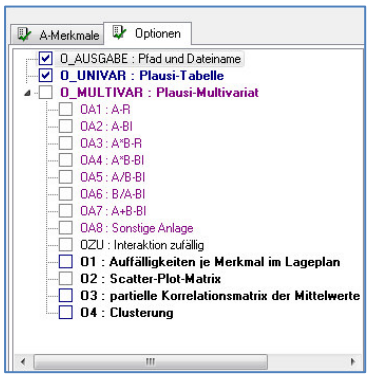
Landesforschungsanstalt für
Landwirtschaft und Fischerei

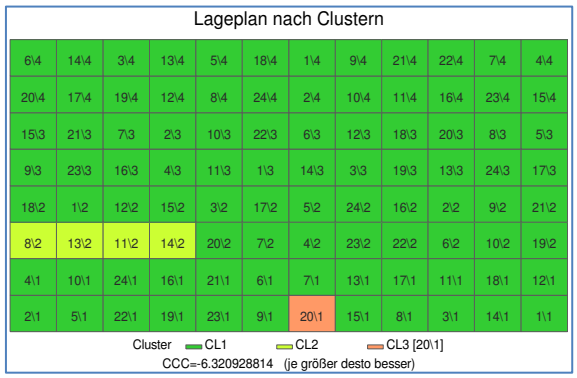
Mecklenburg
Vorpommern

Qualitätscheck von Versuchsdaten

(Benutzerhandbuch für das Software-Produkt PLAUSI)



im Arbeitsthema erstellte Software für Versuchsauswerter



Detektion einer Trockenstelle in einem Maisversuch mit dieser Software

Zwischenbericht 2016

Stand: 29.03.2016

Bearbeiter des Infotextes: Dr. V. Michel, Dr. A. Zenk

Institut für Pflanzenproduktion und Betriebswirtschaft

Dorfplatz 1/OT Gülzow
18276 Gülzow-Prüzen
www.lfamv.de

GLIEDERUNG Seite

1 Zusammenfassung ... 1
2 Einführung ... 1
3 Ziele der Entwicklung ... 2
4 Nutzeroberfläche - Gesamtübersicht ... 2
4.1 Klassifizierungs-Merkmale ... 2
4.2 Analyse-Merkmale ... 3
4.3 Optionen ... 4
5 Univariate Analyse - Statistiken der Einzelmerkmale ... 6
5.1 Aufgabe ... 6
5.2 Nutzerführung ... 6
5.3 Output und allgemeine Hinweise zu ausgewiesenen Kennzahlen ... 7
6 Multivariate Analysen ... 8
6.1 Allgemeiner Hintergrund ... 8
6.2 Auswahl der Merkmale für eine multivariate Auswertung ... 9
6.3 Residuen der Einzelmerkmale im Lageplan ... 9
6.3.1 Aufgabe ... 9
6.3.2 Nutzerführung ... 10
6.3.3 Output ... 11
6.4 Scatter-Plot-Matrix der Prüfgliedmittelwerte und Residuen ... 11
6.4.1 Aufgabe ... 11
6.4.2 Nutzerführung ... 12
6.4.3 Output ... 12
6.5 Partielle Korrelationen der Mittelwerte und Residuen ... 13
6.5.1 Aufgabe ... 13
6.5.2 Nutzerführung ... 13
6.5.3 Output ... 14
6.6 Cluster im Lageplan und Boxplots ... 15
6.6.1 Aufgabe und methodischer Ansatz ... 15
6.6.2 Nutzerführung ... 15
6.6.3 Output ... 17
7 Interpretation und Nutzeffekte an Beispielen ... 20
7.1 Eingabefehler - Detektion durch die ‚univariate Analyse' ... 20
7.2 Parzelleneinkürzung und Korrektur während der Vegetation - Detektion durch die ‚univariate Analyse' ... 21
7.3 Erkennung eines bivariaten Problems (Zielmerkmal ‚Ertrag' und ‚Mangel'-Bonitur) - Detektion durch die Scatterplot-Matrix ... 22
7.4 Erkennung eines Drillfehlers - Detektion durch Boxplots ... 24

7.5 Multivariate Zusammenhangsstruktur - Detektion multivariater versus univariater Auffälligkeiten 26
7.6 Räumlich zusammenhängende Auffälligkeiten - Detektion durch Cluster im Lageplan und Boxplots 29
8 Prozeduren und Formeln, PIAFStat-Codierungen 30
8.1 Statistik der Einzelmerkmale - Plausi-Tabelle 30
8.2 Auffälligkeiten eines Merkmals im Lageplan 30
8.3 Partielle Korrelationsmatrix der Mittelwerte und der Residuen 31
8.4 Cluster im Lageplan und Boxplots 32
8.4.1 Faktorenanalyse 32
8.4.2 Clusteranalyse 34
9 Ausblick 38
10 Überleitung 38

Tabellenverzeichnis **Seite**

Tabelle 1: Messwerte der auffälligen Prüfglieder über alle Wiederholungen 28
Tabelle 2: Kennzeichnung des Signifikanzniveaus 30
Tabelle 3: Klassifizierung und Farbdarstellung der standardisierten Residuen (StdRes) 30
Tabelle 4: Partielle Korrelationskoeffizienten zwischen Merkmalspaaren 31
Tabelle 5: Auszug aus der Faktoranalyse - Erklärung der Faktoren durch Merkmalskoeffizienten bei einem Maisversuch 33
Tabelle 6: Auszug aus der Faktoranalyse - Eigenwerte der Faktoren und erklärte Varianz 33

Abbildungsverzeichnis **Seite**

Abbildung 1: Auswahl des Verfahrens PLAUSI 2
Abbildung 2: Zuordnungen der A-Merkmale 3
Abbildung 3: Zuordnung der Optionen 4
Abbildung 4: Dialogbox des optionalen Blockes ‚Ergebnis-Ausgabe' 4
Abbildung 5: Unteroptionen und optionale Blöcke der Multivariaten Analysen 5
Abbildung 6: Option ‚Plausi-Tabelle' mit Dialogbox des optionalen Blockes ‚Detektionsgrenze' 6
Abbildung 7: Univariate Analyse (Tabelle) 7
Abbildung 8: Unteroption ‚Auffälligkeiten je Merkmal im Lageplan' mit optionalen Blöcken 10
Abbildung 9: ‚Auffälligkeiten je Merkmal im Lageplan' 11
Abbildung 10: Unteroption ‚Scatter-Plot-Matrix' 12
Abbildung 11: ‚Scatter-Plot-Matrix' in einem Maisversuch - links für Mittelwerte, rechts für Residuen 12
Abbildung 12: Unteroption ‚partielle Korrelationsmatrix der Merkmale und der Residuen; optionaler Block ‚formatiert große Korrelationen' 13
Abbildung 13: Partielle Korrelationen über die Prüfgliedmittelwerte - Beispiel für einen multiplen Zusammenhangskomplex bei Silomais 14
Abbildung 14: Multivariater Komplex zwischen Qualitätsmerkmalen bei Silomais 14
Abbildung 15: Unteroption ‚Clusterung' und optionale Blöcke 16

Abbildung 16: Cluster im Lageplan - räumlicher Zusammenhang 17
Abbildung 17: Boxplots der Residuen je Cluster für einen Silomaisversuch 18
Abbildung 18: Protokoll zur Clusterung 19
Abbildung 19: Output der univariaten Analyse - möglicherweise ein Eingabefehler 20
Abbildung 20: Eingabemaske in PIAF - ein Eingabefehler (möglicherweise ein ‚Zahlendreher') 21
Abbildung 21: Output der univariaten Analyse - teilweise zu geringe Pflanzenzahl nach dem Vereinzeln auf eine definierte Soll-Bestandesdichte 21
Abbildung 22: Merkmalsdaten im Lageplan in PIAF - auffällige Parzellen mit zu geringer Bestandesdichte 22
Abbildung 23: Output der univariaten Analyse - eine Parzelle zweimal auffällig 22
Abbildung 24: Scatter-Plot-Matrix und Erkennung eines bivariaten Problems 23
Abbildung 25: Problem erklärender Eintrag im Textbericht als finale Entscheidungsgrundlage 23
Abbildung 26: Univariate Analyse für zwei Merkmale 24
Abbildung 27: ‚Cluster im Lageplan' - zwei auffällige Parzellen 24
Abbildung 28: ‚Cluster nach Boxplots ' - zwei auffällige Parzellen 25
Abbildung 29: Klärender Eintrag im Textbericht → Sortenvertauschung 25
Abbildung 30: Univariate Analyse für Merkmale, die eine Zusammenhangsstruktur aufweisen 26
Abbildung 31: Scatter-Plot-Matrix für Merkmale, die eine Zusammenhangsstruktur aufweisen; links Mittelwerte, rechts Residuen 26
Abbildung 32: Cluster im Lageplan - Erkennung einer multivariat betroffenen Parzelle 27
Abbildung 33: Boxplot (Typ ‚Residuen je Variable') - Unterschied zwischen univariat und multivariat auffälligen Parzellen 28
Abbildung 34: Programmlauf der multivariaten Analyse - Clusterung 32
Abbildung 35: Clusterungsgüte (CCC) in Abhängigkeit von der Anzahl gebildeter Cluster 34
Abbildung 36: Beispiel der Ausgabe im List-Fenster von PIAFStat bei Abbruch der Clusteranalyse 36
Abbildung 37: Dendogramm einer Clusterung (Methode: SINGLE) - Analog zum Beispiel in 6.6.3 und 7.6 37

Abkürzungsverzeichnis

A-Merkmale	Analyse-Merkmale
CCC	Cubic Cluster Criterion
EW	Einzelwerte
K-Merkmale	Klassifikations-Merkmale
LDS	Länderdienststellen
LFA	Landesforschungsanstalt
LSV	Landessortenversuch
MV	Mecklenburg-Vorpommern
MW	Mittelwerte

1 Zusammenfassung

Der biostatistische Verfahrenskomplex PLAUSI ist eine Softwarelösung im Planungs-, Informations- und Auswertungssystem für das Feldversuchswesen (PIAF). PIAF wird umfangreich von Versuchsanstellern der Länder und des Bundes, aber auch in der Privatwirtschaft eingesetzt.

PLAUSI dient dazu, im ersten Schritt der biometrischen und fachlichen Versuchsauswertung einen komprimierten, gestrafft strukturierten Gesamtüberblick über auswertungsrelevante Besonderheiten und Ereignisse eines Einzelversuches zu erhalten. Auffälligkeiten, die die Wiederholbarkeit, Reproduzierbarkeit und Interpretation negativ belasten könnten, sollen effizient, also schnell und sicher, erkannt werden. Insbesondere werden Entscheidungen über die Wertarbeit von Versuchsergebnissen bzw. einzelner Parzellen oder Versuchsblöcke u.ä. objektiviert. Zu diesem Zweck werden modernste biometrische Tools genutzt, wobei insbesondere die intuitive Problemdetektion mit grafisch-visuellen Mitteln im Vordergrund steht. PLAUSI ist seit der Ernte 2015 in vielen Länderdienststellen im Einsatz und hat sich bereits als Instrument zur Unterstützung der Plausibilitätsprüfung / des Qualitätschecks von Versuchsdaten bewährt. In 2016 und 2017 stehen noch einzelne Weiterentwicklungen an.

Das Verfahren PLAUSI wurde durch das Sachgebiet Sortenwesen und Biostatistik der LFA konzipiert. Die bundesweite PIAF-Gemeinschaft hat die Erstellung finanziert, die Umsetzung erfolgte in Zusammenarbeit zwischen der Fa. BioMath (Programmierung) und der LFA (Konzept und fachliche Begleitung). Es steht allen bundesweiten Nutzern, die an der Pflege und Weiterentwicklung von PIAF beteiligt sind, zur Verfügung. Der vorliegende Bericht ist gleichzeitig die Grundlage des Info-Textes für die Softwarelösung PLAUSI.

2 Einführung

Im angewandten Versuchswesen erfolgt die Auswertung der Feldversuche häufig unter Zeitdruck. So ist bei der Sortenprüfung von Winterungen der Zeitraum zwischen Ernte und Aussaat sehr gering, Entscheidungen über Empfehlung und Weiterprüfung einzelner Sorten etc. müssen aber sehr schnell getroffen werden. Beim Versuchsansteller laufen zudem etliche Versuche parallel auf, die alle zeitnah auszuwerten sind. Eine varianzanalytische Auswertung erfolgt i.d.R. nur für ein bzw. sehr wenige Merkmale (z.B. Ertrag). Weitere Merkmale werden lediglich tabellarisch ausgegeben, Merkmale zur Fehler-Indikation (z.B. Mängelbonituren) oft gar nicht berücksichtigt.

Vor der varianzanalytischen Auswertung wäre es für Bewertungen oder für Plausibilitäts-Checks oft wichtig, sich über die Gesamtheit der erfassten Merkmale einen Eindruck zu verschaffen. In vielen Fällen liegen über eine sehr große Zahl von Merkmalen Daten vor (fachlich relevante Merkmale, aber auch Störgrößen wie Mängelbonituren; Textbericht und Notizen, Skizzen und Fotos u.v.m.), die Hinweise auf die Auswertbarkeit eines Zielmerkmals geben könnten - oft werden solche Zusammenhänge aber in der Auswertung nicht wahrgenommen, da sich die Auswertung unter Zeitdruck auf jeweils nur ein Merkmal konzentriert.

Vor diesem Hintergrund hat die PIAF-Koordinierungsgruppe (Länder und Bund) unter Federführung der LFA ein Verfahren konzipiert und in Auftrag gegeben, welches die Arbeit der Auswerter beim Qualitätscheck der Versuchsdaten unterstützen soll.

3 Ziele der Entwicklung

- Entwicklung eines Software-Moduls für PIAFStat zur effizienten, nutzerfreundlichen automatischen Detektion von gestörten Werten oder regelrechten Falschwerten aus Parzellen und räumlich zusammenhängenden Parzellengruppen
  - Auflistung auffälliger Parzellen je Versuchsmerkmal auf Basis explorativer und schließender univariater Datenanalysen
  - multivariate Residual-Analyse auf Merkmalskomplexe
  - farbliche Parzellendarstellung der Residuen von Einzelmerkmalen und von multivariat definierten Clustern im Versuchs-Lageplan
- Unterstützung bei der Ursacheninterpretation von Auffälligkeiten in Einzelversuchen
- Bereitstellung von Informationen, die ggf. die fachliche Entscheidung zur Bereinigung der Versuchsdaten von falschen oder gestörten Daten objektivieren
- Verbesserung der Reproduzierbarkeit von aus Versuchen abgeleiteten fachlichen Aussagen
- Erstellung eines Info- bzw. Hilfe-Textes für das zu entwickelnde Verfahren in PIAFStat (im Wesentlichen in Form dieses Berichtes)

4 Nutzeroberfläche - Gesamtübersicht

Das Verfahren ‚PLAUSI – Vorauswertung zu Auffälligkeiten' wird in PIAFStat im Menüpunkt ‚SAS Ausführen' aus der Verfahrensbibliothek gewählt (Abbildung 1).

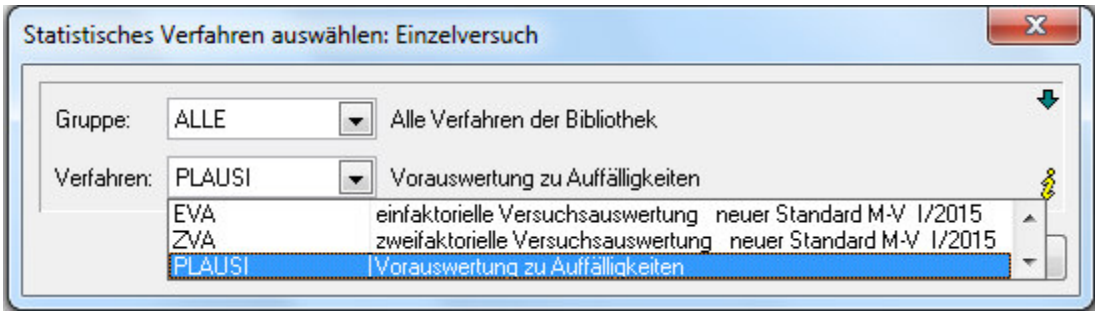


Abbildung 1: Auswahl des Verfahrens PLAUSI

Nach Auswahl des Verfahrens werden in PIAFStat Registerkarten mit interaktiven Eingabemöglichkeiten zur Gestaltung des Verfahrens geöffnet.

4.1 Klassifizierungs-Merkmale

Die Faktorenkombination für die Varianzanalyse wird voreingestellt durch den besonderen Platzhalter [F-ALL] definiert. Es ist daher nicht notwendig für die Klassifizierungs-Merkmale (K-Merkmale) eine eigene Registerkarte anzubieten, wie das in anderen PIAFStat-Verfahren sonst üblich ist (Abbildung 2 → Registerkarte K-Merkmale fehlt).

4.2 Analyse-Merkmale

Auf der Registerkarte A-Merkmale können zwei Merkmalsgruppen festgelegt werden. In der ersten Gruppe ‚**Merkmale für Plausi-Tabelle (univariat)**‘ (Abbildung 2) werden im rechten Bildschirmfenster die Merkmale ausgewählt, die in der univariaten Analyse (entspricht Option ‚Plausi-Tabelle‘; Abbildung 3) berücksichtigt werden sollen. Voreingestellt ist die automatische Auswahl aller Merkmale. Der Nutzer kann Merkmale ausschalten.

In der zweiten Gruppe **‚Merkmale für Plausi-Multivariat‘** (Abbildung 2) sind keine Merkmale voreingestellt zugewiesen. Für die multivariate Auswertung (Option ‚Plausi-Multivariat‘; Abbildung 3) ist hier eine bewusste interaktive Auswahl von Merkmalen erforderlich (Kapitel 6.2).

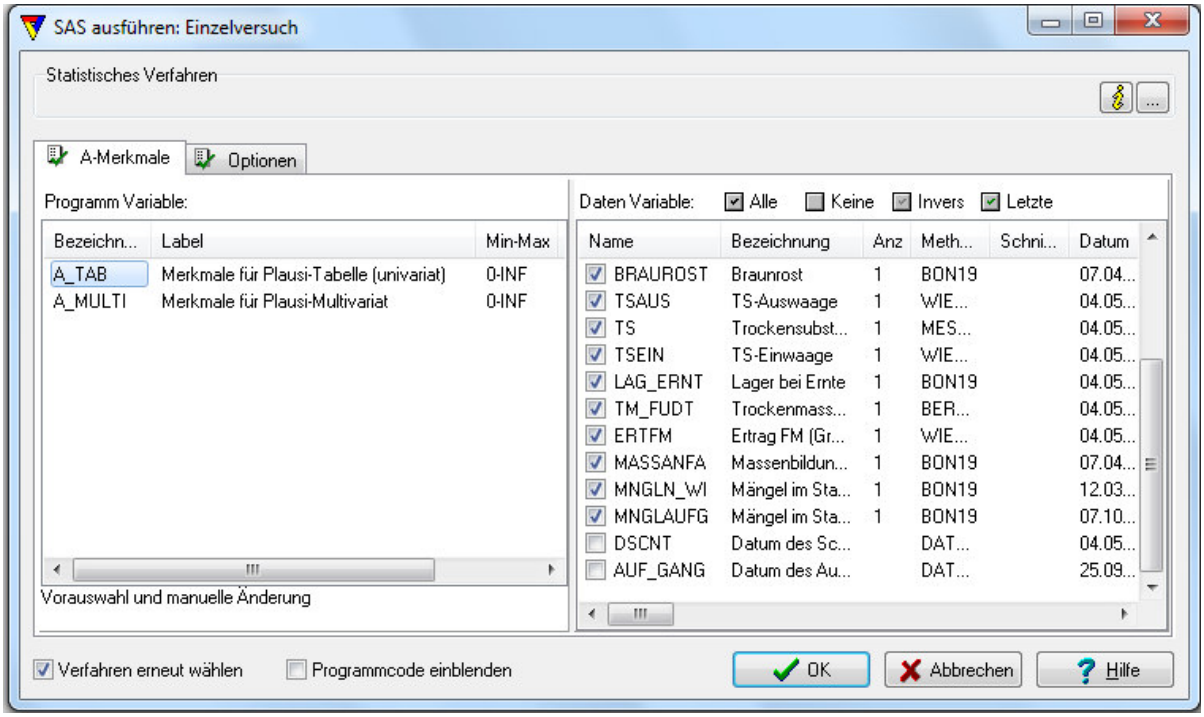


Abbildung 2: Zuordnungen der A-Merkmale

4.3 Optionen

Auf der Registerkarte Optionen (Abbildung 3) können Festlegungen getroffen und Analysen ausgewählt werden.

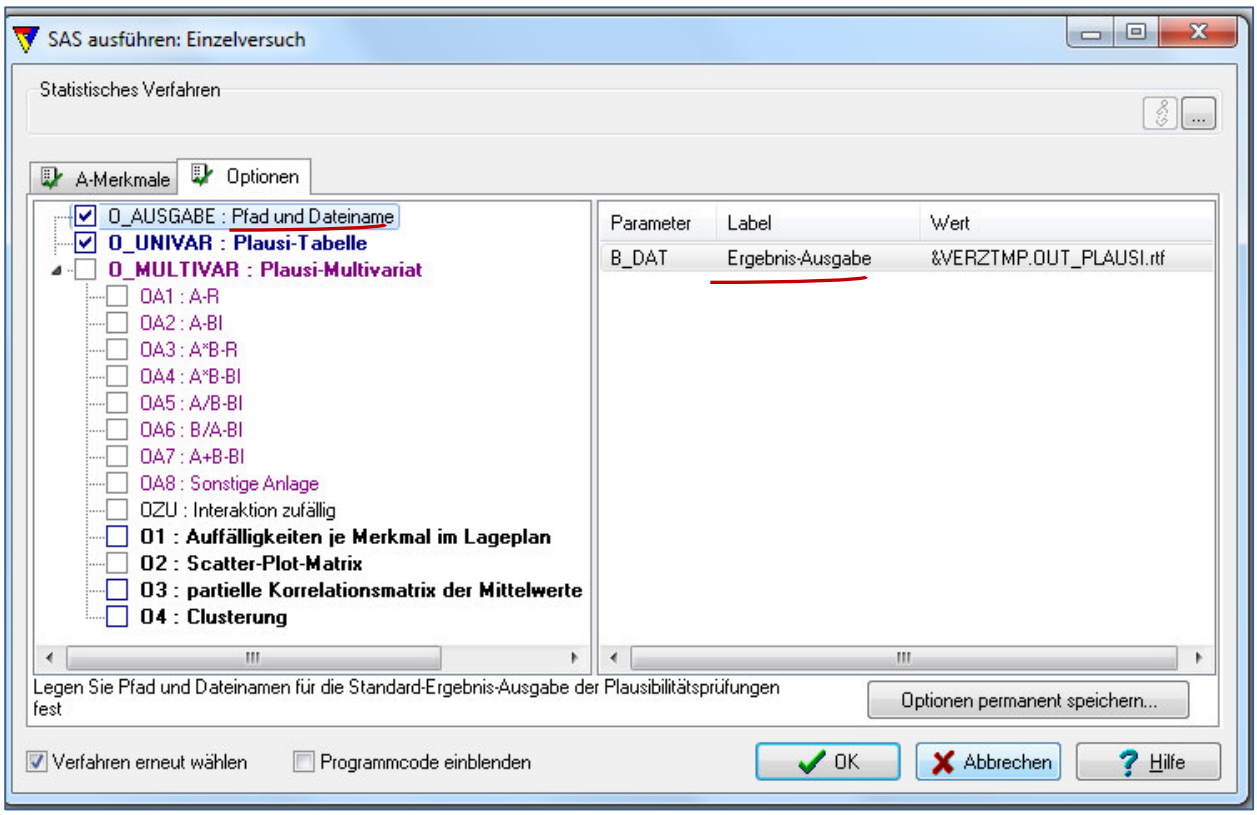


Abbildung 3: Zuordnung der Optionen

Durch die **Option ‚Pfad und Dateiname'** werden Festlegungen für die Ergebnis-Ausgabe-Datei getroffen. Mit Mausklick auf den optionalen Block ‚Ergebnis-Ausgabe' im rechten Bildschirmfenster wird eine Dialogbox geöffnet (Abbildung 4). In dieser Dialogbox kann der Anwender Pfad und Dateinamen ändern.

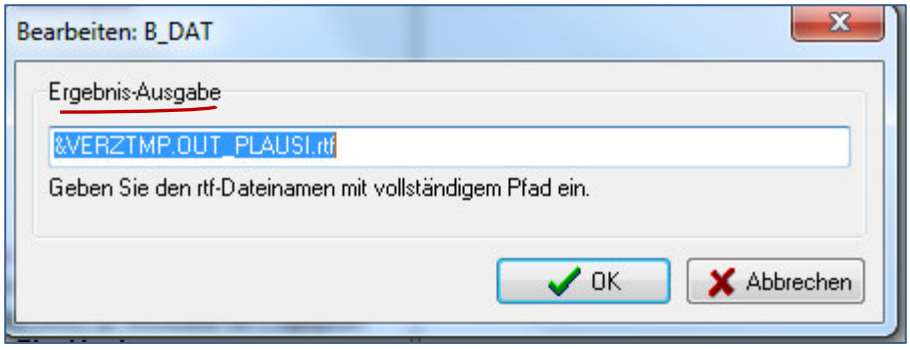


Abbildung 4: Dialogbox des optionalen Blockes ‚Ergebnis-Ausgabe'

Hinweis für Nutzer: Die Verwendung von in PIAFStat definierten globalen Variablen (Makrovariablen) garantiert eine automatische Anpassung der Einträge an eigene Strukturen. Mit der Makrovariable &VERZTMP wird z.B. das in den PIAFStat-Optionen angegebene temporäre Verzeichnis als Pfad festgelegt. Weitere Informationen zu Makrovariablen in PIAFStat können im PIAFStat-Handbuch unter 2.3.1 eingesehen werden. Die Option ‚Pfad und Dateiname' ist nicht deaktivierbar.

Durch Aktivierung der **Option ‚Plausi-Tabelle'** (Abbildung 3) wird die univariate Analyse – Statistik der Einzelmerkmale (Kapitel 5) angewiesen. Bei Mausklick auf die Option ‚Plausi-Tabelle' erscheint im rechten Bildschirmfenster ein optionaler Block ‚Detektionsgrenze' (Kapitel 5.2).

In der **Option ‚Plausi-Multivariat'** können die Analysen ausgewählt werden, mit denen multivariate Zusammenhangsstrukturen zwischen Einzelmerkmalen erkannt werden können. Die mutivariate Analyse beinhaltet vier Unteroptionen (Abbildung 5), die wahlweise, auch einzeln, aktiviert oder deaktiviert werden können. Die Unteroptionen enthalten bis zu fünf optionale Blöcke, in denen der Anwender Festlegungen zu den Analysen treffen kann (Abbildungen 5,8,10,12,15).

Aufgaben, Outputs und Interpretationshilfen zu den einzelnen Optionen/Unteroptionen sowie Erläuterungen zu den optionalen Blöcken werden in den Kapiteln 5, 6 und 7 bereitgestellt.

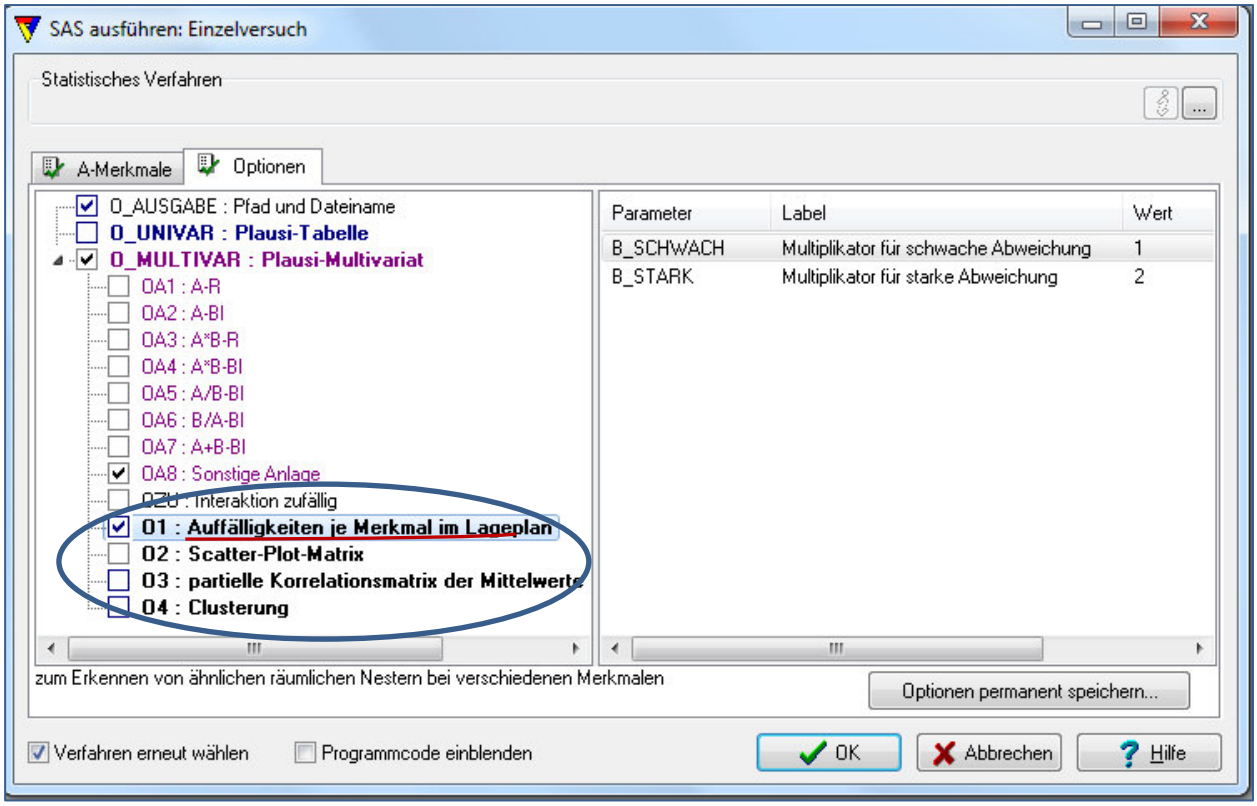


Abbildung 5: Unteroptionen und optionale Blöcke der Multivariaten Analysen

5 Univariate Analyse - Statistiken der Einzelmerkmale

5.1 Aufgabe

Im Ergebnis der Abarbeitung dieser Option wird eine komprimierte Übersicht über die erfassten Merkmale mit einer deskriptiven Beschreibung und einfachen Tests für jedes Einzelmerkmal bereitgestellt. Diese Übersicht soll dem Auswerter sowohl in statistischer wie auch in fachlicher Hinsicht auf Auffälligkeiten im Versuch oder auf regelrechte Falschwerte (Zahlendreher, Dezimalstellenfehler...) hinweisen, bevor weitere vertiefte z.B. varianzanalytische Analysen einzelner ausgewählter Merkmale erfolgen.

Unter 7.1 ist eine typische Situation dargelegt, in der durch diese Tabelle Fehler leicht detektiert werden können.

5.2 Nutzerführung

In Abbildung 6 zeigt das gesetzte Häkchen bei der Option ‚Plausi-Tabelle' an, dass die univariate Analyse aktiviert ist. Durch Klick des Cursors auf diese Option erscheint im rechten Bildschirmfenster der optionale Block ‚Detektionsgrenze'.

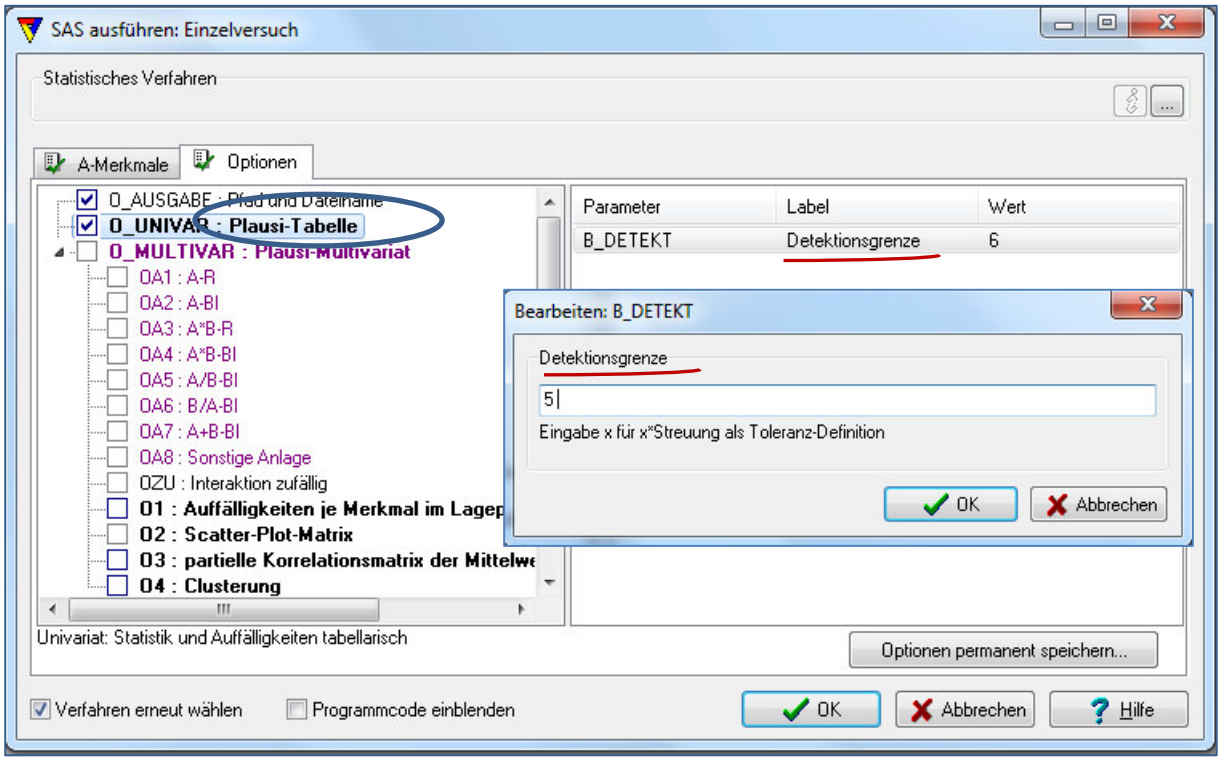


Abbildung 6: Option ‚Plausi-Tabelle' mit Dialogbox des optionalen Blockes ‚Detektionsgrenze'

Parzellen, deren Residuen außerhalb der Spanne $< Mittelwert \pm x * Streuung >$ liegen, werden in der Ergebnis-Tabelle als ‚auffällig' ausgewiesen (Kapitel 5.3 und Abbildung 7). Als Voreinstellung für die Detektionsgrenze wird $x = 5$ vorgeschlagen. Soll x verändert werden, wird durch Klick auf den optionalen Block ‚Detektionsgrenze' die Dialogbox geöffnet (Abbildung 6) und x kann überschrieben werden - temporär oder bei Aktivierung des Buttons 'Optionen permanent speichern' dauerhaft.

5.3 Output und allgemeine Hinweise zu ausgewiesenen Kennzahlen

Die Statistik jedes A-Merkmals (Abbildung 7) umfasst:

- N, Min und Max der Einzelwerte (EW),
- N, Min, Median und Max der Prüfgliedmittelwerte (MW),
- s% und F-Wert der Varianzanalyse (der F-Wert wird bezüglich der Signifikanzniveaus mit Sternen gekennzeichnet),
- Auflistung der Parzellen (Parzellenkürzel), deren Betrag der standardisierten absoluten Residuen größer als eine vorgegebene Grenze ist.

Versuch: LSV KMais Vipperow 2014 / regionale Sorteneignung
Univariate Analyse: Statistiken der A-Merkmale
Der F-Test je Merkmal erfolgte ohne Prüfung der Vorrausetzungen.
Als Auffälligkeitsgrenze wurde verwendet: |std.Res| > 5 .

	Dat	EW N	Min	Max	MW N	Min	Med	Max	s%	FWert	auf-fällig
Abreife Blt 1-9	03.09	48	1.0	7.0	12	2.3	2.9	4.5	46.2	1.1	
Anz.Pfl.Bezugsb	03.09	48	84.0	94.0	12	88.0	91.1	93.0	3.1	1.2	
Beulenbrand %	03.09	48	0.0	2.0	12	0.0	0.0	0.5	516.4	0.9	5\3
Beulenbrand Pfl.	03.09	48	0.0	2.0	12	0.0	0.0	0.5	516.4	0.9	5\3
Maiszünsler %	03.09	48	0.0	7.0	12	0.3	1.5	3.8	101.2	1.5	
Maiszünsler Pfl.	03.09	48	0.0	6.0	12	0.3	1.5	3.3	98.2	1.4	
T.Auss-Ernte	27.10	48	178.0	178.0	12	178.0	178.0	178.0	0.0		
Pfl. n Vereinzel	03.09	48	84.0	94.0	12	88.0	91.1	93.0	3.1	1.2	
Pfl/qm n Vereinz	03.09	48	7.2	7.8	12	7.6	7.7	7.7	1.8	0.7	4\1
Pflanzenlänge cm	03.09	48	250.0	315.0	12	271.0	283.5	304.3	3.4	4.8***	
Kornertrag kg/Pz	27.10	47	15.0	20.0	12	16.5	18.0	19.0	5.1	4.1***	
Kornertrag dt/ha	27.10	47	101.4	142.2	12	110.1	124.7	131.3	5.8	3.1**	
TS- Korn %	27.10	47	68.0	72.9	12	68.6	70.6	72.1	1.3	4.8***	

im oberen Tabellen-Bereich: Indikator-Merkmale für mögliche Störungen
im unteren Tabellen-Bereich: Ziel-Merkmale

Abbildung 7: Univariate Analyse (Tabelle)

Durch die im Output ausgewiesenen Maßzahlen der deskriptiven Statistik für Einzelwerte und Prüfgliedmittelwerte bekommt man leicht einen Überblick über Besonderheiten im Versuch (aufgetretene Krankheiten, Lager u.v.m.). Auch regelrechte Falschwerte (Zahlendreher, Dezimalstellenfehler, weitere Eingabe- oder Messfehler) werden leicht und intuitiv über alle Merkmale hinweg anhand dieser Maßzahlen erkannt.

Darüber hinaus lassen sich an geeigneten Indikatormerkmalen auch Auffälligkeiten ausmachen, die u.U. die Auswertbarkeit tangieren könnten: Ist die Spanne der Einzelwerte weit größer als die der Prüfgliedermittelwerte? Liegt der Median (Zentralwert) nahe beim Minimum oder beim Maximum der Einzelwerte oder Mittelwerte? → Dies zeigt dann an, dass es keine allgemeine symmetrische Variation gibt, sondern eher einzelne wenige Einzelparzellen oder Prüfglieder von der Norm abweichen.

Zur Berechnung des Variationskoeffizienten (s%) und für den F-Test erfolgt keine automatische Prüfung der Modellvoraussetzungen. Ziel ist auch keine Vorwegnahme einer ordnungsgemäßen Varianzanalyse mit Modellierung entsprechend dem Versuchsdesign, sondern erste Hinweise auf mögliche Auffälligkeiten zu erhalten.

Der Variationskoeffizient (s%) gibt Hinweise darauf, ob die Wiederholungen je Prüfglied (zumindest teilweise) stark vom zugehörigen Prüfgliedmittelwert abweichen. Der F-Test zeigt an, ob es signifikante Prüfgliedunterschiede gibt. Bei einzelnen Merkmalen ist dies zu erwarten (Abbildung 7, die Merkmale unter dem Strich) - hier bringt u.U. die adäquate Varianzanalyse erst den korrekten Test. Interessanter für die Plausibilitätsprüfung ist es, wenn Signifikanzen wider Erwarten bei Merkmalen auftreten, die in diesem Kontext eher als Störgrößen anzusehen sind (in Abbildung 7, die Merkmale über dem Strich).

Letztlich werden Parzellen als ‚auffällig' ausgewiesen, deren Residuen außerhalb der Spanne $< Mittelwert \pm x * Streuung >$ liegen. Aus pragmatischer erfahrungsbasierter Sicht wird als (optionaler) Default $x = 5$ vorgeschlagen. Für einzelne, insbesondere für normalverteilte Merkmale wie Ertrag könnte x kleiner gewählt werden. Dies würde aber in der Gesamtübersicht z.B. bei extrem schief verteilten ordinalskalierten Merkmalen zu einer unübersichtlichen großen Zahl auffälliger Werte führen. Letztlich kann dies im Nachgang ohnehin in der eigentlichen Varianzanalyse unter Berücksichtigung des Versuchsdesigns erfolgen - hier sollen zunächst nur besonders grobe Auffälligkeiten und insbesondere mögliche Zusammenhänge zwischen den Merkmalen detektiert werden → sind einzelne Parzellen in mehreren Merkmalen auffällig? Oft gibt es logische Erklärungen dafür, welche bei der weiteren Auswertung berücksichtigt werden sollten. Der Terminus ‚Ausreißer' wurde hier bewusst vermieden, weil zum einen eine Wahrscheinlichkeitsaussage hierüber nicht als hilfreich angesehen wird und zum anderen keinesfalls suggeriert werden soll, dass auffällige Werte zu streichen sind. Vielmehr sollen sie nur Veranlassung geben, vor der Varianzanalyse noch einmal über die Plausibilität im Gesamtkontext aller Merkmale nachzudenken und ggf. auch mal Daten zu hinterfragen und zu kontrollieren.

Unter 7.1 ist eine typische Situation dargelegt, in der durch diese Tabelle Fehler leicht detektiert werden können.

6 Multivariate Analysen

6.1 Allgemeiner Hintergrund

Oft stehen Merkmale hinsichtlich der Wirkung von Störgrößen in engem Zusammenhang miteinander. Manchmal lässt erst eine multivariate Betrachtung der parzellenbezogenen Residuen systematische Probleme / Auffälligkeiten einzelner Parzellen oder Parzellengruppen (z.B. Trockenstelle im Versuch) klar erkennen. So könnte z.B. beim Silomais eine räumlich zusammenhängende Parzellengruppe, deren Residuen durch den Merkmalskomplex {kurz, hoher Trockensubstanzgehalt, wenig Masse, hoher Absterbegrad} auf eine Trockenstelle hindeuten. Die Kombination {hohe Trockensubstanz, viel Zucker, wenig Stärke} wäre ein Hinweis auf eine wenig abgereifte Parzelle (im Vergleich der Wiederholungen des gleichen Prüfgliedes). Haben benachbarte Parzellen die gleiche komplexe Situation, dann ist dies ein Hinweis auf räumliche Probleme in der Versuchsfläche (Heterogenität). Aufgabe nachfolgender multivariater Analysen ist es, derartige Zusammenhangsstrukturen automatisch zu erkennen und ggf. auffällige Strukturen im Versuch effizient zu detektieren und nutzerfreundlich anzuzeigen. Voraussetzung ist eine geeignete Merkmalsauswahl für diese Analyse durch den Auswerter (s. 6.2).

6.2 Auswahl der Merkmale für eine multivariate Auswertung

Die hier bereitgestellten Tools sind nur zielführend, wenn die Auswahl der Merkmale für die multivariate Auswertung begründet erfolgt! Eine automatische Auswahl vieler (≥ 5) oder beliebiger Merkmale lässt kaum geeignete Outputs und Interpretationen erwarten!

Eine zielführende Merkmalsauswahl setzt Vertrautheit mit den Gegebenheiten des konkreten Versuches und pflanzenbauliches Wissen über die Reaktionsnormen voraus. Veranlassung für die multivariate Auswertung und für die konkrete Merkmalsauswahl könnten z.B. Beobachtungen im Verlauf der Vegetation oder auch die Sichtung der Tabelle aus der univariaten Analyse sein (2015 z.B. beim Befall von Wintergetreide mit Gerstengelbverzwergung möglicherweise Zusammenhänge zwischen Befalls- und Ertrags-Residuen). Andererseits wird in MV z.B. bei Mais der Merkmalskomplex {Ertrag, TS-Gehalt, Stärke, Zucker, Energiekonzentration, Pflanzenlänge} regelmäßig gecheckt, da hier a priori Zusammenhangsstrukturen - nicht nur bei Prüfgliedwirkungen, sondern auch bei Residuen - zu erwarten sind. Die multivariate Auswertung erhöht hier die Chance, ggf. Problemparzellen und -bereiche zu erkennen oder zu bestätigen.

Beispiele für Motivationen für die Nutzung multivariater Ansätze (MULTIVAR) und Ansätze zur Auswahl einzubeziehender Merkmale:

- korrelierte Zielmerkmale, z.B. Trockensubstanz-, Stärke- und Zuckergehalt bei Silomais - motiviert aus fachlicher Erwartung,
- Ziel- und Indikatormerkmal, z.B. Ertrag und Mangel oder Ertrag und Pflanzenlänge - motiviert aus UNIVAR + Versuchsunterlagen,
- Zielmerkmale und erhebliche Störgrößen, z.B. Ertrag und Auswinterung von Nachbarparzellen, motiviert aus Wissen um Versuchsbesonderheiten und UNIVAR + Versuchsunterlagen,
- Zielmerkmal und mögliche stetige Kovariable, z.B. Ertrag und Anzahl Maulwurfshaufen - motiviert aus UNIVAR + Versuchsunterlagen → ggf. Vorentscheidung, Kovariable in die nachfolgende varianzanalytische Auswertung einzubeziehen.

6.3 Residuen der Einzelmerkmale im Lageplan

6.3.1 Aufgabe

Diese Option soll die eigentliche multivariate Auswertung mit zunächst univariaten Betrachtungen der für die multivariate Auswertung ausgewählten Merkmale einleiten. Insbesondere soll durch die Darstellungen der Residuen im Lageplan visualisiert werden, ob es zwischen einzelnen Merkmalen Parallelen in der räumlichen Verteilung von Auffälligkeiten gibt bzw. ob Einzelparzellen in verschiedenen Merkmalen besonders auffällig sind.

6.3.2 Nutzerführung

Das Setzen eines Häkchens bei der Unteroption ‚Auffälligkeiten je Merkmal im Lageplan' aktiviert diesen Teil der multivariaten Analysen (Abbildung 8). Über zwei optionale Blöcke können Festlegungen getroffen werden, die die Grenzen für die farblich zu markierenden Parzellen im Lageplan betreffen.

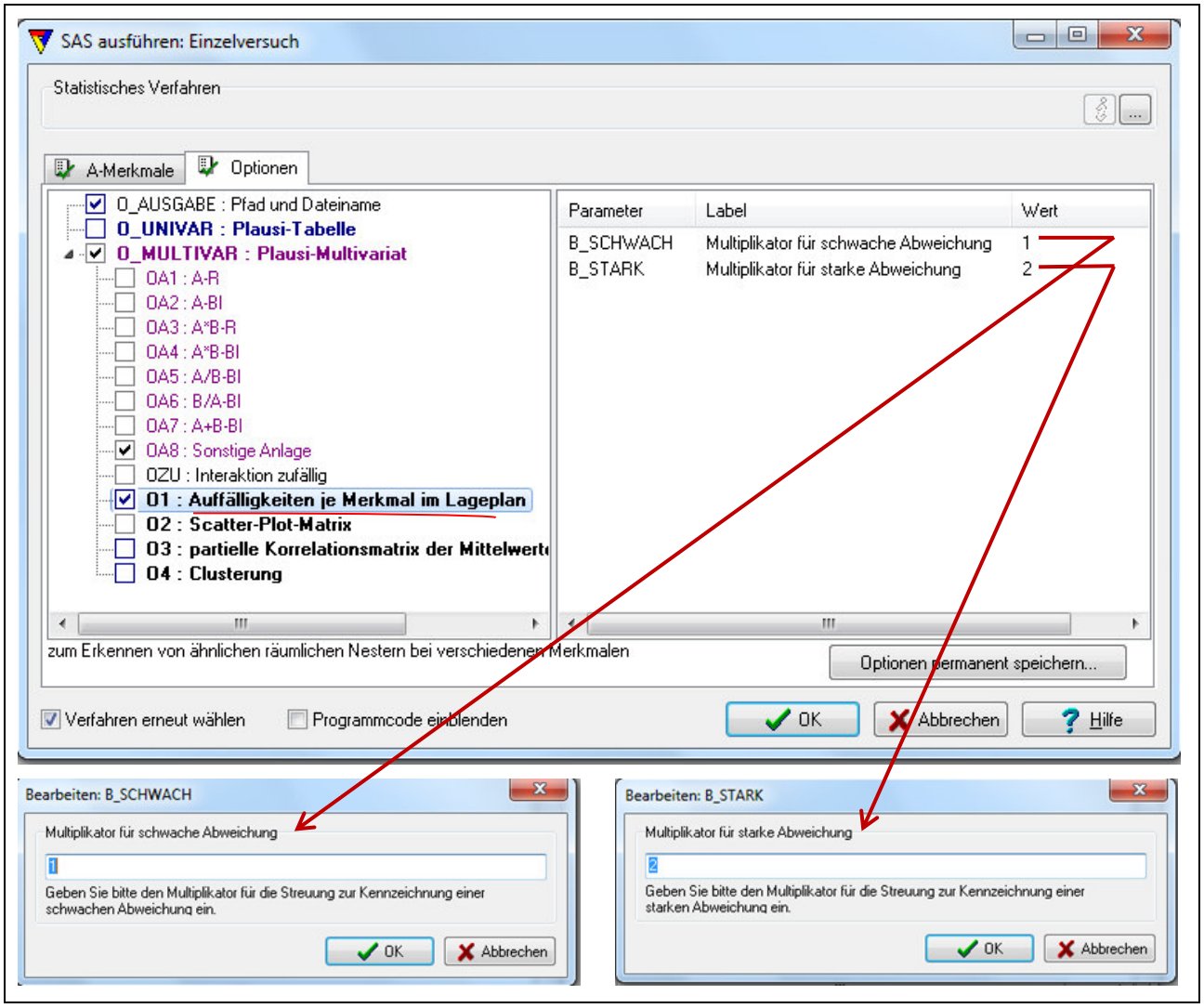


Abbildung 8: Unteroption ‚Auffälligkeiten je Merkmal im Lageplan' mit optionalen Blöcken

6.3.3 Output

Es ist deutlich zu sehen, dass bei den Merkmalen Dürreschaden, Pflanzenlänge, Grünmasseertrag und Trockensubstanzgehalt eine ähnliche größere Stelle in der Fläche auffällig ist (Abbildung 9). Im Detail gibt es allerdings gewisse Abweichungen in der Ausdehnung. Unter 7.6 wird dieses konkrete Beispiel tiefgehender multivariat analysiert und gezeigt, wie diese Parzellengruppe aus multivariater Sicht distinkt abzugrenzen wäre.

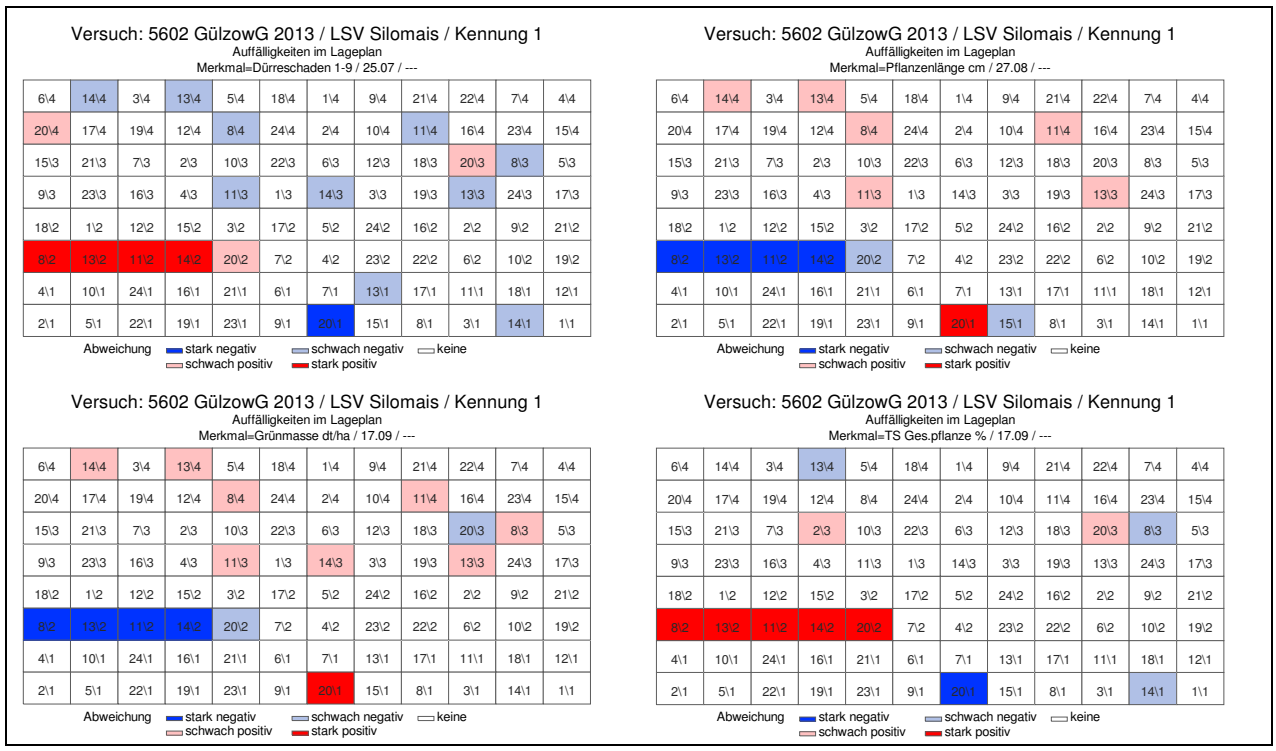


Abbildung 9: ‚Auffälligkeiten je Merkmal im Lageplan'

6.4 Scatter-Plot-Matrix der Prüfgliedmittelwerte und Residuen

6.4.1 Aufgabe

Sowohl für die Prüfgliedmittelwerte als auch für die Residuen (auf Parzellenebene) wird eine Scatter-Plot-Matrix erstellt (Abbildung 11). Die Scatter-Plots auf den Nicht-Diagonal-Feldern visualisieren für alle Merkmalspaare bivariate Zusammenhänge. Sie sollen erkennen lassen, ob es paarweise Korrelationen gibt und ob einzelne Mittelwerte bzw. einzelne Parzellen erhebliche Korrelationsbrecher sind. Liegen auffällige Punkte auf beiden Achsen am Rande der jeweiligen Verteilung (nahe der Eckpunkte des Quadrates), so ist das ein Hinweis auf eine bivariate Zusammenhangsstruktur. Ist die Verteilung insgesamt schmal elliptisch, so besteht zwischen den Merkmalen ein enger korrelativer Zusammenhang - auf Mittelwertebene (fachlicher Natur) oder auch auf Residual-Ebene (ggf. Hinweis auf systematische Störungen).

Die Histogramme auf der Diagonale sind univariater Natur und zeigen an, ob die Einzelmerkmale gut an die Normalverteilung angepasst sind, ob sie Schiefe aufweisen und ob einzelne Mittelwerte oder Parzellen-Residuen erheblich von der allgemeinen Verteilung abweichen.

Die Darstellung auf Prüfgliedmittelwertebene sollte eher fachlich interpretiert werden (Sind die Zusammenhänge wie fachlich erwartet bzw. sind sie erklärbar?), während die Darstellung auf Parzellenebene (Residuen) eher Störungen und deren Zusammenhangsstrukturen anzeigen soll.

6.4.2 Nutzerführung

Sollen die Scatter-Plots ausgegeben werden, wird die Option ‚Scatter-Plot-Matrix' angehakt (Abbildung 10). Weitere Einstellungen des Anwenders sind nicht notwendig.

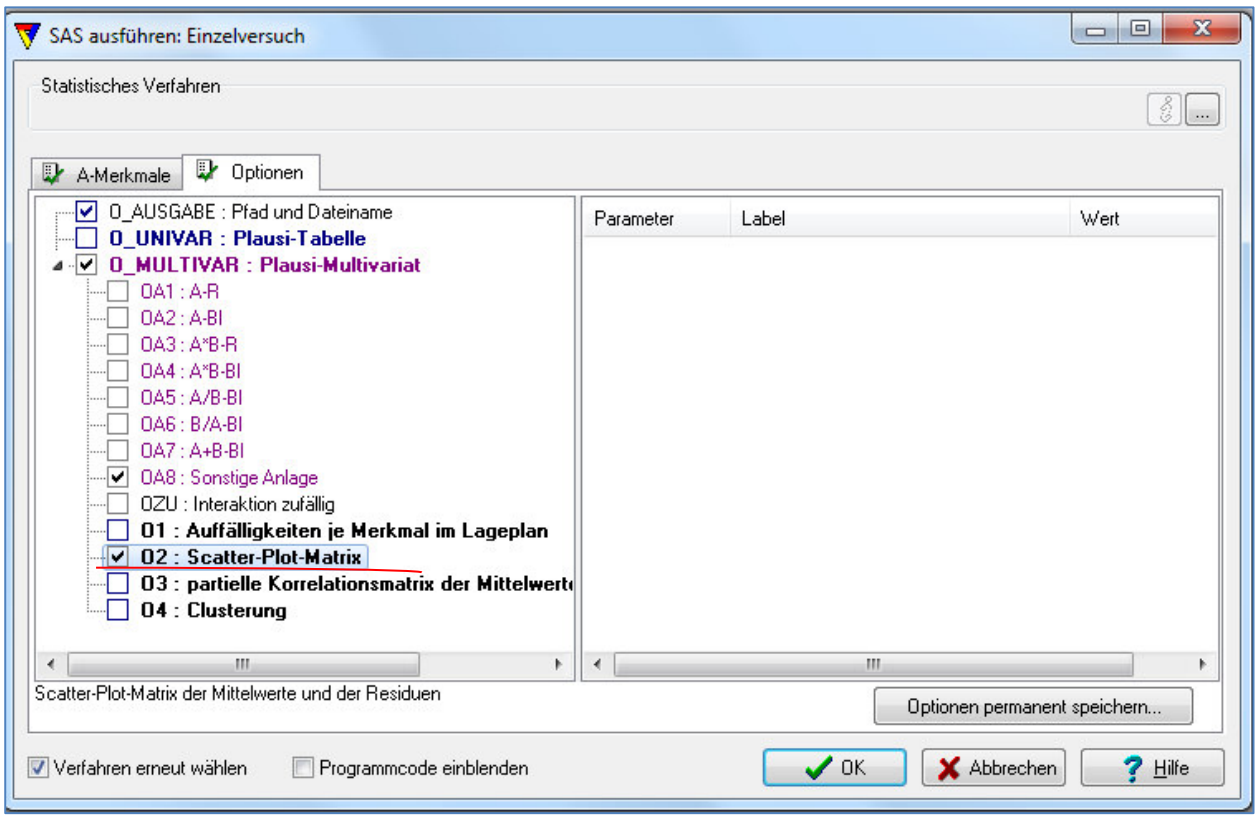


Abbildung 10: Unteroption ‚Scatter-Plot-Matrix'

6.4.3 Output

Im Kapitel 7.3 wird an einem realistischen Beispiel die Detektion von Störungen mit Unterstützung durch diese Darstellung aufgezeigt.

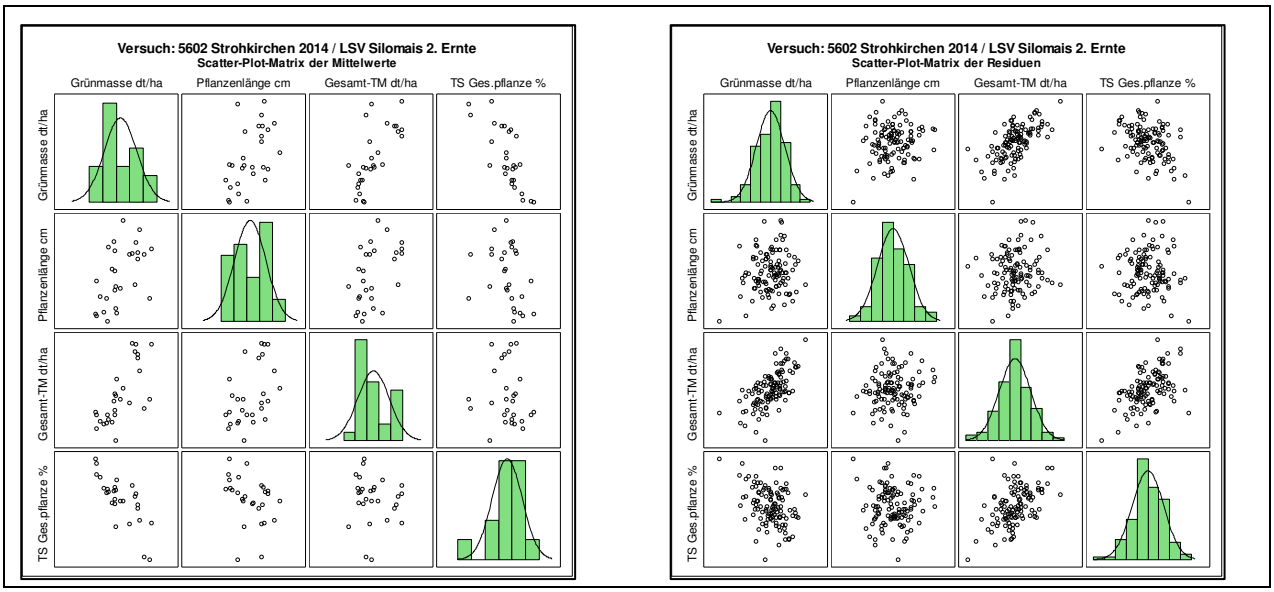


Abbildung 11: ‚Scatter-Plot-Matrix' in einem Maisversuch - links für Mittelwerte, rechts für Residuen

6.5 Partielle Korrelationen der Mittelwerte und Residuen

6.5.1 Aufgabe

Für die Mittelwerte und für die Residuen werden jeweils die partiellen Korrelationskoeffizienten zwischen Merkmalspaaren berechnet. Dies ist eine Ergänzung zur Scatter-Plot-Matrix. Im Gegensatz zu den paarweisen Korrelationen des Scatter-Plots handelt es sich beim partiellen Koeffizienten aber um ein Zusammenhangsmaß, das bei Ausschaltung der Wirkung aller anderen einbezogenen Merkmale ‚übrig' bleibt. Die partiellen Korrelationskoeffizienten erhöhen gegenüber den aus dem Zusammenhang gelösten paarweisen Koeffizienten die Chance, die maßgeblichen, u.U. ursächlichen Korrelationen zu ‚isolieren' und zu interpretieren.,

Die Korrelationskoeffizienten werden tabellarisch dargestellt. Diese Koeffizienten werden in zwei Größen-Gruppen klassifiziert, wobei die Gruppe mit den stärkeren Zusammenhängen farblich herausgehoben wird (Abbildung 13).

6.5.2 Nutzerführung

Durch Aktivierung der Option ‚partielle Korrelationsmatrix der Mittelwerte und der Residuen' (Abbildung 12) können partielle Korrelationskoeffizienten ausgegeben werden. Im zugehörenden optionalen Block ist für die Gruppierung in hohe versus geringe Zusammenhänge standardmäßig die Clusterungsmethode WARD als Default vorgeschlagen.

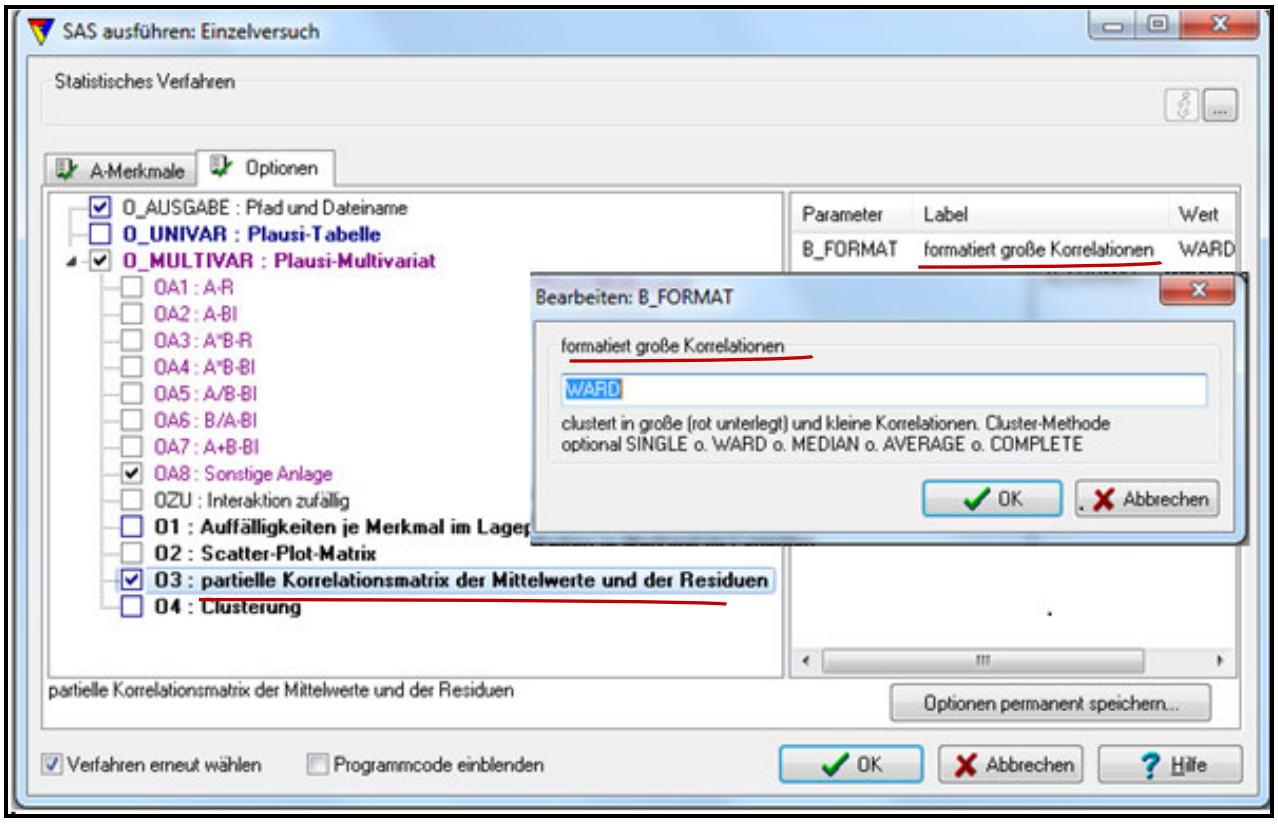


Abbildung 12: Unteroption ‚partielle Korrelationsmatrix der Merkmale und der Residuen; optionaler Block ‚formatiert große Korrelationen'

6.5.3 Output

Die in Abbildung 13 dargestellten partiellen Korrelationen entsprechen dem im Kapitel 7.5 erläuterten realistischen Beispiel für eine multivariate Zusammenhangsstruktur. In Abbildung 31 sind im Gegensatz zu diesen partiellen Korrelationskoeffizienten die paarweisen Korrelationen grafisch in einer Scatter-Plot-Matrix dargestellt. Die hohe Korrelation zwischen Energiekonzentration (NEL je kg Trockenmasse) und Stärke ist in beiden Darstellungen erkennbar. Auch der Zusammenhang zwischen Stärke und Zucker ist bereits in der Scatter-Plot-Matrix erkennbar. Dagegen wird der Zusammenhang zwischen Zucker und Energiekonzentration erst auf der Ebene der partiellen Korrelation erkannt (Abbildung 13). In der paarweisen Korrelation wird er durch die gegenläufige Wirkung des Stärkegehaltes maskiert (Abbildung 14 und 31). Während die paarweisen Korrelationen eher den wahrgenommenen Beziehungen zwischen jeweils zwei Variablen entsprechen, gelangt man über die partiellen Korrelationen häufiger zu den ursächlichen Erklärungen.

Versuch: 5702 Gülzow 2014 / LSV 3. Ernte

partielle Korrelation der Mittelwerte

part. Korr. der Mittelwerte	NEL /kg TM (XL)	Red. Zucker	Stärke-gehalt %	TS Ges.-pflanze %
NEL /kg TM (XL)	.	.	.	.
Red. Zucker	0.51138	.	.	.
Stärkegehalt %	0.81007	-0.72776	.	.
TS Ges.pflanze %	0.16929	-0.29003	-0.30793	.

Abbildung 13: Partielle Korrelationen über die Prüfgliedmittelwerte - Beispiel für einen multiplen Zusammenhangskomplex bei Silomais

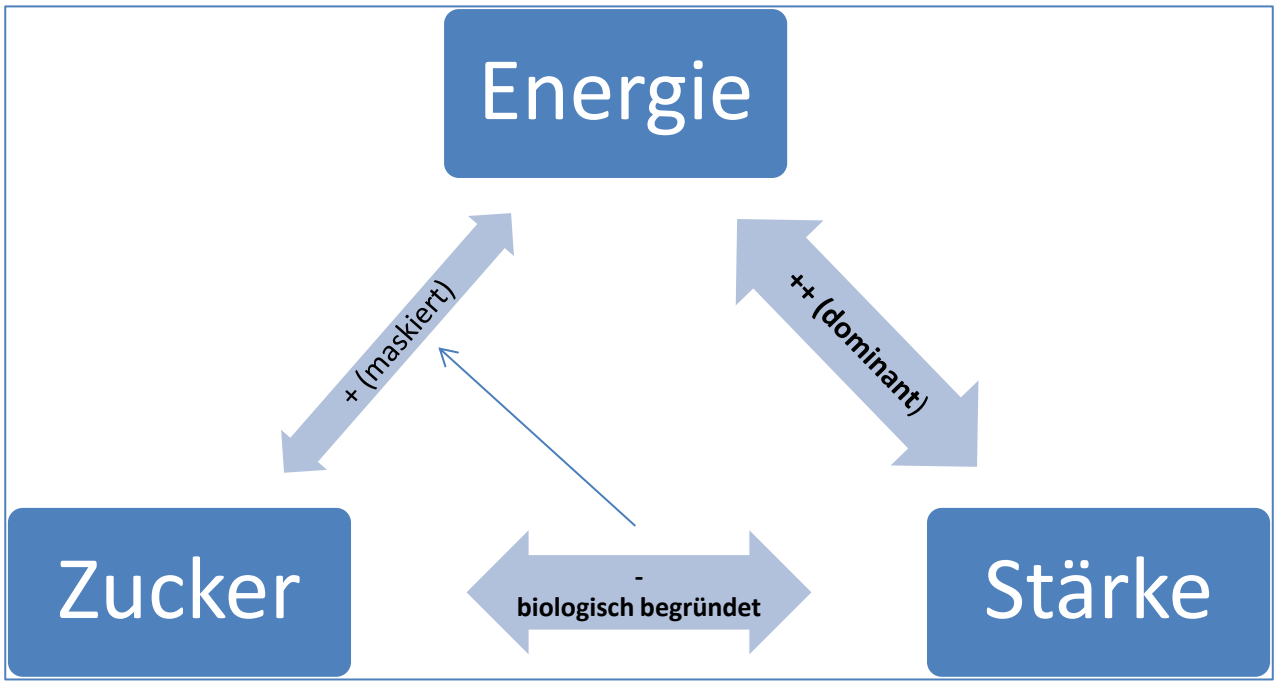


Abbildung 14: Multivariater Komplex zwischen Qualitätsmerkmalen bei Silomais

6.6 Cluster im Lageplan und Boxplots

6.6.1 Aufgabe und methodischer Ansatz

In mehreren Merkmalen auffällige Parzellen und Parzellengruppen sollen distinkt abgegrenzt werden, sofern es eine Zusammenhangsstruktur gibt. Um das zu erreichen, wird eine Clusteranalyse durchgeführt, der eine Faktorenanalyse vorgeschaltet ist (Kapitel 8.4.1 und 8.4.2). Die gefundenen Cluster werden im Lageplan mit disjunkten Farben dargestellt. Jedes Teilstück wird mit dem Parzellenkürzel bezeichnet. Ist die Dimension des Lageplanes zu groß für die Lesbarkeit der Parzellenkürzel wird der Lageplan ohne Bezeichnung ausgegeben, dafür aber zusätzlich eine Tabelle mit der Auflistung der Parzellenkürzel und ihrer Clusterzuordnung.

Die standardisierten absoluten Residuen der Analyse-Merkmale werden jeweils als Boxplot je Cluster dargestellt. Weiterhin werden die standardisierten Residuen der Merkmale in einem Boxplot dargestellt.

6.6.2 Nutzerführung

In fünf optionalen Blöcken (Abbildung 15) kann der Anwender Festlegungen zur Clusteranalyse treffen. Erläuterungen dazu wurden in den Kapitel 8.4.1 und 8.4.2 gegeben.

Lageplan, Boxplots, Protokoll und der SAS-Standardoutput für die Vorzugsvariante werden in den jeweils festgelegten Dateien ausgegeben. Die Lis-Datei enthält nur das Protokoll. Die Nutzereinstellungen der Faktor-Rotations-Methode, der Cluster-Methode sowie der Mindestanzahl der Cluster werden auf Gültigkeit geprüft und bei Ungültigkeit auf die voreingestellten Werte gesetzt sowie im Protokoll dokumentiert. Ist die Anzahl der ausgewählten Merkmale kleiner zwei erfolgt ein Programmabbruch mit Eintrag in das Protokoll.

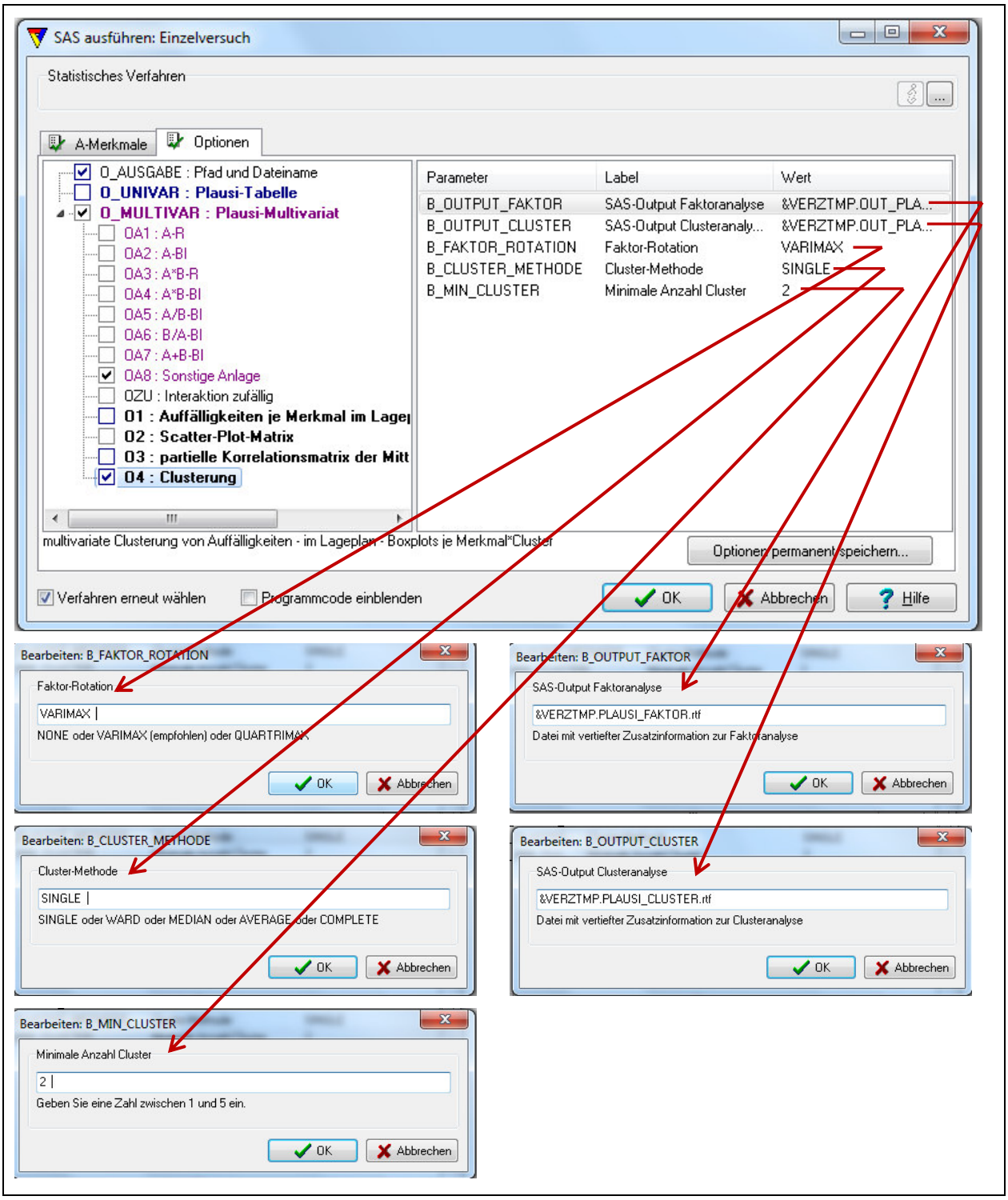


Abbildung 15: Unteroption ‚Clusterung' und optionale Blöcke

6.6.3 Output

Neben den Clustern im Lageplan (Abbildung 16) werden je Merkmal die Cluster auch in Boxplots dargestellt (Abbildung 17). Die in Abbildung 9 bereits angedeuteten Strukturen werden hier merkmalsübergreifend zusammengefasst sehr viel deutlicher ausgegrenzt.

Zusätzlich werden die Residuen jedes Merkmals unabhängig von der Clusterung in einem letzten Boxplot dargestellt.

Eine Interpretation dieser Darstellungen erfolgt in 7.6.

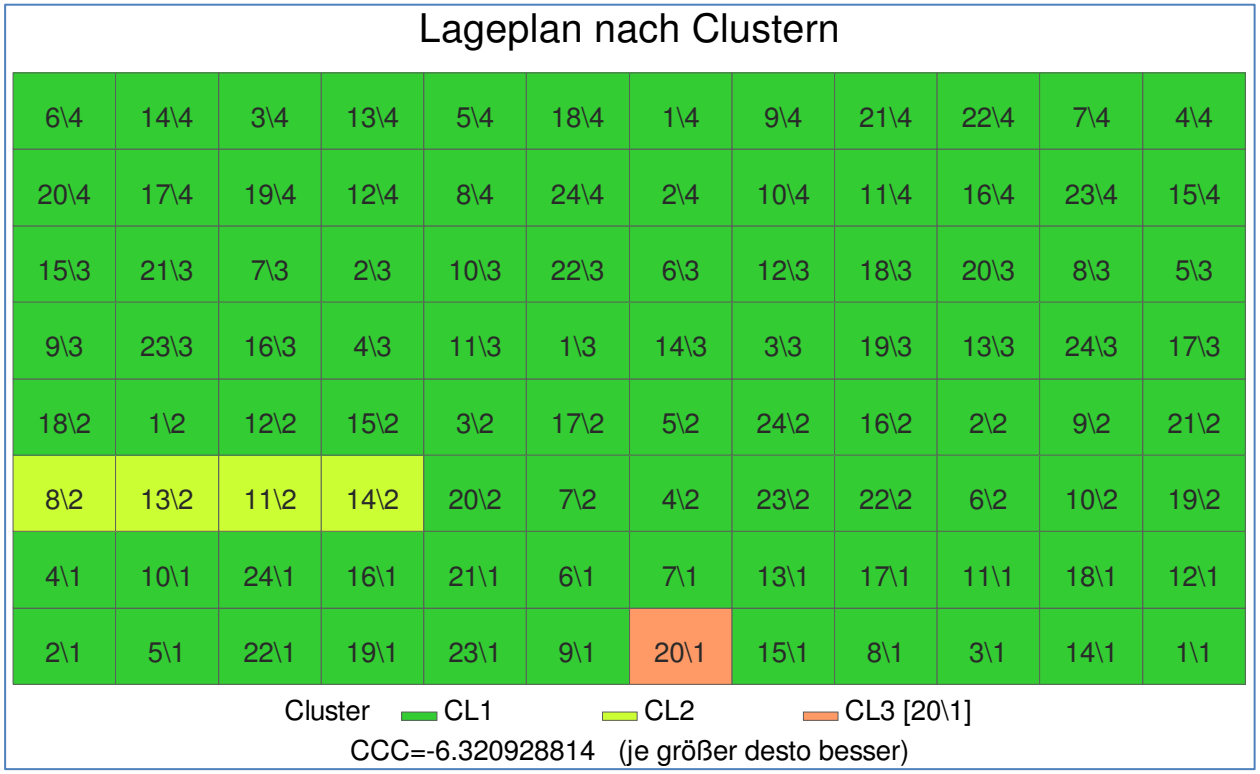


Abbildung 16: Cluster im Lageplan - räumlicher Zusammenhang

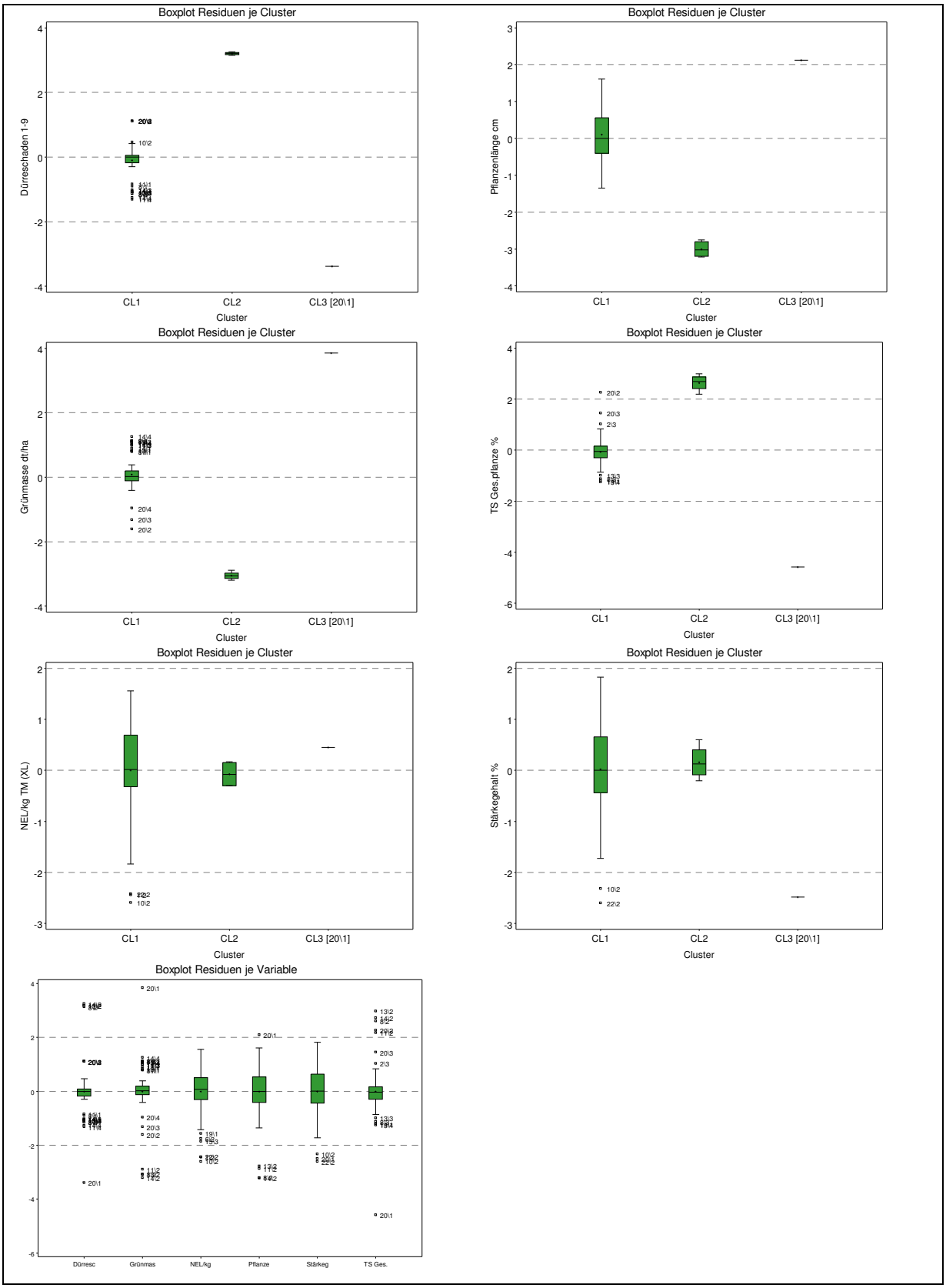


Abbildung 17: Boxplots der Residuen je Cluster für einen Silomaisversuch

Versuch: 5602 GülzowG 2013 / LSV Silomais
Multivariat: Clusterung

Typ	Objekt	Inhalt
Option	Merkmale-Auswahl (6)	Dürreschaden 1-9
		Grünmasse dt/ha
		NEL/kg TM (XL)
		Pflanzenlänge cm
		Stärkegehalt %
		TS Ges.pflanze %
	Merkmale-Analyse (6)	Dürreschaden 1-9
		Grünmasse dt/ha
		NEL/kg TM (XL)
		Pflanzenlänge cm
		Stärkegehalt %
		TS Ges.pflanze %
	Faktor-Rotations-Methode	VARIMAX
	Faktor-Anzahl extrahiert	Kaiserkriterium + 1 = 3
	Faktor-Gewicht verwendet	Wurzel(Eigenwert)
	Cluster-Methode	SINGLE
	Mindestanzahl der Cluster	2
Ergebnis	Max(ccc)	-6.320928814
	Anzahl Cluster	3
Ausgabe	Lageplan und Boxplots	C:\PROGRAMDATA\TEMP\PLAUSI.rtf
	SAS-Output Faktoranalyse	C:\...ATA\TEMP\PLAUSI_FAKTOR.rtf
	SAS-Output Clusteranalyse	C:\...TA\TEMP\PLAUSI_CLUSTER.rtf

Abbildung 18: Protokoll zur Clusterung

Abbildung 18 protokolliert die interaktiven Nutzer-Vorgaben (gewählte Optionen und Voreinstellungen) sowie statistische Eckpunkte der multivariaten Auswertung.

7 Interpretation und Nutzeffekte an Beispielen

7.1 Eingabefehler - Detektion durch die ‚univariate Analyse'

Das Merkmal Pflanzenlänge bei Mais ist i.d.R. in LSV zur Sortenbeschreibung nicht mehr wichtig, da die Sortenrangierungen bekannt sind. Es ist aber zur Fehlerindikation sehr geeignet. Wird das Verfahren bereits nach der Messung lange vor der Ernte ausgewertet, so könnte z.B. eine Vertauschung zweier Sorten in Einzelparzellen beim Maislegen auffällig werden und ließe sich ggf. im Sinne einer korrekten Datenzuordnung zur Auswertung noch bereinigen.

Versuch: LSV KMais D Strohkirchen 2015 / regionale Sorteneignung KENNUNG:1

Statistiken der A-Merkmale

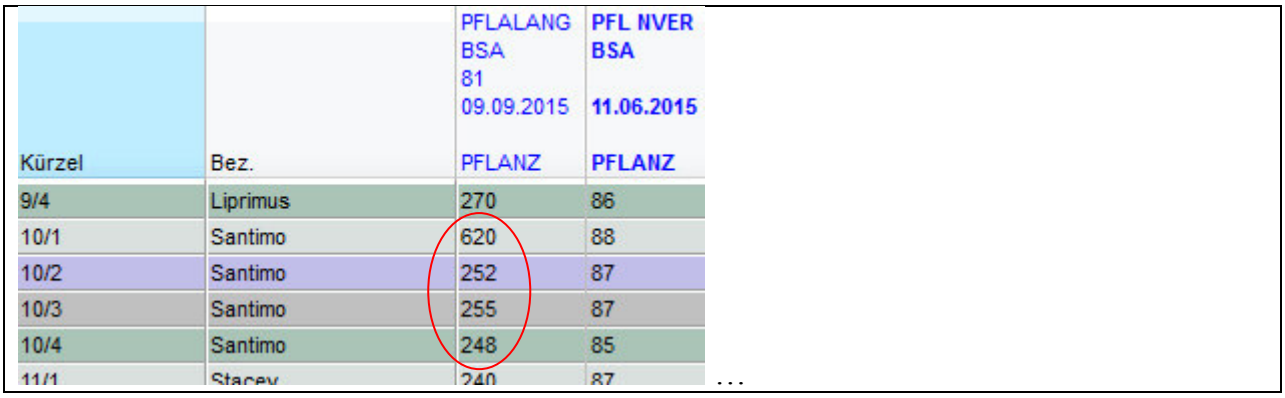
Der F-Test je Merkmal erfolgte ohne Prüfung der Vorrausetzungen.

Als Auffälligkeitsgrenze wurde verwendet: |std.Res| > 9.

	Dat	EW N	Min	Max	MW N	Min	Med	Max	s%	FWert	auf-fällig
Absterbegrad 1-9	09.09	71	2.0	4.0	17	2.0	2.5	3.3	17.6	4.1***	
Kornertrag dt/ha	13.11	71	96.1	131.5	17	107.6	114.6	127.3	5.1	3.1**	
Kornertrag kg/Pz	13.11	71	13.1	17.8	17	13.9	15.4	17.2	4.9	4.3***	
Mängel Aufg. 1-9	29.05	71	1.0	2.0	17	1.0	1.0	1.8	29.6	1.6	
Pfl. n Vereinzel	11.06	71	83.0	88.0	17	84.4	86.3	87.3	1.7	1.6	
Pfl/qm n Vereinz	11.06	71	7.9	8.4	17	8.0	8.2	8.3	1.7	1.6	
Pflanzen/ Parz.	13.11	71	83.0	88.0	17	84.4	86.3	87.3	1.7	1.6	
Pflanzenlänge cm	**09.09**	**71**	**228.0**	**620.0**	**17**	**238.3**	**254.5**	**343.8**	**17.0**	**1.1**	**10\1**
Sollpflanzen	29.05	71	86.0	86.0	17	86.0	86.0	86.0	0.0		
Sollpflanzen qm	29.05	71	8.2	8.2	17	8.2	8.2	8.2	0.0		
TS- Korn %	13.11	71	62.1	70.9	17	62.8	67.6	69.9	1.5	10.8***	

Abbildung 19: Output der univariaten Analyse - möglicherweise ein Eingabefehler

In Abbildung 19 taucht ein extremer Einzelwert von 620 cm auf, der völlig faktisch unmöglich ist. Auch der Vergleich zum maximalen Mittelwert (343,8 cm), welcher wahrscheinlich bereits durch diesen ‚Falschwert beeinflusst ist, unterstreicht die Unplausibilität. Die Folge ist ein für dieses Merkmal außergewöhnlich hoher Variationskoeffizient $s\% = 17$. Entgegen der Erwartung, dass in jedem Maissortenversuch signifikante Sortenunterschied bei der Länge bestehen, ist dies durch diesen einen Ausreißer hier nicht erkennbar. Die Kontrolle in der PIAF-Dateneingabemaske (Abbildung 20) unterstreicht die Unplausibilität des einen Wertes von Parzelle 10/1. denkbar ist, dass es sich um einen ‚Zahlendreher' bei der Eingabe handelt. da dies allerdings nicht mehr prüfbar ist, müsste der Wert gestrichen werden. Wurde die Auswertung allerdings schon vor der Ernte gemacht, so kann die Messung wiederholt werden, sofern sich die Länge nicht mehr geändert hat.



Kürzel	Bez.	PFLALANG BSA 81 09.09.2015 PFLANZ	PFL NVER BSA 11.06.2015 PFLANZ
9/4	Liprimus	270	86
10/1	Santimo	620	88
10/2	Santimo	252	87
10/3	Santimo	255	87
10/4	Santimo	248	85
11/1	Stacey	240	87

...

Abbildung 20: Eingabemaske in PIAF - ein Eingabefehler (möglicherweise ein ‚Zahlendreher')

7.2 Parzelleneinkürzung und Korrektur während der Vegetation - Detektion durch die ‚univariate Analyse'

In Abbildung 21 zeigt sich das Merkmal ‚Pflanzenzahl nach Vereinzeln' (Zeile 2) beim Mais auffällig. Da die Pflanzenzahl eigentlich durch Vereinzelung auf eine vordefinierte, für alle Parzellen einheitliche Anzahl gebracht werden sollte, erscheint die Spanne des Wertebereiches äußerst kritisch. Allerdings traten im Versuch örtlich Lücken auf, die bereits vor dem Vereinzeln von der geplanten Ernteparzelle ausgenommen wurden (Abbildung 22, Zeile 1: Erntefläche nicht einheitlich 12 m², sondern reduziert bis auf 11 m²), sodass einige Parzellen entsprechend geringere Pflanzenzahlen für die Erntefläche aufweisen müssen. Bezieht man die Pflanzenzahl auf einen m² (Abbildung 21, Zeile 3), so reduziert sich das Problem teilweise: der Variationskoeffizient (s%) reduziert sich von 3,1 auf 1,9. Wünschenswert wäre aber ein Wert bei $s\% \approx 0$. Die Sortenmittelwerte liegen zwar bereits relativ einheitlich im Bereich 7,6 bis 7,7 Pflanzen/m², entsprechend gering ist, wie gewünscht der F-Wert ($F = 0{,}8$, d.h. es gibt keine signifikanten Ungleichbehandlungen der Sorten). In Einzelparzellen treten aber Werte bis runter zu 7,3 auf. Diese Parzellen wären zu hinterfragen, sie laufen dem ceteres paribus - Prinzip zuwider, sie sind in einer Ertragsauswertung a priori benachteiligt.
Betroffene Parzellen sind in Abbildung 22 durch PIAF fett markiert. Erfolgte diese Tabellenausgabe bereits vor der Ernte, so können diese Parzellen noch begutachtet und ggf. eingekürzt werden. Alternativ wäre zu entscheiden, ob diese Parzellen nicht wertbar sind.

Versuch: LSV KMais Strohkirchen 2014 / regionale Sorteneignung KENNUNG:1

Statistiken der A-Merkmale
Der F-Test je Merkmal erfolgte ohne Prüfung der Vorrausetzungen.
Als Auffälligkeitsgrenze wurde verwendet: |std.Res| > 5.

	Dat	EW N	Min	Max	MW N	Min	Med	Max	s%	FWert	auf-fällig	
Erntefläche	02.07	48	11.0	12.0	12	11.5	11.9	12.0	2.5	1.1	10\1	3\2
Pfl. n Vereinzel	01.07	48	84.0	94.0	12	88.0	91.1	93.0	3.1	1.2		
Pfl/qm n Vereinz	01.07	48	7.3	7.8	12	7.6	7.7	7.8	1.9	0.8		

Abbildung 21: Output der univariaten Analyse - teilweise zu geringe Pflanzenzahl nach dem Vereinzeln auf eine definierte Soll-Bestandesdichte

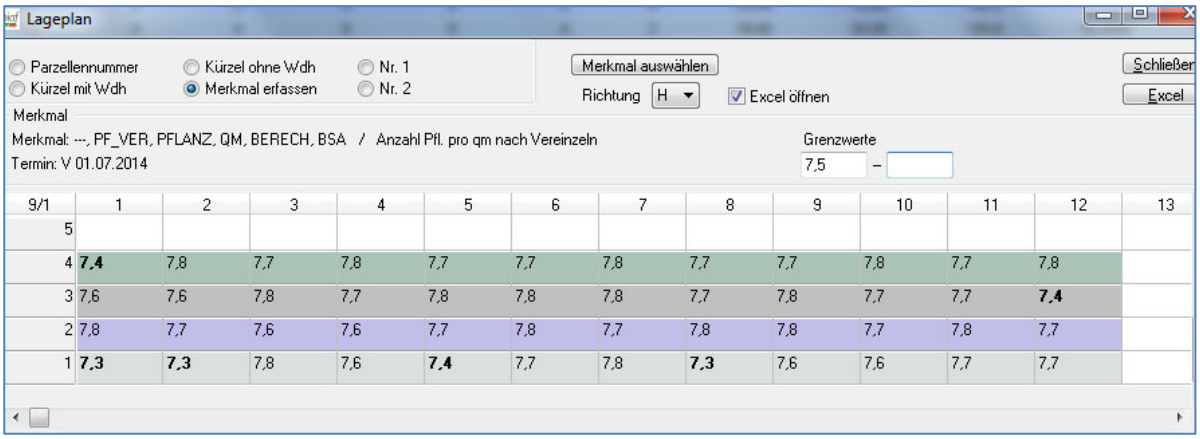


Abbildung 22: Merkmalsdaten im Lageplan in PIAF - auffällige Parzellen mit zu geringer Bestandesdichte

7.3 Erkennung eines bivariaten Problems (Zielmerkmal ‚Ertrag' und ‚Mangel'-Bonitur) - Detektion durch die Scatterplot-Matrix

Zunächst fällt in der ‚univariaten Analyse' auf, dass die Parzelle 7/1 sowohl auffällig beim Ertrag als auch bei einer Mängelbonitur ist (Abbildung 23). Offensichtlich handelt es sich um einen Ausnahmewert, denn Median und Maximum sind bei den Prüfgliedmittelwerten der Mängelbonitur sehr klein, lediglich bei den Einzelwerten taucht ein Maximum von ‚7' auf. Der Variationskoeffizient (s%) ist dadurch sehr hoch. Es besteht der Verdacht, dass dieser Mangel im Ertrag wirksam wurde.

Versuch: LSV KMais D Groß Nemerow 2015 / regionale Sorteneignung KENNUNG:1

Statistiken der A-Merkmale

Der F-Test je Merkmal erfolgte ohne Prüfung der Vorrausetzungen.

Als Auffälligkeitsgrenze wurde verwendet: |std.Res| > 5.

	Dat	EW N	Min	Max	MW N	Min	Med	Max	s%	FWert	auf-fällig
Absterbegrad 1-9	09.09	72	2.0	4.0	17	2.0	2.5	3.4	17.7	4.5***	
Kornertrag dt/ha	13.11	72	72.6	131.5	17	100.9	114.6	127.3	6.3	3.0**	7\1
Mängel Aufg. 1-9	29.05	72	1.0	7.0	17	1.0	1.2	2.2	64.6	0.8	7\1
Pfl. n Vereinzel	11.06	72	83.0	88.0	17	84.4	86.3	87.3	1.6	1.6	
Pfl/qm n Vereinz	11.06	72	7.9	8.4	17	8.0	8.2	8.3	1.7	1.6	
Pflanzenlänge cm	09.09	72	228.0	276.0	17	238.3	254.5	267.8	3.2	4.2***	

Abbildung 23: Output der univariaten Analyse - eine Parzelle zweimal auffällig

Die Scatter-Plot-Matrix (Abbildung 24) zeigt in den Histogrammen (Diagonale), dass es bei den Merkmalen Kornertrag und Mangel jeweils einen Residual-Wert gibt, der nicht zur normalen Variation gehörig zu sein scheint. Auf das Scatter-Plot projiziert wird zudem deutlich, dass der je Merkmal auffällige Wert in beiden Merkmalen der gleiche ist (Parzelle 7/1, s.o.). Bezüglich beider Merkmale findet sich ein Einzelpunkt am Rande der Verteilung, folglich in einer Ecke des Quadrates. Es scheint plausibel, dass der hohe Wert im Merkmal ‚Mangel' mit dem geringen Ertrag eine Sinneinheit ergibt. In der Vorstellung einer Clusterung scheinen alle sonstigen Parzellen ein bivariates Cluster zu bilden, von dem sich 7/1 markant absetzt.

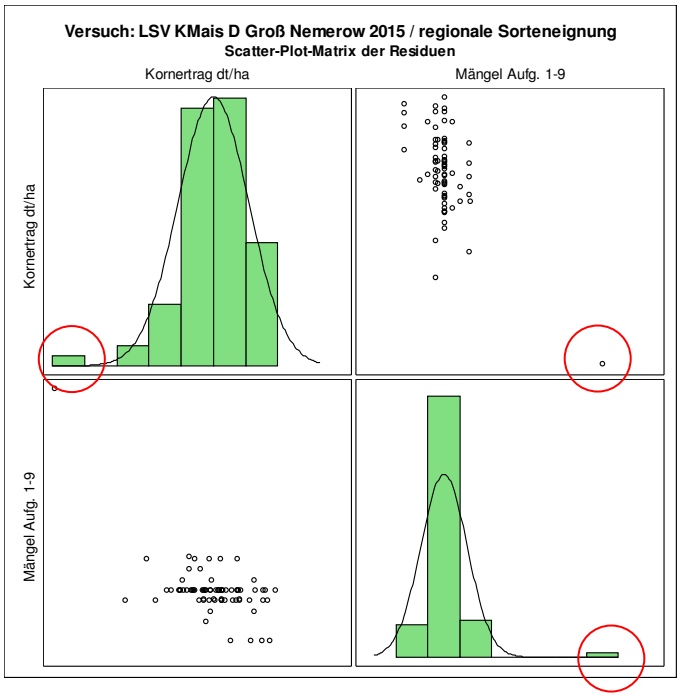


Abbildung 24: Scatter-Plot-Matrix und Erkennung eines bivariaten Problems

Dies ist nicht automatisch ein Indiz zur Streichung - Streichungen sollten sehr vorsichtig, objektiviert und verantwortungsbewusst erfolgen. In jedem Fall sollten weitere Gesichtspunkte hinzugezogen werden - Skizzen, Notizen, Rückfragen beim Techniker, Bonituren u.v.m. In diesem konkreten Fall fand sich ein zuvor übersehener Eintrag im Textbericht (Abbildung 25). Dies war dann Veranlassung, den Ertragswert von 7/1 nicht zu werten.

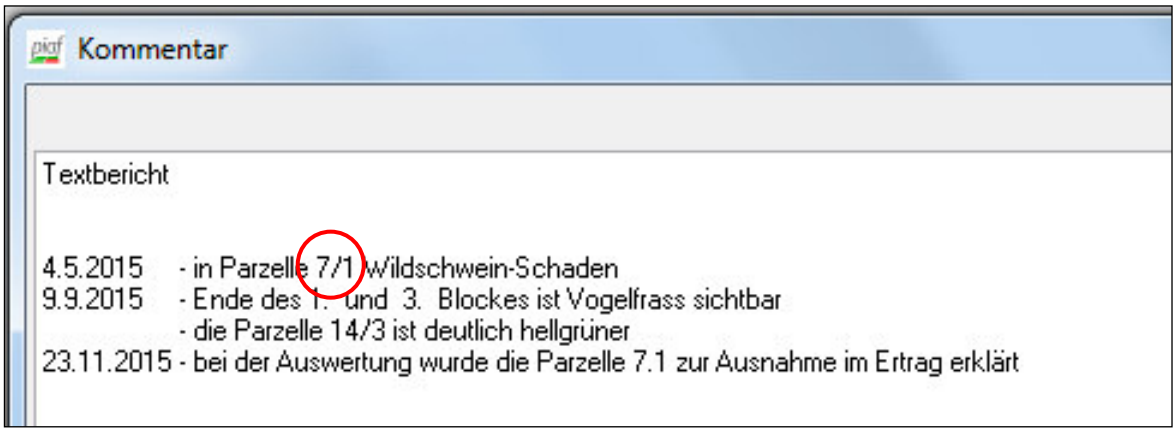


Abbildung 25: Problem erklärender Eintrag im Textbericht als finale Entscheidungs-grundlage

7.4 Erkennung eines Drillfehlers - Detektion durch Boxplots

Die ‚univariate Analyse' (Abbildung 26) zeigt beim Merkmal Pflanzenlänge, dass die Min- bzw. Max-Werte bei den Einzelwerten erheblich (ungewöhnlich) weit entfernt sind von denen der Mittelwerte. das Merkmal Pflanzenlänge zeigt gewöhnlich eine hohe Heritabilität und ist somit geeignet, Sortenvertauschungen gut zu detektieren. Wenn diese Auswertung schon während der Vegetation erfolgte, lohnt sich ein vertieftes Augenmerk auf die Pflanzenlänge, um ggf. an anderen morphologischen Merkmalen zu erkennen, ob z.B. ein Drillfehler / eine Sortenvertauschung bei einzelnen Wiederholungen vorlag.

Vertauschung - Sortentausch / Drillfehler

Versuch: 5702 Strohkirchen 2015 / LSV 3. Ernte KENNUNG:1

Statistiken der A-Merkmale

Der F-Test je Merkmal erfolgte ohne Prüfung der Vorrausetzungen.

Als Auffälligkeitsgrenze wurde verwendet: |std.Res| > 5.

	Dat	EW N	Min	Max	MW N	Min	Med	Max	s%	FWert	auf-fällig
Pflanzenlänge cm	08.09	63	250.0	312.0	21	265.7	281.0	301.3	4.4	2.5**	20\3
TS Ges.pflanze %	14.10	63	37.9	46.0	21	38.4	41.6	45.4	3.1	6.9***	

Abbildung 26: Univariate Analyse für zwei Merkmale

Da auch die Trockensubstanzgehalte mehr als üblich streuten ($s\% = 3{,}1$) wurde im Nachgang eine multivariate Analyse nur für diese beiden Merkmale ausgelöst.

Bei ‚Cluster im Lageplan' (Abbildung 27) werden zwei Parzellen als auffällig im bivariaten Sinne angezeigt: 20/3 und 2/2. Da sie unterschiedliche Cluster bilden (zwei Farben), haben sie ein unterschiedliches Reaktionsmuster und beide scheinen von der allgemeinen Norm abzuweichen.

Lageplan nach Clustern

18\3	8\3	17\3	7\3	21\3	9\3	6\3
10\3	1\3	20\3	19\3	11\3	16\3	2\3
15\3	13\3	14\3	4\3	12\3	5\3	3\3
19\2	16\2	13\2	12\2	6\2	15\2	20\2
5\2	21\2	2\2	8\2	4\2	14\2	18\2
17\2	3\2	9\2	10\2	1\2	7\2	11\2
2\1	4\1	15\1	11\1	17\1	8\1	16\1
7\1	20\1	21\1	5\1	13\1	3\1	10\1
12\1	9\1	1\1	6\1	18\1	19\1	14\1

Cluster CL1 CL2 [2\2] CL3 [20\3]

CCC=-10.93484641 (je größer desto besser)

Abbildung 27: ‚Cluster im Lageplan' - zwei auffällige Parzellen

Die ‚Boxplots nach Clustern' (Abbildung 28) zeigen dann im Detail, wodurch sich die Cluster, die auffälligen Parzellen auszeichnen. 2/2 weicht innerhalb seiner Wiederholungswerte extrem ab, indem sie zu lang und zu trocken ist, 20/3 zeigt genau das entgegengesetzte verhalten. Ein ursächlicher Zusammenhang scheint zu bestehen.

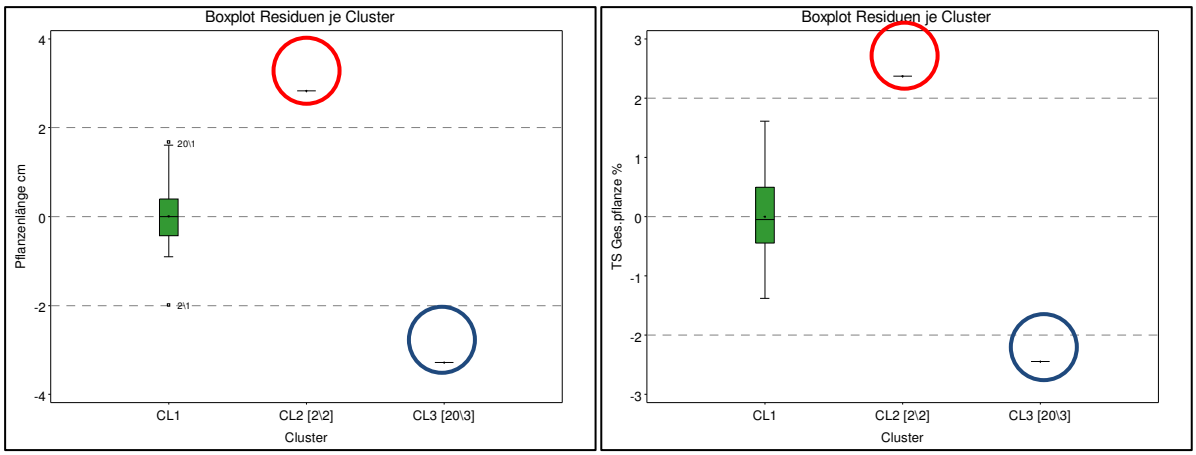


Abbildung 28: ‚Cluster nach Boxplots ' - zwei auffällige Parzellen

Der Textbericht führte zur Klarstellung (Abbildung 29). Ein Drillfehler wurde zwar umgehend in PIAF eingetragen, ist aber entweder nicht ausreichend durchgestellt oder vergessen worden. Empfehlenswert wäre gleich nach Aussaat eine Korrektur der Struktur-Eingaben in PIAF. Das ist aber auch noch nach der Ernte möglich. Der Datenverlust durch Streichung ist nicht empfehlenswert, da fachlich reale Informationen verloren gingen. Es werden nun zwar unbalancierte Blockstrukturen entstehen: Prüfglied 20 und 2 werden jeweils in einer Wiederholung fehlen, in einer anderen doppelt vorhanden sein. Das varianzanalytische univariate Auswertungsverfahren EVA kann damit aber umgehen und die Information optimal nutzen.

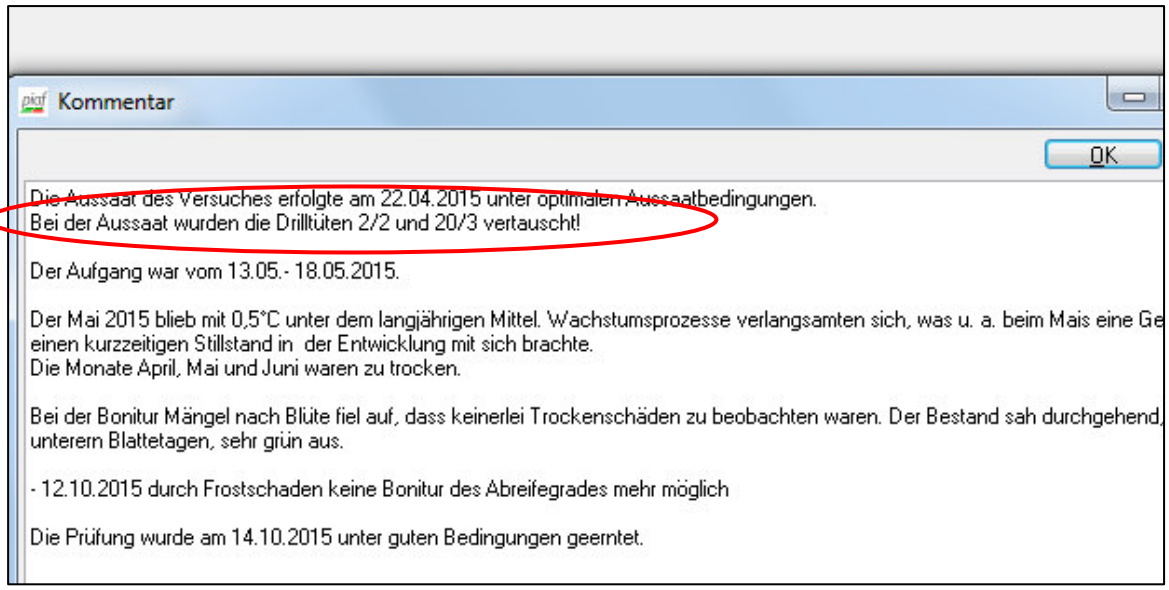


Abbildung 29: Klärender Eintrag im Textbericht → Sortenvertauschung

7.5 Multivariate Zusammenhangsstruktur - Detektion multivariater versus univariater Auffälligkeiten

In Abbildung 30 sind vier Merkmale in der ‚univariaten Analyse' dargestellt, zwischen denen erfahrungsgemäß eine Zusammenhangsstruktur besteht (multipel). Zwei Parzellen werden als auffällig ausgewiesen. In der folgenden multivariaten Analyse sollen diese Parzellen näher beleuchtet werden.

Versuch: 5702 Gülzow 2014 / LSV 3. Ernte KENNUNG:1

Statistiken der A-Merkmale

Der F-Test je Merkmal erfolgte ohne Prüfung der Vorrausetzungen.

Als Auffälligkeitsgrenze wurde verwendet: |std.Res| > 5.

	Dat	EW N	Min	Max	MW N	Min	Med	Max	s%	FWert	auf-fällig
NEL/kg TM (XL)	27.10	80	6.1	7.3	20	6.3	6.7	7.1	1.7	11.7***	
Red. Zucker	27.10	80	6.9	12.5	20	7.4	9.6	11.6	7.0	10.7***	2\3
Stärkegehalt %	27.10	80	22.3	40.8	20	25.7	30.4	37.3	7.6	7.3***	
TS Ges.pflanze %	29.09	80	34.2	45.9	20	35.5	38.6	41.2	3.4	6.9***	19\1

Abbildung 30: Univariate Analyse für Merkmale, die eine Zusammenhangsstruktur aufweisen

Die Scatter-Plot-Matrix zeigt in der Parzellen bezogenen Darstellung (Residuen, Abbildung 31 rechts), dass für eine Parzelle (2/3) ein multivariates Problem vorliegt: viel Zucker & geringe Energiekonzentration & wenig Stärke & geringe Trockensubstanzgehalt. Dies lässt vermuten, dass diese Parzelle weniger ausgereift war, als die anderen Wiederholungen des gleichen Prüfgliedes.

Dagegen scheint die andere auffällige Parzelle (19/1) nur in einem Merkmal auffällig zu sein (univariat): überhöhter Trockensubstanzgehalt.

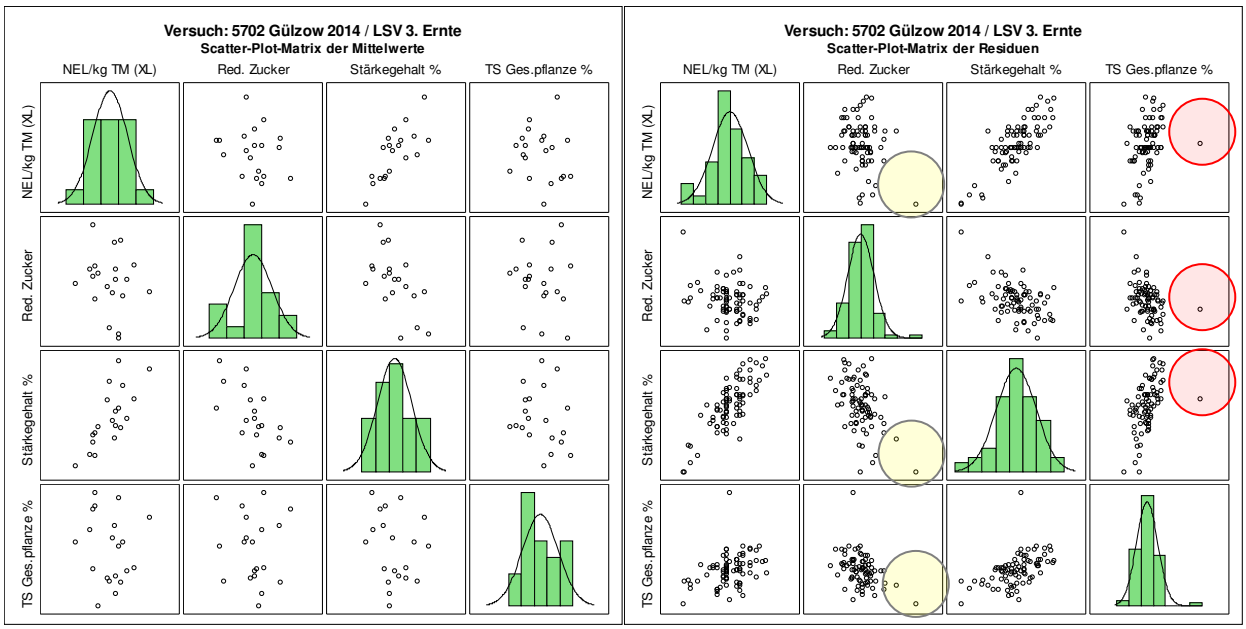


Abbildung 31: Scatter-Plot-Matrix für Merkmale, die eine Zusammenhangsstruktur aufweisen; links Mittelwerte, rechts Residuen

Die Darstellung von Clustern im Lageplan (Abbildung 32) erkennt, dass für die Parzelle 2/3 dieser multivariate Problemkomplex vorliegt. Dagegen wird 19/1 nicht isoliert.

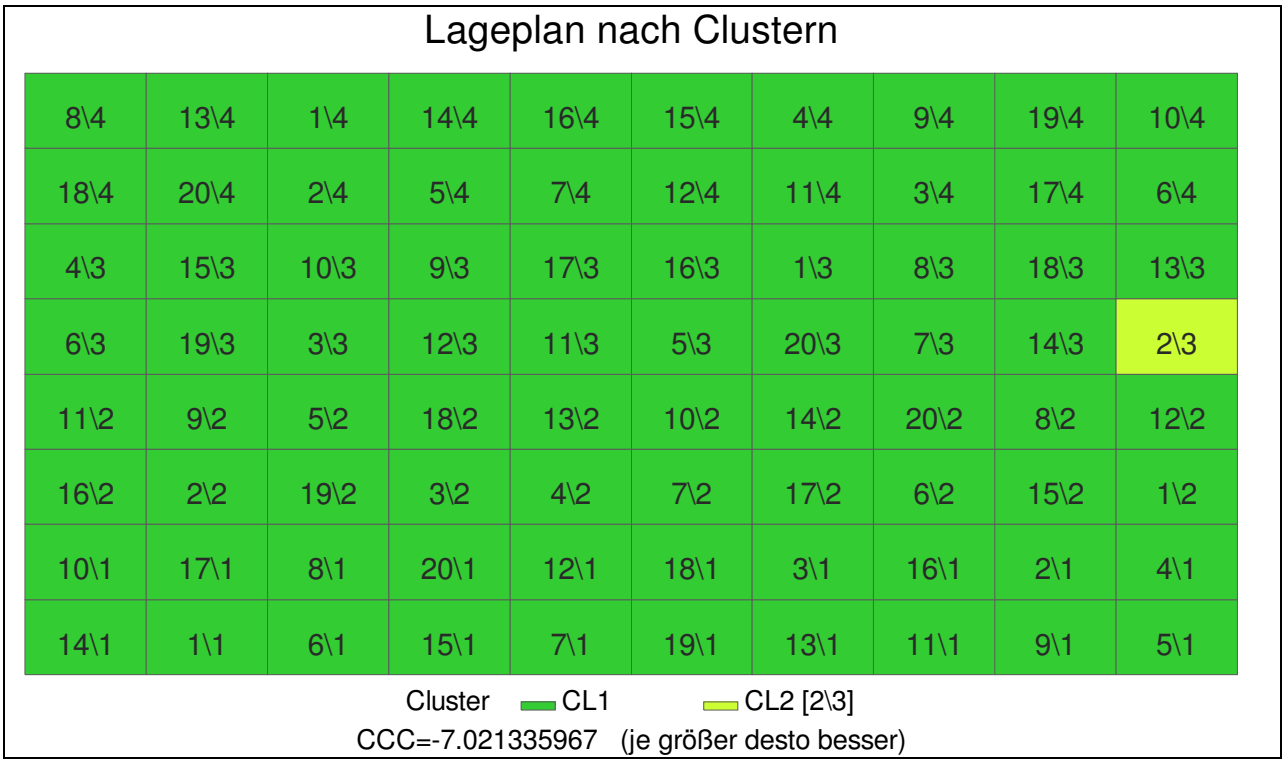


Abbildung 32: Cluster im Lageplan - Erkennung einer multivariat betroffenen Parzelle

Die Darstellung der Residuen je Variable im Boxplot (Abbildung 33) zeigt, dass die Parzelle 19/1 bei univariater Betrachtung die ausgeprägteste Abweichung in einem Einzelmerkmal zeigt. Aus multivariater Sicht zeigt sich aber 2/3 auffälliger.

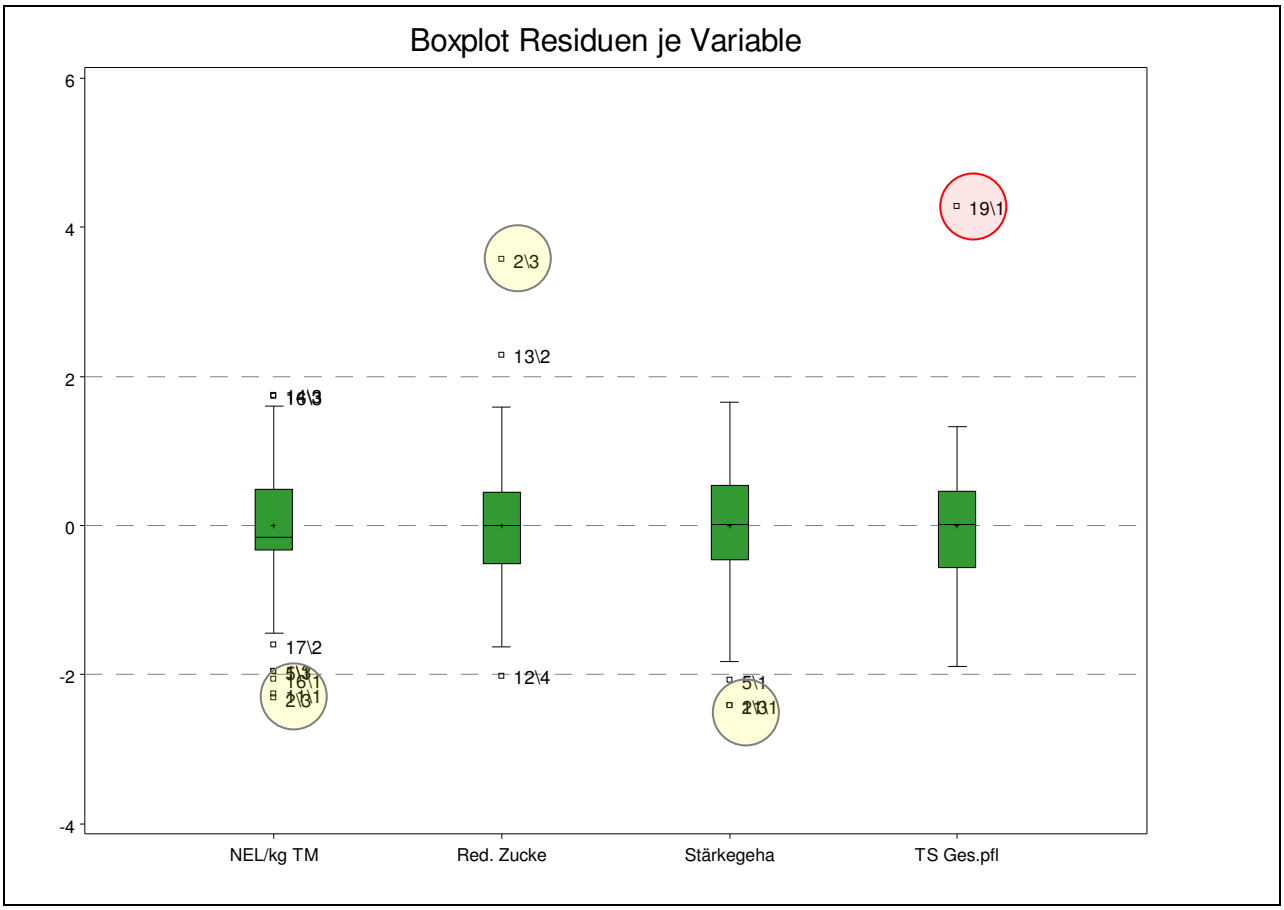


Abbildung 33: Boxplot (Typ ‚Residuen je Variable') - Unterschied zwischen univariat und multivariat auffälligen Parzellen

Die Zusammenhangsstruktur bei 2/3 ist fachlich plausibel. Insofern sollte an deren Messwerten je Merkmal kaum Zweifel bestehen. Vielmehr steht die Frage, ob diese Parzelle aus fachlicher Sicht evtl. zu früh geerntet wurde, die sortentypischen Eigenschaften dadurch nicht zeigen konnte. Über die Wertbarkeit kann nur fachlich entschieden werden. Dagegen erscheint der hohe Trockensubstanzgehalt von 19/1 nicht durch andere Merkmale erklärbar. Hier kann hinterfragt werden, ob der ermittelte TS%-Wert realistisch ist. In Tabelle 1 wurden zu diesem Zweck die Parzellen-Messwerte für beide Prüfglieder zusammengefasst.

Tabelle 1: Messwerte der auffälligen Prüfglieder über alle Wiederholungen

	PG-Nr		Wiederholung			
			1	**2**	**3**	**4**
TS Ges.Pfl. %	2	Torres	40	41	38	41
	19	SY Altitude	46	39	38	39
Red. Zucker	2	Torres	8	8	12	8
	19	SY Altitude	11	12	11	12
Stärk.geh. %	2	Torres	36	38	31	40
	19	SY Altitude	29	28	29	28

7.6 Räumlich zusammenhängende Auffälligkeiten - Detektion durch Cluster im Lageplan und Boxplots

Die nachfolgenden Interpretationen nehmen Bezug auf die bereits in Kapitel 6.6.3 und 6.3.3 dargestellten Outputs.

Abbildung 16 zeigt die räumliche Darstellung dreier gefundener Cluster im Lageplan. Das Gros der Parzellen (Cluster1=grün) stellt quasi die Norm dar. Auffällig sind die ersten vier Parzellen in der zweiten Zeile (Cluster2=gelb). Da diese vier Parzellen alle dem gleichen Cluster zugeordnet sind, scheinen sie nicht nur alle auffällig zu sein, sondern auch den gleichen multivariaten Problemkomplex aufzuweisen. Dies bestätigte in diesem Beispiel die während der Versuchsbetreuung skizzierte optische Einschätzung am Versuch. Im Gegensatz zur etwas diffuser auslaufenden Skizze (also zum subjektiven Eindruck) und auch zu den Residuen der Einzelmerkmale (Abbildung 9) erbringt die multivariate Clusterung eine klare, distinkte Abgrenzung des gestörten Bereiches zum Normbereich. Erfahrungsgemäß zeigen auch die univariaten Auswertungen der Einzelmerkmale in solchen Fällen vergleichbare Tendenzen, welche aber von Merkmal zu Merkmal oft nicht völlig deckungsgleich sind und daher Entscheidungen unsicher gestalten. Somit bietet die Clusterung ein höheres Maß an Objektivität, falls Entscheidungen z.B. zur Nichtwertung einzelner Versuchsbereiche anstehen. Außerdem ist die Einzelparzelle 20/1 auffällig, dabei weist sie einen anderen Verhaltens-Komplex auf, als das Cluster 2.

Die Boxplots der Cluster je Merkmal (Abbildung 17) ermöglichen Deutungen:

- CL2-Parzellen haben erhöhte Dürreschäden,
- CL2-Parzellen haben Ertragseinbußen,
- CL2-Parzellen sind zu kurz,
- CL2-Parzellen sind zu trocken,

→ CL2 scheint eine Trockenstelle zu sein, alle Deutungen passen im Komplex zu dieser Interpretation,

- Parzelle 20/1 (CL3) hat eine anders geartete Struktur und weicht vom mittleren Verhalten der Sorte ab (Sortenverwechslung bei Aussaat?).

8 Prozeduren und Formeln, PIAFStat-Codierungen

Den Plausibilitätsprüfungen ist eine einfaktorielle Varianzanalyse mit der SAS-Prozedur MIXED vorgeschaltet. Die ‚univariate Analyse' erfolgt grundsätzlich als quasi-einfaktorielle Analyse ohne Blockungsstrukturen, wobei die Prüfgliedkombinationen bei mehrfaktoriellen Versuchen nur auf der höchsten Interaktionskombination berücksichtigt werden. Als Klassifikationsvariable wird dazu eine Variable gebildet, die das PIAF-interne Parzellenkürzel verwendet.

Für die ‚multivariate Analyse' sind die gebräuchlichen ein- und zweifaktoriellen Versuchsanlagen unter ‚Optionen' wählbar (Abbildung 3). Sofern als Versuchsanlage ‚Sonstige Anlagen' gewählt wird (z.B. alle Anlagen mit ≥ 3 Faktoren) erfolgt die Varianzanalyse entsprechend dem Ansatz der univariaten Analyse' (s.o.). In der derzeitigen Implementierung (Stand 2016) kann nur ein Blockungsfaktor berücksichtigt werden. Für die Weiterentwicklung in 2017 ist hier eine optionale unbegrenzte Erweiterung vorgesehen, sodass auch lateinische Anlagen, Gitteranlagen, Zeilen-Spalten-Anlagen etc. modelliert werden können.

Die standardisierten absoluten Residuen (nach Pearson) werden über die Optionen OUTPREDM und RESIDUAL des MODEL Statements durch die SAS-Prozedur MIXED berechnet.

8.1 Statistik der Einzelmerkmale - Plausi-Tabelle

Die Berechnung der statistischen Maßzahlen der Einzelwerte und Prüfgliedmittelwerte erfolgt mit der Prozedur MEANS. Der Variationskoeffizient $s\%$ wird wie folgt berechnet:

$$\mathrm{s}\% = \frac{\sqrt{\mathrm{s}_\mathrm{r}^2}}{\bar{y}} \cdot 100\%$$

wobei s_r^2 die berechnete Restvarianz aus der Varianzanalyse und $\bar{y}$ der Gesamtmittelwert ist. Der F-Wert und die absoluten Residuen werden durch die Prozedur MIXED bereitgestellt. Das Signifikanzniveau des F-Wertes wird wie folgt gekennzeichnet:

Tabelle 2: Kennzeichnung des Signifikanzniveaus

Kennzeichnung	Signifikanzniveau
*	p<0.05
**	p<0.01
***	p<0.001

8.2 Auffälligkeiten eines Merkmals im Lageplan

Die Parzellen werden im Lageplan bezüglich der Residuen farblich klassifiziert dargestellt und mit dem Parzellenkürzel bezeichnet. Diese Klassifizierung erfolgt entsprechend nach folgendem Schlüssel (Tabelle 2):

Tabelle 3: Klassifizierung und Farbdarstellung der standardisierten Residuen (StdRes)

Wertebereich StdRes	Farbe	Farbwert - Code
$StdRes < -2$	dunkelblau	CX0033FF
$-2 \leq StdRes < -1$	hellblau	CXB0C0E6
$-1 \leq StdRes \leq 1$	keine	CXFFFFFF
$1 < StdRes \leq 2$	hellrot	CXFFC1C1
$StdRes > 2$	dunkelrot	CXFF0000

8.3 Partielle Korrelationsmatrix der Mittelwerte und der Residuen

Die Klassifizierung der partiellen Korrelationskoeffizienten wird getrennt nach Mittelwerten und Residuen durchgeführt. Ist die Anzahl der Analyse-Merkmale kleiner als 4, wird nur der größte Absolutwert der Gruppe 1 zugeordnet. Alle anderen Werte werden der Gruppe 2 zugeordnet. Ist die Anzahl der Analyse-Merkmale größer 3 werden die Absolutwerte in zwei Gruppen geclustert. Die Gruppe 1 ist die Gruppe mit dem größeren Cluster-Mittelwert, repräsentiert stärkere Zusammenhänge und wird farblich gekennzeichnet. Die Clusterung wird mit der SAS-Prozedur Cluster durchgeführt.

Die partiellen Korrelationskoeffizienten werden im Zuge der Faktorenanalyse bereitgestellt (Tabelle 4) und entsprechend 6.5 im Output des Verfahrens aufgearbeitet.

Tabelle 4: Partielle Korrelationskoeffizienten zwischen Merkmalspaaren

Partial Correlations Controlling all other Variables						
	DURRSCHA	***GM_DTHA***	***NELKGX***	***PFLALANG***	***STAEGEH***	***TSNIGEPF***
DURRSCHA	1.00000	-0.83374	0.13316	0.10170	-0.12655	-0.25235
GM_DTHA	-0.83374	1.00000	0.10963	0.45757	0.01037	-0.65007
NELKGX	0.13316	0.10963	1.00000	-0.31115	0.86147	-0.28293
PFLALANG	0.10170	0.45757	-0.31115	1.00000	0.24582	0.22340
STAEGEH	-0.12655	0.01037	0.86147	0.24582	1.00000	0.51049
TSNIGEPF	-0.25235	-0.65007	-0.28293	0.22340	0.51049	1.00000

8.4 Cluster im Lageplan und Boxplots

In Abbildung 34 ist der Programm-Ablauf zur Bildung der Cluster im Lageplan dargestellt. Das Programm ist durch zwei Abschnitte gekennzeichnet, die Faktorenanalyse (Kapitel 8.4.1) und die Clusteranalyse (Kapitel 8.4.2)

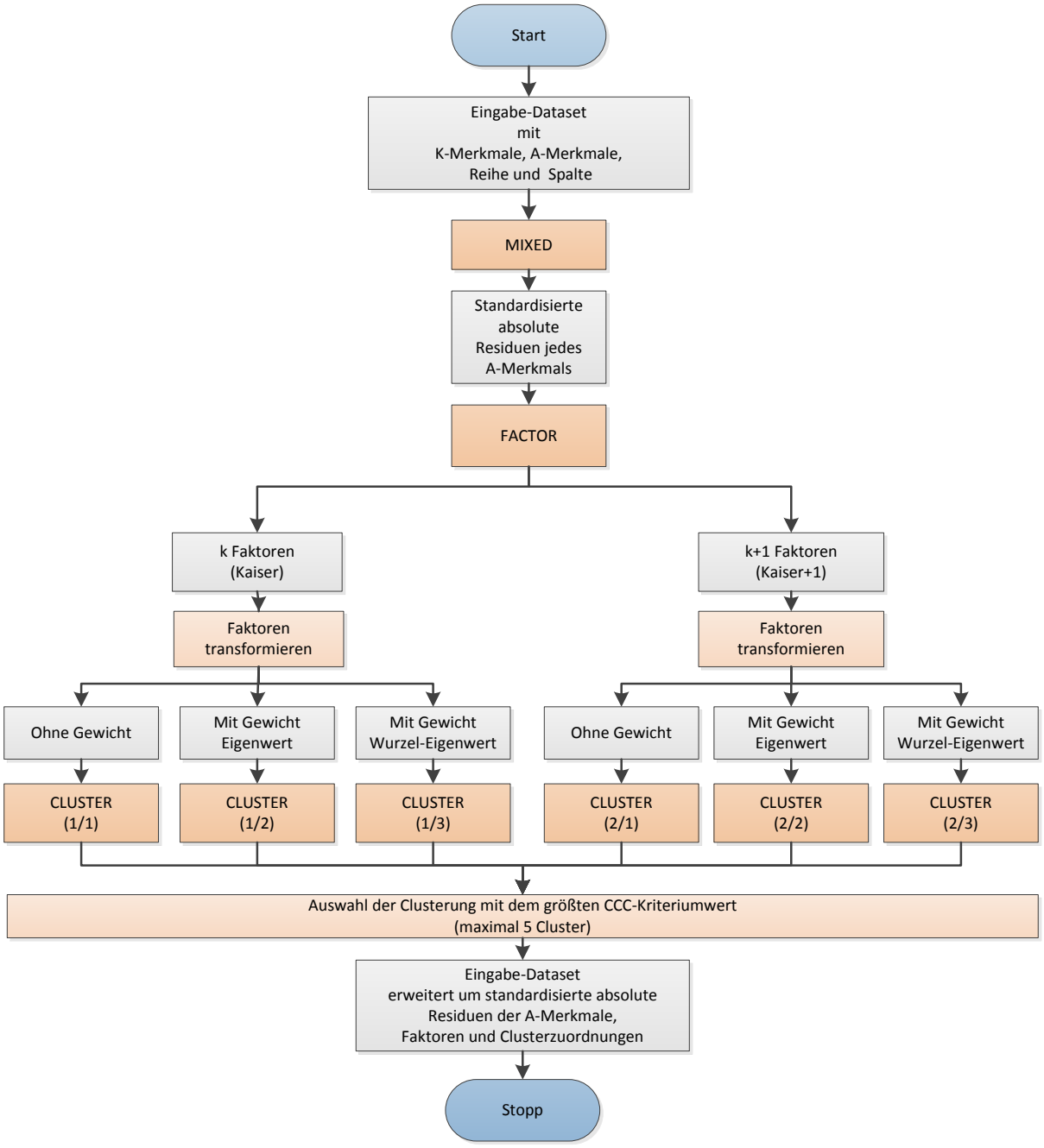


Abbildung 34: Programmlauf der multivariaten Analyse - Clusterung

8.4.1 Faktorenanalyse

Die Faktorenanalyse ist der Clusterung vorgeschaltet, um bei vielen gewählten Merkmalen die Anzahl Variablen zur Clusterung zu vermindern (Verdichtung der Merkmalsanzahl auf eine geringere Faktorenanzahl) und insbesondere um Unkorreliertheit zwischen diesen Variablen (den Faktoren) zu erreichen. Auf Ebene der Originalmerkmale statt der Faktoren würden korrelierte Merkmale ein Übergewicht gegenüber isolierten Merkmalen erhalten. Durch die Vorschaltung der Faktorenanalyse ist es unschädlich, wissentlich hoch korrelierte Merkmale einzubeziehen (z.B. Stärke und Energie).

Tabelle 5: Auszug aus der Faktoranalyse - Erklärung der Faktoren durch Merkmalskoeffizienten bei einem Maisversuch

Rotated Factor Pattern						
	Factor1	***Factor2***	***Factor3***	***Factor4***	***Factor5***	***Factor6***
DURRSCHA	0.99260	-0.02239	0.01828	0.00591	-0.09130	0.07442
GM_DTHA	-0.99665	-0.00749	0.00379	-0.04131	-0.00688	0.06969
NELKGX	-0.07822	0.96518	-0.05974	-0.23506	-0.05899	0.00143
PFLALANG	-0.90271	0.01765	0.42625	0.04257	0.03614	0.00061
STAEGEH	0.14523	0.93361	0.08122	0.30110	0.10004	-0.00336
TSNIGEPF	0.93093	0.16629	0.09172	0.12228	0.28694	-0.00531

Tabelle 5 zeigt z.B. an, dass der Faktor 2 im Wesentlichen durch den hochkorrelierten Merkmalskomplex {Energiekonzentration (NELKGX) und Stärkegehalt} bestimmt ist. Der noch bedeutsamere Faktor 1 ist offensichtlich stark durch einen lokalen Dürreschaden bestimmt, der zum Korrelationskomplex {Dürrebonitur, Ertrag, Länge, Trockensubstanzgehalt} führte.

Nach der Faktoranalyse werden einerseits die Faktoren extrahiert, deren Eigenwerte größer gleich 1 sind (Kaiserkriterium) und zusätzlich als Alternative ein zusätzlicher Faktor extrahiert (Kaiserkriterium+1) (Tabelle 6).

Tabelle 6: Auszug aus der Faktoranalyse - Eigenwerte der Faktoren und erklärte Varianz

Eigenvalues of the Covariance Matrix				
	Eigenvalue	*Difference*	*Proportion*	*Cumulative*
1	14.0127104	10.6936934	0.8046	0.8046
2	3.3190170	3.2357341	0.1906	0.9952
3	0.0832828		0.0048	1.0000

In diesem Beispiel kann die Gesamtvariabilität durch die beiden oben beschriebenen ersten Faktoren (Kaiserkriterium) weitgehend beschrieben werden - der erste Faktor erklärt bereits 80 Prozent und mit dem zusätzlichen zweiten Faktor werden 99,5% der Variabilität erklärt. Die beiden ersten Faktoren können quasi als neues Merkmal mit einer neuen Variablen-bezeichnung verständlicher gemacht werden: Faktor1 ≡ ‚Dürreproblem', Faktor 2 ≡ ‚Energie durch Stärke'.

Danach werden drei Clusteranalysen für die jeweils extrahierten Faktoren ohne Gewicht, mit Gewicht-Eigenwert und mit Gewicht Wurzel-Eigenwert durchgeführt. Die Eigenwerte werden Tabelle 6 entnommen. Die Faktoren mit Gewichten (transformierte Faktoren) werden wie folgt berechnet:

$$FW_{ij}^{trans1} = FW_{ij} \cdot EW_j \quad \text{und} \quad FW_{ij}^{trans2} = FW_{ij} \cdot \sqrt{EW_j}$$

mit $i = 1$…Anzahl Parzellen und $j = 1$…Anzahl Faktoren,
wobei FW_{ij}der Faktorwert und EW_jder Eigenwert ist.

Anmerkungen zur Wahl der Faktor-Rotations-Methode

Über einen optionalen Block (‚Faktor-Rotation', Abbildung 15) kann der Anwender die Faktor-Rotations-Methode festlegen (Voreinstellung: VARIMAX, weitere Möglichkeiten: NONE und QUARTIMAX).

Die Wahl der Faktor-Rotations-Methode hat hier im Wesentlichen nur Einfluss auf die Interpretierbarkeit der Faktoren (Tabelle 5). Der Default ‚VARIMAX' führt gegenüber ‚NONE' (keine Rotation) i.d.R. dazu, dass die Faktoren aus weniger Merkmalen heraus maßgeblich interpretierbar werden. Die Clusterung, also das Hauptziel, wird dadurch nicht relevant beeinflusst. Insofern ist diese Einstellung nur für ‚Insider' von Relevanz, die sich mit dem SAS-Output der Faktorenanalyse befassen und die Faktoren interpretieren.

8.4.2 Clusteranalyse

Für jede der sechs Clusterungen (Varianten 1/1 bis 2/3, Abbildung 34) wird das Maximum des Cubic Cluster Criterion (CCC) über einen Bereich von 2 bis zu maximal 5 Clustern bestimmt. Die Clusterung mit dem größten Maximum des CCC-Wertes (Vorzugsvariante) wird im Lageplan und in den Boxplots verwendet (Abbildung 35).

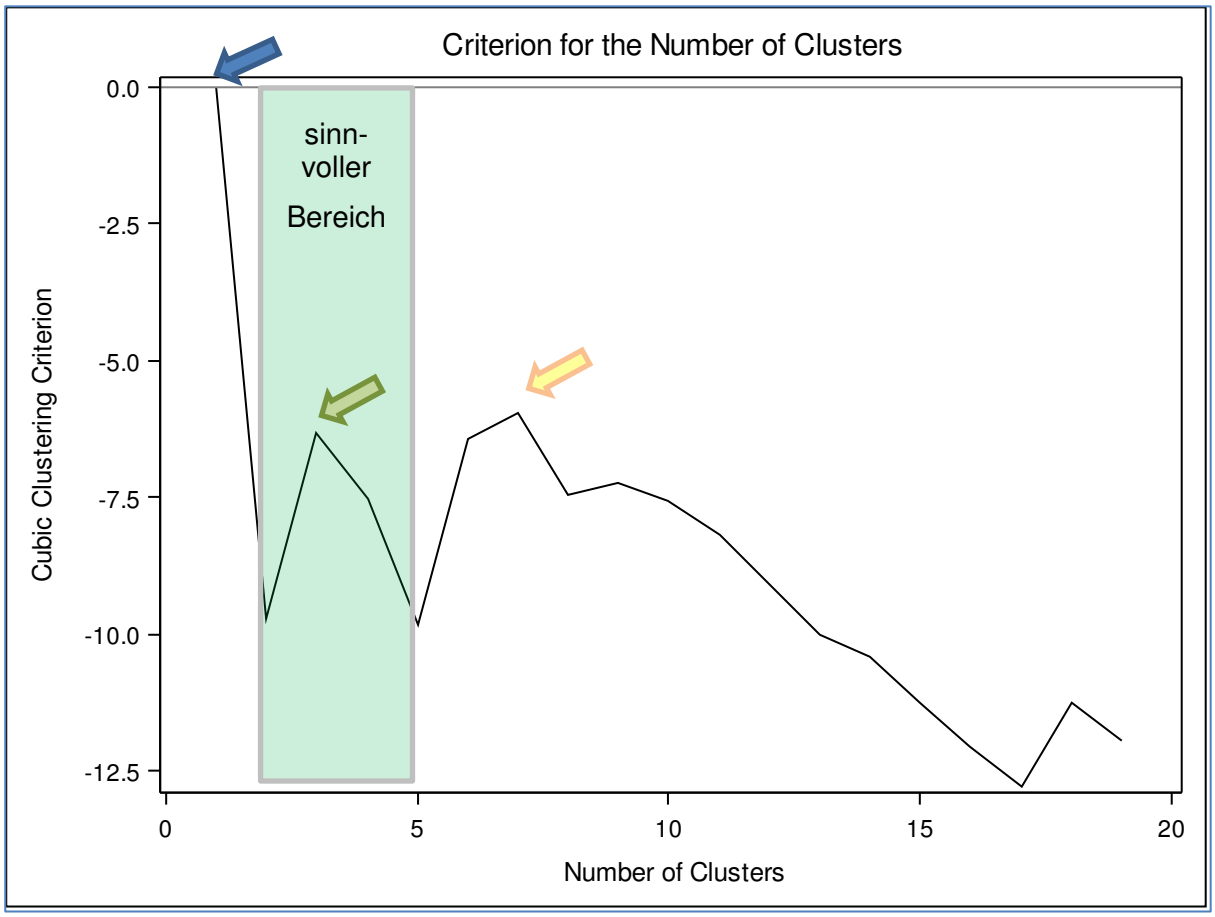


Abbildung 35: Clusterungsgüte (CCC) in Abhängigkeit von der Anzahl gebildeter Cluster

Anmerkungen zur Wahl der Mindestanzahl der Cluster und Clusterungs-Kriterium CCC

Jede Clusterungsmethode (s.u.) optimiert eine adäquate Maßzahl. Diese Maßzahl ist aber jeweils nur für diesen Datensatz interpretierbar - Quervergleiche zu anderen Clusterungen sind nicht möglich. Hierfür wurde das verallgemeinerte ‚Cubic Clustering Criterion', kurz ‚CCC' entwickelt. In diesem Verfahren wird insbesondere die Anzahl ausgegebener Cluster so bestimmt, dass das CCC maximiert wird. CCC-Werte über null deuten auf eine sehr gute Gruppenseparierung hin. Vorstellbar sind solche guten Werte z.B. bei der Separierung von einem Gemisch aus Erbsen und Linsen (Aschenbrödel) aufgrund einiger Kornmerkmale. Residuen in randomisierten Versuchen lassen selten eine solch scharfe Clusterung erwarten. Daher wird i.d.R. die Variante, in der nur ein einziges Cluster gebildet wird, das höchste CCC erzielen. Der Lageplan wird dann komplett ‚grün' und es gibt keine Hinweise auf Besonderheiten, was nicht zur Unterstützung beiträgt.

Daher wurde hier als Default vorgeschlagen, dass mindestens 2 Cluster gebildet werden sollen (also auch, wenn das CCC dann kleiner ist als bei nur einem Cluster). Es wird dadurch eine Gruppierung in mindestens 2 Cluster erzwungen. Eine noch größere Mindestanzahl Cluster kann man erzwingen, indem ein Wert bis max. 5 im optionalen Block eingegeben wird. Mehr als fünf Cluster werden in dem Verfahren PLAUSI a priori ausgeschlossen, da sie kaum noch im Sinne eines Qualitätschecks interpretierbar und nützlich sein dürften. (Abbildung 35)

Die entstehenden Cluster sind dann immer mit Vorsicht zu betrachten, sollen nicht überbewertet werden, keinesfalls sollte voreilig ohne sichere fachliche Gesichtspunkte ein Ausschluss von Parzellen nur aufgrund dieser Clusterung erfolgen! Dessen unbenommen ist dieses Verfahren geeignet, Auffälligkeiten objektiviert zu detektieren, welche sich häufig gut mit Feldbeobachtungen decken.

Ein Vergleich unterschiedlicher Clusterungsmethoden mit dem CCC wird hier nicht empfohlen, da die Methoden in unterschiedlicher Weise dem Anspruch des CCC entsprechen.

Über optionale Blöcke kann der Anwender mit folgenden Einstellungen in das Verfahren eingreifen und dabei die Cluster-Methode (Voreinstellung: SINGLE, weitere Möglichkeiten; WARD, MEDIAN, AVERAGE und COMPLETE) sowie die minimale Anzahl der Cluster (Voreinstellung: 2, weitere Möglichkeiten: 1,3,4 und 5) festlegen (Abbildung 15).

Anmerkungen zur Wahl der Clusterungsmethode

Mit dem Verfahren der Clusteranalyse wird in unserem Fall nach einer geeigneten Bildung von Parzellengruppen bei einer sinnvollen Gruppenanzahl gesucht. Dabei sollen in multivariater Betrachtung einerseits die Parzellen innerhalb einer Gruppe (eines Clusters) relativ ähnlich reagieren und andererseits sollen sich die Cluster deutlich voneinander unterscheiden. Diese beiden Ziele lassen sich allerdings nicht deckungsgleich optimieren, vielmehr wird eine Kompromisslösung über beide Ziele gesucht. Mit der Wahl der Clusterungsmethode wird über die Abwägung zwischen beiden Zielen:

- SINGLE: hier als Default vorgeschlagen zur Detektion von wenigen besonders abweichenden Einzelparzellen oder Parzellengruppen, die sich stark vom „Normbereich“ absetzen; der Normbereich (CL1) wird i.d.R. sehr groß ausfallen und in ihm wird eine relativ große Varianz hingenommen; single linkage=nearst neighbor; Zusammenführung basiert auf euklidischer Distanz zum dichtesten Objekt eines Clusters
- COMPLETE: bildet eher viele kleine homogene Gruppen; detektiert nach unseren Erfahrungen weniger gut bei den hier vorliegenden Problemen; Gegenpol zu SINGLE: euklidische Distanz zum entferntesten Objekt eines Clusters; Ziel: minimierte Varianz im Cluster
- AVERAGE und MEDIAN: Ausreißer und kleine Gruppen vom Pulk sondieren; Tendenz zu mehr Clustern als bei SINGLE; ähnlich WARD; im Ergebnis zwischen SINGLE und COMPLETE: mittlere euklidische Distanz zu allen Objekten eines Clusters; wobei die ‚Mitte‘ durch den arithmetischen MW (bei AVERAGE) oder durch den Zentralwert (bei MEDIAN) definiert ist
- WARD: ausgewogen zwischen Homogenität in Gruppe und Heterogenität zwischen Gruppen, harmonische Gruppengröße; im Allgemeinen oft besonders geeignet, in unserem Fall aber Tendenz zur Verwischung und zu gegenüber SINGLE zu vielen schwerer interpretierbaren Clustern; könnte gelegentlich passen, wenn z.B. in einem Versuch mindestens zwei größere Bereiche erwartet werden, z.B. 1/3 der Fläche trocken, 2/3 normal

Zum Verständnis der Gruppenbildung durch Clusterung eignet sich das Dendogramm (Abbildung 37), welches zwar nicht im Standardausdruck bereitgestellt wird, aber für den Bedarfsfall in einem vom Nutzer definierten Pfad abgelegt ist. In diesem Fall ist das in 6.6.3 und 7.6. gewählte Beispiel verwendet worden. Die im ‚Cluster im Lageplan‘ dargestellte auffällige Einzelparzelle (20/1) und die räumlich zusammenhängende Trockenstelle (8/2, 13/2, 11/2,

14/2) wurden am oberen Ende des Dendogrammes isoliert. Zunächst erstaunt, dass unmittelbar darunter die anderen Widerholungswerte der Prüfglieder 20 bzw. 8, 13, 11, 14 wieder en bloc gebündelt auftauchen. An dieser Stelle muss man sich vor Augen führen, dass keine Originalmerkmale analysiert werden, sondern Residuen. Die Parzellenwerte eines Prüfgliedes sind nicht frei erhoben, sondern stehen als Linearkombinationen in gegenseitigem Wechselbezug. Tritt z.B. bei einem Prüfglied ein stark negativer Ausreißer auf, so ‚zieht' dieser den Prüfglied-Mittelwert zu sich und ‚schiebt' somit die anderen Wiederholungswerte tendenziell in den Bereich numerisch positiver Residuen. Die Clusterung auf Basis eines objektivierten Kriteriums (CCC) erkennt allerdings, dass diese ‚Scheingruppen' real eher zum ‚Normbereich' gehören (Abbildung O). Dies gelingt umso eher, je mehr Wiederholungen vorliegen. Bei typischen zweifaktoriellen Sortenversuchen mit nur zwei Wiederholungen kann es hilfreich sein, die Option ‚Interaktion A*B als zufällig' zu nutzen, sodass für die Residuen i.d.R. vier Parzellenwerte anstelle von nur zwei herangezogen werden.

Wenn die Faktoranalyse oder die Clusterung nicht erfolgreich durchgeführt werden kann, erfolgt im List-Fenster von PIAFStat eine entsprechende Ausgabe (Abbildung 36).

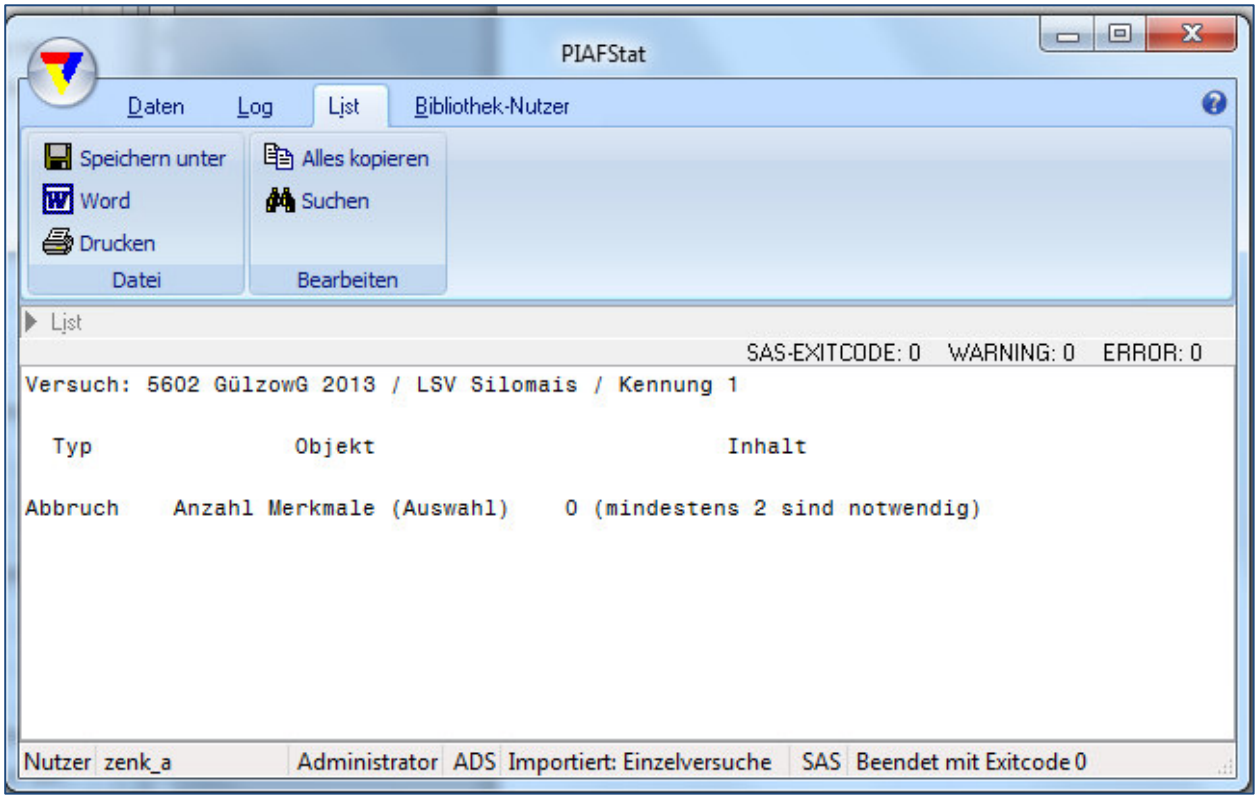


Abbildung 36: Beispiel der Ausgabe im List-Fenster von PIAFStat bei Abbruch der Clusteranalyse

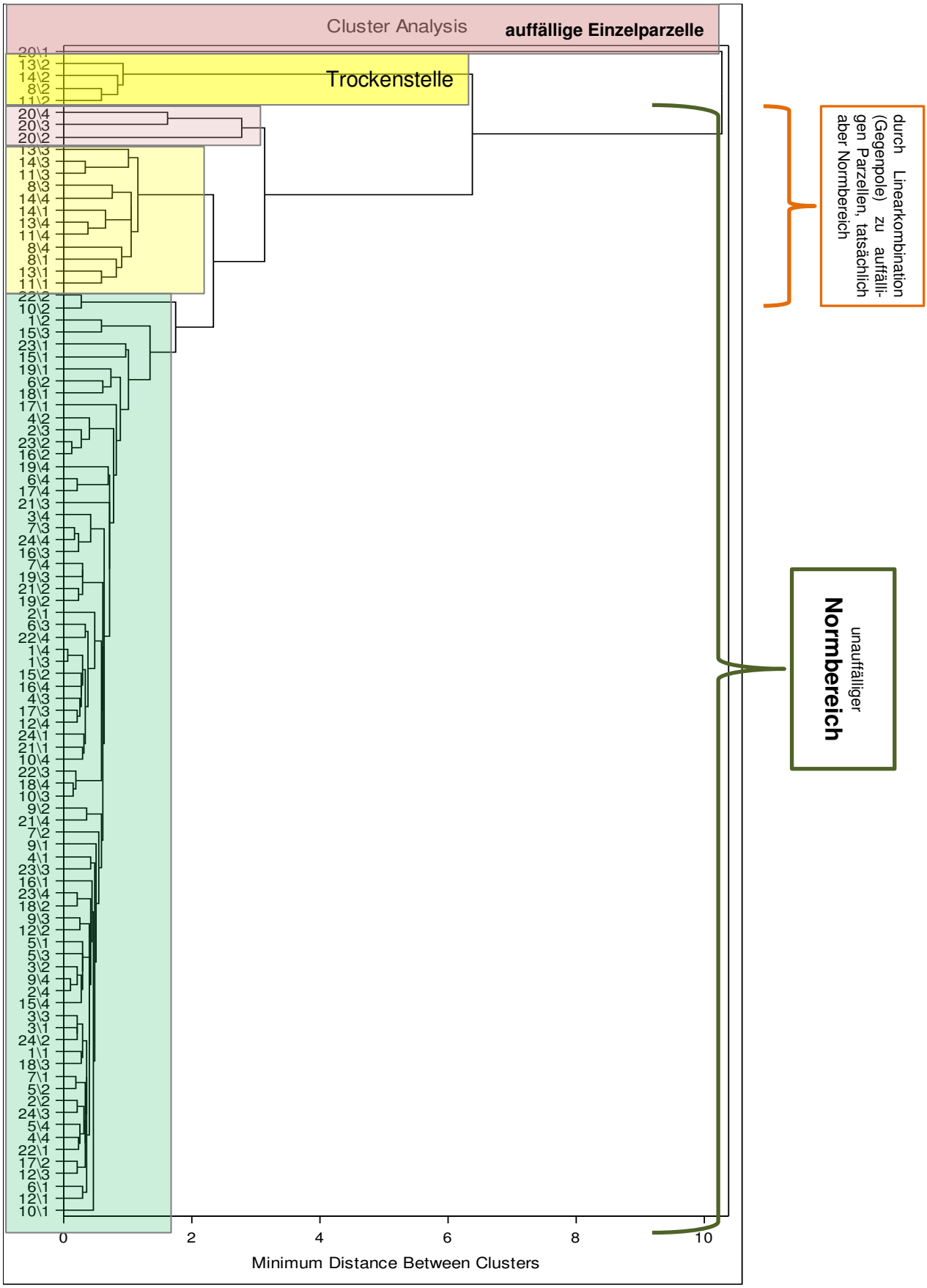


Abbildung 37: Dendogramm einer Clusterung (Methode: SINGLE) - Analog zum Beispiel in 6.6.3 und 7.6

9 Ausblick

Die Maßnahmen zur Weiterentwicklung des PIAFStat-Verfahrens PLAUSI aus 2016 wurden im vorliegenden Bericht teilweise schon berücksichtigt. Diese Weiterentwicklungen betreffen die ‚multivariate Analyse'. Unter ‚Optionen'sind die gebräuchlichen ein- und zweifaktoriellen Versuchsanlagen sowie ‚Sonstige Anlagen' (z.B. alle Anlagen mit ≥ 3 Faktoren) wählbar (Abbildung 3). Dadurch kann der Anwender das Versuchsdesign mit dem Modell der Varianzanalyse in Übereinstimmung bringen (Kapitel 8). In der derzeitigen Implementierung (Stand 2016) wird max. ein Blockungsfaktor berücksichtigt.

Die geplanten Maßnahmen für 2017 sollen in der multivariaten Analyse Blockstrukturen vom Lateinischen Rechteck, von Gitteranlagen, Zeilen-Spalt-Anlagen etc. ermöglichen.

10 Überleitung

Das Verfahren PLAUSI ist eine Softwarelösung für das angewandte Versuchswesen, also nicht unmittelbar für die landwirtschaftliche Praxis. Daher ist im Weiteren die Überleitung in die angewandte Forschung / das angewandte Versuchswesen gemeint. Wenn natürlich durch verbesserte Versuchsauswertung letztlich die landwirtschaftliche Praxis profitiert, handelt es sich bei den nachfolgend genannten Überleitungsaktivitäten also nicht um eine unmittelbare Überleitung in die Praxis. Die Überleitung erfolgte in Form nachfolgender Publikationen und Vorträge:

- Vortrag „Statistische Methoden zur Überprüfung der Qualität von Versuchsdaten und ihre Implementierung in PIAFStat", Herr Schmidtke, Sommertagung der AG Landwirtschaftliches Versuchswesen der Biometrischen Gesellschaft, 25.06.2014, Ihinger Hof,
- Vortrag „Vorstellung des PIAFStat-Verfahrens PLAUSI – Vorauswertung zu Auffälligkeiten", Herr Michel, Sitzung der AG PIAF-Auswertung, 21.04. 2015, Hannover,
- Schulung, Herr Dr. Michel und Frau Dr. Zenk, Sitzung der AG PIAF-Auswertung, 24.02./25.02.2016, Kassel,
- Info-Text für die Softwarelösung PLAUSI,
- Vortrag, Herr Dr. Michel, Sommertagung der AG Landwirtschaftliches Versuchswesen der Biometrischen Gesellschaft, 2016, Osnabrück (angemeldet),
- Einstellung des Verfahrens auf der BioMath-Homepage für alle PIAFStat-Lizenzinhaber, April 2015,
- Installation in MV und anderen LDS, Einsatz für nahezu alle Einzelversuche seit Ernte 2015.