

Detecting errors in biodiversity data based on collectors' itineraries

by A. Townsend Peterson, Adolfo G. Navarro-Sigüenza & Ricardo Scachetti Pereira

Received 24 April 2003

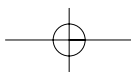
Museum curators have long checked the veracity of specimen data via collectors' field notes, thus establishing whether a particular specimen could have been collected at the claimed place and/or date. In fact, at least one institution (Museum of Vertebrate Zoology, Berkeley, California) invests heavily in its archive of collectors' field notes and has dedicated enormous effort to make this information increasingly useful (<http://elib.cs.berkeley.edu/mvz/>). This source of confirmatory tests of veracity of specimen data, however, would seem at first glance to be available only when detailed (and honest) field notes, journals and catalogues are available.

The idea of checking itinerary information, however, can take on a new relevance with the development of large biodiversity datasets. The example upon which the analyses developed herein are based is that of the birds of Mexico, a databasing project that has covered the Mexican bird holdings of essentially all relevant natural history museums (Peterson *et al.* 1998). This example, however, is increasingly relevant in general: as large biodiversity datasets are assembled from widely distributed sources, such detailed, multi-institutional information will become available for many taxa and regions (e.g., Species Analyst, <http://speciesanalyst.net>; Red Mundial de la Información de la Biodiversidad <http://www.conabio.gob.mx>).

The purpose of this contribution is thus, to develop the framework of a methodology for identifying potential errors in geo-referencing or in the date of collection of specimens by particular collectors. The general approach is that of assembling all of the specimens collected from a particular collector in temporal order, and imposing criteria of maximum radii of likely movement within a day, or over small numbers of days. Although preliminary, and requiring customisation for particular applications, the approach is developed and tested on five Mexican bird collectors.

Methods

Data from 33 core natural history museums in North America and Europe were used in the development of this example. This dataset (*c.*400,000 records) (Peterson *et al.* 1998) was queried for all specimens of each collector, and the top four collectors in terms of productivity were chosen for analysis (Chester C. Lamb, collecting 1920–1969, 41,184 specimens; Wilmot W. Brown, collecting 1890–1953, 18,238 specimens; Mario del Toro Áviles, collecting 1926–1958, 8,744 specimens; A. R.



Phillips, collecting 1923–1989, 7,148 specimens). All of these collectors are now deceased and all were sufficiently productive (Fig. 1) in terms of number of specimens to provide adequate samples for analysis. We also included the 1,458 specimens collected by one of us (AGNS) as an example of specimens for which data were assembled with great care. Specimen collection localities were assigned geographic coordinates based on visual inspection of 1:50,000 topographic maps; specimens for which precise geo-referencing was not possible were excluded from analyses. Records were filtered to keep only those for which both dates and geo-references were available, leaving 40,114 (Lamb), 17,809 (Brown), 7,051 (del Toro Áviles), 6,887 (Phillips) and 1,456 (AGNS) (Fig. 1).

Records of each collector were assembled in Microsoft Excel worksheets and sorted in order of date, and within dates by longitude and then latitude (ideally, localities should have been sorted by geographic proximity, but this was not done for computational simplicity). Distances (km) between one locality (lat1, long1) and the next (lat2, long2) were calculated from the spherical trigonometry formula:

$$\text{Distance} = 6378 \cdot \text{ACOS} \left[\begin{array}{l} \text{SIN}(\text{lat}1/57.2958) \cdot \text{SIN}(\text{lat}2/57.2958) + \\ \text{COS}(\text{lat}1/57.2958) \cdot \text{COS}(\text{lat}2/57.2958) \cdot \\ \text{COS}((\text{long}2/57.2958) - (\text{long}1/57.2958)) \end{array} \right]$$

(drawn from <http://www.auslig.gov.au/geodesy/datums/distance.htm>).

Criteria for identifying potential errors were chosen using a set of assumptions regarding the ability of collectors to move particular distances within particular time intervals. For purposes of demonstration, we used 40 km moved in a single day, 100 km moved from one day to the next, 200 km moved in two days, and 500 km moved in three days as a set of assumptions. These criteria were chosen arbitrarily to represent distances greater than the maximum distances that collectors typically move in particular time intervals. Clearly, if this methodology were to be implemented more broadly, these assumptions could be varied to match particular collectors: collectors moving on horseback could be assigned shorter distance criteria than collectors moving by automobile on more modern roads.

Crude potential errors were summarised as number per 1,000 specimens for which a potential problem was detected (i.e., number of potential errors identified per number of specimens collected on the same day as another specimen, on the day after another specimen, etc.). For a subset of records (Phillips' specimens only), records identified as potential problems were categorised, by laborious visual inspection, into those situations that were most likely to be correct (e.g., collector collected one specimen in one site, and then moved 62 km to another site, and collected more specimens), versus those that were more likely to be erroneous, and into an estimate of actual number of possible errors (e.g., if two sites 1,000 km apart

TABLE 1

Days	AGNS		Brown		Lamb		Phillips		del Toro Áviles	
	No. specimens analysed	No. potential errors	No. specimens analysed	No. potential errors	No. specimens analysed	No. potential errors	No. specimens analysed	No. potential errors	No. specimens analysed	No. potential errors
0	1,220	2	13,551	209	34,346	730	4,586	480	5,079	181
1	154	2	2,365	113	4,339	553	1,147	385	1,316	160
2	11	1	774	23	464	101	305	100	186	15
3	1	0	321	5	165	35	45	23	71	2
0-error	0.0016		0.0154		0.0213		0.1047		0.0357	
1-error	0.0130		0.0478		0.1274		0.3357		0.1216	
2-error	0.0909		0.0297		0.2177		0.3279		0.0807	
3-error	0		0.0156		0.2121		0.5111		0.0282	

were reported as being sampled during seven consecutive days, one error occurred, rather than seven).

Results

The five collectors analysed herein present diverse results in terms of error detection and analysis. Specimens collected by AGNS, as a first example, held a total of five potential errors out of 1,456 specimens (Table 1). Of these specimens with potential errors, four were somewhat long-distance moves in relatively short amounts of time: for instance, after collecting for several days in the Sierra de Juárez, Oaxaca, in early November 1987, AGNS moved 127 km between 5 and 6 November, and collected one bird at La Tinaja, Veracruz (Black Vulture *Coragyps atratus*, MZFC 6395). Checking field notes for this collection indicated that this specimen was a road-kill, and was indeed collected where and when its data indicated; such was the case for one other specimen. Three specimens, however, indeed held errors of geo-referencing (Fig. 1). For example, based on his specimen label data, AGNS moved 541 km between collecting events on 19 and 20 April 1992. Inspection of his field notes indicated that he collected at Chiquihuites, Volcán Tacaná, Chiapas, on 16–19 April, and at Papales, Volcán Tacaná, Chiapas, on 20–21 April. These two localities are not 450 km apart; rather, the geographic reference for Papales proved to represent an error of 5° of longitude. In fact, the two localities are only a few km apart, signifying true errors (post-collection) in data records associated with specimens collected by AGNS.

The other four collectors—in reality the object of this study—showed higher apparent error rates (Table 1). Error rates, calculated under the different distance-time criteria, for Brown ranged 0.015–0.050 per thousand, those for del Toro Áviles 0.028–0.122, for Lamb 0.021–0.218, and for Phillips 0.105–0.511 (see Table 1). Specimens indicated as potentially erroneous revealed many clear cases of error; for example, Phillips' collections in early October 1955 were focused in Sinaloa and Nayarit (Fig. 1), including collections on 8 October 1955. However, also on 8

October 1955, a specimen of Rufous-capped Warbler *Basileuterus rufifrons* (DMNH 26575) was apparently from Acahuizotla, Guerrero, and clearly represents an impossible jump (c.700 km) for a single collector (Fig. 2). Each of the four collectors showed several such clear examples of error; a rough estimate of numbers of actual errors (e.g., two distant localities on the same date, etc., omitting cases in which travel by automobile would have made such distances possible) for Phillips was about 65 probable errors.

Discussion

The approaches explored herein are not new: curators and collection managers of natural history museums have long used collectors' itineraries to assess the veracity of specimen data. What is new is the implementation of these approaches in an exploratory sense, seeking to identify specimens that have high probabilities of error without prior reason to suspect problems. Beyond simple error detection, these methods offer an opportunity to begin to understand objectively the ways in which scientific collections were assembled. In this case, errors of geo-referencing or dating are identified; additional approaches focusing on other sources of error have also been identified, and error-detection tools developed (Chapman 1999). An

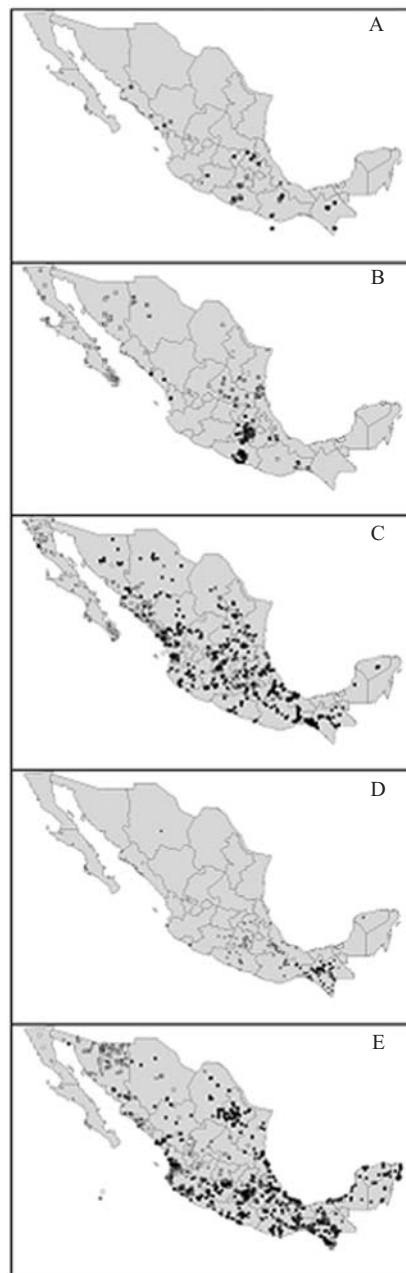


Figure 1. Maps of collecting localities for each of the five collectors analysed in this paper: (A) AGNS, (B) Brown, (C) Lamb, (D) del Toro Áviles, and (E) Phillips. The geo-referencing error detected in AGNS's specimen data is evident off the coast of Guerrero. Collecting localities are roughly classed as to year of collection by five shades from black (oldest specimens) to white (newest specimens).

important point, however, is that these methodologies are not being developed with the aim of designating particular collectors as ‘good’ or ‘bad’, but rather as a way of identifying potential problems among the data records from each collector.

Limitations

The methods explored herein can fail under three circumstances. First, the temporal density of records from a particular collector must be sufficient to permit detection of potential problems. For low-volume collectors, the probability that erroneous records would overlap sufficiently in time to allow detection of problems would be relatively low, so these methods will frequently miss problems with such collectors.

A second limitation is for situations in which specimens are listed as stemming from a single collector, but more than one collector was really involved. Examples, such as Adolphe Boucard across much of the world, are well known as listing the efforts of multiple collectors under a single name. Three of the collectors analysed herein (Brown, Lamb, del Toro Avilés) had no such arrangements. Lamb and often Phillips apparently worked with teams of collectors, although to our knowledge these teams usually remained together in groups (A. R. Phillips pers. comm. to ATP, December 1988); one possible explanation for the complications among Phillips’ specimens, however, is that some of Phillips’ specimens may have been acquired

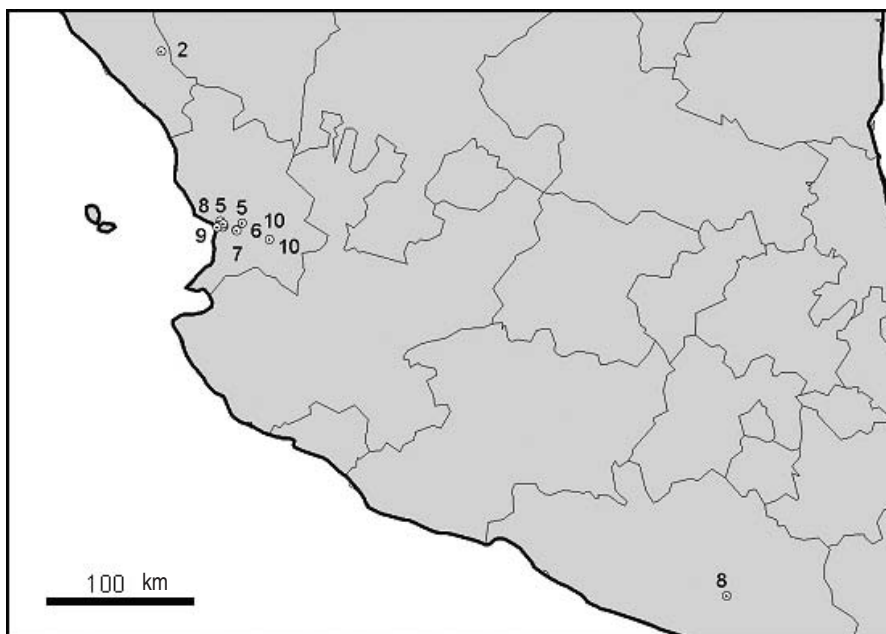
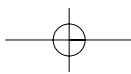


Figure 2. Map of collecting localities for Allan R. Phillips for 2–10 October 1955, showing a locality in Guerrero on 8 October 1955 that is probably erroneous.



from his collectors, who may have continued accumulating specimens (and labelled them under his name) when not in the field with Phillips. More general application of these methods must nevertheless bear in mind that such collectors who included catalogue entries from multiple collectors and preparators will often produce false indications of errors.

The final, and darkest, possibility is that of complete fabrication of data. That is, if a malicious collector simply invented everything, in this way, it would be possible to make the associated information completely consistent. This complete fabrication, rather than errors or fabrications associated with single specimens, would not be detected by the approaches developed herein.

Causes of errors detected

A diversity of factors can produce the types of errors that have been detected in this study. They range from mistakes in data recording to deliberate falsification and, of course, have very different implications as a result. Going beyond simply detecting errors then becomes important: analysis of the types and causes of errors detected permits learning about the methods and motives of each collector, and may allow detection of additional errors that could have important implications for biodiversity inventories.

Errors can be divided into two general types: collector errors and post-collection errors. Post-collection errors can enter the process at a variety of points. In older collections, such as with the Salvin and Godman collections from Mexico and Central America, specimen labels were at times prepared at the museum well after the collector's involvement (Godman 1915), creating considerable opportunity for mistakes. An excellent example of this effect is for the type specimen of the woodnymph *Thaluranina luciae* (Deignan 1961), which is labelled as collected in the Tres Marias Islands off western Mexico, and described as a species new to science; in reality, it was from Brazil, its data having probably been confused between two collections arriving at the U.S. National Museum of Natural History at the same time. Numerous subsequent phases, including re-labelling, cataloguing, computerisation, and geo-referencing, provide additional opportunities for errors to come into existence.

Collector errors are perhaps more complex. Certainly, some of these problems are simple mistakes: a collector noted an incorrect date or locality on the label. Brown, according to Phillips (pers. comm. to ATP, December 1988), had an amusing source of error: labels with locality information were apparently made up prior to collection and placed in the skinning kit. If all of the pre-made labels were not used at a particular site, however, the remainder apparently stayed in the skinning kit, and would be used (accidentally) at succeeding localities. This sort of error would produce a clear signature of past, well-sampled localities appearing as scattered errors at later sites. Such a signature is indeed present in at least some Brown collections: for instance, a specimen of Olive Sparrow *Arremonops [rufivirgatus] sumichrasti* (MCZ 164687) was supposedly collected on 31 January 1931 at

Acapulco, Guerrero, when the collector was collecting numerous specimens from Coyuca, Guerrero (e.g., MCZ 163991). Acapulco, however, was sampled intensively on 9 December 1930–8 January 1931 (e.g., MCZ 163887). Nevertheless, overall, Brown's specimens appear to be relatively clean, without overwhelming numbers of errors produced by his apparent carelessness regarding labels.

A still-worse source of error is that in which labels were not prepared at the time of collection. The famous example here is that of Mario del Toro Áviles, who apparently had more than 10,000 unlabelled specimens in his house when visited by A. R. Phillips (Binford 1989). This utter lack of interest in accuracy of label information has caused numerous problems, including many odd locality and date records (Binford 1989), and even species for which the type locality is most likely inaccurate (Peterson & Nieto-Montes de Oca 1996). This pattern of collector behaviour produces the relatively numerous random locality and date records that characterise del Toro Áviles' specimens, although del Toro Áviles ranked well behind Lamb and Phillips in terms of error rates detected.

The higher error rates detected among the specimens of Lamb and Phillips are quite surprising. In both cases, the collector had been considered to be relatively reliable and no previous commentary had addressed potential problems with their material. This implication of potential problems thus demands a more in-depth, biographical analysis of these collectors.

Most difficult to manage are errors that involve deliberate falsification of data. The most famous such doings are those of the collector Richard Meinertzhagen, who apparently stole specimens, re-labelled them, and sometimes even re-prepared them to give the appearance that he had collected them himself (Knox 1993, Rasmussen & Collar 1999). As mentioned above, such broad falsification would be difficult to detect using the methodologies developed herein.

Nevertheless, errors among the specimens of Allan R. Phillips are suspicious—certainly, potential error rates associated with his specimens are higher than would have been expected, so we must consider additional information regarding Phillips' work as a collector. First, error rates among his specimens are high, indeed many times those of Mario del Toro Áviles, whom Phillips criticised so thoroughly (A. R. Phillips pers. comm. to ATP, December 1988). Second, Phillips was very concerned about specimen data, e.g., in respect of moult patterns, migration records and subspecific variation, so it seems unlikely that he would make so many random errors. Third, comments made by Phillips (pers. comm. to ATP, December 1988) regarding collecting specimens indicated that when he collected specimens during periods for which he had no collecting permits, he would change dates to those time periods for which permits were available. Finally, during museum studies as part of the assembly of the Atlas, several specimens were encountered by AGNS in Phillips' personal collection that were collected by students and staff at Universidad Nacional Autónoma de México and other Mexican institutions, including specimens collected by AGNS himself! The implications of such findings beg

further information, perhaps from Phillips' colleagues and associates, which would illuminate the situation further and might permit a decision as to whether major portions of Phillips' material should be ignored in considerations of the specimen record of Mexican birds.

Implementation

Clearly, a central concern for any broad implementation of these methodologies should be the suites of assumptions regarding the ability of different collectors to move particular distances per unit time. We have, in this prototype of the method, used a 'middle-of-the-road' suite of assumptions, solely for the purposes of demonstration. Clearly, modern collectors (e.g., AGNS) could be treated differently from the former collectors (e.g., Brown), which would result in greater precision in error detection. In this sense, consideration of some biographical information regarding each collector could greatly enrich our error-detection approaches.

Tools for broad implementation of these error-detection approaches are presently under development. A first requirement is that of large quantities of biodiversity information, in which collectors' entire specimen collections are assembled. Such assemblies of information necessarily depend on integrating information from across many institutions, as most collectors' specimens are spread widely among them. Projects such as that of the Atlas of Mexican bird distributions, from which the present dataset is drawn, require intense investment in data management and for that reason there are at present few such projects (Peterson *et al.* 1998, Navarro-Siguenza *et al.* 2002). An alternative approach is that of distributed biodiversity databases, which integrate electronic databases housed at different institutions via the Internet. Nevertheless, a considerable time lag is involved in the creation of such electronic databases, particularly for methods such as this one that require complete information for full functionality.

Acknowledgements

We thank the curators and staff of the following scientific collections for access to specimens and data under their care: American Museum of Natural History; Academy of Natural Sciences of Philadelphia; Bell Museum of Natural History; British Museum (Natural History); California Academy of Sciences; Carnegie Museum of Natural History; Canadian Museum of Nature; Denver Museum of Natural History; Delaware Museum of Natural History; Fort Hays State College; Field Museum of Natural History; Iowa State University; University of Kansas; Los Angeles County Museum of Natural History; Natuurhistorische Museum; Louisiana State University Museum of Zoology; Museum of Comparative Zoology, Harvard University; Moore Laboratory of Zoology, Occidental College; Muséum Nationale d'Histoire Naturelle; Museum of Vertebrate Zoology, Berkeley; Museo de Zoología, Facultad de Ciencias, Universidad Nacional Autónoma de México; University of Nebraska; Royal Ontario Museum; San Diego Natural History Museum; Southwestern College; Texas Cooperative Wildlife Collections; University of Arizona; University of British Columbia Museum of Zoology; University of California Los Angeles; Universidad Michoacana de San Nicolás de Hidalgo; United States National Museum of Natural History; Western Foundation of Vertebrate Zoology; and Peabody Museum, Yale University. This study was supported by the U.S. National Science Foundation. ATP's work (in Brazil) was supported by a grant from the Fundação de Amparo à Pesquisa do Estado de São Paulo.

References:

- Binford, L. C. 1989. *A distributional survey of the birds of the Mexican state of Oaxaca*. Orn. Monogr. No. 43.
- Chapman, A. D. 1999. Quality control and validation of point-sourced environmental resource data. Pp. 409–418 in Lowell, K. (ed.) *Spatial accuracy assessment: land information uncertainty in natural resources*. Ann Arbor Press, Chelsea, MI.
- Deignan, H. G. 1961. Type specimens of birds in the United States National Museum. *Smithsonian Inst. Bull.* 1961: 1–718.
- Godman, F. D. 1915. *Biologia Centrali Americana: introductory volume*. Taylor & Francis, Ltd., London.
- Knox, A. G. 1993. Richard Meinertzhagen—a case of fraud examined. *Ibis* 135: 320–325.
- Navarro-Sigüenza, A. G., Peterson, A.T. & Gordillo-Martinez, A. 2002. A Mexican case study on a centralized database from world natural history museums. *CODATA Journal* 1: 45–53.
- Peterson, A. T., Navarro-Sigüenza, A. G. & Benitez-Diaz, H. 1998. The need for continued scientific collecting: a geographic analysis of Mexican bird specimens. *Ibis* 140: 288–294.
- Peterson, A. T. & Nieto-Montes de Oca, A. 1996. Sympatry in *Abronia* (Squamata: Anguillidae) and the problem of Mario del Toro Aviles' specimens. *J. Herpetology* 30: 260–262.
- Rasmussen, P. C. & Collar, N. J. 1999. Major specimen fraud in the forest owllet *Heteroglaux* (*Athene* auct.) *blewitti*. *Ibis* 141: 11–21.
- Rasmussen, P. C. & Prŷs-Jones, R. P. 2003. History vs mystery: the reliability of museum specimen data. *Bull. Brit. Orn. Cl.* 123A: 66–94.

Addresses: A. Townsend Peterson, Natural History Museum and Biodiversity Research Center, The University of Kansas, Lawrence, Kansas 66045, e-mail: town@ku.edu. Adolfo G. Navarro-Sigüenza, Museo de Zoología, Facultad de Ciencias, Universidad Nacional Autónoma de México, Apartado Postal 70-399, Mexico, D.F. 04510, Mexico. Ricardo Scachetti Pereira, Centro de Referência em Informação Ambiental, Av. Romeu Tórtima 388, Barão Geraldo 13084-520 Campinas S.P., Brazil.

© British Ornithologists' Club 2004

Rediscovery of the White-necked Picathartes *Picathartes gymnocephalus* in Ghana

by Ben D. Marks, Jason D. Weckstein, Kevin P. Johnson,
Mathys J. Meyer, James Braimah & James Oppong

Received 1 July 2003

The White-necked Picathartes *Picathartes gymnocephalus* is endemic to the Upper Guinean forests of West Africa (Fry *et al.* 2000) from Guinea to Ghana. Throughout this range, the rapid fragmentation and destruction of lowland rain forest threatens the survival of this remarkable species (BirdLife International 2000). Recent studies have focused on various demographic and ecological questions regarding populations of *P. gymnocephalus* in Guinea (Halleux 1994), Liberia (Allport 1991), Sierra Leone (Thompson 1993, 2001, Thompson & Fotso 2000), and Ivory Coast (Salewski *et al.* 2000). However, no recent records of this bird are available from Ghana. The most recent published records of *Picathartes* in the country are those of