Universidad de
SanAndrés

Estadística Espacial
Grupo Impares

Trabajo Práctico Final

Autores:
Andres Sarcuno
Antonella Schiavoni
Libia Billordo

Mayo 2021

1 Introducción

En este trabajo se aplicarán diferentes técnicas geoestadísticas con el objetivo de predecir el lugar más apropiado para la instalación de un parque eólico, basándonos en los dataset provistos por el servicio meteorológico nacional argentino al día 20 de Abril del 2021 que consiste de dos archivos con formato txt. Por un lado el dataset de horarios y por otro el de estaciones.

2 Parte Teórica

2.1 Explique qué se entiende por proceso estocástico espacial intrínseco

Un proceso estocástico es una variable aleatoria que depende de un parámetro, en el caso de estadística espacial, este parámetro representa el espacio y sus valores hacen referencia a localizaciones espaciales. Sea X(s) un proceso estocástico, siendo s la localización en el espacio. Para todo valor de s, existe una variable aleatoria, las cuales no necesariamente tienen que ser independientes. Se llama realización del proceso estocástico a la muestra obtenida de variables aleatorias X(s)

Existen procesos estocásticos que carecen de la existencia de varianza por mas que los incrementos tengan varianza finita. Sea X = s: s ϵE S un proceso estocástico no estacionario, se dice que es intrínsecamente estacionario si cumple con las siguientes condiciones:

- E [X(s + h) - X(s) = μ(h)
- V [X(s + h) - X(s)] = σs^2 (h)

En otras palabras, esto quiere decir que ni la esperanza ni la varianza de los incrementos son dependientes de la localización , sino solo del vector h que une los puntos. En el caso de que la media del proceso no fuera constante, correspondería que aplicar la siguiente transformación: $X^t(s+h) - X^t(s) - \mu(h) = X(s+h) - X(s)$ que sólo es función de h. Esta es la forma habitual de representar la estacionariedad intrínseca.

2.2 ¿Que mide un semivariograma?

El semivariograma sirve para medir la autocorrelación espacial o la dependencia espacial. Permite analizar el comportamiento espacial y la estructura de una variable en una área definida. En otras palabras, nos posibilita estimar cuán similares son dos puntos, a partir de medir la distancia entre ellos dos en un espacio determinado.

2.3 ¿En qué casos se utiliza el estimador Kriging ordinario y cuál es su objetivo?

Su objetivo es estimar el valor de una variable en un punto del espacio desconocido, es decir que no existen observaciones para este punto determinado en la muestra, a partir de la muestra de puntos del espacio conocidos. Es un método de interpolación geoestadístico que se aplica cuando se tiene un proceso estacionario, la media se asume desconocida y debe ser estimada, y la varianza es aleatoria.

2.4 ¿Qué se entiende por anisotropía en el análisis espacial?

Un semivariograma se define como anisotrópico si cambia en alguna forma respecto a la dirección que se considere. Si el semivariograma no solo depende de la longitud del vector sino también de la dirección del vector entonces el semivariograma es anisotrópico.

Se reconocen dos tipos de anisotropía: anisotropía geométrica y anisotropía zonal. Anisotropía geométrica ocurre cuando el rango del semivariograma cambia en las distintas direcciones, pero no la varianza sill, por lo tanto, la correlación es más fuerte en una dirección que en otra. Anisotropía zonal existe cuando la varianza estructural del semivariograma cambia con la dirección. Anisotropía geométrica significa que la correlación es más fuerte en una dirección que en otra.

3 Parte Práctica

3.1 Datos

Datos abiertos del Servicio Meteorológico Nacional

3.2 Problemática

Las energías renovables cobran cada vez mayor peso en las agendas políticas de los países a nivel mundial dadas las proyecciones de escasez de recursos energéticos en las próximas décadas y su menor impacto en la contaminación del medio ambiente. Igualmente las energías renovables, como la eólica, suele tener altos costos para bajos niveles de energía producida. Por ello es importante encontrar lugares geográficos en donde poder asegurarse que construir parques eólicos sean costo/efectivos. Los datos abiertos de Servicio Meteorológicos nos permiten observar los promedios de vientos en nuestro país en el espacio tiempo de un día. Dado que el requerimiento principal para que la construcción de un parque eólico sea exitosa es lograr un promedio diario de vientos mayor a 20km/h, buscaremos mediante un modelo de predicción geoestadístico cuáles son las zonas más propicias para crearlos en nuestro país.

3.3 Metodología

3.3.1 Preparación de los Datos

Al observar como estaba compuesto el dataset, convertimos la variable que representa la velocidad del viento en km/h (FF) ya que dicho valor era de tipo Character, y lo convertimos a un valor numérico. Además, borramos las observaciones con datos nulos en la variable viento (FF), y para no perder observaciones de otras variables que pudieran estar correlacionadas con el viento, suplantamos esos valores nulos por el promedio diario de los valores de su respectivo lugar geográfico. Es decir, en el caso de que una observación poseía valores del viento pero no de la humedad, se incluía el promedio diario de la humedad al dataset para las observaciones faltantes de una determinada estación de medición.

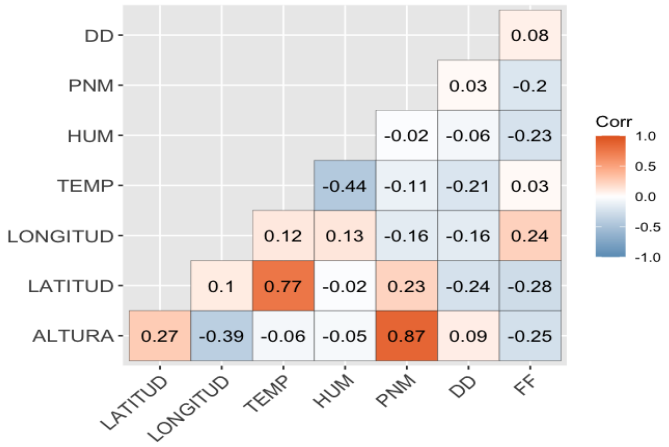
Por otro lado, en el dataset de estaciones, las latitudes y longitudes estaban expresadas en grados y minutos y las convertimos a valores decimales para poder utilizarlas en las librerías geográficas con las cuales íbamos a desarrollar los modelos.

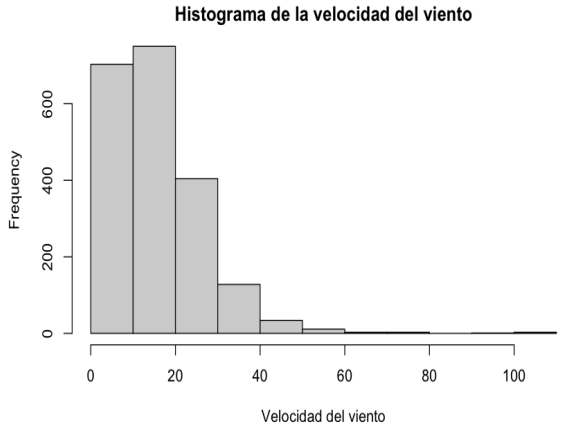
3.3.2 Estadística Descriptiva (clásica y espacial) de los Datos

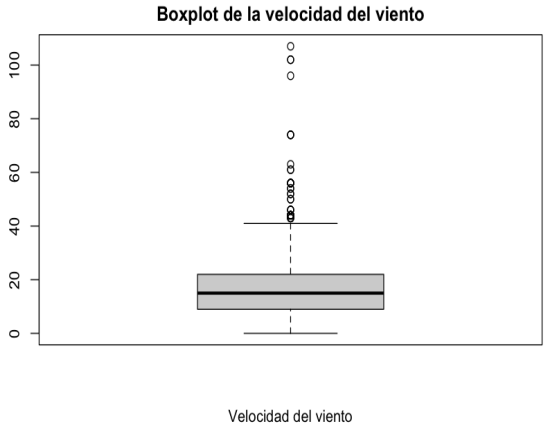
Visualizamos las distribuciones de las variables numéricas de interés mediante histogramas y boxplots. Es así que pudimos observar que la variable viento no tenía una distribución normal y mostraba una gran cantidad de outliers. Esta exploración visual, se confirma con la medida de asimetría (1.588087) y curtosis (7.188948). Nuestra variable de viento es asimétrica a la derecha y tiene una mayor concentración de valores muy cerca de la media de la distribución, y muy lejos de su cola.

Procedemos a analizar la correlación entre las variables, siendo HUMEDAD (negativamente) y HORA (positivamente) las que correlacionan con viento. También observamos que HORA y TEMPERATURA correlacionan negativamente con HUMEDAD. Por ultimo, se observa que HORA y TEMPERATURA correlacionan positivamente.

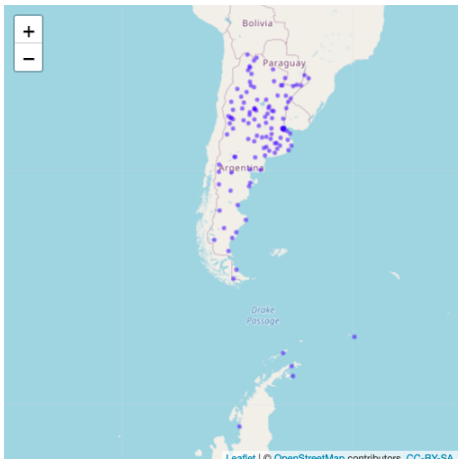
Matriz de Correlación







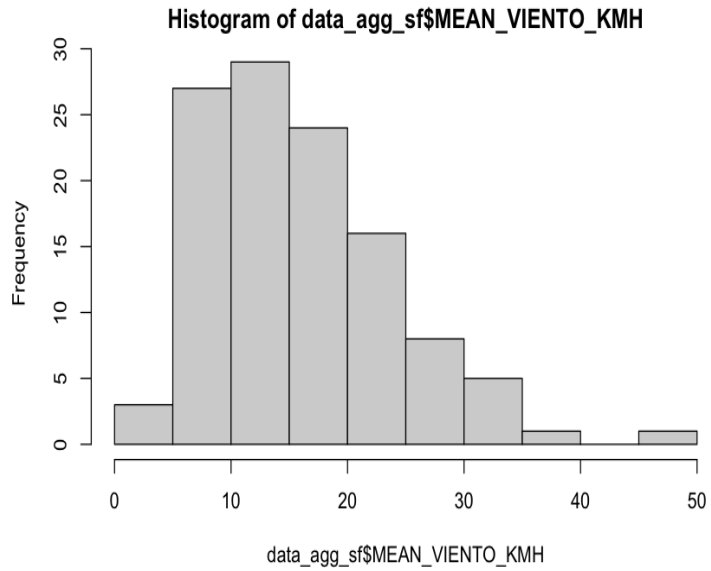
El boxplot arriba graficado pone en evidencia la existencia de outliers. ¿Pero son estos realmente outliers, o pertenecen a observaciones en lugares muy remotos? Esto lo analizaremos luego, al momento de graficar las estaciones en el mapa de Argentina.



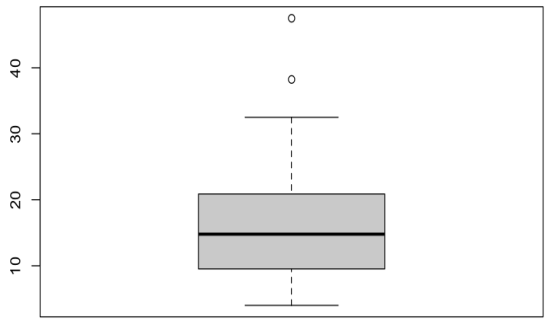
Para comenzar con el análisis exploratorio espacial representamos mediante puntos -en un mapa interactivo de Argentina- los lugares donde teníamos observaciones. Dado el propósito de este estudio, el cual es determinar la ubicación geográfica óptima en base a la variable velocidad del viento, decidimos remover observaciones fuera de Argentina continental, que no aportaban información útil y agregaban ruido a nuestro análisis. Es así que borramos las estaciones que no están en la plataforma continental argentina: - Base Carlini - Base San Martín - Base Marambio - Base Esperanza - Base Orcadas.

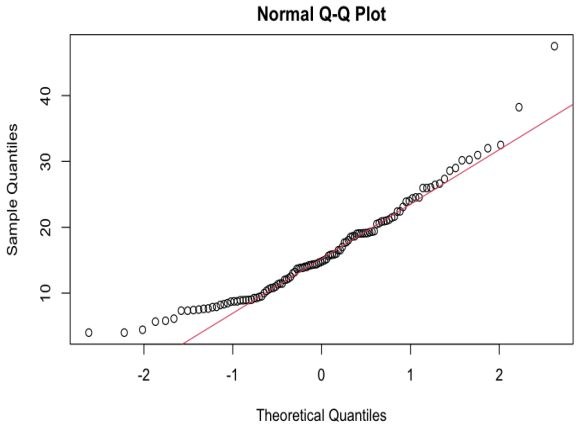
3.4 Análisis del supuesto de distribución normal de los datos

Realizamos el análisis estadístico del promedio diario de la velocidad del viento y de su desvío estándar durante un día de cada una de las estaciones continentales argentinas. De análisis de gráficos como histograma, boxplot, qqplot y el test de Shapiro (arroja un p-value menor a 0.05), observamos que nuestra variable promedio del viento no posee una distribución normal .

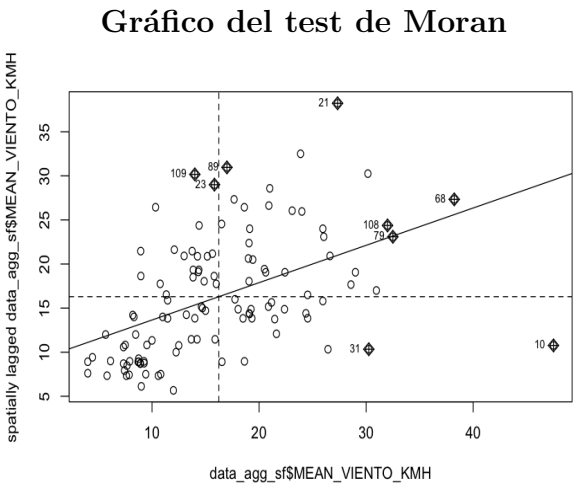


Boxplot



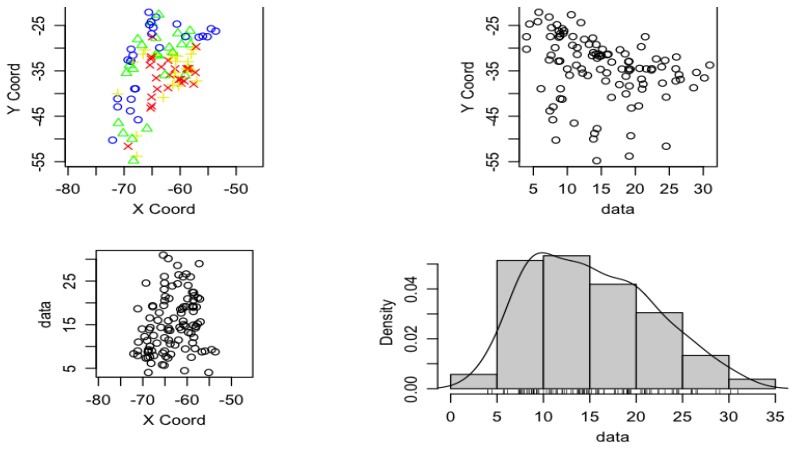


Procedemos a indagar sobre la posible existencia de inliers para -en el caso de encontrarlos- eliminarlos. Utilizamos el test de Moran de para corroborar si el promedio del viento esta o no distribuido de manera aleatoria. Corroboramos que no son normales, y borramos los inliers encontrados en su primera ronda. Realizamos dos rondas más de análisis de inliers pero sólo quitamos del dataset la primera ronda, dado que las otras tampoco afectaban la distribución de la variable promedio de tiempo pero si nos quitaban observaciones.



Igualmente tanto el coeficiente I de Moran(0.483) como el indicador de continuidad C de Geary (0.419) demuestran que los datos presentan una autocorrelacion positiva, que nos permite decir que tienden a una concentración espacial

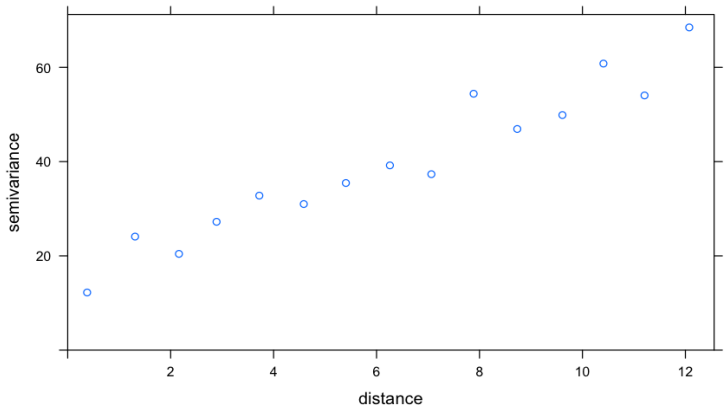
Luego de darle el formato adecuado al dataset, realizamos un análisis exploratorio sencillo para analizar la presencia de estructura y tendencia. Llegamos a la concluimos que pareciera existir una leve tendencia negativa a medida que aumenta los valores en el eje de la coordenada Y. No observamos una presencia de estructura fuerte en estos gráficos.



3.5 Análisis Estructural (Variograma)

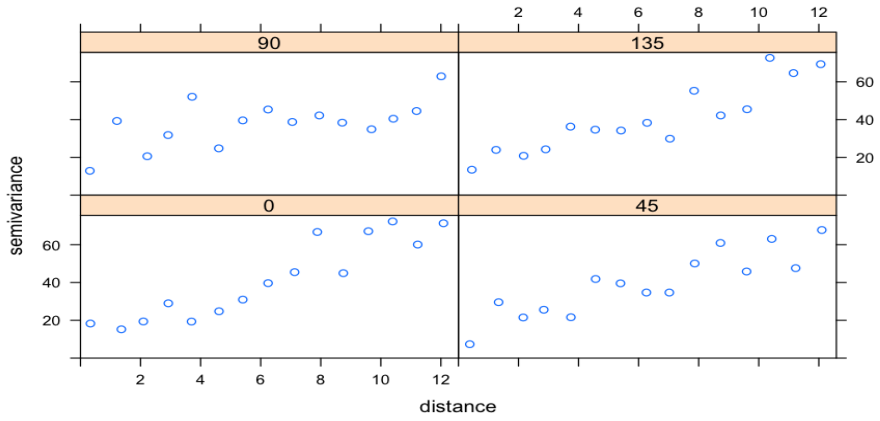
A partir de este momento, asumimos el supuesto de normalidad de los datos en nuestro dataset, así como también asumimos que estamos en presencia de un proceso estacionario. A continuación Para representar a nuestro variograma empírico, a continuación lo graficamos:

Variograma



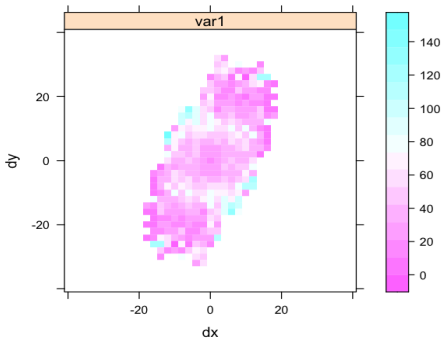
- Con el fin de analizar la isotropía o anisotropía, graficamos el variograma con diferentes grados de direccionalidad, pudiendo observarse la presencia de anisotropía en nuestro variograma dado que los gráficos fluctúan dependiendo de la dirección en la cual se los posiciona.

Variogramas con Direccionalidad



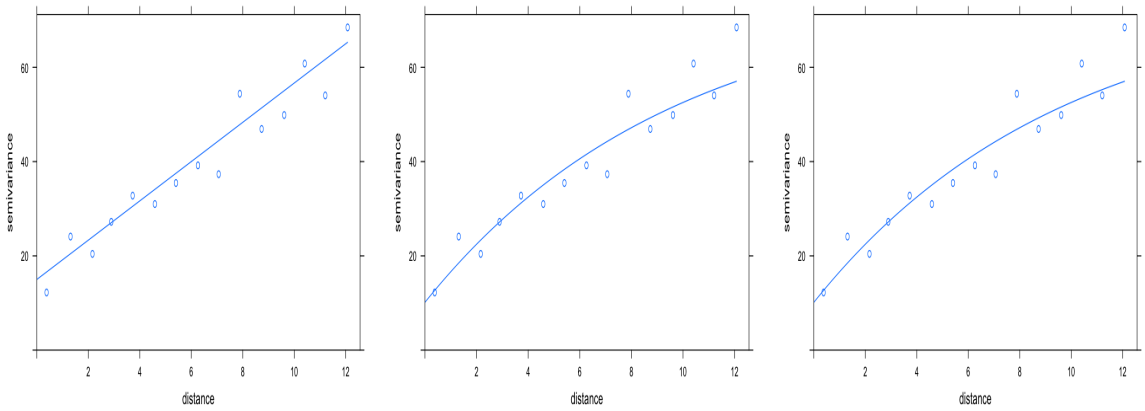
- Gracias al gráfico del variograma nube se puede apreciar una leve tendencia, concluyendo que tenemos una direccionalidad del sudoeste al noreste.

Variograma de Nube



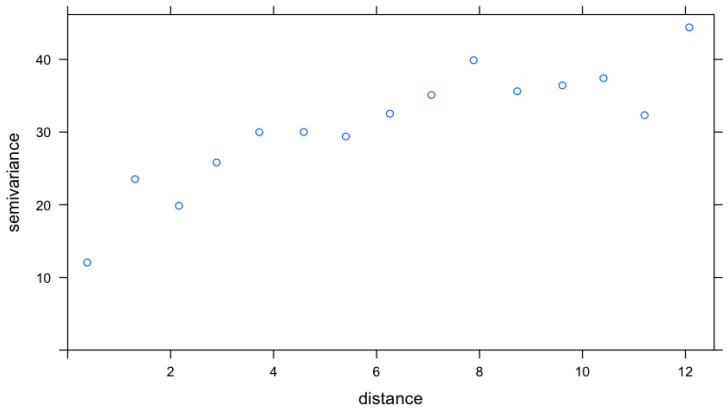
- Ajustamos nuestro modelo empírico a tres teóricos: el lineal, el exponencial, y el Matern, primero probando sin tendencia. Al calcular el error estándar de los residuales, dando el modelo teórico Matern el mejor resultado.

Ajuste variograma a modelos teóricos: lineal, exponencial y matern

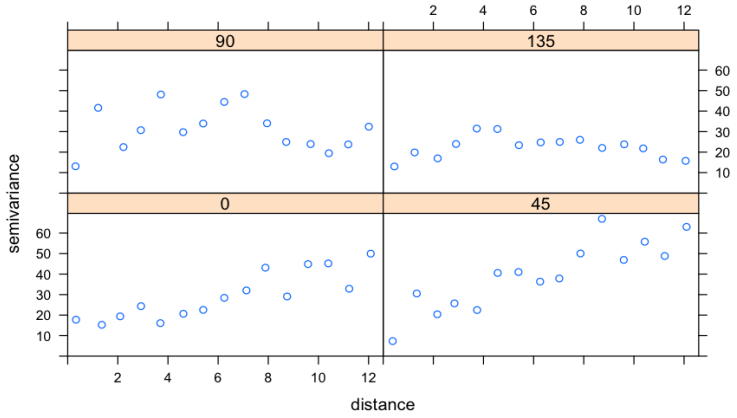


- Luego repetimos los gráficos del variograma pero con tendencia. Mirando el segundo gráfico, observamos una autocorrelación mas fuerte a los 90 y 45 grados, y una autocorrelación menos fuerte a los 135 y 0 grados. A partir del ultimo gráfico del variograma, concluimos que tenemos una direccionalidad del sudoeste al noreste.

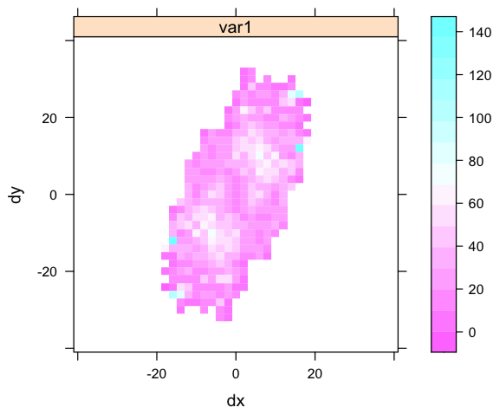
Variograma con tendencia



Variograma con tendencia con direccionalidad

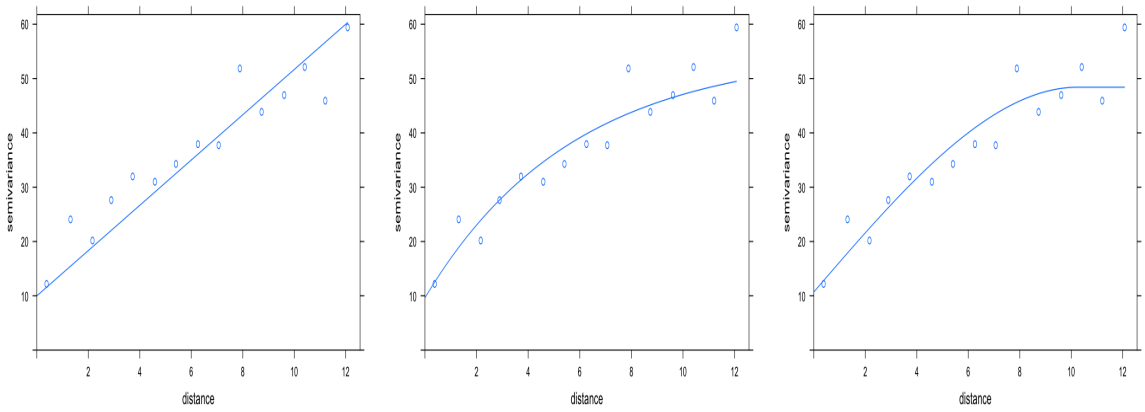


Variograma nube con tendencia



- Ajustamos nuestro variograma contra los siguientes modelos teóricos: lineal, exponencial y esférico, siendo el exponencial el que mejor ajusta según el análisis de residuos (7971.275, 3398.701, 4203.725).

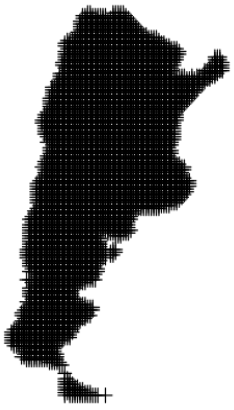
Ajuste variograma a modelos teóricos: lineal, exponencial y esférico



3.6 Predicción (Kriging)

- Comenzamos por armar la grilla que comprendía al territorio de Argentina continental.

Grilla del territorio argentino



- Creamos modelos de Kriging incluyendo 15 observaciones cercanas en base a los diferentes variogramas que presentamos anteriormente tanto sin tendencia como con tendencia.
- A continuación, para elegir cual de todos estos modelos tiene un mejor poder de predicción, utilizamos la técnica de cross validation.

3.7 Cross Validation

- Nuestro modelo de Cross Validation fue basado en 10 folds.

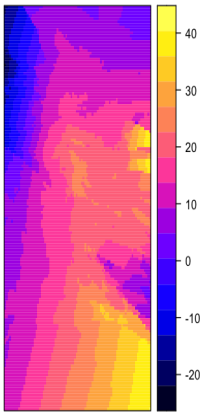
- Armamos una tabla de predicciones con los RMSE para ver cual de los modelos tienen mayor poder de predicción.

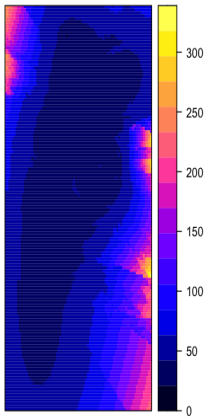
Ranking de modelos por poder de predicción

Modelo Kriging	RMSE
$\mathrm{v1}_m at$	4.980003
$\mathrm{v1}_e xp$	5.016706
$\mathrm{v2}_e xp$	5.086924
$\mathrm{v2}_s ph$	5.186301
$\mathrm{v0}_l inl$	5.201249
$\mathrm{v0}_s ph$	5.229554
$\mathrm{v0}_e xp$	5.277179

Nos quedamos con el modelo que tiene mejores niveles de error que es el matern con tendencia

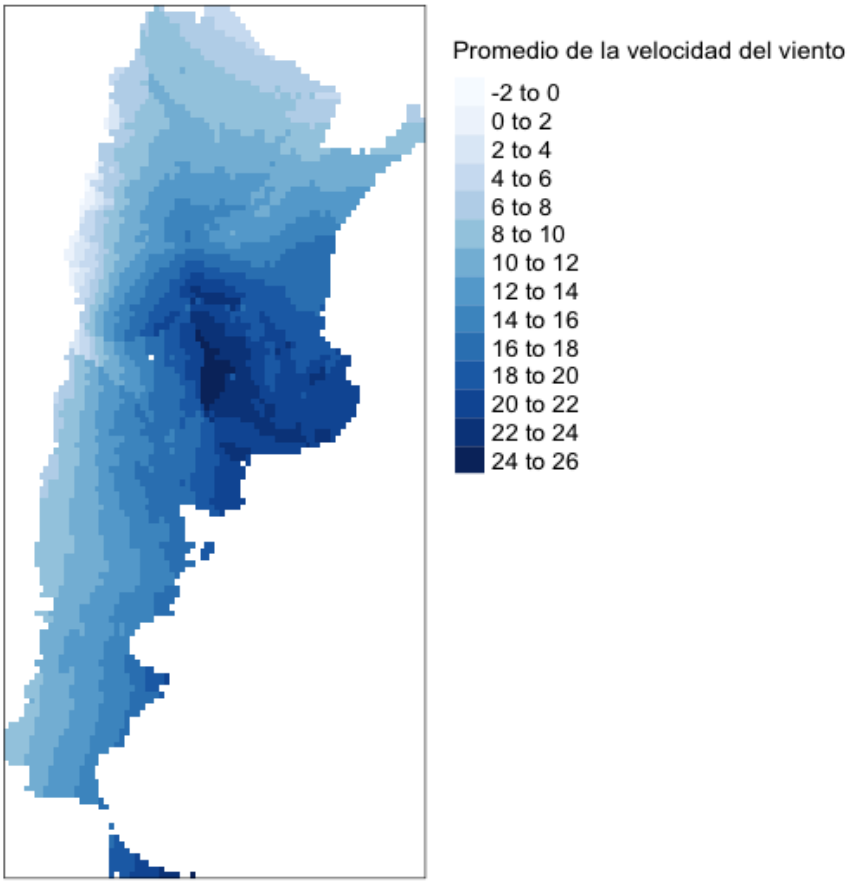
Mapas de predicción y varianza





Finalmente mediante un raster pusimos las predicciones en el mapa. Esto nos permite observar las áreas geográficas dentro de Argentina continental en donde el promedio diario de la velocidad del viento es mayor a 20km/h. La escala muestra como los últimos tres tonos más oscuros de azul cumplen con este requisito.

Mapa de predicción de velocidad de viento promedio



4 Conclusión

En primer lugar, y como suele ocurrir en estos casos, fue muy importante el tiempo dedicado a explorar y transformar (de ser necesario) los datos a utilizar. Dado que nuestro objetivo era buscar el lugar propicio para construir un parque eólico, hemos quitado las observaciones que no provienen de la plataforma continental del país, dada la inviabilidad fáctica de conectar la red troncal energética a un parque eólico construido en una isla. Asimismo, las observaciones de cada estación meteorológica, no capturan mediciones con la misma frecuencia. Por ejemplo, no se dispone de una medición de viento a la misma hora en todas las estaciones. Por ello, hemos realizado nuestro análisis considerando como valor de velocidad de viento de cada estación, su valor promedio del día. Como bien sabemos, era necesario que nuestra variable sujeta análisis (velocidad del viento) tenga una distribución normal. Sin embargo, como surge de la exploración visual y del test de Shapiro, dicha variable no cumple con dicha premisa.

Asimismo, dado que el proceso no es estacionario a causa que la media no es constante se utiliza el predictor Kriging Universal. Para elegir el modelo con mayor poder de predicción, ejecutamos cross validation y seleccionamos el matern con tendencia, dado que tenía mejores niveles de error.

Como conclusión final, y tal como surge del mapa, se observa la zona de la Pampa Húmeda y la zona costera del sur país incluyendo Tierra del Fuego, como los lugares óptimos para instalar un parque eólico, dado que presentarían una velocidad promedio de vientos mayor a 20 Km/H.