Correlations

For several years, we have been working with **Sentinel-1 SAR** data, which is independent of cloud cover and daylight and provides consistent data quality. This data shows the spatial and temporal variability of biomass development. Sentinel-1 delivers measurements that represent a combination of two factors: surface structure and moisture.
In relation to agricultural crops, this most closely corresponds to the development of **fresh biomass**. It is possible to derive several biophysical parameters, such as **yield** or the **leaf area index** (LAI), from this data.
In recent years, we have calibrated our map products so that the values closely match, or correlate with, the corresponding biophysical parameters of agricultural land.

Reference data were used for the calibration. These included point data from field surveys, area data from optical satellite data, yield maps derived from combine harvester data, and, in some cases, soil maps.
All these data sets have different **spatial-temporal** resolutions when compared to one another.

When conducting **field trials**, sample points are visited regularly.
The area studied by a **sample point** is approximately 1 m². The Sentinel-1 data has a spatial resolution of 20 x 20 m, or 400 m²/pixel. Therefore, when comparing satellite data with sample point data, the question arises whether the sample point is truly representative of the area of the corresponding pixel. Inaccuracies in the GPS receiver during field data collection may also play a role, as it is not always certain whether the correct pixel is being measured.

Combine harvester data consists of point datasets that must first be statistically cleaned to create a representative yield map. Implausible values in areas that were driven over multiple times or only partially by the combine harvester must be filtered out as effectively as possible. The point data is then interpolated using a statistical model, and a raster map is created that shows the different yield zones. Ultimately, we are using interpolated or modeled data.

The **grain moisture measurement** during harvesting directly impacts the yield measurement, as the yield is normalized in relation to the moisture content. Therefore, if the measured grain moisture is too high, the yield is reduced because the water content is factored out. Considering the quantities of plant material, sometimes mixed with weeds, processed in a combine harvester and the intensive use of these expensive machines, it's understandable that discrepancies can occur. Furthermore, **moisture calibration** is not performed anew in every field, as actually required.

With **optical satellite data**, we have atmospheric disturbances that can influence the measurement results. In addition, optical satellite data is not always available, so the comparative values are not always up-to-date, leading to discrepancies because agricultural crops are subject to constant change.

A point shapefile was created from the **Sentinel-1 satellite data**. A point was generated at the center of each pixel. For each of these points, the values were read from various raster maps. We then calculated correlations field by field and examined whether we observed similar patterns or whether the values correlated with each other, as it would be expected that the patterns would be visually comparable if the correlations were high.

For the data from the sample points, the corresponding values were read from the Sentinel-1 raster maps to determine whether the correlations for these points were high. However, since these points are only present in small numbers—9 points in an area of up to 10 hectares—it is not possible to compare patterns, but only the point-specific correlations.

We found that this is not always the case. Regardless of whether one compares sample points, a yield map from the end of the season, or an NDVI map recorded as concurrently as possible with a biomass map from Sentinel-1, the correlation can be high or low.

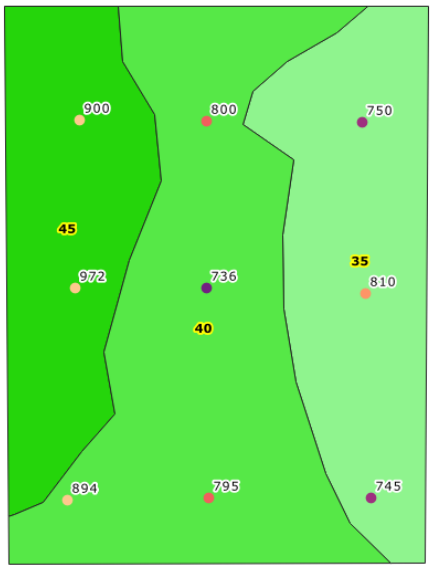
Generally, a correlation $\mathbf{R^2 < 0.5}$ indicates no relationship between the two parameters. A correlation $\mathbf{R^2 > 0.75}$ indicates good agreement between the two parameters, and an $\mathbf{R^2 > 0.9}$ indicates high agreement.

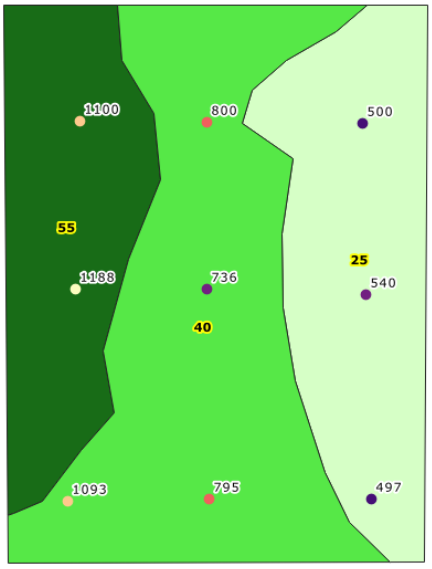
However, it has been shown that different areas with **concurrent data** exhibit varying levels of correlation, and this depends significantly on the **variance of values** within a given area. This is precisely the focus of this article.

When comparing **data sets** collected using the **same method**, the **same measurement systems**, and at the **same time**, and then examining the **same crop** from the same farmer on different plots, we would expect similar results. However, it can still happen that a relatively high correlation is found on one plot and only a relatively low correlation on another.

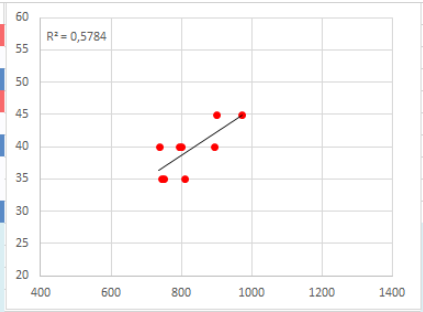
Why is this? With a smaller spread of values within a plot, the **measurement noise** inherent in every measurement has a disproportionately strong effect on the correlation value.

Let's consider the following two examples. Two plots with the same pattern in terms of the values, but with different spreads. Let's assume the plot data comes from an **ESVI** biomass map with a value range of 0-100. The values of the measurement points come from a field survey in which the above-ground parts of the plants were cut and weighed to determine the fresh biomass per square meter. On the left side, we have values from 35 to 45 from the **ESVI biomass map**, and on the right side, we have a range of values from 25 to 55.

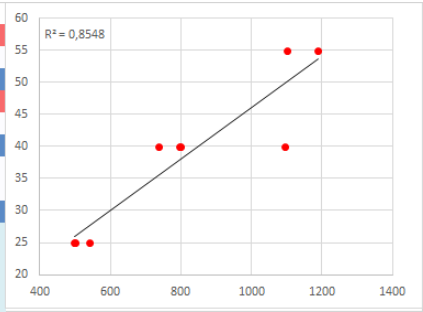




ID	STP	ESVI
1	900	45
2	800	40
3	750	35
4	972	45
5	736	40
6	810	35
7	894	40
8	795	40
9	745	35
Min	736	35
Ave	832	40
Max	972	45
Range	236	10



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



The pattern is the same for both sides. Only the range of values is higher on the right.
At the measurement points, the range of values for the **plant samples** varies between 736 and 972 on the left side and between 497 and 1188 on the right. The gradation of values within the classes is comparable. Nevertheless, the correlation values are **$R^2 < 0.58$** on the left and **R^2 0.85** on the right.

This is a significant difference. On the right, we find a good correlation, while on the left, the correlation is weak.

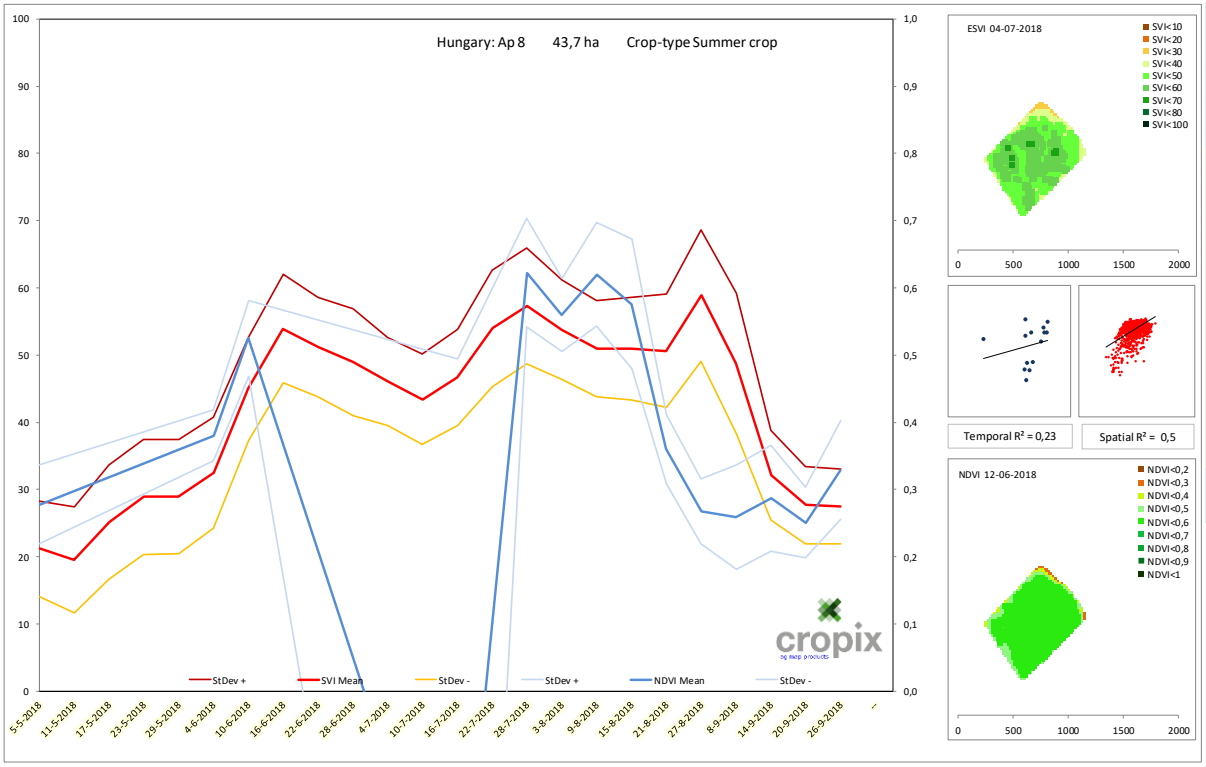
This means that the correlation depends not only on how well two comparable datasets match in their **value distribution**, but above all on the size of the **value spread**.

This, in turn, means that one cannot necessarily conclude that there is no correlation between two datasets simply because the correlation values are low. It depends more on how **homogeneous** or **heterogeneous** individual areas are with respect to their value, which is then reflected in the value spread.

One can even say that on a homogeneous area with a low value spread, no correlation at all can be found between two independent parameters. It is inherent in the nature of things that if the measurement noise, which is present in every measurement, is greater than the value spread, no usable correlation can exist.

And yet, the measurement is correct, and the quality of the data is just as good as the result on a heterogeneous area with a large value spread and good correlation values. Therefore, considering only the correlation values as a benchmark for the comparability of two datasets is unsuitable.

The following graph shows an area in Hungary and its temporal signature for the ESVI values of Sentinel-1 data in red and the signature for the NDVI (Sentinel-2) in blue. The mean value and the simple standard deviation are shown in each case.
On the right is the spatial distribution of a given recording date and its corresponding correlation.



Firstly, the decrease in NDVI values due to partial cloud cover is evident, which is also clearly reflected in the temporal correlation of the mean values. Secondly, the NDVI values appear very homogeneous as early as June. In contrast, the ESVI shows zonal differences at the beginning of July.
Therefore, we have a rather heterogeneous dataset (ESVI) exhibiting some spatial variability and a fairly homogeneous NDVI dataset.
In this case, the spatial correlation analysis does not provide a clear result, and neither dataset can be validated using this method.