# Chapter 63
# A Primer for Meta-Analysis

**Verónica Ferreira and Felix Bärlocher**

**Keywords** Effect size · Litter decomposition · Meta-regression · Publication bias · Sensitivity analysis · Software · Sub-group analysis · Systematic review

## 1 Introduction

Meta-analysis is a statistical tool that combines the results of several conceptually similar studies or experiments by providing a weighted average of the results of the individual studies. This pooled estimate is assumed to be close to the unknown true value. Meta-analysis allows answering the following questions: (1) Is the true effect significantly different from zero? (2) What are the magnitude and direction of the global effect? (3) Are magnitude and direction of the global effect influenced by any characteristics of individual studies or groups of studies? Additionally, it allows identifying knowledge gaps in the research field.

Meta-analysis is especially useful when sample sizes of individual studies are low or their effect sizes small or non-significant. It increases statistical power and tests consistency among individual studies. It also allows the testing of hypotheses that may be difficult to consider in individual studies (e.g. comparisons across biomes).

For meta-analysis to be useful, there needs to be a reasonably large number of empirical studies (although there is no minimum), and the collection of studies needs to be free of publication bias. All studies need to report quantitative measures

V. Ferreira (✉)
MARE – Marine and Environmental Sciences Centre & Department of Life Sciences, University of Coimbra, Coimbra, Portugal
e-mail: veronica@ci.uc.pt

F. Bärlocher
Department of Biology, Mount Allison University, Sackville, NB, Canada
e-mail: fbaerloc@mta.ca

© Springer Nature Switzerland AG 2020

F. Bärlocher et al. (eds.), *Methods to Study Litter Decomposition*,
https://doi.org/10.1007/978-3-030-30515-4_63

of the variables; purely descriptive reports such as case studies cannot be synthetized by meta-analysis.

Meta-analysis has been used to test ecological hypotheses and theories and the effect of covariates that are difficult to examine within a single primary study, assess the impacts of major environmental drivers and the effectiveness of management and conservation strategies, inform environmental risk assessment, and identify research gaps (Koricheva and Gurevitch 2014). Table 63.1 shows examples of meta-analyses in litter decomposition.

Meta-analysis is usually performed in the context of a systematic literature review, which includes *systematically* locating, selecting, and appraising sources (preferably peer-reviewed studies) and then synthetizing data from the selected sources. Each step is clearly documented to allow reproduction (Table 63.2).

**Table 63.1** Examples of meta-analyses in litter decomposition

| References | System | Factor under study | No. of studies | No. of effect sizes | Effect size | Global effect size (95% CL) |
|---|---|---|---|---|---|---|
| 1 | Soil | Nutrient enrichment | 24 | 500 | Response ratio (*R*) | 0.981 |
| 2 | Soil | Microarthropod presence | 30 | 101 | Hedges' *g* | 1.482 |
| 3 | Streams | Litter diversity | 11 | 510 | Signed deviation from additivity | –0.01[c] |
| 4 | Streams | Macroconsumer presence | 17 | 36 | Ln response ratio (ln*R*) | –0.016 (–0.023 to –0.009)[d] |
| 5 | Streams | Nutrient enrichment | 99 | 840 | Response ratio (*R*) | 1.49(1.41 – 1.58)[a] |
| 6 | Streams | Heavy metal contamination | 38 | 133 | Hedges' *g* | –0.81 (–1.02 to –0.61)[b] |
| 7 | Streams | Forest change (4 types) | 24 | 156 | Response ratio (*R*) | 0.82 (0.76 – 0.89)[a] |
| 8 | Streams | Acidification | 17 | 67 | Response ratio (*R*) | 0.37 (0.30 – 0.46)[a] |
| 9 | Streams | Water stress | 9 | 41 | Response ratio (*R*) | 0.69 (0.59 – 0.82)[a] |
| 10 | Streams | Water temperature | 34 | 148 | Hedges' *g* | 1.20 (0.96 – 1.43)[b] |

References: (1) Knorr et al. (2005); (2) Kampichler & Bruckner (2009); (3) Lecerf et al. (2011); (4) Mancinelli et al. (2013); (5) Ferreira et al. (2015); (6) Ferreira et al. (2016a); (7) Ferreira et al. (2016b); (8) Ferreira and Guérold (2017); (9) Sabater et al. (2018); (10) Amani et al. (2019)

[a] $R > 1$, stimulation of decomposition; $R < 1$, inhibition of decomposition

[b] $g > 0$, stimulation of decomposition; $g < 0$, inhibition of decomposition

[c] Negative values indicate synergistic response of litter decomposition to litter diversity

[d] $\ln R < 0$, inhibition of decomposition

# 2 Preparation of the Database

## *2.1 Definition of the Question*

The first step in a systematic literature review is to clearly define the question (or hypothesis) to address. This will determine the scope of the review and literature search (Sect. 2.2), inclusion/exclusion criteria (Sect. 2.3), and the type of information to extract (Sect. 2.5). The question should be broad enough to capture a sufficient number of empirical studies (e.g. 'Does nutrient enrichment affect litter decomposition in running waters?') but not so broad that it will become unmanageable (e.g. 'Does environmental change affect ecosystem functioning?', which would include the effect of any environmental change on any ecosystem function in any ecosystem). A useful strategy to define the question is the PICO method, where the question clearly identifies the population (P), the intervention (I), the control (C; can be implicit), and the outcome (O). In the question 'Does nutrient enrichment affect litter decomposition in running waters?', 'running waters' is the population, 'nutrient enrichment' is the intervention, it is implicit that a non-nutrient enriched condition is the control, and 'litter decomposition' is the outcome.

## *2.2 Intensive and Extensive Literature Search*

The literature search should be intensive and extensive to ideally locate *all* studies that have ever addressed the question of interest. This is generally impossible because some studies may not be published (recent studies or studies not submitted or rejected due to publication bias); may belong to the 'grey literature' (e.g. theses, reports, conference abstracts), which is generally difficult to locate and retrieve; or may be inaccessible for other reasons (e.g. language bias, when published in languages unknown to the meta-analysist) (Sect. 3.6). Clearly, there needs to be a considerable effort to ensure that the studies located and retrieved are a random sample of the studies performed. The literature search protocol will always have an element of subjectivity (e.g. time frame, languages, key words, search paths), but needs to be transparent and may need to be revised repeatedly to address potential biases (Sect. 3.6).

An intensive and extensive literature search may include studies published in the mainstream literature and 'grey literature', in several languages and over a large time frame (but note that methods to determine the outcome of interest may have changed over time, which has to be coded). The literature search should be done via multiple paths, including personal literature databases, reference lists in relevant primary studies and reviews, scientific journal indices, and online databases (e.g. Google Scholar, Web of Science, Scopus). Different search paths generally retrieve different sets of studies and should be used to complement each other. The set of key search words should be clearly defined to allow the search to be reproduced. It may

**Table 63.2** Steps to carry out a meta-analysis in the context of a systematic review

| **Systematic review** | **Meta-analysis** |
|---|---|
| Definition of the question | Estimation of effect sizes and associated variances |
| Intensive and extensive literature search | Selection of the model |
| Definition of the inclusion/exclusion criteria | Estimation of the global effect size and precision |
| Critical appraisal of studies | Evaluation of publication bias |
| Data extraction | Test of heterogeneity and effect of moderators |
| Statistical synthesis (e.g., **meta-analysis**) | Sensitivity analysis |
| Discussion and conclusion based on the most precise studies | Interpretation and presentation of results |

be useful to include a search in libraries and conference abstract books when targeting 'grey literature'. The use of mailing lists (e.g. ECOLOG-L) or direct contact with researchers known to work on the topic of interest may locate additional studies. It is essential to keep track of all steps in the literature search, annotating when and how each study was located and retrieved.

Most of the studies that seem relevant based on title and abstract can generally be retrieved without delay (e.g. by downloading from journal web pages or online databases). An additional effort may be required to retrieve older studies or data from 'grey literature'. This may include contacting the author of the study, the library, or the author of a recent study where the study of interest has been cited.

## 2.3 Definition of Inclusion/Exclusion Criteria

A literature search generally retrieves many studies that are not relevant or useful. The studies to be included in the analysis are selected based on clearly defined inclusion/exclusion criteria. Studies need to report information that will allow estimating effect sizes (and associated variability), which are interpreted as 'dependent variable' in a meta-analysis (Sect. 3.2). For instance, to address the question 'Does nutrient enrichment affect litter decomposition in running waters?', we should only consider empirical studies that report (1) litter decomposition in at least one nutrient-enriched condition and one control (non-enriched) condition, (2) a measure of variation of decomposition values (variance, SD, SE, 95% CL), and (3) sample size. If some of this information is missing, it may be available through a request to the author. If a few studies lack information on variation, they can still be included as the missing values may be imputed (Koricheva et al. 2013; Ferreira et al. 2015).

Additional inclusion/exclusion criteria may be defined. To address the question above, possible inclusion/exclusion criteria could be that the primary studies (1) report decomposition of natural litter rather than artificial substrates and (2) rely on allochthonous rather than autochthonous litter (Ferreira et al. 2015). Inclusion/exclusion criteria may also refer to research methods used (e.g. studies that use a specific method) and may need to be revised to address publication bias issues (Sect. 3.6).

## *2.4 Critical Appraisal of Studies*

The selected studies need to be critically appraised, especially concerning methodological quality and multiple publications. The methodological quality of studies can be coded as a moderator (Sect. 3.5) or used to assess its impact on the analysis (sensitivity analysis; Sect. 3.7). Special care is needed to detect information that has been published multiple times to avoid overweighting these data. The number of studies used in the analysis may thus have to be reduced to avoid counting identical information more than once.

## *2.5 Data Extraction*

### Basic Data to Estimate Effect Size and Associated Variance

Data in ecological studies can be reported in several formats, which will determine the type of effect size (and its variance) that can be estimated (Table 63.3; Borenstein et al. 2009). If data are based on a comparison of two groups of continuous variables, then information on the variable of interest (outcome), measure of variability of the outcome (variance, SD, SE, 95% CL), and sample sizes of control and treatment conditions need to be extracted. If data are reported as a comparison of two groups in terms of categorical variables, then information on sample size and number of cases in the event and non-event situation, in the control and treatment conditions, need to be extracted. If data are reported as the relationship between two continuous variables, then Pearson's r and sample size need to be extracted.

**Table 63.3** Types of data in ecology, types of effect size, and examples of common effect sizes

| Type of data | Type of effect size | Effect size |
|---|---|---|
| Comparison of 2 groups in terms of continuous variables (e.g. control × treatment) | Effect size based on means | Raw mean difference (*D*)<br>Standardized mean difference (Cohens' *d* or Hedges' *g*)<br>Response ratio (*R*) |
| Comparison of 2 groups in terms of categorical variables (2 × 2 contigency table) | Effect size based on binary data | Risk difference (*RD*)<br>Risk ratio (*RR*)<br>Odds ratio (*OR*) |
| Relationship between 2 continuous variables | Effect size based on correlations | Pearson correlation (*r*) |

For details on the estimation of each effect size and associated variance, see Borenstein et al. (**2009**); these are estimated automatically in any software for meta- analysis

## Explanatory Information to Assess the Effect of Study/ Environmental Conditions

Several experimental and environmental explanatory variables – termed moderators in meta-analysis – may affect the response of the dependent variable and explain differences in effect sizes among studies. Meta-analysis allows testing the significance and strength of these moderators ('independent variables') (Sect. 3.5). Only moderators backed up by a hypothesis should be coded. In addition, research procedures likely to affect the response of the variable of interest (e.g. type of study, specific methodologies) should be coded to be later used in sensitivity analyses (Sect. 3.7).

Information on selected moderators (continuous or categorical variables) needs to be extracted from the studies and additional sources (e.g. websites for climatic information) or by contacting the author. For instance, to address the question 'Does nutrient enrichment affect litter decomposition in running waters?', categorical moderators may include type of study, scale of nutrient enrichment and identity of the nutrient used in field manipulative studies, type of aquatic decomposers, type and identity of litter, and climate. Continuous moderators may include the mean dissolved nutrient concentration in the control and the magnitude of the increase in nutrient concentrations compared to the control condition (Ferreira et al. 2015). Additionally, the type of report (i.e. published in the mainstream or 'grey literature'), the type of data (i.e. reported in the study or imputed/estimated), and methodological specifications may be coded (Ferreira et al. 2015).

# 3 Meta-Analysis

## 3.1 Software for Performing Meta-Analysis

Once the database is ready (i.e. study information, basic data for the estimation of effect sizes and associated variance, and information on moderators, generally one case by line; see, e.g. Table S1 in Ferreira et al. 2015), it can be entered into a software spreadsheet. Many options are available (Koricheva et al. 2013), including:

*OpenMEE,* an open-source, user-friendly software designed by ecologists for data in ecology and evolution (Wallace et al. 2017). It offers diverse and advanced statistical options based on packages developed for R without requiring programming skills. The software, as well as the user guide, can be downloaded at http://www.cebm.brown.edu/openmee/

*Comprehensive Meta-Analysis (CMA),* a commercial, user-friendly software developed by specialists in social sciences and medicine (Borenstein et al. 2017). It allows entering data for estimating effect sizes and associated variances in 100 different formats. A trial version, as well as user guides and a large bibliography in meta-analysis, can be downloaded at https://www.meta-analysis.com/

The *metafor package for R,* an open-source, highly versatile package that requires familiarity with R. Codes for performing a meta-analysis are provided in Viechtbauer (2010)

## 3.2 Effect Size and Precision

The effect size reflects the magnitude of the effect of a treatment or the strength of the relationship between two variables; it is estimated for each (case) study and used to estimate the global effect size. It may be necessary to estimate different effect sizes for different studies, depending on the format used to report data (Table 63.3). However, the various types of effect sizes are interconvertible (Table 63.3) so that the analysis is based on a single effect size.

Differences in sample size will affect the precision of the estimated effect size. Provided there is no systematic bias, the precision defines the interval containing the true effect size and indicates how much we can trust the estimated effect size. Effect sizes associated with a larger variance are given less weight in the meta-analysis. Variance estimates are specific to each effect size (Borenstein et al. 2009).

One of the most common effect sizes in ecology is the response ratio $R$ (Koricheva and Gurevitch 2014), the ratio of the variable of interest in the treatment condition to the variable in the control condition (e.g. $R = k_{\text{nutrient-enriched}}/k_{\text{control}}$, $k$ = decomposition rate) (Hedges et al. 1999). Being a ratio, it can only be used if outcomes differ from zero. It is very easy to interpret ($R$ = 1 indicates no treatment effect, while $R > 1$ and $R < 1$ indicate higher and lower values in the treatment than in the control condition, respectively). The analyses are performed with the natural logarithm of $R$ (ln$R$), which is normally distributed even in small samples, but the results can be back transformed to $R$ to facilitate interpretation.

### 3.3 Global Effect Size

If all effect sizes had the same precision, they could be simply averaged to estimate the global effect size. However, effect sizes generally differ in their precision; the global effect size is therefore estimated by a weighted average of the individual effect sizes, with a larger weight given to the more precise effect sizes (i.e. associated with smaller variances).

There are two models to estimate the global effect size in standard weighed meta-analysis:

The *fixed-effect model* assumes a unique and true (fixed!) effect size shared by all studies and that the observed differences among individual effect sizes are due to sampling error. In this case, the weight attributed to each individual effect size is based on the inverse of the within-study variance (or sampling error). This model should be chosen only when the variation among studies is negligible; the goal is to estimate the global effect size for the studies considered and not to extrapolate beyond the analysed studies. These conditions are rarely met in ecological studies, and thus, this is not a model commonly used in ecology.

The *random-effects model* assumes that the true effect size varies among studies and that the analysed studies provide a random sample of the distribution of effect sizes, with the global effect size being the mean of that distribution. In this case, the weight attributed to each individual effect size is based on the inverse of the total variance, which is the sum of the within-study variance (or sampling error) and the between-study variance. This model should be used when the effect sizes are expected to vary among studies (e.g. due to different experimental conditions, ecosystems, species, etc.); the goal is to generalize the global effect size, which is considered to be the mean of the true effect sizes. This is the most common model in meta-analysis in ecology.

In both models, the global effect size is estimated as the sum of individual effect sizes weighted by the inverse of the corresponding variance, corrected by the sum of the weights. The variance of the global effect size is estimated as the inverse of the sum of the weights and can be converted into 95% CL. Significant effect sizes occur when 95% CL do not include 1 for effect sizes based on ratios (e.g. $R$) or 0 for effect sizes based on the natural logarithm of ratios (e.g. ln$R$), differences, or Pearson's r (Borenstein et al. 2009).

The meta-analysis result is presented in a forest plot, which shows the effect sizes of the individual studies considered as well as the global effect size, and their associated variability (generally, 95% CL; Fig. 63.1).

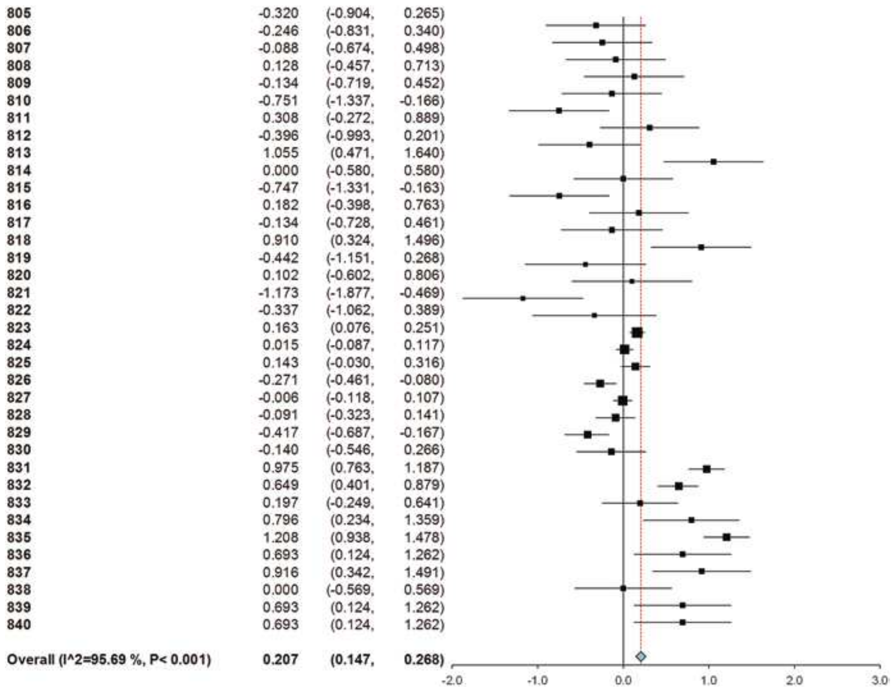



**Fig. 63.1** Forest plot (detail of the last 35 effect sizes) of the response of litter decomposition to nutrient enrichment of stream water in correlative field studies (ln$R$ = ln($k_{\text{nutrient-enriched}}/k_{\text{control}}$); $k$ = decomposition rate; $n$ = 521). Individual effect sizes (squares; the size of the symbols reflects their precision) and associated 95% CL, the global effect size (blue diamond; red line) and associated 95% CL (width of the diamond), and the percentage of total variation due to true variation among effect sizes ($I^2$ = 95.69%) are depicted. A random-effects model was used, with the restricted maximum likelihood method to determine between-study variance. The solid line (ln$R$ = 0) indicates no effect of nutrient enrichment, and ln$R$ > 0 and ln$R$ < 0 indicates stimulation or inhibition, respectively. Significant effect sizes occur when 95% CL do not include zero. To facilitate interpretation, the result is back transformed into $R$: ln$R$ = 0.207 (0.147–0.268) → $R$ = 1.23 (1.16–1.32), which indicates that litter decomposition rates increase significantly by 23% in nutrient-enriched streams (output from OpenMEE, data from Ferreira et al. 2015)

## 3.4 Heterogeneity

In addition to estimating the global effect size, it is useful to identify and quantify heterogeneity in effect sizes. This addresses the sources of differences among studies rather than combining all the effect sizes into a global value.

The heterogeneity, or observed variation among effect sizes, may result from true variation in effect sizes (between-study variance) and from sampling error (within-study variance). The true variation is of primary interest. The observed variation among effect sizes is given by the $Q$-statistic and is estimated as the sum of squared differences between individual effect sizes and the global effect size weighted by the inverse of the variance associated with the individual effect sizes (Borenstein et al. 2009). The $Q$-statistic has a $\chi^2$ distribution ($H_o$: $Q = df$, $df = n - 1$, $n$ = number of effect sizes), with significant $p$-values ($Q > df$) suggesting that the variation among effect sizes is not all due to chance. The causes for heterogeneity are then explored (Sect. 3.5). Non-significant $p$-values ($Q = df$) suggest that the $H_o$ cannot be rejected, which suggests that effect sizes are identical (i.e. there is no between-study variance) and variation among them is due to change (i.e. sampling error); however, a non-significant result may also be due to low statistical power. Thus, it may still be useful to test for moderators when no significant variation is observed among effect sizes (Sect. 3.5). The percentage of observed variation that is due to real differences in effect sizes (i.e. between-study variance) is given by the $I^2$ statistics: $((Q - df)/Q)) \times 100$. $I^2$ values vary between 0% and 100%; $I^2$ values ~25% indicate low heterogeneity, $I^2$ values ~50% indicate moderate heterogeneity, and $I^2$ values ~75% indicate high heterogeneity (Fig. 63.1).

## *3.5 Test of Moderators*

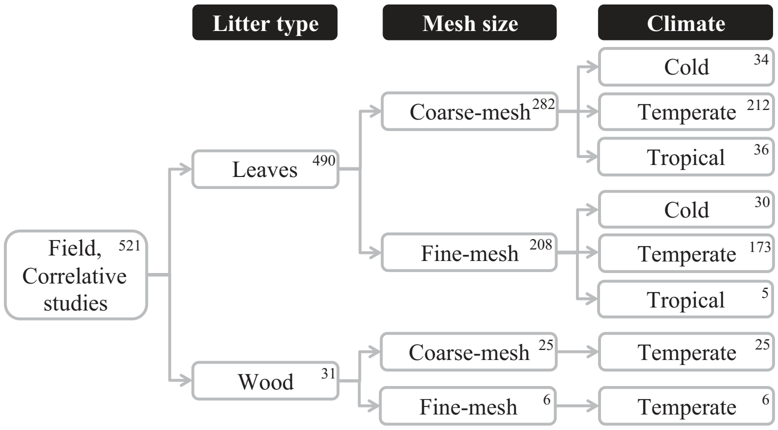
Variation in effect sizes may be due to differences in experimental or environmental characteristics, categorized as moderators. We can assess if they are systematically associated with variation in effect sizes. Moderators are coded, and their effects are assessed in subgroup analysis (categorical moderators) or meta-regression (continuous moderators).

### Subgroup Analysis

In subgroup analysis, effect sizes are grouped by common features (i.e. levels within a given moderator), and the global effect size for each subgroup (levels) is estimated and tested for heterogeneity. A hierarchical approach to test moderators is often useful; effect sizes are stratified, and comparisons of effect sizes among levels of one moderator are made within a level of another moderator (e.g. comparison between the levels 'Coarse mesh' and 'Fine mesh' of the moderator 'Mesh size' within the level 'Leaves' of the moderator 'Litter type'; Table 63.4; Fig. 63.2).

There are multiple models and methods for performing subgroup analysis (Borenstein et al. 2009), which cannot be covered here. In ecology, the most commonly used approaches are the random-effects model with between-study variance pooled for the estimation of the effect size for each subgroup and global effect size and the random-effects model with between-study variance estimated for each subgroup (i.e. not pooled) for the estimation of the effect size for each subgroup and global effect size.

**Table 63.4** Hierarchical diagram for categorical moderators (Ferreira et al. 2015)




The moderator 'Litter type' has two levels ('Leaves' and 'Wood'), the moderator 'Mesh size' has two levels ('Coarse mesh' and 'Fine mesh'), and the moderator 'Climate' has three levels ('Cold', 'Temperate', and 'Tropical'). Values indicate sample size (i.e. number of effect sizes)

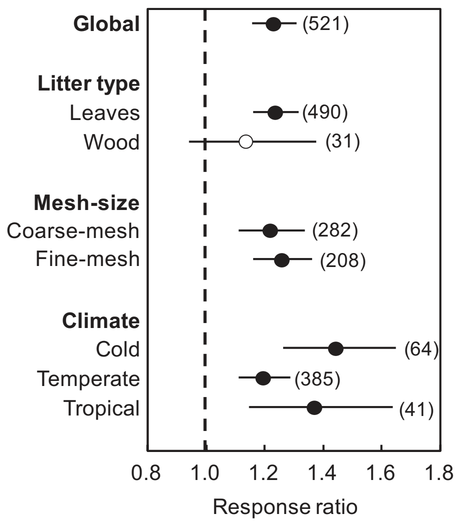



**Fig. 63.2** Subgroup analyses of the response of litter decomposition to nutrient enrichment (response ratio $R$, 95% CL); same data as in Table 63.4 (random-effects model; restricted maximum likelihood method for between-study variance). Global effect size and effect size as a function of litter type (two levels), mesh size (two levels; considering only Leaves since non-significant $R$ was found for Wood), and climate (three levels; considering only Leaves since non-significant $R$ was found for Wood, but considering both mesh sizes since significant $R$ was found for both, with no significant difference between them). The dashed line ($R = 1$) indicates no effect of nutrient enrichment on litter decomposition, and $R > 1$ indicates a stimulation of litter decomposition. The effect of nutrient enrichment is significant when 95% CL do not include 1 (black symbols). Within each moderator (in bold), levels with overlapping 95% CL do not significantly differ. Values in parenthesis indicate sample size (output from OpenMEE, data from Ferreira et al. 2015)

### Meta-Regression

In meta-regression, effect sizes (weighted by their precision) are regressed against continuous moderators to explain observed variation. There are multiple procedures for meta-regression (Borenstein et al. 2009). One is based on the random-effects model. The meta-regression will produce values for the intercept, the slope, and associated statistics. The significance of the slope is assessed by the *Z*-test given by the ratio between the slope and its SE; if several covariates are used simultaneously, then the *Q*-test is used.

The model is tested using the *Q*-statistic (weighted sum of squares; Sect. 3.4). A significant $Q_{\text{model}}$ (i.e. variation explained by the moderator) indicates that the relationship between effect sizes and the moderator is stronger than expected by chance. The goodness of fit test ($Q_{\text{resid}}$) assesses if there is heterogeneity that is not explained by the moderators and can be used to estimate the variance of this unexplained heterogeneity; a significant $Q_{\text{resid}}$ indicates that some between-study variance remains unexplained. The proportion of variation in effect sizes that is explained by the model ($R^2$) is given by the ratio between true variance explained and total true variance (Fig. 63.3).

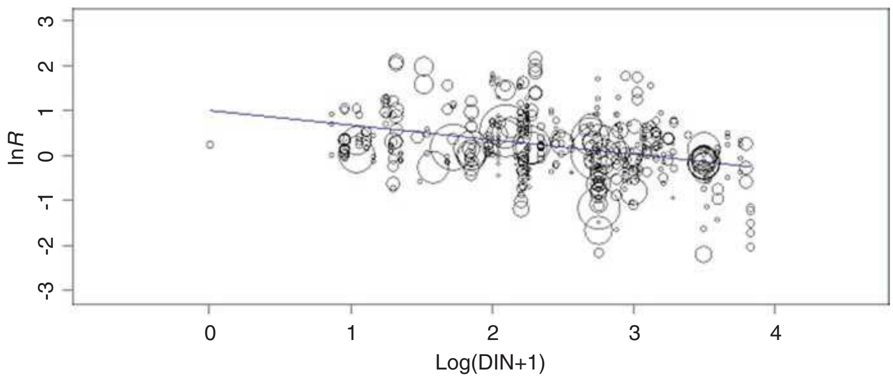



**Fig. 63.3** Meta-regression to assess the effect of dissolved inorganic nitrogen (DIN) concentration in control conditions (Log ($x$ + 1)-transformed) on the response of litter decomposition to nutrient enrichment (ln*R*) ($n$ = 511). The size of the symbols reflects their precision, with effect sizes with larger symbols being more precise and thus given greater weight in the analysis. The response of litter decomposition to nutrient enrichment decreases by 0.326 (slope; $p < 0.001$) for each unit increase in DIN concentration in control conditions suggesting that the response for litter decomposition to nutrient enrichment is stronger in systems with naturally low nutrient concentration. The model explains 13% of the variation (output from OpenMEE, data from Ferreira et al. 2015)

### *3.6 Publication Bias*

A meta-analysis allows a precise synthesis of a selected set of studies, but if the selected studies do not fairly represent all conducted studies, then the estimated global effect size will be biased. Bias may be the result of publication bias, when the results of the primary studies affect their probability of being published, the so-called file drawer problem (e.g. non-significant results or results contrary to current theory are less likely to be submitted to journals). Bias may also occur as a result of dissemination bias, when the presentation of the primary studies affects their probability of being found (e.g. language bias, English vs. other languages; citation bias, highly cited vs. less cited studies; publication site bias, high- vs. low-impact journals). Both publication bias and dissemination bias are known jointly as publication bias.

An intensive and extensive literature search (Sect. 2.2) can overcome many of these potential biases. However, there is always the need to evaluate if the database or subsets of the database (and thus the results of the analyses) are affected by publication bias. There are several methods to assess the potential impact of publication bias in the meta-analysis, which assume that bias is negatively correlated with sample size in primary studies. These methods include:

*Comparing global effect size of published studies* vs. *'grey literature'*. In the absence of bias, the results should not differ significantly. If they differ (and there are no important differences in methodological/environmental characteristics between the two groups), this indicates publication bias, which might be corrected by including the 'grey literature'. The database should still be tested using the methods below.

*Relating effect size and precision* (meta-regression). Without bias, the relationship should be non-significant; if low precision studies have larger effect sizes than more precise studies, the analysis should focus on more precise studies.

*Fail safe numbers* (*Nfs*; e.g. Rosenthal, Orwin, Rosenberg) estimate the number of studies with non-significant results (which may have been missed in the literature search) that are needed to nullify the global effect size (Rosenthal, Rosenberg) or to reduce it to an ecologically non-relevant value (Orwin) (Borenstein et al. 2009). It does not consider missing studies that may report results in the opposite direction. If $Nfs > 5 \times n + 10$ ($n$ = number of effect sizes), the global effect size is robust to publication bias; if the *Nfs* are lower, the literature search and/or inclusion/exclusion criteria may need to be revised.

The *funnel plot* is a scatter plot of effect sizes vs. precision or sample sizes that, in the absence of bias, is symmetrical around the global effect size with a wider distribution of effect sizes for less precise studies. This gives the plot a funnel shape. It is less efficient when the number of effect sizes is low and it does not consider that asymmetry may have other causes (e.g. differences in experimental approaches among studies).

*Duval and Tweedie's trim and fill method*, which imputes the 'missing' effect sizes to the funnel plot and estimates a new global effect size (Borenstein et al. 2009). The comparison of the original global effect size and the new estimate allows assessing the degree to which the original global effect size is affected by publication bias. If publication bias has a strong effect on the results, the literature search and/or inclusion/exclusion criteria may need to be revised.

### 3.7 Sensitivity Analyses

Several decisions may affect the outcome of a meta-analysis. Sensitivity analyses allow assessing if and to what extent these decisions affect the results, i.e. they assess the robustness of the results. Technically, sensitivity analyses imply repeating some analyses using different criteria and performing subgroup analyses or meta-regressions on different data sets or using 'decision' moderators (Ferreira et al. 2015). Sensitivity analyses may include assessing the effect of:

*Considering multiple effect sizes per study*. This is typical of ecological studies but violates the assumption of independence of effect sizes. Thus, it is necessary to show that violating this assumption does not strongly affect the results. To that end, a single effect size per study is computed (using study as moderator in a subgroup analysis), and a new global effect size is estimated and compared with the global effect size estimated from all effect sizes.

*Study quality* by comparing effect sizes in different classes of study quality (subgroup analysis) or assessing the relationship between effect size and study quality (meta-regression).

*Including 'grey literature'*, which may be considered of substandard quality since it did not go through peer review, by comparing effect sizes from published studies vs. 'grey literature' (subgroup analysis).

*Including effect sizes with imputed or recalculated data*, which may be less accurate than reported data, by comparing effect sizes from reported data vs. effect sizes from imputed or recalculated data (subgroup analysis).

*Including particular studies* (e.g. studies that contribute with an exceptionally large number of effect sizes, studies with unusual characteristics), by comparing results with and without these studies.

## 4 Quality in Meta-Analysis

The number of systematic reviews using meta-analysis is increasing in ecology, but not all reports are of high quality (Koricheva and Gurevitch 2014). Although meta-analysis is generally performed in the context of a (systematic) review, its procedures and reporting should follow closely those of a primary study. A traditional literature review generally makes no attempt to locate *all* studies, does not describe why some studies are included and others excluded, and has limited capability to deal with a large number of studies with variable outcomes. This gives it a high degree of subjectivity and low degree of reproducibility.

In a *systematic* literature review, the reviewer needs to keep record of all steps and decisions and report these with sufficient detail to allow reproduction of the results (as in a primary study!). A systematic literature review needs to start with preparing detailed protocols (as in a primary study!) for literature search, inclusion/exclusion criteria, study appraisal, and coding (these protocols can be revised during the process, with all changes being annotated). The Collaboration for Environmental Evidence website (http://www.environmentalevidence.org/) lists examples of protocols for systematic literature reviews.

The number of studies located and retained at each step of the literature search, application of inclusion/exclusion criteria, and study appraisal can be presented in a PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses; Liberati et al. 2009) flow diagram, which can be generated online (PRISMA flow diagram generator: http://prisma.thetacollaborative.ca/).

In order to carry out and report a high-quality meta-analysis, the quality criteria compiled by Koricheva and Gurevitch (2014) should be checked.

## References

Amani, M., Graça, M. A. S., & Ferreira, V. (2019). Effects of elevated atmospheric CO2 concentration and temperature on litter decomposition in streams: A meta-analysis. *International Review of Hydrobiology, 104*, 14–25.

Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester: Wiley.

Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2017). *Comprehensive meta-analysis, version 3*. Englewood: Biostat.

Ferreira, V., & Guérold, F. (2017). Leaf litter decomposition as a bioassessment tool of acidification effects in streams: Evidence from a field study and meta-analysis. *Ecological Indicators, 79*, 382–390.

Ferreira, V., Castagneyrol, B., Koricheva, J., Gulis, V., Chauvet, E., & Graça, M. A. S. (2015). A meta-analysis of the effects of nutrient enrichment on litter decomposition in streams. *Biological Reviews, 90*, 669–688.

Ferreira, V., Koricheva, J., Duarte, S., Niyogi, D. K., & Guérold, F. (2016a). Effects of anthropogenic heavy metal contamination on litter decomposition in streams – A meta-analysis. *Environmental Pollution, 210*, 261–270.

Ferreira, V., Koricheva, J., Pozo, J., & Graça, M. A. S. (2016b). A meta-analysis on the effects of changes in the composition of native forests on litter decomposition in streams. *Forest Ecology and Management, 364*, 27–38.

Hedges, L. V., Gurevitch, J., & Curtis, P. S. (1999). The meta-analysis of response ratios in experimental ecology. *Ecology, 80*, 1150–1156.

Kampichler, C., & Bruckner, A. (2009). The role of microarthropods in terrestrial decomposition: A meta-analysis of 40 years of litterbag studies. *Biological Reviews, 84*, 375–389.

Knorr, M., Frey, S. D., & Curtis, P. S. (2005). Nitrogen additions and litter decomposition: A meta-analysis. *Ecology, 86*, 3252–3257.

Koricheva, J., & Gurevitch, J. (2014). Uses and misuses of meta-analysis in plant ecology. *Journal of Ecology, 102*, 828–844.

Koricheva, J., Gurevitch, J., & Mengersen, K. (2013). *Handbook of meta-analysis in ecology and evolution*. Princeton/London: Princeton University Press.

Lecerf, A., Marie, G., Kominoski, J. S., LeRoy, C. J., Bernadet, C., & Swan, C. M. (2011). Incubation time, functional litter diversity, and habitat characteristics predict litter-mixing effects on decomposition. *Ecology, 92*, 160–169.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., Clarke, M., Devereaux, P. J., Kleijnen, J., & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *PLoS Medicine, 6*, e1000100.

Mancinelli, G., Sangiorgio, F., & Scalzo, A. (2013). The effects of decapod crustacean macroconsumers on leaf detritus processing and colonization by invertebrates in stream habitats: A meta-analysis. *International Review of Hydrobiology, 98*, 206–216.

Sabater, S., Bregoli, F., Acuña, V., Barceló, D., Elosegi, A., Ginebreda, A., Marcé, R., Muñoz, I., Sabater-Liesa, L., & Ferreira, V. (2018). Effects of human-driven water stress on river ecosystems: A metaanalysis. *Scientific Reports, 8*, 11462.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*, 1–48.

Wallace, B. C., Lajeunesse, M. J., Dietz, G., Dahabreh, I. J., Trikalinos, T. A., Schmid, C. H., & Gurevitch, J. (2017). OpenMEE: Intuitive, open-source software for meta-analysis in ecology and evolutionary biology. *Methods in Ecology and Evolution, 8*, 941–947.