# Estadística Espacial

### Parte práctica

# Florencia Ludueña, Paula Luvini, Facundo Marconi Mayo 2021

### Introducción

En el siguiente trabajo buscamos realizar un análisis exhaustivo de datos de temperatura y humedad provenientes del portal de datos del Servicio Meteorológico Nacional (SMN)<sup>1</sup>. Bajo el prior de que temperaturas más templadas y húmedas son propicias para la producción agropecuaria de cultivos tales como la soja, maíz y trigo, quisimos ver de manera descriptiva si dicha relación se ve reflejado en las regiones cultivadas y en el precio por hectáreas de campo de distintas regiones de la República Argentina.

### 1 Análisis y procesamiento de datos

Comenzamos nuestro análisis utilizando la base estaciones meteorológicas que contiene información descriptiva de nombre, provincia, código y localización de cada estación. Utilizando la base departamentos vista en clase, un archivo con información geográfica que contienen colecciones de multipolígonos y nos brinda información con respecto al sistema de coordenadas de referencias utilizando WGS 84, analizamos la distribución de las estaciones a lo largo de las diferentes provincias y departamentos. Observamos una gran concentración de estaciones en la provincia de Buenos Aires, por lo que acotamos gráficamente al área metropolitana para poder visualizar las etiquetas claramente.

Contamos con información de temperatura, humedad relativa, velocidad del viento, nubosidad total y precipitaciones promedio mensual histórica que abarcan el período 1981-2020. Las estaciones que cuentan con esta información son en total 71 de las 123 que se encuentran actualmente registradas por el SMN.

Comenzamos con un análisis de datos de temperaturas. Disponemos, además, de un datasets de datos diarios del año 2020 y 2021, el cual contiene valores máximos y mínimos. Sacamos la mediana de la temperatura máxima y mínima anual. Sin embargo, consideramos, dado nuestro foco de interés, que los datos climáticos que abarquen una mayor cantidad de años serán más apropiados y menos ruidosos dado que cualquier año atípico es suavizado con el promedio total. A diferencia de los datos diarios, los datos históricos no disponen la información de la estación Antártida. Dado que el objetivo final es el de analizar la relación con precio de hectáreas de campo/producción agrícola, los valores

<sup>&</sup>lt;sup>1</sup>https://datosabiertos.aeroterra.com/

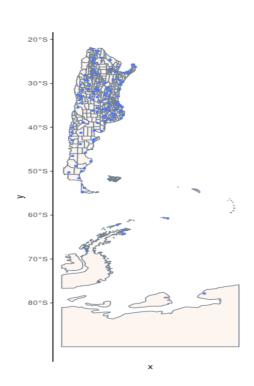


Figure 1: Estaciones meteorológicas

# 

Figure 2: Estaciones meteorológicas de PBA y AMBA

de la Antártida no son relevantes a nuestro problema, no afectan las conclusiones y nos genera ruido ya que corrompe la simetría de la distribución con valores extremadamente bajos de temperatura. Por este motivo decidimos sacar los datos de la Antártida.

Dado que los datos son de tipo data frame, para tratarlas fue necesario convertir las columnas de coordenadas a un objeto sf, con su respectivo crs. Por otro lado, transformamos los datos a la clase "geodata". Abajo y a la izquierda, el gráfico muestra los diferentes valores en función de la longitud. En esta se observa una marcada tendencia de cómo cambia el valor de norte-sur en línea con el gráfico anterior. Arriba y a la izquierda, se observan los datos en función de la latitud, no evidenciando ninguna tendencia marcada. Al graficar las temperaturas asignando una escala de colores según los valores, se observa un patrón muy claro en el que las temperaturas más bajas se encuentran en el sur, mientras que los valores más altos en el norte del país. Esto lo identificamos en el gráfico de la derecha.

Para realizar un análisis exploratorio análogo de los datos de humedad, también fue necesario transformar los datos. Observamos que, hay un comportamiento un poco menos marcado en términos Norte y Sur que el que había en el caso de la temperatura. Podemos ver que los casos de humedad más baja se encuentran tanto en la Patagonia como en algunas zonas cuyanas. Por otro lado, valores de humedad por encima del 70% pueden encontrarse tanto en Tierra del Fuego, Río Negro como en La Pampa o Jujuy. En todo caso, una escala de humeda se encuentra en un sentido más E-O.

Graficando los valores de humedad contra la latitud y longitud observamos que, a diferencia de la temperatura, no se observa una tendencia marcada.

Figure 3: Análisis exploratorio de la temperatura media

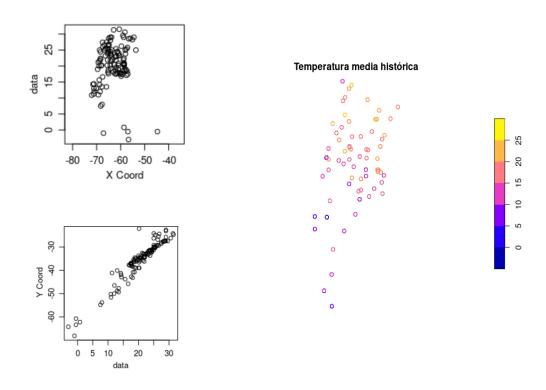
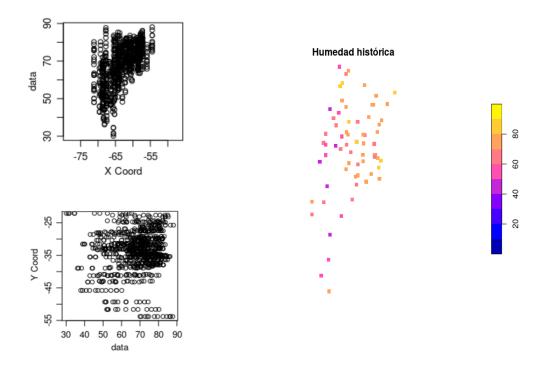


Figure 4: Análisis exploratorio de humedad



Dado que al momento de realizar Kriging el supuesto de normalidad es crucial, analizamos si se cumple. En primer lugar, en el Q-Q Plot de temperatura se observa colas

Figure 5: Análisis de normalidad

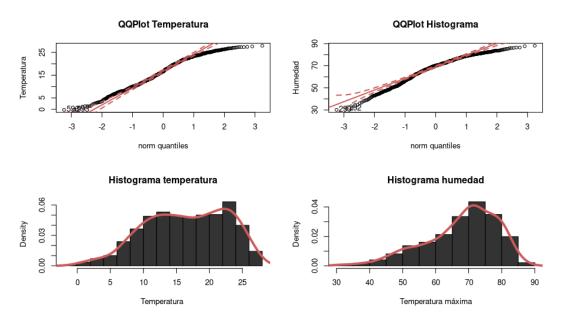


Figure 6: Histogramas y QQ-Plot de Temperatura y Humedad

pesadas, que sobrepasan los márgenes de normalidad. El histograma de valores muestra una distribución con una cola pesada hacia izquierda, debido a pocos valores cercanos al cero. Con respecto a la humedad. al igual que en el caso de la temperatura, los valores se encuentran concentrados en la cola derecha del histograma, por encima del 60% de humedad. Se observa también una cola pesada a izquierda en el histograma, y colas pesadas en el Q-Q Plot.

Para testear la normalidad, utilizamos el test de Shapiro-Wilk y el test de Kolmogorov-Smirnov. El segundo test compara si dos distribuciones son iguales, se procedió entonces a crear dos vectores normales aleatorios del largo de ambas bases con la media y desvío muestral respectivo. En ambos test, se parte de la hipótesis nula de que la muestra sigue una distribución normal. Al visualizar la tabla se observa que ambas presentan la evidencia suficiente para rechazar la hipótesis nula con un nivel de significación menor al 0.01 en el test de Shapiro Wilk, y menor al 0.05 en el de Kolmogorov-Smirnov, por lo que no se puede suponer la normalidad de los datos.

Si bien la normalidad es un requisito para varios tests y aplicaciones, si nos referimos al libro de Oliver y Webster encontramos lo siguiente: "Geostatistical analysis does not require data to follow a normal distribution. However, variograms comprise sequences of variances, and these can be unstable where data are strongly skewed and contain outliers. If your data do not have a near-normal distribution and have a skewness coefficient outside the limits  $\pm 1$ , because of a long tail, you should consider transforming them". <sup>2</sup>De

<sup>&</sup>lt;sup>2</sup>El análisis geoestadístico no requiere que la información siga una distribución normal. De todas manera, los variogramas son una secuencia de varianzas y las mismas pueden ser inestables cuando los datos están fuertemente sesgados y contienen outliets. Si tus datos no tienen una distribución cercana a la normal y tienen un sesgo a coeficientes por fuera de los límites +-1, por una cola pesada, deberías considerar transformarlos.

Table 1: Test de normalidad

Data	Test Shapiro - Wilk	$Kolmogorov ext{-}Smirnov$
Temperatura	W=0.9748	D = 0.0692
	p-value < 7.8e-12***	p-value = 0.036*
Humedad	W = 0.953	D = 0.0693
	p-value = 1.2e-16***	p-value = 0.03577*
Note:		·*** 0.001 ·** 0.01 ·* 0.05 ·. 0.1 · 1

esta manera, procedimos a dos tipos de transformaciones: normalizar las observaciones por un lado, y por otro, desplazar el dominio fuera del 0 y aplicar una log-transformación. Luego de ambas transformaciones, volver a testear normalidad pero siguió rechazando la hipótesis nula.

Table 2: Test de normalidad sobre transformaciones

Data	Test Shapiro - Wilk	Kolmogorov-Smirnov
Temperatura normalizada	W=0.9748 $p-value = 7.8e-12***$	D = 0.0585 $p-value = 0.1139$
	P (www. 1100.12	P cause office
Log- transformación	W = 0.793	D = 0.12172
Temperatura Temperatura	p-value = 2.2e-16***	p-value = 8.12e-5***
Humedad $Normal$ -	W=0.95327	D = 0.07177
izada	p-value = 1.23e-15	p-value = 0.026*
Log-	W = 0.9047	D = 0.10048
$trans formaci\'on \ Humedad$	p-value = 2.2e-16***	p-value = 4.3e-4***
Note:		·*** 0.001 ·** 0.01 ·* 0.05 ·. 0.1 · 1

Al comparar los variogramas, tras haberlos escalado y al respetarse las mismas tendencias, decidimos continuar con los datos sin normalizar para poder mantener una fácil interpretación.

## 2 Autocorrelación Espacial

Habiendo realizado el análisis exploratorio lo primero que debemos hacer es confirmar que nuestros datos presentan algún tipo de autocorrelación entre sí que justifique llevar a cabo el análisis. Con tal fin llevamos a cabo el test Global de Moran tanto para la humedad como para temperatura. En ambos casos encontramos que los datos se encuentran espacialmente correlacionados y que hace sentido continuar con el análisis.

Table 3: Test de Autocorrelación Espacial

	Humedad	Temperatura	
Test de Moran	Moran I= 0.566956313	Moran I = $0.806799284$	
	p-value $< 3.691e$ - $16***$	p-value < 2.2e-16***	
Note:		·*** 0.001 ·** 0.01 ·* 0.05 ·. 0.1 · 1	

Por otro lado, revisamos el scatter plot del test de Moran para ambas variables. En el scatter plot podemos ver que la mayoría de nuestros puntos se encuentran en los cuadrantes superior derecho e inferior izquierdo. En ambos casos significa que tenemos puntos con autocorrelación espacial positiva, es decir que en el vecindario hay valores similares de humedad y temperatura. Cabe destacar que en el caso de humedad hay una mayor cantidad de puntos en los otros dos cuadrantes donde hay puntos registrados disímiles a sus vecinos. Esto hace sentido si pensamos que los valores de humedad tienen mayor dispersión que los de temperatura.

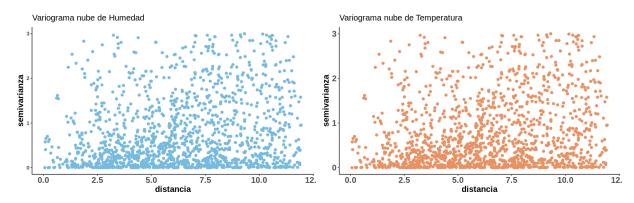
### 3 Variograma

En primer lugar vamos a observar el variograma nube. El mismo se obtiene al graficar todas las posibles diferencias al cuadrado de los pares de observaciones  $(Z(s_i) - Z(s_j))^2$  contra su vector de separación  $h_{ij}$ . El variograma nube es un gráfico bastante "ruidoso" pero que nos permite identificar con qué datos estamos trabajando en términos de distancia y de semivarianza. En este caso podemos ver cómo a distancias muy pequeñas hay pocos pares de puntos con gran semivarianza en los datos de temperatura, al contrario de lo que pasa con la humedad.

Podemos también mirar el variograma nube de Cressie-Hawkins, que en lugar de elevar las diferencias al cuadrado obtiene su raíz, cosa que lo hace más robusto a outliers:  $(Z(s_i) - Z(s_j))^{(1/2)}$ . Para ambos casos observamos que las tendencia y diferencias entre los datos de humedad y temperatura se mantienen.

El paso siguiente a nuestro análisis sería obtener el variograma empírico de nuestras observaciones. Uno de los requerimientos mínimos más importantes que teníamos a la

Figure 7: Variograma nube



hora de analizar el variograma era contar con isotropía. Un proceso es isotrópico cuando el vector h del que depende el variograma  $\gamma(\mathbf{h})$  sólo se ve afectado por su magnitud. En este caso donde estamos analizando variables climáticas de todo el país es importante determinar si alejarme de norte a sur o en estas direcciones va a afectar a mi análisis, es decir si la dirección del vector cambia el análisis.

La anisotropía puede ser analizada mirando el mapa direccional del variograma. El mapa nos muestra una grilla con vectores de separación h, centrados en el (0,0) y mostrados sobre x e y, lo que hace que veamos conjuntamente la distancia y la dirección. También podemos definir un variograma direccional para distintos ángulos como se ve en los gráficos a continuación. Esto no es más que considerar los ángulos de  $0^{\circ}$ ,  $45^{\circ}$ ,  $90^{\circ}$  y  $135^{\circ}$  respecto del eje x seteando este factor como parámetro en la función. En R lo único que debemos hacer es reemplazar el parámetro de rango con tres parámetros: el rango en la dirección de mayor autocorrelación, el del ángulo de la dirección principal y el ratio de anisotropía.

De esta manera, nuestros primeros variogramas resultan ser anisotrópicos. Para solucionar este problema podemos realizar alguna transformación logarítmica o realizar algún estudio de la tendencia, como indican Bivand et.al (2008). En el análisis exploratorio habíamos observado que en la temperatura media había una marcada tendencia respecto a la longitud. En la humedad no era tan claro pero había una débil tendencia respecto a la latitud. A continuación entonces procedimos a calcular nuestros variogramas con estas dos tendencias y realizamos el posterior análisis, comparando los variogramas direccionales y el mapa del variograma. Efectivamente, en los mapas de isotropía, cosa que se ve claramente en los mapas de variograma.

Es así que continuamos nuestra estimación con los variogramas con tendencia. Si queremos observar en detalle los variogramas direccionales con los que vamos a trabajar son también los que se encuentran a continuación.

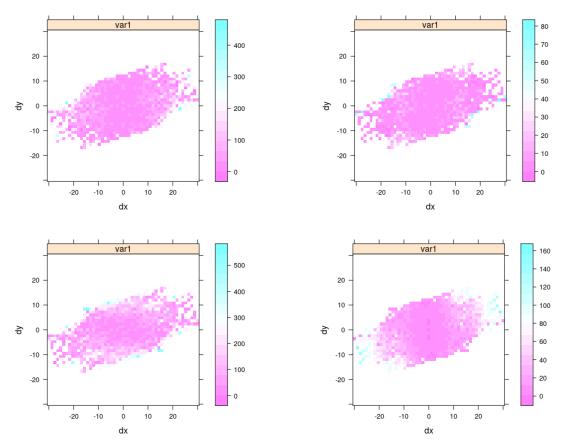
Una vez que obtuvimos los variogramas empíricos y el problema de la anisotropía fue solucionado debemos ajustar nuestros datos a un variograma teórico. La selección del mismo la realizamos a partir de una aproximación de la forma de la estimación no paramétrica del semivariograma empírico. Tenemos en cuenta el comportamiento del nugget, la meseta, el rango y de los factores de imprecisión.

En el caso de la humedad seleccionamos una función esférica con parámetros de meseta en 30, rango de 12, nugget de 4 y dirección en 45° para la humedad. Por otro lado,

Figure 8: Mapa de variogramas

Variograma Humedad con tendencia

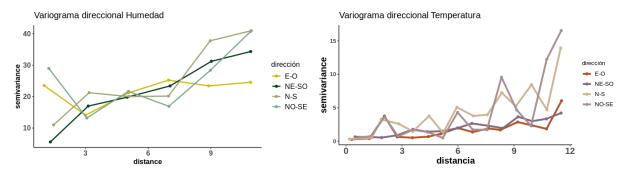
Variograma Temperatura con tendencia



Variograma Humedad sin tendencia

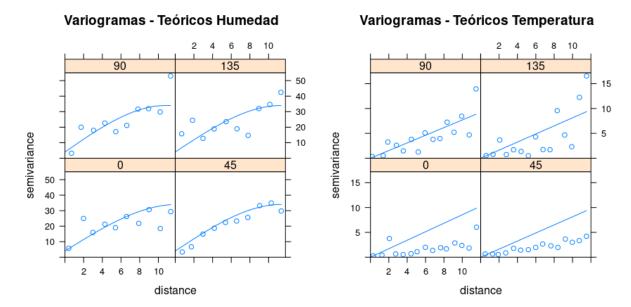
Variograma Temperatura sin tendencia

Figure 13: Variograma direccional



seleccionamos que los m intervalos disjuntos  $I_j = (h_j, h_{j+1})$  o intervalos de clase tengan un valor de 1.2. La tendencia agregada en este caso es respecto a la longitud. Para la temperatura la función seleccionada fue la exponencial, con una meseta de 5.5, rango de 9, nugget de 0.2 y tamaño del intervalo también de 1.2. Agregamos asimismo una tendencia respecto a la latitud de los datos. A continuación podemos observar los resultados para los variogramas direccionales de haber ajustado la humedad y la temperatura con los parámetros anteriores.

Figure 14: Variograma teórico



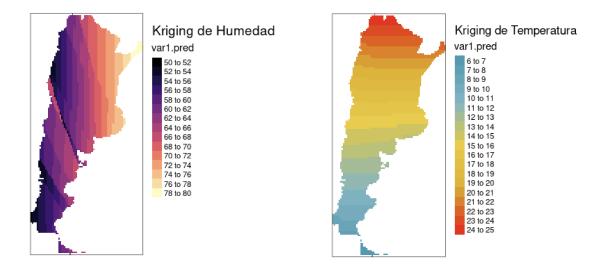
### 4 Kriging

Tras haber ajustado el variograma teórico el siguiente paso en el análisis es realizar el kriging que no es otra cosa que la herramienta geoestadística de predicción. Es un método de interpolación mediante el cual podemos conocer los valores de las variables en puntos donde no hay una muestra disponible. De esta forma podremos contar con los valores interpolados de humedad y de temperatura de todo el país, con la utilización de tan sólo 71 puntos correspondientes a las estaciones meteorológicas disponibles.

Desde ya que esta tarea no está exenta de complicaciones, especialmente por el hecho de que estamos intentando explicar el comportamiento meteorológico de todo el extenso territorio del país con menos de 100 coordenadas en el mapa. Es esperable entonces observar un considerable error o falta de precisión en la predicción final. Como dicen Oliver y Webster (2015): "When one solves the kriging equations in practice one usually finds that only points near to the target carry significant weight and that most can be disregarded. Kriging is seen to be a local predictor."<sup>3</sup>

Aún así y teniendo lo anterior en cuenta, vemos a continuación que las predicciones finales tanto de temperatura como de humedad, y contando con nuestro conocimiento acerca de los distintos climas de Argentina, son en líneas generales correctas. Por ejemplo, identificamos que las zonas con menor porcentaje de humedad corresponden a la Patagonia y al Noroeste. Por el contrario, la provincia de Misiones es la que mayor porcentaje de humedad presenta, con niveles en torno al 80%. Para el caso de la temperatura, vemos que el mapa muestra la transición entre las temperaturas más frías del país en Tierra del Fuego y el sur de Santa Cruz, pasando por un centro templado y el norte del país con temperaturas medias del orden de los 24°.

<sup>&</sup>lt;sup>3</sup>Cuando uno resuelve las ecuaciones de kriging en la práctica suele encontrar que sólo los puntos cercanos al objetivo tienen peso significativo y que la mayoría pueden ser descartados. El kriging es visto como un predictor local.



### 5 Análisis y procesamiento de datos de agropecuarios

Una vez realizado el análisis y los pasos necesarios para realizar el Kriging de los datos de temperatura y humedad, realizamos un procedimiento análogo para los precios de campos. Disponemos de datos de producción y de observaciones de precios de campo por hectárea. <sup>4</sup>

Para una disciplina que surgió de la mano de la industria minera, las aplicaciones con las que las geoestadística puede ser de utilidad en la actualidad son varias. En este caso, para ayudarnos a identificar posibles impactos que aparecen a la hora de tomar decisiones en agricultura. Es así que nos preguntamos qué papel jugarían las condiciones climáticas de las distintas regiones del país a la hora de decidir producir ciertos cultivos en ciertos lugares. Desde un principio teníamos bien en claro que el clima no es el único factor ni el más importante en un sector agropecuario que depende de variables externas como el dólar, retenciones y precios internacionales, además claro de la calidad del suelo. De todas maneras, el análisis nos pareció muy relevante porque condiciones más o menos cálidas o más o menos húmedas determinan que ciertas regiones sean aptas para algunos cultivos. En el caso de la soja por ejemplo y según InfoAgro: "Las temperaturas óptimas para el desarrollo de la soja están comprendidas entre los 20 y 30º C, siendo las temperaturas próximas a  $30^{\circ}$  C las ideales para su desarrollo. El crecimiento vegetativo de la soja es pequeño o casi nulo en presencia de temperaturas próximas o inferiores a  $10^{\circ}$  C, quedando frenado por debajo de los  $4^{\circ}$  C. Sin embargo, es capaz de resistir heladas de -2 a - $4^{\circ}$  C sin morir. Temperaturas superiores a los  $40^{\circ}$  C provocan un efecto no deseado sobre la velocidad de crecimiento, causando daños en la floración y disminuyendo la capacidad de retención de legumbres." <sup>5</sup>.

Queda claro entonces que el clima es un factor más que relevante y a considerar para la distribución de los cultivos en el país. Por otro lado, considerando que es una variable que podría afectar más en los próximos años como consecuencia del cambio climático vale la pena realizar el consecuente análisis. Tal es el caso que la prevalencia de fuertes sequías o de fatales inundaciones pueden provocar la pérdida de campañas enteras. Esto desde ya se intensificaría aún más en los años venideros con desastrosas consecuencias para la economía.

<sup>&</sup>lt;sup>4</sup>Agrofy.com.ar

<sup>&</sup>lt;sup>5</sup>https://www.infoagro.com/herbaceos/industriales

En primer lugar debemos mencionar que los principales cultivos desde desde la década de 1980 a la actualidad estuvieron repartidos entre el maíz, el girasol, el trigo y la soja, con esta última dominando la escena predominantemente. Si tomamos en cuenta los datos de la Dirección de Estimaciones Agrícolas<sup>6</sup> podemos ver cómo se distribuyó entre provincias esta producción en la última campaña de 2019/20, expresadas en toneladas.

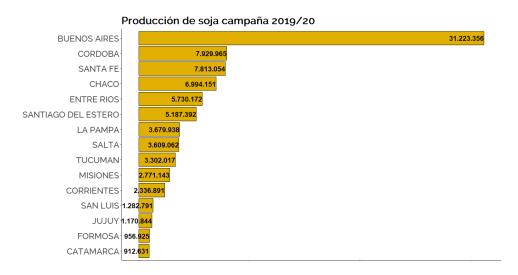


Figure 15: Producción de soja por provincia

Es así que las provincias de Buenos Aires, Córdoba y Santa Fe llevan la delantera en términos de producción total de soja, seguidas muy de cerca por Chaco. En estas primeras tres provincias podemos comparar los mapas de kriging de temperatura y de humedad y observar que tienen una temperatura media templada, de entre 15º y 20º, así como humedad en torno al 64% y 72%. En el caso de las siguientes provincias más importantes en producción podemos destacar a Chaco, Santiago del Estero y Entre Ríos. En las dos primeras idenfiticamos niveles de humedad similares a los de la zona núcleo aunque de temperaturas más altos. En el caso de Entre Ríos destacamos que la humedad es como mínimo mayor al 70%.

Otra variable muy relevante y distinta a la producción, que puede verse afectada más fácilmente por variables de coyuntura, es el precio de la tierra por hectárea. El precio del alquiler de la tierra va a ayudarnos de manera muy sencilla a poder ponderar algunas variables importantes como la calidad del suelo. Es así que en la zona núcleo de la soja conformada por el norte de la provincia de Buenos Aires y el sur de Santa Fe encontraremos los campos más caros en términos de USD/hectárea.

Realizando un análisis exploratorio de nuestra base de datos podemos confirmar que contamos con observaciones de precios de campo representativos. Cuando miramos el histograma de precios podemos ver que los datos tienen una cola a derecha: la mayor parte de las observaciones son en cierto punto baratas pero contamos con un par de ellas de muy alto costo que se separan del resto marcadamente y están más dispersos.

Para ver de manera más clara cuáles son las zonas de mayor precio decidimos hacer un análisis de clusters utilizando como método *kmeans* sobre los precios y sobre el raster resultante del kriging de precios. Mirando el gráfico de *elbow* o codo vimos que los clusters identificados no deberían ser más de tres o cuatro. De esta manera decidimos pedirle al

<sup>&</sup>lt;sup>6</sup>Datos.gob.ar/agroindustria-agricultura

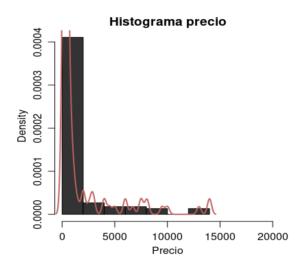


Figure 16: Precios por hectárea

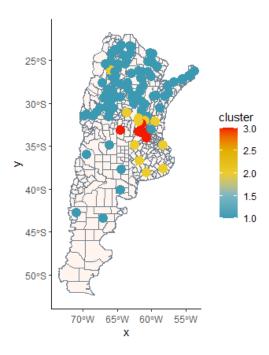
algoritmo que encuentre tres clusters en total y los graficamos en un mapa en la figura 17. Omitimos poner muchos datos de la zona patagónica o Antartida debido a que no es zona productiva de los cultivos tradicionales, que puede verse influenciado el precio por turismo o actividad petrolera, por lo que en el mapa tampoco graficamos Tierra del Fuego. Es así que identificamos tres grupos bastante marcados, la mayoría de los precios están agrupados en el grupo celeste que corresponde a los más baratos. El grupo amarillo está compuesto por campos productivos más caros concentrados en la provincia de Buenos Aires, sur de Santa Fe y Córdoba. Notablemente, el último grupo es el de la zona núcleo sojera, un sector espacialmente reducido ubicado en la zona norte de Buenos Aires y sur de Santa Fe.

Table 4: Centers de cluster de precios por hectárea

Celeste	Amarillo	Rojo
455.90	4,307.27	10,177.27

De esta manera, con estos datos y los mapas que obtuvimos del kriging vemos cuáles son los lugares más codiciados para tener cultivos tradicionales. Y también observamos que, pese a no tener la riqueza en el suelo que se observa en la zona pampeana, otras provincias como Chaco, Santiago del Estero y Entre Ríos también pueden tener una participación importante en la producción agropecuaria. Esto último no sería posible si no tuviesen las condiciones climáticas descriptas anteriormente. Es decir, si bien mediante precios podemos saber cuáles son los mejores suelos entendemos que esto no es lo único relevante a la hora de decidir tener ciertos cultivos y que la temperatura y humedad de la zona son más que relevantes.

Figure 17: Clusters de Precios por Hectárea



### 6 Conclusiones

En resumen, en este trabajo realizamos un análisis exhaustivo de dos variables climáticas muy importantes para la producción agrícola: la humedad y la temperatura. Mediante la utilización de datos históricos del SMN conseguimos estas variables para 71 estaciones meteorológicas de todo el país, que utilizamos para llevar a cabo el *kriging* y la consecuente predicción de estas dos variables en todo el territorio nacional. Con las salvedades ya mencionadas del caso nos parece que nuestra predicción es bastante buena y que el mapa final de temperaturas y humedad se asemeja bastante a la realidad climatológica del país.

Desde luego para llegar a este último resultado fueron muchos los obstáculos con los que nos topamos y que debimos resolver. La presencia de no-normalidad, de tendencia en nuestros datos, la anisotropía inicial, sin contar con los tests de autocorrelación global que realizamos y el final ajuste del variograma teórico.

Los resultados finales fueron a nuestro parecer satisfactorios y creemos que son muy relevantes a la hora de analizar la distribución de la producción de cultivos en el país. De esta manera, mirando tanto a la temperatura como a la humedad presente identificamos las zonas más propicias para la producción de soja, procedimiento que se puede llevar a cabo para cualquier otro cultivo. Lo relevante de estos mapas que conseguimos puede ser muy útil a la hora de analizar la introducción de nuevos cultivos en algunas áreas y tener la necesidad de con una fugaz revisión identificar posibles lugares. Además, con el advenimiento de mayores consecuencias del cambio climático esto será con los años más relevante. Poder realizar este mapa de temperaturas cada año será clave para predecir los impactos que temperaturas más extremas tendrán en los cultivos así como qué zonas serán las más afectadas.

Por último, complementamos este análisis incluyendo datos de precios de campo por hectárea para confirmar dónde se encuentra la zona núcleo de la soja y, como ya men-

cionamos, esta última hace mucho sentido por las condiciones climáticas presentes en la zona.

# Bibliografía

Banerjee, Sudipto, Bradley P Carlin, and Alan E Gelfand. *Hierarchical modeling and analysis for spatial data*. CRC press, 2014.

Bivand, Roger S et al. Applied spatial data analysis with R. Vol. 2. Springer, 2013.

Hengl, Tomislav. "A practical guide to geostatistical mapping". In: (2009).

Oliver, Margaret A and Richard Webster. "The Variogram and Modelling". In: *Basic Steps in Geostatistics: The Variogram and Kriging*. Springer, 2015, pp. 15–42.