Herausforderungen bei der Validierung KI-basierter Merkmalserfassung im pflanzenbaulichen Versuchswesen:

Beispiel: Zählung Ährentragender Halme mit Smartphone-App Count-My-Crop

Dr. Andreas Büchse, BASF SE, 67056 Ludwigshafen am Rhein, DLG Ausschuss Versuchswesen in der Pflanzenproduktion

Dr. Gregor Heine, Landwirtschaftskammer Nordrhein-Westfalen, 50765 Köln-Auweiler

Prof. Dr. Gregor Fischer, Prof. Dr. Klaus-Dieter Ruelberg, Technische Hochschule Köln, Procedeon Image Science GmbH, 51491 Overath

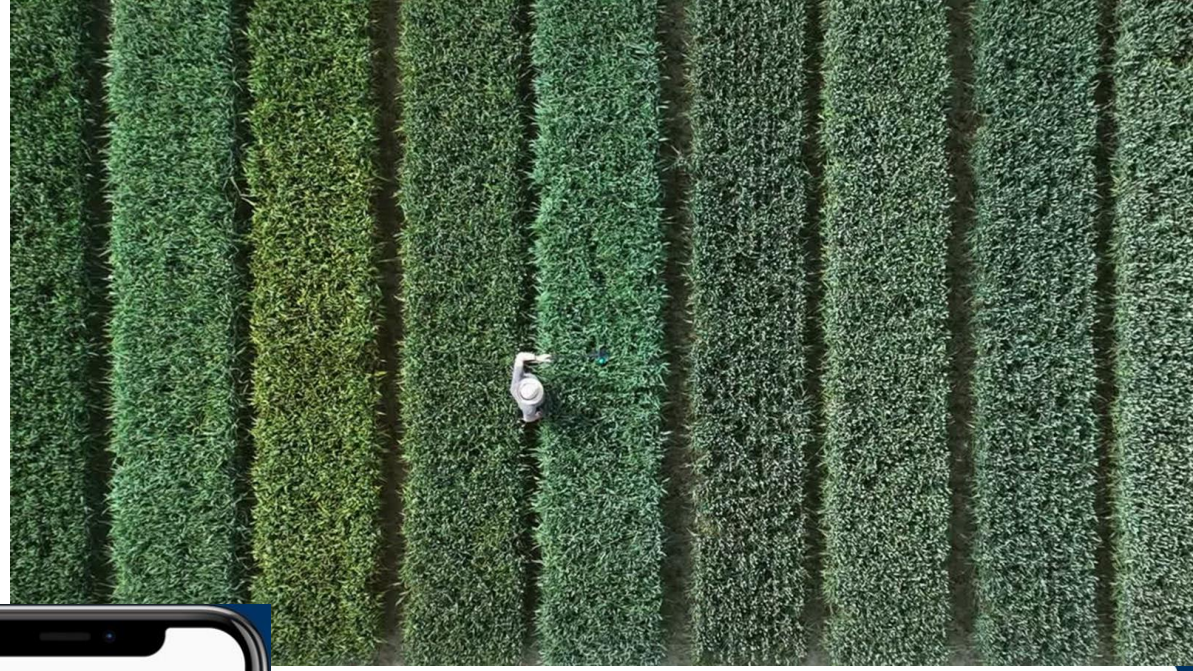
andreas.buechse@basf.com

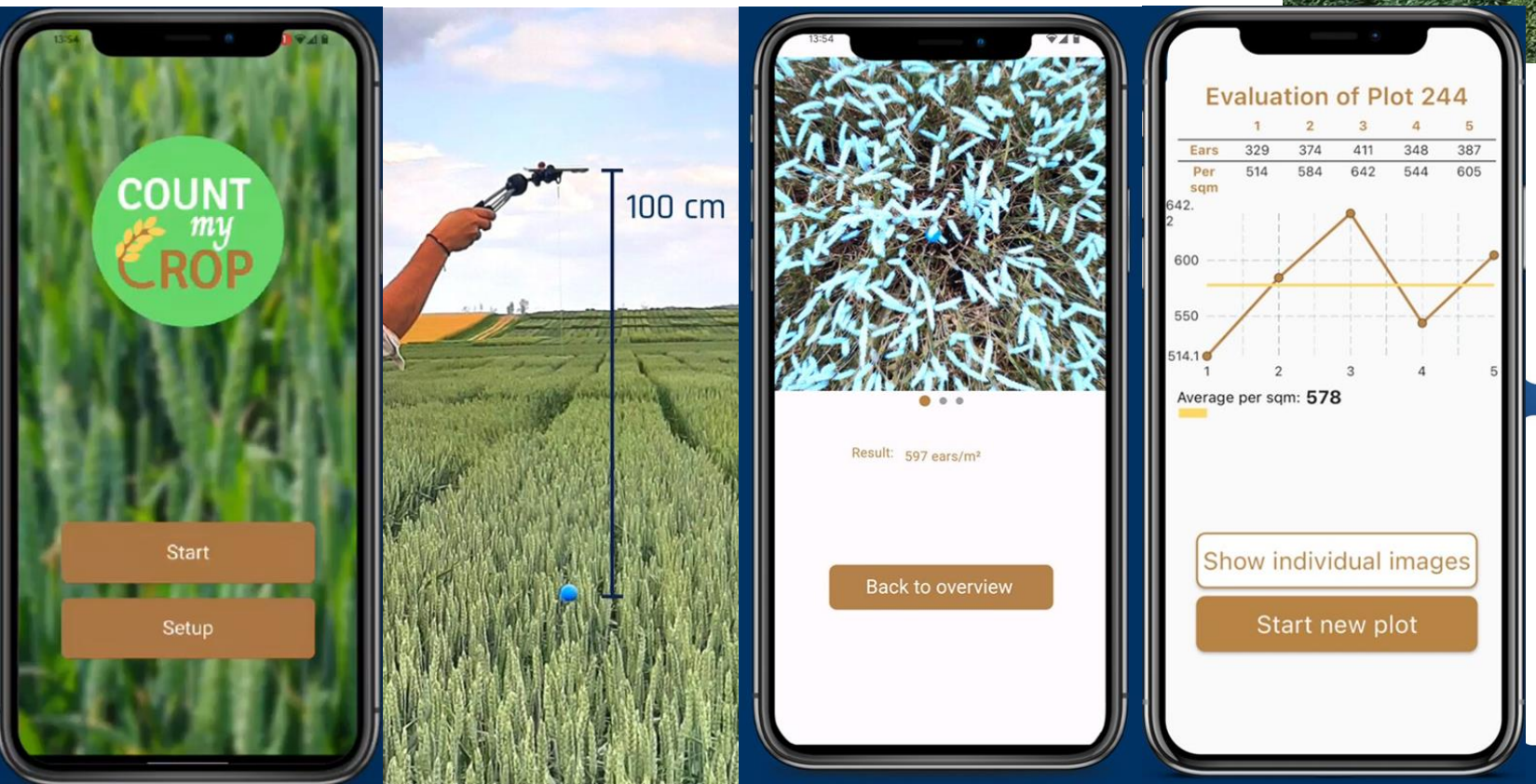
Sommertagung AG „Landwirtschaftliches Versuchswesen“, 27. und 28. Juni 2024 Fachhochschule Südwestfalen

Basierend auf: DLG-Technikertagung, Hannover, 30.-31. Januar 2024, 54. Fachtagung des DLG-Ausschusses Versuchswesen in der Pflanzenproduktion

Disclaimer - Zielstellung

- Wir zeigen Erfahrungen und Erkenntnisse bei der Validierung der Smartphone-App „***Count my Crop***“ zur Erfassung des Merkmals „Ährentragende Halme“
- Die **Prinzipien** sind auf viele Validierungen von KI-Algorithmen zur Nutzung im Pflanzenbaulichen Versuchswesen **übertragbar.**





Ausgangslage

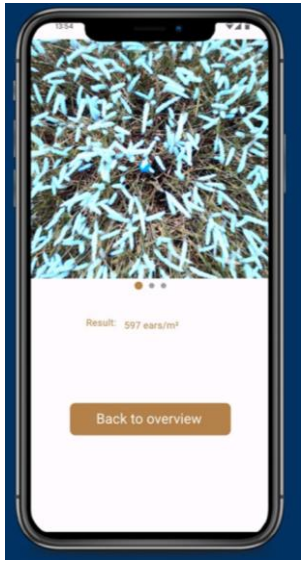
- Das Merkmal „Ährentragende Halme“ dient der Charakterisierung von Getreidesorten
- Die Erhebung des Merkmals ist im Rahmen der Durchführung der Wertprüfungen zur Sortenzulassung vorgeschrieben

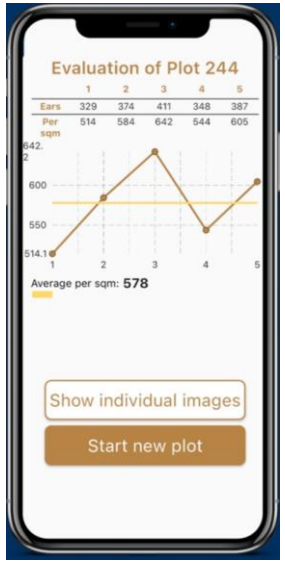
Referenzmethode „Meterzählung“

- Zählung der Ährentragenden Halme in 1 laufenden Meter einer „repräsentativen Drillreihe“ und Umrechnung auf Ähren/m²
- Die manuelle Zählung ist anstrengend und zeitintensiv!

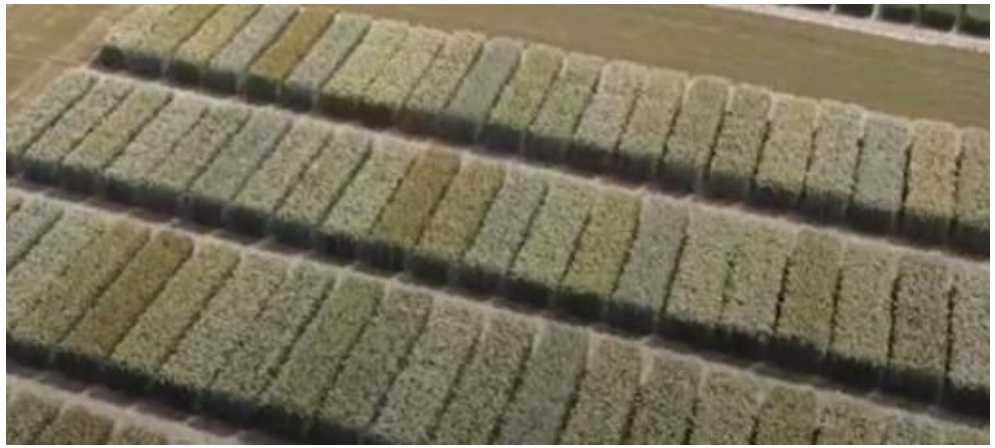
Alternativmethode „Count my Crop“

- Digitale Bildanalyse-App der Firma Procedeon Image Science GmbH (https://procedeon.de/de/)
- Pro Parzelle 3-5 Bilder mit Smartphone, 1 m über Bestand, Fläche jeweils 80*80 cm, Mittelwert aus 3-5 Bildern.





Methodenvergleich: Evaluation „*Count my Crop*“ in LSV Winterweizen Erkelenz 2023



10	20	29	23	5	24	15	19	7	12	13	6	4	9	1	2	8	3	28	16	22	17	21	27	30	25	18	26	11	14
4	21	24	11	22	16	6	23	9	10	25	18	26	20	28	1	15	12	17	14	2	13	27	8	7	19	29	3	5	30
7	12	18	17	15	23	16	28	8	13	24	4	6	5	3	10	20	29	22	19	25	27	2	14	9	26	21	30	1	11
20	12	15	19	1	3	9	29	11	4	25	8	27	13	5	2	10	22	21	16	14	6	24	28	7	30	18	17	23	26
3	19	14	22	8	9	15	13	29	5	18	23	11	7	2	26	20	21	1	30	6	16	27	24	12	10	28	25	4	17
19	13	5	24	23	15	20	27	12	10	6	11	28	7	26	14	9	4	30	1	22	21	18	2	16	3	17	8	29	25

3 Wiederholungen · 2 Stufen = 6 Blocks

10 Sorten ausgewählt → 60 Parzellen

Meterzählung
jeweils 1 Meter
4 Personen, 3 Termine
(06.06., 09.06., 16.06.)
5 Bonituren insgesamt

300 Werte Referenz
30 Werte je Sorte
6 Werte je Sorte·Bonitur
5 Werte je Parzelle

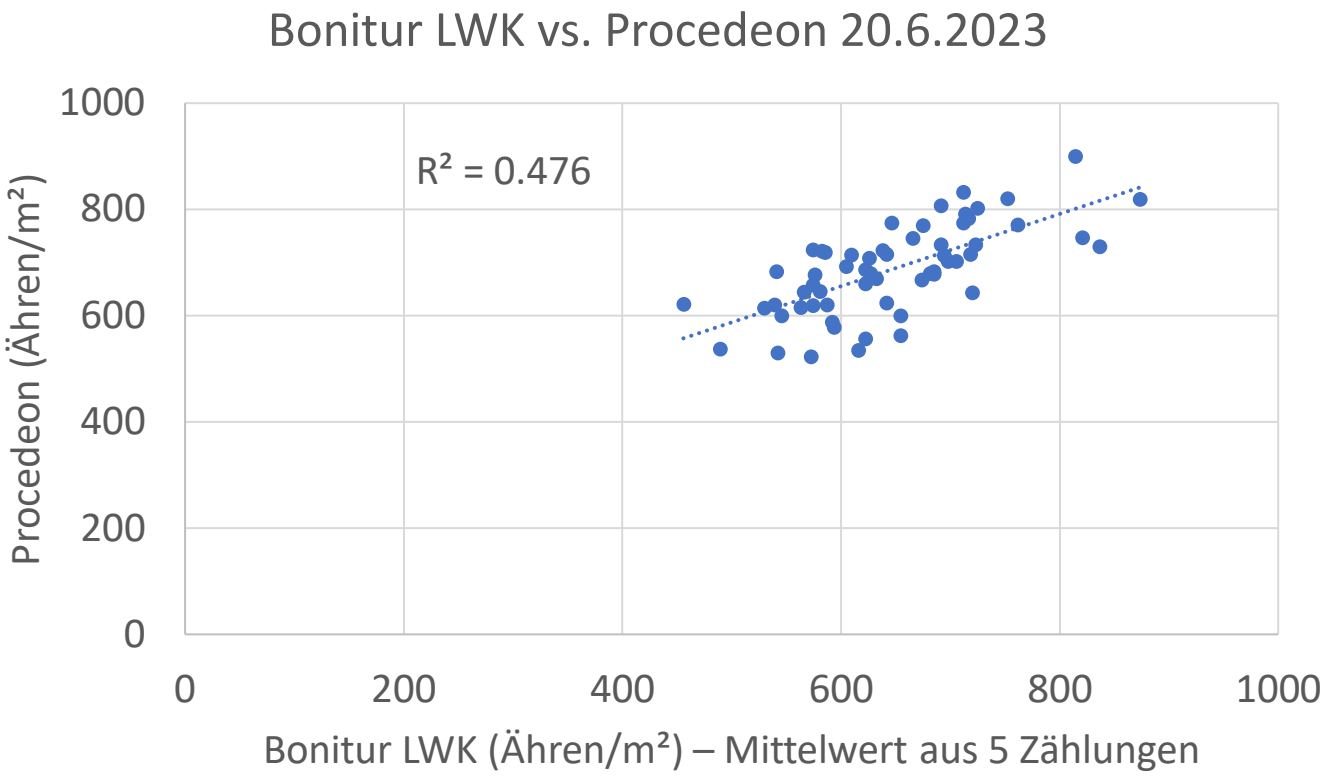
Count my Crop
3 Termine
(20.06., 27.06., 04.07.)
5 Bilder je Parz·Termin

180 Werte *Count my Crop*
18 Werte je Sorte
6 Werte je Sorte·Termin
5 Bilder je Parz·Termin

Empfehlung Procedeon: 5 Bilder je Parzelle aggregieren

Ergebnis Methodenvergleich 2023

- Beide Methoden nach den Standard-Vorgaben durchgeführt
- Referenzmethode jedoch mit 5-fach erhöhter Wiederholungszahl (5 mal 1 Meter anstatt nur 1 mal 1 Meter)
- Auswertung ad hoc: Streudiagramm und Lineare Regression

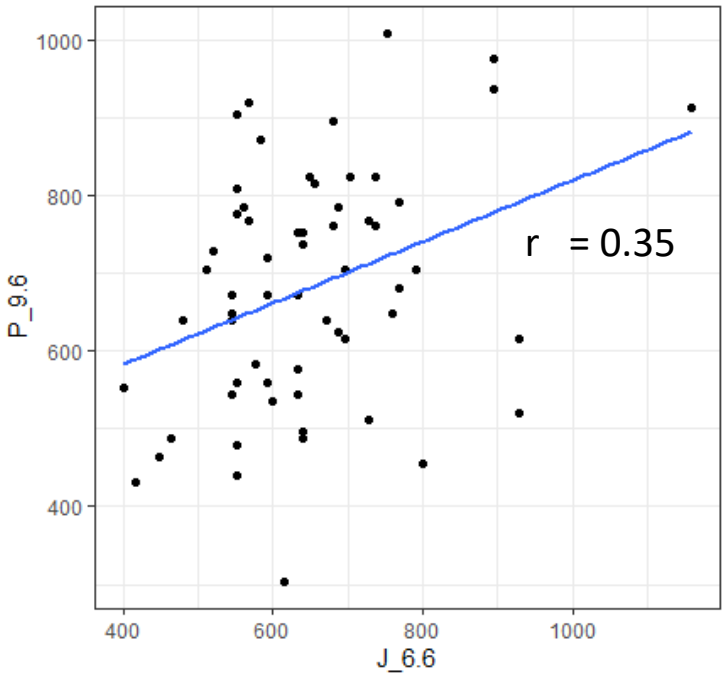


- Korrelation lediglich r = 0.69
- Bestimmtheitsmaß R^2 = 0.48
- Unbefriedigende Methoden-Übereinstimmung

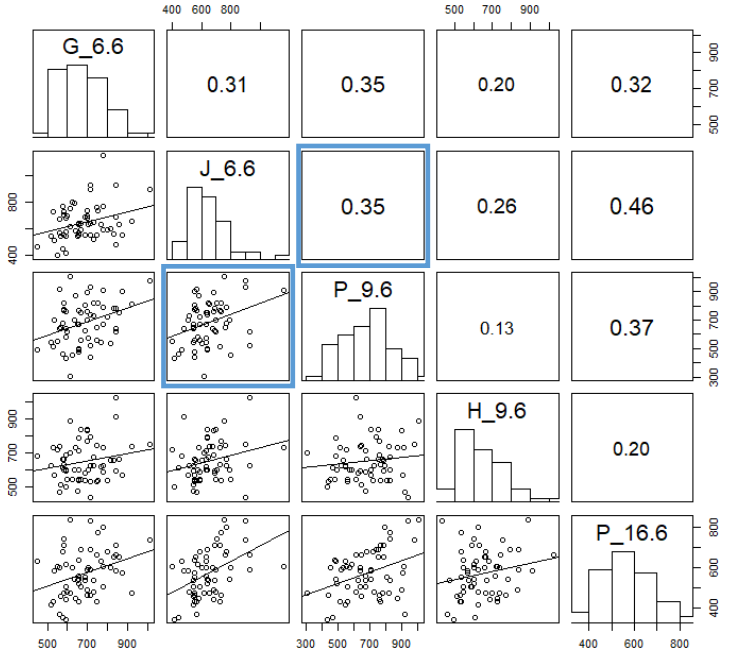
- Warum ist das so schlecht?
- Firma Procedeon hat KI auf vielen verschiedenen Weizensorten trainiert.
- Taugt die App doch nichts?
- Haben die LWK Mitarbeiter falsch gezählt?

Korrelation zwischen Einzel-Bonituren bei „Meterzählung“

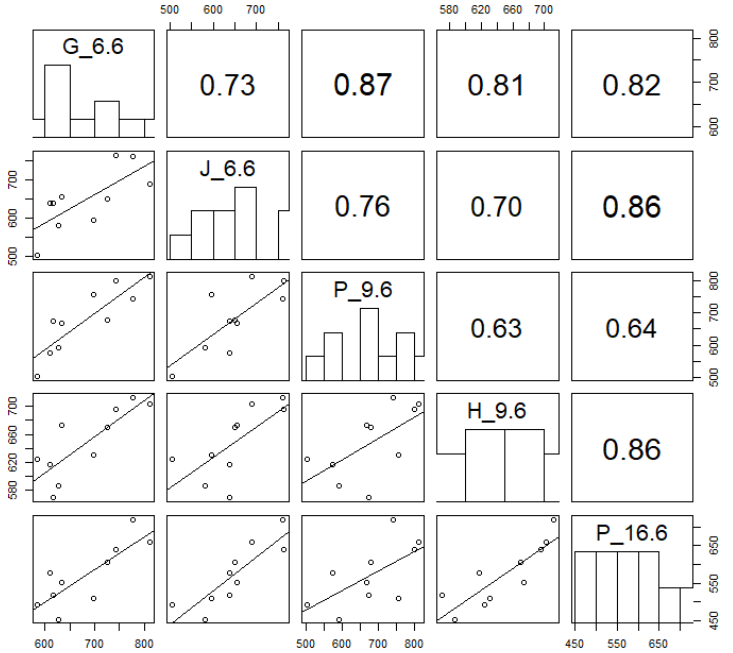
Bonitur J_06.06. vs. P_09.06.
60 Parzellen



Korrelation grundsätzlich niedrig
(unterschiedliche Personen, Termine, Reihen)



10 Sortenmittelwerte
(jeweils 6 Parzellen je Sorte·Bonitur)



Korrelation zwischen Bonituren (und Methoden) ist abhängig von Aggregationsebene

Einzelparzellen Korrelationskoeffizient im Mittel r = 0.30; Sortenmittelwerte Korrelationskoeffizient im Mittel r = 0.77.
Referenzmethode hat geringe Wiederholbarkeit und Reproduzierbarkeit, Methodenvergleich daher schwierig.

Warum ist die Wiederholbarkeit der Meterzählung so gering?

Es wurde an 5 verschiedenen Positionen innerhalb der Parzelle gezählt.

+ das ist **gut**, um einen möglichst guten Schätzwert für die mittlere Ährenzahl einer Parzelle und damit einer Sorte zu erhalten!
- das ist **schlecht**, um die Zuverlässigkeit einer Methode zu quantifizieren.

Getreidebestand = Drillsaat + Bestockung = Zufallsprozess, Ganzzahlige Zähldaten
Merkmal Poisson-verteilt $\boldsymbol{\mu = \sigma^2 = \lambda}$.
Wenn 640 Ähren/qm und Reihenabstand 12.5 cm -> 80 Ährentragende Halme je 100 cm Strecke
Standardabweichung von $\sigma = \sqrt{\mu} = \sqrt{80} = 8.94$
Variationskoeffizient $\sigma/\mu = \sqrt{80}/80 = 11.2\%$.

σ^2Parzelle (bzw. Sorte) ≈ 64 (σ_Parzelle ≈ 8, Parzelle 64 - 96 Ähren/m bzw. 512 - 768 Ähren/qm)
σ^2Poisson ≈ 80

$$\rho = \frac{\sigma^2_{Parzelle}}{\sigma^2_{Parzelle}+\sigma^2_e} = 64/(64+80) = 0.44$$

Selbst in günstigem Szenario mit großen Sortenunterschieden wird Korrelation zwischen Einzelreihen der Referenzmethode meist < 0.5 betragen.
Bodenunterschiede, Drillfehler, Mäusefraß etc. kommen noch dazu!
Was bedeutet das für den Methodenvergleich?

Unterschiedliche Positionen

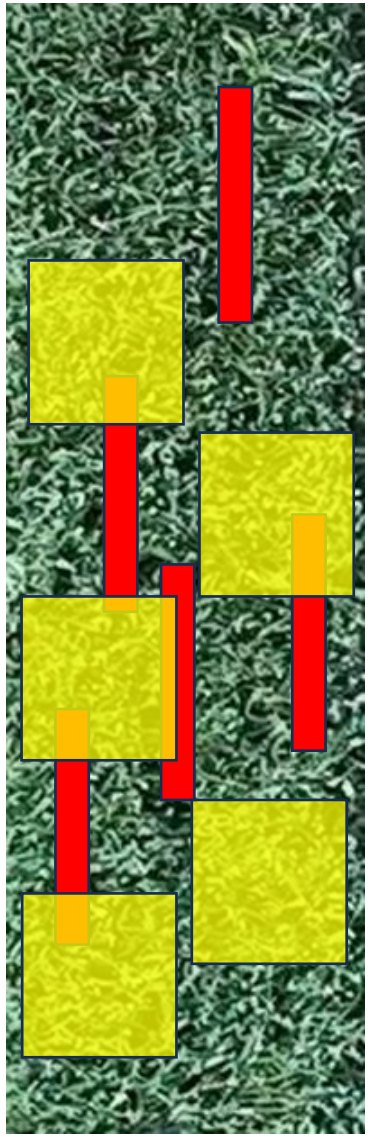
… für CMC-Werte wurde an fünf zufällig ausgewählten Positionen ein Bild gemacht.

- **Unterschiedliche Positionen**

- **Unterschiedliche Form**
  - Meterzählung = lange Rechtecke, CMC = Quadrate

- **Unterschiedliche Fläche**
  - CMC-Bild jeweils 0.64 qm. In Summe 3.2 qm.
  - Meterzählung jeweils 0.125 qm. In Summe 0.625 qm.

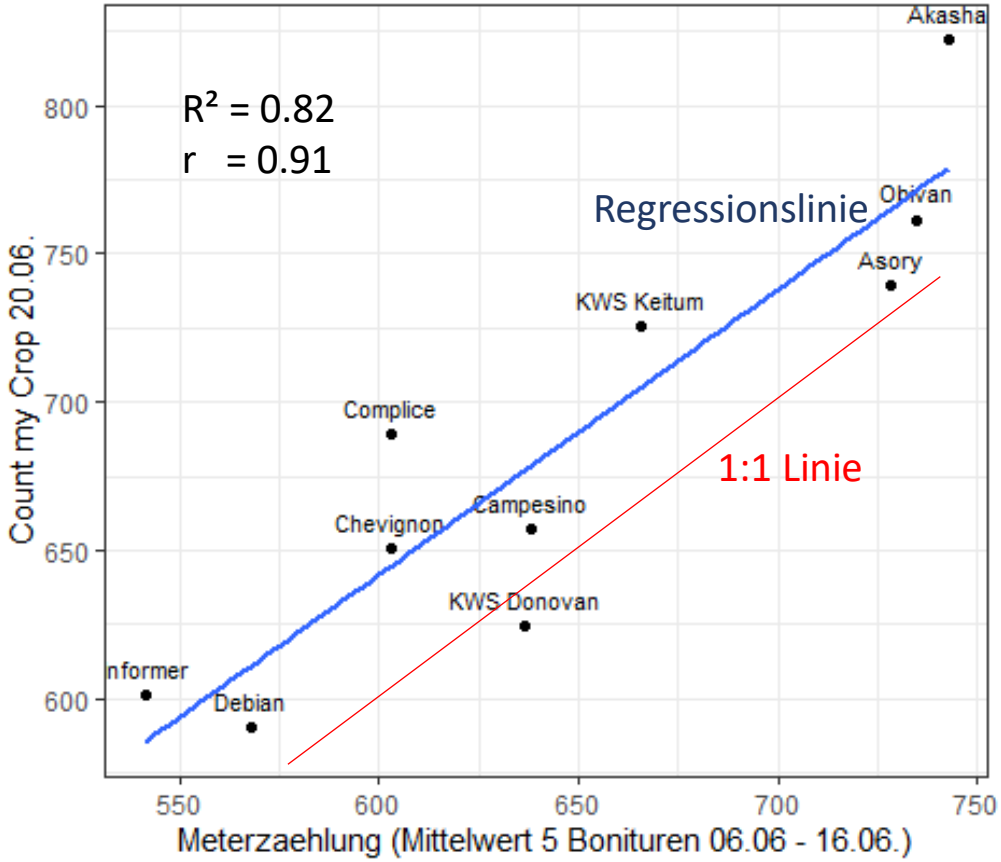
Heterogenität innerhalb einer Parzelle wird zum Problem für den Methodenvergleich



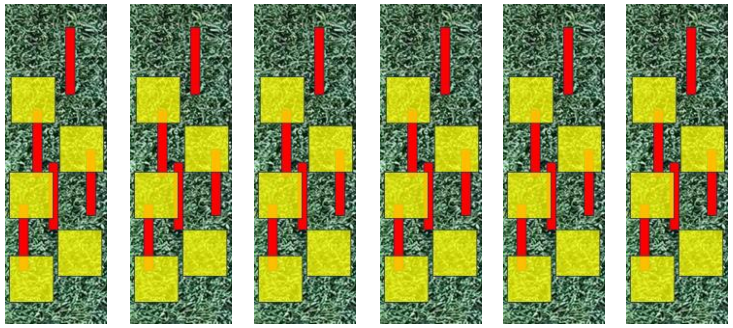
Sortenmittelwerte (aus jeweils 30 Meter, 30 Bilder)

Sorte	Meterzählung (5 Bonituren 06.06.-16.06.)	CMC 20.06.	Delta CMC_20.06. - Meterzählung
Akasha	743	822	79
Asory	729	739	11
Campesino	638	657	19
Chevignon	603	651	48
Complice	603	689	86
Debian	568	590	23
Informer	541	602	60
KWS Donovan	637	625	-12
KWS Keitum	666	725	59
Obivan	735	762	27
Gesamtmittel	646.2	686.3	40.1
StdAbw	71.0	75.1	31.6

Hinreichende Korrelation? Bias?



Varianzkomponenten, Fehlerfortpflanzung und erwartete Korrelation auf Ebene der Sortenmittelwerte



Je Sorte

- 6 Parzellen mit 5 Meterzählungen (à 0.125qm) = 30 Werte
- 6 Parzellen mit Mittelwerten von je 5 Bildern (à 0.64qm) der CMC-App.

Varianzkomponenten Meterzählung

	Varianz	Std.Abw.
Sorte	4563.3	67.55
Bonitur	1793.2	42.35
Bonitur:Sorte	0.0	0.00
Block	0.0	0.00
Block:Sorte	565.1	23.77
Residual	11538.0	107.41

Varianzkomponenten Count My Crop

	Varianz	Std.Abw.
Sorte	3075.7	55.46
Bonitur	980.1	31.31
Bonitur:Sorte	550.8	23.47
Block	93.9	9.69
Block:Sorte	546.2	23.37
Residual	2149.8	46.37

Verhältnis der Residualvarianz entspricht sehr genau dem Flächenverhältnis: 1/5
Sortenvarianz bei CMC etwas geringer (Termineffekt?)

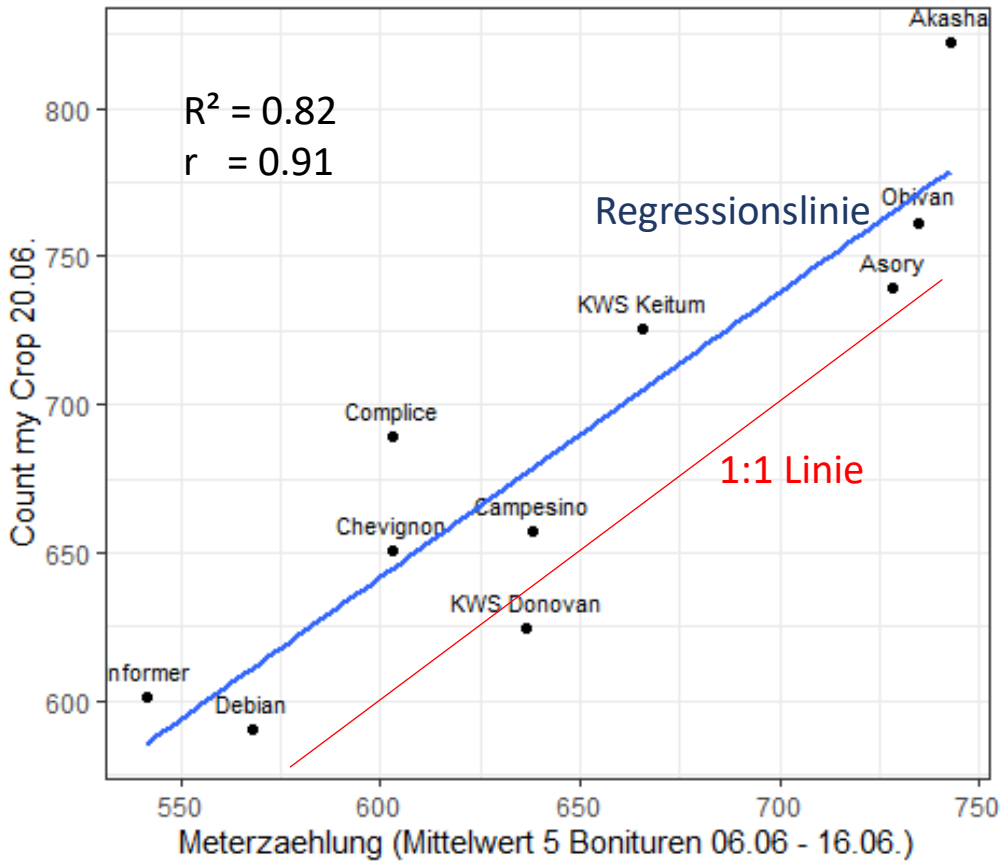
Der Korrelationskoeffizient ergibt sich als Quotient zwischen der Kovarianz zweier Merkmalsreihen und der Wurzel aus dem Produkt der beiden merkmalsspezifischen Varianzen.

$$\rho(x,y) = \frac{Cov(x,y)}{\sqrt{Var(x) \cdot Var(y)}}$$

$$\rho = \frac{4563.3 + \frac{565.1}{6}}{\sqrt{(4563.3 + \frac{565.1}{6} + \frac{11538}{30}) \cdot (4563.3 + \frac{565.1}{6} + \frac{2150}{6})}} = \frac{4657.5}{\sqrt{5042.1 \cdot 5015.8}} = 0.926$$

Unter der Annahme, dass Interaktion Sorte*Methode = 0, beträgt die erwartete Korrelation Meter ~ CMC auf Ebene Sortenmittelwerte r = 0.926

Korrelation auf Ebene der Sortenmittelwerte



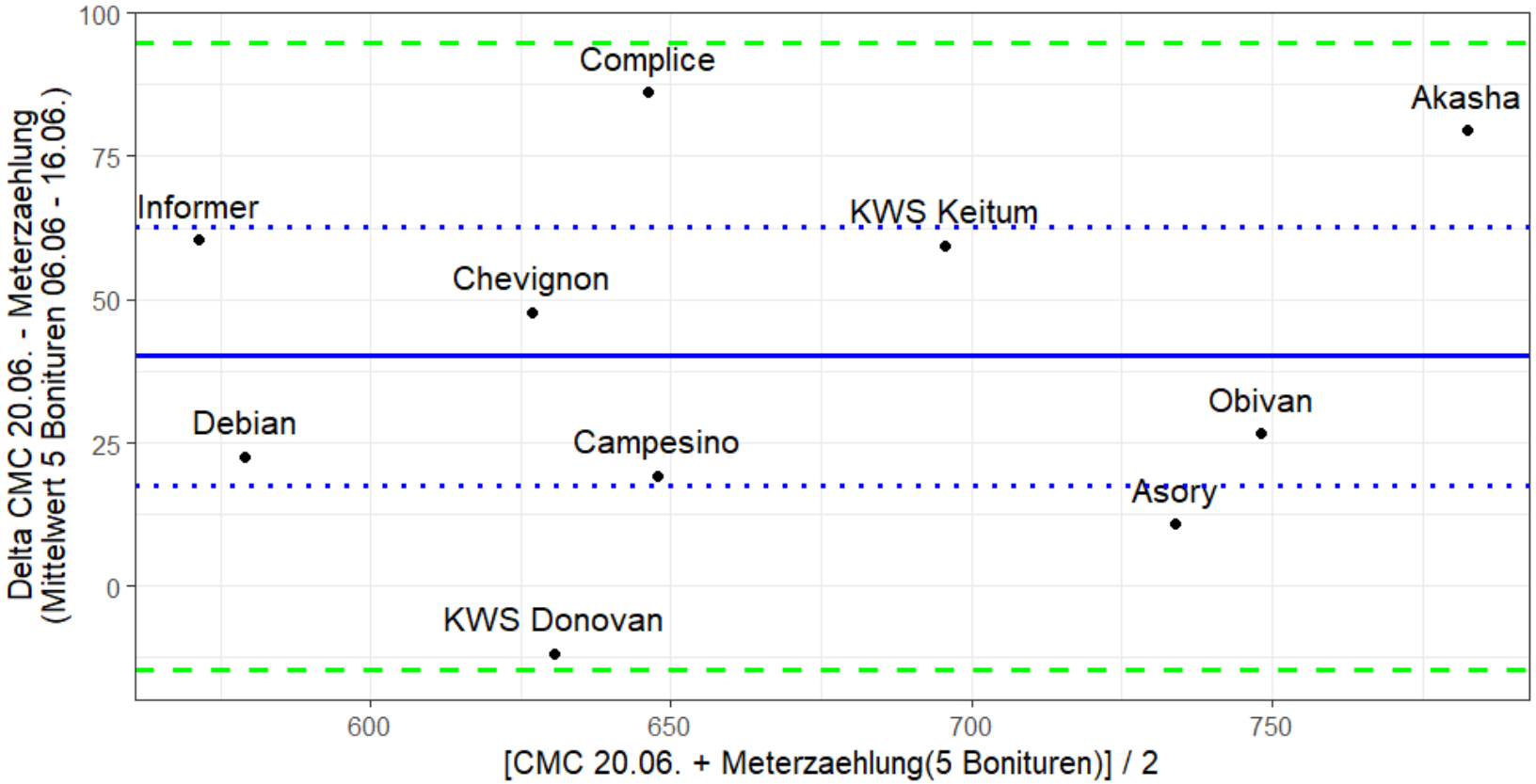
Erwartete Methodenkorrelation Meterzählung ~ CMC im Versuch auf Ebene der Sortenmittelwerte r = 0.926.

Tatsächliche Phänotypische Korrelation im Versuch r = 0.91.

=> Schlussfolgerung: CMC und Meterzählung haben grundsätzlich eine hohe Korrelation.

Bland-Altman-Plot : Methoden-Differenz gegen Mittelwert beider Methoden

(Bland JM, Altman DG 1986: *Comparing methods for assessing agreement between two methods of clinical measurement. Lancet. 307-310, https://www-users.york.ac.uk/~mb55/meas/ba.htm*)



Signifikanter Bias 40 Ähren/qm.

Alle Sorten im 95% Vertrauensintervall. Keine Sorte wird schlechter gemessen als erwartet.

Abweichungen teilweise erheblich, bis zu 86 Ähren (=14%).

Abweichungen nicht signifikant , aber Vertrauensintervall sehr breit.

Blaue Linie = Mittlere Differenz, blaue gepunktete Linie = 95%-Konfidenzintervall für die mittlere Differenz; grüne Linien = 95%-Vorhersageintervall für einzelnen Sorten. Winterweizen, Erkelenz 2023

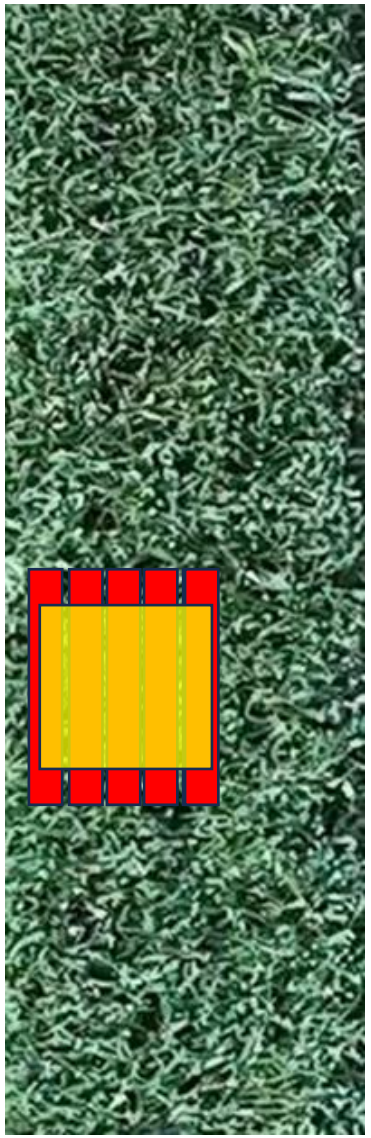
Versuche zur Validierung Count My Crop in 2024 – Veränderte Methodik

… verschiedene Dienststellen (NRW, Bayern, Thüringen, Hessen, Nds, SH, Pflanzenschutzfirmen …)

Empfohlenes Vorgehen

- Meterzählung und CMC-Bild an gleicher Position
- Position festlegen. CMC-Bild machen (am besten mehrfach).
- Danach die ausgewählte Fläche (80 x 80 cm) manuell zählen. (Die manuelle Messung sollte nach den CMC-Messungen durchgeführt werden, weil die manuelle Zählung die räumliche Ährenverteilung (Handhabung der Halme, Sortierung…) und damit die Ährenanzahl verändern kann.)
- Alle Ähren zählen (auch unterständige)

80 x 80 cm Zählrahmen



Effekt des veränderten Experimentellen Designs

*- **schlecht**, um repräsentative Stichprobe einer Parzelle zu erhalten!*

*+ **gut** um die Zuverlässigkeit der Methode zu quantifizieren.*

Wenn beides, Methodenvalidierung und Sortenvergleich im Fokus stehen, dann sind an weiteren Positionen in der Parzelle CMC-Bilder zu machen und/oder Meterzählungen durchzuführen. Die 80 x 80 cm Fläche mit Bild und Meterzählung kann dann zur Adjustierung für einen Gesamtmittelwert genutzt werden.

Take Home Messages

Wenn eine neue Methode gegen eine Referenzmethode verglichen werden soll,

- so müssen die **möglichen Varianzursachen** bekannt sein. Oftmals ist die **Referenzmethode fehlerhaft, und KEIN GOLDSTANDARD**. Es sind Maßnahmen zu treffen, um Referenzwerte hoher Qualität zu haben (das gilt auch für das Training eines KI-Algorithmus). Z.B. viele Wiederholungen machen!

- sollte berechnet werden welche Methoden-Korrelation aufgrund **Fehlerfortpflanzung** möglich ist.

- sollten **nach Möglichkeit die gleichen Pflanzen, die gleichen Positionen, das gleiche Material** gemessen werden. Bei zerstörenden Prüfungen müssen geeignete Sub-Samples hergestellt werden (Fehlerfortpflanzung auch hier beachten!)

- Streudiagramm (x-y-Plot) allein ist nicht aussagekräftig. **Bland-Altman-Plot** ist besser.
    - Sind die ermittelten Methoden-Abweichungen noch im Rahmen?
    - Gibt es einen signifikanten Bias?
    - Ist der Bias linear?

- **Versuchsplanung** ist wichtig
    - Was ist das Ziel der Messung? Gute Werte für Sorte? Oder Methodenvalidierung?
    - Welche Präzision ist erforderlich? Welche Kapazität steht zur Verfügung?
    - Allokation von zu prüfendem Material (Anzahl Sorten, Orte, … ?) und Anzahl Wiederholungen.