TEXTE
144/2025

Final report

Risk management for plant protection products: Higher uncertainties by combining multiple measures? A statistical analysis

by:
Ludwig A. Hothorn
Retired from Leibniz University Hannover

publisher:
German Environment Agency

German Environment Agency

Umwelt Bundesamt

TEXTE 144/2025

Project No. 190640
FB001845/ENG

Final report

Risk management for plant protection products: Higher uncertainties by combining multiple measures? A statistical analysis

by

Ludwig A. Hothorn
Retired from Leibniz University Hannover

On behalf of the German Environment Agency

Imprint

Publisher
Umweltbundesamt
Wörlitzer Platz 1
06844 Dessau-Roßlau
Tel: +49 340-2103-0
Fax: +49 340-2103-2285
buergerservice@uba.de
Internet: www.umweltbundesamt.de

Report performed by:
Leibniz University Hannover (retired)
c/o L. A. Hothorn
Im Grund 12
31867 Lauenau
Germany

Report completed in:
May 2025

Edited by:
Section IV 1.3-2 Environmental Exposure and Groundwater Risks of Plant Protection Products
Dr. Ulrike Krug und Dr. Konstantin Kuppe

DOI:
https://doi.org/10.60810/openumwelt-7932

ISSN 1862-4804

Dessau-Roßlau, November 2025

Abstract: Risk management for plant protection products: Higher uncertainties by combining multiple measures? A statistical analysis

In the environmental risk assessment of plant protection products, the risk of spray drift losses to the off-field can be reduced by adding drift reduction measures. The question is how the combination of several measures affects the overall risk and, in particular, the associated uncertainty. Various statistical methods have been used to quantify both the overall risk and its uncertainty, in particular the increase in power and the width of a 95% confidence interval where the additional measures are modelled as qualitative factors. It can be concluded that for most of the scenarios considered, power is reduced or only marginally increased. I.e., adding further measures tends to increase uncertainty. This is mainly due to the fact that the power of the 'distance' factor is already very high and it is naturally difficult to increase it further. This relationship is even stronger for the confidence interval model, where it actually increases as additional factors are included in the experimental design with increasing sample size. These trends have also been confirmed empirically using selected experimental data from the SETAC DRAW database, albeit only selectively.

Kurzbeschreibung: Risikominderung beim Einsatz von Pflanzenschutzmitteln: Höhere Unsicherheiten durch die Kombination mehrere Maßnahmen? Eine statistische Analyse

Bei der Umweltverträglichkeitsprüfung von Pflanzenschutzmitteln kann das Risiko der Abdrift von Spritzmitteln in Nichtzielflächen durch weitere technische Maßnahmen zur Verringerung der Abdrift reduziert werden. Die Frage ist, wie sich die Kombination mehrerer Maßnahmen auf das Gesamtrisiko und insbesondere auf die damit verbundene Unsicherheit auswirkt. Es wurden verschiedene statistische Methoden angewandt, um sowohl das Gesamtrisiko als auch dessen Unsicherheit zu quantifizieren, insbesondere die Zunahme der Güte eines Tests und die Breite eines 95%igen Konfidenzintervalls, wenn die zusätzlichen Maßnahmen als qualitative Faktoren modelliert werden. Es kann festgestellt werden, dass für die meisten der betrachteten Szenarien die Güte verringert oder nur geringfügig erhöht wird. D.h., die Hinzufügung weiterer Maßnahmen erhöht tendenziell die Unsicherheit. Dies ist hauptsächlich darauf zurückzuführen, dass die statistische Güte des Faktors „Abstand“ bereits sehr hoch ist und es offensichtlich schwierig ist, sie weiter zu erhöhen. Diese Beziehung ist beim Konfidenzintervallmodell sogar noch stärker ausgeprägt, da sie mit zunehmender Stichprobengröße steigt, wenn zusätzliche Faktoren in den Versuchsplan aufgenommen werden. Diese Tendenzen wurden auch empirisch anhand ausgewählter experimenteller Daten aus der SETAC DRAW-Datenbank bestätigt, wenn auch nur punktuell.

Table of content

List of figures

List of tables

List of abbreviations

Abbreviation	Explanation
BBA	(ehemalige) Biologische Bundesanstalt
Power π	Statistical term: power
f-	Statistical term: false negative error rate
f+	Statistical term: false positive error rate
H_0	Statistical term: null hypothesis
H_1	Statistical term: alternative hypothesis
CI	Statistical term: Confidence interval
μ	Statistical term: expected value
ANOVA	Analysis of variance
ES	Effect size
k	Number of factor levels
VT	Variance term
df	Statistical term: degree of freedom
α	Statistical term: level of false positive rate, commonly 5%
ED	Statistical term: effective distance
A, B, A*B	Definition for factor A, factor B and their interaction A*B
FWER	Statistical term: familywise error rate
N	Statistical term: total sample size
n_i	Statistical term: sample size per level i

Summary

In the environmental risk assessment of plant protection products, the risk from spray drift can be reduced by considering drift-reducing spray technologies. Additional risk mitigation measures may be considered to further reduce spray drift to acceptable levels. It is being discussed whether the existing risk management options should be extended, i.e. whether the possibility of cumulating more than two risk mitigation measures should be used. The aim of this report is to answer the question how the combination of several measures affects the assessment of the overall risk and the associated uncertainty in the assessment.

Various statistical methods are used to quantify both the overall risk and its uncertainty. In particular, finding appropriate combinations of several measures, such as nozzle type, drift reduction class of the nozzle, boom height, speed or pressure, to influence the primary relationship "spray drift deposition - distance" so that drift is minimized for different crops. As a first step, statistical models were derived from existing spray drift experimental data. Based on these models and associated parameter assumptions, uncertainty was quantified using the concepts of i) maximum power of the underlying F-test in the analysis of variance with the power defined as (1 minus false negative decision rate) and ii) minimum 95% confidence interval for the estimated distance for, e.g., 90% reduction in drift deposition. The different measures are modeled as qualitative factors. Primarily, the usual design of a completely randomized experiment with a predetermined total sample size is considered. Second, another design is considered where the total sample size increases with additional factors.

As a result of various calculations and simulations, it can be concluded that in most of the scenarios considered, power is reduced or only marginally increased. In other words, the addition of further measures tends to increase uncertainty. This is mainly due to the fact that the power of the 'distance' factor is already very high and it is naturally difficult to increase it further.

This relationship is even stronger in the model of the width of the confidence interval, where it even increases when additional factors are included in the model with increasing sample size, particularly if qualitative interactions are present. These tendencies have also been impressively confirmed empirically using selected experimental data from the SETAC DRAW database, albeit only selectively.

Zusammenfassung

Im Rahmen der Umweltrisikobewertung von Pflanzenschutzmitteln wird das Risiko für Sprühabdrift abgeschätzt. Eine Maßnahme zur Verminderung des Risikos durch Abdrift ist der Einsatz abdriftmindernder Sprühtechnologien. Durch die Anwendung solcher Risikominderungsmaßnahmen kann eine Zulassung von risikoreicheren Pflanzenschutzmitteln möglich werden, da das prognostizierte Risiko für Nichtzielflächen durch die Maßnahmen auf ein annehmbares Maß verringert wird. Derzeit befinden sich die Interessensgruppen im Austausch darüber, ob die bestehenden Risikomanagementoptionen erweitert werden sollten, d.h. ob die Möglichkeit der Kumulierung von mehr als zwei Risikominderungsmaßnahmen genutzt werden sollte. Das Ziel dieses Berichts ist es, die Frage zu beantworten, wie sich die Kombination mehrerer Maßnahmen auf die Bewertung des Gesamtrisikos und die damit verbundene Unsicherheit bei der Bewertung auswirkt.

Zur Quantifizierung sowohl des Gesamtrisikos als auch seiner Unsicherheit werden verschiedene statistische Methoden verwendet. Insbesondere geht es darum, geeignete Kombinationen mehrerer Maßnahmen, wie z. B. Düsentyp, Reduktionsetikett, Gestängehöhe, Geschwindigkeit oder Druck, zu finden, um die primäre Beziehung „Deposition durch Sprühabdrift und Abstand“ so zu beeinflussen, dass die Abdrift für verschiedene Kulturen minimiert wird. In einem ersten Schritt wurden statistische Modelle aus vorhandenen Versuchsdaten zur Abdrift abgeleitet. Auf der Grundlage dieser Modelle und der zugehörigen Parameterannahmen wurde die Unsicherheit quantifiziert, indem i) die maximale Güte des zugrundeliegenden F-Tests in der Varianzanalyse, wobei die Güte definiert ist als (1 minus der falsch-negativ Fehlerrate) und ii) das minimale Konfidenzintervall für den geschätzten Abstand, für z. B. 90% Verringerung der Sprühdeposition verwendet wurden. Die verschiedenen Maßnahmen werden als qualitative Faktoren modelliert.

In erster Linie wird das übliche Design eines vollständig randomisierten Experiments mit einer vorgegebenen Gesamtstichprobengröße betrachtet. In zweiter Linie wird ein anderes Design betrachtet, bei dem die Gesamtstichprobengröße durch zusätzliche Faktoren entsprechend erhöht wird.

Als Ergebnis verschiedener Berechnungen und Simulationen lässt sich feststellen, dass in den meisten der betrachteten Szenarien die Güte verringert oder nur geringfügig erhöht wird. Mit anderen Worten: Die Hinzufügung weiterer Maßnahmen erhöht tendenziell die Unsicherheit. Dies ist vor allem darauf zurückzuführen, dass die Güte des Faktors „Abstand“ bereits sehr hoch ist und es daher schwierig ist, sie weiter zu erhöhen.

Noch stärker ist dieser Zusammenhang im Modell der Breite des Konfidenzintervalls, wo diese sogar noch deutlich zunimmt, wenn mit zunehmendem Stichprobenumfang zusätzliche Faktoren in das Modell aufgenommen werden und/oder qualitative Wechselwirkungen vorliegen. Diese Tendenzen wurden auch empirisch anhand ausgewählter experimenteller Daten aus der SETAC DRAW-Datenbank eindrucksvoll bestätigt, wenn auch nur selektiv.

1 Introduction

When a competent authority assesses a plant protection product (PPP), principally three regulatory decision can be taken:

- the product can be approved if an unacceptable risk can be excluded for all areas of risk assessment; and
- there is a risk but the product can be authorized if the risk is minimized by appropriate risk mitigation measures; or
- authorization is refused because the risks are unacceptably high.

In the environmental risk assessment, a risk may be posed by spray drift, i.e., the deposition of droplets containing active ingredients outside the target area during spray application. As a result, pesticides may be carried into water bodies and other non-target areas. In the PPP authorization process, two main measures are established to reduce drift:

- the use of drift-reducing spray technology, and
- increasing the distance from non-target areas.

The implementation of further risk mitigation measures may allow the authorization of high-risk plant protection products as the predicted risk for non-target areas may be reduced to an acceptable level. In this context, it is being discussed whether the existing system of risk management options should be extended, i.e., whether the possibility of cumulating more than two risk mitigation measures (the 'toolbox' approach) should be realized. The question of how the combination of several measures would affect the assessment of the overall risk and the associated uncertainty in the assessment is becoming increasingly important, as each individual mitigation option is associated with some uncertainty. One way to approach this question is to derive selected models to quantify the uncertainty. Based on these results, recommendations for the overall uncertainty are given for selected examples.

This report considers the statistical design and evaluation of spray drift experiments. A particular objective is to find an appropriate combination of several measures, such as nozzle type, drift reduction class of the nozzle, boom height, speed or pressure to influence the ‘spray drift deposition-distance’ relationship so that the drift is minimized for different crops and their specific conditions (such as hedgerows). The initial proposal entailed the analysis of a maximum of five factors. Due to the considerable statistical demands, the decision was taken to restrict the analysis to a maximum of three factors.

The first step of the analysis is to derive a statistical model. Based on this model, a number of analyses are performed to derive generally valid conclusions or, if this is not possible due to the complexity, at least to identify probable trends for the uncertainties. Both the model and the conclusions are essentially based on assumed model parameters. These may or may not be appropriate, which is a critical point for the modelling.

‘Uncertainty’ is quantified by maximizing the power of the underlying statistical tests, such as F-test in analysis of variance taking multiple measures into account. This widely used approach is limited by the inherently high power of the spray drift deposition-distance dependency modelled as primary factor and the simple 100% limit of power.

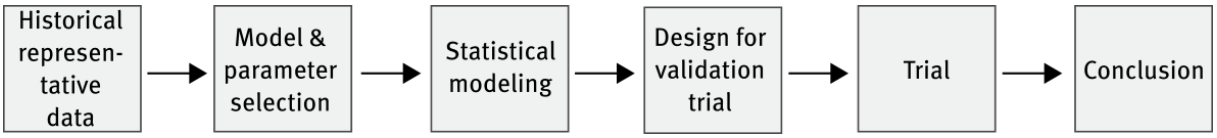
The following statistical methods were used in the report: i) power of the F-test in factorial designs (calculated or simulated depending on the various scenarios, ii) modelling of qualitative

and quantitative interaction effects in factorial designs, iii) width of the 95% confidence interval of the distance estimated for 90% spray drift reduction based on a three-parametric log-logistic model.

2 Representative Data

Representative data from historical trials are well suited for model and parameter assumptions. As a next step, specially designed field trials for validation would be recommended for final quantitative and qualitative, reproducible conclusions (Figure 1).

Figure 1: Flowchart to illustrate a design for a statistical evaluation process of experimental data



Source: own illustration, L. A. Hothorn.

A design for such a validation trial should at least include

- the type and number of measures considered (in addition to the primary factor 'spray drift deposition distance'),
- the number of factor levels (e.g. number and type of nozzles),
- the calculated number of plots (to ensure a given power),
- a completely randomized design on the trial area (or a block design, if applicable) and
- the number of independent trials within a crop or between crops.

However, in this project it was not possible to collect data after a validation step and thus data from a validation trial could not be used for the statistical analysis here. Therefore, historical representative data were used for statistical modelling and the validation step was omitted.

Two data sets were available for the project. The first data set included spray drift deposition values from spray drift trials conducted by the Biologische Bundesanstalt für Land- und Forstwirtschaft (BBA) in Germany during the years 1989 to 1992 and 1996 to 1999 (Ganzelmeier et al. 1995, Rautmann et al. 2001), hereafter referred to as BBA data. These experiments were conducted under controlled field conditions and without drift reduction nozzles. The BBA data have been used to establish the basic drift values for different crops, which provide a reference for spray drift modelling and the development of risk mitigation measures.

Additional data for the statistical analysis of this project were obtained from the SETAC DRAW database, which contains spray drift data from various European countries (hereafter referred to as DRAW data). The SETAC DRAW working group was established to facilitate the understanding of spray drift to improve the regulatory basis for risk assessment (Mel's Vineyard 2023). One objective was to develop a database of existing spray drift trials that were conducted under diverse conditions. Access to this database only became available toward the end of the project, limiting the amount of data that could be used (as detailed below). The advantage of this structured dataset is the availability of different reduction levels of the nozzles, different experimenters and widely varying sample sizes.

The SETAC DRAW database contains multi-regional spray drift data in a standardized format, i.e. data from eight European countries (Belgium, Denmark, Germany, France, Italy, the Netherlands, Poland and the United Kingdom) with one or more test series documented for each country (Miller 2024). Each test series contains a certain number of trials (i.e. replicates). The trials in a

series were usually carried out at the same test site with a specific crop, nozzle type, pressure, driving speed, boom width and/or height etc.. Meteorological parameters were recorded. Spray drift deposition was collected in Petri dishes at different distances. The Petri dishes will be referred to as plots. A trial consisted of several plots, e.g., 6 to 10 plots in the German SETAC-DRAW subset or 21 plots in the French subset. For each distance (e.g., 3, 5, …, 30 m) and each plot of a field trial, a single value is defined for the percentage of spray drift deposition. For example: for a drift measurement in 3 m distance a deposition of 4.69% of the application rate was determined for plot 1.

To demonstrate a multi-factorial design, the structure of the data subset for a selected crop is important. The data subset should be randomly selected from various crops, plant protection products, trial vendors, time periods, independent replicated trials, etc..

The selection of the subsets from the SETAC DRAW database for this project was unsystematic due to time constraints - with the sole aim of mapping a data structure suitable for the experimental design. From the German subset, the trials DE_5_002, DE_5_004, DE_5_005, DE_5_007, DE_5_008, DE_5_010, DE_5_011, DE_5_012 were selected, from the French subset, the trial FR_1_005, FR_1_006 and FR_1_022.

Many of these trials were pooled together in the considered crop-specific database for different trial providers, locations, times and in particular nozzles, drift reduction classes of the nozzles, boom heights, speed or pressure etc. together with many covariates such as wind speed. Small data subsets were randomly selected from this huge database in such a way that multi-factorial designs could be systematically analyzed in terms of the task at hand.

3 Statistical Approaches

The term 'higher uncertainties' is translated into a statistical property according to the EFSA recommendations (EFSA, 2018). In frequentist statistics, two properties are used: i) maximum power, ii) smaller width of confidence intervals. These approaches are described in detail below.

3.1 Uncertainty quantified by the concept of power

For a selected point-zero null-hypothesis test, the power π is defined as

$$\pi = (1 - f^{-}),$$

with the false negative error rate f-.

The underlying Neyman-Pearson hypotheses are statistical translations of Popper's falsification principle 'we can never claim an effect directly, only to demonstrate the unlikeliness of its opposite'. This system is binary and asymmetric. For a claim of superiority in a two-sample comparison Group 1 against Group 2 based on expected values μ_i, the alternative hypothesis $H_1: \mu_1 - \mu_2 > 0$ can be proven by rejection of the opposite null hypothesis $H_0: \mu_1 - \mu_2 = 0$ alone. Each of both decisions is uncertain, i.e. erroneous for a small error rate:

- erroneous rejection of H_1 with a false-negative error rate f- and
- erroneous rejection of H_0 with a false-positive error rate f+.

A powerful test reveals a small false negative error rate, i.e., this decision reveals a 'lower uncertainty'. Note that the range of power π is specific, varying from low f+ = α, i.e., usually 5% (i.e. the α-level) to up to 100%. This power concept is a statistical approach for quantifying 'uncertainty'.

As the power of the 'distance' factor is naturally very high, the power approach can only be used to a limited extent for other factors in the multi-factorial design. In addition, 'power' as a criterion is more of a statistical issue than a directly biological-agricultural interpretable variable. Therefore, the width of the 95% confidence interval for the model-based predicted distance for 90% reduction was established (see chapter 3.2).

In addition to the frequentist approach used here, there is also the Bayesian approach as used to evaluate a selected SETAC-DRAW data subset (Chapple, 2022). This alternative approach was not used in this report as the necessary priors were not available and the experimental design is still extremely challenging.

Key message: Concept of power

The power of a hypothesis test (π) quantifies statistical uncertainty and depends on the false negative error rate (f−). Low uncertainty is expressed by a π-value close to 100%.

General limitation: As the power of the primary factor 'distance' is high, a further power increase by additional factors is hard to demonstrate.

3.2 Uncertainty quantified by the concept of confidence interval width

A second concept is the minimum width of a 95% confidence interval (CI) which is a continuous and unlimited measure. In this concept the main question is, how power and confidence interval are related. For selected tests, such as the t-test, power and confidence interval width are related quantities, i.e., the test decision (p-value < 0.05) and the exclusion of the value zero (i.e., the

value of the null hypothesis) from the confidence interval are 'compatible'. Here, however, the test (in this case the F-test in the ANOVA) and the confidence interval are not compatible or even the confidence intervals are not available.

The definition of a two-sided confidence interval requires the definition of an appropriate effect size (ES; difference to total mean) and the definition of a variance term (VT; describes the uncertainty):

$$CI = ES \pm VT$$

with

CI = confidence interval

ES = effect size

VT = variance term

Various effect sizes and variance terms exist for several tests. For the two-sided t-test with equal sample sizes it is simply:

$$CI = (\mu_1 - \mu_2) \pm t_{df,1-\alpha/2}\sqrt{2s/n}$$

with $df = 2n - 2$

and

df = the degree of freedom

n = the sample size

s = pooled standard deviation s

α = pre-defined false positive error rate (e.g. α = 0.05)

In this report, more complex confidence intervals are used, but the principle of uncertainty remains the same. Confidence intervals can be estimated not only for mean differences, but also for predicted values of a nonlinear model, such as the parameter ED90 (see below).

Key message: Concept of confidence interval width

The width of a confidence interval is a second measure of uncertainty:

- small width reflects lower uncertainty and
- wider width indicates higher uncertainty.

3.3 Multi-factorial design

3.3.1 Uncertainty quantified by the power approach in factorial designs

The main question is: how to model up to five measures?

From a statistical point of view, these measures, i.e. independent variables, are termed factors and are classified into:

- qualitative factors (such as nozzle types),
- quantitative covariates (such as distance),

- the statistical modelling of both types jointly.

On a second level factors are classified into:

- factors with many levels (in this case e.g., nozzle types),
- factors with with few levels (in this case, e.g., pressure).

A commonly used model is the completely randomized multi-factorial fixed effect analysis of variance (ANOVA). Each measure is considered as a factor with qualitative levels. The measure 'nozzle' represents a qualitative factor, whereas boom height is originally a quantitative covariate with, for example, levels 0.5 m and 1.65 m. A quantitative covariate can be transformed into a qualitative factor with levels such as 'low' and 'high'. Such a transformation may result in a loss of information (Greenland, 1995), or it may simplify the evaluation and interpretation. The joint evaluation of qualitative factors and quantitative covariates is possible (Piepho, 2018). However, estimating power in such a setup is rather challenging and is not considered here.

In factorial designs the power can be directly calculated (Spangl, 2023). However, for the assumption of homoscedastic errors only. Homoscedastic means homogeneous, i.e. not equal but similar, variances per treatment level. Therefore, a simulation study was used allowing heteroscedastic errors and even non-Gaussian distributed variables (although the latter was not used in this report).

Random experiments for various designs and pre-defined parameter sets for location, scale, sample size, number of levels, effect size, etc. are generated.

Here, designs were used for one to five factors with various sample sizes.

3.3.2 The specific nature of the primary factor 'distance'

In multi-factorial designs, there is usually one dominating 'primary' factor. Other 'secondary' factors are much less significant and describe modifications only. An example for a primary factor is the 'distance' in spray drift experiments, whereas 'boom height' would be a secondary factor. The primary factor 'distance' is quite specific:

- quantitative (3, 5, ...,100 m),
- extremely powerful by design and objective, i.e., high values for small distances up to near-to-zero at larger distances,
- heterogeneous variances,
- the definition as percentage change from baseline and values of 0 m cause a conflict, and
- skewed distribution.

As an example, Figure 2 shows the boxplots for the BBA orchard (late)[1] data to illustrate the specific nature of these data. From 30 available trials for the crop "orchard late", a subset of 28 trials with the distances '3 m', '5 m', '7.5 m', '10 m', '15 m', '20 m', '30 m', '50 m' was analyzed. The data subset contains the trials: "Vers1", "Vers101", "Vers102", "Vers103", "Vers104", "Vers16", „Vers163", "Vers165", "Vers166", Vers167", "Vers168", "Vers169", "Vers17", "Vers18", "Vers19", "Vers2", "Vers20", "Vers22", "Vers23", "Vers3", "Vers4", "Vers43", "Vers44", "Vers45",

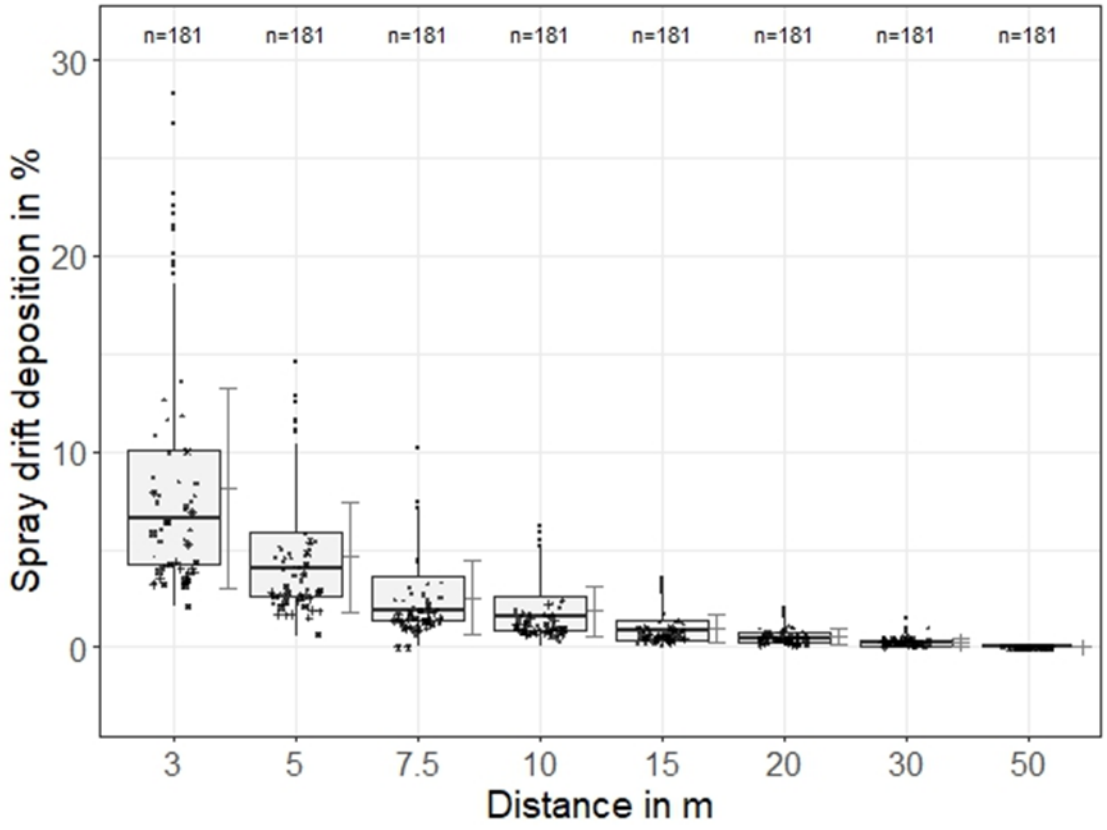
[1] "Orchard late" refers to the late application time of plant protection products; here to the plant growth stages 85 (advanced ripening or fruit coloration) and 91 (shoot development completed; foliage still green).

"Vers46", "Vers47", "Vers48", "Vers5". Those trials that did not contain drift deposition values for all distances were not included in the data subset (i.e. “Vers6” and “Vers164”).

The 28 trials are characterized by different sample sizes (n_i = 4, …, 10) which resulted in a sample size of n = 181 for the whole subset.

Figure 2: Boxplots for selected BBA orchards data (orchards late) to demonstrate specific data conditions.

For details on the selected BBA spray drift data, please refer to the text.



Source: own illustration, L. A. Hothorn.

The specific nature of the data in Figure 2 can be described as follows. The primary feature is the pronounced monotonic non-linear decline with increasing distance. The secondary feature is the decreasing variance with increasing distance and hence decreasing values. The third feature is the skewed distribution.

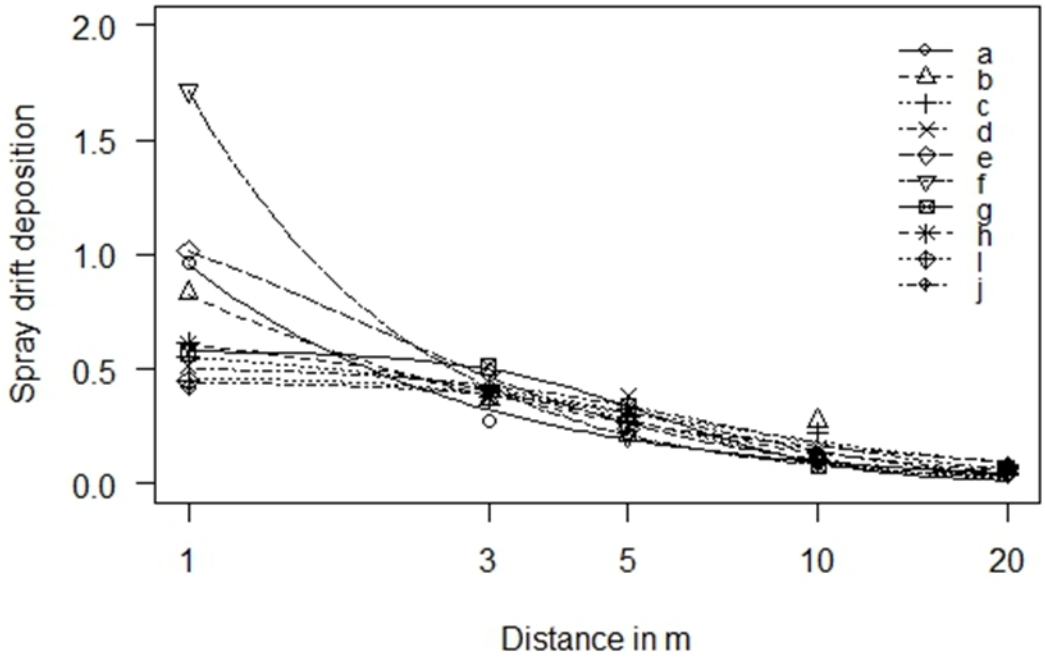
The statistical evaluation of test data from the ’Biologische Bundesanstalt’ (BBA) shows the extremely large effect of the ’distance’ factor, which resulted in post-hoc power values of > 90%, as well as corresponding power estimates in simulation studies. Since the power is limited to 100%, a further increase in power by additional factors is naturally restricted. This specificity of the primary factor ‘distance’ is a strong determining and limiting aspect in the report and clearly determines the selection of statistical methods.

As a consequence of the specific nature of the data described above, a second approach is used to analyze the primary covariate ‘distance’: the consideration as a quantitative covariate and its fit by means of a non-linear model. A method to model a primary covariate jointly with secondary factors and covariates is not available, neither for evaluation nor for power considerations (Ritz, 2015). Therefore, a two-step approach is used:

1. a per-plot fit of a nonlinear model followed by an estimation of a certain characteristic, such as an effective distance for 90% reduction (ED-90) compared to the estimated lower asymptote of the non-linear model (see the plot-specific model fits in Figure 3, see further details in chapter 4), and
2. statistical tests using this estimated confidence intervals between the levels of additional factor(s)

Figure 3: Example for plot-specific model fits (three-parametric log-logistic model) for the German SETAC-DRAW data experiment with the trial number DE_5_011 with its 10 plots (plots a-j).

The spray drift deposition is given in [%].



Source: own illustration, L. A. Hothorn.

3.3.3 Interactions

For all multi-factorial designs, the term 'interaction' is the most interesting issue. It is the dominating effect. The power decreases seriously if an interaction exists between factors.

In the following, interaction is explained for a simple two-factor design. Simple means a primary factor A with three levels [a1, a2, a3] and a secondary factor B with two levels [b1, b2]. Examples for treatment levels are e.g., various nozzle types, multiple distances, boom height etc. Commonly, treatment interaction is tested by an F-test: F(A*B) in addition to the main effect factor terms F(A) and F(B). The trivial cases, namely 'no effect in A', 'no effect in B' are not considered.

The following effects are possible (see Figure 4):

- Additive effect only:
  There is a significant main effect A and a significant main effect B, but a non-significant interaction effect [A*B]. Here A and B are additive, the B levels are either superior or

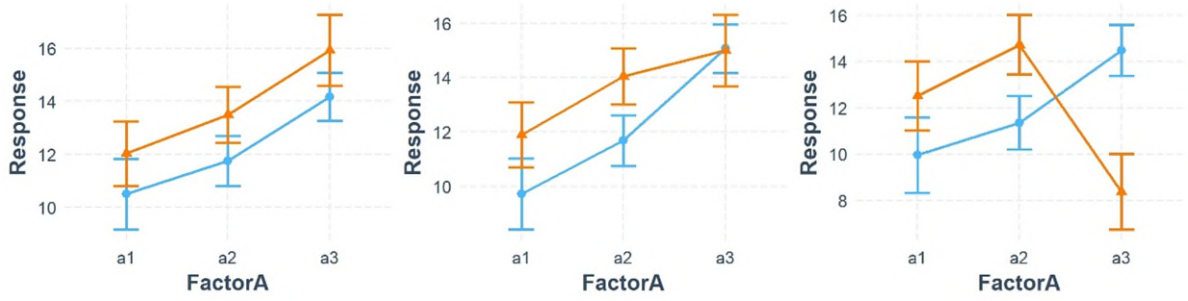
inferior, but additive (or proportional). For example, if we assume a straight line for [a1 → a2 → a3] for level b1, then there is also a straight line for level b2.

- Quantitative interaction:

  The 2nd case is: significant main effect A, significant main effect B and significant interaction effect [A*B]. I.e., at least one (any-one, up to all) combination [A by B] behaves different. For example, the effect a3 in b1 is smaller than in b2 and the other combinations a1-by-b1, a2-by-b1, a1-by-b2, a1-by-b2. Or, the line is linear for level b1 - but not linear for level b2.

- Qualitative interaction:
  In addition to the quantitative interaction, the direction at level a3-by-b1 changes to descending.

Figure 4: Examples of interaction plots with Factor A represented on the x-axis and factor B shown as blue (Factor b1) and orange (Factor b2) curves.



additive effect only **quantitative interaction** **qualitative interaction**

Source: own illustration, L. A. Hothorn.

Two kinds of interaction are possible. They are different in both analysis and interpretation in terms of directional decisions: quantitative vs. qualitative interaction. In the case of quantitative interaction, the effect direction remains similar - despite non-additivity. In the case of qualitative interaction however, the direction is changed.

An example could be: factor A with 20 levels as nozzle types and factor B as the boom height with two levels [low, high]. A qualitative interaction exists if, for example, the nozzles A13, A20 are the best for low boom height, but belong to the worst for high boom height. Such a formal scenario is a disaster for interpretation, but also for the power of multifactorial designs.

Note, that the usual F-test [A*B] is not able to differentiate between qualitative vs. quantitative interaction. This is a much more complex issue and not within the scope of this report. Further information is published (Kitsche, 2014). See also an overview for interactions in factorial designs in agriculture (Kitsche, 2015).

Nevertheless, even in the ANOVA F-test approach the interaction is central: the power is reduced in the absence of interaction and the effect of the secondary factor(s) is only additive. An extreme example: the power can reduce from high (e.g. 90%) to low (5%, α-level) if [a1 to a2] and [a2 to a3] are monotonically increasing at level b1, but monotonically decreasing at level b2. In this counterintuitive case, the effect sizes cancel each other out.

Key message: Interactions

The mere possibility of interactions in multi-factorial designs can significantly increase the uncertainty of the statements.

There are two types of interactions:

- Quantitative interaction: The direction of the main effects does not change although it could change in magnitude.
- Qualitative interaction: In addition to quantitative interaction, the direction of the effect is reversed for a at least one combination. Qualitative interactions in particular can massively increase the uncertainty of the overall statement.

Besides quantitative and qualitative interactions, additive effects are possible.

3.3.4 Impact of interactions in multifactorial designs

Experimenters assume the interactions to be negligible in multifactorial designs – to avoid suffering a power loss due to further factors. The reality is quite different: designs with many factors (2, 3, 4, etc.) can be highly challenging, both from the point of view of analysis, interpretation, and post hoc power.

Adding more factors increases the risk of interactions per se. Multi-factorial designs should therefore be used with caution. This fact is often ignored in the practice of agricultural field trials. From this perspective, the uncertainty can increase the more factors are considered (data-dependent).

On the other hand, if the primary interest lies in the principal effect of numerous factors, a highly incomplete factorial design with only two levels per factor (i.e. a 'yes or no' option) can be utilized. This approach inherently disregards any interaction effects.

Unfortunately, the approach currently typically used for group comparisons is a stepwise one: if the interaction is not significant, then the group comparisons between the levels of the primary factor of interest are made for the data pooled across the levels of the secondary factor (i.e. for all data).

If the interaction is significant, the group comparisons between the interesting levels of the primary factor are made separately per level of the secondary factor. I.e., these comparisons are performed with reduced sample size n, i.e., n/2 for two levels of the secondary factor, n/4 for four levels of the secondary factor etc.. Alone this sample size reduction has a dramatic impact on power, particularly for multifactorial design with many interactions and many levels of the secondary factors. Imagine a counter-intuitive example with a strong primary factor (similar to the factor 'distance' here) and two additional secondary factors. Assume a large total sample size of 72 and three levels per factor. Without interaction you compare e.g., treatment level a1 with level a2 with n=24 each - which promises a low uncertainty of the statement. However, with interactions of the three factors (each with three levels) this comparison is performed with only n=8, each on the separate sub-data sets of the secondary treatment levels. The result is a substantially increased uncertainty!

With an increasing number of factors and an increasing number of levels, the sample size n_i decreases. Thus, uncertainty increases due to a reduced sample size n_i.

Although widely used, this two-step procedure is problematic from a statistical point of view. The reason is the dichotomization of the interaction (i.e., a binary classification of the interaction

into either a significant interaction effect exists or no significant interaction effect), the lack of a control of the familywise error rate (FWER), the difficult interpretation, and the data-dependent reduction in sample sizes by splitting the randomized design in sub-designs.

For designs using only a small number of factors with a small number of levels, such as 2 or 3 factors with 3 or 4 levels, a solution is the transformation of the multi-factorial design into a pseudo-one-factor design of the so-called 'cell means model'. A cell means model for a two-way layout consists of one factor where the factor levels are the levels of interaction. Using the above teaching example with factor A with a1, a2, a3 and factor B with b1, b2, the cell means model uses a transformed factor AB with the levels [a1b1, a2b1, a3b1, a1b2, a2b2, a3b2] This transformation into a pseudo-one-factor design is intuitively used by practitioners. I.e. the levels of the interaction term are compared only, instead of main effects and interaction effects in the common ANOVA-table. The control of the FWER is thus defined, the interaction can be modelled elementary by the group comparisons and, above all, simulated confidence intervals are available through the use of multiple contrast tests in comparison to the overall mean (Pallmann, 2016).

Another disadvantage of the F-tests within the ANOVA is that they do not provide confidence intervals, only p-values.

3.3.5 Error rate control in multifactorial design

A further issue in multifactorial F-tests is the missing control of the FWER. By the argument of orthogonal sum-of-squares decomposition, each F-test is tested at elementary level α. But for k tests, each at level α, the FWER is inflated in principle. Although this approach is standard in statistics and applications, there are publications with a FWER control based on the Bonferroni inequality, i.e., each F-test is tested at level α (Cramer, 2016) or assuming a multivariate t-distribution (Hothorn, 2022). This largely unused approach is important in the power consideration, as the cell mean model contains precisely this FWER control (Hothorn, 2022a). Otherwise, power comparisons are unfair.

Notice, interactions are a central point when considering the power of multifactorial designs. Confidence intervals can be estimated in the cell means model for interactions. This model is not available for the F-test in standard ANOVA.

3.3.6 The impact of design on power

A central point in multifactorial designs is the way in which the sample size is defined. This belongs to a central theorem and is of central importance for the power analysis, as the following applies to all tests: 'a higher sample size results in higher power.'

A statistician usually assumes a given total number of samples N_{total}, e.g., for a field trial a total of N plots (i.e., the replicates in agriculture field trials in the completely randomized design) are available. In experiments or field trials, N_{total} can be determined by e.g., limited financial or natural resources.

In multifactorial designs, a modified N_{total} is used depending on the number of factors. If only one factor is considered, the sample size N_{total} remains N_{total} for this factor. This is referred to as one-factor design. In a two- factor design, N_{total} is divided into N/2 for each factor; in a three- factor design into N/3 for each factor and so on. Generally, in a multifactorial design N_{total} is divided into N/k for each factor k on each factor level. This means that the more factors are considered, the lower the power per factor. This is a central property of this approach, denoted here as N_{total}-design.

The opposite approach starts with the consideration of a first factor A in a one-factor design based on a sample size n (to be precise $n = n_1 + n_2 + \cdots + n_i$ (for q levels in factor A)). Then a second factor B also with a sample size n is added which results in a two-factor design. Equal sample sizes n for the two factors, what is referred to as balanced design, are assumed for the sake of simplicity. For two factors, this results in $N_{total}=2n$ and in general $N_{total}=k*n$ for k factors. This means that the power increases monotonically with further factors considered. Again, this is a central property of this second approach, denoted here as $n_{elementary}$-design.

The $n_{elementary}$ approach is implied in the BBA data and the DRAW data sets, simply by compiling very different field trials from different experimenters, years, crops, districts - for other objectives than power considerations of multifactorial trials. From a statistical point of view, however, this is unfair: regardless of all other influencing variables, the power increases with the inclusion of other factors. Note that adding factors could also lead to independent randomized trials, even if they were conducted together in time and location. These should not be pooled as in the $n_{elementary}$-design, but the independent trials should be analysed as random in a meta-analysis (Griffin, 2021).

Key message: The impact of design on power

The difference between a completely randomized design with a predetermined total sample size and matched subsets with an increasing total number of cases is essential.

The central point of this report is to differentiate between the two design variants $n_{elementary}$-design vs. N_{total}-designs, precisely when considering power.

In the N_{total}-design, the total sample size is constant. As N_{total} is divided by the number of factors, n_i decreases as the number of factors increases. Thus, the power decreases the more factors are considered.

In the $n_{elementary}$ design, the total sample size N_{total} increases with an increasing number of factors. Thus, the power of the $n_{elementary}$-design naturally increases as the number of factors increases. This design is used in meta-analysis over compiled experiments.

4 Modelling

4.1 Modelling the relationship between spray drift deposition and distance

The central relationship in drift data is the dependency between spray drift deposition and the distance (central in the sense of the highest power). The modelling of this quantitative dependency is highly dependent on the data structure: per plot, per trial or per experiment. A uniform model is preferable, which is also robust in case of non-ideal relationships (as long as these do not have an inverse upward direction). Empirical studies on various real data (including model selection methods) showed that a three-parametric log-logistic model is practically suitable:

$$f(distance) = \frac{d}{1 + \exp\left(b(\log(distance) - e\right)}$$

with

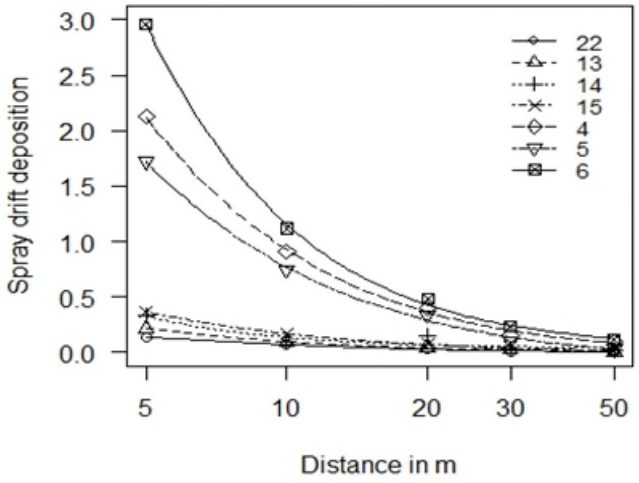
d - upper asymptote,
b - steepness and
e - effective distance.

The less parameter there are to estimate, the more robust such non-linear models are. Therefore, the lower asymptote in a four-parameter log-logistic model was assumed to be zero (i.e., near-to-zero values for larger distances).

For example, the model fit for per-trial French data is reasonable (Figure 5):

Figure 5: Model fit for French SETAC-DRAW data: per trial (separate fits for the trial numbers 4,…,22).

Each point in the diagram represents the predicted mean values of the data from 21 plots per trial. The spray drift deposition is given in [%].



Source: own illustration, L. A. Hothorn.

The model fit is possible per-plot, pooled over the plots within trial or modeling the plots as random factor within a mixed effect model (Ritz, 2013) whereas the first approach was used

here. The model fit itself is not the aim of the analysis but the estimation of derived parameters for further evaluation. The effective distance (for e.g., 90% reduction) or the area under the curve (AUC) were considered. The latter can be used as an integral measure for spray drift reduction.

Key message: Three-parametric log-logistic model

The relationship between spray drift deposition and distance is modelled with a three-parametric log-logistic model.

4.2 Properties of the area under the curve approach

The 'effective distance' approach is certainly simple and can be interpreted in an adapted way. But due to the unfavorable signal-to-noise ratio at the larger distances of interest, it is rather unstable and thus naturally results in wide confidence intervals. Due to its integral character, the AUC approach is much more stable. In simplified terms, it can be understood as a weighted average over all experimental intervals. Mean values in themselves have the effect of minimizing variance. One can simply estimate a three-parametric log-logistic model for each replicate in the experiment, calculate an AUC for each model and calculate a mean AUC and its confidence interval across all replicates. Empirically, different conditions (nozzle, pressure, tests) could thus be distinguished statistically quite well.

5 Summary of Results

5.1 Simulation results based on BBA data

5.1.1 The power approach based on assumptions derived from BBA data

Power was estimated for selected factorial designs and parameter setups for the N_{total}-design. The interpretation is simple: 'the higher the power, the lower the uncertainty of the decision'.

Table 1 summarizes the results for one factor, two factor and three factor designs - reflecting the addition of further measures - with four patterns of the effect of secondary factors: without, additive only, quantitative interaction and qualitative interaction. Table 1 considers one, two and three factors only. The trend remains the same for four and five factors (not shown), i.e. factorial designs with one, two or three risk mitigation measures are considered.

The analysis in Table 1 and Table 2 is based on BBA orchard data related to the experiments conducted with the nozzles Albuz gelb (with n = 7) and ATR gelb (with n = 10). The following factors were considered:

- Factor distance with levels: 3 m, 5 m, 7.5 m, 10 m, 15 m, 20 m, 30 m
- Factor nozzle with the levels: 1) Albuz gelb, 2) ATR gelb
- Factor speed with the levels: low (5.5 km/h) and high (6.0 km/h)

Table 1: Statistical power in selected factorial designs for the N_{total} = const. scenario based on selected BBA spray drift data.

For details on the selected BBA spray drift data, please refer to the text.

Factorial design	Independent factor (risk mitigation measures)	Secondary effect	Power π
One factor design	distance	Without	0.81
Two factor design	distance, nozzle	Without	0.80
Three factor design	distance, nozzle, speed	Without	0.77
One factor design	distance	Additive	0.83
Two factor design	distance, nozzle	Additive	0.79
Three factor design	distance, nozzle, speed	Additive	0.78
One factor design	distance	Quantitative Interaction	0.83
Two factor design	distance, nozzle	Quantitative Interaction	0.81
Three factor design	distance, nozzle, speed	Quantitative Interaction	0.76
One factor design	distance	Qualitative Interaction	0.83
Two factor design	distance, nozzle	Qualitative Interaction	0.26
Three factor design	distance, nozzle, speed	Qualitative Interaction	0.47

As shown in Table 1, there is a slight decrease in power when a second or even a third factor is added, meaning that the uncertainty increases slightly with the addition of further measures.

However, when there is a qualitative interaction, this pattern changes fundamentally: the addition of further factors drastically reduces the power. The crux of the matter is that such interactions cannot be ruled out a priori in factorial designs and that their probability increases with the number of factors added. In other words, the uncertainty can increase significantly depending on the data. It is important to note that the loss in power for qualitative interactions does not necessarily decrease monotonically with the inclusion of more factors, due to the increasing complexity of the model.

In contrast, the power of the $n_{elementary}$-design naturally increases as the number of factors increases (Table 2). This is simply a function of the increasing total sample size.

Table 2: Statistical power in selected factorial designs for the $n_{elementary}$ scenario (with increasing N_{total} and n_i = const.) based on selected BBA spray drift data.

Factorial design	Independent factor (risk mitigation measures)	Additive effect?	Power π
One factor design	distance	No	0.43
Two factor design	distance, nozzle	No	0.79
Three factor design	distance, nozzle, speed	No	0.99
One factor design	distance	Yes	0.45
Two factor design	distance, nozzle	Yes	0.90
Three factor design	distance, nozzle, speed	Yes	0.99
One factor design	distance	Interaction	0.43
Two factor design	distance, nozzle	Interaction	0.90
Three factor design	distance, nozzle, speed	Interaction	0.99

5.1.2 Concept of confidence interval width

From a statistical point of view, the basic non-linear residue-distance relationship is very dominant by design alone (many distances, value range up to near zero) and therefore challenging for power studies alone. Therefore, a derived measure is used: the upper confidence limit of the estimated effective distance (using nonlinear three-parametric log-logistic model) for 90% reduction (referred to as ED90, where 90 is a chosen threshold) (Ritz, 2015).

For this derived variable, the width of CI of the ED90 is used to characterize uncertainty in a simulation study for factorial designs. Here, uncertainty decreases, the smaller the confidence intervals are. I.e., the limitation of the power approach taking only values between 5 and 100% is overcome here, whereas the width of the confidence is variable- and context dependent.

For selected conditions, the CI width is estimated for one factor, two factor and three factor designs for both design types (N_{total}, $n_{elementary}$) for simulated data according to the assumptions in Table 3. Table 3 considers one, two and three factors only. The half width of the confidence interval is presented since the one-sided upper confidence limit is of interest.

An increase in CI width (i.e. an increase in uncertainty) with an increasing number of factors can be seen in Table 3, even for the unfair $n_{elementary}$ design. For the fair N_{total} design, the increase is so

massive that even the significance in the three factor design is lost (in this specific parameter setup). From the perspective of the CI width criterion, the addition of further factors leads to an increase in uncertainty. The decision 'significant' and 'not significant' (e.g. by means of a p-value) is compatible with the interval inclusion of the value zero in the confidence interval. That means, p-value as well as confidence interval result in the same conclusion. In Table 3, the increase of the confidence interval width by adding a second and third factor is demonstrated where the decision is still significant. However, the three-factor design shows (in line 3) such a strong increase in width that there is even a qualitative reversal: no significant effect anymore.

Table 3: Influence of the number of factors on the half width of the confidence intervals for ED90

Type of design	Factorial design	Half width of confidence interval [m]	Decision
N_{total} with k =1	One factor design	17.6	Significant
N_{total} with k = 2	Two factor design	29.9	Significant
N_{total} with k = 3	Three factor design	47.7	Not significant
$n_{elementary}$ with k=1	One factor design	17.6	Significant
$n_{elementary}$ with k=2	Two factor design	20.8	Significant
$n_{elementary}$ with k=3	Three factor design	23.1	Significant

5.2 Results from selected case studies of the SETAC DRAW database

The extraction of data subsets from real data and their analysis will only ever produce locally valid statements. With this restriction in mind, French and German data subsets were selected from the SETAC-DRAW database (FR, DE). As the evaluation of the 'distance' factor yields small (and therefore not really comparable) p-values, the F-value of the ANOVA test statistic was chosen instead. The rule is: the higher the F-value, the lower the uncertainty.

Statistically, under these conditions the F value increases proportionally to the sample size cases, e.g. double the sample size results in double the F value and thus an n-ratio of 2-fold. If a lower n-ratio is observed empirically, this indicates that the addition of further factors increases the uncertainty despite an increased N_{total}.

5.2.1 Analysis of the French subset

The French data were selected because of a high number of replicates (i.e. 21 plots) and a relatively high variability. Scenarios were selected for one factor, two factor and three factor designs to characterize the impact of design and sample size on the F-test value for the residue-distance relationship.

Out of 22 trials in the French dataset, trial no. 4, 5, 6, 13, 14, 15 and 22 were selected for the analysis. Spray drift deposition values for the distances 5 m, 10 m, 20 m, 30 m, 50 m were considered.

Using the example of trial no. 5, trial no. 6 and trial no. 22, the general results are explained. Some characteristics of these trials are the following:

- Trial 5: N_{total} = 105, nozzle AXI 110 02, no drift reduction

Trial 6: N_{total} = 104, nozzle AXI 110 02, no drift reduction

- Trial 22: N_{total} = 104, nozzle AVI 110 02, 75% drift reduction

Table 4 contains several conclusions for the French subset:

3. For the one-factor design with trial no. 6, the F-value increases monotonically with increasing sample size and decreases monotonically with decreasing sample size. I.e., the total sample size is most important for un-certainty.
4. The F-values for the one-factor design of trials no. 5, 6 or 22 vary a bit – depending on the specific real data, i.e., different trials naturally have different residue-distance dependencies. This speaks for natural variance differences of test repetitions.
5. The combination of trials generating a two-factor design doubles the total sample size, but does not necessarily double the F-value compared to the one-factor design. In particular, the three factor design with N = 313 does not reveal the highest F-value for these real data. I.e. depending on the specific real data conditions, the uncertainty does not decrease as would actually be expected from the increased total sample size.

Table 4: Influence of the total sample size and the number of factors on the F-value (based on data from the SETAC-DRAW database-French subset)

Design	Trials used in analysis	Total sample size	F-value	F-ratio: Empirical ratio**	n-ratio: Expected ratio due to sample size***
One factor design	5 alone	105	66.7	-	-
One factor design	6 alone	104	63.1	-	-
One factor design	6 pseudo half*	50	29.0	0.5	0.5-fold
One factor design	6 pseudo double*	208	127.5	2.0	2-fold
One-factor design	22 alone	104	66.8	-	
Two factor design	5 & 6	209	118.8	1.8	2-fold
Two factor design	5 & 22	209	90.0	1.3	2-fold
Two factor design	6 & 22	208	77.6	1.2	2-fold
Three factor design	5 & 6 & 22	313	108.3	1.7	3-fold

* "pseudo half" and "pseudo double" indicate that the sample size available from trial 6 has been artificially halved or doubled by simulation;

** Empirical ratio = [Actual F-value] / [F-value of means value of included one- factorial designs];

*** Expected ratio = [Actual sample size] / [Sample size of one-factorial design]. In the FR subset of the SETAC DRAW data base the independent three factors 'distance', 'reduction' and 'trial replication' were available

5.2.2 Analysis of the German subset

The German data were selected because of a common number of replicates (i.e. 6 to 10 plots) and a relatively small variability. Scenarios were selected for one factor, two factor, three factor

and four factor designs to characterize the impact of design and sample size on the F-test value for the residue-distance relationship.

The German dataset consists of the subsets DE1 to DE12, each comprising a certain number of trials. Out of 51 trials in DE5, trials DE_5_002, DE_5_004, DE_5_005, DE_5_007, DE_5_008, DE_5_010, DE_5_011, DE_5_012 were selected for the analysis and the spray drift deposition values for the distances 1 m, 3 m, 5 m, 10 m, 20 m were considered.

Some characteristics of these trials are the following:

- Trial DE_5_002: N_{total} = 50, nozzle AVI 04, 75% drift reduction, pressure 200 kPa
- Trial DE_5_004: N_{total} = 50, nozzle AVI 04, 50% drift reduction, pressure 500 kPa
- Trial DE_5_005: N_{total} = 50, nozzle AVI 04, 50% drift reduction, pressure 500 kPa
- Trial DE_5_007: N_{total} = 50, nozzle IDKT 120 04, 50% drift reduction, pressure 200 kPa
- Trial DE_5_008: N_{total} = 50, nozzle IDKT 120 04, 50% drift reduction, pressure 200 kPa
- Trial DE_5_0010: N_{total} = 50, nozzle IDKT 120 04, 50% drift reduction, pressure 500 kPa
- Trial DE_5_0011: N_{total} = 50, nozzle IDKT 120 04, 50% drift reduction, pressure 500 kPa
- Trial DE_5_0012: N_{total} = 50, nozzle IDKT 120 04, 50% drift reduction, pressure 500 kPa

Table 5 contains a single conclusion for the German subset DE5. The main issue in considering one-factor, two-factor, three-factor and four-factor designs are the mas-sive increases in total sample sizes: 50 → 150 → 300 → 1500. The test theory states that F values are proportional to the sample size: for N = 50 → 150 → 300 → 1500 follows F-ratio = → 3-fold → 2-fold → 2fold → 5-fold. But empirically, these F-value ratios are lower in these real data subsets: compare column 5 and 6 in Table 5. Again, depending on the specific real data conditions, the uncertainty does not decrease as much as it would be expected due to the increased total sample size.

Table 5: Expected and actual effect of the sample size on the F-values.

Data basis is the SETAC-DRAW database - German DE5 subset.

Design	Total sample size n	F-value	F-ratio: Empirical ratio*	n-ratio: Expected ratio due to sample size**
One factor design	50	18.4	-	-
Two factor design	150	84.6	4.6	3-fold
Two factor design	150	48.8	2.7	3-fold
Three factor design	300	85.0	4.6	6-fold
Four factor design	1500	382.9	20.9	30-fold

* Empirical ratio = [Actual F-value] / [F-value of One factorial design]

** Expected ratio = [Actual sample size] / [Sample size of One factorial design]

5.3 Summary of the results

The results presented in chapter 5.1 and 5.2 can be summarized as follows.

1. Effect of sample size on uncertainty: A higher sample size usually reduces the uncertainty. This was shown by the power approach based on the $n_{elementary}$-design where the power increases with an increasing number of factors. Each factor contributes to an increasing sample size (chapter 5.1.1). Furthermore, a reduced uncertainty was demonstrated by an increasing F-value of the ANOVA test as a result of an increasing sample size (chapter 5.2).
2. Effect of the number of factors on uncertainty: The width of the confidence intervals, which indicates the degree of uncertainty, tends to increase with an increasing number of factors (chapter 5.1.2) Also in the ANOVA test based on the French SETAC DRAW dataset, most setups show a significantly lower F-ratio the more factors are considered even though the sample size increased. The reduced F-ratio and thus the increase in uncertainty is a result of an increasing number of factors (chapter 5.2.1).
3. Effect of the nature of the interaction between the factors: Qualitative interactions of the factors have the highest impact on uncertainty and reduced the statistical power (chapter 5.1.1).

6 Conclusion

Depending on the data conditions, the uncertainty in multifactorial designs can be reduced or increased by adding further factors - however, most statistical arguments favor an increase.

The first argument is that a dominant primary factor can superimpose most of the other factors. The results showed that the 'residue-to-distance' dependency, statistically modeled as the factor ‘distance’, has a comparatively high power.

As the power is only possible up to 100%, there is little scope for further power increase by further measures.

The second argument is the complex influence of the sample size. In general, higher sample size results in higher power of a test. Furthermore, the elementary sample size is reduced proportionally as more and more factors (with the respective levels) have to divide up the total sample size N_{total} in a field trial. On the other hand, the total sample size increases if one simply combines several experiments with different measures each is implicitly the case in the SETAC-DRAW database.

The third argument is the uncertainty-increasing influence of interactions, especially qualitative interactions. The more factors are considered in a multifactorial design, the higher the probability that such interactions will occur. This was demonstrated using a DE-data set from SETAC-DRAW.

It has been shown, not only by means of simulation models, that the uncertainty is usually increased by adding further factors (usually, but not in principle, i.e., examples with decreasing uncertainty can be shown). A similar behavior was observed in both the BBA data and selected SETAC-DRAW DE-data subsets.

The increase in uncertainty has been further demonstrated by adding further factors using the new 'width of the confidence interval' approach.

7 List of references

Benford, D., Halldorsson, T., Jeger, M. J., Knutsen, H. K., More, S., Naegeli, H., Noteborn, H., Ockleford, C., Ricci, A., Rychen, G., Schlatter, J. R., Silano, V., Solecki, R., Turck, D., Younes, M., Craig, P., Hart, A., von Goetz, N., Koutsoumanis, K., … Hardy, A. (2018). Scientific opinion on the principles and methods behind EFSA's guidance on uncertainty analysis in scientific assessment. EFSA Journal, 16(1), 5122. https://doi.org/10.2903/j.efsa.2018.5122

Chapple, A. C. (2022, 29. August). Towards a regulatory spray drift model for risk assessment purposes [Vortragsfolien]. EWM Workshop, York. Abgerufen von https://esdac.jrc.ec.europa.eu/public_path/shared_folder/PesticidesModelling/EMW-10/%2308%20Chapple%20regulatory%20spray%20drift%20model%20for%20risk%20assessment.pdf

Cramer, A. O. J., van Ravenzwaaij, D., Matzke, D., Steingroever, H., Wetzels, R., Grasman, R. P., Waldorp, L. J., & Wagenmakers, E. J. (2016). Hidden multiplicity in exploratory multiway ANOVA: Prevalence and remedies. Psychonomic Bulletin & Review, 23(2), 640–647. https://doi.org/10.3758/s13423-015-0913-5

Greenland, S. (1995). Avoiding power loss associated with categorization and ordinal scores in dose–response and trend analysis. Epidemiology, 6(4), 450–454. https://doi.org/10.1097/00001648-199507000-00025

Griffin, J. W. (2021). Calculating statistical power for meta-analysis using metapower. Quantitative Methods in Psychology, 17(1), 24–39. https://doi.org/10.20982/tqmp.17.1.p024

Hothorn, L. A. (2022). Hidden multiplicity in the analysis of variance (ANOVA): Multiple contrast tests as an alternative. bioRxiv. https://doi.org/10.1101/2022.01.15.476452

Hothorn, L. A. (2022a). Simultaneous confidence intervals for the interpretation of primary and secondary effects in factorial designs without a pre-test on interaction. arXiv preprint arXiv:2204.08336. https://doi.org/10.48550/arXiv.2204.08336

Kitsche, A., & Hothorn, L. (2014). Testing for qualitative interaction using ratios of treatment differences. Statistics in Medicine, 33(9), 1477–1489. https://doi.org/10.1002/sim.6048

Kitsche, A., & Schaarschmidt, F. (2015). Analysis of statistical interactions in factorial experiments. Journal of Agronomy and Crop Science, 201(1), 69–79. https://doi.org/10.1111/jac.12076

Mel's Vineyard (2023): Advancing Options for Management and Mitigation of Spray Drift. https://www.spraydriftmitigation.info/setac-draw-workshop

Miller, P.C.H. (2024): The SETAC DRAW Guidance Protocol for Arable Crop Deposition Spray Drift Trials. In: ASPECTS148: International Advances in Pesticide Application, 2024, 162–172, International Advances in Pesticide Application

Pallmann, P., & Hothorn, L. A. (2016). Analysis of means: A generalized approach using R. Journal of Applied Statistics, 43(8), 1541–1560. https://doi.org/10.1080/02664763.2015.1117584

Piepho, H. P., & Edmondson, R. N. (2018). A tutorial on the statistical analysis of factorial experiments with qualitative and quantitative treatment factor levels. Journal of Agronomy and Crop Science, 204(5), 429–455. https://doi.org/10.1111/jac.12267

Rautmann, D., Streloke, M., & Winkler, R. (2001). New basic drift values in the authorization procedure for plant protection products. Mitteilungen aus der Biologischen Bundesanstalt für Land- und Forstwirtschaft, (383), 133–141. Biologische Bundesanstalt für Land- und Forstwirtschaft, Berlin

Ritz, C., Baty, F., Streibig, J. C., & Gerhard, D. (2015). Dose-response analysis using R. PLOS ONE, 10(12), e0146021. https://doi.org/10.1371/journal.pone.0146021

Ritz, C., Gerhard, D., & Hothorn, L. A. (2013). A unified framework for benchmark dose estimation applied to mixed models and model averaging. Statistics in Biopharmaceutical Research, 5(1), 79–90. https://doi.org/10.1080/19466315.2012.757559

Spangl, B., Kaiblinger, N., Ruckdeschel, P., & Rasch, D. (2023). Minimal sample size in balanced ANOVA models of crossed, nested, and mixed classifications. Communications in Statistics – Theory and Methods, 52(6), 1728–1743. https://doi.org/10.1080/03610926.2021.1938126