

JMP: Verbessern von Prozessen mit statistischen Modellen

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

Zusammenfassung

Statistische Modelle dienen immer der Beschreibung von Zusammenhängen. In vielen Fällen solchen, bei denen der Einfluss bestimmter Faktoren auf abhängige Zielgrößen bestimmt wird. Das Einbeziehen von Wechselwirkungen, quadratischen Termen oder anderer Transformationen in die Analyse vertiefen das Verständnis, das man vom Zusammenwirken der Faktoren und der Auswirkung auf abhängige Variablen hat. Wenn das Modell zur Optimierung der Zielgrößen dienen soll, ist es in vielen Fällen wichtig, die Faktoren zu identifizieren, die den größten Einfluss auf die Ergebnisse haben; auch dazu dienen statistische Modelle. Wie aber findet man das richtige Modell? Sind klassische statistische Modelle oder Data Mining Verfahren vorzuziehen? Gibt es zuverlässige Vorgehensweisen, das richtige Modell zu finden? Wie optimiere ich einen Prozess, für den mehrere Zielgrößen unterschiedliche Spezifikationen erfüllen sollen? Anhand praktischer Beispiele werden verschiedene Analysesituationen dargestellt, die jeweils zu erfolgreichen Prozessverbesserungen führten.

Schlüsselwörter: Modellierung, Lineares Modell, pönalisierte Regression, Prognose, Optimierung

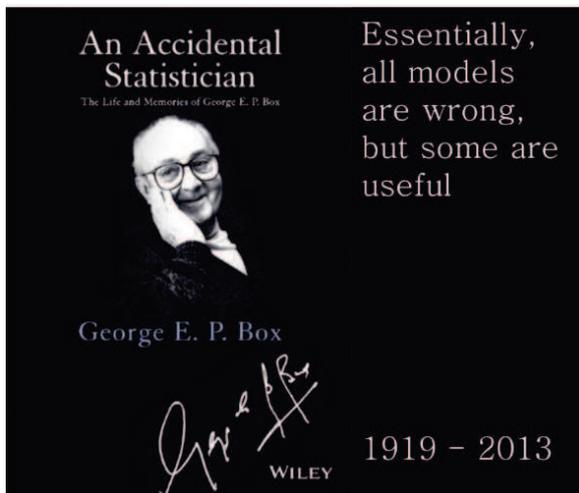
1 Einleitung



Statistik wird von Statistikern gemacht (ich bin auch einer), aber nicht nur. Die Anwendung von Statistik gehört zu vielen Berufsbildern, ohne dass Statistik ein fundamentaler Bestandteil der jeweiligen Ausbildung gewesen wäre. Sonst wären die Betreffenden ja Statistiker geworden und nicht Biologen, Chemiker, Ärzte, Ingenieure, oder Anderes.

Angewandte Statistik beruht auf Daten und angefangen bei der Planung der Datenerhebung über die Erhebung selbst, die Speicherung, Aufbereitung bis hin zur Analyse der Daten ist ein langer Weg zu gehen, ehe eine statistische Aussage getroffen werden kann. Angesichts ihres Wissens um die vielfältigen Implikationen dieser Verarbeitungskette erheben Statistiker oft die Warnung, Statistik gehöre nicht in Laienhände, denn da könne so viel Unheil angerichtet werden. Natürlich ist diese Warnung berechtigt – ne-

benbei stärkt sie auch die Bedeutung des Berufsstands – aber Statistik gehört nun mal auch in die Hände von Spezialisten fast aller Fachbereiche.



Hinzu kommt, dass statistische Verfahren immer leichter verfügbar werden und teilweise schon versteckt zum Einsatz gelangen, denkt man nur an die vielen Fitness- und Gesundheitsapps auf den Smartphones dieser Welt.

Die nebenstehende Aussage von George Box hat mich anfangs eher verstört. Aber vielleicht liegt in der Konsequenz ein Ansatz, der hilft, den rigor statisticae mit der laienhaften Anwendung zu versöhnen.

2 Uneinheitliche Ergebnisse legen mangelnde Datenlage nahe

Die Daten dieses Beispiels stammen aus Versuchen zur Fettverdauung, genauer Möglichkeiten zur Steigerung der Lipaseaktivität, die in Blutproben gemessen wird. Es werden insgesamt sechs Faktoren überprüft, die diese Aktivität beeinflussen können. Ein aktuelles Versuchsdesign, das den Einfluss einzelner Faktoren mit sehr geringen Versuchszahlen zu ermitteln erlaubt, ist das Definitive Screening Design [1].

Tabelle 1: Versuchsergebnisse

TRIS	pH	L-ASP	P5P	OG	MDH	Elevated Serum 30C
60	7	160	0,18	12	3375	120,58
60	7	240	0,1	20	375	118,95
60	7,8	240	0,02	4	3375	108,00
60	8,6	160	0,02	20	1875	107,20
60	8,6	200	0,18	4	375	106,65
80	7	160	0,02	4	375	110,67
80	7,8	200	0,1	12	1875	125,15
80	8,6	240	0,18	20	3375	127,01
100	7	200	0,02	20	3375	117,23
100	7	240	0,18	4	1875	112,01
100	7,8	160	0,18	20	375	112,68
100	8,6	160	0,1	4	3375	108,56
100	8,6	240	0,02	12	375	98,25

Die dreizehn Versuche sind mit Ihren Ergebnissen in Tabelle 1 dargestellt. Ziel des Screenings ist es ja, die wirksamen Faktoren herauszubekommen, wobei möglicherweise auftretende Wechselwirkungen und nichtlineare Effekte die Entscheidung erschweren. Bei diesen geringen Versuchszahlen lassen sich natürlich auch keine Wirkungsflächenmodelle anpassen,

d.h. Modelle, die alle Haupteffekte, deren quadratische Terme und paarweise Wechselwirkungen beinhalten.

Eine sinnvolle Analyse baut also von vornherein darauf auf, dass sich einzelne Faktoren als unwirksam herausstellen. Die üblichen Verfahren zur Auswahl von Faktoren sind schrittweise Regression und pönalisierte Regression.

Tabelle 2: Terme, die bei verschiedenen Selektionsverfahren in das Modell aufgenommen wurden

Term	Methode		
	DSD	Elastic Net	Schrittwe. R.
Achsenabschnitt	1	1	1
L-ASP	0	1	1
MDH	1	1	1
OG	1	1	1
P5P	1	1	1
pH	1	1	1
TRIS	0	1	1
L-ASP*L-ASP	0	1	0
MDH*MDH	0	0	1
OG*MDH	0	0	1
P5P*MDH	1	0	0
P5P*OG	0	0	1
P5P*P5P	0	1	1
pH*MDH	1	0	0
pH*OG	0	0	1
pH*P5P	1	0	0
pH*pH	0	1	0
TRIS*TRIS	0	1	0

Speziell zugeschnitten auf die Anwendung bei Definitive Screening Designs ist ein mehrstufiges Verfahren unter Einbeziehung schrittweiser Regression, das auch zur Anwendung kam. Vergleicht man die Ergebnisse (Tabelle 2), so sieht man, dass die verschiedenen Vorgehensweisen zu sehr unterschiedlichen Modellen führen.

Die meisten nehmen alle Haupteffekte in das Modell auf, aber Wechselwirkungen und quadratische Terme werden höchst unterschiedlich bewertet. Auf den ersten Blick erkennt man diesen Sachverhalt in den Analysediagrammen (Abbildung 1), die den geschätzten Einfluss der Faktoren auf die Zielgröße leicht erkennen lassen. Auf Grund der vorhandenen Datenbasis kann man also weder Effekte ausschließen noch ein vernünftiges Modell konstruieren. Es ist also erforderlich, mit weiteren Experimenten die Datenbasis zu erweitern.

Analysediagramm - Vorhersageanalyse

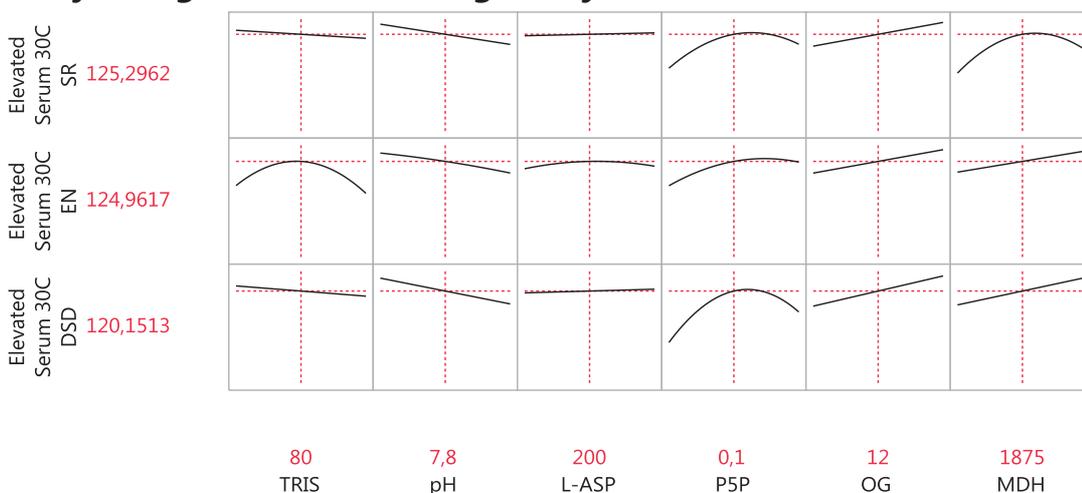


Abbildung 1: Analysediagramm, Hauptfaktoren und quadratische Effekte verschiedener Selektionsverfahren für Screeningversuche

Tabelle 3: Terme verschiedener Selektionsverfahren nach Erweiterung der Daten

Term	Methode	
	Elastic Net	Schritt w. R.
Achsenabschnitt	1	0
MDH	1	1
OG	1	1
P5P	1	1
pH	1	1
TRIS	1	1
MDH*MDH	1	0
OG*MDH	1	1
OG*OG	1	1
P5P*P5P	1	1
pH*MDH	1	1
pH*OG	0	1
pH*P5P	1	1
pH*pH	1	1
TRIS*TRIS	1	0

Nach einer Erweiterung auf insgesamt 35 Versuche zeigen die beiden hier anwendbaren Verfahren schrittweise Regression und pönalisierte Regression fast identische Ergebnisse (Tabelle 3). Bei der nun vorhandenen Datenmenge lässt sich auch ein vollständiges Oberflächenmodell anpassen und als Referenz heranziehen. Das Analysediagramm (Abbildung 2) zeigt in diesem Fall nicht nur die Übereinstimmung der verschiedenen Modelle, es zeigt auch, dass die reduzierten Modelle sich in Ihrer Prognosefähigkeit nicht von dem vollständigen Modell unterscheiden. Auch wenn man die Modelle nutzt, um die Zielgröße zu optimieren – hier ist Maximieren gefragt – kommen alle zu identischen Ergebnissen.

Analysediagramm - Vorhersageanalyse

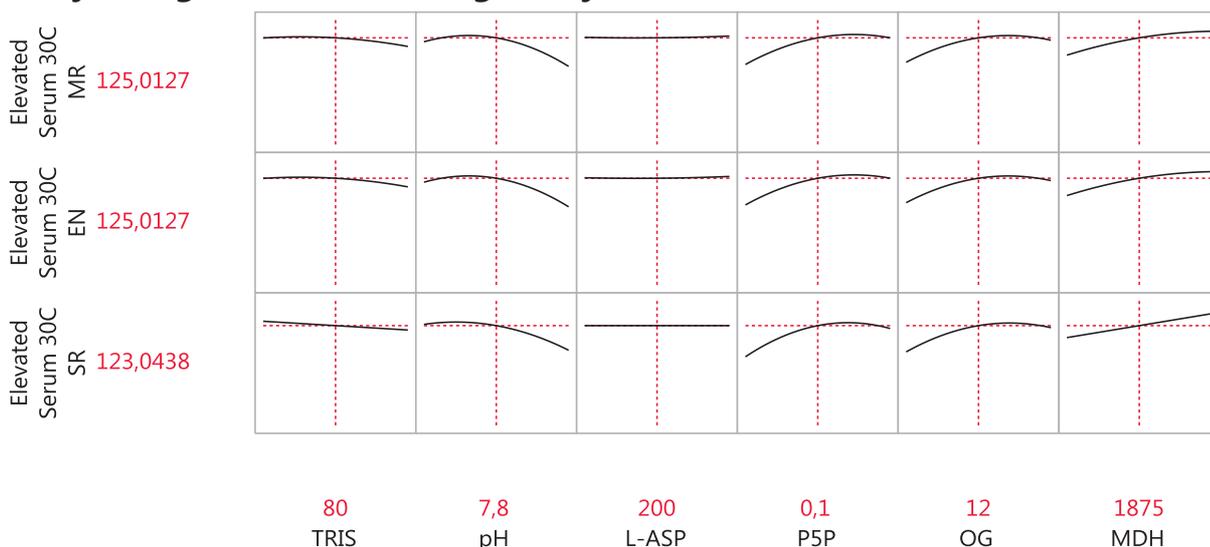


Abbildung 2: Analysediagramm, Hauptfaktoren und quadratische Effekte verschiedener Selektionsverfahren nach Erweiterung der Daten

Die einfache Verfügbarkeit mehrerer Modellierungstechniken eröffnet den Weg zu einer leicht verständlichen Bewertung der Modelle. Wenn Modelle, die mit unterschiedlichem Vorgehen erstellt wurden, sich in ihrem Ergebnis deutlich unterscheiden, liegt

ein grundsätzliches Problem vor. In diesem Fall lag es, wie leicht zu erkennen war, an der nicht ausreichenden Datenbasis.

3 Datengetriebene Versuchsplanung

Produktionsprozesse werden zunehmend elektronisch gesteuert, was bedeutet, dass sowohl Plan- als auch Ist-Daten in größerem Umfang zur Verfügung stehen. Natürlich ist das für quantitative Analyse erfreulich und auch notwendig, aber nicht immer hinreichend. Das folgende Beispiel stammt aus der Produktion von LCD Bildschirmen.

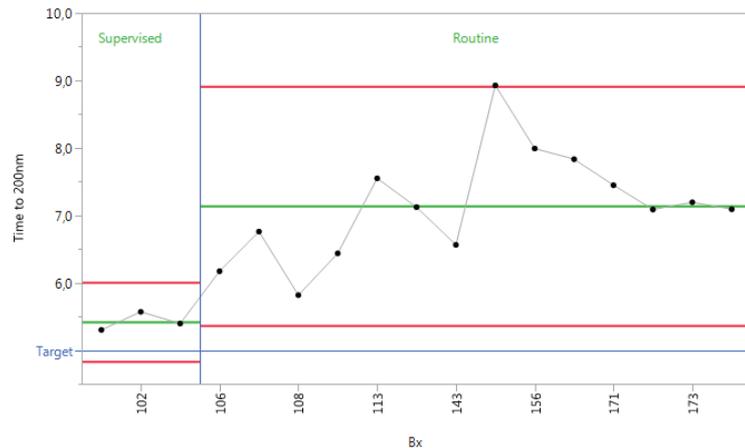


Abbildung 3: Perlmahlwerk **Abbildung 4:** Durchlaufzeiten der Perlmühle

Ein Schritt des Produktionsprozesses besteht daraus, die verwendeten Kristalle in Perlmühlen auf die Größe von 200 Nanometer (nm) zu mahlen (Abbildung 1). Dieser Schritt erwies sich wegen sehr stark variierender Mahlzeiten als Engpass in der Prozesskette. Die für die Produktionsplanung durchgeführten Läufe zeigten noch akzeptable Ergebnisse, aber bei den ersten Chargen wurden drastische Verschlechterungen beobachtet (Abbildung 4). Von 17 Faktoren waren die Einstellungen für jede Charge bekannt.

Tabelle 4: Faktorenbewertung durch Entscheidungsbaum

Spaltenbeiträge				
Term	Anzahl Teilungen	SQ		Anteil
Temp(Pot)	1	4,88633907		1,0000
Pig Bx	0	0		0,0000
%Pigment	0	0		0,0000
Pump rpm	0	0		0,0000
Temp(Out)	0	0		0,0000
Temp(CW)	0	0		0,0000
Mill rpm	0	0		0,0000
%Beads	0	0		0,0000
P1	0	0		0,0000
P2	0	0		0,0000
P3	0	0		0,0000
P4	0	0		0,0000
P(In)	0	0		0,0000
P(Out)	0	0		0,0000
Time PreMix	0	0		0,0000
Transfer Time	0	0		0,0000
Temp PreMix	0	0		0,0000

Die statistische Analyse brachte zunächst keinen weiteren Aufschluss, da alle Analysen nur je eine Variable als einflussreich zeigten, beispielhaft an der Ergebnisdarstellung der Bewertung durch die Bildung eines Entscheidungsbaumes, aber auch lineare Modelle kamen zu derselben Einschätzung.

Tabelle 5: Faktorenbewertung durch Bootstrap (Random) Forest

Spaltenbeiträge				
Term	Anzahl Teilungen	SQ		Anteil
Temp(Pot)	1905	1,27706767		0,3139
Temp(Out)	1607	1,04789632		0,2576
%Beads	1034	0,49355042		0,1213
P(In)	794	0,30726988		0,0755
%Pigment	771	0,30402484		0,0747
P(Out)	450	0,17106896		0,0421
Temp(CW)	345	0,09428938		0,0232
Temp PreMix	271	0,06755552		0,0166
Pig Bx	233	0,0581695		0,0143
P1	268	0,05712317		0,0140
Pump rpm	209	0,04231738		0,0104
Mill rpm	166	0,02673738		0,0066
P2	148	0,0257638		0,0063
P3	139	0,02572694		0,0063
Transfer Time	134	0,02517231		0,0062
Time PreMix	124	0,02286862		0,0056
P4	118	0,02138769		0,0053

Allein mit einem Bootstrap Forest konnte eine sinnvolle Einschätzung unter Einbeziehung aller Variablen erzielt werden. Um eine zuverlässige Aussage zu erhalten, wurden die ersten sieben Variablen ausgewählt und dafür ein Versuchsplan nach einem Definitive Screening Design erstellt.

Für die daraus gewonnenen Daten zeigt die Analyse lediglich drei wesentliche Faktoren, so dass sich aus den vorliegenden Daten ein vollständiges Modell mit Termen erster Ordnung schätzen ließ, das eine hohe Anpassung zeigte. Darauf aufbauend ließen sich optimale Einstellungen ableiten, die homogene, kurze Mahlzeiten erwarten ließen. Das technische Problem wurde in drei Schritten gelöst:

- Analyse bestehender Daten (wichtig: geeignete Methode finden)
- Geplante Versuche durchführen (DSD als effizientes Design)
- Optimierung und Verifizierung

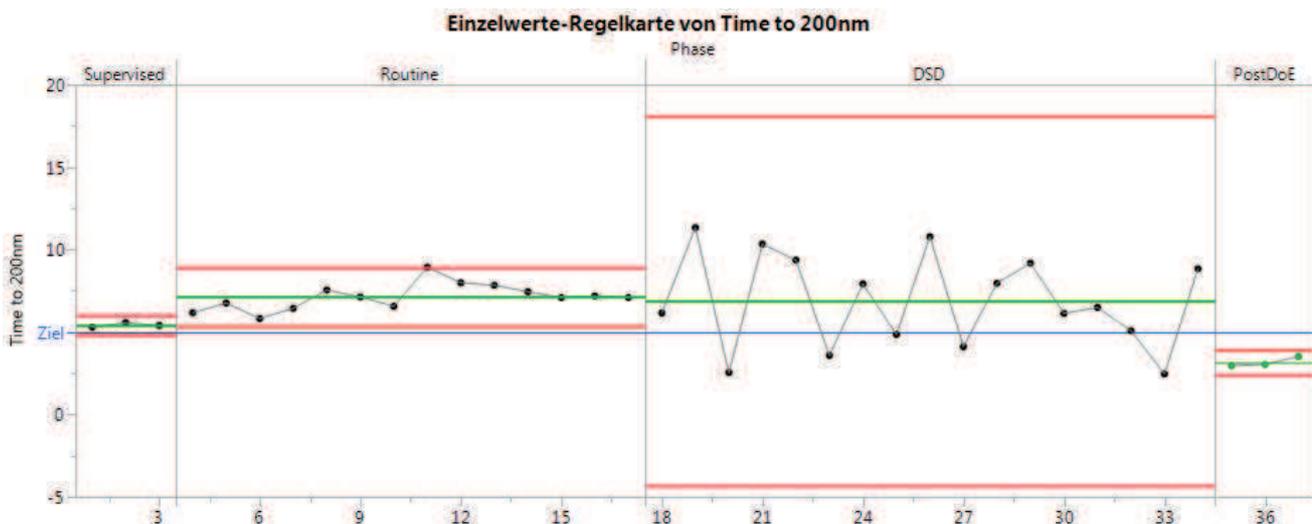


Abbildung 5: Laufzeiten Kugelmühle in verschiedenen Entwicklungsphasen

4 Analyse und Simulation

In den vorangegangenen Beispielen war es jeweils erforderlich, die Datenbasis durch zusätzliche Versuche zu erweitern. Das muss nicht zwingend notwendig sein. Manchmal reicht auch die statistische Analyse bestehender Daten. In dem zugrunde liegenden Beispiel aus der Produktion von Formteilen aus weißem Kunststoff ist es immer wieder zu Produktionsabbrüchen gekommen, die durch Verklumpungen des Kunststoffes in den Gussformen verursacht wurden.

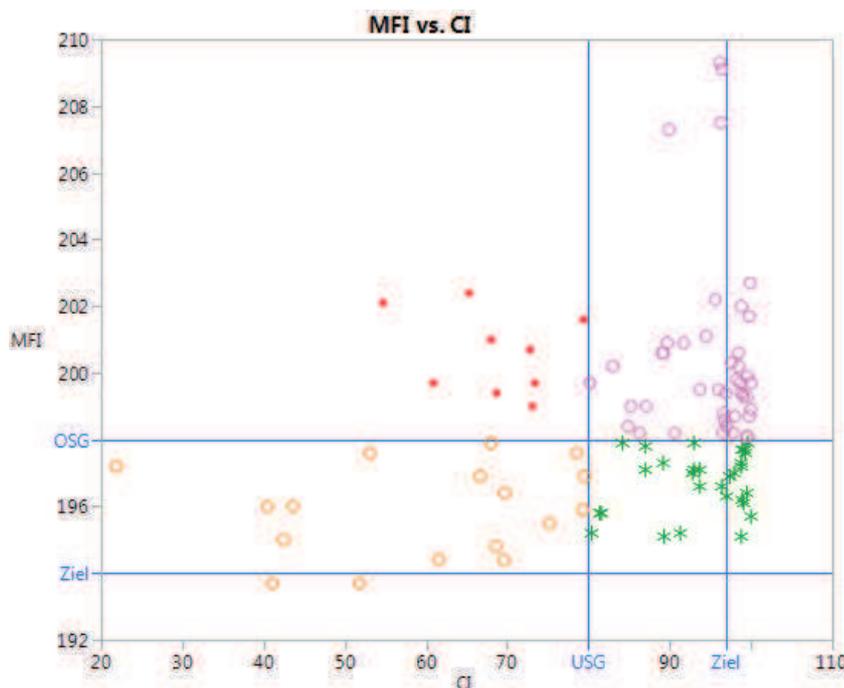


Abbildung 6: Zielgrößen und Spezifikationsgrenzen

Da der weiße Farbstoff die Fließfähigkeit des Kunststoffes herabsetzt, liegt es nahe, zwei Zielgrößen zu beachten: die Fließfähigkeit (Melt Flow Index, MFI) und die Farbdichte (Color Index, CI). In einem dreistufigen Prozess wird dem Farbstoff Wasser zugesetzt, beides mit dem flüssigen Polymer vermischt und granuliert. Wenn die fertige Mischung zu zähflüssig ist, wird ein Fließmittel zugesetzt.

Für den Fließindex gibt es eine obere Spezifikationsgrenze (OSG), für den Farbindex eine untere (USG). Wie man sieht (Abbildung 6), hält nur ein Bruchteil der Chargen beide Grenzen ein (Sterne). Zusätzlich zu diesen Zielgrößen kennt man pro Charge insgesamt acht Faktoren, die die Produktion beschreiben und für eine Analyse zur Verfügung stehen. Kann man die Abweichungen auf bestimmte Kombinationen dieser Faktoren zurückführen?

Quelle	Log-Wertigkeit	P-Wert
M%	29,626	0,00000
SA*SA	24,353	0,00000
Xf*Xf	20,412	0,00000
SA*SA*SA	16,130	0,00000
Xf	8,624	0,00000
Xf*Xf*Xf	2,040	0,00912
SA	1,754	0,01760

Abbildung 7: Terme eines linearen Modells

In diesem Fall führen alle Modellierungsverfahren zu vergleichbaren Ergebnissen. Schrittweise Regression, Bootstrap Forest oder pönalisierte Regression weisen auf dieselben Faktoren und Interaktionen hin.

Ein eher umfassendes Modell, das für beide Zielgrößen gute Prognosen liefert, wird die Basis für ein lineares Modell, das zur Optimierung herangezogen wird. Darin sind als Faktoren:

- M% der Anteil des zugesetzten Fließmittels
- SA die Leistung des Rührwerks für die farbige Emulsion und
- Xf die Menge des zugesetzten Farbstoffes.

Es wurden keine Wechselwirkungen als relevant eingeschätzt, aber sowohl XF als auch SA treten in quadratischen und kubischen Termen auf.

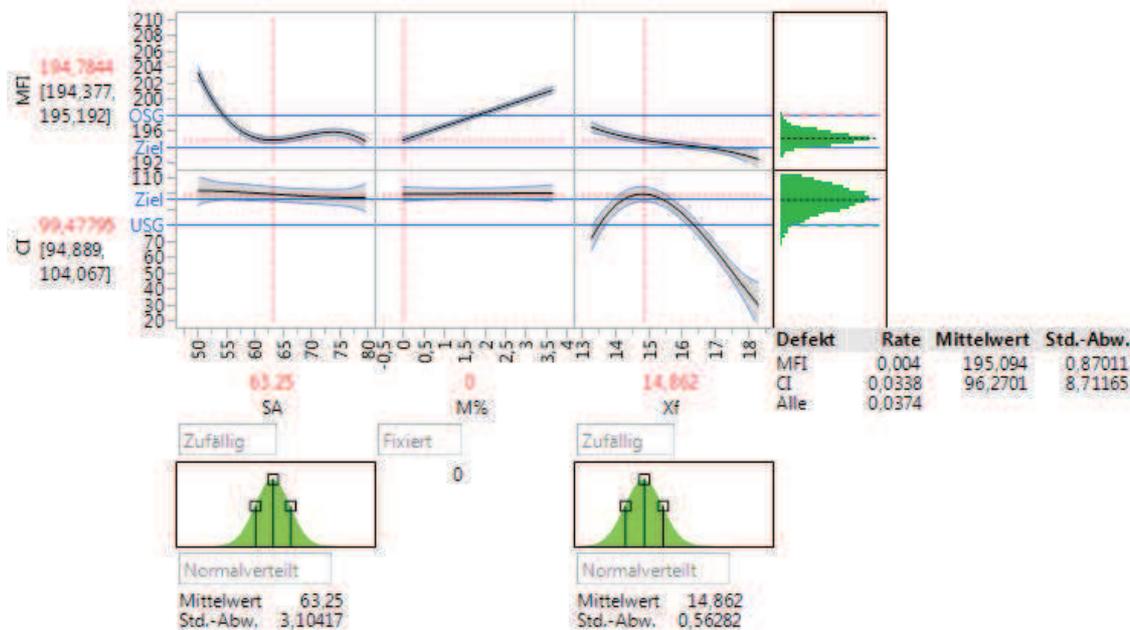


Abbildung 8: Optimale Betriebsparameter und Simulation der Ausschussraten

Nutzt man dieses Modell zur Optimierung s. (Abbildung 8: Analysediagramm) so stellt sich überraschenderweise heraus, dass die besten Ergebnisse erzielt werden, wenn kein Fließmittel (M%) zugesetzt wird. Simuliert man diese Eingangsgrößen und lässt sinnvolle Schwankungen der Stellgrößen zu, dann bringt man die Ausschussrate unter vier Prozent. Kein Vergleich zu den fast 70% Problemfällen, die die Ausgangslage der Analyse darstellten.

Fazit

Nicht immer ist zu Beginn einer Analyse erkennbar, ob die vorliegenden Informationen ausreichen. Auch ergibt sich der gesamte Analysebedarf oft erst während der Auswertungen. Auf jeden Fall aber schafft die Modellierung technischer Prozesse ein tieferes Verständnis für die Zusammenhänge zwischen den unabhängigen Faktoren und den abhängigen Ergebnissen. Dieses bessere Verständnis ist die Basis für Optimierung, deren Ergebnisse man durch Simulation quantitativ bewerten kann. Statistik ist sicher nicht die Lösung der Probleme, aber sie liefert den Schlüssel dazu.

Literatur

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