Proc MIXED in SAS 9.2 – hat sich die Auswertung von mehrjährigen Serien landwirtschaftlicher Sortenversuche verbessert?

Andrea Zenk
Landesforschungsanstalt für
Landwirtschaft und Fischerei M-V
Dorfplatz 1
18276 Gülzow
a.zenk@lfa.mvnet.de

Christian Marschall Volker Michel
Landesforschungsanstalt für Landwirtschaft und Fischerei M-V
Dorfplatz 1
18276 Gülzow
c.marschall@lfa.mvnet.de v.michel@lfa.mvnet.de

Zusammenfassung

Eine Schwerpunktaufgabe der Landesforschungsanstalt für Landwirtschaft und Fischerei Mecklenburg-Vorpommern ist die Erarbeitung von regionalen Sortenempfehlungen für die Praxis (www.lfamv.de). Grundlage dafür ist die frühzeitige und schätzgenaue regionale Ertragseinstufung von Sorten, die durch mehrjährige Auswertung von Landessortenversuchen und anderen Offizialprüfungen wie den Wertprüfungen des Bundessortenamtes erarbeitet wird. Seit Ernte 2005 wurden dazu biometrische Verfahren der von der Universität Hohenheim und der Landesforschungsanstalt für Landwirtschaft und Fischerei Mecklenburg-Vorpommern entwickelten ‚Hohenheim-Gülzower Serienauswertung' genutzt, die auf der 9. KSFE in zwei Vorträgen vorgestellt wurden [3,7].
Eine zentrale Funktion in den Verfahren besitzt die SAS-Prozedur MIXED. Trotz der schon im Modell vorgenommenen Möglichkeiten der Performance-Steigerung [3] kann es bei sehr großen Datenmengen und komplexer sowie stark unbalancierter Datenstruktur zu Problemen hinsichtlich Speicherplatzbedarf und zu erheblichen Laufzeiten kommen.
Der Vortrag gibt einen Überblick über die Erfahrungen mit den Verfahren der ‚Hohenheim-Gülzower Serienauswertung' und der SAS-Prozedur MIXED in Mecklenburg-Vorpommern. Dabei wird auf Veränderungen der PROC MIXED in SAS 9.2 eingegangen. Es werden Möglichkeiten zur Vermeidung bzw. Verringerung der Probleme diskutiert.

Schlüsselwörter: Proc MIXED, SAS 9.2, Sortenversuche, Landwirtschaft

1 Einleitung

Das Sachgebiet Sortenwesen der Landesforschungsanstalt für Landwirtschaft und Fischerei Mecklenburg-Vorpommern (LFA) erstellt jedes Jahr Sortenempfehlungen für ca. 12 Kulturarten. Diese Empfehlungen unterstützen die Landwirte von Mecklenburg-

Vorpommern in der Sortenwahl für den Anbau in der Praxis. Landessortenversuche sind die Grundlage dieser regionalen Sortenempfehlungen. Ihre mehrjährige Auswertung unter Einbeziehung von Wertprüfungen (WP) des BSA und anderer Offizialprüfungen erfolgt seit 2005 in Mecklenburg-Vorpommern mit der Hohenheim-Gülzower Serienauswertung [3,7].
Die Hohenheim-Gülzower Serienauswertung wurde in Kooperation der Universität Hohenheim und der LFA in 3 Verfahren realisiert (Tabelle 1). Dabei haben die Verfahren ‚PHI' und ‚VK' eine vorbereitende Funktion für die Optimierung der Schätzung der Sortenleistungen. Das Verfahren ‚MW' erfüllt die zentrale Aufgabe des Verfahrenskomplexes – die optimierte Schätzung der Sortenleistungen und deren Schätzgenauigkeit.

Tabelle 1: Die Verfahren der Hohenheim-Gülzower-Serienauswertung

Abarbeitung	Kurzbezeichnung	Langbezeichnung
1	PHI	Optimale Datentransformation
2	VK	Bestimmung der Varianzkomponenten
3	MW	Bestimmung der Mittelwerte

- Wichtigste Prozedur ist die SAS-Prozedur MIXED. Schon bei der Erarbeitung der Verfahren war bekannt, dass die Prozedur MIXED bei sehr großen Datenmengen und stark unbalancierter Datenstruktur zu Speicherplatz-Problemen und erheblichen Laufzeiten führt. Diese Probleme zu minimieren, war ein wichtiges Ziel bei der Entwicklung der Verfahren. So wurden in den Modellen Restriktionen festgelegt, die die Lauffähigkeit der Verfahren überhaupt ermöglichten.
- Jedoch zeigten die Erfahrungen der ersten Jahre, dass an diesen Problemen weiter zu arbeiten ist.

2 Datengrundlage

2.1 Umfang der jährlichen Auswertungen in MV

Jährlich kommen bei 12 Kulturarten jeweils ein bis sechs Merkmale zur Auswertung. Im Jahr 2009 wurden je 31 Auswertungen mit den Verfahren PHI und VK und 54 Auswertungen mehrjähriger Datensätze mit dem Verfahren MW durchgeführt.
Für die Berechnung des optimalen Transformationsparameters Phi und der Varianzkomponenten wurden dabei jeweils neunjährige Datensätze genutzt. Diese Auswertungen können ausserhalb der Saison, also nicht unmittelbar nach der Ernte, stattfinden. Da durch die Langjährigkeit der Daten der Parameter Phi und die Varianzkomponenten eine gute Stabilität erhalten, werden diese Berechnungen nur alle drei Jahre durchgeführt.
Beim Verfahren MW zur Berechnung der Mittelwerte wurde der Datensatz auf die sechs bzw. sieben letzten Jahre reduziert.

Die zu verarbeitende Datenmenge variierte sehr stark. Größte Datenmengen mit ca. 5.000 bis 10.000 Beobachtungen lagen bei Winterweizen, Winterraps, Wintergerste, Winterroggen, Wintertriticale und Silomais vor, also bei den 6 der 12 Kulturarten, die auch die größte Bedeutung im Praxisanbau besitzen.

2.2 Beschreibung der Daten

Eine beispielhafte Beschreibung der Daten, die zur Auswertung vorliegen, soll am aktuellen Datensatz zu Winterweizen erfolgen.

Tabelle 2: Beschreibung des Datensatzes ‚Winterweizen'

Faktor	Anzahl Klassen	Beschreibung
Jahr	9	2001-2009
Ort	39	...
Region	5	3 4 5 45 82
Sortengruppe	9	A B B/C BH C CH CK E -
Sorte	1121	...
Versuchstyp	4	…

Analyse-Merkmal: Ertrag in dt/ha
Beobachtungen insgesamt: 9964

In den Modellen (siehe unten) sind als Faktoren Sorte (s), Jahr, Ort, Region (r) und Versuchstyp (typ) berücksichtigt. Zusätzlich kann der Faktor Sorte in Sortengruppe (gr) geschachtelt sein, wobei unter Sortengruppe eine Einteilung der Sorten in distinkte Klassen zu verstehen ist (Qualitätsklassen bei Winterweizen).
Der Faktor Versuchstyp ist ein Blockungsfaktor, der verschiedene Versuche am selben Ort im selben Jahr (z. B. Landessortenversuch und Wertprüfung) unterscheidet.
Für die Sortenberatung wurde Deutschland in Regionen (Anbaugebiete) eingeteilt. Sie unterscheiden sich durch standortkundliche Gegebenheiten und deren Relevanz für sortenspezifische Reaktionen und basieren auf Boden-Klima-Räumen [5]. Jeder Ort wird genau einem Anbaugebiet zugeordnet.
Ist einer der Faktoren nur mit einer Klasse belegt, werden die Modelle (siehe unten) automatisch angepaßt, d. h. um diesen Faktor reduziert.

Die mehrjährigen Daten der Sortenversuche sind in vielerlei Hinsicht unbalanciert. So gibt es in jeder mehrjährigen Serie Sorten, die nur in einigen Jahren vorkommen. Gleiches trifft auch für Orte und Versuchstypen zu. Die Unbalanciertheit der Daten ist extrem. Wären alle 1121 Sorten in den 9 Jahren an den 39 Orten geprüft worden, dann hätte der vorgestellte Datensatz zu Winterweizen **393.471** Beobachtungen.

3 Zu den Verfahren

Auf Grundlage der Arbeiten von Michel et al. [1,2] wurden an der Universität Hohenheim Modelle für die optimale Auswertung von Sortenversuchen entwickelt. Diese Modelle wurden von Möhring in den SAS-Makros %boxcox, %varianzkomponenten und %lsm umgesetzt [3]. Dabei wurden alle Möglichkeiten zur Minderung der Rechenzeit und des Speicherbedarfs in der Prozedur MIXED ausgeschöpft [6]. Dazu gehören die Optionen notest und DDFM=residual in der Model-Anweisung, die Verwendung von Subjects (subject=s) sowie die Reduktion der festen Umwelt- und Blockeffekte auf die Interaktionen der größten Ordnung.
Die Makros sind die Basis der PIAFStat-Verfahren PHI, VK und MW. Die Aufsplittung in drei Verfahren wurde gewählt, da dies eine Optimierung der Performance für jeden Teilschritt ermöglichte.

Verfahren PHI
Das Verfahren PHI mit dem SAS-Makro %boxcox enthält folgende Prozedur MIXED:

```
Proc mixed data=a1 method=ml lognote;
class gr s jahr r ort typ;
model Ertrag=gr jahr r ort jahr*r jahr*ort jahr*ort*typ gr*jahr gr*r
gr*ort gr*jahr*r gr*jahr*ort /ddfm=residual;
random int jahr ort jahr*r jahr*ort /subject=s;
random r/ subject=s type=UN(1);
parms /parmsdata=dat_parms;
run;
```

Besonderheit: Diese Prozedur wird im Makro mehrmals aufgerufen, da zur Bestimmung des optimalen Transformationsparameters PHI ein iteratives Verfahren verwendet wird und für jeden Schritt die Prozedur ablaufen muss. In der Konsequenz entstehen extrem hohe Laufzeiten.

Verfahren VK
Das Makro %varianzkomponenten wurde für die Varianzkomponentenschätzung optimiert. Die entscheidende Prozedur MIXED lautet:

```
Proc mixed sigiter lognote asycov cl alpha=0.05;
class  gr s jahr r ort typ;
model Ertrag= gr*r gr*Jahr*r gr*jahr*r*ort typ(r*jahr*ort)
/ddfm=residual notest;
random int jahr ort jahr*r jahr*ort /subject=s;
random r/ type=UN(1) subject=s solution;
run;
```

Verfahren MW
Im Makro %lsm sind zwei MIXED-Prozeduren enthalten:

```
*Bestimme die umweltbedinge Varianz von Sortenvergleichen;
Proc mixed Data=x_daten2 sigiter lognote noinfo;
by s;
class=gr r s jahr ort typ;
model Ertrag= gr*r s(gr*r) jahr*ort(r) typ(jahr*ort*r)/
ddfm=residual noint notest;
random jahr ort jahr*r jahr*ort ;
parms/ parmsdata=x_var1 noiter;
*Estimate-Statements;
%estimate_LSM(&anzahl_r);
run;

*Bestimme lsmeans gr*r und s(gr*r);
ods output lsmeans=x_lsm;
Proc mixed Data=x_daten sigiter lognote noprofile;
class=gr r s jahr ort typ;
model Ertrag= gr*r s(gr*r) jahr*ort(r) typ(jahr*ort*r)/
ddfm=residual noint notest;
random jahr ort jahr*r jahr*ort ;
parms/ parmsdata=x_var1 noiter;
lsmeans gr*r;
lsmeans s(gr*r);
lsmeans typ(jahr*ort*r);
run;
```

Im Verfahren MW ist es die zweite MIXED-Prozedur, bei der der Fehler ‚Out off memory' auftrat.
Mit den erstellten Makros und den damit entwickelten Verfahren PHI, VK und MW wurde eine routinemäßige gemeinsame, gewichtete Auswertung von Sortenversuchen innerhalb der PIAF-Stat-Umgebung möglich.

- Veränderungen an den Modellen in den Verfahren stehen zurzeit nicht zur Disposition. Deshalb wurde das Augenmerk auf andere Möglichkeiten der Optimierung der Verfahren gelegt.

4 Praxis-Erfahrungen

Die hier aufgeführten Auswertungen wurden an einem PC Intel Pentium 4 CPU mit 3 GHz und 2 GB RAM vorgenommen. Dieser fünf Jahre alte PC diente in den letzten Jahren als Arbeitsplatz-PC für die Arbeit mit der Hohenheim-Gülzower Serienauswertung.
Trotz der Optimierungen in den Makros traten in den letzten Jahren bei den Auswertungen mit sehr großen Datenmengen und stark unbalancierter Datenstruktur Probleme hinsichtlich Speicherplatzbedarf und Rechenzeit auf.

1. Problem: Der Speicherplatzbedarf - Fehlermeldung: out of memory

Bei sehr großen, sehr unbalancierten Datensätzen ist es immer wieder zum Abbruch insbesondere des Verfahren MW gekommen.

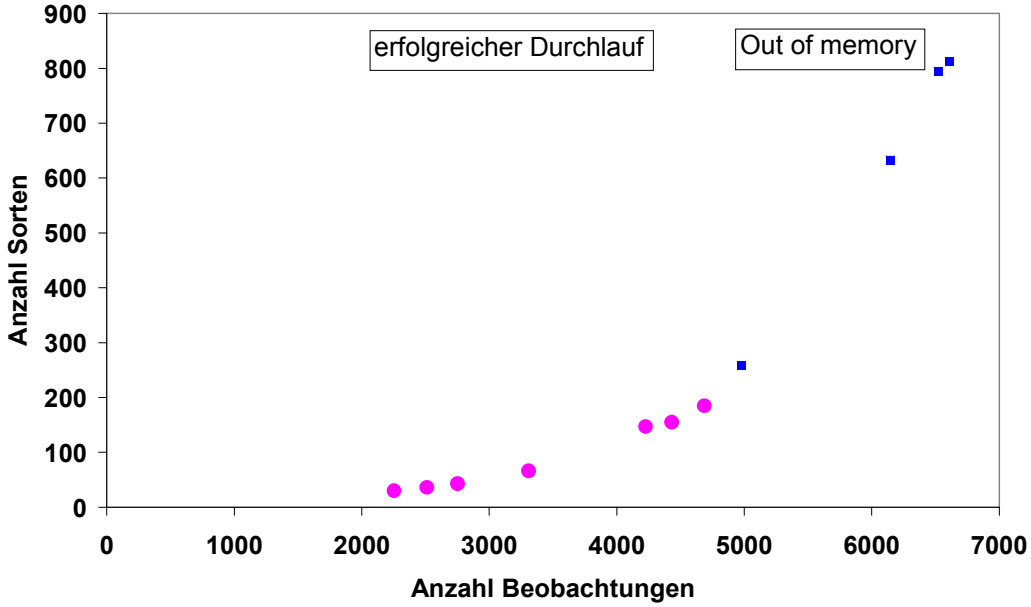

Abbildung 1: Grenze für Fehler ‚Out of memory' (Verfahren MW)

Um trotzdem zu Ergebnissen zu kommen, muss eine Reduzierung der Daten vorgenommen werden. Dazu wurden in den Verfahren Möglichkeiten der Datenreduktion integriert, die vom Anwender interaktiv festgelegt werden können. Insbesondere durch das Festlegen einer Mindestanzahl Versuche, in denen eine Sorte geprüft worden sein soll, kann die Unbalanciertheit der Daten verringert und ein Abbruch der Verfahren vermieden werden. Leider gehen dabei Informationen über z. B. sehr junge, noch selten geprüfte Sorten verloren. Bei Winterweizen mussten so im Verfahren MW alle Sorten aus dem Datensatz gelöscht werden, die insgesamt weniger als achtmal im Datensatz vorhanden waren.

2. Problem: Die Rechenzeit

Tabelle 3: Auswertungen 2009

Verfahren	Zeitpunkt	Kulturart	Datensätze	cpu time (h:min:s)
PHI	Januar 2009	Silomais	4608	10:02:02
VK	Januar 2009	Silomais	4608	1:04:13
MW	Juli-Dezember 2009	Winterraps	6340	1:31:18

Insbesondere das Verfahren PHI benötigte extrem lange Rechenzeiten. Aber auch die Zeiten für das Verfahren MW zu reduzieren, wäre sehr wichtig, da für diese Berechnungen nur ein schmales Zeitfenster von der Ernte bis zur Veröffentlichung der Sortenempfehlung zur Verfügung steht.

5 Verbesserungen durch SAS 9.2

Der Umstieg auf SAS 9.2 im Frühjahr 2009 versprach deutliche Verbesserungen in vielen Bereichen. Um zu prüfen, ob sich das bei der Prozedur MIXED bestätigt, wurden Vergleichsrechnungen durchgeführt.
Der Speicherbedarf hat sich bei Nutzung von SAS 9.2 nur marginal verringert (Tabelle 4). Auch beim Abbruch der Verfahren (Out of memory, Speichermangel) hat sich keine Veränderung in SAS 9.2 gegenüber SAS 9.1.3 gezeigt.

Tabelle 4: Vergleich memory von SAS 9.13 zu SAS 9.2

Verfahren	Anzahl Beobachtungen	Anzahl Sorten	memory (MB)		Differenz	
			SAS 9.13	SAS 9.2	absolut (MB)	relativ (%)
PHI	6731	225	163	154	9	5,5
VK	6731	225	61	55	6	9,8
MW	4687	185	775	770	5	0,6

Tabelle 5: Vergleich cpu time von SAS 9.13 zu SAS 9.2

Verfahren	Anzahl Beobachtungen	Anzahl Sorten	cpu time (h:min)		Differenz	
			SAS 9.13	SAS 9.2	absolut (h:min)	relativ (%)
PHI	6731	225	24:12	16:29	7:42	32
VK	6731	225	1:45	1:30	0:14	13
MW	4687	185	2:49	2:01	0:48	28

Anders die Rechenzeit. Hier hat sich der Umstieg von SAS 9.1.3 auf 9.2 gelohnt. Die Zeitersparnis lag bei 13 bis 32 % (Tabelle 5).
Mit steigender Anzahl Sorten im Datensatz, d. h. mit steigender Unbalanciertheit, wird die Zeitersparnis z. B. beim Verfahren MW immer größer, wie Abbildung 2 zeigt.

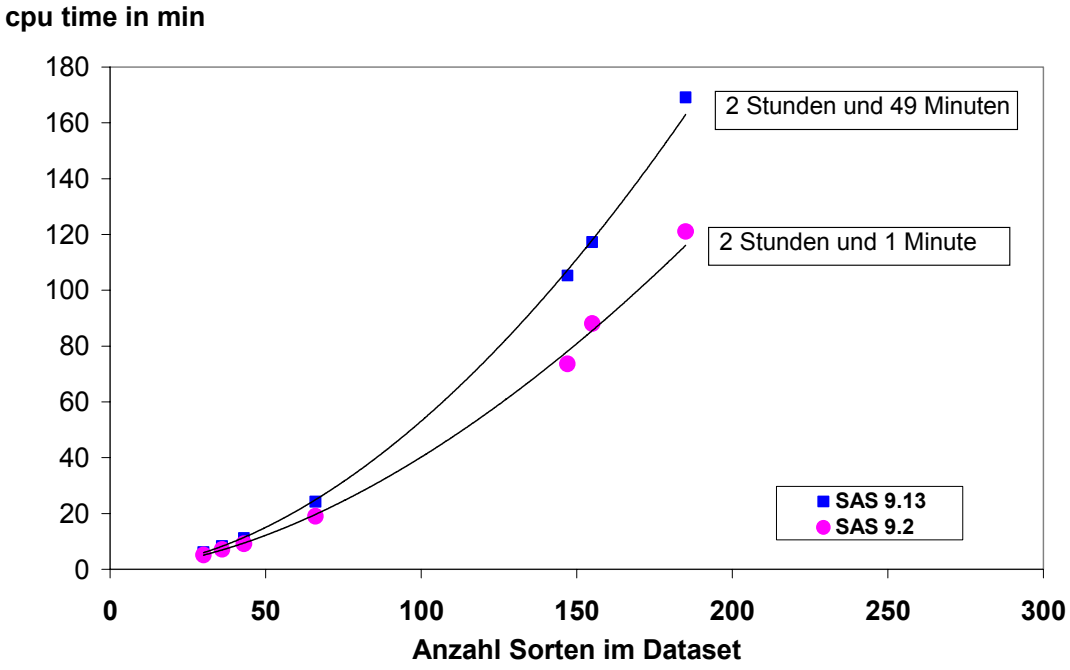

Abbildung 2: Entwicklung der cpu time bei steigender Anzahl Sorten im Datensatz (Verfahren MW)

6 Weitere Lösungsansätze

6.1 Neue Strategien der Datenreduktion innerhalb des Verfahrens MW

Im Verfahren MW werden die Möglichkeiten zur Reduktion des Datensatzes erweitert, um zielorientiert und effizient zu Ergebnissen zu kommen.
Erstens ist vorgesehen, über eine Excel-Datei mit den Sortimenten der drei letzten LSV-Jahre die Sorten zu definieren, die keinesfalls aus dem Datensatz gelöscht werden sollen. So wird gewährleistet sein, dass Mittelwerte auch für junge Sorten geschätzt werden können. Zweitens soll es im Verfahren möglich werden, eine kombinierte Datenreduktion zu wählen, z. B.: es soll keine Reduktion in den letzten zwei Jahren erfolgen, in den Vorjahren aber werden alle Sorten gelöscht, die nicht in mindestens 10 Versuchen insgesamt geprüft wurden. Auch so wird die Unbalanciertheit eingeschränkt, die neu geprüften Sorten bleiben davon jedoch unberührt.

6.2 Moderne Hardware; neue Betriebssysteme

Bei Einsatz moderner PC-Technik ist eine Verbesserung der Performance durch derzeit übliche Multicore-Prozessoren zu erwarten.
Ältere Betriebssysteme wie Windows NT und auch die SAS-Software vor SAS 9.2 erkannten einen Multicore-Prozessor als mehrere Einzelkernprozessoren, dadurch sind zwar auch alle Kerne nutzbar, spezielle Mehrkernprozessor-Optimierungen aber können nicht greifen. Mit Version 9.2 nutzt SAS Intel's Math Kernel Library (MKL). Mehr als 100 SAS-Prozeduren profitieren davon, darunter auch die Prozedur MIXED [4]. Das

lässt erwarten, dass durch die neue Technik die Probleme hinsichtlich Performance und Abbruch der Verfahren deutlich abnehmen.
Erste Tests wurden mit zwei neuen PC durchgeführt. Diese PC waren folgendermaßen ausgestattet:
PC neu 1: Intel QuadCore Q9550 Prozessor mit 2,83 GHz und 6 GB RAM
PC neu 2: Intel iCore i7 920 mit 2,66 GHz und 6 GB RAM.
Obwohl jeweils ein 64Bit Betriebssystem genutzt wurde, ist die 32Bit Variante der SAS Software 9.2 zum Einsatz gekommen.
Es hat sich gezeigt, dass sich die cpu time nochmals um mehr als 50% verringerte (Tabelle 6). Zudem konnten größere Datenmengen ausgewertet werden, ohne dass ein Abbruch wegen Speichermangel erfolgte. Für 164 Sorten wurden so Mittelwerte ermittelt, die bisher bei alter Technik durch die notwendige Datenreduktion nicht in den Ergebnissen enthalten waren. Die cpu time nahm hier aber durch die deutlich höheren Unbalanciertheit der Daten wieder stark zu. Dabei war der PC neu 2 der leistungsstärkere PC.

Tabelle 6: Vergleich des Verfahrens MW mit alter und neuer Technik

Anzahl Beobachtungen	Anzahl Sorten	cpu time (h:min)			Differenz relativ (%)	
		PC alt	PC neu 1	PC neu 2	PC neu 1 zu PC alt	PC neu 2 zu PC neu 1
4687	185	02:01	00:51	00:43	58	18
5439	349	Abbruch	03:45	02:46	-	23

6.3 Prozedur HPMIXED

HPMIXED ist eine neue Prozedur in SAS 9.2. Sie kann für gemischte lineare Modelle mit sehr vielen fixen oder zufälligen Faktoren und sehr große Datensätze zum Einsatz kommen. Aber nicht alle Möglichkeiten der Prozedur MIXED sind auch bei HPMIXED gegeben. Das betrifft z. B. die TYPE-Option im RANDOM-Statement, die DDFM-Option oder das TEST- Statement [6]. Da die Prozedur HPMIXED aber erheblich schneller ist, wird z. Z. geprüft, ob und wie es möglich ist, in den Verfahren PHI, VK und / oder MW die Prozedur MIXED durch HPMIXED zu ersetzen.

7 Fazit

Aufgrund unserer Erfahrungen können wir folgende Empfehlungen für den Einsatz der Hohenheim-Gülzower Serienauswertung mit PIAFStat geben:

- Vorteilhaft ist es, moderne Hardware und neueste SAS-Version einzusetzen.
- Die Auswertung der mehrjährigen Daten mit den Verfahren PHI und VK sollte außerhalb der Saison ohne Zeitdruck erfolgen.
- Für das Verfahren MW empfehlen wir, auf die neuen Reduktionsmöglichkeiten der Daten im Verfahren zurückzugreifen. Dadurch werden sowohl neue Sorten

als auch langjährig vielfach geprüfte Sorten nicht ungewollt von der Auswertung ausgeschlossen.

Die Entwicklungen hinsichtlich der Prozedur HPMIXED in künftigen SAS-Versionen wollen wir verfolgen, um sich ergebende Möglichkeiten in den Verfahren umzusetzen.

Literatur

[1] Michel, V. ; Zenk, A.; Graf, R.; Möhring, J.; Büchse, A.; Piepho, H.P. (2007): The "Hohenheim-Gülzow-method" for analysis of series of trials as basic procedure for PIAFStat and SAS in a regionalized field trial system. Proceedings of the International Symposium "Agricultural Field Trials – Today and Tomorrow", Stuttgart, Grauer-Verlag, 2007, S. 136-141

[2] Michel, V. ; Zenk, A.; Möhring, J.; Büchse, A.; Piepho, H.P. (2007): Die Hohenheim-Gülzower Sereinauswertung als bundesweites Basisverfahren im regionalisierten Sortenwesen, Mitteilungen der Landesforschungsanstalt für Landwirtschaft und Fischerei Mecklenburg-Vorpommern, Gülzow, 2007, Heft 37, S.72-82

[3] Möhring, J.; Büchse, A. und Piepho, H.-P. (2005): Auswertung von landwirtschaftlichen Sortenversuchen mit PROC MIXED – Spagat zwischen Theorie und Praxis. Proceedings der 9. KSFE Berlin, Shaker-Verlag, 2005, S.279-288

[4] Ralston, C.E. (2008): Making sense of Multi-core Processor technology for SAS Environment. Proceedings of the SAS Global Forum, Paper 388-2008

[5] Roßberg, D., V. Michel, R. Graf & R. Neukampf (2007): Boden-Klima-Räume und Anbaugebiete als Basis des regionalisierten Sortenwesens in Deutschland. Mitteilungen der Landesforschungsanstalt für Landwirtschaft und Fischerei Mecklenburg-Vorpommern, Gülzow, 2007, Heft 37, 24-30

[6] SAS Institute, Inc., SAS/STAT User's Guide, Version 9.2, Cary, NC, USA 2009

[7] Zenk, A.; Möhring, J. und Michel, V. (2005): Einbindung neuer Methoden zur Routineauswertung von landwirtschaftlichen Versuchen mit Hilfe von SAS-Makros. Proceedings der 9. KSFE Berlin, Shaker-Verlag, 2005, S.407-417