·技术 / TECHNOLOGY·

生物多样性数据 Gap 分析工作流

黄雄伟[1,2]，林聪田[1]，纪力强[1*]

1. 中国科学院动物研究所，北京 100101
2. 中国科学院大学，北京 100049

摘　要： 生物多样性科学研究、保护与持续利用离不开准确、详实的生物多样性分布数据。近年来，随着计算机信息技术与生物多样性信息学的发展，生物多样性数字化数据得到快速、大量积累。同时也难免存在一些问题，例如数据来源众多、标准不统一、质量参差不齐，影响其共享与充分利用，并且在类群覆盖、采样时间和地理分布上存在一定的空缺 (Gaps)，导致目前的生物多样性数据能否全面客观地描述生物多样性存在疑问。生物多样性数据 Gap 分析(BiodiversityDataGapAnalysis，BDGA)的目的就是对现有数据进行完整性和可靠性评估，为今后的生物多样性调查与研究工作提供决策支持，帮助相关研究人员在应用生物多样性数据时对数据客观性和可用性进行判断。本研究旨在探讨如何建立 BDGA 工作流，着重阐述 BDGA 的基本方法、相关工具与流程。

关键词： 生物多样性信息；数据集；数据评估；Gap 分析

doi: 10.11871/j.issn.1674-9480.2017.04.007

Workflow of Biodiversity Data Gap Analysis

Huang Xiongwei[1,2], Lin Congtian[1,2], Ji Liqiang[1*]

1. Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China
2. University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: The biodiversity scientific research, conservation and sustainable use increasingly depend on the accurate and detailed biodiversity distribution data. With the development of internet and information technology, large amounts of primary digital biodiversity data have been accumulatedon a global scale.However, the

基金项目：中国科学院信息化建设专项：(Y329C31114)“动物学科领域基础科学数据整合与集成应用”；国家科技部国家科技基础条件平台项目“国家基础科学数据共享服务平台”(DKA2017-12-02-10)

通讯作者：纪力强 (ji@ioz.ac.cn)

existing biodiversity data have some flaws, such as different sources of biodiversity dataresulting in the lack of unified standards, the poor data quality which affected the effectively sharing and utilization, and some gaps in the taxonomic, temporal and geographical coverage. These flaws obstruct the comprehensive and objective evaluation of biodiversity. Biodiversity Data Gap Analysis (BDGA) is to assess the completeness and reliability of available biodiversity data, which aims to provide decision supports for further biodiversity surveys and researches, and support data consumers with prior knowledge estimation. The purpose of this paper is to explore how to construct the BDGA workflow, which focuses on the basic methods, related tools and processes of BDGA.

Keywords: Biodiversityinformatics; dataset; data assessment; data gap analysis

引言

生物多样性对于人类具有十分重要的价值，是人类社会赖以生存和发展的基础，人类的生活与生物多样性的维持密切相关[1]。科学准确的生物多样性分布信息是生物多样性科学研究、保护与持续利用的重要基础[2-4]。信息技术的蓬勃发展使得生物多样性及其相关学科研究领域取得了长足进步，包括中国在内的各国的生物多样性信息管理系统建设工作自上个世纪九十年代开展以来，至今已取得了丰硕成果[5]，全球和区域水平的生物多样性数据得到大量积累，生物多样性数据库不断建立并得到完善[6]。国际上出现了如全球生物多样性信息机构 (Global Biodiversity Information Facility，GBIF)、网络生命大百科 (Encyclopedia of Life，EOL)、全球生物物种名录 (Catalogue of Life, CoL) 等，国内也建设了国家标本资源共享平台 (National Specimen Information Infrastructure, NSII)、中国动物主题库、中国生物物种名录等一系列生物多样性数据源。

但是，现存的生物多样性数据存在着一些问题，例如本底不清，数据缺失严重。主要原因是过去的调研工作主要致力于生物多样性的组成，缺乏详实的分布数据[1]。纵观国际，与欧美等科学强国相比，我国在生物多样性的分类、地理分布和时间序列等方面数据的完整性、精确性和可靠性还有很大的提升空间[6]。现存的生物多样性数据能否全面客观描述和评估中国的生物多样性存在疑问，从而影响后续利用。生物多样性数据 Gap 分析 (BiodiversityDataGapAnalysis, BDGA) 的主要目的就是评估生物多样性数据是否满足用户的使用需求，在没有足够数据的情况下，使得数据的使用者注意到数据本身是否存在偏差，这样的偏差是否影响数据的有效利用，避免造成决策失误，浪费宝贵而紧缺的生物多样性保护资源[7]。关于 BDGA 的系统性研究近几年才刚刚起步，相应的评估方法、流程和工具还比较少，有待于进一步开发。

北美和欧洲生物多样性数字化工作起步较早，积累的生物多样性数据较丰富，具备开展 BDGA 工作的条件。Condé 等 (1995) 针对欧洲生物多样性数据源进行了较简单的 Gap 分析，彼时数据库较少并且没有连接，也没有数据交换标准，主要通过调研和目录数据分析，发现 gap 主要存在数据的定量上，并且几乎没有无脊椎动物数据[8]。Soberón 等 (1996) 对墨西哥生物多样性进行 BDGA，主要从世界各大博物馆中获取数据，发现“收藏家综合征 (collector syndrome)”，即标本采集地趋向于聚集在道路附近或者在相关科研院所附近，说明采样站点的可及性是影响数据分布的重要因素[9]，类似的研究结果还有“公路效应 (highway effect)”[10] 和“公路图效应 (road-map effect)”[11]。

随着信息技术水平的发展，数据库成为了 BDGA 的主要数据来源之一，例如 Chavan 等 (2010) 对 GBIF 数据库的数据进行了完整性统计评估，发现数据在空间、时间和类群覆盖度上有很大偏差，北半球的数据几乎集中在北美和欧洲，无脊椎动物和无花植物覆盖度很低，有大量数据没有时间信息[12]。Gaiji 等 (2013) 重新评估了 GBIF 数据的完整性和覆盖度，运用了

Hadoop 和 Hive 表等数据库分析和可视化方法[13]。研究显示此时的类群完整性相比 2010 年有所提高，类群覆盖维度的数据 Gap 可能正在被填补，但仍然有大量数据没有时间信息，影响了超过 1/3 的数据[14]。Wetzel 等 (2014) 详细分析了欧洲生物多样性数据库的 Gaps，并将分析范围扩展到全球数据库。该项目系统地评估了空间、时间、类群上的 Gaps，以及数据可访问性、趋势和数据质量等因素，发现类群覆盖度上存在明显不足，主要在特定区域的目标物种上 (如传粉者)，空间覆盖度上东欧明显不足[15]。该研究强调了“未知数据 (dark data)”[16] 的重要性，发现有 2/3 的数据存在访问障碍，形成“假 Gaps”。

技术进一步发展与成熟以后，BDGA 趋向更加复杂与精细化。从单维度的类群、地理的 BDGA 到多维度的类群和时空分布数据的综合分析。类群方面，从分析单一动植物类群发展到多类群；时间序列方面，倾向于挖掘全时间序列的数据；地区方面，出现从单一国家或区域到面向全球的分析[17-19]。“大数据时代”的到来使得 BDGA 朝着数据更加完整精细的方向发展，可加入分析的环境因素更为丰富，分析算法技术角度也更加多样。例如 Meyer 等 (2015) 整合了 GBIF 数据库中全球共 21,170 种脊椎动物共 1.57 亿个分布记录数据进行了 BDGA，该研究使用了四种不同的分辨率进行地图映射，发现在全球范围内数字化的生物多样性信息在许多大型新兴经济体中比热带地区物种丰富的发展中国家更加不足[17]。但这项研究的数据来源单一，其它来源的可用数据可能会影响其分析结果。

国内有以植物为主的 BDGA 研究发表，如 Yang 等 (2013) 对中国 2370 个县的维管植物分布进行了 BDGA，表明中国 91% 的县级行政区采样强度还远远不够，存在着较大 Gaps，同时发现海拔和年度降雨天数最适合用于预测物种丰富度，但这种预测的可信度受到数据完整性的强烈影响[20]。Yang 等 (2014) 又进一步对中国维管植物数据进行了 BDGA，分析了生物多样性信息的采样强度与可及性、人口密度等六种社会经济因素的相关性，以了解现有数据的潜在偏差，结果发现海拔高度对采样密度有积极效应，而采样强度在人口稠密地区反而不足[21]。针对于中国动物分布数据的 BDGA，尚未见有系统报道。

进行 BDGA 方法论的研究和相关工具的开发，对生物多样性数据的完整性和可靠性进行评估，是目前生物多样性信息学研究的重要课题之一。通过 BDGA，了解中国乃至全球现有数据的完整性和可靠性，可为今后的生物多样性调查与研究工作提供决策支持，帮助相关研究人员等应用生物多样性数据时进行数据客观性和可用性判断，为决策者制定生物多样性保护的相关政策提供参考和依据。

1 BDGA 工作流

BDGA 工作流可分为数据整合、Gap 分析和结果及评价三个部分，每个部分的相应流程如图 1 所示。

1.1 范围与期望设定

进行 Gap 分析之前，应首先根据研究目的确定 BDGA 的范围，并设定期望。可以从区域范围、时间跨度以及目标类群三个方面确定 BDGA 的范围。区域范围最大可达到全球，也可以是大洲、国家和地区，甚至是某一个自然保护区。物种分布数据的产生时间对于数据质量与获取难度具有很大影响，因此必须确定一个时间跨度，以确定获取的数据符合分析目标的要求。另外，还必须根据计划确定研究的目标物种或类群。数据期望需要符合实际情况，因为数据可访问性受到许多障碍的限制。例如数据集过时，从业人员或者研究兴趣导致的数据偏差，或者地理覆盖导致的数据集的不一致[22]。

1.2 数据源确定、可用性分析及预处理

用于 BDGA 的物种分布数据主要有两种类型：一类是基于野外调查或者标本采集地点的经纬度数据形成的物种分布点记录，是特定物种在特定时间点出现在特定地理位置的直接证据。另一类数据是物种分布范围地图，此类型数据是在物种分布点记录的基础上，同时考虑物种的适宜生境等不同来源的数据，结合各领域专家的研究成果确定的物种分布信息，最后

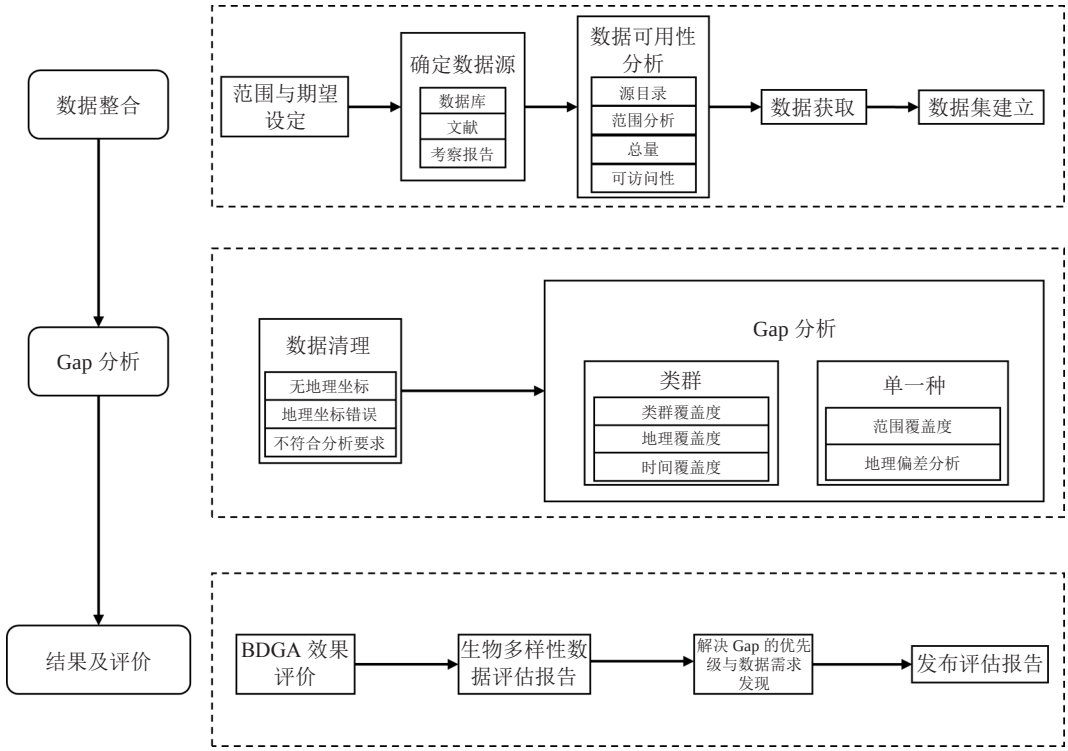

图1　BDGA 工作流示意图

Fig. 1　Workflow of BDGA

形成物种分布范围多边形地图。另外还可从文献、考察报告等未数字化的资源中获得大量的物种分布信息，有条件的情况下可将这些数据整理并加入分析。

根据研究策略与计划确定物种分布的数据源以后，需要进行数据可用性分析。数据可用性分析主要从以下几个方面入手：数据源目录、数据覆盖范围、数据总量和可访问性 [7]。通过直接下载或者数据服务接口技术来获取数据，并利用数据库相关技术来整合各种来源的生物多样性数据，建立 BDGA 数据集。

获取到的物种分布原始数据在进行 Gap 分析之前应先进行数据清理，剔除明显无法用于后续分析的数据。例如部分没有可靠的完整学名或者物种分类还存在争议的记录；部分没有标明数据来源的记录；另有一些数据获取时间太早 (如在 1850 年以前)，可信度极低，应予以剔除；数据缺少地理坐标或者地理坐标错误，并且没有相关可地标化的地理信息的；含有其他错误或者不符合分析要求的数据。随后将符合分析条件的数据整合并通过 GIS 分析软件形成栅格化地图并用于后续分析。

1.3　Gap 分析

Gap 分析一般针对两个不同方面进行，一种是针对类群，另一种是针对单个物种。针对类群进行分析，主要着眼于分析物种分布数据的类群、地理和时间覆盖度 (图 2) [19]。而针对单个物种进行分析，则主要着眼于单个物种分布数据的范围覆盖程度以及数据存在的地理偏差估计 (图 3) [18]。

1.3.1　针对类群的 Gap 分析

按类群进行的 Gap 分析主要评估生物多样性数据的三个维度——即类群覆盖度、地理覆盖度和时间覆盖度 [19]。类群覆盖度，即有多少不同类群的现有物种被记录过，各区域之间的生物多样性状况被记录的详实程度 [23-24]；地理覆盖度，即特定地理区域被物种分布记录的覆盖程度，影响物种分布模型的可行性与可靠性 [25-26]；时间覆盖度，即物种记录在时间序列上的连续程度，这对于监测物种对环境变化的响应是十分必要的信息 [27]。

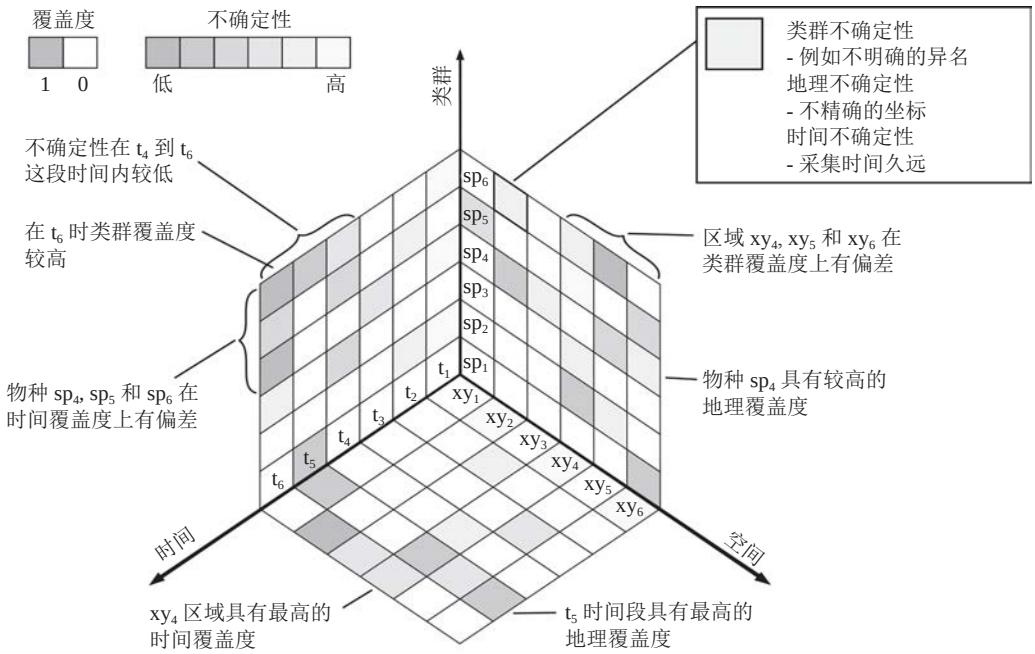

图2　针对类群的 Gap 分析的分析框架[18]

Fig. 2　Framework of DGA for taxonomic group

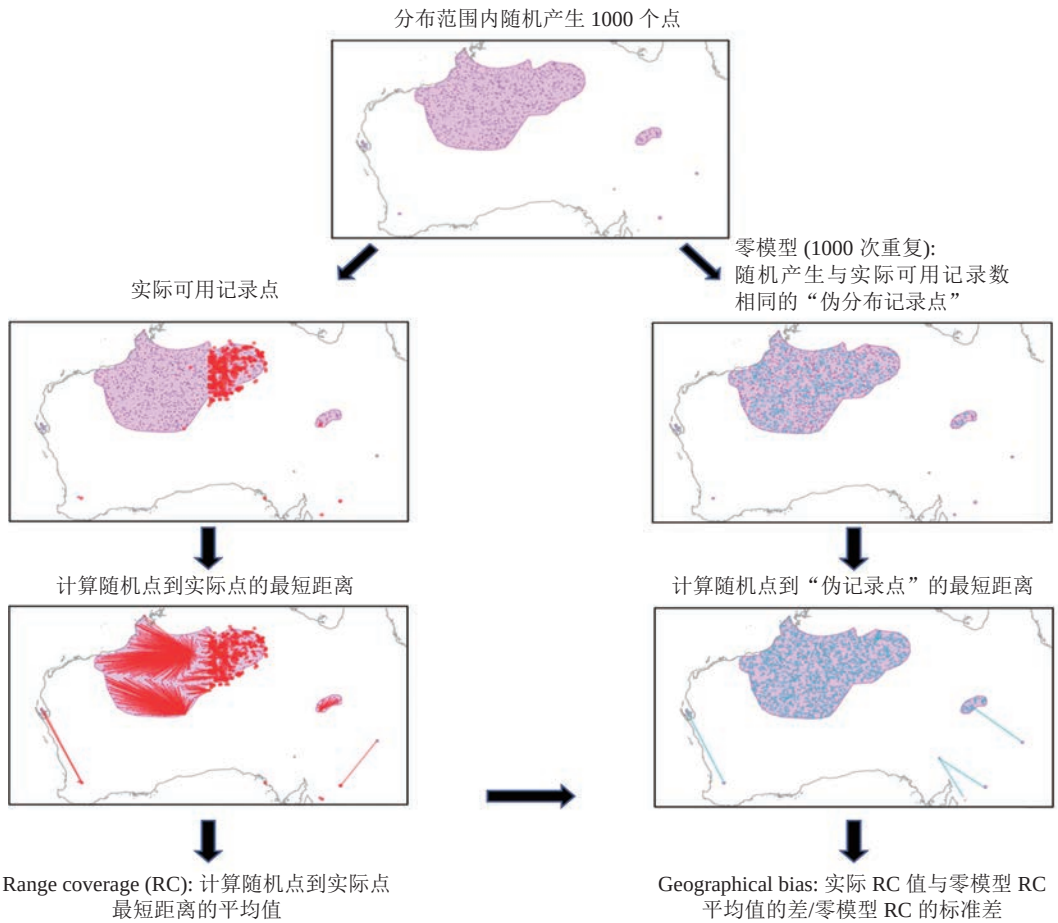

图3　针对单个物种的 Gap 分析流程[18]

Fig. 3　Analysis flow of DGA for in dividual species

类群覆盖度：关于类群覆盖度的研究已有不少，例如“census-”，“inventory-”或“survey completeness”[28]。可从三个方面评估类群覆盖度，一是在有权威的生物物种名录的地区，将具有物种分布数据记录的物种或类群与标准名录进行对比分析来估计类群覆盖度；二是利用物种分布模型，在栅格地图中表示出实际记录的物种丰富度与模型估算得到的物种丰富度的比值，再通过计算 Moran’s I 等指数，估计各区域类群覆盖的均匀程度[29-30]。另外可通过时间累积曲线反映数据的类群覆盖度在时间序列上的变化情况[20, 31]。在没有独立的基准信息的情况下，大多数估计类群覆盖度的研究均是基于现有的可用物种分布点记录数据,计算相应的指数以估计类群覆盖度。

地理覆盖度：可以从三个方面分析数据的地理覆盖度，从整体上看，可评估采样点数和物种数的关系，如果具有极少采样点的物种数极多，或者只有极少数的物种采样点数量很多，可说明数据存在较大偏差。另外可从给定区域栅格图的每个栅格所包含的采样点数量，通过计算 Moran’s I 评估采样点的空间聚集程度，可以了解物种分布数据在地理上的分布均匀程度。还可通过时间累积曲线计算每个栅格被采样点覆盖的百分比，了解物种分布数据在时间维度上的地理覆盖度。

时间覆盖度：对每个物种或者每个栅格计算数据采集时间间隔的平均年份数的负值 (精确到月份)。这个指标表示给定的点在时间上的间隔，若该值的绝对值越大表示时间覆盖度低，也即表示在相当大的时间间隔里没有任何数据，存在较大的时间序列上的 Gaps。从两个方面评估数据时间覆盖度侧重点有所不同，对每个物种均计算时间覆盖指数，可以了解数据整体的时间覆盖度，若大量物种的时间覆盖指数偏低，则说明数据在时间覆盖度上存在较大偏差；对给定区域的每个栅格计算时间覆盖度，如果区域的时间覆盖指数偏低，则说明该区域的数据在时间覆盖度上存在较大偏差。

1.3.2 针对单个物种的 Gap 分析

除了简单的记录计数以外，范围覆盖度 (Range coverage，RC) 和地理偏差值 (Geographical bias，GB) 是可用于评估物种分布数据偏差的两个指标[18]。这两个指标在计算时使用物种分布范围地图为基准数据，用于评估特定物种分布点记录的 Gaps。其中，范围覆盖度是用于衡量物种的分布范围被可用的分布点记录所覆盖的程度；地理偏差值则描述了该物种分布点记录在其分布范围内的非随机性水平，二者结合可较为清晰地表示出单一物种分布点记录的 Gap 大小。由于这两个指标均是以物种分布范围地图为基准数据，因此分布范围地图的准确性将极大影响指标的可信度。并且这两个指标的隐含假设是对于一个给定物种，在可用记录的时间和分布范围内，在技术上至少能被记录一次。然而现实中，分布范围地图在精细的空间分辨率下倾向于高估物种的分布区域 (即包含有物种从未被记录过的点)[32-33]。

范围覆盖度和地理偏差值的计算方法如下：

$$range\ coverage = -\mathrm{MMD} = -\frac{1}{1000}\sum_{i=1}^{1000}\mathrm{MinDistRP}_i$$

其中，MMD (mean minimum distance) 为 1000 个随机点和 n 个可用物种分布点记录之间最小距离的平均值，$\mathrm{MinDistRP}_i$ 表示第 i 个随机点到离其最近的物种分布点记录之间的最小距离。因此该绝对值越小表示范围覆盖度越高。

$$geographical\ bias = \frac{\mathrm{MMD}_{\mathrm{observed}} - \mathrm{mean}\ (\mathrm{MMD}_{\mathrm{null\ model}})}{\mathrm{SD}\ (\mathrm{MMD}_{\mathrm{null\ model}})}$$

地理偏差值 (geographical bias) 实际上是一个标准化效应指数，可用于定量表示在特定物种分布范围内可用的分布点记录的偏差程度，其值的计算联系了实际的 MMD 和在随机取样下的潜在 MMD 的零模型。具体计算方法是首先在目标物种的分布范围内随机取 n 个 (n 为实际可用记录数) “伪分布记录点 (pseudo-records)”，然后计算 MMD 值，重复 1000 次，取平均值和标准差。然后用实际的 MMD 减去上述平均值除以上述标准差后可得地理偏差值。该值越大表示实际采样点越集中于物种实际分布范围的某一部分，导致现有数据在地理上存在较大的偏差，数据过于集中，还需要进一步采样以减小偏差。

1.4 BDGA 效果评价

完成 BDGA 以后还应对其结果进行相应评估，以确认 BDGA 结果的可靠性，提出解决数据 Gap 的相应方法，并尝试提出未来的 BDGA 的可能改进方向。对应于 BDGA 分析的期望设定，BDGA 效果评价标准主要是分析是否达到了预期效果，分析的结果与目的的一致性。如果 BDGA 分析结果“符合预期”，则需要仔细验证，是否数据仍存在明显的 Gaps (例如在特定的时间下某些数据来自不可能获得数据的地区)。若 BDGA 分析结果“不符合预期”，则应进行后续分析，讨论原设定的期望是否难以实现，是否考虑到现有资源状况有无法克服的困难妨碍了 BDGA 取得适当结果，或者讨论已发表的 BDGA 研究是否遇到类似的问题。BDGA 效果评价将为未来的 BDGA 奠定基础，可能产生新的见解与帮助[7]。

2 BDGA 研究前景与展望

处于“大数据时代”的今天，生物多样性数据量越来越大，共享程度越来越高，是充分挖掘生物多样性数据应用价值的最好时机。但尽管近年来生物多样性数据的数字化与共享程度一直在加强，全球范围的生物多样性数据状况仍然在类群、地理分布和时间序列上体现出较大的偏差，严重影响使用这些数据进行的生物多样性研究和保育工作，因为大多数物种分布模型的方法对记录数与数据质量高度敏感[34]。更多地将国家与地区的数据源整合到全球的数据共享网络可以在一定程度上进行补偿，但研究表明这些数据源显示出了相似的偏差[20, 30]。例如 GBIF 中中国大陆部分的数据量较少，而实际上中国多年来通过野外调查、考察和监测已积累了大量生物多样性数据，在生物多样性数据的数字化和信息系统建设方面工作起步也较早[35-36]。然而，不同时期建立的生物多样性数据系统没有统一标准，各系统信息传递与整合困难，遑论有效管理与支持决策分析[5]。此外，我国生物多样性数据的质量与更新周期也有待提高[3]。另外，还有许多生物多样性分布数据尚未数字化，存在于各种文献著作中。可以预见，生物多样性数据在今后的数十年将仍然会存在严重的 Gaps[19]。鉴于为完成生物多样性保护的目标确保充足并持续的资金与资源支持存在相当的困难[37-39]，因此推进未数字化信息的数字化以及已数字化数据的共享是当前改善生物多样性数据状况应该予以协调与优先考虑的工作[17]。

现存的生物多样性数据数量少且质量不佳，因此需要利用 BDGA 在应用这些数据进行相关研究与制定政策时分析数据的不足之处，以及造成偏差的原因，分析数据在应用时会对结果产生什么样的影响 (例如数据偏差对进行物种分布模型预测的准确性影响程度)，以便尝试加以改进。另外需要开发更加严谨精确的方法或者评估指标，在基底数据不足的情况下，尽可能了解数据的可用程度，评估不同指标对生物多样性评价及物种分布预测的影响。这一方面能够有助于研究人员更精确地进行物种分布格局的预测，从而进行更精准的保育工作；一方面也能够为今后的生物多样性保育研究作指导，以免有限的资源被浪费在重复取样等方面。在数据逐步完善的过程中，BDGA 是一个需要定期持续进行的过程，其流程、方法和途径并不需要彻底更新。正相反，任何后续的 BDGA 应该从早期的 BDGA 中获取经验，即早期的 BDGA 结果是未来 BDGA 研究的基准。这就要求在进行新的 BDGA 之前，应该对以前的 BDGA 结果进行深入评估，以了解以往 BDGA 结果的积极或者消极的方面，有什么遗漏之处，并在进行后续的 BDGA 时加以纠正。这些都有助于 BDGA 成为生物多样性研究与保育工作的有效工具[7]。

参考文献

[1] 武建勇, 薛达元, 王爱华,等. 生物多样性重要区域识别——国外案例、国内研究进展[J]. 生态学报, 2016, 36(10):3108-3114.

[2] 纪力强. 生物多样性信息系统建设的现状及CBIS简介[J]. 生物多样性, 2000, 8(1):41-49.

[3] 黄晓磊, 乔格侠. 生物多样性数据共享和发表:进展和建议[J]. 生物多样性, 2014, 22(3):293-301.

[4] 马克平. 生物多样性信息学在中国快速发展[J]. 生物多

样性, 2014, 22(3):251-252.

[5] 戴小廷. 近二十年来生物多样性信息系统的研究进展[J]. 信息技术, 2012(6):55-59.

[6] 马克平. 亚洲植物多样性数字化计划[J]. 生物多样性, 2017, 25(1):1-2.

[7] Ariño A H, Chavan V, Otegui J. Best practice guide for data gap analysis for biodiversity stakeholders[R]. GBIF Secretariat, Copenhagen, Denmark, 2016.

[8] Condé, S, Roekaerts M, Vignault M P, et al. Databases on species, habitats and sites: survey and analysis 1995-96[C]. //Copenhagen: European Topic Centre on Nature Conservation, 1995.

[9] Soberón J, Llorente J, Benítez H. An International View of National Biological Surveys[J]. Annals of the Missouri Botanical Garden, 1996, 83(4): 562-573.

[10] Soberón J M, Llorente J B, Oñate L. The use of specimen-label databases for conservation purposes: an example using Mexican Papilionid and Pierid butterflies[J]. Biodiversity & Conservation, 2000, 9(10): 1441-1466.

[11] Crisp M D, Laffan S, Linder H P, et al. Endemism in the Australian Flora[J]. Journal of Biogeography, 2001, 28(2):183-198.

[12] Chavan V, Gaiji S, Hahn A, et al. State-of-the-Network 2010: Discovery and Publishing of Primary Biodiversity Data through the GBIF Network[R]. Copenhagen: Global Biodiversity Information Facility, 2010.

[13] Gaiji S, Chavan V, Ariño A H, et al. Content assessment of the primary biodiversity data published through GBIF network: Status, Challenges and Potentials[J]. Biodiversity Informatics, 2013, 2(2): 94-172.

[14] Otegui J, Ariño A H, Encinas M A, et al. Assessing the Primary Data Hosted by the Spanish Node of the Global Biodiversity Information Facility(GBIF)[J]. Plos One, 2013, 8(1): e55144.

[15] Wetzel F, Hoffmann A, Kroupa A, et al. EU BON Deliverable 1.1: Gap analysis and priorities for filling identified gaps in data coverage and quality, Berlin, Germany, 2014.

[16] Heidorn P B. Shedding Light on the Dark Data in the Long Tail of Science[J]. Library Trends, 2008, 57(2):280-299.

[17] Meyer C, Kreft H, Guralnick R P, et al. Global priorities for an effective information basis of biodiversity distributions[J]. Nature Communications, 2015, 6: 8221.

[18] Meyer C, Jetz W, Guralnick R P, et al. Range geometry and socio‐economics dominate species‐level biases in occurrence information[J]. Global Ecology & Biogeography, 2016a, 25(10): 1181-1193.

[19] Meyer C, Weigelt P, Kreft H. Multidimensional biases, gaps and uncertainties in global plant occurrence information[J]. Ecology Letters, 2016b, 19(8): 992-1006.

[20] Yang W, Ma K, Kreft H. Geographical sampling bias in a large distributional database and its effects on species richness–environment models[J]. Journal of Biogeography, 2013, 40(8): 1415-1426.

[21] Yang W, Ma K, Kreft H. Environmental and socio‐economic factors shaping the geography of floristic collections in China[J]. Global Ecology and Biogeography, 2014, 23(11): 1284-1292.

[22] Fry C, McColl V, Tomlinson P, et al. Analysis of baseline data requirements for the SEA directive–final report[J]. TRL Ltd and Collingwood Environmental Planning, 2002.

[23] Funk V A, Richardson K S, Ferrier S. Survey-gap analysis in expeditionary research: where do we go from here?[J]. Biological Journal of the Linnean Society, 2005, 85(4): 549-567.

[24] Hortal J, Lobo J M, JIMÉNEZ‐VALVERDE A. Limitations of biodiversity databases: case study on seed‐plant diversity in Tenerife, Canary Islands[J]. Conservation Biology, 2007, 21(3): 853-863.

[25] Kadmon R, Farber O, Danin A. A systematic analysis of factors affecting the performance of climatic envelope models[J]. Ecological Applications, 2003, 13(3): 853-867.

[26] Feeley K J, Silman M R. Keep collecting: accurate species distribution modelling requires more collections than previously thought[J]. Diversity and Distributions, 2011, 17(6): 1132-1140.

[27] Brummitt N, Bachman S P, Aletrari E, et al. The Sampled Red List Index for Plants, phase II: ground-truthing specimen-based conservation assessments[J]. Philosophical Transactions of the Royal Society of London B: Biological Sciences, 2015, 370(1662): 20140015.

[28] Colwell R K, Coddington J A. Estimating terrestrial biodiversity through extrapolation[J]. Philosophical Transactions of the Royal Society of London B: Biological Sciences, 1994, 345(1311): 101-118.

[29] Soberón J, Jiménez R, Golubov J, et al. Assessing completeness of biodiversity databases at different spatial scales[J]. Ecography, 2007, 30(1): 152-160.

[30] Sousa‐Baena M S, Garcia L C, Peterson A T. Completeness of digital accessible knowledge of the plants of Brazil and priorities for survey and inventory[J]. Diversity and Distributions, 2014, 20(4): 369-381.

[31] Hortal J, Lobo J M. An ED-based protocol for optimal sampling of biodiversity[J]. Biodiversity and Conservation, 2005, 14(12): 2913-2947.

[32] Hurlbert, A.H. & Jetz, W. (2007) Species richness, hotspots, and the scale dependence of range maps in ecology and conservation. Proceedings of theNational Academy of Sciences of the United States of America, 104, 13384–9.

[33] Hawkins, B.A., Rueda, M. & Rodríguez, M.Á. (2008) What Do Range Maps and Surveys Tell Us About Diversity Patterns? Folia Geobotanica, 43,345–355.

[34] Guisan A, Zimmermann N E, Elith J, et al. What Matters for Predicting the Occurrences of Trees: Techniques, Data, or Species' Characteristics?[J]. Ecological Monographs, 2007: 615-630.

[35] 王长永, 曹学章, 薛达元. 中国生物多样性数据资源现状分析[J]. 中国环境科学, 1998, 18(5):387-390.

[36] 李勇. 植物标本数字化与生物多样性信息整合——以天津自然博物馆为例[J]. 科学教育与博物馆, 2015(1):55-60.

[37] Vollmar A, Macklin J A, Ford L. Natural history specimen digitization: challenges and concerns[J]. Biodiversity informatics, 2010, 7(2).

[38] Bradley R D, Bradley L C, Garner H J, et al. Assessing the value of natural history collections and addressing issues regarding long-term growth and care[J]. BioScience, 2014, 64(12): 1150-1158.

[39] Costello M J, Appeltans W, Bailly N, et al. Strategies for the sustainability of online open-access biodiversity databases[J]. Biological Conservation, 2014, 173: 155-165.

收稿日期：2017 年 4 月 30 日

黄雄伟：中国科学院动物研究所，博士研究生，主要研究方向为生物多样性信息学。
E-mail: *huangxiongwei@ioz.ac.cn*

林聪田：中国科学院动物研究所，工程师，主要研究方向为生物多样性信息学。
E-mail: *linct@ioz.ac.cn*

纪力强：中国科学院动物研究所，研究员，博士生导师，主要研究方向为生物多样性信息学。
E-mail: *ji@ioz.ac.cn*