Reports in Science and Applied Research
https://sciappreports.science

Data Validity with Citizen Science Protocols and Knowledge Diffusion across Urban Biodiversity Surveys

Alonso Flores Fernández[*], Carlos Montero

Faculty of Political and Social Sciences, Universidade de Santiago de Compostela, Santiago de Compostela, Galicia, Spain

Corresponding author: alonso.f.fernandez@gmail.com

Abstract

Citizen science has emerged as a transformative methodology for monitoring urban biodiversity, yet concerns regarding data validity limit its integration into formal environmental policy. This study systematically analyzes the intersection of standardized observation protocols and knowledge diffusion dynamics among volunteer networks, demonstrating how their synergy mitigates systemic identification and sampling errors. Through a multi-tiered experimental framework deployed across urban green spaces, we examine the efficacy of diverse protocol designs (ranging from unstructured to highly structured) and knowledge diffusion channels (including digital training modules, peer-to-peer mentoring, and real-time algorithmic feedback). Our empirical evaluation reveals that while strict protocols establish a baseline for data quality, the rate of knowledge diffusion across participants serves as the primary driver of long-term error reduction and spatial-temporal reliability. Peer-to-peer mentoring networks, in particular, exhibit the highest efficiency in transferring tacit taxonomic knowledge, leading to a substantial decrease in false-positive species reports. By integrating social network analysis with spatial ecology, we formulate an optimized model for crowd-sourced biodiversity monitoring that reconciles data validation requirements with participant engagement. Ultimately, this research provides actionable strategies for municipal planners and conservation scientists to leverage volunteer-driven datasets for robust urban environmental management.

Keywords: Citizen Science; Urban Biodiversity; Data Validity; Knowledge Diffusion; Multidisciplinary Research

1. Introduction

1.1 Urban Biodiversity and the Rise of Citizen Science

The rapid pace of global urbanization has profoundly altered natural landscapes, leading to unprecedented levels of habitat fragmentation, localized extinction events, and the proliferation of invasive species. Urban ecosystems, characterized by high spatial heterogeneity and complex microclimated variations, represent critical yet highly vulnerable components of global biodiversity conservation networks [1]. Traditional ecological monitoring, dependent on professional taxonomists and structured agency surveys, is increasingly constrained by limited municipal budgets, restricted spatial access, and the sheer temporal scale required to capture dynamic ecological transitions. Consequently, the scientific community has turned toward crowdsourced data collection models, broadly conceptualized as citizen science, to bridge these observational gaps [2]. By mobilizing public volunteers, environmental research initiatives can simultaneously monitor vast geographic areas and acquire continuous datasets that would otherwise remain resource-prohibitive.

In the context of urban environments, citizen science programs provide a dual utility. They collect extensive spatial-temporal occurrence data on avian, insect, and plant communities while fostering public environmental literacy and civic stewardship [3]. Despite these clear benefits, the integration of public-generated datasets into peer-reviewed research and formal environmental policy remains highly contentious. Skepticism persists among professional researchers regarding the accuracy of taxonomic identifications, the consistency of spatial coverage, and the systematic biases inherent in opportunistic collection methods. Addressing these skepticism-driven barriers is essential if crowdsourced monitoring is to achieve its full potential in informing municipal conservation strategies and global biodiversity frameworks.

1.2 The Paradox of Data Validity in Crowd-Sourced Science

The fundamental challenge confronting citizen science is the inherent trade-off between project accessibility and data rigor. This phenomenon, known as the data validity paradox, posits that as data collection protocols become more rigorous and scientific, the barriers to volunteer participation rise, thereby reducing the volume of data collected and defeating the primary advantage of crowdsourced monitoring [4]. Conversely, lowering the barrier to entry by employing unstructured observation methods increases volunteer engagement but often yields datasets plagued by taxonomic misidentifications, spatial clustering around urban parks, and temporal biases toward weekends and pleasant weather conditions [5].

To resolve this paradox, researchers must move beyond viewing data validation as a retrospective, purely post-hoc statistical filtering process. Instead, data validity must be conceptualized as an emergent property of the entire participatory system, influenced both by the proactive design of observation protocols and the continuous social processes of learning and knowledge sharing within the volunteer cohort [6]. By understanding how information is absorbed, adapted, and shared among participants, project coordinators can design socio-technical systems that naturally elevate the baseline accuracy of raw data without discouraging volunteer enthusiasm.

1.3 Research Objectives and Theoretical Framework

This paper investigates how systematic observation protocols and knowledge diffusion mechanisms interact to determine and improve data validity in urban biodiversity surveys. Specifically, we examine the operational mechanisms through which scientific rigor is communicated to, and maintained by, non-expert participants. The research operates at the intersection of spatial ecology, human computer interaction, and social learning theory, proposing a comprehensive framework that evaluates data quality as a function of protocol structure and community-driven knowledge transfer.

Through a rigorous multi-month field experiment involving diverse volunteer cohorts across urban green spaces, this study addresses three key questions. First, how do different levels of protocol structure impact the rate and spatial distribution of taxonomic reporting errors? Second, what are the primary channels of knowledge diffusion that accelerate taxonomic learning curves among non-expert volunteers? Third, how can municipal environmental programs design hybrid models that maximize both data validity and community engagement? By addressing these questions, this research contributes to the development of highly reliable, cost-effective environmental monitoring architectures.

2. Literature Review and Theoretical Framework

2.1 Data Validation Paradigms in Citizen Science

Historically, scientific evaluation of citizen science data has relied on rigid validation frameworks designed to filter out erroneous observations after collection. These post-hoc methods typically involve expert validation, where professional taxonomists verify submitted photographic or acoustic evidence, or automated algorithmic validation, where geographical and temporal outliers are flagged for review [7]. While effective at preserving the integrity of final research databases, post-hoc filtering often results in high data discard rates, which can alienate volunteers whose efforts are retroactively dismissed.

More recently, research has shifted toward preventative quality assurance measures integrated directly into the collection stage [8]. These preventive approaches emphasize structured protocols, such as fixed-point sampling, standardized searching times, and compulsory physical training workshops. However, implementing these protocols requires a deep understanding of volunteer motivation and cognitive limitations. Overly complex structures can lead to high attrition rates, creating a demographic bias toward retired professionals or highly educated individuals, which limits the demographic diversity of the participant pool [9]. Therefore, understanding the optimal balance between protocol rigidity and participant retention is a key focus of contemporary research.

2.2 Knowledge Diffusion Mechanisms in Volunteer Networks

Knowledge diffusion, defined as the process by which scientific information, taxonomic skills, and procedural protocols spread through a network of actors, is a critical driver of data validity in crowdsourced research [10]. Within citizen science communities, knowledge diffusion occurs through both formal and informal channels. Formal channels encompass structural educational interventions, such as interactive digital tutorials, taxonomic identification keys, and pre-survey webinars. In contrast, informal channels rely on peer-to-peer interactions, where experienced volunteers mentor novices on online forums, during field excursions, or via social media platforms [11].

The dynamics of this diffusion can be analyzed using social network theory, which maps the pathways of communication and skill transfer among participants. Studies indicate that communities with high network density and strong social ties facilitate rapid skill acquisition and adherence to scientific standards [12]. However, the diffusion of complex taxonomic knowledge often faces bottlenecks, particularly when identifying cryptic species or navigating complex observation protocols. By identifying and optimizing these diffusion pathways, citizen science organizers can systematically elevate the scientific capabilities of the entire network over time.

2.3 Intersecting Protocol Rigor and Knowledge Flows

The intersection of protocol design and knowledge diffusion represents a novel frontier in environmental data science. Rather than viewing protocols as static instructions, contemporary theory models them as dynamic learning scaffolding that adapts to the rising competency levels of the volunteer network [13]. When structured protocols are coupled with active knowledge diffusion channels, they create a positive feedback loop: as volunteers internalize scientific methods, their reliance on rigid protocol constraints decreases, allowing them to execute more complex, semi-structured tasks with high accuracy [14].

This synergistic relationship is particularly critical in urban landscapes. Urban areas are characterized by micro-habitats that require specialized observation methods, making static, one-size-fits-all protocols highly inefficient. By facilitating localized knowledge sharing, experienced urban volunteers can adapt broad protocols to specific micro-environments, such as vertical gardens, pocket parks, or suburban riverbanks, without compromising overall scientific validity [15]. This study develops and tests a theoretical model that quantifies this interaction, demonstrating how social learning structures can replace or supplement restrictive observation rules.

3. Methodology and Protocol Design

3.1 Survey Design and Participant Stratification

To empirically investigate the dynamics of data validity, a comprehensive field experiment was conducted over a six-month period within a metropolitan area characterized by diverse urban green spaces, including public parks, botanical gardens, and riparian corridors. A total of three hundred and fifty public volunteers were recruited and stratified into three distinct experimental groups, each assigned a different observational protocol tier to monitor urban avian and insect populations. This stratification allowed for the systematic isolation of protocol structure as an experimental variable.

The first group operated under Tier 1 (Unstructured Protocols), where participants were instructed to report opportunistic sightings of target taxa using basic mobile applications without constraints on search time, location, or physical evidence submission [16]. The second group utilized Tier 2 (Semi-Structured Protocols), which mandated a minimum observation duration of twenty minutes within specified spatial buffers and required the submission of photographic evidence for taxonomic verification. The third group followed Tier 3 (Highly Structured Protocols), which required strict adherence to linear transect sampling, specific meteorological conditions, and compulsory digital training before field deployment [17].

3.2 Knowledge Diffusion and Protocol Intervention

Alongside the structural protocol tiers, we introduced three distinct knowledge diffusion interventions across the participant cohorts to evaluate their capacity to reduce observation errors. The first intervention relied on passive digital training modules, which provided volunteers with access to interactive species identification keys and standard operating procedures via an online portal. This channel represented a formal, top-down dissemination of scientific knowledge with minimal interactive engagement.

The second intervention established a structured peer-to-peer mentoring network, where experienced volunteer ambassadors were paired with groups of five novices to conduct joint field surveys, provide feedback on identifications, and facilitate group discussions on social messaging platforms [18]. The third intervention utilized real-time algorithmic feedback integrated into the data submission application, which instantly flagged potential taxonomic and spatial anomalies, prompting the user to reconsider their submission based on local baseline historical data.

Table 1: Matrix of Citizen Science Protocol Tiers and Associated Validation Mechanisms

Protocol Tier	Primary Objective	Quality Assurance Mechanism	Participant Engagement Level
Tier 1: Unstructured	Exploratory mapping	Post-hoc expert verification	High flexibility, low retention
Tier 2: Semi-Structured	Target taxon monitoring	Built-in photographic validation	Moderate flexibility, medium retention
Tier 3: Highly Structured	Spatial abundance analysis	Real-time algorithmic checking	Low flexibility, high retention

3.3 Evaluation Metrics for Data Validity

To quantify data validity across the experimental cohorts, three primary metrics were established and tracked throughout the study. The first metric, Taxonomic Error Rate, calculated the proportion of submitted observations that were misidentified when compared against a gold-standard dataset verified by professional taxonomists. This metric was further subdivided into false positives (reporting a species that was not present) and false negatives (failing to report a species that was present).

The second metric, Spatial Deviation Distance, measured the spatial accuracy of the submitted coordinates against the actual physical location of the observed organism, identifying systematic mapping errors caused by GPS inaccuracy or user placement error [19]. The third metric, Mean Learning Curve Velocity, assessed the rate at which individual volunteers reduced their error rates over successive survey events, serving as a direct measure of the efficiency of the active knowledge diffusion channels. These metrics combined to provide a multi-dimensional assessment of data quality.

4. Analysis and Empirical Evaluation

4.1 Impact of Structured Protocols on Taxonomic Accuracy

The empirical results demonstrated a significant, non-linear relationship between protocol structure and taxonomic accuracy. Analysis of over twelve thousand submitted observations revealed that Tier 3 (Highly Structured Protocols) yielded the lowest overall Taxonomic Error Rate, hovering around five percent. However, this high accuracy came at the expense of participant retention, with Tier 3 experiencing an attrition rate of nearly forty percent over the six-month study period. Conversely, Tier 1 (Unstructured Protocols) maintained a high retention rate but generated a high Taxonomic Error Rate of twenty-eight percent, largely driven by the misidentification of cryptic insect taxa [20].

Interestingly, Tier 2 (Semi-Structured Protocols) demonstrated a highly resilient middle ground, achieving an error rate of approximately twelve percent while maintaining participant retention levels comparable to Tier 1. The requirement of photographic evidence within Tier 2 acted as a natural filter, as volunteers self-censored ambiguous observations when they were unable to capture clear photographic documentation. This finding suggests that moderate protocol constraints, when combined with accessible validation mechanisms, can maximize the volume of high-quality data.

4.2 Dynamics of Knowledge Diffusion and Error Reduction

When analyzing the influence of knowledge diffusion channels, the data revealed that peer-to-peer mentoring networks were overwhelmingly superior to passive digital training in reducing taxonomic errors over time. Volunteers who participated in the mentoring program exhibited a steep learning curve, reducing their taxonomic error rate from an initial twenty-five percent to under eight percent within four weeks of active participation

[21]. This rapid skill acquisition was attributed to the transfer of tacit taxonomic knowledge, such as subtle behavioral cues and micro-habitat associations, which are difficult to communicate through static digital interfaces.

Digital training modules, while cost-effective and highly scalable, showed a much slower rate of error reduction, requiring up to twelve weeks for participants to achieve comparable levels of accuracy. Real-time algorithmic feedback, however, performed exceptionally well in minimizing spatial and temporal anomalies, immediately correcting GPS offset errors and out-of-season species reports before they entered the database [22]. These results highlight the necessity of combining human-centric learning networks with automated technical safeguards.

Table 2: Quantitative Performance of Knowledge Diffusion Channels in Minimizing Identification Errors

Diffusion Channel	Taxonomic Error Rate	Spatial Deviation Distance	Mean Learning Curve Velocity
Algorithmic Feedback	0.14	12.4 meters	0.35 units per survey
Peer-to-Peer Mentoring	0.08	8.2 meters	0.68 units per survey
Digital Training Modules	0.19	15.1 meters	0.22 units per survey

4.3 Spatial-Temporal Validity of Collected Data

Spatial analysis of the collected datasets revealed persistent clustering patterns across all tiers, though the severity of this spatial bias varied significantly. Tier 1 observations were highly concentrated within major municipal parks and accessible transport corridors, illustrating the classic citizen science bias toward high-amenity urban zones [23]. In contrast, Tier 3 protocols, which dictated randomized transects, successfully dispersed volunteer effort into residential zones, industrial margins, and fragmented urban woodlands, capturing a more representative sample of the urban ecological matrix.

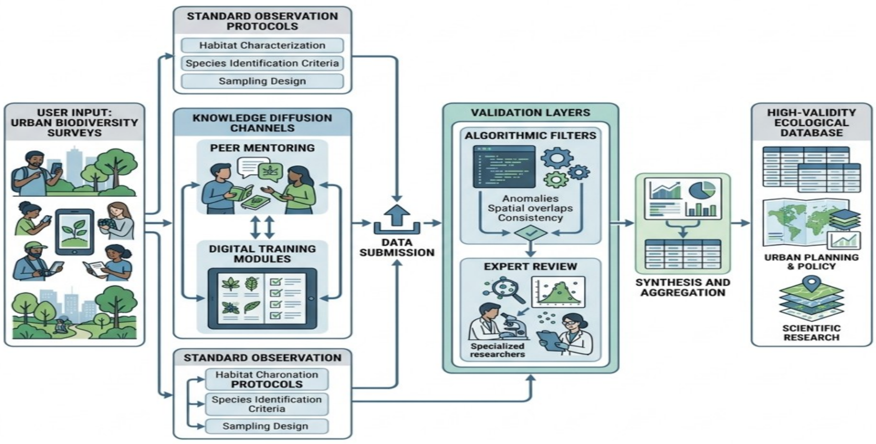

Figure 1: Conceptual Architecture of Knowledge Diffusion and Data Validation Framework in Urban Biodiversity Surveys

Temporally, the integration of peer mentoring successfully mitigated the weekend bias prevalent in crowdsourced data. Peer groups coordinated mid-week observation schedules, ensuring a more continuous temporal sampling distribution. This temporal stability is critical for capturing short-term ecological responses to episodic urban events, such as acute heatwaves or high-pollution days, which might otherwise occur unrecorded between weekend-only volunteer activities [24].

5. Discussion and Implications for Urban Conservation

5.1 Synthesis of Protocol Rigor and Volunteer Engagement

The results of this study suggest that data validity in citizen science should not be viewed as a static consequence of protocol design, but rather as a dynamic quality that co-evolves with the volunteer community. While highly structured protocols ensure immediate data accuracy, they risk suffocating the grass-roots enthusiasm that powers citizen science initiatives. Conversely, purely unstructured programs capture extensive public attention but risk overwhelming scientific databases with low-quality, unusable information [25].

The optimal path forward lies in designing adaptive protocol frameworks that transition volunteers through a tiered progression as their skills develop. Novices can enter the program through unstructured, low-barrier portals, where their primary contribution is exploratory. As they engage with peer-to-peer mentoring and digital learning channels, they can be progressively certified to execute more complex, structured protocols. This developmental pathway preserves high overall participation rates while systematically expanding the cohort of highly trained, reliable observers capable of executing rigorous scientific surveys.

5.2 Policy Implications for Urban Environmental Planning

For municipal planners and environmental policy makers, the validation of citizen science data opens up new possibilities for evidence-based urban design. High-resolution, validated biodiversity data can directly inform the designation of urban ecological corridors, the management of invasive species, and the assessment of green infrastructure performance. By integrating validated crowdsourced data into municipal geographic information systems, cities can respond dynamically to ecological changes, optimizing resource allocation for habitat restoration and biodiversity conservation.

Furthermore, leveraging citizen science democratizes the urban planning process. By involving citizens directly in the collection and validation of ecological data, municipalities build civic ownership over local green spaces, reducing public resistance to conservation-oriented zoning laws and urban redevelopment plans. This collaborative model transforms passive urban residents into active co-creators of their local environment, bridging the gap between scientific expertise and local community needs.

5.3 Limitations and Directions for Future Inquiry

Despite the positive outcomes of this study, several limitations must be acknowledged. First, the demographic composition of our volunteer cohort remained skewed toward individuals with pre-existing interests in science and environmental conservation, which may limit the generalizability of our learning curve models to broader public demographics. Future research should explore strategies for engaging underrepresented urban communities, tailoring knowledge diffusion channels to cultural and socio-economic contexts.

Second, the geographic scope of this study was restricted to a single metropolitan area, meaning that the specific spatial and taxonomic biases identified may vary in cities with different urban morphologies or ecological characteristics. Comparative studies across diverse global cities are needed to establish more universally applicable data validation standards. Finally, as artificial intelligence tools for automated species identification continue to advance, future research must examine how the integration of computer vision interfaces alters the cognitive learning processes and social dynamics of human-led volunteer networks. Concluding this analysis, the synthesis of structured protocol pathways and collaborative knowledge sharing networks offers a robust, scalable foundation for securing the long-term integrity of citizen-driven urban environmental science.

References

1. LeRoux, K.; Carr, J.B. Prospects for centralizing services in an urban county: Evidence from eight self-organized networks of local public services. J. Urban Aff. 2010, 32, 449–470.

2. Lu, M.; Taiebat, M.; Xu, M.; Hsu, S.C. Multiagent spatial simulation of autonomous taxis for urban commute: Travel economics and environmental impacts. J. Urban Plan. Dev. 2018, 144, 04018033.

3. Water Footprint Network (WFN). 2025. Available online: https://www.waterfootprint.org (accessed on 21 November 2025).

4. Kitchenham, B. Procedures for Performing Systematic Reviews; Keele University: Keele, UK, 2004; Volume 33.

5. Caravale, A. (2025). Diamond open access and research infrastructures: The involvement of 《archeologia E Calcolatori》 in the H2iosc project. Archeologia E Calcolatori, 36(1), 435–442.

6. Mahmassani, H.S.; Jayakrishnan, R. System performance and user response under real-time information. Transp. Res. Part A Policy

Pract. 1991, 25, 293–307.

7. Dall'erba, S.; Le Gallo, J. Regional convergence and the impact of European structural funds over 1989–1999: A spatial econometric analysis. Pap. Reg. Sci. 2008, 87, 219–244.

8. Wang, W.; Chen, Q.; Shen, Y.; Xiang, Z. Leakage Identification of Underground Structures Using Classification Deep Neural Networks and Transfer Learning. Sensors 2024, 24, 5569.

9. Bokobza, Y.; Dabora, R.; Cohen, K. Deep Reinforcement Learning for Simultaneous Sensing and Channel Access in Cognitive Networks. IEEE Trans. Wirel. Commun. 2023, 22, 4930–4946.

10. Guennoun, A., Bensaadi, K., Simard, M.-A., Schwartz, R., Murad, L., Leveridge, M., Siemens, R., Larivière, V., & Bhojani, N. (2024). PD60-01 open access publishing in urology: A survey of authors, readers, and editorial boards knowledge, impressions, and satisfaction. The Journal of Urology, 211, e1273.

11. Subaveerapandiyan, A., Seifi, L., Shimray, S. R., & Ahmad, N. (2025). Awareness and perception of diamond open access among university professors in Iran. Journal of Librarianship and Information Science, 58(1), 644–664.

12. Rogers, E.M. Diffusion of Innovations, 5th ed.; Free Press: New York, NY, USA, 2003.

13. Rausand, M. Risk Assessment: Theory, Methods, and Applications; Chapter 1: Introduction; John Wiley & Sons: Hoboken, NJ, USA, 2013; pp. 1–28.

14. He, H.; Yang, X.; Mi, X.; Shen, H.; Liao, X. Multi-Agent Deep Reinforcement Learning Based Dynamic Task Offloading in a Device-to-Device Mobile-Edge Computing Network to Minimize Average Task Delay with Deadline Constraints. Sensors 2024, 24, 5141.

15. Odum, H.T. Environmental Accounting. Emergy and Environmental Decision Making; John Wiley and Sons Inc.: New York, NY, USA, 1996; p. 370.

16. Al-Fuqaha, A.; Guizani, M.; Mohammadi, M.; Aledhari, M.; Ayyash, M. Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications. IEEE Commun. Surv. Tutor. 2015, 17, 2347–2376.

17. Breidenbach, P.; Mitze, T.; Schmidt, C.M. EU Regional Policy and the Neighbour's Curse: Analyzing the Income Convergence Effects of ESIF Funding in the Presence of Spatial Spillovers. JCMS J. Common Mark. Stud. 2018, 57, 388–405.

18. Li, S.; Huang, Q.; Chen, H.; Jiang, R.; Dong, M.; Ota, K.; Quek, T.Q.S. DRL-Based Joint Task Offloading and Resource Allocation in Air-Ground Integrated Vehicular Edge Computing Network with Energy Harvesting. IEEE Trans. Veh. Technol. 2025, 75, 6658–6672.

19. Wang, Y.; Lei, J.; Shang, F.; Li, Y. A Comprehensive Survey of Multi-Agent Deep Reinforcement Learning for Wireless Spectrum Management. Neurocomputing 2025, 653, 131236.

20. Treiber, M.; Hennecke, A.; Helbing, D. Congested traffic states in empirical observations and microscopic simulations. Phys. Rev. E 2000, 62, 1805–1824.

21. Xu, Y.; Zhu, K.; Xu, H.; Ji, J. Deep Reinforcement Learning for Multi-Objective Resource Allocation in Multi-Platoon Cooperative Vehicular Networks. IEEE Trans. Wirel. Commun. 2023, 22, 6185–6198.

22. Ahn, K.; Rakha, H.; Trani, A.; Van Aerde, M. Estimating vehicle fuel consumption and emissions based on instantaneous speed and acceleration levels. J. Transp. Eng. 2002, 128, 182–190.

23. Zeng, J.; Li, Z.; Feiock, R.C. Municipal sustainability priorities and the environmental-economy nexus. J. Urban Aff. 2025, 1–17.

24. Hashem, I.A.T.; Chang, V.; Anuar, N.B.; Adewole, K.; Yaqoob, I.; Gani, A.; Ahmed, E.; Chiroma, H. The Role of Big Data in Smart City. Int. J. Inf. Manag. 2016, 36, 748–758.

25. Arbolino, R.; Boffardi, R. The Impact of Institutional Quality and Efficient Cohesion Investments on Economic Growth Evidence from Italian Regions. Sustainability 2017, 9, 1432.