Using Herbarium Specimens to Explore Species Diversity and Collecting Completeness Across Plant Habit Groups in Tropical Africa.

By

Stephen Garrett

(Student Number: s1669270)

Thesis submitted in partial fulfilment of the requirements for the MSc in Biodiversity and Taxonomy of Plants

At

And

Supervisors: Antje Ahrends and David Harris

August 2017

Acknowledgements.

Supervisors Dr. Antje Ahrends and Dr. David Harris offered invaluable guidance throughout this thesis. The statistical methodology was only achievable with the support of Ahrends. Harris supplied the extensive dataset used in this study and analysis was dependent on the high caliber of his floristic expertise. Both gave instrumental advice at every stage and dedicated many hours to offering thoughtful feedback. Thanks go to Dr. Maria Vorontsova, Martin Xanthos, and Dr. Nick Hind for offering their time and expertise and helping wade through the otherwise overwhelming processes of *Poaceae*, *Cyperaceae*, and *Asteraceae* identification, respectively. Further acknowledgements must go to the collectors from the study area (Appendix 3) who contributed to the dataset used, especially Jean-Marie Moutsambé, E. Kami and S.T. Ndolo Ebika. Federico Fabriani is thanked for insight he provided through previous work he conducted on the same dataset. Dr. Zoë Goodwin and Max Brown require mention and my eternal gratitude for their patience and generosity in guiding me through training in R, Excel and Brahms. Finally, I thank Thomas Vowles for dragging me across the finish line.

Abstract.

Tropical floras show high rates of biodiversity and species richness, yet they are often the most threatened, remain under-studied and are far from being fully understood. Accurate assessment is imperative for effective monitoring and management of tropical plant communities, and this requires quantification of the plant diversity using species inventories. Fieldwork constraints result in collecting bias favouring certain habit groups over others. This study aimed to assess collecting completeness, measure species diversity, and explore collecting bias between different habit-types within the flora of a forested area extending across Republic of Congo, Cameroon and Central African Republic. An additional aim of the study was to explore the suitability of herbarium specimens for estimating total species richness. Rarefied and extrapolated species accumulation curves, as well as asymptotic species diversity models were used. The results indicated that herbarium specimen data are a robust tool for exploring species diversity when appropriate methods are applied. The bias-corrected form of the Chao1 model was seen to adequately account for different sampling intensities across and under collecting within habit groups. This study reached the conclusion that collecting completeness has not been reached for the flora of the study area, nor for any of the habit-types explored. Herbs were seen to have the least species richness representation already collected, and yet they were estimated to have the highest level of species richness. Shrubs appeared to be the least species rich group. At >96%, trees showed the highest rate of collecting completeness.

Keywords: Tropical floras, high rates of biodiversity and species richness, threatened, understudied, quantification of plant diversity, species inventories, fieldwork constraints, collecting bias, habit groups, collecting completeness, species diversity, Sangha-Trinational, Republic of Congo, Cameroon, Central African Republic, suitability of herbarium specimens, rarefaction, extrapolation, species accumulation curves, asymptotic species diversity, Chao1, iNEXT, R, herbs, shrubs, climbers, trees.

List of Figures.

List of Tables

Table of Contents

1 Introduction

1.1 Background

Tropical floras show the highest rates of biodiversity and species richness for vascular plants (Lieberman & Lieberman 2007; Schemske 2002; Gentry 1988) on our planet and yet are often the most threatened (Prance et al. 2000; Colwell et al. 2008), faced with land-use change (Millington and Jepson 2008) and the adverse effects of climate change (Corlett 2012). It is therefore crucial that tropical plant diversity is more thoroughly inventoried so that efforts can be efficiently focused to most effectively combat and ameliorate threats to global biodiversity (Dick & Kress 2009).

Tropical floras are understudied (Prance et al. 2000; Ruokolainen et al. 2005) with an estimated 40% of vascular plant species in the tropics still undescribed (Hubbell et al. 2008). Tropical flora collections often have low taxonomic resolution (Dick & Kress 2009). Amazonian forest inventory plots of trees greater than ten centimeters in diameter at breast height (10cm dbh) reveal that 20% of the specimens collected are never identified to species level (Ruokolainen et al. 2005).

Accurate assessment of species diversity is imperative for effective monitoring and management of tropical plant communities. Assessing species diversity requires some quantification of that diversity (Magurran 2004). To this end, a range of different species diversity indices have been developed, which offer quantitative measures reflecting the number of species, their abundance, and their evenness (Gotelli & Colwell 2011). Of these, the number of observed or expected species in a given area (= species richness) is probably the most popular and simplest way of describing biodiversity at community and regional scales (Gotelli & Colwell 2001). The documentation of tropical plant diversity through species inventories is the foundation for any biodiversity research aiming to explore plant species diversity (Shen et al. 2003), and any species inventory and subsequent assessment of plant biodiversity requires initial data gathering through fieldwork.

Difficulties in undertaking fieldwork in tropical forests are extensive and include insufficient knowledge, funding constraints, physical inaccessibility, inaccessibility due to political strife, and a simple lack of time to do it all. These difficulties all inhibit biodiversity inventorying

and therefore difficult choices are required and research prioritizing is necessary (Dick and Kress, 2009).

Bias is often introduced to broad scale biodiversity studies of tropical forests by inventory focus being set on trees greater than 10cm dbh. This results in the inevitable absence of robust studies of understory shrubs (here defined as woody plants with stems of a dbh less than 10cm), herbaceous plants, epiphytes and climbers (both woody and herbaceous) (Gentry & Dodson 1987; Knapp 2017; Prance et al. 2000).

The sheer scope of highly diverse tropical plant communities makes inventorying all individuals and even all species an impossible aim. To overcome collecting incompleteness various statistical tools for estimating an assemblage's species diversity have been developed (Chao 2006; Gotelli & Colwell 2011).

1.2 General Collecting vs Plot-Based Collecting

General collecting as referred to in this thesis, includes any plant specimen collected outside of plot-based work. Plot-based collecting is defined here as survey field work undertaken within fixed area plots. The definitions used within this study follow Garcillán & Ezcurra 2011 and Fabriani 2015.

1.3 Species Accumulation Curves

Species accumulation curves (SAC) plot the cumulative number of species against a unit of measurement chosen to represent sampling effort (Colwell et al. 2012; Ugland et al. 2003; Chao 2006; Fisher et al. 1943). The randomly-sampled, independent sample unit can be chosen to show individuals, or a group of individuals from a single collection effort such as traps, nets, quadrants, plots or a fixed period (Colwell et al. 2012). When individuals are used, the sample unit is termed 'individual-based', and when grouped individuals are used, the sample unit is referred to as 'sample-based'. In addition, a distinction is made between 'incidence data' and 'abundance data'. Most general collecting herbarium collections depict 'incidence data', also known as 'presence/absence data' or 'occurrence data', and simply indicate the existence of certain taxa at a defined location. 'Abundance data' has some link

between a taxon existing and its abundance. The data used in this study is individual-based incidence data.

Simple SACs can represent the observed accumulation as a product of the collecting order but more often they are produced as smooth, averaged curves through interpolation, also known as rarefaction (Ugland et al. 2003; Chao 2006). Rarefaction is a technique that allows for the comparison of species richness between assemblages in the presence of unequal sampling intensity. The process of creating rarefaction curves involves resampling the observed data multiple times and then plotting the average number of species found in *n* samples, with *n* ranging from one to the total number of samples (Gotelli & Colwell 2011). Rarefaction curves are generally seen to grow rapidly at first, as many of the common taxa are found, and then to flatten off as only rare taxa remain to be added (Colwell 2013; Gotelli & Colwell 2001). When the curve plateaus, the asymptote has been reached. The asymptote is a reflection of the estimated total species richness for the assemblage. The asymptote can either be achieved within the observed data or can be estimated by extrapolation. If the asymptote is reached before the reference sample size has been exhausted, collecting completeness is suggested to be complete. Extrapolation to an asymptote suggests how much more collecting effort might be required before collecting completeness has been reached (Soberón & Llorente 1993; Ugland et al. 2003).

SACs can be based on three different estimator methods: using mathematical functions set to the observed data, parametric models incorporating abundance across the dataset, and non-parametric models (Magurran 2004; Gotelli & Colwell 2011). Non-parametric estimators are best suited for most data types (Walther & Moore 2005; Chao 2006).

1.4 Study Area

The study area for this thesis is a geographical rectangle within the Sangha-Trinational area in tropical central Africa defined by 1.950 × 3.250 latitude and 15.200 × 16.999 longitude. The study site extends across three countries: Cameroon, Congo and Central African Republic. The study area is shown in Figure 1, which depicts maps created in ArcMap10 (ESRI 2011) .

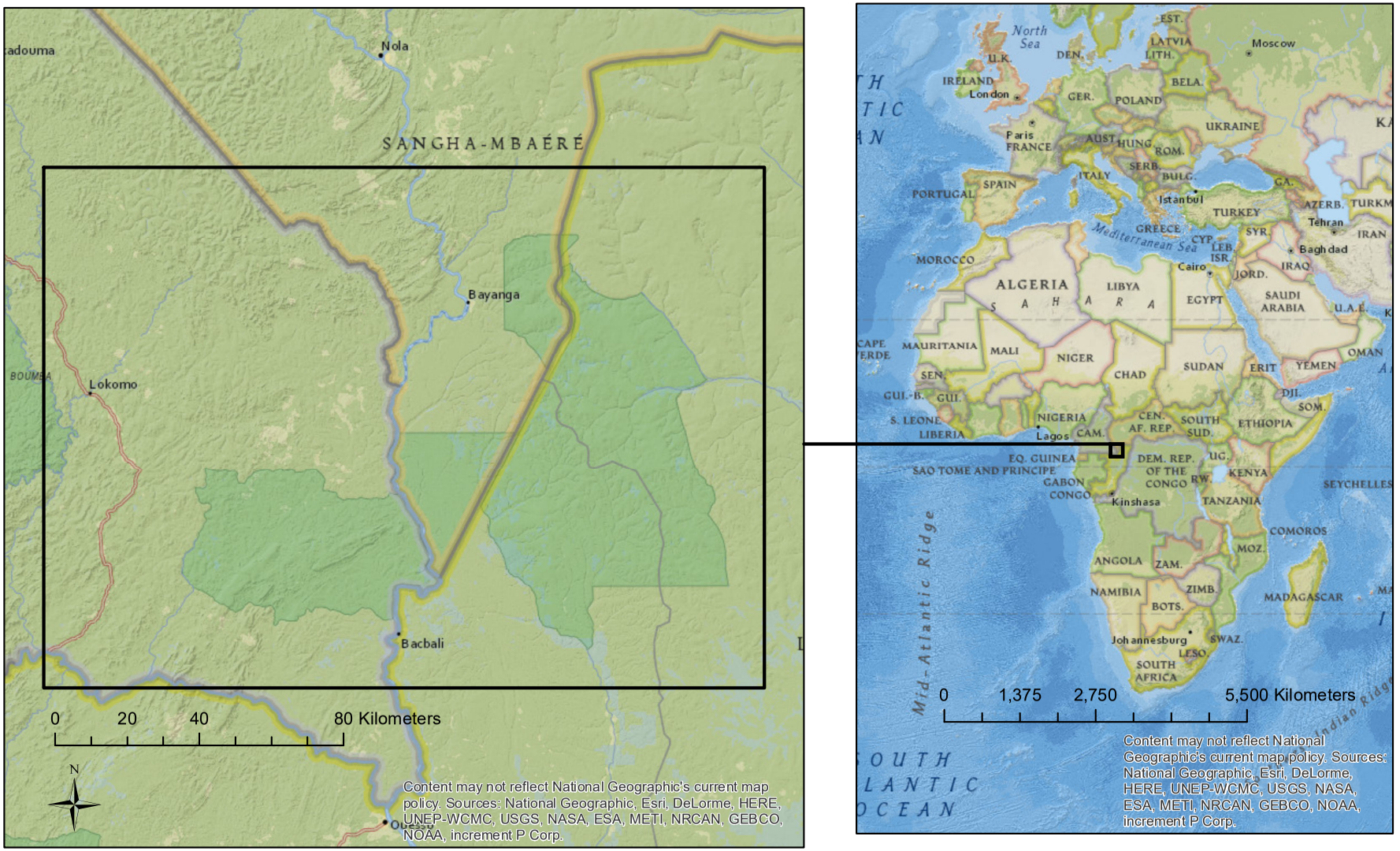

Figure 1. Maps showing the study area and its location within Africa. The study area is within the Sangha-Trinational area spanning the borders of Republic of Congo, Cameroon and Central African Republic. Maps created in ArcMap10 (ESRI 2011).

2 Aims and Objectives

This study focuses on exploring the flora of an area using a single dataset consisting of 19,394 vascular plant specimens collected from within the Sangha-Trinational area in Tropical Africa. This study aims to assess species richness, both observed and expected, using herbarium specimens and non-parametric estimators. Specimens were collected using general collecting methods rather than plot-based collecting methods. The objectives of this study are to (1) investigate the observed and expected species richness for different plant growth forms, (2) compare collection completeness for these different groups, (3) assess whether there is any evidence of collection bias for particular groups, and (4) evaluate the suitability of general collecting for estimating true species richness in a tropical floristic assemblage.

2.1 What is the Species Richness of the Flora?

2.1.1 All General Collecting Specimens

Whilst a raw, observed species richness figure of the vascular plants of the Dzanga-Sangha Reserve can be derived from its checklist (Harris 2002), no analysis exploring both observed and estimated species richness of the selected area has been undertaken. Based on the trend of tropical floras being undersampled (Prance et al. 2000), hypothesis $H_{1.1}$ was set:
($H_{1.1}$) Expected species richness is much greater than the observed species richness for the selected area.

2.1.2 Trees (dbh >10cm)

Research into the flora of the study area has been focused on trees (dbh >10cm) based on collections using plot-based and general collecting methods (Wortley & Harris 2015; Harris, Moutsamboté, et al. 2011; Harris & Wortley 2008; Peccoud et al. 2013). A recent study based on the same dataset used in this thesis concluded that for general collecting of combined tree and shrub taxa there is still a large gap between observed and estimated species richness (Fabriani 2015). In order to explore whether there are any differences in collection completeness between large woody groups and smaller, often understory woody taxa, this thesis chose to separate trees (dbh >10cm) and shrubs (woodiness dbh <10cm). Given that there has been a focus on collecting trees in the study area, as well as more generally in tropical forests (Prance et al. 2000), hypothesis $H_{1.2}$ is formulated as follows:

($H_{1.2}$) Observed and expected species richness for trees (dbh >10cm) are better matched than for other life forms.

2.1.3 Shrubs (woodiness dbh <10cm)

Understory taxa in global tropical forests are particularly undersampled (Prance et al. 2000). It was suggested (D.J. Harris pers. comm.) that both species richness and collecting completeness of the shrub group of this study were expected to be less than that of the tree group. Based on the above, hypothesis $H_{1.3}$ was set:

($H_{1.3}$) Expected species richness is greater than the observed species richness for the shrubs of the selected area.

2.1.4 Herbs

Collecting completeness of herbaceous taxa in global tropical forests is lower than that of woody groups (Knapp 2017; Prance et al. 2000). Based on this, hypothesis $H_{1.4}$ was set:

($H_{1.4}$) A large discrepancy exists between expected and observed species richness of herbs of the selected area.

2.1.5 Climbers

It was suggested that the observed species richness of taxa with a climbing habit was expected to be particularly low and less than the other habit groups (D.J. Harris pers. comm.). Based on this, hypothesis $H_{1.5}$ was set:

($H_{1.5}$) Expected species richness is greater than the observed species richness for the climbers of the selected area.

2.2 How Well Collected is the Flora?

2.2.1 All General Collecting Specimens

Taxonomic information covering the taxa of the area has progressed significantly due to work using the dataset relevant to this study. Prior to the checklist of the vascular plants of the Dzanga-Sangha Reserve compiled by Harris using this dataset (Harris 2002), the most comprehensive floristic information of the area was offered by the incomplete Flora of Tropical Africa (Oliver 1868). Additionally, some taxonomic and descriptive, but not distributional, information on taxa found within the area was partially covered in the Flora of West Tropical Africa (Hutchinson & Dalziel 1928), and by the incomplete volumes of Flore

d'Afrique Centrale (Bamps 1972), Flore du Cameroun (Aubreville 1961), and Flore du Gabon (Aubreville 1963). Despite over 25 years of intensive cataloging of the study area's flora, tropical floristic diversity patterns suggest that collecting completeness has yet to be reached for the area (Prance et al. 2000). Below is hypothesis $H_{2.1}$:

($H_{2.1}$) Collecting completeness of the vascular plants of this project's study area has not been reached.

2.2.2 Trees (dbh >10cm)

Since the production of the checklist of the vascular plants of the Dzanga-Sangha Reserve (Harris 2002), much of Harris' and his colleagues' work on the flora of the area has focused on trees using both general collecting and plot-based collecting (Wortley & Harris 2015; Harris, Moutsamboté, et al. 2011; Harris & Wortley 2008; Peccoud et al. 2013). Based on this biased research focus, hypothesis $H_{2.2}$ was set:

($H_{2.2}$) Collecting completeness of the trees (dbh >10cm) of this project's study area is greater than the overall flora's collecting completeness.

2.2.3 Shrubs, Herbs and Climbers

Compared with that of trees, the collecting completeness of shrubs, herbs, and climbers within the study area were poorly documented before this study. Relevant work includes the checklist of the vascular plants of the Dzanga-Sangha Reserve (Harris 2002), and various monographic work represented by revisions and new species descriptions (Lachenaud & Harris 2010; Dhetchuvi et al. 2011; Harris, Cheek, et al. 2011; Ley & Harris 2014; Ndolo Ebika et al. 2015). Due to the ease of collecting samples of taxa within the shrub group, the associated collection completeness was expected to be high (D.J. Harris pers. comm.). However, with less evidence of active or completed research into this group or taxa within it, within the area, it can be expected that if collection completeness has not yet been reached, it will be lower than that of trees. As discussed in section 1.4 of this paper, collecting completeness of herbaceous taxa in tropical forests is notoriously lower than that of woody groups (Knapp 2017; Prance et al. 2000). D.J. Harris further conveyed that collecting completeness and possibly species richness of climbers were expected to be particularly low and less than the other habit groups (D.J. Harris pers. comm.). Based on these groups appearing in less active or completed research than trees, and also on the expert expectations, hypotheses $H_{2.3}$, $H_{2.4}$, and $H_{2.5}$ were set:

($H_{2.3}$) Collecting completeness of the shrub (woodiness dbh <10cm) group within the

study area has been, or is close to being, reached.

($H_{2.4}$) Collecting completeness of the herbaceous-habit group within the study area is low and has not been reached.

($H_{2.5}$) Collecting completeness of the climbing-habit group within the study area is low and has not been reached.

2.3 Further Exploration Using the Herbaceous Flora

2.3.1 Poorly Identified Group (*Poaceae*)

Specimens of taxa from groups that are taxonomically challenging and/or which present difficulties in identification can have various effects on SACs. Changes in species concepts which affect the overall number of species represented by the specimens will alter SACs, and therewith conclusions on collecting completeness and observed and estimated species richness. Identification difficulties can manifest as specimens with incorrectly assigned species names or as specimens missing data at the species level. Both examples of identification difficulty will falsely influence analysis types used in this project, the former producing skewed results by the inclusion of false data and the latter effecting such analysis by data exclusion. When botanical research is carried out across a regional flora, scientists working on such projects will bring with them their prior knowledge and contribute with specific topics of expertise. Identification difficulty and the personal skills of the people involved will result in groups of records at specific taxa levels harboring more errors and gaps than others. The original dataset of 19,394 specimens used for this study includes ~1500 specimens across 102 different plant families not identified to the species level. Plant families with the highest representation of identification only to family or genus level are *Rubiaceae* Juss., *Malvaceae* Juss., *Annonaceae* Juss., and *Poaceae* Barnhart. It is expected that effort to overcome identification difficulties (see 3.5.1) will lead to different, more reliable results. These differences are expected to be significant and visible when analysis is carried out on small scale taxa groups, but possibly with knock-on effects up to habit-based group analysis.

($H_{3.1}$) SACs, collecting completeness and species richness results will differ when taken before and then after effort has been expended to overcome identification difficulties.

2.3.2 Monographic Collection (*Aframomum K.Schum.*)

Specimens of taxa collected, using general collecting methods, to contribute towards monographic works are thorough in their collecting completeness, but do not necessarily

represent abundance realities (Colwell et al. 2012). Monographic works require exhaustive collecting to capture the entire range of phenotypic variation present within a species. As such, species identified within collections generated for monographic works are often represented by a greater number of specimens than expected in collections established with unbiased general collecting for less focused inventory work. Based on this, hypothesis $H_{3.2}$ was set:

($H_{3.2}$) Monographic collections from general collecting methods will exhibit collecting completeness.

2.3.3 Verifying Techniques by Exploring Bad Singletons and Doubtful Doubletons

D.J. Harris identified many of the herb species recorded as singletons and doubletons as 'false', i.e. as not truly rare, or undersampled taxa (See Appendix 2). The reason for the presence of these false singletons and false doubletons was collecting bias which resulted in decisions to collect some taxa only once or twice even though they were encountered more frequently. Some explanation for these decisions are presented in Appendix 2. SACs created in this study and the species diversity measures used to extrapolate them to their asymptotes rely strongly on the influence of these singletons and doubletons. Based on the known collecting bias outlined above, $H_{3.3}$ was set:

($H_{3.3}$) The removal of "bad singletons" and "doubtful doubletons" will substantially alter SACs, and therefore conclusions on collecting completeness and observed and estimated species richness.

2.3.4 Verifying Sample Unit Choice by Comparing SACs Created Using Individual-Based and Sample-Based Sample Units.

Specimens as the sample unit for this dataset was deemed an appropriate choice as the dataset benefited from uniform collecting methods which aimed to avoid multiple successive collections of spatially clustered taxa (D.J. Harris pers. comm.). Sample-based sample units benefit datasets by accounting for collecting bias which has resulted in multiple records of truly rare, but spatially aggregated taxa (Gotelli & Colwell 2011). Because this type of potential collecting bias is not expected, Hypothesis $H_{3.4}$ was set:

($H_{3.4}$) Expected Species Diversity measures will be similar using either individual-based or sample-based sample units.

3 Methods and Materials

3.1 The Data and the Sample Unit

Data used in this project was collected, databased and curated by or with an experienced botanist and taxonomist (David J. Harris), who has invested more than 15 years carrying out field work in the study area (Harris 2002; Harris & Wortley 2008; Dhetchuvi et al. 2011; Harris, Moutsamboté, et al. 2011; Ley & Harris 2014; Ndolo Ebika et al. 2015; Lachenaud & Harris 2010; Peccoud et al. 2013; Wortley & Harris 2015). When doing general collecting, D.J. Harris, the major collector of the specimens defining the dataset used here, collected samples of all plants he encountered within the study area unless he knew he had collected them previously at least thrice. The total dataset comprises over 25,000 herbarium specimen records extracted from Harris' personal database, stored using the database managing software BRAHMS (Botanical Research and Herbarium Management System)(Filer 2012). Of these 25,000+ specimens, 19,394 are from within the study area. Additionally, plot collected voucher records were added from a supplementary dataset stored in Harris' personal Microsoft Excel Version 15.36 (2017) spreadsheet.

The dataset used in this study represents all records from fieldwork undertaken on the flora of the study area. The dataset has been used in compiling checklists (Harris 2002; Harris et al. 2012), identification guides (Harris, Moutsamboté, et al. 2011; Wortley & Harris 2015) and it contributed towards monographic works (Peccoud et al. 2013; Ley & Harris 2014; Lachenaud & Harris 2010; Ndolo Ebika et al. 2015).

The data used is incidence based data. The sample unit chosen for this study is individual specimens. This removes problems associated with assessing variable collecting effort across different collecting days or in different geographical areas. However, a potential problem arising from using specimens as the sample unit arises if taxa that are rare across the whole study area are collected more than once on the same collecting outing. In this situation, the rare taxa will appear more common than if the samples were aggregated by collecting day or geographical area. The project has therefore done an additional analysis for herbs using a sample-based approach, with individuals being grouped by sampling day. This was to assess

the magnitude of differences that arise when using sample-based versus individual-based approaches

3.2 Defining the Habit Groupings

Habit types were scored at the species level in this dataset by a regional expert, D.J. Harris. The categories included for this dataset are: tree, shrub, herb, climber, hemi-epiphyte, and hemi-parasite. Trees were defined as woody species with a dbh greater than ten centimeters. Shrubs were defined as woody species that are not recorded as having a dbh greater or equal to ten centimeters in this region. Hemi-epiphytes and hemi-parasites were included in the shrub group as each contained too few data alone to offer reliable results. All species within these two groups within this dataset comprise woody taxa (See Appendix 1). Herbs were defined as all vascular herbaceous species. Climbers were defined as woody and herbaceous climbing taxa. This dataset's climbing-habit group includes the subgroups lianas and vines. However, these have not been scored and are therefore only analyzed as a combined climbing-habit group. Not all habit-type groups are mutually exclusive in this dataset, with some taxa from the herb and shrub groups also found in the climbing-habit group.

Over 91% of the species within RAINBIO[1] can be categorized by nine growth form types: tree, shrub, herb, liana, vine, aquatic herb, epiphyte, mycoheterotroph and parasite (Sosef et al. 2017; Gilles et al. 2016). The assignment of these habit-type labels within RAINBIO is based on R-scripts searching for habit keywords in the description field of specimens, and on expert analysis of a portion not covered by the R-script habit assignation. Difficulties in multiple keywords appearing in descriptions were overcome by awarding the habit-type of a species to the category with the most representative specimens, and by allowing secondary habit-type assignations if the majority habit-type appeared in less than half the total species' specimen's description (Gilles et al. 2016). RAINBIO further categorizes these habit groups by woodiness (tree, shrub, liana), herbaceous (herb, vine, aquatic herb, and mycoheterotroph), and mixed (epiphyte and parasite). Species of mycoheterotrophs and aquatic herbs are found sunk within this dataset's herb group. Lianas and vines are combined in our dataset into our climbing-habit group.

[1] a "unique, comprehensive, mega-database of georeferenced records for vascular plants in continental tropical Africa" covering 25,356 native, tropical African vascular plant species representing ca. 89% of the known plant species in the area (Gilles et al. 2016).

3.3 Data Management Plan

A data management plan is key to ensure research data quality, traceability and repeatability. Sections 3.3.1 – 3.3.5 detail the data management standards employed in this project.

3.3.1 Identifying and Cleaning Errors and Missing Data within the Dataset

When a digitalized specimen ("record") is missing information in one or more critical data fields, the individual record must be traced back to the original hardcopy (unmounted material, mounted herbarium specimen and/or collecting notes by the collector) and the relevant information collected from there and digitally added. Some missing data, such as nomenclatural data, can be resolved without tracing back to the original hardcopy. With a collection of over 25,000 specimens, there are complications in tracing individual original hardcopy data sources. Records with error(s) or missing critical data not able to be resolved without the original hardcopy were excluded from the study.

Errors were limited in this dataset due to extensive prior analysis and consequently prior cleaning; and due to the dataset benefiting from a single, dedicated collector (D.J. Harris) controlling new entries, additions and revisions. Many error-types were avoided by the functions within the database managing software BRAHMS (Filer 2012).

3.3.2 Geographic Missing Data and Error

Records with error or missing data in the locality fields were automatically excluded as data exports from the Brahms database were restricted by latitude and longitude data to records falling within the prescribed area spanning across Cameroon, Republic of Congo and Central African Republic. Locality information was run through BG-BASE's geographic tool set to verify that latitude and longitude data fit within the corresponding country (Walter et al. 2017).

3.3.3 Taxonomic Missing Data and Error

Taxonomic checks were performed by D.J. Harris using International Plant Names Index (IPNI, 2015), the World Checklist of Selected Plant Families (WCSP, 2017), and various

monographs, checklists and floras e.g. (Hutchinson & Dalziel 1928). Due to the inbuilt functions of BRAHMS, uniformity in the spelling of taxa levels was ensured. The use of synonyms, unaccepted names, and other taxonomic anomalies were cleaned and checked by Harris. Records were checked against the African Plants Database (African Plants Database 2015). Specimens only identified to genus level were mostly excluded from this study. It was possible to keep specimens (records) with only genus identification if they had been further informally identified as unique species by D.J. Harris. Products of this informal method of identifying distinct, separate species without using accepted binomials are termed 'morphospecies' (D.J. Harris pers. comm.). There were 335 specimens representing 68 different morphospecies included in this study, a situation which would not be possible with any reliability using a dataset without such controlled input parameters.

3.3.4 Data Loop

Any data cleaning efforts need to be fed back to the original dataset so that corrections can be made to the permanent data source. Any changes or additions to the dataset will require rerunning of any analysis using the relevant data.

3.3.5 Filtering, Exporting and Preparing Data

Data were exported from BRAHMS (Filer 2012) using a geographical filter restricting the data to a geographical rectangle in tropical central Africa defined as 1.950 x 3.250 latitude / 15.200 x 16.999 longitude. Exports were then set by filtering for the scoring of the relevant habit-groups from within the study area. Specimens collected using plot collection methods were later excluded from this study to focus analysis on a collection of specimen records collected using general collecting methods. The dataset restricted to only specimens collected using general collecting methods contained 6,334 specimens representing 1,673 species. The collectors from within the study area were not restricted and included D.J. Harris, G. Moukassa, F.O. Nzolani Silaho, S.T. Ndolo Ebika, and 17 others. Data exported from BRAHMS (Filer 2012) were prepared in R (R Core Team 2017).

3.4 Species Richness and Species Accumulation Curves (SAC)

Species richness is a measurement of the number of species in a community. Shannon and Simpson diversity are two additional, widely used indices of species diversity that each combine information on richness and abundance (Magurran 2004). Shannon diversity uses natural and exponential logarithms and Simpson diversity is the inverse form. Due to the inherent interactions of richness and evenness, Simpson and then Shannon diversity measures will stabilize faster than richness (Colwell 2013). The Chao1 equation accounts for undersampling bias, and the species richness results it provides are the most intuitive for exploring the dataset type used in this study (Walther & Moore 2005; Gotelli & Colwell 2011; Chao 2006).

Species diversity extrapolation has been achieved with various success using mathematical functions, non-parametric estimation, and parametric models incorporating abundance across the dataset (Magurran 2004; Gotelli & Colwell 2011).

Rigid mathematical functions can be set to simple interpolation SACs to produce the extrapolation. However, these techniques remain statistically unsound, produce conflicting results depending on the equation chosen, do not account for undersampling bias, and fail to ensure that the rarefied interpolation curve smoothly meets the corresponding extrapolation curve at the reference sample size (Gotelli & Colwell 2011).

Non-parametric estimators are based on sample coverage principles (Good 2000; Good 1953). For most data types, species diversity estimation performance is best using non-parametric estimators, namely the Chao and Jackknife estimator models (Walther & Moore 2005; Chao 2006). The problem that rarefaction and extrapolation confidence intervals do not meet smoothly (and instead narrow at the observed sample limit and then widen again) has been addressed for the Chao1 (Chao 1984) non-parametric estimator thanks to new work by Colwell et al. (2012), incorporating probability distribution models. Reliable extrapolation can be achieved for up to double, with some datasets even triple, the size of the observed sample unit (Colwell et al. 2012; Chao & Jost 2012). Calculations of unconditional variance are possible for non-parametric estimators. These are able to combine variance derived from within the reference sample and extrapolated variance considering all samples leading to the

species diversity asymptote (Gotelli & Colwell 2011). Consequently, non-parametric estimators were deemed most appropriate for this project.

Of all non-parametric estimators, the statistical development of Chao 2 is probably most advanced. For the Chao1 model, the sample coverage principles were adapted by Colwell et al. (2012) to provide a statistical tool to calculate the estimated behaviour of the SAC past the observed sample size. Unlike its predecessors, the Chao model (Chao 1984) overcomes difficulties in producing smooth, reliable confidence intervals across the observed sample size limit. Because of the strength of the output confidence intervals and the smoothness of the rarefied and extrapolated curves it produces, the Chao1 model (Chao 1984) allows rigorous statistical comparison of different samples, along the length of their curves (Colwell et al. 2012).

The Chao1 model (Chao 1984) uses the following equation, which uses doubletons to reduce the otherwise exponential effect of singletons to predict asymptotic species richness:

$$\mathbf{S_{EST} = S_{OBS} + ((N - 1) / N) . (S^2 / (2 . D))}$$

S: singleton		$\mathbf{S_{OBS}}$: observed species richness
D: doubleton	**N**: sample size	$\mathbf{S_{EST}}$: estimated species richness

The term ‘singleton’ refers to a species represented by only a single sample unit. The term ‘doubletons’ refers to a species represented by two specimens. Infrequently sampled species are expressed as ‘singletons’ within a given dataset. In the case of general collecting, or more widely, individual-based incidence data, singletons equate to a species being represented by only a single specimen. The more singletons there are in a given dataset the larger the actual species richness is estimated to be using Chao 2 (as the presence of singletons may indicate undersampling). Specifically, a term is added to observed species richness which consists of the number of singletons squared, divided by twice the number of doubletons. (Doubletons are species that have been collected twice.) The presence of doubletons in the dataset thus mediates the exponential effect of singletons and therewith standardises for natural rarity.

Assumptions of rarefaction must be met to produce comparable SACs across different groups. Those assumptions are sampling sufficiency, uniform and independent sampling

methodology, taxonomic similarity, and random placement of species throughout each assemblage (Gotelli & Colwell 2011).

Species accumulation curves presented in this study were produced in the R (R Core Team 2017) package "iNEXT" (Chao et al. 2014) which uses the Chao1 model (Chao 1984).

3.5 Methods for Further Exploration Using the Herbaceous Flora Group

3.5.1 Overcoming Identification Difficulties

As the herbaceous plant family with the highest representation of identification difficulty within this dataset, *Poaceae* was chosen to explore differences in species diversity and SAC results produced before and after some of the missing identification was tackled. Five days in the herbarium at Royal Botanic Gardens, Kew were spent determining 59 different species-level identifications for the 108 grass specimens from D.J. Harris' collection. Before the identification efforts at Kew, 19 species were identified across 26 specimens from within the study area from within the dataset. Afterwards there were 41 species identified across 82 specimens.

Differences in apparent species diversity and collecting completeness taken before and after the identification efforts, were explored using the R (R Core Team 2017) package "iNEXT" (Chao et al. 2014) which uses the Chao1 model (Chao 1984).

3.5.2 Exploring Monographic Collections

Aframomum K.Schum. specimens were selected as the monographic collection from within the dataset to be analyzed. These were isolated so that the genus' species diversity and SAC could be explored.

3.5.3 Exploring Bad Singletons and Doubtful Doubletons

Lists of singletons and doubletons from within the herb group were generated using R (R Core Team 2017). This list was then checked by D.J. Harris to identify any known false

singletons and doubletons. Reasons given (D.J. Harris pers. comm.) for this known collecting bias include:

- Taxon collection difficulty (e.g. *Lasiomorpha senegalensis*)
- Known to be introduced, weedy, planted, or cultivated and therefore of only passing interest in the original study of the native flora of the area.
- Other unspecified reasons for only 'needing' one of them

Data for two new study groups were then created using R (R Core Team 2017). The first was a copy of the herb group with all the false singletons and doubletons removed, and the second was another copy of the herb group this time with the false singletons and doubletons turned into triple-tons and quadruple-tons respectively.

3.5.4 Verifying Sample Unit Choice by Comparing SACs Created Using Individual-Based and Sample-Based Sample Units

To test for potential collecting bias, two SACs for the herbs of the study area, using specimens (individual-based) and date (sample-based) as the different sample units, were compared. Differences in apparent species diversity and collecting completeness between the two sample units were explored using the R (R Core Team 2017) package "iNEXT" (Chao et al. 2014) which uses the Chao1 model (Chao 1984).

4 Results

4.1 What is the Species Richness of the Flora?

Figure 2 below shows the species accumulation curves (SAC) for each of the habit groups within the study area. The graph was created in the R (R Core Team 2017) package "iNEXT" (T. C. Hsieh et al. 2016) and plots specimens as the sampling unit against species diversity using Chao1 (Chao 1984) estimated species richness.

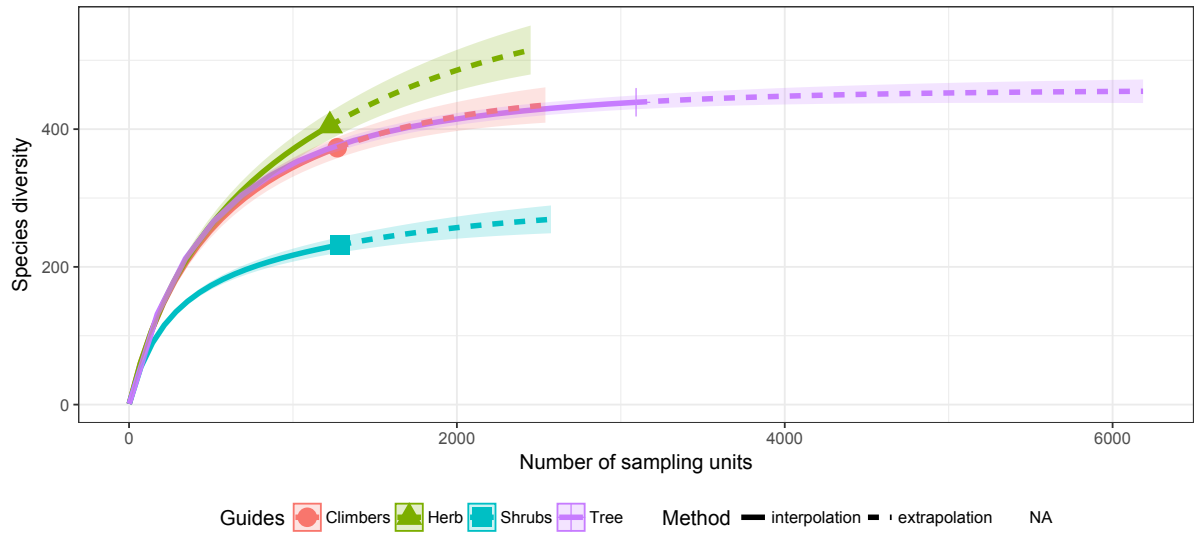

Figure 2. Species Accumulation Curves for each of the habit groups from within the study area. The graph was created in the R package "iNEXT" (T.C. Hsieh et al. 2016) *and plots specimens as the sampling unit against species diversity, using Chao1 estimated species richness* (Chao 1984).

Species richness represented by SACs which have not reached an asymptote are best compared by the shapes of the individual curves. Figure 2 shows that for this study area and dataset herbaceous taxa are both observed and expected to be the most species rich group. Climbers and trees have very similar species richness values and almost equally steep SACs. However, climbers are slightly more species rich than trees with a somewhat steeper gradient in the SAC. Shrubs appear to be the least species rich group and they also exhibit the shallowest SAC.

4.1.1 All General Collecting Specimens with Habit-Types Combined

($H_{1.1}$) Expected species richness is much greater than the observed species richness for the selected area.

Table 1 below shows the observed and expected species diversity results with confidence intervals obtained using the R package "iNEXT" (T. C. Hsieh et al. 2016). These results are for all general collecting specimens combined from across all habit-types. Observed species richness indicates the number of species within the study area already represented by the dataset. Expected species richness represents the estimated number of species expected, using the Chao1 model (Chao 1984), which extrapolates to an asymptote (T. C. Hsieh et al. 2016). Shannon and Simpson diversity are two additional, widely used indices of species diversity that each combine information on richness and abundance (Magurran 2004). The 95% Lower and Upper Species Richness figures represent the lower and upper confidence intervals given by Chao1 model (Chao 1984) extrapolations. Shannon and Simpson Diversity values are also given.

Table 1. iNEXT (T. C. Hsieh et al. 2016) *species diversity results produced in R* (R Core Team 2017) *for all General Collecting specimens from within the study area.*

Diversity	Observed	Expected	95% Lower	95% Upper
Species Richness	**1399.000**	**1759.957**	**1677.680**	**1866.525**
Shannon diversity	**944.515**	**1101.228**	**1073.117**	**1129.339**
Simpson diversity	**679.716**	**761.186**	**731.167**	**791.205**

These results show that based on the identified samples in the dataset an estimated 360 (278-467) species are still to be discovered in the study area, i.e. currently only 79.49% (74.95-83.39%) of the assemblage has been collected and identified. Figure 2 shows that with an extrapolation limit of double the sample unit size, no species richness asymptote is reached for the majority of the habit-types (with the exception of trees). Based on these results, hypothesis $H_{1.1}$ can be accepted.

4.1.2 Trees (dbh >10cm)

($H_{1.2}$) Observed and expected species richness for trees (dbh >10cm) are closely matched.

Table 2 shows the diversity results for trees.

Table 2. iNEXT (T. C. Hsieh et al. 2016) *species diversity results produced in R* (R Core Team 2017) *for General Collecting tree (dbh >10cm) specimens from within the study area.*

Diversity	Observed	Expected	95% Lower	95% Upper
Species Richness	**439.000**	**456.634**	**447.078**	**477.496**
Shannon diversity	**324.433**	**350.873**	**340.323**	**361.424**
Simpson diversity	**251.522**	**273.608**	**261.358**	**285.859**

According to these calculations an estimated 17 (8-38) species are yet to be discovered in the study area, whereas 96.14% (91.94-98.19%) have already been collected and identified. Based on these results, hypothesis $H_{1.2}$ can be accepted.

4.1.3 Shrubs (woodiness dbh <10cm)

($H_{1.3}$) Expected species richness is greater than the observed species richness for the shrubs of the selected area.

Table 3 shows the diversity results for shrubs.

Table 3. iNEXT (T. C. Hsieh et al. 2016) *species diversity results produced in R* (R Core Team 2017) *for General Collecting shrub (woodiness dbh <10cm) specimens from within the study area.*

Diversity	Observed	Expected	95% Lower	95% Upper
Species Richness	**232.000**	**289.972**	**262.095**	**343.668**
Shannon diversity	**142.247**	**160.853**	**150.610**	**171.095**
Simpson diversity	**83.795**	**89.492**	**83.795**	**100.775**

According to these calculations an estimated 57 (30-111) species are yet to be discovered in the study area, whereas 80.01% (67.51-88.52%) have already been collected and identified. Based on these results, hypothesis $H_{1.3}$ can be accepted.

4.1.4 Herbs

($H_{1.4}$) Discrepancy exists between expected and observed species richness of herbaceous taxa of the selected area.

Table 4 shows the diversity results for herbs.

Table 4. iNEXT (T. C. Hsieh et al. 2016) *species diversity results produced in R* (R Core Team 2017) *for General Collecting herbaceous specimens from within the study area.*

Diversity	Observed	Expected	95% Lower	95% Upper
Species Richness	**405.000**	**588.123**	**527.906**	**677.844**
Shannon diversity	**276.341**	**360.979**	**339.828**	**382.131**
Simpson diversity	**187.367**	**220.840**	**198.704**	**242.977**

According to these calculations an estimated 183 (122-272) species are yet to be discovered in the study area, whereas 68.86% (59.75-76.72%) have already been collected and identified. Based on these results, hypothesis $H_{1.4}$ can be accepted.

4.1.5 Climbers

($H_{1.5}$) Expected species richness is greater than the observed species richness for the climbers of the selected area.

Table 5 shows the diversity results for herbs.

Table 5. iNEXT (T. C. Hsieh et al. 2016) *species diversity results produced in R* (R Core Team 2017) *for general collecting climbing-habit specimens from within the study area.*

Diversity	Observed	Expected	95% Lower	95% Upper
Species Richness	**373.000**	**456.242**	**424.880**	**506.562**
Shannon diversity	**266.386**	**325.881**	**309.317**	**342.445**
Simpson diversity	**188.643**	**221.204**	**198.541**	**243.868**

According to these calculations an estimated 83 (51-133) species are yet to be discovered in the study area, whereas 81.75% (73.63-87.79%) have already been collected and identified. Based on these results, hypothesis $H_{1.5}$ can be accepted.

4.2 How Well Collected is the Flora?

Collecting completeness has been represented below in Table 6 by the percentage of the expected species richness, already captured as the observed species richness. All figures have been taken from the Observed and Expected Species Richness results obtained through the R (R Core Team 2017) package "iNEXT' using the Chao1 model (Chao 1984). As discussed throughout Section 4.1, for most habit-type groups, the species diversity asymptote was not reached using an extrapolation limit of double the size of the sample units (specimens).

Table 6. Collection Completeness of each habit-type group represented by the percentage of the Expected Species Richness already portrayed by the Observed Species Richness.

	All General Collecting	**Trees (dbh >10cm)**	**Shrubs (woodiness dbh <10cm)**	**Herbs**	**Climbers**
Collection Completeness	79.49%	96.14%	80.01%	68.86%	81.75%

These results reveal that trees (dbh >10cm) as a group have the highest level of collection completeness, followed, each with ~15% less collection completeness, by climbers, shrubs, and all general collecting specimens combined. With a further ~10% less collecting completeness, herbs come in last with the lowest representation of the group's assemblage's species richness covered by the observed species richness apparent in the dataset.

4.2.1 All General Collecting Specimens

($H_{2.1}$) Collecting completeness of the vascular plants of this project's study area has not been reached.

Figure 3 shows the species accumulation curves for all general collecting specimens within the study area, plotting specimens as the sampling unit against Chao1 (Chao 1984) estimated species richness (T. C. Hsieh et al. 2016)

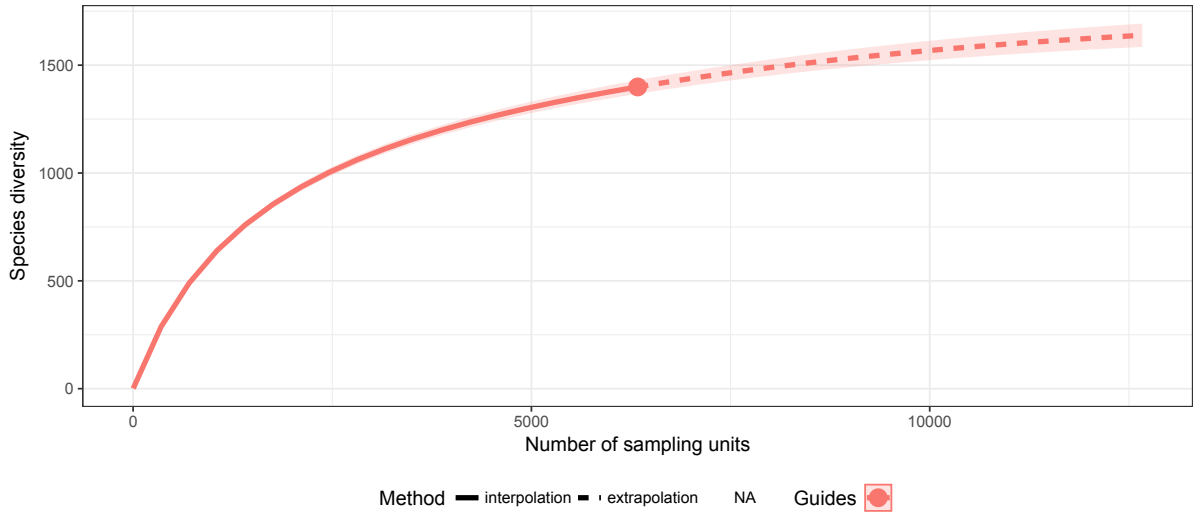

Figure 3. The species accumulation curve for all general collecting specimens, habit-types combined, from within the study area. The graph was created in the R (R Core Team 2017) *package "iNEXT"* (T.C. Hsieh et al. 2016) *and plots specimens as the sampling unit against species diversity, using Chao1 estimated species richness (Chao 1984).*

The nature of the solid interpolation curve reveals that the species richness asymptote for all general collecting specimens has not yet been reached. Figure 3 further shows by the continual growth of the dotted extrapolated curve that even with an extrapolation limit of double the sample unit size, the species richness asymptote for all general collecting specimens has not been reached. Because the asymptote has not quite been reached, the expected species richness indicated by the graph is not fully representative of the whole assemblage.

Based on these results, $H_{2.1}$ can be accepted.

4.2.2 Trees, Shrubs, Herbs, and Climbers

($H_{2.2}$) Collecting completeness of the trees (dbh >10cm) of this project's study area is greater than any other individual habit-based group explored, as well as greater than the overall flora's collecting completeness.

($H_{2.3}$) Collecting completeness of the shrub (woodiness dbh <10cm) group within the study area has been, or is close to being, reached.

($H_{2.4}$) Collecting completeness of the herbaceous-habit group within the study area is low and has not been reached.

($H_{2.5}$) Collecting completeness of the climbing-habit group within the study area is low

and has not been reached.

Figure 2 shows that the interpolated species richness curve for the general collecting tree group (purple) has come close to an asymptote. The extrapolated tree curve (purple dotted) shows that with an extrapolation limit of double the sample unit size, the species richness asymptote for tree specimens has been reached and therefore the expected species richness indicated by the SAC is representative of trees throughout the whole assemblage. In contrast to this, the interpolated and extrapolated species richness curves for the shrub (blue), herb (green), and climber (red) groups indicate that they have not reached their asymptotes with an extrapolation limit of double the sample unit size.

As seen in section 4.1.2, roughly 96.14% (91.94-98.19%) of the tree taxa within the assemblage has been collected and identified, making this group's collecting completeness greater than any other individual habit-based group explored, as well as greater than the overall flora's collecting completeness. Sections 4.1.3 and 4.1.5 showed that whilst less than the tree group, 80.01% (67.51-88.52%) of the shrub, and 81.75% (73.63-87.79%) of the climbing taxa within the assemblage have been collected and identified. Whereas section 4.1.4 showed that only 68.86% (59.75-76.72%) of the herbaceous taxa within the assemblage have been collected and identified, making this group's collecting completeness less than any other individual habit-based group explored, as well as less than the overall flora's collecting completeness combined.

Based on these results, hypothesis $H_{2.3}$ must be rejected and hypotheses $H_{2.2}$, $H_{2.4}$, and $H_{2.5}$ can be accepted.

4.3 Further Exploration Using the Herbaceous Flora Group

4.3.1 Poorly Identified Groups (*Poaceae*)

($H_{3.1}$) Analyses carried out at both the smaller-scale, taxa level and larger-scale, habit-based level will produce different SAC, collecting completeness and species richness results when taken before and then after effort has been expended to overcome identification difficulties.

Figure 4 and Table 7 below show iNEXT results and two different SACs, taken from the data before and also using the data available after effort was expended to overcome identification difficulties. The SACs are restricted to the grass (*Poaceae* Barnhart) specimens within the defined area of this thesis, plotting specimens as the sampling unit against Chao1 estimated species richness *(T. C. Hsieh et al. 2016; T.C. Hsieh et al. 2016)*.

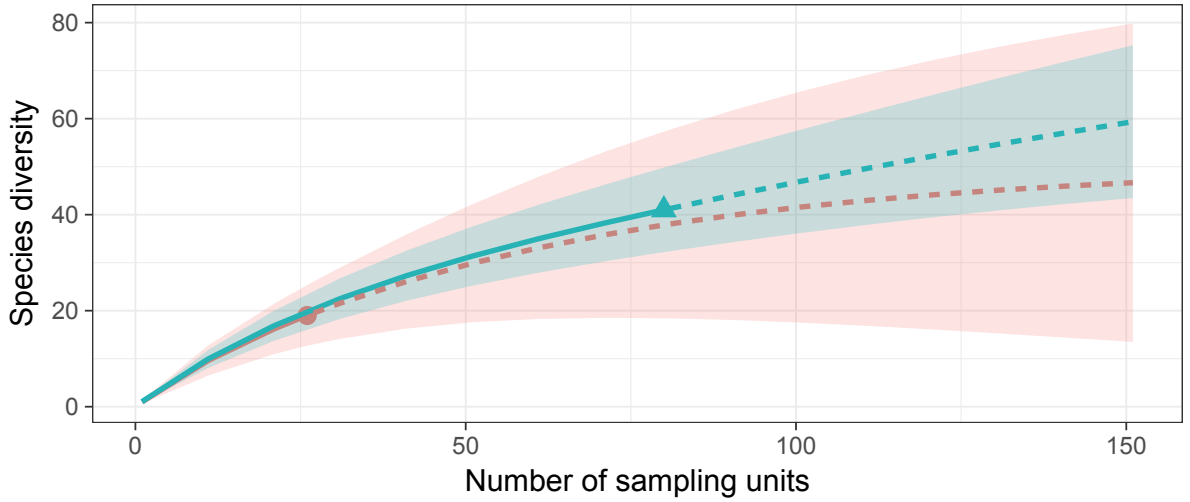

Figure 4. Species Accumulation curves of Poaceae Barnhart specimens before (red) and after (blue) identification effort. The graph was created in the R package "iNEXT" (T.C. Hsieh et al. 2016) *and plots specimens as the sampling unit against species diversity, using Chao1 estimated species richness (Chao 1984).*

Table 7 iNEXT (T. C. Hsieh et al. 2016) *species diversity results produced in R* (R Core Team 2017) *for general collecting* Poaceae Barnhart *specimens from within the study area.*

	Diversity	Observed	Expected	95% Lower	95% Upper
Before	**Species richness**	**19.000**	**50.410**	**26.841**	**144.820**
After	**Species richness**	**41.000**	**112.100**	**63.001**	**270.775**

The curves of Figure 4 and the results in Table 7 show that collecting completeness is low for *Poaceae Barnhart* specimens both before (37.69% (13.12-70.79%)) and after (36.57%(15.14-65.08%)) identification efforts. Figure 4 and Table 7 also reveal a weakness in reliability of both before and after groups, as both exhibit wide ranging confidence intervals. The SACs of

the two groups represent similar species richness patterns, with the "after' group showing slightly narrower, but still weak confidence intervals.

4.3.2 Monographic Collection (*Aframomum K.Schum.*)

($H_{3.2}$) Monographic collections from general collecting methods will exhibit collecting completeness.

Figure 5 below shows the SAC for a monographic collection of *Aframomum K.Schum.* specimens collected using general collecting methods from within the defined area of this thesis, plotting specimens as the sampling unit against Chao1 estimated species richness *(T. C. Hsieh et al. 2016; T.C. Hsieh et al. 2016)*.

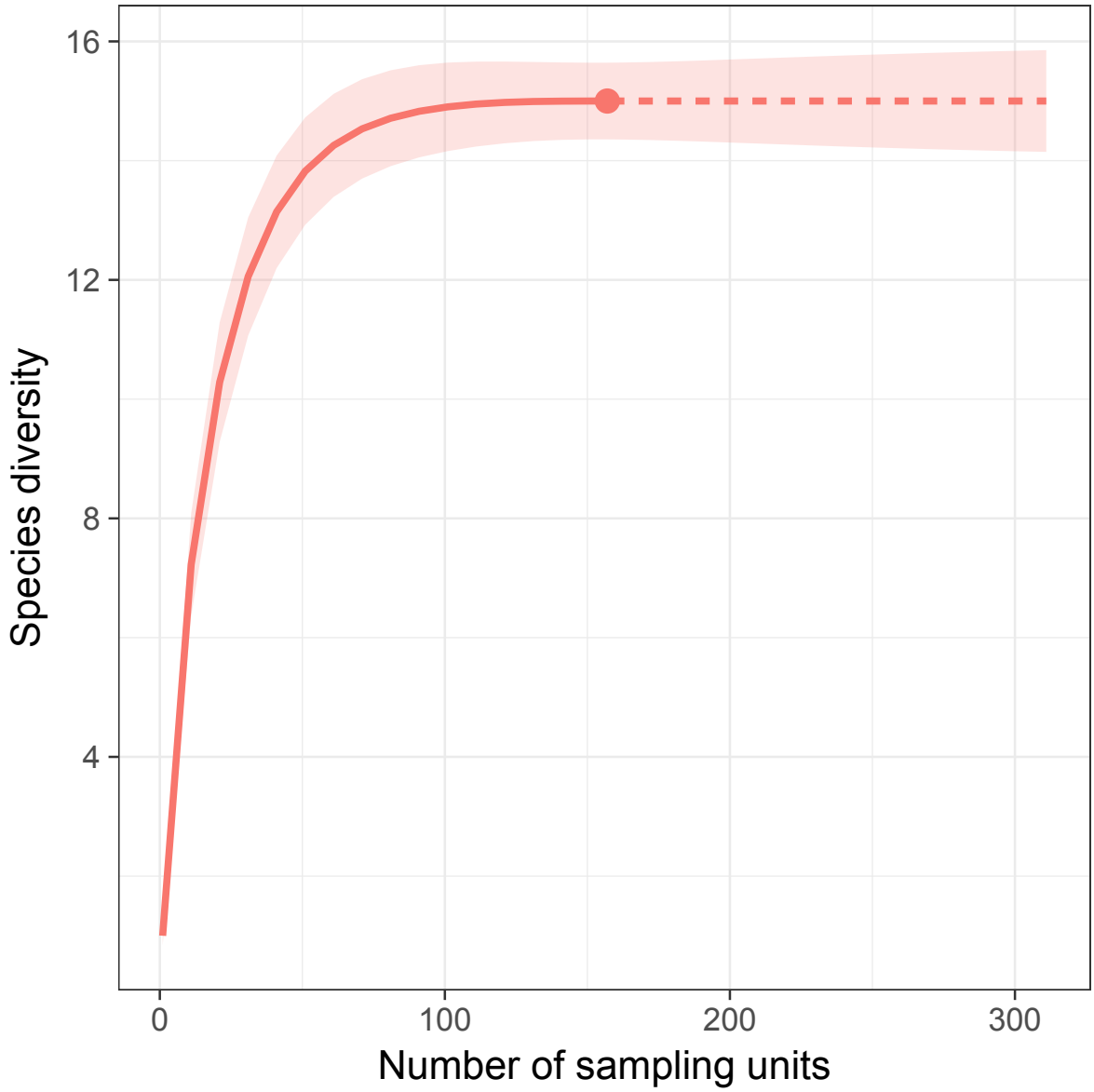

Figure 5. Species accumulation curve for a monographic collection of Aframomum K.Schum. ***specimens*** *collected using general collecting methods from within the study area. The graph was created in the R package "iNEXT"* (T.C. Hsieh et al. 2016) *and plots specimens as the sampling unit against species diversity, using Chao1 estimated species richness* ***(Chao 1984)****.*

Table 8 below shows the observed and expected species diversity results with confidence intervals obtained using the R (R Core Team 2017) package "iNEXT" (T. C. Hsieh et al. 2016). These results are for general collecting *Aframomum K.Schum.* specimens.

Table 8. iNEXT (T. C. Hsieh et al. 2016) *species richness results produced in R* (R Core Team 2017)*, using Chao1 methods* (Chao 1984)*, for general collecting Aframomum K.Schum. specimens from within the study area.*

Diversity	Observed	Estimated	95% Lower	95% Upper
Species Richness	**15.000**	**15.000**	**15.000**	**15.764**

The curves of Figure 5 and the results of Table 8 show that collecting completeness has been reached (100% (95.15-100%)) for the monographic *Aframomum K.Schum.* collection.

4.3.3 Verifying Techniques by Exploring false singletons "Bad Singletons" and false doubletons "Doubtful Doubletons"

($H_{4.1}$) The removal of "bad singletons" and "doubtful doubletons" will drastically alter SAC, and therefore estimates of collecting completeness and expected species richness.

Table 9 below shows the observed and expected species diversity results with confidence intervals obtained using the R (R Core Team 2017) package "iNEXT" (T. C. Hsieh et al. 2016). These results are for three versions of the general collecting herbaceous-habit specimens. The "Original" version contains the herb specimen data as obtained from D.J. Harris' dataset. The "Bad single/doubletons removed" version has all singletons and doubletons identified as false by D.J. Harris using the criteria set out in Section 3.6, removed. The "Bad single/doubletons added again twice" version is the "Original" herb data with all singletons and doubletons identified as false by D.J. Harris using the criteria set out in Section 3.6, added again twice with unique collection identifiers assigned. This final version effectively sees all the bad singletons and doubtful doubletons as triple-tons and quadruple-tons respectively.

Table 9. Chao1 (Chao 1984) *species richness for three versions of the general collecting Herb specimens from within the study area.*

Version of Herbs	Diversity	Observed	Estimator	LCL	UCL
Original	**Species richness**	**405.000**	**588.123**	**527.906**	**677.844**
Bad single/doubletons removed	**Species richness**	**355.000**	**506.744**	**452.247**	**591.780**
Bad single/doubletons added again twice	**Species richness**	**405.000**	**556.760**	**502.258**	**641.805**

Table 10 below shows the collecting completeness of each version of the herb specimens. Collecting completeness has been represented by the percentage of the expected species richness, already captured as the observed species richness. All figures have been taken from the observed and expected species richness results obtained through the R (R Core Team 2017) package "iNEXT' (T. C. Hsieh et al. 2016) using the Chao1 model (Chao 1984). For all three versions, the species diversity asymptote was not reached using an extrapolation limit of double the size of the sample units (specimens).

Table 10. Collection Completeness, with confidence intervals, of three versions of the herb specimen data, represented by the percentage of the Expected Species Richness already portrayed by the Observed Species Richness.

	Original	Bad single/doubletons removed	Bad single/doubletons added again twice
Collection Completeness	**68.86%**	**70.06%**	**72.84%**
Confidence Intervals	59.75 - 76.72%	59.99 - 78.50%	63.10 – 82.63%

These results show that estimates of collection completeness increase when the "bad singletons" and "doubtful doubletons" are addressed. This increase is minimal and compared to the results seen in Table 6, all three versions of the herb-habit group exhibit the lowest levels of collection completeness across the different habit-type groups.

Figure 6 below shows the SACs for each of the habit groups within the study area (as seen in Figure 2) as well as the two new versions of the herb group as discussed above. The graph was created in the R (R Core Team 2017) package "iNEXT" (T. C. Hsieh et al. 2016) and plots specimens as the sampling unit against species diversity using Chao1 (Chao 1984) estimated species richness.

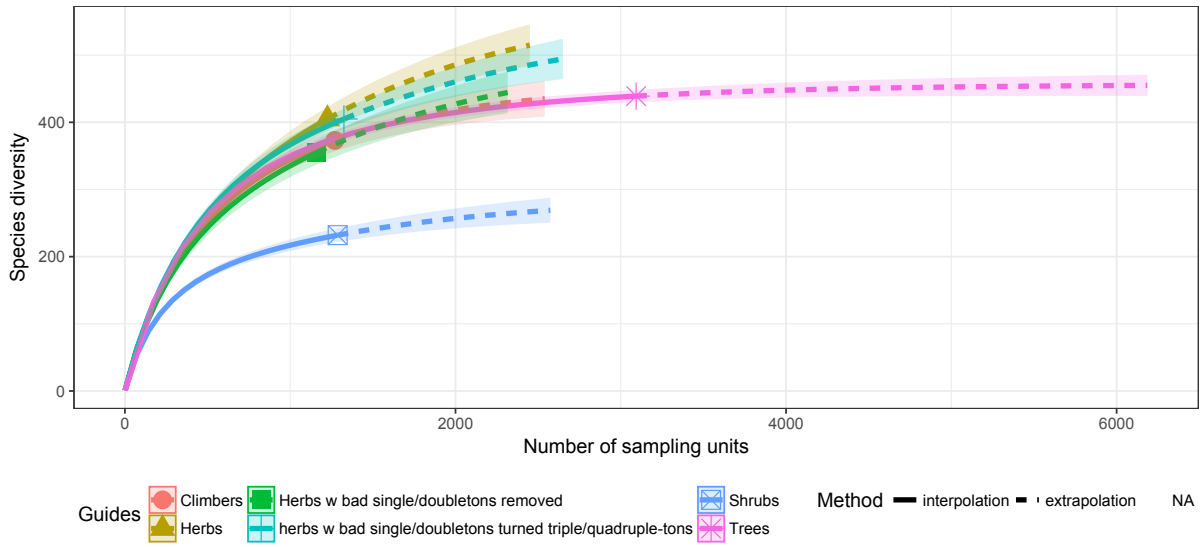

Figure 6. Species Accumulation Curves for each of the habit groups from within the study area, as well as the two additional versions of the herb group as discussed in this Section. The two new versions are the herb group with the 'false' singletons and 'false' doubletons changed to triple-tons and quadruple-tons respectively, and the herb group with the 'false' singletons and 'false' doubletons removed. The graph was created in the R (R Core Team 2017) *package "iNEXT"* (T.C. Hsieh et al. 2016) *and plots specimens as the sampling unit against species diversity, using Chao1 estimated species richness (Chao 1984).*

Figure 6 reveals that corrections made within the herb group, addressing the problems of collection bias resulting in bad singletons and doubtful doubletons, do not alter the SACs so drastically as to change the order of most-to-least species rich habit-groups. The shape and steepness of the 'Herbs' (brown) and 'Herbs w bad single/doubletons turned triple/quadruple-tons' (pale blue) SACs are similar, whilst the 'Herbs w bad single/doubletons removed'(green) SAC drops significantly.

4.3.4 Verifying Sample Unit Choice by Comparing SACs Created Using Individual-Based and Sample-Based Sample Units

Table 11 shows the two different diversity results for herbs using either individual-based or sample-based sample units.

Table 11. Chao1 model (Chao 1984) *iNEXT* (T. C. Hsieh et al. 2016) *species diversity results produced in R* (R Core Team 2017) *for general collecting herb specimens from within the study area, using both individual-based and sample-based sample units..*

Herb sample unit	Diversity	Observed	Estimated	LCL	UCL
date	Species richness	401.000	575.273	517.505	661.684
specimens	Species richness	405.000	588.123	527.906	677.844

Estimated species richness compared between the two sample units shows that there is less than three percent (575.273/588.123*100) difference between the asymptotic species richness estimations of the two sample-unit types.

5 Discussion

The aim of this research project was to assess species richness and sample completeness for vascular plants within a diverse lowland tropical forest area. The results highlighted that despite 25 years' worth of collecting effort, vascular plants have still not been exhaustively sampled. In addition, there was evidence for taxonomic bias with groups such as trees being better sampled than herbs and other under-storey growth types. Potential methodological issues (sections 5.1 to 5.5) and wider implications of this work (sections 5.8 to 5.10) are discussed below.

5.1 The Input Data

The dataset used in this study was compiled over 25 years and provides an extensive inventory of the plant diversity of the study area. Whilst this unique resource was mostly well suited for analyzing the questions addressed in this thesis, a potential caveat is that species diversity information inputted over such an extended period is at risk of over-representing current diversity levels due to failing to account for species turnover (Mora et al. 2011). Comprehensive floristic inventories of an area are necessary before any changes in species diversity dynamics can be assessed.

5.2 Identification of the specimens

Identification effort at Kew covered 108 specimens and resulted in the determination of 59 different *Poaceae* species. This allowed for the inclusion of an extra 56 specimens representing 22 species not previously included, from within the study area. After the Kew identification, 1406 specimens without any species level identification and 1199 morphospecies remained within the dataset. All together specimens with missing or unaccepted identification information at the species level or higher represented 417 genera from over 100 plant families. If the rate of identification (21.6 specimens/day), as exhibited by the week at Kew spent on *Poaceae* specimens, is applied across the remaining specimens without accepted identification to species level, it would take over 120 days to complete the identification of this dataset. This is not taking into account differences in identification difficulty between, or the steep learning curves inherent in tackling, different family or genus

level groups. This rate of identification further fails to account for the fact that the *Poaceae* identification was possible only due to the combined input of an MSc taxonomy student (S.S. Garrett) and a regional plant generalist expert (D.J. Harris), and the advice of a world *Poaceae* expert (Dr. Maria Vorontsova).

Over half of the estimated undescribed plant species are believed to already exist in herbaria (Bebber et al. 2010). It is unknown whether any undescribed species remain lurking within the dataset used in this study, however the dataset has been used for new plant descriptions previously (Lachenaud & Harris 2010). The time-consuming process of identification is significantly amplified when specimens have the potential to link with poorly circumscribed or unknown taxon names. The expertise required to unravel such identification and nomenclatural chaos is high. *Poaceae* specimens with species determinations, and the species they represent more than doubled for the study area after identification efforts at Kew. Without further exploration of this, it is difficult to suggest whether an enhanced understanding of the species diversity of the study area would be most effectively achieved through additional collecting or by focusing efforts on identifying the 2605 specimens without accepted species-level information already collected from the area. Such recommendations are beyond the scope of this study but could have wide-reaching implications for future studies of regional species diversity and in tackling the estimated 70,000 plant species still to be described

5.3 Habit-Type Classification

In this study habit-types were categorized by a regional expert, and the categories used did not exactly overlap with the categories used in RAINBIO (Gilles et al. 2016) – an Africa wide database of plant records. Uniformity to habit classification would facilitate wider ranging comparison of the results of this study to similar future analysis on other datasets. Future work of this type on this dataset would benefit from habit-type classification synchronizing with the more comprehensive nine growth form types used in RAINBIO (Gilles et al. 2016). Standardization incorporating multiple languages, of habit-type keywords tied to specimen/record data at the time of collection, and subsequently digitalization, would facilitate accurate and efficient analyses of this type across larger datasets. Fabriani 2015 found that classification of specific habit-types using keywords from within the dataset led to

34% inaccuracy (Fabriani 2015). Habit-type classification used in this study was therefore only possible using a regional flora expert's (D.J. Harris) time and expertise.

5.4 Selection of the Sample Unit

When extrapolating richness estimates a decision has to be made as to whether to treat the data as individual-based or sample-based. In this study, all data were treated as individual-based incidence data, thus the sample units are individual specimens. This could lead to biases when rare species exhibit spatial clustering in the environment or when sampling has been spatially clustered. Sample-based approaches whereby the data are aggregated by time (e.g. days) or geographical location would effectively filter out the effects of multiple spatially or temporally aggregated collections of truly rare species (Gotelli & Colwell 2011). However, in this study specimens as the sample unit were deemed an appropriate choice as the dataset benefited from uniform collecting, aiming to avoid multiple successive collections of spatially clustered taxa (D.J. Harris pers. comm.). A sensitivity analysis highlighted that there was less than three percent difference in the results between individual-based and sample-based approaches. Specimens are simply the most intuitive unit when analyzing herbarium data, and there was little difference in the results between sample-based and individual-based approaches.

5.5 Assumptions of Rarefaction

For SACs to be reliable and comparable across different samples, the assumptions of rarefaction (Gotelli & Colwell 2011; Colwell & Coddington 1994) must be met. Sampling sufficiency is a sample unit size of at least half the expected species richness asymptote. All samples explored in this study, except for the *Poaceae* group, meet sampling sufficiency assumptions. The expectation of uniform sampling methods within and across the samples requires consistent collecting approaches. Plot-based sampling was excluded from this study and all general collecting specimens were collected with uniform methodologies by all 21 collectors. The assumption of taxonomic similarity refers to an inability to clearly compare SACs between taxonomic groups which are extremely dissimilar e.g. algae and vascular plants.

The SACs of *Poaceae* specimens shown in Figure 4 and the species diversity results shown in Table 7 must be viewed with caution. Both have less than half the expected species richness represented by the reference sample and therefore both fail the assumption of rarefaction that sampling sufficiency has been achieved (Gotelli & Colwell 2011; Colwell & Coddington 1994).

5.6 Species Diversity and Accumulation Curve

Extrapolation was limited to double the sample unit size for all SACs throughout this study. Advice in the literature clearly supports this limit, but there is also mention of "up to triple the sample unit size" being reliable (Colwell et al. 2012; Chao & Jost 2012). We are not aware of any work verifying the statistical soundness of an extrapolation limit of double (or triple) the sample unit size. In recent work e.g.(Longino & Colwell 2011) SACs exploring multiple groups plotted against each other tend to be extrapolated to double the largest sample unit size (Colwell 2013). Reduced reliability in the extrapolations exceeding double their sample unit size is reflected in wider confidence intervals. It is important to note that comparison of various SACs in this study can only be made at the point of the lowest extrapolation limit. This results in a necessary loss of data for any curve extending past this limit. Such loss of data was a main critique of using SACs to compare species diversity between different communities (Magurran 2004). These issues were before Colwell et al. (2012) provided the statistical tools to link interpolation and extrapolation. The Chao1 model (Chao 1984) is an asymptotic estimator (Colwell 2013). Therefore, whilst the SACs have been set in this study with an extrapolation limit of double the sample unit size, the iNEXT (T. C. Hsieh et al. 2016) species diversity results produced in R (R Core Team 2017) are comparable.

Calculations for the expected asymptote using the non-parametric Chao1 model (Chao 1984) equation produce the expected species richness results seen in Tables 1-5 and Tables 7-9. The extrapolated SACs presented in this study use probability distribution models fitted to the interpolated curves to predict the shape of extrapolation curves leading to the calculated asymptote (Colwell et al. 2012).

5.7 Potential caveats and gaps in this work

The expected species richness asymptotes for the three herb groups used in section 4.3.3 have less than 15% difference between them. The results indicate that herbs remain as the habit group with the lowest collecting completeness and the highest expected species richness irrespective of potential collecting bias. The group consisting of herbs with bad singletons and doubtful doubletons removed is the least representative of any species diversity reality as 50 species and associated specimens have been effectively deleted despite it being known that they exist in the study area. The herb group most representative of the area's species diversity is the group with the bad singletons and doubtful doubletons transformed into frequently sampled species. The results of the SACs compared between habit-types (Figures 2 and 6) can be viewed as reliable as the Chao1 model (Chao 1984) has been seen to adequately account for potential collecting bias.

5.8 Suggested Future Research

Concepts and techniques of testing for the precision and accuracy of the estimator chosen, as discussed and performed by Walther and Moore 2005, are beyond the scope of this study. Future work could use similar techniques as well as simulations to test the performance of the species richness indicator (Chao 1984) used (Walther & Moore 2005).

5.9 Implications for Collecting Strategies

Collecting completeness has not been reached for the flora of the study area, nor for any of the habit-types explored in this study. The results shown in Table 6 indicate that herbs are the group with the least species richness representation already collected, and yet Figure 2 and Table 4 showed that they have the highest level of expected species richness. Shrubs appear to be the least species rich group and they also exhibit the shallowest SAC. Trees show the highest rate of collecting completeness. An increasing focus on ecosystem services results in collecting bias favouring trees with direct links to economy or carbon stocks (Gibbs et al. 2007; Bustamante et al. 2016). If biodiversity is to be accurately assessed, however, this study shows that it is in habit-type groups other than trees where the highest levels of unknown species richness are. These results are in line with global patterns of collecting bias (Gentry &

Dodson 1987; Knapp 2017; Prance et al. 2000) and suggest that future collecting priorities for efforts to assess vascular plant biodiversity be set using richness estimators.

Furthermore, the importance of identifying collections cannot be over-estimated as unidentified collections are effectively meaningless to studies of species diversity. With only common taxa identified, we fail to recognise what rare species are present, and also there may be a tendency to under-estimate richness with non-parametric estimators when the data is skewed towards species that occur frequently.

5.10 Implications for Conservation Priority Setting

National and global conservation priority setting is often based on species diversity assessments taken from counts of species in an area, or expert opinion. A study (Ahrends et al. 2011) focusing on the Eastern Arc Mountains in Tanzania found that alongside funding, observed species richness confounds issues inherent in identifying conservation priority areas. As seen by the SACs and species richness estimations presented in Section 4, predicted species richness far exceeds levels observed using only reference sample sizes. Approaches to assessing species diversity used to inform conservation priority setting would benefit from further analysis using species richness estimator models so that clearer reflections of species diversity are taken into account. To facilitate this however, adequate levels of initial inventorying are necessary to ensure the sampling sufficiency assumption of rarefaction is met.

6 Conclusion

This research project utilized over six thousand general collecting records to study species richness, sample completeness and taxonomic bias within an area of lowland tropical forest straddling the Republic of Congo, Cameroon and Central African Republic. The study showed that collecting completeness has not been reached for the vascular flora of the study area. There was also evidence for a taxonomic bias, with herbs having the least species richness representation already collected, whilst being estimated to have the highest level of species richness. Shrubs appeared to be the least species rich group and they also exhibited the shallowest SAC. Trees showed the highest rate of collecting completeness and had already come very close to collecting completeness.

The project also highlighted a range of methodological issues that deserve attention when using herbarium data to explore species diversity. Firstly, decisions must be made as to whether to treat the data as sampled-based (e.g. aggregated by sampling days or other units) or individual-based. A sample-based approach is typically more conservative in that the species accumulation curves will rise less steeply, but an individual-based approach is more intuitive. This study showed that for herbs the difference between the two estimations was less than 3%, thus largely negligible within the context of a highly biodiverse system. Secondly, the project showed that the assumptions of rarefaction must be understood and met to produce reliable and comparable species accumulation curves. Finally, the project highlighted that due to the taxonomic collecting bias it is informative to analyse species richness separately for different habit-groups. Habit-type classification used in this study was only possible using a regional flora expert's time and expertise. Standardization incorporating multiple languages, of habit-type keywords tied to specimen/record data at the time of collection, and subsequently digitalization, would facilitate accurate and efficient analyses of this type across larger datasets.

Overall, the project has shown that herbarium specimen data are a robust tool for exploring species diversity when appropriate methods are applied. The Chao1 method (Chao 1984) adequately accounted for potential collecting bias and produced reliable species accumulation curves and estimates of asymptotic species diversity. These methods however assume that collected specimens of rare species are identifiable. Identification efforts for large datasets are massive. It is difficult to suggest what would be more profitable to building on our

understanding of the species diversity of the study area: to undertake additional collecting, or to focus efforts on identifying the 2605 specimens without accepted species-level information already collected from the area. It seems clear that both approaches to improving our understanding of tropical floras across the world are vital. Finally, the use of species diversity estimators is fundamental to more accurately explore global biodiversity.

7 References and Bibliography

African Plants Database, 2015. African Plants Database (version 3.4.0). Available at: http://www.ville-ge.ch/musinfo/bd/cjb/africa/.

Ahrends, A. et al., 2011. Funding begets biodiversity. *Diversity and Distributions*, 17(2), pp.191–200.

Aubreville, 1961. *Flore du Cameroun*, National Herbarium of the Netherlands.

Aubreville, 1963. *Flore du Gabon*, National Herbarium of the Netherlands.

Bamps, 1972. *Flore d'Afrique Centrale*, National Botanic Garden of Belgium.

Bebber, D.P. et al., 2010. Herbaria are a major frontier for species discovery. *Proceedings of the National Academy of Sciences of the United States of America*, 107(51), pp.22169–22171.

Bustamante, M.M.C. et al., 2016. Toward an integrated monitoring framework to assess the effects of tropical forest degradation and recovery on carbon stocks and biodiversity. *Global Change Biology*, 22(1), pp.92–109.

Chao, A., 1984. Nonparametric Estimation of the Number of Classes in a Population. *Scandinavian Journal of Statistics*, 11(4), pp.265–270.

Chao, A. et al., 2014. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecological Monographs*, 84, pp.45–67.

Chao, A., 2006. Species Estimation and Applications. *Encyclopedia of Statistical Sciences*, (1), pp.1–23. Available at: http://doi.wiley.com/10.1002/0471667196.ess5051.

Chao, A. & Jost, L., 2012. Coverage-based rarefaction and extrapolation: Standardizing samples by completeness rather than size. *Ecology*, 93(12), pp.2533–2547.

Colwell, R.K., 2013. EstimateS 9.1.0 User's Guide.

Colwell, R.K. et al., 2008. Global warming, elevational range shifts, and lowland biotic attrition in the wet tropics. *Science*, 322(5899), pp.258–261.

Colwell, R.K. et al., 2012. Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology*, 5(1), pp.3–21.

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

Corlett, R.T., 2012. Climate change in the tropics: The end of the world as we know it?

Biological Conservation, 151(1), pp.22–25. Available at: http://dx.doi.org/10.1016/j.biocon.2011.11.027.

Dhetchuvi, J.B., Wortley, A.H. & Harris, D.J., 2011. A new species of Aframomum (Zingiberaceae) from Central Africa. *Phytotaxa*, 28, pp.31–34.

Dick, C.W. & Kress, W.J., 2009. Dissecting Tropical Plant Diversity with Forest Plots and a Molecular Toolkit. *BioScience*, 59(9), pp.745–755.

ESRI, 2011. ArcGIS Desktop: Release 10.

Fabriani, F., 2015. *" What is the best sampling strategy to characterise a tropical forest and estimate its species richness ?" GENERAL COLLECTING vs PLOT SAMPLING*. The University of Edinburgh.

Filer, D.L., 2012. BRAHMS (Botanical Research And Herbarium Management System) Version 7.0 Software. Available at: http://herbaria.plants.ox.ac.uk/bol/.

Fisher, R.A., Corbet, A.S. & Williams, C.B., 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, pp.42–58.

Garcillán, P.P. & Ezcurra, E., 2011. Sampling procedures and species estimation: Testing the effectiveness of herbarium data against vegetation sampling in an oceanic island. *Journal of Vegetation Science*, 22(2), pp.273–280.

Gentry, 1988. Tree species richness of upper Amazon forests. *Proceedings of the National Academy of Sciences of the United States of America*, 85, p.156–159.

Gentry, A. & Dodson, C., 1987. Contribution of Nontrees to Species Richness of a Tropical Rain Forest. *Biotropica*, 19(2), pp.149–156.

Gibbs, H.K. et al., 2007. Monitoring and estimating tropical forest carbon stocks: making REDD a reality. *Environmental Research Letters*, 2(2007), p.45023.

Gilles, D. et al., 2016. RAINBIO: a mega-database of tropical African vascular plants distributions. *PhytoKeys*, 74, pp.1–18. Available at: http://phytokeys.pensoft.net/articles.php?id=9723.

Good, I.J., 1953. The Population Frequencies of Species and the Estimation of Population Parameters. *Biometrika*, 40(3–4), pp.237–264.

Good, I.J., 2000. Turing's Anticipation of Empirical Bayes in Connection with the Cryptanalysis of the Naval Enigma. *Journal of Statistical Computation and Simulation*, 66(2), pp.101–111.

Gotelli, N. & Colwell, R., 2011. Chapter 4: Estimating species richness. *Biological Diversity. Frontiers in Measurement and Assessment*, (2), pp.39–54. Available at:

http://www.uvm.edu/~ngotelli/manuscriptpdfs/Chapter 4.pdf.

Gotelli, N.J. & Colwell, R.K., 2001. Quantifying biodiversity: procedures and pitfalls in the measurment and comparison of species richness. *Ecology Letters*, 4(May 1988), pp.379–391.

Harris, D.J., Moutsamboté, J.-M., et al., 2011. *An introductions to the Trees from the North of the Republic of Congo.*, Royal Botanic Gardens Edinburgh.

Harris, D.J. et al., 2012. Phytogeographical analysis and checklist of the vascular plants of Loango National Park, Gabon. *Plant Ecology and Evolution*, 145, pp.242–257.

Harris, D.J., 2002. *The Vascular Plants of the Dzanga-Sangha Reserve.*, Meise, Belgium: National Botanic Garden of Belgium.

Harris, D.J., Cheek, M. & Onana, J.M., 2011. Zingiberaceae 232-233. In *The Plants of Mefou proposed National Park, Yaoundé, Cameroon, A Conservation Checklist*. Royal Botanic Gardens, Kew, pp. 232–233.

Harris, D.J. & Wortley, A.W., 2008. *Sangha Trees.*, Royal Botanic Gardens Edinburgh.

Hsieh, T.C., Ma, K.H. & Chao, A., 2016. iNEXT: An R package for interpolation and extrapolation of species diversity (Hill numbers). *To appear in Methods in Ecology and Evolution*.

Hsieh, T.C., Ma, K.H. & Chao, A., 2016. iNEXT: interpolation and extrapolation for species diversity. R package version 2.0.8. *R-project*, pp.1–18. Available at: http://chao.stat.nthu.edu.tw/blog/software-download/%0Ahttp://chao.stat.nthu.edu.tw/blog/software-download.

Hubbell et al., 2008. How many tree species are there in the Amazon and how many of them will go extinct? In *Proceedings of the National Academy of Sciences, 105(Supplement 1), 11498.*

Hutchinson, J. & Dalziel, J., 1928. *Flora of West Tropical Africa*, The Crown Agents for the Colonies.

Knapp, S., 2017. Using the "Natural History Large Hadron Collider" to tell us about plant diversity. *BioMed Central: The Open Access Publisher*. Available at: http://blogs.biomedcentral.com/on-biology/2017/03/07/using-the-natural-history-large-hadron-collider-to-tell-us-about-plant-diversity/.

Lachenaud, O. & Harris, D.J., 2010. Three new species of Psychotria and Chassalia (Rubiaceae) from Central Africa. *Edinburgh Journal of Botany*, 67(2), pp.219–233.

Ley, A.C. & Harris, D.J., 2014. Flower morphological diversity in Aframomum (Zingiberaceae) from Africa – the importance of distinct floral types with presumably

specific pollinator associations, differential habitat adaptations and allopatry in speciation and species maintenance. *Plant Ecology and Evolution*, 147, pp.33–48. Available at: doi: 10.5091/plecevo.2014.824.

Lieberman, M. & Lieberman, D., 2007. Nearest-neighbor tree species combinations in tropical forest: The role of chance, and some consequences of high diversity. *Oikos*, 116(3), pp.377–386.

Longino, J.T. & Colwell, R.K., 2011. Density compensation, species composition, and richness of ants on a neotropical elevational gradient. *Ecosphere*, 2(3), p.art29.

Magurran, A.E., 2004. *Measuring biological diversity.*, Blackwell Publishers.

Mora, C. et al., 2011. How many species are there on earth and in the ocean? *PLoS Biology*, 9(8), pp.1–8.

Ndolo Ebika, S.T. et al., 2015. Hemi-epiphytic Ficus (Moraceae) in a Congolese forest. *Plant Ecology and Evolution*, 148(3), pp.377–386.

Oliver, D., 1868. *Flora of Tropical Africa*, Royal Botanic Gardens, Kew.

Peccoud, J. et al., 2013. Multi-locus phylogenies of the genus Barteria (Passifloraceae) portray complex patterns in the evolution of myrmecophytism. *Molecular Phylogenetics and Evolution*, 66, pp.824–832.

Prance, Beentje & Dransfield, 2000. The Tropical Flora Remains Undercollected. *Annals of the Missouri Botanical Garden*, 87(1), pp.67–71.

R Core Team, 2017. R: A language and environment for statistical computing. Available at: https://www.r-project.org/.%0A.

Ruokolainen, Tuomisto & Kalliola, 2005. Landscape Heterogeneity and Species Diversity in Amazonia. In *Tropical Rainforests: Past, Present, and Future (Bermingham, Dick and Moritz C, eds.)*. University of Chicago Press., pp. 251–270.

Schemske, 2002. *Ecological and Evolutionary Perspectives on the Origins of Tropical Diversity.*, University of Chicago Press.

Shen, T., Chao, A. & Lin, C.F., 2003. Predicting the number of new species in further taxonomic sampling. *Ecology*, 84(3), pp.798–804.

Soberón, J. & Llorente, J., 1993. The use of species accumu- lation functions for the prediction of species richness. *Conserv. Biol.*, 7(3), pp.480–488.

Sosef, M.S.M. et al., 2017. Exploring the floristic diversity of tropical Africa. *BMC Biology*, 15(1), p.15. Available at: http://bmcbiol.biomedcentral.com/articles/10.1186/s12915-017-0356-8.

Ugland, K.I., Gray, J.S. & Ellingsen, K.E., 2003. The species – accumulation curve and

estimation of species richness. *Journal of Animal Ecology*, 72(Rosenzweig 1995), pp.888–897.

Walter, K.S., O'Neal, M.J. & Akbalik, M., 2017. BG-BASE Collections Management Software Version 9.0. Available at: www.bg-base.com.

Walther, B.A. & Moore, J.L., 2005. The concept of bias, precison and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimators. *Ecography*, 28(July), pp.815–829.

Wortley, A.H. & Harris, D.J., 2015. Sangha Trees: an identification and training guide to the trees of the northern Republic of Congo. *In Watson M.F., Lyal H.C. & Pendry C.A. (eds). Descriptive Taxonomy*, The Founda, pp.127–145.

8 Appendix

Table 12. **(Appendix 1)** *Hemi-Epiphyte and Hemi-Parasite Species within the Study Area.*

Hemi-Epiphytes	Hemi-Parasites
Ficus adolfi-friderici Mildbr.	Agelanthus brunneus (Engl.) Balle & N.HallÃ©
Ficus ardisioides Warb. subsp. ardisioides	Agelanthus dichrous (Danser) Polhill & Wiens
Ficus barteri Sprague	Globimetula braunii (Engl.) Danser
Ficus burretiana Mildbr. & Hutch.	Helixanthera mannii (Oliv.) Danser
Ficus calyptrata Vahl	Helixanthera subalata (De Wild.) Wiens & Polhill
Ficus conraui Warb.	Phragmanthera batangae (Engl.) Balle
Ficus craterostoma Warb. ex Mildbr. & Burret	Phragmanthera capitata (Spreng.) Balle
Ficus cyathistipula Warb. subsp. cyathistipula	Phragmanthera polycrypta (Didr.) Balle
Ficus dryepondtiana De Wild.	Tapinanthus ogowensis (Engl.) Danser
Ficus elasticoides De Wild.	Viscum congolense De Wild.
Ficus kamerunensis Mildbr. & Burret	
Ficus lingua De Wild. & T.Durand subsp. lingua	
Ficus lutea Vahl	
Ficus natalensis Hochst. subsp. leprieurii (Miq.) C.C.Berg	
Ficus natalensis Hochst. subsp. natalensis	
Ficus ovata Vahl	
Ficus polita Vahl subsp. polita	
Ficus preussii Warb.	
Ficus recurvata De Wild.	
Ficus sansibarica Warb. subsp. macrosperma (Mildbr. & Burret) C.C.Berg	
Ficus subcostata De Wild.	
Ficus thonningii Blume	

Ficus tremula subsp. kimuenzensis (Warb.) C.C.Berg	
Ficus wildemaniana De Wild. & T.Durand	

Table 13 and Table 14. **(Appendix 2)** *Bad Singletons, Doubtful Doubletons and some Reasons Accounting for the Collecting Bias Resulting in their Existence.*

Bad Singletons	**Reason for Badness**
Adiantum vogelii Mett. ex Keys	Not specialising in ferns
Asparagus drepanophyllus Welw. ex Baker	Known to be common and thought to be only one species, unusual, sterile and not interesting.
Azolla pinnata R.Br.	Floating weed, not my thing. Assumed to be one common species
Bolbitis gaboonensis (Hook.) Alston	Not focusing on these common ferns, one to show broadmindednesss.
Cajanus cajan	Introduced crop
Capsicum annuum L.	Introduced crop.
Chamaecrista kirkii (Oliv.) Standl.	Weed.
Commelina diffusa Burm.f. subsp. diffusa	Everywhere in tropical Africa, one is enough.
Costus afer Ker Gawl.	Big and fleshy, much more common than a singleton would suggest.
Crinum natans Baker	Special, beautiful and aquatic - in deep fast flowing streams.
Crotalaria ochroleuca G.Don	weed in fields
Crotalaria retusa L.	weed in field
Crotalaria spectabilis Roth	weed or crop in field
Eichhornia crassipes (Mart.) Solms	Assume one clone all over Africa, weed, introduced, invasive. Water.
Eleusine indica (L.) Gaertn.	Planted as lawn.
Leonotis nepetifolia (L.) R.Br. var. nepetifolia	cultivated
Ludwigia sp. decurrens Walter	water plant, the petals fall off, I keep meaning to collect -

	usually in deep soft mud with tetse flies.
Mikania microptera DC.	I am sceptical of the identifications.
Mimosa pigra L.	Weed.
Ocimum gratissimum L.	Cultivated
Paspalum conjugatum	Cultivated.
Passiflora foetida L.	Cultivated.
Paullinia pinnata L.	Common and only one species.
Physalis angulata L.	Cultivated
Portulaca grandiflora Hook.	Cultivated
Rhipsalis baccifera (J.S.Muell.) Stearn	Epiphyte and only one species
Rhipsalis cassutha Gaertn.	Epiphyte and only one species.
Selaginella soyauxii Hieron.	Not my interest.
Stachytarpheta cayennensis (Rich.) Vahl	Weed.
Thalia geniculata L.	common and thought to be only one species. Wet places.
Torenia thouarsii (Cham. & Schltdl.) Kuntze	Weed, and small. Pain to identify.
Vernonia amygdalina Delile	Cultivated.
Zornia latifolia Sm.	Introduced weed.

Doubtful Doubles	**Reason for Badness**
Talinum fruticosum (L.) Juss.	weed on lawn
Thonningia sanguinea Vahl	"twice is enough-no reason to recollect"
Crinum jagus (J.Thomps.) Dandy	Too wet
Platycerium stemaria (P.Beauv.) Desv.	Too much hassle to recollect
Gloriosa superba L.	too gaudy to collect twice
Mimosa pudica L.	too common to collect
Sauvagesia erecta L.	too common for recollecting
Ipomoea mauritiana Jacq.	too common for me
Hyptis lanceolata Poir.	too common for m
Desmodium adscendens (Sw.) DC.	too common a weed
Commelina diffusa Burm.f.	too common
Paspalum paniculatum L.	is this not singleton weed

Chromolaena odorata (L.) R.M.King & H.Rob.	invasive
Bacopa egensis (Poepp. & Endl.) Pennell	common weed
Centrosema pubescens Benth.	common weed
Lasimorpha senegalensis Schott	a hassle to recollect
Pueraria phaseoloides (Roxb.) Benth.	yuck weed

Appendix 3 – List of Collectors from the area

Carroll, R.W.

Fangounda, J.

Fay, J.M.

Gentry, A.

Goldsmith, M.

Harris, D.J.

Iyenguet, C.F.

Kami, E.

Koni, D.

Kuroda, S.

Letouzey, R.

Madzoké Bola

Mbani, O.A.

Medjibe, V.P.

Moukassa, G.

Ndolo Ebika, S.T.

Nzolani Silaho, F.O.

Remis, M.J.

Schmidt, R.

Thomas, D.W.

Wraber