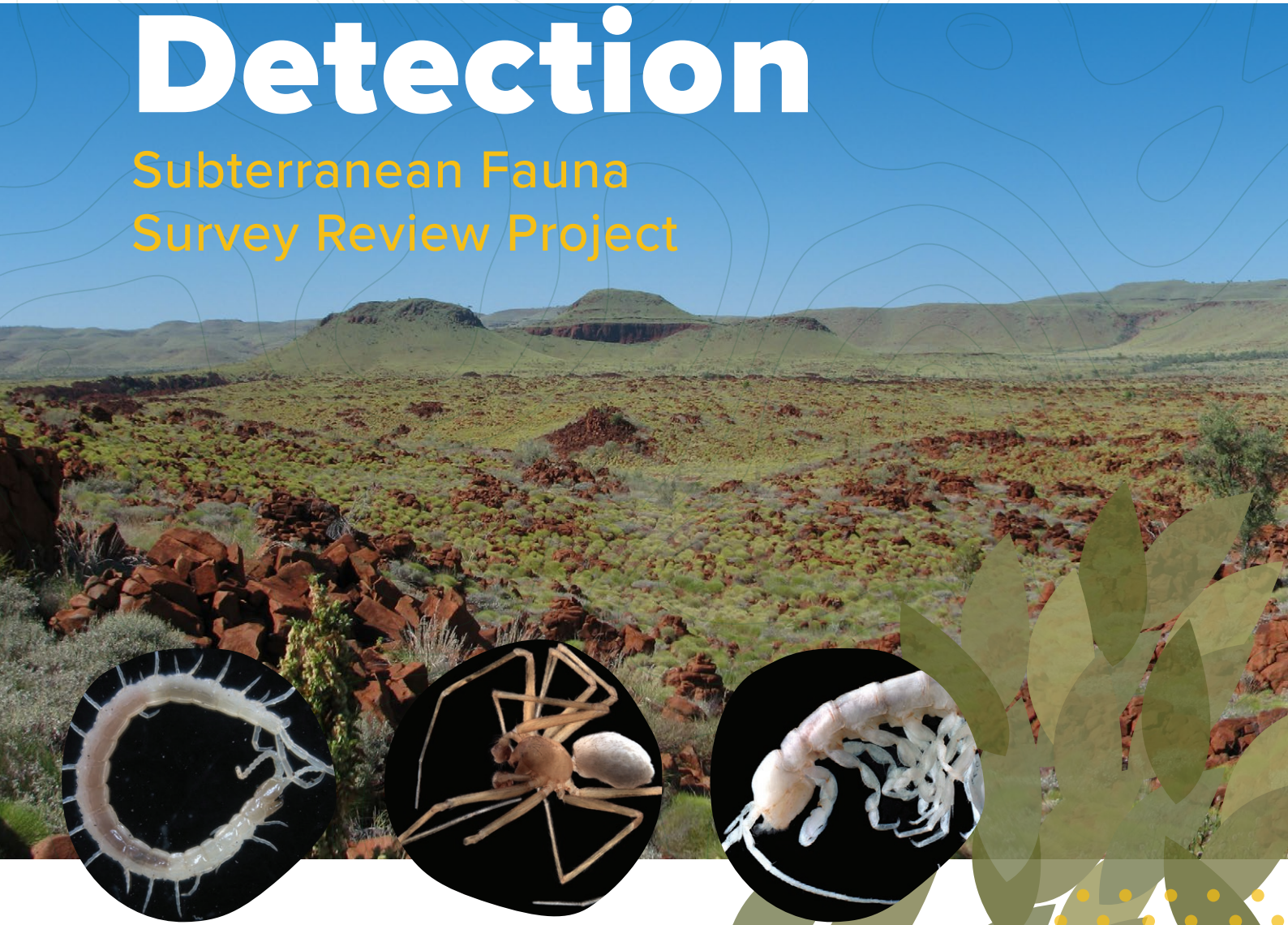
Optimising Species Detection

Subterranean Fauna Survey Review Project



The Western Australian Biodiversity Science Institute

Report commissioned by:

The Western Australian Biodiversity Science Institute

Project funded by:

BHP

Cameco

Department of Water and Environmental Regulation, Government of Western Australia

FMG Fortescue

RioTinto

The Western Australian Biodiversity Science Institute

Report authors: Dr Volker W. Framenau, Cameron McMains, Dr Mariana Campos

Harry Butler Institute, Murdoch University

In collaboration with:

Bennelongia Environmental Consultants

Legal notice

This document should be cited as:

Framenau, V.W., McMains, C. and Campos, M. (2021). *Optimising species detection, subterranean fauna survey review project.* The Western Australian Biodiversity Science Institute, Perth Western Australia.

ISBN 978-0-646-84669-9

IMAGES: Jane McRae, Bennelongia, Department of Biodiversity, Conservation and Attractions, and Steve Dillon

Acknowledgements

The authors of this report and The Western Australian Biodiversity Science Institute would like to thank the following for their contribution:

- The Subterranean Fauna Research Program Steering Committee and the Survey Review Group for their input and support, with special gratitude to Lesley Gibson (DBCA), whose organisational skills and ongoing enthusiasm for this project kept it moving along despite some challenging circumstances, including a world-wide pandemic.
- Stuart Halse (Bennelongia Environmental Consultants), Lesley Gibson (DBCA), Chad Hewitt (Harry Butler Institute), Claire Stevenson (DWER), Dean Main (Rio Tinto), Jared Nelson (FMG), Tanya Carroll (BHP) and Owen Nevin (WABSI) for reviewing the report. Preeti Castle (WABSI) for assistance with structure, layout and in the publication of this report.
- Biologic Environmental Survey (Brad Durrant), Biota Environmental Sciences (Garth Humphreys), Ecologia Environment (Shaun Grein), Phoenix Environmental Sciences (Jarrad Clark), Stantec (Nick Stevens), and Subterranean Ecology (Stefan Eberhard).
- WA Museum (Mark Harvey, Nik Tatarnic, Bill Humphreys) and the Department of Water and Environmental Regulation/IBSA (Clayton Waghorn) for the preparation of data and productive discussions.
- Erich Volschenk (Alacran Environmental Sciences), Nick Stevens (Bestiolas Consulting, Murdoch University), and Andy Austin (University of Adelaide) for their time and insights.
- Atlas Iron, Dacian Gold, Chevron Australia, Citic Pacific, Hancock Prospecting, Hastings, and Mineral Resources and Mineral Resources for permission to use their survey data.

Contents

List of Tables

List of Figures

List of Figures (continued)

1 Executive Summary

Subterranean fauna has been a key environmental factor in the assessment of development proposals in Western Australia since the mid-1990s. Recently, the Western Australian Biodiversity Science Institute (WABSI) has responded to regulatory uncertainties in the assessment of impacts on subterranean fauna by developing a research program that aims to greatly improving our knowledge of subterranean fauna, its environment and response to disturbance. This comprehensive research program focuses on five areas where knowledge gaps were identified: (1) species delineation; (2) best practice sampling and survey protocols; (4) improved understanding of abiotic and biotic habitat requirements; (4) resilience to disturbance and (5) data consolidation.

As a first step to developing best practice sampling and survey protocols, the Harry Butler Institute, Murdoch University (HBI), in collaboration with Bennelongia Environmental Consultants (Bennelongia), was commissioned by WABSI in October 2019 to review historical data to better understand sampling efficiency when targeting subterranean fauna and to highlight areas of improvement. The scope of work of the Subterranean Fauna Review Project – Optimising Species Detection (the 'Project') was divided into two tasks with eight subtasks: 1) collation of data (determine data sources; identify a common set of data parameters to be collated; reconcile nomenclature; determine method of data capture and storage), and 2) statistical analyses to compare detection rates (determine the level of data interrogation; compare detection rate based on sampling strategy; report on results; produce recommendations).

Following extensive consultation with proponents and environmental consultants in Western Australia, a database (the 'Project database') containing about 11,000 troglofauna and 6,500 stygofauna sample sites from ten IBRA regions in Western Australia was compiled. More than 50,000 samples, and almost three times more stygofauna than troglofauna samples, resulted in more than 25,000 records of subterranean fauna with over 224,000 collected specimens. More than 55% of stygofauna records and around 40% of troglofauna records were identified at the species level, either as described species (28.6%% of stygofauna, 13% of troglofauna) or by para-taxonomic morphocodes.

As the original data collection was generally not designed with comprehensive statistical analyses in mind, the Project database had a large number of missing values for many variables that were initially targeted for analyses of survey efficiency and sampling protocols. In addition, survey data were highly biased towards the Pilbara IBRA region. Of those variables that were deemed suitable for analysis throughout the whole dataset (site data: IBRA region, altitude, latitude, longitude, depth to bottom, total visits; sample data: sample type; visit date; conductivity, pH, temperature), many were strongly correlated limiting their analytical power.

The Project database is dominated by rare taxa; 64.2% of stygofauna taxa, and 79.5% of troglofauna taxa were found at three or fewer sites (bores/holes/wells). More than 90% of sites were visited three or fewer times suggesting that the dataset contains largely baseline studies rather than monitoring programs. A strong positive linear relationship between the number of times a site was visited and the number of novel identified lowest taxonomic unit (LTUs) was recorded for both stygo- and troglofauna, and within the limits of the analysis (up to 10 visits per site), the number of novel LTUs collected does not reach a plateau. This is reflected in the modelled taxon accumulation curves (novel taxa by cumulative taxa) which flatten at very high counts of taxa found, into the hundreds of taxa. Sample-based accumulation curves were not calculated as it was often difficult to ascertain the number of zero-samples (i.e., samples that did not record specimens) of a survey.

A comparison of the efficacy of different trap types included by-catch of non-target methods (i.e., scrapes for stygofauna and haul nets for troglofauna). Stygofauna were collected by nets (14,873 records) an order of magnitude higher than by scrapes (1,906 records), though scrapes are not designed to target stygofauna. Troglofauna were collected by three methods: traps (3,674 records), scrapes (4,350 records) and nets as by-catch (1,130 records). All three methods were similar in the mean number of troglofauna organisms retrieved per sample. Sampling method is a significant determinant of the community found for troglofauna, but less so for stygofauna. Scrapes and nets collect similar organisms (not surprising as they are essentially the same sampling method), but troglofauna traps collect a distinct assemblage. However, troglofauna communities were not significantly different between traps of different depth order.

The 7-day cumulative rainfall showed a slight negative relationship with both stygofauna and troglofauna LTU richness, while the 30-day and storm metrics had no significant relationship. There was no overall significant relationship between length of the sampling interval and the dissimilarity of samples; however, where a significant relationship does exist, it is strongly positive for both stygofauna and troglofauna, which means in these cases temporally distant sites are also ecologically distant. Community composition between months of the year varies. For stygofauna, January samples are significantly more diverse than in any other month; for troglofauna, the most diverse sampling month was March.

Recommendations in relation to the objective of the Project are derived from all aspects of this study, i.e., problems during the acquisition of the data, data quality (how data are being collected and stored, including minimal taxonomic standards), and results of the data analyses, specifically in relation to sampling efficacy. The recommendations fall into five categories: (1) database structure; (2) standardising data; (3) governance of taxonomy; (4) improving sampling; and (5) experimental sample programs.

The Project database structure was chosen to facilitate data analyses and may not meet the objectives of a public subterranean fauna data depository. It is recommended to create a database working group of survey and database experts that critically reviews the current database structure and develops a fit-for-purpose data solution. This database solution may incorporate other survey components, such as short-range endemic invertebrates (SREs) or aquatic invertebrates to provide a more comprehensive data platform for environmental assessments.

The analyses were limited by a lack of clear definitions of many categorised variables and inconsistencies in the data collected during surveys. These issues highlight the need for standardised data collection parameters and data delivery format. A standard data collection sheet would include which parameters are mandatory, recommended, and optional for collection; details regarding the sampling methodology (which is often described in the report but not detailed in the data spreadsheets); and a specific format that would enable automated incorporation into the large database. Compliance with data standards requires regulatory oversight, for example, non-acceptance of survey assessments if minimum data standards are not met (e.g., missing mandatory values).

The taxonomic tables of a biodiversity database are of crucial importance for correctly analysing survey data in relation to taxon richness and evenness, rarity (and therefore conservation significance), distribution ranges, habitat preferences, and sampling design. It is therefore recommended, at a minimum, to implement standardised taxonomic principles in a future database, ideally the designation of a publicly available reference specimens ("type") for each parataxonomic morphospecies, accompanied by a diagnosis (morphological and/or molecular) and detailing who recognised the new species and when.

There are a few key recommendations concerning current sampling methods and regimes for consideration. These are: (1) use both traps and scrapes for troglofauna surveys as they are complementary in the communities/taxa they collect and include any stygofauna by-catch in the analysis; (2) as there are no obvious differences in the troglofauna communities of traps of different depth, multiple traps can be installed for increased sampling effort but are unlikely to increase diversity per sample; (3) include any by-catch of troglofauna scrapes in the analysis of a stygofauna survey; (4) collect in different months if possible and space survey phases temporally as far apart as possible; (5) collect in January for stygofauna (although possibly prohibitive due to high temperatures in Pilbara) and in March for troglofauna; and (6) collect as many times as practicable, but observe the rate of novel taxa over previously collected taxa to indicate whether a minimum target community has been documented.

Many target variables for the analysis of sampling efficiency could not be analysed as the initial data were not collected with these analyses in mind. It is therefore recommended to further explore the influence of core variables (such as appropriately defined trap designs or sampling regime, geology, time since bore installation) with statistically sound experimental sample programs that control for correlated variables.

2 Introduction

Subterranean fauna has been a key environmental factor in the assessment of development proposals in Western Australia since the mid-1990s (EPA 2016a, c, d). Techniques to survey subterranean fauna are generally limited to sampling exploration or production bores and respective survey strategies have shown to be relatively inefficient, with many species often detected in a single bore only (Eberhard et al. 2009a). However, an impact assessment requires knowledge of the distribution of species within and beyond a development footprint, and therefore it is important to determine the range of each species and the availability of suitable habitat beyond the impact area. Limitations in the ability to survey subterranean fauna result in uncertainties modelling their distribution patterns and are compounded by a lack of knowledge of habitat preferences and responses to impacts. In addition, subterranean fauna also pose specific problems in delineating species (Halse 2018). Uncertainties in environmental assessments that include subterranean fauna as a key environmental factor have led to delays in developments and investment decisions and some projects being rejected as the objectives of the EPA in relation to subterranean fauna were not met for assumed rare species (e.g. EPA 2016b).

Due to uncertainties in delineating subterranean fauna distributions and a lack in knowledge on how species may respond to potential impacts, subterranean fauna was presented as a research priority to the Western Australian Biodiversity Science Institute (WABSI) in early 2017. A series of workshops involving end-users and researchers were organised with the aim to develop a program of research to close knowledge gaps. The intent of the research program was to provide the framework for the development of research activities and to encourage collaboration. A clear consensus on five broad focus areas to progress included (Gibson 2018): (1) species delineation (i.e. what is a species and how much genetic differentiation is ecologically/evolutionary important); (2) best practice sampling and survey protocol (review and refine techniques to develop optimal sampling methods, incl. stratified sampling); (3) improved understanding of abiotic and biotic habitat requirements (e.g., best habitat predictors for species, continuity of habitat); (4) resilience to disturbance (e.g., response to disturbance, migration; possibility of translocation); and (5) data consolidation (how to consolidate data, data ownership, standardised taxonomic classifications).

One of the five broad areas to be addressed was the identification of best practice survey and sampling protocols to optimise the efficiency of survey and monitoring (Gibson 2018). A review and refinement of survey and sampling methods is required to ensure contemporary approaches are efficient, effective, and reproducible so that the subterranean fauna of an area is accurately documented. As a first step towards improving current practice, a review of the historical survey effort and sampling techniques were proposed. Subterranean fauna surveys undertaken as a part of environmental impact assessments provided the primary source of information for this Subterranean Fauna Review – Optimising Species Detection (the 'Project').

The scope of work of the Project was divided into two tasks with eight subtasks:

i. **Collation of data**

- Determine data sources with the project working group (i.e., funding partners)
- Liaise with the project working group and environmental consultants to identify a common set of data parameters to be collated
- Establish a procedure for reconciling nomenclature with advice from the WA Museum (e.g., specimen codes, name changes)
- Determine method of data capture and storage in consultation with Department of Water and Environmental Regulation (e.g., database, custodian)

ii. **Statistical analyses (dependent on the available data)**

- Liaise with the project working group and consultants to determine the level of data interrogation required and identify the answers most desired from the available data
- Compare detection rate based on the sampling strategy used (techniques and effort)
- Report on the outcomes of the analyses
- Produce recommendations for improving data collection and reporting.

This work aimed to collate and review a large historical set of data from Western Australia to satisfy the main objective of a better understanding of best practices to maximise sampling efficiency.

3 Materials and Methods

3.1 Data acquisition

3.1.1 Subterranean fauna survey data

Data from funding partners (refer acknowledgments) and non-funding partners was sourced and a final selection of data of troglofauna and stygofauna records to be included in the analyses was made based on data quality, completeness and confidentiality. These data were collected between 2001 and 2018 from almost 11,000 sites with troglofauna records and almost 6,500 sites with stygofauna records in ten IBRA regions throughout Western Australia (Table 1).

The following dataset were requested but were not suitable for statistical analyses to meet the objectives of the Project:

- The Western Australian Museum's (WAM) Arachnology/Myriapodology database (provided to WABSI on 3 March 2019) represented a specimen database with generally insufficient information on collection type and effort, both of which are important for the data analyses.
- Research data from the South Australian Museum and Adelaide University were considered, but not incorporated as no consolidated databases were available. Similar to the WA Museum database, published data sets are specimen-based and not suitable for statistical analyses of survey effort (A. Austin, pers. comm. to VWF, February 2019).
- Data from the Index of Biodiversity Surveys for Assessment (IBSA) (https://www.wa.gov.au/service/environment/environmental-impact-assessment/program-index-of-biodiversity-surveys-assessments; accessed 4 December 2020) were not suitable for incorporation into the Project database, as at the time of data compilation IBSA only had survey reports, but no datasets available online (VWF meeting with Clayton Waghorn, 9 June 2019). However, some data from these reports have been captured in the data acquisition phase.

3.1.2 External environmental data

Rainfall data of all WA meteorological stations for the period between 1 January 2001 and 31 December 2018 were acquired from the Bureau of Meteorology in February 2020 covering all analysed sampling events (see Table 1).

The IBRA dataset (GIS shapefiles) was provided under a creative commons license by the Australian Federal Government Department of Agriculture, Water and the Environment (see http://www.environment.gov.au/land/nrs/science/ibra/ibra7-codes; accessed 4 December 2020).

The analysis of geology on sampling efficacy was a key question to be addressed. In addition to the lack of geological data in many source data sets or a lack of standardised geological categories across the data, a number of other factors prevented a detailed analysis:

1. Publicly available data layers (e.g., surface geology) were considered not to accurately reflect subterranean conditions. Other geo-spatial layers, e.g., those developed by Mokany et al. (2018), were only developed for the Pilbara and not available for other regions covered in the Project database.
2. Geological data compilation by detailed screening of survey reports or standardisation of datasets offered by the different funding partners was considered to exceed the scope of this study.
3. There was a high data bias in geology towards target ores (e.g., BIF/CID in the Pilbara) or known biodiversity hotspots of subterranean fauna (calcretes, karst in Yilgarn), with often little variation within regions. Datasets of areas with weathered volcanic or utramafic geologies are rare in the Project database and further increased the data bias.

Table 1. Data analysed for the Subterranean Fauna Survey Review Project – Optimising Species Detection; sites include those where subterranean fauna was recovered

Data source	Nearest town/location	IBRA regions#	Number of sites (troglofauna)	Number of sites (stygofauna)	Earliest sample	Latest sample
BHP	Newman, Wittenoom	PIL, GAS, GSD	4,762	1,980	24 Aug 2003	25 Aug 2016
Cameco	Telfer, Wiluna	LSD, MUR	186	348	10 Mar 2009	3 Jul 2015
Chevron Australia	Mardie	CAR	-	44	28 Apr 2011	6 June 2017
Dacian Gold	Laverton	MUR	15	77	10 Feb 2016	15 Dec 2017
Fortescue Metals Group	Mallina, Marble Bar, Newman, Nullagine, Pannawonica, Paraburdoo, Port Hedland, Tom Price, Wittenoom	PIL	2,832	1,468	11 Mar 2005	12 Apr 2016
Hancock Prospecting	Newman, Nullagine, Wittenoom	PIL	464	351	25 Mar 2008	11 Feb 2015
Hastings	Paraburdoo	GAS	132	154	16 May 2015	14 Aug 2018
Pilbara Stygofauna Survey	Balla Balla, Bamboo, Dampier, Goldsworthy, Horseshoe, Karratha, Mallina, Marble Bar, Mardie, Newman, Nullagine, Onslow, Onslow (Old), Pannawonica, Paraburdoo, Peak Hill, Port Hedland, Roebourne, Shay Gap, Shellborough, Telfer, Tom Price, Wittenoom	PIL, GSD, DAL, GAS, CAR, LSD	-	539	1 Jan 2001	2 Aug 2006
Rio Tinto	Newman, Pannawonica, Paraburdoo, Tom Price	PIL, GAS	2,228	1,107	1 Mar 1998	14 Dec 2017
Stantec	Browns Range, Lake Maitland, Lake Way, Yakabindie	MUR, YAL, CAR, TAN, GES	325	400	1 Jan 2006	8 Aug 2017
Sum		**10 regions**	**10,994**	**6,468**	**Earliest: 1 Jan 2001**	**Latest: 14 Aug 2018**

#IBRA (Interim Biogeographic Regionalisation for Australia) regions – CAR, Carnarvon; DAL, Dampierland; GAS, Gascoyne; GES, Geraldton Sandplains; GSD, Great Sandy Desert; LSD, Little Sandy Desert; MUR, Murchison; PIL, Pilbara; TAN, Tanami; YAL, Yalgoo (see http://www.environment.gov.au/land/nrs/science/ibra/ibra7-codes; accessed 4 December 2020)

3.2 Database compilation

The final database ('Project database') was assembled in two steps:

1. **Initial data compilation including Quality Assurance and Quality Control (QAQC)**

Subterranean fauna data were collated into a pre-existing SQL database with Microsoft (MS) Access® frontend maintained by one of the Project partners (Bennelongia). Additional datasets were received in multiple formats, generally in MS Excel® or comma-separated (csv) tables and were manipulated to match the table formats of the SQL database and imported to it. Survey reports associated with survey data in the SQL database were screened in many cases to complement the data in the database; however, initial data collection and reporting during environmental surveys was generally not designed to fulfil the statistical requirements for the analyses here. Therefore, the time required to extract information from reports to fill data gaps was in many cases prohibitive or these data were simply not reported.

Common problems detected during the QAQC process were:

- Missing data, e.g., sites present in the taxon record tables were absent from the site tables (so collecting events had no location and associated details).
- Many survey datasets recorded only the subterranean animals collected; there were no zero-sample data which are crucial for analysing sampling effort (i.e., sample-based accumulation curves).
- Results of sub-samples were combined, e.g., when setting multiple troglofauna traps, or combining results of troglofauna traps and scraping from the same bore.
- Use of sampling trip as a date marker caused problems when bores were re-sampled during the same field trip.
- It often remained unclear from both databases and reports how sampling was conducted and the associated sampling effort.
- Lack of formal species descriptions made data compilation difficult (i.e., the possibility that the same species was reported under different parataxonomic names).

2. **Final preparation of the database as required for statistical analysis**

The structure of the SQL database was reviewed and initial data exploration was undertaken for missing connections or data sets. The structure of the database was altered from SQL/MS Access® to nine csv-files. This reduced the connection steps between the different files (Figure 1). This format was also chosen to minimise storage requirements (see Table 2) and duplication of data while maximising cross-platform usability.

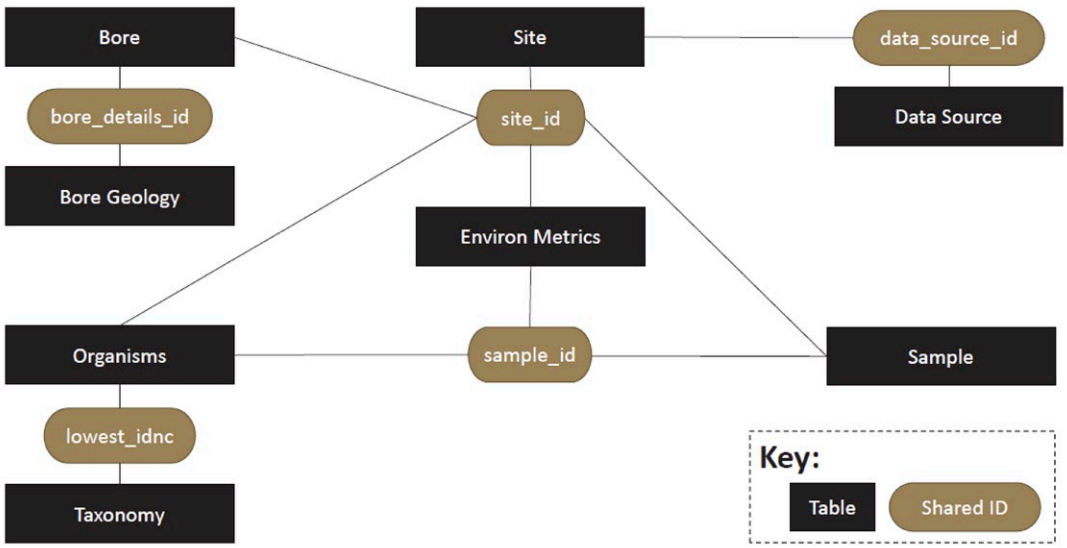


Figure 1. Final Project database structure. Each dark square represents a separate table (csv-file), and the grey-brown fields show how these are connected (shared id). The table 'source reports' is not displayed above as irrelevant for the analyses (see Table 2 for summary statistics of each table and Appendix 1 for Project database metadata)

Table 2. Summary statistics for the tables of the Project database

Name	Size	Rows	Columns
Bore.csv	650 KB	7,459	18
Bore_geology.csv	27 KB	139	12
Environ_metrics.csv	307 KB	6,689	9
Organisms.csv	10,465 KB	90,702	28
Data_source.csv	4 KB	13	11
Sample.csv	5,539 KB	50,834	26
Site.csv	3,929 KB	19,380	26
Source_reports.csv	41 KB	443	7
Taxonomy.csv	767 KB	3,792	26

3.3 Reconciliation of taxonomic nomenclature

Initial data compilation included a thorough review of subterranean fauna taxonomy and para-taxonomy (see also Mokany *et al*. 2018). Troglofauna and stygofauna nomenclature of the final Project database were reviewed a second time by taxonomic experts, Volker Framenau (VWF) and Stuart Halse (SH) respectively, to rectify data entry errors and consider taxonomic changes that may have occurred since a record was first entered into the Bennelongia SQL database. The Atlas of Living Australia (https://www.ala.org.au) served as a reference source for published names; in some cases, recourse to original taxonomic literature was necessary. Unpublished parataxonomic morphocodes are not open to an objective review as there are no standards governing those. Therefore, the original taxonomy as provided in the initial SQL database was used in the analysis. For example, *Prethopalpus scanloni* Baehr and Harvey, 2012, *Prethopalpus scanloni* sl (=*sensu latu*), and *Prethopalpus* nr (=near) *scanloni* were considered three different taxonomic units, although it is not clear if all *P. scanloni* sl or *P.* nr *scanloni* belong to the same species, or if their concepts overlap.

After the nomenclature was considered final for the purposes of the Project, a WA Museum taxonomist with expertise in subterranean fauna was consulted to discuss taxonomic considerations.

3.4 Data exploration and analyses

Different statistical analyses used different subsets of the database defined by parameters which had adequate completeness, variability, and both statistical and biological relevance to the selected research questions with focus on sampling efficacy. Lack of data or data variability and autocorrelation meant that many analyses in relation to biogeographic region could not be conducted, as comparably few datasets were available from outside the Pilbara region.

Temporary preparation and manipulation of parameters were done within the scope of each of the separate analyses to facilitate meaningful and statistically sound results. Methods of data manipulation are described at the beginning of each Results section for each of the analyses.

3.4.1 Data exploration

The data exploration began by defining the variables to be analysed and aspects of their representation in the database, namely:

- number of samples where the data for the variable was recorded
- number of records within each level of discrete variable or numerical distribution of continuous variables
- correlation between potential explanatory variables
- spatio-temporal patterns of variable completeness (e.g., was the variable only reported in a certain area).

Variables with more than 60% null values were considered unsuitable for analyses.

Numeric and integer variables were transformed where required (e.g., log-transformation of abundance). For discrete variables, some values were combined to minimise differences in comparative sample sizes (e.g., combining trap 1, 2, 3, and 4 for troglofauna traps). Sparse values (generally those comprising fewer than 150 sites or samples) within discrete variables were excluded from the analyses.

An exploration of correlation between site data variables (IBRA region, altitude, latitude, longitude, depth to bottom, total visits) and sample data (sample type, visit date, conductivity, pH, and temperature) showed high levels of correlation. Therefore, only one deemed biologically most appropriate for a specific analysis was selected for model inclusion.

Spatio-temporal data exploration was an essential requirement prior to data analyses. This exploration assessed whether different regions could be compared with one another based on the number of records in each, and the temporal distribution of their visits. If sample sizes between regions range from very small to very large, the number of samples itself becomes a significant factor in interpreting the analyses where region is used as a predictor. In the case of this analysis, the Pilbara had many times more samples than any other region. Similarly, if the time interval of sampling does not overlap, then time becomes a significant variable in the analysis when region is used as a predictor. For these reasons, all regions were included in the analyses (biases by region are assumed to be small because the comparative number of samples from regions outside the Pilbara was small) but region was never used as a predictor.

With some exceptions, data were analysed separately for stygofauna and troglofauna. Analyses were undertaken at the lowest taxonomic unit (LTU) identified. If an organism was only identified to family, its family level identification was used as a unique identifier when calculating diversity and richness metrics, and the same for class, species binomial, morphotype, etc. This means that numerous taxonomic levels were used to describe the community within sites, although ca. 41% (troglofauna) and 57% (stygofauna) of identifications overall were to species or morphospecies level (Table 4). In this way, more identifications, and thus biological diversity, can be incorporated into the statistical models. It is recognised that this may inflate overall metrics of biodiversity, because an identification to family level is considered unique even if that family is already represented in the community by something identified to species, but is standard practice in applied ecology of poorly known taxa due to the way that it increases data availability and statistical power (e.g., Jones 2008).

3.4.2 Data analyses

The following analyses were conducted:

- incidence of rare organisms (frequency of occurrence by number of sites where organism was found)
- visits per site
- model of richness by number of visits to a site
- taxon accumulation curves (troglofauna and stygofauna combined)
- sample method efficacy
- community composition and dispersion[1] by sample type (Anderson *et al*. 2006)
- community composition by trap order (when more than one trap was placed at the same site in the same visit)
- sampling interval against richness
- community composition by month.

For an analyses of rainfall data each sampled bore was matched to its closest meteorological station. Analyses of rainfall included:

- cumulative rainfall 7 days prior to sampling against richness
- cumulative rainfall 30 days prior to sampling against richness
- days since storm (where 'storm' was arbitrarily classified as the top 1% event for all records kept from each meteorological station) against richness.

Due to overlapping methodologies, specifically as many stygofauna net hauls in uncased bores collect a considerable number of troglofauna species, data analyses were conducted both on the target trapping method and on the taxa actually found. For example, the analysis of sample method efficacy for stygofauna included a comparison of net and scrape, as troglofauna scrapes sometimes collect stygofauna.

3.5 Analytical software

All data exploration, clean up and statistical analyses were conducted in R version 3.5.2 (R Core Team 2018), including the following software extensions:

- data cleaning and visualisation were undertaken with the packages *Tidyverse* (Wickham *et al*. 2019) and *janitor* (Firke 2020)
- *Parallel* (R Core Team 2018) was used for performance parallel computing
- *vegan* (Oksanen *et al*. 2020) for multivariate statistical analyses
- *broom* (Robinson *et al*. 2002) for model parameter extraction
- *lubridate* (Grolemund & Wickham 2011) for handling and analysing time series and date data types
- *GGpubR* (Kassambara 2020) for data visualisation formatting
- *KableExtra* (Zhu 2020) for table formatting
- *KnitR* (Xie 2020) for presenting analyses in HTML.

[1] Multivariate dispersion (variance) of a group of samples is calculated in a number of statistical analyses in this report by the average distance of group members to the group centroid or spatial median in multidimensional space. To test if the dispersions of groups are different, the distances of group members to the group centroid are subject to an Analysis of Variance (ANOVA). An additional use of this function is assessing beta diversity (Anderson et al. 2006).

4 Results

4.1 Database coverage

The Project database contains data sets from more than 17,000 subterranean fauna sample sites (holes/bores/wells), with the majority of sites located in and around the Pilbara region (Table 1; Figure 2). Data of 51,657 samples were analysed, with almost three times more troglofauna samples than stygofauna samples (Table 3).

Table 3. Summary statistics of data analysed for samples by data source

Data source	Samples	
	troglofauna	stygofauna
BHP	16,617	4,386
Cameco	950	781
Chevron Australia	-	129
Dacian Gold	15	141
Fortescue Metals Group	7,931	2,552
Hancock Prospecting	1,399	681
Hastings	302	253
Pilbara Stygofauna Survey	-	1,112
Rio Tinto	10,186	2,922
Stantec	470	830
Total	**37,870**	**13,787**

The data set analysed included almost 28,000 records (i.e., sum of separate LTU occurrences in each sample), of which about two thirds were stygofauna (Table 4). However, stygofauna were relatively more abundant in samples as the estimated total number of specimens of stygofauna collected were almost by a magnitude larger than that of troglofauna (Table 4). More than 50% of stygofauna records and more than 40% of troglofauna records were identified at the species level, either as described species or by para-taxonomic morphocodes; more than a quarter of all stygofauna records belonged to described species, but only 13% of troglofauna records were described (Table 4).

According to the Project database, less than 10% of all stygofauna records were submitted to the WAM, but almost a quarter of troglofauna (Table 4). However, this is likely an underestimate as not all records may have been reliably marked as such in the respective source databases.

The database included a total of more than 1,000 stygo- and troglofauna LTUs respectively, and the majority of these were species-level LTUs. However, the number of described species as percentage of species-level LTUs differed considerably, with many more stygofauna species being officially described (Table 5).

Table 4. Stygo- and troglofauna records (and estimated number of specimens) analysed in the Project database, including percentage lodged with WA Museum (WAM) and percentage of described species

	Total records (est. no. of specimens)	Higher level identification	Species-level identification	No. of records of specimens belonging to described species	Lodged with WAM
Stygofauna	18,231 (197,351)	7,794 (42.8%)	10,437 (57.2%)	5,220 (28.6%)	2,206 (8.3%)
Troglofauna	9,730 (26,955)	5,747 (59.1%)	3,983 (40.9%)	1,266 (13.0%)	2,293 (23.6%)
Total	**27,961 (224,306)**	**13,541**	**14,420**	**6,486**	**4,499**

Table 5. Summary of taxonomic units in the Project database

	LTUs (total)	LTUs (higher level)	LTUs (species level)	Described species (% of species level LTU)
Stygofauna	1,405	570	852	278 (36.2%)
Troglofauna	1,315	633	694	78 (11.2 %)

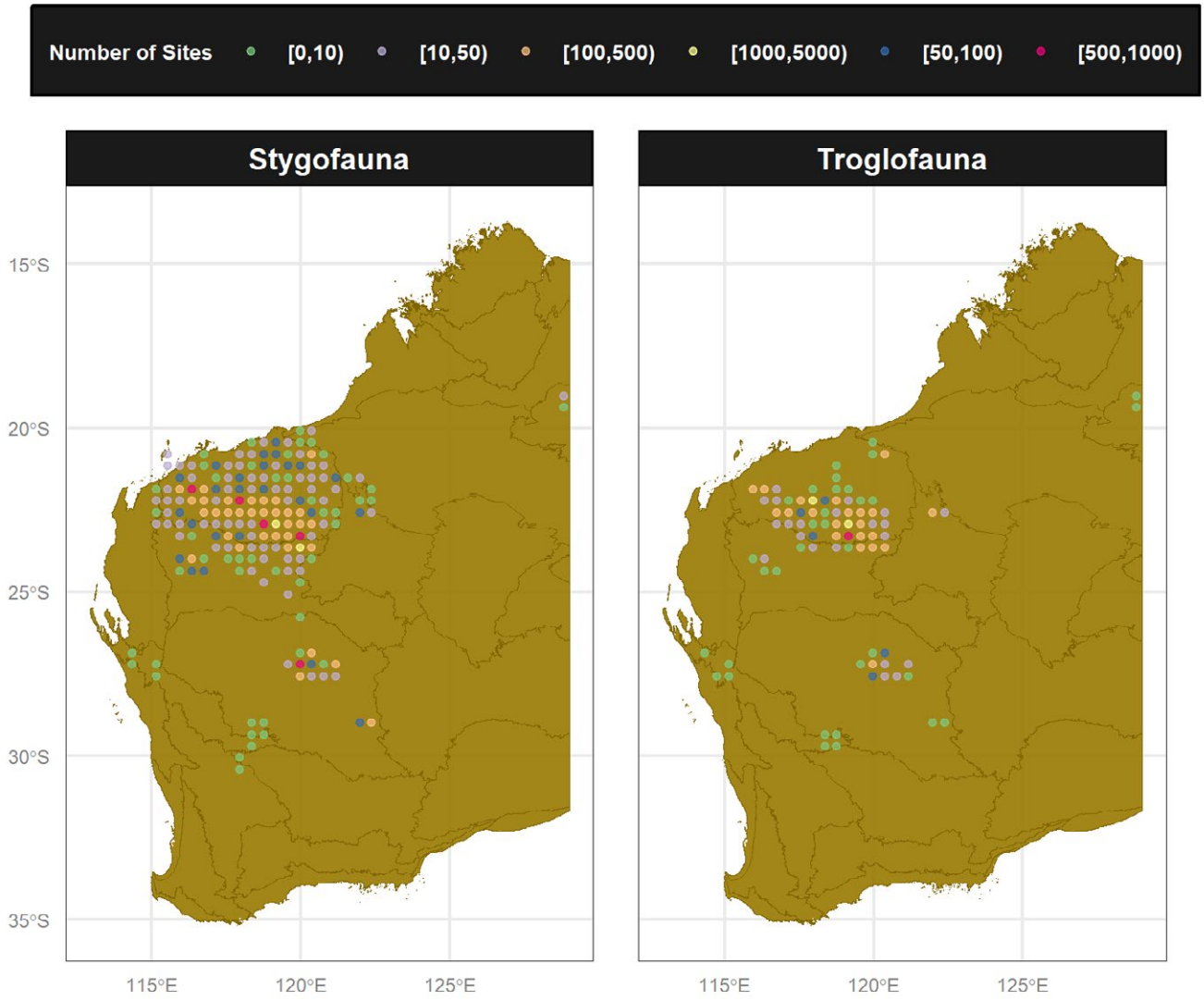


Figure 2. Sites in the Project database where stygofauna and troglofauna were recorded

4.2 Data exploration

Inconsistencies between the data fields in the source datasets meant that the Project database had a large number of missing values for many variables and had an uneven distribution of values within these variables, exemplified by variables of two tables, bores and organisms (Figure 3).

Given the data at hand, many variables that were considered at the outset of this study could not be incorporated in the analyses (Table 6).

In addition, survey data were highly biased towards the Pilbara bioregion (Figure 4).

Of those variables that were deemed most complete for analysis throughout the whole dataset (site data: IBRA region, altitude, latitude, longitude, depth to bottom, total visits; sample data: sample type; visit date; conductivity, pH, temperature), many were strongly correlated limiting their analytical power (Table 7).

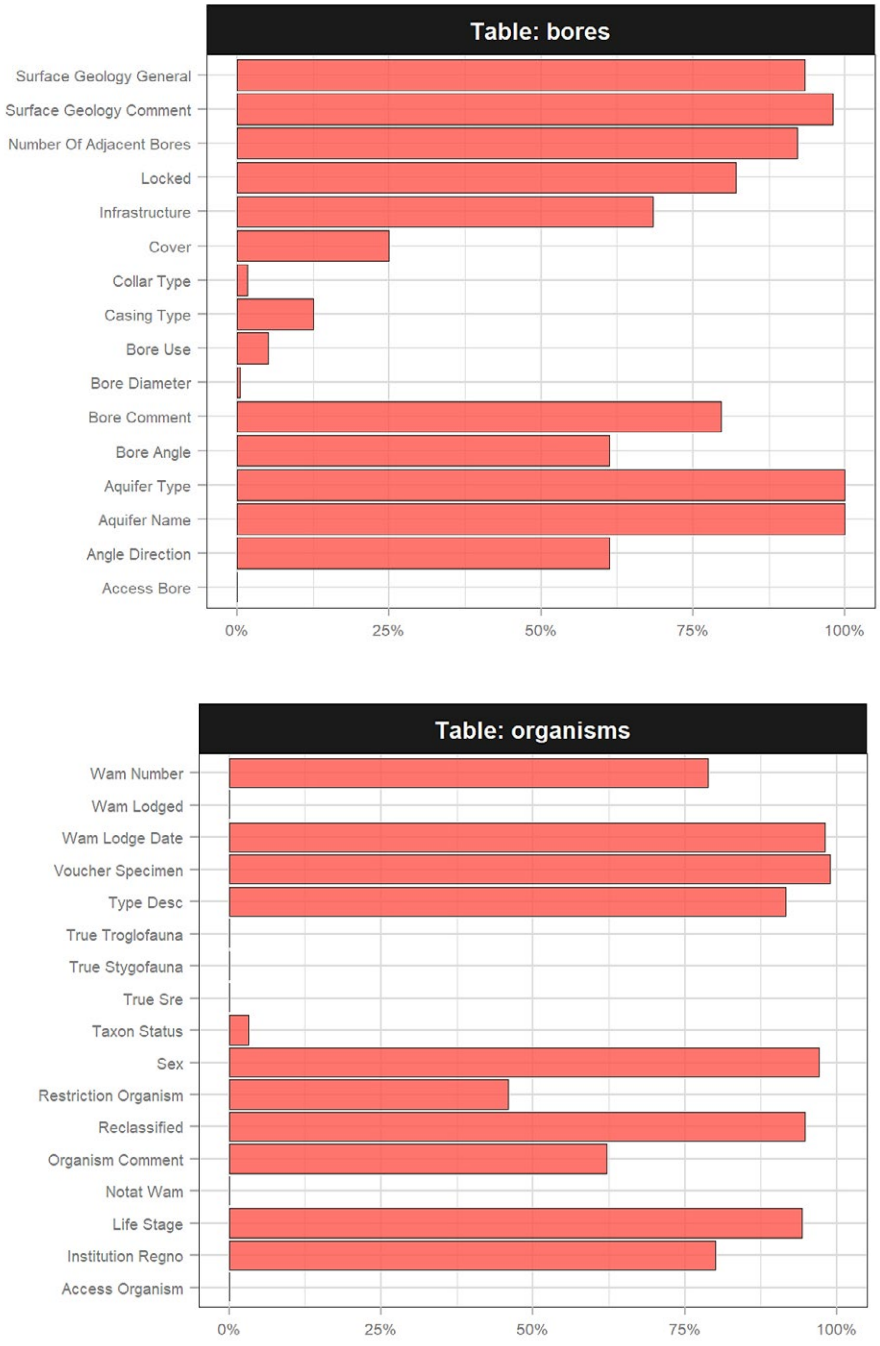


Figure 3. Percentage of missing values in variables within two of the Project database tables

Table 6. Data variables proposed for analysis and their consideration in the current study (see also Figure 1 and Appendix 1 for Project database structure and content)

Category	Variable	Analysed	Comment
Specimen information	specimen code		data fragmented; no statistical analysis but important for curation (data quality and transparency); no 'organism_id' in 'organism' table in Project database
	species code	x	highly fragmented (not standardised) para-taxonomies for undescribed species; accepted 'as is'; unified species codes desirable, incl. minimal taxonomic standards (e.g., diagnosis, public reference specimens; identifier); coded as 'low-est_idnc' in Project database.
	DNA barcode		not considered, as this is a diagnostic tool; but can be incorporated in a database as species diagnosis (% divergence) or specimen identification; no data on eDNA surveys in database – this will require capture of absence/ presence data
	preservation and storage	x	very fragmented data, e.g., WAM registration number only for very few specimens; completeness analysed for spiders
Specimen information (continued)	date recorded	x	sample (not specimen) parameter; records in 'sample' table of Project database; analysed by month and period between samples
	location	x	site (not specimen) parameter; in 'sites' database as geographic coordinates; analysed in distance of bores
	number of individuals per sample		inconsistent between data sources; often estimates; sample poorly defined; in 'organism' table as 'number_identified' and 'life_stage'; analysed based on absence/presence
Sampling methods	trap type	x	poorly standardised categorical variable with a lot of variation (designs); analyses of three broad cat-egories only ('net', 'scrape', 'trap')'; 'sample_type_name' in 'sample' table of Project data-base
	bait type		not in database
	haul net mesh size		not in database
	trap depth	x	continuous variable, but often categorised and analysed as such (e.g., 1-2-3 for troglofauna traps, not depth)
	haul depth		not in database; "depthtobottom" and "depthtowater" in "site_visit" table
Characterisation of bores sampled	type		too fragmented and poorly categorised data
	age		not captured in "bore" or "site-visit" table of database
	depth of hole		"depthtobottom" in "site_visit" table
	angle of hole		lack of data and highly biased dataset (lack of variation – almost all 90 degrees bores); 'bore_angle' in 'bore' table of database
	hole diameter		not sufficient variation for analysis; 'bore_diameter' in 'bore' table
	depth to water table		not captured in database
	drill core samples collected		not captured in database
	physicochemical information		only available for stygofauna (pH, salinity, DO), lack of variation; in table 'environ_metrics' of database

(Table 6 continued following page)

Table 6. Data variables proposed for analysis and their consideration in the current study (see also Figure 1 and Appendix 1 for Project database structure and content)

Category	Variable	Analysed	Comment
Sampling effort	number of bores sampled		requires definition of reference area; not analysed (but distance of bores analysed)
	number of sampling events per bore	x	core analysis; but sampling event poorly defined, here 'site_visit' used
	timing of sampling	x	analysed by month
	haul/scrape/trap number per bore during single survey event		not captured in database
	seasons sampled		season poorly defined category state-wide (e.g., wet/dry; four seasons, indigenous seasons); data analysed by months
	duration of trap deployment		low variance in variable
Survey area information	hydrogeological setting		not in database
	geomorphology/ geology		no standardised dataset available between sources; surface geology not considered to provide reliable data for subterranean environment
	topography		important predictor for troglofauna dissimilarity in Pilbara (Mokany *et al*. 2018); not analysed as little variation outside that region
	climate information including prevailing weather (e.g., rainfall during and preceding)	x	cumulative rainfall prior to sampling analysed for 7 days, 30 days and storm event
Sampling configuration	distance between bores sampled	x	analysed based on geographic coordinates; data not presented here as trivial: community dispersion increases with distance of bores
	number of bores sampled inside/ outside of the impact		not captured in database and potentially changing between surveys (e.g., follow-up survey for project area expansion)
	stratified according to geology and/or hydrology		not analysed as stratigraphic information not captured in database; no external and standardised data available for all samples

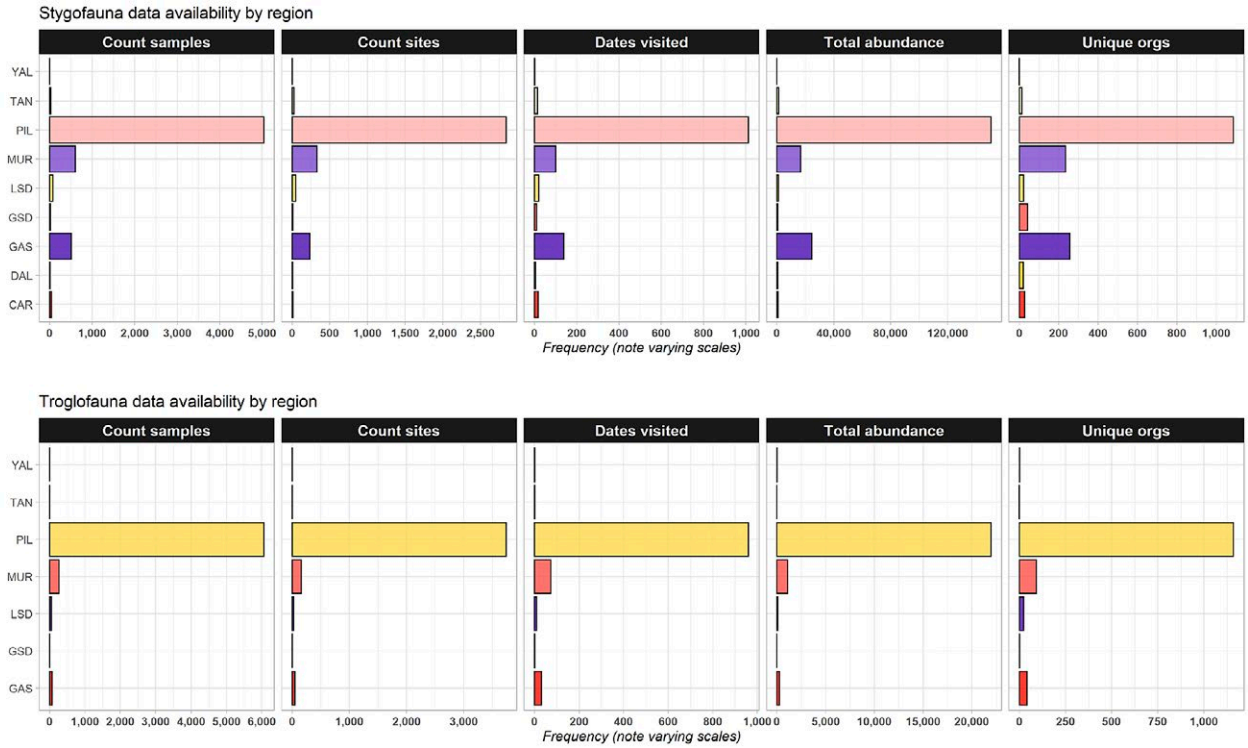


Figure 4. Data availability for Interim Biogeographic Regionalisation for Australia (IBRA) for stygofauna (top) and troglofauna (bottom)

Table 7. Spearman correlation test result (P-value) for non-independence. Asterisks (***) indicate a significant relationship between two variables

	Altitude	Conductivity (ms/cm)	Depth to bottom	IBRA code	pH
Altitude		0 ***	0 ***	0 ***	0.643
Conductivity (ms/cm)	0 ***		0 ***	0.288	0.829
Depth to bottom	0 ***	0 ***		0 ***	0.942
IBRA code	0 ***	0.288	0 ***		0.16
pH	0.643	0.829	0.942	0.16	
Sample type name	0 ***	0 ***	0 ***	0 ***	0.951
Site lat	0 ***	0 ***	0 ***	0 ***	0.51
Site long	0 ***	0 ***	0 ***	0 ***	0.456
Temperature (C)	0 ***	0 ***	0 ***	0.62	0.122
Total visits	0 ***	0 ***	0 ***	0 ***	0.071 *
Visit date	0 ***	0 ***	0 ***	0 ***	0.277

4.3 Data analyses

After data exploration was finalised, the scope of analyses was refined based on the completeness of available data. The main themes of data analyses were:

- Frequency of observations – rare taxa (section 4.3.1)
- LTU richness analysis against number of site visits (sections 4.3.2 and 4.3.3)
- Sample method efficacy (sections 4.3.4 and 4.3.5)
- Influence of rainfall (section 4.3.6)
- Influence of timing of sampling (section 4.3.7).

4.3.1 Frequency of observations – rare taxa

The subterranean fauna data of the Project database is dominated by rare taxa (Figure 5); 63.4% of stygofauna taxa, and 78.6% of troglofauna taxa were found in three or fewer sites/bores (Table 8).

The definition of rarity is arbitrary, but for the purposes of this Project, separating species in any analysis to give consideration to 'rare taxa' (i.e., rare species or morphospecies) was unsuitable, since the majority of taxa were found in three or fewer sites. The most common taxa for both groups, troglofauna and stygofauna, were found in 24 sites (Figure 5).

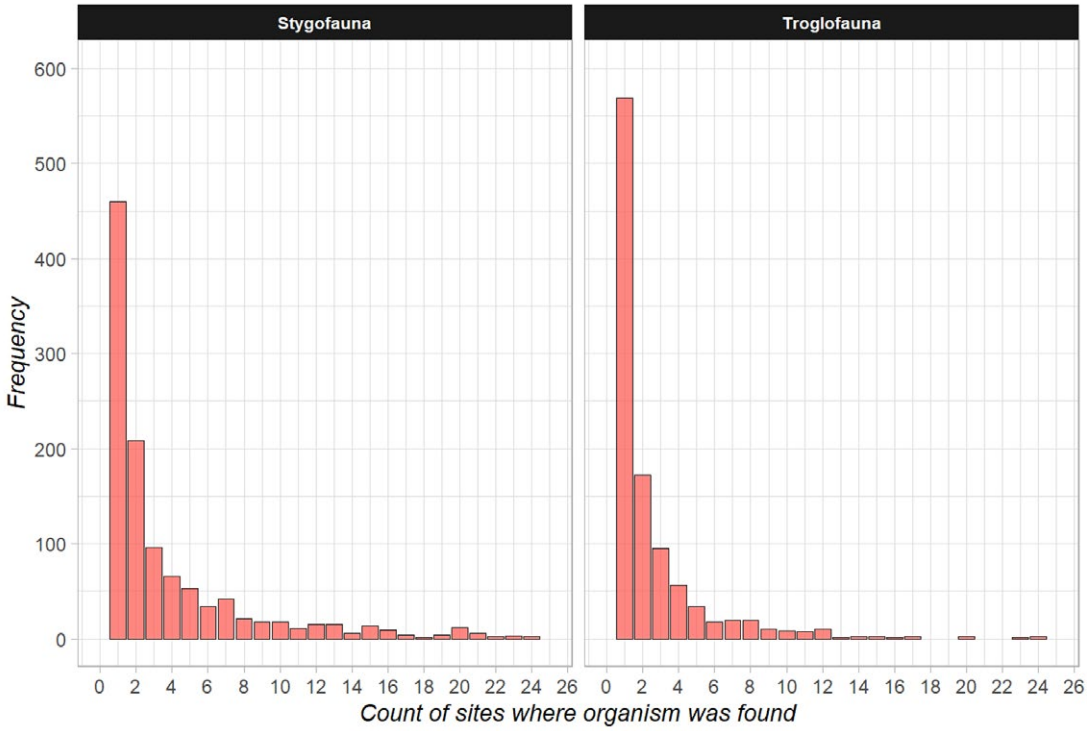


Figure 5. Incidence of rare organisms in the Project database

Table 8. Incidence of rare taxa within the Project database

Group	Total LTU richness	Sites found (bores/ holes/wells)	Taxa	Percent of total	Cumulative percentage
Stygofauna	1,193	1	469	39.3	39.3
		2	204	17.1	56.4
		3	93	7.8	64.2
		4	65	5.4	69.7
		5	57	4.8	74.4
Troglofauna	1,164	1	623	53.5	53.5
		2	201	17.3	70.8
		3	101	8.7	79.5
		4	61	5.2	84.7
		5	40	3.4	88.1

The number of sampling visits per site (bore/hole/well) is concentrated at the lower end of the scale (Figure 6, Table 9). More than 90% of sites were sampled three or fewer times, indicating that the dataset contains baseline survey data rather than monitoring data.

The resulting data frames for richness and total site visits were compiled, and their distributions were examined so that the appropriate transformations, models, and linking functions could be identified. Stygofauna and troglofauna richness, as well as the total visits per site, followed a Poisson distribution.

Having a limited number of sites sampled four or more times (less than 5% of sites) restricts the ability to conduct analyses of species accumulation per visit and statistical tests against species richness, including variability in timing, rainfall, and other variables.

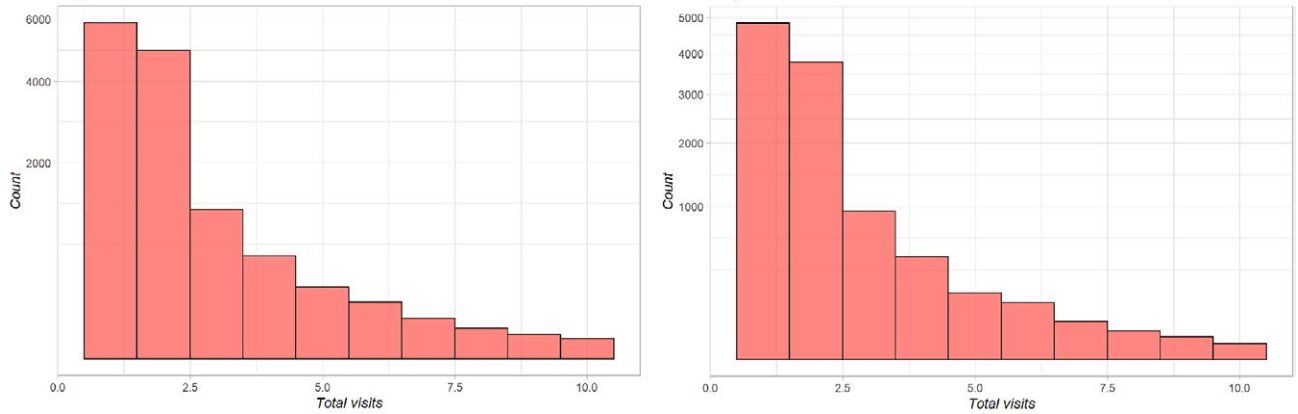


Figure 6. Number of visits per site (holes/wells/bores) where stygofauna (left) and troglofauna (right) were sampled. Figure was truncated at 10 visits

Table 9. Frequency of visits to sites for stygofauna and troglofauna sampling

Total visits	Stygofauna			Troglofauna		
	Frequency	Percentage	Cumulative percentage	Frequency	Percentage	Cumulative percentage
1	5,854	44.7	44.7	4,825	46.2	46.2
2	4,927	37.6	82.2	3,773	36.1	82.3
3	1,155	8.8	91.1	937	9.0	91.3
4	550	4.2	95.2	449	4.3	95.6
5	267	2.0	97.3	190	1.8	97.4
6	168	1.3	98.6	138	1.3	98.7
7	86	0.7	99.2	63	0.6	99.3
8	49	0.4	99.6	35	0.3	99.7
9	31	0.2	99.8	22	0.2	99.9
10	22	0.2	100.0	11	0.1	100.0

4.3.2 LTU richness against number of site visits

A comparison between the community (quantified by LTU richness) and effort (quantified by number of sampling visits) was undertaken to address the question of how many samples should be taken at each site to obtain a representative sample of the community.

These analyses were conducted separately for stygofauna and troglofauna. Richness was calculated at the site level as the total number of unique LTUs obtained from each visit to the site.

Using generalised linear models (GLMs), a significant relationship was found between LTU richness and the number of visits to a site (Table 10, Table 11). A clear and strong positive linear relationship between the number of times a site was visited, and the number of novel taxonomic units recorded was found for both stygo- and troglofauna (Figure 7, Figure 8).

Table 10. ANOVA results for site richness by total visits for stygofauna collection

Term	df	sumsq	meansq	statistic	p-value
as.factor(total_visits)	9	33,910.73	3767.85915	160.1534	0
Residuals	6,435	151,393.44	23.52656	NA	NA

Table 11. ANOVA results for site richness by total visits for troglofauna collection

Term	df	sumsq	meansq	statistic	p-value
as.factor(total_visits)	9	2,894.417	321.601880	141.704	0
Residuals	11,217	25,457.347	2.269533	NA	NA

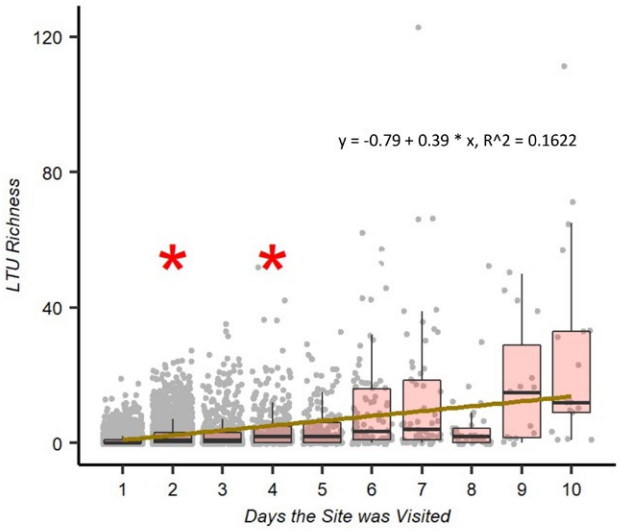


Figure 7. Stygofauna LTU richness relationship to number of site visits, with trendline and equation for Poisson generalised linear model (GLM) displayed on the plot. The asterisk indicates where the change in richness from the previous category (number of days) is statistically insignificant (Tukey HSD test, p-value>0.1)

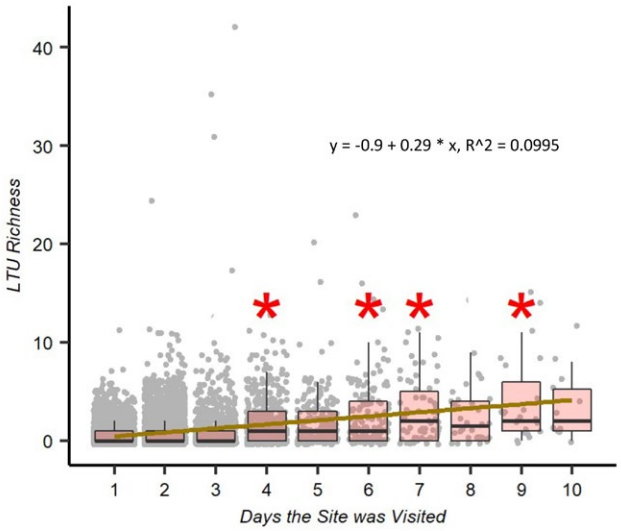


Figure 8. Troglofauna LTU richness relationship to number of site visits, with trendline and equation for Poisson generalised linear model (GLM) displayed on the plot. The asterisk indicates where the change in richness from the previous category (number of days) is statistically insignificant (Tukey HSD test, p-value>0.1)

The initial assumption was that the relationship in both models would be logarithmic (i.e., follow an accumulation curve), traditionally observed between sample effort and richness. However, within the bounds of survey effort represented in this dataset, the relationship was linear for stygofauna and troglofauna. The results indicate that there is no point within the range of sampling effort considered here (up to 10 visits) at which the rate of total taxonomic richness at a site per total number of visits begins to decline, although it is notable that the confidence intervals increase with an increased number of visits per site (Figure 7; Figure 8).

There is an inherent bias in the results of Figure 7 and Figure 8, because sites with initially higher diversity will often have more subsequent sampling events (sometimes requested by regulators), and sites that yield little diversity may have fewer visits. Therefore, the number of site visits should be appropriately adjusted for sites with high richness within the existing sampling regime. However, the timeframe in which samples are collected likely influences diversity as collections further apart in time often increase in diversity (see section 4.3.7).

4.3.3 Taxon accumulation curves

Taxa accumulation curves were calculated with troglofauna and stygofauna combined, because the number of organisms within each group was too low at many sites to allow for separate analysis and because trapping methods collect a considerable amount of non-target by-catch. Sites visited fewer than three times were excluded from the analyses as they do not suit the purpose of this analysis. In the model, very low richness sites where only one new taxon is found each visit skew the results while offering little analytical benefit, therefore, they have been excluded.

Taxon accumulation curves were calculated by first randomly reordering organisms within sampling days and sites, then assigning a +1 to novel nomenclature and a 0 to nomenclature that had occurred prior. A cumulative sum of these 1's and 0's was generated to track accumulation of novel nomenclature. When tracked alongside the row number within sites, an accumulation curve by effort (quantified as number of organisms found at time(t)) was generated.

The results show that the mean percentage of novel species with a subsequent visit increases by some 32.3% to 49.7%, depending on the number of total visits to that site (Table 12). In these analyses we have separated the sites with a different count of total visits (from 3 to 10) to account for the bias of having more site visits to a site with higher perceived richness (Table 12, Figure 9).

Table 12. Visit and site level taxonomic richness summary statistics sorted by the total number of visits to a site. Sites visited fewer than three times have been omitted because they do not suit this analysis

Total visits	Total frequency	Mean novel at last visit (%)	Mean site richness	Peak site richness
3	1,256	49.7	2.8	24
4	579	46.0	3.7	36
5	270	41.9	3.9	21
6	170	40.5	5.5	28
7	86	36.4	6.7	65
8	49	32.9	4.4	22
9	31	32.3	3.6	22
10	22	40.0	11.5	44

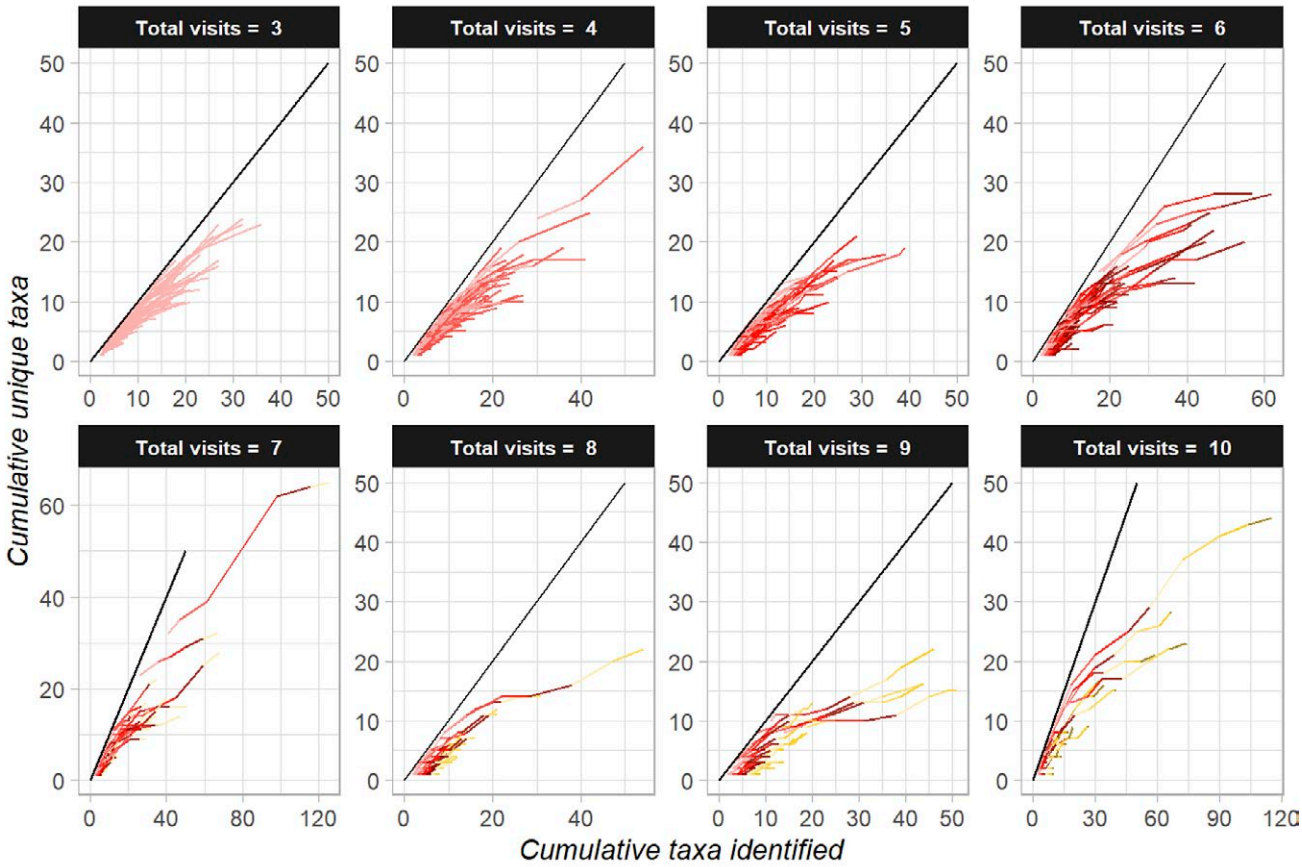


Figure 9. Cumulative unique taxa against cumulative total taxa identified in sites with 3 to 10 total visits. Each individual site is represented by a coloured line and the black line represents 1:1 increase in novel taxa vs all taxa

The graphical representation of the taxa accumulation curves separately by number of visits (Figure 9) alludes to the variability within groups: some of the lines (each representing a site) in the categories of "Total visits = 3" and "Total visits = 4" show little flattening of the curve, while others do. These two categories which show little flattening also happen to be the categories with the most samples and therefore can be inferred from most strongly.

The modelled taxa accumulation curves (novel taxa by cumulative taxa) using a log Poisson model indicate that there is a tendency towards flattening of the curve at very high counts of cumulative sums of taxa found, generally in the hundreds of taxa (Figure 10). The curves were split between sites with different number of visits to account for the potential bias of higher expected richness in sites visited more frequently. The maximum steepness of the curve for sites with 3, 4, 5, 8, and 9 visits (which are all clustered closely together) starts to drop off at around 20 taxa.

In most cases the community was not fully surveyed. Continued effort returned new LTUs at each new sampling visit, although it did decrease with an increasing number of visits and plateaued at about 20%. Zero increase in LTU richness was achieved in seldom cases after 15 visits in this dataset; however, the sample size for sites with more than 10 visits is not large enough to ensure this is not an outlier.

The increased LTU richness with increased effort was a relationship that held regardless of perceived site richness that may have influenced how many visits were made to the site.

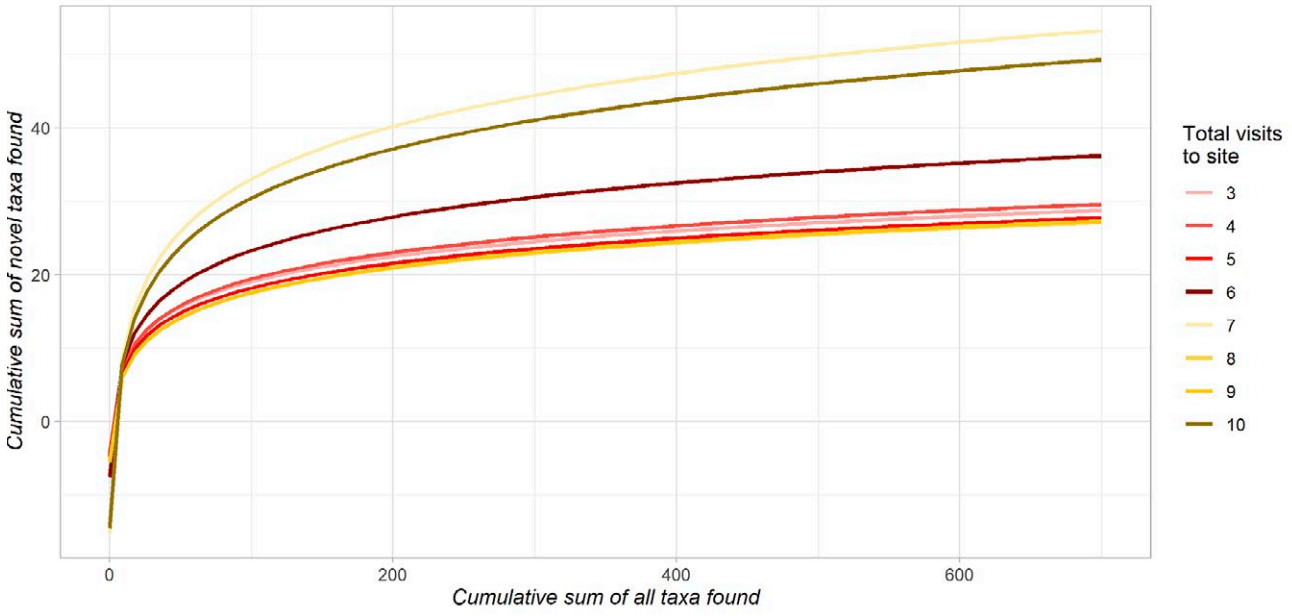


Figure 10. Taxon (in LTUs) accumulation curves for 3 to 10 visits per site

We also analysed the percent increase in LTU richness against visit order (Figure 11, Figure 12). In this analysis, LTU richness is represented as a percent increase – because it is a percent, it becomes a binomial and requires an alternate model; here a Probit GLM was used (Table 13).

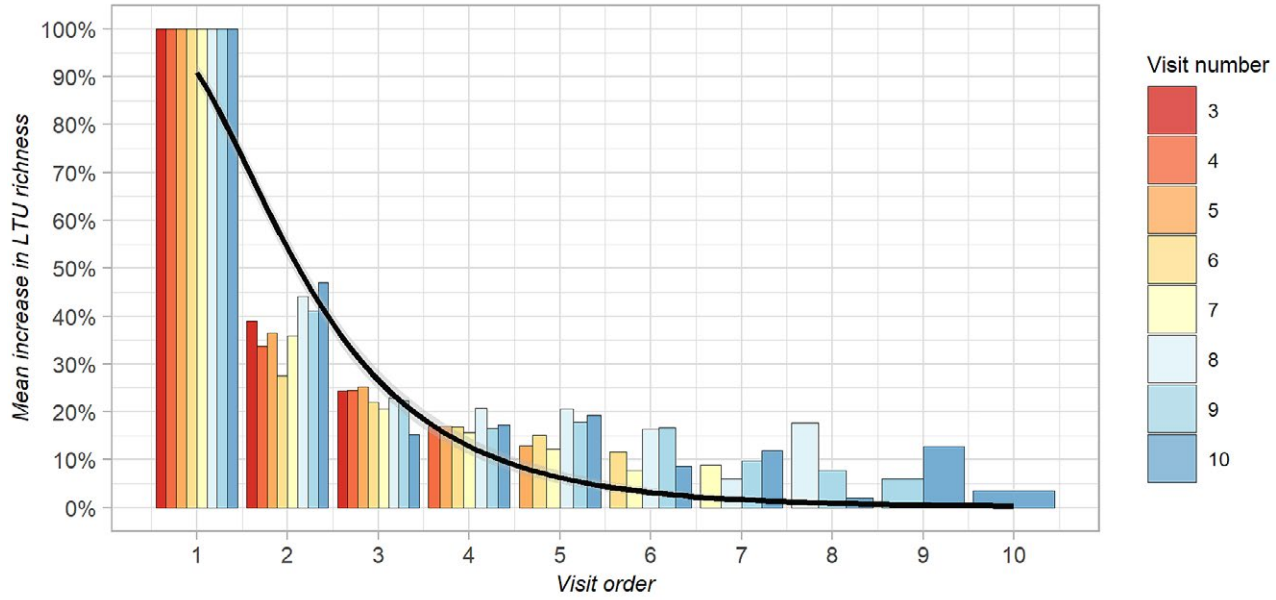


Figure 11. Mean increase in LTU richness with increasing site visits

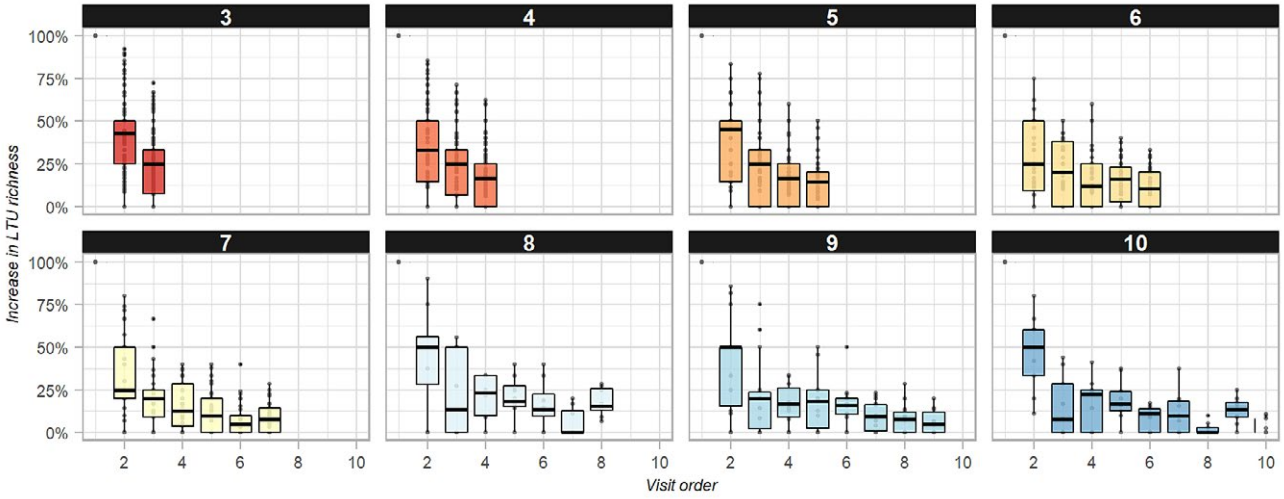


Figure 12. Boxplots of increase in LTU richness with increasing site visits, grouped by total number of site visits

Table 13. Probit GLM results for paired running count of record at site with running novel records at site against visit order. (R2 = 0.4118, n(sites) = 2,463)

Term	estimate	standard error	statistic	p-value
Intercept	1.337159	0.0189158	70.69001	0
log(taxa_accum_visit by visit_order)	-1.194253	0.0153415	-77.84439	0

4.3.4 Sample method efficacy: richness and abundance

For analyses of sample method efficacy, sparse sample types (those with fewer than 500 total organisms captured) were excluded, which included the sample method "net and scrape". The analysis combined both 1) the sampling methods targeting the capture of a group (stygofauna, troglofauna), and 2) those methods which captured that group as by-catch.

The patterns resulting from the analysis are of richness and abundance of sample types for stygofauna (net or scrape) and troglofauna (net, scrape or trap). Traps are not targeting stygofauna and by-catch of stygofauna was too low for consideration in the analyses.

Stygofauna were collected by nets (14,873 records) an order of magnitude higher than by scrapes (1,906 records). Therefore, more stygofauna were collected by nets than by scrapes, also resulting in more organisms per sample and mean taxa by sample in nets (Figure 13).

There were 21 orders of stygofauna represented in the database. Nine orders collected by nets were not collected by scrapes (Figure 14). However, the most common orders broadly overlapped indicating that there is limited bias between nets and scrapes, although scrapes likely only sampled the upper levels of the groundwater through incidental interaction. The abundance of organisms retrieved by nets was also much higher than that retrieved by scrapes; not surprising as stygofauna are essentially only by-catch of scrapes, which target troglofauna (Figure 15).

Troglofauna were collected by three methods: traps (3,674 records), scrapes (4,350 records) and nets as by-catch (1,130 records). All three methods were similar in the mean number of troglofauna organisms retrieved per sample (Figure 16). The mean number of taxa per sample was also similar (between 1 and 2 for all three sampling types). Nets yielded approximately half the number of unique records as traps did, and scrapes were slightly more effective at collecting unique organisms than traps (Figure 16). These results were expected, as troglofauna are considered by-catch in nets.

There were two in 21 orders of troglofauna (Enchytraeida – oligochaete worms, Lithobiomorpha – stone centipedes) that were collected exclusively in traps (Figure 17); however, the categorisation as terrestrial, troglo- or stygofauna of the Enchytraeida is often ambiguous and the records of both may also be an artefact of their rare presence in subterranean fauna. Traps collect a higher abundance of troglofauna than scrapes or nets (Figure 18).

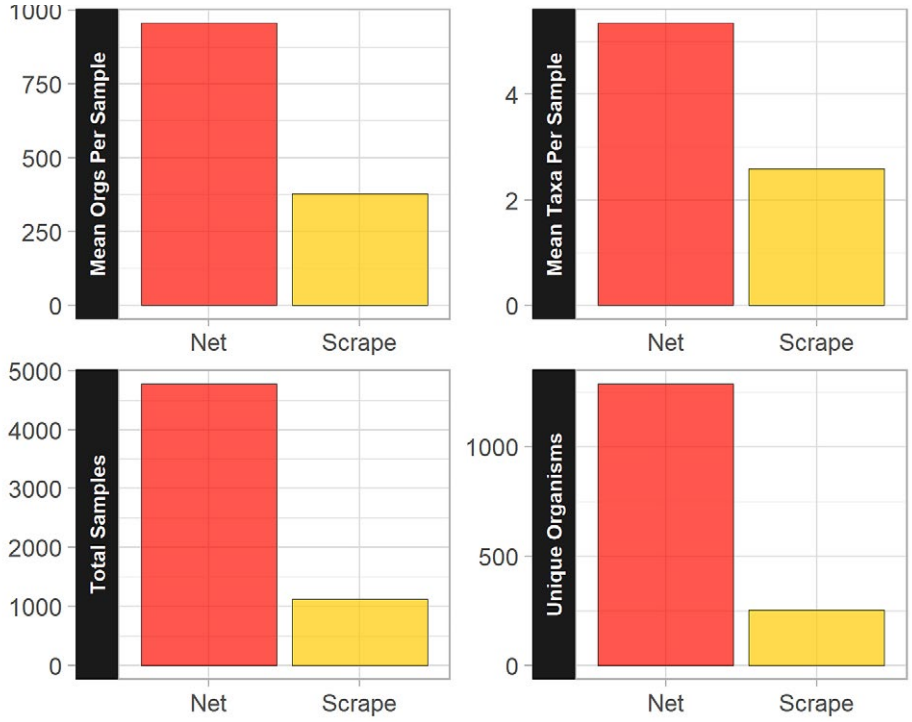


Figure 13. Patterns in richness and abundance of stygofauna by sampling method

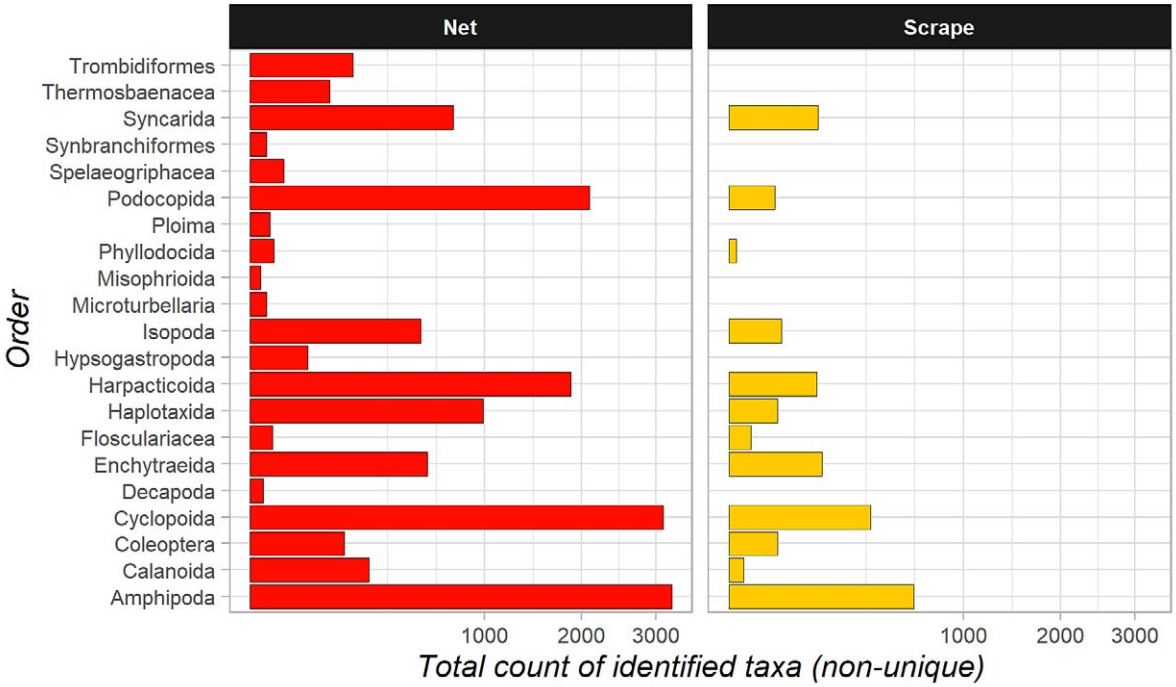


Figure 14. Abundance of identified taxa within orders of stygofauna collected by net and scrape

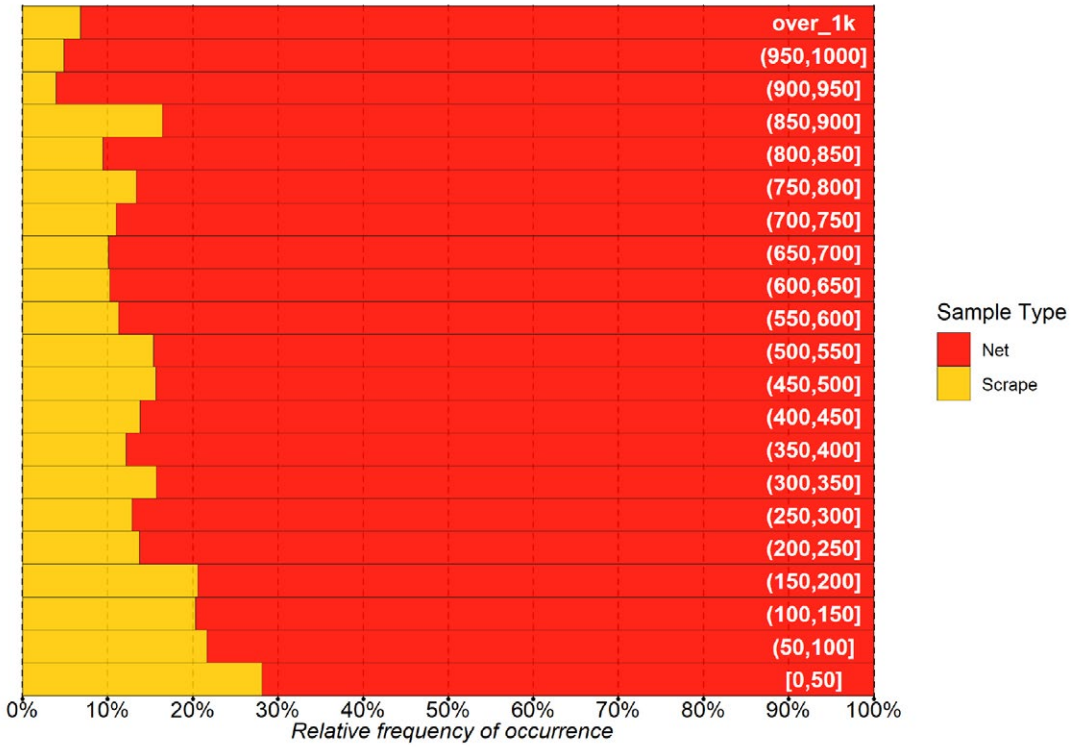


Figure 15. Total stygofauna abundance retrieved by sample type

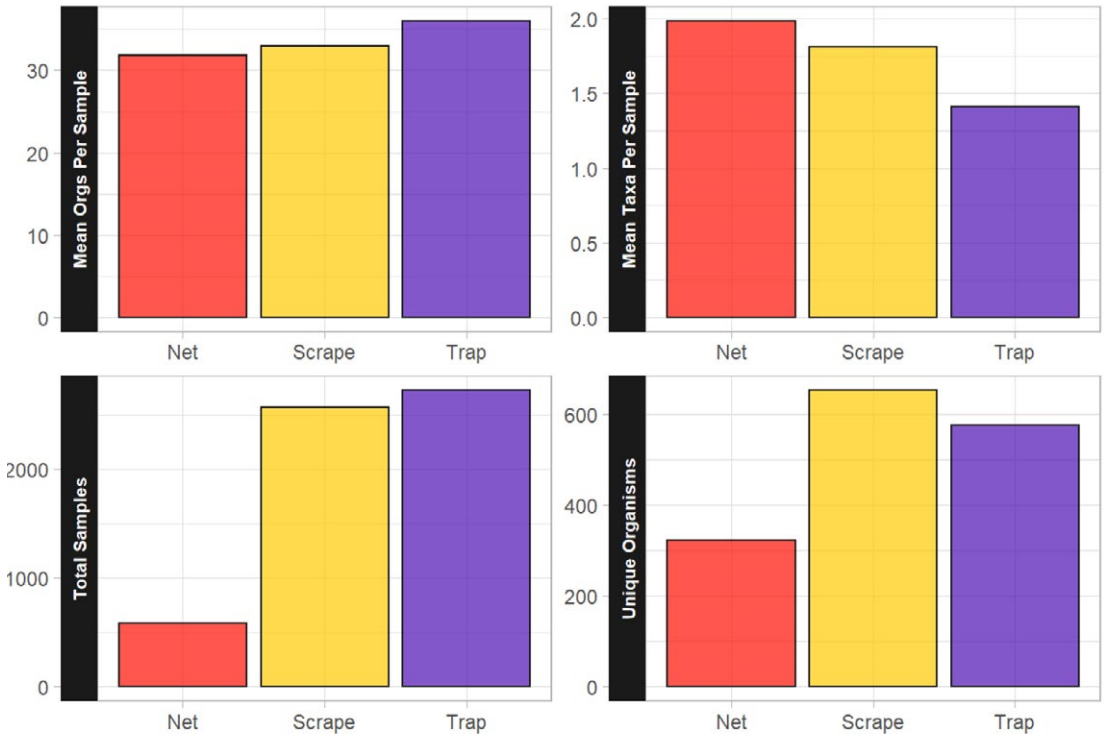


Figure 16. Patterns in richness and abundance of troglofauna by sampling method

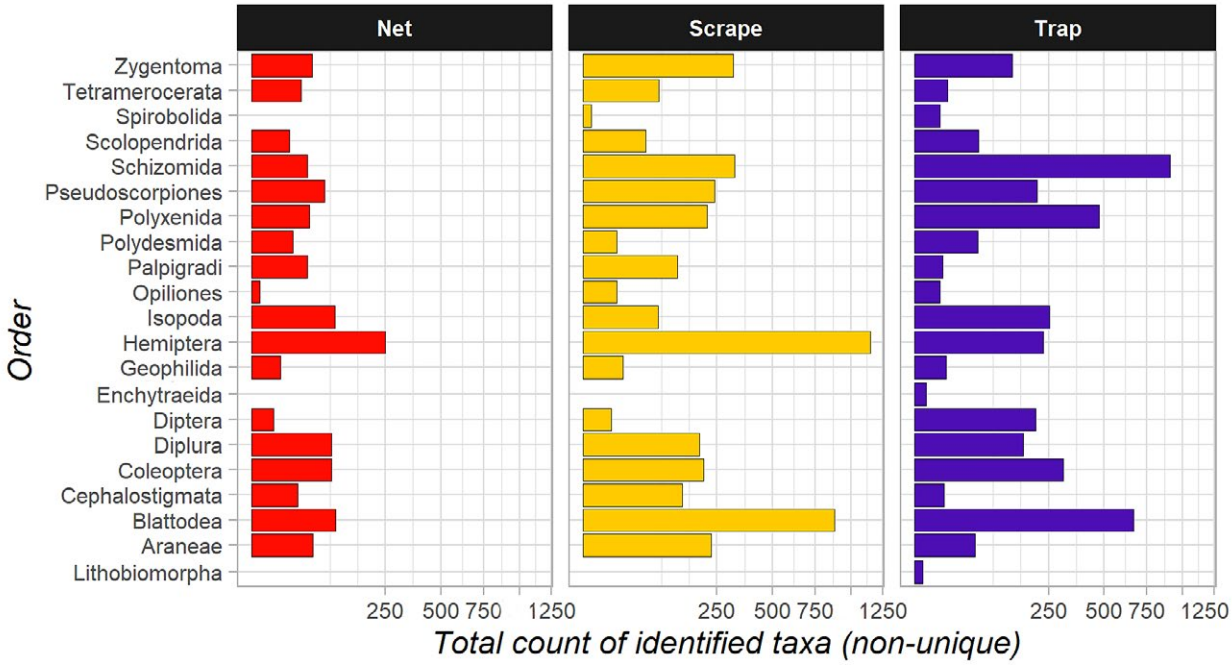


Figure 17. Abundance of identified taxa within orders of troglofauna collected by two sampling methods

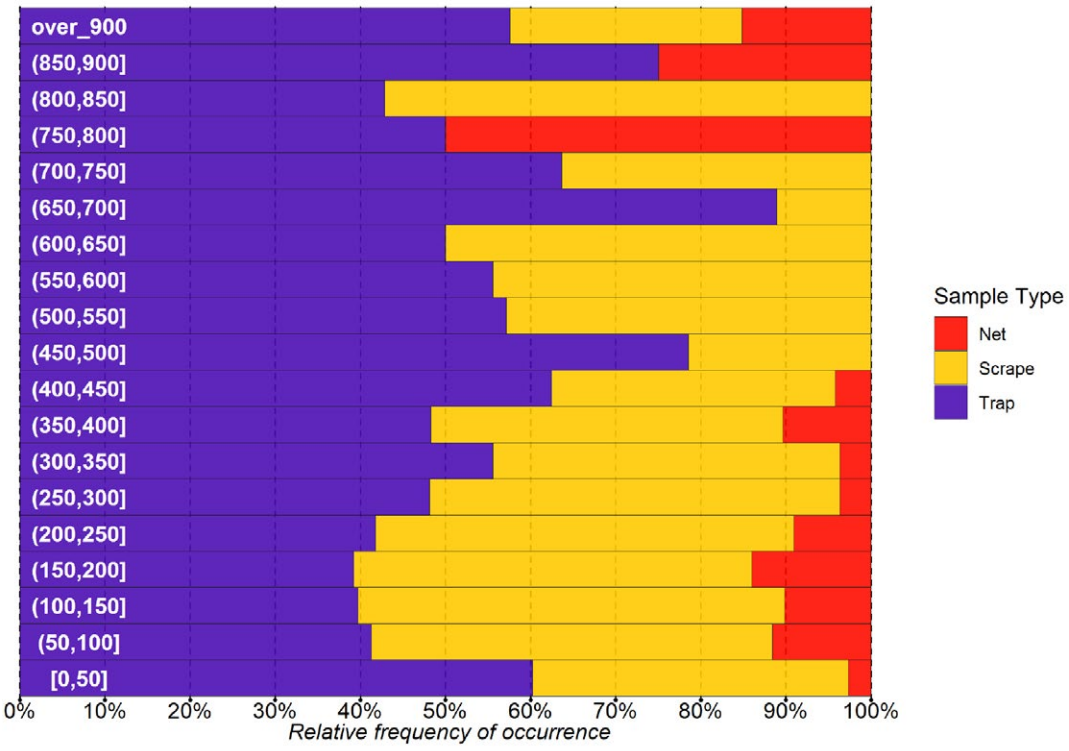


Figure 18. Total troglofauna abundance retrieved by sample type

4.3.5 Sample method efficacy: community composition and dispersion

The analysis of community composition against sampling method aims to determine whether different collection methods are required to collect a representative sample of each community. Statistically, this would mean that the communities found in each of the different sampling methods would be significantly different.

The analysis method selected for this was a PERMANOVA, a statistical method uniquely suited to ecological data that requires no assumptions of normality of distribution and accounts for the ambiguity of a zero in ecological data (does not always represent true absence).

In the case of the Project database, it was necessary to select a taxonomic level that would be appropriate for the analysis first, as the species-based matrix traditionally used did not apply in this case due to the low level of taxonomic resolution in the dataset. To select the most suitable taxonomic level, ANOVAs were run for stygofauna and troglofauna separately.

Stygofauna showed almost all variation at the class level (Table 14), with a sharp dip at order and an increase in variation again at the level of family. Going from family to genus increased the explained variation, but not significantly compared to the increase in degrees of freedom. Species level variation was greater than the number of samples, so it could not be used. Therefore, stygofauna PERMANOVA analysis was run at the family level.

For troglofauna, most of the variation was seen at the class level (Table 15) but using order would increase the variation significantly due to its increase in degrees of freedom. Family showed a much weaker relationship with collection method and genus level identifications cannot be analysed due to the poor taxonomic resolution of troglofauna at that taxonomic level. Therefore, order was used for the troglofauna PERMANOVA analysis.

With rank selected, an analysis data frame was created where community was quantified across each sample at that rank. Because there was a wide range of values for abundance of organisms, these were log + 1 transformed, then rounded up to the next whole value (e.g., 1.01 to 1.99 all became 2 in this method). This maintained the count type of the value, and zero as a zero, while reducing the overall range of the values. It was a necessary step because PERMANOVA may not be sensitive to distribution, but it is sensitive to overdispersion.

Table 14. ANOVA results for variation in trap type by taxonomic ranks within stygofauna

Term	degrees of freedom	sumsq	meansq	statistic	p-value
class	8	35.731020	4.4663775	73.674168	0
order	10	7.531912	0.7531912	12.424103	0
family	31	21.378918	0.6896425	11.375850	0
genus	119	41.107368	0.3454401	5.698132	0
species	521	163.290418	0.3134173	5.169908	0
residuals	9,378	568.526116	0.0606234	NA	NA

Table 15. ANOVA results for variation in trap type by taxonomic ranks within troglofauna

Term	degrees of freedom	sumsq	meansq	statistic	p-value
class	6	161.05032	26.8417202	101.098080	0
order	9	155.51930	17.2799223	65.084017	0
family	24	98.28001	4.0950005	15.423627	0
genus	40	38.45392	0.9613481	3.620872	0
species	404	314.29518	0.7779584	2.930144	0
residuals	3,348	888.89996	0.2655018	NA	NA

The data frame was then reduced to only samples that had multiple collection types at the same site within a two-year span. Restricting the data in this manner reduced dispersion and controlled for other variables that may affect community composition such as geology and climate. Samples which returned no organisms were also removed from the analyses. Finally, the community matrix was submitted to calculations of ecological distance using the Chao method.

Overlaps in the notches on the boxplots (Figure 19, Figure 20) signify that the two means do overlap and therefore, the dispersion between groups is not significantly different.

The results for stygofauna (Table 16, Figure 19) and troglofauna (Table 17, Figure 20) show that scrapes collected the most dispersed (= diverse) samples, although the difference to nets is not significant in stygofauna. However, it is also worth noting that Figure 14 and Figure 17 demonstrated that the detection of orders in relation to sampling method differed mainly in the rare taxonomic orders.

Table 16. Stygofauna analysis of variance results for PERMDISP[2] calculated dispersion within groups (sampling methods) where null hypothesis is that dispersion does not differ between groups

Term	degrees of freedom	sumsq	meansq	F value	Pr(>F)
groups	1	0.410853	0.4108530	3.169962	0.0760437
residuals	292	37.845583	0.1296082	NA	NA

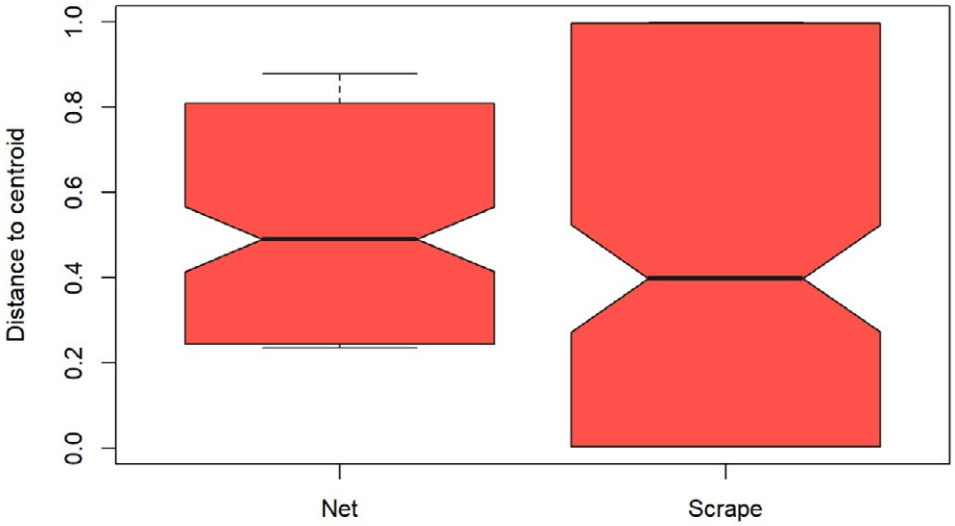


Figure 19. Stygofauna community dispersion between sample methods

[2] Anderson's (2006) procedure for the analysis of multivariate homogeneity of dispersion.

Table 17. Troglofauna analysis of variance results for PERMDISP2 calculated dispersion within groups (sampling methods) where null hypothesis is that dispersion does not differ between groups

Term	degrees of freedom	sumsq	meansq	F value	Pr(>F)
groups	2	0.0643778	0.0321889	3.953608	0.0193313
residuals	2,057	16.7473864	0.0081417	NA	NA

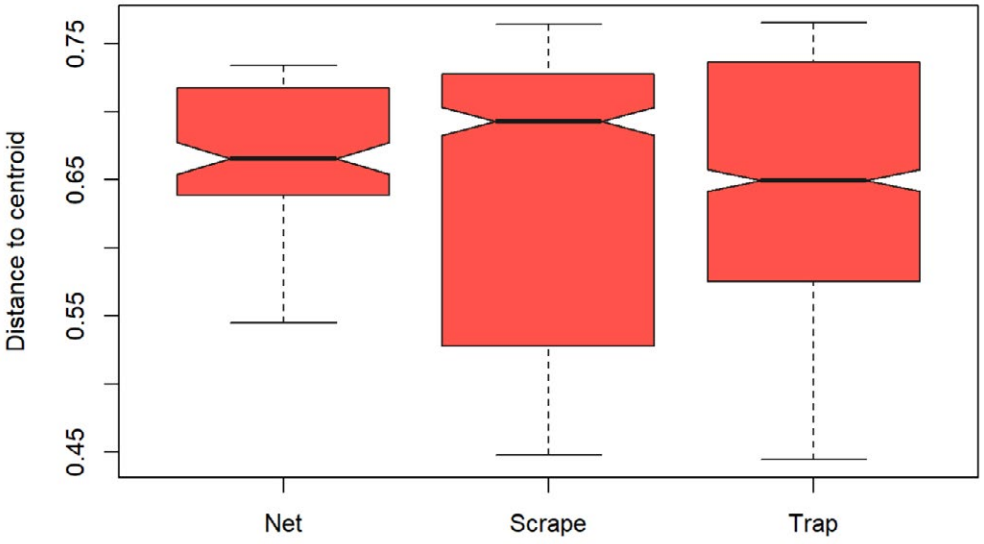


Figure 20. Troglofauna community dispersion between sample methods

The PERMANOVA above answered the question of whether overall community composition collected by different sampling methods was significantly different, regardless of location. The results of the PERMANOVA below indicates whether the observed community composition within the same site was influenced by sampling method, with site_id added as factor (Table 18, Table 19).

Sampling method is a significant determinant of the observed community composition of troglofauna (Table 19), but less so for stygofauna (Table 18). Scrapes and nets collect similar organisms (not surprising as they are essentially the same sampling method), but traps collect a distinct assemblage.

Graphical representation of the PERMANOVA analyses (Figure 21, Figure 22) show the distance and overlap between communities.

Table 18. PERMANOVA within site comparison of collection method differences in stygofauna community composition at family level. Null hypothesis is that communities do not differ

Term	Degrees of freedom	SumsofSqs	Meansq	F model	R^2	Pr(>F)
site id	87	67.625856	0.7773087	9.672058	0.7871725	0.7664671
Site id:sample type	88	8.800737	0.1000084	1.244405	0.1024416	0.7664671
residuals	118	9.483238	0.0803664	NA	0.1103859	NA
total	293	85.909831	NA	NA	1.0000000	NA

Table 19. PERMANOVA within site comparison of collection method differences in troglofauna community composition at order level. Null hypothesis is that communities do not differ

Term	Degrees of freedom	SumsofSqs	Meansq	F model	R^2	Pr(>F)
site id	772	502.8628	0.6513767	2.491281	0.5365176	0.001996
Site id:sample type	793	288.2514	0.3634948	1.390236	0.3075430	0.001996
residuals	559	146.1576	0.2614626	NA	0.1559394	NA
total	2,124	937.2717	NA	NA	1.0000000	NA

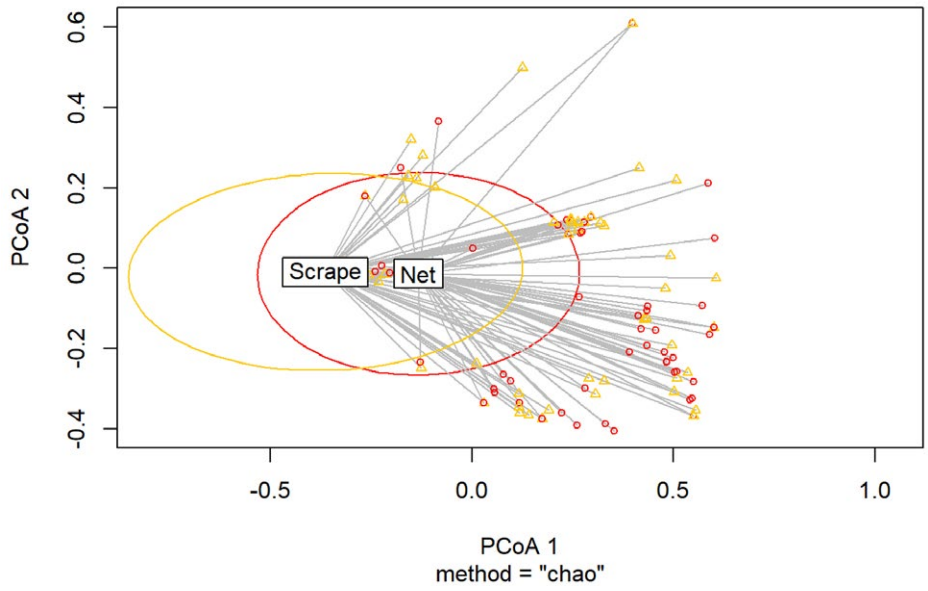


Figure 21. Dispersion of stygofauna by sample collection method

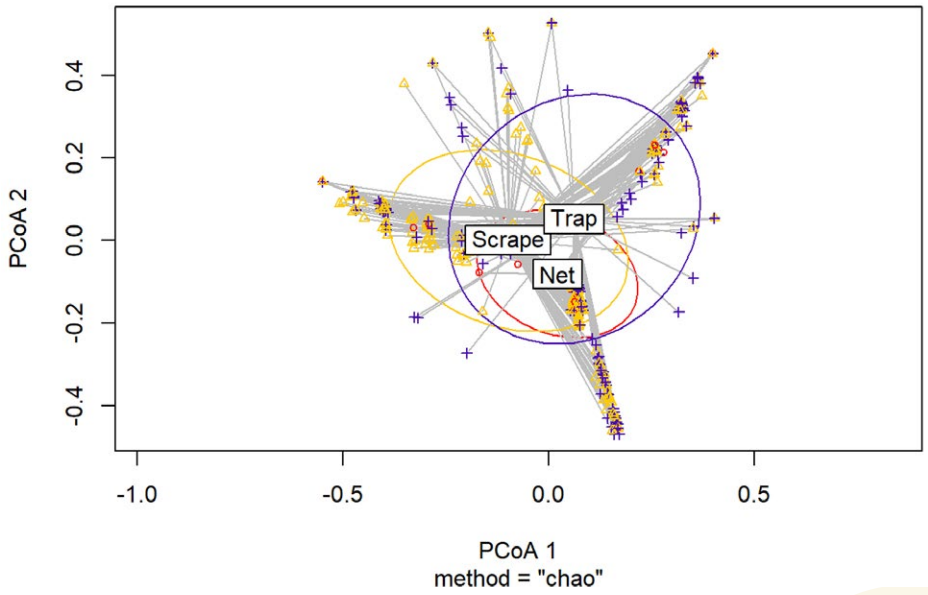


Figure 22. Dispersion of troglofauna by sample collection method

New PERMANOVA coefficients were calculated for stygofauna (Figure 23) and troglofauna families (Figure 24) and show which families affect selectivity of sampling type the most. For stygofauna, these include copepods in the family Cyclopidae, amphipods in the family Paramelitidae, annelid worms (Phreodrilidae), parabathynellids and candonid ostracods (Figure 23).

For troglofauna selectivity of sample types was mainly influenced by cockroaches (Nocticolidae), pin-cushion millipedes (Lophoproctidae), shot-tailed whipscorpions (Hubbardiidae), subterranean planthoppers (Meenoplidae) and fungus gnats (Sciaridae) (Figure 24).

On some occasions, multiple traps were set at different depths at the same site. The trap depth order (as recorded by the consultants undertaking the work) was analysed by PERMANOVA against troglofauna community composition as represented by taxonomic order. Trap depth could not be analysed, as it was not recorded consistently for the traps. The results showed community dispersion was not significantly different between traps of different depth order (Figure 25).

PERMANOVA coefficients were also calculated for troglofauna families in relation to trap depth order (Figure 26) and showed that short-tailed whipscorpions (Hubbardiidae), flies (Sciaridae), cockroaches (Nocticolidae) and pin-cushion millipedes (Lophoproctidae) affect selectivity of trap order the most.

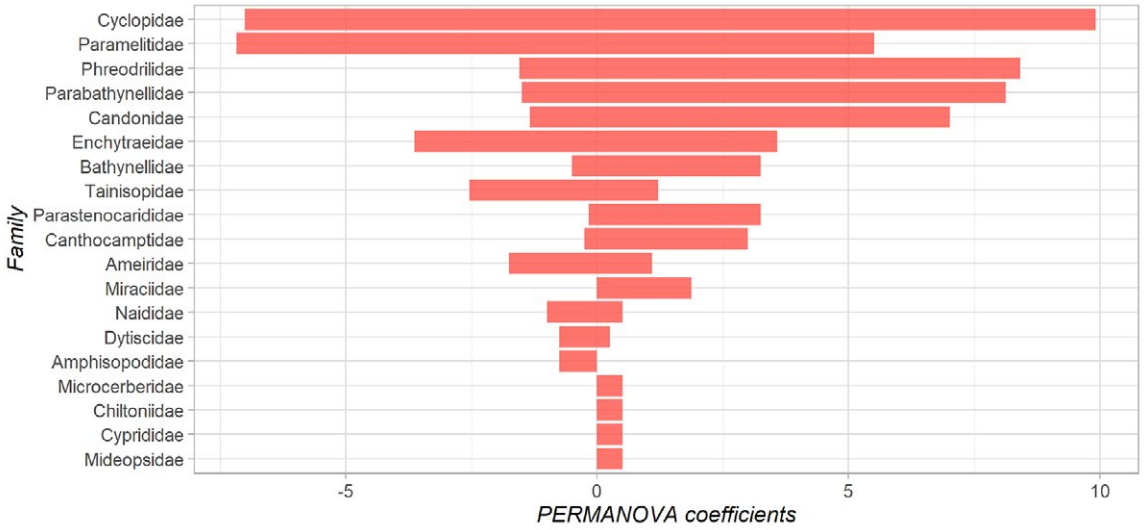


Figure 23. Sampling method impact on stygofauna (by family)

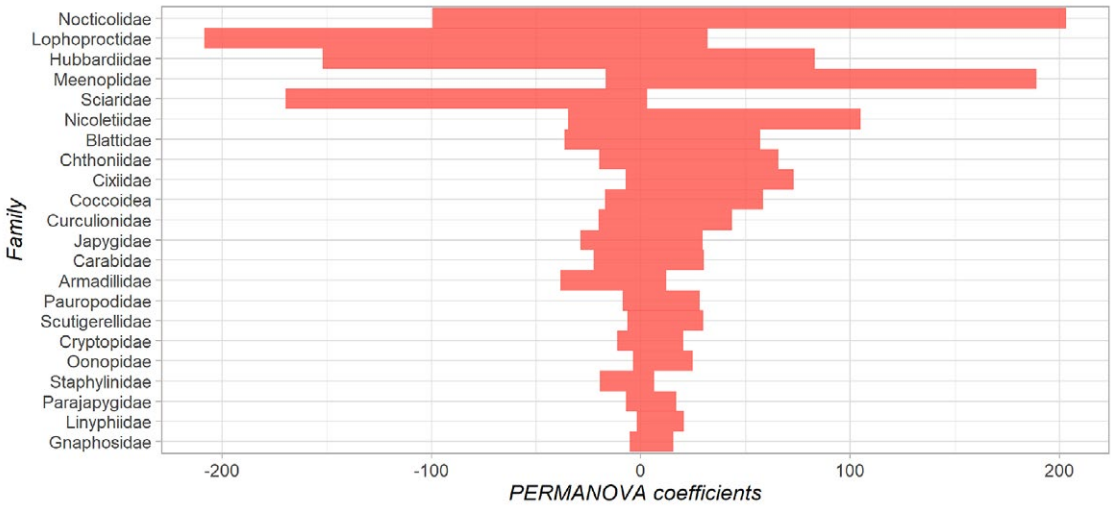


Figure 25. Sampling method impact on troglofauna (by family)

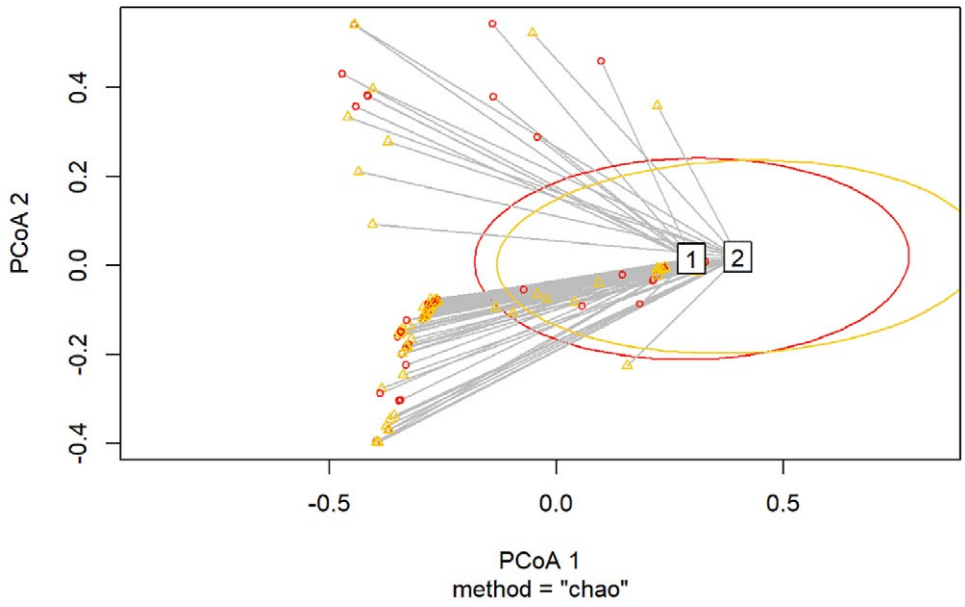


Figure 25. Troglofauna community dispersion by trap order for two traps

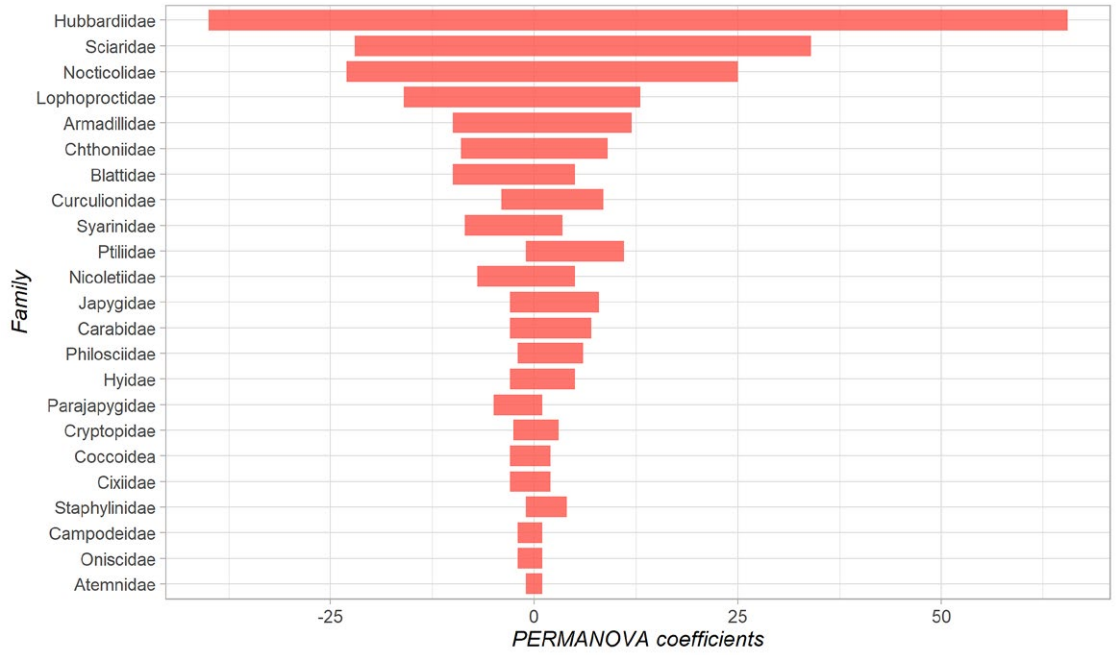


Figure 26. Trap order impact on troglofauna (by family)

4.3.6 Influence of rainfall

All daily rainfall measurements that did not occur within a 30-day buffer from the sampling date were removed, as were low quality rainfall measurements ('low quality' as attributed by the Bureau of Meteorology). Three tailored metrics of rainfall were used against LTU richness to answer whether rainfall has an observable effect on sample richness, i.e., 7-day and 30-day total cumulative rainfall prior to a sampling visit and storm events.

Storm events were identified as events where the 3-day rainfall total is greater than the station's local mean 3-day rainfall) and the daily rainfall is also greater than the station local mean daily rainfall (which is akin to the top 1% events for that station). The number of days between a storm event and a sampling event were calculated for the 'storm' metric.

LTU richness is a count metric, so a Poisson GLM was used to quantify the relationship between richness and the rainfall variables. Separate models were run for each of the rainfall variables because their interaction was not relevant to the interpretation of the analysis. Including them in the same model would have also made comparison of their relative importance difficult since interest was in their distinct, not additive, relationship with richness.

The 7-day cumulative rainfall showed a slight negative relationship with both stygofauna LTU richness (Figure 27) and troglofauna LTU richness (Figure 28), while the 30-day and the storm metrics had no significant relationship. This result is marginal and influenced by the high number of samples used in this analysis and should be interpreted with caution.

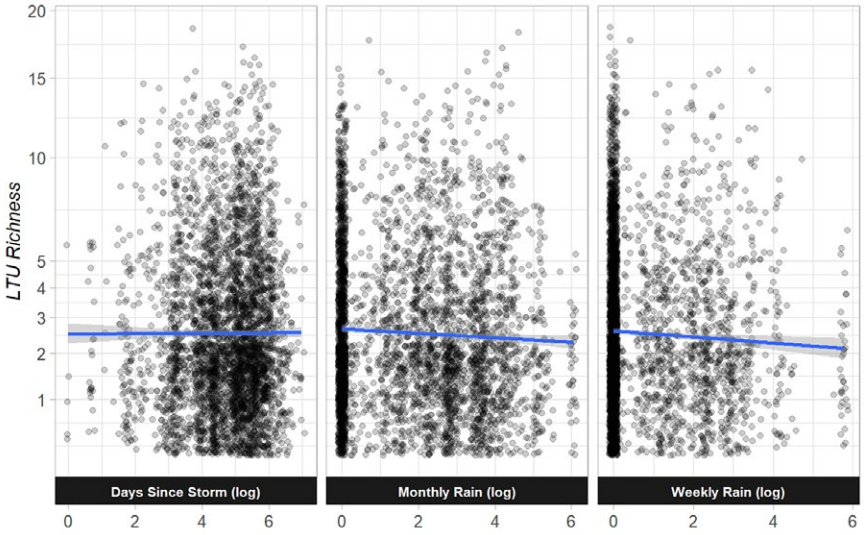


Figure 27. Stygofauna LTU richness plotted against rainfall metrics (days since storm, 30-day and 7-day cumulative rainfall)

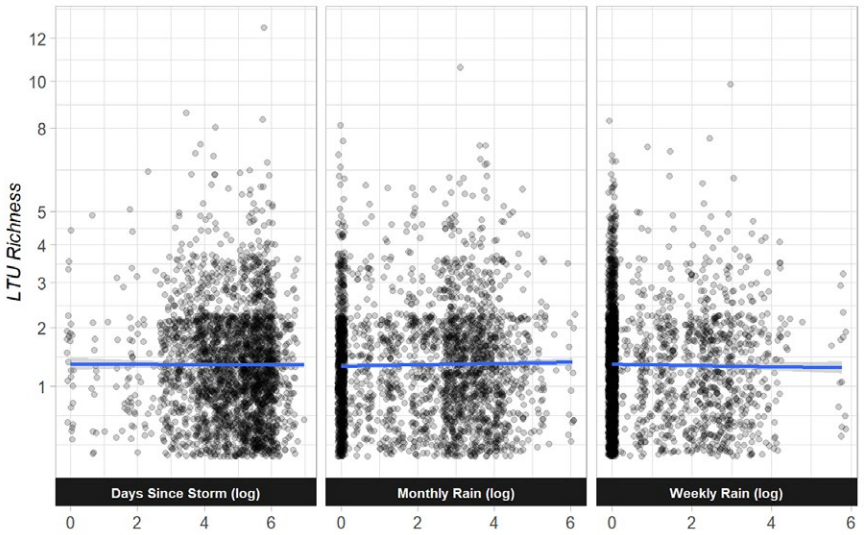


Figure 28. Troglofauna LTU richness plotted against rainfall metrics (days since storm, 30-day and 7-day cumulative rainfall)

4.3.7 Influence of timing of sampling

Timing of sampling was examined against community composition at LTU level to answer whether: (1) samples collected further apart in time were more different to each other; and (2) the community composition changes across months. Only sites with three or more sampling visits producing at least one count of a unique organism were used in these analyses.

The analysis followed three steps. Firstly, ecological distance matrices using the Chao method were calculated for all samples within sites where sampling occurred three times or more, with a matrix calculated for each site independently. Sites with no overlap between any of the visits were removed because a null matrix cannot be used in the Mantel test (final step).

Secondly, a separate matrix was calculated for those sampling trips where there was at least some overlap between visits. In this second matrix, each row and cell combination contained the time between those samples as a decimal fraction of a year (to keep the data range lower than using absolute values).

Thirdly, the two matrices were compared using a Mantel test and Pearson correlation, effectively calculating the linear relationship between temporal and ecological distance between samples collected at the same site. The results of the Mantel test and Pearson correlation show the likelihood of a linear relationship, indicated by the significance, and the direction of the relationship, indicated by the coefficient.

The results indicate no overall significance between length of the sampling interval and the dissimilarity of samples (high diversity; Figure 29); however, where a significance relationship does exist, it is strongly positive for both stygofauna and troglofauna (Figure 30), which means temporally distant sites are also ecologically distant sites. This suggests that spacing out sampling trips aids in obtaining a more representative sample of the full community.

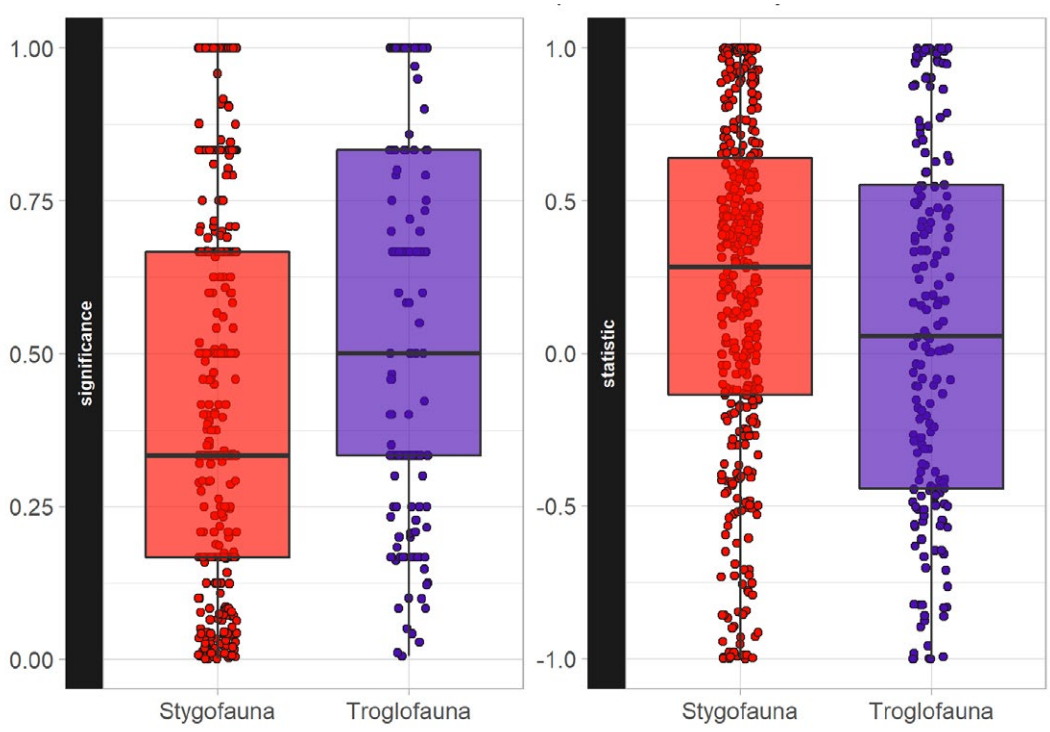


Figure 29. Mantel test result for time between sampling events and community difference

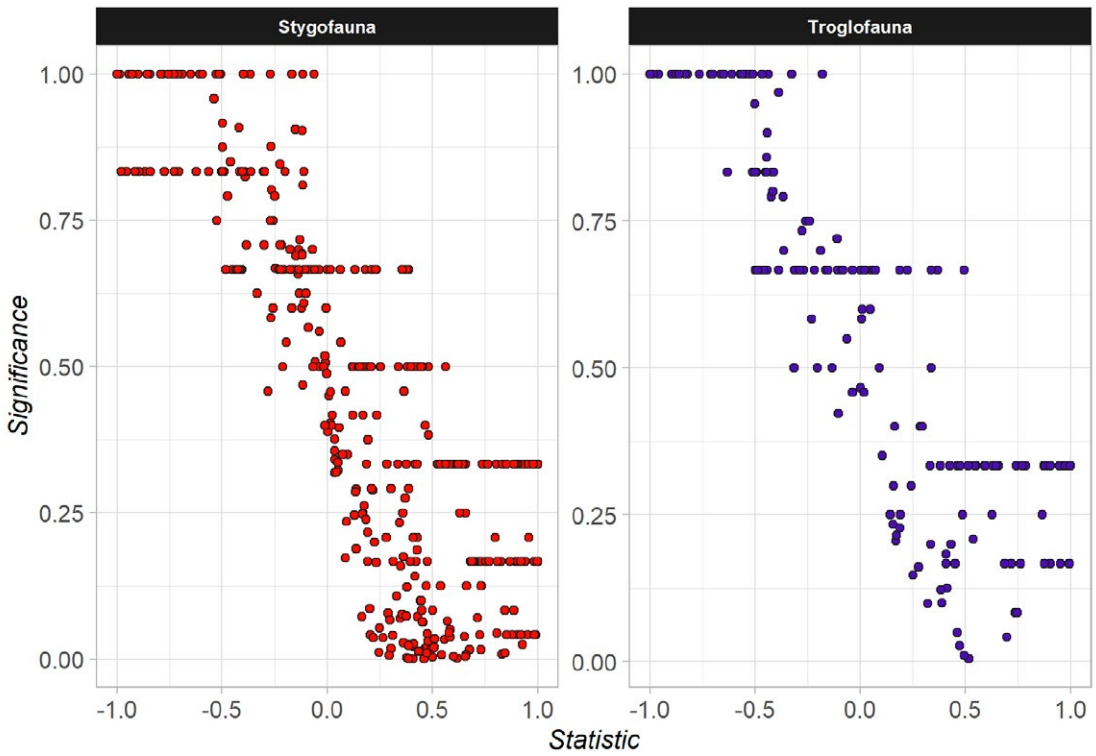


Figure 30. Relationship between Mantel test significance and statistic

A community composition analysis by month was undertaken with months used as a proxy for seasonality, since seasons are hard to define in data analyses that encompass a large geographic area with different climatic regions.

For this analysis, community composition was quantified at the family level, not LTU, because the test was run once for all sites with stygofauna and once for all with troglofauna. Sites with fewer than two visits were discounted in this analysis because the comparison is constrained within sites.

When looking at the results (Figure 31, Figure 32), dispersion between months varies. However, the dispersion between months is far lower than it is for collection method (Figure 21, Figure 22). For stygofauna, samples in January were more dispersed (diverse) than any other month (Figure 31). For troglofauna, the most diverse sampling month is March (Figure 32). These differences were statistically significant (stygofauna: PERMDISP2, $F = 5.08$, $df = 11$, $p < 0.001$; troglofauna, $F + 2.68$, $df = 11$, $p < 0.01$).

The PERMANOVA results displayed in Figure 33 and Figure 34 show the divergence of communities between months. This PERMANOVA analysis addressed whether there are highly dissimilar months in which to schedule sampling visits. The analyses were constrained within site to ensure that regional differences in composition did not influence the relationship between month and diversity.

The differences in community composition per month are less evident for stygofauna (Figure 33) than they are for troglofauna (Figure 34). Some overlap exists in both groups, but the months influence community composition significantly.

These results indicate that if only two surveys are to be undertaken, these should be aimed at different months if possible.

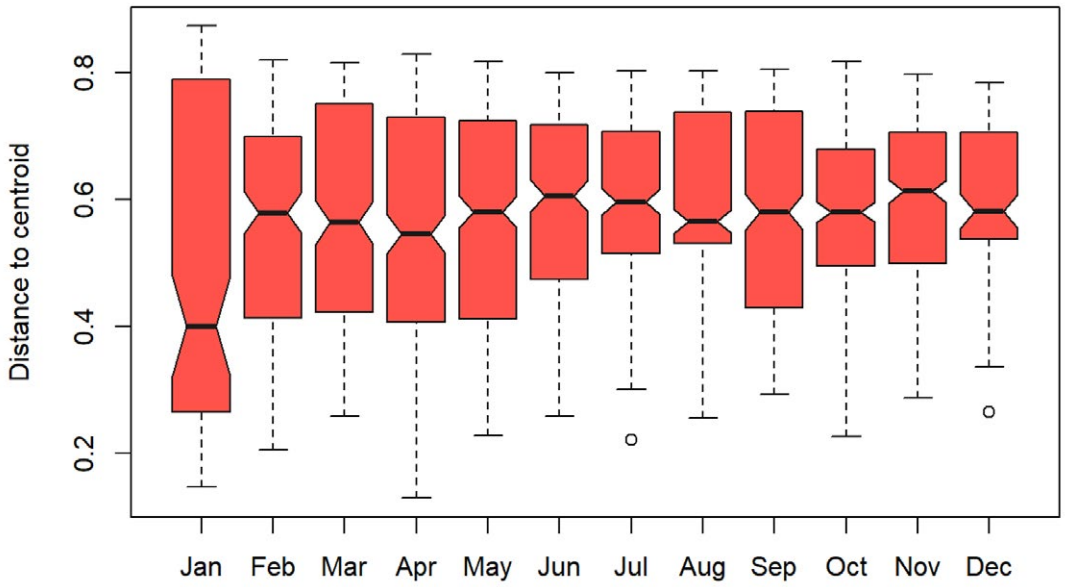


Figure 31. Stygofauna community dispersion between months

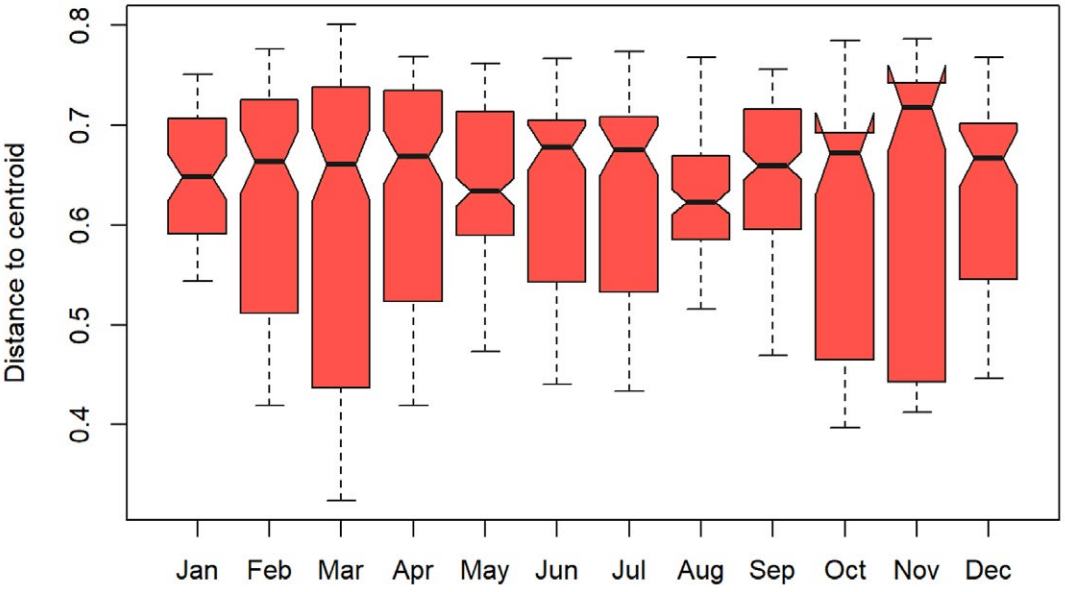


Figure 32. Troglofauna community dispersion between months

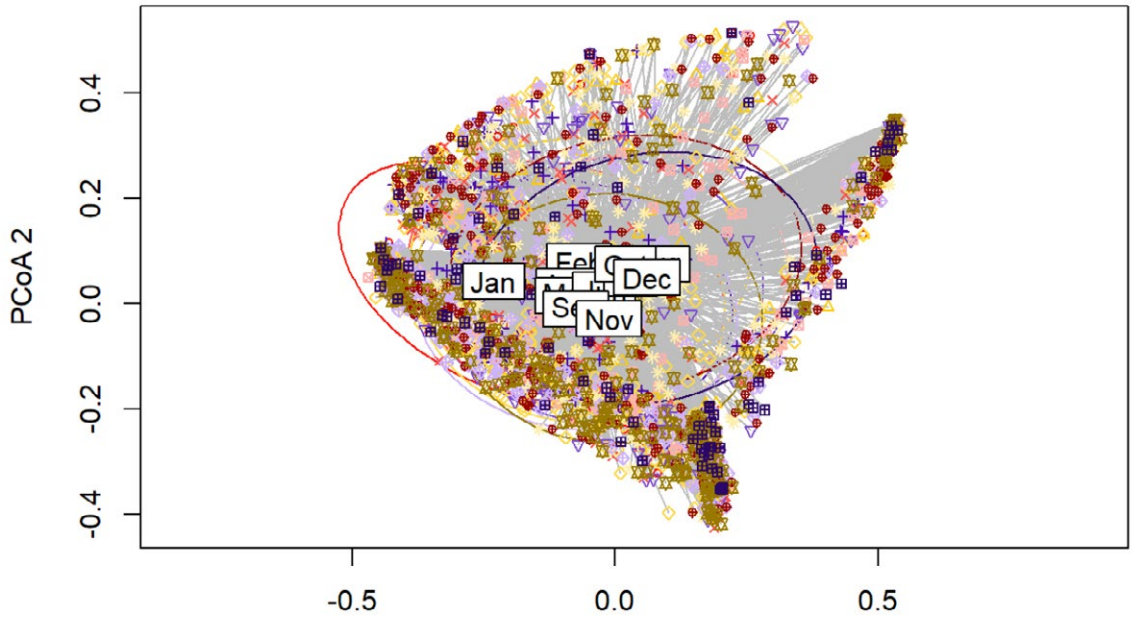


Figure 33. Stygofauna community dispersion by month

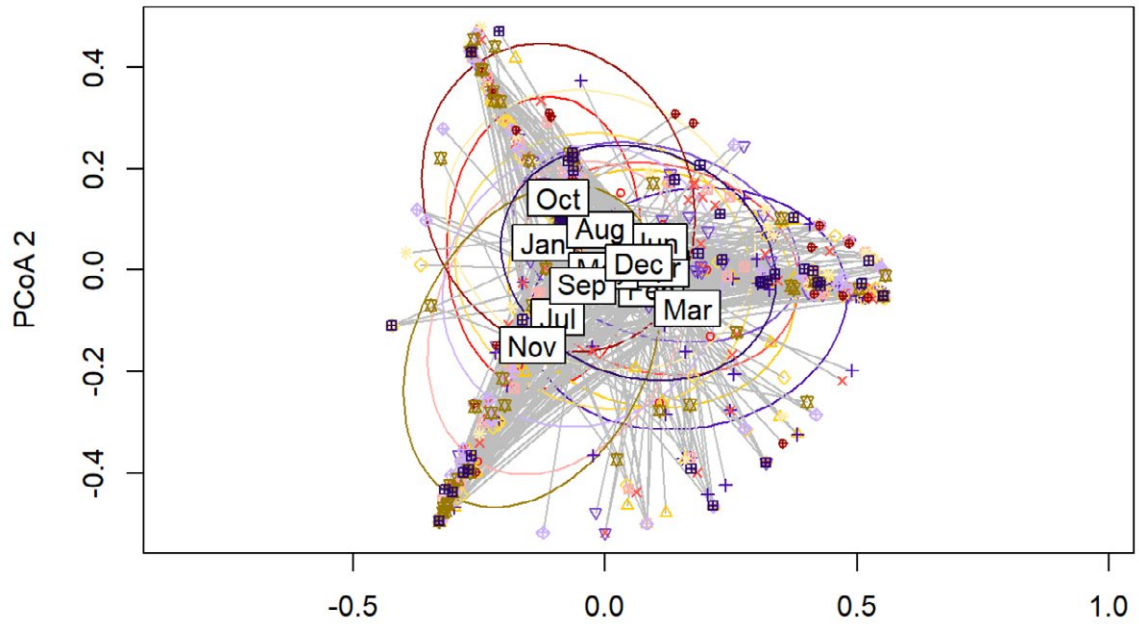


Figure 34. Troglofauna community dispersion by month

5 Discussion

5.1 Data coverage

The main aim of this study was to aggregate and explore existing historical data to better understand sampling efficiency of subterranean fauna using current approaches, and to highlight areas to improve survey protocols for environmental impact assessments. In addition, the analyses aimed to provide guidance for the next phases of the subterranean research program. The current analyses combined subterranean survey data from a total of 17,462 site (bores/holes/wells) in 10 IBRA regions in Western Australia, surveyed between 2001 and 2018. Of these, almost 11,000 were stygofauna and almost 6,500 were troglofauna collection sites, resulting in more than 50,000 collecting events and almost 28,000 troglofauna and stygofauna records.

In comparison, a study by Mokany *et al*. (2018) who aimed to model subterranean biodiversity patterns on a dataset for the Pilbara region only, was based on troglofauna data from 8,605 drill holes (sampled 2005–2015), with troglofauna recorded in 3,470 of those. Stygofauna results were based on a sample from 4,334 bores or drill holes (1997–2015) with stygofauna recorded in 2,585. In total, Mokany *et al.*'s (2018) study included 7,507 troglofauna and 11,813 stygofauna records, respectively.

The database resulting from this project is, to our knowledge, the largest subterranean fauna dataset compiled for Australia. The Queensland Subterranean Aquatic Fauna Database (https://www.data.qld.gov.au/dataset/queensland-subterranean-aquatic-fauna-database; accessed 7 December 2020) currently contains 456 fauna records from 728 visits at 602 sites.

According to the Project database, only a small fraction of specimens (8.3% of stygofauna; 23.6% of troglofauna) were submitted to the WA Museum. However, within the context of a largely undescribed fauna, verifications of species identifications require publicly accessible reference specimens. The data highlight the current poor lodgement of specimens to the WA Museum (even considering an underestimate of submissions), without which any inferences drawn from the analyses can ultimately not be verified.

5.2 Data quality

Already at the acquisition stage (i.e., when datasets of varying quality were incorporated into the SQL database), it was clear that many of the source datasets committed to this study did not fulfill the criteria for analyses focusing on sampling method and effort. This was not necessarily a surprise, as the various source databases were never designed to allow for those types of analyses, including across a wide geographic and geological scale. Even where data collection standards exist (e.g., those of the major proponents), these were generally designed based on individual project needs and standards (e.g., geological nomenclature) that are often not consistent across projects or proponents. This was compounded by the multitude of different data collection and storage systems used by consultancies that range from sophisticated GIS-based systems, custom-made MS Access® databases to simple MS Excel® or csv-spreadsheets. This confirmed the identified need of this Project to inform regulators of best practice survey and data-capture methods. Specific limitations in the source dataset in relation to the analyses were:

- **Missing data.** Source data sets did not include key variables for analyses, and these were not recoverable from reports. This included specific sampling designs (e.g., how many stygofauna hauls of what mesh diameter were conducted per sample; age of bore; haul or scrape depth etc.). Even if the database included fields for variables, these were inconsistently recorded.
- **Lack of structure in the source databases.** A data coherence issue (e.g., pooling of sample data for one bore by different sample methods) prevented populating or connecting the tables in the Project database as required and therefore excluded some datasets from the analyses.
- **Lack of clear definitions of categorical variables.** One of the main problems concerned poor definitions of many categorical variables, particularly those that may have influenced sampling success. Even if data were present, it was not categorised for analyses. For example, there are two basic types of scrapes for troglofauna surveys, one that principally utilises a little modified stygofauna haul net (Halse & Pearson 2014), and a second that adds a scraping attachment above each net that comprises numerous strands of fishing line which dislodge additional troglofauna in their reach (e.g. Subterranean Ecology 2011). An analysis differentiating between these different types of scrapes was not possible as they were not coded in the source datasets. Characterising geology was also highly variable between source datasets and therefore an analysis on geology not conducted. There was also inconsistent or no information, for example, on the mesh-diameter for stygofauna samples or the type of bait used in troglofauna traps although considerable time and effort was expended to source these data from the survey reports. There was also often no definition on what a single survey effort/sample is and the database

lacks information about if this total sampling effort was successfully conducted at each bore. There was often only a generic statement of methods in the report. Six stygofauna hauls are recommended by the EPA (2016c) and should be considered as one sample. But there is no clear definition what a single troglofauna trap sample is (more than one trap in a bore is often used), or how many scrapes constitute one troglofauna sample (assumed four). Only with a clear definition of a survey unit and a proper documentation of how much of that has been conducted at each bore, can sampling efficacy be analysed.

- **Lack of variation.** Many variables that were initially considered for the analysis only included a small range of values. For example, bore diameter, if recorded, was highly biased to 150 mm bores in the database with too few samples for any other bore diameter. Data were also highly biased towards the Pilbara bioregion.

Exploratory analyses also showed that many of the variables targeted for analyses were correlated so that a multivariate analysis on the whole dataset could not be employed. However, the data set will allow future expanded analyses of subsets, for example a regional analysis of Pilbara data, which was beyond the scope of this study.

5.3 Taxonomy and nomenclature

A large number of records in the Project database (57.2% stygofauna, 40.9% troglofauna) were identified at the species level. This is a substantial number, considering that only about 36% of the stygofauna species and 11% of the troglofauna species in the database are described. The use of para-taxonomic morphospecies codes is prevalent in the Project database, but with this come serious problems when inconsistent codes are used.

It was not part of the scope of this project to provide solutions to the 'taxonomic impediment', i.e., the lack of taxonomic research to accurately document the distribution of subterranean fauna. However, some consideration is given here to standards required to allow for consistent compilation of taxonomic data, i.e., 1) appropriate documentation of species (and higher taxa) in the taxonomic tables of a database irrespective of them being described or not, and 2) appropriate documentation of identifications of specimens of fauna records with reference to species tables (independent if identified by morphology or molecular data).

There are a number of key aspects of a stable nomenclature at the species level, i.e., 1) the designation of a type specimen as reference for each recognised species; 2) its lodgement in a public institution; 3) a diagnosis (i.e., how the species differs from others); and 4) the author of a new species (i.e., a statement of who recognised the new species and when). If a number of para-taxonomic nomenclatures are used, the author of a particular system should be stated, such as the WA Museum for a three-letter species code within taxonomic orders (e.g., *Draculoides* 'SCH014' in the Schizomida) or Bennelongia's designation of a 'B-code'.

Without adherence to these principles and their appropriate documentation in a future subterranean fauna database, any decisions made based on surveys documented in the database are not reproducible and therefore do not follow accepted scientific principles. Based on the poor WA Museum submission rate it appears that even the most important taxonomic aspect, the public lodgement of a reference specimen, is generally not adhered to. Appropriate documentation of taxonomic decisions is particularly important for subterranean fauna, as there are unique challenges, both at the morphological and molecular level (Halse 2018).

5.4 Data analyses

Despite insufficient data to analyse many factors as proposed (see Table 6 in chapter 4.2) there are some key results that inform the development of efficient, repeatable and effective survey protocols for subterranean fauna.

5.4.1 Troglofauna survey methods

Troglofauna surveys are conducted using two principal methods, trapping and scraping, although the Project database shows net hauls collect a considerable amount of troglofauna as by-catch (more than 10% of all troglofauna records). The Project database did not contain information that allowed analyses of variations in each method, for example the number of scrapes per sample, or if a scraping device was used, or in the case of troglofauna, which substrate or bait was used. In addition, there was too little variation in the data to allow for some analyses (e.g., how long should a troglofauna trap be installed?). The analyses pooled data for all scrapes, traps and net hauls to compare their efficacy.

All three methods were similar in the mean number of troglofauna taxa retrieved per sample; however, scrapes were slightly more effective at collecting unique taxa than traps, and traps collected a higher abundance of troglofauna than scrapes (or nets). Sampling method is a significant determinant of the community found for troglofauna; scrapes and nets collect similar taxa (not surprising as they are essentially the same sampling method), but traps collect a distinct assemblage.

There are few studies that compare subterranean sampling methods and survey efficiency. With an analysis of 10,895 sampling events targeting troglofauna in the Pilbara and Yilgarn, Halse and Pearson (2014) assessed the efficiency of bore scraping vs. trapping. Scraping collected more specimens than trapping (in contrast to the results here) and more than twice as many troglofauna species per sample (scrapes performed slightly better in the analyses here). Most orders of troglofauna were collected in greater numbers by scraping than trapping, although there was a collecting bias in some groups. This bias matches the result here whereby different communities were collected using different methods. This suggests that any troglofauna survey should employ both sampling techniques (i.e., scrapes and traps).

5.4.2 Stygofauna survey methods

Net hauls are the principal method to sample stygofauna in Western Australia and no other technique (e.g., pumping, interval sampling) was captured in the Project database. However, similar to troglofauna, about 10% of all stygofauna was collected as incidental by-catch when scraping (i.e., a survey method not targeting them), possibly by dipping the scraping device into the groundwater.

A stygofauna study in the calcretes of the Yilgarn investigated the effectiveness of three sampling methods for stygofauna: haul net sampling, pumping with a 12-V impeller pump, and a discrete interval sampler (Allford *et al*. 2008). More than 150 samples were taken over 16 months from 55 uncased bore holes. No significant taxonomic bias was detected across the sampling methods; however, sampling using a haul net was found to be the most efficient method for capturing the available taxa per unit time when sampling bores less than 10 m deep, with pumping being the least efficient. In contrast, a study in New South Wales found that ten net hauls alone only collected about 64% of taxa, increasing to 92.5% when combined with analysing the first 100 L of pumping (Hancock and Boulton 2009). Pumping also collected a larger number of species than net hauls in a study on sampling efficiency of stygofauna in the Pilbara, although this difference was not statistically significant (Eberhard *et al*. 2009b).

5.4.3 Species accumulation

A clear and strong positive linear relationship between the number of site visits and the number of novel taxonomic units was evident for both stygo- and troglofauna, which indicates that taxonomic richness continued to increase with each successive visit. The modelled taxon accumulation curves for troglofauna and stygofauna combined indicated that there is a tendency towards flattening of the curve at very high counts of cumulative sum of taxa found, into the hundreds of taxa (Figure 10). Within the limitations of this dataset, at no point was the entire community surveyed in most cases. Zero increase in LTU richness was achieved rarely after 15 visits in this dataset, although the increase in richness slowed with a larger number of visits. This is similar to results of previous studies. Halse and Pearson (2014) showed that yields from troglofauna traps and scrapes in both the Pilbara and Yilgarn were low, with species accumulation curves for some survey areas not plateauing even after more than 100 samples. However, preliminary survey data may inform how many samples are needed to capture a specific proportion of the fauna in a region.

Halse *et al*. (2018) analysed troglofauna survey results from 150 drill holes in the Pilbara, sampled three times each, and showed that while abundant species had a 61% probability of being collected at least twice, the species always collected as single animals had only a 6% probability of being collected twice. Many holes yielded no fauna even if repeatedly sampled. Similarly, for stygofauna, Pilbara-based sampling programs using haul nets have shown that the first sample from a bore captures 46% of all high abundance species and 23% of species found in low abundance (e.g. Eberhard *et al*. 2009a). Six samples collected over three to four years captured more than 80% of all species present at the bore holes sampled and more than 90% of the abundant species. However, no study has shown a plateauing accumulation curve in subterranean fauna surveys in WA.

The Project database is biased towards regions with a comparatively high incidence of subterranean fauna, although there are areas where there are overall very low numbers in samples (e.g. Karanovic et al. 2013). A set minimum level of sampling effort as initially recommended by the EPA (40 stygofauna and 60 troglofauna samples from impact areas) (EPA 2016c) may not provide sufficient data for an assessment, as subsequently recognised by the regulators with a less prescriptive sampling regime (EPA 2016d). A more appropriate course of action may be to adjust sampling effort to initial capture rates, thereby following a more precautionary than risk-based approach to sampling, consistent with suggestions by Eberhard et al. (2009a). On the other hand, as it is unlikely that the full complement of a subterranean fauna community is ever captured with reasonable survey effort, a trade-off between the logistical constraints of multiple sampling visits and the representation of a community is required.

5.4.4 Temporal analyses

Current guidelines recommend that subterranean fauna surveys be conducted “ideally in two different seasons” with a minimum of three-months spacing for stygofauna (EPA 2016c). It was not possible to test a two-phase survey scenario over the complete dataset, as ‘phase’ within a survey program was not defined in the Project database, and seasons differ considerably between regions over Western Australia. However, two separate

temporal analyses were conducted, i.e., simply looking at the effect on the length of the period between sampling visits and the monthly dispersion (i.e., taxon dissimilarity) of the communities throughout the year. The effects of rainfall of the nearest weather station to a sample site were also analysed.

With sites pooled, there was no overall significant difference between length of sampling interval and the dissimilarity of samples; however, where a relationship did exist at the site level, it showed a higher dissimilarity the longer the samples were spaced temporally. Without knowing prior to a sampling program if this relationship exists at a particular site, prudent survey design should temporally space collecting trips as much as possible to take advantage of this potential effect.

Survey month influenced subterranean fauna community composition, with January recovering the most diverse samples for stygofauna – the prevalent rainfall pattern in the Pilbara may contribute to this pattern – and in March for troglofauna; however, months like October and November also show high community dispersion in troglofauna. There was no clear pattern in the analysis of rainfall on the LTU richness of samples; however, the analyses were somewhat arbitrary (7-, 30- and storm days). Rainfall is likely to influence subterranean fauna through the transport of nutrients into their habitat and changes in groundwater levels (e.g. Mokany et al. 2918), but the influence of these changes on species richness is currently not clear. Sample phases should ideally be conducted in different months, with consideration given to sampling in January for stygofauna and March for troglofauna.

6 Recommendations

Recommendations are derived from all aspects of the Project, specifically learnings associated with the acquisition of the data, data quality (how data are being collected and stored, including minimal taxonomic standards), and results of the data analyses specifically in relation to sampling efficacy. The recommendations assume that a public subterranean fauna database will be established of which the Project database may provide the core element, and which will be used to collate data for future analyses.

Recommendations fall into five areas: (1) database structure; (2) standardising data (3) governance of taxonomy; (4) improving sampling; and (5) experimental sample programs.

6.1 Database structure

The Project database reflects the table structure of the initial SQL database maintained by one of the Project partners, which also includes data of surveys other than subterranean, i.e., terrestrial and aquatic invertebrates and birds. It reflects the database needs of a single company and has been designed with specific applications in mind. It was used as the primary data source in this Project as it most likely represents the most comprehensive data collection of subterranean fauna in WA. However, the database structure may not meet the objectives of a public subterranean fauna database and an expert review of the data structure is recommended.

6.2 Standardising data

The analyses greatly suffered from a poor definition of many variables and the lack of standardised data collation methods. These issues highlight the need for standardised data collection parameters and data delivery format. A standard data collection sheet would include which parameters are essential, recommended, and optional for collection; details regarding the sampling methodology (which is often described in the report but not detailed in the data spreadsheets); and a specific format that would enable automated incorporation into a large database. Similar sampling and data standards, for example, exist for the sampling of aquatic macroinvertebrate for Australian rivers through AUSRIVAS (https://ausrivas.ewater.org.au/index.php/manuals-a-datasheets; accessed 11 December 2020).

Following the establishment of a survey database, it is therefore recommended:

- Standardising the collection of data (including details on sampling methods), and a determination of mandatory, important, or optional parameters. Within a database, these could be governed by highly regulated look-up fields.
- Templates for data collection and data submission to the regulators.
- Non-acceptance of survey assessments if minimum data standards are not met (e.g., missing mandatory values).

6.3 Governance of taxonomy

The taxonomic tables of a biodiversity database are of crucial importance for correctly analysing survey data in relation to taxon richness and evenness, rarity (and therefore conservation significance), distribution ranges, habitat preferences, and sampling design. It is therefore recommended to implement standard taxonomic principles in a future database, at least:

- each morphospecies be based on a publicly available, unambiguous reference specimens ("type"), accompanied by a diagnosis (morphological and/or molecular), and detailing who recognised the new species and when.
- each taxon at the species level be documented online (including molecular data) to facilitate identification and alignment of different para-taxonomic systems.
- each para-taxonomic system be clearly defined and explained and who the custodian of the system is; ideally only one system to be used per taxonomic group throughout the state, created by a subject matter expert, and if available, the WA Museum morphospecies code to be used.
- each identification of a specimen be based on the availability of a taxon in the database and accompanied by information on who identified a specimen and when and what morphospecies reference system was used.
- funding be provided to continuously update the taxonomic part of the database (e.g., add new species using criteria above or replace morpho-codes with available species names once species are described), and maintain taxonomic consistency between the database and WA Museum data.

- each specimen be unambiguously identified by a specimen code, either designated by the survey consultant, the (para-)taxonomist, or a subterranean fauna database code. This code should be lodged with the specimen to the WA Museum to allow cross-reference to survey data.

6.4 Improving sampling

There are a few key recommendations when current sampling methods and regimes are concerned:

- The use of both traps and scrapes for troglofauna surveys will maximise taxon richness and community representation; also record stygofauna by-catch.
- Sampling troglofauna at varying depths using traps does not appear to influence community composition.
- Sample in different months if possible and as far apart as in time as possible.
- Sample as many times as practicable (since a species accumulation plateau is never reached), observe the rate of novel taxa over previously collected taxa to indicate whether a minimum target community has been documented.

6.5 Experimental survey design

The analyses showed that despite the collation of a large dataset, the Project database suffered from a substantial amount of missing data, inconsistent data fields, and many variables that could not be used in the analysis due to biases and intercorrelations. Whilst standardised data collection will help to mitigate some of these problems, appropriately designed experimental studies are recommended to specifically address some of the key questions not able to be addressed here, such as the difference between troglofauna scrapes with and without scraping attachment, stratification of subterranean fauna or the effect of bore age.

7 References

- Allford, A., Cooper, S. J. B., Humphreys, W. F. & Austin, A. D. 2008. Diversity and distribution of groundwater fauna in a calcrete aquifer, does sampling method influence the story? *Invertebrate Systematics* **22**: 127–138.
- Anderson, M. J. 2006. Distance-based tests for homogeneity of multivariate dispersions. *Biometrics* **62**: 245–253.
- Anderson, M. J., Ellingsen, K. E. & McArdle, B. H. 2006. Multivariate dispersion as a measure of beta diversity. *Ecology Letters* **9**: 683–693.
- Eberhard, S. M., Halse, S. A., Williams, M. R., Scanlon, M. D., Cocking, J. & Barron, H. H. 2009a. Exploring the relationship between sampling efficiency and short-range endemism for groundwater fauna in the Pilbara region. *Freshwater Biology* **54**: 885–901.
- Eberhard, S. M., Halse, S. A., Williams, M. R., Scanlon, M. D., Cocking, J. & Barron, H. J. 2009b. Exploring the relationship between sampling efficiency and short-range endemism for groundwater fauna in the Pilbara region, Western Australia. *Freshwater Biology* **54**: 885–901.
- EPA. 2016a. *Environmental Factor Guideline: Subterranean fauna*. Environmental Protection Authority, Perth, WA. Available at: http://www.epa.wa.gov.au/sites/default/files/Policies_and_Guidance/Guideline-Subterranean-Fauna-131216_3.pdf (accessed 20 December 2016).
- EPA. 2016b. *Report and recommendations of the Environmental Protection Authority. Yeelirrie Uranium Project - Cameco Australia Pty Ltd*. Environmental Protection Authority, Perth, WA. Available at: https://www.epa.wa.gov.au/sites/default/files/EPA_Report/Rep%201574%20Yeelirrie%20PER%20030816.pdf (accessed 11 December 2020).
- EPA. 2016c. *Technical Guidance: Sampling methods for subterranean fauna*. Environmental Protection Authority, Perth, WA. Available at: http://www.epa.wa.gov.au/sites/default/files/Policies_and_Guidance/Tech%20guidance-%20Sampling-Subt-fauna-Dec-2016.pdf (accessed 20 December 2016).
- EPA. 2016d. *Technical Guidance: Subterranean fauna survey*. Environmental Protection Authority, Perth, WA. Available at: http://www.epa.wa.gov.au/sites/default/files/Policies_and_Guidance/Technical%20Guidance-Subterranean%20fauna-Dec2016.pdf (accessed 20 December 2016).
- Firke, S. 2020. *Simple Tools for Examining and Cleaning Dirty Data. R package version 2.0.1*. Available at: https://CRAN.R-project.org/package=janitor
- Gibson, L. 2018. *Shedding new light on the cryptic world of subterranean fauna. A research program for Western Australia*. Western Australian Biodiversity Science Institute, Perth, WA.
- Grolemund, G. & Wickham, H. 2011. Dates and imes made easy with lubridate. *Journal of Statistical Software* **40**: 10.18637/jss.v040.i03.
- Halse, S. A. 2018. Subterranean fauna of the arid zone. In: Lambers, H. (ed.) *On the ecology of Australia's arid zone*. Springer International Publishing AG, pp. 215–241.
- Halse, S. A., Curran, M., Carroll, T. & Barnett, B. 2018. What does sampling tell us about the ecology of troglofauna? *ARPHA Conference Abstracts* **1**: doi: 10.3897/aca.1.e29829.
- Halse, S. A. & Pearson, G. B. 2014. Troglofauna in the vadose zone: comparison of scraping and trapping results and sampling adequacy. *Subterranean Biology* **13**: 17–34.
- Hancock, P. J. & Boulton, A. 2009. Sampling groundwater fauna: efficiency of rapid assessment methods tested in bores in eastern Australia. *Freshwater Biology* **54**: 902–917.
- Jones, F. C. 2008. Taxonomic sufficiency: the influence of taxonomic resolution on freshwater bioassessment using benthic macroinvertebrates. *Environmental Reviews* **16**: 45–69.
- Karanovic, T., Eberhard, S. M., Perina, G. & Callan, S. 2013. Two new subterranean ameirids (Crustacea: Copepoda: Harpacticoida) expose weaknesses in the conservation of short-range endemics threatened by mining developments in Western Australia. *Invertebrate Systematics* **27**: 540-566.

- Kassambara, A. 2020. *ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0*. Available at: https://CRAN.R-project.org/package=ggpubr
- Mokany, K., Harwood, T. D., Halse, S. A. & Ferrier, S. 2018. Riddles in the dark: Assessing diversity patterns for cryptic subterranean fauna of the Pilbara. *Diversity and Distributions* **25**: 240–254 https://doi.org/10.1111/ddi.12852
- Mokany, K., Harwood, T. D., Halse, S. A. & Ferrier, S. 2918. Riddles in the dark: assessing diversity patterns for cryptic subterranean fauna of the Pilbara. *Diversity and Distributions* **25**: 240–254.
- Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E. & Wagner, H. 2020. *vegan: Community Ecology Package*. Available at: https://cran.r-project.org/web/packages/vegan/index.html
- R Core Team. 2018. *A language and environment for statistical computing*. R Foundation for Statistical Computing. Available at: https://www.R-project.org
- Robinson, D., Hayes, A. & Couch, S. 2002. broom: *Convert Statistical Objects into Tidy Tibbles. R package version 0.7.0*. Available at: https://CRAN.R-project.org/package=broom
- Subterranean Ecology. 2011. *Fortescue Metals Group. Solomon Project: Regional Subterranean Fauna Survey*. Subterranean Ecology Pty Ltd, Stirling, WA. Unpublished report prepared for Fortescue Metals Group. Available at: https://epa.wa.gov.au/sites/default/files/Proponent_response_to_submissions/1386%20-%205%20Appendix%20A%20-%20Regional%20Subterranean%20Fauna%20Survey%20-%20Final%20Report.pdf
- Wickham, H., Averick, M., Bryan, J., Chang, W., D'Agostino McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughn, D., Wilke, C., Woo, K. & Yutani, H. 2019. Welcome to the Tidyverse. *Journal of Open Source Software* **4**: https://www.R-project.org/
- Xie, Y. 2020. *A General-Purpose Package for Dynamic Report Generation in R*. Available at: https://cran.r-project.org/web/packages/knitr/
- Zhu, H. 2020. *Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.2.1*. Available at: https://CRAN.R-project.org/package=kableExtra

Appendix 1
Metadata of Project database

File	Variable	Type	Description
bore	access_bore	factor	description of the record access restrictions at the bore level
	aquifer_name	character	labels the aquifer in which the bore is located
	aquifer_type	character	labels the type of aquifer in which the bore is located
	bore_comment	character	description of the bore, free entry field
	bore_details_id	integer	identifies unique bore
	bore_use	character	initial purpose of bore
	bore_diameter	integer	diameter of bore opening
	collar_type	factor	material used at bore opening
	casing_type	factor	material used along bore walls
	cover	logical	presence of covering over bore
	locked	integer	presence of lock or lock code
	number_of_ adjacent_bores	integer	field count of nearby bores
	infrastructure	character	description of nearby infrastructure
	surface_geology_ comment	character	description of the visible surface geology, free entry field
	surface_geology_ general	character	field description of visible geology
	site_id	integer	identifies unique site
	bore_angle	double	angle of bore from surface
	angle_direction	double	cardinal direction of bore angle
	bore_comments	character	free description field
bore_ geology	access_geology	factor	description of the record access restrictions at the geology level
	bore_geology_id	integer	identifies unique geology
	screen_interval	integer	unable to determine
	depth_bg_ ltotopofscreen	double	unable to determine
	depth_bg_ ltobottomofscreen	double	unable to determine
	depth_as_ ltotopofscreen	double	unable to determine
	depth_as_ ltobottomofscreen	double	unable to determine
	general_geology	character	categorical descriptor of geology
	formation_name	character	name of geological formation
	geology_source_ notes	character	citation of geology
	site_id	integer	unique site identifier
	bore_details_id	integer	identifies unique bore

File	Variable	Type	Description
environ_ metrics	access_metrics	factor	description of the record access restrictions at the metric level
	sample_id	integer	identifies unique sample
	conductivity_ms_ cm	double	bore water conductivity in mS/cm
	p_h	double	bore water pH
	temperature_c	double	bore water temperature in C
	oxygen_mg_l	double	bore water dissolved oxygen in mg/L
	oxygen_percent	double	bore water oxygen saturation
	salinity_mg_l	double	bore water salinity in mg/L
	site_id	integer	unique site identifier
organisms	access_organism	factor	description of the record access restrictions as the organism level
	taxon_sample_id	integer	unique organism within sample identifier
	sample_id	integer	unique sample identifier
	lowest_idnc	character	unique taxonomic classification identifier
	true_troglofauna	logical	T/F designates organism as troglofauna
	true_stygofauna	logical	T/F designates organism as stygofauna
	true_sre	logical	T/F designates organism as short-range endemic
	true_burrow	logical	T/F designates organism as burrowing
	number_identified	integer	raw count of indicated organism in sample
	organism_ comment	character	free description of organism specimen
	wam_lodged	logical	T/F organism sent to WA Museum
	wam_number	integer	unique organism identifier used by WA Museum
	notat_wam	logical	T/F organism has taxonomic registration with WA Museum
	wam_lodge_date	date	date that the specimen was lodged at the WA Museum
	institution_regno	character	taxonomic registration at WA Museum
	taxon_status	character	validity of taxonomic identification in the literature
	sex	character	sex (or sex related information) about the organism sample
	life_stage	character	life stage of organism identified
	reclassified	character	historic details of organism identification
	restriction_ organism	character	known regulatory restrictions around organism
	type_desc	factor	identifies the broad environmental type of the organism
	voucher_specimen	character	specimen identifier at registering institution
	sub_site_id	integer	identifies unique location in bore
	site_visit_id	integer	unique site visit identifier
	site_id	integer	unique site identifier

File	Variable	Type	Description
data_ source	data_source_id	integer	unique data source identifier
	data_source	character	name of the company or project that provided the data
	duration	character	duration of the projects provided by the data source
	objectives	character	objectives of the data source
	funding_sources	character	sources of funding for the data source
	geographic_data_ statement	character	coordinate system and projection used
	data_source_ parameters	character	general description of the data source
	geographic_extent	character	geographic extent of the data source
	total_sites	integer	the number of sites that are connected to the data source
	total_sites partially_restricted	integer	the number of sites with partially restricted data
	total_sites_public	integer	the number of sites with no data restrictions
	total_sites_ restricted	integer	the number of sites which are completely restricted
	permission	character	restrictions and permissions on use of the data
sample	access_sample	factor	description of the record access restrictions at the sample level
	depth_to_bottom	double	measured depth in metres to end of bore at sampling
	depth_to_water	double	measured depth in metres to groundwater at sampling
	collected_by1	factor	anonymised sample collector ID
	collected_by2	factor	anonymised sample collector ID, if more than one collector
	consultant_name	character	the consultant who collected the sample
	general_ description	character	free entry description of the sample quality or collection
	invertebrates_ sampled	logical	true if the sample collected invertebrates
	land_tenure	character	description of land use at sample collection
	sample_id	integer	unique sample identifier
	sample_date	date	date that the sample was processed
	sample_collected	date	date that the sample was collected
	sample_notes	character	free description of sample collected
	sample_type_id	integer	unique identifier of the sample type
	site_visit_comment	integer	description of the site or sample during the visit
	sorted_by	factor	anonymised sample sorter ID
	sre_sampled	logical	true if the sample intended to collect a short-range endemic taxon
	stygofauna_ sampled	logical	true if the sample intended to collect stygofauna
	troglofauna_ sampled	logical	true if the sample intended to collect troglofauna
	stygo_sample	logical	true if the sample intended or resulted in the collection of stygofauna

File	Variable	Type	Description
	troglo_sample	logical	true if the sample intended or resulted in the collection of troglofauna
	sub_site_id	integer	identifies unique location in the bore
	sub_site_code	factor	sample collection type within the bore or at the site
	trap_depth	integer	depth (for trap only) at which sample was collected
	collected_by1	integer	personnel identifier 1
	collected_by2	integer	personnel identifier 2
	sorted_by	integer	personnel identifier of sample sorter
	sample_type_ name	factor	categorical description of sample collection method
	site_visit_id	integer	unique site visit identifier
	site_id	integer	unique site identifier
	visit_date	date	date that the site was visited for sample collection
site	access_site	factor	description of the record access restrictions at the site level
	altitude_device	factor	measurement tool used to collect altitude at the site
	altitude	double	measured altitude at the site in metres
	brief_location	character	description of the location of the site
	field_bore_codes	character	bore identifier from the source report
	site_id	integer	unique site identifier
	site_code	character	site identifier from source report
	site_name	character	plain text common name of site
	ibra_code	character	biogeographic region code from IBRA7
	ore_body	factor	common name of the ore body at the site
	station_id	integer	bureau of meteorology rainfall station ID for the site closest to the nearest sample collection site
	station_name	character	bureau of meteorology name of the rainfall measurement site nearest to the sample collection site
	data_source_id	integer	unique project identifier
	site_type_desc	integer	describes the context of the site in regards to biodiversity sampling
	accuracy_ confidence	integer	indicator of location accuracy
	subterranean_site	logical	true if the site contained records of stygofauna or troglofauna
	site_lat	double	decimal latitude at site
	site_long	double	decimal longitude at site
	site_comment	character	free descriptor of site characteristics
	accuracy_site	integer	unable to determine
	locality_name	character	common name of the location or group of sites
	locality_lat	double	decimal latitude of centroid of location of group of sites
	locality_long	double	decimal longitude of centroid of location of group of sites
	stygo_site	logical	T/F indicator if stygofauna were searched for or found at site
	troglo_site	logical	T/F indicator if troglofauna were searched for or found at site

File	Variable	Type	Description
taxonomy	access_taxon	factor	description of the record access restrictions at the taxonomy level
	lowest_name	character	scientific name of lowest identification of the organism
	level_id	character	lowest taxonomic level at which the record was identified
	lowest_idnc	character	unique identifier for lowest level taxonomic classification
	family_code	character	unique identifier for the family identification
	kingdom	character	taxonomic kingdom
	phylum	character	taxonomic phylum
	subphylum	character	taxonomic subphylum
	class	character	taxonomic class
	subclass	character	taxonomic subclass
	infraorder	character	taxonomic infraorder
	order	character	taxonomic order
	suborder	character	taxonomic suborder
	superfamily	character	taxonomic superfamily
	family	character	taxonomic family
	subfamily tribe	character	taxonomic subfamily tribe
	subfamily	character	taxonomic subfamily
	genus	character	taxonomic genus
	species	character	taxonomic species
	tribe	character	taxonomic tribe
	type_desc	factor	ecological group of the record
	authority	character	taxonomic authority for nomenclature
	registered_species	character	regulatory restrictions on the species
	restriction_notes	character	free entry description of restrictions on the taxon
	taxonomic_notes	character	free description of taxonomic identification
	restriction_taxon	character	known regulatory restrictions around taxon

www.wabsi.org.au

Contact: subtfauna@wabsi.org.au

Proudly supported by:

CSIRO

Curtin University

ECU Australia Edith Cowan University

Murdoch University

The University of Western Australia

Western Australian Museum

Department of Biodiversity, Conservation and Attractions

Department of Energy, Mines, Industry Regulation and Safety

Department of Primary Industries and Regional Development

Department of Jobs, Tourism, Science and Innovation

Department of Water and Environmental Regulation