

Correlaciones

Durante varios años, hemos trabajado con **datos SAR de Sentinel-1**, que son independientes de la nubosidad y la luz diurna y proporcionan una calidad de datos consistente. Estos datos muestran la variabilidad espacial y temporal del desarrollo de la biomasa. Sentinel-1 proporciona mediciones que representan una combinación de dos factores: la estructura de la superficie y la humedad.

En relación con los cultivos agrícolas, esto se corresponde con mayor precisión con el desarrollo de la **biomasa fresca**. A partir de estos datos, es posible derivar varios parámetros biofísicos, como el **rendimiento** o el **índice de superficie foliar (LAI)**.

En los últimos años, hemos calibrado nuestros productos cartográficos para que los valores coincidan estrechamente o se correlacionen con los parámetros biofísicos correspondientes de las tierras agrícolas.

Para la calibración se utilizaron datos de referencia. Estos incluyeron datos puntuales de estudios de campo, datos de área de datos satelitales ópticos, mapas de rendimiento derivados de datos de cosechadoras y, en algunos casos, mapas de suelos.

Todos estos conjuntos de datos tienen diferentes **resoluciones espaciales y temporales** al compararse entre sí.

Durante los **ensayos de campo**, se visitan regularmente los **puntos de muestreo**. El área estudiada por un punto de muestreo es de aproximadamente 1 m². Los datos de Sentinel-1 tienen una resolución espacial de 20 x 20 m, o 400 m²/píxel. Por lo tanto, al comparar los datos satelitales con los datos de puntos de muestreo, surge la pregunta de si el punto de muestreo es realmente representativo del área del píxel correspondiente. Las imprecisiones del receptor GPS durante la recopilación de datos de campo también pueden influir, ya que no siempre es seguro que se esté midiendo el píxel correcto.

Los **datos de la cosechadora** se componen de conjuntos de datos puntuales que primero deben depurarse estadísticamente para crear un mapa de rendimiento representativo. Los valores inverosímiles en **áreas sobre** las que la cosechadora ha pasado varias veces o solo parcialmente deben filtrarse de la forma más eficaz posible. Los datos puntuales se interpolan mediante un modelo estadístico y se crea un mapa ráster que muestra las diferentes zonas de rendimiento. En definitiva, utilizamos datos interpolados o modelados.

La medición de la **humedad del grano** durante la cosecha influye directamente en la medición del rendimiento, ya que este se normaliza en función del contenido de humedad. Por lo tanto, si la humedad del grano medida es demasiado alta, el rendimiento se reduce debido a que se descarta el contenido de agua. Considerando la cantidad de material vegetal, a veces mezclado con malezas, que se procesa en una cosechadora y el uso intensivo de estas costosas máquinas, es comprensible que puedan surgir discrepancias. Además, la **calibración de la humedad** no se realiza de nuevo en cada campo, como realmente se requiere.

Con los **datos satelitales ópticos**, existen perturbaciones atmosféricas que pueden influir en los resultados de las mediciones. Además, los datos satelitales ópticos no siempre están disponibles, por lo que los valores comparativos no siempre están actualizados, lo que genera discrepancias debido a que los cultivos agrícolas están sujetos a cambios constantes.

Se creó un shapefile de puntos a partir de los **datos del satélite Sentinel-1**. Se generó un punto en el centro de cada píxel. Para cada uno de estos puntos, se extrajeron los valores de varios mapas ráster. A continuación, calculamos las correlaciones campo por campo y examinamos si observamos patrones similares o si los valores se correlacionaban entre sí, ya que se esperaba que los patrones fueran visualmente comparables si las correlaciones fueran altas.

Para los datos de los puntos de muestra, se extrajeron los valores correspondientes de los mapas ráster de Sentinel-1 para determinar si las correlaciones entre estos puntos eran altas. Sin embargo, dado que estos puntos solo están presentes en pequeñas cantidades (9 puntos en un área de hasta 10 hectáreas), no es posible comparar patrones, sino solo las correlaciones específicas de cada punto.

Observamos que esto no siempre es así. Independientemente de si se comparan puntos de muestra, un mapa de rendimiento del final de la temporada o un mapa del NDVI registrado de la forma más concurrente posible con un mapa de biomasa de Sentinel-1, la correlación puede ser alta o baja.

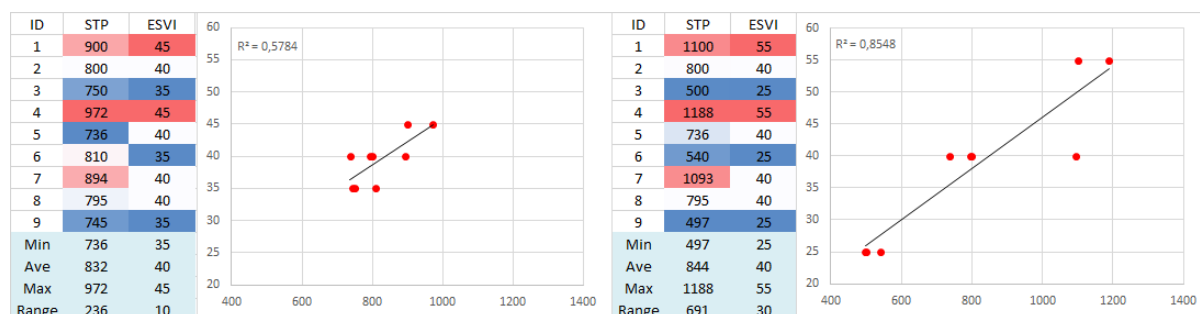
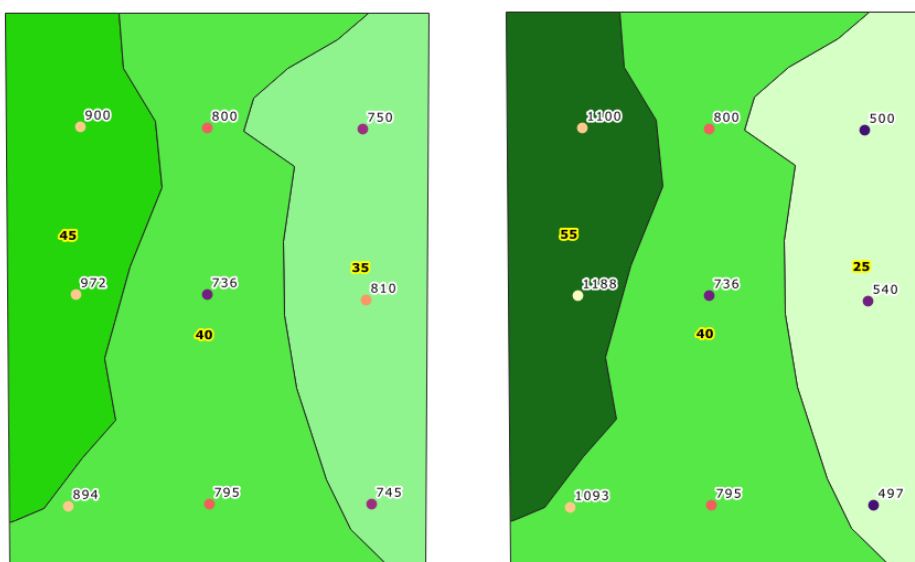
Generalmente, una correlación $R^2 < 0,5$ indica que no hay relación entre los dos parámetros. Una correlación $R^2 > 0,75$ indica una buena concordancia entre ambos parámetros, y una $R^2 > 0,9$ indica una alta concordancia.

Sin embargo, se ha demostrado que diferentes áreas con datos **concurrentes** presentan distintos niveles de correlación, lo cual depende significativamente de la **varianza de los valores** dentro de un área determinada. Este es precisamente el enfoque de este artículo.

Al comparar conjuntos de **datos** recopilados con el **mismo método**, los **mismos sistemas de medición** y al **mismo tiempo**, y luego examinar el **mismo cultivo** del mismo agricultor en diferentes parcelas, cabría esperar resultados similares. Sin embargo, aún puede ocurrir que se encuentre una correlación relativamente alta en una parcela y solo una correlación relativamente baja en otra.

¿Por qué? Con una menor dispersión de valores dentro de una parcela, el **ruido de medición** inherente a cada medición tiene un efecto desproporcionadamente fuerte en el valor de correlación.

Consideremos los dos ejemplos siguientes: dos parcelas con el mismo patrón en cuanto a valores, pero con diferentes dispersiones. Supongamos que los datos de la parcela provienen de un mapa de biomasa **ESVI** con un rango de valores de 0 a 100. Los valores de los puntos de medición provienen de un estudio de campo en el que se cortaron y pesaron las partes aéreas de las plantas para determinar la biomasa fresca por metro cuadrado. En el lado izquierdo, se muestran valores de 35 a 45 del mapa de **biomasa del ESVI**, y en el lado derecho, un rango de valores de 25 a 55.



El patrón es el mismo en ambos lados. Solo el rango de valores es mayor en el lado derecho. En los puntos de medición, el rango de valores de las **muestras de plantas** varía entre 736 y 972 en el lado izquierdo y entre 497 y 1188 en el derecho. La gradación de valores dentro de las clases es comparable. Sin embargo, los valores de correlación son $R^2 < 0,58$ en el lado izquierdo y $R^2 0,85$ en el derecho.

Esta es una diferencia significativa. En el lado derecho, se observa una buena correlación, mientras que en el izquierdo, la correlación es débil.

Esto significa que la correlación depende no solo de la coincidencia de dos conjuntos de datos comparables en su **distribución de valores**, sino sobre todo de la amplitud de la **dispersión de valores**.

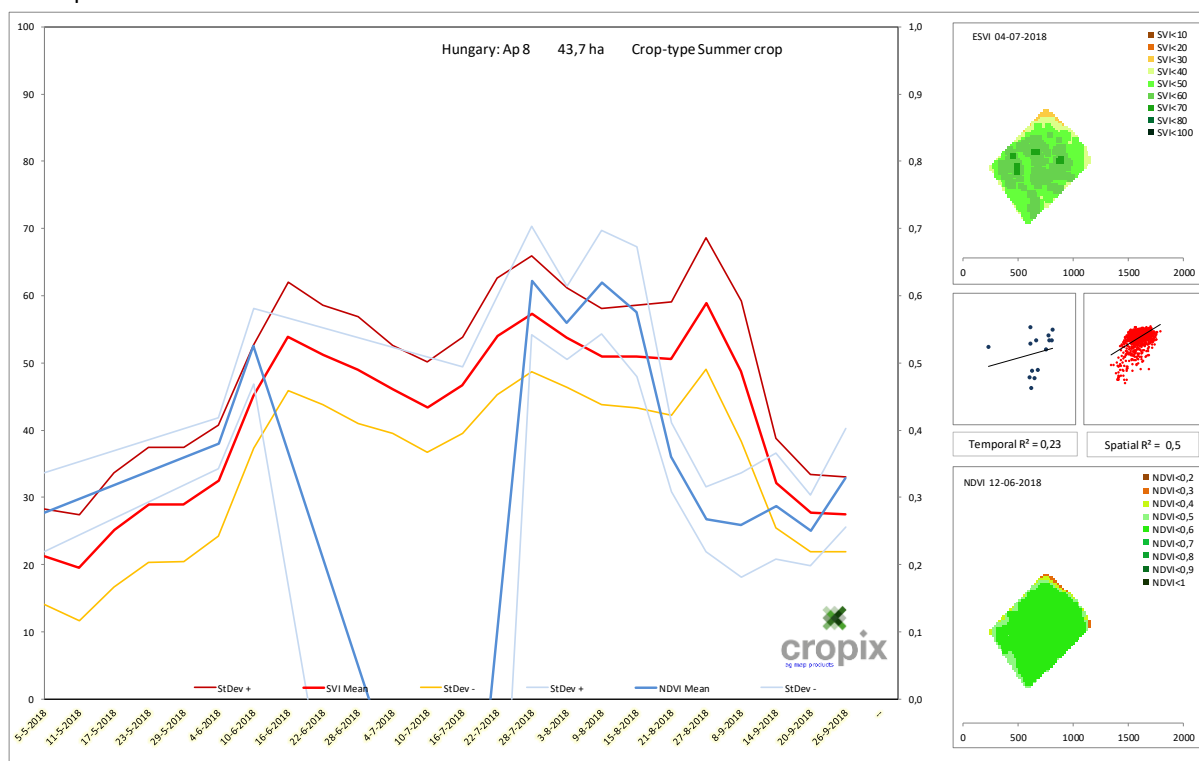
Esto, a su vez, significa que no se puede concluir necesariamente que no existe correlación entre dos conjuntos de datos simplemente porque los valores de correlación sean bajos. Depende más de la **homogeneidad o heterogeneidad** de las áreas individuales con respecto a su valor, lo que se refleja en la dispersión de valores.

Incluso se puede afirmar que en un área homogénea con una dispersión de valores baja, no se puede encontrar correlación alguna entre dos parámetros independientes. Es inherente a la naturaleza de las cosas que, si el ruido de medición, presente en cada medición, es mayor que la dispersión de valores, no puede existir una correlación utilizable.

Sin embargo, la medición es correcta y la calidad de los datos es tan buena como la del resultado en un área heterogénea con una gran dispersión de valores y buenos valores de correlación. Por lo tanto, considerar únicamente los valores de correlación como punto de referencia para la comparabilidad de dos conjuntos de datos no es adecuado.

El siguiente gráfico muestra un área en Hungría y su firma temporal para los valores ESVI de los datos de Sentinel-1 en rojo y la firma para el NDVI (Sentinel-2) en azul. Se muestran el valor medio y la desviación estándar simple en cada caso.

A la derecha se muestra la distribución espacial de una fecha de registro determinada y su correlación correspondiente.



En primer lugar, es evidente la disminución de los valores del NDVI debido a la nubosidad parcial, lo que también se refleja claramente en la correlación temporal de los valores medios. En segundo lugar, los valores del NDVI parecen muy homogéneos ya en junio. En cambio, el ESVI muestra diferencias zonales a principios de julio.

Por lo tanto, tenemos un conjunto de datos bastante heterogéneo (ESVI) que presenta cierta variabilidad espacial y un conjunto de datos del NDVI bastante homogéneo.

En este caso, el análisis de correlación espacial no proporciona un resultado claro, y ninguno de los conjuntos de datos puede validarse con este método.