InvAlloR-manuscript

Table of contents

**Main text**

**Title:** InvAlloR: A database and R-package for automating biomass estimations from allometric equations for freshwater invertebrates using taxonomic backbones and geographic/habitat similarity

**Authors:** James G. Hagan, Ronald Bergmann, Vincent Dolmans, Joren Snoeks, Bram Vanschoenwinkel

1. Department of Marine Sciences, University of Gothenburg, Box 461, SE-40530, Gothenburg, Sweden
2. Gothenburg Global Biodiversity Centre, Box 461, SE-40530, Gothenburg, Sweden
3. Community Ecology Lab, Department of Biology, Vrije Universiteit Brussel (VUB), Pleinlaan 2, 1050 Brussels, Belgium.
4. Centre for Environmental Management, University of the Free State, Bloemfontein, South Africa.

**Author contributions:** JH conceived the idea with input from JS and BV. JH, VD and JS compiled the database. JH and RB wrote the code for the R-package. JH analysed the data. JH wrote the first draft, and all authors revised the manuscript.

**Data availability statement:** All code and data to reproduce the analysis in this article can be found in a Github repository: https://github.com/haganjam/InvAlloR-journal-article.

**Words:** **Figures:** **Tables:** **References:**

### 0.1 Abstract

Biomass data are increasingly used to gain better insights into the functioning of freshwater ecosystems. However, because weighing organisms is notoriously labour intensive, most workers rely on allometric equations to combine available information on body dimensions and species abundance into proxies of biomass. This approach is prone to errors due to taxonomic or geographic generalisation and application beyond the body size range used to make the equation. Here, we compiled a database of more than 350 length-biomass equations for freshwater invertebrates and developed an R-package (InvAlloR) that automatically selects the most suitable length-biomass equation for a range of freshwater invertebrates based on taxonomic similarity and geographic/habitat similarity.

### 0.2 Introduction

Estimating individual-, population- and community-level biomass has become standard practice, both in terrestrial and aquatic ecology. In freshwater ecology in particular, many commonly investigated topics rely on biomass information. These include the value of invertebrates as food sources for higher trophic levels (Deboelpaep et al. 2020), estimating ecosystem-level secondary production patterns (Patrick et al. 2019) and obtaining more representative measures of abundance than individual counts (Méthot et al. 2012).

For freshwater invertebrates, biomass is typically measured by weighing several individuals of a taxon and multiplying the average biomass by the number of individuals of that taxon (O’Gorman et al. 2017). However, accurately measuring biomass is challenging. Depending on the preservation method, the biomass of preserved samples can change over time (Howmiller 1972; Leuven, Brock, and Druten 1985; Giguère et al. 1989). Moreover, weighing specimens is cumbersome, time-consuming (Rogers, Buschbom, and Watson 1977) and requires high-precision balances which are often not available under field conditions. Finally, many freshwater invertebrates are so small that individual specimens cannot be weighed using standard high-precision balances. Instead, specimens are typically pooled until their combined weight allows for measurements to be taken. Thus, using taxon-specific equations relating some linear body size dimension (usually body length) to biomass has become the favoured approach (Gsell et al. 2016; Hébert et al. 2021).

Length-biomass equations are useful because they allow researchers to accurately convert length to biomass without directly weighing specimens. However, despite widespread use and several compilations (Meyer 1989; Benke et al. 1999; Towers, and Veltman 1994), there is no database of length-dry biomass equations for freshwater invertebrates. Moreover, the existing compilations are published as tables in papers and not as user-friendly databases. Therefore, researchers still typically compile their own set of length-dry biomass equations for their study and set of species (O’Gorman et al. 2017; Hébert et al. 2021). A centralised database that can easily be searched and used would save researchers considerable time and improve reproducibility in many analyses requiring biomass estimations.

Using length-biomass equations for freshwater invertebrates also comes with unique challenges. Although there are many published length-biomass equations, it is rare to find species-specific or genus-specific equations for all target species (O’Gorman et al. 2017; Hébert et al. 2021). Moreover, invertebrate taxonomy is often uncertain and there are many cases where taxa are not identified to species but to a lower taxonomic resolution. Thus, in addition to finding relevant equations, several decisions need to be made about how to choose an equation when a taxon-specific equation is unavailable, or the target taxon is poorly identified. These decisions may differ between researchers which can lead to results that are not reproducible or comparable.

The appropriateness of an equation for a given taxon is not only based on taxonomic relatedness. Other factors such as geographic proximity and habitat similarity of the taxon for which the equation was created are also important (Benke et al. 1999; Méthot et al. 2012). Indeed, using equations for the same taxon from a different ecoregion or reference site can lead to over- or under-estimations of dry biomass (Mehler, Acharya, and Sada 2015; Méthot et al. 2012; Ruiz-Lupión, Gómez, and Moya-Laraño 2020). However, incorporating geographic proximity and/or habitat similarity when choosing equations is rarely done when researchers select length-dry biomass equations. Most researchers simply rely on existing compilations which are biased towards North America and Europe (Meyer 1989; Benke et al. 1999; Towers, and Veltman 1994).

Here, we introduce a centralised database of >350 length-biomass equations for freshwater invertebrates worldwide incorporated in (InvAlloR)[https://github.com/haganjam/InvAlloR], an R-package that can find appropriate body length-dry biomass equations based on taxonomic relatedness and a combination of geographic proximity and habitat similarity. Taxonomic relatedness is derived from the Global Biodiversity Information Facility (GBIF), Integrated Taxonomic Information System (ITIS) and Catalogue of life (COL) taxonomic backbones. Geographic proximity and habitat similarity is based on geographic location and global freshwater ecoregions (Abell et al. 2008) which classify the globe into biogeographic regions, major habitat types and ecoregions. InvAlloR thus allows users to provide a table with taxon names, life-stage information, geographical coordinates and body length data and obtain dry biomass data generated from an equation developed for a taxonomically similar taxon from a geographically similar region.

### 0.3 Database overview

#### 0.3.1 *Database structure*

There are three components to the database. The first is the equation database which stores the equations. In the equation database, there are three pieces of data associated with each equation: (i) Each equation has a unique ID number (ID, [Figure 1](#fig-1)). (ii) Each equation has a taxon name (*Taxon*, [Figure 1](#fig-1)) for which the equation was developed. This taxon name can be any taxonomic rank below or including order. (iii) Each taxon name has associated life-stages (Life-stage, [Figure 1](#fig-1)). We provide a classification of accepted life-stages for broad groups of freshwater invertebrates ([Table 1](#tbl-1)). We refer to the taxon names in the equation database as equation taxa hereafter. Note that we treat different life stages of the same taxon as different equation taxa.

In the equation database, we only allow equations that predict dry biomass. This is because wet versus dry mass conversion factors are not available for many taxa in the database. Dry biomass is also less sensitive than wet biomass to water content which, in turn, reflects environmental conditions at the time of sampling. It is therefore better suited to comparisons among species. Moreover, we also only allow equations that predict dry biomass from a single linear body size dimension as most equations utilise one dimension. The exact linear body size dimension differs depending on the broad group of freshwater invertebrates. For most taxa, this is body length (BL) ([Table 1](#tbl-1)). But, for some taxa like molluscs, this is either shell width (SW) or shell height (SH) ([Table 1](#tbl-1)). For examples of the linear body size dimension for different groups of freshwater invertebrates used in the equation database, see Figure S1.

The second component of the database is the higher-level taxonomic graphs based on the GBIF, ITIS and COL taxonomic backbones which depict the taxonomic relationships among taxa as nodes (i.e. a taxon name) and edges (i.e. the connection between, for example, a family and a genus) ([Figure 1](#fig-1)). For example, a higher taxonomic graph for a family contains all genera and species within those genera that fall under that family name. We build these higher-level taxonomic graphs by harmonising all the equation taxa to the GBIF, ITIS and COL taxonomic backbones using the Biodiversity Data Cleaning (bdc) package in R. We cleaned the equation taxa names using the bdc\_clean\_names() function and then harmonised the names using the bdc\_query\_names\_taxadb() function. The bdc package uses the taxadb package for harmonisation to the different taxonomic backbones. Along with name harmonisation, the bdc\_query\_names\_taxadb() function extracts higher taxonomic names for each equation taxon. These higher taxon names are generally families or orders depending on the taxonomic backbone.

For each of the three taxonomic backbones, the unique higher taxon names (order or family) from all the equation taxa are extracted as some equation taxa in the equation database can be associated with the same higher taxon (e.g. two species from the same family). Next, the filter\_rank() function from the taxadb package is used to get all descendant names from each unique higher taxon name. These descendant names are then processed into an igraph object (igraph package in R) where the taxonomic distance between names from different ranks is set at one (e.g. distance between family and order is one, distance between family and genus is one). The one exception is that the distance between genus and species is set by the user (default = 0.5) to reflect the general rule that species within a genus tend to be closely related. These igraph objects preserve the relatedness implied by the taxonomic backbone such that two families from the same order are separated by a taxonomic distance of two (i.e. path from family 1 to family 2 must go through the common order), ([Figure 1](#fig-1)). Thus, for each equation taxon, we generate higher-level taxonomic graphs from each of the three taxonomic backbones (GBIF, ITIS and COL) as some taxa are only present in certain taxonomic backbones. Using these higher-level taxonomic graphs from the different taxonomic backbones places each equation within a set of closely related taxa names ([Figure 1](#fig-1)).

|  |
| --- |
| Figure 1: The three components of the InvAlloR database. **(a)** Each equation in the equation database has a unique ID number, a taxon name for which the equation was developed and life-stage information. **(b)** Each taxon in the equation database is placed in a higher-level taxonomic graph which describes how different taxa are taxonomically related based on three taxonomic backbones: GBIF, ITIS and COL. For example, taxon names present in the equation database are highlighted in red in the graphical representations of higher-level taxonomic graphs. **(c)** Each unique equation ID number is also associated with a latitude-longitude coordinates on a freshwater ecoregion map (Abell et al. 2008). Given a target taxon name and latitude-longitude coordinates, equations from closely related taxa and geographical locations can be identified in the database. |

The third component of the database is the latitude-longitude coordinates associated with each equation ID. These coordinates reflect the geographical area from which the specimens were collected to produce the length-dry biomass equation. We couple these coordinates for each equation ID to a global map of freshwater biogeographical realms, major habitat types and ecoregions (Abell et al. (2008)), (Figure 1c). Length-dry biomass relationships in freshwater invertebrates can vary depending on habitat conditions and biogeographical realm (Mehler, Acharya, and Sada 2015; Méthot et al. 2012; Ruiz-Lupión, Gómez, and Moya-Laraño 2020). This information on the biogeographical realm, major habitat type and ecoregion allows users to select equations for a target taxon (i.e. a name for which dry biomass data is required) with similar geographic locations and habitat conditions if the user provides latitude-longitude information for the target taxon.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1: Classification used in the InvAlloR database to describe the major groups of freshwater invertebrates globally. The groups 1 and 2 are not strict taxonomic entities. Also included is the life-stage classification and the body-size measurement used in the equations (BL - body length, SH - shell height and SW - shell width). The ‘Example taxon’ column gives an example of a taxa present in the equation database corresponding to that group (if NA then the database does not contain equations for that group).   | Group1 | Group2 | Life\_stage | Body\_size | Order | Example\_taxon | | --- | --- | --- | --- | --- | --- | | Insecta | Hemimetabolous | nymph, adult | BL | Ephemeroptera | Coloburiscus humeralis | |  |  |  |  | Plecoptera | Zelandobius | |  |  |  |  | Megaloptera | Archichauliodes diversus | |  |  |  |  | Trichoptera | Olinga feredayi | |  |  |  |  | Odonata | Erythrodiplax justiniana | |  |  |  |  | Hemiptera | Notonecta | |  | Holometabolous | larva, pupa, adult | BL | Diptera | Austrosimulium | |  |  |  |  | Lepidoptera | Pyralidae | |  |  |  |  | Coleoptera | Tropisternus | | Crustacea | Copepoda | nauplius, copepodite, adult | BL | Calanoida | Argyrodiaptomus azevedoi | |  |  |  |  | Cyclopoida | Thermocyclops decipiens | |  |  |  |  | Harpacticoida | NA | |  | Malacostraca | adult | BL | Amphipoda | Hyalella azteca | |  |  |  |  | Isopoda | Asellidae | |  |  |  |  | Decapoda | Aegla neuquensis | |  | Branchiopoda | adult | BL | Diplostraca | Daphnia gessneri | |  |  |  |  | Anostraca | NA | |  |  |  |  | Notostraca | Triops cancriformis | |  |  |  |  | Clam shrimps | NA | |  | Ostracoda | adult | BL |  | NA | | Arachnida |  | adult | BL |  | NA | | Rotifera |  | adult | BL |  | Asplanchna | | Tardigrada |  | adult | BL |  | NA | | Collembola |  | adult | BL |  | Hypogastrura manubrialis | | Nematoda |  | none | BL |  | NA | | Platyhelminthes | Turbellaria | none | BL |  | Girardia | | Annelida | Oligochaeta | none | BL |  | Limnodrilus | |  | Hirudinea | none | BL |  | NA | | Mollusca | Gastropoda | none | SH |  | Tarebia granifera | |  | Bivalvia | none | SW |  | Ptychobranchus fasciolaris | |

Classification used in the InvAlloR database to describe the major groups of freshwater invertebrates found in freshwater habitats globally. ‘Group 1’ and ‘Group 2’ are not strict taxonomic entities. Also included is the life-stage classification and the body-size measurement used in the equations: BL - body length, SH - shell height and SW - shell width. The ‘Example taxon’ column gives an example of a taxa present in the equation database corresponding to that group (if NA then the database does not contain equations for that group).

The datafiles that comprise the database and the code used to generate the datafiles are available in a Github repository: https://github.com/haganjam/InvAlloR. The database will be updated periodically as more equations become available. Upon acceptance, we will archive a version of the database on Figshare (https://figshare.com/).

#### 0.3.2 *Taxonomic coverage*

The current version of InvAlloR contains 366 individual body-length dry biomass equations from 17 published studies. These 366 equations cover 296 unique taxa names which come from 116 different families and 32 different orders. Most taxa are *Insecta* (144), but *Mollusca* (44) and *Crustacea* (20) are also well-represented ([Figure 2](#fig-2)). In contrast, *Annelida* and *Platyhelminthes* are poorly represented (four taxa, [Figure 2](#fig-2)) and, currently, the database is missing representatives of *Rotifera*, *Tardigrada*, *Nematoda* and *Arachnida*.

|  |
| --- |
| Figure 2: The number of families (bar-height) and the number of taxa (numbers above bars) represented in InvAlloR by order (vertical x-axis label) and major group (horizontal x-axis labels and different colours). Reported numbers are averaged across GBIF, ITIS and COL taxonomic backbones as they differ slightly in their taxonomic assignments. Annel. - Annelida, Platy. - Platyhelminthes. |

#### 0.3.3 *Geographic coverage*

The equations in InvAlloR come from 23 different geographical locations (@[Figure 3](#fig-3)) which cover all continents barring Africa and Antarctica. However, more than 40% of all equations come from the Nearctic and, specifically, North America (@[Figure 3](#fig-3)). Despite coverage of most continents, equations from three of the seven freshwater biogeographical realms (Abell et al. 2008) are missing: Afrotropics, Indo-Malay and Oceania (@[Figure 3](#fig-3)). For the other four biogeographical realms, equations for more than 35% of the 32 orders in the database (@[Figure 3](#fig-3)) are represented.

|  |
| --- |
| Figure 3: **(a)** Locations from which the length-biomass equations in InvAlloR were derived. The size of the points indicates the number of unique equations from that location. Colours represent different biogeographic realms corresponding to **(b)** which shows the proportion of orders (see [Figure 2](#fig-2)) in InvAlloR for the seven recognised biogeographic realms for freshwater taxa (Abell et al. 2008). |

### 0.4 InvAlloR: Using the database

To use the database to obtain dry biomass data for a given taxon name, we have written an R-package (InvAlloR, https://github.com/haganjam/InvAlloR) directly downloadable from Github through R (soon available on CRAN). This package loads the database into the R-environment from the online repository.

There are two workflows that the R-package provides and both are conducted using the function: get\_trait\_from\_taxon(). The first workflow (argument: workflow = "workflow1") takes a vector of target taxon name(s) (i.e. names for which dry biomass data are required), cleans and harmonises the names to the GBIF, ITIS and COL taxonomic backbones using the bdc\_clean\_names() function and the bdc\_query\_names\_taxadb() function from the bdc package in R. It then searches the database for any equation that was developed for a taxon within some predefined maximum taxonomic distance from the target taxon name(s). For example, if *Brachionus calyciflorus* and a maximum taxonomic distance of three are provided, the function provides a data.frame of equations developed for taxa that are within three taxonomic units of *B. calyciflorus* based on the higher-level taxonomic graphs from GBIF, ITIS and COL that include *B. calyciflorus*. Practically, this means that equations developed for taxa in similar genera or families are outputted along with metadata related to life-stage (see [Table 1](#tbl-1)), geographic location, habitat type (biogeographical realm, major habitat type and ecoregion from (Abell et al. 2008) and various other information. Users can then inspect the data and choose the equations that they deem most relevant. In addition, links to the original sources are provided for all pieces of data. This workflow replaces the need for researchers to search the literature and compile their own sets of equations.

The second workflow (workflow2), which also uses the get\_trait\_from\_taxon() function (argument: workflow = "workflow2"), automates this process. In workflow2, users provide a data.frame that includes target taxon names, the appropriate life-stage for each target taxa (matched with the life-stage classification in [Table 1](#tbl-1)), latitude-longitude coordinates and body length data corresponding to the length measurements accepted in the equation database for different taxa ([Table 1](#tbl-1)). The function then finds the equations that best match the target taxon names using the following set of hierarchical rules ([Figure 4](#fig-4)):

First, as with workflow1, the function selects a set of equations developed for taxa within some predefined maximum taxonomic distance from the target taxon names. Second, the function removes equations if the body size of the target taxon does not fall within the range (± 30%) of body sizes that the equation was developed for. This is important because allometric equations tend to be non-linear (West and Brown 2005). For example, if an equation was developed on specimens ranging from 2 to 5 mm but the body length of the target taxon is 8.5 mm, that equation will be removed. Third, any equations that do not match the life-stage of the target taxon are also removed. If there are multiple equations left over after these three steps, the function will choose the equation with the lowest taxonomic distance. Thus, the function focuses on first selecting equations based on taxonomic distance, the equation’s body size range and the life-stage as these are typically considered to be the most important factors for selecting equations (Benke et al. 1999).

When there are multiple equations with the same taxonomic distance and life stage and which match the body size of the target taxon, the function uses two other information sources to choose the best matching equation. First, the biogeographic realm, major habitat type and ecoregion of our target taxon is determined using the latitude-longitude coordinates provided by the user based on a freshwater ecoregion map (Abell et al. 2008). Then, whether there are matches (0 - no match, 1 - match) between the biogeographic realm (), major habitat type () and ecoregion () with the equation are tested. For each match a score of one third is given. Therefore, if the target taxon matches the biogeographic realm, major habitat type and ecoregion of an equation, a score of one is given. Second, the coefficient of determination ( value) of the equation (value between 0 and 1) reflecting the reliability of the equation is used as an additional score. The scores from the biogeographic realm, major habitat type and ecoregion along with the coefficient of determination of the equation are summed into a score for each equation () and the equation with the maximum score is then selected and outputted:

We chose to weight the biogeographic realm, major habitat type and ecoregion matches as equivalent to the value because, in our view, it is better to have a reliable equation from a different geographic/habitat condition than an unreliable equation with similar geographic/habitat conditions. In the unlikely case that there are still multiple equations, we choose the first equation in the list. We choose the first equation in the list rather than taking the average because the average might represent something that does not exist in nature. All relevant metadata like the original source from which the equation was derived are provided for users to check that the method uses reasonable equations and/or default lengths for calculating dry biomass.

### 0.5 Testing InvAlloR

To test the accuracy of InvAlloR, we compiled 206 direct measurements of body length and dry biomass from 73 taxa of freshwater invertebrates from 10 published studies (see Table S2 references). We used workflow2 to automatically estimate dry biomass data for these measurements. From this, we obtained estimates of dry biomass data for 38 of the 73 taxa (ca. 52%), and 114 individual dry biomass estimates in total, given the constraints on equation choice in workflow2 and database incompleteness. We then correlated the dry biomass estimates () from workflow2 with the direct dry biomass measurements () from the literature using Pearson’s correlation coefficient (log10-log10 scale). Moreover, we directly quantified the accuracy of workflow2 by calculating the percentage prediction error () as:

For comparison, we estimated dry biomass for the 145 direct measurements of body length from the 44 species using standard, order-level length-biomass equations from the literature. We then calculated PPE using the order-level length-biomass equations.

The dry biomass estimated by InvAlloR was highly correlated with the 114 measured dry biomass values from 39 species that we compiled from the literature (Pearson’s r = 0.98, CI95%: 0.97 - 0.99, log10-log10 scale). Moreover, most estimates fell close to the 1:1 line on the log10-log10 scale ([Figure 5](#fig-5)) but there were exceptions (e.g. Coleoptera and Odonata). The mean percentage prediction error (PPE) of the dry biomass estimated using InvAlloR from the measured dry biomass was 46.8%, more than 11% lower than the PPE of the dry biomass estimated using order-level allometric equations (Mean: 57.9%). Moreover, the maximum observed PPE of dry biomass estimated using InvAlloR from the measured dry biomass was 204%, considerably lower than the maximum PPE obtained using order-level allometric equations (maximum: 331%). In addition to comparing the results from workflow2 with direct measurements of body length and dry biomass, we tested whether workflow2 led to similar dry biomass estimates as those based on human experts that selected equations from the literature and found similarly good performance (Figure S2).

|  |
| --- |
| Figure 4: How InvAlloR selects equations. We start with the full database (eight circles). Then, all equations where the taxonomic distance from the target taxon is greater than the user-defined maximum are excluded. If workflow1 is chosen, equations within the maximum taxonomic distance are outputted. If workflow2 is chosen, all equations that do not match the body size of the target taxon are excluded. Next, equations without matching life stages are excluded. If there are multiple equations leftover, the equation with the minimum taxonomic distance is chosen. If multiple equations are still leftover, we use other information to select the equation (i.e. geographic/habitat information, coefficient of determination). |

|  |
| --- |
| Figure 5: The relationship between measured dry biomass (mg, log10) and the estimated dry biomass (mg, log10) from InvAlloR of 114 individuals of 38 species from 10 studies (different colours) grouped by taxonomic order. Lines between points connect measurements from the same taxon from the same study. The black, dashed line is the 1:1 line. The Pearson correlation coefficient (r) on a log10-log10 scale is also shown. |

All analyses were conducted in R v4.2.3 (R Core Team 2023). In addition to any R-packages cited previously, we used dplyr (Wickham et al. 2023), tidyr (Wickham, Vaughan, and Girlich 2024) and readr (Wickham, Hester, and Bryan 2024) for data handling, ggplot2 (Wickham 2016) for plotting. We used renv (Ushey and Wickham 2024) for package version management, and the accompanying renv lockfile can be consulted to retrieve and restore exact versions of all dependencies.

### 0.6 Conclusions

Our database of length-biomass equations and complementary R-package (InvAlloR) provides an automated equation selection procedure that can estimate dry biomass using body size data for freshwater invertebrate taxa that approximates measured dry biomass ([Figure 5](#fig-5)) and is more accurate than order-level length-biomass equations. Moreover, the method is fully reproducible which makes it a more systematic and standardised way to estimate biomass of freshwater invertebrates than individual researchers compiling their own sets of length-biomass equations. Finally, the method used to select equations using is general and can be used for any taxon-specific trait and is particularly useful for taxonomic groups (i) without robust phylogenies and (ii) where identifications are often uncertain. For example, our approach could be extended and used for classifying taxa into feeding groups or habitat preferences based on data from existing trait databases or to link taxonomically similar species to continuous traits such as body size or macronutrient content that are already available in the literature (Moreau et al. 2021).

**Acknowledgements:** We thank Lucas Van Passel for testing early versions of the R-package and Eoin O’Gorman for providing data used to test our approach. JH was supported by a doctoral studentship to JH from the University of Gothenburg. VD is supported by a predoctoral scholarship from the Fonds Wetenschappelijk Onderzoek in Belgium.

### 0.7 Reference list

Abell, Robin, Michele L. Thieme, Carmen Revenga, Mark Bryer, Maurice Kottelat, Nina Bogutskaya, Brian Coad, et al. 2008. “Freshwater Ecoregions of the World: A New Map of Biogeographic Units for Freshwater Biodiversity Conservation.” *BioScience* 58 (5): 403–14. <https://doi.org/10.1641/B580507>.

Benke, Arthur C., Alexander D. Huryn, Leonard A. Smock, and J. Bruce Wallace. 1999. “Length-Mass Relationships for Freshwater Macroinvertebrates in North America with Particular Reference to the Southeastern United States.” *Journal of the North American Benthological Society* 18 (3): 308–43. <https://doi.org/10.2307/1468447>.

Deboelpaep, Evelien, Tina Coenegracht, Lore De Wolf, Alexandre Libert, Bram Vanschoenwinkel, and Nico Koedam. 2020. “Bio-Energetic Data Show Weak Spatial but Strong Seasonal Differences in Wetland Quality for Waders in a Mediterranean Migration Bottleneck.” *Freshwater Biology* 65 (9): 1529–42. <https://doi.org/10.1111/fwb.13518>.

Giguère, Louis A., J.-F. St-Pierre, B. Bernier, A. Vézina, and j.-G. Rondeau. 1989. “Can We Estimate the True Weight of Zooplankton Samples After Chemical Preservation?” *Canadian Journal of Fisheries and Aquatic Sciences* 46 (3): 522–27. <https://doi.org/10.1139/f89-070>.

Gsell, Alena Sonia, Ulrike Scharfenberger, Deniz Özkundakci, Annika Walters, Lars-Anders Hansson, Annette B. G. Janssen, Peeter Nõges, et al. 2016. “Evaluating Early-Warning Indicators of Critical Transitions in Natural Aquatic Ecosystems.” *Proceedings of the National Academy of Sciences* 113 (50): E8089–95. <https://doi.org/10.1073/pnas.1608242113>.

Hébert, Marie-Pier, Vincent Fugère, Beatrix E. Beisner, Naíla Barbosa da Costa, Rowan D. H. Barrett, Graham Bell, B. Jesse Shapiro, Viviane Yargeau, Andrew Gonzalez, and Gregor F. Fussmann. 2021. “Widespread Agrochemicals Differentially Affect Zooplankton Biomass and Community Structure.” *Ecological Applications* 31 (7): e02423. <https://doi.org/10.1002/eap.2423>.

Howmiller, Richard P. 1972. “Effects of Preservatives on Weights of Some Common Macrobenthic Invertebrates.” *Transactions of the American Fisheries Society*, October. <https://www.tandfonline.com/doi/abs/10.1577/1548-8659%281972%29101%3C743%3AEOPOWO%3E2.0.CO%3B2>.

Leuven, Rob S. E. W., Theo C. M. Brock, and Hans A. M. van Druten. 1985. “Effects of Preservation on Dry- and Ash-Free Dry Weight Biomass of Some Common Aquatic Macro-Invertebrates.” *Hydrobiologia* 127 (2): 151–59. <https://doi.org/10.1007/BF00004193>.

Mehler, Knut, Kumud Acharya, and Donald W. Sada. 2015. “Spatial and Temporal Pattern in Length-Mass Regressions of Freshwater Gastropods in Nevada Spring Ecosystems.” *Malacologia* 58 (12): 167–77. <https://doi.org/10.4002/040.058.0205>.

Méthot, Ginette, Christiane Hudon, Pierre Gagnon, Bernadette Pinel-Alloul, Alain Armellin, and Anne-Marie Tourville Poirier. 2012. “Macroinvertebrate Sizemass Relationships: How Specific Should They Be?” *Freshwater Science* 31 (3): 750–64. <https://doi.org/10.1899/11-120.1>.

Meyer, Elisabeth. 1989. “The Relationship Between Body Length Parameters and Dry Mass in Running Water Invertebrates.” *Archiv Für Hydrobiologie*, December, 191–203. <https://doi.org/10.1127/archiv-hydrobiol/117/1989/191>.

Moreau, Axelle, Christine Dupuy, Pierrick Bocher, and Sébastien Farau. 2021. “Morphological, Calorific and Nutritive Characteristics of 656 Freshwater Invertebrates Taxa.” *Biodiversity Data Journal* 9 (October): e70214. <https://doi.org/10.3897/BDJ.9.e70214>.

O’Gorman, Eoin J., Lei Zhao, Doris E. Pichler, Georgina Adams, Nikolai Friberg, Björn C. Rall, Alex Seeney, Huayong Zhang, Daniel C. Reuman, and Guy Woodward. 2017. “Unexpected Changes in Community Size Structure in a Natural Warming Experiment.” *Nature Climate Change* 7 (9): 659–63. <https://doi.org/10.1038/nclimate3368>.

Patrick, C. J., D. J. McGarvey, J. H. Larson, W. F. Cross, D. C. Allen, A. C. Benke, T. Brey, et al. 2019. “Precipitation and Temperature Drive Continental-Scale Patterns in Stream Invertebrate Production.” *Science Advances* 5 (4): eaav2348. <https://doi.org/10.1126/sciadv.aav2348>.

Rogers, L. E., R. L. Buschbom, and C. R. Watson. 1977. “Length-Weight Relationships of Shrub-Steppe Invertebrates1.” *Annals of the Entomological Society of America* 70 (1): 51–53. <https://doi.org/10.1093/aesa/70.1.51>.

Ruiz-Lupión, Dolores, José María Gómez, and Jordi Moya-Laraño. 2020. “Masslength Allometry Covaries with Ecosystem Productivity at a Global Scale.” *Global Ecology and Biogeography* 29 (1): 87–101. <https://doi.org/10.1111/geb.13010>.

Towers, Dale J., Henderson ,Ian M., and Clare J. and Veltman. 1994. “Predicting Dry Weight of New Zealand Aquatic Macroinvertebrates from Linear Dimensions.” *New Zealand Journal of Marine and Freshwater Research* 28 (2): 159–66. <https://doi.org/10.1080/00288330.1994.9516604>.

Ushey, Kevin, and Hadley Wickham. 2024. “Renv: Project Environments.” <https://CRAN.R-project.org/package=renv>.

Wickham, Hadley. 2016. “Ggplot2: Elegant Graphics for Data Analysis.” <https://ggplot2.tidyverse.org>.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. “Dplyr: A Grammar of Data Manipulation.” <https://CRAN.R-project.org/package=dplyr>.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. “Readr: Read Rectangular Text Data.” <https://CRAN.R-project.org/package=readr>.

Wickham, Hadley, Davis Vaughan, and Maximilian Girlich. 2024. “Tidyr: Tidy Messy Data.” <https://CRAN.R-project.org/package=tidyr>.