Korrelationen

Seit mehreren Jahren arbeiten wir mit **Sentinel-1 SAR** Daten, die unabhängig sind von Bewölkung und Tageslicht und regelmäßig Daten in gleichbleibender Qualität liefern. Diese Daten zeigen die räumliche und zeitliche Variabilität der Biomasseentwicklung. Sentinel-1 liefert Messwerte, die eine Kombination aus den beiden Faktoren Struktur der Oberfläche und deren Feuchte abbilden.
In Bezug auf landwirtschaftliche Kulturen entspricht das am ehesten der Entwicklung von **frischer Biomasse**.
Es ist möglich einige bio-physikalische Parameter wie **Ertrag** oder den **Blattoberflächen Index** (LAI) von diesen Daten abzuleiten.
In den vergangenen Jahren haben wir unsere Kartenprodukte dahingehend kalibriert, dass die Werte mit den entsprechenden bio-physikalischen Parametern von landwirtschaftlichen Flächen gut übereinstimmen, bzw. korrelieren.

Für die Kalibrierung wurden Referenzdaten herangezogen. Zum einen punktuelle Daten aus Feldbegehungen und zum anderen Flächendaten von optischen Satellitendaten oder Ertragskarten, die von Mähdrescherdaten abgeleitet wurden, und zum Teil auch von Bodenkarten.
All diese Daten haben im Vergleich untereinander eine unterschiedliche **räumlich-zeitliche Auflösung**.

Wenn man **Feldversuche** macht, dann hat man **Stichprobenpunkte**, die man regelmäßig besucht.
Die Untersuchungsfläche eines Stichprobenpunktes umfasst ca. 1 m². Die Daten von Sentinel-1 haben eine räumliche Auflösung von 20x20 m, also 400 m²/Pixel. Vergleicht man also Satellitendaten mit Daten von Stichprobenpunkten stellt sich die Frage, ob der Stichprobenpunkt tatsächlich repräsentativ für die Fläche des entsprechenden Pixels ist. Möglicherweise kommen hier und da auch noch Ungenauigkeiten des GPS Empfängers bei der Felddatenerhebung zum tragen, da nicht immer sicher ist, ob man im richtigen Pixel ist.

Mähdrescherdaten sind Punktdatensätze, die zunächst statistisch bereinigt werden müssen, um daraus eine repräsentative Ertragskarte zu erstellen. Unplausible Werte in **Teilbereichen**, die vom Mähdrescher mehrmals oder nur zum Teil überfahren wurden, müssen herausgefiltert werden so gut es geht. Anschließend werden die Punktdaten mit einem statistischen Model interpoliert und es wird daraus eine Rasterkarte erstellt, die die verschiedenen Ertragszonen ausweist. Wir verwenden letztlich also interpolierte bzw. modellierte Daten.

Die **Kornfeuchtemessung** beim Drusch wirkt sich direkt auf die Ertragsmessung aus, da der Ertrag in Bezug auf die Feuchte normalisiert wird. Wird also eine zu hohe Kornfeuchte gemessen, veringert sich der Ertrag, weil der Wassergehalt herausgerechnet wird. Wenn man sich vorstellt welche Mengen an Pflanzenmaterial, teilweise mit Unkraut versetzt, in einem Mähdrescher verarbeitet werden und wie intensiv diese teuren Geräte genutzt werden, versteht man, dass es da zu Abweichungen kommen kann. Auch die **Feuchte-Kalibrierung** wird nicht wie eigentlich gefordert an jedem Feld neu ausgeführt.

Bei **optischen Satellitendaten** haben wir atmosphärische Störungen, die das Messergebnis beeinflussen können. Zudem sind optische Satellitendaten nicht immer verfügbar, so dass die Vergleichswerte nicht immer aktuell sind, was zu Abweichungen führt, da landwirtschaftliche Kulturen ständiger Veränderung unterliegen.

Von den **Satellitendaten von Sentinel-1** wurde ein Punkt-shapefile erstellt. Dabei wurde in der Mitte jedes Pixels ein Punkt erzeugt. Für jeden dieser Punkte wurden die Werte von verschiedenen Rasterkarten ausgelesen. Wir haben also Feld für Feld Korrelationen gebildet und uns angesehen, ob wir in den Feldern ähnliche Muster sehen, bzw. ob die Werte miteinander korrelieren, denn es wäre eigentlich erwartbar, dass die Muster optisch vergleichbar sind, wenn die Korrelationen hoch sind.

Bei den Daten von den Stichprobenpunkten wurde mit den Punktdaten auf den Rasterkarten von Sentinel-1 die entsprechenden Werte ausgelesen, um festzustellen, ob die Korrelationen für diese Punkte hoch sind. Da diese Punkte aber nur in geringer Zahl vorhanden sind, also 9 Punkte auf einer Fläche von bis zu 10 ha, lassen sich dort keine Muster vergleichen, sondern lediglich die punktuellen Korrelationen.

Dabei haben wir festgestellt, dass das nicht immer der Fall ist. Unabhängig davon man Stichprobenpunkte, eine Ertragskarte vom Ende der Saison, oder eine NDVI Karte, die möglichst zeitgleich erfasst wurde, mit einer Biomassekarte von Sentinel-1 vergleicht, kann die Korrelation hoch oder gering sein.

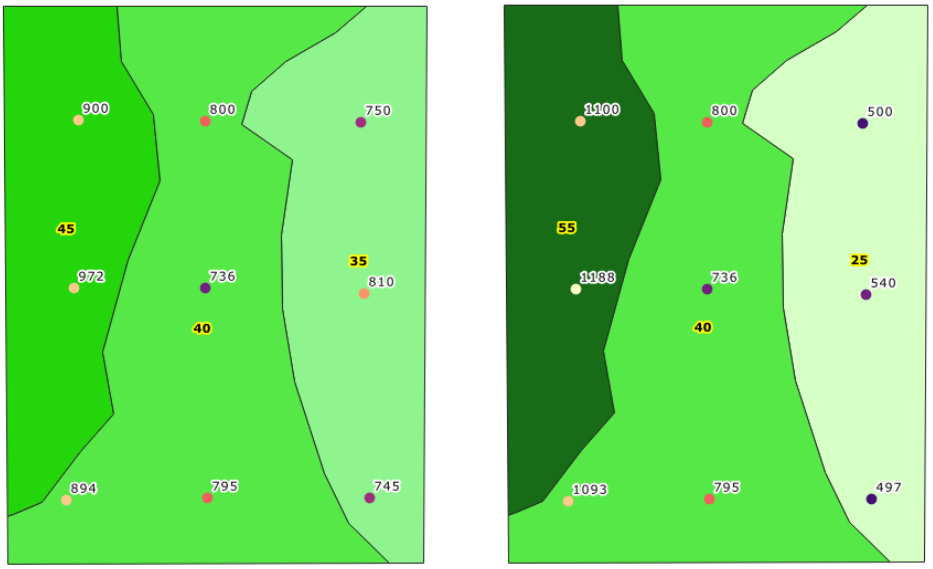
Generell gilt, dass bei einer Korrelation **$R^2 < 0.5$** keine Abhängigkeit zwischen den beiden Parametern besteht. Liegt die Korrelation bei **$R^2 > 0.75$** gibt es eine gute Übereinstimmung zwischen den beiden Parametern und wenn **$R^2 > 0.9$** ist, haben wir bereits eine hohe Übereinstimmung.

Es hat sich allerdings gezeigt, dass unterschiedliche Flächen mit **zeitgleichen** Daten unterschiedlich hohe Korrelationen hervorbringen und das hängt maßgeblich davon ab wie groß die **Wertespreitung** innerhalb einer Fläche ist. Und genau darum geht es in diesem Artikel.

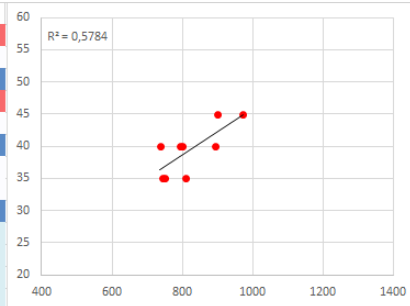
Wenn wir **Datensätze** miteinander vergleichen, die mit **gleicher Methode**, den **gleichen Messsystemen** und zur **gleichen Zeit** erfasst wurden, und wir dann die **gleiche Kultur** von ein und demselben Landwirt auf unterschiedlichen Flächen betrachten, dann erwarten wir eigentlich ähnliche Ergebnisse. Es kann allerdings dennoch passieren, dass man auf einer Fläche eine relativ hohe Korrelation findet, und auf einer anderen Fläche nur eine relativ geringe Korrelation.

Warum ist das so? Bei einer geringeren Wertespreitung innerhalb einer Fläche, wirkt sich das **Messrauschen**, das jede Messung mit sich bringt, überproportional stark auf den Korrelationswert aus.

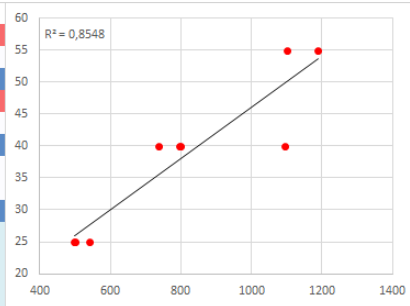
Betrachten wir einmal die beiden folgenden Beispiele. Zwei Flächen mit gleichem Muster in Bezug auf die Ausprägung der Werte, jedoch mit unterschiedlicher Wertespreitung. Nehmen wir an die Flächendaten stammen von einer **ESVI** Biomassekarte mit einem Wertebereich von 0-100. Die Werte der Messpunkte stammen von einer Feldbegehung bei der die oberirdischen Pflanzenteile abgeschnitten und gewogen wurden, um die frische Biomasse / m² zu ermitteln. Auf der linken Fläche haben wir Werte von 35 – 45 von der **ESVI Biomassekarte** und auf der rechten Fläche haben wir einen Wertebereich von 25 – 55.



ID	STP	ESVI
1	900	45
2	800	40
3	750	35
4	972	45
5	736	40
6	810	35
7	894	40
8	795	40
9	745	35
Min	736	35
Ave	832	40
Max	972	45
Range	236	10



ID	STP	ESVI
1	1100	55
2	800	40
3	500	25
4	1188	55
5	736	40
6	540	25
7	1093	40
8	795	40
9	497	25
Min	497	25
Ave	844	40
Max	1188	55
Range	691	30



Das Muster ist bei beiden Flächen gleich. Lediglich die Wertespreitung ist auf der rechten Seite höher.
Bei den Messpunkten variiert der Wertebereich für die **Pflanzenproben** zwischen 736-972 auf der linken Fläche und auf der rechten Fläche von 497-1188. Die Abstufung der Werte innerhalb der Klassen ist vergleichbar.
Dennoch sind die Korrelationswerte links bei **$R^2 < 0,58$** und rechts **R^2 0,85**.

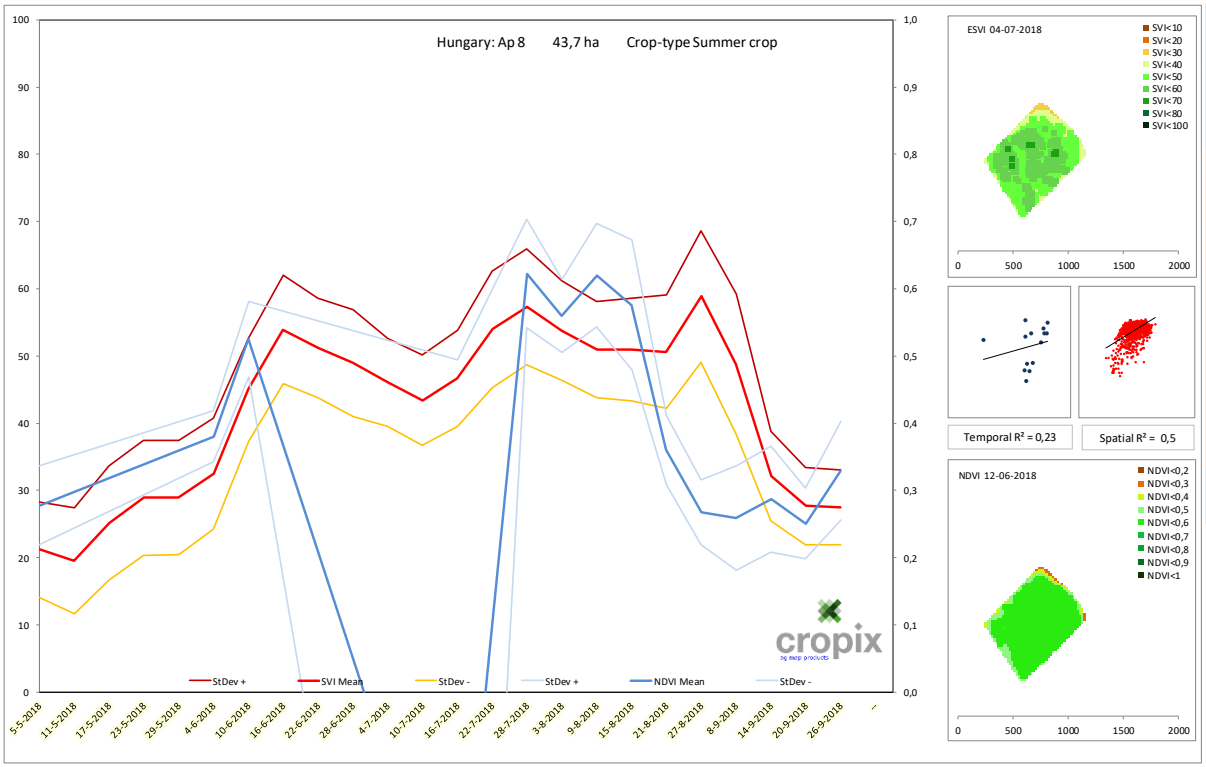
Das ist ein deutlicher Unterschied. Rechts finden wir eine gute Korrelation, links nur eine geringe Korrelation.

Das heißt, dass die Korrelation nicht nur davon abhängt wie gut zwei vergleichbare Datensätze in ihrer **Werteverteilung** übereinstimmen, sondern vor allem wie groß die **Wertespreitung** ist.
Das heißt wiederum, dass man nicht zwingend sagen kann, dass es keine Abhängigkeit zwischen zwei Datensätzen gibt, wenn die Korrelationswerte gering sind. Es ist eher davon abhängig wie **homogen** oder **heterogen** einzelne Flächen in Bezug auf ihre Wertigkeit sind, was sich dann in der Wertespreitung niederschlägt.

Man kann sogar sagen, dass auf einer homogenen Fläche mit einer geringen Wertespreitung sich überhaupt keine Korrelation zwischen zwei unabhängigen Parametern finden lässt. Es liegt in der Natur der Sache, dass wenn das Messrauschen, das bei jeder Messung vorhanden ist, größer ist als die Wertespreitung es keine brauchbare Korrelation geben kann.

Und dennoch ist die Messung korrekt und die Qualität der Daten ebenso gut, wie das Ergebnis auf einer heterogenen Fläche mit großer Wertespreitung und guten Korrelationswerten. Insofern ist die ausschließliche Betrachtung der Korrelationswerte als Meßlatte für die Vergleichbarkeit zweier Datensätze ungeeignet.

Die folgende Graphik zeigt eine Fläche in Ungarn und deren zeitliche Signatur für die ESVI Werte von Sentinel-1 Daten in rot und die Signatur für den NDVI (Sentinel-2) in blau. Dabei sehen wir jeweils den Durchschnittswert und die einfache Standardabweichung.
Auf der rechten Seite die räumliche Ausprägung eines Aufnahmedatums und die dazugehörige Korrelation.



Man sieht zum einen den Abfall der NDVI Werte weil es teilweise Bewölkung gab, was sich auch in der zeitlichen Korrelation der Mittelwerte deutlich zeigt. Zum anderen sieht man dass die NDVI Werte bereits im Juni sehr homogen erscheinen. Demgegenüber zeigt der ESVI Anfang Juli zonale Unterschiede.
Wir haben hier also im Vergleich einen eher heterogenen Datensatz (ESVI), der eine gewisse räumliche Variabilität aufweist und einen recht homogenen Datensatz von dem NDVI.
Die räumliche Korrelationsanalyse liefert uns in diesem Fall kein eindeutiges Ergebnis und keiner der beiden Datensätze lässt sich daher mit dieser Methode validieren.