# Auswertung von landwirtschaftlichen Sortenversuchen mit PROC MIXED – Spagat zwischen Theorie und Praxis

Jens Möhring
Dr. Andreas Büchse
Prof. H.-P. Piepho
Universität Hohenheim, Fachgebiet Bioinformatik
Fruhwirthstr. 23
Stuttgart
moehring@uni-hohenheim.de
buechse@uni-hohenheim.de
piepho@uni-hohenheim.de

**Zusammenfassung**

Die regionale Auswertung von landwirtschaftlichen Sortenversuchen kann mit gemischten Modellen erfolgen. Um die Effizienz der Auswertung im deutschen Sortenprüfwesen zu erhöhen, können vorgelagerte Wertprüfungen und Informationen aus benachbarten Anbaugebieten genutzt werden. Zu diesem Zweck wurde ein entsprechend erweitertes Gemischtes Modell vorgeschlagen, welches eine gewichtete Auswertung über die Anbaugebiete hinweg erlaubt mit dem Ziel, eine möglichst genaue Ertragsschätzung für das jeweilige Zielanbaugebiet zu erhalten. Als Zielkriterium für die Ableitung eines optimalen Schätzverfahrens diente der mittlere quadratische Vorhersagefehler (MSEP = mean squared error of prediction). Dieser erfasst nicht die Varianz eines Schätzers allein, sondern die Summe aus quadrierter Verzerrung und Varianz. Für die Umsetzung in der SAS-Prozedur MIXED wurden die Tests der festen Effekte ausgeschaltet, um die Rechenzeit zu reduzieren. Zudem wurde die Interumweltinformation ignoriert, was eine stark Ressourcen sparende subject-Syntax erlaubt. Als Sortenschätzwert wird ein gewichtetes Mittel verwendet, wobei die Gewichte von der genetischen Korrelation zwischen den Anbaugebieten und somit von der zu schätzenden Varianz-Kovarianz-Matrix der Sorte*Anbaugebiets-Varianz abhängen. Zur Erleichterung der Umsetzung in die Praxis wurden Makros erstellt, welche derzeit in die SAS-basierte Auswertungsoberfläche PIAF-Stat integriert werden.

**Schlüsselworte:** Proc MIXED, Gemischte Modelle, Sortenversuche, subject-Option, Landwirtschaft, regionale Auswertung, MSEP

# 1 Einleitung

In Deutschland gibt es ein neutrales Offizialprüfwesen für landwirtschaftliche Nutzpflanzen, welches über die Zulassung von Sorten entscheidet und Anbauempfehlungen für besonders geeignete Sorten gibt. Dieses Prüfwesen lässt sich in zwei zeitlich aufeinander folgende Abschnitte unterteilen: (1) die für die Zulassung benötigte und vom Bundessortenamt durchgeführte Wertprüfung (WP) sowie Registerprüfung und (2) die von den Länderdienststellen durchgeführten Landessortenversuche (LSV), mit denen regionale Anbauempfehlungen ermöglicht werden. Sowohl Zulassung als auch Anbauempfehlung beruhen auf Prüfergebnissen der Sorten in speziell hierfür angelegten Versuchsserien, die z.B. bei Winterweizen aus einer dreijährigen WP an bundesweit 15-32 Versuchsorten und den sich daran anschließenden 2-3 jährigen LSV mit jeweils etwa 120 Versuchen besteht. Bei der Kulturart Winterweizen wird jeder Versuch als Spaltanlage mit 2 Intensitäten (mit und ohne Fungizidbehandlung) in 2 Wiederholungen angelegt und separat ausgewertet, so dass für jede Sorte pro Versuch und Intensität ein Mittelwert vorliegt. Diese Mittelwerte werden in einem zweiten Schritt einer Serienauswertung unterzogen (Zweischrittanalyse – two-stage analysis, Frensham et al., 1997). Bei anderen Nutzpflanzenarten ist das Prüfwesen in Deutschland ähnlich organisiert.

Der Landwirt erwartet ein effizientes Prüfwesen und eine seinen lokalen/ regionalen Anbaubedingungen angepasste Sortenempfehlung. Eine solche Empfehlung kann über die Zusammenfassung von agrarökologisch, also klimatisch und standörtlich ähnlichen Gebieten zu sogenannten Anbaugebieten und der Auswertung aller Versuche in diesem Anbaugebiet gegeben werden. In Deutschland wurden z.B. für Weizen 23 Anbaugebiete definiert, in denen sich etwa 5 Versuchsorte je Anbaugebiet befinden. Für die Auswertung der Daten in einem Anbaugebiet kann ein gemischtes Modell verwendet werden, dessen allgemeine Form in Matrizenschreibweise in Gl. 1 wiedergegeben ist:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e}\,, \tag{1}$$

wobei $\boldsymbol{y}$ der Beobachtungsvektor, $\boldsymbol{e}$ der Fehlervektor, $\boldsymbol{X}$ und $\boldsymbol{Z}$ Designmatrizen sowie $\boldsymbol{\beta}$ und $\boldsymbol{u}$ die Vektoren der festen und zufälligen Effekte sind. Im betrachteten Fall handelt es sich um ein gemischtes Modell mit den Faktoren Sorte ($s$), Jahr ($j$) und Ort ($o$), wobei die Orte innerhalb des jeweils betrachteten Anbaugebiete, dem Zielanbaugebiet, liegen. In skalarer Schreibweise kann das Modell wie folgt geschrieben werden:

$$y_{ijk} = s_i + j_j + o_k + (sj)_{ij} + (so)_{ik} + (jo)_{jk} + (sjo)_{ijk} + e_{ijk} \quad (2)$$

wobei $s_i$ der Haupteffekt für die $i$-te Sorte ist, $j_j$ der Haupteffekt für das $j$-te Jahr und $o_k$ der Haupteffekt für den $k$-ten Ort. Der Effekt $(sj)_{ij}$ ist beispielsweise die Interaktion der $i$-ten Sorte mit dem $j$-ten Jahr. Die Dreifachinteraktion $(sjo)_{ijk}$ der $i$-ten Sorte im $j$-ten Jahr am $k$-ten Ort ist, da bei der Auswertung Mittelwerte verrechnet werden (Zweischrittanalyse), mit dem Restfehler $e_{ijk}$ vermengt und nicht separat schätzbar. Die Dreifachinteraktion kann jedoch geschätzt werden, wenn die Standardfehler für die Versuche vorhanden sind. Dann kann eine gewichtete Analyse durchgeführt werden, die eine Trennung erlaubt (Piepho und Michel, 2001).

Die bei einer solchen Analyse erhaltenen Sortenschätzwerte besitzen allerdings auf Grund der hohen Genotyp*Umwelt-Interaktion einen relativ großen Fehler, da die Schätzung auf nur wenigen Versuchen eines Anbaugebietes und nur auf Versuchsergebnissen aus zwei oder drei Jahren beruhen. Deshalb ist es sinnvoll, weitere Versuche in die Auswertung einzubeziehen. Um den Schätzfehler zu reduzieren, ist es insbesondere von Vorteil, zeitlich vorgelagerte WP in der Auswertung zu berücksichtigen, denn diese liefern Informationen für weitere Prüfjahre. Zusätzlich sollten Versuche aus Nachbaranbaugebieten einbezogen werden, da diese ebenfalls Informationen über die Sortenleistung im Zielanbaugebiet liefern. Der Informationsgewinn ist um so höher, je ähnlicher sich die Anbaugebiete sind, also um so höher die genetische Korrelation ist. Im vorliegenden Beitrag wird ein Verfahren vorgestellt, mit dem Versuche verschiedener Anbaugebiete und zeitlich versetzter Serien (WP und LSV) integriert werden können. Dabei erhalten Daten aus Nachbargebieten ein geringeres Gewicht als die Versuche im Zielanbaugebiet, werden aber wie die WP ebenfalls zur Auswertung herangezogen. Grundlage der gewichteten Schätzung ist ein gemischtes Modell, wie im folgenden erläutert wird.

# 2 Material und Methoden

Die Auswertung aller Versuche eines Zielanbaugebietes mit dem oben beschriebenen Modell führt zu einer unverzerrten Schätzung der Sortenleistung im Zielanbaugebiet. Deren Varianz (Prüfgenauigkeit) ergibt sich aus der Anzahl Versuche je Anbaugebiet. Wird die Datenbasis auf mehrere Anbaugebiete erweitert, so stehen für die Auswertung mehr Versuche aus mehr Jahren zur Verfügung und entsprechend wird die Varianz kleiner. Weil sich die Sortenleistung aus anderen Anbaugebieten aber

systematisch von denen des Zielanbaugebiete unterscheiden, müssen diese Informationenquellen gewichtet werden. Eine solche gewichtete Schätzung kann verzerrt sein, besitzt jedoch auf Grund der breiteren Datenbasis eine kleinere Varianz. Im Extrem werden alle Anbaugebiete gemeinsam ausgewertet, wobei regionale Unterschiede ignoriert werden und zu einer systematischen Abweichung führen.

Das oben vorgestellte Auswertungsverfahren für ein Zielanbaugebiet minimiert die Varianz für eine unverzerrte Schätzung. Im Gegensatz dazu soll im nachfolgend vorgestellten Ansatz die erwartete quadrierte Abweichung (MSEP = mean squared error of prediction) minimiert werden, also der Summe aus quadrierter Verzerrung und Varianz. Bei der Hinzunahme weiterer Anbaugebiete in die Auswertung gibt es gegenläufige Trends. Je mehr Versuche ausgewertet werden, desto kleiner ist die Varianz, aber desto größer ist die zu erwartende Verzerrung. Da die Reduktion der Varianz mit zunehmender Versuchszahl immer weiter abnimmt (abnehmender Grenznutzen), gibt es ein Optimum, an dem der MSEP minimiert wird. Je nach Größe der zu erwartenden Verzerrung und der Reduktion der Varianz durch mehr Versuche kann also die Hinzunahme weiterer Anbaugebiete den MSEP reduzieren oder nicht. Bei einer Reduktion ist es effizient, diese zusätzlichen Anbaugebiete in die Auswertung einzubeziehen. Nimmt man die Sorten als zufällige Stichprobe einer Grundgesamtheit aller möglichen, von den Züchtern anmeldbaren Sorten, so wird der MSEP bei Schätzung von BLUPs (best linear unbiased predictor) für die Sorten automatisch minimiert (Piepho und Möhring, 2005). Bei allen im folgenden betrachteten Modellen wird deshalb die Sorte als zufällig angenommen. Die jeweiligen Sortenschätzwerte ergeben sich durch Addition des Regionenmittels.

Um die vorgelagerten WP-Ergebnisse zu nutzen, muss ein geeignetes Modell angepasst werden. Auf Grund der Struktur des Sortenprüfwesens, also der Anlage mehrerer Versuchsserien, kann es vorkommen, dass eine Sorte in einem Jahr an einem Ort mehrmals geprüft wird (z.B. in WP und LSV). Diese Unterschiede (z.B. der Unterschied zwischen den Versuchsserien WP und LSV) werden über einen Blockungsfaktor Versuchstyp ($t$), der in Jahr*Ort geschachtelt ist, modelliert. Es ergibt sich folgendes Modell

$$y_{ijkl} = s_i + j_j + o_k + (sj)_{ij} + (so)_{ik} + (jo)_{jk} + (sjo)_{ijk} + (tjo)_{l(jk)} + e_{ijkl} \qquad (3)$$

Die Dreifachinteraktion $(sjo)_{ijk}$ ist nun vom Restfehler trennbar und verbleibt somit explizit im Modell. Bei der Erweiterung der Daten um die Versuche aus anderen An-

baugebieten muss deren Herkunft berücksichtigt werden. Die Orte eines fest definierten Anbaugebietes werden als Zufallsstichprobe für dieses Anbaugebiet betrachtet, Unterschiede zwischen Anbaugebieten werden über einen fixen Faktor für das Anbaugebiet (*r*) modelliert.

$$y_{ijklm} = s_i + j_j + r_m + o(r)_{k(m)} + (sj)_{ij} + (sr)_{im} + (sor)_{ik(m)} + (jr)_{jm} + (jor)_{ik(m)} \quad (4)$$
$$+ (sjr)_{ijm} + (sjor)_{ijk(m)} + (tjor)_{l(ikm)} + e_{ijkml}$$

Durch die gemeinsame Auswertung von WP und LSV entsteht ein stark unbalancierter Datensatz. Da die SAS-Prozedur MIXED unbalancierte Daten handhaben kann, wurde diese zur Auswertung herangezogen (SAS-Institute, Inc., Cary, NC, USA). Der entsprechende SAS-Code lautet:

```
Proc mixed Data=Datei;
class S J R O Typ;
model y=R /ddfm=kr;
random S J O(R) J*R J*O(R) S*J S*R S*O(R) S*J*R S*J*O(R) Typ(J*O);
run;
```

In der Modell-Anweisung finden sich die festen Effekte, in der random-Anweisung die zufälligen. Die Default-Einstellung für die Freiheitsgradapproximation wurde von `ddfm=contain` (Containment) auf `ddfm=kr` (Kenward-Roger) umgestellt, da diese Approximation adäquate Freiheitsgrade berechnet und zudem berücksichtigt, dass die Varianzkomponenten ebenfalls Schätzwerte sind und somit einen Fehler besitzen (Kenward und Roger, 1997). Allerdings kommt es bei realen Datensätzen zu einem Ressourcenproblem (Speicherplatz, out of memory), d.h. eine Auswertung mit oben angegebenem Programmcode ist in SAS derzeit nicht standardmäßig möglich.

# 3 Ergebnisse/ Lösungsweg

Als ersten Schritt zur Lösung der Ressourcenprobleme kann man in der Prozedur MIXED die Tests der festen Effekte, hier die Tests auf Unterschiede zwischen den Anbaugebieten, ausschalten. Ziel der Versuchsserien ist die Bewertung der Sorten. Ein Test ob sich Anbaugebiete im durchschnittlichen Ertragsniveau unterscheiden ist hierbei uninteressant. Somit wird auch eine Approximation der Freiheitsgrade nicht mehr benötigt, wobei sich die Bestimmung der Freiheitsgrade nicht ganz abschalten lässt, da sie bei der Berechnung von Mittelwerten und Effektschätzungen per Voreinstellung zur Anwendung kommt. Allerdings gibt es deutlich schnellere Methoden zur

Freiheitsgradbestimmung, z.B. die sogenannte Residual-Methode. Hier werden bei allen Tests die Freiheitsgrade des Restfehlers verwendet, welche sich als $n$ - rank($\boldsymbol{X}$) berechnen lässt, wobei $n$ die Anzahl der Beobachtungen und $\boldsymbol{X}$ die Designmatrix der festen Effekte ist. Auch diese Option kann somit ohne Informationsverlust in Bezug auf die Sortenschätzwerte geändert werden. Verzichtet man des weiteren auf die Nutzung der Interumweltinformation (analog zur Interblockinformation in Versuchsanlagen mit unvollständigen Blöcken, siehe Kempton und Fox, 1997, S.60), so kann man die Umwelteffekte und den Blockfaktor als fix nehmen. Die Interumweltinformation ist relativ gering, da die Varianz der reinen Umwelt-Effekte (Jahr, Ort und Jahr*Ort) relativ groß ist.

Ein solches Vorgehen hat den Vorteil, dass nun nur noch mit Sorte gekreuzte Effekte zufällig sind und eine `subject`-Schreibweise verwendet werden kann. Diese reduziert den Rechenbedarf erheblich, da nun statt der gesamten Varianz-Kovarianz-Matrix der zufälligen Effekte (G) nur noch die Varianz-Kovarianz-Matrizen für jede Sorte einzeln invertiert werden müssen. Der entsprechende Aufruf in SAS sieht wie folgt aus:

```
Proc mixed Data=Datei;
class S J R O Typ;
model y=J R O(R) J*R J*O(R) Typ(J*O)/ notest ddfm=residual;
random int J R O(R) J*R J*O(R)/subject=S;
run;
```

Dieser Programmaufruf umgeht das oben beschriebene Ressourcenproblem und führt, natürlich in Abhängigkeit von der Datensatzgröße, innerhalb einer vertretbaren Zeit zur Konvergenz und damit zu BLUP-Schätzwerten für die Sortenleistung.

Dem hier verwendeten gemischten Modell liegen drei Annahmen zu Grunde. Diese sind Additivität, Normalverteilung und Varianzhomogenität. Zur Erzielung der ersten beiden Annahmen kann eine Transformation erfolgen, worauf hier aber nicht näher eingegangen werden soll. Die Annahme der Varianzhomogenität ist aus fachlichen Gründen insbesondere für die Sorte*Region-Varianz fraglich. Betrachtet man zunächst die genetische Korrelation ($r_G$) zwischen dem Zielanbaugebiet und einem seiner Nachbaranbaugebiete (Gl. 5), so erkennt man, dass diese eine Funktion der Sortenvarianz und der Sorte*Region-Varianz ist:

$$r_G = \frac{\mathrm{cov}(G_{i1}, G_{i2})}{\sqrt{\mathrm{var}(G_{i1})\,\mathrm{var}(G_{i2})}} = \frac{\sigma_S^2}{\sqrt{(\sigma_S^2 + \sigma_{SR_1}^2)(\sigma_S^2 + \sigma_{SR_2}^2)}} \quad . \tag{5}$$

$G_{i1}$ ist hierbei die *i*-te Sorte in Region 1, $\sigma_S^2$ ist die Sortenvarianz und $\sigma_{SR_1}^2$ die Sorte*Region-Varianz in Anbaugebiet 1. Es ist zu erwarten, dass nicht alle Nachbaranbaugebiete dem Zielanbaugebiet gleichermaßen ähnlich sind. Dies kann, wie Gl. 5 zeigt, durch eine Regionen spezifische Sorte*Region-Varianz berücksichtigt werden. Alternativ könnte man beispielsweise auch eine völlig unstrukturierte Varianzstruktur annehmen (Piepho, 1998). Ein Vergleich dieser Modelle mittels Likelihood-Quotienten-Test führt bei großen Datensätzen fast immer zu dem Ergebnis, dass komplexe Strukturen die Daten besser beschreiben als einfachere. Allerdings kommt es insbesondere bei unstrukturierter Varianz-Kovarianz-Matrix ohne gute Startwerte oft zu Konvergenzproblemen. Zusätzlich steigt mit zunehmender Komplexität auch die Rechenzeit, wie Tab. 1 exemplarisch für die Sorte*Ort-Varianz zeigt.

**Tab.1:** Rechenzeit und Güte verschiedener Modellerweiterungen der Sorte*Ort-Varianz mit der SAS-Prozedur MIXED

| Varianzstruktur (S*O) § | Vergleich zur Standardeinstellung (VC) | | | |
|---|---|---|---|---|
| | Rechenzeit (in h:min) | Anzahl zusätzlicher Varianzparameter für S*O | 2*ΔLog-Likelihood | $\chi^2_{FG;0,95}$-Wert |
| VC=default | 0:15 | 0 | 0 | - |
| FA1(1) | 3:16 | 37 | 61,8 | 52,19 |
| FA(1) | 12:10 | 74 | 120,7 | 95,08 |

§ VC: einfache Varianzkomponenten. FA1(1)/FA(1): Faktoranalytisches Modell mit einem Faktor und homogener/heterogener Restvarianz (Piepho, 1998)

Beides spricht für einen Kompromiss zwischen dem statistisch wünschenswerten und dem in der Praxis rechenbaren, weshalb wir uns bei großen Datensätzen für eine Matrix mit homogener Kovarianz und heterogenen Varianzen entschieden haben. Diese lässt sich in SAS mit folgendem Programmcode umsetzen:

```
Proc mixed Data=Datei;
class S J R O Typ;
model y=J R O(R) J*R J*O(R) Typ(J*O)/ notest ddfm=residual;
random int J O(R) J*R J*O(R)/subject=S;
random R/subject=S type=UN(1);
run;
```

Tatsächlich werden, um die Regionenmittelwerte im Rahmen der LSMEANS-Anweisung einfach schätzbar zu machen, die festen Jahres- und Ortseffekte sowie die Interaktion Jahr*Region aus dem Modell eliminiert, da diese im Effekt Jahr*Ort(Region) subsummiert sind und dann aus den unbalancierten Daten nur noch die Effekte Jahr*Ort(Region) geschätzt werden müssen. Ohne diesen Kunstgriff wären die Kleinstquadratmittelwerte nur durch die Definition geeigneter Koeffizienten mittels der OBSMARGINS-Option der LSMEANS-Anweisung zu erhalten (Oliver Schabenberger, SAS Institute, pers. Mitt.).

# 4 Diskussion

Die Annahme zufälliger Sorten sowie die damit verbundene Schrumpfung der BLUP ist für die Auswertung von Sortenversuchen ungewöhnlich. Im Gegensatz zur hier gemachten Annahme einer Zufallsstichprobe können Sorten auch als gezielt ausgewählte und durch ihre Eigenschaften und ihre Genetik definierte Stufen des Faktors Sorte betrachtet werden, womit sie dann fest wären. Auch hierfür lassen sich mit zuvor bestimmten Gewichten Schätzwerte ermitteln. Diese sind dann nicht geschrumpft (unpublizierte Resultate). Zur Berechnung der optimalen Gewichte selbst wird jedoch auch hier mit zufälligen Sorten gerechnet, da die genetische Varianz-Kovarianz-Struktur für die Berechnung des Zielkriteriums MSEP weiterhin benötigt wird.

Alternativ zu SAS gibt es speziell für gemischte Modelle Programme (z.B. ASReml, Gilmour et al., 1999), die zum Teil deutlich schneller sind. In der vorliegenden Arbeit wurde jedoch bewusst SAS verwendet, da dieses in den Länderdienststellen über PIAF-Stat bereits verfügbar ist. PIAF-Stat ist ein auf dem Planungs- und Erfassungssystem PIAF (Planung, Information und Auswertung von Feldversuchen) aufbauendes und auf SAS beruhendes System zur Auswertung von Feldversuchen. Deshalb ist die Entwicklung effizienter Auswertungsroutinen auf Basis von PROC MIXED von vorrangiger Bedeutung für das deutsche Sortenprüfwesen. Um die Umsetzung des oben beschrieben Gemischten Modells in PIAF-Stat-Verfahren zu erleichtern, wurden von uns Makros entwickelt. Diese beziehen also WP und LSV-Daten aus ver-

schiedenen Anbaugebieten ein und minimieren den MSEP. In der Ausgabe werden gewichtete Mittelwerte als Sortenschätzwert angegeben.

## Literatur

[1] Frensham, A., B. Cullis, A. Verbyla. 1997. Genotype by environment variance heterogeneity in a two-stage analysis. Biometrics 53, 1373-1383.

[2] Gilmour, A.R., B.R. Cullis, S.J. Welham, R. Thompson. 1999. ASREML reference manual. NSW Agriculture Biometric Bulletin No.3. NSW Agriculture, Locked Bag 21, Orange NSW, 2800, Australia, 210ff.

[3] Kenward, M.G., J.H. Roger. 1997. Small sample inference for fixed effects from restricted maximum likelihood. Biometrics 53, 983-997.

[4] Kempton, R.A., P.N. Fox. 1997. Statistical methods for plant variety evaluation. Chapman & Hall, London.

[5] Piepho, H.-P. 1998. Empirical best linear unbiased prediction in cultivar trials using factor analytic variance-covariance structures. Theoretical and Applied Genetics 97, 195-201.

[6] Piepho, H.-P., V. Michel. 2001. Überlegungen zur regionalen Auswertung von Landessortenversuchen. Informatik, Biometrie und Epidemiologie in Medizin und Biologie 31, 123-139.

[7] Piepho, H.-P., J. Möhring. 2005. Best linear unbiased prediction of cultivar effects for subdivided target regions. Crop Science 45 (in press)

[8] SAS Institute, Inc., SAS/STAT *User´s Guide, Version 8*, Cary, NC, USA 1999.