Project Number 289706
Start date of the project: 01/12/2011, duration: 54 months

Deliverable 9.4

Environmental Risk Assessment of Genetically Modified Organisms: Protocols for statistical aspects of non-target effects field studies

Authors:

Hilko van der Voet, Paul W. Goedhart, Johannes W. Kruisselbrink

Organisation name of lead contractor for this deliverable:

DLO: Wageningen University and Research centre,
Plant Research International, Biometris, Wageningen, Netherlands
http://www.biometris.nl

in collaboration with AMIGA partners conducting the field trials
and collecting the data:
INIA, SAU, AU, LU, DLO, TEAGASC, WU, TI

May 2016

Dissemination Level: Public

Contents

Abstract

The EFSA Guidance on the environmental risk assessment (ERA) of genetically modified (GM) plants gives broad guidance on the design and analysis of field experiments. The AMIGA research project aims at providing more detailed guidance in the form of protocols for design and analysis. This report provides statistical elements for such protocols.

The protocol for experimental design specifies all elements that are needed to perform a prospective power analysis. This includes the specification of a list of endpoints and their hierarchical relations, the specification of intended levels of analysis, and the specification of provisional limits of concern to be used as trigger values for further investigation.

The AMIGA Power Analysis software is presented, and examples of its use are given. Two scenarios are illustrated, one concerning the replication in a field trial in a single environment, the other regarding the number of environments that would be needed in a multi-environment context.

The protocol for statistical analysis presents a flow chart for approaches to be used, and shows how to prepare graphical representations of the results of the analysis. Emphasis is placed on showing estimates and confidence intervals for effects such as the ratio of expected abundances or the fold change. Interpretation is mainly by comparing the estimates to the limits of concern (equivalence tests) rather than by calculating p values from traditional tests to detect differences (difference tests).

The proposed statistical analyses are illustrated with data from the AMIGA field work on maize in four different countries over several years, and on potato in two different countries in two different years. It is indicated that, depending on the research questions and expert choices, many different ways of analysing the data are possible. Choices, such as those of limits of concern, typically are provisional, and open to changes based on motivated reasoning.

1 Introduction

The EFSA Guidance on the environmental risk assessment (ERA) of genetically modified (GM) plants (EFSA 2010) gives broad guidance on the design and analysis of field experiments. The AMIGA research project aims at providing more detailed guidance in the form of ***protocols for design and analysis*** (Arpaia et al. 2014). This report provides statistical elements for such protocols.

For the design of experiments we focus on methods to perform a prospective power analysis as required by Perry et al. (2009) and EFSA (2010). An important element is the specification of non-zero effect sizes that are of sufficient interest. Such effect sizes (here termed limits of concern) are essential for power analyses, but also allow to rephrase the testing procedure in terms of equivalence tests.

A commonly encountered problem in entomology is the occurrence of many taxa with many zero catches (per plot), and perhaps only a few specimens overall. A typical way of addressing this problem is omit such data from the analysis, but this raises the question of criteria for when to omit a taxon. In this research we have investigated alternative methods of statistical analysis, and will advise on the method to be used (see also Goedhart et al. 2014). Rare taxa are generally perceived to be of minor concern for ecological functions, and a flexible system to allow for this fact will be presented. On the other hand, within any system, risk assessors or risk managers should always have the possibility to overrule the proposed simple settings of limits of concern, and specify their own well-motivated values.

The ecological situation and the practical possibilities for conducting field trials may be diverse across multiple biogeographical regions in Europe. As a consequence data may be different between field trials, e.g. because some taxa are only observed in certain regions, or because the identification of sampled arthropods is dependent on the varying expertise of local experts. We propose a hierarchical analysis to deal with such issues.

This report (D9.4) describes the results of research in AMIGA Work Package 9, Tasks 4 and 5. These tasks generalize the work on single-environment trials (earlier reported in deliverable D9.2, see van der Voet & Goedhart 2014, Goedhart & van der Voet 2014) to the general situation where multiple trials are performed in different biogeographical regions and in different years. Therefore we focus on integration methods to summarise conclusions from statistical analyses over multiple environments or over other factors of interest. The report is accompanied by the AMIGA software for power analysis as a separate deliverable D9.5. Suggestions on how the results of this research could be of importance for an update of the EFSA guidance document are given in a separate deliverable D9.6.

The examples presented in this report are based on experiments performed in the AMIGA project, for maize in Spain by INIA (Cristina Chueca and team), in Slovakia by SAU (Ľudovít Cagáň and team), in Denmark by AU (Gabor Lövei and team) and in Sweden (Tina D'Hertefeldt and team), and for potato in the Netherlands by DLO (Bert Lotz, Geert Kessel and team) and in Ireland by TEAGASC (Ewen Mullins and team). In addition, data on arthropods in the potato field were obtained from WU (Joop van Loon, Jenny Lazebnik) and

data on soil micro-organisms from Thünen Institut (Christoph Tebbe, Astrid Näther). We gratefully acknowledge the input and help from these partners which was crucial in this work.

2 Motivation for a protocol for the statistical aspects of designing experimental studies of non-target effects

2.1 Prospective power analysis

The costs of field experiments are high and therefore field experiments should only be conducted if they contribute sufficiently to answering the research questions. For more detailed research questions the necessary size and costs of a field experiment to answer the questions will become higher, as can be studied in a prospective power analysis. Therefore the choice of the research questions is essentially a form of risk-benefit reasoning, and has to be done with great care before the actual planning of the field study.

A prospective power analysis is an important instrument in the preparation of an intended field study. It is a prescribed part of the EFSA ERA guidance (EFSA 2010). EFSA asks for a power analysis of the difference tests, whereas in this document we argue that for non-target ERA equivalence testing is more relevant than difference testing. In the AMIGA project we developed software for the power analysis of both difference and equivalence tests (see section 3.2), which allows to compare the two types of power analysis. A power analysis for a difference test requires to specify an effect size, which according to EFSA (2010) will usually be identical to the limit of concern. In our experience with the new AMIGA Power Analysis tool, a power analysis for an equivalence test where the effect size is set to 0 will often give similar results as the power analysis for a difference test with the effect size equal to the limit of concern. Therefore, it does not seem very important which type of power analysis is performed, but in this document we will focus on equivalence tests.

To perform a prospective power analysis the following should be known (at least in the form of prior estimates):

1. the list of endpoints and a suitable hierarchy to order them
2. a specification of the anticipated statistical distribution for each endpoint
3. prior estimates of the mean and variation of the endpoints
4. limits of concern for the endpoints
5. a significance level for the tests (typically $\alpha = 0.05$)
6. a specification of the design structure of the intended experiment (e.g. randomised blocks) including other varieties and or agricultural treatments in the experiment
7. the intended method of data analysis

In the next sections we consider some of these points in more detail.

2.2 Research questions and a hierarchy of endpoints

In the design of experiments it is essential to have a clear description of the research questions and of the proposed methodology to answer these questions. In a field study of non-target effects the main research question is whether the genetically modified (GM) crop is substantially equivalent to the comparator variety (CMP) with respect to the non-target fauna in the agro-ecosystem. For an operational procedure it is needed then to specify a list of endpoints that will be measured in the experiment. Here, ‘endpoint’ can be understood at

several levels. For example, the endpoint 'Carabidae' may refer to the total of pitfall trap catches per plot over the field season in an intended single-environment experiment, but it may also refer to the catch per plot at one specific sampling time in spring (a more refined level) or the average catch per plot over multiple environments (a more integrated level). In general, it will be possible to arrange all these possible levels hierarchically, as shown for a simplified example in Figure 1.

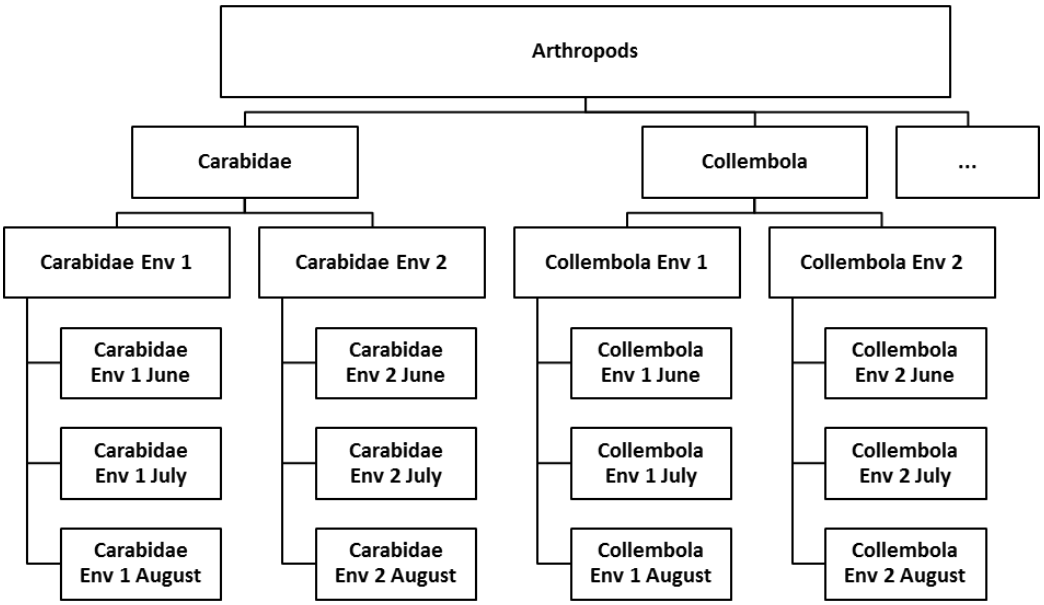

Figure 1. Simplified example of a hierarchy of endpoints in which different endpoints are sampled in different environment at different points in time during the season.

'Environment' here can denote another site or another year, or both. The tree also shows further integration of endpoints into a larger 'Arthropods' category. Risk assessors should establish at which level they will pose their research question. For example,

- is there a potential concern if the GM crop would affect the carabids in August in one specific environment, or
- is it sufficient to consider the year total of carabids for this environment, or
- is it sufficient to consider the average carabid counts over multiple environments, or
- can the research question be framed in terms of counts for functional groups, like predators and herbivores, or even all arthropods?

In the data analysis we can distinguish three parts:

1. pre-processing of the data, e.g. logarithmic transformations, but also integration steps such as summing pitfall trap catches over all time points in the field season;
2. the intended method of statistical analysis (SA) to estimate effects from the data, as will be further discussed in chapters 4 and 5;
3. the intended method of equivalence analysis (EA) to integrate estimated effects to higher levels in the hierarchy.

Figure 2 gives two simple examples, A and B, of the structure of an intended data-analysis for a single field experiment. In hierarchy A it is assumed that the measured endpoints are restricted to ten arthropod taxa, and that data will be collected at seven time points during the field season. A possible choice may be not to study the endpoints at the time-point level, but only at the level of the season total counts. This is especially practical for rare taxa. Summing is indicated by the 'Sum' pre-processing step; the underlining of 'timepoint' is meant to indicate that in this step some kind of summary over time-points is made. After this, the data will be analysed in a statistical analysis (SA step) to provide estimates and confidence intervals for the ten effects. The statistical analysis would normally involve an ANOVA type of analysis. Underlining of 'Data' indicates that Data are summarized to give effects. After this step each of the effects can be judged for equivalence. In the final step, denoted by EA_{all}, the equivalence for the individual taxa are combined in an overall NTO equivalence. EA_{all} denotes that each individual endpoint should meet its equivalence criterion.

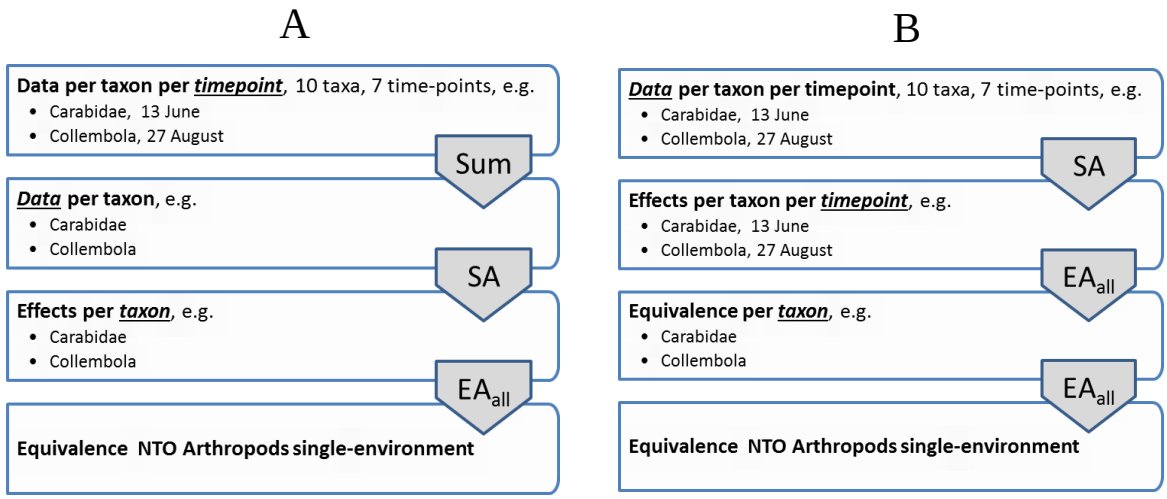

Figure 2. Two simple examples of alternative logical trees for analysis of Arthropod count data in a single-environment NTO study of a GM maize. Sum = summation of data. SA = statistical analysis. EA_{all} = equivalence analysis requiring all concern quotients to be within limits.

Hierarchy B in Figure 2 presents an alternative. Here the data are analysed at the time-point level, and the effects at all time-points are required to fulfil equivalence criteria. Note that the expected counts in the statistical analysis will be much lower as compared to hierarchy A, and therefore it will be more difficult to have sufficient power for all 10 x 7 = 70 endpoints. In fact, scheme B may not be practical at all, when it is expected that some species are not present at all (expected counts zero) during parts of the field season. In principle, the scheme could be adapted by specifying for each taxon the relevant time intervals during the season. The first EA_{all} step in hierarchy B could be replaced by an EA_{av} step implying that each taxon should on average meet the equivalence limits during the growing season. This is much less strict than an EA_{all} equivalence analysis.

Power analysis is related to estimating effects from data. Therefore the logical trees for analysis are relevant to see at which level the endpoints have to be ordered when performing a prospective power analysis. E.g. in hierarchies A and B the effects per taxon (summed over time-points) are relevant; in hierarchy B the effects per taxon per time-point could also be of

interest. If this is the case, then the power analysis should be performed at the level of the time-point.

The key message of this example is that alternative logical hierarchies for the analysis are possible, and that these choices can have a big impact on the number of required replications and thus on the cost-benefit reasoning relevant for the planning of field studies. Hierarchy B for example will require more replications than hierarchy A because it is required that equivalence is met for every time-point rather than for the sum across time-points.

Further details of the data analysis methods are discussed in relation to the statistical analysis protocol (Chapters 4 and 5).

2.3 Limits of concern

EFSA (2010), in the general section on problem formulation, states the following regarding limits of concern (LoC):

> Finally, for each measurement endpoint, the level of environmental protection to be preserved is expressed through the setting of 'limits of concern' which may take one of two forms. For studies in the environment(s) that are controlled [...] the limits of concern will usually be trigger values which, if exceeded, will either lead to conclusions on risks or the need for further assessment in receiving environment(s). For field studies, the limits of concern will reflect more directly the minimum effect that is considered to potentially lead to harm [...]. If these limits are exceeded, then detailed quantitative modelling of exposure may be required to scale up effects at the field level both temporally and spatially. Limits of concern can be defined by e.g. literature data, modelling, existing knowledge and policy goals.

It is not entirely clear which distinction is meant between the use of LoC in controlled studies (semi-field trials, e.g. using cages in the field) and in field studies. In both cases LoC is functioning as a trigger value for further attention. The word 'potentially' makes clear that exceeding the LoC does not necessarily indicate a harm. In fact, the EFSA Guidance is not fully consistent, because in other places (e.g. Glossary, p. 111) LoCs are defined as 'the minimum ecological effects that are deemed biologically relevant and that are deemed of sufficient magnitude to cause harm'. Therefore we should distinguish:

1. Limits of Concern as system-based and context-dependent concepts, indicating limits of harm to the environment: these are useful concepts, but will require further modelling to quantify, and
2. Limits of Concern as pragmatic trigger values for further assessment after field trial data analysis: these are used for equivalence tests, exceeding LoC is not necessarily indicating harm to the environment.

In this report LoC is used in the second meaning.

It is not easy to set values for the LoCs. These values will preferably be based on ecological expertise and, according to EFSA (2010), 'can be defined by e.g. literature data, modelling, existing knowledge and policy goals'. In the AMIGA project LoCs were tentatively set to 0.5 (i.e. 50% decrease) and 2 (i.e. 100% increase) for continuous nonnegative data, and for counts at abundance levels of 10 or higher. These values are provisional, and open for discussion.

Therefore all results which depend on these LoCs (such as all equivalence test results) should be seen as the results of a scenario study using these provisional LoCs. Alternative scenarios can be considered if other appropriate LoC values would be proposed. Note that LoCs should in principle be defined separately for each endpoint; therefore the choice of the same LoCs for all endpoints in this report should not be read as a general suggestion.

A main problem with count data is the increasing variability at low abundances, see for example Figure 1 in van der Voet and Goedhart (2015) showing the relation between the coefficient of variation (CV) and the expected abundance. This has often led researchers to omit low abundance data from their analyses (e.g. Prasifka et al. 2008). Here we suggest another approach, based on the idea that observing more variable effects at low abundances is just a consequence of the statistical properties, and therefore will and should not be seen by ecologists as reasons of concern. Thus, a flexible system of assigning Limits of Concern for taxa with low abundance may be used to reflect the biological ranges of no concern. The system we propose employs a threshold abundance value μ_0 below which scaling of the LoCs is applied. For observed means *m* below the threshold μ_0 the logarithms of the limits of concern are scaled by a factor $\sqrt{\mu_0/m}$,where *m* is the average of the counts for the GMO and the CMP. This implies that boundaries for equivalence testing become wider for lower abundances, corresponding to less concern at these low levels. For example, with LoCs at 0.5 and 2 and μ_0 = 10, the adapted LoCs are 0.38 and 2.7 for taxa with a mean abundance of 5 per plot, and 0.11 and 9.0 for taxa with an abundance of 1 per plot. Figure 3 displays the upper limit of concern on the original ratio scale for two cut-off values 10 and 5. Also see Figure 14 for how this would work for real data.

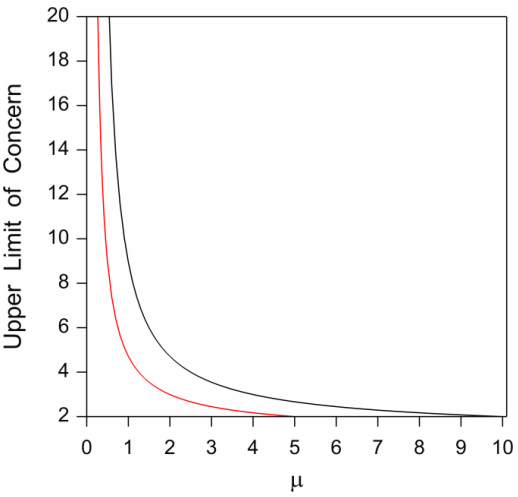

Figure 3. Upper Limit of Concern on the original ratio scale as a function of μ employing scaling on the log scale with a factor $\sqrt{(10/\mu)}$ for $\mu < 10$ (black line) and a factor $\sqrt{(5/\mu)}$ for $\mu < 5$ (red line).

The use of $\sqrt{1/m}$ in the scaling factor for the logarithms of the LoCs can be motivated in the following way. Suppose we have two samples each of size *n* from a Poisson distribution with means μ_1 and μ_2 respectively. The maximum likelihood estimator for the log-ratio $\Delta = \log(\mu_1/\mu_2)$ is given by $\log(X_m/Y_m)$ in which X_m and Y_m are the respective sample means. Suppose for simplicity that $\mu_1 = \mu_2 = \mu$. The large sample variance of this estimator then equals $2/(n\mu)$. Consequently the asymptotic standard error on the log-ratio scale is

proportional to $\sqrt{1/\mu}$ and the length of the confidence interval is thus also proportional to $\sqrt{1/\mu}$. This implies that, on the log-ratio scale, the length of the confidence interval in case μ=1 is a factor $\sqrt{10}$ larger than for μ=10. It is then natural to use $\sqrt{1/m}$ as a scaling factor for the logarithm of the LoCs for means smaller than 10.

This is basically statistical large sample theory and therefore a small simulation study was carried out to see whether this approach is also useful for small samples. Suppose that for μ=10 and a certain sample size *n* we have derived symmetric limits of concern, on the log scale, such that approximately 98% of simulated intervals for the log-ratio Δ lie within these limits of concern. LoCs for other values of μ are then obtained by multiplication of these LoCs for μ=10 with a factor of $\sqrt{10/\mu}$. The simulated coverage for this new situation should then be similar to 98%. This was done for values of μ equal to 6, 4, 2, 1 and 0.5 and sample size *n* equal to 6, 10, 20, 40 and 80. The simulated coverages and the original LoCs on the log scale for μ=10 are given in Table 1. Note that the lower and upper LoC were taken to be symmetric, i.e. $\text{Log}(\text{LoC}_{\text{low}}) = -\text{Log}(\text{LoC}_{\text{upp}})$.

The results in Table 1 indicate that for large sample sizes, e.g. *n* =80, the proposed multiplication of the LoCs give similar coverage probabilities for all values of μ. However, for smaller sample sizes the coverage probabilities become smaller for smaller values of μ indicating that the proposed multiplication factor $\sqrt{\mu_0/m}$ for the logarithm of the LoCs might still be too small. Further research could indicate a more precise multiplication factor. However the proposal does work as a first approximation, is simple to apply and will therefore be used in the sequel.

Table 1. Percentage coverage for the log-ratio of two independent samples from the Poisson distribution with mean μ. See text for further explanation

%Coverage	***n* = 6**	***n* = 10**	***n* = 20**	***n* = 40**	***n* = 80**
$\text{Log}(\text{LoC}_{\text{upp}})$	**0.793**	**0.615**	**0.421**	**0.305**	**0.227**
μ = 10	97.3	97.7	97.0	97.4	98.9
μ = 6	97.3	97.0	96.6	97.9	99.1
μ = 4	95.5	97.3	96.7	98.0	98.6
μ = 2	93.8	95.1	96.5	96.8	98.8
μ = 1	88.4	93.8	95.0	96.7	98.7
μ = 0.5	80.8	89.0	92.8	95.2	98.5

2.4 Intended data analysis

Already at the planning stage of an experiment, or a series of experiments, it is required to specify the statistical analysis of the data that will be observed later. The statistical analysis should be guided by the hierarchy of endpoints defined earlier, e.g. should the statistical analysis take place for each taxon at each time-point, for each taxon for the sums over the field-season, for the sums over the taxa in functional groups, etc.?. Specification of the statistical analysis method iso needed to specify the correct information for the prospective

power analysis. The proposed methods of statistical analysis are further discussed in section 4.1.

3 Protocol and software for the statistical aspects of designing experimental studies of non-target effects

Attention is required before a field trial is performed to ensure that the experiment will be meaningful to answer research questions. We present relevant points from a statistical viewpoint as a checklist. This list is an update of the checklist given in Deliverable D9.2a.

3.1 Checklist

1. Describe the **questions** the experiment is meant to answer, in words.
2. Decide on the **site(s)** and the **year(s)/season(s)** for the experiment based on the questions to be answered and feasibility.
3. Decide on the **spatio-temporal and/or taxonomic integration levels** at which conclusions regarding the research questions have to be formulated.
4. Prepare the **list of endpoints**. This may typically be organised in a hierarchy of endpoints. At the lower levels endpoints may be open to the possibilities of observation in the field (e.g. it may be impossible to predict if individual arthropod taxa will or will not be present under the conditions of the experiment).
5. Construct a **logical tree for the analysis** of all observed endpoints, showing how *data* may be pre-processed (**data pre-processing steps**), how *effects* will be estimated from the data by statistical analysis (**statistical analysis steps**), and how conclusions on *equivalence* will follow from the collection of effects and the limits of concern (**equivalence analysis steps**). The branches of the trees may have equal or different schemes for the subtrees. See sections 6.2.1 and 6.3.1 for examples of logical trees for analysis. In general, many different trees will be possible; the chosen tree should therefore be motivated.
6. For count or fraction data, a typical way of **pre-processing** the data is to **sum** over primary levels, e.g. over individual time points to obtain year totals, or over individual taxa to obtain totals for functional groups.
7. Indicate the **nature of the statistical analysis steps** in the logical tree as being a statistical analysis (**SA**, where the effects are calculated at the same level as the data), a statistical hierarchical analysis (**SHA**, where the analysed data are at a lower level of integration than the estimated effects) or a statistical meta-analysis (**SMA**, where effect estimates of a previous analysis are integrated to a higher level). More guidance on SA is provided in section 5.2, more guidance on SHA and SMA in section 5.3.
8. Indicate the **nature of the equivalence analysis integration steps** in the logical tree as requiring equivalence conclusion to be valid for all members (**EA_{all}**) or as allowing members to compensate for each other by averaging concern quotients (**EA_{av}**). More guidance is given in section 5.5.
9. For each endpoint to be used in the statistical analysis classify the **measurement type**, e.g. non-negative continuous data, count data or fractions (percentage) data.
10. For each endpoint to be tested formulate the **Limits of Concern (LOCs)**. For each endpoint one lower and/or one upper LOCs can be set. For non-negative continuous and count data these will typically be ratios of GMO divided by CMP true values. The LOCs define the **null hypotheses for equivalence tests**.
11. For endpoints which are counts, decide **how to address low abundance data**. Either define a criterion to omit uninformative low counts (e.g. abundance below 5, or CV larger than 100%), or adopt a rule that LOCs will be set as a function of the expected or observed abundance in the field. A proposed scaling factor for the log(LOC) expressed as a factor is $\sqrt{\mu_0/m}$, where m is the expected or observed abundance in the experiment, and μ_0 is a chosen threshold abundance below which concern limits will be widened (e.g. $\mu_0 = 10$).
12. Set the **significance levels (**α) for statistical testing. Conventionally the level (size) will be 0.05. In the TOST approach to equivalence testing (Schuirmann 1987) the significance level

of a two-sided confidence interval is twice the significance level for the equivalence test, therefore employing a two-sided (1-2α) confidence interval corresponds with (1-2α) confidence equivalence tests.

13. Set the **required power** of the equivalence tests to detect a zero differences. Under the null hypothesis the effect sizes will be equal to the LoCs. Conventional values for power are between 70 and 90%.
14. Describe the structure of the proposed **experimental design**, e.g. completely randomized, randomized blocks, split-plot, balanced incomplete blocks.
15. Describe the **experimental units** (typically plots or sub-plots), and give details of the **blocking structure** (e.g. 4 main plots per randomized block, each main plot split into 3 sub-plots) and the **treatment structure** (e.g. three types of spraying and four crop varieties). Also describe if interactions should be included.
16. Describe whether **repeated measurements** will be taken from the same experimental unit.
17. Provide a **model formula** specifying how the data will be analysed, using the syntax of one of the common software tools for statistical analysis (e.g. SAS, GenStat, R), for example *block/plot/subplot + treatment + variety*. Include terms and a correlation structure for repeated measurements if used. Indicate which factors are random rather than fixed.
18. For each primary endpoint provide **prior estimates of central value and variation** for a measurement on a single experimental unit. For non-negative continuous and count data the prior estimates for central values will typically be expected values or geometric means, and the prior estimates for variation will typically be coefficients of variation. Such values can be derived from previous experiments or based on expert knowledge.
19. For each endpoint specify the simplest **statistical analysis method** that will be used (the analysis method may need to be adapted if there are unexpected deviations in the execution of the field study or unexpected data). See the statistical analysis protocol for details.
20. Based on the prior estimates **estimate the power of the proposed design as a function of replication,** for the equivalence test at the chosen integration level(s), see point 3. For this the AMIGA Power Analysis software may be used (see 3.2).
21. From the power curves derive the **replication** of the comparison of GMO to CMP in the proposed design.
22. If the calculated minimal replication cannot be realized in practice, the **power is insufficient**. In such case adapt the design or reformulate the research questions.
23. **Randomise** the treatments over the experimental units taking proper account of the design.

3.2 AMIGA Power Analysis program and R scripts for data analysis

To guide the design of ERA field trials, specific methods for power analysis for statistical tests based on field trial count data have been developed in the AMIGA project, as described in this report (see also Goedhart et al. 2014, van der Voet and Goedhart 2015). This has resulted in publically available software for this purpose in the form of the AMIGA Power Analysis tool (Deliverable 9.5).

To perform power analysis for ERA field trials, a complete definition of the field trial and envisioned method of analysis is required, along with a data model for the comparator variety for each endpoint. So the tool effectively requires all information described by items 1-19 of the checklist of section 3.1, and systematically asks for this information using the following steps:

1. **Define the endpoints and their limits of concern:** A general issue for establishing sample sizes of ERA field trials is that statistical power can only be evaluated given specified relevant effect sizes. These relevant effect sizes should bound a range of GMO vs. comparator differences that are not considered to be of any biological

concern. Hence, it is considered of no importance when the statistical power for effect sizes within these so-called limits of concern (LoCs) is low. Without specification of LoCs by experts no power analysis can be performed. In AMIGA tentative limits of concern have been set for non-target arthropods and soil organisms at two-fold increases or decreases with respect to the comparator abundance provided this abundance is not too low.

2. **Define the endpoint data models:** The Power Analysis tool adopts a uniform way to describe the data model of the comparator variety in terms of a statistical distribution (e.g. Poisson or overdispersed Poisson), a mean and a CV.
3. **Specify the experimental setup:** All elements relevant for the analysis are also relevant for the power analysis. Therefore, a complete description of the experimental setup must be provided, including the experimental design, varieties additional to the GMO and CMP, additional agricultural treatment factors, and possible interactions between these additional factors and the GMO or comparator.
4. **Specify the method of analysis and power analysis:** Specify which statistical method(s) will be used for equivalence and difference testing and specify the range of replications for which the power should be computed. The power can be calculated by Monte Carlo simulation or by an approximate method (Lyles et al. 2007).

Having this information, the Power Analysis tool computes the power for different numbers of replications and various levels of difference between the LoCs (checklist item 19) for all specified endpoints individually, and combined using the concern quotient CQ (see Section 5.5). From the results, the required number of replications can be derived (checklist item 20, 21). An additional feature of the tool is that it can produce a data template and analysis script (written in the programming language R) that can be used directly for analysis of the specified field trial.

For the purpose of the experimental design of ERA NTO field trials the methods and software can be used in two scenarios. The first scenario is a single-environment field trial (i.e. a field trial in single location/year) in which the aim is to determine the number of replicates needed to obtain sufficient statistical power. The second scenario is a multi-environment field trial in which the aim is to determine the number of environments (i.e. location/year combinations) in order to obtain sufficient statistical power. A case study based on historical data provided by Prasifka et al. (2008) has been performed and is reported here. More details will be conveyed later (Kruisselbrink et al, in prep.).

With regard to the risk management question, it is not always clear if the ERA should lead to a general conclusion for Europe, to specific conclusions for each biogeographical zone, or to specific conclusions for each field site. Obviously, the requirements with respect to the overall needed number of replications may depend on the question at hand.

4 Motivation for a protocol for statistical equivalence analysis of experimental studies of non-target effects

4.1 Methods of statistical analysis

In field studies for environmental risk assessment of GMOs typically counts of various taxa are observed, sometimes supplemented with continuous non-negative data and/or percentage data. Observed counts are generally log transformed by entomologists, typically after the addition of one to avoid taking the logarithm of zero, to achieve homogeneity of variance after which statistical methods based on the normal distribution, such as analysis of variance, are used. Alternatively the squared root transform of counts is taken. In other fields of ecological research counts are nowadays statistically analysed by log linear models which rely on distributions specific for count data such as the Poisson, the overdispersed Poisson and the negative binomial distribution (McCullagh & Nelder 1989). Log-linear models for ecological count data have been advocated for many years, see e.g. Sileshi (2006), Ver Hoef and Boveng (2007), O'Hara and Kotze (2010) and Szöcs and Schäfer (2015). In a simulation study Goedhart and van der Voet (2014) found that the transformation approach has good properties when it comes to difference testing but that confidence intervals for the true ratio of the mean of the GMO and the CMP have poor coverage probabilities. The coverage probability of the log-linear model employing the overdispersed Poisson distribution is generally satisfactory even when data are simulated according to other count distributions. They therefore recommend a statistical analysis according to the overdispersed Poisson model for count data when it comes to equivalence testing and this method of analysis will therefore be used in the sequel.

Continuous non-negative data are usually analysed by first applying a log-transformation and then doing an ANOVA type of analysis. Percentages data are commonly analysed by means of logistic regression which employs the (overdispersed) binomial distribution.

4.2 Adapted limits of concern for count data of non-abundant taxa

In section 2.3 a flexible system of assigning Limits of Concern for taxa with low abundance was proposed to reflect the biological ranges of no concern. Here we present more fully the proposed system including additional considerations for plotting and interpretation.

In the proposal Limits of Concern are based on a threshold abundance value below which scaling is applied, and special rules for very low abundances (zero counts for GMO or CMP):

1. Below a limit abundance value, e.g. μ_0=10, it is proposed to apply a scaling to the LoCs for taxa. The scaling factor is $\sqrt{\mu_0/m}$, to be applied to the logarithms of the LoCs, in which *m* is the combined mean of the GMO and CMP. This implies that boundaries become wider for lower abundances, corresponding to less concern at these low levels. For example, with basic LoCs at 0.5 and 2 and a threshold of μ_0=10 the adapted LoCs are 0.38 and 2.7 for taxa with an abundance of 5 per plot, and 0.11 and 9.0 for taxa with an abundance of 1 per plot.

2. If for either the GMO or CMP no single specimen is found (zero average), then the ratio is zero or infinite, which cannot be analysed on a logarithmic scale. It is proposed to re-calculate the ratio with the zero average replaced by the lowest possible value, which is one over the number of replications. This ratio (without a confidence interval) will only be displayed in case it falls outside the scaled LoCs.

In a graphical representation of the results a background colouring may be applied to the area within the (scaled) LoCs to indicate its meaning as equivalence area, i.e. the observed data do not indicate concern under the specified criteria. On the other hand, no background colouring is applied to the area outside the LoCs, because in the proposed system the LoCs only act as a trigger for further consideration, and values outside the LoCs do not indicate the presence of environmental harm.

4.3 Summarising over different dimensions

In the design phase of the experiment a hierarchical tree of endpoints has been prepared (section 2.2). The statistical analysis of equivalence should follow this tree by stepwise integration of each sub-collection of endpoints to its connecting node endpoint. In general there are three types of analysis:

1. data pre-processing
2. data analysis
3. combination of effects in an equivalence analysis

In a typical environmental risk assessment study which accounts for biogeographical variation count data are obtained for many taxa, in multiple years and from multiple locations (sites). However, the data selected for analysis may have different taxa for different space-time combinations. A general question is how data should be integrated to obtain an overall conclusion for the risk assessment (see Figure 4).

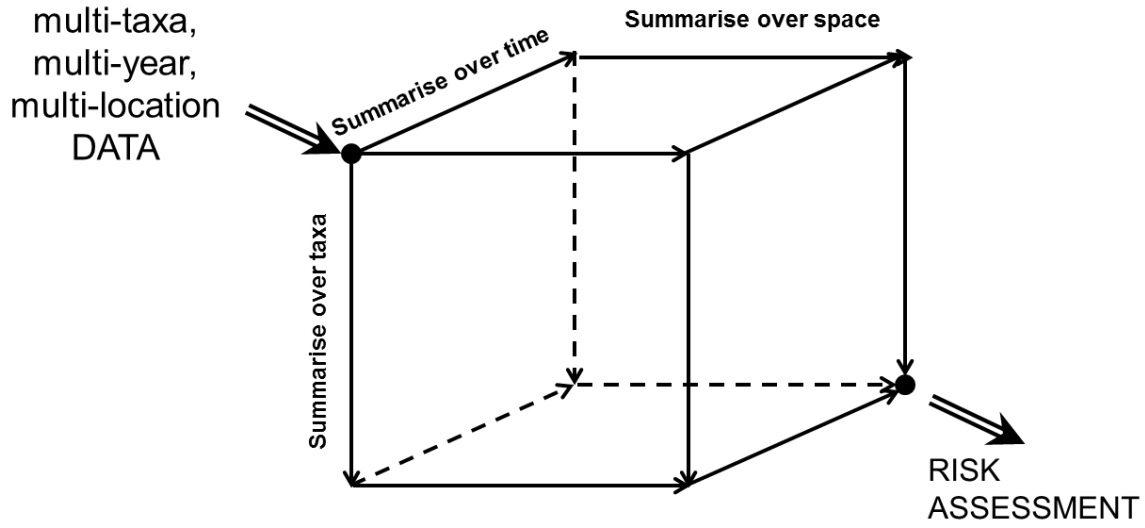

Figure 4. Possible routes for integration over endpoints.

As can be seen in Figure 4, there are six possible ways to summarise conclusions:

1. First summarise over time, then over space, then over taxa;
2. First summarise over time, then over taxa, then over space;
3. First summarise over space, then over time, then over taxa;
4. First summarise over space, then over taxa, then over time;
5. First summarise over taxa, then over time, then over space;
6. First summarise over taxa, then over space, then over time.

For each integration step, there are in principle several methods for summarising:

1. Pre-processing of the data, e.g. summing counts over taxa or summing over time for the same experimental unit;
2. Joint data analysis – this is applicable for summarising over time or over space, but only when the same taxa are addressed;
3. Multi-criteria decision analysis applied to effect estimates – this method is applicable for all forms of summarising.

We further distinguish between various forms of data analysis. In the hierarchies A and B given in Figure 2 the statistical analysis SA estimates Effects from Data without further integration: in hierarchy A Data per taxon were summarized to an Effect per taxon, while in hierarchy B Data per time-point were summarized to an Effect per time-point. In hierarchy B however the statistical analysis could also, in one go, summarize over time-points giving a single effect per taxon. Such an analysis will be termed a statistical hierarchical analysis, or SHA for short, because it estimates effects for a higher level in the hierarchy.

Individually estimated effects, e.g. for several taxa, can be combined in a single effect and a corresponding confidence interval by means of a statistical meta-analysis, or SMA for short. This provides an objective way of combining information from separate effects, while taking into account the different standard errors for the individual effects see e.g. Hardy and Wright (1996). There are basically two versions of meta-analysis. The "fixed" version assumes that estimated effects have a common mean and individually known variances. The overall effect is then simply the weighted average of the individual effects, in which the individual variances are used as weights. The "random" version on the other hand allows for heterogeneity of the individual effects by introducing a between individuals component of variance. The statistical technique REML can then be used to estimate the overall effect and to provide a confidence interval for the overall effect. We used the "random" version throughout this report and applied it to the estimated log-ratio effects. Note that a meta-analysis implicitly assumes that the individual effects are statistically independent. This might be a strong assumption when combining information for different taxa within the same experiment. Also note that a SMA implicitly assumes that negative effects, e.g. for a taxon, can be compensated by positive effects for another taxon.

In the equivalence analysis, the simplest option is to require that effects are within limits of concern for all endpoints (EA_{all}). Alternatively equivalence could be met on average and this is denoted by EA_{av}. Finally, the equivalence analysis (EA) can be performed both on the

estimated effects (point estimates), and on the confidence limit which gives rise to the most concern. The latter EA analysis can be termed a worst case analysis.

Table 2 summarizes the different possible steps in building a hierarchy for the analysis of observed data. Note that these steps also form a hierarchy in the sense that an element cannot be followed by an element which is in a class above the class of the current element.

Table 2. Elements of the hierarchy for data analysis and integration of equivalence

Element	*Explanation*
Data pre-processing	
SUM	Summing the data. For example summing counts of a taxon over different points in time, or summing counts of taxa within the same functional group to give a single count for the functional group
Statistical analysis	
SA	Statistical Analysis of data resulting in estimated effects at the same level of the hierarchy. i.e. without integration of other levels in the hierarchy. For example estimation of the effect for a single taxon per time-point.
SHA	Statistical Hierarchical Analysis of data resulting in estimated effects at a higher level of the hierarchy. i.e. including integration of other levels in the hierarchy. For example estimation of the effect for a single taxon summarized over time-points.
SMA	Statistical Meta-Analysis which combines individual effects into a single combined effect. For example combining effects for taxa within the same functional group to give a single effect for the functional group, or combining effect for individual environments to give a single effect across environments
Equivalence analysis (multi-criteria decision analysis)	
EA_{all}	Equivalence analysis of estimated effects in which all estimated effects should meet the equivalence criterion. This step can be present several time, for example when moving from 1) equivalence per functional group per year per site to 2) equivalence per year per site to 3) equivalence per site to 4) overall equivalence
EA_{av}	Equivalence analysis of estimated effects in which the average of estimated effects should meet the equivalence criterion.

Note that summarising by means of SHA, SMA or EA_{av} implies that we are interested in an average effect. In contrast, EA_{all} considers all individual effects on their own.

4.3.1 Multi-criteria decision analysis

A statistical analyses result in estimated effects, i.e. differences between GMO and CMP at an appropriate scale (often the log scale). These effects, and their confidence limits, can be standardized by scaling to a no-concern yardstick which represents a minimum limit of potential biological relevance, i.e. the Limit of Concern (LoC). For count data, if Q is the estimated ratio for GMO vs. CMP, and if lower and upper Limits of Concern are also expressed as ratios LoC_{low} and LoC_{upp} (which are assumed to be respectively below 1 and

above 1, e.g. 0.5 and 2), then for further integration ***concern quotients*** *CQ* can be calculated, which are non-negative scores that express absence of concern for values up to 1:

$$\text{Two-sided:} \quad CQ = \max\left[\frac{\log(Q)}{\log(LoC_{low})}, \frac{\log(Q)}{\log(LoC_{upp})}\right]$$

$$\text{One-sided left:} \quad CQ = \max\left[\frac{\log(Q)}{\log(LoC_{low})}, \quad 0\right]$$

$$\text{One-sided right:} \quad CQ = \max\left[0, \frac{\log(Q)}{\log(LoC_{upp})}\right]$$

For integration over time, space and/or endpoints the *CQ* values can be used as input in a multi-criteria decision analysis (MCDA) model.

In the simple approach for equivalence analysis (EA) that is proposed in this report, and for which examples are given in Chapter 6, *CQ* values up to 1 are considered acceptable, and values above 1 are a trigger for further investigation. Two options may be considered for combining effects of multiple indicators:

1. EA_{all} by taking the maximum, i.e. no compromising, the worst case defines the overall concern, or
2. EA_{av} by taking the average, i.e. bad scores for one indicator can be compensated by good scores for another.

The proposed MCDA approach is a special and simple case of a more flexible MCDA method, the Balance Of Acceptability model (van der Voet et al. 2014). This model allows specification of weights to be applied in a weighted rather than unweighted average. This is open for further refinement, e.g. weighting could be related to perceived ecological relevance, or statistical precision. Further it is possible to specify a smooth transition between given acceptable and unacceptable *CQ* values, and intermediate approaches between EA_{all} and EA_{av} by specification of a compensability parameter.

4.4 Confidence intervals vs. tests, graphical summaries

Often the final aim of an NTO study is framed as testing hypotheses about unintended differences. This is then presented as a list of test results such as *p* values resulting from difference tests. However, this way of presentation obscures the magnitude of the observed differences, the precision of these estimates and the criteria (limits of concern) against which the differences should be interpreted. It is much more insightful to present the results as confidence intervals for the true effects, together with the limits of concern.

Confidence intervals for effects and limits of concern can be displayed for multiple endpoints together in a single graph. This works fine for effects that are at the same scale, for example the ratio scale which compares the GMO mean to the CMP mean. A more general way of plotting allows a simultaneous display of endpoints measured at different scales, and this may also be useful when limits of concern are very diverse among endpoints. For this, the effect estimates can be scaled. The scaled dimensionless measure is called the ***LoC Scaled***

Difference (*LoCSDIF*). Depending on the LoCs scaling may be the same or different for negative and positive deviations, but for one-sided problems, i.e. when there is only one LoC, the same scaling factor is used for both positive and negative deviations. For count data, if Q is the estimated ratio for GMO vs. CMP, and if lower and upper Limits of Concern are also expressed as ratios LoC_{low} and LoC_{upp} (which are assumed to be respectively below 1 and above 1, e.g. 0.5 and 2), then the *LoCSDIF* is defined as follows

$$\text{Two-sided:} \quad LoCSDIF = \begin{cases} \dfrac{\log(Q)}{-\log(LoC_{low})} & \text{if } Q < 1 \\ \dfrac{\log(Q)}{\log(LoC_{upp})} & \text{if } Q \geq 1 \end{cases}$$

$$\text{One-sided left:} \quad LoCSDIF = \frac{\log(Q)}{-\log(LoC_{low})}$$

$$\text{One-sided right:} \quad LoCSDIF = \frac{\log(Q)}{\log(LoC_{upp})}$$

An artificial example of plots on the *Q* scale and the *LoCSDIF* scale is shown in Figure 5 with unequal limits of concern for three taxa. In the upper plot the Aaidae and Beidae taxa are seen to be significantly different from zero because their intervals do not overlap with the equality line at 1. But the four-fold decrease for Beidae is not considered a concern, whereas the two-fold increase for Aaidae is a concern. In a similar way the larger (three-fold) increase for Ceidae is not considered a concern. The ordering of concerns is easier seen in the lower plot for the LoC scaled differences. Note that for Ceidae scaling on the right is done with the upper LoC which is 16, while scaling on the left employs the lower LoC which is 0.5.

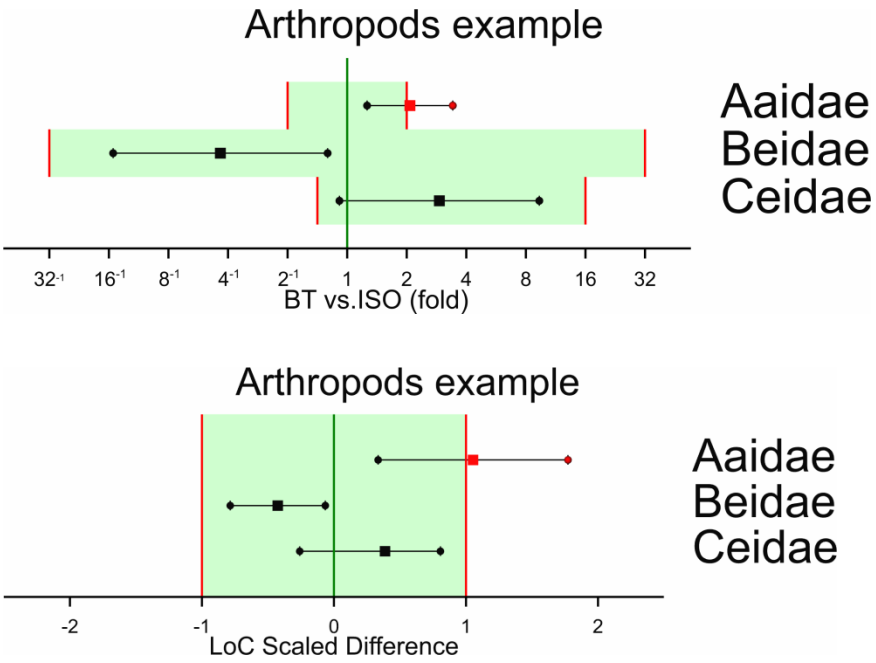

Figure 5. Artificial example of graphical representation of comparative analysis results. Point estimates of the expected ratio of test variety (here BT maize) and comparator variety (here isogenic maize) with 90% confidence limits, and shown in relation to provisional limits of concern.

Real examples of plots showing both types of graphical representation are shown in chapter 6.

5 Protocol for statistical equivalence analysis of non-target effects

In this chapter we present a protocol for the statistical analysis of data from ERA field trials. In principle, the methods of statistical analysis have already been decided at the time of planning the experiment, but it may be needed to update the methods based on the context or unexpected findings.

5.1 General

1. Check and if necessary update the **list of endpoints** that was established in the design phase. Motivate any change.
2. Check and if necessary update the **logical tree for the analysis** of all observed endpoints. Motivate any change.
   The logical tree for analysis shows how *data* may be pre-processed (**data pre-processing steps**), how *effects* will be estimated from the data by statistical analysis (**statistical analysis steps**), and how conclusions on *equivalence* will follow from the set of all estimated effects and the limits of concern (**equivalence analysis steps**). The branches of the trees may have equal or different schemes for the subtrees. See sections 6.2.1 and 6.3.1 for examples of logical trees for analysis. In general, many different trees will be possible; therefore the chosen tree should be motivated.
   a. For count or fraction data, a typical way of **pre-processing** the data is to **sum** over primary levels, e.g. over individual time points to obtain year totals, or over individual taxa to obtain totals for functional groups.
   b. Indicate the **nature of the statistical analysis steps** in the logical tree as being a statistical analysis (**SA**, where the effects are calculated at the same level as the data), a statistical hierarchical analysis (**SHA**, where the analysed data are at a lower level of integration than the estimated effects) or a statistical meta-analysis (**SMA**, where effect estimates of a previous analysis are integrated to a higher level). More guidance on SA is provided in section 5.2, more guidance on SHA and SMA in section 5.3.
   c. Indicate the **nature of the equivalence analysis integration steps** in the logical tree as requiring equivalence conclusion to be valid for all members (**EA_{all}**) or as allowing members to compensate for each other by averaging of concern quotients (**EA_{av}**). More guidance is given in section 5.5.
3. Graphical summaries of results are to be prepared for estimated effects (section 5.4) and, if deemed useful, for LoC-scaled differences (section 5.5).

5.2 Statistical analysis of single endpoints

The basic approach is to calculate estimates and 90% confidence intervals for effects (GMO vs. CMP differences, expressed on an appropriate scale), and then compare these to the (possibly provisional) limits of concern which were specified during the design of the experiment (see Figure 6 for a flowchart for the statistical analysis of a single endpoint).

1. The method of statistical analysis depends on the type of endpoint. For continuous endpoints with necessarily positive values it is recommended to perform an analysis on the log transformed data. For discrete endpoints such as count data and fraction data it is recommended to perform an analysis on the original scale using an appropriate statistical distribution and link function.

2. Analyse the transformed data by linear models: ANOVA if the design is balanced, or by a mixed model (REML) if they are not.
3. Analyse the untransformed data by generalized linear models (GLM), or by a generalized linear mixed models (GLMM) in case there are additional random effects in the model. Allow for over-dispersion in counts and fractions.
4. Check whether statistical assumptions are reasonable, e.g. as follows:
    a. Outliers: check data points with large standardised residuals. Compare analyses with and without such data points in a sensitivity analysis.
    b. A QQ plot of the residuals should show approximately a straight line
    c. A plot of residuals vs. fitted values can be used to check if there is heteroscedasticity.
5. If statistical assumptions are unreasonable, then an ad-hoc strategy will have to be followed. For example, another variance function might be more appropriate or non-parametric tests may be used. This protocol further assumes that the model fits sufficiently well.
6. Extract the estimated difference between the GMO and CMP from the statistical model, e.g. the log-ratio for count data, along with the standard error of the estimate. Employs these to calculate a two-sided 90% interval taking account of the degrees of freedom for residual. Display the confidence interval in a graph along with the LoCs. Note: for visual display it is recommended to calculate and display both limits, even if there is only one LoC, for either a decrease or an increase.

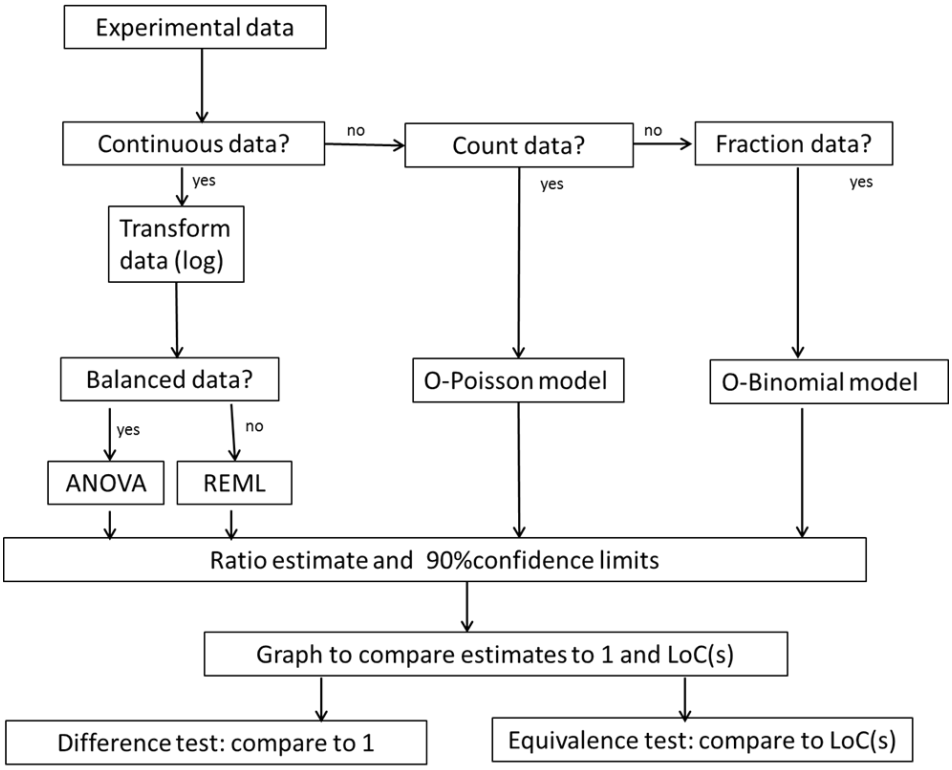

Figure 6. Flow chart for the statistical analysis of single endpoints. ANOVA = analysis of variance, REML = residual maximum likelihood. O-Poisson = over-dispersed Poisson model. O-Binomial = over-dispersed binomial model. LoC = Limit of Concern.

5.3 Statistical analysis integrating multiple endpoints

1. Integration over multiple endpoints may be automatically performed in a **statistical hierarchical analysis** (SHA) model as described in section 5.2. Perform a **statistical meta-analysis** (SMA) if described in the logical tree for analysis. For this, consider the estimated

effects with their standard errors (at an appropriate scale, e.g. the log scale) as input for the meta-analysis. Consider the level over which an average is taken as a random factor.
2. From the output, construct an **estimate and a 90% confidence interval for the overall effect**.
3. The use of SHA or SMA is only logical if limits of concern are defined for the integrated output or if they are equal for all individual endpoints.

5.4 Graphical representation of effects

1. For each endpoint, **plot point estimates and 90% confidence intervals of estimated effects**, together with lines for the equality ratio 1, and the LoCs. In most cases plots on a logarithmic scale are advised. The 90% limits of the interval represent a 5% significance level for equivalence testing in a two one-sided tests (TOST) approach.
2. Prepare **one or more graphs**, depending on the number of endpoints, and possible groupings in the hierarchy which are of interest as specified in the logical tree for analysis.
3. Compare the intervals to the LoCs to obtain **conclusions regarding equivalence** of the GMO and the CMP.
4. If of interest, compare the intervals to zero to obtain conclusions regarding the **statistical significance of the difference** between the GMO and the comparator. Note that this implicitly employs a significance level of 10% for a two-sided difference test.
5. Optionally, **confidence intervals can be displayed on the LoC Scaled Difference (LoCsDIF) scale**. This possibly allows an easier comparison in case (scaled) limits of concern are not the same for various endpoints.

5.5 Integration of endpoints, overall equivalence analysis

1. For each estimated effect and its corresponding limit(s) of Concern (LoC) calculate the **concern quotient** (*CQ*):

$$\text{Two-sided:} \quad CQ = \max\left[\frac{\log(Q)}{\log(LoC_{low})}, \frac{\log(Q)}{\log(LoC_{upp})}\right]$$

$$\text{One-sided left:} \quad CQ = \max\left[\frac{\log(Q)}{\log(LoC_{low})}, \quad 0\right]$$

$$\text{One-sided right:} \quad CQ = \max\left[0, \frac{\log(Q)}{\log(LoC_{upp})}\right]$$

   Use the same formulae to convert the confidence limits for log(Q) to the *CQ* scale, using the lower confidence limit Q_{low} in combination with LoC_{low}, and the upper confidence limit Q_{upp} in combination with LoC_{upp}.
2. For each **equivalence analysis (EA) step in the logical tree for the analysis**, check whether the intended equivalence criterion is that 1) all member endpoints should comply to $CQ \leq 1$ (EA_{all}) or 2) *CQ*s can be averaged (EA_{av}).
3. For each **EA_{all} step**, integrate over individual endpoints *i* by $CQ = \max_i(CQ_i)$ and $CQ_{upp} = \max_i(CQ_{upp,i})$.
4. For each **EA_{av} step**, integrate over individual endpoints *i* by $CQ = \text{mean}(CQ_i)$ and $CQ_{upp} = \text{mean}(CQ_{upp,i})$.

6 Statistical analysis examples

6.1 Power analysis arthropods based on historical data

6.1.1 Power analysis case study

For this case study, we consider a hypothetical new ERA field trial designed to test a new variety against a comparator. We consider the test variety to be a Bt crop, for which it is feasible to apply a no-spraying treatment regime, which is not feasible for the comparator variety. The field trial is setup to compare the Bt and comparator varieties, but it also includes two agricultural treatment regimens (default spraying and no spraying). It is anticipated that no spraying leads to undesirable effects for the comparator variety. For this new field trial we will focus on two particular scenarios for power analysis:

Scenario 1: power analysis for a new experiment on a new location/year: In this scenario a hypothetical new experiment for a single future location/year combination is to be designed. For this new location/year, we consider a completely randomized design and the aim is to determine the number of replicates necessary to detect "critical" changes in abundance of the NTOs of interest.

Scenario 2: power analysis for choosing the number of environments for a multi-environment setting: In this scenario, we assume that there is a fixed design for each location/year combination (or environment), which is a completely randomized design with a replication size of four. Given this fixed design, the aim is to determine the number of environments that are required to detect "critical" changes in abundance of the NTOs of interest.

The background data used for this case study are the abundance data of the NTOs for the comparator variety of Prasifka et al. (2008). This is a study on field corn over five locations and three years per location, with different numbers of replications per trial. Counts were recorded for over 100 taxa using pitfall traps, sticky cards, and visual counting as sampling methods. The taxa were mainly single species, but also some species groupings and miscellaneous categories are present. All sampling methods were done for multiple points in time for each plot, and also multiple samples were taken per plot. Both the number of points in time as well as the number of samples per plot varied per location and year. The case study is restricted to the abundance data of the pitfall traps and focuses on 15 selected NTOs, with average counts larger than 10, and use these data to obtain a priori data models for the endpoints within the power analysis. Table 3 summarizes the design configurations for the different sites and years of field trial data used by Prasifka et al. (2008).

Table 3. Experimental setup of each location/year combination used in the field trials presented by Prasifka et al. 2008.

Location	Year	Replicates	Pitfall traps		
			Periods	Traps	Days
Maryland	2000	3	8	10	7
Maryland	2001	3	11	8	7
Maryland	2002	3	6	8	7
Nebraska	2001	2	10	4	1
Nebraska	2002	2	10	4	1
Nebraska	2003	2	4	10	1
Iowa-1	2001	3	10	10	7
Iowa-1	2002	3	10	10	7
Iowa-1	2003	4	10	11	1
Iowa-2	2001	2	10	4	1
Iowa-2	2002	2	10	4	1
Iowa-2	2003	2	4	10	1
Illinois	2000	4	4	4	3
Illinois	2001	4	4	4	3
Illinois	2002	4	4	4	3

6.1.2 Methods

6.1.2.1 Define the endpoints and limits of concern

As a pre-processing step, the counts (recorded over multiple time periods and multiple pitfall traps per plot) are aggregated by taking total counts. Hence, the total abundance pitfall trap counts of the 15 selected NTOs form the endpoints for the scenarios of the case study.

For the purpose of the present study, limits of concern are considered equal for all endpoints, and are pragmatically set to levels of -50% decrease and +100% increase in abundance which corresponds with the two-fold differences considered by Perry et al. (2003). This implies LoCs of 0.5 and 2 for the ratio between the mean of the test variety (μ_T) and the mean of the comparator variety (μ_C), employing the notation used by the Power Analysis tool.

6.1.2.2 Define the endpoint data models

The overdispersed Poisson distribution is chosen as the distribution type for all endpoints in this case study. A priori estimates of the endpoint means and CVs, along with CVs that describe the between location/year variation, are derived from the data using the following GLMM per endpoint:

Response variable:	Total count y_i of endpoint *i*
Probability distribution:	Poisson-LogNormal with $y_i \sim \text{Poisson}(\lambda_i)$ and $\lambda_i \sim \text{LogNormal}(\mu_i, \sigma_i^2)$
Link function:	Log
Random component:	Site x Year

In addition an offset is used to account for differences in effort per plot. From such models, the estimated mean of endpoint *i* is then derived as:

$$\text{Mean}_i = \exp(\mu_i + \sigma_i^2/2),$$

and, using $\mathbb{V}\text{ar}(y_i) = \tau^2\lambda_i$, the within location/year CV is derived as:

$$CV_{\text{within},i} = \sqrt{\tau^2/\lambda_i}\,.$$

The variation between location/year can be obtained by means of:

$$CV_{\text{between},i} = \sqrt{\exp(\sigma^2) - 1}\,.$$

Applying this model for each endpoint, using the sampling periods and samples per plot as specified in Table 3, yields a successful fit for the selected 15 endpoints. Table 4 presents the selected endpoints along with the extracted mean, CV and between site/year CV.

Table 4. The extracted means and CVs for the selected endpoints for the hypothetical new field trials.

Endpoint	Mean	CV (%)	CV between site/year (%)
Click beetles	10.5	83.2	197.3
Millipedes	15.0	112.4	146.9
Aphids	15.9	61.3	114.5
Springtails globular	16.2	96.9	133.8
Antlike flower	16.3	63.7	306.1
Ground beetle larvae	17.4	122.5	362.9
Sowbugs	28.1	144.3	133.7
Predaceous mites	30.1	48.9	241.5
Oribatid mites	44.6	128.5	90.6
Japanese beetles	47.5	87.8	112.4
Springtails hypogastrurids	96.0	94.7	514.0
Ants	131.0	108.2	66.8
Springtails isotomids	132.8	44.8	472.8
General collembola	706.7	51.3	121.1
Springtails entomobryids	710.4	51.2	152.9

6.1.2.3 Specify the experimental setup

In the case study there is one additional treatment (spraying) with two levels (spraying and no spraying). Given that the test variety is a Bt crop, it is reasonable to assume that the test variety does not respond to spraying. However, it is assumed that the comparator will respond to spraying. The contrast of interest is therefore between the mean of the Bt crop with and

without spraying, and the comparator with spraying. Table 5 summarizes the experimental structure of the two scenarios.

The case study assumes that the design within each site/year is completely randomized. However, for scenario 2, each site/year should be seen as a block with respect to the overall (power) analysis. Hence, in the context of the power analysis, scenario 2 is a randomized complete block design, for which the between location/year variation derived in the previous section can be used per endpoint.

Table 5. Summary of the experimental structure of the scenarios of the case study. The basic field trial consists of four treatment levels: test with default spraying, test without spraying, comparator with defaults.

			Replicates per location/year		**Location/year replications**	
Variety	**Spraying**	**Comparison**	**Scenario 1**	**Scenario 2**	**Scenario 1**	**Scenario 2**
Test	Default	Test-variety	*To be assessed*	4	1	*To be assessed*
Test	None	Test-variety	*To be assessed*	4	1	*To be assessed*
Comparator	Default	Comparator	*To be assessed*	4	1	*To be assessed*
Comparator	None	Exclude	*To be assessed*	4	1	*To be assessed*

6.1.2.4 Specify the method of analysis and power analysis

Following the recommendations of Goedhart et al. (2014), the log-normal model is used for difference tests and the overdispersed Poisson model is used for equivalence testing. The significance level is set to the standard level of 0.05 and the desired power is 0.8. The replication sizes of interest are different for the two scenarios and are chosen such that the required total number of plots to conduct the experiments are aligned, being 32, 64, 96, 128, 160 and 192 plots. For scenario 1, in which one replicate consists of four plots, this yields replication sizes 8, 16, 24, 32, 40 and 48. For scenario 2, in which on replicate consists of 16 plots, this yields the replication sizes 2, 4, 6, 8, 10 and 12. The power is approximated by means of Monte Carlo simulation using 100 simulations per effect/replication level. Table 6 summarizes the analysis and power analysis settings used for the two scenarios of the case study.

Table 6. Power analysis settings used for the two scenarios of the case study.

Power analysis settings	**Scenario 1**	**Scenario 2**
Significance level	0.05	0.05
Tested replications	8, 16, 24, 32, 40, 48	2, 4, 6, 8, 10, 12
Analysis method difference tests	Log-Normal	Log-Normal
Analysis method equivalence tests	Overdispersed Poisson	Overdispersed Poisson
Power calculation method	Monte-Carlo Simulation (100 per effect/replication)	Monte-Carlo simulation (100 per effect/replication)

6.1.3 Results

6.1.3.1 Power analysis single environment

Figure 7 shows a graphical summary of the combined power analysis results for all endpoints. Table 7 shows the power of the difference tests at the limits of concern for various numbers of replications and Table 8 shows the power for the equivalence tests for the no difference case. It appears that 32 replications are needed to obtain powers larger than 0.8 for the difference test for all endpoints in scenario 1, while more than 48 replicates are required for the equivalence test.

Figure 8 shows the power (represented by a colour) for all endpoints visualized in terms of mean and CV for the difference tests at the lower LoC for 16 replicates. The main cause for a low power for an endpoint is a low mean and/or a high CV.

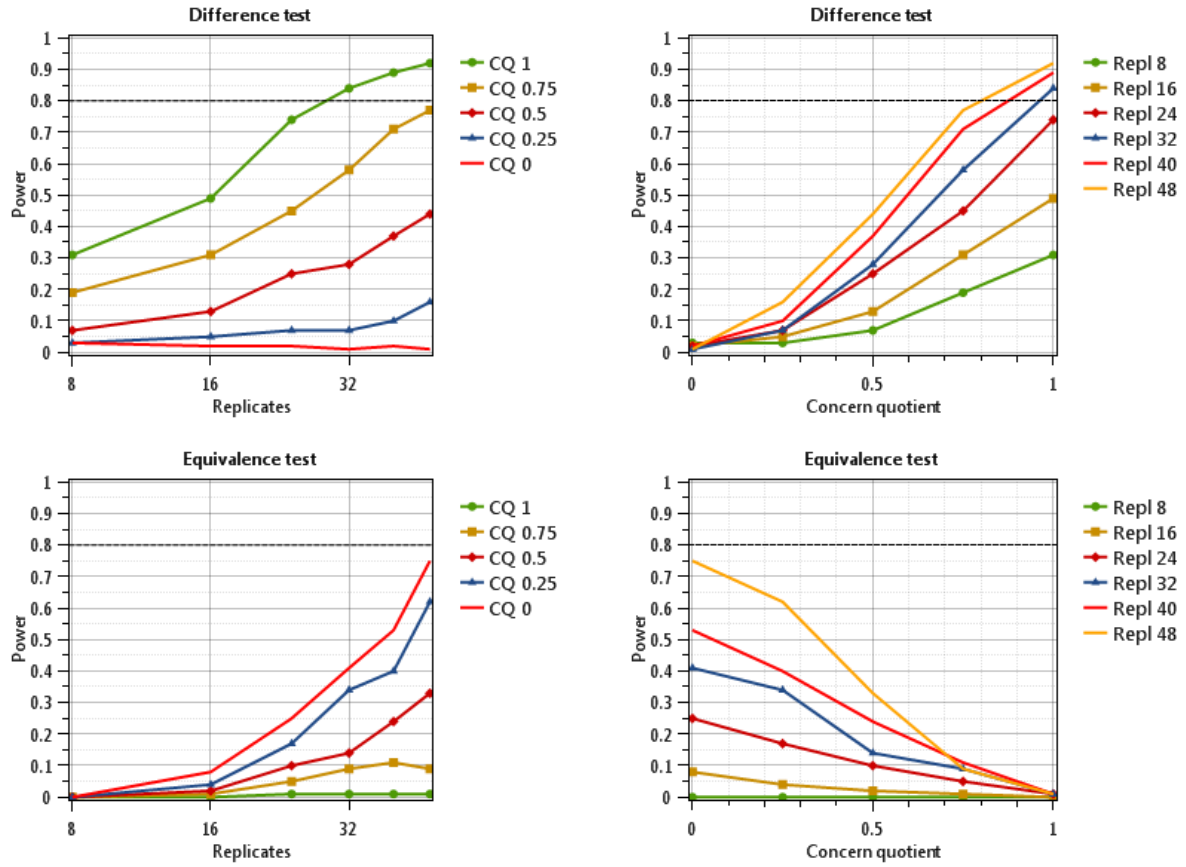

Figure 7. Powers for scenario 1: Summary of the combined power analysis for all endpoints for various levels of the concern quotient aggregated by taking the minimum power over all endpoints. Top row: the results for the difference tests. Bottom row: the results for the equivalence test. Left column: the power versus the number of replicates. Right column: the power versus the concern quotient.

Table 7. Results scenario 1: Power of the difference tests at the limits of concern (i.e., CQ=1) for various numbers of replications. Powers larger than 0.8 are given in green.

Endpoint	Overall mean	CV CMP (%)	8	16	24	32	40	48
Click beetles	10.5	83.2	0.36	0.75	0.90	0.98	0.98	1
Millipedes	15.0	112.4	0.33	0.53	0.83	0.88	0.89	0.97
Aphids	15.9	61.3	0.58	0.88	1	1	1	1
Springtails globular	16.2	96.9	0.35	0.71	0.86	0.9	0.92	0.98
Antlike flower	16.3	63.7	0.69	0.93	1	1	1	1
Ground beetle larvae	17.4	122.5	0.34	0.63	0.76	0.84	0.95	0.97
Sowbugs	28.1	144.3	0.36	0.49	0.75	0.90	0.89	0.92
Predaceous mites	30.1	48.9	0.73	1	1	1	1	1
Oribatid mites	44.6	128.5	0.31	0.56	0.74	0.89	0.94	0.96
Japanese beetles	47.5	87.8	0.43	0.80	0.96	0.97	0.99	1
Springtails hypogastrurids	96.0	94.7	0.40	0.83	0.96	0.98	0.99	0.99
Ants	131.0	108.2	0.35	0.76	0.87	0.94	0.99	1
Springtails isotomids	132.8	44.8	0.81	1	1	1	1	1
General collembola	706.7	51.3	0.76	1	1	1	1	1
Springtails entomobryids	710.4	51.2	0.67	1	1	1	1	1

Table 8. Results scenario 1: Power of the equivalence tests at the level of no difference (i.e., CQ=0) for various numbers of replications. Powers larger than 0.8 are given in green.

Endpoint	Overall mean	CV CMP (%)	8	16	24	32	40	48
Click beetles	10.5	83.2	0.28	0.73	0.90	0.97	0.99	0.99
Millipedes	15.0	112.4	0.01	0.22	0.54	0.78	0.83	0.96
Aphids	15.9	61.3	0.68	0.97	1	1	1	1
Springtails globular	16.2	96.9	0.04	0.48	0.84	0.83	0.94	0.99
Antlike flower	16.3	63.7	0.55	0.92	0.99	1	1	1
Ground beetle larvae	17.4	122.5	0.01	0.14	0.37	0.68	0.79	0.87
Sowbugs	28.1	144.3	0.01	0.08	0.25	0.41	0.53	0.75
Predaceous mites	30.1	48.9	0.91	1	1	1	1	1
Oribatid mites	44.6	128.5	0.00	0.16	0.34	0.62	0.72	0.86
Japanese beetles	47.5	87.8	0.16	0.57	0.87	0.95	0.99	1
Springtails hypogastrurids	96.0	94.7	0.06	0.51	0.82	0.96	0.94	1
Ants	131.0	108.2	0.03	0.31	0.56	0.83	0.88	0.94
Springtails isotomids	132.8	44.8	0.90	1	1	1	1	1
General collembola	706.7	51.3	0.84	0.98	1	1	1	1
Springtails entomobryids	710.4	51.2	0.87	1	1	1	1	1

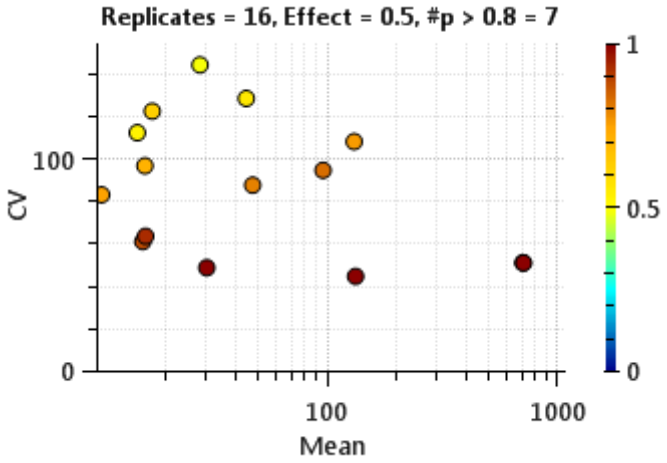

Figure 8. Results scenario 1: The power (represented by a colour) for all endpoints visualized in terms of mean and CV for the difference tests at the lower LoC for 16 replicates illustrating the main cause of a low power, being a low mean and/or a high CV.

6.1.3.2 Power analysis multiple environments

For scenario 2, the same output is obtained as for scenario 1. Figure 9 shows a graphical summary of the combined power analysis results for all endpoints. Table 9 shows the power of the difference tests at the limits of concern (i.e., CQ=1) for various numbers of replications and Table 10 shows the power for the equivalence tests at the level of no difference (i.e., CQ=0). These results are similar to the results of scenario 1: 10 environments, with a total of

40 replicates, is sufficient for an overall power of 0.8 for the difference and for the equivalence test.

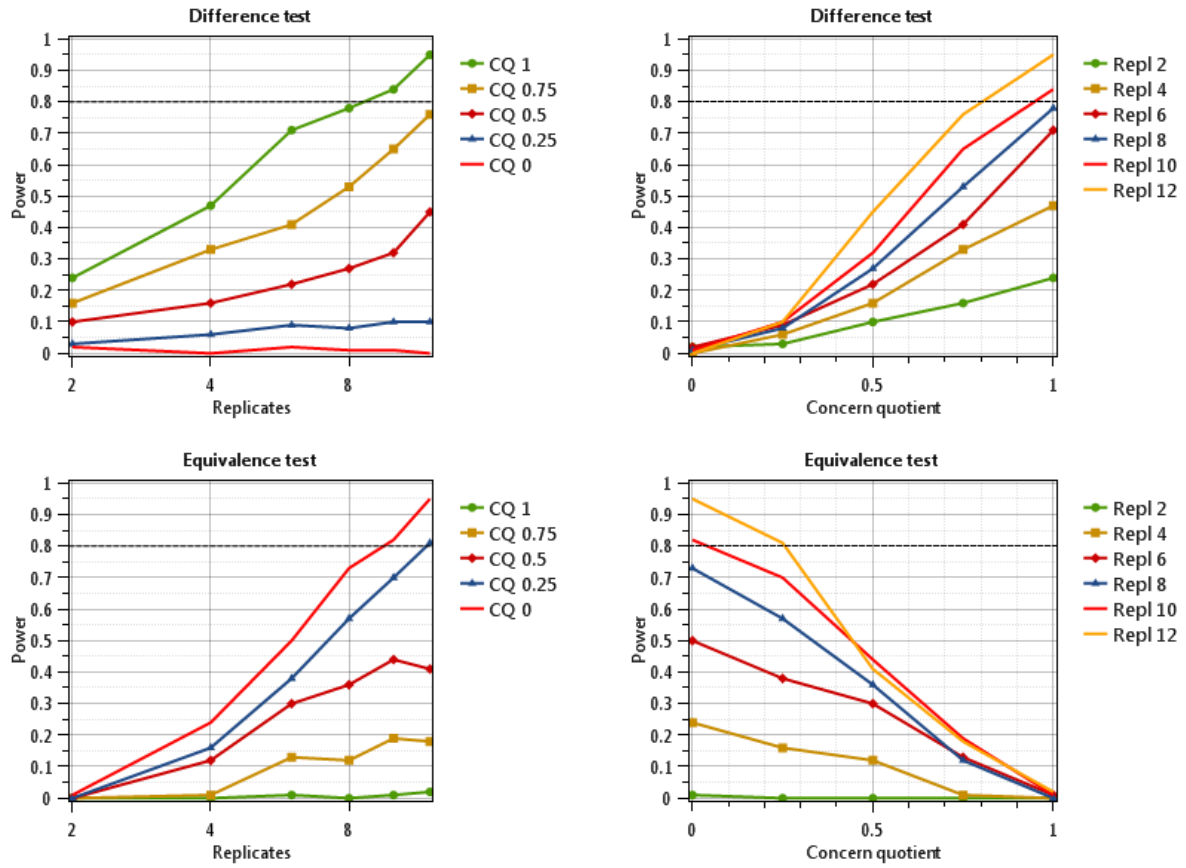

Figure 9. Results scenario 2: Summary of the combined power analysis for all endpoints for various levels of the concern quotient aggregated by taking the minimum power over all endpoints. Top row: the results for the difference tests. Bottom row: the results for the equivalence test. Left column: the power versus the number of replicates. Right column: the power versus the concern quotient.

Table 9. Results scenario 2: Power of the difference tests at the limits of concern (i.e., CQ=1) for various numbers of replications. Powers larger than 0.8 are given in green.

Endpoint	Overall mean	CV CMP (%)	2	4	6	8	10	12
Click beetles	10.5	83.2	0.49	0.80	0.87	0.98	0.96	0.99
Millipedes	15.0	112.4	0.44	0.47	0.79	0.92	0.94	0.96
Aphids	15.9	61.3	0.62	0.86	0.97	0.99	1	1
Springtails globular	16.2	96.9	0.32	0.63	0.83	0.90	1	0.99
Antlike flower	16.3	63.7	0.53	0.84	0.95	0.98	1	1
Ground beetle larvae	17.4	122.5	0.31	0.56	0.76	0.83	0.89	0.98
Sowbugs	28.1	144.3	0.24	0.55	0.75	0.78	0.84	0.95
Predaceous mites	30.1	48.9	0.69	0.96	1	1	1	1
Oribatid mites	44.6	128.5	0.38	0.64	0.71	0.84	0.92	0.95
Japanese beetles	47.5	87.8	0.45	0.70	0.87	0.93	0.99	0.99
Springtails hypogastrurids	96.0	94.7	0.46	0.62	0.85	0.92	0.97	0.97
Ants	131.0	108.2	0.38	0.63	0.82	0.90	0.97	1
Springtails isotomids	132.8	44.8	0.71	0.93	1	1	1	1
General collembola	706.7	51.3	0.69	0.98	0.99	1	1	1
Springtails entomobryids	710.4	51.2	0.68	0.95	1	1	1	1

Table 10. Results scenario 2: Power of the equivalence tests at the level of no difference (i.e., CQ=0) for various numbers of replications. Powers larger than 0.8 are given in green.

Endpoint	Overall mean	CV CMP (%)	2	4	6	8	10	12
Click beetles	10.5	83.2	0.31	0.88	0.97	1	1	1
Millipedes	15.0	112.4	0.10	0.53	0.81	0.96	1	1
Aphids	15.9	61.3	0.69	0.99	1	1	1	1
Springtails globular	16.2	96.9	0.14	0.74	0.91	0.98	0.99	1
Antlike flower	16.3	63.7	0.71	1	1	1	1	1
Ground beetle larvae	17.4	122.5	0.14	0.68	0.88	0.99	1	1
Sowbugs	28.1	144.3	0.02	0.24	0.50	0.73	0.82	0.98
Predaceous mites	30.1	48.9	0.96	1	1	1	1	1
Oribatid mites	44.6	128.5	0.01	0.28	0.56	0.77	0.86	0.95
Japanese beetles	47.5	87.8	0.29	0.82	0.97	0.99	1	1
Springtails hypogastrurids	96.0	94.7	0.31	0.96	1	1	1	1
Ants	131.0	108.2	0.04	0.51	0.76	0.91	0.96	0.98
Springtails isotomids	132.8	44.8	1	1	1	1	1	1
General collembola	706.7	51.3	0.87	1	1	1	1	1
Springtails entomobryids	710.4	51.2	0.85	1	1	1	1	1

6.1.3.3 Discussion

We can align the results of scenario 1 and scenario 2 by expressing the replication levels in terms of the total number of plots. Figure 10 shows the number of endpoints with power

larger than 0.8 as a function of the number of plots for difference and equivalence tests, and for both scenarios.

A noteworthy detail of Figure 10 is the order of the three lines of the difference test for the lower LoC, the equivalence test, and the difference test for the upper LoC. Here, it can be seen that the difference test on the lower LoC yields lower numbers of endpoints with power larger than 0.8 (here called 'analysable endpoints') than the difference test for the upper LoC. Apparently, the power to detect a two-fold reduction in abundance is lower than the power to detect a two-fold increase in this scenario. An explanation for this might be that a two-fold decrease implies lower abundances which are apparently less informative that larger abundances.

Table 11 shows the number of analysable endpoints expressed in terms of the total number of plots for both scenarios. From this table, it can be observed that scenario 2 yields more analysable endpoints for the equivalence test for the same number of plots. An explanation for this might be that different site/year combinations will result in experiments with relatively high counts, due to between site/year variation, which are apparently more informative.

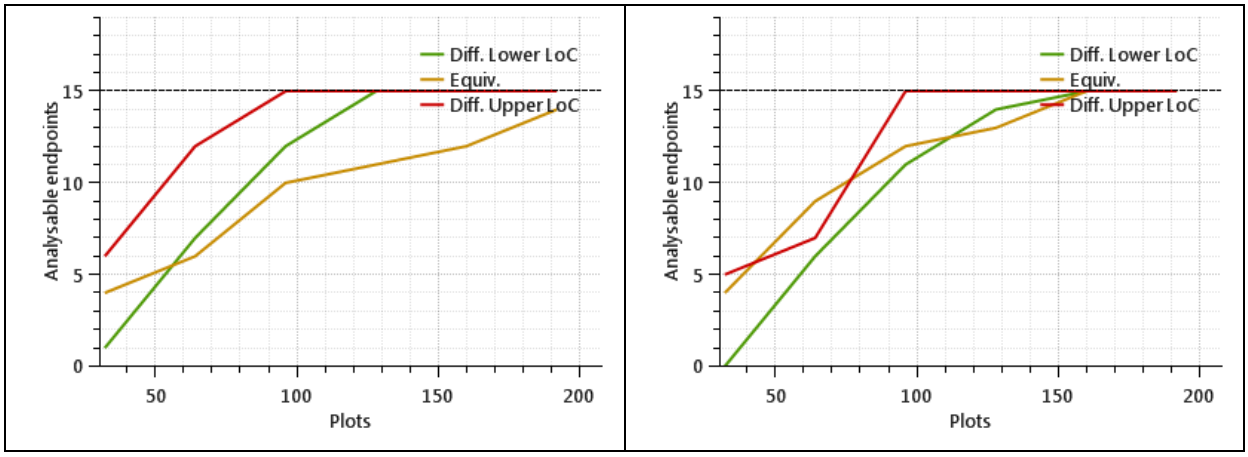

Figure 10. Number of analysable endpoints (i.e. with a power > 0.8) for each number of replicates for the difference tests at the lower and upper LoC and the equivalence test at the point of no-difference. Left: results of scenario 1. Right: results of scenario 2.

Table 11. Number of analysable endpoints (i.e. with a power > 0.8) for scenario 1 for different numbers of replicates calculated for the difference tests at the lower and upper LoC and the equivalence test at the point of no-difference.

	Scenario 1			Scenario 2		
Nplots	**Difference test LoC_{low}**	**Equivalence test**	**Difference test LoC_{upp}**	**Difference test LoC_{low}**	**Equivalence test**	**Difference test LoC_{upp}**
32	1	4	6	0	4	5
64	7	6	12	6	9	7
96	12	10	15	11	12	15
128	15	11	15	14	13	15
160	15	12	15	15	15	15
192	15	14	15	15	15	15

6.2 NTOs in maize in Spain, Slovakia, Denmark, Sweden

6.2.1 Examples of trees of endpoints

Figure 11 shows an example of a hierarchical tree for the analysis of maize NTO data in AMIGA. The attention here is restricted to the arthropod count data of Spain and Slovakia, and the profiling data of the soil micro-organisms in the four countries Spain, Slovakia, Denmark and Sweden. A first decision was to sum the arthropod counts over the time-points within a season, because not enough power was expected for low abundance taxa at single time points. It was also decided not to focus on single years, but to further sum the counts over the available years when the experimental design was continued on the same plots. The data were then analysed per taxon and per site (see 6.2.2.1 and 6.2.2.2). The soil data were analysed per functional group per site and per year (6.2.2.3). The effects were then summarised by various forms of equivalence analysis (6.2.2.4). The arthropod data were first summarised for functional groups per site, then for functional groups and finally to the whole arthropod group. The soil data were summarised per functional group per site per year, and then per functional group, and finally for the whole rhizosphere micro-organism category. The last step is then to assess the equivalence across the two NTO categories.

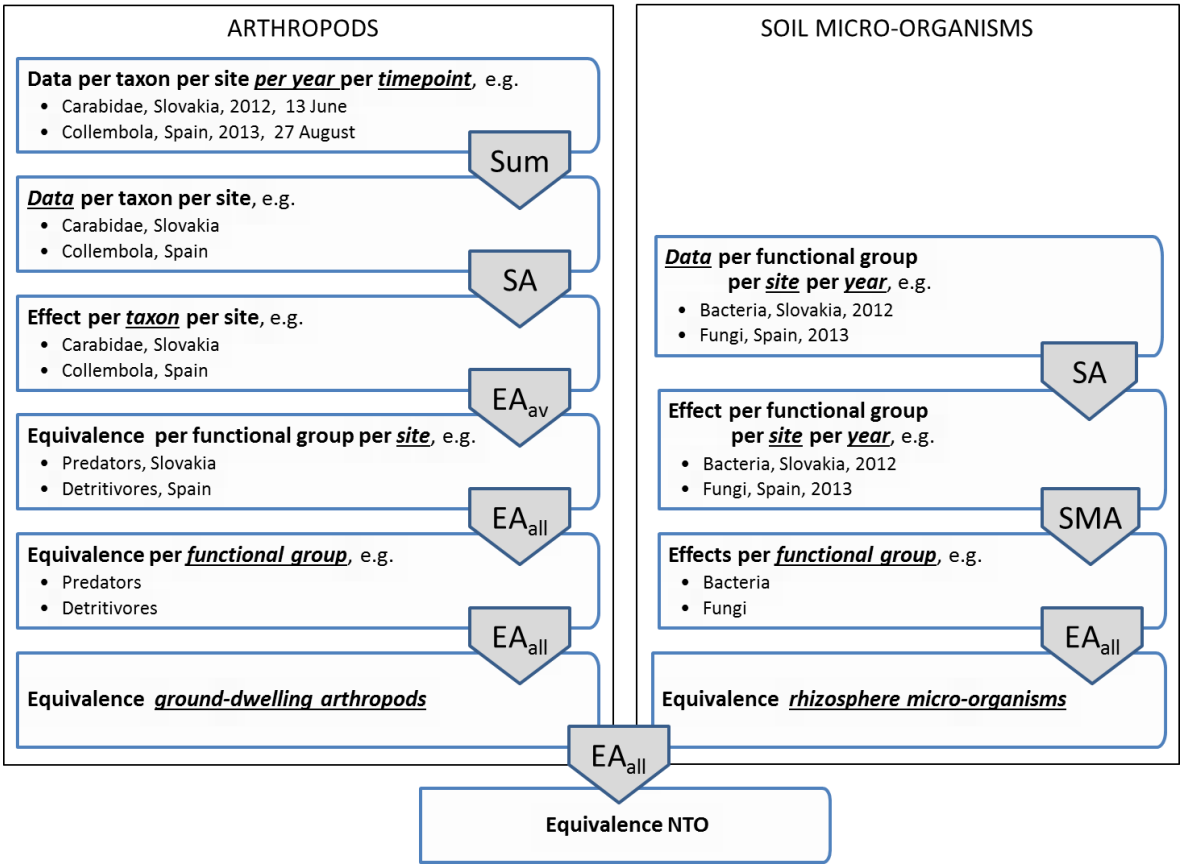

Figure 11. Example 1 of a logical tree for analysis of the data in a NTO study of a GM maize. SUM = summation of data. SA = statistical analysis. SHA = statistical hierarchical analysis. SMA = statistical meta-analysis. EA_{av} = equivalence analysis with averaging of concern quotients. EA_{all} = equivalence analysis requiring all concern quotients to be within limits.

Many other analysis schemes are possible. For example Figure 12 shows an alternative scheme where the arthropod data are first summarised over functional categories per site. This would be relevant for example if national decisions have to be made.

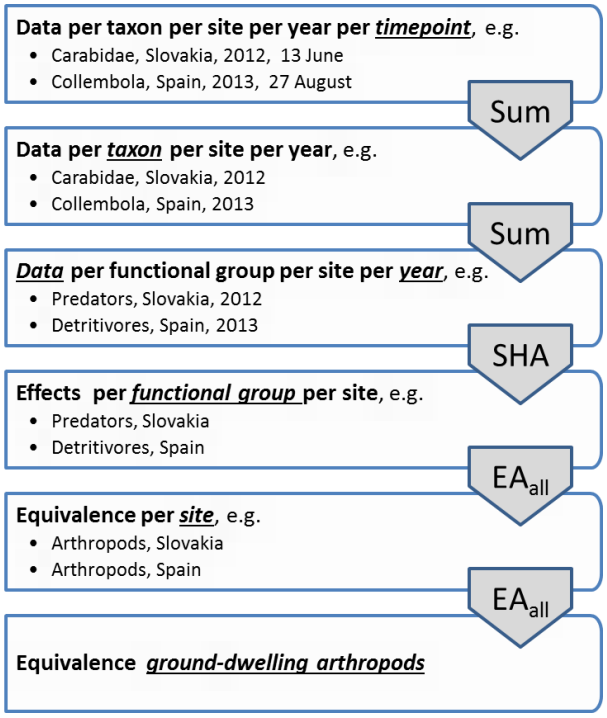

Figure 12. Example 2 of a logical tree for analysis of the data in a NTO study of a GM maize, as an alternative to the subtree for Arthropods in Figure 11). SUM = summation of data. SA = statistical analysis. SHA = statistical hierarchical analysis. SMA = statistical meta-analysis. EA_{av} = equivalence analysis with averaging of concern quotients. EA_{all} = equivalence analysis requiring all concern quotients to be within limits.

6.2.2 Example analyses NTO field study maize

6.2.2.1 Arthropods maize Spain

The purpose was to compare two varieties of maize, BT and ISO. A field trial was performed in Seseña, Spain, in 2012, 2013 and 2014 using the same experimental design in the three years. The experimental design was a randomized block design with rows as blocks and two replicates of ISO and BT were randomized within each row (Figure 13). In each of the 20 plots two pitfall traps were placed. Count data for arthropods were obtained in 9 sampling periods per year. The taxa were grouped into 5 functional groups (herbivores, predators, parasitoids, detritivores, other). Statistical analyses were performed using GenStat 18th edition (VSN 2012).

Following the logical tree for analysis in Figure 11 counts were summed over the two traps per plot and over the nine sampling periods per year. For the multi-year analysis counts were also summed over the three years.

	Col 1	Col 2	Col 3	Col 4
Row 1	BT	BT	ISO	ISO
Row 2	BT	ISO	ISO	BT
Row 3	BT	ISO	BT	ISO
Row 4	BT	ISO	ISO	BT
Row 5	ISO	BT	BT	ISO

Figure 13. Scheme showing the experimental design of the maize field study in Spain.

The mean overall count (*m*) and the means for the two maize varieties were calculated. For taxa where both variety means were positive an over-dispersed Poisson (OP) log-linear model was fitted with on the logarithmic scale additive effects for block and variety. This model assumes that the variance of a count is proportional to the mean and that differences are additive on the log-scale. The dispersion factor was estimated with the Pearson statistic *X*. It was set to 1 if the estimate was lower than 1. From this the coefficient of variation (*CV*), expressed as a percentage, was estimated as $CV = 100\sqrt{X/m}$ in which *m* is the mean of the counts. The log-linear model directly estimates the log-ratio (or log of fold change) of the mean of the BT and ISO variety and its associated standard error. These were used to construct a two-sided 90% confidence interval for the log-ratio by calculating {log-ratio ± *tvalue* × se}, in which *tvalue* is according to the Student distribution with appropriate degrees of freedom. Note that a 90% interval is used, rather than a 95% interval, to enable equivalence testing at a 5% significance level by means of the two one-sided tests (TOST) approach. Estimates of fold change and confidence intervals are graphically depicted on the log-scale in Figure 14, together with the equivalence region between the limits of concern. A further scaling to LoC scaled differences (*LoCSDIFF*) leads to the representation in Figure 15.

For taxa with a zero average for the GMO or CMP it is not possible to estimate the log-ratio using the OP model. For these taxa, we calculated a ratio where the zero count was replaced with the lowest possible mean value based on a count 1, e.g. 0.1 in the current case with 10 replications per year. If these estimates fell outside the limits of concern they were included in the graphical display to focus attention on these possibly relevant changes. However, this situation did not occur with the current dataset.

In the example of Figure 14 and Figure 15 the point estimates for all taxa are within the equivalence region. Possibly decreased levels of Eurobella and Chironomidae, and possibly increased levels of Collembola, Corulophidae, Chrysopidae and Formicidae cannot be excluded. Note that there are also several significant differences (Aphididae, Phytoseiidae, Opiliones, Corylophidae, Chironomidae), but the points estimates for all of them as well as the confidence intervals for the first three mentioned taxa are all fully within the equivalence region.

Finally, it should be stressed again that the LoC values used in this case study are tentative. In practice LoC values can always be discussed, and based on the context given other values, which would then lead to adapted graphs.

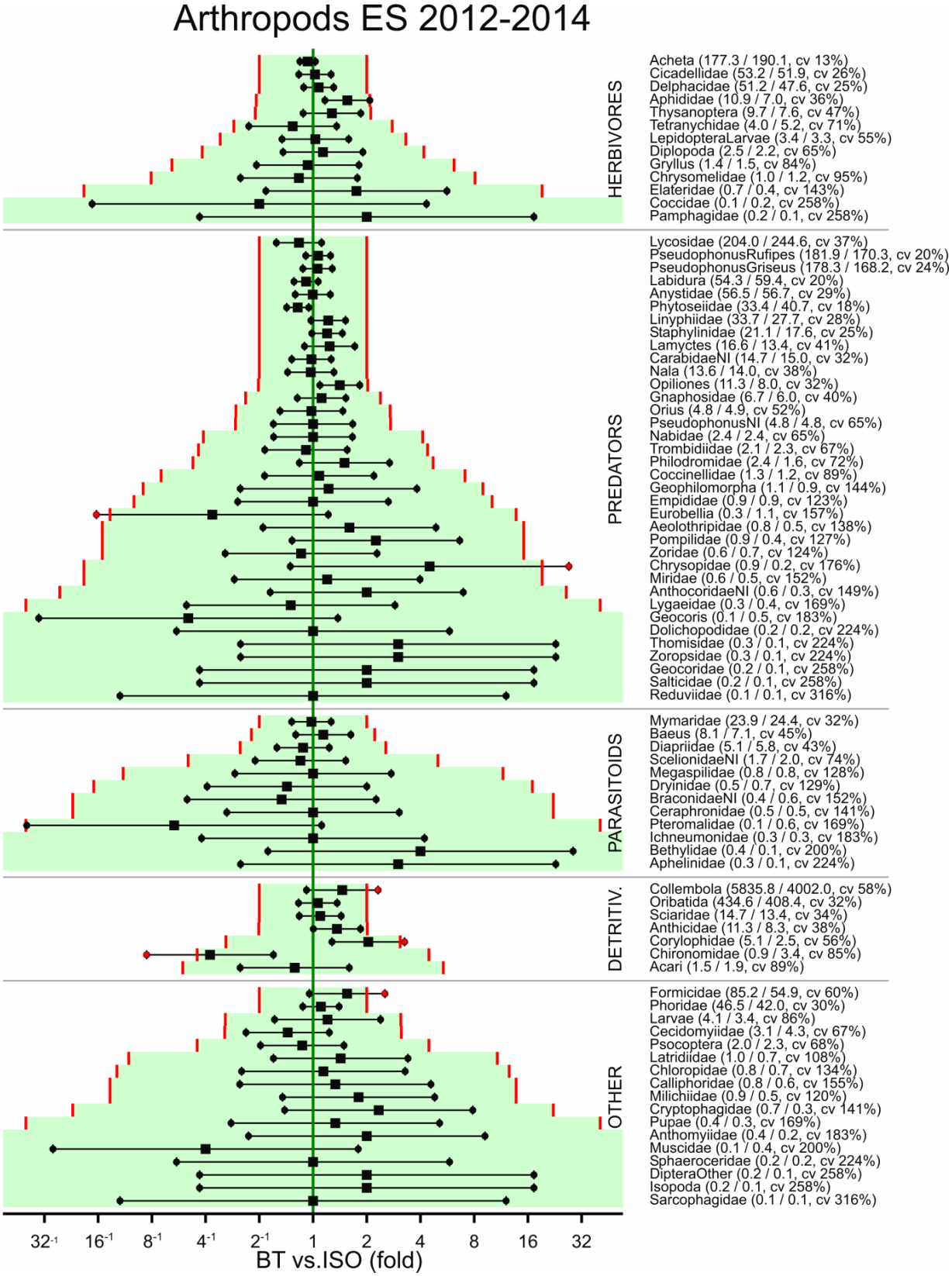

Figure 14. Arthropods in field trials Spain, summed over 2012-2014. Fold changes GMO vs. CMP for all taxa with positive means for both varieties. LoCs are factors 0.5 and 2 for taxa with means of 10 and higher, and log(LoC) is scaled by $\sqrt{\mathbf{10/m}}$ for lower means. Results sorted according to LoC within functional groups. Means over the ten plots for test and comparator group and CV are indicated in brackets.

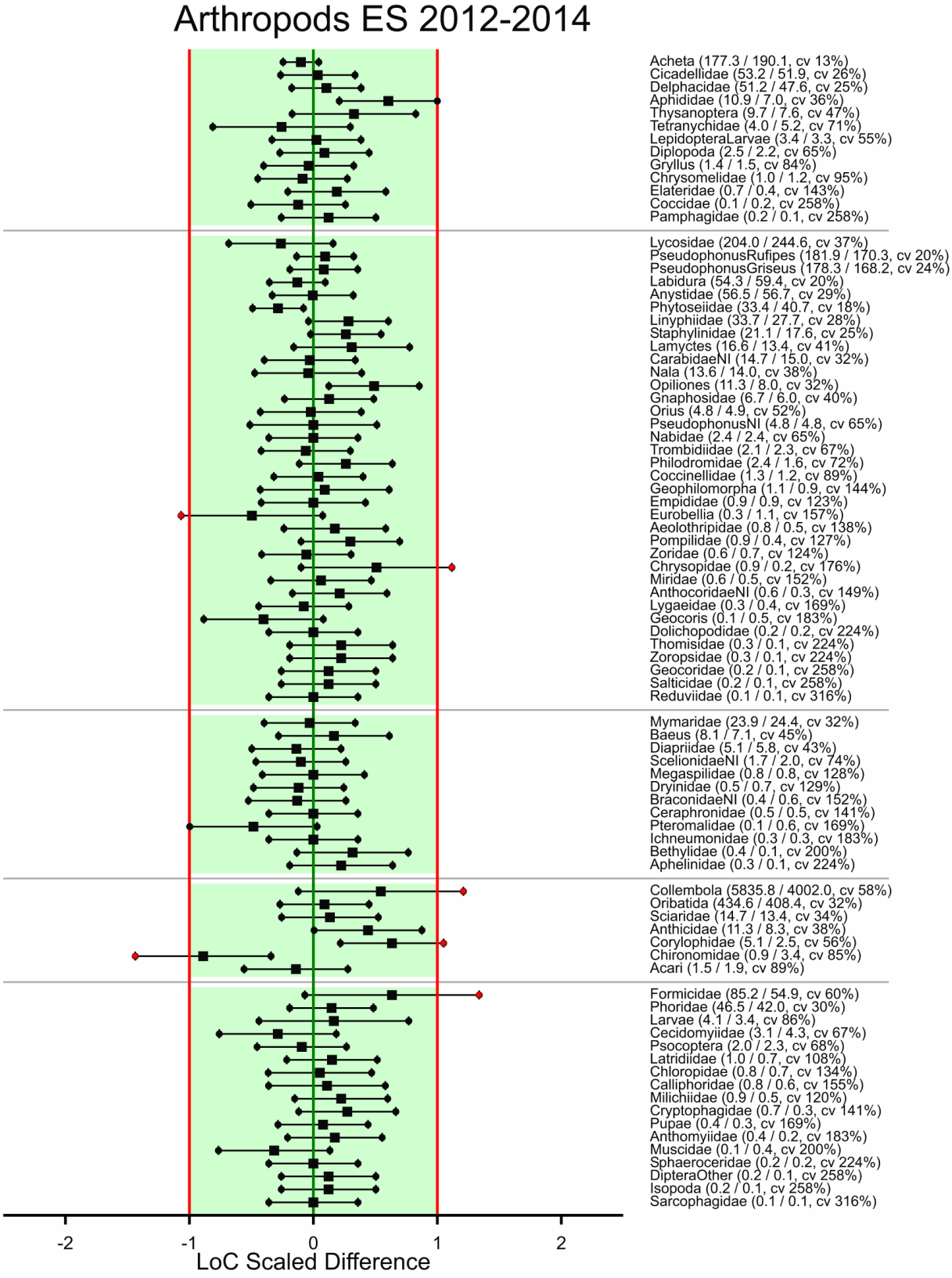

Figure 15. Arthropods in field trials Spain, summed over 2012-2014. LoC scaled differences GMO vs. CMP for all taxa with positive means for both varieties. LoCs are factors 0.5 and 2 for taxa with means of 10 and higher, and log(LoC) is scaled by $\sqrt{\mathbf{10/m}}$ for lower means. Results sorted according to LoC within functional groups. Means over the ten plots for test and comparator group and CV are indicated in brackets.

6.2.2.2 Arthropods maize Slovakia

The purpose was to compare two varieties of maize, BT and ISO. A field trial was performed in Borovce, Slovakia, in 2012, 2013, 2014 and 2015 using the same experimental design in the three years. The experimental design was a block design with two plots with ISO and BT per block in a systematic order, see Figure 16

	Col 1	Col 2	Col 3	Col 4
Row 1	ISO	BT	ISO	BT
Row 2	BT	ISO	BT	ISO
Row 3	ISO	BT	ISO	BT
Row 4	BT	ISO	BT	ISO
Row 5	ISO	BT	ISO	BT

Figure 16. Scheme showing the experimental design of the field study in Slovakia.

In each of the 20 plots pitfall traps were placed. Count data for arthropods were obtained using pitfall traps in around 9 sampling periods per year. Using the same methods as for Spain, but for a different list and grouping of arthropod taxa, we arrive at the results in Figure 17. All point estimates of the fold change lie in the equivalence region. The only uncertainty is about the Collembola species Xenylla welchi, which might be decreased. The Elateridae were significantly increased, but within the limits of concern.

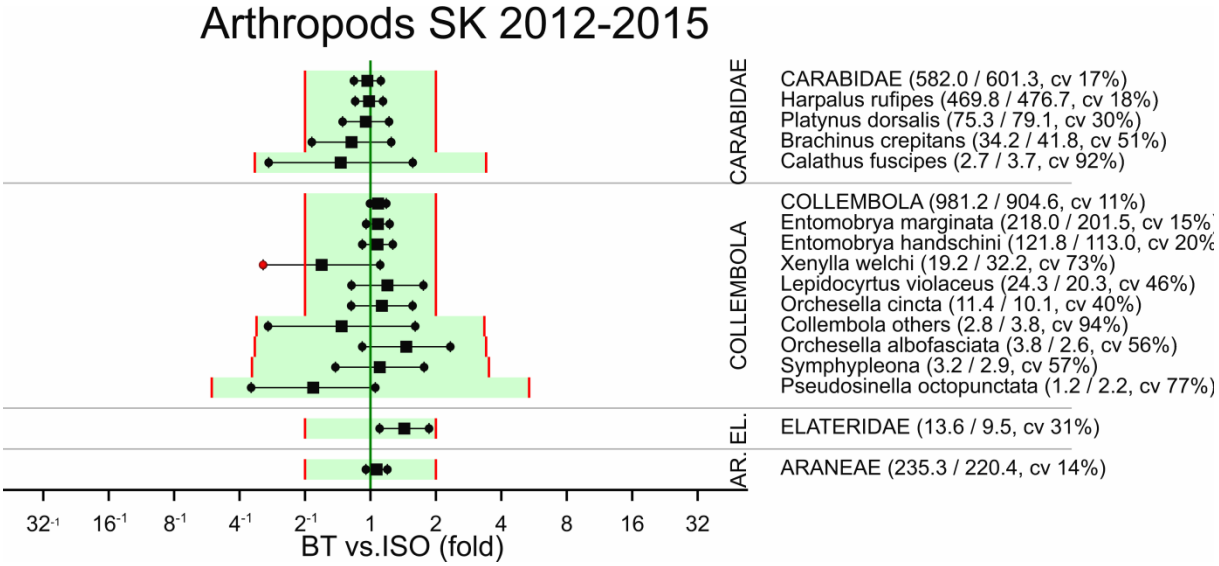

Figure 17. Arthropods in field trials Slovakia, summed over 2012-2015. Fold changes GMO vs. CMP for all taxa with positive means for both varieties. LoCs are factors 0.5 and 2 for taxa with means of 10 and higher, and log(LoC) is scaled by $\sqrt{\mathbf{10}/\boldsymbol{m}}$ for lower means. Results sorted according to LoC within groups. Means over the ten plots for test and comparator group and CV are indicated in brackets.

6.2.2.3 Soil micro-organisms Spain, Slovakia, Denmark, Sweden

Total soil DNA was extracted from the rhizospheres and this DNA served as the raw material to assess the diversity of bacteria, archaea and fungi. Similarly, as an indicator for functional

diversity, the nirS/nirK genes, which encode for bacterial denitrification, were sequenced. The five response variables (Bacteria, Archaea, Fungi, nirK and nirS) were log transformed to stabilize the variance. Each response variable was subjected to analysis of variance, separately for each experiment, and 90% confidence intervals for the log-difference between the BT and ISO variety were derived from this statistical analysis. LoCs were again tentatively set to 0.5 and 2. The intervals are given in Figure 18. All intervals, with the exception of Fungi in Slovakia 2013, lie well within the LoCs.

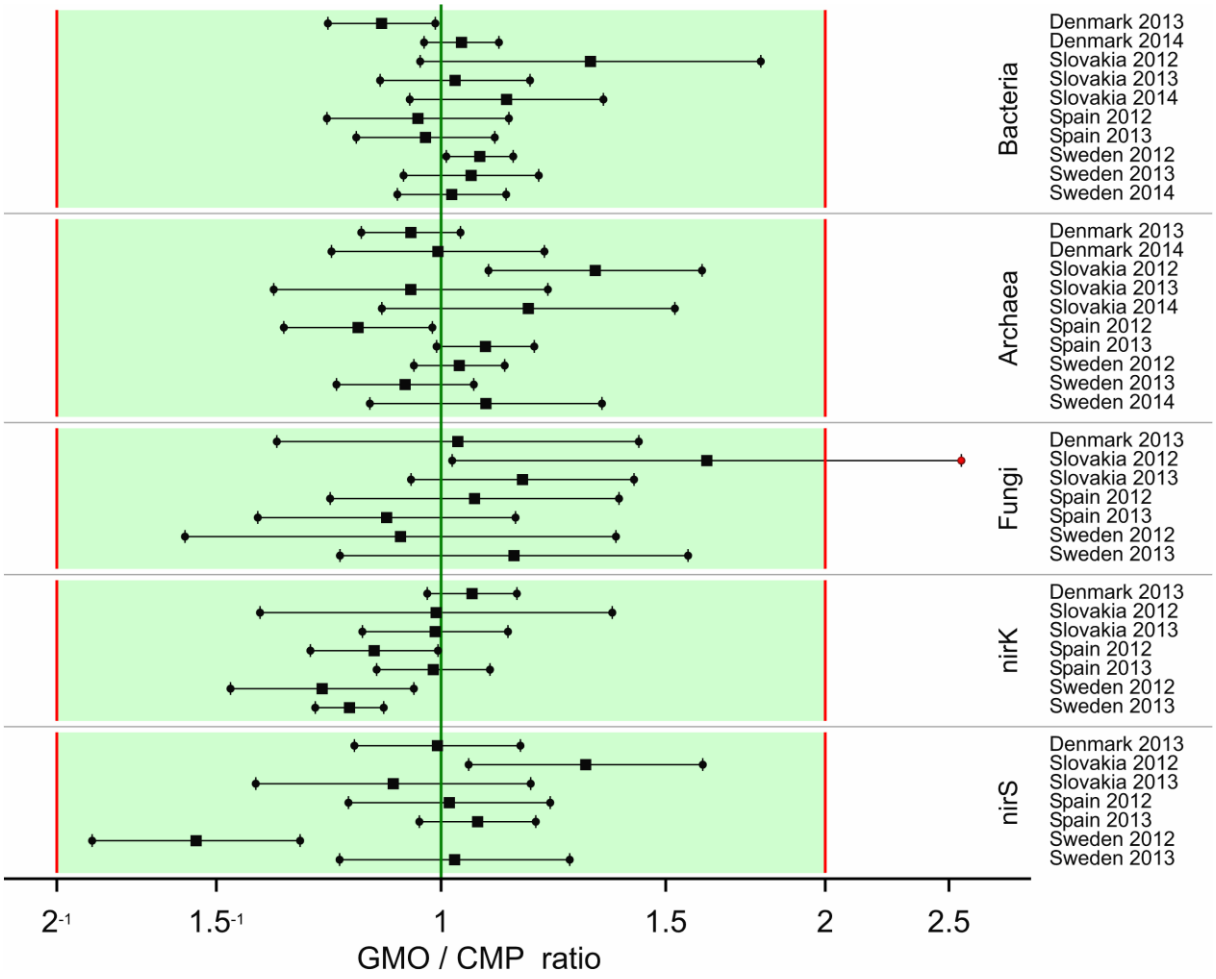

Figure 18. 90% confidence intervals for the ratio between the mean of the GMO and the CMP for soil micro-organisms in Denmark, Slovakia, Spain and Sweden in separate trials. LoCs are tentatively set to 0.5 and 2.

6.2.2.4 Integrated analysis

For the integration we follow the steps of Figure 11, but note that this is just an example of a possible tree for analysis.

6.2.2.4.1 Integration for Spanish arthropod data

The first integration step is an averaging equivalence analysis on the taxa within functional groups (step EA_{av}). Table 12 shows the calculation of concern quotients (CQ) for all taxa, where some of CQ_{upp} values still exceed 1, indicating a possible need for closer inspection.

Table 13 shows the average CQs for the functional groups, showing that no CQ values exceed 1, i.e. all values are within the equivalence range.

Table 12. Calculation of concern quotients CQ for the Spanish arthropod data. Limits of concern and estimates are expressed on the natural log scale. CQ values larger than one are given in light red.

Group	Taxon	LoC_{low}	LoC_{upp}	left	esti	right	CQ	CQ_{upp}
HERBIVORES	Acheta	-0.69	0.69	-0.17	-0.07	0.03	0.10	0.24
HERBIVORES	Cicadellidae	-0.69	0.69	-0.18	0.02	0.23	0.04	0.34
HERBIVORES	Delphacidae	-0.69	0.69	-0.12	0.07	0.27	0.11	0.38
HERBIVORES	Aphididae	-0.73	0.73	0.15	0.44	0.73	0.60	1.00
HERBIVORES	Thysanoptera	-0.75	0.75	-0.13	0.24	0.62	0.33	0.83
HERBIVORES	Tetranychidae	-1.02	1.02	-0.83	-0.26	0.30	0.26	0.81
HERBIVORES	LepidopteraLarvae	-1.20	1.20	-0.40	0.03	0.46	0.02	0.38
HERBIVORES	Diplopoda	-1.43	1.43	-0.39	0.13	0.64	0.09	0.45
HERBIVORES	Gryllus	-1.82	1.82	-0.73	-0.07	0.59	0.04	0.40
HERBIVORES	Chrysomelidae	-2.09	2.09	-0.94	-0.18	0.57	0.09	0.45
HERBIVORES	Elateridae	-2.96	2.96	-0.61	0.56	1.73	0.19	0.58
HERBIVORES	Coccidae	-5.66	5.66	-2.85	-0.69	1.46	0.12	0.50
HERBIVORES	Pamphagidae	-5.66	5.66	-1.46	0.69	2.85	0.12	0.50
PREDATORS	Lycosidae	-0.69	0.69	-0.47	-0.18	0.11	0.26	0.68
PREDATORS	PseudophonusRufipes	-0.69	0.69	-0.09	0.07	0.23	0.10	0.33
PREDATORS	PseudophonusGriseus	-0.69	0.69	-0.13	0.06	0.25	0.08	0.36
PREDATORS	Labidura	-0.69	0.69	-0.25	-0.09	0.07	0.13	0.35
PREDATORS	Anystidae	-0.69	0.69	-0.23	0.00	0.22	0.01	0.33
PREDATORS	Phytoseiidae	-0.69	0.69	-0.34	-0.20	-0.06	0.29	0.49
PREDATORS	Linyphiidae	-0.69	0.69	-0.03	0.20	0.42	0.28	0.60
PREDATORS	Staphylinidae	-0.69	0.69	-0.02	0.18	0.38	0.26	0.55
PREDATORS	Lamyctes	-0.69	0.69	-0.11	0.21	0.54	0.31	0.78
PREDATORS	CarabidaeNI	-0.69	0.69	-0.27	-0.02	0.23	0.03	0.40
PREDATORS	Nala	-0.69	0.69	-0.33	-0.03	0.27	0.04	0.47
PREDATORS	Opiliones	-0.71	0.71	0.09	0.35	0.60	0.49	0.85
PREDATORS	Gnaphosidae	-0.87	0.87	-0.20	0.11	0.42	0.13	0.49
PREDATORS	Orius	-1.00	1.00	-0.43	-0.02	0.38	0.02	0.43
PREDATORS	PseudophonusNI	-1.00	1.00	-0.51	0.00	0.51	0.00	0.51
PREDATORS	Nabidae	-1.42	1.42	-0.51	0.00	0.51	0.00	0.36
PREDATORS	Trombidiidae	-1.48	1.48	-0.62	-0.09	0.44	0.06	0.42
PREDATORS	Philodromidae	-1.55	1.55	-0.18	0.41	0.99	0.26	0.64
PREDATORS	Coccinellidae	-1.96	1.96	-0.62	0.08	0.79	0.04	0.40
PREDATORS	Geophilomorpha	-2.19	2.19	-0.94	0.20	1.34	0.09	0.61
PREDATORS	Empididae	-2.31	2.31	-0.97	0.00	0.97	0.00	0.42
PREDATORS	Eurobellia	-2.62	2.62	-2.79	-1.30	0.20	0.50	1.07
PREDATORS	Aeolothripidae	-2.72	2.72	-0.65	0.47	1.59	0.17	0.58
PREDATORS	Pompilidae	-2.72	2.72	-0.27	0.81	1.89	0.30	0.70
PREDATORS	Zoridae	-2.72	2.72	-1.13	-0.15	0.83	0.06	0.42

PREDATORS	Chrysopidae	-2.96	2.96	-0.29	1.50	3.30	0.51	1.12
PREDATORS	Miridae	-2.96	2.96	-1.02	0.18	1.38	0.06	0.47
PREDATORS	AnthocoridaeNI	-3.27	3.27	-0.55	0.69	1.94	0.21	0.59
PREDATORS	Lygaeidae	-3.71	3.71	-1.63	-0.29	1.06	0.08	0.44
PREDATORS	Geocoris	-4.00	4.00	-3.54	-1.61	0.32	0.40	0.88
PREDATORS	Dolichopodidae	-4.90	4.90	-1.76	0.00	1.76	0.00	0.36
PREDATORS	Thomisidae	-4.90	4.90	-0.94	1.10	3.13	0.22	0.64
PREDATORS	Zoropsidae	-4.90	4.90	-0.94	1.10	3.13	0.22	0.64
PREDATORS	Geocoridae	-5.66	5.66	-1.46	0.69	2.85	0.12	0.50
PREDATORS	Salticidae	-5.66	5.66	-1.46	0.69	2.85	0.12	0.50
PREDATORS	Reduviidae	-6.93	6.93	-2.49	0.00	2.49	0.00	0.36
PARASITOIDS	Mymaridae	-0.69	0.69	-0.28	-0.02	0.23	0.03	0.40
PARASITOIDS	Baeus	-0.80	0.80	-0.22	0.13	0.49	0.17	0.61
PARASITOIDS	Diapriidae	-0.94	0.94	-0.47	-0.13	0.21	0.14	0.50
PARASITOIDS	ScelionidaeNI	-1.61	1.61	-0.75	-0.16	0.42	0.10	0.46
PARASITOIDS	Megaspilidae	-2.45	2.45	-1.01	0.00	1.01	0.00	0.41
PARASITOIDS	Dryinidae	-2.83	2.83	-1.37	-0.34	0.69	0.12	0.48
PARASITOIDS	BraconidaeNI	-3.10	3.10	-1.63	-0.41	0.81	0.13	0.52
PARASITOIDS	Ceraphronidae	-3.10	3.10	-1.11	0.00	1.11	0.00	0.36
PARASITOIDS	Pteromalidae	-3.71	3.71	-3.69	-1.79	0.11	0.48	1.00
PARASITOIDS	Ichneumonidae	-4.00	4.00	-1.44	0.00	1.44	0.00	0.36
PARASITOIDS	Bethylidae	-4.38	4.38	-0.58	1.39	3.36	0.32	0.77
PARASITOIDS	Aphelinidae	-4.90	4.90	-0.94	1.10	3.13	0.22	0.64
DETRITIVORES	Collembola	-0.69	0.69	-0.08	0.38	0.84	0.54	1.21
DETRITIVORES	Oribatida	-0.69	0.69	-0.19	0.06	0.31	0.09	0.45
DETRITIVORES	Sciaridae	-0.69	0.69	-0.18	0.09	0.36	0.13	0.52
DETRITIVORES	Anthicidae	-0.70	0.70	0.00	0.31	0.61	0.44	0.88
DETRITIVORES	Corylophidae	-1.12	1.12	0.24	0.71	1.18	0.63	1.05
DETRITIVORES	Chironomidae	-1.50	1.50	-2.15	-1.33	-0.51	0.89	1.44
DETRITIVORES	Acari	-1.68	1.68	-0.94	-0.24	0.47	0.14	0.56
OTHER	Formicidae	-0.69	0.69	-0.05	0.44	0.93	0.63	1.34
OTHER	Phoridae	-0.69	0.69	-0.13	0.10	0.34	0.15	0.48
OTHER	Larvae	-1.13	1.13	-0.50	0.19	0.87	0.17	0.77
OTHER	Cecidomyiidae	-1.14	1.14	-0.86	-0.33	0.21	0.29	0.76
OTHER	Psocoptera	-1.50	1.50	-0.68	-0.14	0.40	0.09	0.45
OTHER	Latridiidae	-2.38	2.38	-0.51	0.36	1.22	0.15	0.51
OTHER	Chloropidae	-2.53	2.53	-0.92	0.13	1.19	0.05	0.47
OTHER	Calliphoridae	-2.62	2.62	-0.95	0.29	1.52	0.11	0.58
OTHER	Milichiidae	-2.62	2.62	-0.39	0.59	1.57	0.22	0.60
OTHER	Cryptophagidae	-3.10	3.10	-0.37	0.85	2.06	0.27	0.66
OTHER	Pupae	-3.71	3.71	-1.06	0.29	1.63	0.08	0.44
OTHER	Anthomyiidae	-4.00	4.00	-0.83	0.69	2.22	0.17	0.55
OTHER	Muscidae	-4.38	4.38	-3.36	-1.39	0.58	0.32	0.77
OTHER	Sphaeroceridae	-4.90	4.90	-1.76	0.00	1.76	0.00	0.36
OTHER	DipteraOther	-5.66	5.66	-1.46	0.69	2.85	0.12	0.50

OTHER	Isopoda	-5.66	5.66	-1.46	0.69	2.85	0.12	0.50
OTHER	Sarcophagidae	-6.93	6.93	-2.49	0.00	2.49	0.00	0.36

Table 13. EA_{av} method applied to Spanish Arthropod data.

Group	**CQ**	**CQ_{upp}**
HERBIVORES	0.16	0.53
PREDATORS	0.16	0.55
PARASITOIDS	0.14	0.54
DETRITIVORES	0.41	0.87
OTHER	0.17	0.60

6.2.2.4.2 Integration for Slovakian arthropod data

In the Slovakian field trial the identification of arthropods has been done in another way and another grouping of arthropod taxa has been used. However, the same method can be applied to calculate CQ values, and average them over the taxa in a group. The results are given in Table 14, showing that no CQ values exceed 1, i.e. all values are within the equivalence range.

Table 14. EA_{av} method applied to Slovakian Arthropod data.

Group	**CQ**	**CQ_{upp}**
PREDATORS: Carabidae, Araneae	0.15	0.54
DETRITIVORES: Collembola	0.27	0.72
HERBIVORES: Elateridae	0.52	0.90

6.2.2.4.3 Integration for arthropod functional groups

The next step is an equivalence analysis over sites where all members should comply to their LoCs (EA_{all}). Therefore, we take the maximum available CQ value for each functional group. Obviously, no concern is indicated, because this was already the case at the previous level.

Table 15. EA_{all} method applied to Arthropod data of Spanish and Slovakian sites.

Group	**CQ**	**CQ_{upp}**
HERBIVORES	0.52	0.90
PREDATORS	0.16	0.55
PARASITOIDS	0.14	0.54
DETRITIVORES	0.41	0.87
OTHER	0.17	0.60

6.2.2.4.4 Integration for arthropods

Again taking the maximum over the functional groups we arrive at final values for the Arthropod group as a whole. Obviously, no concern is indicated, because this was already the case at the previous level.

6.2.2.4.5 Integration for soil micro-organism data

A meta-analysis of the soil effects over sites and years for each endpoint results in the 90% confidence intervals given in Figure 19. This reveals that, still employing the same LoCs of 0.5 and 2, all intervals lie well within the LoCs. This would imply that there is no concern across environments. Further integration over endpoints, employing an equivalence analysis (EA), does not present any problems, and leads to the conclusion of no concern for the category of soil micro-organisms.

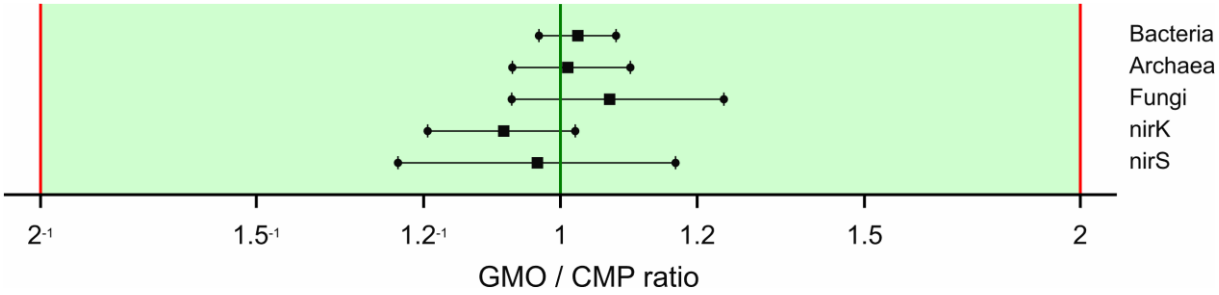

Figure 19. 90% confidence intervals for the ratio between the mean of the GMO and the CMP resulting from a meta-analysis for each response variable over sites and years for soil micro-organisms in Denmark, Slovakia, Spain and Sweden. LoCs are tentatively set to 0.5 and 2.

Table 16. EA_{all} method applied to Soil biology data per group of sites in Denmark, Slovakia, Spain and Sweden, 2012-2014.

Group	**CQ**	**CQ_{upp}**
Bacteria	0.02	0.07
Archaea	0.01	0.09
Fungi	0.07	0.22
nirK	0.08	0.18
nirS	0.03	0.22

6.2.2.4.6 Overall integration maize NTO data

The final step is to integrate using EA_{all} the conclusions over the NTO categories, in this example the arthropod data and the soil micro-organism data. Obviously the final conclusion is that of equivalence, because this was already the case for each of these categories.

Table 17. NTO in maize. EAall method applied to Arthropod and Soil micro-organisms data

Level of assessment	**CQ**	**CQ_{upp}**
Arthropods	0.52	0.90
Soil micro-organisms	0.08	0.22
NTO	0.52	0.90

6.3 NTOs in potato Ireland and Netherlands

Field trials with potato varieties were performed in Ireland and the Netherlands both in 2013 and in 2014. The main purpose was to compare a GMO potato variety, called A15-13, with its

comparator, or CMP for short, which is Desiree. A third variety SarpoMira was also included in the trial and all three varieties were subjected to three agricultural treatments: Weekly spraying, No spraying and spraying according to IPM. Completely randomized block designs were employed and a fresh randomization was carried out for each of the four experiments. The number of replications six in Ireland and seven in the Netherlands. Pitfalls were placed in every plot and emptied three times during each trial. Arthropods were identified and counted in each pitfall. Taxa were grouped into the six functional groups Predators, Detritivores, Parasitoids, Fungivores, Herbivores, Hyperparasitoids and a seventh group Unknown with remaining taxa.

6.3.1 Examples of trees to analyse NTOs in potato trials

Figure 20 shows three examples of hierarchies for analysing the NTO data. Details of the steps depicted in Figure 20 and their implicit assumptions are detailed below. For hierarchy A in Figure 20 the steps are as follows

A.1 SUM: the first step in hierarchy A is to sum the count data for each individual taxon over the three time points which results in a single count for every taxon for each plot per site per year. This was done because not enough power was expected at single time points especially for the less abundant species. Summing disregards interactions with time-point within experiments.

A.2 SA: counts of single taxa within experiments are statistically analysed to give effects for each taxon per site per year. This enables us to inspect the effect for every single taxon per site per year. This is useful when national decision regarding individual taxa need to be made under different conditions as represented by years.

A.3 SMA: effects for taxa within the same functional group are combined per site per year using a meta-analysis. This assumes that a negative effect for a taxon can be compensated by a positive effect for another taxon within the same functional group. Effects with large standard errors, e.g. due to low abundances, have a lower weight in the meta-analysis. This implies that the overall effect is dominated by effects with small standard errors and these are generally taxa with high abundances.

A.4 EA_{all} the combined effects for the functional groups are first evaluated for each combination of sites and years. This involves combining the Concern Quotient (CQ) derived from step A.3 over the functional groups. This would give a single result for each site for each year and national decision could be based on that.

A.5 EA_{all} these combined CQs are then assessed over years for each site.

A.6 EA_{all} and finally the CQ for sites are combined into a single judgement.

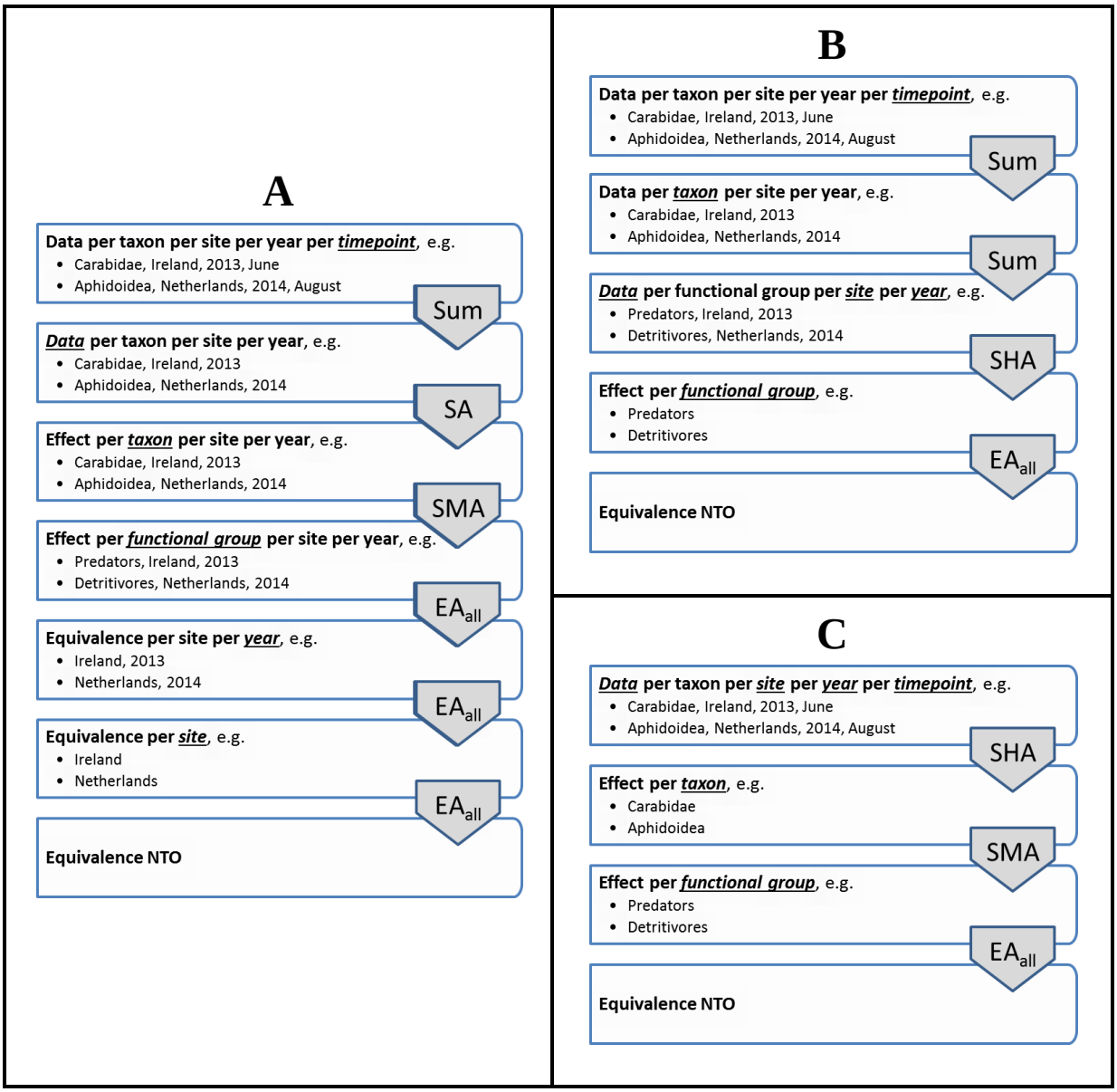

Figure 20. Examples of logical trees for the analysis of NTO data for potato field trials in Ireland and the Netherlands in 2013 and 2014.. Sum = summation of data. SA = statistical analysis. SHA = statistical hierarchical analysis. SMA = statistical meta-analysis. EAav = equivalence analysis with averaging of concern quotients. EAall = equivalence analysis requiring all concern quotients to be within limits.

For hierarchy B in Figure 20 the steps are as follows

B.1 SHA: this is identical to step A.1 described above.

B.2 SUM: the taxa are further summed to form counts for functional groups. This implicitly assumes that individuals of different species within the same functional group are equally valuable. It also presumes that there is no interest in individual taxa.

B.3 SHA: a statistical hierarchical analysis is performed to estimate the effect for each functional group while averaging over years and sites. This implicitly assumes that there is only interest in a cross-environment estimate of effects, and that negative effects in one environment can be compensates by positive effects in another environment. It also assumes that national decisions are not of interest.

B.4 EA_{all} the CQ obtained in the previous step is assessed over functional groups

For hierarchy C in Figure 20 the steps are as follows

C.1 SHA: a statistical hierarchical analysis is performed to estimate the effect for each taxon while averaging over time-points, years and sites. This implicitly assumes there is only interest in a cross-environment estimate of effects, and that negative effects in one environment can be compensates by positive effects in another environment. It also assumes that national decisions are not of interest.

C.2 SMA: effects for taxa within the same functional group are combined. This assumes that a negative effect for a taxon can be compensated by a positive effect for another taxon within the same functional group. Effects with large standard errors, e.g. due to low abundances, have a lower weight in the meta-analysis. This implies that the overall effect is dominated by effects with small standard errors and these are generally taxa with high abundances.

C.3 EA_{all}: the CQ obtained in the previous step is assessed over functional group

6.3.2 Example analysis for NTOs in potato trials

The analysis according to hierarchy A provides the most detail and is presented below. In step A.1 the counts for each taxon are summed over the time-points. Two pitfalls in the Irish trial in 2013 were missing at the second time point. To enable summing over time points, these missing count were imputed using the log-linear model "Block + Treatment" for the time point in question. The same was done for a single missing pitfall at the second time point in the Dutch trial in 2014. For the Dutch 2013 trial 13 out of 63 traps were missing for the first and third time point. Therefor for this trial the first and third time point were discarded.

6.3.2.1 Effects per taxon per site per year

In step A.2 each taxon was statistically analysed separately for each experiment. The statistical model "Block + Treatment" results in an estimate of the log-ratio for the GMO-IPM versus CMP-Weekly comparison. However when there is no interaction between Variety and the Spraying treatment, the effective level of replication can be increased by a factor of three by investigating the difference between the GMO and CMP averaged over the three Spraying treatments. This can be accomplished by fitting the main effects model "Block + Spraying + Variety". It is customary to use this main effects model in case the interaction is not significant. However the interaction between Variety and Spraying has four degrees of freedom and also involves the additional variety SarpoMira which is of no interest for the main comparison between the GMO and CMP. So it is possible that an interaction between Spraying and the GMO/CMP is swamped by complete absence of an interaction with SarpoMira or the other way around. This problem can be settled by excluding the additional variety from significance testing of the interaction. The remaining interaction is then between GMO/CMP on the one hand and Spraying on the other hand. Moreover, the Spraying treatment None can be fully responsible for the remaining interaction in which case we would like to compare the GMO and CMP averaged over the two Spraying treatments IPM and Weekly. These considerations were formalized in the following procedure:

1. Test for the interaction between GMO/CMP and Spraying (with three levels) which has two degrees of freedom. In case this interaction is not significant compare the GMO and CMP averaged over the three Spraying levels. Otherwise go to 2.
2. Test for the interaction between GMO/CMP and the Spraying levels Weekly and IPM; this interaction has one degree of freedom. In case this interaction is not significant compare the GMO and CMP averaged over the two Spraying levels Weekly and IPM. Otherwise go to 3.
3. Fit the full model "Block + Treatment" and compare GMO-IPM vs CMP-Weekly.

When either the GMO-IPM or the CMP-Weekly treatment has a zero mean count, the estimated log-ratio, i.e. $\Delta = \log(Q)$, equals plus or minus infinity and a confidence interval for the ratio cannot be constructed. For many of these cases both mean counts will be small. However a zero mean count for the GMO can be combined with a large mean for the CMP or vice versa. For these situations the zero mean count is replaced by the smallest positive mean possible and the ratio, without a confidence interval is calculated. The smallest positive mean possible equals 1 over the number of replications, e.g. 1/6 for 6 replications. The calculated ratio can be interpreted as a lower bound for the true ratio. In many cases this results in a small ratio which is of no interest. Therefore only calculated ratios outside the LoCs will be signalled in the graphical displays. Limits of concern were tentatively set to 0.5 and 2 for each taxon and the logarithm of LoC was multiplied by $\sqrt{10/m}$ whenever the combined mean m of the GMO and CMP is below 10. The confidence interval for each effect with the associated LoCs are given in Figure 21 to Figure 24. Note that a confidence interval for each functional group is also given; this is for the sum over the taxa within each group. There are two abundant species with an estimated effect which is outside the tentative LoCs: Poduromorpha in IR-2013 and Mesostigmata in IR-2014. Most intervals fall completely within the LoCs.

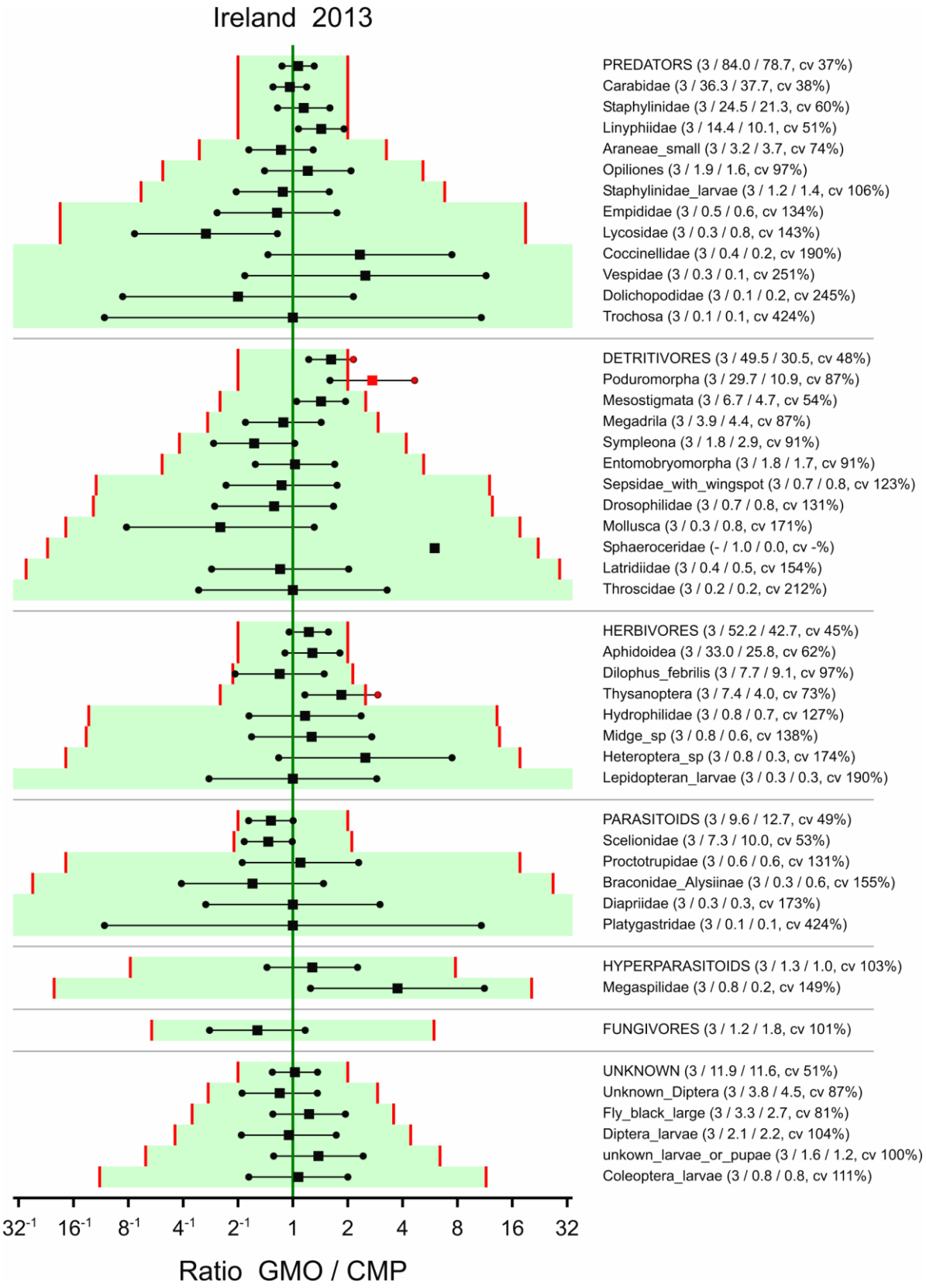

Figure 21. Arthropods in potato trial in Ireland 2013. 90% confidence intervals for the ratio between GMO and CMP averaged over Spraying treatment if possible. The number of Spraying treatments over which is averaged, the means for the GMO and CMP and the *CV* are added in parenthesis. Limits of Concern equal 0.5 and 2 and log(LoC) is scaled by $\sqrt{\boldsymbol{10/m}}$ for combined means *m* lower than 10.

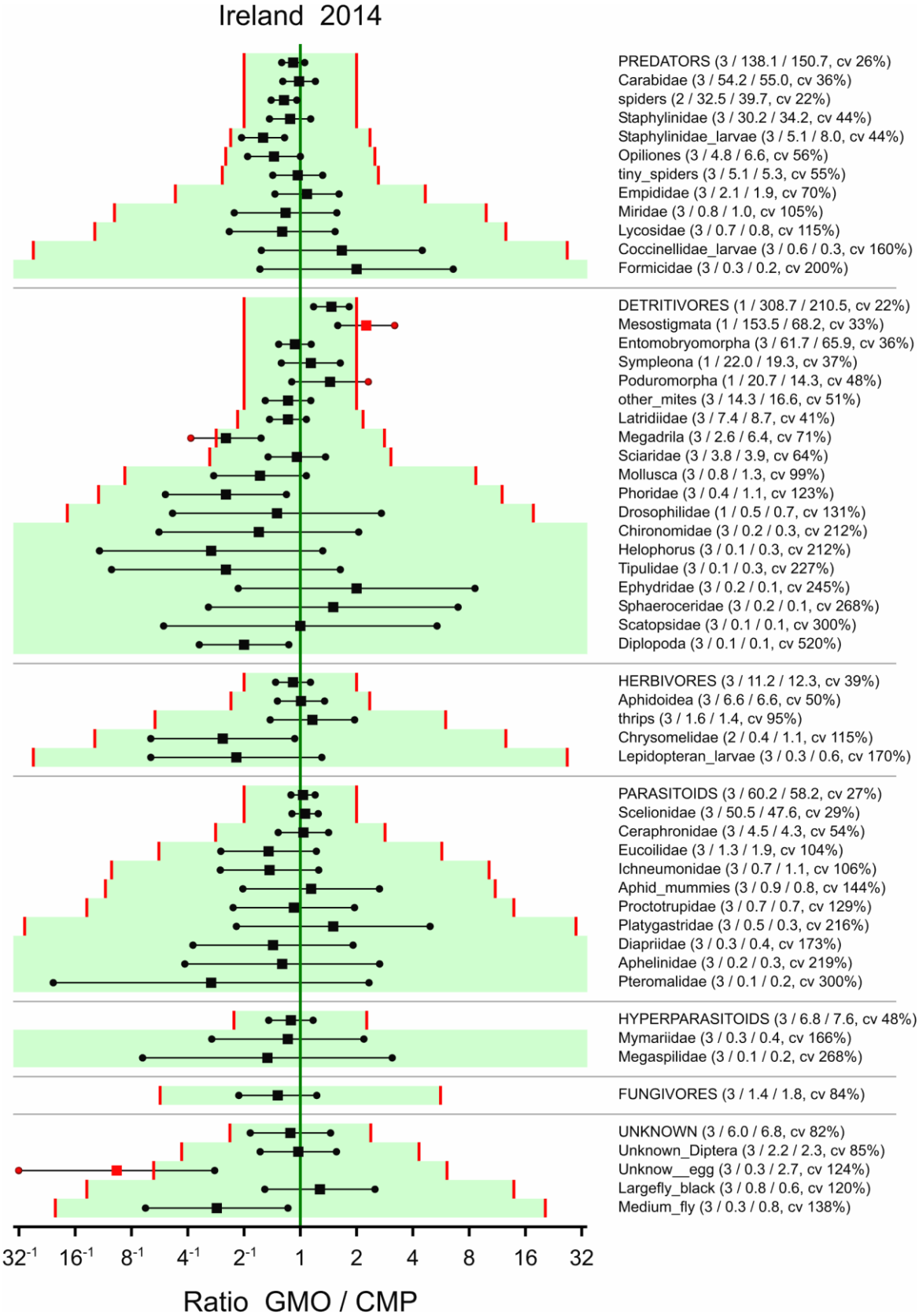

Figure 22. Arthropods in potato trial in Ireland 2014. 90% confidence intervals for the ratio between GMO and CMP averaged over Spraying treatment if possible. The number of Spraying treatments over which is averaged, the means for the GMO and CMP and the *CV* are added in parenthesis. Limits of Concern equal 0.5 and 2 and log(LoC) is scaled by $\sqrt{\boldsymbol{10/m}}$ for combined means *m* lower than 10.

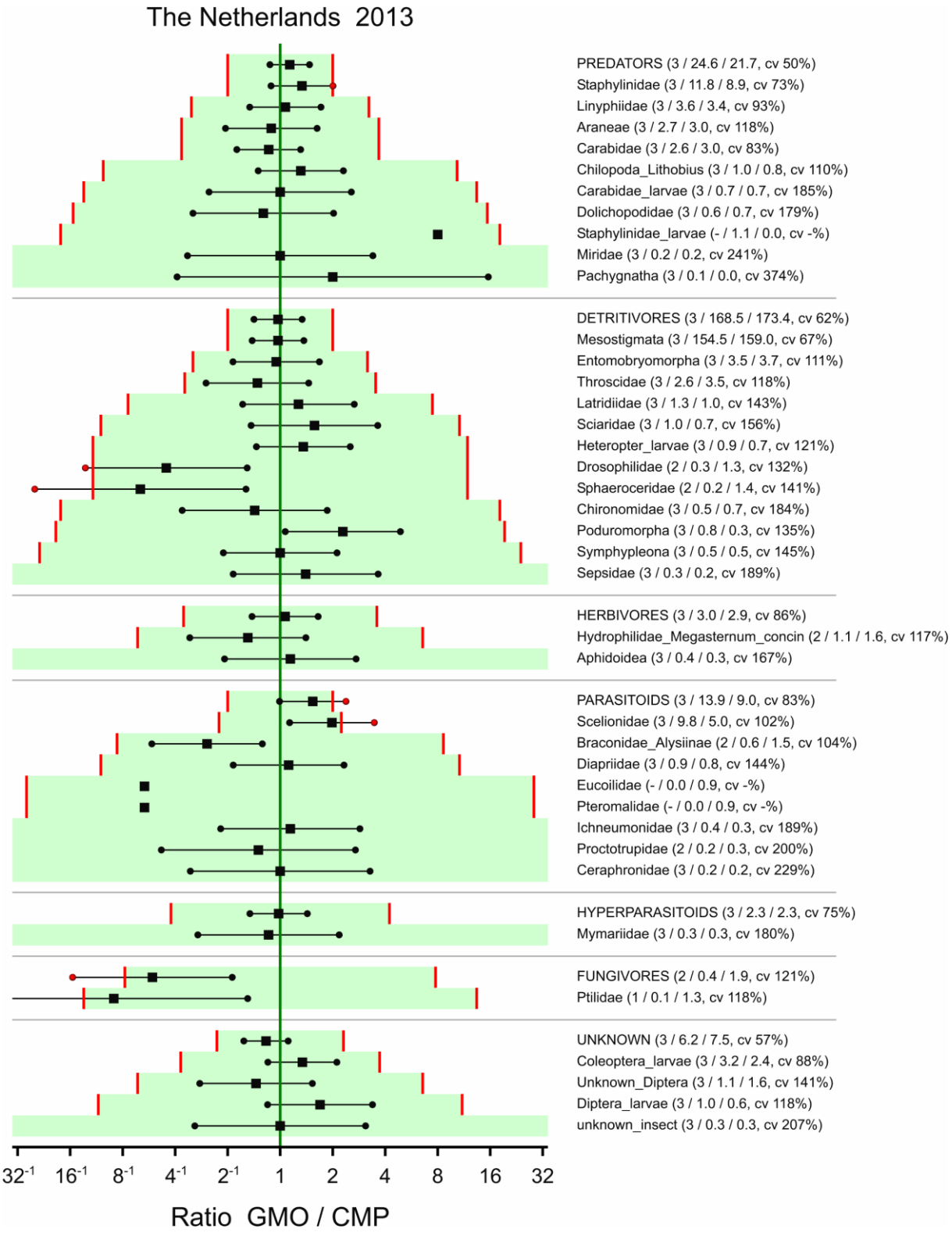

Figure 23. Arthropods in potato trial in The Netherlands 2013. 90% confidence intervals for the ratio between GMO and CMP averaged over Spraying treatment if possible. The number of Spraying treatments over which is averaged, the means for the GMO and CMP and the CV are added in parenthesis. Limits of Concern equal 0.5 and 2 and log(LoC) is scaled by $\sqrt{\boldsymbol{10/m}}$ for combined means *m* lower than 10.

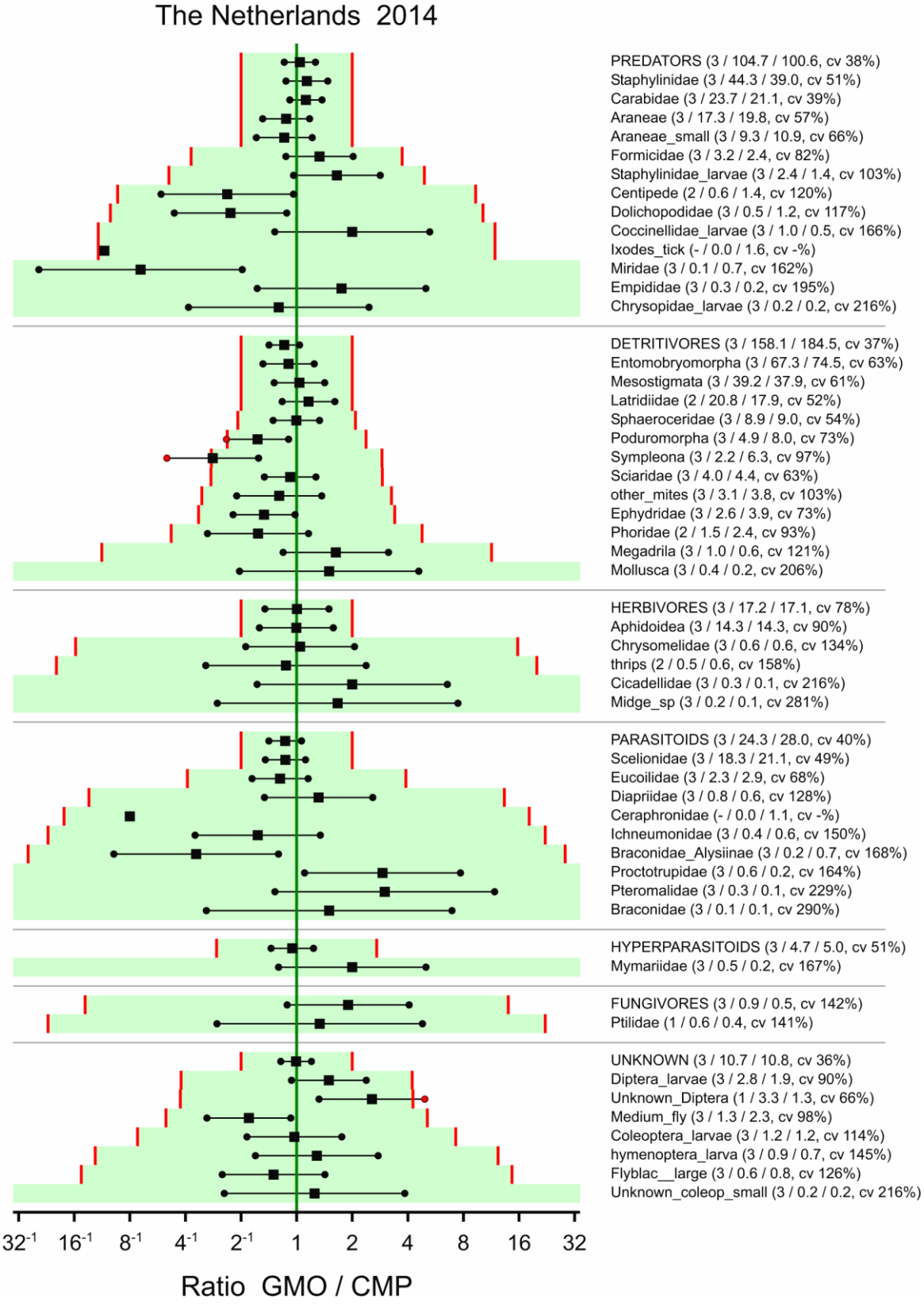

Figure 24. Arthropods in potato trial in The Netherlands 2014. 90% confidence intervals for the ratio between GMO and CMP averaged over Spraying treatment if possible. The number of Spraying treatments over which is averaged, the means for the GMO and CMP and the *CV* are added in parenthesis. Limits of Concern equal 0.5 and 2 and log(LoC) is scaled by $\sqrt{\boldsymbol{10/m}}$ for combined means *m* lower than 10.

6.3.2.2 Integrated analysis

Step A.3 involves a meta-analysis for each functional group for each site/year combination. This is not useful when there are only a few taxa within a functional group. The meta-analysis was therefore only carried out for those functional groups with 4 or more species. In case the functional group has three or less species the estimated effect for the sum was taken. This was thus done for Hyperparasitoids and Fungivores in all four experiment and Herbivores in The Netherlands 2013. Limits of concern for estimated overall effect for the meta-analysis were again tentatively set to 0.5 and 2, and for the estimated effect for the sum as before. Confidence intervals are given in Figure 25. All intervals are within the LoCs except for the Fungivores interval for the trial in The Netherlands in 2013. Note that this involves very few individuals and also note that in the Netherlands 2014 the estimated effect for Fungivores has an opposite sign.

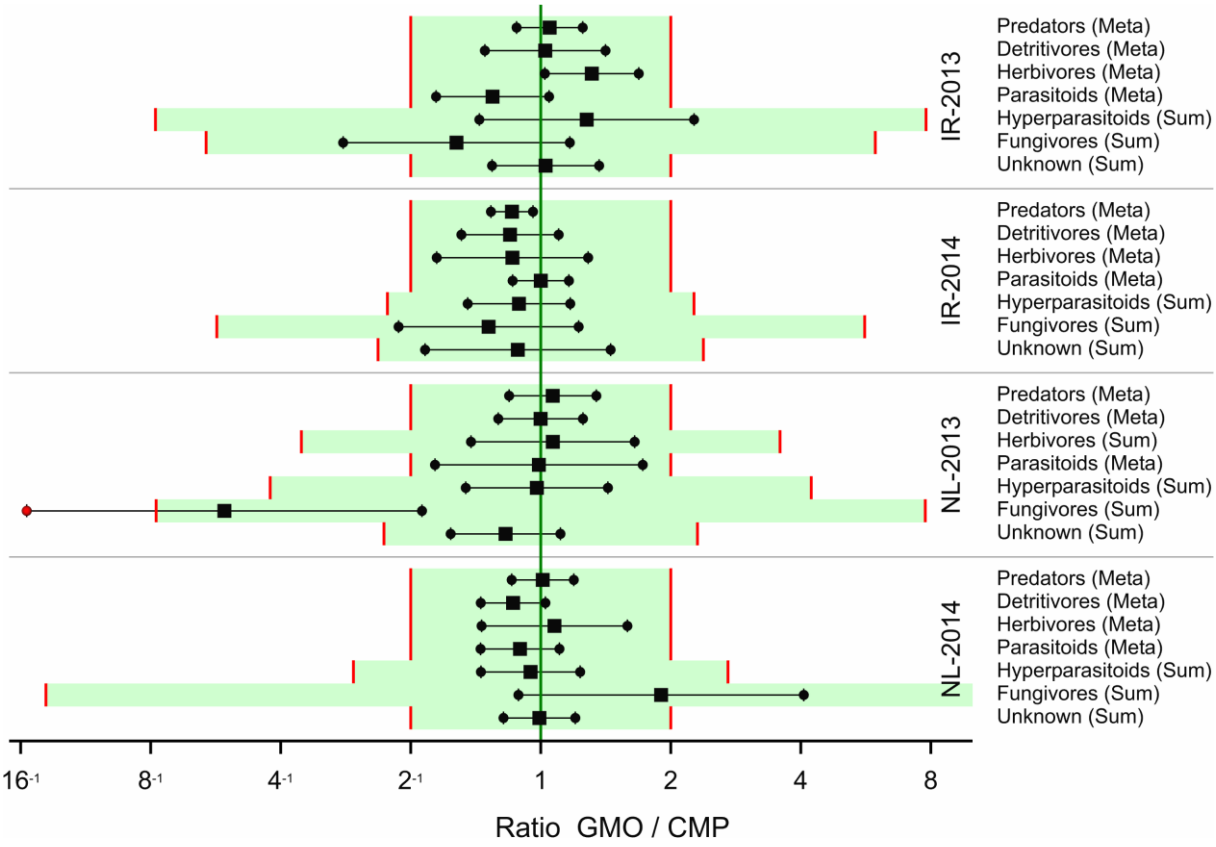

Figure 25. 90% confidence interval resulting from a meta-analysis for functional groups for arthropods data in potato field trials. Meta-analysis is only performed for those functional groups with 4 or more taxa. For other groups the interval for the sum counts is given.

An equivalence analysis (E_{all}) on the estimated effects would reveal no concern because all the estimated effects are within the LoCs. In a worst case scenario there would however be concern with respect to Fungivores in the Netherlands 2013 trial and no concern in Ireland and the Netherlands 2014.

7 References

Arpaia S, Messéan A, Birch NA, Hokkanen H, Härtel S, van Loon J, Lövei GL, Park J, Spreafico H, Squire GR, Steffan-Dewenter I, Tebbe C, van der Voet H (2014). Assessing and monitoring impacts of genetically modified plants on agro-ecosystems: the approach of AMIGA project. Entomologia, 2: 154. http://dx.doi.org/10.4081/entomologia.2014.154.

EFSA (2010). EFSA Panel on Genetically Modified Organisms (GMO). Guidance on the environmental risk assessment of genetically modified plants. EFSA Journal, 8(11): 1879. [111 pp.], doi:10.2903/j.efsa.2010.1879.

Goedhart PW, van der Voet H (2014). Environmental Risk Assessment of Genetically Modified Organisms: Simulation study to investigate properties of difference and equivalence tests. Deliverable 9.2b, AMIGA project, project number 289706. Available at http://www.amigaproject.eu/documents/deliverables.

Goedhart PW, van der Voet H, Baldacchino F, Arpaia S (2014). A statistical simulation model for field testing of non-target organisms in environmental risk assessment of genetically modified plants. Ecology and Evolution, 4: 1267–1283. http://dx.doi.org/10.1002/ece3.1019.

Hardy RJ & Thompson SG (1996). A likelihood approach to meta-analysis with random effects. Statistics in Medicine, 15, 619-629.

Lyles RH, Lin H-M & Williamson JM (2007). A practical approach to computing power for generalized linear models with nominal, count, or ordinal responses. Statistics In Medicine, 26(7): 1632-1648.

McCullagh P & Nelder JA (1989). Generalized Linear Models, second edition. Chapman and Hall. London.

O'Hara RB & Kotze DJ (2010). Do not log-transform count data. Methods in Ecology and Evolution, 1(2): 118-122.

Perry JN, Rothery P, Clark SJ, Heard MS & Hawes C (2003). Design, analysis and statistical power of the Farm-Scale Evaluations of genetically modified herbicide-tolerant crops. Journal of Applied Ecology, 40: 17-31.

Perry JN, ter Braak CJF, Dixon PM, Duan JJ, Hails RS, Huesken A, Lavielle M, Marvier M, Scardi M, Schmidt K, Tothmeresz B, Schaarschmidt F & van der Voet, H (2009). Statistical aspects of environmental risk assessment of GM plants for effects on non-target organisms. Environmental Biosafety Research, 8: 65-78.

Prasifka JR, Hellmich RL, Dively GP, Higgins LS, Dixon PM, Duan JJ (2008). Selection of nontarget arthropod taxa for field research on transgenic insecticidal crops: using empirical data and statistical power. Environ Entomol. 2008 Feb;37(1):1-10.

Schuirmann DJ (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. Journal of Pharmacokinetics and Biopharmaceutics, 15(6): 657-680.

Sileshi G (2006). Selecting the right statistical model for analysis of insect count data by using information theoretic measures. Bulletin of Entomological Research, 96: 479-488.

Szöcs E & Schäfer RB (2015). Ecotoxicology is not normal. Environmental Science and Pollution Research, 22(18): 13990-13999.

van der Voet H, Goedhart PW (2014). Environmental Risk Assessment of Genetically Modified Organisms: Statistical aspects of a protocol for single-environment GMO field studies. Deliverable 9.2a, AMIGA project, project number 289706. Available at http://www.amigaproject.eu/documents/deliverables.

van der Voet H, Goedhart PW (2015). The power of statistical tests using field trial count data of non-target organisms in environmental risk assessment of genetically modified plants. Agricultural and Forest Entomology 17: 164–172. http://dx.doi.org/10.1111/afe.12092.

van der Voet, H., van der Heijden, G.W.A.M., Kruisselbrink, J.W., Tromp, S.O., Rijgersberg, H., van Bussel, L.G.J., van Asselt, E.D., van der Fels-Klerx, H.J. (2014). A decision support tool for assessing scenario acceptability using a hierarchy of indicators with compensabilities and importance weights. Ecological Indicators, 43: 306-314. http://dx.doi.org/10.1016/j.ecolind.2014.02.022.

Ver Hoef JM & Boveng PL (2006). Quasi-poisson vs. negative binomial regression: How should we model overdispersed count data? Ecology, 88(11): 2766-2772.

VSN International (2012). GenStat for Windows 15th Edition. VSN International, Hemel Hempstead, United Kingdom. Web page: www.GenStat.co.uk.