Big Data in Biodiversity Science

Subjects: Environmental Sciences | Agriculture, Dairy & Animal Science
Contributor: Adekunle Adebowale, Muxe Gladmond Dlomu

Biodiversity refers to the variety of genes, species and ecosystems of life on Earth, and is the source of many essential goods and services (e.g., food, timber, medicine, nutrient recycling, crop pollination) that support human well-being and quality of life. Despite several international treaties, efforts and commitments to curb its loss, biodiversity continues to decline at a rate above species discovery rate, largely due to anthropogenic factors. To assess the status and trends (local and global) in biodiversity requires a vast amount of relevant information on the distribution and abundance of different species across varying spatial and temporal scales. In other words, relevant data need to be collected, collated, and analyzed.

Keywords: big data ; biodiversity ; data curation ; data generation ; cyber infrastructure ; data access ; science communication

1. Overview

Despite best efforts, the loss of biodiversity has continued at a pace that constitutes a major threat to the efficient functioning of ecosystems. Curbing the loss of biodiversity and assessing its local and global trends requires a vast amount of datasets from a variety of sources. Although the means for generating, aggregating, and analyzing big datasets to inform policies are now within the reach of the scientific community, the data-driven nature of a complex multidisciplinary field such as biodiversity science necessitates an overarching framework for engagement. In this review, we propose such a schematic based on the life cycle of data to interrogate the science. The framework considers data generation and collection, storage and curation, access and analysis, and finally, communication as distinct yet interdependent themes for engaging biodiversity science for the purpose of making evidence-based decisions.

2. Biodiversity

Biodiversity refers to the variety of genes, species, and ecosystems of life on Earth, and is the source of many essential goods and services (e.g., food, timber, medicine, nutrient recycling, crop pollination) that support human well-being and quality of life [1]. Despite several international treaties, efforts, and commitments to curb its loss, biodiversity continues to decline at a rate above species discovery rate, largely due to anthropogenic factors [2]. To assess the status and trends (local and global) in biodiversity requires a vast amount of relevant information on the distribution and abundance of different species across varying spatial and temporal scales [3]. In other words, relevant data need to be collected, collated, and analyzed.

The last two and half decades have witnessed an exponential increase in the generation and analysis of data in virtually all domains of human engagement such that the term ‘big data’ was coined to distinguish the data explosion era from what went on before [4][5]. Scholz (2017) [6] tracked the origin of the term to the 1960s and 1970s and summarized its appearances in documents from the US Congress publications to various academic and non-academic works spanning a period from 1961 through to 1979. These early usages had little bearing on how it is conceived today. In its more contemporary form, several authors, for example, [5][7], have traced the emergence of the term from the world of commerce, whose main interest in big data was, and still is, driven by the need to monitor and improve performance. The concept has since spread to several areas of endeavor including, but not limited to, the healthcare industry, the agricultural industry, the education industry, the media sector, governance, the banking and finance sector, astronomy, climate change, and biodiversity management. As with concepts of such diverse application, and to which several distinct domains can lay claim, there is no universally satisfactory definition of big data. However, there is a consensus as to the key elements of its essence: big data is characterized by the three Vs of huge volume, high velocity, and diverse variety [5][8]. The volume component refers to the size of data generated, considered in petabytes or higher units of data; the velocity component suggests a rate of generation that is real-time or nearly so, thereby contributing to the huge volume;

and the variety indicates a mixture of structured, semi-structured and unstructured pieces of information [4][9]. Two other possible Vs, veracity and variability, are sometimes included.

Within the context of biodiversity, big data is defined as a “techno-political tool to manage the distribution of biological species”, and as “the intensive data accumulation of digitized information on biodiversity, corresponding to a spatial and temporal description of species distribution” [10]. While these rather similar definitions are limited in their scope, because they ignore some other aspects embedded in biodiversity [1], the first part, nevertheless, provides a historical anchor for situating the deliberate integration of big data and biodiversity within a techno-political agenda. This agenda, which could be viewed in simple terms as the implementation of policy supported by an evidence base, started in the mid-1960s [11], much in the tradition of the data-intensive research of the physical sciences (e.g., The Manhattan Project). Big data is central to biodiversity science because, at its barebone level, biodiversity involves species and their distributions across space and time. For instance, 36.5% of global plant species are considered “exceedingly rare” [12], suggesting a need for conservation planning to, at least, take such metrics into account. For our purpose, we assume the view that Biodiversity Big Data (BBD) as a concept encompasses a cyclical scheme that involves the generation, curation, processing, analysis, and communication of biodiversity information, at huge volumes and diverse varieties, with the purpose of making an informed decision for biodiversity management.

The emergence of big data (BD) as a discipline has raised some philosophical questions, challenging established ways of knowing in various domains of knowledge, including biodiversity science. Some advocates of BD [13][14] have been quick to declare the end of theory; they oppose the need for model building or hypothesis formulation due to the power of data analytics to detect patterns and provide new insights from big data independent of human bias. This view has been robustly contested in the BD literature and is shown to be based on fallacious thinking, whilst recognizing the inherent potential of analyzing vast amount of data. Kitchin (2013) [4] and others [5][12][15][16] have shown that BD, however exhaustive, is still representational (a sample) and is therefore subject to the vagaries of sampling bias. Data collection and analysis are shaped by the theories underpinning the systems of collection and the algorithms of analytics. The emergent patterns are, thus, not free of human bias as they are interpreted within frameworks. In addition, there is the real possibility of random correlations between variables with no underlying causal linkage. Succi and Coveney (2019) [17] suggest that the pattern recognition power of BD analysis could provide a basis for further engaging theories in making sense of the patterns that would be otherwise undiscernible to the human mind. One application of BD raised in the work [17], and which is relevant to biodiversity science, is its ability to handle some of the sensitive aspects of non-linearity or chaos found in many complex systems [18]; a concept that underlies Spatio-temporal organization and weather events and is best encapsulated by the popular phrase ‘the butterfly effect'.

In addressing “the datafication of biodiversity” [10], it was convincingly demonstrated that the process of transforming ecological and other records of living forms into biodiversity data not only changes the nature of the information, it also corresponds to a politically-driven shift in priorities for ecological research from local concerns to a global outlook, resulting in the birth of global biodiversity. The key element was to underpin sustainability policy with a strong evidence base. The authors highlighted the positive role played by the creation of the global biodiversity information facility (GBIF) —one of the largest biodiversity databases in the world—in bridging the divide between science and politics for the global good. The aim of GBIF was to facilitate the translation of good science to good government policy [19]. While this datafication process provided one approach to viewing the global environmental landscape and developing some of the tools for effective monitoring, it nevertheless came at the expense of biological context. As argued by Bowker (2000) [20], BBD production often results in the loss of ecological meaning as species become disconnected from their ecological context in the process of achieving uniformity and compatibility of data format in a single database. This poses a peculiar danger of database creation becoming an end in itself [20]. A similar line of reasoning was extended further by detailing how the real-world ecological niche of various organisms, captured by numerous information records, are reduced to a two-dimensional world of “rows and columns” [10], thereby creating an artificial data niche detached from the biophysical realities of the organisms supposedly represented. This is what was construed as the datafication of ecological records [10].

The discourse around the emergence of BD in biodiversity would be incomplete without consideration for the infrastructures that make it possible to generate, store and analyze BD. These infrastructures vary from instruments capable of recording tens of petabytes of information (e.g., radio-telescopes) to next-generation sequencers for sequencing whole genomes (about 3 billion nucleotide base-pairs in one human being for instance), to remote sensing devices for collecting vast amount of environmental data. In addition to these are the rapidly increasing computer storage capabilities, including storage in the clouds, the increasing computational power of PCs, coupled with innovations in statistical computing that allow new ways to analyze and visualize BD.

3. Conclusions

The continuous loss of biodiversity affects ecosystem functioning, of which humans are an integral part. To stem the tide, evidence-based decision-making processes should become the normative mode of operation. This is only possible on the back of adequate and quality data that is well-analyzed and accurately interpreted. This review presents the data life cycle as an umbrella framework for critically engaging the subject of big data in biodiversity science with the goal of making informed decisions in biodiversity management. Although we present the framework in what appears to be a logical flow starting from data generation, through storage, to analysis, and finally to communication, any of the themes could, arguably, be a starting point for engagement depending on context. The themes and associated sub-themes are all interlinked and dependent on each, and not necessarily in the neat order we have arranged them. Data collection could be informed by the analysis of previously available datasets, which may identify specific data gaps. In turn, data analysis is underpinned by access to some sets of data in the first place. For informed policy decisions on biodiversity issues, the insight gained through analysis must be effectively communicated to stakeholders and policy makers. Infrastructural developments to drive innovative data collection, the storage of massive datasets, and the performance of relevant analyses are critical to the smooth operation of the scheme. The interlinked nature of the scheme suggests that there will be some element of redundancies for quality assurance. As summarized in **Figure 1**, such overlaps are reflected in the similarity of challenges and opportunities across some themes.

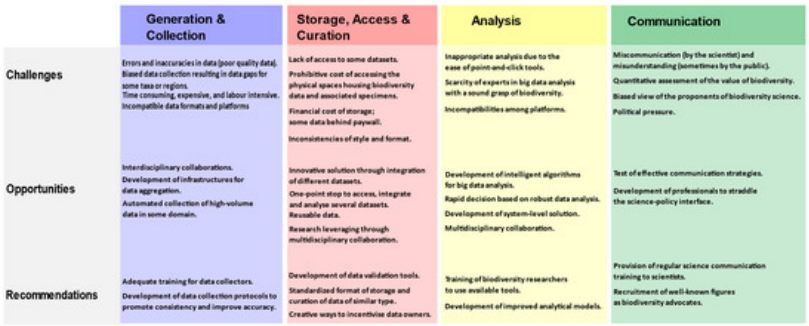

Figure 1. Summary of challenges and opportunities across BBD themes.

The cyclic nature of the scheme also connotes the potential reusability of biodiversity data. Indeed, this is a necessity due to the historical element inherent to biodiversity datasets, and the logistical and financial constraints of data collection. Because biodiversity scientists are usually directly involved in every theme of the scheme except, perhaps, for the policy formulation and decision-making phase, the need for deliberate constructive engagement between scientists and policy makers becomes non-negotiable. A good starting point for such engagement is the recognition by both sets of players that they belong in the same domain, even if their roles are different. Critical to those roles is good quality big data and what can be done with it.

References

1. Rands, M.R.W.; Adams, W.M.; Bennun, L.; Butchart, S.H.M.; Clements, A.; Coomes, D.; Entwistle, A.; Hodge, I.; Kapos, V.; Scharlemann, J.P.W. Biodiversity conservation: Challenges beyond 2010. Science 2010, 329, 1298–1303.
2. Dietz, S.; Adger, W.N. Economic growth, biodiversity loss and conservation effort. J. Environ. Manag. 2003, 68, 23–35.
3. Jetz, W.; McGeoch, M.A.; Guralnick, R.; Ferrier, S.; Beck, J.; Costello, M.J.; Fernandez, M.; Geller, G.N.; Keil, P.; Merow, C.; et al. Essential biodiversity variables for mapping and monitoring species populations. Nat. Ecol. Evol. 2019, 3, 539–551.
4. Kitchin, R. Big data and human geography: Opportunities, challenges and risks. Dialogues Hum. Geogr. 2013, 3, 262–267.
5. Kitchin, R. Big Data, new epistemologies and paradigm shifts. Big Data Soc. 2014, 1, 2053951714528481.
6. Scholz, T.M. Big Data in Organizations and the Role of Human Resource Management; Peter Lang International Academic Publishers: New York, NY, USA, 2017.
7. Diebold, F.X. A Personal Perspective on the Origin(s) and Development of ‘Big Data’: The Phenomenon, the Term, and the Discipline, Second Version (26 November 2012). PIER Working Paper No. 13-003. Available online: https://ssrn.com/abstract=2202843 (accessed on 5 May 2021).

8. Swan, M. Philosophy of big data: Expanding the human-data relation with big data science services. In Proceedings of the 2015 IEEE First International Conference on Big Data Computing Service and Applications 2015, Redwood City, CA, USA, 30 March–2 April 2015; pp. 468–477.
9. Boyd, D.; Crawford, K. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information. Commun. Soc. 2012, 15, 662–679.
10. Devictor, V.; Bensaude-Vincent, B. From ecological records to big data: The invention of global biodiversity. Hist. Philos. Life Sci. 2016, 38, 1–23.
11. Aronova, E.; Baker, K.S.; Oreskes, N. Big science and big data in biology: From the international geophysical year through the international biological program to the long term ecological research (LTER) Network, 1957—Present. Hist. Stud. Nat. Sci. 2010, 40, 183–224.
12. Enquest, J.B.; Feng, X.; Boyle, B.; Maitner, B.; Newman, E.A.; Jørgensen, P.M.; Roehrdanz, P.R.; Thiers, B.M.; Burger, J.R.; Corlett, R.T.; et al. The commonness of rarity: Global and future distribution of rarity across land plants. Sci. Adv. 2019, 5, eaaz0414.
13. Anderson, C. The end of theory: The data deluge makes the scientific method obsolete. Wired Mag. 2008, 16, 7–16.
14. Prensky, M.H. Sapiens digital: From digital immigrants and digital natives to digital wisdom. Innov. J. Online Educ. 2009, 5, EJ834248.
15. Amin, A.; Thrift, N. Cities: Reimagining the Urban; Polity Press: Cambridge, UK, 2002; p. 192.
16. Crawford, K. The hidden biases in big data. Harv. Bus. Rev. 2013, 1, 814.
17. Succi, S.; Coveney, P.V. Big data: The end of the scientific method? Philos. Trans. R. Soc. A 2019, 377, 20180145.
18. Bond, W.J.; Maze, K.; Desmet, P. Fire life histories and the seeds of chaos. Ecoscience 1995, 2, 252–260.
19. Kelmelis, J.A.; Snow, M. Proceedings of the US Geological Survey Global Change Research Forum; US Government Printing Office: Washington, DC, USA, 1993.
20. Bowker, G.C. Biodiversity datadiversity. Soc. Stud. Sci. 2000, 30, 643–683.