tier3 solutions

Kolberger Str. 61-63
51381 Leverkusen
Germany

Background Information for the Revision of the Guidance Document Risk Assessment for Birds and Mammals

– Based on experiences of daily practical work on higher tier risk assessments and field studies –

Authors

Dr. Christian Wolf, Dr. María M. Benito, Ralf Dittrich, Dr. Olaf Fuelling, Dr. Benedikt Giessing, Dr. Ines Hotopp, Markus Persigehl, Dr. Andrea Rossbach, Dr. Anja Russ

2018-12-19

Table of Contents

List of Figures

List of Tables

List of Appendix Tables

1 Introduction

1.1 About tier3 solutions GmbH

The tier3 solutions GmbH is an independent and privately owned GLP-certified contract research organization (CRO), offering a competent and adaptable service portfolio for the environmental safety of agrochemicals. Since founding of the company in 2011, tier3 solutions GmbH has continuously extended its scope of work and expertise. Beside the regulatory affairs and science support, tier3 solution GmbH is specialised to execute higher tier GLP-studies for the risk assessment of terrestrial invertebrates and vertebrates. The claim of our experts is to improve the exploratory power of field studies through scientifically accepted and tested methods. Therefore, our experts of the field team are working strongly together with our statisticians, modellers and the regulatory affairs team.

1.2 About this document

The regulatory-, risk assessment-, statistician/modelling- and field-team of tier3 solutions is working on a daily basis (some colleagues for decades) with the respective guidance on risk assessment for birds and mammals. Some members of the team were already involved in the development of current and former guidance documents. This experience was used to extract thoughts, ideas and concepts which may serve as a pool of information potentially useful for the team of experts working on the revision of the actual EFSA Guidance Document Guidance Document on Risk Assessment for Birds & Mammals. We have collected our actual ideas on further methods but also parts of our currently used toolbox for field studies and data analysis. We hope that this collection can provide some help and background for the development of future guidance for bird & mammal risk assessments.

2 Higher tier risk assessment – proposals to the refinement steps

2.1 Identification of focal species

Authors: María M. Benito and Ralf Dittrich

2.1.1 Introduction and context

The EFSA (2009) Guidance Document for Birds and Mammals (EFSA (2009) GD), in the context of higher tier risk assessment of PPPs, offers the possibility to refine the exposure element by using a "focal species" (FS), whenever an active substance fails when the 'generic focal species' is used. Thus, the current text provides guidance on the methods used to identify this real species, and establishes the basis for a correct selection.

The agricultural landscape holds a wide range of both bird and mammal species that may be exposed by the use of PPPs. There is, however, a great variation in the use of agricultural land by different species. Some species live their entire life in agricultural habitats while others are present only during breeding, wintering or migration. Another important factor in determining the presence and the densities of birds and mammals is the actual crop. Wildlife preference for different crop types varies between species, geographical areas and seasons. Therefore, the EFSA (2009) GD established some criteria in order to be able to select relevant standard species for higher tier risk assessment.

However, while the background and foundations of this refinement option are well established in the EFSA (2009) GD, the experience accumulated during the last 10 years has identified gaps in the existing procedures and, therefore, in the usefulness of the results for an actual refinement. Based on that information, this is now a unique opportunity to improve the methodology and implement solutions to certain observed flaws.

2.1.2 Identification of issues in the EFSA (2009) GD and proposals for improvement

The selection of focal species is the first key point in all the procedures intended to refine the exposure element of the higher tier risk assessment, since they are the most appropriate species for further options such as e.g. radio-tracking and dietary studies. Therefore, a special attention should be placed into the revision of focal species in the new guidance document; it could be used to clarify several issues that have arisen in the experimental practise of the last decade.

In order to improve the degree of realism added to the risk assessment, an overview is presented below of the most important issues identified, a clear definition of their problematic as well as our suggestions for improvement.

Table 1: Focal species selection: The most important issues identified and our suggestions for GD improvement

Issue	GD says:	Problem definition	Action / GD revision proposal
Concept of FS	"…species that actually occurs in the crop when the pesticide is being used" (6.1.3, p.85)	"Occurrence" in the crop does not specify behaviour, which may strongly influence the exposure level.	Regarding exposure of species, focus should be, in a crop with - Spray application: on all species present in the crop, regardless of behaviour - Seed treatment/Pellets: particularly on species foraging in the crop (so far, the parameter "percentage of foraging individuals" is barely considered)
		Distribution and density of potential focal species at the intended GAP sometimes not taken into account.	**Study area selection** should be based on ornithological / mammalogical literature sources: - expected species and their spatial and temporal distribution at the GAP time. E.g. winter distribution vs. breeding distribution - preference for areas with known high diversity of species
	"…representative of all other species from the feeding guild" (6.1.3, p85)	FS sometimes not accepted as protective for other species within the regulatory authorities. How to deal with the request for 'other potentially more sensitive species', when actually no other species are known to occur or when information on which species would be more sensitive is missing?	Reinforce and clarify the standards given in the GD. Encourage the comparison with data from equivalent studies (Lahr et al. 2018). In other higher tier studies (e.g. acute effects): justify the selection of focal species with previous FS/pilot studies, and/or additionally with scientific literature on the biology and abundance of the species.
Field selection	"…the appropriate crop, its correct growth stage and at a time of the year that is relevant to the proposed use." (6.1.3.1, p.85)	The observed densities of birds/mammals are low in many crops and BBCH stages (e.g. bare soil) and hence criticised as insufficient data in the study evaluation process.	The attractiveness of each target crop as foraging/breeding habitat is highly dependent on its conditions (plant structure and food availability). For instance, low bird numbers in the bare soil period is a likely and realistic outcome for some crops,

Issue	GD says:	Problem definition	Action / GD revision proposal
Methods			not due to incorrect study design or field selection. This can be acknowledged by comparing results (e.g. these compiled by Lahr et al. 2018) of crops at same BBCH with equivalent crop structure when season is similar. Thus, the **selection of FS** can be supported by the outcome of FS studies with comparable **crop structure**. For example for maize, sugar beet and potato, crops with low attractiveness to birds at 1. Bare soil period (BBCH 00 - 09) 2. Period BBCH 10-19
	"…necessary to have a range of fields that are representative of where the pesticide is used…" (Appendix M, p.1)	Influence of the agricultural practices and/or former crop on the results. For example, for the bare soil period (i.e. freshly drilled fields), minimum tillage cultivation techniques (harvest remains still available) and/or former crops with especially attractive seeds (e.g. oilseed rape) can increase bird presence (relevant in the case of spray application) but can lead to an overestimation of the real exposure (in the case of seed treatment/pellets).	The justification of the **study area / field selection** should consider the agricultural practice and the presence of former crop rests in the study fields. For many crops, minimum tillage and abundant harvest rests are representative in the European agriculture, and therefore they should be at least partially included in the selection of study fields. For example, select 50% of fields with such characteristics. Alternatively, a more detailed guideline or examples of good praxis should be included in the revision of the guidance document.
		Influence of landscape composition on the results.	**Field selection** should consider differences in habitat requirements (hedges/open landscape/forest) for different species. For example, for birds, select 50% of fields with and 50% without hedges at the border.

Issue	GD says:	Problem definition	Action / GD revision proposal
Methods	"…essential to ensure that there are sufficient sites visited." (6.1.3.3, p. 86)	Current guidance document lacks information/proposal regarding sample size of sites (in order to provide representative results).	Minimum number of 20 fields is recommended, visited at least twice (Gregory et al. 2004). Smaller sample size or just one visit would be possible for small mammals or if justified by other already available information in literature and/or former studies.
	"Cropping details and surrounding habitats should be included in the final report." (Appendix M, p.1)	Need of a harmonised method to report and consider such factors in a final analysis (Lahr et al. 2018).	**Cropping details** should allow the categorization of the field by its cultivation technique or former crop (see above, "Study area/field selection"). **Surrounding habitats** should be recorded in a way that allows categorization into landscape type but also to allow potential further analysis (see below, "Analysis"). **Farming practices** at least during the study period should be recorded and presented, to allow characterisation of representativeness of GAP.
	"The identification of focal species using targeted observation data can involve one of two methods, i.e. the transect method and the field survey method." (6.1.3.1, p.85) "Survey techniques: Basically there are two techniques for birds – namely the transect method and the field survey (point count)	Field methods should be adapted to the time of the year and the habitat and not applied arbitrarily.	The decision about the appropriate method for focal species selection is crucial because the probability to detect species with respect to the method applied differs regarding the conditions of the crop. Therefore, the **crop structure** at a specific time defines the most suitable method. See Table A 1 - Table A 4 for exemplified proposals. Additionally, see Table 2 and Table 3 for a summary of key points for possible methods.

Issue	GD says:	Problem definition	Action / GD revision proposal
Methods	method." (Appendix M, p.2)		
	Identification of focal mammalian species: no guideline provided.	Not all the mammalian species listed in Appendix A occur all over Europe. Sometimes they are replaced by another one without additional data from literature and/or field studies.	Information should be provided on similarities and differences in foraging behaviour and habitat preferences of e.g. : striped field mouse (*Apodemus agrarius*), Savi's pine vole (*Microtus savii*) or Iberian hare (*Lepus granatensis*)
Analysis	"…their surrounding habitats (e.g. what crops were being grown, presence of woodlands, hedgerows etc.) should be included in the final report." (Appendix M, p.1) "Justification…on a comparison of agricultural landscape including size of fields, presences of hedgerows, field boundaries as well as climatic conditions." (6.1.3.2, p.86)	Need of a harmonised method to consider environmental factor in a final analysis (Lahr et al. 2018).	Description of field surroundings up to a buffered distance of e.g. 100 m from the field limits, and presentation of resulting areas of each habitat type. Such information would allow comparisons of area representativeness for extrapolation among different MS and if required, further statistical analysis. A key code with clear definitions of used habitats should be also present, for comparison reasons.

Selection of FS	"…species with FO >20% considered to be of high priority... However before deciding which species 'covers' all other species, it is necessary to consider issues such as … body weight…". (Appendix M, p.3)	Frequency of observation vs. body weight: How to take a final decision when a more common species is heavier than a rarely observed but lighter species? (Rarity could indicate that the crop may be not relevant for the population of this species, especially if it is otherwise common in the agricultural landscape.)	Smaller species are to be considered the worst case, as higher body weight lowers the estimated theoretical risk. However, there are special situations in which a case-by-case approach might be more appropriate, if properly justified: whenever the heavier species fulfils all the other criteria required by the guidelines (correct feeding strata, feeding guild, etc.), it may be more accurate and protective to use the more frequent species as focal species, if it can still be considered to show a higher potential exposure than the rarely observed but otherwise common species.

Table 2: Summary of key parameters for FS methods (birds)

	Transect count	Point count (Scan sampling)	Mist-netting
Method description	Counts of birds observed or heard inside a pre-defined area to both sides (or only one side) of a track walked by foot through (or alongside) the crop	Counts of birds observed or heard inside a pre-defined area from an observation point (e.g. car serving as a hide) located next to the crop	Counts of birds trapped in fine invisible vertical nets (mist nets) set up between poles within the crop. Additionally, possible to mark trapped birds with individually numbered metal rings, and to measure, age and sex them before release
Main endpoints	Species composition No. of individuals/species Abundance (individuals/ area/species) Frequencies of occurrence	Species composition No. of individuals/species Abundance (individuals/ area/species) Proportion of foraging individuals (behaviour observations) Frequencies of occurrence	Species composition No. of individuals/species 'Relative abundance' (No. of individuals/net-length/trapping time) Frequencies of occurrence
Number of sampling units (fields, plots...)	20	20	20
Sampling strategy	Recommended: stratified by habitat (e.g. half fields with hedges and half without)		
Distance between units	Min. 250 m		
Number of sessions per sampling unit	2 (max. 4)	2 (max. 4)	2 (max. 4)
Duration	One transect length (200 - 300m) Defined walking speed	4 hours/session, 10 min. count periods plus initial settling time of 5 min.	4-5 hours
Observation period	Preferably early morning and/or evening (recommended: one session at each time)		Preferably early morning
Additional	Can be combined with "Territory mapping technique", for breeding birds	Birds flushed when approaching the point, recorded separately but included	Nets to be checked every 30 to 60 minutes
	Record some measure of distance to each bird ("Distance sampling technique": for detection probability and estimation of bird densities) Record whether detection was by sight or sound Aerial species (birds flying over census area), consider how to treat them		Net height: as high as the crop, covering the entire range between ground and top of crop for species using the uppermost or lowest part

Table 3: Summary of key parameters for FS methods (mammals)

	Transect count (Spot-light count)	Point count (Scan sampling)	Live trapping
Method description	Counts of lagomorph species on a landscape level, made from a slow moving car in darkness. A strong spot-light is directed towards the fields. Lagomorph eyes reflect the light and can be counted per habitat (crop). Night-vision devices can be also used instead of spot-lights.	Counts of mammals at night inside a pre-defined area from an observation point located next to the crop using night-vision devices. The method is ideal for medium sized species and limited (but not impossible) for small species.	Live trapping of small mammals (rodents and shrews) during the night. Especially in early crop stages, set traps in-crop to identify the focal species and off-crop to proof the occurrence of potential focal species. In addition, individuals can be marked, and sex, body weight and repro-condition recorded.
Main endpoints	Abundance (individuals/ area/species) Frequencies of occurrence No. of individuals/species (method usually applied on hares)	Species composition No. of individuals/species Abundance (individuals/ area/species) Proportion of foraging individuals (behaviour observations) Frequencies of occurrence	Species composition No. of individuals/species 'Relative abundance' (No. of individuals/trap-night) Frequencies of occurrence
Number of sampling units (fields, plots...)	≥20	20	20
Distance between units	Min. 250 m		
Number of sessions per sampling unit	2 (max. 4)	2 (max. 4)	2 (max. 4)
Duration	One transect length depends on the frequency of the target crop Recommended driving speed (approx. 5 km/h)	4 hours/session, 10 min. count periods plus initial settling time of 5 min.	From dusk to dawn. If the trapping of shrews is intended, care must be taken to check the traps frequently.
Observation period	Spot-light and night-vision observation work only during darkness which limits the observation period in mid-summer.		Preferably, sunset to sunrise (most species are circadian, crepuscular or night-active).

2.1.3 Conclusion

In the context of PPPs risk assessment, using real focal species remains the unique way to test realistic exposure scenarios in the field. Therefore, the selection of focal species is an essential point for any further higher tier risk assessment procedure. Since the publication of the EFSA GD in 2009 providing a first guideline to that selection, broad experience has been accumulated; it would be advisable to use it to improve the methodology and implement solutions to some unclear issues.

2.1.4 References

EFSA 2009. European Food Safety Authority; Guidance Document on Risk Assessment for Birds & Mammals on request from EFSA. EFSA Journal 2009; 7(12):1438.

Gregory RD, Gibbons DW, Donald PF. 2004. Bird census and survey techniques. In: Sutherland W.J, Newton I, Green R.E, editors. Bird ecology and conservation: a handbook of techniques. Cambridge, UK: Cambridge University Press. pp 17–55.

Lahr J, Krämer W, Mazerolles V. et al. 2018. Data collection for the estimation of ecological data (specific focal species, time spent in treated areas collecting food, composition of diet), residue level and residue decline on food items to be used in the risk assessment for birds and mammals. EFSA Supporting Publications 15: 1513E.

2.2 Radio tracking studies and evaluating observational data (PT-factor)

Authors: Ralf Dittrich, Benedikt Giessing and María M. Benito

2.2.1 Introduction and outlook

PT studies became a classical method in the refinement of the higher tier risk assements (RA) of plant protection products (PPPs) since the introduction of the last version of the EFSA (2009) GD.

We aim to cover the following points: methodological aspects, an assessment of the variability of PT data and the utilisation of PT values for long term RA. Based on long-time experience, important points of the data collection will be considered and improvements in the revision of the EFSA (2009) GD proposed. Second, the utilisation of the generated PT values needs to discussed more closely. Of create importance is the recurrent consumer issue which is linked to the discussion about the correct way to estimate a long term PT (21-days or the toxicologically relevant time period). Additionally alternative ways to estimate a long term PT should be discussed.

Due to the complexity of this topic and the currently presented contributions to this topic (e.g. Ludwigs 2018, HSE postion paper 2018) we are still working on this topic and may provide it beginning 2019 separately.

2.2.2 References

EFSA 2009. European Food Safety Authority; Guidance Document on Risk Assessment for Birds & Mammals on request from EFSA. EFSA Journal 2009; 7(12):1438.

Ludwigs J-D. 2018. Appropriate PT estimates for vertebrate risk assessment – what does radio-tracking actually reveal?. Platform presentation at 18th International Fresenius ECOTOX Conference, Mainz.

HSE postion paper on the use of Monte-Carlo simulated PT value. 2018. Chemical Regulation Division, UK Health and Safety Executive. Distributed at 18th International Fresenius ECOTOX Conference, Mainz.

2.3 Information on composition of vertebrate diet (PD-factor)

Author: Benedikt Giessing

2.3.1 Introduction and context

The assessment of the proportion of food types birds obtain from treated areas (PD) can still be treated as a significant refinement tool. In most cases this approach offers quite detailed results but requires only a comparable moderate investment. The information given in the EFSA (2009) GD (Appendix Q) regarding PD is still valid. Hence, the main intend of this document is to give additions to this document.

Faeces analysis are still treated as the most appropriate and less invasive approach in order to assess the PD for a certain species in a treated area. In order to improve the value of this approach it is suggested how the likelihood that food items found in faeces samples of a selected species originate from the focal crop.

A considerable issue of deriving proportions of different diet types originally ingested from their content in faeces is differential digestibility of diet types, i.e. the proportions ingested differ from their remains found in the faeces. However, this issue can best be overcome by calibration trials with captive birds (Southerland 2004). The generation of food type specific correction factors that can be applied to the quantity of remains found in the faeces samples to calculate the proportion ingested is the aim of these trials. These correction factors may be transferable to closely related species. Hence, it might not be necessary to conduct calibration trials for all bird species in order to get correction factors for different diet types. However, especially for bird species that belong to different feeding guilds the transfer of correction factors is awkward. In the current literature correction factors for only a few bird species and only a moderate selection of food types are published. In order to be able to make use of faeces samples for additional bird species, correction factors for these species should be derived using calibration trials. In order to encourage the conduct of calibration trials guidance how to conduct such trials is given in the appendix.

2.3.2 Identification of issues in the EFSA (2009) GD and improved proposals

Table 4: PD-factor: The most important issues identified and revision proposals

Issue	GD says:	Problem definition	Action / GD revision proposal
Collection of faeces	"If radio-tracking is applied simultaneously to the collection of diet samples, the source (e.g. a specific crop) of the food items found in the sample can be identified. Appendix Q. For collecting faeces, birds can be kept in a clean bird bag or held over a polythene sheet during handling (Sutherland, 2004). Droppings can often also be collected in the field, e.g. where birds perch, roost and at nests." Appendix Q.	It is not mentioned in the GD how the radio-tracking approach can be used to identify the source of the food items found in diet samples. Moreover, the approaches how to collect faeces given in the GD do not describe if and in which way identified food items can be assigned to a certain source. There is the need to give advice which basic requirements have to be met in order to link food items to potential sources	Radio-tracking can be used to support the identification of the source (e.g. a specific crop) of food items found in a diet sample if: (1) the radio-tagged individual has been tracked continuously insight the specific habitat (e.g. a specific crop) for (2) a period that exceed passage time of the majority of food types in the diet of the respective species (3) before it is observed defecating and (4) the respective dropping can be found. These requirements can also be adopted to purely visual approaches (i.e. without radio-tracking) (see 'Refined selection of faeces')

Issue	GD says:	Problem definition	Action / GD revision proposal
Derivation of Correction factors	"A considerable difficulty [of deducing diet proportions from the proportion of their remains found in faeces samples] is the differential digestibility of different food types. Calibration trials with captive birds can help to overcome this difficulty" that "few remains may be found either because few items were eaten or because food items were almost completely digested" (Appendix Q)	No proposals or recommendations are given how 'calibration trials with captive birds' can be conducted in order to provide useful correction factors that can be applied to diet proportions found in faeces sampled in order to deduce their proportions originally ingested	In order to provide guidance how a study should be design in order to derive correction factors for different food types of a certain bird species an example is given (see appendix 'Suggestions for the approach to generate correction factors')
Methodo-logical prospects	The current guidance document lacks a promising approach for DNA-based diet analysis	Current scientific research has revealed a new method for diet analysis that has already been used to analyse the content of faeces. Since this technique was not available during the compilation of the GD, it is not mentioned there	Meta-barcoding is a promising new technique that can be used to identify the content of faeces using the DNA of ingested species. Currently effort is made in order to improve quantification of the faeces content.

2.3.3 Refined selection of vertebrate faeces

In order to increase the likelihood that food items found in the faeces of an individual of the selected species originate from the focal crop simultaneous radio-tracking was suggested in the guidance document. However, it is not mentioned how the radio-tracking approach can be used in order to link the food items found in faeces of a tracked individual to the data obtained by radio-tracking. To our understanding some conditions have to be met in order to use radio-tracking for identifying the source (e.g. a specific crop) of the food items found in the sample. Correct assignment of the food items to a specific habitat can be conducted if:

(1) the radio-tagged individual has been tracked continuously insight the specific habitat (e.g. a specific crop) for
(2) a period that exceed passage time of the majority of food types in the diet of the respective species
(3) before it is observed defecating and
(4) the respective dropping can be found

Moreover, under certain conditions even purely visual approaches can also be adopted to get faeces samples whose content reflects the diet selection in a specific crop. For example, if crop structure allows (e.g. freshly drilled cereal field) it may be possible to observe individuals of some species (e.g. skylarks) continuously in a field for a considerable period of time until they defecate and to find the respective faeces sample. Also in this case observation period has to exceed passage time of the majority of food types in the diet of the respective species in order to allow for the conclusion that the food items in the droppings originate from this field. Similar conclusions may be possible for other conditions as well. For example finches or buntings, which are observed foraging in a field and change into a hedge adjacent to the field occasionally can be treated as obtaining their food mainly from the field. Hence, the content of faeces samples gathered from these individuals reflects primarily their food selection inside the field.

These two examples (radio-tracking and the 'visual' approach) were used to illustrate how faeces can be obtained that most probably contains food items taken in a certain habitat. However, there may be other prevailing circumstances that justify this assumption as well (e.g. faeces from bird species known to have small home ranges occurring in a crop that is surrounded by unsuitable habitat). Hence the aim of this section is mainly to illustrate that effort should be expended to justify that the content of faeces collected can be linked to a certain source (i.e. crop) or can be treated as being representative for a certain source.

2.3.4 A methodological prospect: Meta-barcoding

While visual analyses are highly labour intensive and may sometimes lack sufficient resolution, recent DNA-based approaches potentially provide more accurate methods for dietary studies. A suite of approaches have been used based on the identification of consumed species by characterization of DNA present in faecal samples. In one approach, a standardized DNA region (DNA barcode) is PCR amplified, amplicons are sequenced and then compared to a reference database for identification. The recent development of next generation sequencing (NGS) has made this approach much more powerful, by allowing the direct characterization of dozens of samples with several thousand sequences per PCR product, and has the potential to reveal many consumed species simultaneously

(DNA meta-barcoding). Continual improvement of NGS technologies, on-going decreases in costs and current massive expansion of reference databases make this approach promising.

However, given the fact that there still exists a number of potential biases even a well-designed dietary barcoding study is likely to only provide semi-quantitative data on the diet of a species (Pompanon et al., 2012).

2.3.5 References

EFSA 2009. European Food Safety Authority; Guidance Document on Risk Assessment for Birds & Mammals on request from EFSA. EFSA Journal 2009; 7(12):1438.

EFSA 2013. European Food Safety Authority; Guidance on tiered risk assessment for plant protection products for aquatic organisms in edge-of-field surface waters. EFSA Journal 2013; 11(7):3290.

Flinks H. 2013. Tatort Weidetor: Warum Kotanalysen für die Ökologie der Vögel wichtig sind. Falke 60:280-284.

Glutz von Blozheim UN, BAUER K. 1997. Handbuch der Vögel Mitteleuropas; Volume 14/III. Wiesbaden, Germany: Aula-Verlag.

Green RE. 1978. Factors affecting the diet of farmland Skylarks, *Alauda arvensis*. Journal of Animal Ecology 47:913-928.

Green RE, TYLER GA 1989. Determination of the diet of the stone curlew (*Burhinus oedicnemus*) by faecal analysis. J. Zool. 217:311-320.

Pompanon F, Deagle BE, Symondson WOC, Brown DS, Jarman SN, Taberlet P. 2012. Who is eating what: diet assessment using next generation sequencing. Molecular Ecology 21:1931-1950.

Sutherland WJ. 2004. Diet and foraging behaviour. In: Bird Ecology and Conservation. Oxford, UK: Oxford University Press: 233 -250.

2.4 Field effect studies to investigate acute risks for vertebrates

Authors: Ralf Dittrich, Ines Hotopp and María M. Benito

2.4.1 Introduction

Vertebrate risks assessments of plant protection products (PPPs) may indicate an acute risk to wild birds and mammals or predict effects on population development. This risk might be concluded from (too) conservative assumptions on the exposure side of the equation for the risk evaluation, due to the lack of better data. The EFSA (2009) GD on the risk assessment (RA) mentions that one option to demonstrate acceptable risk is to conduct so-called field effects studies (section 6.4, p. 101). General recommendations are given about the required study design but no detailed guidance, instructions or quality criteria are provided. Here, we want to highlight three complementary ways to improve the quality (and therefore the usefulness and acceptance) of acute field effects studies: combined extensive-intensive study design, specific tools for improved statistical evaluation and estimation of the power analysis.

An optimal study design combines the 'extensive' landscape approach that uses a broad geographical area or a high number of agricultural fields in different study sites, with the 'intensive' approach that uses radio-tracking techniques in a control/treatment design. This double approach covers the natural variation in parameter estimates and enables the identification of possible treatment effects. The radio-tracking technique is sensitive enough to monitor the fate of single individuals within a population over a long time period and to detect their carcasses in case of mortality. In comparison, carcass search as an alternative method is much less exact, because the number of exposed individuals is unknown and the actual detection rate of mortality is difficult to estimate. In addition, the area which needs to be covered by the carcass search is unknown.

Individuals which disappear without confirmed mortality, i.e. signal loss, are in most cases the critical point in the evaluation of radio-tracking study results. These losses are often regarded as undetected mortality events, even though field experience suggests that this is an unlikely possibility: acute mortality induced by PPPs normally occurs within the study area, which can be searched extensively in case of signal loss. More likely the signal losses derive from natural dispersal events (the tracked animal moved out of the area covered by the study) or in rare occasions, from radio-tag malfunction. In any case, these possible differences in the encounter rate between treatment and control sites can be always analysed statistically, given a sufficient sample size, in order to improve the value and credibility of the studies.

In the context of a good study design, we also propose an improved statistical evaluation that can considerably increase the detectability of real effects in comparison to earlier studies. The Kaplan-Meier survival curve and the Cox proportional-hazard model are nowadays standard and highly recommended methods for the analysis of survival data. The Kaplan-Meier estimator is used to estimate the survival function from lifetime data, and can be used to measure the fraction of individuals living for a certain amount of time after a treatment. The Cox model is a well-known statistical technique commonly used in medical research. It provides an estimate of the treatment effect on survival adjusted for other explanatory variables. Therefore, the effects of the treatment can be compared to the effects of other covariates and the assessment of the results is simplified.

Additionally, an essential information before the beginning of a study is the minimum number of individuals needed in order to detect actual treatment effects in the statistical analysis afterwards. Below, we present a case study based on data from real field studies, with the aim to provide guidance about this procedure. Specifically, data was obtained from generic radio telemetry studies on untreated populations of wood mouse and of several bird species. We have used statistical simulations to add acute effects for different scenarios of PPP effects to the treatment group. The first results showed that the minimum sample size is highly dependent, first, on the pattern of dispersive behaviour of each species at the respective time in the year, and second, on the action mode and persistence of residues of each PPP. Also, the required sample size was found to be reducible by increasing the general encounter rate via improvements in the field observation method. Further analyses were then oriented to the detection probability of treatment effects depending on: (i) number of individuals radio-tracked (ii) differences in presence between species and (iii) differences between action modes of PPPs.

2.4.2 Identification of issues in the EFSA (2009) GD and proposals for improvement

Table 5: Acute effect studies: The most important issues identified and our suggestions for GD improvement

Issue	GD says:	Problem definition	Action / GD revision proposal
Study method	"Note that, although the lack of vegetative cover makes it easier to find carcasses in newly sown fields, it may also make intoxicated animals more likely to seek cover away from the field." (5.2.3, p.60)	There is low acceptance of carcass search as a reliable method to quantify mortality. The detection probability depends on parameters which are difficult to estimate.	Radio-tracking should be always preferred over carcass search as a method to detect and quantify acute mortality following the application of PPPs.
	"The choice of methods and their detailed implementation in each case should be driven by the study objectives, including the type of effects that are of interest and the degree of certainty required in detecting and quantifying them." (6.4, p103)	It is not specified how to reach an agreement about the degree of certainty required. Therefore, the acceptance of effect studies is inconsistent between the risk managers.	A binding instruction should be given how to define the degree of certainty required according to the type of effect that is of interest. The studies would be designed and conducted accordingly and its acceptance considerably increased.
	"Radio-tracking to monitor activity and survival of tagged individuals (e.g. Prosser et al., 2006). The number of individuals should be sufficient to measure the level of mortality with the desired level of certainty…" (6.4, p. 103)	Radio-tracking is a highly efficient method to detect mortality. With a correct sample size, the survival rate is properly estimated as a combination of mortality and signal loss, due to natural dispersal.	Between 65 and 130 individuals in a treatment/control design are sufficient to reach a power >0.8 to detect a difference in survival rate of 20% (Dittrich et al. 2018).
Study design	"…classical field 'effect' studies can be used to refine assessments on the acute risk of seed treatments. Quality criteria should be applied to the studies regarding the relevance of the species that are present (e.g. diet, use of field), the representativeness	There is no detailed guidance how to apply quality criteria to such studies. Therefore, the evaluation differs between the risk managers.	An official checklist of quality criteria would facilitate the acceptance of field effect studies. Alternatively, examples of good praxis should be presented in the revised guidance.

Issue	GD says:	Problem definition	Action / GD revision proposal
	of the field situation and the power of the study to detect effects..." (5.2.3, p.60)		
	"Care is required to ensure that the methods chosen for detecting effects in field studies are appropriate to the study objectives and provide adequate statistical power to be useful for risk assessment and decision-making." (6.4, p.105)	It is possible to establish adequate study design based on expert advice. The feedback by the risk manager differs according to their interest and knowledge.	Examples, criteria checklists or detailed instructions would be helpful to reach a uniform evaluation of the studies conducted. Simulations, based on generic data, can be used to estimate the statistical power and should be encouraged.
Number of study sites/ fields	"An 'extensive' approach with multiple field study sites is recommended in preference to 'intensive' approaches where fewer sites are studied in more detail. More work (research and/or a workshop) would be desirable to develop guidance on how to determine an appropriate number of sites. In the meantime, expert statistical advice should be sought case-by-case on this issue." (6.4, p.104)	There is not enough guidance yet on how to determine an appropriate number of sites, according to the type of effects of interest and to the degree of certainty required in detecting and quantifying them.	Ideally, a combination of intensive and extensive approach should be used. An acute field study should be conducted in two regions with a high number of study fields: typically around 20-30 treatment/control fields in total. The exact number depends on the species and crop studied. This design covers the natural variability in exposure scenarios and a sufficient number of individuals can be monitored.

Table 6: Summarising the key parameters of field 'acute effects' study on birds or mammals

Method	Radio-tracking, combined with behavioural observations
Main endpoints	Survival of individuals until the end of the observation period (survival time) Occurrence of sublethal effects
Number of study fields	20 - 30 (10 - 15 control + 10 - 15 treated). Final number depends on density of the focal species. A surplus of control fields improves the statistical power. In high value crops, it might be impossible to find adequate control fields, so the design needs to be adjusted.
Number of regions	2
Number of individuals	65 – 130, according to expected survival time (mortality + signal loss) of the focal species
Observation period	Depends on the mode of action of PPP (min. 10 - 14 days)
Interval between survival checks	Directly after the application the animals should be checked on a daily basis, afterwards every second day can be sufficient
Statistical analysis	Kaplan-Meier survival curve and Cox proportional-hazard model

2.4.3 Material and methods

Two general scenarios were assumed in the timing of acute effects on the population in the first ten days after application of PPP. Scenario I was defined as an exponential decline with a reduction by 20% within ten days after application in the treatment group. Scenario II was defined as a linear reduction by 20% within ten days (Figure 1).

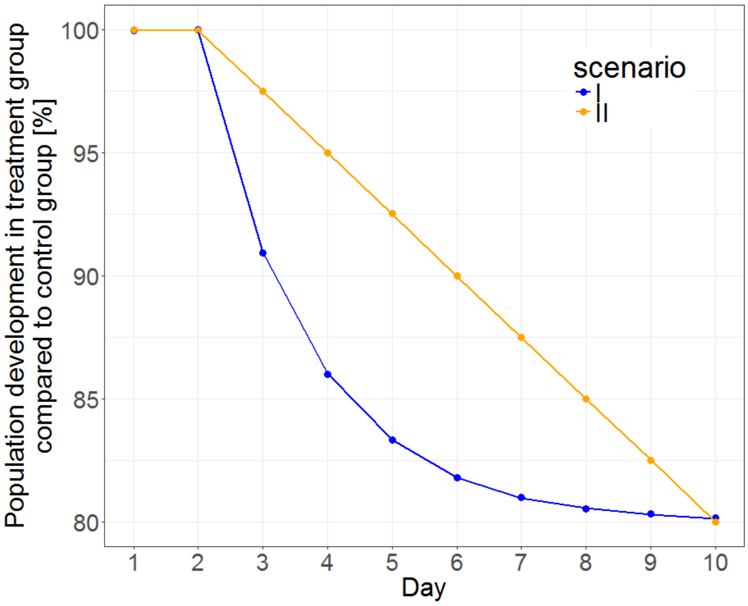


Figure 1: Assumed timing of acute effects after application of PPP

2.4.3.1 Estimation of statistical power through simulation

In order to determine what sample size is needed so that actual treatment effects can be found in the statistical output, generic data were used and PPP effects were simulated according to scenarios I and II.

Information about the presence of radio-tagged individuals of different bird species and one small mammal species were collected in the working routine of five telemetry studies (2012 – 2017). For each individual the presence time during 12 days after tagging was determined. The data were combined based on similarity of species (small or medium size), crop and study season. A specified number of individuals was sampled from the pooled data with replacement. For half of these individuals, scenarios I and II of treatment effect were applied. The resulting datasets were analysed using the Cox proportional-hazard model and model formulas included the factor treatment for all species. For small birds and mammals, region and sex were added as fixed effects; for small birds the species was added to account for differences between insectivorous and omnivorous birds. The simulation was carried out 10 times with 1000 runs each. The fraction of times that the factor treatment was significant per 1000 runs was counted and the mean and standard deviation were calculated. The resulting mean value represents the mean power of the scenario. The significance level was set to p=0.10 in order to allow for a more conservative analysis.

2.4.3.2 Survival analysis

Statistical survival models deal with the analysis of time duration, i.e. survival times, between the entry to the study and a subsequent event, such as death or signal loss (*time*-to-event). The typical observations in survival studies are right-censored, with events (death or signal loss) at times t_1, t_2, etc. Right-censoring occurs because most of the birds are still alive at the end of the study (the censoring time); meaning that the point of death is unknown. Thus, we know that for a censored individual the data point (time of death) is greater than the observation time. The number of checked individual birds at time *i* is n_i, and the number of events at time *i* is d_i.

2.4.3.2.1 Kaplan-Meier survival curve

From the set of observed survival times in our sample of individuals, we can estimate the proportion of the population which would survive a given length of time under the same circumstances. This method is called the Kaplan–Meier estimate of the survivor function. This non-parametric method produces a table and a graph which are referred to as the "life table" and "survival curve" respectively.

The Kaplan–Meier estimate of the survivor function is a step function, in which the estimated survival probabilities are constant between adjacent death times and only decrease at each death (including signal loss). The cumulative proportion of individuals surviving with increasing study day is shown by groups of individuals in the respective group.

To determine the Kaplan–Meier estimate of the survivor function, a series of time intervals was created. Each of these intervals was constructed in a way that one observed death or loss of signal was contained in the interval, and the time of this death/loss was taken to occur at the start of the interval.

The Kaplan-Meier or product estimator arises naturally as a Maximum Likelihood Estimator (MLE) when the survival function is a step function $\mathrm{S(t)} = \prod_{i::\mathrm{t_i}<\mathrm{t}} \alpha_\mathrm{i}$. The likelihood is then

$$\mathrm{L}(\alpha) = \prod_{i}(1 - \alpha_\mathrm{i})^{\mathrm{d_i}} \alpha_\mathrm{i}^{\mathrm{n_i}-\mathrm{d_i}},$$

and the MLE is $\hat{\alpha}_\mathrm{i} = 1 - \frac{\mathrm{d_i}}{\mathrm{n_i}}$. Therefore, the KM survival function is $\mathrm{S(t)} = \prod_{i::\mathrm{t_i}<\mathrm{t}}(1 - \frac{\mathrm{d_i}}{\mathrm{n_i}})$

Where:

α_i is the probability of survival at t_i given a bird is alive before time t_i
d_i is the number of deaths(signal lost) at time t_i
n_i is the number of animals at risk prior to t_i.

An important part of the survival analysis is the creation of plot with the survival curves plot for each group of interest, e.g. treatment with PPP. However, the comparison of the survival curves of the two groups should be based on a formal non-parametric statistical test, the log-rank test, and not upon visual impressions.

However, the log-rank test cannot be used to explore (and adjust for) the effects of several variables, such as species and sex. Adjustment for variables that presumably affect survival may improve the precision of the estimation of the treatment effect. Therefore the determination of the significance was done in the second part of survival analysis by using the Cox model.

2.4.3.2.2 Cox proportional hazard model

The Cox model is a well-known statistical technique, used in medical research for exploring the relationship between the survival of a patient and several explanatory variables (Chow et al., 2008). In our case study, it was used to estimate the treatment effect of the PPP on the survival time of birds after adjustment for other explanatory variables. The model allows the isolation of treatment effects from the effects of other variables, and additionally, allows the estimation of the hazard (or risk) of death for an individual, given their prognostic variables. The result is given as a hazard ratio, which is defined as the proportion of the hazard in the affected group to the hazard in the reference group.

The Cox model regresses the survival times (or more specifically, the so-called hazard function) on the explanatory variables. The hazard function is the probability that an individual will experience an event (for example, death) within a given time interval. A hazard is defined as the rate at which events happen, so that the probability of an event happening in a given time interval. Although the hazard may vary over time, proportional hazard models for survival analysis assume that the hazard in one group is a constant proportion of the hazard in the other group. This proportion is the hazard ratio.

Following the Cox model, the estimated hazard for individual i with covariate vector $\mathrm{x_i}$ has the form

$$\hat{\mathrm{h}}_{\mathrm{i}}(\mathrm{t}) = \hat{\mathrm{h}}_0(\mathrm{t})\exp(\mathrm{x_i}'\hat{\beta}),$$

where $\hat{\beta}$ is found by maximising the partial likelihood, while $\hat{\mathrm{h}}_0$ follows from the Nelson-Aalen estimator, $\hat{\mathrm{h}}_0(\mathrm{t_i}) = \frac{\mathrm{d_i}}{\sum_{\mathrm{j:t_j \geq t_i}} \exp(\mathrm{x_j}'\hat{\beta})}$

with $\mathrm{t_1}, \mathrm{t_2}, \cdots$ the distinct death event times and $\mathrm{d_i}$ the number of deaths at $\mathrm{t_i}$.

Similarly, the survival function is assumed as: $\hat{\mathrm{S}}_{\mathrm{i}}(\mathrm{t}) = \hat{\mathrm{S}}_0(\mathrm{t})^{\exp(\mathrm{x_i}'\hat{\beta})}$

with $\hat{\mathrm{S}}_0(\mathrm{t}) = \exp(-\hat{\Lambda}_0(\mathrm{t}))$ and $\hat{\Lambda}_0(\mathrm{t}) = \sum_{\mathrm{j:t_j \leq t}} \hat{\mathrm{h}}_0\,(\mathrm{t_j})$.

Like this the final model from a Cox regression analysis yields an equation for the hazard as a function of several explanatory variables.

2.4.4 Results

2.4.4.1 Estimation of statistical power

The minimum sample size required to detect a specified effect size with a power of 0.8 is highly dependent on the effect size, on the standard deviation (SD) of the survival times in the control group and, to a lesser extent, on the effect scenario (Table 7).

Table 7: Sample size and statistical power for different species

Species	N individuals (treatment + control)	Mean power scenario I[1]	SD power scenario I[1]	Mean power scenario II[2]	SD power scenario II[2]	Mean SD of survival times in control groups (days)
Medium granivorous bird 'pigeon'	66	0.84	0.04	0.80	0.05	0.37
Small insectivorous/ omnivorous bird 'wagtail/lark'	80	0.86	0.11	0.80	0.12	1.08
Small omnivorous mammal 'mouse'	132	0.84	0.13	0.80	0.11	2.2

For focal species with a high survival rate, i.e. a low rate of signal loss, the minimum sample size is lower as for species with a low survival rate, i.e. a high rate of signal loss, as shown by the comparison of the estimates for the number of individuals necessary and the mean standard deviation of the survival times in the control group for 'pigeon' and 'mouse' (Table 7 and Figure 2)

2.4.4.2 Kaplan-Meier estimate of the survivor function

The visual comparison between the survival curves of treatment and control fields showed that both the estimated survival curves and the confidence intervals overlap only partly between the two groups, especially for the 'mouse' (Figure 2). This suggests a difference in survival, as expected due to the simulated effect by treatment.

Additionally, Figure 2 allows the comparison of the estimated survival curves for small insectivorous/omnivorous birds and a small omnivorous mammal. The bird and mammal species have different survival probabilities throughout the study period, with mammals showing a clearly lower estimated survival (it is noted that methodological improvements in the encounter rate are feasible).

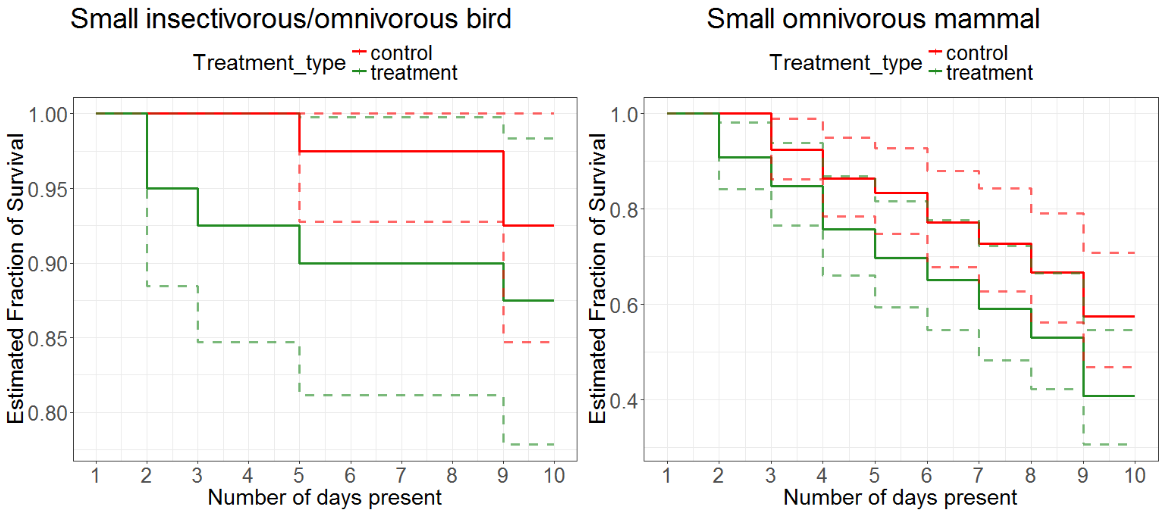


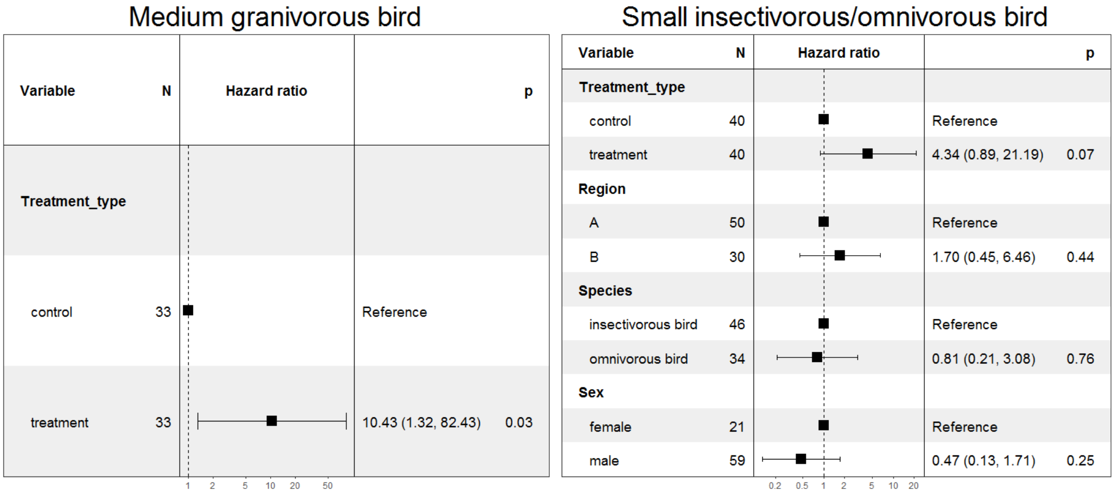
Figure 2: Estimated Kaplan-Meier (K-M) survival curve of scenario I for combined datasets of small insectivorous/omnivorous birds and small omnivorous mammals for different groups (T = treatment, C = control group).

Dashed lines indicate the 95% confidence intervals (CI) for the estimated K-M curves.

2.4.4.3 Cox proportional hazard model

The Cox proportional hazard regression was used to model the bird/mammal survival in relation to the treatment with PPP and further parameters when available (e.g. species, sex and region).

Figure 3 and Figure 4 give the summary of the full model including all predictors for the three focal species groups. A hazard ratio with value of 1 indicates that the hazard is the same for the two groups tested. If the ratio is significantly different from one, there is a significantly different survival (death + signal loss) between the two groups. For example, for small insectivorous/omnivorous birds in Figure 3, the hazard ratio for treatment is 4.34 and significant (p = 0.07), meaning that individuals in the control group have a higher hazard rate than individuals in the treatment group.



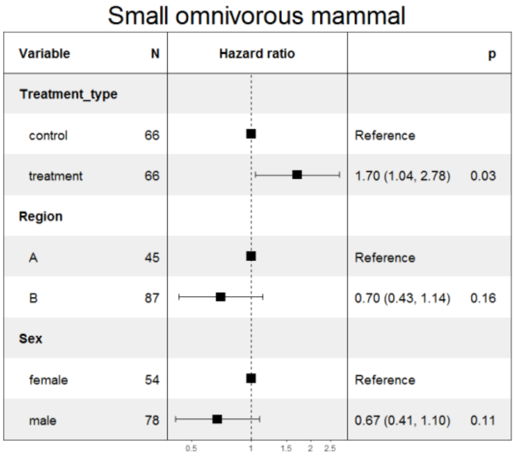
Medium granivorous bird

Variable	N	Hazard ratio		p
Treatment_type				
control	33		Reference	
treatment	33		10.43 (1.32, 82.43)	0.03

Small insectivorous/omnivorous bird

Variable	N	Hazard ratio		p
Treatment_type				
control	40		Reference	
treatment	40		4.34 (0.89, 21.19)	0.07
Region				
A	50		Reference	
B	30		1.70 (0.45, 6.46)	0.44
Species				
insectivorous bird	46		Reference	
omnivorous bird	34		0.81 (0.21, 3.08)	0.76
Sex				
female	21		Reference	
male	59		0.47 (0.13, 1.71)	0.25

Figure 3: Cox proportional-hazard model of scenario I for medium granivorous birds and small insectivorous/omnivorous birds



Variable	N	Hazard ratio	p
Treatment_type			
control	66	Reference	
treatment	66	1.70 (1.04, 2.78)	0.03
Region			
A	45	Reference	
B	87	0.70 (0.43, 1.14)	0.16
Sex			
female	54	Reference	
male	78	0.67 (0.41, 1.10)	0.11

Figure 4: Cox proportional-hazard model of scenario I for small omnivorous mammals

That is, treatment birds have an estimated more than 4 times increased risk of death or signal loss than control birds, after adjustment for the other explanatory variables in the model. This corresponds to the simulated lower survival scenario due to the PPP application. Additionally, significant differences in the hazard rate could be found by sex for small insectivorous/omnivorous birds and by region for omnivorous mammals.

For all birds and mammal groups, an increased hazard was found in treatment compared to control fields in the simulated datasets (Figure 3 and Figure 4). Thus, the statistical analysis was able to detect the post-application effect imposed by the assumed PPP scenario.

2.4.5 Conclusion

By means of the radio-tracking method, the survival of a large number of individuals can be monitored. The advantage of the radio-tracking approach lies in the high probability of finding every fatality, including those animals leaving the close proximity of the treated field and dying elsewhere. Whenever carcasses are found, they are still in adequate conditions for subsequent analyses in the lab. Other individuals will leave the study area due to dispersal events and the radio signal cannot be detected anymore, but this eventuality can be covered with adequate study design and sample size as estimated by power analysis. Therefore, the survival rate measured by means of radio-tracking is a combination of acute mortality, where carcasses can be found, and signal losses, most probably due to dispersal or radio-tag malfunction.

For analysis of the survival data, we propose to use the Kaplan-Meier survival curve and the Cox proportional-hazard model, which provide a helpful estimate of the potential effect of the treatment on survival after adjustment for other explanatory variables. For example, in some cases measured survival time can be more affected by the site or the sex than by the treatment.

The number of individuals necessary to detect a specified effect size with a certain power depends on the effect size, on the standard deviation of the survival times in the control group and, to a lesser extent, on the effect scenario. Based on simulations with generic data, the minimum sample size required to detect a specified effect size can be calculated and used for the planning a field 'acute effects' study. Field studies based on these estimations can be considered statistically robust and reliable enough to find effects in the field.

- The radio-tracking method provides a high probability of finding every fatality
- Statistical methods recommended for analysis of the survival data are the Kaplan-Meier survival curve and the Cox proportional-hazard model
- Based on simulations statistically robust and reliable field 'acute effects' study can be conducted

2.4.6 References

Dittrich R, Hotopp I, Benito Martinez M, Wolf C. 2018. Optimising design and analysis of acute effect field studies. Poster, SETAC Conference (Rom).

EFSA 2009. European Food Safety Authority; Guidance Document on Risk Assessment for Birds & Mammals on request from EFSA. EFSA Journal 2009; 7(12):1438.

Chow SC, Jones, B, Liu J-P, Peace KE. 2008. Sample Size Calculations in Clinical Research, 2nd edition. Boca Raton (FL) USA: Chapman & Hall/CRC Biostatistics Series.

Prosser PJ, Hart ADM, Langton SD, Mckay HV and Cooke AS. 2006. Estimating the rate of poisoning by insecticide-treated seeds in a bird population. Ecotoxicology 15:657-654.

2.5 Nest monitoring – An approach to identify reproductive risks of birds

Author: Anja Ruß, Benedikt Gießing, Ralf Dittrich

2.5.1 Introduction and context

The current EFSA (2009) GD offers the possibility of conducting field effects studies to refine the risk of PPPs in general, but states that "this refinement step is not really practical […] to assess potential effects on reproduction." Quite contrary, nest monitoring studies provide an ideal tool to verify the results of avian reproductive tests according to OECD and EPA standards under real world conditions. With a more detailed guidance, field nest monitoring studies can be a valuable tool in the assessment of the PPP risk to birds, as well as part of a post-registration monitoring. Nest monitoring studies are especially valuable if they are appropriately designed to take into account variation in environmental conditions (weather, food availability, predation etc.). To accomplish this, agreed standards should be defined in the context of the revision of the EFSA (2009) GD 'Risk assessment for Birds and Mammals'. Here, we propose a framework regarding study area selection, proper study conduct, relevant parameters as endpoints and suitable statistical tools.

2.5.2 Identification of issues in the EFSA (2009) GD and improved proposals

Table 8: Nest monitoring: The most important issues identified and revision proposals

Issue	GD says:	Problem definition	Action / GD revision proposal
'Nest monitoring studies' for detecting effects of PPPs in the field	Available methods for detecting effects in the field include: "Monitoring of reproductive performance of birds". And it is emphasised that "Large samples of nests are required to ensure that an adequate number are active at the time of pesticide application" (6.4.3 p. 103).	It is not mentioned in the GD how monitoring of reproductive performance of birds can be conducted	Nest monitoring can be a valuable tool for detecting effects of PPPs in the field if the respective methodological guidance is considered appropriately: • Focal species selection should be conducted according to the instructions given (see also separate chapter for 'Focal species selection') • Study site selection should identify an area with a high abundance of the focal species occurring in the respective crop • Sample size required in order to be able to detect potential effects should be calculated by appropriate statistics (e.g. 'power analysis') prior the study • Timing of application should be scheduled that a relevant number of nests is active. However, spread in timing of nesting attempts offers the possibility to investigate the effect of the PPPs on different nest stages and identify the potentially most vulnerable stage • Nest search should consider the given instruction in order to be effective but as less invasive as possible • Nest fate determination should be based on a pre-defined catalogues of parameters taken at each nest and a standardised criteria list of observations • Endpoints of the study have to consider the relevant aspects of avian reproduction (e.g. nest

Issue	GD says:	Problem definition	Action / GD revision proposal
Nest monitoring studies' for detecting effects of PPPs in the field			abundance, clutch initiation date, no. of breeding attempts, clutch size, fertile eggs/nest, unhatched eggs/nest, chicks fledged/chicks hedged, growth rate) and have to be statistically tested appropriately

2.5.3 Nest monitoring studies in EFSA (2009) GD

In the current EFSA (2009) GD (section 6.4.3.), methods for detecting effects of PPPs in the field mention the possibility to monitor reproductive performance of birds and state the need for a large sample size of nests to ensure that a sufficient number of nests are active during PPP application but give no further guidance on the study conduct.

2.5.4 Points to consider

2.5.4.1 Selection of focal species and study site

In general, selecting the **focal species** is essential in field effects studies for the outcome and the appropriateness of the study. Depending on the crop, focal species are those that breed inside the study field and might be directly exposed to the application of PPP as well as species that breed in adjacent habitats and using the study field to forage during the reproductive phase. Thus, monitoring of the study field and adjacent habitat for nesting attempts is necessary to cover all possibly affected bird species. Further advice on the selection of focal species is given in the respective section 2.1.

As important as the selection of the focal species is the selection of the **study sites**. Ideally, the study sites are chosen by conducting a 'pilot' study in different potential areas during the breeding season prior to the nest monitoring study. Based on the abundance of breeding birds recorded in the different inspected areas during this pilot study and the expectation that the pattern of the abundance will be similar during the next season the most suitable study sites can be selected. During the 'pilot' study, the occurrence and abundance of the focal species is assessed at least twice per potential study area to get a reasonable estimate of the presence of the species (or these species) in general and their occurrence in the crop(s) planned to be investigated for the nest monitoring study. Thus, potential study areas are determined by distribution and abundance of the focal species and the actual study fields are selected shortly before the start of the study depending on the presence of adult individuals in the crop.

2.5.4.2 Sample size

Ideally, the required sample size is determined by a power analysis based on a 'pilot' study (see section 2.6.8.1). In the case that a previous study is not available, a minimum of 5 study fields/orchards per treatment should be investigated in order to reach a sufficient sample size of nests. Especially in open cup nests the predation rate is comparably high and approximately twice as many nests have to be included in the study to account for losses due to predation. Accordingly, study sites need to be large enough to accommodate a sufficient number of active nests of the focal species.

2.5.4.3 Timing of application

If possible, the timing of PPP application should be scheduled that a relevant number of nests is active which might vary for different species. As already mentioned in the EFSA (2009) GD, "only a proportion of birds will be exposed and furthermore, for those which are exposed, the peak exposure may not occur during the most sensitive reproductive phase". However, this is not a shortcoming of nest monitoring studies but an advantage as the spread in timing of nesting attempts over the entire breeding season offers the possibility to investigate the effect of the PPP application on different nest stages and identify the potentially most vulnerable stage (see section2.5.4.7).

Depending on the time of application in the breeding season, nests which are completed prior to the PPP application may serve as additional control nests. However, the success rate of nests is closely correlated to the time of breeding with early nests having a higher probability to be successful than later nests (e.g. Verhulst & Nilsson 2008). This has to be considered when early (control) nests are compared to late (exposed) nests (see also example below, section 2.5.4.4).

2.5.4.4 Nest searches

Systematic **nest searches** are conducted repeatedly by experienced/trained staff every 7-10 days to guarantee the detection of breeding attempts of species with short breeding duration. The time of the first nest search needs to be appropriately chosen to detect also early breeding attempts. It should be aimed to find nests at an early stage of activity in order to determine its fate appropriately. In orchards and woodland areas close to the study fields, natural cavities have to be checked for cavity-nesting species using an endoscope camera. The actual number of nest searches has to be adapted to the study design. If nests in a specified interval should be analysed also one nest search can be sufficient.

In addition to systematic nest searches, "watching back" is a suitable method to find active nests. By following the movements and behaviour of birds seen it is possible to identify key locations of activity which can lead to find the nest (Ferguson-Lees et al., 2011). Furthermore, radio-tracking of adults also aids in finding nests, while ground nests in uniform open habitat can be found by dragging a long (approximately 40m) rope over the vegetation and flushing the incubating bird. In open habitat, nests can be marked with small poles a few metres away from the nest to facilitate relocation. For a more detailed description of nest searching methods see Ferguson-Lees et al. (2011).

In case cavity-nesting species (e.g. blue tit (*Cyanistes caeruleus*), great tit (*Parus major*), pied flycatcher (*Ficedula hypoleuca*)) were identified as focal species, **nest boxes** can be installed in the study orchard or at the field margins. It is important to use nest boxes that can be opened easily without chasing away a potentially incubating bird on the nest when the nest box is checked. Mounting poles to support the nest boxes proved to be useful and allow an equal density of nest boxes/ha in different study sites. Nest boxes should be established at least one year prior to the study in order to accustom the birds to the nest boxes. The diameter of the nest box' entrance hole determines the suitability of the nest box for different species. Especially in orchards, nest predation by mammals (e.g. dormouse) can be severe. To hinder predator access to the nest box, slippery tape can be applied around the entrance hole as well as plastic covers at the side of the nest boxes

2.5.4.5 Nest fate

Each nest deemed to be active (i.e. prepared for egg laying, with eggs or chicks), the locations (GPS coordinates) and the current stage of the breeding attempt following Sutherland et al. (2004) is recorded. The incubation stage of bird eggs is estimated by candling (Lokemeon and Koford, 1996). The **fate of active nests** is monitored until the juveniles fledge or the nest becomes inactive for other reasons (e.g. chicks predated or dead). Active nests are checked on regular intervals with timing adapted to the status of the nest to keep the disturbance as low as possible. Nests with eggs are checked at longer intervals (3-5 days) to minimize disturbance in this early stage when nest abandonment is more likely than during later stages. Nests with chicks can be checked every 2-3 days and nests with older chicks close to fledging every 1-2 days in order to verify that the chicks left the nest successfully. Flushing birds which sit on the nests or disturbing adults during feeding young must be avoided and the nest content is then not checked. Additionally, nests checks one day (approx. 24 hours) before and one day (approx. 24 hours) after application of PPP increases the certainty in nest fates regarding potential effects. In case nests were empty earlier than expected, all signs (e.g. egg shells, nestling body parts, tracks left by predators) are recorded to determine the probable cause of failure (Table 9). A nest is categorized 'successful' if at least one nestling fledged irrespective of the fate of the other eggs/nestlings.

Remaining eggs which do not show any signs of development according to candling (Lokemeon and Koford, 1996) after all other eggs have hatched are categorized as infertile eggs. In case an egg disappears between nest controls without any signs of predation, it is probably ejected by the adults due to unfavourable weather conditions and sanitary behaviour (Shitikov et al. 2018) and is also counted as infertile egg.

Table 9: Proposed criteria for the standardisation of nest fates

Nest status	Other evidence	RESULT
Nest with eggs found to be empty before the hatching date	no further evidence needed	Predated
Nest with eggs found to be empty after the assumed hatching date (in between of two controls)	Rests of eggs, removed vegetation or nest material, blood...	Predated
Nest with eggs found to be empty after the assumed hatching date (in between two controls)	No hints	Empty
Nest with chicks was found to be empty before the earliest leaving age	Rests of chicks, removed vegetation or nest material, blood...	Predated
Nest with chicks was found to be empty before the earliest leaving age	No hints	Empty
Nest with chicks was active on the last visit before the earliest leaving age and empty on the first visit after the earliest leaving age	Evidence of fledging (e.g. fledglings seen/heard, droppings in and at the edge of the nest-cup, adults uttering alarm calls)	Successful
Nest with chicks was active on the last visit before the earliest leaving age and empty on the first visit after the earliest leaving age	No evidence if chicks are fledged or predated	Probably Successful
Nest with chicks was active on the last visit before the earliest leaving age and empty on the first visit after the earliest leaving age	Rests of chicks, removed vegetation or nest material, blood...	Predated
Nest was found to be active after the earliest leaving age and in next check it is empty	Evidence of fledging (e.g. fledglings seen/heard, droppings in and at the edge of the nest-cup, adults uttering alarm calls)	Successful
Nest was found to be active after the earliest leaving age and in next check it is empty	No evidence or equivocal	Likely Successful
Eggs left in the nest after incubation ceased, cold and/or wet	Two or more visits with same status	Abandoned
Empty nest with no content at any check, wet or with leaves inside	Two or more visits with same status	Inactive
Dead Chicks in the nest	No traces of predation	Dead
Destroyed nest		Predated
Nest with previous breeding activity not found again	Evidence of failure (checked by same person who found it)	Predated
Nest with previous breeding activity but no information on contents, found empty or not found	e.g. female on the nest + no info on contents + found empty in later control	Unknown

2.5.4.6 Endpoint 'nest survival'

The proportion of the number of successful nests divided by the total number of monitored nests is known as the **apparent nest survival**. It gives general information about a possible treatment effect but does not consider further parameters which affect the survival of a nest such as time of breeding or weather conditions and has, thus, only limited explanatory power.

Mayfield (1975) recognized that the appropriate sampling unit was not the nest as used in the proportional apparent nest survival, but the number of days the nest was active and therefore exposed to the hazards of predation, parasitism, bad weather conditions or other negative factors. The **daily survival rate (DSR)** is different for nests which failed early, e.g. just after egg laying, compared to nests failed in a later nesting stage. Therefore it is important to take into account how many days a nest was active after being found. By estimating the sum of exposure days of all nests and total number of failed nests Mayfield equation for daily survival rate can be applied:

$$DSR = \frac{1-(total\ number\ of\ failed\ nest)}{(total\ number\ of\ exposure\ days)}$$

The DSR estimates the rate at which a nest will survive from one day to the next day. To calculate the survival probability of nests (NSP) which estimates the probability that the nest will survive the entire nesting period (laying, incubation, nestlings) the daily survival rate is raised to the power of the number of days of the entire nesting period:

$$NSP = DSR^{days\ of\ nesting}$$

The Mayfield estimator assumes that the hazard rate is constant throughout the nest period. Like this is it still limited to test for an application effect. Therefore the DSR is modelled as the response variable as a function of temporal variation in nest survival and covariates representative of individual nests which allows to incorporated greater detail. This represents a substantial improvement over traditional estimation methods (Dinsmore et al. 2002). The resulting generalized linear model is described by Shaffer and Burger (2004) as the **logistic-exposure model**.

A case study shall illustrate how powerful Shaffer’s logistic-exposure model is. Breeding parameters were monitored for blue and great tits in 10 conventional (applied 1 or 2 times with a PPP) and in three organic pome fruit orchards in 2013 and 2014 in the UK. Birds used natural cavities and nest boxes in the study orchards for breeding. The fate of 156 and 309 active nests of great and blue tit, respectively, was monitored at fixed intervals. A logistic exposure model was fitted to the nest survival of each species, taking into account other potentially relevant factors, e.g. time and year. The logistic exposure model of nest survival revealed for both species a significant effect of the breeding date: survival probability decreased as breeding season progressed, i.e. later initiated nest had a lower success rate (Figure 5).

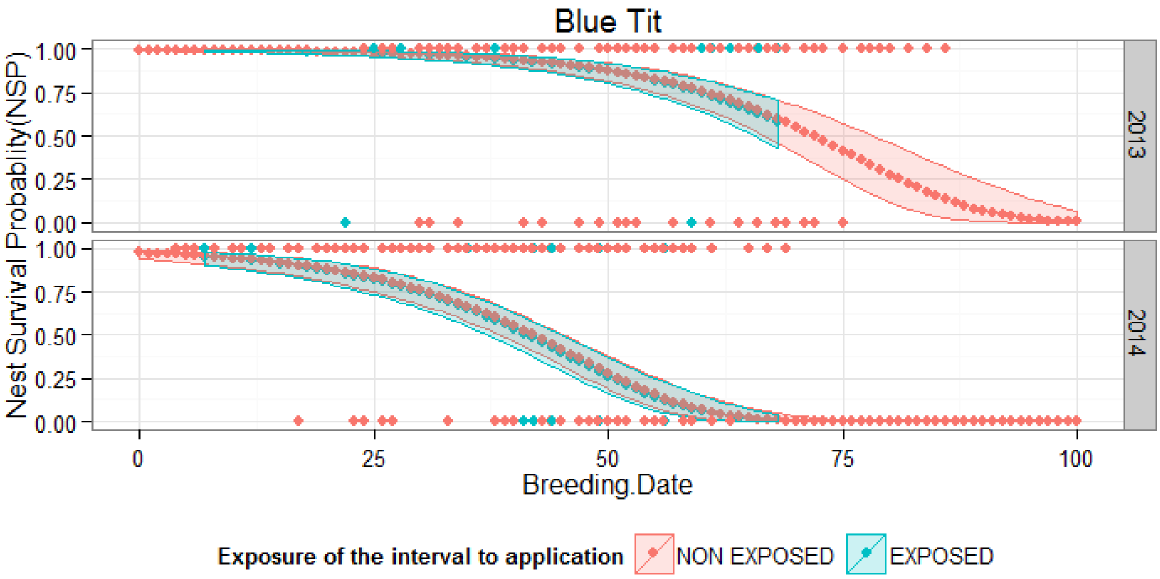


Figure 5: Prediction plots of nest survival probability (NSP) for blue tit, years 2013 and 2014.
Each dot represents a single interval of monitored nest. Dots above the graph (at NSP = 1) indicate successful intervals, while failed intervals are shown below (NSP = 0). Bands are confidence intervals for the estimated NSP drawn from the model.

The observed decline in survival was most probable linked to a decline in food availability. Furthermore, for both species a significant difference was found between years, with nests from 2014 having a lower survival probability than those from 2013, most likely due to less favourable weather conditions in 2014. As shown by the overlap of the estimated confidence intervals for exposed and unexposed nests, the survival probability was not significantly affected by the PPP application (Figure 5). Despite the natural variability, the main pattern underlying the nest survival could be identified and an effect by the PPP was ruled out.

2.5.4.7 Endpoints according to 'avian reproductive tests'

While the fate of the nest can be seen as a summary of the entire reproductive cycle, other endpoints are more linked to one of the five different phases of the reproductive cycle as proposed by Bennett et al. (2005). The current EFSA (2009) GD already recognised the importance to differentiate between these phases: "The screening and Tier 1 assessments do not distinguish between different phases of reproduction. In reality, different phases of reproduction may differ both in their exposure and their toxicological sensitivity to the pesticide. […] These factors may be addressed by phase-specific risk assessment. To gain the full benefits of this approach requires detailed data that may not be available in some cases (e.g. time of application of the pesticide, time of breeding phases for focal species etc.). However, the phase specific approach may be an effective approach if these data are available." Table 10 provides an overview of the reproductive phases and the corresponding endpoints which can be used to indicate phase specific effects of the PPP application in nest monitoring studies.

Table 10: Phase specific effects of PPP application and proposed endpoints tier1-testing of avian reproduction

(OECD Guidline for testing chemicals 206; 4 Apr. 1984; EPA Ecological effects est guidelines: OPPT850.2300; Avian reproduction test; APA 712-C-96-141; Apr. 1996) and in nest monitoring studies. Table modified from Bennet et al. 2005 and Bouvier at al. 2005

Breeding phase	Phase-specific effect of concern	Endpoints in tier1 testing avian reproduction	Endpoints in nest monitoring study
Pair formation/breeding site selection	Adult behavioural effects leading to territory abandonment or delayed breeding	Mortality (and behaviour) of adults, Adult body weight prior to breeding	Adult body weight prior to egg laying (probably obtained from accompanying trapping study); number of pairs/ha; second brood initiation (no. of pairs initiating a second brood after a first brood with at least one fledged young)
Copulation and egg laying (5 days pre-laying through end of laying)	Adult behavioural effects leading to reduced clutch size or abandonment of nesting attempt	Number of eggs laid per hen	Clutch size per nest; date of clutch initiation (date at which first egg was laid)
	Reduced egg shell thickness	Egg shell thickness per pen each 14 days, Number cracked eggs at Day 0 of incubation	Eggshell thickness (not yet routinely measured in nest monitoring studies but can be implemented)
	Reduced fertility	Viability	Proportion of fertile eggs/nest
Incubation and hatching	Adult behavioural effects leading to abandonment of nesting attempt		Proportion of abandoned nests
	Embryotoxicity leading to reduced hatchability	Hatchability, number of embryos that mature , embryos that pip shell, embryos that liberate themselves	Proportion of unhatched eggs/nest after incubation
Juvenile growth and survival until fledging	Adult behavioural effects leading to brood abandonment or abnormal parental care	Change in adult body weight each 14 days	Growth rate of nestlings, nestling weight at age 8 days, fledging rate (rate of chicks fledged relative to the number of chicks hatched)

Breeding phase	Phase-specific effect of concern	Endpoints in tier1 testing avian reproduction	Endpoints in nest monitoring study
	Reduced juvenile survival from direct exposure	Not applicable since juveniles are not exposed in avian reproduction test	Growth rate of nestlings, nestling weight at age 8 days, fledging rate (rate of chicks fledged relative to the number of chicks hatched)
	Reduced juvenile survival and growth from *in ovo* exposure	Percentage of 14d-survivors, body weight of 14d-survivors, cumulative mortality until 5 days of age and 5-14 days	Growth rate of nestlings, nestling weight at age 8 days, fledging rate (rate of chicks fledged relative to the number of chicks hatched)
Post-fledging survival	Reduced fledging survival from direct exposure	Not applicable	Nestlings growth rate as an indication of post-fledging survival probability
	Reduced juvenile survival and growth from *in ovo* exposure	body weight of 14d-survivors	Nestlings growth rate as an indication of post-fledging survival probability

In the nest monitoring studies, endpoints should be included which are also considered in the laboratory avian reproduction tests according to OECD or EPA: number of laid eggs (clutch size), infertile eggs and hatched eggs; embryo development, weight of chicks at the age of 8 or 14 days depending on the species, and survival of chicks. Unlike the tier1 tests, nest monitoring studies additionally cover the period of parental care of the chicks after hatching, a phase of great importance for altricial species like most European farmland birds. Furthermore, nest monitoring studies also cover the direct exposure of chicks which might be more vulnerable to negative effects of the PPP.

In contrast to lab studies whose high degree of environmental standardisation can lead to spurious findings and little external validity (Richter, 2009), nest monitoring studies can provide, with the help of appropriate statistics, the assessment of the impacts of relevant environmental factors (including PPPs) on reproduction, in spite of the high natural variability.

2.5.5 Conclusion

Nest monitoring is a commonly used tool in population ecology research; therefore it offers a wide range of well-established methodologies and statistical approaches approved by the scientific community. Additionally, sufficient sample sizes are possible to obtain. This allows proper statistical testing in order to account for the natural variability. Furthermore the farmland bird species, actually relevant for a certain crop and GAP, differ in their biology from the standard test species like the bobwhite quail or mallard. Following a properly defined study protocol, the reproductive performance following a PPP application can be shown for both, endpoints used in avian reproduction tests and for further key parameters like nest survival and number of fledglings per pair. By means of advanced statistics, the potential effect of the PPP and other relevant environmental parameters can be quantified. These results and the comparison with published data prove their potential for extrapolation to the general situation in the studied crop in a certain zone. With a detailed guidance, field nest monitoring studies can be a valuable tool in risk assessment of the PPP risk to birds, and therefore they should be part of the revision of the EFSA (2009) GD 'Risk assessment for Birds and Mammals'.

2.5.6 References

Bennet RS, Dewhurst IC, Fairbrother A, Hart ADM, Hooper MJ, Leopold A, Mineau P, Mortensen SR, Shore RF, Springer TA. 2005. A new interpretation of avian and mammalian reproduction toxicity test data in ecological risk assessment. Ecotoxicology 14:801-815.

Bouvier J-C, Toubon J-F, Boivin T, Sauphanor B. 2005. Effects of apple orchard management strategies on the great tit (*Parus major*) in southeastern France. Environmental Toxicology and Chemistry 24:2846-2852.

Dinsmore SJ, White GC, Knopf FL. 2002. Advanced Techniques for Modeling Avian Nest Survival. Ecology 83:3476-88.

EFSA 2009. European Food Safety Authority; Guidance Document on Risk Assessment for Birds & Mammals on request from EFSA. EFSA Journal 2009; 7(12):1438.

Ferguson-Lees J, Castell R, Dave L. 2011. A Field Guide to Monitoring Nests. Norfolk, UK: British Trust for Ornithology. 272 p.

Lokemoen JT, Koford RR. 1996. Using candlers to determine the incubation stage of passerine eggs. Journal of Field Ornithology 67:660-668.

Mayfield HF. 1975. Suggestions for calculating nest success. Wilson Bulletin 87:456–466.

Richter S, Helene G, Joseph P, Würbel H. 2009. Environmental standardization: cure or cause of poor reproducibility?. Nature Methods 6 (4):257-261.

Shaffer TL, Burger AE. 2004. A Unified Approach to Analyzing Nest Success. The Auk 121:526–540.

Shitikov D, Samsonov S, Marakova T. 2018. Cold weather events provoke egg rejection behavior in open-nesting passerines. Ibis. https://doi.org/10.1111/ibi.12695.

Sutherland JS, Newton I, Green RE. 2004. Bird Ecology and Conservation. Oxford, UK: Oxford University Press.

Verhulst, S, Nilsson J-Å. 2008. The timing of birds' breeding season: a review of experiments that manipulated timing of breeding. Philosophical Transactions of the Royal Society 363:399 410.

2.6 Long-term field effect studies – Determination of risks on the population level

Author: Olaf Fülling, Ralf Dittrich and Ines Hotopp

2.6.1 Introduction

The higher tier risk assessment for Plant Protection Products (PPP) is based on the calculation of a Toxicity Exposure rate (TER). The TER is a model that incorporates different factors, for example the food intake rate, the residue unit dose or the proportion of treated food in the animal's diet. These values for the model can be assumed, adopted from the results of laboratory studies or be refined values from field studies. The advantage of this process is a well-defined trigger value and specified safety margins for acute and long-term risks. Still, the TER is a model that needs assumptions on e.g. animal foraging behaviour and space use, contamination of mobile and immobile food items and the interaction of these factors exactly as described in the TER-formula.

Field studies, so called field effect(s) studies, on the other hand are designed to observe possible acute as well as chronic (adverse) effects of a PPP under realistic conditions. Thus, such studies do not need any model assumptions on animal behaviour, food intake rates or residues. The realistic conditions, however, should be representative for the use of the test item. To be more conservative, the field study can be designed as a worst but still realistic case.

2.6.2 Identification of issues in the EFSA (2009) GD and improved proposals

Table 11: Long-term studies: The most important issues identified and revision proposals

Issue	GD says:	Problem definition	Action / GD revision proposal
Number of study sites	"The 'extensive' approach uses simple techniques such as carcass searching and census methods but employs a large number of sites to cover a broad spectrum of use conditions." (6.4.2, p. 102)	Simple methods like carcass searching have their well-known short comes. More up-to-date methods are available to achieve the goal.	We suggest an updated definition of the extensive approach using radio tracking for acute risks and capture-recapture designs for chronic risks.
		The 'large number' of study sites is not specified.	A statistical power analysis can calculated the number of individuals to be tracked (acute) or the number of sites to be trapped on (chronic).
	"The 'intensive' approach on the other hand involves more detailed investigations but on a smaller number of sites, or on one site only."	Conducting detailed investigations on a large number of sites (calculated by a power analysis) might not be feasible. A lower number of sites might	Data on e.g. population size, body weight, reproduction or proportion of juveniles should be investigated by comparing treated

Issue	GD says:	Problem definition	Action / GD revision proposal
Number of study sites	(6.4.2, p. 102)	be selected. Only one field is never sufficient.	and reference sites. When, for practical reason, the number of sites has to be below the results of a power analysis, MDDs can be used to evaluate the weight of evidence of the particular study.
	"The recommen-dations of the 1988 workshop tended to favour the intensive approach. However, this should be reconsidered in the light of developments since that time. […]It is concluded that an 'extensive approach' with suitable methods and an appropriate number of sites is preferable to field studies with fewer sites." (6.4.2, p. 102)	The current GD prefers on of the two approaches over the other one. The intensity of methods used in 'extensive' approaches and the number of sites used for 'intensive' approaches has increased and better statistical method gained wider use.	Both, the intensive and the extensive approach should be considered valuable. The definition of extensive and intensive needs to be updated and quality measurements like power analysis or MDDs should be mandatory. However, even studies with a power below 80% or MDDs above 50% add information to the regulatory process.
Statistical power	"To enable the design of a study of appropriate power, it is desirable to know in advance the levels of effects that are considered acceptable, as well as the degree of certainty that is required to prevent the acceptable limit being exceeded. Since such questions address risk management, it is desirable to discuss them in advance with the relevant authorities." (6.4.1. p. 102)	As field studies have the highest level of realism in the process of PPP risk assessment, they should not simply rejected just because a power analysis > 80% cannot be shown.	A power analysis is the appropriate method to calculate the number of study fields resp. individuals needed. However, depending on the archived safety margin (TER>1) of the lower tier results, a power below 80% can be sufficient. In this case discussion with the authorities prior to a field study, as by the current GD suggested, is strongly recommended.
			We modified the

Statistical power

Issue	GD says:	Problem definition	Action / GD revision proposal
			concept of MDDs for field studies on mammals. As there are parallels to aquatic mesocosm studies, intensive studies on grassland as a surrogate crop can be evaluated by MDDs.
Worst case scenario/surrogate crop	"Furthermore, this cannot be addressed by selecting 'worst-case' sites, as it is not possible to know in advance which sites will have high residues or which species will be most sensitive, nor is it possible to ensure that individuals of sensitive species with high PT will be present." (6.4.2, p. 102) "Pen tests […] Such tests are only rarely conducted with mammals and birds, and there is no currently-recognised standard method. (6.4.5, p. 104)	The current GD is rather sceptical concerning 'worst case' scenarios.	We suggest field studies on common voles (and to a certain extend on rabbits) on surrogate crop grassland as an alternative to pen studies covering a worst case scenario. Common voles use grassland (meadows, fallow land) as their preferred habitat. One (sub) population is usually restricted to a single grassfield and individual home ranges are small. When grassland is used as a surrogate crop (e.g. for fungicide spray applications) the study design will allow sufficient replicates (see MDD calculations), absolutely untreated controls, maximized exposure of the preferred food source (the grass), high abundance of fully exposed individuals and high recapture rates providing histories of individual

Issue	GD says:	Problem definition	Action / GD revision proposal
Worst case scenario/surrogate crop			measurements. In contrast to a pen study grassland studies can be conducted under different climatic conditions, with a number of replicates. Specific capture-mark-recapture statistics allow calculating in-situ recruitment, immigration, emigration and mortality. The design can even provide a risk envelope for other herbivorous small mammals.

2.6.3 Preparations

To be the most realistic and representative approach, a field effects study needs to be well planned and prepared. Depending on the target crop and the distribution of the focal species, the study can be conducted in one geographical region or in regions that differ in climate and/or landscape structure. Within each region a sufficient number of study sites needs to be selected. The current birds and mammals guidance (EFSA 2009) distinguishes between an intensive and an extensive approach which differ in the number of study sites and the effort taken at each site. Here we followed this discrimination but with respect to the development of technical and statistical methods we increased the effort and methodological requirements for both the intensive as well as the extensive approach. After determining the sufficient number of study sites real fields have to be selected. This process starts with GIS investigation of landscape pattern and average field sizes. Once a geographical region or area is selected, fields have to be investigated on the ground to ensure the presence of the focal species.

2.6.4 Number of study fields

In contrast to a laboratory setup the test sites in a field study will never be exactly the same. There is variation in e.g. shape, soil or adjacent habitats that can only be standardized to certain extend by the site selection. To some extent agricultural practice can be controlled during the experiment but not, for example, the weather conditions following an application. Due to this variation of natural conditions, it is a necessity to run a field effect study on a sufficient number of study fields. What a sufficient number of fields means depends on the selected study design, i.e. whether an intensive or an extensive approach has been chosen. The sufficient number of study fields can either be calculated *a priori* by a power analysis (section 2.6.8.1) or *a posteriori* evaluated by MDDs (section 2.6.8.2).

2.6.5 Study design

The study design, i.e. whether to prefer an intensive or extensive approach, depends on the data needed, respectively the endpoints to be considered, but also on the studied species and crop of concern. Furthermore, practical issues like the permission to apply the test item can be relevant. An extensive approach in the study design should be considered if the focal species has established populations in the crop of concern and the PPP is regularly used by the farmers. In case of species with high mobility, like wood mice or hares, the same individual is less likely to be encountered frequently. Thus, an extensive study design can benefit from a high number of study sites (for statistical power) and can deal with a lower number of individual encounter histories. When expanding a field study to a larger area (landscape level), the PPP test item for an effect study should be frequently used by farmers to provide a sufficient number of treated fields within the study area.

In contrast, the intensive approach should be preferred when the focal species does not have established populations in the crop of concern, e.g. voles in vegetables, or the test item is not regularly used. In this case the study needs to be conducted in a surrogate crop, typically meadows, so a high exposure of an established population is guaranteed. Hence, a lower number of sites is sufficient to detect effects, especially when the spatial activity of the studied species is restricted to the study field. Low mobility of studied species increases the probability of frequently encountering the same individual for example in capture-mark-recapture studies on common voles. Thus, the intensive approach suits as well rare crops or when testing a new product which needs exceptional permits to be applied on farmland. The requirements and options for a statistical analysis are provided in sections 2.6.6 and 2.6.7.

2.6.6 Extensive landscape approach

EFSA (2009) GD: The 'extensive' approach uses simple techniques such as carcass searching and census methods but employs a large number of sites to cover a broad spectrum of use conditions. In contrast to that definition given by Somerville & Walker (1990) and employed by the EFSA (2009) GD the proposed extensive, landscape approach uses a large number of sites determined by a statistical power analysis. The same study methods are used as in the intensive approach but the number of trapping sessions is lower and the time interval between sessions is longer. Typically the focus of the study design is on: population effects following the application period and second: the population development at the end of the reproductive phase (note, this can be adjusted according to the expected toxicological effects). The natural variability in the parameter estimates is unproblematic in the analysis due to the high number of study fields.

The study fields should be distributed in different regions (2-3) to facilitate the interpretation and extrapolation of the study outcome. The final distance between the regions is less important but there should be differences in the climatic conditions as this influences the food availability and the reproduction success. These factors are relevant for the study outcome, independent from the toxicological effects by the PPP application. This allows to generalise the study results. Furthermore, it is easier to find suitable study fields.

Table 12: Summarising the key parameters of field effect study on small mammals following the extensive, landscape approach

Method	Capture-Mark-Recapture with individual markings
Main endpoints	Population size and dynamics (e.g. as MNA) individual body weight dynamics reproductive performance (e.g. as juvenile/adult ratio)
Number of study fields	36 (18 control + 18 treated)
Number of regions	3
Number of traps per field	60 to 100 multi-capture live traps
Observation period	1 trapping sessions before the (1st) application, followed by 1 trapping sessions in in temporal proximity to the last application and two further trapping sessions until the end of the species' breeding season
Trapping sessions	3 trapping nights per trapping session
Interval between trapping sessions	Regular intervals, e.g. 90 days (depending on the species)
Statistical analysis	Mixed models (with Poisson distribution for population size and Normal distribution for body weight); confounding factors included

2.6.7 Intensive approach

Our proposal for a modern extensive, landscape approach (section 2.6.6) is suitable for focal species which have established populations in a crop which is regularly treated with the PPP of concern. However, some species, like voles and rabbits, are less mobile but can reach high local population densities. For such species an intensive study approach might be a suitable alternative, especially when the study is conducted in a surrogate crop in order to maximise the exposure. In this case the study system is comparable with mesocosm or bee tunnels which use a comparable number of study sites/units. According to the current EFSA (2009) GD an intensive approach to a field effect study 'involves more detailed investigations but on a smaller number of sites, or on one site only.' Both mammalian species mentioned above, the common vole and the European rabbit, can occur on arable land but they prefer permanent grassland (meadows, pastures or fallow land) for foraging. Voles can live entirely on grassland while rabbits prefer to have adjacent shrubs, hedges or other vegetation to cover their warren entrances. Besides this difference, both species can be studied on grassland as a surrogate crop. A surrogate crop offers some advantages: (I) higher densities of voles and rabbits compared to arable (especially tilled farmland); (II) high exposure when PPP is applied by a boom sprayer directly onto the grass layer and (III) 'clean' control fields without any agro-chemical product applied. Especially as for many high-valued crops it can be impossible to find suitable control crops due to long term contracts of the farmer and structural difference in organic farming practise. Thus, field studies in grassland allow a worst-case scenario in terms of high population densities, highest exposure to the test item without interception, with little or none alternative food sources and 'clean', completely untreated control fields. In the following we will use the example of field effect studies on common voles in grassland to describe what we consider a state of the art field effect study following an up-to-date intensive approach.

For an intensive approach we conduct field studies on common voles using capture-mark-release-recapture (CMR) techniques to monitor population density and age structure but also individual condition and body weights. EFSA (2009) GD explicitly recommends CMR to identify possible adverse effects of a PPP on small mammal reproduction and CMR is also a preferred tool of small mammal ecology (e.g. Fuelling and Halle 2001; Flowerdew et al. 2004; Briner et al. 2007; Bonnet et al. 2013; Hein & Jacob 2018). Hein and Jacob (2018) employed CMR to assess the effect of a rodenticide on common vole populations. While they used four treated and four control plots (fields) and observed the vole populations for two subsequent years, we employ six or seven treated fields and an equal number of control fields for testing PPP effects on common voles in one growing period. The higher number of study fields allows to gain more statistical power for observations within one growing season. This meets the requirements for testing PPPs as each PPP is used only during a relatively short period of the crop development. The number of 12 (6 treated + 6 control) or 14 (7+7) fields is based on experience and verified by the calculation of Minimal Detectable Differences (MDDs) for field effect studies on common voles in grassland. The MDD for statistical comparisons of the population size (as Minimum Number Alive, MNA) was calculated *a posteriori* for four field studies with a total of 31 trapping sessions. On average the applied statistical analysis was able to detect an 18.0% MDD between treatment and control populations. In other words when analysed by a Generalised Linear Mixed Model (GLMM) differences between treatment and control in population size of more than 18 % were considered significant. In the section (2.6.8.2) on Minimal Detectable Differences we will provide evidence that 18% MDD is a considerably good result. When grassland as a surrogate crop is selected the study fields should be selected to be as similar as possible in terms of surrounding habitats and any agricultural activities (mulching or mowing) have to be conducted on all fields in the same manner. It is not necessary to do placebo spray applications with water on the control fields as such activities have no disturbing effects on vole behaviour (Jacob & Hempel 2003). Besides all care taken, there will be differences between study fields. This can be covered by appropriate statistics. The statistical analysis to identify possible adverse effects of a PPP and to address natural differences between study fields will be described in section 2.6.8 . Another issue we consider important for our field effect studies on voles is the study duration and the period between trapping sessions. When tier1 data suggests any adverse effects on the reproductive fitness, the study period should cover in minimum one reproductive season (e.g. from late spring/early summer to the end of the breeding season in late autumn). There should be at least two trapping sessions before the (first) test item application to settle the trapping results and to allow for a before-after application comparison within the same study field. For the statistical analysis of the trapping data it is desirable to sample in a regular pattern with short trapping sessions and biologically meaningful intervals. This will result in a so-called robust design (Pollock 1982) with a short trapping session, usually three to five nights (8 to 12 hours each) and longer intervals between each trapping session. For a study on common voles (and some other small mammal species) this should be a period of three weeks or 21 days because this time span fits the reproductive biology of common voles. Pregnancy in this species lasts usually 19 to 21 days (Dieterlen 2005, Frank 1956a, Reichenstein 1964). According to Reichenstein (1964) it takes another 21 days on average for the new-born pubs to reach a body weight of 10 g. This is the weight when the juveniles leave their burrow and can be trapped and individually marked outside. Our data indicate that trapping sessions with a regular interval in between give much better results (resulting in lower MDDs) than irregular trapping sessions.

Table 13: Summarising the key parameters of field effect study on small mammals following the intensive approach

Method	Capture-Mark-Recapture with individual markings
Main endpoints	Population size and dynamics (e.g. as MNA) individual body weight dynamics reproductive performance (e.g. as juvenile/adult ratio) individual survival
Number of study fields	12 to 14 (6-7 control + 6-7 treated)
Number of traps per field	60 to 100 multi-capture live traps
Observation period	Minimum 2 trapping sessions before the (1st) application until the end of the species' breeding season (approx. end of October for many small mammals in the Central Zone)
Trapping sessions	3 to 5 trapping nights (8 to 12 h each) per trapping session
Interval between trapping sessions	Regular intervals every 20 to 30 days (depending on the species)
Statistical analysis	Mixed models (with Poisson distribution for population size and Normal distribution for body weight); confounding factors included

2.6.8 Statistical analysis

In ecological field studies, increasingly complex data sets are obtained whose analysis requires sophisticated statistical approaches. One major challenge is the lack of statistical independence in the replicates of field studies (Hurlbert 1984). In the case of birds and mammals field effect studies this pseudoreplication arises from e.g. repeated trapping sessions per study field. These study designs would lead to biased parameter estimates and increased type I errors in regression models if not handled appropriately. However, this kind of pseudoreplication can be dealt with by applying mixed-effects models (Pinhero 2000). Further details are given in the section 2.7 'Validation of the study design and statistical analysis of field study data' in the sub-section 2.7.3 'Statistical methods for the analysis of field study data'.

2.6.8.1 Power analysis with simulated fields effects study datasets

An overview to the topic of power analysis is given in the section 2.7 'Validation of the study design and statistical analysis of field study data' in the sub-section 2.7.2 'Power and MDD – validation of the study design *a priori* and *a posteriori*'. Here, we go further into the option of applying effects on simulated field study datasets to obtain a power estimate.

There are no standard tools for power analysis with advanced linear statistic. By conducting simulations we are able to get an estimation of the power of the linear statistic. Therefore, we propose a combined approach. By using the power analysis for non-parametric t-test we have a conservative measure of the sample size needed and by conducting simulations we get an estimation of the power of the linear statistic.

For the simulation a dataset should be generated with count data following a Poisson distribution for multiple fields and sessions that accounted for natural differences between the fields. Then the simulation can apply a pre-defined treatment effect be reducing the count number on treatment sites. The obtained dataset is analysed using a GLMM with Poisson distribution and the fields a random effect. In a Monte Carlo-approach the dataset generated and analysed 10 * 1000 times. The number of times - the power - a significance in the treatment-session interaction (the effect is variable over time and pre-application sessions are included) is detected, will be counted for 1000 runs. Doing this for 10 times allowed for the calculation of mean, standard deviation, and quantiles of the power. Obviously the relevant effect size needs to be defined before the simulations are done. According to EFSA, 2011: "A biologically relevant effect can be defined as an effect considered by expert judgement as important and meaningful for human, animal, plant or environmental health. It therefore implies a change that may alter how decisions for a specific problem are taken."

2.6.8.2 Minimal Detectable Difference for field effect studies

In section 2.7 'Validation of the study design and statistical analysis of field study data' in the sub-section 2.7.2 'Power and MDD – validation of the study design *a priori* and *a posteriori*' it is explained that a power analysis needs either standard deviations from previous studies or a pilot study and such requirements cannot always be fulfilled. There is, furthermore, a trade-off between high statistical power by increased number of study sites (as in the extensive approach) and depth of individual information (e.g. body weight, reproduction, survival) obtained by intensive and frequently repeated observations of a smaller number of study sites. In case such individual information is needed, the sample size required by a power analysis might not be practically feasible but the calculation of MDDs can at least enlighten the statistical value of an intensive study approach with a low(er) number of study sites.

One possible example for the use of MDDs is a field effect study focussing on the common vole. For the common vole it is especially difficult to get reliable standard deviations as the voles undergo multi-annual cycles (Lambin et al. 2006) with huge difference between low and peak years (Jacob et al 2013). A power analysis that based on standard deviations from different phases in the multiannual vole cycle may result in impractically high number of predicted study sites. On the other hand, pilot studies cannot be conducted in the same year if a long-term study covering the entire breeding season is intended. Furthermore, vole studies often aim on population data and in addition on individual body weights, several reproductive parameters and individual survival.

In the guidance document for aquatic organisms EFSA (2013) provides a ranking with four MDD classes. The best achievable class is class IV with MDD <50% and defined as "Small effects can be determined statistically". For terrestrial field effects studies there is no such ranking provided. We calculate MMD% for four real field data sets (Figure 6) on common voles. Each study was conducted on 12 to 14 sites with seven to nine trapping sessions (a total of 31 sessions) and approx. 21 days between each session. The trapping data were analysed by a GLMM with a Poisson distribution. O'Hara et al. (2010) demonstrated that the Poisson distribution should be the only distribute to assume when analysing count data, e.g. trapping data. During all except one of the 31 trapping sessions the MDDs were above the 50% threshold (Table 13) and had an overall mean of 18.0% MDD.

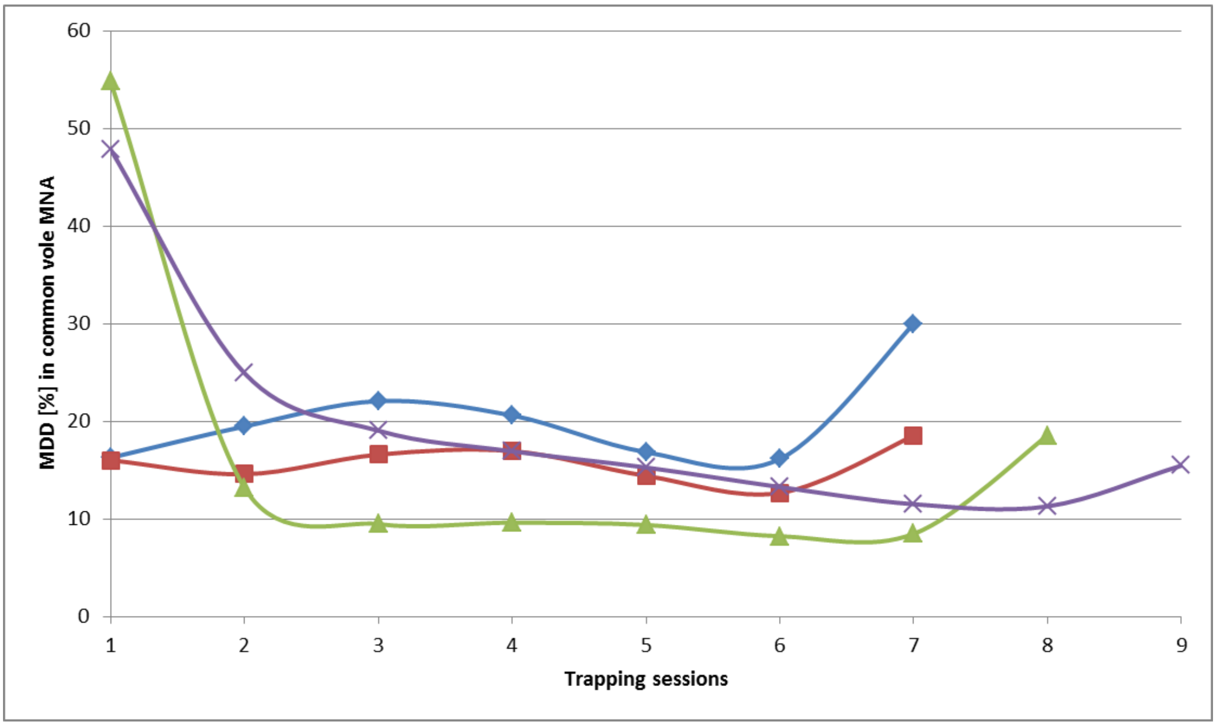


Figure 6: Minimal detectable differences in population size calculated for four different studies on common voles.

The minimal detectable difference (in %) that could be achieved in four studies on common voles to identify significant differences in population size (as MNA) was calculated for each trapping session of each particular study. For the comparison the trapping data were (re-)analysed by a linear mixed model (GLMM) based on a Poisson distribution. The results show that with regular trapping intervals (as done in all four studies) it takes about two trapping sessions to achieve MDDs below 50%. On average the field effect studies on common voles achieved a MDD of 18% meaning that differences of more than 18% between treatment and control in the GLMM could be identified as significant.

Compared to a 50% threshold for aquatic studies an average of 18.0% MDD is a very acceptable result. To evaluate what an 18.0% difference means in common vole populations, we calculated the deviation from the mean for each population on the control study fields. Only data from control fields were used to avoid any possible effect from the treatments applied during the studies. The mean value was calculated from the six respectively seven control fields (i.e. common vole populations) for every trapping session and study separately. Data from the same session were obtained within less than a month but the total study time was always more than six month. The different studies were conducted in different years and, thus, by calculating deviations just within studies and sessions the variation in vole densities over longer time periods was excluded.

Table 14: Minimal Detectable differences for 31 trapping sessions of four different common vole studies.

Trapping	MDD% for MNA			
Sessions	**Study A**	**Study B**	**Study C**	**Study D**
1	47.84	54.80	16.00	16.26
2	24.92	13.18	14.57	19.47
3	19.03	9.46	16.58	22.06
4	16.92	9.61	16.94	20.59
5	15.26	9.37	14.40	16.83
6	13.26	8.22	12.65	16.13
7	11.50	8.46	18.52	29.97
8	11.28	18.55		
9	15.45			

In 26 different common vole populations (from control fields only) the minimal deviation from the mean population size within the same trapping session and the same study (i.e. same time and area) was 17.39% and the mean deviation for all sessions was 52.76% (Table 15). In other words, a deviation of about 17% is a small (the smallest found) difference between natural common vole populations, while the average difference is around 50%. Thus, the test procedure applied in field effect studies, which is able to detect minimal difference of 18% between treatment and control populations, is a very sensitive tool. It can be assumed that all difference in population size between treatment and control that are below 18% are ecological irrelevant.

We are aware that this is a comparison within the same data set of four large field effect studies. However, the results are promising and it seems worth to apply this procedure to a larger set of field data. EFSA has access to additional data which could be analysed to get a sound evaluation of MDDs.

Table 15: Population abundance deviations in % from the mean calculated separately for each trapping session

Trapping	Deviation from mean MNA			
Session	**Study A**	**Study B**	**Study C**	**Study D**
1	109.18	101.28	32.09	45.67
2	98.40	67.58	28.54	62.89
3	90.88	63.42	26.03	50.16
4	73.81	52.96	31.76	47.12
5	83.85	33.46	35.02	39.56
6	75.22	24.57	22.98	29.99
7	69.82	26.25	17.39	32.94
8	62.33	46.85		
9	53.31			

2.6.9 References

Bonnet T, Crespi L, Brunetau L, Bretagnolle V,Gauffre B. 2013. How the common vole copes with modern farming: Insights from a capture–mark–recapture experiment. Agriculture Ecosystems & Environment 177:21-27.

Briner T, Favre N, Nentwig W, Airoldi JP 2007. Population dynamics of *Microtus arvalis* in a weed strip. Mammalian Biology. 72 (2):106–115.

Dieterlen F. 2005. Feldmaus *Microtus arvalis* (Pallas 1778). In: Die Säugetiere Baden-Württembergs (Vol. 2). Ulm, Germany: Ulmer Verlag:297 – 311.

EFSA 2009. European Food Safety Authority; Guidance Document on Risk Assessment for Birds & Mammals on request from EFSA. EFSA Journal 2009; 7(12):1438.

EFSA 2011. Scientific Committee; Statistical Significance and Biological Relevance. EFSA Journal 2011; 9(9):2372. 17pp

EFSA 2013. European Food Safety Authority; Guidance on tiered risk assessment for plant protection products for aquatic organisms in edge-of-field surface waters. EFSA Journal 2013; 11(7):3290.

Flowerdew JR, Shore RF, Simon M, Poulton C, Sparks TH. 2004. Live trapping to monitor small mammals in Britain. Mammal Rev. 34(1):31–50.

Fuelling O, Halle S. 2001. Breeding suppression in free-ranging grey-sided voles under the influence of predator odour. Oecologia 138:151–159.

Hein S, Jacob J. 2018. Population recovery of a common vole population (*Microtus arvalis*) after population collapse. Pest Management Science. doi:10.1002/ps.5211.

Hurlbert SH. 1984. Pseudoreplication and the design of ecological field experiments. Ecological Monographs 54(2):187-211.

Jacob J, Hempel N. 2003. Effects of farming practice on the spatial behaviour of common voles. Journal of Ethology 21:45-50.

Jacob J, Manson P, Barfknecht R, Fredricks T. 2013. Common vole (*Microtus arvalis*) ecology and management: implications for risk assessment of plant protection products. Pest Management Science 70(6):869-878.

Lambin X, Bretagnolle V., Yoccoz NG 2006. Vole population cycles in northern and southern europe: Is there a need for different explanations for single pattern? Journal of Animal Ecology 75:340–349.

O'Hara RB, Kotze J. 2010. Do not log-transform count data. Methods in Ecology and Evolution 1: 118–122.

Pacheco M, Kajin M, Gentile R, Zangrandi PL, Vieira MV, Cerqueira R. 2013. A comparison of abundance estimators for small mammal populations. Zoologia 30(2):182–190.

Peters B, Gao Z, Zumkier U. 2016. Large-scale monitoring of effects of clothianidin-dressed oilseed rape seeds on pollinating insects in Northern Germany: effects on red mason bees (*Osmia bicornis*). Ecotoxicology 25:1679–1690.

Pinheiro J, Bates D. 2000. Mixed-effects models in S and S-PLUS. New York, USA: Springer-Verlag New York. 528 p

Pollock KH. 1982. A capture-recapture sampling design robust to unequal catchability. Journal for Wildlife Management 46 752—757.

Reichstein H. 1964. Untersuchungen zum Körperwachstum und zum Reproduktionspotential der Feldmaus, *Microtus arvalis* (Pallas, 1779). Z. wiss. Zoologie 170:112-222.

Somerville L, Walker CH (eds.). 1990. Pesticide effects on terrestrial wildlife. London, UK: Taylor and Francis. 404 p.

2.7 Validation of the study design and statistical analysis of field study data

Authors: Ines Hotopp and Anja Ruß

2.7.1 Introduction

Ecological data is often complex and thus in need of a specialised evaluation. Sophisticated methods exist for the adequate analysis of different kinds of field study data.

The number of individuals or sites required to be able to find an effect in a PT or field effects study can be determined with a power analysis. In case this is not possible the calculation of the minimal detectable difference (MDD) a posteriori can help to validate the study. Both approaches depend on the statistical method used to or intended to be used to analyse the data. Here, mixed models are often the appropriate choice

2.7.2 Power and MDD – validation of the study design *a priori* and *a posteriori*

"Care is required to ensure that the methods chosen for detecting effects in field studies are appropriate to the study objectives and provide adequate statistical power to be useful for risk assessment and decision-making." (EFSA 2009)

A study should be designed in a way that possible effects can be found. This requires intensive planning before the study start and, furthermore, the choice of adequate statistical methods. Even the best and most adequate methods are not going to be able to detect an existing effect, if the power of the study, e.g. the capability of the study design to find an effect, was too low. Therefore, ways have to be found to ensure the validity of the planed study. A power analysis is the formal tool to provide the answer to the question of how many individuals or sites should be used in the planed study design to be able to find a certain effect with the chosen statistical methods with a given probability. To get a reliable answer data is required that gives information about the natural variability that can be expected. Thus, data from prior studies or pilot studies is needed. This is not always possible. The construct of the minimum detectable differences (MDD) was developed to find out what minimal effect would have been possible to detect with the recorded data after the study was conducted and thus prove the validity of the study.

2.7.2.1 Power analysis for field studies

According to the EFSA (2009) GD the results of field effects studies are not used to refine the toxicity exposure ratio (TER) but provide a weight of evidence argument. To do the weighting the study design, the representativeness of the location(s), agricultural practice, weather conditions etc. should be taken into account and, of course, the statistical value of the test design. As field effect studies are expensive to conduct there is a high risk that they are underpowered and fail to reliably answer the regulatory question. Hence, it has to be ensured that the field study is sufficiently robust to confide in the detection of "no effect found". Consequently, the current guidance document demands an adequate statistical power.

"[…] need sufficient number and size of sites, and sufficient variety of ecological conditions, to ensure opportunity for sensitive species to be present and to be exposed in a representative range of conditions, and to give adequate statistical power to detect effects and/or quantify their frequency."

However, what 'adequate' is needs a precise definition. In ecology the statistical power increases more by sampling a new location than repeating the measurement on an already sampled point (A. Zuur, pers.). Important points to increase the power are:

- Increase number of replicates (sites)
- Decrease standard deviation (SD) in the data set, e.g. minimise effects of other biological relevant parameters

There are certain parameters of vertebrate field studies which change during the course of the year. These are for example: body weight and population size. By focusing the field phase of the study on certain time windows the variance in the data set will be reduced for these parameters. This can be the critical phase in which the toxicological effects can be expected. For example directly after the application the weight development and at the end of the reproductive phase the population size.

A power analysis in *senso stricto* is conducted *a priori* to the study and has the goal to find the sample size (either number of fields or number of individuals, depending on the study type) needed to detect an predefined and biologically meaningful effect size (difference of endpoint between treatment and control group) with a sufficient power – usually 80% (Fairweather 1991). To determine the necessary number of study fields or individuals, the variance (as Standard Deviation) either from previous studies or a pilot study is needed.

Recently, the evaluation of field studies by member states focussed on the power of the test design beyond the actual outcome of such studies:

"For higher tier effect studies (field, semi-field), a power analysis of the test protocol is always requested and applicant should provide an argumentation for justifying the assumed variance used in the power analysis." (EFSA 2015)

A power analysis depends on the statistical test that is used to analyse the data. Several software options exists that can be used to calculate the required sample sizes for t-tests, ANOVA, and other simple statistical tests. Methods to perform a power analysis for (G)LMMs (Kain et al. 2015) and Survival models (e.g. R package 'powerSurvEpi' by Qiu et al. 2018) have been developed as well. Besides the application of software solutions it can be sensible to apply a simulated effect on half of an existing datasets in Monte Carlo runs and calculate the power by counting the number of times the statistical model was able to detect the effect. However, none has become a standard method yet as their applicability is highly dependent on the study design.

Obviously the relevant effect size needs to be defined before a power analysis can be done. According to EFSA (2011): "A **biologically relevant** effect can be defined as an effect considered by expert judgement as important and meaningful for human, animal, plant or environmental health. It therefore implies a change that may alter how decisions for a specific problem are taken." This approach should be followed in the revision of the GD.

2.7.2.2 Minimal detectable difference

A power analysis needs either standard deviations from previous studies or a pilot study and such requirements cannot always be fulfilled. In case a power analysis was not or could not be performed *a priori* to a study, the minimal detectable difference (MDD) can be calculated *a posterior* to assess the statistical value of one particular study. The MDD concept was first developed by Brock et al. (2014) for aquatic mesocosm/microcosm studies using t-tests. Peters et al. (2016) provided in the supplementary material to their article a method to apply the MDD concept to linear mixed models. The idea of MDD is to calculate (often in percent, MDD%) the minimal difference between the (predicted) control value (e.g. population size) and the (predicted) treatment that could be identified as a significant difference for the analysed data set. A MDD, however, applies only to the one data set (one specific study) for which it was calculated and cannot be extrapolated to other studies. A power analysis, in contrast, is made for a specific study design and applies to all studies conducted according to that design.

One possible example for the use of MDDs is a field effect study focussing on the common vole. For the common vole it is especially difficult to get reliable standard deviations as the voles undergo multi-annual cycles (Lambin et al. 2006) with huge difference between low and peak years (Jacobs et al. 2013). A power analysis that based on standard deviations from different phases in the multiannual vole cycle may result in impractically high number of necessary study sites. On the other hand, pilot studies cannot be conducted in the same year if a long-term study covering the entire breeding season is intended.

In the guidance document for aquatic organisms (EFSA 2013) provides a ranking with four MDD classes (Table 16). The best achievable class is class IV with MDD <50% and defined as "Small effects can be determined statistically". For terrestrial field effects studies there is no such ranking provided so far.

Table 16: Classes of minimum detectable differences (MDD) as proposed in the EFSA Aquatic Guidance Document (Brock et al., 2015)

MDD class	%MDD	Comment
0	> 100%	No effects can be determined statistically
I	90 – 100%	Only large effects can be determined statistically
II	70 – 90 %	Large to medium effects can be determined statistically
III	50 – 70%	Medium effects can be determined statistically
IV	< 50%	Small effects can be determined statistically

2.7.3 Statistical methods for the analysis of field study data

2.7.3.1 General remarks

ANOVA, t-test, and u-test are not considered to be adequate for analyzing complex ecological field data anymore. This kind of data often violates the assumptions of these simple tests (i. e. normality, homogeneity of variances, independence of data). The assumption of normality is commonly violated for most types of ecological data besides weight data (Figure 7). Field effects study data that contains data points for multiple sessions within one study field violates the assumption of independence (Hurlbert 1984). Statistical models can be adapted for different data distributions and are able to handle dependencies in the data.

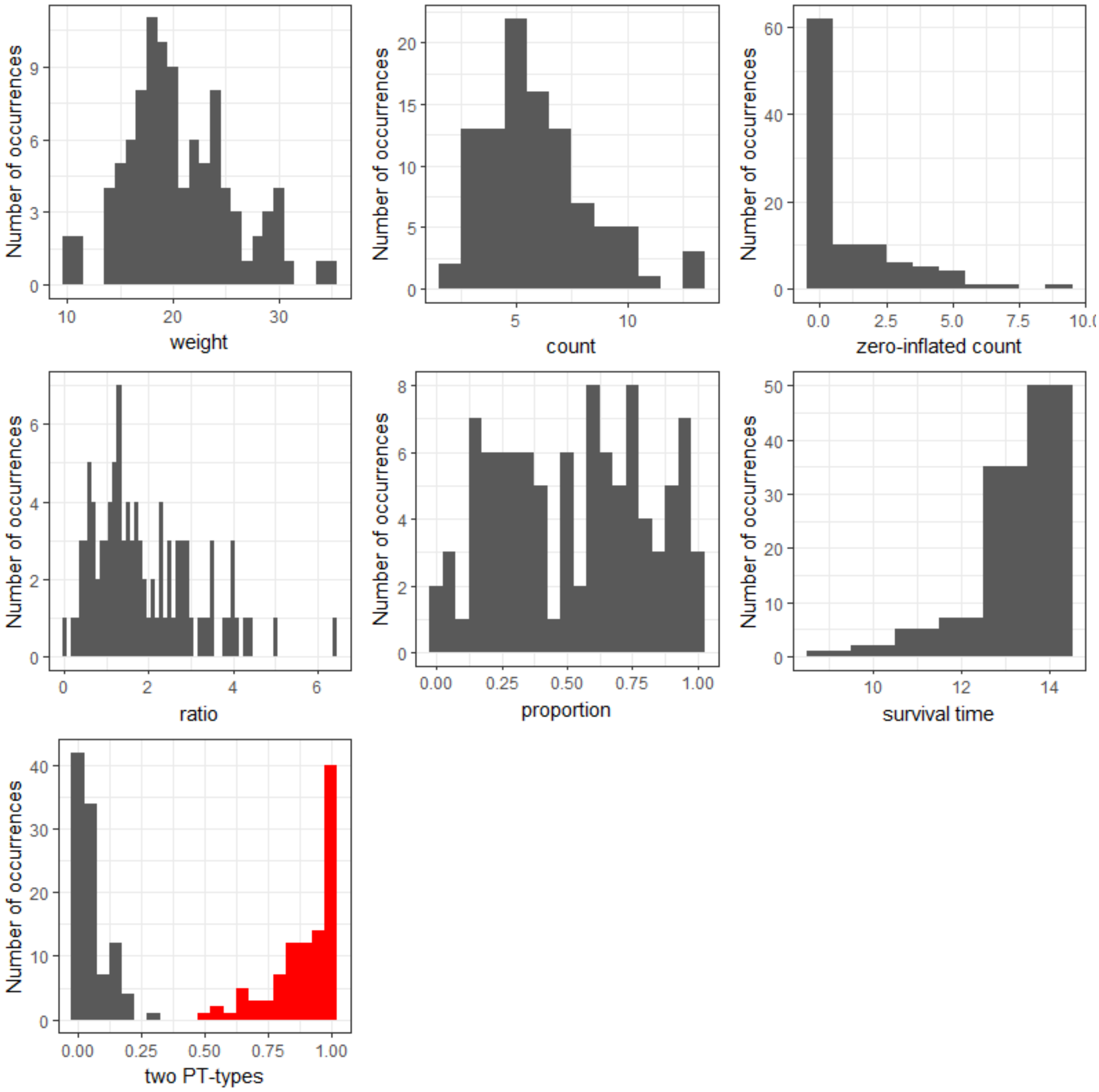


Figure 7: Exemplary histograms for different data types.

In simple statistical tests the response variable (e.g. weight or count) is related to one explanatory variable (the treatment group).

Response variable ~ explanatory variable

Statistical models are able to incorporate more than one explanatory variable such as the treatment group, time, sex, and precipitation and furthermore interactions between explanatory variables (e.g. a time dependent treatment effect).

Response variable ~ explanatory variable 1 * explanatory variable 2 + explanatory variable 3

The explanatory variables are also called 'fixed effects'. To include further explanatory variables apart from the treatment groups reduces otherwise unexplained variability in the data and increases the probability to find a real treatment-related effect.

Besides the fixed effects many of the models include random effects. Models that include fixed and random effects are called mixed-effects models. With random effects the model accounts adequately for individual differences between groups (e.g. study sites, study areas).

Response variable ~ explanatory variable 1 * explanatory variable 2 + explanatory variable 3 +
random effect

In Table 17 the abbreviations for common effect models are given. Table 18 shows which mixed-effects model is appropriate to be used for which type of data.

Table 17: Abbreviations used for common effect models.

Abbreviation	Full name	Application with distributions
LM	linear model	normal
GLM	generalized linear model	normal (defaults to LM) Poisson gamma beta binomial
GAM	generalized additive model	normal Poisson gamma beta binomial
LMM	linear mixed model	normal
GLMM	generalized linear mixed model	normal (defaults to LMM) Poisson gamma beta binomial
GAMM	generalized additive mixed model	normal Poisson gamma beta binomial

Table 18: Possible applications for selected mixed-effects models.

Statistical model	Data type
GLMM with Poisson family	count data
LMM	weight data

GLMM with gamma family	ratios
GLMM with binomial family	proportions
GLMM for zero-inflated data with a family depending on the data type	Absence/presence data of rare species (zero-inflated data)
Hurdle model	zero-inflated count data
Cox proportional hazards regression model	Survival data

Model type, explanatory variables and their usage (categorical or continuous) in the model, interactions, random effects and the distribution of the response variable are to be clearly specified to ensure transparency and reproducibility. To verify that assumptions such as independence and absence of residual patterns are not violated, the fitted model needs to be validated (Zuur & Ieno, 2016). If the data include temporal (or spatial) aspects, autocorrelation functions and/or variograms should be used to assess independence of residuals.

2.7.3.2 Linear and generalized mixed-effects models in field effects studies

Field effects studies investigate possible effects of plant protection products under realistic field conditions. In contrast to a simple but much less realistic laboratory experiment, the measurements in a realistic field study can be affected by natural factors (confounding effects) in addition to the experimental factor ‘treatment’. Some of these factors can be measured like the amount of precipitation or the time (date, session number) or are known (species, sex). Other factors, which may occur at only some study fields, cannot be measured easily and thus, cause unobserved differences between the experimental units (study fields). An example for such “unseen” and uncontrollable (i.e. naturally occurring) influence could be the activities of predators.

Mixed Models are used to account adequately for these individual differences between study sites (random effects). These models are specifically designed to incorporate (a) other explanatory variables than just the treatment and (b) a natural (random) difference between study fields, which means they can unmask side effects of the realistic field situation and therefore provide a very sensitive tool to detect any treatment related effects.

For the practical application the following points should be considered:

An important prerequisite for the calculation of a statistical model is the correct selection of a distribution family. Such a distribution can be, for example, Poisson distribution for count data (like number of captures or MNA) or Binomial for presence/absence data.

Once the distribution family is found the models have to be created. This is a crucial point where statistical and biological knowledge interacts. Such a model is a summary of all biological meaningful and practically measureable information, represented by explanatory variables (e.g. weather or treatment) that may explain the measured response variable (e.g. weight or MNA). Time is often used with ascending powers to test if the response variable changes just linear over time or shows one or more peaks and nadirs (lowest points). For example the population of voles in a long-term study is expected to increase from a low spring population towards a peak in late summer and starts to decrease afterwards. This would be a quadratic time response.

The output table of a mixed model shows the (different) tested model(s) in column(s) with the first column showing (sometimes cryptic shortcut) names of the explanatory variables and the following columns showing the results (the model estimates with their standard deviation) for each tested factor in each model. An example is given in Table 19.

Table 19: Example for the result table of a statistical analysis with a GLMM

The tested models are GLMMs with an arbitrary distribution used for a data set consisting of 312 observations for multiple time points (e.g. sessions) and treatment and control groups at 12 fields. The response variable is an arbitrary variable called 'response'; fixed effects are time and treatment for model 1 and the time-treatment interaction for model 2. The random effect is the field.
Model 1 = response ~ treatment + time + (1|field)
Model 2 = response ~ treatment * time + (1|field)

	Model 1	Model 2
Intercept	3.08 (0.05) p < 0.05	3.06 (0.05) p < 0.01
Treatment – time interaction		-0.25 (0.07) p < 0.005
Treatment	-0.10 (0.05) not significant	-0.12 (0.06) not significant
Time	0.5 (0.1) p < 0.01	0.52 (0.1) p <0.05
AIC	375.25	365.34
Number of observations	312	312
Number of levels in random effect "field"	12	12
Variance in random effect "field"	0.01	0.01

To answer the question if there is a significant effect of the PPP treatment on, e.g. the population development measured, one will look at the plotted data first. This visual observation of the MNA on treated and on control fields will give an impression, either 'Yes', there seems to be a difference or 'No', there is no obvious difference. To confirm (or not) this visual impression statistically, mixed models will be setup and the best fitting model selected by its AIC value (Akaike Information Criterion). The AIC is a method to select among the set of models (fitted with the maximum likelihood method) the one model which explains the observed (measured) data best. The model with the lowest AIC value is the best. Note that differences in the AIC value below 4 are usually regarded as negligible, i.e. the models are equally good. The AIC can only be used to select between models within the same distribution family and dataset used.

In case the best model from the result table (lowest AIC) shows no significances at the chosen significance level (usually 0.05), there is no statistical difference between treatments.

In case of a real treatment effect some facts have to be considered: The study fields are arranged into two groups (treatment and control) from the first to the last day of the field test. A significant difference in factor 'treatment' indicates, therefore, a difference between these two groups at the beginning of the study period. A real treatment effect caused by the application of the test item should occur only on the fields of the treatment group but not before the first application. Thus, a real treatment effect is indicated by a significant treatment-time interaction. However, if there is significance in the treatment-time interaction, it is necessary to check the plot of measured data and estimated model results again. The significance may be caused by a difference that existed before the application and is reduced in the course of the study. Furthermore, an effect during or after the application may be positive (e.g. for MNAs the treatment data is higher than the control data) and is, therefore, not considered as an adverse effect.

2.7.4 References

Bonnet T, Crespi L, Brunetau L, Bretagnolle V,Gauffre B. 2013. How the common vole copes with modern farming: Insights from a capture–mark–recapture experiment. Agriculture Ecosystems & Environment 177:21-27.

Brock TCM, Hammers-Wirtz M, Hommen U, Preuss TG, Ratte HT, Roessink I, Strauss T, Van den Brink PJ. 2014. The minimum detectable difference (MDD) and the interpretation of treatment-related effects of pesticides in experimental ecosystems. Environ Sci Pollut Res 22:1160-1174.

EFSA 2009. European Food Safety Authority; Guidance Document on Risk Assessment for Birds & Mammals on request from EFSA. EFSA Journal 2009; 7(12):1438.

EFSA 2011. Scientific Committee; Statistical Significance and Biological Relevance. EFSA Journal 2011; 9(9):2372. 17pp

EFSA 2013. European Food Safety Authority; Guidance on tiered risk assessment for plant protection products for aquatic organisms in edge-of-field surface waters. EFSA Journal 2013; 11(7):3290.

EFSA 2015. European Food Safety Authority; Technical report on the outcome of the pesticides peer review meeting on general recurring issues in ecotoxicology. EFSA supporting publication 2015:924. 62pp.

Fairweather PG. 1991. Statistical Power and Design Requirements for Environmental Monitoring. Marine and Freshwater Research 42:555-567.

Hurlbert SH. 1984. Pseudoreplication and the design of ecological field experiments. Ecological Monographs 54(2):187-211.

Jacob J, Manson P, Barfknecht R, Fredricks T. 2013. Common vole (*Microtus arvalis*) ecology and management: implications for risk assessment of plant protection products. Pest Management Science 70(6):869-878.

Lambin X, Bretagnolle V., Yoccoz NG 2006. Vole population cycles in northern and southern europe: Is there a need for different explanations for single pattern? Journal of Animal Ecology 75:340–349.

Kain MP, Bolker BM, McCoy MW. 2015. A practical guide and power analysis for GLMMs: detecting among treatment variation in random effects. PeerJ 3:e1226. doi: 10.7717/peerj.1226

Lauenstein G, Barten R. 2011. Management von Feldmäusen in der Landwirtschaft. Unna, Germany: Frunol Delicia GmbH. 1-160p.

Peters B, Gao Z, Zumkier U. 2016. Large-scale monitoring of effects of clothianidin-dressed oilseed rape seeds on pollinating insects in Northern Germany: effects on red mason bees (*Osmia bicornis*). Ecotoxicology 25:1679–1690.

Qiu W, Chavarro J, Lazarus R, Rosner B, Jing M. 2018. Package 'powerSurvEpi'. CRAN repository. https://CRAN.R-project.org/package=powerSurvEpi. Accessed 14th December 2018.

Zuur AF, Ieno EN, Freckleton R. 2016. A protocol for conducting and presenting results of regression-type analyses. Methods in Ecology and Evolution 7(6):636-645.

3 Appendices

Appendix- I Tables to section 2.1 - Identification of focal species

Table A 1: Birds: Proposal for methodological approaches to FS selection depending on crop and crop stage

Crop	BBCH (Definition)				
	00-08 (Pre-emergence)	**09-14/15** (First leaves)	**15-29** (Leaf development)	**30-59** (Stem, inflorescence)	**>60** (Flowering, fruit, ripening)
Bare soil	Point count*	-	-	-	-
Grassland	Transect count				
Maize	Point count	Point count	Transect count	Transect count Mist-netting	Mist-netting
Potatoes		Point count	Transect count	Transect count	Transect count
Sugar beet		Point count	Transect count**	Transect count	Transect count
Sunflower		Point count	Transect count	Transect count Mist-netting	Mist-netting
Cereal	Point count	Point count Transect count	Transect count	Transect count	Transect count
Oilseed rape		Point count Transect count	Transect count	Transect count	Mist-netting

* Point count (EFSA (2009)GD) = Scan sampling technique

**until BBCH 31 ground cover is not described, however the structure of beet plants makes observations difficult between and within rows already before

When 2 methods are proposed, both can be applied

Table A 2: Birds (cont.)

Crop	Stage				
	Without leaves	**Flowering**		**Foliage developm.**	**Full foliage**
Orchards	Mist-netting Transect count	Mist-netting (Transect count)		Mist-netting (Transect count)	
	Without leaves	**First leaves**	**Leaf developm.**	**Flowering**	**Ripening**
Vineyards	Mist-netting Transect count	Mist-netting (Transect count)		Mist-netting (Transect count)	

Table A 3: Mammals: Proposal for methodological approaches to FS selection depending on crop and crop stage

Crop	BBCH (Definition)				
	00-08 (Pre-emergence)	**09-14/15** (First leaves)	**15-29** (Leaf development)	**30-59** (Stem, inflorescence)	**>60** (Flowering, fruit, ripening)
Bare soil	Trapping in-crop and off-crop	-	-	-	-
Grassland	Trapping in-crop; Point count* or Transect count** (depending on the vegetation height)				
Maize	Trapping in-crop and off-crop Point count* Transect count**	Trapping in-crop and off-crop Transect count**	Trapping in-crop Transect count**	Trapping in-crop	Trapping in-crop
Potatoes					
Sugar beet					
Sunflower					
Cereal					
Oilseed rape					

* Point count (EFSA GD) = Scan sampling technique

** Transect count (EFSA GD) = Spot-light or night-device count

Table A 4: Mammals (cont.)

Crop	Stage				
	Without leaves	**Flowering**		**Foliage developm.**	**Full foliage**
Orchards	Trapping in-crop Transect count*	Trapping in-crop Transect count*		Trapping in-crop Transect count*	
	Without leaves	**First leaves**	**Leaf developm.**	**Flowering**	**Ripening**
Vineyards	Trapping in-crop Transect count*	Trapping in-crop Transect count*		Trapping in-crop Transect count*	

* not conducted from a moving car but on foot in a 90° angle to the tree lines

Appendix- II Suggestions for the approach to generate correction factors

For the example given in this approach an omnivorous passerine, that is known to feed also on cereal seeds has been chosen. This involves the need to test seeds and arthropods. For insectivorous species the entire procedure is easier because of the lack of the need to test seeds.

General approach

For the feeding trials birds should be moved into a 'trial-cage' (e.g. for small passerines a cage with 80cm width x 50cm depth x 60cm height would be sufficient) where they are kept individually. The 'trial-cage' laterals should preferably consist of plastic. This construction minimises the potential loss of food, food remains and faecal samples.

At the beginning of the feeding trial the cage should not contain any food remains, no faeces remains or any other remains from former trials. Water should be offered to the bird always *ad libitum*.

Feeding trials should be conducted in the morning after approximately 12 hours of food deprivation in order to approach emptiness of the digestive tract.

The diet that is offered during the trial is referred to as 'trial food'.

The durations for the trials (period for the presentation of the trial food and subsequent period for collecting the faeces samples) has to be adapted appropriately. For example if periods are too short (e.g. not enough different food items are ingested to get sufficient remains of them in the faeces) they will be changed appropriately. Periods for each trial and the cause (if periods have changed in relation to a former trial) should be recorded.

Before the trials all different food types offered have to be weighed with an analytical balance. For this purpose the entire amount of one food type should be weighed and the number of the items (e.g. seeds) counted. Both values (weight for all items and number of items) should be recorded in order to calculate the mean weight per item.

For food types other than seeds (e.g. arthropods) their length and weight, respectively, should also be taken and recorded. At least 10 specimens of each food type should be measured for smaller food types (e.g. <3mm) of which more than 10 items are offered. If less than 10 items are offered (e.g. for larger specimens) then all should be measured. Characteristic parts likely to be recovered in faeces samples should also be measured in order to be able to deduce the length of the entire arthropod from the size of these structures.

Estimating the proportion of husk removed from seeds before ingestion

The aim of this trial is to get an approximate value for the average proportion of the husk of seeds removed by the trial species before ingestion ('de-husking rate'). If the species is likely feeding on very different types of seeds, typical characteristic specimen should be tested in separate because the 'de-husking rate' might be seed type specific.

The trial can be scheduled as follows. The duration of each phase should be documented.

'Trial-food-offered phase': During this feeding trial only one seed type with husks will be offered to the trial species ad libitum at the beginning.

'Trial-food-derivation phase': Then the trial food should be taken away and for the subsequent time of the trial either no food or a different food (that produces easily recognizable remnants in the faeces) should be offered in order to facilitate normal digestion. Afterwards the trial will be stopped by collecting all faeces samples and offering the standard food to the bird.

After the trial the husk on the cage bottom should be collected and weighed. Moreover ten seeds of the seed type that was offered to the bird should be de-husked manually and the husk weighed and divided by ten in order to get information of the 'average husk weight per seed'. The number of seeds ingested by the birds can be calculated from the difference between the number of seeds offered and the number that remained when the seeds were removed. By dividing the 'weight of the husks found on the ground' by the product of the two factors 'number of seeds ingested' and the 'average husk weight per seed', the 'average de-husking rate' can be calculated for the specific seed type.

Detectability of seeds ingested by birds in their faeces

The aim of this trial is to find out if remains of seeds ingested can be found in the faeces.

From the trial described above all faeces samples from the bottom of the cage will be collected. The content of the samples will be analysed as described in 'Analysis of faeces samples'. It will be checked if any structures can be found that can clearly be assigned to the ingested seed type.

If no structures can be found in the faeces samples that can be identified as parts of the seeds ingested unambiguously, or if structures cannot be quantified properly (i.e. the surface area or the number of items cannot be measured) or if the individuals tested show a considerable between and/or within variability regarding the relation between number of seeds ingested and the amount of remains found in the faeces (e.g. because of different de-husking rates), no reliable correction factor might be derivable for this seed type.

However, if remains will be found that can clearly be assigned to the seed type ingested and that can be quantified properly and if the tested individuals show a tolerable de-husking variability, the following trials should be conducted in order to assess the proportion of food type specific remains in the faeces in relation to the proportion of these food types in the diet ingested.

Comparison between the quantity of selected food items ingested and the quantity of their remains found in the faeces

The aim of this trial is to get information about the quantity of remains of different food types found in the faeces in comparison to the quantity of these food types originally ingested in order to derive food type specific correction factors. It is suggested to start with a first 'few items' approach in order to test the general feasibility of this approach for a few food types only (see 'few items' approach). If this 'few items' approach turns out to be successful a 'multiple items' approach (see 'multi items' approach) can be conducted with pre-defined food types (see Table A 5).

'Few items' approach

'Trial-food-offered phase': At least three different food types should be offered to the bird *ad libitum* at the beginning: (i) crop seeds (e.g. cereals), (ii) arthropods and (iii) a type of seeds that is known to cause quantifiable remains in the faeces of granivorous birds (e.g. *Fallopia convolvulus* or

Caryophyllaceae). The type of seeds and arthropods should be the same for all trials of this 'few items' approach.

'Trial-food-deprivation phase': Subsequently the food offered initially should be removed and either no other food or a different food (that produces easily recognizable remnants in the faeces) should be offered in order to facilitate normal digestion. Afterwards the trial should be stopped by collecting all faeces samples and offering a standard food to the bird.

The duration of each phase should be recorded.

The number of seeds and arthropods ingested by the birds can be calculated from the difference between the number of items offered and the number that remained at the end of the trial, respectively.

All faeces samples in the cage should be collected. The content of each of the samples should be analysed as described in 'Analysis of faeces samples'.

The quantity (number or surface area) of structures recovered has to be divided by the number of items eaten by the bird. The obtained value (the quantity of recovered structure per items eaten) is the 'food item specific correction factor' that is aimed at.

If it turns out that a bird does not eat all offered food types in sufficient quantities (i.e. no remains could be found in the faeces of the initial trial) the schedule how the different food items are provided has to be changed. In this case the food items are offered in a new trial in modified quantities. The offered quantities should be recorded for each trial. The offered quantities that finally lead to the sufficient ingestion of all food types (i.e. remains of all food types would be discovered in the faeces) should be recorded. These quantities might differ between the different bird individuals tested.

'Multiple items' approach

This approach should be conducted in the same way as the 'few items' approach (see 'few items' approach) but with more diet items offered (see Table A 5 as an example). The selection of the food types offered can be based on the following criteria. Each 'diet category' likely to be taken by the trial bird species living in the wild should be presented by one of the diet types offered in the trial. However, it is mainly important that the presented items represent diet types with similar expected recovery rates in the faeces of the trial bird species rather than the application of any classification criteria. Recovery rates are expected to be similar for items with similar dimensions, shapes and composition. E.g. Coleoptera larvae might be taken by the trial bird species in the wild, but they may not be tested in the trial. However, the larvae of Lepidoptera may be tested and because Coleoptera larvae are similar to Lepidoptera larvae it is expected that the correction factor obtained for Lepidoptera larvae in the trial can also be assigned to Coleoptera larvae. In order to test representatives of all diet items likely to be taken by the trial bird species a comprehensive literature research of the diet known to be taken by the trial bird species should be conducted.

The list of diet types in Table A 5 gives an example for the trial food that can be offered to a theoretical trial bird species. The diet types were derived from Green (1978) (plant diet types) and from Glutz & Bauer (1997) (animal diet types). The plant diet types used by Green (1978) showed different recovery rates for Skylarks (*Alauda arvensis*) and cover the most common potential plant diet types also likely to be taken by other omnivorous songbirds occurring in agricultural areas in Europe. The animal diet types taken from Glutz & Bauer (1997) are common prey types likely to differ from each other regarding their recovery rate (Arachnida).

In this 'multiple item' approach members of the following food types will be offered (see Table A 5).

Table A 5: Diet types suggested to be tested in a 'multiple item' approach for an omnivorous passerine

Food type	Example[1]
Plant diet	
cereal grains	oat (*Avena sativa*) seeds
'small' grass seeds (approx. from 2.5 to 4.5 mm length)	Seeds of *Poa annua* and/or *Dactylis glomerata*
'small' dicotyledonous seeds (approx. from 2.5 to 4.5 mm length)	Seeds of Polygonaceae (*Polygonum lapathifolium/maculos* or *Fallopia convolvulus*)
'very small' dicotyledonous seeds (approx. from 1.0 to 1.5 mm)	Seeds of Caryophyllaceae *(Cersatium* spec. and/or *Stellaria media*)
monocot leaf	wheat leaf
dicot cotyledons and leaves	sugar beet cotyledons
Animal diet	
spiders (Arachnida)	Lycosidae
beetles (Coleoptera)	Curculionidae, Scarabaeideae, Elateridae
Orthoptera	grasshoppers (Acrididae)or crickets (Gryllidae)
bugs (Heteroptera)	Nabidae and/or Miridae
butterfly larvae (Lepidoptera larvae)	larvae of Galleriinae (e.g. *Galleria mellonella*)
adult butterfly	Lepidoptera
crane flies (Tipulidae)	*Tipula* spec.
sawflies, wasps, ants, bees (Hymenoptera)	Apoidea
snails and slugs (Gastropoda)	Stylommatophora (preferably Succineidae)

[1] The examples show candidates of the different food types, that may be easy accessible. Of course it is possible to replace them by other candidates if they are more convenient to be obtained.

If it turns out that a trial bird does not eat all offered food types in sufficient quantities (i.e. no remains could be found in the faeces of the initial trial) the schedule how the different food items are provided may have to be changed. In this case the food items can for example be offered in a new trial in modified quantities. The offered quantities that finally lead to the sufficient ingestion of all food types (i.e. remains of all food types would be discovered in the faeces) might differ between the different bird individuals in the trial.

Sample collection and storage

Single faeces samples can be taken from the bottom of the cage by tweezers, a knife or a scalpel. Samples should preferably be gathered as completely as possible. Each single sample should be transferred and subsequently stored in a small lockable vessel separately. The vessel can be filled up with table salt (NaCl) for preservation. No further arrangements are required to preserve the samples until their analysis. All vessels should be labelled appropriately with a sticker. Further details of the sampling (e.g. trial, date) should be recorded on an appending data sheet, with the respective sample ID on it.

Analysis of faeces samples

Faecal samples collected during the trials should be analysed separately. For analysing each faecal sample, water can be added until the salt is entirely dissolved. Subsequently the matrix can be surveyed using a binocular microscope (magnification 20x and subsequently 40x). Microscopic observations (max. magnification 400x) may also be used to assign the remains found in the samples to the food type originally ingested by the bird (reflected light microscopy and transmission light microscopy; see also Flinks, 2013).

Remains from the arthropods fed during the trial and recovered in the faeces samples should be assigned to the level that was used for the trial food types that were offered to the birds during the feeding trial. The size of characteristic parts of invertebrates (e.g. chitin fragments of arthropods) should be measured with a measuring ocular, which can reach an accuracy of ± 0.1mm. The obtained sizes are to be compared to the measurements from specimens that were measured before the trial in order to deduce the dimensions of the entire food items ingested.

In order to quantify the number of food items (e.g. number of invertebrates) within each faeces sample, all food fragments recovered in the sample should be counted and recorded, and the minimum number of individuals required to account for the number of assigned remains should be calculated. For example, two right mandibles and one left mandible of a beetle species can be attributed to (at least) two individuals. The number of each of such characteristic structures (as e.g. shell pieces, chelicera, first tibia, mandible, elytron or wings) should also be assessed.

Additionally the area of each fragment assigned to a specific invertebrate group should be assessed in order to be able to calculate the entire surface area of the remains of each invertebrate type in the sample.

All remains of seeds should also be identified to the seed types offered as good as possible. For each remain the part of the seed (e.g. pericarp, seed coat, aleurone layer) and the surface area (with a measuring ocular, which can reach an accuracy of ± 0.1mm) should be recorded.

Data evaluation and statistics

Use of faeces samples for the determination of correction factors *f* for each food type

As described in previous section the content should be assessed for each single faeces sample separately. However, for the determination of the correction factor *f* for each food type the number (n_r) or area (a_r) of remnants recovered in all faeces together will be divided by total number ingested (n_t) during the "Trial-food-offered phase" of the respective trial. Eventually the correction factor *f* describes the number or area of remnants in the faeces per food item eaten. Where different types of remnants are counted for a diet type, the *f* will be given for each type separately.

$$f = \frac{n_r}{n_t} \tag{1}$$

$$f = \frac{a_r}{n_t} \tag{2}$$

In order to apply the obtained correction factors f for each food type to faeces samples obtained from wild birds of the species that served as trial birds (or closely related species which may digest their food similarly) to estimate their diet composition, the following calculations have to be performed (according to Green & Tyler, 1989).

PD calculations by numbers or areas of food types

If *f* is the number or area of remnants recovered in the faeces per food item eaten and there are a total of k food types, then the proportion p_j, in the diet by numbers of food type *j*, from a faecal sample from a wild bird will be given (according to Green & Tyler, 1989) by

$$p_j = \frac{(n_j/f_j)}{\sum_{i=1}^{k}(n_i/f_i)} \tag{3}$$

where *n* are the counts of fragments for the different food types in the faecal sample.

PD calculations by mass of food types

In order to estimate the composition of the diet in terms of fresh or dry weight rather than numbers of food items, estimates of the mean mass per food item m can be incorporated in the formula (1) (according to Green & Tyler, 1989)

$$p_j = \frac{(m_j n_j/f_j)}{\sum_{i=1}^{k}(m_i n_i/f_i)} \tag{4}$$

Statistics

At least descriptive statistics, such as mean value, standard error of the mean, the 50th and 90th percentile, should be applied to calculate the required results from the trial data.