g., buffer sizes). Integrating such expert feedback into future system iterations can enhance both acceptance and effectiveness in operational contexts.

8 Outlook

The results of this thesis demonstrate that automated Outlier Detection methods, particularly unsupervised Machine Learning models, can substantially enhance the identification of implausible records in large-scale bird sighting datasets. Both statistical filters and ML models showed value in different contexts, and expert reviewers found the system promising for practical use. Nonetheless, this work represents only an initial step toward integrating Anomaly Detection into the operational workflows of citizen science platforms such as *ornitho.ch* and *ornitho.de*. This chapter outlines several avenues for future research and development, aiming to further improve accuracy, usability, scalability, and real-world integration.

8.1 Integration into Operational Platforms

One of the most immediate next steps is the deployment of the developed system within an actual data review environment. Currently, the framework operates as an external tool, but integration into platforms like *ornitho.ch* would allow real-time application of filters and ML predictions. Reviewers could benefit from automated plausibility flags to prioritize their work, especially for the increasing number of daily reports. Integration would also facilitate the collection of expert decisions, enabling continuous improvement of the models through feedback loops.

Such integration must be accompanied by a robust UI/UX design process. As suggested by the reviewers, the existing Gradio-based interface could be enhanced with interactive visualizations, compact layouts, and context-aware basemaps. A web-based widget embedded in the reviewer portal could display ML and filter predictions alongside traditional sighting metadata.

8.2 Supervised Learning and Label Expansion

While this thesis focused on unsupervised approaches due to the lack of extensive labeled data, Supervised Learning remains a promising direction for future work. With a large enough dataset of confirmed false sightings - either manually annotated or derived from reviewer logs - supervised models could be trained to capture more subtle patterns and learn species-specific error distributions.

To support this, future initiatives should aim to systematically label historical outliers and true positives. Importantly, as proposed by Reviewer $\mathbf{R_1}$, expanding the gold standard to include expert-detected errors - not just synthetically manipulated records - would significantly improve label diversity and realism.

8.3 Species-specific Feature Engineering

Both reviewers noted that the effectiveness of filters - particularly those involving land cover - depends heavily on species-specific ecology. This was also confirmed by quantitative evaluation. Future versions of the framework should adapt features dynamically based on species traits. For instance:

- **Adaptive habitat buffers:** Instead of using a fixed 1 km buffer for all species, buffer sizes could reflect typical home range, foraging radius, or habitat breadth.
- **Sighting confidence estimation:** Differentiating between exact locations and sightings reported from a distance (e.g., over water bodies or valleys) would allow more accurate spatial plausibility checks. This could be inferred from metadata (e.g., optical devices used, observer comments) or modelled probabilistically.
- **Taxon-specific thresholds:** Rather than applying uniform decision thresholds across all species, models and filters could adapt sensitivity based on ecological variance or historical error rates.

This level of refinement requires additional species-level metadata (e.g., mobility, niche width), which could be extracted from established ecological databases or inferred via clustering techniques.

8.4 Scaling to Additional Species

The current implementation focuses on 27 ornithologist-selected species. To generalize the system for operational use, the coverage must be significantly expanded. A scalable training pipeline should be developed to automatically generate filters and train ML models for hundreds of species. This could be done incrementally, prioritizing taxa with high observation frequency or known reporting biases.

Particular attention should be given to rare species, whose low observation frequency may limit model generalization. Semi-supervised or transfer learning approaches could be investigated to support learning in data-scarce regimes.

8.5 Incorporating Reviewer Metadata and Trust Scores

Currently, all records are treated equally in terms of their assumed reliability. However, citizen science platforms often collect reviewer metadata-such as user ID, experience level, or historical accuracy. Future models could integrate this information as a feature or as a weighting factor:

- **Reviewer credibility scores:** If a reviewer has consistently produced high-quality data (e.g., few manual corrections required), their sightings could be down-weighted in Anomaly Detection or prioritized differently.
- **Model calibration:** These scores could also be used to calibrate the output probability of ML models, shifting thresholds based on expected trustworthiness.

This approach would allow more nuanced filtering strategies, enabling the system to adapt based on both the content and source of the sighting. However, for this approach, data privacy must be ensured consistently.

8.6 Feature Innovation and Temporal Dynamics

Additional features could further enhance detection capabilities. Potential extensions include:

- **Weather and climate features:** Integrating high-resolution climate data (e.g., temperature anomalies, wind conditions) could improve models' ability to distinguish between true shifts (e.g., early migration due to warm weather) and implausible sightings.
- **Observer history features:** Anonymized summaries of each observer's past reports (e.g., spatial range, frequency, taxonomic focus) may provide useful context for sighting plausibility.
- **Temporal change features:** Inspired by Siebold (2025), who used Bayesian Change Point Detection to analyze shifts in citizen science bird sighting data, future work could integrate temporal change point detection to inform model retraining intervals or flag sudden shifts that may indicate either ecological changes or observer behavior shifts.

Such features may also support longitudinal analysis of anomaly rates, offering insights into trends in observer behavior, habitat use, or climate-driven range shifts.

8.7 Human-Centered Design and Gamification

To maximize reviewer engagement and encourage consistent participation, gamification elements could be incorporated into the reviewing interface. Based on reviewer feedback, possible features include:

- **Leaderboards:** Displaying the number of reviewed entries per user could foster friendly competition.
- **Achievement badges:** Awarding badges for milestone completions (e.g., “100 sightings reviewed”) could increase motivation.
- **Progress tracking:** Showing progress bars or review summaries may help reviewers plan their workload and feel more rewarded.
- **Instant feedback:** When possible, providing post-review insights (e.g., "Your review matched the expert consensus") could build trust and provide educational value.

Such enhancements are particularly relevant given that manual data validation is often seen as tedious. Increasing enjoyment and transparency could boost review volume and consistency.

8.8 Toward Hybrid Decision Systems

Finally, this thesis supports the development of hybrid systems that integrate multiple detection layers: ML models, ecological filters, and expert judgment. The most promising architecture may involve:

- **Model ensembles:** Aggregating predictions from multiple ML algorithms (e.g., KNN, AutoEncoder, DBSCAN) can reduce variance and increase robustness.
- **Rule-based overrides:** Allowing critical ecological rules (e.g., no winter presence in alpine zones) to override or flag ML decisions.
- **Interactive dashboards:** Presenting both aggregated risk scores and individual filter/method outputs to support transparent decision-making.

Such systems can ensure high precision, preserve expert trust, and adapt to different reviewer profiles or taxonomic groups.

8.9 Conclusion

This thesis lays the groundwork for automated, interpretable, and ecologically grounded Anomaly Detection in bird sighting data. However, realizing its full potential requires significant next steps: operational integration, feature refinement, and species-specific customization. By extending the framework in these directions-and incorporating real-world usage feedback-future systems can substantially improve the reliability, scalability, and usability of biodiversity monitoring tools in the age of citizen science.

Bibliography

Backstrom, L. J., Callaghan, C. T., Worthington, H., Fuller, R. A. & Johnston, A. (2025), 'Estimating sampling biases in citizen science datasets', *Ibis* **167**(1), 73–87.

Baker, E., Drury, J., Judge, J., Roy, D., Smith, G. & Stephens, P. (2021), 'The verification of ecological citizen science data: Current approaches and future possibilities.'.

Bandaragoda, T., Ting, K., Albrecht, D., Liu, F. T. & Wells, J. (2014), Efficient anomaly detection by isolation using nearest neighbour ensemble, Vol. 2015.

Bonter, D. N. & Cooper, C. B. (2012), 'Data validation in citizen science: a case study from project feederwatch', *Frontiers in Ecology and the Environment* **10**(6), 305–307.

Bourgeois, Q., Kaptijn, E., Verschoof-van der Vaart, W. & Lambers, K. (2024), 'Assessing the quality of citizen science in archaeological remote sensing: results from the heritage quest project in the netherlands', *Antiquity* **98**(402), 1662–1678.

Breunig, M. M., Kriegel, H.-P., Ng, R. T. & Sander, J. (2000), Lof: Identifying density-based local outliers, *in* 'Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data', ACM, pp. 93–104.

Campbell, C., Barve, V., Belitz, M. W., Doby, J. R., White, E., Seltzer, C., Di Cecco, G., Hurlbert, A. H. & Guralnick, R. (2023), 'Identifying the identifiers: How inaturalist facilitates collaborative, research-relevant data generation and why it matters for biodiversity science', *BioScience* **73**(7), 533–541.

Campello, R. J., Moulavi, D. & Sander, J. (2013), Density-based clustering based on hierarchical density estimates, *in* 'Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)', Springer, pp. 160–172.

Cavadino, I., Port, G., Mill, A., Clover, G., Roy, H. & Jones, H. (2024), 'Slugs count: Assessing citizen scientist engagement and development, and the accuracy of their identifications', *People and Nature* **6**(5), 1822–1837.

Chandola, V., Banerjee, A. & Kumar, V. (2009), 'Anomaly detection: A survey', *ACM Computing Surveys (CSUR)* **41**(3), 1–58.

De Sherbinin, A., Bowser, A., Chuang, T.-R., Cooper, C., Danielsen, F., Edmunds, R., Elias, P., Faustman, E., Hultquist, C., Mondardini, R. et al. (2021), 'The critical importance of citizen science data', *Frontiers in Climate* **3**, 650760.

Di Febbraro, M., Bosso, L., Fasola, M., Santicchia, F., Aloise, G., Lioy, S., Tricarico, E., Ruggieri, L., Bovero, S., Mori, E. et al. (2023), 'Different facets of the same niche: Integrating citizen science and scientific survey data to predict biological invasion risk under multiple global change drivers', *Global Change Biology* **29**(19), 5509–5523.

Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. (1996), A density-based algorithm for discovering clusters in large spatial databases with noise, *in* 'Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD)', AAAI Press, pp. 226–231.

Figuerola-Ferrando, L., Linares, C., Zentner, Y., López-Sendino, P. & Garrabou, J. (2024), 'Marine citizen science and the conservation of mediterranean corals: the relevance of training, expert validation, and robust sampling protocols', *Environmental Management* **73**(3), 646–656.

Johnston, A., Matechou, E. & Dennis, E. B. (2023), 'Outstanding challenges and future directions for biodiversity monitoring using citizen science data', *Methods in Ecology and Evolution* **14**(1), 103–116.

Johnston, A., Moran, N., Musgrove, A., Fink, D. & Baillie, S. R. (2020), 'Estimating species distributions from spatially biased citizen science data', *Ecological Modelling* **422**, 108927.

Kelling, S., Johnston, A., Bonn, A., Fink, D., Ruiz-Gutierrez, V., Bonney, R., Fernandez, M., Hochachka, W. M., Julliard, R., Kraemer, R. & Guralnick, R. (2019), 'Using semistructured surveys to improve citizen science data for monitoring biodiversity', *BioScience* **69**(3), 170–179.

Kelling, S., Yu, J., Gerbracht, J. & Wong, W.-K. (2011), Emergent filters: Automated data verification in a large-scale citizen science project, *in* '2011 IEEE Seventh International Conference on e-Science Workshops', pp. 20–27.

Kessel, A.-L., Sahri, S., Groppe, S., Groppe, J., Khorashadizadeh, H., Pignal, M., Perez Pimparé, E. & Vignes-Lebbe, R. (2025), 'Impact of chatbots on user experience and data quality on citizen science platforms', *Computers* **14**(1), 21.

La Sorte, F. A. & Somveille, M. (2020), 'Survey completeness of a global citizen-science database of bird occurrence', *Ecography* **43**(1), 34–43.

Liu, F. T., Ting, K. M. & Zhou, Z.-H. (2008), Isolation forest, *in* '2008 Eighth IEEE International Conference on Data Mining', IEEE, pp. 413–422.

Lotfian, M., Ingensand, J. & Brovelli, M. A. (2021), 'The partnership of citizen science and machine learning: benefits, risks, and future challenges for engagement, data collection, and data quality', *Sustainability* **13**(14), 8087.

Lotfian, M., Ingensand, J., Ertz, O., Oulevay, S. & Chassin, T. (2019), Auto-filtering validation in citizen science biodiversity monitoring: A case study, *in* 'Proceedings of the ICA', Vol. 2, Copernicus GmbH, pp. 1–5.

Nizan, O. & Tal, A. (2024), k-nnn: nearest neighbors of neighbors for anomaly detection, *in* 'Proceedings of the IEEE/CVF Winter conference on applications of computer vision', pp. 1005–1014.

Parris, K. M., Steven, R., Vogel, B., Lentini, P. E., Hartel, J. & Soanes, K. (2023), 'The value of question-first citizen science in urban ecology and conservation', *Conservation science and practice* **5**(6), e12917.

Pocock, M. J., Adriaens, T., Bertolino, S., Eschen, R., Essl, F., Hulme, P. E., Jeschke, J. M., Roy, H. E., Teixeira, H. & De Groot, M. (2024), 'Citizen science is a vital partnership for invasive alien species management and research', *Iscience* .

Ramaswamy, S., Rastogi, R. & Shim, K. (2000), Efficient algorithms for mining outliers from large data sets, *in* 'Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data', pp. 427–438.

Rempel, R. S., Jackson, J. M., Van Wilgenburg, S. L. & Rodgers, J. A. (2019), 'A multiple detection state occupancy model using autonomous recordings facilitates correction of false positive and false negative observation errors.', *Avian Conservation & Ecology* **14**(2).

Sheard, J. K., Adriaens, T., Bowler, D. E., Büermann, A., Callaghan, C. T., Camprasse, E. C., Chowdhury, S., Engel, T., Finch, E. A., von Gönner, J. et al. (2024), 'Emerging technologies in citizen science and potential for insect monitoring', *Philosophical Transactions of the Royal Society B* **379**(1904), 20230106.

Siebold, M. (2025), Spatio-temporal shifts in citizen science data: Detecting disruptions in bird sightings with change point analysis, Master's thesis, Hamburg University of Applied Sciences (HAW Hamburg). Master's thesis.
URL: *https://www.mars-group.org/img/student-work/theses/siebold$_m$s$_t$hesis.pdf*

Yepmo, V., Smits, G., Lesot, M.-J. & Pivert, O. (2024), 'Leveraging an isolation forest to anomaly detection and data clustering', *Data & Knowledge Engineering* **151**, 102302.

Zhou, C. & Paffenroth, R. C. (2017), Anomaly detection with robust deep autoencoders, *in* 'Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', pp. 665–674.

Zhou, Y., Xia, H., Yu, D., Cheng, J. & Li, J. (2024), 'Outlier detection method based on high-density iteration', *Information Sciences* **662**, 120286.

Zhu, R. & Newman, G. (2024), 'Methodological overview for ebird, inaturalist, gap data, and wildlife richness', *Contemporary Landscape Performance Methods and Techniques* pp. 77–83.

A Appendix

A.1 Applied Tools

Table A.1 lists the tools and resources used for this Master thesis.

Table A.1: Used Tools and Resources

Tool	Usage
LaTeX	Typesetting and layout tool used for the creation of this document
Corine Land Cover dataset (CLC)	Creation of the Land-Cover-feature
Digital Elevation Model (EU-DEM)	Creation of the Altitude-feature
Gradio	Python package for building the presented review interface
ChatGPT 4o	Translation tasks for the creation of this document

Erklärung zur selbständigen Bearbeitung

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne fremde Hilfe selbständig verfasst und nur die angegebenen Hilfsmittel benutzt habe. Wörtlich oder dem Sinn nach aus anderen Werken entnommene Stellen sind unter Angabe der Quellen kenntlich gemacht.

Ort	Datum	Unterschrift im Original