Escuela de Ingeniería
Ingeniería Civil en Computación

Distribución de horas frío acumuladas dentro de campos de cerezos usando procesos Gaussianos

Duvan García
Profesor Guía: Cristóbal Quiñinao

Memoria para optar al título de Ingeniero Civil en Computación

Rancagua, Chile
Enero 2023

Resumen

La plantación de árboles de cerezos para la producción y exportación de su fruto se ha convertido en una importante parte de la agricultura en la región de O'Higgins, dejando así en segundo plano a otros tipos de cultivos. Este fenómeno se explica en parte debido al feroz crecimiento de la demanda por este fruto rojo que ha estado ocurriendo en el mercado asiático durante los últimos años.

Generalmente los datos reportados por los sistemas de monitoreo existentes corresponden a métricas que en la mayoría de los casos por sí solas no tienen ninguna utilidad para los agricultores, así, la toma de decisiones sobre el proceso productivo dentro del campo se basa principalmente en la experiencia que se ha ganado en años anteriores y no en la predicción o análisis de algún modelo matemático-probabilístico basado en los datos recolectados. Una métrica muy importante para los encargados de los campos es la cantidad de horas frío acumuladas por los árboles, esta información les llega comúnmente de datos recolectados por estaciones meteorológicas cercanas, lo cual tiene múltiples desventajas.

Para entregar información valiosa con respecto a cómo se distribuyen las horas frío acumuladas dentro de los campos se propone el uso de procesos Gaussianos los cuales corresponden a una técnica de Machine Learning especialmente diseñada para resolver problemas de regresión, en palabras simples esto se traduce en la posibilidad de predecir horas frío acumuladas en lugares donde no existen sensores. Los datos usados para entrenar al modelo provienen de sensores dispuestos en posiciones fijas dentro de los campos.

La técnica de procesos Gaussianos muestra un buen despeño en la predicción de temperatura y más importante aún en la predicción de horas frío acumuladas para distintas locaciones en el campo, además de alertar de zonas con alta varianza/incertidumbre las cuales serían ideales para la colocación de nuevos sensores. Como resultado de este método es posible darle información significativa a los agricultores, como por ejemplo qué porcentaje de árboles han alcanzado las horas frío requeridas, esto para distintos umbrales de horas y en distintos momentos del proceso productivo.

Índice

1. Introducción

1.1. Antecedentes generales

Como es bien sabido los cerezos lideran el crecimiento del sector frutícola en la región de O'Higgins, este hecho se puede corroborar dando un vistazo al Catastro Frutícola del Centro de Información de Recursos Naturales [Esp21] cuya última publicación para nuestra región fue en septiembre del 2021. Según este informe las hectáreas con cultivos de cerezos pasaron de 13699 en 2018 a 22966 en 2021 lo que se traduce en un crecimiento de 67.7% en 3 años. Dentro de la región destaca la provincia del Cachapoal, la cual tiene la mayor cantidad de superficie en explotación (tomando en cuenta todo tipo de cultivos), un 61% del total regional, y a su vez dentro de la cual el cerezo es el cultivo líder con más de 8 mil hectáreas.

De las más de 100 mil toneladas de cerezas que se producen anualmente en los campos de la región apenas un 9.3% tiene como destino el mercado Chileno mientras que un 90.3% tiene como destino la exportación a otros mercados. Este fruto representa la segunda exportación más importante hacia China, el principal socio comercial de nuestro país. La ventas de este preciado producto al gigante asiático generan más de 1300 millones de dólares anualmente por lo cual la tecnificación y en general el uso de la tecnología para la profesionalización del trabajo se vuelve cada vez más necesario para mantener un producto competitivo a nivel internacional.

Hablando de variedades en la región se cuenta con cuatro variedades principales de cerezos en los campos, estas son Lapins, Santina, Bing, y Regina. De acuerdo a estimaciones hechas para la región de O'Higgins [Olm21] los tipos Lapins y Santina requieren 550 horas frío mientras que la variedad Bing requiere unas 600 horas, por otro lado Regina según la experiencia de los agricultores locales [Red21] requiere entre 800 y 1000 horas.

Como parte del proyecto "Transferencia y Adopción de Tecnologías para la Gestión de Riesgo en el Proceso productivo de la Cereza: Hacia una agricultura de precisión para la Región de O'Higgins", iniciativa adjudicada por la Universidad de O'Higgins y apoyada por el Centro de Estudios Avanzados en Fruticultura (CEAF), se busca la sensorización de campos en la región para la toma de decisiones mediante la predicción del comportamiento de las cosechas.

1.2. Justificación del problema

La región de O'Higgins se caracteriza por su amplio sector agronómico, diversas especies son cultivadas especialmente para su exportación a mercados de Asia y Europa. A pesar de ser

una importante industria productiva en muchos de los campos de cultivo en la región aún se utilizan métodos rudimentarios para la estimación de cambios fenológicos en las plantaciones. Dicho lo anterior la presente memoria pretende ser un aporte en la tecnificación y uso de modelos matemáticos avanzados para la estimación y predicción de indicadores fenológicos importantes en el proceso productivo.

Por la propia naturaleza de este proyecto piloto en la región de O'Higgins que incluye sensorización de campos se requiere la exploración de los datos recolectados por estos dispositivos con el fin de chequear su calidad e integridad. Luego el paso a seguir es la integración de los datos hacia algún modelo matemático que permita hacer predicciones a nivel de cambios fenológicos. Para los agricultores un modelo que permitiera conocer la distribución de horas frío acumuladas en los campos y hacer estimaciones a corto plazo de la evolución de esta métrica sería sumamente útil. Como los que trabajan en la industria agrícola de la cereza saben, la ventaja o no de unos pocos días en el inicio de la cosecha significa la perdida o ganancia de millones de pesos, esto debido a lo exigente que son las fechas de entrega y en general el proceso de venta en los puertos de China.

Con la técnica a usar sobre los datos recolectados en los campos de cerezas además de querer hacer predicciones con respecto a las horas frío acumuladas también se quiere marcar un precedente en el uso de este tipo de métodos de Machine Learning en la agricultura regional, lo que eventualmente puede derivar en más posibles aplicaciones dentro de distintas fases del proceso productivo.

1.3. Objetivos

El objetivo general del presente trabajo de título es investigar, implementar y testear modelos predictivos, llámese procesos Gaussianos, para pronosticar la distribución de indicadores fenológicos en plantaciones de cerezas dentro de la región. Debido a la existencia de múltiples indicadores fenológicos se decidió desde un principio enfocar este estudio en la horas frío acumuladas dentro de los campos, esto debido a que este indicador es uno de los más importantes que los agricultores toman en consideración, además de ser el que les permite estimar cuando sus cultivos van a despertar de su sueño y empezar a florecer.

Este desafío pretende por una parte chequear la integridad y suficiencia de los datos, sugerir posibles mejoras en el proceso de recolección de estos y por último usarlos para alimentar los modelos con los que se espera obtener información relevante con respecto a las horas frío acumu-

ladas por los árboles de cerezos, lo que posibilitaría el pronóstico de cambios fenológicos a corto plazo. Así, en el desarrollo del objetivo general se proponen los siguientes objetivos específicos:

1. Realizar una descripción de cada campo con sus características individuales y particularidades.
2. Chequear la integridad y calidad de los datos, detectar gaps de información, anomalías, puntos outliers, etc.
3. Examinar y testear el rendimiento de la técnica de pronóstico de Series de Tiempo (Time Series forecasting) SARIMA para rellenar gaps de datos en los registros de los sensores.
4. Usar modelos de procesos Gaussianos con datos georefenciados para así predecir como se distribuye la temperatura dentro de los campos, lo que de forma subsecuente deriva en la habilidad de poder predecir la distribución de horas frío.
5. Generar un modelo de horas frío que simule las hileras de árboles en los campos, usando como punto de partida las distribuciones de temperatura que estima el modelo de procesos Gaussianos.
6. Reconocer los sectores con más niveles de incertidumbre dentro de los campos, lo cual habilita la colocación de nuevos sensores de forma estratégica reduciendo así los niveles de incertidumbre y mejorando las predicciones.

1.4. Metodología

La metodología a usar está basada en KDD o Knowledge Discover Database, la cual se compone de varios pasos o etapas. KDD a diferencia de otros métodos, los cuales se centran de forma casi exclusiva en el data mining, incluye otros pasos relevantes como por ejemplo la preparación, selección, y limpieza de los datos, además de la incorporación de conocimientos específicos del área e interpretación de los resultados obtenidos del data mining. Todos estos pasos extras además del data mining son esenciales para asegurar que conocimiento útil sea extraído de los datos [Fay01].

Debido a que no existe una estandarización en aspectos como por ejemplo la distribución de los sensores en el campo, su altura con respecto al suelo y su sincronización al momento de muestrear, varias suposiciones deberán ser hechas al comenzar a trabajar. La primera suposición a hacer es que todos los sensores de un mismo tipo (temperatura ambiente, humedad del suelo, humedad de la hoja, etc) y en un mismo campo se encuentran a la misma altura con respecto al suelo. La segunda suposición es que todos los sensores de un mismo tipo se encuentran bajo

condiciones similares, por ejemplo, no hay sensores que queden completamente bajo la sombra de un árbol mientras que otros estén totalmente expuestos al sol. Finalmente la última suposición a hacer es que la medición de un sensor para para una variación de tiempo Δt lo suficientemente pequeña sufre un cambio despreciable, lo cual es necesario para promediar por horas las mediciones hechas por estos dispositivos. Habiendo hecho las suposiciones anteriores la etapas a seguir se listan a continuación:

- Implementación de los modelos predictivos previo estudio de la literatura correspondiente.
- Exploración de los datos para tomar las muestras que presenten la mayor integridad (menor cantidad de gaps, anomalías, etc).
- Preparación de los datos; realizar las transformaciones que sean necesarias sobre estos.
- Realización de múltiples pruebas y simulaciones usando las mediciones georeferenciadas capturadas por los sensores.
- Análisis de los resultados.

1.5. Resultados esperados

Los resultados esperados van en concordancia con los objetivos planteados; se listan a continuación:

1. Identificar particularidades en el comportamiento de las muestras de datos extraídas del interior de cada campo en estudio.
2. Identificar cuáles sensores presentan un mal desempeño en su funcionamiento (e.g. los que sufren prolongadas y/o frecuentes caídas).
3. Evaluar el desempeño del modelo de forecasting SARIMA sobre los gaps de información existentes en los registros de los sensores.
4. Obtener las distribuciones de temperatura a lo largo del tiempo en los campos, se espera generar simulaciones tipo timelapse que permitan vizualizar el constante cambio en terreno de esta métrica.
5. Obtener la distribución de horas frío acumuladas al final del periodo de tiempo en estudio para cada campo, así, será posible dar una estimación sobre el porcentaje de árboles con su meta de horas frío cumplida para distintos umbrales de horas frío y en distintos momentos del proceso productivo.

6. Identificar sectores con altos niveles de incertidumbre en los campos para guiar y recomendar la colocación de nuevos sensores que ayuden a mejorar las predicciones.

2. Marco Teórico

Muchos de los algoritmos de Machine Learning clásicos funcionan mediante la resolución de un problema de regresión en el cual se intenta modelar la relación que existe entre una serie de variables que se toman como input y una variable continua dependiente que se toma como output. Para poder llegar al mejor modelo que ajusta un conjunto determinado y finito de datos de entrenamiento estas técnicas frecuentemente intentan minimizar una media de error en una secuencia de pasos iterativa. Así, el problema fundamental a resolver se convierte en uno de optimización en el que se busca minimizar un cierto error mediante un algoritmo iterativo. Luego con este modelo obtenido se hace la mejor predicción para nuevos datos, los cuales se denominan datos de prueba.

En el campo del aprendizaje supervisado tradicionalmente modelos paramétricos han sido usados para resolver problemas de regresión, pero para datasets complejos estos modelos paramétricos simples pueden carecer de poder expresivo para capturar la amplia variedad de relaciones posibles entre las variables de entrada y salida [Ras04]. Cuando se trata de problemas con información georeferenciada generalmente métodos de interpolación espacial de datos (spatial data interpolation methods) son usados, los cuales usualmente funcionan bien cuando los puntos de datos están distribuidos equitativamente a lo largo del área de estudio y su densidad es buena [CPG21]. Sin embargo, estos métodos de interpolación son desafiados cuando se usan para predicciones en regiones donde los datos son escasos y están agrupados solo en partes del área de estudio [CPG21].

En las próximas secciones se hará una revisión general de la teoría que hay detrás de los procesos Gaussianos, que se levantan como una alternativa flexible, no paramétrica y con un enfoque Bayesiano para resolver problemas de regresión antes delegados a algoritmos tradicionales. Así, se pondrá a prueba y comprobará que son un poderoso y expresivo método para modelar funciones desconocidas [SSK18].

2.1. Procesos Gaussianos (PG)

Los procesos Gaussianos son procesos estocásticos, lo que corresponde a familias de variables aleatorias indexadas por tiempo o espacio, así, cada colección finita de estas variables indexadas sigue una distribución normal multivariable. Las distribuciones de los procesos Gaussianos son las distribuciones conjuntas de las infinitas variables aleatorias involucradas, lo que desde otra perspectiva se puede ver como distribuciones de probabilidad sobre funciones. Tales distribuciones

están definidas por una media (o función promedio) $m(x)$ y una función de covarianza o Kernel $K(x,x')$ capaz de generar una matriz de covarianza positiva definida, lo que se puede denotar como:

$$f(x) \sim \mathcal{GP}(m(x), K(x,x'))$$

Además, como se mencionó anteriormente cualquier colección finita de variables aleatorias $X = \{x_1, \ldots, x_n\}$ pertenecientes al dominio de entrada siguen una distribución Gaussiana multivariable, lo que se denota comúnmente como:

$$f(X) \sim \mathcal{N}(m(X), K(X,X))$$

Se va a explorar un algoritmo de regresión Bayesiana que se basa en Kernel y que se conoce como Regresión de Proceso Gaussiano (Gaussian Process Regression), pero antes de llegar a ese punto es necesario revisar algunos tópicos importantes como la distribución Gaussiana multivariable y su condicionamiento.

2.1.1. Distribución Gaussiana multivariable

Se dice que un valor vectorial x perteneciente a $\mathbb{R}^n$ sigue o se distribuye de forma normal/-Gaussiana con promedio μ perteneciente a $\mathbb{R}^n$ y matriz de covarianza Σ perteneciente al espacio de las matrices definidas positiva de $n \times n$ si se cumple que:

$$p(x;\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right),$$

lo que quiere decir que la probabilidad de x dado μ y Σ (probabilidad condicional) es Gaussiana. Formulado de otra manera se dice que un vector x sigue una distribución normal multivariable si cada combinación lineal de sus componentes se distribuye normalmente. Notar que si el vector x es de una dimensión queda la clásica distribución Gaussiana univariable. La parte fundamental de la distribución corresponde sin duda a la matriz de covarianza, la cual se define como:

$$\Sigma = \begin{bmatrix} cov(x_1,x_1) & \cdots & cov(x_1,x_D) \\ cov(x_2,x_1) & \cdots & cov(x_2,x_D) \\ \vdots & \ddots & \vdots \\ cov(x_D,x_1) & \cdots & cov(x_D,x_D) \end{bmatrix},$$

con x_k correspondiente a una variable aleatoria del vector x que sigue la distribución Gaussiana multivariable.

La matriz de covarianza Σ cumple tres propiedades, primero, es simétrica lo que significa que

la parte triangular de arriba es igual a la parte triangular de abajo, segundo, la diagonal contiene las varianzas de cada variable aleatoria, y tercero, es una matriz definida positiva lo que implica que tiene inversa.

2.1.2. Condicionamiento de una distribución Gaussiana multivariable

Supóngase que la distribución conjunta de dos variables x e y está siguiendo una distribución Gaussiana multivariable de la siguiente manera:

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_y \end{bmatrix} \right),$$

ahora se quiere saber cómo es la distribución que sigue x dado un valor particular de y, lo cual se puede expresar matemáticamente como:

$$x|y = p_{x|y}(x|y) \sim \mathcal{N}(m, P),$$

en este caso se necesita calcular la media m y la varianza P para encontrar la distribución deseada. Las expresiones que permiten llegar a estos dos valores se pueden encontrar en diferentes libros de la literatura con su respectiva demostración, en específico se destaca la demostración 4.3.4 del libro "Machine Learning. A probabilistic perspective" de Kevin P. Murphy [Mur12] por su detalle y claridad. Así, las expresiones para calcular m y P son:

$$m = \mu_x + \Sigma_{xy}\Sigma_y^{-1}(y - \mu_y) \tag{1}$$

$$P = \Sigma_x - \Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx} \tag{2}$$

2.1.3. El problema de regresión

Supóngase que se tiene un conjunto de muestras de puntos de una cierta función desconocida como sigue:

$$\{(x_1, f(x_1)), (x_2, f(x_2)), \ldots, (x_n, f(x_n))\},$$

ahora a partir de estas muestras se quiere poder dar valores estimados para nuevos puntos x_k, pero antes se deben hacer ciertas suposiciones para continuar con el proceso. Primero se hace la suposición de que las muestras vienen de una distribución Gaussiana, segundo, que las correlaciones que existen entre los puntos están dadas por las distancias a las que están unos de otros, así, los puntos más cercanos están más correlacionados y viceversa, y tercero, que los puntos que se quieren predecir vienen de una distribución Gaussiana.

Como ya se hizo la suposición de que las muestras vienen de una distribución Gaussiana se puede escribir la siguiente distribución conjunta:

$$\begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} K_{11} & K_{12} & \cdots & K_{1n} \\ K_{21} & K_{22} & \cdots & K_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ K_{n1} & K_{n2} & \cdots & K_{nn} \end{bmatrix} \right) \iff f \sim \mathcal{N}(0, K),$$

Nota: Como se señala en [SSK18] la función promedio previa (prior) se define con frecuencia como $m(x) = 0$ con el fin de evitar cálculos posteriores complejos y solo hacer la inferencia vía la función de covarianza o kernel. En la práctica la función promedio previa se logra fijar en 0 sustrayendo el promedio previo de todas las observaciones.

ahora dado un x^* se quiere predecir su correspondiente $f(x^*)$, simplemente para simplificar la notación se llamará a este término como f_*. Como previamente se hizo la suposición de que las predicciones también vienen de una distribución Gaussiana directamente se puede inferir que f_* sigue una distribución Gaussiana de media 0 y una covarianza la cual se denotará como K_{**}, formalmente:

$$f_* \sim \mathcal{N}(0, K_{**}),$$

ahora se puede proceder a realizar una distribución conjunta entre f y f_* ya que vienen de la misma distribución Gaussiana según lo supuesto:

$$\begin{bmatrix} f_* \\ f \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K_{**} & K_{*1} & K_{*2} & \cdots & K_{*n} \\ K_{1*} & \multicolumn{4}{c}{\multirow{4}{*}{$\begin{bmatrix} K_{11} & K_{12} & \cdots & K_{1n} \\ K_{21} & K_{22} & \cdots & K_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ K_{n1} & K_{n2} & \cdots & K_{nn} \end{bmatrix}$}} \\ K_{2*} & & & & \\ \vdots & & & & \\ K_{n*} & & & & \end{bmatrix} \right) \iff \begin{bmatrix} f_* \\ f \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K_{**} & K_*^T \\ K_* & K \end{bmatrix} \right),$$

ahora se procede a condicionar la distribución y como se vio anteriormente en las ecuaciones (1) y (2) las expresiones que permiten calcular el promedio y la varianza de f_* son (haciendo el reemplazo correspondiente):

$$\mu_* = K_*^T K^{-1} f$$
$$\sigma_*^2 = K_{**} - K_*^T K^{-1} K_*$$

Se puede repetir el proceso para un infinito número de puntos y es la razón por la cual se puede pensar en un proceso Gaussiano como una extensión de una distribución Gaussiana multivariable a un infinito número de dimensiones ya que se está usando un infinito número de puntos para

describir la función. Finalmente f se puede describir como:

$$f(x) \sim \mathcal{GP}(0, K(x, x_*))$$

La expresión anterior se lee como "$f(x)$ sigue un proceso Gaussiano descrito por una funcion promedio igual a 0 y una función de covarianza igual a K".

2.1.4. Estimación de la probabilidad máxima

Los Kernels usados para construir la matriz de covarianza usan parámetros los cuales son comúnmente llamados hiperparámetros, estos deben ser optimizados (encontrar los hiperparámetros ideales) de manera que la probabilidad de un buen ajuste de los datos de entrenamiento se maximice.

Dado que los procesos Gaussianos asumen que los datos de entrenamiento vienen de una distribución Gaussiana multivariable es posible expresar la probabilidad L basado en la función de densidad de probabilidad de la distribución Gaussiana multivariable:

$$L(y|\mu, \sigma^2, \theta) = \frac{1}{\sqrt{(2\pi\sigma^2)^n |K|}} \exp\left[-\frac{1}{2\sigma^2}(y-\mu)^T K^{-1}(y-\mu)\right],$$

Nota: L (al igual que $\ln(L)$) es una función cóncava, en especial se destaca la demostración de [Pré01] donde se prueba que la función de distribución de probabilidad normal estándar multivariable es cóncava para valores de argumento grandes. Además el método de demostración empleado también permite la derivación de declaraciones similares para otros tipos de funciones de distribución de probabilidad multivariable. Como sabemos que cualquier máximo local de una función cóncava es igual al máximo global se vuelve evidente que el valor óptimo o solución de la primera derivada de la expresión es único (no hay óptimos locales).

donde θ es un vector con los hiperparámetros del Kernel para los cuales se quiere encontrar los valores óptimos. Para evitar errores de redondeo es común que se prefiera maximizar el logaritmo de la probabilidad L, así, desarrollando la expresión se tiene que:

$$\begin{aligned} \ln(L) &= \ln\left[\frac{1}{\sqrt{(2\pi\sigma^2)^n |K|}} \exp\left[-\frac{1}{2\sigma^2}(y-\mu)^T K^{-1}(y-\mu)\right]\right] \\ &= -\frac{n}{2}\ln[2\pi] - \frac{n}{2}\ln\left[\sigma^2\right] - \frac{1}{2}\ln|K| - \frac{1}{2\sigma^2}(y-\mu)^T K^{-1}(y-\mu) \end{aligned} \tag{3}$$

Ahora bien, todavía se puede seguir simplificando esta expresión dado que es posible encontrar los valores óptimos de σ^2 y μ de forma analítica, lo que se debe hacer es igualar la derivada parcial con respecto a una variable a 0 para luego despejar el valor óptimo de esta variable. Así, primero

se deriva con respecto a σ^2 para luego reemplazar en (3):

$$\frac{\partial \ln(L)}{\partial \sigma^2} = 0 \implies \sigma^2 = \frac{1}{n}(y-\mu)^T K^{-1}(y-\mu)$$

Nótese cómo en la expresión que describe el valor óptimo de σ^2 está presente la media μ. Ahora se puede proceder a reemplazar en la ecuación (3) y eliminar los términos constantes para llegar al problema de optimización que se requiere resolver:

$$\theta = \underset{\theta}{\operatorname{argmax}} \left[-\frac{n}{2} \ln\left(\sigma^2\right) - \frac{1}{2} \ln |K| \right] \tag{4}$$

Al no poderse llegar a una expresión analítica para poder encontrar el valor óptimo de θ lo que se hace recurrentemente es emplear un algoritmo de optimización numérica sobre la expresión (4), tales algoritmos ya están implementados en distintas bibliotecas para diversos lenguajes de programación.

2.1.5. La función de Kernel

Como se mencionó previamente la función de covarianza también conocida como Kernel se utiliza para construir la matriz de covarianza la cual mide la similaridad entre los diferentes puntos de datos del dominio sobre el cual se trabaja, un valor alto indica más similaridad y viceversa. No cualquier función puede servir de Kernel ya que esta debe ser capaz de construir una matriz simétrica, que en su diagonal contenga las varianzas y que sea definida positiva (lo que significa que debe poseer inversa).

Las funciones de Kernel pueden ser categorizadas en dos tipos, las estacionarias y las no estacionarias. Un Kernel estacionario es una función de la distancia de las entradas, lo que quiere decir que el valor generado solamente depende de la distancia entre las entradas que se dan y que es invariante ante las traslaciones de estas. Por otra parte los Kernels no estacionarios dependen de forma directa de las entradas por lo cual una traslación sí afecta su valor, generalmente estos Kernels contienen productos punto dentro de sus expresiones.

Una característica importante que vale la pena mencionar es que cuando las entradas del Kernel tienen dos o más características o dicho de otra forma cuando las entradas del Kernel son vectores en vez de escalares se asume que la función de Kernel multidimensional $K(x_i, x_j)$ es una serie de multiplicaciones de funciones de Kernel de una dimensión por cada una de las dimensiones, el enunciado anterior se puede expresar como:

$$K(x_i, x_j) = \prod_{k=1}^{m} K(x_i^k, j_i^k),$$

donde m representa la cantidad de dimensiones de cada entrada.

A continuación se pasarán a describir dos de los Kernels más usados, estos son el Kernel RBF y el Cuadrático Racional, estás dos funciones de Kernel están íntimamente ligadas la una con la otra por lo cual es más adecuado describirlas juntas que de forma aislada. Otros Kernels comúnmente usados y que no están listados aquí se describen en el Apéndice A para la revisión de los interesados.

2.1.5.1. Kernel RBF.

El Radial Basis Funcion (RBF) también conocido como squared-exponential es uno de los Kernels más ampliamente usado dentro de la comunidad de Machine Learning. La expresión para el Kernel RBF está dada por:

$$K(x, x') = \sigma_f^2 \exp\left(-\frac{||x - x'||^2}{2l^2}\right)$$

El hiperparámetro l llamado *lengthscale* determina que tanto hay que moverse en un determinado eje en el espacio de la entrada para que los valores de la función ya no estén relacionados.

En la Subsección B.1 del Apéndice se representa una matriz de covarianza de 25×25 construida con RBF con los limites señalados en la figura, se ve el comportamiento de esta para un valor de $\sigma_f^2 = 1$ y $l = 1$. La matriz obtenida después aumentar el hiperparámetro l hasta 1.8 se encuentra en la Subsección B.2 del Apéndice; se puede apreciar claramente el efecto que tiene el aumento de l sobre la covarianza, al aumentar el hiperparámetro l sucede que el nivel de correlación aumenta para puntos más distantes.

Si se usa como segundo término de entrada el 0 en la función de covarianza es posible ver como se distribuye la correlación en base a la distancia:

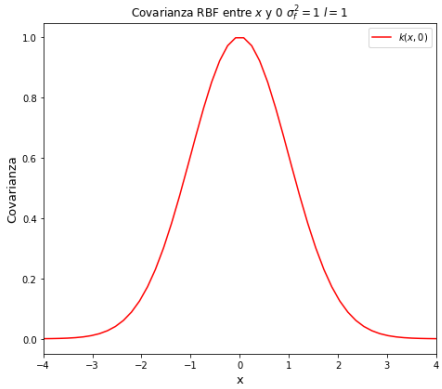


Figura 1: Covarianza en base a distancia. RBF $\sigma_f^2 = 1$, $l = 1$.

2.1.5.2. Kernel Cuadrático Racional.

La expresión para el Kernel Cuadrático Racional está dada por:

$$K(x, x') = \sigma_f^2 \left(1 + \frac{||x - x'||^2}{2\alpha l^2}\right)^{-\alpha}$$

Este Kernel es equivalente a sumar infinitamente muhos Kernel RBF con distintos *lenghscales*. El parámetro α determina el peso relativo de diferentes *lenghscales*. Cuando α tiende a infinito el Kernel Cuadrático Racional es idéntico al Kernel RBF.

2.1.6. Ejemplo de ajuste

Para poner a prueba la Regresión de Procesos Gaussianos o GPR por sus siglas en inglés se empleará en la predicción de la siguiente función en los reales:

$$f(x) = \sin(x) \log\left(x^{1.5} + 0.001\right)$$

El intervalo a trabajar es el $[0, 22]$, el Kernel seleccionado para actuar como función de covarianza es el Cuadrático Racional. Los pasos del proceso predictivo y el resultado de este se muestran en la siguiente figura:

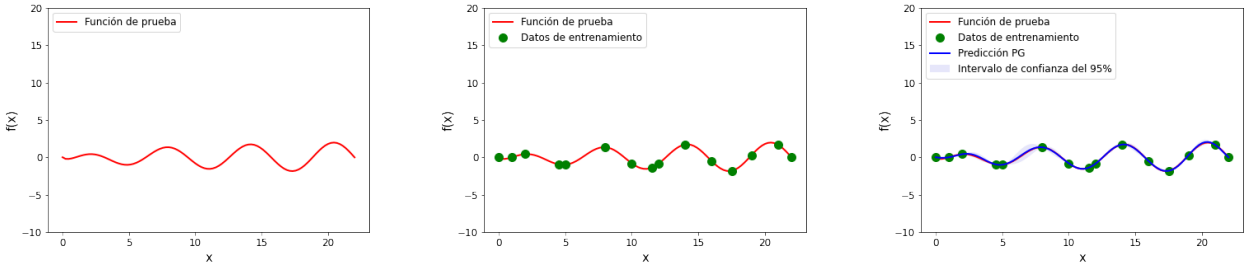


Figura 2: Proceso predictivo modelo GPR; a la izquierda se define la función, al centro se escogen los datos de entrenamiento y por último a la derecha se muestra la predicción.

Como se puede apreciar el ajuste hecho es casi perfecto teniendo además un nivel de confianza muy sólido. En la mayoría de los modelos y los procesos Gaussianos no son la excepción gran parte del desempeño depende de la calidad de los datos de entrenamiento por lo que tener muestras representativa de la función es fundamental. Es notorio como el nivel de confianza aumenta cuando los puntos están más cerca entre sí y disminuye cuando están más alejados. Así, los procesos Gaussianos además de predecir valores nuevos dan un nivel de confianza para estas nuevas predicciones lo cual los convierte en una poderosa herramienta para la toma de decisiones basadas en factores de riesgo.

2.1.7. Adaptación del modelo a datos georeferenciados

Hasta ahora se ha visto la aplicación del algoritmo de GPR para funciones tradicionales en una dimensión, ¿pero qué pasaría si se quisiera extrapolar esta técnica a datos georeferenciados?. Como se verá más adelante los datos disponibles corresponden a registros de temperatura provenientes de sensores de los cuales se conocen sus coordenadas exactas. Como el modelo no entiende de coordenadas los datos con los que se planea trabajar corresponden a las distancias en metros entre los sensores y los puntos sobre los cuales se quieren hacer predicciones, tal proceso se ilustra en la siguiente figura:

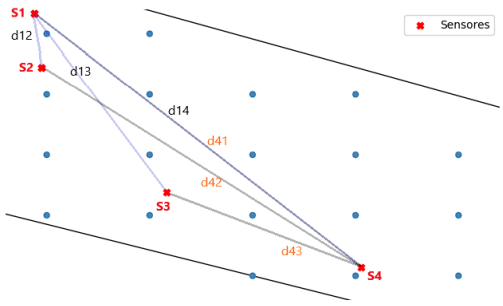


Figura 3: Obtención de set de entrenamiento para datos georeferenciados; estos corresponden a las distancias entre los sensores y las mediciones registradas.

La figura (3) muestra el proceso de obtención de datos de entrenamiento para alimentar el modelo GPR, como se puede deducir de la figura las distancias de un cierto sensor al resto de ellos (junto con la medición del sensor seleccionado) corresponden a una fila de entrenamiento. Así, siguiendo el ejemplo de la figura (3) las filas de entrenamiento serían:

Tabla 1: Datos entrenamiento ejemplo figura (3)

x_1	x_2	x_3	x_4	y
d_{11}	d_{12}	d_{13}	d_{14}	T_1
d_{21}	d_{22}	d_{23}	d_{24}	T_2
d_{31}	d_{32}	d_{33}	d_{34}	T_3
d_{41}	d_{42}	d_{43}	d_{44}	T_4

donde d_{ij} corresponde a la distancia entre el sensor i y j, y T_i a la temperatura del sensor i. Ya teniendo los datos de entrenamiento se procede a tomar nuevos puntos como se muestra en la figura (4) y entregárselos al modelo para que seguir con los pasos descritos en la sección 2.1.3.

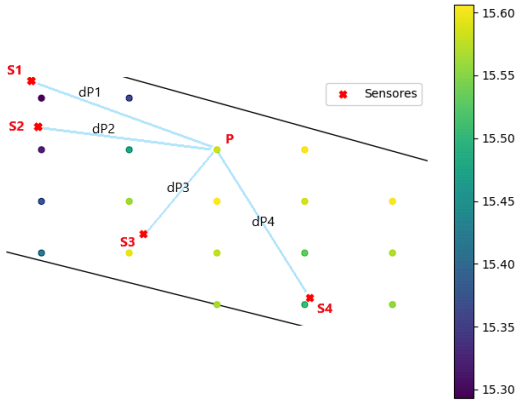


Figura 4: Proceso predictivo modelo GPR para datos georeferenciados.

2.2. Modelo ARIMA

2.2.1. Series de Tiempo

Primeramente antes de introducir el modelo ARIMA es necesario comprender qué es exactamente una Serie de Tiempo. En sencillas palabras una Serie de Tiempo corresponde a una sucesión de datos medidos en determinados instantes de tiempo y ordenados cronológicamente. Los datos pueden estar o no espaciados en intervalos iguales de tiempo aunque en general se suele trabajar con datos igualmente espaciados, además la Serie de Tiempo puede ser de una o más variables.

Los objetivos del análisis de una Serie de Tiempo corresponden a describir características im-

portantes de los patrones que esta contiene, explicar cómo los valores pasados afectan los valores futuros y finalmente hacer una predicción (forecasting) de valores futuros.

Existen dos tipos básicos de modelos para las Series de Tiempo, por una parte se tienen los modelos que intentar relacionar los valores presentes de una serie con los valores y errores de predicción pasados, a estos modelos se les llama ARIMA. Por otra parte están los modelos de regresión ordinarios que usan los índices de tiempo simplemente como variable x (variable independiente).

Las características más importantes de una Serie de Tiempo a tener en consideración al momento de realizar un análisis son [Uni22]:

- **Tendencia:** Indica si las medidas en promedio tienden a incrementar o descrecer en el tiempo.
- **Estacionalidad:** Indica si hay un patrón de altos y bajos que se repite regularmente, el cual está relacionado al tiempo del calendario como los días, semanas, meses, años, etc.
- **Outliers:** Indica si hay datos que están aislados del resto (se puede deber por ejemplo a errores de medición).
- **Ciclo a largo plazo o periodo no relacionado a la estacionalidad:** Indica si hay periodos que no siguen un comportamiento estacional.
- **Varianza Constante:** Indica si la varianza es contante en la serie a lo largo del tiempo.
- **Cambios abruptos:** Indica si hay cambios abruptos ya sea en el nivel de la serie o la varianza.

ARIMA es una acrónimo de Autoregressive Integrated Moving Average y es probablemente el modelo más ampliamente usado para aproximar Times Series, de ahí su preferencia en el presente trabajo. Como se puede ver en su nombre el modelo consta de tres componentes, un proceso autoregresivo (AR), una operación de diferenciación (I) para eliminar la tendencia y convertir la serie en estacionaria, y por último el modelo de medias móviles (MA). Las distintas componentes de este modelo serán explicadas en las siguientes secciones.

2.2.2. Modelo Autoregresivo (AR)

Un modelo de autoregresión consiste en una combinación lineal de valores pasados de una serie los cuales son usados para hacer predicciones. Matemáticamente un modelo AR de orden p

o $\mathcal{AR}(p)$ se puede expresar como:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \epsilon_t,$$

donde p es el orden, c es una constante y ϵ_t es ruido blanco.

2.2.3. Modelo de medias móviles (MA)

En vez de usar valores pasados de la variable a predecir en una regresión, el modelo de medias móviles usa errores de predicción pasados en un modelo tipo regresión el cual puede ser expresado matemáticamente de la siguiente manera:

$$y_t = c + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_q \epsilon_{t-q},$$

donde c es una constante, ϵ_t es ruido blanco y q es el orden de la regresión $\mathcal{MA}(q)$.

2.2.4. Modelo ARMA

El modelo ARMA es simplemente la combinación de los modelos anteriores, se denota como $\mathcal{ARMA}(p,q)$ donde p es el orden del modelo autoregresivo y q el orden del modelo de medias móviles y se puede expresar como:

$$y_t = c + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_q \epsilon_{t-q} + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p}$$

El modelo es capaz de tomar en cuenta los errores pasados gracias a modelo de medias móviles (MA) y los valores pasados gracias al modelo de autoregresión (AR).

2.2.5. Modelo ARIMA

Como los modelos autoregresivos (AR) normalmente se restringen a datos estacionarios antes de empezar a trabajar con ellos es necesario transformarlos en estacionarios mediante diferenciación, si es que así lo requieren. Diferenciar los datos simplemente significa sustraer un punto x_{t-1} de x_t. Así, un modelo ARIMA se denota como $\mathcal{ARIMA}(p,d,q)$, donde p corresponde al orden del modelo AR, d a el grado de diferenciación, y por último q al orden del modelo MA, y se puede expresar como:

$$y_t = c + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_q \epsilon_{t-q} + \phi_1 y'_{t-1} + \phi_2 y'_{t-2} + \ldots + \phi_p y'_{t-p},$$

donde c es una constante, ϵ_t es ruido blanco y p y q son los ordenes de los modelos autoregresivo (AR) y de medias móviles (MA) respectivamente.

2.2.6. SARIMA (Modelo ARIMA estacional)

Muchas veces en una Serie de Tiempo existe un patrón regular que se repite cada S periodos, donde S representa el número de periodos hasta que el patrón se repite de nuevo. Un ejemplo de este fenómeno podría ser la cantidad de chaquetas cortaviento vendidas mensualmente en una cierta tienda, es muy probable que haya personas que las adquieran por diferentes razones durante todo el año pero en cierta época a portas del invierno su venta crece fuertemente, así, en este caso S sería igual a 12 por los doce meses del año.

En el modelo SARIMA o ARIMA estacional los modelos AR y MA que lo componen también son estacionales, lo que en términos prácticos se traduce en que se predicen datos y errores respectivamente con información registrada en tiempos con lags que son múltiplos de S, por ejemplo, en el caso de las chaquetas cortaviento se podría predecir la venta mensual para el mes próximo x_t mirando el mismo mes del año pasado, dato que se denota x_{t-12}.

Un aspecto interesante del modelo SARIMA es que también incorpora el comportamiento no estacional a la hora de hacer predicciones, esto debido a que factores a corto plazo también contribuyen al modelo. Siguiendo con el ejemplo de la chaquetas podría darse el caso hipotético de que los materiales para confeccionarlas se hayan encarecido en el mercado internacional en los últimos meses lo que haya provocado una fuerte subida de precios y por ende un consumo más bajo por parte de los clientes a modo de ahorro. En el ejemplo anterior un modelo ideal tendría que tomar en consideración las ventas del año pasado en el mismo mes pero también la tendencia de los últimos meses.

Así, SARIMA es un modelo multiplicativo donde el primer factor corresponde a la componente no estacional y el segundo a la componente estacional, lo que se puede denotar de forma abreviada de la siguiente manera:

$$\mathcal{ARIMA}(p,d,q) \times (P,D,Q)S,$$

donde p es el orden del modelo AR no estacional, d la diferenciación no estacional, q el orden del modelo MA no estacional, P es el orden del modelo AR estacional, D la diferenciación estacional, Q el orden del modelo MA estacional y por último S es la cantidad de periodos que le toma al patrón estacional volver a repetirse.

3. Descripción de los datos

El set de datos disponible proviene de la información recolectada por múltiples sensores dispuestos en varios campos de la región de O'Higgins, tales sensores miden métricas tales como la temperatura ambiente, de la hoja, del suelo, la humedad ambiente, etc. La siguiente figura muestra una tabla que describe los datos en cuestión, nótese como los 3 primeros atributos forman una llave primaria compuesta:

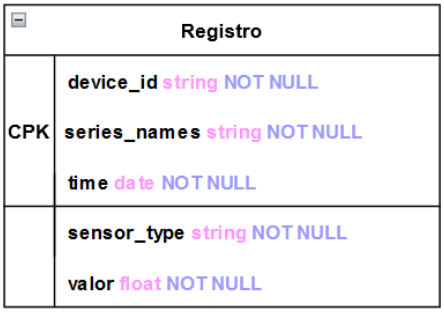


Figura 5: Tabla derivada del modelo entidad relación que describe la estructura de datos del dataset de los registros de los sensores.

La base de datos cuenta con 513.041 filas y 5 columnas o atributos, esta masiva cantidad de filas se debe a que están incorporados los registros de todos los sensores dispuestos en cada uno de los campos. Cada atributo y sus posibles valores se pasan a describir a continuación:

1. **series_names:** Tipo de medición que está haciendo el sensor.
    - **temp_environment:** Indica medición de la temperatura ambiente.
    - **temp_leaf:** Indica medición de la temperatura de la hoja.
    - **temp_soil:** Indica medición de la temperatura del suelo.
    - **humidity_environment:** Indica medición de la humedad ambiente.
    - **moisture_leaf:** Indica medición de la humedad de la hoja.
    - **water_soil:** Indica medición de la humedad del suelo.
    - **conduct_soil:** Indica medición de la conductividad del suelo.
2. **device_id:** Identificador único del sensor.
3. **sensor_type:** Tipo de sensor.
    - **LSE01:** Sensor de humedad, electroconductividad y temperatura de suelo.
    - **LSN50v2-S31B:** Sensor de humedad y temperatura ambiente.

- **LLMS01:** Sensor de humedad y temperatura de la hoja.

4. **time:** Fecha y hora del registro.
5. **valor:** Valor del registro.

4. Caracterización de los campos

4.1. Campo CEAF Rengo

Este campo se encuentra en la comuna de Rengo a un costado de la ruta Panamericana Sur. Es el más avanzado de todos en cuanto a tecnificación ya que está a cargo del Centro de Estudios Avanzados en Fruticultura (CEAF). Sus coordenadas, forma y la disposición de los sensores se muestran en la siguiente figura:

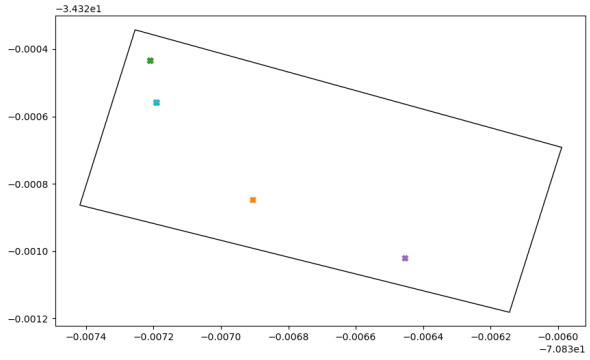


Figura 6: Campo Rengo CEAF

Tabla 2: Medidas campo CEAF Rengo

Perímetro (m)	361
Área (m^2)	7090
Área (ha)	0.71

En la figura se ven varias locaciones, las locaciones púrpura y naranja tienen un sensor cada una mientras que la celeste y verde tienen tres sensores cada una. Para este estudio los sensores relevantes y a tomar en consideración son los que miden temperatura ambiente, para esta medición solo hay 3 sensores operativos a plenitud los cuales están en la posiciones verde, celeste y púrpura mientras tanto el sensor de la posición naranja tiene una cantidad de registros muy pequeña debido a sus fallas por lo cual se decidió simplemente no tomarlo en cuenta.

4.2. Campo Graneros

Este campo se encuentra en la comuna de Graneros a un costado de la Ex Ruta 5 a una distancia aproximada de 32 kilómetros del primer campo. Sus coordenadas, forma y la disposición de los sensores se muestran en la siguiente figura:

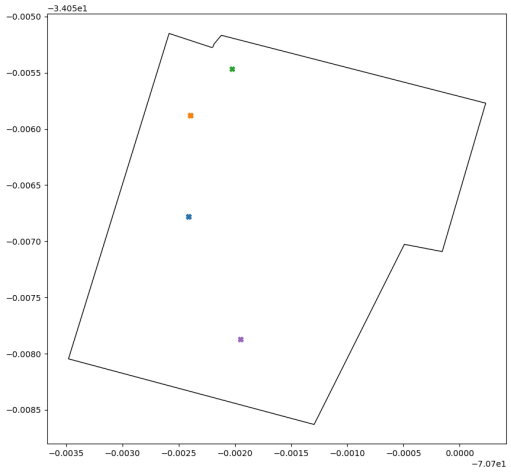


Figura 7: Campo Graneros

Tabla 3: Medidas campo Graneros

Perímetro (m)	1199
Área (m^2)	83817
Área (ha)	8.38

La locación verde tiene un sensor, la púrpura dos, la azul tres y por último la locación naranja tiene cuatro sensores. En este caso por un error humano ocurre la situación opuesta que en el caso de CEAF donde hay una locación que quedó sin sensor de temperatura ambiente, aquí hay dos sensores de temperatura ambiente en la misma locación (locación naranja).

4.3. Campo Peumo

El campo de Peumo se localiza a una costado de la Ruta 66. Sus coordenadas, forma y la disposición de los sensores se muestran en la siguiente figura:

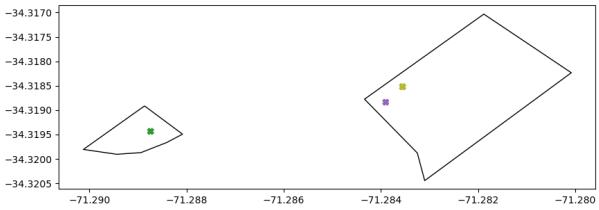


Figura 8: Campo Peumo

Tabla 4: Medidas campo Peumo

	Sector 1 (Este)	Sector 2 (Oeste)
Perímetro (m)	1103	448
Área (m^2)	68529	10345
Área (ha)	6.85	1.03

Como se puede apreciar el campo de Peumo está a su vez dividido en dos sectores, el sector Oeste cuenta con 3 sensores los cuales están posicionados en una sola locación, la locación verde, por otra parte el sector Este cuenta con la locación púrpura con 2 sensores y la locación amarilla con 3 sensores.

4.4. Campo Requínoa

Este campo se encuentra en la comuna de Requínoa a un costado de la Ruta H-409. Sus coordenadas, forma y la disposición de los sensores se muestran en la siguiente figura:

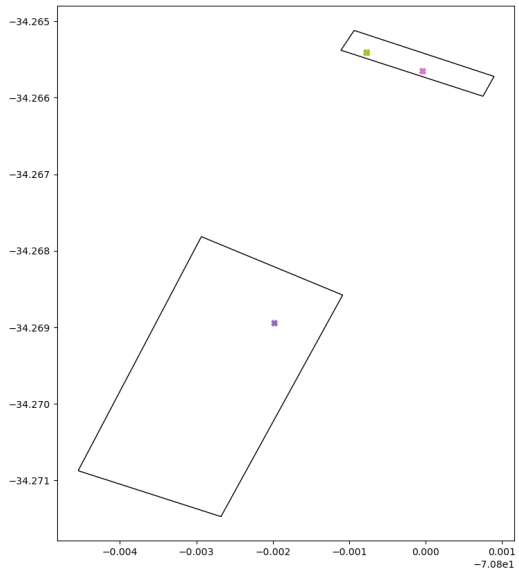


Figura 9: Campo Requinoa

Tabla 5: Medidas campo Requínoa

	Sector 1 (Norte)	Sector 2 (Sur)
Perímetro (m)	429	1098
Área (m^2)	5857	67672
Área (ha)	0.59	6.77

El campo de Requínoa esta dividido en 2 sectores al igual que en el caso anterior, el sector Sur cuenta con 1 sensor obviamente en una sola locación mientras que el sector Norte cuenta con dos locaciones, la locación rosa con 2 sensores y la locación amarilla con 3 sensores.

4.5. Campo Rengo

Este campo se encuentra a un costado de la Ruta Panamericana Sur en la comuna de Rengo. Sus coordenadas, forma y la disposición de los sensores se muestran en la siguiente figura:

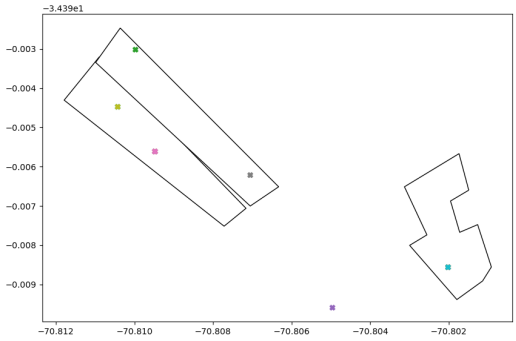


Figura 10: Campo Rengo

Tabla 6: Medidas campo Requínoa

	Sector 1 (Oeste-Sur)	Sector 2 (Este)	Sector 3 (Oeste-Norte)
Perímetro (m)	1284	1088	1324
Área (m^2)	57045	42784	54226
Área (ha)	5.70	4.28	5.42

Este campo cuenta con 6 locaciones distintas donde hay dispuestos sensores lo que lo convierte en el campo mejor sensorizado de todos. Dentro del campo hay 3 sectores, un sector al Este y dos sectores en el Oeste lo cuales están muy cerca el uno del otro, por simplicidad y conveniencia estos se pasarán a nombrar como sector Oeste-Norte y sector Oeste-Sur. Como se puede apreciar en el sector Este hay solo una locación, la locación celeste la cual cuenta con 3 sensores. Por otro lado el sector Oeste-Sur cuenta con dos locaciones, la locación rosa con 3 sensores y la locación amarilla con 2 sensores. Finalmente el sector Oeste-Norte cuenta con dos locaciones, la locación gris y verde con un sensor cada una. Además de los sensores dentro de los tres sectores antes descritos existen un sensor exterior en la parte sur (locación púrpura).

5. Resultados y discusión

5.1. Modelo SARIMA y gaps de datos

Es común para el sistema de recolección de datos mediante sensores en los campos sufrir caídas de vez en cuando debido a diversas razones, por ejemplo, cortes del suministro eléctrico, caídas del servicio de internet o por las propias fallas de los sensores. Una prolongada caída se extendió entre las fechas (en formato timestamp) 2022-09-12 22:00 UTC y 2022-09-15 14:00 UTC debido a una caída en el servicio de internet, el cual solo después de unos dos días y medio se pudo reponer. En el registro de los sensores de puede ver la disrupción del servicio y el severo gap de datos que existe y que corta la serie en dos partes, en este caso se usará un sensor del campo CEAF como referencia:

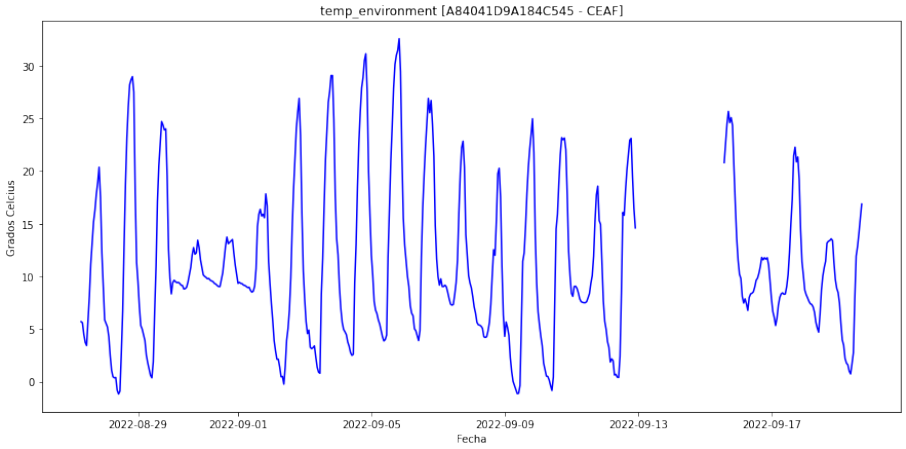


Figura 11: Severo gap de datos registrado por uno de los sensores del campo Rengo CEAF.

Las librerías que implementan ARIMA reciben una Serie de Tiempo para luego resolver un problema de optimización que permita encontrar los parámetros óptimos del modelo. Su funcionamiento se basa en probar y comparar múltiples modelos y elegir el que minimiza un criterio de selección de modelo (model selection criterion) previamente establecido. Un criterio de selección evalúa si un modelo ajustado ofrece un balance óptimo entre la bondad de ajuste y la complejidad del modelo de manera que modelos que son demasiado simplistas para describir los datos o innecesariamente complejos son descalificados [CN19]. El criterio de selección de modelo más comúnmente usado es el Akaike information criterion (AIC), este estimador estrega un score que permite comparar que tan bien 2 o más modelos ajustan un conjunto de datos.

Así, el problema de encontrar el modelo óptimo que describe la Serie de Tiempo se reduce al clásico problema de minimización de una función, la cual tiene por entradas los paramétros del

modelo y como salida el score entregado por el criterio de selección de modelo. Dependiendo de lo larga que sea la serie a ajustar el tiempo que les toma a los algoritmos implementados encontrar el mejor modelo puede variar enormemente, en este caso se usaran los últimos 400 valores antes del gap de datos que se vio previamente, con tal cantidad de valores el ajuste toma de 4 a 5 minutos.

Al hacer el ajuste usando el criterio de selección de modelo AIC y luego de predecir los datos faltantes la serie queda de la siguiente manera:

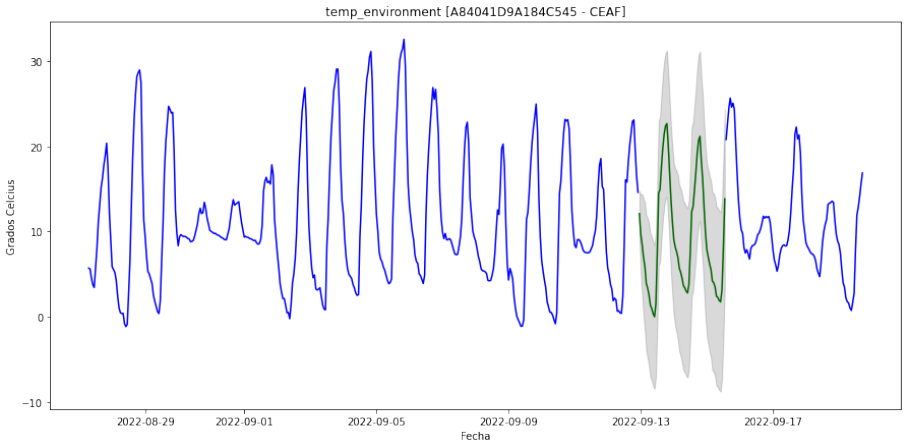


Figura 12: Gap de datos "parchado" usando SARIMA con el criterio de selección de modelo AIC.

donde la línea verde representa los datos predichos y el área de color gris el intervalo de confianza del 95 %. Al ver cómo la serie predicha se integra dentro del gap de datos da la sensación de que la predicción calza de forma prácticamente perfecta dentro de este, de alguna u otra forma la serie luce como si ninguna disrupción hubiera ocurrido. Usando la notación introducida en el marco teórico se puede expresar el modelo SARIMA ajustado como:

$$\mathcal{ARIMA}(2,0,2) \times (2,1,0)24$$

Aunque el resultado anterior dejó la sensación de una predicción precisa/exacta es necesario contrastar una serie predicha con la registrada por el sensor debido a que el rendimiento de SARIMA para predecir datos faltantes depende directamente de los datos previos, si estos registros son demasiado irregulares y/o erráticos, como en este caso, la precisión en la predicción de SARIMA se verá fuertemente afectada. A continuación usando otro segmento de la serie del mismo sensor de CEAF se compara la serie de datos predicha por el modelo SARIMA con la serie de datos registrada, adicionalmente se incluyen los datos registrados por la estación meteorológica más cercana INIA Rayentué - Rengo ubicada a unos 0.22 kilómetros de distancia:

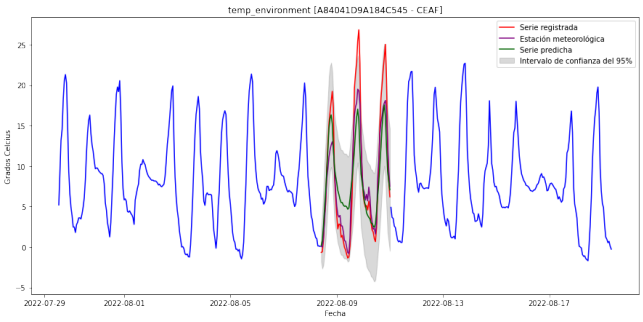


Figura 13: Serie predicha, registrada por el sensor y por la estación más cercana.

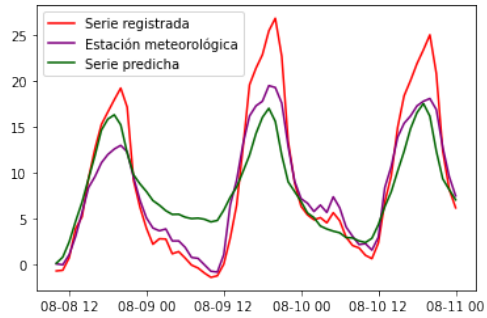


Figura 14: Serie predicha, registrada por el sensor y por la estación más cercana [zoom].

Aquí se puede apreciar como la irregularidad y lo errático de los datos previos afecta gravemente la precisión y el nivel de incertidumbre de la serie predicha dando como resultado un nivel de error más que severo. Si comparamos el acumulado de errores absolutos de la serie predicha versus la estación meteorológica se puede ver que el error acumulado de la estación es casi un 38% menor.

El ejemplo recién expuesto de error entre una serie predicha y la secuencia real registrada no es aislado y se extrapola a todos los gaps de datos de todos lo sensores. En este caso la motivación principal para usar SARIMA era poder llenar los datos faltantes en los sensores para así luego calcular las horas frío acumuladas o al menos hacer una estimación, pero debido a lo grosero que resultan ser los errores de las series predichas ese camino se vuelve inviable a razón del severo nivel de distorsión que causaría incluirlas en el cálculo de la horas frío acumuladas. Es necesario recordar lo sensible que es el cálculo de las horas frío para pequeñas variaciones de temperatura, a modo de ejemplificación, la minúscula diferencia de 0.4 grados que existe entre los 6.9 y 7.3 grados decide si la hora en cuestión se cuenta o no como hora frío. Al no ser raro que las diferencias entre una serie predicha por SARIMA y los valores reales sean de 4, 5 o incluso 6 grados se puede entender de mucho mejor forma por qué no tendría sentido integrar estos resultados al cálculo de este importante indicador como lo es las horas frío acumuladas.

Así, el uso de SARIMA para rellenar datos faltantes de estos sensores de temperatura se debería restringir solo a situaciones donde la cantidad de gaps de información sea pequeña con respecto al total de registros disponibles, a diferencia de este caso donde los datos faltantes de algunos sensores dentro del periodo de estudio pueden llegar a sumar varias semanas lo que representa porcentualmente un enorme número de datos perdidos. Por otro lado si es esencial rellenar gaps de datos de largos periodos de tiempo, ya sean días o incluso semanas, una opción más viable es usar los datos proporcionados por la estación meteorológica más cercana.

5.2. Distribución de temperatura

La primera tarea a realizar es predecir cómo se distribuye la temperatura en los campos para instantes de tiempo puntuales, esto usando los datos proporcionados por los sensores los cuales están dispuestos en lugares fijos en los terrenos. Los sensores toman mediciones aproximadamente cada 20 minutos pero debido a que no están sincronizados entre sí para este estudio los instantes de tiempo puntuales a considerar serán las horas, así, las series de temperatura se conformarán por los valores promedio por hora registrados por los sensores. Debido a la forma recién descrita en la que se trabajarán los datos el conteo de horas frío para las distintas locaciones se vuelve sumamente directo. Luego de esta primera tarea el objetivo es obtener las progresiones en el tiempo de las distribuciones de temperatura para cada campo.

5.2.1. Campo CEAF Rengo

Para poder predecir la distribución de temperatura usando una regresión de proceso Gaussiano (GPR) es necesario contar con datos de entrenamiento. En este caso se quiere calcular la distribución de temperatura para a una cierta hora en un cierto día. Así, los datos de entrenamiento a usar son las temperaturas promedio registradas por lo sensores (a una cierta hora) junto con las distancia que hay entre ellos:

$$\{d_{i1}, d_{i2}, \ldots, d_{in}\} \rightarrow \{T_i\}$$

donde d_{ik} representa la distancia entre el sensor i-ésimo y k-ésimo y T_i es la temperatura promedio registrada por el sensor i-ésimo durante una hora determinada. Para más detalles dirigirse a la sección 2.1.7 “Adaptación del modelo a datos georeferenciados”.

Luego de definir los datos de entrenamiento es necesario crear una grilla sobre el campo cuyos puntos actuarán como datos de prueba. Un grilla de ejemplo para el campo CEAF se incluye en la Subsección C.1 del Apéndice. La fecha y hora elegida para hacer la predicción de la distribución de temperatura (promedio) sobre el campo es el 21 de julio de 2022 entre las 16 y 17 horas, esta fecha y hora no tienen ninguna característica en especial, simplemente se eligieron de forma discrecional a modo de probar el desempeño del modelo. Aplicando el modelo y haciendo las predicciones correspondientes se obtiene la siguiente distribución de temperatura sobre el campo:

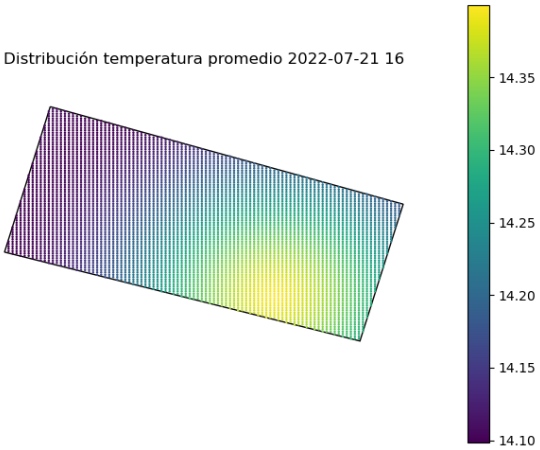


Figura 15: Distribución de temperatura para instante de tiempo puntual campo CEAF Rengo.

A primera vista pareciera que la temperatura varía enormemente entre el color morado y amarillo, pero al ver el rango de menos de 1 grado es claro que este no es el caso. Para ver la progresión en el tiempo de la predicción de distribución de temperatura no tiene sentido usar diferentes rangos para cada instante de tiempo diferente por lo cual lo ideal es tomar como rango la temperatura mínima y máxima registrada en el campo durante el periodo de estudio.

Como parte de los resultados de la presente memoria se adjunta el archivo de vídeo de nombre "CEAF_TEMPERATURA_TIMELAPSE.mp4", el cual es un timelapse de la distribución de temperatura en el tiempo generada a partir de la predicciones de las regresiones de procesos Gaussianos. En esta secuencia se puede ver la utilidad de haber utilizado esta técnica de aprendizaje de máquinas, los sensores dispuestos en terrero sirvieron como datos de entrenamiento con los cuales se pudo estimar la temperatura de cada punto/locación individualmente en el campo.

5.2.2. Campo Graneros

Los datos de entrenamiento son definidos de la misma forma que en el caso anterior y de la misma manera se aplica la grilla sobre el polígono que representa el campo. Una grilla de ejemplo se encuentra en la Subsección C.2 del Apéndice. Para el día 21 de julio a las entre las 16 y 17 horas la distribución de la temperatura (promedio) sobre el campo es como sigue:

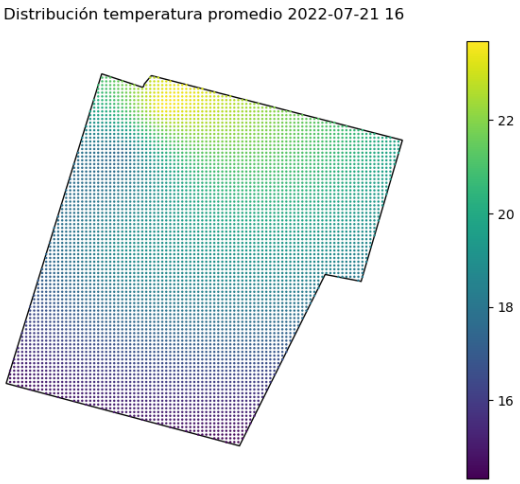


Figura 16: Distribución de temperatura para instante de tiempo puntual campo Graneros.

Al igual que en el caso anterior carece de sentido ver una distribución de temperatura para un instante aislado de tiempo, por lo cual se adjunta un timelapse con el nombre "GRANEROS_TEMPERATURA_TIMELAPSE.mp4" como parte de lo resultados.

Nota: Al ser el proceso exactamente igual que en el caso de CEAF y Graneros, y en aras de mantener la brevedad del texto, para los campos de Peumo, Requínoa y Rengo los resultados en formato timelapse (vídeo) serán adjuntados junto con el documento con la misma convención de nombre usada hasta ahora.

5.3. Distribución de horas frío acumuladas

Las horas frío (HF) son una importante medición en el mundo de la agricultura, se definen como la cantidad de horas bajo 7 grados Celsius a las que las plantaciones han estado expuestas. Gracias al cálculo de estas horas frío es posible saber qué cultivos van a despertar primero de su sueño para florecer. Ya que mediante el algoritmo de regresión de procesos Gaussianos previamente implementado es posible hacer una estimación de temperatura en cualquier locación del campo ahora se cuenta con una poderosa herramienta que permite predecir las horas frío acumuladas a lo largo y ancho del terreno, lo que a su vez da la posibilidad de predecir qué sectores van a florecer primero y cuáles después.

5.3.1. *Campo CEAF Rengo*

Para el campo CEAF las estadísticas de las horas frío acumuladas predichas por el modelo y registradas por lo sensores desde el 20 de julio hasta el 3 de octubre de 2022 se listan a continuación junto con el gráfico que muestra la distribución de las horas frío acumuladas predichas, en este caso y en los siguientes se usa una representación que simula hileras de árboles:

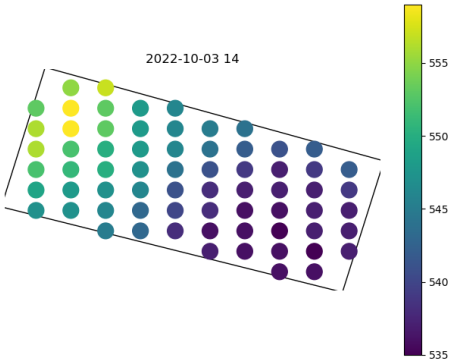


Figura 17: CEAF [Distribución de horas frío acumuladas predichas]

Tabla 7: CEAF [HF acumuladas]

$\max_h$	559
$\min_h$	535
$\overline{h}$	543.69
σ_h	6.76
$\overline{h}_{reg}$	548.33
$\sigma_{h_{reg}}$	22.42

Nótese cómo de similar es el promedio de las horas frío acumuladas predichas por el modelo versus el promedio de las horas frío acumuladas registradas por los sensores, y en cambio lo diferente que son sus desviaciones estándar. Aunque la desviación estándar de las horas frío acumuladas registradas por los sensores es un dato empírico el hecho de que sea calculada usando una pequeña muestra de apenas 3 locaciones/sensores hace que sea muy poco fiable. Supóngase por ejemplo el caso en el que se quisiera saber la desviación estándar de las alturas de un grupo de 100 personas y para estimarlo apenas se eligieran a 3 de ellas, dependiendo de cuales personas se hayan elegido el resultado podría variar muchísimo y estar más o menos alejado del valor real. Por otra parte la desviación estándar de las horas frío acumuladas predichas se calcula tomando decenas y decenas de muestras obtenidas gracias a la ayuda de los procesos Gaussianos, a lo largo y ancho de todo el campo, por lo cual se puede estar altamente seguro de que es un número muchísimo más representativo y cercano al valor real.

Los agricultores miden las horas frío entre el 1 de mayo y el 31 de julio pero lamentablemente los datos disponibles solo están desde la segunda quincena de julio, debido a esto para el siguiente análisis se supondrá que los datos medidos están dentro de la temporada y que la variedad Lapins la cual requiere unas 550 horas frío es la tratada. Tomando en cuenta la distribución de las horas

frío predichas la siguiente tabla muestra el porcentaje de árboles (representados por las distintas locaciones) que han acumulado más de un umbral establecido de horas frío, incluyendo el umbral de las 550 horas del tipo Lapins. Tomando el promedio de horas frío acumuladas durante los últimos días es posible hacer una estimación sobre el porcentaje de árboles con sus horas frío listas para los próximos días, tal estimación también se agrega en la tabla:

Tabla 8: CEAF [Porcentaje de árboles despiertos según umbral HF]

Umbral Horas Frío	Porcentaje de árboles despiertos	Estimación 1 día	Estimación 2 días	Estimación 3 días
530	97.86%	98.65%	99.18%	99.51%
535	90.07%	92.95%	95.14%	96.75%
540	70.74%	76.8%	82.09%	86.54%
545	42.32%	49.7%	57.10%	64.26%
550	17.53%	22.75%	28.75%	35.41%
555	4.72%	6.85%	9.67%	13.27%
560	0.79%	1.30%	2.07%	3.19%

Nota: El umbral de horas frío representa la meta de horas que se quiere alcanzar, supóngase que un agricultor requiere 550 horas acumuladas para sus cultivos, tal cifra correspondería a su umbral. Así, dependiendo del umbral o meta necesaria para los cultivos la cantidad de árboles despiertos (árboles con una cantidad de horas frío acumuladas mayor al umbral) varía. En la tabla de arriba y en las posteriores se muestran varios umbrales a modo de simular distintos escenarios.

5.3.2. Campo Graneros

Para el campo Graneros las estadísticas de las horas frío acumuladas predichas por el modelo y registradas por lo sensores desde el 20 de julio hasta el 3 de octubre de 2022 se listan a continuación junto con el gráfico que muestra la distribución de las horas frío acumuladas predichas:

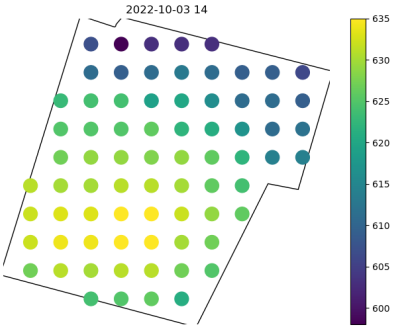


Figura 18: Graneros [Distribución de horas frío acumuladas predichas]

Tabla 9: Graneros [HF acumuladas]

$\max_h$	635
$\min_h$	598
$\overline{h}$	622.50
σ_h	9.57
$\overline{h}_{reg}$	613
$\sigma_{h_{reg}}$	28.85

En la tabla se puede detectar un interesante fenómeno el cual es que a diferencia del campo CEAF la media de horas frío registradas es menor a la media de horas frío predichas, esta situación depende de cómo el modelo interpretó los datos de entrenamiento de cada campo lo que está estrechamente ligado a las posiciones de los sensores y los valores que registran. Al igual que en el caso de CEAF la desviación estándar registrada y predicha difieren bastante.

La tabla muestra el porcentaje de árboles despiertos (con su meta de horas frío completada) para distintos umbrales de horas frío, basándose en la distribución de las horas frío acumuladas predichas. Para unas cerezas tipo Bing las cuales requieren unas 600 horas la mayoría del campo ya estaría despierto y floreciendo. Al igual que en el caso anterior se agregan las estimaciones para los siguientes 3 días:

Tabla 10: Graneros [Porcentaje de árboles despiertos según umbral HF]

Umbral Horas Frío	Porcentaje de árboles despiertos	Estimación 1 día	Estimación 2 días	Estimación 3 días
600	99.06%	99.42%	99.65%	99.79%
605	96.63%	97.72%	98.50%	99.04%
610	90.43%	93.02%	95.04%	96.56%
615	78.34%	83.02%	87%	90.28%
620	60.30%	66.73%	72.71%	78.09%
625	39.70%	46.42%	53.25%	59.98%
630	21.66%	27.02%	32.96%	39.37%
635	9.57%	12.82%	16.77%	21.42%
640	3.37%	4.87%	6.87%	9.43%

5.3.3. Campo Peumo

Para el campo Peumo las estadísticas de las horas frío acumuladas predichas por el modelo y registradas por lo sensores desde el 20 de julio hasta el 3 de octubre de 2022 se listan a continuación junto con el gráfico que muestra la distribución de las horas frío acumuladas predichas:

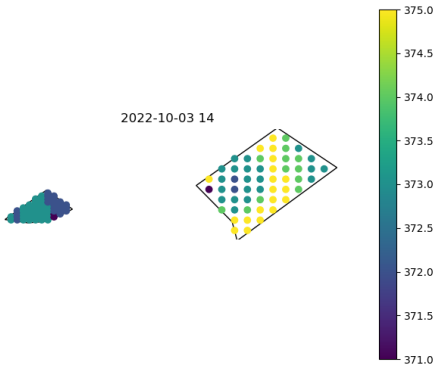


Figura 19: Peumo [Distribución de horas frío acumuladas predichas]

Tabla 11: Peumo [HF acumuladas]

	Sector 1 (Este)	Sector 2 (Oeste)
$\max_h$	375	373
$\min_h$	371	371
$\overline{h}$	373.83	372.56
σ_h	1.01	0.56
$\overline{h}_{reg}$	378	375
$\sigma_{h_{reg}}$	1	0

Los dos sectores tienen valores similares excepto en las desviaciones estándar. Para el sector 1 la desviación estándar de las horas frío registradas equivale a 1 mientras que para el sector 2 equivale a 0, esto se debe a que en el sector 2 solo hay un sensor y por definición la desviación estándar de la horas frío registradas siempre será 0. Para la desviación estándar predicha el valor del sector 1 es mucho más cercano al del sector 2, el cual tiene un valor bastante más realista que simplemente 0.

Como las desviaciones estándar son bastante pequeñas en comparación con los campos anteriores los umbrales de horas frío simulados son mucho más acotados y cercanos entre sí. Simulando el caso de árboles con un bajo requerimiento de horas frío los porcentajes de árboles despiertos según umbral y sector son:

Tabla 12: Peumo [Porcentaje de árboles despiertos según umbral HF]

Umbral Horas Frío	Porcentaje de árboles despiertos sector 1	Porcentaje de árboles despiertos sector 2
370	99.99%	100%
371	99.75%	99.73%
372	96.5%	84.13%
373	79.44%	21.60%
374	43.32%	0.51%
375	12.33%	0%
376	1.58%	0%

Para los interesados, en la Subsección D.1 del Apéndice se pueden encontrar dos tablas con las predicciones a tres días para cada sector.

5.3.4. Campo Requínoa

Para el campo Requínoa las estadísticas de las horas frío acumuladas predichas por el modelo y registradas por lo sensores desde el 20 de julio hasta el 3 de octubre de 2022 se listan a continuación junto con el gráfico que muestra la distribución de las horas frío acumuladas predichas:

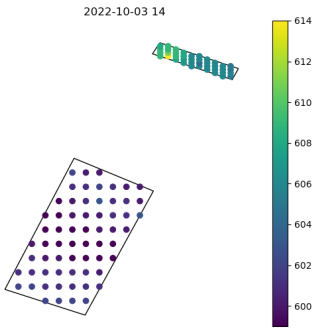


Figura 20: Requínoa [Distribución de horas frío acumuladas predichas]

Tabla 13: Requínoa [HF acumuladas]

	Sector 1 (Norte)	Sector 2 (Sur)
$\max_h$	614	603
$\min_h$	604	599
$\overline{h}$	607.05	600.39
σ_h	1.90	1.11
$\overline{h}_{reg}$	609	598
$\sigma_{h_{reg}}$	27	0

Se puede notar en la tabla la abismal diferencia de las desviaciones estándar entre las horas frío registradas por los sensores de cada sector del campo, lo cual se debe a la existencia de un solo sensor en el sector 2 del campo, lo que hace que por definición la desviación estándar sea 0 en tal sector. Luego de hacer las estimaciones de horas frío acumuladas para las hileras de árboles se puede ver como las desviaciones estándar de ambos sectores son mucho más cercanas y por lo demás realistas.

Los porcentajes de árboles con sus horas frío cumplidas para distintos umbrales se ven a continuación. Nótese cómo el sector 2 llega mucho más rápido a porcentajes cercanos a 0, esto debido a la menor desviación estándar predicha con respecto al sector 1:

Tabla 14: Requínoa [Porcentaje de árboles despiertos según umbral HF]

Umbral Horas Frío	Porcentaje de árboles despiertos sector 1	Porcentaje de árboles despiertos sector 2
598	100%	98.43%
600	99.99%	63.73%
602	99.61%	7.35%
604	94.58%	0.06%
606	70.97%	0%
608	30.85%	0%
610	6.03%	0%
612	0.46%	0%

Para los interesados, en la Subsección D.2 del Apéndice se pueden encontrar dos tablas con las predicciones a tres días para cada sector.

5.3.5. Campo Rengo

Para el campo Rengo las estadísticas de las horas frío acumuladas predichas por el modelo y registradas por lo sensores desde el 20 de julio hasta el 3 de octubre de 2022 se listan a continuación junto con el gráfico que muestra la distribución de las horas frío acumuladas predichas:

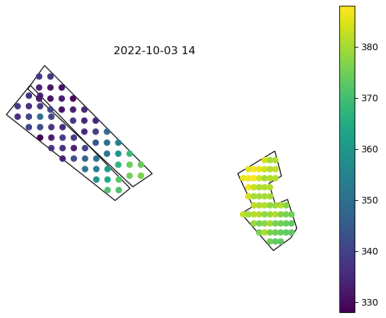


Figura 21: Rengo [Distribución de horas frío acumuladas predichas]

Tabla 15: Rengo [HF acumuladas]

	Sector 1 (Oeste-Sur)	Sector 2 (Este)	Sector 3 (Oeste-Norte)
$\max_h$	372	388	376
$\min_h$	328	373	329
$\overline{h}$	344.18	379.70	345.06
σ_h	12.08	3.95	15.90
$\overline{h}_{reg}$	336	375	348.50
$\sigma_{h_{reg}}$	11	0	22.5

Se puede ver la similitud entre los valores del sector 1 y el sector 3 mientras que los valores del sector 2 están más disparados. Como es de esperar debido a que en el sector 2 solo existe un sensor la desviación estándar para las horas frío acumuladas registradas es 0 lo que mediante el uso de los procesos Gaussianos se puede revertir y así obtener una estimación mucho mas cercana a la realidad.

Debido a que la acumulación de horas frío en el sector 2 es mucho más alta que el caso de los sectores 1 y 3 el rango del umbral de horas frío acumuladas que se va a simular va a ir desde las 320 a las 390 horas, tal rango es el más amplio de todos los campos:

Tabla 16: Rengo [Porcentaje de árboles despiertos según umbral HF]

Umbral Horas Frío	Porcentaje de árboles despiertos sector 1	Porcentaje de árboles despiertos sector 2	Porcentaje de árboles despiertos sector 3
320	97.73%	100%	94.25%
325	94.38%	100%	89.65%
330	87.98%	100%	82.82%
335	77.64%	100%	73.65%
340	63.53%	100%	62.48%
345	47.29%	100%	50.15%
350	31.5%	100%	37.8%
355	18.52%	100%	26.59%
360	9.52%	100%	17.37%
365	4.24%	99.99%	10.49%
370	1.63%	99.3%	5.84%
375	0.54%	88.3%	2.98%
380	0.15%	46.97%	1.4%
385	0.04%	8.98%	0.6%
390	0.01%	0.46%	0.24%

Para los interesados, en la Subsección D.3 del Apéndice se pueden encontrar tres tablas con las predicciones a tres días para cada sector.

5.4. Localización óptima para nuevos sensores

Al hacer una predicción el proceso Gausiano es capaz de generar la varianza correspondiente a ella. Una varianza alta se puede interpretar como un alto nivel de incertidumbre para la predicción mientras que una varianza baja como un bajo nivel de incertidumbre. Si este proceso se sistematiza para cada locación en el terrero será posible determinar cuales áreas son las que poseen una mayor y menor varianza, y por consiguiente será posible decidir a ciencia cierta cuáles son los sectores óptimos en el campo para la colocación de nuevos sensores.

5.4.1. *Campo CEAF Rengo*

Las estadísticas de la distribución de varianza acumulada junto con un gráfico que permite ver la distribución en terreno están a continuación:

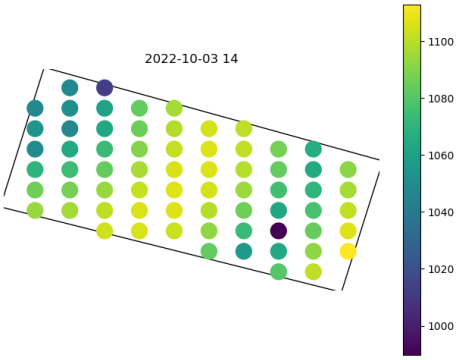


Figura 22: Distribución de varianza acumulada

Tabla 17: CEAF [Estadísticas de varianza acumulada]

$\max_{\sigma^2_{ac}}$	1112.80
$\min_{\sigma^2_{ac}}$	989.74
$\overline{\sigma^2_{ac}}$	1083.08
$\sigma_{\sigma^2_{ac}}$	23.19

Se ve como la incertidumbre en este caso representada por la acumulación de varianza es alta en el interior del campo (tonalidades amarillas y verdes), además de ser alta en la orilla Este. La mancha amarilla de incertidumbre que se encuentra en el centro probablemente responde al hecho de que los sensores están bastante cerca de las orillas y no se adentran lo suficiente en el terreno. Lo ideal sería poner un dispositivo de monitoreo en el centro de esta región coloreada lo cual se podría implementar en una próxima temporada de cerezas.

5.4.2. *Campo Graneros*

Las estadísticas de la distribución de varianza acumulada junto con un gráfico que permite ver la distribución en terreno están a continuación:

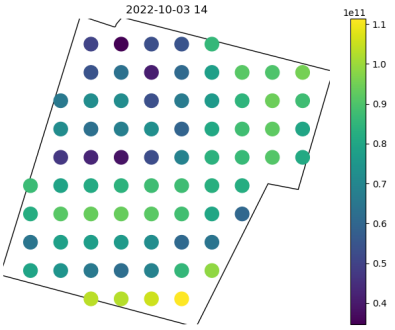


Figura 23: Graneros [Distribución de varianza acumulada]

Tabla 18: Graneros [Estadísticas de varianza acumulada]

$\max_{\sigma^2_{ac}}$	1.11e11
$\min_{\sigma^2_{ac}}$	3.46e10
$\overline{\sigma^2_{ac}}$	7.55e10
$\sigma_{\sigma^2_{ac}}$	1.66e10

En el campo Graneros no se aprecian zonas con un excesivo nivel de varianza, esto probablemente responde a la mayor cantidad de sensores operativos en terreno como también a la amplia y representativa área que estos cubren. Aun así, se puede notar que en la zona noroeste hay una región de varianza más eleveda con respecto al resto del campo, lo cual posiblemente se debe a que los sensores están posicionadas más cerca de la mitad Oeste del terreno, así, si se quisieran poner nuevos sensores en el campo Graneros esta región sería la ideal para tal cometido.

5.4.3. Campo Peumo

Las estadísticas de la distribución de varianza acumulada junto con un gráfico que permite ver la distribución en terreno están a continuación:

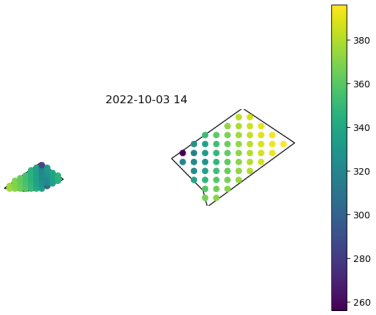


Figura 24: Peumo [Distribución de varianza acumulada]

Tabla 19: Peumo [Estadísticas de varianza acumulada]

	Sector 1 (Este)	Sector 2 (Oeste)
$\max_{\sigma^2_{ac}}$	396.19	375.97
$\min_{\sigma^2_{ac}}$	256.08	286.12
$\overline{\sigma^2_{ac}}$	362.52	341.91
$\sigma_{\sigma^2_{ac}}$	26.05	18.08

En este caso la distribución de varianza acumulada obtenida refleja claramente la mala disposición que tienen los sensores, los cuales están concentrados hacia las esquinas de los polígonos que representan los sectores del campo. La zona Este del sector más grande registra la varianza acumulada más severa de tal forma que sería primordial para una próxima temporada poner sensores en tal región.

5.4.4. Campo Requínoa

Las estadísticas de la distribución de varianza acumulada junto con un gráfico que permite ver la distribución en terreno están a continuación:

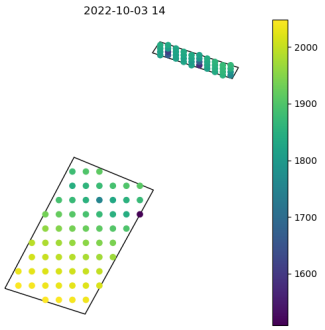


Figura 25: Requínoa [Distribución de varianza acumulada]

Tabla 20: Requínoa [Estadísticas de varianza acumulada]

	Sector 1 (Norte)	Sector 2 (Sur)
$\max_{\sigma^2_{ac}}$	1872.34	2048.33
$\min_{\sigma^2_{ac}}$	1570.75	1508.58
$\overline{\sigma^2_{ac}}$	1808.57	1940.97
$\sigma_{\sigma^2_{ac}}$	66.38	88.49

La zona sur del sector más grande registra una gran cantidad de varianza acumulada, lo que se debe a que este sector solo cuenta con un sensor en la zona norte. Para este campo al menos un sensor más en la zona sur de su sector más grande sería lo ideal para así bajar el nivel de incertidumbre que allí existe.

5.4.5. Campo Rengo

Las estadísticas de la distribución de varianza acumulada junto con un gráfico que permite ver la distribución en terreno están a continuación:

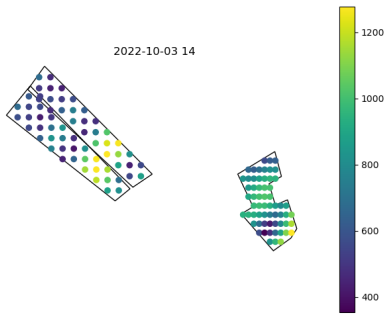


Figura 26: Rengo [Distribución de varianza acumulada]

Tabla 21: Rengo [Estadísticas de varianza acumulada]

	Sector 1 (Oeste-Sur)	Sector 2 (Este)	Sector 3 (Oeste-Norte)
$\max_{\sigma^2_{ac}}$	1272.72	1277.30	1277.46
$\min_{\sigma^2_{ac}}$	410.18	353.72	438.18
$\overline{\sigma^2_{ac}}$	735.28	840.45	705.520
$\sigma_{\sigma^2_{ac}}$	258.83	184.90	267.06

El campo de Rengo es el más privilegiado en cuanto a sensores activos lo que se puede notar claramente en la distribución de varianza acumulada, donde las zonas en las cuales esta es excesivamente alta son muy acotadas. Así, para el caso de este campo en particular los sensores existentes resultan ser suficientes debido a que mantienen la varianza estable a lo largo y ancho del terreno, situación que no se da en los demás campos.

5.5. Importancia de cada sensor

Para poner a prueba la precisión de las estimaciones y además medir la importancia de cada sensor, en busca de los sensores que más y menos aportan al proceso predictivo, se va a dejar un sensor sin usar con el objetivo de usar los dispositivos restantes para estimar y el sensor apartado para contrastar la diferencia entre la medición real y la predicha. Para este cometido se usará primeramente el campo de Rengo debido a que es el que más sensores posee, de sus 6 sensores 5 se usaran para estimar y el último para contrastar y calcular la diferencia entre la predicción y el dato registrado. Los resultados obtenidos se muestran en una tabla a continuación:

Tabla 22: Rengo [Predicciones vs Mediciones Tº]

Sensor	$\overline{\Delta t}$	$\sigma^2_{\Delta t}$	$min(\Delta t)$	$max(\Delta t)$
A84041E8E184C530	0.6345	0.9201	0.0	6.10
A84041F61184C52E	0.9721	1.2990	0.0	9.61
A84041883184C52D	1.0217	1.7651	0.0	15.79
A840415C8184C534	0.8236	1.1640	0.0	9.94
A840414E6184C543	0.6952	1.1300	0.0	13.08
A84041A15184C538	0.6938	1.2111	0.0	10.41

No está de más recalcar que Δt representa la diferencia (en valor absoluto) entre la temperatura registrada y la predicha, por ende $\overline{\Delta t}$ representa la media de las diferencias registradas, $\sigma^2_{\Delta t}$ la desviación estándar de estas diferencias y por último $min(\Delta t)$ y $max(\Delta t)$ las diferencias mínimas y máximas registradas.

Como se puede ver el sensor terminado en “C530” registra una media y desviación estándar de las diferencias entre los valores reales y los predichos (por los sensores restantes) bastante baja, por lo cual si fuera necesario remover un sensor del campo por algún motivo este sería el sensor ideal para hacerlo debido al bajo impacto en la precisión del modelo que generaría. Por otro lado el sensor terminado en “C52D” registra una media y desviación estándar de las diferencias entre los valores reales y los predichos bastante alta con respecto a los demás, en este caso este dispositivo sería la última opción a la hora de sacar un sensor debido al alto impacto en las precisiones de las predicciones del modelo que generaría su ausencia.

El mismo análisis es válido y extrapolable para todos los demás sensores de cada uno de los campos. Siguiendo la misma nomenclatura de colores que en el caso anterior, de color verde están los sensores cuyas medias y desviaciones estándar de las diferencias entre las predicciones y los datos son menores con respecto a los demás dispositivos, y en color rojo los sensores cuyas medias y desviaciones estándar de las diferencias entre las predicciones y los datos son mayores con respecto a los demás dispositivos. En caso de que ningún sensor posea la media y desviación estándar más alta/baja (de forma simultánea) el dato a dirimir la decisión del sensor a sacar o mantener será la media, aunque no hay que preocuparse demasiado ya que en la mayoría de los casos la media y desviación estándar más alta (o más baja) se dan de forma simultánea.

Las tablas con los resultados para los campos de CEAF, Graneros, Peumo y Requínoa se listan

a continuación:

Tabla 23: CEAF [Predicciones vs Mediciones Tº]

Sensor	$\overline{\Delta t}$	$\sigma^2_{\Delta t}$	$min(\Delta t)$	$max(\Delta t)$
A840415C6184C536	0.6639	1.0330	0.0	12.04
A84041885184C53C	0.7618	1.1325	0.0	7.56
A84041D9A184C545	0.7367	1.1737	0.0	13.14

Tabla 24: Graneros [Predicciones vs Mediciones Tº]

Sensor	$\overline{\Delta t}$	$\sigma^2_{\Delta t}$	$min(\Delta t)$	$max(\Delta t)$
A840413C3184C535	0.9135	1.3644	0.0	8.54
A84041452184C52F	0.5842	0.6889	0.0	5.80
A8404158A184C540	1.2992	1.9345	0.0	12.79
A84041A97184C53E	0.8128	0.8266	0.0	7.39
A84041B53184C532	0.8122	0.8266	0.0	7.39

Tabla 25: Peumo [Predicciones vs Mediciones Tº]

Sensor	$\overline{\Delta t}$	$\sigma^2_{\Delta t}$	$min(\Delta t)$	$max(\Delta t)$
A84041447184C53A	0.4600	0.6654	0.0	13.10
A84041868184C542	0.5291	0.7349	0.0	6.55
A84041E12184C53D	0.4510	0.6430	0.0	6.55

Tabla 26: Requínoa [Predicciones vs Mediciones Tº]

Sensor	$\overline{\Delta t}$	$\sigma^2_{\Delta t}$	$min(\Delta t)$	$max(\Delta t)$
A840411B3184C544	1.2113	1.5174	0.0	9.10
A840414B6184C539	0.7937	1.0887	0.0	6.75
A84041BA5184C541	1.0734	1.5861	0.0	9.19

6. Conclusiones

Los indicadores agronómicos mediante los cuales los agricultores de la sexta región se guían en la toma de decisiones del proceso productivo provienen fundamentalmente de la red de estaciones meteorológicas instaladas a lo largo del territorio. Estas estaciones meteorológicas las cuales se pueden encontrar a varios kilómetros de distancia de los campos solo dan visión una general sobre la situación dentro de las comunas, así, su capacidad de medir el comportamiento específico al interior de los campos es baja, además de dar nula información sobre cómo los indicadores se distribuyen en terreno.

Debido a las limitaciones que presenta el sistema de estaciones meteorológicas la puesta en marcha de sensores dentro de los campos mismos para la medición de indicadores agronómicos se ha levantado como un enfoque pionero en la sexta región. Así, los objetivos de esta memoria se centraron en el uso de los datos reportados por sensores dispuestos en 5 campos de la región de O'Higgins para dar información relevante con respecto a la distribución de hora frío acumuladas dentro de estos. Para los agricultores conocer la distribución de horas frío acumuladas dentro de su campo significa valiosa información con respecto a la cronología en que los distintos sectores del terreno van a florecer, además de poder tener estimaciones diarias con respecto al porcentaje de árboles con sus horas frío listas y poder hacer una estimación de la evolución de este para los próximos días.

Los tres primeros objetivos específicos planteados tenían que ver con lo relativo a la exploración, tratamiento y procesamiento de los datos recolectados por los sensores. Primeramente se exploró y encontró grandes diferencias en la velocidad de acumulación de horas frío entre los campos, este fenómeno no solo se dio entre los campos sino que también entre los distintos sectores de un mismo campo. Así, los campos de CEAF y Graneros acumularon sustancialmente muchas más horas frío que los campos de Peumo, Requínoa y Rengo, las razones de estas diferencias radican en la geografía y clima de cada sector, pero en especial en el tratamiento específico que recibe cada campo. Luego se procedió a identificar sensores con mal desempeño en su funcionamiento ya sea porque porque sufren prolongadas o frecuentes caídas, en el caso de sensores que tenían muy pocos registros en relación con los otros dispositivos de su mismo campo se decidió por simplemente marginarlos. Por último se probó el desempeño del modelo SARIMA para rellenar gaps de datos en las secuencias de registros de los sensores, este mostró una rendimiento inferior a estaciones meteorológicas cercanas pero aún así aceptable para gaps pequeños de tiempo.

Los últimos tres objetivos específicos tuvieron que ver netamente con el uso de procesos Gaussianos para obtener la distribución de distintos indicadores dentro de los campos. Se generaron simulaciones de las variaciones de temperatura en el tiempo dentro de los distintos campos para luego obtener la distribución de horas frío acumuladas dentro de estos. En complemento a los resultados anteriores se pudo establecer los sensores más relevantes al momento de entrenar los modelos y los sectores óptimos para la colocación de nuevos dispositivos, esto gracias a que los procesos Gaussianos por su propia naturaleza entregan la varianza o nivel de incertidumbre para cada una de sus predicciones.

En nuestro país las horas frío se miden típicamente entre el 1 de mayo y el 31 de julio de todos los años pero debido a que los datos disponibles solo estaban a partir de finales de julio se tuvo que trabajar en base a escenarios hipotéticos con umbrales de horas frío hipotéticos. A pesar de lo anterior el objetivo de esta memoria era mostrar la aplicación de una técnica de Machine Learning como lo son los procesos Gaussianos para datos provenientes de sensores colados dentro de campos de cerezos en la región de O'Higgins con el objetivo de generar información valiosa para un usuario final como lo podrían ser los agricultores. De esta forma este trabajo puede ser tomado como una prueba de concepto (proof of concept) de lo que en un futuro con los recursos y el desarrollo necesario podría ser un sistema integrado de sensores que reportara información valiosa como esta a agricultores de la región usando datos provenientes de sus propios campos.

Bibliografía

[Fay01] Usama Fayyad. «Knowledge Discovery in Databases: An Overview». En: *Relational Data Mining*. Ed. por Saso Dzeroski y Nada Lavrac. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, págs. 28-47. ISBN: 978-3-662-04599-2. DOI: 10.1007/978-3-662-04599-2_2. URL: https://doi.org/10.1007/978-3-662-04599-2_2.

[Pré01] András Prékopa. «On the concavity of multivariate probability distribution functions». En: *Operations Research Letters* 29.1 (2001), págs. 1-4. ISSN: 0167-6377. DOI: https://doi.org/10.1016/S0167-6377(01)00070-0. URL: https://www.sciencedirect.com/science/article/pii/S0167637701000700.

[Ras04] Carl Edward Rasmussen. «Gaussian Processes in Machine Learning». En: *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures*. Ed. por Olivier Bousquet, Ulrike von Luxburg y Gunnar Rätsch. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, págs. 63-71. ISBN: 978-3-540-28650-9. DOI: 10.1007/978-3-540-28650-9_4. URL: https://doi.org/10.1007/978-3-540-28650-9_4.

[Mur12] Kevin P. Murphy. *Machine learning : a probabilistic perspective*. Adaptive computation and machine learning series. Cambridge, MA: MIT, 2012. ISBN: 978-0-262-01802-9. URL: https://www.worldcat.org/title/machine-learning-a-probabilistic-perspective/oclc/781277861?referer=br&ht=edition.

[SSK18] Eric Schulz, Maarten Speekenbrink y Andreas Krause. «A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions». En: *Journal of Mathematical Psychology* 85 (2018), págs. 1-16. ISSN: 0022-2496. DOI: https://doi.org/10.1016/j.jmp.2018.03.001. URL: https://www.sciencedirect.com/science/article/pii/S0022249617302158.

[CN19] Joseph E. Cavanaugh y Andrew A. Neath. «The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements». En: *WIREs Computational Statistics* 11.3 (2019), e1460. DOI: https://doi.org/10.1002/wics.1460. eprint: https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/wics.1460. URL: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wics.1460.

[CPG21] Tao Cui, Dan Pagendam y Mat Gilfedder. «Gaussian process machine learning and Kriging for groundwater salinity interpolation». En: *Environmental Modelling & Software* 144 (2021), pág. 105170. ISSN: 1364-8152. DOI: https://doi.org/10.1016/j.envsoft.2021.105170. URL: https://www.sciencedirect.com/science/article/pii/S1364815221002139.

[Esp21] Patricia Larrañaga Espínola. *CATASTRO FRUTÍCOLA. PRINCIPALES RESULTADOS*. CIREN. 2021. URL: https://bibliotecadigital.odepa.gob.cl/bitstream/handle/20.500.12650/71122/Ohiggins202109.pdf.

[Olm21] Dra. Michelle Morales Olmedo. *Requerimiento de frío en cerezos y cambio climático*. CEAF. 2021. URL: https://bibliotecadigital.odepa.gob.cl/bitstream/handle/20.500.12650/71122/Ohiggins202109.pdf.

[Red21] Redagrícola. *Experiencia con Regina en distintas zonas productivas*. 2021. URL: https://web.archive.org/web/20210511180527/https://www.redagricola.com/cl/experiencia-con-regina-en-distintas-zonas-productivas.

[Uni22] The Pennsylvania State University. *STAT 510 | Applied Time Series Analysis. 1.1 overview of time series characteristics*. 2022. URL: https://online.stat.psu.edu/stat510/lesson/1/1.1 (visitado 30-11-2022).

Apéndice A. Kernels

A.1. Kernel Constante

La expresión para el Kernel constante está dada por

$$K(x, x') = C$$

Donde C es un real positivo, generalmente se usa para sumarlo a otro Kernel modificando así la media del proceso Gaussiano o para multiplicarlo a otro Kernel con el fin de escalar su magnitud.

A.2. Kernel Lineal

La expresión para el Kernel lineal está dada por

$$K(x, x') = \sigma_f^2 (x - c)(x' - c)$$

El desplazamiento c determina la coordenada x del punto por el que pasan todas las líneas de la distribución posterior, en este punto la función tendrá varianza cero (a menos que agregue ruido). Además la varianza de salida σ_f^2 determina la distancia promedio de la función a su media (Cada Kernel tiene este parámetro al frente; es solo un factor de escala). Cabe mencionar que este Kernel es no estacionario ya que posee un producto punto en su expresión.

A.3. Kernel Periódico

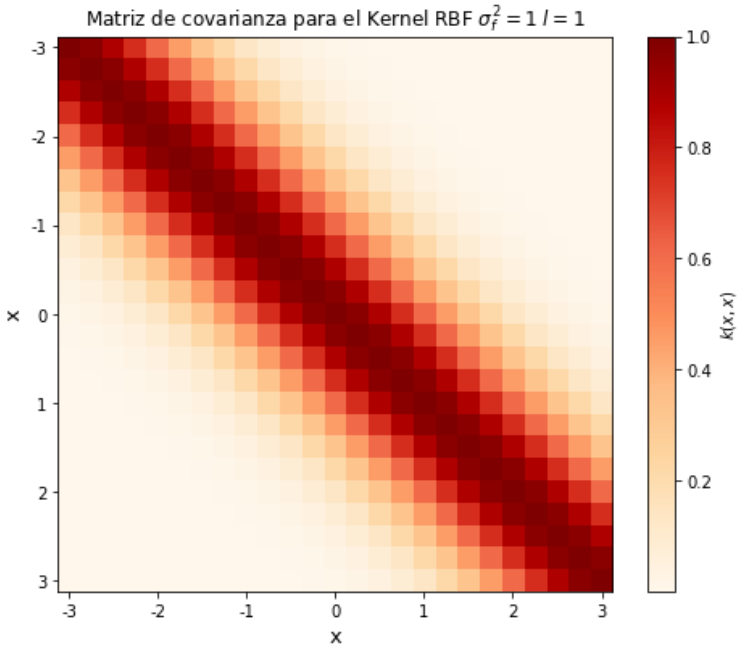
La expresión para el Kernel Periódico está dada por

$$K(x, x') = \sigma_f^2 \exp\left(-\frac{2}{l^2} \sin^2 \left(\pi \frac{||x - x'||}{p} \right) \right)$$

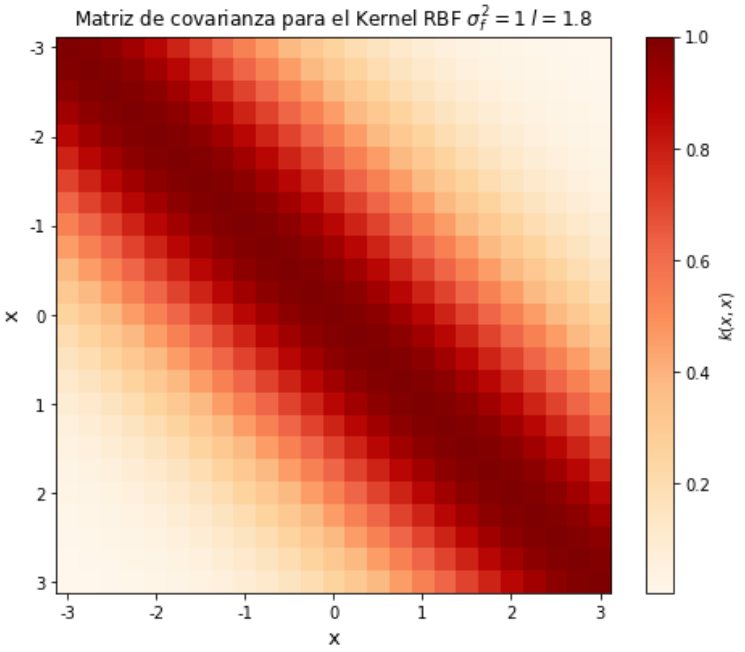
Este Kernel permite modelar funciones que se repiten así mismas, sus hiperparámetros son un *lengthscale* representado por l y una periodicidad representada por p.

Apéndice B. Matrices

B.1. Matriz de covarianza 25×25 Kernel RBF $\sigma_f^2 = 1, l = 1$

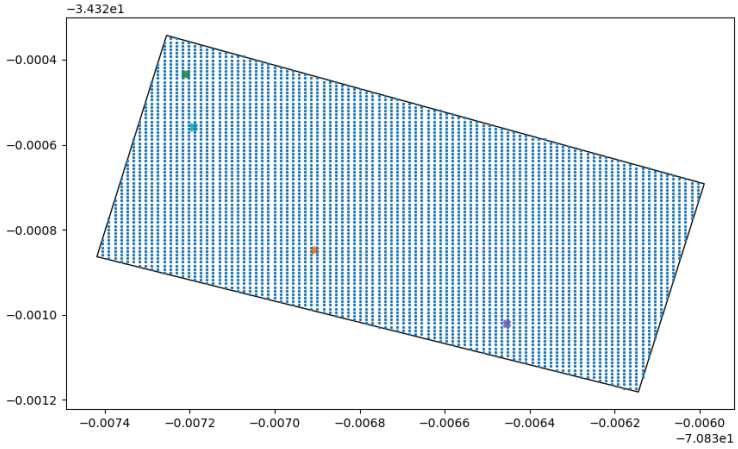


B.2. Matriz de covarianza 25×25 Kernel RBF $\sigma_f^2 = 1, l = 1.8$

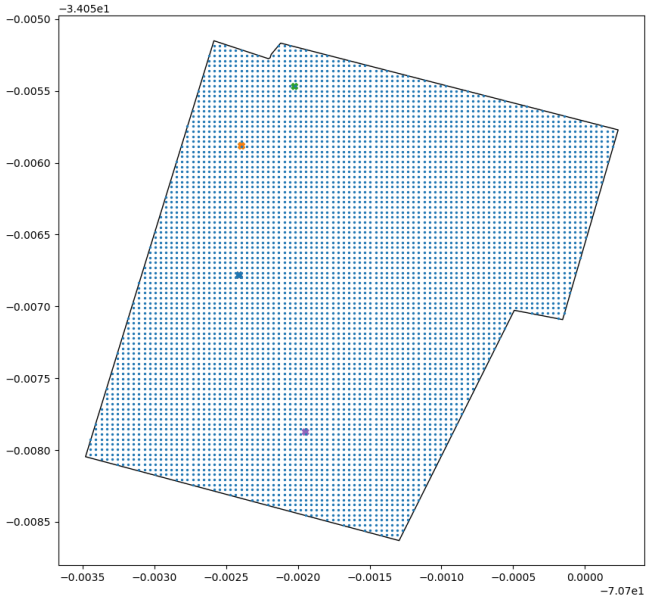


Apéndice C. Grillas

C.1. Grilla sobre el campo CEAF



C.2. Grilla sobre el campo Graneros



Apéndice D. Estimación de árboles con su meta de horas frío completada

A continuación se listan distintas tablas con las estimaciones a 3 días del porcentaje de árboles con su meta de horas frío completada para los campos de Peumo, Requínoa y Rengo. Se simulan distintos escenarios con diferentes umbrales de horas frío.

D.1. Peumo

Tabla D.1: Peumo Sector 1 [Estimación a 3 días del porcentaje de árboles despiertos según umbral HF]

Umbral Horas Frío	Porcentaje de árboles despiertos	Estimación 1 día	Estimación 2 días	Estimación 3 días
370	99.99%	100%	100%	100%
371	99.75%	99.87%	99.93%	99.97%
372	96.5%	97.78%	98.64%	99.19%
373	79.44%	84.61%	88.84%	92.16%
374	43.32%	51.18%	59.01%	66.49%
375	12.33%	16.84%	22.29%	28.63%
376	1.58%	2.56%	3.98%	6%

Tabla D.2: Peumo Sector 2 [Estimación a 3 días del porcentaje de árboles despiertos según umbral HF]

Umbral Horas Frío	Porcentaje de árboles despiertos	Estimación 1 día	Estimación 2 días	Estimación 3 días
370	100%	100%	100%	100%
371	99.73%	99.98%	100%	100%
372	84.13%	95.68%	99.24%	99.92%
373	21.60%	47.15%	73.98%	91.26%
374	0.51%	3.16%	12.65%	33.41%
375	0%	0.01%	0.17%	1.34%
376	0%	0%	0%	0%

D.2. Requínoa

Tabla D.3: Requínoa Sector 1 [Estimación a 3 días del porcentaje de árboles despiertos según umbral HF]

Umbral Horas Frío	Porcentaje de árboles despiertos	Estimación 1 día	Estimación 2 días	Estimación 3 días
598	100%	100%	100%	100%
600	99.99%	100%	100%	100%
602	99.61%	99.91%	99.98%	100%
604	94.58%	98.12%	99.47%	99.88%
606	70.97%	84.76%	93.32%	97.58%
608	30.85%	48.95%	67.27%	82.15%
610	6.03%	14.03%	27.25%	44.77%
612	0.46%	1.65%	4.87%	11.82%

Tabla D.4: Requínoa Sector 2 [Estimación a 3 días del porcentaje de árboles despiertos según umbral HF]

Umbral Horas Frío	Porcentaje de árboles despiertos	Estimación 1 día	Estimación 2 días	Estimación 3 días
598	98.43%	99.65%	99.94%	99.99%
600	63.73%	81.38%	92.40%	97.58%
602	7.35%	18.14%	35.59%	56.80%
604	0.06%	0.33%	1.50%	5.15%
606	0%	0%	0%	0.03%
608	0%	0%	0%	0%
610	0%	0%	0%	0%
612	0%	0%	0%	0%

D.3. Rengo

Tabla D.5: Rengo Sector 1 [Estimación a 3 días del porcentaje de árboles despiertos según umbral HF]

Umbral Horas Frío	Porcentaje de árboles despiertos	Estimación 1 día	Estimación 2 días	Estimación 3 días
320	97.73%	97.82%	97.91%	97.99%
325	94.38%	94.57%	94.75%	94.92%
330	87.98%	88.31%	88.63%	88.94%
335	77.64%	78.13%	78.61%	79.09%
340	63.53%	64.15%	64.77%	65.38%
345	47.29%	47.95%	48.61%	49.27%
350	31.50%	32.09%	32.68%	33.28%
355	18.52%	18.97%	19.42%	19.88%
360	9.52%	9.80%	10.09%	10.38%
365	4.24%	4.39%	4.55%	4.71%
370	1.63%	1.70%	1.77%	1.84%

Tabla D.6: Rengo Sector 2 [Estimación a 3 días del porcentaje de árboles despiertos según umbral HF]

Umbral Horas Frío	Porcentaje de árboles despiertos	Estimación 1 día	Estimación 2 días	Estimación 3 días
370	99.30%	99.61%	99.79%	99.89%
375	88.30%	91.81%	94.46%	96.39%
380	46.97%	55.04%	62.90%	70.25%
385	8.98%	12.73%	17.45%	23.14%
390	0.46%	0.81%	1.38%	2.28%

Tabla D.7: Rengo Sector 3 [Estimación a 3 días del porcentaje de árboles despiertos según umbral HF]

Umbral Horas Frío	Porcentaje de árboles despiertos	Estimación 1 día	Estimación 2 días	Estimación 3 días
320	94.25%	94.53%	94.81%	95.07%
325	89.65%	90.09%	90.52%	90.94%
330	82.82%	83.46%	84.07%	84.68%
335	73.65%	74.47%	75.27%	76.06%
340	62.48%	63.43%	64.38%	65.31%
345	50.15%	51.15%	52.16%	53.16%
350	37.80%	38.76%	39.73%	40.70%
355	26.59%	27.43%	28.27%	29.13%
360	17.37%	18.02%	18.69%	19.38%
365	10.49%	10.95%	11.43%	11.93%
370	5.84%	6.14%	6.45%	6.77%