Definitive Screening Design eine Lösung für (fast) alle Fragestellungen?

Bernd Heinen
SAS Institute GmbH
In der Neckarhelle 162
69118 Heidelberg
Bernd.Heinen@jmp.com

Zusammenfassung

Definitive Screening Designs (DSD) haben seit ihrer Einführung eine gewisse praktikable Reife erlangt. Zudem wurden angepasste Auswertungsstrategien entwickelt, die erlauben, über die Haupteffekte hinaus auch Effekte zweiter Ordnung zu schätzen. Können diese Designs eine Art Standard für die Versuchsplanung werden? DSDs sind die Designs mit den geringsten Anforderungen an den Versuchsumfang, es gibt allerdings Einschränkungen bei den zugelassenen Faktoren und auch den möglichen Auswertungsmodellen. Einschränkungen und Vorteile werden in diesem Vortrag gegeneinander abgewogen.

Schlüsselwörter: Definitive Screening Design, DSD, Dummy Faktoren, Zweistufige Analyse

1 Grundlagen

Screening Experimente werden durchgeführt, um festzustellen, welche aus einer Kandidatenliste von möglichen Faktoren überhaupt einen Einfluss auf die Zielgröße(n) haben, ohne die Art dieses Zusammenhangs feststellen zu wollen. Aber ein Faktor soll auch dann auffallen, wenn sein Effekt nicht linear ist oder er an einer Wechselwirkung beteiligt ist. Die meisten Versuchspläne beinhalten daher zusätzliche Experimente am Mittelpunkt des Versuchsraums. Definitive Screening Designs (DSD) [1] ist eine Klasse von Designs, die den kleinstmöglichen Versuchsumfang liefern, gleichzeitig aber einige Vorteile gegenüber klassischen Versuchsplänen bieten [2]. Seit ihrer Einführung 2011 haben praktische Erfahrungen und weitere Studien zu neuen Erkenntnissen geführt, die eine Entscheidung für die Anwendung dieser Versuchspläne präziser fassen und auch tiefere Einblicke in die Analyse der Versuchsdaten geben. Gleichzeitig zeichnen sich auch die Grenzen der aus diesen Versuchsplänen ableitbaren Erkenntnisse deutlicher ab. Eine Einschränkung liegt in der Konstruktion der Versuchspläne, sie sind nur für stetige oder binäre Faktoren geeignet. Bei der geringen Versuchszahl der DSDs sind gut interpretierbare Ergebnisse nur für stetige Zielgrößen zu erwarten. Daher rührt eine weitere Einschränkung, müssen Entscheidungen bezüglich kategorialer Zielgrößen getroffen werden, ist ein optimales Design die bessere Lösung. Im Folgenden bezeichnet m die Zahl der Faktoren und n die Zahl der Versuche. Für gerade m hat ein DSD 2m+1 Versuche; für ungerade m empfiehlt sich, einen Dummy Faktor hinzuzunehmen und dessen

Ausprägungen für die Versuchsdurchführung zu ignorieren. Damit kommt man bei ungeraden m auf 2(m+1) +1 Versuche. DSDs sind durch Faltung erzeugte Pläne, wobei jeder Faktor zwei Versuche an seinem Mittelpunkt hat, die mit komplementären Eckpunkten der übrigen Faktoren kombiniert sind. Zusätzlich gibt es einen Mittelpunktsversuch. Das Schema ist in Tabelle 1 für m=6 dargestellt.

Tabelle 1: Definitive Screening Design für sechs Faktoren

Run	X1	X2	X3	X4	X5	X6
1	0	1	1	1	1	1
2	0	-1	-1	-1	-1	-1
3	1	0	-1	1	1	-1
4	-1	0	1	-1	-1	1
5	1	-1	0	-1	1	1
6	-1	1	0	1	-1	-1
7	1	1	-1	0	-1	1
8	-1	-1	1	0	1	-1
9	1	1	1	-1	0	-1
10	-1	-1	-1	1	0	1
11	1	-1	1	1	-1	0
12	-1	1	-1	-1	1	0
13	0	0	0	0	0	0

Die Auswertung geplanter Versuche erfolgt üblicherweise durch ein lineares Modell. Eigentlich will man in einem Screeningprozess ja lediglich Aussagen über die Wirksamkeit einzelner Faktoren treffen, aber natürlich kann ein Faktor ja auch über eine Wechselwirkung oder einen quadratischen Effekt wirken. Für die Anpassung eines kompletten Oberflächenmodells mit einfachen Interaktionen und quadratischen Termen reicht aber die Versuchszahl nicht aus. Also braucht man ein Anpassungsverfahren mit Variablenselektion. Power Studien sind zu einer Vielzahl von Designs und Variablenselektionsverfahren ausgeführt worden, die Resultate führten zu Empfehlungen, dass die Zahl der Experimente das Doppelte bis Dreifache der zu schätzenden Terme betragen sollte. DSDs liefern nun sehr spärliche Versuchsdaten, aber da die Haupteffekte orthogonal sind und weniger als die Hälfte der Experimente ausmachen, kann man erwarten, dass alle aktiven Haupteffekte auch erkannt werden, solange keine Wechselwirkung aktiv ist. Nicht erkannte Wechselwirkungen erhöhen aber den Schätzer der Fehlervarianz, was dazu führen kann, das aktive Haupteffekte nicht erkannt werden. Simulationsstudien [3] haben ergeben, dass aus DSDs mit den üblichen Selektionsverfahren maximal noch eine Wechselwirkung geschätzt werden kann, vorausgesetzt ihr Effekt ist größer als das doppelte der Standardabweichung für zweifach Interaktionen und dreifach größer als die Standardabweichung reiner quadratischer Effekte. Wird die Zahl aktiver Effekte zweiter Ordnung größer, kann man diese Effekte mit einem DSD nicht mehr zuverlässig schätzen, gleich, welche Selektionsmethode man wählt. Was kann man also maximal aus einem DSD ableiten?

Daten von durch Faltung erzeugten Versuchsplänen lassen sich in zwei orthogonale Versuchsräume aufteilen [4], die von Miller und Sitter [5] als „odd space“ für Haupteffekte, dreifach Interaktionen und weitere Effekt mit ungeraden Exponenten und „even space“ für den konstanten Term, zweifach Interaktionen und weitere Effekte mit geraden Exponenten bezeichnet werden. DSDs haben nun den speziellen Vorteil gegenüber allgemeinen Faltungsdesigns, dass die Haupteffekte orthogonal sind. Außerdem lassen sie sich leicht konsistent erweitern, indem man, wie schon zuvor bei ungerader Zahl der Faktoren erwähnt, Dummy Faktoren bei der Generierung hinzufügt, die den Versuchsumfang konsistent erweitern (um den Preis zusätzlicher Kosten).

Tabelle 2: Definitive Screening Design für sechs Faktoren und zwei Dummy Faktoren

Run	X1	X2	X3	X4	X5	X6	D1	D2
1	0	1	1	1	1	1	1	1
2	0	-1	-1	-1	-1	-1	-1	-1
3	1	0	1	1	-1	1	-1	-1
4	-1	0	-1	-1	1	-1	1	1
5	1	-1	0	1	1	-1	1	-1
6	-1	1	0	-1	-1	1	-1	1
7	1	-1	-1	0	1	1	-1	1
8	-1	1	1	0	-1	-1	1	-1
9	1	1	-1	-1	0	1	1	-1
10	-1	-1	1	1	0	-1	-1	1
11	1	-1	1	-1	-1	0	1	1
12	-1	1	-1	1	1	0	-1	-1
13	1	1	-1	1	-1	-1	0	1
14	-1	-1	1	-1	1	1	0	-1
15	1	1	1	-1	1	-1	-1	0
16	-1	-1	-1	1	-1	1	1	0
17	0	0	0	0	0	0	0	0

Wieder am Beispiel von 6 Faktoren gezeigt, werden dann 17 statt 13 Versuche erzeugt, wobei bei der Versuchsdurchführung die Dummy Faktoren ignoriert werden. Wenn man die gesamte Design Matrix nun in die beiden Teile für die echten und die Dummy Faktoren zerlegt, so wird der Teil, der für die Dummy Faktoren steht, ja nicht zur Schätzung von Parametern benötigt. Er kann also zur Schätzung des Versuchsfehlers herangezogen werden. Zusammengefasst kann man davon ausgehen, dass man die relevanten Hauptfaktoren, einige Wechselwirkungen und einige quadratische Terme des zugrundeliegenden Modells schätzen kann, solange die Gesamtzahl der Terme die Versuchszahl nicht übersteigt. Immer unter der Berücksichtigung, dass man ja Screeningversuche durchführt, also an einer zuverlässigen Einschätzung der Haupteffekte interessiert ist, zeigen Simulationsstudien [3], wie stark eventuell vorhandene Wechselwirkungen die Möglichkeiten begrenzen, solche Effekte zuverlässig zu schätzen.

In einem minimalen DSD mit sechs Faktoren und sechs aktiven Wechselwirkungen lassen sich 54.264 Modelle bilden, von denen sich 24 Prozent nicht schätzen lassen. Fügt man allerdings zwei Dummy Faktoren hinzu, was die Versuchszahl um 4 erhöht, dann sind in der gleichen Situation nur 0,38 Prozent der Modelle nicht schätzbar. Mit m Faktoren und m_d Dummy Faktoren gilt als Faustregel: bis zu einem Anteil von $m_m = \frac{m+ m_d}{2}$ wirksamer Wechselwirkungen lassen sich die daraus bildbaren möglichen Modelle zuverlässig schätzen, darüber hinaus wird auch das Schätzen der Haupteffekte unzuverlässig.

Nachdem nun die Grenzen der schätzbaren Modelle ausgelotet sind, stellt sich die Frage, wie kann man die zugrundeliegenden Modelle am besten schätzen? Die Unabhängigkeit der Haupteffekte von anderen Effekten und die Idee der Aufspaltung des Modellraums in orthogonale Teilräume legen ein schrittweises Vorgehen nahe [4]:

1. Bestimmen der signifikanten Haupteffekte
2. Berechnen der Residuen zu diesem reduzierten Modell
3. Bestimmen aller Terme 2. Ordnung, die aus den identifizierten Haupteffekten gebildet werden können. Anpassen dieser Terme an die Residuen und bestimmen der signifikanten Effekte.

Tabelle 3: Anteil nicht schätzbarer Modelle

			95% Limits	
m_f	n_{so}	n_{models}	LB p_{sing}	UB p_{sing}
0	3	1330	0.3759	0.3759
0	4	5985	1.5038	1.5038
0	5	20,349	4.6734	4.6734
0	6	54,264	24.0141	24.0141
2	4	5985	0.0000	0.0000
2	5	20,349	0.0147	0.0147
2	6	54,264	0.3778	0.3778

In Tabelle 3 bezeichnen:

m_f die Zahl der hinzugefügten Dummy Faktoren
n_{so} die Zahl signifikanter Terme zweiter Ordnung
n_{models} die Zahl der resultierenden Modelle
p_{sing} den prozentualen Anteil nicht schätzbarer Modelle (hier exakt bestimmbar, daher sind die Grenzen der Konfidenzintervalle jeweils identisch)

Dieses Verfahren setzt die Annahme voraus, dass die aktiven Faktoren des Modells einer strengen Hierarchie folgen, d.h. Terme zweiter Ordnung sind nur für signifikante Haupteffekte zugelassen. Diese Annahme kann man dahingehend erweitern, dass man alle Terme zweiter Ordnung zulässt, aber nur Modelle bis maximal m_m Termen zweiter Ordnung bildet. Für die Konstruktion des Modells werden zunächst alle Haupteffekte

geschätzt und nur die signifikanten (t-Test, $\alpha = 0{,}05$) behalten. Die Fehlervarianz wird unter Einbeziehung der übrigen Faktoren neu geschätzt, was die Zahl der Freiheitsgrade entsprechend erhöht. Die Residuen des zuvor gefundenen Haupteffekmodells bilden den Datenvektor (gerader Modellraum) zur Schätzung der Terme zweiter Ordnung. Ist die Quadratsumme dieses Vektors signifikant größer als der Versuchsfehler (F-Test, $\alpha = 0{,}20$), wird angenommen, dass es wirksame Effekte zweier Ordnung gibt. Jeder Term wird einzeln geschätzt und die Anpassung mit der kleinsten Fehlerquadratsumme ausgewählt. Ist der F-Test signifikant ($\alpha = 0{,}20$) wird angenommen, dass es noch mehr Effekte gibt und alle Modelle mit zwei Termen werden gebildet. Wieder wird das mit der geringsten Fehlerquadratsumme ausgewählt und der F-Test berechnet. Dieses Verfahren wird fortgesetzt bis entweder der F-Test nicht mehr signifikant ist oder die Maximalzahl von m_m Effekten zweiter Ordnung ins Modell aufgenommen wurde.

Simulationsstudien dieses Vorgehens zeigen, dass die Aufnahme zweier extra Faktoren in das Design die Power zur für alle Terme wesentlich erhöht. Wenn man Vererbung annimmt und diese auch zutrifft, können Terme zweiter Ordnung bis zu einer Zahl von m_m sicher identifiziert werden. Selbst wenn mehr als m_m Terme zweiter Ordnung aktiv sind, lassen sich alle aktiven Haupteffekte auch zuverlässig finden. Für drei oder weniger aktive Haupteffekte lässt sich ein vollständiges Oberflächenmodell schätzen. Allerdings wächst die Laufzeit für die Analyse exponentiell mit der Zahl der Faktoren in einem Versuch. Realistische Obergrenzen für diese Auswertung liegen bei 10 oder 12 Faktoren, je nach Rechnerkapazität. Zwar sind auch darüber hinaus DSDs die effizientesten Designs, aber sie erfordern dann andere Auswertungsstrategieen.

Für das in Schritt 3 vorgeschlagene Selektionsverfahren gibt es ja auch hinreichend bekannte Alternativen wie schrittweise Regression oder Lasso. Die Power für dieses neue Verfahren wurde in Simulationsstudien mit einigen Alternativen verglichen. Für die Simulationen wurden unterschiedliche Signal to Noise Ratios von zwei oder drei angenommen. Die Power des hier vorgeschlagenen Verfahrens, basierend auf 1.000 Simulationen ist herkömmlichen Verfahren in allen Fällen deutlich überlegen. Ein Teilbereich der Simulationen kombiniert 6 Faktoren mit unterschiedlichen Mengen wirksamer Terme und vergleicht die Power des hier vorgestellten Selektionsverfahrens mit der eines Elastic Net.

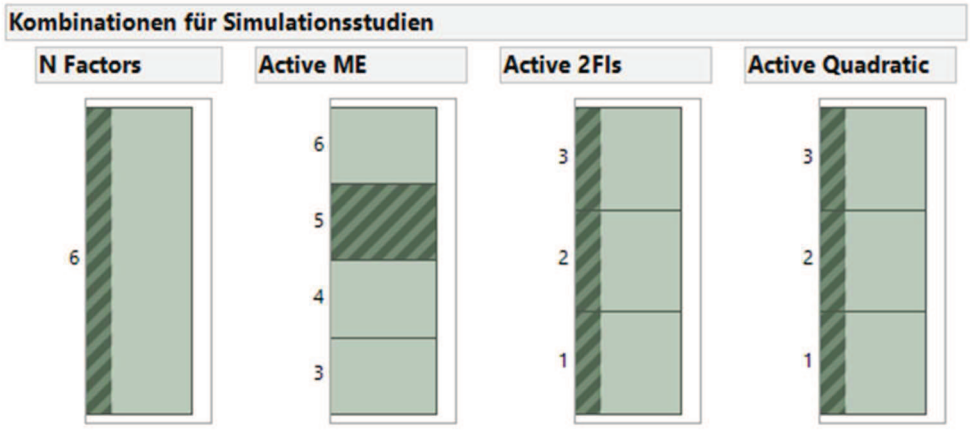

Abbildung 1: Versuchsplan der Simulationsstudie

Die Simulation erfüllt einen vollfaktoriellen Versuchsplan, wobei ein DSD mit 2 Dummy Faktoren, also 17 Versuchen, zugrunde gelegt wurde.

Die besten Ergebnisse, die bei einer Analyse mittels Elastic Net erzielt wurden, bei jeweils nur einem aktiven Term der Haupteffekte, Wechselwirkungen und quadratischen Terme, wird mittels der neuen Analyse auch noch bei mittleren Zahlen aktiver Terme erreicht.

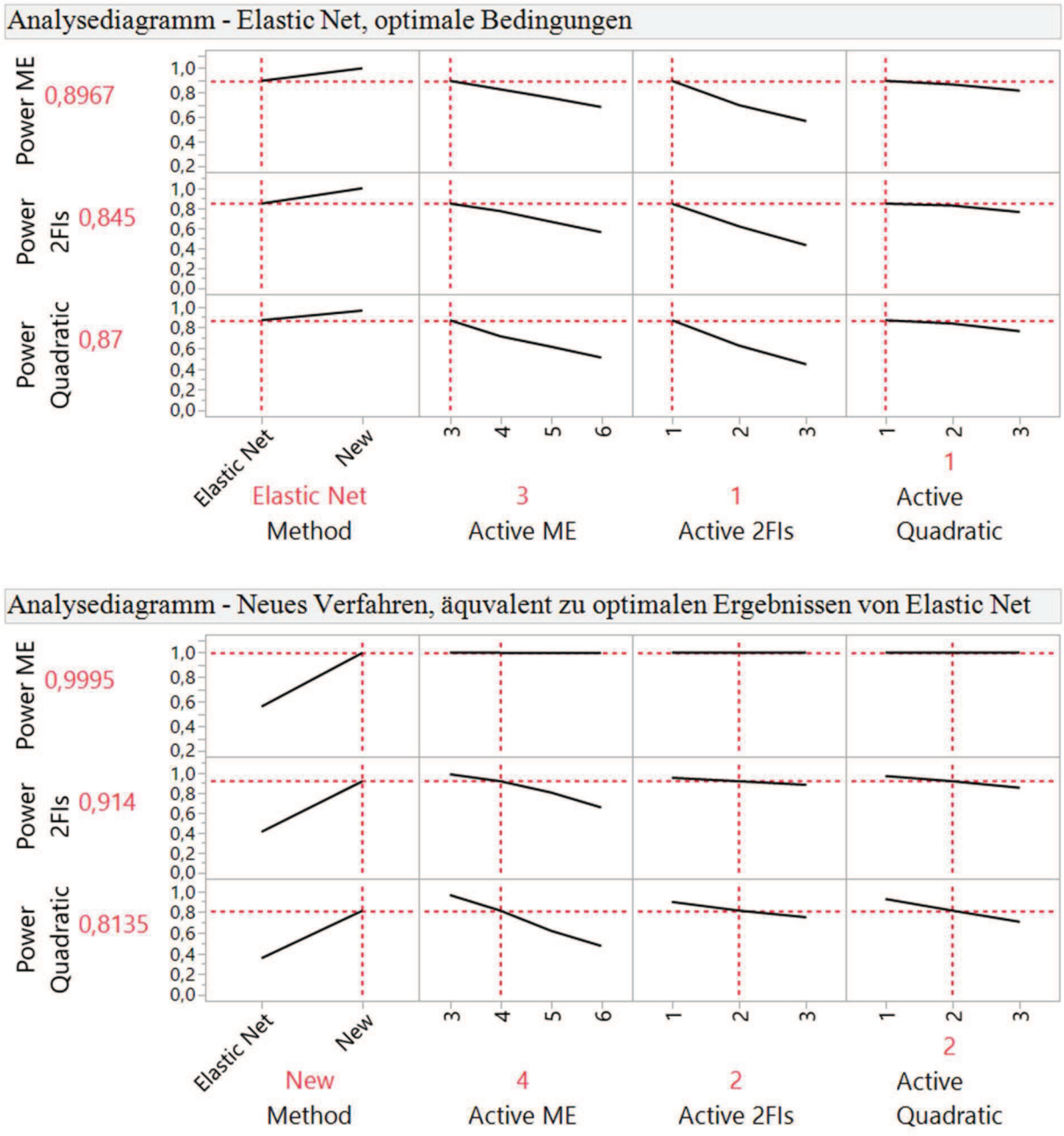

Abbildung 2: Powervergleich zweier Selektionsverfahren

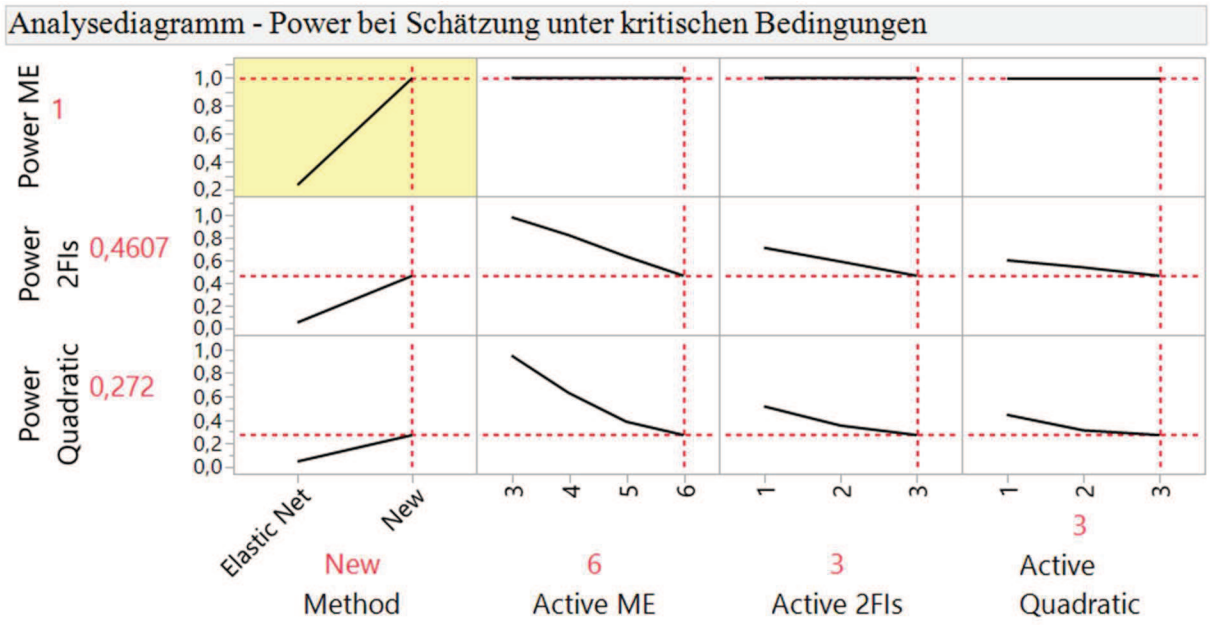

Abbildung 3: Power bei Auswertungen unter kritischen Bedingungen

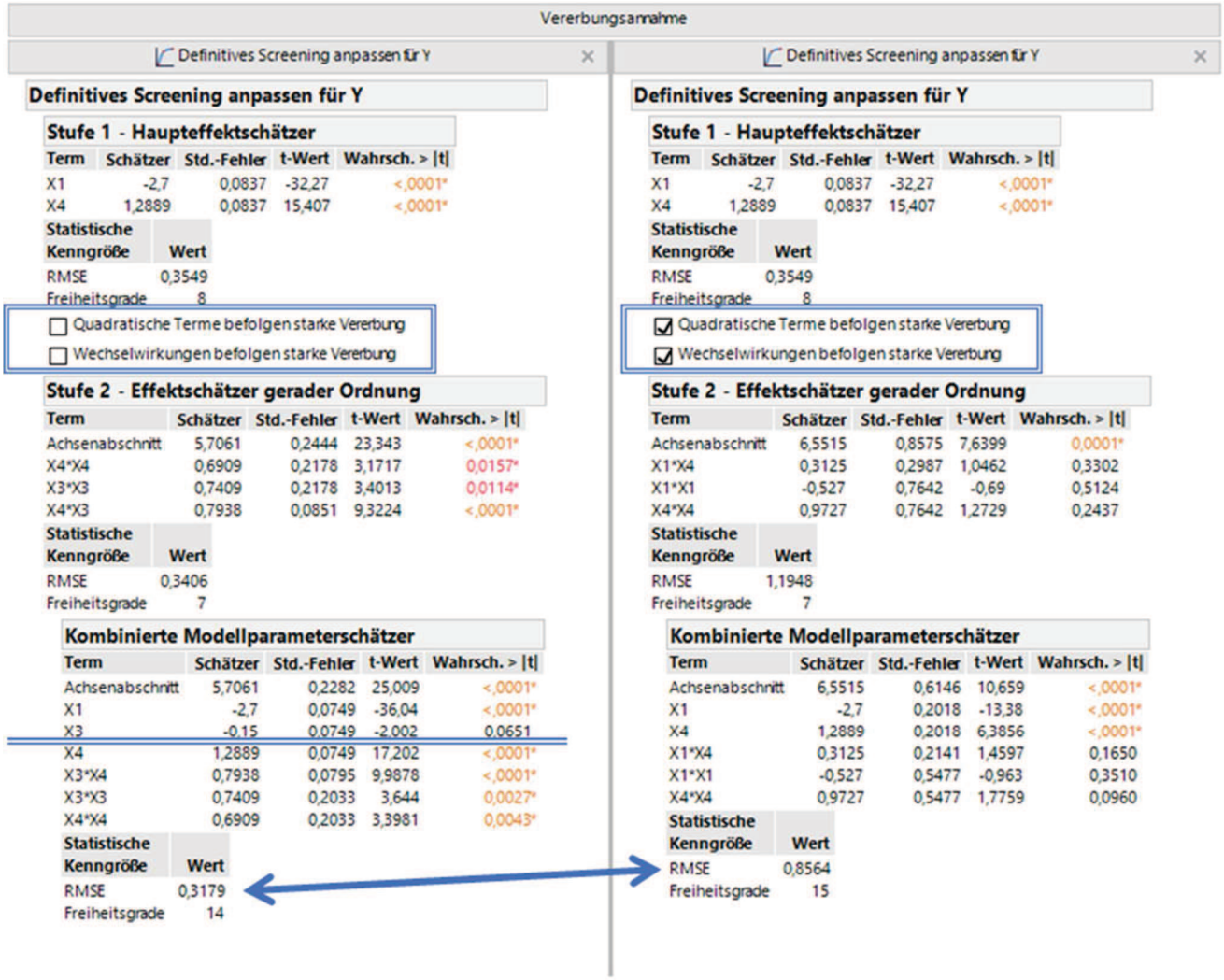

Abbildung 4: Vergleich hierarchischer und nicht hierarchischer Modelle

Im ungünstigsten Fall vieler aktiver Terme reduziert sich die Power für Terme zweiter Ordnung dramatisch, allerdings bleibt sie für Haupteffekte bei eins. Das Elastic Net weist unter denselben Bedingungen für Haupteffekte nur eine Power von 0,2 auf (Abb 3, hinterlegtes Feld oben links) und die Power für alle anderen Terme sinkt gegen Null.

In der praktischen Anwendung lassen sich auf Grund der Orthogonalität der Haupteffekte mit den Effekten 2. Ordnung Modelle mit oder ohne die Vererbungsannahme leicht vergleichen. In diesem Fall zeigt sich im Gesamtmodell, dass der Faktor X3 nur knapp an dem Testkriterium für die Haupteffekte gescheitert ist. Außerdem ist er in je einer signifikanten Wechselwirkung und einem quadratischen Term enthalten. Für das Gesamtmodell ohne Vererbungsannahme (Abb. 4, links) wird ein deutlich kleinerer Restfehler errechnet als für das Modell mit Vererbungsannahme. Hier scheint das Modell ohne Vererbungsannahme eher gerechtfertigt.

DSDs bieten Versuchspläne mit minimalem Versuchsumfang, die unter Umständen auch das Schätzen vollständiger Oberflächenmodelle zulassen. Daher gehören sie zu Recht zur ersten Wahl bei der Planung von Versuchen. Allerdings können sie eben auf Grund der geringen Versuchszahlen nicht allen Ansprüchen genügen. Ein Versprechen halten sie aber sicher: der Einfluss von Haupteffekten wird zuverlässig erkannt und präzise geschätzt. Sollten sich drei oder weniger Haupteffekte als einflussreich erweisen, kann für diese Faktoren ein komplettes Oberflächenmodell geschätzt werden.

Wenn man alle Ausführungen kurz zusammenfasst so ergeben sich die folgenden Empfehlungen und Bedingungen:

- DSDs sind nur geeignet für stetige Zielgrößen
- DSDs können nur für stetige Faktoren und bivariate Faktoren gebildet werden
- Der Effekt der zu untersuchenden Faktoren muss mindestens doppelt so groß sein, wie der Restfehler.
- Für die Generierung des Designs sollten auf jeden Fall zwei Dummy Faktoren zu den echten Faktoren hinzugefügt werden.
- Die Analyse sollte unbedingt über das spezielle Analyseverfahren erfolgen (bis zu 12 Faktoren).
- Auf jeden Fall sollte eine hierarchische Auswertung erfolgen, um einen Anhaltspunkt für die Konsistenz der gefundenen Modelle zu erhalten.
- Die Zahl der aktiven Terme zweiter Ordnung darf jeweils die Hälfte der Zahl der Gesamtfaktoren (reale plus Dummy) nicht übersteigen.

Sollten die beiden letztgenannten Punkte Anlass zu der Vermutung geben, dass mehr Terme zweiter oder höherer Ordnung eine entscheidende Rolle spielen oder der Modellvergleich zwischen hierarchischem und nicht eingeschränktem Modell (fachlich) nicht plausibel sein, bleibt nur, das Design geeignet zu erweitern, um ausreichende Daten für eine eindeutige Entscheidung zu erhalten.

Literatur

[1] Jones, B., and Nachtsheim, C. J. (2011), A Class of Three-Level Designs for Definitive Screening in the Presence of Second-Order Effects, Journal of Quality Technology, 43, 1–15

[2] Ramsey, P., Weese, M., Montgomery, D., Model Selection Strategies for Definitive Screening Designs Using JMP Pro and R, Discovery Summit 2015 , San Diego, https://community.jmp.com/t5/Discovery-Summit-2015/Model-Selection-Strategies-for-Definitive-Screening-Designs/ta-p/22854

[3] Jones, B., Nachtsheim, C.J. (2017) Effective Design-Based Model Selection for Definitive Screening Designs, Technometrics, 59:3, 319-329

[4] Heinen, B., Jones, B., Auswertung von Definitive Screening Designs, Proceedings der 20. KSFE Schmalkalden. Shaker-Verlag, Aachen 2016

[5] Miller, A., and Sitter, R. R. (2005), Using Folded-Over Nonorthogonal Designs, Technometrics, 47, 502–513