Biodiversity Baselines: Tracking Insects in Kruger National Park with DNA Barcodes

Appendix A: Supplementary Protocols

1. *DNA extraction protocol*

Alkaline lysis was used to generate DNA extracts (Truett et al. 2000). Briefly, 50 µL of lysis buffer (25 mM NaOH, 0.2 mM EDTA) was added to each sample before incubation at 95° C for 30 min. After cooling the samples to room temperature, 50 µL of neutralization buffer (40 mM Tris-HCl at pH 5.0) was added. The samples were centrifuged at 1000 g for 1 min to allow the solution to mix, after which 80 µL of the DNA extract from each sample was transferred to a well in a new 96-well plate. A 1 µL aliquot was immediately used to fuel a PCR reaction. The remaining DNA extract was stored at -20° C.

2. *DNA sequencing protocol*

Each 6 µl PCR reaction included 5% trehalose (Fluka Analytical), 1× Platinum Taq reaction buffer (Invitrogen), 2.5 mM $MgCl_2$ (Invitrogen), 0.1µM of each primer (Integrated DNA Technologies), 50µM of each dNTP (KAPA Biosystems), 0.15 units of Platinum Taq (Invitrogen), 1µl of DNA extract, and Hyclone ultra-pure water (Thermo Scientific). The PCR protocol involved two rounds of amplification (PCR1 and PCR2). PCR1 used a single primer cocktail (C_LepFolF, C_LepFolR) with a 30 bp adapter sequence that provided binding sites for UMIs introduced in PCR2 (Hamady et al. 2008, Marx 2016). The thermocycling regime for PCR1 involved initial denaturation for 2 min at 94° C followed by 20 cycles of denaturation for 40 s at 94° C, then annealing for 1 min at 51° C, and extension for 1 min at 72° C. After a 1:1 dilution with ddH_2O, PCR1 products were used as the template for PCR2 which employed primers consisting of a terminal 5 bp pad sequence (GGTAG), a 16 bp Unique Molecular Identifier (UMI), and a 30 bp forward and reverse adapter to match the primer tails from PCR1. PCR2 employed a thermocycling regime similar to PCR1 except the annealing temperature was raised to 64° C.

A 1 µl aliquot from each of 9216 PCR2 reactions was used to create an amplicon pool that was prepared for sequence characterization using standard protocols before being loaded onto a SMRT cell (Hebert et al. 2018). Each of the resultant circular consensus sequences (CCS) represented a concatenated set of sub-reads. These CCS varied in quality, largely depending on the number of component subreads. Only those passing a 99.99% cut off were retained for analysis.

The sequences from each run were analyzed on mBRAVE (mbrave.net) using a standard pipeline that employed BOLD (boldsystems.org) reference libraries for data validation (Ratnasingham & Hebert 2007). Each CCS was assigned to its source well by examining its UMI (see Hebert et al. 2018) and 30 bp was trimmed from its 5′ and 3′ termini to remove primers before further truncation to 640 bp. After trimming, the quality of each CCS was assessed; those with a minimum QV < 50, length < 600 bp or > 1% of their bases with a QV< 20 or > 0.5% of their bases with a QV< 10 were excluded. All remaining sequences were de-replicated based on perfect string identity and each distinct CCS was examined for similarity (>90%) to three reference libraries imported from BOLD to identify 'non-target' CCS reads. Sequences were first screened for bacterial endosymbionts, then for vertebrates, and finally for non-arthropod invertebrates. Sequences passing this screen were then examined for frameshift mutations and/or stop codons to exclude reads deriving from NUMTS.

3. Measuring species diversity

True species diversity (^{q}D) is defined as the inverse of mean species proportional abundance (Hill 1973, Tuomisto 2010a). The mean is calculated with the exponent q that alters the relative weight given to rare vs. abundant species, for example, $q < 1$ de-emphasizes abundance differences among species giving rare species more weight, $q = 1$ gives each species a weight relative to its proportional abundance, and $q > 1$ enhances abundance differences by assigning abundant species more weight. For a given number of species, diversity is highest when species abundances are equal while the larger the value of q, the closer the mean is to the proportional abundance of the most abundance species, lowering the species diversity. When this aspect of community structure was visualised by plotting the site-wise diversity values against q, ecoregion diversity patterns across the park remained (Figure A13).

We quantified species diversity with $q = 0$ which weights species equally and represents the species richness of the data set. Subsequently, to examine the relative contributions of explanatory factors to variation in beta diversity, compositional similarity was measured using the Sorenson index, which measures the turnover of species between effective compositional units (or true beta diversity ^{0}D). As a linear transformation of the beta component of species diversity, it can be used to obtain results that are consistent with species richness used to measure the alpha and gamma components (Tuomisto, 2010b).

Appendix B: Supplementary Tables and Figures

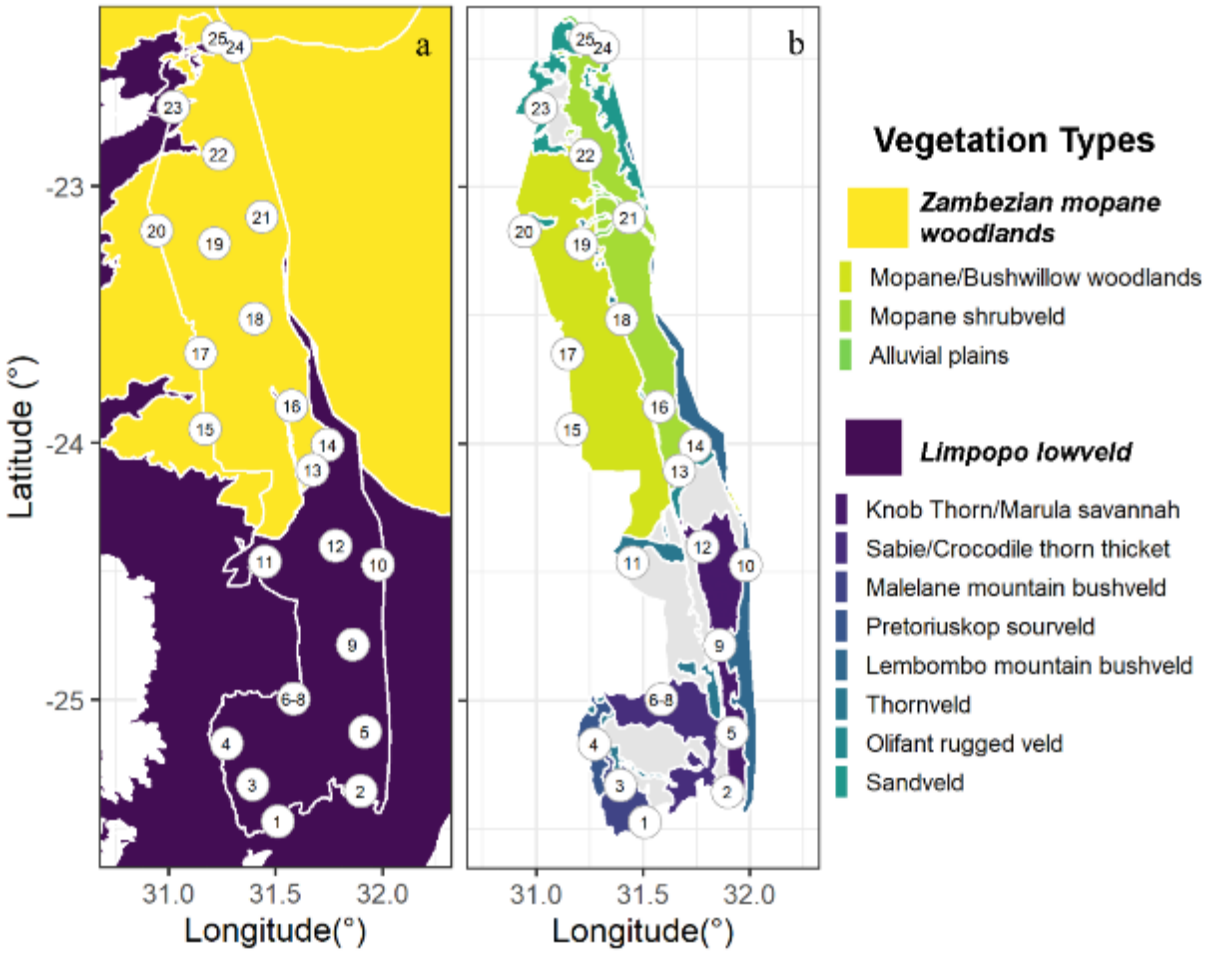

Figure 1: Locations of the 25 Malaise traps deployed across the 22 management sections in the two ecoregions (a) and 11 vegetation types (b) of Kruger National Park.
Single traps were deployed in 20 sections (1–Malelane, 2–Crocodile Bridge, 3–Stolznek, 4–Pretoriuskop, 5–Lower Sabie, 9–Tshokwane, 10–Nwanetsi, 11–Kingfisherspruit, 12–Satara, 13–Houtboschrand, 14–Olifants, 15–Phalaborwa, 16–Letaba, 17–Mahlangeni, 18–Mooiplaas, 19–Woodlands, 20–Shangoni, 21–Shingwedzi, 22–Vlakteplaas, 23–Punda Maria), but three traps were placed in Skukuza (6–8) and two in Pafuri (24, 25). Trap numbers increase with latitude across the lowveld largely in the south comprising eight vegetation types (Knob Thorn/Marula savannah (2, 5, 9, 12), Sabie/Crocodile thorn thicket (6, 7, 8), Malelane mountain bushveld (1, 3), Lembombo mountain bushveld (10), Olifant rugged veld (14), Pretoriuskop sourveld (4), sandveld (23), and thornveld (11)) and the woodlands in the north with three vegetation types (Mopane/Bushwillow woodlands (15, 17, 19, 20, 22), Mopane shrubveld (13, 16, 18, 24), and Alluvial plains (21, 25)).

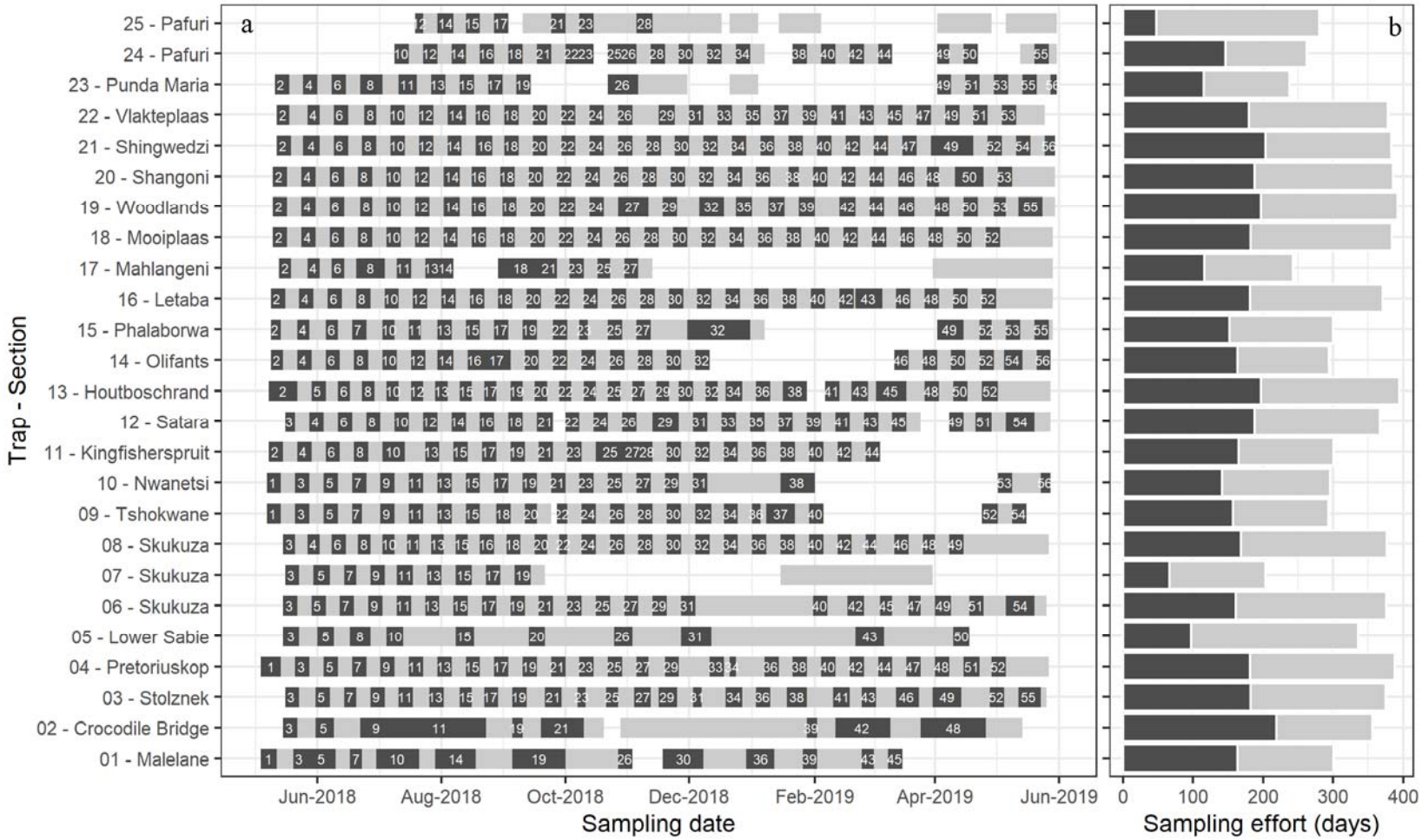

Figure 2: Dates of sample collection (a) and the range in sampling days (b) at each of 25 sites in Kruger National Park. The darker bars indicate the biweekly samples chosen for analysis with the 56 sampling weeks indicated.

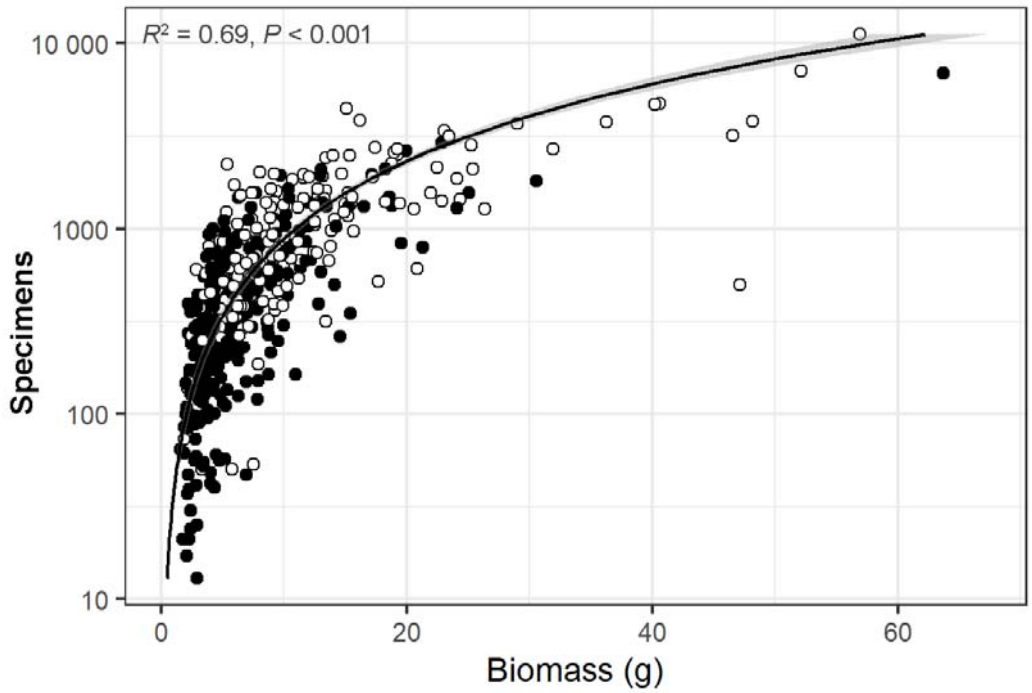

Figure 3: Increase in insect specimen count with biomass for 506 Malaise trap samples from Kruger National Park. Whole samples were weighed but not all specimens counted due to excess individuals of a few very common morphospecies indicated in white.

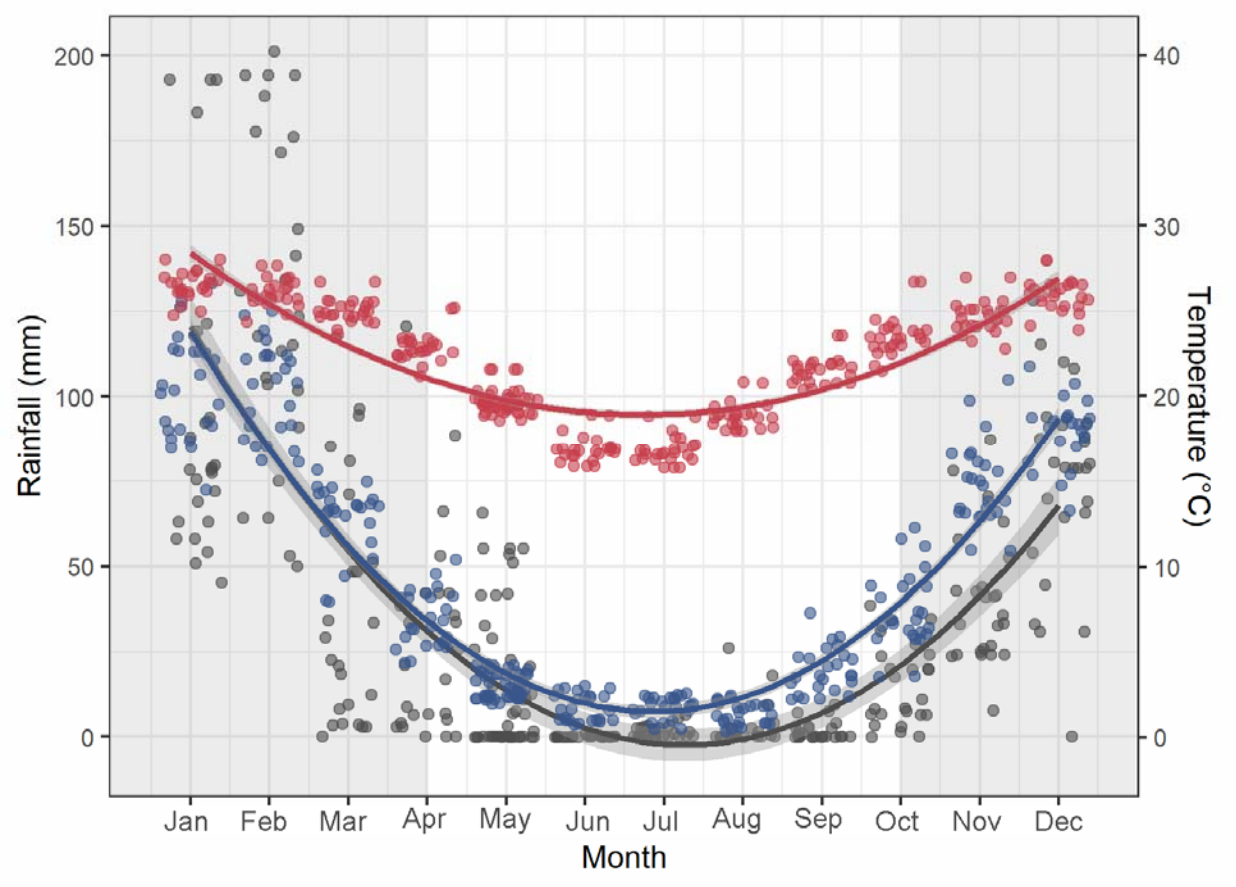

Figure 4: Mean monthly climatic data across the 25 sites in Kruger National Park. Rainfall measured during the program sampling period (2018–2019) (black, left axis), and long-term rainfall (blue, left axis) and temperature (red, right axis) data averaged from the years 1970–2000 (WorldClim v2; Fick & Hijmans 2017). Grey shading indicates the wet season (Oct–Apr) (Smit et al. 2013).

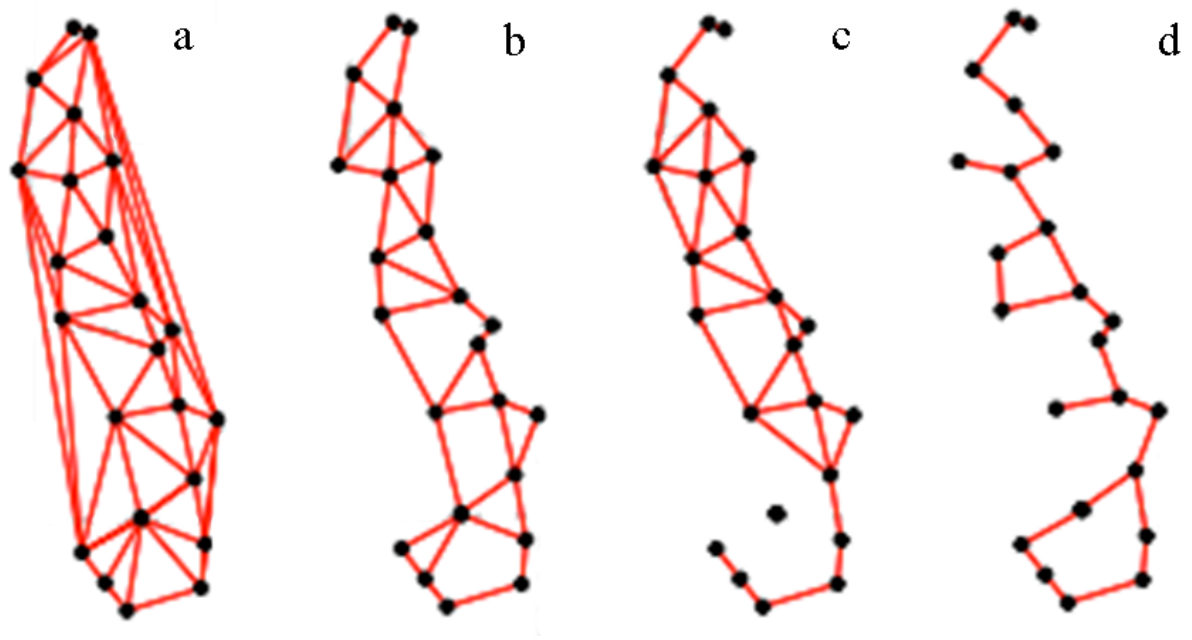

Figure 5: Connectivity matrices based on a decreasing order of nestedness—Delaunay triangulation (a), Gabriel graph (b), sphere of influence (c), and relative neighbourhood graph (d) —representing large- to fine-scale spatial effects.

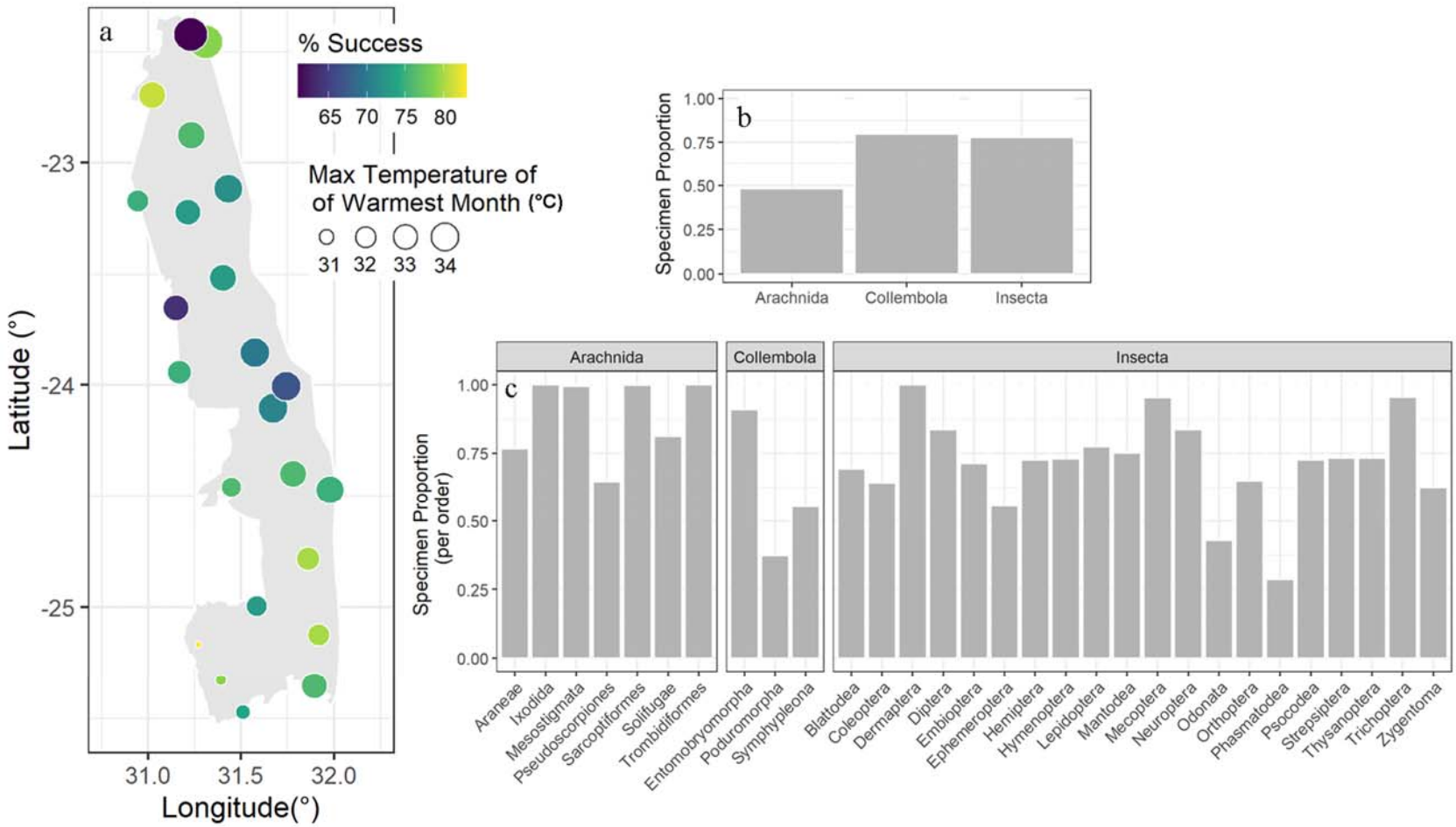

Figure 6: Sequencing success as it relates to the maximum temperature of the warmest month across the 25 sites in Kruger National Park (a) as well as the arthropod class (b) and order (c). About 55% of the specimens of Arachnida, largely mites, were not assigned to an order. Temperature data averaged from the years 1970–2000 (WorldClim v2; Fick & Hijmans 2017).

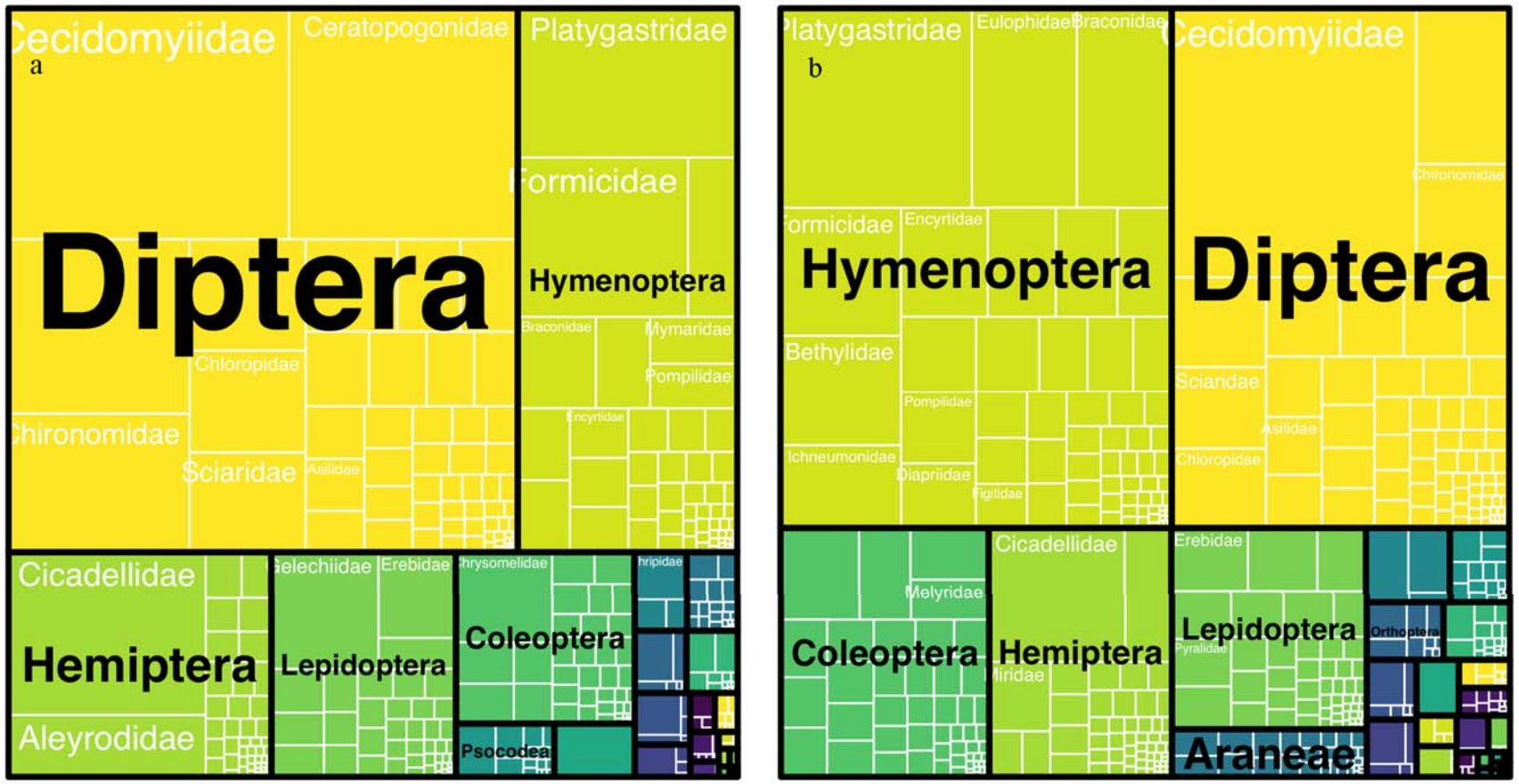

Figure 7: Block diagram showing the relative number of specimens (a) and BINs (b) in each order and family. The results are based on 272,773 specimens with barcodes that represented 19,730 BINs belonging to 32 orders of arthropods (shown in black) and 458 families (shown in white).

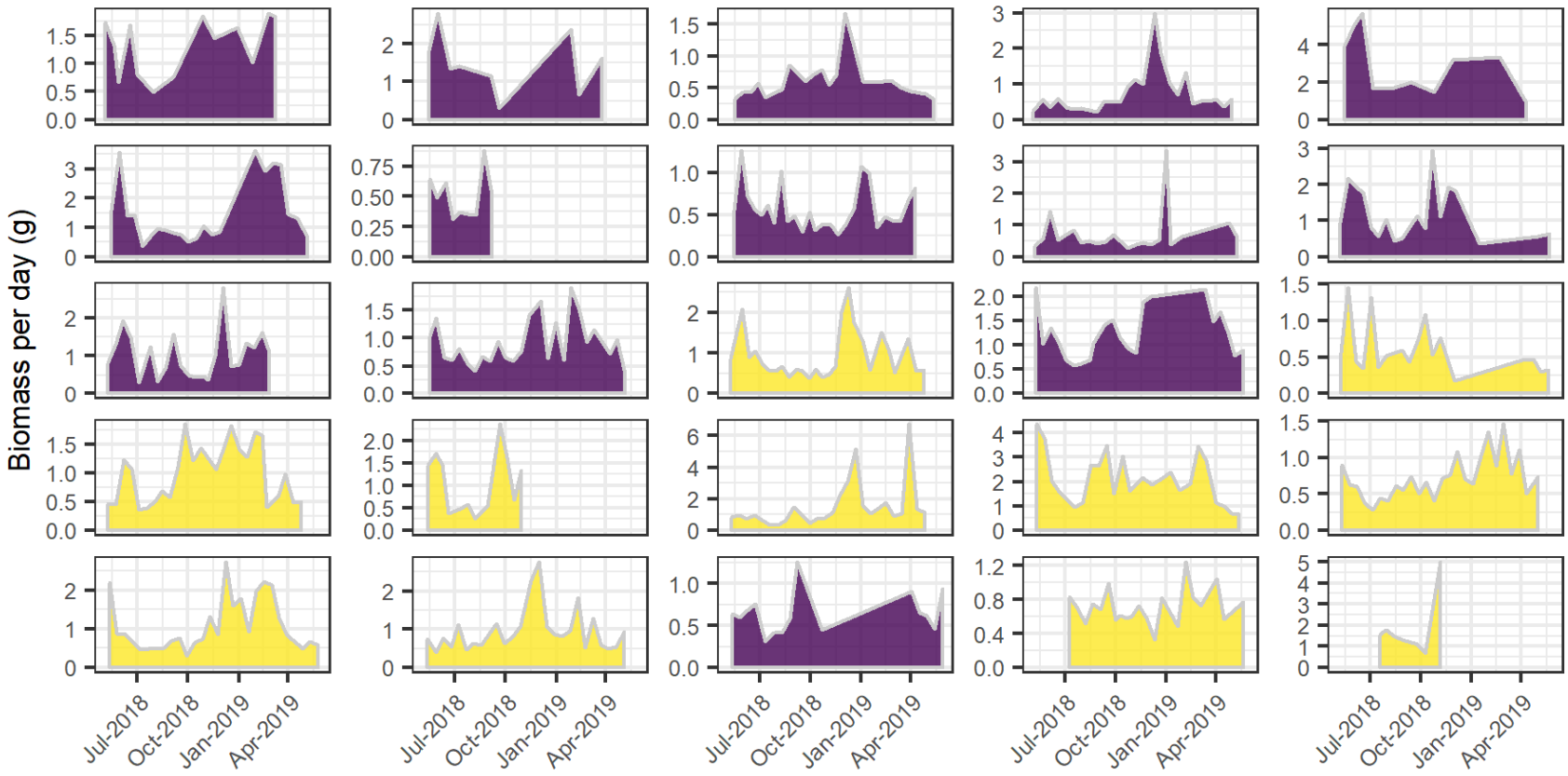

Figure 8: Daily variation in biomass of insects across the 25 sites (left to right, top to bottom, see Figure B1 for site order) in the lowveld (purple) and woodlands (yellow) ecoregions in Kruger National Park. Biomass per day calculated by dividing the observed biomass per sample by the sampling duration.

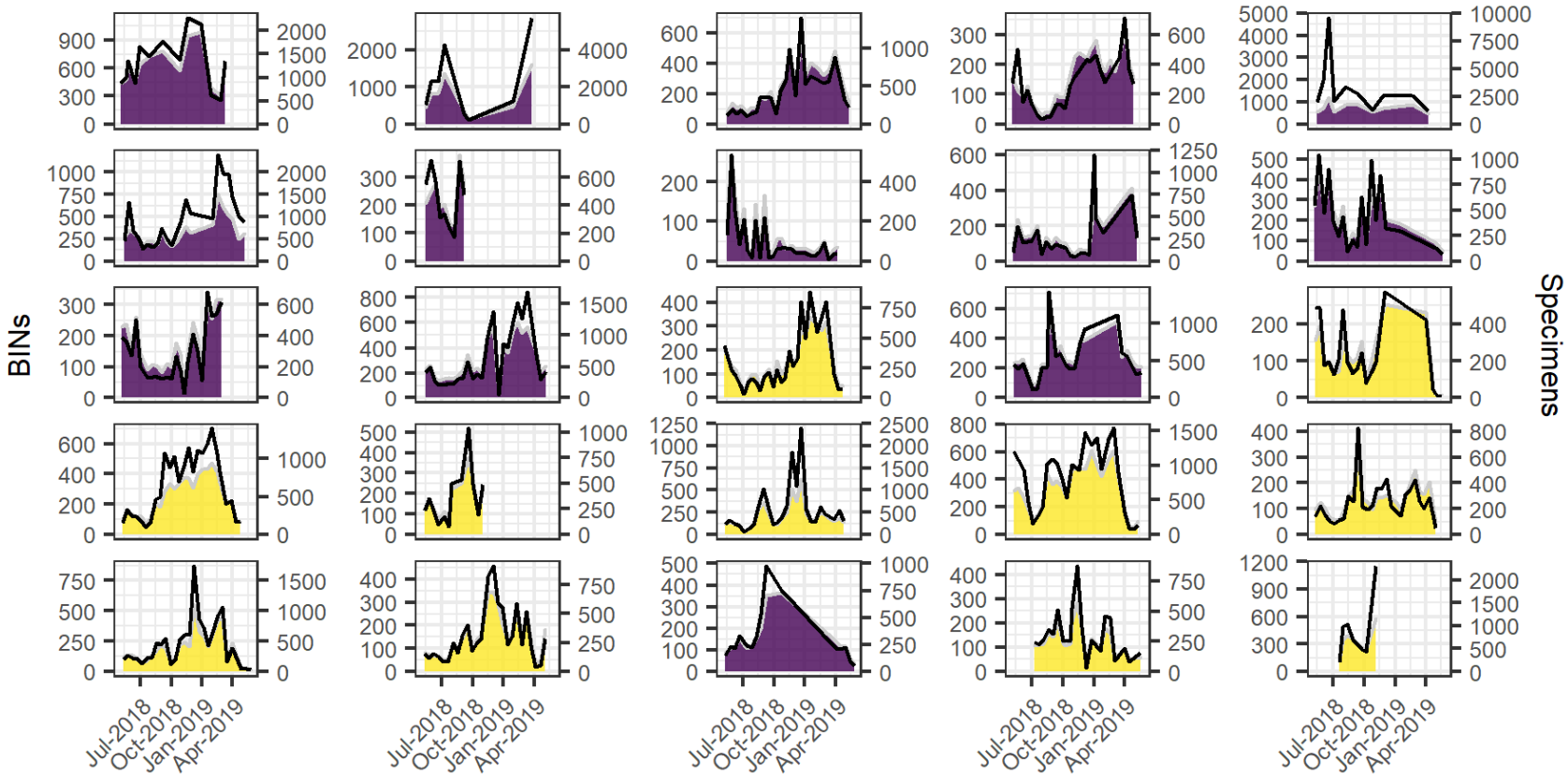

Figure 9: Sample variation in the 19,730 BINs (coloured area, left axis) and 367,743 specimens (black line, right axis) of insects across the 25 sites (left to right, top to bottom, see Figure B1 for site order) in the lowveld (purple) and woodlands (yellow) ecoregions in Kruger National Park.

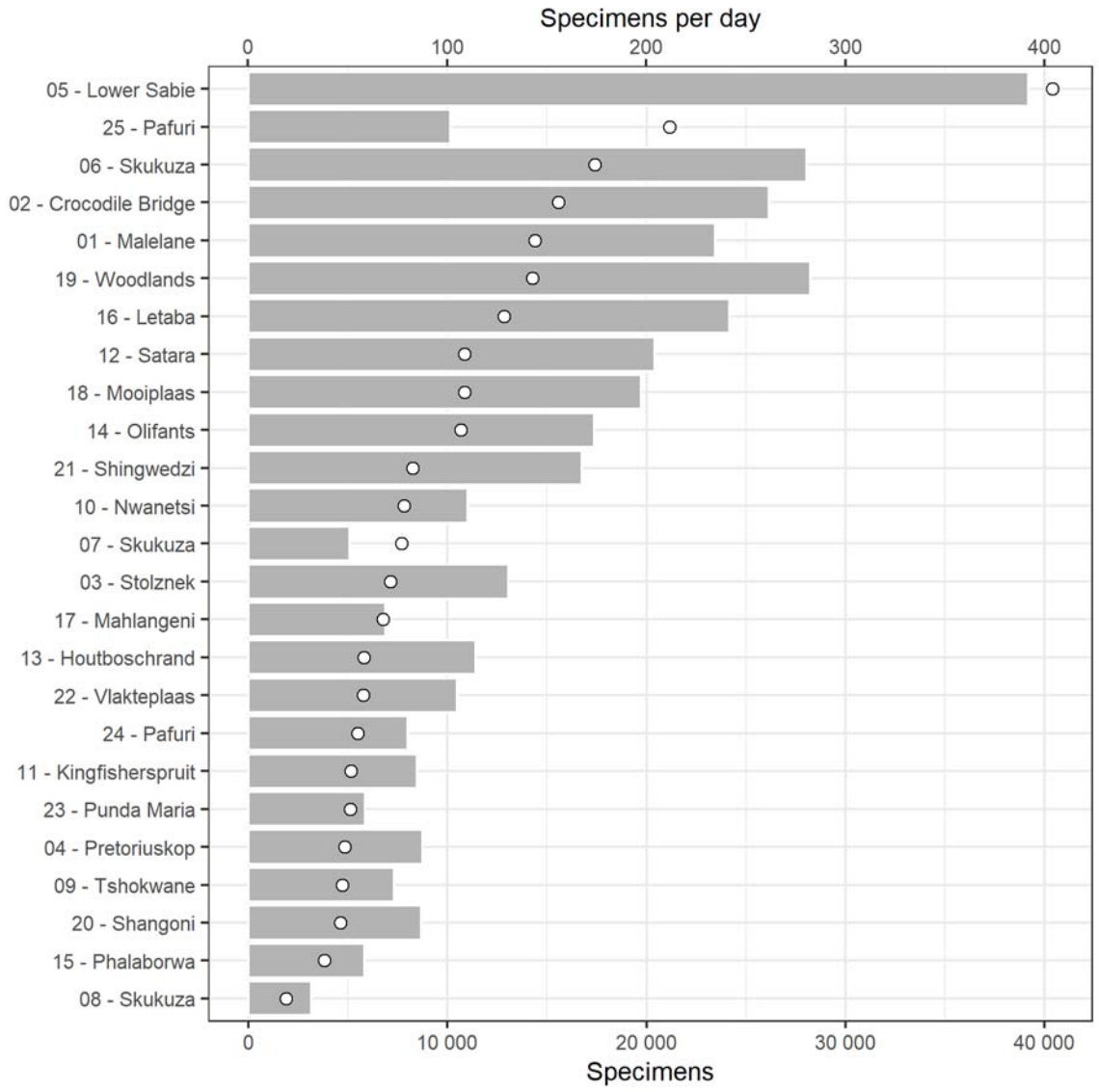

Figure 10: The total insect specimen count (bars, bottom axis) and average daily catch (points, top axis) for each of the 25 sites in Kruger National Park. Sites are ordered (top to bottom) by decreasing number of specimens per day, calculated by dividing the total number of specimens by the duration of collection.

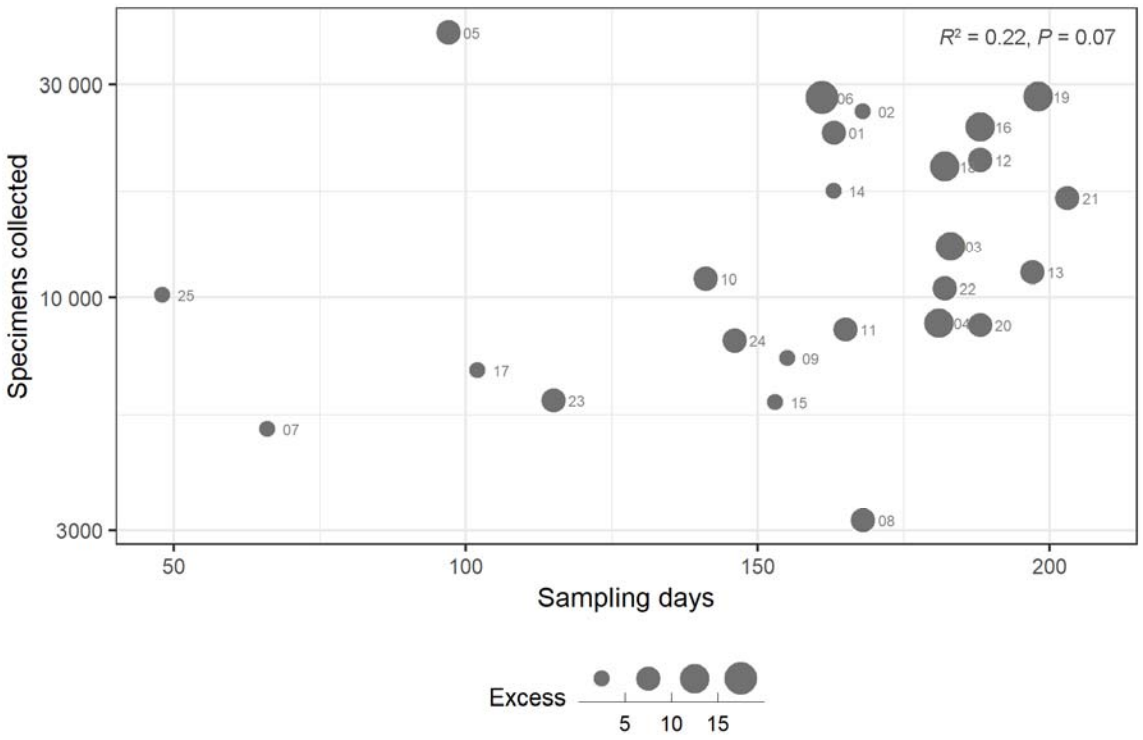

Figure 11: Insect specimens collected per site in relation to sampling days and samples with excess individuals (range: 0–18). Site indicated by the plotted number as outlined in Figure B1

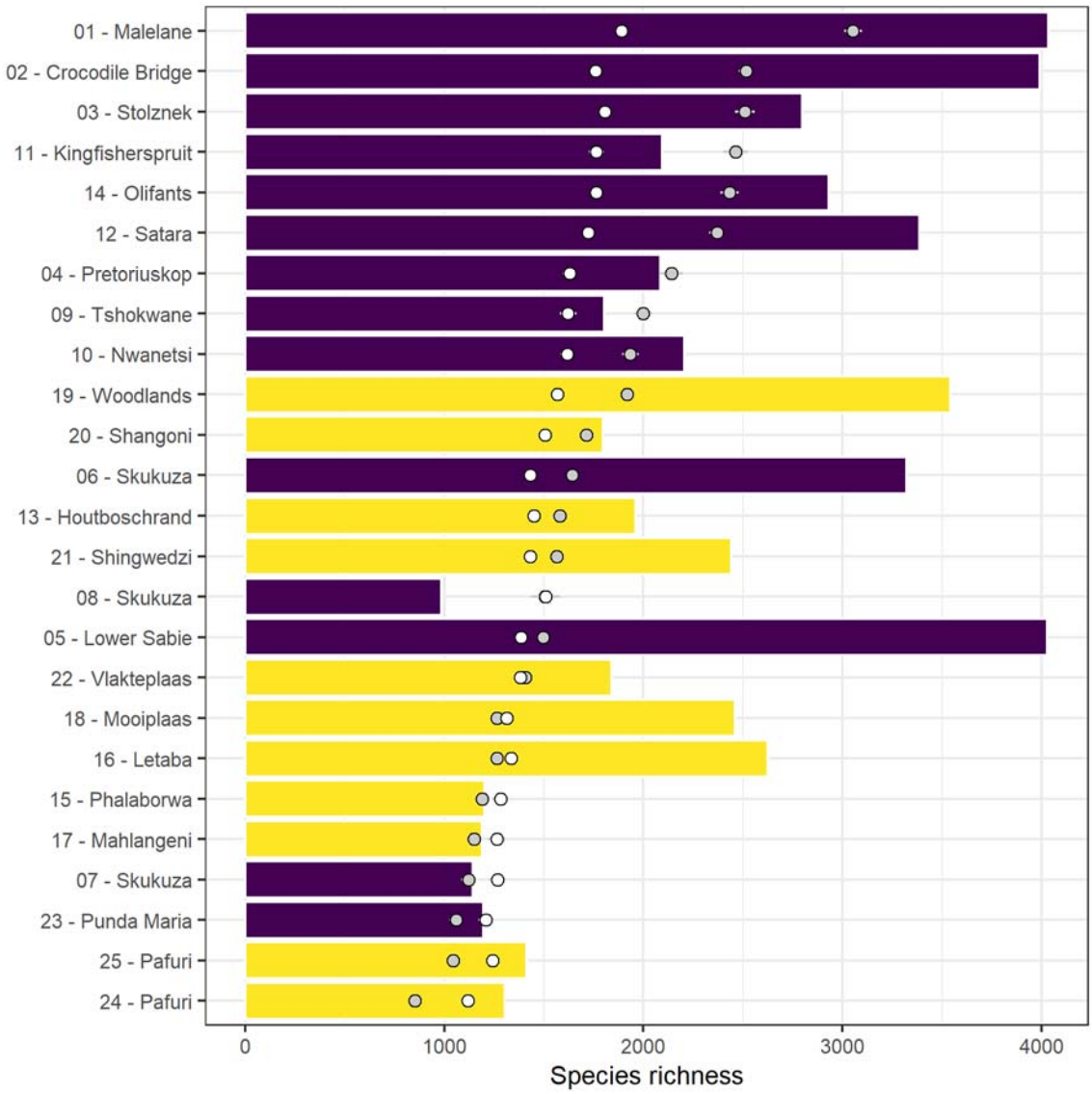

Figure 12: Observed (bars), coverage-based rarefied (grey points), and sample-size-based rarefied (white points) species richness at each of 25 sites in the lowveld (purple) and woodlands (yellow) ecoregions in Kruger National Park. Sites are ordered (top to bottom) by decreasing coverage-based rarefied species richness.

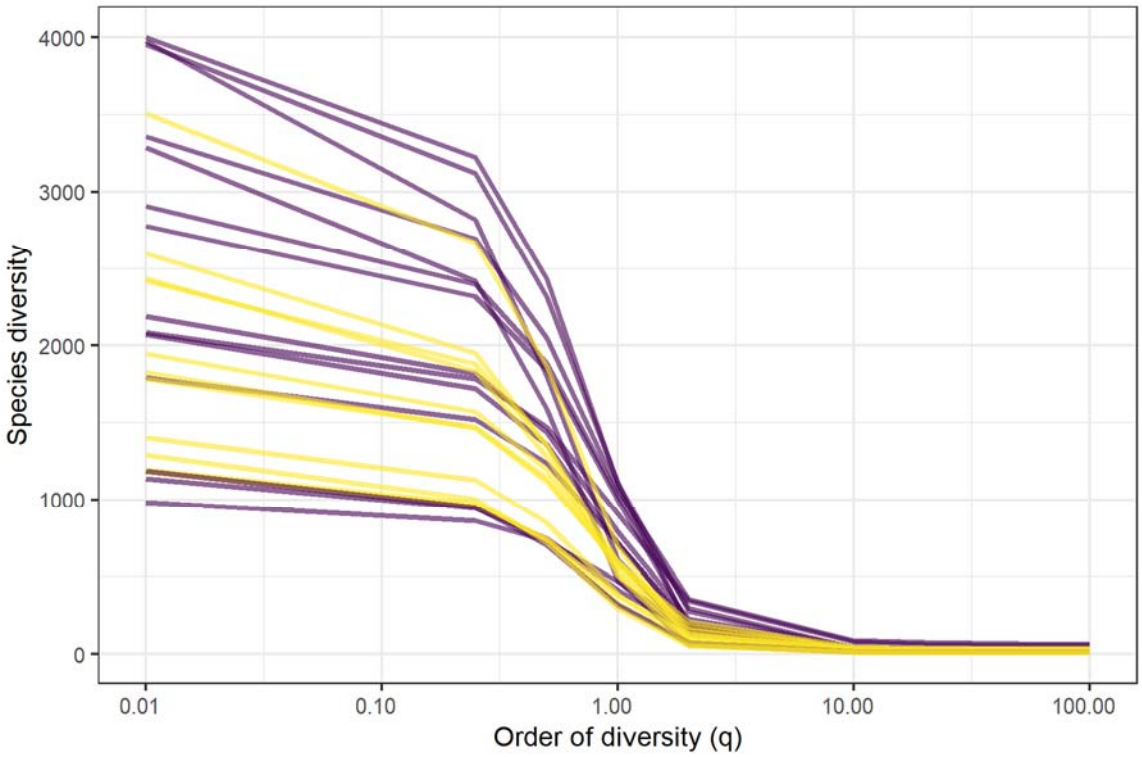

Figure 13: Trends in species diversity at 25 sites in the Kruger National Park as a function of q. Increasing q gives increasing weight to the most abundant species when calculating mean proportional species abundance, which leads to a smaller effective number of species (i.e., the number of species if all were equally abundant).

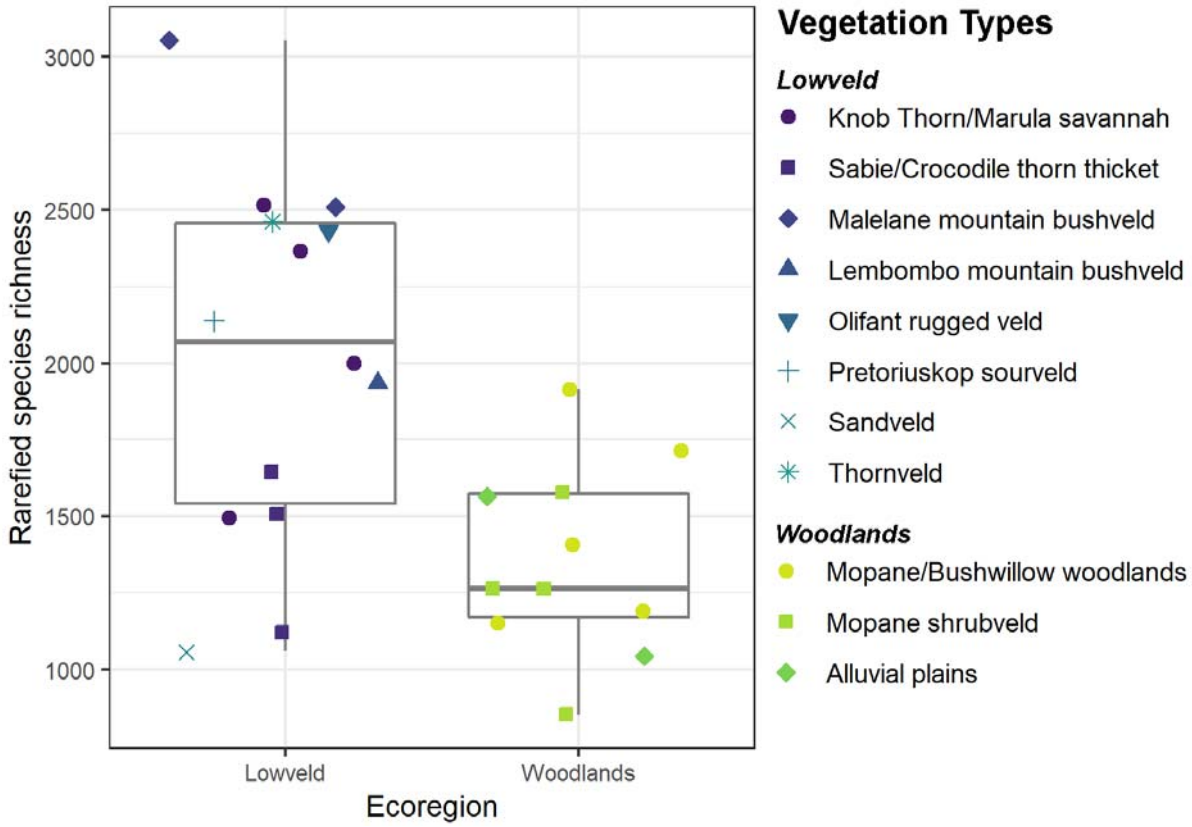

Figure 14: Comparison of coverage-based rarefied species richness of insects between ecoregions and among vegetation types at 25 sites in the Kruger National Park.

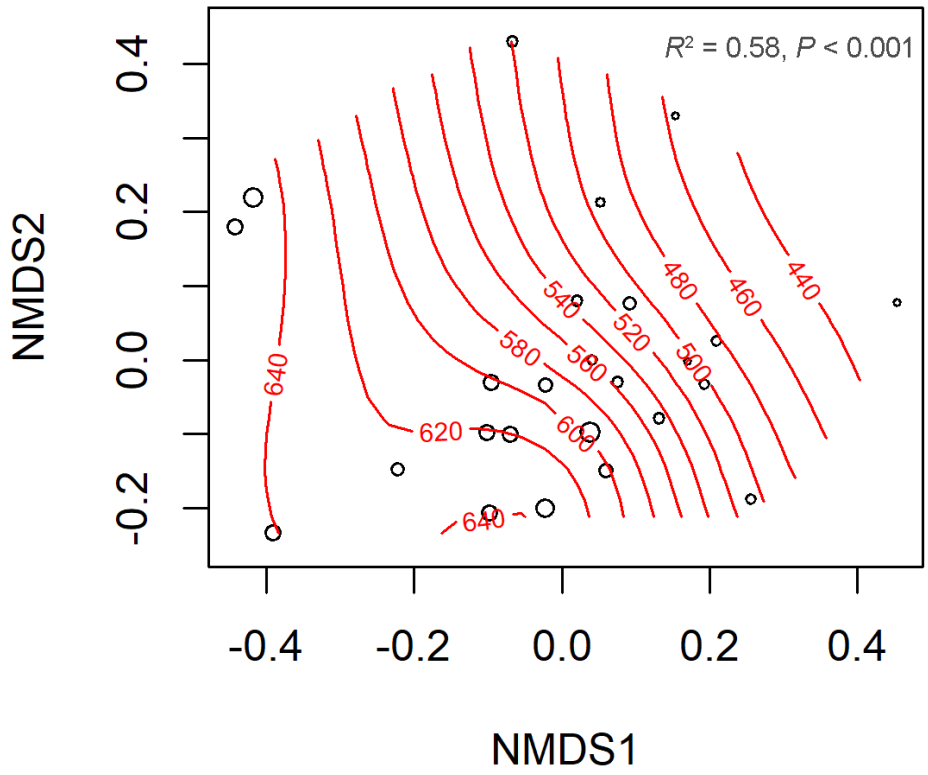

Figure 15: The gradient in mean annual rainfall (mm) correlated with the two-dimensional NMDS ordination as determined by *Envfit* analysis (see Figure 5).

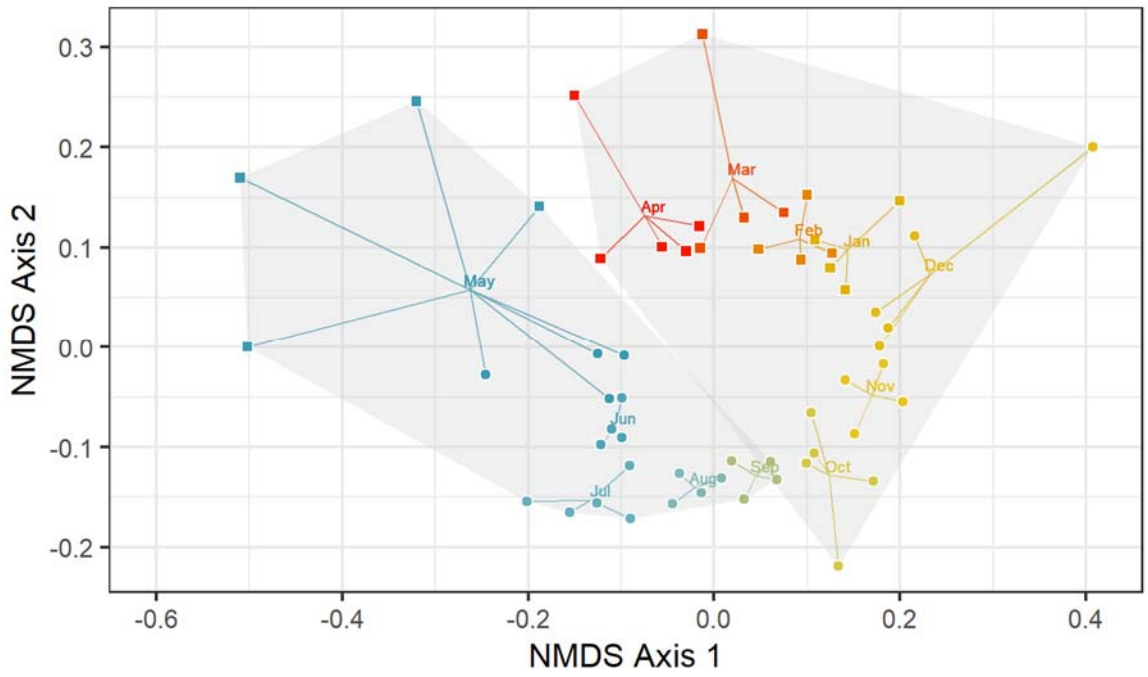

Figure 16: Two-dimensional NMDS ordination of faunal dissimilarity (1 - Sørensen index) between the 56 weeks of sampling from 2018 (circle) to 2019 (square) for all 25 sites in Kruger National Park. The centroid and spread of weeks in a month are indicated by colour with the grey hull indicating clustering by season.

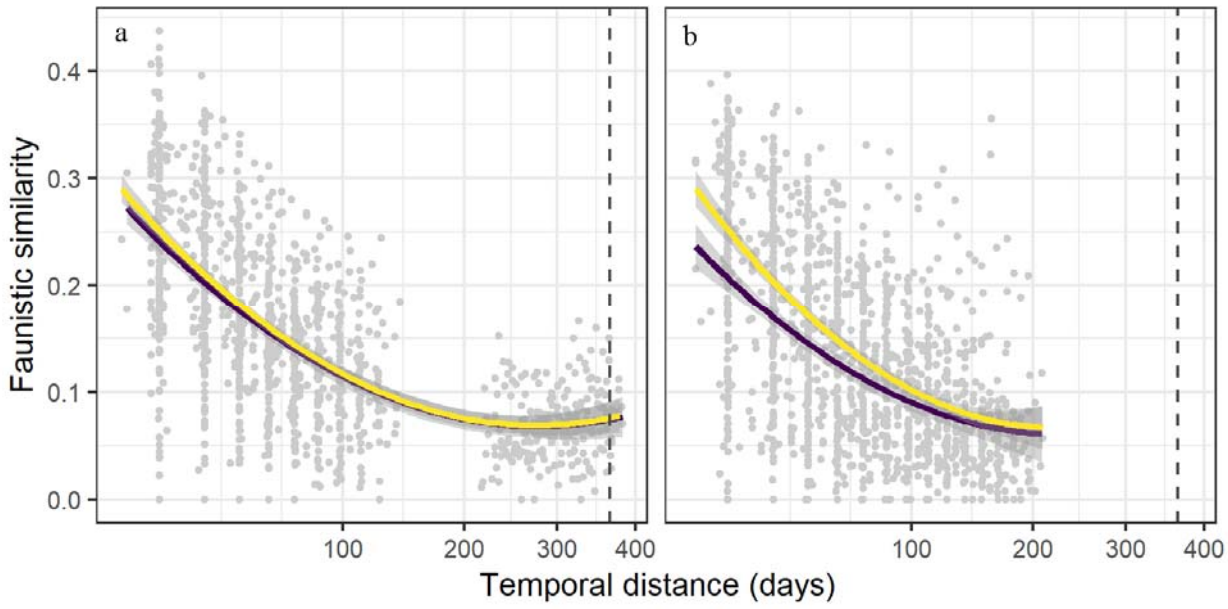

Figure 17: Pairwise similarity values (Sørensen index) between Malaise trap samples of insects over temporal distance in the dry (a) and wet (b) season in the lowveld (purple) and woodlands (yellow). A similarity value of 0 indicates no species overlap between samples. The x-axes are square root transformed to reduce the impact on the model of fewer measurements at larger distances and long intervals. Dashed line indicates 365 days—samples a year apart.

Appendix References

Fick SE, Hijmans RJ (2017) Worldclim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology* 37(12):4302–4315.

Hamady M, Walker JJ, Harris JK, Gold J, Knight R (2008) Error-correcting barcoded primers allow hundreds of samples to be pyrosequenced in multiplex. *Nat Methods* 5:235–237.

Hebert PDN, Braukmann T, Prosser S, Ratnasingham S, deWaard J, Ivanova N, Janzen D, Hallwachs W, Naik S, Sones J, Zakharov E (2018) A Sequel to Sanger: amplicon sequencing that scales. *BMC Genomics* 19:219. doi:10.1186/s12864-018-4611-3.

Hill MO (1973) Diversity and evenness: a unifying notation and its consequences. *Ecology* 54:427–432

Marx V (2016) PCR: the price of infidelity. *Nat Methods*13:475–476.

Ratnasingham S, Hebert PDN (2007) BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). *Molecular Ecology Notes* 7:355–364. doi:10.1111/j.1471-8286.2007.01678.x.

Smit I, Riddell E, Cullum C, Koedoe P (2013) Kruger National Park research supersites: Establishing long-term research sites for cross-disciplinary, multiscaled learning. *Koedoe* 55(1), Art. #1107, 7 pages. doi:10.4102/koedoe. v55i1.1107

Truett GE, Heeger P, Mynatt RL, Truett AA, Walker JA, Warman ML (2000) Preparation of PCR-quality mouse genomic DNA with hot sodium hydroxide and Tris (HotSHOT). *BioTechniques* 29:52–54.

Tuomisto H (2010a). A consistent terminology for quantifying species diversity? Yes, it does exist. *Oecologia* 164(4): 853–860. doi:10.1007/s00442-010-1812-0

Tuomisto H (2010b). A diversity of beta diversities: straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity *Ecography* 33(1): 2–22. doi:10.1111/j.1600-0587.2009.05880.x