Signifikanzen grafisch veranschaulichen

Jörg Sellmann
Julius Kühn-Institut
Stahnsdorfer Damm 81
14532 Kleinmachnow
joerg.sellmann@julius-kuehn.de

Zusammenfassung

Das Ergebnis der MEANS- oder LSMEANS-Anweisung linearer Modelle sind entweder Tabellen mit p-Werten für die paarweisen Vergleiche und/oder Buchstaben aus der LINES-Anweisung. P-Werte kleiner als das gewählte Signifikanz-Niveau werden dann üblicherweise mit einem Stern gekennzeichnet. Gibt es nun eine Möglichkeit, diese Sterne oder die Buchstaben in Grafiken einzubetten?
Mittels SGANNO= können der SGPLOT-Grafik relativ frei gestaltbare Elemente hinzugefügt werden. Im Fall nur einer Kategorie ist es ein einfaches Vorgehen, die zuvor mittels ODS output= erstellten Datensätze so zu erweitern, dass die gewünschten Sterne/Buchstaben in die Balken-, Streuungs- oder Boxplot-Grafiken integriert werden können.
Kommt zu der Kategorie noch eine Gruppierung GROUP= hinzu, ist ein wenig mehr zu investieren, doch auch hier lassen sich die Sterne oder Buchstaben in die Grafik übernehmen. Das Zauberwort heißt DISCRETEOFFSET.
Hier werden Makros vorgestellt, die im Anschluss an PROC MIXED, PROC GLM oder PROC GLIMMIX die Datensätze für die „Signifikanz-Grafik“ generieren.

Schlüsselwörter: Signifikanz, Lines, SGPLOT, SGANNO, Discreteoffset

1 Einführung

Im landwirtschaftlichen Versuchswesen des Julius Kühn-Instituts (JKI) werden die geplanten Feldversuche oder die Erhebungen zu Pflanzenschutzintensitäten mittels linearer Modelle in SAS ausgewertet. Die daraus resultierenden Veröffentlichungen und Berichte enthalten eine große Anzahl von Tabellen und Grafiken mit dem Ziel, Unterschiede zwischen Methoden und Strategien aufzuzeigen. Den Signifikanz-Stern und/oder die Signifikanz-Buchstaben mittels SAS-Mitteln in die Tabellen zu bekommen ist kein schwieriges Unterfangen.
Anders sieht es bei den Grafiken aus. Hier hieß der Weg, mittels eines geeigneten Zeichenprogramms die Sterne oder Buchstaben in die SAS-Grafik nachträglich einzuzeichnen.
Anhand eines realen mehrjährigen Feldversuchs, hier aber mit simulierten Ertragsdaten, soll ein automatisiertes Vorgehen in SAS gezeigt werden, das eine nachträgliche Bearbeitung der Grafiken überflüssig macht.
Das Ziel ist es, das Zusammenspiel der linearen Modellierung, von ODS output und deren einfache Überführung in anschauliche Grafiken zu demonstrieren.

2 Die 6 Makros

Im Rahmen dieses Beitrages werden folgende Makros genutzt, wobei V für vertikale und H für horizontale Grafiken steht, z. B. VBAR, VBOX bzw. HBAR, HBOX.

Für alpha-numerische Kategorien, ohne Gruppe:

```
%MACRO significanceLettersV (labdat, daten, xc1val, labelval);
%MACRO significanceLettersH (labdat, daten, xc1val, labelval);
```

Für numerische Kategorien, ohne Gruppe:

```
%MACRO significanceLettersVN (labdat, daten, x1val, labelval);
%MACRO significanceLettersHN (labdat, daten, x1val, labelval);
```

Kategorie und Gruppe (2-fache Klassifizierung):

```
%MACRO significanceLettersV2(result, category, group);
%MACRO significanceLettersH2(result, category, group);
```

Die Unterscheidung des Typs der Kategorie ist wesentlich. Eine Verwechslung führt zu Warnungen im log (`WARNING: The XC1 option is the incorrect data type. The option will be ignored.`) und zu Fehlern in der Grafik. Ebenso ist die Unterscheidung in H und V zwingend, da bei H die Markierungen rechts neben der Grafik, bei V oberhalb angezeigt werden.

Jedes der ersten vier Makros erzeugt einen Annotate-Datensatz, der die Positionen und die Signifikanz-Level der Kategorien als Buchstaben/Buchstabenkombinationen enthält. Es werden die folgenden Parameter verwendet:

labdat	(Ausgabe-)Label-Dataset
daten	Daten mit den Kategorien und Level
xc1val	Kategorie-Variable (alphanumerisch) bzw.
x1val	Kategorie-Variable (numerisch)
labelval	Signifikanz-Level-Variable oder -Kombination

Wird als Grundlage ein per %MULT-Makro erzeugter Datensatz LETTERS genutzt, so gilt

- 2. Parameter letters
- 3. Parameter label
- 4. Parameter upcase(catx('00'x, of col:))

Wird als Grundlage ein mittels

```
ODS OUTPUT LSMLines=Lines
```

und der Option LINES erzeugter Datensatz Lines der Prozedur GLM genutzt, dann gilt analog

- 2. Parameter lines (where=(not missing(<class>))),
- 3. Parameter die Klasse <class> aus der class Anweisung
- 4. Parameter reverse(catx('00'x, of Line:))

Für andere lineare Modelle sind die entsprechenden Optionen und Parameter durch eigenes Probieren mittels `ODS TRACE ON; PROC …; RUN; ODS TRACE OFF;` leicht zu finden.

Der Kern der ersten 4 Makros besteht im Hinzufügen der Spalten

```
Function="Text", XC1/X1, X1space, Y1space, YC1,
Anchor, Justify, Label, Textcolor, Textweight
```

zum übergebenen Datensatz (2. Parameter). Beispielhaft sei hier das erste Makro angegeben:

```
%MACRO significanceLettersV (labdat, daten, xc1val, labelval);
 data &labdat.;
 retain    function "text" X1SPACE 'datavalue'
           Y1 100 Y1SPACE 'datapercent'
           textweight 'BOLD' textcolor 'blue'
           justify 'center' anchor 'TOP';
 set &daten.;
 XC1=&xc1val.;
 label1=&labelval.;
 drop label;
 rename label1=label;
 run;

%MEND significanceLettersV;
```

Die beiden letzten Makros für die 2-fache Klassifizierung fügen dann lediglich die Spalte `discreteoffset` zum Datensatz hinzu. Zugrunde liegt dabei die Formel

```
discreteoffset=0.85 * (-0.5 + 1/symget('grp')*(lfdnr-0.5));
```

wobei `grp` die (gleiche) Anzahl Gruppen pro Kategorie ist und `lfdnr` die laufende Nummer innerhalb der Kategorie. Sowohl `grp` als auch `lfdnr` werden innerhalb des Makros generiert. Für die horizontale Version wird noch eine empirisch ermittelte minimale Verschiebung von `-0.1` hinzugerechnet.

3 Anwendungsfall Einfaktoriell

3.1 Vergleich zur Kontrolle: Stern

Im ersten Schritt stellen wir die Frage, ob die Erträge zum Ausgangsjahr 2008 hin signifikant verschieden sind.

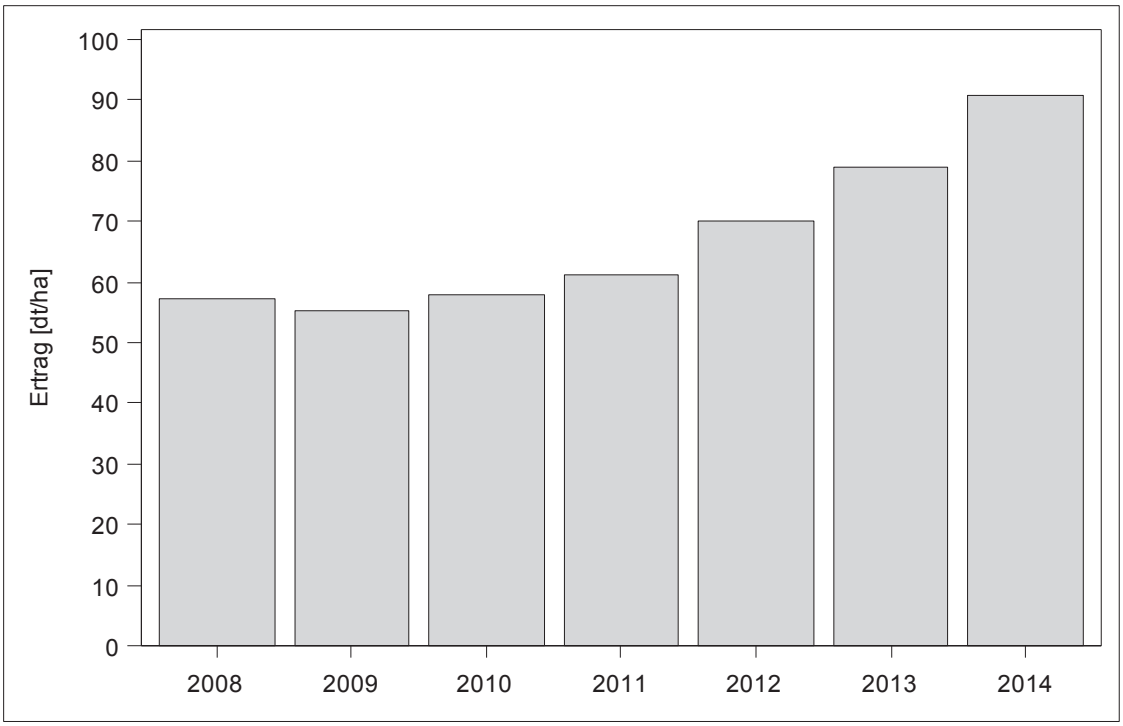

Abbildung 1: (simulierte) Erträge der Jahre 2008-2014

```
proc mixed data=Vergleich nobound;
ods output diffs=diffs;
class Jahr;
model Ertrag = Jahr / ddfm=kr;
random Jahr;
lsmeans Jahr / adjust=Simulate(report seed=423309000)
pdiff=Control('2008');
run;

data diffs;
set diffs;
if adjp<0.05 then sign='*';
run;
```

Die Differenzen zu 2008 werden im Datensatz DIFFS abgelegt (PROC MIXED) und bei Signifikanz mit einem Stern gekennzeichnet (DATA diffs):

Tabelle 1: Datensatz DIFFS

Obs	Effect	Jahr	_Jahr	Estimate	StdErr	DF	tValue	Probt	Adjustment	Adjp	sign
1	Jahr	2009	2008	-1.8112	1.5687	133	-1.15	0.2503	Simulate	0.7178	
2	Jahr	2010	2008	0.9183	1.5687	133	0.59	0.5593	Simulate	0.9806	
3	Jahr	2011	2008	4.2328	1.5687	133	2.70	0.0079	Simulate	0.0387	*
4	Jahr	2012	2008	13.0827	1.5687	133	8.34	<.0001	Simulate	<.0001	*
5	Jahr	2013	2008	21.7351	1.5687	133	13.86	<.0001	Simulate	<.0001	*
6	Jahr	2014	2008	33.7535	1.5687	133	21.52	<.0001	Simulate	<.0001	*

Anschließend werden mit Hilfe der Makros die Annotate-Datensätze LABEL4 bzw. LABEL5 erstellt und in die Grafiken mittels SGANNO= integriert:

```
/* vertikal, numerisch */
%significanceLettersVN(label4, diffs, jahr, sign));

proc sgplot data=Vergleich sganno=label4 noautolegend;
vbar Jahr / response=Ertrag stat=mean limitstat=clm;
yaxis values=(50 to 100 by 10) label="Ertrag [dt/ha]";
xaxis display=(Nolabel);
run;
```

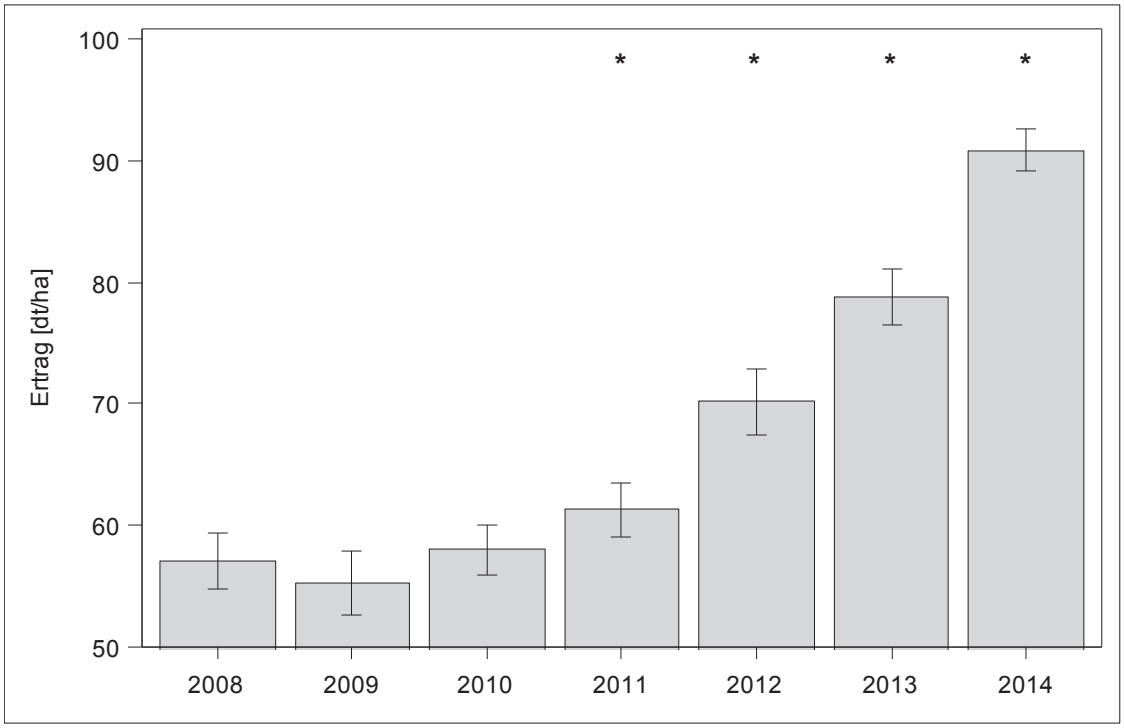

Abbildung 2: Erträge und Signifikanzen im Vergleich mit 2008

```
/* horizontal, numerisch und verschöneren */
%significanceLettersHN(label5, diffs, jahr, sign));

proc format;
value jahr
 2008="Versuchsstart   2008"
 2009="1. Versuchsjahr 2009"
 2010="2. Versuchsjahr 2010"
 2011="3. Versuchsjahr 2011"
 2012="4. Versuchsjahr 2012"
 2013="5. Versuchsjahr 2013"
 2014="6. Versuchsjahr 2014"
;
run;
```

```
proc sgplot data=Vergleich sganno=label5 noautolegend;
hbar Jahr / response=Ertrag stat=mean limitstat=clm;
xaxis values=(50 to 100 by 10) label="Ertrag [dt/ha]";
yaxis display=(Nolabel);
format jahr jahr.;
run;
```

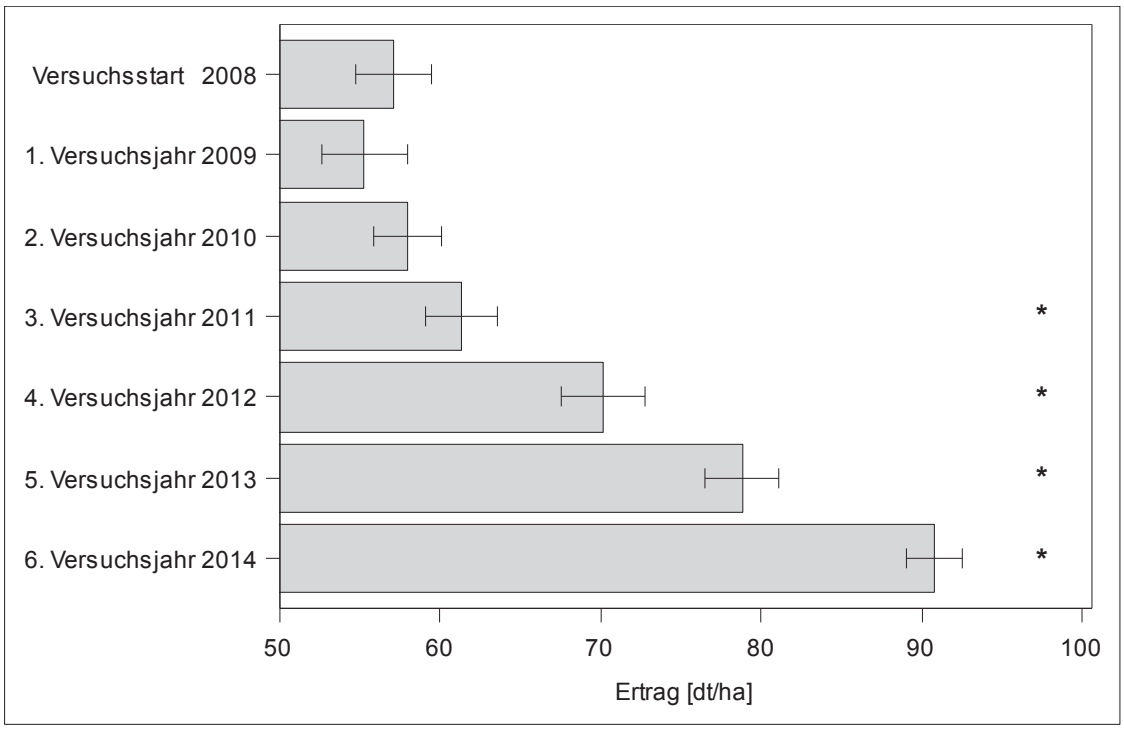

Abbildung 3: Horizontale und „aufgehübschte“ Version von Abb. 2

3.2 Vergleich untereinander: Buchstaben

Im nächsten Beispiel wollen wir die (neu simulierten) Erträge der Jahre untereinander vergleichen. Wir nutzen wiederum PROC MIXED und in Ermangelung einer LINES-Anweisung das MULT-Makro von PIEPHO [1]. Dabei ist die ODS-Anweisung in der verwendeten Form zwingend notwendig.
Anschließend erzeugen wir den Annotate-Datensatz LABEL3. Der ungewöhnlich aussehende 4. Parameter ergibt die Verknüpfung der COL-Spalten aus dem Datensatz LETTERS, der das Resultat des Makros MULT ist.

```
proc mixed data=Ertraege nobound;
ods output diffs=diffs lsmeans=lsmeans;
class Jahr;
model Ertrag = Jahr / ddfm=kr;
random Jahr;
lsmeans Jahr / adjust=Simulate (report seed=423309000) pdiff;
%mult(trt=Jahr, p=adjp);
run;

%significanceLettersVN
 (label3, letters, label, upcase(catx('00'x, of col:)));
```

Wir erhalten das folgende Letter-Display:

trt	label	lsmean	letters			
Jahr	2008	68.918854		a		
	2009	63.571044			b	c
	2010	59.817458				c
	2011	60.930565				c
	2012	61.88053			b	c
	2013	66.014845		a	b	
	2014	70.274788		a		

Die Grafiken mit den Buchstaben werden dann wie folgt gebildet, von denen die zweite, der mittleren Ertrags-Größe nach sortierte, hier gezeigt werden soll (Abb. 4).

```
proc sgplot data=Ertraege sganno=label3 noautolegend;
vbar Jahr / response=Ertrag stat=mean limitstat=clm;
yaxis values=(50 to 80 by 10) label="Ertrag [dt/ha]";
xaxis display=(Nolabel);
run;

proc sgplot data=Ertraege sganno=label3 noautolegend;
vbar Jahr / response=Ertrag stat=mean limitstat=clm
categoryorder=respdesc;
yaxis values=(50 to 80 by 10) label="Ertrag [dt/ha]";
xaxis display=(Nolabel);
run;

proc sgplot data=Ertraege sganno=label3 noautolegend;
vbox Ertrag / category=Jahr meanattrs=(symbol=plus);
yaxis values=(40 to 90 by 10) label="Ertrag [dt/ha]";
xaxis display=(Nolabel);
run;
```

Hinweis: In SAS 9.4 M0 und M1 kann es bei Verwendung von categoryorder und/oder Formaten zum Fehlen von Buchstaben kommen. Ab Release M2 ist der vom Autor vermutete Bug offensichtlich stillschweigend behoben.

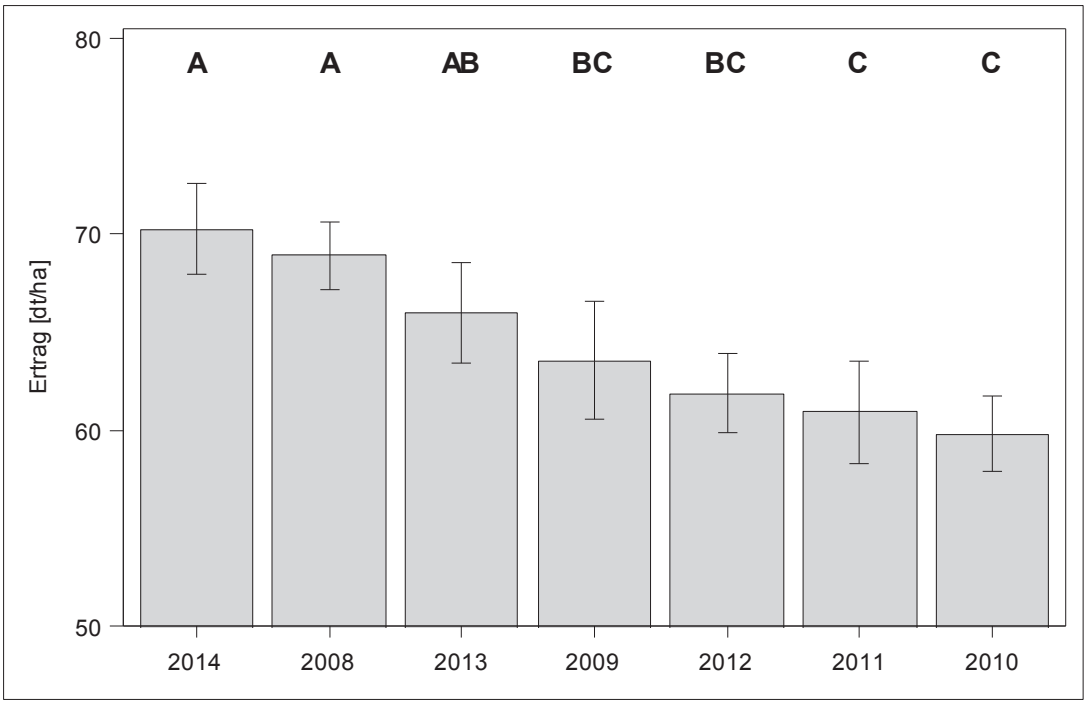

Abbildung 4: sortierte Erträge der Jahre und die Signifikanz-Level

4 Anwendungsfall Zweifaktoriell

Zu den Jahren 2008-2014 nehmen wir jetzt noch als zweiten Faktor eine (fiktive) Anbau- und Pflanzenschutz-Strategie in 3 Varianten hinzu. Wir wollen aber nicht zwei beliebige Kombinationen von Jahr und Strategie miteinander vergleichen, sondern immer nur die Strategien auf gleicher Jahresstufe bzw. die Jahre innerhalb einer Strategie.

4.1 Vergleich gegen Null: Stern

Ausgangspunkt sind die pro Jahr und Strategie (simulierten) erzielten Mehrerträge (Abb. 5). Die Frage ist, ob die pro Jahr und Strategie erzielten Mehrerträge gegenüber der Kontrolle, jeder für sich allein betrachtet, signifikant größer als Null sind. Die Ergebnisse werden mittels ODS in den Datensatz TTESTS geschrieben.

```
proc sort data=Strategien;
by Strategie Jahr;
run;

proc ttest data=Strategien sides=upper;
ods output ttests=ttests;
by Strategie Jahr;
var Mehrertrag;
run;
```

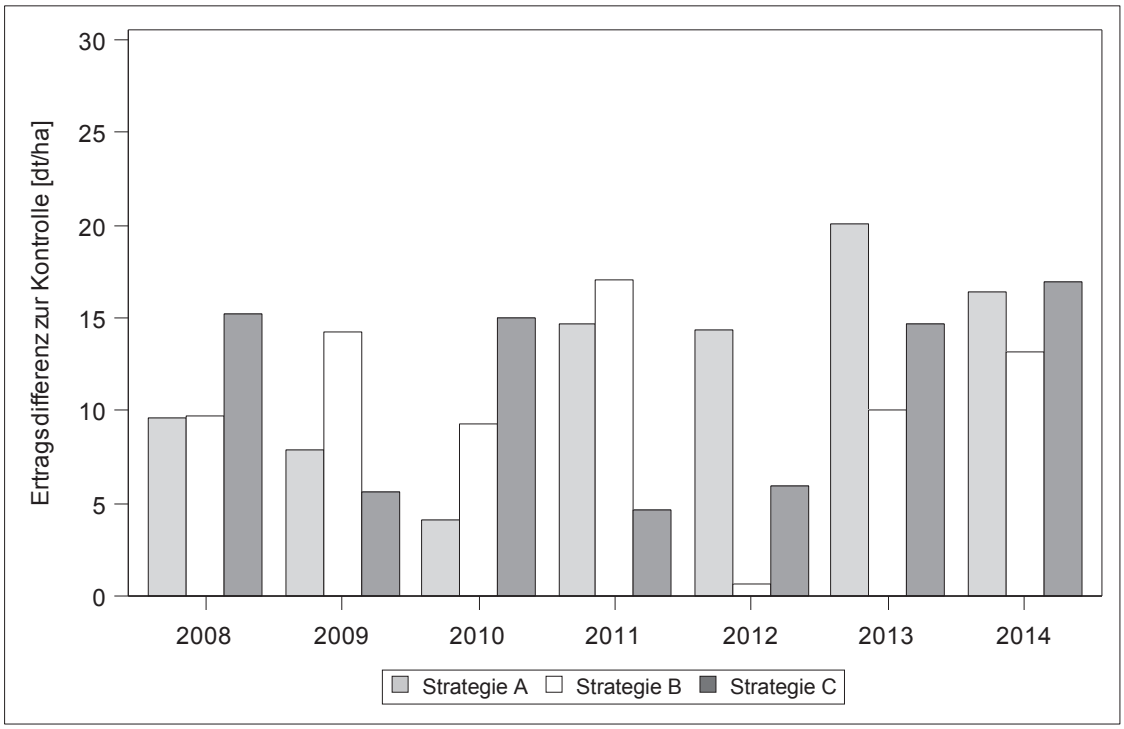

Abbildung 5: simulierte Mehrerträge der Jahre und Strategien

Werden in der bewährten Weise dem Datensatz TTESTS die Sterne hinzugefügt, ergibt sich folgende Tab. 2:

Tabelle 2: Datensatz TTESTS (Auszug)

Strategie	Jahr	Variable	tValue	DF	Probt	sign
Strategie A	2008	Mehrertrag	2.34	19	0.0153	*
Strategie A	2009	Mehrertrag	1.08	19	0.1464	
…						
Strategie B	2008	Mehrertrag	1.34	19	0.0974	
…						
Strategie C	2008	Mehrertrag	2.17	19	0.0215	*
…						

Das unter 3.1 skizzierte Vorgehen führt hier jedoch zu einer falschen Grafik (Abb. 6), da pro Kategorie (Jahr) alle Markierungen übereinander gelegt werden.

```
%significanceLettersVN(label1, ttests, Jahr, Sign);

proc sgplot data=Strategien sganno=label1;
vbar Jahr / group=Strategie response=Mehrertrag
            stat=mean limitstat=clm groupdisplay=cluster;
yaxis max=50 label="Ertragsdifferenz zur Kontrolle [dt/ha]";
refline 2008 to 2013 by 1 /axis=x discreteoffset=0.5
                           lineattrs=(pattern=dot);
label Strategie = '00'x;
run;
```

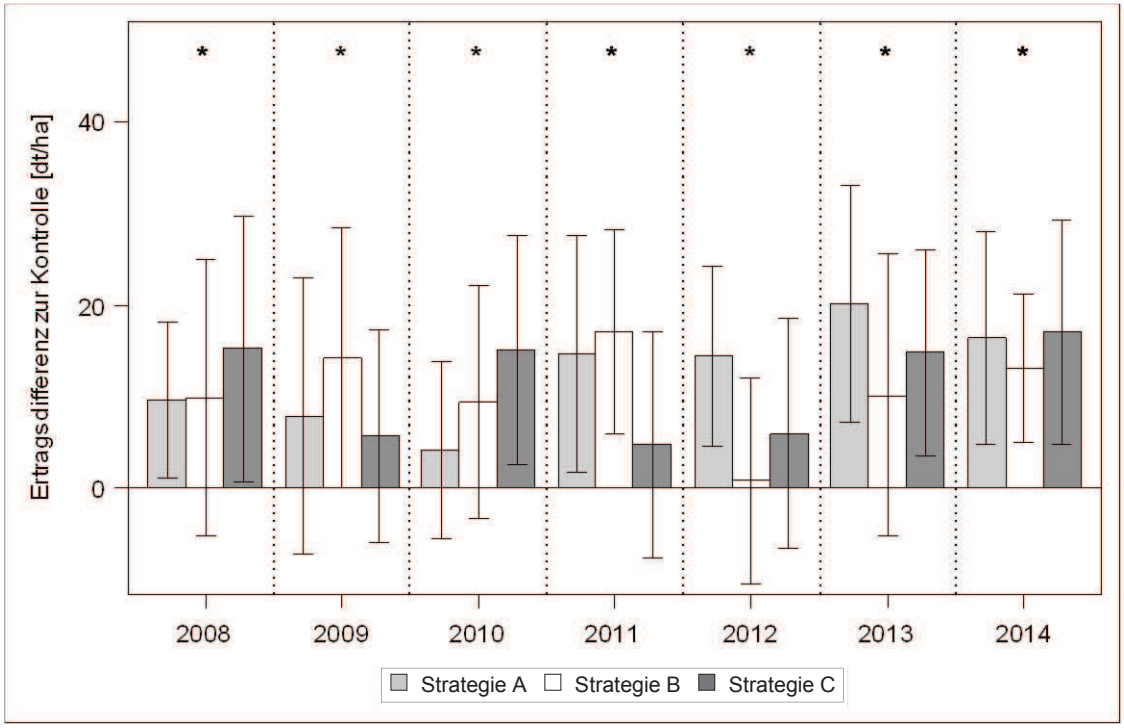

Abbildung 6: falsche Darstellung der Signifikanzen

Mittels des Makros für die 2-fache Klassifizierung

```
%significanceLettersV2(label1, Jahr, Strategie);
```

wird der DISCRETEOFFSET hinzugefügt und die gleiche Grafik-Anweisung führt jetzt zur gewünschten richtigen Abb. 7:

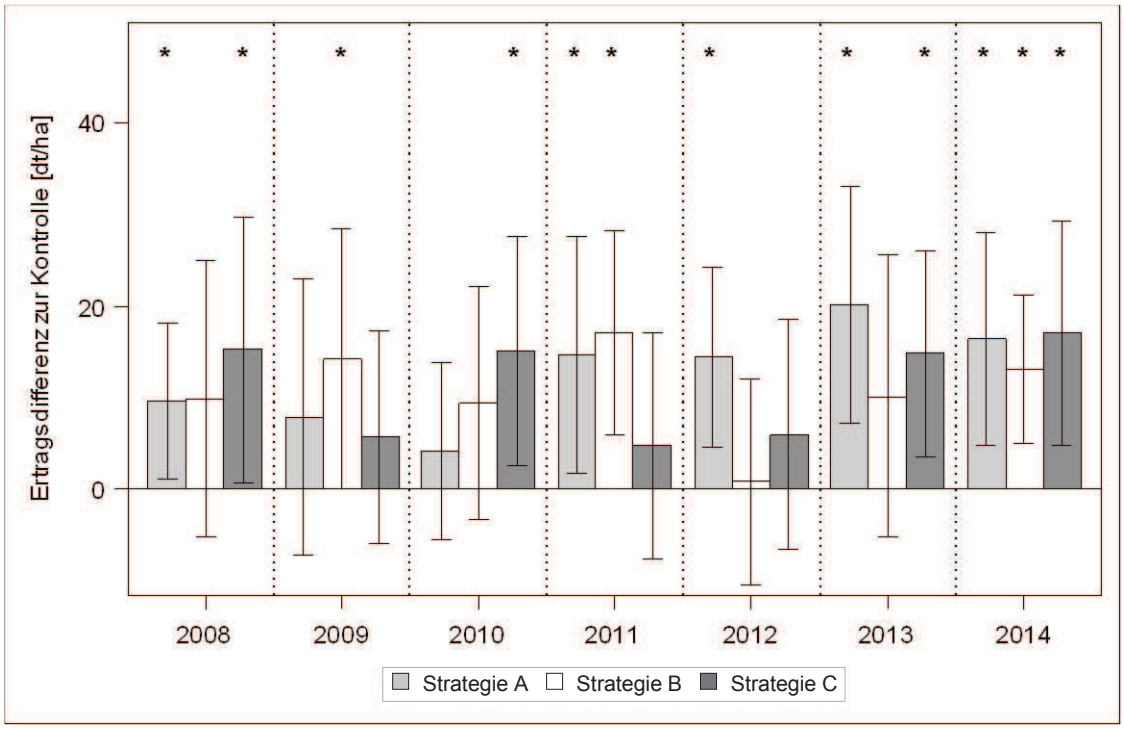

Abbildung 7: Mehrerträge pro Jahr/Strategie und Signifikanz gegen Null

Mit den Anweisungen

```
%significanceLettersHN(label3, ttests, Jahr, Sign);

%significanceLettersH2(label3, Jahr, Strategie);

proc sgplot data=Strategien sganno=label3;
hbar Jahr / group=Strategie response=Mehrertrag stat=mean
            limitstat=clm groupdisplay=cluster;
xaxis max=50 label="Ertragsdifferenz zur Kontrolle [dt/ha]";
refline 2008 to 2013 by 1 /axis=y discreteoffset=0.5
                           lineattrs=(pattern=dot);
label Strategie = '00'x;
run;
```

erhalten wir eine horizontale Ausrichtung der Grafik.

4.2 Vergleich untereinander: Buchstaben

In nächsten Schritt vergleichen wir innerhalb eines Jahres die Strategien untereinander. Dazu führen wir unter Kenntnis der Nachteile dieses Vorgehens eine jahresweise Analyse mittels PROC GLM durch. Wichtig ist hier wieder der ODS Output. Anschließend greifen wir auf die bekannten Makros zurück und erstellen die Grafik.

```
proc glm data=strategien;
ods output lsmlines=lines;
by jahr;
class Strategie;
model Ertrag = Strategie;
lsmeans Strategie / adjust=Simulate pdiff lines;
run;

/* Label erstellen ohne Berücksichtigung der Gruppierung */
%significanceLettersVN(label2, lines (where=(not
missing(Strategie))), Jahr, reverse(catx('00'x,of Line:)));

/* Gruppierung einrechnen */
%significanceLettersV2(label2, Jahr, Strategie);

proc sgplot data=Strategien sganno=label2;
vbar Jahr / group=Strategie response=Ertrag stat=mean
            limitstat=clm groupdisplay=cluster;
xaxis display=(nolabel);
yaxis values=(50 to 100 by 10) label="Ertrag [dt/ha]";
refline 2008 to 2013 by 1 /axis=x discreteoffset=0.5
                           lineattrs=(pattern=dot);
label Strategie = '00'x;
run;
```

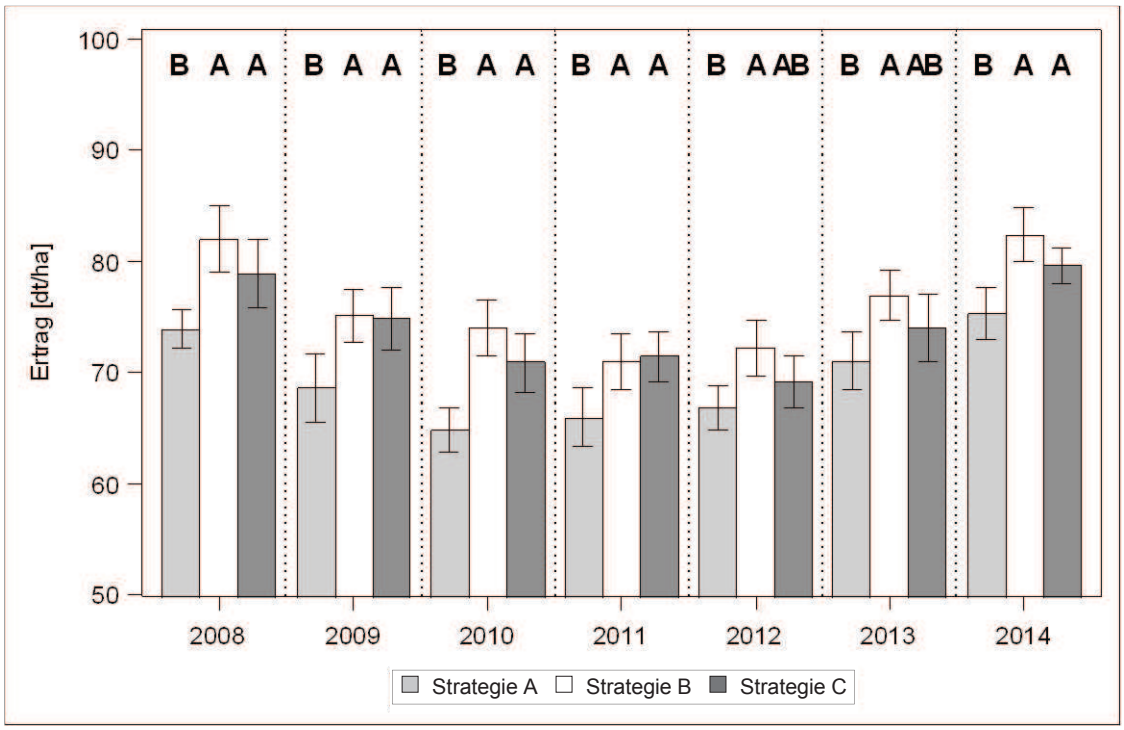

Abbildung 8: Erträge der Strategien und deren Signifikanz pro Jahr

Die Buchstaben in Abb. 8 gelten immer nur innerhalb eines Jahres (=Kategorie). Sie sind nicht zum Vergleich zwischen den Jahren verwendbar.

5 PROC PLM und By-Annotate

Arbeitet man mit der „einfachen" Prozedur MIXED und wenigstens 2 Faktoren, so fehlen dem Autor immer wieder einfache Dinge wie `Lines` oder `SliceBy/SliceDiff`. Es gibt zwar `Slice`, die Prozedur führt aber immer alle Vergleiche durch. Eine mögliche Lösung bietet hier die Prozedur PLM (vermutlich **P**ostprocessing **L**inear **M**odels [2]). Prinzipiell kann in allen Verfahren der linearen Analyse (MIXED, GLM, GENMOD, GLIMMIX, …) das Modell für eine spätere Analyse mittels `store` gesichert werden:

```
proc mixed data=Strategien;
class Strategie jahr;
model ertrag=Strategie jahr Strategie*jahr /ddfm=kr;
lsmeans Strategie*jahr / slice=jahr diff;
store ertraegeResults;
run;
```

Mit PROC PLM können dann im Nachgang u. a. `Sliceby` und `Lines` gerechnet werden:

```
proc plm restore=ertraegeResults;
ods output lsmLines=lsmLines sliceLines=sliceLines;
lsmeans Strategie / adjust=Simulate pdiff lines;
slice Strategie*jahr / sliceby=Strategie pdiff lines;
run;
```

Ein weiteres Handicap in SAS ist, dass die Verwendung von Annotate-Datensätzen in Kombination mit `by`-Statements syntaktisch zwar korrekt, semantisch jedoch fehlerhaft ist. So führt das folgende Statement zu einer Überlagerung aller 3 Ebenen des Annotate-Datensatzes:

```
proc sgplot data=Strategien sganno=label5 noautolegend;
by Strategie;
title #byval1;
vbar jahr / response=ertrag stat=mean limitstat=clm;
xaxis display=(nolabel);
yaxis max=100 label="Ertrag [dt/ha]";
run;
```

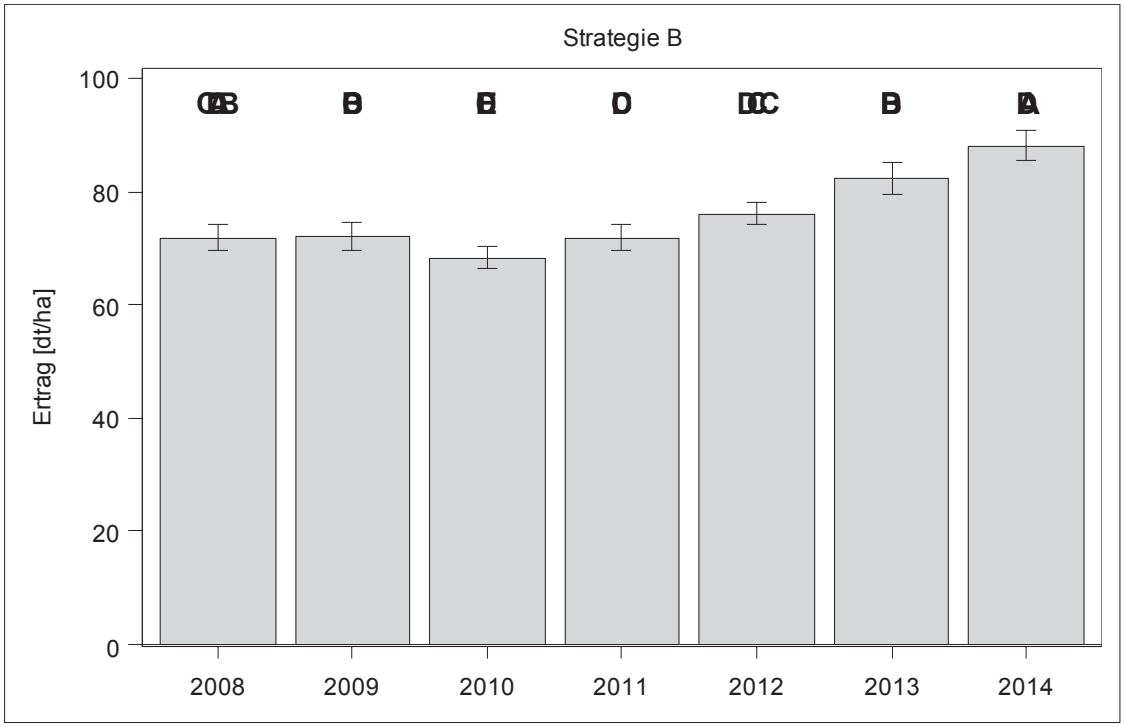

Abbildung 9: Fehlerhaftes `sganno` mit `by`

Im Rahmen dieses Beitrags sollen die Signifikanzbuchstaben aus `PLM/Sliceby` in die Grafik integriert und gleichzeitig das genannte Handicap überwunden werden. Dazu ergänzen wir das PLM-Statement um die Zeile

```
%significanceLettersVN(
 label5(rename=(slice=Strategie)),
 sliceLines(where=(not missing(effect))),
 jahr,
 catx('00'x, of Line:)
);
```

Mit einer kleinen Makro-Schleife generieren wir dann die drei korrekten Grafiken.

```
data label6; set label5; strategie=substr(strategie,11,11); run;

%macro plot(byval);
 data label6by;
 set label6;
 where Strategie="&byval.";
 run;

 proc sgplot data=Strategien sganno=label6by noautolegend;
 title "&byval.";
 where Strategie="&byval.";
 vbar jahr / response=ertrag stat=mean limitstat=clm;
 xaxis display=(nolabel);
 yaxis max=100 label="Ertrag [dt/ha]";
 run;
%mend;

data _null_;
set Strategien;
by Strategie;
if first.Strategie then call execute('%PLOT('||Strategie||')');
run;
```

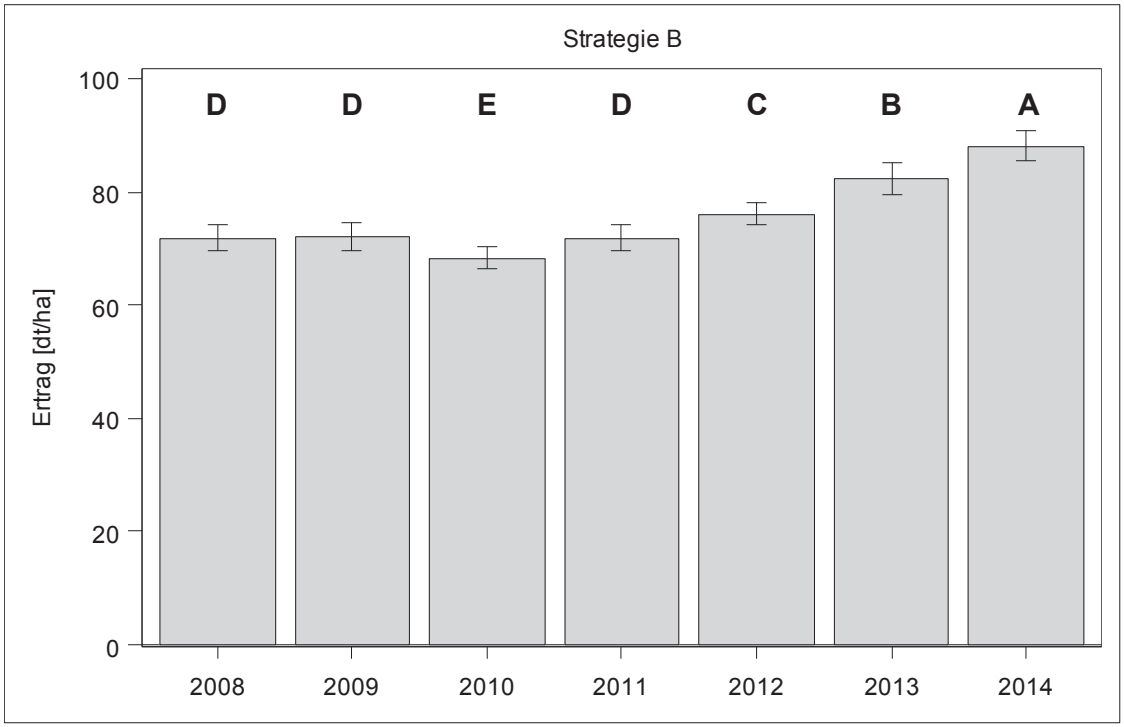

Abbildung 10: korrektes `sganno` mit „Pseudo“ `by`

6 Programme und Grafiken

Alle verwendeten Programme und Makros finden Sie unter:

http://sf.julius-kuehn.de/sas

Die in diesem Beitrag verwendeten Grafiken wurden mittels den Anweisungen

```
ods listing style=base.template.Style gpath='d:\zzz';
ods graphics / outputfmt=emf width=560px height=360px;
```

als skalierbare schwarz-weiße EMF-Grafiken parallel zur Bildschirmausgabe im Ordner `gpath` abgelegt [3].

Hinweis: In SAS 9.4 scheint sich aber ein Bug in das EMF-Format eingeschlichen zu haben, denn bei der Konvertierung des Word-Dokuments nach PDF verschwanden einige Legendeneinträge. Mit SAS 9.3 war alles noch in Ordnung.

Literatur

[1] H.P. Piepho (2012): A SAS macro for generating letter displays of pairwise mean comparisons. Communications in Biometry and Crop Science 7 (1), 4–13.

[2] R. Tobias (2010): Introducing PROC PLM and Postfitting Analysis for Very General Linear Models in SAS/STAT® 9.22. SAS Global Forum 2010, Paper 258-2010.

[3] E. Moll, J. Sellmann (2014): Ausgabe von Grafiken in editierfähige EMF-Formate unter SAS 9.3 und SAS 9.4. In: A. Koch, R. Minkenberg (Hrsg.): KSFE 2015 - Proceedings der 19. Konferenz der SAS-Anwender in Forschung und Entwicklung (KSFE); Shaker Verlag, Aachen (2015), S. 235 - 242.