

How Much of the Forest Do Citizen Scientists Actually Record? A Decade of iNaturalist Sampling in the Hyrcanian Forests

Saman Ghasemian Sorboni ^{1,*} and Motahareh Taghipour ²

¹ Department of Land, Environment, Agriculture and Forestry (TeSAF), University of Padova, Viale dell'Università, 16, 35020 Legnaro PD, Italy

² Department of Biology (DiBio), University of Padova, Via U. Bassi 58/B, 35131 Padova, Italy

* Correspondence: saman.ghasemiansorboni@studenti.unipd.it

Abstract

Citizen-science records increasingly inform biodiversity research and conservation planning, yet how much of a large forest they actually cover is rarely measured. We analyzed 1,700 research-grade iNaturalist records from the Hyrcanian Forest, northern Iran (2015–2025), on a 10 × 10 km grid, comparing observed coverage with null models. Annual records rose from 48 to 386, but occupied only 43.7% of forest cells, 41% fewer than an even spread would give, and 56% of cells had no record. Fifteen observers (top 10%) provided 69% of records. In a presence–background model with cluster-robust errors, only elevation was independently associated with record presence (odds ratio 0.92 per 100 m); distances to roads and settlements were not. Because steep terrain limits where people can walk and record, similar elevation-related gaps may occur in other mountain forests, where record counts could overstate coverage. Models using such records should adjust for elevation bias.

Keywords: Forest monitoring; Community-based monitoring; Citizen science; Global Biodiversity Information Facility (GBIF); Elevation gradient; Sampling bias; Species Distribution Modeling (SDM); Ecological informatics; Spatial coverage; Mountain forests

1. Introduction

Citizen-science platforms now supply a large share of new species occurrence records [1]. Research-grade iNaturalist observations that carry open licences are shared with the Global Biodiversity Information Facility (GBIF) [2], where

they are used for distribution modelling, conservation assessment and monitoring [3]. These records are collected opportunistically, so where, when and what is recorded depends on the observers [4]. Documented biases include a preference for certain taxa, concentration of records near roads, cities and trails, strong inequality among observers, and under-recording of remote or high-elevation areas [5]. Because such biases affect inference from the data, they need to be described for each region before records are used [6].

The Hyrcanian Forest is a Tertiary relict broadleaved forest along the southern Caspian coast, listed as a UNESCO World Heritage Site [7]. Hyrcanian ecoregion is dominated by trees such as Oriental beech (*Fagus orientalis*), chestnut-leaved oak (*Quercus castaneifolia*), and common hornbeam (*Carpinus betulus*), holds many endemic and relict species [8-9]. The forest extends from the Caspian lowlands to the upper forest limit over steep terrain, which makes systematic field surveys expensive and gives opportunistic records an obvious role. Some studies in Iran already use GBIF occurrences for species distribution models and conservation planning [10-12]. Such models assume that records sample a species' environmental range reasonably well.

We analyzed iNaturalist records from the Hyrcanian Forest for 2015–2025 to answer four questions. How much of the forest has been recorded, and did coverage grow with the number of records? How unequally are records contributed among observers? Are records biased by elevation or by distance to roads and settlements? How did the timing and taxonomic composition of records change?

2. Materials and Methods

2.1. Study area and iNaturalist data

Observations from the Hyrcanian Forest of Gilan, Mazandaran and Golestan provinces were downloaded from iNaturalist for 2015–2025 [13]. We retained research-grade records and removed records with obscured coordinates and records flagged as captive or cultivated. Remaining records were clipped to the forest boundary, leaving 1,700 records (Fig. 1a). Records with a reported positional accuracy $\leq 1,000$ m, or with no accuracy reported (10.1% of records), were used for the grid analyses. The 1,172 records with accuracy ≤ 250 m formed the subset for the environmental bias model.

All layers were projected to a custom Albers equal-area projection centered on the study region (standard parallels 36° and 38.5° N, central meridian 53° E, WGS84), so that grid cells have equal area and distances are in meters. The forest boundary covered 18,834 km². We laid a 10×10 km grid over the boundary and retained the 412 cells that intersect it. Because edge cells lie partly outside the forest, coverage is also reported for the 187 core cells with at least 50% of their area inside the boundary (Fig. 1b). Each record was assigned to the cell containing it.

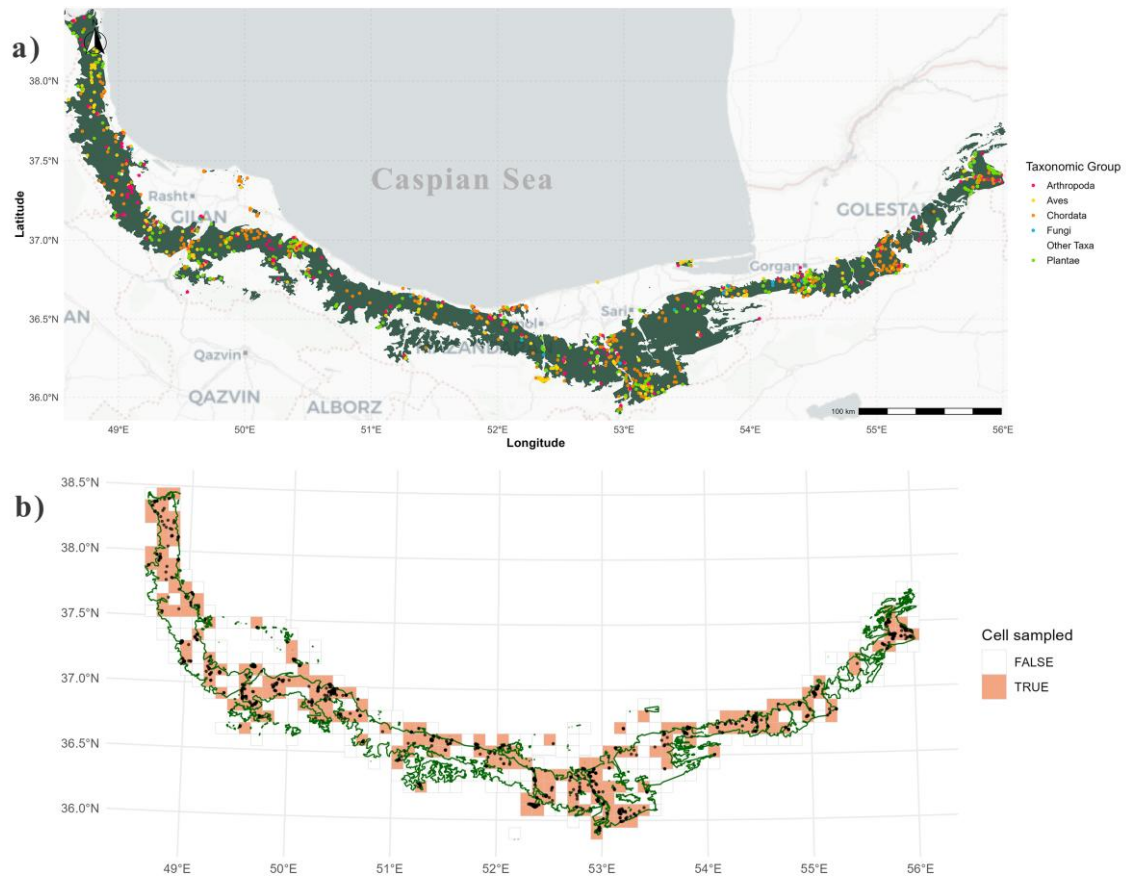


Figure 1. Study area and spatial coverage of iNaturalist records in the Hyrcanian Forest, northern Iran, 2015–2025. (a) Research-grade records ($n = 1,700$) inside the forest boundary ($18,834 \text{ km}^2$), coloured by taxonomic group. (b) The $10 \times 10 \text{ km}$ grid used for the coverage metrics, in an Albers equal-area projection. Of the 412 cells that intersect the forest, 180 (shaded) contain at least one record (black dots).

2.2. Spatial coverage and observer contributions

Spatial coverage was measured as the cumulative number of cells with at least one record, expressed as a percentage of all forest cells and of core cells. The annual novelty rate was the percentage of a year's records falling in cells with no record in any earlier year; 2015 served as the baseline and was excluded. We compared both metrics with two null models that kept the observed number of records per year. In null A, records were placed uniformly at random over the forest area (200 repetitions), which indicates how the observed footprint compares with spatially even sampling. In null B, year labels were permuted among the observed record locations (1,000 permutations), which preserves the observed spatial clustering and tests whether the accumulation of new cells differs from that expected if year were unrelated to location. Null envelopes are the 2.5th–97.5th percentiles.

Inequality in contributions was summarized with the Gini coefficient of records per observer (iNaturalist user ID) and the share of records contributed by the top 1%, 5% and 10% of observers. For each year we also counted active, new and returning observers (new = first record in the dataset), and we counted the cells recorded by the ten most prolific observers.

Records were assigned to plants, vertebrates (birds, mammals, reptiles, amphibians, ray-finned fishes), arthropods (insects, arachnids), fungi or other taxa using iNaturalist's iconic taxon. Seasonal timing was visualized with density ridge plots of day of year. Seasonal evenness for each year was Shannon's H of monthly record counts divided by $\ln(12)$, where 1 indicates equal activity in all months.

2.3. Elevation and distance to roads and settlements

We generated uniform random background points across the forest, equal in number to the ≤ 250 m subset ($n = 1,172$). For each observed and background point we extracted elevation from a 30-arc-second SRTM Digital Elevation Model [14] and the Euclidean distance to the nearest road and nearest city. Observed and background distributions were compared using medians and the two-sample Kolmogorov–Smirnov statistic (D). We then fitted a logistic regression of presence (1) versus background (0) on elevation (per 100 m) and the two distances (per km). Multicollinearity was assessed with variance inflation factors. Because records within a cell are not independent, standard errors were clustered by 10×10 km cell, and odds ratios are reported with 95% confidence intervals from these robust errors.

3. Results

3.1. Records and forest coverage

After filtering, 1,700 research-grade records fell within the forest boundary. Annual records ranged from 26 to 111 in 2015–2021, then rose to 153 in 2022, 305 in 2023 and 430 in 2024, and fell to 386 in 2025. The number of cells recorded each year grew more slowly than the number of records, from 21 cells in 2015 to 62 in 2025. Records per occupied cell increased from 3.5 in 2022 to 6.2–6.4 in 2023–2025.

By 2025, cumulative coverage reached 180 of 412 forest cells (43.7%) and 112 of 187 core cells (59.9%). Coverage was still increasing, with 12–24 new cells added per year in 2022–2025. Under null A, the same number of records distributed uniformly across the forest would have occupied 306 cells on average (95% range 297–315), well above the 180 observed cells.

3.2. Contributions by observer

Records were unevenly distributed among the 148 observers (Gini = 0.76; Fig. 2a). The top 10% of observers ($n = 15$) contributed 69.2% of records, and 35.1% of observers contributed one record. The ten most prolific observers recorded in 87 of the 180 sampled cells and were the first to record 37.8% of those cells.

Active observers increased from 12–13 in 2020–2021 to 33 in 2022 and 52 in 2025, with about half being new to the dataset in 2024 (22 of 43) and 2025 (26 of 52; Fig. 2b). The five most active observers contributed 77% of records in 2023 but about 50% in 2024–2025.

The novelty rate fell from 85.5% in 2016 to 22–23% in 2020–2021 and then remained between 29% and 39% in 2022–2025 (Fig. 3). In 2022–2025 it exceeded the 95% ranges of both null models. Mean novelty in 2023–2025 was 3.6–8.6% under null B and 3.6–14.5% under null A.

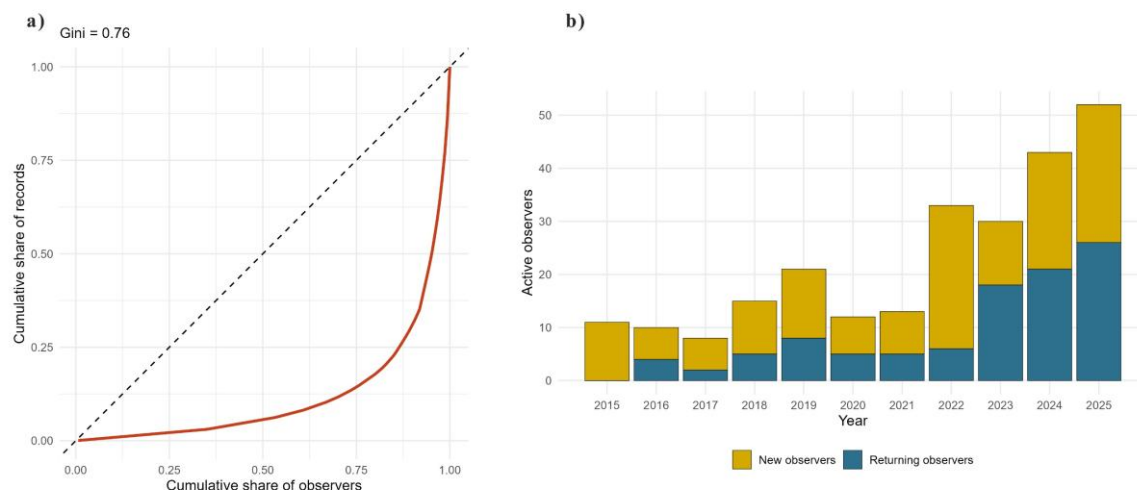


Figure 2. Contributions by observer. (a) Lorenz curve of records per observer (148 observers, 1,700 records). The dashed line marks equal contributions; the Gini coefficient is 0.76. (b) Active observers per year, split into new observers (first record in the dataset) and returning observers (at least one earlier record). All observers in 2015 count as new because the dataset starts that year.

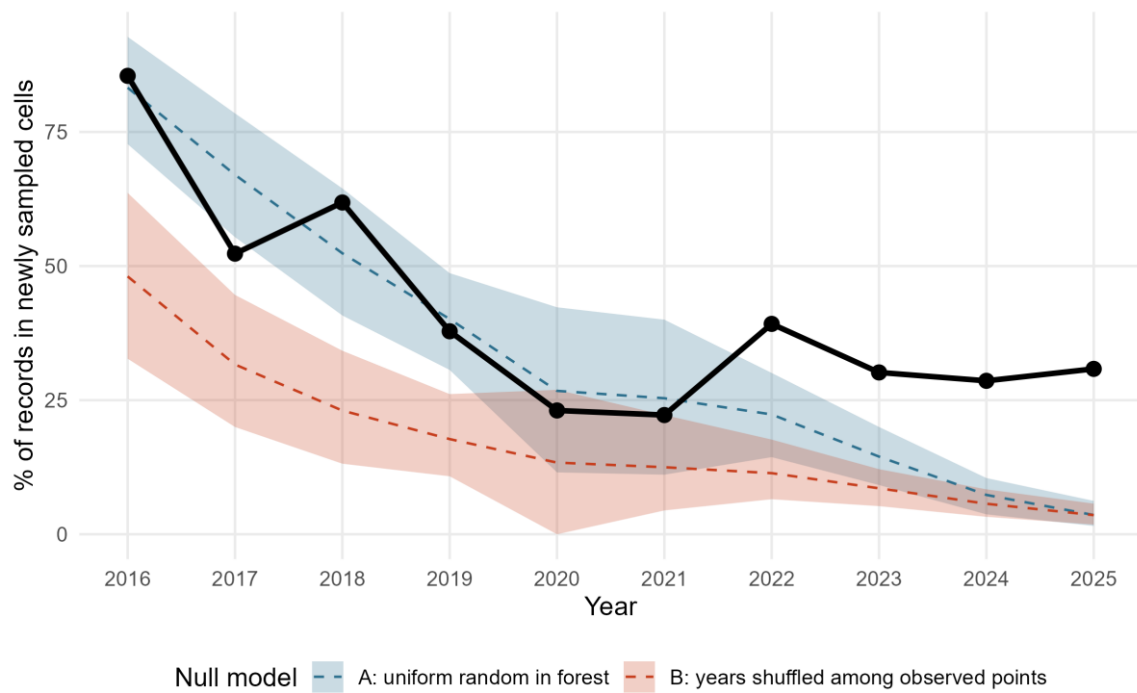


Figure 3. Percentage of each year’s records that fell in 10×10 km cells with no earlier record (black line), compared with two null models that keep the observed number of records per year. Null A places records uniformly at random over the forest area (200 repetitions). Null B shuffles year labels among the observed record locations (1,000 permutations). Dashed lines show null means, and shaded bands show the 2.5th–97.5th percentiles. 2015 is the baseline year and is not plotted.

3.3. Environmental and Accessibility Biases

Records occurred at lower elevations than background points (median 443 m vs. 961 m; KS $D = 0.38$) and closer to roads (1.88 vs. 4.13 km; $D = 0.28$) and settlements (7.29 vs. 11.5 km; $D = 0.27$; all $p < 0.001$).

In the joint logistic model ($VIF \leq 1.41$), elevation was the only predictor with a confidence interval excluding 1 (OR = 0.918 per 100 m, 95% CI 0.847–0.996, $p = 0.040$), equivalent to 8.2% lower odds of a record relative to background per 100 m of elevation. Distance to roads (OR = 0.969 per km, 95% CI 0.893–1.05, $p = 0.449$) and to settlements (OR = 0.987 per km, 95% CI 0.935–1.04, $p = 0.624$) had no detectable independent effect.

3.4. Timing and taxonomic composition of records

In 2015–2021 (26–111 records per year), peak timing changed from year to year (Fig. 4a). Seasonal evenness was 0.62–0.79 in these years, 0.83 in 2022, and 0.89–0.93 in 2023–2025 (Table 1). Records in 2023–2024 were spread across much of the year, while 2025 had a pronounced September–October peak.

Plants contributed the most records in May–June and October, vertebrates in Feb–April, June, September–October, and arthropods made up a larger share in June–July and September. Fungi were recorded mainly in October–December (Fig. 4b).

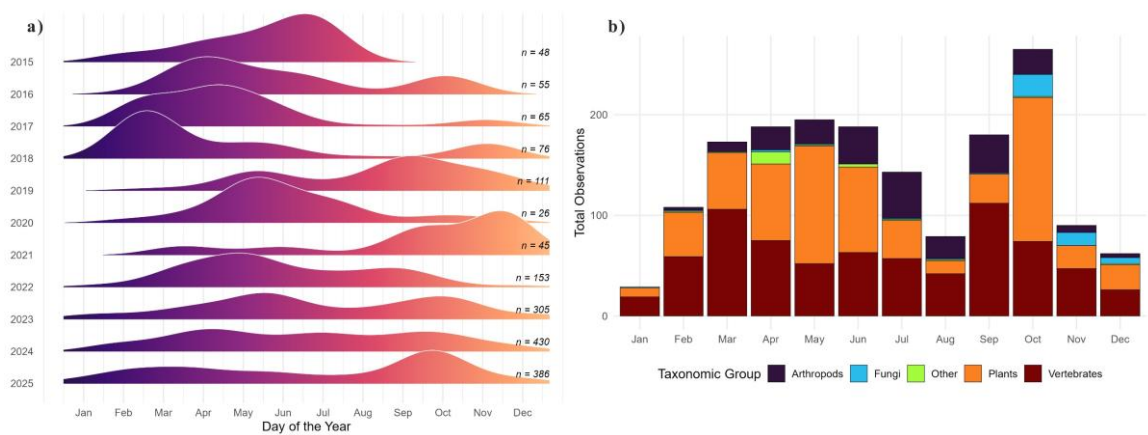


Figure 4. Timing and taxonomic composition of records. (a) Distribution of records over the day of year for each year, as kernel density ridges. Each ridge is scaled to its own year, so heights are not comparable between years; n is the number of records, and color follows the day of year. (b) Monthly record counts for all years combined, by taxonomic group. Vertebrates include birds, mammals, reptiles, amphibians and ray-finned fishes; arthropods include insects and arachnids.

Table 1. Annual sampling metrics, 2015–2025. New observers are those with no earlier record in the dataset. Cumulative cells are cells with ≥ 1 record since 2015 (of 412). Novelty rate is the percentage of records in cells with no earlier record (2015 is the baseline). Seasonal evenness is Shannon H of monthly counts divided by $\ln 12$.

Year	Active Obs.	New Obs.	Cumulative Cells	Novelty Rate (%)	Seasonal Evenness
2015	11	11	21	na	0.62
2016	10	6	41	85.5	0.76
2017	8	6	58	52.3	0.66
2018	15	10	71	61.8	0.72
2019	21	13	91	37.8	0.79
2020	12	7	96	23.1	0.71
2021	13	8	102	22.2	0.64
2022	33	27	122	39.2	0.83
2023	30	12	144	30.2	0.92
2024	43	22	168	28.6	0.93

2025	52	26	180	30.8	0.89
------	----	----	-----	------	------

4. Discussion

Data volume in the Hyrcanian Forest grew faster than the area it covers. Annual records rose from 48 in 2015 to 386 in 2025, whereas the number of 10×10 km cells recorded per year rose about threefold (21 to 62), so records per occupied cell increased from 3.5 in 2022 to 6.2–6.4 in 2023–2025. By 2025 the records covered 43.7% of forest cells (59.9% of core cells), about 40% fewer than the 306 cells expected if the same records were spread evenly over the forest. Spatial concentration of this kind is typical of citizen-science datasets [15]. However, we found no spatial plateau. Cumulative coverage rose every year, and the novelty rate stayed near 30% in 2023–2025. New cells continued to be added, but more slowly than records accumulated.

Contributions were highly unequal (Gini = 0.76), with 15 observers (the top 10%) providing 69% of records, a skew also reported for iNaturalist as a whole [4]. The ten most prolific observers recorded in 87 of the 180 occupied cells and were first to record 38% of them, so a small group accounts for much of both the volume and the geographic reach. The 2023 increase was concentrated among few people (the five most active observers provided 77% of that year’s records), whereas the 2024–2025 records came from a wider group of 43–52 observers, about half of them new to the dataset. Whether these newcomers continue to contribute cannot be judged from two years. Observed novelty exceeded the expectation from shuffling years among the observed locations (null B) in most years. We read this descriptively rather than as evidence about site fidelity, because the shuffle removes the link between an observer, a trip and a year.

Elevation was the only variable independently associated with record presence. The odds of a record fell by 8.2% per 100 m, about 35% per 500 m. Distances to roads and settlements differed between records and background points when examined alone, but not in the joint model. This does not show that roads are unimportant. The road interval still allows odds up to about 11% lower per kilometer, and road and city layers that omit forest tracks and villages would weaken any distance effect. Elevation may also stand in for factors we did not measure, such as travel time, trail availability and distance to the coastal towns where many observers live. We therefore describe the pattern as a bias toward lower-elevation forest rather than toward urban edges as such.

Seasonal evenness of records rose from 0.62–0.79 in 2015–2021 to 0.89–0.93 in 2023–2024, consistent with more observers recording across the year. Two caveats apply. Annual samples before 2022 were small and Shannon evenness is biased low in small samples [16]. The 2025 records also show a clear September–October peak. Moreover, the seasonal pattern of citizen-science records combines species phenology, detectability and observer availability, so these

data describe when people record, not when species are active [17]. The monthly differences between taxonomic groups are what would be expected given differences in when each group is conspicuous, but we did not test this.

Several other limitations apply. Records are opportunistic, and we have no measure of effort per observer (hours or trips), so lower coverage may reflect observer numbers as much as observer choices. Removing obscured records excludes threatened taxa, and the accuracy threshold limited the bias model to 1,172 records. Coverage metrics depend on the 10 km cell size, and we did not test alternatives. Distances and elevation entered the model as linear terms, and the background points represent forest area rather than observer effort. Observers' own locations and travel routes were not available. Species distribution models built from these records should correct for the bias toward low elevations, for example with bias layers or target-group background points, otherwise higher-elevation habitat will be under-represented [18].

The Hyrcanian Forest is large, and most of it is poorly recorded. Over eleven years, the 1,700 records average about 0.4 records per 10×10 km cell per year, and 56% of forest cells have no record. About half of the observers active in 2024–2025 were new to the dataset, so outreach has a realistic chance of changing where records are made. We suggest that campaigns aim to move observers into unrecorded forest, not only recruit more of them. Options include guided walks and short bioblitzes in unrecorded cells led by rangers, universities or hiking clubs, training on taking identifiable photographs with location enabled, and a public map of unrecorded cells that observers can use to choose where to go [5,19]. Remote and higher-elevation areas also call for attention to safety, permits in protected areas and avoiding disturbance, so outreach should come with a code of conduct.

Our results have implications beyond this forest, although we tested only one. First, record counts are a poor measure of coverage. Records rose eightfold between 2015 and 2025, and records per occupied cell nearly doubled after 2022, yet 56% of forest cells remained empty. A growing dataset can therefore keep sampling the same places, and reporting the share of cells recorded, together with a comparison against an even-sampling null, would show this in any region. Second, the decline in records with elevation may apply wherever steep terrain limits where people can record. We present this as a hypothesis for other mountain forests. Third, the workflow uses only public occurrence records and freely available map layers, so it can be repeated for other forests and platforms. Elevation, the only variable associated with record presence here, is a sensible candidate covariate for bias layers in species distribution models of mountain forests. Comparative studies can test whether the pattern holds elsewhere and whether outreach changes it.

Conclusion

Between 2015 and 2025, annual iNaturalist records from the Hyrcanian Forest rose from 48 to 386, but coverage did not keep pace. The 1,700 records occupied 43.7% of forest cells, 41% fewer than an even spread would give, and 56% of cells still have no record. Fifteen observers supplied 69% of records, and the odds of a record fell by 8.2% with every 100 m of elevation, whereas distance to roads and settlements was not independently associated with presence. Record counts alone therefore say little about how much of a forest has been sampled. We recommend that studies using opportunistic records report coverage against an even-sampling null, and that species distribution and macroecological models of mountain forests include elevation in their bias correction. Because the approach uses only public occurrence data and free map layers, it can be repeated elsewhere to test whether the elevation pattern is general. In the Hyrcanian Forest, directing observers to unrecorded cells is the clearest route to better coverage; whether campaigns achieve this remains to be tested.

Author Contributions: S.G.S.: Conceptualization, methodology, software, validation, formal analysis, investigation, writing—original draft, writing—review and editing, visualization. M.T.: Software, data curation, writing—original draft. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: All data are available from the corresponding author upon reasonable request.

Acknowledgments: We thank iNaturalist and its contributors for providing open-access biodiversity data that made this study possible. We thank Anthropic for providing access to Claude Sonnet 5.5, which was used for language editing of this draft. All content was reviewed and verified by the authors, who take full responsibility for the work.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Chandler, M.; See, L.; Copas, K.; Bonde, A.M.Z.; López, B.C.; Danielsen, F.; Legind, J.K.; Masinde, S.; Miller-Rushing, A.J.; Newman, G.; et al. Contribution of Citizen Science towards International Biodiversity Monitoring. *Biol. Conserv.* **2017**, *213*, 280–294. <https://doi.org/10.1016/j.biocon.2016.09.004>.
2. López-Guillén, E.; Herrera, I.; Bensid, B.; Gómez-Bellver, C.; Ibáñez, N.; Jiménez-Mejías, P.; Mairal, M.; Mena-García, L.; Nualart, N.; Utjés-Mascó, M.; et al. Strengths and Challenges of Using iNaturalist in Plant Research with Focus on Data Quality. *Diversity* **2024**, *16*, 42. <https://doi.org/10.3390/d16010042>.

3. Mason, B.M.; Mesaglio, T.; Barratt Heitmann, J.; Chandler, M.; Chowdhury, S.; Gorta, S.B.Z.; Grattarola, F.; Groom, Q.; Hitchcock, C.; Hoskins, L.; et al. iNaturalist Accelerates Biodiversity Research. *BioScience* **2025**, *75*, 953–965. <https://doi.org/10.1093/biosci/biaf104>.
4. Di Cecco, G.J.; Barve, V.; Belitz, M.W.; Stucky, B.J.; Guralnick, R.P.; Hurlbert, A.H. Observing the Observers: How Participants Contribute Data to iNaturalist and Implications for Biodiversity Science. *BioScience* **2021**, *71*, 1179–1188. <https://doi.org/10.1093/biosci/biab093>.
5. Sorboni, S.G.; Pourebrahim, S.; Chen, J.E.; Hadipour, M. Anthropogenic Accessibility and Observer Inequality Dictate Spatial Biodiversity Patterns in Malaysia. *Sci. Rep.* **2026**, *16*, 58296. <https://doi.org/10.1038/s41598-026-58296-2>.
6. Bird, T.J.; Bates, A.E.; Lefcheck, J.S.; Hill, N.A.; Thomson, R.J.; Edgar, G.J.; Stuart-Smith, R.D.; Wotherspoon, S.; Krkosek, M.; Stuart-Smith, J.F.; et al. Statistical Solutions for Error and Bias in Global Citizen Science Datasets. *Biol. Conserv.* **2014**, *173*, 144–154. <https://doi.org/10.1016/j.biocon.2013.07.037>.
7. Hosseini, S.M. Outstanding Universal Values of Hyrcanian Forest, the Newest Iranian Property, Inscribed in the UNESCO's World Heritage List. *Tour. Res.* **2019**, *1*, 1–17.
8. Ghasemian Sorboni, S.; Ghahremaninejad, F.; Hadipour, M.; Carrer, M. Projecting Future Climatic Refugia for *Fagus Orientalis* in the Hyrcanian Ecoregion. *J. Nat. Conserv.* **2026**, *95*, 127444. <https://doi.org/10.1016/j.jnc.2026.127444>.
9. Tohidifar, M.; Moser, M.; Zehzad, B.; Ghadirian, T. Biodiversity of the Hyrcanian Forests: A Synthesis Report; Caspian Hyrcanian Forest Project, FRWO: Tehran, Iran, **2016**. <https://doi.org/10.13140/RG.2.2.31436.00649>.
10. Ghaitaranpour, M.; Poorbagher, H.; Eagderi, S. Modelling the Spatial Distribution of Narrow-Barred Spanish Mackerel (*Scomberomorus commerson*; Lacepede 1800) in the Persian Gulf Using Logistic Regression. *Iran. J. Fish. Sci.* **2020**, *29*, 201–211.
11. Oveisi, M.; Hazrati, Z.; Hejazi, M.; Arvin, K.; Mueller-Schaerer, H. Climate Change–Driven Shifts in Weed and Invasive Alien Plant Distributions in Iran: A Species Distribution Modelling Approach. *Iran. J. Weed Sci.* **2026**, *21*, 39–61. <https://doi.org/10.22034/ijws.2026.371350.1511>.
12. Ghasemian Sorboni, S.; Hadipour, M.; Chen, J.E.; Pourebrahim, S. Evaluating Climate Change Impacts on Habitat Suitability for Iberian Freshwater Crab “*Potamon Ibericum* Bieberstein, 1808” in Iran. *Front. Ecol. Evol.* **2025**, *13*, 1608518. <https://doi.org/10.3389/fevo.2025.1608518>.
13. iNaturalist. iNaturalist Observations Export. Available online: <https://www.inaturalist.org/observations/export> (accessed on 14 April 2026).
14. NASA JPL. NASA Shuttle Radar Topography Mission Global 30 Arc Second; NASA EOSDIS Land Processes Distributed Active Archive Center: Sioux Falls, SD, USA, **2013**. <https://doi.org/10.5067/MEaSURES/SRTM/SRTMGL30.003>.
15. Tiago, P.; Ceia-Hasse, A.; Marques, T.A.; Capinha, C.; Pereira, H.M. Spatial Distribution of Citizen Science Casuistic Observations for Different Taxonomic Groups. *Sci. Rep.* **2017**, *7*, 13130. <https://doi.org/10.1038/s41598-017-13130-8>.

16. Konopiński, M.K. Shannon Diversity Index: A Call to Replace the Original Shannon's Formula with Unbiased Estimator in the Population Genetics Studies. *PeerJ* **2020**, *8*, e9391. <https://doi.org/10.7717/peerj.9391>.
17. Rosário, I.T.; Tiago, P.; Chozas, S.; Capinha, C. When Do Citizen Scientists Record Biodiversity? Non-random Temporal Patterns of Recording Effort and Associated Factors. *People Nat.* **2025**, *7*, 860–870. <https://doi.org/10.1002/pan3.70017>.
18. Johnston, A.; Moran, N.; Musgrove, A.; Fink, D.; Baillie, S.R. Estimating Species Distributions from Spatially Biased Citizen Science Data. *Ecol. Model.* **2020**, *422*, 108927. <https://doi.org/10.1016/j.ecolmodel.2019.108927>.
19. Tiago, P.; Evaristo, I.; Pinto, B. The Role of BioBlitzes in Citizen Science: Insights from Participants and Experts. *Front. Environ. Sci.* **2024**, *12*, 1347428. <https://doi.org/10.3389/fenvs.2024.1347428>.