Prueba técnica para cargo de Data Analyst en Spaceag

Matías Felipe Rebolledo González

Índice general

Índice de cuadros

Índice de gráficos

1. Enunciado.

Encontrar la relación entre el NDVI obtenido de imágenes satelitales del servidor sentinelhub y el rendimiento promedio obtenido en una temporada para una hacienda de paltos en la zona de Ica, Perú. Idealmente buscar una relación causal mediante un modelo que explique y pronostique el comportamiento del rendimiento promedio de paltos a partir del NDVI. En el gráfico 1 presentamos la disposición de las parcelas dentro del predio del estudio.

2. Preprocesamiento.

Se configuró la descarga de 204 imágenes desde el servidor sentielhub definiendo una fecha inicial y una final. Las fechas usadas tomaron la hora específica definida en el excel inicial. Se utilizó un notebook jupyter tal como se sugirió en el protocolo de la prueba.

Posteriormente de las 204 imágenes fueron seleccionadas 89 que correspondian a la fecha y hora que se estipuló en el mismo excel que contenía información sobre las fechas de interés para el análisis.

Una vez descargadas se diseñó un proceso usando herramientas Unix y librería gdal para preparar las imágenes antes de analizarlas en software. Entre los pasos más importantes, se truncaron las imágenes para valores inferiores a 0.2 del NDVI. Es decir, se convirtieron las celdas con valores inferiores a ese umbral, debajo del cual se representa suelo, a valores NoData.

Esto trajo sí como consecuencia que al menos 10 días del periodo, no contenían información para valores inferiores a ese umbral y, una vez que se separaron las fechas en sus respectivas parcelas, esas 10 imágenes se reflejaron en 262 imágenes por día y por parcela sin información. Debido a este motivo, el proceso se hizo varias veces probando umbrales un poco menores para poder construir una grilla completa sin datos faltantes para cada parcela y cada día y posteriormente dar una menor ponderación a valores que estén por debajo del umbral de interés. Este proceso se repitió hasta un umbral de NDVI de -0.01 en el cual no hubo ninguna imagen parcela/día sin valores faltantes, lo que permitió calcularle a todas las 2937 imágenes las estadísticas descriptivas.

Una métrica adicional que se obtuvo para cada imagen y poder modelar con ella fue la mediana, la que se obtuvo exportando a csv todas las imágenes y calculado este indicador para cada una de ellas. Muchos pasos de preprocesamiento se ejecutaron en paralelo para disminuir el tiempo de ejecución.

Posteriormente para efectos descriptivos decidimos agrupar las 33 parcelas en 5 zonas y poder presentar resultados más resumidos en este informe. Algo similar se procedió a hacer con días agrupándolos en meses para analizar la estacionalidad del indicador. Primero mediante un geoproceso de dissolve agrupamos las 33 parcelas en 5 zonas

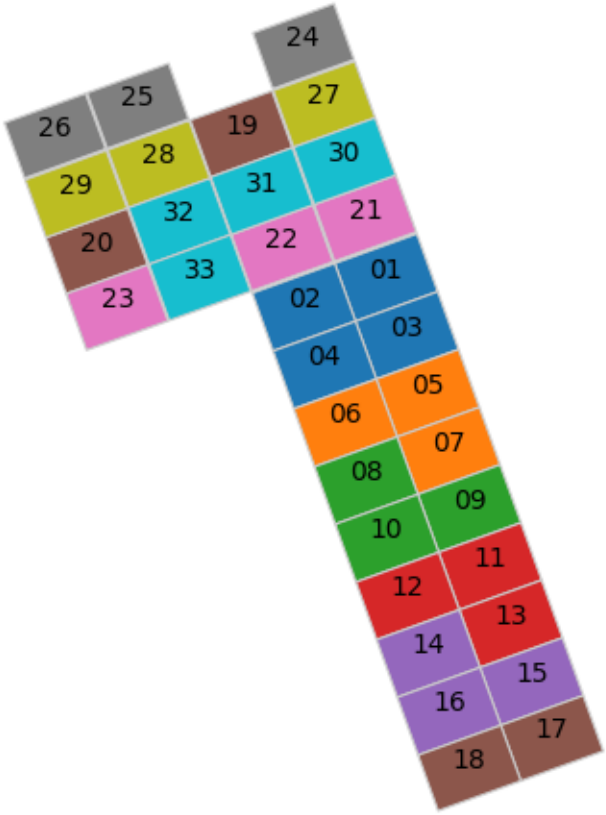

Gráfico 1: Parcelas de la zona de estudio.

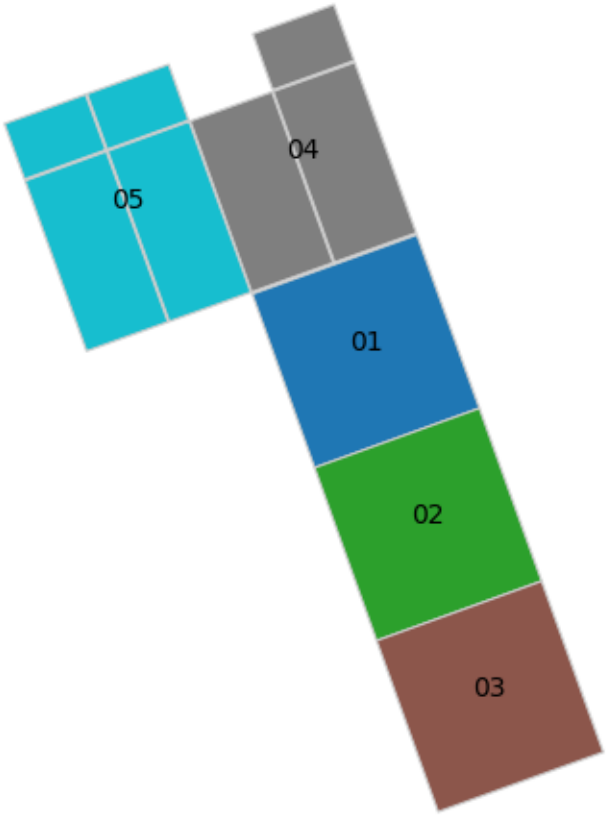

Gráfico 2: Parcelas agrupadas en zonas mediante dissolve.

creando identificadores y asignando un nuevo para cada zona (ver gráfico 2).

3. Análisis descriptivo.

3.1. Análisis de corte transversal.

Posteriormente extrajimos de las 2937 imágenes las estadísticas descriptivas incluyendo la mediana con lo cual hicimos un archivo tabular de 2937 x 8 con el cual hicimos 8 archivos de 89 x 33. El nivel inicial de NDVI escogido para truncar las imágenes fue de -0.01 el cual nos permitirá visualizar el porcentaje de información que cae por

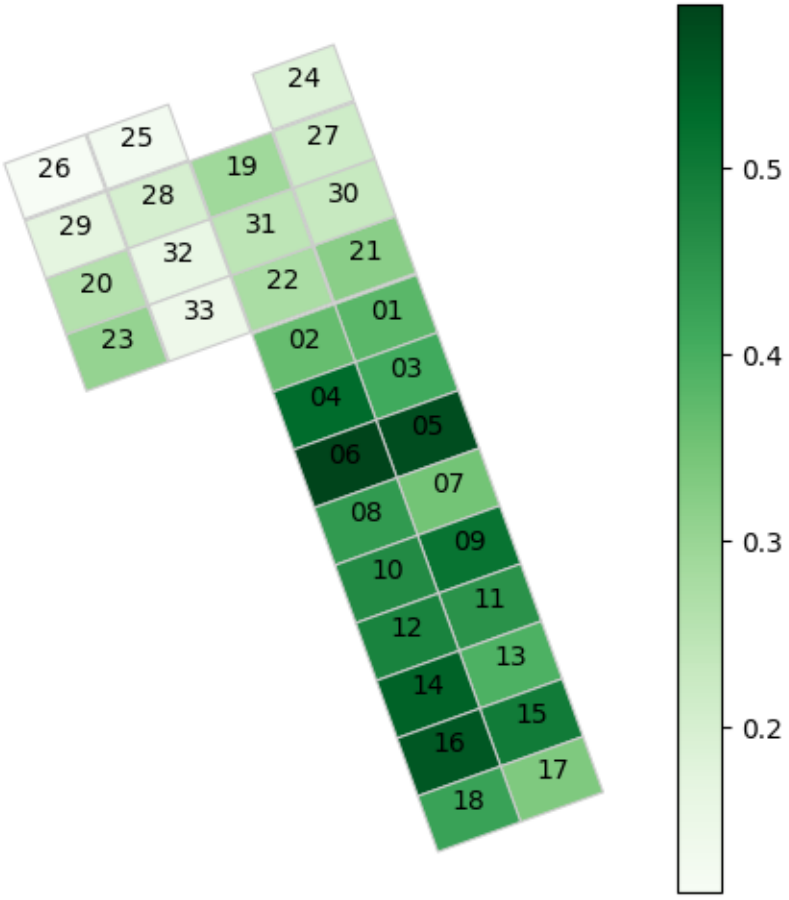

Gráfico 3: NDVI-mediana agregado en el tiempo y por parcelas.

debajo de ese umbral en cada parcela y con esto identificar las parcelas que tienen un bajo nivel de NDVI.

Poteriormente agregamos las métricas de todas las imágenes para poder visualizarlas más fácilmente. Elegimos dos métricas para visualizar, a saber el valor NDVI-mediana y NDVI-porcentaje-corte. El primer indicador es la mediana del NDVI por imagen el que a su vez fue agregado, primero mensualmente y luego anualmente, usando la misma mediana para poder presentarlo por parcela. El segundo indicador corresponde al porcentaje de celdas que sí tienen información sobre el umbral NDVI de -0.01 para tener una idea de cuáles parcelas tienen una distribución con mayores valores de este indicador, lo que pueda representar una mayor cobertura foliar.

En el gráfico 3 bservamos que la distribución del NDVI-mediana resulta mayor en las primeras parcelas de la muestra, mientras que disminuye considerablemente hacia las últimas parcelas. La dispersión de este indicador a

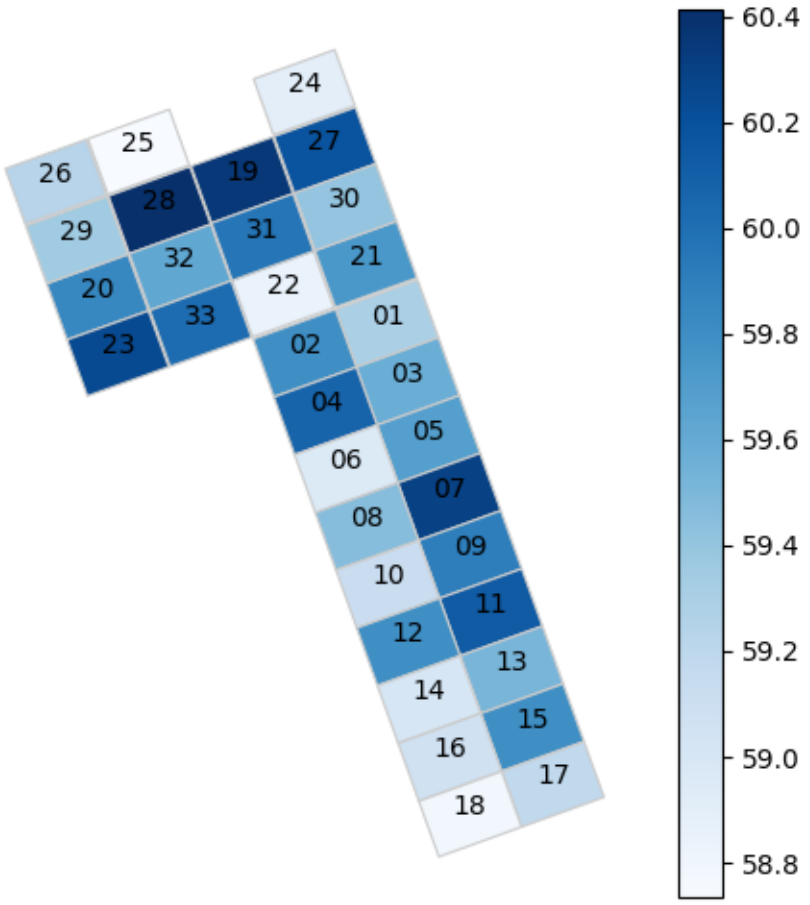

Gráfico 4: Porcentaje-corte-NDVI mayores a -0.01 agregado en el tiempo y por parcelas.

lo largo de cada parcela es considerable partiendo de casi 0.6 como valor máximo en las primeras parcelas hasta un valor cercano a 0.1 en las últimas.

En el gráfico 4 vemos que la dispersión de celdas por sobre el umbral de -0.01 es menor que en el caso de NDVI-mediana. También fue obtenido calculando valores de la mediana para este porcentaje a nivel de meses, luego de años y finalmente en todo el periodo a nivel de parcelas. A un mayor valor de este porcentaje, mayor es la cantidad de celdas que están por sobre el umbral. Es interesante ver que hay parcelas donde el porcentaje es alto, pero la mediana es baja, tal como, la parcela 28. Es implica que el NDVI no estaría relacionado con la cobertura foliar en sí, es decir, que la calidad del NDVI no refleja directamente cuánto abarca el follaje de este huerto, sino con su intensidad dentro del mismo follaje independiente del tamaño.

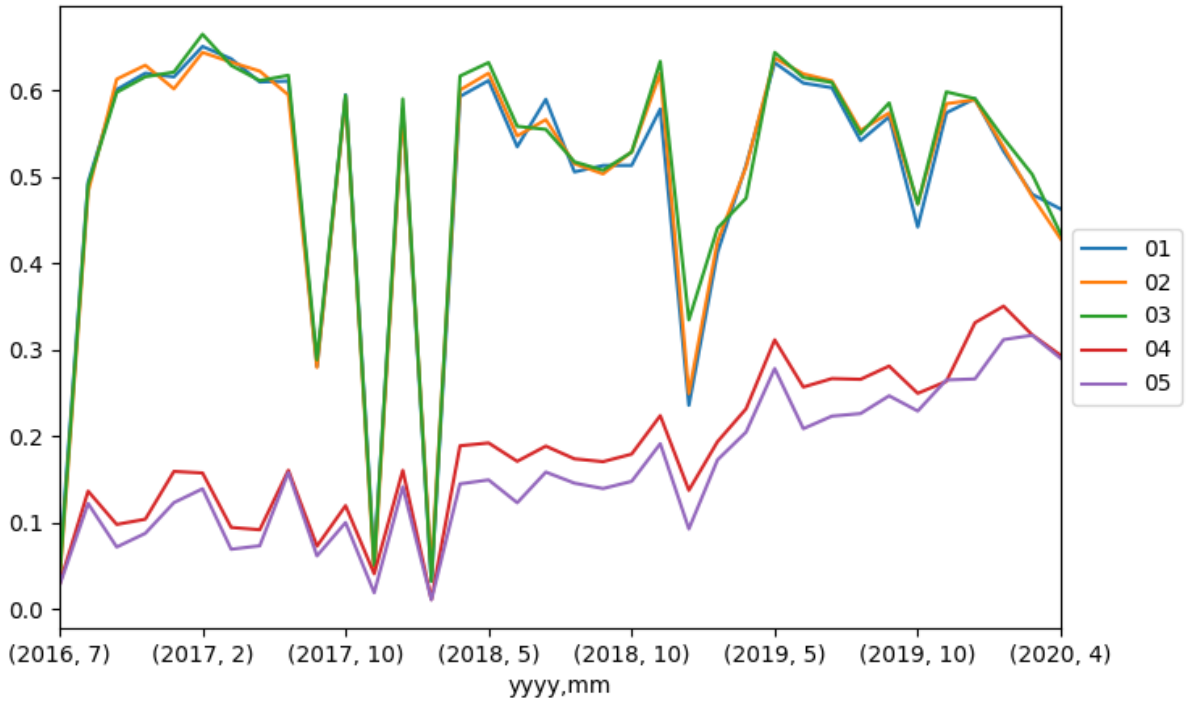

Gráfico 5: Serie temporal mensual NDVI-mediana según zona.

3.2. Análisis temporal.

El primero paso del análisis temporal fue observar la serie NDVI-mediana agregada a nivel mensual (ver gráfico 5). Observamos que las series de las zonas 1, 2 y 3, son las que presentan un mayor nivel del indicador. Además observamos un comportamiento estacional que tiende a sugerir que en temporada de verano, existe una disminución del NDVI, lo que pudiera deberse a aspectos climáticos y fisiológicos de los árboles frutales. En el verano entre 2017 y 2018 ser observó la mayor caída en el indicador. Los árboles frutales requieren de una cantidad de horas de frio para lograr inducir las yemas, lo que se adscribe a un fuerte aspecto de fisiología de la producción. En un contexto de cambio climático, este fenómeno puede estar fuermente influenciado por el cambio de corrientes en zonas de verano. Pero estamos señalando estos aspectos como conjeturas. Otra posibilidad que explique la fuerte diferencia entre las zonas 1, 2 y 3 versus las 4 y 5, es que puedan haber dos tipos de suelo distintos entre esas dos áreas, puesto que el huerto cuenta solamente con una varidad en todas las parcelas.

Posteriormente decidimos separar cada año en 2 temporadas producto de la observación de las series que construimos a nivel de las 5 zonas que definimos previamente. La primera temporada va desde Abril hasta Septiembre y la segunda va desde Octubre hasta Marzo para comparar estas dos temporadas a nivel de las 5

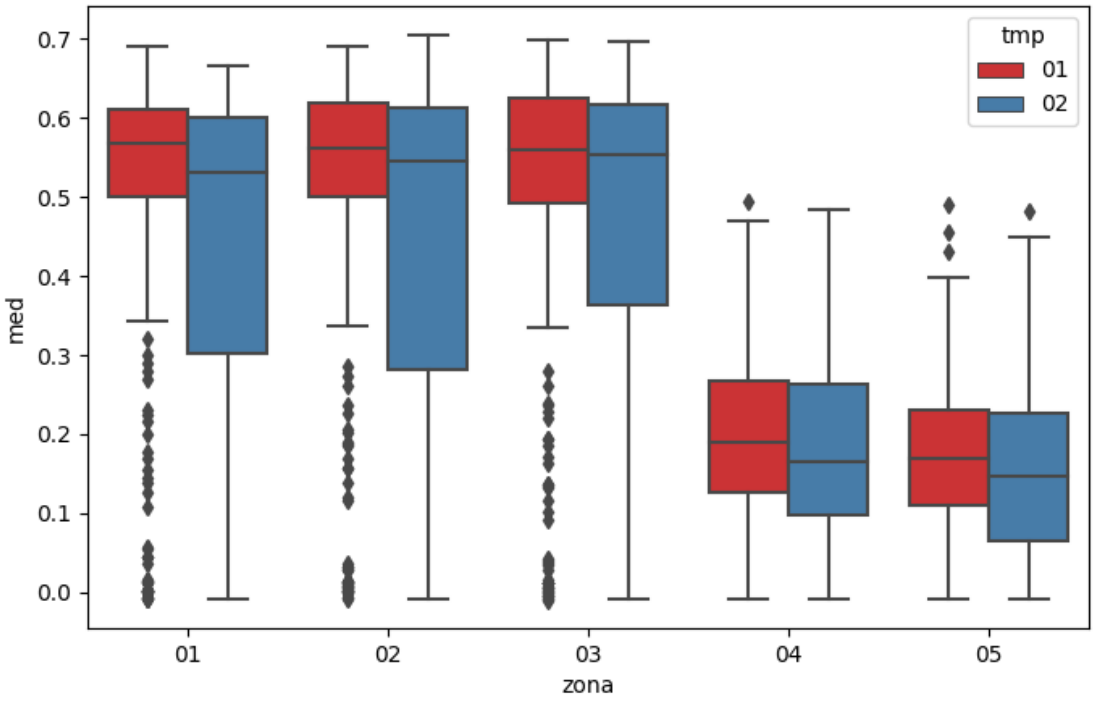

Gráfico 6: Gráfico de cajas NDVI-mediana según temporada y zona.

zonas definidas.

En el gráfico 6 se abrevia la temporada como “tmp” y se separa la dispersión del NDVI entre las 5 zonas del predio. Se observa que la temporada 2 que va desde Octubre a Marzo presenta una mayor dispersión, lo que puede estar en función de los fenómenos de inestabilidad climática, principalmente, el Niño y la Niña, que no necesariamente coinciden con la longitud de 1 año ni de una temporada sino que pueden ser más largos. Esta mayor variabilidad se presenta justamente en las zonas donde la productividad medida usando el NDVI es mayor, lo que podría traducirse en una inestabilidad fisiológica de los frutales.