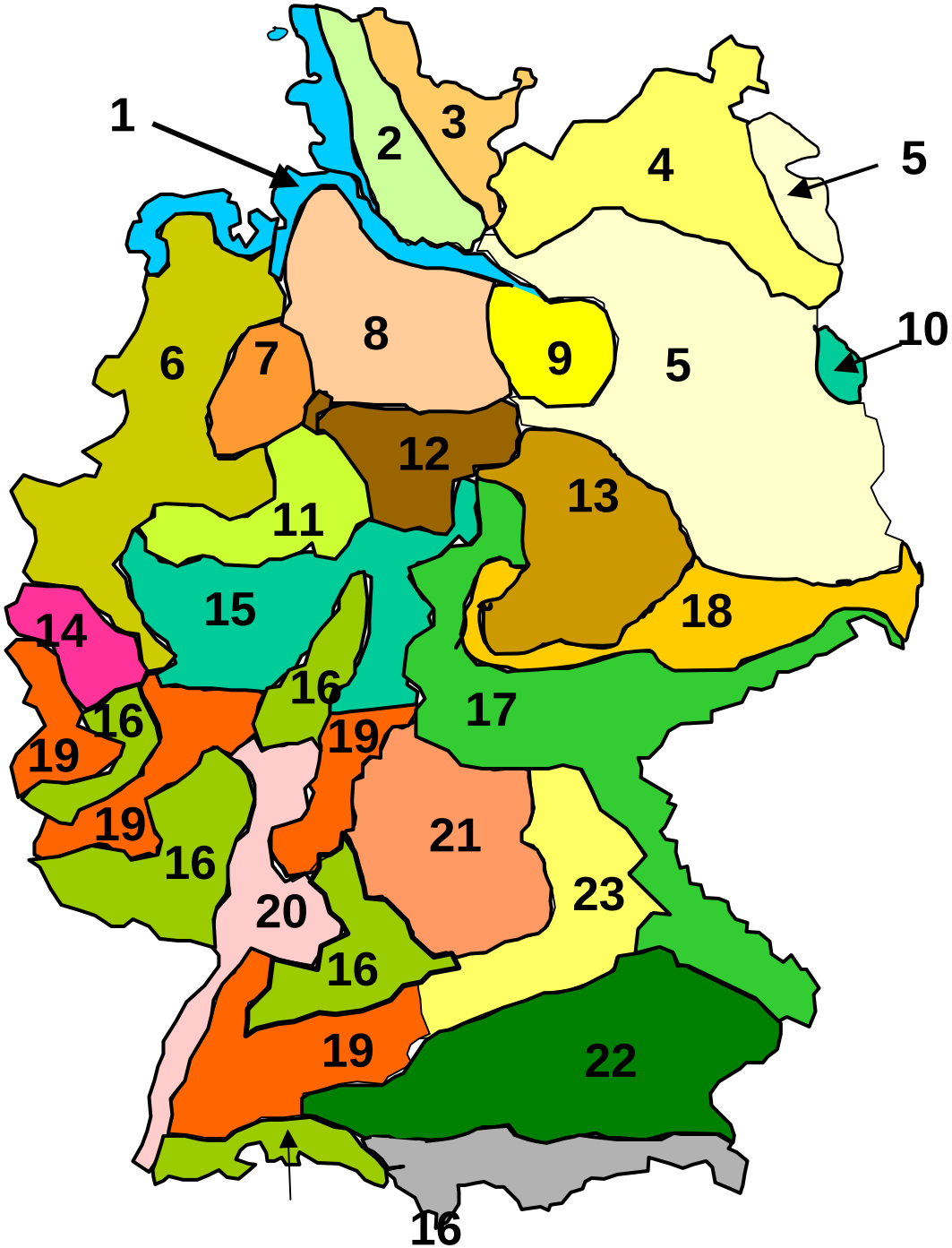
# Auswertung von Sortenversuchen mit überlappenden Großräumen. I. Idee und Ansatz.

Prof. Dr. H.-P. Piepho

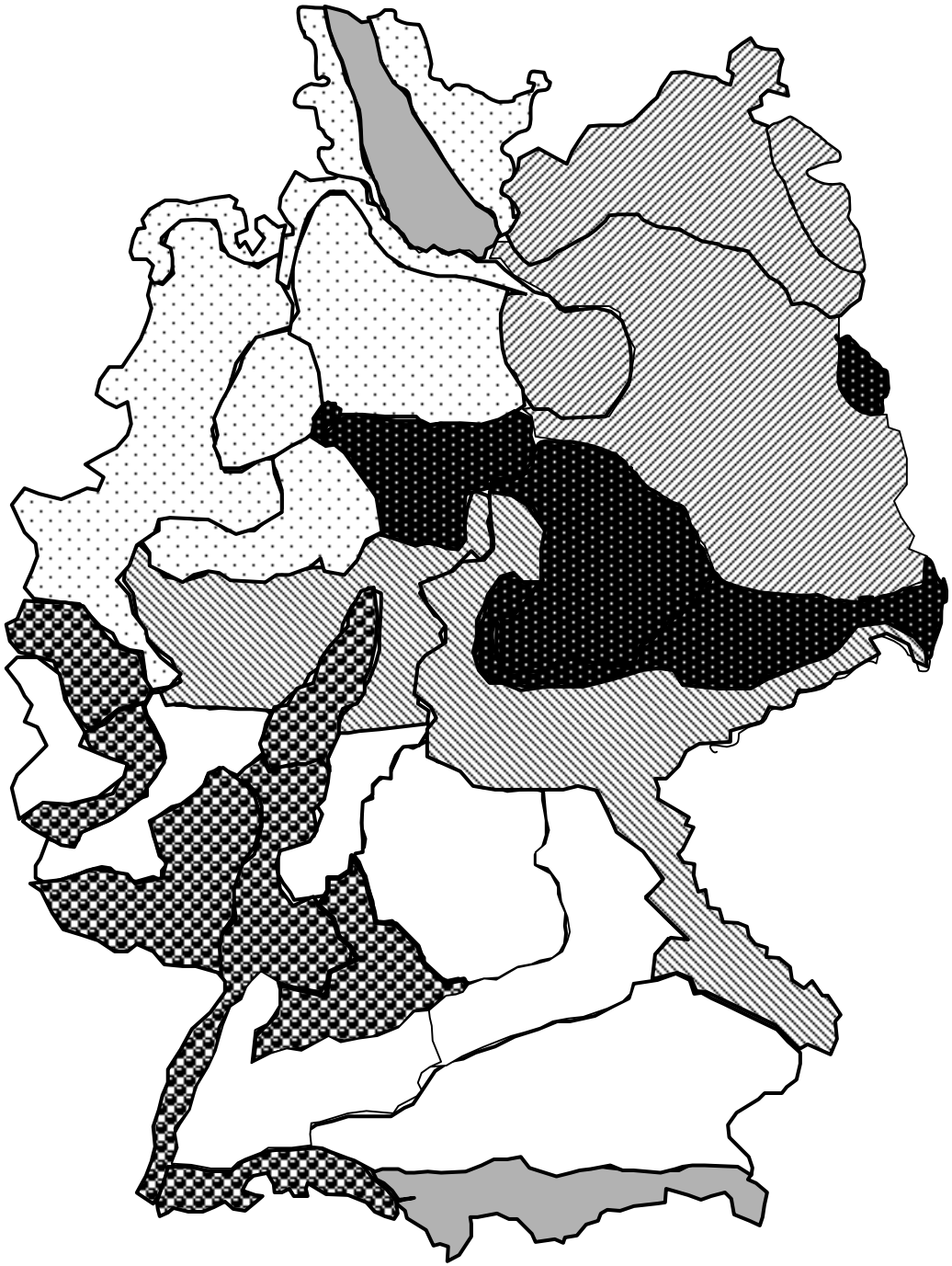
Universität Hohenheim, Fachgebiet Bioinformatik

# Übersicht

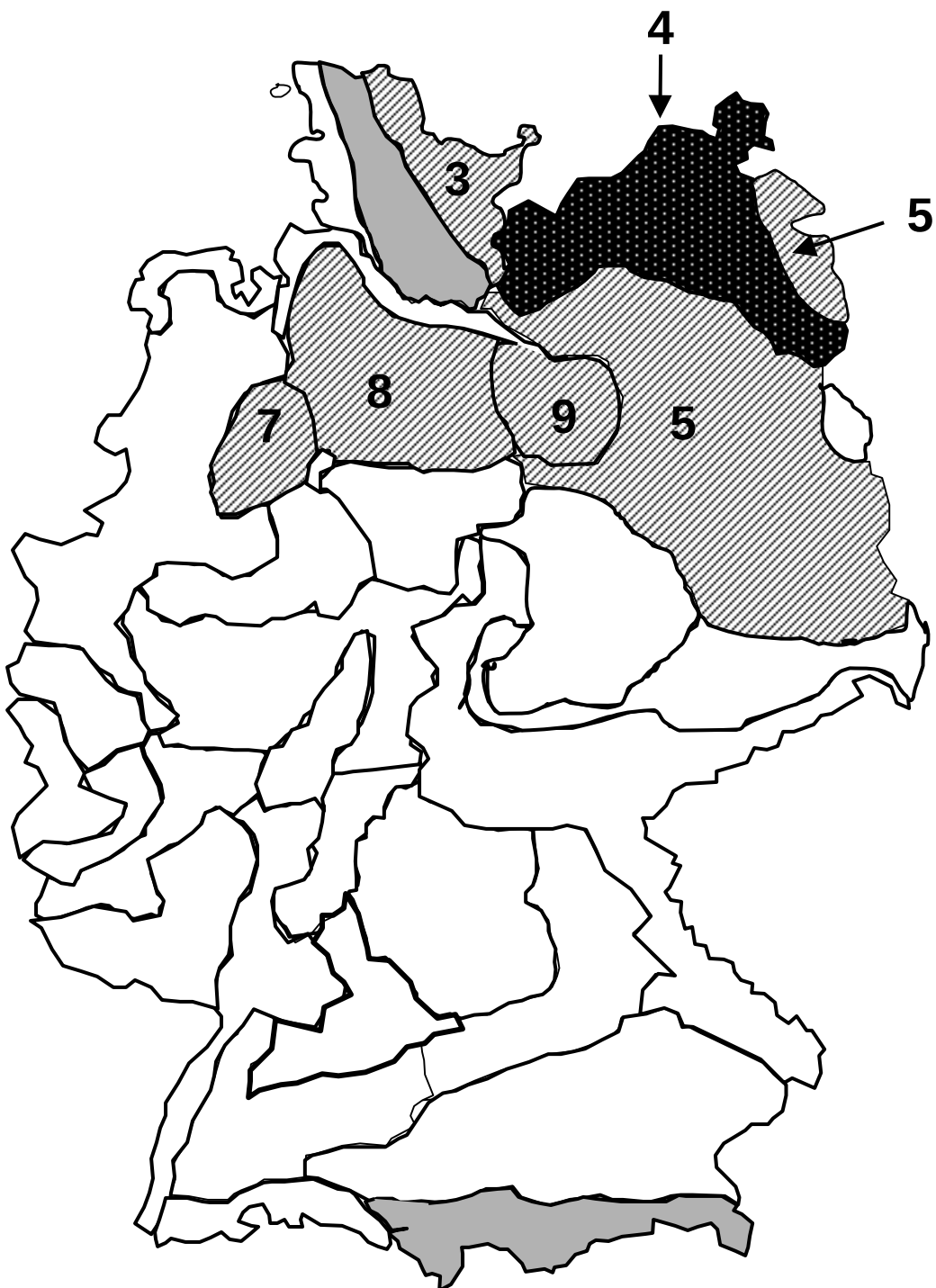
- Das Konzept „Überlappende Großräume“
- Mittlerer Vorhersagefehler
- Optimale Gewichte
- Ergebnisse bei Weizen
- Selektion – ist sie ignorierbar?
- Schlussfolgerungen




# Anbaugebiete Getreide Deutschland 17.06.2003



Großräume
Getreide
Vorschlag
Jentsch




Zielanbaugebiet

Nachbaranbaugebiet

Keine Daten

Andere Anbaugebiete

# Großraum für Anbaugebiet 4

# Überlappende Großräume

- 23 Anbaugebiete
- Ertragsschätzung für Zielgebiet (eines der 23)
- Nutzung der Information von Nachbargebieten

$\Rightarrow$ Auswertungsverfahren entwickeln

$\Rightarrow$ Genauigkeitsgewinn abschätzen

$\Rightarrow$ Optimale Verteilung der Prüfkapazität

} exemplarisch

# Grundlage des Ansatzes

- In jedem Anbaugebiet Serienauswertung
- Gemischtes Modell mit Faktoren
  SORTE (fix), JAHR, ORT und ggf. TYP
- Mittelwerte einer Sorte in den 23 Anbaugebieten

$y_1$ = Ertrag der Sorte im Zielgebiet

$y_2$, $y_3$, ...., $y_{23}$ = Erträge der Sorten in den anderen Anbaugebieten

# Daten nur aus Zielgebiet

Ertragsschätzung:

$y_1$ = Ertrag der Sorte im Zielgebiet

# Daten auch aus Nachbargebieten (1)

Ertragsschätzung: $\bar{y} = \frac{y_1 + y_2 + y_3 + ... + y_R}{R}$

$y_1$ = Ertrag der Sorte im Zielgebiet

$y_2$, $y_3$, ...., $y_R$ = Erträge der Sorten in Nachbargebieten

# Daten auch aus Nachbargebieten (2)

Ertragsschätzung:

$$\bar{y}_w = w_1 \times y_1 + w_2 \times y_2 + w_3 \times y_3 + \ldots + w_R \times y_R$$

$w_r$ = Gewichte

Ungewichtet: $w_r = 1/R$

# Was wird geschätzt?

*E* = Erwartungswert der Sorte in Zielgebiet

= Durchschnitt über alle Orte und unendlich viele Jahre in Zielgebiet

$\bar{y}_w$ ist nur Schätzwert für *E*

Ziel: $(\bar{y}_w - E)^2$ minimieren

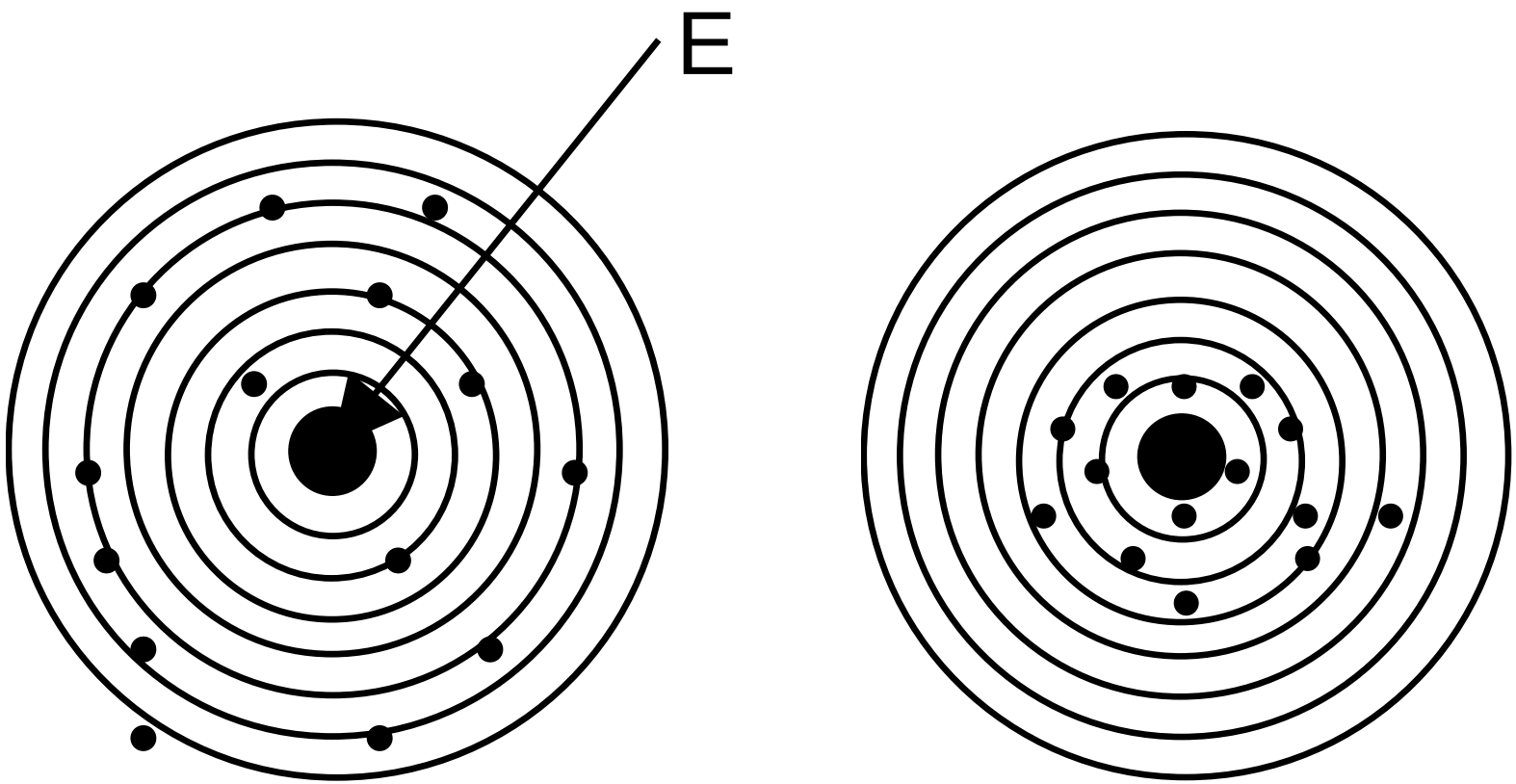
# Zwei Varianten

- Optimale Gewichte ohne Restriktion

  $\Rightarrow$ BLUP, Schrumpfung

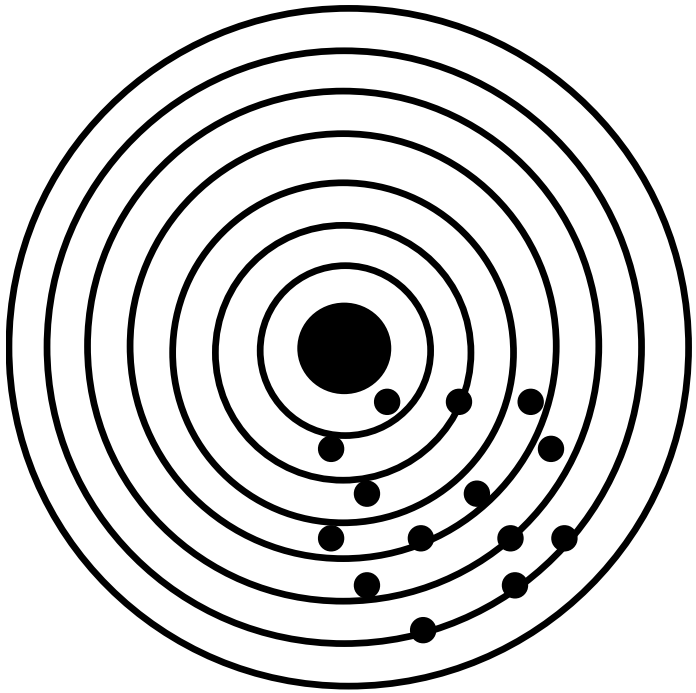
- Optimale Gewichte mit Restriktion $\sum w_r = 1$

$\Rightarrow$ kein gewöhnliches BLUP, keine Schrumpfung

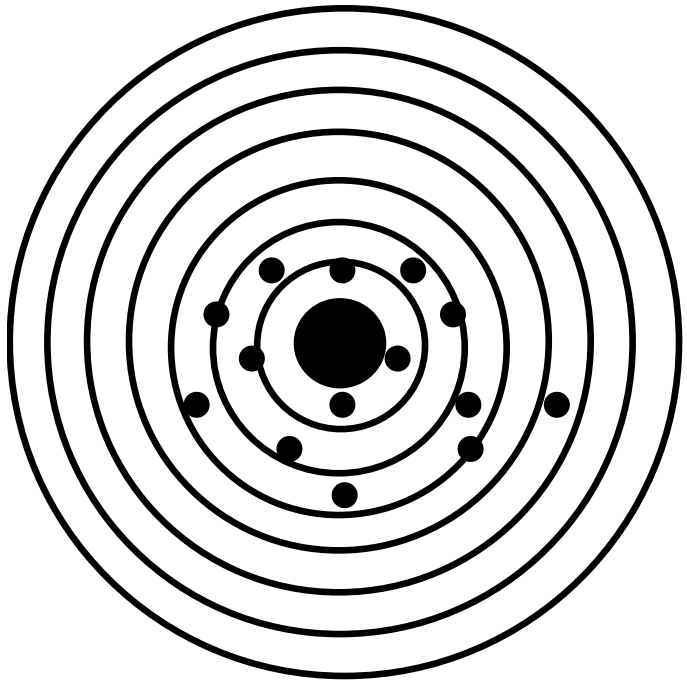



Großer Standardfehler

Kleiner Standardfehler



Mit Verzerrung



Ohne Verzerrung

# Vorhersagefehler

„Vorhersagefehler“ = „Standardfehler“ + „Verzerrung“

$$\left(\bar{y}_w - E\right)$$

# Mittlerer quadratischer Vorhersagefehler (MSEP)

1. Betrachte Sortendifferenzen

2. Quadriere Vorhersagefehler

3. Betrachte Erwartungswert des MSEP über alle Sortenpaare (Sorteneffekte als zufällig betrachtet)

# Zweischritt-Analyse

1. Serien-Auswertung über Orte und Jahre je Anbaugebiet (Sorten formal als fest angenommen)
   $\Rightarrow$ Adjustierte Sortenmittelwerte je Region

2. Kombination der Sortenmittelwerte über Anbaugebiete:

$$\boldsymbol{w}'\hat{\boldsymbol{\eta}}_{\boldsymbol{i}} \quad , \quad \boldsymbol{\eta_i} = \left(\eta_{i1}, \eta_{i2}, \ldots, \eta_{iR}\right)'$$

$\eta_{ir}$ = Erwartungswert der Sorte *i* im *r*-ten Anbaugebiet

# Varianz-Kovarianz-Struktur und MSEP

$$\text{var}\left(\hat{\boldsymbol{\eta}}_i\right) = \boldsymbol{\Sigma}_g + \boldsymbol{\Sigma}_e$$

$\Sigma_g$ = genetische Varianz-Kovarianz-Matrix

$\Sigma_e$ = umweltbedingte Varianz-Kovarianz-Matrix

$$MSEP = \left(\boldsymbol{w} - \boldsymbol{u}_r\right)' \boldsymbol{\Sigma}_g \left(\boldsymbol{w} - \boldsymbol{u}_r\right) + \boldsymbol{w}' \boldsymbol{\Sigma}_e \boldsymbol{w}$$

# Optimale Gewichte (ohne Restriktion)

$$\hat{\boldsymbol{w}}' = \boldsymbol{u}_r' \boldsymbol{\Sigma}_g \left( \boldsymbol{\Sigma}_g + \boldsymbol{\Sigma}_e \right)^{-1}$$

$\Sigma_g$ = genetische Varianz-Kovarianz-Matrix

$\Sigma_e$ = umweltbedingte Varianz-Kovarianz-Matrix

$\boldsymbol{u}_r$ = Einheitsvektor für *r*-tes Anbaugebiet

$\Rightarrow$ Das ist BLUP! MSEP minimal!

# Optimale Gewichte (mit Restriktion)

- Lagrange-Multiplikator für Nebenbedingung $\sum w_r = 1$

  $\Rightarrow$ Gleichungssystem

$$\hat{\boldsymbol{v}}' = \left(2\boldsymbol{u}_r'\boldsymbol{\Sigma}_g \mid 1\right)\begin{pmatrix} 2\left(\boldsymbol{\Sigma}_g + \boldsymbol{\Sigma}_e\right) & \mathbf{1} \\ \mathbf{1}' & 0 \end{pmatrix}^{-1}$$

$\boldsymbol{v} = (\boldsymbol{w}, \lambda)$

$\Sigma_g$ = genetische Varianz-Kovarianz-Matrix

$\Sigma_e$ = umweltbedingte Varianz-Kovarianz-Matrix

$\boldsymbol{u}_r$ = Einheitsvektor für *r*-tes Anbaugebiet

**Tab.1**: Varianzkomponentenschätzungen für Weizen aus allen Daten und aus den LSV-Daten (Ertrag, Intensität 1, homogene S*R-Varianz)

| Varianzkomponente | Alle Daten | Qualitätsklassen gewichtetes Mittel | Nur LSV |
|---|---|---|---|
| Sorte | **14,28** | **9,65** | **13,98** |
| Sorte*Region | 1,32 | 1,16 | 1,44 |
| Sorte*Jahr | 3,90 | 3,15 | 2,34 |
| Sorte*Ort | 2,69 | 2,53 | 3,09 |
| Sorte*Jahr*Ort | 5,94 | 6,28 | im Restfehler enthalten |
| Sorte*Region*Jahr | 2,92 | 2,59 | 3,35 |
| Jahr | 10,93 | 12,41 | 14,28 |
| Ort | 28,75 | 28,75 | 37,59 |
| Jahr*Ort | 41,82 | 43,86 | 18,59 |
| Jahr*Ort*Typ | 21,70 | 19,85 | 54,87 |
| Jahr*Region | 21,75 | 20,76 | 30,22 |
| Restfehler | 14,28 | 12,87 | 20,61 |

**Tab. 2**: Genetische Korrelation zwischen den sechs Anbaugebieten (Anbaugebiet 4 und Nachbarn) (Weizen)

| Anbaugebiet | Korrelation mit Anbaugebieten | | | | | |
|---|---|---|---|---|---|---|
| | 3 | 4 | 5 | 7 | 8 | 9 |
| 3 | 1 | 0,99 | 1,00 | 1,00 | 0,95 | 1,00 |
| 4 | | 1 | 0,99 | 0,99 | 0,94 | 0,99 |
| 5 | | | 1 | 1,00 | 0,95 | 1,00 |
| 7 | | | | 1 | 0,95 | 1,00 |
| 8 | | | | | 1 | 0,95 |
| 9 | | | | | | 1 |

**Tab. 3:** Gewichte der Nachbaranbaugebiete und des Zielanbaugebiet 4 bei gemeinsamer Auswertung der Anbaugebiete 3-5 und 7-9 sowie Angaben zur Zahl der WP- und LSV-Orte (Angaben gerundet, wobei nicht auf Null abgerundet wurde) (Weizen)

| Anbaugebiet | 3 | **4** | 5 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|
| Gewicht | 0,18 | **0,35** | 0,19 | 0,10 | 0,09 | 0,08 |
| Zahl WP-Versuche | 3 | **6** | 1 | 1 | 1 | 0 |
| Zahl LSV-Versuche | 12 | **15** | 21 | 9 | 9 | 9 |

**Tab. 4:** Gewichte der Nachbaranbaugebiete und des Zielanbaugebiet 23 bei gemeinsamer Auswertung der Anbaugebiete 19 und 21-23 sowie Angaben zur Zahl der WP- und LSV-Orte (Angaben gerundet, wobei nicht auf Null abgerundet wurde) (Weizen)

| Anbaugebiet | 19 | 21 | 22 | **23** |
|---|---|---|---|---|
| Gewicht | 0,10 | 0,14 | 0,63 | **0,12** |
| Zahl WP-Versuche | 2 | 3 | 7 | **1** |
| Zahl LSV-Versuche | 9 | 9 | 18 | **3** |

**Tab. 5:** Varianz durch das Prüfsystem, quadrierte Verzerrung und MSEP bei separater Auswertung von Anbaugebiet 4 und gemeinsamer Auswertung mit Nachbaranbaugebieten (Weizen)

| | Nur AG 4 nur LSV | Nur AG 4 | AG 4 + Nachbarn |
|---|---|---|---|
| | ---------- (dt/ha)² ---------- | | |
| Varianz des Prüfsystems | 7,18 | 4,37 | 2,11 |
| Bias² (Verzerrung)² | 0 | 0 | 0,35 |
| MSEP (=Summe) | 7,18 | 4,37 | 2,46 |
| Selektionserfolg (p=20%) | 3,86 | 4,08 | 4,25 |

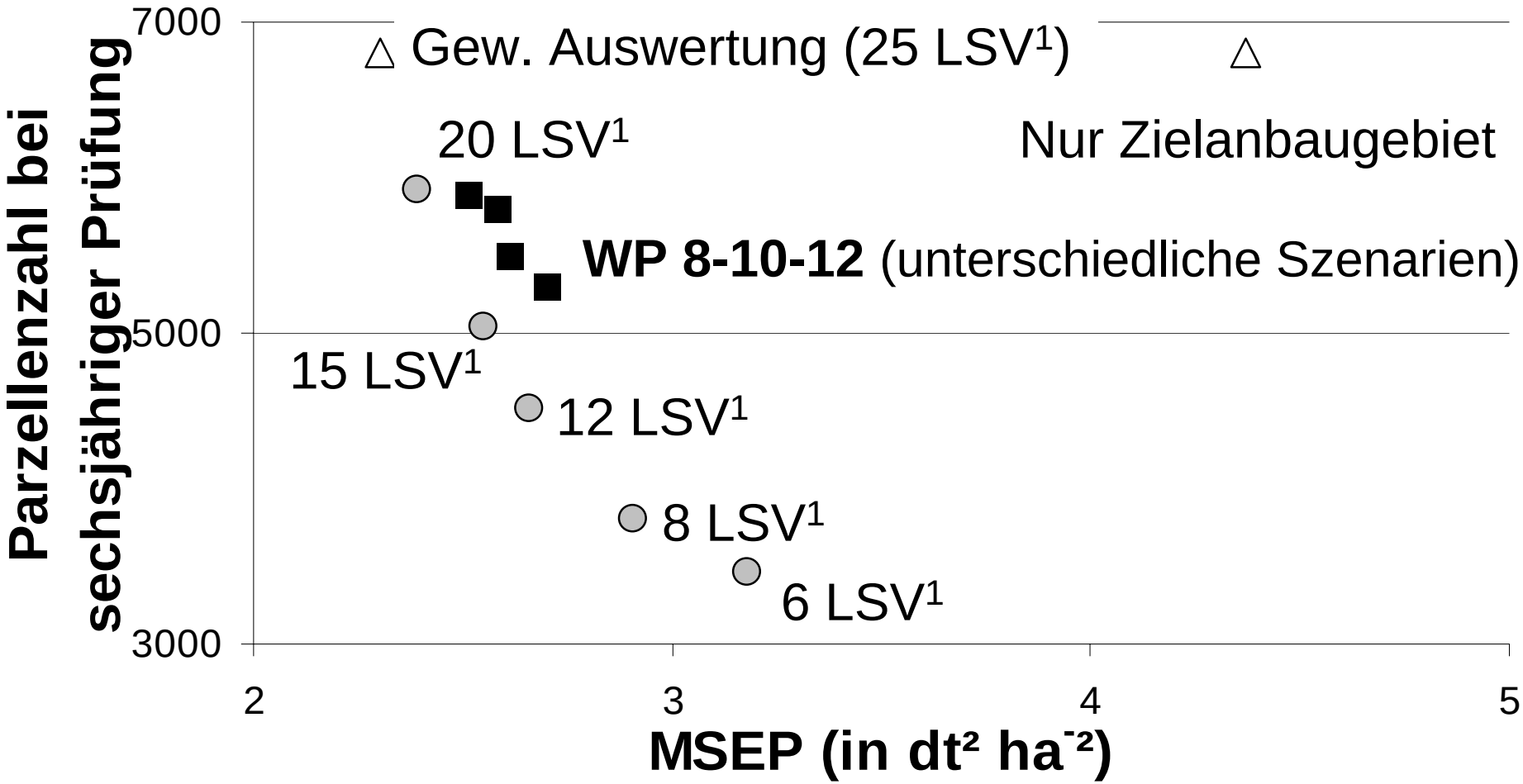



**Abb.1**: MSEP von Zielanbaugebiet 4 mit fünf Nachbaranbaugebieten bei Reduktion der WP oder LSV (Weizen).

([1]jährliche Versuchsanzahl in AG4 und Nachbarn)

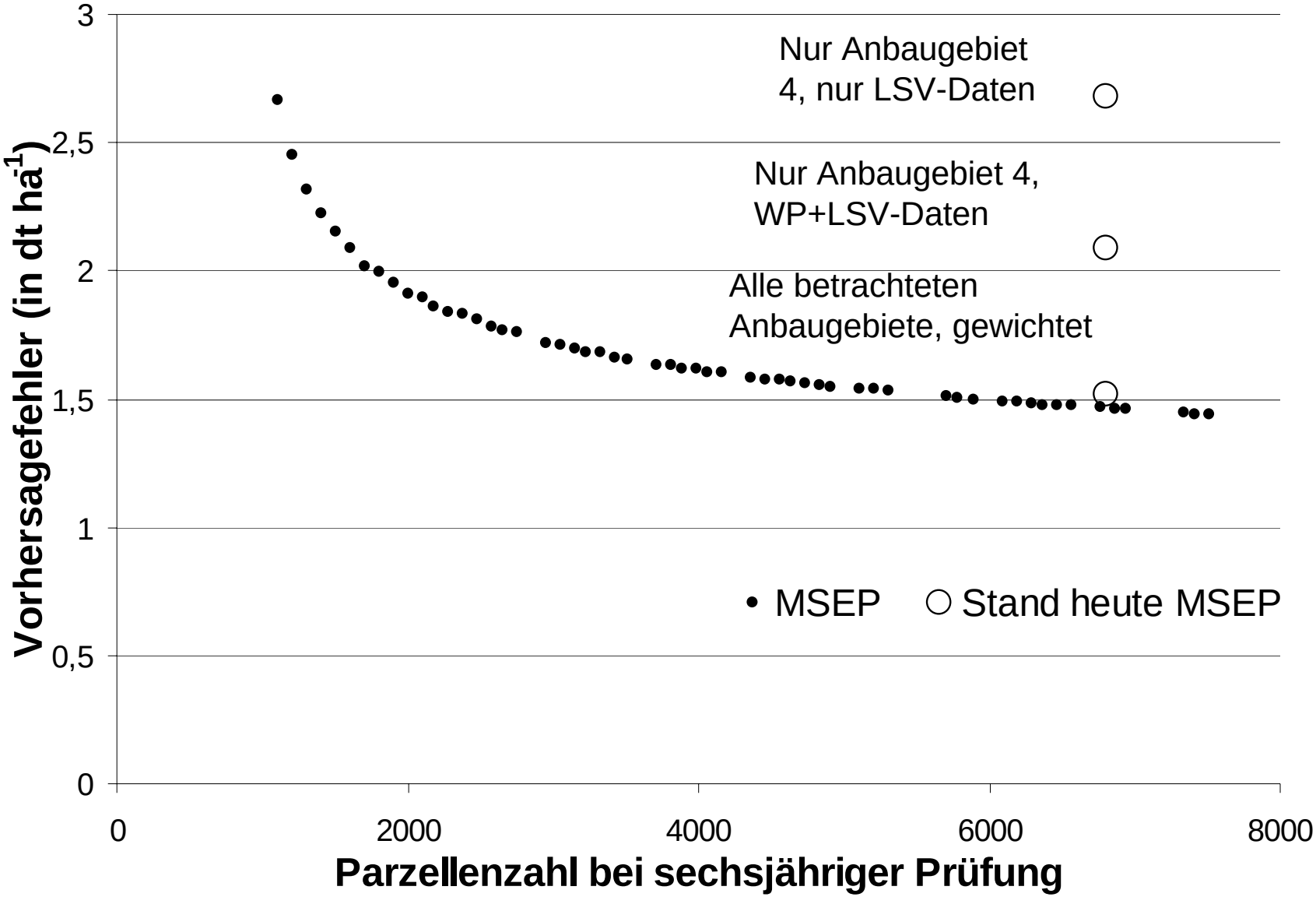



**Abb.2**: Vorhersagefehler in Abhängigkeit vom Auswertungs-verfahren und der Prüfkapazität (Zielanbaugebiet 4) (Weizen)

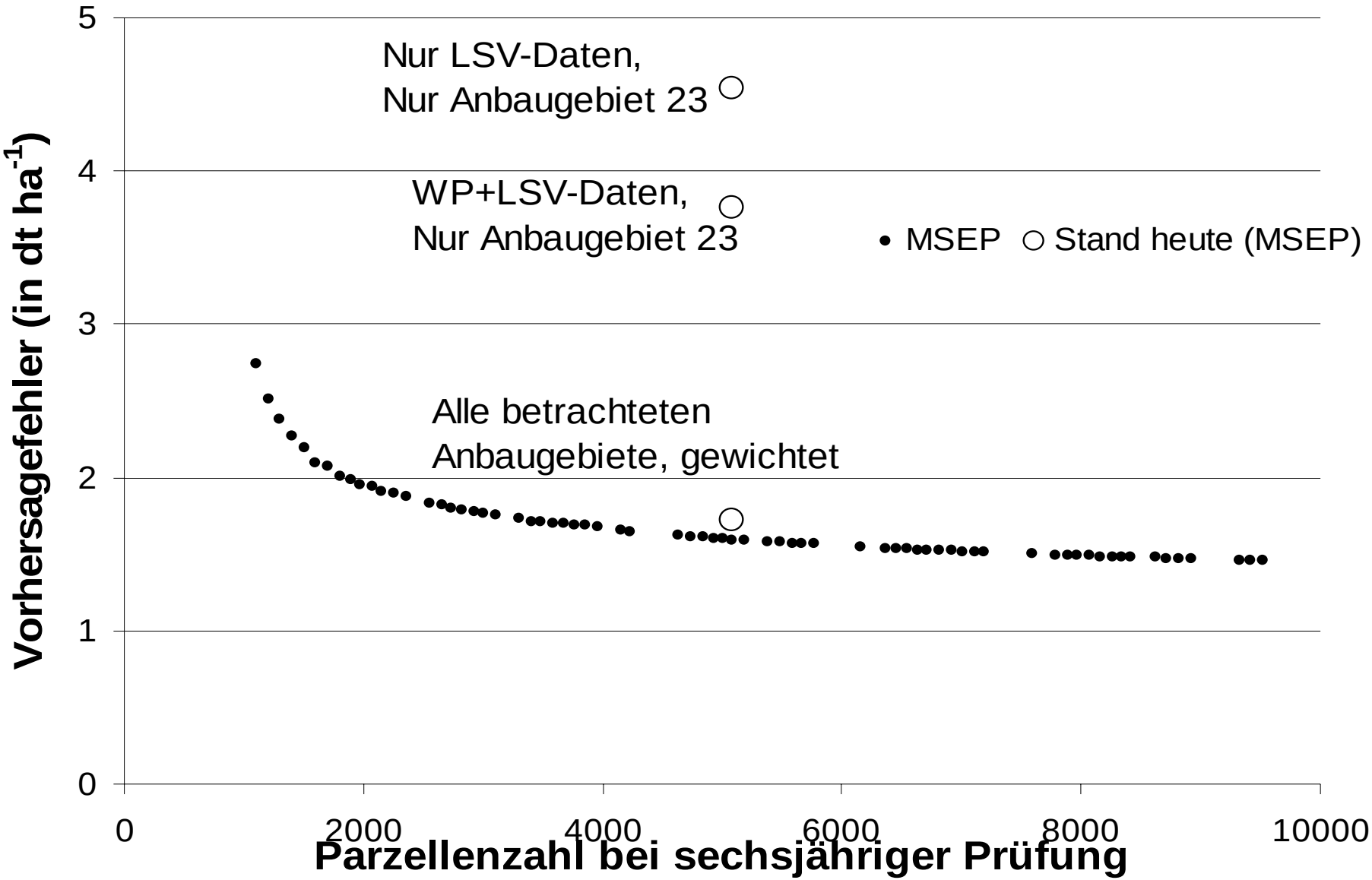



**Abb.3**: Vorhersagefehler in Abhängigkeit vom Auswertungs-verfahren und der Prüfkapazität (Zielanbaugebiet 23) (Weizen)

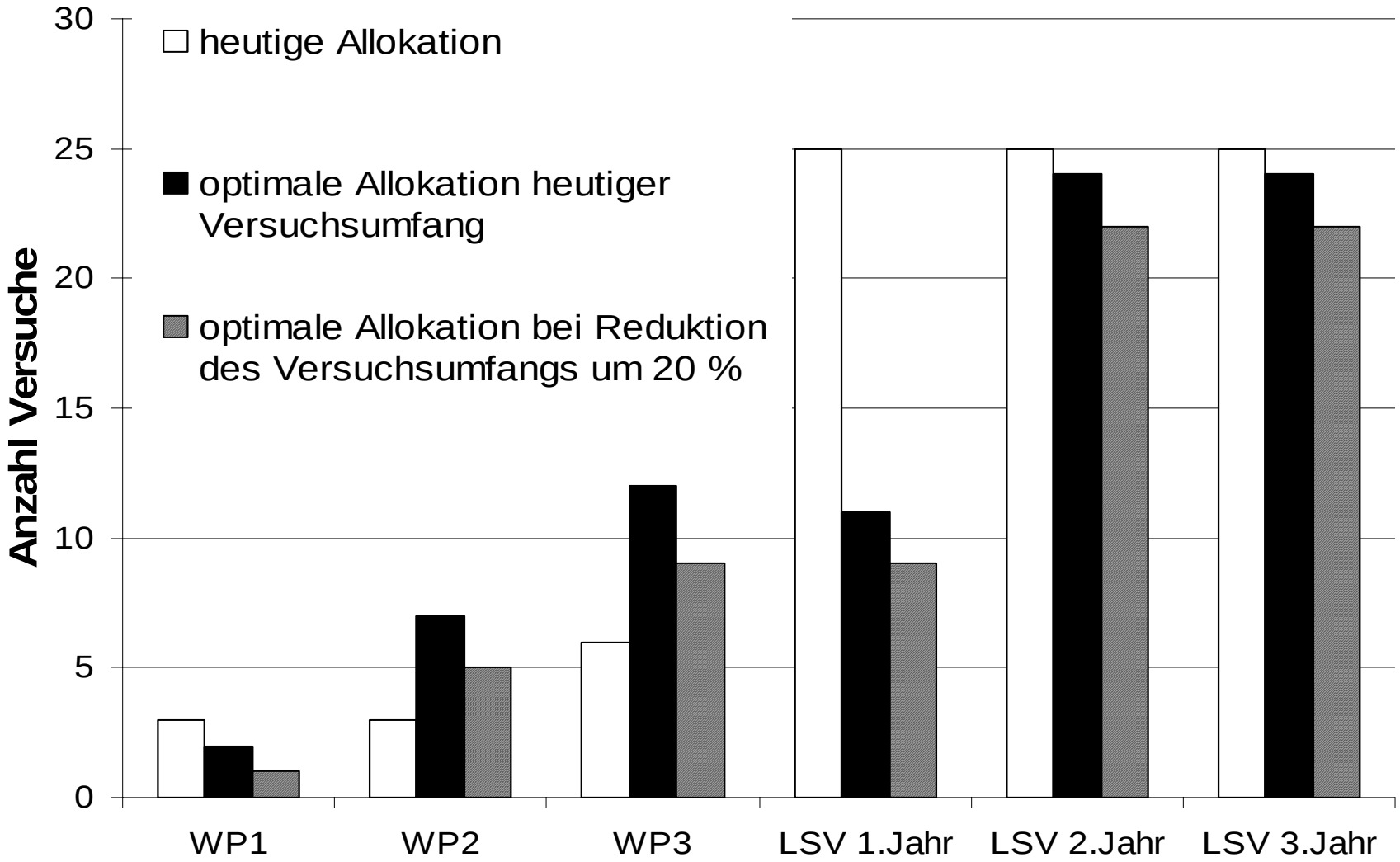



**Abb.4**: Allokation der Versuche beim Zielanbaugebiet 4 und dessen Nachbaranbaugebieten (Weizen)

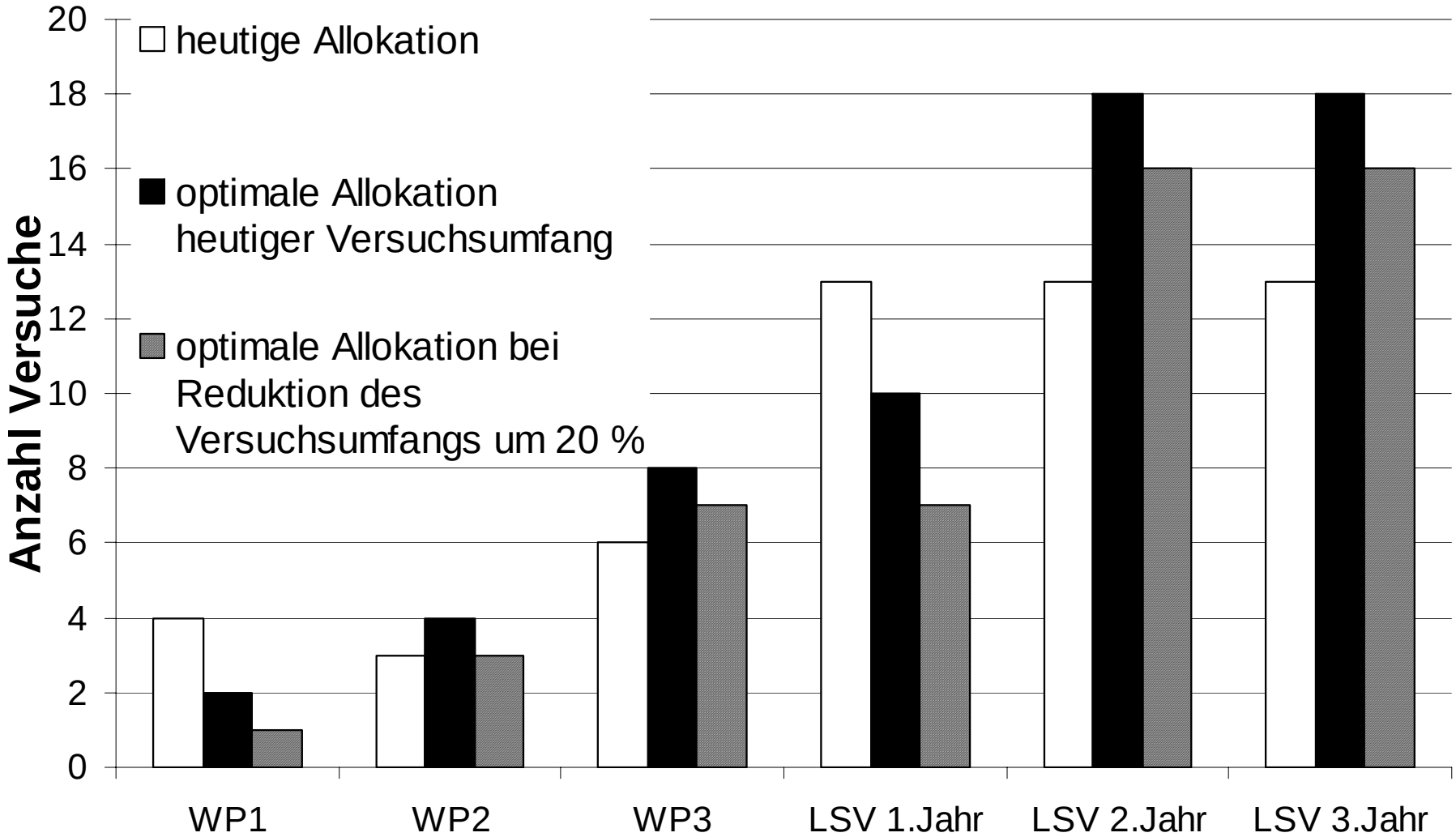



**Abb.5**: Allokation der Versuche beim Zielanbaugebiet 23 und dessen Nachbaranbaugebieten (Weizen)

# Orte wechseln erhöht Genauigkeit

**Beispiel:**

- Ein Anbaugebiet, 6 Jahre, 4 Wiederholungen
- Varianzkomponenten Winterweizen

**3 Orte pro Jahr**

Orte gleich: *s.e.d.* = 2,16
Orte wechselnd: *s.e.d.* = 1,80

**6 Orte pro Jahr**

Orte gleich: *s.e.d.* = 1,81
Orte wechselnd: *s.e.d.* = 1,61

# Selektion - ist sie ignorierbar?

$Y$ = Komplette Daten ohne Fehlwerte =($Y_{obs}$, $Y_{mis}$)

$Y_{obs}$ = Beobachtene unvollstäbndige Daten

$Y_{mis}$ = Fehlende Daten

$R$ = Fehlmuster

[Rubin (1976)]

# Fehlmuster

- **Missing completely at random (MCAR):**

  $R$ ist unabhängig von $Y_{obs}$ und $Y_{mis}$

- **Missing at random (MAR):**

  $R$ ist unabhängig von $Y_{mis}$

- **Missing not at random (MNAR):**

  $R$ ist abhängig von $Y_{mis}$ [Rubin (1976)]

# Ignorierbarkeit

- **MCAR, MAR:**

Fehlmuster ignorierbar bei likelihood-basierter Analyse

- **Sortenversuche:**

Alle Daten seit Anmeldung bei BSA in Auswertung einbeziehen und mit REML rechnen
$\Rightarrow$ Selektion ignorierbar

# Schlußfolgerungen (1 von 2)

- Neuer Auswertungsansatz (überlappende Anbaugebiete) hat hohes Potential
- 1/3 der Sorte*Ort Varianz kann durch Anbaugebiete erklärt werden (Weizen)
  $\Rightarrow$ Anbaugebiete sehr aussagekräftig
  $\Rightarrow$ aber: 2/3 der Varianz innerhalb Anbaugebiete
- Für drei Zielgebiete hohe genetische Korrelationen zu Nachbarn (Weizen)
  $\Rightarrow$ Wahl der Nachbarn war sehr gut!

# Schlußfolgerungen (2 von 2)

- Ansatz mit überlappenden Großräume sichert optimale Gewichtung und damit optimale Auswertung
- Anspruch an biometrische Auswertung steigt
- Kürzung Anzahl WP-Standorte vermindert Nutzen integriertes System
  $\Rightarrow$ im Sinne des Gesamtsystems besser LSV-Versuche als WP-Versuche einsparen
- Je mehr Versuche aus Nachbargebieten einbezogen, desto höher der Genauigkeitsgewinn