

PIAFStat-Verfahrensgruppe ‚Hohenheim-Gülzower-Serienauswertung‘ (HGM)

STAND 11.11.2024

Die drei Verfahren PHI, VK und MW nutzen z.T. gleiche Funktionen. Daher sollten alle drei Verfahren immer gemeinsam aktualisiert werden, sonst kann es zu Fehlern oder zum Absturz des Runs kommen. Alte individuell eingerichtete Konfigurationen (Optionseinstellungen) sollten neu erstellt werden.

NEUERUNGEN in den drei Kern-Verfahren ab 4 / 2024

- Outputs erfolgen für die aktuellen MS-Office-Versionen von Word, Excel und Access
- die Option O53 ‚Gewichtung nach SE‘ kann in jetzt auch genutzt werden, wenn (noch) nicht in allen einbezogenen Versuchen SE gespeichert sind (betrifft VK und MW)
- Ein weiteres Verfahren DAT wird nun mitgeliefert: zum Auslesen einer Serien-ADS als externe Excel-Datei
- nutzerfreundlichere und zielführendere Optionen, Struktur, Bezeichnung und Nummerierung der Optionen zwischen PHI, VK und MW abgeglichen
- Excel-Tabellen jetzt gleichwertige Alternative zu Access-Tabellen
- verbesserte Formatierungen, strukturiere Tabellen und verständlichere Bezeichnungen im Output
- Unterstützung bei Plausibilitäts-Checks:
 - im Excel-Output von MW (3 Tabellenreiter: ‚Plausibilitätsprüfung‘ ‚Erläuterung Plausibilität‘ ‚Histogramme‘)
 - in PHI Option O33 von Herrn Eckl hilfreich, um schon vor Beginn der Prozeduren auf Auffälligkeiten aufmerksam zu werden
- Parameter der Ökostabilität werden stabil ausgegeben; Detektion von „Intensivsorten“ und „Extensivsorten“
- Handbücher bezüglich Optionen in der Nutzeroberfläche aktualisiert

Das zentrale Ziel, die Schätzung von Sorteneffekten bzw. -unterschieden, wird mit dem Verfahren **MW** erzielt. **PHI** und **VK** sind vorgeschaltete Verfahren, wobei VK zwingend nötig ist, PHI dagegen optional (s.u.). Das Hilfs-Verfahren ‚DAT‘ wandelt eine ADS in eine externe Excel-Datei, welche anstelle oder in Kombination mit einer ADS in den Kernverfahren genutzt werden kann.

1. PHI

Bei der Auswertung von mehrortigen langjährigen Versuchsserien im Sortenwesen wird eine Vielzahl von Versuchen einbezogen, die sich in der mittleren Ausprägung eines Merkmales häufig extrem, z.T. um ein Vielfaches, unterscheiden. Viele Versuchsansteller haben auch in balancierten Datenstrukturen schon mal festgestellt, dass es bei der Mittelwertbildung (z.B. von Erträgen) über mehrere Versuche (Versuchsserie) einen Unterschied macht, ob man (a) zuerst die absoluten Werte mittelt und erst dann die Serienmittel in Relativzahlen wandelt oder ob man (b) zuerst in den Einzelversuchen Relativzahlen bildet und diese dann mittelt. Im letzteren Fall wird man feststellen, dass sich die Sorteneffekte ertragschwächerer Versuche im Serienmittel stärker niederschlagen (und umgekehrt).

Die Frage ist nun, ob einer dieser beiden Ansätze oder aber auch ein Ansatz zwischen beiden Extremen folgendes Ziel besser erfüllt: Vorhersage der wahrscheinlichsten Sorteneffekte in einem neuen Jahr an einem beliebigen Standort im Anbaubereich mit noch nicht absehbarem (Ertrags-)Niveau. Dieses Ziel soll durch das Vorschalten des Verfahrens PHI vor die eigentliche Mittelwertschätzung unterstützt werden. Im unbalancierten Fall (bei nicht-

orthogonal geprüften Sorten) ist ein Tabellenkalkulationsansatz nicht realistisch bzw. suboptimal wegen erheblichen Datenverlustes. Der Situation angemessen ist das gemischte lineare Modell der HGM. Ohne Transformation der Daten entspricht es dem additiven Ansatz von (a).

Das Verfahren PHI untersucht nun, ob eine Vorab-Transformation der Urdaten die Modellvoraussetzungen (insbesondere den Komplex Additivität, Varianzhomogenität und Normalverteilung der Zufallseffekte) besser erfüllt als der Ansatz (a). Es wird ein Transformationsparameter phi (bzw. φ) gesucht, welcher gegebene Daten optimal an das gemischte lineare Modell anpasst und die Modellvoraussetzungen bestmöglich erfüllt. Der geschätzte, ausgegebene Parameter phi (φ) kann dann in den Verfahren ‚VK‘ und ‚MW‘ eingesetzt werden. Dies trägt zur optimierten unverzerrten Mittelwertschätzung im Sinne oben formulierter Problemstellung bei.

Das Verfahren ‚PHI‘ kann optional eingesetzt werden. Alternativ kann in VK und MW auf eine Transformation verzichtet werden, was dem Vorgehen (a) mit einem $\varphi = 1$ bzw. additivem Datenverhalten entspricht. Dem Vorgehen (b) entspräche ein $\varphi = 0$ (\equiv Log-Transformation). Dies ergibt sich bei rein multiplikativem Datenverhalten, wenn Varianzen proportional z.B. zum Ertragsniveau steigen. Häufig ist eine intermediäre Transformation optimal ($0 \leq \varphi \leq 1$). Ein intermediäres $\varphi = 0,5$ entspricht z.B. der Wurzeltransformation. Bei einzelnen Merkmalen, welche z.B. an eine obere nicht überschreitbare Grenze stoßen, sind auch Transformationen mit $\varphi > 1$ plausibel. Dies betrifft oft Relativzahlen wie z.B. die sog. Siebsortierung (%) bei Gerste (denn je höher die Werte sind und an die 100%-Grenze ‚stoßen‘, desto kleiner werden die Varianzen).

Ein $\varphi < 0$ ist theoretisch denkbar, wird aber kaum auftreten und sollte hinterfragt werden. Dabei unterstützt die grafische Ausgabe eines Konfidenzintervalls für φ . Sollte $\varphi = 0$ im Intervall enthalten sein, empfiehlt es sich, im Weiteren mit $\varphi = 0$ zu arbeiten.

2. VK

Das Leistungspotential einer Sorte kann nicht fehlerfrei ermittelt werden, sondern wird außer von den Fehlern in den Einzelversuchen besonders durch Wechselwirkungen mit der Umwelt (Jahre, Orte, Regionen ...) beeinflusst. Ausgehend von dem Ziel: „Vorhersage der wahrscheinlichsten Sorteneffekte in einem neuen Jahr an einem beliebigen Standort im Ziel-Anbaugebiet“ stellt sich die Frage, wie die Einzelversuchsergebnisse über alle Orte und Jahre je Sorte verdichtet werden sollten, wenn die einzelnen Sorten z.B. in den Jahren sehr unterschiedlich oft geprüft wurden (z.B. WP- versus LSV-Jahre). Bei der mit der HGM vorgeschlagenen vollen Ausnutzung der verfügbaren Datenbestände (LSV+WP+EUSV; >>3-jährig; Einbeziehung von Nachbarregionen ...) ist die Besetzung der Sorten in den einzelnen Umwelten regelmäßig extrem differenziert und sehr sortenspezifisch, also unbalanciert.

VK schätzt Varianzen, welche das Ausmaß der verschiedenen Sorte×Umwelt-Wechselwirkungen quantifizieren. Diese Varianztabelle wird auf dem Computer als *.sas7bdat-Datei gespeichert und nachfolgend im Verfahren MW genutzt. Das gemischte Modell führt mit diesen Varianzen in MW zu einer bestmöglichen Verdichtung / Gewichtung der Informationen aus den Urdaten für die Schätzung der sortenspezifischen Erwartungswerte.

Ein spezielles ‚Angebot‘ der HGM ist die optionale Einbeziehung von Daten aus standortkundlich ähnlichen Nachbargebieten für die Aussagen für ein Zielgebiet. Ergebnisse aus Nachbargebieten werden dann geringer gewichtet als Ergebnisse aus Zielgebieten. Die Varianzschätzungen optimieren diese Gewichtungsabstufung sortenspezifisch (!) in der Weise, dass der Gewinn an Präzision für das Zielgebiet jeweils stärker ist als ein möglicher Nachteil durch spezifische Effekte in Nachbargebieten. Die Sortenspezifika schlägt sich dabei in der Weise nieder, dass bei sehr jungen Sorten mit noch sehr wenigen Ergebnissen ($N \leq 8$) Versuche aus Nachbargebieten nur wenig geringer gewichtet werden als die des Zielgebietes, während mit zunehmender Datenbasis im Zielgebiet ($N \geq 16$ mehrjährig geprüfte Sorten) die Daten aus Nachbargebieten einen immer geringeren Einfluss erhalten.

Ist für das Verfahren MW die Berücksichtigung der Einzelversuchsfehler (SE) zur optimierten Gewichtungsabstufung in der Serienauswertung vorgesehen, so muss dies bereits in VK vorbereitet werden, da die Varianzstruktur beeinflusst wird. Die Gewichtungsabstufung in MW hängt von den Relationen der Sorte×Umwelt - Varianzen zur Restvarianz in Einzelversuchen ab.

3. MW

Hauptziel der Auswertung ist die Schätzung bzw. Vorhersage der wahrscheinlichsten Sorteneffekte in einem neuen Jahr an einem beliebigen Standort im Ziel-Anbaugebiet. Die einbezogenen Jahre und Orte bilden die Stichprobe, nicht den Aussagebereich. Diese Schätzwerte sind unter Berücksichtigung aller anderen wertbestimmenden Sorteneigenschaften die beste Grundlage für die Sortenwahl für ein neues Jahr, dessen konkrete Spezifik (Witterung, Befallssituation ...) zur Saatzeit ja noch nicht absehbar ist.

Dabei können mit MW alle verfügbaren langjährigen Daten - z.B. inklusive der den LSV zeitlich vorgelagerten WP und EUSV - einbezogen werden (trotz der dann hochgradig unbalancierten Datenstrukturen). Außerdem können benachbarte standortkundlich ähnliche Nachbargebiete mit einfließen, wobei diese sortenspezifisch ein optimal verringertes Gewicht erhalten (s.a. Erläuterungen zu VK). Eine durchgängig orthogonale Bezugsbasis ist nicht erforderlich – das gemischte Modell nutzt automatisch sämtliche vorhandenen Brückensorten zwischen Versuchen. Viele Brücken (z.B. VRS) stabilisieren das Ergebnis. Für ein vorzugebendes Zielgebiet werden Sorten-Erwartungswerte in der Weise bestimmt, dass jede Sorte auf eine einheitliche Ebene / Vergleichsbasis / mittleres absolutes Ertragsniveau so projiziert wird, dass alle Sorten trotz unterschiedlicher Datenbasis optimal und unverzerrt miteinander vergleichbar sind (Ismeans, Adjustierung). Eine Relativierung erfolgt bei Bedarf im Nachgang außerhalb dieses Verfahrens durch den Anwender.

ausgewählte Nutzer-Optionen im Verfahren MW:

- ✓ Nutzung einer Transformation (Basis ist PHI; muss mit gleichem Wert in VK berücksichtigt sein)
- ✓ Nutzung von Nachbargebieten (muss schon in VK vorgegeben sein)
- ✓ Nutzung der Gewichtung nach Versuchspräzision in Einzelversuchen (SE)
(muss schon in VK vorgegeben sein)
- ✓ Vorbestimmte Sortierung der Sorten in den Outputs

wesentlichste Outputs von,MW:

- adjustierter Mittelwert je Sorte
- adjustierter Mittelwert je Einzelversuch
- Standardfehler jedes Mittelwertes als Genauigkeitsmaß
- Ökovalenz (Ökostabilität) je Sorte
- Ökoregression (Reaktion einer Sorte auf gegensätzliche Gunstbedingungen) je Sorte
- Plausibilitätscheck – Visualisierung und Zuordnung auffälliger Urwerte

Hinweis zur Anzahl einbezogener Jahre

Das Verfahren MW kann zeitlich auf diejenigen letzten Jahre begrenzt werden, in denen auch die älteren der aktuell noch relevanten Sorten schon im Datensatz enthalten sind. Da möglichst viele ‚Sortenbrücken‘ zwischen einzelnen Versuchen und Jahrgängen die Adjustierung fördern, sollte man eher etwas mehr als zu wenige Jahre einbeziehen – bei den meisten Kulturarten sind die letzten 6 bis 7 Jahre eine sinnvolle Größenordnung in MW.

Die Verfahren PHI und MW sind umso nützlicher, je besser sie die Grundgesamtheit treffen – sie sollten also nicht auf die u.U. reduzierte Stichprobe der in MW verwendeten Jahre begrenzt werden. Viele Jahre zu nutzen ist insofern besonders wichtig, als die Jahresvarianzen und die Interaktionen von Sorten mit Jahren i.d.R. besonders groß sind. Sofern man viele Jahre im Datenbestand gepflegt hat, werden die Parameter (φ und die *Varianz – Komponenten*) mit ca. 12 einbezogenen Jahren i.d.R. sehr gut und stabil geschätzt. Mit so langen Datenbeständen kann man daher MW auch mal mit den Parametern aus PHI und VK einer vorherigen Auswertung ablaufen lassen, z.B., wenn bei Winterungen eine möglichst schnelle Neuempfehlung schon nach der aktuellen Ernte ansteht und die Auswertungsaufwand verkürzt werden soll.

4. DAT

Für sehr langjährige überregionale Auswertungsserien wird die Harmonisierung in PIAF zunehmend komplex und kompliziert, und die Ausgabe einer Serien-ADS für die HGM ist zeitaufwändig. Daher lesen einige Nutzer die älteren Daten aus und legen sie z.B. als externe Excel-Datei ab. DAT liest Daten aus einer Serien-ADS und speichert sie in einer *.xlsx-Datei („externe Daten“) so ab, dass sie von den Hohenheim-Gülzower-Verfahren unbearbeitet weiterverwendet werden kann. Eine Harmonisierung ist dann u.U. nur noch jährlich für die einjährigen neuen Versuche nötig, was sehr unkompliziert ist. In den drei Kernverfahren kann man dann entscheiden, ob nur eine ADS eingelesen wird oder nur eine externe Excel-Datei oder ob eine ADS mit einer externen Excel-Datei kombiniert wird.

5. Ergänzende Excel-Tabellen

In den HGM-Verfahren können ergänzende Informationen genutzt werden, um die Schätzungen oder die Outputs zu verbessern. Sie müssen in Excel-Tabellen mit einer bestimmten Struktur. Deshalb werden mit den HGM-Verfahren Beispiel-Dateien als Templates mit heruntergeladen:

daten.xlsx	Statt Daten aus einer ADS einzulesen, können externe Daten genutzt werden.
gruppen.xlsx	Zuordnung von Sorten zu Gruppen; nur bei Differenzierung in Regionen relevant; nur sinnvoll, wenn die Gruppen im Auswertungsmerkmal nicht ohne weiteres vergleichbar sind (z.B. Protein zwischen E- und C-Weizen); soll überhöhte Gewichtung von Nachbargebieten vermeiden
phistart.xlsx	Sofern für die Merkmale Vorkenntnisse zum Transformationsparameter phi vorliegen, kann die Vorgabe nahegelegener Startwerte das Verfahren PHI beschleunigen.
reg.xlsx	Zuordnung von BKR zu Anbaugebieten; nur bei Differenzierung in Regionen relevant
reglim.xlsx	Wenn eine Reihe von Anbaugebieten gemeinsam ausgewertet wird, die jeweils definierten Zielgebiete aber nur eine Teilmenge der anderen als Nachbargebiet nutzen sollen, so wird dies in dieser Datei definiert. Der Ablauf von VK und mehrerer MW-Runs ist damit effizienter.
sorti.xlsx	Enthält für ausgewählte Sorten die Vorgabe der gewünschten Sortierreihenfolge bei der Mittelwertausgabe. Zudem kann festgelegt werden, dass für diese Sortenauswahl eine Schätzung und Ausgabe auch dann erfolgt, wenn sie nicht die ansonsten vorgegebene Mindestanzahl an Versuchsdaten erfüllt.