Analysen zur Anwendung der ‚Hohenheim - Gülzower - Serienauswertung' im regionalisierten Sortenversuchswesen in Mecklenburg - Vorpommern

D i s s e r t a t i o n

zur Erlangung des akademischen Grades
Doctor rerum agriculturarum
(Dr. rer. agr.)

eingereicht an der
Lebenswissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

von
Dipl. Ing. agr. Volker Michel
geboren am 13. April 1963 in Rostock

…

Versuchsserien - Theorie, Literatur

Landwirtschaftliche Feldversuche werden als Versuchstypus eingeordnet, der für die Prüfung unter praxisnahen Bedingungen steht. Damit ist der geplante Aussagebereich a priori sehr viel weiter gefasst, als es bei Modellversuchen, Gefäßversuchen, in Klimakammern u.ä. der Fall ist. Gegenüber letzteren nimmt im Feldversuch die Anzahl steuerbarer Konstantfaktoren ab und die Repräsentativität für die Praxis nimmt zu (Rasch, 1987).

Hierbei steht insbesondere die Unvorhersagbarkeit der Witterung oder unter vielem anderen z.B. auch des Schaderregerdrucks während des Anbaus eines Feldversuches zu Buche. Insofern impliziert die Anforderung an Repräsentativität für die Praxis geradezu die Notwendigkeit der Anlage von Feldversuchen in mehreren Jahren. Bereits Kuckuck und Mudra (1950) gehen für die Pflanzenzüchtung davon aus, dass Neuzüchtungen nur auf Grund mehrjähriger Versuchsergebnisse zu beurteilen sind.

Auch die Repräsentativität einer einortigen Versuchsanstellung wird häufig den Erfordernissen der Praxis nicht gerecht. In aller Regel sollen abgeleitete Aussagen

nicht nur für genau diesen Schlag gelten, sondern für ein mehr oder weniger breit definiertes Anbaugebiet. Selbst in relativ kleinräumig definierten Regionen sollte man nicht von Homogenität des Einflussfaktors *Standort* ausgehen. Insofern erfordert auch die Frage, ob die Behandlungen / die Prüfglieder in einer Region einheitlich reagieren bzw. wie bedeutsam Interaktionen sind, häufig nicht nur eine *mehrjährige*, sondern gleichzeitig auch *mehrortige* Versuchsanstellung.

Auch Cochran und Cox (1957) unterstellten, dass viele landwirtschaftliche Versuchsanstellungen in der Hoffnung durchgeführt werden, dass die Ergebnisse nicht nur für die Umweltbedingungen in den jeweiligen Einzelversuchen Gültigkeit haben, sondern auf die praktische Landwirtschaft übertragbar sind. Die Gesamtfragestellung erfordert somit i.d.R. eine Versuchsserie, in der der einjährige Feldversuch an einem Standort, hier im Weiteren als Einzelversuch bezeichnet, *nur* ein Element ist. Nach Mudra (1952) sollte es die Regel sein, Versuche über allgemein interessierende Fragen nicht als Einzelversuche, sondern als Versuchsserie zu planen und auszuwerten. Der induktive Schluss auf die Grundgesamtheit *Anbaugebiet und dessen Klima* setzt eine ausreichend repräsentative Stichprobe von Einzelversuchen voraus (Bätz, 1984). Ebenso argumentiert Rasch (1987), dass bei biologischen Problemen aufgrund der großen Anzahl nicht kontrollierbarer Umweltbedingungen die Antwort auf eine Versuchsfrage i.d.R. nur durch eine Versuchsserie möglich ist. Er definiert eine Versuchsserie als eine Serie von Einzelversuchen, die mit der gleichen Fragestellung, mit den gleichen Prüfgliedern räumlich oder / und zeitlich getrennt durchgeführt wird.

Methodische Grundlagen für die Auswertung von Versuchsserien wurden bereits von Yates und Cochran (1938) veröffentlicht. Mudra (1949) stellt die ‚Differenzmethode' in Einzelversuchen mit Randomisationseinschränkungen wie auch in so genannten Streuversuchen (viele Orte ohne Wiederholungen) einer varianzanalytischen Auswertung gegenüber. Ziel der *Differenzmethode* war die Ausschaltung der von Mudra beobachteten häufigen Korreliertheit von Prüfgliedeffekten in Blöcken oder an Orten mit einem damals erforderlichen einfachen mathematischen Ansatz. Dieser Ansatz lag in der Bestimmung der Fehlervarianz einer Differenz aus den Einzeldifferenzen von Prüfgliedpaaren anstelle der Bestimmung aus der Fehlervarianz eines Mittelwertes. Einen solchen Ansatz beschrieb bereits Student (1923). Patterson (1997) beschreibt dies als ‚Methode der direkten Differenzen'. Auch Forkman (2013) greift wieder auf diese Ansätze zurück. Er verfolgt damit vor allem das Ziel, in unbalancierten Datensätzen ausschließlich direkt (in jeweils einem Versuch) gemessene Differenzen zu einer oder zu mehreren durchgängig geprüften Referenzsorten in der Auswertung zu verwenden.

Umfassende Zusammenfassungen zum internationalen Stand der Versuchsserienauswertung erfolgten u.a. durch Cochran und Cox (1957), Bätz (1984) und Patterson (1997). Auf die besondere Bedeutung der Entwicklung des REML-Algorithmus (restricted bzw. residual maximum likelihood) durch Patterson und Thompson (1971) wird im Zusammenhang mit Unbalanciertheit in Versuchsserien eingegangen (s.u.).

Cochran und Cox (1957) demonstrieren an Beispieldaten Probleme und Auswertungsansätze für Versuchsserien u.a. unter Berücksichtigung möglicher heterogener Prüffaktor × Umwelt - Interaktionen, heterogener Fehlervarianzen in Einzelversuchen oder ungleicher Strukturen in den Einzelversuchen.

Bätz (1984) formuliert folgende Schwerpunkte für die Auswertung einer Versuchsserie:

1. Analyse der Ursachen der Prüfglied/Umwelt-Wechselwirkungen,
2. Beurteilung der Verwendbarkeit und des Informationswertes von Versuchen, Versuchsorten und Versuchsjahren,
3. Beurteilung der Ökostabilität (Ertragssicherheit) von Prüfgliedern,
4. Qualifizierung der Entscheidung für einzelne Prüfglieder,
5. Berücksichtigung der Wechselwirkung Prüfglieder/Umwelt bei der Ableitung von Anbauempfehlungen.

In den 80'er Jahren erfolgte die Versuchsserienauswertung vorrangig durch Varianzanalyse mit anschließendem Mittelwertvergleich. Die Beurteilung der Prüfglieder und deren Wechselwirkung wurden als fachlicher Schwerpunkt betrachtet, während die Beurteilung der Orts- oder Jahresunterschiede eine untergeordnete Rolle spielte (Autorenkollektiv, 1987). Unabhängig von der Hypothesenwahl ‚*fix*' oder ‚*zufällig*' für einzelne Faktoren wurde die Frage der Varianzkomponentenschätzung gegenüber Tests zurückgestellt. Umfassend wurde die Berechnung der MQ-Werte (mittlere Abweichungsquadrate) in Zusammenhang mit der Wahl des ‚richtigen F-Tests' auch für zufällige Effekte im Modell diskutiert.

Mudra (1952) beschreibt die Versuchsserien-Auswertung „mehrjähriger Versuche von einem Ort" sowie methodisch analog die von „einjährigen Versuchen aus verschiedenen Orten". Die zusammenfassende Auswertung von mehrjährigen und gleichzeitig mehrortigen Versuchen bezeichnet Mudra noch als äußerst kompliziert. Auch Bätz (1984) beschreibt, dass zu diesem Zeitpunkt mit den verfügbaren Auswertungsverfahren nur 2-fach klassifizierte Auswertungen erfolgen würden (z.B. Versuche und Prüfglieder). Die Auswertungen erfolgten überwiegend auf Basis der Prüfgliedmittelwerte aus den Einzelversuchen unter Berücksichtigung der zu ‚poolenden' Fehlervarianzen der Einzelversuche. Die Auswertung von Versuchsserien für einen Prüffaktor wird durch ein Autorenkollektiv (1987) für verschiedene Modellansätze beschrieben. Richter et al. (1999) beschreiben die Auswertung von Versuchsserien zweifaktorieller Einzelversuche auf der Basis der Einzelwerte.

Vor dem Hintergrund eingeschränkter Informationstechnologie war es insbesondere im Falle unbalancierter Daten über einen großen Zeitraum ein sehr erfolgreicher pragmatischer Ansatz, auch Einzelversuche einer mehrjährigen mehrortigen Serie als statistisch unabhängig zu betrachten. Dabei werden die Variablen *Ort* und *Jahr* in einer einzigen Variablen *Versuch* bzw. *Umwelt* zusammengefasst. Die Auswertung erfolgte dann häufig nach der von Yates (1933) vorgeschlagenen und von Patterson (1978) vertieften FITCON-Methode (method of fitting constants). Patterson (1997) zählt dazu im weiteren Sinne alle für diesen Zweck angepassten linearen Modelle mit einem einzelnen Fehlerterm. Im Sortenwesen erfolgt dies i.d.R. durch eine Mittelwert basierte Serienauswertung über Umwelten (Versuche), wobei die Haupteffekte *Prüfglied* und formal auch *Versuch* fix gesetzt werden und nur deren Wechselwirkung als Fehlerterm zufällig ist. Bereits 1978 beschreibt Silvey, dass diese Methode in der offiziellen Sortenprüfung Großbritanniens routinemäßig eingesetzt wird. Auch vom Bundessortenamt wird dieses Verfahren bereits langjährig für die Erstellung der deutschen Beschreibenden Sortenlisten verwendet (Laidig, 2013). Mit diesem vereinfachten Modellierungs-Ansatz wurden hinsichtlich der begrenzenden Rechnerleistungen Auswertungen komplex strukturierter großer unbalancierter Datensätze rein technisch ermöglicht, die ohne diese Vereinfachungen oft nicht erfolgreich durchlaufen würden. Patterson (1997) beschreibt erheblich kürzere Computer-Rechenzeiten bei FITCON gegenüber REML. Allerdings stellten Piepho und Michel

(2001) an einem langjährigen Datensatz mit Rapssorten auch fest, dass die Annahme der Unabhängigkeit von Versuchen nicht tatsächlich erfüllt war, sondern dass Versuche im gleichen Jahr bzw. am gleichen Ort korrelierte Effekte aufweisen. Sie diskutierten - mit Verbesserung der Rechnerleistungen nun mögliche - Verbesserungen in der der Sachlage adäquaten Modellbildung, wobei auch hier noch von relativ einfachen Varianz-Kovarianz-Strukturen, insbesondere von homogenen Varianzen sämtlicher Zufallseffekte ausgegangen wurde. Auswertungsansätze auf Basis faktoranalytischer Varianz-Kovarianz-Strukturen diskutieren Denis et al. (1997), Gilmour et al. (1998), Piepho und van Eeuwijk (1999) und Piepho (1999). Kelly et al. (2007) beschreiben faktoranalytische Modelle als einfache Form zur Approximation der komplett unstrukturierten Form der genetischen Varianz-Kovarianz-Matrix, die zur besten Modellanpassung führen und in Züchtungsprogrammen auch das Ziel ‚Selektion der besten Genotypen' am besten erreichen.

Die Auswertung von Versuchsserien in der Pflanzenzüchtung diskutiert van Eeuwijk (2007) für Datenstrukturen von Genotyp × Umwelt - Mittelwerten. Dabei werden insbesondere folgende Problemkreise angerissen: Definition von Effekten im Modell als fix oder zufällig, Plausibilität additiver Modelle, Reaktionsnormen der Genotyp × Umwelt - Wechselwirkungen, multiplikative Modelle für die Wechselwirkung, Testung, Modelle zur Berücksichtigung von Varianzheterogenität.

Aufgrund der Bedeutung von Versuchsserien, in denen Einzelversuche *nur* Element der Gesamtauswertung sind, bemängeln bereits Nelder (1986) wie auch Richter et al. (1999), dass Literatur und Software-Lösungen überwiegend für Einzelversuche bereit stehen, für Versuchsserien aber nur bruchstückhaft vorhanden seien. Richter et al. (1999) leiten Auswertungs-Algorithmen für mehrortige und/oder mehrjährige Versuchsserien mit zwei Prüffaktoren (im Einzelversuch) detailliert her, wobei bezüglich der Faktoren Orte und Jahre alle Konstellationen für die Definition *fix* versus *zufällig* Berücksichtigung finden. Über alle Einzelversuche wird hierbei eine einheitliche Anlage in vollständigen Blocks, Varianzhomogenität sowie insgesamt vollständige Balanciertheit der Daten angenommen. Insofern greifen diese Methoden eher im Bereich der Forschung und anbautechnischer, befristeter Versuchsprojekte. In Sortenprüfsystemen können diese Voraussetzungen i.d.R. als nicht erfüllt gelten.

Einschritt- oder Zweischrittanalyse

Grundsätzlich können Versuchsserien in Form einer Einschritt- oder einer Zwei- bzw. Mehr-Schritt-Analyse ausgewertet werden (Piepho et al., 2012). Die Einschrittanalyse setzt auf den Daten der Parzellen in den Einzelversuchen auf. Aus theoretischer Sicht liefert sie unter Annahme des Einschritt-Analysemodells exakte Schätzwerte für alle fixen Effekte (BLUE) und für alle zufälligen Effekte (BLUP). Smith et al. (2001a) bezeichnen die Einschrittanalyse als den ‚*goldenen Standard*'. Dies wird durch Simulationsergebnisse von Welham et al. (2010) bestätigt.

Bei der Zweischrittanalyse werden zunächst die Einzelversuche entsprechend ihren jeweiligen Besonderheiten ausgewertet und Sortenmittelwerte mit ihren Standardfehlern berechnet und diese dann im zweiten Schritt über die gesamte Versuchsserie verrechnet. Die Einzelversuche können sich bereits planungsseitig, z.B. in der Wahl der Versuchsanlage oder sogar der Anzahl Prüffaktoren, unterscheiden. Hinzu kommen u.U. weitere Unterschiede im Zuge der Auswertung der Einzelversuche, z.B. in Folge einer versuchsspezifischen Modellselektion oder der Nutzung einer räumlichen Modellierung des Fehlers.

Die Güte einer Zweischrittanalyse muss sich an der Einschrittanalyse bemessen, sofern die Einschrittanalyse alle Besonderheiten der Einzelversuche (z.B. die Unterschiedlichkeit von Fehler- und Blockvarianzen zwischen den Versuchen) in der Modellbildung vollständig berücksichtigt. Letzteres erreicht aber bei großen unbalancierten Datensätzen und komplexer Modellstruktur oft seine Grenzen. Sehr lange Rechenzeiten oder auch Abbruch wegen zu geringer Speicherkapazität treten nicht selten auf, wenn die MIXED Prozedur von SAS® genutzt wird. Dann ist häufig die Nutzung einer Zweischrittanalyse eine bessere Alternative als ein reduziertes Einschritt-Modell, welches relevante Gegebenheiten unberücksichtigt lässt. Die Zweischrittanalyse kann also erforderlich werden, wenn die Versuchsanlagen und Auswertungsmodelle zwischen den Versuchen erheblich variieren, allerdings zu Lasten der Effizienz im Vergleich zu einer adäquaten Einschrittanalyse.

Frensham et al. (1997) untersuchen die Frage der Genotyp × Umwelt - Varianzheterogenität in einer Zweischrittanalyse. In den meisten Fällen sei es nötig, eine Versuchsniveau abhängige Genotyp × Umwelt - Interaktion zu berücksichtigen.

Es ist also wichtig, für die Zweischrittanalyse eine Methode zu finden, die die Einschrittanalyse möglichst gut reproduziert. Eine zentrale Rolle kommt hierbei der Methode zu, mit der die Mittelwerte aus dem ersten Schritt im zweiten Schritt gewichtet werden. Möglich ist hierbei als einfachste Variante die Arbeit ohne Gewichtung, was einer Gleichgewichtung aller Einzelversuchs-Sortenmittelwerte unter Annahme von Varianzhomogenität gleichkommt. Methodische Ansätze der differenzierten Gewichtung für unterschiedliche Situationen diskutieren u.a. Smith et al. (2001b), Möhring und Piepho (2009) und Welham et al. (2010). Piepho et al. (2012) stellen eine Methode der Mehr-Schritt-Analyse vor, in der die vollständigen Varianz-Kovarianz-Matrizen der adjustierten Mittelwerte der Einzel-Umwelten in der Serienauswertung einbezogen werden. Diese Methode kann die Ergebnisse der Einschrittanalyse vollständig reproduzieren, wenn man bekannte Varianzkomponenten annimmt.

Der Aspekt differenzierter Gewichtung der Sortenmittelwerte aus Einzelversuchen wird in dieser Arbeit vertieft und in seiner Wirkung auf die Serienmittelwerte an realen Daten untersucht.

Hypothese ‚fix' oder ‚zufällig' für Versuche bzw. Umwelten

Die Planung und Auswertung von Versuchsserien steht maßgeblich im Zusammenhang mit der Hypothese ‚*fix*' oder ‚*zufällig*' für die Versuche bzw. für Orte oder / und für Jahre (Bätz, 1984). Die Fragen der Repräsentanz, des Aussagebereiches bzw. der Verallgemeinerungsfähigkeit hängen damit zusammen. Häufig, insbesondere bei Versuchsanstellungen mit Beratungsauftrag, deckt sich die von Bätz (1984) formulierte Definition des Aussagebereiches ‚*Anbaugebiet und dessen Klima*' mit dem Ziel der Versuchsanstellung. Auch Richter et al. (1999) gehen davon aus, dass sich im Zuge der Serienauswertung im Allgemeinen eine zusammenfassende Beurteilung über alle Umwelten anschließen soll. Dies impliziert dann bereits eine Zielstellung der Versuchsanstellung, für die i.d.R. die Hypothese ‚*zufällig*' für die Umwelten angestrebt werden sollte. Nach Buhtz und Bätz (1984) kann die Hypothese ‚*zufällig*' angenommen werden, wenn die Orte eine zufällige und repräsentative Stichprobe aus einem Anbaugebiet darstellen und die Jahre hinsichtlich ihres Witterungsverlaufes ebenfalls eine zufällige und repräsentative Stichprobe ergeben. Unterschiedliche Konstellationen werden von den Autoren am Beispiel der Pflanzenzüchtung und Sortenprüfung dargestellt und hinsichtlich des sich ergebenden Aussagebereiches systematisiert (Tab. 1). Auch vom Autor dieser Arbeit werden hier

im Weiteren überwiegend Beispiele aus dem Sortenversuchswesen aufgegriffen, wobei aber synonym für den Faktor *Sorte* auch andere Faktoren stehen könnten.

Aussagebereich von Feldversuchen nach Buhtz und Bätz (1984)

Modell Nr.	**Stichprobe (Einzelversuch)**	**Hypothese für Orte**	**Hypothese für Jahre**	**Aussagebereich (Grundgesamtheit)**
1.	Orte in einem Jahr	zufällig	(fix)	Anbaugebiet mit Witterungsbedingungen, die denen des Versuchsjahres entsprechen
2.	Orte in einem Jahr	fix	(fix)	Mittlere Anbaubedingungen der Orte bei Witterungsbedingungen, die denen des Versuchsjahres entsprechen
3.	Ort in mehreren Jahren	(fix)	zufällig	Ort und dessen Klima
4.	Ort in mehreren Jahren	(fix)	fix	Ort mit Witterungsverhältnissen, die dem „Mittel“ der Versuchsjahre entsprechen
5.	Orte und Jahre	zufällig	zufällig	Anbaugebiet und dessen Klima
6.	Orte und Jahre	fix	fix	Mittlere Anbaubedingungen der Orte mit Witterungsverhältnissen, die dem „Mittel“ der Versuchsjahre entsprechen
7.	Orte und Jahre	zufällig	fix	Anbaugebiet mit Witterungsverhältnissen, die dem „Mittel“ der Versuchsjahre entsprechen
8.	Orte und Jahre	fix	zufällig	Mittlere Anbaubedingungen der Orte und deren Klima

( ) ‚Jahr‘ bzw. ‚Ort‘ in nur einer Stufe, also in der Modellgleichung vernachlässigbar

Die von Bätz (1984) formulierte Zielstellung ‚*Schluss auf die Grundgesamtheit Anbaugebiet und dessen Klima‘* (Zeilennummer ‚5‘ in Tab. 1) trifft uneingeschränkt auf das in dieser Arbeit diskutierte Sortenprüfsystem mit dem Ziel der Sortenberatung zu. Die anderen in Tab. 1 aufgeführten Ansätze passen in einigen Fällen für geplante Versuchsserien z.B. zu anbautechnischen Fragen, bei denen im Gegensatz zum Sortenversuchswesen die Fragestellung a priori zeitlich und räumlich eng gefasst wird oder bei denen die Versuchsorte bewusst so distinkt gewählt werden, dass die Reaktionen auf jeden Standort a priori im Zentrum der Fragestellung stehen und eine Mittelwertbildung über Orte selbst bei nicht signifikanter Interaktion eher nur eine Zusatzinformation darstellt.

Auch wenn häufig die Hypothese ‚*zufällig*‘ für Jahre und Orte anzustreben ist, so ist aber der Idealfall einer tatsächlich zufälligen Wahl der Orte im absoluten Sinne nicht realisierbar. Bestenfalls kann ein Standortnetz etabliert werden, das fachlich-subjektiv als *hinreichend* repräsentativ sowohl für das Mittel als auch für die Variation der Standortbedingungen eingeschätzt wird. Es wird bei der Hypothesenwahl also immer ein gewisser Widerspruch zwischen Anforderung aus der Zielstellung und tatsächlicher ‚*Natur*‘ eines Faktors hinsichtlich Auswahl / Randomisation der Um-

welten bestehen, sobald ein Umwelt-Faktor, speziell aber der Faktor *Ort,* als zufällig eingestuft wird. Insofern verwendet bereits Bätz (1984) die relativierende, pragmatische Formulierung „…*ausreichend* repräsentative Stichprobe…". Es gibt keine rein statistischen Kriterien für diese Entscheidung, der Aussagebereich muss vorrangig durch sachlogische Erwägungen bestimmt werden (Autorenkollektiv, 1987).

In frühen Literaturquellen neigte man im Zweifelsfall eher zu einer formalen Anwendung der Hypothese ‚*fix*'. So argumentiert z.B. Bätz (1984), dass bei geringer Anzahl von Jahren (oder Orten) eher der Ansatz ‚*fix*' verwendet werden sollte. Für zufällige Effekte mit weniger als fünf bis zehn Stufen, welche einen zu testenden fixen Effekt nicht enthalten, kann es nach Piepho et al. (2003) vorteilhaft sein, diesen als fix zu definieren. Ursache ist die bei wenigen Stufen u.U. sehr geringe Genauigkeit der Varianzschätzungen, wie es in Analogie hierzu auch bei der Nutzung der Inter-Block-Information in Versuchsanlagen mit unvollständigen Blocks diskutiert wird. Bei geringer Anzahl von Stufen eines seiner Natur nach zufälligen Faktors oder Effektes im Modell wird also u.U. eine formale Umdefinition zu ‚fix' pragmatisch sinnvoll sein. Unabhängig von einer derartigen Umdefinition ist es für die Versuchsplanung ebenso wie für die Interpretation und Bewertung der Ergebnisse wichtig, dass sich der Versuchsansteller in allen Phasen darüber klar ist, ob die Zuordnung zu ‚fix' oder ‚zufällig' der ‚*Natur*' des Faktors entspricht. Andernfalls kann z.B. bei Definition von ‚Jahr' als fix die Genauigkeit der Mittelwertschätzung von Sorten nach nur zwei- oder dreijähriger Auswertung u.U. als hoch eingeschätzt werden, obgleich die (dann unbekannte) Vorhersagegenauigkeit für neue (zufällige) Jahre gering ist.

Um ein kleines Anbaugebiet zu repräsentieren, reichen häufig vier bis sechs Orte aus, die Anzahl Jahre sollte aber lt. Autorenkollektiv (1987) nicht vorab festgelegt, sondern durch ein sequentielles Prinzip bestimmt werden. Hierbei soll aufgrund der rückwirkenden Beschreibung der Umweltbedingungen beurteilt werden, ob die vorliegenden Versuche die Grundgesamtheit ausreichend repräsentieren oder ob weitere Versuche in zusätzlichen Jahren erforderlich sind. Allerdings wird dieses Prinzip als subjektiv-pragmatisch eingeschränkt, da es praktisch schwierig ist, eine Klassifikation der Witterungsbedingungen vorzunehmen. Nach Richter et al. (1999) kann die Repräsentativität von Jahren erst nach Ablauf der Versuchsdauer beantwortet werden.

Den Ansatz, Aussagen eher für vordefinierte (also fixe) Umwelten abzuleiten, zeigt ein durch Hamblin et al. (1980) beschriebenes Vorgehen, wonach aufgrund von Vorauswertungen solche Standorte auszuwählen seien, die für den *Durchschnitt* des Anbaugebiets repräsentative Ergebnisse liefern. Kienzl (1974) schlägt solche Vorauswahlen bei bayerischen Landessortenversuchen nicht nur für die Versuchsorte vor, sondern schließt auch ‚*untypische*' Jahre von der Auswertung aus. Nach Ansicht des Autor vorliegender Arbeit ist Repräsentativität von Umwelten aber letztlich nicht loszulösen von der Hypothese ‚*zufällig*', die Ausgrenzung z.B. von Jahren sollte nur in extremen Einzelfällen erfolgen. Die einbezogenen Umwelten sollten nicht nur den (vermeintlichen) Durchschnitt, sondern auch die Variabilität im Anbaugebiet und insbesondere die Reaktion der Prüfglieder auf diese uneingeschränkt repräsentieren. Auch Richter et al. (1999) stellen die Reaktion der Prüfglieder auf die durch die Standorte verursachte Variabilität in den Mittelpunkt des Interesses. Speziell auch das Verhalten z.B. von Sorten in besonderen Situationen, die u.U. vermeintlich ‚untypisch', aber doch Teil der Umweltvariabilität sind und die sich also in ähnlicher Weise wiederholen können, sind für die Nutzung in der Beratung besonders wichtig. Auch ein Landwirt will nicht nur in einem *typischen* Jahr

die richtige Sortenwahl treffen, sondern im langjährigen Mittel, das auch besondere Jahre einschließt.

In diesem Sinne repräsentieren Versuchsjahre die klimatische Bandbreite und sollten m.E. als zufällig angesehen werden, solange nicht willkürlich Jahre ausgewählt oder ausgeschlossen werden, z.B. weil sie in der Rückschau als *typisch* oder *untypisch* klassifiziert werden. Da das Jahr mit seiner Witterung von Natur aus zufälligen Charakter trägt, kommt es vorrangig auf eine entsprechend große Anzahl Jahre an, um die Witterung repräsentativ zu erfassen (Autorenkollektiv, 1987). Dass wenige Versuchsjahre die klimatische Bandbreite nur sehr unvollkommen bzw. vage repräsentieren können, schlägt sich in den statistischen Maßzahlen dann angemessen nieder, wenn das Jahr als zufällig definiert ist. Der Witterungsverlauf ist in seinem Einfluss auf Pflanzenbestände derart mannigfaltig, dass sich - außer in Extremsituationen - nicht an einigen ausgewählten, oft über lange Zeiträume kumulierten Witterungskennzahlen festmachen lässt, ob die einbezogenen Jahre in ihrer Gesamtheit als repräsentativ oder Einzeljahre als ‚typisch' gelten können. Insofern können die Formulierungen „Witterungsbedingungen, die denen des Versuchsjahres entsprechen" oder „Witterungsverhältnissen, die dem Mittel der Versuchsjahre entsprechen" (Tab. 1) m.E. nur als sehr unscharf bzw. subjektiv angesehen werden.

In jüngerer Literatur gibt eher die Anforderung aus der Zielstellung das Primat für die Hypothesenwahl, sofern die Umweltfaktoren *hinreichend* als repräsentativ angesehen werden können. Smith et al. (2005) plädieren dafür, in der Pflanzenzüchtung und Sortenprüfung Genotypen als zufällig zu definieren, da sich so der ‚Selektionsfehler' minimieren lässt. Auch Van Eeuwijk (2007) unterstützt eine sehr weitgehende Wahl der Hypothese ‚*zufällig*' für den Faktor ‚*Genotyp*', sofern das Interesse nicht vorrangig auf die Ausprägung jedes einzelnen Genotyps (als Individuum), sondern auf das Gesamtsortiment gerichtet ist. Inzwischen ist auch in der Pflanzenzüchtung das Interesse an BLUP für Genotypen gestiegen, da sich die Präzision im Vergleich zu BLUE-Schätzungen erhöht (Piepho et al., 2008). Eher als in der Phase der Züchtung treten bei der amtlichen Sortenprüfung damit allerdings Zielkonflikte auf: aus Sicht der zu beratenden Landwirte kommt es vorrangig darauf an, aus der Gesamtheit der Sorten eine herausragende Leistungspitze zu finden, selbst wenn im Sinne dieses Gesamtzieles Individual-Sorten, die nicht als Teil einer ‚Population' von Sorten angesehen werden sollten, dabei im zufälligen Ansatz durch Schrumpfungsschätzung (siehe Abschnitt 6.4.3) ‚gestört' geschätzt werden könnten. Die den Züchtungsfortschritt hervorbringenden Züchter erwarten dagegen von der offiziellen Sortenprüfung nachvollziehbarer Weise eine Schätzmethodik, die ihre jeweilige Sorte als Individuum, den Faktor Sorte also als fix, ansieht. Eine Schrumpfungsschätzung zum Sortenmittel findet bei ihnen kaum Akzeptanz. Richter et al. (1999) gehen davon aus, dass man üblicherweise die ‚*eigentlichen'* Prüffaktoren als fixe Faktoren behandelt. In diesem Zusammenhang soll festgestellt werden, dass eine Vielzahl der methodischen Arbeiten zur Versuchsserienauswertung aus dem Bereich der Züchtung- und Sortenprüfung stammt, dass hierbei wiederum oft die Arbeit des selektierenden Züchters in den Vordergrund gestellt wird - so auch bei van Eeuwijk (2007). Eine vertiefte Diskussion für den Bereich der amtlichen Sortenprüfung erfolgt im Abschnitt 6.4.

Während im Frühstadium eines züchterischen Selektionsprozesses häufig ein einzelnes Zielmerkmal im Fokus steht, erfolgt die Bewertung von ‚fertigen' Sorten im Sortenprüfsystem grundsätzlich auf Basis einer simultanen Gesamtsicht auf alle wertbestimmenden Eigenschaften (ggf. mittels einer Indexberechnung). Beim Winterweizen wird jede Sorte in der *Beschreibenden Sortenliste* (Bundessortenamt, 2014)

durch 18 im Feld erfasste Merkmale und 12 Qualitätsmerkmale beschrieben. Für jedes dieser 30 Merkmale erfolgte dafür eine univariate Auswertung. Bei der Analyse jedes *Einzelmerkmals* kommt es auf eine möglichst präzise Schätzung / Vorhersage der Sortenunterschiede auf einer stetigen Skala an, damit in der Synthese die *Gesamtbewertung über alle Merkmale* bestmöglich fundiert ist. Sortenrangfolgen werden also erst in der multivariaten Sicht bedeutsam, sind je Einzelmerkmal aber i.d.R. wenig relevant. Ein statistischer Signifikanztest von Sorten gegeneinander ist auf der univariaten Ebene ebenso kaum hilfreich. Des Weiteren sind u.a. auch aus genannten Gründen für die univariaten Auswertungen parametrische Verfahren, welche Informationen auf einer stetigen Skala nutzen und Effekte auf einer stetigen Skala schätzen, prinzipiell zielführender als nichtparametrische Verfahren, welche auf Ranginformationen basieren - solange die Voraussetzungen für parametrische Verfahren hinreichend gegeben sind. In diesem Zusammenhang soll eine treffende Aussage von Stroup (2014) zitiert werden: „… they are focused only on testing, not on estimation. In most plant and soil science research, the question is not, 'Is there a treatment difference?'. Instead, it is, 'We know there is a difference. How big is it?' ".

Balanciertheit

Mudra (1952) geht davon aus, dass Versuche aus beliebig vielen Orten und Jahren zusammengefasst werden können, wenn die Zahl der Versuchsglieder und die Art der Anlage einheitlich sind. Damit deutet er bereits ein sehr häufiges Problem an, dass diese im besten Fall gegebene *Balanciertheit* der Daten nicht zwangsläufig gegeben sein muss - sei es durch Ausfälle von Prüfgliedern oder durch bewusst geplante Lücken z.B. im Zuge der Streichung bzw. Neuaufnahme von Prüfgliedern über die Jahre. 1949 formuliert Mudra es so, dass in der Serienauswertung nur Prüfglieder einbezogen werden können, die in allen Jahren und an allen Orten geprüft wurden, wobei es gleichgültig sei, ob in einzelnen Versuchen weitere Prüfglieder geprüft wurden. Die Abschlussformulierung beinhaltet das noch heute häufig verwendete *Schneiden orthogonaler Kerne* - es werden dabei Versuche oder Prüfglieder in der Weise von der Auswertung ausgeschlossen, dass die verbliebene Datenstruktur orthogonal / balanciert ist. Bei der Versuchsserienauswertung mit dem Softwareprodukt DAVEP in der DDR war das Bilden orthogonaler Kernstrukturen zwingende Voraussetzung für die Auswertung (Franko et al., 1982).

Möhring et al. (2004a) und Piepho und Möhring (2006) raten inzwischen bei Verwendung gemischter Modelle im Sortenversuchswesen vom willkürlichen Ausschluss von Prüfgliedern ab. Hintergrund sind ihre Analysen und Simulationen zur Fragestellung, inwieweit die durch Selektion hervorgerufene Unbalanciertheit zu Verzerrungen bei der Schätzung von Varianzkomponenten und Sorteneffekten führt. Da im Verlauf der Sortenprüfung in jedem Jahr auch Selektion stattfindet, kann von vollständig zufälligen Fehlstellen im mehrjährigen Datensatz nicht ausgegangen werden. Die Selektion findet zwar nicht nur auf Basis des einen gerade in der univariaten Auswertung betrachteten Merkmals statt, sondern auf Basis der Gesamtheit aller wertbestimmenden oder für das Sortenregister relevanten morphologischen u.a. Eigenschaften. Andererseits ist die Selektion häufig auch nicht völlig unabhängig von diesem Einzelmerkmal. Die ‚missing completely at random'-Annahme (MCAR) in der Definition nach Little und Rubin (2002) ist also nicht erfüllt. Allerdings kann die schwächere ‚missing at random'-Annahme (MAR) als erfüllt gelten, wenn zum einen ein separierbares Modell für den Fehlwertmechanismus sowie für die Merkmalsdaten angenommen werden kann und zum anderen die im Selektionsprozess involvierten Daten vollständig in die Auswertung einfließen.

Piepho und Möhring (2006) kommen im Speziellen zu folgenden Aussagen: (1) Ein verzerrender Einfluss der Selektion kann hinreichend vernachlässigt werden, sofern alle im Gesamtzeitraum angefallenen Daten (auch die der herausgenommenen Sorten!) in die Auswertung einfließen. (2) Bei Unbalanciertheit im Zusammenhang mit Selektion sei der REML-Algorithmus dem ML-Algorithmus vorzuziehen und der BLUP-Ansatz (best linear unbiased prediction) sei dem BLUE-Ansatz (best linear unbiased estimation) vorzuziehen. Probleme, die in der Bestimmung von Sorteneffekten durch BLUP - also mit dem zufälligen Ansatz für Sorteneffekte - allerdings liegen können, werden in dieser Arbeit aufgezeigt und diskutiert (s. 6.4).

Die Problematik des Anstrebens absoluter Balanciertheit durch ‚Schneiden' orthogonaler Kerne soll an folgendem Szenarium verdeutlicht werden: wenn in einer langjährigen vielortigen Serie eine Sorte in einem einzelnen Versuch fehlt, so wäre dieser Versuch (oder je nach Modell sogar der Versuchsort über alle Jahre oder das Versuchsjahr über alle Orte) komplett auszuschließen oder alternativ müsste diese Sorte durchgehend aus dem Datensatz entfernt werden. In jedem Fall entstünde ein erheblicher Informationsverlust, welcher auch aus Gründen der wirtschaftlichen Effizienz kaum zu vertreten ist, wenn die Grundgesamtheit das Anbaugebiet und dessen Klima ist.

Insofern waren Auswertungsalgorithmen zur Abschwächung der von Mudra (1952) noch gestellten Balanciertheitsforderung gesucht. Patterson (1978) beschreibt Methoden, die bei unbalancierten Datenstrukturen ‚least squares'-Schätzungen (LS Means) für Sorten gestatten, wobei neben ‚direkten' auch ‚indirekte' Sortenvergleiche ermöglicht werden. Unter direkten Sortenvergleichen werden hier Vergleiche zwischen zwei Sorten verstanden, die in allen einbezogenen Versuchen gemeinsam geprüft wurden. Bei indirekten Vergleichen standen diese Sorten dagegen nicht in allen Versuchen und insbesondere wurden sie nicht in allen Versuchen *gemeinsam* geprüft. Indirekte Vergleiche setzten ‚Drittsorten' voraus, die quasi 'Brücken' zwischen Versuchen bilden, in denen das Sortenpaar nicht gemeinsam stand. Ein Mindestmaß an solchen Brückensorten ist in Sortenprüfsystemen i.d.R. automatisch vorhanden. Da allerdings der Umfang an solchen Brücken die Präzision von indirekten Sortenvergleichen maßgeblich mitbestimmt, sind für Deutschland die bundesweit abgestimmten Verrechnungs- und Vergleichssorten (s.a. 4.3) eine entscheidende Basis für indirekte Sortenvergleiche, wie sie z.B. die Einstufungen in der Beschreibenden Sortenliste (Bundessortenamt, 2014) liefern.

Da häufig mehr als ein zufälliger Effekt in einem linearen (additiven) Modell zu berücksichtigen ist, kommen zunehmend gemischte Modelle (mixed models) zum Einsatz. Die Schätzmethode für Varianzkomponenten nach dem ANOVA-Ansatz (analysis of variance, eingeführt durch Fisher (in Fisher und Mackenzie, 1923)) ist für Serien von Sortenversuchen aufgrund der erheblichen Unbalanciertheit nicht adäquat (Kempton, 1984). Im Fall unbalancierter Daten beeinflussen die Relationen zwischen den Varianzkomponenten der zufälligen Effekte der Modellgleichung z.T. erheblich die Mittelwertschätzung. Nach Patterson und Thompson (1971) ist unter Nutzung gemischter Modelle der Restricted Maximum Likelihood Algorithmus (REML) für die Auswertung unbalancierter Daten geeignet. Die Varianzkomponenten werden, Normalverteilung für zufällige Effekte zugrunde legend (in generalisierten linearen gemischten Modell (GLMM) auch andere Verteilungsfunktionen), mit dem REML - Algorithmus geschätzt (bei mehr als einer Varianzkomponente iterativ). Die REML - Schätzungen für Varianzen und Kovarianzen werden in die Mixed-Model-Gleichungen eingesetzt, wonach nach den unbekannten fixen Parametern aufgelöst und empirische gewichtete LS-Mittelwerte (eWLS-Schätzer; LS-Mittelwerte

(LS Means) = kleinste Quadrat-Mittelwerte) geschätzt werden können. Im Zusammenhang mit modernen Softwarelösungen und leistungsfähiger Rechentechnik können diese Algorithmen nun zunehmend auch sehr große Datenumfänge mit unbalancierten Strukturen und mehreren fixen und zufälligen Effekten in der Modellgleichung verarbeiten. Bei der traditionellen Varianzanalyse unter Nutzung der Algorithmen der ANOVA (*analysis of variance)* erfüllt die Varianzschätzung zwar die Anforderung der Erwartungstreue (REML dagegen nicht absolut), die Mittelwertschätzung ist bei unbalancierten Daten aber fraglich. Robinson (1987) fasst die Literaturdiskussion zum Vergleich des ML-Ansatzes (maximum likelihood) mit REML zusammen. Als Nachteil von ML gegenüber REML stellt sie besonders heraus, dass im balancierten Fall die Varianzen von der ANOVA-Schätzung abweichen.

Van Eeuwijk (2007) gibt auf dem Internationalen Symposium ‚Agricultural Field Trials - Today and Tomorrow' in Stuttgart-Hohenheim einen Überblick zur Umsetzung der REML-Methodologie im Sortenwesen und verweist auf gängige Softwarelösungen für gemischte Modelle (z.B. SAS®, ASREML®, GenStat®).

Silvey (1978) stellt fest, dass die offizielle Sortenprüfung in Großbritannien fast ausnahmslos lückige Tabellen hervorbringt. Sie zeigt, dass der Sortenwechsel sich derart beschleunigte, dass eine ausnahmslos in allen Versuchen mitgeprüfte Kontrollsorte in langjährigen Datensätzen nicht mehr verfügbar ist. Daher wurde in Großbritannien zu dieser Zeit die Mittelwertbildung über Relativzahlen zu einer Kontrollvariante durch die Kleinst-Quadrat-Schätzung mittels der ‚fitting constants' - Methode (s.o.) abgelöst. Diese Problematik der systematischen Lückigkeit ist zum Beispiel bei der Planung der in dieser Arbeit diskutierten Landessortenversuche ein nicht zu umgehendes, dem Prüfsystem immanentes Charakteristikum: Landessortenversuche stellen keine zeitlich begrenzte Versuchsserie dar, sondern werden kontinuierlich über die Jahre fortgeführt. Dabei verlassen naturgemäß Sorten das System, während Neuzulassungen aufgenommen werden. Auch die Verweildauer der Sorten in der Serie ist sehr variabel. Bei der mehrjährigen Auswertung regionaler Sortenversuche müssten nach Analysen des Autors zur Erzwingung balancierter Reststrukturen häufig über 50% der vorhandenen Daten beratungsrelevanter Sorten ungenutzt bleiben. Die Problematik der Unbalanciertheit wird in dieser Arbeit anhand von Datenstrukturen im Sortenprüfsystem im Abschnitt 4.4 vertieft.

Forkman (2013) beschreibt Probleme in der Akzeptanz von verallgemeinerten Kleinst-Quadrat-Schätzungen, wenn der Datensatz unbalanciert ist. Es könne z.B. zur Schätzung einer mittleren Differenz zwischen einer eingeschränkt geprüften Testsorte zu einer durchgängig geprüften Referenzsorte kommen, die außerhalb der in direkten Vergleichen gemessenen Differenzen liege (also nur in Versuchen, die beide Sorten enthalten). Eine derartige Beschreibung einer Sorte sei Akteuren mit besonderem Interesse an dieser konkreten Sorte kaum vermittelbar. Er stellt die von ihm so bezeichnete ‚reference treatment method' vor, die je Testsorte nur die direkten gemessenen Differenzen zu genau einer für alle Vergleiche definierten Referenzsorte auswertet und weitere Ergebnisse unberücksichtigt lässt. Alle Vergleiche von Testsorten untereinander erfolgen indirekt über die Einzelvergleiche zur Referenzsorte. Forkman schränkt ein, dass selbst für die Vergleiche zur Referenzsorte die Testeffizienz leidet, zugunsten der fachlichen Nachvollziehbarkeit für Nichtstatistiker. Insbesondere sinkt aber die Präzision der Vergleiche von Testsorten untereinander. Nach Einschätzung des Autors dieser Arbeit ist die Fokussierung auf Vergleiche zu einer Referenzsorte bzw. auf indirekte Vergleiche von Testsorten über eine Referenzsorte grundsätzlich suboptimal. Oft stellen über lange Zeiträume vorgesehene Referenzsorten im mehrjährigen Kontext nicht mehr den fachlichen Maßstab

für neue Sorten dar, sondern sind methodisch motiviert (Brücke zwischen Versuchen, insbesondere Versuchen in verschiedenen Jahren).

Analyse der Prüfglied × Umwelt - Interaktionen

Nach Kuckuck und Mudra (1950) interessieren in der Züchtung häufig nicht nur Mittelwerte über die Orte bzw. Jahre sondern auch ihr Verhalten an den einzelnen Standorten bzw. in den Jahren.

Die erweiterte Auswertung von Versuchsserien insbesondere hinsichtlich der Prüfglied × Umwelt-Wechselwirkungen im Sortenwesen der DDR wurde maßgeblich durch Bätz (1984) geprägt. Bei Mittelwertvergleichen wurden Sorte × Umwelt-Wechselwirkungen bei der Durchführung der Tests im Kontext der Frage ‚*Wogegen ist zu testen?*' berücksichtigt. Bei Annahme ‚zufällig' für einen Effekt erfolgte ein Signifikanztest für diese Varianzkomponente, die Varianzkomponentenschätzung selbst war eher nachrangig. Die Möglichkeiten der Datenverdichtung, der Mittelwertbildung über Orte und/oder Jahre wurden von diesen Testergebnissen abhängig gemacht, d.h. bei Signifikanz jeweiliger Wechselwirkungen als ‚unzulässig' eingestuft. Obgleich dieser Entscheidungsprozess von Bätz (1984) so dargelegt ist, räumt er doch auch ein, dass auch bei signifikanter Wechselwirkung zusammenfassende Aussagen für die Grundgesamtheit von primärem Interesse sein können.

Dagegen werden in jüngeren Auswertungen mit gemischten Modellen alle im Modell als ‚*zufällig* ' definierten Effekte konsequent als solche behandelt. D.h. insbesondere, dass i.d.R. nicht vorrangig eine Testung dieser Effekte erfolgt, sondern eine Schätzung der Varianzkomponenten. Die Bewertung der Größenordnung der Varianzen und insbesondere ihrer Relationen zueinander nimmt z.B. bei Laidig et al. (2008) wie auch in der vorgelegten Arbeit umfassenden Raum ein. Eine hohe Sorte × Umwelt - Varianz schränkt die Vorhersagegenauigkeit ein bzw. erhöht den Standardfehler von Sortenmittelwerten bzw. Mittelwertdifferenzen. Eine Mittelwertbildung über die Umwelten erfolgt dessen ungeachtet. Allerdings sollte speziell bei hoher Sorte × Ort - Interaktion innerhalb eines Anbaugebietes über die Abgrenzung des Anbaugebietes, seine ausreichende innere Homogenität, sowie über die Repräsentativität der Versuchsorte für dieses Anbaugebiet nachgedacht werden.

Großes Interesse findet die Ausweisung von statistischen Kennziffern der Umweltstabilität / Ökostabilität von Prüfgliedern. Die Verwendung der Ökovalenz wird u.a. durch Wricke (1967) und die der Ökoregression u.a. durch Utz (1972) diskutiert. Eine umfassende methodische Zusammenfassung erfolgte durch Bätz (1984). Er stellt am Beispiel des Merkmals Ertrag im Sortenwesen die Ertragsstabilität als ein besonderes Ziel der Sortenwahl heraus. Piepho (2005) diskutiert diese Stabilitätsmaße als Parameter eines gemischten Modells. Damit können mit Hilfe von Software für gemischte Modelle Stabilitätsanalysen auch bei unbalancierten Daten auf einfachem Wege durchgeführt werden. Mühleisen et al. (2014a) vergleichen unter Nutzung gemischter Modelle die Ertragsstabilität von Hybrid- und Liniensorten bei selbstbefruchtenden Getreidearten. Mühleisen et al. (2014b) fanden für Wertprüfungen mit Wintergerste, dass eine präzise Bewertung der Ertragsstabilität einzelner Sorten eine Prüfung in mindestens 40 Test-Umwelten erfordert.

Die Ökovalenz charakterisiert die Umweltstabilität, z.B. die Ertragssicherheit einer Sorte. In der vergleichenden Sortenbewertung gilt eine Sorte dann als stabil, wenn sich ihre Leistung proportional zum Ertragspotenzial der Umwelt, z.B. ausgedrückt im (ggf. adjustierten) Versuchsmittel, verhält. Dahinter steht das agronomische bzw. dynamische Konzept der Ökostabilität (Thomas, 2006). Dies kommt in vergleichs-

weise stabilen Sorte × Umwelt - Wechselwirkungseffekten bzw. in stabilen Rängen (bzw. bei Darstellung von Relativerträgen in stabilen Relativzahlen) einer Sorte über alle Versuche zum Ausdruck. Das statische oder biologische Konzept der Ökostabilität (Becker, 1981), nach dem eine Sorte unter allen Bedingungen gleiche Absoluterträge (o.ä.) realisiert, ist für die landwirtschaftliche Praxis nicht realistisch - eine derartige Form der absoluten Stabilität, z.B. durch absolute Trockentoleranz, ist kaum denkbar.

Die Ökoregression kann Hinweise darauf geben, warum Sorten u.U. vom agronomischen Konzept der Ertragsstabilität abwichen. Grundlage der Berechnung ist z.B. die sortenspezifische lineare Abhängigkeit der Leistung von den zugehörigen (ggf. adjustierten) Versuchsmitteln (Mittel aller Sorten je Einzelversuch). Sorten mit einem Anstieg über 1,0 weisen in ertragsstarken Versuchen tendenziell überdurchschnittliche Sorte × Umwelt - Wechselwirkungseffekte auf, sie konnten die günstigen Bedingungen überdurchschnittlich verwerten und werden häufig - mit gebotener Vorsicht - als ‚Intensivsorten' klassifiziert. Sorten mit einem Anstieg < 1,0 werden dagegen eher als ‚Extensivsorten' angesehen. Einschränkend muss gerade bezüglich der Begriffe *Extensivsorte* versus *Intensivsorte* betont werden, dass allein der Parameter der Ökoregression noch keine ursächliche Wirkung aufdeckt.

In der Bewertung der Eignung von Sorten für die Anbaugebiete in Mecklenburg-Vorpommern werden erste Ansätze genutzt, um Kennziffern der Umweltstabilität im Zuge der Anwendung der Hohenheim-Gülzower-Serienauswertung zu erzeugen (s. Anlage 9). Um das Risiko von Überinterpretationen zu verringern, wird die Ökovalenz - subjektiv erfahrungsbasiert - nur für Sorten ausgewiesen, von denen mindestens zwanzig Versuchsergebnisse vorliegen. Die Ökoregression wird nur für Sorten ausgewiesen, bei denen sich der Regressionskoeffizient signifikant von ‚1' unterscheidet. Diese Parameter dienen aber ausschließlich der Einbeziehung in die interne Gesamtbewertung. Aufgrund des erhöhten Abstraktionsniveaus werden diese Parameter im Beratungsmaterial für die breite Praxis nicht dargestellt. Die derzeitigen Ansätze erfordern nach Einschätzung des Autors dieser Arbeit aufgrund der Komplexität des Auswertungsmodells weitere methodische Untersuchungen, die nicht Teil dieser Arbeit werden sollen.

Prüfglied-Umwelt-Interaktionen stellen häufig, insbesondere auch im Sortenwesen, Wechselwirkungseffekte zwischen fixen und zufälligen Faktoren dar. Hierbei muss zwischen der *Modellformulierung mit unabhängigen* und der *Modellformulierung mit abhängigen Wechselwirkungen* unterschieden werden (Searle et al., 1992; Richter et al., 1999). Bei unabhängigen Wechselwirkungen wird die Annahme gemacht, dass alle zufälligen Effekte (worin alle Kombinationseffekte zufälliger und fixer Faktoren eingeschlossen sind) unabhängig sind. Bei abhängigen Wechselwirkungen wird die Bedingung aufgenommen, dass die Summe über die zufälligen Wechselwirkungseffekte bei Summation über den Index eines fixen Faktors null ergibt. Damit sind die zufälligen Wechselwirkungseffekte auf verschiedenen Stufen einer fixen und gleichen Stufe eines zufälligen Faktors nicht mehr unabhängig. Beim Test der fixen Effekte gibt es zwischen beiden Ansätzen keinen Unterschied. Basford et al. (2004) weisen für züchterische Parameter darauf hin, dass zwar die Schätzung des Selektionsgewinns unabhängig von der Wahl der Formulierung ist, dass aber die genetische Korrelation und die Heritabilität unterschiedlich ausfällt - sie empfehlen für diese Zwecke das Modell mit abhängigen Wechselwirkungen. Die schätzbaren Funktionen sind bei dem Model mit abhängigen Wechselwirkungen durch den ‚Null-Summen-Ansatz' leichter nachvollziehbar. Die in dieser Arbeit verwendete Software SAS® nutzt aber grundsätzlich das Modell mit unabhängigen Wechselwirkungen. Nur

schätzbare Funktionen, die unabhängig von der Annahme bezüglich der Wechselwirkung sind (Mittelwerte oder Differenzen), stellen interpretierbare Schätzungen dar.

Aus einer Vielzahl von Gesprächen und Diskussionen mit Versuchsanstellern, Sortenprüfern, mit der Versuchsauswertung betrauten Mitarbeitern und Biometrikern hat der Autor dieser Arbeit den Eindruck gewonnen, dass häufig nicht für jeden konkreten Einzelfall eine ergebnisoffene Abwägung z.B. dahingehend stattfindet, wie weit der Aussagebereich definiert werden soll, ob Testung oder Schätzung im Vordergrund stehen, welche Hypothesen für die Umweltfaktoren (z.B. Jahre und Orte) zu wählen sind oder welches Ausmaß der Unbalanciertheit hinnehmbar ist. Vielmehr zeichnen sich hierbei zwei subjektive Grundmuster im methodischen Herangehen ab, zu denen die jeweiligen Bearbeiter eher tendieren:

Die eine, tendenziell eher traditionelle ‚Schule' neigt dazu, überschaubare Datensätze in streng orthogonaler Struktur zu nutzen oder diese künstlich zu bilden. Dies erfolgt selbst dann, wenn wie im hier beschriebenen Fall des Sortenprüfsystems ca. 50% der beratungsrelevanten Daten ‚abgeschnitten' und per Definition nur die letzten drei Versuchsjahre genutzt werden. Letzteres hat dann u.a. wiederum zur Folge, dass Jahre eher wie Stufen eines fixen Faktors betrachtet werden. Ebenso werden Orte eher als fix betrachtet und häufig bereits so gewählt, dass jeder Ort auch innerhalb eines Anbaugebietes eine klar abgegrenzte Standortcharakteristik aufweist. Die Interpretation erfolgt dann vorrangig zurückgerichtet auf die kurze Versuchsperiode bzw. auf die Einzeljahresreaktionen und bezogen auf die konkreten Versuchsstandorte. Verallgemeinerungen für einen übergeordneten Aussagebereich (Anbaugebiet und dessen Klima) erfolgen mit Vorbehalt und sehr vorsichtig. Es besteht oft eine gewisse Scheu, sich von der erinnernden Beschreibung der selbst betreuten Versuche und ihrer sehr konkreten Umstände zu lösen und diese Versuche bzw. Orte und Jahre abstrahierend als eine Stichprobe der Grundgesamtheit einzuordnen. Den Aussagebereich entsprechend den Notwendigkeiten weiter zu fassen und z.B. Empfehlungen für andere Orte im Anbaugebiet in neuen Jahren abzuleiten, erscheint oft zu gewagt. Auch wird Testung häufig wie selbstverständlich als zentrale Aufgabe der biometrischen Auswertung betrachtet.

Das zweite beobachtete und auch in dieser Arbeit favorisierte Herangehen ist eher dadurch gekennzeichnet, die vorliegenden Daten umfassend zu nutzen, die gegebenenfalls systemimmanente Unbalanciertheit hinzunehmen und die immanenten Gegebenheiten des Datensatzes durch Modellwahl, ggf. auch durch Transformation etc. bestmöglich zu berücksichtigen. Die Hypothesenwahl wird vorrangig von der ‚Natur' der Faktoren und dem Ziel der Versuchsanstellung abgeleitet und neigt für die Umweltfaktoren eher zum Ansatz ‚zufällig', da i.d.R. eine Aussage für ein Anbaugebiet und dessen Klima als Zielstellung gesehen wird. Demzufolge werden möglichst viele Jahre und Orte einbezogen und bereits bei der Wahl der Orte wird der Anspruch ‚*repräsentativ*' in den Vordergrund gestellt. Sofern nicht Grundsatzaussagen zur Wirkung von Faktoren und deren Stufen Zielstellung sind, sondern Beratungsaussagen, hat Schätzung im Falle immer wieder auftretender Zielkonflikte Vorrang vor Testung von Prüfgliedeffekten und -differenzen.

Die Modellbildung für hier diskutierte Sortenprüfsysteme wird - ausgehend von einfachsten Ansätzen und nachfolgender sukzessiver Aufnahme von relevanten Besonderheiten des Prüfsystems - im Abschnitt 5.3 dargelegt, wobei durchweg von Zweischrittanalysen und vom zufälligen Ansatz für Umwelten ausgegangen wird.

Die in dieser Gesamtschau auf die Versuchsserienauswertung aufgezeigten Aspekte tangieren alle den Gegenstand und die Zielstellungen (s. Abschnitt 2) der vorgelegten Arbeit bzw. der darin dargelegten methodischen Ansätze der Hohenheim-Gülzower-Serienauswertung. Die Problematik der systematischen Unbalanciertheit in Sortenprüfsystemen wird aufgezeigt und in ihrer Auswirkung auf die Schätzung von Sortenmittelwerten unter unterschiedlichen Szenarien untersucht. Bei Anwendung einer Zweischrittanalyse wird das Ausmaß der Differenziertheit der Präzision der Sortenmittelwerte aus Einzelversuchen gezeigt und Methodik und Nutzen einer Gewichtung mittels der Kehrwerte ihrer Fehlervarianzen diskutiert. Bezüglich der Einhaltung der Modellvoraussetzungen werden Ansätze zur Datentransformation diskutiert und in ihrer Auswirkung auf die Schätzung von Sorteneffekten untersucht. Bezüglich der Sorte $\times$ Umwelt - Wechselwirkungen werden zum einen unterschiedliche Regionalisierungsstrategien untersucht - insbesondere unter dem Aspekt, wie Nachbargebiete mit zu einem Zielgebiet korrelierten Sorteneffekten in die Auswertung für dieses Zielgebiet einfließen können. Zum anderen werden die Varianzkomponenten des Komplexes der Sorte $\times$ Umwelt - Wechselwirkungen für eine breite Palette von Pflanzenarten und Merkmalen in ihrer Bedeutung untersucht und interpretiert. Für das Sortenprüfsystem wird diskutiert, inwieweit eine BLUP-Schätzung, insbesondere der Aspekt der ‚Schrumpfung' für Sorteneffekte geeignet bzw. tragfähig ist. Letztlich wird hinterfragt, ob die mit dem Methodenkomplex der Hohenheim-Gülzower-Serienauswertung vorgenommen Schätzungen von Varianzen und Sorteneffekten hinreichend valide sind.