Collecting biodiversity

Padmattie Persaud Haripersaud

Collecting biodiversity Padmattie Persaud Haripersaud

Collecting biodiversity

Plant collecting in Guyana

Padmattie Persaud Haripersaud

The research that was carried out in this thesis was carried out within the framework of the Plant Ecology and Biodiversity group, Institute of Environmental Biology, Utrecht University

www.bio.uu.nl/~boev/

Padmattie Persaud Haripersaud

Collecting biodiversity

ISBN: 978-90-393-5105-5

Cover: Photo *Eperua falcata* flower made by Lubbert Westra in Guyana

Printed by Ponsen & Looijen of GVO printers & designers B.V.

Designed by Kooldesign Utrecht

Collecting biodiversity

Over het botanisch verzamelen

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit Utrecht
op gezag van de rector magnificus, prof. Dr. J.C. Stoof,
ingevolge het besluit van het College voor Promoties
in het openbaar te verdedigen
op maandag 21 september 2009 des middags te 2.30 uur

door
Padmattie Persaud Haripersaud

geboren op 15 oktober 1967 te Annandale, Demerara Guyana

Promotor: Prof. Dr. M.J.A. Werger

Co-promotor: Dr. H. ter Steege

Universiteit Utrecht

NWO

Netherlands Organisation for Scientific Research

Contents

Chapter 1

General introduction

1

Species distribution information is fundamental for biodiversity studies. One of the main purposes of biodiversity studies is to assist the decision-making process for conservation and preservation of habitats and species (Mace 2004). Yet there is scarcity of species distribution information for most species, especially those occurring in the tropics. Gathering of new species information is becoming increasingly difficult mainly because of the declining number of professional taxonomists and fewer resources (Erkens & Baas 2008). Species information for specific geographical areas is scattered in institutions worldwide. This information is now becoming linked together through initiatives such as the Global Biodiversity Information Facility (GBIF) and is freely available on the internet. The GBIF portal now contains more than 171 million specimens. These specimens are an excellent data source for biodiversity studies because of their large geographical and environmental coverage and in some cases they are the only source of distribution information available. Not only species data are becoming rapidly available on the internet, but also environmental data (FAO 2002; Hijmans 2005; FAO 2006). In addition, improved modeling software packages that are specialized in linking these data sources are also freely available (Elith *et al.* 2006; Phillips *et al.* 2006). This free availability of information and software has substantially increased the potential use of the specimen data and ecologists are increasingly interested in this data source for biodiversity research. However this free availability of the data also presents the danger that they might be used without understanding the pitfalls associated with the way in which these data were collected and their taxonomic and spatial quality.

There are some drawbacks to the use of herbarium data in describing biodiversity. The original aim of plant collecting was to describe new species and prepare regional floras and monographs and specimens were collected in a specialized way so as to maximize the number of species found. Herbarium specimen data does not represent the community structure well (van Gemerden *et al.* 2005) because usually only one specimen per species is collected during an expedition regardless of whether the species is common or rare. Further, the

collecting localities are often not randomly distributed across the study area (Reddy & Davalos 2003; Kadmon *et al.* 2004; Loiselle *et al.* 2008). Moreover, botanists are known to show preferences for certain taxa, easily accessible areas and to plan their expeditions during the dry season (ter Steege & Persaud 1991; Reddy & Davalos 2003; Kadmon *et al.* 2004; Loiselle *et al.* 2008). As a result when herbarium data is used for evaluating biodiversity studies, their potential limitations must be addressed.

The Guianas

The Guianas are made up of three countries Guyana (formerly British Guiana), Suriname and French Guiana (an overseas Department of France). These countries are located in the north-eastern part of South America (Fig. 1). The area is very sparsely populated (the 2009 estimates of the population for Guyana being 772,298, Suriname, 481,267 and French Guiana 209,000) (www.cia.gov/library/publications/the-world-factbook/geos/ns.html) and most of the population resides along the coast. Geographically, the area is delimited in the north by the Atlantic Ocean, Brazil in the south and Venezuela in the west (Fig. 1). The total area is about 461,768 km^2 (Guyana, 214,970 km^2; Suriname, 163,270 km^2; French Guiana 83,534 km^2). The Guianas, together with parts of Brazil, Venezuela and Colombia form the Guiana Shield. This shield lies on an old Precambrian geological formation. The shield area is characterized by low disturbance partly because of the low population density and few natural disasters and by low soil fertility and productivity and few roads (Hammond 2005). In addition, the area has a high biodiversity and many of the species are habitat specialists (Fanshawe 1952; Richards 1996; Hammond 2005).

Forest regions and species composition of the Guianas

Distinct forest regions have been recognized in the Guianas based mainly on soil, climatic and attitudinal variation and forest composition (Fanshawe 1952; ter Steege & Zondervan 2000). Five forest regions were described by ter Steege and Zondervan (2000), three of which extend across all three of the Guianas – forests in the coastal plain, forests on the white sand formation and forests on the southern peneplain. Two additional forest regions are recognized in Guyana – the forests in the North West District and forests in the Pakaraima-Central Guyana Upland region. Within each of these regions a number of forest

formations are found which might overlap among the regions and share many species in common (Fanshawe 1952; Lindeman 1953; Lindeman and Moolenaar 1959). More than 7,000 angiosperms are known to occur in the Guianas. About one-third of the species have a wide distribution range and occur in all three of the Guianas (Funk *et al.* 2007) but many others are thought to be endemics or restricted range species (Fanshawe 1952; Richards 1996; ter Steege *et al.* 2000).

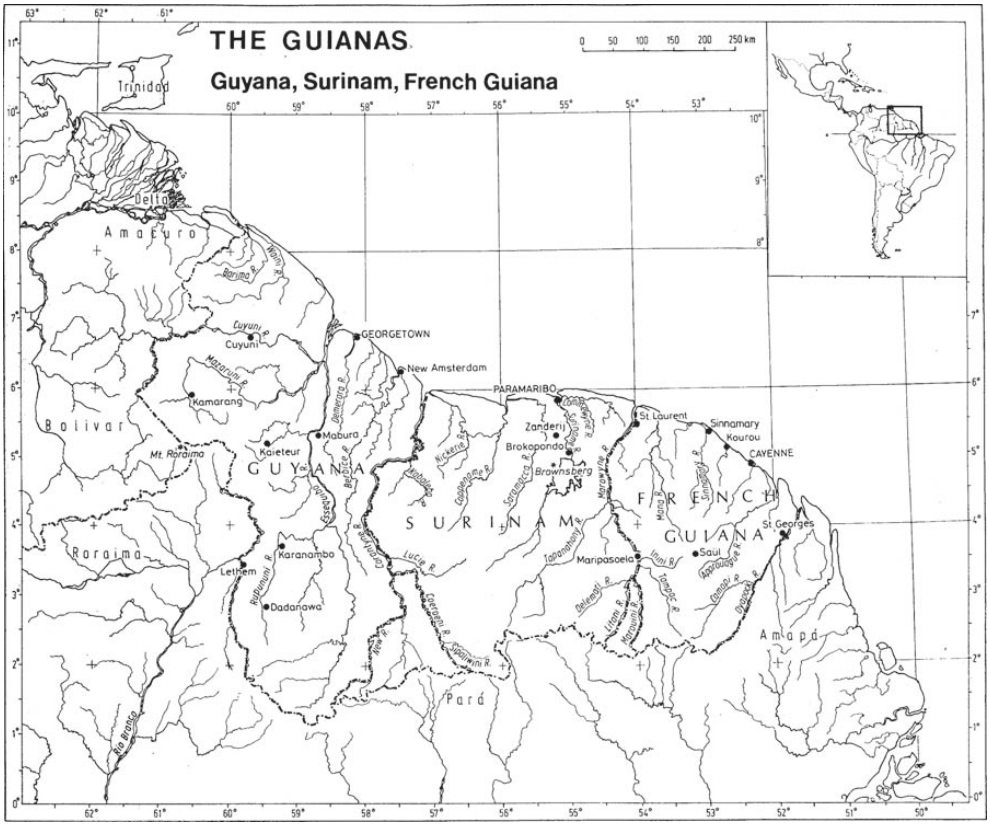

Fig. 1. Map of the Guianas, indicating the most commonly known locations and rivers. (Source: http://www.bio.uu.nl/~herba/public.html).

Aim of the study

As discussed above, no random sampling strategy was applied when herbarium were specimens collected and this might influence the usefulness of the collections for biodiversity studies that require that species data is collected by random sampling. Centuries of botanical collecting in the Guianas has resulted in close to 200,000 angiosperm specimens. Although many studies have used herbarium data to describe biodiversity in the Guianas, most of the studies were focused in one of the countries in the Guianas even though these countries have many species and forest regions in common (o.a Poncy *et al.* 1998; Funk *et al.* 1999; ter Steege *et al.* 2000; Funk and Richardson 2002) and the Flora of the Guianas project dates back to 1984. This is the first time that the specimens of all three of the Guianas are combined in one database and biodiversity studies are undertaken at the scale of the Guianas using this data source. None of the studies has assessed the extent of the biases associated with herbarium data collected in the Guianas. This thesis is based on the idea that if herbarium data is used for biodiversity purposes the quality of the species data may compromise the reliability of the results. This thesis has two main aims. Firstly, to quantify the biases associated with the data when it is used for a number of applications and to address the consequences of the biases. Secondly, it aims to develop models (or ways) to correct for these biases and describe biodiversity across the Guianas.

Chapter 2 is a historical description of the data source. It describes the historical accumulation of species and specimens in the Guianas, the distribution of specimens among species, genera and families, growth forms and countries.

Chapter 3 quantifies biases in herbarium data in time and space. The consequences of these biases when the data is used in applications such as species richness estimation, species distribution mapping, taxonomic and phenological studies are discussed.

Chapter 4 presents a simple simulation model which uses the species abundance distribution of trees in a large area, based on plots data, to simulate the abundance distribution in a herbarium, following botanical collecting strategies.

Chapter 5 examines the use of herbarium data collected at different intensities in assessing the roles of dispersal and the environment in shaping the floristic composition.

Chapter 6 determines the species richness and endemism patterns based on collections and two environmental datasets both including climatic and altitudinal and one including also soil variables. The species richness patterns using the Guianan data are verified with patterns obtained with a limited dataset of the full Neotropics.

Chapter 7 summarizes the findings and discusses the potential use of herbarium data collected in the Guianas for biodiversity purposes.

Chapter 2

Species abundance, distribution and diversity in time and space after centuries of collecting in the Guianas

2

with Hans ter Steege, Jean-Jacques de Granville, Hervé Chevillotte and Michel Hof

Abstract

For centuries botanists have been exploring different areas of the Guianas. In this chapter after augmenting botanical specimen data from different sources, we used the comprehensive database to examine the historical accumulation of species and specimens and the pattern of geographical expansion of collecting localities. We then looked at how specimens in the database were divided among families, genera and species, growth forms and countries. The 7,146 species in the database were distributed in unequal proportions over many families and genera and growth forms. Although many species collected were common to Guyana, Suriname and French Guiana, unique species for the countries were also collected. Despite the high collecting effort many areas still remain under-collected.

Centuries of botanical collecting in the Guanas

Botanical knowledge of the Guianas (Guyana, Suriname and French Guiana) has accumulated for more than four centuries (Ek 1990, 1991; Hof unpublished). However, specimens from botanical expeditions are scattered in herbaria worldwide and it is only during the past two decades that specimen label information is becoming digitally available. To our knowledge this is the first attempt to compile the digital data from primary sources to form a comprehensive database for the Guianas. Here we describe the historical variation in the accumulation of specimen and species data and the pattern of expansion of the geographical area in which the specimens were collected. We then analyse how the specimens in the database are divided among species, growth forms and countries.

Building a digital Herbarium

The angiosperm database of the Nationaal Herbarium Nederland, Utrecht branch (c. 115,000 specimens) forms the backbone of this study, and was augmented with data from herbaria which house specimens collected in the Guianas and from botanists who have worked there. Many herbaria have contributed digital databases - the Institut de Recherche pour le Développement, IRD, Centre de Cayenne, Cayenne (c. 79,000 specimens), the New York Botanical Garden (http://www.nybg.org/bsci/res/resproj.html) and the Missouri Botanical Garden. Also included were digitized accession records of the Jenman Herbarium and the Jonah Boyan Herbarium of Guyana, and of Lands Bosbeheer and Bosbeheer Suriname of Suriname. Finally, we included lists of taxa collected by different botanists from the Smithsonian Institution, Washington, D.C. (Hollowell *et al.* 2000; 2003; 2004). Duplicate records, i.e. of specimens with the same taxonomic identity, collection date, botanist and number, were removed from the database. The species names, as shown on each label, were updated, based on the Smithsonian 2005 Web Listing of Plants of the Guiana Shield and the Guianas (www.mnh.si.edu/biodiversity/bdg/planthtml/index.html) and the W3tropicos website (www.tropicos.org). The Angiosperm Phylogeny Group II system was used for the systematic classification of families and genera (www.mobot.org/MOBOT/research/APweb).

Only specimens that were identified to the species level were used in the analysis and infra-specific information was not used. Introduced species were removed from the database because most of these grow in populated areas and the focus was on natural ecosystems of the Guianas. We used the W3tropicos and the Smithsonian's 2005 Web Listing of Plants of the Guiana Shield websites to determine whether species occurring in the original database were introduced or naturally occurring. Information on latitude and longitude was taken from the labels when available. National gazetteers were used to fill the gaps when information on latitude and longitude was lacking but a locality name was present on the label, or were traced from field notes. Specimens that lacked information on the date of collection, specimen identity and/or the collection locality were not used for the analysis.

After compiling the data from different primary sources and removing the duplicate specimens and introduced species, the database contained a total of

190,398 specimens. Of these specimens, 168,487 contained complete species identification and locality data and were used for analysis. Although this database does not contain all of the specimens, we feel it covers more than 85% of all angiosperm specimens from the Guianas. The specimens were collected by 560 botanists (Fig. 1). Some spent most of their careers collecting in the region and 40 (about 7%) collected more than 1,000 specimens. However the majority did not collect as many specimens and 232 (about 42%) collected 10 or less specimens.

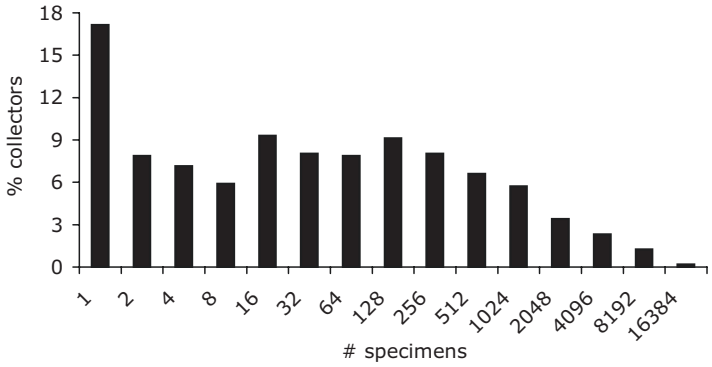

Fig. 1. The frequency distribution of specimens among 560 botanists who collected plant specimens in the Guianas.

Historical development of the specimen data

The oldest known specimen of the Guianas was collected by A. Chastelein in Suriname in 1661 (Ek 1991). The oldest known botanical expeditions in Guyana and French Guiana date back to the eighteenth century (Ek 1990; Hof unpublished). However, there is not much information on the physical location of specimens collected during the seventeenth and eighteenth centuries. The oldest specimen in the database was collected in 1804 (Fig. 2).

The specimens in the database were collected between 1804 and 2004 (Fig. 2 and 3). The number of specimens increased gradually between 1804 and 1953 and then very rapidly until 2004 (Fig. 2). The geographical area in which the specimens were collected expanded from 1804 to 2004. Some areas, especially those closer to cities and research institutions, were repeatedly visited by botanists.

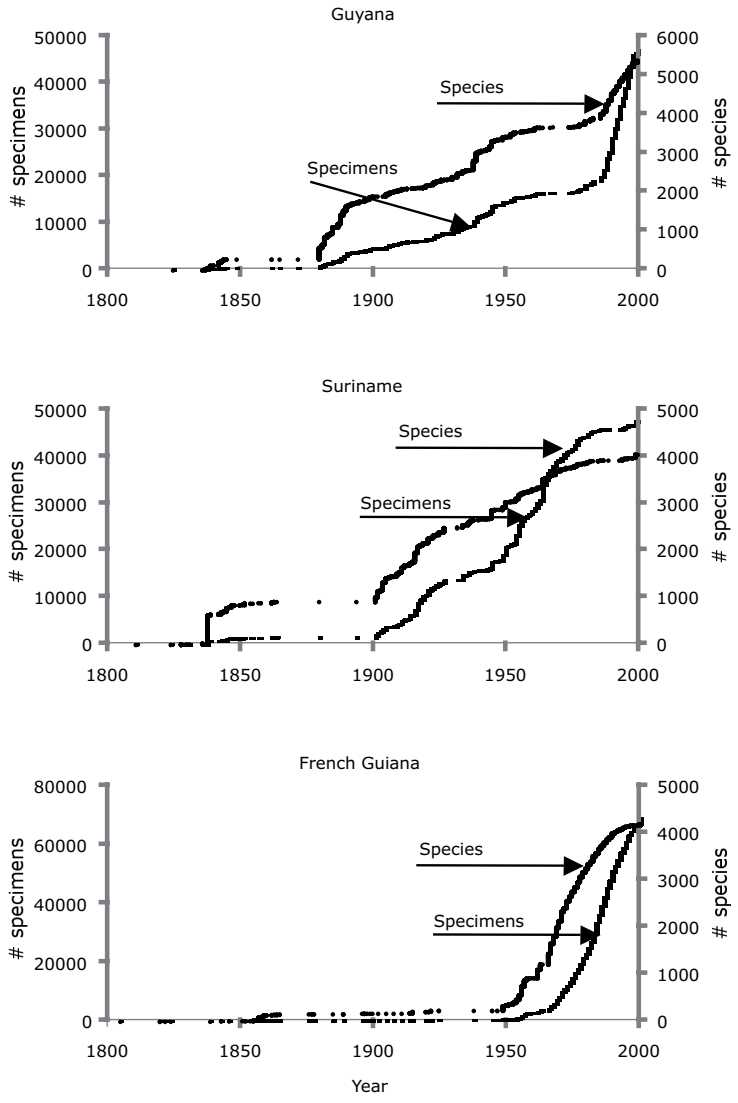

Fig. 2. The number of plant specimens and species in herbaria collected over time. Species and specimens for Guyana, Suriname and French Guiana were orded according to the year in which they were collected and the cumulative number of specimens or species was plotted as a function of the year in which they were collected. Species and specimens increased unevenly through time and accumulation in French Guiana occurred mainly after the 1950's.

Between 1804 and 1853, despite the low number of the specimens (1,728 or 1%) that was collected, 1107 (15%) species present in the database were collected. There were two main peaks in species accumulation during this time period. The first peak occurred in 1837 due to the collecting effort by F.L. Splitberger in Suriname, mainly in the vicinity of Paramaribo and the plantations. The second peak was due to the collecting effort of Robert H. Schomburgk in the following year. He was given the assignment by the British Royal Geographical Society to survey and mark off the boundaries of British Guiana (now Guyana) and during these boundary expeditions he collected most of his specimens (Ek, 1990). The low number of specimens collected before 1854 is not only because of the low collecting effort but also the label information of some of the older specimens was either illegible or insufficient. Most of the specimens were collected in Suriname and the least in French Guiana. The database contains only 18 specimens collected in French Guiana before 1854. The main reason is that we did not access the specimens of the Museum d'Histoire Naturelle (Paris) where most of these historical specimens are stored because no digital data was available from the herbarium at the time we were compiling the database. In the period between 1767 and 1876 cartographers were appointed to demarcate the boundaries of French Guiana, and botanical collecting (mostly lower plants) was a mere side activity. The more serious botanists were at that time given the task to research useful plants for food and medicine and forestry botany was not a priority (Hof, unpublished).

By the end of 1903, 7,979 (5%) specimens and 2,503 (35%) of the species in the database had been collected (Fig. 1) and the geographical area which the botanical expeditions covered increased (Fig. 2). Most of the specimens collected between 1854 and 1903 were collected in Guyana and the least in French Guiana. During this period the collection effort of G.S. Jenman in various geographical areas in Guyana (1879-1887) was responsible for the main peaks in new species accumulation. A second peak in species accumulation was caused in 1903 by G.M. Versteeg collecting along the Gonini River in Suriname. French Guiana continued to be poorly collected. P.A. Sagot and E.M. Mélinon were the most important contributors to the botanical knowledge of French Guiana but their activities were restricted to the coast as the interior was considered unsafe for travelling and collecting due to disorder after the discovery of gold (Hof, unpublished).

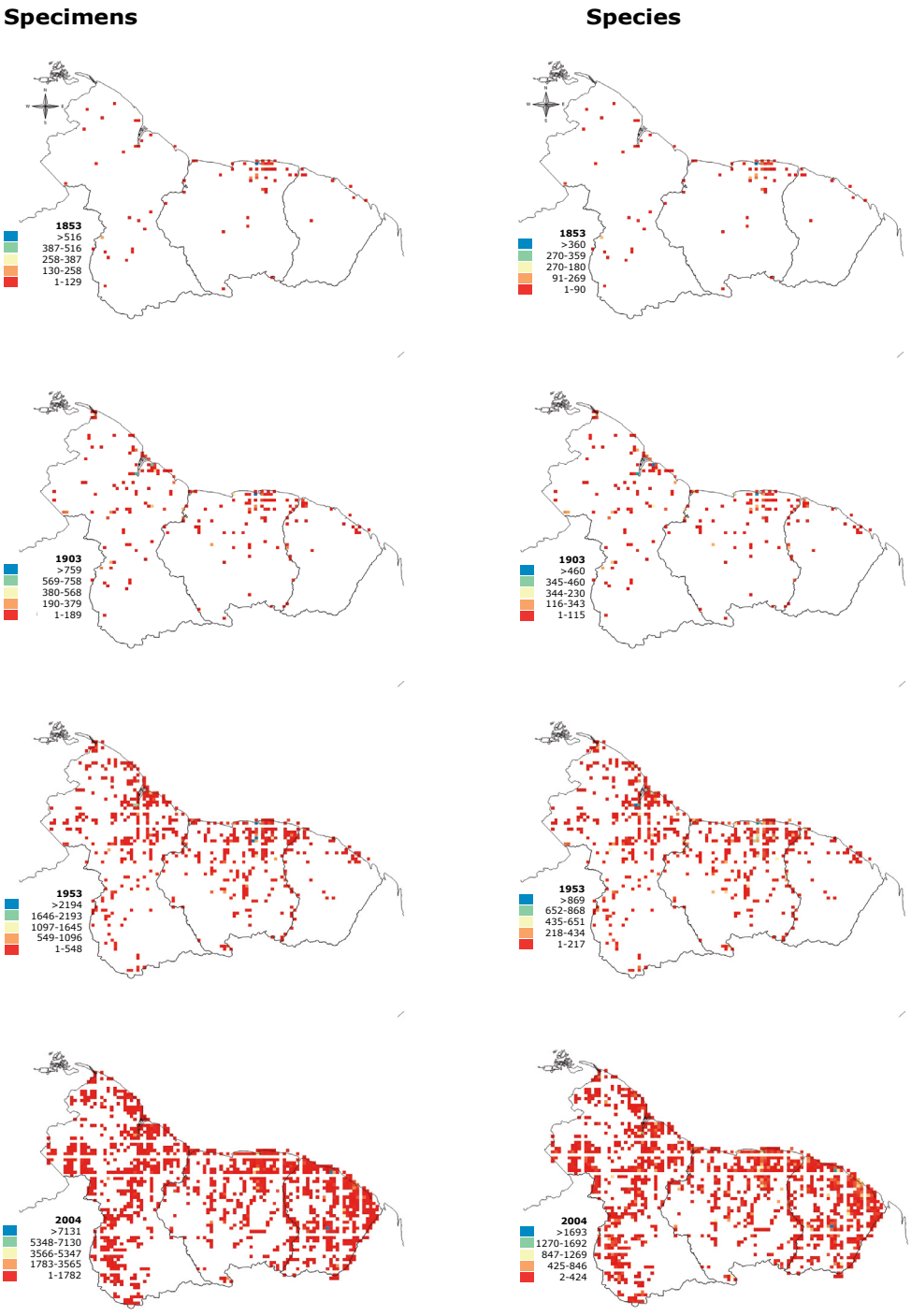

Fig. 3. The number of specimens (N) and species (S) collected in the Guianas over cumulative 50-year time periods (1804-1853, 1804-1903, 1804-1953 and 1804-2004). The area is divided in 5 x 5 arc-minutes resolution and the number of species and specimens per grid cell were counted. In addition to the area visited, both the species and specimens increased over time.

By the end of 1953, 39,666 (24%) specimens and 4,658 (65%) of the species in the database had been collected. There were four main peaks in species accumulation during this time period. The first peak was caused by the effort of botanists collecting under the 'Boschwezen' number series, mainly in the Zanderij, Sectie O, Brownsberg areas and along the Nickerie River in Suriname. The second peak occurred in 1937 due to the collecting effort of N.Y. Sandwith around the Mazaruni Station and of A.C. Smith collecting around the Kuyuwini landing and in the Rupununi area in Guyana. In the same year, H.E. Rombouts collected specimens during his expedition to the Marowyne, Lawa and Litani rivers, causing a large increase of the number of new species known for Suriname. A third peak occurred in 1938 due to the collecting effort of A.C. Smith in the Rupununi area and the Kanuku and Iramaikpang Mountains areas and in the same year due to the effort of botanists collecting under the 'Forest Department' number series in the Moraballi Creek area and around Mazaruni Station in Guyana. The last peak in this period occurred in 1944 caused by the effort of B. Maguire in several geographical areas of Suriname and the Forest Department along many tributaries of the Essequibo River in Guyana.
In the period 1954 to 2004 collecting strategies changed from enriching species lists of individual botanists to enriching species lists of target geographical areas and a large number of specimens in the database were collected in this period. There were three major peaks in species accumulation. The first peak occurred in 1963 mainly due to the collecting effort of B. Maguire along with a number of botanists to the Wilhelmina Mountains of Suriname. A second peak occurred in 1987 mainly due to the efforts of J.J. Pipoly particularly in the Ayanganna Mountains and M.J. Jansen-Jacobs in the Rupununi area of Guyana. The third peak occurred in 1989, mainly due to collecting efforts of J.W. Hahn in the Paramakatoi area, L.J. Gillespie in the Kaieteur and Kanuku Mountains area, M.J. Jansen-Jacobs in the Gunns strip area in Guyana and J.-J. de Granville in the Monts Atachi Bacca area in French Guiana. During this period large scale forestry inventories were carried out in all three countries and permanent research centres were set up. Of the 2,485 new species that were added to the database during this time period, 1,060 were first collected in French Guiana, 960 came from Guyana and 501 came from Suriname.

A rich flora with over 7,000 species of flowering plants

The specimens are not distributed evenly among 183 families, 1,525 genera and 7,146 species (Appendix 2.1). The best represented families in the herbarium are Fabaceae, Rubiaceae, Melastomataceae, Poaceae and Cyperaceae. The ten most specimen-rich families account for 71,101 (about 42%) specimens and 3,045 (43%) species in the database but there are also 12 families with just one or two specimens. A few species were collected in high frequencies (Fig. 4). The highest frequencies for a single species were recorded in French Guiana – for 65 species more than 100 specimens per species were collected whereas only seven and five species were collected at such high frequencies in Guyana and Suriname, respectively. On the other hand, many species were collected in low frequencies (Fig. 4).

More species were represented by one specimen (singletons) in Guyana (1,295 or 24% of all species) than in Suriname (801 or 20%) and French Guiana (655 or 15%). The oldest singleton in our database was collected in 1827 but it must be noted that most of the singletons have been collected after the 1980's (Fig. 5). The singletons were collected all over the Guianas (Fig. 6). The number of singletons in the herbarium may be related to the number of specimens collected for a specific area, however. This is because only one specimen per species is usually collected per expedition and a singleton in the herbarium database is a sampling artefact. To correct for sampling artefact, we first fitted a logarithmic function through the relation of the number of singletons to the number of specimens and then calculated the residuals from this line. Nine 1-degree grid cells showed an unusually high number of singletons for the number of specimens collected (Fig. 6). One of these sites is in Saül of French Guiana, five sites are in the Pakaraimas area and three sites are in the Rupununi Savannah area of Guyana. We may conclude that such areas have a very characteristic species composition, and they are known to have a high concentration of endemics (Berry *et al.* 1995).

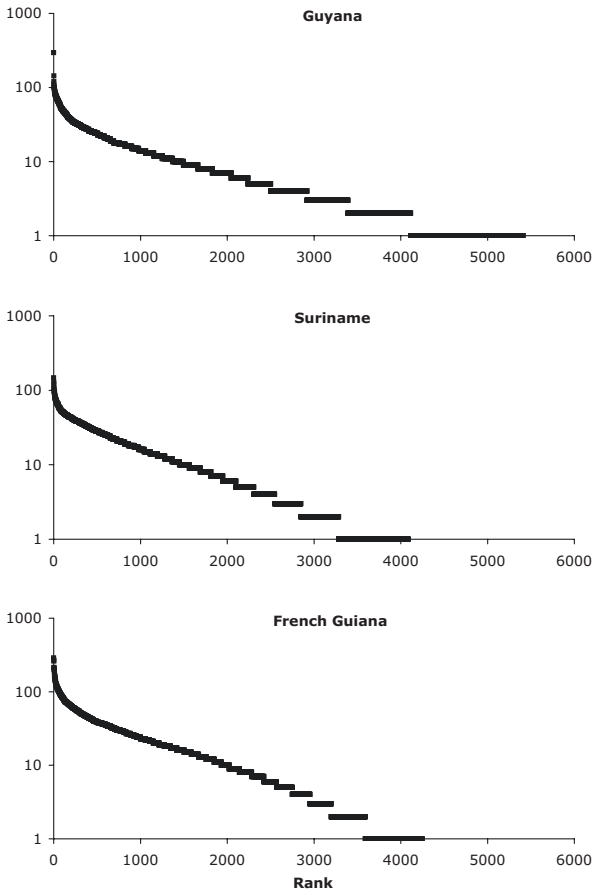

Fig. 4. Rank abundance curves for herbarium specimens collected in Guyana, Suriname and French Guiana. Species were ranked according to their increasing abundance and the abundance (log scale) was plotted against the species rank.

Most specimens are trees

Specimens were not equally distributed among climbers, epiphytes, herbs, palms, shrubs and trees (Fig. 7). The highest fraction of the specimens collected was trees. Palms are the least represented in the herbarium and this is probably because generalist botanists avoid collecting palms due to the physical difficulty associated with collecting the specimens. Also palm specimens are represented by few species in the Guianas (Funk *et al.* 2007). Among the three countries there are no substantial differences in the fractions of climbers, epiphytes, herbs, palms, shrubs and trees collected.

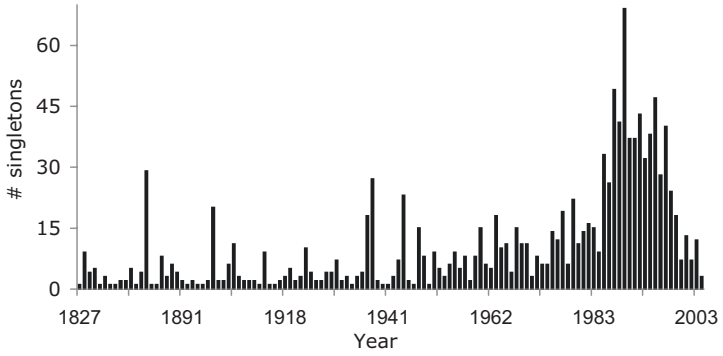

Fig. 5. The number of singletons collected per year in the Guianas. More singletons were collected from the 1980's.

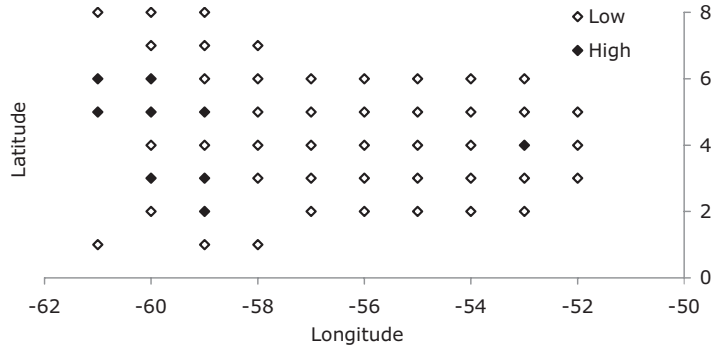

Fig. 6. The spatial distribution of singletons across the Guianas. The singletons were grouped into 1 degree grid cells. After fitting a logarithmic regression line through the graph of the number of singletons plotted against the number of specimens, we calculated the residuals from this line. Nine sites showed an unusually high number of singletons (black diamonds) for the number of specimens collected.

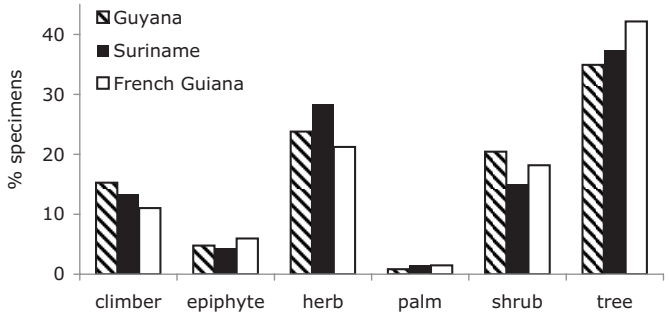

Fig. 7. The percentage of specimens represting the various growth forms forms for herbarium specimens collected in Guyana, Suriname and French Guiana. Trees were the most collected and palms were the least collected of the growth forms.

Variation in species composition among the Guianas

Not all the species in the database were collected in all three countries. The number of species collected was highest in Guyana and lowest in Suriname (Fig. 1 and 8). A total of 2,520 (or 35%) of the species were collected in all three countries. Guyana and Suriname show the highest similarity in collected species and French Guiana and Guyana the least. About 25% of the species in the database were unique to Guyana while a lower number of unique species (6 and 11% respectively) were collected in Suriname and French Guiana (Fig. 8). Unique species do not necessarily occur in low frequencies in the database as the number of specimens the unique species are between 1 and 72 (average 4).

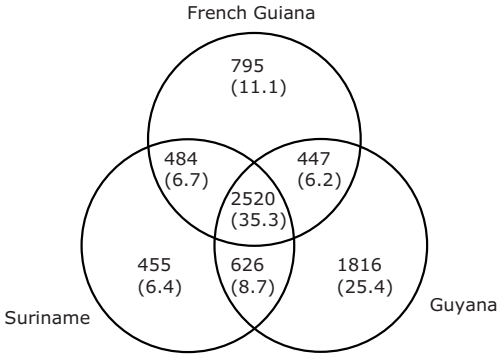

Fig. 8. Venn diagram showing the number of unique species and species shared between Guyana, Suriname and French Guiana. About 32 % of the species were collected in all three of the countries and the highest number of unique species was collected in Guyana.

Geographical distribution of specimens

To maximize the chances of finding new species, botanical expeditions tend to go to under-collected sections of the Guianas. As a result, the geographical area in which the specimens were collected gradually expanded from 1804 to 2004 (Fig. 3). Also some areas, especially those that are close to cities or research stations, were revisited. The result is that by the end of 2004 the number of specimens collected per 5 arc-minutes grid cell (c. 10 x 10 km) ranged between 0 and 7,131 while the number of species ranged between 0 and 1,693.

The number of specimens collected strongly determined the number of species found (Fig. 9). Our knowledge on the geographical distribution of species still remains incomplete as botanical expeditions did not go to all parts of the Guianas since so far only 1,504 of the 5,345 grid cells (or 28.1%) were sampled.

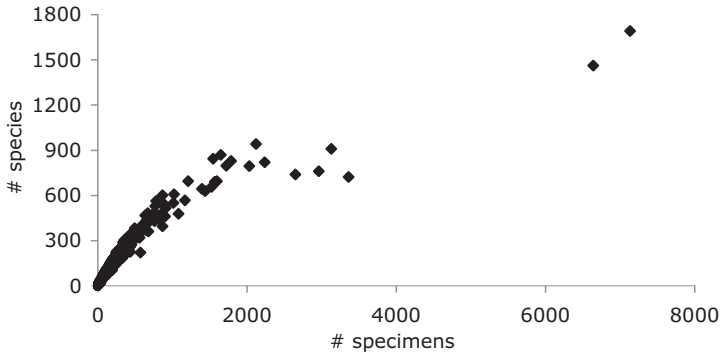

Fig. 9. The relationship between the number of specimens and the number of species per grid cell. The number of specimens collected strongly determined the number of species collected.

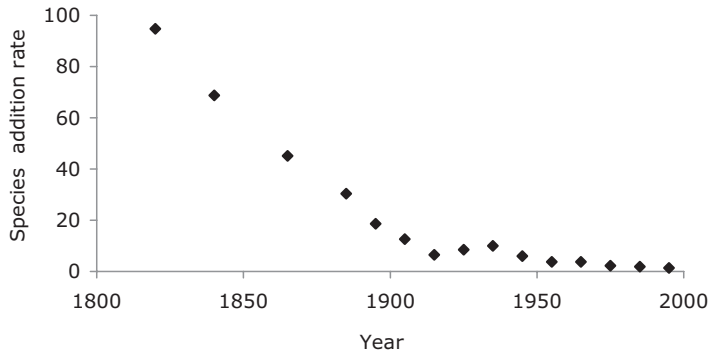

Fig. 10. The rate of addition of new species per collection to the herbarium.

Addition rate of new species is declining

The addition of new species to the historical curve was initially very rapid when new species were constantly being discovered (Fig. 2). However towards the end of the 20th century the rate of addition of new species slowed down considerably possibly because most of the species are occurring in the parts of the Guianas where the botanists had visited were already collected. By 1995 the rate of addition of new species was reduced to 1.4 for every 100 specimens collected (Fig. 10). Since most of the Guianas still remains under-collected, choosing new areas with very different habitats than those already sampled might increase the rate of addition of new species. To be cost effective, most botanical expeditions are planned for the dry season when most of the species are known to be

flowering or fruiting. Different strategies might be employed to find new species such as collecting at a different time of the year. This increases the chances of collecting species with a different phenology pattern. Still some species occur in low abundance in nature and some are less showy than others and these factors may contribute to making it difficult to find the remaining species.

Concluding remarks

Peaks in the historical species accumulation curves correspond mostly to botanists exploring new areas. The database shows a high variation in the distribution of species among genera, families and growth forms and this may be due in part by variation in collecting strategy and habitat. Our knowledge of species occurring in the Guianas has increased considerably due to the centuries of botanical collecting but this knowledge is expected to improve as most of the geographical areas still remain under-collected.

Chapter 3

Using a herbarium database to demonstrate collector bias in time and space

3

With Hans ter Steege, Jean-Jacques de Granville, Hervé Chevillotte and Michel Hof

Abstract

The digitizing of herbarium specimen label information has increased the value of herbarium databases for biodiversity assessments. However there is concern that the bias in collecting specimens might impact biodiversity model outputs based on herbarium databases. Yet ecologists continue to make use of herbarium databases for biodiversity models without addressing these impacts. Here we assess the extent of possible biases associated with the angiosperm herbarium database for specimens collected in Guyana, Suriname and French Guiana, and their consequence for estimating species richness, producing species distribution maps, and in taxonomic and phenological studies.

Due to historical bias in the herbarium database, standard methods for estimating species richness (e.g. Michaelis-Menten model) give unrealistic estimates of species richness. Here we propose a new method, which combines the Michaelis-Menten and the Arrhenius models and gives a more realistic estimate of species richness. Collecting effort is strongly biased towards rivers and roads. However, this bias might not affect species distribution models as the collection localities were not demonstrated to show environmental bias in the study area. Although most botanists show taxonomic bias in their collecting effort, this does not have serious consequences when the herbarium database is used for biodiversity studies. The herbarium database shows bias towards specimens collected in the dry season, and data on flowering (but not fruiting) period show strong resemblance to that collected in a field surveys. However, if the herbarium database is used to assess phenological patterns, then the bias towards collecting in the dry season must be addressed.

Key words: angiosperm herbarium database, the Guianas, species richness estimation, species distribution maps, phenology, taxonomy, bias, space, time.

Introduction

Over the past centuries, an estimated 2.5 billion biological specimens (mostly plants and animals) have been stored in herbaria and natural history museums worldwide (Graham *et al.* 2004). About 37 million of these (some for specialized geographic areas), collected over a two-century period, are stored in herbaria and natural history museums in the Netherlands (Nederlands Centrum voor Biodiversiteit, unpublished). Arguably, such specimen databases cannot be omitted when addressing large-scale biodiversity issues. This is because herbarium specimens represent, for example, all growth forms, show large geographical coverage (in remote areas herbarium specimens are often the sole data source) and high taxonomic resolution (Hoff 1996; Hijmans *et al.* 2000; Funk & Richardson 2002; but see Tobler 2007). Furthermore, over the last two decades specimen label information is becoming rapidly digitally available and this has increased the value of herbarium databases for biodiversity assessments.

In spite of the clear strengths, herbarium data also have inherent weaknesses. Herbarium specimens were initially collected for the purpose of documenting new species. One of the main weaknesses is that the ad-hoc nature of collecting strategies, or even of documenting the label information (Soberon *et al.* 2000; Reddy & Davalos 2003; Kadmon *et al.* 2004; Schulman *et al.* 2007; Tobler *et al.* 2007). This is because at the time of collection, it was generally not anticipated that herbarium data would be used for large-scale biodiversity studies or for providing a foundation for land management decisions. Specimens were generally collected just to describe new species, to establish the flora of given areas and to establish rich herbarium collections for the purpose of research and a large number of botanists contributed to the database, each with his/her own collecting objectives and methods. The ad-hoc strategy of collecting specimens from the plant community was used to maximise the number of species collected during expeditions (Chapter 6). This non-random method of collecting makes the application of statistical testing problematic.

In this paper we evaluate collection bias in the herbarium database of a well-collected area, the Guianas, of which most specimens have been identified

by specialists in the framework of the Flora of the Guianas project. We will describe how the large herbarium database was developed over time and test its potential for biodiversity research. Below we describe the assumptions that herbarium databases need to satisfy before they can be used for a number of the new applications, later we will test whether our database satisfies these assumptions and propose some solutions to deal with these biases. We'll specifically examine historical bias, geographical bias, taxonomical bias and seasonal bias as potential sources of error in herbarium databases.

Historical bias

Species accumulation curves are widely used for the estimation of species richness in an area (Colwell & Coddington 1994; Gotelli & Colwell 2001; Magurran 2004). The shape of a species accumulation curve depends on the order in which specimens are added to the curve – a different order will give a different shape and therefore usually a different richness estimate (Colwell & Coddington 1994). Herbarium databases show a historical bias in collecting effort and different incremental subsets of the database (e.g. specimens collected between 1804 and 1904 or between 1804 and 1954 for our Guianas database) will give different species abundance distributions. Several models have been used to describe the form of the accumulation curves (Raaijmakers 1985; Rosenzweig 1995; Flather 1996). One of the common models used is the Michaelis-Menten (MM) model (Raaijmakers 1985; Colwell & Coddington 1994; Flather 1996; Gotelli & Colwell 2001; Magurran 2004), of the form $S_n = S_{max} *N/(b + N)$, where S_n is the estimated species richness, S_{max} is the asymptotic number of species, b is the collecting effort necessary to collect 50% of S_{max} and N is the total number of specimens of the sample. The performance of the MM model depends on the species abundance distribution in the database, and in a herbarium database a large number of species are represented in low frequency. S_{max} estimates tend to increase with N. We argue that for the Michaelis-Menten model to give realistic outputs, different incremental subsets of the database should provide the same estimate of S_{max} as the full database and we test this argument with three different incremental subsets of the database and the full database.

Geographical bias

Large scale predictive species distribution models assume that collecting effort shows no environmental bias (Reddy & Davlos 2003; Kadmon 2004). Rainforests are generally difficult to access and collecting effort tends to be concentrated along rivers and roads. Knowledge of species ranges increases over time as collecting effort expands over wider geographical and environmental areas. It is well known that most specimens are collected within the first 5 km from the nearest river or road (Hijmans 2000; Reddy & Davlos 2003; Kadmon 2004). We therefore quantify to what extent collecting effort in the Guianas was restricted to the vicinity of rivers and roads and how this affects the geographical and environmental coverage of the whole study area.

Taxonomical bias

Botanical specialists are selective and may show preference for collecting specimens belonging to their target taxa. This introduces taxonomic bias. Furthermore, specialists will visit areas where their target taxa are more likely to occur. If a botanist shows a preference for certain taxa, it is expected that the taxon composition of the database of this botanist would be different from that of the herbarium database. With the Guianas database we examine the extent of this bias and discuss the implications.

Seasonal bias

Flowering and fruiting in the tropics show a wide range of patterns. Community patterns of flowering can range from one single peak in the dry season (ter Steege & Persaud 1991; Haugassen & Peres 2005) to several peaks per year or show no clear pattern. Community fruiting patterns show peaks correlating strongly with the wet season (Sabatier 1985, ter Steege & Persaud 1991; Zang & Wang 1995; Haugassen & Peres 2005). When seed/fruit traps are used in dispersal studies, it is assumed that all seeds and fruits have an equal chance of arriving in the seed traps (Sabatier 1985; Hubbell 2001). Transect studies, counting flowers, fruits and seeds, along a path in the forest make the same assumption that each species has an equal chance of being 'found' by the investigator (Sabatier 1985). Herbarium specimens also contain phenological information, as the collection date of a flowering or fruiting specimen is recorded. Consequently, herbarium specimens have been used to describe

phenological patterns in tropical rain forests (a.o. ter Steege & Persaud 1991). Like the standard methods, this method assumes an equal chance of species being found flowering or fruiting. However, this assumption is only met if specimen collecting is random with regard to the phenological pattern, thus random throughout the year, and to the stage in the reproductive cycle. This will rarely be the case, as plants without flowers or fruits will presumably less often be collected. Most of the Guianas have two dry and two wet seasons and flowering usually occurs in the dry seasons (ter Steege & Persaud 1991) and we question whether specimens have been collected randomly with regard to this pattern or if botanists have avoided collecting during the wet seasons, creating seasonal bias in the data.

Taxonomic data collected in the Guianas has been used for biodiversity assessment studies (ter Steege *et al.* 2000; Funk & Richardson 2002; Hoff *et al.* 2002; Lim *et al.* 2002; Clarke & Funk 2005; Funk *et al.* 2005). Although many of these studies have acknowledged that bias associated with the data might influence the results of their studies, none of these studies have quantified it. In this study we assess biases in the angiosperm herbarium database collected in the Guianas (Guyana, Suriname and French Guiana) when the database is used for species richness estimation, species distribution mapping, taxonomic and phenological studies. We will then address the consequences of the biases when the herbarium database is used for these applications and offer potential solutions.

Materials and methods

Data preparation

The angiosperm database of the Nationaal Herbarium Nederland, Utrecht branch, forms the backbone of this study (c. 115,000 specimens) and was augmented with databases from botanists and herbaria that are working on, or housing specimens collected in the Guianas. Many herbaria have contributed digital databases - Institut de Recherche pour le Developpement, IRD, Centre de Cayenne, Cayenne (c. 79,000 specimens); the New York Botanical Garden (www.nybg.org/bsci/res/resproj.html) and the Missouri Botanical Garden (www.mnh.si.edu/biodiversity/bdg/). The accession records of the Jenman Herbarium, Jonah Boyan Herbarium of Guyana and Lands Bosbeheer and Bosbeheer

Suriname of Suriname that were prepared during the colonial times were digitized by the staff of the Nationaal Herbarium Nederland. Finally, we included lists of taxa collected by different botanists from the Smithsonian Institution, Washington, D.C. (Hollowell *et al.* 2000; 2003; 2004). Duplicate records, i.e. of specimens with the same collection date, botanist and number, were removed. The species names, as shown on each label, were updated, based on the Smithsonian's 2005 Web Listing of Plants of the Guiana Shield and the Guianas (www.mnh.si.edu/biodiversity/bdg/planthtml/index.html) and the W3tropicos website (www.tropicos.org). The Angiosperm Phylogeny Group II system was used for the systematic classification of families and genera (www.mobot.org/MOBOT/research/APweb/).

Only specimens that were identified to the species level were used in the analysis and infra-specific information was not used (Table 1). Introduced species were removed from the database. We used the W3tropicos and the Smithsonian's 2005 Web Listing of Plants of the Guiana Shield and the Guianas websites to determine whether species occurring in the original database were introduced or naturally occurring. We chose to exclude introduced species from the analysis because most of these grow in populated areas and play only a minor role in natural ecosystems of the Guianas. Information on latitude and longitude was copied from the labels when available. National gazetteers were used to fill the gaps when information on latitude and longitude was lacking but a locality name was present on the label, or were traced from field notes. Specimens that lacked information on date of collection, specimen identity and/or the collection locality were not used for the analysis (Table 1).

Historical bias

To address the extent of the historical bias we compared the observed species accumulation curve with a rarefaction curve. The rarefaction curve uses a re-sampling procedure to calculate the statistical expectation of the species accumulation curve and the variance of species richness after 1,000 randomizations of the database. We chose the rarefaction method to calculate the expected curve because it takes into account rare and common species in the database and unequal specimen accumulation (Koellner *et al.* 2004). Through interpolation, the rarefaction method makes it possible to compare the observed and expected species richness for a given number of specimens. To

form the species accumulation curve, the specimens were first ordered according to their collection year. The cumulative number of species discovered was then plotted against cumulative number of specimens collected (effort). We calculated the rarefaction curve and the 95% confidence intervals.

Estimating species richness

We estimated the total hypothetical species richness of the Guianas using incremental subsets of the database collected during ten cumulative time periods with 20 year increments beginning in 1804 and ending in 2004. We used two methods to estimate the species richness. For the first method we chose the MM model to extrapolate the species accumulation curve to an asymptote (Colwell & Coddington 1994; Gotelli & Entsminger 2001). To get a mean species accumulation curve for each time period, we calculated the average diversity for different abundance levels after 1000 Monte Carlo randomizations using the software EcoSim (Gotelli & Entsminger 2000). We used the quasi-Newton non-linear regression approach in Statistica (Statsoft 1994) to fit the MM model (Keating and Quinn 1998).

For a second method of estimating species richness we assumed that species richness estimates for incremental subsets of the database increases with the total area visited as we suspected that S_{max}, as estimated using the MM model, increased with the increasing area visited during incremental 20 year periods. The oldest law of biodiversity (Rosenzweig 1999) states that the number of species in an area is a function of the size of that area, $S=cAz$ where S is the total number of species in an area A and c and z are constants. Here we assume that S_{max} would be a similar function of the area actually visited ($A_{visited}$). Thus $S_{max}=cA_{visited}^{z}$. We calculated the area visited during each time period by creating 5 km buffer zones around the locality of each specimen and then summing up the total area of the buffers. We chose the 5 km zones because botanists tend to cover about this distance per day on a collecting expedition and also because most specimens were collected within 5 km of access roads and rivers (see below). We then used a power regression to estimate S_{max} through extrapolation for the total land mass of the Guianas.

Geographical bias

To examine the magnitude of geographical bias towards more accessible areas, we compared the distance between original collection localities and randomly generated localities, respectively, and the nearest river or road using the method described by Reddy & Davalos (2003). The coordinates of the rivers and roads were downloaded from biogeo.berkeley.edu/bgm/gdata.php. We generated the same number of random point localities as in the database (n = 7,253) using the 'Generate randomly-distributed points' script in ArcView (ESRI 1999; Lead 2005). The distance between the original and random localities to the nearest river or road was calculated using the Geo-processing wizard in ArcView. We then performed the Mann-Whitney test to determine whether there were significant differences between the distances from the original and random localities to the nearest river or road.

Roadside bias would only have serious consequences if the observed collecting localities do not properly represent the ecological conditions of the Guianas. We therefore examined whether the climatic and altitudinal conditions of the localities collected (the observed localities) were different from those existing elsewhere in the Guianas (the expected localities). We used the altitude and 19 bioclimatic variables for the current conditions with a 5 arc-minutes resolution from the WORLDCLIM database (www.worldclim.org; Hijmans *et al.* 2005). To avoid inter-correlation among or between the variables and to avoid redundancy, we performed a principal components analysis (PCA) on the variables, we retained the first four component scores which jointly explained about 93% of the overall variation in the ecological conditions across the Guianas. We divided each of the four PCA component scores obtained from each of the two datasets (i.e. based on the observed and expected localities) into 10 equal interval groups (bins) based on the range each component score. The difference between the two datasets was tested using the Kolmogorov-Smirnov test (Kadmon *et al.* 2004). If the datasets are significantly different then the climatic and altitudinal conditions of the collecting localities are different from those existing elsewhere in the Guianas, indicating that the collecting localities do not properly represent the ecological conditions of the Guianas.

Taxonomical bias

To examine botanists' bias towards certain growth forms or plant families we compared the composition of the collecting lists of botanists to the herbarium dataset using a multinomial probability distribution. For instance, to test whether a given botanist was biased, we compared the number of specimens per species of this botanist $\{n_1,...,n_S\}$, where ni is the number of specimens for taxon (growth form or family) *i*, to the full list of specimens in the herbarium $\{N_1,...,N_S\}$. We were asking whether the botanist provided a biased or an unbiased sample of the full dataset, and this for each of the taxa. This probability is given by the multinomial distribution P(n1,...,nS | p1,...,pS), where $p\alpha_i = N\alpha_i/sum\alpha_i$ Ni. A result was interpreted as significantly different (over-collected or under-collected) if the actual abundance of taxon or growth form *i* in the collection was outside the central 95% of simulated abundances.

Seasonal bias

To examine seasonal bias in the herbarium database, we grouped the specimens according on their month of collection. We compared this with the mean monthly rainfall data derived from the WORLDCLIM database. The mean monthly rainfall was defined as an average of the rainfall for all the localities where the specimens were collected. We chose to use the mean monthly rainfall for the collection localities for the whole of the Guianas because we found that the rainfall patterns were quite similar among the countries. Pearson correlation was used to examine the relationship between the monthly number of specimens and the number of field days per month (based on the information on the specimen labels) and the total monthly rainfall, assuming that the relationships tested are linear. The relationship between the number of species and specimens collected per month was also tested through Pearson correlation, assuming that the relationship is linear. We then examined the correlation between the flowering and fruiting patterns per month of the herbarium database and data from autecological flowering and fruiting records for 190 species collected in Guyana over a century period (ter Steege & Persaud 1991). The trees involved in the autecological studies were marked and information on their flowering and fruiting was collected on a standard year-round basis.

Results and discussion

The historical accumulation of species and specimens in the herbarium

Our working database comprised a total of 190,398 specimens, collected by 560 botanists. Only 168,487 of the specimens containing complete information, represented species considered to be naturally occurring in the Guianas, and were subsequently used during all the analyses except for the analyses involving seasonal bias (Table 1). The specimens belong to 183 families, 1,525 genera and 7,148 species.

Collection effort was not evenly distributed over time and space in the Guianas (Fig. 1 and Chapter 2). The species accumulation curve was irregular, showing a series of temporary plateaus, especially during the first century of collecting (Fig. 1). These temporary plateaus occur when there was slow accumulation of specimens due to a lack of collecting effort or because effort being concentrated in a limited area so that after continued collecting new species became increasingly more difficult to find. The exploring of new areas was associated with an increase in the number of species until most of the species in that new area had been collected, leading to a new temporary plateau.

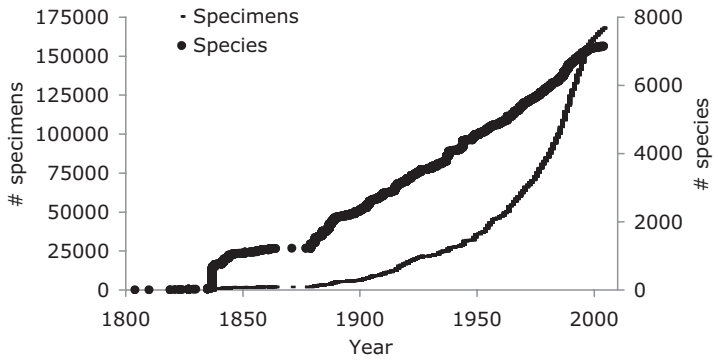

Fig 1. The effects of the results of the historical process of compiling specimens and species for the Guianas from the period 1804 to 2004.

The species accumulation curve did not mirror closely the rarefaction curve (Fig. 2). The slope of the species accumulation curve was initially much lower than that of the rarefaction curve, with estimates of species richness derived

from the rarefied curve being higher than the observed species richness. This is because the species accumulation curve represents a single ordering of specimens by their collection year while the rarefaction curve is based on random sampling of all of the species collected, from an initially (at low specimen numbers) much larger biogeographical range.

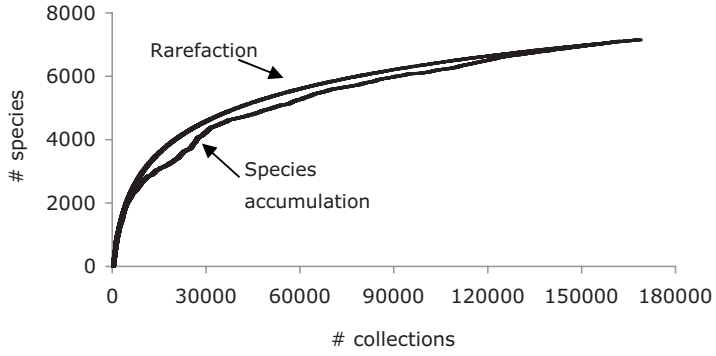

Fig. 2. Species collection and rarefaction curves for herbarium specimens with sufficient label information. The solid line shows the species accumulation curve based on specimens ordered in the way they were successively collected.

S_{max} and hence the estimate of species richness increased with incremental subsets of the database (Table 2). The increase in S_{max} can be explained by the increasing geographical area covered by botanists (Chapter 2). At the end of 2004 S_{max} was estimated to be 7,507 while the number of observed species was 7,143 (Fig. 3). We expect the MM model to substantially underestimate species richness for the Guianas because much of the area still remains under-collected or not collected at all (Fig. 4). When S_{max} was modeled as a function of the area visited and species richness was estimated through extrapolation of the Arrhenius model to the whole land mass of the Guianas, we estimated that a total of 11,266 species would be present (Fig. 3). This is about 37% more species than known today.

The species accumulation curve did not reach an asymptote although the curve clearly indicates that the efficiency of discovering new species decreased steadily with the growth of the herbarium (Fig. 2). Based on specimen data collected after 1995, from Fig. 1 we estimated that about 1.4 species are added for every 100 specimens collected (Chapter 2). It is possible that many of the species that are still to be discovered are rare. How can we collect those rare species or can

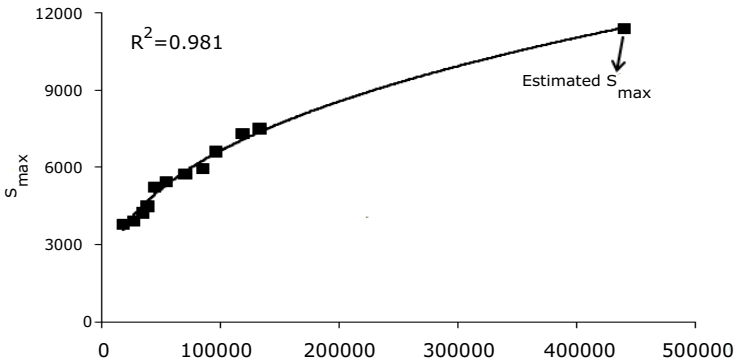

Fig. 3. The Arrhenius relationship between Smax and the area visited and the estimated Smax for the total area visited. The estimated Smax is calculated through extrapolation based on power regression.

we collect them at all? It is a well known fact that *'common species are rare and rare species are common'* (Forysth and Miyata, 1984). This is illustrated by the fact that rank distributions mostly follow a near log-series (Fisher *et al.* 1943; Hubbell 2001). At large scales tree communities appear to follow this rule very well (Hubbell *et al.* 2008). This suggests that many of the last species to be found are expected to be in the tail of the log-series. They are rare to very rare and the chance of collecting them is very small indeed. Some gain may be made if areas are visited which have an ecology that is different from the (main) areas that have been collected thus far. However, just collecting under-collected areas of tropical rain forests may not result in collection of all species. We need to accept, that a number of such species will never be found by further inventories and botanical explorations – they are simply too rare.

Distribution of collections; data is strongly geographically biased

The observed frequencies of collecting localities near rivers and roads were greater than expected from randomly generated localities (Fig. 4 and 5). Geographical collecting bias was strongest for distances of 0-5 km from the nearest river, accounting for about 66% of the observed localities while less (about 39%) of the randomly generated localities were in this zone. The number of species collected was strongly correlated with the number of specimens collected and strongly decreased with increasing distance from the nearest river or road (Fig. 6). In fact about 62% of all the specimens and about 87% of all the species in the Guianas were collected within 0-5 km of easy access points.

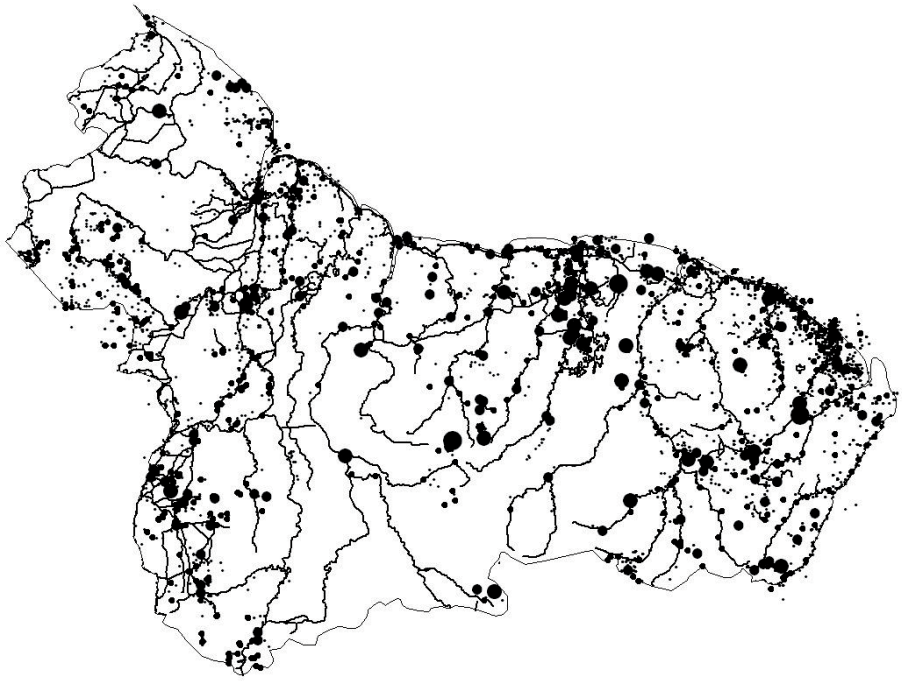

Fig. 4. Spatial distribution of road and river networks (black lines) and botanical collection localities (black dots) in the Guianas. Most of the collection localities are centered along rivers and roads.

The results confirm our expectations that localities of the specimens made in the Guianas occur mainly along easy access routes and that the number of specimens and species collected declines with increasing distance from

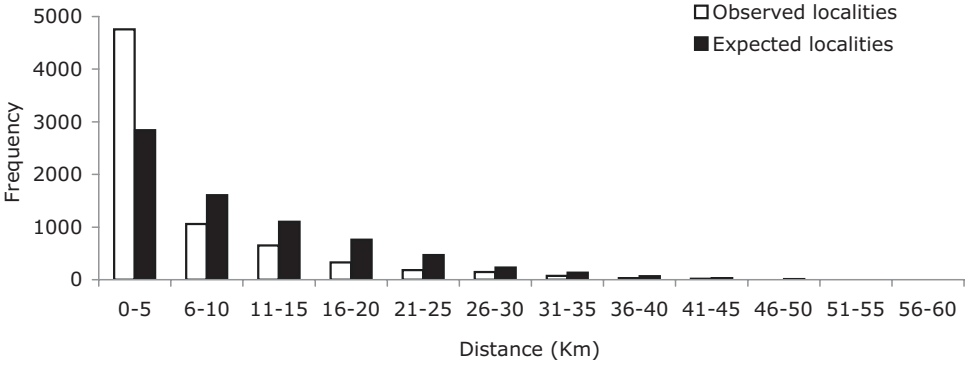

Fig. 5. The distribution of the number of observed and randomly generated localities at different distances from the nearest river or road. The distance between the observed and randomly generated localities from the nearest river or road is significantly different ($p < 0.0001$).

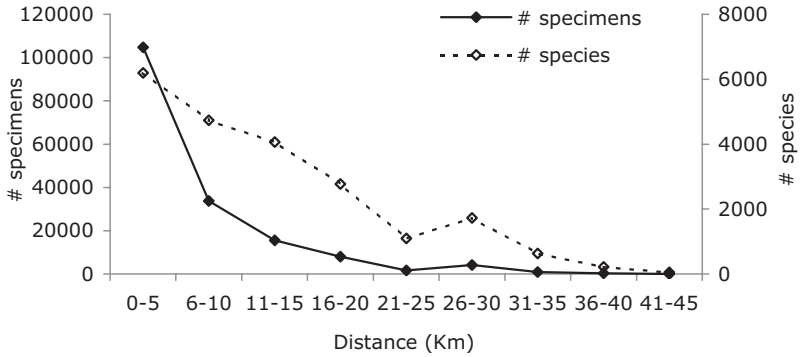

Fig. 6. Number of specimens (whole line) and species (broken line) collected at varying distances from the nearest river or road. The number of specimens and species decreased as the distance between the nearest river or road increased.

these routes. Botanists hardly ever collect the same species twice during the same expedition. A botanist uses rivers and roads to get to the destination of the expeditions and starts collecting upon arrival. Collecting activities for the expedition usually starts closest to the nearest river or road. At the beginning all species are new for the expedition and are therefore collected once they are flowering or fruiting. However as the expedition continues areas deeper into the forest are searched for new species but once the forest remains the same, it becomes more and more difficult to find species that were not already collected. This explains why most of the specimens were collected within 0-5 km from the nearest river or road. The low number of species discovered with increased distance from the rivers and roads is in fact the result of poor collecting effort with increasing distance from the nearest river or road.

The consequence of geographical bias is that knowledge of species distribution decreases with increasing distance from rivers and roads. Species distribution maps derived from herbarium data may reveal the distribution of collecting effort of botanists rather than actual distribution patterns, and this is highly scale dependent.

Predictive modeling techniques have been used to deal with gaps in collecting data by establishing a relationship between the species data and environmental (mainly climatic and altitudinal) data. However, the roadside bias will seriously impact the model predictions if the climatic conditions of the localities where specimens were collected were different from those existing elsewhere in the Guianas. This was not the case in the Guianas (Fig. 7). For all four of the

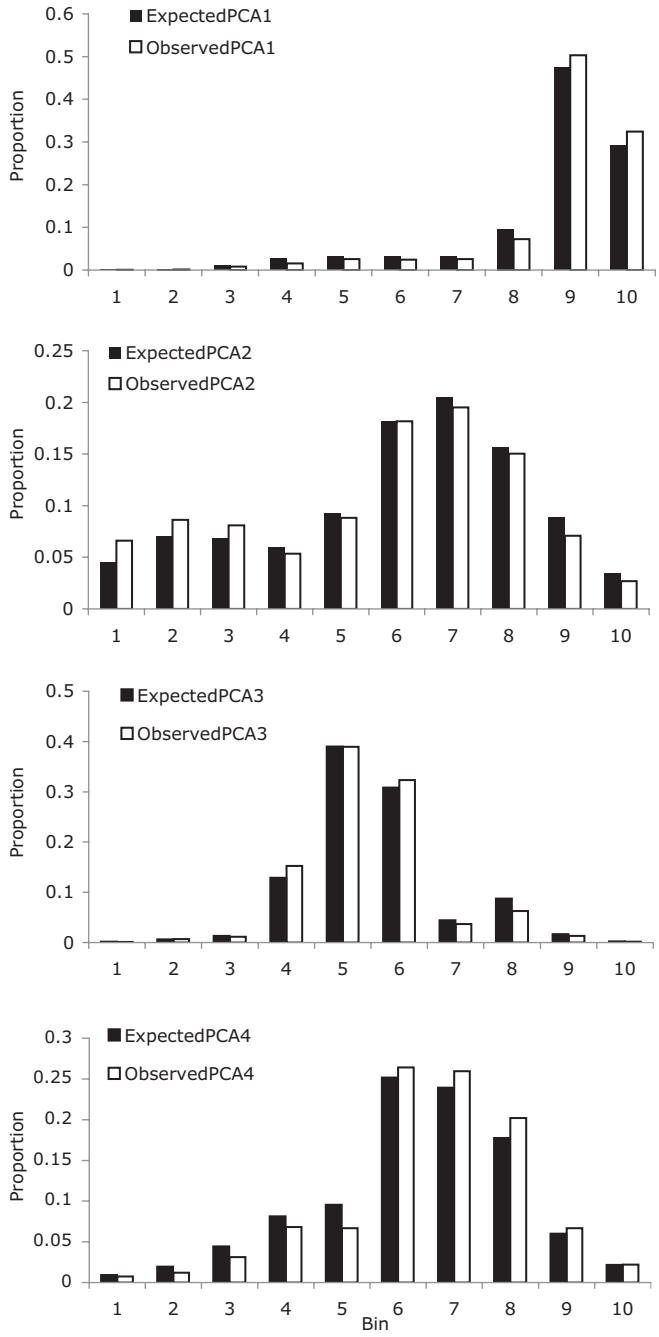

Fig. 7. Comparison of the observed distributions of four PCA variables with the countrywide distributions. For all four PCA variables the difference between the two data sets was not significant when the Kolmogorov-Smirnov test was applied (D=0.4, P>0.3).

PCA variables there were no significant differences between the two datasets suggesting that the altitudinal and climatic conditions of the localities where specimens were collected were not different from those existing elsewhere in the Guianas. Our results suggest that although the herbarium data is biased towards rivers and roads it is not biased with respect to climate and altitude and therefore the accuracy of models using altitude and climatic data in the Guianas may not be compromised. This is probably because the river (especially) and road networks are well distributed across geographical climate area.

Taxonomic and growth form bias

Botanists exhibited a bias towards the families as they collected proportionally more of some than of other families present in the herbarium (Appendix 3.1). Furthermore, all botanists showed a preference for collecting one or more growth forms (Table 3). The consequence of the bias is that the species list of the target taxa becomes larger and the range size of the individual species may be more predictable. A better knowledge of species responses to the environment leads to more stable models when species data is used in applications such as ecological niche modeling (Lim *et al.* 2002; Hortal *et al.* 2007).

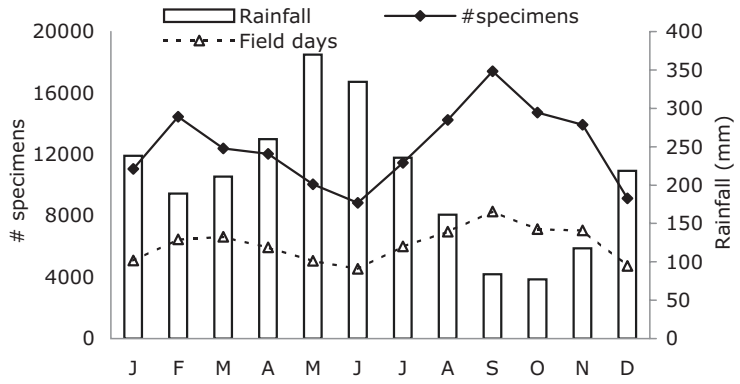

Fig. 8. Rainfall, the number of days the botanists spent in the field and the number of specimens collected per month. During the drier months botanists spent more days in the field and more specimens were collected. The Pearson correlation between the number of specimens collected and the number of field days per month and the total monthly rainfall are negative ($P<0.0001$).

Phenology: collecting is strongly seasonally biased

The number of field days and the number of specimens collected were not uniformly distributed throughout the year. The number of days spent in the field and the number of specimens collected peaked in September, which corresponds to the first month of the long dry season. The lowest number of days in the field and specimens collected were in June corresponding to the rainy season (Fig. 8). The monthly number of specimens and the number of field days per month were negatively correlated with the total monthly rainfall ($P < 0.001$, Fig. 8). The number of species collected per month was determined strongly and positively by the number of specimens collected (Fig. 9). These results therefore agree with our expectation that a higher number of specimens and thus species are collected in the dry seasons than in the wet seasons (Fig.9).

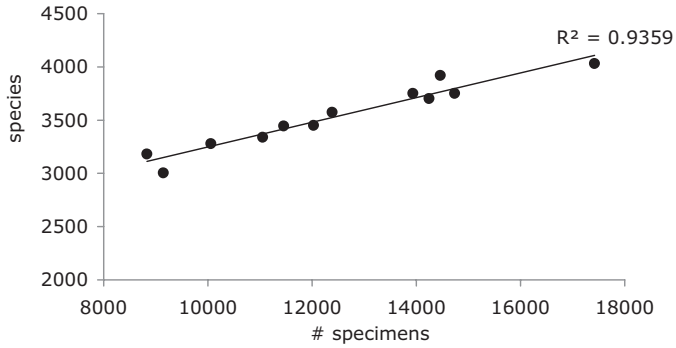

Fig. 9. The relationship between the number of specimens and species collected per month. The number of species collected was significantly correlated with the number of specimens collected per month.

Many field studies have shown that phenology patterns are closely correlated with rainfall (e.g. ter Steege & Persaud 1991). However these studies are based on standard year-round observations of the same trees. When phenology data from the herbarium database and the autecological records were compared we found a strong correlation ($R^2=0.756$) with the flowering data but not the fruiting data (Fig. 10). Although the herbarium database showed a bias toward collecting in drier months, it still reflects the actual flowering patterns, but not the fruiting patterns. This is probably because a sufficient number of flowering specimens was collected during the dry season to give the strong correlation with the flowering trees of autecological data. The results show that bias in

collecting effort in the dry season can lead to significant associations in the herbarium database between flowering and rainfall data. Since the number of specimens collected determined the number of species discovered, a high collecting effort in the dry season would lead to a large number of species being collected in the dry season. The opposite is probably true for the fruiting data. Therefore if herbarium database is used to describe phenology patterns, bias in collecting effort in relation to dry and wet season must be corrected for.

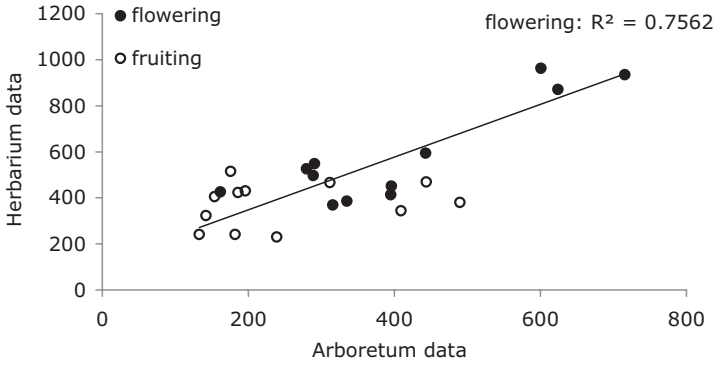

Fig. 10. Correlation between data collected from autecological records from Guyana (ter Steege & Persaud 1991) and herbarium data. Strong correlation can be seen with the flowering data but not with the fruiting data. The correlation with the fruiting data was not significant.

Concluding remarks

We have shown with the herbarium dataset of the Guianas that the number of species collected is strongly determined by the number of specimens. The dataset was biased in collector effort historically, geographically, taxonomically and seasonally. Most of the biases are caused by unequal collecting effort in space and time. Ecologists have to be aware of the biases.

Table 1. The level of completion of label information of herbarium specimens collected in Guyana (GU), Suriname (SU) and French Guiana (FG) during the period 1803 to 2004.

Status	FG	SU	GU	Total
Total original specimens	75,691	56,915	57,792	190,398
Species identity	74,958	52,748	55,486	183,192
Species identity & geographical information	74,378	49,032	52,146	175,556
Species identity & temporal information	74,575	50,394	50,987	175,956
Species identity & temporal & geographical information	74,351	48,284	49,666	172,301
Introduced species	1,368	177	2,269	3,814
Specimens used in this paper	72,983	48,107	47,397	168,487

Table 2. The estimated number of species for cumulative time periods. Only the final six of the ten 20-year cumulative time periods are shown here. N and S are the number of specimens and species respectively. The S_{max} and b were estimated using the Michaelis-Menten $S_n = S_{max} *N/(b + N)$.

Cumulative time period	Observed S	Observed N	S_{max}	b
To 1904	2595	8736	3805	4232
To 1924	3428	20745	4243	5503
To 1944	4379	31327	5249	7157
To 1964	5094	54708	5738	8712
To 1984	6123	100433	6614	11462
To 2004	7143	168456	7507	14972

Table 3. The multinomial probability results indicating that botanists showed a preference for collecting the one or more life forms relative to other life forms when the botanists' species lists were compared to the total of the herbarium. A value greater that 994 indicates a bias towards a particular life form.

Main botanist	climber	epiphyte	herb	palm	shrub	tree
Clarke, H.D.	976	999	1	1	999	997
Cremers, G.	100	999	999	807	997	1
Donselaar, J. van	1	979	999	1	1	999
Granville, J.J. de	1	999	999	999	999	1
Jansen-Jacobs, M.J.	999	6	999	994	864	1
Jenman, G.S.	999	948	999	43	3	1
Lanjouw, J.; Lindeman, J.C.	235	214	999	1	1	1
Lindeman, J.C.	1	1	1	2	1	999
Maguire, B.	844	983	999	689	998	1
Mori, S.A.	94	1	1	1	1	999
Oldeman, R.A.A.	999	8	1	1	999	999
Pipoly, J.J.	87	1	1	1	999	1
Prévost, M.F.	999	1	1	1	4	999
Sabatier, D.	1	1	1	1	1	999

Chapter 4

Never the same species twice: a model of botanical collector's behaviour in the field

With Hans ter Steege, Olaf S. Bánki, Feike Schieving

4

Abstract

Because of their sheer numbers natural history museum specimens cannot be ignored when answering one of the fundamental questions in science 'what determines species diversity?'. The non-random nature of collecting specimens does not allow most statistical tests to be applied to herbarium data, however. Here we present a simple simulation model, which allows for any natural species abundance distribution the generation of the abundance distribution in a herbarium, following sample collecting strategies. We show that, in essence, the strategy of "never collect the same species twice" is enough to generate the collection structure as found in a herbarium. We illustrate this using real plot and specimen data from two well collected areas, one in central Guyana, one in Suriname.

Introduction

The question 'what determines species diversity' is still among the major questions in biological science (Pennisi 2005). Because of their sheer numbers (2.5 billion specimens are stored in natural history museums worldwide (Graham *et al.* 2004)), natural history museum collections cannot be ignored when answering this question. The original goal of collecting these specimens was to describe the wealth of diversity of plants and animals in nature, and to produce floras and monographs of families or genera. Thus collecting strategy should meet thus these objectives - to collect as many new species as possible. So, as time and the number of collections that can be made on a collecting trip are limited, collectors strive to never collect the same species twice and the phrase from the leading botanist "oh that one we collected already" must ring very familiar in the ears of participants of such trips – it does in ours. To increase the overall number of species per trip even more, collectors will tend to move to another area when the time to find new species becomes too long.

The consequence of this strategy is that, in principle, a large number of species is collected per expedition (Appendix 4.1). The herbaria today are the result of many short and long expeditions. Because the expeditions did not always have the full knowledge of what was already present in the herbarium, most species are represented by more than one specimen. However, as all collectors use the same search strategy, the herbarium today is characterised by an overrepresentation of rare species and an underrepresentation of common species, compared to abundance distributions in the field. As an example, in central Guyana the five most common tree species in the field *(Mora gonggrijpii, Eperua falcata, Chlorocardium rodiei, Dicymbe altsonii and Swartzia leiocalycina* (for nomenclature see Boggan *et al.* 1997)) make up 43% of all individuals over 30 cm dbh in the field (ter Steege *et al.* 2000) but account for only 6% of all herbarium specimens made of trees in that area (Ek & ter Steege 1998). In fact, the focus on rare species has lead to a staggering amount of species with only one collection (singletons) in herbaria, far more than any model of relative distribution (e.g. log-normal, log series) predicts (Chapter 2), thus providing a significant overestimation of the plant diversity of the collected region.

Early work of Fisher (Fisher *et al.* 1943) and Preston (Preston 1962a, b) suggested that fundamental mathematical distributions underlie the structure of communities in the field. Hubbell (Hubbell 2001) unified these distributions in the so-called zero sum multinomial (ZSM), which resembles the lognormal in local communities and the log-series in metacommunities of large areas. Data suggest that the log-series distribution indeed fits the community structure of Amazon trees very well (Hubbell *et al.* 2008). The ZSM has two free parameters θ, which is asymptotically equal to Fisher's α (one of the parameters of the log-series) and m, which can be associated with the input of new species (in the local community: immigration rate from the metacommunity; in the metacommunity: speciation). In the log-series Fisher's α is almost equal to the number of species with one individual (singletons), which is actually the first term of the log-series. The ZSM becomes similar to the log-series when m approaches one. In that case α is also equal to the number of singletons.
As the collecting strategy used by collectors is far from random sampling, statistical testing of the herbarium data is problematic and this has hampered the estimation of diversity. Here we present a model that explains how the

distribution of numbers of individuals of species present in herbaria develops in a non-random but predictable fashion from the log-series in the field.

Material and methods

The model consists of two parts 1) the construction of the relative abundance distribution in the field and 2) the sampling of that distribution by the collectors. The data for part 1 consisted of plot data (1-ha plots; full inventories of all trees with dbh > 10 cm) from two areas: Mabura Hill, central Guyana (hereafter Mabura) and the bauxite mountain region of North Eastern Suriname (hereafter Bauxite). For both areas we constructed a hypothetical species abundance distribution as follows (see Appendix 4.2) for a graphical lay-out of this part of the model). We constructed the relative abundance distribution from the plot data (RADplot) and calculated α and m. Based on the average number of trees in the 1-ha plots and the total area encompassing all specimens collected from that area, we calculated the total number of trees > 10 cm dbh in the area (Jm, the meta-population) and constructed the abundance distribution of all species and individuals of the total area (ZSM_{area}). Both calculations were made with Matlab scripts provided by Brian McGill (McGill *et al.* 2006).
In part 2 of the model we sampled individual collections from the ZSMarea (simulating botanical collecting) using Matlab scripts, written for this purpose. We simulated 'expedition type collectors', collectors that carry out large collecting trip and collect many specimens (100 – 400 specimens typically) and modeled three scenarios:

1) never collect the same species twice, keep searching for new species, regardless of the time (number of search loops) it takes;
2) as (1) but when this took more than a pre-set number of search loops in the programme, collect whatever is the next individual (some species are now collected more than once). This simulates stress or anxiety. Although the preference in real expeditions is one individual per species, there is usually also a target number of specimens per expedition, that needs to be met, relaxing rule 1;
3) as (2) but sample an equal number of specimens from 4 different sub-areas, which have no species in common (ZSM's based on α of the individuals but with similar α and m). This simulates the sampling of α-diversity (the behaviour of moving to a new terrain when the time

required to find new species gets too long).
The number of specimens that were collected per expedition in an area also resembles a ZSM (Appendix 4.2), as do most things in life (Nekola & Brown 2007). We made a relative abundance distribution of specimens collected per expedition and combined this RAD with scenario 1.

4) all collectors were allowed to collect from the ZSM_{area} the number of specimens they had collected for their expedition in real life and never collect the same species twice.

Finally we tested our model with varying number of collectors and varying number of collections to understand better the effect of collecting intensity on the structure of our herbarium, using scenario 1 as the collecting scenario. For the varying number of collectors we used the Bauxite ZSM_{area} and collected 3,000 specimens. The number of collectors varied from 1, who collected 3,000 specimens ($\alpha_{collectors} = 0$) and thus collected 3,000 species based on scenario 1. At the opposite were 3,000 collectors, each collecting 1 individual ($\alpha_{collectors} = \infty$). The latter is in fact a random sampling of the ZSM, in theory producing a smaller copy of it within the same α (c. 145 for Bauxite). $\alpha_{collectors}$ used were 400, 200, 100, 50, 10 (close to the actual value), 5, 2, 1. For the number of varying collections we sampled the ZSM_{area} of Bauxite with a $\alpha_{collectors}$ of 10 (which is similar to the empirical one of this $_{area}$, N = 2998, 57 collectors). We reduced the number of collections made in the $_{area}$, while keeping $\alpha_{collectors}$ constant by removing the collector with the most collections (629) for a total of 2,369 collections (56 collectors), and repeated this to obtain a number of collections of 1,115 (51 collectors). To obtain a higher number we added 1 collector with 1,300 collections to the number of collectors of Bauxite (N = 4298, 58 collectors). We ran the model with only scenario 1. We finally constructed species accumulation curves with Ecosim (Gotelli & Entsminger 2001), using 1000 randomisations.

Results

To characterize the structure of the ecosystem, in Mabura six 1-ha plots were inventoried for a total of 3,086 trees (> 10 cm dbh) and 112 species (Bánki & ter Steege unpublished data). M_{plots} was 0.934, which suggests a near log-series distribution, consistent with a α almost equal to Fisher's α and the number of

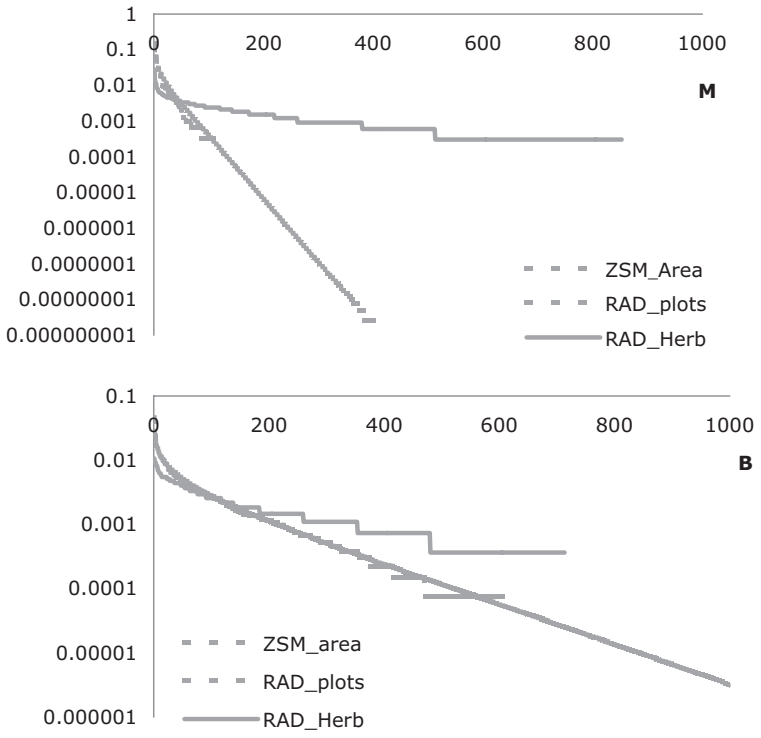

Figure 1. There is great difference in the relative abundance distributions of plot data and herbarium data. Relative abundance distribution of plots (RAD_{plots}) in Mabura (M) and Bauxite (B). The modelled relative abundance distribution (ZS_{Marea}) for each area follows the RADplots quite well. The relative abundance distribution of the herbarium (RAD_{herb}) is much flatter with less dominance and much more rare species (large tail), especially in Mabura Hill (A). X-axis species rank in abundance, Y-axis relative abundance (100*Nspeciesαi/Ntot).

singletons (Table 1). The relative abundance distribution of the plots (RAD_{plots}) and the calculated Zero Sum Multinomial (ZSM_{Mabura}) were also quite similar (Fig. 1A). In Bauxite 23 plots were inventoried for a total of 13,241 trees (>10 cm dbh) and 605 species (Bánki & ter Steege unpublished data). M_{plots} was 0.954, which also suggests a near log-series distribution, again consistent with a α almost equal to Fisher's α and the number of singletons (Table 1). The RAD_{plots} and $ZSM_{Bauxite}$ were also very similar here (Fig. 1B).

To characterize the herbarium, in Mabura a total of 3,302 botanical specimens were collected by 47 collectors (period 1846 - 2004), including 853 species (Appendix 4.2). The largest collection for one collector was 690 specimens and there were six collectors with only one specimen. In Bauxite a total of 2,727

specimens were collected by 46 collectors (period 1841 - 2003), including 713 species. The largest collection was 837 specimens (but in this case probably collected on numerous occasions: collector BW, Forest Dept Suriname) and there were also six collectors with only one specimen (see Appendix 4.2). Roughly half of the collectors collected one specimen per species, adhering strictly to scenario 1. The number of specimens of these collectors varied from 1 to 118 specimens. If the number of specimens per collector increased (often species were collected more than once) however, for 75% of the collectors in Mabura and 84% of the collectors in Bauxite the S/N ratio was larger than 0.75. The same species were more often collected by those collectors who visited the area on numerous occasions or among the specimens of the Forest Departments of Guyana (FD) and Suriname (BW), in fact collaboration of various collectors and forest scientists. The lowest S/N ratio was 0.41 in Mabura (John Pipoly, N = 199, S = 83) and 0.35 in Bauxite (BW: N = 837, S = 299).

The $RAD_{herbarium}$ for each of the areas was much flatter than the RAD_{plots} and ZSM_{plots} (Fig. 1A and B) consistent with the expected under-collecting of common species (less dominance) and over-collecting of rare species (long tail). Complete random collecting from the metacommunity in the field (ZSM_{area}) would in principle have led to a $RAD_{herbarium}$ very comparable to RAD_{plots} and ZSM_{area}. The ten most common species in the area (Eperua falcata, E. grandiflora, Catostemma fragrans, Licania buxifolia, Dicymbe altsonii, Oxandra asbeckii, Talisia squarrosa, Eschweilera sagotiana and Chlorocardium rodiei) amounted to 69% of all individuals in the plots but only 4% of the number of herbarium specimens. Similarly, the ten most common species in Bauxite (Lecythis corrugata, Eperua falcata, Micrandra brownsbergensis, Eschweilera sp., Elvasia elvasioides, Croton argyrophylloides, Qualea rosea, Astrocaryum sciophilum, Quararibea duckei and Bocoa prouacensis) accounted for 22% of the individuals of the plots but only 2% of the herbarium specimens. α calculated for the herbarium specimens for Mabura was 337, an extremely high number, compared to the α of the plots (Table 1). The same is true for Bauxite, where α of the herbarium specimens was 584.

So far we characterized the original herbarium structure and the structure of the ecosystem from which this collection was obtained. Simulated herbarium structures based on the sampling scenarios introduced above show that in the case of a relatively few large scale expeditions with the rule 'never the same

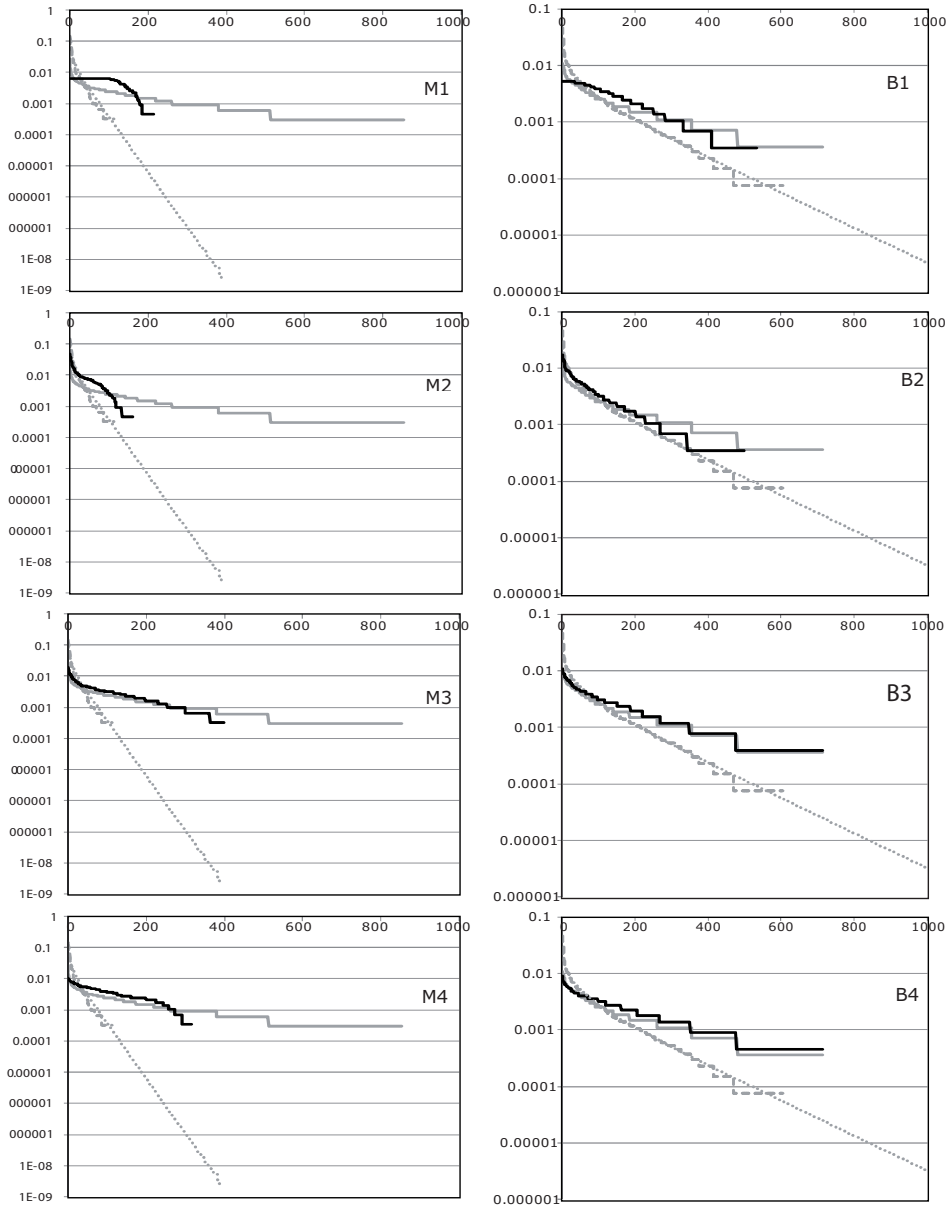

Figure 2. Output of simulations based on scenario 1 to 4 (top to bottom) in Mabura (M) and Bauxite (B). Light grey curves as in Fig. 1, black simulation result. Axis legends as in Fig. 1.

species twice' (Scenario 1: Mabura 15 collectors, 150 specimens, Fig. 2.M1; Bauxite: 15 collectors, 200 specimens, Fig. 2.B1), the relative abundance distributions of the herbaria are rather flat. This was especially clear in Mabura

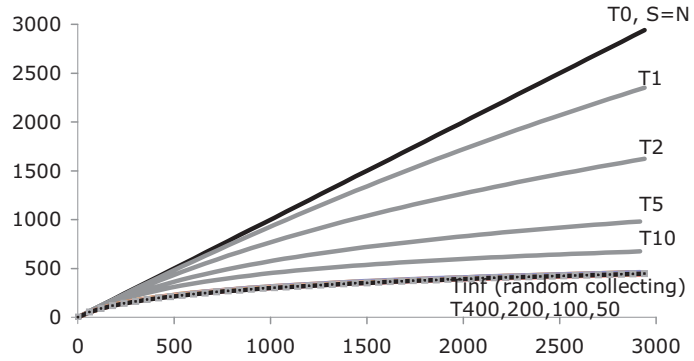

Figure 3. Species accumulation curves based on sampling the ZSM_{area} of Bauxite with collectors with differing relative abundance distributions of the collections per collector. Legend: T0; $\alpha_{collectors}$ = 0, 1 collector collects all specimens, hence S=N, Tinf; $\alpha_{collectors}$ = infinite, 3000 collectors all collecting 1 random specimen. All other $\alpha_{collectors}$ fall within this range (as should their curves); The ones with $\alpha_{collectors}$ = 50 – 400 are very close to $\alpha_{collectors}$ = infinite.

(Fig. 2.M1). Here all collectors collect at least the most common species once; hence, the number of specimens of all common species is equal to the number of collectors (15). Rarer species are collected less than the number of collectors and this causes the horizontal line to drop off. As there are not that many very common species in Bauxite, this happened much earlier there (Fig. 2.B1). Now extending the collecting strategy with a stress factor (scenario 2), results in an upward curve at the position of the common species, which are now collected more often than the number of collectors in the area (Fig. 2.M2 and B2). The same result can be obtained modeling scenario 1 and adding a few collectors that collect just a few specimens. Hence the more collectors visiting an area, the more often the most common species will be collected (approaching scenario 4). Allowing the collectors to utilize the α-diversity of an area (scenario 3) produced the long tail that is so characteristic of the relative abundance distribution of the herbarium (Fig. 2.M3 and B3). Simulating the collecting with the rule 'never the same species twice' with the actual numbers of collectors and their actual collection sizes (Appendix 3) (Scenario 4) also results in a relative abundance distribution that closely resembles that of the herbaria (Fig. 2.M4 and B4). This is the result of the collectors that collect many specimens and find several rare species (as they never collect the same species twice) and collectors that collected only 1 or 2 specimens and invariably end up with the most common species.

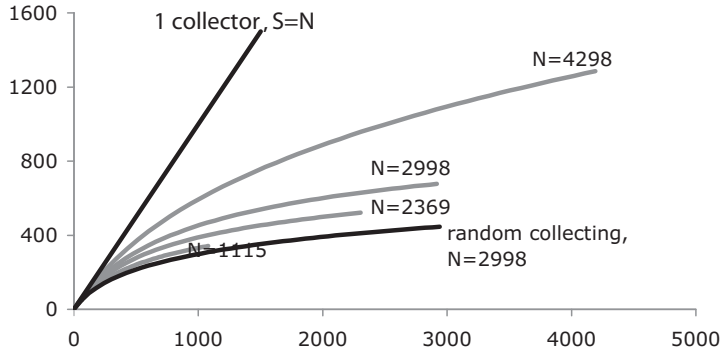

Figure 4. Species accumulation curves based on sampling the ZSMarea of Bauxite with collectors with differing relative number collections and relative abundance distribution among the collector with constant theta (c. 10). Legend: 1 collector, 1 collector collects all specimens, hence S=N; random collecting 2998 random draws from the ZSMarea; All other curves are based on randomisation of model out with rule 'never the same species twice' and varying numbers of collections (1115, 2369, 2998, 4298).

Collecting the ZS_{Marea} of Bauxite with varying relative abundance distribution of the number of collections over the collectors, with $\alpha_{collectors} = 0$ (1 collector collects all 3,000 specimens), to $\alpha_{collectors} = \infty$ (3,000 collectors each collecting one specimen) resulted in greatly varying number of species and species accumulation curves. When high $\alpha_{collectors}$ were used (from infinite (=random sampling) to 50), i.e. many collectors making small number of collections, almost similar accumulation curves were produced with almost similar number of species collected (c. 460, Fig. 3). When $\alpha_{collectors}$ was lower than 50, the accumulation curves became steeper, due to a larger number of collectors collecting large numbers of specimens and thus species. The highest number of species was, obviously, the single collector with 3,000 specimens (and thus 3,000 species). Search time for this simulation was (expectedly) very long.

Collecting the ZSM_{area} of Bauxite with varying number of collections but with constant $\alpha_{collectors} = 10$, also resulted in widely varying species accumulation curves (Fig. 4). The lowest number of species was found with a small number of collections and the number of species logically increased when the number of collections increased. It increased more rapidly with increasing numbers, though, hence the steepness of the accumulation curve depends on the number of collections made. This is the result of the fact that with large number of total collections, the collector with the highest number of collections strongly determines the total number of species.

Discussion

We have two models that are quite good in producing the relative abundance distribution in the herbarium from the actual abundance distribution of trees in the field – scenario 3 and scenario 4. Although scenario 4, combining the actual number of collectors and their actual collection sizes, is arguably the most parsimonious model, we argue that it can only be partly correct. Firstly, the data clearly show that more than half of the collectors did not follow scenario 1 (see Appendix 4.1), collecting more than one individual of several species. We know that this is caused by the need to collect a certain amount specimens during the expeditions (a common goal is often 400-500 specimens) and the time to find new species may become too long. Resident collectors for instance are often hired to collect a fixed number of specimens per month and cannot fill this quota with unique species. Botanists may also collect the same species more than once if, for instance, they find a new individual of that species at a different place, find a more representative individual, or just forgot they collected the species already. Secondly, Scenario 3 (including habitat diversity into the collecting strategy) definitely is employed by botanists who specifically aim to collect in all different vegetation types in an area during their expeditions (pers. obs.) and thus find a larger number of rare species.

Our simple model shows that just one rule 'never collect the same species twice' is responsible for the flat relative abundance distribution of herbarium specimens, while a significant addition to the flatness (tail with rare species) is caused by trying to include as many vegetation types as possible. The fact that common species are still relatively abundant in the herbarium is caused by 1) the fact that they tend to be collected by all collectors, as opposed to the rare species and 2) the fact that some botanist spend only a small amount of time in the field, and then collect mainly the most common species in flower or fruit. Many estimators of diversity such as Fisher's α and Chao's estimator are very sensitive to the large number of singletons and the use of these estimators on herbarium data will thus provide a serious overestimate of diversity.

Sampling the modeled relative abundance distribution of the field with a log series of specimens collected per collector may allow a mathematical model to reverse engineer the relative abundance distribution of the field (and hence diversity measures) from herbarium specimens. Our results, however, show that this reverse engineering will suffer from non-linear relationships. Only with

relatively high $\alpha_{collectors}$, that is many $\alpha_{collectors}$ with a small number of specimens per trip, $\alpha_{herbarium}$ approaches the α of the ZSM_{area}. As the $\alpha_{collectors}$ of our total herbarium is rather high (98 for the collectors in Guyana and 104 for those of Suriname), we might assume that the αherbarium can be used to estimate the number of species in the field. The $\alpha_{herbarium}$ values are extremely high though (598 for Guyana; 583 for Suriname). With an average number of trees per ha of 500, and a surface area of 21 mln ha for Guyana (85% of which is forest), this leads to a number of species in the order of 9,500 for Guyana alone. The total estimate for the Amazon (Hubbell *et al.* 2008) was 12,000. These numbers are difficult to reconcile. We therefore conclude that our herbarium still contains too many singletons to make accurate estimates. Perhaps, as seems to be the case for the many forest stations, collectors had prior knowledge of what species were already present in the herbarium and they extended the rule never the same species twice over different expeditions.

Sampling from the ZS_{Marea} shows us one more interesting result. Species accumulation curves do not level off. After the most common species have been accumulated in the re-sampling of the herbarium specimens, rare species are added at a almost constant rate. Hence the curves neigh to a straight line upwards. Larger samples, with more collectors having large collections, will increase faster in the beginning, suggesting higher species richness. As all lines will finally end up in the total number of species of the full region (1,252 for Bauxite) when all individuals (46,670,615,338 for Bauxite) have been sampled, this clearly is wrong. A similar result will be obtained if we regard the number of species versus the number of individuals sampled as a species area curve (saturated landscape = fixed number of individuals per area). Even when all trees have been sampled the line is not horizontal – it is a power function. Clearly, the Michaelis-Menten curve is thus an inappropriate model for such data, despite its abundant use and recommendation for the estimation of species richness in plot and herbarium data (Colwell & Coddington 1994; Gotelli & Colwell 2001; Magurran 2004; Colwell 2005). Note however that some of these authors do point out some of the inherent weaknesses of these curves. Nevertheless, they are used very often and seemingly without attention for these weaknesses by many of us: searching for "Species accumulation curves" in Google Scholar (articles only!) resulted in 334,000 hits (20,300 since 2004). The combination of Michaelis-Menten and Biodiversity in almost 1000! The

curves present a very easy way of visualizing the desired and 'expected' levelling off, but in fact they do not level off to a horizontal line – the desired maximum number of species in an area. As the species accumulation curves are also sensitive to the total number of individuals collected in an area, it is difficult to compare samples of unequal sizes. Only when a smaller sample has higher accumulation rates, we can argue that it has been sampled from a more diverse region (or perhaps it had much more effective collectors).

Acknowledgments

We like to thank Brian McGill for sending Matlab scripts to calculate α and m and construct the ZSM of the field, and Sylvia Mota de Oliveira, Heinjo During and Marinus Werger for comments on previous versions of the manuscript. PPH was supported by WOTRO grant W 84-580; OSB by WOTRO grant W 84-581.

Table 1. Original plot and botanical collection data for Mabura Hill, Guyana and Bauxiet Mts, Suriname.

	Mabura Hill	Bauxite Mts
1-ha plots (#)	6	23
# individuals	3086	13241
Tree density (trees / ha)	514	576
# species	112	605
# singletons	27	135
α	23.01	131.5
Fishers α (for all plots)	22.78	130.7
m	0.934	0.954
ZSM input parameters		
Area for collections (km^2)	7500	810000
Jm calc	385750000	46631347826
α αcalc	25	140
Mαcalc	0.93	0.95
ZSMα_{area} (modeled)		
N	397898675	46631347826
S	400	2661
Singletons	17	55
Fishers α	24.0660785	135.4

Chapter 5

Using herbarium data to assess the roles of dispersal and environmental constraints in shaping the floristic composition of the Guianas

5

With Hans ter Steege, Jean-Jacques de Granville, Hervé Chevillotte and Michel Hof

Abstract

1. After decades of intense debate, it is still not clearly understood to what extent dispersal-limited neutral dynamics and processes inherent to the biology of species (such as environmental determined processes) influence species composition across landscapes. So far, much of the data used to test the neutral and niche theories were based on permanently censused tree plots which typically extend over small scales and contain only a small fraction of the total species pool. Herbarium records may provide a useful complementary source of information to test ecological theories, because they typically contain extensive information over large spatial scales, and they may also be merged with more comprehensive species lists for large areas.

2. Using herbarium data that were collected at different intensities per 0.5 degree grid cell, we firstly used Mantel tests to examine to what extent geographical distance and environmental differences explain variation in floristic distribution patterns across the Guianas (Guyana, Suriname and French Guiana). Secondly, we used the variation partitioning approach to examine what percentage of the variation in species composition can be explained by geographical distance and environmental differences. Thirdly, we examined whether species with life history traits that assist in dispersal (dispersal mode and growth form) showed a lower rate of distance decay of floristic similarity than other species.

3. Variation in floristic composition was strongly determined by geographical distance, altitude and temperature but less by rainfall and seasonality of

precipitation. The total variation in species composition explained by these factors ranged between 17.8 and 24.2% and the amount explained by geographical distance and environmental factors was higher with higher collecting intensity. Species with assisted modes of dispersal showed a lower rate of distance decay than those that are assumed to be poor dispersers. Trees, shrubs and palms showed a higher rate of distance decay than herbs, climbers and epiphytes.

Key words: herbarium data, floristic similarity, the Guianas, dispersal limitation, environmental differences

Introduction

Although dispersal and niche processes have formally been recognized as fundamental in shaping species composition among sites across landscapes, there is no agreement about the relative importance of these processes, even after decades of intense discussion (Condit *et al.* 2002; Hubbell 2001; Mcgill *et al.* 2005). The neutral theory predicts that floristic similarity decreases (or decays) between sites in a community as a function of geographical distance (Hubbell 2001). This theory demonstrates that biological patterns emerge even if it is assumed that all individuals are ecologically equivalent and differences in community composition are caused by random dispersal, birth and death. Before the 'success' of Hubbell's neutral theory, Nekola and White (1999) examined the effect of geographical distance on floristic similarity, based on comparisons with nearly complete floras. They concluded that distance decay of floristic similarity was caused by dispersal limitation, which is in fact in complete agreement with the predictions of the neutral theory. Further, they found that distance decay of similarity was correlated with plant growth form, dispersal type and the rarity of the species in biological communities. So far, much of the data from tropical forest communities to test the neutral theory has come from tree plots which typically contain only a small fraction of the total regional species assemblage (Chust *et al.* 2006; Condit *et al.* 2002; Duque *et al.* 2002; Phillips *et al.* 2003; Potts *et al.* 2002; Tuomisto *et al.* 2003a; Tuomisto *et al.* 2003b; Vormisto *et al.* 2004). In tropical areas, arguably the most species rich terrestrial communities, well identified plot data are still rare (see refs above) and an analysis with complete floras has not yet been carried out.

One of the drawbacks in rich tropical areas is that many regional floras are still incomplete. For instance, for the Flora Neotropica only 98 angiosperm families of the Neotropics have been published, while for the Flora of the Guianas 30% of all angiosperm families or 21% of all currently known species have been published. Yet, the Guianas are relatively well collected, and many of the herbarium specimens have been reviewed by specialists. We propose that herbarium databases provide a useful data source to test the neutral theory because large 'plots' can be constructed on large spatial scales. Herbarium databases are more comprehensive than plot data because botanists put much emphasis on collecting rare species while plot data are dominated by common species and rare species are few. Furthermore, herbarium databases comprise all plant groups, not only trees. As these groups vary in attributes related to dispersal – dispersal mode, growth form, wood density and seed size – certain predictions of the neutral theory can be tested.

We suggest, as predicted by the neutral theory, that floristic similarity between sites in the Guianas decreases with increasing geographical distance between the sites and that the rate of distance decay of similarity in this area is associated with the dispersal mode, seed size, wood density and growth form of the species. If dispersal mode is important for dispersal distance (Hubbell 2001) then species that have better dispersal abilities (e.g. wind-dispersed species) are expected to be more wide-spread and show a slower rate of distance decay than species that are poor dispersers (e.g. mammal-dispersed species) (Nekola & White 1999). If seed size is a limiting factor for dispersal then species with low seed weights are expected to be more wide-spread and show a slower rate of distance decay than species with high seed weights. Species with high wood density are characteristically slow growing, mature at a later age and produce heavier seeds than species with low wood density. Wood density is therefore partly collinear with seed mass. We thus expect that species with low wood density will be more wide-spread and show a slower rate of distance decay than species with high wood density.

Alternatively, differences in floristic similarity across the Guianas can be the result of environmental differences. Niche theory predicts that floristic composition varies with environmental conditions as a result of species-specific adaptations to the environment (Hubbell 2001; Tilman 1982). The ecological niche of a species is the suite of environmental conditions necessary for the

maintenance of the species population. The distance decay of floristic similarity for growth forms is expected to be influenced by both environmental conditions and dispersal mode (see above). In the moist tropics, herbs, epiphytes and lianas typically have small seeds with good dispersal abilities (mainly wind or bird-dispersed) and we expect that they will be more evenly distributed across the landscape than trees, shrubs and palms which typically have larger seeds that are poorly dispersed (by mammals or autochory).
In this paper we examine whether herbarium data of a well collected tropical area can be used to test the predictions of neutral and niche theory. More specifically we aim to: (1) assess to what extent geographical distance and environmental factors explain the floristic composition across the Guianas; (2) quantify the fraction of the variation in species composition that can be explained by geographical distance, environmental factors and a combination of these factors; (3) examine whether better dispersers show a slower rate of distance decay in similarity than poor dispersers; (4) examine the extent to which geographical distance and environmental factors can explain the floristic composition depends on collecting intensity.

Methods

Data preparation

The Angiosperm dataset for the Guianas was extracted from the database of the Nationaal Herbarium Nederland, Utrecht Branch. This herbarium is very up-to-date and most of the material has been reviewed by specialists, as part of the edition of the Flora of the Guianas. The database was augmented by data from other herbaria of the Flora of the Guianas Consortium. The species names, as shown on each label, were updated, based on the Smithsonian's 2005 Web Listing of Plants of the Guiana Shield and the Guianas (www.mnh.si.edu/biodiversity/bdg/planthtml/index.html) and the W^{3}tropicos website (www.mobot.org/W3T/Search/vast.html).
We overlaid 0.5 degree grid cells on the map of the Guianas and spatially aggregated the specimen occurrence data into these grid cells. Three working datasets were constructed by selecting grids cells with: (i) more than 500 specimens (referred to hereafter as Grid500); (ii) more than 1,500 specimens (referred to hereafter as Grid1500) and (iii) more than 3,000 specimens

(referred to hereafter as Grid3000). In all of the datasets, species were grouped according to their growth form (climber, herb, epiphyte, palm, shrub and tree). A fourth dataset (referred to hereafter as FiveSites) was constructed by extracting the tree data from the database and aggregating them into 1 degree grid cells around five localities – the Mabura Hill area in Guyana, the combined data from three major bauxite mountains in Suriname (Lely, Brownsberg and Nassau), and Nouragues, Saül, and Piste de Saint-Elie in French Guiana. Collection effort around these five localities was exhaustive because each of these sites has been or is associated with temporary or permanent research stations. All species datasets were transformed from abundance to presence/absence data. These datasets represent a gradient in collecting intensity with collecting intensity increasing from the first to the fourth dataset. The extent to which geographical distance and environmental factors can explain the floristic composition depends on collecting intensity. It is expected that the higher the collecting intensity the greater the extent to which geographical distance and/or environmental factors would explain floristic composition.

Only tree species from the Grid500, Grid1500 and Grid3000 datasets were used when testing the effect of dispersal mode on distance decay. Species were subdivided into groups according to their dispersal syndromes (mammal or wind dispersal), wood density class (high (>0.7g cm^{-3}) and low (≤0.7g cm^{-3})) and seed dry mass class (≤3g, and >3g). Since most of the data on biological characteristics cannot be gathered from herbarium records, we compiled the life history traits for species from existing literature sources (Chave *et al.* 2006; Hammond & Brown 1995b; Hammond *et al.* 1996; Mori *et al.* 1996; Mori *et al.* 1997; Mori & Brown 1998; van Roosmalen 1985). For species without published information on seed dry mass or wood density class, we used the class value of the genus. We justify this choice because studies have shown a strong correlation between plant characteristics (e.g. seed size and wood density) and phylogeny (Chave *et al.* 2006; Hammond & Brown 1995b; Hammond *et al.* 1996; Mori *et al.* 1996; Mori *et al.* 1997; Mori & Brown 1998; van Roosmalen 1985).

Using the variables for the current conditions (between ~1950 and 2000) from the WORLDCLIM dataset (www.worldclim.org) together with the specimen occurrence records, we calculated average values for the altitudinal and climatic variables at all locations. The climate data of interest were mean annual precipitation (mm year-1), rainfall seasonality (Coefficient of Variation), and

mean annual temperature (°C). The geographical distance matrices were based on latitude and longitude of the centre of each grid cell.

Statistical analysis

To assess to what extent geographical distance and environmental factors explain variation in floristic composition across the Guianas we used the Mantel approach (Legendre & Legendre 1998). Floristic dissimilarity matrices were calculated separately for each dataset using the Bray-Curtis index.
The environmental distance matrices were based on altitude and climate (mean annual rainfall, precipitation seasonality and mean annual temperature) and were computed using Euclidean distance. The geographical distance matrices were calculated using the latitude and longitude coordinates of the centre of each grid cell. The geographical distance matrices were then logarithmically (ln)-transformed since floristic similarity is expected to decrease logarithmically with increasing geographical distance (Hubbell 2001). We first used the simple Mantel statistic (r) to assess the Pearson correlation between floristic, ln-transformed geographical or environmental distance matrices. We then used partial Mantel tests to examine whether floristic and environmental distance matrices were still correlated after the effect of geographic distance was removed. Only environmental variables that were significantly correlated with the floristic distance matrices were used in the partial Mantel tests. The functions 'mantel' and 'mantel.partial' found in the vegan library of R were used (Oksanen *et al.* 2007; R Developement Core Team 2006). All analyses were repeated for all four datasets in order to understand how collection intensity influenced the correlation between floristic, geographic and environmental distance matrices.
The original geographical coordinates were subjected to spatial decomposition using the Principal Coordinates of Neighbouring Matrices (PCNM) analysis. This was necessary because the geographical coordinates give linear trends in species composition in either the latitudinal or longitudinal direction across the sites (or some additive combination of the two trends). The wavelengths of the PCNMs range across all spatial scales included in the grid distribution. To do this we used the 'pcnm' function in the SpacemakerR library (Dray *et al.* 2006). We then selected the PCNMs and the environmental variables that contributed significantly ($p < 0.5$ after 999 permutations) using the forward selection method. The function 'forward.sel' in the 'packfor' library was used. Only those

variables that were significant were used in further analysis (Legendre 2008). We used the 'variation partitioning method' to quantify the fraction of the variation in species composition among the grid cells that was explained by the explanatory variables and their combined effects (Borcard *et al.* 1992;Legendre *et al.* 2005; Legendre 2008). The two explanatory variables were environment (climate and altitude) and distance (PCNMs). The variation partitioning was computed using the function 'varpart' of the vegan library. This function performs redundancy analysis (RDA) and reports the adjusted coefficients of determination (R_a^2). The R_a^2 gives an unbiased estimator of the contribution of each set of the explanatory variables in explaining the species composition (Legendre 2008). The statistical significance of the fractions of variation was tested using RDA and ANOVA functions of the vegan library. Only the first three (Grid500, Grid1500 and Grid3000) datasets were used during this analysis. The FiveSites dataset was not used because there were too few sites included in the dataset and this could lead to over-fitting of the data.

To examine whether species with functional traits that assist in dispersal show a lower rate of distance decay than species that are poor dispersers, we compared the rate of decay in similarity per ln-transformed unit distance. For each life history trait - based on dispersal syndromes (mammal or wind dispersal), wood density class (high (>0.7 g cm^{-3}) and low (≤0.7 g cm^{-3})) and seed dry mass class (≤3 g, and >3 g), separate floristic distance matrices were calculated using the Bray-Curtis index. The slopes of the linear regression lines of the similarities against the ln-transformed geographical distances were calculated. We then examined whether the slopes associated with each functional trait (e.g. mammal versus wind dispersal) were significantly different from each other with Analysis of Covariance (ANCOVA) using R.

To examine whether the rate of distance decay in floristic similarity depends on the growth form sampled, we compared the slopes of the linear regression lines of the floristic similarity values against the ln-transformed geographic distances in a similar manner as with the life history traits. Only the first three (Grid500, Grid1500 and Grid3000) datasets were used in this analysis.

Table 1. Mantel correlations between floristic and geographic or environmental distance matrices. Floristic distance matrices were calculated using the Bray-Curtis index and were based on presence-absence data, while geographical or environmental distance matrices were calculated using the Euclidian distance. The Mantel statistic Mantel r was based on Pearson's correlations using 1,000 permutations. The significant p-values are denoted by * where *P < 0.05, **P < 0.01, ***P < 0.001)

	Grid500	Grid 1500	Grid 3000	FiveSites
Simple Mantel tests				
ln-transformed Geographical distance	0.507***	0.690***	0.797***	0.942**
Altitude	0.420 ***	0.577***	0.425**	-0.325
Annual rainfall	0.045	0.158	0.400*	-0.437
Mean annual temperature	0.342 ***	0.540***	0.519***	-0.019
Rainfall seasonality	-0.111	0.246	0.568*	-0.119
Partial Mantel tests removing the effect of geographical distance				
Altitude	0.368***	0.523***	0.248*	
Mean annual temperature	0.295***	0.477***	0.282	
Annual rainfall			0.175	
Rainfall seasonality			0.281	
Partial Mantel test removing the effect of altitude				
Temperature	0.038	0.190*	0.410**	

Table 2. The partitioning of variation of the floristic community composition in response to the explanatory variables distance and the environment. The explained variation is divided into the components solely environment, solely geographical distance and shared geographical distance/environment. All fractions are significant ($p \leq 0.05$ for Grid1500 and Grid3000 data sets and $p \leq 0.01$ for the Grid500 data set) after 999 permutations.

	Environment	Geographical distance	Shared	Residuals
Grid500	0.01	0.10	0.07	0.8
Grid1500	0.04	0.06	0.09	0.81
Grid3000	0.06	0.08	0.10	0.76

Table 3. The slopes of linear regression lines for distance decay curves of geographical distance. The data sets contained grid cells with more than 500, 1,500 or 3,000 specimens, respectively. The slopes are calculated based on the ln-geographical distance against Bray-Curtis similarity.

Slope	Grid 500	Grid 1500	Grid 3000
Dispersal mode			
Wind	-0.088	-0.094	-0.101
Animal	-0.071	-0.133	-0.129
Seed mass			
Low	-0.053	-0.104	-0.100
High	-0.073	-0.139	-0.137
Wood density			
Low	-0.073	-0.141	-0.136
High	-0.073	-0.139	-0.144

Table 4. The slopes of linear regression lines for distance decay curves for the three data sets. The data sets contained grid cells with more than 500, 1,500 or 3,000 specimens respectively. The slopes are calculated based on the ln-geographical distance against Bray-Curtis similarity.

Slope	Grid 500	Grid 1500	Grid 3000
Climber	-0.059	-0.097	-0.087
Epiphyte	-0.046	-0.113	-0.107
Herb	-0.059	-0.098	-0.089
Palm	-0.070	-0.187	-0.151
Shrub	-0.077	-0.127	-0.111
Tree	-0.075	-0.139	-0.139

Results

A total of 7,143 species and 169,182 specimens were recorded in 164 grid cells. There were large variations in the number of species (between 1 and 2,069) and specimens (between 1 and 10,523) per grid cell. Only 81 (49%) of these grid cells contained more than 500 specimens, while only 28 (17%) and 15 (9%) grid cells contained more than 1,500 and 3,000 specimens, respectively. After discarding the grid cells with less than 500 specimens, a total of 6,965 species remained in the dataset and this was used during the analysis. About 34% of all species were collected only once or twice. While trees outnumbered the other growth forms (2,294 species and 59,633 specimens), palms were the least represented (80 species and 1,909 specimens). For the FiveSites dataset about 32.5% of the 568 tree species occurred in at least four of the five sites.

Mantel correlation with distance matrices

Bray-Curtis similarity ranged between 0.05 and 0.67 among the pairs of grid cells (average = 0.24). For the FiveSites dataset, Bray-Curtis similarity ranged between 0.38 and 0.816 among the pairs of grid cells (average = 0.58). For the FiveSites dataset, floristic distance matrices correlated strongly and significantly ($p < 0.01$) with geographical (Mantel $r = 0.94$) but not with the environmental distance matrices (Table 1). Floristic distance matrices correlated strongly and significantly ($p < 0.001$) with the geographical and environmental distance matrices for the Grid500, Grid1500 and Grid3000 datasets (Table 1). However, the most strongly correlated variable was the geographical distance. The lowest correlations were found in the Grid500 dataset for the environmental and geographical distance matrices. Annual rainfall and rainfall seasonality correlated significantly ($p < 0.05$) with floristic distances only for the Grid3000 dataset. When the effect of geographical distance was controlled, the Mantel statistic for the correlation between altitudinal distance matrix and the floristic distance matrix was reduced although it remained significant for the Grid500, Grid1500 and Grid3000 datasets (Table 1). When the effect of geographical distance was controlled, the Mantel statistic for the correlation between the temperature distance matrix and the floristic distance matrix was reduced although it remained significant for the Grid500 and Grid1500 datasets when the effect of geographical distance was controlled (Table 1). When the effect of geographical distance was controlled, the Mantel statistic for the correlation between climatic distance matrices and the floristic distance matrix did not remain significant

for Grid3000 dataset (Table 1). When the effect of altitude was controlled, the Mantel statistic for the correlation between the temperature distance matrix and the floristic distance matrix was reduced although it remained significant (except for the Grid500 dataset).

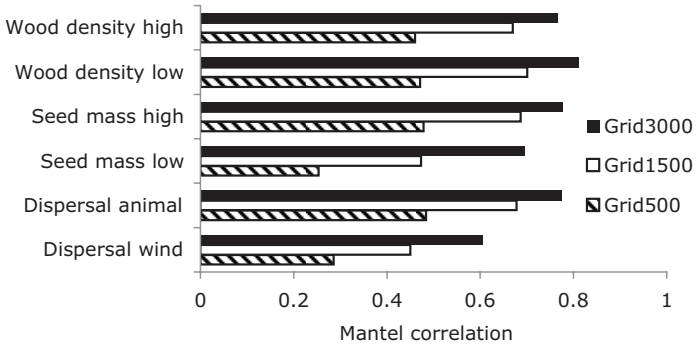

Fig. 1. Mantel correlations between floristic and ln-transformed geographic distance matrices. Floristic distance matrices for the life history traits are based on presence-absence data and were calculated using the Bray-Curtis index while geographical distance matrices were calculated using the Euclidian distance. The mantel statistic Mantel r was based on Pearson's correlations using 999 permutations. The significant p-values are ***P < 0.001).

Variation partitioning

The total variation in species composition explained by geographical distance (PCNMs) and the environment was highest (24.2%) in the Grid3000 dataset and lowest (17.8%) in the Grid500 dataset (Table 2). Solely geographical distance and a joint effect of geographical distance and environmental variation explained a higher fraction of the variation regardless of the dataset used. The fraction of variation explained solely by environmental variation and a joint effect of geographical distance and environmental variation was lowest for the Grid500 and highest for the Grid3000 dataset.

Effect of dispersal mode, seed mass, wood density and growth form on distance decay rates

Floristic distance matrices for animal-dispersed species correlated more strongly with the geographical distance matrices than floristic distance matrices for wind-dispersed species when datasets of similar collecting intensity were compared ($p \leq 0.001$) (Fig. 1). Animal- dispersed species showed higher rates of distance decay of floristic similarity (i.e. higher value of slope) with all except the Grid500

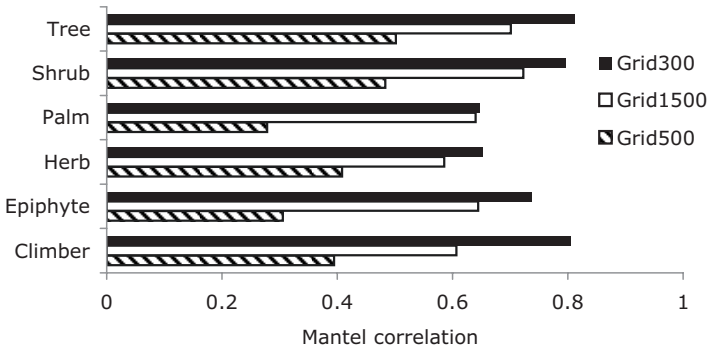

Fig. 2. Mantel correlations between floristic and ln-transformed geographic distance matrices. Floristic distance matrices for the growth forms are based on presence-absence data and were calculated using the Bray-Curtis index while geographical distance matrices were calculated using the Euclidian distance. The mantel statistic Mantel r was based on Pearson's correlations using 999 permutations. The significant p-values are ***P < 0.001).

dataset ($p \leq 0.001$) (Table 3).

Floristic distance matrices for species with high seed masses correlated more strongly with the geographical distance matrices than floristic distance matrices for species with low seed masses when datasets of similar collecting intensity were compared ($p \leq 0.001$) (Fig. 1). Species with high seed masses showed a higher rate of distance decay of floristic similarity than species with low seed masses ($p \leq 0.001$) (Table 3).

The correlation between the floristic distance matrices for species with high wood density and the geographical distance matrices was similar to the correlation between the floristic distance matrices for species low wood density and the geographical distance matrices when datasets of similar collecting intensity were compared ($p \leq 0.001$) (Fig. 1). Species with high wood density showed very similar rates of distance decay of floristic similarity compared to species with low wood density with all except the Grid500 dataset ($p \leq 0.001$).

For each life history trait, the results of the Mantel correlations between floristic similarity and geographical distance clearly differed among the datasets and were highest when the Grid3000 dataset was used and lowest when the Grid500 dataset was used (Table 3, Fig. 1 & 2).

Climbers, epiphytes and herbs showed slower distance decay rates than palms, shrubs and trees when datasets of similar collecting intensity were compared (Table 4). Climbers and herbs showed similar rates of distance decay of floristic

similarity. The rate of distance decay of floristic similarity was lower in the Grid500 than in the other datasets.

Discussion

Floristic variation in relation to geographic distance, altitude and climate

The analysis showed that the ln-transformed geographical distance was the key factor influencing the floristic composition. This was regardless of the dataset used (Mantel r ranged between 0.51 and 0.94). The joint effects of geographical distance and environment and geographical distance alone were much more important in explaining the variation in species composition than solely environmental conditions, except for Grid3000 dataset. Not only geographical distance but also, to a lesser extent, the environmental conditions are important in shaping the species composition. This can be observed from the significant correlations between altitudinal or temperature distance matrices and floristic distance matrices even when the effect of geographical distance was removed through partial Mantel tests (but not for the FiveSites dataset). Solely environmental conditions explained between 1-6% of the floristic variation. How can we explain the high value of the Mantel statistic between species composition and the explanatory variables and the low variation in species composition explained? The Mantel approach and variation partitioning approach address different issues. The Mantel test quantifies statistically the extent to which floristic and explanatory distance matrices (which are independent variables) are linearly correlated with each other. These correlations can be high even though overall floristic similarity between sites is low. In the variation partitioning approach the species dataset is the dependent variable and the explanatory variables are independent variables. This method considers the amount of variation in the species data that can be explained by each independent variable and combinations of them. Much of the residual variance might be explained by other environmental factors than the ones measured here. The four datasets represent a gradient in collecting intensity with collecting intensity increasing from the first (Grid500) to the fourth (FiveSites) dataset, and a gradient in spatial resolution with the largest number of grid cells sampled in Grid500 and the lowest in FiveSites. The Mantel statistic between species composition and the explanatory variables and the amount of variation in species composition explained generally increased with increasing collecting

intensity and reduced spatial resolution. This is probably because the dataset with the lowest collecting intensity has more grid cells with incomplete species lists. Many of the species were rare. Better collected grid cells will tend to have more 'complete' species lists of the area and therefore more species in common. This will lead to a higher Mantel statistic between species composition and the explanatory variables, and a larger percentage of variation in species composition explained.

How does herbarium data compare with plot studies that used similar variables to explain floristic variation? Generally the amount of variation explained with plot data tends to be higher than with herbarium data – between 46-50% (Chust *et al.* 2006), 38% (Borcard *et al.* 1992) and 22-38% (Svenning *et al.* 2004), compared with 19-24% (this study). These studies attributed most of the variation to geographical distance only or a combination of geographical distance and environmental or historical factors. This high degree of explained variation could be due to plots sites being chosen based on the criterion that conditions are 'uniform' (which leads to high floristic similarity) and/or environmental data measured are more precise.

Our results show that distance was a more important factor in describing variation in floristic composition in the Guianas than the climatic and altitudinal factors we examined. The large fraction of the joint geographical distance/ environment component and the significant partial Mantel correlations between floristic composition, and temperature and altitudinal distance matrices indicate that dispersal patterns occur in combination with the environmental gradients across the Guianas. A large proportion of the floristic variation remained unexplained, however, and this may be a result of several factors. Firstly, collecting intensity per grid cell may not have been sufficient for the Grid500, Grid1500 and Grid3000 datasets. Secondly, only few environmental variables have been measured in this study and other attributes (e.g. slopes, different vegetation types etc.) and soil variables might influence the distribution of species and floristic composition of forests in the Guianas. Thirdly, we have chosen a coarse scale for this study whereas processes that explain variation in species composition might occur at a much finer scale.

Effect of dispersal mode, rarity and growth form on distance decay rates

Distance decay of floristic similarity was a function of dispersal mode (except when the Grid500 dataset was used) and seed mass (for all datasets). We did not show clearly that floristic similarity was a function of wood density.

We predicted that distance decay of similarity is a function of dispersal mode, seed mass and wood density. Species with high dispersal ability are expected to have wider niche breadths and therefore show lower rates of distance decay than those with low dispersal ability. Our results generally support these predictions (except for the dataset Grid500 and wood density). The reason for the lack of support when using the Grid500 dataset is probably the unequal collecting intensity among the grid cells. Although studies, e.g. Nekola and White (1999), have found that distance decay was related with dispersal mode, other studies showed no support for this expectation (Eriksson & Jacobsson 1998; Chust *et al* 2006).

The rate of distance decay depended on the growth form sampled. Trees, shrubs and palms showed higher rates of distance decay than herbs, climbers and epiphytes. We expected herbs, climbers and epiphytes to be more evenly distributed in the landscape than trees, shrubs and palms and our results support our expectations. Trees, shrubs and palms characteristically have heavier seeds which are typically dispersed over shorter distances by mammals or by autochory and require more time before they can be dispersed to all suitable habitats for their establishment. The lower Mantel correlations and the lower overall similarity for the climbers, epiphytes and herbs suggest that grid cells are more variable in their composition of these species than for trees, shrubs and palms.

We conclude that herbarium data can be used to test the predictions of neutral theory and niche theory. Floristic composition across the Guianas is strongly determined by both geographical distance and environmental factors. However, to draw meaningful conclusions about floristic similarity it is important to compare areas where collecting effort is high. When collecting effort is low only a small fraction of the species pool is documented and this leads to a low similarity among sites.

Chapter 6

Using species distribution models to determine species richness and endemism patterns for the Guianas

With Hans ter Steege, Niels Raes, Jean-Jacques de Granville, Hervé Chevillotte and Michel Hof

Abstract

When setting aside areas for conservation, information on species richness and endemism patterns is important. Yet this information to assist in selecting protected areas is scarce. Using species distribution models to address the problem of scarcity of species richness data, we: (1) identified the general patterns of species richness using two environmental datasets (one containing altitudinal, climatic and soil variables and the other containing only altitudinal and climatic variables); (2) identified the general patterns of weighted endemism; (3) determined whether patterns of species richness were different when herbarium data were modeled at the scale of the Guianas or the Neotropics (only the taxa Inga and Lecythidaceae were modeled).

The coastal area of all three countries and the interior of French Guiana were predicted to be more species rich than the rest of the Guianas. The south-eastern part of Guyana and the south of Suriname were predicted to be poor in species. The species richness patterns using two environmental datasets were highly correlated. The Pakaraima and the Kanuku mountains of Guyana were predicted to have the highest levels of endemism. For both Inga and Lecythidaceae, the general patterns of species richness were the same when SDMs are based on herbarium data from the Guianas or the Neotropics.

Key words: Species distribution modeling, species richness, endemism, soil, climate, altitude, Guianas, Neotropics.

Introduction

The Guianas (Guyana, Suriname and French Guiana), together with the Amazon basin, form the largest area of contiguous rainforest in the world. The southern part of the Guianas, in combination with the adjacent areas of Brazil, has been classified by Mittermeier et. al. (1998) as a major tropical wilderness area, because of its low human population density and large proportion (>75%) of intact vegetation. Major tropical wilderness areas are vast storehouses of biodiversity, which provide refuge for many endemic and rare species, and are of high significance when determining conservation areas. Given the accelerated demands on the pristine vegetation in the Guianas for mining and logging activities (Hammond 2005), knowledge of the patterns of species richness and endemism is crucial for effective ecosystem management. However, these patterns remain unknown for most areas.

Angiosperm biodiversity across the Guianas (gamma diversity) is the result of species richness (alpha diversity) and turnover in species composition among habitats (beta diversity). All three of these diversity elements are of great relevance in conservation planning decisions. For the Guianas, plot and inventory studies show that tree alpha-diversity increases from west to east (ter Steege *et al.* 2003). Also the south of Guyana and Suriname were shown to have a higher alpha-diversity than the north (which includes the coast) (ter Steege *et al.* 2003). Herbarium databases may provide a useful complementary source of information to examine species richness and endemism patterns, because they typically contain extensive information on all growth forms collected over large geographical areas. However, the problem with using herbarium data is the inadequate coverage of geographical areas.

To fill the gaps in poorly surveyed areas methods such as species distribution modeling (SDM), have become available to determine species ranges based on observed patterns of occurrence and ecological preferences (Araújo & Guisan 2006; Phillips *et al.* 2006; Raes *et al.* 2009). Therefore, the quality of the species and environmental data used is crucial for the accuracy of SDMs. The main concern about species data derived from herbarium databases is that the sampling localities of specimens are geographically biased towards areas that are easily accessible by rivers and roads or closer to populated areas (Reddy & Davlos 2003; Kadmon 2004, Chapter 3). As a consequence, collecting effort may not be random with respect to the environmental conditions (Chapter 3) and

this may affect the accuracy of SDMs (Kadmon *et al.* 2004; Hortal *et al.* 2008; Loiselle *et al.* 2008). We demonstrated that, although there is geographic bias in sampling effort in the Guianas (Chapter 3), the specimen localities nevertheless adequately represented the range of environmental conditions that prevail in the Guianas, as the river and road network is well distributed across the area. The Guianas occupy only a small part of the Neotropics and many species that occur in the Guianas also occur in the rest of the Neotropics (ter Steege 2003; Hopkins 2007). Nevertheless, it may be that the general patterns of species richness will be different when SDMs are based on herbarium data from the Guianas and the Neotropics.

The environmental variables used in SDMs should fully describe the ecological conditions under which a species persists (Araujo & Guisan 2006). Climatic, altitudinal and soil variables are commonly used for SDMs. Field studies have shown that at small spatial scales, variation in tree diversity is strongly related to soil variables (ter Steege & Hammond 2001). Compared to other regions there is relatively little variation in climatic variables across the Guianas and this might limit the use of climatic variables in SDMs. We expect that SDMs using only altitude and climatic data will yield less stable models than those involving altitude, climatic and soil data.

A species is considered to be endemic if its distribution range is restricted to a specific area. Areas where a large number of endemics occur are of high priority when choosing areas for conservation. Published research has suggested that some areas in the Guianas have high concentrations of endemics. In Guyana, the Pakaraima mountains region and Central Guyana are the areas of high endemism (ter Steege *et al.* 2000). In French Guiana, Saül is suggested to have the highest number of endemic species (Conservation International 2003). In Suriname, the Tafelberg mountain area, Eilerts de Haan Gebergte and the Sipaliwini area have large numbers of endemic species (Conservation International 2003). Some areas in the Guianas are still poorly surveyed and therefore the distribution range of many species is incompletely documented. This could lead to an over-estimation of the endemism richness patterns. By using SDMs to fill the gaps in poorly surveyed areas, better estimates of endemism patterns can be obtained.

To address the problem of scarcity of species distribution data and to determine spatial patterns of biodiversity in the Guianas, we use species distribution

modeling to: ((1) identify the general patterns of species richness using two environmental datasets (one containing altitudinal, climatic and soil variables and the other containing only altitudinal and climatic variables); (2) identify the general patterns of weighted endemism; (3) determine whether different species richness patterns are obtained when herbarium data from the Guianas or the Neotropics are used (only the taxa Inga and Lecythidaceae were modeled).

Methodology

At the scale of the Guianas

Angiosperm occurrence data were extracted for the Guianas from the database of the 'Nationaal Herbarium Nederland', Utrecht Branch. This database was supplemented by data from other herbaria of the Flora of the Guianas Consortium. Coordinates for the localities were copied from the specimen labels (and converted to decimal degrees when necessary) or came from national gazetteers when only descriptive locality information was available. Only species with five or more specimens with unique localities were used (Chapter 2).
The ecological landscape for the Guianas (defined as 0.0° to 9.0° N, 51.0° to 62.0° W for the purpose of this paper) was defined by 17 soil variables that were obtained from the FAO Geonetwork database (www.fao.org/geonetwork/srv/en/main.home), altitude (derived from a digital elevation model, DEM) and 19 bioclimatic variables for the current conditions (~1950-2000) from the WORLDCLIM database (www.worldclim.org) (Hijmans et. al. 2005a). We chose these data sources because they are the best data available that cover the geographical extent of the Guianas. Soil, DEM and climatic variables are known to be related to species distribution patterns. All variables were selected at 5 arc-minutes spatial resolution and were re-sampled to match the spatial extent of the FAO soil data layers which were more restricted in coverage than the DEM and bioclimatic data layers. This resulted in a total of 37 data layers. We reduced the number of data layers by removing those that were highly inter-correlated and are therefore redundant in describing the environment (Peterson 2008; Raes *et al.* 2009). Reduction of the soil variables was performed using a principal components analysis (PCA) on the complete set of soil variables. We retained the first four component scores of the PCA which jointly explained about 78% of the overall variation among the soil variables. We performed a

Pearson's correlation on the 19 bioclimatic variables and the DEM and selected the nine least correlated variables ($r < 0.7$). Therefore after removing redundant variables 13 data layers remained to describe the ecological landscape of the Guianas – four principal component scores summarizing soil variation, DEM and 8 bioclimatic variables (Figure S6.1). We then made two datasets from these data layers. The first dataset contained all of the 13 data layers (referred hereafter Soil_DEM_Clim dataset) while the second dataset (referred hereafter DEM_Clim dataset) contained DEM and eight bioclimatic variables but not the four soil variables.

Environmental bias in sampling effort

To examine whether the herbarium data showed an environmental bias in sampling effort, we compared the differences in the distribution of the environmental variables between the grid cells that were visited by botanists and those that prevail for the whole of the Guianas. We divided each of the 13 environmental variables for each of the two datasets (i.e. based on the observed and the whole of the Guianas) into 10 equal interval groups (bins) based on the range of each variable. The difference between the two datasets was tested using the Kolmogorov-Smirnov test (Kadmon *et al.* 2004; Loiselle 2008). If the datasets are significantly different then the environmental conditions of the collecting localities are different from those existing elsewhere in the Guianas, indicating that the collecting localities not properly represent the ecological conditions of the Guianas.

Species distribution modeling

Using the two environmental datasets and the species distribution data we developed SDMs using the maximum entropy method (MaxEnt Version 3.4; Phillips *et al.* 2006). We chose MaxEnt because it is specialized in modeling species probability distribution using presence-only data such as herbarium data, its performance level is high even with few species localities (Hernandez *et al.*2006; Wisz 2008) and it is demonstrated to outperform all other modeling applications available today (Elith *et al.* 2006). This method aims at estimating species probability distributions by determining the probability distribution of maximum entropy (close to uniform), subject to the constraint that the expected average value for each environmental variable should match the observed

average value (Phillips *et al* 2006). We used the same MaxEnt parameters as Raes *et al.* (2009). For each model the MaxEnt parameters were adjusted such that all species presence records were used to build each species model by setting the 'random test percentage' to zero. This is because area under the curve values do not apply when pseudo-absences are used, hence the use of null-models, and these do not require a test percentage (Raes & ter Steege 2007) The features in MaxEnt selected were linear feature when less than 10 specimens per species was used, quadratic feature when 10-14 specimens per species was used, and hinge feature when 15 or more specimens per species was used (Raes & ter Steege 2007; Raes *et al.* 2009). We removed duplicate specimens of all species in the grid cells and each species was modeled with the Soil_DEM_Clim dataset and the DEM_Clim dataset.

Collection bias correction and validating the SDMs against a null model

To evaluate the accuracy of the SDMs we used the threshold independent and prevalence insensitive area under the curve (AUC) of the receiver operating characteristic (ROC) plot (Raes & ter Steege 2007) constructed by MaxEnt. We used the method presented in Raes & ter Steege (2007) to test the AUC value of each SDM developed against a bias corrected null-model of AUC values expected by chance. For each SDM developed with n specimens, the AUC value was tested against the upper 95% one-sided confidence interval (CI) AUC value derived from the AUC values of 1,000 X n randomly drawn and modeled points. To correct for possible geographical bias in the random points were selected from grid cells in which collections were made. Of the 5,345 grid cells falling within the boundaries of the Guianas, 1,504 (or 28.1%) had collections and from these grid cells the random points were drawn.

We developed null-distributions for 5-35 records at a continuous interval, for 40-50 records at 5 record-intervals, for 60-100 records at 10 record-intervals and for 150-250 records at 50 record-intervals. These gave a total of 42 distributions. Using these null-distributions and the two environmental datasets we developed null SDMs using MaxEnt in a similar manner as with the observed species distribution data.

For each of the 42 distributions the 1,000 AUC values were ranked and the 950th of the confidence interval (C.I.) value (equal to the 95% one-sided C.I.)

was selected. We developed three data series of C.I. values – 5-9, 10-14 and 15 or more records based on the modeling features of MaxEnt (see Phillips *et al.* 2006; Raes *et al.* 2009) and used curve fitting to find a curve that best fits the data. The fitted AUC values from the null models were then used to determine the significance of the AUC for the 4,110 species modeled with the two environmental datasets. A significant SDM was recorded if the AUC of the observed species was higher than the fitted AUC of the null model.

Species richness and endemism patterns

To determine species richness patterns we first converted the continuous MaxEnt prediction values for each species into binary data (presence/absence) and then counted the number of predicted presences per grid cell. The criterion for conversion to presence/absence data was based on the '10 percentile training presence logistic threshold' (Raes *et al.* 2009). This threshold value was chosen because it is practical to assume that about 10% of the species data is either wrongly identified or geo-referenced (Raes *et al.* 2009). For each species a presence was scored if the MaxEnt prediction was greater than the threshold value of the species while an absence was scored if the prediction was less than the threshold value. To determine the total predicted species richness for each environmental dataset, the number of predicted presences per grid cell was counted for all significant species. To determine whether the predicted was lower than the observed species richness, we plotted the number of species predicted as a function of the number of species observed per grid cell. To determine whether the two environmental datasets (Soil_DEM_Clim and the DEM_Clim datasets) gave different predicted species richness patterns, we compared the species richness patterns using linear regression analysis.

To determine the endemism patterns, we first calculated the corrected weighted endemism index (CWEI) for each species (see Crisp *et al.* 2001 for details) and then summed up the CWEI values for each species per grid cell. This index is calculated in three steps. First, the inverse range size of each species was calculated. Second, the inverse range size of all species per grid cell was summed up. Third, the value for each grid cell derived from the second step was divided by the total number of species predicted per grid cell. The predicted presence/absence data for all significant SDMs based on the Soil_DEM_Clim dataset were used to calculate the CWEI.

At the scale of the Neotropics

We used herbarium data for the family Lecythidaceae and the genus Inga for the Neotropics, to examine whether the species richness patterns would be different when SDMs were modeled at the scale of the Guianas or the Neotropics. We chose these taxa because their occurrence and identification are well documented for the Neotropics. We supplemented the species data of the Guianas (see above) with the species dataset from the Lecythidaceae webpage (sweetgum.nybg.org/lp/index.html) and the Inga data from the BRAHMS website (dps.plants.ox.ac.uk/bol/samples/inga.aspx). Only species occurring in five or more unique localities were used.

The ecological landscape for the Neotropics (-23.46° to 23.46° N, -34.79° to110.96° W) was defined by the 37 environmental variables as in the case of the Guianas (see above). All variables were selected at 5 arc-minutes spatial resolution and were re-sampled in a similar manner as with the Guianas data. As with the Guianas data, we reduced the number of data layers by removing those that were highly inter-correlated and are therefore redundant in describing the environment (Peterson 2008, Raes *et al.* 2009). After performing a PCA on the soil variables and Pearson's correlation with the bioclimatic variables with the altitudinal and bioclimatic data, we redefined the ecological landscape of the Neotropics by 11 data layers – five principal component scores summarizing soil variation and six bioclimatic variables. The six bioclimatic variables used were BIO1 (Annual Mean Temperature), BIO2 (Mean Diurnal Range), BIO4 (Temperature Seasonality), BIO12 (Annual Precipitation), BIO15 (Precipitation Seasonality) and BIO18 (Precipitation of Warmest Quarter). No null model was developed to test the significance of the SDMs. To model each species we used MaxEnt with the maximum iterations of 1,000, removing the duplicate specimens of all species in the grid cells and using the default setting and automatic features.

To determine the species richness pattern for each taxa, the continuous MaxEnt prediction values of the grid cells for each species were converted into binary data (presence/absence) and the number of predicted presences per grid cell was counted in a similar manner to the Guianas data. To compare whether the species richness patterns were different when SDMs were modeled at the scale of Neotropics or of the Guianas, we first extracted the species richness data for the Guianas area from the Neotropical dataset. For each taxa, we then compared

the species richness of extracted data with the Guianas dataset using linear regression analysis.
All GIS manipulations were carried out with Manifold GIS (ver 7x, Manifold Net Ltd).

Results

At the scale of the Guianas

Our original herbarium database for the Guianas consisted of 7,148 species. However, only 4,110 of these species provided a basis for SDMs in this study, as they were represented in five or more unique localities. The largest number of these species (3,513) was collected in Guyana and the smallest number (3,307) was collected in Suriname. The number of specimens per species varied between 5 and 185 (average 25), giving a total of 102,845 unique specimens. Of the 5,345 grid cells falling within the boundaries of the Guianas, 1,504 (or 28.1%) had collections (Fig. 1). The number of species per grid cell varied between 5 and 1,710 (average 80). Only about 18% of all the species used for the SDMs were not collected along the coast. The area used for the SDMs was a larger rectangular area than the total geographical area of the Guianas (0.0-9.0 N, 51.0-62.0 W) enclosing all the country boundaries and comprising 10,259 grid cells.
The Lecythidaceae dataset contained 7,888 specimens representing 93 species with 5 to 1,581 specimens per species. The Inga dataset contained 8,406 specimens representing 201 species, with 5 to 362 specimens per species.

Environmental bias in sampling effort

The sampling localities were not evenly distributed across the study area (Fig. 1). The highest number of species per grid cell was observed in French Guiana, the north-western part of Suriname and along the coast of all three countries. Additionally, sampling was concentrated in the Rupununi savannas in the south-west of Guyana and the Pakaraima Mountains of western Guyana. The south-eastern part of Guyana and the south of Suriname were poorly sampled. Despite the non-random distribution of the collection sites, there was no significant variation between the environmental conditions of the sampled localities and all grid cells of the Guianas (Fig. S6.2).

Significance of the SDMs

Of the 4,110 species, with five or more unique collection localities, 2,934 (or 71.4% and 41% of the total 7,146 species) yielded a significant SDM with the Soil_Alt_Clim dataset. With the Alt_Clim dataset 2291 (or 55.7% and 32% of the total species) yielded a significant SDM. The number of species predicted was lower than the number of species observed for four grid cells when the Soil_Alt_Clim set was used (Fig. 2) and for 12 grid cells when the Alt_Clim dataset was used (data not shown). A large number of species were predicted to occur in many grid cells in which few species were collected (Fig. 1, 3 and 4). About 55% of grid cells within the boundaries of the Guianas showed a predicted species richness of greater than 1,000 species per grid cell compared with about 7% of the observed grid cells.

Species richness and endemism patterns

The map of predicted species richness, summing up all of the 2,934 species maps, showed a clear pattern (Fig. 3). Species richness was predicted to be highest along the coast (Fig. 3) where the sampling effort was high (Fig. 1). The interior of the *arrondissement* of Cayenne in French Guiana was predicted to have higher species richness than the rest of the interior of the Guianas based on the Soil_Alt_Clim dataset (Fig. 3). The lowest species richness was predicted

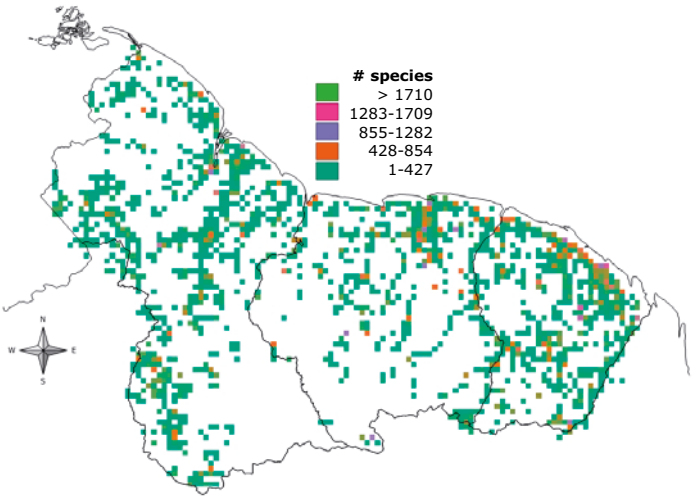

Fig. 1. The observed number of species per 5 x 5 arc-minutes grid cells in the Guianas. The south-east of Guyana and south of Suriname were the least collected.

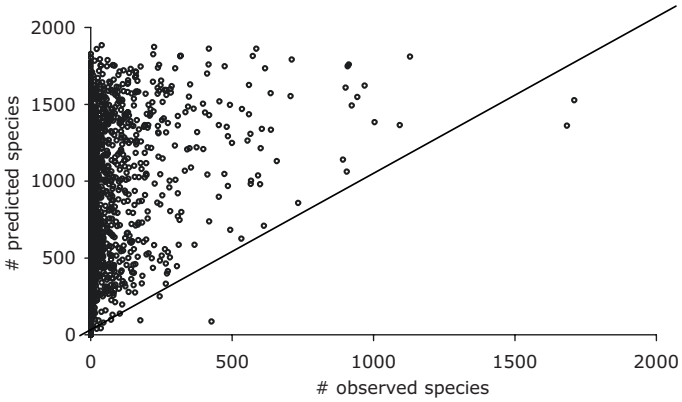

Fig. 2. The predicted number of species plotted against the observed number of species per grid cell. The four points below the diagonal line represent grid cells that were under-predicted.

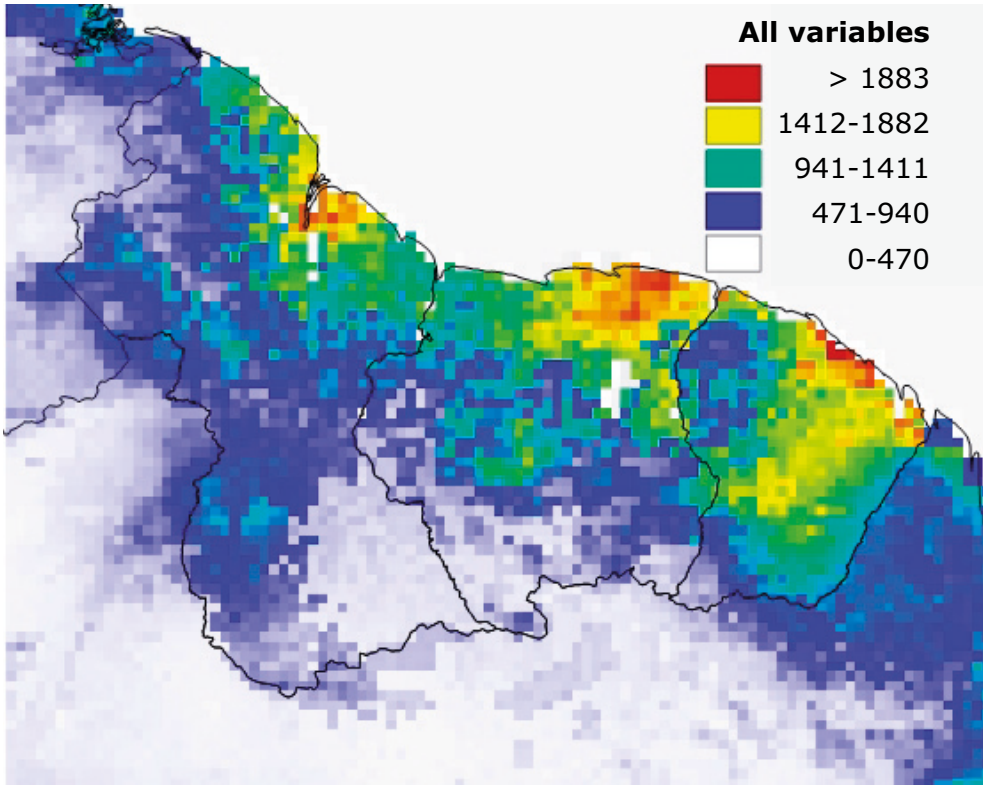

Fig. 3. Predicted botanical richness per 5 x 5 arc-minutes grid cells in the Guianas. The SDMs are based on all climatic, altitude and soil variables. Generally, French Guiana is predicted to have the highest botanical richness. The south of Guyana and Suriname are predicted to show a low botanical richness.

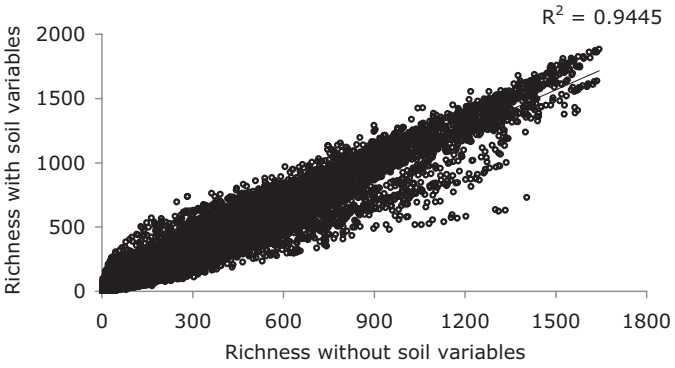

Fig. 4. Correlation between the number of predicted species per grid cell when the Soil_Alt_Clim and the AltαClim datas were used.

to be in the south-eastern part of Guyana and the south of Suriname (Fig. 3) where the sampling effort was low (Fig. 1). There was a strong correlation (R^2=0.94) between the predicted species richness values for the Alt_Clim with the Soil_Alt_Clim datasets (Fig. 4). The predicted number of species per grid cell was lower when the Alt_Clim dataset was used however, than when the Soil_Alt_Clim dataset was used.

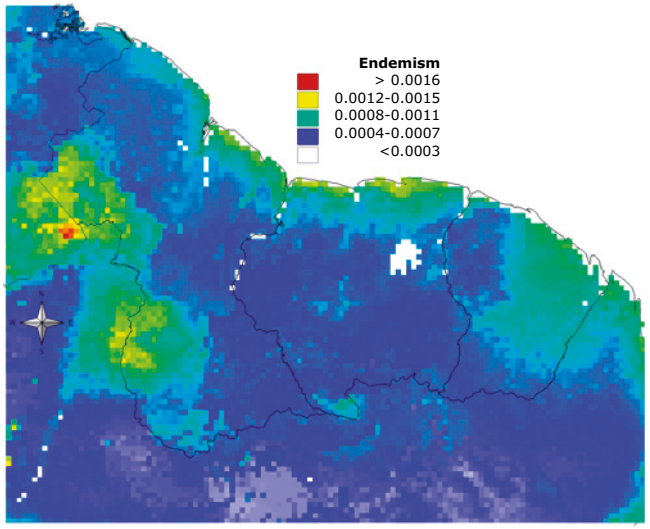

Fig. 5. Predicted corrected weighted endemism pattern per 5 x 5 arc-minutes grid cells in the Guianas. The SDMs are based on all climatic, altitude and soil variables.

The Pakaraima, Roraima and Kanuku mountain areas of Guyana and the coast of Guyana and Suriname showed the highest levels of weighted endemism compared to the rest of the Guianas (Fig. 5). Central Guiana and the Rupununi areas of Guyana, the Tafelberg mountain area and Sipaliwini Savannas in Suriname, and most of the *arrondissement* of Cayenne of French Guiana showed intermediate levels of weighted endemism.

Predicted species richness at the scale of the Neotropics

The predicted patterns of species richness were similar for the genus Inga, regardless if the specimen data used was restricted to the Guianas or covered the entire Neotropics (Fig. 6 and 7). A similar result was found for the Lecythidaceae (Fig. 8 and 9). For both Inga and Lecythidaceae the number of species per grid cell (only for the Guianas grid cells) predicted at the scale of the Neotropics and the Guianas were highly correlated (R^2=0.48 and 0.91 respectively). However, the number of species predicted per grid cell was lower when the herbarium data used for the SDMs came from the Guianas than when they were from the Neotropics.

The map of the predicted patterns of species richness for Inga and Lecythidaceae were very similar to that produced with the 2,934 species when the Soil_Alt_Clim dataset was used (Fig. 10 and 11). Here too, high species richness was predicted along the coastal areas and the interior of French Guiana. The rest of the Guianas was predicted to be botanically poorer and the poorest areas were predicted to be the south-eastern part of Guyana and the south of Suriname (Fig. 5-8).

Discussion

SDMs built with the Soil_Alt_Clim dataset were more accurate than those built with the Alt_Clim dataset, since more significant SDMs (2934 versus 2291) were obtained and fewer grid cells (4 versus 12) were under-predicted when compared to the observed species richness. This is probably because the climatic conditions across the Guianas are not highly variable and as a result these variables alone have a low predictive power for the SDMs (Saatchi 2008; Beurmann 2008). When these variables were combined with soil variables, there was more differentiation in the environment across the Guianas, leading to a higher predictive power of the Soil_Alt_Clim dataset and more significant

SDMs. The number of species predicted to occur per grid cell using the two environmental datasets was highly correlated. This is not surprising because there was a high overlap between the two environmental datasets as only four of the 13 data layers used were related to soil variables. Thus, we may conclude that the predicted patterns of species richness based on the two environmental datasets are similar.

The coastal area was predicted to have higher species richness than the interior of the countries. We think that this prediction might not be accurate for two main reasons. Firstly, the sampling intensity along the coast was high and only about 18% of the species used for the SDMs were not found along the coast. As a result more species that were collected along the coast were represented by five or more specimens in our original herbarium dataset and were therefore selected for modeling. Secondly, as the environmental conditions along the coast are distinct from those of the interior of the Guianas and within the coastal zone there is little environmental variation, most of the species occurring on the coast will also be correctly predicted. These factors together might have led to high predicted species richness along the coast.

The interior of the *arrondissement* of Cayenne in French Guiana was predicted to have higher species richness than the rest of the interior of the Guianas and this may be due to two factors. Firstly, this area is better sampled and more specimens per species for this area were used for the SDMs than for the rest of the Guianas. Secondly, the environmental variables used showed very little variation within the interior *arrondissement* of Cayenne implying that once species collected in this area give significant SDMs, it is likely that such species will be predicted to occur in a large number of grid cells.

The rest of the interior of the Guianas was predicted to have a relatively low number of species per grid cell. There are three main reasons for this. Firstly, these areas are poorly sampled and many of the species were not represented by enough specimens to be modeled. Secondly, many of the species that occur in these areas exist in very narrow climatic regimes and are therefore very difficult to model. Thirdly, because of the relatively strong altitudinal gradients there was high variation in the environmental variables. This implies that once species collected in this area give significant SDMs, it is not likely that such species will be predicted to occur in many of the grid cells in the area.

The south-eastern part of Guyana and the south of Suriname were predicted to have low species richness. There are several explanations for this species richness pattern. Sampling effort might play a role. The south of Suriname was extremely poorly sampled while south-eastern part of Guyana is not sampled at all. Even though the collecting localities were not environmentally biased, this area might still have a distinct environment form the rest of the Guianas and as a result the probability that species occurring in the rest of the Guianas be predicted in these areas is low. History might play a role but SDMs used in this research do not take this into account. Beerling and Mayle (2006) suggested that the area was grassland 21 thousand years ago. This area is now covered with forest (Huber *et al.* 1995; ter Steege 2000). This area might actually be species poor and many of the species that occur in the rest of the Guianas might not occur here. That this area might be species poor is supported by the fact that the herbarium data for Inga and Lecythidaceae from the Neotropics support the richness patterns predicted with the Guianas data well. The number of species predicted per grid cell was higher with the Neotropics than with the Guianas data. This was especially with the Inga herbarium data for the Neotropics.

The species distribution map of Hopkins (2007) using a smaller herbarium data containing 1,582 species showed that the southern part of Guyana and Suriname were predicted to be more species rich than the rest of the Guianas. Further, French Guiana was predicted to show low to intermediate species richness patterns. The main reason for the discrepancy between Hopkins' and our maps is the scale of the studies. Hopkins used a coarse spatial resolution (10 grid cells) and we used a relatively fine spatial resolution (5 arc minutes). At a coarse spatial resolution many different habitats are included in each grid cell. Since many species have limited distributions, in a heterogeneous landscape the number of species predicted per grid cell may be small at a fine scale, but species beta diversity across the landscape may be high. In a heterogeneous environment, the higher variation in environmental variables within a coarse scale grid cell will lead to higher gamma diversity. Hopkins' map predicted proportionally low to intermediate numbers of species per grid cell along the coast and all of French Guiana. In these areas although the number species predicted per grid cell is high, the similarity in species composition between grid cells is high and this will lead to lower gamma diversity.

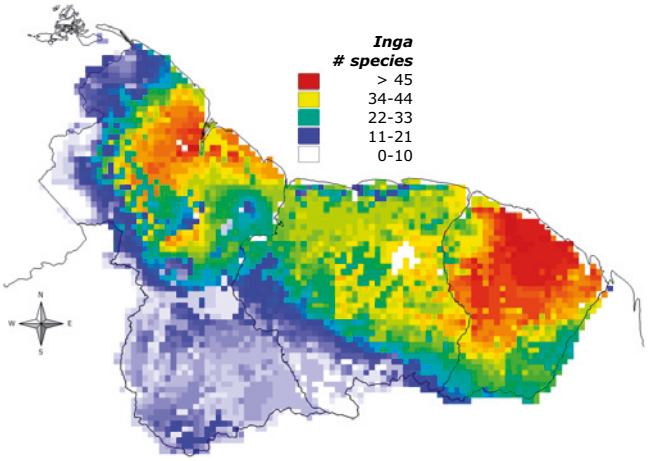

Fig. 6. Predicted Inga richness per 5 x 5 arc-minutes grid cells in the Guianas. The species data came from herbarium data for species only occurring in the Guianas.

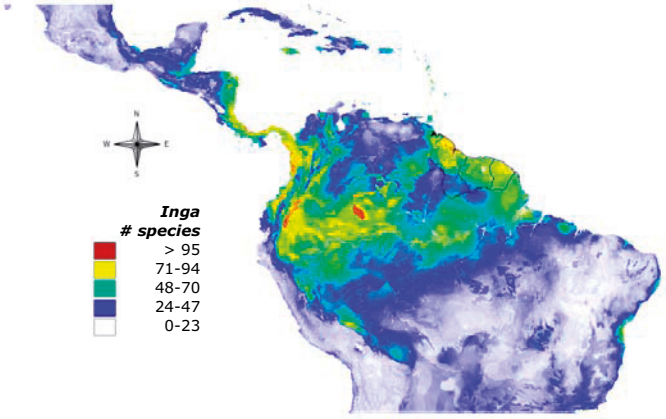

Fig. 7. Predicted Inga richness per 5 x 5 arc-minutes grid cells in the Neotropics. The species data came from herbarium data for the Neotropics.

Using plots data, ter Steege *et al.* (2003) suggested that tree alpha diversity increased from Guyana to French Guiana and from the north to the south. The plots did not sample the diversity of landscapes across the Guianas as they were all concentrated in the central but not the south of Guyana and Suriname. Therefore, while the south might have a high tree alpha diversity based on 1 ha plots this area might have low beta diversity resulting in an overall low gamma diversity.

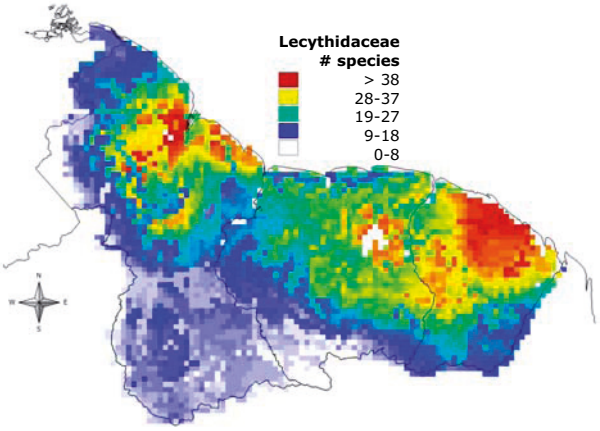

Fig. 8. Predicted Lecythidaceae richness per 5 x 5 arc-minutes grid cells in the Guianas. The species data came from herbarium data for species only occurring in the Guianas.

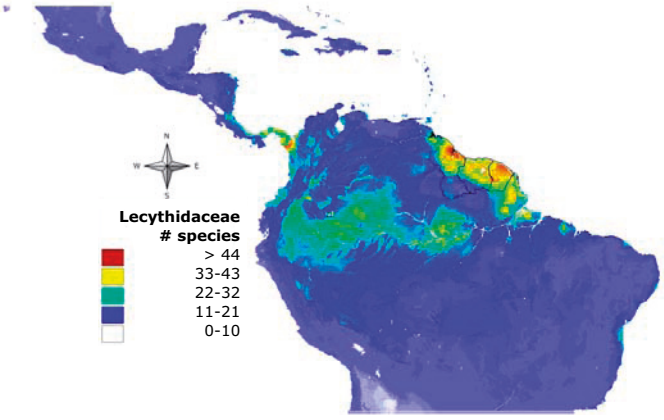

Fig. 9. Predicted Lecythidaceae richness per 5 x 5 arc-minutes grid cells in the Neotropics. The species data came from herbarium data for the Neotropics.

The highest levels of corrected weighted endemism were predicted in the Pakaraima Mountain, Kanuku Mountain areas and the coast of Guyana, the Sipaliwini area and the coast of Suriname. The Tafelberg mountain area and the *arrondissement* of Cayenne in French Guiana showed intermediate levels of endemism. The high endemism patterns for these areas are associated with low dispersal ability, low disturbance ratios and specialization on specific soil types.

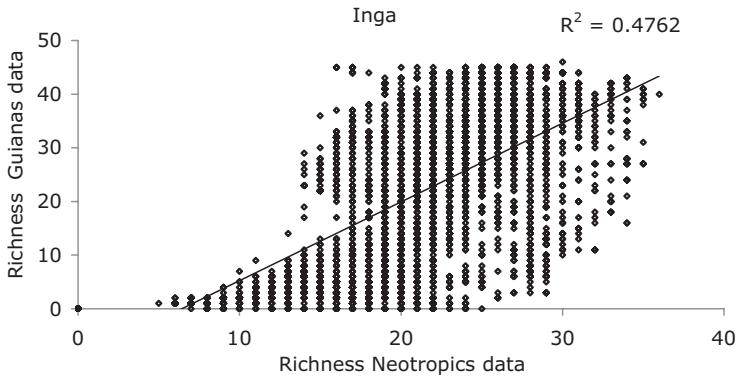

Fig. 10. Correlation between the number of Inga species predicted per grid cell for the Neotropics and for the Guianas.

Furthermore, altitude attributes to the high endemism patterns (Fanshawe 1952; ter Steege 2000). These patterns reflect to some extent the findings of published research. For example, the research of ter Steege *et al.* (2000) suggests that there were two concentrations of endemics for Guyana, the white sands area (which in some cases extends to the coast) and the head of the Mazaruni – Mt. Roraima area (which also includes the Pakaraima mountains area). For French Guiana, Saül was found to have the highest degree of endemics (Granville 1988) but our results do not support this. In Suriname, the Eilerts de Haan Gebergte and the Sipaliwini area were suggested to have a large number of endemic species (Conservation International 2003) and our results support to some extent these findings.

Not all species were successfully modeled, because of the low number of specimens per species and possibly because some species occur in the Guianas have very high local abundances but restricted species ranges making them very difficult to model. These factors might have compromised the accuracy of the predicted botanical richness maps for the Guianas. The patterns of species richness were not different when herbarium data were modeled at the scale of the Guianas or the Neotropics.

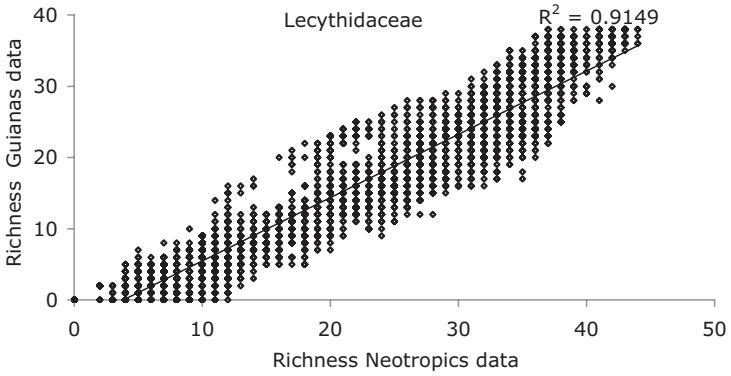

Fig. 11. Correlation between the number of Lecythidaceae species predicted per grid cell for the Neotropics and for the Guianas.

Chapter 7

General summary and discussion

Primary species occurrence data are rapidly becoming available on the internet. It is expected that in the next decade about 1 billion species occurrence records collected worldwide will be available on the internet (Guralnick & Hill 2009). The data are becoming an increasingly important source of species information for ecologists. However it is felt that before herbarium databases can be used for biodiversity studies biases associated with the databases must be assessed (Graham *et al.* 2004; Hortal *et al.* 2008). To this end, the primary occurrence data of plants collected in the Guianas (Guyana, Suriname and French Guiana) were used to assess the extent of biases associated with herbarium databases. After getting an understanding of the biases, the database was used to: (a) develop a model to simulate relative abundance distributions in the herbarium; (b) to assess the roles of dispersal and environmental constraints in shaping the floristic composition of the Guianas; and (c) determine species richness and endemism patterns across the Guianas.

The Guianas was chosen as the study area because for more than a century the area has been the focal point of the herbarium of the Utrecht University and a large amount of specimens have been accumulated from the area (Ek 1990; Ek 1991; Hof unpublished). The specimens of this herbarium form the backbone of the herbarium data used in this study. Over the decades specimen identifications were updated by specialist botanists and many duplicate specimens collected by botanists associated with other institutions were deposited in this herbarium. The specimen data of the Utrecht herbarium were supplemented with those of other herbaria from the Flora of the Guianas project and species lists of botanists who collected in the area. This is the most comprehensive and updated source of angiosperm data available for the Guianas.

The herbarium database – rich in species

Botanists have been collecting specimens in the Guianas for more than four centuries (Ek, 1990; Ek 1991; Hof *unpublished)* but the physical location of most of the old specimens is unknown. The specimens in the database used in this thesis were collected by 560 botanists between the period 1804 to 2004

(Chapter 2). The database contained 7,146 species and 168,487 specimens with complete label information. The specimens were not distributed equally among families, genera and growth forms and countries. The five most collected families were Fabaceae, Rubiaceae, Melastomataceae, Poaceae and Cyperaceae. The ten most specimen-rich families accounted for 71,101 (about 42%) records and 3045 (43%) species in the database. The largest number of species was collected in Guyana and the smallest number in Suriname. Although about 35% of the species in the database were collected in all three countries, 42.6% were collected in only one country. Only a few species were represented by a large number of specimens but about 38% of the species were represented by less than five records.

Between 1804 and 2004 the geographical area in which the specimens were collected gradually expanded although some areas such as those close to research stations and cities were revisited several times (Chapter 2). When all of the collecting localities were aggregated into grid cells of 5 X 5 arc-minutes spatial resolution (about 10 X 10 km) only about 28% of the grid cells had collections (Chapter 6). Towards the end of the period of collecting the rate of addition of new species to the herbarium was reduced to 1.4 for every 100 specimens collected, even though specimens were collected in new areas (Chapter 2). The fact that the rate of discovery of new species has slowed down considerably suggests that most of the (regionally) common species in the Guianas have been collected.

A herbarium database – rich in biases

One of the major concerns raised in biodiversity studies based on herbarium databases is the existence of biases associated with collecting the specimens (Soberon *et al.* 2000; Reddy & Davalos 2004; Graham 2004). In this study, the extents of the historical, geographical, taxonomic and seasonal bias were examined. It was found that the number of specimens collected almost always determines the number of species found (Chapter 3). The herbarium database showed historical bias in collecting. As a result more specimens and subsequently more species were collected when different (incremental) time periods, were examined. The consequence of this bias is that when species richness is estimated, using data of different (incremental) time periods, different richness estimates are obtained (Chapter 3 and 4). This is because

species accumulation curves do not attain an asymptote. However, many of the species richness estimates used, e.g. the Michaelis-Menten model, assume asymptotic behaviour.

Using a combination of the Michaelis-Menten and the Arrhenius models, a total of about 12000 angiosperm species were estimated to occur in the Guianas (Chapter 3). The species not yet collected may be rare in nature or might have very restricted ranges perhaps in areas not yet visited. Abundance distributions suggested by Hubbell (2001; Hubbell *et al.* 2007) predict that many species will be very rare in nature. The chance of finding them, with 'ad-hoc collecting expeditions' or systematic sampling (too time consuming) is small. We must accept the fact that many of these rare species will indeed never be collected.

Botanists showed a strong geographical bias towards collecting close to rivers and roads and as a result more specimens and therefore more species were collected along rivers and roads (Chapter 3). However, from a comparison with the environmental conditions of rest of the Guianas it was shown that the collection localities were not environmentally biased. Since the collecting effort represents the environmental conditions well, it is expected that this geographical bias should have no implications when the species data is used in species distribution modeling (SDM) to predict species richness patterns.

More specimens and species were collected during the drier months of the year. The collecting effort almost entirely explained the pattern of annual flowering in the herbarium data (Chapter 3). This suggests that the use of phenology data from the herbarium is problematic. Still the flowering data from the Guianian herbarium specimens show a strong relationship with those collected from independent autecological records. This is perhaps because the collectors used prior knowledge of flowering in planning their expeditions, thereby increasing the possibility of collecting species in flower. For fruiting records this is less the case. Perhaps the fact that fruiting takes place in the wet season (and flowering in the dry) plays a role here. Nevertheless, phenology studies based on herbarium data should not ignore the potential implications of seasonal bias in collecting effort on the accuracy of such studies.

The biases associated with herbarium data collected in the Guianas makes the database suitable for some but not all biodiversity studies. The data is suitable for SDMs and for estimating species richness. Seasonal bias in collecting effort is important if the data is used for phenology study. If this bias is not corrected for,

the phenological data will reflect collecting effort during drier months more than actual phenology patterns.

Never the same species twice - Modeling botanists' collecting strategies

The sheer number of herbarium specimens makes it impossible to ignore this data source when answering the fundamental question: what determines species diversity (i.e. species richness)? The problem with using herbarium data to answer this question is that the dominance diversity curves based on species in the herbarium do not represent that of the community structure well, due to the non-random strategy of collecting. The dominance diversity curve based on herbarium data is flatter than that of plots, indicating that more species are represented by fewer specimens in the herbarium than in the field (Chapter 4). The relative abundance of species in a given area, represented in the herbarium depends among other factors, on the number of botanists visiting the area and how much time was spent collecting. The non-random strategy of collecting makes it unrealistic to apply statistical tools that assume random sampling. A model that explains how the species relative abundance in herbaria develops in a non-random but predictable fashion from the log-series and based on plots data is presented in Chapter 4. The model consisted of two parts. The first part used plots data (from the Mabura Hill area in Guyana and the bauxite mountain region of North Eastern Suriname) to determine the relative abundance distribution (RAD) of all species in an area and used a zero-sum multinomial (ZSM) distribution to describe the community structure of the area. In the second part, the results of the first part were used to simulate collecting behavior based on herbarium data for the two areas, using four different collecting scenarios (Chapter 4).

The main strategy of "never collect the same species twice" generated species abundance distribution patterns comparable to those in the herbarium (Chapter 4). The model predicted even better if some factor of anxiety was built in, reproducing the relative abundance of common species in the herbarium. The long tail of the herbarium was reproduced when the botanists were modeled to collect from different areas with different species, simulating the visiting of different habitats. In another scenario botanists using the strategy 'never collect the same species twice' sampling from the ZSM of the area sampled the same

number of specimens that they collected from the area in reality. The dominance diversity curve resulting from this strategy was very similar to that of the herbarium.
Although it was possible to model the RAD of the herbarium based on species distribution in the field, it was not possible to do the opposite, because of the large number of rare species in the herbarium.
Sampling the ZSM of the area suggests that the resulting species accumulation curve does not reach an asymptote. This is because in the beginning there is a rapid increase in the rate of sampling of new species as the most common species are sampled first. However as sampling continues, the rare species continue to be sampled at a constant rate, resulting in a positive slope in the species accumulation curve. This suggests that the Michaelis-Menten model is a fundamentally wrong model to estimate species richness and that models that estimate species richness based on species accumulation curves must address these weaknesses.

Floristic similarity across the Guianas – a role for distance and ecology

In this thesis for the first time herbarium data are used to test the predictions of neutral and niche theory in explaining similarity across a landscape. Neutral theory predicts that floristic similarity decreases (or decays) between sites in a community as a function of geographical distance (Hubbell 2001). Niche theory predicts that floristic composition between sites varies with environmental conditions as a result of species-specific adaptations to the environment (Hubbell 2001; Tilman 1982). Herbarium data have an advantage over plot data because they cover large spatial scales and are more comprehensive, covering all growth forms. When assessing the variation in species compositional similarity among sites, it is assumed that the species information was collected through random sampling and this is not the case with herbarium data (Chapters 3 and 4). The relative abundance of species in herbarium data is not a good representation of that in the field (Chapters 3 and 4). However, as botanists spend time finding more species rather than measuring more individuals (of the same species) they are more effective in finding many species. Therefore in Chapter 5, presence/absence data and sites of different collecting intensities were used to: (a) assess to what extent geographical distance and environmental factors

explain floristic composition across the Guianas; (b) quantify the fraction of variation in species composition that can be explained by geographical distance, environmental factors and a combination of these factors. We then examined whether the decrease in similarity of any two sites with distance is slower for better dispersers than poor dispersers, as predicted by neutral theory. The environmental variables used were, altitude, temperature, rainfall and seasonality of precipitation.

Mantel tests showed that floristic distance matrices were strongly correlated with geographical distance, altitude and temperature distance matrices and to a lesser extent to rainfall and seasonality of precipitation distance matrices. The total variation in species composition, explained by geographical distance and/or the environment, depended on the collecting intensity. The amount of variation explained was high, when sites with high collecting effort were compared and low, when sites with low collecting effort were compared. The decrease in similarity of any two sites with distance is slower for better dispersers than poor dispersers. The decrease in similarity between two sites with distance is slower for wind dispersed species than for animal dispersed species. The decrease in similarity between two sites with distance is slower for species with low seed mass than for species with high seed mass. Climbers, epiphytes and herbs showed a slower rate of distance decay in similarity than palms, shrubs and trees. These results suggest that species that are better dispersers are therefore more evenly distributed across the landscape while those that are poor dispersers are more clustered. The predictions of neutral theory better support an explanation of similarity across a landscape than niche theory.

Herbarium databases can be used to fuel the discussions about the relative role of dispersal-limited and the niche-related processes in determining species composition across landscapes. However, it is necessary to use ‘sites’ with high collecting effort. Since the collecting effort determines the number of species found, comparing sites with unequal collecting effort blurs an analysis based on similarity due to sampling effects. Also, presence/absence data must be used, since, as mentioned above, species abundance in the herbarium not a good representation of that in the field.

Species richness and endemism patterns

As mentioned above, despite the centuries of collecting in the Guianas, only about 28% of the area, based on grid cells of 5 X 5 arc-minutes spatial resolution (about 10 X 10 km), has been sampled (Chapter 6). Since collection effort determines the number of species found (Chapter 3), it would not be appropriate to use species data from the herbarium to describe species richness (Chapter 6) To fill gaps in poor survey effort, species distribution modeling was used to determine potential species richness and endemism patterns based on species occurrence data and soil, bioclimatic and altitudinal variables (Chapter 6). The species richness map based on about 41% of all the species in the herbarium showed that the coastal zone, where sampling effort was high, was predicted to have the highest species richness of the Guianas. The *arrondissement* of Cayenne in French Guiana, where sampling effort was high, showed higher richness than the rest of the Guianas. The south-eastern part of Guyana and the southern part of Suriname, where sampling effort was low, were predicted to be species poor.

This pattern does not meet our expectations and we believe it does not describe the pattern of diversity well. The unexpectedly high species richness in the coastal zone is probably due to the high collection intensity in this area in combination with the requirements of the model. Most of the species (72%) that were used in the SDMs occurred in this area. Secondly, the environmental variation along the coast is very little but still very distinct from the rest of the Guianas. Therefore once a species collected in this area was successfully modeled there was a high chance that it was predicted to occur in most of the grid cells along the coast. In contrast, in regions characterized by more variable environments, species stand a lower chance of being predicted to occur throughout the region. The *arrondissement* of Cayenne in French Guiana, was also predicted to show high species richness, possibly because of high sampling effort and low variation in environmental variables. The rest of the Guianas was predicted to be poor. This could be because the area was poorly sampled, causing many of the species that occur in these areas to fail to meet the sampling requirements for modeling, which was 5 specimens. This reduced the predicted species richness. There is relatively high variation in the environmental variables in the interior of the Guianas and therefore even species that were successfully modeled might only be predicted to occur in a limited number of

grid cells. The south-eastern part of Guyana and the south of Suriname were predicted to have low species richness. The combination of under-sampling and deviating environmental conditions results in a very low proportion of the regional species pool to be included in the model in these areas. Alternatively, these areas might be simply species poor and this is supported by the fact that the herbarium data for Inga and Lecythidaceae from the Neotropics support the richness patterns predicted with the Guianas data well.

When areas across the Guianas (based mainly on country boundaries and forest region (ter Steege and Zondervan 2000)) were arranged using Detrended correspondence analysis (DCA), the Pakaraima mountain area was shown to be very distinct and there was a strong separation of the sites in the north and those in the south (Fig. 1). This comparison among sites was made under the assumption that common species determine the diversity patterns (Lennon *et al.* 2004) and that for a given area the most common species have been collected (Chapter 4). The DCA results support the view that the south and the north (which includes the coast) of the Guianas are floristically very different from each other. The high species turnover across the Guianas results in the overall high species richness.

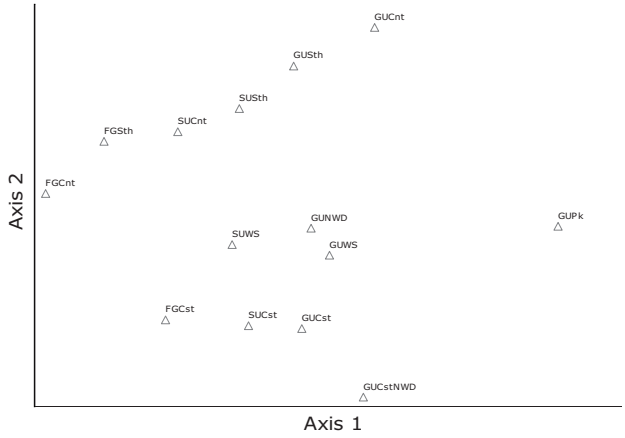

Fig. 1. Detrended Correspondence Analysis ordination of the most common species found in the forest regions of the Guianas.

The highest endemism were predicted in the Pakaraima Mountain, Kanuku Mountain areas and the coast of Guyana, and in the Sipaliwini area and the coast of Suriname (Chapter 6). The Tafelberg mountain area and the *arrondissement* of Cayenne in French Guiana showed intermediate levels of endemism. These predictions are according to expectations (Granville 1988; ter Steege *et al.* 2000; Conservation International 2000) except that ter Steege *et al.* (2000) predicted high levels of endemism in central Guyana. Based on the modeled species and endemism richness patterns and the DCA analysis, it would appear that the Pakaraima mountain area is floristically different from the rest of the Guianas (Fanshawe 1952; Granville 1988; Berry *et al.* 1995).

References

Araujo, M. B., and A. Guisan. 2006. Five (or so) challenges for species distribution modeling. Journal of Biogeography 33:1677-1688.

Beale, C. M., J. J. Lennon, and A. Gimona. 2008. Opening the climate envelope reveals no macroscale associations with climate in European birds. Proceedings of the National Academy of Sciences of the United States of America 105:14908-14912.

Beerling, D. J., and F. E. Mayle. 2006. Contrasting effects of climate and CO2 on Amazonian ecosystems since the last glacial maximum. Global Change Biology 12:1977-1984.

Berry, P. E., O. Huber, and H. B.K. 1995. Floristic analysis and phytogeography. In: P. E. Berry, H. B.K., and K. Yatskievych, editors. Flora of the Venezuelan Guayana, pp 161–191. Missouri Botanical Garden Press, St. Louis.

Borcard, D., P. Legendre, and P. Drapeau. 1992. Partialling out the Spatial Component of Ecological Variation. Ecology 73:1045-1055.

Buermann, W., S. Saatchi, T. B. Smith, B. R. Zutta, J. A. Chaves, B. Mila, and C. H. Graham. 2008. Predicting species distributions across the Amazonian and Andean regions using remote sensing data. Journal of Biogeography 35:1160-1176.

Chave, J., H. C. Muller-Landau, T. R. Baker, T. A. Easdale, H. Ter Steege, and C. O. Webb. 2006. Regional and phylogenetic variation of wood density across 2456 neotropical tree species. Ecological Applications 16:2356-2367.

Chust, G., J. Chave, R. Condit, S. Aguilar, S. Lao, and R. Perez. 2006. Determinants and spatial modeling of tree beta-diversity in a tropical forest landscape in Panama. Journal of Vegetation Science 17:83-92.

CIA. 2001. The World Factbook. in.

Clarke, H. D., and V. A. Funk 1998. A preliminary analysis of the plant diversity of the Iwokrama Forest, Guyana. Smithsonian Institution, Washington, DC.

Clarke, H. D., and V. A. Funk. 2005. Using checklists and collections data to investigate plant diversity: II. An analysis of five florulas from northeastern South America. Proceedings of the Academy of Natural Sciences of Philadelphia 154:29-37.

Colwell, R. K., and J. A. Coddington. 1994. Estimating Terrestrial Biodiversity through Extrapolation. Philosophical Transactions of the Royal Society of London Series B-Biological Sciences 345:101-118.

Condit, R., N. Pitman, E. G. Leigh, J. Chave, J. Terborgh, R. B. Foster, P. Nunez, S. Aguilar, R. Valencia, G. Villa, H. C. Muller-Landau, E. Losos, and S. P. Hubbell. 2002. Beta-diversity in tropical forest trees. Science 295:666-669.

Conservation International 2003. Conservation priorities for the Guiana Shield. Conservation International, Washington, DC.

Costa, G. C., C. Nogueira, R. B. Machado, and G. R. Colli. 2007. Squamate richness in the Brazilian Cerrado and its environmental-climatic associations. Diversity and Distributions 13:714-724.

Crisp, M. D., S. Laffan, H. P. Linder, and A. Monro. 2001. Endemism in the **Australian flora. Journal of Biogeography 28:183-198.**

Dray, S., P. Legendre, and P. R. Peres-Neto. 2006. Spatial modeling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). Ecological Modeling 196:483-493.

Duivenvoorden, J. F., J. C. Svenning, and S. J. Wright. 2002. Ecology - Beta diversity in tropical forests. Science 295:636-637.

Duque, A., M. Sanchez, J. Cavelier, and J. F. Duivenvoorden. 2002. Different floristic patterns of woody understorey and canopy plants in Colombian Amazonia. Journal of Tropical Ecology 18:499-525.

Ek, R. C. 1990. Index of Guyana plant collectors. Koeltz Scientific Books, Koenigstein.

Ek, R. C. 1991. Index of Suriname plant collectors. Koeltz Scientific Books, Koenigstein.

Ek, R. C., and H. Ter Steege. 1998. The Flora of the Mabura Hill Area, Guyana. Botanische Jahrbücher für Systematik Pflanzengeschichte und Pflanzengeographie 120:461-502.

Elith, J., C. H. Graham, R. P. Anderson, M. Dudik, S. Ferrier, A. Guisan, R. J. Hijmans, F. Huettmann, J. R. Leathwick, A. Lehmann, J. Li,

L. G. Lohmann, B. A. Loiselle, G. Manion, C. Moritz, M. Nakamura, Y. Nakazawa, J. M. Overton, A. T. Peterson, S. J. Phillips, K. Richardson, R. Scachetti-Pereira, R. E. Schapire, J. Soberon, S. Williams, M. S. Wisz, and N. E. Zimmermann. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29:129-151.

Eriksson, O., and A. Jakobsson. 1998. Abundance, distribution and life histories of grassland plants: a comparative study of 81 species. Journal of Ecology 86:922-933.

Erkens, R. H. J., and P. Baas. 2008. Utrecht: rise and fall of a great herbarium. Taxon 57:1024-1026.

Fanshawe, D. B. 1952. The Vegetation of British Guyana. A Preliminary Review. Imperial Forestry Institute, Oxford.

FAO. 2002. Terrastat; global land resources GIS models and databases for poverty and food insecurity mapping in Land and Water Digital Media.

FAO. 2006. Global Forest Resource Assessment 2005. in FAO Forestry Paper.

Ferrier, S., G. Watson, J. Pearce, and M. Drielsma. 2002. Extended statistical approaches to modeling spatial pattern in biodiversity in northeast New South Wales. I. Species-level modeling. Biodiversity and Conservation 11:2275-2307.

Fielding, A. H., and J. F. Bell. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation 24:38-49.

Fisher, R. A., A. S. Corbet, and C. B. Williams. 1943. The relation between the number of species and the number of individuals in a random sample of an animal population. Journal of Animal Ecology 12:42-58.

Flather, C. H. 1996. Fitting species-accumulation functions and assessing regional land use impacts on avian diversity. Journal of Biogeography 23:155-168.

Forsyth, A., and K. Miyata. 1984. Tropical Nature. Macmillan Publishing Company, New York.

Funk , V. A., T. Hollowell, P. E. Berry, C. Kelloff, and S. N. Alexander. 2007. Checklist of the Plants of the Guiana Shield (Venezuela: Amazonas, Bolivar, Delta Amacuro; Guyana, Surinam, French Guiana) Contributions from the United States National Herbarium, Washington, DC.

Funk, V. A., and K. S. Richardson. 2002. Systematic data in biodiversity studies: Use it or lose it. Systematic Biology 51:303-316.

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

Funk, V. A., M. F. Zermoglio, and N. Nasir. 1999. Testing the use of specimen collection data and GIS in biodiversity exploration and conservation decision making in Guyana. Biodiversity and Conservation 8:727-751.

Gotelli, N. J., and R. K. Colwell. 2001. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. Ecology Letters 4:379-391.

Gotelli, N. J., and G. L. Entsminger. 2000. EcoSim: Null models software for ecology. in. Acquired Intelligence Inc. & Kesey-Bear.

Graham, C. H., J. Elith, R. J. Hijmans, A. Guisan, A. T. Peterson, and B. A. Loiselle. 2008. The influence of spatial errors in species occurrence data used in distribution models. Journal of Applied Ecology 45:239-247.

Graham, C. H., S. Ferrier, F. Huettman, C. Moritz, and A. T. Peterson. 2004. New developments in museum-based informatics and applications in biodiversity analysis. Trends in Ecology & Evolution 19:497-503.

Granville, J.-J. d. 1988. Phytogeographical characteristics of the Guianan forests. Taxon 37:578–594.

Guisan, A., and W. Thuiller. 2007. Predicting species distribution: offering more than simple habitat models (vol 8, pg 993, 2005). Ecology Letters 10:435-435.

Guralnick, R., and A. Hill. 2009. Biodiversity informatics: automated approaches for documenting global biodiversity patterns and processes. Bioinformatics 25:421-428.

Hammond, D. S. 2005. Tropical forests of the Guiana Shiled. Ancient forests in modern world. CABI Publishing, Oxfordshire.

Hammond, D. S., and V. K. Brown. 1995. Seed Size of Woody-Plants in Relation to Disturbance, Dispersal, Soil Type in Wet Neotropical Forests. Ecology 76:2544-2561.

Hammond, D. S., S. Gourlet-Fleury, P. vanderHout, H. terSteege, and V. K. Brown. 1996. A compilation of known Guianan timber trees and the

significance of their dispersal mode, seed size and taxonomic affinity to tropical rain forest management. Forest Ecology and Management 83:99-116.

Haugaasen, T., and C. A. Peres. 2005. Tree phenology in adjacent Amazonian flooded and unflooded forests. Biotropica 37:620-630.

Hernandez, P. A., C. H. Graham, L. L. Master, and D. L. Albert. 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography 29:773-785.

Hijmans, R. J., S. E. Cameron, J. L. Parra, P. G. Jones, and A. Jarvis. 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25:1965-1978.

Hijmans, R. J., K. A. Garrett, Z. Huaman, D. P. Zhang, M. Schreuder, and M. Bonierbale. 2000. Assessing the geographic representativeness of genebank collections: the case of Bolivian wild potatoes. Conservation Biology 14:1755-1765.

Hoff, M. 1996. Study of a flora: Who has collected what in French Guiana since Aublet? . Acta Botanica Gallica 143:199-218.

Hoff, M., G. Cremers, and J. F. Brulard. 2002. Botanical prospections in French Guiana; historical geography of the floristic exploration. Acta Botanica Gallica 14:245-274.

Hollowell, T., V. A. Funk , C. Kelloff, and G. Gharbarran. 2000. Smithsonian Plant Collections, Guyana: 1986 - 1987, John J. Pipoly III. Biological Diversity of the Guianas Program, Smithsonian Institution, Washington, DC.

Hollowell, T., L. J. Gillespie, V. A. Funk , and C. Kelloff. 2003. Smithsonian plant collections, Guyana : 1989-1991, Lynn J. Gillespie. Contributions from the United States National Herbarium, Washington, DC.

Hollowell, T., T. McDowell, V. A. Funk , C. Kelloff, and D. Gopaul. 2004. Smithsonian Plant Collections, Guyana: 1990 - 1991, Tim McDowell. Contributions from the United States National Herbarium, Washington, DC.

Hopkins, M. J. G. 2007. Modeling the known and unknown plant biodiversity of the Amazon Basin. Journal of Biogeography 34:1400-1411.

Hortal, J., A. Jimenez-Valverde, J. F. Gomez, J. M. Lobo, and A. Baselga. 2008. Historical bias in biodiversity inventories affects the observed environmental niche of the species. Oikos 117:847-858.

Hortal, J., J. M. Lobo, and A. Jimenez-Valverde. 2007. Limitations of biodiversity databases: Case study on seed-plant diversity in Tenerife, Canary Islands. Conservation Biology 21:853-863.

Hubbell, S. P. 2001. The Unified Neutral Theory of Biodiversity and Biogeography. Princeton University Press, Princeton, NJ.

Hubbell, S. P., F. L. He, R. Condit, L. Borda-de-Agua, J. Kellner, and H. ter Steege. 2008. How many tree species and how many of them are there in the Amazon will go extinct? Proceedings of the National Academy of Sciences of the United States of America 105:11498-11504.

Huber, O., G. Gharbarran, and V. A. Funk 1995. Preliminary Vegetation Map of Guyana. in. Biological Diversity of the Guianas Program, Smithsonian Institution, Washington, DC.

Janzen, D. H. 1969. Seed eaters vs. seed size, number, toxicity and dispersal. Evolution 23:1-27.

Jarvis, A., M. E. Ferguson, D. E. Williams, L. Guarino, P. G. Jones, H. T. Stalker, J. F. M. Valls, R. N. Pittman, C. E. Simpson, and P. Bramel. 2003. Biogeography of wild Arachis: Assessing conservation status and setting future priorities. Crop Science 43:1100-1108.

Jones, M. M., H. Tuomisto, D. Borcard, P. Legendre, D. B. Clark, and P. C. Olivas. 2008. Explaining variation in tropical plant community composition: influence of environmental and spatial data quality. Oecologia 155:593-604.

Jones, M. M., H. Tuomisto, D. B. Clark, and P. Olivas. 2006. Effects of mesoscale environmental heterogeneity and dispersal limitation on floristic variation in rain forest ferns. Journal of Ecology 94:181-195.

Jurasinski, G., V. Retzer, and C. Beierkuhnlein. 2009. Inventory, differentiation, and proportional diversity: a consistent terminology for quantifying species diversity. Oecologia 159:15-26.

Kadmon, R., O. Farber, and A. Danin. 2004. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. Ecological Applications 14:401-413.

Keating, K. A., and J. F. Quinn. 1998. Estimating species richness: the Michaelis-Menten model revisited. Oikos 81:411-416.

Koellner, T., A. M. Hersperger, and T. Wohlgemuth. 2004. Rarefaction method for assessing plant species diversity on a regional scale. Ecography 27:532-544.

Kuper, W., J. H. Sommer, J. C. Lovett, and W. Barthlott. 2006. Deficiency in African plant distribution data - missing pieces of the puzzle. Botanical Journal of the Linnean Society 150:355-368.

Lead, S. 2005. Generate randomly-distributed points. in.

Legendre, L., and P. Legendre. 1998. Numerical ecology. Elsevier, Amsterdam.

Legendre, P. 2008. Studying beta diversity: ecological variation partitioning by multiple regression and canonical analysis. Journal of Plant Ecology-Uk 1:3-8.

Legendre, P., D. Borcard, and P. R. Peres-Neto. 2005. Analyzing beta diversity: Partitioning the spatial variation of community composition data. Ecological Monographs 75:435-450.

Legendre, P., and M. J. Fortin. 1989. Spatial Pattern and Ecological Analysis. Vegetatio 80:107-138.

Legendre, P., and L. Legendre. 1998. Numerical ecology. Elsevier, Amsterdam.

Leibold, M. A., M. Holyoak, N. Mouquet, P. Amarasekare, J. M. Chase, M. F. Hoopes, R. D. Holt, J. B. Shurin, R. Law, D. Tilman, M. Loreau, and A. Gonzalez. 2004. The metacommunity concept: a framework for multi-scale community ecology. Ecology Letters 7:601-613.

Lennon, J. J., P. Koleff, J. J. D. Greenwood, and K. J. Gaston. 2004. Contribution of rarity and commonness to patterns of species richness. Ecology Letters 7:81-87.

Leonard, R., P. Legendre, M. Jean, and A. Bouchard. 2008. Using the landscape morphometric context to resolve spatial patterns of submerged macrophyte communities in a fluvial lake. Landscape Ecology 23:91-105.

Lim, B. K., A. T. Peterson, and M. D. Engstrom. 2002. Robustness of ecological niche modeling algorithms for mammals in Guyana. Biodiversity and Conservation 11:1237-1246.

Lindeman, J. C. 1953. The vegetation of the coastal region of Suriname. Kemink, Utrecht.

Lindeman, J. C., and S. P. Moolenaar. 1959. Preliminary survey of the vegetation types of northern Suriname. In d. H. I.A. and L. J., editors. The Vegetation of Suriname. Mededelingen van het Botanisch Museum en Herbarium van de Rijksuniversiteit, Utrecht.

Lobo, J. M., A. Baselga, J. Hortal, A. Jimenez-Valverde, and J. F. Gomez. 2007. How does the knowledge about the spatial distribution of Iberian dung beetle species accumulate over time? Diversity and Distributions 13:772-780.

Lobo, J. M., and F. Martin-Piera. 2002. Searching for a predictive model for species richness of Iberian dung beetle based on spatial and environmental variables. Conservation Biology 16:158-173.

Loiselle, B. A., P. M. Jorgensen, T. Consiglio, I. Jimenez, J. G. Blake, L. G. Lohmann, and O. M. Montiel. 2008. Predicting species distributions from herbarium collections: does climate bias in collection sampling influence model outcomes? Journal of Biogeography 35:105-116.

Mace, G. M. 2004. The role of taxonomy in species conservation. Philosophical Transactions of the Royal Society of London Series B-Biological Sciences 359:711-719.

Magurran, A. E. 2004. Measuring biological diversity. Blackwell Publishing, Oxford.

McGill, B. J., E. A. Hadly, and B. A. Maurer. 2005. Community inertia of quaternary small mammal assemblages in North America. Proceedings of the National Academy of Sciences of the United States of America 102:16701-16706.

McGill, B. J., B. A. Maurer, and M. D. Weiser. 2006. Empirical evaluation of neutral theory. Ecology 87:1411-1423.

Mittermeier, R. A., N. Myers, J. B. Thomsen, G. A. B. da Fonseca, and S. Olivieri. 1998. Biodiversity hotspots and major tropical wilderness areas: Approaches to setting conservation priorities. Conservation Biology 12:516-520.

Mori, S.A. and G. T. Prance: The Lecythidaceae Pages. The New York Botanical Garden, Bronx, New York.

Mori, S. A., and J. L. Brown. 1998. Epizoochorous dispersal by barbs, hooks, and spines in a lowland moist forest in central French Guiana. Brittonia 50:165-173.

Mori, S. A., G. Cremers, C. Gracie, J. J. de Granville, S. V. Heald, and J. D. Mitchell. 2002. Guide to the vascular plants of central French Guiana. Part 2. Dicotyledons. Memoirs of the New York Botanical Garden, NY.

Mori, S. A., G. Cremers, C. Gracie, J. J. d. Granville, M. Hoff, and J. D. Mitchell. 1997. Guide to the vascular plants of central French Guiana. Part 1. Pteridophytes, gymnosperms, and monocotyledons, NY.

Murray-Smith, C., N. A. Brummitt, A. T. Oliveira-Filho, S. Bachman, J. Moat, E. M. N. Lughadha, and E. J. Lucas. 2009. Plant Diversity Hotspots in the Atlantic Coastal Forests of Brazil. Conservation Biology 23:151-163.

Nekola, J. C., and J. H. Brown. 2007. The wealth of species: ecological communities, complex systems and the legacy of Frank Preston. Ecology Letters 10:188-196.

Nekola, J. C., and P. S. White. 1999. The distance decay of similarity in biogeography and ecology. Journal of Biogeography 26:867-878.

Oksanen, J., R. Kindt, P. Legendre, and R. B. O'Hara. 2007. vegan: Community Ecology Package. R package version. in.

Pearman, P. B., and D. Weber. 2008. Common species determine richness patterns in biodiversity indicator taxa: Errata. Biological Conservation 141:5-5.

Pennington, T. D. 1997. The Inga database. Royal Botanical Gardens, Kew

Pennisi, E. 2005. What determines species diversity. Science 309:90-90.

Peres-Neto, P. R., P. Legendre, S. Dray, and D. Borcard. 2006. Variation partitioning of species data matrices: Estimation and comparison of fractions. Ecology 87:2614-2625.

Peterson, A. T., and Y. Nakazawa. 2008. Environmental datasets matter in ecological niche modeling: an example with Solenopsis invicta and Solenopsis richteri. Global Ecology and Biogeography 17:135-144.

Phillips, O. L., R. Vasquez Martinez, P. Nunez Vargas, A. Lorenzo Monteagudo, M. E. Chuspe Zans, W. Galiano Sanchez, A. Pena Cruz, M. Timana, M. Yli-Halla, and S. Rose. 2003. Efficient plot-based floristic assessment of tropical forests. Journal of Tropical Ecology 19:629-645.

Phillips, S. J., R. P. Anderson, and R. E. Schapire. 2006. Maximum entropy modeling of species geographic distributions. Ecological Modeling 190:231-259.

Ponder, W. F., G. A. Carter, P. Flemons, and R. R. Chapman. 2001. Evaluation of museum collection data for use in biodiversity assessment. Conservation Biology 15:648-657.

Potts, M. D., P. S. Ashton, L. S. Kaufman, and J. B. Plotkin. 2002. Habitat patterns in tropical rain forests: A comparison of 105 plots in Northwest Borneo. Ecology 83:2782-2797.

Preston, F. W. 1962. The canonical distribution of commonness and rarity. Ecology 43:185-215.

R Development Core Team, R. D. C. 2007. R:A language and environment for statistical computing. in. R Foundation for Statistical Computing, Vienna.

Raaijmakers, J. G. W. 1987. Statistical-Analysis of the Michaelis-Menten Equation. Biometrics 43:793-803.

Raes, N., M. C. Roos, J. W. F. Slik, E. E. van Loon, and H. ter Steege. 2009. Botanical richness and endemicity patterns of Borneo derived from species distribution models. Ecography 32:180-192.

Raes, N., and H. ter Steege. 2007. A null-model for significance testing of presence-only species distribution models. Ecography 30:727-736.

Reddy, S., and L. M. Davalos. 2003. Geographical sampling bias and its implications for conservation priorities in Africa. Journal of Biogeography 30:1719-1727.

Richards, P. W. 1996. The tropical rain forest. Cambridge University Press, Cambridge, Massachusetts.

Rosenzweig, M. L. 1995. Species diversity in space and time. Cambridge **University Press, Cambridge, Massachusetts.**

Rosenzweig, M. L. 1999. Ecology - Heeding the warning in biodiversity's basic law. Science 284:276-277.

Ruokolainen, K., and H. Tuomisto. 2002. Beta-diversity in tropical forests. Science 297.

Saatchi, S., W. Buermann, H. Ter Steege, S. Mori, and T. B. Smith. 2008. Modeling distribution of Amazonian tree species and diversity using remote sensing measurements. Remote Sensing of Environment 112:2000-2017.

Sabatier, D. 1985. Saisonnalité et déterminisme du pic de fructification en forêt guyanaise. Revue d'Ecologie (Terre et Vie) 40:289-320.

Schmidt, M., H. Kreft, A. Thiombiano, and G. Zizka. 2005. Herbarium collections and field data-based plant diversity maps for Burkina Faso. Diversity and Distributions 11:509-516.

Schulman, L., K. Ruokolainen, L. Junikka, I. E. Saaksjarvi, M. Salo, S. K. Juvonen, J. Salo, and M. Higgins. 2007. Amazonian biodiversity and protected areas: Do they meet? Biodiversity and Conservation 16:3011-3051.

Soberon, J. M., J. B. Llorente, and L. Onate. 2000. The use of specimen-label databases for conservation purposes: an example using Mexican Papilionid and Pierid butterflies. Biodiversity and Conservation 9:1441-1466.

Stevens, P. F. (2001 onwards). Angiosperm Phylogeny Website. Version 7, May 2006.

Svenning, J. C., D. A. Kinner, R. F. Stallard, B. M. J. Engelbrecht, and S. J. Wright. 2004. Ecological determinism in plant community structure across a tropical forest landscape. Ecology 85:2526-2538.

Ter Steege, H., M. J. Jansen-Jacobs, and V. K. Datadin. 2000. Can botanical collections assist in a National Protected Area Strategy in Guyana? Biodiversity and Conservation 9:215-240.

Ter Steege, H., and C. A. Persaud. 1991. The phenology of Guyanese timber species: a compilation of a century of observations. Vegetatio 95:177-198.

Ter Steege, H., N. Pitman, D. Sabatier, H. Castellanos, P. Van der Hout, D. C. Daly, M. Silveira, O. Phillips, R. Vasquez, T. Van Andel, J. Duivenvoorden, A. A. De Oliveira, R. Ek, R. Lilwah, R. Thomas, J. Van Essen, C. Baider, P. Maas, S. Mori, J. Terborgh, P. N. Vargas, H. Mogollon, and W. Morawetz. 2003. A spatial model of tree alpha-diversity and tree density for the Amazon. Biodiversity and Conservation 12:2255-2277.

Ter Steege, H., and R. Zagt. 2002. Ecology - Density and diversity. Nature 417:698-699.

Ter Steege, H., and G. Zondervan. 2000. A preliminary analysis of large-scale forest inventory data of the Guiana Shield. In: H. Ter Steege, editor. Plant diversity in Guyana. With recommendation for a protected areas strategy, pp 35-54. Tropenbos Foundation, Wageningen.

Tilman, D. 1982. Resource Competition and Community Structure. Princeton University Press, Princeton.

Tobler, M., E. Honorio, J. Janovec, and C. Reynel. 2007. Implications of collection patterns of botanical specimens on their usefulness for conservation planning: an example of two neotropical plant families

(Moraceae and Myristicaceae) in Peru. Biodiversity and Conservation 16:659-677.

Toriola, D. 1998. Fruiting of a 19-year old secondary forest in French Guiana. Journal of Tropical Ecology 14:373-379.

Tuomisto, H., K. Ruokolainen, M. Aguilar, and A. Sarmiento. 2003. Floristic patterns along a 43-km long transect in an Amazonian rain forest. Journal of Ecology 91:743-756.

Tuomisto, H., K. Ruokolainen, and M. Yli-Halla. 2003. Dispersal, environment, and floristic variation of western Amazonian forests. Science 299:241-244.

Van Gemerden, B. S., R. S. Etienne, H. Olff, P. Hommel, and F. Van Langevelde. 2005. Reconciling methodologically different biodiversity assessments. Ecological Applications 15:1747-1760.

Van Roosmalen, M. G. M. 1985. Fruits of the Guianan Flora. Institute of Systematic Botany, University of Utrecht, Utrecht.

Vormisto, J., H. Tuomisto, and J. Oksanen. 2004. Palm distribution patterns in Amazonian rainforests: What is the role of topographic variation? Journal of Vegetation Science 15:485-494.

Wisz, M. S., R. J. Hijmans, J. Li, A. T. Peterson, C. H. Graham, and A. Guisan. 2008. Effects of sample size on the performance of species distribution models. Diversity and Distributions 14:763-773.

Zhang, S. Y., and L. X. Wang. 1995. Comparison of 3 Fruit Census Methods in French-Guiana. Journal of Tropical Ecology 11:281-294.

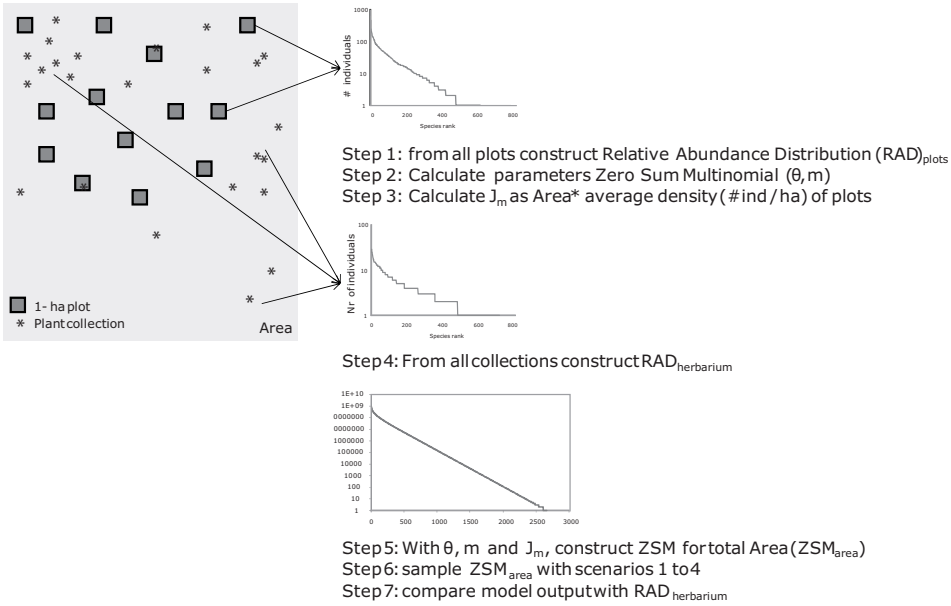

Fig. S4.1. Schematic representation of the model to simulate collector's behaviour.

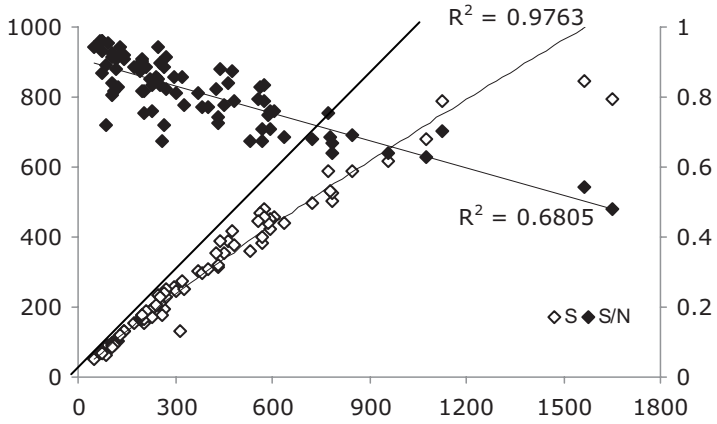

Fig. S4.2. Number of specimens (x-axis) and species (left y-axis) collected on big expeditions in the Guianas from 1953 to 2004 (straight line: S=N). The efficiency to collect each species only once (S/N ratio – right y-axis)) decreases with the size of the collection made (x-axis), (one outlier left out, Jansen-Jacobs 2003).

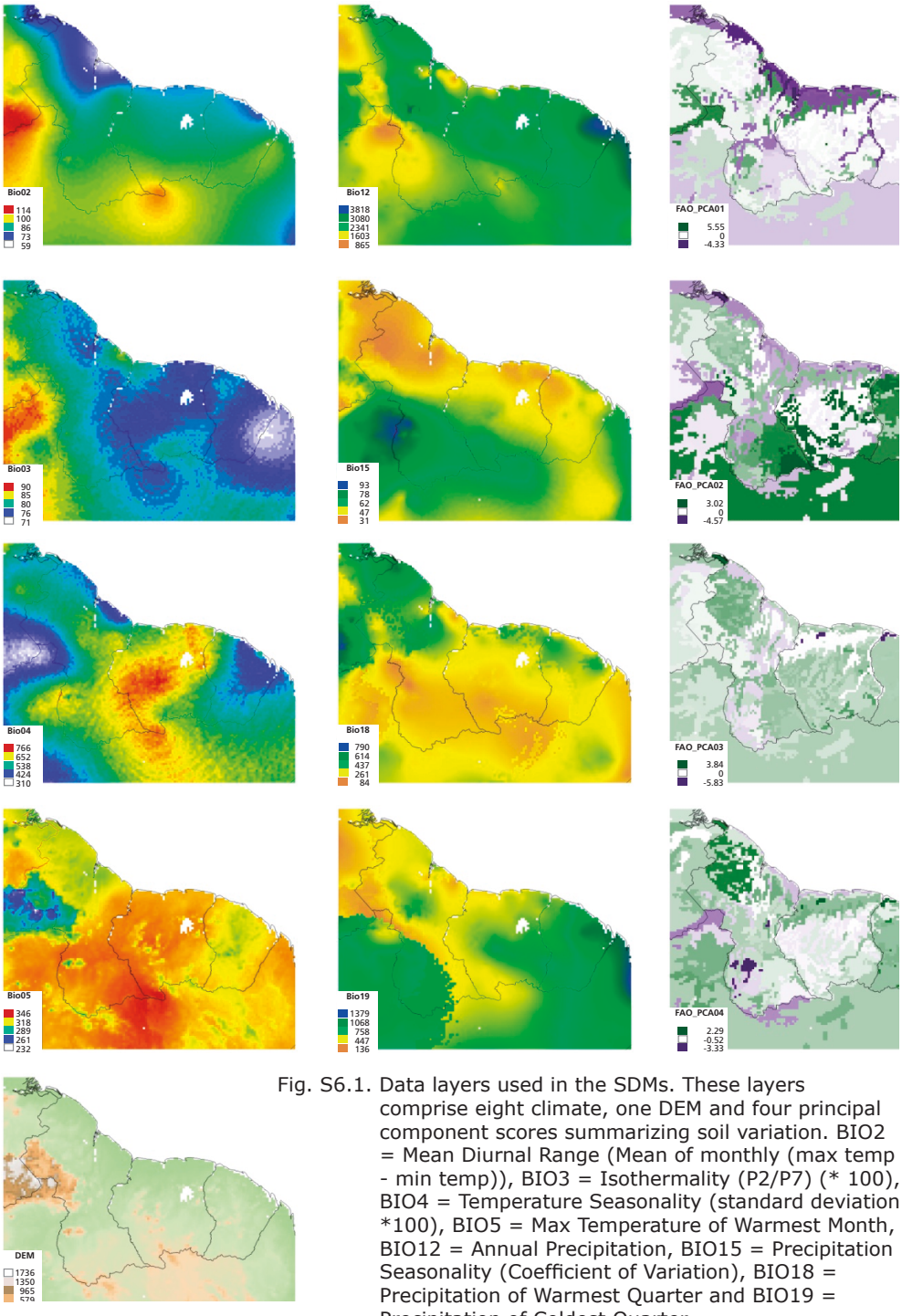

Fig. S6.1. Data layers used in the SDMs. These layers comprise eight climate, one DEM and four principal component scores summarizing soil variation. BIO2 = Mean Diurnal Range (Mean of monthly (max temp - min temp)), BIO3 = Isothermality (P2/P7) (* 100), BIO4 = Temperature Seasonality (standard deviation *100), BIO5 = Max Temperature of Warmest Month, BIO12 = Annual Precipitation, BIO15 = Precipitation Seasonality (Coefficient of Variation), BIO18 = Precipitation of Warmest Quarter and BIO19 = Precipitation of Coldest Quarter.

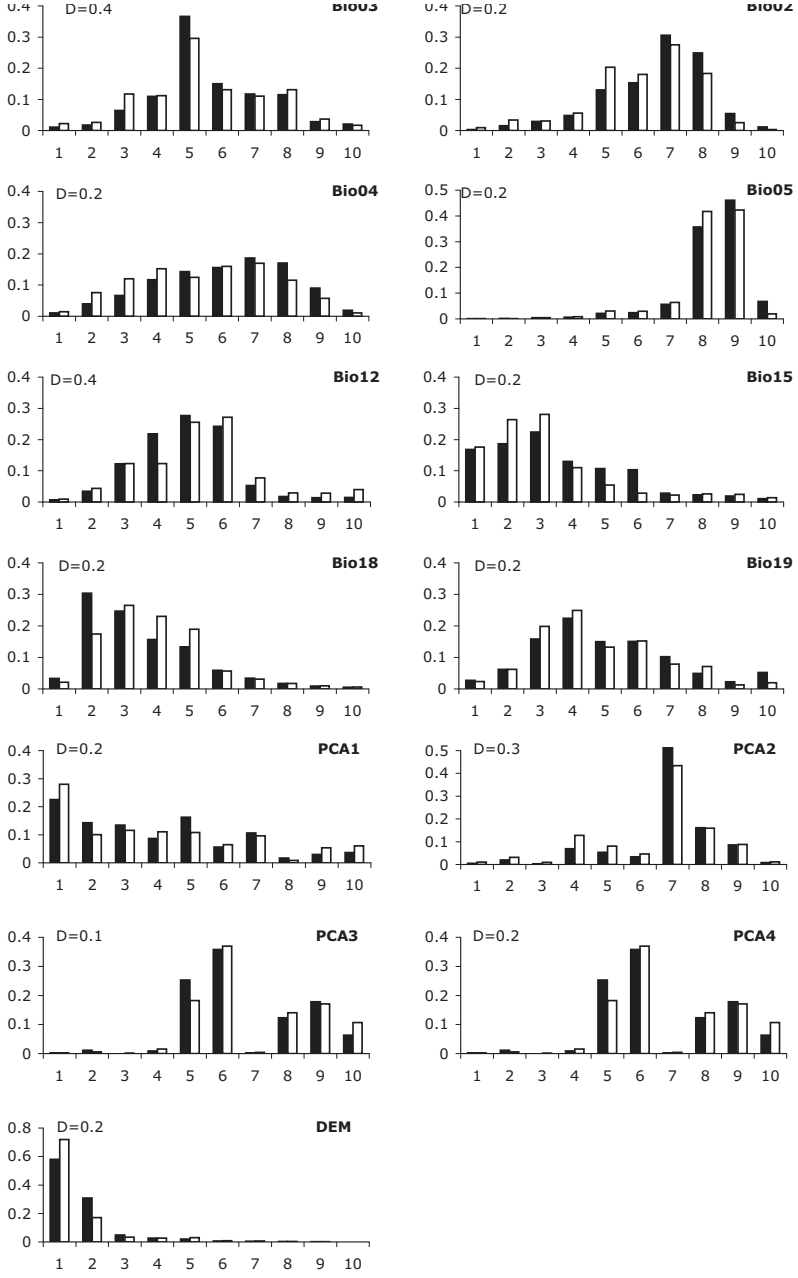

Fig. S6.2. Examining whether sampling effort is random with respect to the environmental conditions in the Guianas. Each of the 13 environmental variables was divided into 10 equal-interval bins. The difference between the frequency distribution between the observed and expected grid cells were tested using the Kolmogorov-Smirnov test. No differences were found between the observed and expected distributions ($P<0.001$ in all cases).

Appendix 2.1 The number of genera, species and specimens collected per family in the Guianas.

	Family	w		
A	Acanthaceae	19	70	1230
	Achariaceae	3	8	260
	Adoxaceae	1	1	2
	Agavaceae	1	1	18
	Aizoaceae	1	1	41
	Alismataceae	2	10	109
	Amaranthaceae	6	13	284
	Amaryllidaceae	8	9	203
	Anacardiaceae	8	21	841
	Anisophylleaceae	1	1	2
	Annonaceae	20	145	4658
	Apiaceae	2	3	73
	Apocynaceae	47	194	4196
	Aquifoliaceae	1	14	231
	Araceae	22	142	3010
	Araliaceae	2	10	173
	Arecaceae	22	80	2123
	Aristolochiaceae	1	18	130
	Asteraceae	76	147	2913
	Avicenniaceae	1	2	94
B	Balanophoraceae	1	1	95
	Bataceae	1	1	26
	Begoniaceae	1	10	496
	Bignoniaceae	37	128	2399
	Bixaceae	1	1	59
	Bonnetiaceae	2	6	73
	Boraginaceae	4	47	1143
	Brassicaceae	3	14	214
	Bromeliaceae	19	110	1570
	Burmanniaceae	9	20	451
	Burseraceae	6	57	2138
C	Cactaceae	7	13	177
	Campanulaceae	4	7	190
	Canellaceae	1	1	9
	Cannaceae	1	1	28
	Cardiopteridaceae	1	1	120
	Caricaceae	2	2	45
	Caryocaraceae	2	8	267
	Caryophyllaceae	2	2	47
	Cecropiaceae	3	30	513
	Celastraceae	12	55	1085
	Celtidaceae	4	5	179
	Chloranthaceae	1	1	1
	Chrysobalanaceae	7	122	3578
	Clusiaceae	18	121	3292
	Cochlospermaceae	1	2	49
	Combretaceae	6	26	815
	Commelinaceae	11	19	409
	Connaraceae	4	24	421
	Convolvulaceae	14	73	1282
	Costaceae	2	14	592
	Crassulaceae	1	1	18
	Cucurbitaceae	11	34	630
	Cunoniaceae	1	4	11
	Cyclanthaceae	8	18	489

	Cyperaceae	34	255	6430
	Cyrillaceae	1	1	53
D	Dichapetalaceae	2	7	412
	Dilleniaceae	5	20	445
	Dioscoreaceae	1	22	169
	Dipterocarpaceae	1	1	10
	Droseraceae	1	7	112
E	Ebenaceae	2	17	267
	Elaeocarpaceae	1	26	342
	Eremolepidaceae	1	1	1
	Ericaceae	12	26	326
	Eriocaulaceae	7	41	646
	Erythroxylaceae	1	24	448
	Euphorbiaceae	48	160	3942
	Euphroniaceae	1	2	4
F	Fabaceae	117	642	15310
G	Gelsemiaceae	1	1	5
	Gentianaceae	14	36	1838
	Gesneriaceae	19	54	1448
	Goupiaceae	1	1	251
H	Haemodoraceae	2	2	149
	Heliconiaceae	1	16	904
	Hernandiaceae	2	5	73
	Humiriaceae	6	19	552
	Hydrocharitaceae	2	2	42
	Hydroleaceae	1	2	60
I	Icacinaceae	5	10	248
	Iridaceae	5	6	60
	Ixonanthaceae	2	4	16
K	Krameriaceae	1	2	8
L	Lacistemataceae	1	4	236
	Lamiaceae	7	23	602
	Lauraceae	17	136	2851
	Lecythidaceae	7	58	3334
	Lentibulariaceae	2	39	856
	Limnocharitaceae	1	1	20
	Linaceae	2	7	112
	Loganiaceae	5	32	691
	Loranthaceae	7	38	602
	Lythraceae	6	12	163
M	Magnoliaceae	1	1	1
	Malpighiaceae	21	113	2326
	Malvaceae	41	149	3362
	Marantaceae	11	66	1976
	Marcgraviaceae	5	10	464
	Mayacaceae	1	3	67
	Melastomataceae	40	287	7896
	Meliaceae	7	39	1533
	Menispermaceae	13	35	417
	Menyanthaceae	1	1	78
	Molluginaceae	3	3	44
	Monimiaceae	1	3	65
	Moraceae	15	70	1996
	Myristicaceae	4	19	834
	Myrsinaceae	4	41	826
	Myrtaceae	15	169	3240
N	Najadaceae	1	1	2
	Nartheciaceae	1	2	11
	Nyctaginaceae	5	15	411
	Nymphaceae	1	1	2
	Nymphaeaceae	3	10	249
O	Ochnaceae	1	1	1
	Olacaceae	9	16	654
	Onagraceae	1	20	838
	Opiliaceae	1	2	30

	Orchidaceae	119	505	4317
	Oxalidaceae	3	8	128
P	Passifloraceae	4	58	1283
	Pedaliaceae	2	2	4
	Phyllanthaceae	7	28	520
	Phytolaccaceae	6	11	220
	Picramniaceae	1	5	81
	Piperaceae	2	96	4163
	Poaceae	81	304	7155
	Podostemaceae	8	36	307
	Polygalaceae	6	48	1132
	Polygonaceae	5	27	672
	Pontederiaceae	3	8	136
	Portulacaceae	2	6	96
	Primulaceae	1	1	1
	Proteaceae	3	6	150
	Putranjivaceae	1	2	175
Q	Quiinaceae	3	18	425
R	Rafflesiaceae	1	2	12
	Rapateaceae	9	22	346
	Rhabdodendraceae	1	1	81
	Rhamnaceae	5	11	141
	Rhizophoraceae	3	8	308
	Rosaceae	2	3	30
	Rubiaceae	76	457	13034
	Ruppiaceae	1	1	9
	Rutaceae	16	39	425
S	Sabiaceae	1	3	15
	Salicaceae	10	37	1700
	Santalaceae	1	1	1
	Sapindaceae	16	117	2201
	Sapotaceae	11	112	3000
	Sarraceniaceae	1	1	8
	Schlegeliaceae	1	6	203
	Scrophulariaceae	17	41	781
	Simaroubaceae	4	12	303
	Siparunaceae	1	8	600
	Smilacaceae	1	19	302
	Solanaceae	13	77	2287
	Strelitziaceae	2	2	90
	Styracaceae	1	8	32
	Symplocaceae	1	4	120
T	Taccaceae	1	1	8
	Ternstroemiaceae	1	11	99
	Theaceae	1	1	28
	Theophrastaceae	1	3	142
	Thurniaceae	1	2	83
	Thymelaeaceae	3	7	38
	Trigoniaceae	1	8	156
	Triuridaceae	4	6	95
	Turneraceae	3	17	405
	Typhaceae	1	1	9
U	Ulmaceae	1	1	1
	Urticaceae	4	11	179
V	Velloziaceae	1	2	6
	Verbenaceae	14	47	1461
	Violaceae	10	33	1891
	Viscaceae	2	34	405
	Vitaceae	1	11	462
	Vochysiaceae	5	36	751
W	Winteraceae	1	2	2
X	Xyridaceae	3	35	756
Z	Zingiberaceae	1	7	344
	Zygophyllaceae	1	1	4

Appendix 3.1 The multinomial probability results indicating that botanists showed a preference for collecting the one or more plane families relative to other families when the botanists' lists were compared to the rest of the herbarium. A value greater that 994 indicates a bias towards a particular family. The first table contains six botanists and the second table contains the other seven.

	Family	Clarke, H.D.	G. Cremers	Donselaar, J. van	Granville, J.J. de	Jansen-Jacobs, M.J.	Jenman, G.S.	Lanjouw, J. Lindeman, J.C.	Lindeman, J.C.	Maguire, B.	Mori, S.A.	Oldeman, R.A.A.	Pipoly, J.J.	Prévost, M.F.	Sabatier, D.
A	Acanthaceae	70	716	1	999	511	926	65	199	105	340	980	1	933	1
	Achariaceae	23	14	822	999	745	19	1	844	490	553	999	1	415	282
	Agavaceae	1	1	1	945	584	655	1	1	730	1	1	1	1	1
	Aizoaceae	1	607	1	159	1	1	993	1	1	1	1	468	428	1
	Alismataceae	79	145	1	65	999	999	493	1	127	1	1	1	370	1
	Amaranthaceae	1	979	37	1	655	591	995	114	543	11	63	116	510	1
	Amaryllidaceae	47	894	716	827	858	992	152	180	1	56	6	1	434	1
	Anacardiaceae	16	1	201	1	16	994	116	15	2	977	26	998	431	999
	Anisophylleaceae	1	1	1	1	1	1	1	790	1	1	1	1	1	1
	Annonaceae	999	1	4	458	236	1	1	1	2	997	376	11	999	999
	Apiaceae	1	93	1	642	1	164	286	999	1	1	67	1	835	1
	Apocynaceae	1	1	544	1	234	999	310	1	957	421	998	1	999	149
	Aquifoliaceae	703	1	402	1	27	977	965	1	727	1	1	999	1	23
	Araceae	491	999	1	999	2	21	1	998	694	1	1	270	418	1
	Araliaceae	1	1	40	129	263	867	1	40	144	861	424	394	139	399
	Arecaceae	1	601	1	999	998	45	1	318	708	1	4	1	1	1
	Aristolochiaceae	194	416	1	568	109	995	128	2	1	176	9	1	999	1
	Asteraceae	408	661	58	1	661	999	822	1	2	1	1	999	13	1
	Avicenniaceae	1	47	1	1	1	338	994	1	1	1	375	912	725	476
B	Balanophoraceae	636	969	1	980	220	104	523	1	744	1	147	1	1	1
	Bataceae	1	1	1	1	1	535	922	1	1	1	406	634	1	1
	Begoniaceae	6	573	1	999	107	2	28	588	610	2	999	1	988	1
	Bignoniaceae	1	623	30	1	681	997	9	53	47	392	167	2	994	1
	Bixaceae	987	1	1	10	742	847	392	891	280	1	129	350	1	1
	Bonnetiaceae	903	1	1	1	1	736	1	338	594	190	1	999	1	1
	Boraginaceae	694	260	934	1	999	426	231	1	1	741	900	627	729	4
	Brassicaceae	6	787	829	8	374	238	928	673	19	232	17	1	211	993
	Bromeliaceae	999	999	999	999	1	130	53	17	510	13	1	302	1	1
	Burmanniaceae	3	997	982	320	618	18	807	1	999	114	2	1	70	1
	Burseraceae	297	1	977	1	1	1	1	435	1	999	281	2	999	999
C	Cactaceae	414	998	999	37	919	373	1	999	160	1	262	43	31	1
	Campanulaceae	354	20	26	768	664	361	760	121	1	1	329	35	501	1
	Canellaceae	1	1	1	1	1	1	1	1	1	999	1	1	1	982
	Cannaceae	1	402	1	115	434	1	625	1	1	1	401	998	1	1
	Cardiopteridaceae	1	25	135	1	33	1	137	1	1	995	743	1	582	991
	Caricaceae	332	569	1	323	613	1	1	991	421	719	793	1	788	792
	Caryocaraceae	687	1	6	1	200	670	689	769	10	792	160	791	901	932
	Caryophyllaceae	1	1	418	1	255	336	807	560	391	1	1	440	386	1
	Cecropiaceae	97	4	680	5	541	19	405	742	19	125	897	864	993	122
	Celastraceae	999	1	45	54	373	202	27	972	969	999	999	450	1	881
	Celtidaceae	421	28	168	1	716	187	1	999	165	739	822	40	513	144
	Chloranthaceae	1	1	1	1	1	1	1	928	1	1	1	1	1	1
	Chrysobalanaceae	811	1	825	1	4	248	30	1	45	998	846	999	292	999
	Clusiaceae	985	1	607	1	154	534	258	987	999	900	30	999	756	2
	Cochlospermaceae	288	217	1	797	984	870	1	72	392	1	1	1	361	1
	Combretaceae	971	1	4	1	333	849	619	1	1	174	999	354	458	995
	Commelinaceae	2	942	420	984	995	973	781	3	226	1	214	3	993	1
	Connaraceae	161	636	990	142	986	787	162	85	798	999	999	523	41	50
	Convolvulaceae	1	985	473	1	993	958	969	507	1	522	279	21	995	1
	Costaceae	855	999	1	999	673	91	222	2	33	1	664	267	720	1

	Crassulaceae	1	567	1	1	1	1	1	21	1	1	1	1	995	1
	Cucurbitaceae	85	989	3	951	45	168	259	1	155	857	989	14	999	6
	Cunoniaceae	786	1	1	1	1	1	1	3	1	1	1	1	1	1
	Cyclanthaceae	1	999	1	999	1	1	94	1	944	230	934	136	102	1
	Cyperaceae	1	999	999	67	231	997	999	27	981	1	1	1	1	1
	Cyrillaceae	952	1	1	1	1	996	1	3	1	1	172	999	1	1
D	Dichapetalaceae	977	292	873	939	436	166	195	1	347	446	989	1	671	863
	Dilleniaceae	30	4	999	1	19	975	999	896	975	3	38	999	76	37
	Dioscoreaceae	91	399	55	668	428	999	60	799	587	423	52	53	799	1
	Dipterocarpaceae	797	1	1	1	1	1	1	750	1	1	1	999	1	1
	Droseraceae	1	501	983	1	944	699	997	1	995	1	1	404	1	1
E	Ebenaceae	836	7	49	909	93	380	850	1	597	999	933	140	999	999
	Elaeocarpaceae	44	3	999	2	73	573	871	998	14	999	767	23	331	994
	Eremolepidaceae	1	1	1	1	1	1	1	999	1	1	1	987	1	1
	Ericaceae	997	21	1	928	2	230	1	1	995	328	620	999	1	1
	Eriocaulaceae	1	715	885	1	998	986	999	1	999	7	1	407	2	1
	Erythroxylaceae	115	162	823	998	945	702	884	997	22	94	999	929	131	2
	Euphorbiaceae	993	16	147	1	952	699	180	308	7	979	923	1	999	957
	Euphroniaceae	1	1	1	1	1	1	1	765	1	1	1	997	1	1
F	Fabaceae	1	1	785	1	981	979	160	1	1	1	1	1	999	999
G	Gelsemiaceae	1	1	1	1	1	1	1	996	1	1	1	1	1	1
	Gentianaceae	461	999	699	999	1	149	986	1	157	325	96	193	69	1
	Gesneriaceae	999	999	1	999	2	2	8	58	889	1	999	886	596	1
	Goupiaceae	163	44	10	1	65	432	83	1	1	598	696	1	9	337
H	Haemodoraceae	1	523	1	918	927	718	597	25	981	21	915	1	436	1
	Heliconiaceae	999	998	1	999	995	147	25	49	59	1	1	999	23	1
	Hernandiaceae	205	604	268	7	598	475	322	9	1	739	506	620	998	1
	Humiriaceae	281	1	999	1	8	243	976	226	506	629	1	999	43	999
	Hydrocharitaceae	1	1	1	1	1	975	965	896	1	1	1	1	1	1
	Icacinaceae	862	50	1	28	32	265	617	427	153	50	818	976	928	991
I	Iridaceae	1	420	345	1	995	563	377	602	1	1	1	694	992	1
	Ixonanthaceae	1	1	1	1	1	1	1	1	958	677	1	1	1	999
K	Krameriaceae	829	1	1	1	999	1	1	1	1	1	1	1	1	1
L	Lacistemataceae	29	136	992	327	163	172	483	1	12	982	999	755	824	923
	Lamiaceae	1	922	586	73	646	320	449	999	1	217	214	14	363	1
	Lauraceae	999	1	999	1	1	1	1	3	78	999	892	13	78	999
	Lecythidaceae	626	1	999	1	1	1	1	999	1	999	4	1	999	999
	Lentibulariaceae	2	999	731	4	999	403	999	920	999	1	1	37	1	1
	Limnocharitaceae	1	1	1	1	1	901	994	999	695	1	1	1	1	1
	Linaceae	300	24	979	2	35	739	1	1	1	263	460	132	1	999
	Loganiaceae	13	13	988	446	76	433	831	999	310	915	902	1	870	110
	Loranthaceae	571	222	749	1	475	999	872	45	731	1	881	999	4	1
	Lythraceae	111	387	65	571	999	962	65	663	1	226	21	673	799	1
M	Magnoliaceae	974	1	1	1	1	1	1	141	1	1	1	1	1	1
	Malpighiaceae	999	16	1	1	963	999	646	1	974	164	999	999	947	1
	Malvaceae	1	1	138	1	999	981	151	29	1	3	25	1	712	422
	Marantaceae	1	999	1	999	972	1	1	58	1	6	66	4	880	1
	Marcgraviaceae	955	142	1	348	110	998	6	240	997	926	845	918	56	1
	Mayacaceae	1	1	967	989	942	974	971	38	636	1	1	1	1	1
	Melastomataceae	1	999	2	999	42	25	134	243	999	7	999	999	1	1
	Meliaceae	440	1	239	128	233	1	57	3	76	326	999	1	995	999
	Menispermaceae	981	731	791	175	999	445	627	993	1	997	775	263	2	2
	Menyanthaceae	1	591	1	281	899	152	636	150	1	1	1	1	777	1
	Molluginaceae	1	998	476	45	268	334	960	197	1	1	1	1	1	1
	Monimiaceae	998	660	1	996	992	1	322	1	992	547	1	1	1	1
	Moraceae	1	1	961	53	1	1	16	256	5	997	918	10	999	999
	Myristicaceae	284	1	1	1	342	2	485	999	14	999	868	3	177	999
	Myrsinaceae	454	28	305	999	98	737	791	398	149	122	999	999	31	1
	Myrtaceae	944	1	999	130	102	4	998	997	999	493	82	983	67	2
N	Najadaceae	1	1	1	1	1	1	1	999	1	1	1	1	1	1
	Nartheciaceae	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	Nyctaginaceae	1	97	156	867	2	493	999	1	50	997	295	998	194	553
	Nymphaceae	1	1	1	1	1	998	1	996	1	1	1	1	1	1
	Nymphaeaceae	1	116	335	8	259	994	936	1	57	1	122	17	1	1
O	Ochnaceae	1	1	1	1	1	973	1	51	1	1	1	1	1	1
	Olacaceae	190	3	502	947	1	91	812	1	730	589	995	30	286	843
	Onagraceae	1	84	974	138	554	919	999	999	8	1	2	52	103	1
	Opiliaceae	1	1	584	92	412	1	604	2	1	1	1	1	1	997

	Orchidaceae	614	999	999	999	199	999	981	985	985	1	1	2	1	1
	Oxalidaceae	840	15	106	228	999	912	118	1	316	392	68	345	301	1
P	Passifloraceae	962	999	1	15	877	999	49	79	1	1	60	143	999	1
	Pedaliaceae	1	1	1	1	1	994	1	1	1	1	1	1	1	1
	Phyllanthaceae	908	63	763	1	999	616	945	908	35	5	169	93	429	537
	Phytolaccaceae	265	11	20	2	522	961	104	839	78	111	104	83	720	1
	Picramniaceae	999	72	1	683	989	1	1	1	1	437	942	1	505	1
	Piperaceae	999	999	1	999	82	1	1	193	93	998	278	399	774	1
	Poaceae	1	999	999	1	61	994	999	1	999	1	1	1	1	1
	Podostemaceae	153	34	999	18	544	999	14	2	759	1	158	1	63	1
	Polygalaceae	258	997	978	1	999	900	998	1	515	2	648	976	105	1
	Polygonaceae	827	1	978	1	298	986	962	725	941	1	841	786	11	16
	Pontederiaceae	48	368	300	256	997	985	101	417	89	1	43	1	65	1
	Portulacaceae	1	984	1	907	988	332	1	1	163	1	33	1	415	1
	Primulaceae	1	1	1	1	1	1	1	165	1	1	1	1	1	1
	Proteaceae	303	10	501	13	987	706	281	1	453	18	560	74	438	977
	Putranjivaceae	13	1	157	7	1	1	1	218	39	985	133	1	26	999
Q	Quiinaceae	5	1	682	529	1	8	98	999	14	999	969	1	956	999
	Rafflesiaceae	770	666	1	744	1	1	1	999	1	1	927	1	1	1
R	Rapateaceae	2	999	2	742	48	922	23	1	999	5	78	950	484	1
	Rhabdodendraceae	454	1	1	55	1	1	1	1	211	674	690	998	791	946
	Rhamnaceae	161	340	99	216	999	341	104	1	1	741	447	1	222	740
	Rhizophoraceae	155	5	91	4	197	794	604	454	64	133	565	337	910	70
	Rosaceae	500	1	1	320	1	501	1	986	1	1	715	1	538	884
	Rubiaceae	999	309	1	999	578	1	1	891	942	2	999	999	1	1
	Ruppiaceae	1	760	1	1	1	1	991	1	1	1	1	1	1	1
	Rutaceae	886	78	6	226	684	207	159	1	110	672	987	1	414	1
S	Sabiaceae	1	1	1	1	1	942	1	10	740	1	1	1	1	1
	Salicaceae	1	13	830	998	985	31	62	1	2	999	999	1	998	534
	Santalaceae	1	1	1	1	1	1	1	999	1	1	1	1	1	1
	Sapindaceae	6	1	993	1	314	17	63	1	77	990	999	39	415	999
	Sapotaceae	843	1	140	1	1	1	101	999	15	999	1	1	999	999
	Sarraceniaceae	840	1	1	1	1	1	1	999	1	1	1	1	1	1
	Schlegeliaceae	803	49	1	183	3	901	47	1	257	785	987	508	99	258
	Scrophulariaceae	1	685	462	1	999	994	773	428	1	1	1	73	891	1
	Simaroubaceae	2	41	991	1	20	808	870	38	997	30	1	28	865	997
	Siparunaceae	894	2	185	699	788	1	71	744	12	999	973	1	999	999
	Smilacaceae	961	169	913	262	929	805	245	451	977	261	958	28	154	1
	Solanaceae	13	246	1	847	97	973	205	156	560	950	190	225	999	1
	Strelitziaceae	1	1	788	416	68	119	1	9	480	117	987	211	1	195
	Styracaceae	1	1	567	1	384	1	1	999	981	434	1	1	546	997
	Symplocaceae	1	814	624	1	531	62	431	999	337	212	570	999	98	816
T	Taccaceae	1	1	1	1	1	1	1	314	1	1	1	991	1	1
	Ternstroemiaceae	95	1	989	2	440	773	999	1	999	88	1	177	144	1
	Theaceae	1	1	597	1	1	1	921	945	1	1	1	1	1	1
	Theophrastaceae	999	342	1	331	998	1	1	973	718	154	849	80	1	275
	Thurniaceae	156	998	1	864	1	864	238	936	908	135	71	551	1	1
	Thymelaeaceae	1	673	1	999	336	1	1	1	1	746	1	1	1	1
	Trigoniaceae	848	42	673	52	946	689	527	1	420	655	507	864	988	58
	Triuridaceae	1	817	500	980	223	1	979	35	183	588	163	1	148	1
	Turneraceae	40	165	458	146	999	998	314	160	216	5	3	891	303	1
	Typhaceae	1	741	1	505	1	1	1	96	824	1	1	1	1	1
U	Ulmaceae	1	1	980	1	1	1	1	848	1	1	1	1	1	1
	Urticaceae	88	638	1	996	193	725	1	1	866	546	690	1	970	1
V	Velloziaceae	1	1	1	1	1	1	1	122	1	1	1	991	1	1
	Verbenaceae	999	44	1	1	989	999	456	1	24	9	411	207	930	1
	Violaceae	848	9	63	999	664	1	74	21	173	293	999	304	427	840
	Viscaceae	999	357	11	1	999	999	500	1	906	360	236	999	1	1
	Vitaceae	477	151	4	341	605	222	784	15	633	2	972	101	5	1
	Vochysiaceae	7	1	418	1	4	2	754	29	12	994	295	8	842	999
W	Winteraceae	1	1	1	1	1	1	1	543	1	1	1	1	1	1
X	Xyridaceae	18	93	999	1	877	287	999	1	999	1	1	83	1	1
Z	Zingiberaceae	999	912	17	994	949	13	83	999	990	8	818	148	485	3
	Zygophyllaceae	1	1	1	1	1	995	1	176	1	1	1	1	1	1

Appendix 4.1 Number of collections and species collected by a few large 'expedition type' collectors in French Guiana, Suriname and Guyana. Most of these collecting trips were made within the framework of the Flora of the Guianas. Each collector tried to collect the highest amount of species possible, resulting in very high S/N ratios. 20% of the collecting trip had an S/N of over 90%, while 80% had an S/N of over 75%, clearly supporting the idea that collectors maximize for the number of species within a collecting trip.

Collector	Year	N	S	S/N
Lindeman, J.C.	1953	1650	795	0.48
Lindeman, J.C.	1954	1563	845	0.54
Lindeman, J.C.	1955	123	102	0.83
Maas, P.J.M.	1965	464	389	0.84
Lindeman, J.C.	1967	76	71	0.93
Granville, J.J. de	1969	245	208	0.85
Granville, J.J. de	1970	596	422	0.71
Granville, J.J. de	1971	206	155	0.75
Maas, P.J.M.	1971	52	49	0.94
Granville, J.J. de	1972	726	495	0.68
Granville, J.J. de	1973	848	587	0.69
Granville, J.J. de	1974	300	257	0.86
Maas, P.J.M.	1974	175	155	0.89
Granville, J.J. de	1975	453	353	0.78
Lindeman, J.C.	1975	268	193	0.72
Lindeman, J.C.; Stoffers, A.L.	1975	262	176	0.67
Granville, J.J. de	1976	192	168	0.88
Mori, S.A.	1976	567	383	0.68
Granville, J.J. de	1977	76	73	0.96
Lindeman, J.C.	1977	189	168	0.89
Maas, P.J.M.	1977	200	163	0.82
Granville, J.J. de	1978	74	71	0.96
Granville, J.J. de	1979	304	246	0.81
Maas, P.J.M.	1979	776	586	0.76
Granville, J.J. de	1980	786	501	0.64
Lindeman, J.C.	1980	71	68	0.96
Lindeman, J.C.; Görts-van Rijn, A.R.A.	1980	604	458	0.76
Granville, J.J. de	1981	592	450	0.76
Lindeman, J.C.; Roon, A.C. de	1981	225	191	0.85
Maas, P.J.M.	1981	440	387	0.88
Granville, J.J. de	1982	274	226	0.82
Mori, S.A.	1982	533	360	0.68
Stoffers, A.L.	1982	374	303	0.81
Granville, J.J. de	1983	326	254	0.78
Mori, S.A.	1983	230	192	0.83
Granville, J.J. de	1984	1077	679	0.63
Granville, J.J. de	1985	961	617	0.64
Jansen-Jacobs, M.J.	1985	480	378	0.79
Granville, J.J. de	1986	569	402	0.71
Mori, S.A.	1986	273	249	0.91
Granville, J.J. de	1987	787	525	0.67
Jansen-Jacobs, M.J.	1987	590	442	0.75
Mori, S.A.	1987	85	76	0.89
Granville, J.J. de	1988	117	103	0.88
Maas, P.J.M.	1988	574	479	0.83
Mori, S.A.	1988	93	89	0.96
Granville, J.J. de	1989	640	439	0.69
Jansen-Jacobs, M.J.	1989	563	468	0.83
Mori, S.A.	1989	205	168	0.82
Mori, S.A.	1990	255	213	0.84
Granville, J.J. de	1991	228	173	0.76
Jansen-Jacobs, M.J.	1991	476	416	0.87
Mori, S.A.	1991	107	98	0.92

Görts-van Rijn, A.R.A.	1992	384	296	0.77
Jansen-Jacobs, M.J.	1992	575	455	0.79
Mori, S.A.	1992	195	177	0.91
Granville, J.J. de	1993	269	239	0.89
Mori, S.A.	1993	433	314	0.73
Granville, J.J. de	1994	118	107	0.91
Jansen-Jacobs, M.J.	1994	558	443	0.79
Mori, S.A.	1994	141	130	0.92
Granville, J.J. de	1995	402	311	0.77
Jansen-Jacobs, M.J.	1995	1126	790	0.70
Mori, S.A.	1995	86	62	0.72
Jansen-Jacobs, M.J.	1996	210	186	0.89
Granville, J.J. de	1997	143	130	0.91
Jansen-Jacobs, M.J.	1997	250	236	0.94
Mori, S.A.	1997	253	227	0.90
Granville, J.J. de	1998	243	208	0.86
Mori, S.A.	1998	76	66	0.87
Granville, J.J. de	1999	199	180	0.90
Jansen-Jacobs, M.J.	1999	431	321	0.74
Mori, S.A.	1999	107	90	0.84
Granville, J.J. de	2000	323	277	0.86
Mori, S.A.	2000	106	98	0.92
Granville, J.J. de	2001	128	121	0.95
Mori, S.A.	2001	109	100	0.92
Granville, J.J. de	2002	779	534	0.69
Mori, S.A.	2002	114	93	0.82
Jansen-Jacobs, M.J.	2003	316	133	0.42
Mori, S.A.	2003	107	86	0.80
Granville, J.J. de	2004	427	352	0.82

Appendix 4.2 All collectors and their number of collections made during long and short expeditions in Mabura Hill, Guyana and the Bauxite Mts are NE-Suriname.

Mabura		**Bauxite**	
Clarke, H.D.	690	BW	837
FD	278	Lanjouw, J.; Lindeman, J.C.	291
Jansen-Jacobs, M.J.	272	LBB	286
Polak, A.M.	271	Donselaar, J. van	280
Hoffman, B.	246	Lindeman, J.C.; Stoffers, A.L.	203
Mutchnick, P.	237	Andel, T.R. van	167
Pipoly, J.J.	199	Mori, S.A.	117
Maas, P.J.M.	152	Lindeman, J.C.	77
Steege, H. ter	131	Tresling, J.H.A.T.	53
Chanderbali, A.	125	Maguire, B.	52
McDowell, T.	103	Emden, W.C. van	40
Mori, S.A.	99	Cowan, R.S.	30
Pennington, R.T.	84	Unknown	30
Gillespie, L.J.	71	Collector indigenous	28
Stoffers, A.L.	49	Lanjouw, J.	25
Acevedo R., P.	42	Tjon-Lim-Sang, R.J.M.	22
Scharf, U.	27	WE	22
Jenman, G.S.	25	Hulk, J.F.	21
Hahn, W.J.	24	Schulz, J.P.	18
Smith, A.C.	21	Stahel, G.	14
Cruz, J.S. de la	21	Maas, P.J.M.	11
University Guyana - Neotropical Botany	19	Scharf, U.	10
Raes, N.	18	Lindeman, J.C.; Cowan, R.S.	9
Schomburgk, R.H.	17	WH	9
Ehringhaus, C.	15	Zaandam, C.J.	9
Maguire, B.	12	BBS	8
Redden, K.M.	7	Mennega, A.M.W.	8
Henkel, T.W.	5	Versteeg, G.M.	8
Grewal, M.S.	4	Kock, C.	6
Bartlett, A.W.	4	Lems, K.	4
Abraham, A.A.	4	Lindeman, J.C.; Mennega, E.A.	4
Kelloff, C.L.	4	Jonker, F.P.	3
Gleason, H.A.	3	Kramer, K.U.	3
Schomburgk, M.R.	3	Sauvain, M.	3
Sandwith, N.Y.	3	Wessels Boer, J.G.	3
Rombouts, H.E.	2	Christenhusz, M.J.M.	2
Christenson, E.A.	2	Florschütz, P.A.	2
Stockdale, F.A.	2	Lindeman, J.C.; Görts-van Rijn, A.R.A.	2
Persaud, C.A.	2	Lindeman, J.C.; Roon, A.C. de	2
Unknown	2	Wullschlägel, H.R.	2
Arets, E.J.M.M.	1	Evans, R.J.	1
Fanshawe, D.B.	1	Focke, H.C.	1
Granville, J.J. de	1	Gonggrijp, J.W.	1
Knapp, S.	1	Prance, G.T.	1
Davis, T.A.W.	1	Troon, F. van	1
Ek, R.C.	1	Webster, G.L.	1
Kennedy, H.	1		
Total	**3,302**		**2,727**

Samenvatting

Primaire vindplaatsgegevens van planten en dieren komen in toenemende mate beschikbaar op het Internet. Naar verwachting zullen er binnen tien jaar zo'n één miljard vindplaatsgegevens uit de hele wereld op het Internet beschikbaar zijn (Guralnick & Hill 2009). Dit soort gegevens wordt steeds belangrijker voor biologen die geinteresseerd zijn in de verspreidingspatronen van soorten. Echter, voordat deze gegevens gebruikt kunnen worden, moet vastgesteld worden of en in hoeverre zij te lijden hebben onder vertekening die te maken heeft met de manier waarop soorten verzameld worden (Graham *et al.* 2004; Hortal *et al.* 2008). Het doel van de voorliggende studie was daarom om deze vertekening of bias te onderzoeken aan de hand van de oorspronkelijke vindplaatsgegevens van de planten in de herbariumdatabase van de Guianas (Guyana, Suriname en Frans Guiana). Vervolgens werd deze database gebruikt om (a) een model te ontwikkelen dat simuleert hoe vaak verschillende soorten in het herbarium vertegenwoordigd zijn; (b) de relatieve bijdrage van verspreidingsvermogen en milieufactoren in de samenstelling van de flora van de Guianas te bepalen en (c) patronen van soortenrijkdom en endemisme in de Guianas vast te stellen.
De Guianas werden gekozen als studiegebied, omdat hier al meer dan een eeuw intensief door het Herbarium van de Universiteit van Utrecht planten verzameld worden waardoor het een groot aantal collecties uit het gebied bezit (Ek 1990; Ek 1991; Hoff niet gepubliceerd). De collecties uit dit herbarium vormen de ruggengraat van de gegevens die gebruikt zijn voor deze studie. In de loop der tijd hebben gespecialiseerde botanici regelmatig de soortsidentificaties in dit herbarium aan de hand van voortschrijdende taxonomische inzichten geactualiseerd. In het Herbarium van Utrecht bevinden zich ook vele duplicaten van door niet-Utrechtse botanici in de Guianas verzamelde planten. De gegevens uit het Herbarium van Utrecht werden aangevuld met gegevens afkomstig van andere deelnemende herbaria uit het Flora of the Guianas projekt en van soortenlijsten van botanici die in het gebied verzameld hebben. De in deze studie gebruikte database is daarom de meest complete en geactualiseerde lijst van angiospermen die beschikbaar is voor de Guianas.

De herbarium database – rijk aan soorten

Al meer dan vier eeuwen lang zijn botanici bezig herbariumcollecties aan te leggen van de planten van de Guianas (Ek, 1990; Ek 1991; Hoff niet gepubliceerd), al is de precieze plek waar de oudere collecties verzameld zijn meestal onbekend. De collecties in de database, die voor deze studie is gebruikt, zijn tussen 1804 en 2004 door in totaal 560 botanici bijeengebracht. De database omvat 168.487 afdoende gedocumenteerde collecties met daarin 7.146 soorten. Deze collecties zijn niet evenredig verdeeld over de families, geslachten, groeivormen en landen. De vijf meest soortenrijke families in de database zijn de Fabaceae, Rubiaceae, Melastomataceae, Poaceae en Cyperacea. De tien families met de meeste gegevens beslaan ongeveer 42% van alle collecties en 43% van alle soorten in de database. Het hoogste aantal soorten werd verzameld in Guyana en het laagste aantal in Suriname. Hoewel ongeveer 35% van de soorten in alle drie de landen verzameld werden, werd 42,6% in slechts één land verzameld. Slechts enkele soorten zijn vertegenwoordigd met een groot aantal collecties; daarentegen zijn er van 38% van de soorten minder dan vijf collecties beschikbaar.

Tussen 1804 en 2004 werd het geografische gebied dat door botanici werd bestreken in hun verzameltochten geleidelijk groter, al werden sommige gebieden zoals rondom onderzoeksstations en steden het meest intensief bezocht (hoofdstuk 2). Van slechts 28% van alle gridcellen van 5 x 5 boogminuten (ongeveer 10 x 10 km) zijn collecties bekend (hoofdstuk 6). De snelheid waarmee nieuwe soorten aan de database werden toegevoegd nam af tot 1,4 voor iedere 100 collecties tegen het einde van de waarneemperiode, zelfs al werden er nieuwe gebieden ontsloten die nog niet eerder verzameld waren (hoofdstuk 2). Deze afname suggereert dat de meeste (regionaal) algemene soorten in de Guianas nu wel 'gevonden' en in de database vertegenwoordigd zijn.

De herbarium database – rijk aan vertekening

Een van de belangrijkste bezwaren die tegen het gebruik van herbariumdatabases voor biodiversiteitsonderzoek kan worden ingebracht betreft de veronderstelde statistische vertekening in herbariumdatabases als gevolg van de wijze waarop planten verzameld worden (Soberon *et al.* 2000; Reddy & Davalos 2004; Graham 2004). In deze studie is er gekeken naar

de mate van historische, geografische, taxonomische en seizoensgebonden vertekening. Eén van de belangrijkste conclusies is dat het aantal soorten dat in een gebied gevonden is bijna altijd bepaald wordt door het aantal collecties: hoe meer collecties, hoe meer soorten (hoofdstuk 3). Er was sprake van historische vertekening in verzamelen: de relatie tussen het aantal collecties en het aantal daarin aangetroffen soorten was afhankelijk van de lengte van de periode waarin die collecties gemaakt waren. Als gevolg hiervan kunnen schattingen van soortenrijkdom die op verschillende verzamelperioden gebaseerd zijn, niet makkelijk met elkaar vergeleken worden. Dit komt doordat soortenaccumulatiecurves geen asymptoot bereiken, maar het aantal soorten altijd blijft toenemen met het aantal collecties. Veel modellen die gebruikt worden om soortenrijkdom te schatten, zoals bijvoorbeeld het Michaelis-Menten model, gaan wel uit van asymptotisch gedrag.

Met behulp van een verbeterd model, een combinatie van het Michaelis-Menten model en het Arrhenius model, kan het aantal angiosperme soorten in de Guianas op ongeveer 12.000 geschat worden (hoofdstuk 3). Van soorten, die nu nog niet in de database voorkomen, kan verondersteld worden dat ze van nature zeldzaam zijn, of dat ze tot zeer kleine gebieden beperkt zijn, mogelijk in gebieden die nog niet door botanici bezocht zijn. Dit is in overeenstemming met de theoretische talrijkheidsverdelingen van soorten in de natuur, zoals die door Hubbell (2001; Hubbell *et al.* 2007) zijn opgesteld, en waarin voorspeld wordt dat veel soorten in de natuur zeer zeldzaam voorkomen. De kans is klein om uiteindelijk alle soorten te verzamelen door middel van 'ad-hoc verzamelexpedities' of systematische steekproeven. Vele soorten zijn zo zeldzaam dat ze mogelijk nooit verzameld zullen worden.

Botanici hebben een sterke voorkeur voor gebieden die dicht bij rivieren en wegen liggen, waardoor er hier relatief veel soorten verzameld zijn (hoofdstuk 3). Het kon echter worden aangetoond, dat, als de milieuvariabelen in dergelijke gebieden vergeleken worden met die van willekeurige plekken in de Guianas als geheel, er geen vertekening als gevolg van verschillen in milieuomstandigheden in de herbariumdatabase optrad. Aangezien de verzamelinspanning dus representatief was voor de milieuomstandigheden in de Guianas, kan worden aangenomen dat de geografische vertekening in de database nauwelijks

implicaties zal hebben voor de betrouwbaarheid van de resultaten van soortsverspreidingsmodellen (species distribution models of SDMs), die gebaseerd zijn op de database en die patronen in soortenrijkdom voorspellen. Het blijkt dat er meer collecties en meer soorten verzameld zijn gedurende de droge dan de natte maanden van het jaar. De fenologie van de bloei kan bijna volledig verklaard worden door de verzamelinspanning (hoofdstuk 3), waaruit geconcludeerd zou kunnen worden dat gebruik van fenologische informatie die gebaseerd is op herbariumgegevens problematisch is. Toch is er een goede correlatie tussen de bloeigegevens gebaseerd op herbariumgegevens en onafhankelijke, in het veld verzamelde bloeigegevens. Dit zou verklaard kunnen worden uit de gedachte dat verzamelaars hun expedities zodanig afstemmen op de hun bekende bloeiperiodes, dat de kans om bloeiende planten tegen te komen groter is. Dit geldt veel minder voor vruchtdragende collecties, wellicht gerelateerd aan het feit dat vruchten vooral in het natte seizoen gevonden worden (en bloemen in het droge seizoen). Hoe dan ook, in fenologische studies, die gebaseerd zijn op herbariumgegevens, zal rekening gehouden moeten worden met de vertekening, die veroorzaakt wordt door de seizoensgebondenheid van de verzamelinspanning.
Deze vertekeningen in de herbariumdatabase van de Guianas hebben gevolgen voor sommige maar ni*et all*e biodiversiteitstoepassingen. De gegevens zijn geschikt voor SDMs en voor het schatten van soortenrijkdom. Seizoensinvloeden op de verzamelinspanning hebben hun gevolgen voor fenologische studies. Indien geen correctie wordt toegepast, zullen fenologische gegevens eerder iets zeggen over de verzamelinspanning gedurende droge maanden dan over de bloei van soorten.

Niet tweemaal dezelfde – over hoe botanici verzamelen

Bij het beantwoorden van de fundamentele vraag waardoor het aantal soorten in een gebied bepaald wordt, is het is onmogelijk voorbij te gaan aan het enorme aantal collecties dat wordt bewaard in herbaria. Een belangrijk probleem van herbariumgegevens is, dat de hieruit bepaalde dominantie-diversiteitscurves niet representief zijn voor die van de natuurlijke gemeenschappen waaruit ze afkomstig zijn. Dit komt doordat soorten niet willekeurig verzameld worden. De dominantie-diversiteitscurve van herbariumgegevens is ‘platter’ dan die van proefperken, ofwel, in het herbarium bevinden zich meer soorten die ieder

vertegenwoordigd zijn met minder individuen dan in het veld (hoofdstuk 4). De relative abundantie van soorten die in een bepaald gebied verzameld zijn hangt onder andere samen met het aantal botanici dat in het gebied op bezoek is geweest en de hoeveelheid tijd die dezen aan verzamelen hebben besteed. De niet-willekeurige manier van verzamelen leidt ertoe, dat statistische methoden die van aselecte steekproeven uitgaan, niet bruikbaar zijn. In hoofdstuk 4 wordt een model van de relatieve abundantieverdeling van soorten in herbaria gepresenteerd. Dit model wordt op een niet-willekeurige maar voorspelbare manier uit de log-serie afgeleid, gebaseerd op gegevens van proefperken.
Het model bestaat uit twee delen. Het eerste deel maakt gebruik van proefperkgegevens (van het Mabura Hill gebied in Guyana en de bauxietbergen in noordoost Suriname) om de relatieve abundantieverdeling van alle soorten in een bepaald gebied vast te stellen, en van een 'zero sum multinomiale verdeling' om de structuur van de soortengemeenschap in het gebied te beschrijven. In het tweede deel worden de resultaten van het eerste deel toegepast om het verzamelgedrag van botanici te simuleren, gebaseerd op de herbariumgegevens van de twee gebieden en met gebruikmaking van vier verschillende scenarios. De belangrijkste strategie, namelijk om nooit dezelfde soort tweemaal te verzamelen, genereerde relatieve abundantieverdelingen die goed vergelijkbaar waren met die van het herbarium. Het resultaat werd nog beter, als er een soort 'bezorgdheidsfactor' werd ingebouwd, die de relatieve abundantie van talrijke soorten in het herbarium reproduceerde. De lange staart van de relatieve soortsabundantieverdeling kon gereproduceerd worden indien door het model werd aangenomen dat botanici verschillende habitats met verschillende soortensamenstelling bemonsteren. In een ander scenario werd het aantal collecties per botanicus gelijkgesteld aan het werkelijk aantal gemaakte collecties, onder dezelfde aanname om nooit dezelfde soort tweemaal te verzamelen en gebruikmakend van de zero-sum multinomiale verdeling van het gebied . De gesimuleerde dominantie-diversiteitscurve die het gevolg was van dit scenario leek zeer sterk op de werkelijke curve van de herbariumdatabase. Ook al bleek het mogelijk om de relatieve abundantieverdeling van het herbarium te reconstrueren uit de soortensamenstelling in het veld, het was omgekeerd niet mogelijk om de soortensamenstelling in het veld te reconstrueren vanuit de herbariumdatabase. Dit is het gevolg van het grote aantal zeldzame soorten.

Bemonstering van de zero-sum multinomiale verdeling van het gebied gaf aan dat de resulterende soortenaccumulatiecurve geen asymptoot bereikt. In het begin is er een snelle toename van het aantal soorten in de steekproef doordat de algemenere soorten snel 'aangetroffen' worden. Naarmate de bemonstering doorgaat, blijven er nieuwe, zeldzame soorten opduiken, hetgeen resulteert in een positieve helling in de soortenaccumulatiecurve. Dit suggereert dat het Michaelis-Menten model een fundamenteel verkeerd model is om soortenrijkdom te schatten en dat de modellen die soortenaccumulatiecurves gebruiken om soortenrijkdom te schatten rekening moeten houden met deze zwaktes.

Floristische gelijkenis binnen de Guianas – de invloed van afstand en ecologie

In deze thesis wordt voor het eerst gebruik gemaakt van herbariumgegevens om te testen hoe de neutrale theorie en de niche theorie de floristische similariteit over een landschapsgradient voorspellen. De neutrale theorie voorspelt dat floristische similariteit van twee plekken afneemt naarmate de afstand ertussen toeneemt (Hubbell 2001). De nichetheorie daarentegen voorspelt dat de floristische samenstelling varieert al naar gelang de milieuomstandigheden ter plekke, als gevolg van soortsspecifieke aanpassingen aan het milieu (Hubbell 2001; Tilman 1982). Herbariumgegevens hebben het voordeel boven proefperkgegevens dat ze grotere spatiele schalen beslaan en meer groeivormen omvatten. Daarentegen is een van de aannames bij het bepalen van floristische similariteit dat de informatie door middel van aselecte steekproeven wordt verzameld – iets wat niet opgaat voor herbariumgegevens (hoofdstuk 3 en 4). De relatieve abundantie van soorten in het herbarium is geen goede afspiegeling van abundantie in het veld (hoofdstuk 3 en 4). Botanici investeren hun tijd in het vinden van nieuwe soorten in plaats van het meten van meer individuen van dezelfde soort, waardoor ze erg effectief zijn in het vinden van veel soorten, maar niet in het bepalen van relatieve abundantie. Daarom worden presentie/absentiegegevens en gebieden met verschillende verzamelintensiteiten in hoofdstuk 5 gebruikt om (a) te bepalen in hoeverre geografische afstand (ofwel verspreidingsvermogen) en variatie in milieufactoren bijdragen tot verschillen in floristische samenstelling in de Guianas; en (b), de relatieve bijdrage van afstand, milieuvariatie en hun combinatie aan de

variatie in soortensamenstelling te bepalen. Daarna werd bekeken of het zo was dat de verschillen tussen twee gebieden kleiner zijn voor soorten die goede verspreiders zijn dan voor soorten die slechte verspreiders zijn, zoals dat door de neutrale theorie wordt voorspeld. De milieufactoren die in beschouwing werden genomen waren hoogte, temperatuur, regenval en het seizoenspatroon in de regenval.
Door middel van Manteltests kon worden aangetoond dat floristische verschillen sterk gecorreleerd waren met geografische afstand en verschillen in hoogte en temperatuur, en tot op zekere hoogte ook met verschillen in regenval en regenvalpatroon. De mate waarin verschillen in soortensamenstelling door deze factoren verklaard konden worden hing af van de verzamelintensiteit. Hoe intensiever bepaalde gebieden verzameld waren, hoe hoger de mate waarin de floristische verschillen verklaard konden worden. De floristische similariteit tussen twee willekeurige gebieden was hoger voor goed verspreide soorten, zoals windverspreiders of soorten met lichte zaden, dan voor slecht verspreide soorten, zoals door dieren verspreide soorten of soorten met zware zaden. Lianen, epifyten en kruiden vertoonden een kleiner afstandseffect in similariteit dan palmen, struiken en bomen. Deze resultaten suggereren dat betere verspreiders gelijkmatiger over het landschap verdeeld zijn terwijl slechte verspreiders meer geclusterd voorkomen. De voorspellingen van de neutrale theorie zijn beter in staat deze patronen te verklaren dan de niche theorie. Herbariumgegevens kunnen derhalve gebruikt worden in de discussie over de relatieve bijdragen van beperkte verspreiding en nichedifferentiatie aan de samenstelling van soortengemeenschappen op landschapsschaal. Dan is het wel belangrijk om gebruik te maken van intensief verzamelde gebieden. Aangezien verzamelinspanning in hoge mate bepalend is voor het aantal soorten dat gevonden wordt, is het onzuiver om gebieden met sterk verschillende verzamelintensiteiten te vergelijken. Verder is het van belang om presentie/absentiegegevens te gebruiken en niet abundantiegegevens, aangezien, zoals hierboven al werd aangetoond, de abundantie van soorten in het herbarium geen goede afspiegeling is van de werkelijke abundantie.

Soortenrijkdom en patronen van endemisme

Zoals hierboven al vermeld werd zijn er uit slechts 28% van de gridcellen van 5 x 5 boogminuten (ongeveer 10 x 10 km) collecties bekend, ondanks het feit dat

er al eeuwen planten verzameld worden in de Guianas (hoofdstuk 6). Aangezien de verzamelinspanning bepalend is voor het aantal aangetroffen soorten (hoofdstuk 3), zou het niet gerechtvaardigd zijn om gegevens uit het herbarium te gebruiken om soortenrijkdom te beschrijven. Om de door onvoldoende onderzoeksinspanning veroorzaakte leemten in de verspreidingspatronen van planten te vullen, zijn de potentiele diversiteit en patronen van endemisme gemodelleerd met behulp van soortsverspreidingmodellen (SDMs), die het voorkomen van planten modelleren op basis van bekende verspreidingsgegevens en bodem-, klimatologische en hoogtegegevens (hoofdstuk 6). De resulterende kaart van diversiteit, gebaseerd op ongeveer 41% van de in het herbarium aanwezige soorten, laat zien dat de kustzone de hoogste soortenrijkdom van de Guianas bezit. In dit gebied is de verzamelintensiteit altijd hoog geweest. Het intensief verzamelde *arrondissement* van Cayenne in Frans Guiana liet een hogere soortenrijkdom zien dan de rest van de Guianas. Daarentegen was de voorspelde soortenrijkdom in het zuidoostelijke deel van Guyana en het zuidelijke deel van Suriname – beide met een lage verzamelintensiteit – laag. Deze uitkomsten komen niet overeen met onze verwachtingen en we denken dat de daadwerkelijke patronen van diversiteit door het model niet juist voorspeld worden. De onverwacht hoge diversiteit in de kuststrook is waarschijnlijk het gevolg van de hoge verzamelintensiteit ter plaatse, in combinatie met de eigenschappen van het model zelf. De meeste soorten (72%) die voor het model gebruikt werden komen in dit gebied voor. Verder is de milieuvariatie langs de kust relatief gering, maar wel afwijkend van de rest van de Guianas. Dat betekent dat als een bepaalde in dat gebied verzamelde soort succesvol gemodelleerd kon worden, er een grote kans was dat de soort voorspeld werd in de gehele kuststrook voor te komen. In gebieden met een hogere milieuvariatie, daarentegen, is er een kleinere kans dat soorten voorspeld worden over het gehele gebied voor te komen. Het *arrondissement* van Cayenne heeft een hoge diversiteit door de combinatie van een homogeen milieu en hoge verzamelintensiteit. De rest van de Guianas is relatief arm, doordat het gebied matig verzameld is waardoor veel soorten niet in het model mee mogen doen omdat de kritische grens van vijf collecties die vereist is om te kunnen modelleren niet werd bereikt. Dit drukt de voorspelde diversiteit. Het is ook mogelijk dat sommige gebieden eenvoudigweg arm aan soorten zijn.
In een analyse met Detrended Correspondence Analysis kwam naar voren

dat van alle onderscheiden biogeografische regios (ter Steege and Zondervan 2000) in de Guianas, het Pakaraimagebergte in Guyana het meest afwijkend was. Gebieden in het noorden en in het zuiden waren duidelijk verschillend in floristische samenstelling. Deze analyse stoelt op de aanname dat vooral de algemene soorten bepalend zijn voor de diversiteitspatronen (Lennon *et al.* 2004), en dat in een gegeven gebied de meeste algemene soorten wel verzameld zijn (hoofdstuk 4).

De meeste endemische soorten werden voorspeld voor te komen in het Pakaraimagebergte, het Kanukugebergte en de kust van Guyana; en in het Sipaliwinigebied en de kust van Suriname (hoofdstuk 6). De Tafelberg in Suriname en het *arrondissement* van Cayenne in Frans Guiana kennen een intermediair niveau van endemisme. Deze voorspellingen komen overeen met eerdere verwachtingen (Granville 1988; ter Steege *et al.* 2000; Conservation International 2000), behalve dat de hoge mate van endemisme in centraal Guyana (ter Steege *et al.* 2000) niet door het model voorspeld werd.

Acknowledgements

Although I grew up in a country with more than 70% forest, I saw the forest for the first time during my university days. Doing my MSc field work under the Tropenbos-Guyana Programme in Mabura Hill has certainly formed an everlasting impression on me. I could never get enough of the treasures of the amazing forest – the huge impressive trees with fruits and flowers of all sizes, shapes and colours. Every day was exciting with the unforgettable sounds of birds, insects and monkeys, the unexpected animals hurrying to hide or boldly monitoring every move.

This research would not have been possible without Hans ter Steege who secured the funding. Hans, thank you for giving me the opportunity to do the study. Your infectious enthusiasm, your innovative ideas and your knowledge of the study area have contributed greatly to the success of the study. I thank Marinus Werger for agreeing to be my promotor and for your invaluable comments on the individual chapters. I was always nervous when you had one of my manuscripts but you eased my anxiety by returning them within the shortest time. You gave me a lot of confidence, thank you.

Without data there is no thesis. I have used information from specimen labels from plants collected by botanists. Many of these dedicated botanists have travelled to very remote and breathtaking areas and often stayed under the most rudimentary circumstances in their quest for discovering new species. I express sincere gratitude for your contribution and show deepest respect for your profession.

Many herbaria and botanists have provided their digital specimen data and in the process saved me an enormous amount of time and effort. To all of you, thank you. Many persons have worked on the NWO-groot project for the NHN-Utrecht branch – Bert, Eric, Fenneke, Harm-Willem, Job, Johan, Jorit, Leon, Luc, Mark, Renske and Zé, - you guys worked hard and showed so much enthusiasm, thank you.

I am not a plant collector but I am grateful to Marion Jansen-Jacobs for allowing me to go with her on two of her expeditions. This has given me the much needed insight on what botanists do when they are in the field. I thank also all of the staff of the former Utrecht herbarium for their support and kindness during my years there.

I am grateful to many people who have given me invaluable advice on various chapters of this thesis. I especially thank Heinjo During who has kindly reviewed most of the original manuscripts and gave me invaluable advice on the analyses. I thank also Jerome Chave, Marion Jansen-Jacobs, Paul Maas, Jenny Ordoñez, Niels Raes and James Weedon for their advice and help on various aspects. I thank Pete and Bertus, the computer elves, for helping me with my numerous troubles.

As a Guyanese girl, born and bred in a small village, I was one of the lucky ones to go to university. Neither of my parents finished high school but they understood the value of good education and thanks to their support and encouragement during my formative years I have come this far. I would like to thank my family - especially my mummy- for encouraging me and for respecting the difficult fact that I have chosen to live so far away from them. I also thank my family-in-love for their interest, support and encouragement throughout the years.

One of the disadvantages of using existing data is that I spent more time at my desk than the average PhD and could only dream of the place names on the specimen labels. Happily, the nice thing about working in such a multi-cultural group is that discussions at the lunch on their own experiences in far away countries and in the Netherlands. Thanks for sharing your experiences and contributing to us all having good laughs during the lunch.

I thank Pieter for being the nicest room mate I could wish for. I thank also my Bunnik-Rhenen train friends, my Rhenen neighbours, the families van de Brink and van Leerdam for their interest and encouragement throughout the years.

During the study I went back to Guyana twice and I thank my friends Michelle (and of course Peter) and Raquel, still remaining there, for welcoming me in their homes and for their friendship and un-ending frank conversations which I miss awfully. Friends like you don’t grow on a tree.

Roderick, thank you for kindness, understanding and gentle firm support during the last years. I value your good listening, clear-thinking, fairness and encouraging me to always stand on my own two feet.

To all of you who have contributed in some way to this thesis but I have not mentioned, I thank you.

Curriculum vitae

Padmattie Persaud Haripersaud was born on 15 October 1967 in Annandale, East Coast Demerara Guyana. She attended St. Joseph's High School and the Bishops High School. She completed her BSc and MSc at the University of Guyana. She worked in Guyana between 1989 and 2002. In 1989 she assisted in collecting and analyzing water samples for a potable water quality assessment project. Between 1991 and 1993 she worked in a plant tissue culture laboratory where she assisted in initiating plant tissue cultures, supervised the daily lab activities and monitored field trials. Between 1998 and 2000 she was curator of a herbarium and during this time she initiated digitizing of herbarium records; organized tree identification courses; participated in many national expeditions; participated in plant collection expedition to south Guyana with the Nationaal Herbarium Nederland- Utrecht. Between 2000 and 2002 she worked as the national coordinator for the UNDP-Programme on Forest project where she organized international and national workshops, facilitated international consultants, prepared project reports and newsletters; managed the national forest certification process.