Biological Journal of the Linnean Society, 2005, **85**, 549–567. With 20 figures

Survey-gap analysis in expeditionary research: where do we go from here?

V. A. FUNK[1]*, KAREN S. RICHARDSON[2] and SIMON FERRIER[3]

[1]*US National Herbarium, Department of Botany, National Museum of Natural History, Smithsonian Institution MRC 166, Washington D.C. 20013–7012, USA*
[2]*Cooperative Research Centre for Rainforest Ecology and Management, Department of Zoology and Entomology, University of Queensland, St Lucia, Qld. 4072, Australia*
[3]*New South Wales Department of Environment and Conservation, PO Box 402, Armidale, NSW 2305, Australia*

Received 29 March 2004; accepted for publication 23 October 2004

Research expeditions into remote areas to collect biological specimens provide vital information for understanding biodiversity. However, major expeditions to little-known areas are expensive and time consuming, time is short, and well-trained people are difficult to find. In addition, processing the collections and obtaining accurate identifications takes time and money. In order to get the maximum return for the investment, we need to determine the location of the collecting expeditions carefully. In this study we used environmental variables and information on existing collecting localities to help determine the sites of future expeditions. Results from other studies were used to aid in the selection of the environmental variables, including variables relating to temperature, rainfall, lithology and distance between sites. A survey gap analysis tool based on 'ED complementarity' was employed to select the sites that would most likely contribute the most new taxa. The tool does not evaluate how well collected a previously visited site survey site might be; however, collecting effort was estimated based on species accumulation curves. We used the number of collections and/or number of species at each collecting site to eliminate those we deemed poorly collected. Plants, birds, and insects from Guyana were examined using the survey gap analysis tool, and sites for future collecting expeditions were determined. The south-east section of Guyana had virtually no collecting information available. It has been inaccessible for many years for political reasons and as a result, eight of the first ten sites selected were in that area. In order to evaluate the remainder of the country, and because there are no immediate plans by the Government of Guyana to open that area to exploration, that section of the country was not included in the remainder of the study. The range of the ED complementarity values dropped sharply after the first ten sites were selected. For plants, the group for which we had the most records, areas selected included several localities in the Pakaraima Mountains, the border with the south-east, and one site in the north-west. For birds, a moderately collected group, the strongest need was in the north-west followed by the east. Insects had the smallest data set and the largest range of ED complementarity values; the results gave strong emphasis to the southern parts of the country, but most of the locations appeared to be equidistant from one another, most likely because of insufficient data. Results demonstrate that the use of a survey gap analysis tool designed to solve a locational problem using continuous environmental data can help maximize our resources for gathering new information on biodiversity. *Biological Journal of the Linnean Society*, 2005, **85**, 549–567.

ADDITIONAL KEYWORDS: collecting expeditions – ED complementarity – environmental diversity.

INTRODUCTION

Expeditionary research (defined here as travel into poorly known and difficult to access places for the purpose of collecting biological specimens and/or data) has never been more important. There are many areas of the world (terrestrial and marine) about which we have little or no information. The pressure to understand the distribution of plants and animals is greater than ever before, with conservation planners seeking

*Corresponding author. E-mail: funkv@si.edu

to use the best available data for delimiting areas for national protected areas (Wilson, 2000). However, good expeditionary research is not easy to accomplish. There are several reasons for this: trips are expensive, resources are limited, experienced people hard to find, regulations and paperwork have increased, and time is short.

In the nineteenth and early twentieth centuries, expeditions were difficult and expensive and years of travel were required. Many of us have been enthralled by the accounts of naturalists and collectors on those magnificent voyages, such as Alexander von Humboldt (1814–1826) Joseph Dalton Hooker (1844–1847, 1853–1855, 1855–1860), Charles Darwin (1845, 1859), Alfred Russel Wallace (1869, 1876) and many others. The cost of these legendary trips was great. Sometimes sets of specimens were sold to help defray the cost or collectors were employed by museums. However, governments sponsored many of the really long trips (e.g. Voyage of the Beagle 1831–1836, the US Exploration Expedition 1838–1842, Lewis and Clark 1804–1806).

Expeditionary research then experienced a period when air travel increased the speed of trips and low in-country costs kept the price reasonable. As a result, more expeditions of shorter duration and more limited scope were undertaken. However, in the last couple of decades in-country costs have increased to the point where it is more expensive to mount an expedition into the interior of a remote country than it is to do field work in a more developed one. For example, the Smithsonian Institution has had an active field programme in Guyana for over 15 years. The cost of an expedition into the interior, where there is only transportation via chartered aircraft, river and walking, now averages $US10 000–25 000.

Undertaking an expedition to a biologically diverse or otherwise interesting area requires a commitment of resources from one's home institution. Whenever specimens are collected, the process of turning those collections into museum specimens involves preparing, cataloguing, processing, identifying, distributing duplicate material or sets to experts for further identification, adding the information to the database and barcoding, integrating the material into the museum's collection, possibly making the data available via the internet, and most important, maintaining the information and collections over the long term.

With some groups like birds and mammals, specimens are prepared in the field and one can identify nearly all of them in a short amount of time. With a concerted effort, groups such as plants and some insects (i.e. dragonflies, ants) may reach a 60–70% identification level after approximately 2 years. With other groups such as Microlepidoptera or aquatic Coleoptera one might find that over 90% of the species are new to science. Of course, collections and databases must be maintained and improved through the years. On average, the cost of taking care of the collections once they are back at the home institution exceeds the cost of collecting them in the first place. Several international initiatives to link museum collections online, including the Global Biodiversity Information Facility (GBIF), have highlighted the efforts, shortcomings and costs of compiling such databases and continue to stress the importance of field exploration and collection management (Edwards, Lane & Nielsen, 2000; Pennisi, 2000; Sugden & Pennisi, 2000).

Competent and experienced professionals, who will travel to remote areas for months or even years, are difficult to find and are under-appreciated by the systematic community and administrators. Collecting is more than just a rote activity because it requires commitment, training and an 'eye' for deciding what to collect – not to mention the fact that it is often dangerous and time consuming. A great deal of experience is necessary to enable one to carry out a successful expedition, yet it is hard for such individuals to survive in our academic system. Spending months in the field every year makes it difficult to publish at the same rate as colleagues who remain at home or only visit easy-to-access places. Getting to 'know' a group of organisms, to the point where one can become the expert on that group at a relatively high taxonomic level, takes years of field and museum work, and today students are encouraged to spend time in the lab rather than the field.

Finding money to do this type of fieldwork is difficult and only a few funding bodies (e.g. National Geographic Society) will still accept proposals for fieldwork in places where the major justification is 'we don't know what's there'. As a result, researchers are finding it hard to get the attention of funding agencies, all too necessary for recognition and promotion. Furthermore, the pool of individuals who are young enough to do this type of work, as well as being competent and willing, shrinks every year. It is estimated that there are only approximately 6000 PhD-level researchers worldwide active in the exploration and description of the world's flora and fauna (Wilson, 2000) and many of these are over 50 years of age. Even at museums, the emphasis for younger scientists is now on molecular phylogenies and obtaining grants, and the credit one receives for major fieldwork and its companion discipline, alpha taxonomy, has decreased. In recent years, the National Science Foundation (NSF) has recognized this shortcoming and has tried to promote the resurgence of systematic research and fieldwork by funding the Partnership for Enhancing

Expertise in Taxonomy (PEET; Pennisi, 2000) and Biotic Surveys and Inventories (BS & I). Unfortunately, the former has become a largely molecular exercise and the latter is poorly funded compared to other programmes.

Although the regulations for collecting change from country to country, in general it is now much more difficult, and sometimes impossible, to obtain permits. For years, scientists and conservation groups have stressed that biodiversity is valuable for numerous reasons. As a result of our public relations efforts, many countries are now afraid that expeditions are just an excuse for bioprospecting and as a result there are many difficulties to overcome prior to being issued collecting and export permits. In some countries the governments equate collecting permits with commercial permits and charge accordingly. A sad result of all of this is that it is often easier to get permission to cut down a forest than it is to take samples from it for scientific purposes. Even when permits are possible, there are often long lists of time-consuming and expensive requirements. The list of countries that are friendly to scientists vs. ones that should be avoided is constantly changing and such information is available only from other scientists who work in the countries in question.

Time has become the enemy in conservation and the longer it takes to provide the data necessary for decision-making the more biological diversity that is destroyed. We cannot wait until we know everything about every place to make decisions on what to conserve. Biological survey data are extremely useful if one wishes to determine how to maximize the total diversity conserved. Even though time is limited and answers to questions concerning conservation are needed soon, there is still time to incorporate information from new expeditions, especially if these data are from critical areas that will vastly increase our ability to either model, or predict in some fashion, the diversity of an area.

All of these considerations mean that once the resources and personnel have been identified, the decisions on where to send expeditions must be made with care. In this study we attempt to answer the following question: If one is interested in obtaining an overall knowledge of the biodiversity (or of a taxon) of an area, and if there are insufficient data, then where should survey data be gathered?

In the past, expeditions have been sent to areas believed to be 'interesting', based on the knowledge of an individual, or perhaps because no one has ever been there or because of the idea that 'it is remote and therefore it must be interesting'. However, new ideas are emerging as to how we should determine the location of future field expeditions in order to gain the most information for the money spent. Recent efforts by a number of groups and institutions to compile databases of museum collections and make those data available have given the biodiversity research community the ability to use tools such as GIS to ask questions that were not possible to address just a short time ago (Richardson & Funk, 1999; Edwards *et al.*, 2000; Pennisi, 2000; Araújo *et al.*, 2001; Bell, 2003; Faith, 2003; Suarez & Tsutsui, 2004).

One of the results of GIS studies is the documentation of collecting bias. Indeed, there is evidence of bias even in areas where the flora and fauna are considered relatively well known. Bias can be geographical, the product of more easily accessed areas being favoured. It can also result from collecting that is taxonomically incomplete, with only easy-to-study species being included, thus giving undue weight to a few taxa. It can be temporal, based on one survey, during one season (usually the dry) (Faith & Walker, 1996; Ferrier, 1997; Funk, Zermoglio & Nassir, 1999). One way of addressing geographical biases is to design regional surveys that attempt to sample the heterogeneity of that region.

Regional survey design

The design of regional surveys of large heterogeneous areas, such as a country, is critical for assessing biodiversity (Ferrier & Smith, 1990; Austin & Heyligers, 1991; Balmford & Gaston, 1999). Most of the world's regions identified as being of high priority for conservation action, particularly those in the tropics, are relatively data-poor (Myers *et al.*, 2000; Brooks *et al.*, 2001; Olson & Dinerstein, 2002). The paucity of data in these regions reinforces the need to conduct carefully designed and cost-efficient surveys that yield the greatest potential future use of data (Margules & Austin, 1994; Balmford & Gaston, 1999). The few studies of regional survey techniques have demonstrated the usefulness of several key elements including: (1) consideration of efficiency due to the high costs of surveys; (2) clear purpose for data use, and (3) repeatable, explicit sampling procedures (Gillison & Brewster, 1985; Ferrier, 1990; Austin, 1991; Austin & Heyligers, 1991; Margules & Austin, 1994; Balmford & Gaston, 1999). There is also concurrence amongst these studies that survey sites should be spread representatively across major environmental gradients within a region to capture the heterogeneity within the region.

Several approaches to sampling environmental gradients have been suggested including: (1) gradsect sampling, whereby the steepest gradients are sampled to ensure the greatest range of organismal diversity (Gillison & Brewster, 1985), and (2) division of environmental variables (e.g. mean annual rainfall, lithology) into a small number of discrete classes, deriving all possible combinations of these classes to sample

using GIS (Ferrier & Smith, 1990). A shortcoming of the latter technique is the arbitrary division of the continuous environmental space described by the variables into classes and the lack of consideration of the similarity between any two classes, as each class is treated equally and separately. The problem with both techniques is the potential lack of spread of sites geographically if environmental gradients are steep and clustered.

A new tool developed by the GIS Unit at the New South Wales National Parks and Wildlife Service (now Department of Environment and Conservation) in Armidale (NSW NPWS, 1998; Ferrier, 2002) addresses these shortcomings by extending an analysis technique pioneered by Faith & Walker (1996). Referred to henceforth as 'the survey gap analysis tool', it analyses the survey coverage of a region in relation to the underlying continuous environmental and geographical space, rather than in terms of arbitrary classes. Faith & Walker's (1996) environmental diversity (ED) measure, based on the p-median criterion, was developed for selecting sets of sites that represent regional biodiversity by providing best possible coverage of regional environmental variation. It functions by measuring how well a set of sites covers the continuous environmental space and evaluating the potential improvement that any given site would make if added. The technique, based on the finding that sampling different parts of the overall environmental space yields a good representation of the biological diversity of a region (Faith & Walker, 1996), can equally be applied to the problem of selecting survey sites.

In the context of the survey gap analysis tool, the objects in the ED analysis are the sites (geographical locations) and pattern-relationships among these objects (as represented in this case by a matrix of pairwise dissimilarities/distance among sites rather than an ordination) are assumed to indicate relative numbers of underlying features, corresponding to species (Faith & Walker, 1996; Ferrier, 2002; Faith, Ferrier & Walker, 2004). The pattern provides predictions of the degree of biodiversity complementarity of a site to any given set of sites (e.g. Faith *et al.*, 2004). Hence the ED complementarity of a site (with reference to some set of other sites) is larger to the extent that it reduces the value of the p-median (the overall ED values) by a larger amount when added to the set. Within a set, the ED complementarity value of a site is larger to the extent that its removal from the set increases the p-median score (Faith *et al.*, 2004).

In this study the survey gap analysis tool is used to examine the distributions of existing collecting localities for plants and animals across Guyana in conjunction with data on geology, rainfall, elevation, and geographical distance to determine areas of the country that have the most potential to provide the maximum amount of new information. The method is tested using data from recent and historical collections made in Guyana (Fig. 1).

Guyana is bordered by the Atlantic Ocean to the north, Venezuela to the west, Brazil to the south-west

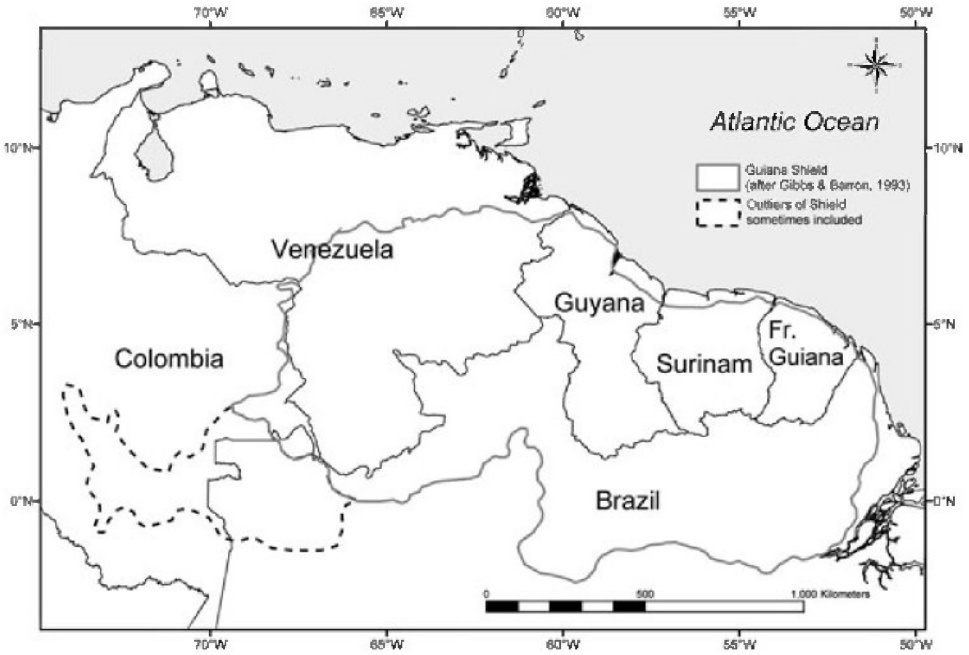


Figure 1. North-eastern South America, showing the Guiana Shield and the country of Guyana.

Table 1. Number of families, genera and species for each taxonomic group

Group	#Families	#Genera	#Species	#Records
Plants	214	1474	6016	34 228
Insects (termites and butterflies)	64	392	715	9414
Birds	7	129	316	1111
Total	279	1869	7047	44 753

Table 2. Weightings for each variable

Variable	Weight
December rainfall	1.0
October rainfall	1.0
Maximum temperature of the warmest period	1.0
Mean temperature of the wettest quarter	1.0
Lithology	1.0
Geographical distance	0.5

and south, and Surinam to the east. It is situated just north of the Equator and most of it fits between 56.5 and 61.33 degrees W. It covers 215 000 km^2, about the size of Idaho (USA) and slightly smaller than Great Britain. It has approximately 800 000 inhabitants, most of whom live along the coast near the capital, Georgetown. The country sits on the Guiana Shield, which is endowed with a diverse array of habitats and vegetation types. There are mangroves along the coast, the vast Rupunini savannahs in the south-west, smaller white sand savannahs east of the Essequibo, marsh and swamp forests, herbaceous swamps, dense tropical forests, dry deciduous forests, the famous Greenheart forests and of course the spectacular Pakaraima Mountains, with peaks such as Ayanganna (2100 m), Wokoman (2000 m), and Roraima (2800 m) that have montane forests on the slopes and tepui vegetation at the summits. Except for much of the coast, the Greenheart forests, and most of the savannahs, the country is relatively undisturbed and there are many areas that are still untouched. Current estimates place the amount of native vegetation coverage at around 70% (J. Singh, pers. comm.). However, road building, gold, diamond, and bauxite mining, along with foreign logging concerns and the wildlife trade are all increasing at an alarming rate and as yet very little land is under any sort of protection.

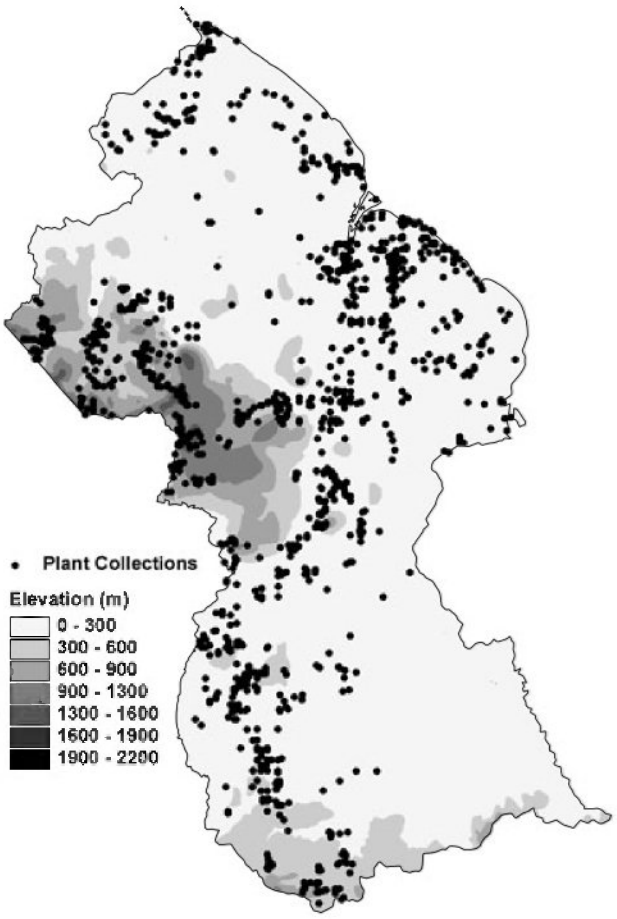
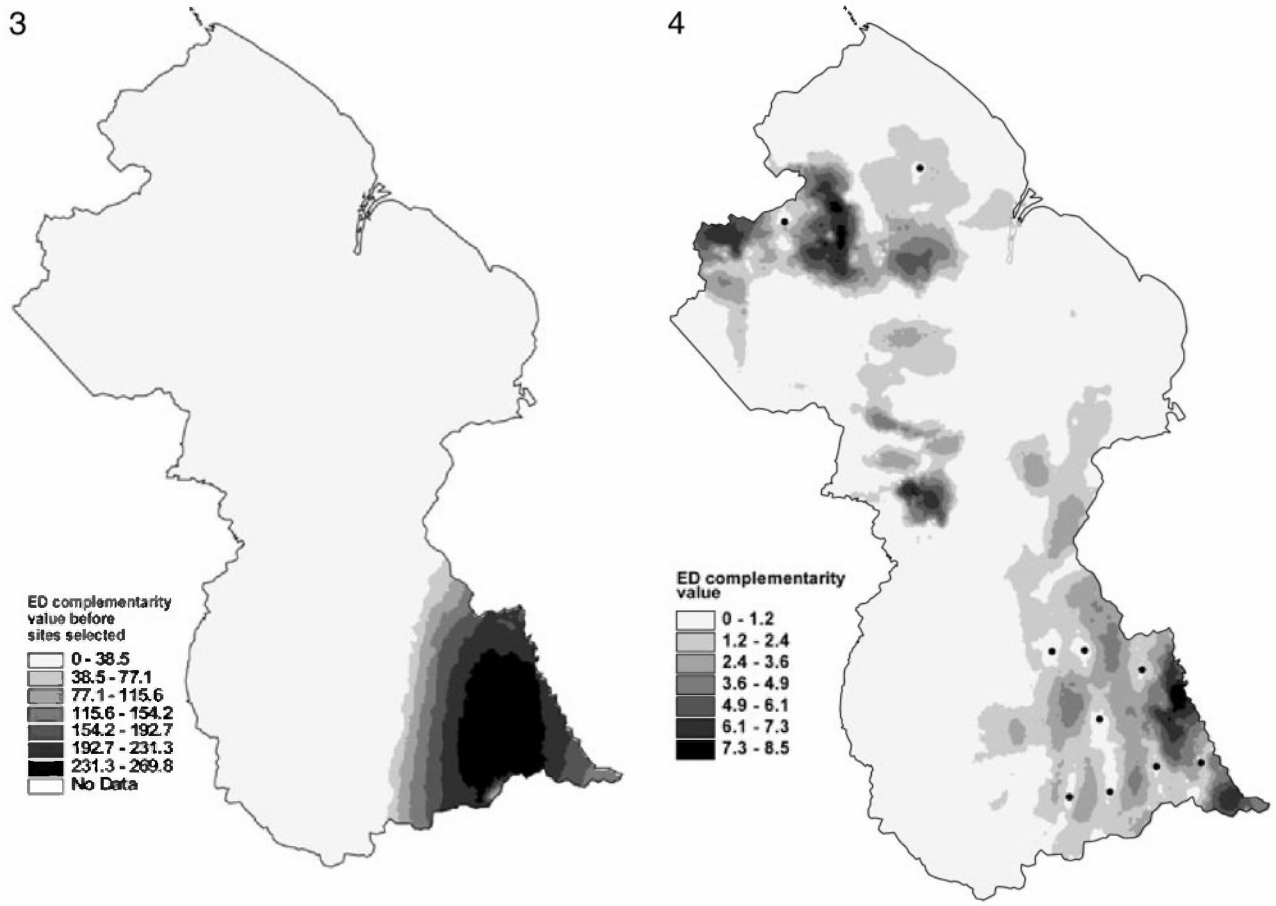


Figure 2. Each dot represents a site in Guyana where at least one plant has been collected.

METHODS

DATA

The collections data were provided by the Biological Diversity of the Guiana Shield Program (BDG) of the Smithsonian Institution (http://www.nmnh.si.edu/biodiversity/bdg). For nearly 20 years the BDG has had an active field programme in Guyana and has developed a database of its collections. In addition, data were collected from some of the historical collections from Guyana housed at museums and herbaria around the world (see Funk *et al.*, 1999 and Funk & Richardson, 2002 for a detailed discussion of the database). One important feature of the BDG database is that all species records are 'specimen-based',



Figures 3–4. Guyana with all plant localities, using the entire country in the analysis. Fig. 3. Areas in most need of plant collecting. Fig. 4. Dots indicate the location of the ten sites where plants should be collected to find the largest number of new records.

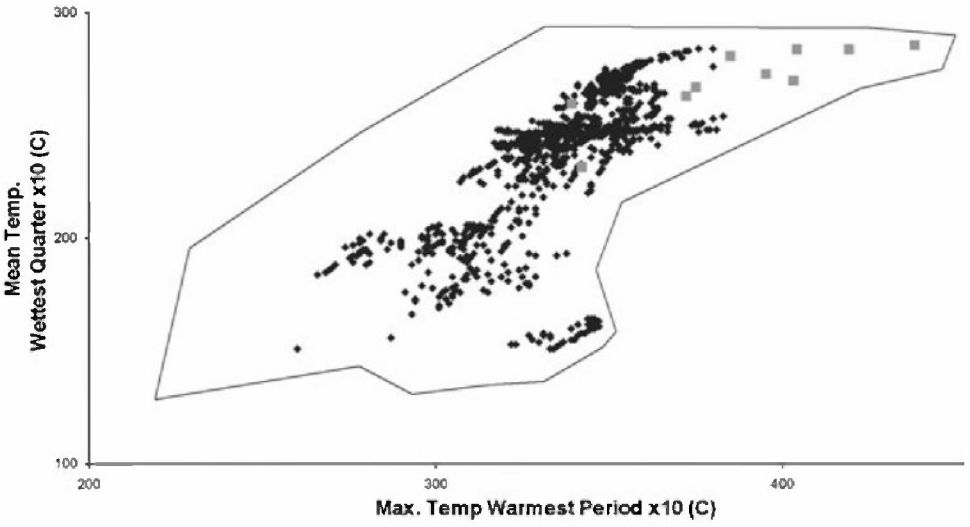
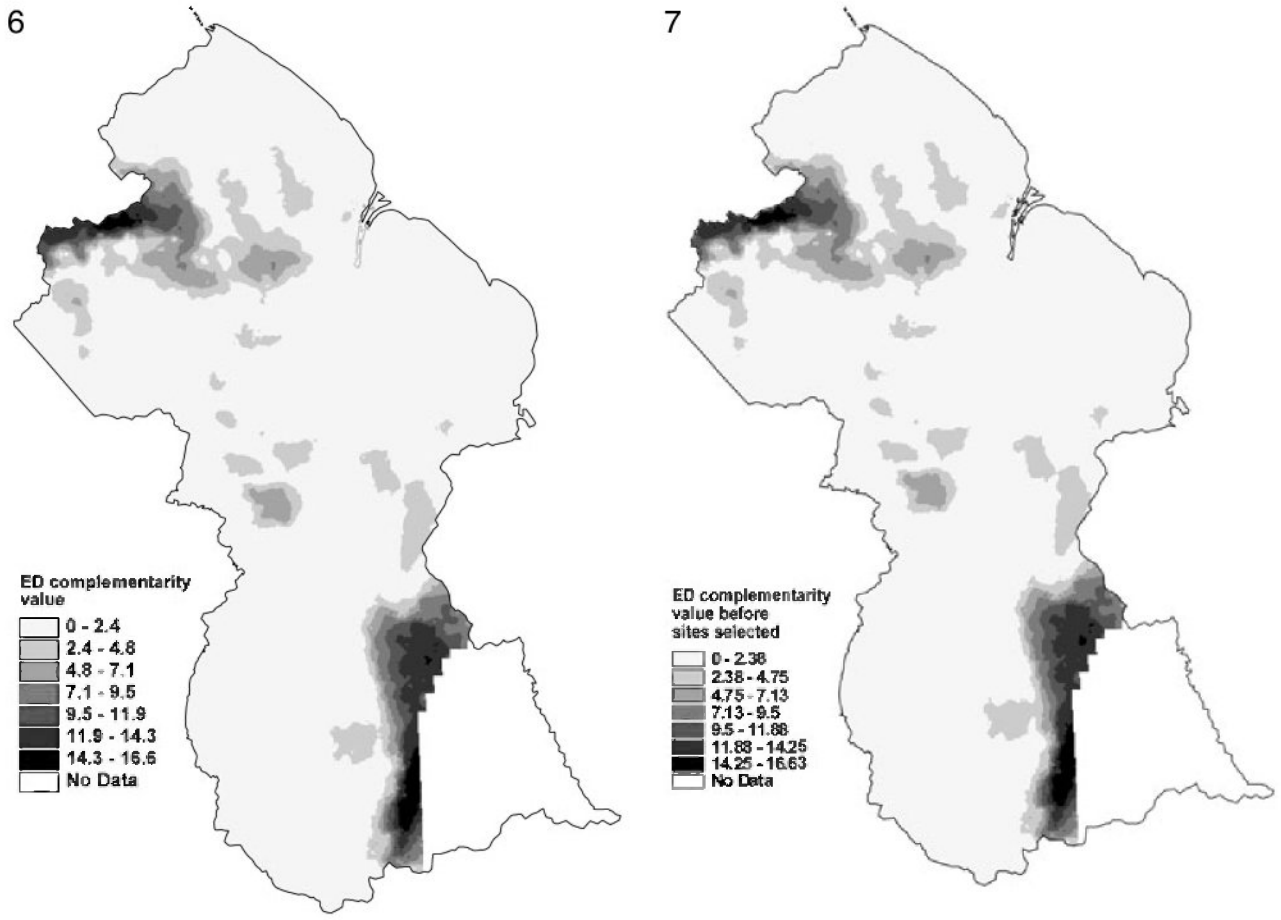


Figure 5. An example of environmental space with two variables. Dots indicate collecting sites and the grey squares are the proposed ten collecting sites.



Figures 6–7. Guyana with the south-east corner masked. Fig. 6. Areas in most need of plant collecting. Fig. 7. Dots indicate the location of the ten sites where plants should be collected to find the largest number of new records.

i.e. no observational data are included. As a result, every record has a voucher for which the identification can be verified (a few of the bird records are based on vocal recordings). The majority of the data for the historical collections come from seven institutions (in alphabetical order by country): the Royal Ontario Museum (ROM), Canada; the Royal Botanic Gardens, Kew (K), and the Natural History Museum (BM), England; the University of Guyana (GNM, BRG), Guyana; the University of Utrecht Herbarium (U), the Netherlands; and the Smithsonian Institution (US), and American Museum of Natural History (AMNH), USA.

Data were entered into an Access database and only species records that were geocoded in the database were included in this study. Table 1 lists the number of families, genera, species, and records for each group studied. The plant database has 34 228 collections from 1235 locations, representing 6016 species; an average of 5.7 collections per species. For birds the database has 9414 records (it does not include duplicate species from the same site collected during the same trip) from 192 sites and 715 species. The average number of collections per site cannot be calculated for birds because we did not have duplicate site records in the database. Finally, the termite database has 80 sites, 1111 collections and 316 species; an average of *c.* 3.5 collections per species.

SURVEY GAP ANALYSES

The survey gap analysis tool (NSW NPWS, 1998; Ferrier, 2002) was used to identify gaps in the historical and current collections data. The tool runs as an extension to ArcView (ESRI, 1999) and uses the p-median algorithm to measure how well a set of collecting sites represents the environmental and geographical variation of the region. In the case of the survey gap analysis tool, a combination of the two

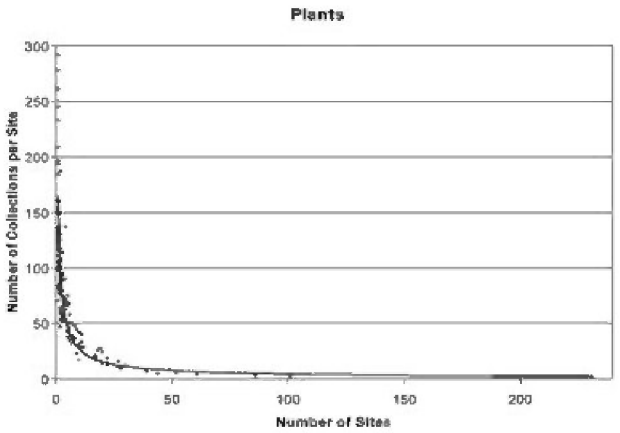


Figure 8. The number of plant collections per site ranges from 231 sites that have only one collection to one site with 292 collections (two sites, one with 448 collections and one with 695, were not included in the graph).

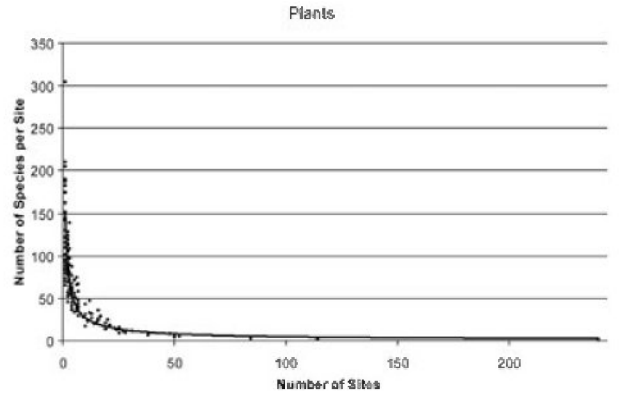


Figure 9. The number of times each species has been collected, ranging from 240 sites that have only one species to one site with 485 species.

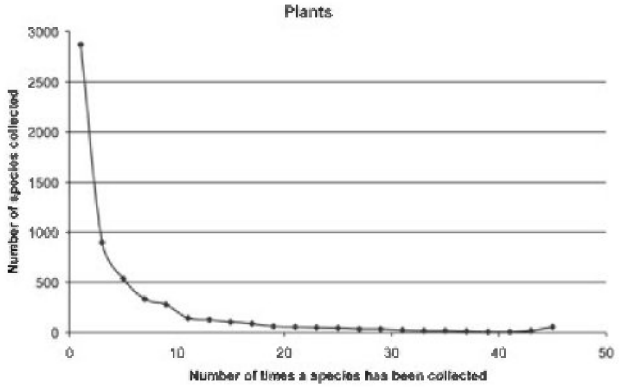


Figure 10. The number of times a plant species has been collected.

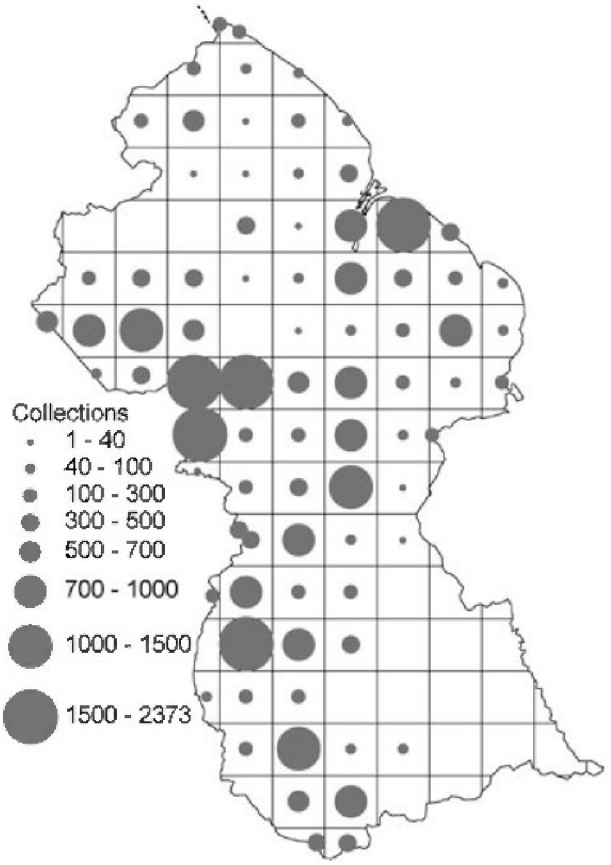


Figure 11. Guyana with a 50 km^2 grid. The size of the circles indicates the extent to which the grid square has been well collected for plants.

variants of ED, discrete and continuous (see Faith & Walker, 1996), is used. The tool selects random but uniform 'candidate sites' throughout the geo-space and these form the 'demand points' that are the basis for calculating ED's p-median scores. This approach solves a problem discussed in the literature (see Faith, 2003) with regard to discrete ED: demand points are not uniformly distributed in the space and so may create biases. Our uniform distribution of demand points avoids this and also sidesteps some of the computationally demanding requirements of continuous ED. The tool has four input layers:

(1) *Survey domain*: a grid that defines the surveyable area of the region of interest. In this case, the domain is the country of Guyana; in some instances, it is the country minus its south-east corner.
(2) *Existing survey sites*: the location of any existing survey sites for the biological group of interest.
(3) *Candidate sites*: the set of sites located randomly throughout the surveyable domain of the region of

Table 3. Number of plant collections made at each site

No. of plant collected	No. of sites	No. of plant collected	No. of sites	No. of plant collected	No. of sites
1	231	48	8	98	1
2	101	49	5	99	2
3	86	50	7	101	1
4	61	51	6	102	2
5	44	52	3	104	1
6	52	53	4	105	2
7	39	54	2	106	1
8	49	55	3	107	2
9	28	56	3	108	2
10	28	57	3	111	1
11	26	58	6	114	3
12	30	59	2	116	1
13	22	60	3	117	2
14	20	61	2	118	1
15	22	62	4	121	1
16	27	63	2	122	1
17	10	64	2	124	1
18	17	65	3	125	2
19	22	66	5	126	1
20	17	67	4	128	1
21	17	68	6	129	1
22	13	69	5	130	2
23	9	70	4	131	1
24	20	71	1	132	1
25	11	72	4	134	1
26	18	74	3	135	1
27	19	75	5	136	1
28	11	76	2	137	4
29	12	77	2	148	1
30	10	78	2	150	2
31	10	79	3	151	1
32	7	80	2	154	1
33	11	81	1	159	1
34	11	82	2	160	1
35	6	83	2	161	1
36	8	84	1	184	1
37	8	85	2	187	2
38	5	86	2	194	1
39	7	87	2	197	1
40	11	89	3	209	1
41	6	90	4	233	1
42	10	91	2	245	1
43	5	92	2	250	1
44	5	93	3	261	1
45	9	94	2	278	1
46	6	95	3	292	1
47	2	97	1	448	1
				695	1

Table 4. Number of plants species collected at each site

No. of species collected	No. of times collected	No. of species collected	No. of times collected	No. of species collected	No. of times collected
240	1	10	43	2	87
114	2	7	44	4	88
84	3	4	45	1	89
50	4	2	46	3	90
52	5	12	47	2	91
50	6	7	48	1	92
38	7	6	49	1	95
48	8	6	50	1	96
25	9	5	52	3	97
28	10	2	53	3	98
27	11	3	54	2	99
31	12	3	55	1	101
24	13	5	56	1	102
19	14	2	57	2	106
23	15	7	58	2	107
25	16	4	59	3	109
10	17	3	60	1	110
21	18	4	61	2	113
21	19	4	62	2	115
14	20	4	63	1	116
19	21	3	64	2	119
15	22	1	65	2	120
11	23	6	66	1	122
16	24	7	67	1	123
20	25	4	68	2	124
16	26	2	69	2	129
13	27	1	70	1	130
9	28	1	71	3	139
17	29	5	72	1	142
7	30	3	73	1	144
10	31	1	74	1	152
13	32	6	75	1	162
12	33	4	77	1	174
5	34	2	78	1	175
6	35	3	79	1	182
16	36	1	80	1	187
4	37	2	81	1	188
6	38	3	82	1	190
7	39	1	83	1	205
7	40	2	84	1	210
4	41	1	85	1	305
6	42	2	86	1	485

interest. A sample of 10 000 candidate sites was used throughout the survey domain. This represented a good compromise between computational speed and accuracy.

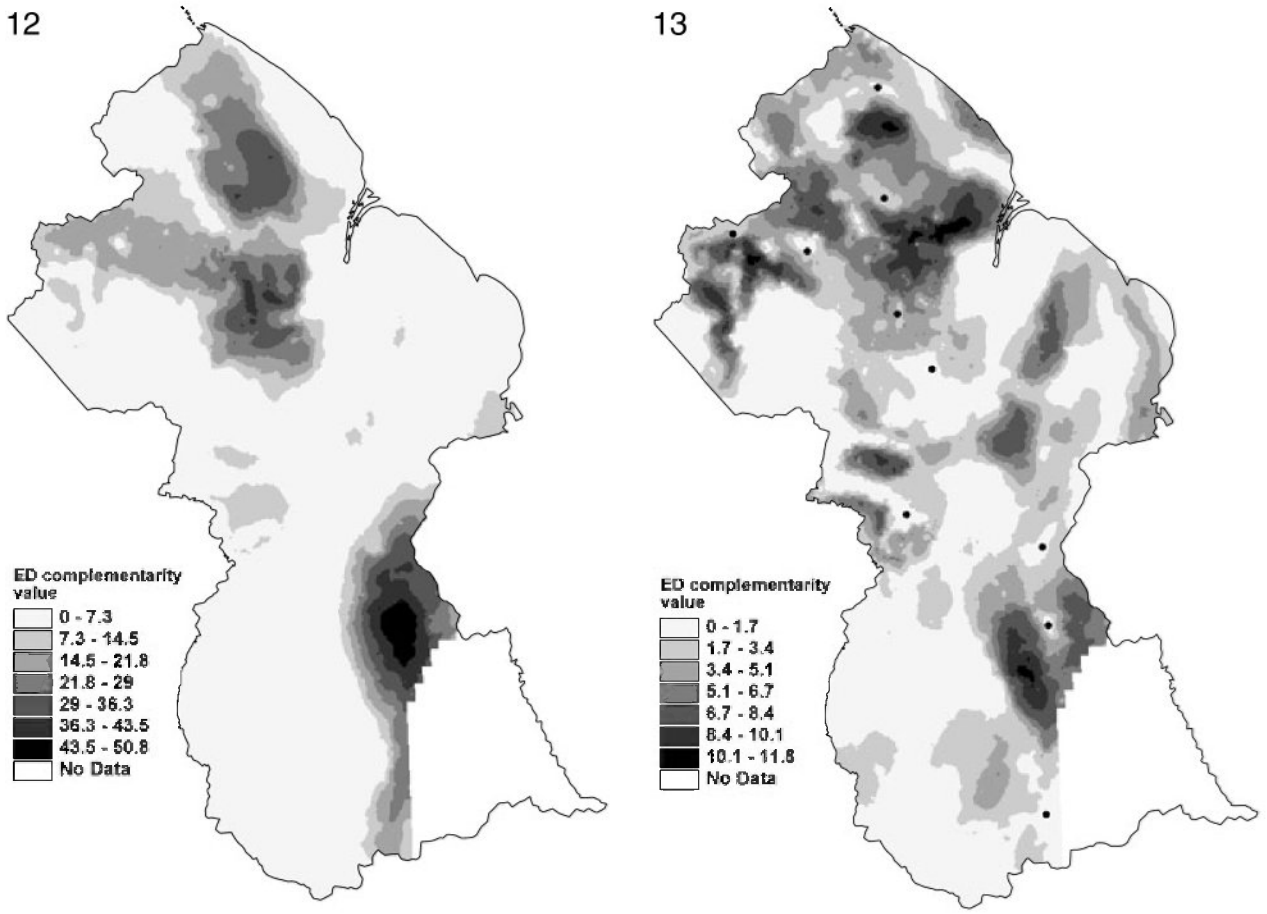
(4) *Environmental layers*: the climate, terrain and substrate layers used to define the environmental space; each of the layers used is input as a grid layer.

The indices used to evaluate the survey coverage and priority for further surveys are all derived from estimates of the likely biological dissimilarity between pairs of sites. A simple linear model based on environmental values at those sites is used:

Biological dissimilarity $= W_1D_1 + W_2D_2 + \ldots W_nD_n$ where D_x is the absolute difference between the values of variable x at the two sites. W_x is the specified

Table 5. Number of plant collections and number of species collected from each of the 50 km^2 grid cells (ordered by number of collections)

50 km Grid no.	No. of collections	No. of species	50 km Grid no.	No. of collections	No. of species
60	0	0	49	106	94
73	0	0	176	110	103
74	0	0	123	112	95
75	0	0	87	117	112
104	0	0	175	134	122
137	0	0	121	165	143
177	0	0	20	174	156
178	0	0	94	194	161
179	0	0	132	210	190
187	0	0	107	216	181
191	0	0	173	220	181
192	0	0	136	222	176
193	0	0	202	232	131
204	0	0	46	235	216
205	0	0	33	246	201
206	0	0	133	257	193
207	0	0	203	296	356
208	0	0	89	304	259
220	0	0	76	328	238
221	0	0	147	359	291
222	0	0	246	370	290
230	0	0	116	383	331
233	0	0	80	384	217
234	0	0	160	387	281
61	1	1	88	392	313
65	1	1	190	398	587
90	1	1	64	454	384
105	1	1	93	467	275
149	3	2	245	493	333
62	3	3	47	610	443
163	4	3	100	613	436
145	4	4	119	622	406
77	8	8	103	655	414
48	14	13	231	691	452
35	43	42	189	701	510
91	45	39	134	746	462
115	48	39	108	752	373
109	51	45	101	786	558
162	54	51	78	806	535
218	57	52	174	809	556
34	68	63	161	813	565
50	71	65	232	912	618
106	74	63	92	914	563
219	74	72	120	962	542
122	78	64	217	1272	785
201	89	85	148	1449	776
95	94	81	102	1472	800
135	94	84	117	1648	925
63	99	91	188	1658	910
146	104	101	79	1713	876
216	105	95	131	1853	1045
			118	2373	987



Figures 12–13. Guyana plant data analysed using only sites that have 40 or more collections. Fig. 12. Areas in most need of plant collecting. Fig. 13. Dots indicate the location of the ten sites where plants should be collected to find the largest number of new records.

weight to be applied to that difference. The weights are provided by the user for each survey analysis. These are based on expert knowledge of the relative importance of each variable in driving biological variation.

Alternatively, the weights and variables can be estimated using a new technique called Generalized Dissimilarity Modelling (see Ferrier, 2002), which models the observed compositional dissimilarities through a regression on environment distances. This new technique is still being integrated into the survey gap analysis tool; once integrated, it will allow the user to determine the variables that best explain the variance in species compositional dissimilarity between sites and will produce a transformed layer for those variables with the correct weighting as predicted by the model (Ferrier, pers. comm.).

Faith & Walker (1996) and Faith *et al.* (2004) describe ED complementarity value in detail. Briefly, it is a measure of the extent to which a selected subset of sites (in this case existing survey sites) covers the space defined by a larger set of sites (in this case the candidate sites). It is calculated as the sum of the distances (in this case the predicted biological dissimilarity) between each candidate site and the nearest existing survey site. Smaller values indicate better coverage of the region of interest. The ED complementarity values generated by the survey gap analysis tool are calculated as the difference between values for a pair of sites. The first value is simply the value achieved with all existing survey sites. The second value, for a given candidate site, is the value that would be achieved if that site were to be surveyed. The difference between these two values therefore measures the improvement in survey coverage that would be achieved by surveying the site in question. The tool uses a greedy algorithm and after each iteration the ED complementarity value is recalculated to reflect the selection of a site. The tool can be used to

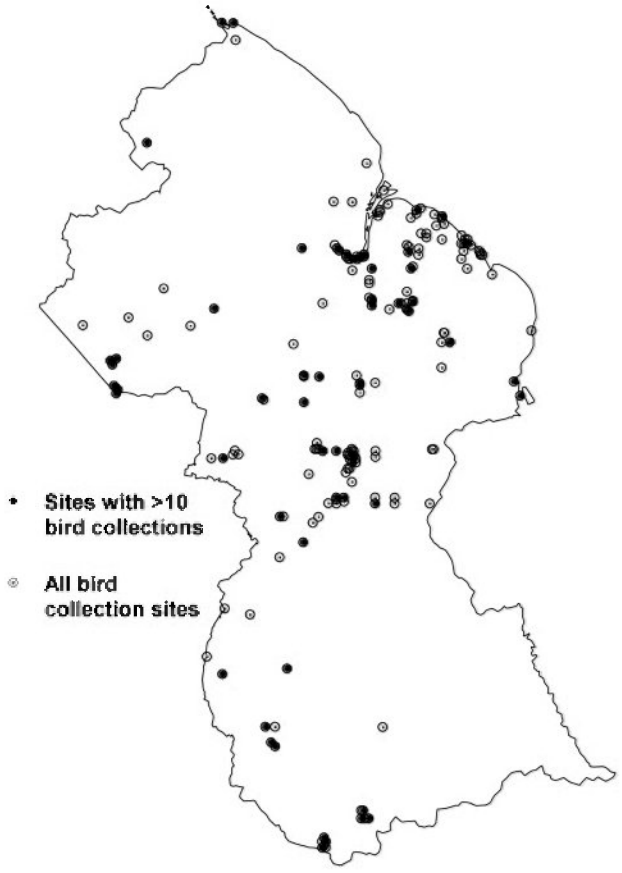


Figure 14. Guyana. Solid dots indicate the bird collecting sites with more than ten birds; the open circles indicate sites with less than ten.

evaluate the relative adequacy of existing survey coverage and to provide guidance in selecting new survey sites.

In this study, the tool was used to select ten new survey sites for each taxonomic group. The choice of ten sites was made because most resident collectors sent to Guyana can plan and execute approximately five major expeditions a year (spending 5–6 weeks in the field for each trip) and such collectors are often appointed for 2 years. So ten sites, or 2 years of work, is the goal of a resident collector. In practice, any number of sites can be chosen and the choice may be made based on the cost in selecting the best set (see Faith & Walker, 1996). Potential survey sites are selected by picking the site with the highest ED complementarity value first. The values are then re-calculated and the site with the next highest value is selected until the number of desirable potential survey sites is reached. The iterative calculation of the ED complementarity value results in a series of sites that are complementary in nature.

For the purpose of this study, several abiotic variables were evaluated, including at a 1 km^2 grid-cell resolution: temperature seasonality, maximum temperature of the warmest period, mean temperature of the wettest quarter, precipitation during the driest quarter, elevation, vegetation classes across Guyana, lithology, and rainfall during the extreme months (January, June, July, October, December). For categorical variables such as vegetation and lithology, an expert scored the dissimilarity between each pair of classes on a scale of 0–1 and these dissimilarities were used in the analyses. Various permutations of these variables were run in sets of five and ultimately the following were selected: December rainfall, October rainfall, maximum temperature of the warmest period, mean temperature of the wettest quarter, and lithology. The geographical distance between existing collecting sites was also used as a variable. A weight was assigned to each variable by an expert as to their importance in the determining the distribution of a taxonomic group (Table 2). Ten thousand candidate sites were located randomly within Guyana and the resulting ED complementarity values for the candidate sites were interpolated, using the spline interpolation function in ArcView across the entire surface of the country. Ten potential survey sites were selected for each taxonomic group and the location of these sites were mapped over the resulting ED complementarity value maps.

RESULTS

There was one issue of general interest relating to all groups investigated. It is well known that the south-east corner of Guyana, the New River Triangle and the surrounding area, inaccessible for political reasons for many years, has little or no reliable biodiversity information available. In the preliminary analyses using all plant data (or data from any taxa) for the entire country, the south-east corner always came up as the most important area for investigation because it has the greatest potential for new records and species no matter what environmental variables were used.

Figure 2 (all maps were drawn using ArcMap; ESRI, 2002) shows the localities where plants were collected. It is no surprise that the survey gap analysis of all plant data (Fig. 3; ED complementarity value range of 0–270) strongly indicated the south-east area as the most important. When ten sites were selected as the most important collecting sites (Fig. 4), eight of the ten were in the south-east. In fact, at least eight of the first ten sites selected for all groups investigated were found there. After the first ten plant collecting sites were selected, the ED complementarity value range dropped to 0–8.5, which indicates the impor-

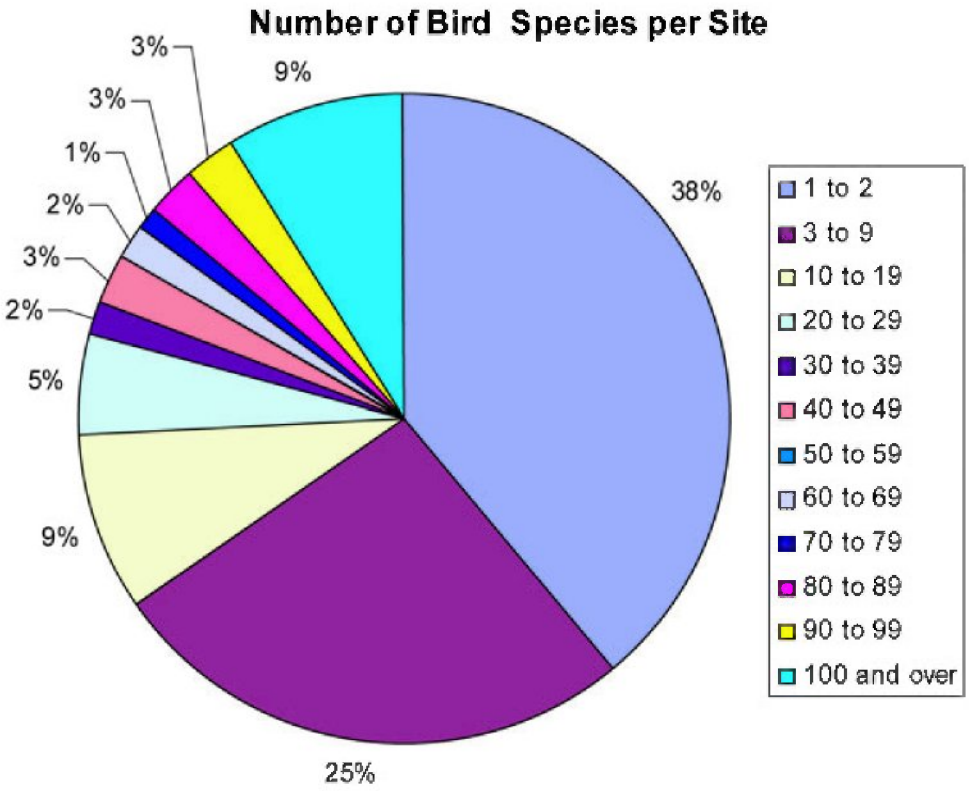


Figure 15. Pie chart showing the number of bird species per site.

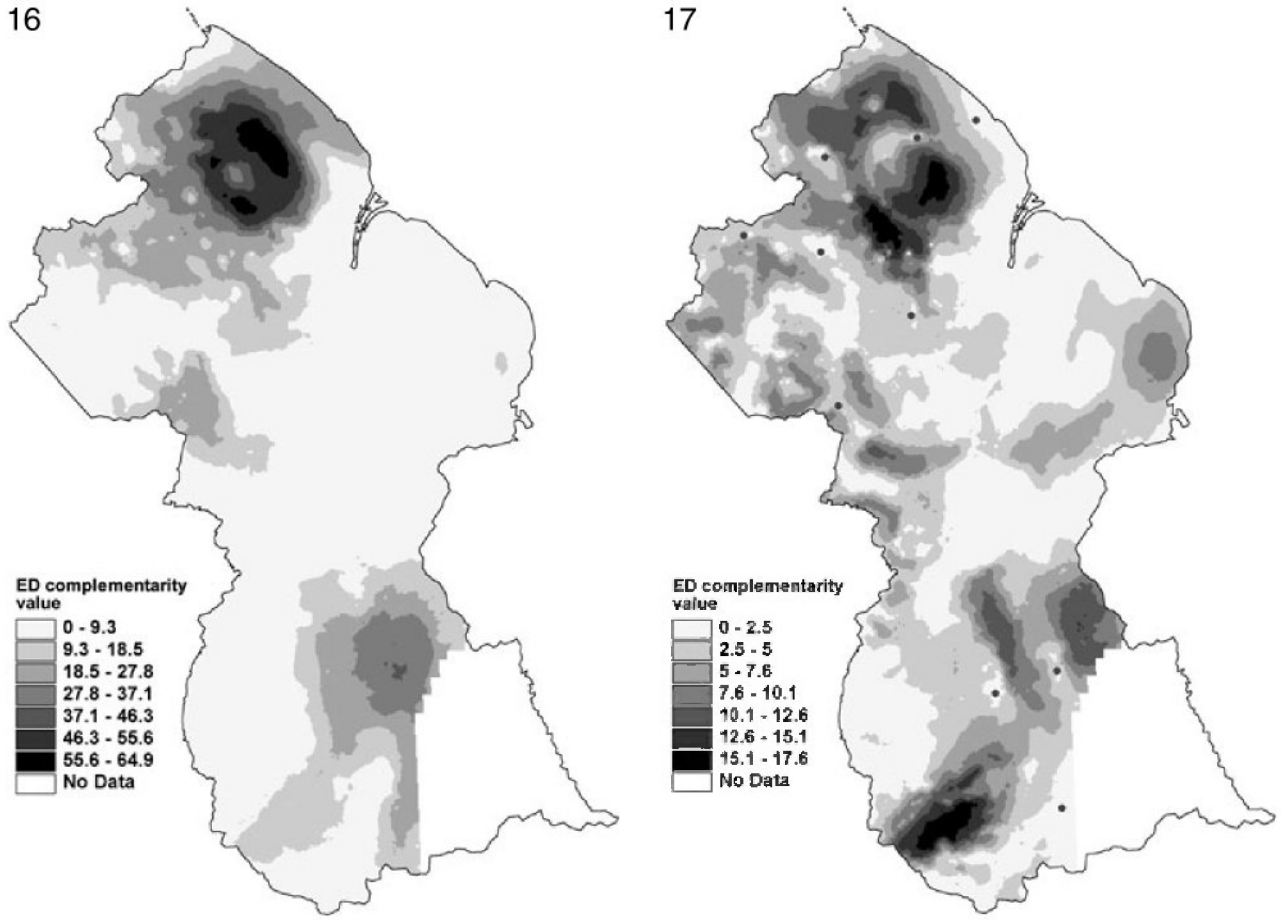
tance of these first ten sites in covering the environmental space. To illustrate how this method works, Figure 5 shows the environmental space created by two abiotic variables: mean temperature during the wettest quarter, and maximum temperature during the warmest period. The black dots indicate the environmental space that had already been sampled and the grey squares show where one should sample to improve the coverage of these two variables.

In Figure 4, the dark areas indicate the next areas that will need to be collected, the ones that become important after the sampling of the first ten sites. However, one should not get the impression that the situation is not improving, because based on the ED complementarity values there is a dramatic improvement in the coverage of the environmental space. The implication of these results is that as soon as the south-east becomes open to collectors, numerous expeditions should take place. It would be best if the location of the expeditions were determined by using the differences in environmental space, but when, where, and how this area is made available is up to the Government of Guyana. Although it is clear that this area should be the number one priority for research expeditions, the fact that it is currently off-limits means that we need to investigate the sites of future expeditions in a way that does not include this part of Guyana. In all the remaining examples, the south-east corner is therefore 'masked' or eliminated from consideration.

PLANTS

The plant data were re-examined with the south-east corner masked. As one might expect, a different pattern emerged. Figure 6 illustrates this pattern with some areas still indicated on the border of the south-east but with new areas in the north and west (ED complementarity value range 0–17.6). Of the ten sites selected (Fig. 7), three remain down on the border with the south-east, but six are in the northern Pakaraima Mountains, mostly north and west of the Mazaruni River and north of the Cuyuni River. The final site is in the extreme southern Pakaraimas.

There is one factor that is not taken into consideration with the survey gap analysis tool, but is important in determining where to collect. The current tool assumes that sites have been collected with a more or less equivalent amount of collecting effort; in other words, that the collecting localities are equally well known. Although this may be the case in some regions, unfortunately Guyana has never been collected in a



Figures 16–17. Guyana bird data analysed using only sites that have ten or more collections. Fig. 16. Areas in most need of bird collecting. Fig. 17. Dots indicate the location of the ten sites where birds should be collected to find the largest number of new records.

systematic fashion. Collectors have tended to focus on groups or locations reflecting their own interest; rarely are there enough collections from an area for it to be considered 'well known'.

In order to evaluate collecting intensity, the data were examined in three ways. First we looked at the number of collections from each site (Table 3); the results are illustrated in Figure 8. The values ranged from 231 sites that had only one collection, to one site with 484 collections and one with 695. These latter two numbers are so much higher than all the others (the next highest is 292) that they were left off of the graph (Fig. 8). Examination of Figure 8 reveals a definite upward turn in the slope at around 40 collections per site; this was selected as an important threshold in knowledge about the site. However, of the *c.* 1400 collecting sites for plants, only 245 had 40 or more collections.

The second method for examining collecting intensity was to look at the number of times each species was collected at each site (Table 4) and in total (Figs 9, 10). The pie chart (Fig. 9) shows that half of the species have only been collected once or twice, 66% have been collected 1–4 times and 75% have been collected 1–6 times. If one uses the criterion that a species needs to be collected ten times or more in order to be relatively sure of its distribution (see Funk & Richardson, 2002; for a discussion) then Figure 10 shows that 80% of the species have not been collected ten times or more.

Lastly, the collecting sites can be aggregated within a larger grid size. Aggregation of several collecting sites in an area may also provide a solution to the collecting effort problem in Guyana. Grids of 1–50 km^2 were examined; only the results of the 50 km^2 grid are presented. A 50 km^2 grid was imposed over the country and all plants collected within each grid cell were tabulated in two ways: the number of collections and the number of species (Table 5). Figure 11 is colour coded and sized to show the number of

Figure 18. Guyana. Dots indicate insect collecting sites.

records for all collecting sites in the grid. Usually, the number of species was more or less close to the number of collections. However, one exception is Georgetown, where there are 695 collections but only 485 species. This discrepancy is most likely due to many older collections having only 'Georgetown' on the label.

Of the 103 cells, nearly 25% have no collections. Of course, such areas are not a problem because they will be picked up by the survey gap analyses. If we examine the remaining 79 grid cells, ten have 15 or fewer collections (13%) while eight (10%) have over 1000, which might be considered an adequate number of collections to get a general idea of what might be found in that grid cell. Actually, in north-eastern South America, an area of 50 km^2 might be expected to have from 1500 to 2000 species (Kelloff & Funk, 2004), and yet only one grid cell has over 1000 species. As tempting as aggregating data may be, the variability of the abiotic variables within each 50 km^2 grid makes interpretation of the results difficult. However, the large discrepancy in the numbers of collections and taxa from each grid cell demonstrates how important consideration of collecting effort may be.

In the final analyses, the raw collecting points were used, but only including sites that had at least 40 collections. The original analyses were re-run; Figures 12 and 13 show the areas deemed important and the ten sites selected. Note the shift in the ED complementarity values before and after we removed the least well known sites (0–17 and 0–51, respectively). The range is greater because we have fewer sites. The location of the ten sites selected has also changed. Three are still in the south-east but one has moved farther north. Three are still north of the Mazaruni and Cuyuni Rivers, but one is now in the north and two are in the east-central Pakaraimas. The ED complementarity value has decreased from 0–51 in Figure 12 to 0–10 in Figure 13.

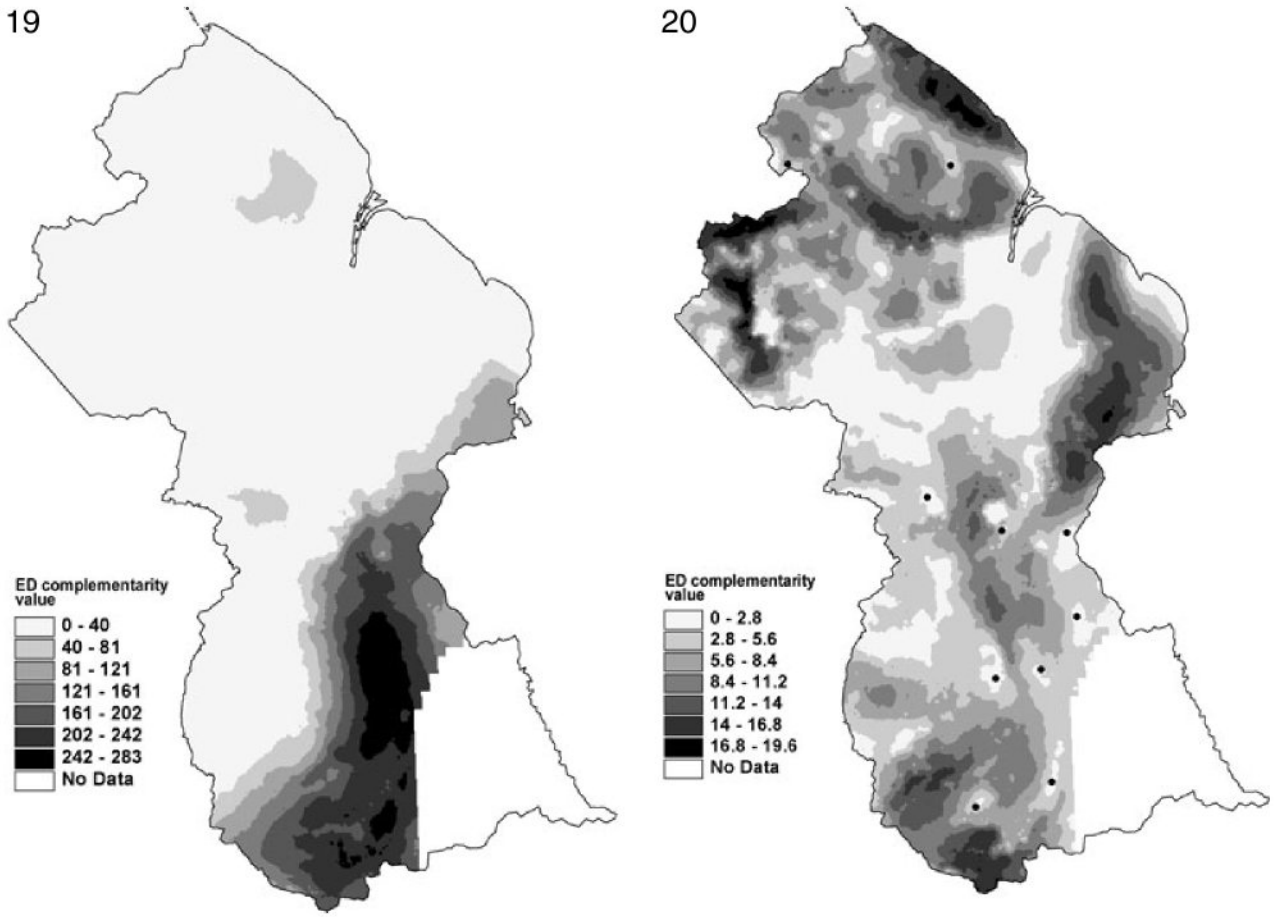
Although many different groups have been examined using a variety of environmental variables, we will only briefly illustrate two others – birds and insects.

Birds

The bird data were examined for the 'number of species per site' but not for the 'number of collections per site' because the bird data did not have duplicate collections of the same species listed. We had available 192 collecting sites (Fig. 14) but many of them had only a few species for the site. Figure 15 shows the number of species collected at each site with 38% (75) of the sites having only one or two species, about one third having ten or more and 16 sites (9%) that had more than 100 species recorded. After careful consideration it was decided to use only the 66 sites (Fig. 15) that had ten or more bird species, because it was felt that fewer than ten did not provide a good basis for evaluating the bird fauna. Figure 16 shows the ED complementarity values for the country, with a range of 0–214. The most important area to sample is the extreme north of the country, which has only two sites with more than ten species. Figure 17 shows the ten sites selected for the next bird expeditions; three are in the north, two in the western Pakaraimas and the other five are spread out along the border with Surinam. The ED complementarity value has decreased and now ranges between 0 and 26. Many of the areas in the north are still indicated as needing work, as well as areas in the south and the Kanuku Mountains in the south-west.

Termites and butterflies

The weakest dataset available is for insects. There are quite a few collecting sites (Fig. 18), but none has more than 15 collections and most have less. Because all the



Figures 19–20. Guyana termite data analysed using all information. Fig. 19. Areas in most need of termite collecting. Fig. 20. Dots indicate the location of the ten sites where insects should be collected to find the largest number of new records.

sites are equally poorly sampled there was no need to adjust for the number of collections. The analysis (Fig. 19) shows the whole of the southern and eastern part of the country in need of work, along with areas in the central- and north-west. The ED complementarity value range is higher than it is for any other group (0–283), but after the ten sites were selected (Fig. 20) it decreases to 0–20. While the value is still high by the standard of the other maps, it is much improved. It seems that the sites selected in the southern part of the country are nearly equidistant from one another and therefore not as controlled as the placement of the sites in the other groups. Perhaps this is because of the lack of data.

We have illustrated higher-level groups of organisms: plants, birds, and insects. However, one can work at any taxonomic level and we have tried using plant families and orders. The only note of caution is that the survey gap analysis method assumes that it is equally likely that a member of the group will be in any habitat. Therefore, if the group in question is confined to areas of higher elevation or savannahs, etc. one must exclude any areas where it is unlikely that the organism will be found.

CONCLUSION

Collecting new specimens or data from the field is a very expensive and time-consuming endeavour. Efforts have to be made to use funds allocated to this task in the most efficient manner. Results from this study show that by considering the distribution of existing collections in terms of environmental space (abiotic variables) and geographical space (geographical distance) and by using a method that is based on the principle of complementarity (which selects potential survey sites to maximize the potential difference in environmental and geographical space between sites) a clearer idea of collecting priorities emerges. This approach can be applied at any spatial scale,

although it is most useful at the regional scale (e.g. a country). The results also demonstrate that for a country the size of Guyana there are numerous undersampled areas. Collecting effort must therefore be taken into consideration for those expeditions where collections were not undertaken in a systematic manner. With the call to endeavour to survey the planet from pole to pole (Wilson, 2000), repeatable, cost-efficient and purposeful surveys are needed. We believe that techniques such as the one described in this study will provide researchers with the best tools to continue the tradition of expeditionary research.

ACKNOWLEDGEMENTS

We thank the following persons and institutions: Tom Hollowell for his help with the figures; the Government of Queensland for funding a fellowship (VAF); the Smithsonian Institution's Biological Diversity of the Guiana Shield Program for the use of their data; the Cooperative Research Centre for Rainforest Ecology and Management for staff assistance; Glenn Manion for help with developing the method and with its application, and Dan Faith for comments on the manuscript. This is publication number 86 in the Biological Diversity of the Guiana Shield Program series.

REFERENCES

Araújo MB, Densham PJ, Lampinen R, Hagemeijer WJM, Mitchell-Jones AJ, Gasc JP, Humphries CJ. 2001. Would environmental diversity be a good surrogate for species diversity? *Ecography* **24:** 103–110.

Austin MP. 1991. Vegetation: data collection and analysis. In: Margules CR, Austin MP, eds. *Nature conservation: cost effective biological surveys and data analysis*. East Melborne: Australia CSIRO, 37–41.

Austin MP, Heyligers HP. 1991. New approaches to vegetation survey design: gradsect sampling. In: Margules CR, Austin MP, eds. *Nature conservation: cost effective survey and data analysis*. East Melborne: Australia: CSIRO, 31–37.

Bell G. 2003. The interpretation of biological surveys. *Proceedings of the Royal Society of London B* **270:** 2531–2542.

Balmford A, Gaston KJ. 1999. Why biodiversity surveys are good value. *Nature* **398:** 204–205.

Brooks T, Hannah L, da Fonseca GAB, Mittermeier RA. 2001. Prioritizing hotspots, representing transitions. *Trends in Ecology and Evolution* **16:** 673.

Darwin C. 1845. *Journal of researches into the natural history and geology of the various countries visited by H.M.S. Beagle under the command of Captain Fitz Roy R.N.* London: John Murray.

Darwin C. 1859. *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. London: John Murray.

Edwards JL, Lane MA, Nielsen ES. 2000. Interoperability of biodiversity databases: Biodiversity information on every desktop. *Science* **289:** 2312–2314.

ESRI. 1999. *Arcview 3.2. Environmental systems*. Research Institute, Redlands, CA (http://www.esri.com).

ESRI. 2002. *Arcmap 8.3. Environmental systems*. Research Institute, Redlands, CA (http//: http://www.esri.com).

Faith DP. 2003. Environmental diversity (ED) as surrogate information for species-level biodiversity. *Ecography* **26:** 374–379.

Faith DP, Ferrier S, Walker PA. 2004. The ED strategy: how species-level surrogates indicate general biodiversity patterns through an 'environmental diversity' perspective. *Journal of Biogeography* **31:** 1207–1217.

Faith DP, Walker PA. 1996. Environmental diversity: on the best-possible use of surrogate data for assessing the relative biodiversity of sets of areas. *Biodiversity and Conservation* **5:** 399–415.

Ferrier S. 1997. Biodiversity data for reserve selection: making best use of incomplete information. In: Pigram JJ, Sundell RC, eds. *National parks and protected areas: selection, delimitation, and management*. Armidale, Australia: Centre for Water Policy Research, 315–329.

Ferrier S. 2002. Mapping spatial pattern in biodiversity for regional conservation planning: where to from here? *Systematic Biology* **51:** 331–363.

Ferrier S, Smith AP. 1990. Using geographical information systems for biological survey design, analysis and extrapolation. *Australian Biologist* **3:** 105–116.

Funk VA, Richardson K. 2002. Systematic data in biodiversity studies: use it or lose it. *Systematic Biology* **51:** 303–316.

Funk VA, Zermoglio MF, Nassir N. 1999. Testing the use of specimen based collecting data and GIS in biodiversity exploration and conservation decision-making in Guyana. *Biodiversity and Conservation* **8:** 727–751.

Gillison AN, Brewster KRW. 1985. The use of gradient directed transects or gradsects in natural resource surveys. *Journal of Environmental Management* **20:** 103–127.

Hooker JD. 1844–1847. *The botany of the Antarctic voyage of HM discovery ships Erebus and Terror in the years 1839–1843*, Vol. I. *Flora Antarctica.* London: Lovell Reeve.

Hooker JD. 1853–1855. *The botany of the Antarctic voyage of HM discovery ships Erebus and Terror in the years 1839–1843*, Vol. II. *Flora Novae-Zelandiae*. London: Lovell Reeve.

Hooker JD. 1855–1860. *The botany of the Antarctic voyage of HM discovery ships Erebus and Terror in the years 1839–1843*, Vol. III. *Flora Tasmaniae.* London: Lovell Reeve.

von Humboldt A. 1814–1826. *Personal narrative of travels to the equinoctial regions of the new continent during the years1799–1804*. London: Longman, Hurst, Rees, Orme & Brown.

Kelloff C, Funk VA. 2004. Phytogeography of Kaieteur Falls, Potaro Plateau, Guyana: flora distributions and affinities. *Journal of Biogeography* **31:** 501–513.

Margules CR, Austin MP. 1994. Biological models for monitoring species decline: the construction and use of data bases. *Philosophical Transactions of the Royal Society of London B* **344:** 69–75.

Myers N, Mittermeier RA, Mittermeier CG, da Fonseca GAB, Kent J. 2000. Biodiversity hotspots for conservation priorities. *Nature* **403:** 853–858.

NSW NPWS. 1998. Vertebrate fauna Survey. NSW comprehensive regional assessment project report. Sydney: NSW NPWS.

Olson DM, Dinerstein E. 2002. The Global 200: Priority ecoregions for global conservation. *Annals of the Missouri Botanical Garden* **89:** 199–224.

Pennisi E. 2000. Taxonomic revival. *Science* **289:** 2306–2308.

Richardson KS, Funk VA. 1999. An approach to designing a systematic protected area system in Guyana. *Parks* **9:** 7–16.

Suarez AV, Tsutsui ND. 2004. The value of museum collections for research and society. *Bioscience.* **54:** 66–74.

Sugden A, Pennisi E. 2000. Diversity digitized. *Science* **289:** 2305.

Wallace AR. 1869. *The Malay Archipelago; the land of the orang-utan and the bird of paradise. A narrative of travel with studies of man and nature, 2 volumes*. London: Macmillan.

Wallace AR. 1876. *The geographical distribution of animals*, 2. New York: Harper and Brothers.

Wilson EO. 2000. A global biodiversity map. *Science* **289:** 2279.